The Role of Genomic Data in Stratifying Patients within Predictive Models for Breast Cancer Survival Outcome

A Thesis submitted by
Lydia King

Supervisor: Dr. Emma Holian
Co-supervisors: Prof. Róisín Dwyer & Dr. Simone Coughlan

School of Mathematical and Statistical Sciences
University of Galway

Ollscoil na Gaillimhe
University of Galway

April, 2024

Table of Contents

Abstract

Genomic instability (GI), defined as an increased tendency for genomic alterations to occur, is a common feature of cancers and is recognised as a “facilitating” hallmark of cancer. Genomic alterations include base substitutions, indels, rearrangements and copy number alterations (CNAs). CNAs in cancer have been extensively profiled but due to the complexity of cancer genomes, frequent deviations from diploidy, i.e. having two sets of homologous chromosomes, and the presence of both tumour and non-tumour cells, many studies have been limited to reporting total copy number, the sum of the copy numbers of the two homologous chromosomes. Determining the CNA landscape of each homologous chromosome, i.e. allele-specific copy number, is important for the characterisation of certain genomic aberrations and the inference of their clonal history.

Breast cancer is largely dominated by CNAs, rather than mutations in a single gene, with increasing evidence suggesting that the genomic landscape of the tumour is associated with survival and incorporating this information into treatment decisions is beneficial to the patient.

This thesis uses total and allele-specific CNA data to explore the CNA landscape of breast tumours and their associations with survival. Focusing on observations from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort, we define novel metrics for total CNA measurements, estimating the distribution of these metrics allowing for missing values.

Analysing distributions of the CNA metrics comparing groups of patients stratified by molecular classifications indicates that subtypes associated with worse survival outcomes tend to have significantly higher levels of GI, and higher deletion burden, than subtypes associated with better survival outcomes.

Further investigation of these CNA metrics in the context of survival indicates that for molecular classifications displaying low levels of GI, the CNA metrics can partition patients based on survival outcome and aid in the identification of patients who may be more at risk. CNA metrics consistently selected as useful predictors for survival outcome include CNA metrics measuring the copy number deletion landscape, further indicating that deletions are more harmful than amplifications.

Differential gene expression analysis is carried out to investigate the effect that CNAs have on gene expression. Genes observed to be dysregulated in patients with decreased survival outcomes are known to facilitate cell proliferation, tumour progression and invasion. Investigating the direct relationship between a gene’s CNA state and its expression, using a modified limma pipeline, two differentially expressed gene sets are produced, with some degree of congruence observed when comparing to published predictive and prognostic assays and additional genes emerging as new focus.

Deriving allele-specific copy number profiles applying Allele-Specific Copy number Analysis of Tumours (ASCAT), models are proposed and assessed to identify and model features of changepoints in these profiles, including allele independent (AI) models and allele dependent (AD) models. Application of the AD models to defined intervals, including gene regions and genomic segments of specified length, identifies a number of gene and non-gene regions of interest.

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor Dr. Emma Holian. I could not have carried out this research and completed this thesis without your knowledge, dedication, guidance and unwavering support week in, week out. It has been a privilege working with you. I would also like to thank my co-supervisors, Prof. Róisín Dwyer and Dr. Simone Coughlan. I really appreciate all the guidance and support you've both given me. A massive thank you to Prof. Levi Waldron and Marcel Ramos, you both welcomed me with open arms and made my time in New York unforgettable.

To Prof. Aaron Golden, thank you for setting in motion my PhD, introducing me to Emma, and sometimes knowing what was best for me, even when I didn't know it myself. From the MSc to PhD, you have always had my back and for that I am eternally grateful.

Thank you to my GRC, Prof. Aaron Golden (again), Dr. Andrew Simpkin and Dr. Andrew Flaus for your support and advice each year.

To my fellow bioinformatics and statistics PhD students and colleagues, thank you for being welcoming, supportive and most importantly always up for laugh. I am forever grateful that I spent my PhD in such a positive environment.

Sarah and Shane, who have been there since day one, I would have been lost without both of you! Sarah, queen of data vis, gnome aficionado and most importantly my long-suffering housemate for much of the PhD, thank you from the bottom of my heart for all your love and support, for always agreeing to a bottle of wine or a spice bag, and for introducing me to Made in Chelsea, it's gotten me through some rough times! Shane, thank you for being someone who I could bounce anything off, someone who always listened and tried to understand, you've helped me in a million different ways, and for that I thank you immensely. I am so thankful we all ended up in ADB-1019 at the same time and I'm incredibly lucky to consider you both my friends.

To everyone in the CRT, particularly the CRT management, Dr. Sandra Healy and Prof. Cathal Seoighe, and those in the Galway CRT office, you've all played such a massive role in my PhD journey, and I am so lucky to have been part of the wonderful CRT community. It has been a privilege to work in an environment where everyone has a diverse skillset, a willingness to share their knowledge and an ability to recognise when it's time to call it a day and have some fun. A massive thank you to Anna, Sophie and Amanda, I really appreciate all the support you've given me, all the laughs we've shared and all the hugs that got me through the day. I could always rely on you for anything, and I hope you know what amazing, kind and generous people you are. Dónal and Micheál, thank you both for being a constant source of kindness, reassurance and laughter! I am going to miss all our random conversations, particularly about rowing, a topic I have significantly more knowledge about now than when I started the PhD. Siobhán, though COVID got in the way, I am so grateful to have had the opportunity to get to know you. It was such a joy to talk to someone who seemed to understand exactly where I was coming from, our coffee dates were just what I needed, and I already miss them terribly. A massive thank you to Anthony and Declan, without your help I'm not sure I would have finished this PhD, if you know, you know.

To all my friends, particularly Ciara, Ailbhe, and Kevin. Thank you for being by my side every step of the way, through the good times but also when things weren't going my way and you had to listen to my frustrated ranting. You all encouraged me to keep going and always left me with a smile on my face.

To my parents Clarissa and David, I can't thank you enough for everything you have done for me. Without your constant encouragement, support and love, I would not be who I am today. I know how lucky I am to have such amazing parents and I hope you know just how much I love you. Thank you to my siblings, Abbie and David. I credit much of this finished PhD thesis to you Abbie, your constant "are you not finished your thesis yet" or "I think you've been slacking" or "this is getting a bit ridiculous now" motivated me to finish, if only to make you stop.

To Galway, the place I have called home for the past five and a half years, it was so easy to fall in love with you and there are so many things I will miss including Charlie Byrne's Bookshop, Knocknacarra parkrun, Xian spice bags, the Prom, Bierhaus, Caribou and the Secret Garden.

Lastly, to Patsy, your house was the first place I lived in Galway, and you were the first friend I made, thank you for everything.

Declaration of Authorship

I hereby declare that this thesis titled, 'The Role of Genomic Data in Stratifying Patients within Predictive Models for Breast Cancer Survival Outcome' submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy is entirely my own work and I have acknowledged any assistance or contributions and cited the published work of others where applicable. The research contained within this thesis has emanated from research supported by a research grant from Science Foundation Ireland (SFI) and the National Breast Cancer Research Institute (NBCRI) under Grant number 18/CRT/6214. This work has not been submitted by me or another person for the purpose of obtaining any other degree.

Lydia King

Signature

08/04/2024

Date

List of Figures

List of Tables

1 Introduction

Breast cancer is a highly heterogeneous disease, meaning that there is a high degree of genotypic and phenotypic diversity within and between tumours, with much of this heterogeneity being attributed to the high frequency of mutations within the genome, also referred to as genomic instability (GI) (Duijf et al., 2019; Guo et al., 2023). Breast cancer classification and selection of treatment regimen are currently based on clinical and histopathological features (Dawson et al., 2013; Rakha et al., 2023), with recent research focused on utilising markers derived from genomics data to expand our understanding of the molecular mechanisms underlying breast cancer and to improve patient outcome by identifying patients who may not be well classified by the standard tissue-based biomarkers (Curtis et al., 2012; Dawson et al., 2013; Hamdan et al., 2019; Ochoa and Hernández-Lemus, 2023).

Copy number variations (CNVs) and copy number alterations (CNAs), forms of GI, are changes in the copy number of a DNA sequence in the form of either a gain or loss, occurring in germline and somatic cells, respectively (Shlien and Malkin, 2009; Ha and Shah, 2013; Luo, 2019). In the context of cancer, focus is primarily given to CNAs, their potential to initiate cancer through activation of oncogenes and inactivation of tumour suppressor genes, and their associations with disease progression and survival (Stephens et al., 2009; Pereira et al., 2016; Hieronymus et al., 2018; Smith and Sheltzer, 2018; Stopsack et al., 2019; Tao et al., 2023).

The aim of this thesis is to assess whether CNA information, in isolation or in combination with clinical and gene expression data, improves predictive models of overall survival (OS) and disease-specific survival (DSS) for breast cancer patients. We explore this by proposing several novel CNA metrics that quantify the levels of CNAs across the whole genome and across chromosome arms and assess the role the distribution of these metrics have in the context of OS and DSS. We go further, examining the role of the allele-specific CNA landscape, exploring statistical models to detect and model features of changepoints in allele-specific CNA profiles.

1.1 Breast Cancer in the Clinical and Research Setting

Breast cancer is one of the most common malignancies affecting women worldwide and is one of the leading causes of cancer related death among this group (Torre et al., 2017; Sung et al., 2021). Cancer that develops in breast cells typically forms in either the lobules (lobular carcinoma) or the milk ducts (ductal carcinoma). Cancer cells that remain in the milk ducts or lobules and do not grow into or invade normal tissues within or beyond the breast are termed non-invasive, also sometimes called carcinoma in situ ("in the same place") or pre-cancers. Invasive breast cancer, where there is spread of cancer cells outside of the ducts and lobules into the surrounding normal tissue, is most commonly observed in breast cancer patients (Libson and Lippman, 2014; Akram et al., 2017).

Tests used to diagnose breast cancer include mammograms, ultrasounds and biopsies (Bevers et al., 2009). Breast cancer classification and treatment generally follows an integrative approach whereby both clinical information and tissue-based biomarkers are used (Dawson et al., 2013; Russnes et al., 2017). These clinical and histopathological features include age, histological grade, tumour size, nodal status, oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) status, amongst others (Russnes et al., 2017; Rakha et al., 2023). Current classification of breast cancer in the clinical setting is based on immunohistochemical staining determining ER, PR and HER2 status, with measurement of ER, PR and HER2 being mandatory in all newly diagnosed breast cancer cases (Nicolini et al., 2018). Based on hormone receptors (HR), i.e. combinations of ER and PR, and HER2 positivity, patients are classified as HR+/HER2-, HR-/HER2-, HR+/HER2+ or HR-/HER2+ (Blows et al., 2010).

In published research, gene expression and CNA data have been used to produce molecular classifications of breast cancer along with a number of prognostic and predictive assays, providing information about likely survival outcome and response to therapy (Perou et al., 2000; Curtis et al., 2012; Nicolini et al., 2018). Molecular-based classifications, being evaluated in the research setting, but not yet common place in routine clinical use, include the Prediction Analysis of Microarray 50 (PAM50) intrinsic subtypes and Integrative Clusters (IntClust) (Perou et al., 2000; Sørlie et al., 2003; Curtis et al., 2012). PAM50 is a 50-gene signature that classifies breast cancer into five molecular intrinsic subtypes, Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like, that have been shown to have both prognostic and predictive power. Briefly, using complementary DNA (cDNA) microarrays on 65 breast cancer samples Perou et al. (2000) identified a subset of 496 genes whose variation in expression was significantly greater between samples from different tumours than between samples from the same tumour. Performing hierarchical clustering with this "intrinsic" gene subset resulted in the samples being split into four groupings related to different biological features (ER+/Luminal-like, Basal-like, HER2-enriched and Normal-like). Subsequently, using the same intrinsic gene set, Sørlie et al. (2001) performed hierarchical clustering on 85 breast tumour cDNA microarrays. In addition to classifying samples into Luminal-like, Basal-like, HER2-enriched and Normal-like groups, the Luminal-like group was further divided into at least two subgroups, each with a distinctive gene expression profile. Parker et al. (2009) later developed the 50-gene subtype predictor (PAM50) by performing gene set reduction on the 496 intrinsic genes (Perou et al., 2000; Sørlie et al., 2001), along with an additional 1,410 identified in three other microarray studies (Sorlie et al., 2003; Perreard et al., 2006; Hu et al., 2006). Claudin-low, a sixth subtype of breast cancer identified using gene expression data in a separate study (Herschkowitz et al., 2007; Prat et al., 2010), is also considered an intrinsic subtype (Fougner et al., 2020). The Cancer Genome Atlas Network (2012), integrating DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays, also observed the existence of four main breast cancer classes, each of which shows significant molecular heterogeneity in terms genetic and epigenetic alterations and are highly correlated with PAM50 subtype. IntClust derived from gene expression and CNA data classifies breast cancer into ten integrative clusters, IntClust 1-10, each with distinct CNA landscape, risk patterns and prognosis (Curtis et al., 2012). Prognostic biomarkers include ER, PR, HER2 and Ki67 status, Urokinase plasminogen activator/plasminogen activator inhibitor 1 (uPA/PAI-1), Oncotype DX, MammaPrint, Prosigna and Breast Cancer Index (BCI), while predictive biomarkers include ER status, PR status, HER2 status, deficiency in DNA damage response (DDR), mutational status of ER, amongst others (Mulligan et al., 2014; Nicolini et al., 2018).

1.2 Molecular Taxonomy of Breast Cancer International Consortium Data

The data used in this thesis are from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study (Curtis et al., 2012). The datasets, collected from five centres in the United Kingdom and Canada between 1977-2005, are well-annotated and contain clinical, transcriptomic and genomic data for approximately 2,000 breast cancer cases. The processed METABRIC datasets are publicly available from cBioPortal (http://www.cbioportal.org/study?id=brca_metabric) (Cerami et al., 2012; Gao et al., 2013). For the focus of this thesis, only the clinical, transcriptomic and CNA data are used.

The clinical data includes information on approximately 25 variables including age at diagnosis, Nottingham Prognostic Index (NPI), number of lymph nodes positive, tumour size, ER, PR and HER2 status, tumour stage, histological grade, PAM50 subtype (with Claudin-low) and IntClust classification, where IntClust 4 is split into IntClust 4ER+ and 4ER-, resulting in 11 IntClusts (Table 1.1). Figure 1.1 displays the distribution of PAM50 subtypes within each IntClust. Available treatments to this cohort of patients were hormone therapy, chemotherapy, radiotherapy and breast surgery, as summarised in Table 1.2. Enrolment of patients into the METABRIC study predated availability of trastuzumab (Herceptin). The survival outcomes recorded are OS, defined as the time from breast cancer diagnosis to death from any cause, DSS, defined as the time from breast cancer diagnosis to death from cancer, and recurrence-free survival (RFS), defined as the time from breast cancer diagnosis to relapse (Table 1.3). From the OS and DSS variables, 5- and 10-year OS and DSS are generated for each patient, e.g. if a patient experienced an event at 5 years and 1 month (61 months), then they would be censored (0) at 5 years (60 months). While most clinical variables are recorded for a large proportion of patients, some missingness exists.

Copy number for tumours observed in the METABRIC cohort were measured with the Affymetrix single-nucleotide polymorphism (SNP) 6.0 array, with pre-processing and quality control steps implemented by Curtis et al. (2012) to obtain $\log_2$ intensity values. For each tumour sample, the $\log_2$ ratio (the ratio between the observed $\log_2$ intensity value and the expected $\log_2$ intensity value) for each probe was calculated by subtracting a "normal" pooled reference, generated using the HapMap (International HapMap Consortium, 2003) and matched normal datasets, from the tumour sample $\log_2$ intensities. After computing the $\log_2$ ratios for each probe, Curtis et al. (2012) applied the circular binary segmentation (CBS) algorithm (Olshen et al., 2004; Venkatraman and Olshen, 2007), using DNAcopy (Venkatraman and Olshen, 2023), to each sample to detect changepoints and divide the genome into regions of equal copy number. CNAs were then called using selected thresholds for gains and losses across the whole genome, and genes affected by CNAs identified by gene annotation using hg18 (Curtis et al., 2012). The summary CNA data contains patient-specific somatic CNA calls for each of the 22,544 annotated genes and has values indicating homozygous deletion (-2), hemizygous deletion (-1), diploidy (0), single copy gain (+1) and high-level amplification (+2).

The transcriptomic data include $\log_2$ transformed and normalised gene expression and z-score data, measured with the Illumina HT-12v3 array. The log intensity z-score data contain information on the number of standard deviations away a gene's expression is from its mean expression across all profiled samples. This measure is useful to determine whether a gene in one patient's tumour sample is up or down-regulated relative to all other tumour samples.

Table 1.1: Selected clinical characteristics of the METABRIC patients.

Clinical Characteristics	N = 2,509[1]
Age	61 (51, 70)
NA	11
NPI	4.04 (3.05, 5.04)
NA	222
Lymph Nodes Positive	0 (0, 2)
NA	266
Tumour Size	22 (17, 30)
NA	149
ER Status	
Negative	644 (26%)
Positive	1,825 (74%)
NA	40
PR Status	
Negative	940 (47%)
Positive	1,040 (53%)
NA	529
HER2 Status	
Negative	1,733 (88%)
Positive	247 (12%)
NA	529
Tumour Stage	
0	24 (1.3%)
1	630 (35%)
2	979 (55%)
3	144 (8.1%)
4	11 (0.6%)
NA	721

[1] Median (IQR); n (%)

Clinical Characteristics	N = 2,509[1]
Histological Grade	
1	214 (9.0%)
2	976 (41%)
3	1,198 (50%)
NA	121
PAM50	
Basal	209 (11%)
Claudin-low	218 (11%)
HER2	224 (11%)
Luminal A	700 (35%)
Luminal B	475 (24%)
Normal	148 (7.5%)
NA	535
IntClust	
1	139 (7.0%)
2	72 (3.6%)
3	290 (15%)
4ER-	83 (4.2%)
4ER+	260 (13%)
5	190 (9.6%)
6	85 (4.3%)
7	190 (9.6%)
8	299 (15%)
9	146 (7.4%)
10	226 (11%)
NA	529

[1] Median (IQR); n (%)

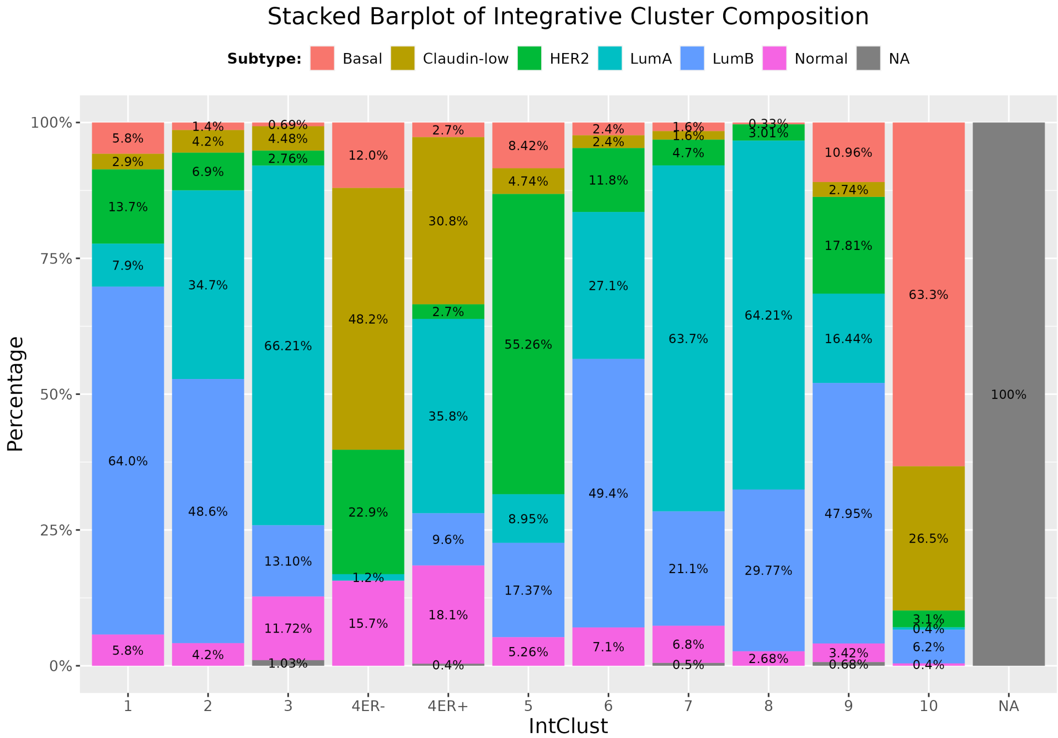


Figure 1.1: Stacked barplot indicating the PAM50 composition of each Integrative Cluster. The x-axis denotes the Integrative Clusters and the y-axis percentages.

Table 1.2: Treatment characteristics of the METABRIC patients.

Treatment Characteristics	**N = 2,509**[1]
Hormone Therapy	
No	764 (39%)
Yes	1,216 (61%)
NA	529
Chemotherapy	
No	1,568 (79%)
Yes	412 (21%)
NA	529
Radiotherapy	
No	807 (41%)
Yes	1,173 (59%)
NA	529
Breast Surgery	
Breast Conserving	785 (40%)
Mastectomy	1,170 (60%)
NA	554

[1] n (%)

Table 1.3: Survival characteristics of the METABRIC patients.

Survival Characteristics	**N = 2,509**[1]
Overall Survival	
Deceased	1,144 (58%)
Living	837 (42%)
NA	528
Disease-specific Survival	
Died of Disease	646 (33%)
Died of Other Causes	497 (25%)
Living	837 (42%)
NA	529
Recurrence-free Survival	
Not Recurred	1,486 (60%)
Recurred	1,002 (40%)
NA	21

[1] n (%)

These publicly available data are highly curated and periodically updated with additional information or datasets. Throughout this thesis the processed METABRIC data downloaded from cBioPortal are used “as is” and with the training and test sets defined in Curtis et al. (2012), combined. Unless otherwise stated the results discussed in this thesis are based on the processed METABRIC data downloaded from cBioPortal in 2021.

To obtain allele-specific CNA profiles, 1,992 Affymetrix SNP 6.0 CEL files available for 1,992 patients in the METABRIC study were accessed from the European Genome Phenome archive (study accession EGAS00000000083) (Curtis et al., 2012; Freeberg et al., 2022).

1.3 Structure of Thesis

Chapter 2 discusses CNAs as a measure of GI, with an introduction to published approaches for quantifying GI, and GI patterns in breast cancer. With application to the METABRIC cohort, novel CNA metrics are proposed, to measure individual patient CNA burden, accounting for the type, magnitude and location of the CNA. The distributions of the CNA metrics observed for the cohort are summarised, with an assessment of any effect from missing values. Distributions of CNA metrics are produced and summarised given location, e.g. global, or chromosome-arm specific, and are analysed comparing patients grouped by pre-defined breast cancer molecular classifications, such as PAM50 and IntClust.

Chapter 3 investigates whether there is an association between the CNA metrics and survival outcomes, within the METABRIC cohort. A number of parametric, semi-parametric and non-parametric survival models are applied. Applications of survival trees demonstrate splits of patients into classification nodes using the molecular classifications and clinicopathological variables, while introducing the proposed CNA metrics as candidate predictors.

Chapter 4 examines the effect of CNAs on gene expression, initially describing differential gene expression using limma and several expression-based predictive and prognostic assays for breast cancer. Differential gene expression analysis is carried out comparing gene expression between stratified groups of patients shown to have different survival outcomes, informed by models incorporating the CNA metric information. To finish, a comparative study is conducted, to compare the listing of differently expressed genes arrived at in this thesis, having incorporated CNA metric information, to the listings of prognostic and predictive gene sets previously derived and in use.

Chapter 5 focuses on allele-specific CNA profiles, CNA changepoints, and their identification and classification. The chapter reviews allele-specific copy number profiling using Allele-Specific Copy number Analysis of Tumours (ASCAT). Extraction of allele-specific CNA profiles using the PennCNV and ASCAT software is discussed and applied to the METABRIC cohort. Approaches to identify and model features of changepoints in allele-specific CNA profiles are proposed and include an extensive simulation study.

Chapter 6 details application of allele-specific models to the METABRIC data. The models are applied to defined intervals, corresponding to gene regions and whole genome segments, and the genes and segments identified as containing CNA changepoints of significant length examined in the context of survival.

2 Copy Number Alterations as a Measure of Genomic Instability

GI can be defined as an increased tendency for genomic alterations to occur. Genomic alterations, also termed genomic aberrations, include base substitutions, small insertions or deletions (indels), rearrangements, CNAs and even gain or loss of entire chromosomes and/or whole genome duplication (Kalimutho et al., 2019; Duijf et al., 2019). Two of the most well-characterised forms of GI are chromosomal instability (CIN) and microsatellite instability (MSI). CIN refers to changes in either chromosome number (numerical CIN), and/or structure (structural CIN), while MSI refers to the accumulation of mutations, usually point mutations or small indels, in microsatellite regions, i.e. regions of the genome displaying nucleotide repeats of about 1-6 bases in length (Kalimutho et al., 2019; Li et al., 2020).

GI can occur as the result of defects in mechanisms including DNA replication, DNA damage repair, transcription, mitotic chromosome segregation, and telomere maintenance (Lee et al., 2016; Kalimutho et al., 2019; Duijf et al., 2019). GI is a common feature of cancers and is recognised as a "facilitating" hallmark of cancer, enabling the activation of the eight functional hallmarks needed for tumour growth and progression (Hanahan, 2022). While the degree of GI is variable within and between cancer types, the GI profile of a tumour can be thought of as the accumulation of genomic alterations, which have the potential to promote oncogenesis, affect progression and influence patient prognosis (Lee et al., 2016; Kalimutho et al., 2019). In addition, the GI profile of a tumour can reflect the tumour's evolutionary history and future evolutionary potential (Pladsen et al., 2020).

In this chapter, patterns of GI in breast cancer and several measures of GI utilised in the literature are discussed, a number of metrics based on CNAs are proposed, the effect of missing values on the CNA metric distributions assessed, and the distributions of these CNA metrics within pre-defined breast cancer molecular classifications (PAM50 and IntClust) analysed.

2.1 Genomic Instability in Breast Cancer

Breast cancer genomes are often tetraploid (4n) or near-triploid (3n), commonly have distinct gene expression patterns and often display specific numerical and complex structural chromosomal aberrations (Duijf et al., 2019). The exercise of classifying tumour samples or patients into groups of homogeneous gene expression patterns or CNA profiles identifies distinct genomic alteration patterns in breast cancer and provides a number of breast cancer classifications (Russnes et al., 2017). Here we focus on patterns and classifications based on CNA landscape.

Hicks et al. (2006) described four distinct patterns of genomic alterations in breast cancer, termed "flat", "simplex", "complex I" and "complex II" (Figure 2.1). These patterns were identified using high-resolution comparative genome hybridisation arrays on 243 tumours selected from two breast cancer cohorts (140 samples were from the Cancer Center of the Karolinska Institute, while 103 were from the Oslo Micrometastasis study).

Tumours displaying the "simplex" pattern have large segments of duplication and deletion that usually span entire chromosome arms or even chromosomes (Figure 2.1A). Frequent copy number changes observed within tumours displaying the

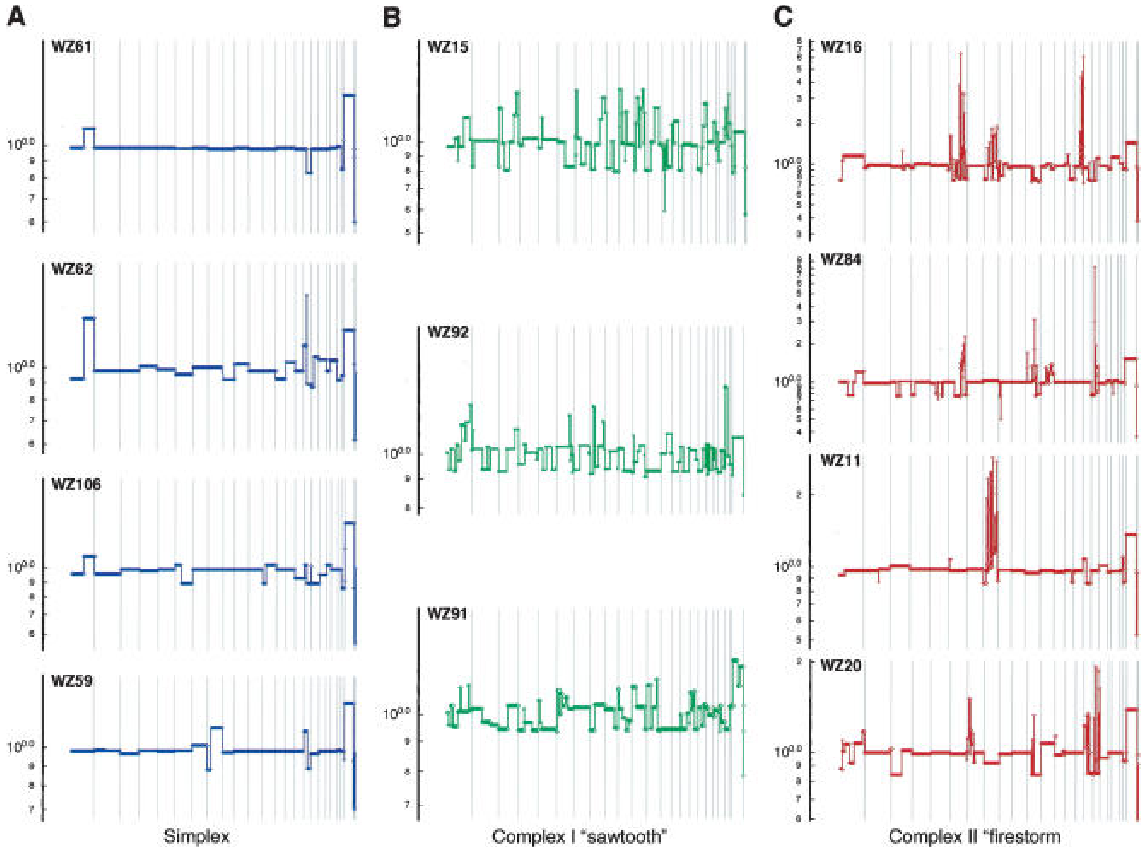


Figure 2.1: Distinct patterns of genomic rearrangements in breast cancer, taken from Hicks et al. (2006). Segmentation profiles for individual tumours representing each category: (A) "Simplex" pattern (B) "Complex I"/"sawtooth" pattern (C) "Complex II"/"firestorm" pattern. The x-axis denotes chromosomes 1-22, X and Y, ordered from left to right, and the y-axis displays the geometric mean value of two experiments on a log scale.

"simplex" pattern include gain of chromosomes 1q, 8q and/or 16p and loss of chromosomes 16q, 8p and/or 22. These tumours are usually ER+ and of the Luminal subtype. Tumours displaying the "complex I" pattern, also termed the "sawtooth" pattern, have complex patterns of narrow, low-amplitude gains and losses. These gains and losses usually span short chromosome regions and are often alternating, resulting in regions with many copy number transitions (Figure 2.1B). These events commonly affect all chromosomes and lead to the majority of the genome undergoing copy number changes. Recurring copy number changes observed within tumours displaying the "complex I" pattern include regions of gain on chromosome 10p, and regions of loss on chromosomes 3p, 4p, 4q, 5q, 14q, 15q, and 17q. These tumours are usually triple-negative (ER-/PR-/HER-) and correspond to the Basal subtype. Tumours displaying the "complex II" pattern, also known as the "firestorm" pattern, resemble the "simplex" pattern except that the tumours contain at least one localised region of clustered narrow peaks of amplification, i.e. each amplification cluster is restricted to an individual chromosome or chromosome arm. These regions of amplification are referred to as amplicons and are usually separated by regions displaying normal copy number or deletions (Figure 2.1C). Recurrently amplified sites include FGFR1, MYC, CCND1, MDM2, ERBB2 (HER2), and ZNF217.

These tumours are usually of the Luminal B and HER2 subtype. Tumours displaying the “flat” pattern have no clear amplifications or deletions except copy number polymorphisms (Hicks et al., 2006; Russnes et al., 2017).

To relate these patterns to clinical outcome, Hicks et al. (2006) developed the Firestorm Index. Using this metric, it was observed that the “complex I” and “complex II” patterns were associated with more aggressive disease and worse survival outcomes.

Morganella et al. (2016) provides a classification based on rearrangement signatures derived from 560 breast cancer whole genome sequences. Six rearrangement signatures (RS1-RS6) were created, based on whether the rearrangement was a deletion, tandem duplication, inversion, or translocation, the size of the rearrangement and also whether the rearrangements occurred in close proximity to each other. Interestingly, Russnes et al. (2017) noted that these signatures relate back to the classification described in Hicks et al. (2006). RS1 and RS3 are characterised by tandem duplications similar to the “complex I” or “sawtooth” pattern, RS4 and RS6 by clustered rearrangements similar to the “complex II” or “firestorm” pattern, RS5 by deletions, and RS2 by translocations, similar to the “simplex” pattern. For most tumours, the genomic landscape of rearrangements is composed of combinations of these signatures (Morganella et al., 2016; Russnes et al., 2017).

Curtis et al. (2012) used gene expression data along with CNA data, of 1,992 breast cancer samples from the METABRIC cohort, to identify ten distinct subtypes of breast cancer. Initially, using Analysis of Variance (ANOVA) and the Kruskal-Wallis test, they identified genes where the presence of a CNA influenced the expression of that gene, i.e. where overexpression is associated with copy number gain or amplification and underexpression with copy number loss. This method, by definition, captures genomic drivers, oncogenes and tumour suppressor genes whose expression is associated with copy number changes. The 1,000 most significant cis-driven genes, in terms of Bonferroni corrected p-values, were inputted as explanatory variables in a joint latent variable framework for integrative clustering. The most parsimonious solution, with reference to copy number profiles, risk patterns, and prognosis, classified tumours into ten distinct groups (IntClust 1-10, Figure 2.2).

2.2 Measures of Genomic Instability

To explore the impact of GI in cancer, a number of genomic and transcriptomic signatures have been created to quantify levels of GI in tumours, and their prognostic and predictive power assessed. These metrics, summarised in Table 2.1, are described below.

2.2.1 Expression Based Signature CIN25 and CIN70

It has been well documented that correspondence exists between gene expression changes and CNAs in regions relevant to those genes (Pollack et al., 2002; Stranger et al., 2007; Curtis et al., 2012; Bhattacharya et al., 2020). Carter et al. (2006) derived two expression-based signatures, reflecting CIN in tumours, termed CIN25 and CIN70, using 25 and 70 genes, respectively. These signatures were developed using integrated gene expression data from 18 studies, across nine cancer types, totalling 1,944 samples. These signatures are based on a functional aneuploidy measure (FA) calculated across cytobands, i.e. genomic regions corresponding to

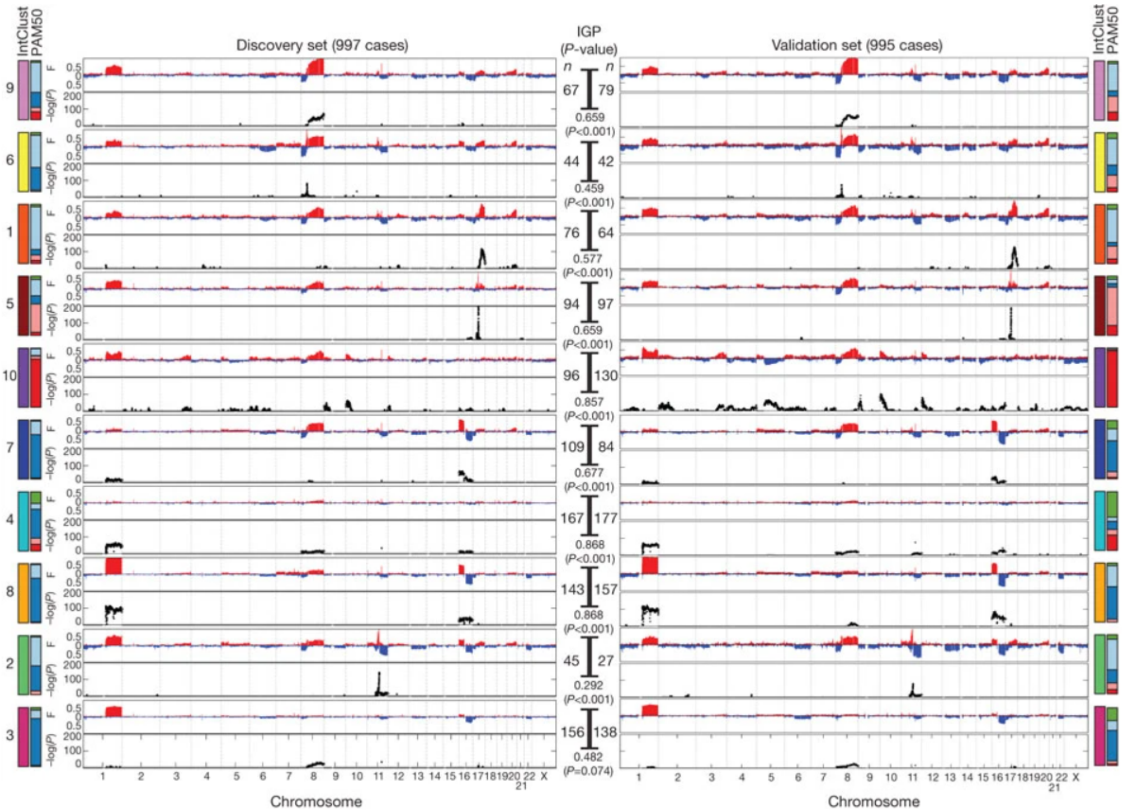


Figure 2.2: Distinct copy number profiles of the Integrative Clusters, taken from Curtis et al. (2012). Frequencies of CNAs are displayed on the upper y-axis of each section and the subtype-specific association (-log10 p-value) of aberrations is displayed on the bottom y-axis. Regions of copy number gain are indicated in red and regions of loss in blue. The distribution of PAM50 subtypes within each cluster is also shown.

Table 2.1: Summary of existing measures of Genomic Instability.

GI Measure	Cancer Type(s)	Input Data	Platform(s) Used in Study	Author
CIN25 and CIN75	Breast cancer, lung adenocarcinoma, small-cell lung cancer, mesothelioma, prostate cancer, B-cell lymphoma, ovarian cancer, glioma, medulloblastoma	Gene expression data	Affymetrix Human Genome U133A microarray Affymetrix Human Genome U133+2 microarray Rosetta 25k microarray	Carter et al. (2006)
Chromosomal Instability Score	Breast cancer	Copy number data	Affymetrix GeneChip Mapping 100K microarray	Smid et al. (2011)
Centromere and Kinetochore Gene Expression Score	Breast, lung, ovarian, liver, pancreatic, colon, nasopharyngeal, gastric, cervical, head and neck, prostate, brain.	Gene expression data	Affymetrix Human Genome U133+2 microarray	Zhang et al. (2016)
Chromosomal Instability Index	Colorectal cancer	Copy number data	Affymetrix Genome-wide Human SNP array 6.0	Song et al. (2017)
Whole Arm Aberration Index and Complex Arm-Wise Aberration Index	Breast cancer	Copy number data	Custom ROMA 85k microarray Agilent Human Genome CGH 244K microarray Custom Human 30K 60-mer oligo microarray	Russnes et al. (2010)
Firestorm Index	Breast cancer	Copy number data	Custom ROMA 85k microarray	Hicks et al. (2006)
Copy Number Alteration Burden	Prostate cancer, breast cancer	Copy number data	Agilent Human CGH Whole Genome microarray Affymetrix Genome-wide Human SNP array 6.0	Hieronymus et al. (2014) Hieronymus et al. (2018) Zhang et al. (2018)
Copy Aberration Regional Mapping Analysis Scores	Breast cancer	Copy number data	Affymetrix Genome-wide Human SNP array 6.0	Pladsen et al. (2020)
Genomic Instability Index	Breast cancer	Copy number data	Custom Human 30K 60-mer oligo microarray	Chin et al. (2007)
Genomic Identification of Significant Targets in Cancer	Glioma	Copy number data	Affymetrix Human Mapping 50K Xba240 SNP array Affymetrix Human Mapping 50k Hind240 SNP array	Beroukhim et al. (2007)

the approximate location of bands seen on Giemsa-stained chromosomes. For a given dataset, a cytoband specific t-statistic compares normalised gene expression measurements mapped to a specific cytoband, group B, to the normalised gene expression measurements for the genes mapped to all other cytobands, group G:

$$t = \frac{\mu_B - \mu_G}{\sqrt{(\frac{\sigma_B^2}{N_B}) + \frac{\sigma_G^2}{N_G})}} \tag{2.1}$$

where μ_B, μ_G are the observed means, σ_B^2, σ_G^2, are the observed variances, and N_B, N_G, the number of genes, for groups B and G.

The total FA (tFA) for each sample, within each of the 18 datasets, was defined as the sum of all FA magnitudes (the absolute t statistics), across each cytoband with more than 10 genes recorded, in that sample. For all genes within each dataset, the correlation coefficient across all samples between each gene's expression vector (the vector containing that genes expression for all samples in that dataset) and the tFA vector (the vector containing the tFA for each sample in that dataset) was computed. Genes in each dataset were then ranked based on the value of the correlation coefficient. Following normalisation of ranks within each dataset, the total of the ranks of a gene within three selected datasets was used as the final integrated ranking for the gene. The top 25 and 70 genes from this ranking formed the CIN25 and CIN70 signature, respectively.

tFA was found to be significantly correlated with aneuploidy assessed using CNA profiles and structural chromosomal aberrations from spectral karyotyping on NCI-60 cell lines. Furthermore, the CIN25 and CIN70 genes showed significant deviation in their expression relative to the remainder of the transcriptome and were enriched for regulators of mitotic spindle assembly, the mitotic checkpoint, and the DNA damage checkpoint.

To explore the prognostic power of CIN25 and CIN70, patients were split into two groups, patients with total expression, i.e. sum of the log-ratio measures, above the mean signature expression, and patients below the mean signature expression, in all samples from that dataset. This indicated that the CIN25 signature was a significant predictor of clinical outcome in 12 out of 18 cancer datasets and the CIN70 signature was a significant predictor of clinical outcome in 13 out of 18 datasets. Comparing the CIN25 and CIN70 signatures between primary and metastatic tumours also indicated metastatic samples display higher levels of the CIN signatures compared to primary tumours.

2.2.2 Chromosomal Instability Score

Smid et al. (2011) used SNP copy number data from 313 primary lymph-node negative breast cancers to study the prognostic relevance of CIN within breast cancer subtypes. In this study, by measuring the loss, gain, or diploid status of SNPs within 100-kilobase (kb) genomic windows, a measure for CIN was defined as the total number of chromosomal segments showing a gain or loss. Hierarchical clustering of patients using this CIN metric identified four main groups showing varying degrees of chromosomal abnormalities. In addition, it was found that high CIN score was significantly associated with worse prognosis in ER+, Luminal B, and HER2 subtypes, but not in ER- patients.

2.2.3 Centromere and Kinetochore Gene Expression Score

Centromeres and kinetochores play essential roles in cell division and their protein level is usually tightly regulated (Allshire and Karpen, 2008). Their dysfunction can result in a number of misregulation effects, including missegregation and mislocalisation to non-centromeric chromatin, generating neo-centromeres, dicentric behaviour and chromosome bridges, that drive aneuploidy and CIN (gains and losses) (Allshire and Karpen, 2008; Zhang et al., 2016). To capture this misregulation, Zhang et al. (2016) developed the centromere and kinetochore gene expression score (CES) that quantifies the misexpression of 14 centromere and kinetochore genes in cancers. To arrive at this scoring mechanism, expression profiles of 31 candidate centromere and kinetochore genes were analysed, 15 of these genes were observed to be significantly misregulated and of these, 14 were found to be associated with poor patient survival and correlated with cancer progression, in an analysis of 18 different cancer datasets from The Cancer Genome Atlas (TCGA). CES is calculated as the sum of the $\log_2$ mRNA expression level of the 14 centromere and kinetochore genes. It was shown that high CES significantly correlated with increased CIN and accurately predicts patient outcome in terms of OS, distant metastasis-free survival and relapse-free survival. This study also reported that high CES cell lines were sensitive to genotoxic drugs, such as campthothecin, topotecan and irinotecan.

2.2.4 Chromosomal Instability Index

The CIN index is a measurement that quantitatively characterises genome-wide CNAs. The CINdex algorithm uses segmented copy number data to calculate global measures of GI across chromosomes and at a higher resolution across cytobands. The first step in calculating CIN index involves calling segments as either gain or loss. A segment with mean signal intensity greater than an assigned threshold, t_{gain}, is called as a gain, whereas a segment with mean signal intensity smaller than an assigned threshold, t_{loss}, is called as a loss. In Song et al. (2017) the biologically experimental values of t_{gain} and t_{loss} are 2.5 and 1.5, respectively. Subsequently, the amplitude of change is scaled to make maximal losses and maximal gains comparable in magnitude. To do this, the amplitude of each loss segment, a, is converted to the new value, a', based on the relationship given by:

$$(t_{loss} - a)/a = (a' - t_{gain})/(A - t_{gain}) \tag{2.2}$$

where A is maximum gain amplitude across all samples and segments and t_{loss} and t_{gain} are the assigned thresholds for calling losses and gains, respectively.

The chromosome-specific instability index for each sample is calculated using:

$$CIN_i = (\sum_k a_k + \sum_j a'_j)/N \tag{2.3}$$

where N is the number of SNP probes on chromosome i, a is the amplitude of gain segments and a' is the amplitude of loss segments.

Applying the same calculation at the cytoband level provides the cytoband-specific instability index. The CINdex Bioconductor package (Song et al., 2022) implements this algorithm and generates a chromosome and cytoband CIN value for each sample. The package also enables comparison of CIN index values between groups of patients to identify differentially altered chromosomes or cytobands. Genes within these differentially altered regions can then be identified and pathway enrichment performed.

2.2.5 Whole Arm Aberration Index and Complex Arm-Wise Aberration Index

Russnes et al. (2010) developed two algorithms to characterise levels of genomic distortion using array comparative genomic hybridisation (aCGH) data. These algorithms are termed the Whole Arm Aberration Index (WAAI) and the Complex Arm Aberration Index (CAAI), where WAAI aims to capture whole-arm deviations from normal copy number, i.e. whole-arm gains/losses, and CAAI aims to capture the degree of local distortion i.e complex rearrangements.

WAAI is calculated across each chromosome arm for each sample. The first step in generating the WAAI values is to use the Piecewise Constant Fitting (PCF) algorithm to fit a piecewise constant regression function to the log-transformed aCGH data for each sample. As a result, a fitted value, termed "PCF-value", is obtained for each probe. The centred PCF-values were then divided by the residual standard deviation to produce normalised PCF (NPCF)-values and a new variable s was obtained by averaging the NPCF-values over all probes. If $s > 0$, WAAI was the 5% quantile of NPCF and if $s \leq 0$, WAAI was the 95% quantile of NPCF. Chromosome arms with WAAI ≥ 0.8 were called as whole-arm gains, and chromosome arms with WAAI ≤ -0.8 were called as whole arm losses.

CAAI is also calculated across each chromosome arm for each sample. In the original paper a threshold of 0.5 was applied to create a two-category CAAI variable, whereas it is possible to use the CAAI as a continuous variable (Pladsen et al., 2020). The first step in generating the CAAI variable is to use the PCF algorithm to fit a piecewise constant regression function to the log-transformed aCGH data for each sample. Then for each breakpoint (chromosomal position affected by rearrangements) identified by PCF, three scores, P, Q and W, were calculated. To produce the CAAI variable from the original paper, P, Q and W are defined as follows:

$$P = tanh\left(\frac{\alpha}{L1 + L2}\right) \tag{2.4}$$

$$Q = tanh(|H2 - H1|) \tag{2.5}$$

$$W = 0.5\left[1 + \frac{tanh(10(P - 0.5))}{tanh(5)}\right] \tag{2.6}$$

where α is a constant. For any given breakpoint, $L1$ and $L2$ denote the number of nucleotides in each segment and $H1$ and $H2$ denote their scaled PCF-values.

Pladsen et al. (2020), proposed a refined version of the CAAI variable, where P, Q and W are defined as:

$$P = tanh\left(\frac{\alpha}{L1 + L2}\right) \tag{2.7}$$

$$Q = tanh(\beta \cdot |H1 - H2|) \tag{2.8}$$

$$W = 0.5\left[1 + \frac{tanh(10P - 5)}{tanh(5)}\right] \tag{2.9}$$

where α and β are constants $2{\cdot}10^6$ and $\frac{1}{1.2}$, respectively. For any given breakpoint, $L1$ and $L2$ denote the size of the segments joined and $H1$ and $H2$ denote their height, i.e. total copy number.

These three scores P, Q and W, reflect the proximity to neighbouring breakpoints, the magnitude of change and a weight of importance. Subsequently, CAAI is defined as the maximal value of $\sum W \cdot min(P, Q)$ across all breakpoints within a region of predefined size, i.e. 20 megabases (Mb).

Applying CAAI and WAAI to data from 595 breast cancer patients from four clinical cohorts (MicMa cohort, WZ cohort, Chin-UCAM cohort and Ull cohort), patients were split into eight subgroups each with distinct patterns of genomic alterations. CAAI was observed to be highly prognostic for DSS and OS in breast cancer. In addition, CAAI also correlates with expression-based prognostic signatures including MammaPrint and OncotypeDX. Subsequently, Vollan et al. (2015) validated CAAI as an independent prognostic indicator in breast cancer and also showed that CAAI could act as a prognostic indicator in high-grade serous ovarian cancer.

2.2.6 Firestorm Index

Hicks et al. (2006) noted that the "complex I"/"sawtooth" and "complex II"/"firestorm" patterns often correlated with aggressive disease and worse survival in diploid tumours. To confirm this, the authors created a metric which separates the highly rearranged "complex I"/"sawtooth" and "complex II"/"firestorm" from the "flat" and "simplex" patterns. To distinguish the "complex II"/"firestorm" pattern from the "simplex" pattern this metric considered both the tightly packed spacing of the firestorm events and the total number of events. This metric, termed the Firestorm index (F), is obtained by summation across the reciprocals of the mean of lengths of all adjacent segment pairs:

$$F = \sum_i \frac{2}{l_i^L + l_i^R} \tag{2.10}$$

where i corresponds to the set of all discontinuities or breaks with a magnitude above the threshold of 0.1, l_i^L and l_i^R correspond to the number of probes in the nearest discontinuity to the left or right, respectively, or to a chromosome boundary, whichever is closer.

This metric can distinguish the "complex II"/"firestorm" pattern from the "simplex" pattern and assigns high F values to the complex patterns. The "complex I"/"sawtooth" pattern will have a high F value as a result of the high number of events across a large number of chromosomes, while the "complex II"/"firestorm" pattern will have a high F value due to the sparse events occurring in close proximity. Hicks et al. (2006) also reported a strong association between F and survival outcomes.

2.2.7 Copy Number Alteration Burden

CNA Burden is defined as a measure of the percentage of the genome affected by CNAs, calculated as the summation of the lengths of all CNA (gain and loss) segments as a percentage of the total length of the autosomal genome.

A number of studies have reported an association between CNA burden and recurrence, metastasis, OS and DSS (Hieronymus et al., 2014, 2018; Zhang et al., 2018). Hieronymus et al. (2014) showed that CNA burden is prognostic for prostate cancer recurrence and metastasis, Hieronymus et al. (2018) observed that CNA burden is associated with both OS and DSS in a range of cancers including breast, endometrial, renal, thyroid, and colorectal cancer and Zhang et al. (2018) showed that there is a significant association between CNA burden and OS and DSS in breast cancer cohorts.

2.2.8 Copy Aberration Regional Mapping Analysis Scores

Pladsen et al. (2020) developed the Copy Aberration Regional Mapping Analysis (CARMA) algorithm which identifies multiple local copy number features, or "motifs", across a pre-defined region and combines these to create regional metrics. CARMA takes allele-specific copy number profiles as inputs and produces six metrics that aim to capture the degree of amplification (AMP), deletion (DEL), complexity, i.e. chromothripsis and chromoplexy (STP and CRV), loss of heterozygosity (LOH) and allelic imbalance or asymmetry (ASM). Together, these metrics consider copy number magnitude, the spatial distribution of copy number breakpoints, allelic imbalance and regional fluctuations in copy number.

These scores are defined using continuous functions on genomic loci, i.e. positions on a chromosome, $t_1, \ldots, t_i$, over a region R. Here, $f(t)$ is the median centred total copy number in locus t and is calculated by $f(t) = f_A(t) + f_B(t) - m$, where $f_A(t)$ and $f_B(t)$ are piecewise constant functions representing the allele-specific copy number profiles of the major allele and minor allele, respectively, and m is chosen as the median observed copy number.

The degree of amplification AMP is defined as:

$$AMP = \int_R \{f(t)_+\}^2 \tag{2.11}$$

where $f(t)_+$ corresponds to the regions where the median centred total copy number is greater than 0. Alternatively, one can think of this metric as $AMP = \sum L_+ \times H_+^2$, where L_+ is a vector containing the scaled lengths of segments where an amplification is present, relative to the median copy number, and H_+ is a vector containing the corresponding copy number magnitudes. AMP will take value 0 where the total copy number is equal to the median copy number and greater than 0 when there are some gains and no losses relative to the median.

Similarly, the degree of deletion is defined as:

$$DEL = \int_R \{f(t)_-\}^2 \tag{2.12}$$

where $f(t)_-$ corresponds to the regions where the median centred total copy number is less than 0. DEL will take value 0 where the total copy number is equal to the median and greater than 0, where there are some losses and no gains relative to the median.

The complexity scores are defined as:

$$STP = \int_R \{Df(t)\}^2 dt \tag{2.13}$$

$$CRV = \int_R \{D^2 f(t)\}^2 dt \tag{2.14}$$

where $Df(t)$ is the first derivative and $D^2f(t)$ is the second derivative, reflecting the change in total copy number and the oscillation in total copy number, respectively. STP will take value 0 where there is constant total copy number and greater than 0 where there is gradually increasing or decreasing copy number, or where there are fluctuations between smaller and larger copy numbers. CRV will take value 0 where the copy number is constant, be close to 0 where there is gradually increasing or decreasing copy number and greater than 0 where there are fluctuations between smaller and larger copy numbers.

Loss of heterozygosity is defined as:

$$LOH = \int_R \{1_0(f_B(t))\} dt \tag{2.15}$$

where $f_B(t)$ is the piecewise constant function representing the copy number profile of the minor allele, 1_0 is an indicator variable informing whether or not the minor allele is lost. If the copy number of the minor allele is 0 at locus t then $1_0 = 1$, otherwise $1_0 = 0$. Alternatively, one can think of this metric as $LOH = \sum L[minor = 0]$, where L is a vector containing the scaled lengths of segments where the minor allele has been lost, i.e. copy number is 0. LOH takes a value greater than 0 where the minor allele has been lost, with the magnitude of the metric reflecting the proportion of the region with LOH.

Allelic imbalance or asymmetry is defined as:

$$ASM = \int_R \{(f_A(t) - f_B(t))^2\} dt \tag{2.16}$$

where $f_A(t)$ and $f_B(t)$ are the piecewise constant functions representing the copy number profile of the major allele and the minor allele, respectively, so that ASM > 0 in regions of allelic imbalance.

The authors applied CARMA to four breast cancer cohorts, METABRIC (n = 1,943), Oslo2 (n = 276), OsloVal (n = 165), and ICGC (n = 553). To standardise the scores, all six scores were $\log_2$-transformed and normalised by dividing by the 99th percentile in the METABRIC discovery set. The authors showed that the CARMA metrics correlated with the CAAI and CINdex metrics and provided significantly more detail about the copy number profile, enabling identification of alterations that may not be captured by the other methods. For example, in a region where there is loss of one allele and gain of the other, the CINdex would indicate that no alteration has occurred, whereas the LOH and ASM metrics would capture this event. The authors also considered the distribution of CARMA scores within breast cancer subtype classifications (PAM50 and IntClust) and noted that the CARMA scores captured differences in the genomic landscapes of the distinct subtypes.

Examining whether the CARMA metrics were significantly associated with survival outcome, univariate Cox proportional hazards regression models were fitted for each metric, and these models indicated that all CARMA metrics were associated with DSS. To assess if the presence or absence of the copy number motifs was significantly associated with survival outcome, the information provided by all six CARMA metrics was combined into two prognostic indices, the CARMA Prognostic Index (CPI) and the weighted CPI ($\text{CPI}_{\text{weighted}}$). Briefly, using a discovery set and test set from the observed cohort, the CPI index was produced by fitting multivariate Cox regression models for DSS outcome and progression-free survival outcome, using the unweighted mean of the six CARMA metrics as predictors. The fitted model was then applied to the test set, producing a single unweighted prognostic value for each patient in the test set.

Patients were stratified into low, intermediate, and high-risk groups, of equal cohort size, based on their CPI value. These tertile groups are defined as having CPI score, 1, 2, and 3, respectively. The $\text{CPI}_{\text{weighted}}$ score was produced using the 252 arm-wise CARMA scores directly as predictors and fitting a Cox regression model with LASSO penalty to the discovery set. Coefficients derived from the model were then used as weights to calculate the $\text{CPI}_{\text{weighted}}$. Both CPI and $\text{CPI}_{\text{weighted}}$ were shown to be significantly associated with DSS and progression-free survival before and after adjusting for other relevant clinical variables. Patients in the high-risk CPI group displayed significantly worse DSS and progression-free survival.

2.2.9 Genomic Instability Index

The Genomic Instability Index (GII) is defined as the fraction of the genome with CNAs. Chin et al. (2007) proposed two GII metrics, calculated based on the fraction of the genome that was altered using common regions of alteration (CRA), regions that were altered in at least 5% of tumours, and on the fraction of altered probes. As expected, a very strong correlation was observed between the two GII metrics (Spearman rank correlation 0.96). Using hierarchical clustering, on the CRA from 171 primary breast tumours, the authors identified a novel subtype of high-grade ER-breast cancer, characterised by a low GII. With this index the authors documented regions across the breast cancer genome that frequently contain CNAs and have corresponding dysregulated expression. Furthermore, they identified regions of the genome that were frequently amplified and correlated with poor prognosis, with some of these regions not previously identified.

2.2.10 Genomic Identification of Significant Targets in Cancer

Beroukhim et al. (2007) introduced the Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm which differs from the previously mentioned GI measures/algorithms in that it identifies regions within the genome that are significantly altered across multiple samples. GISTIC produces multiple outputs, including a categorical value (0, 1 or 2) of aberration for each region and each sample. The GISTIC algorithm first assigns a score (G score) to each aberration, which reflects the aberration amplitude and the frequency with which the aberration occurs across samples, the significance of each aberration is assessed using permutation tests based on the overall pattern of aberrations observed across the genome. Regions with false discovery rates below a given threshold are declared to be significant aberration regions. For each significant aberration region, GISTIC defines a "peak region", containing the highest frequency and amplitude of aberration, and determines whether the signal is due to broad events, focal events, or both. GISTIC has been applied to multiple cancer types and has identified a number of new targets of deletions and amplifications such as EHMT1 in medulloblastoma and CDK8 in colorectal carcinoma (Mermel et al., 2011).

In subsequent years, the GISTIC algorithm underwent a number of methodological improvements resulting in GISTIC2.0 (Mermel et al., 2011). These improvements address challenges relating to modelling of complex cancer genomes that contain a mixture of CNA types occurring at distinct background rates and the ability of copy-number algorithms to provide a priori statistical confidence.

2.2.11 Summary

As discussed, a large number of measures to quantify GI in tumours exist in the literature. All these measures, except for CES, use regions of altered copy number as a measure of CIN and should, in theory, be comparable when dealing with simple copy number deviations. For more complex copy number patterns, such as copy-neutral loss of heterozygosity, certain measures perform better, i.e. CARMA LOH and ASM metrics capture the event, while CINdex does not. These measures derived and evaluated using array-based data, aCGH and SNP arrays, or whole genome sequencing data, are limited in their accessibility and use as access to raw or segmented array/whole genome sequencing data is required. In addition, they can often be hard to interpret. As a result, we aim to create easily interpretable GI measures that can be calculated using publicly available summary CNA data. These CNA metrics aim to capture the main aspects of CNAs, including magnitude, type and genomic location.

2.3 Proposed Copy Number Alteration Metrics

The following CNA metrics, termed CNA Score and Burden metrics, aim to quantify levels of GI from summary CNA data and consider the magnitude of the CNA and if the CNA is an amplification or deletion. The CNA calls range from -2 to +2 indicating homozygous deletion (-2), hemizygous deletion (-1), diploidy (0), single copy gain (+1) and high-level amplification (+2).

2.3.1 Copy Number Alteration Score Metrics

Absolute CNA Score (Equation 2.17) for a sample or patient is the summation across all genes, $g \in 1 : G$, of the absolute magnitudes of all calls, irrespective of type, while CNA Amp Score and CNA Del Score, capture the total magnitudes of amplifications only (Equation 2.18) and the total magnitudes of deletions only (Equation 2.19). The Difference Score (Equation 2.20) measures the difference between the magnitudes of CNA Amp Score and CNA Del Score. The last two proposed CNA Score metrics measure the percentage of a patient's total CNA Score that is classified as amplifications (Equation 2.21) and the percentage classified as deletions (Equation 2.22). Notably, the percentage CNA metrics are correlated with each other, i.e. if for a patient the percentage amplified is 80%, then the percentage deleted will be 20%.

$$\text{Absolute CNA Score} = \sum_{g=1}^{G} \mid CNA\ call_g \mid \tag{2.17}$$

$$\text{CNA Amp Score} = \sum_{g=1}^{G} \mid CNA\ Amp\ call_g \mid \tag{2.18}$$

$$\text{CNA Del Score} = \sum_{g=1}^{G} |\, CNA\ Del\ call_g \,| \tag{2.19}$$

$$\text{Difference Score} = CNA\ Amp\ Score - CNA\ Del\ Score \tag{2.20}$$

$$\text{Percentage Amp Score} = \frac{CNA\ Amp\ Score}{Absolute\ CNA\ Score} \times 100 \tag{2.21}$$

$$\text{Percentage Del Score} = \frac{CNA\ Del\ Score}{Absolute\ CNA\ Score} \times 100 \tag{2.22}$$

2.3.2 Copy Number Alteration Burden Metrics

Further, we propose calculation of several CNA Burden metrics, measured for each patient. It is important to note that our CNA Burden metrics (Equations 2.23-2.28) differ from the CNA Burden metric (Hieronymus et al., 2014) mentioned in Section 2.2.7. While both metrics aim to measure the percentage of the genome affected by CNAs, our CNA Burden metric uses publicly available gene level summary CNA data to calculate the percentage of genes containing an alteration, whereas the pre-existing metric uses the CNA segment lengths obtained from segmented CNA data to calculate the percentage of the genome affected by CNAs. Therefore, the focus here is on the presence of a CNA for each gene, while the CNA Burden metric utilised by Hieronymus et al. (2014) focuses on the lengths of these altered segments in relation to the total length of the autosomal genome. The proposed CNA Burden metric here also differs from the proposed CNA Score metric in several ways, including considering the presence or absence of a CNA rather than the magnitudes, and the scale and range of measurement, i.e. summation versus percentage.

Absolute CNA Burden (Equation 2.23) reflects the percentage of genes recorded containing an alteration. Similarly, the CNA Amp Burden metric (Equation 2.24) and CNA Del Burden metric (Equation 2.25) capture the percentage of genes containing an amplification and deletion, respectively. The Difference Score (Equation 2.26) measures the difference between the CNA Amp Burden and CNA Del Burden. The last two proposed CNA Burden metrics measure the percentage of a patient's total CNA Burden that is classified as amplifications (Equation 2.27) and the percentage classified as deletions (Equation 2.28).

$$\text{CNA Burden} = \frac{\sum_{g=1}^{G} Alt_g}{G} \times 100 \tag{2.23}$$

$$\text{CNA Amp Burden} = \frac{\sum_{g=1}^{G} AltAmp_g}{G} \times 100 \tag{2.24}$$

$$\text{CNA Del Burden} = \frac{\sum_{g=1}^{G} AltDel_g}{G} \times 100 \tag{2.25}$$

$$\text{Difference Burden} = CNA\ Amp\ Burden - CNA\ Del\ Burden \tag{2.26}$$

$$\text{Percentage Amp Burden} = \frac{CNA\ Amp\ Burden}{CNA\ Burden} \times 100 \tag{2.27}$$

$$\text{Percentage Del Burden} = \frac{CNA\ Del\ Burden}{CNA\ Burden} \times 100 \tag{2.28}$$

Alt corresponds to the alteration status (0 or 1) for each gene g, $AltAmp$ corresponds to the amplification status (0 or 1) for each gene g, and $AltDel$ corresponds to the deletion status (0 or 1) for each gene g.

2.4 Application of CNA Metrics to the METABRIC Cohort

These CNA Score and Burden metrics are calculated for all breast cancer patients in the METABRIC cohort for which CNA data were available (n = 2,173). The metrics are calculated globally, i.e. over all 22,544 genes recorded, and, more locally, for each of the 42 chromosome arms, to account for the genomic location of the CNA. It should be noted that chromosomes differ in length and number of genes, with chromosome 1 being the longest autosomal chromosome and chromosome 22 being the shortest autosomal chromosome, meaning the CNA Score metrics are not comparable across chromosomes.

2.4.1 Observed Distributions for Global CNA Metrics

The observed distributions of the global CNA Score and Burden metrics are explored and summarised in Tables 2.2-2.5, along with density plots and histograms (Figure 2.3 and Figure 2.4).

A large proportion of patients display some level of GI, 99.95% with Absolute CNA Score > 0 and 95% with Absolute CNA Score > 100. The distribution of the CNA Del Score is broader than the distribution of the CNA Amp Score, standard deviation 3,150.91 compared to 2,252.74, with a higher maximum score value of 14,530, indicating that a patient's genome may undergo higher levels of deletion than amplification (Table 2.2). This feature is also indicated in the Difference Score distribution, where there is a higher density of patients displaying negative difference values, mean -378.86 and median -8, indicating higher levels of deletion than amplification. Similar trends are observed in the CNA Burden distributions, where the standard deviations of the CNA Amp and Del distributions are 8.54 and 13.89, respectively, with a higher maximum burden value of 64.17 for the CNA Del distribution (Table 2.4).

To determine the impact of missingness on the CNA metrics, an assessment using only complete-case (CC) data, i.e. including only patients that have CNA information for all 22,544 genes recorded, versus using all available data to produce the CNA metrics is carried out. When using all available data, CNA metrics for all $2,173$ patients are produced, while CNA metrics calculated using only CC data discard patients displaying an NA value in any of the genes, leaving $2,091$ patients for which CNA metrics are calculated. It should be noted for the CNA Burden calculation, G refers to the number of genes recorded for each patient and ranges from $22,466$ to $22,544$, when using all available data, and is $22,544$ when using the CC data. The main advantage of using CC data is simplicity, as statistical analysis is more straightforward with CC data. Disadvantages of using CC data stem from the potential loss of information in discarding incomplete cases.

The effect of any missingness is assessed by comparing the features of the observed distributions for CC data and all data, using comparative density plots and estimating the overlapping area of the two kernel densities, using the R `overlap()` function (Pastore et al., 2022). This function is used to estimate the proportion of overlapping area between two densities, i.e. where the integral of the minimum between two densities is divided by the integral of the maximum of the two densities. This proportion is then multiplied by 100 to calculate the percentage overlap.

Figure 2.3 and Table 2.6 indicate that the two density plots for each global CNA Score metric are similar and have a high percentage overlap. The lowest percentage overlap, observed within Absolute CNA Score and CNA Del Score were 96.55% and 97.15%, respectively. High concordance is also observed in the comparison between the CNA Burden metrics (Figure 2.4 and Table 2.7). The CNA Burden metrics displaying the lowest percentage overlap are CNA Burden and CNA Del Burden with 96.76% and 97.18%, respectively.

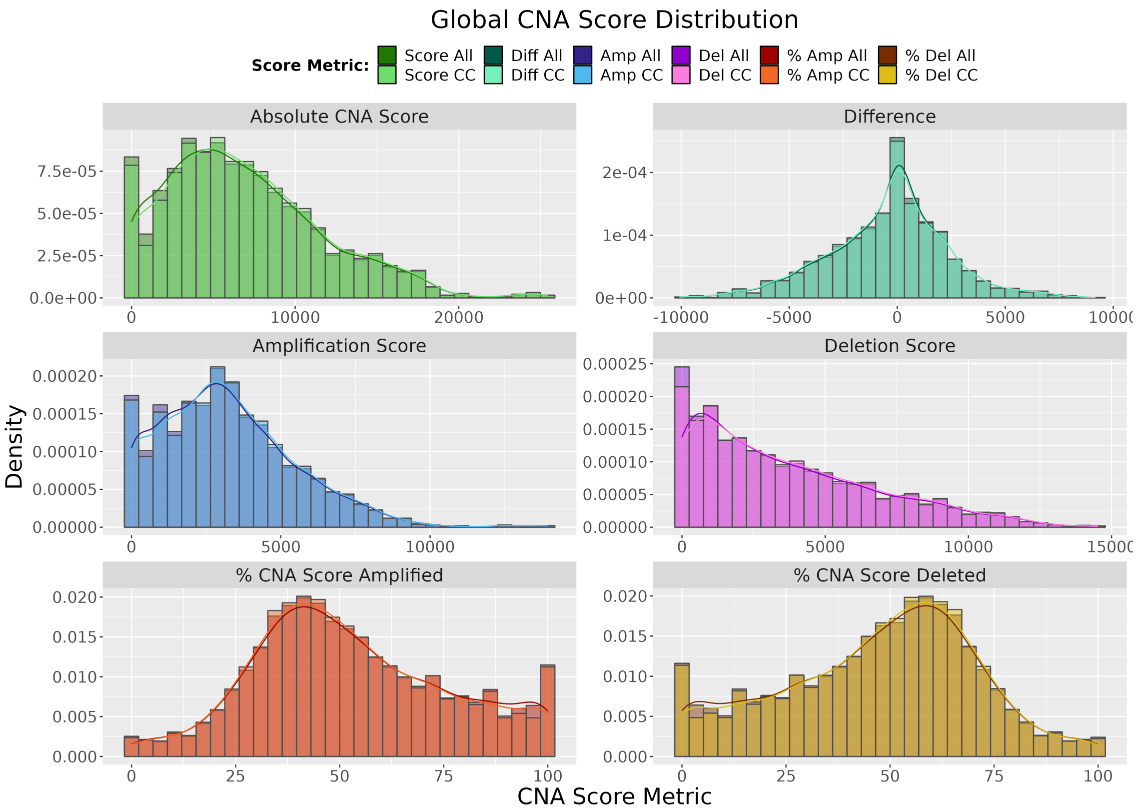


Figure 2.3: Density plots for each global CNA Score metric. Each facet contains density plots for both the complete-case CNA Score metric and the CNA Score metric calculated using all available data.

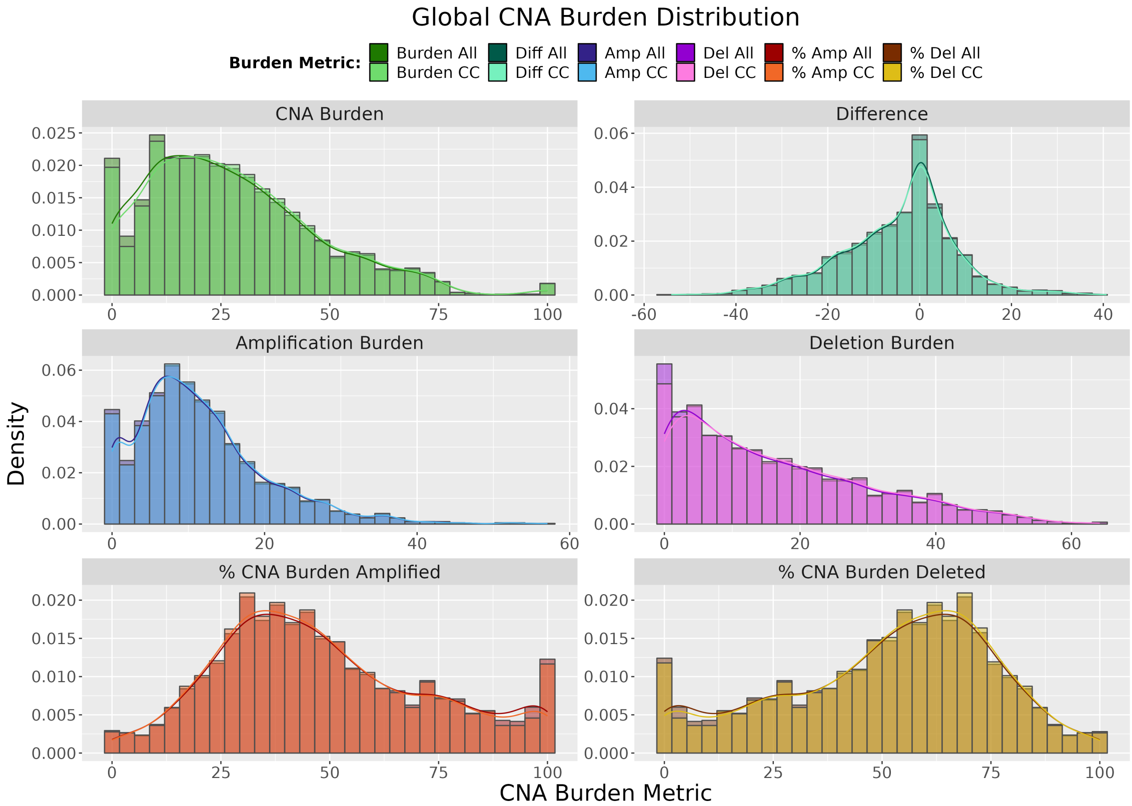


Figure 2.4: Density plots for each global CNA Burden metric. Each facet contains density plots for both the complete-case CNA Burden metric and the CNA Burden metric calculated using all available data.

Table 2.2: Summary statistics of the CNA Score metrics where all available data are used.

Summary Statistics of CNA Score Metrics (All)						
CNA Score Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
Absolute CNA Score	2,173	0.00	6,835.01	6,093.00	25,434.00	4,731.11
CNA Amp Score	2,173	0.00	3,228.08	2,960.00	13,939.00	2,252.74
CNA Del Score	2,173	0.00	3,606.94	2,809.00	14,530.00	3,150.91
Difference Score	2,173	−10,088.00	−378.86	−8.00	9,179.00	2,760.93
Percentage Score Amp	2,173	0.00	52.83	49.46	100.00	23.46
Percentage Score Del	2,173	0.00	47.13	50.50	100.00	23.45

Table 2.3: Summary statistics of the CNA Score metrics where only complete cases are used.

Summary Statistics of CNA Score Metrics (CC)						
CNA Score Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
Absolute CNA Score	2,091	0.00	7,007.52	6,313.00	25,434.00	4,722.01
CNA Amp Score	2,091	0.00	3,289.06	3,020.00	13,939.00	2,251.11
CNA Del Score	2,091	0.00	3,718.45	2,922.00	14,530.00	3,153.92
Difference Score	2,091	−10,088.00	−429.39	−56.00	9,179.00	2,780.66
Percentage Score Amp	2,091	0.00	52.23	48.85	100.00	23.07
Percentage Score Del	2,091	0.00	47.72	51.10	100.00	23.07

Table 2.4: Summary statistics of the CNA Burden metrics where all available data are used.

Summary Statistics of CNA Burden Metrics (All)						
CNA Burden Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
CNA Burden	2,173	0.00	27.42	24.28	100.00	19.35
CNA Amp Burden	2,173	0.00	11.51	9.91	57.09	8.54
CNA Del Burden	2,173	0.00	15.91	12.42	64.17	13.89
Difference Burden	2,173	−54.10	−4.39	−1.90	40.73	12.54
Percentage Burden Amp	2,173	0.00	48.69	44.65	100.00	24.36
Percentage Burden Del	2,173	0.00	51.26	55.31	100.00	24.36

Table 2.5: Summary statistics of the CNA Burden metrics where only complete cases are used.

Summary Statistics of CNA Burden Metrics (CC)						
CNA Burden Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
CNA Burden	2,091	0.00	28.09	25.10	100.00	19.33
CNA Amp Burden	2,091	0.00	11.69	10.15	57.09	8.54
CNA Del Burden	2,091	0.00	16.40	12.93	64.17	13.90
Difference Burden	2,091	−54.10	−4.71	−2.41	40.73	12.58
Percentage Burden Amp	2,091	0.00	47.93	44.12	100.00	23.88
Percentage Burden Del	2,091	0.00	52.02	55.88	100.00	23.88

Table 2.6: The percentage overlap between the global complete-case and all-case CNA Score metric densities. Metrics are ordered and coloured by percentage overlap.

CNA Score Metric	% Overlap
Absolute CNA Score	96.55
CNA Del Score	97.15
CNA Amp Score	97.35
% CNA Score Amp	97.58
% CNA Score Del	97.69
Difference Score	97.69

Table 2.7: The percentage overlap between the global complete-case and all-case CNA Burden metric densities. Metrics are ordered and coloured by percentage overlap.

CNA Burden Metric	% Overlap
CNA Burden	96.76
CNA Del Burden	97.18
CNA Amp Burden	97.24
% CNA Burden Amp	97.45
% CNA Burden Del	97.45
Difference Burden	97.87

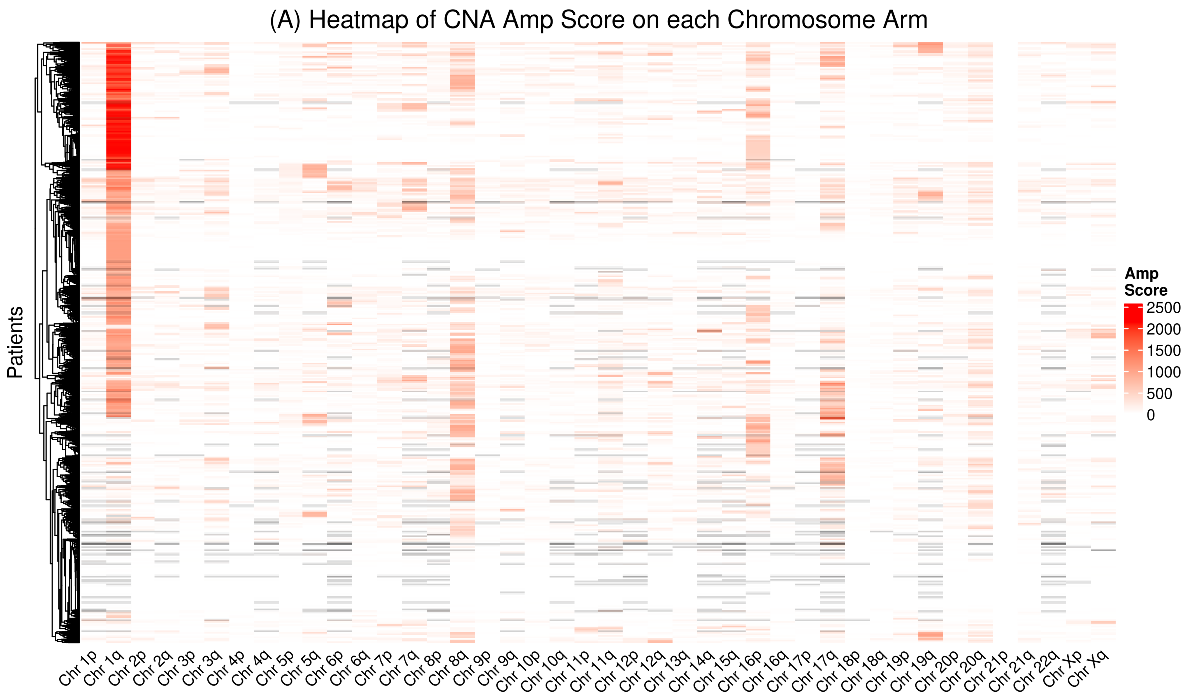
2.4.2 Observed Distributions for Chromosome Arm CNA Metrics

Similarly, the observed distributions for the chromosome arm CNA metrics are inspected and an assessment of the effects of missingness is carried out. Figures 2.5 and 2.6 display the heatmaps of the chromosome arm CNA Amp and Del Score metrics calculated using compete cases only and all available data, the grey indicates missing data. Figures 2.7 and 2.8 display the heatmaps of the chromosome arm CNA Amp and Del Burden metrics calculated using compete cases only and all available data. Comparing A and B in each of these figures, it is observed that similar clusters of patients are generated based on both the all-case and CC chromosome arm CNA metrics. When comparing the overlap in distributions it is observed that only 10 out of the 252 chromosome arm CNA Score and Burden metrics display percentage overlap below 80%. The lowest overlap in the CNA Score and Burden metric distributions is in the Percentage Amp metrics on 9p, 11.00% and 10.84%, respectively (Table 2.8 and 2.9). The second and third lowest overlap between the CC and all-case CNA Score and Burden metric distributions is observed for chromosome arms 9p, 16.69% and 16.83%, and 7p, 42.65% and 45.26%. Density plots, focusing on chromosome 9p and 7p, are provided in Figure 2.9 and indicate that high density regions, such as those located around 0, display high levels of mismatch.

Figures 2.5-2.8 also highlight previously documented patterns of CNAs in breast cancer, including high levels of amplifications on chromosome 1q, 8q and 16p and high levels of deletions on 8p, 16q and 17p (Jönsson et al., 2010; Curtis et al., 2012). Similar to the global CNA Score and Burden metrics, it appears that deletions are more widespread across the genome, affecting greater numbers of chromosomes, than amplifications.

Rather than extensively presenting details of distributions for each of the 42 chromosome arms, we select chromosome arm 1q for more detailed illustration and discussion (see Supplementary Information for remaining chromosome arms). Chromosome arm 1q is frequently altered in breast cancer and shows interesting features in this analysis. The summary statistics (Tables 2.10-2.13) and distributions (Figures 2.10 and 2.11) indicate that the copy number landscape of chromosome 1q is dominated by amplifications, median CNA Score values 0 and 954 and median CNA Burden values 0 and 78.71%, for deletions and amplifications, respectively. The maximum values of the CNA Burden and CNA Amp, 100% and 99.91%, and the Difference Score distribution being nearly entirely positive also suggests that almost all of the alterations observed are amplifications. Figure 2.10 indicates that three of the CNA Score metrics on chromosome 1q, Absolute CNA Score, CNA Amp Score and Difference Score, have trimodal distributions. Three peaks correspond to cases where patients have low, moderate and high levels of GI. For the CNA Burden metrics the majority of the metric distributions are bimodal, with peaks corresponding to patients with low and high GI.

Overall, assessing distributions of the global CNA metrics comparing CC patient data to all-patient data, i.e. including those with some missingness prevalent, shows that missing values have only a minor impact on these distributions. The impact of missingness is greater in the chromosome arm CNA metric distributions, however, as the majority of distributions displayed greater than 80% overlap, it is determined that including all cases, across the global and chromosome arm-specific metrics, is unlikely to invoke bias in the form of underestimating the CNA metrics. Furthermore, imputation of missing CNA values was considered but not performed as the impact of the missing values on the CNA metric distributions was shown to be minor across the majority of CNA metrics.



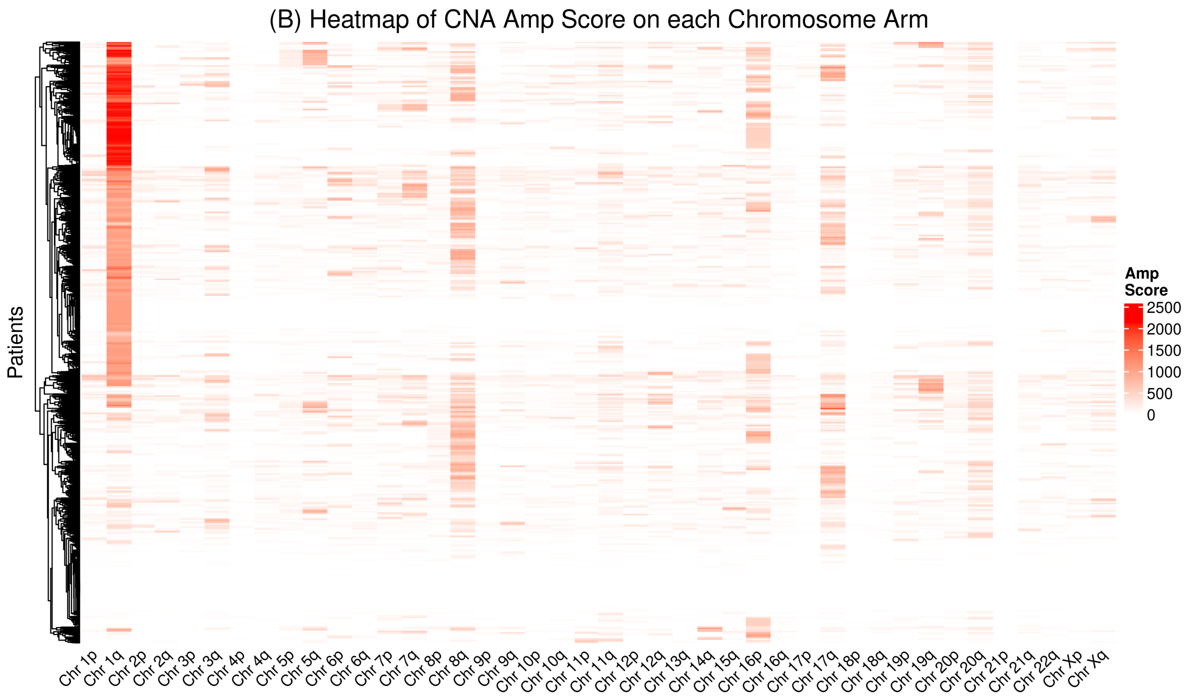


Figure 2.5: Heatmap of CNA Amp Score across chromosome arms with (A) consideration to complete-case METABRIC patients only (n = 2,091) and (B) consideration to all METABRIC patients including those presenting with some missing data (n = 2,173). Grey indicates missing values.

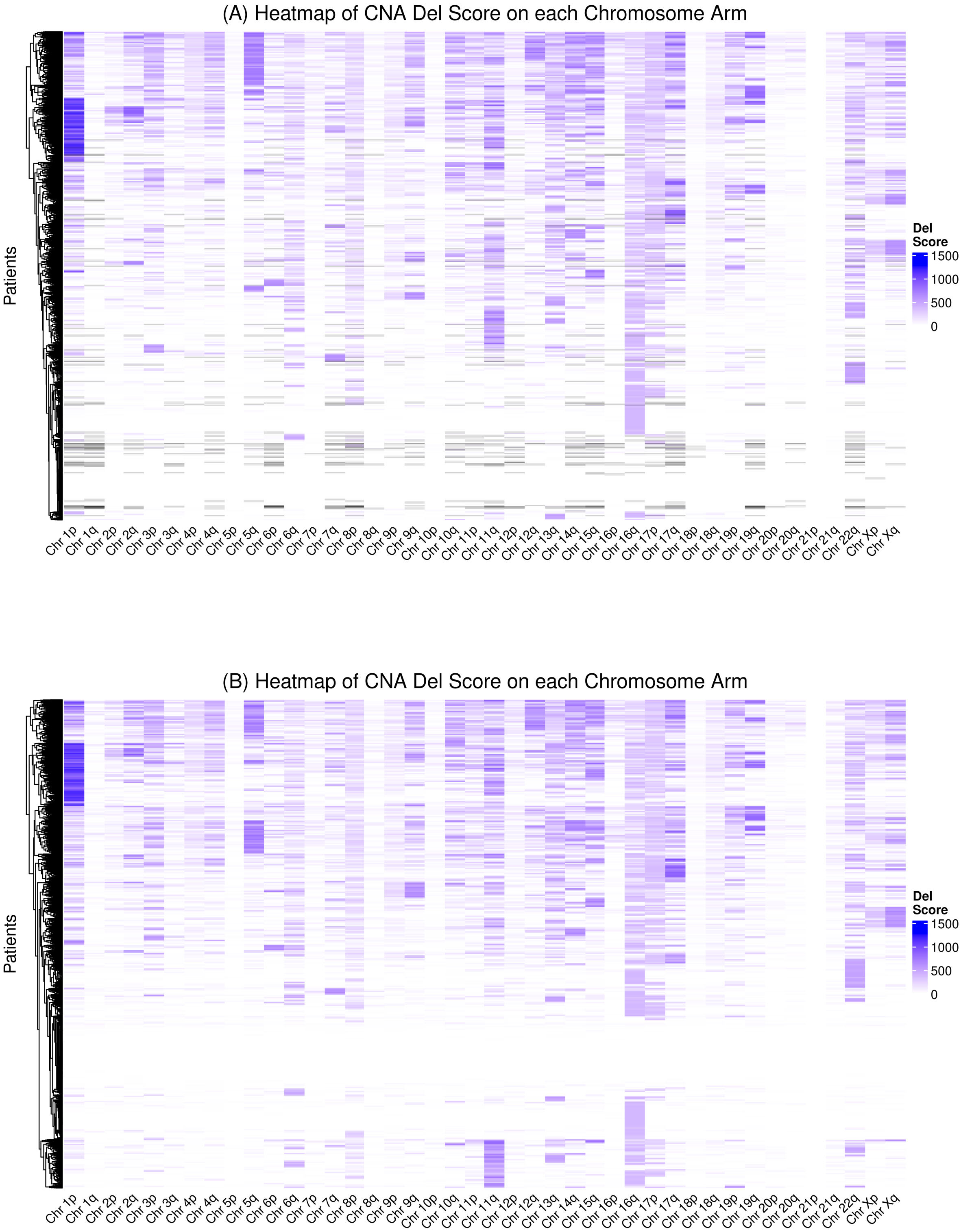


Figure 2.6: Heatmap of CNA Del Score across chromosome arms with (A) consideration to complete-case METABRIC patients only (n = 2,091) and (B) consideration to all METABRIC patients including those presenting with some missing data (n = 2,173). Grey indicates missing values.

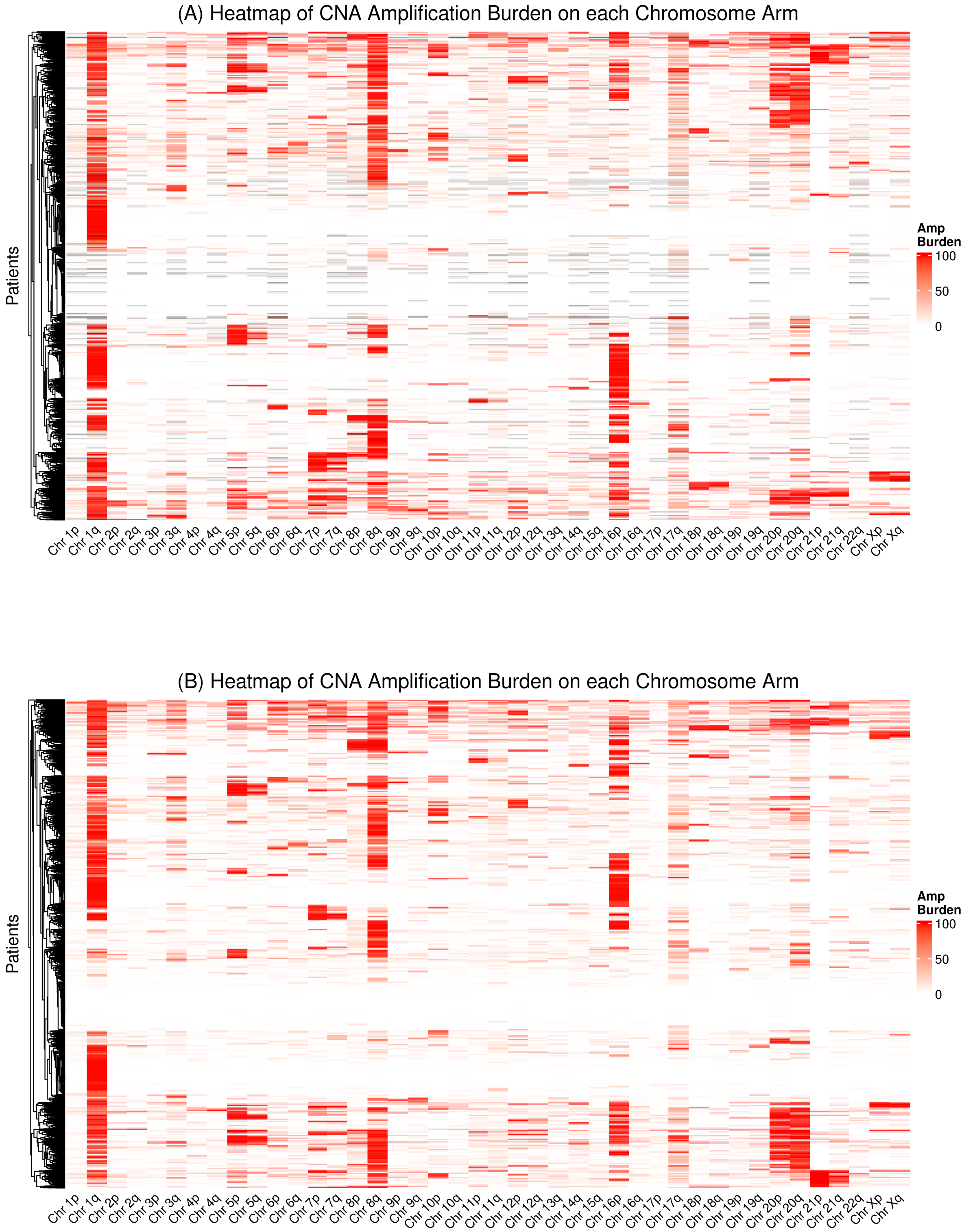
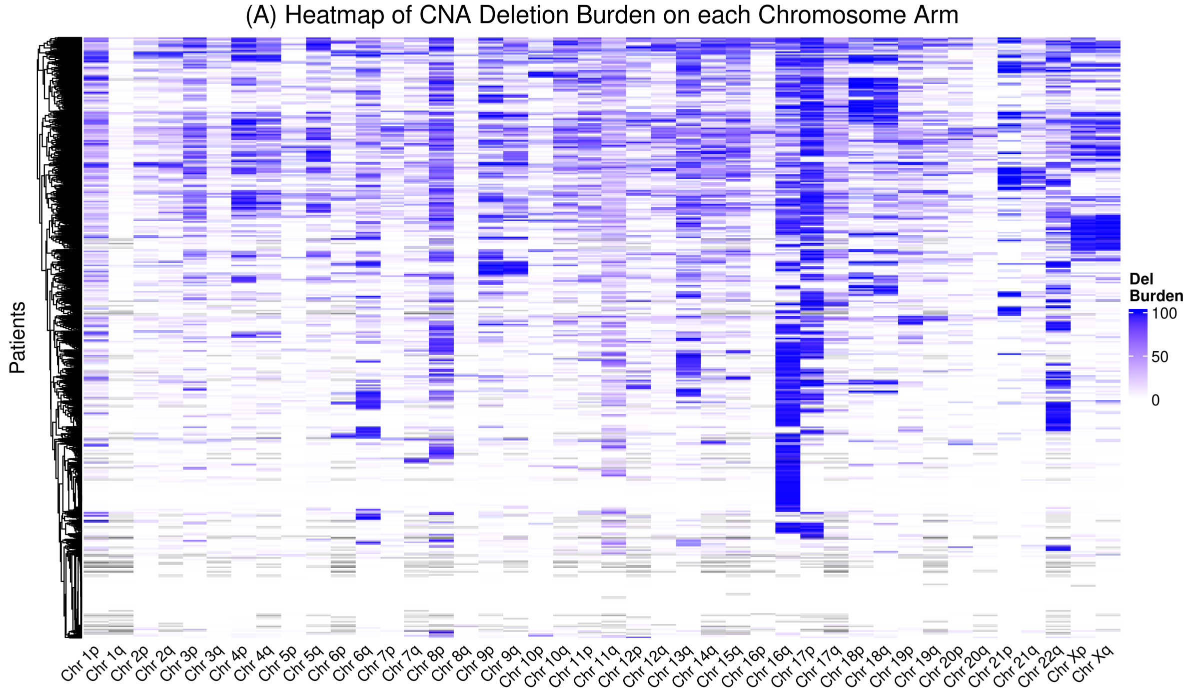


Figure 2.7: Heatmap of CNA Amp Burden across chromosome arms with (A) consideration to complete-case METABRIC patients only (n = 2,091) and (B) consideration to all METABRIC patients including those presenting with some missing data (n = 2,173). Grey indicates missing values.



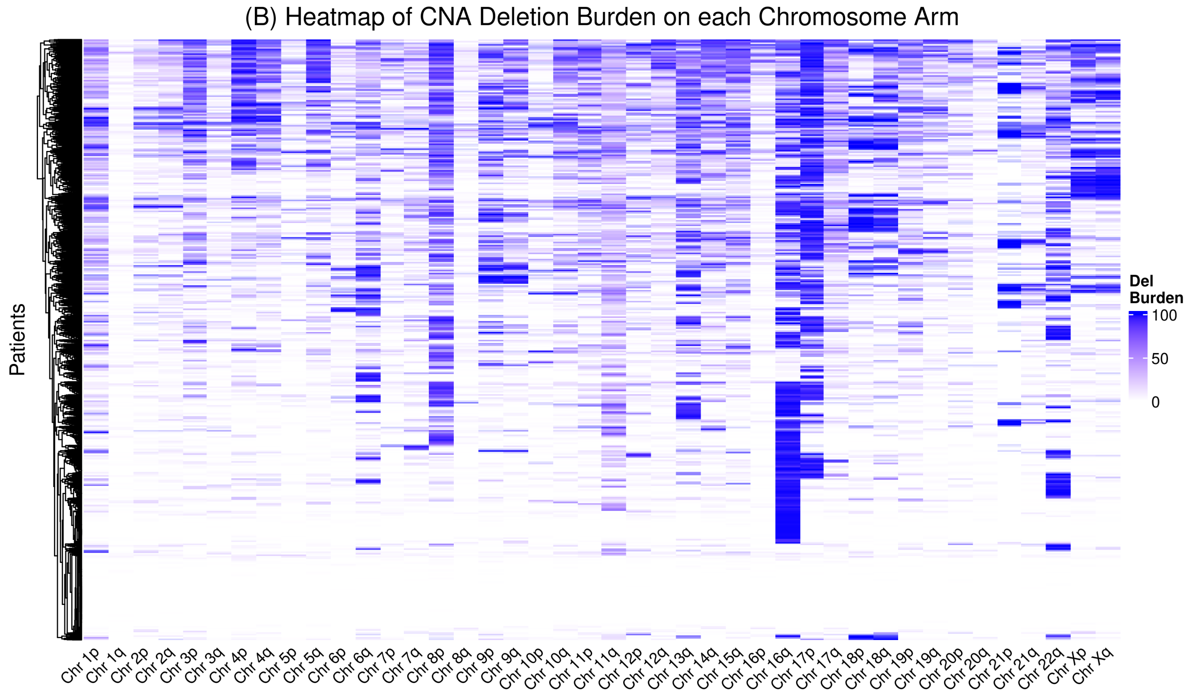


Figure 2.8: Heatmap of CNA Del Burden across chromosome arms with (A) consideration to complete-case METABRIC patients only (n = 2,091) and (B) consideration to all METABRIC patients including those presenting with some missing data (n = 2,173). Grey indicates missing values.

Table 2.8: Chromosomes arms with poor overlap between complete-case patient and all-patient CNA Score metrics. Metrics are ordered and coloured by percentage overlap.

CNA Score Metric	% Overlap
% CNA Score Amp 9p	11.00
CNA Amp Score 9p	16.69
% CNA Score Del 7p	42.65
% CNA Score Amp 18q	55.59
Difference 18p	74.15
% CNA Score Del 8q	75.14
% CNA Score Amp 22q	76.71
Difference Xq	77.85
Difference 18q	78.80
Difference 19p	79.10

Table 2.9: Chromosomes arms with poor overlap between complete-case patient and all-patient CNA Burden metrics. Metrics are ordered and coloured by percentage overlap.

CNA Burden Metric	% Overlap
% CNA Burden Amp 9p	10.84
CNA Amp Burden 9p	16.83
% CNA Burden Del 7p	45.26
% CNA Burden Amp 18q	65.25
CNA Amp Burden Xq	67.35
Difference 19p	70.39
% CNA Burden Amp 22q	72.01
Difference 18p	76.29
Difference Xq	77.29
Difference 18q	79.68

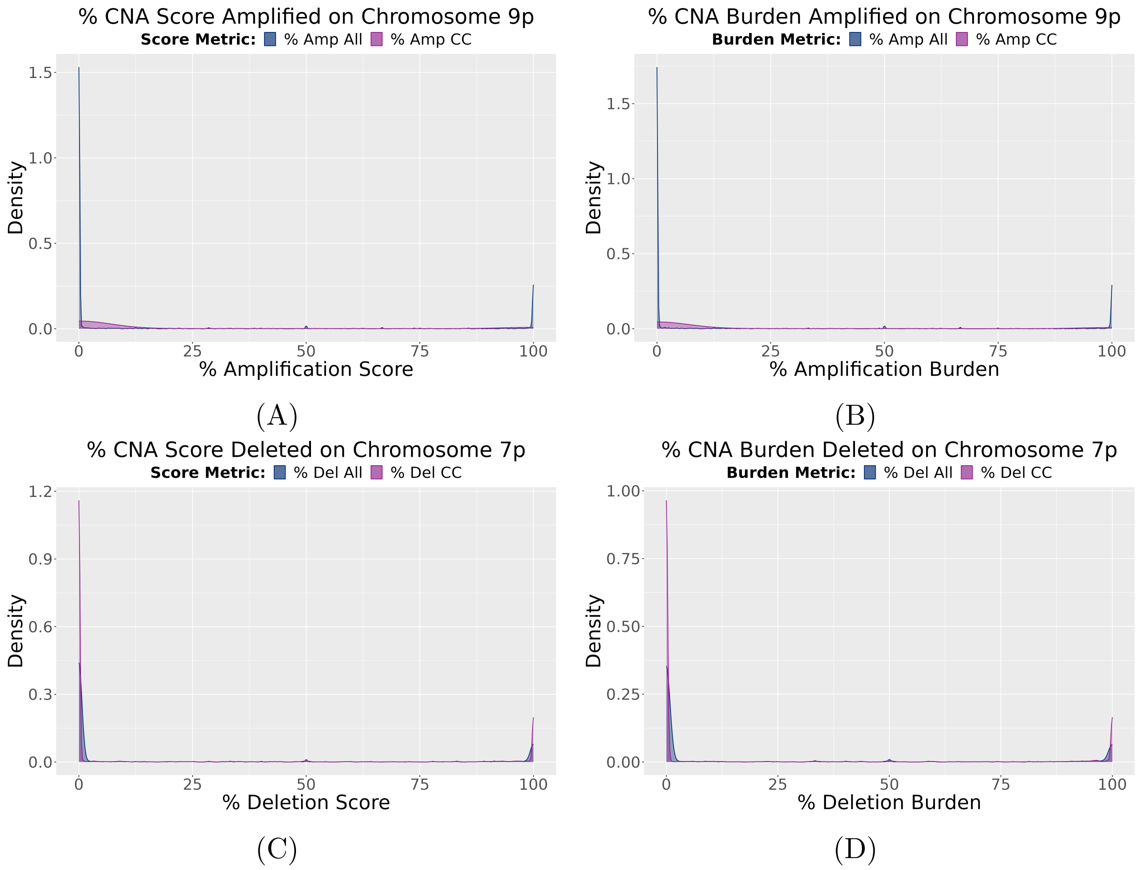


Figure 2.9: Density plots for selected chromosome arm CNA metrics. (A) Percentage CNA Score Amp on chromosome 9p, (B) Percentage CNA Burden Amp on chromosome 9p, (C) Percentage CNA Score Del on chromosome 7p and (D) Percentage CNA Burden Del on chromosome 7p. Each plot contains density plots for both the complete-case metric and the metric calculated using all available data.

Table 2.10: Summary statistics of the CNA Score metrics on chromosome 1q where all available data is used.

Summary Statistics of CNA Score Metrics on Chromosome 1q (All)						
CNA Score Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
CNA Score	2,173	0.00	859.64	985.00	2,138.00	728.17
CNA Amplification Score	2,173	0.00	845.61	954.00	2,138.00	731.76
CNA Deletion Score	2,173	0.00	14.03	0.00	876.00	57.50
Difference Score	2,173	−702.00	831.59	935.00	2,138.00	739.81
Percentage Score Amplified	2,173	0.00	79.52	99.91	100.00	37.96
Percentage Score Deleted	2,173	0.00	12.25	0.00	100.00	29.78

Table 2.11: Summary statistics of the CNA Score metrics on chromosome 1q where only complete cases are used.

Summary Statistics of CNA Score Metrics on Chromosome 1q (CC)						
CNA Score Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
CNA Score	2,091	0.00	877.88	1,011.00	2,138.00	730.20
CNA Amplification Score	2,091	0.00	863.52	976.00	2,138.00	734.10
CNA Deletion Score	2,091	0.00	14.36	0.00	876.00	58.56
Difference Score	2,091	−702.00	849.16	962.00	2,138.00	742.63
Percentage Score Amplified	2,091	0.00	80.03	99.92	100.00	37.67
Percentage Score Deleted	2,091	0.00	11.74	0.00	100.00	29.27

Table 2.12: Summary statistics of the CNA Burden metrics on chromosome 1q where all available data is used.

Summary Statistics of CNA Burden Metrics on Chromosome 1q (All)						
CNA Burden Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
CNA Burden	2,173	0.00	58.95	80.58	100.00	43.16
CNA Amplification Burden	2,173	0.00	57.69	78.71	99.91	43.50
CNA Deletion Burden	2,173	0.00	1.26	0.00	81.79	5.35
Difference Burden	2,173	−73.67	56.44	76.66	99.91	44.48
Percentage Burden Amplified	2,173	0.00	79.25	99.91	100.00	38.08
Percentage Burden Deleted	2,173	0.00	12.51	0.00	100.00	30.00

Table 2.13: Summary statistics of the CNA Burden metrics on chromosome 1q where only complete cases are used.

Summary Statistics of CNA Burden Metrics on Chromosome 1q (CC)						
CNA Burden Metric	**n**	**min**	**mean**	**median**	**max**	**sd**
CNA Burden	2,091	0.00	59.88	82.73	100.00	42.95
CNA Amplification Burden	2,091	0.00	58.59	80.58	99.91	43.32
CNA Deletion Burden	2,091	0.00	1.29	0.00	81.79	5.44
Difference Burden	2,091	−73.67	57.30	79.46	99.91	44.35
Percentage Burden Amplified	2,091	0.00	79.72	99.91	100.00	37.81
Percentage Burden Deleted	2,091	0.00	12.05	0.00	100.00	29.54

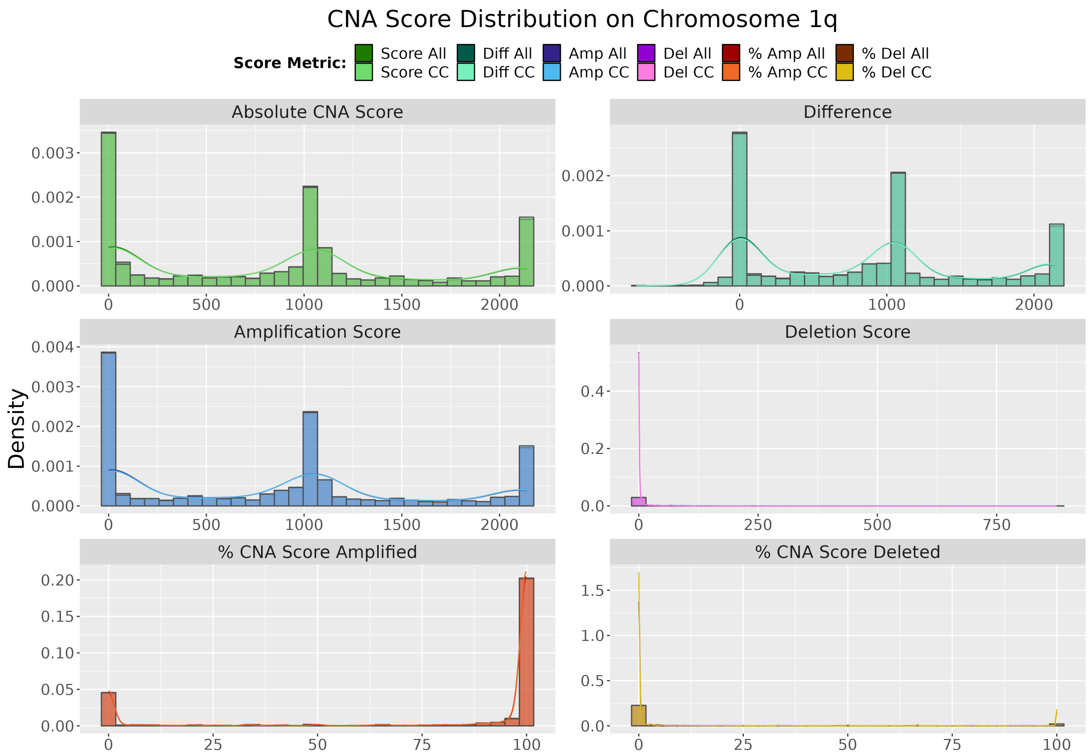


Figure 2.10: Density plots for each CNA Score metric on chromosome 1q. Each facet contains density plots for both the complete-case CNA Score metric and the CNA Score metric calculated using all available data.

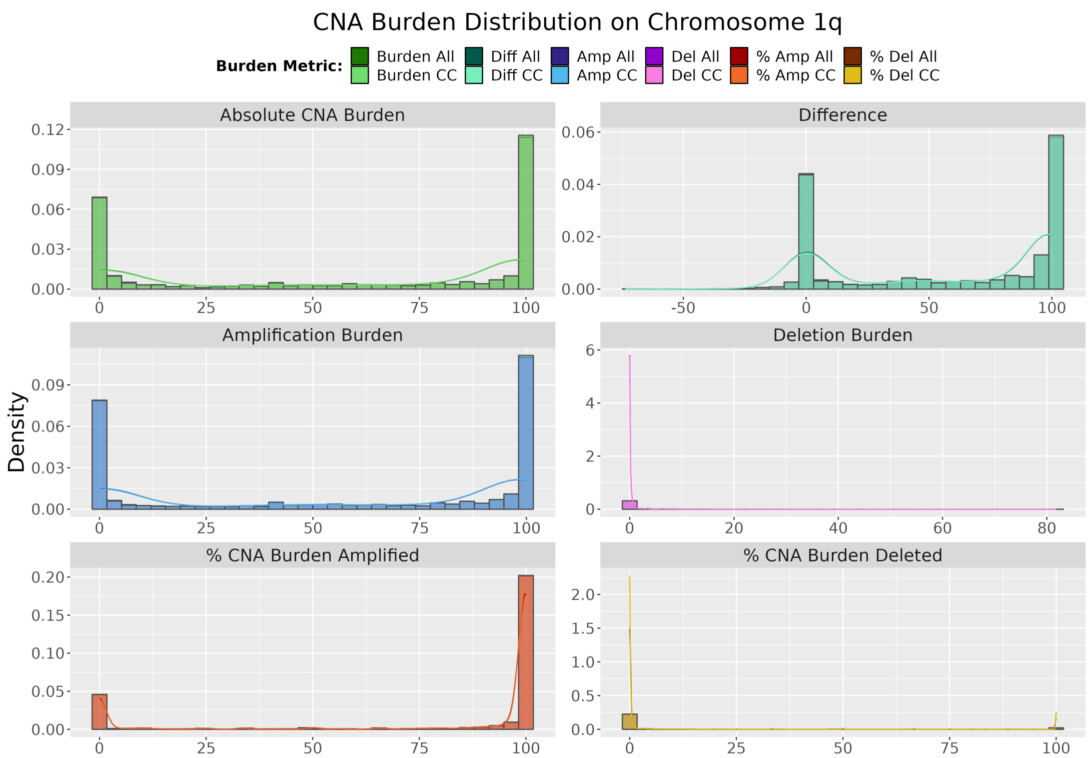


Figure 2.11: Density plots for each CNA Burden metric on chromosome 1q. Each facet contains density plots for both the complete-case CNA Burden metric and the CNA Burden metric calculated using all available data.

2.5 CNA Metric Distributions within Molecular Subtype Classifications

The calculated CNA Score and Burden metrics are cross-referenced against breast cancer subtype classifications, PAM50 subtype and IntClust, to determine if observed distributions of metrics differed comparing these stratified cohorts of patients. Within the METABRIC cohort there are 1,974 patients for which CNA data and PAM50 subtype information are available and 1,980 patients for which CNA data and IntClust information are available (Table 1.1). The 529 patients missing both PAM50 and IntClust information did not have any gene expression data available meaning they could not be allocated PAM50 or IntClust, while six patients had IntClust information but were categorised as PAM50 “NC” and subsequently recoded as NA. We present distributions of the global CNA metrics across molecular classifications in Section 2.5.1 and distributions of the chromosome arm CNA metrics across molecular classifications in Section 2.5.2.

2.5.1 Observed Distributions for Global CNA Metrics across Molecular Subtype Classifications

The observed distribution of the six CNA Score metrics, for patients stratified by PAM50 subtype, is displayed in Figure 2.12, and Figure 2.13 displays the six CNA Burden metrics. Also provided with Figures 2.12 and 2.13 are the Benjamini-Hochberg (BH) adjusted p-values for the Kruskal-Wallis test for any difference in the distribution of a metric comparing the groups of patients stratified by PAM50 subtype.

These visualisations and accompanying statistical tests (Figure 2.12 and Figure 2.13) indicate that some significant difference exists comparing each of the CNA metric distributions across PAM50 subtype (Kruskal-Wallis adjusted $p < 0.0001$). Dunn’s Test, a post-hoc test for Kruskal-Wallis, is then applied to each CNA metric performing pairwise comparisons to determine which groups are significantly different in mean rank scores (Tables 2.14 and 2.15). The distribution of CNA Score and Burden metrics in Basal patients is significantly different from all other subtypes (Tables 2.14 and 2.15). Basal patients display the highest Absolute CNA Scores and CNA Burden across all subtypes ($p < 0.0001$) indicating higher levels of GI when compared to other subtypes. In line with this, the Basal patients display the highest CNA Amp and Del Score and Burden across all subtypes ($p < 0.001$ for each comparison).

The HER2 and Luminal B subtypes have the 2nd and 3rd highest Absolute CNA Score and Burden, CNA Amp Score and Burden and CNA Del Score and Burden, respectively. While the shape and spread of these metric distributions appear quite similar, the HER2 subtype displays slightly higher levels of Absolute CNA Score and Burden ($p = 0.01$). This is also observed for the CNA Del metrics, where HER2 patients have higher levels of deletions than Luminal B patients ($p = 0.01$), but not the CNA Amp metrics ($p = 0.41$ and $p = 0.47$). When comparing the Luminal A and Luminal B patients, it is observed that Luminal B patients have significantly higher levels of instability across all metrics ($p < 0.0001$). The Luminal A, Normal and Claudin-low patients display the lowest CNA Score and Burden metrics. The Normal and Claudin-low subtypes display no significant difference for the total, amplification and deletion CNA Score and Burden metrics ($p > 0.05$). All Luminal

A and Claudin-low densities, apart from the CNA Amp metric distributions, are not significantly different from each other in these subtypes ($p > 0.05$). Luminal A patients display significantly higher levels of amplifications when compared with Normal and Claudin-low subtypes ($p < 0.0001$, Tables 2.14 and 2.15).

Focusing on the direct comparison of levels of amplification to deletion within each PAM50 subtype, Figures 2.14 and 2.15, the subtypes known to be associated with poorer survival outcome, i.e. Basal, HER2 and Luminal B, have significantly higher levels of deletion burden than amplification burden ($p < 0.0001$). Conversely PAM50 subtypes with better survival prognosis either display significantly more amplifications, Luminal A ($p < 0.01$), or no significant difference in the levels of amplifications and deletions, Normal and Claudin-low ($p > 0.05$).

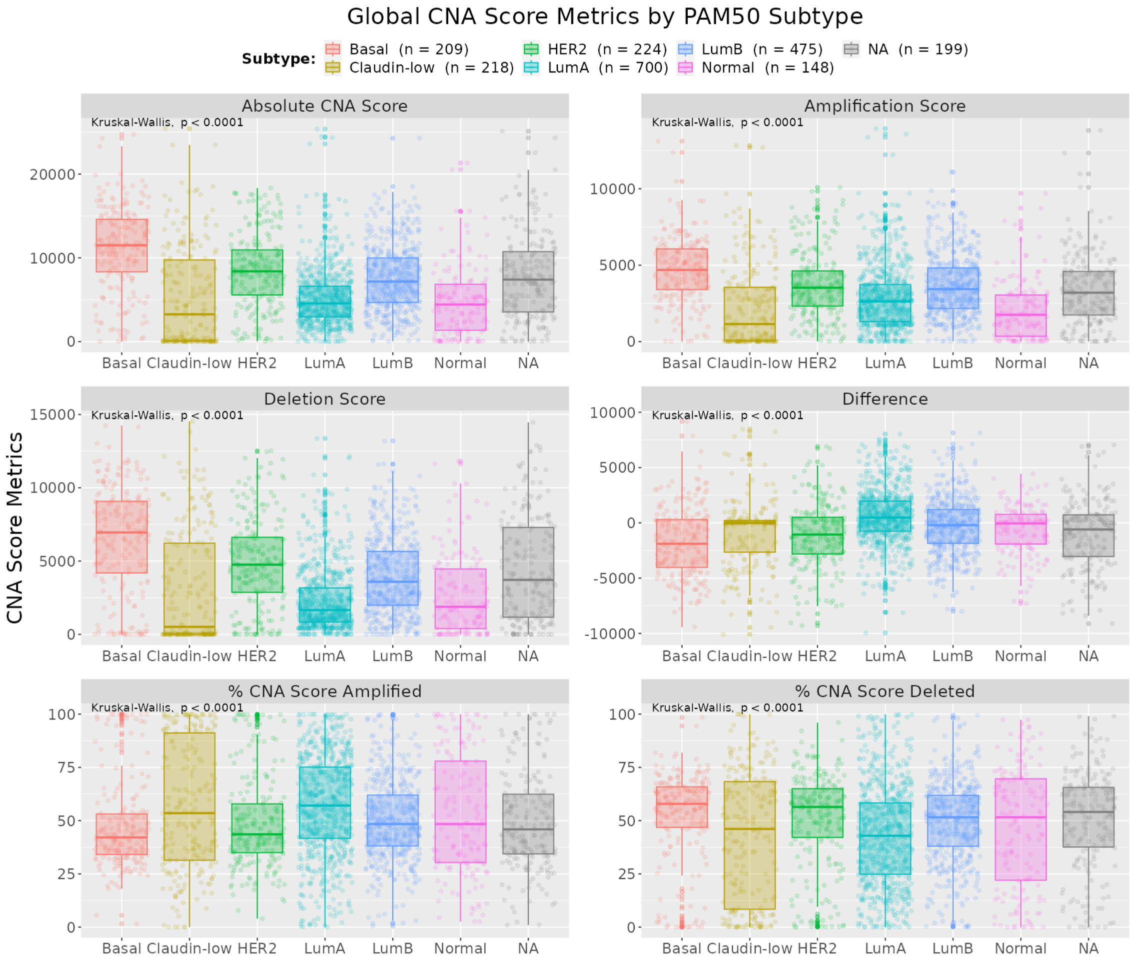


Figure 2.12: Boxplots for each CNA Score metric by PAM50 subtype. Each facet contains boxplots for the CNA Score metrics calculated using all available data accompanied by Benjamini-Hochberg adjusted Kruskal-Wallis p-values. NA denotes METABRIC patients missing PAM50 information.

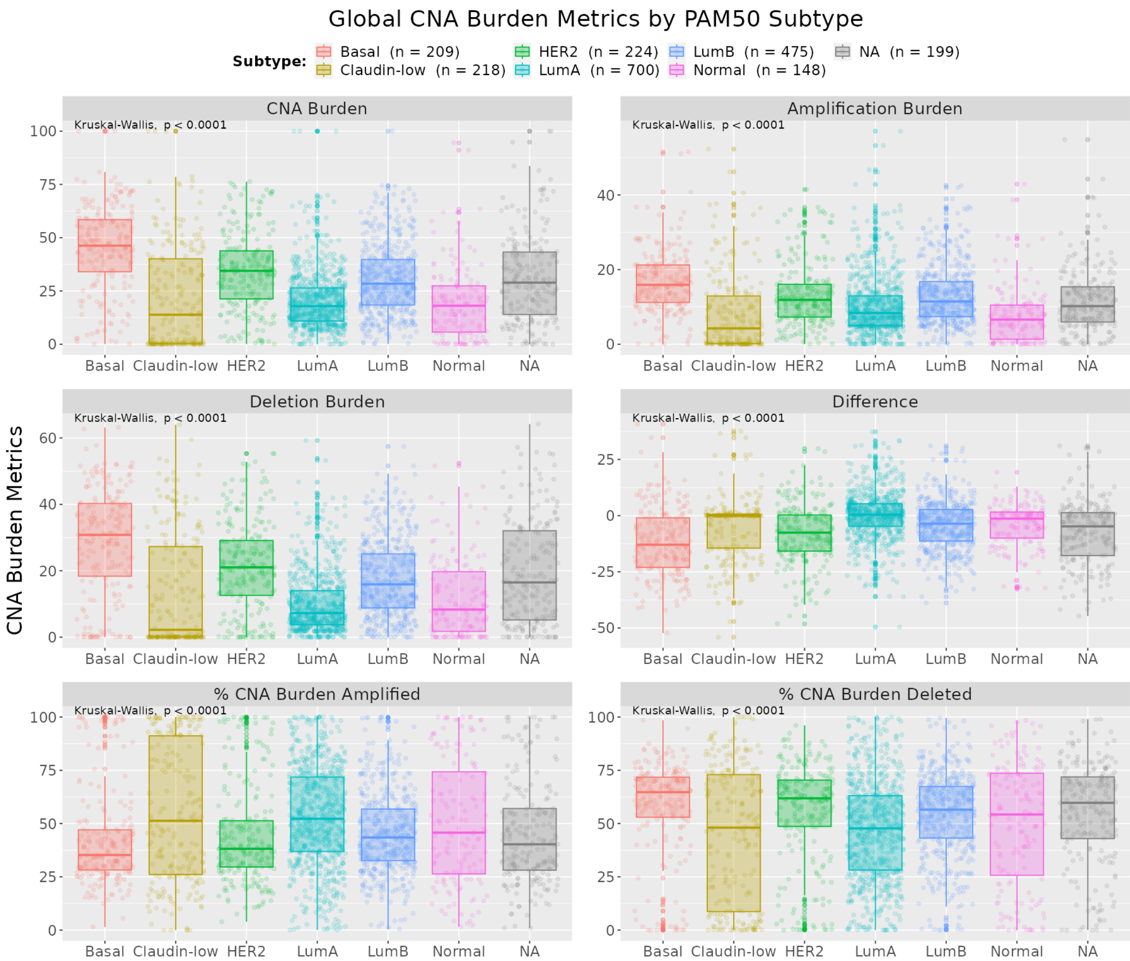


Figure 2.13: Boxplots for each CNA Burden metric by PAM50 subtype. Each facet contains boxplots for the CNA Burden metrics calculated using all available data accompanied by Benjamini-Hochberg adjusted Kruskal-Wallis p-values. NA denotes METABRIC patients missing PAM50 information.

Table 2.14: Comparisons of CNA Score metric distributions by PAM50 subtype. Z statistic and Benjamini-Hochberg adjusted p-value for each Dunn's test are shown.

Comparison of CNA Score Metrics by PAM50 Subtype			
Comparisons	Absolute CNA Score Z (adj p-value)	CNA Amp Score Z (adj p-value)	CNA Del Score Z (adj p-value)
Basal - LumA	15.14 (<0.0001)	11.09 (<0.0001)	14.32 (<0.0001)
Basal - Claudin-low	13.05 (<0.0001)	12.88 (<0.0001)	11.94 (<0.0001)
Basal - Normal	12.12 (<0.0001)	11.67 (<0.0001)	9.93 (<0.0001)
HER2 - LumA	10.11 (<0.0001)	5.65 (<0.0001)	10.26 (<0.0001)
HER2 - Normal	8.36 (<0.0001)	7.67 (<0.0001)	6.85 (<0.0001)
LumB - Normal	7.44 (<0.0001)	8.37 (<0.0001)	5.64 (<0.0001)
Basal - LumB	7.26 (<0.0001)	5.6 (<0.0001)	6.46 (<0.0001)
Basal - HER2	4.34 (<0.0001)	4.58 (<0.0001)	3.55 (<0.0001)
HER2 - LumB	2.28 (0.01)	0.3 (0.41)	2.4 (0.01)
LumA - Normal	1.21 (0.13)	4.19 (<0.0001)	-0.68 (0.27)
Claudin-low - Normal	0.37 (0.36)	0.06 (0.48)	-0.83 (0.23)
Claudin-low - LumA	-0.9 (0.2)	-4.81 (<0.0001)	-0.35 (0.36)
Claudin-low - LumB	-8.08 (<0.0001)	-9.56 (<0.0001)	-7.58 (<0.0001)
Claudin-low - HER2	-8.89 (<0.0001)	-8.48 (<0.0001)	-8.56 (<0.0001)
LumA - LumB	-9.94 (<0.0001)	-6.88 (<0.0001)	-9.97 (<0.0001)

Table 2.15: Comparisons of CNA Burden metric distributions by PAM50 subtype. Z statistic and Benjamini-Hochberg adjusted p-value for each Dunn's test are shown.

Comparison of CNA Burden Metrics by PAM50 Subtype			
Comparisons	Absolute CNA Burden Z (adj p-value)	CNA Amp Burden Z (adj p-value)	CNA Del Burden Z (adj p-value)
Basal - LumA	15.38 (<0.0001)	10.76 (<0.0001)	14.33 (<0.0001)
Basal - Claudin-low	12.96 (<0.0001)	12.7 (<0.0001)	11.98 (<0.0001)
Basal - Normal	11.99 (<0.0001)	11.19 (<0.0001)	9.95 (<0.0001)
HER2 - LumA	10.28 (<0.0001)	5.3 (<0.0001)	10.26 (<0.0001)
HER2 - Normal	8.16 (<0.0001)	7.18 (<0.0001)	6.86 (<0.0001)
Basal - LumB	7.43 (<0.0001)	5.4 (<0.0001)	6.45 (<0.0001)
LumB - Normal	7.13 (<0.0001)	8.01 (<0.0001)	5.67 (<0.0001)
Basal - HER2	4.41 (<0.0001)	4.59 (<0.0001)	3.56 (<0.0001)
HER2 - LumB	2.38 (0.01)	0.08 (0.47)	2.39 (0.01)
LumA - Normal	0.83 (0.23)	3.91 (<0.0001)	-0.67 (0.27)
Claudin-low - Normal	0.32 (0.38)	-0.26 (0.43)	-0.85 (0.23)
Claudin-low - LumA	-0.54 (0.32)	-4.92 (<0.0001)	-0.39 (0.35)
Claudin-low - LumB	-7.8 (<0.0001)	-9.55 (<0.0001)	-7.63 (<0.0001)
Claudin-low - HER2	-8.73 (<0.0001)	-8.28 (<0.0001)	-8.59 (<0.0001)
LumA - LumB	-10.03 (<0.0001)	-6.73 (<0.0001)	-10 (<0.0001)

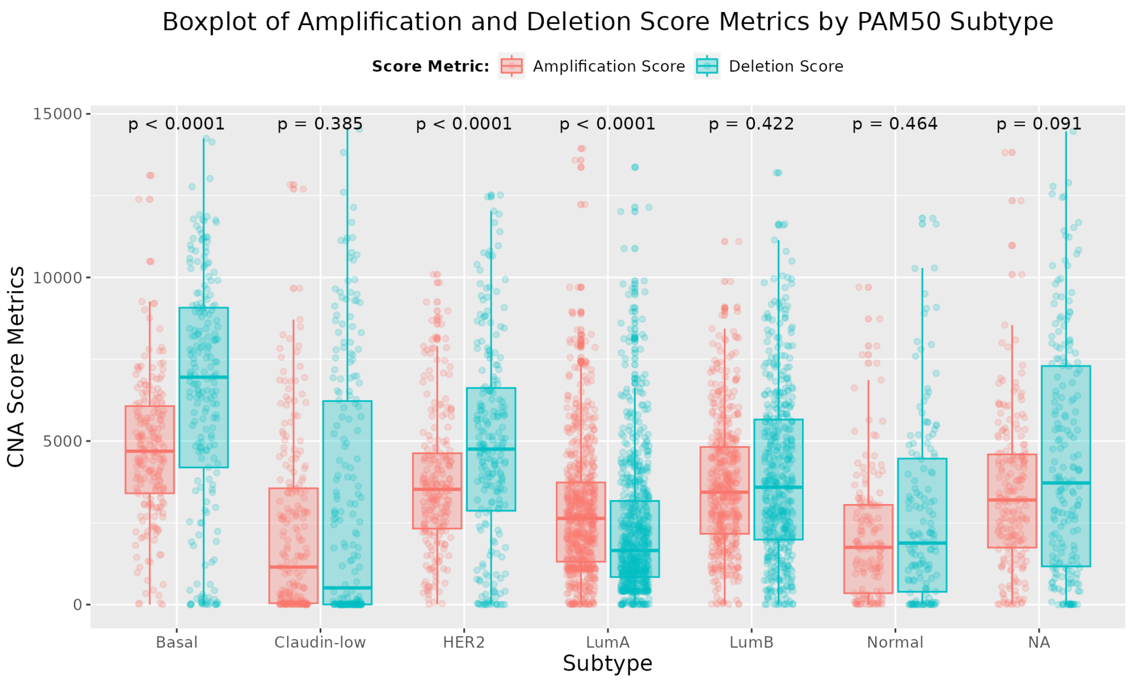


Figure 2.14: Boxplots for each CNA Amp and CNA Del Score metric by PAM50 subtype. Includes Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

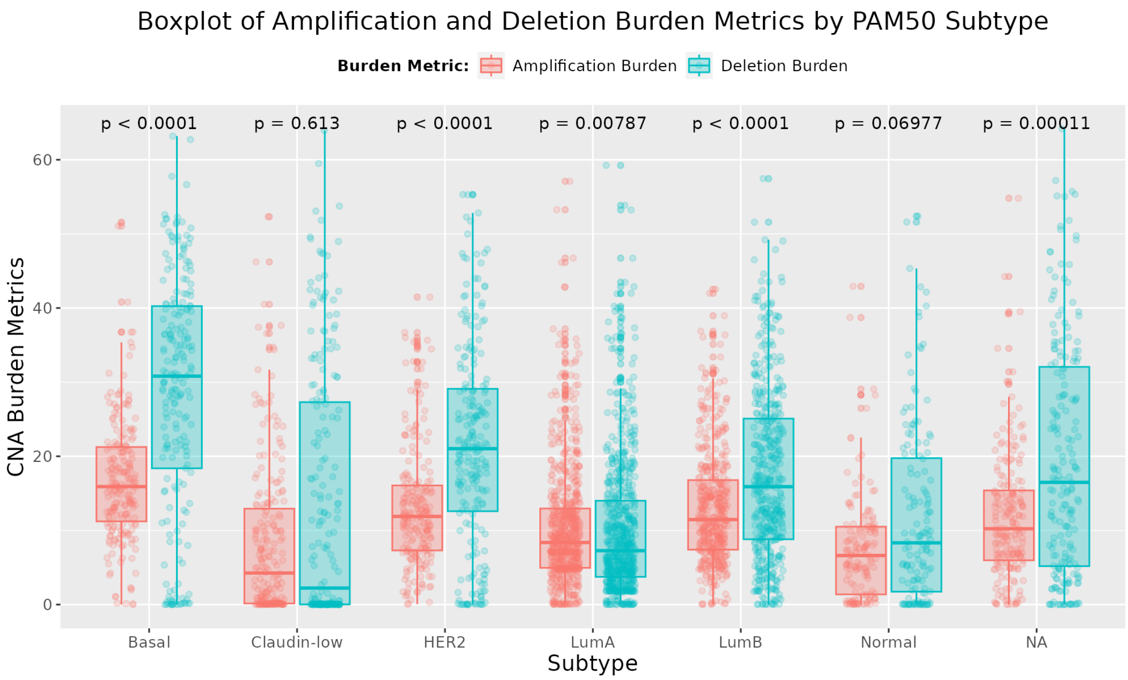


Figure 2.15: Boxplots for each CNA Amp and CNA Del Burden metric by PAM50 subtype. Includes Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

The observed distribution of the six CNA metrics, for patients stratified by IntClust is provided (Figures 2.16 and 2.17), accompanied by BH adjusted Kruskal-Wallis p-values. These figures indicate some significant difference exists comparing each of the CNA metric distributions across IntClusts (Kruskall-Wallis adjusted $p < 0.0001$ for all CNA metrics).

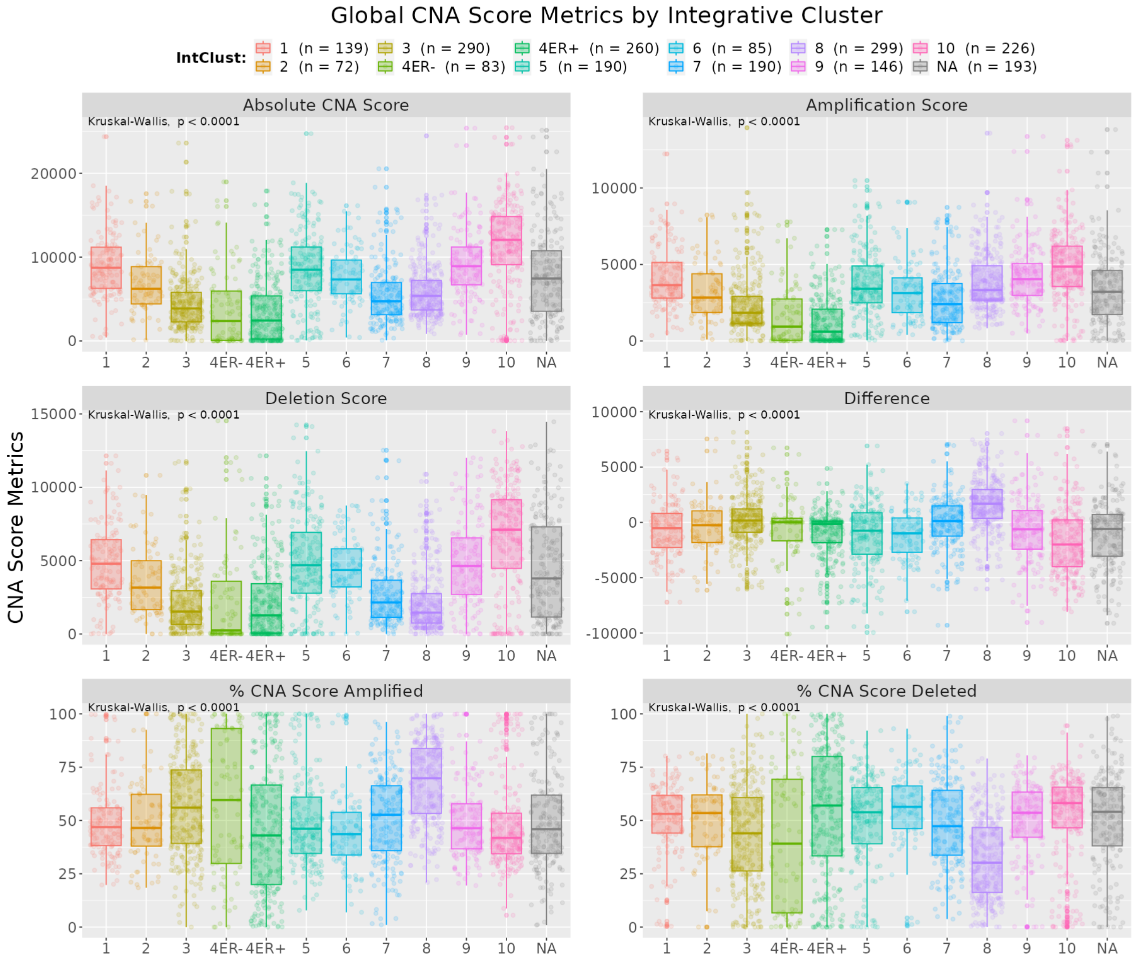


Figure 2.16: Boxplots for each CNA Score metric by IntClust. Each facet contains boxplots for the CNA Score metrics calculated using all available data and Benjamini-Hochberg adjusted Kruskal-Wallis p-values. NA denotes METABRIC patients missing IntClust information.

IntClust 10 displays the highest absolute, amplification and deletion CNA Score and Burden metrics (Dunn's test $p < 0.0001$ for most comparisons, Tables 2.16 and 2.17). This is not surprising as IntClust 10 is primarily made up of Basal patients (Figure 1.1). IntClust 1, 5 and 9 also display high levels of GI and are primarily composed of the PAM50 subtypes that have higher levels of GI, i.e. Basal, HER2 and Luminal B subtype. The densities of these distributions, apart from the CNA Amp metric distribution, are not significantly different from each other ($p > 0.05$). The levels of amplification burden observed in IntClust 5 are significantly lower than the levels observed in IntClust 1 and 9 ($p = 0.01$ and $p = 0.02$, respectively). IntClust 3, 4ER-, ER+, and 7 are primarily composed of Luminal A, Normal and Claudin-low subtypes (Figure 1.1). These subtypes display low levels of GI and as such the boxplots display the lowest CNA Score and Burden metrics, except

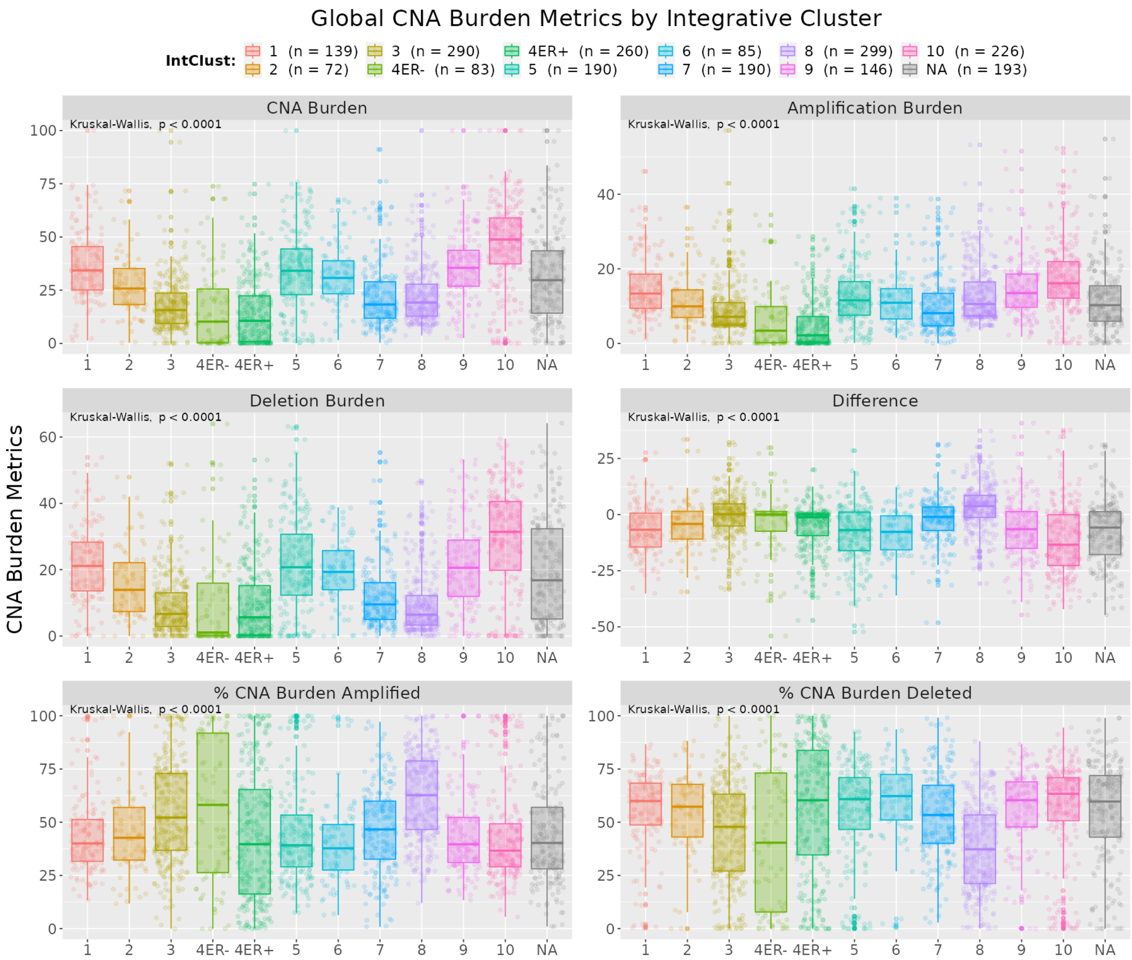


Figure 2.17: Boxplots for each CNA Burden metric by IntClust. Each facet contains boxplots for the CNA Burden metrics calculated using all available data and Benjamini-Hochberg adjusted Kruskal-Wallis p-values. NA denotes METABRIC patients missing IntClust information.

for CNA Del Score and Burden where IntClust 8 displays lower levels of deletions, when compared with IntClust 7. The distributions of IntClust 4ER- and 4ER+ do not significantly differ from each other, while IntClust 3 displays significantly more amplifications than IntClust 4ER+ and 4ER- ($p < 0.0001$ and $p < 0.01$, respectively), and IntClust 7 has significantly higher distributions across all metrics ($p < 0.01$, Tables 2.16 and 2.17).

Focusing on the direct comparison of levels of amplification to deletion within each IntClust, Figures 2.18 and 2.19, IntClust 2, 5 and 10, three clusters associated with poorer survival outcome, have significantly higher levels of deletions than amplifications ($p = 0.034$, $p < 0.0001$ and $p < 0.0001$, respectively). IntClust 1, 4ER+ and 6 also display significantly more deletions than amplifications but correspond to intermediate or good survival outcome ($p < 0.0001$, $p = 0.0005$ and $p < 0.0001$, respectively). The remaining IntClust classifications observed by Curtis et al. (2012) to have favourable survival outcomes, i.e. IntClust 3, 4ER-, 7 and 8, show no significant difference in the levels of amplifications and deletions, IntClust 3, 4ER-, 7 ($p > 0.05$) or significantly more amplifications than deletions, IntClust 8 ($p < 0.0001$).

Table 2.16: Comparisons of CNA Score metric distributions by Integrative Cluster. Z statistics and Benjamini-Hochberg adjusted p-values for each Dunn's test are shown.

Comparison of CNA Score Metrics by IntClust

Comparisons	Absolute CNA Score Z (adj p-value)	CNA Amp Score Z (adj p-value)	CNA Del Score Z (adj p-value)
10 - 4ER+	18.03 (<0.0001)	18.57 (<0.0001)	14.05 (<0.0001)
10 - 3	15.8 (<0.0001)	13.13 (<0.0001)	13.77 (<0.0001)
1 - 4ER+	12.74 (<0.0001)	13.79 (<0.0001)	9.65 (<0.0001)
10 - 4ER-	12.19 (<0.0001)	11.64 (<0.0001)	10.89 (<0.0001)
10 - 7	11.6 (<0.0001)	9.6 (<0.0001)	9.66 (<0.0001)
10 - 8	11.36 (<0.0001)	4.54 (<0.0001)	13.66 (<0.0001)
1 - 3	10.67 (<0.0001)	8.96 (<0.0001)	9.29 (<0.0001)
1 - 4ER-	9.1 (<0.0001)	9.03 (<0.0001)	8.17 (<0.0001)
1 - 7	7.53 (<0.0001)	6.31 (<0.0001)	6.15 (<0.0001)
5 - 7	6.9 (<0.0001)	4.81 (<0.0001)	5.79 (<0.0001)
1 - 8	6.82 (<0.0001)	1.56 (0.06)	9.16 (<0.0001)
2 - 4ER+	6.46 (<0.0001)	7.54 (<0.0001)	4.84 (<0.0001)
5 - 8	6.11 (<0.0001)	-0.55 (0.3)	9.13 (<0.0001)
10 - 2	5.76 (<0.0001)	5.06 (<0.0001)	4.68 (<0.0001)
2 - 4ER-	4.87 (<0.0001)	5.02 (<0.0001)	4.75 (<0.0001)
2 - 3	4.73 (<0.0001)	3.65 (<0.0001)	4.47 (<0.0001)
6 - 7	4.55 (<0.0001)	2.11 (0.02)	4.81 (<0.0001)
10 - 5	4.41 (<0.0001)	4.58 (<0.0001)	3.62 (<0.0001)
10 - 6	4.3 (<0.0001)	5.26 (<0.0001)	2.54 (0.01)
6 - 8	3.69 (<0.0001)	-2.19 (0.02)	7.17 (<0.0001)
1 - 2	3.29 (<0.0001)	3.06 (<0.0001)	2.54 (0.01)
10 - 9	2.8 (<0.0001)	1.96 (0.03)	2.83 (<0.0001)
3 - 4ER+	2.79 (<0.0001)	6.13 (<0.0001)	0.65 (0.29)
2 - 7	2.62 (0.01)	1.88 (0.03)	2.29 (0.01)
1 - 6	1.79 (0.04)	3.11 (<0.0001)	0.43 (0.36)
2 - 8	1.69 (0.05)	-2.17 (0.02)	4.35 (<0.0001)
3 - 4ER-	1.3 (0.11)	2.64 (0.01)	1.41 (0.1)

Comparison of CNA Score Metrics by IntClust

Comparisons	Absolute CNA Score Z (adj p-value)	CNA Amp Score Z (adj p-value)	CNA Del Score Z (adj p-value)
1 - 5	1.19 (0.12)	1.89 (0.03)	0.83 (0.23)
5 - 6	0.87 (0.2)	1.67 (0.05)	-0.25 (0.41)
4ER- - 4ER+	0.6 (0.28)	1.55 (0.06)	-0.95 (0.2)
1 - 9	-0.03 (0.49)	-0.27 (0.4)	0.31 (0.4)
5 - 9	-1.24 (0.12)	-2.21 (0.02)	-0.51 (0.34)
2 - 6	-1.45 (0.08)	-0.1 (0.46)	-1.93 (0.03)
7 - 8	-1.52 (0.07)	-5.87 (<0.0001)	2.73 (<0.0001)
6 - 9	-1.83 (0.04)	-3.38 (<0.0001)	-0.17 (0.43)
2 - 5	-2.5 (0.01)	-1.69 (0.05)	-2 (0.03)
3 - 7	-2.79 (<0.0001)	-2.36 (0.01)	-2.91 (<0.0001)
1 - 10	-2.79 (<0.0001)	-2.23 (0.02)	-2.45 (0.01)
4ER- - 7	-3.21 (<0.0001)	-4.17 (<0.0001)	-3.4 (<0.0001)
2 - 9	-3.35 (<0.0001)	-3.31 (<0.0001)	-2.31 (0.01)
4ER- - 8	-4.54 (<0.0001)	-8.81 (<0.0001)	-1.56 (0.07)
3 - 8	-4.87 (<0.0001)	-9.28 (<0.0001)	-0.22 (0.42)
4ER+ - 7	-5.22 (<0.0001)	-7.79 (<0.0001)	-3.43 (<0.0001)
4ER- - 6	-6.59 (<0.0001)	-5.34 (<0.0001)	-6.96 (<0.0001)
3 - 6	-6.93 (<0.0001)	-4.02 (<0.0001)	-7.29 (<0.0001)
8 - 9	-6.97 (<0.0001)	-1.9 (0.03)	-8.95 (<0.0001)
4ER+ - 8	-7.54 (<0.0001)	-15.2 (<0.0001)	-0.87 (0.23)
7 - 9	-7.67 (<0.0001)	-6.69 (<0.0001)	-5.91 (<0.0001)
4ER- - 5	-8.59 (<0.0001)	-7.92 (<0.0001)	-7.91 (<0.0001)
4ER+ - 6	-8.74 (<0.0001)	-8.16 (<0.0001)	-7.64 (<0.0001)
4ER- - 9	-9.21 (<0.0001)	-9.35 (<0.0001)	-7.98 (<0.0001)
3 - 5	-10.37 (<0.0001)	-7.65 (<0.0001)	-9.27 (<0.0001)
3 - 9	-10.89 (<0.0001)	-9.43 (<0.0001)	-9.08 (<0.0001)
4ER+ - 5	-12.63 (<0.0001)	-12.97 (<0.0001)	-9.65 (<0.0001)
4ER+ - 9	-12.98 (<0.0001)	-14.32 (<0.0001)	-9.45 (<0.0001)

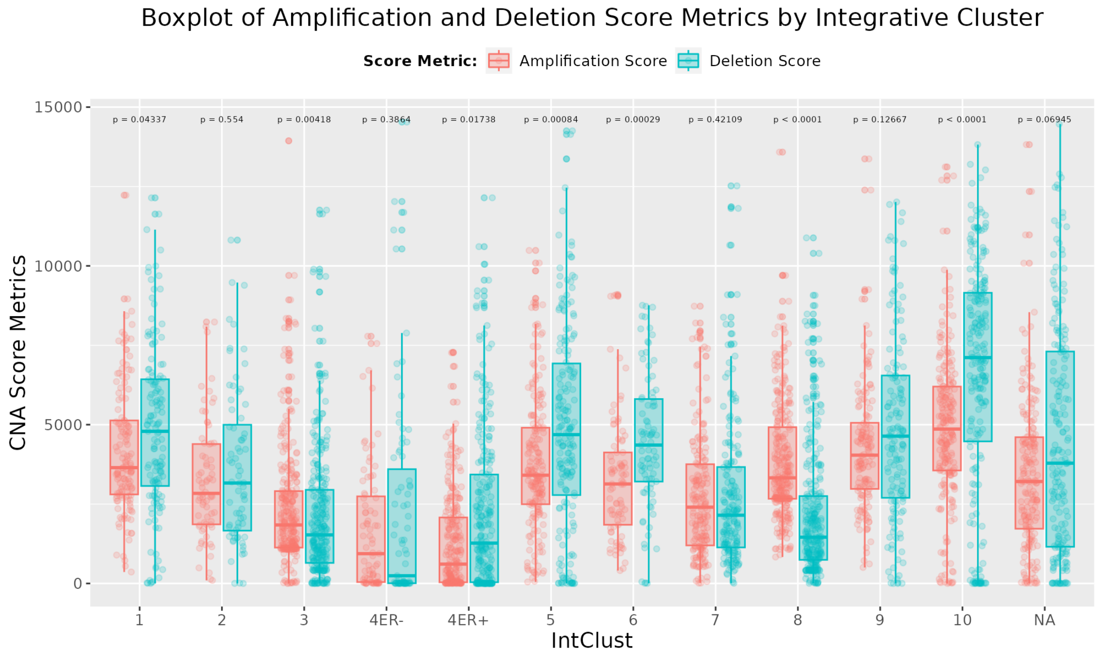


Figure 2.18: Boxplots for each CNA Amp and CNA Del Score metric by Integrative Cluster. Includes Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

Table 2.17: Comparisons of CNA Burden metric distributions by Integrative Cluster. Z statistics and Benjamini-Hochberg adjusted p-values for each Dunn's test are shown.

Comparison of CNA Burden Metrics by IntClust

Comparisons	Absolute CNA Burden Z (adj p-value)	CNA Amp Burden Z (adj p-value)	CNA Del Burden Z (adj p-value)
10 - 4ER+	17.74 (<0.0001)	18.09 (<0.0001)	14.07 (<0.0001)
10 - 3	15.48 (<0.0001)	11.89 (<0.0001)	13.8 (<0.0001)
10 - 8	12.5 (<0.0001)	5.44 (<0.0001)	13.65 (<0.0001)
1 - 4ER+	12.4 (<0.0001)	13.63 (<0.0001)	9.67 (<0.0001)
10 - 4ER-	11.96 (<0.0001)	11.05 (<0.0001)	10.91 (<0.0001)
10 - 7	11.64 (<0.0001)	9.37 (<0.0001)	9.67 (<0.0001)
1 - 3	10.3 (<0.0001)	8.16 (<0.0001)	9.32 (<0.0001)
1 - 4ER-	8.82 (<0.0001)	8.69 (<0.0001)	8.19 (<0.0001)
1 - 8	7.7 (<0.0001)	2.59 (0.01)	9.16 (<0.0001)
1 - 7	7.48 (<0.0001)	6.35 (<0.0001)	6.17 (<0.0001)
5 - 8	6.99 (<0.0001)	0.02 (0.49)	9.13 (<0.0001)
5 - 7	6.75 (<0.0001)	4.34 (<0.0001)	5.81 (<0.0001)
2 - 4ER+	6.35 (<0.0001)	7.52 (<0.0001)	4.87 (<0.0001)
10 - 2	5.68 (<0.0001)	4.75 (<0.0001)	4.67 (<0.0001)
6 - 7	4.79 (<0.0001)	2.42 (0.01)	4.83 (<0.0001)
2 - 4ER-	4.76 (<0.0001)	4.81 (<0.0001)	4.77 (<0.0001)
6 - 8	4.72 (<0.0001)	-1.04 (0.16)	7.18 (<0.0001)
10 - 5	4.6 (<0.0001)	4.85 (<0.0001)	3.62 (<0.0001)
2 - 3	4.6 (<0.0001)	3.13 (<0.0001)	4.5 (<0.0001)
10 - 6	4.09 (<0.0001)	4.77 (<0.0001)	2.53 (0.01)
1 - 2	3.15 (<0.0001)	2.96 (<0.0001)	2.53 (0.01)
10 - 9	3.1 (<0.0001)	2.29 (0.01)	2.84 (<0.0001)
3 - 4ER+	2.81 (<0.0001)	6.91 (<0.0001)	0.64 (0.29)
2 - 7	2.73 (<0.0001)	2.02 (0.03)	2.32 (0.01)
2 - 8	2.54 (0.01)	-1.25 (0.12)	4.36 (<0.0001)
1 - 6	1.52 (0.07)	2.86 (<0.0001)	0.42 (0.36)
3 - 4ER-	1.3 (0.11)	2.92 (<0.0001)	1.41 (0.1)

Comparison of CNA Burden Metrics by IntClust

Comparisons	Absolute CNA Burden Z (adj p-value)	CNA Amp Burden Z (adj p-value)	CNA Del Burden Z (adj p-value)
1 - 5	1.27 (0.11)	2.37 (0.01)	0.83 (0.23)
4ER- - 4ER+	0.63 (0.28)	1.8 (0.04)	-0.96 (0.2)
5 - 6	0.52 (0.31)	0.99 (0.17)	-0.27 (0.41)
1 - 9	0.15 (0.44)	0.25 (0.41)	0.32 (0.4)
7 - 8	-0.48 (0.32)	-4.78 (<0.0001)	2.71 (0.01)
5 - 9	-1.13 (0.14)	-2.13 (0.02)	-0.5 (0.34)
6 - 9	-1.41 (0.09)	-2.66 (0.01)	-0.15 (0.44)
2 - 6	-1.55 (0.07)	-0.23 (0.42)	-1.94 (0.03)
2 - 5	-2.28 (0.01)	-1.2 (0.13)	-1.99 (0.03)
3 - 7	-2.44 (0.01)	-1.42 (0.09)	-2.92 (<0.0001)
1 - 10	-2.88 (<0.0001)	-1.98 (0.03)	-2.44 (0.01)
4ER- - 7	-2.96 (<0.0001)	-3.77 (<0.0001)	-3.4 (<0.0001)
2 - 9	-3.05 (<0.0001)	-2.78 (<0.0001)	-2.29 (0.01)
3 - 8	-3.3 (<0.0001)	-6.99 (<0.0001)	-0.25 (0.41)
4ER- - 8	-3.49 (<0.0001)	-7.57 (<0.0001)	-1.58 (0.07)
4ER+ - 7	-4.9 (<0.0001)	-7.57 (<0.0001)	-3.43 (<0.0001)
4ER+ - 8	-6.04 (<0.0001)	-13.75 (<0.0001)	-0.9 (0.22)
4ER- - 6	-6.57 (<0.0001)	-5.26 (<0.0001)	-6.99 (<0.0001)
3 - 6	-6.91 (<0.0001)	-3.64 (<0.0001)	-7.32 (<0.0001)
7 - 9	-7.42 (<0.0001)	-6.17 (<0.0001)	-5.91 (<0.0001)
8 - 9	-7.65 (<0.0001)	-2.33 (0.01)	-8.94 (<0.0001)
4ER- - 5	-8.22 (<0.0001)	-7.15 (<0.0001)	-7.93 (<0.0001)
4ER+ - 6	-8.75 (<0.0001)	-8.31 (<0.0001)	-7.67 (<0.0001)
4ER- - 9	-8.77 (<0.0001)	-8.55 (<0.0001)	-7.99 (<0.0001)
3 - 5	-9.86 (<0.0001)	-6.19 (<0.0001)	-9.3 (<0.0001)
3 - 9	-10.29 (<0.0001)	-8 (<0.0001)	-9.1 (<0.0001)
4ER+ - 5	-12.16 (<0.0001)	-12.23 (<0.0001)	-9.68 (<0.0001)
4ER+ - 9	-12.42 (<0.0001)	-13.55 (<0.0001)	-9.46 (<0.0001)

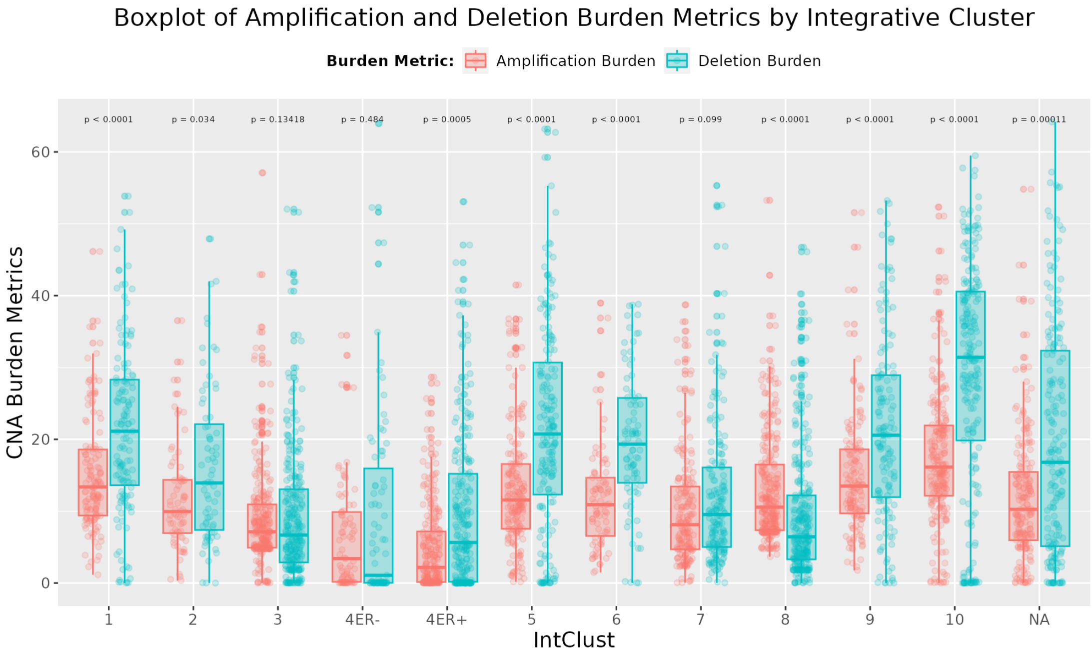


Figure 2.19: Boxplots for each CNA Amp and CNA Del Burden metric by Integrative Cluster. Includes Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

2.5.2 Observed Distributions for Chromosome Arm CNA Metrics across Molecular Subtype Classifications

A similar association analysis is conducted for the 42 chromosome arm CNA Score and Burden metrics. In general, we observe similar effects comparing stratified subgroups of patients in the chromosome arm metrics as observed in the global metrics.

Although Basal patients displayed widespread GI, some noteworthy alterations primarily observed in Basal patients include high levels of amplifications on chromosome 3q and 10p and deletions on chromosome 3p, 4p, 5q and 15q (Figure 2.20). Figure 2.20 indicates that some significant difference exists comparing each of the selected CNA Burden metric distributions across PAM50 subtype ($p < 0.0001$). Applying Dunn's Test to each CNA Burden metric indicates Basal patients display the highest CNA Burden metric across all subtypes, with $p < 0.0001$ for all selected CNA metrics and comparisons, indicating higher levels of GI on the specified chromosome arms when compared to other PAM50 subtypes (Table 2.18). The highlighted chromosome arms correspond largely to those frequently altered in tumours displaying the "complex I" pattern, which are usually Basal tumours, observed in Hicks et al. (2006).

Noteworthy alterations observed in HER2 patients include high levels of amplification on chromosome 1q, 8q and 17q, where the HER2 gene is located, and high levels of deletions on chromosome 8p, 17p and 17q (Figure 2.21). For each selected CNA Burden metric some significant difference exists comparing each of the CNA metric distributions across PAM50 subtype ($p < 0.0001$). Performing pairwise comparisons indicates that HER2 patients display higher CNA Burden metric across the majority of selected CNA Burden metrics and PAM50 subtypes ($p < 0.0001$) except when comparing to Basal patients (Table 2.18 and 2.19). The distributions of the selected CNA Burden metrics in the HER2 and Basal subtypes do not significantly differ from each other ($p > 0.05$) indicating similar levels of GI. Exceptions include CNA Amp Burden on chromosome 1q and CNA Amp Burden on chromosome 17q, where HER2 patients display lower and higher levels of amplifications, respectively, when compared with Basal patients ($p = 0.04$ and $p < 0.0001$). Interestingly high levels of deletions are also observed on chromosome 17q in HER2 patients, indicating amplification of HER2 locus may be correlated with widespread chromosome arm instability.

For the Luminal patients, high levels of GI are documented on chromosome 1q and 16p (amplifications) and chromosome 16q (deletions) (Figures 2.21 and 2.22). Luminal B patients display higher levels of whole genome instability than Luminal A patients (Tables 2.18-2.20). In particular, Luminal B patients display significantly more amplifications on chromosome 8q and 17q (Table 2.19) and deletions on chromosome 11q and 13q ($p < 0.0001$, Table 2.20). Luminal A patients display more amplifications on chromosome 16p, and more deletions on chromosome 16q than Luminal B patients ($p < 0.001$, Table 2.20).

Some alterations consistently observed across the PAM50 subtypes associated with poorer survival, i.e. Basal, HER2 and Luminal B, include high levels of amplification on chromosome 8q and 17q and high levels of deletions on chromosome 8p, 13q and 17p.

The observed patterns of instability, measured by our CNA Score and Burden metrics, within the IntClusts largely matched with what Curtis et al. (2012) documented previously. Other chromosome arms to note include, 3p and 4p, which display high levels of deletions for IntClust 10, and IntClust 5 and 10, respectively (Figure 2.23 and Table 2.21).

Overall, patients exhibiting the highest levels of GI across most of the chromosome arms correspond to the PAM50 and IntClusts associated with reduced survival. This is consistent with findings in the previous section, where patients with higher measures of GI generally have reduced survival.

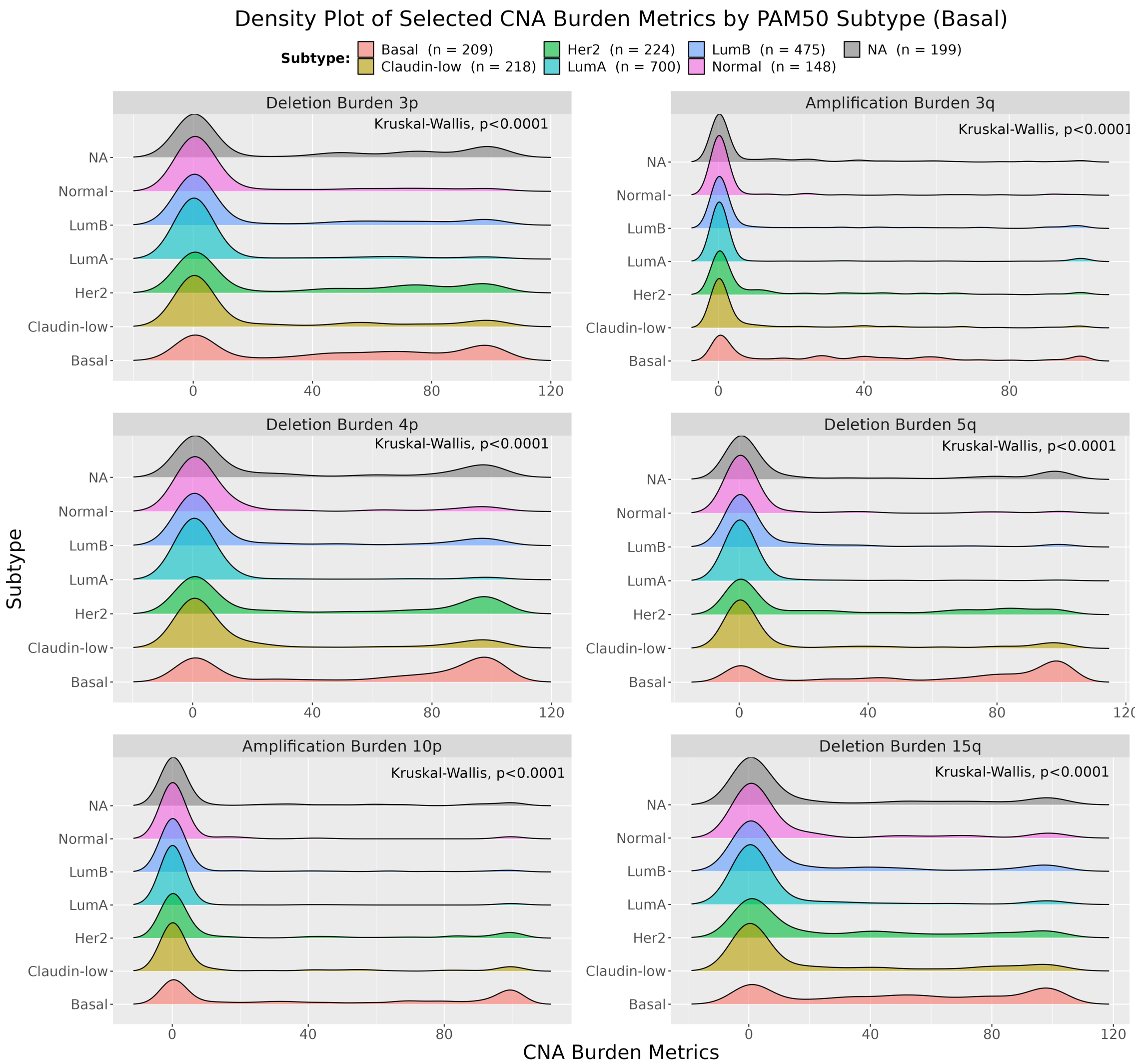


Figure 2.20: Density plots for each selected chromosome arm CNA Burden metrics, with a focus on the Basal subtype. Chromosome arms where Basal patients display high GI are selected. Each facet contains boxplots for the chromosome arm CNA Burden metrics calculated using all available data and Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

Table 2.18: Comparisons of selected chromosome arm CNA Burden metric distributions by PAM50 subtype, with a focus on the Basal subtype. Chromosome arms where Basal patients display high GI are selected. Z statistics and Benjamini-Hochberg adjusted p-values are shown.

Comparison of CNA Burden Metrics by PAM50 Subtype (Basal)						
Comparisons	CNA Del Burden 3p Z (adj p-value)	CNA Amp Burden 3q Z (adj p-value)	CNA Del Burden 4p Z (adj p-value)	CNA Del Burden 5q Z (adj p-value)	CNA Amp Burden 10p Z (adj p-value)	CNA Del Burden 15q Z (adj p-value)
Basal - LumA	14.16 (<0.0001)	10.07 (<0.0001)	15.66 (<0.0001)	17.82 (<0.0001)	17.27 (<0.0001)	14.52 (<0.0001)
Her2 - LumA	9.16 (<0.0001)	4.13 (<0.0001)	11.26 (<0.0001)	10.84 (<0.0001)	7.69 (<0.0001)	8.14 (<0.0001)
Basal - LumB	8.66 (<0.0001)	8.03 (<0.0001)	9.97 (<0.0001)	13.96 (<0.0001)	13.43 (<0.0001)	10.08 (<0.0001)
Basal - Normal	8.21 (<0.0001)	8.76 (<0.0001)	9.01 (<0.0001)	11.55 (<0.0001)	11.53 (<0.0001)	8.77 (<0.0001)
Basal - Claudin-low	7.98 (<0.0001)	8.99 (<0.0001)	9.34 (<0.0001)	10.6 (<0.0001)	9.91 (<0.0001)	9.23 (<0.0001)
Claudin-low - LumA	4.43 (<0.0001)	-0.99 (0.17)	4.25 (<0.0001)	4.88 (<0.0001)	5.18 (<0.0001)	3.24 (<0.0001)
Her2 - Normal	4.42 (<0.0001)	4.39 (<0.0001)	5.64 (<0.0001)	6.31 (<0.0001)	4.42 (<0.0001)	3.99 (<0.0001)
Basal - Her2	4.3 (<0.0001)	4.95 (<0.0001)	3.85 (<0.0001)	5.95 (<0.0001)	8.02 (<0.0001)	5.4 (<0.0001)
Her2 - LumB	3.77 (<0.0001)	2.35 (0.01)	5.65 (<0.0001)	7.23 (<0.0001)	4.24 (<0.0001)	3.91 (<0.0001)
LumB - Normal	1.73 (0.05)	2.92 (<0.0001)	1.49 (0.08)	0.87 (0.19)	1.32 (0.09)	1.12 (0.15)
Claudin-low - Normal	1.02 (0.16)	0.66 (0.25)	0.59 (0.28)	2.02 (0.03)	2.62 (0.01)	0.46 (0.32)
Claudin-low - LumB	-0.65 (0.26)	-2.49 (0.01)	-0.94 (0.19)	1.62 (0.06)	1.89 (0.03)	-0.69 (0.26)
LumA - Normal	-2.59 (0.01)	1.63 (0.06)	-2.95 (<0.0001)	-1.82 (0.04)	-1.35 (0.1)	-2.24 (0.02)
Claudin-low - Her2	-3.78 (<0.0001)	-4.14 (<0.0001)	-5.62 (<0.0001)	-4.77 (<0.0001)	-1.98 (0.03)	-3.93 (<0.0001)
LumA - LumB	-6.68 (<0.0001)	-2.14 (0.02)	-6.84 (<0.0001)	-4.14 (<0.0001)	-4.15 (<0.0001)	-5.18 (<0.0001)

Table 2.19: Comparisons of selected chromosome arm CNA Burden metric distributions by PAM50 subtype, with a focus on the HER2 subtype. Chromosome arms where HER2 patients display high GI are selected. Z statistics and Benjamini-Hochberg adjusted p-values are shown.

Comparison of CNA Burden Metrics by PAM50 Subtype (HER2)						
Comparisons	CNA Amp Burden 1q Z (adj p-value)	CNA Del Burden 8p Z (adj p-value)	CNA Amp Burden 8q Z (adj p-value)	CNA Del Burden 17p Z (adj p-value)	CNA Amp Burden 17q Z (adj p-value)	CNA Del Burden 17q Z (adj p-value)
Basal - Claudin-low	7.49 (<0.0001)	6.95 (<0.0001)	7.35 (<0.0001)	7.07 (<0.0001)	4.34 (<0.0001)	7.64 (<0.0001)
LumA - Normal	5.64 (<0.0001)	-1.94 (0.03)	0.97 (0.21)	-1.33 (0.11)	0.09 (0.46)	-3.93 (<0.0001)
LumB - Normal	3.79 (<0.0001)	2.96 (<0.0001)	6.47 (<0.0001)	2.39 (0.01)	6.1 (<0.0001)	-0.59 (0.3)
Basal - Normal	3.41 (<0.0001)	5 (<0.0001)	7.07 (<0.0001)	4.53 (<0.0001)	3.53 (<0.0001)	6.2 (<0.0001)
LumA - LumB	2.59 (0.01)	-7.64 (<0.0001)	-8.77 (<0.0001)	-5.82 (<0.0001)	-9.52 (<0.0001)	-5.06 (<0.0001)
Basal - Her2	1.86 (0.04)	-0.8 (0.23)	1.26 (0.14)	-0.62 (0.27)	-4.12 (<0.0001)	1.24 (0.13)
Her2 - Normal	1.77 (0.04)	5.8 (<0.0001)	6.03 (<0.0001)	5.16 (<0.0001)	7.32 (<0.0001)	5.17 (<0.0001)
Basal - LumB	0.12 (0.45)	3.11 (<0.0001)	1.81 (0.05)	3.15 (<0.0001)	-2.34 (0.01)	8.69 (<0.0001)
Basal - LumA	-1.82 (0.04)	9.04 (<0.0001)	8.53 (<0.0001)	7.71 (<0.0001)	4.71 (<0.0001)	12.97 (<0.0001)
Her2 - LumB	-2.08 (0.03)	4.14 (<0.0001)	0.37 (0.36)	3.96 (<0.0001)	2.49 (0.01)	7.43 (<0.0001)
Claudin-low - Normal	-3.36 (<0.0001)	-1.27 (0.12)	0.46 (0.35)	-1.85 (0.04)	-0.38 (0.38)	-0.69 (0.28)
Her2 - LumA	-4.21 (<0.0001)	10.29 (<0.0001)	7.18 (<0.0001)	8.69 (<0.0001)	10 (<0.0001)	11.77 (<0.0001)
Claudin-low - Her2	-5.74 (<0.0001)	-7.88 (<0.0001)	-6.2 (<0.0001)	-7.82 (<0.0001)	-8.58 (<0.0001)	-6.52 (<0.0001)
Claudin-low - LumB	-8.74 (<0.0001)	-5.07 (<0.0001)	-6.85 (<0.0001)	-5.17 (<0.0001)	-7.52 (<0.0001)	-0.22 (0.41)
Claudin-low - LumA	-11.2 (<0.0001)	0.51 (0.3)	-0.5 (0.35)	-0.99 (0.17)	-0.63 (0.3)	3.64 (<0.0001)

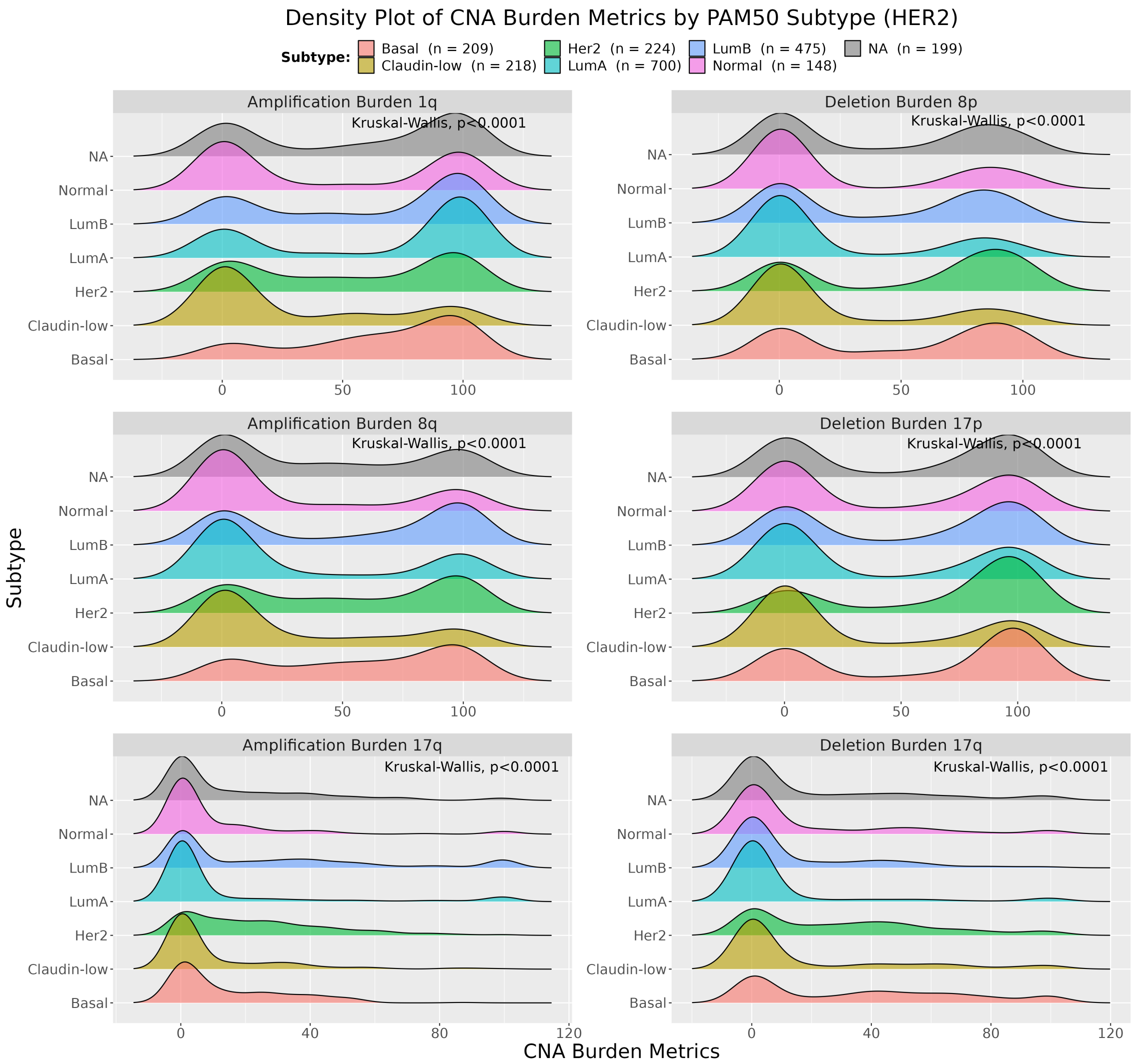


Figure 2.21: Density plots for each selected chromosome arm CNA Burden metrics, with a focus on the HER2 subtype. Chromosome arms where HER2 patients display high GI are selected. Each facet contains boxplots for the chromosome arm CNA Burden metrics calculated using all available data and Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

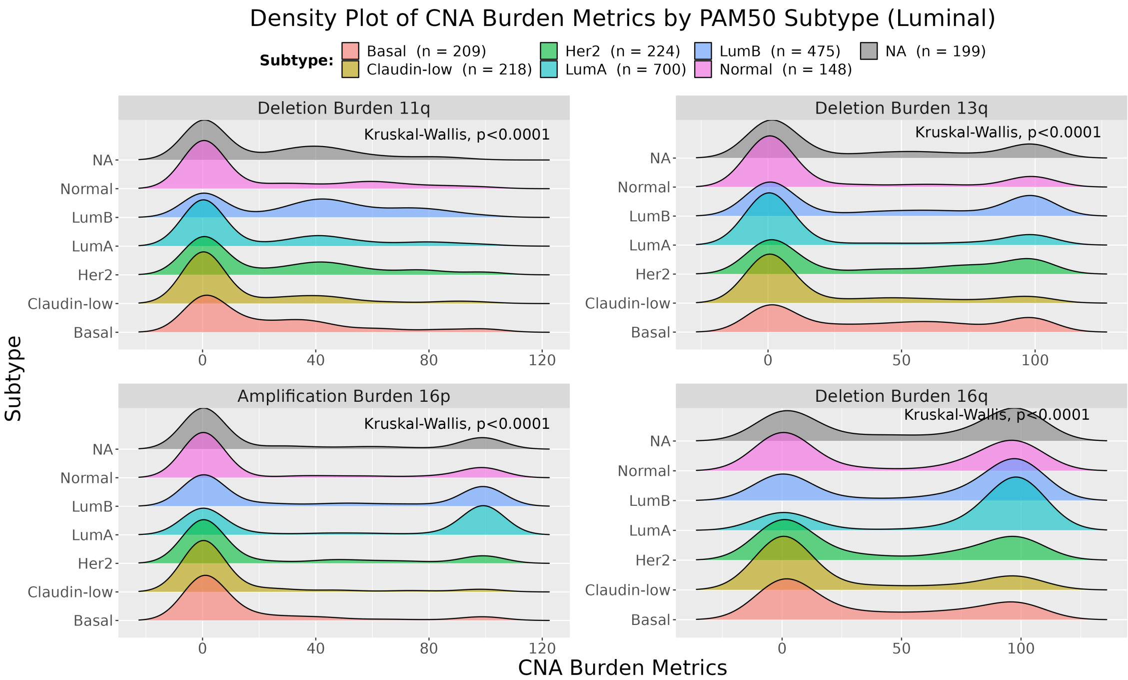


Figure 2.22: Density plots for each selected chromosome arm CNA Burden metrics, with a focus on the Luminal subtype. Chromosome arms where Luminal patients display high GI are selected. Each facet contains boxplots for the chromosome arm CNA Burden metrics calculated using all available data and Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

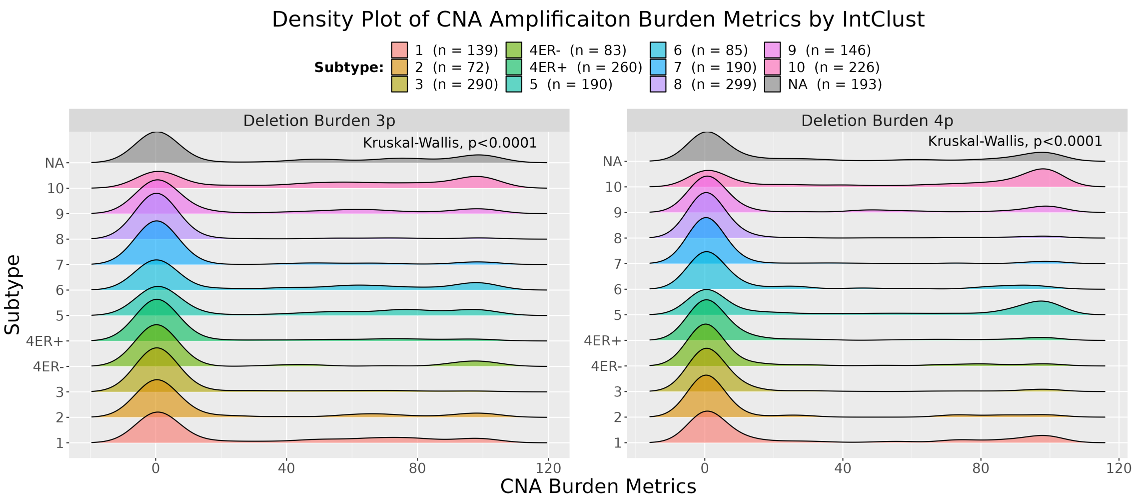


Figure 2.23: Density plots for each selected chromosome arm CNA Burden metrics across Integrative Cluster. Each facet contains boxplots for the chromosome arm CNA Burden metrics calculated using all available data and Benjamini-Hochberg adjusted Kruskal-Wallis p-values.

Table 2.20: Comparisons of selected chromosome arm CNA Burden metric distributions by PAM50 subtype, with a focus on the Luminal subtype. Chromosome arms where Luminal patients display high GI are selected. Z statistics and Benjamini-Hochberg adjusted p-values are shown.

Comparison of CNA Burden Metrics by PAM50 Subtype (Luminal)

Comparisons	CNA Del Burden 11q Z (adj p-value)	CNA Del Burden 13q Z (adj p-value)	CNA Amp Burden 16p Z (adj p-value)	CNA Del Burden 16q Z (adj p-value)
LumB - Normal	6.29 (<0.0001)	5.08 (<0.0001)	4.39 (<0.0001)	3.6 (<0.0001)
Basal - Claudin-low	4.93 (<0.0001)	7.3 (<0.0001)	3.35 (<0.0001)	3.67 (<0.0001)
Basal - LumA	3.43 (<0.0001)	9.08 (<0.0001)	-6.88 (<0.0001)	-8.18 (<0.0001)
Her2 - LumA	3.01 (<0.0001)	7.63 (<0.0001)	-6.93 (<0.0001)	-8.33 (<0.0001)
Basal - Normal	2.32 (0.01)	5.59 (<0.0001)	0.72 (0.25)	-0.19 (0.49)
Her2 - Normal	1.98 (0.03)	4.44 (<0.0001)	0.82 (0.24)	-0.15 (0.47)
Basal - Her2	0.41 (0.36)	1.35 (0.12)	-0.1 (0.46)	-0.05 (0.48)
LumA - Normal	-0.23 (0.41)	-1.27 (0.13)	6.85 (<0.0001)	6.89 (<0.0001)
Claudin-low - Normal	-2.13 (0.02)	-0.99 (0.19)	-2.32 (0.01)	-3.53 (<0.0001)
Claudin-low - LumA	-2.66 (0.01)	0.12 (0.48)	-11.17 (<0.0001)	-12.9 (<0.0001)
Basal - LumB	-4.13 (<0.0001)	1.48 (0.1)	-4.04 (<0.0001)	-4.34 (<0.0001)
Claudin-low - Her2	-4.59 (<0.0001)	-6.06 (<0.0001)	-3.51 (<0.0001)	-3.79 (<0.0001)
Her2 - LumB	-4.72 (<0.0001)	-0.09 (0.46)	-4.02 (<0.0001)	-4.38 (<0.0001)
Claudin-low - LumB	-10.02 (<0.0001)	-7.13 (<0.0001)	-8.07 (<0.0001)	-8.75 (<0.0001)
LumA - LumB	-10.32 (<0.0001)	-9.97 (<0.0001)	3.47 (<0.0001)	4.78 (<0.0001)

Table 2.21: Comparisons of selected chromosome arm CNA Burden metric distributions by Integrative Cluster. Z statistics and Benjamini-Hochberg adjusted p-values are shown.

Comparison of CNA Burden Metrics by IntClust

Comparisons	CNA Del Burden 3p Z (adj p-value)	CNA Del Burden 4p Z (adj p-value)
10 - 8	14.4 (<0.0001)	14.56 (<0.0001)
10 - 3	12.98 (<0.0001)	13.57 (<0.0001)
10 - 4ER+	11.7 (<0.0001)	12.51 (<0.0001)
10 - 7	11.62 (<0.0001)	12.94 (<0.0001)
5 - 8	8.56 (<0.0001)	9.49 (<0.0001)
10 - 4ER-	8.06 (<0.0001)	9.43 (<0.0001)
1 - 8	7.56 (<0.0001)	6.8 (<0.0001)
10 - 2	6.78 (<0.0001)	8.34 (<0.0001)
5 - 7	6.51 (<0.0001)	8.49 (<0.0001)
1 - 3	6.39 (<0.0001)	6 (<0.0001)
10 - 9	6.35 (<0.0001)	6.34 (<0.0001)
6 - 8	5.98 (<0.0001)	2.93 (<0.0001)
1 - 7	5.83 (<0.0001)	6.17 (<0.0001)
1 - 4ER+	5.43 (<0.0001)	5.26 (<0.0001)
10 - 5	4.83 (<0.0001)	4.09 (<0.0001)
6 - 7	4.67 (<0.0001)	2.68 (0.01)
10 - 6	4.19 (<0.0001)	7.26 (<0.0001)
1 - 4ER-	3.91 (<0.0001)	4.5 (<0.0001)
1 - 2	2.92 (<0.0001)	3.74 (<0.0001)
2 - 8	2.68 (0.01)	1.18 (0.16)
4ER+ - 8	2.42 (0.01)	1.71 (0.06)
4ER- - 8	1.89 (0.05)	0.59 (0.32)
5 - 9	1.81 (0.05)	2.45 (0.01)
2 - 3	1.79 (0.05)	0.58 (0.32)
2 - 7	1.64 (0.07)	1.05 (0.19)
1 - 9	1.53 (0.08)	0.74 (0.27)
3 - 8	1.42 (0.1)	0.96 (0.21)

Comparison of CNA Burden Metrics by IntClust

Comparisons	CNA Del Burden 3p Z (adj p-value)	CNA Del Burden 4p Z (adj p-value)
7 - 8	1.35 (0.11)	0.11 (0.48)
2 - 4ER+	1.1 (0.17)	0.07 (0.48)
6 - 9	1.03 (0.18)	-1.84 (0.05)
4ER+ - 7	0.84 (0.23)	1.42 (0.11)
4ER- - 7	0.82 (0.23)	0.48 (0.33)
2 - 4ER-	0.73 (0.26)	0.51 (0.33)
5 - 6	0.45 (0.35)	3.99 (<0.0001)
1 - 6	0.29 (0.41)	2.46 (0.01)
4ER- - 4ER+	0.23 (0.43)	-0.57 (0.31)
3 - 7	-0.09 (0.46)	0.74 (0.27)
1 - 5	-0.16 (0.44)	-1.64 (0.07)
3 - 4ER-	-0.94 (0.2)	0.04 (0.48)
3 - 4ER+	-1.04 (0.18)	-0.78 (0.27)
2 - 9	-1.68 (0.06)	-3.16 (<0.0001)
2 - 6	-2.39 (0.01)	-1.28 (0.13)
4ER- - 9	-2.62 (0.01)	-3.91 (<0.0001)
2 - 5	-3.19 (<0.0001)	-5.24 (<0.0001)
4ER- - 6	-3.25 (<0.0001)	-1.86 (0.05)
4ER+ - 9	-3.76 (<0.0001)	-4.5 (<0.0001)
4ER+ - 6	-4.24 (<0.0001)	-1.72 (0.06)
4ER- - 5	-4.25 (<0.0001)	-6.14 (<0.0001)
7 - 9	-4.26 (<0.0001)	-5.46 (<0.0001)
1 - 10	-4.58 (<0.0001)	-5.43 (<0.0001)
3 - 9	-4.71 (<0.0001)	-5.24 (<0.0001)
3 - 6	-5.01 (<0.0001)	-2.28 (0.02)
8 - 9	-5.89 (<0.0001)	-6.05 (<0.0001)
4ER+ - 5	-6.16 (<0.0001)	-7.71 (<0.0001)
3 - 5	-7.25 (<0.0001)	-8.59 (<0.0001)

2.6 Conclusions

GI plays an important role in the initiation and progression of cancer and can influence patient prognosis. There are myriad ways to try to quantify the levels of GI within tumour samples, from both gene expression and CNA data, a number of these and their limitations have been discussed in this chapter.

We proposed a number of novel CNA Score and Burden metrics quantifying GI of a patient, calculated globally across the full genome and for each chromosome arm. These comprehensible metrics, applicable to publicly available data, quantify GI in totality for all aberration types, and also quantify GI attributed to amplifications and deletions.

It was observed that the presence of missing values, in existence for some patients, has a negligible effect on both the global and chromosome arm CNA metric distributions. As a result, the approach of using all of the available patient CNA Score and Burden data, as opposed to including only complete cases, is adopted in the downstream analysis. Analysing distributions of the CNA metrics comparing groups of patients stratified by molecular classifications PAM50 and IntClust offered interesting observations. We see concordance with characteristic genomic aberrations documented previously, such as high quantification of deletion burden on chromosome 5q within Basal tumours and high quantification of amplification on chromosome 17q within HER2 tumours. Focusing on the direct comparison of levels of amplification metrics to deletion metrics, we observe the subtypes associated with worse OS and DSS tend to have significantly higher quantified deletion burden. As the deletion landscape of breast cancer has been poorly characterised, our novel findings that PAM50 and IntClust classifications with poorer OS and DSS have higher levels of GI, in particular higher CNA deletion burden, encourage further investigation.

In the next chapter we investigate how the quantified levels of CNA Scores and Burden correlate with survival outcomes.

3 Association of Copy Number Alteration Signatures and Survival Outcomes

Time-to-event analysis is often used to estimate the extent to which potential biomarkers can differentiate on outcomes of disease progression, survival, recurrence, and response to treatment (Kleinbaum and Klein, 2012). Time-to-event analysis focuses on the length of time until the occurrence of an event. This event may be the time to death from the disease, time to death from other causes, time to relapse or time to response to treatment. The data utilised in survival analysis are often termed survival data and have two key characteristics: that the response variable, representing the time to event occurrence, is a non-negative discrete or continuous random variable and that censoring is present. Censoring occurs when there is some information about an individual's time-to-event/survival time, but the time is not precisely observed. There are several reasons why an observation may be censored, such as the case where an individual has withdrawn from the study, where an individual is lost to follow-up during the study period, or where an individual has not experienced the event in the study period. These cases are all examples of right censoring. Other forms of censoring include left censoring, where the event of interest has already occurred before the study starts, and interval censoring, where the event occurs at an unknown time in an interval. Thus, survival data comprise a measurement indicating survival time and a binary indicator of the outcome occurring (Kleinbaum and Klein, 2012; Moore, 2016).

The primary goals of survival analysis are to estimate and interpret survival and/or hazard functions from sample survival data, to compare survival and/or hazard functions across stratified subcohorts of patients and to assess the relationship of candidate predictor variables in modelling time to event (Kleinbaum and Klein, 2012). Statistical methods for modelling the survival and hazard functions can be categorised into parametric, semi-parametric and non-parametric techniques (Kleinbaum and Klein, 2012). Parametric approaches are used when the survival time is assumed to follow an identifiable probability distribution such as the Weibull distribution (Crowther and Lambert, 2014). The most common non-parametric approach in application is the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function with log-rank tests (Bland and Altman, 2004) to compare survivor functions and the most commonly applied semi-parametric approach is the Cox proportional hazards (CPH) regression (Cox, 1972), used for modelling the association between survival time and one or more predictor variables. Survival trees, a more recent non-parametric approach, employ machine learning techniques to sequentially partition patients into groups that display similarity in survival functions (Lee and Lim, 2019).

This chapter presents an overview of survival analysis, discusses parametric, semi-parametric, and non-parametric survival approaches, with application to the METABRIC cohort, to measure any association of the derived and quantified CNA Score and Burden metrics on survival outcomes. This will assess the prognostic potential of the CNA metrics, alone and in combination with selected clinical and molecular features, and as such help identify patients who may be at a higher risk.

3.1 Survival Analysis Methods

Two critical functions in survival analysis are the survivor function ($S(t)$) and hazard function ($h(t)$) (Kleinbaum and Klein, 2012). The survivor function is a function of time t and denotes the probability that a person survives longer than some specified time t. Another way to think about the survivor function is the probability of a patient not experiencing an event. This function provides survival probabilities for different values of t, giving a full summary of the survival distribution, and can be specified by:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)du \tag{3.1}$$

where T represents a random variable denoting time to event occurring, e.g. the time from diagnosis to occurrence of death or disease relapse, and $F(t)$ is the cumulative distribution function for T. Theoretically, as t goes from 0 to infinity, the survivor function can be graphed as a smooth curve, and has the following properties:

- Is a non-increasing function.
- The probability of surviving past time 0 is 1, i.e. under the assumption the cohort under observation present at the beginning of the study without occurrence of the event, $S(t) = S(0) = 1$.
- As time goes to infinity the survivor curve tends towards 0, i.e. at time $t = \infty$, $S(t) = S(\infty) = 0$. However, as observed in a cure model, it is possible for $S(\infty) > 0$.

In contrast to the survivor function, the hazard function ($(h(t))$) gives the instantaneous rate of failure per unit time, given that the individual has survived up to time t and is specified as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{3.2}$$

where Δt denotes a small interval of time. For a particular value of t, $h(t)$ has the following characteristics:

- $h(t)$ is always non-negative, i.e. $h(t) \ge 0$ for all $t > 0$.
- $h(t)$ has no upper bound.

Like the survivor function, the hazard function can be graphed where t is on the x-axis and $h(t)$ is on the y-axis. In this case, $h(t)$ does not have to start at 1, it may start anywhere and fluctuate up and down over time. These characteristics result in the hazard function taking on different forms depending on the nature of the study. For example, in a study where an individual remains healthy for the course of study, their instantaneous potential for event occurrence remains constant so that in this case, $h(t)$ remains the same constant value for all t. When the hazard function is constant, the survival model is called exponential. Another example might be a study of cancer patients and their response to treatment, where the event of interest is death due to disease. In this case the patient's potential for the event occurring, i.e. death, increases as t increases. This is called an increasing Weibull model. Other types of hazard functions include the decreasing Weibull model ($h(t)$ decreases over time), the lognormal survival model ($h(t)$ increases and then decreases over time) and the bathtub model ($h(t)$ decreases and then increases over time).

Even though these two functions differ in the fact that $S(t)$ directly describes survival and $h(t)$ is a measure of instantaneous potential, if $S(t)$ or $h(t)$ is known, then the other may be determined. The relationship between $S(t)$ and $h(t)$ can be expressed using:

$$S(t) = exp(-\int_0^t h(u)du) \tag{3.3}$$

$$h(t) = -(\frac{dS(t)/dt}{S(t)}) \tag{3.4}$$

In parametric survival models the response, survival time, is assumed to follow an identifiable probability distribution. Parametric models are used in cases where information about the event process in a population point towards a particular distribution and once this probability density function, $f(t)$, is defined for survival time, the survivor and hazard functions can be calculated using Equations 3.1 and 3.2. Some commonly used distributions include exponential, Weibull, log-logistic, generalised gamma, and lognormal. The main appeal in using parametric survival models include their simplicity and completeness. In the case where the assumed parametric model is correct, the parameter estimates can completely specify the survival and hazard functions. However, the use of parametric survival models is not always appropriate or sufficiently flexible and, in these cases, semi-parametric or non-parametric survival models may be used (Kleinbaum and Klein, 2012).

Semi-parametric models have both parametric and non-parametric components. An example of a semi-parametric survival model is the CPH model (Cox, 1972; Kleinbaum and Klein, 2012). The CPH model uses non-parametric methods to estimate some baseline hazard, i.e. leaves the distribution of the baseline hazard function unspecified, and parametric methods to estimate the influence of predictor variables. The CPH model enables the incorporation of one or more predictor variables and as such is commonly used to investigate the association between survival and one or more categorical or continuous predictor variables, while accounting for possible confounders. Some examples of clinical predictor variables used in the context of breast cancer survival include number of lymph-nodes positive, tumour size, tumour grade, tumour subtype, ER/PR/HER2 status and type of treatment.

Although this model is semi-parametric and as such does not require knowledge of the underlying distribution, it does have required assumptions. Two of the main assumptions are that hazard functions of different individuals are assumed to be proportional over time t and the relationship between the log hazard and each predictor is assumed to be linear. Violations of these assumptions may lead to inaccurate results, i.e. biased effect estimates, and reduced statistical power (Zeng et al., 2022). These assumptions can be tested using the Schoenfeld and Martingale residuals (Patil and Dessai, 2019).

The CPH model facilitates quantification of the differences in survival distribution between groups. This is done by estimating the hazard ratio, defined as the ratio of the event rate at any given time in one group relative to another group. The CPH regression model can be written as follows:

$$h(t, X) = h_0(t) \times exp(\sum_{i=1}^{p} \beta_i X_i) \tag{3.5}$$

where $h(t)$ is the expected hazard at time t, $h_0(t)$ is the baseline hazard function that determines the shape of the survivor function and represents the hazard when all of the predictors are equal to zero, the X_i represent the predictor variables in the model and the β_i are the regression coefficients. The predicted hazard, i.e. $h(t)$, is the product of the baseline hazard $h_0(t)$ and the exponential function of the linear combination of the predictors, having a multiplicative or proportional effect on the predicted hazard.

Since there is a corresponding relationship between $S(t)$ and $h(t)$, when a CPH model is applied survival estimates can be obtained that adjust for the explanatory variables used as predictors in the model. Using Equation 3.3 the survivor function corresponding to the CPH regression model can be written as follows:

$$S(t|X) = exp(-\int_0^t h(u)du \times exp(\sum_{i=1}^{p} \beta_i X_i)) = S_0(t)^{exp(\beta' X)} \tag{3.6}$$

where $S_0(t)$ is the baseline survivor function and $exp(\beta' X)$ is called the prognostic index. This survivor function is the basis for producing adjusted survival curves.

The CPH model is popular in application for many reasons. One of the main reasons is that even though the baseline hazard, $h_0(t)$, is an unspecified function, the CPH model is robust and therefore will closely approximate the results for the correct parametric model, i.e. provide good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves. Other properties of the CPH model that make it appealing include that the model will always produce non-negative hazard ratios, an estimated baseline hazard function is not necessary for the estimation of the hazard ratio and there are many widely available computer packages supporting application of CPH models. Of course, there are cases where the assumptions of the CPH model are not met and, in those cases, non-parametric methods may be considered (Kleinbaum and Klein, 2012).

In cases where parametric and semi-parametric survival models are not appropriate or sufficiently flexible, non-parametric methods, not making any specific assumptions about the distribution of survival time, are applicable, we focus on two such approaches, the Kaplan-Meier (KM) estimator and survival trees.

3.1.1 Kaplan-Meier Estimator

In the case where there are no censored observations in the survival data, the survivor function can be estimated using the empirical survivor function, where, as the sample size increases the function will approach the shape of the population's true survivor function. In this case the survivor function can be estimated using:

$$\hat{S}(t) = \frac{\text{Number of observations with } T > t}{\text{Total sample size}} \tag{3.7}$$

However, cases where there are no censored observations in the data are rare and so an extension of this method, termed the KM estimate, was developed to analyse survival data with censoring (Kaplan and Meier, 1958). The survivor function is estimated using the KM estimator and is defined as the fraction of observations that survived for a certain amount of time under the same circumstances, e.g. after treatment or disease diagnosis, and is given by the following formula:

$$\hat{S}(t) = P(T > t) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i}) \tag{3.8}$$

where n_i is the number of subjects at risk at time t_i, and d_i is the number of subjects who fail at time t_i.

The KM estimator is a continuous decreasing step function that starts at a survival probability of 1 and then steps down as you move from one ordered failure time to another, i.e. each step down represents the occurrence of the event of interest for at least one observation. Censoring times only affect the estimate by reducing the size of the risk set for the next event, and thereby increasing the height of the next step down. To visualise this function, KM survival curves (plots of the KM estimator over time) are used. On a survival curve the y-axis represents the probability that the subject has not yet experienced the event of interest after surviving up to time t and the x-axis represents time t.

As an example, Figure 3.1 provides the KM survival curves for DSS time, i.e. time from breast cancer diagnosis to the date of death from the disease, estimated from application to the HER- and HER2+ stratified patients observed in the METABRIC cohort. Figure 3.1 illustrates the role of HER2 status as a biomarker: patients with HER2- breast cancer display more favourable DSS than patients with HER2+ breast cancer. The median survival time is 286 months for the HER2- group compared to 125 months for the HER2+ group.

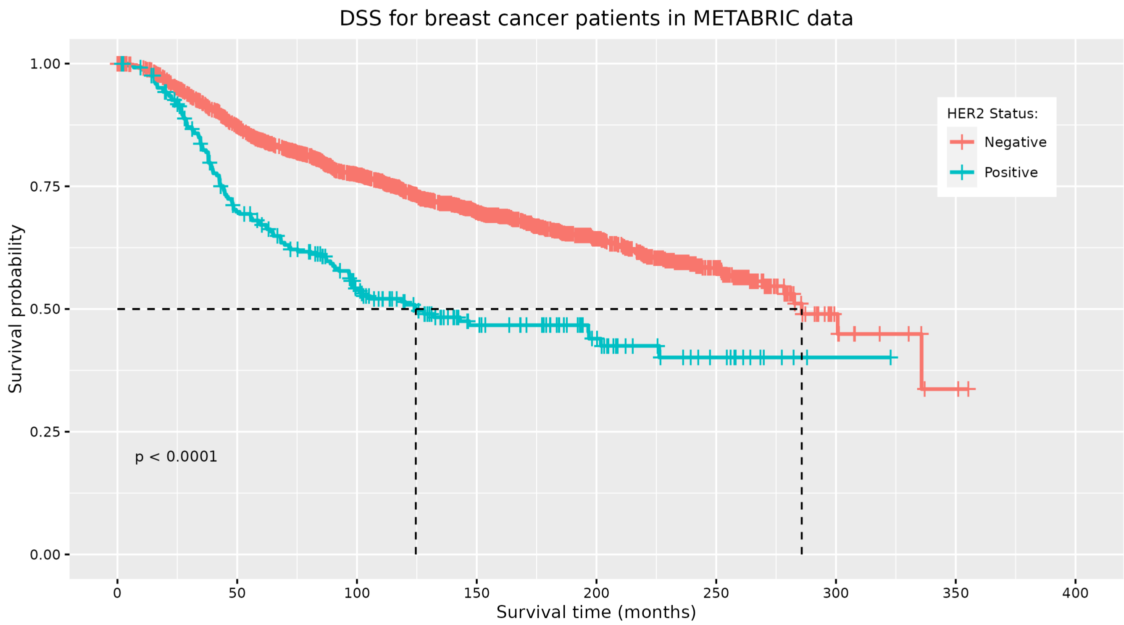


Figure 3.1: Kaplan-Meier plot for disease-specific survival in METABRIC patients stratified by HER2 status. Lines corresponding to median disease-specific survival time for each group and the p-value associated with the log-rank test are displayed.

To determine whether two or more survival curves are significantly different from each other, a large-sample χ^2 test (the log-rank test) tests the null hypothesis of no difference between the populations in the probability of an event at any time point. In Figure 3.1, the $p < 0.0001$ indicates a significant difference in DSS comparing HER2- and HER2+ patients.

3.1.2 Recursive Partitioning Survival Trees

Recursive partitioning techniques or tree-based methods were first developed by Morgan and Sonquist (1963) but were popularised in the 1980s following the development of the Classification and Regression Tree (CART) by Breiman et al. (Breiman et al., 1984; Bou-Hamad et al., 2011). More recently, conditional inference trees (CTREE) have been developed to resolve some of the limitations of CART including overfitting and selection bias towards variables with many possible splits, i.e. continuous variables (Hothorn et al., 2006). These non-parametric techniques are useful for identifying important predictors and structure in a dataset.

These techniques recursively partition the data to form groups, called nodes, containing individuals with homogenous response values. The predicted response value for each node is then generally either the mean or mode dependent on whether the partitioning variable is continuous or categorical. Within the CART methodology, splits within the tree are arrived at by minimising a measure of node impurity. For categorical response variables measures of impurity may be the Gini index or information index. The Gini index denotes the probability that a random observation is misclassified when chosen randomly, while the information index relates to how much information is gained by splitting a set of data on a particular feature. For continuous response variables this measure may be the sum of squared deviations from the mean. The CTREE methodology differs in that the splits within the tree are arrived at using p-values from permutation-based significance tests.

Using the CART or CTREE methodologies to produce predictive models has a number of advantages which are detailed below:

1. No distributional assumptions are made.
2. The predictor variables used can be continuous, interval or categorical.
3. Robust to outliers, collinearities and heteroscedasticity.
4. Can detect interactions and structure in a highly complex dataset.
5. Transformations of the data do not change the structure of the tree.
6. Can use the same variable multiple times, i.e. at different branches in the tree.

To apply CART and CTREE methodologies available R packages include rpart (Therneau et al., 2022) and partykit (Hothorn et al., 2006; Hothorn and Zeileis, 2015).

The rpart procedure (Therneau et al., 2022) implements many of the ideas found in CART and builds classification models (predicts a continuous value based on the predictor variables) or regression models (predicts discrete labels or categories based on the predictor variables) which can be represented as binary trees. The rpart algorithm is an iterative algorithm where the tree is built by first identifying a single variable which best splits the data into two groups. The same process is then applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size or until no improvement can be made. To split the data, rpart uses one of several measures of impurity such as the Gini index of a node and then chooses the split with maximal impurity reduction.

The CTREE procedure, applied in partykit (Hothorn et al., 2006; Hothorn and Zeileis, 2015), carries out variable selection and splitting in two steps, mitigating the tendency towards predictor variables with many possible splits or many missing values. Rather than employing information measures for predictor selection, CTREE applies a significance test procedure. The conditional distribution of statistics measuring the association between responses and predictor variables is responsible for the unbiased selection of variables that are measured on different scales. In addition, this algorithm applies multiple testing procedures to determine if there is no significant association between any of the predictors and the response and as such decides when the recursion should be halted.

3.2 CNA Metrics Stratify Luminal Breast Cancer Patients to Explain Survival Outcome

Approximately 70% of breast cancers are classified as Luminal A or B, characterised by increased levels of ER and PR (Tishchenko et al., 2016). Luminal B tumours tend to grow faster than Luminal A tumours, are of a higher grade, have a slightly worse prognosis and usually require more aggressive treatments. It has been suggested that the relationship between Luminal A and Luminal B tumours may be a continuum rather than a strict division of subtypes (Wirapati et al., 2008; Curtis et al., 2012; Tishchenko et al., 2016). It has also been hypothesised that Luminal A tumours may evolve into Luminal B tumours as a result of stochastic acquisitions of mutations in genes associated with worse prognosis, including HER2 and tumour protein p53 (TP53) (Ulrich, 2013). This ambiguity in Luminal classification may account for the variation that exists in DSS outcome for some Luminal A patients (Tishchenko et al., 2016; Sung et al., 2016; Kumar et al., 2019; Wang and Lee, 2023).

With focus and application to the Luminal METABRIC patients (n = 1,175, data downloaded from cBioPortal in 2019) we aim to explore whether the metrics of GI, specifically Absolute CNA Score, can add value in modelling OS and DSS within this group.

3.2.1 Preliminary Survival Analysis using Absolute CNA Score and Quartiles

To explore estimated survival curves applying KM, the continuous Absolute CNA Score (Equation 2.17) was categorised into 4 levels: Q1, Q2, Q3, Q4. Each patient is recorded as one of these levels based on whether their Absolute CNA Score was in the first quartile (lowest GI) to fourth quartile (highest GI) relative to the observed Luminal METABRIC cohort (Figure 3.2).

KM fitted to patients stratified by the four levels of Absolute CNA Score Quartile indicate significant differences between the four estimated OS curves (log-rank test $p < 0.0001$, Figure 3.3). The Absolute CNA Score Quartiles are associated with OS in Luminal breast cancer patients. Those in Absolute CNA Score Q4 (highest GI) have worse survival outcomes than patients with less GI in Absolute CNA Score Quartiles 1-3 (Q1-3). For DSS outcomes, KM fitted to patients stratified by the four levels of Absolute CNA Score Quartile indicates significant differences between the four estimated DSS curves (log-rank $p < 0.0001$, Figure 3.4). Luminal breast cancer patients in Q4 have worse DSS outcomes than patients in Q1-3.

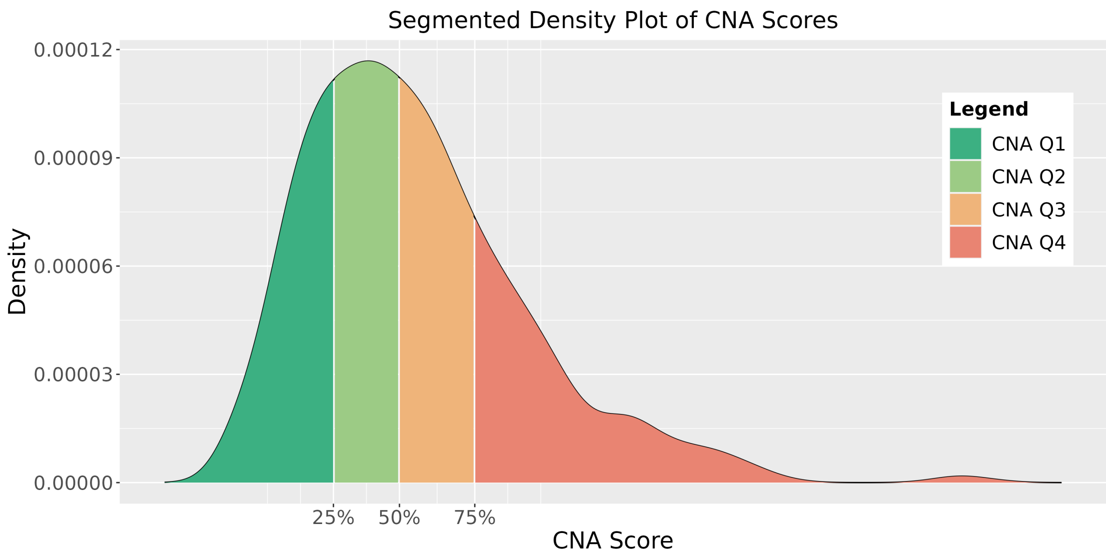


Figure 3.2: Density plot of Absolute CNA Score distribution for METABRIC Luminal cases. Absolute CNA Score Quartiles 1-4 indicated by legend colours.

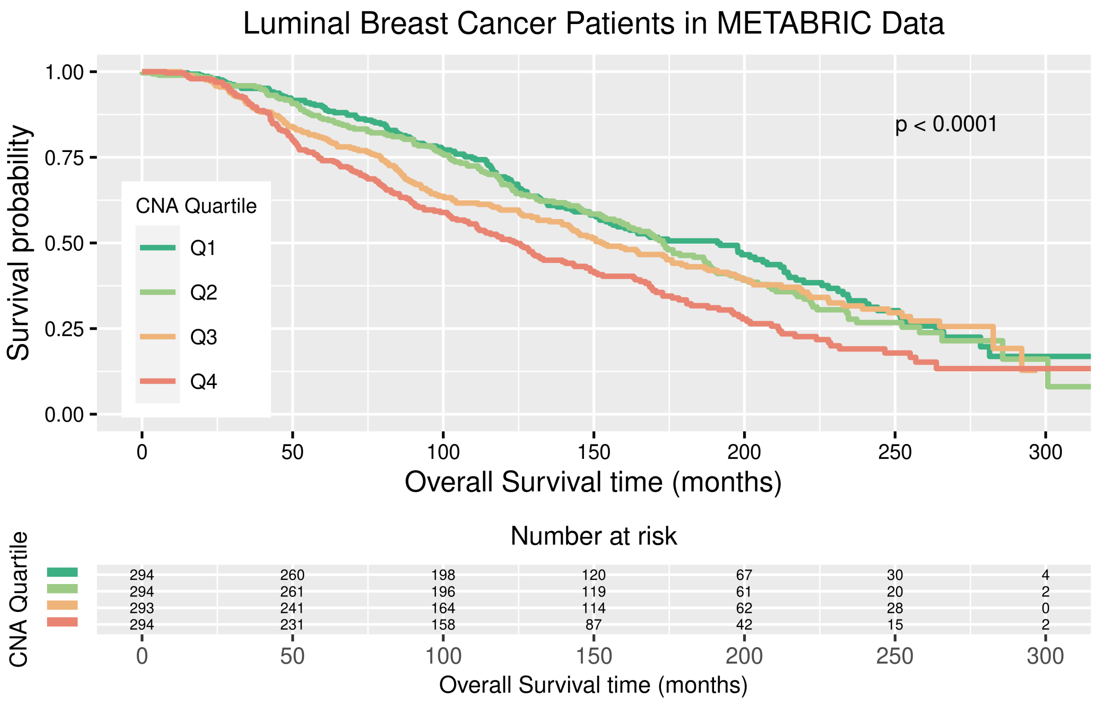


Figure 3.3: Kaplan-Meier plots for overall survival for METABRIC Luminal breast cancer patients in each Absolute CNA Score Quartile. The p-value associated with the log-rank test and a risk table displaying the number of patients at risk at each time interval is displayed.

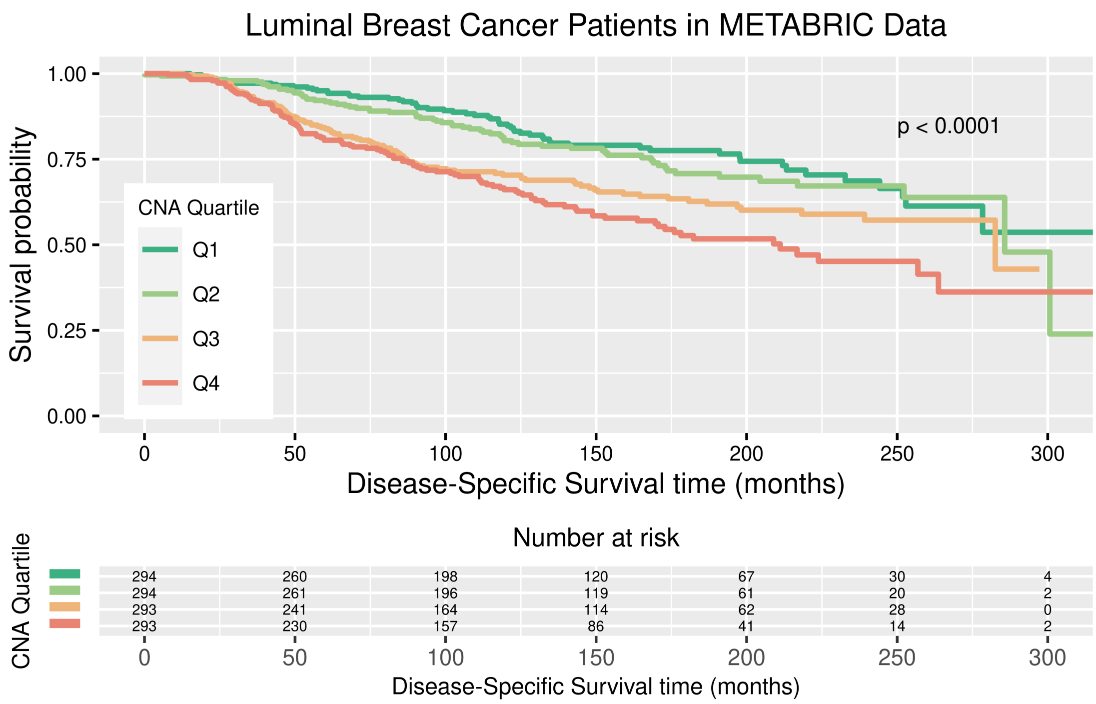


Figure 3.4: Kaplan-Meier plots for disease-specific survival for METABRIC Luminal breast cancer patients in each Absolute CNA Score Quartile. The p-value associated with the log-rank test and a risk table displaying the number of patients at risk at each time interval is displayed.

To maintain the information from the uncategorised, quantitative measure of Absolute CNA Score, univariate Cox models were fitted for OS and DSS. The results obtained indicate that the Absolute CNA Score is associated with both OS (Hazard Ratio (HR) = 1.00 [1.00-1.00], $p < 0.001$) and DSS (HR = 1.00 [1.00-1.00], $p < 0.0001$).

3.2.2 Analysis of Potential Confounding Variables and Multivariable Cox Models

Having observed that Absolute CNA Score can independently help stratify Luminal patients into groups of similar survival outcome, it is important to assess whether Absolute CNA Score can add value in combination with other biomarkers. KM plots and univariate Cox models were used to determine if any of the 23 available clinical variables (list in Appendix A) were associated with survival outcome. It was found that 19 of the clinical variables considered were associated with OS and 18 were associated with DSS (Tables 3.1 and 3.2). The clinical variables found to be significant within the univariate analysis were examined for possible associations with the Absolute CNA Scores and Absolute CNA Score Quartiles, applying tests such as the χ^2 test, Fisher's exact test, Kruskal-Wallis test and Pearson's correlation, as appropriate to the variable type. These tests indicated that the Absolute CNA Scores and CNA Score Quartiles were significantly associated with a number of clinical variables (Tables 3.3 and 3.4).

Since highly correlated predictors may lead to unreliable and unstable estimates of regression coefficients (Keith, 2019), a refined selection of variables were considered based on understanding of the clinical definition of the variable, e.g. HER2 Status and HER2 SNP6 use different methods to capture similar information (discussed further in King et al. (2021a)). Eight candidate clinical predictors remained: PAM50 subtype, histological grade, tumour size, number of lymph nodes positive, age, HER2 status, PR status and histological subtype. These eight candidate clinical predictors were considered along with either the Absolute CNA Score or Absolute CNA Score Quartile variable in multivariable CPH models for OSS and DSS. Under the assumption of proportional hazards, the results indicated that the Absolute CNA Score metric was significantly associated with the outcome in a model for DSS along with six clinical predictors: PAM50 subtype, histological grade, tumour size, number of positive lymph nodes, age at diagnosis, and HER2 status, both using the categorical CNA Score Quartiles (Table 3.5) and the original continuous Absolute CNA Scores (Table 3.6).

Table 3.1: OS Univariate Cox models for each clinical variable. Likelihood ratio test (LRT) and Wald test p-values and Benjamini-Hochberg adjusted p-values are displayed.

Univariate Cox models for each clinical variable for OS, within the Luminal METABRIC cohort				
Clinical Variable	**LRT**	**Wald Test**	**Adjusted LRT**	**Adjusted Wald Test**
ER Status	0.82	0.82	0.82	0.82
Cellularity	0.45	0.46	0.47	0.48
Laterality	0.44	0.44	0.47	0.48
Cancer Type Detailed	0.03	0.15	0.04	0.18
Chemotherapy	0.06	0.05	0.06	0.06
Histological Subtype	<0.0001	0.04	<0.0001	0.05
ER Immunohistochemistry	0.04	0.03	0.05	0.04
HER2 SNP6	0.03	0.02	0.04	0.03
Radiotherapy	0.02	0.02	0.03	0.03
HER2 Status	<0.0001	<0.0001	<0.0001	<0.0001
Histological Grade	<0.0001	<0.0001	<0.0001	<0.0001
PR Status	<0.0001	<0.0001	<0.0001	<0.0001
Hormone Therapy	<0.0001	<0.0001	<0.0001	<0.0001
Three Gene Classification	<0.0001	<0.0001	<0.0001	<0.0001
Integrative Cluster	<0.0001	<0.0001	<0.0001	<0.0001
PAM50	<0.0001	<0.0001	<0.0001	<0.0001
Breast Surgery	<0.0001	<0.0001	<0.0001	<0.0001
Inferred Menopausal State	<0.0001	<0.0001	<0.0001	<0.0001
Clinical Stage	<0.0001	<0.0001	<0.0001	<0.0001
NPI	<0.0001	<0.0001	<0.0001	<0.0001
Positive Lymph Nodes	<0.0001	<0.0001	<0.0001	<0.0001
Tumour Size	<0.0001	<0.0001	<0.0001	<0.0001
Age at Diagnosis	<0.0001	<0.0001	<0.0001	<0.0001

Table 3.2: DSS Univariate Cox models for each clinical variable. Likelihood ratio test (LRT) and Wald test p-values and Benjamini-Hochberg adjusted p-values are displayed.

Univariate Cox models for each clinical variable for DSS, within the Luminal METABRIC cohort

Clinical Variable	LRT	Wald Test	Adjusted LRT	Adjusted Wald Test
Radiotherapy	0.94	0.94	0.97	0.97
ER Status	0.97	0.97	0.97	0.97
Histological Subtype	<0.0001	0.69	0.01	0.76
Cancer Type Detailed	0.31	0.53	0.35	0.61
Laterality	0.4	0.4	0.43	0.48
Cellularity	0.11	0.13	0.13	0.17
Inferred Menopausal State	0.01	0.02	0.02	0.02
HER2 SNP6	0.02	0.01	0.02	0.02
ER Immunohistochemistry	<0.0001	<0.0001	0.01	<0.0001
Hormone Therapy	<0.0001	<0.0001	<0.0001	<0.0001
Breast Surgery	<0.0001	<0.0001	<0.0001	<0.0001
HER2 Status	<0.0001	<0.0001	<0.0001	<0.0001
Age at Diagnosis	<0.0001	<0.0001	<0.0001	<0.0001
PR Status	<0.0001	<0.0001	<0.0001	<0.0001
Chemotherapy	<0.0001	<0.0001	<0.0001	<0.0001
Histological Grade	<0.0001	<0.0001	<0.0001	<0.0001
Three Gene Classification	<0.0001	<0.0001	<0.0001	<0.0001
Integrative Cluster	<0.0001	<0.0001	<0.0001	<0.0001
PAM50	<0.0001	<0.0001	<0.0001	<0.0001
Clinical Stage	<0.0001	<0.0001	<0.0001	<0.0001
Positive Lymph Nodes	<0.0001	<0.0001	<0.0001	<0.0001
Tumour Size	<0.0001	<0.0001	<0.0001	<0.0001
NPI	<0.0001	<0.0001	<0.0001	<0.0001

As the focus and application here is only within Luminal cancers, PAM50 subtype has only two levels Luminal A and Luminal B. In the fitted models an indicator variable assumes the reference group to be Luminal A, and the estimated effect using this indicator is for Luminal B relative to Luminal A. In the model using Absolute CNA Scores the reference group is Luminal A, histological grade 1, HER2-negative patients (Table 3.6). For Luminal A patients, Absolute CNA Score is associated with DSS (HR = 1.00 [1.00-1.00], $p < 0.001$). For Luminal B patients, the effect of Absolute CNA Score on DSS is estimated by fitting interaction effects between Absolute CNA Score and PAM50 subtype. This indicates that the association between Absolute CNA Score is significantly different for Luminal B patients compared to Luminal A patients (HR = 1.00 [0.99-1.00], $p < 0.012$). Setting Luminal B as the reference group indicates that for Luminal B patients, Absolute CNA Score is not associated with DSS.

In the model using Absolute CNA Score Quartiles the reference group is Luminal A, histological grade 1, HER2-negative patients with Absolute CNA Scores with lowest GI level, Absolute CNA Score Q1 (Table 3.5). Comparing Absolute CNA Score Q4 to CNA Q1, within Luminal A patients, shows a significant increased risk in DSS (HR = 2.32 [1.36-3.94], p = 0.002). Comparing Absolute CNA Score Q3 to Absolute CNA Score Q1, within Luminal A patients, shows a significant increased risk in DSS (HR = 2.15 [1.33-3.49], p = 0.002). There was no evidence of a significant effect on risk comparing Absolute CNA Score Q2 Luminal A patients to Absolute CNA Score Q1 Luminal A patients (HR = 1.37 [0.83 - 2.27], p = 0.219). For Luminal B patients, the effect of Absolute CNA Score Quartile on DSS differs in comparison to Luminal A patients estimated by fitting interaction effects between Absolute CNA Score Quartiles and PAM50 subtype, the effect is a reduction in the estimated difference comparing Absolute CNA Score Quartile within Luminal B.

Table 3.3: Association tests between CNA Score and selected clinical variables. Kruskal-Wallis and Pearson's Correlation p-values and Benjamini-Hochberg adjusted p-values are displayed.

Association between CNA Score metric and Clinical variables, within the Luminal METABRIC cohort			
Clinical Variable	**Kruskal-Wallis Test**	**Pearson's Correlation**	**Adjusted P-value**
Age at Diagnosis	NA	0.57	0.57
Breast Surgery	0.33	NA	0.35
Inferred Menopausal State	0.29	NA	0.32
ER Immunohistochemistry	0.13	NA	0.15
Positive Lymph Nodes	NA	0.01	0.01
Radiotherapy	0.01	NA	0.01
Chemotherapy	<0.0001	NA	<0.0001
Tumour Size	NA	<0.0001	<0.0001
Cancer Type Detailed	<0.0001	NA	<0.0001
Clinical Stage	<0.0001	NA	<0.0001
Hormone Therapy	<0.0001	NA	<0.0001
HER2 Status	<0.0001	NA	<0.0001
Histological Subtype	<0.0001	NA	<0.0001
PR Status	<0.0001	NA	<0.0001
HER2 SNP6	<0.0001	NA	<0.0001
NPI	NA	<0.0001	<0.0001
PAM50	<0.0001	NA	<0.0001
Histological Grade	<0.0001	NA	<0.0001
Three Gene Classification	<0.0001	NA	<0.0001
Integrative Cluster	<0.0001	NA	<0.0001

Plotting the adjusted survival curves for Luminal A and Luminal B patients within the different Absolute CNA Score Quartiles illustrates how these estimated effects differ between the two subtypes. Adjusted survival curves represent the estimated effect of Absolute CNA Score Quartiles by plotting the predicted survival curves for Luminal A and Luminal B patients for each Absolute CNA Score Quartile, having adjusted for the effects of the other covariates in the multivariable Cox model, where other covariates are fixed at the median/mode values of those variables (Figure 3.5). Here we see that DSS curves comparing Absolute CNA Score Quartiles within Luminal A show significant differences while the differences observed in DSS curves comparing Absolute CNA Score Quartiles within Luminal B are small and non-significant.

Table 3.4: Association tests between CNA Quartiles and selected clinical variables. Chi-Squared, Fisher's Exact and Kruskal-Wallis test p-values and Benjamini-Hochberg adjusted p-values are displayed.

Association between categorised CNA Score metric quartiles and Clinical variables, within the Luminal METABRIC cohort

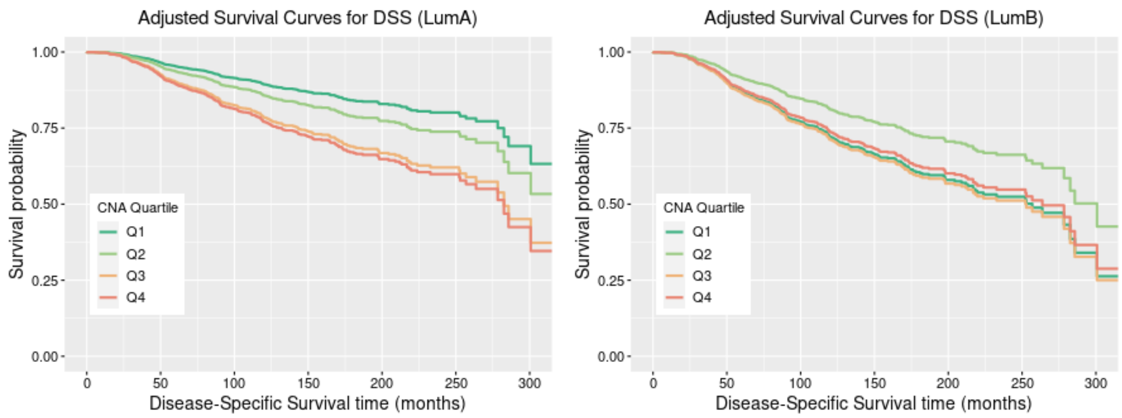
Clinical Variable	Chi-Squared Test	Fisher's Exact Test	Kruskal-Wallis Test	Adjusted P-value
Breast Surgery	0.56	0.57	NA	0.56
ER Immunohistochemistry	0.51	0.48	NA	0.53
Inferred Menopausal State	0.07	0.06	NA	0.07
Age at Diagnosis	NA	NA	0.01	0.02
Radiotherapy	0.01	0.01	NA	0.01
Chemotherapy	<0.0001	<0.0001	NA	0.01
Clinical Stage	<0.0001	<0.0001	NA	<0.0001
Positive Lymph Nodes	NA	NA	<0.0001	<0.0001
Cancer Type Detailed	<0.0001	<0.0001	NA	<0.0001
Hormone Therapy	<0.0001	<0.0001	NA	<0.0001
HER2 Status	<0.0001	<0.0001	NA	<0.0001
Histological Subtype	<0.0001	<0.0001	NA	<0.0001
PR Status	<0.0001	<0.0001	NA	<0.0001
HER2 SNP6	<0.0001	<0.0001	NA	<0.0001
Tumour Size	NA	NA	<0.0001	<0.0001
NPI	NA	NA	<0.0001	<0.0001
PAM50	<0.0001	<0.0001	NA	<0.0001
Histological Grade	<0.0001	<0.0001	NA	<0.0001
Three Gene Classification	<0.0001	<0.0001	NA	<0.0001
Integrative Cluster	<0.0001	<0.0001	NA	<0.0001

Table 3.5: Final multivariable CPH model for DSS. Including selected clinical variables, Absolute CNA Score Quartiles and an interaction term, within the Luminal METABRIC cohort.

Clinical Variable	Beta	SE	HR	95% CI	P-value
PAM50:					
Luminal A (Ref)	-	-	-	-	-
Luminal B	1.069	0.299	2.912	(1.619 - 5.237)	<0.001
Histological Grade:					
1 (Ref)	-	-	-	-	-
2	0.381	0.254	1.464	(0.889 - 2.410)	0.134
3	0.528	0.262	1.696	(1.014 - 2.837)	0.044
Tumour Size	0.015	0.003	1.015	(1.010 - 1.020)	<0.001
Positive Lymph Nodes	0.050	0.008	1.051	(1.034 - 1.069)	<0.001
Age at Diagnosis	0.018	0.005	1.018	(1.008 - 1.029)	<0.001
HER2 Status:					
Negative (Ref)	-	-	-	-	-
Positive	0.541	0.202	1.717	(1.157 - 2.550)	0.007
CNA Quartile:					
CNA Q1 (Ref)	-	-	-	-	-
CNA Q2	0.315	0.256	1.370	(0.829 - 2.265)	0.219
CNA Q3	0.767	0.247	2.152	(1.326 - 3.493)	0.002
CNA Q4	0.839	0.272	2.315	(1.360 - 3.942)	0.002
CNA Q2:LumB	-0.764	0.395	0.466	(0.215 - 1.010)	0.053
CNA Q3:LumB	-0.730	0.364	0.482	(0.236 - 0.983)	0.045
CNA Q4:LumB	-0.909	0.370	0.403	(0.195 - 0.831)	0.014
Likelihood Ratio Test p-value					<2e-16
Wald Test p-value					<2e-16
Score (log-rank) Test p-value					<2e-16
SE: Standard Error; HR: Hazard Ratio; CI: Confidence Interval					

Table 3.6: Final multivariable CPH model for DSS. Including selected clinical variables, Absolute CNA Score and an interaction term, within the Luminal METABRIC cohort.

Clinical Variable	Beta	SE	HR	95% CI	P-value
PAM50:					
Luminal A (Ref)	-	-	-	-	-
Luminal B	0.896	0.225	2.450	(1.576 - 3.808)	<0.001
Histological Grade:					
1 (Ref)	-	-	-	-	-
2	0.431	0.253	1.539	(0.937 - 2.525)	0.088
3	0.636	0.260	1.888	(1.135 - 3.141)	0.014
Tumour Size	0.013	0.002	1.013	(1.009 - 1.018)	<0.001
Positive Lymph Nodes	0.049	0.009	1.050	(1.033 - 1.068)	<0.001
Age at Diagnosis	0.016	0.005	1.016	(1.006 - 1.027)	0.002
HER2 Status:					
Negative (Ref)	-	-	-	-	-
Positive	0.568	0.201	1.765	(1.191 - 2.615)	0.005
CNA Score	6.05e-05	1.83e-05	1.000	(1.000 - 1.000)	<0.001
CNA Score:LumB	-6.77e-05	2.69e-05	0.999	(0.999 - 1.000)	0.012
Likelihood Ratio Test p-value					<2e-16
Wald Test p-value					<2e-16
Score (log-rank) Test p-value					<2e-16
SE: Standard Error; HR: Hazard Ratio; CI: Confidence Interval					



(A) Luminal A Patients

(B) Luminal B Patients

Figure 3.5: Adjusted survival curves for estimated Absolute CNA Score Quartile effects within each Luminal PAM50 subtype. (A) Luminal A and (B) Luminal B breast cancer patients.

3.2.3 Implementation of Recursive Partitioning Survival Trees

Although the fitted models give strong indication that Absolute CNA Score, both as a continuous and categorical metric, can add value in modelling survival outcome in Luminal breast cancer, diagnostic tests indicate that the proportional hazards assumption may not be met (Supplementary Information). As an alternative approach, recursive partitioning survival trees are fitted using the rpart and ctree algorithms. Recursive partitioning trees can explore the association between Absolute CNA Score and survival and examine any interactions between the six clinical variables that were significant in models including Absolute CNA Score variables. The predetermined Absolute CNA Score Quartiles can be fitted in the model as a predictor, but recursive partitioning trees also offer the added benefit of determining the optimum cut-off in Absolute CNA Score, implicit in the partitioning algorithm.

Survival trees considering Absolute CNA Score Quartiles, with additional applications provided in Appendix A, suggest a similar partitioning with Absolute CNA Score Q1 and Q2 versus Absolute CNA Score Q3 and Q4, consistent with the effects estimated by the CPH model (Figure 3.6). Figure 3.6 shows the survival tree fitted using the ctree algorithm, which indicates that for Luminal A patients who have 0-1 positive lymph nodes, tumour size less than 31mm and age of diagnosis less than 71.4 years, DSS outcome can be stratified by Absolute CNA Score Quartile, where patients with high GI show reduced survival probability than those with a lower GI ($p = 0.032$).

Survival trees considering Absolute CNA Score, with additional applications provided in Appendix A, also suggest a similar partitioning consistent with the effects estimated by the CPH model and the survival trees considering Absolute CNA Score Quartiles (Figure 3.7). In Figure 3.7, the survival tree fitted using the ctree algorithm, Luminal A patients who have 0-1 positive lymph nodes, tumour size less than 31mm and age of diagnosis less than 71.4 years, DSS outcome can be stratified by Absolute CNA Score with optimised Absolute CNA Score cut-off point value 5,882 ($p = 0.005$). Including the continuous Absolute CNA Score, rather than the categorised Absolute CNA Score Quartile, allows a more nuanced investigation of the optimal Absolute CNA Score cut-off point. While the estimated optimal cut-off of 5,882 is close to 5,547, the boundary between Absolute CNA Score Quartile 2 and 3, utilising the Absolute CNA Score cut-off results in 20 patients being reclassified to the low-risk group.

Overall, the survival trees indicate that the Absolute CNA Score metric, implemented either as predetermined categorised quartiles or original continuous variable, can stratify subsets of patients based on DSS and therefore help identify Luminal A patients who are at elevated risk.

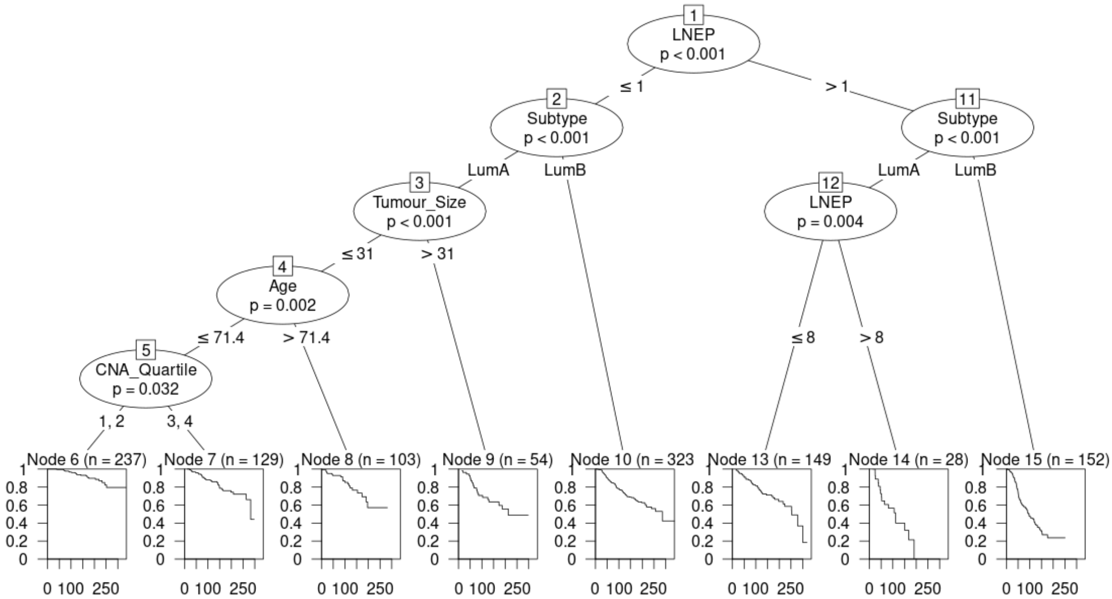


Figure 3.6: Recursive partitioning survival tree for disease-specific survival using clinical variables and Absolute CNA Score Quartile as candidate predictors. Fitted using the ctree algorithm.

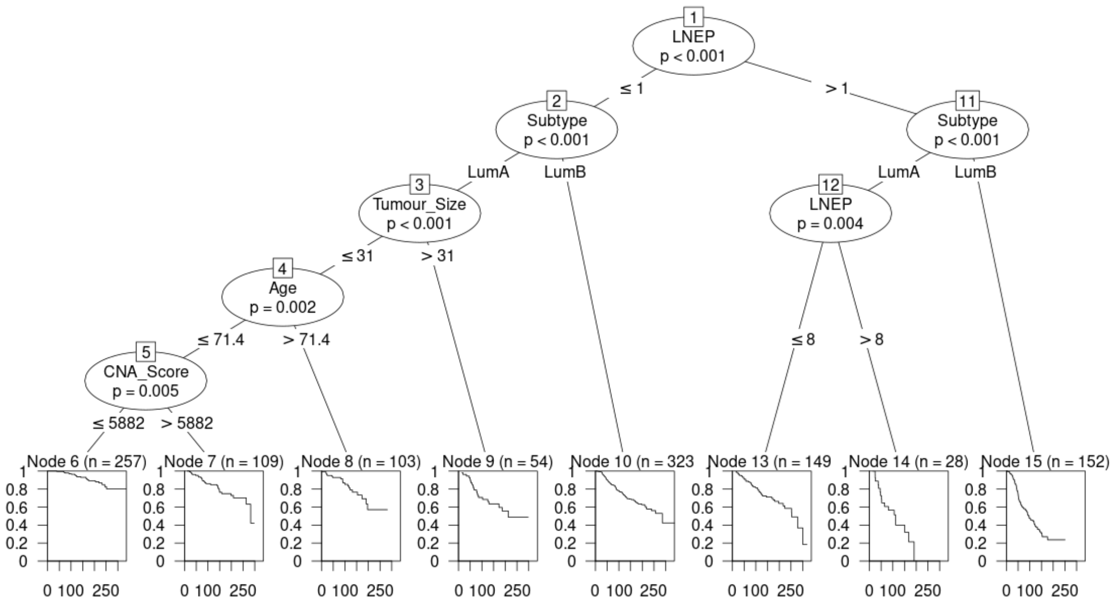


Figure 3.7: Recursive partitioning survival tree for disease-specific survival using clinical variables and Absolute CNA Score as candidate predictors. Fitted using the ctree algorithm.

3.3 Analysis of Global CNA Metrics across All METABRIC Patients

Expanding the study focus to all patients, i.e. all PAM50 subtypes (Basal, HER2, Luminal B, Luminal A, Normal and Claudin-low), associations between the six global CNA Score metrics and six global CNA Burden metrics with survival are examined in this section. The global CNA metrics are initially included with PAM50 subtype or IntClust molecular classifications to assess whether the CNA metric information can add additional prognostic value to the molecular classifications, and then included with a selection of clinical variables to explore interactions between the clinical variables and CNA metrics. Given the large number of candidate predictors under consideration, the fact that the CPH assumption may not be met, and the benefit of the partitioning trees in determining optimal cut-off, recursive partitioning survival trees are implemented.

3.3.1 CNA Metric Survival Trees, in Combination with Molecular Classification Predictors

A range of survival trees for OS, DSS, 5-year OS and DSS, and 10-year OS and DSS are produced. These trees include the six global CNA Score metrics: Absolute CNA Score, CNA Amp Score, CNA Del Score, Difference Score, Percentage Amp Score and Percentage Del Score, or the six global CNA Burden metrics: CNA Burden, CNA Amp Burden, CNA Del Burden, Difference Burden, Percentage Amp Burden and Percentage Del Burden, with PAM50 subtype or IntClust molecular classifications. Depending on the algorithm used (rpart or ctree) a number of different global CNA Score and Burden metrics are selected as useful predictors when modelling survival outcomes. The survival trees for DSS outcomes are displayed and discussed below, while the survival trees for OS outcomes are provided in Appendix B.

Initially survival trees including only PAM50 subtype (Figure 3.8) or IntClust (Figure 3.9) as candidate predictors are fitted, indicating which subtypes display similarity in survival outcome and providing information on partitions in trees where only molecular classification is included.

(A) **(B)**

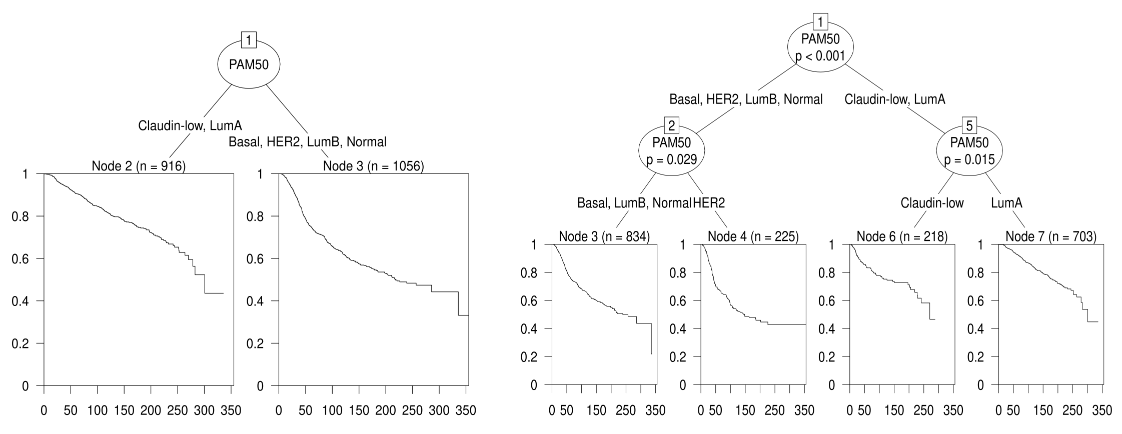


Figure 3.8: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype as a candidate predictor. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A) **(B)**

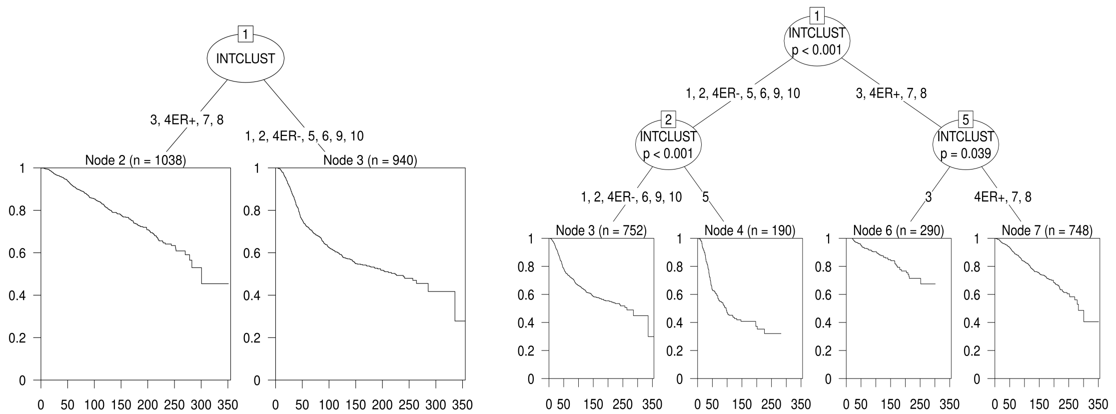


Figure 3.9: Recursive partitioning survival trees for disease-specific survival using Integrative Cluster as a candidate predictor. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

Focusing on survival trees for DSS, 5-year DSS and 10-year DSS, that have the six CNA Score metrics and PAM50 molecular classification as candidate predictors, indicates CNA Del Score to be a consistently significant factor (Figure 3.10, 3.11 and 3.12). While variation in the predictors used to partition the data is observed across survival outcome (DSS, 5-year DSS or 10-year DSS) and algorithm (rpart or ctree), CNA Del Score appears to add additional prognostic value to the PAM50 subtypes, primarily Luminal A and Claudin-low subtypes. Patients within these subtypes, displaying higher levels of CNA Del Score, have poorer outcome with respect to DSS, 5-year and 10-year DSS. Consistency is observed in the CNA Del Score cut-off points chosen by the ctree algorithm across the three survival outcomes. The optimal cut-off point for CNA Del Score for trees produced using the ctree algorithm is 3,286 in Luminal A and Claudin-low patients (Figure 3.10B), 3,286 in Luminal A patients (Figure 3.11B), and 3,138 in Luminal A, Claudin-low and Normal patients (Figure 3.12B), for DSS, 5-year DSS and 10-year DSS.

When considering the survival trees for DSS, 5-year DSS and 10-year DSS, generated using the six CNA Burden metrics and PAM50 molecular classification as candidate predictors, similar tree structures are observed (Figures 3.13, 3.14, and 3.15). CNA Del Burden is again consistently selected as an important predictor in the context of DSS, partitioning Luminal A and Claudin-low subtypes using a cut-off of 18.28% (Figure 3.13B), 5-year DSS, partitioning Luminal A subtype using a cutoff of 14.55% (Figure 3.14B), and 10-year DSS, partitioning Luminal A, Claudin-low and Normal subtypes using a cutoff of 14.02% (Figure 3.15B).

In the survival trees for DSS, 5-year DSS and 10-year DSS, generated using the six CNA Score metrics and IntClust molecular classification as candidate predictors, CNA Del Score consistently appears as an important predictor for survival outcome, in patients corresponding to IntClust 3, 4ER+, 7 and 8 (Figures 3.16-3.18). Again, the optimal cut-off points are fairly consistent, in all but one tree, at values 1,469.5 and 1,469 for DSS (Figure 3.16), 1,933 and 3,722 for 5-year DSS (Figure 3.17), and 1,469.5 and 1,989 for 10-year DSS (Figure 3.18). Patients within IntClust 3, 4ER+, 7 and 8, displaying levels of CNA Del Score above the optimal cut-off point have worse DSS, 5 and 10-year DSS, than patients displaying levels of CNA Del Score below the optimal cut-off point.

Considering survival trees for DSS, 5-year DSS and 10-year DSS, with the six CNA Burden metrics and IntClust molecular classification as candidate predictors, indicate similar tree structures (Figures 3.19-3.21). Again, all trees initially split on IntClust, with CNA Del Burden consistently selected as an additional significant predictor in the context of the DSS, 5-year DSS and 10-year DSS.

It appears that the CNA Del Score and Burden metrics are associated with DSS, 5-year DSS and 10-year DSS. The majority of the survival trees showed that the CNA Del metrics are useful in splitting Luminal A and Claudin-low patients, and IntClust 3, 4ER+, 7 and 8 patients, into groups with distinct survival outcomes. Interestingly, a known feature of these PAM50 subtypes and IntClusts is that they display low genomic instability and good prognosis (Curtis et al., 2012), Section 2.5. This may explain why the optimal cut-off points for the CNA Del Score and Burden are quite low. In Chapter 2, we observed that PAM50 subtypes associated with poorer survival (Basal, HER2 and Luminal B) have significantly higher levels of deletions. Here, we observe that within PAM50/IntClust classifications associated with good prognosis, that have CNA Del Score and Burden over an optimised threshold, patients having poorer survival outcomes, again indicating that high levels of deletions are more detrimental than other forms of alterations.

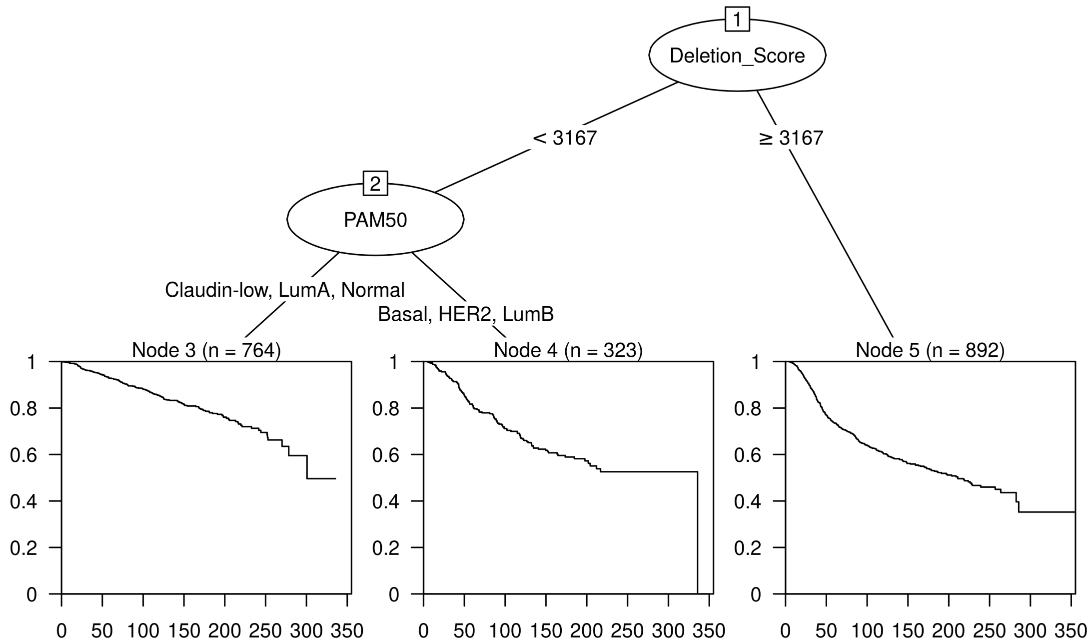
3.3.2 CNA Metric Survival Trees, in Combination with Molecular Classification and Clinical Predictors

To assess how the addition of clinical variables alters the observed partitioning and explore interactions between the clinical variables and CNA metrics in modelling DSS, 5-year DSS and 10-year DSS, survival trees including the six CNA Burden metrics, IntClust or PAM50 molecular classification, and selected clinical variables as candidate predictors are fitted (Figures 3.22-3.27). To avoid overcrowding, these trees are limited to a depth of four, where depth is defined as the number of layers or levels in the tree. While the CNA Score and CNA Burden trees partition the data similarly, it is observed that there is more consensus among the CNA Burden trees produced across different survival times and using different algorithms. Based on this and the fact that CNA Burden is a standardised metric, i.e. all patients have CNA Burden in the range 0 to 100, we show only the survival trees including CNA Burden, IntClust or PAM50 molecular classification, and selected clinical variables, as candidate predictors. The clinical variables selected are number of positive lymph nodes, NPI, ER Status, PR Status, HER2 Status, age, tumour size, tumour stage, tumour grade and cancer type.

Figures 3.22-3.24 display survival trees for DSS, 5-year DSS and 10-year DSS, that have the six CNA Burden metrics, PAM50 molecular classification and the selected clinical variables as candidate predictors. It is observed that total CNA Burden, CNA Del Burden and Percentage Amp Burden, appear as significant predictors in the context of the DSS, in addition to PAM50 subtype and a number of clinical variables including NPI, number of positive lymph nodes, age and ER status.

For DSS, CNA Burden provides additional information for patients who have NPI < 5.05, and for patients who have NPI < 5.05 and ≤ 1 positive lymph node, for the rpart and ctree algorithms, respectively (Figure 3.22). The CNA Burden threshold for both partitions are 24.91% and 24.90%. For the 5-year DSS survival trees, fitted with the rpart algorithm, CNA Burden with threshold 24.56% is used

(A)



(B)

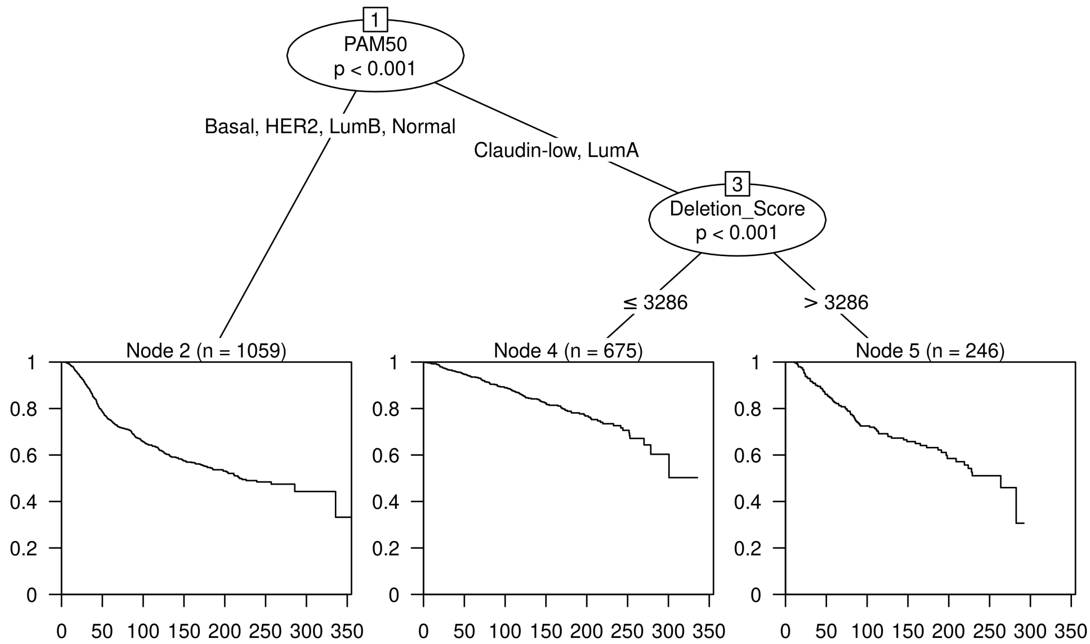
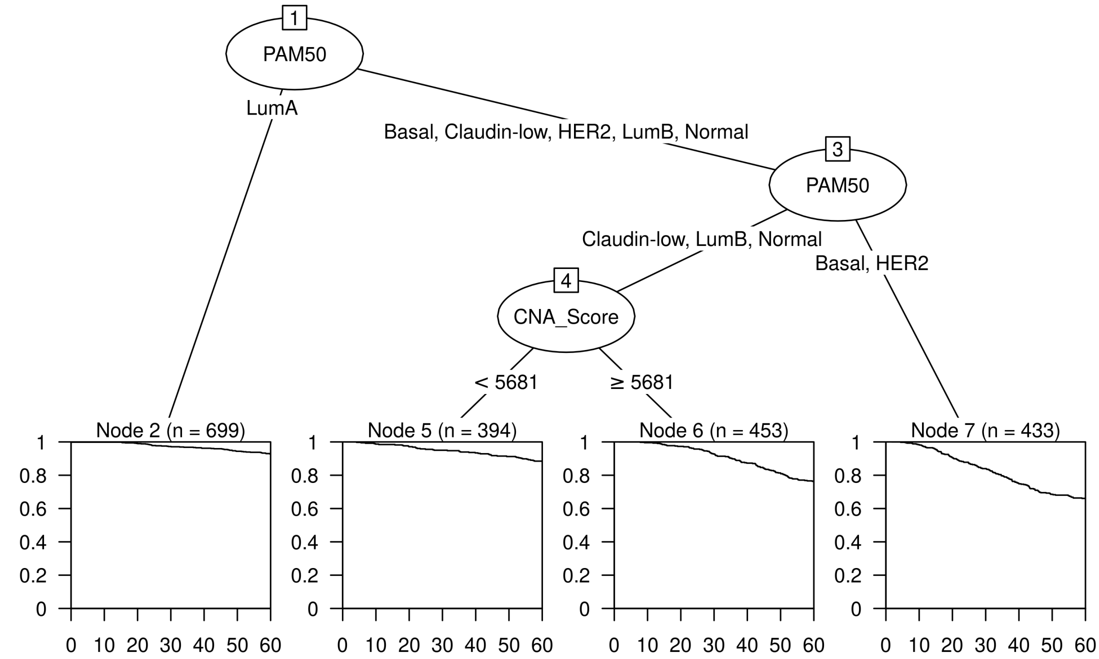


Figure 3.10: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype and the six CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

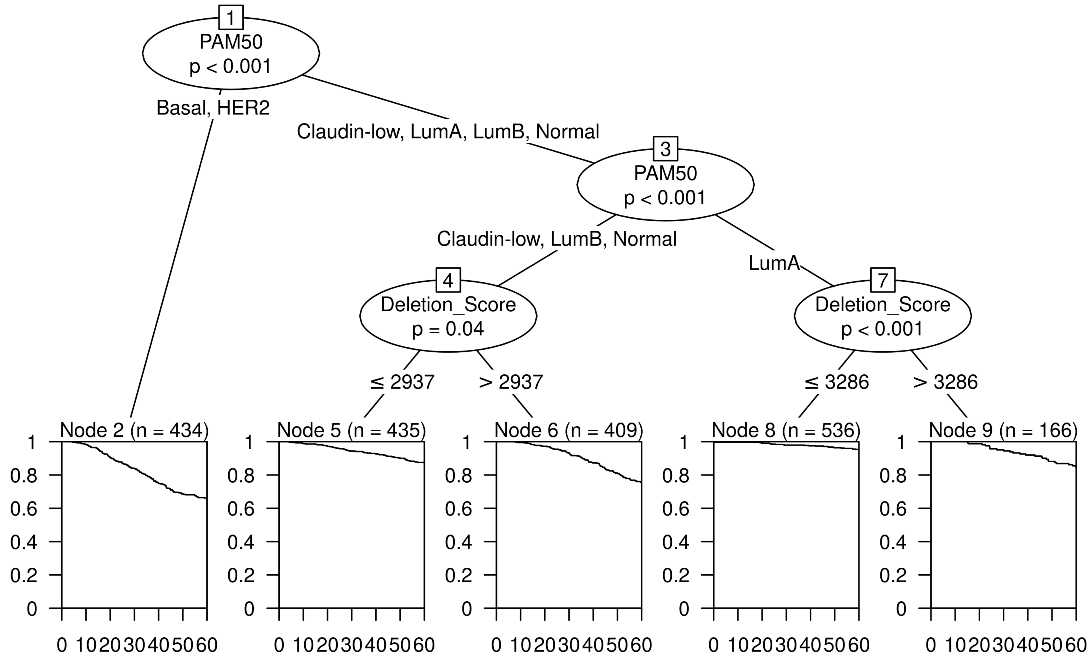
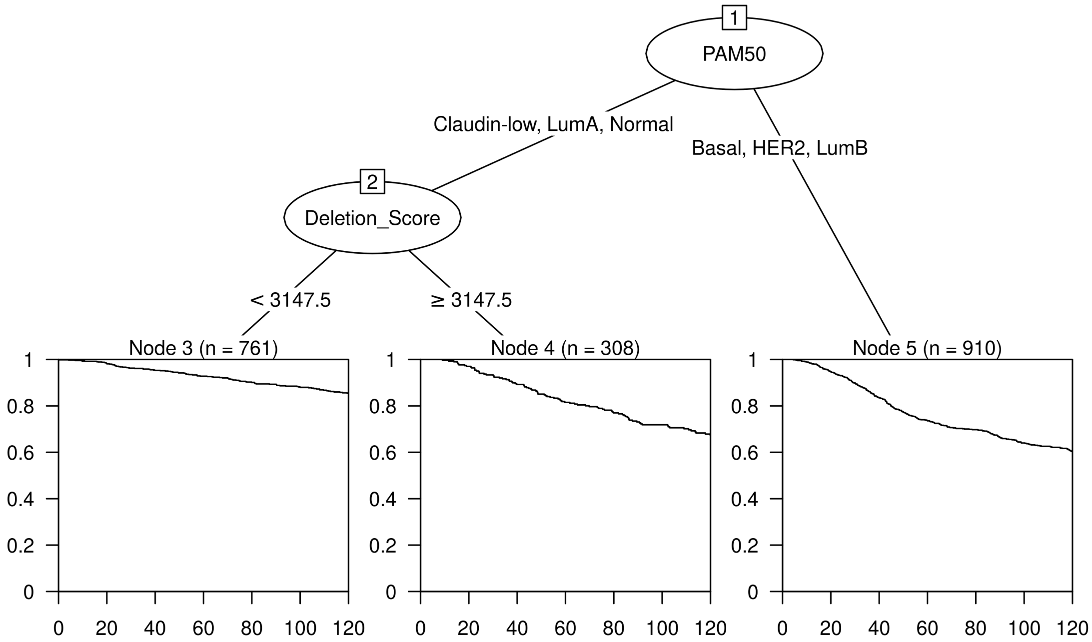


Figure 3.11: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype and the six CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

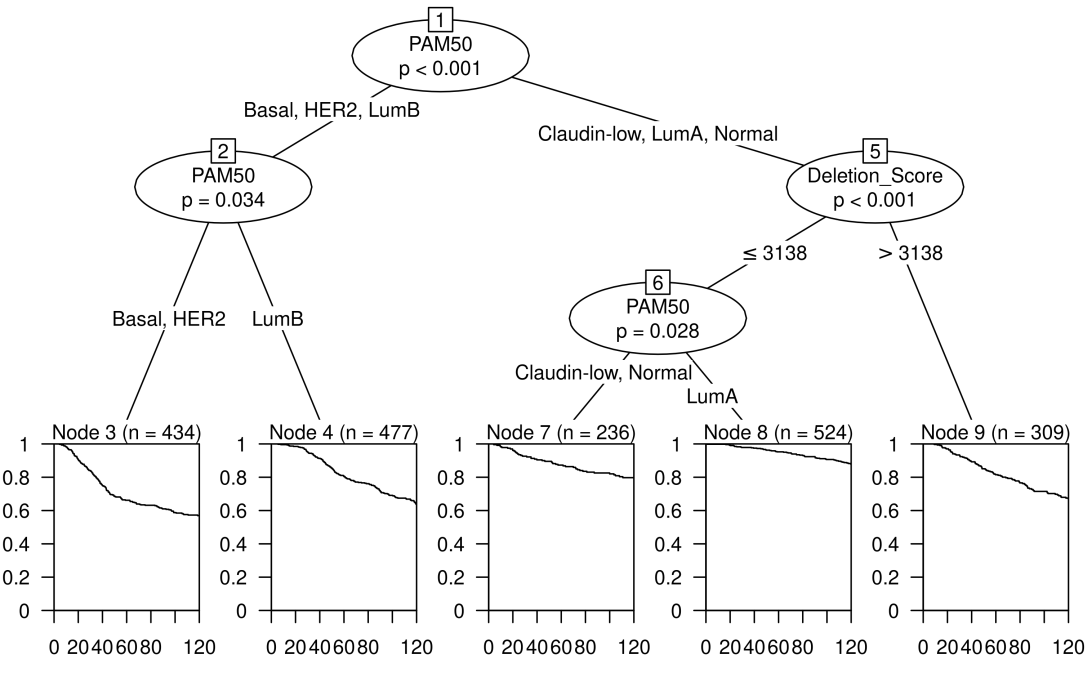
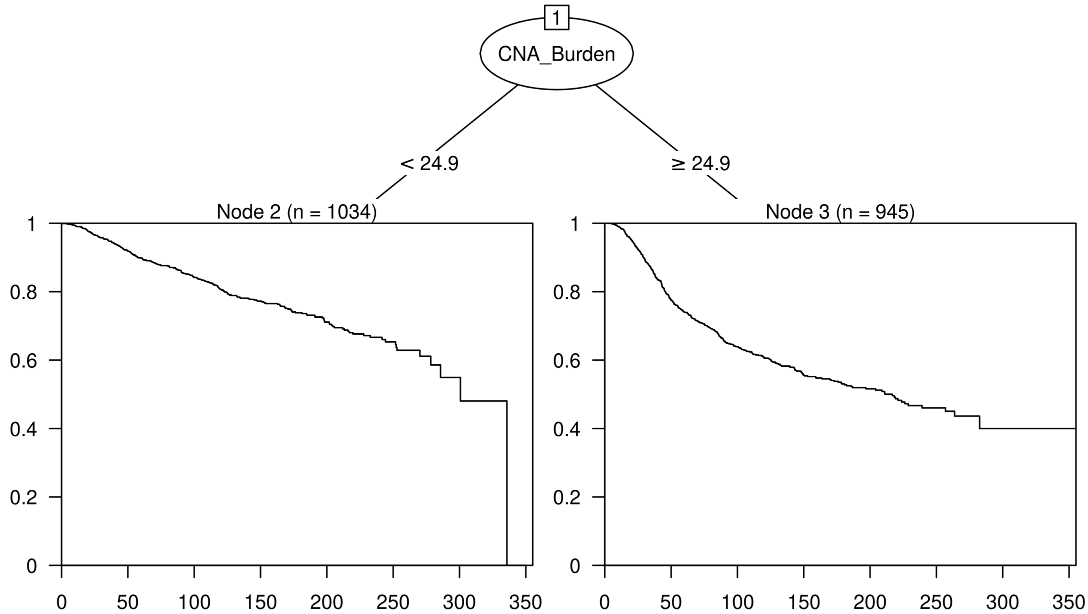


Figure 3.12: Recursive partitioning survival trees for ten-year disease-specific survival using PAM50 subtype and the six CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

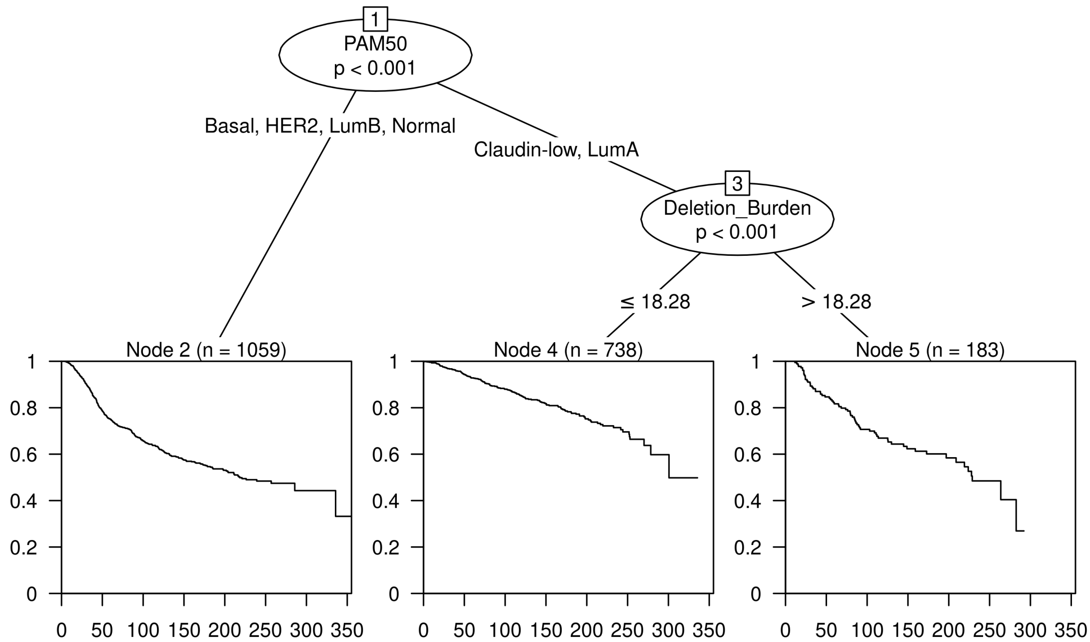
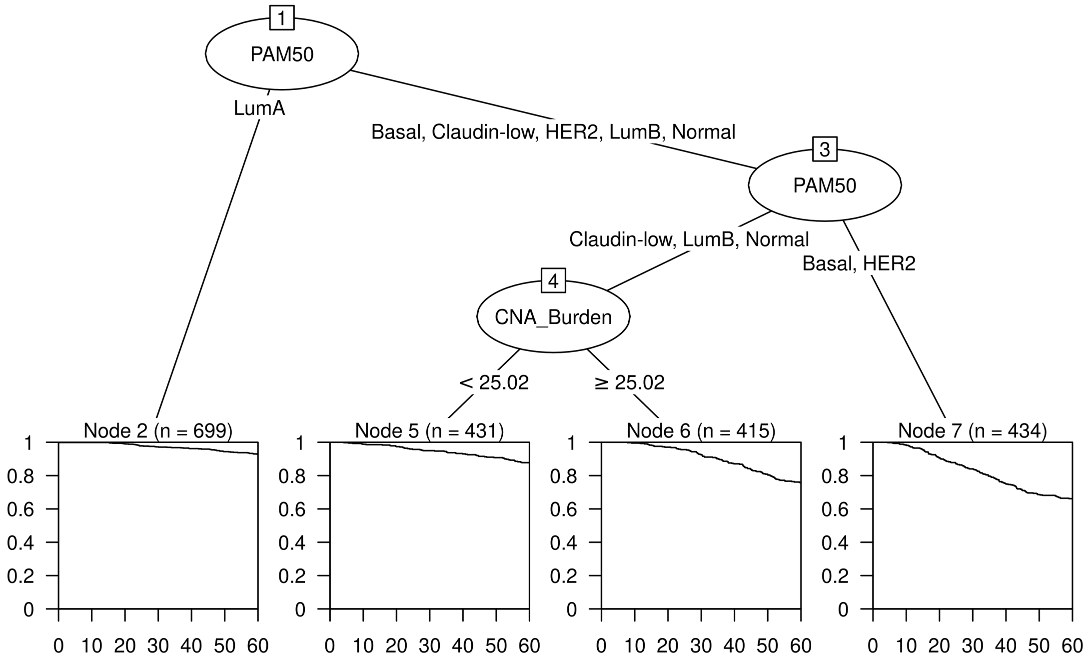


Figure 3.13: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype and the six CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

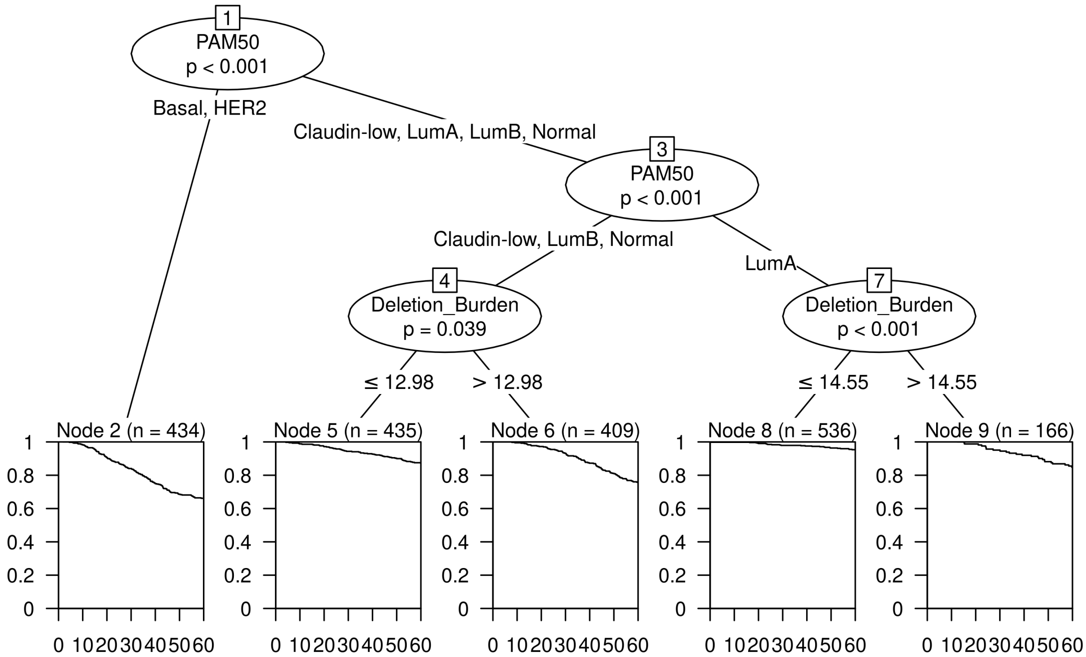
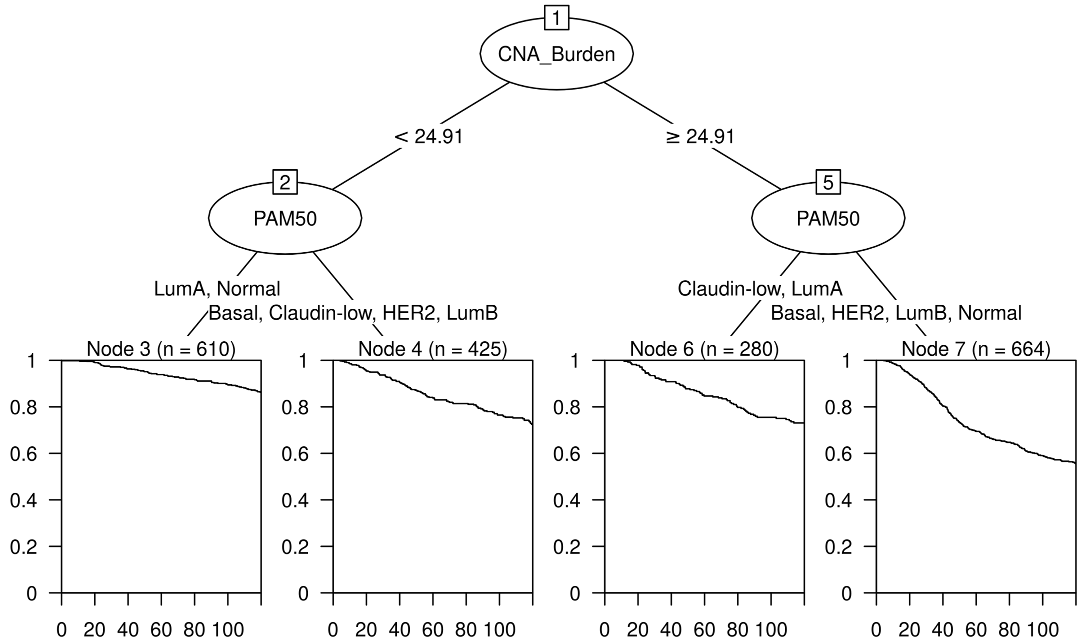


Figure 3.14: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype and the six CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

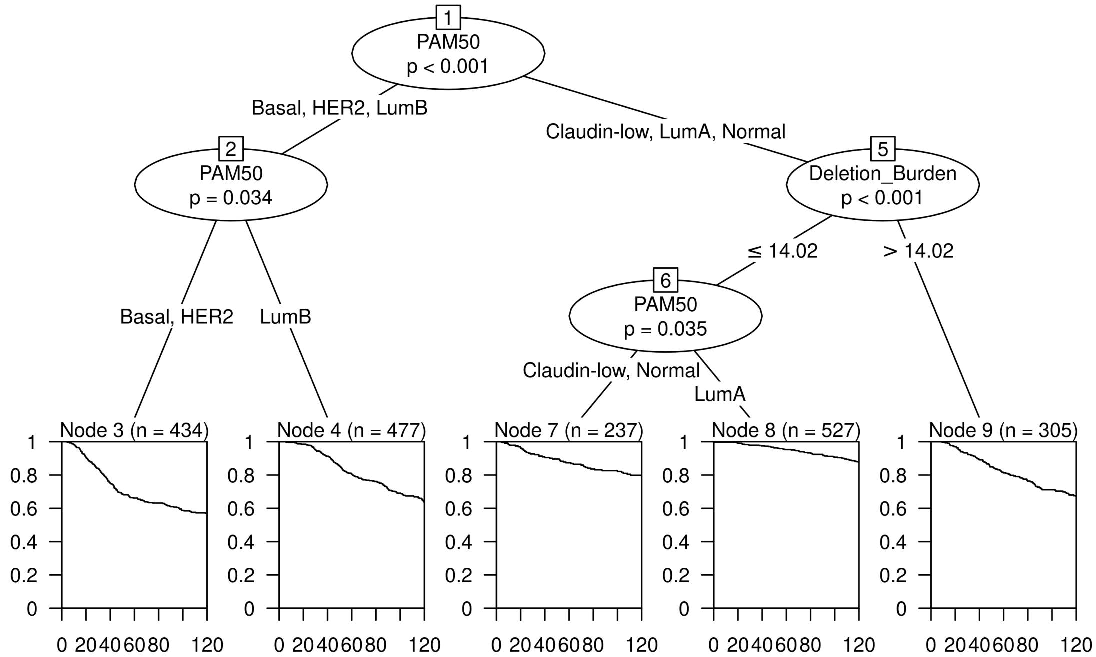
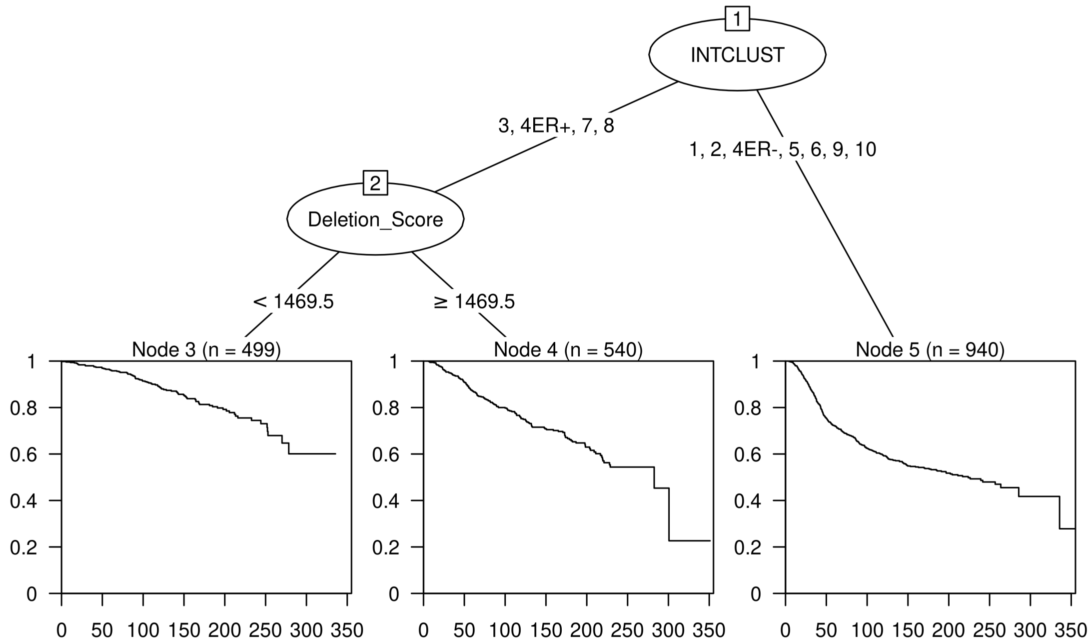


Figure 3.15: Recursive partitioning survival trees for ten-year disease-specific survival using PAM50 subtype and the six CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

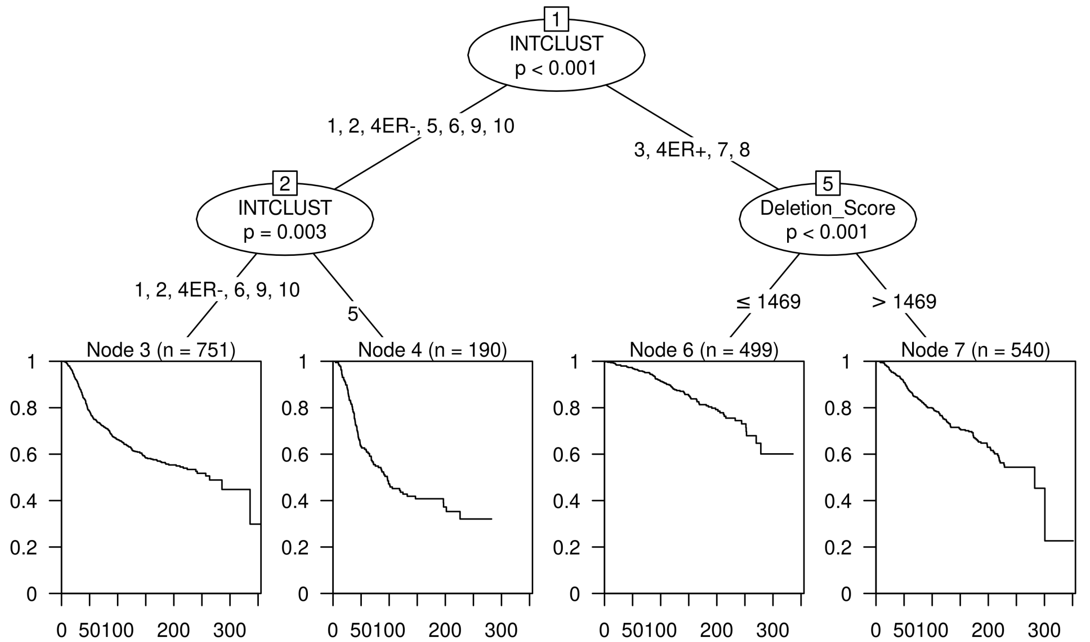
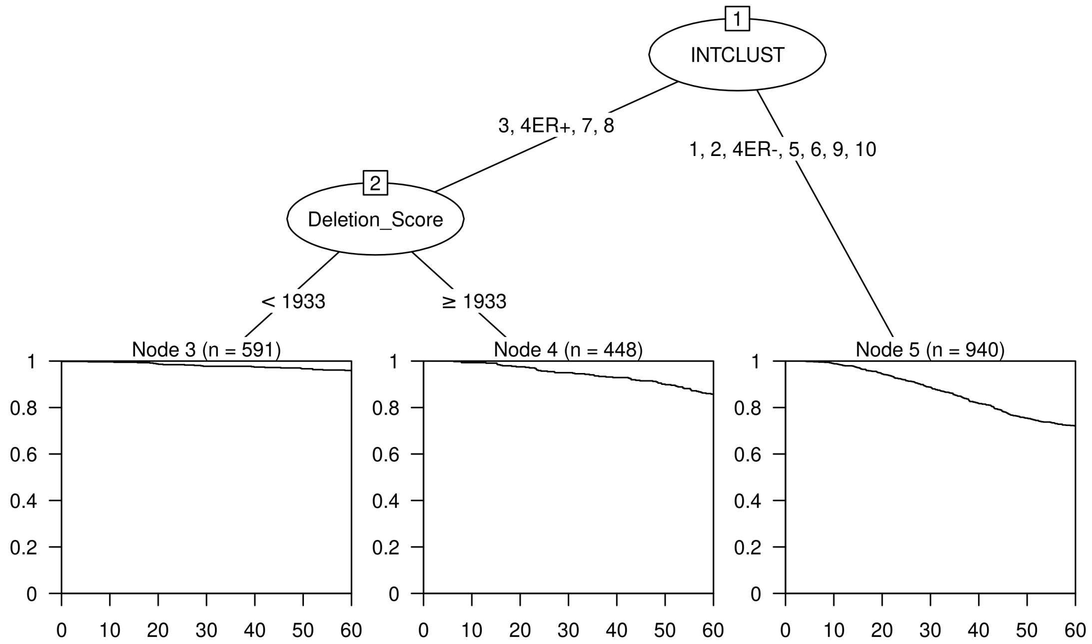


Figure 3.16: Recursive partitioning survival trees for disease-specific survival using IntClust and the six CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

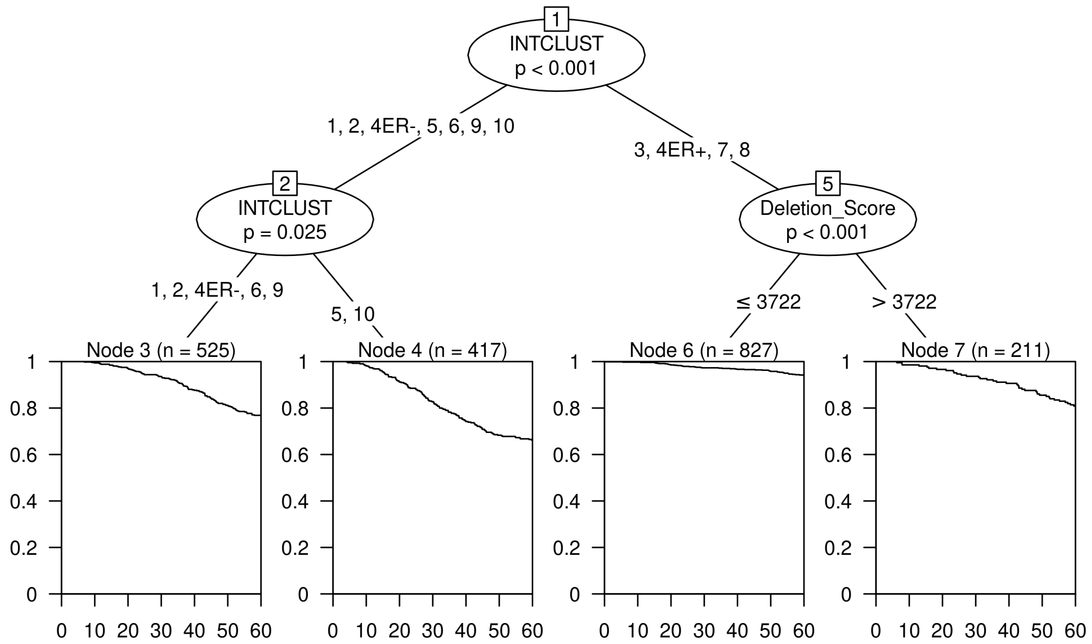
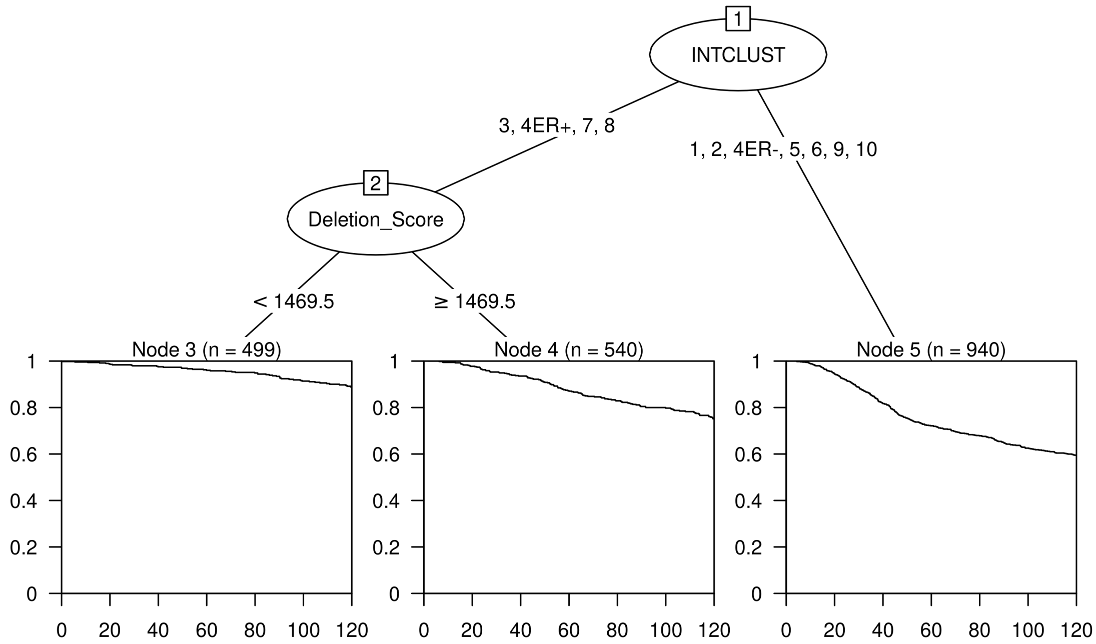


Figure 3.17: Recursive partitioning survival trees for five-year disease-specific survival using IntClust and the six CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

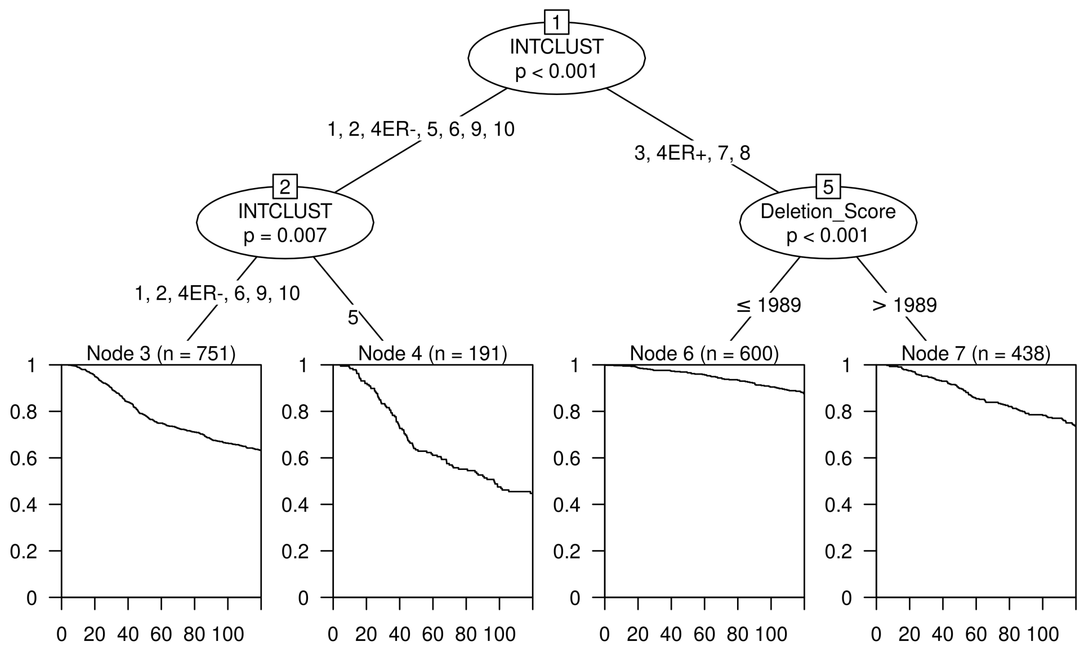
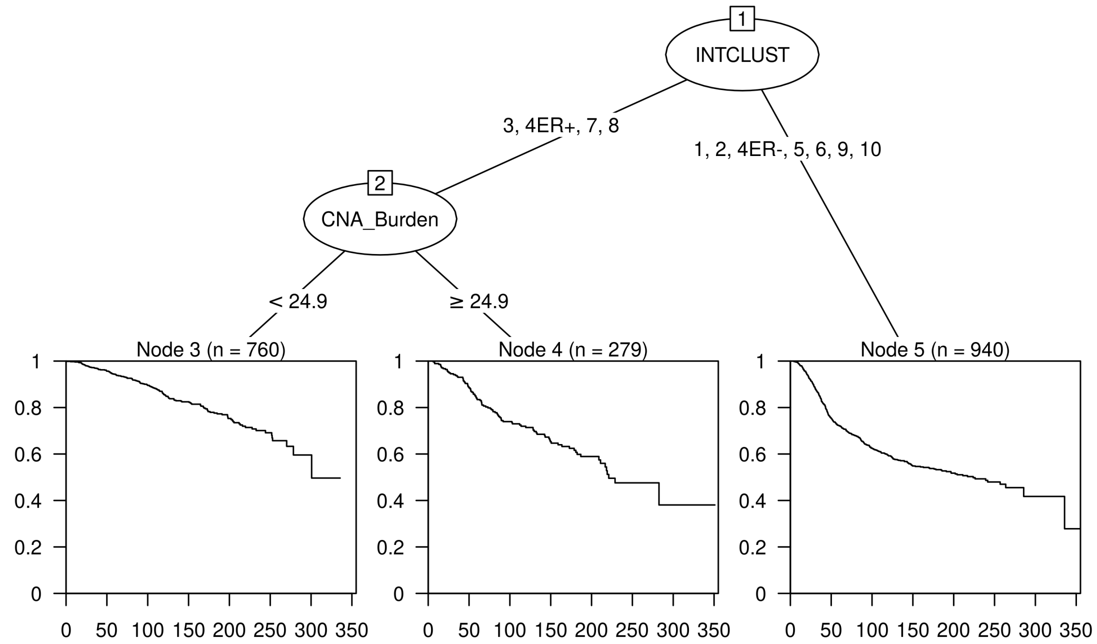


Figure 3.18: Recursive partitioning survival trees for ten-year disease-specific survival using IntClust and the six CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

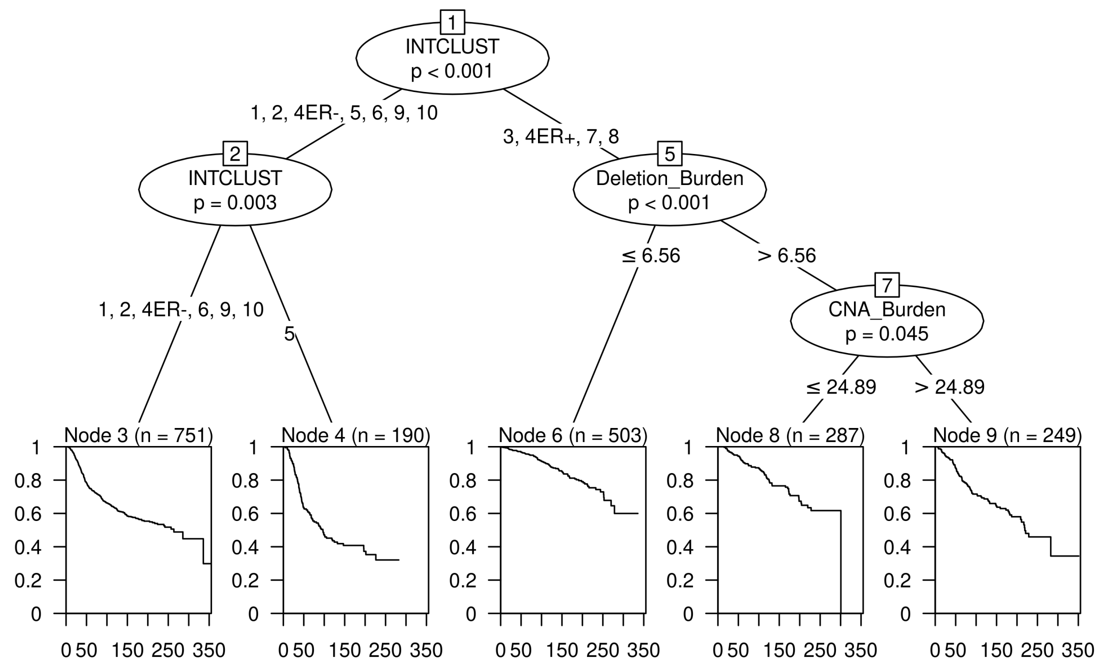
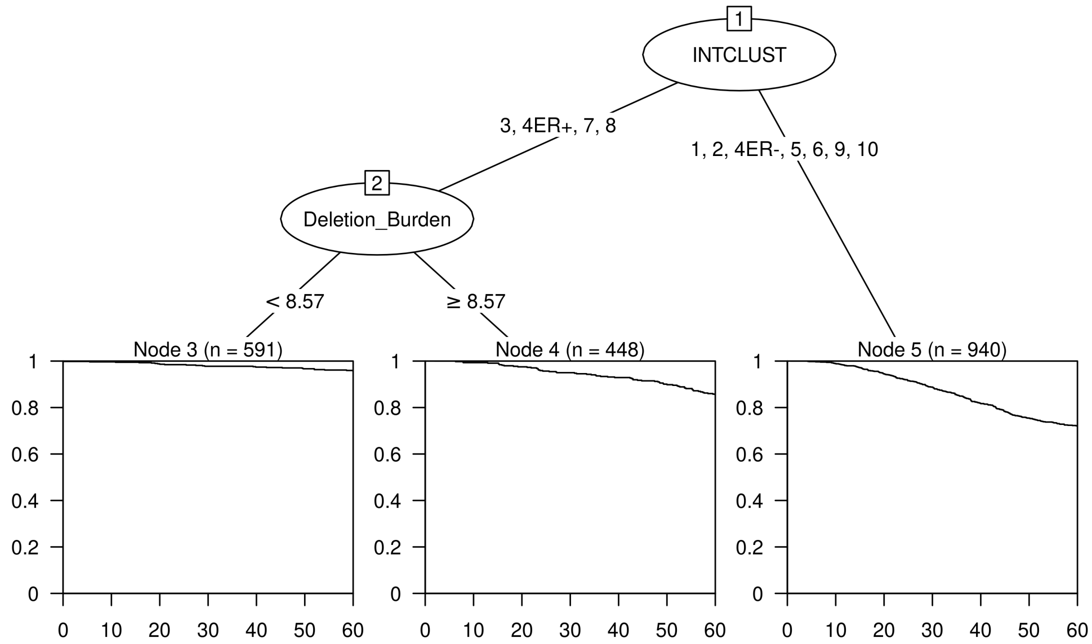


Figure 3.19: Recursive partitioning survival trees for disease-specific survival using IntClust and the six CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

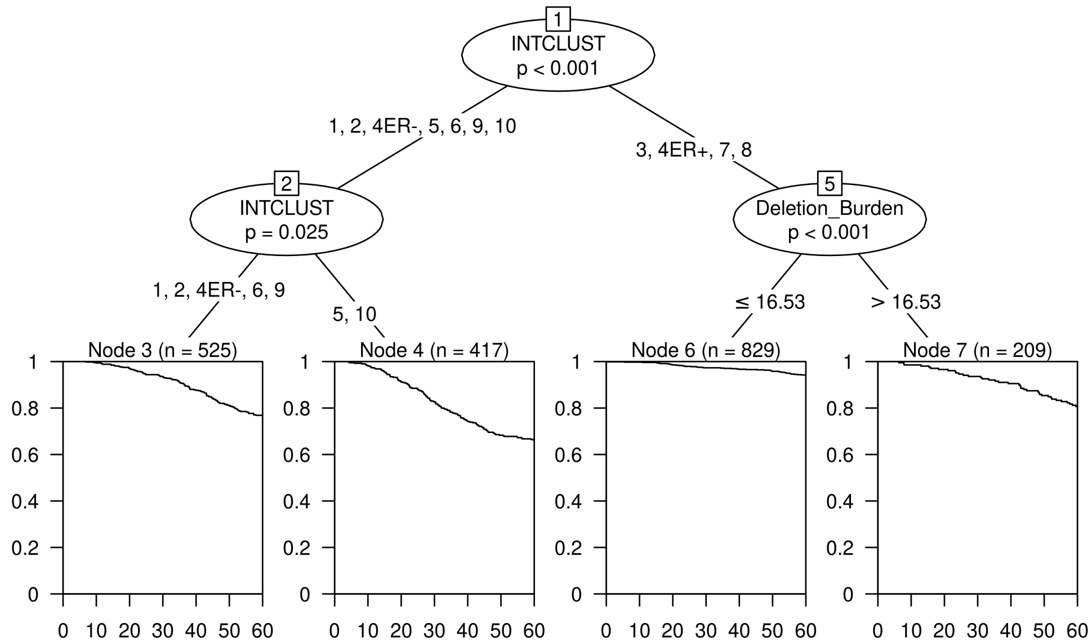
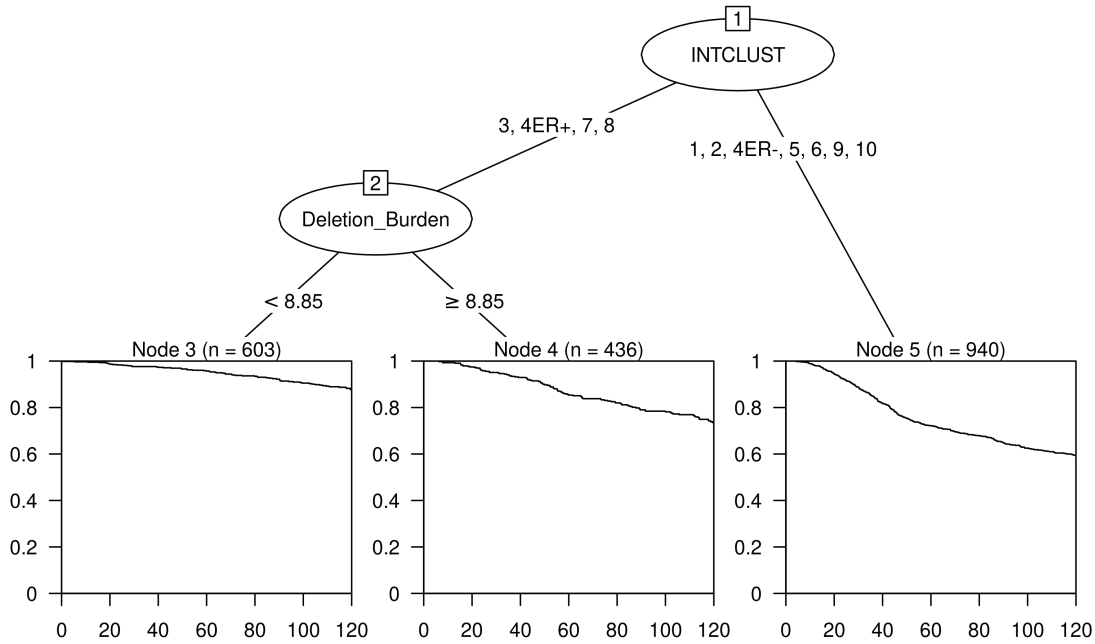


Figure 3.20: Recursive partitioning survival trees for five-year disease-specific survival using IntClust and the six CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

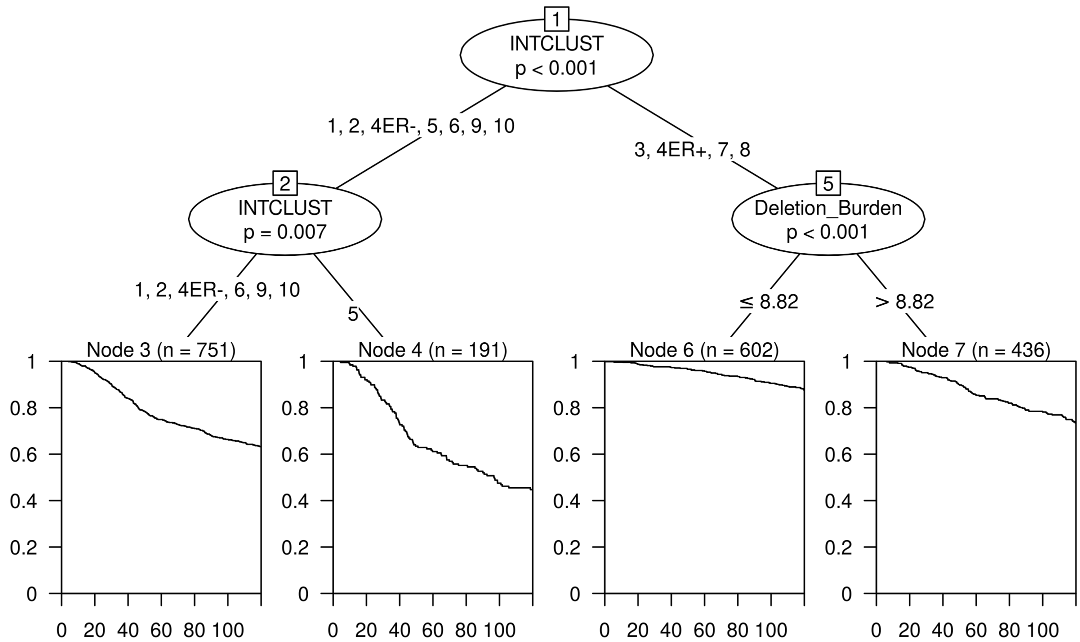
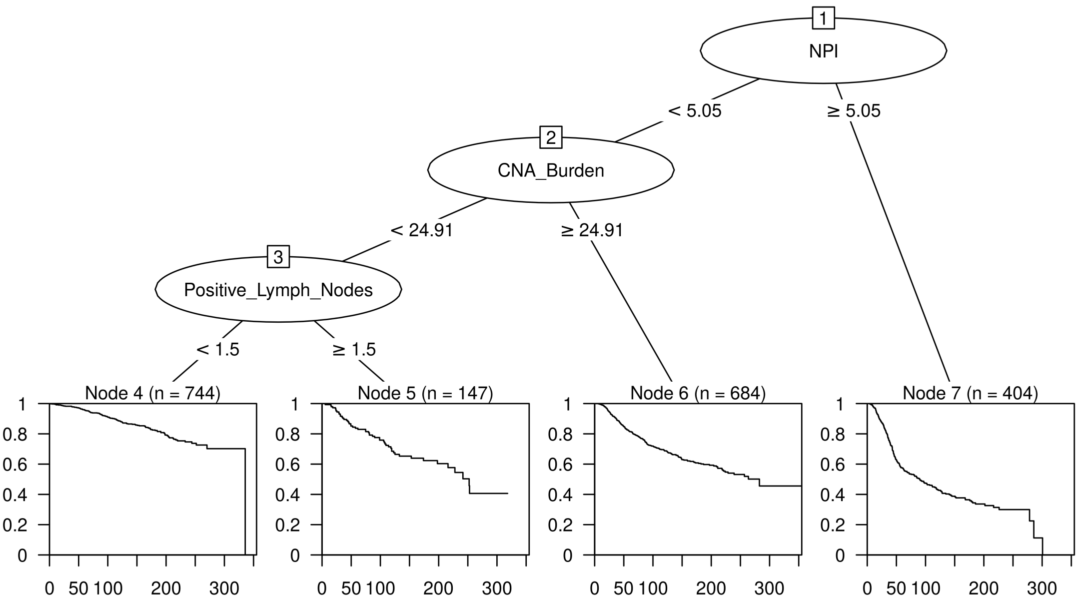


Figure 3.21: Recursive partitioning survival trees for ten-year disease-specific survival using IntClust and the six CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

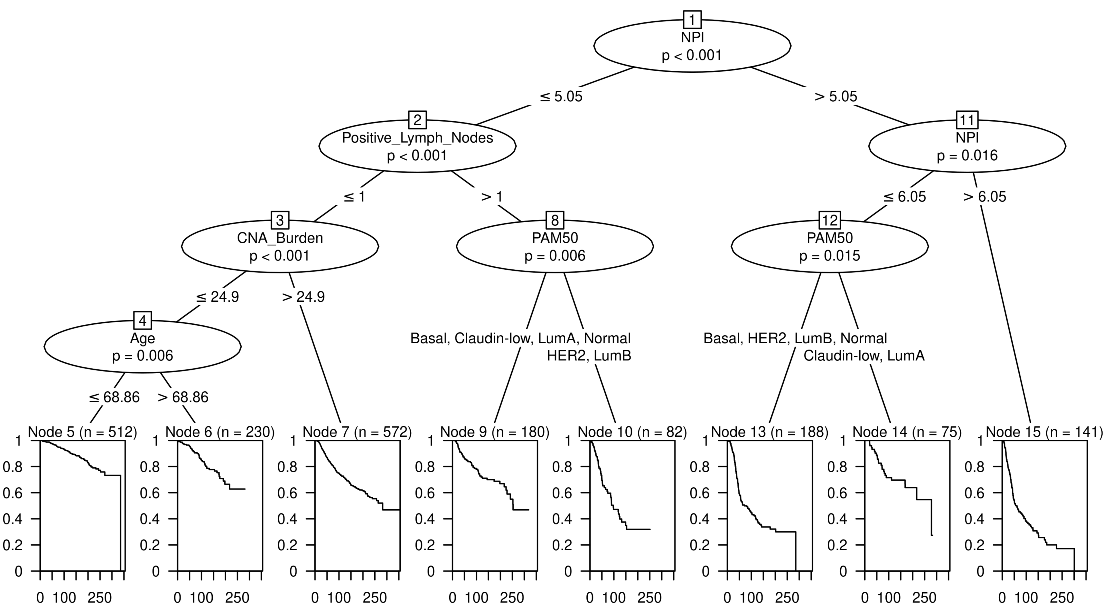
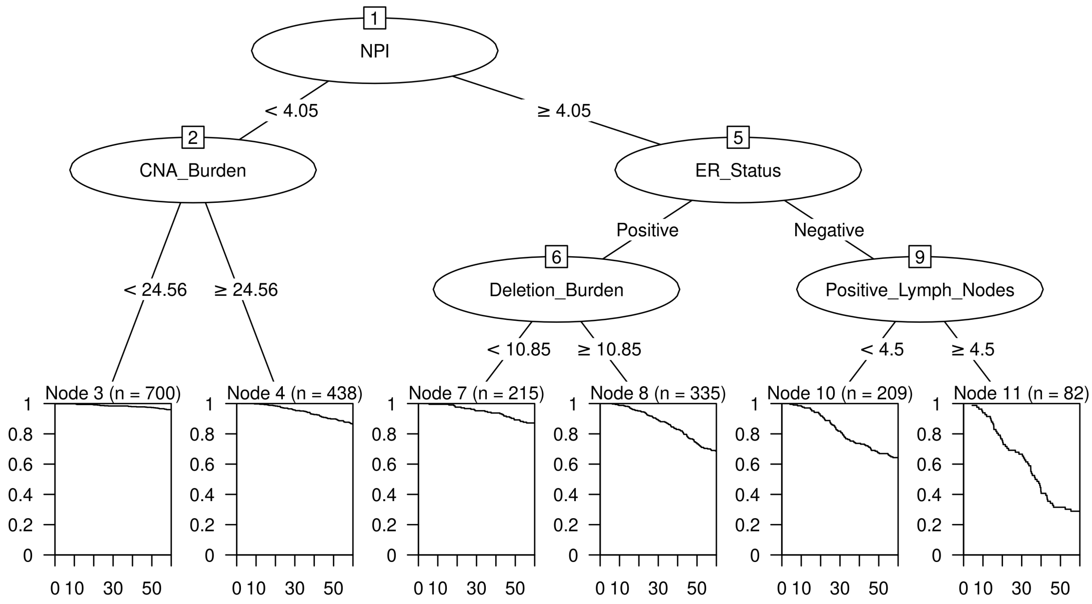


Figure 3.22: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype, the six CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

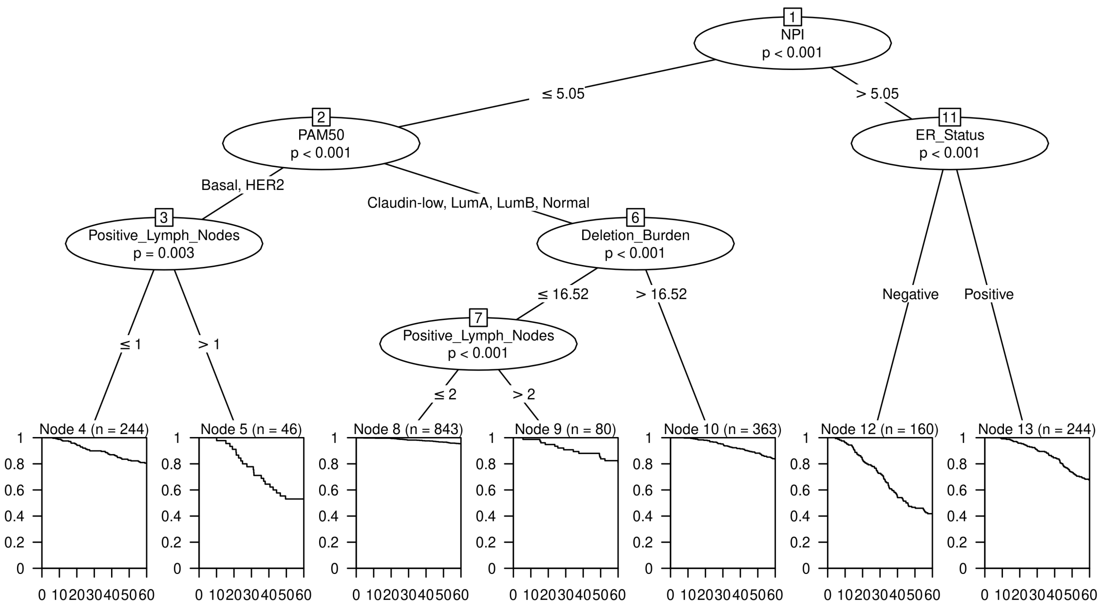


Figure 3.23: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype, the six CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

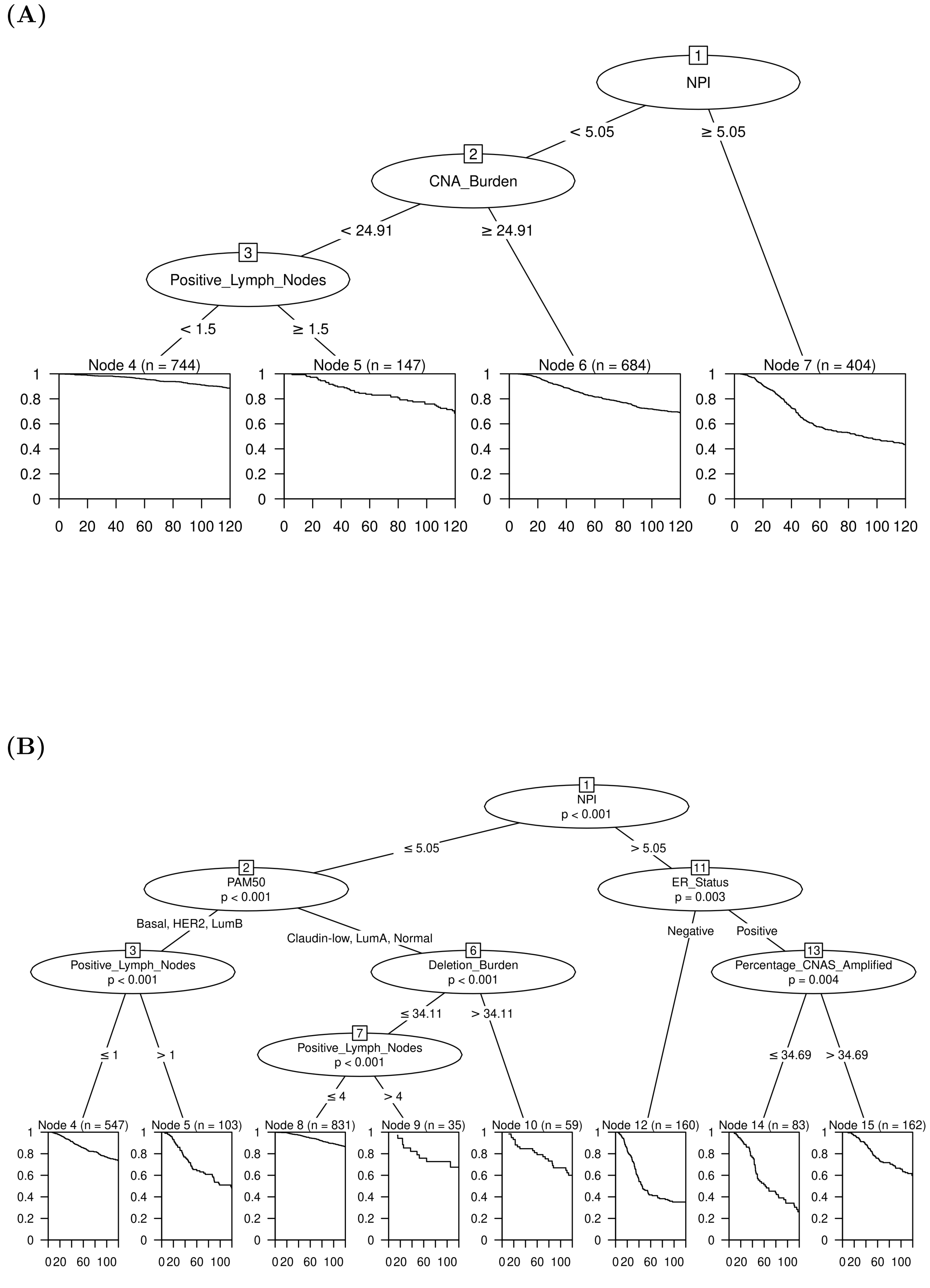
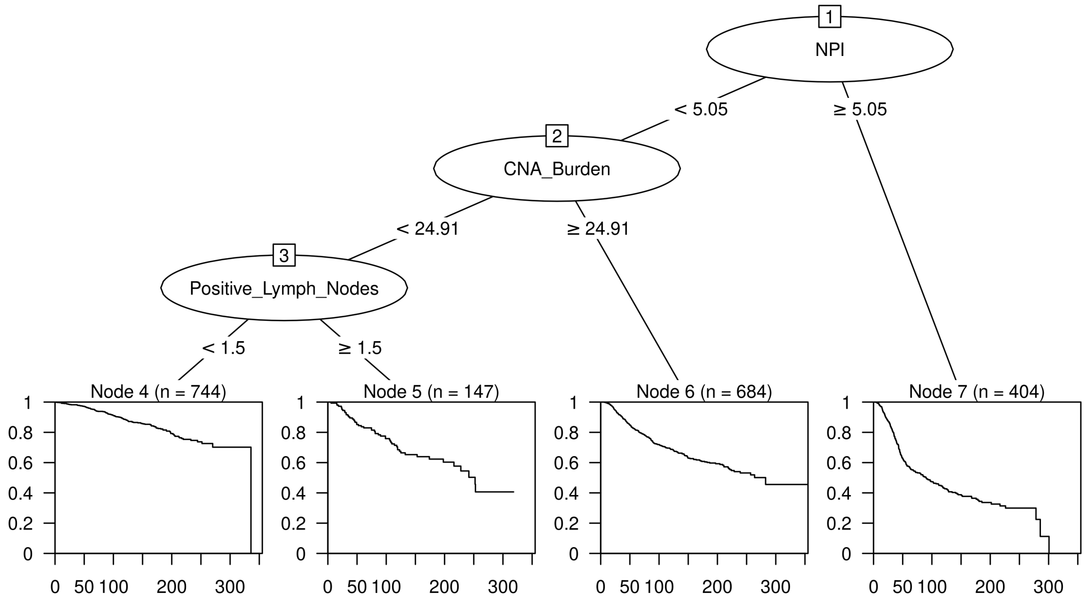


Figure 3.24: Recursive partitioning survival trees for ten-year disease-specific survival using PAM50 subtype, the six CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

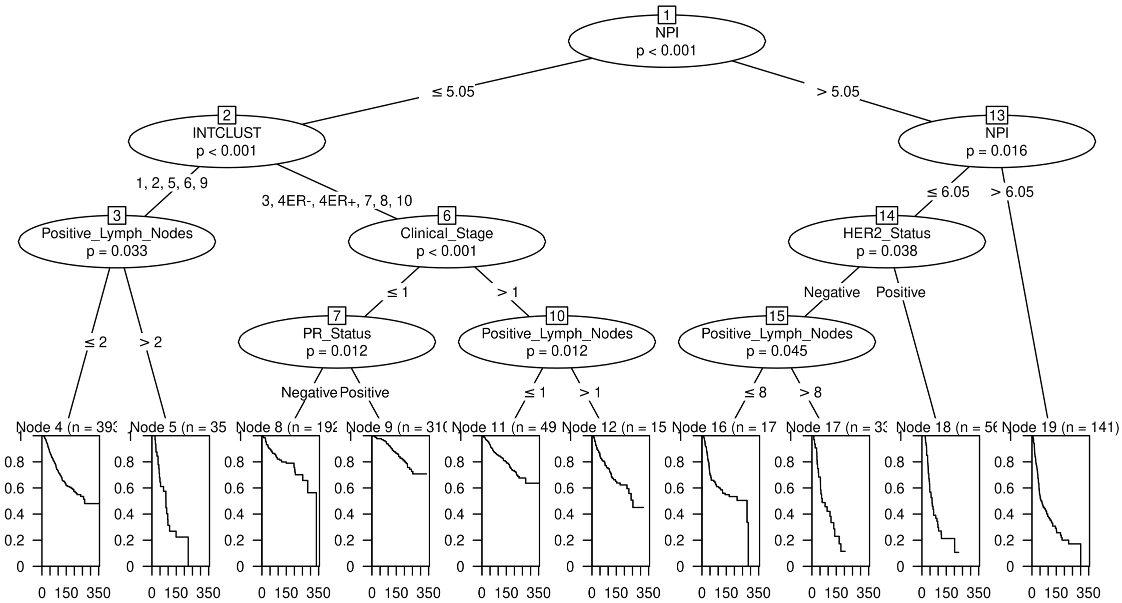
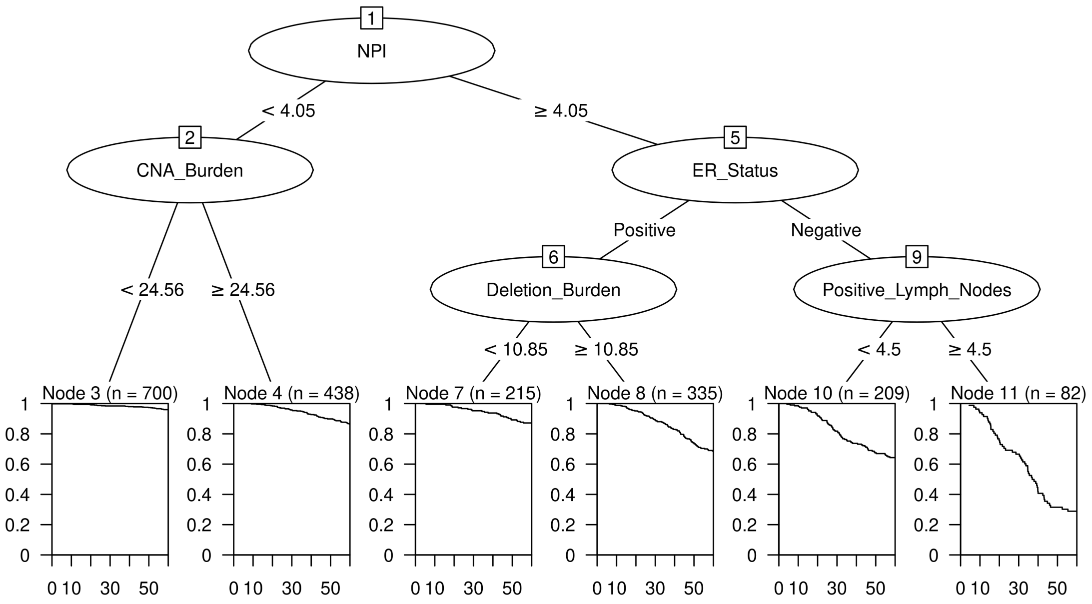


Figure 3.25: Recursive partitioning survival trees for disease-specific survival using IntClust, the six CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

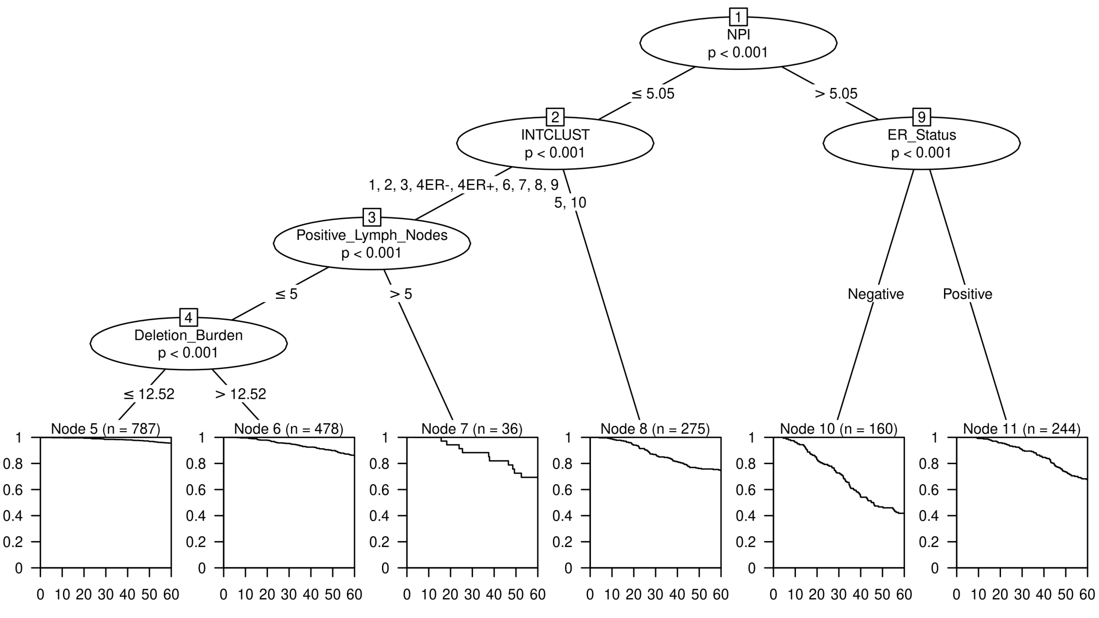
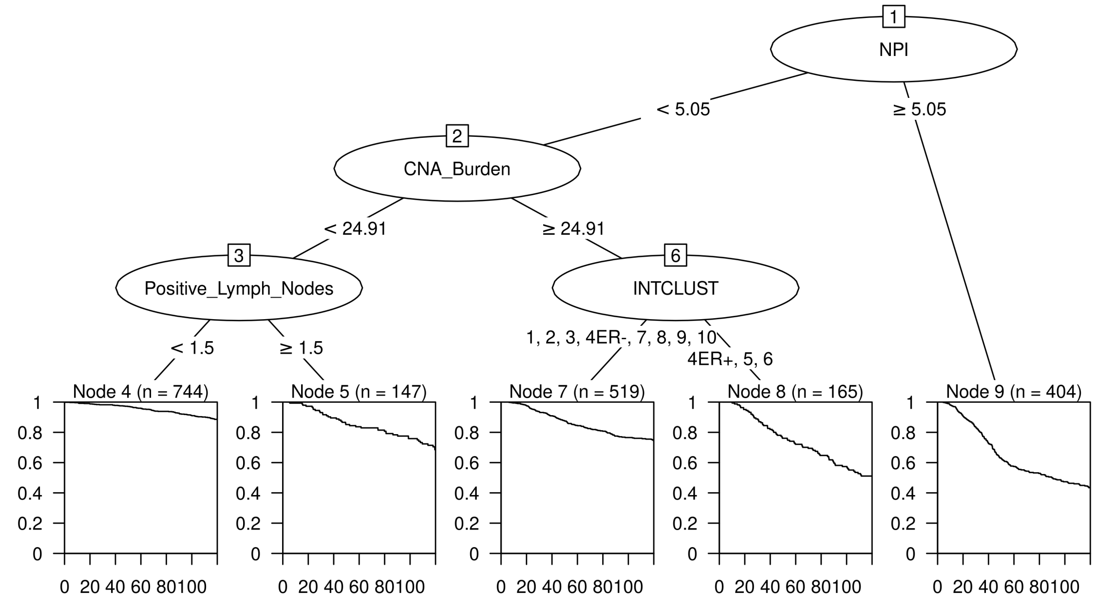


Figure 3.26: Recursive partitioning survival trees for five-year disease-specific survival using IntClust, the six CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

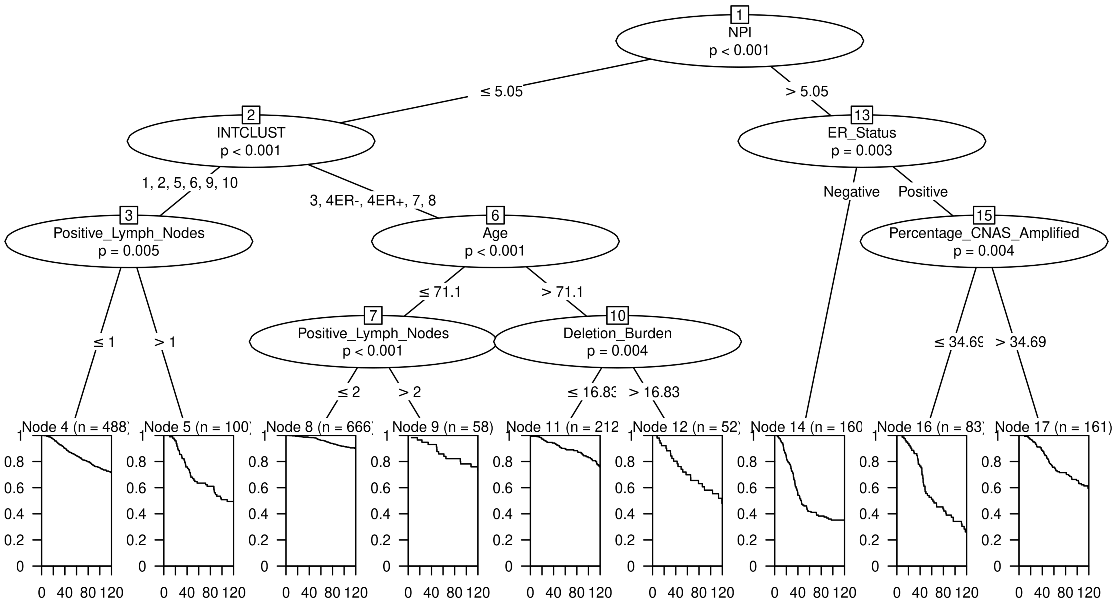


Figure 3.27: Recursive partitioning survival trees for ten-year disease-specific survival using IntClust, the six CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

to partition the patients with NPI < 4.05 and CNA Del Burden is used to partition patients with NPI ≥ 4.05 and ER positivity, with threshold 10.85% (Figure 3.23A). For the 5-year DSS survival trees, fitted with the ctree algorithm, CNA Del Burden, with threshold 16.52%, is used to split patients with NPI ≤ 5.05 who correspond to the Claudin-low, Luminal A, Luminal B and Normal PAM50 subtypes (Figure 3.23B). Similar partitions are observed in the 10-year DSS trees (Figure 3.24), where CNA Burden and CNA Del Burden are utilised in sub-partitions of the trees. Interestingly, even with the addition of traditionally used clinical variables, the CNA Burden metrics still appear useful in stratifying patients based on DSS, 5-year and 10-year DSS. In particular, CNA Del Burden is again used to partition Luminal A, Claudin-low and Normal patients. In all cases, patients in the partition corresponding to the lower GI have better disease-specific survival outcomes.

Focusing on survival trees for DSS, 5-year DSS and 10-year DSS, that have the six CNA Burden metrics, IntClust molecular classification and the selected clinical variables as candidate predictors, similar tree structures are observed. Again, CNA Burden, CNA Del Burden and Percentage Amp Burden appear to be useful in stratifying patients in the context of disease-specific survival (Figures 3.25-3.27).

3.4 Analysis of Chromosome Arm CNA Metrics across All METABRIC Patients

In addition to expanding the study focus to all patients, we also broaden the analysis by using the 42 chromosome arm CNA Score and Burden metrics as candidate predictors in the survival trees. These chromosome arm CNA metrics are initially included with PAM50 or IntClust molecular classifications to assess whether the chromosome arm CNA metric information can add additional prognostic value to the molecular classifications, and then included with a selection of clinical variables to explore interactions between the clinical variables and CNA metrics.

3.4.1 Chromosome Arm CNA Metric Survival Trees, in Combination with Molecular Classification Predictors

There was less consistency observed in the survival trees produced using the chromosome arm CNA metrics than in the survival trees using the global CNA metrics. This may be due to the increased number of candidate predictors used in the chromosome arm CNA metric survival trees (43 candidate predictors) compared to the global CNA metric survival trees where seven predictors are included. Including the 42 chromosome arm CNA metrics, along with PAM50 or IntClust molecular classification, as candidate predictors enables us to determine if the CNA Score or Burden on specific chromosome arms is useful in stratifying patients on DSS, 5-year DSS and 10-year DSS outcomes.

Focusing on the survival trees including the CNA Score metrics (Figure 3.28-3.30), it is noted that all trees initially split on PAM50 subtype followed by either one or more of the CNA Score metrics, or PAM50 subtype again. While variation is observed between trees for DSS, 5-year DSS and 10-year DSS, CNA Score metrics corresponding to chromosome 3p and 18q appear most frequently as important predictors for DSS.

For example, the CNA Difference Score and CNA Del Score landscape on chromosome 3p is useful in stratifying patients (Figures 3.28 and 3.30). Luminal A and Claudin-low patients with a 3p CNA Difference Score < -5.5 or CNA Del Score > 184, have reduced DSS when compared to patients with CNA Difference Score ≥ -5.5 or CNA Del Score ≤ 184, respectively (Figure 3.28). CNA Difference Score and CNA Del Score on chromosome 3p also appear as significant predictors of 10-year DSS survival, with thresholds -6.5 and 6, respectively (Figure 3.30).

Chromosome 18q CNA Score metrics also appear as useful predictors of survival across PAM50 subtypes, mainly Claudin-low, Luminal A, Luminal B and Normal patients (Figures 3.29 and 3.30). Claudin-low, Luminal A, Luminal B and Normal patients with Percentage Del Score on chromosome 18q $> 55.56\%$ have decreased 5-year DSS. Patients with Percentage Del Score on chromosome 18q $\leq 55.56\%$ are further partitioned based on CNA Del Score on chromosome 4p and then on PAM50 subtype (Figure 3.29). In terms of 10-year DSS, CNA Del Score on chromosome 18q is useful in stratifying Claudin-low, Luminal A and Normal patients with CNA Del Score on chromosome 3p ≤ 6 and CNA Amp Score on 11q ≤ 298 (Figure 3.30). Patients in these groups, with 18q CNA Del Score > 191 have worse 10-year DSS survival outcomes than patients with CNA Del Score below this threshold. Other chromosome arm CNA Scores observed include CNA Del Score on chromosome 11p, partitioning Luminal A patients with a threshold of 38.5 and CNA Score on chromosome 17p, partitioning Claudin-low, Luminal B and Normal patients with a threshold of 5 (Figure 3.29).

The survival trees generated using the CNA Burden metrics and PAM50 subtype as candidate predictors partition the patients similarly to the CNA Score survival trees, with high congruence seen in the chromosome arm metrics highlighted as useful predictors of survival (Figure 3.31-3.33). For example, the CNA Del Burden landscape on chromosome 3p is again identified as a useful predictor in stratifying patients (Figure 3.31 and 3.33). Luminal A and Claudin-low patients with a 3p CNA Del burden higher than 30.21% have reduced DSS when compared to patients with lower 3p CNA Del burden (Figure 3.31B). This highlights that irrespective of using CNA Del Score on chromosome 3p with threshold 184 (Figure 3.28B) or CNA Del Burden with a threshold of 30.21% (Figure 3.31B), Claudin-low and Luminal A patients are partitioned into Node 4 with $n = 794$ and Node 5 with $n = 128$.

The majority of the chromosome CNA Score and Burden metrics selected as useful predictors corresponded to either the deletion or difference metrics. This is similar to what was observed in the survival trees including the global CNA Score and Burden metrics as candidate predictors: patients that have chromosome arm CNA Score and Burden metrics above the optimised threshold have worse survival outcomes.

Focusing on the survival trees including IntClust classification and CNA Score metrics (Figure 3.34-3.36), it is observed that all trees initially split on IntClust classification, grouping IntClust 1, 2, 4ER-, 5, 6, 9, 10 together and IntClust 3, 4ER+, 7 and 8 together, followed by CNA Del Score on chromosome 18q, and in some cases, other chromosome arm CNA Score metrics. CNA Del Score on chromosome 18q, with a threshold of ≈ 160, consistently appears as an important predictor for DSS, 5-year DSS and 10-year DSS. For example, CNA Del Score on chromosome 18q partitions IntClust 3, 4ER+, 7 and 8 patients into two groups, ≥ 160.5 and < 160.5 (Figure 3.34A and Figure 3.36A). Similar to what is observed in previously fitted survival trees, patients that have CNA Del Score above an optimised threshold have poorer survival outcomes. In trees where patients were partitioned using additional chromosome arm CNA Score metrics, CNA Score on chromosome 1p, CNA Amp Score on 6q, CNA Amp Score on 12p, CNA Del Score on 4p, CNA Score on 11q and Difference Score on 1p, appear as useful predictors.

The survival trees including the CNA Burden metrics and IntClust molecular classification as candidate predictors, are similar to the trees including the CNA Score metrics (Figure 3.37-3.39).

The chromosome arms that were selected across the survival trees as useful predictors in the context of DSS, 5-year DSS and 10-year DSS, were chromosome arms 1p, 3p, 4p, 6q, 11p, 11q, 17p, 17q and 18q. However, CNA Del and Difference metrics on chromosome 3p, and CNA Del metrics on 18q are the predictors that appear most frequently across the chromosome arm CNA Score and Burden metric survival trees. In agreement with the global CNA Score Burden survival trees, the patients are partitioned initially on PAM50 subtypes and IntClusts, where subtypes or clusters that display low genomic instability and generally good prognosis are grouped together. Subsequently these patients are partitioned, using one or more of the CNA metrics (primarily CNA Del metrics) at an optimal cut-off point.

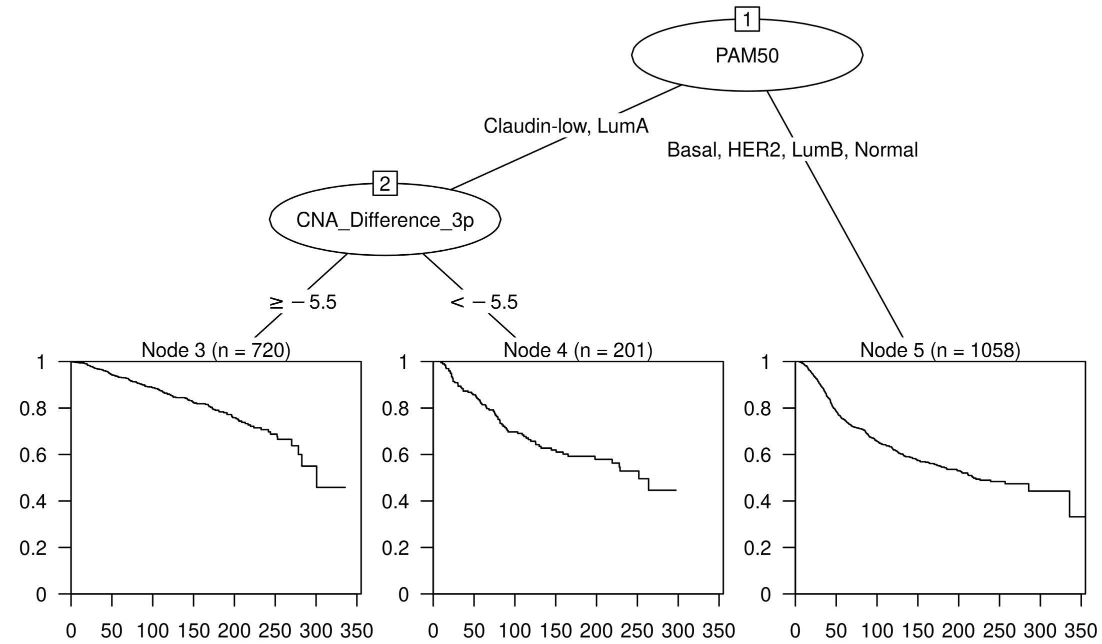
3.4.2 Chromosome Arm CNA Metric Survival Trees, in Combination with Molecular Classification and Clinical Predictors

To assess how the addition of clinical variables alters the observed partitioning and explore interactions between the clinical variables and CNA metrics in modelling DSS, 5-year DSS and 10-year DSS, survival trees including the six CNA Burden metrics, IntClust or PAM50 molecular classification, and selected clinical variables, are fitted (Figures 3.40-3.45).

The survival trees utilising the 42 chromosome arm CNA Burden metrics, PAM50 subtype, and selected clinical variables as candidate predictors indicate that a number of chromosome arm CNA Burden metrics, PAM50 subtype and several of the selected clinical variables are identified as useful predictors of DSS, 5-year DSS and 10-year DSS. All trees initially partition on NPI, with thresholds ranging from 4.05 to 5.05, and then on one or more clinical predictors, PAM50 subtype, or chromosome arm CNA Burden metric. For example, the ctree survival tree modelling DSS (Figure 3.40B), partitions patients into six groups, also referred to as nodes, with Node 5 displaying the best DSS. Node 5 corresponds to patients with NPI ≤ 5.05, number of positive lymph nodes ≤ 1, CNA Burden on chromosome 1p ≤ 15.81 and CNA Burden on chromosome 4p ≤ 34.09. Despite the addition of the clinical variables, CNA Del Burden metrics on chromosome 3p and chromosome 18q are still observed in survival trees modelling 5-year DSS and 10-year DSS, particularly in Claudin-low, Luminal A, Luminal B and Normal patients (Figures 3.41 and 3.42). In addition, CNA Amp Burden on chromosome 9q, CNA Difference on chromosome 5q, CNA Difference on chromosome 17p, CNA Burden on 11q, CNA Burden on 16p and CNA Amp Burden on 16p also appear as useful predictors in subgroups of patients. The survival trees including the 42 chromosome arm CNA Burden metrics, IntClust molecular classification, and the selected clinical variables also consistently partition on NPI, with thresholds ranging from 4.05 to 5.05 (Figures 3.43-3.45).

Variation is observed when comparing the chromosome arm CNA Burden metrics and clinical variables selected by the recursive partitioning survival trees considering

(A)



(B)

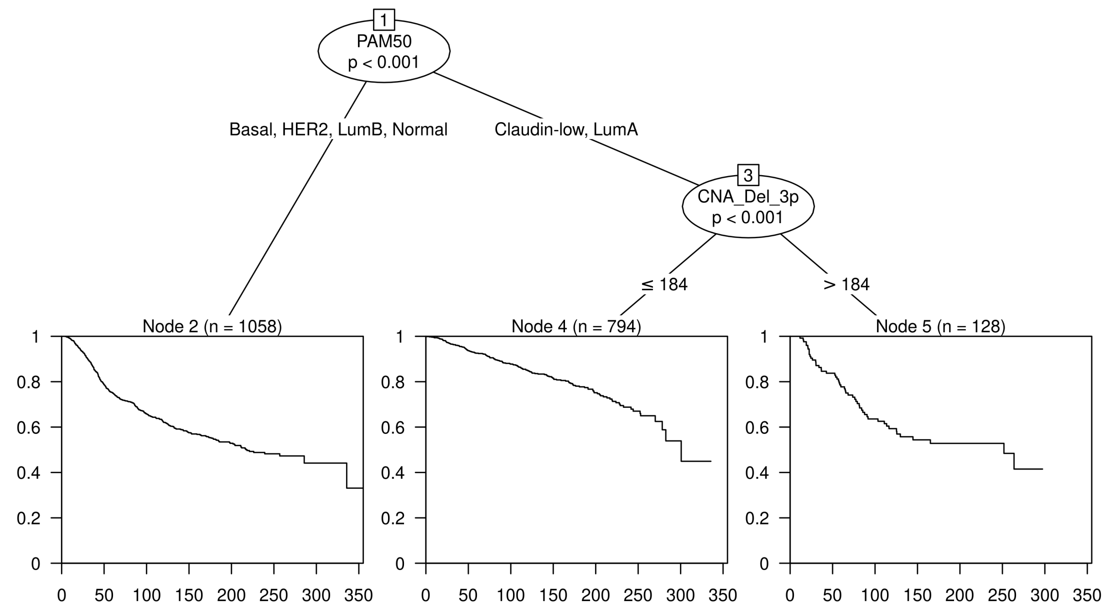
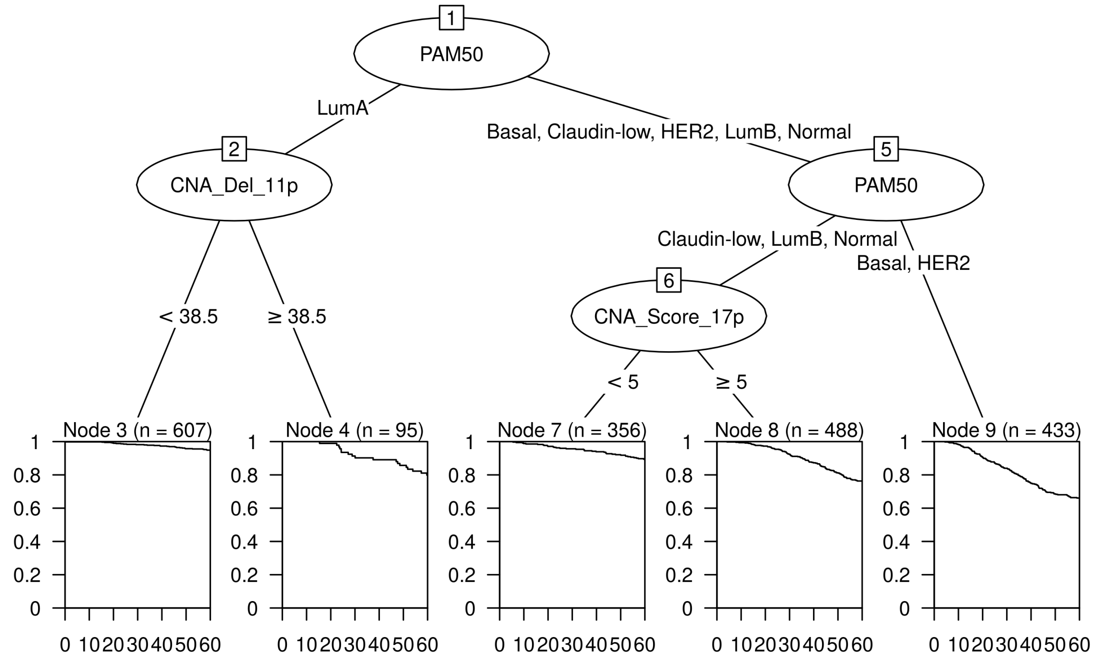


Figure 3.28: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

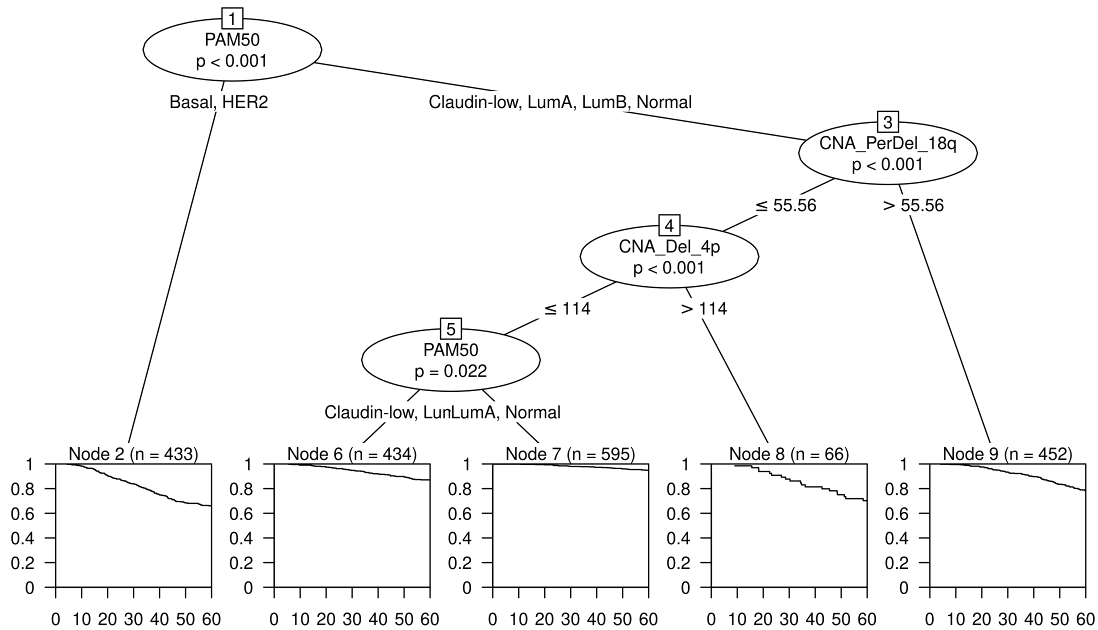
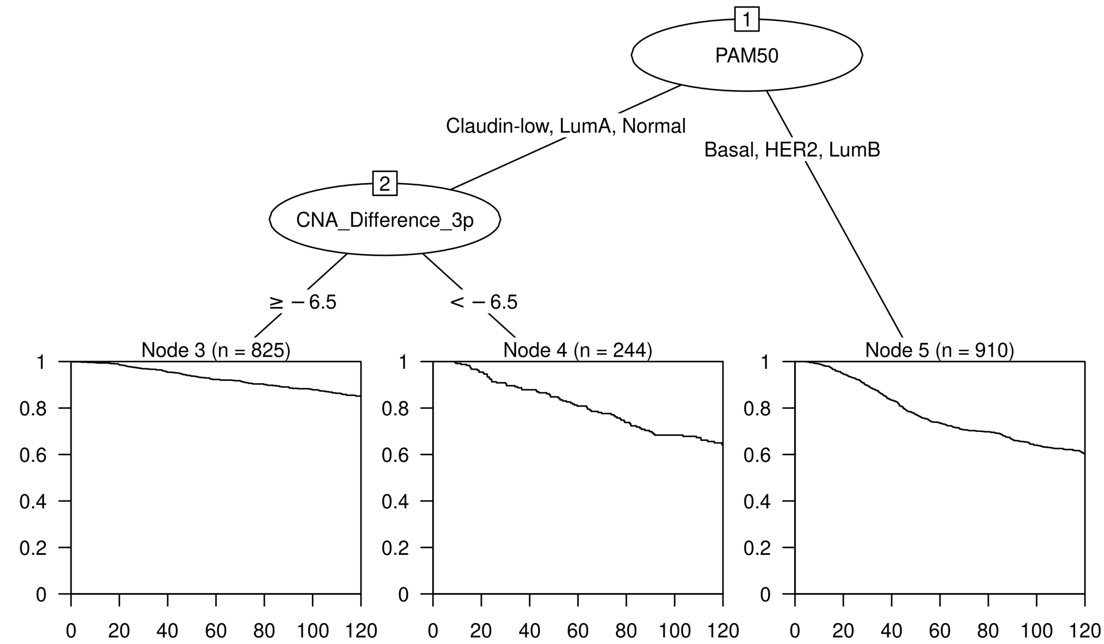


Figure 3.29: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

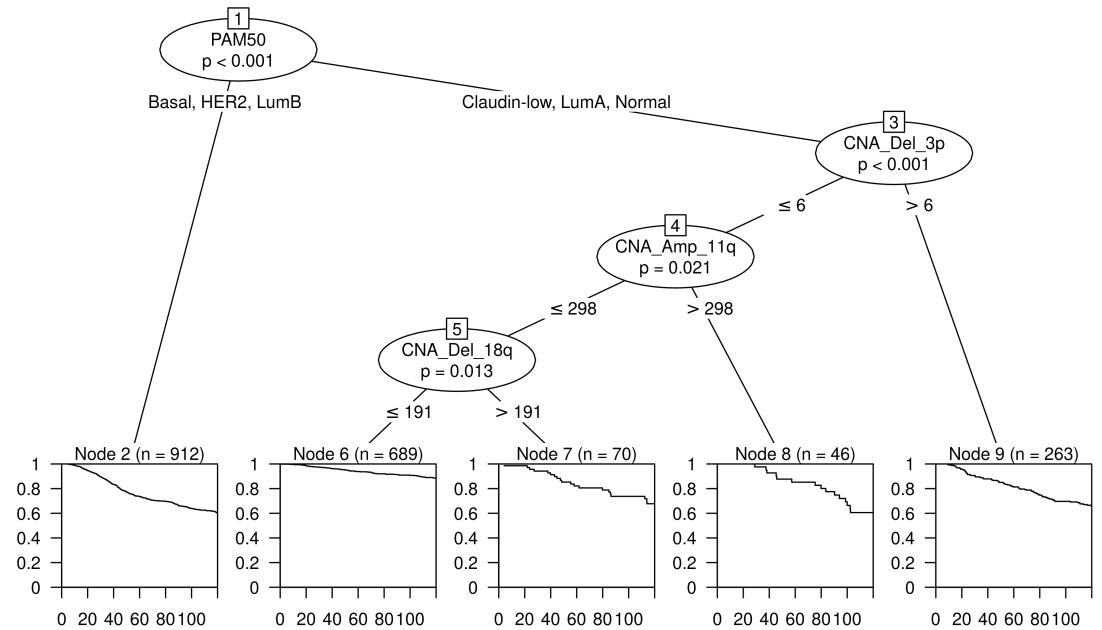
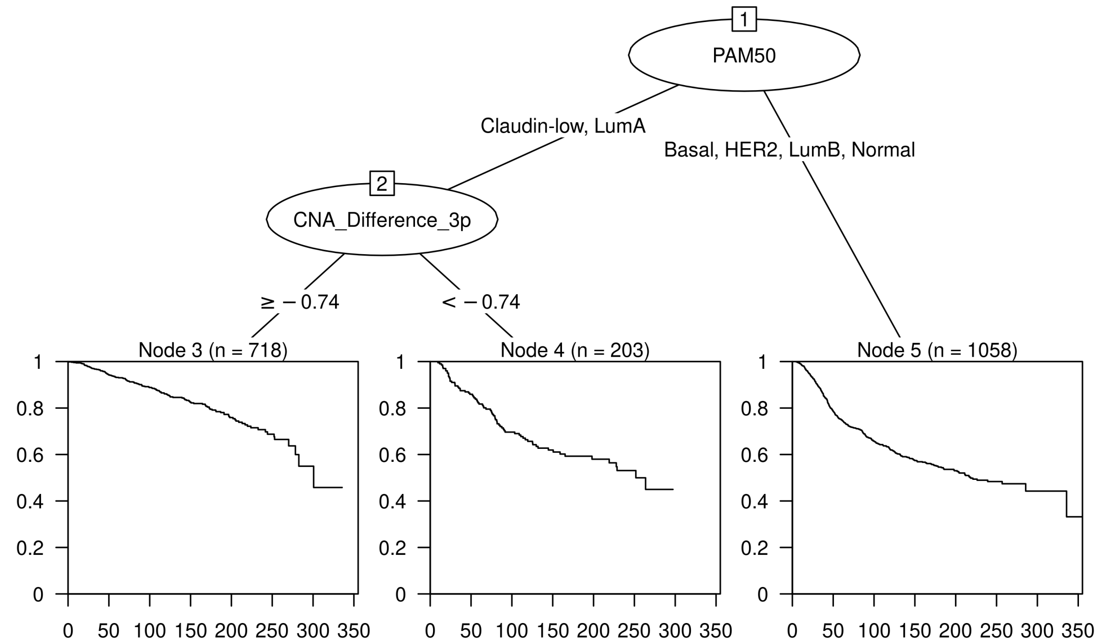


Figure 3.30: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

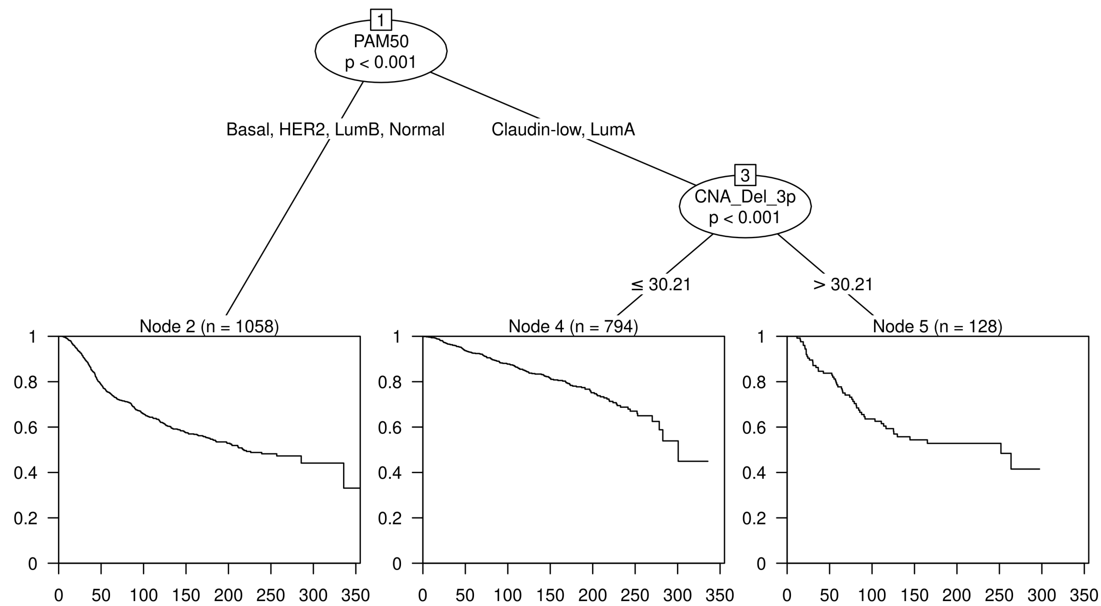
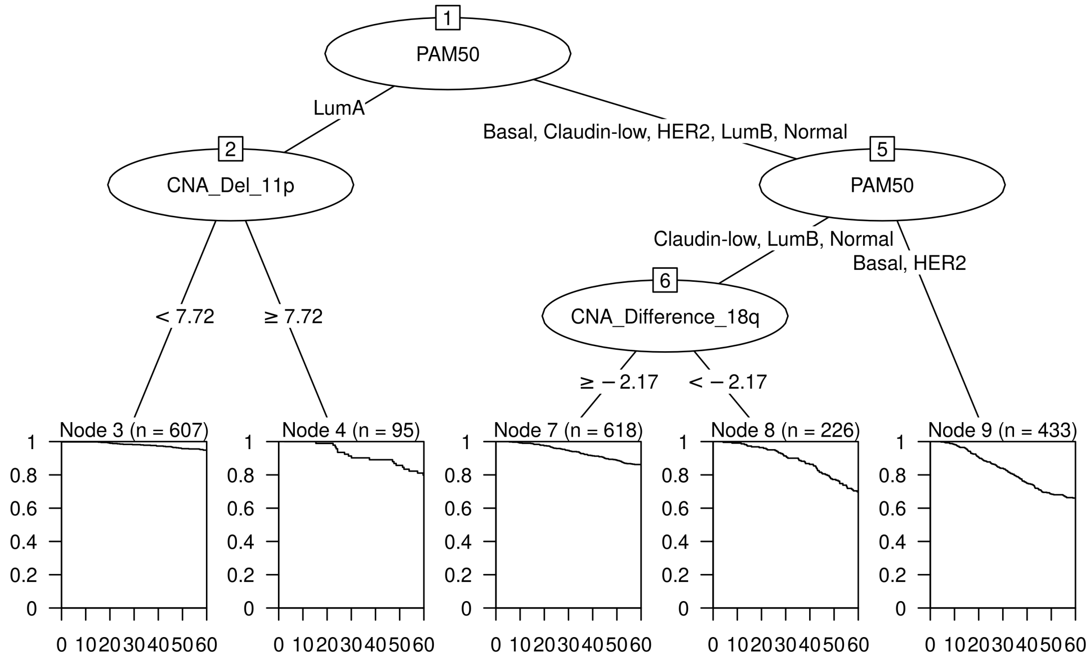


Figure 3.31: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

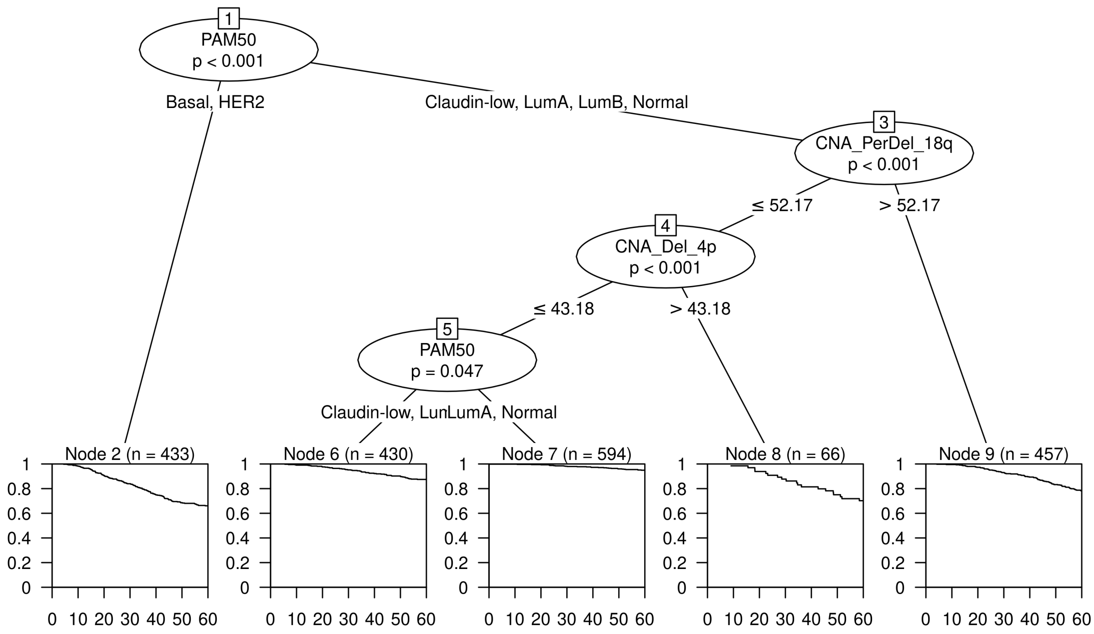
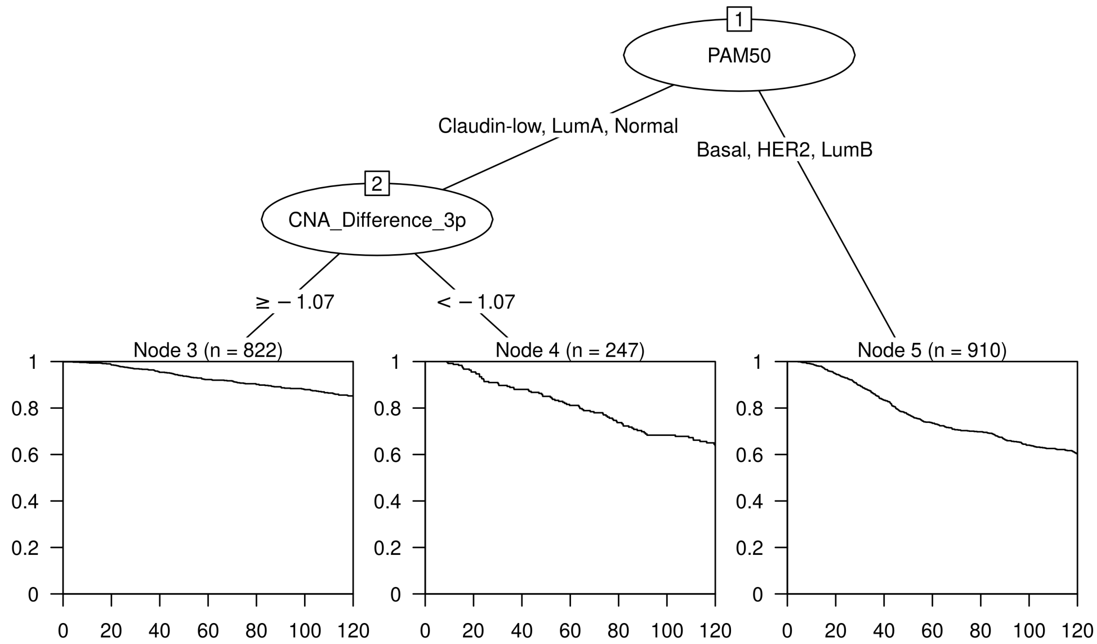


Figure 3.32: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

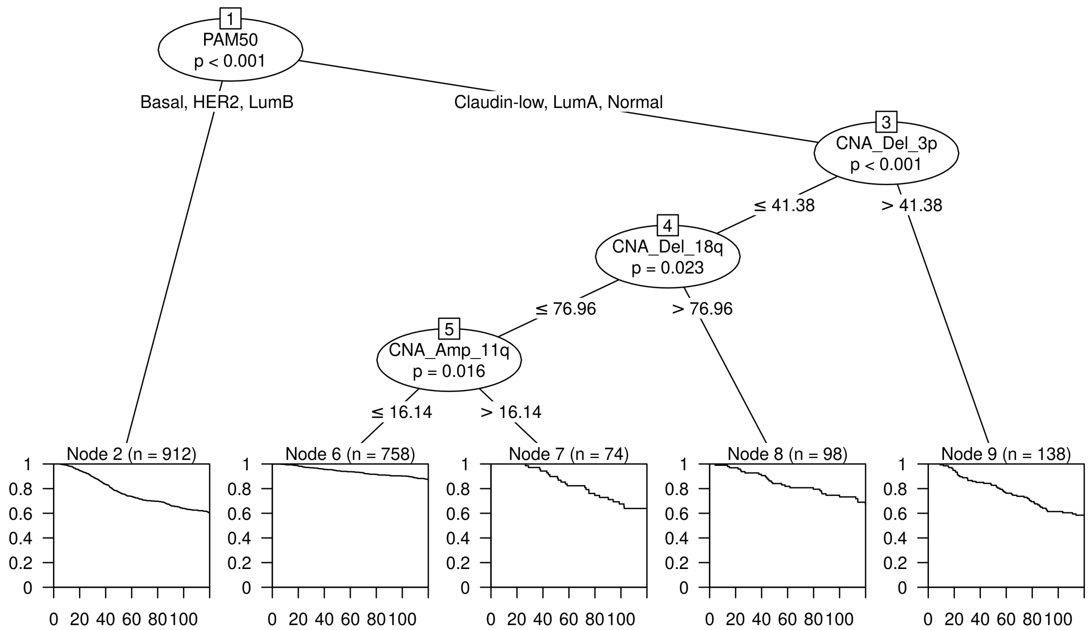
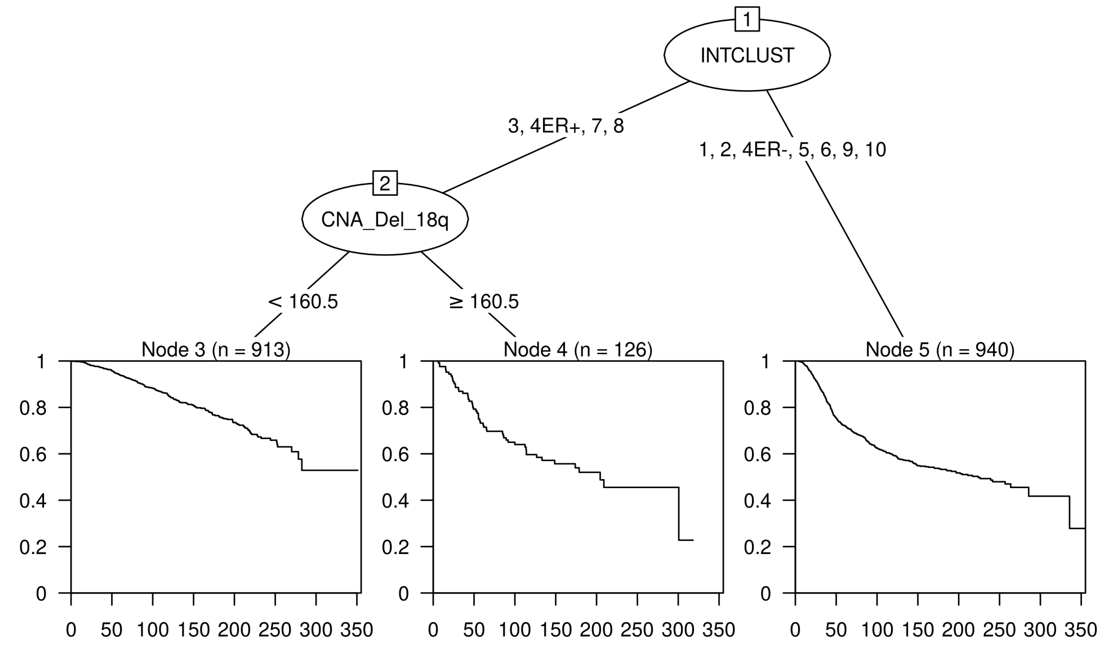


Figure 3.33: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

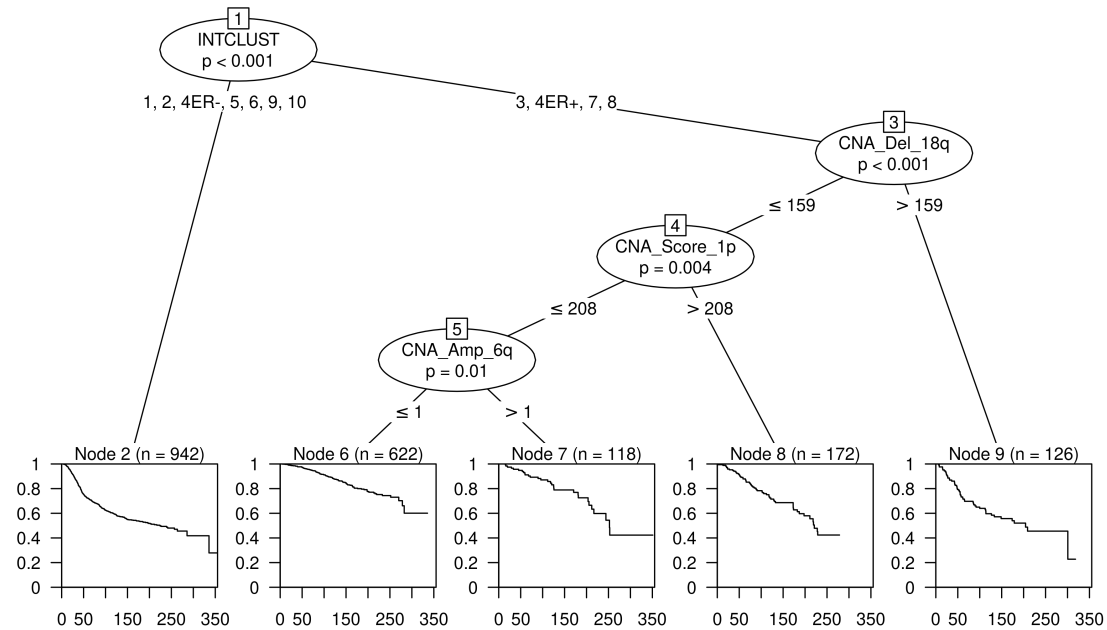
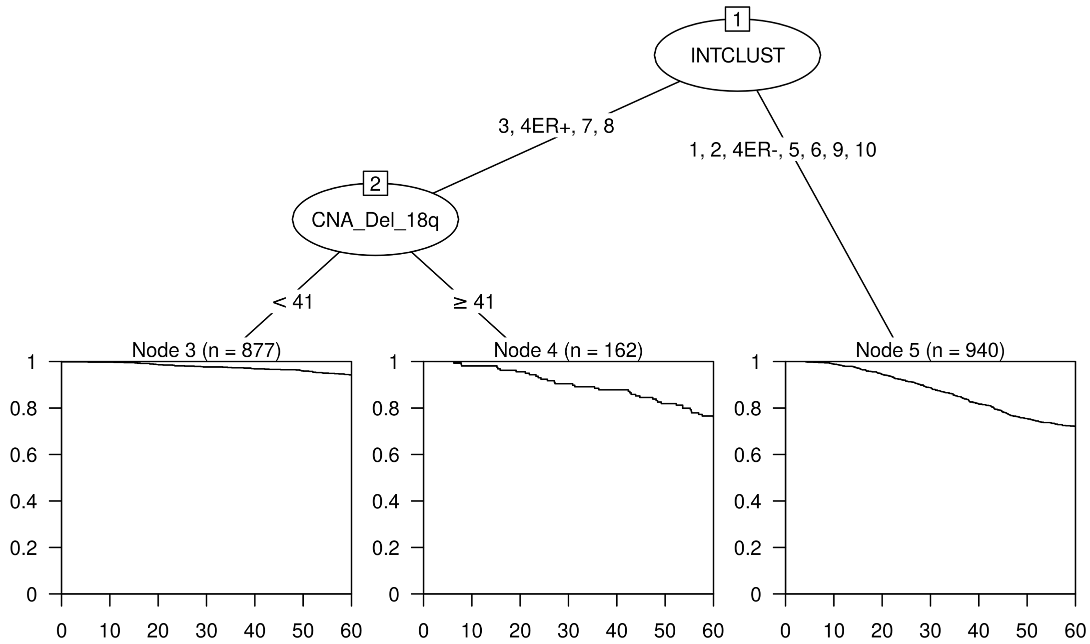


Figure 3.34: Recursive partitioning survival trees for disease-specific survival using Integrative Cluster and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

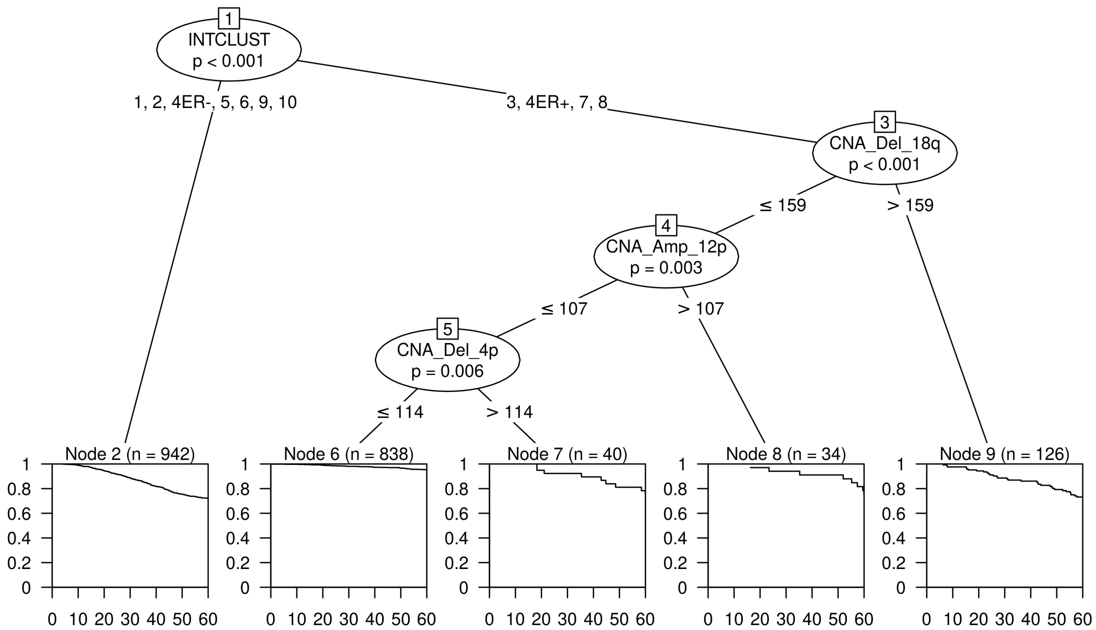
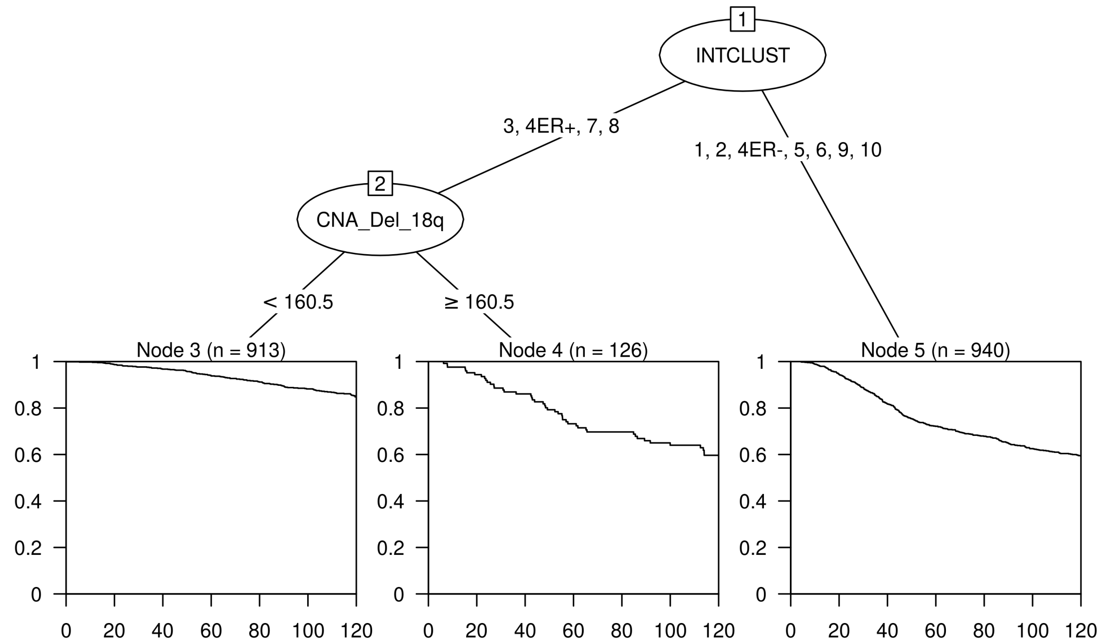


Figure 3.35: Recursive partitioning survival trees for five-year disease-specific survival using Integrative Cluster and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

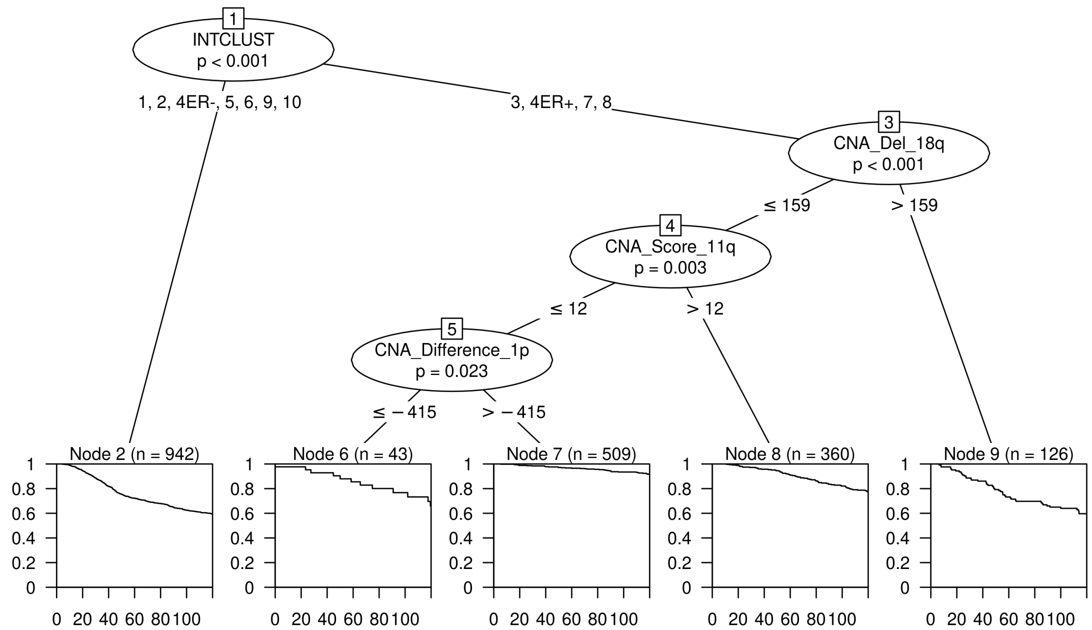


Figure 3.36: Recursive partitioning survival trees for five-year disease-specific survival using Integrative Cluster and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

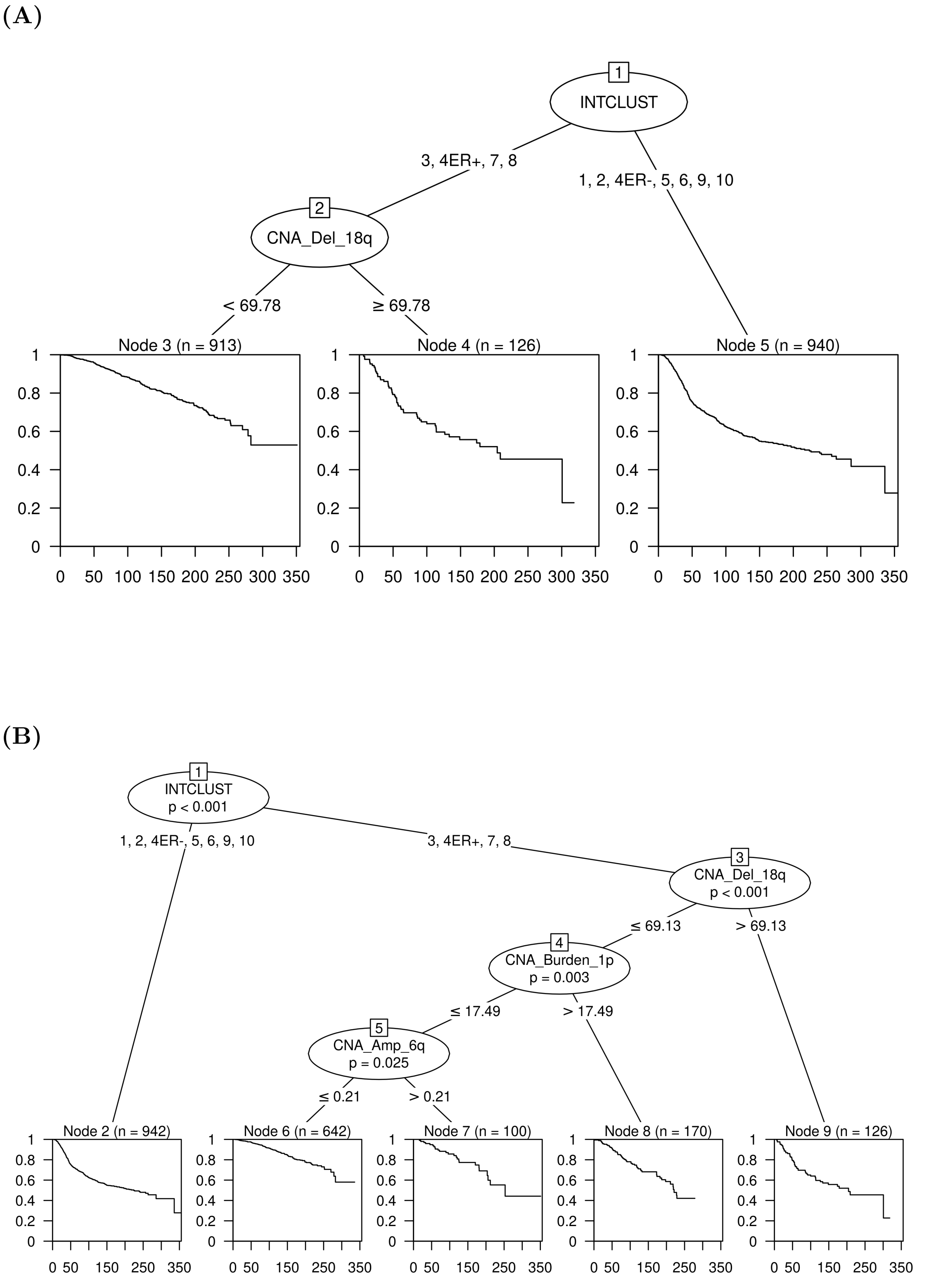
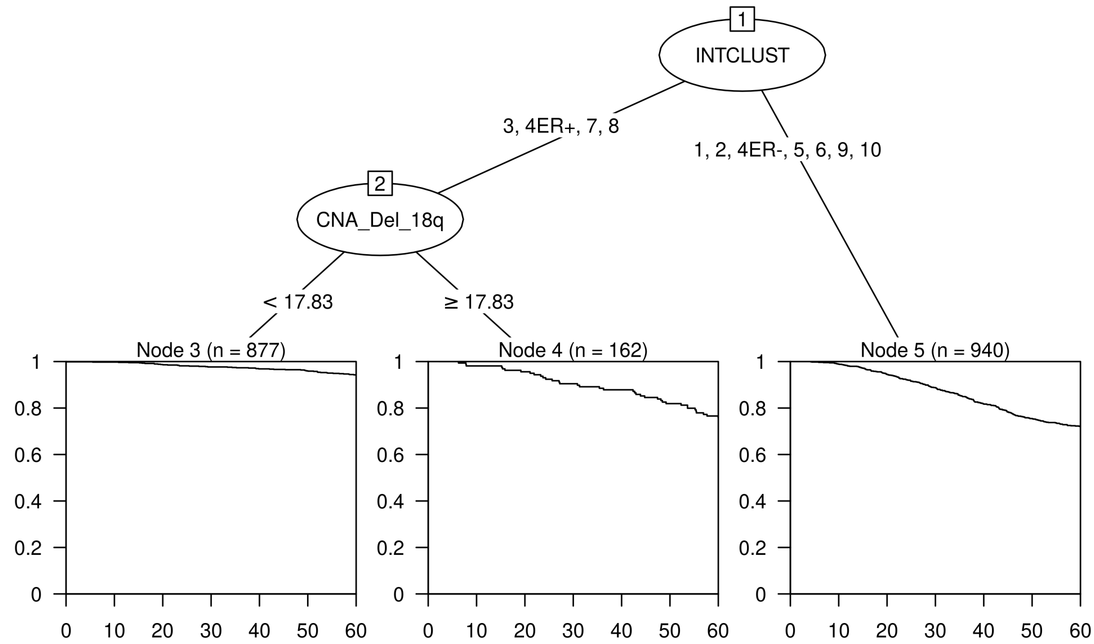


Figure 3.37: Recursive partitioning survival trees for disease-specific survival using Integrative Cluster and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

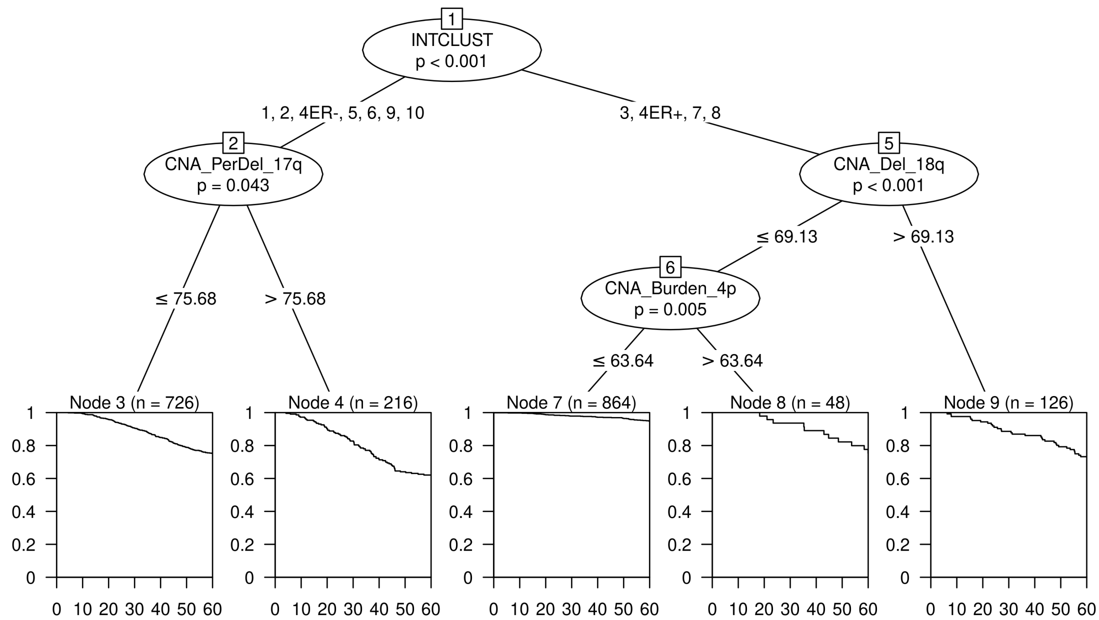
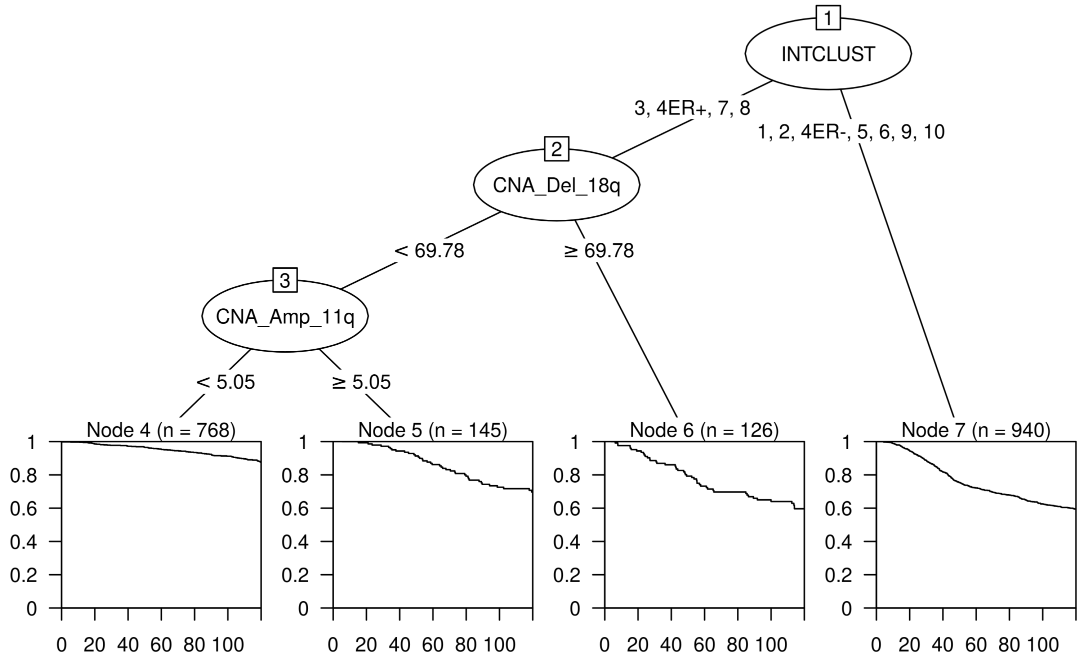


Figure 3.38: Recursive partitioning survival trees for five-year disease-specific survival using Integrative Cluster and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

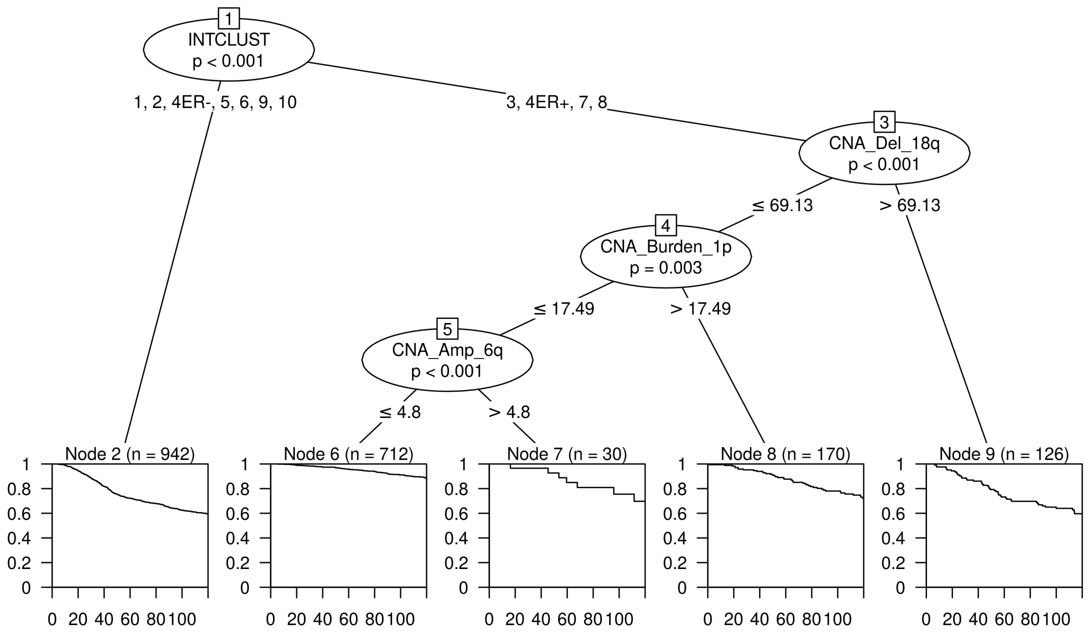
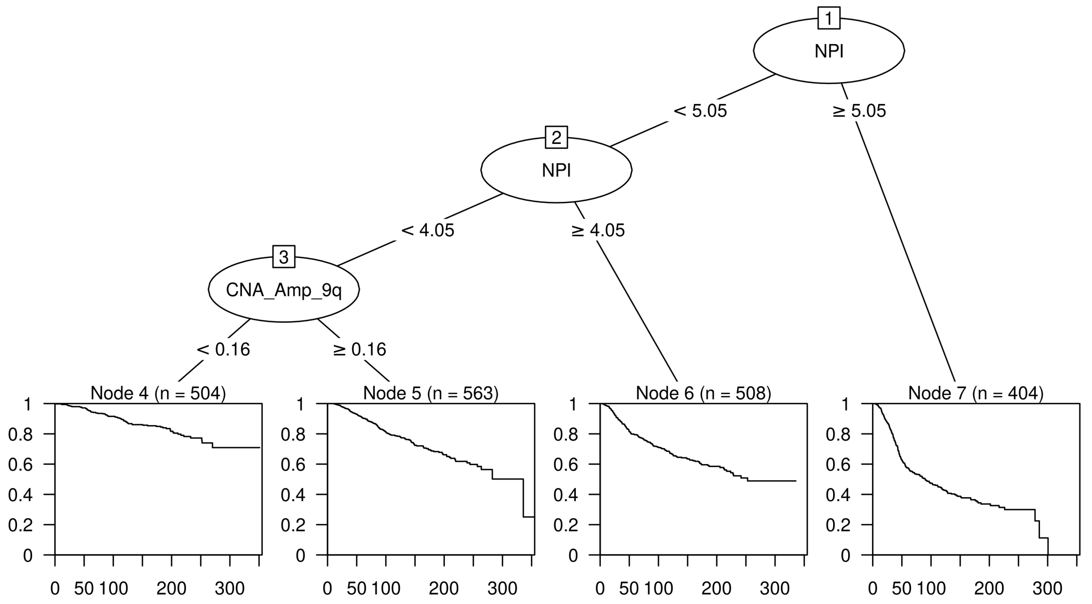


Figure 3.39: Recursive partitioning survival trees for five-year disease-specific survival using Integrative Cluster and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

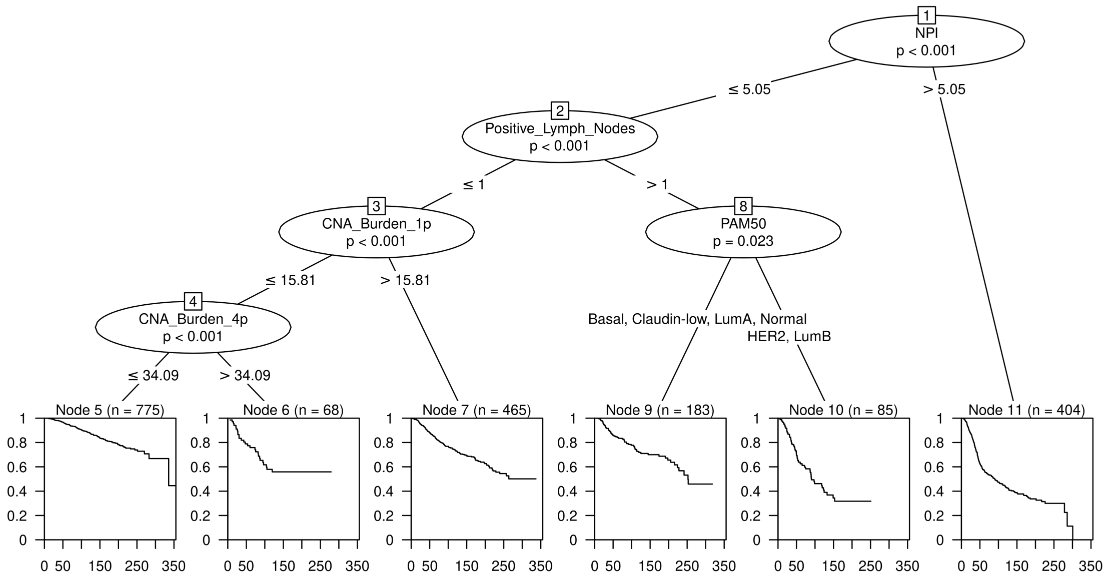
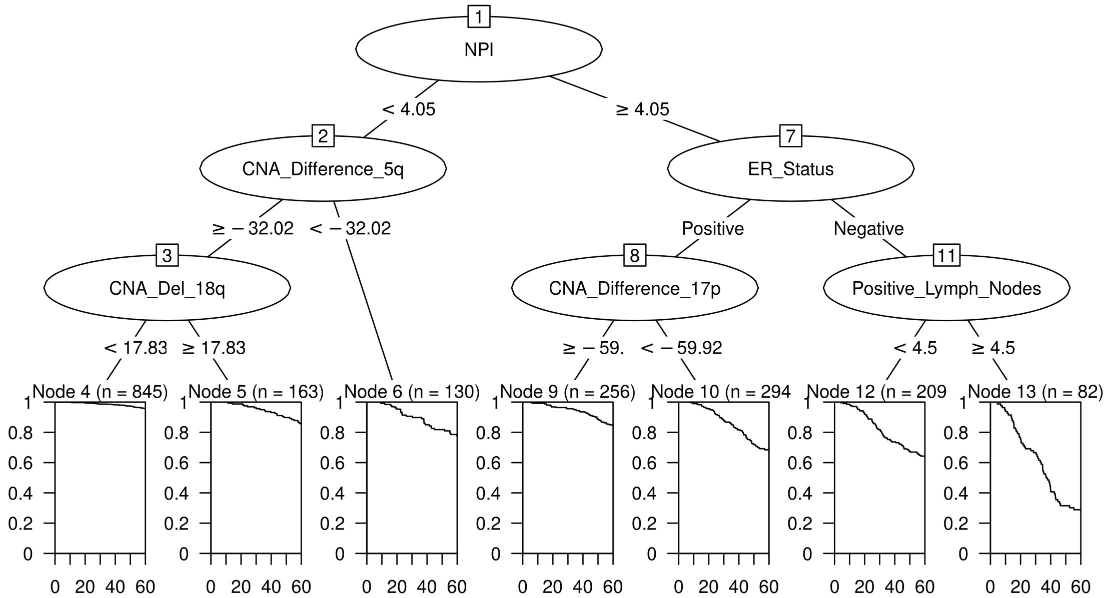


Figure 3.40: Recursive partitioning survival trees for disease-specific survival using PAM50 subtype, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

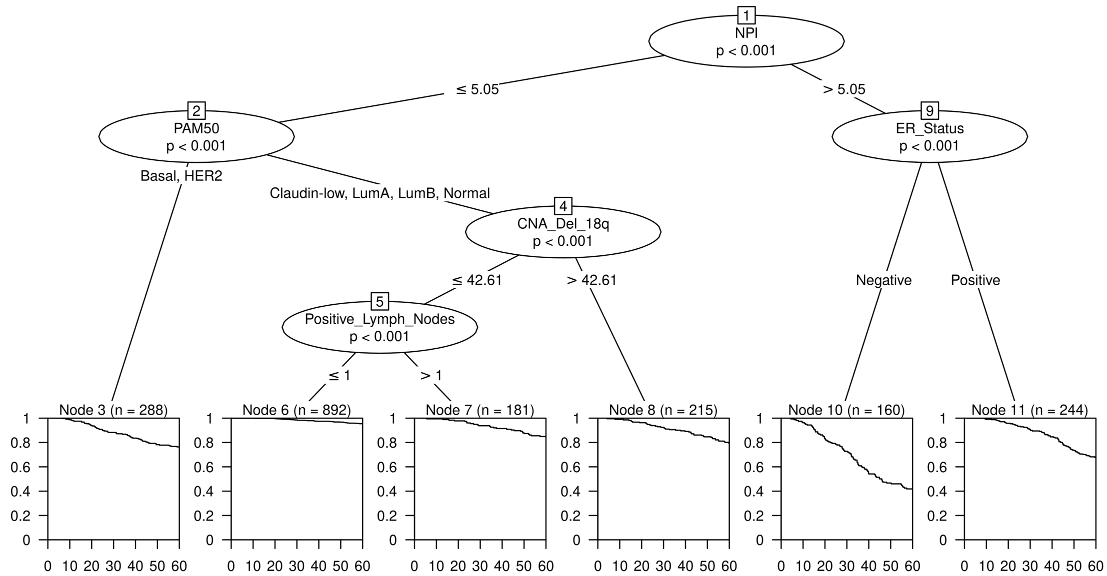


Figure 3.41: Recursive partitioning survival trees for five-year disease-specific survival using PAM50 subtype, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)

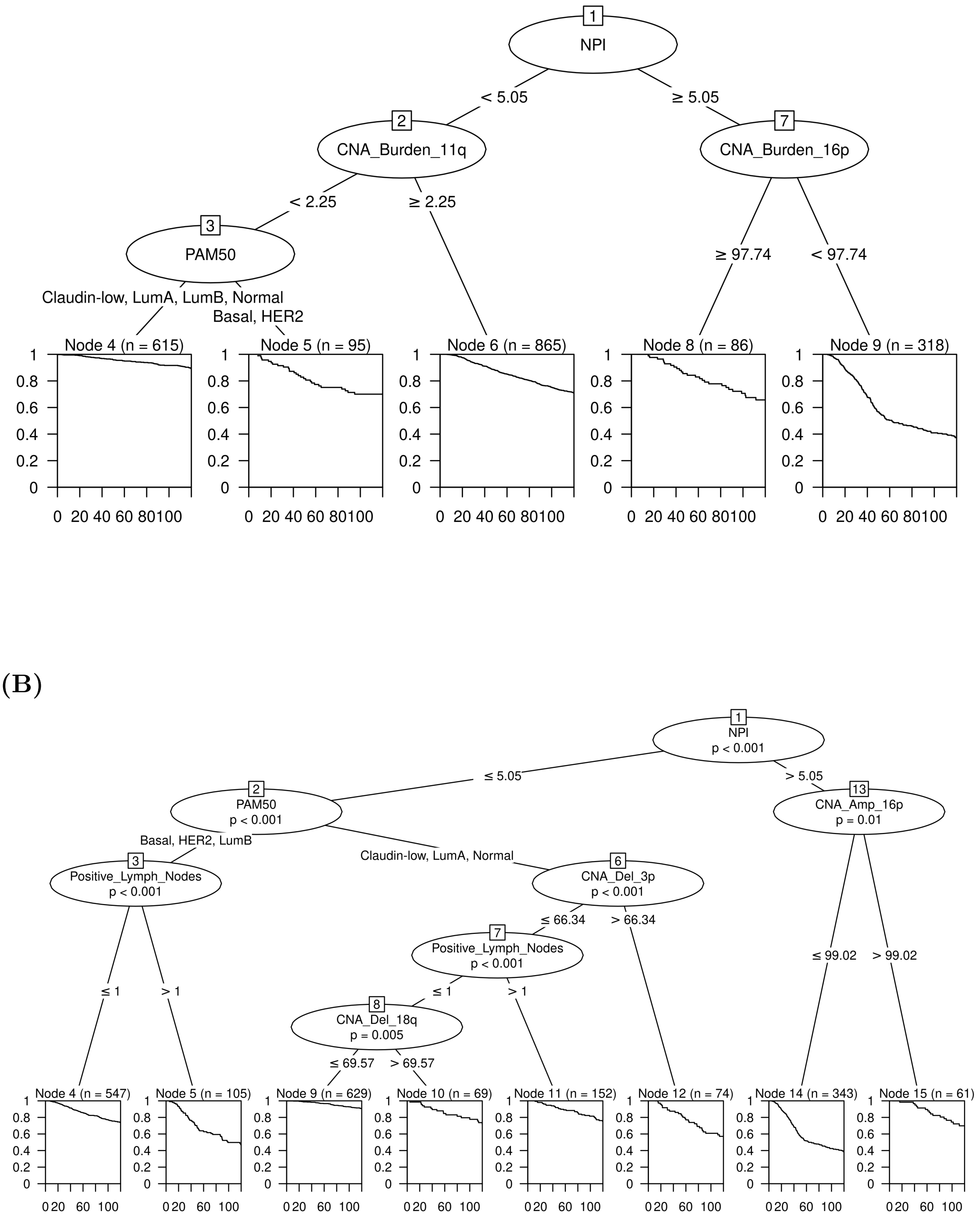
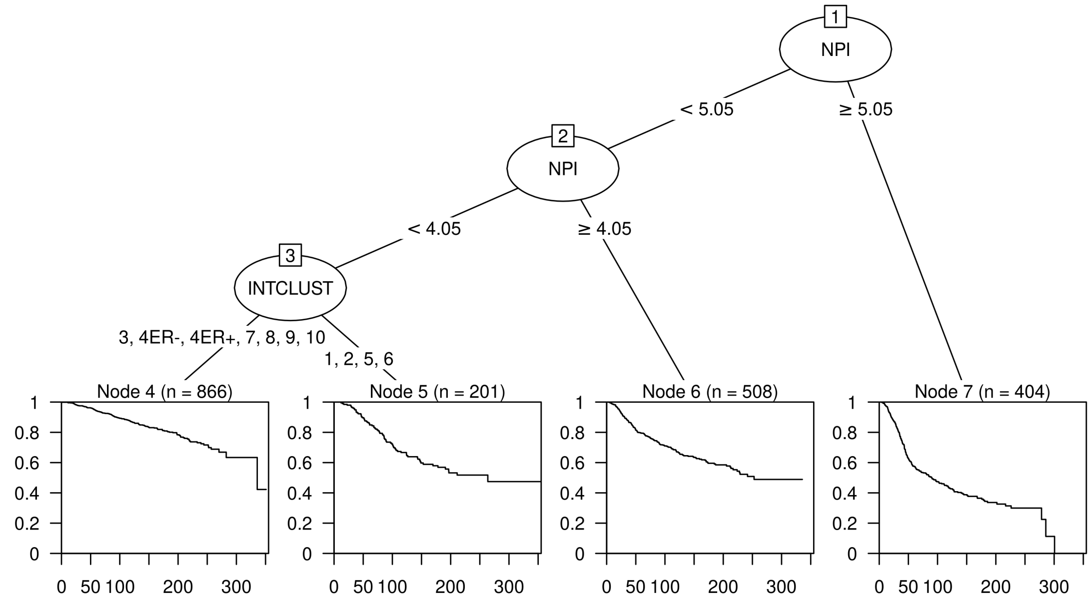


Figure 3.42: Recursive partitioning survival trees for ten-year disease-specific survival using PAM50 subtype, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

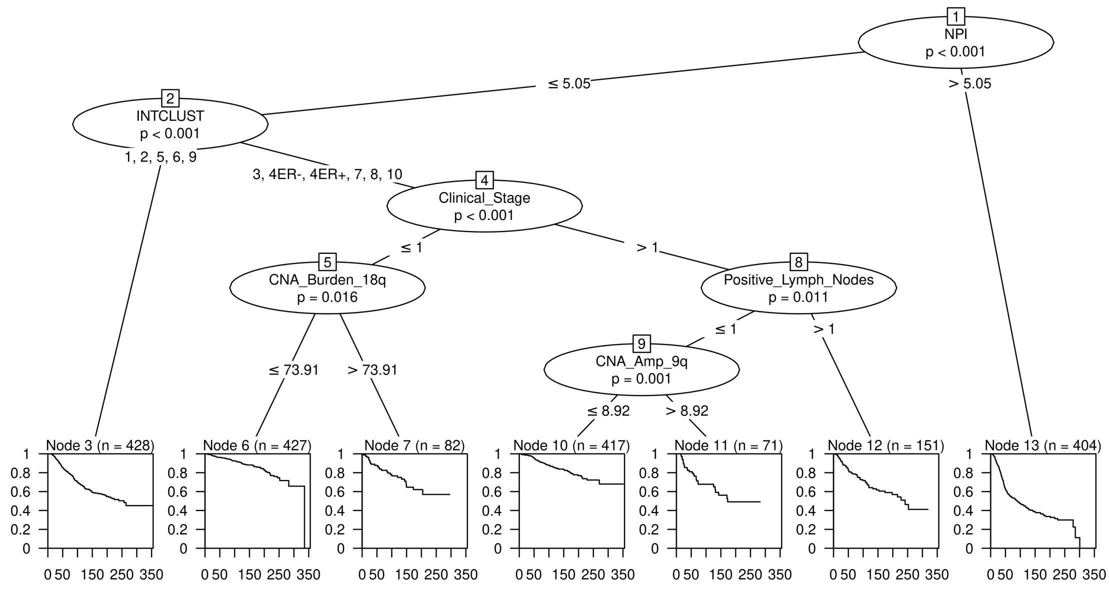
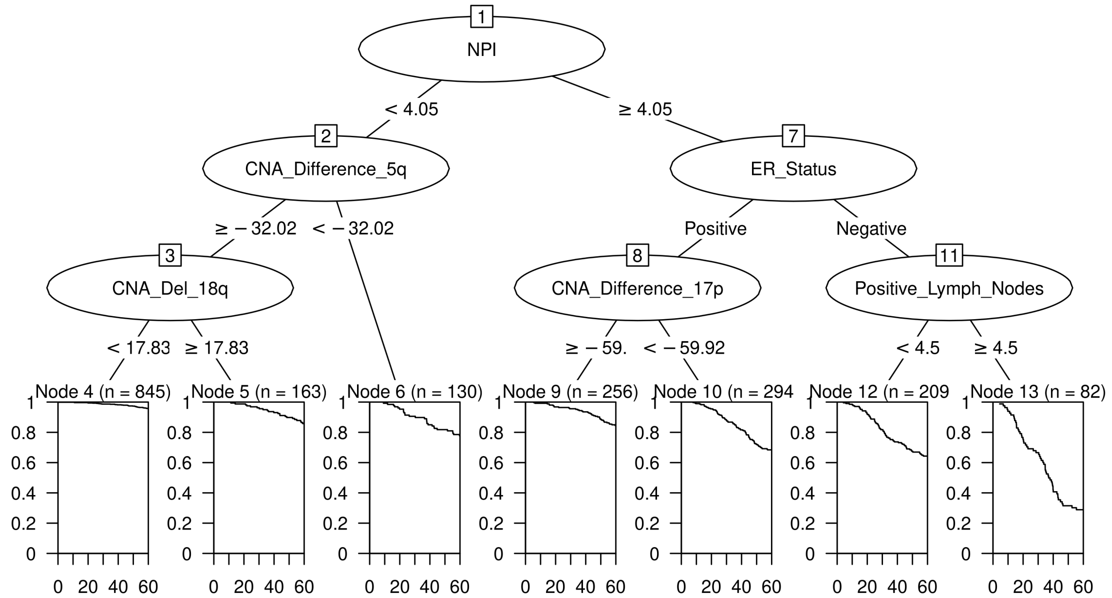


Figure 3.43: Recursive partitioning survival trees for disease-specific survival using IntClust, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

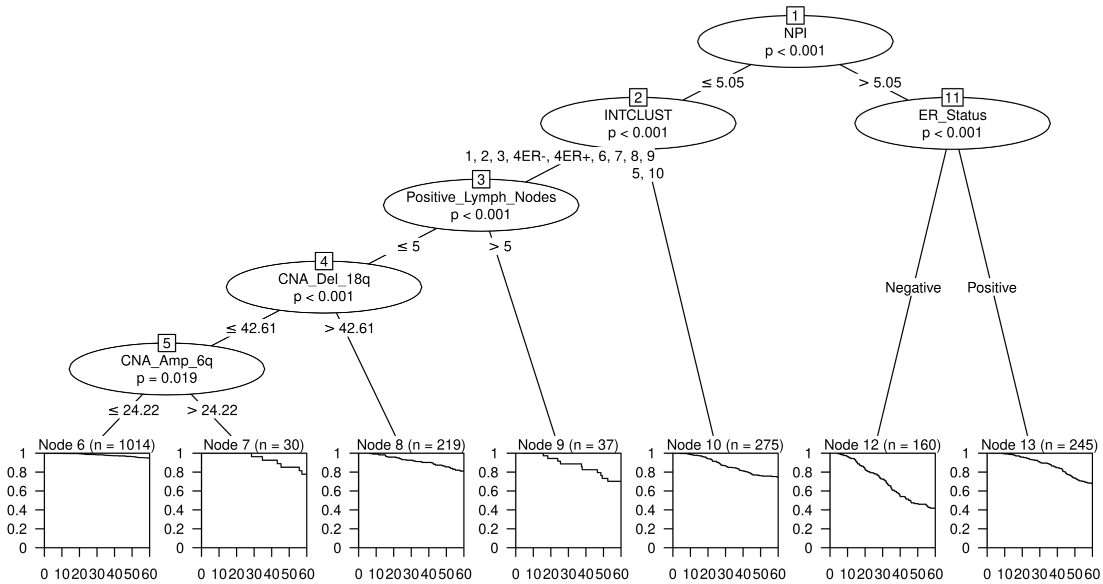
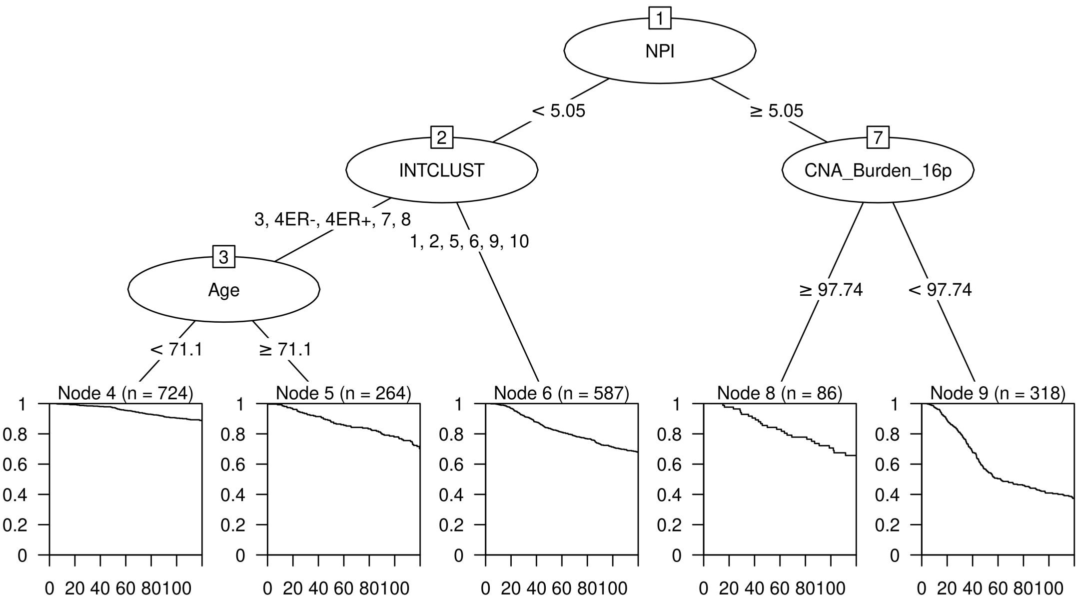


Figure 3.44: Recursive partitioning survival trees for five-year disease-specific survival using IntClust, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

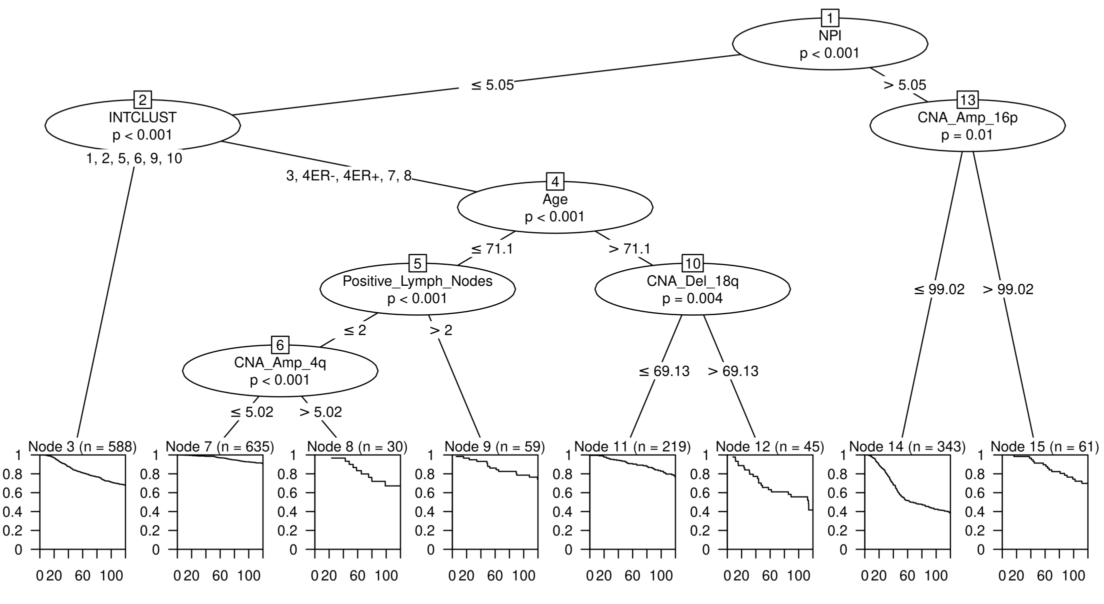


Figure 3.45: Recursive partitioning survival trees for ten-year disease-specific survival using IntClust, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

IntClust, rather than PAM50 subtype. The clinical variables used to partition the patients include combinations of NPI, number of positive lymph nodes, age, clinical stage and ER Status, while the chromosome arm metrics used to partition the data include CNA Burden on 18q, CNA Amp Burden on 9q, CNA Difference Burden on 5q, CNA Del Burden on 18q, CNA Difference on 17p, CNA Amp Burden on 6q, CNA Burden on 16p, CNA Amp Burden on chromosome 4q and CNA Amp Burden on 16p. Noticeably CNA Del Burden on chromosome 3p is absent from the survival trees considering IntClust molecular classification and initially partitioning on NPI alters the IntClust partition from consistently grouping IntClust 3, 4ER+, 7 and 8 together to a range of partitions.

Interestingly, the chromosome arm CNA Burden metrics appear as useful predictors in partitions of PAM50 subtypes and IntClusts, but also in trees and partitions where these molecular classifications were not identified as significant predictors. This indicates that the CNA Burden metrics can provide additional survival information in groups of patients split on molecular classifications and groups of patients who are not.

3.4.3 Heatmaps of CNA State across Selected Chromosome Arms

Heatmaps of the CNA landscape of chromosome 3p, chromosome 18q and chromosome 11q, with patients partitioned into nodes corresponding to Figures 3.31B (ctree), 3.37A (rpart) and 3.32A (rpart), respectively, are produced. These heatmaps provide detail of the CNA state for each of the 609, 230 and 492 genes recorded on chromosomes 3p, 18q and 11p.

Figure 3.46, the heatmap of CNAs across chromosome 3p, shows that the Claudin-low and Luminal A patients corresponding to Node 5 have high levels of deletions across the majority of chromosome 3p, Node 4, also containing Claudin-low and Luminal A patients, consists of a small proportion of patients with high levels of amplification with the remainder being relatively stable. Node 2, containing Luminal B, HER2, Normal and Basal patients, consists of patients displaying variation in levels of GI across chromosome 3p. Figure 3.47, the heatmap of CNA calls across chromosome 18q, displays a similar pattern, where IntClust 3, 4ER+, 7 and 8 patients corresponding to Node 4 have high levels of deletions across the majority of chromosome 18q. Node 3 consists of a small proportion of IntClust 3, 4ER+, 7 and 8 patients with high levels of amplification with the remainder being relatively stable. Node 5 consists of IntClust 1, 2, 4ER-, 5, 6, 9 and 10 patients displaying variation in levels of GI across chromosome 18q. Figure 3.48 displays the heatmap of CNAs across chromosome 11p. Focusing on the nodes corresponding to Luminal A patients, Nodes 3 and 4, it is observed that patients in Node 4, with worse survival outcomes, have high levels of deletions across the majority of chromosome 11p.

There are two important aspects of the data used to produce these heatmaps, including that the CNAs are only recorded for annotated genes and that the data used are total CNA data meaning it is not possible to determine whether the hemizygous CNAs observed across the chromosome arms are occurring contiguously on one homologous chromosome or if the CNAs are occurring randomly across the two. This aspect is discussed further in the upcoming chapters.

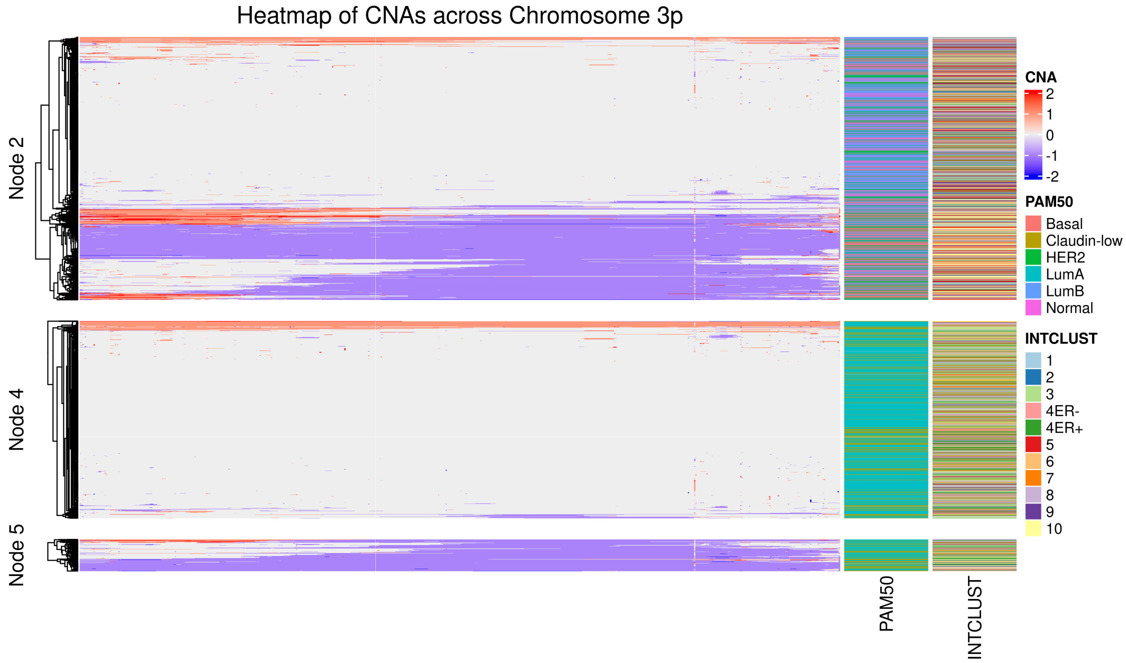


Figure 3.46: Heatmap of CNAs across Chromosome 3p. The heatmap depicts the CNA state for each gene across Chromosome 3p, partitioning the patients into the nodes corresponding to Figure 3.31B.

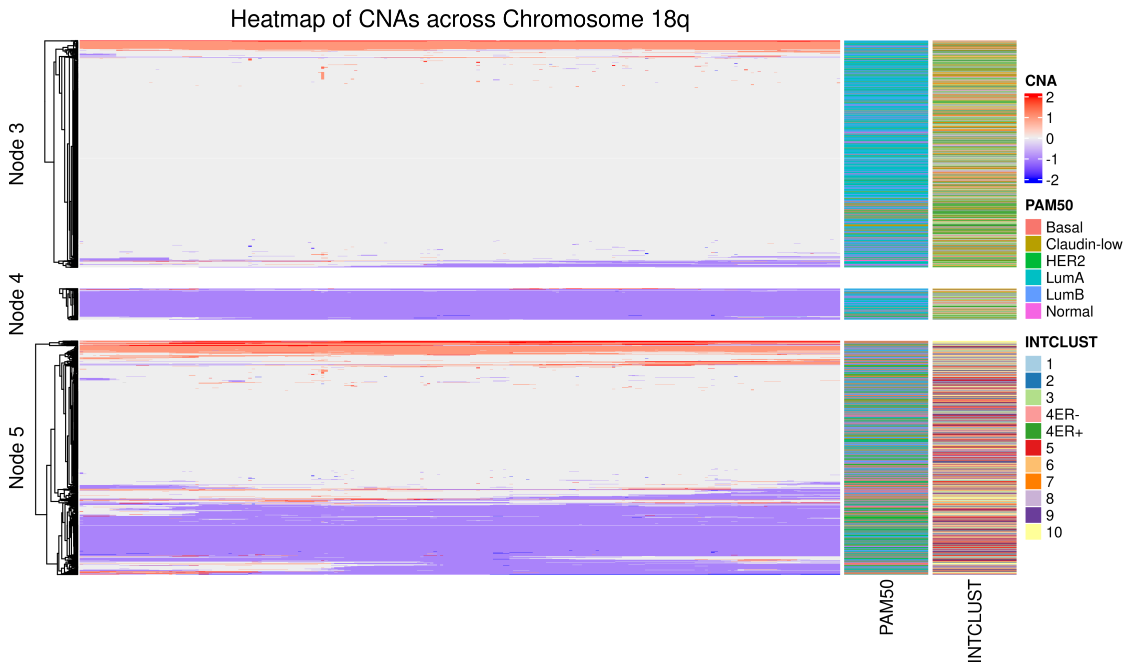


Figure 3.47: Heatmap of CNAs across Chromosome 18q. The heatmap depicts the CNA state for each gene across Chromosome 18q, partitioning the patients into the nodes corresponding to Figure 3.37A.

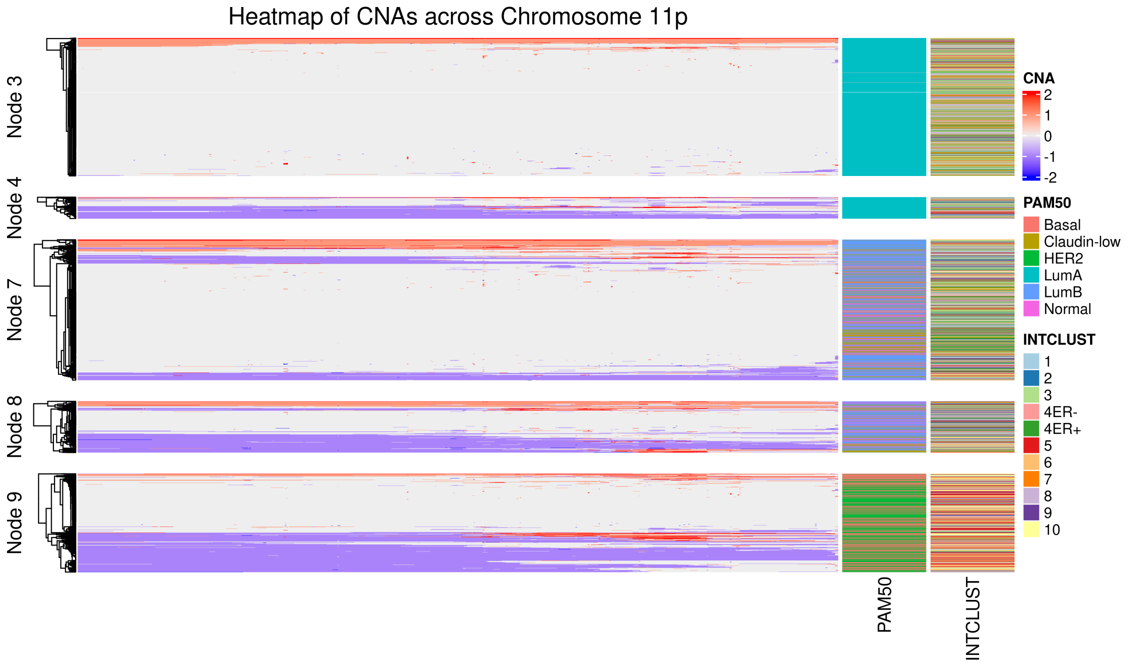


Figure 3.48: Heatmap of CNAs across Chromosome 11p. The heatmap depicts the CNA state for each gene across Chromosome 11p, partitioning the patients into the nodes corresponding to Figure 3.32A.

3.5 GNOSIS: an R Shiny app supporting cancer genomics survival analysis with cBioPortal

As shown above, and in previous chapters, exploratory, statistical and survival analysis of cancer genomic data is extremely important and can lead to new discoveries, such as the identification of novel genomic prognostic markers, that have the potential to advance our understanding of cancer and ultimately benefit patients. These analyses are often performed on data available from a number of consortia websites, such cBioPortal (Cerami et al., 2012; Gao et al., 2013), which is one of the best known and commonly used consolidated curations that hosts data from large consortium efforts. While cBioPortal provides both graphical user interface (GUI)-based and representational state transfer mediated means for researchers to explore and analyse clinical and genomics data, its capabilities have their limitations and oftentimes, to explore specific hypotheses, users need to perform a more sophisticated 'off site' analysis that typically requires users to have some prior programming experience.

To overcome these limitations and provide a GUI that facilitates the visualisation and interrogation of cancer genomics data, particularly cBioPortal-hosted data, using standard biostatistical methodologies, we developed an R Shiny app called GeNomics explOrer using StatistIcal and Survival analysis in R (GNOSIS). GNOSIS was initially developed as part of our study, using the METABRIC data, to investigate whether survival outcomes are associated with genomic instability in Luminal breast cancers (King et al., 2021a) and was further developed to enable the exploration, analysis and incorporation of a diverse range of genomic features with clinical data in a research or clinical setting.

GNOSIS leverages a number of R packages and provides an intuitive GUI with multiple tab panels supporting a range of functionalities, including data upload and initial exploration, data recoding and subsetting, data visualisations, statistical analysis, mutation analysis and, in particular, survival analysis to identify prognostic markers. In addition, GNOSIS also helps researchers carry out reproducible research by providing downloadable input logs (Shiny_Log.txt) and R scripts (ggcode.zip) from each session.

3.5.1 Layout and Functionality

The current version of GNOSIS has 11 tabs and allows users to carry out a comprehensive visual exploration, statistically robust survival analysis and mutation analysis in a simple, efficient and reproducible way. GNOSIS installs and loads up a number of R packages, primarily shiny, tidyverse, ggplot2, survival, survminer, rpart, partykit and maftools (Hothorn et al., 2006; Hothorn and Zeileis, 2015; Wickham, 2016; Mayakonda et al., 2018; Wickham et al., 2019; Kassambara et al., 2021; Therneau and Atkinson, 2022; Chang et al., 2022; Therneau, 2023) (see full list in Section 3.5.2), and provides users with tabs for data upload, exploration, data subsetting and recoding, visualisations, comprehensive survival analysis, association testing and mutation analysis. In addition, GNOSIS records all user activity and provides downloadable .txt files and R scripts to facilitate reproducibility. Figure 3.49 shows the GNOSIS front-end with the specific entry points and 'tabs' marked in red.

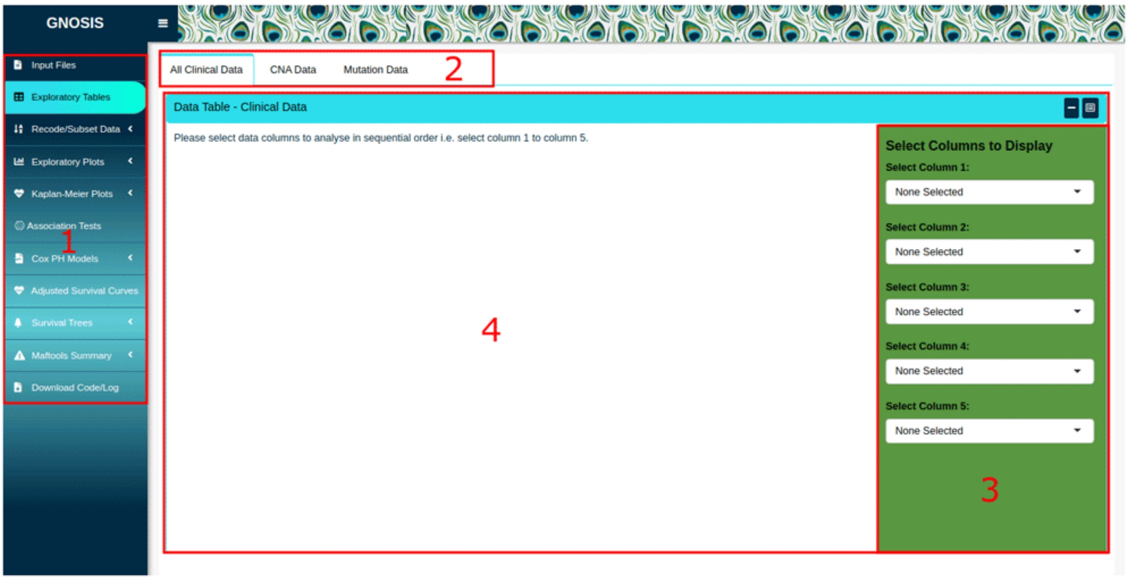


Figure 3.49: GNOSIS GUI with highlighted interface elements. (1) The Exploratory Tables tab is selected in the tab sidebar. (2) Within tab panels allowing multiple operations to be carried out and viewed in the one tab. (3) Box sidebar allowing users to select inputs, alter arguments and customise and export visualisations. (4) Viewing panel displaying output.

3.5.1.1 Data upload and formatting (Tab 1-Tab 3)

Users can upload comma-, semicolon- or tab-delimited files containing clinical, summary CNA and/or mutation data using the Input Files tab. In addition to providing users with a space to upload their data of interest, the Input Files tab also provides users with a preview of the data to ensure that the data has been read in correctly. Although GNOSIS was built using data downloaded from cBioPortal, and so the default settings are suited to these file types, users can upload clinical or summary genomics data files from other sources. In the case where users are uploading non-cBioPortal data, care should be taken to set appropriate default values and that the uploaded data contains the columns required by GNOSIS. More specifically, the clinical patient and sample data should contain a column named "PATIENT_ID" and the CNA data should contain a column called "Hugo_Symbol". As these are core named data types for all subsequent analytics, warning messages will be produced, and downstream analysis will not be possible if they are missing.

After the data are successfully uploaded and previewed, further exploration of selected columns can be done using the Exploratory Tables tab. In this tab, users can select and view up to five columns in each file uploaded. It should be noted that the columns should be selected in sequential order; if this is not adhered to an error message will be displayed.

After data upload and initial exploration, and before more extensive data analysis, users are encouraged to carry out data pre-processing or cleaning. This ensures that the data is in the desired format for downstream analysis. The Recode/Subset Data tab enables users to pre-process the clinical data by providing information on the variables present in the data, their type and factor levels and by allowing users to change selected variables to numeric or factors, subset the data based on several categorical variables, and carry out survival variable recoding. Where CNA data is uploaded, users may generate and segment a number of CNA metrics for each patient (Absolute CNA Score, CNA Amp Score and CNA Del Score), as well as select and extract specific genes for further analysis. The GUI is updated with the changes in real time, meaning that users can check that their alterations have been implemented correctly. Users also have the option to save the formatted dataframe for future use. Figure 3.50 shows examples of how the uploaded clinical data can be examined, and filters applied to extract a subset and Figure 3.51 shows the resulting subset following calculation of global CNA metrics and subsequent quartile segmentation of the Absolute CNA Scores.

3.5.1.2 Data visualisation (Tab 4)

After initial data upload, basic exploration and data cleaning, the Exploratory Plots tab can be used to produce a range of visualisations including boxplots, scatterplots, barplots, histograms and density plots. These visualisations are generated using the ggplot2 R package (Wickham, 2016). For all visualisations, users can use the box sidebar to make a number of selections, including selecting which variables to interrogate, choosing whether to include or omit NA values, choosing whether to display a legend, choosing the legend position and changing the plot title, x- and y-axis titles and legend titles, among others. There are also a number of plot-specific options available to users including the ability to produce boxplots where the sample size is reflected in the width of the boxplot, the ability to produce scatterplots where the points are coloured by an additional variable, and the ability to produce plain, segmented and faceted histograms and density plots. For example, Figure 3.52 displays a segmented density plot of the absolute CNA scores. As GNOSIS aims to aid users in their research, facilitate reproducibility and support users in developing their programming skills, all visualisations and the R code to produce them can be downloaded as .pngs or .svgs in specified dimensions and R scripts respectively.

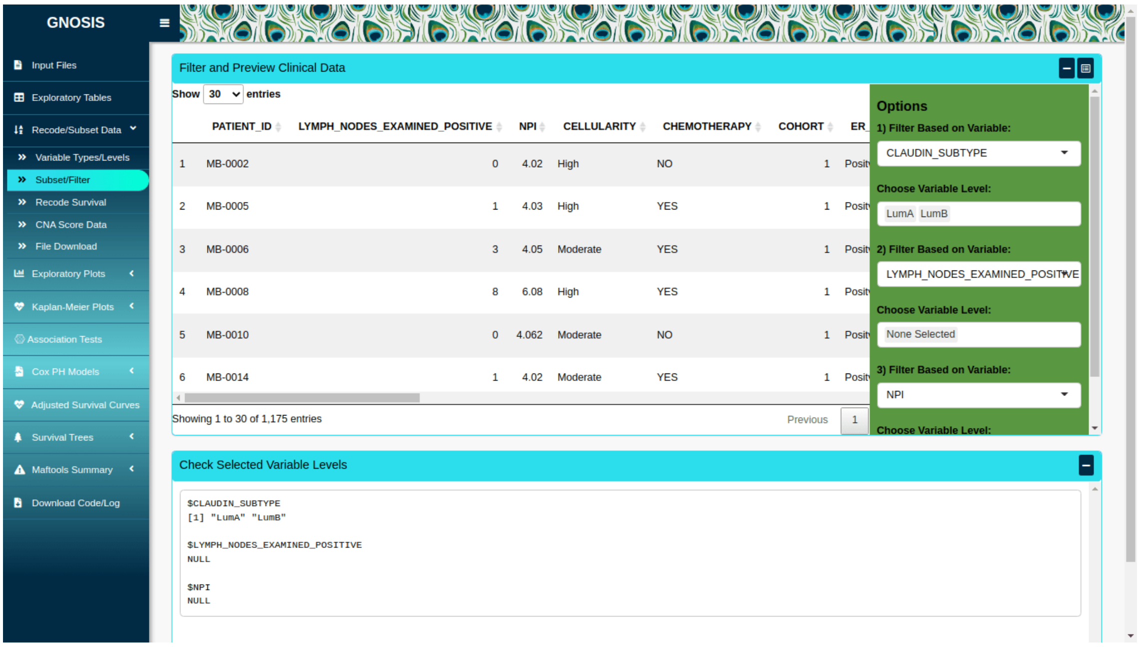


Figure 3.50: The Recode/Subset tab. Data is being subsetted based on PAM50 subtype with Luminal A and Luminal B subtypes selected.

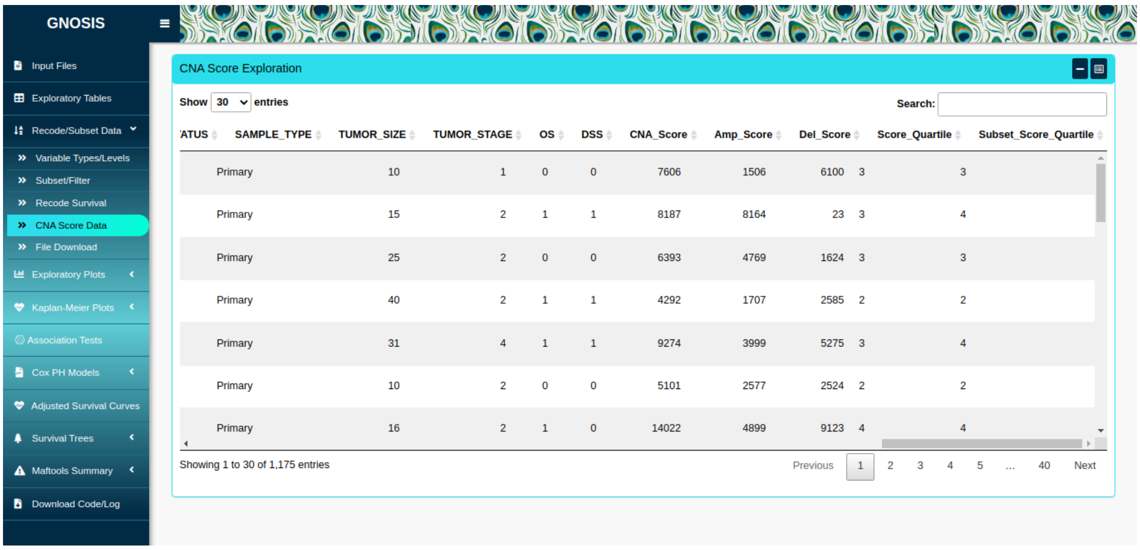


Figure 3.51: The dataset after CNA metrics have been calculated and quartile segmented.

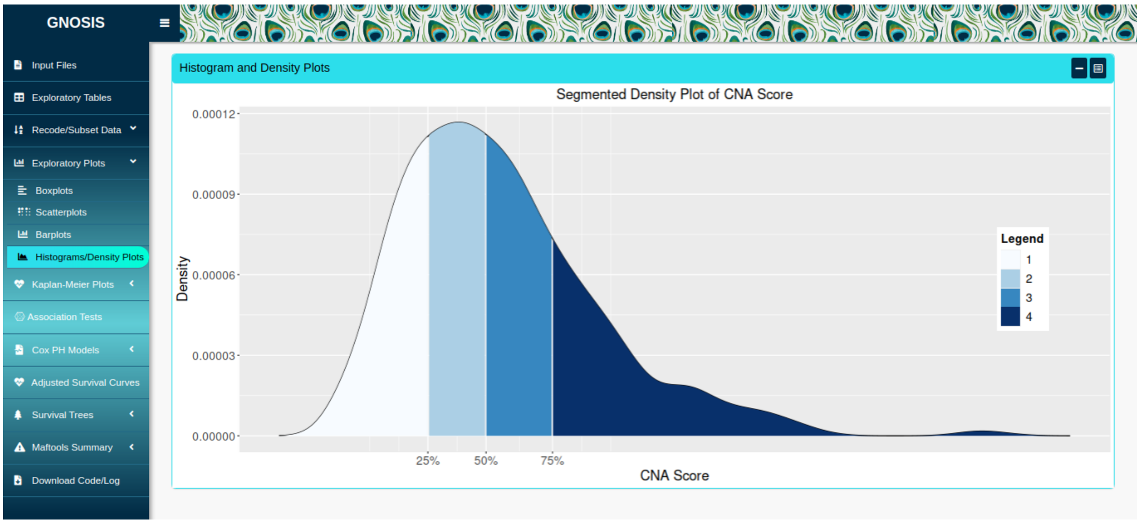


Figure 3.52: A density plot of the resulting quartile segmentation.

3.5.1.3 Statistical and survival analysis (Tab 5 - Tab 9)

The primary function offered by GNOSIS is comprehensive statistical and survival analysis. GNOSIS provides users with a number of step-wise tabs, including the KM Plots tab, Association Tests tab, Adjusted Survival Curves tab and Survival Trees tab, enabling users to carry out a complete and statistically robust survival analysis of the data under scrutiny.

Initially the KM Plots tab provides a space to produce survival curves and the corresponding log-rank tests to identify survival-associated categorical variables, both visually and statistically (Figure 3.53). This tab contains three sub-tabs allowing users to produce KM plots and log-rank tests for selected clinical variables, for segmented CNA metrics and for clinical variables of interest split based on treatment assignment (i.e. where patients received different treatments, e.g. split into patients who received chemotherapy and patients who did not) simultaneously. Within each sub-tab, the box sidebar allows users to indicate which columns contain the survival time, event status (OS, DSS or RFS) and the clinical or CNA variable of interest. Importantly, when generating KM curves for variables split by treatment assignment, the selected treatment variable must be a binary variable of the form YES/NO. Again, the KM plots produced in this tab can be customised and exported as .pngs or .svgs using the sidebar options.

The next tab, the Association Tests tab, utilises a number of association tests to determine if there exists a relationship between selected variables and enables users to detect potential confounding variables in their analysis. As is the case in most tabs, users select the variables of interest within the box sidebar and view the output in the main panel space. Statistical association tests available in GNOSIS include the χ^2 test, Fisher's exact test, simulated Fisher's exact test, ANOVA, Kruskal-Wallis test, pairwise t-test and Dunn's test. It is important that users know which statistical test(s) are most appropriate to answer their research question(s), how to interpret the output of the selected test(s) correctly and how to check that the relevant assumptions of the selected test(s) are met. To aid users in this, information buttons containing links to useful resources are available throughout the app. Briefly, the χ^2 test is used to assess the association between two categorical variables with sufficient cell sizes (Figure 3.54) and Fisher's exact test can be used when any cell size is sufficiently small. ANOVA can be used to test whether there is a difference in means between groups and the Kruskal-Wallis test may be used in the situation where the assumptions of the ANOVA test are not met. Pairwise comparisons can also be carried out using the t-test and/or Dunn's test. In all cases, results of each individual association test are displayed in the main panel alongside the adjusted p-values calculated using the Benjamini-Hochberg (BH) p-value adjustment.

The CPH models tab allows users to use univariate and multivariable Cox models to identify survival-associated variables and assess whether the assumptions of these models are met (Figure 3.55). While KM curves and log-rank tests are only suitable for categorical variables, the CPH model accepts both categorical and continuous variables and extends survival analysis methods to simultaneously assess the effect of a number of selected variables on survival time. Within the univariate Cox models and multivariable Cox models sub-tabs, the box sidebar enables users to select which columns contain the survival time, event status (OS, DSS or RFS), and the variables to be included in the models. For each univariate Cox model fitted the output is displayed along with a summary table containing the BH adjusted p-values and for each multivariable Cox model fitted the output is displayed. The PH assumption of the fitted multivariable Cox models can be assessed using graphical diagnostics based on the scaled Schoenfeld residuals. Again, these plots can be customised and exported as .pngs or .svgs using the sidebar options.

Figure 3.53: Kaplan-Meier plot for disease-specific survival for each CNA Quartile group. The p-value associated with the log-rank test and a risk table displaying the number of patients at risk at each time interval is displayed.

Following multivariable Cox model selection, users may want to produce corresponding adjusted survival curves, which are survival curves adjusted for the covariates included in the multivariable Cox model. This functionality is provided in the Adjusted Survival Curves tab, where users are provided with spaces to view the multivariable Cox model fitted in the previous tab, to set up the 'new data' data frame including the grouping variable, variable of interest and the variables to be

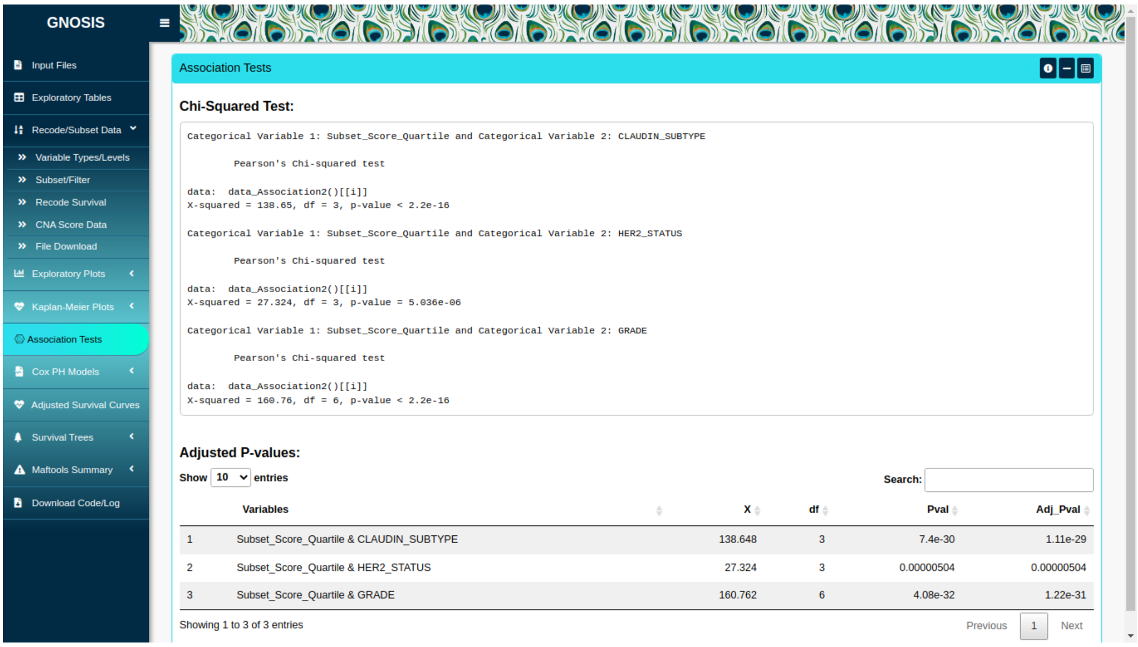


Figure 3.54: Example of a χ^2 analysis of the data. Individual χ^2 tests displayed in top box and table with adjusted p-values displayed in bottom box.

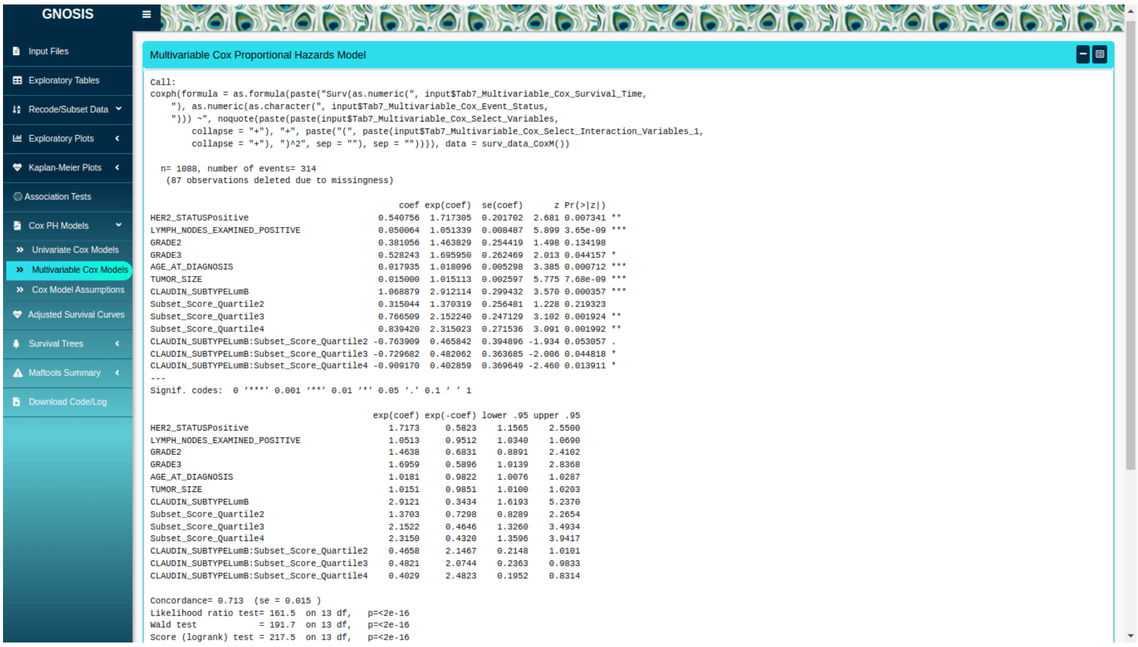


Figure 3.55: Example of an implementation of a multivariable Cox model.

kept constant and to view or download the adjusted survival curves in a number of ways. All covariates included in the selected multivariable Cox model should be included in the 'new data' data frame and when computing adjusted survival curves, the value chosen for a covariate being adjusted is the mean or median for continuous variables and the mode for categorical variables.

If the PH assumption of the selected multivariable Cox model is violated, the Survival Tree tab provides users with a space to apply recursive partitioning survival trees to their data. Users can use the rpart (Therneau and Atkinson, 2022) or ctree (Hothorn and Zeileis, 2015; Hothorn et al., 2006) algorithms with customised parameters. Like the KM Plots tab within each sub-tab, users can use the box sidebar to indicate which columns contain the survival time, event status (OS, DSS or RFS) and the clinical or CNA variable of interest. The main outputs of this tab are survival trees containing the selected variables along with the corresponding KM curves (Figure 3.56). Similar to previous tabs, the survival trees and corresponding KM curves can be exported as .pngs or .svgs with specified a plot width and height. It should be noted that the ctree algorithm requires the selected categorical variables to be factors.

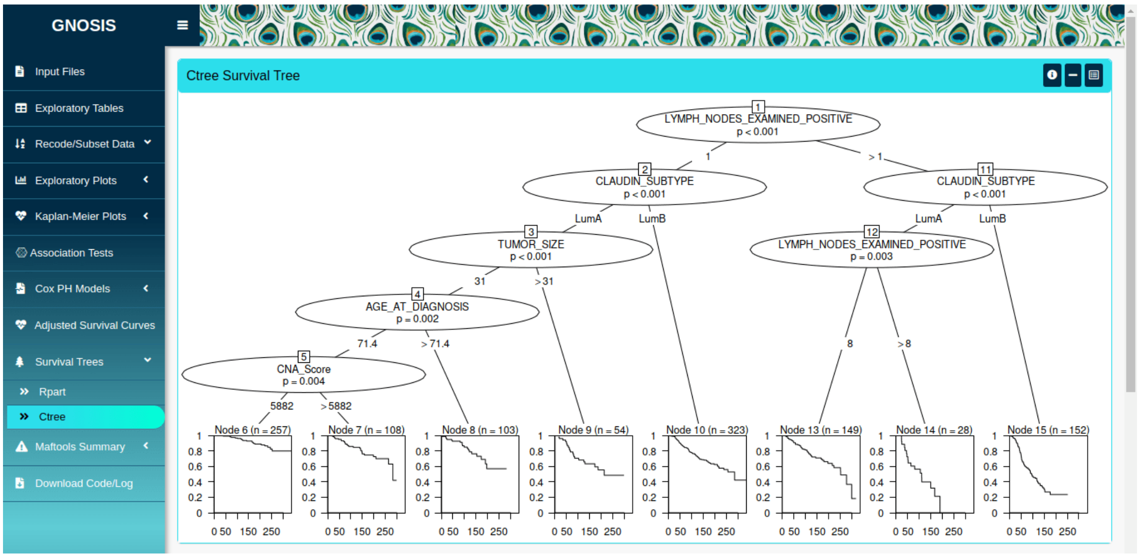


Figure 3.56: Example output of a ctree survival tree analysis.

3.5.1.4 Mutation Analysis (Tab 10)

An additional function of GNOSIS is the ability to perform mutation analysis. The Mutation Analysis tab in GNOSIS allows users to summarise, analyse and visualise mutation annotation format (MAF) files using maftools (Mayakonda et al., 2018). MAF files are tab-delimited text files containing aggregated mutation information and are commonly available as part of the cBioPortal downloads. The Mutation Analysis tab in GNOSIS provides users with two sub-tabs, MAF Text Summary and MAF Visual Summary. The MAF Text Summary sub-tab allows users to produce and view text summaries of the MAF files including the MAF summary, sample summary, gene summary and summary of the associated clinical data, if provided. These summaries contain information on the number of mutations, type of mutations and genes affected by these mutations. The MAF Visual Summary sub-tab enables users to examine the mutational landscape of the tumours in a graphical way (Figure 3.57). The plots available include MAF summary plots, oncoplots, oncostrips, graphs displaying transition and transversion rates, lollipop plots for up to three genes simultaneously, mutation load plots and somatic interaction plots, all derived from the original maftools package. All the visualisations produced in this tab can be customised and exported in .png or .svg format with specific dimensions. It should be noted that if clinical data is provided users need to make sure the column named "Tumor_Sample_Barcode" is present, if this column is not provided the clinical data will not be loaded in by maftools.

Figure 3.57: Sample output from use of the maftools package. A MafSummary plot is displayed.

3.5.1.5 Downloadables (Tab 11)

The Download Code/Log tab in GNOSIS facilitates reproducible research by providing a space for users to view and download a log containing information on all the inputs selected throughout the session and also an R script containing code to reproduce the outputs displayed in the app (Figure 3.58).

3.5.2 Operation

GNOSIS works on R versions $\geq$ 4.0.0 and depends on a number of R packages including BiocManager, shiny, shinymeta, shinydashboard, dashboardthemes, shinydashboardPlus, shinyWidgets, shinycssloaders, shinylogs, fontawesome, DT, cBioPortalData, tidyverse, ggplot2, fabricatr, reshape2, operator.tools, rpart, rpart.plot, partykit, coin, survminer, survival, stats, rstatix, DescTools, car, compareGroups, R.utils, RColorBrewer and maftools (Hothorn et al., 2006; Wickham, 2007; Hothorn et al., 2008; Subirana et al., 2014; Hothorn and Zeileis, 2015; Wickham, 2016; Brown, 2017; Mayakonda et al., 2018; Wickham et al., 2019; Fox and Weisberg, 2019; Sali and Attali, 2020; Ramos et al., 2020; Kassambara et al., 2021; Cheng and Sievert, 2021; Chang and Borges Ribeiro, 2021; Granjon, 2021; Bengtsson, 2022; Chang et al., 2022; Therneau and Atkinson, 2022; Blair et al., 2022; Meyer and Perrier, 2022; Milborrow, 2022; Neuwirth, 2022; Lilovski, 2022; Therneau, 2023; Morgan and Ramos, 2023; Perrier et al., 2023; Iannone, 2023; Xie et al., 2023; R Core Team, 2023; Kassambara, 2023; Signorell, 2023) which are automatically installed and loaded when running GNOSIS manually from RStudio. Should GNOSIS be run using the `runGitHub()` function, shiny must be installed beforehand.

GNOSIS is available on shinyapps.io and GitHub. This enables users to access GNOSIS via a web browser or run GNOSIS locally by downloading, extracting and launching the app manually in RStudio, or running the app in RStudio using: `shiny::runGitHub(repo='GNOSIS',username = 'Lydia-King',ref="GNOSIS_Software_Tool_Article")`.

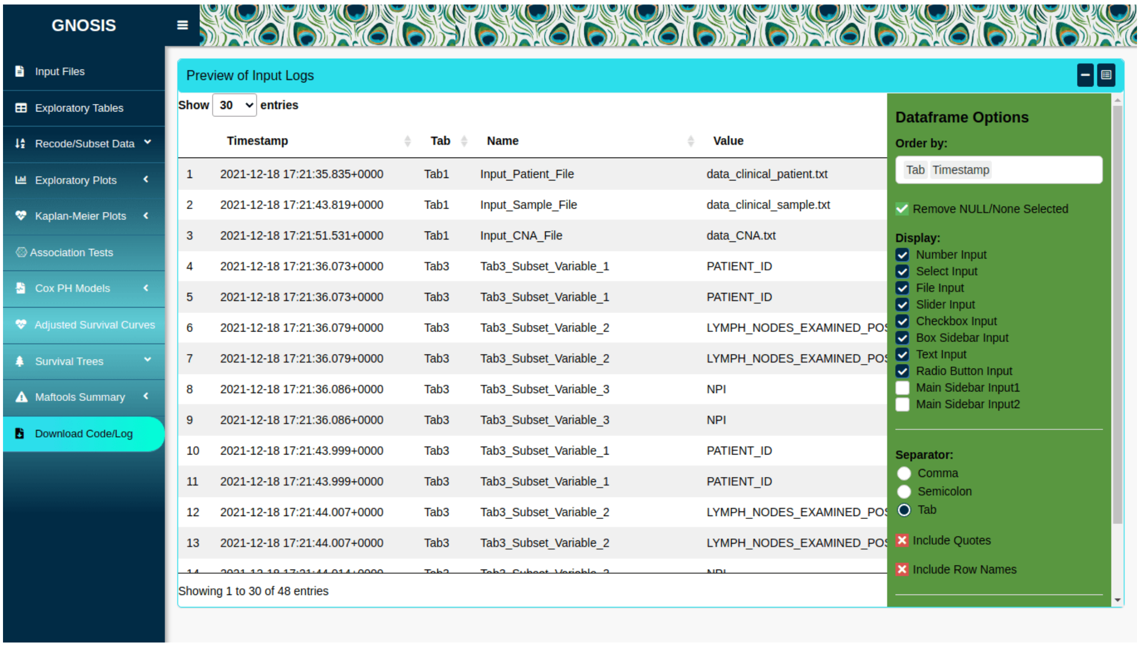


Figure 3.58: Dataframe containing log of inputs selected. This can be downloaded as a .txt file. Option to download R script containing code run in app also available.

3.5.3 Use Cases

GNOSIS was developed as part of a study to carry out an exploratory and statistically robust survival analysis on the METABRIC Luminal breast cancer cohort (King et al., 2021a). Using the wide variety of functions that GNOSIS offers, we were able to efficiently determine that CNAs reflecting GI in Luminal breast cancers are associated with survival. This work acts as a use case and demonstrates the utility and capability of GNOSIS to facilitate oncogenomic analysis.

3.5.4 Data Availability

The data utilised in the study discussed in the previous section, King et al. (2021a) are available for download on cBioPortal as well as Zenodo (King et al., 2021b). In addition, instructional videos providing a walkthrough of GNOSIS and example Rmarkdown files and R scripts containing the code to run the analysis presented are provided on Zenodo (King, 2022) and on the project's GitHub, respectively. More information on the nature of the underlying and extended data can be found in King et al. (2022).

3.5.5 Integration of cBioPortalData

In version v1.0.3 of GNOSIS, King et al. (2022), GNOSIS has an Input Files tab that accesses files locally on the user's file system and allows users to upload the clinical patient and sample data, summary CNA data and mutation data manually. This requires users to download the data from cBioPortal and then upload this data to GNOSIS. To bypass this step and make GNOSIS more efficient, the Bioconductor package cBioPortalData (Ramos et al., 2020) was integrated into GNOSIS. The cBioPortalData package allows users to access study datasets from cBioPortal, either from the pre-packaged zip/tar files or from an API interface. As a result, version v1.0.4, available on Bioconductor, has an updated Input Files tab containing two sub-tabs, one to upload the data, either manually or using the cBioPortalData API, and one to preview the data. Within the Input Files tab users can view all the available cBioPortal studies and select which study to download and analyse by clicking on the row corresponding to the study of interest (Figure 3.59).

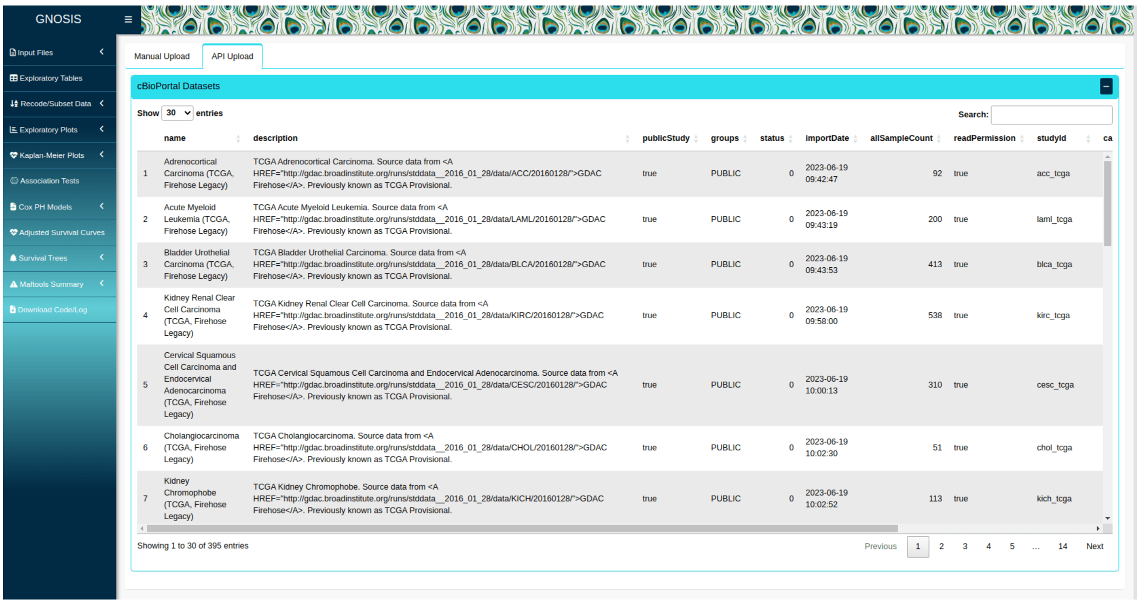


Figure 3.59: Datatable containing list of cBioPortal studies users can select.

3.6 Conclusions

Using a range of these semi-parametric and non-parametric survival models, it was observed that a subset of both global and chromosome arm CNA Score and Burden metrics are useful predictors of disease-specific survival outcomes.

Focused analysis of Luminal METABRIC patients showed that absolute CNA Score metric, implemented either as predetermined categorised quartiles or original continuous variable, can stratify subsets of patients based on disease-specific survival and identify Luminal A patients who are at elevated risk, results published in *Survival Outcomes are Associated with Genomic Instability in Luminal Breast Cancer* (King et al., 2021a).

Extensive study was then given to CNA Score and Burden metrics, calculated globally and for each chromosome arm, for all PAM50 subtypes and IntClusts. Interestingly a large proportion of the predictors selected as useful predictors for disease-specific survival outcomes were CNA Del Score and Burden metrics, from the global CNA metrics, and CNA Del Score and Burden on chromosome 3p and 18q, from the chromosome arm CNA metrics. These results further suggest that deletions are more harmful than amplifications and can help identify patients with worse survival outcomes. It is also noted that the CNA metrics can provide additional information to already used molecular classifications and clinical variables.

In addition, accessibility of results is supported by building and publication of GNOSIS (King et al., 2022), an R Shiny app that enables the tractable and efficient exploratory analysis of cBioPortal clinical and genomic data products in a reproducible manner.

The associations observed between the CNA landscape of tumours, particularly the deletion landscape, and survival could potentially be the result of gene expression changes caused by the observed CNAs. In the next chapter we will explore the possibility of incorporating gene expression data, enabling us to examine how the presence of CNAs globally, across chromosome arms and in specific genes influences gene expression.

4 Effect of Copy Number Alterations on Gene Expression

It has been reported in literature that CNAs can promote tumour progression by altering gene expression levels (Pollack et al., 2002; Stranger et al., 2007; Curtis et al., 2012; Bhattacharya et al., 2020). Here, we utilise differential gene expression analysis (DGEA) to explore the impact of CNAs on gene expression. DGEA identifies differences in gene expression comparing conditions or states, e.g. healthy/disease or treatment/control states, allowing identification of differentially expressed genes (DEGs) and biological pathways that may be perturbed. DGEA has been used to compare gene expression patterns in breast cancer facilitating the formation of the IntClust molecular classification and a number of prognostic and predictive assays (Curtis et al., 2012; Nicolini et al., 2018).

Microarrays and RNA sequencing are the two most common technologies used to study transcriptional activity (Harrington et al., 2000; Wang et al., 2009). Microarrays contain thousands of probes, usually oligonucleotide or complementary DNA probes, anchored to a glass slide at defined positions. Fluorescently labelled RNA or DNA in observed tissue samples hybridise to the probes present on the array and hybridization intensities measured for each probe are converted to a quantitative read-out of relative gene expression levels. This allows simultaneous measurement of the expression level of thousands of genes and direct comparison of different tissue samples via different fluorescent labelling on a single hybridization assay (Harrington et al., 2000; Trevino et al., 2007). There are numerous microarray platforms available for carrying out gene expression analysis. In the METABRIC study (Curtis et al., 2012), the microarray platform used for measuring gene expression was the HumanHT-12 BeadChip (v3) produced by Illumina, which supports highly efficient whole-genome expression studies and expression-based quantitative trait loci studies. This RNA microarray contains more than 48,000 probes that provide genome-wide transcriptional coverage of more than 25,000 RefSeq and UniGene annotated genes, including well-characterised genes, gene candidates, and splice variants (Illumina, 2010). Microarrays can be used for other purposes, such as genotyping and for the detection of CNAs. To carry out copy number and genotype analysis in the METABRIC study, the Affymetrix Genome-Wide Human SNP Array 6.0 array was utilised (Curtis et al., 2012), containing over 1.8 million probes, approximately 906,600 probes for single nucleotide polymorphisms (SNPs) and 946,000 probes for the detection of copy number variation. These probes are evenly distributed across the entire genome, facilitating measurement of copy number, allele-specific copy number, and copy number-neutral LOH (Affymetrix, 2009).

This chapter provides an overview of approaches to DGEA, including a common R programming package limma (Ritchie et al., 2015). DGEA is applied to compare gene expression between groups of stratified patients identified as having similarity in survival curves, derived by incorporating the CNA information, in Section 3.4, i.e. comparing patients of particular survival tree nodes of interest. DGEA is also applied to compare gene expression between different gene CNA states, i.e. homozygous deletion (-2), hemizygous deletion (-1), diploidy (0), single copy gain (+1) and high-level amplification (+2). To finish, the differentially expressed gene sets emerging from this analysis, are compared to previously defined breast cancer prognostic and predictive gene sets, i.e. Oncotype DX, MammaPrint, Prosigna and

BCI, and the molecular classification gene sets, i.e. PAM50 and IntClust.

4.1 Differential Gene Expression Analysis using Limma

Limma is a Bioconductor/R software package that facilitates analysis of data generated from microarray and RNA-sequencing gene expression experiments (Ritchie et al., 2015). In its implementation, limma fits a linear model to each gene simultaneously, taking as input a matrix of expression values, with rows corresponding to genes and columns corresponding to RNA samples, and a user-specified model design matrix. These linear models are incredibly flexible, capable of handling complex experimental designs, and can be used to test various hypotheses. In addition, limma can distinguish and estimate different sources of variability, e.g. between genes, between samples, variations in quality of data sources, and technical or biological heterogeneity. Limma applies adjustments for different sources of variability, implementing information borrowing between genes and use of observation weights and variance modelling, enabling robust conclusions for statistical testing, particularly when sample sizes are small.

A general limma pipeline, for microarray analysis, begins with preprocessing, including background correction and normalisation, followed by creation of the design matrix and generation of array weights. Array weights correspond to the relative reliability of each microarray based on how well the expression values from that array follow the linear model, where arrays that have larger residuals, i.e. larger deviations, are assigned lower weights. The linear model is then fitted for each gene given a series of arrays, with the option to apply array weights. To test specific hypotheses, contrasts are fitted using the model estimates, and log-odds of differential expression, moderated t-statistics, and moderated F-statistic are computed. The moderated statistics borrow information from across genes and samples, facilitated using empirical Bayes methods to obtain posterior variance estimators (Smyth, 2004). The resulting estimated variance for each gene is then an informed balance between the gene-wise estimator obtained from the data for that gene alone and the global variability across all genes estimated by pooling the ensemble of all genes. Finally, a number of functions can be used to summarise and visualise the results of the estimated linear models, hypothesis test, and apply p-values adjustments for multiple testing, e.g. topTable and volcano plots.

The basic limma pipeline, fitted with array weights and applying the same common design matrix for all genes, is used in Section 4.2.1 to identify up- or down-regulated genes between survival tree nodes. A single design matrix, including the predictor variable indicating survival tree node membership for the individual patient, is defined and applied to every gene in the dataset. The contrast matrix specifies which survival tree node comparisons are of interest. Using the topTable results, a gene is called as differentially expressed if the adjusted p-value is below 0.05 and the absolute log-fold change is above 0.58, i.e. greater than a 1.5-fold change (McCarthy and Smyth, 2009).

In Section 4.2.2, to model the direct relationship between CNA states and gene expression, a modified limma pipeline allowing a different design matrix to be fitted for each gene is implemented. The modification, and additional tailored R programming, is necessary in this application since a patient's CNA state may differ across genes. In addition, since some genes may not exhibit any CNAs or have very few, resulting in cases where the sample size of the particular CNA state may be too small to support inference, genes not altered with sufficient frequency (< 1%) in the CNA state of interest were filtered out for that comparison only.

4.2 Application to METABRIC cohort

DGEA to compare gene expression between METABRIC patients stratified by the global and chromosome arm CNA metric informed survival profiles was carried out. We first focus on presenting four DGEA applications of interest from the survival trees produced as a result of incorporation of the global CNA metric information and then on presenting three DGEA applications of interest from the survival trees produced as a result of incorporation of the chromosome arm CNA metric information. These survival trees, produced for a range of survival times (DSS, 5- and 10-year DSS) and algorithms (rpart and ctree), were selected for DGEA as they partitioned the METABRIC patients first on PAM50 or IntClust and subsequently on a single global or chromosome arm CNA metric, simplifying comparison and inference. The gene expression data, downloaded from cBioPortal in 2023 and used here, include 18,739 genes for which CNA, gene expression and genomic location data were available.

4.2.1 Differential Gene Expression Analysis of Global CNA Metric Survival Tree Nodes

Focusing on the survival trees including PAM50 subtype and the six global CNA Burden metrics as candidate predictors (Figure 4.1A), the DSS ctree survival tree indicates that for patients within specific PAM50 subypes (Claudin-low and Luminal A), survival profiles could be stratified into two nodes: Node 5 with poorer DSS outcome compared to Node 4 with better DSS outcome, where global CNA Del Burden performed as the classifier into the two nodes, patients with global CNA Del Burden above a value of 18.28% in Node 5. Applying DGEA to compare gene expression of patients classified into Node 5 (higher GI and lower DSS outcome) and patients classified into Node 4 (lower GI and better DSS outcome), reveals 77 down-regulated genes and 37 up-regulated genes. A number of genes including CXCL10 and CXCL9 are identified as significantly up-regulated in Node 5, while genes including PIP, ANKRD30A, and REEP6, are among those identified as significantly down-regulated in Node 5 compared to Node 4 (Figure 4.1A).

The 5-year DSS ctree survival tree (Figure 4.1B) indicates that for Luminal A patients, survival profiles could be stratified into two nodes: Node 9 with global CNA Del Burden $> 14.55\%$ and poorer DSS outcome, compared to Node 8 with global CNA Del Burden $\leq 14.55\%$ and better DSS outcome. Applying DGEA to compare gene expression of patients classified into Node 9 (higher GI and lower DSS outcome) and patients classified into Node 8 (lower GI and better DSS outcome), reveals 15 down-regulated genes and 4 up-regulated genes. Genes SLC7A5, PITX1 and S100P, are identified as significantly up-regulated in Node 9, while genes PIP, FCGBP and SCUBE2, are among those identified as significantly down-regulated in Node 9 compared to Node 8 (Figure 4.1B).

Focusing on the survival trees including IntClust and the six global CNA Burden metrics as candidate predictors, the DSS rpart survival tree (Figure 4.2A), indicated that for patients within specific IntClusts (3, 4ER+, 7 and 8), survival profiles could be stratified into two nodes: Node 4 with poorer DSS outcome compared to Node 3 with better DSS outcome, where global CNA burden performed as the classifier into the two nodes, patients with global CNA Burden above a value of 24.90% in Node 4. Applying DGEA to compare gene expression of patients classified into Node 4 (higher GI and lower DSS outcome) and patients classified into Node 3 (lower GI and better DSS outcome), reveals 13 down-regulated genes and 3 up-regulated genes. Genes UBE2C and S100P, are identified as significantly up-regulated in Node 4, while genes PIP, CYBRD1, IRX2, are among those identified as significantly down-regulated in Node 4 compared to Node 3 (Figure 4.2A).

The 10-year DSS ctree survival tree (Figure 4.2B), showed that for patients of certain IntClusts (IntClust 3, 4ER+, 7 and 8), their 10-year survival patterns could be stratified, using global CNA Del Burden, into two nodes: Node 7 with CNA Del Burden > 8.82% and poorer 10-year DSS outcome compared to Node 7, where patients have CNA Del Burden ≤ 8.82% and better 10-year DSS outcome. Applying DGEA to compare those with higher deletion burden, Node 7, to lower deletion burden, Node 6, identifies 18 differently expressed genes including UBE2C, PIP, and IRX2 (Figure 4.2B).

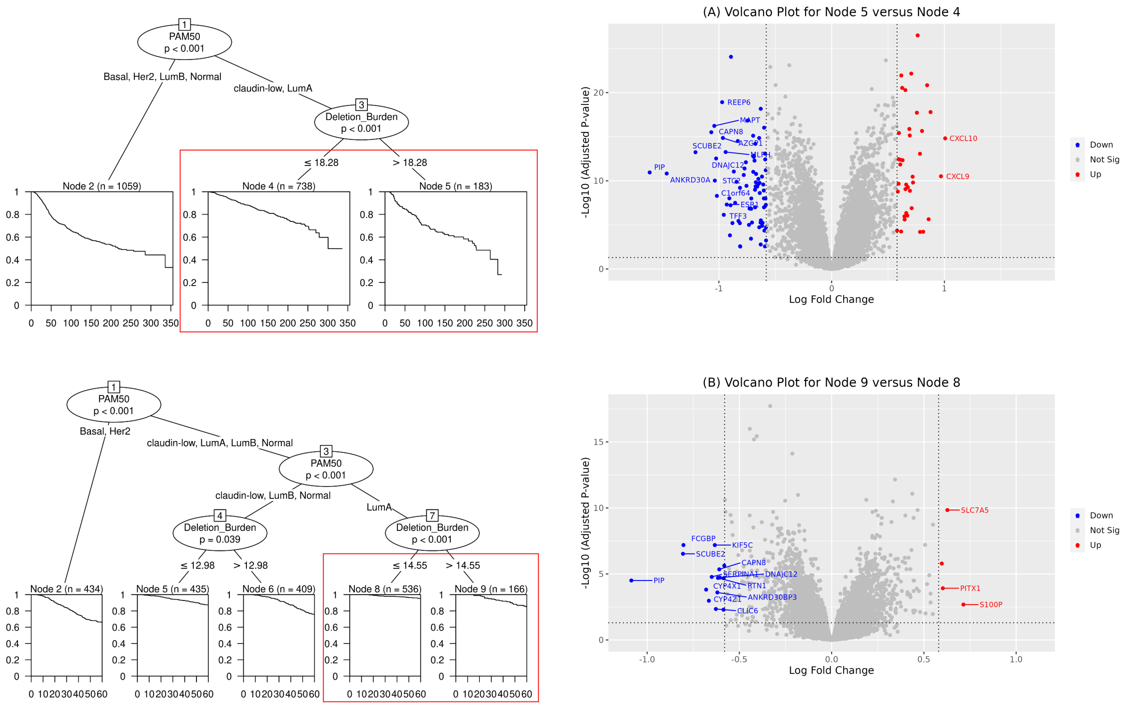


Figure 4.1: Volcano plots resulting from DGEA applied to compare nodes informed by global CNA Burden metrics and PAM50 subtype. Plots show differentially expressed genes between (A) Node 4 and Node 5 of the ctree DSS survival tree and (B) Node 8 and Node 9 of the ctree 5-year DSS survival tree.

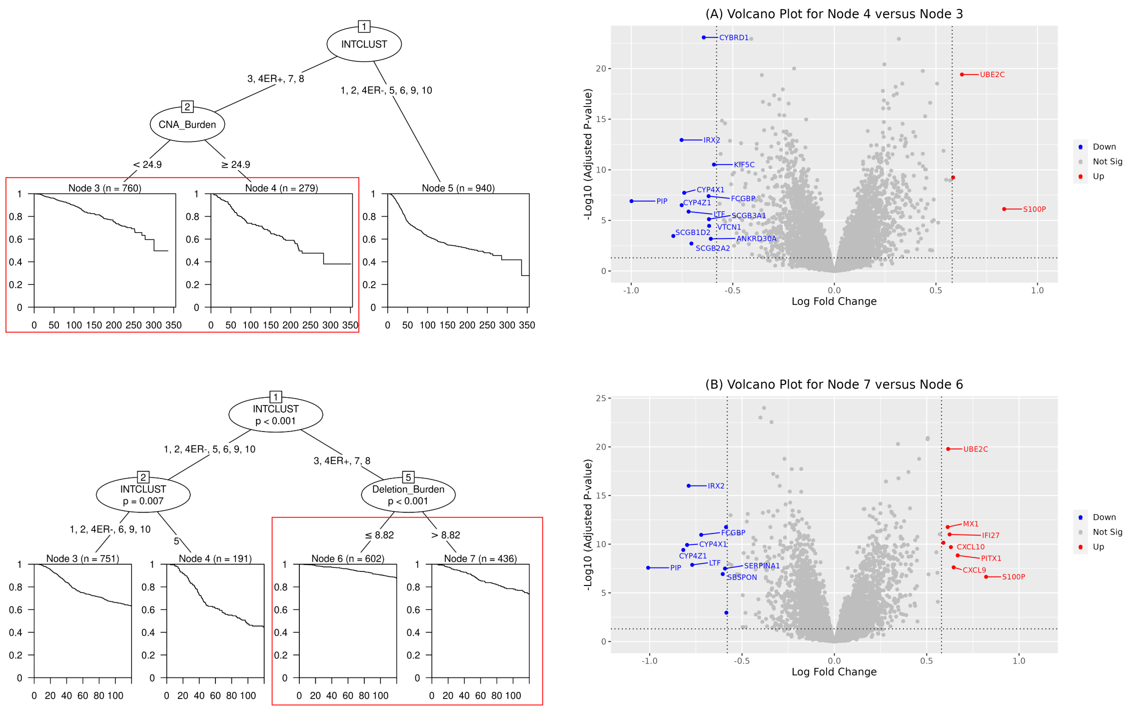


Figure 4.2: Volcano plots resulting from DGEA applied to compare nodes informed by global CNA Burden metrics and IntClust. Plots show differentially expressed genes between (A) Node 3 and Node 4 of the rpart DSS survival tree and (B) Node 6 and Node 7 of the ctree 10-year DSS survival tree.

4.2.2 Differential Gene Expression Analysis of Chromosome Arm CNA Metric Survival Tree Nodes

Section 3.4 provided survival trees where splits into stratified groups of patients were informed by chromosome arm CNA burden metrics. DGEA is applied to three survival trees of particular interest. The DSS ctree survival tree utilising the PAM50 subtype molecular classification (Figure 3.31) indicated that Luminal A and Claudin-low patients can be further stratified into two nodes: Node 4 and Node 5, classified by CNA Del Burden of chromosome 3p. Higher chromosome 3p CNA Del Burden, above an optimised threshold of 30.21%, partitioned patients into Node 5 with poorer DSS survival profile. DGEA comparing gene expression between Node 4 (low CNA Del Burden on chromosome 3p) and Node 5 (high CNA Del Burden on chromosome 3p), indicates that the genes LZTFL1, IMPDH2, ZMYND10, LRIG1, P4HTM, FLNB, MST1, GPD1L, KCTD6, ACOX2, LTF, located on chromosome 3p, are down-regulated in Node 5 compared to Node 4 (Figure 4.3A and B).

The DSS rpart survival tree utilising the IntClust molecular classification (Figure 3.37) indicated that for IntClust 3, 4ER+, 7 and 8 patients, CNA Del Burden on chromosome 18q partitions patients into Node 3, CNA Del Burden $< 69.78\%$ with better DSS outcome and Node 4, CNA Del Burden $\geq 69.78\%$ with poorer DSS outcome. DGEA comparing gene expression between Node 3 (low CNA Del Burden on chromosome 18q) and Node 4 (high CNA Del Burden on chromosome 18q), indicates that 23 genes are down-regulated and 19 are up-regulated in Node 4 compared to Node 3. There is one gene located on chromosome 18q that is classified as down-regulated in patients with higher levels of deletions on chromosome 18q, BCL2.

The 5-year DSS rpart survival tree utilsiing the PAM50 subtype molecular classification (Figure 3.32) indicated that for Luminal A patients, CNA Del Burden on chromosome 11p, above 7.72%, is associated with poorer 5-year DSS outcome. DGEA comparing gene expression between Node 3 (low CNA Del Burden on chromosome 11p) and Node 4 (high CNA Del Burden on chromosome 11p), indicates that no genes, located on chromosome 11p, are differentially expressed in Node 4 compared to Node 3.

Comparing the gene expression profiles of the survival tree nodes for focused examples of interest, produced using the global and chromosome arm CNA Burden metrics, indicates there is widespread differential gene expression between Node partitions.

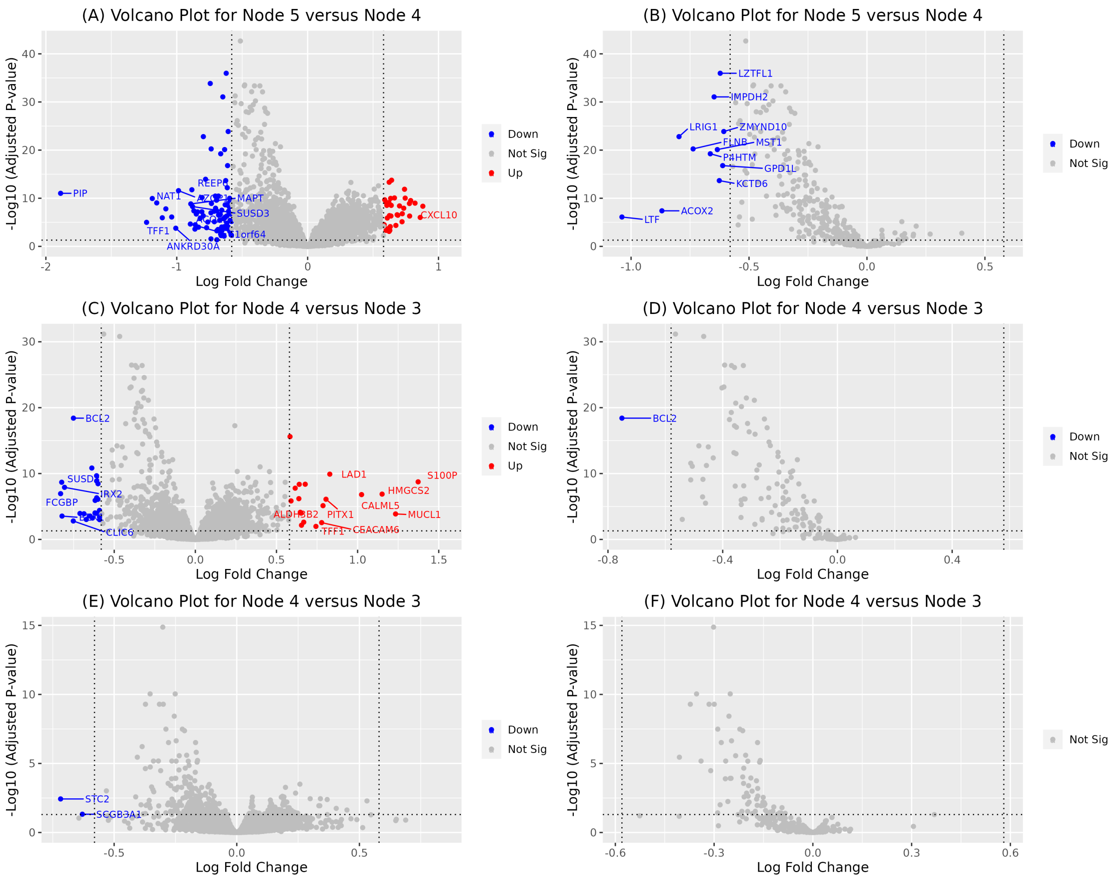


Figure 4.3: Volcano plots resulting from DGEA applied to compare Nodes informed by chromosome arm specific CNA Burden metrics. Plots show differentially expressed genes between: (A) Node 5 and Node 4 of the ctree DSS survival tree, (C) Node 4 and Node 3 of the rpart 5-year DSS survival tree, and (E) Node 4 and Node 3 of the rpart 10-year DSS survival tree. Plots (B), (D) and (F) correspond to plots (A), (C) and (E), but only show the genes present on the chromosome arm of interest.

4.2.3 Differential Gene Expression Analysis of CNA States

The CNA metrics are a cumulative measure over all genes. To explore the direct relationship between the gene's individual CNA and the gene's expression we propose, fit and use a modified DGEA model. Applications of DGEA using limma usually applies the same design matrix for all genes:

$$\text{Gene Expression}_g = \mathbf{X}\beta_g + \epsilon_g$$

where $\mathbf{X}$ is the design matrix and is the same for each gene g, β is the vector of parameters and ϵ is the error vector. However, when the CNA states of the gene is the explanatory variable, a patient may have different CNA states for different genes, requiring gene-specific model design matrices. For example, patient MB.0010 may have an amplification (+2) in the gene ACTL8 and should be placed in the amplification group with the other patients displaying an amplification in this gene, but this patient may also have a hemizygous deletion (-1) in the gene MFSD2 and therefore should also be placed in the deletion group with the other patients displaying a deletion in this gene. Here, the CNA state of a gene is one of *amplification, gain, neutral, hemizygous deletion* or *homozygous deletion*, and is considered as predictor in a model with gene expression as the response, proposing the following model specification:

$$\text{Gene Expression}_{gi} = \beta_{0g} + \beta_{1g}\text{CNA State}_{gi} + \epsilon_{gi}$$

where CNA State$_{gi}$, for gene g and patient i, can either take on one of three states, i.e. amplification, neutral or deletion, or one of five states, i.e. amplification, gain, neutral, hemizygous deletion and homozygous deletion, depending on the specification. This five-state specification corresponds to the CNA States assigned by (Curtis et al., 2012). A three-state specification is derived from this, grouping hemizygous and homozygous deletions together, and gains and amplifications together.

Focusing first on the three-state model specification, and the contrast of patients who exhibit a CNA Amplification in a particular gene, to those who exhibit no alteration, CNA Neutral, 772 genes including GRB7, ERBB2 (HER2) and PGAP3, are significantly up-regulated for the amplification state (Figure 4.4A). Comparing the expression of genes with CNA Deletions to genes without alteration, CNA Neutral, 362 genes are differentially expressed (Figure 4.4C). In this case, 345 are down-regulated, including FOXA1, IL6ST and EEF1A2. These plots suggests that the presence of a CNA in a gene has the potential to impact the expression of that gene. Figure 4.5A, B, C and D show that in the five-state specification many genes that have a gain or amplification present are significantly up-regulated, while Figure 4.5E, F, G and H show that in the five-state specification many genes that have a hemizygous or homozygous deletion present are significantly down-regulated.

The number of genes filtered out of the three-state model results, comparing patients who exhibit a CNA Deletion in a particular gene, to those who exhibit no alteration (i.e. CNA Neutral) is low (418 out of 18,739 filtered from the results). Similarly, the number of genes filtered out of the three-state model results, comparing patients who exhibit a CNA Amplification in a particular gene, to those who exhibit no alteration (i.e. CNA Neutral) is also negligible with 99 genes out of 18,739 being filtered from the results. However, in the five-state specification the impact of small sample size alteration groups is much more pronounced, particularly in the amplification and homozygous deletion groups. The number of genes filtered out when comparing the CNA Gain to CNA Neutral group, CNA Amplification to CNA Neutral group, CNA Hemizygous Deletion to CNA Neutral group and CNA Homozygous Deletion to CNA Neutral group are 129, 9,780, 422 and 18,617, respectively.

Two gene sets, ModLim3 and ModLim5, containing differentially expressed genes of sufficiently large sample size, i.e. genes with an adjusted p-value below 0.05 and absolute log fold change above 0.58 in at least one contrast, in the three- and five-state specification are defined.

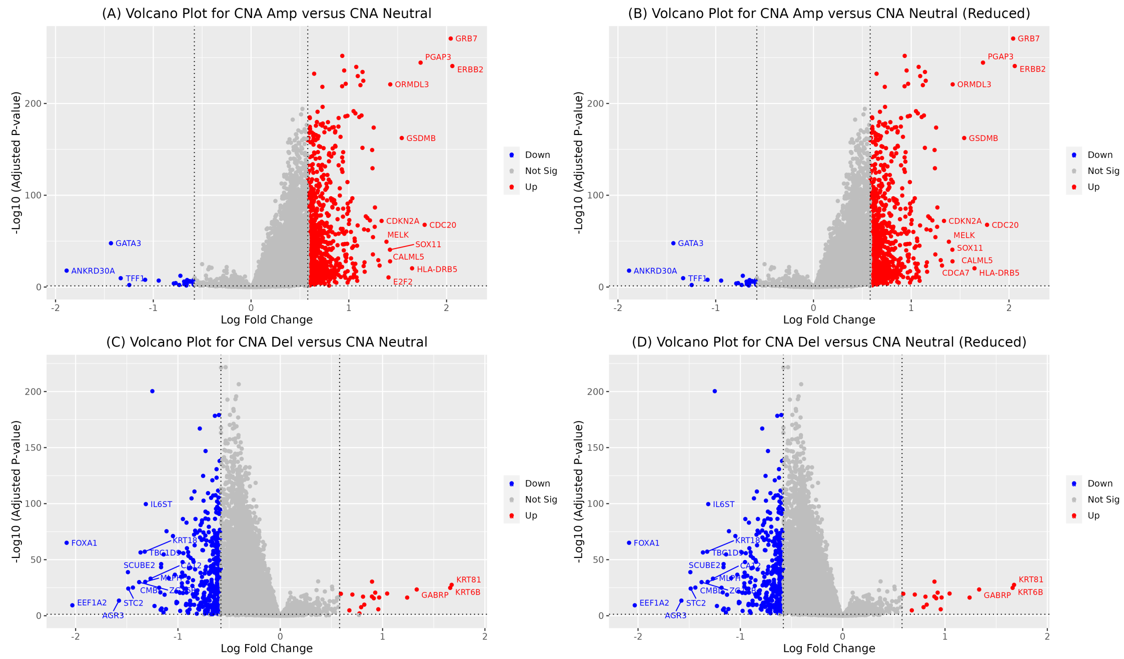


Figure 4.4: Volcano plot showing differentially expressed genes comparing CNA three-state specifications. Plots denote (A) the CNA Amplification and CNA Neutral states, where all genes are shown (B) the CNA Amplification and CNA Neutral states, where only the genes displaying sufficient numbers of patients with an amplification are shown (C) the CNA Deletion and CNA Neutral states, where all genes are shown (D) the CNA Deletion and CNA Neutral states, where only the genes displaying sufficient numbers of patients with an deletion are shown.

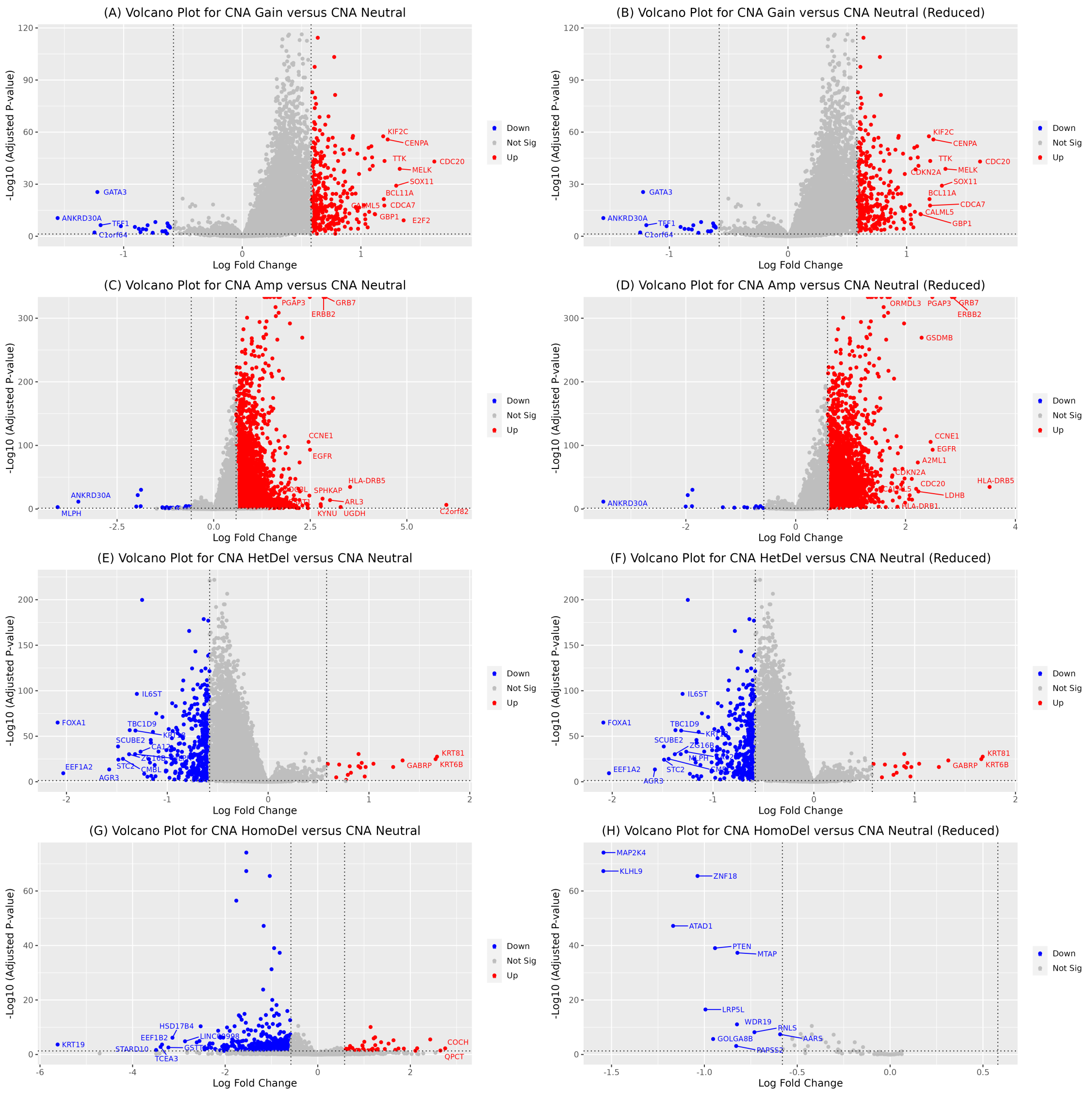


Figure 4.5: Volcano plot showing differentially expressed genes comparing CNA five-state specifications. Plots denote (A) the CNA Gain and CNA Neutral states, where all genes are shown (C) the CNA Amplification and CNA Neutral, where all genes are shown (E) the CNA Hemizygous Deletion and CNA Neutral states, where all genes are shown and (G) the CNA Homozygous Deletion and CNA Neutral states, where all genes are shown. Figures (B), (D), (F) and (H) correspond to (A), (C), (E) and (G), where only the genes displaying sufficient numbers of patients with a CNA are shown.

4.3 Comparative Study

To explore the level of congruence and whether new gene identifications arise in our ModLim3 and ModLim5 differentially expressed gene sets, a comparative study is conducted against some established molecular classification, prognostic and predictive assays reported in the literature. The prognostic and predictive assays selected for comparison are Oncotype DX, MammaPrint, Prosigna (PAM50) and BCI, and molecular classifications, PAM50 and IntClust.

Oncotype DX is a reverse transcription polymerase chain reaction (RT-PCR) based 21 gene signature used to predict the probability of disease recurrence and to help identify patients who are likely to benefit from adjuvant chemotherapy. Of these 21 genes, 16 are linked to cancer and 5 are used as controls (see Appendix C). The expression of these 21 genes is measured and based on the relative expression of the cancer associated genes to the control genes, a score called recurrence score is calculated. This recurrence score, which ranges from 0 to 100, categorises patients into 3 subgroups. Patients with recurrence score less than 18 are assigned as low risk of recurrence, patients with recurrence score 18-30 are assigned as intermediate risk of recurrence and patients with recurrence score greater than 30 are assigned as high risk of recurrence (Paik et al., 2004; Nicolini et al., 2018).

MammaPrint is a microarray-based assay that predicts the probability of disease recurrence and helps guide treatment decision making in breast cancer patients (van ’t Veer et al., 2002; van de Vijver et al., 2002; Knauer et al., 2010; Nicolini et al., 2018). This assay uses the combined expression of 70 genes (the Amsterdam 70-gene expression profile, Appendix C) to categorise tumours as low or high risk for disease recurrence or metastasis. MammaPrint can be used as a prognostic and predictive biomarker in patients with newly diagnosed lymph node-negative or lymph node-positive (1-3 metastatic nodes) invasive breast cancer.

The Prosigna test, formerly known as PAM50, is a microarray-based assay that predicts the probability of disease recurrence, helps guide treatment decision making in hormone-receptor-positive, HER2-negative breast cancer patients and can classify breast cancers into intrinsic molecular subtypes. The assay measures the expression of 58 genes; the 50 PAM50 genes and eight control genes for normalisation (see Appendix C). Based on the relative expression of these genes, a risk score ranging from 0-100 is produced. Using this risk score, patients with node-negative breast cancer are split into low (0-40), intermediate (41-60), or high risk (61-100) categories, while node-positive breast cancer patients are split into low (0-40) or high risk (41-100) categories (Duffy et al., 2017; Nicolini et al., 2018).

BCI is a microarray-based assay that predicts outcome and helps guide adjuvant treatment decision making in lymph node-negative, HR+ and HER2- patients. This assay measures the expression of 11 genes, 7 test genes and 4 control genes (see Appendix C). The development of this assay was based on the combination of two previously identified molecular assays, the HOXB13:IL17BR ratio and the Molecular Grade Index. The HOXB13:IL17BR ratio is a two-gene expression assay that can predict disease-free survival in early-stage breast cancer (Ma et al., 2004) and the Molecular Grade Index is a five-gene expression assay, comprising genes related to histological grade and tumour progression that helps predict risk of distant metastasis in ER+, lymph node-negative patients (Ma et al., 2008).

Table 4.1 provides information on the number of genes measured within each assay/molecular classification and any differences in availability compared to the processed METABRIC CNA and gene expression data, while Figures 4.6 and 4.7 provide results of comparative study.

High congruence is observed when comparing ModLim3 and ModLim5 gene sets (Figure 4.6). Approximately 92% of genes identified as differentially expressed in the three-state specification were also differentially expressed in the five-state specification. Owing to this high congruence, focus was given to the ModLim5 gene set and its congruence with established molecular classification, prognostic and predictive gene sets.

Table 4.1: Table containing gene set information for each assay and availability in the METABRIC data.

Gene Assay	**No. Genes Measured**	**No. of genes in METABRIC data in common with comparative assay**	**Missing Genes**
Oncotype DX	21	21	-
MammaPrint	70 (66 unique)	62	EBF4 (missing from gene expression data) SERF1A (missing from gene expression data) KDM7A (missing from gene expression data) GPR126 (missing from gene expression and CNA data) LOC100288906 (missing from gene expression and CNA data) LOC730018 (missing from gene expression and CNA data) AA555029_RC (missing from gene expression and CNA data) LOC100131053 (missing from gene expression and CNA data)
PAM50_Prosigna	50 (58)	58	-
BCI	11	11	-
IntClust	1,000	959	see Appendix D for full list

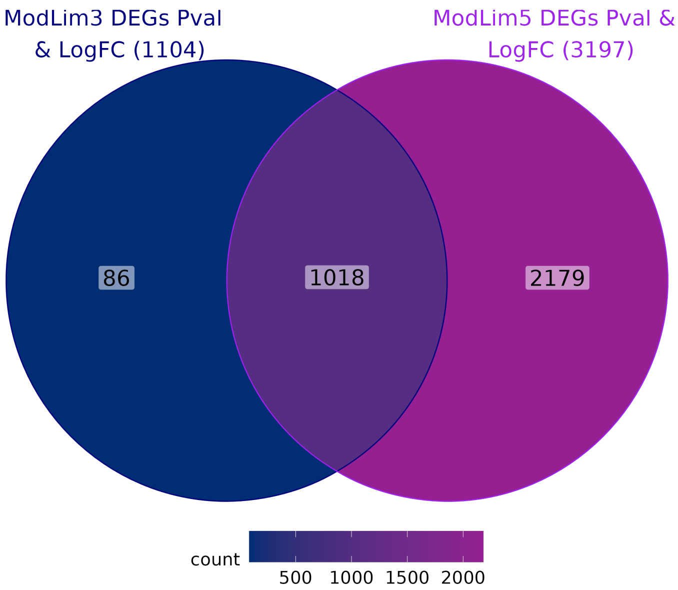


Figure 4.6: Venn diagram showing gene set congruence between the ModLim3 and ModLim5 gene sets.

Figure 4.7: Venn diagram showing gene set congruence between ModLim5 differentially expressed genes and prognostic and predictive assay genes. (A) OncoType DX, (B) MammaPrint, (C) Prosigna/PAM50, (D) BCI, (E) IntClust and (F) IntClust (bound to n = 1,000).

Focusing on the congruence between the prognostic and predictive assays, OncoType DX, MammaPrint, Prosigna and BCI, and the ModLim5 gene set, it is observed that 11 out of 21, 28 out of 66, 33 out of 58, and 3 out of 8 genes, respectively, are present in the ModLim5 gene set. This modest overlap is expected for a number of reasons, primarily due to the difference in objectives of each study. While the prognostic and predictive assays contain genes identified as useful in stratifying patients based on survival outcome and/or response to therapy, the ModLim5 gene set comprises genes where the presence of a CNA influences gene expression.

As expected over half of the IntClust differentially expressed gene set is present in the ModLim5 differentially expressed gene set (Figure 4.7E and F). Congruence between these gene sets was expected for a number of reasons, mainly that the IntClust gene set was also produced using the METABRIC CNA and gene expression data and although the authors used a different approach (ANOVA and the Kruskal-Wallis test), the idea was similar. Interestingly, 281 genes present in the IntClust gene set are not found to be differentially expressed in our analysis. Reasons for this include absence from dataset, difference in method, Kruskal-Wallis versus modified limma, and also the difference in thresholds applied, adjusted p-values versus adjusted p-values and log-fold change. Indeed when only the adjusted p-value threshold is applied in our DGEA, 13,806 genes are differentially expressed, with 954 IntClust genes overlapping. Overall this DGEA analysis identified an additional 2,478 genes whose expression is influenced by CNAs and which should be considered for further investigation as candidate biomarkers for breast cancer treatment and outcome.

4.4 Conclusions

The literature reports that CNAs can influence gene expression and that some genes are more affected by the presence of a CNA than others. Both CNA and gene expression data have been used, separately and in tandem, to form molecular classification and prognostic and predictive assays of genes known to display differential expression and correlate with survival.

In this chapter we compared gene expression profiles between METABRIC patients stratified by similarity in survival profiles as derived by global and chromosome arm specific metrics, i.e. comparing patients in particular survival tree nodes of interest. Under selected thresholds, a number of genes are found to be up- or down-regulated between the survival tree nodes. Genes observed to be up-regulated in patients in survival tree nodes associated with poorer survival outcomes include UBE2C, CXCL10 and S100P, while genes observed to be down-regulated in patients in survival tree nodes associated with poorer survival outcomes include PIP, BCL2 and IRX2. In cancer, overexpression of UBE2C, CXCL10 and S100P has been shown to facilitate cell proliferation, tumour progression and invasion, and correlate with worse survival outcomes (Andersen et al., 2011; Dastsooz et al., 2019; Huang et al., 2021). Similarly, underexpression of PIP, BCL2 and IRX2 can facilitate tumour invasion and is associated poorer survival outcomes and response to therapy (Dawson et al., 2010; Werner et al., 2015; Urbaniak et al., 2018).

To investigate the direct relationship of a gene's CNA state to the gene's expression, a modified limma pipeline was employed, comparing gene expression profiles across patients based on the CNA state for each gene. From this analysis, it is evident that a large number of genes have altered gene expression when there is a CNA present. As expected, genes containing an amplification were often seen to be up-regulated, while genes containing a deletion were often seen to be down-regulated. Overall, using specified thresholds and considering sample size restrictions, 1,104 genes were differentially expressed in the three-gene specification, ModLim3, and 3,197 genes were differentially expressed in the five-gene specification, ModLim5.

A comparative study to explore the extent of overlap between the ModLim5 gene set and molecular classification, prognostic and predictive assays published in the literature indicated a moderate degree of congruence, identifying some of the same genes, but also identifying additional genes to be considered for further investigation as candidate biomarkers for breast cancer treatment and outcome.

The CNA data utilised up to this point is total CNA data. In the next chapter we will consider allele-specific data and explore how to accurately identify and characterise CNA changepoints in allele-specific copy number profiles of breast cancer patients.

5 Modelling Allele-Specific Copy Number Associated Changepoints

CNAs in cancer have been extensively studied, however due to the complexity of cancer genomes, such as frequent deviations from diploidy and the presence of both tumour and non-tumour cells, many studies have been limited to reporting total CNAs across the genome. Total copy number profiling estimates the sum of the copy numbers of the two homologous chromosomes and as such only provides aggregate information. Determining the copy number landscape of each homologous chromosome, i.e. allele-specific copy number profiling, is important for the characterisation of certain types of genomic aberrations within tumour genomes, such as copy neutral loss of heterozygosity and the inference of their clonal history (Van Loo et al., 2010; Chen et al., 2015).

In this chapter, measurements of allele-specific copy number profiling for the METABRIC cohort, using ASCAT (Van Loo et al., 2010), is presented and a modelling framework is proposed for the detection and classification of changepoints in observed allele-specific copy number profiles, with a simulation study carried out to assess modelling approaches. These analyses aim to capture large scale changes in CNA state that occur on individual alleles, providing information on CNA events that may occur preferentially in certain genomic regions and whose downstream influence requires more investigation.

5.1 Allele-specific Copy Number Profiling using ASCAT

A wide range of software is available facilitating measurement of allele-specific copy number, such as ASCAT (Van Loo et al., 2010), Tumor Aberration Prediction Suite (TAPS) (Rasmussen et al., 2011), Parent-Specific-Copy-Number (PSCN) (Olshen et al., 2011), Patchwork (Mayrhofer et al., 2013), Falcon (Chen et al., 2015), Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing (FACETS) (Shen and Seshan, 2016) and SPICE-pipeline (Ciani et al., 2022). Giving consideration to software packages available for analysis of microarray data, specifically Affymetrix SNP6 data without matched normals, the type of output, and quality of documentation, ASCAT was deemed as most suitable for allele-specific copy number calling in this study.

The ASCAT algorithm was applied first in Van Loo et al. (2010), to produce genome-wide allele-specific copy number profiles (ASCAT profiles) for 91/112 breast carcinoma samples genotyped on Illumina 109K SNP arrays. The paper reported that both the estimated percentages of aberrant cells and tumour ploidy were significantly different across the breast cancer subtypes, with the Luminal A subtype displaying the highest percentage of aberrant tumour cells and lowest ploidy. The frequency of gains and losses observed across the genome closely matched with previously reported patterns, but when stratifying by molecular subtype higher frequencies of gains and losses were observed in HER2 and Normal-like subtypes than reported previously. LOH was most frequently observed on chromosome arms 8p, 11q, 16q, and 17p, while copy number-neutral events were observed across the genome with a frequency of 20% or higher. In addition, genomic regions with higher frequencies of losses were more likely to contain copy number-neutral events, this was particularly evident on chromosomes 1p, 2, 3, 4q, 9q, 15, and 19p.

As discussed in Section 2.2.8, Pladsen et al. (2020) utilised allele-specific copy number profiles, generated using ASCAT, to produce six metrics: AMP, DEL, STP, CRV, LOH, and ASM. The AMP, DEL, STP and CRV metrics were calculated using the sum of the allele-specific copy number, and all six metrics were combined into two prognostic indices, CPI and $\text{CPI}_{\text{weighted}}$. Notably, combining the metrics into a single index results in loss of valuable information, specifically the type of CNA observed and which allele the CNA is observed on. Other studies utilising ASCAT include Cutcutache et al. (2016); Pereira et al. (2016); Steele et al. (2022); Tao et al. (2023) and Glodzik et al. (2023).

5.2 Generation of Allele-specific Copy Number Profiles for METABRIC Patients

To produce allele-specific copy number profiles from Affymetrix SNP 6.0 arrays, preprocessing using PennCNV (v1.0.5), a software tool to detect any form of copy number variation, is applied to generate signal intensity files, followed by ASCAT (v3.0.0), allele-specific copy number calling on these signal intensity files.

5.2.1 PennCNV

PennCNV (Wang et al., 2007) infers copy number calls for individual genotyped samples measured on genotyping arrays, such as Illumina and Affymetrix arrays, using a hidden Markov model, integrating multiple sources of information.

PennCNV is applied to 1,992 Affymetrix SNP 6.0 CEL files, from the METABRIC study (study accession EGAS00000000083) (Curtis et al., 2012; Lappalainen et al., 2015), to generate signal intensity files. Two signal intensities of particular importance, as they serve as input variables to ASCAT, are Log R Ratio (LRR), the normalised measure of total signal intensity, and B-Allele Frequency (BAF), the normalised measure of relative signal intensity ratio of the B and A alleles. Generating the cross-marker normalised signal intensity data from these Affymetrix SNP 6.0 CEL files is a multi-step process that can be carried out using substep 1.1 to substep 1.4 of the PennCNV-Affy pipeline. These substeps include generating genotyping calls from the CEL files, allele-specific signal extraction, generating the canonical genotype clustering file and LRR and BAF calculation. When preprocessing the Affymetrix SNP 6.0 CEL files, for input into ASCAT, the PennCNV pipeline recommended by the creators of ASCAT, and subsequently applied here, contains three substeps, substeps 1.1, 1.2 and 1.4. More detailed information can be found in the comprehensive guide to using PennCNV-Affy at: https://penncnv.openbioinformatics.org/en/latest/user-guide/affy/.

An example of the output obtained from the 3-substep PennCNV-Affy pipeline is provided in Table 5.1, where rows display SNP/CN probes, and for each probe, columns provide the genomic location (chromosome and position) and the LRR and BAF values for samples.

Table 5.1: The first 15 rows of the output obtained from the PennCNV-Affy pipeline. Only 9 columns are shown, corresponding to the first 3 METABRIC samples.

Name	Chr	Position	MB.0000.LRR	MB.0000.BAF	MB.0002.LRR	MB.0002.BAF	MB.0005.LRR	MB.0005.BAF
SNP_A-2131660	1	1156131	0.0784	0.0645	0.3256	0.4680	-0.0204	0.6749
SNP_A-1967418	1	2234251	-0.4757	0.8708	-0.0664	0.9642	0.4691	0.9017
SNP_A-1969580	1	2329564	-0.1603	0.9783	0.0210	0.8927	-0.0135	0.8405
SNP_A-4263484	1	2553624	-0.2693	0.9257	-0.1307	0.1196	-0.1347	0.0278
SNP_A-1978185	1	2936870	0.2567	0.0000	0.4048	0.0000	0.5406	0.0390
SNP_A-4264431	1	2951834	-0.0046	0.9343	0.5220	0.4813	-0.0062	0.8009
SNP_A-1980898	1	3095126	0.3309	1.0000	0.5433	1.0000	0.5810	1.0000
SNP_A-1983139	1	3165267	0.1509	0.0717	0.0452	0.0578	-0.4310	0.0693
SNP_A-4265735	1	3302871	0.0285	0.0227	0.2208	0.5460	-0.2972	0.2305
SNP_A-1995832	1	3705226	0.1941	0.5224	0.0320	0.5746	0.3774	0.6883
SNP_A-1995893	1	3720965	0.1675	0.5941	0.2923	0.9567	0.4228	0.7709
SNP_A-1997689	1	3763164	-0.4370	0.9560	-0.3390	0.8172	-0.3035	0.8170
SNP_A-1997709	1	3763567	-0.1209	0.0034	-0.3431	0.0792	-0.5607	0.0846
SNP_A-1997896	1	3766240	0.1209	0.0107	0.0167	0.0702	0.3122	0.6201
SNP_A-1997922	1	3766286	-0.0838	0.0018	-0.0365	0.9161	-0.5464	0.0287

5.2.2 ASCAT

ASCAT, applicable to Illumina SNP arrays, Affymetrix SNP arrays and high-throughput sequencing data, is used to derive allele-specific copy number profiles of tumour cells, accounting for non-aberrant cell admixture and aneuploidy. ASCAT produces profiles that map the distribution of allele-specific gains and losses, revealing LOH and copy number-neutral events, across the genome (Van Loo et al., 2010). ASCAT is available as an R package and can be found on GitHub at: https://github.com/VanLoo-lab/ascat.

ASCAT can be run in a number of different ways such as with or without matched normals or with or without a logR correction step. The ASCAT pipeline implemented here is for Affymetrix Genome-Wide Human SNP Array 6.0 CEL files, without matched normals, with a logR correction (GC content and replication timing) and where all samples are from females, adapted from the `ASCAT_fromCELfiles.R` script available on GitHub. Implementing this pipeline 1,984 ASCAT profiles were generated from 1,992 samples.

Outputs obtained from the ASCAT pipeline include a dataframe containing the copy number segments of each sample, an example of which is provided in Table 5.2, allele-specific copy number profiles for each sample (Figure 5.1), together with quality control metrics from ASCAT profiles including information on the purity and ploidy of each sample. Table 5.1 shows the copy numbers for each allele, nMajor and nMinor, which take on values 0 (indicating deletion), 1 (indicating normal copy number of allele), or any positive whole number greater than 1 (indicating increasing levels of amplifications). In Figure 5.1 the Minor and Major allele are coloured blue and red, respectively. The allele-specific copy number profile illustrated in Figure 5.1A, observed for sample MB.0000, provides an example of a tumour sample that has very little GI, as very few chromosomes contain CNAs. Samples such as this are labelled as "non-aberrant". The allele-specific copy number profile observed for sample MB.0025 provides an example of a tumour sample with moderate GI (Figure 5.1B). In the observed sample, some chromosomes contain no CNAs (chromosomes 4, 9 and 13), while others contain numerous CNAs (chromosomes 1, 3, 17). The amplification and deletion observed separately on chromosome 3 are examples of genomic aberrations that would not be detected in total CNA data, as the total copy number is 2. The allele-specific copy number profile observed for sample MB.0062 provides an example of a tumour sample with widespread GI, the majority of chromosomes contain at least one CNA and there are high levels of fluctuations between the CNA states (Figure 5.1C).

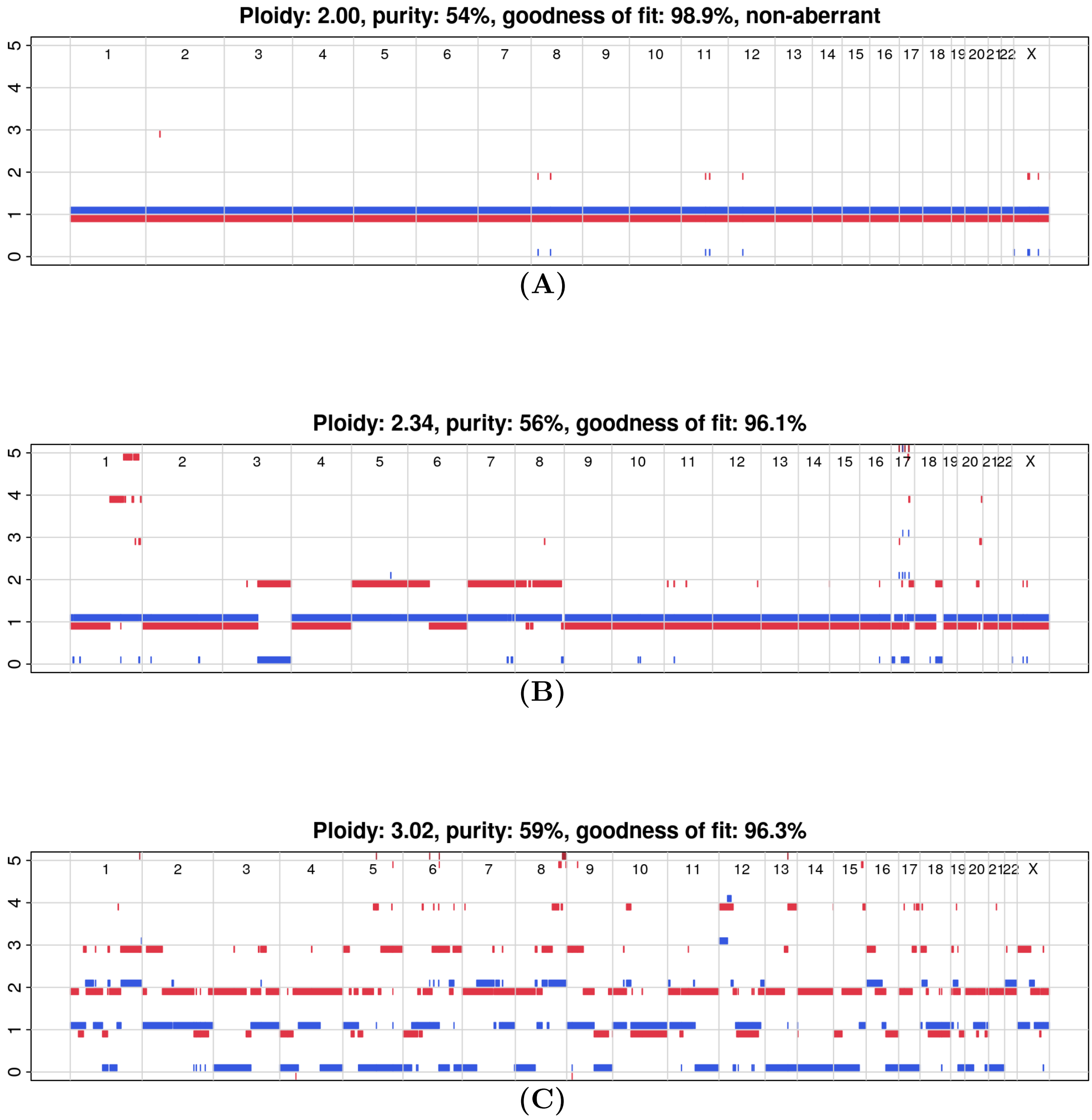


Figure 5.1: Allele-specific copy number profiles of (A) sample MB.0000, (B) sample MB.0025 and (C) sample MB.0062. Ploidy is defined as the amount of DNA relative to a haploid genome and purity is defined as the percentage of tumour cells. These ploidy and purity estimates are generated by creating a grid of possible values and evaluating the goodness-of-fit for both parameters (Van Loo et al., 2010).

5.2.3 Reformatting ASCAT Output for Downstream Analysis

It is useful to recode the copy numbers so they correspond to deletion, neutral and amplification states. In this case any probe that has a copy number greater than 2 will be assigned a 2, resulting in only three possible copy number values, 0, 1 and 2. Table 5.2 shows that after reformatting, the copy number of each allele, nMajorRF and nMinorRF, is bound in the range [0-2].

Table 5.2: The first 20 rows of the ASCAT segments file containing the allele-specific copy number calls for each sample. nMajor and nMinor refer to the copy numbers of the Major and Minor allele, while nMajorRF and nMinorRF refer to the reformatted copy numbers of the Major and Minor allele.

sample	chr	startpos	endpos	nMajor	nMinor	nMajorRF	nMinorRF
MB.0000	1	61735	152555527	1	1	1	1
MB.0000	1	152555706	152586540	0	0	0	0
MB.0000	1	152586576	152761923	1	1	1	1
MB.0000	1	152761939	152768700	0	0	0	0
MB.0000	1	152773905	249224388	1	1	1	1
MB.0000	2	12784	32630548	1	1	1	1
MB.0000	2	32635284	33331778	3	1	2	1
MB.0000	2	33333871	243089456	1	1	1	1
MB.0000	3	60345	197896118	1	1	1	1
MB.0000	4	12281	191027923	1	1	1	1
MB.0000	5	15532	180790320	1	1	1	1
MB.0000	6	149661	171051005	1	1	1	1
MB.0000	7	43259	159127004	1	1	1	1
MB.0000	8	31254	10555654	1	1	1	1
MB.0000	8	10555762	11310012	2	0	2	0
MB.0000	8	11322997	11701198	1	0	1	0
MB.0000	8	11701253	41811521	1	1	1	1
MB.0000	8	41811769	49326113	2	0	2	0
MB.0000	8	49328524	51634513	2	1	2	1
MB.0000	8	51634751	146298155	1	1	1	1

5.3 Classification of Changepoints in Allele-specific CNA Profiles

Our aim here is to model the allele-specific copy number profile across the genome by modelling features of common changes in the profile by detection of changepoints, also known as breakpoints, in the allele-specific CNA profiles. We define a copy number changepoint as a point along a chromosome, corresponding to an individual allele, where there is a change in copy number, i.e. if there is a copy number change from 2 to 0, there exists a point along that chromosome where the change has occurred (Figure 5.2). Figure 5.2, an annotated allele-specific profile, displays the different copy number states, Neutral (N) for a copy number of 1, Deletion (D) for a copy number of 0 and Amplification (A) for a copy number of 2. This figure highlights a copy number changepoint (CP) where the CNA state has gone from A to D. Based on the copy number either side of the changepoint we categorise each changepoint as either Neut/Amp (1 to 2), Neut/Del (1 to 0), Amp/Neut (2 to 1), Del/Neut (0 to 1), Amp/Del (2 to 0) or Del/Amp (0 to 2). Where no changepoint occurs, the category is labelled NoChangepoint. The detection of changepoints is based on length of the alteration segment to the left of the changepoint (TS) and the length of the alteration segment to the right of the changepoint (TE). For each copy number changepoint, for each sample, we record the changepoint location (chromosome, genomic position and allele), the TS and TE lengths, and the assigned category (Table 5.3). In the case where there is a gap between the segments, and as such between the copy number changes, the midpoint of the segment is used as the genomic position of the changepoint. Summary statistics of the TS and TE lengths for each category in the ASCAT data are shown in Tables 5.4 and 5.5.

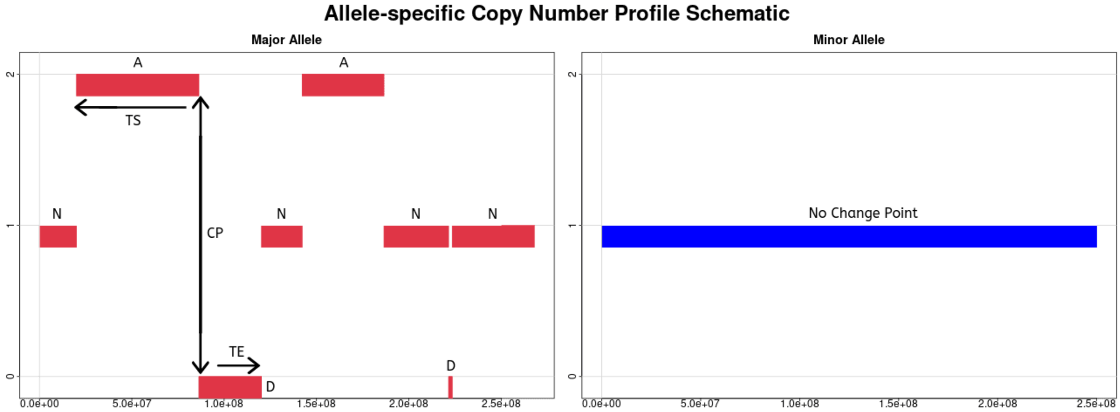


Figure 5.2: Copy number profiles of the Major and Minor alleles denoting changepoints, alteration states and TS/TE lengths.

Further examples of simulated allele-specific copy number profiles are provided in Figure 5.3. Profile A illustrates both the Major and Minor allele having a copy number of 1 across the observed region, both alleles are categorised as NoChangepoint. Profile B illustrates the Major allele having an amplified segment flanked by two neutral segments, categorised as Neut/Amp and Amp/Neut, and the Minor allele categorised as NoChangepoint. Profile C illustrates the Major allele having an amplified segment followed by a deleted segment, what we define as an Amp/Del flashpoint pattern, flanked by two neutral segments, categorised as Neut/Amp, Amp/Del and Del/Neut, and the Minor allele categorised as NoChangepoint. Profile D illustrates the Major allele displaying an Amp/Del flashpoint pattern, flanked by two neutral segments, and the Minor allele displaying an oscillating pattern of deleted and neutral segments, categorised as Neut/Del and Del/Neut.

Table 5.3: First 15 rows of the reformatted ASCAT segments file containing the allele-specific copy number calls for each sample. TS and TE displayed in bases.

Sample	Chr	Changepoint	Allele	TS	TE	Category
MB.0000	1	152555616.5	Major	0	30941.5	Neut/Del
MB.0000	1	152586558	Major	30941.5	0	Del/Neut
MB.0000	1	152761931	Major	0	9371.5	Neut/Del
MB.0000	1	152771302.5	Major	9371.5	0	Del/Neut
MB.0000	2	32632916	Major	0	699908.5	Neut/Amp
MB.0000	2	33332824.5	Major	699908.5	0	Amp/Neut
MB.0000	3		Major	0	0	NoChangepoint
MB.0000	4		Major	0	0	NoChangepoint
MB.0000	5		Major	0	0	NoChangepoint
MB.0000	6		Major	0	0	NoChangepoint
MB.0000	7		Major	0	0	NoChangepoint
MB.0000	8	10555708	Major	0	760796.5	Neut/Amp
MB.0000	8	11316504.5	Major	760796.5	0	Amp/Neut
MB.0000	8	41811645	Major	0	9822987	Neut/Amp
MB.0000	8	51634632	Major	9822987	0	Amp/Neut

Table 5.4: Summary statistics of the ASCAT segment kilobase lengths where the length of the neutral segments are set to 0. n refers to the number of changepoints.

Summary Statistics of Segment Kilobase Lengths							
		TS			TE		
Category	n	mean	median	sd	mean	median	sd
NoChangepoint	50,876	0.00	0.00	0.00	0.00	0.00	0.00
Neut/Amp	36,987	0.00	0.00	0.00	17,389.76	2,991.80	32,087.38
Neut/Del	37,178	0.00	0.00	0.00	10,066.80	1,894.52	20,992.81
Amp/Neut	35,274	13,107.90	2,693.65	25,539.58	0.00	0.00	0.00
Del/Neut	36,998	7,927.00	1,874.60	16,355.71	0.00	0.00	0.00
Amp/Del	9,657	28,164.46	10,199.99	38,383.29	11,611.34	1,556.43	23,121.62
Del/Amp	9,513	8,873.94	1,501.95	18,248.95	34,969.62	11,638.10	45,777.10

Table 5.5: Summary statistics of the ASCAT segment kilobase lengths where the length of the neutral segments are recorded as greater than 0. n refers to the number of changepoints.

Summary Statistics of Segment Kilobase Lengths							
		TS			TE		
Category	n	mean	median	sd	mean	median	sd
NoChangepoint	50,876	0.00	0.00	0.00	0.00	0.00	0.00
Neut/Amp	36,987	27,641.08	12,194.52	35,853.79	17,389.76	2,991.80	32,087.38
Neut/Del	37,178	31,128.85	15,232.80	38,124.67	10,066.80	1,894.52	20,992.81
Amp/Neut	35,274	13,107.90	2,693.65	25,539.58	27,983.83	11,654.74	36,586.19
Del/Neut	36,998	7,927.00	1,874.60	16,355.71	36,775.67	17,536.79	44,674.59
Amp/Del	9,657	28,164.46	10,199.99	38,383.29	11,611.34	1,556.43	23,121.62
Del/Amp	9,513	8,873.94	1,501.95	18,248.95	34,969.62	11,638.10	45,777.10

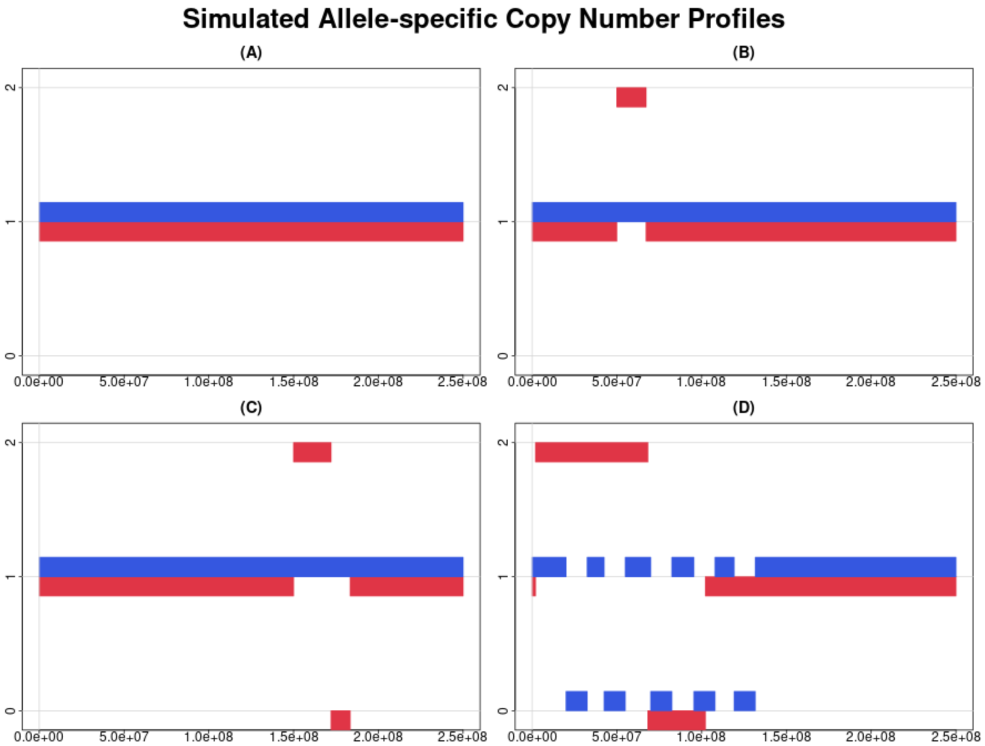


Figure 5.3: Simulated allele-specific copy number profiles. Minor allele in blue and Major allele in red.

5.4 Proposed Models for Changepoints in Allele-specific CNA Profiles

Our modelling approach has two aims: to enable the detection of changepoints and to model the features of changepoints across the observed allele-specific CNA profile, with two alleles labelled as Major and Minor, for all observed patients.

Given a specified interval of observation d where a changepoint is observed in an allele, the type of changepoint (Category) is recorded as one of six categories: Neut/Amp, Neut/Del, Amp/Neut, Del/Neut, Amp/Del and Del/Amp. The features of the observed changepoint are recorded as two continuous response variables TS and TE, representing the lengths of the segments to the left (TS) and to the right (TE) of the changepoint. These lengths are recorded irrespective of the interval boundaries, i.e. lengths $> d$ may be observed.

Where the category contains a Neutral flanking segment (Neut/Amp, Neut/Del, Amp/Neut, Del/Neut) this length can be recorded or set to 0. For the specified interval of observation, if no changepoints of any kind are observed in the interval, the category NoChangepoint can be recorded, within which $TS = TE = 0$.

5.4.1 Allele Independent (AI) Model

Treating the information from both alleles as independent, the following models the average TS length feature, the first response variable, given the type of changepoint:

$$\begin{aligned} TS_{ij} = \beta_0 + \beta_1 NeutAmp_{ij} + \beta_2 NeutDel_{ij} + \beta_3 AmpNeut_{ij} + \\ \beta_4 DelNeut_{ij} + \beta_5 AmpDel_{ij} + \beta_6 DelAmp_{ij} + \epsilon_{ij} \end{aligned} \tag{5.1}$$

For an observed interval d, where there are n_1 number of changepoints on the Major allele for individual i and n_2 number of changepoints on the Minor allele for individual i, the information on both alleles for an individual are included as independent observations, so that $j \in [1 : (n_1 + n_2)]$ for individual i.

Similarly, for each of the changepoints j, the average TE length feature, the second response variable, for each type of changepoint is specified as:

$$\begin{aligned} TE_{ij} = \beta_0 + \beta_1 NeutAmp_{ij} + \beta_2 NeutDel_{ij} + \beta_3 AmpNeut_{ij} + \\ \beta_4 DelNeut_{ij} + \beta_5 AmpDel_{ij} + \beta_6 DelAmp_{ij} + \epsilon_{ij} \end{aligned} \tag{5.2}$$

In the above intercept model specification, where regions containing no changepoints are observed in the observation range for individual i, and the NoChangepoint observations are included in the data, the intercept, β_0, serves as the baseline category (NoChangepoint). Since the mean TS and TE lengths for the baseline group, NoChangepoint, are equal to 0, the coefficients of each of the six changepoint categories correspond directly to the mean response lengths for those categories.

Alternatively, where the NoChangepoint observations are excluded from the observed dataset, the intercept is removed from the model to maintain this interpretation and the models are specified as:

$$\begin{aligned} TS_{ij} = \beta_1 NeutAmp_{ij} + \beta_2 NeutDel_{ij} + \beta_3 AmpNeut_{ij} + \\ \beta_4 DelNeut_{ij} + \beta_5 AmpDel_{ij} + \beta_6 DelAmp_{ij} + \epsilon_{ij} \end{aligned} \tag{5.3}$$

$$\begin{aligned} TE_{ij} = \beta_1 NeutAmp_{ij} + \beta_2 NeutDel_{ij} + \beta_3 AmpNeut_{ij} + \\ \beta_4 DelNeut_{ij} + \beta_5 AmpDel_{ij} + \beta_6 DelAmp_{ij} + \epsilon_{ij} \end{aligned} \tag{5.4}$$

We refer to this as the Allele-Independent Non-Intercept Model (AINIM) and the former as the Allele-Independent Intercept Model (AIIM).

For both AINIM and AIIM, TS and TE are also jointly modelled using the multivariate response vector, $Y_{ij} = (TS_{ij}, TE_{ij})$, which captures the covariance structure between the two response variables, $\Sigma = \sigma^2 I$.

The assumptions of the proposed linear models include linearity, normality for any fixed value of the predictor variable, homoscedasticity, and independence.

These models were fit using two R functions, `lm()` from the stats package (R Core Team, 2023) and `MCMCglmm()` from the MCMCglmm package (Hadfield, 2010). While the `lm()` function can be used to fit basic linear regression models, `MCMCglmm()` (with default priors) should produce similar results but provides more flexibility.

5.4.1.1 Illustration of Allele Independent (AI) Model

The AI Models, AIIM (Equations 5.1 and 5.2) and AINIM (Equations 5.3 and 5.4), are fitted to a simulated dataset using the two response variables in a univariate approach and multivariate approach. The simulated dataset is simulated as n = 20 patient tumour samples, of which 20% have allele-specific copy number profile A, 40% have allele-specific copy number profile C and 40% have allele-specific copy number profile D. As a result, samples 1 to 4 have profile A, samples 5 to 12 have profile C, and samples 13 to 20 have profile D (Table 5.7). For all samples, the lengths of the neutral segments, in addition to the lengths of the amplified and deleted segments, are simulated from a truncated Normal distribution with parameters specified in Table 5.6.

Table 5.6: Parameters of truncated Normal distributions used to simulate segment length and properties of simulated data. a and b correspond to the lower and upper bound.

Truncated Normal Distribution Parameters and Properties			
	Major Allele	**Minor Allele**	**Properties**
Profile A	No Breakpoint = 0	No Breakpoint = 0	P_A= 20%
Profile C	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 22,777, σ = 35,235, a = 1, b = 250,000) Del ~ TN(μ = 9,769, σ = 19,739, a = 1, b = 250,000)	No Breakpoint = 0	P_C = 40%
Profile D	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 68,331, σ = 35,235, a = 1, b = 250,000) Del ~ TN(μ = 29,307, σ = 19,739, a = 1, b = 250,000)	Neutral ~ TN(μ = 31,129, σ = 38,125, a = 1, b = 250,000) Del ~ TN(μ = 8,997, σ = 18,675, a = 1, b = 250,000)	P_D = 40%

Table 5.7A provides a snapshot of simulated data values where lengths of neutral segments are set to 0, while Table 5.7B provides information on simulated data values where lengths of neutral segments are retained as greater than 0. Table 5.8 provides summary statistics for the full simulated datasets. From these tables we can see the sample size, mean, median and standard deviation of the TS and TE variables corresponding to the different categories of changepoints. These tables highlight that each simulated sample contributes at least two changepoint observations to the dataset and that the number of changepoints observed is dependent on the cumulative length of the region.

Table 5.7: Structure of single simulated dataset. Simulated sample 1, sample 5 and sample 13, displaying the possible allele-specific copy number profiles, are shown.

(A) Dataset where neutral segment length recorded as length 0.

Example of Possible Simulated Samples					
Sample	**Category**	**Allele**	**Changepoint**	**TS**	**TE**
Sample 1	NoChangepoint	Major	NA	0	0
Sample 1	NoChangepoint	Minor	NA	0	0
Sample 5	Neut/Amp	Major	11,721	0	16,834
Sample 5	Amp/Del	Major	28,555	16,834	37,629
Sample 5	Del/Neut	Major	66,184	37,629	0
Sample 5	NoChangepoint	Minor	NA	0	0
Sample 13	Neut/Amp	Major	11,721	0	62,388
Sample 13	Amp/Del	Major	74,109	62,388	57,167
Sample 13	Del/Neut	Major	131,276	57,167	0
Sample 13	Neut/Del	Minor	13,831	0	7,165
Sample 13	Del/Neut	Minor	20,996	7,165	0
Sample 13	Neut/Del	Minor	49,893	0	15,046
Sample 13	Del/Neut	Minor	64,939	15,046	0
Sample 13	Neut/Del	Minor	121,281	0	18,587
Sample 13	Del/Neut	Minor	139,868	18,587	0
Sample 13	Neut/Del	Minor	143,139	0	31,489
Sample 13	Del/Neut	Minor	174,628	31,489	0

(B) Dataset where neutral segment lengths are retained.

Example of Possible Simulated Samples					
Sample	**Category**	**Allele**	**Changepoint**	**TS**	**TE**
Sample 1	NoChangepoint	Major	NA	0	0
Sample 1	NoChangepoint	Minor	NA	0	0
Sample 5	Neut/Amp	Major	11,721	11,720	16,834
Sample 5	Amp/Del	Major	28,555	16,834	37,629
Sample 5	Del/Neut	Major	66,184	37,629	183,816
Sample 5	NoChangepoint	Minor	NA	0	0
Sample 13	Neut/Amp	Major	11,721	11,720	62,388
Sample 13	Amp/Del	Major	74,109	62,388	57,167
Sample 13	Del/Neut	Major	131,276	57,167	118,724
Sample 13	Neut/Del	Minor	13,831	13,830	7,165
Sample 13	Del/Neut	Minor	20,996	7,165	28,897
Sample 13	Neut/Del	Minor	49,893	28,897	15,046
Sample 13	Del/Neut	Minor	64,939	15,046	56,342
Sample 13	Neut/Del	Minor	121,281	56,342	18,587
Sample 13	Del/Neut	Minor	139,868	18,587	3,271
Sample 13	Neut/Del	Minor	143,139	3,271	31,489
Sample 13	Del/Neut	Minor	174,628	31,489	75,372

Table 5.8: Summary statistics by category of the simulated dataset. In (A) the lengths of the neutral segments are recorded as length 0 and in (B) the lengths of the neutral segments are retained as length greater than 0.

(A) Dataset where neutral segment length recorded as length 0.

Summary Statistics for Simulated Dataset by Category							
		TS			TE		
Category	n	mean	median	sd	mean	median	sd
NoChangepoint	16	0.00	0.00	0.00	0.00	0.00	0.00
Neut/Amp	16	0.00	0.00	0.00	65,128.12	62,935.50	29,793.65
Neut/Del	31	0.00	0.00	0.00	18,026.29	16,923.00	8,384.48
Del/Neut	47	20,715.09	18,088.00	12,250.80	0.00	0.00	0.00
Amp/Del	16	65,128.12	62,935.50	29,793.65	25,924.62	26,401.00	16,606.11

(B) Dataset where neutral segment lengths are retained as length greater than 0.

Summary Statistics for Simulated Dataset by Category							
		TS			TE		
Category	n	mean	median	sd	mean	median	sd
NoChangepoint	16	0.00	0.00	0.00	0.00	0.00	0.00
Neut/Amp	16	36,903.69	22,390.50	27,352.69	65,128.12	62,935.50	29,793.65
Neut/Del	31	32,220.32	28,897.00	27,073.83	18,026.29	16,923.00	8,384.48
Del/Neut	47	20,715.09	18,088.00	12,250.80	70,450.91	63,241.00	51,771.99
Amp/Del	16	65,128.12	62,935.50	29,793.65	25,924.62	26,401.00	16,606.11

Fitting the univariate AIIM to the data, using the `lm()` function, produces model parameter estimates provided in Table 5.9A. Table 5.10A and Figure 5.4A summarise interval estimates for these parameters. Table 5.10A shows agreement between the parameter estimates and the mean lengths of the TS and TE recorded in Table 5.8 across all categories, indicating that our fitted models are estimating the parameters as intended. The parameter estimates and confidence intervals demonstrate that the mean lengths of TS for the Del/Neut and Amp/Del categories and the mean lengths of the TE for the Neut/Del, Neut/Amp and Amp/Del categories are all significantly greater than 0. Tables 5.9B, 5.10B and Figure 5.4B provide the model parameter and interval estimates produced using the dataset where the lengths of the neutral segments are retained as length greater than 0. The parameter estimates and confidence intervals demonstrate that the mean lengths of TS and TE for all categories, excluding the NoChangepoint category, are estimated to be significantly different from 0 (Table 5.10B and Figure 5.4B). Retaining the lengths of the neutral segments, which are often extremely variable in length, results in an increased variance within the dataset, leading to wider, less precise, interval estimates for all categories. Although these intervals are wider, it is important to note that apart from the detection of the neutral segments, the conclusions do not change, i.e. the categories detected as having mean length(s) significantly greater than 0 are the same.

Table 5.9: Univariate Allele-Independent Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	Direction	n	Beta	P
(Intercept)	TS	16	-0.000	1.000
CategoryAmp/Del	TS	16	6.513×10^{4}	1.184×10^{-27}
CategoryDel/Neut	TS	47	2.072×10^{4}	1.823×10^{-7}
CategoryNeut/Amp	TS	16	-1.770×10^{-12}	1.000×10^{0}
CategoryNeut/Del	TS	31	-1.705×10^{-12}	1.000×10^{0}
(Intercept)	TE	16	1.361×10^{-11}	1.000×10^{0}
CategoryAmp/Del	TE	16	2.592×10^{4}	6.309×10^{-8}
CategoryDel/Neut	TE	47	-8.414×10^{-12}	1.000×10^{0}
CategoryNeut/Amp	TE	16	6.513×10^{4}	3.330×10^{-28}
CategoryNeut/Del	TE	31	1.803×10^{4}	1.024×10^{-5}

(B) Parameter Estimates

Coefficients	Direction	n	Beta	P
(Intercept)	TS	16	4.148×10^{-11}	1.000×10^{0}
CategoryAmp/Del	TS	16	6.513×10^{4}	1.382×10^{-14}
CategoryDel/Neut	TS	47	2.072×10^{4}	8.958×10^{-4}
CategoryNeut/Amp	TS	16	3.690×10^{4}	2.257×10^{-6}
CategoryNeut/Del	TS	31	3.222×10^{4}	2.127×10^{-6}
(Intercept)	TE	16	5.510×10^{-11}	1.000×10^{0}
CategoryAmp/Del	TE	16	2.592×10^{4}	3.486×10^{-2}
CategoryDel/Neut	TE	47	7.045×10^{4}	1.006×10^{-10}
CategoryNeut/Amp	TE	16	6.513×10^{4}	4.023×10^{-7}
CategoryNeut/Del	TE	31	1.803×10^{4}	9.089×10^{-2}

Table 5.10: Univariate Allele-Independent Intercept Model parameter estimates and intervals fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	0.000	-6.398×10^{3}	6.398×10^{3}
Amp/Del	16	TS	6.513×10^{4}	5.873×10^{4}	7.153×10^{4}
Del/Neut	47	TS	2.072×10^{4}	1.698×10^{4}	2.445×10^{4}
Neut/Amp	16	TS	-1.770×10^{-12}	-6.398×10^{3}	6.398×10^{3}
Neut/Del	31	TS	-1.705×10^{-12}	-4.596×10^{3}	4.596×10^{3}
NoChangepoint	16	TE	1.361×10^{-11}	-6.293×10^{3}	6.293×10^{3}
Amp/Del	16	TE	2.592×10^{4}	1.963×10^{4}	3.222×10^{4}
Del/Neut	47	TE	5.198×10^{-12}	-3.672×10^{3}	3.672×10^{3}
Neut/Amp	16	TE	6.513×10^{4}	5.884×10^{4}	7.142×10^{4}
Neut/Del	31	TE	1.803×10^{4}	1.351×10^{4}	2.255×10^{4}

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	4.148×10^{-11}	-1.040×10^{4}	1.040×10^{4}
Amp/Del	16	TS	6.513×10^{4}	5.473×10^{4}	7.553×10^{4}
Del/Neut	47	TS	2.072×10^{4}	1.465×10^{4}	2.678×10^{4}
Neut/Amp	16	TS	3.690×10^{4}	2.650×10^{4}	4.730×10^{4}
Neut/Del	31	TS	3.222×10^{4}	2.475×10^{4}	3.969×10^{4}
NoChangepoint	16	TE	5.510×10^{-11}	-1.701×10^{4}	1.701×10^{4}
Amp/Del	16	TE	2.592×10^{4}	8.918×10^{3}	4.293×10^{4}
Del/Neut	47	TE	7.045×10^{4}	6.053×10^{4}	8.037×10^{4}
Neut/Amp	16	TE	6.513×10^{4}	4.812×10^{4}	8.213×10^{4}
Neut/Del	31	TE	1.803×10^{4}	5.809×10^{3}	3.024×10^{4}

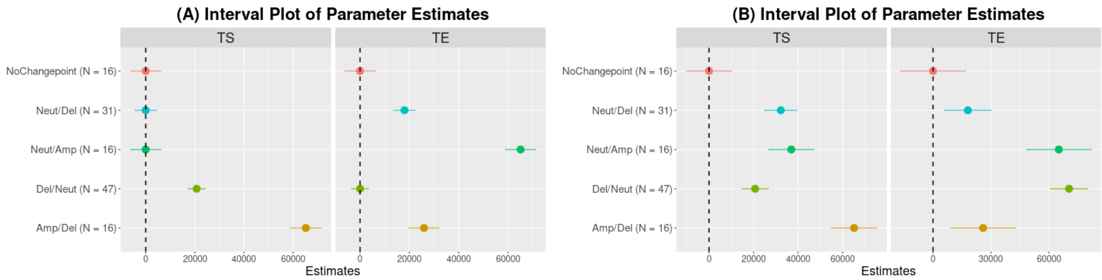


Figure 5.4: Interval plot of univariate Allele-Independent Intercept Model parameter estimates fitted using `lm`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Focusing on the application in which a variant of the dataset excludes all NoChangepoint observations, and fitting the univariate AINIM gives model estimates, as provided in Table 5.11, and confidence interval estimates (Table 5.12 and Figure 5.5).

The results of the AINIM are similar to the AIIM, with the parameter estimates corresponding to the mean lengths of the TS and TE recorded in Table 5.8 across all categories, and the interval plots, highlighting the categories where the lower bound of the confidence interval is greater than 0. Utilising the dataset where the neutral lengths are recorded as 0, the mean lengths of TS for the Del/Neut and Amp/Del categories and the mean lengths of the TE for the Neut/Del, Neut/Amp and Amp/Del categories are all significantly greater than 0. When utilising the dataset where the neutral lengths are retained as greater than 0, the mean lengths of TS and TE for all categories are significantly greater than 0. The main consequences of removing the NoChangepoint observations are the reduction in sample size and the increased width of the confidence intervals in the AINIM, compared to the AIIM, due to the removal of a large number of constant 0 values leading to an increase in variance in the dataset.

The results obtained for the univariate AIIM and AINIM, fitted using the `MCMCglmm()` function, show similar behaviour, with the same categories being highlighted as significant (Appendix E).

Table 5.11: Univariate Allele-Independent Non-Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	Direction	n	Beta	P
CategoryAmp/Del	TS	16	6.513×10^{4}	1.171×10^{-35}
CategoryDel/Neut	TS	47	2.072×10^{4}	1.286×10^{-17}
CategoryNeut/Amp	TS	16	1.364×10^{-12}	1.000×10^{0}
CategoryNeut/Del	TS	31	−0.000	1.000
CategoryAmp/Del	TE	16	2.592×10^{4}	1.053×10^{-11}
CategoryDel/Neut	TE	47	−0.000	1.000
CategoryNeut/Amp	TE	16	6.513×10^{4}	3.005×10^{-36}
CategoryNeut/Del	TE	31	1.803×10^{4}	3.576×10^{-11}

(B) Parameter Estimates

Coefficients	Direction	n	Beta	P
CategoryAmp/Del	TS	16	6.513×10^{4}	1.381×10^{-20}
CategoryDel/Neut	TS	47	2.072×10^{4}	6.138×10^{-9}
CategoryNeut/Amp	TS	16	3.690×10^{4}	1.884×10^{-9}
CategoryNeut/Del	TS	31	3.222×10^{4}	1.748×10^{-12}
CategoryAmp/Del	TE	16	2.592×10^{4}	5.655×10^{-3}
CategoryDel/Neut	TE	47	7.045×10^{4}	5.083×10^{-24}
CategoryNeut/Amp	TE	16	6.513×10^{4}	1.510×10^{-10}
CategoryNeut/Del	TE	31	1.803×10^{4}	7.335×10^{-3}

Table 5.12: Univariate Allele-Independent Non-Intercept Model parameter estimates and intervals fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
Amp/Del	16	TS	6.513×10^{4}	5.828×10^{4}	7.197×10^{4}
Del/Neut	47	TS	2.072×10^{4}	1.672×10^{4}	2.471×10^{4}
Neut/Amp	16	TS	1.364×10^{-12}	-6.845×10^{3}	6.845×10^{3}
Neut/Del	31	TS	0.000	-4.918×10^{3}	4.918×10^{3}
Amp/Del	16	TE	2.592×10^{4}	1.919×10^{4}	3.266×10^{4}
Del/Neut	47	TE	0.000	-3.928×10^{3}	3.928×10^{3}
Neut/Amp	16	TE	6.513×10^{4}	5.840×10^{4}	7.186×10^{4}
Neut/Del	31	TE	1.803×10^{4}	1.319×10^{4}	2.286×10^{4}

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
Amp/Del	16	TS	6.513×10^{4}	5.400×10^{4}	7.626×10^{4}
Del/Neut	47	TS	2.072×10^{4}	1.422×10^{4}	2.721×10^{4}
Neut/Amp	16	TS	3.690×10^{4}	2.578×10^{4}	4.803×10^{4}
Neut/Del	31	TS	3.222×10^{4}	2.423×10^{4}	4.021×10^{4}
Amp/Del	16	TE	2.592×10^{4}	7.729×10^{3}	4.412×10^{4}
Del/Neut	47	TE	7.045×10^{4}	5.983×10^{4}	8.107×10^{4}
Neut/Amp	16	TE	6.513×10^{4}	4.693×10^{4}	8.332×10^{4}
Neut/Del	31	TE	1.803×10^{4}	4.954×10^{3}	3.110×10^{4}

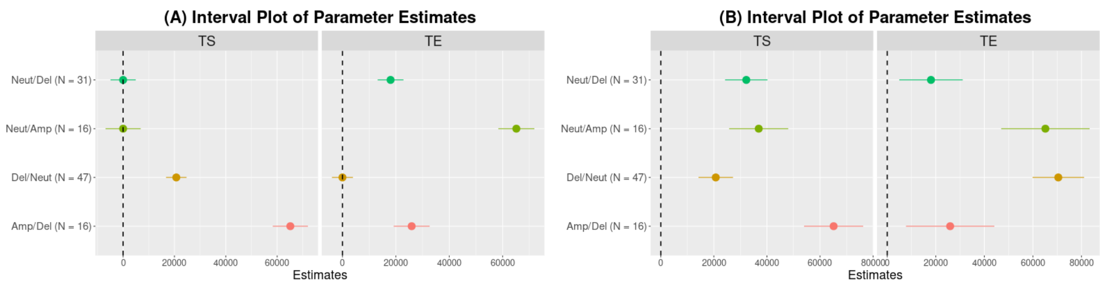


Figure 5.5: Interval plot of univariate Allele-Independent Non-Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Fitting the multivariate AIIM and AINIM to the data identifies a limitation of the `predict()` function previously used to produce confidence intervals (Tables 5.13-5.16 and Figures 5.6 and 5.7). Obtaining confidence intervals for multivariate models fitted with the `lm()` function is not supported. As a result, only point estimates are shown in Tables 5.13-5.16 and Figures 5.6 and 5.7, with these point estimates mirroring those obtained from the univariate models.

The `MCMCglmm()` function overcomes this limitation, and produces similar results, by enabling the production of confidence intervals for both the multivariate models using the `predict.MCMCglmm()` function (Appendix E).

Overall, when comparing model fits where the neutral segment lengths are retained as greater than 0 and where the neutral segment lengths recorded as 0, there is increased detection of the changepoints with a neutral length and an increase in the width of the confidence intervals. Although these intervals are wider, it is important to note that apart from the detection of the neutral segments, the conclusions do not change, i.e. the categories detected as having mean length(s) significantly greater than 0 are the same in both instances. As the primary interest here is in CNA changepoints, focus is given to the dataset where the neutral segment lengths are recorded as 0.

Notably, the `lm()` and `MCMCglmm()` functions perform similarly across the univariate AI models, but limitations in the `predict()` functions result in only the `predict.MCMCglmm()` function being capable of producing confidence intervals for multivariate models.

Table 5.13: Multivariate Allele-Independent Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	Direction	n	Beta	P
(Intercept)	TS	16	-0.000	1.000
CategoryAmp/Del	TS	16	6.513×10^{4}	1.184×10^{-27}
CategoryDel/Neut	TS	47	2.072×10^{4}	1.823×10^{-7}
CategoryNeut/Amp	TS	16	-1.770×10^{-12}	1.000×10^{0}
CategoryNeut/Del	TS	31	-1.705×10^{-12}	1.000×10^{0}
(Intercept)	TE	16	1.361×10^{-11}	1.000×10^{0}
CategoryAmp/Del	TE	16	2.592×10^{4}	6.309×10^{-8}
CategoryDel/Neut	TE	47	-8.414×10^{-12}	1.000×10^{0}
CategoryNeut/Amp	TE	16	6.513×10^{4}	3.330×10^{-28}
CategoryNeut/Del	TE	31	1.803×10^{4}	1.024×10^{-5}

(B) Parameter Estimates

Coefficients	Direction	n	Beta	P
(Intercept)	TS	16	4.148×10^{-11}	1.000×10^{0}
CategoryAmp/Del	TS	16	6.513×10^{4}	1.382×10^{-14}
CategoryDel/Neut	TS	47	2.072×10^{4}	8.958×10^{-4}
CategoryNeut/Amp	TS	16	3.690×10^{4}	2.257×10^{-6}
CategoryNeut/Del	TS	31	3.222×10^{4}	2.127×10^{-6}
(Intercept)	TE	16	5.510×10^{-11}	1.000×10^{0}
CategoryAmp/Del	TE	16	2.592×10^{4}	3.486×10^{-2}
CategoryDel/Neut	TE	47	7.045×10^{4}	1.006×10^{-10}
CategoryNeut/Amp	TE	16	6.513×10^{4}	4.023×10^{-7}
CategoryNeut/Del	TE	31	1.803×10^{4}	9.089×10^{-2}

Table 5.14: Multivariate Allele-Independent Intercept Model parameter estimates and intervals fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	0.000	NA	NA
Amp/Del	16	TS	6.513×10^{4}	NA	NA
Del/Neut	47	TS	2.072×10^{4}	NA	NA
Neut/Amp	16	TS	-1.770×10^{-12}	NA	NA
Neut/Del	31	TS	-1.705×10^{-12}	NA	NA
NoChangepoint	16	TE	1.361×10^{-11}	NA	NA
Amp/Del	16	TE	2.592×10^{4}	NA	NA
Del/Neut	47	TE	5.198×10^{-12}	NA	NA
Neut/Amp	16	TE	6.513×10^{4}	NA	NA
Neut/Del	31	TE	1.803×10^{4}	NA	NA

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	4.148×10^{-11}	NA	NA
Amp/Del	16	TS	6.513×10^{4}	NA	NA
Del/Neut	47	TS	2.072×10^{4}	NA	NA
Neut/Amp	16	TS	3.690×10^{4}	NA	NA
Neut/Del	31	TS	3.222×10^{4}	NA	NA
NoChangepoint	16	TE	5.510×10^{-11}	NA	NA
Amp/Del	16	TE	2.592×10^{4}	NA	NA
Del/Neut	47	TE	7.045×10^{4}	NA	NA
Neut/Amp	16	TE	6.513×10^{4}	NA	NA
Neut/Del	31	TE	1.803×10^{4}	NA	NA

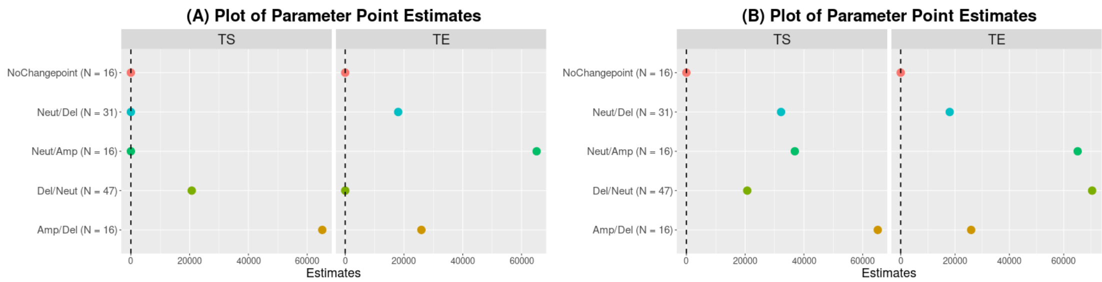


Figure 5.6: Plot of multivariate Allele-Independent Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Table 5.15: Multivariate Allele-Independent Non-Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	Direction	n	Beta	P
CategoryAmp/Del	TS	16	6.513×10^{4}	1.171×10^{-35}
CategoryDel/Neut	TS	47	2.072×10^{4}	1.286×10^{-17}
CategoryNeut/Amp	TS	16	1.364×10^{-12}	1.000×10^{0}
CategoryNeut/Del	TS	31	−0.000	1.000
CategoryAmp/Del	TE	16	2.592×10^{4}	1.053×10^{-11}
CategoryDel/Neut	TE	47	−0.000	1.000
CategoryNeut/Amp	TE	16	6.513×10^{4}	3.005×10^{-36}
CategoryNeut/Del	TE	31	1.803×10^{4}	3.576×10^{-11}

(B) Parameter Estimates

Coefficients	Direction	n	Beta	P
CategoryAmp/Del	TS	16	6.513×10^{4}	1.381×10^{-20}
CategoryDel/Neut	TS	47	2.072×10^{4}	6.138×10^{-9}
CategoryNeut/Amp	TS	16	3.690×10^{4}	1.884×10^{-9}
CategoryNeut/Del	TS	31	3.222×10^{4}	1.748×10^{-12}
CategoryAmp/Del	TE	16	2.592×10^{4}	5.655×10^{-3}
CategoryDel/Neut	TE	47	7.045×10^{4}	5.083×10^{-24}
CategoryNeut/Amp	TE	16	6.513×10^{4}	1.510×10^{-10}
CategoryNeut/Del	TE	31	1.803×10^{4}	7.335×10^{-3}

Table 5.16: Multivariate Allele-Independent Non-Intercept Model parameter estimates and intervals fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
Amp/Del	16	TS	6.513×10^4	NA	NA
Del/Neut	47	TS	2.072×10^4	NA	NA
Neut/Amp	16	TS	1.364×10^{-12}	NA	NA
Neut/Del	31	TS	0.000	NA	NA
Amp/Del	16	TE	2.592×10^4	NA	NA
Del/Neut	47	TE	0.000	NA	NA
Neut/Amp	16	TE	6.513×10^4	NA	NA
Neut/Del	31	TE	1.803×10^4	NA	NA

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
Amp/Del	16	TS	6.513×10^4	NA	NA
Del/Neut	47	TS	2.072×10^4	NA	NA
Neut/Amp	16	TS	3.690×10^4	NA	NA
Neut/Del	31	TS	3.222×10^4	NA	NA
Amp/Del	16	TE	2.592×10^4	NA	NA
Del/Neut	47	TE	7.045×10^4	NA	NA
Neut/Amp	16	TE	6.513×10^4	NA	NA
Neut/Del	31	TE	1.803×10^4	NA	NA

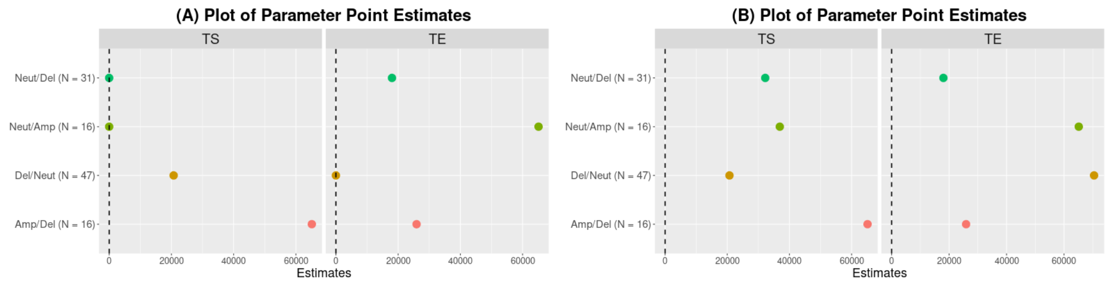


Figure 5.7: Plot of multivariate Allele-Independent Non-Intercept Model parameter estimates fitted using `lm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

5.4.2 Allele-Dependent (AD) Models

While these models are shown to be performing well in estimating the changepoint category features, the AI model does not consider the information regarding the specific allele, Major or Minor, on which the changepoint is observed. To allow flexibility in estimating allele-specific effects, we formulate an Allele-Dependent (AD) model framework, fitting an interaction term allowing for features of changepoints to be specific to the Major/Minor alleles, broadly speaking, $TS \sim Category + Allele + Category : Allele$.

The AD Intercept Model, ADIM, for the TS response variable is specified as:

$$\begin{aligned} TS_{ij} = \beta_0 &+ \beta_1 NeutAmp_{ij} + \beta_2 NeutDel_{ij} + \beta_3 AmpNeut_{ij} + \beta_4 DelNeut_{ij} + \\ &\beta_5 AmpDel_{ij} + \beta_6 DelAmp_{ij} + \beta_7 AlleleMinor_{ij} + \\ &\beta_8 NeutAmp_{ij} : AlleleMinor_{ij} + \beta_9 NeutDel_{ij} : AlleleMinor_{ij} + \\ &\beta_{10} AmpNeut_{ij} : AlleleMinor_{ij} + \beta_{11} DelNeut_{ij} : AlleleMinor_{ij} + \\ &\beta_{12} AmpDel_{ij} : AlleleMinor_{ij} + \beta_{13} DelAmp_{ij} : AlleleMinor_{ij} + \epsilon_{ij} \end{aligned} \tag{5.5}$$

where the term $AlleleMinor_{ij}$, corresponds to an indicator term with value 1 if the observed changepoint j comes from the Minor allele, and the estimated coefficient β_7 corresponds to the estimated difference in response length for the Minor allele compared to the Major allele, within the NoChangepoint category, β_0. For the

NoChangepoint category by definition, lengths are 0 for both alleles and therefore the difference between alleles is 0, $\beta_0 = \beta_7 = 0$. The ADIM for the TE response variable is the same as shown above.

The Non-Intercept specification of the AD model, ADNIM, for the TS response variable, omitting specification of $AlleleMinor_{ij}$, follows as:

$$\begin{aligned} TS_{ij} = {} & \beta_1 NeutAmp_{ij} + \beta_2 NeutDel_{ij} + \beta_3 AmpNeut_{ij} + \beta_4 DelNeut_{ij} + \\ & \beta_5 AmpDel_{ij} + \beta_6 DelAmp_{ij} + \beta_7 NeutAmp_{ij} : AlleleMinor_{ij} + \\ & \beta_8 NeutDel_{ij} : AlleleMinor_{ij} + \beta_9 AmpNeut_{ij} : AlleleMinor_{ij} + \\ & \beta_{10} DelNeut_{ij} : AlleleMinor_{ij} + \beta_{11} AmpDel_{ij} : AlleleMinor_{ij} + \\ & \beta_{12} DelAmp_{ij} : AlleleMinor_{ij} + \epsilon_{ij} \end{aligned} \tag{5.6}$$

Again, for both ADIM and ADNIM, TS and TE are also jointly modelled using the multivariate response vector, $Y_{ij} = (TS_{ij},\ TE_{ij})$, with the usual error term assumptions.

5.4.2.1 Illustration of Allele-Dependent (AD) Model

With application to the same dataset, the variant in which lengths of neutral segments are recorded as 0, Table 5.17 provides a summary of the data, differentiated by the allele on which the changepoint is observed.

Table 5.17: Summary statistics by category and allele of the simulated dataset.

Summary Statistics for Simulated Dataset by Allele and Category

		TS			TE		
	n	mean	median	sd	mean	median	sd
Category (Minor Allele)							
NoChangepoint	12	0.00	0.00	0.00	0.00	0.00	0.00
Neut/Del	31	0.00	0.00	0.00	18,026.29	16,923.00	8,384.48
Del/Neut	31	18,026.29	16,923.00	8,384.48	0.00	0.00	0.00
Category (Major Allele)							
NoChangepoint	4	0.00	0.00	0.00	0.00	0.00	0.00
Neut/Amp	16	0.00	0.00	0.00	65,128.12	62,935.50	29,793.65
Del/Neut	16	25,924.62	26,401.00	16,606.11	0.00	0.00	0.00
Amp/Del	16	65,128.12	62,935.50	29,793.65	25,924.62	26,401.00	16,606.11

Applying univariate ADIMs, for the two responses TS and TE, provides model parameter and interval estimates (Table 5.18 and Figure 5.8). Table 5.18 shows agreement between the parameter estimates and the mean lengths of the TS and TE recorded in Table 5.17, across all categories and alleles, indicating that our fitted univariate ADIMs seem to be estimating the parameters as intended. Table 5.18 and Figure 5.8 demonstrate that the mean lengths of TS for the Del/Neut and Amp/Del categories on the Major allele, the mean length of TS for the Del/Neut categories on the Minor allele, the mean lengths of TE for the Neut/Amp and Amp/Del categories on the Major allele and the mean lengths of TE for the Neut/Del categories on the Minor alleles, are all significantly greater than 0.

Table 5.18: Univariate Allele-Dependent Intercept Model estimates and intervals fitted using `lm()` and where neutral lengths are recorded as 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

Parameter Estimates and Confidence Intervals									
Coefficients	**Allele**	**Direction**	**n**	**Beta**	**P**	**Category**	**Fit**	**LB**	**UB**
(Intercept)	Major	TS	4	1.296×10^{-11}	1.000×10^{0}	NoChangepoint	1.296×10^{-11}	-1.269×10^{4}	1.269×10^{4}
CategoryAmp/Del	Major	TS	16	6.513×10^{4}	2.709×10^{-15}	Amp/Del	6.513×10^{4}	5.878×10^{4}	7.147×10^{4}
CategoryDel/Neut	Major	TS	16	2.592×10^{4}	4.382×10^{-4}	Del/Neut	2.592×10^{4}	1.958×10^{4}	3.227×10^{4}
CategoryNeut/Amp	Major	TS	16	-1.064×10^{-11}	1.000×10^{0}	Neut/Amp	2.324×10^{-12}	-6.347×10^{3}	6.347×10^{3}
AlleleMinor	Minor	TS	12	-1.186×10^{-11}	1.000×10^{0}	NoChangepoint	1.106×10^{-12}	-7.328×10^{3}	7.328×10^{3}
CategoryDel/Neut:AlleleMinor	Minor	TS	31	-7.898×10^{3}	NA	Del/Neut	1.803×10^{4}	1.347×10^{4}	2.259×10^{4}
CategoryNeut/Del:AlleleMinor	Minor	TS	31	NA	NA	Neut/Del	2.334×10^{-12}	-4.560×10^{3}	4.560×10^{3}
(Intercept)	Major	TE	4	1.620×10^{-11}	1.000×10^{0}	NoChangepoint	1.620×10^{-11}	-1.269×10^{4}	1.269×10^{4}
CategoryAmp/Del	Major	TE	16	2.592×10^{4}	4.382×10^{-4}	Amp/Del	2.592×10^{4}	1.958×10^{4}	3.227×10^{4}
CategoryDel/Neut	Major	TE	16	-8.414×10^{-12}	1.000×10^{0}	Del/Neut	7.791×10^{-12}	-6.347×10^{3}	6.347×10^{3}
CategoryNeut/Amp	Major	TE	16	6.513×10^{4}	2.709×10^{-15}	Neut/Amp	6.513×10^{4}	5.878×10^{4}	7.147×10^{4}
AlleleMinor	Minor	TE	12	-2.858×10^{-12}	1.000×10^{0}	NoChangepoint	1.335×10^{-11}	-7.328×10^{3}	7.328×10^{3}
CategoryDel/Neut:AlleleMinor	Minor	TE	31	3.034×10^{-12}	NA	Del/Neut	7.966×10^{-12}	-4.560×10^{3}	4.560×10^{3}
CategoryNeut/Del:AlleleMinor	Minor	TE	31	NA	NA	Neut/Del	1.803×10^{4}	1.347×10^{4}	2.259×10^{4}

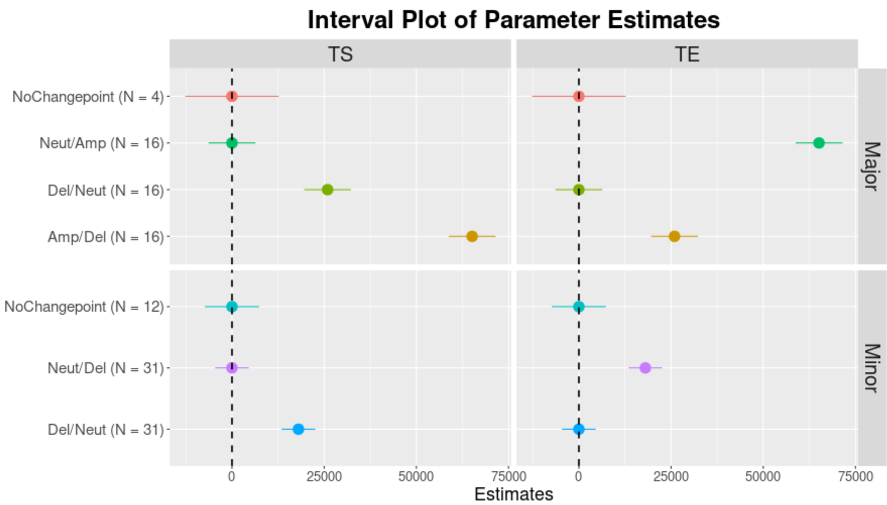


Figure 5.8: Interval plot of univariate Allele-Dependent Intercept Model parameter estimates fitted using `lm()` and where neutral lengths are recorded as 0.

Notably, the average of the parameter point estimate for the TS and TE lengths for each category is equal to the parameter point estimates in our AI models. This is expected as the average TS or TE length for each category should be equal, but the addition of the interaction term differentiates which allele the changepoint is observed on, providing more nuanced information about the average lengths of alterations on each allele. In addition, the AD models can highlight changepoints occurring preferentially on one allele, for example, Figure 5.8 indicates that in this dataset the Neut/Amp changepoints are only observed on the Major allele, a detail not provided in the AI models.

Applying univariate ADNIMs, for the two responses TS and TE, provides model parameter and interval estimates (Table 5.19 and Figure 5.9). Again, agreement between the parameter estimates and the mean lengths of the TS and TE recorded in Table 5.17 across all categories and alleles is observed. All categories containing a mean TS or TE length significantly greater than 0 on each allele are detected successfully, with the confidence intervals being slightly wider and less precise, for the ADNIMs compared to the ADIMs.

Table 5.19: Univariate Allele-Dependent Non-Intercept Model parameter estimates and intervals fitted using `lm()` and where neutral lengths are recorded as 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

Parameter Estimates and Confidence Intervals									
Coefficients	**Allele**	**Direction**	**n**	**Beta**	**P**	**Category**	**Fit**	**LB**	**UB**
CategoryAmp/Del	Major	TS	16	6.513×10^{4}	6.383×10^{-36}	Amp/Del	6.513×10^{4}	5.836×10^{4}	7.189×10^{4}
CategoryDel/Neut	Major	TS	16	2.592×10^{4}	1.317×10^{-11}	Del/Neut	2.592×10^{4}	1.916×10^{4}	3.269×10^{4}
CategoryNeut/Amp	Major	TS	16	8.154×10^{-13}	1.000×10^{0}	Neut/Amp	8.154×10^{-13}	-6.766×10^{3}	6.766×10^{3}
CategoryDel/Neut:AlleleMinor	Minor	TS	31	NA	NA	Del/Neut	1.803×10^{4}	1.317×10^{4}	2.289×10^{4}
CategoryNeut/Del:AlleleMinor	Minor	TS	31	NA	NA	Neut/Del	-9.095×10^{-13}	-4.861×10^{3}	4.861×10^{3}
CategoryAmp/Del	Major	TE	16	2.592×10^{4}	1.317×10^{-11}	Amp/Del	2.592×10^{4}	1.916×10^{4}	3.269×10^{4}
CategoryDel/Neut	Major	TE	16	-5.662×10^{-12}	1.000×10^{0}	Del/Neut	-5.662×10^{-12}	-6.766×10^{3}	6.766×10^{3}
CategoryNeut/Amp	Major	TE	16	6.513×10^{4}	6.383×10^{-36}	Neut/Amp	6.513×10^{4}	5.836×10^{4}	7.189×10^{4}
CategoryDel/Neut:AlleleMinor	Minor	TE	31	NA	NA	Del/Neut	2.922×10^{-12}	-4.861×10^{3}	4.861×10^{3}
CategoryNeut/Del:AlleleMinor	Minor	TE	31	NA	NA	Neut/Del	1.803×10^{4}	1.317×10^{4}	2.289×10^{4}

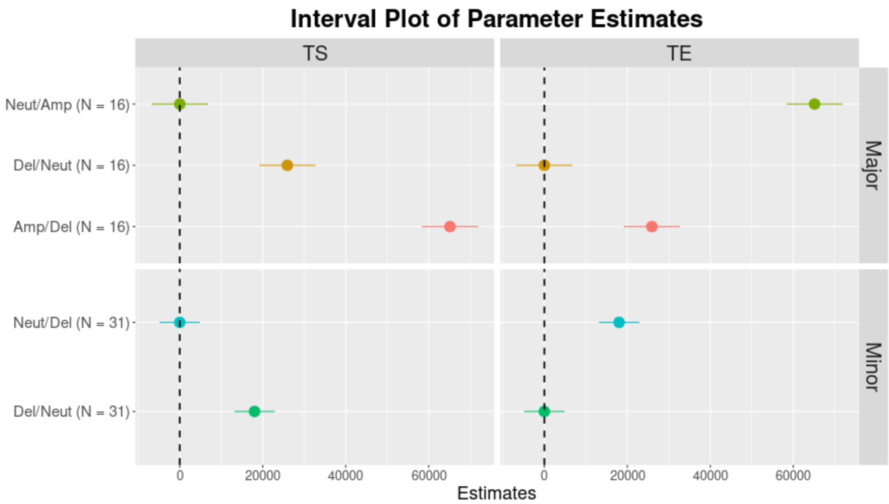


Figure 5.9: Interval plot of univariate Allele-Dependent Non-Intercept Model parameter estimates fitted using `lm()` and where neutral lengths are recorded as 0.

As expected the results of fitting the multivariate ADIM (Table 5.20 and Figure 5.10) and ADNIM (Table 5.21 and Figure 5.11) to the simulated data mirror those obtained from the univariate models, with the exception of the confidence intervals, which are absent in the models fitted using the `lm()` function. Multivariate ADIM and ADNIM, fitted using the `MCMCglmm()` function, are provided in Appendix E.

Overall, these models indicate that there is minimal difference in the categories detected using the Intercept and Non-Intercept models. However, due to sample size concerns in datasets where very few changepoints are present, the decision was made to include the NoChangepoint observations. In addition, these AD models provide valuable information regarding which allele the CNA changepoint has occurred on, highlighting instances where changepoint categories occur preferentially on one allele and instances where average lengths of alteration segments, TS and TE, may differ between alleles.

Table 5.20: Multivariate Allele-Dependent Intercept Model parameter estimates and intervals fitted using `lm()` and where neutral lengths are recorded as 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

Parameter Estimates and Confidence Intervals									
Coefficients	**Allele**	**Direction**	**n**	**Beta**	**P**	**Category**	**Fit**	**LB**	**UB**
(Intercept)	Major	TS	4	1.296×10^{-11}	1.000×10^{0}	NoChangepoint	1.296×10^{-11}	NA	NA
CategoryAmp/Del	Major	TS	16	6.513×10^{4}	2.709×10^{-15}	Amp/Del	6.513×10^{4}	NA	NA
CategoryDel/Neut	Major	TS	16	2.592×10^{4}	4.382×10^{-4}	Del/Neut	2.592×10^{4}	NA	NA
CategoryNeut/Amp	Major	TS	16	-1.064×10^{-11}	1.000×10^{0}	Neut/Amp	2.324×10^{-12}	NA	NA
AlleleMinor	Minor	TS	12	-1.186×10^{-11}	1.000×10^{0}	NoChangepoint	1.106×10^{-12}	NA	NA
CategoryDel/Neut:AlleleMinor	Minor	TS	31	-7.898×10^{3}	3.483×10^{-1}	Del/Neut	1.803×10^{4}	NA	NA
CategoryNeut/Del:AlleleMinor	Minor	TS	31	NA	NA	Neut/Del	2.334×10^{-12}	NA	NA
(Intercept)	Major	TE	4	1.620×10^{-11}	1.000×10^{0}	NoChangepoint	1.620×10^{-11}	NA	NA
CategoryAmp/Del	Major	TE	16	2.592×10^{4}	4.382×10^{-4}	Amp/Del	2.592×10^{4}	NA	NA
CategoryDel/Neut	Major	TE	16	-8.414×10^{-12}	1.000×10^{0}	Del/Neut	7.791×10^{-12}	NA	NA
CategoryNeut/Amp	Major	TE	16	6.513×10^{4}	2.709×10^{-15}	Neut/Amp	6.513×10^{4}	NA	NA
AlleleMinor	Minor	TE	12	-2.858×10^{-12}	1.000×10^{0}	NoChangepoint	1.335×10^{-11}	NA	NA
CategoryDel/Neut:AlleleMinor	Minor	TE	31	3.034×10^{-12}	1.000×10^{0}	Del/Neut	7.966×10^{-12}	NA	NA
CategoryNeut/Del:AlleleMinor	Minor	TE	31	NA	NA	Neut/Del	1.803×10^{4}	NA	NA

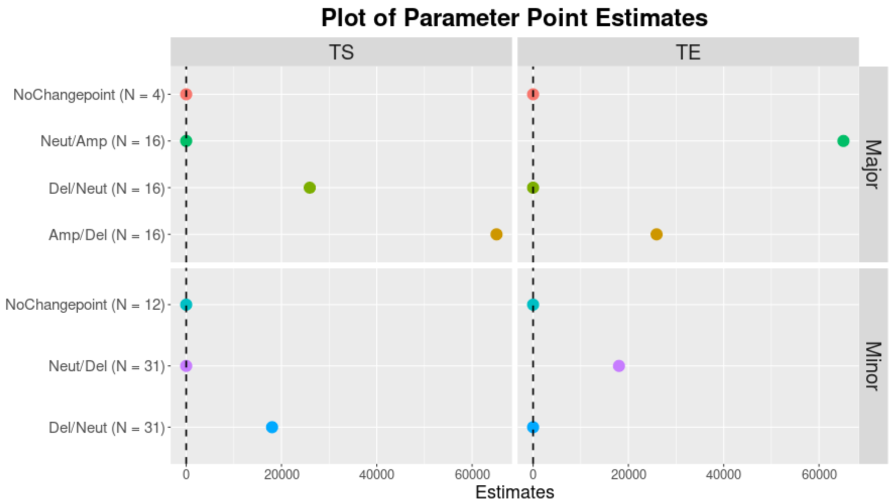


Figure 5.10: Plot of multivariate Allele-Dependent Intercept Model parameter estimates fitted using `lm()` and where neutral lengths are recorded as 0.

Table 5.21: Multivariate Allele-Dependent Non-Intercept Model parameter estimates and intervals fitted using `lm()` and where neutral lengths are recorded as length 0. Fit, LB and UB correspond to the parameter estimates and associated 95% confidence intervals.

Parameter Estimates and Confidence Intervals									
Coefficients	**Allele**	**Direction**	**n**	**Beta**	**P**	**Category**	**Fit**	**LB**	**UB**
CategoryAmp/Del	Major	TS	16	6.513×10^{4}	6.383×10^{-36}	Amp/Del	6.513×10^{4}	NA	NA
CategoryDel/Neut	Major	TS	16	2.592×10^{4}	1.317×10^{-11}	Del/Neut	2.592×10^{4}	NA	NA
CategoryNeut/Amp	Major	TS	16	8.154×10^{-13}	1.000×10^{0}	Neut/Amp	8.154×10^{-13}	NA	NA
CategoryDel/Neut:AlleleMinor	Minor	TS	31	NA	NA	Del/Neut	1.803×10^{4}	NA	NA
CategoryNeut/Del:AlleleMinor	Minor	TS	31	NA	NA	Neut/Del	-9.095×10^{-13}	NA	NA
CategoryAmp/Del	Major	TE	16	2.592×10^{4}	1.317×10^{-11}	Amp/Del	2.592×10^{4}	NA	NA
CategoryDel/Neut	Major	TE	16	-5.662×10^{-12}	1.000×10^{0}	Del/Neut	-5.662×10^{-12}	NA	NA
CategoryNeut/Amp	Major	TE	16	6.513×10^{4}	6.383×10^{-36}	Neut/Amp	6.513×10^{4}	NA	NA
CategoryDel/Neut:AlleleMinor	Minor	TE	31	NA	NA	Del/Neut	2.922×10^{-12}	NA	NA
CategoryNeut/Del:AlleleMinor	Minor	TE	31	NA	NA	Neut/Del	1.803×10^{4}	NA	NA

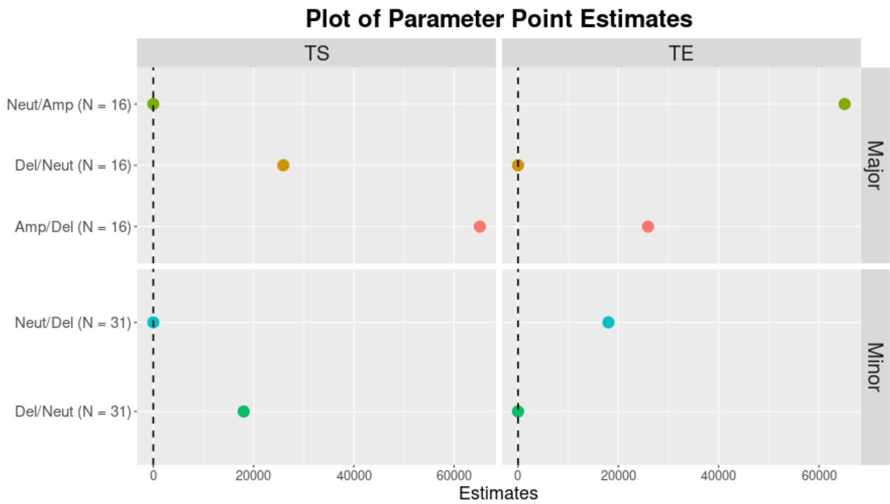


Figure 5.11: Plot of multivariate Allele-Dependent Non-Intercept Model parameter estimates fitted using `lm()` and where neutral lengths are recorded as 0.

5.5 Simulation Study

To explore the behaviour of the proposed AD models, a variety of scenarios are simulated, with varying sample sizes and profile compositions.

Scenario 1 assigns:

- a percentage of samples to have one of the two alleles displaying an amplified segment within the observed region (CNA profile B), and
- a percentage of samples to have no CNAs, i.e. no changepoints, in the observed region, for both alleles (CNA profile A).

Scenario 2 assigns:

- a percentage of samples to have one of the two alleles displaying an amplified segment followed by a deleted segment (what we define as an Amp/Del flashpoint pattern) within the observed region (CNA profile C), and
- a percentage of samples to have one of the two alleles displaying an amplified segment within the observed region (CNA profile B).

Scenario 3 assigns:

- a percentage of samples to have allele-specific copy number profile D, where one allele displays an Amp/Del flashpoint pattern flanked by two neutral segments and the other allele displays an oscillating pattern of deleted and neutral segments,
- a percentage of samples to have one of the two alleles displaying an amplified segment followed by a deleted segment (Amp/Del flashpoint pattern), within the observed region (CNA profile C), and
- a percentage of samples to have no CNAs, i.e. no changepoints, in the observed region, for both alleles (CNA profile A).

In setting parameters to generate the profiles, we assume a region similar in length to chromosome 1 ($\approx 250,000$kb) as the genomic region. Lengths are generated from truncated Normal distributions, mean lengths of CNA segments are similar to those observed in the ASCAT data (Tables 5.4 and 5.5), as specified in Table 5.22.

Table 5.22: Parameters of truncated Normal distributions used to simulate segment lengths and properties of simulated scenarios. a and b correspond to the lower and upper bound.

Scenario Distribution Parameters and Properties			
	Major Allele	**Minor Allele**	**Properties**
Scenario 1			
Profile A	No Breakpoint = 0	No Breakpoint = 0	P = 10%, 20%, ..., 90%
Profile B	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 15,249, σ = 28,815, a = 1, b = 250,000)	No Breakpoint = 0	n = 20, 50, 80, 100, 200, 500
Scenario 2			
Profile B	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 15,249, σ = 28,815, a = 1, b = 250,000)	No Breakpoint = 0	P = 10%, 20%, ..., 90%
Profile C	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 22,777, σ = 35,235, a = 1, b = 250,000) Del ~ TN(μ = 9,769, σ = 19,739, a = 1, b = 250,000)	No Breakpoint = 0	n = 20, 50, 80, 100, 200, 500
Scenario 3			
Profile A	No Breakpoint = 0	No Breakpoint = 0	P = 20%
Profile C	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 22,777, σ = 35,235, a = 1, b = 250,000) Del ~ TN(μ = 9,769, σ = 19,739, a = 1, b = 250,000)	No Breakpoint = 0	P_1= 10%, 20%, .., 70% P_2 = 100% - P - P_1
Profile D	Neutral ~ TN(μ = 27,641, σ = 35,854, a = 1, b = 250,000) Amp ~ TN(μ = 68,331, σ = 35,235, a = 1, b = 250,000) Del ~ TN(μ = 29,307, σ = 19,739, a = 1, b = 250,000)	Neutral ~ TN(μ = 31,129, σ = 38,125, a = 1, b = 250,000) Del ~ TN(μ = 8,997, σ = 18,675, a = 1, b = 250,000)	n = 20, 50, 80, 100, 200, 500

For scenario 1, datasets are generated varying the size of the dataset, n, ranging from 20 to 500, and varying the percentage of the dataset having a CNA, P, ranging from 20% to 90% of samples displaying allele-specific copy number profile B. For each specification, 20 replicated datasets were simulated. Scenario 2 consists of allele-specific copy number profiles B and C, at varying percentages P and $1-P$, each with 20 replications. For scenario 3, three possible allele-specific copy number profiles, A, C and D, setting the percentage of samples displaying profile A at $P = 20\%$ and with varying percentages of samples displaying allele-specific copy number profile C and D, P_1 and P_2, each with 20 replications, are generated. In all datasets the lengths of the neutral segments are recorded as length 0 and the NoChangepoint category is retained.

Univariate and multivariate AD models (ADIM) are fitted to all simulated datasets and assessed. An illustration of the distribution of point estimates across the replicated datasets, for example in simulated scenario 1, is provided in Figures 5.12-5.13. The TS and TE point estimates for each category and allele, over the 20 replications, indicate the variability in simulated datasets and that lengths simulated as 0 are observed to have points estimates around 0.

The significance of a changepoint category is assessed based on the confidence interval of the parameter being strictly positive, i.e. the value of the lower bound of the interval, $LB > 0$. The proportion of datasets observed to indicate significance of that parameter when fitted with univariate `lm()` and univariate `MCMCglmm()` are provided in Figure 5.14 and Figure 5.15, respectively. For scenario 1, the categories with simulated TS length, Amp/Neut, and TE length, Neut/Amp, are detected consistently on the Major allele across all simulated sample sizes and profile percentages. For scenario 2, with more categories and complex structure, more variability is observed but as the sample size increases the effect of this variability decreases and the

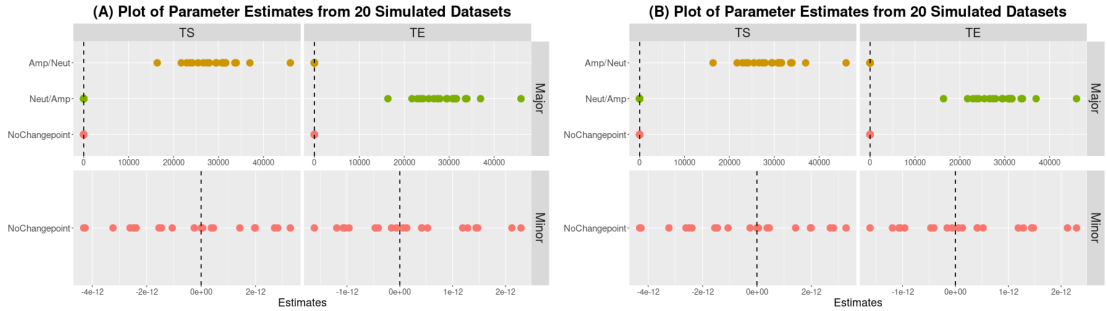


Figure 5.12: Plot of univariate Allele-Dependent Intercept Model parameter estimates for simulated scenario 1 with $n = 50$ and $P = 20$ fitted using the (A) `lm()` function and (B) `MCMCglmm()` function. Scales on the x-axis vary between alleles.

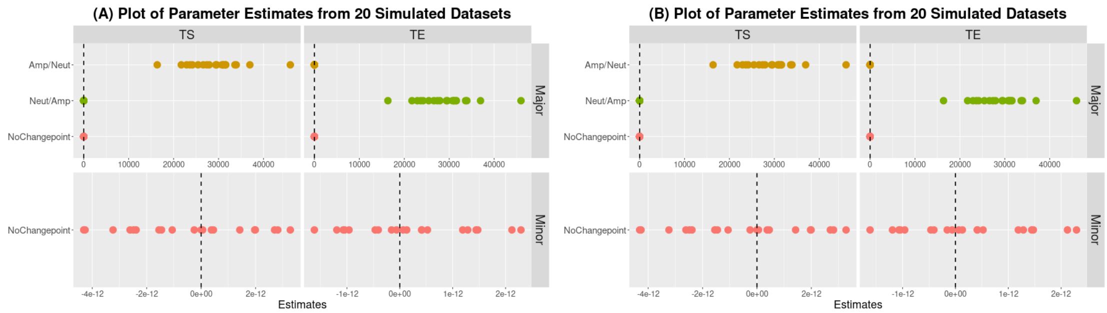


Figure 5.13: Plot of multivariate Allele-Dependent Intercept Model parameter estimates for simulated scenario 1 with $n = 50$ and $P = 20$ fitted using the (A) `lm()` function and (B) `MCMCglmm()` function. Scales on the x-axis vary between alleles.

types of changepoints simulated in the scenario are detected as significant. Similarly for the complexity of scenario 3, the TS Amp/Del and Del/Neut and TE Amp/Del and Neut/Amp on the Major allele, and the TS Del/Neut and TE Neut/Del on the Minor allele, are consistently detected across all sample sizes and profile percentages.

Fits produced from the multivariate AD models, using the `MCMCglmm()` function produces similar results to the univariate models and shows that as the sample size increases, the types of changepoints simulated in the scenario are detected as significant (Figure 5.16).

In the multivariate MCMCglmm fit of ADIM, we observe decreased True Positive rate for smaller sample sizes, evident in particular for scenario 2. Taking a closer look at individual datasets in scenario 2 reveals that for sample size $n = 20$, where 90% of samples display copy number profile C, the Amp/Neut category occurs only twice and as such the TS is not detected in approximately 35% of datasets (Table 5.23). In addition, where 10% of samples display copy number profile C, the Amp/Del category appears only twice and the TE is not detected in approximately 55% of datasets (Table 5.23). This indicates that the variability is due to the small sample size and decreased profile percentages of those specific categories.

Overall, it appears that for the majority of categories and scenarios, as the profile percentage or sample size increases, the detection of each category simulated goes to 1. Notably, by using intervals to assess significance, we can also identify regions where changepoints over a certain length occur, i.e. over $1,000$kb. This is done by checking whether the lower bound of the confidence interval is greater than 1,000 ($LB > 1000$).

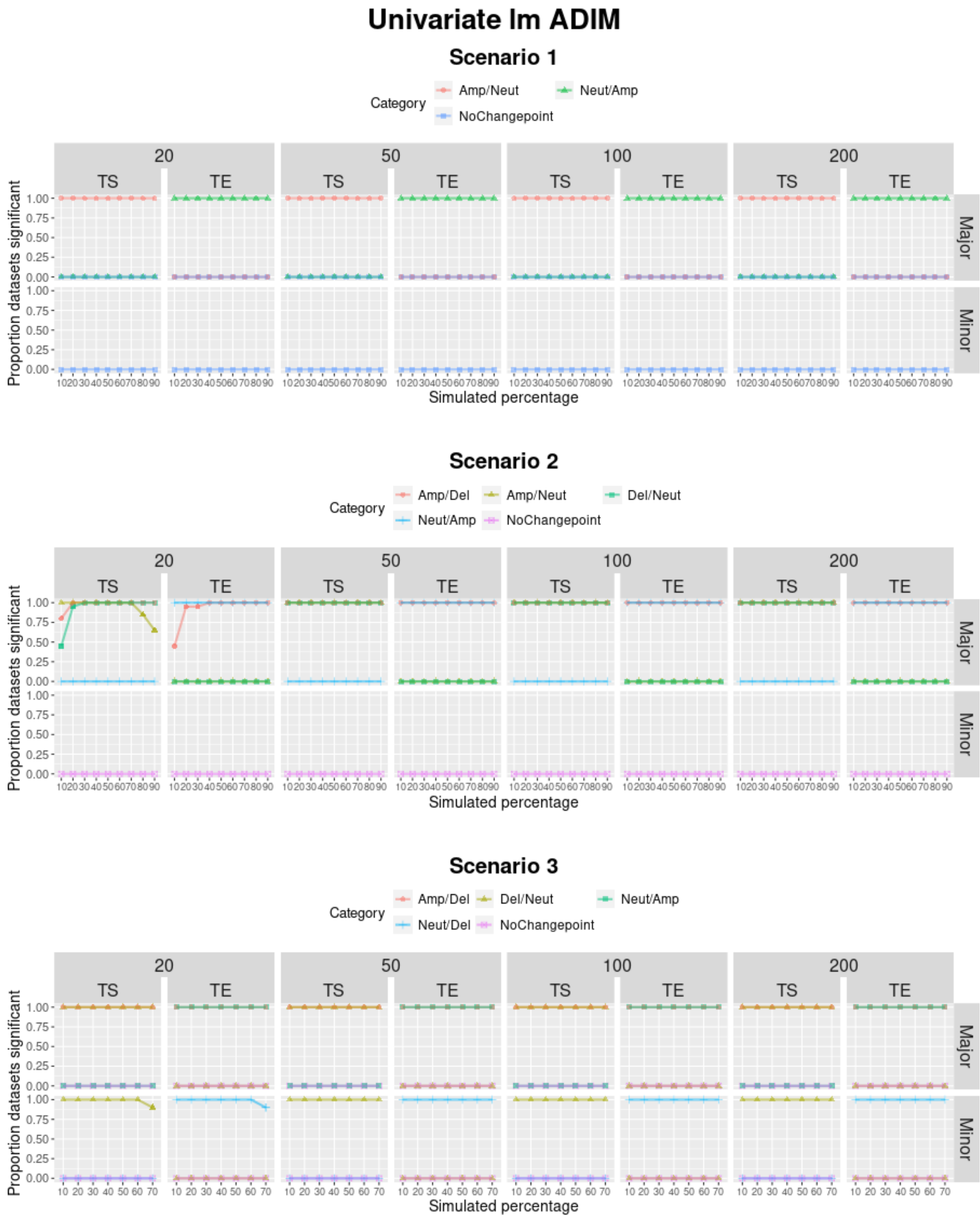


Figure 5.14: Plot displaying the proportion of the 20 simulated datasets, for each sample size and percentage, where the category was detected by our proposed univariate Allele-Dependent Intercept Model. Fitted using the `lm()` function. Significance of a changepoint category assessed using $LB > 0$.

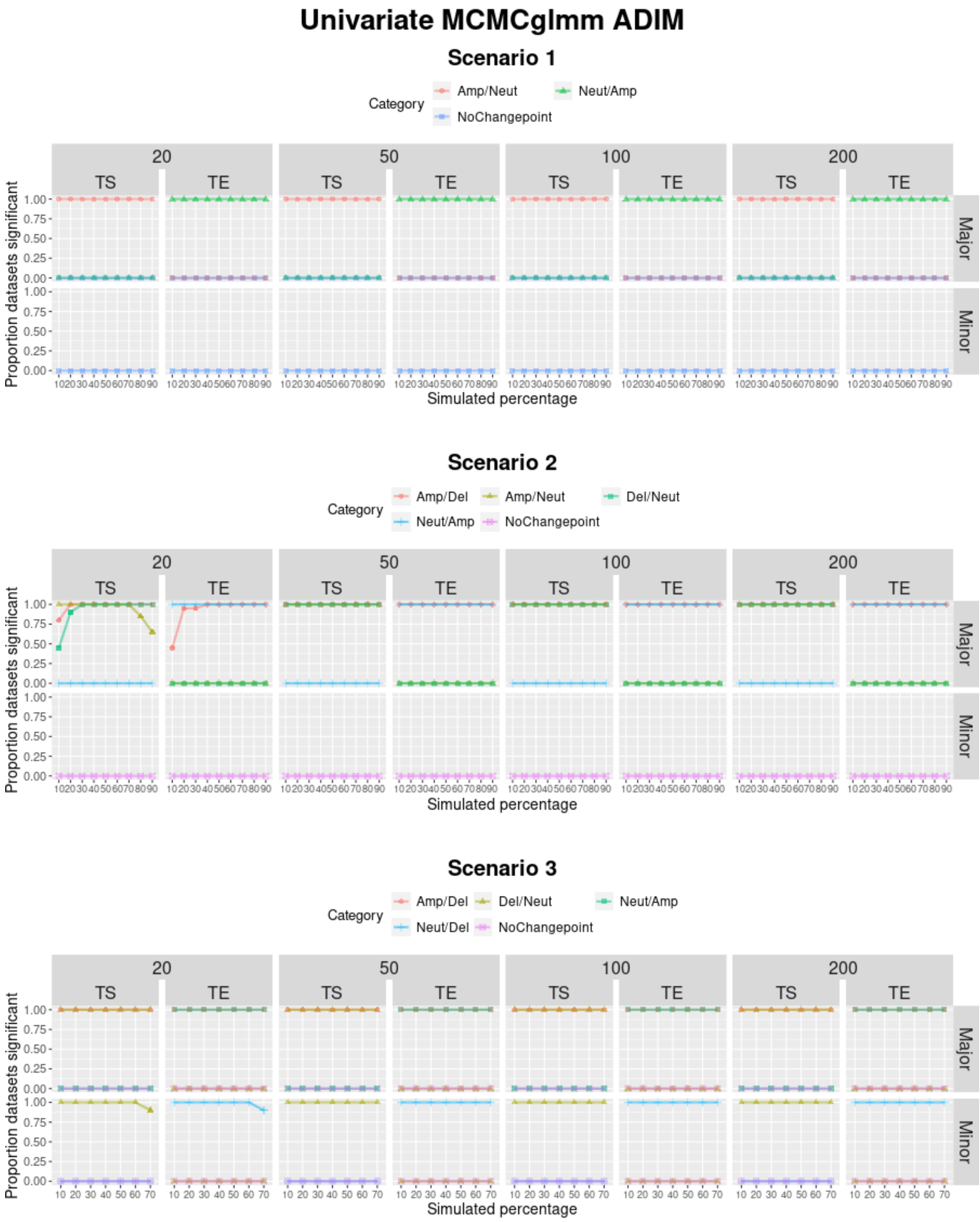


Figure 5.15: Plot displaying the proportion of the 20 simulated datasets, for each sample size and percentage, where the category was detected by our proposed univariate Allele-Dependent Intercept Model. Fitted using the `MCMCglmm()` function. Significance of a changepoint category assessed using $LB > 0$.

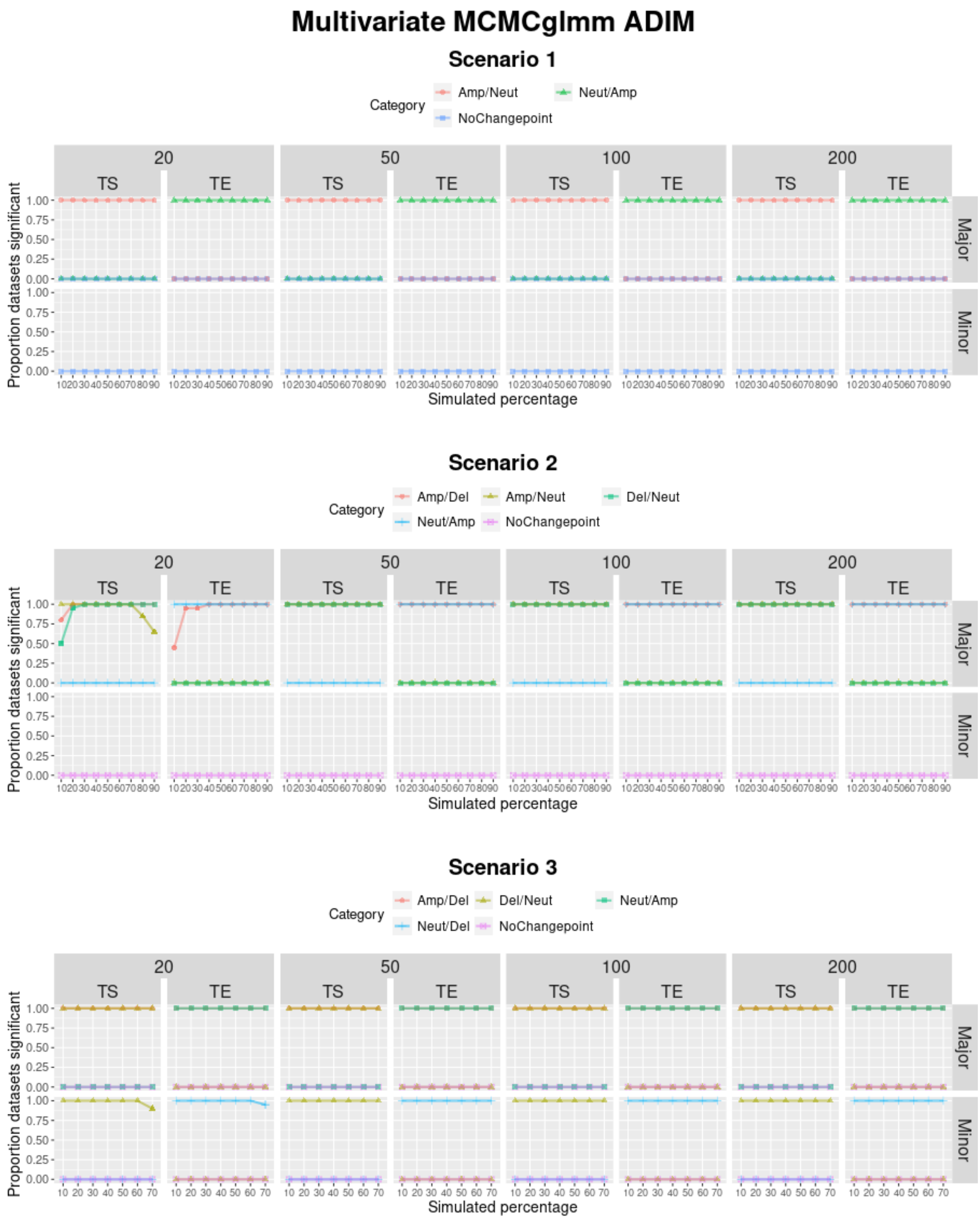


Figure 5.16: Plot displaying the proportion of the 20 simulated datasets, for each sample size and percentage, where the category was detected by our proposed multivariate Allele-Dependent Intercept Model. Fitted using the `MCMCglmm()` function. Significance of a changepoint category assessed using $LB > 0$.

Table 5.23: Model estimates for (A) TS Amp/Neut category and (B) TE Amp/Del category, where $n = 20$, across 20 simulated datasets.

(A) Univariate TS Parameter Estimates							
Dataset	**Category**	**Allele**	**Dir**	**n**	**Fit**	**LB**	**UB**
Dataset_1	Amp/Neut	Major	TS	2	26440.0	3246.2859	49633.71
Dataset_2	Amp/Neut	Major	TS	2	8765.5	-9085.6445	26616.64
Dataset_3	Amp/Neut	Major	TS	2	18242.0	-1471.8902	37955.89
Dataset_4	Amp/Neut	Major	TS	2	73338.0	55465.8272	91210.17
Dataset_5	Amp/Neut	Major	TS	2	10325.0	-12526.2699	33176.27
Dataset_6	Amp/Neut	Major	TS	2	20821.5	-1507.9160	43150.92
Dataset_7	Amp/Neut	Major	TS	2	23088.5	4970.2298	41206.77
Dataset_8	Amp/Neut	Major	TS	2	36372.5	11592.1352	61152.86
Dataset_9	Amp/Neut	Major	TS	2	67177.5	47494.3472	86860.65
Dataset_10	Amp/Neut	Major	TS	2	25001.0	9077.6980	40924.30
Dataset_11	Amp/Neut	Major	TS	2	25617.5	8804.8800	42430.12
Dataset_12	Amp/Neut	Major	TS	2	14224.0	-3060.4766	31508.48
Dataset_13	Amp/Neut	Major	TS	2	16377.5	2329.8033	30425.20
Dataset_14	Amp/Neut	Major	TS	2	23839.0	2021.7412	45656.26
Dataset_15	Amp/Neut	Major	TS	2	25224.5	4698.9172	45750.08
Dataset_16	Amp/Neut	Major	TS	2	21790.0	-721.8147	44301.81
Dataset_17	Amp/Neut	Major	TS	2	35030.5	18237.5442	51823.46
Dataset_18	Amp/Neut	Major	TS	2	17080.5	-578.4325	34739.43
Dataset_19	Amp/Neut	Major	TS	2	42653.0	25901.3583	59404.64
Dataset_20	Amp/Neut	Major	TS	2	56049.5	35919.1199	76179.88

(B) Univariate TE Parameter Estimates							
Dataset	**Category**	**Allele**	**Dir**	**n**	**Fit**	**LB**	**UB**
Dataset_1	Amp/Del	Major	TE	2	9314.5	-4723.4427	23352.44
Dataset_2	Amp/Del	Major	TE	2	13636.0	-3998.2505	31270.25
Dataset_3	Amp/Del	Major	TE	2	29605.0	14222.7267	44987.27
Dataset_4	Amp/Del	Major	TE	2	3132.0	-15889.5476	22153.55
Dataset_5	Amp/Del	Major	TE	2	26741.5	12189.7868	41293.21
Dataset_6	Amp/Del	Major	TE	2	37306.0	19687.2627	54924.74
Dataset_7	Amp/Del	Major	TE	2	30011.0	11844.2085	48177.79
Dataset_8	Amp/Del	Major	TE	2	8894.5	-10033.1042	27822.10
Dataset_9	Amp/Del	Major	TE	2	15455.0	2322.6625	28587.34
Dataset_10	Amp/Del	Major	TE	2	15975.0	-2501.1921	34451.19
Dataset_11	Amp/Del	Major	TE	2	7817.0	-4869.1098	20503.11
Dataset_12	Amp/Del	Major	TE	2	10833.5	-488.3804	22155.38
Dataset_13	Amp/Del	Major	TE	2	13238.5	-6191.8651	32668.87
Dataset_14	Amp/Del	Major	TE	2	5658.5	-10632.7464	21949.75
Dataset_15	Amp/Del	Major	TE	2	29063.5	7699.6489	50427.35
Dataset_16	Amp/Del	Major	TE	2	8816.5	-8771.2345	26404.23
Dataset_17	Amp/Del	Major	TE	2	39099.5	23041.1222	55157.88
Dataset_18	Amp/Del	Major	TE	2	22013.5	4745.2476	39281.75
Dataset_19	Amp/Del	Major	TE	2	34083.0	14180.0149	53985.99
Dataset_20	Amp/Del	Major	TE	2	14516.0	-1089.0303	30121.03

5.6 Conclusions

Allele-specific copy number profiling provides information on genome wide copy number for each allele and tackles some of the limitations of total copy number profiling, including masking of changepoints and certain types of genomic aberrations.

To use allele-specific copy number profiles, produced using ASCAT, to identify and characterise copy number associated changepoints based on the lengths of the flanking alteration segments, TS and TE, AD models, ADIM and ADNIM are proposed, including interaction terms, to detect the presence of these changepoints and their performance assessed across a number of scenarios.

It was observed that AI and AD models performed as expected when applying them to simulated datasets, where consideration was given to the form of data the models are applied to, including recording neutral segment length and retention of the NoChangepoint category. It was noted that while the confidence intervals were wider in cases where the neutral segments were retained as lengths greater than 0, and also in cases where the NoChangepoint category was excluded, the simulated categories are significantly greater than 0 in all cases and the overall conclusions the same. As our focus is on flanking alteration segments, the decision was made to record neutral segment lengths as 0, and due to sample size concerns, the decision was made to include NoChangepoint observations.

The AI and AD models perform similarly, estimating the mean lengths of TS and/or TE for the simulated categories as significantly greater than 0. Importantly, the AD models provide more detailed information regarding which allele the changepoint has occurred on, enabling identification of cases where a changepoint category occurs preferentially on one allele and cases where the average lengths of alteration segments may differ between alleles.

Assessing how the AD models, fitted using the `lm()` and `MCMCglmm()` functions, perform across datasets of varying sample sizes and profile percentages, indicates that the univariate ADIM, fitted using the `lm()` and `MCMCglmm()` functions, and the multivariate ADIM, fitted using the `MCMCglmm()` function, performed consistently well across sample sizes and profile percentages. The multivariate AD models, fitted using the `lm()` function, were lacking confidence intervals, a result of limitations in the software.

In the next chapter the selected models, univariate ADIM fitted using `lm()` and multivariate ADIM fitted using `MCMCglmm()`, will be applied to the allele-specific copy number profiles generated for the METABRIC data. In this application, the observed interval d is either a gene region, i.e. between the start and end position of a gene, or segments of the genome, generated using a segmentation or sliding window approach. The resulting significant regions will then be analysed, identifying regions of the genome where allele-specific changepoints occur and further investigation is warranted.

6 Application of Allele-specific models to the METABRIC data

In this Chapter, allele-specific copy number profiles for 1,984 patients in the METABRIC cohort are produced, the frequency of changepoints across defined intervals examined, and the selected statistical models, ADIM fitted using the `lm()` and `MCMCglmm()` functions, applied to identify regions containing CNA changepoints of significant length. Here, the observed interval d is either a gene region, i.e. between the start and end position of a gene, referred to as gene-centric, or genomic segments of specified length generated using a segmentation approach.

6.1 Gene-centric Application of Allele-specific Profile Analysis

Gene-centric changepoints (changepoints that occur in genes) in allele-specific CNA profiles are examined using visualisations and application of AD models.

6.1.1 Gene-centric Allele-specific CNA State Heatmaps

The CNA landscape of allele-specific copy number profiles, in terms of the CNA state of each gene, on each allele, are visualised using heatmaps.

In Section 3.4.3, heatmaps were implemented based on aggregate total CNA data, of particular interest was chromosome 3p, where the patients were partitioned into nodes informed by chromosome arm CNA Burden metrics (Figure 3.46). Allele-specific profiles are now provided, that is, for the 3p Major allele (Figure 6.1) and for the 3p Minor allele (Figure 6.2). The CNA states correspond to -1 (deletion), 0 (neutral), 1 (gain) and 2 (amplification).

Figure 6.1 indicates subsets of patients displaying amplifications in almost every gene on the Major allele across chromosome 3p, while Figure 6.2 indicates clusters of patients displaying either amplifications or deletions in almost every gene on the Minor allele across chromosome 3p. These heatmaps indicate that amplifications are more prevalent on the Major allele, while deletions tend to occur more on the Minor allele. This is not surprising given that the Minor allele is defined as the allele with the lowest copy number across the genome. Notably, the deletions observed across chromosome 3p utilising total CNA states occur on a single allele, the Minor allele (Figure 3.46).

Combining the allele-specific CNA states, by adding the individual allele states for the gene, shown in Figure 6.3, shows some similarities to that generated using total copy number data (Figure 3.46). In both heatmaps, the Claudin-low and Luminal A patients corresponding to Node 5 have high levels of deletions across chromosome 3p. The majority of Node 4, also containing Claudin-low and Luminal A patients, display little to no copy number changes across chromosome 3p and Node 2, containing Luminal B, HER2, Normal and Basal patients, consists of patients displaying variation in levels of GI across chromosome 3p. There are some noteworthy differences, however. The allele-specific data contains more patients displaying high levels of amplifications, possibly indicative of genome duplication, across chromosome 3p, particularly evident for Node 2 patients (Figures 3.46 and

6.3). These allele specific heatmaps also highlight the existence of copy number neutral changes.

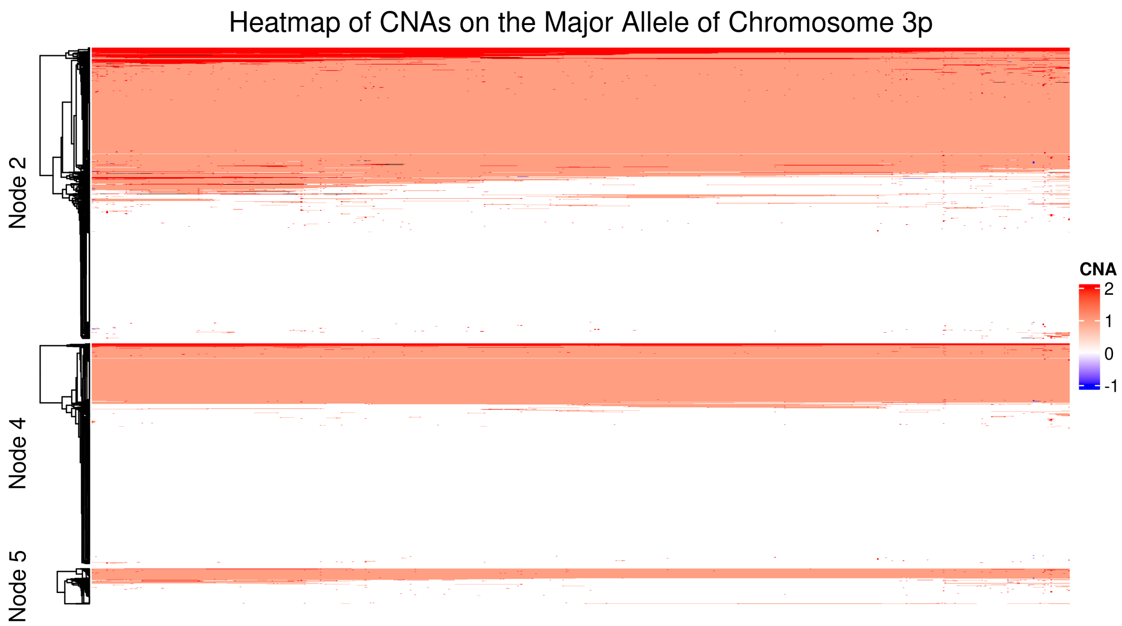


Figure 6.1: Heatmap of CNAs across the Major Allele of Chromosome 3p. The heatmap depicts the CNA state for each gene across Chromosome 3p, partitioning the patients into the nodes corresponding to Figure 3.31. NAs, depicting multiple states, are coloured in black.

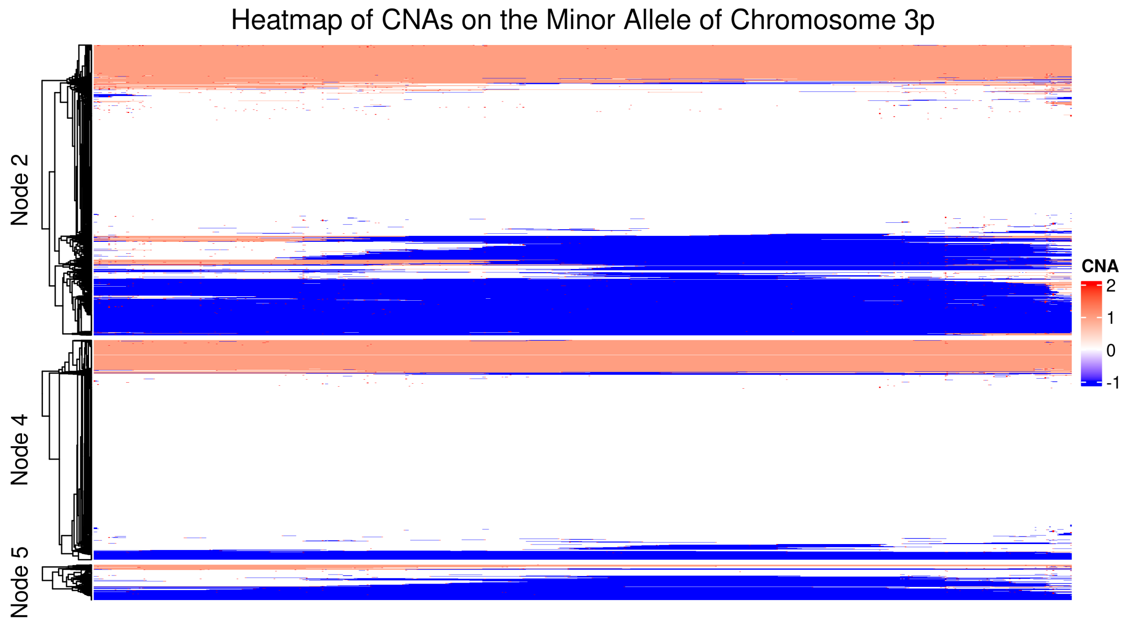


Figure 6.2: Heatmap of CNAs across the Minor Allele of Chromosome 3p. The heatmap depicts the CNA state for each gene across Chromosome 3p, partitioning the patients into the nodes corresponding to Figure 3.31. NAs, depicting multiple states, are coloured in black.

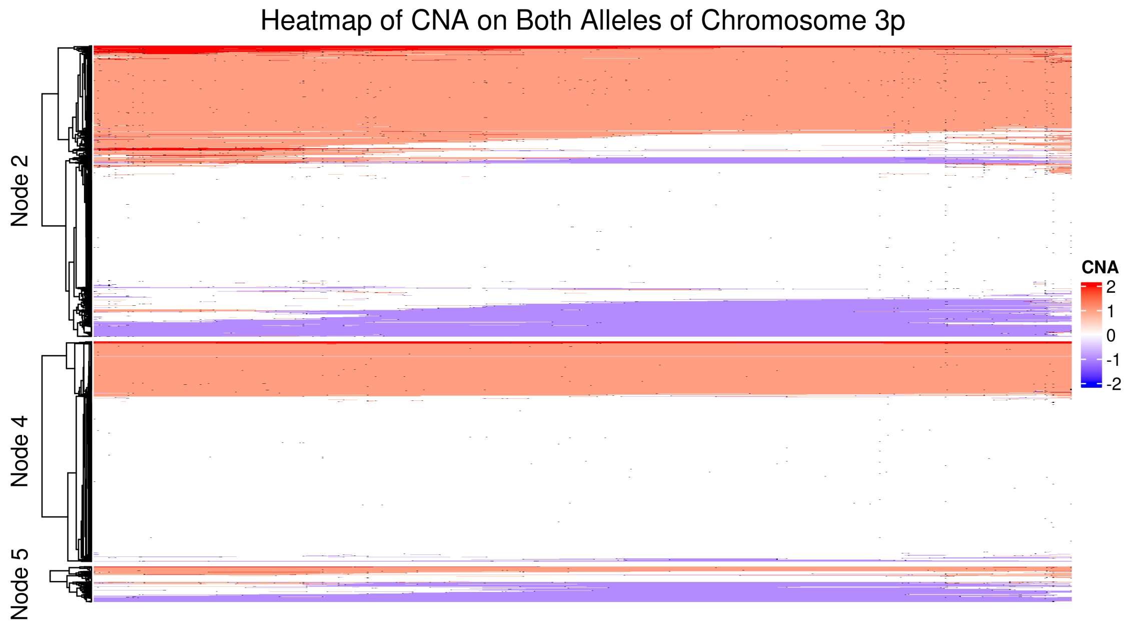


Figure 6.3: Heatmap of CNAs on both the Major and Minor alleles of Chromosome 3p. The heatmap depicts the CNA state for each gene across Chromosome 3p, partitioning the patients into the nodes corresponding to Figure 3.31. NAs, depicting multiple states, are coloured in black.

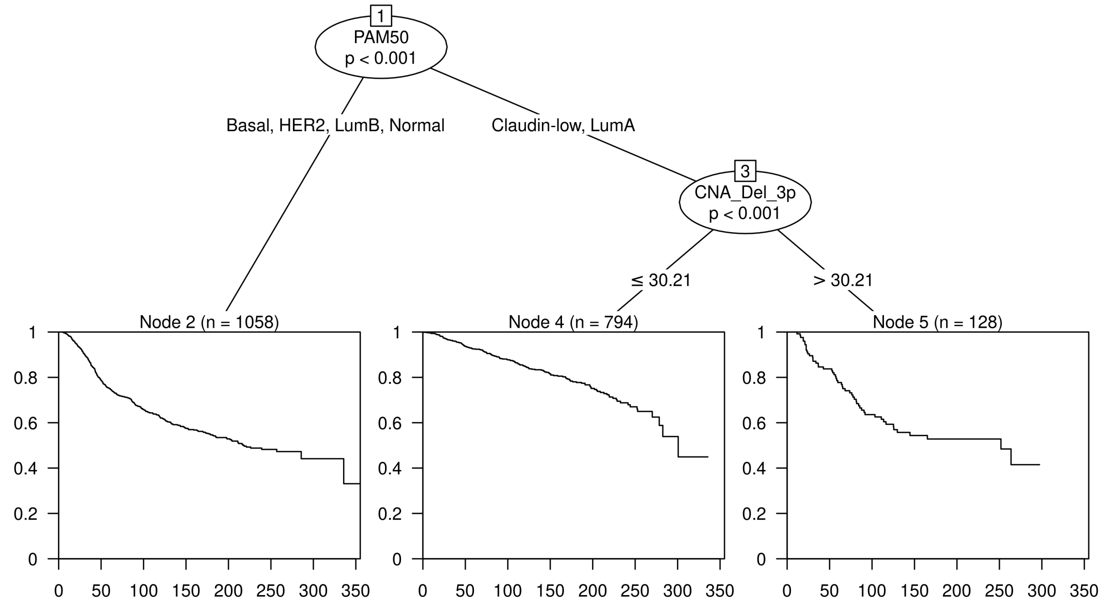
Chromosome arms 18q and 11p were also a point of focus in analysing total CNAs in Section 3.4.3 (Figure 3.47-3.48). Allele-specific heatmaps produced for chromosome 18q and 11p, also indicate that the Major allele is dominated by amplification events, while the Minor allele, although displaying both amplifications and deletions, is primarily dominated by deletion events (Appendix F). Comparatively, ASCAT estimates fewer patients with widespread deletions across chromosomes and estimates more patients with widespread amplifications across chromosomes.

6.1.2 Gene-centric Allelle-specific Changepoints across Chromosome Arms

Figures 6.1-6.3 indicate that amplifications and deletions may occur because the whole length of the gene is amplified or deleted, denoted in red and blue, or because alterations (amplifications and/or deletions) occur at some point(s) within the gene, denoted in black.

The frequency of changepoint events in genes on chromosome 3p, determined by the summation of observed changepoint counts over all patients, within each gene, is provided in Figure 6.4 where panels provide survival tree node and changepoint category information. These changepoint events, with the potential to disrupt gene function, tend to be observed with similar frequencies across all categories within each node, except for the Amp/Del and Del/Amp categories, which occur less frequently. Again, it is observed that Neut/Del and Del/Neut events occur more often on the Minor allele and Amp/Neut and Neut/Amp events occur more often on the Major allele.

For Node 2, comprising 1,044 patients, the frequency of observed changepoints is quite low, with the gene displaying the highest total number of changepoints being



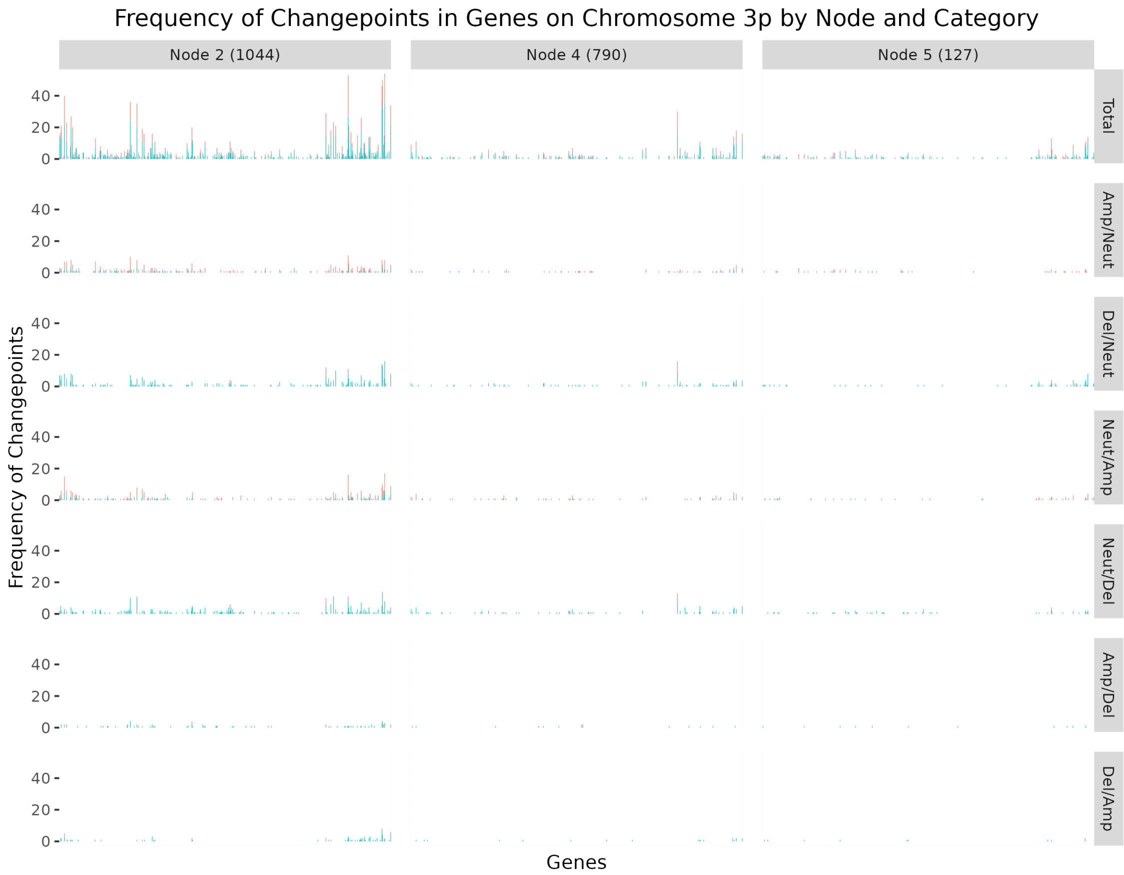


Figure 6.4: Frequency of changepoints in genes across chromosome 3p, split by Node and Category, and coloured by allele. For patients stratified into distinct survival patterns, Nodes 2, 4, and 5 corresponding to Figure 3.31 (figure panel columns), and for each changepoint category (panel rows), the frequency of that changepoint, observed in the each gene across chromosome 3p (x-axis) is plotted. Frequencies of the Major allele are coloured pink and the Minor allele coloured blue.

CADM2, displaying 54 changepoints (Table 6.1). Node 4 and Node 5, comprising 790 and 127 patients, also display low numbers of changepoints. The gene with the highest frequencies of changepoints in Nodes 4 and 5 are SFMBT1 with 30 changepoints, and CADM2 with 14 changepoints. While it appears that some genes are more susceptible to containing a changepoint, there seems to be no characteristic changepoint pattern that distinguishes patients within nodes.

Table 6.1: Top 10 genes on chromosome 3p with highest frequency of changepoints for patients in (A) Node 2, (B) Node 4, and (C) Node 5.

(A) Frequency of Changepoints within Genes (Node 2)			
Gene	**Major**	**Minor**	**Total**
CADM2	19	35	54
FHIT	26	27	53
ROBO1	13	37	50
ROBO2	14	32	46
SUMF1	19	21	40
TBC1D5	12	24	36
ZNF385D	15	20	35
EPHA3	12	22	34
SFMBT1	14	15	29
GRM7	8	19	27

(B) Frequency of Changepoints within Genes (Node 4)			
Gene	**Major**	**Minor**	**Total**
SFMBT1	15	15	30
CADM2	8	10	18
EPHA3	5	11	16
ROBO2	4	10	14
FHIT	2	9	11
SUMF1	5	6	11
CHL1	4	5	9
ROBO1	1	8	9
PTPRG	1	7	8
CACNA1D	4	3	7

(C) Frequency of Changepoints within Genes (Node 5)			
Gene	**Major**	**Minor**	**Total**
CADM2	4	10	14
FHIT	6	7	13
ROBO1	3	8	11
FOXP1	4	5	9
ROBO2	5	4	9
ARHGEF3	2	4	6
PTPRG	4	2	6
RBMS3	3	3	6
THRB	2	3	5
ZNF385D	1	4	5

Chromosome 18q (Appendix F) and 11p (Figure 6.5) also indicate low frequencies of changepoints within genes across the chromosome arms. Figure 6.5 displays a prominent peak in Node 3, count 113, and Node 7, count 161, corresponding to the gene TRIM5.

These results indicate that while a large number of genes on chromosomes 3p, 18q and 11p, in subsets of patients, are affected by amplifications or deletions, as seen in Figure 6.3 where a large number patients display widespread amplifications and/or deletions, the changepoints are generally not occurring in the genes (Figures 6.4 and 6.5 and Tables 6.1 and 6.2). The distribution of changepoints across all genes annotated in the CNA data, for which genomic location information could be obtained, is provided in Figure 6.6, again indicating that apart from a few genes harbouring large numbers of changepoints, changepoints are not frequently observed in genes. Genes displaying high numbers of changepoints include TRIM5 on chromosome 11, LCE1E on chromosome 1 and OPHN1 on the X chromosome, with 400 (204 on Major allele and 196 on Minor allele), 328 (188 on Major allele and 140 on Minor allele), and 289 (121 on Major allele and 168 on Minor allele) changepoints observed across all patients, where more than one changepoint can occur in a patient (Figure 6.6 and Table 6.3).

While Figure 6.6 and Table 6.3 indicate which genes most frequently harbour changepoints, no indication is given regarding how much of the surrounding genome is of altered copy number, i.e. the number of bases affected by an amplification or deletion. To detect changepoints, based on the TS and TE lengths, we apply the ADIM to each gene.

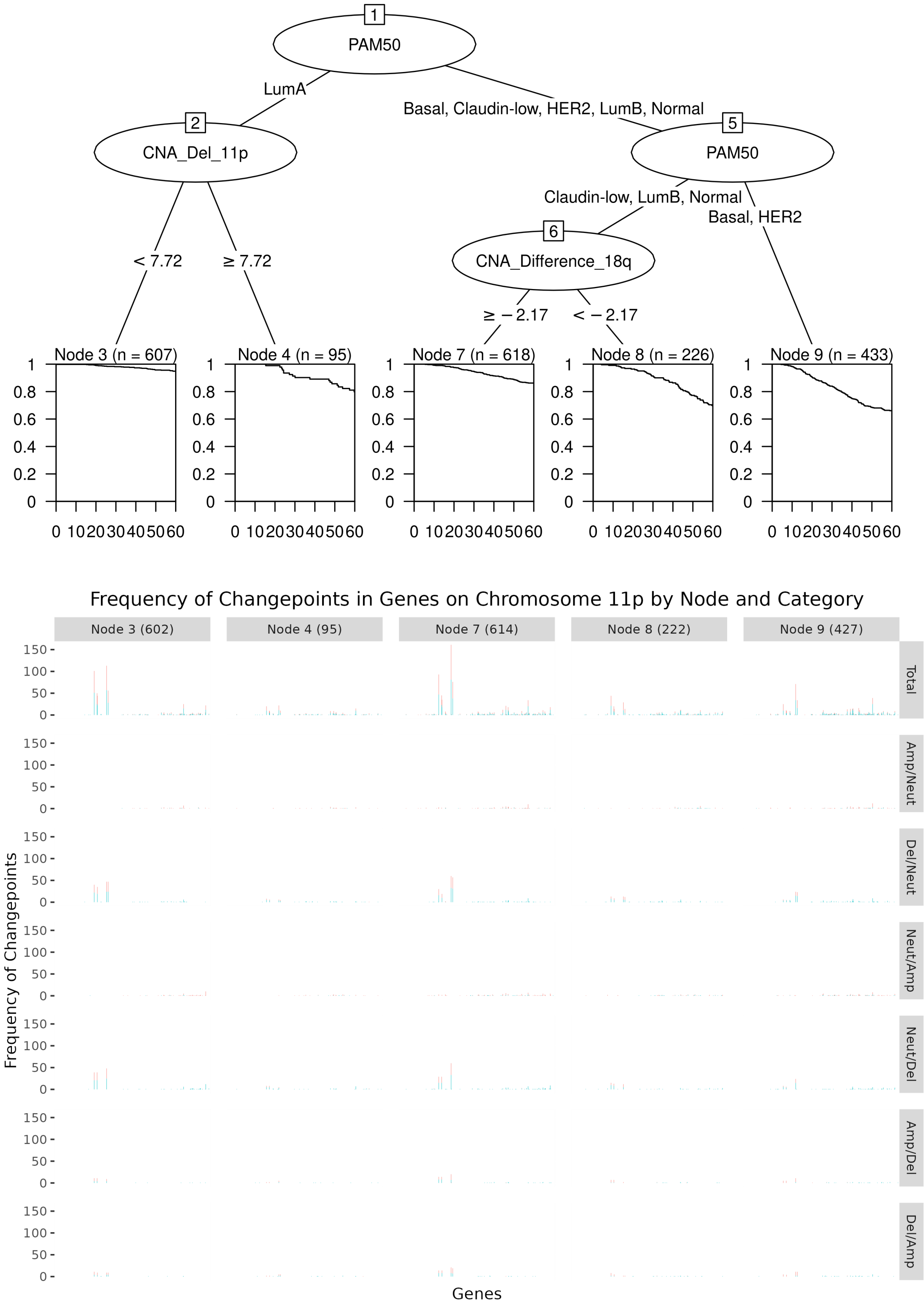


Figure 6.5: Frequency of changepoints in genes across chromosome 11p, split by Node and Category, and coloured by allele. For patients stratified into distinct survival patterns, Nodes 3, 4, 7, 8 and 9, corresponding to Figure 3.32 (figure panel columns), and for each changepoint category (panel rows), the frequency of that changepoint, observed in the each gene across chromosome 11p (x-axis) is plotted. Frequencies of the Major allele are coloured pink and the Minor allele coloured blue.

Table 6.2: Top 10 genes on chromosome 11p with highest frequency of changepoints for patients in (A) Node 3 and (B) Node 7.

(A) Frequency of Changepoints within Genes (Node 3)

Gene	Major	Minor	Total
TRIM5	56	57	113
MMP26	50	51	101
OR52N1	28	28	56
OR51A4	25	25	50
OR51A2	22	22	44
LRRC4C	10	15	25
PTPRJ	10	12	22
ANO3	4	2	6
NELL1	1	5	6
NUP160	3	3	6

(B) Frequency of Changepoints within Genes (Node 7)

Gene	Major	Minor	Total
TRIM5	80	81	161
MMP26	46	47	93
OR52N1	38	38	76
OR51A4	22	23	45
OR51A2	17	18	35
LRRC4C	14	20	34
NELL1	9	12	21
LUZP2	7	11	18
PTPRJ	6	12	18
NAV2	7	5	12

Table 6.3: Frequency of changepoints within genes, the top 20 genes are shown.

Frequency of Changepoints within Genes
Rows 1 to 10

Gene	Chr	Major	Minor	Total
TRIM5	11	204	196	400
LCE1E	1	188	140	328
OPHN1	X	121	168	289
MMP26	11	146	141	287
DLG2	11	97	170	267
CSMD1	8	141	124	265
ALG1L2	3	133	129	262
KANSL1	17	120	123	243
SHANK2	11	97	145	242
EYS	6	70	171	241

Frequency of Changepoints within Genes
Rows 11 to 20

Gene	Chr	Major	Minor	Total
NRG1	8	90	121	211
PTPRD	9	85	116	201
OR52N1	11	99	93	192
HYDIN	16	149	40	189
ADAM5	8	98	89	187
MACROD2	20	72	115	187
LCE1E	1	142	44	186
ALG1L2	3	79	98	177
LCE3B	1	88	82	170
KANSL1	17	113	56	169

Notably, different genes have different base lengths, meaning that in the gene-centric approach the observed interval d will have variable lengths. Consequently, genes of larger length may display higher frequencies of changepoints, while genes of shorter length may display low numbers of changepoints. In addition, not all categories of changepoints may occur within d and observation of small numbers of changepoints within d may not support model fitting.

Out of 22,544 genes considered, 10,966 genes do not display any changepoint, meaning no model is fit for those genes. Fitting multivariate ADIM, using the `MCMCglmm()` function, is supported for only 2,290 genes. The 9,288 genes for which the `MCMCglmm()` function is not supported require the use of a stronger or proper prior to be implemented.

We first test whether the mean lengths of the changepoint alterations, TS and TE, are greater than 10kb, where significance is determined by the lower bound of the confidence interval for the relevant length, TS or TE, being greater than 10kb, $LB > 10$kb. Figure 6.7, the tile plot displaying the results from applying the ADIM to 2,290 genes, indicates the widespread presence of changepoints with TS or TE

greater than 10kb. Significance in TS but not TE is indicated in purple, significance in TE but not TS is indicated in blue and significance in both is indicated in green. The number of observations in each category, for each gene, is provided by the diameter size of the point, larger points indicating larger sample size. For a large number of genes, the mean TS for the Amp/Neut and Del/Neut categories, the mean TE for the Neut/Amp and Neut/Del categories, and the mean TS or TE for the Del/Amp and Amp/Del categories are significantly greater than 10kb, $LB >$ 10kb. We also observe that deletion events, i.e. Neut/Del, Del/Neut, Amp/Del and Del/Amp, occur more often on the Minor allele than on the Major allele. A number of chromosomes, including chromosome 1, 3, 8, 16, 17 and 23, contain genes harbouring a large number of changepoints with TS or TE length greater than 10kb.

Next we test whether the mean lengths for each changepoint category (TS or TE) are greater than 10,000kb, significance determined by the lower bound of the confidence interval for the relevant TS/TE length being greater than 10,000kb, $LB > 10,000$kb (Figure 6.8). Chromosomes containing genes with changepoints of significant length, at larger sample sizes, include chromosomes 1, 6, 8, 11 and X.

Table 6.4A indicates which genes harbour changepoints with a TS and/or TE $> 10,000$kb on average, with a within category sample size greater than 30 observations. Noteworthy genes identified include OR52N1 on chromosome 11, TRIM5 on chromosome 11, and ALG1L2 on chromosome 3, all containing a significant Del/Amp on the Major allele. The gene with the largest number of changepoints, with average TS and/or TE length $> 10,000$kb, is LCE1E with $n = 142$.

As the `MCMCglmm()` function did not support model fitting for 9,288 genes, univariate ADIM fitted using the `lm()` function is applied, producing model estimates for all 22,544 genes.

Testing whether the mean lengths of the changepoint alterations, TS and TE, are greater than 10kb, $LB >$ 10kb, and whether the mean lengths of the changepoint alterations, TS and TE, are greater than 10,000kb, $LB > 10,000$kb (Figures 6.9 and 6.10) produces similar results for the 2,290 genes considered previously, but also provides information on an additional 9,288 genes, highlighting additional genes with significant changepoints, one of which is observed on chromosome 1, where a Del/Amp changepoint on the Major allele is observed. Table 6.4B indicates which genes harbour changepoints with a TS and/or $TE > 10,000$kb on average, with a within category sample size greater than 30 observations, fitted using the `lm()` function. The only difference observed between the results produced using `lm()` (Table 6.4A) and `MCMCglmm()` (Table 6.4B) is the inclusion of LCE3B, a gene where the MCMCglmm model did not support fitting (Table 6.4A). Chromosomes containing genes of interest include chromosome 1, 3, 6, 8, 9, 10, 11, 20 and X.

Exploring further the identified point of focus, Del/Amp changepoint on the Major allele in gene OR52N1, KM survival curves for DSS outcome are produced and compared for patients who exhibit or do not exhibit this particular changepoint profile (Figure 6.11A). Similarly, DSS KM survival curves are produced, for two further points of focus, Del/Amp changepoint on the Major allele in gene TRIM5 (Figure 6.11B) and Del/Amp changepoint in gene ALG1L2 (Figure 6.11C). Applying the log-rank test for each point of focus indicates patients with a Del/Amp changepoint in ALG1L2 have worse DSS outcomes than patients that do not exhibit that changepoint, $p = 0.03$. However, when multiple testing correction is applied, this p-value becomes non-significant at $\alpha = 0.05$.

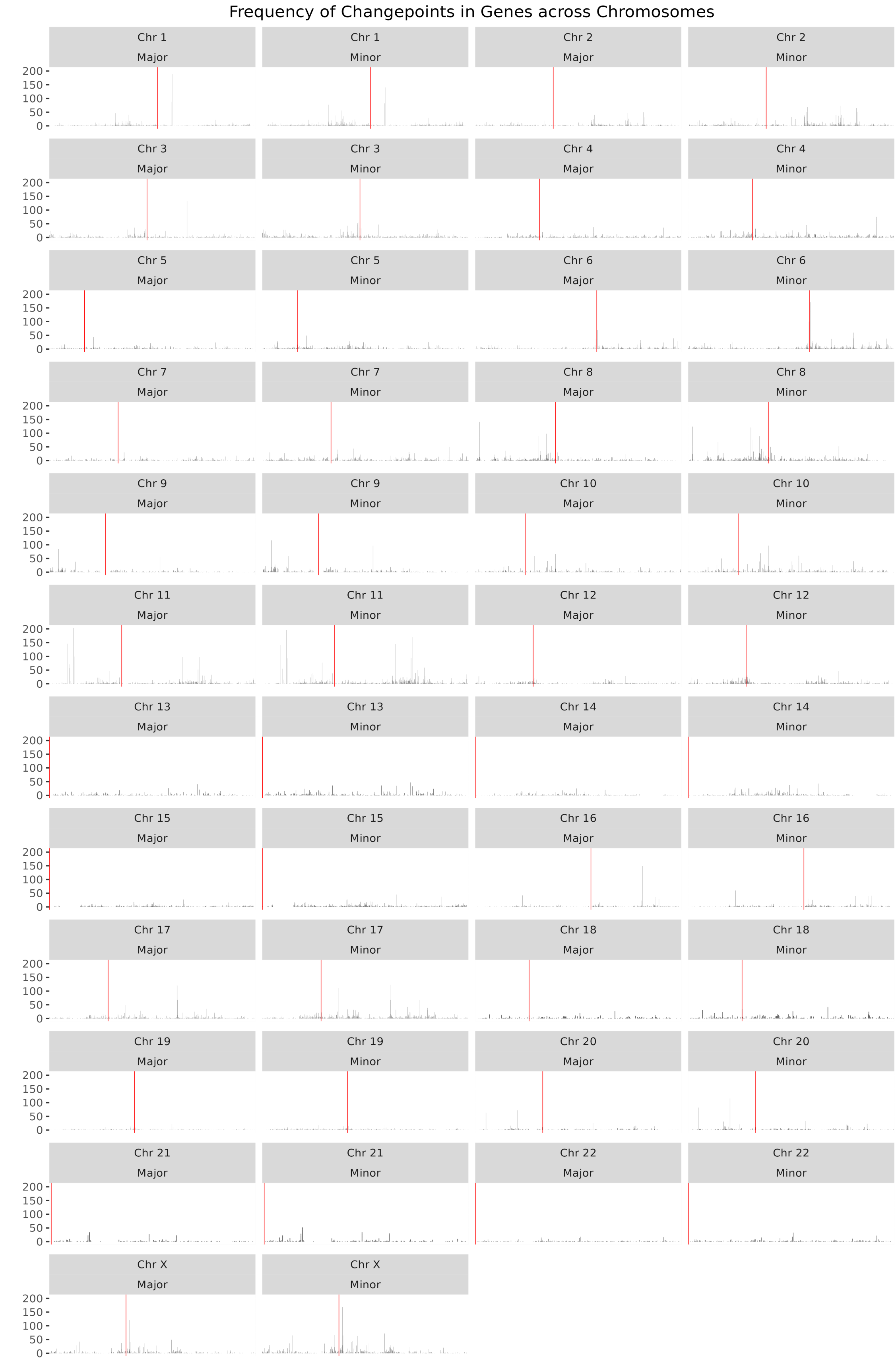


Figure 6.6: Frequency of changepoints in genes across each chromosome and allele. Scale of y-axis is the same across all chromosomes and the red line indicates the midpoint of the centromere.

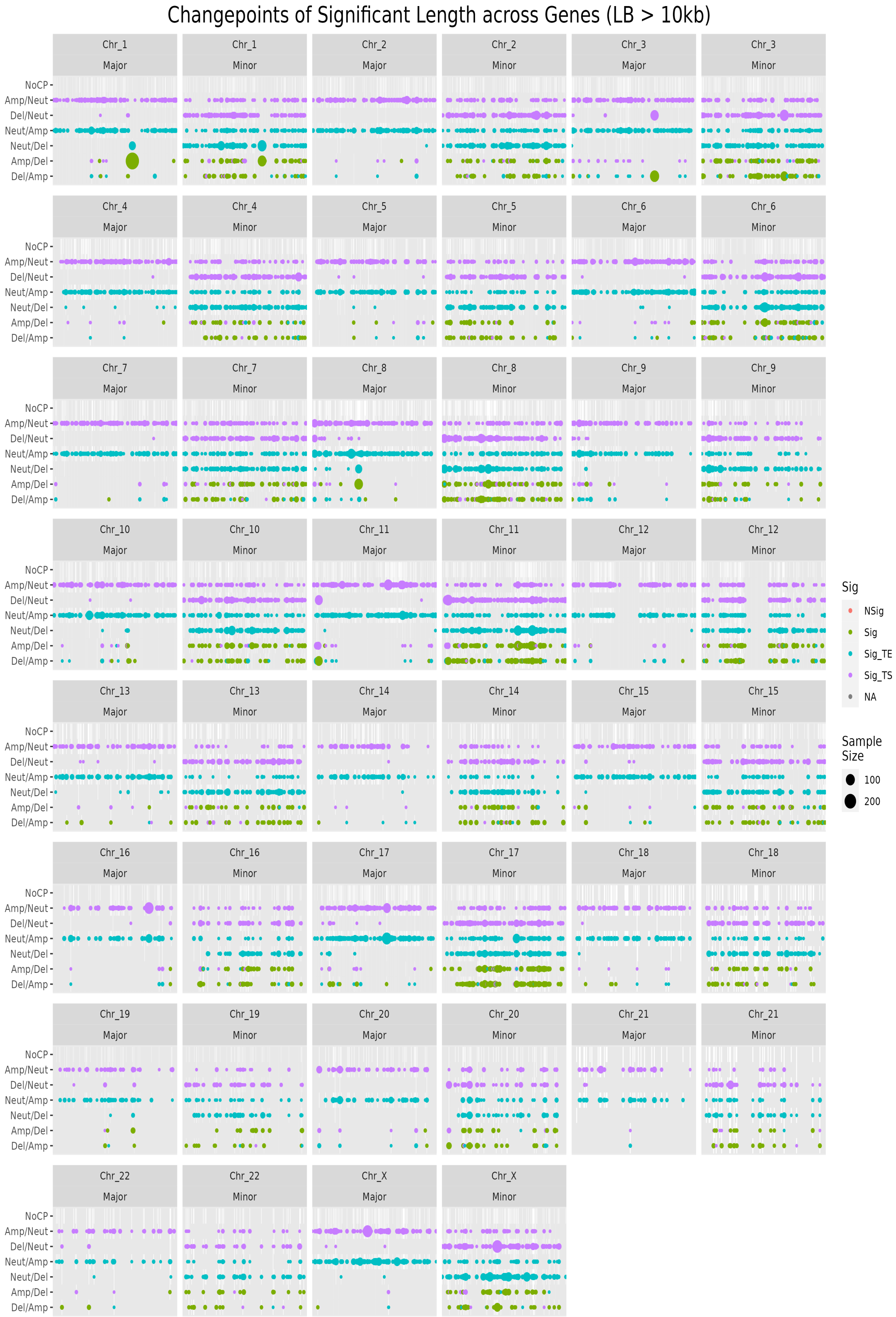


Figure 6.7: Application of multivariate Allele-Dependent Intercept Model to each gene, providing confidence intervals for each category and allele. Each panel, corresponding to chromosome and allele, displays significance of changepoint determined by $LB > 10$kb. NoCP corresponds to NoChangepoint. Fitted using the `MCMCglmm()` function.

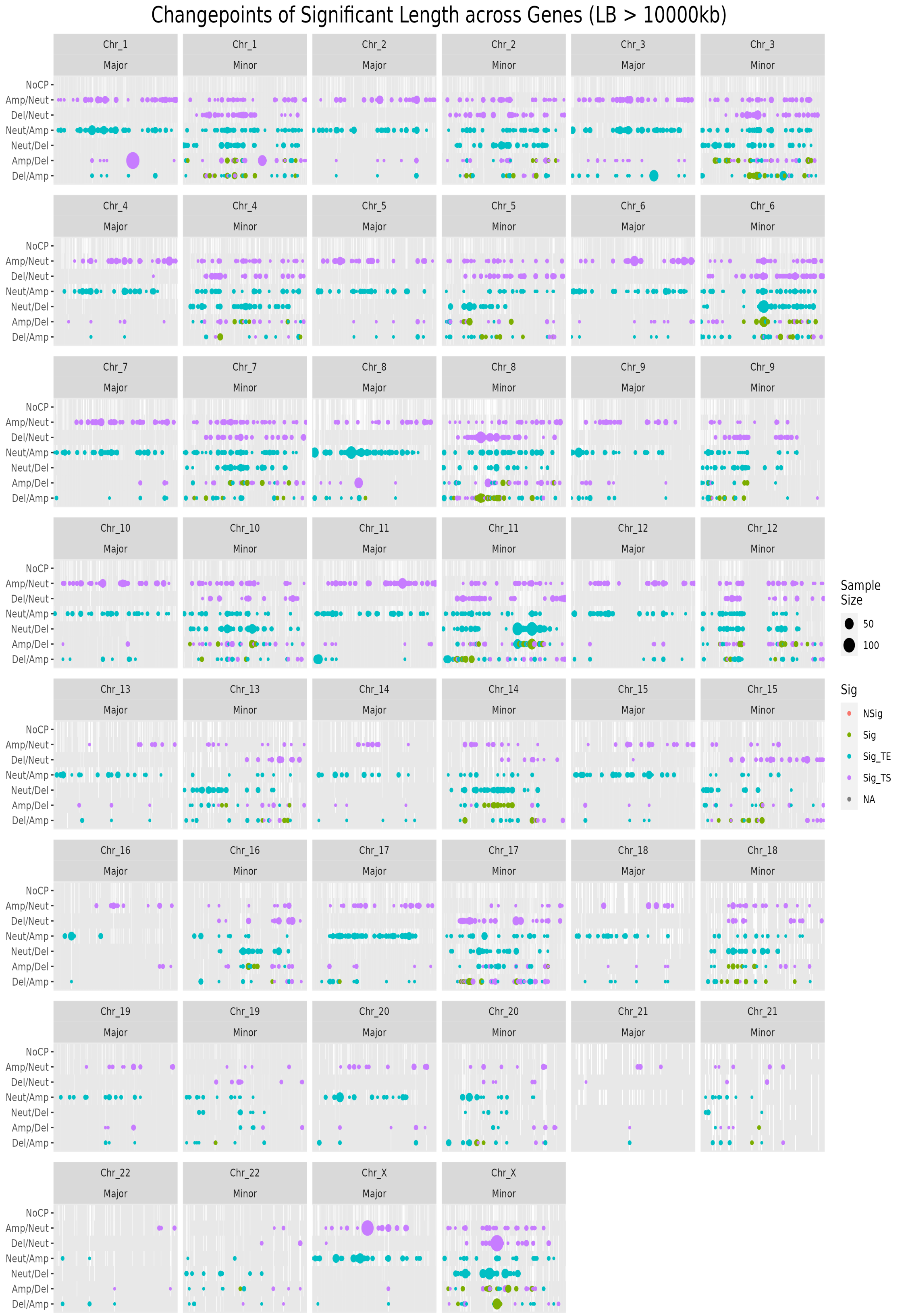


Figure 6.8: Application of multivariate Allele-Dependent Intercept Model to each gene, providing confidence intervals for each category and allele. Each panel, corresponding to chromosome and allele, displays significance of changepoint determined by $LB > 10,000$kb. NoCP corresponds to NoChangepoint. Fitted using the `MCMCglmm()` function.

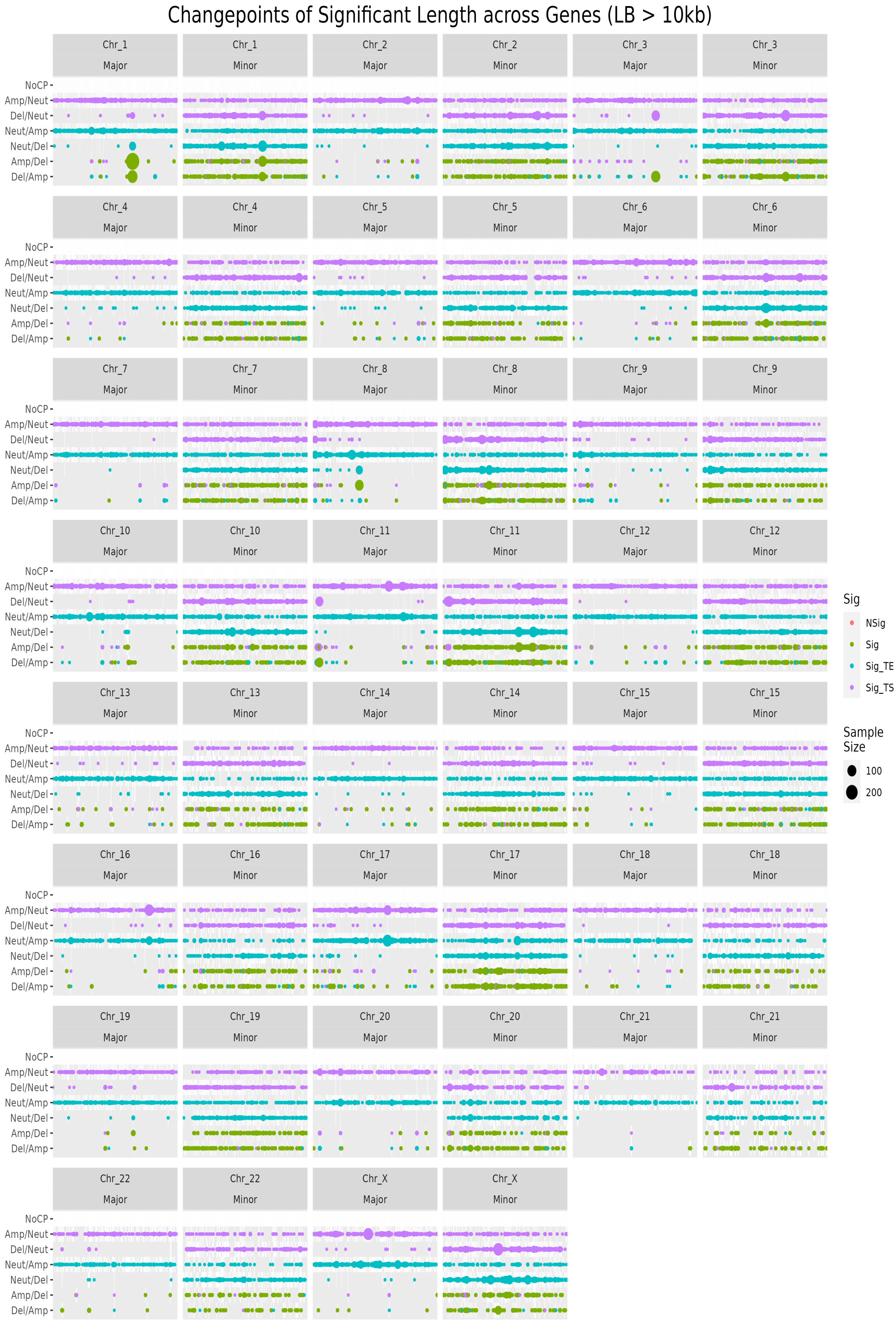


Figure 6.9: Application of univariate Allele-Dependent Intercept Model to each gene, providing confidence intervals for each category and allele. Each panel, corresponding to chromosome and allele, displays significance of changepoint determined by $LB > 10$kb. NoCP corresponds to NoChangepoint. Fitted using the `lm()` function.

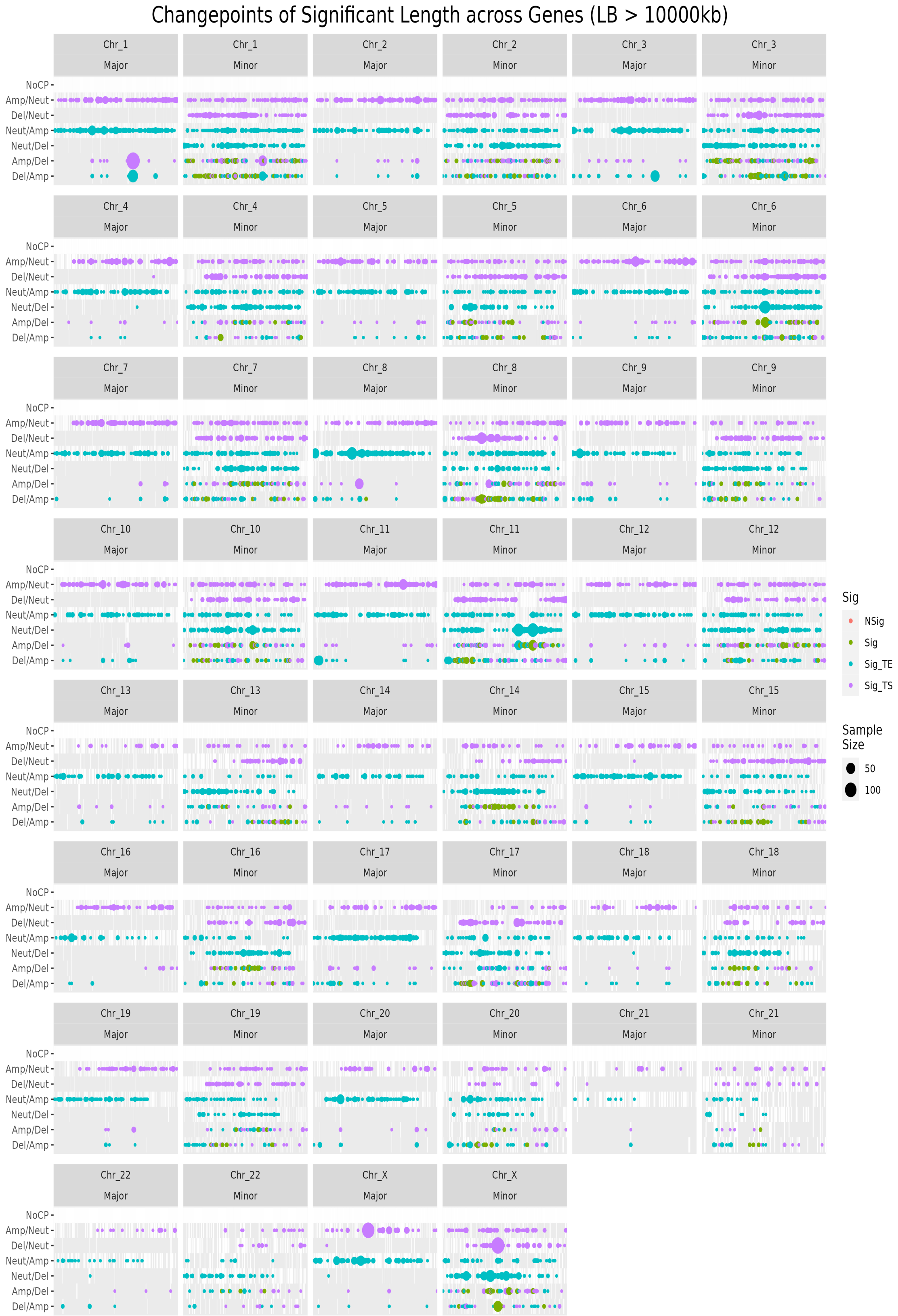


Figure 6.10: Application of univariate Allele-Dependent Intercept Model to each gene, providing confidence intervals for each category and allele. Each panel, corresponding to chromosome and allele, displays significance of changepoint determined by $LB > 10,000$kb. NoCP corresponds to NoChangepoint. Fitted using the `lm()` function.

Table 6.4: Top 20 genes containing largest changepoints of significant length with $n > 30$ and $LB > 10,000$kb. Models fitted using (A) `MCMCglmm()` and (B) `lm()` functions.

(A) Genes Containing Changepoints with Large CNAs (MCMCglmm)

Gene	Chr	Allele	Category	n	Direction	Fit	LB	UB
OR52N1	Chr_11	Major	Del/Amp	32	TE	102072.38	100628.80	103594.15
TRIM5	Chr_11	Major	Del/Amp	33	TE	99001.52	97448.98	100549.56
ALG1L2	Chr_3	Major	Del/Amp	52	TE	64209.25	63630.05	64814.50
ALG1L2	Chr_3	Minor	Del/Amp	31	TE	59239.36	58471.93	59963.17
PRIM2	Chr_6	Minor	Neut/Del	54	TE	43160.36	41489.02	44955.94
EYS	Chr_6	Minor	Neut/Del	70	TE	32790.35	31265.44	34366.85
PTPRD	Chr_9	Major	Neut/Amp	33	TE	29744.87	27557.61	31996.55
TENM4	Chr_11	Minor	Neut/Del	39	TE	29670.26	28635.39	30668.16
SHANK2	Chr_11	Minor	Amp/Del	33	TE	27178.81	25623.73	28801.81
CSMD1	Chr_8	Major	Neut/Amp	36	TE	26206.80	24161.67	28350.61
LCE1E	Chr_1	Major	Amp/Del	142	TS	26145.01	24328.77	27869.26
ADAM5	Chr_8	Major	Amp/Del	43	TS	25870.88	25193.66	26502.90
NRG1	Chr_8	Major	Neut/Amp	66	TE	25617.32	24137.16	27119.08
LCE1E	Chr_1	Minor	Amp/Del	44	TS	21611.82	18265.79	24758.43
SHANK2	Chr_11	Minor	Neut/Del	72	TE	21546.11	20504.38	22606.24
NRG1	Chr_8	Minor	Del/Neut	55	TS	20259.90	19802.39	20733.36
MACROD2	Chr_20	Major	Neut/Amp	33	TE	18939.02	18007.48	19871.15
DLG2	Chr_11	Minor	Neut/Del	83	TE	16967.40	16179.02	17826.49
PFKFB1	Chr_23	Major	Neut/Amp	37	TE	15899.73	15375.31	16427.05
EYS	Chr_6	Major	Amp/Neut	39	TS	15681.58	14541.24	16840.22
AGBL4	Chr_1	Major	Neut/Amp	33	TE	13085.08	11216.35	14727.44
OPHN1	Chr_23	Minor	Del/Neut	132	TS	12373.20	11997.76	12730.38
PFKFB1	Chr_23	Minor	Neut/Del	58	TE	12204.63	11788.63	12598.30
DLG2	Chr_11	Major	Amp/Neut	42	TS	12120.25	11113.47	13158.54
OPHN1	Chr_23	Major	Amp/Neut	116	TS	11633.17	11251.74	12027.07
NRG3	Chr_10	Minor	Neut/Del	31	TE	11293.13	10593.74	11981.01
HDAC8	Chr_23	Minor	Neut/Del	31	TE	10691.12	10159.36	11184.20
STARD8	Chr_23	Minor	Del/Neut	53	TS	10512.35	10332.99	10713.15

(B) Genes Containing Changepoints with Large CNAs (lm)

Gene	Chr	Allele	Category	n	Direction	Fit	LB	UB
OR52N1	Chr_11	Major	Del/Amp	32	TE	102085.79	100589.56	103582.01
TRIM5	Chr_11	Major	Del/Amp	33	TE	99002.76	97420.96	100584.56
ALG1L2	Chr_3	Major	Del/Amp	52	TE	64196.16	63610.32	64782.00
ALG1L2	Chr_3	Minor	Del/Amp	31	TE	59243.01	58484.26	60001.76
LCE3B	Chr_1	Major	Del/Amp	68	TE	55344.63	53618.81	57070.45
PRIM2	Chr_6	Minor	Neut/Del	54	TE	43181.24	41451.83	44910.65
EYS	Chr_6	Minor	Neut/Del	70	TE	32828.66	31248.70	34408.63
PTPRD	Chr_9	Major	Neut/Amp	33	TE	29746.90	27548.40	31945.40
TENM4	Chr_11	Minor	Neut/Del	39	TE	29677.27	28670.24	30684.30
SHANK2	Chr_11	Minor	Amp/Del	33	TE	27188.63	25570.08	28807.18
CSMD1	Chr_8	Major	Neut/Amp	36	TE	26201.59	24114.19	28289.00
LCE1E	Chr_1	Major	Amp/Del	142	TS	26156.68	24385.10	27928.27
ADAM5	Chr_8	Major	Amp/Del	43	TS	25873.25	25217.12	26529.38
NRG1	Chr_8	Major	Neut/Amp	66	TE	25631.60	24141.38	27121.82
LCE1E	Chr_1	Minor	Amp/Del	44	TS	21624.40	18441.82	24806.98
SHANK2	Chr_11	Minor	Neut/Del	72	TE	21555.90	20460.14	22651.67
NRG1	Chr_8	Minor	Del/Neut	55	TS	20262.63	19779.57	20745.69
MACROD2	Chr_20	Major	Neut/Amp	33	TE	18933.61	17982.14	19885.08
DLG2	Chr_11	Minor	Neut/Del	83	TE	16973.08	16135.20	17810.95
PFKFB1	Chr_23	Major	Neut/Amp	37	TE	15894.95	15367.21	16422.70
EYS	Chr_6	Major	Amp/Neut	39	TS	15687.13	14586.35	16787.92
AGBL4	Chr_1	Major	Neut/Amp	33	TE	13072.33	11345.30	14799.37
OPHN1	Chr_23	Minor	Del/Neut	132	TS	12378.88	12005.90	12751.86
PFKFB1	Chr_23	Minor	Neut/Del	58	TE	12209.63	11788.11	12631.14
DLG2	Chr_11	Major	Amp/Neut	42	TS	12129.45	11086.15	13172.76
OPHN1	Chr_23	Major	Amp/Neut	116	TS	11631.57	11233.70	12029.44
NRG3	Chr_10	Minor	Neut/Del	31	TE	11288.56	10593.66	11983.46
HDAC8	Chr_23	Minor	Neut/Del	31	TE	10698.71	10181.24	11216.19
STARD8	Chr_23	Minor	Del/Neut	53	TS	10513.67	10325.24	10702.10

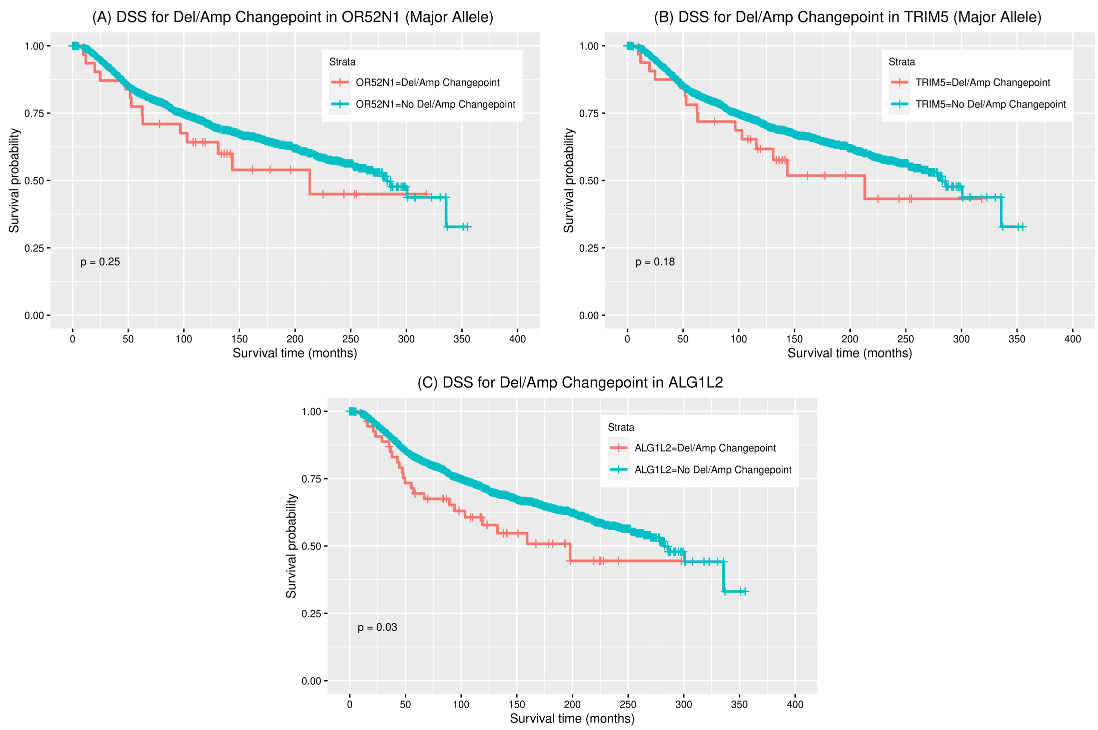


Figure 6.11: Survival curves for changepoints in selected genes. (A) OR52N1 (B) TRIM5 and (C) ALG1L2.

6.2 Whole-genome Allele-specific Changepoints across Chromosomes

Genomic regions attributed to genes, as analysed in the last section, make up only a small proportion of the total length of the human genome. The ASCAT data provides us with whole genome allele-specific copy number profiles, from which the gene regions were extracted previously. In this section, we carry out a broader analysis across the whole genome by segmenting the genome into consecutive equal distances of a pre-determined value d. Applying ADIM within each segmented region detects changepoints across the whole genome region with significant length of the changepoint states, TS and/or TE.

6.2.1 Genome Segmentation

Figure 6.12 displays the frequency and category of changepoints across the genome. To determine over what distance d the AD model will be applied, consideration is given to number of options including a sliding window approach, a per chromosome/chromosome arm approach or a segmentation approach, where the genome is split into segments of constant length using varying distances, e.g. setting d= 5,000kb, 5 million bases. For the genome segmentation method, segmentation is applied across each chromosome, starting at the first observed genomic location and ending at the last observed genomic location. Each segment will be the same length across the chromosome except for the last segment which will be of varying length.

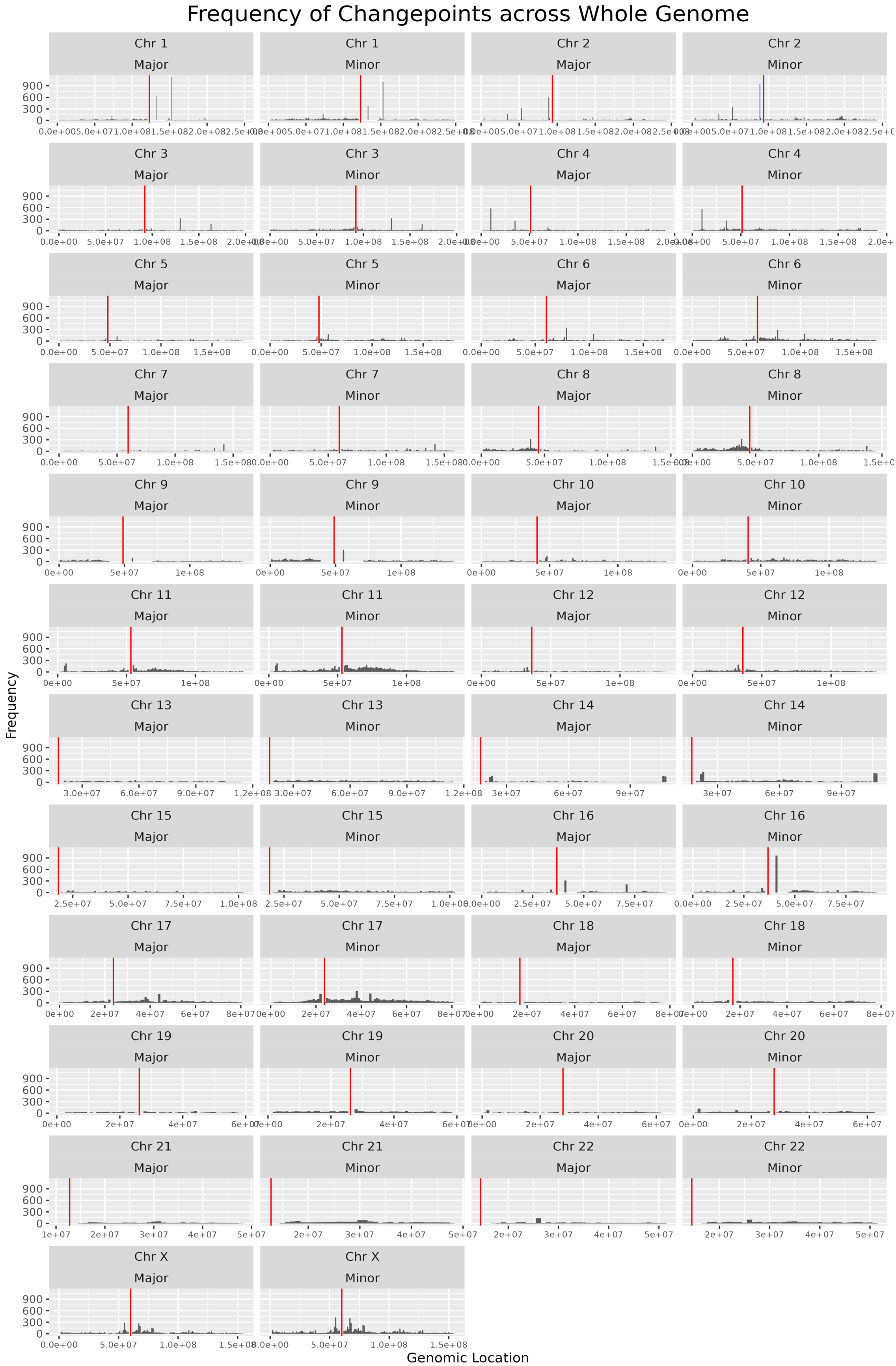


Figure 6.12: Frequency of changepoints across the whole genome for each chromosome and allele. Scale of y-axis is the same across all chromosomes and the red line indicates the midpoint of the centromere.

6.2.2 Outcomes of Selected Models to Segmented Regions

Out of 603 segmented regions of d=5,000kb, nine do not display any changepoint and are omitted from the analysis. For the remaining 594 genomic segments, the AD model is successfully applied to 591 genomic segments using the `MCMCglmm()` function. The three failed segments are segment 29 on chromosome 1, segment 28 on chromosome 10 and segment 8 on chromosome 16, and these likely require a stronger or proper prior to be implemented. The tile plot highlighting segments across the genome containing changepoints with TS or TE greater than 10kb indicates that most segments contain at least one changepoint with average length greater than 10kb (Figure 6.13). While the presence of changepoints is widespread, chromosomes containing changepoints with a large number of observations and TS and/or TE lengths significantly greater than 10kb include chromosome 1, 2, 8, 11, 16, 17 and X (Figure 6.13).

Similarly, the tile plot showing segments across the genome containing changepoints with TS and/or TE lengths greater than 10,000kb indicates that despite the increased average length threshold from 10kb to 10,000kb, a large number of segments still contain at least one changepoint of average length greater than 10,000kb in at least one patient (Figure 6.14). Chromosomes containing notable changepoints - those with a large number of observations and average length significantly greater than 10,000kb - include chromosome 1, 8, 11, 16 and X (Figure 6.14).

Fitting the multivariate ADIM MCMCglmm indicates 591 unique segments across the genome contain at least one changepoint with an average length significantly greater than 10kb, with 588 having an average length over 10,000kb. Focusing on the AD models imposing sample size filtering, $n > 100$ and $n > 200$, results in 97 and 22 unique segments across the genome containing changepoints with an average length significantly greater than 10kb and 38 and 11 unique segments across the genome containing changepoints with an average length significantly greater than 10,000 kb.

Table 6.5 provides information on the 20 non-unique genomic segments containing changepoints with average alteration length $> 10,000$kb, filtered for $n > 200$ observed changepoints in the segment, and is identical to the table that would be produced for the top 20 non-unique genomic segments containing changepoints with average alteration length $>$ 10kb and $n > 200$. Observations of note include, segment 27 on chromosome 1 containing 611 changepoint observations with significant mean amplification lengths on the Major allele (estimated mean TE = 86,130kb), segment 31 on chromosome 1 containing 405 and 395 changepoint observations with significant mean amplification lengths on the Major allele (estimated mean TE = 66,886kb and TS = 56,691kb) and segment 9 on chromosome 16 containing 607 changepoint observations with significant mean deletion lengths on the Minor allele (estimated mean TS = 44,768kb). As expected, chromosome 1, the longest chromosome, contains a number of genomic segments containing changepoints with large alterations on average.

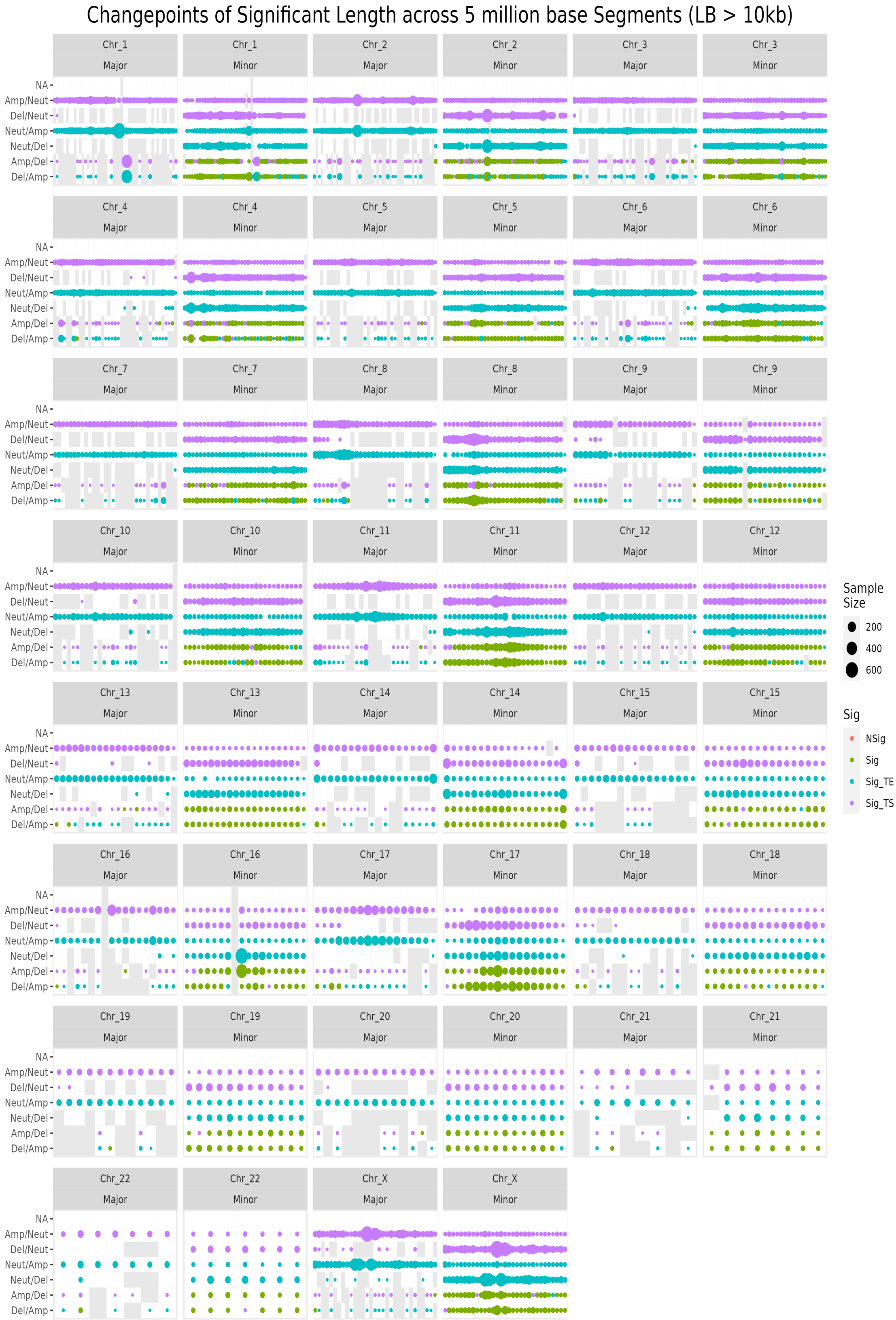


Figure 6.13: Application of multivariate Allele Dependent Intercept Model to each segment, providing prediction intervals for each category and allele. Each panel, corresponding to chromosome and allele, displays significance of changepoint, determined by $LB > 10$kb. NoCP corresponds to NoChangepoint. Fitted using the `MCMCglmm()` function.

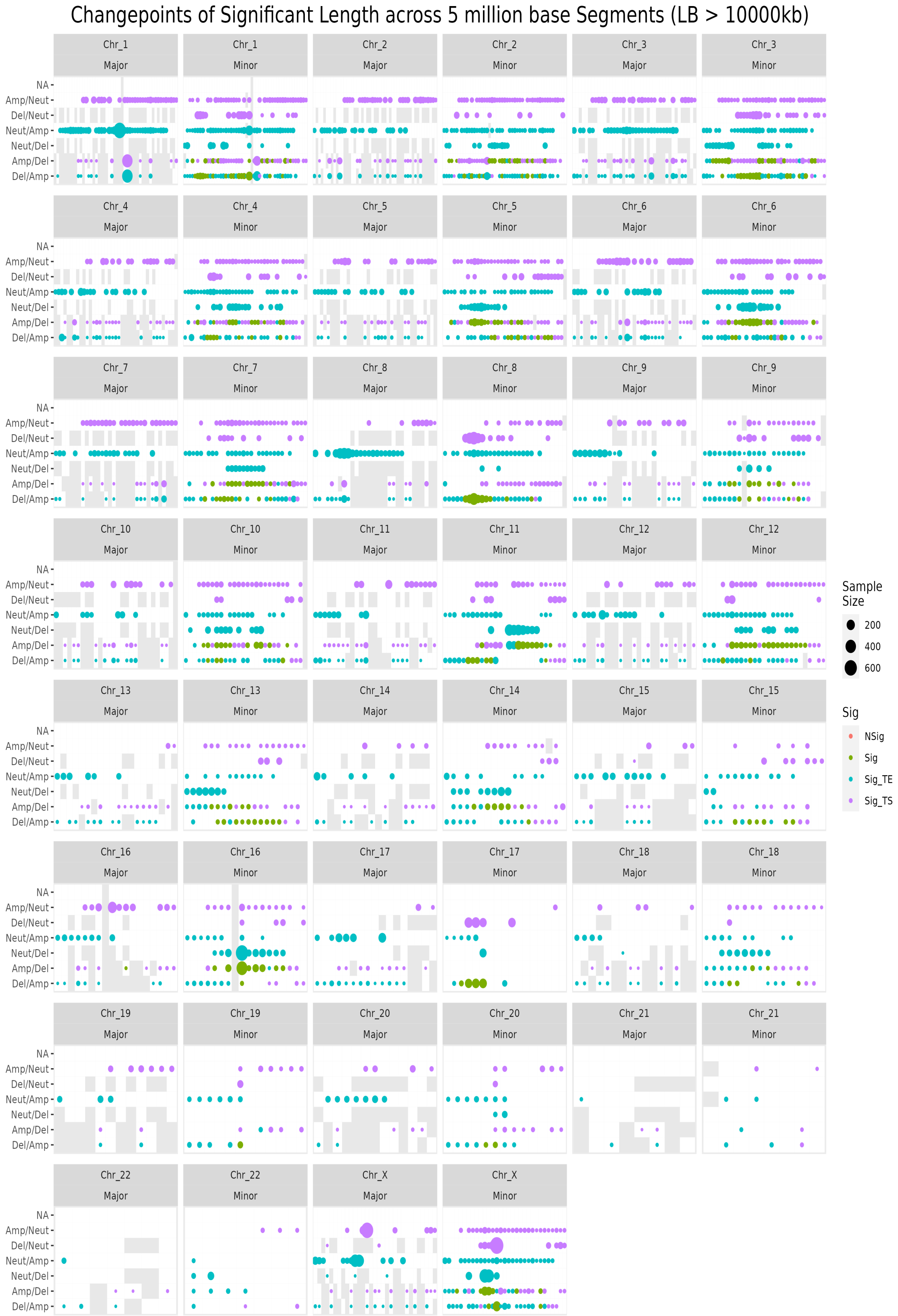


Figure 6.14: Application of multivariate Allele Dependent Intercept Model to each segment, providing prediction intervals for each category and allele. Each panel, corresponding to chromosome and allele, displays significance of changepoint, determined by $LB > 10,000$kb. NoCP corresponds to NoChangepoint. Fitted using the `MCMCglmm()` function.

Exploring further the identified points of focus, Neut/Amp changepoint on the Major allele in segment 27 on chromosome 1, Del/Amp changepoint on the Major allele in segment 31 on chromosome 1 and Neut/Del changepoint on the Minor allele in chromosome 16 segment 9, KM survival curves for DSS outcome are produced and compared for patients who exhibit or don't exhibit this particular changepoint profile, Figure 6.11. Applying the log-rank test for each point of focus, indicates patients with a Neut/Amp changepoint in chromosome 1 segment 27, and patients with a Neut/Del changepoint in chromosome 16 segment 9, have better DSS outcomes than patients that do not exhibit that changepoint ($p < 0.0001$). This is counter-intuitive as large genomic CNA burden is often associated with worse survival outcome. However, there are a number of explanations for this: the KM curves only consider one type of changepoint at a time, i.e. compares survival of patients who have specific CNA changepoint with patients who do not, even though there could be other changepoints events occurring in a genomic region simultaneously, the KM curves do not consider any other variables such as clinical variables, and whole genome duplication events or changepoints with large lengths occurring in adjacent segments (resulting in CNAs spanning the length of the segment) are not detected.

Table 6.5: Genomic segments containing changepoints with $n > 200$ and $LB > 10,000$kb from models fitted using `MCMCglmm()` function.

Segments Containing n > 200 Changepoints with CNAs > 10,000kb

Chr	Segment	Allele	Category	n	Direction	Fit	LB	UB
Chr_1	Seg_27	Major	Neut/Amp	611	TE	86129.93	84503.49	87744.89
Chr_1	Seg_31	Major	Del/Amp	405	TE	66886.34	65511.59	68373.03
Chr_1	Seg_31	Major	Amp/Del	395	TS	56691.46	54708.60	58549.17
Chr_16	Seg_9	Minor	Neut/Del	607	TE	44768.47	44273.21	45272.50
Chr_16	Seg_9	Minor	Amp/Del	225	TE	42367.76	41456.35	43144.38
Chr_8	Seg_8	Major	Neut/Amp	235	TE	39656.52	37510.54	41750.72
Chr_16	Seg_9	Minor	Amp/Del	225	TS	37930.30	37510.47	38388.41
Chr_16	Seg_9	Major	Amp/Neut	259	TS	37888.49	37497.53	38278.06
Chr_8	Seg_7	Major	Neut/Amp	205	TE	30156.30	28406.19	31723.52
Chr_11	Seg_17	Minor	Neut/Del	217	TE	23199.41	22298.65	24041.52
Chr_11	Seg_15	Minor	Neut/Del	260	TE	19723.00	18749.49	20792.80
Chr_17	Seg_5	Minor	Del/Neut	217	TS	18830.97	18578.32	19094.22
Chr_8	Seg_7	Minor	Del/Neut	250	TS	18345.07	17842.03	18857.78
Chr_8	Seg_8	Minor	Del/Neut	353	TS	15479.94	14830.76	16187.34
Chr_X	Seg_11	Major	Neut/Amp	339	TE	14856.14	14077.89	15590.49
Chr_X	Seg_11	Minor	Neut/Del	425	TE	13853.58	13175.03	14539.80
Chr_X	Seg_12	Minor	Neut/Del	386	TE	13428.15	12715.79	14180.54
Chr_X	Seg_12	Major	Neut/Amp	322	TE	13368.52	12528.00	14163.56
Chr_X	Seg_14	Major	Amp/Neut	603	TS	12423.92	11940.68	12838.79
Chr_X	Seg_14	Minor	Del/Neut	738	TS	12319.03	11912.35	12715.60

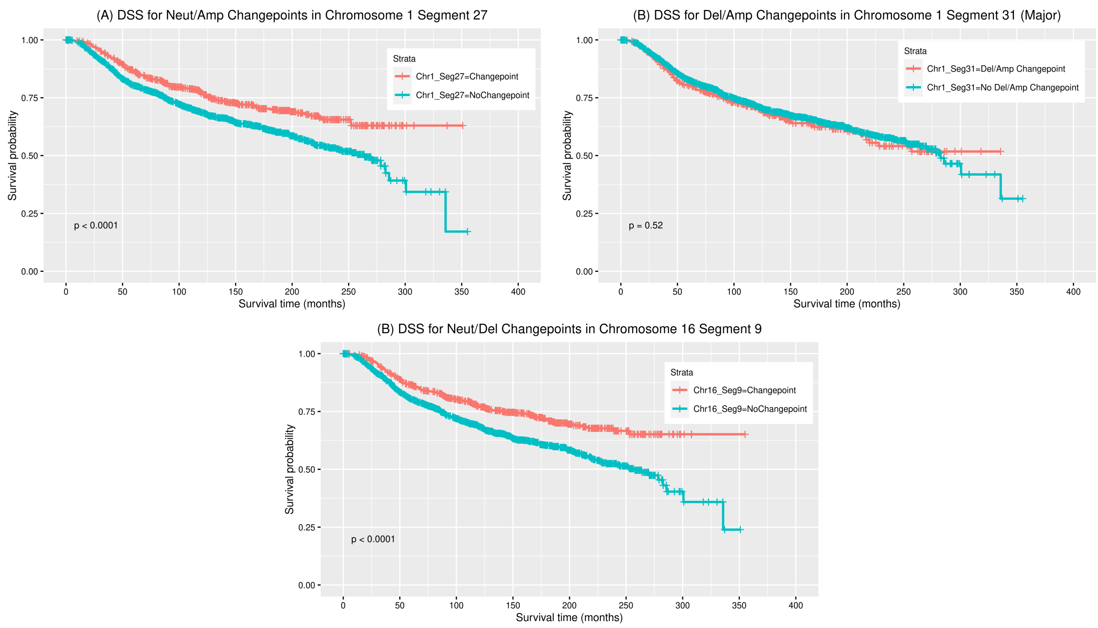


Figure 6.15: Survival Curves for changepoints in (A) Chromosome 1 Segment 27 (B) Chromosome 1 Segment 31 and (C) Chromosome 16 Segment 9.

6.3 Conclusion

Allele-specific copy number profiling provides information on genome wide copy number for each allele and tackles some of the limitations of total copy number profiling, including masking of changepoints and being unable to detect certain types of genomic aberrations, such as LOH and copy number-neutral events.

In this chapter we produced allele-specific copy number profiles, ASCAT profiles, for 1,984 METABRIC patients. Comparing allele-specific copy number profiles to the total copy number profiles, as produced in Chapter 3, using heatmaps of CNA states, similarities were observed, but the allele-specific copy number offered additional insight, by displaying high level of amplifications, possibly indicative of whole genome or chromosome duplication.

The ADIM was developed in Chapter 5 to identify genomic regions displaying significant changepoints along the allele-specific profile, with significant lengths of state before (TS) or after (TE) the changepoint. ADIM is applicable within a pre-defined genomic region d, and in applying to the METABRIC cohort, we focused on two approaches, a gene-centric application, where d is defined as the region of the gene, and assumed different in size for different genes, and a whole-genome segmentation application, where d is set to a fixed length and the application searches over consecutive regions of length d. These applications ensure whole-genome coverage and identifies specific genes and non-gene regions of interest. KM curves were produced to take an exploratory look at how these changepoints may influence survival outcome.

Genes including OR52N1, TRIM5, ALG1L2, LCE3B, PRIM2 and EYS were among the genes displaying changepoints, with > 30 observations, that had an average TS and/or TE length greater than 10,000kb. These genes contain both a changepoint and region(s) of altered copy number, potentially disrupting gene function. Whole genome segments, including chromosome 1 segment 27, chromosome

1 segment 31 and chromosome 6 segment 19, were among the genomic segments displaying changepoints, with > 200 observations, that had an average TS and/or TE length greater than 10,000kb.

Overall, it is clear that CNAs affect a large proportion of the cancer genome, with these CNAs occurring through whole-region duplication/deletion, or via a copy number change at some point along the genome, resulting in a CNA associated changepoint.

7 Conclusions and Future Work

The aim of this thesis was to assess the role of genomic data (CNA data) in stratifying patients within predictive models for breast cancer OS and DSS outcomes. While studies of CNAs in breast cancer have been published in the literature and a large number of metrics defined to measure GI, limitations such as requiring access to raw or segmented data to calculate these measures, along with the complexity of some of these measures, have hindered their widespread use in research and clinical settings. This thesis has revealed the potential in incorporating CNA information, with lots of avenues for further research questions.

In Chapter 2 we proposed a number of easy to interpret GI measures that can be calculated using publicly available summary CNA data. These CNA Score and Burden metrics captured the main aspects of CNAs, including magnitude, type and genomic location. Exploring the distributions of these CNA metrics, overall and stratified by PAM50 and IntClust molecular classifications, highlighted characteristic genomic aberrations documented previously, such as 5q deletions in Basal tumours and 17q amplification in HER2 tumours. In published research, subtypes associated with worse OS and DSS, i.e. Basal, HER2 and Luminal B subtypes, have higher GI than subtypes associated with better survival outcomes (Curtis et al., 2012). This result was echoed in the analysis here, when applying the new CNA metrics, but, in addition, we also reveal that subtypes associated with worse OS and DSS, tend to have significantly higher levels of deletions in genes than amplifications.

Chapter 3 focused on how these CNA Score and Burden metrics are associated with survival, primarily DSS. Applying a combination of KM estimators, CPH regression and recursive partitioning survival trees, it was found that the global Absolute CNA Score metric was associated with DSS in Luminal A breast cancer patients. Patients with higher Absolute CNA Score values, i.e. Absolute CNA Score Quartile 4, indicative of higher levels of GI, had worse survival outcomes than patients with less GI, in Absolute CNA Score Quartiles 1-3 (Q1-3). This is encouraging, given reports in literature that the Luminal A and Luminal B subtypes may not be distinct subtypes, with ambiguity existing in DSS outcome for Luminal A patients (Tishchenko et al., 2016; Sung et al., 2016; Kumar et al., 2019; Wang and Lee, 2023), as it suggests that a simple measure of gene-centric CNAs across the genome can help identify Luminal A patients who are at elevated risk. These results are published in *Survival Outcomes are Associated with Genomic Instability in Luminal Breast Cancer* (King et al., 2021a). We expanded the analysis further to assess the association between the 12 CNA Score and Burden metrics, formulated in this thesis, and DSS outcome, with consideration to PAM50 and IntClust molecular classifications. It was revealed that global CNA Del Score and Burden metrics, and chromosome arm CNA Del Score and Burden, specific to chromosomes 3p and 18q, play a role in stratifying patients on DSS outcome, primarily within Luminal A and Claudin-low patients. This again suggests that deletions are more harmful than amplifications and measuring these using CNA metrics can help identify patients with poorer survival outcomes. To facilitate navigation and exploration of the potential to unlock these discoveries, an R shiny app, GNOSIS, was developed to support the tractable and efficient exploration and application of survival analysis to cBioPortal clinical and genomic data products. This development is presented, accessible and published with peer review in *GNOSIS: an R Shiny app supporting cancer genomics* *survival analysis with cBioPortal* (King et al., 2022).

While it was shown that our CNA Score and Burden metrics are accessible, easy to interpret and can provide valuable insights about how GI can impact survival outcomes, when comparing to pre-existing CNA measures, a number of limitations exist. These limitations, not encountered in most of the CNA measures discussed in Chapter 2, include that our CNA metrics are calculated using CNA summary data for annotated genes only, meaning the length of the CNA is not considered and CNAs occurring outside of gene regions are not included.

To assess the effect that CNAs may have on gene expression, DGEA was carried out comparing the gene expression of groups of patients, stratified by similarities in survival outcome, where CNA metrics have a role to play in the stratification. This analysis, presented in Chapter 4, identified genes displaying significant differential gene expression between patient groups. This research revealed genes up-regulated among patients in survival tree nodes associated with poorer DSS outcomes include UBE2C, CXCL10 and S100P, while genes observed to be down-regulated among patients in survival tree nodes associated with poorer DSS outcomes include PIP, BCL2 and IRX2. Misexpression of these genes has been documented in literature as facilitating cell proliferation, tumour progression and invasion and as being correlated with survival (Andersen et al., 2011; Dastsooz et al., 2019; Huang et al., 2021; Dawson et al., 2010; Werner et al., 2015; Urbaniak et al., 2018). To investigate the direct relationship of a gene's CNA state to the gene's expression, this work employed a modified limma pipeline, comparing gene expression profiles across patients given the CNA state of the gene. A large number of genes were identified where the presence of a CNA in the gene led to an up- or down-regulation of that gene. As expected, if a deletion occurred the gene expression was usually down-regulated while if an amplification occurred the gene expression was usually up-regulated. Overall, using specified thresholds and considering sample size restrictions, 1,104 genes were differentially expressed in the three-gene specifications, ModLim3, and 3,197 genes were differentially expressed in the five-gene specification, ModLim5. Comparing these gene-sets to prognostic and predictive assays published in the literature indicated a moderate degree of congruence, identifying some of the same genes, but also identifying additional genes to be considered for further investigation as candidate biomarkers for breast cancer treatment and outcome.

While we have shown that total copy number has a role in predictive models for survival outcome, there are a number of limitations associated with measuring copy number as a total across the two alleles, including masking of certain types of genomic aberrations and CNA changepoints. The aim of Chapters 5 and 6 was to generate allele-specific copy number data for the METABRIC patients and to detect regions of the genome where copy number changes of significant length occurred. To accomplish this a number of models, AIIM, AINIM, ADIM and ADNIM, were proposed and their performance assessed in a simulation study. Overall, based on a number of considerations, the ADIM model was selected and applied to the allele-specific copy number profiles. Unlike Pladsen et al. (2020), we retained valuable information on the type of CNA observed and the allele upon which the copy number change occurred, for each patient. Furthermore, rather than simply identify changepoints based on their observed frequency, van den Broek and van Lieshout (2023), our proposed models are based on TS and TE lengths for each changepoint category and are able to identify the changepoint categories accompanied by significantly large TS or TE lengths. Applying ADIM to the METABRIC cohort, Chapter 6, highlighted a number of genes and genomic regions where CNA changepoints, with large average TS or TE alterations, occurred. With application focusing on genomic lengths of genes, genes identified as containing changepoints with significant TS and/or TE lengths include OR52N1, TRIM5 and ALG1L2. Applying the ADIM with segmentation applied to the entire genomic region, genomic segments containing changepoints with significant TS and/or TE lengths include chromosome 1 segment 27, chromosome 1 segment 31 and chromosome 16 segment 9. Interestingly, comparing survival curves, for patients with/without the identified CNA changepoint of focus indicates that changepoints occurring in gene regions may be associated with poorer DSS, while changepoints occurring in genomic segments may be associated with improved DSS. These results with opposite direction of effect suggest that there is much still unknown and yet to explore with regards to CNAs, their changepoints and survival.

7.1 Future Work

While this thesis offers a comprehensive analysis of total and allele-specific copy number in the METABRIC cohort, and their associations with survival, the research offers opportunity for further work.

Throughout this thesis we only consider OS and DSS outcomes, and not Recurrence-free survival (RFS). As RFS outcomes are also available for a large proportion of the METABRIC patients, a similar analysis assessing the association between our CNA Score and Burden metrics, and the allele-specific copy number changepoints, and RFS may lead to new insights. In addition, treatment information, including whether a patient received chemotherapy, radiotherapy, hormone therapy, and type of surgery, were not included in our analyses. Inclusion of this information may highlight CNA motifs that confer resistance to certain therapies. Although it should be noted that the METABRIC patients were enrolled between 1977 and 2005 and treatment options and standard of care have changed since e.g. use of Herceptin in HER2+ patients. Therefore, it may be interesting to produce and compare CNA Score and Burden metrics across patients in another more recent breast cancer dataset, for which detailed treatment information is available.

Further to this, assessing the performance of our predictive models including the CNA Score and Burden metrics, molecular subtype classifications and selected clinical information as candidate predictors, and comparison with performance of the predictive models only including the molecular subtype and clinical information, could be carried out.

The gene expression analyses carried out using limma (the survival tree node analysis and CNA state analysis) were performed including only gene expression and CNA state in our models. In addition, no batch correction was carried out. Future work may include application of other gene expression methodologies, e.g. Significance Analysis of Microarrays, expansion to consider other variables of interest, implementation of batch correction techniques, and the possibility of combining gene expression data and our CNA metrics in predictive models for survival outcome.

In application of the changepoint detection, the definition of a NoChangepoint region is a region of no alteration in CNA state in an allele, i.e. a constant CNA state observed within that region. This NoChangepoint region, of course, would be a constant state of neutral "normal" copy number, but would also represent a constant state of increased copy number, amplification, or a constant state of decreased copy number, deletion. A constant state of amplification or constant state deletion is still a region of copy number alteration, in contrast to normal copy number. While the application of the model to consecutive segmented regions as a form of search across the genome for changepoints does not necessarily require this as a point of focus, an expansion of the modelling approach could be considered, in order to include the three Nochangepoint categories, NoChangepoint, NoChangepoint_Amp and NoChangepoint_Del. In addition, during preprocessing of the ASCAT data the copy numbers of each allele were bound in the range [0-2], resulting in copy number changes for amplified regions being missed. These changepoints could potentially provide valuable insights and their impact on survival and treatment response should be investigated in future work.

Application of the AD models, were under usual default modelling assumptions of `lm()` and default priors when fitting `MCMCglmm()`. Although the `MCMCglmm()` function enables fitting of random effects, applicability of random effects to the METABRIC data were not given further consideration within the scope of this body of research. Further work could give opportunity to explore suitability of assumptions and assumed priors, and application of random effects structures if required.

The methods and results in this thesis, researching the role of CNA metrics and allele specific information in prognostic models, has led to a number of interesting outcomes and produced lists of candidate genes or genomic regions from which further exploration, in a research or clinical setting, can be carried out. One of the most important avenues of future work is the validation of these results in another breast cancer dataset, confirming our findings are dataset independent and of potential benefit. The code to run the analyses presented in this thesis is provided on GitHub at: https://github.com/Lydia-King/PhD_Thesis.

Bibliography

Affymetrix. *Genome-Wide Human SNP Array 6.0*, 2009. URL `https://www.cancer.gov/ccg/research/structural-genomics/tcga/using-tcga-data/technology/affymetrix-snp6-data-sheet`.

M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan. Awareness and current knowledge of breast cancer. *Biol Res*, 50(1):33, Oct 2017.

R. C. Allshire and G. H. Karpen. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat Rev Genet*, 9(12):923–937, Dec 2008.

K. Andersen, H. Mori, J. Fata, J. Bascom, T. Øyjord, G. M. Mælandsmo, and M. Bissell. The metastasis-promoting protein S100A4 regulates mammary branching morphogenesis. *Dev Biol*, 352(2):181–190, Apr 2011.

H. Bengtsson. *R.utils: Various Programming Utilities*, 2022. URL `https://CRAN.R-project.org/package=R.utils`. R package version 2.12.2.

R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff, and W. R. Sellers. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A*, 104(50):20007–20012, Dec 2007.

T. B. Bevers, B. O. Anderson, E. Bonaccio, S. Buys, M. B. Daly, P. J. Dempsey, W. B. Farrar, I. Fleming, J. E. Garber, R. E. Harris, A. S. Heerdt, M. Helvie, J. G. Huff, N. Khakpour, S. A. Khan, H. Krontiras, G. Lyman, E. Rafferty, S. Shaw, M. L. Smith, T. N. Tsangaris, C. Williams, T. Yankeelov, T. B. Bevers, B. O. Anderson, E. Bonaccio, S. Buys, M. B. Daly, P. J. Dempsey, W. B. Farrar, I. Fleming, J. E. Garber, R. E. Harris, A. S. Heerdt S, M. Helvie, J. G. Huff, N. Khakpour, S. A. Khan, H. Krontiras, G. Lyman, E. Rafferty, S. Shaw, M. L. Smith, T. N. Tsangaris, C. Williams, and T. Yaneeklov. NCCN clinical practice guidelines in oncology: breast cancer screening and diagnosis. *J Natl Compr Canc Netw*, 7(10):1060–1096, Nov 2009.

A. Bhattacharya, R. D. Bense, C. G. Urzúa-Traslaviña, E. G. E. de Vries, M. A. T. M. van Vugt, and R. S. N. Fehrmann. Transcriptional effects of copy number alterations in a large set of human cancers. *Nat Commun*, 11(1):715, Feb 2020.

G. Blair, J. Cooper, A. Coppock, M. Humphreys, A. Rudkin, and N. Fultz. *fabricatr: Imagine Your Data Before You Collect It*, 2022. URL `https://CRAN.R-project.org/package=fabricatr`. R package version 1.0.0.

J. M. Bland and D. G. Altman. The logrank test. *BMJ*, 328(7447):1073, May 2004.

F. M. Blows, K. E. Driver, M. K. Schmidt, A. Broeks, F. E. van Leeuwen, J. Wesseling, M. C. Cheang, K. Gelmon, T. O. Nielsen, C. Blomqvist, P. Heikkilä, T. Heikkinen, H. Nevanlinna, L. A. Akslen, L. R. Bégin, W. D. Foulkes, F. J.

Couch, X. Wang, V. Cafourek, J. E. Olson, L. Baglietto, G. G. Giles, G. Severi, C. A. McLean, M. C. Southey, E. Rakha, A. R. Green, I. O. Ellis, M. E. Sherman, J. Lissowska, W. F. Anderson, A. Cox, S. S. Cross, M. W. Reed, E. Provenzano, S. J. Dawson, A. M. Dunning, M. Humphreys, D. F. Easton, M. García-Closas, C. Caldas, P. D. Pharoah, and D. Huntsman. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med*, 7(5):e1000279, May 2010.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Stat Surv*, 5(none), 2011. doi: 10.1214/09-ss047.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. CRC Press, 1984.

C. Brown. *operator.tools: Utilities for Working with R's Operators*, 2017. URL `https://CRAN.R-project.org/package=operator.tools`. R package version 1.6.3.

S. L. Carter, A. C. Eklund, I. S. Kohane, L. N. Harris, and Z. Szallasi. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet*, 38(9):1043–1048, Sep 2006.

E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5):401–404, May 2012.

W. Chang and B. Borges Ribeiro. *shinydashboard: Create Dashboards with 'Shiny'*, 2021. URL `https://CRAN.R-project.org/package=shinydashboard`. R package version 0.7.2.

W. Chang, J. Cheng, JJ. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. *shiny: Web Application Framework for R*, 2022. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.7.4.

H. Chen, J. M. Bell, N. A. Zavala, H. P. Ji, and N. R. Zhang. Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res*, 43(4): e23, Feb 2015.

J. Cheng and C. Sievert. *shinymeta: Export Domain Logic from Shiny using Meta-Programming*, 2021. URL `https://CRAN.R-project.org/package=shinymeta`. R package version 0.2.0.3.

S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. van de Wiel, A. R. Green, I. O. Ellis, P. L. Porter, S. Tavaré, J. D. Brenton, B. Ylstra, and C. Caldas. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol*, 8(10):R215, 2007.

Y. Ciani, T. Fedrizzi, D. Prandi, F. Lorenzin, A. Locallo, P. Gasperini, G. M. Franceschini, M. Benelli, O. Elemento, L. L. Fava, A. Inga, and Demichelis F. Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. *Cell Syst*, 13(2):183–193.e7, 2022. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2021.10.001. URL `https://www.sciencedirect.com/science/article/pii/S2405471221003835`.

D. R. Cox. Regression models and life-tables. *J R Stat Soc Series B Methodol*, 34(2): 187–220, 1972. ISSN 00359246. URL `http://www.jstor.org/stable/2985181`.

M. J. Crowther and P. C. Lambert. A general framework for parametric survival analysis. *Stat Med*, 33(30):5280–5297, Dec 2014.

C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, S. Aparicio, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, Apr 2012.

I. Cutcutache, A. Y. Wu, Y. Suzuki, J. R. McPherson, Z. Lei, N. Deng, S. Zhang, W. K. Wong, K. C. Soo, W. H. Chan, L. L. Ooi, R. Welsch, P. Tan, and S. G. Rozen. Abundant copy-number loss of CYCLOPS and STOP genes in gastric adenocarcinoma. *Gastric Cancer*, 19(2):453–465, Apr 2016.

H. Dastsooz, M. Cereda, D. Donna, and S. Oliviero. A Comprehensive Bioinformatics Analysis of UBE2C in Cancers. *Int J Mol Sci*, 20(9), May 2019.

S. J. Dawson, N. Makretsov, F. M. Blows, K. E. Driver, E. Provenzano, J. Le Quesne, L. Baglietto, G. Severi, G. G. Giles, C. A. McLean, G. Callagy, A. R. Green, I. Ellis, K. Gelmon, G. Turashvili, S. Leung, S. Aparicio, D. Huntsman, C. Caldas, and P. Pharoah. BCL2 in breast cancer: a favourable prognostic marker across molecular subtypes and independent of adjuvant therapy received. *Br J Cancer*, 103(5):668–675, Aug 2010.

S. J. Dawson, O. M. Rueda, S. Aparicio, and C. Caldas. A new genome-driven integrated classification of breast cancer and its implications. *EMBO J*, 32(5): 617–628, Mar 2013.

M. J. Duffy, N. Harbeck, M. Nap, R. Molina, A. Nicolini, E. Senkus, and F. Cardoso. Clinical use of biomarkers in breast cancer: Updated guidelines from the european group on tumor markers (egtm). *Eur J Cancer*, 75:284–298, 2017. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2017.01.017. URL `https://www.sciencedirect.com/science/article/pii/S0959804917300758`.

P. H. G. Duijf, D. Nanayakkara, K. Nones, S. Srihari, M. Kalimutho, and K. K. Khanna. Mechanisms of Genomic Instability in Breast Cancer. *Trends Mol Med*, 25(7):595–611, Jul 2019.

C. Fougner, H. Bergholtz, J. H. Norum, and T. Sørlie. Re-definition of claudin-low as a breast cancer phenotype. *Nat Commun*, 11(1):1787, Apr 2020.

J. Fox and S. Weisberg. *An R Companion to Applied Regression.* Sage, Thousand Oaks CA, third edition, 2019. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

M. A. Freeberg, L. A. Fromont, T. D'Altri, A. F. Romero, J. I. Ciges, A. Jene, G. Kerry, M. Moldes, R. Ariosa, S. Bahena, D. Barrowdale, M. C. Barbero, D. Fernandez-Orth, C. Garcia-Linares, E. Garcia-Rios, F. Haziza, B. Juhasz, O. M. Llobet, G. Milla, A. Mohan, M. Rueda, A. Sankar, D. Shaju, A. Shimpi, B. Singh, C. Thomas, S. de la Torre, U. Uyan, C. Vasallo, P. Flicek, R. Guigo, A. Navarro, H. Parkinson, T. Keane, and J. Rambla. The European Genome-phenome Archive in 2021. *Nucleic Acids Res*, 50(D1):D980–D987, Jan 2022.

J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*, 6(269):pl1, Apr 2013.

D. Glodzik, P. Selenica, R. A. Rogge, I. M. Silverman, D. Mandelker, S. Harris, J. Zhao, M. Zinda, A. Veloso, N. Malani, N. Riaz, M. Koehler, R. D. Daber, V. Johnson, V. Rimkunas, and J. S. Reis-Filho. Detection of Biallelic Loss of DNA Repair Genes in Formalin-Fixed, Paraffin-Embedded Tumor Samples Using a Novel Tumor-Only Sequencing Panel. *J Mol Diagn*, 25(5):295–310, May 2023.

D. Granjon. *shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'*, 2021. URL `https://CRAN.R-project.org/package=shinydashboardPlus`. R package version 2.0.3.

L. Guo, D. Kong, J. Liu, L. Zhan, L. Luo, W. Zheng, Q. Zheng, C. Chen, and S. Sun. Breast cancer heterogeneity and its implication in personalized precision therapy. *Exp Hematol Oncol*, 12(1):3, Jan 2023.

G. Ha and S. Shah. Distinguishing somatic and germline copy number events in cancer patient DNA hybridized to whole-genome SNP genotyping arrays. *Methods Mol Biol*, 973:355–372, 2013.

J. D. Hadfield. Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J Stat Softw*, 33(2):1–22, 2010. URL `https://www.jstatsoft.org/v33/i02/`.

D. Hamdan, T. T. Nguyen, C. Leboeuf, S. Meles, A. Janin, and G. Bousquet. Genomics applied to the treatment of breast cancer. *Oncotarget*, 10(46):4786–4801, Jul 2019.

D. Hanahan. Hallmarks of Cancer: New Dimensions. *Cancer Discov*, 12(1):31–46, Jan 2022.

C. A. Harrington, C. Rosenow, and J. Retief. Monitoring gene expression using DNA microarrays. *Curr Opin Microbiol*, 3(3):285–291, Jun 2000.

J. I. Herschkowitz, K. Simin, V. J. Weigman, I. Mikaelian, J. Usary, Z. Hu, K. E. Rasmussen, L. P. Jones, S. Assefnia, S. Chandrasekharan, M. G. Backlund, Y. Yin, A. I. Khramtsov, R. Bastein, J. Quackenbush, R. I. Glazer, P. H. Brown, J. E. Green, L. Kopelovich, P. A. Furth, J. P. Palazzo, O. I. Olopade, P. S. Bernard, G. A. Churchill, T. Van Dyke, and C. M. Perou. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol*, 8(5):R76, 2007.

J. Hicks, A. Krasnitz, B. Lakshmi, N. E. Navin, M. Riggs, E. Leibu, D. Esposito, J. Alexander, J. Troge, V. Grubor, S. Yoon, M. Wigler, K. Ye, A. L. Børresen-Dale, B. Naume, E. Schlicting, L. Norton, T. Hägerström, L. Skoog, G. Auer, S. Månér, P. Lundin, and A. Zetterberg. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res*, 16(12):1465–1479, Dec 2006.

H. Hieronymus, N. Schultz, A. Gopalan, B. S. Carver, M. T. Chang, Y. Xiao, A. Heguy, K. Huberman, M. Bernstein, M. Assel, R. Murali, A. Vickers, P. T. Scardino, C. Sander, V. Reuter, B. S. Taylor, and C. L. Sawyers. Copy number alteration burden predicts prostate cancer relapse. *Proc Natl Acad Sci U S A*, 111 (30):11139–11144, Jul 2014.

H. Hieronymus, R. Murali, A. Tin, K. Yadav, W. Abida, H. Moller, D. Berney, H. Scher, B. Carver, P. Scardino, N. Schultz, B. Taylor, A. Vickers, J. Cuzick, and C. L. Sawyers. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *Elife*, 7, Sep 2018.

T. Hothorn and A. Zeileis. partykit: A modular toolkit for recursive partytioning in R. *J Mach Learn Res*, 16:3905–3909, 2015. URL `https://jmlr.org/papers/v16/hothorn15a.html`.

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *J Comput Graph Stat*, 15(3):651–674, 2006. doi: 10.1198/106186006x133933.

T. Hothorn, K. Hornik, M. A. van de Wiel, and A. Zeileis. Implementing a class of permutation tests: The coin package. *J Stat Softw*, 28(8):1–23, 2008. doi: 10.18637/jss.v028.i08.

Z. Hu, C. Fan, D. S. Oh, J. S. Marron, X. He, B. F. Qaqish, C. Livasy, L. A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. G. Ewend, L. R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Ruiz Orrico, D. Dreher, J. P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. F. Quackenbush, M. J. Ellis, O. I. Olopade, P. S. Bernard, and C. M. Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7:96, Apr 2006.

H. Huang, W. Zhou, R. Chen, B. Xiang, S. Zhou, and L. Lan. CXCL10 is a Tumor Microenvironment and Immune Infiltration Related Prognostic Biomarker in Pancreatic Adenocarcinoma. *Front Mol Biosci*, 8:611508, 2021.

R. Iannone. *fontawesome: Easily Work with 'Font Awesome' Icons*, 2023. URL `https://CRAN.R-project.org/package=fontawesome`. R package version 0.5.1.

Illumina. *HumanHT-12 v3 Expression BeadChip*, 2010. URL `https://www.illumina.com/Documents/products/datasheets/datasheet_humanht_12.pdf`.

International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, Dec 2003.

P. L. Jerevall, X. J. Ma, H. Li, R. Salunga, N. C. Kesty, M. G. Erlander, D. C. Sgroi, B. Holmlund, L. Skoog, T. Fornander, B. Nordenskjöld, and O. Stål. Prognostic utility of HOXB13:IL17BR and molecular grade index in early-stage breast cancer patients from the Stockholm trial. *Br J Cancer*, 104(11):1762–1769, May 2011.

G. Jönsson, J. Staaf, J. Vallon-Christersson, M. Ringnér, K. Holm, C. Hegardt, H. Gunnarsson, R. Fagerholm, C. Strand, B. A. Agnarsson, O. Kilpivaara, L. Luts, P. Heikkilä, K. Aittomäki, C. Blomqvist, N. Loman, P. Malmström, H. Olsson, O. T. Johannsson, A. Arason, H. Nevanlinna, R. B. Barkardottir, and A. Borg. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res*, 12(3):R42, 2010.

M. Kalimutho, K. Nones, S. Srihari, P. H. G. Duijf, N. Waddell, and K. K. Khanna. Patterns of Genomic Instability in Breast Cancer. *Trends Pharmacol Sci*, 40(3): 198–211, Mar 2019.

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

A. Kassambara. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2023. URL `https://CRAN.R-project.org/package=rstatix`. R package version 0.7.2.

A. Kassambara, M. Kosinski, and P. Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2021. URL `https://CRAN.R-project.org/package=survminer`. R package version 0.4.9.

T. Z. Keith. *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling.* Routledge, 2019.

L. King. Lydia-king/gnosis: Gnosis, May 2022. URL `https://zenodo.org/record/6543416`.

L. King, A. Flaus, E. Holian, and A. Golden. Survival outcomes are associated with genomic instability in luminal breast cancers. *PLoS One*, 16(2):e0245042, 2021a.

L. King, A. Flaus, E. Holian, and A. Golden. Data associated with "survival outcomes are associated with genomic instability in luminal breast cancers"., Dec 2021b. URL `https://zenodo.org/record/5791192`.

L. King, A. Flaus, S. Coughlan, E. Holian, and A. Golden. GNOSIS: an R Shiny app supporting cancer genomics survival analysis with cBioPortal. *HRB Open Res*, 5:8, 2022.

D. G. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text.* Springer, 2012.

M. Knauer, S. Mook, E. J. Rutgers, R. A. Bender, M. Hauptmann, M. J. van de Vijver, R. H. Koornstra, J. M. Bueno-de Mesquita, S. C. Linn, and L. J. van 't Veer. The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer. *Breast Cancer Res Treat*, 120(3):655–661, Apr 2010.

N. Kumar, D. Zhao, D. Bhaumik, A. Sethi, and P. H. Gann. Quantification of intrinsic subtype ambiguity in Luminal A breast cancer and its relationship to clinical outcomes. *BMC Cancer*, 19(1):215, Mar 2019.

I. Lappalainen, J. Almeida-King, V. Kumanduri, A. Senf, J. D. Spalding, S. Ur-Rehman, G. Saunders, J. Kandasamy, M. Caccamo, R. Leinonen, B. Vaughan, T. Laurent, F. Rowland, P. Marin-Garcia, J. Barker, P. Jokinen, A. C. Torres, J. R. de Argila, O. M. Llobet, I. Medina, M. S. Puy, M. Alberich, S. de la Torre, A. Navarro, J. Paschall, and P. Flicek. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet*, 47(7):692–695, Jul 2015.

J. K. Lee, Y. L. Choi, M. Kwon, and P. J. Park. Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies. *Annu Rev Pathol*, 11:283–312, May 2016.

S. Lee and H. Lim. Review of statistical methods for survival analysis using genomic data. *Genomics Inform*, 17(4):e41, Dec 2019.

K. Li, H. Luo, L. Huang, H. Luo, and X. Zhu. Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int*, 20:16, 2020.

S. Libson and M. Lippman. A review of clinical aspects of breast cancer. *Int Rev Psychiatry*, 26(1):4–15, Feb 2014.

N. Lilovski. *dashboardthemes: Customise the Appearance of 'shinydashboard' Applications using Themes*, 2022. URL `https://CRAN.R-project.org/package=dashboardthemes`. R package version 1.1.6.

F. Luo. A systematic evaluation of copy number alterations detection methods on real SNP array and deep sequencing data. *BMC Bioinformatics*, 20(Suppl 25): 692, Dec 2019.

X. J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. T. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, and D. C. Sgroi. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, 5(6):607–616, Jun 2004.

X. J. Ma, R. Salunga, S. Dahiya, W. Wang, E. Carney, V. Durbecq, A. Harris, P. Goss, C. Sotiriou, M. Erlander, and D. Sgroi. A five-gene molecular grade index and HOXB13:IL17BR are complementary prognostic factors in early stage breast cancer. *Clin Cancer Res*, 14(9):2601–2608, May 2008.

A. Mayakonda, D. Lin, Y. Assenov, C. Plass, and P. H. Koeffler. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*, 2018. doi: http://dx.doi.org/10.1101/gr.239244.118.

M. Mayrhofer, S. DiLorenzo, and A. Isaksson. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biol*, 14(3):R24, Mar 2013.

D. J. McCarthy and G. K. Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, Mar 2009.

C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, 12(4): R41, 2011.

F. Meyer and V. Perrier. *shinylogs: Record Everything that Happens in a 'Shiny' Application*, 2022. URL `https://CRAN.R-project.org/package=shinylogs`. R package version 0.2.1.

S. Milborrow. *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*, 2022. URL `https://CRAN.R-project.org/package=rpart.plot`. R package version 3.1.1.

D. Moore. *Applied Survival Analysis Using R*. Springer, 2016.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *J Am Stat Assoc*, 58(302):415–434, 1963. doi: 10.1080/01621459.1963.10500855.

M. Morgan and M. Ramos. *BiocManager: Access the Bioconductor Project Package Repository*, 2023. URL `https://CRAN.R-project.org/package=BiocManager`. R package version 1.30.21.

S. Morganella, L. B. Alexandrov, D. Glodzik, X. Zou, H. Davies, J. Staaf, A. M. Sieuwerts, A. B. Brinkman, S. Martin, M. Ramakrishna, A. Butler, H. Y. Kim, Å. Borg, C. Sotiriou, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, H. G. Stunnenberg, M. J. van de Vijver, J. W. Martens, A. L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, J. Sale, C. Rada, M. R. Stratton, E. Birney, and S. Nik-Zainal. The topography of mutational processes in breast cancer genomes. *Nat Commun*, 7:11383, May 2016.

J. M. Mulligan, L. A. Hill, S. Deharo, G. Irwin, D. Boyle, K. E. Keating, O. Y. Raji, F. A. McDyer, E. O'Brien, M. Bylesjo, J. E. Quinn, N. M. Lindor, P. B. Mullan, C. R. James, S. M. Walker, P. Kerr, J. James, T. S. Davison, V. Proutski, M. Salto-Tellez, P. G. Johnston, F. J. Couch, D. Paul Harkin, and R. D. Kennedy. Identification and validation of an anthracycline/cyclophosphamide-based chemotherapy response assay in breast cancer. *J Natl Cancer Inst*, 106(1): djt335, Jan 2014.

E. Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2022. URL `https://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-3.

A. Nicolini, P. Ferrari, and M. J. Duffy. Prognostic and predictive biomarkers in breast cancer: Past, present and future. *Semin Cancer Biol*, 52(Pt 1):56–73, Oct 2018.

S. Ochoa and E. Hernández-Lemus. Molecular mechanisms of multi-omic regulation in breast cancer. *Front Oncol*, 13:1148861, 2023.

A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, Oct 2004.

A. B. Olshen, H. Bengtsson, P. Neuvial, P. T. Spellman, R. A. Olshen, and V. E. Seshan. Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics*, 27(15):2038–2046, Aug 2011.

S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, 351(27):2817–2826, Dec 2004.

J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*, 27(8):1160–1167, Mar 2009.

M. Pastore, P. Alaimo Di Loro, M. Mingione, and A. Calcagni. *overlapping: Estimation of Overlapping in Empirical Distributions*, 2022. URL `https://CRAN.R-project.org/package=overlapping`. R package version 2.1.

V. Patil and S. Dessai. Testing and interpreting assumptions of cox regression analysis. *Cancer Res Stat Treat*, 2(1):108, 2019. doi: 10.4103/crst.crst_40_19.

B. Pereira, S. F. Chin, O. M. Rueda, H. K. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S. J. Sammut, D. W. Tsui, B. Liu, S. J. Dawson, J. Abraham, H. Northen, J. F. Peden, A. Mukherjee, G. Turashvili, A. R. Green, S. McKinney, A. Oloumi, S. Shah, N. Rosenfeld, L. Murphy, D. R. Bentley, I. O. Ellis, A. Purushotham, S. E. Pinder, A. L. Børresen, H. M. Earl, P. D. Pharoah, M. T. Ross, S. Aparicio, and C. Caldas. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*, 7:11479, May 2016.

C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000.

L. Perreard, C. Fan, J. F. Quackenbush, M. Mullins, N. P. Gauthier, E. Nelson, M. Mone, H. Hansen, S. S. Buys, K. Rasmussen, A. R. Orrico, D. Dreher, R. Walters, J. Parker, Z. Hu, X. He, J. P. Palazzo, O. I. Olopade, A. Szabo, C. M. Perou, and P. S. Bernard. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res*, 8(2):R23, 2006.

V. Perrier, F. Meyer, and D. Granjon. *shinyWidgets: Custom Inputs Widgets for Shiny*, 2023. URL `https://CRAN.R-project.org/package=shinyWidgets`. R package version 0.7.6.

A. V. Pladsen, G. Nilsen, O. M. Rueda, M. R. Aure, Ø. Borgan, K. Liestøl, V. Vitelli, A. Frigessi, A. Langerød, A. Mathelier, OSBREAC, O. Engebråten, V. Kristensen, D. C. Wedge, P. Van Loo, C. Caldas, A. L. Børresen-Dale, H. G. Russnes, and O. C. Lingjærde. DNA copy number motifs are strong and independent predictors of survival in breast cancer. *Commun Biol*, 3(1):153, Apr 2020.

J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lønning, R. Tibshirani, D. Botstein, A. L. Børresen-Dale, and P. O. Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968, Oct 2002.

A. Prat, J. S. Parker, O. Karginova, C. Fan, C. Livasy, J. I. Herschkowitz, X. He, and C. M. Perou. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*, 12(5):R68, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

E. A. Rakha, G. M. Tse, and C. M. Quinn. An update on the pathological classification of breast cancer. *Histopathology*, 82(1):5–16, Jan 2023.

M. Ramos, L. Geistlinger, S. Oh, L. Schiffer, R. Azhar, H. Kodali, I. de Bruijn, J. Gao, V. J. Carey, M. Morgan, and L. Waldron. Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO Clin Cancer Inform*, 4:958–971, Oct 2020.

M. Rasmussen, M. Sundström, H. Göransson-Kultima, J. Botling, P. Micke, H. Birgisson, B. Glimelius, and A. Isaksson. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol*, 12(10):R108, Oct 2011.

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

H. G. Russnes, H. K. M. Vollan, O. C. Lingjærde, A. Krasnitz, P. Lundin, B. Naume, T. Sørlie, E. Borgen, I. H. Rye, A. Langerød, S. F. Chin, A. E. Teschendorff, P. J. Stephens, S. Månér, E. Schlichting, L. O. Baumbusch, R. Kåresen, M. P. Stratton, M. Wigler, C. Caldas, A. Zetterberg, J. Hicks, and A. L. Børresen-Dale.

Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci Transl Med*, 2(38):38ra47, Jun 2010.

H. G. Russnes, O. C. Lingjærde, A. L. Børresen-Dale, and C. Caldas. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am J Pathol*, 187(10):2152–2162, Oct 2017.

A. Sali and D. Attali. *shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating*, 2020. URL `https://CRAN.R-project.org/package=shinycssloaders`. R package version 1.0.0.

R. Shen and V. E. Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res*, 44 (16):e131, Sep 2016.

A. Shlien and D. Malkin. Copy number variations and cancer. *Genome Med*, 1(6): 62, Jun 2009.

A. Signorell. *DescTools: Tools for Descriptive Statistics*, 2023. URL `https://CRAN.R-project.org/package=DescTools`. R package version 0.99.49.

M. Smid, M. Hoes, A. M. Sieuwerts, S. Sleijfer, Y. Zhang, Y. Wang, J. A. Foekens, and J. W. Martens. Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes. *Breast Cancer Res Treat*, 128(1): 23–30, Jul 2011.

J. C. Smith and J. M. Sheltzer. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. *Elife*, 7, Dec 2018.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

L. Song, K. Bhuvaneshwar, Y. Wang, Y. Feng, I. M. Shih, S. Madhavan, and Y. Gusev. CINdex: A Bioconductor Package for Analysis of Chromosome Instability in DNA Copy Number Data. *Cancer Inform*, 16:1176935117746637, 2017.

L. Song, K. Bhuvaneshwar, Y. Wang, Y. Feng, I. M. Shih, S. Madhavan, and Y. Gusev. *CINdex: Chromosome Instability Index*, 2022. R package version 1.24.0.

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–10874, Sep 2001.

T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A. L. Børresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–8423, Jul 2003.

T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. nning, P. O. Brown, A. L. rresen Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–8423, Jul 2003.

C. D. Steele, A. Abbasi, S. M. A. Islam, A. L. Bowes, A. Khandekar, K. Haase, S. Hames-Fathi, D. Ajayi, A. Verfaillie, P. Dhami, A. McLatchie, M. Lechner, N. Light, A. Shlien, D. Malkin, A. Feber, P. Proszek, T. Lesluyes, F. Mertens, A. M. Flanagan, M. Tarabichi, P. Van Loo, L. B. Alexandrov, and N. Pillay. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, Jun 2022.

P. J. Stephens, D. J. McBride, M. L. Lin, I. Varela, E. D. Pleasance, J. T. Simpson, L. A. Stebbings, C. Leroy, S. Edkins, L. J. Mudie, C. D. Greenman, M. Jia, C. Latimer, J. W. Teague, K. W. Lau, J. Burton, M. A. Quail, H. Swerdlow, C. Churcher, R. Natrajan, A. M. Sieuwerts, J. W. Martens, D. P. Silver, A. d, H. E. Russnes, J. A. Foekens, J. S. Reis-Filho, L. van 't Veer, A. L. Richardson, A. L. Børresen-Dale, P. J. Campbell, P. A. Futreal, and M. R. Stratton. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010, Dec 2009.

K. H. Stopsack, C. A. Whittaker, T. A. Gerke, M. Loda, P. W. Kantoff, L. A. Mucci, and A. Amon. Aneuploidy drives lethal progression in prostate cancer. *Proc Natl Acad Sci U S A*, 116(23):11390–11395, Jun 2019.

B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, Feb 2007.

I. Subirana, H. Sanz, and J. Vila. Building bivariate tables: The compareGroups package for R. *J Stat Softw*, 57(12):1–16, 2014. URL `https://www.jstatsoft.org/v57/i12/`.

H. Sung, M. Garcia-Closas, J. Chang-Claude, F. M. Blows, H. R. Ali, J. Figueroa, H. Nevanlinna, R. Fagerholm, P. Heikkilä, C. Blomqvist, G. G. Giles, R. L. Milne, M. C. Southey, C. McLean, A. Mannermaa, V. M. Kosma, V. Kataja, R. Sironen, F. J. Couch, J. E. Olson, E. Hallberg, C. Olswold, A. Cox, S. S. Cross, P. Kraft, R. M. Tamimi, A. H. Eliassen, M. K. Schmidt, M. K. Bolla, Q. Wang, D. Easton, W. J. Howat, P. Coulson, P. D. Pharoah, M. E. Sherman, and X. R. Yang. Heterogeneity of luminal breast cancer characterised by immunohistochemical expression of basal markers. *Br J Cancer*, 114(3):298–304, Feb 2016.

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 Cancer J Clin*, 71(3): 209–249, May 2021.

Z. Tao, S. Wang, C. Wu, T. Wu, X. Zhao, W. Ning, G. Wang, J. Wang, J. Chen, K. Diao, F. Chen, and X. S. Liu. The repertoire of copy number alteration signatures in human cancer. *Brief Bioinform*, 24(2), Mar 2023.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct 2012.

T. Therneau and B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2022. URL `https://CRAN.R-project.org/package=rpart`. R package version 4.1.19.

T. M. Therneau. *A Package for Survival Analysis in R*, 2023. URL `https://CRAN.R-project.org/package=survival`. R package version 3.5-3.

T. M. Therneau, E. J. Atkinson, and Mayo Foundation. An introduction to recursive partitioning using the rpart routines. Technical report, Technical report Mayo Foundation, 2022.

S. Tian, P. Roepman, L. J. Van't Veer, R. Bernards, F. de Snoo, and A. M. Glas. Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. *Biomark Insights*, 5:129–138, Nov 2010.

I. Tishchenko, H. H. Milioli, C. Riveros, and P. Moscato. Extensive Transcriptomic and Genomic Analysis Provides New Insights about Luminal Breast Cancers. *PLoS One*, 11(6):e0158259, 2016.

L. A. Torre, F. Islami, R. L. Siegel, E. M. Ward, and A. Jemal. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev*, 26(4):444–457, Apr 2017.

V. Trevino, F. Falciani, and H. A. Barrera-Saldaña. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med*, 13(9-10):527–541, 2007.

P. Ulrich. *Cancer Genomics: Molecular classification, prognosis and response prediction*, volume 9789400758421. Springer Netherlands, Dordrecht, 1. aufl. edition, 2013. ISBN 9400758413.

A. Urbaniak, K. Jablonska, M. Podhorska-Okolow, M. Ugorski, and P. Dziegiel. Prolactin-induced protein (PIP)-characterization and role in breast cancer progression. *Am J Cancer Res*, 8(11):2150–2164, 2018.

M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, Dec 2002.

E. van den Broek and S. van Lieshout. *GeneBreak: Gene Break Detection*, 2023. URL `https://bioconductor.org/packages/GeneBreak`. R package version 1.31.0.

P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A. L. Børresen-Dale, and V. N. Kristensen. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*, 107(39):16910–16915, Sep 2010.

L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871): 530–536, Jan 2002.

E. S. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, Mar 2007.

E. S. Venkatraman and A. B. Olshen. *DNAcopy: DNA Copy Number Data Analysis*, 2023. URL `https://bioconductor.org/packages/DNAcopy`. R package version 1.75.5.

H. K. Vollan, O. M. Rueda, S. F. Chin, C. Curtis, G. Turashvili, S. Shah, O. C. Lingjærde, Y. Yuan, C. K. Ng, M. J. Dunning, E. Dicks, E. Provenzano, S. Sammut, S. McKinney, I. O. Ellis, S. Pinder, A. Purushotham, L. C. Murphy, V. N. Kristensen, METABRIC Group, J. D. Brenton, P. D. Pharoah, A. L. Børresen-Dale, S. Aparicio, and C. Caldas. A tumor DNA complex aberration index is an independent predictor of survival in breast and ovarian cancer. *Mol Oncol*, 9(1): 115–127, Jan 2015.

K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*, 17(11):1665–1674, Nov 2007.

S. Wang and D. Lee. Identifying prognostic subgroups of luminal-A breast cancer using deep autoencoders and gene expressions. *PLoS Comput Biol*, 19(5):e1011197, May 2023.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

S. Werner, H. Stamm, M. Pandjaitan, D. Kemming, B. Brors, K. Pantel, and H. Wikman. Iroquois homeobox 2 suppresses cellular motility and chemokine expression in breast cancer cells. *BMC Cancer*, 15:896, Nov 2015.

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL `http://www.jstatsoft.org/v21/i12/`.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D Robinson, D. P. Seidel, V. Spinu, K. Takahashi,

D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *J Open Source Softw*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schütz, D. R. Goldstein, M. Piccart, and M. Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*, 10(4):R65, 2008.

Y. Xie, J. Cheng, and X. Tan. *DT: A Wrapper of the JavaScript Library 'DataTables'*, 2023. URL `https://CRAN.R-project.org/package=DT`. R package version 0.28.

Z. Zeng, Y. Gao, J. Li, G. Zhang, S. Sun, Q. Wu, Y. Gong, and C. Xie. Violations of proportional hazard assumption in Cox regression model of transcriptomic data in TCGA pan-cancer cohorts. *Comput Struct Biotechnol J*, 20:496–507, 2022.

L. Zhang, N. Feizi, C. Chi, and P. Hu. Association Analysis of Somatic Copy Number Alteration Burden With Breast Cancer Survival. *Front Genet*, 9:421, 2018.

W. Zhang, J. H. Mao, W. Zhu, A. K. Jain, K. Liu, J. B. Brown, and G. H. Karpen. Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat Commun*, 7:12619, Aug 2016.

Appendix A

Appendix A contains a list of the 23 clinical variables considered at the beginning of Chapter 3 and the rpart survival trees including selected clinical variables and Absolute CNA Scores/Quartiles as candidate predictors.

List of the 23 clinical variables mentioned in section 3.2: ER_IHC, HER2_SNP6, HORMONE_THERAPY, INFERRED_MENOPAUSAL_STATE, INTCLUST, CLAUDIN_SUBTYPE, THREEGENE, RADIO_THERAPY, HISTOLOGICAL_SUBTYPE, BREAST_SURGERY, CANCER_TYPE_DETAILED, HER2_STATUS, GRADE, PR_STATUS, LYMPH_NODES_EXAMINED_POSITIVE, NPI, AGE_AT_DIAGNOSIS, TUMOR_SIZE, TUMOR_STAGE, CELLULARITY, LATERALITY, ER_STATUS, CHEMOTHERAPY.

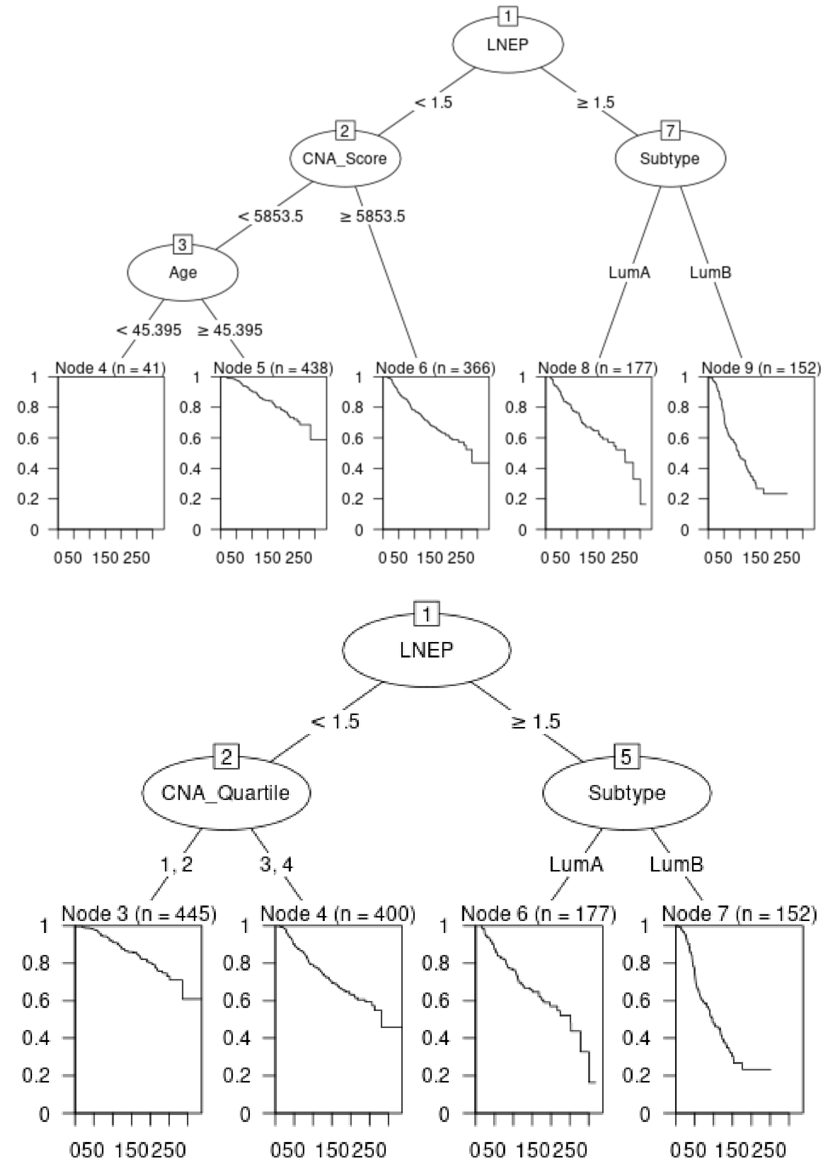


Figure A1: Recursive partitioning survival trees, fitted using the rpart algorithm, for disease-specific survival using clinical variables and CNA Score and CNA Quartile as candidate predictors.

Appendix B

Appendix B contains the rpart and ctree survival trees for OS outcomes. These survival trees are fitted under a number of scenarios (1) including PAM50 and IntClust only, (2) including Global CNA metrics and PAM50 or IntClust, (3) including Global CNA metrics, PAM50 subtype or IntClust, and a selection of clinical variables, (4) including Chromosome arm CNA metrics and PAM50 or IntClust, and (5) including Chromosome arm CNA metrics, PAM50 subtype or IntClust, and a selection of clinical variables.

(A) **(B)**

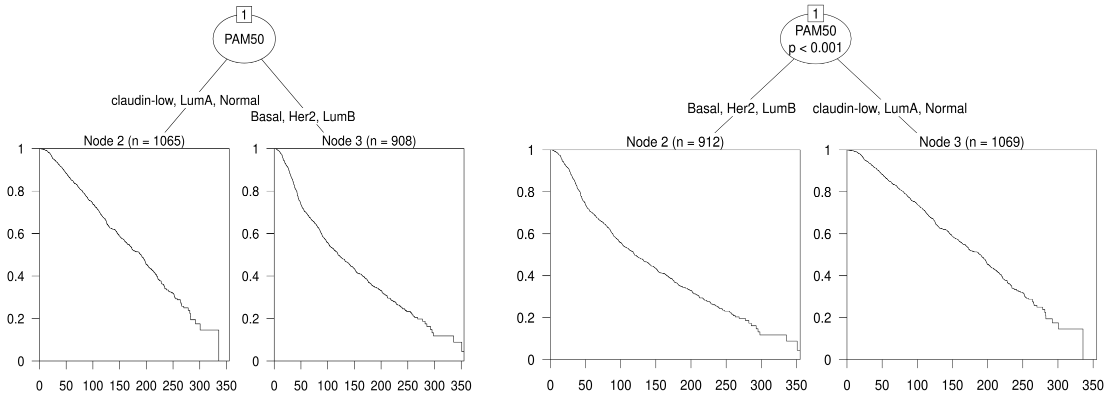


Figure B1: Recursive partitioning survival trees for overall survival using PAM50 Subtype as a candidate predictor. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A) **(B)**

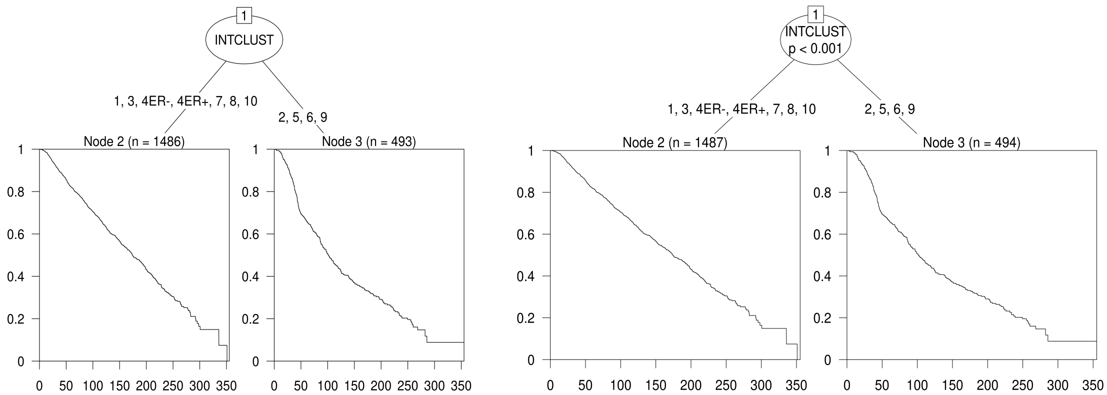
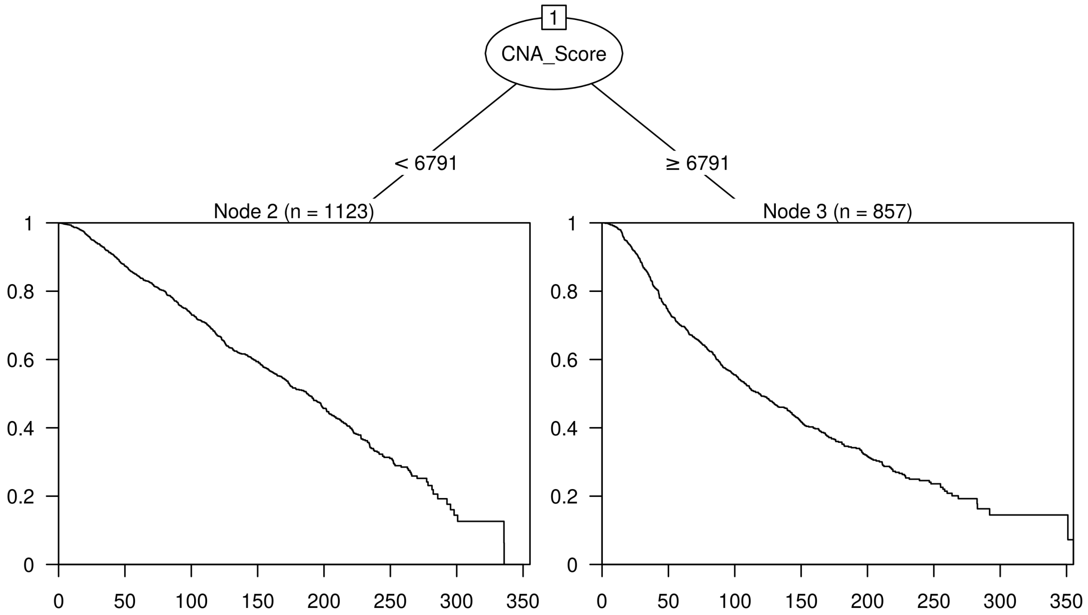


Figure B2: Recursive partitioning survival trees for overall survival using Integrative Cluster as a candidate predictor. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

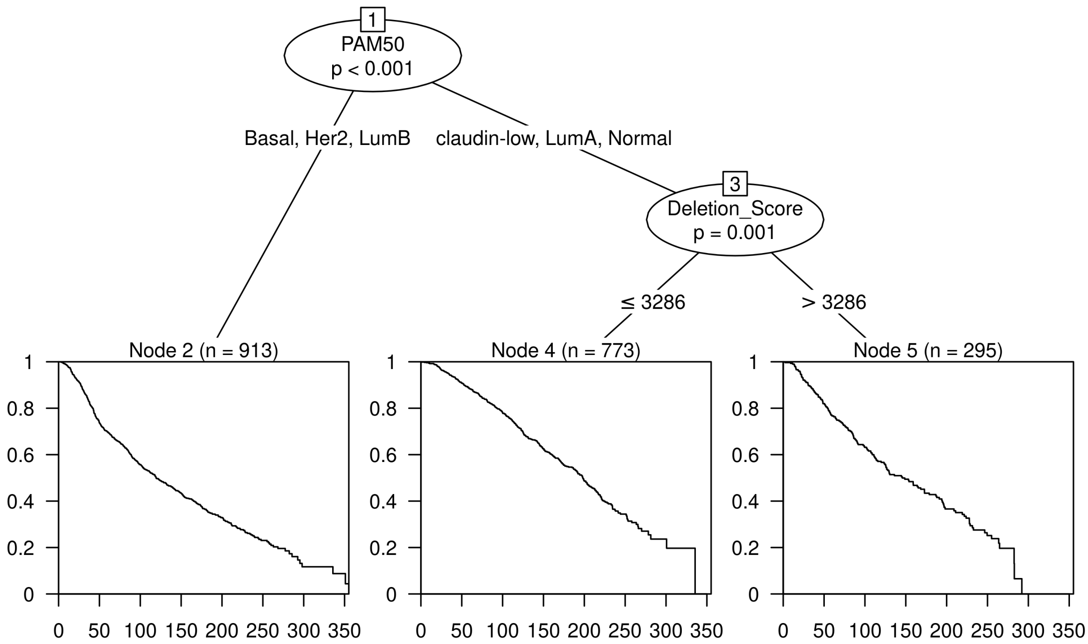
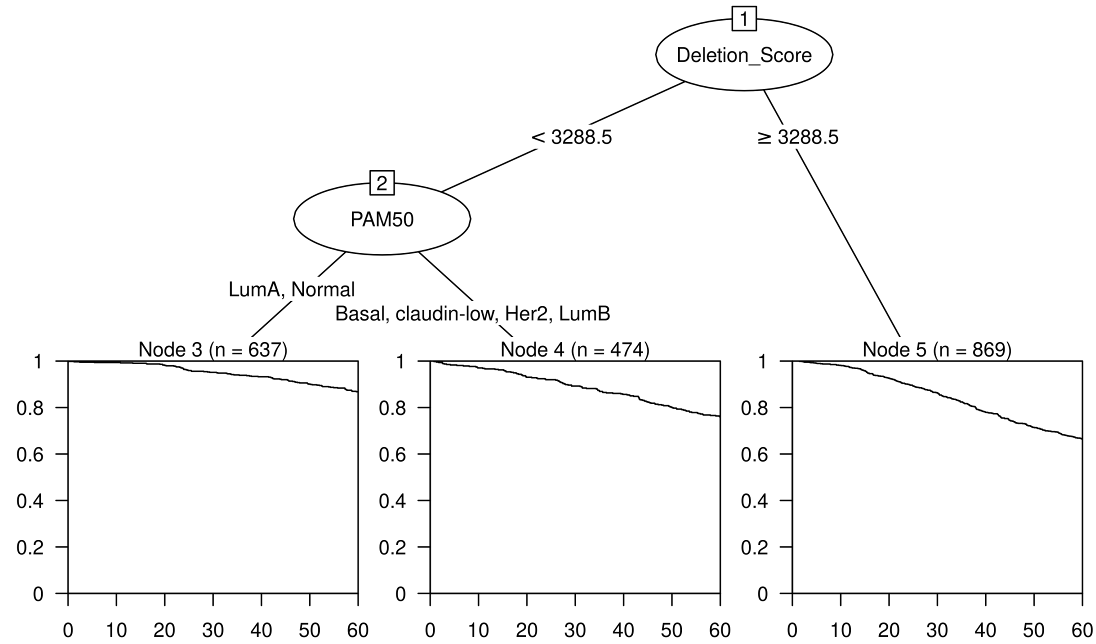


Figure B3: Recursive partitioning survival trees for overall survival using PAM50 and the 6 CNA Score metrics as candidate predictors. Trees fitted using the rpart algorithm are displayed on the top and trees fitted using the ctree algorithm are displayed on the bottom.

(A)



(B)

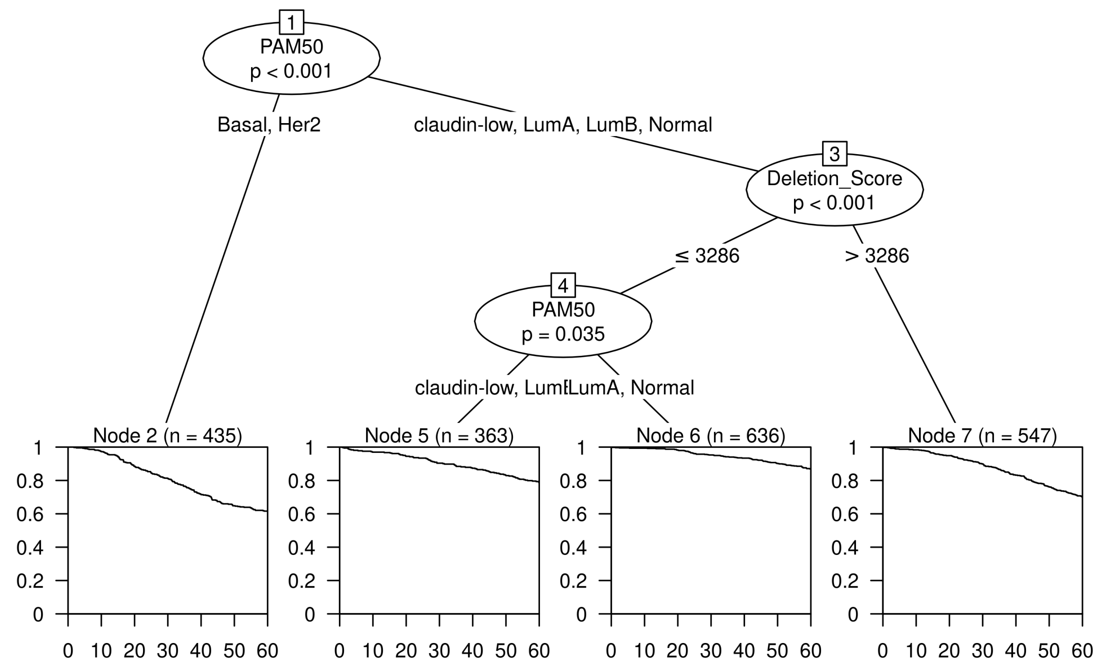
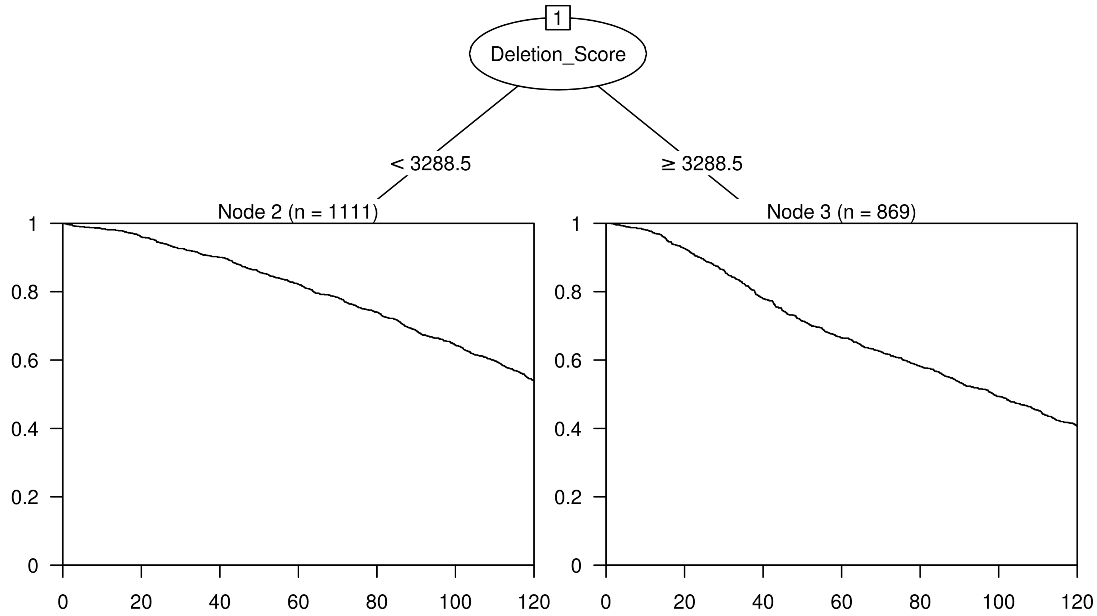


Figure B4: Recursive partitioning survival trees for five-year overall survival using PAM50 and the 6 CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

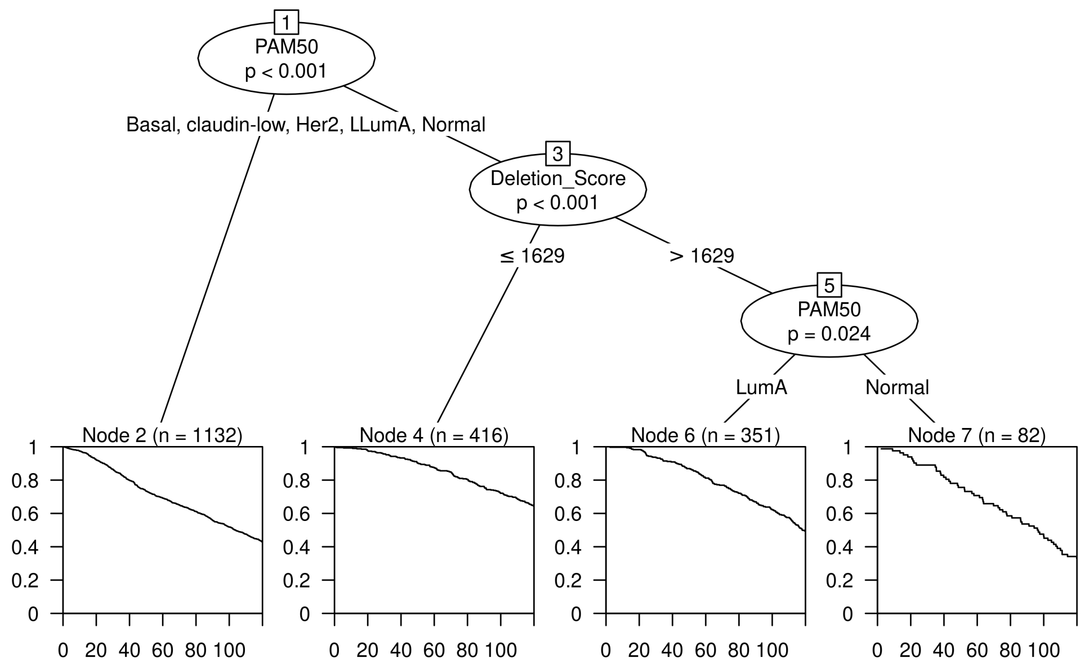
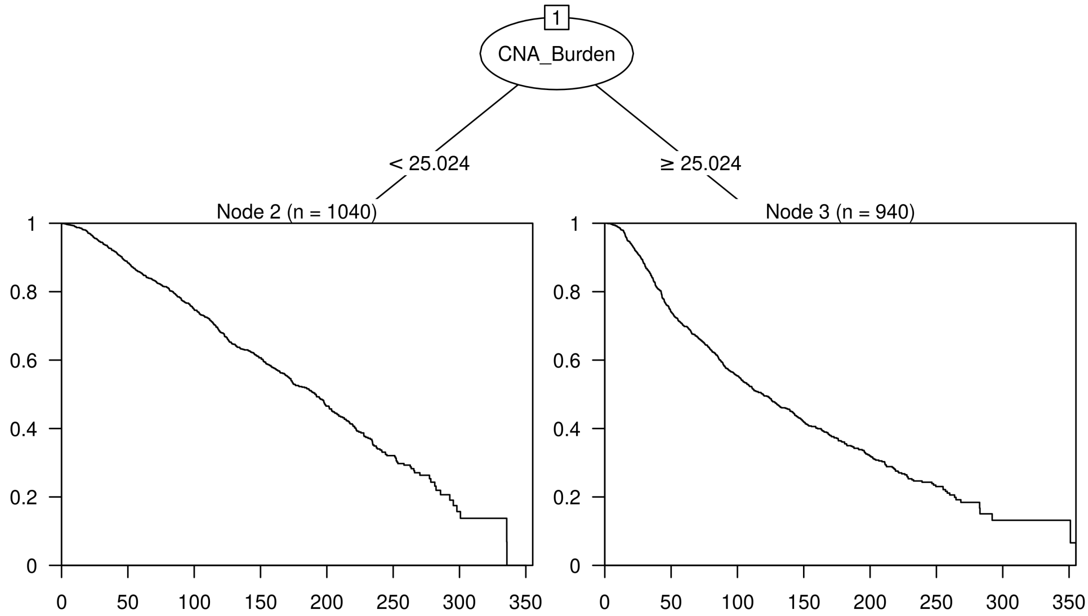


Figure B5: Recursive partitioning survival trees for ten-year overall survival using PAM50 and the 6 CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

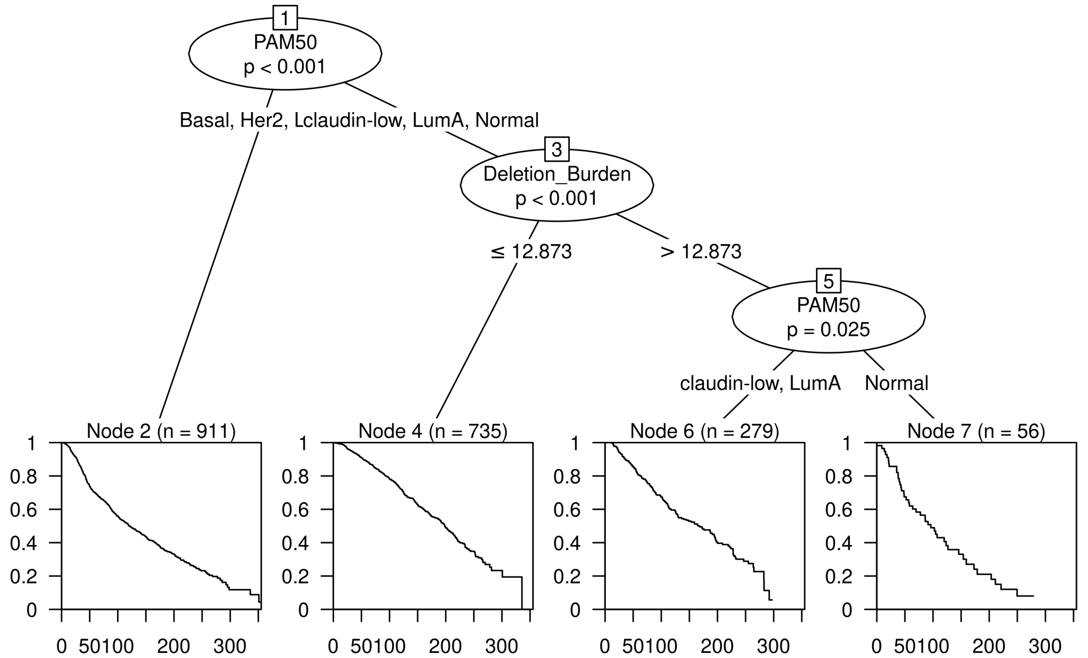
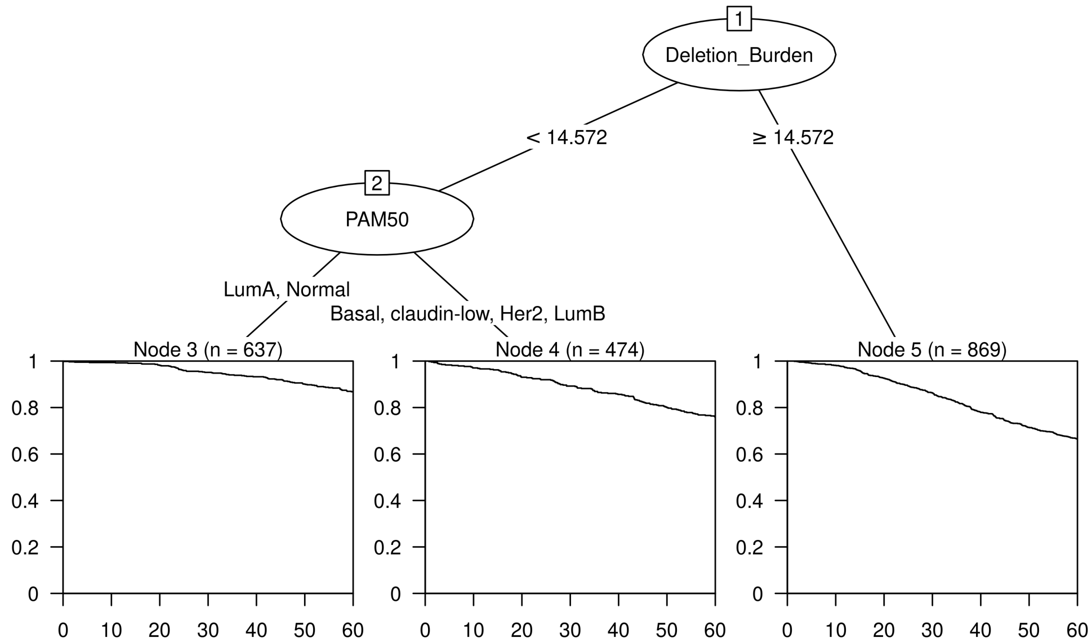


Figure B6: Recursive partitioning survival trees for overall survival using PAM50 and the 6 CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

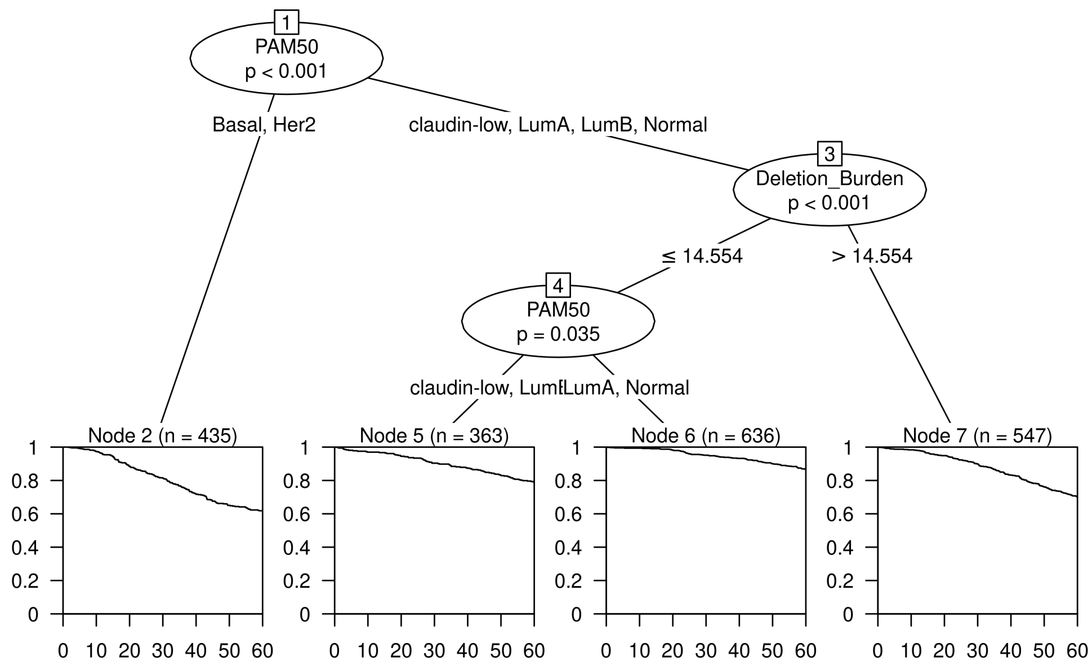
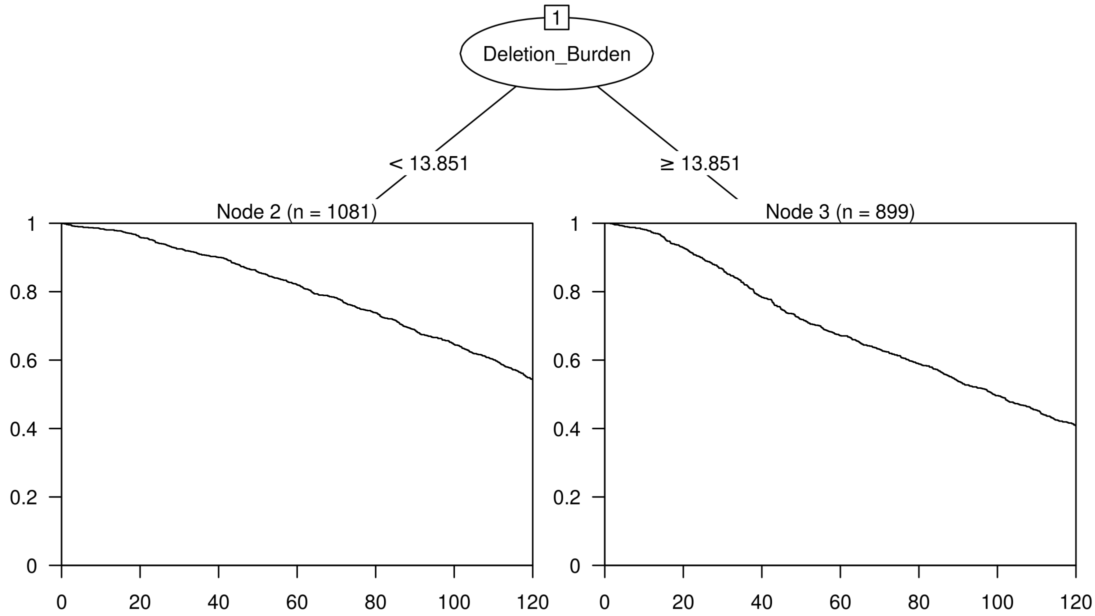


Figure B7: Recursive partitioning survival trees for five-year overall survival using PAM50 and the 6 CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

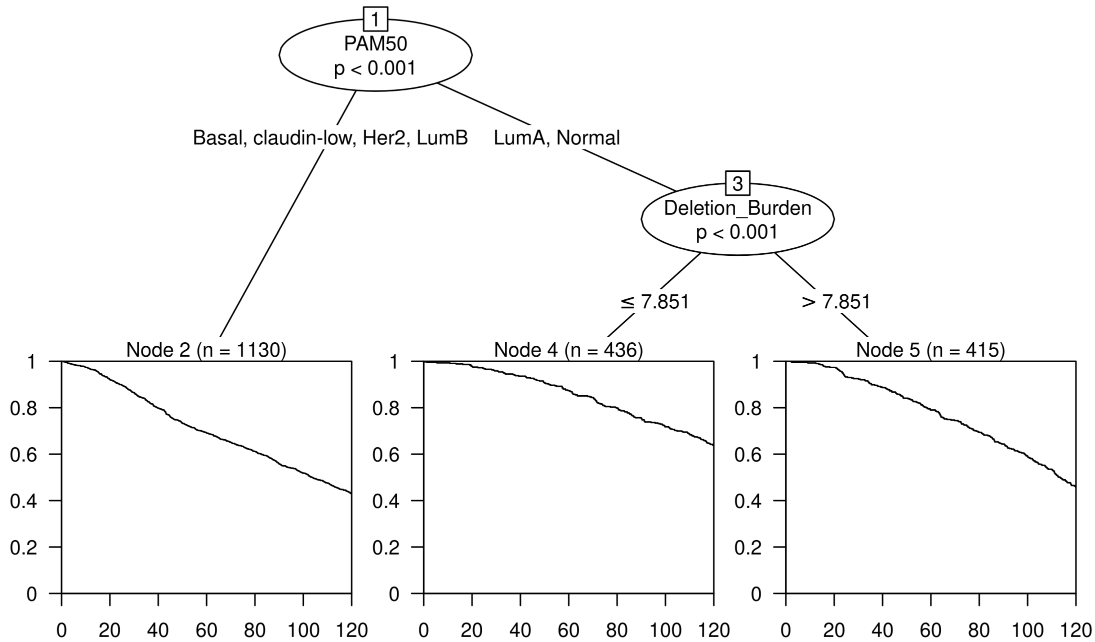
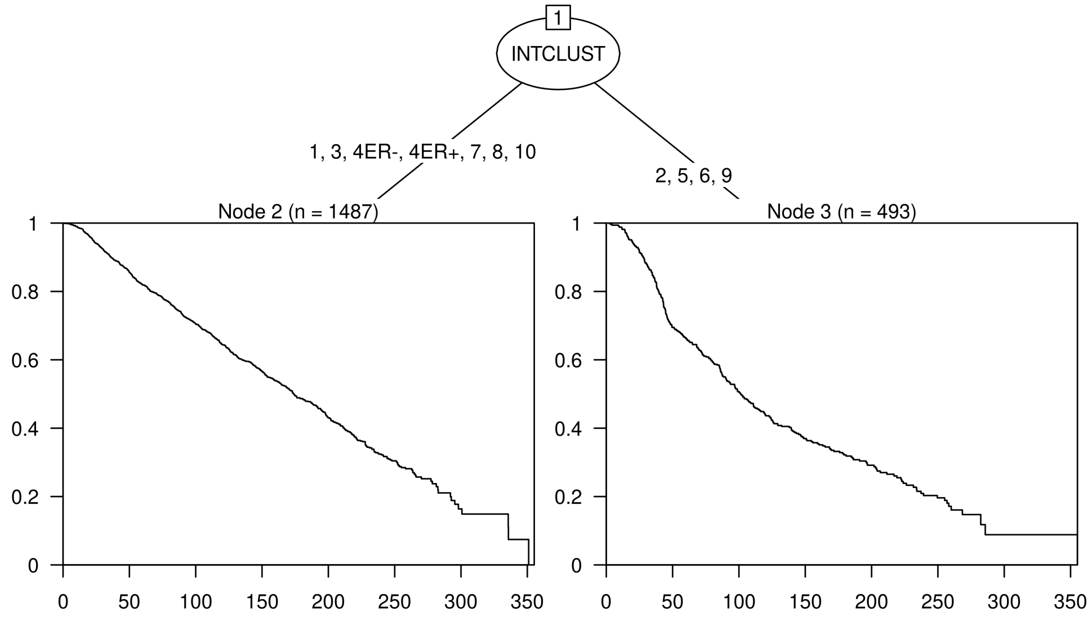


Figure B8: Recursive partitioning survival trees for ten-year overall survival using PAM50 and the 6 CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

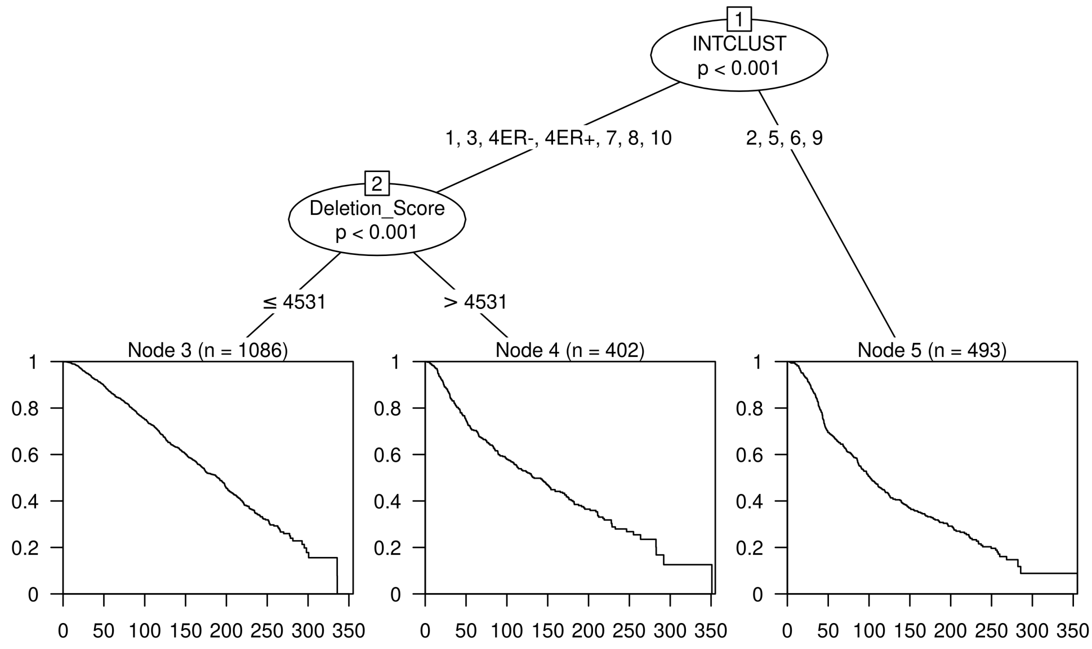
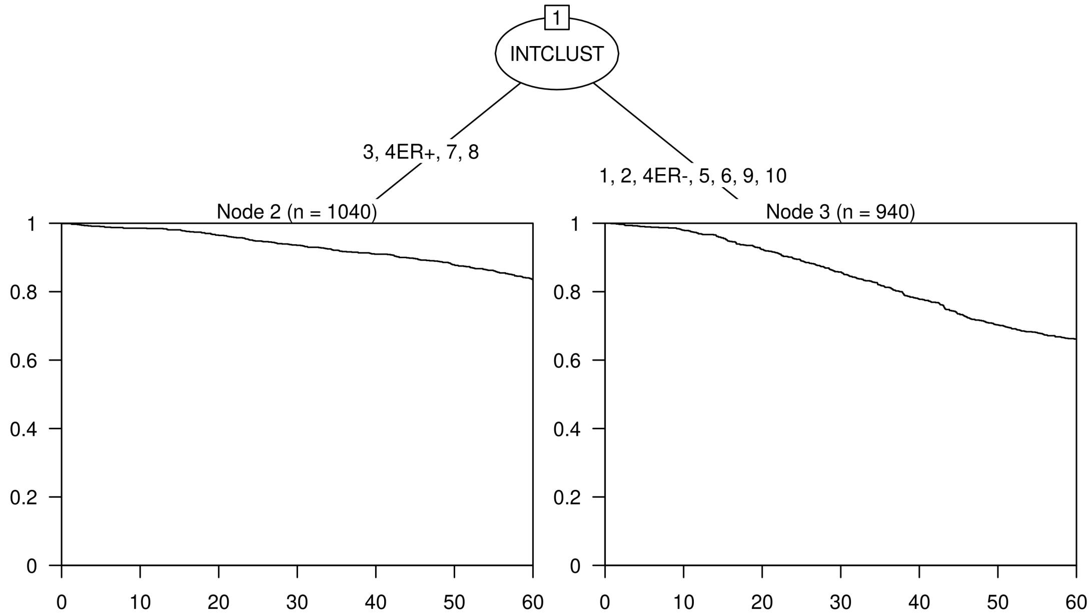


Figure B9: Recursive partitioning survival trees for overall survival using IntClust and the 6 CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

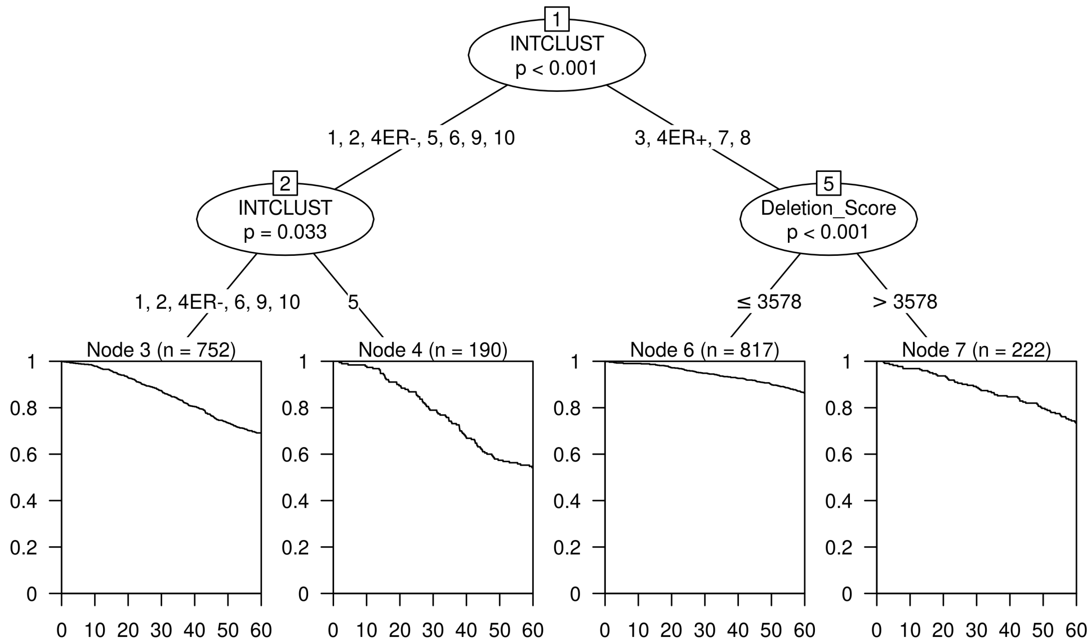
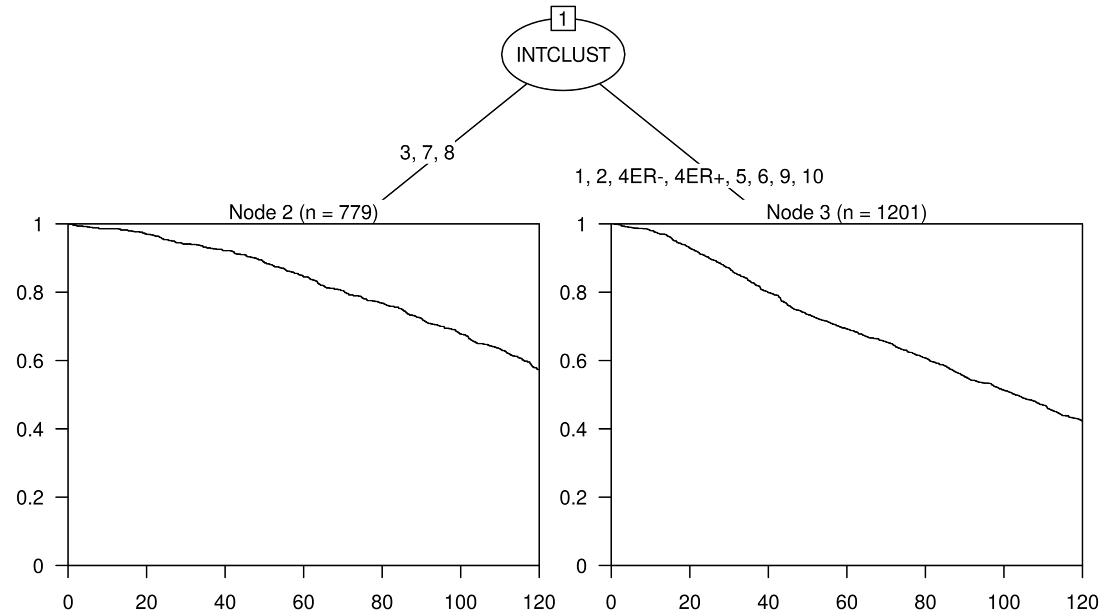


Figure B10: Recursive partitioning survival trees for five-year overall survival using IntClust and the 6 CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

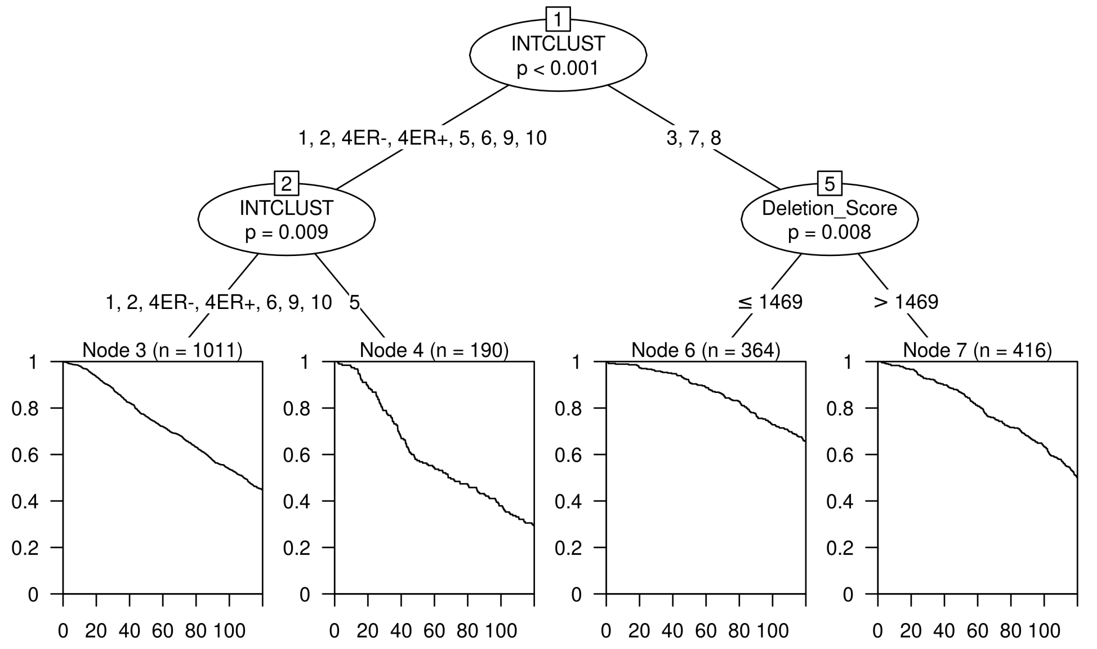
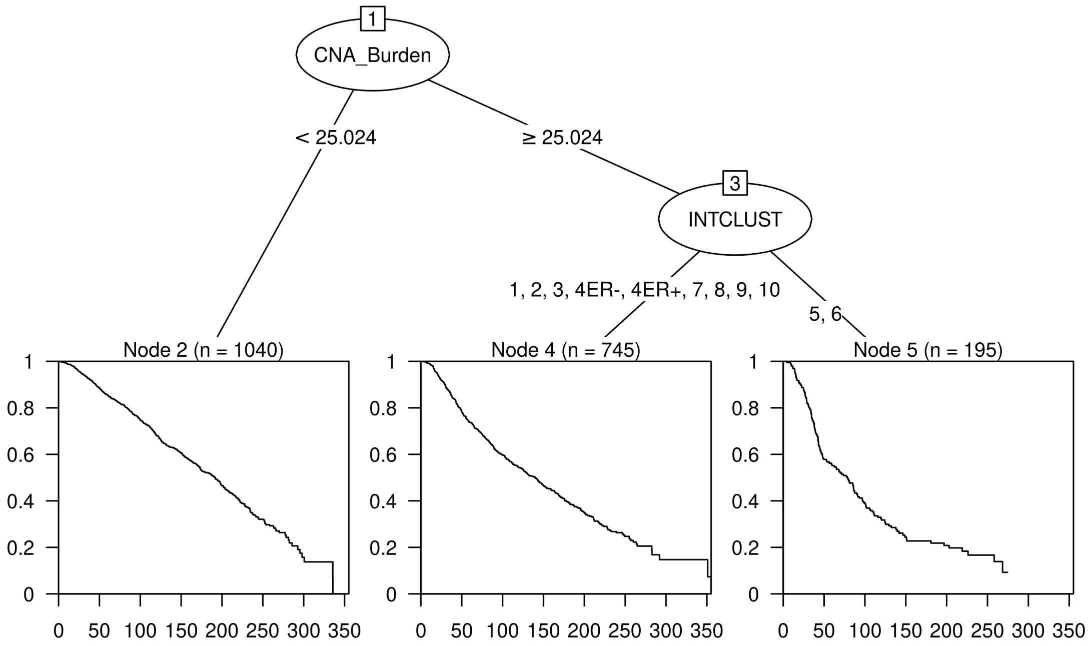


Figure B11: Recursive partitioning survival trees for ten-year overall survival using IntClust and the 6 CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

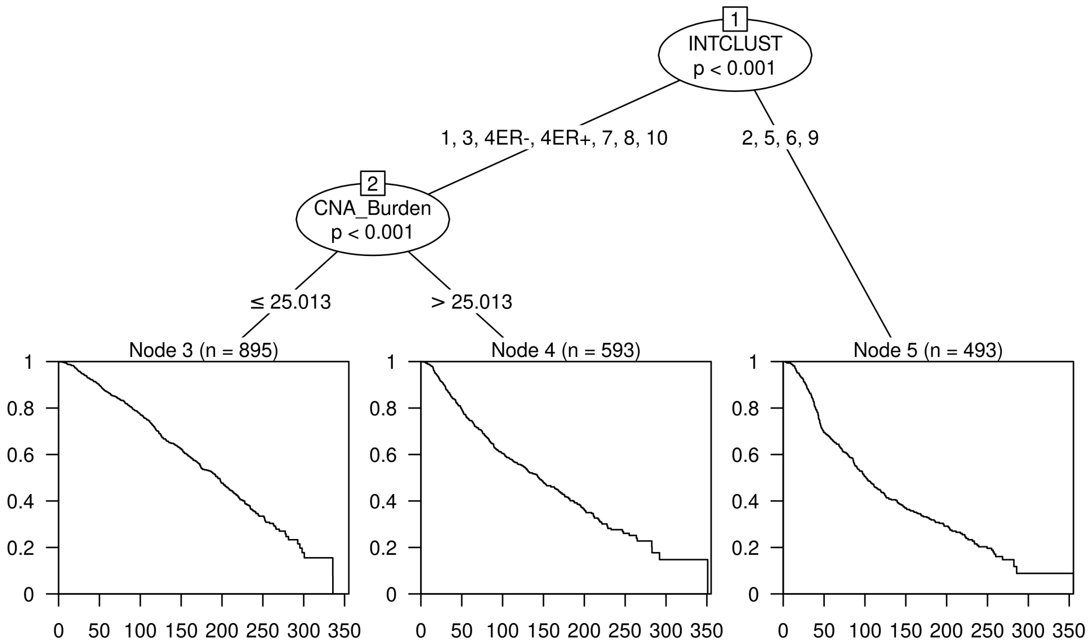
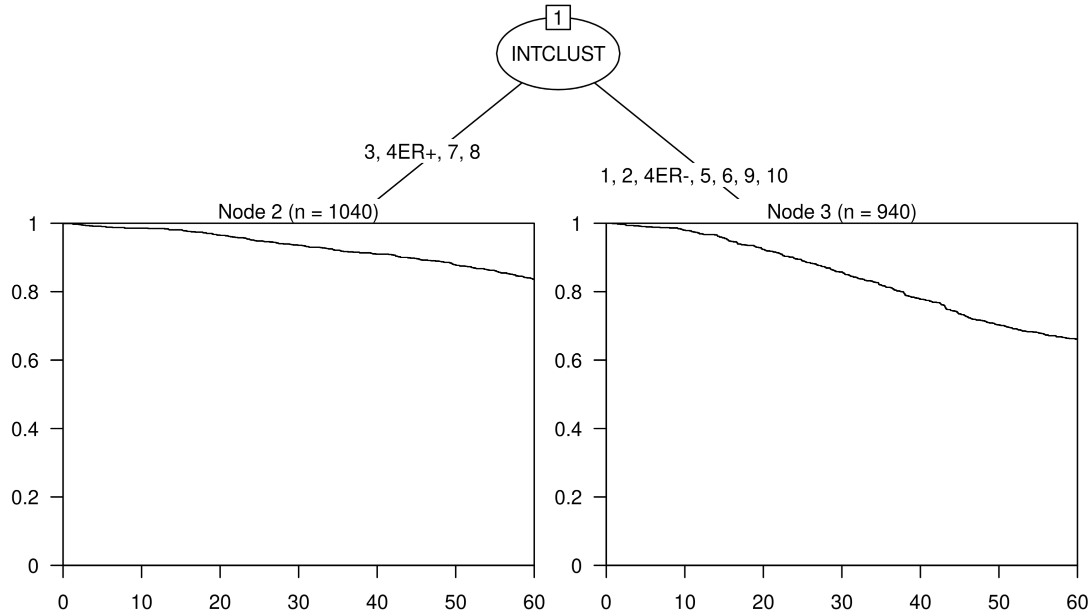


Figure B12: Recursive partitioning survival trees for overall survival using IntClust and the 6 CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

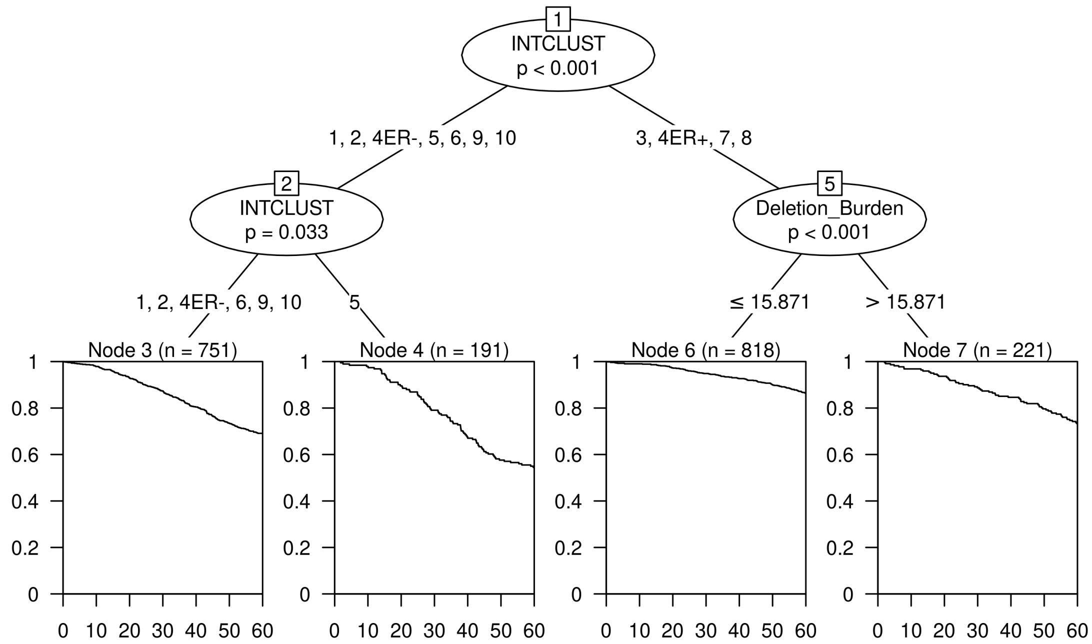
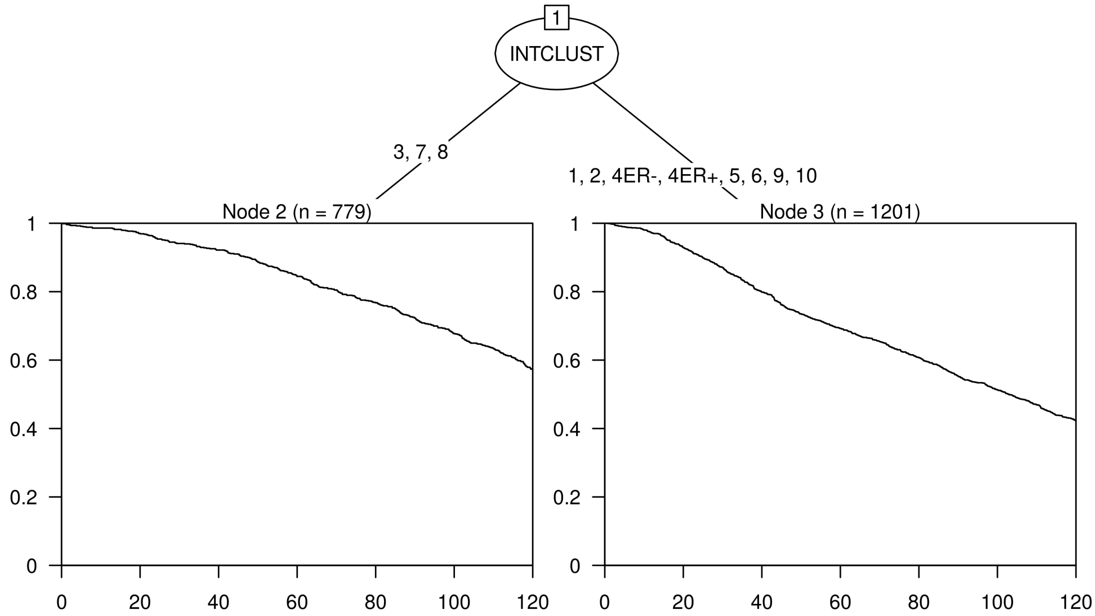


Figure B13: Recursive partitioning survival trees for five-year overall survival using IntClust and the 6 CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

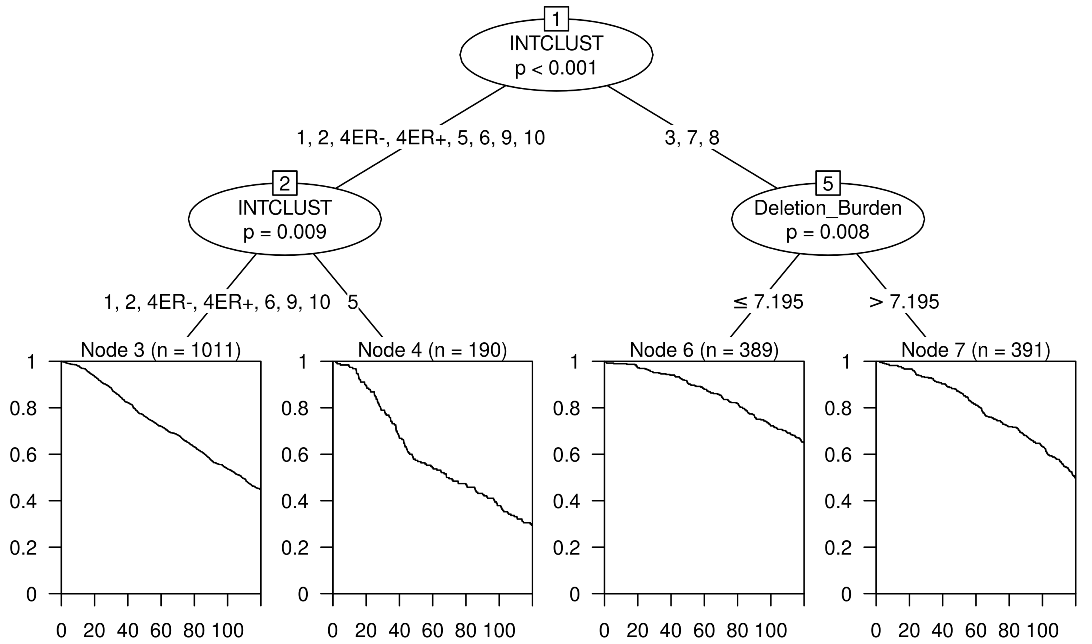
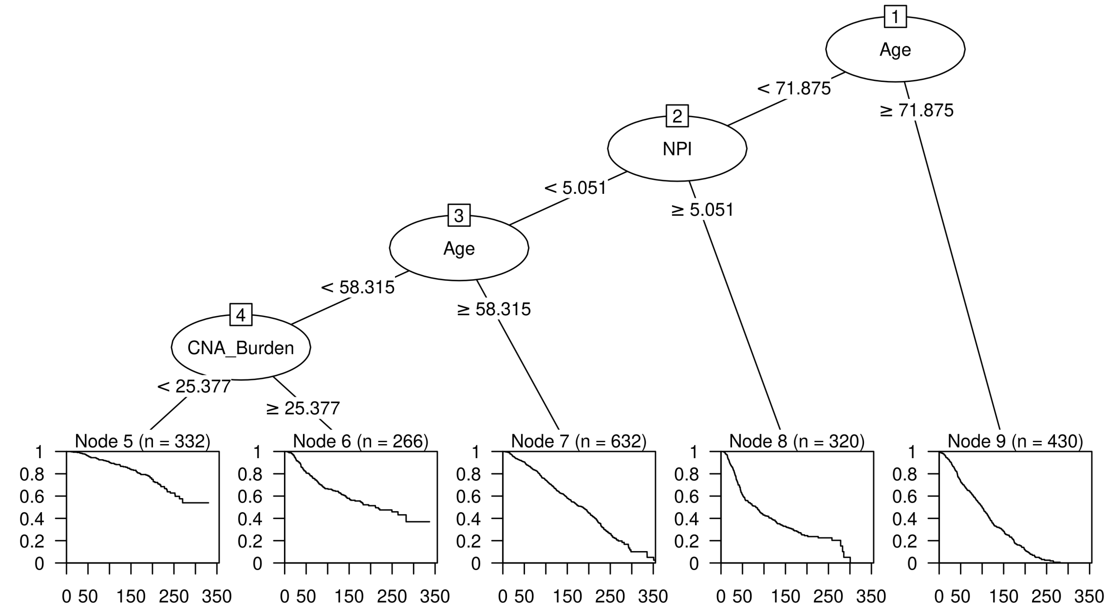


Figure B14: Recursive partitioning survival trees for ten-year overall survival using IntClust and the 6 CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

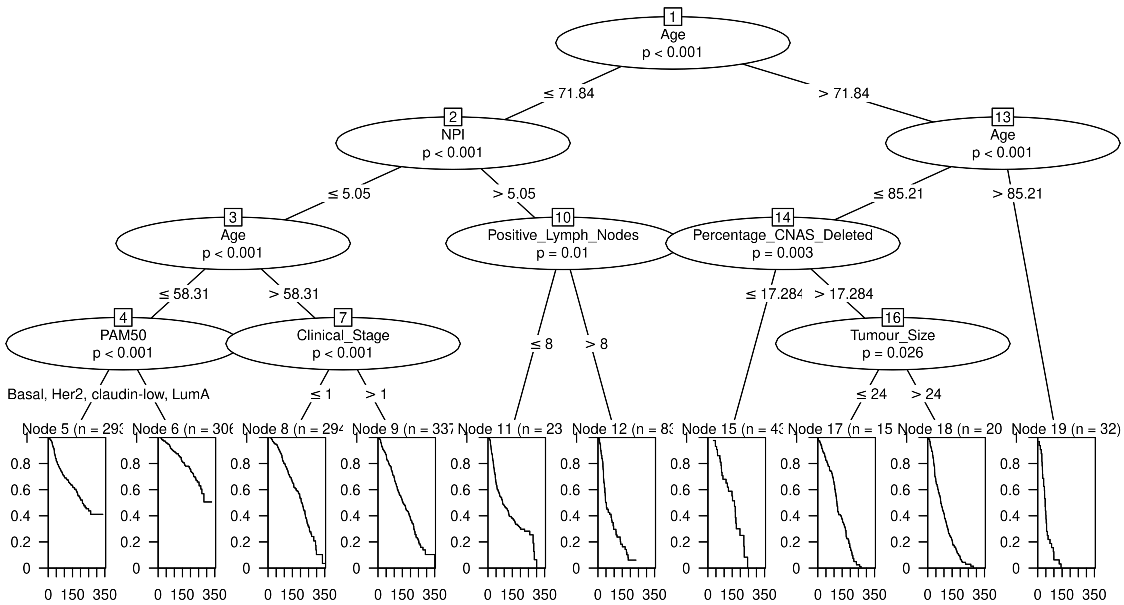
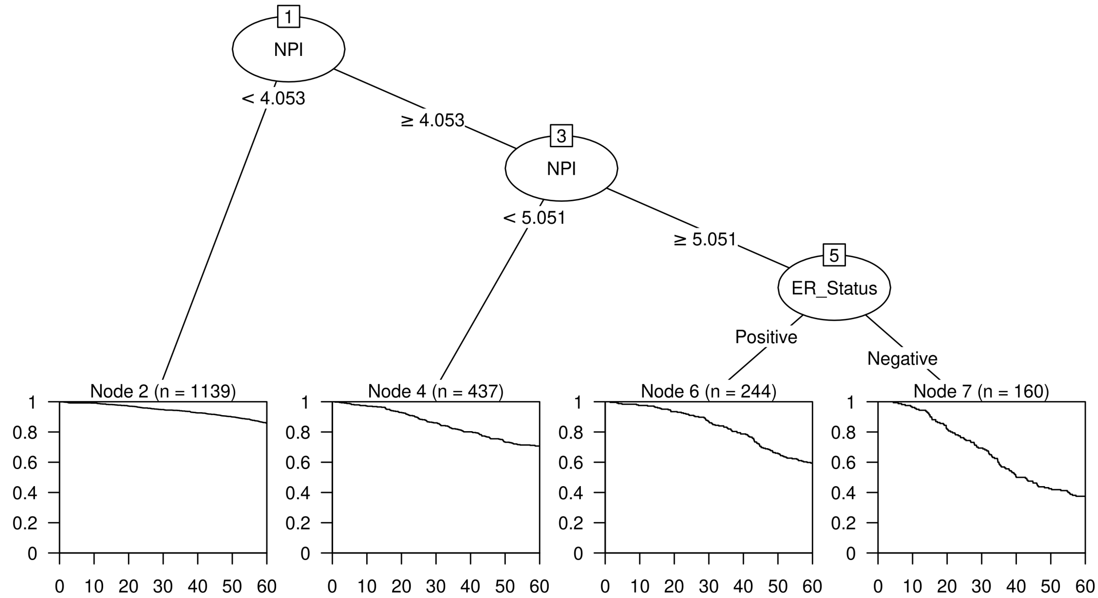


Figure B15: Recursive partitioning survival trees for overall survival using PAM50, the 6 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

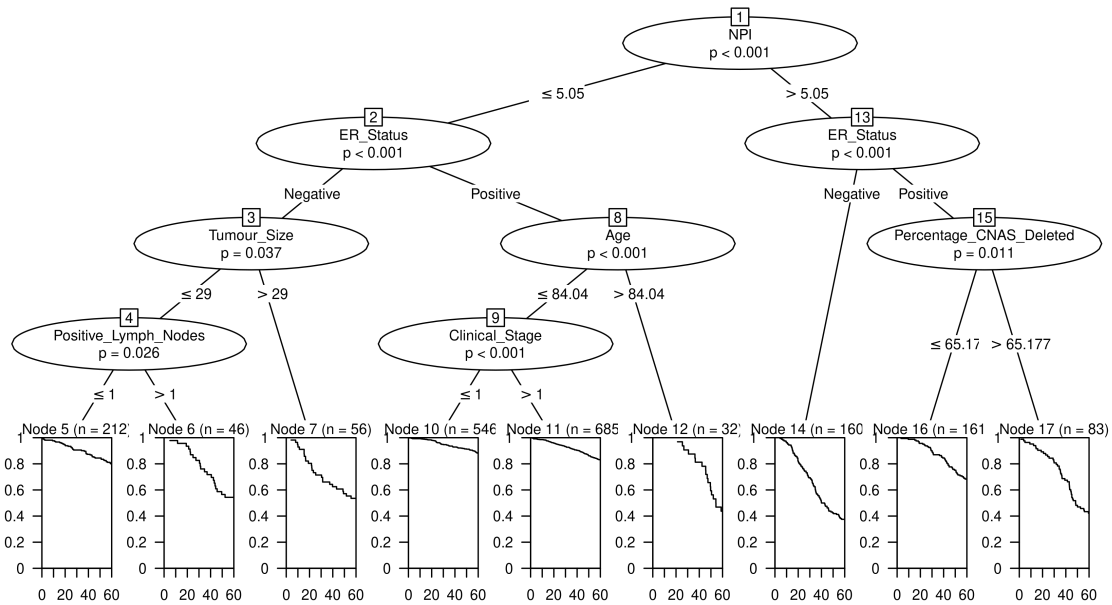
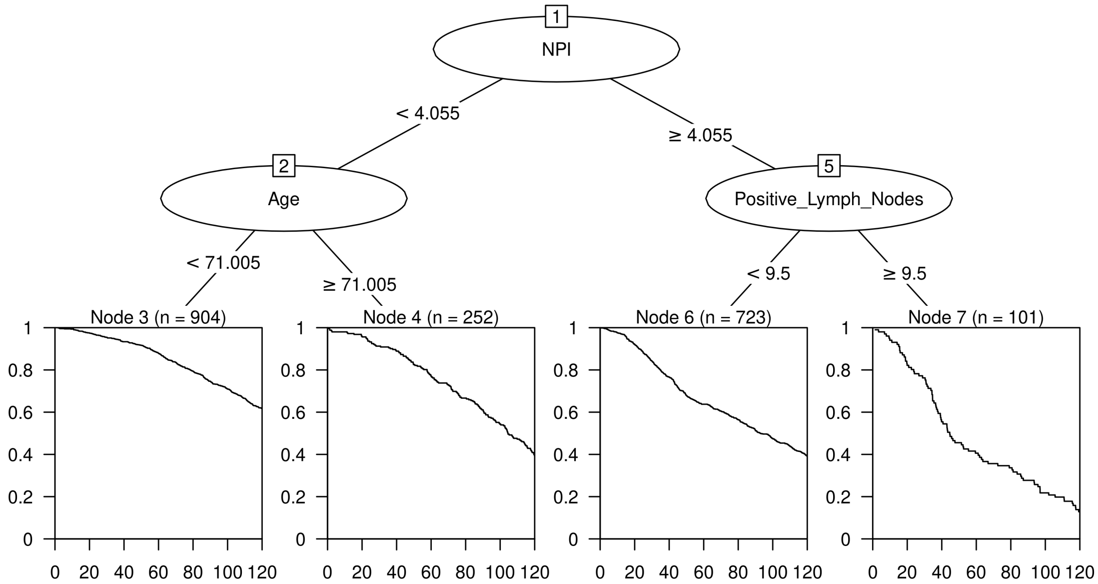


Figure B16: Recursive partitioning survival trees for five-year overall survival using PAM50, the 6 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

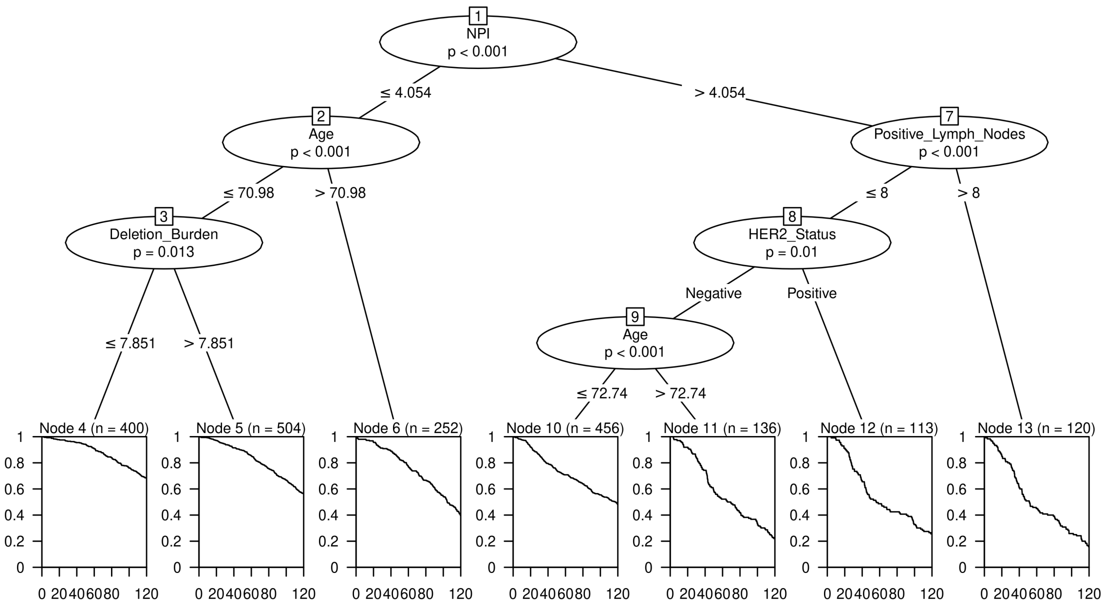
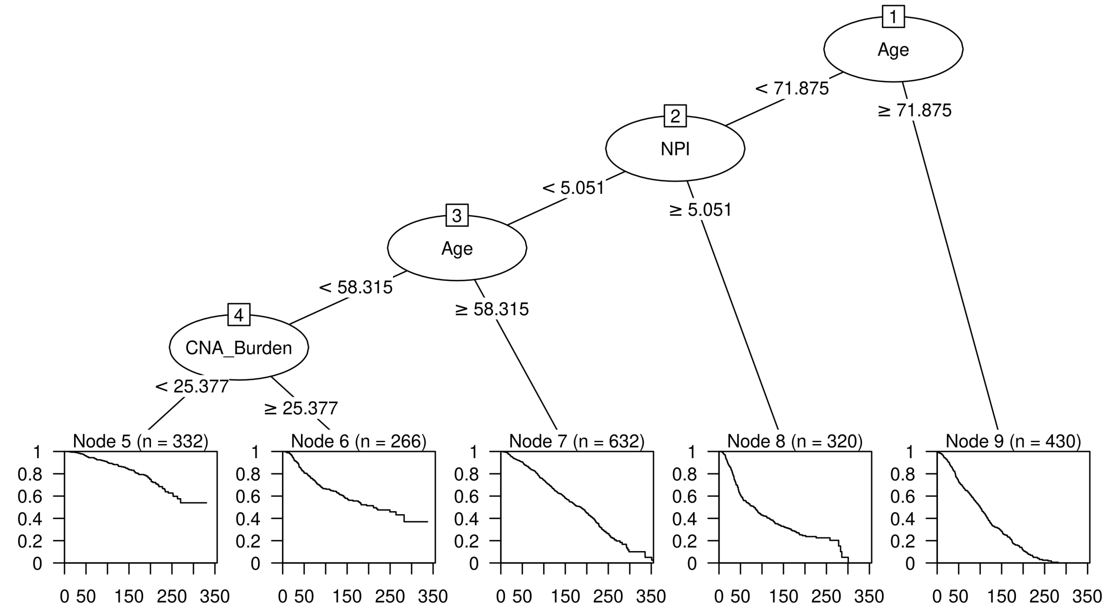


Figure B17: Recursive partitioning survival trees for ten-year overall survival using PAM50, the 6 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

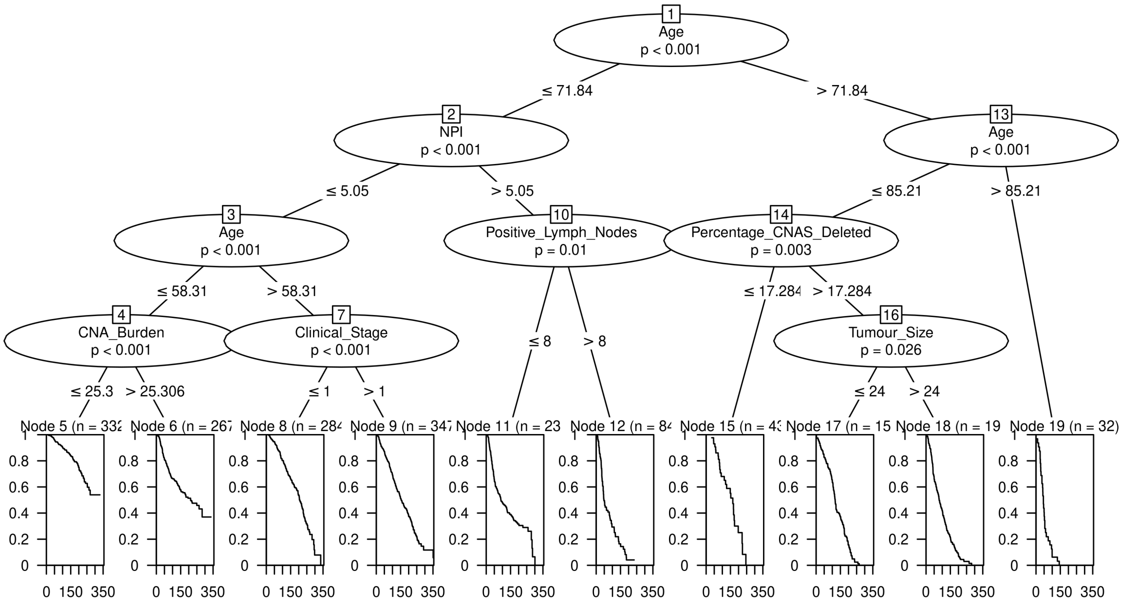
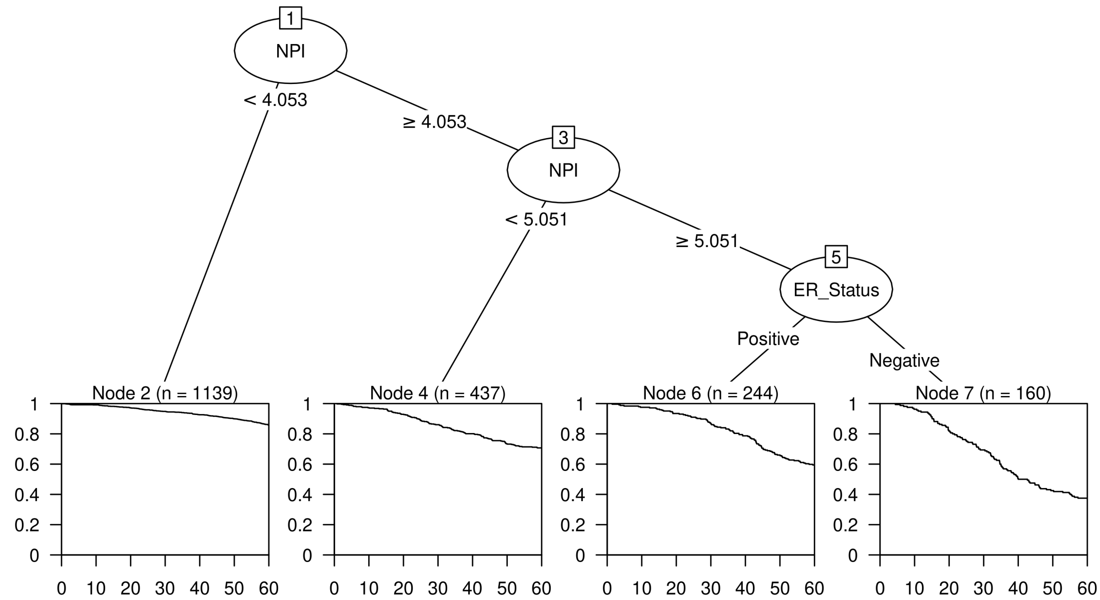


Figure B18: Recursive partitioning survival trees for overall survival using INT-CLUST, the 6 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

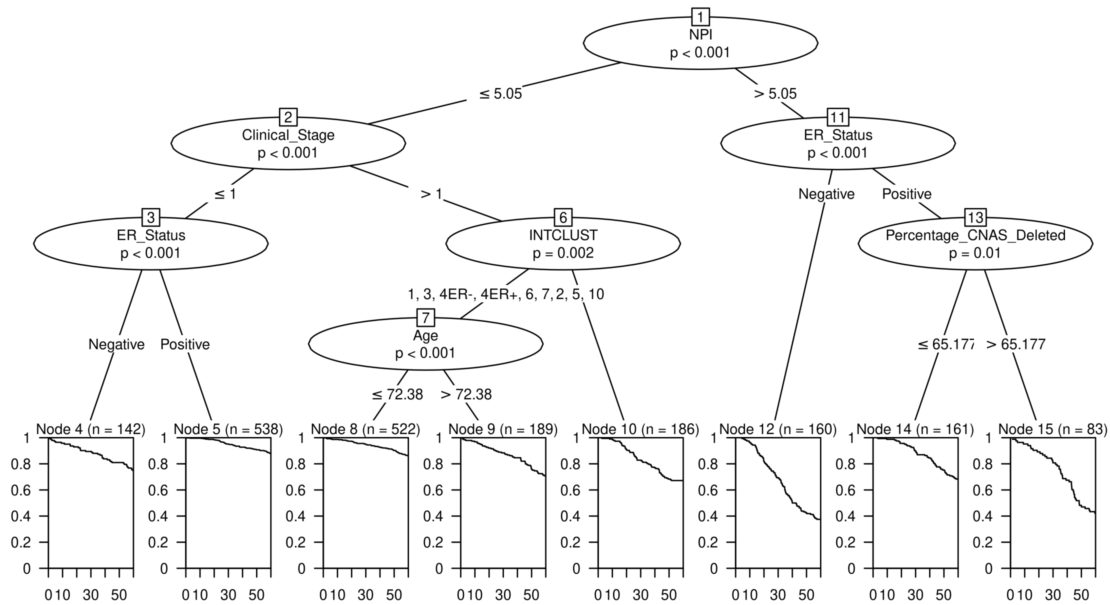
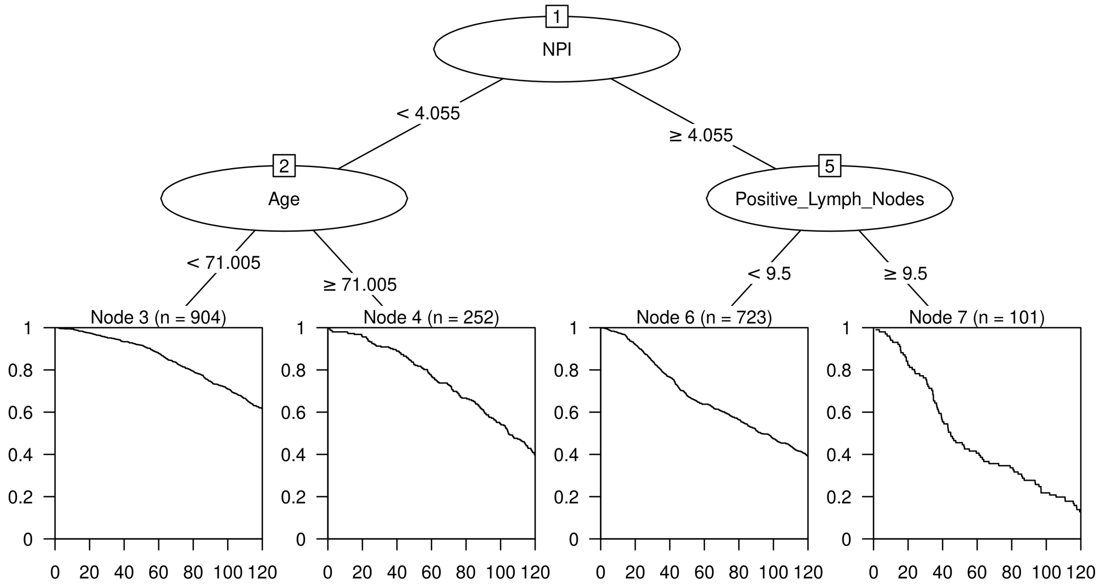


Figure B19: Recursive partitioning survival trees for five-year overall survival using INTCLUST, the 6 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

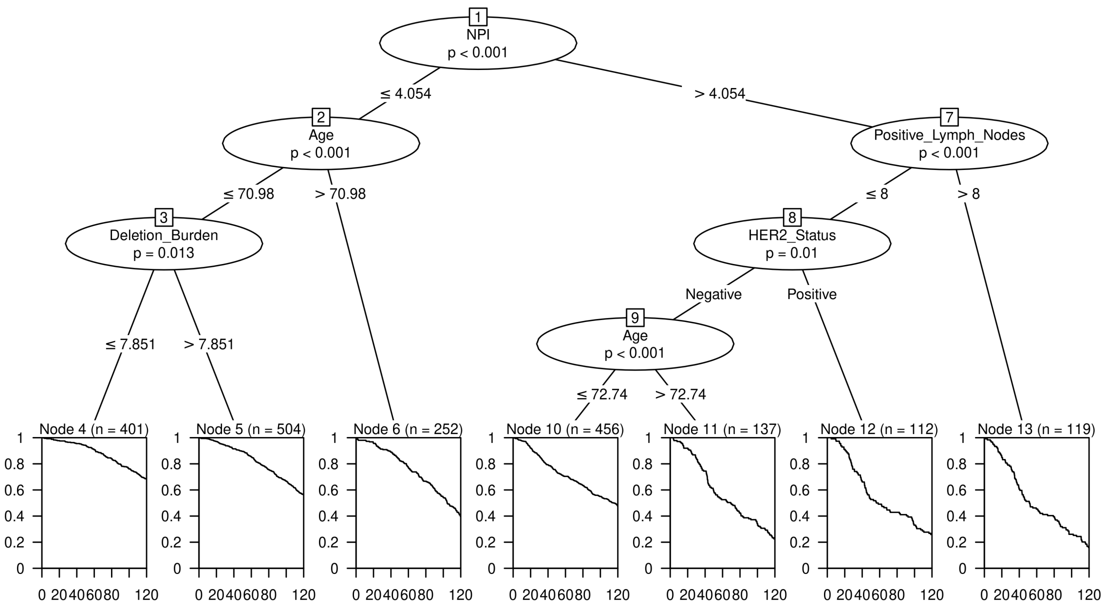
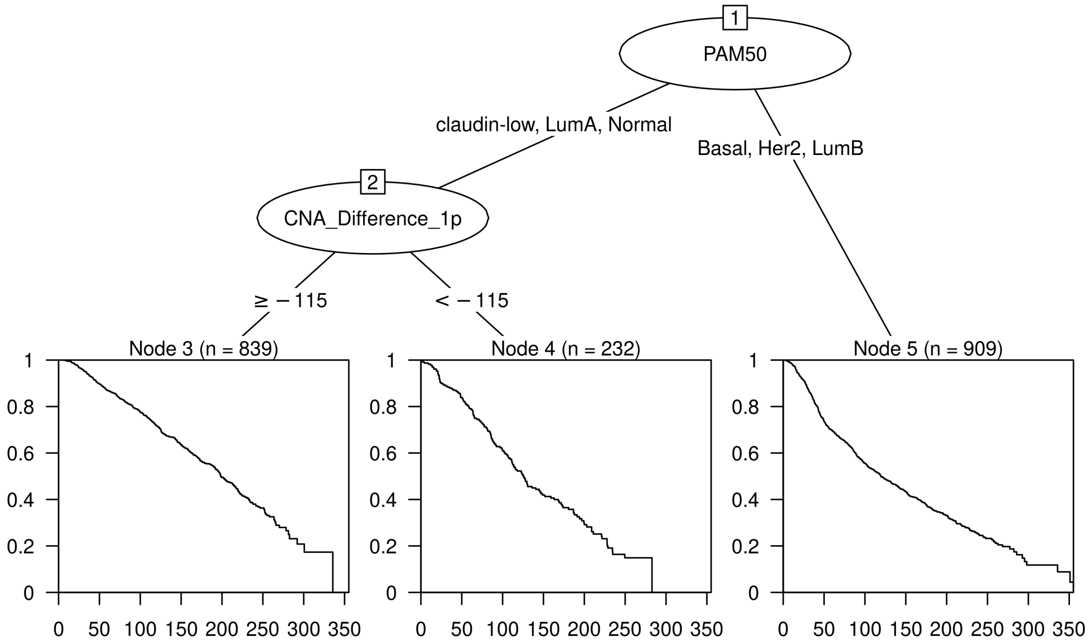


Figure B20: Recursive partitioning survival trees for ten-year overall survival using INTCLUST, the 6 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

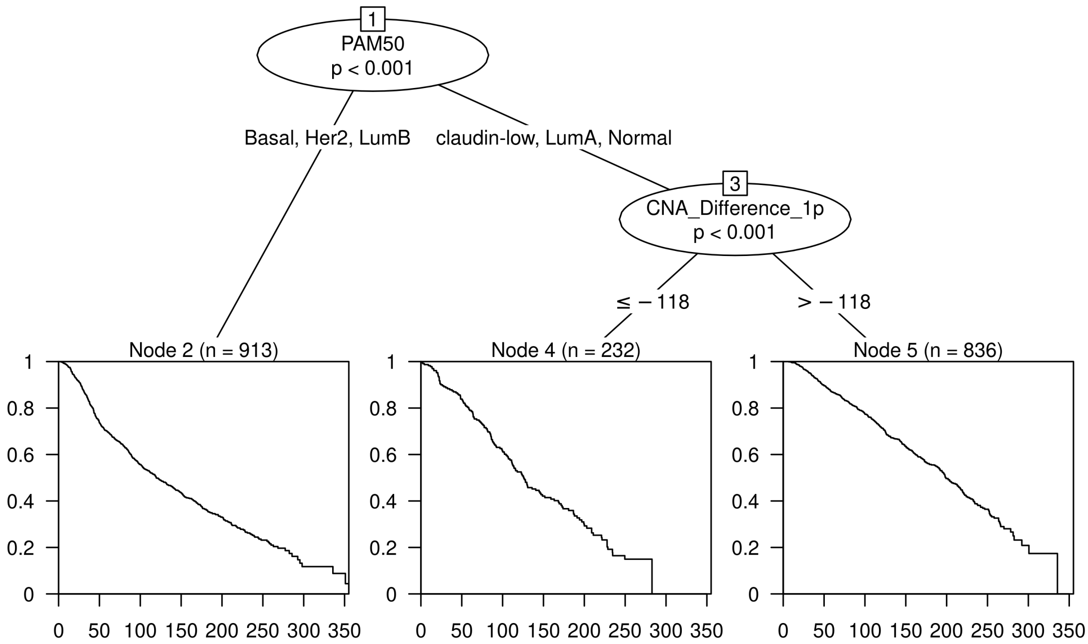
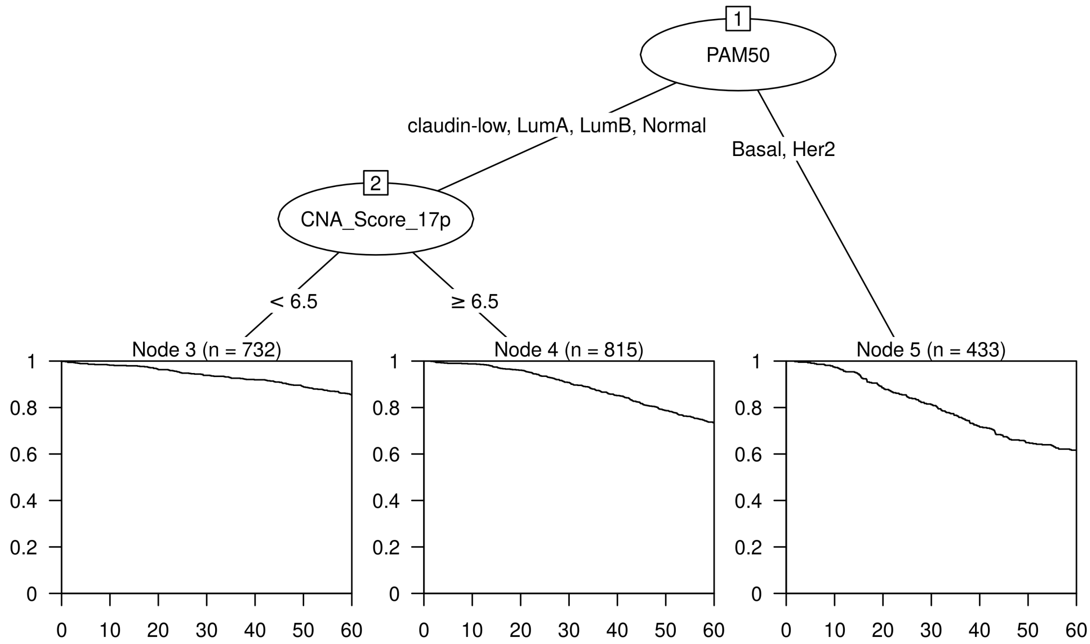


Figure B21: Recursive partitioning survival trees for overall survival using PAM50 and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

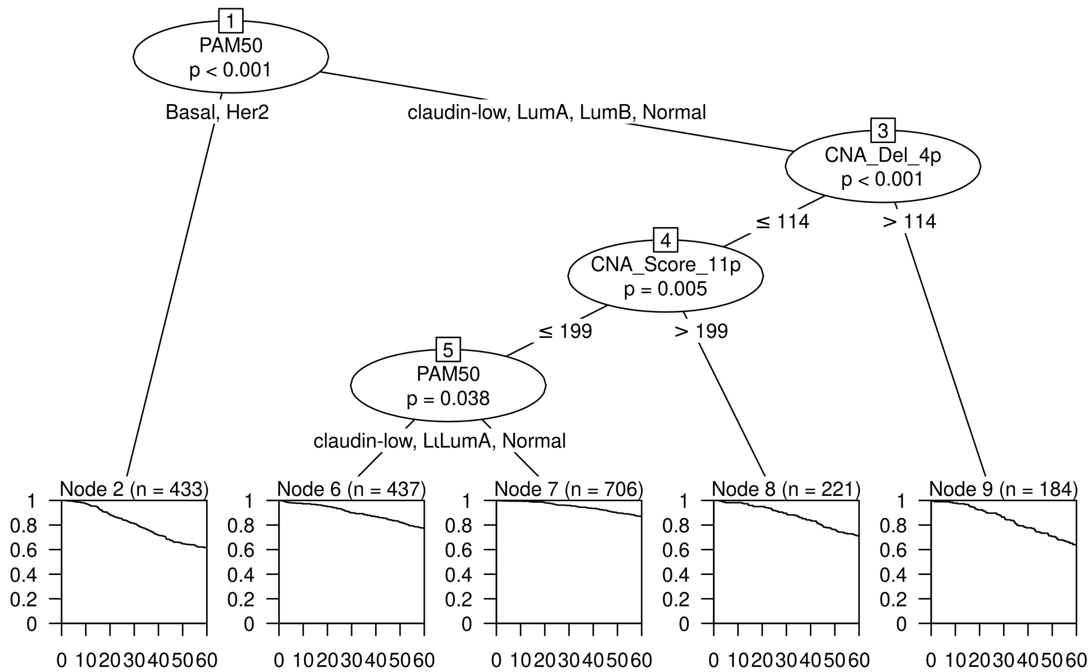
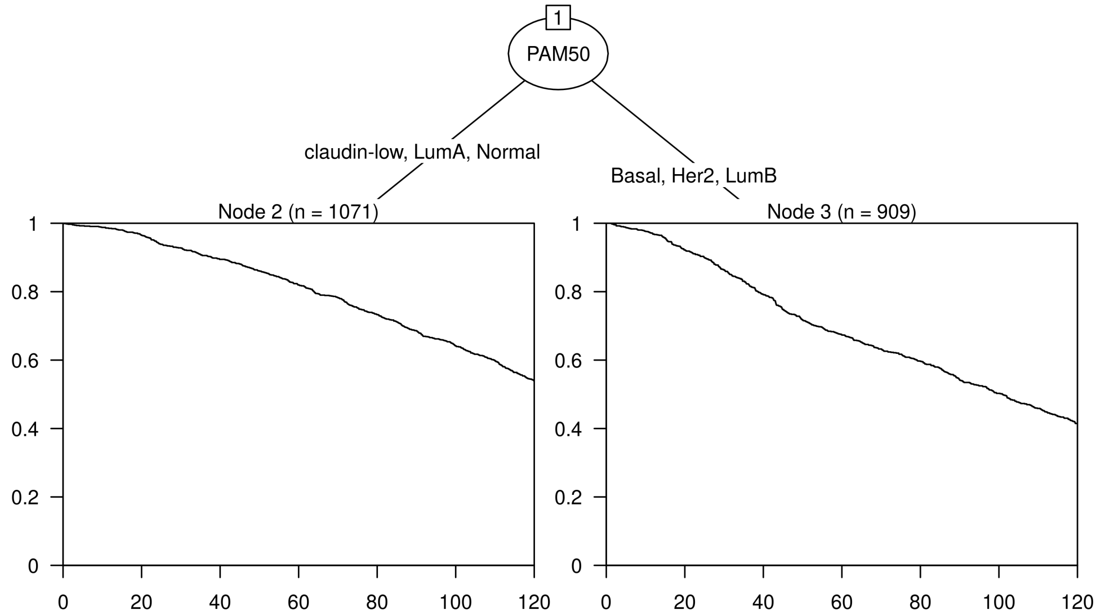


Figure B22: Recursive partitioning survival trees for five-year overall survival using PAM50 and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

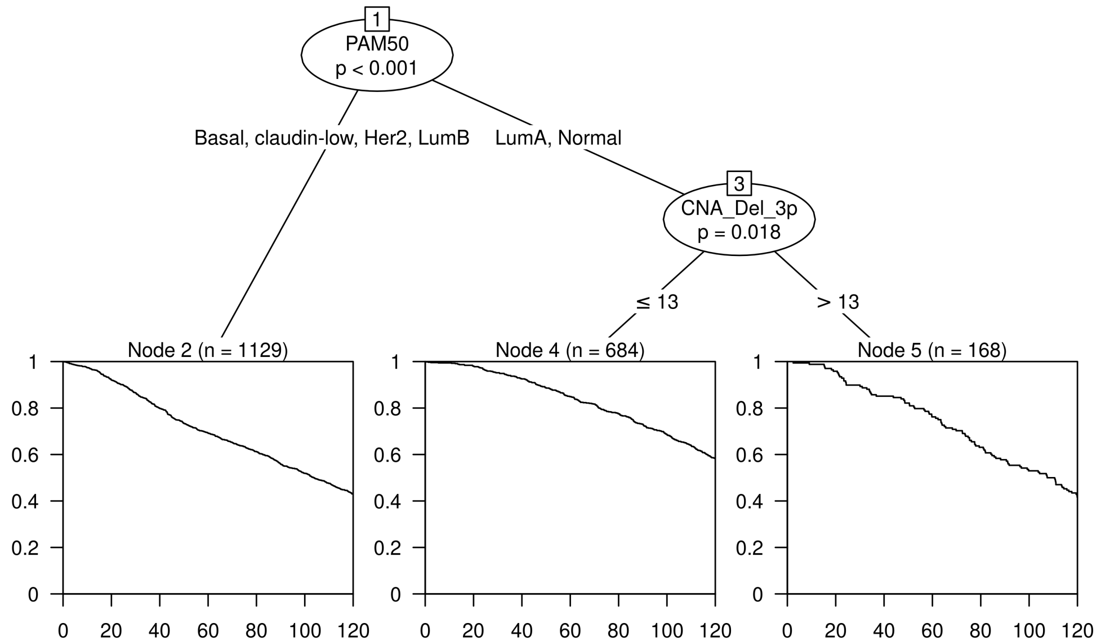
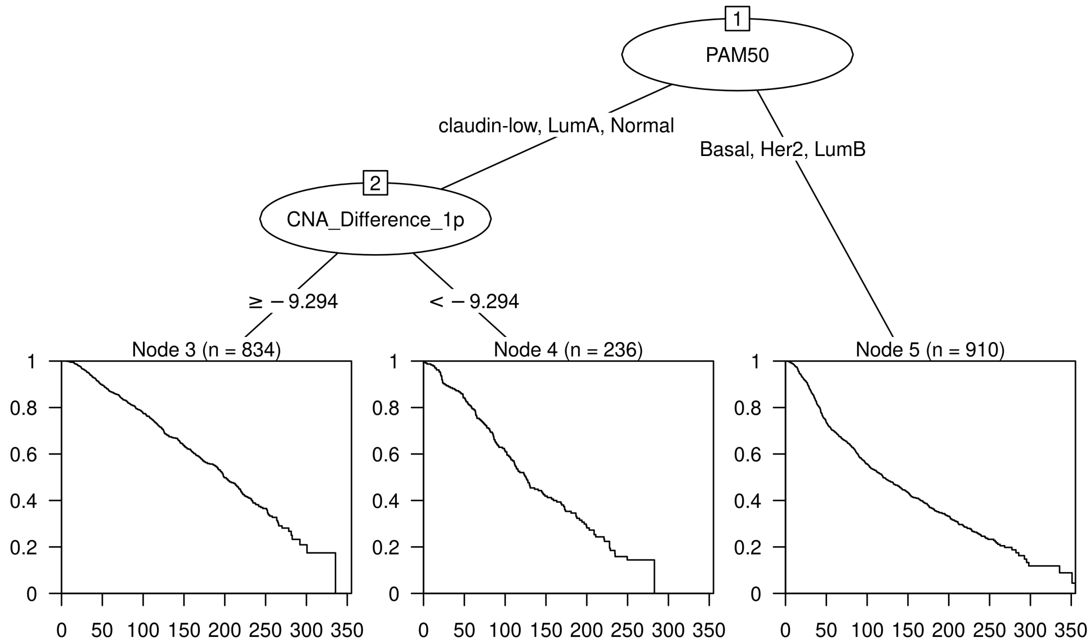


Figure B23: Recursive partitioning survival trees for five-year overall survival using PAM50 and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

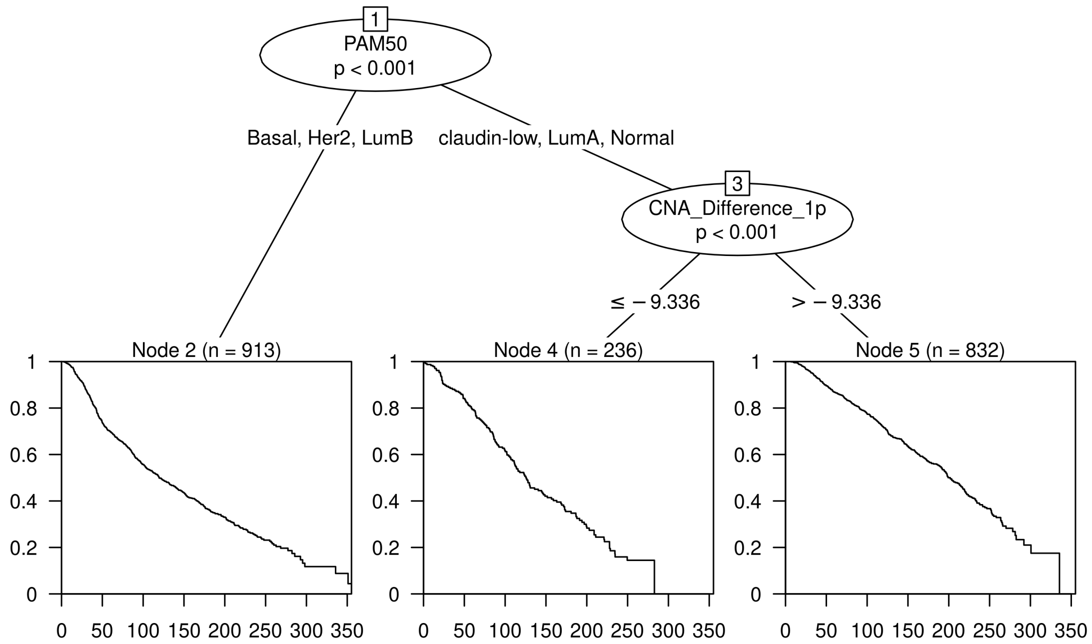
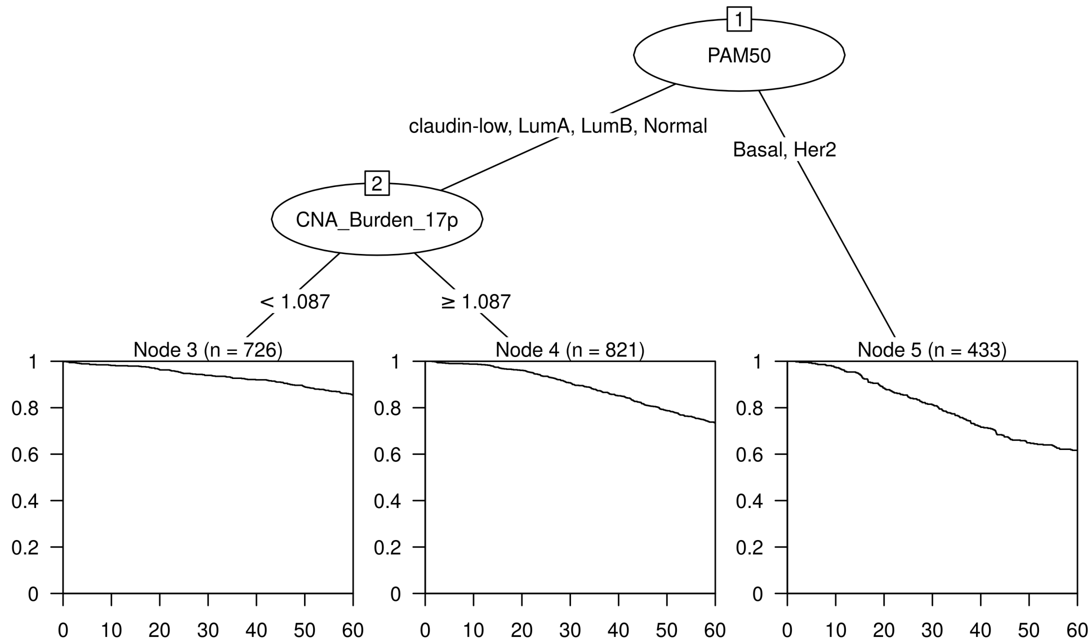


Figure B24: Recursive partitioning survival trees for overall survival using PAM50 and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

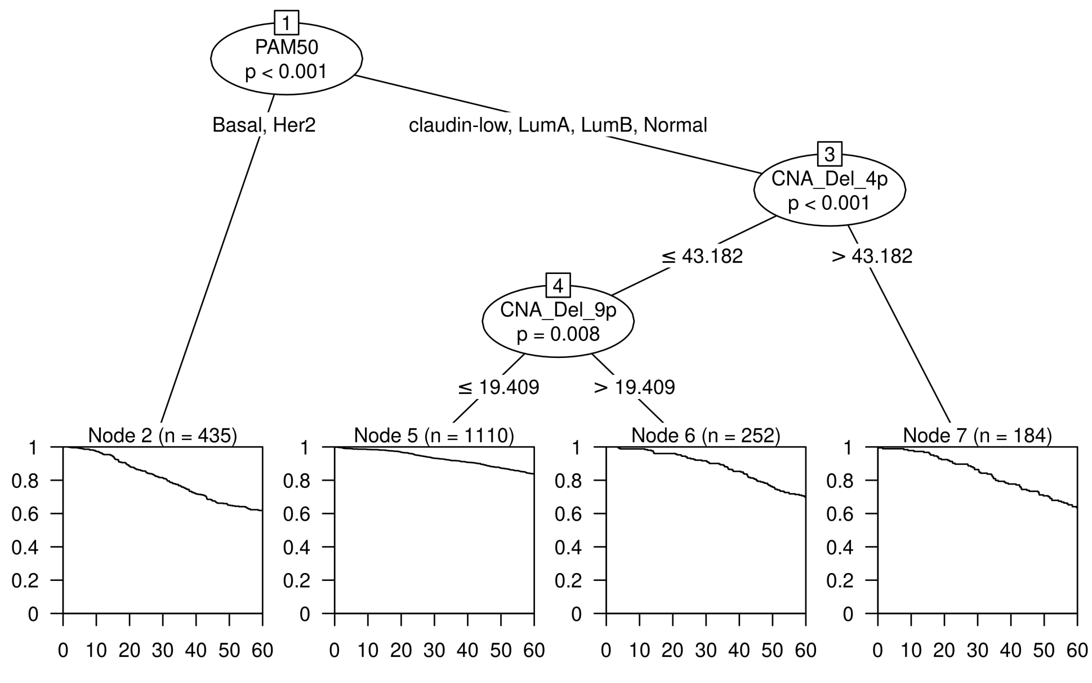
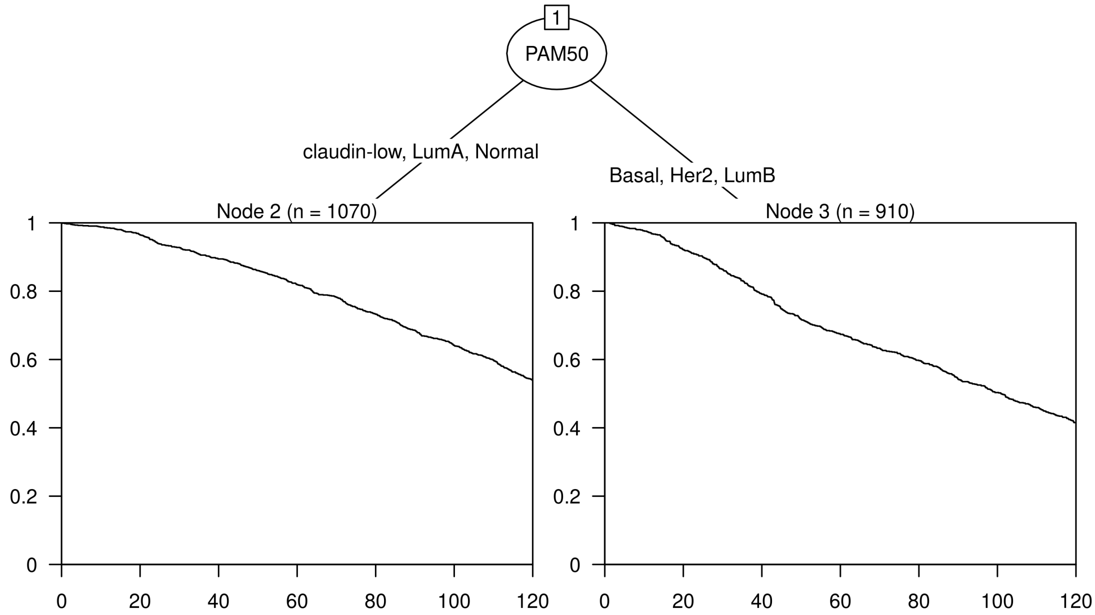


Figure B25: Recursive partitioning survival trees for five-year overall survival using PAM50 and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

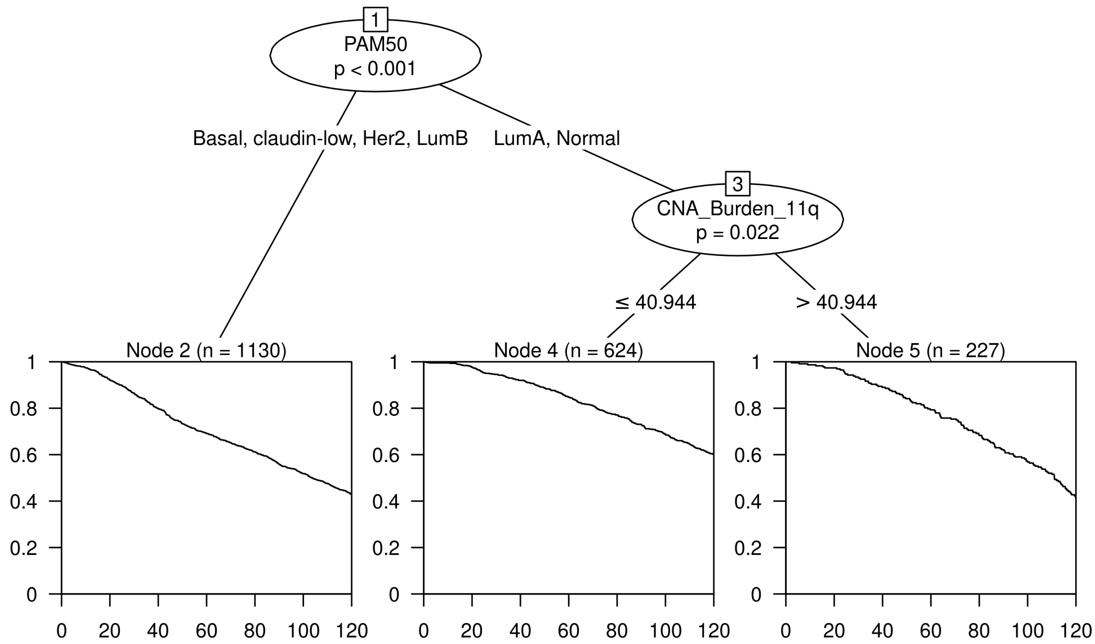
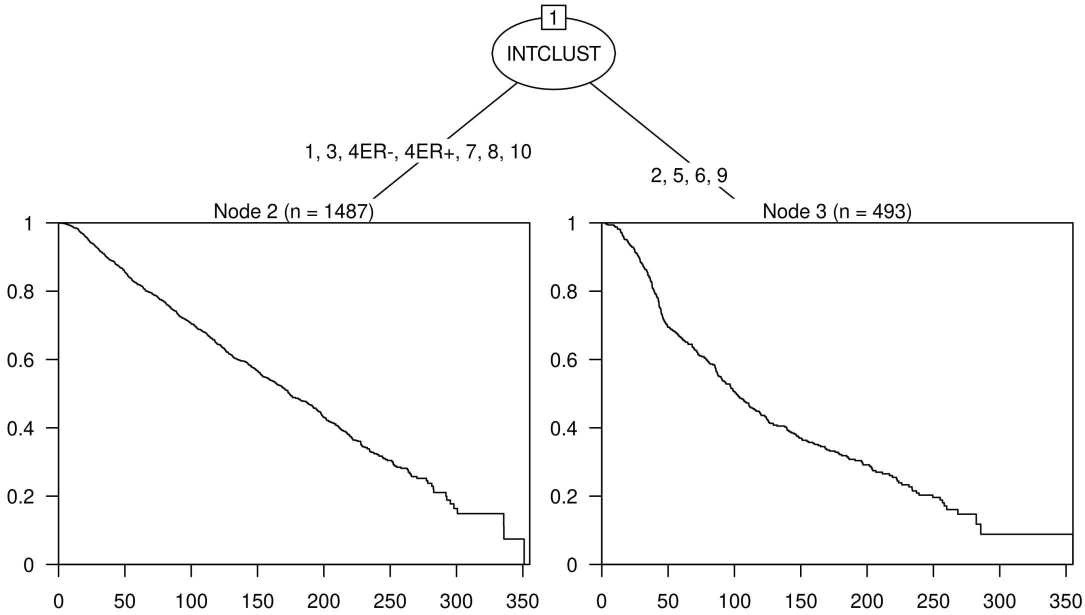


Figure B26: Recursive partitioning survival trees for five-year overall survival using PAM50 and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

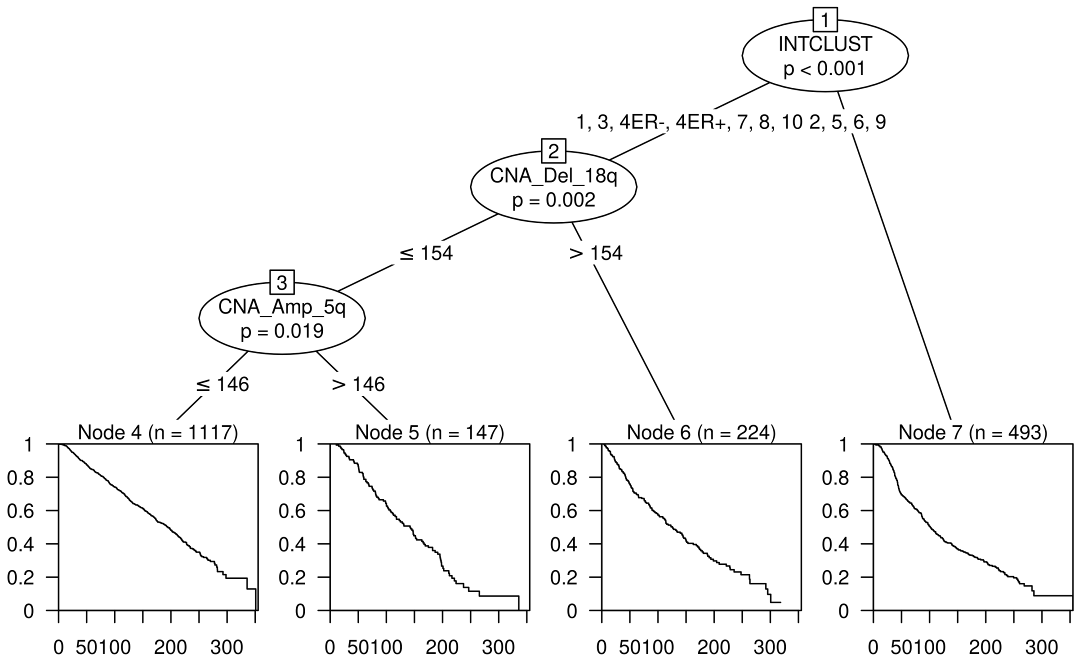
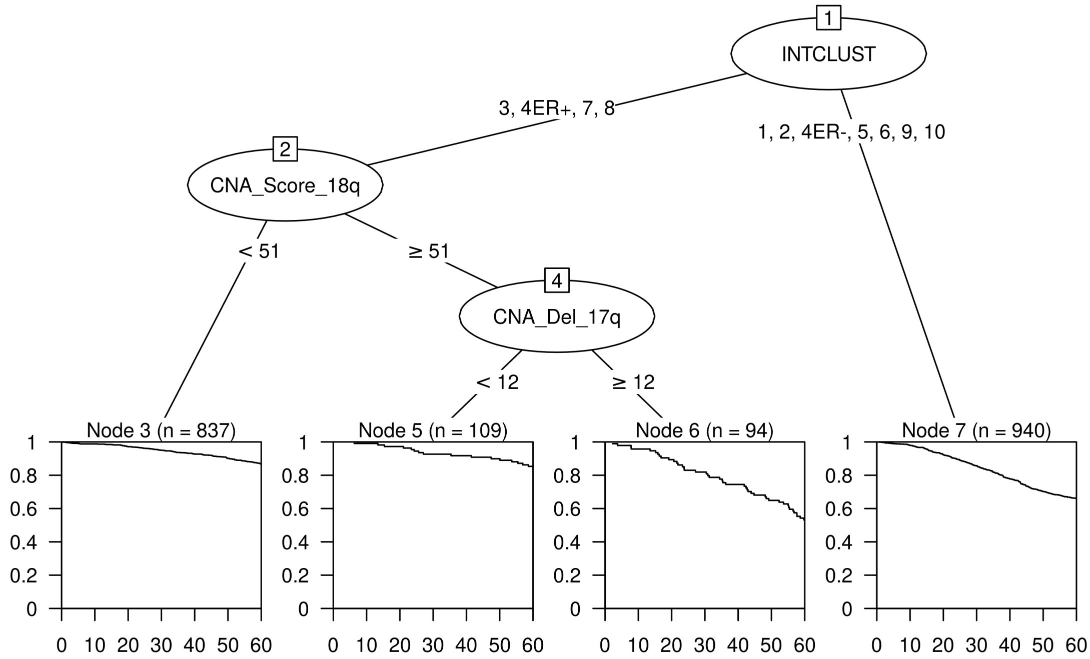


Figure B27: Recursive partitioning survival trees for overall survival using Integrative Cluster and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

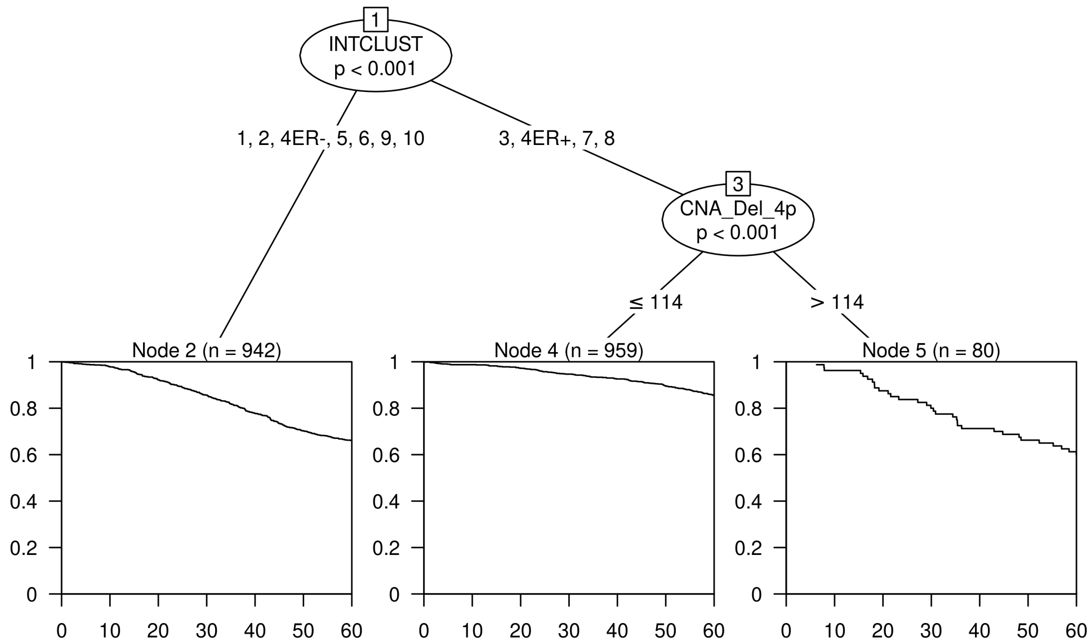
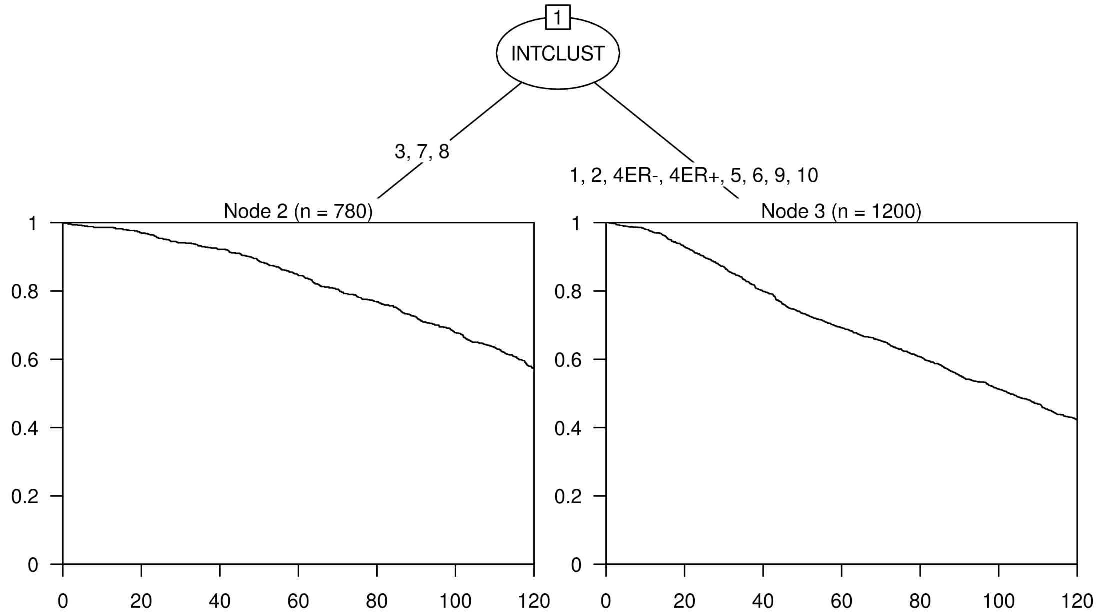


Figure B28: Recursive partitioning survival trees for five-year overall survival using Integrative Cluster and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

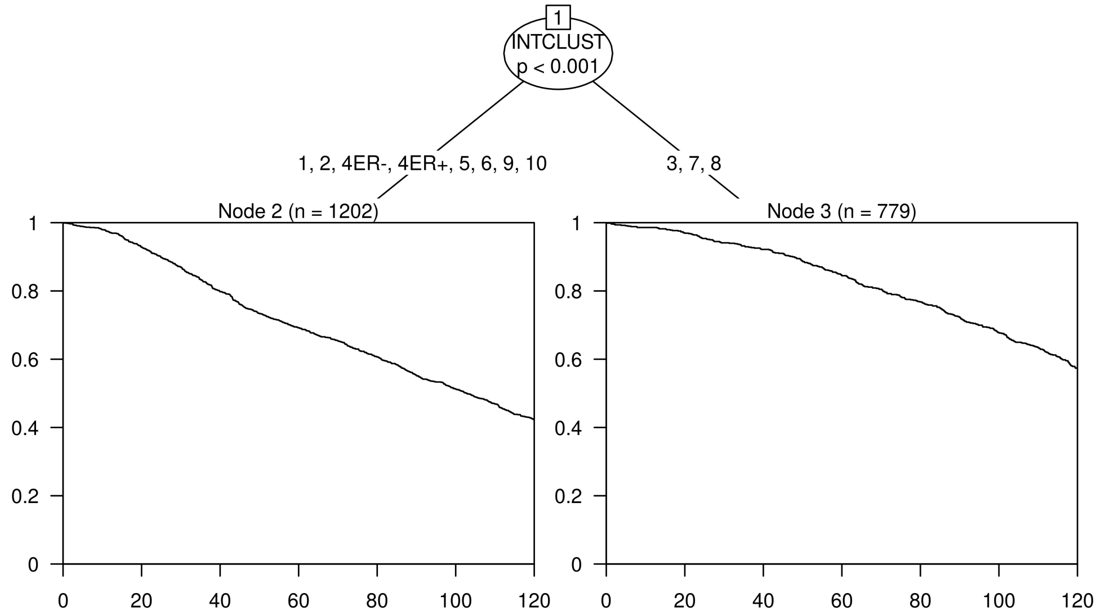
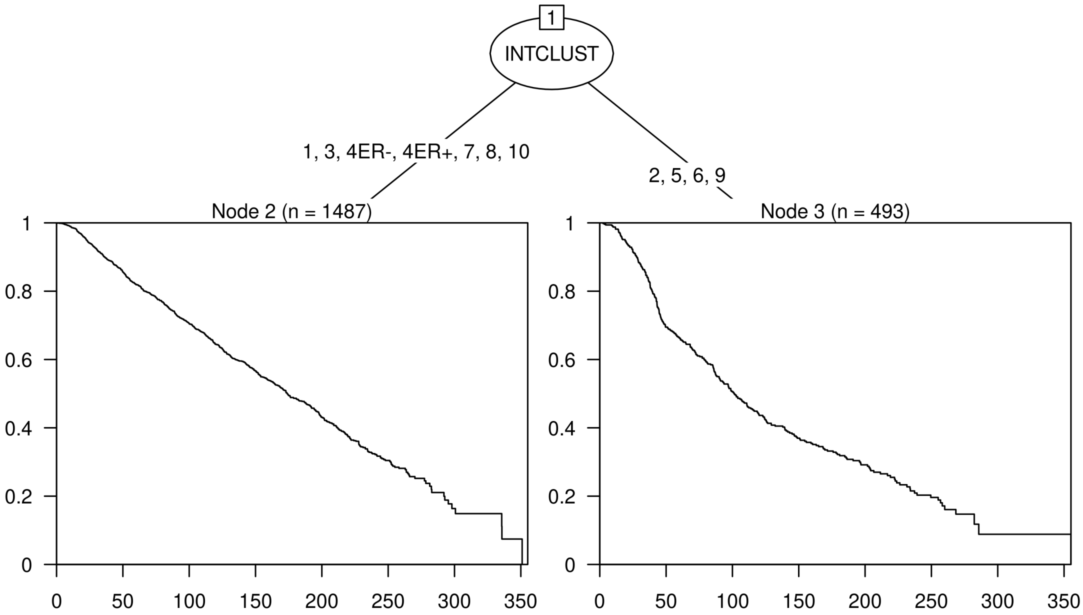


Figure B29: Recursive partitioning survival trees for five-year overall survival using Integrative Cluster and the 42 chromosome arm CNA Score metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

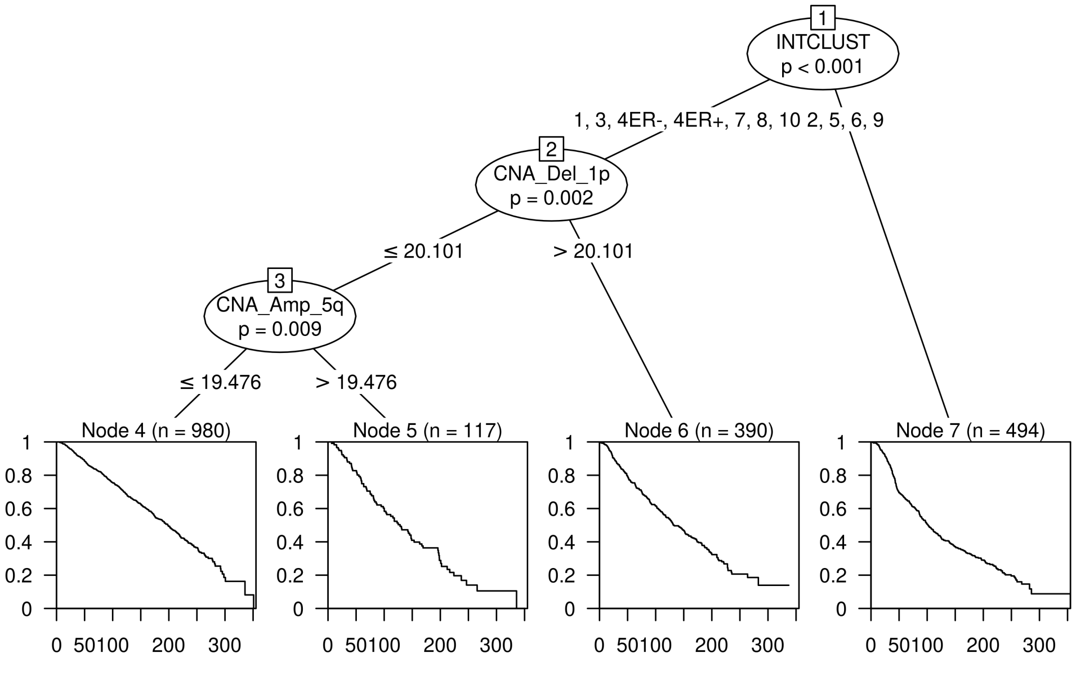
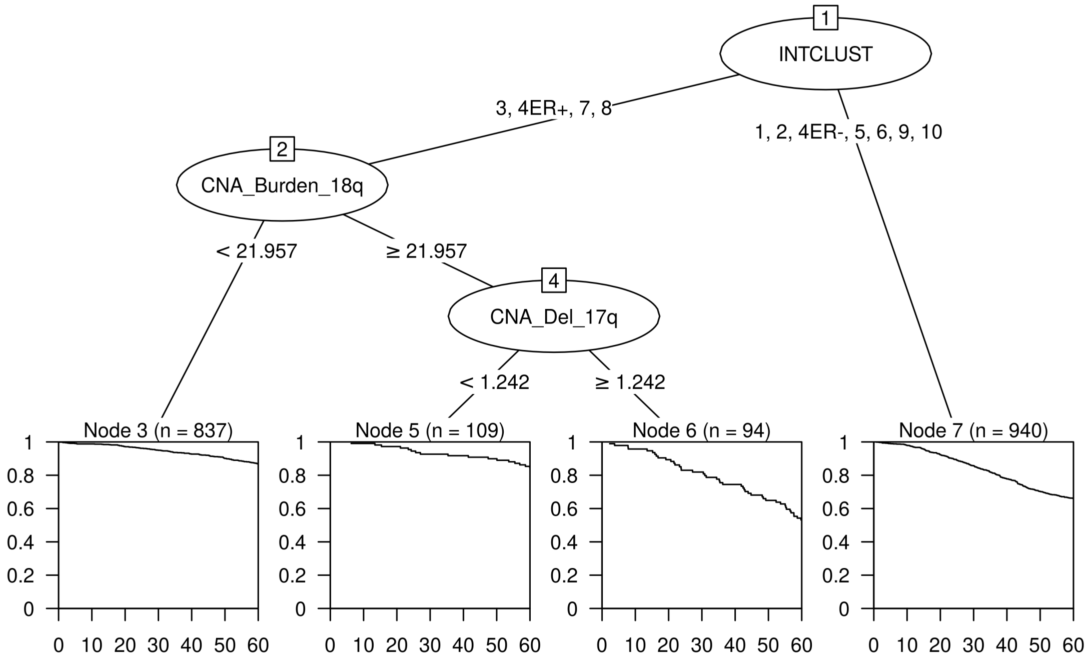


Figure B30: Recursive partitioning survival trees for overall survival using Integrative Cluster and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

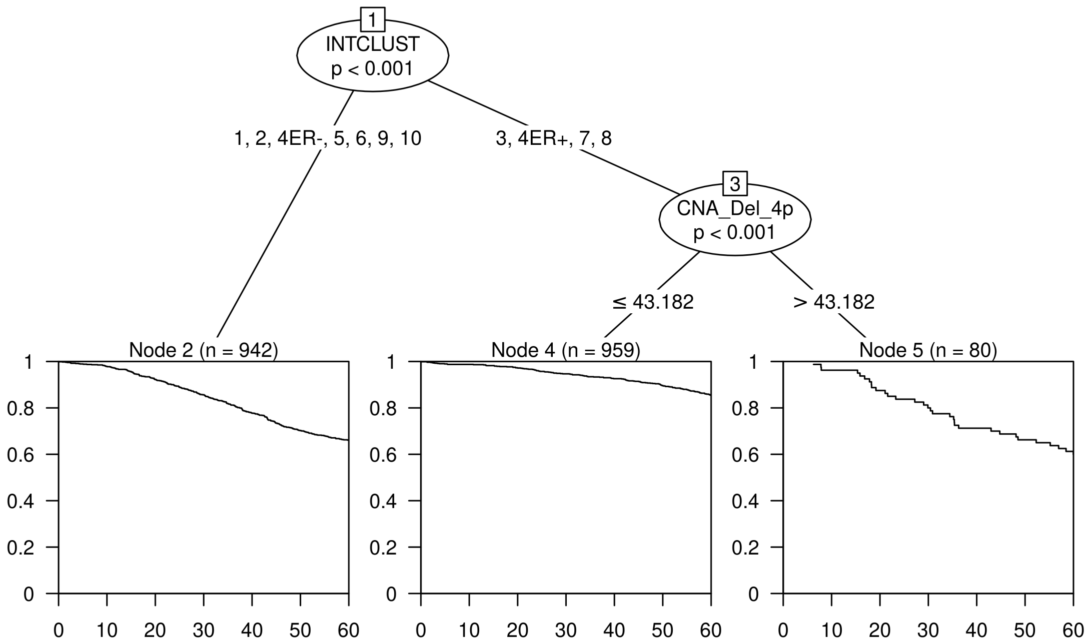
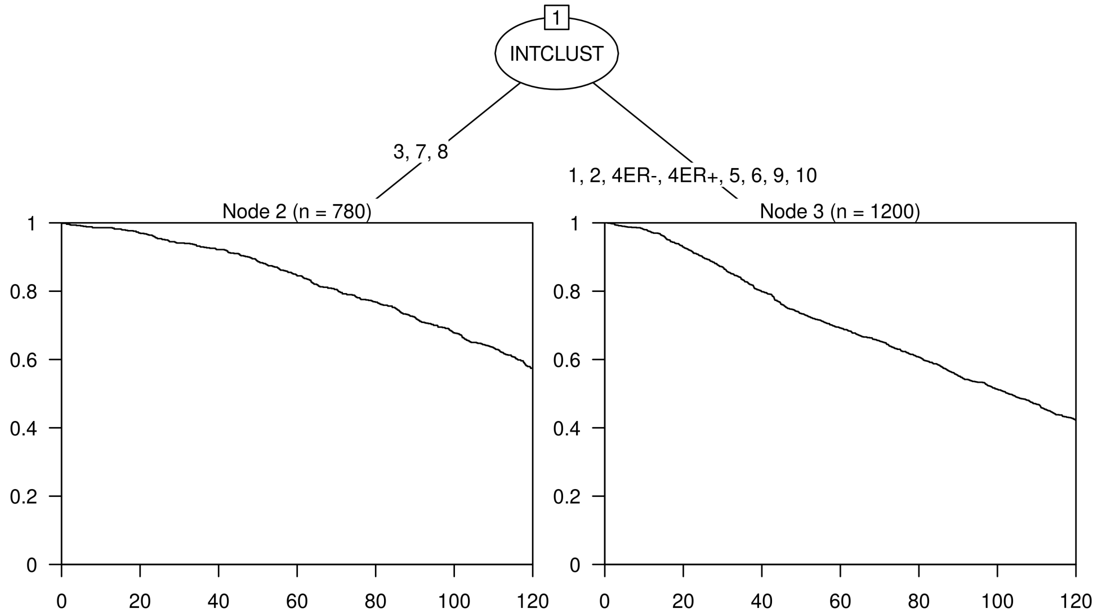


Figure B31: Recursive partitioning survival trees for five-year overall survival using Integrative Cluster and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

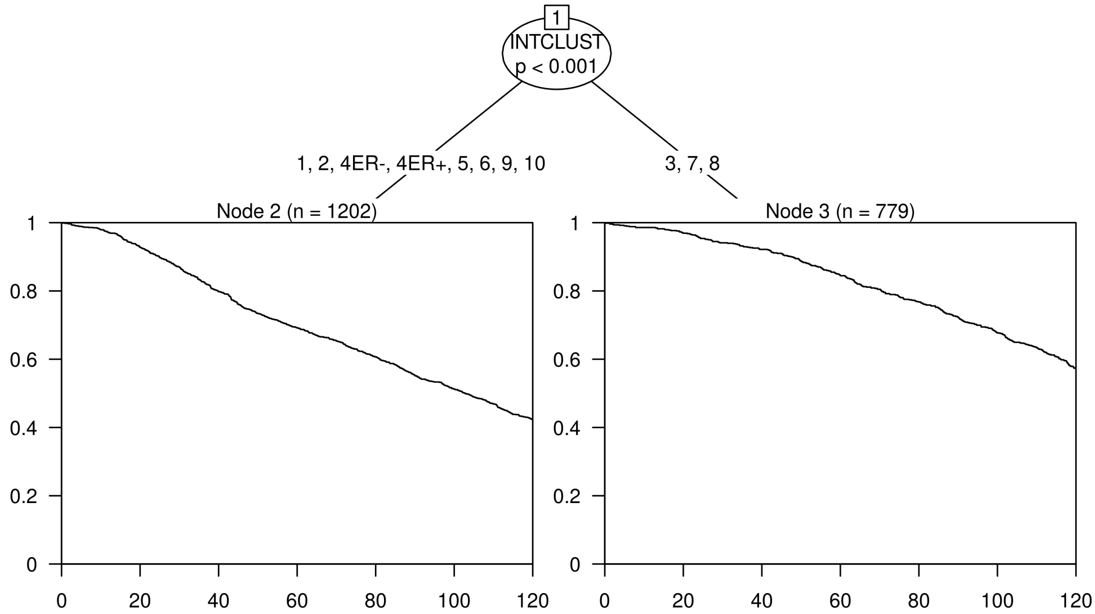
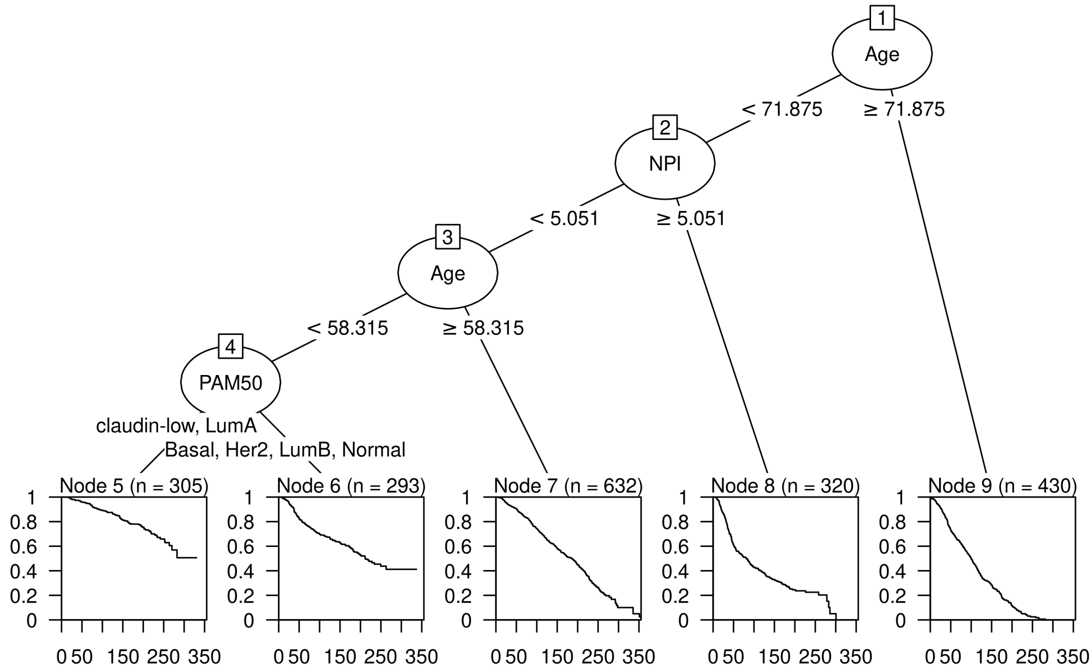


Figure B32: Recursive partitioning survival trees for five-year overall survival using Integrative Cluster and the 42 chromosome arm CNA Burden metrics as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

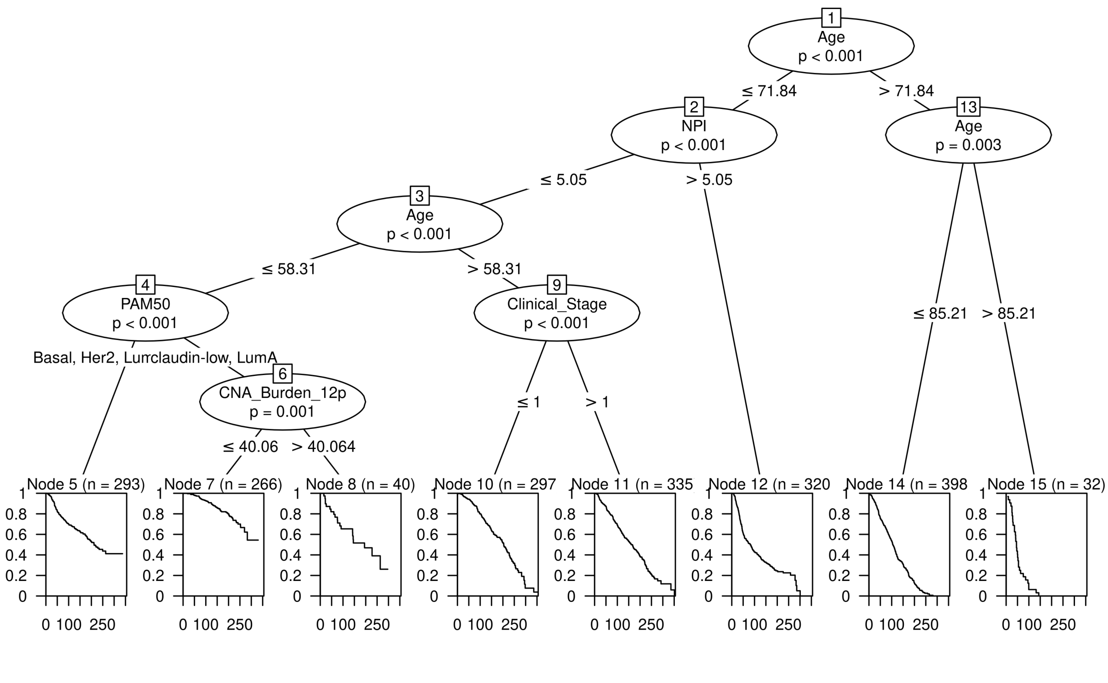
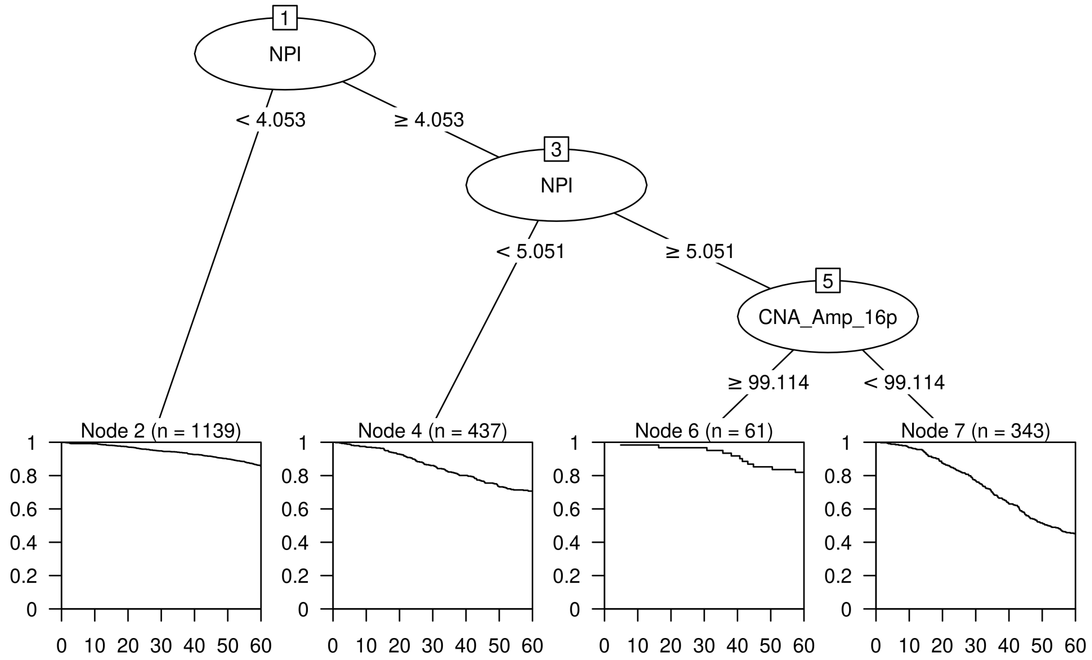


Figure B33: Recursive partitioning survival trees for overall survival using PAM50, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

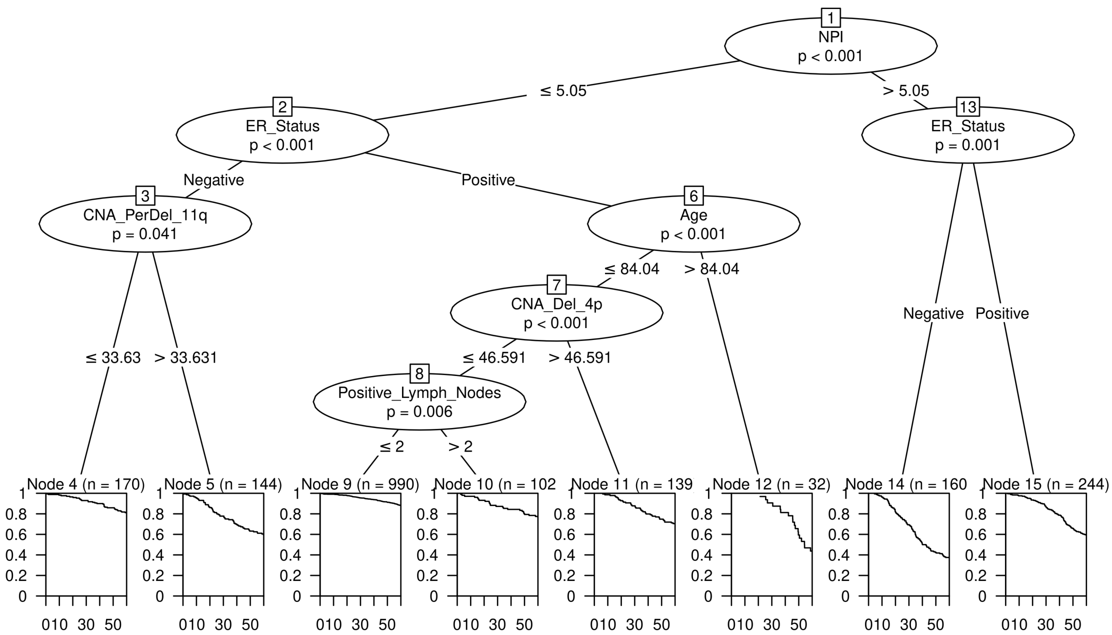
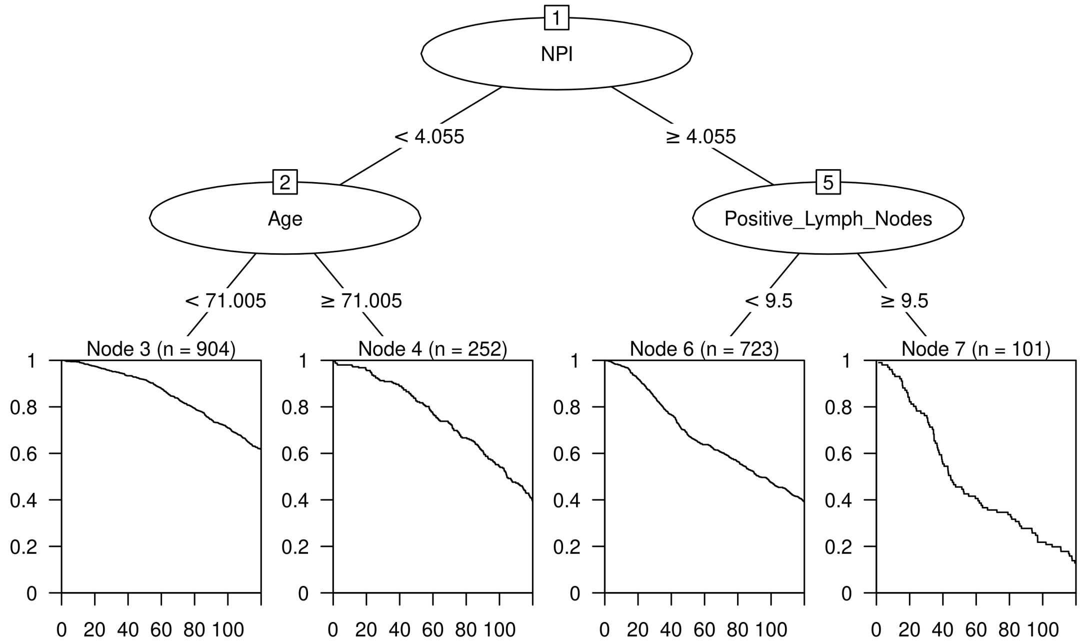


Figure B34: Recursive partitioning survival trees for five-year overall survival using PAM50, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

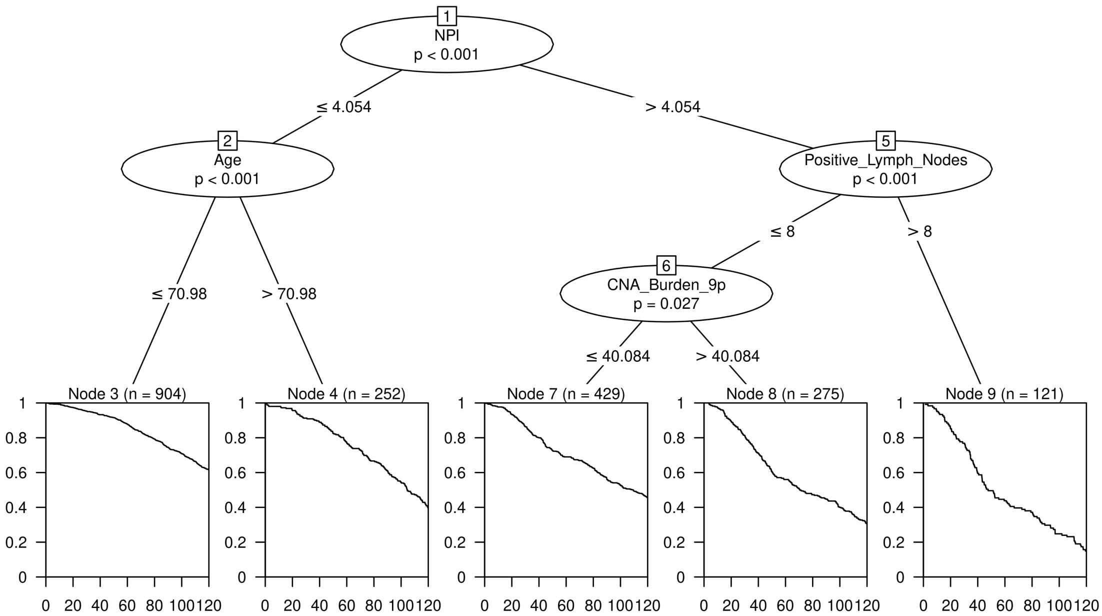
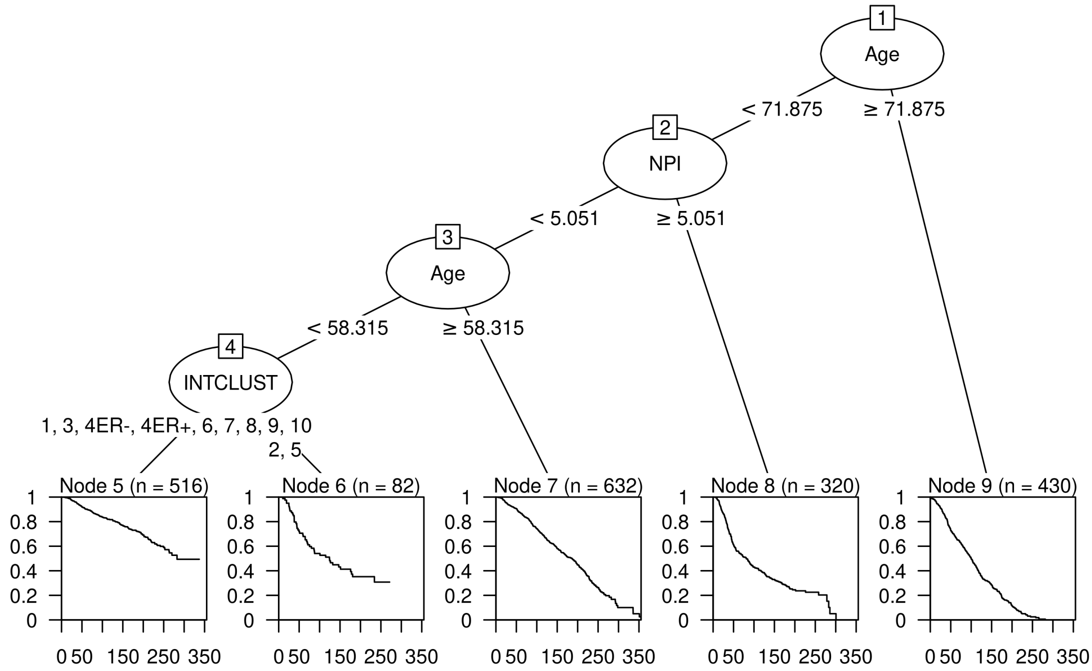


Figure B35: Recursive partitioning survival trees for ten-year overall survival using PAM50, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

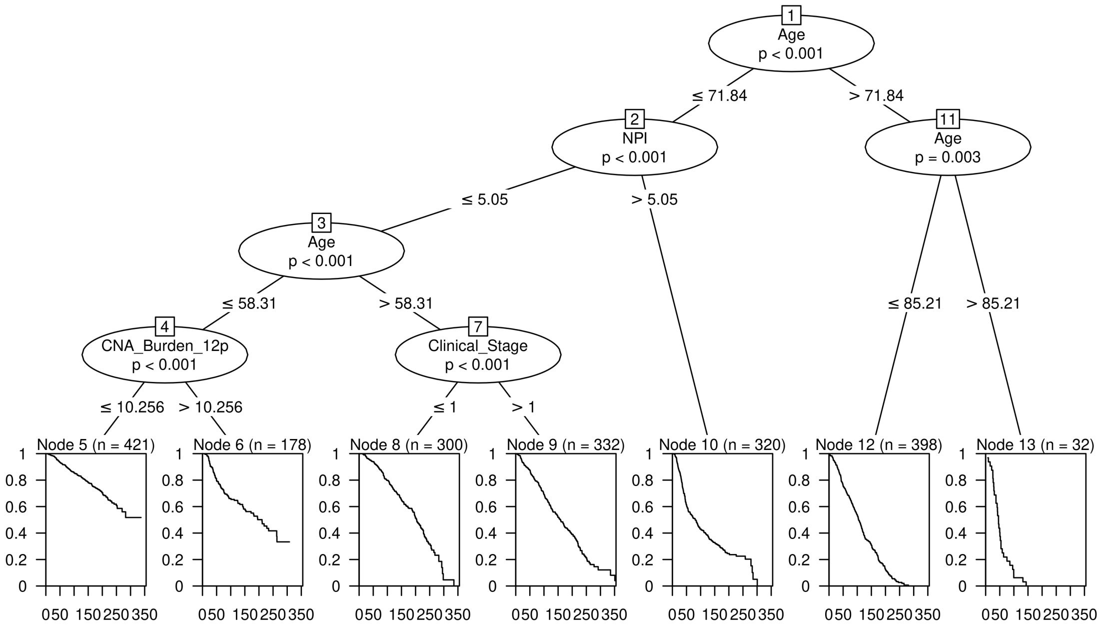
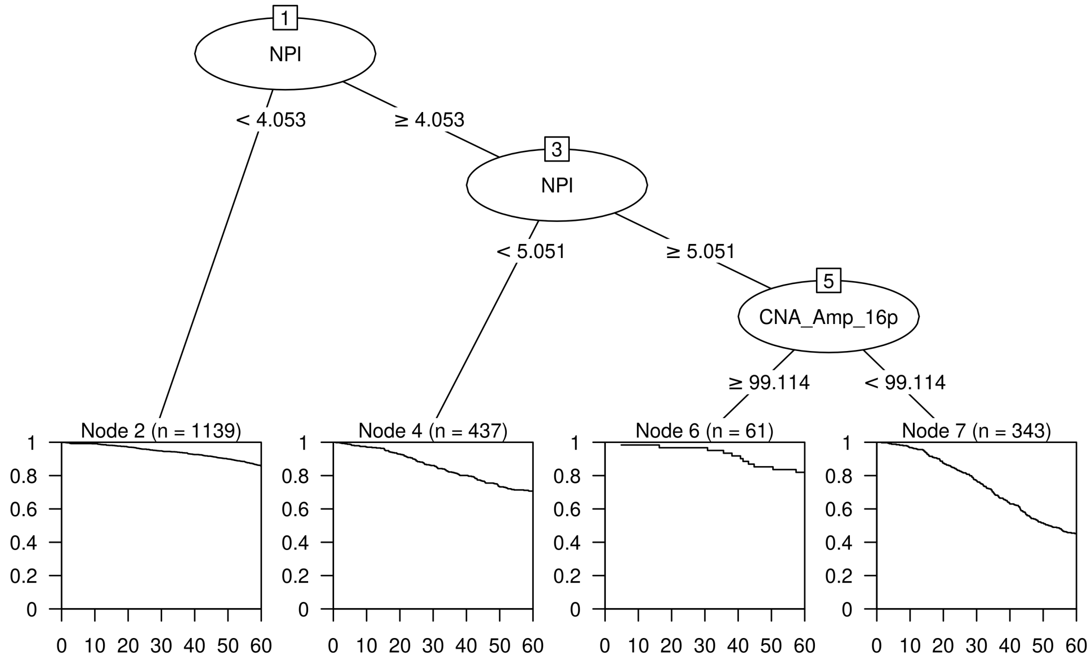


Figure B36: Recursive partitioning survival trees for overall survival using INTCLUST, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

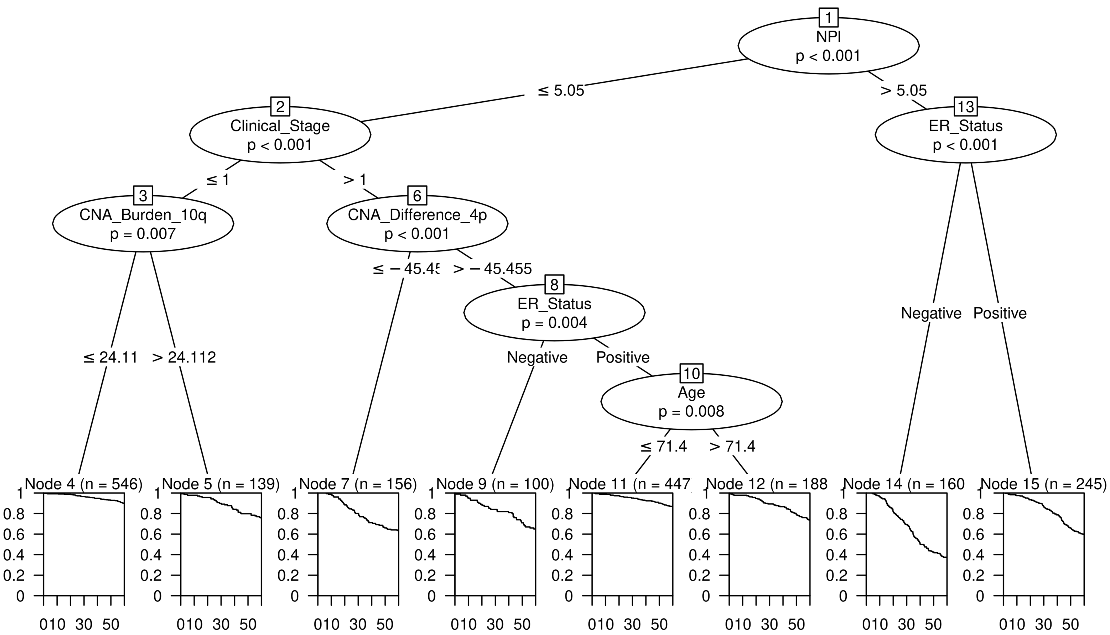
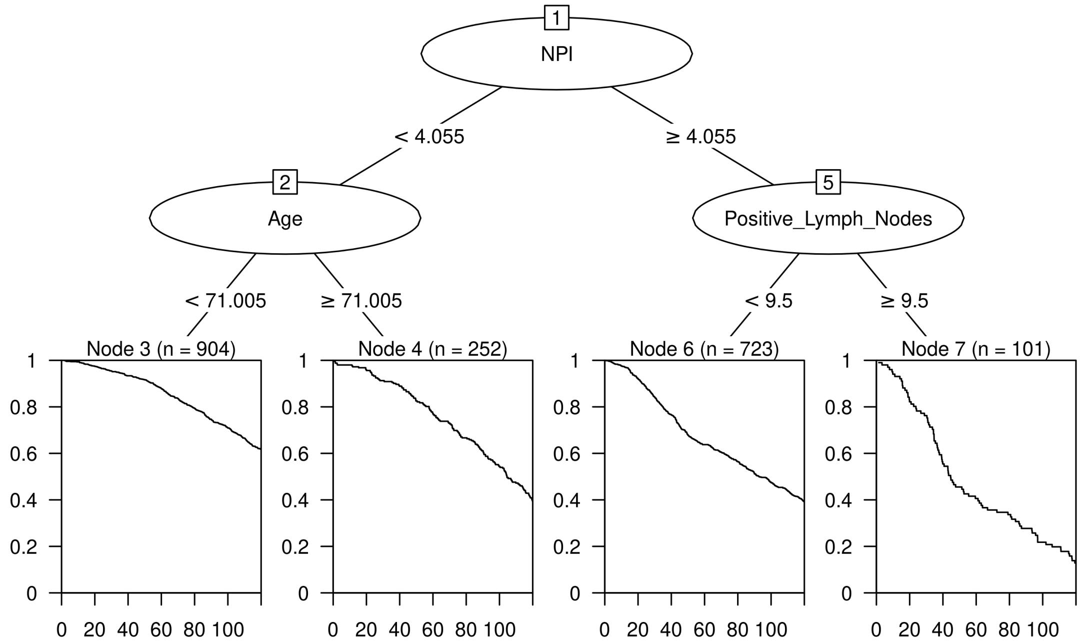


Figure B37: Recursive partitioning survival trees for five-year overall survival using INTCLUST, the 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

(A)



(B)

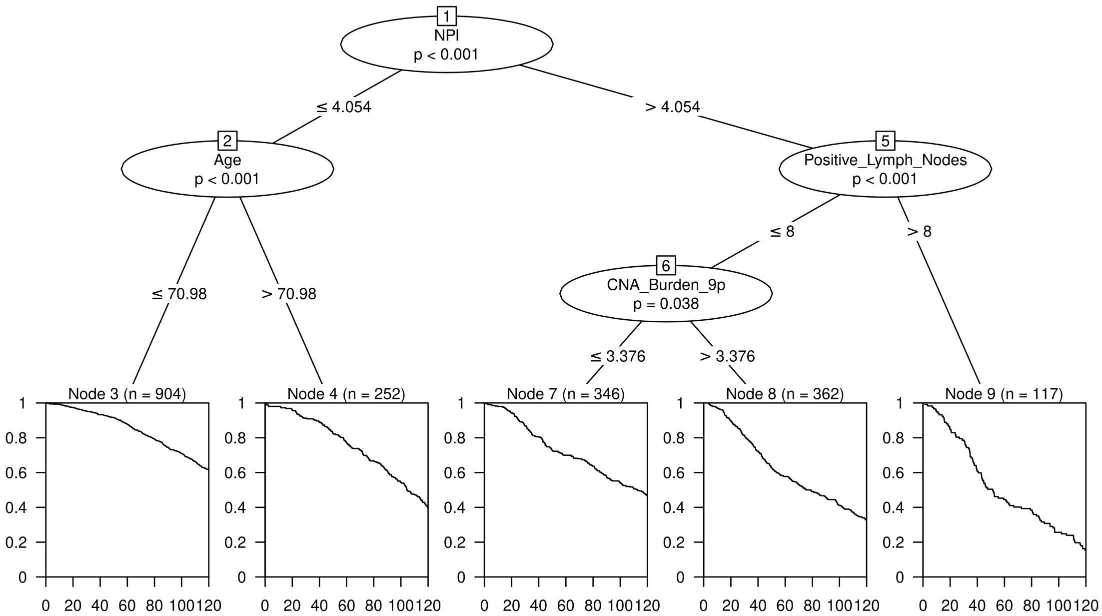


Figure B38: Recursive partitioning survival trees for ten-year overall survival using INTCLUST, te 42 CNA Burden metrics and a number of clinical variables as candidate predictors. (A) Trees fitted using the rpart algorithm and (B) trees fitted using the ctree algorithm.

Appendix C

Appendix C contains lists of genes measured on the Oncotype DX assay, MammaPrint assay, Prosigna (PAM50) assay and BCI assay.

Table C1: The 21 genes included in the Oncotype DX assay (Paik et al., 2004).

Gene Symbol	Oncotype DX Category
ACTB	Reference gene
GAPDH	Reference gene
GUSB	Reference gene
RPLPO	Reference gene
TFRC	Reference gene
MKI67	Proliferation-related gene
AURKA	Proliferation-related gene
BIRC5	Proliferation-related gene
CCNB1	Proliferation-related gene
MYBL2	Proliferation-related gene
MMP11	Metastasis-related gene
CTSL2	Metastasis-related gene
GRB7	HER2-related gene
ERBB2	HER2-related gene
ESR1	Hormone-related gene
PGR	Hormone-related gene
SCUBE2	Hormone-related gene
BCL2	Hormone-related gene
GSTM1	Hormone-related gene
BAG1	Hormone-related gene
CD68	Hormone-related gene

Table C2: The 70 genes included in the MammaPrint assay (van 't Veer et al., 2002; Tian et al., 2010).

Gene Symbol	MammaPrint Category
BBC3	Evading apoptosis
EGLN1	Evading apoptosis and sustained angiogenesis
FLT1	Evading apoptosis and sustained angiogenesis
HRASLS	Evading apoptosis
STK32B	Evading apoptosis
TGFB3	Insensitivity to anti-growth signals and self-sufficiency in growth signals
RASSF7	Insensitivity to anti-growth signals
DCK	Insensitivity to anti-growth signals
MELK	Insensitivity to anti-growth signals
EXT1	Insensitivity to anti-growth signals
ESM1	Self-sufficiency in growth signals
IGFBP5	Self-sufficiency in growth signals
FGF18	Self-sufficiency in growth signals and sustained angiogenesis
SCUBE2	Self-sufficiency in growth signals
WISP1	Self-sufficiency in growth signals

GNAZ	Self-sufficiency in growth signals
EBF4	Self-sufficiency in growth signals
MTDH	Self-sufficiency in growth signals
PITRM1	Self-sufficiency in growth signals
QSCN6L1 (QSOX1)	Self-sufficiency in growth signals
CCNE2	Limitless replicative potential
ECT2	Limitless replicative potential
CENPA	Limitless replicative potential
LIN9	Limitless replicative potential
KNTC2 (NDC80)	Limitless replicative potential
MCM6	Limitless replicative potential
NUSAP1	Limitless replicative potential
ORC6L	Limitless replicative potential
TSPYL5	Limitless replicative potential
RUNDC1	Limitless replicative potential
PRC1	Limitless replicative potential
RFC4	Limitless replicative potential
RECQL5	Limitless replicative potential
CDCA7	Limitless replicative potential
DTL	Limitless replicative potential
COL4A2	Tissue invasion and metastasis and sustained angiogenesis
GPR180	Tissue invasion and metastasis and sustained angiogenesis
MMP9	Tissue invasion and metastasis and sustained angiogenesis
GPR126	Tissue invasion and metastasis
RTN4RL1	Tissue invasion and metastasis
CDC42BPA	Tissue invasion and metastasis
DIAPH3	Tissue invasion and metastasis
PALM2	Tissue invasion and metastasis
ALDH4A1	Sustained angiogenesis
AYTL2 (LPCAT2)	Sustained angiogenesis
OXCT1	Sustained angiogenesis
PECI	Sustained angiogenesis
GMPS (LOC728564)	Sustained angiogenesis
GSTM3	Sustained angiogenesis
SLC2A3	Sustained angiogenesis
LOC100288906	Unknown function
C9orf30	Unknown function
C20orf46	Unknown function
ZNF533	Unknown function
C16orf61	Unknown function
SERF1A	Unknown function
LOC730018	Unknown function
LOC100131053	Unknown function
AA555029_RC	Unknown function
LGP2 (DHX58)	Miscellaneous
NMU	Miscellaneous
UCHL5	Miscellaneous
JHDM1D	Miscellaneous
AP2B1	Miscellaneous
MS4A7	Miscellaneous
RAB6B	Miscellaneous

Table C3: The 58 genes included in the Prosigna assay (Duffy et al., 2017)

Gene Symbol	**Gene Symbol**
ACTB	FOXC1
GUSB	GPR160
MRPL19	GRB7
PSMC4	KIF2C
PUM1	KNTC2
RPLP0	KRT14
SF3A1	KRT17
TFRC	KRT5
ACTR3B	MAPT
ANLN	MDM2
BAG1	MELK
BCL2	MIA
BIRC5	MKI67
BLVRA	MLPH
CCNB1	MMP11
CCNE1	MYBL2
CDC20	MYC
CDC6	NAT1
CDCA1	ORC6L
CDH3	PGR
CENPF	PHGDH
CEP55	PTTG1
CXXC5	RRM2
EGFR	SFRP1
ERBB2	SLC39A6
ESR1	TMEM45B
EXO1	TYMS
FGFR4	UBE2C
FOXA1	UBE2T

Table C4: The 11 genes included in the Breast Cancer Index assay (Jerevall et al., 2011).

Gene Symbol	**BCI Category**
ACTB	Reference gene
HMBS	Reference gene
SDHA	Reference gene
UBC	Reference gene
HOXB13	H:I index gene
IL17BR	H:I index gene
BUB1B	Molecular Grade Index gene
CENPA	Molecular Grade Index gene
NEK2	Molecular Grade Index gene
RACGAP1	Molecular Grade Index gene
RRM2	Molecular Grade Index gene

Appendix D

Appendix D contains a list of genes that are present in the METABRIC CNA and gene expression data utilised in this thesis but missing from the IntClust gene list.

Table D1: Genes present in our analysis but missing from the IntClust gene set (Curtis et al., 2012).

ProbeID	Gene	Gene Description
ILMN_2044617	MTERFD1	MTERF domain containing 1
ILMN_1679867	LOC642255	Heat shock transcription factor 1
ILMN_1720819	LOC653566	Signal peptidase complex subunit 2 homolog (S. cerevisiae)
ILMN_1783469	LOC642197	Family with sequence similarity 82, member B
ILMN_1675406	PPAPDC1B	Phosphatidic acid phosphatase type 2 domain containing 1B
ILMN_1685774	LOC647340	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1
ILMN_1665423	ZFP91	Zinc finger protein 91 homolog (mouse)
ILMN_1764323	LOC124512	Chromosome 17 open reading frame 95
ILMN_1763955	LOC653119	Block of proliferation 1
ILMN_1665483	KIAA0020	KIAA0020
ILMN_1651899	LOC653314	
ILMN_1780141	TMEM66	Transmembrane protein 66
ILMN_1769118	38595	Septin 9
ILMN_1699253	LOC729317	Voltage-dependent anion channel 2
ILMN_1693862	MGC70857	Chromosome 8 open reading frame 82
ILMN_1784436	KIAA1688	KIAA1688 protein
ILMN_1796235	CIRH1A	Cirrhosis, autosomal recessive 1A (cirhin)
ILMN_1785660	SRPR	Signal recognition particle receptor (docking protein)
ILMN_1687921	LOC339123	Jumonji domain containing 8
ILMN_2402930	LOC440926	H3 histone, family 3A
ILMN_1746706	LOC653103	Ankyrin repeat domain 11
ILMN_2112599	C16orf80	Chromosome 16 open reading frame 80
ILMN_1879344	HS.571404	Calpain 1, (mu/I) large subunit
ILMN_1675542	LOC729148	Nuclear undecaprenyl pyrophosphate synthase 1 homolog (S. cerevisiae)
ILMN_1740351	KIAA0174	KIAA0174
ILMN_1814812	LOC650546	Ubiquitin specific peptidase 32
ILMN_1790162	LOC441155	Zinc finger CCCH-type containing 11A
ILMN_2059211	KIAA0195	KIAA0195
ILMN_2172269	TMEM183B	Transmembrane protein 183B
ILMN_1655403	LOC730083	Exoribonuclease 2
ILMN_1763404	LOC653226	Homo sapiens clone 24452 mRNA sequence.
ILMN_1700461	AARSD1	Alanyl-tRNA synthetase domain containing 1
ILMN_1759991	MGC3731	Nucleolar protein 12
ILMN_1733757	LOC374395	Transmembrane protein 179B

Continued on next page

Table D1 – continued from previous page

ProbeID	Gene	Gene Description
ILMN_1753790	ZNF259	Zinc finger protein 259
ILMN_1746206	AZI1	5-azacytidine induced 1
ILMN_1763663	FLJ20718	HEAT repeat containing 3
ILMN_1655819	LOC728919	Anaphase promoting complex subunit 11
ILMN_1686401	LOC728739	Programmed cell death 2
ILMN_1669424	LOC646531	Y box binding protein 1
ILMN_2179726	LOC90835	Chromosome 16 open reading frame 93

Appendix E

Appendix E contains (1) the results obtained for the univariate Allele-Independent Intercept Model and Allele-Independent Non-Intercept Model fitted using the `MCMCglmm()` function (2) the results obtained for the univariate Allele-Dependent Intercept Model and Allele-Dependent Non-Intercept Model fitted using the `MCMCglmm()` function and (3) the results obtained for the multivariate Allele-Dependent Intercept Model and Allele-Dependent Non-Intercept Model fitted using the `MCMCglmm()` function.

Table E1: Univariate Allele-Independent Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	Direction	n	Beta	P
(Intercept)	TS	16	1.260×10^{2}	9.644×10^{-1}
CategoryAmp/Del	TS	16	6.496×10^{4}	2.222×10^{-4}
CategoryDel/Neut	TS	47	2.061×10^{4}	2.222×10^{-4}
CategoryNeut/Amp	TS	16	-1.659×10^{2}	9.836×10^{-1}
CategoryNeut/Del	TS	31	-1.134×10^{2}	9.702×10^{-1}
(Intercept)	TE	16	1.363×10^{2}	9.409×10^{-1}
CategoryAmp/Del	TE	16	2.589×10^{4}	2.222×10^{-4}
CategoryDel/Neut	TE	47	-1.562×10^{2}	9.600×10^{-1}
CategoryNeut/Amp	TE	16	6.497×10^{4}	2.222×10^{-4}
CategoryNeut/Del	TE	31	1.789×10^{4}	2.222×10^{-4}

(B) Parameter Estimates

Coefficients	Direction	n	Beta	P
(Intercept)	TS	16	5.458×10^{2}	9.049×10^{-1}
CategoryAmp/Del	TS	16	6.427×10^{4}	2.222×10^{-4}
CategoryDel/Neut	TS	47	2.014×10^{4}	1.333×10^{-3}
CategoryNeut/Amp	TS	16	3.631×10^{4}	2.222×10^{-4}
CategoryNeut/Del	TS	31	3.159×10^{4}	2.222×10^{-4}
(Intercept)	TE	16	1.198×10^{3}	8.920×10^{-1}
CategoryAmp/Del	TE	16	2.464×10^{4}	4.000×10^{-2}
CategoryDel/Neut	TE	47	6.908×10^{4}	2.222×10^{-4}
CategoryNeut/Amp	TE	16	6.336×10^{4}	2.222×10^{-4}
CategoryNeut/Del	TE	31	1.676×10^{4}	1.129×10^{-1}

Table E2: Univariate Allele-Independent Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	1.260×10^{2}	-5.886×10^{3}	6.703×10^{3}
Amp/Del	16	TS	6.509×10^{4}	5.886×10^{4}	7.128×10^{4}
NA	47	TS	2.073×10^{4}	1.719×10^{4}	2.466×10^{4}
NA	16	TS	-3.983×10^{1}	-6.358×10^{3}	6.269×10^{3}
NA	31	TS	1.267×10^{1}	-4.499×10^{3}	4.751×10^{3}
NoChangepoint	16	TE	1.363×10^{2}	-5.955×10^{3}	6.546×10^{3}
Amp/Del	16	TE	2.603×10^{4}	1.973×10^{4}	3.244×10^{4}
NA	47	TE	-1.982×10^{1}	-3.518×10^{3}	3.806×10^{3}
NA	16	TE	6.511×10^{4}	5.904×10^{4}	7.150×10^{4}
NA	31	TE	1.803×10^{4}	1.348×10^{4}	2.248×10^{4}

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	5.458×10^{2}	-9.992×10^{3}	1.065×10^{4}
Amp/Del	16	TS	6.482×10^{4}	5.418×10^{4}	7.454×10^{4}
NA	47	TS	2.069×10^{4}	1.467×10^{4}	2.632×10^{4}
NA	16	TS	3.686×10^{4}	2.638×10^{4}	4.707×10^{4}
NA	31	TS	3.213×10^{4}	2.448×10^{4}	3.932×10^{4}
NoChangepoint	16	TE	1.198×10^{3}	-1.582×10^{4}	1.769×10^{4}
Amp/Del	16	TE	2.584×10^{4}	8.035×10^{3}	4.200×10^{4}
NA	47	TE	7.027×10^{4}	6.030×10^{4}	7.984×10^{4}
NA	16	TE	6.456×10^{4}	4.804×10^{4}	8.158×10^{4}
NA	31	TE	1.796×10^{4}	5.832×10^{3}	3.009×10^{4}

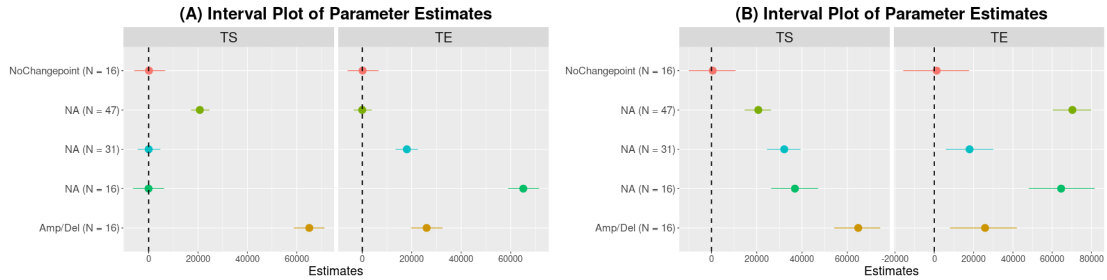


Figure E1: Interval plot of univariate Allele-Independent Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Table E3: Univariate Allele-Independent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	Direction	n	Beta	P
CategoryAmp/Del	TS	16	6.504×10^{4}	2.222×10^{-4}
CategoryDel/Neut	TS	47	2.066×10^{4}	2.222×10^{-4}
CategoryNeut/Amp	TS	16	-8.122×10^{1}	9.844×10^{-1}
CategoryNeut/Del	TS	31	-1.825×10^{1}	9.831×10^{-1}
CategoryAmp/Del	TE	16	2.602×10^{4}	2.222×10^{-4}
CategoryDel/Neut	TE	47	4.771×10^{1}	9.916×10^{-1}
CategoryNeut/Amp	TE	16	6.504×10^{4}	2.222×10^{-4}
CategoryNeut/Del	TE	31	1.803×10^{4}	2.222×10^{-4}

(B) Parameter Estimates

Coefficients	Direction	n	Beta	P
CategoryAmp/Del	TS	16	6.507×10^{4}	2.222×10^{-4}
CategoryDel/Neut	TS	47	2.072×10^{4}	2.222×10^{-4}
CategoryNeut/Amp	TS	16	3.671×10^{4}	2.222×10^{-4}
CategoryNeut/Del	TS	31	3.212×10^{4}	2.222×10^{-4}
CategoryAmp/Del	TE	16	2.564×10^{4}	6.667×10^{-3}
CategoryDel/Neut	TE	47	7.025×10^{4}	2.222×10^{-4}
CategoryNeut/Amp	TE	16	6.446×10^{4}	2.222×10^{-4}
CategoryNeut/Del	TE	31	1.803×10^{4}	8.000×10^{-3}

Table E4: Univariate Allele-Independent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates and Confidence Intervals					
Category	n	Direction	Fit	LB	UB
Amp/Del	16	TS	6.504×10^4	5.845×10^4	7.193×10^4
Del/Neut	47	TS	2.066×10^4	1.675×10^4	2.473×10^4
Neut/Amp	16	TS	-8.122×10^1	-6.912×10^3	7.011×10^3
Neut/Del	31	TS	-1.825×10^1	-5.097×10^3	4.822×10^3
Amp/Del	16	TE	2.602×10^4	1.917×10^4	3.274×10^4
Del/Neut	47	TE	4.771×10^1	-3.848×10^3	3.868×10^3
Neut/Amp	16	TE	6.504×10^4	5.851×10^4	7.177×10^4
Neut/Del	31	TE	1.803×10^4	1.287×10^4	2.257×10^4

(B) Parameter Estimates and Confidence Intervals					
Category	n	Direction	Fit	LB	UB
Amp/Del	16	TS	6.507×10^4	5.420×10^4	7.658×10^4
Del/Neut	47	TS	2.072×10^4	1.468×10^4	2.758×10^4
Neut/Amp	16	TS	3.671×10^4	2.606×10^4	4.787×10^4
Neut/Del	31	TS	3.212×10^4	2.372×10^4	4.012×10^4
Amp/Del	16	TE	2.564×10^4	6.767×10^3	4.197×10^4
Del/Neut	47	TE	7.025×10^4	5.988×10^4	8.110×10^4
Neut/Amp	16	TE	6.446×10^4	4.712×10^4	8.269×10^4
Neut/Del	31	TE	1.803×10^4	4.417×10^3	2.999×10^4

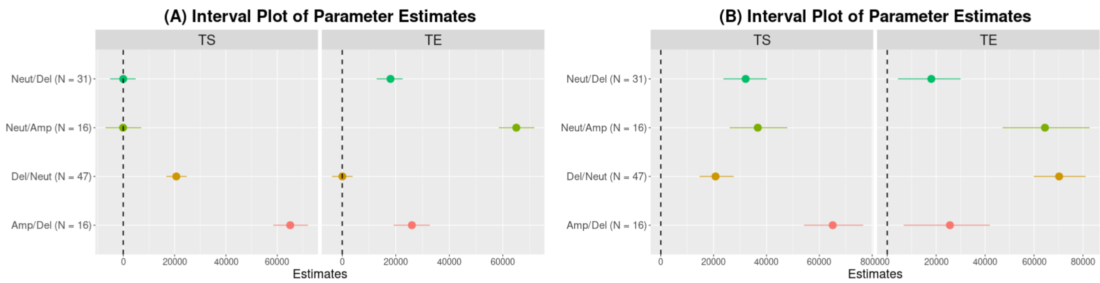


Figure E2: Interval plot of univariate Allele-Independent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Table E5: Multivariate Allele-Independent Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates			
Coefficients	n	Beta	P
(Intercept)	16	1.803×10^4	2.222×10^{-4}
traitTS:CategoryNoChangepoint	16	-1.802×10^4	2.222×10^{-4}
traitTE:CategoryNoChangepoint	47	-1.795×10^4	2.222×10^{-4}
traitTS:CategoryAmp/Del	16	4.698×10^4	2.222×10^{-4}
traitTE:CategoryAmp/Del	31	7.860×10^3	4.356×10^{-2}
traitTS:CategoryDel/Neut	16	2.628×10^3	3.804×10^{-1}
traitTE:CategoryDel/Neut	16	-1.805×10^4	2.222×10^{-4}
traitTS:CategoryNeut/Amp	47	-1.798×10^4	2.222×10^{-4}
traitTE:CategoryNeut/Amp	16	4.707×10^4	2.222×10^{-4}
traitTS:CategoryNeut/Del	31	-1.800×10^4	2.222×10^{-4}

(B) Parameter Estimates			
Coefficients	n	Beta	P
(Intercept)	16	1.840×10^4	2.667×10^{-3}
traitTS:CategoryNoChangepoint	16	-1.842×10^4	2.178×10^{-2}
traitTE:CategoryNoChangepoint	47	-1.855×10^4	6.978×10^{-2}
traitTS:CategoryAmp/Del	16	4.647×10^4	2.222×10^{-4}
traitTE:CategoryAmp/Del	31	7.512×10^3	4.738×10^{-1}
traitTS:CategoryDel/Neut	16	2.306×10^3	7.253×10^{-1}
traitTE:CategoryDel/Neut	16	5.184×10^4	2.222×10^{-4}
traitTS:CategoryNeut/Amp	47	1.851×10^4	2.356×10^{-2}
traitTE:CategoryNeut/Amp	16	4.654×10^4	2.222×10^{-4}
traitTS:CategoryNeut/Del	31	1.374×10^4	4.711×10^{-2}

Table E6: Multivariate Allele-Independent Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	1.244×10^{1}	-6.566×10^{3}	6.360×10^{3}
NA	16	TE	5.372×10^{1}	-6.266×10^{3}	6.529×10^{3}
Amp/Del	16	TS	6.501×10^{4}	5.860×10^{4}	7.139×10^{4}
NA	47	TE	2.066×10^{4}	1.678×10^{4}	2.433×10^{4}
NA	31	TS	3.493×10^{1}	-4.639×10^{3}	4.563×10^{3}
NoChangepoint	16	TE	8.056×10^{1}	-6.333×10^{3}	6.162×10^{3}
NA	16	TS	6.510×10^{4}	5.896×10^{4}	7.156×10^{4}
Amp/Del	16	TE	2.589×10^{4}	1.964×10^{4}	3.229×10^{4}
NA	47	TS	-1.397×10^{1}	-3.451×10^{3}	3.766×10^{3}
NA	31	TE	1.803×10^{4}	1.349×10^{4}	2.250×10^{4}

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NoChangepoint	16	TS	-2.014×10^{1}	-1.018×10^{4}	1.054×10^{4}
NA	16	TE	3.691×10^{4}	2.587×10^{4}	4.704×10^{4}
Amp/Del	16	TS	6.488×10^{4}	5.431×10^{4}	7.551×10^{4}
NA	47	TE	2.071×10^{4}	1.471×10^{4}	2.661×10^{4}
NA	31	TS	3.214×10^{4}	2.455×10^{4}	3.927×10^{4}
NoChangepoint	16	TE	-1.435×10^{2}	-1.658×10^{4}	1.681×10^{4}
NA	16	TS	6.494×10^{4}	4.758×10^{4}	8.145×10^{4}
Amp/Del	16	TE	2.592×10^{4}	8.986×10^{3}	4.282×10^{4}
NA	47	TS	7.024×10^{4}	6.009×10^{4}	8.019×10^{4}
NA	31	TE	1.840×10^{4}	6.098×10^{3}	3.028×10^{4}

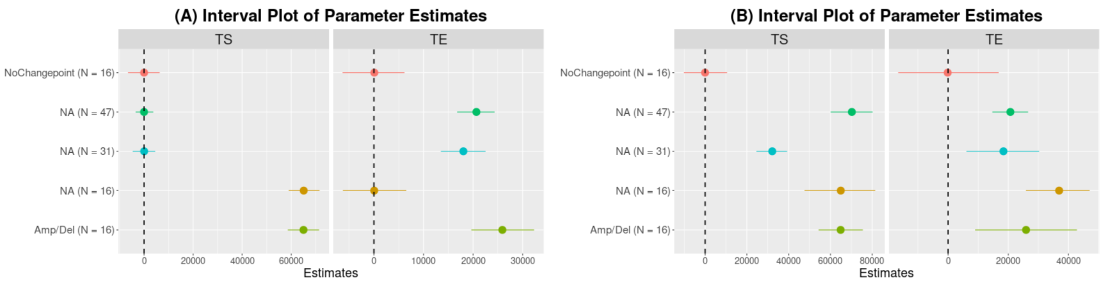


Figure E3: Interval plot of multivariate Allele-Independent Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Table E7: Multivariate Allele-Independent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates

Coefficients	n	Beta	P
traitTS:CategoryAmp/Del	16	6.510×10^{4}	2.222×10^{-4}
traitTE:CategoryAmp/Del	47	2.592×10^{4}	2.222×10^{-4}
traitTS:CategoryDel/Neut	16	2.073×10^{4}	2.222×10^{-4}
traitTE:CategoryDel/Neut	31	-2.382×10^{1}	9.978×10^{-1}
traitTS:CategoryNeut/Amp	16	1.201×10^{2}	9.840×10^{-1}
traitTE:CategoryNeut/Amp	47	6.506×10^{4}	2.222×10^{-4}
traitTS:CategoryNeut/Del	16	-8.614	9.889×10^{-1}
traitTE:CategoryNeut/Del	31	1.809×10^{4}	2.222×10^{-4}

(B) Parameter Estimates

Coefficients	n	Beta	P
traitTS:CategoryAmp/Del	16	6.493×10^{4}	2.222×10^{-4}
traitTE:CategoryAmp/Del	47	2.577×10^{4}	6.222×10^{-3}
traitTS:CategoryDel/Neut	16	2.065×10^{4}	2.222×10^{-4}
traitTE:CategoryDel/Neut	31	7.025×10^{4}	2.222×10^{-4}
traitTS:CategoryNeut/Amp	16	3.680×10^{4}	2.222×10^{-4}
traitTE:CategoryNeut/Amp	47	6.447×10^{4}	2.222×10^{-4}
traitTS:CategoryNeut/Del	16	3.218×10^{4}	2.222×10^{-4}
traitTE:CategoryNeut/Del	31	1.815×10^{4}	5.333×10^{-3}

Table E8: Multivariate Allele-Independent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

(A) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NA	16	TS	1.201×10^{2}	-7.027×10^{3}	6.776×10^{3}
Amp/Del	16	TS	6.510×10^{4}	5.830×10^{4}	7.201×10^{4}
NA	47	TS	2.073×10^{4}	1.663×10^{4}	2.461×10^{4}
NA	31	TS	-8.614	-4.967×10^{3}	4.868×10^{3}
NA	16	TE	6.506×10^{4}	5.821×10^{4}	7.153×10^{4}
Amp/Del	16	TE	2.592×10^{4}	1.955×10^{4}	3.269×10^{4}
NA	47	TE	-2.382×10^{1}	-3.797×10^{3}	3.989×10^{3}
NA	31	TE	1.809×10^{4}	1.313×10^{4}	2.293×10^{4}

(B) Parameter Estimates and Confidence Intervals

Category	n	Direction	Fit	LB	UB
NA	16	TS	3.680×10^{4}	2.585×10^{4}	4.841×10^{4}
Amp/Del	16	TS	6.493×10^{4}	5.393×10^{4}	7.641×10^{4}
NA	47	TS	2.065×10^{4}	1.447×10^{4}	2.749×10^{4}
NA	31	TS	3.218×10^{4}	2.441×10^{4}	4.045×10^{4}
NA	16	TE	6.447×10^{4}	4.647×10^{4}	8.200×10^{4}
Amp/Del	16	TE	2.577×10^{4}	7.431×10^{3}	4.374×10^{4}
NA	47	TE	7.025×10^{4}	5.977×10^{4}	8.106×10^{4}
NA	31	TE	1.815×10^{4}	5.017×10^{3}	3.064×10^{4}

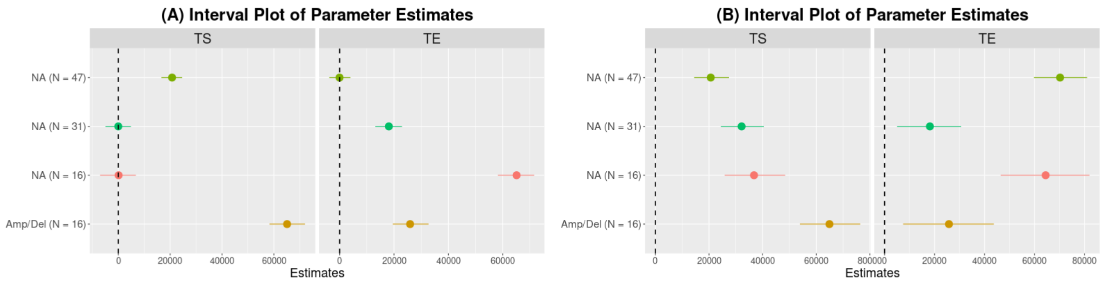


Figure E4: Interval plot of multivariate Allele-Independent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`. In (A) neutral lengths are recorded as length 0 and in (B) neutral lengths are retained as greater than 0.

Table E9: Univariate Allele-Dependent Intercept Model parameter estimates and confidence intervals fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Parameter Estimates and Confidence Intervals

Coefficients	Allele	Direction	n	Beta	P	Category	Fit	LB	UB
NoChangepoint	Major	TS	4	4.559×10^{2}	9.502×10^{-1}	NoChangepoint	4.559×10^{2}	-1.189×10^{4}	1.345×10^{4}
NoChangepoint	Minor	TS	12	6.455×10^{4}	2.222×10^{-4}	NoChangepoint	-5.273×10^{1}	-7.835×10^{3}	7.253×10^{3}
Amp/Del	Major	TS	16	2.551×10^{4}	1.333×10^{-3}	Amp/Del	6.501×10^{4}	5.865×10^{4}	7.101×10^{4}
Del/Neut	Major	TS	16	-4.897×10^{2}	9.484×10^{-1}	Del/Neut	2.597×10^{4}	1.937×10^{4}	3.230×10^{4}
Del/Neut	Minor	TS	31	3.306×10^{1}	9.898×10^{-1}	Del/Neut	1.807×10^{4}	1.350×10^{4}	2.249×10^{4}
Neut/Amp	Major	TS	16	-5.086×10^{2}	9.444×10^{-1}	Neut/Amp	-3.380×10^{1}	-6.021×10^{3}	6.512×10^{3}
Neut/Del	Minor	TS	31	-7.393×10^{3}	3.813×10^{-1}	Neut/Del	-1.967×10^{1}	-4.483×10^{3}	4.422×10^{3}
NoChangepoint	Major	TE	4	3.588×10^{2}	9.524×10^{-1}	NoChangepoint	3.588×10^{2}	-1.271×10^{4}	1.228×10^{4}
NoChangepoint	Minor	TE	12	2.555×10^{4}	2.222×10^{-4}	NoChangepoint	3.523×10^{1}	-7.474×10^{3}	7.272×10^{3}
Amp/Del	Major	TE	16	-3.823×10^{2}	9.391×10^{-1}	Amp/Del	2.591×10^{4}	1.979×10^{4}	3.220×10^{4}
Del/Neut	Major	TE	16	6.470×10^{4}	2.222×10^{-4}	Del/Neut	-2.349×10^{1}	-6.325×10^{3}	6.359×10^{3}
Del/Neut	Minor	TE	31	1.801×10^{4}	2.222×10^{-4}	Del/Neut	-1.022×10^{1}	-4.383×10^{3}	4.533×10^{3}
Neut/Amp	Major	TE	16	-3.235×10^{2}	9.564×10^{-1}	Neut/Amp	6.506×10^{4}	5.875×10^{4}	7.117×10^{4}
Neut/Del	Minor	TE	31	3.368×10^{2}	9.600×10^{-1}	Neut/Del	1.805×10^{4}	1.372×10^{4}	2.269×10^{4}

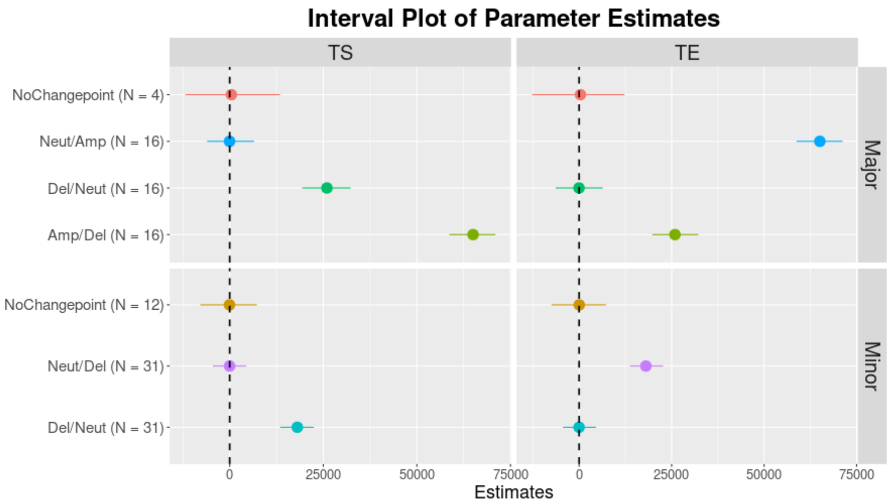


Figure E5: Interval plot of univariate Allele-Dependent Intercept Model parameter estimates fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Table E10: Univariate Allele-Dependent Non-Intercept Model parameter estimates and confidence intervals fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Parameter Estimates and Confidence Intervals									
Coefficients	**Allele**	**Direction**	**n**	**Beta**	**P**	**Category**	**Fit**	**LB**	**UB**
Amp/Del	Major	TS	16	6.509×10^{4}	2.222×10^{-4}	Amp/Del	6.509×10^{4}	5.819×10^{4}	7.181×10^{4}
Del/Neut	Major	TS	16	2.580×10^{4}	2.222×10^{-4}	Del/Neut	2.580×10^{4}	1.902×10^{4}	3.257×10^{4}
Del/Neut	Minor	TS	31	1.262	9.920×10^{-1}	Del/Neut	1.806×10^{4}	1.335×10^{4}	2.326×10^{4}
Neut/Amp	Major	TS	16	7.699×10^{3}	1.262×10^{-1}	Neut/Amp	1.262	-7.016×10^{3}	6.512×10^{3}
Neut/Del	Minor	TS	31	-7.735×10^{3}	6.933×10^{-2}	Neut/Del	-3.684×10^{1}	-4.937×10^{3}	4.859×10^{3}
Amp/Del	Major	TE	16	2.590×10^{4}	2.222×10^{-4}	Amp/Del	2.590×10^{4}	1.896×10^{4}	3.231×10^{4}
Del/Neut	Major	TE	16	-3.338×10^{1}	9.867×10^{-1}	Del/Neut	-3.338×10^{1}	-6.750×10^{3}	6.495×10^{3}
Del/Neut	Minor	TE	31	6.505×10^{4}	2.222×10^{-4}	Del/Neut	2.901×10^{1}	-4.822×10^{3}	4.968×10^{3}
Neut/Amp	Major	TE	16	1.795×10^{4}	4.444×10^{-4}	Neut/Amp	6.505×10^{4}	5.861×10^{4}	7.165×10^{4}
Neut/Del	Minor	TE	31	6.239×10^{1}	9.844×10^{-1}	Neut/Del	1.802×10^{4}	1.330×10^{4}	2.293×10^{4}

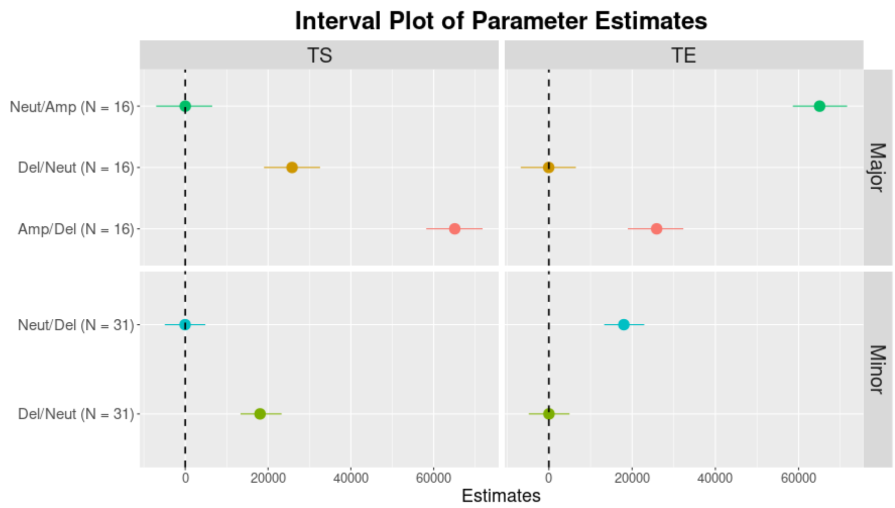


Figure E6: Interval plot of univariate Allele-Dependent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Table E11: Multivariate Allele-Dependent Intercept Model parameter estimates and confidence intervals fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Parameter Estimates and Confidence Intervals

Coefficients	Allele	n	Beta	P	Category	Fit	LB	UB
(Intercept)	NA	4	1.810×10^{4}	1.911×10^{-2}	NoChangepoint	2.169×10^{2}	-1.217×10^{4}	1.291×10^{4}
traitTS:CategoryNoChangepoint	NA	12	-1.788×10^{4}	6.800×10^{-2}	NoChangepoint	-1.719×10^{1}	-7.249×10^{3}	7.289×10^{3}
traitTE:CategoryNoChangepoint	NA	16	-1.794×10^{4}	2.222×10^{-4}	Neut/Amp	3.049×10^{1}	-6.514×10^{3}	6.550×10^{3}
traitTS:CategoryAmp/Del	NA	16	4.702×10^{4}	2.222×10^{-4}	Amp/Del	6.512×10^{4}	5.859×10^{4}	7.142×10^{4}
traitTE:CategoryAmp/Del	NA	31	7.872×10^{3}	3.493×10^{-1}	Del/Neut	2.587×10^{4}	1.937×10^{4}	3.213×10^{4}
traitTS:CategoryDel/Neut	NA	16	7.771×10^{3}	3.400×10^{-1}	Neut/Del	-2.010×10^{1}	-4.597×10^{3}	4.521×10^{3}
traitTE:CategoryDel/Neut	NA	31	-1.802×10^{4}	3.778×10^{-2}	Del/Neut	1.799×10^{4}	1.317×10^{4}	2.234×10^{4}
traitTS:CategoryNeut/Amp	NA	4	-1.807×10^{4}	2.978×10^{-2}	NoChangepoint	1.524×10^{2}	-1.267×10^{4}	1.255×10^{4}
traitTE:CategoryNeut/Amp	NA	12	4.695×10^{4}	2.222×10^{-4}	NoChangepoint	7.559×10^{1}	-7.592×10^{3}	7.314×10^{3}
traitTS:CategoryNeut/Del	NA	16	-1.788×10^{4}	9.200×10^{-2}	Neut/Amp	6.505×10^{4}	5.887×10^{4}	7.163×10^{4}
traitTS:AlleleMinor	NA	16	-2.341×10^{2}	9.653×10^{-1}	Amp/Del	2.597×10^{4}	1.972×10^{4}	3.248×10^{4}
traitTE:AlleleMinor	NA	31	-7.684×10^{1}	9.800×10^{-1}	Del/Neut	7.285×10^{1}	-6.002×10^{3}	6.846×10^{3}
traitTS:CategoryDel/Neut:AlleleMinor	NA	16	-7.648×10^{3}	3.542×10^{-1}	Neut/Del	1.802×10^{4}	1.370×10^{4}	2.262×10^{4}
traitTE:CategoryDel/Neut:AlleleMinor	NA	31	1.046×10^{1}	9.853×10^{-1}	Del/Neut	6.476	-4.636×10^{3}	4.699×10^{3}

Figure E7: Interval plot of multivariate Allele-Dependent Intercept Model parameter estimates fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Table E12: Multivariate Allele-Dependent Non-Intercept Model parameter estimates and confidence intervals fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Parameter Estimates and Confidence Intervals								
Coefficients	**Allele**	**n**	**Beta**	**P**	**Category**	**Fit**	**LB**	**UB**
traitTS:CategoryAmp/Del	NA	16	6.502×10^{4}	2.222×10^{-4}	Neut/Amp	-2.880	-6.639×10^{3}	6.616×10^{3}
traitTE:CategoryAmp/Del	NA	16	2.585×10^{4}	2.222×10^{-4}	Amp/Del	6.502×10^{4}	5.823×10^{4}	7.186×10^{4}
traitTS:CategoryDel/Neut	NA	31	2.582×10^{4}	2.222×10^{-4}	Del/Neut	2.582×10^{4}	1.874×10^{4}	3.241×10^{4}
traitTE:CategoryDel/Neut	NA	16	8.955×10^{1}	9.760×10^{-1}	Neut/Del	-5.230×10^{1}	-4.557×10^{3}	5.229×10^{3}
traitTS:CategoryNeut/Amp	NA	31	-2.880	9.978×10^{-1}	Del/Neut	1.801×10^{4}	1.318×10^{4}	2.299×10^{4}
traitTE:CategoryNeut/Amp	NA	16	6.504×10^{4}	2.222×10^{-4}	Neut/Amp	6.504×10^{4}	5.772×10^{4}	7.168×10^{4}
traitTS:CategoryNeut/Del	NA	16	7.749×10^{3}	1.182×10^{-1}	Amp/Del	2.585×10^{4}	1.859×10^{4}	3.224×10^{4}
traitTE:CategoryNeut/Del	NA	31	1.810×10^{4}	2.222×10^{-4}	Del/Neut	8.955×10^{1}	-6.553×10^{3}	6.707×10^{3}
traitTS:AlleleMinor	NA	16	-7.801×10^{3}	7.289×10^{-2}	Neut/Del	1.806×10^{4}	1.315×10^{4}	2.264×10^{4}
traitTE:AlleleMinor	NA	31	-3.897×10^{1}	9.769×10^{-1}	Del/Neut	5.058×10^{1}	-4.709×10^{3}	5.163×10^{3}

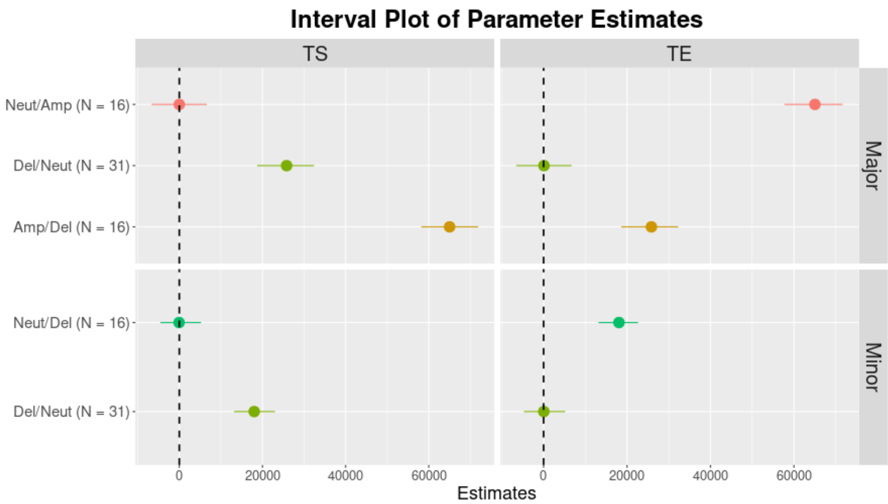


Figure E8: Interval plot of multivariate Allele-Dependent Non-Intercept Model parameter estimates fitted using `MCMCglmm()`, where neutral lengths are recorded as length 0.

Appendix F

Appendix F contains allele-specific heatmaps produced for chromosome 18q and 11p and a plot of the frequency of changepoints in genes across chromosome 18q.

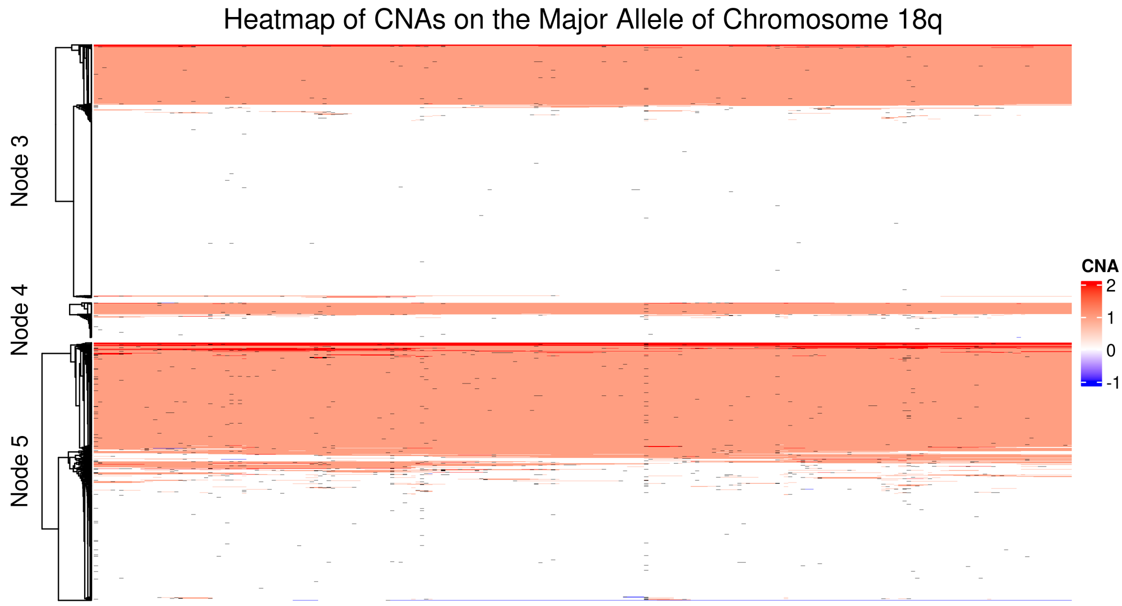


Figure F1: Heatmap of CNAs across the Major Allele of Chromosome 18q. The heatmap depicts the CNA state for each gene across Chromosome 18q, partitioning the patients into the nodes corresponding to Figure 3.37. NAs, depicting multiple states, are coloured in black.

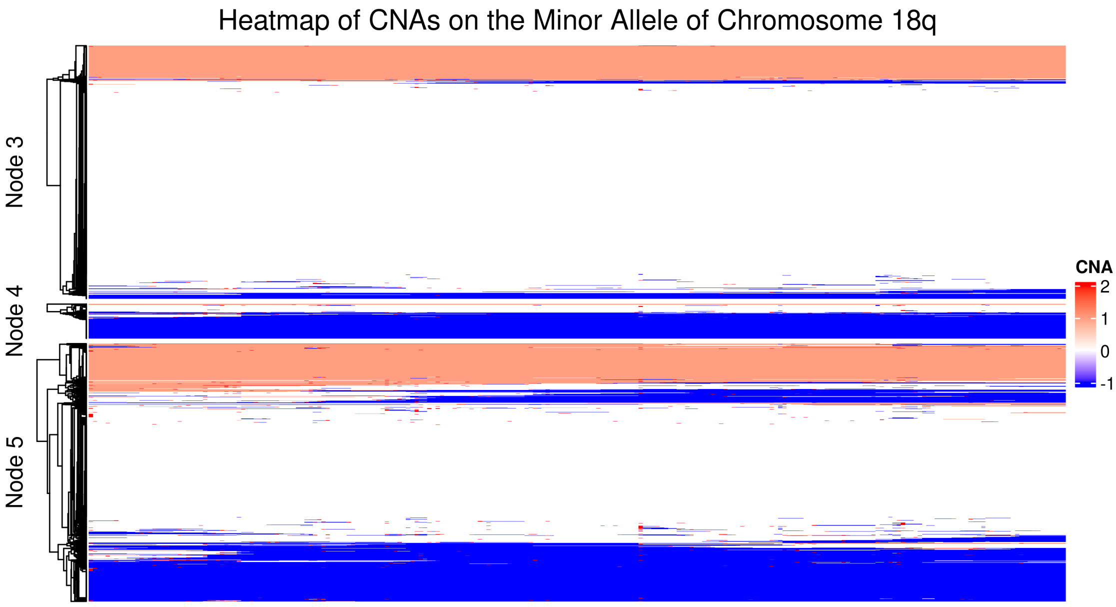


Figure F2: Heatmap of CNAs across the Minor Allele of Chromosome 18q. The heatmap depicts the CNA state for each gene across Chromosome 18q, partitioning the patients into the nodes corresponding to Figure 3.37. NAs, depicting multiple states, are coloured in black.

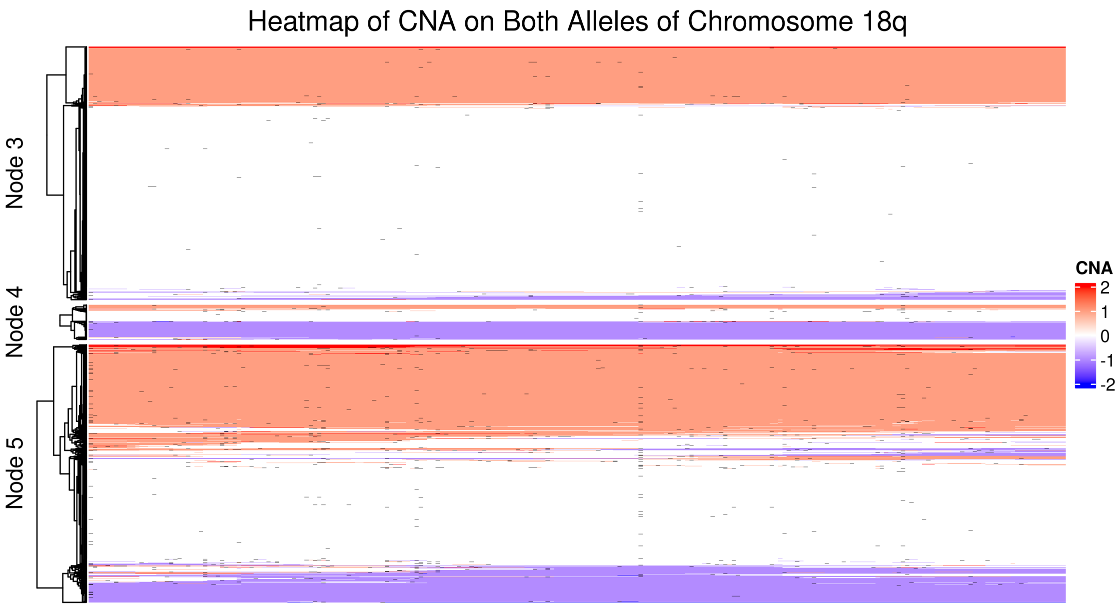


Figure F3: Heatmap of CNAs across both alleles of Chromosome 18q. The heatmap depicts the CNA state for each gene across Chromosome 18q, partitioning the patients into the nodes corresponding to Figure 3.37. NAs, depicting multiple states, are coloured in black.

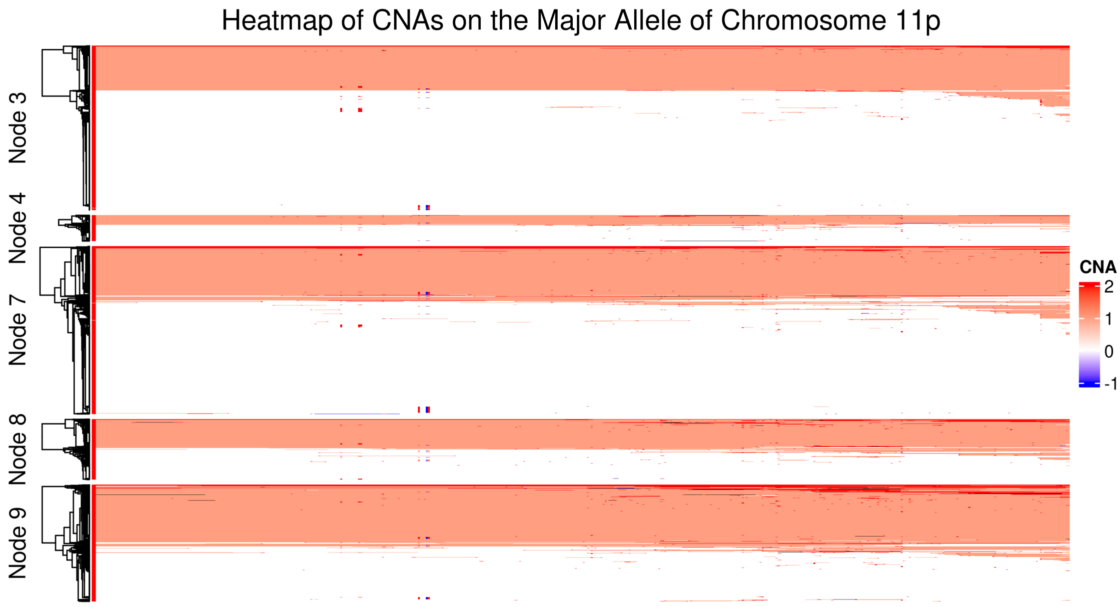


Figure F4: Heatmap of CNAs across the Major Allele of Chromosome 11p. The heatmap depicts the CNA state for each gene across Chromosome 11p, partitioning the patients into the nodes corresponding to Figure 3.32. NAs, depicting multiple states, are coloured in black.

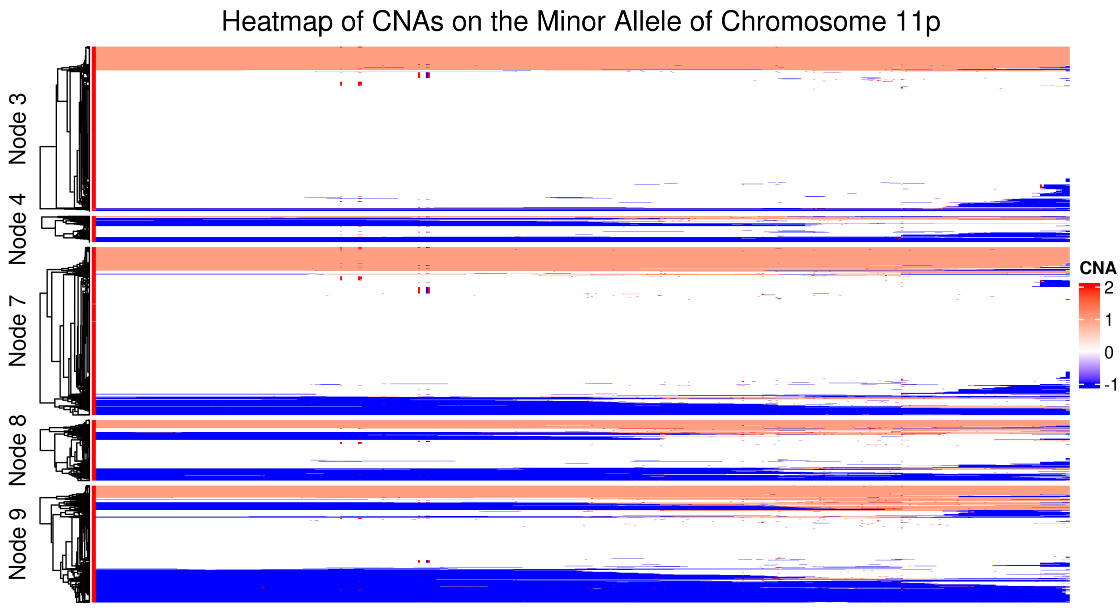


Figure F5: Heatmap of CNAs across the Minor Allele of Chromosome 11p. The heatmap depicts the CNA state for each gene across Chromosome 11p, partitioning the patients into the nodes corresponding to Figure 3.32. NAs, depicting multiple states, are coloured in black.

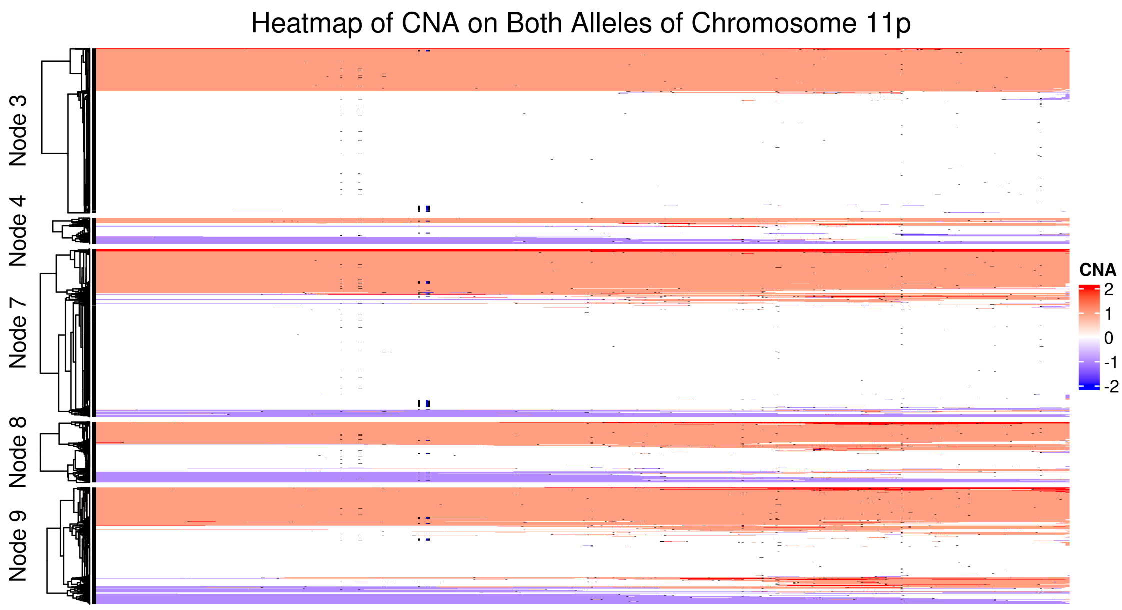


Figure F6: Heatmap of CNAs across both alleles of Chromosome 11p. The heatmap depicts the CNA state for each gene across Chromosome 11p, partitioning the patients into the nodes corresponding to Figure 3.32. NAs, depicting multiple states, are coloured in black.

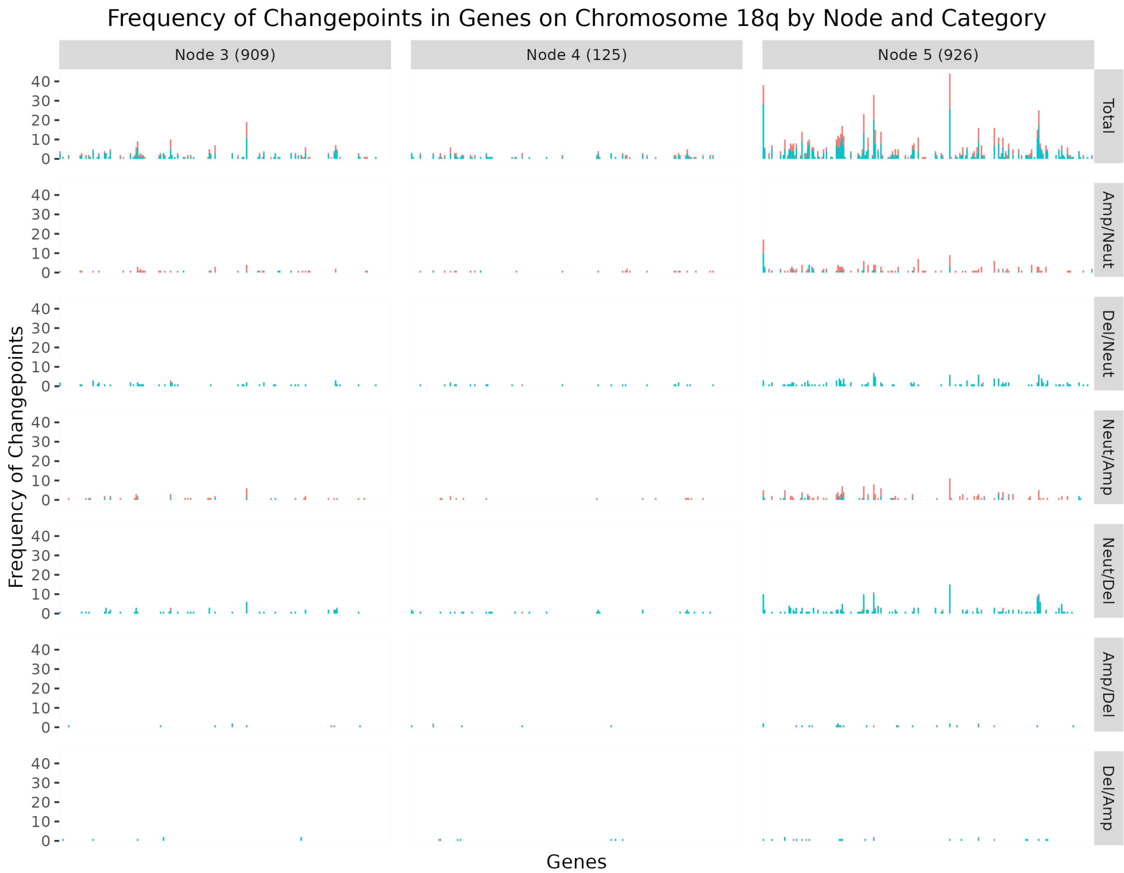


Figure F7: Frequency of changepoints in genes across chromosome 18q, split by Node and Category, and coloured by allele.