Development of Novel Methods to Study Chromosomal Instability and Genetic Heterogeneity for Biomarker Identification

by

Noshad Hosseini

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2025

Doctoral Committee:

Assistant Professor Marcin Cieslik, Chair
Professor Veera Baladandayuthapani
Professor Arul Chinnaiyan
Associate Professor Analisa Difeo
Professor Ryan Mills

Noshad Hosseini

noshadh@umich.edu

ORCID iD: 0000-0003-3761-0198

© Noshad Hosseini 2025

Acknowledgements

First and foremost, I want to thank Dr. Marcin Cieslik for his guidance, support and trust he dedicated to me and his patience throughout my entire Ph.D. journey. His professionalism, thoroughness and expertise in the field inspired me to improve my work and it will inspire me for the rest of my life. I also want to thank my lab mates who made my experience in Cieslik lab enjoyable.

Next, I'd like to thank the members of my dissertation committee, Dr. Arul Chinnaiyan, Dr. Ryan Mills, Dr. Veera Baladandayuthapani and Dr. Analisa Difeo. They have provided me with valuable comments to improve my work, and to continue my journey in academia.

I would also like to thank members of Michigan Center for Translational Pathology (MCTP), specially Dr. Arul Chinnaiyan, Dr. Mohan Saravana Dhanasekaran and Dr. Abhijit Parolia, who trusted me and included me in their projects and constantly provided me with their advice.

I would like to thank my wife, who has patiently supported me in difficult times and encouraged me not to give up. I am also deeply grateful to my parents, who made countless sacrifices to help me settle in a foreign country and whose encouragement has been the foundation of my success, no matter the path I chose to follow.

I am also grateful to my brother and sister-in-law who always had my back and never let me feel alone in this path. My brother, though miles away, has always shared his experience, knowledge and support, paving the way for my achievements and making my journey possible.

I would like to express my appreciation to Dr. Margit Burmeister for her support and encouragement throughout my PhD journey. Her kindness and understanding during challenging times allowed me to stay focused on my research, and her help has been invaluable.

Table of Contents

Acknowledgements ii

List of Figures ix

Abstract xix

Chapter 1 Introduction 1

1.1 Motivation 1

1.2 Genomics of Cancer 1

1.2.1 Cancer initiation and progression 1

1.3 Detection of alterations in tumors 2

1.3.1 Alignment 3

1.3.2 Sequence altering variations detection 3

1.3.3 Detection of Copy number variations 4

1.4 Chromosomal instability in cancer 5

1.4.1 Copy number variations 5

1.4.2 Chromosomal Instability 6

1.5 Genomic Heterogeneity 8

1.6 References 8

Chapter 2 Mathematical Modeling of Genome Contiguity and Breakage Patterns Characterize Chromosomal Instability 12

2.1 Introduction 12

2.1.1 I CPC-CPTAC LUAD cohort 13

2.2 Results 13

2.2.1 Chromosomal instability in Lung Adenocarcinoma 13

2.2.2 Prognostic role of genome contiguity 16

2.2.3 Genetics underpinning genome contiguity 19

2.2.4 Breakage Intensity Clustering - a novel measure of CIN identifies NKX2-1-dependent and -independent paths to aggressive disease 23

2.2.5 IGF2BP3 is a robust biomarker of genome fragmentation 30

2.3 Discussion 36

2.4 Methods 38

2.4.1 Identifying regions of recurrent amplification 38

2.4.2 Instability measure according to CNA segment length 38

2.4.3 Breakage Intensity-Clustering (BIC) 38

2.5 References 40

Chapter 3 Proteogenomic Analysis of Chemo-Refractory High-Grade Serous Ovarian Cancer 44

3.1 Introduction 44

3.2 Results 45

3.2.1 Study cohort and proteogenomic profiling 45

3.2.2 Genomic alterations associated with chemo-refractoriness 47

3.2.3 Chr17-LOH is associated with chemo-sensitivity 50

3.2.4 TP53 inactivation is associated with chemo-sensitivity 53

3.2.5 Genomic scars associated with chemo-response 63

3.3 Discussion 68

3.4 Methods 69

3.4.1 Patient selection and sequencing 69

3.4.2 WGS Data Processing and Alignment 70

3.4.3 WGS Data QC 71

3.4.4 WGS Data Variant/Mutation Calling 71

3.4.5 CNV analysis 73
3.4.6 Instability measures 76
3.4.7 RNA Sequencing Processing, Alignment and Quantification 77
3.4.8 RNA-sequencing Data Quality Control 78
3.4.9 TP53 Signatures 80
3.5 References 80
Chapter 4 Divergent Clonal Evolution and Early Dissemination Promotes Genetic Heterogeneity of Metastases in Castration Resistant Prostate Cancer 85
4.1 Introduction 85
4.2 Results 87
4.2.1 Multi-site sequencing of a rapid autopsy cohort of mCRPC 87
4.2.2 Sequencing a single metastatic site overestimates clonality of mutations 92
4.2.3 Most mutations in mCRPC are subclonal and restricted to a single site 96
4.2.4 Single-site sequencing underestimates patient and cohort-level mutation rates 101
4.2.5 Most CNVs occur after metastasis but pathogenic alterations are largely truncal 108
4.2.6 The majority of tumor clones are detected in only a single site 116
4.2.7 Position of driver mutations within the phylogenetic tree is highly variable 120
4.2.8 Single-site sequencing significantly underestimates clonal heterogeneity 122
4.2.9 High prevalence of polyclonal and polyphyletic metastatic seeding 125
4.2.10 Early metastatic dissemination in a case of neuroendocrine mCRPC 128
4.3 Discussion 136
4.4 Methods 141
4.4.1 Sample collection 141
4.4.2 Preparation of whole exome and RNA sequencing libraries 141
4.4.3 Comprehensive genomics analysis pipeline 142
4.4.4 Read processing and data quality control 142

4.4.5 Somatic Variant Calling 142
4.4.6 Somatic Variant Calling Quality Control 143
4.4.7 Absolute Copy Number Analysis 143
4.4.8 Whole Exome Sequencing (WES) Quality Control 145
4.4.9 RNA-seq expression quantification 145
4.4.10 Tumor mutational burden calls 145
4.4.11 Cancer Cell Fraction (CCF) calculation 146
4.4.12 Mutation classification according to presence and clonality in patients 147
4.4.13 Mutational cluster detection 148
4.4.14 Categorizing clones according to their position in phylogenetic trees 148
4.4.15 Measuring the extent of "illusion of clonality" 149
4.4.16 Estimating event detection rates 150
4.4.17 Unobserved mutation extrapolation model 150
4.4.18 Reconstruction of clonal phylogenetic trees 151
4.4.19 LOH permutation model 152
4.5 References 153
Chapter 5 MHCnvex: A Probabilistic Method For the Detection of MHC Class I and II Copy-Number Alterations 162
5.1 Introduction 162
5.2 Results 163
5.2.1 MHCnvex significantly removes the coverage noise in the MHC region 163
5.2.2 MHCnvex significantly reduces the number of segments in the MHC region 165
5.2.3 MHCnvex shows consistent results using WXS and WGS data 166
5.2.4 MHCnvex LOH calls correlate with changes in RNA expression 167
5.3 Discussion 169
5.4 Methods 169

5.4.1 Data acquisition 169
5.4.2 Overview of the algorithm 170
5.4.3 WGS analysis 174
5.4.4 Haplotyping and alignment 174
5.5 References 174
Chapter 6 Concluding Remarks 176

List of Figures

Figure 2.1: Copy number profile of 3 tumors from CPTAC-ICPC cohort. This figure highlights the CIN heterogeneity among LUAD tumors. The top figure shows a very stable tumor with few breakpoints, the middle figure highlights a very highly shattered genome with the majority of short segments, and the bottom figure depicts a tumor that is in general stable, but has been shattered locally in some regions. 14

Figure 2.2: wGII distribution of tumors in CPTAC cohort. The dotted line represents the threshold which leads to best separation of patients based on their overall 15

Figure 2.3: Overall survival of CPTAC patients grouped according to their wGII with threshold of wGII=0.35 16

Figure 2.4: Overall survival of CPTAC patients grouped according to their contiguity score. ... 17

Figure 2.5: Overall survival of APOLLO-LUAD patients grouped according to their contiguity score. 18

Figure 2.6: Overall survival of CPTAC-CCRCC patients grouped according to their contiguity score. 18

Figure 2.7: Number of detected SVs per each contiguity group, where as expected, the most fragmented group (Unstable) had significantly higher number of SVs 19

Figure 2.8: Regions of recurrent amplification in each contiguity group. As expected, a higher rate of recurrent amplification is observed in the fragmented group (Top) versus the full stable group (bottom), as well as certain loci of known oncogenes such as *TERT*, *MYC* and *NKX2-1*.. 20

Figure 2.9: Amplified segments on chromosome 5 short arm. Each horizontal line shows an amplified segment. Each row of x-axis represents one patient, and y-axis represents coordinates along chromosome 5. Dotted line shows the *TERT* loci. 21

Figure 2.10: RNA expression of TERT gene stratified by genome contiguity groups. Each dot represents TERT expression for one sample 22

Figure 2.11: TERT downstream activity stratified by genome contiguity groups. As the activity of TERT increases, the downstream activity will increase too. 22

Figure 2.12: Association between TERT expression and overall survival of LUAD patients. Higher TERT is associated with the worst outcome. 23

Figure 2.13: Factors impacting length of a segment in genome 24

Figure 2.14: Methods used to calculate the positional and intensity aspects of breakpoints, which are impacting length of segments 25

Figure 2.15: Classification of tumors according to their breakpoint intensity (x-axis) and clustering score (Y-axis) to derive BIC groups. Tumors are categorized into three groups: tumors with highly clustered breakpoints (Contiguous), tumors with low clustering and intensity of breakpoints (Fragmented), and tumors with low clustering and high intensity of breakpoints (Intense) 25

Figure 2.16: Kaplan-Meier survival curves showing the overall survival of patients, stratified by BIC. Tumors with high clustering show exceptionally good survival independent of their breakpoint intensity. In groups with low clustering, high intensity of breakpoints results in worse outcomes. 26

Figure 2.17: SV counts for each sample in BIC groups. P values are based on the Wilcoxon Rank Sum test. 27

Figure 2.18: The proportion of tumors with NKX2-1 amplification (red) and null expression (blue). Tumors with amplified NKX2-1 show significant enrichment in the BIC-intense group. 28

Figure 2.19: RNA and protein expression for each sample (represented by a point) stratified by *NKX2-1* amplification status. 28

Figure 2.20: Proliferative activity of tumors stratified by NKX2-1 amplification (left) and expression (right) status. Proliferative activity is measured by ssGSEA analysis of RNA expression for the HALLMARK_G2M_CHECKPOINT pathway. P values are based on the Wilcoxon Rank Sum test. 29

Figure 2.21: Predicted identity of LUAD tumors stratified by BIC status 29

Figure 2.22: Log2 fold changes of genes from differential expression analysis comparing BIC-contiguous to BIC-intense tumors at both RNA (x-axis) and protein (y-axis) levels, showing IGF2BP3 to be highly differentially expressed in the BIC-intense group 30

Figure 2.23: Corresponding IGF2BP3 IHC and H&E staining shows heterogeneous protein expression of IGF2BP3 in case-3 (partial stable LUAD), uniformly strong expression in case-1 (unstable), and complete abrogation of IGF2BP3 expression in case-4 (full stable). (IHC; H&E; X=200um) 31

Figure 2.24: Independent validation in 3 patient TMAs reveals higher IGF2BP3 expression in metastatic LUAD compared to primary LUAD and primary LSCC. Also noteworthy is the absence of IGF2BP3 in benign lung parenchyma cores. 32

Figure 2.25: Kaplan-Meier survival curves showing the overall survival of non-squamous patients treated with MPDL3280A stratified by IGF2BP3 RNA expression in the OAK trial. ... 33

Figure 2.26: Kaplan-Meier survival curves showing overall survival of non-squamous patients treated with docetaxel stratified by IGF2BP3 RNA expression in the OAK trial. 33

Figure 2.27: Boxplot showing PD-L1 (CD274) protein expression stratified by BIC status. 34

Figure 2.28: Scatter plot showing the correlation of IGF2BP3 protein expression and average IGF2BP3 promoter methylation 35

Figure 2.29: Association of BIC groups with smoking status (left), tumor WGD status (middle), and tumor stage (right). 35

Figure 2.30: Association of the most aggressive, BIC-intense group, with mutations in TP53 (left), EGFR (middle) and KRAS (right). Each column shows all samples in each instability group. The block size in each column shows the proportion of patients with mutated (top) and WT (bottom) status. 35

Figure 3.1: Availability of the proteomic, phospho-proteomic, WGS, and RNA-seq data across the cohorts. 46

Figure 3.2: Boxplots showing the distribution of sample storage times across three biospecimen source sites: Fred Hutchinson Cancer Research Center (FHCRC), Mayo Clinic, and University of Alabama at Birmingham (UAB). The sample sizes for the three source sites are 62 FHCRC, 50 Mayo and 46 UAB. 47

Figure 3.3: OncoPrint of likely pathogenic genetic variants in sensitive and refractory tumors. Genes are ordered by mutation frequency in sensitive tumors. Chr17-LOH binary variable indicates whether more than 25% of Chr17 has LOH. nTAI denotes the number of chromosome arms with telomeric allelic imbalance. 48

Figure 3.4: Association between mutations in BRCA1/2 and tumor response. The p value is based on the chi-square test. 48

Figure 3.5: Sample-wise plot showing copy-number aberrations in sensitive and refractory tumors. Copy-number (CN) aberrations are quantified as changes relative to the modal ploidy (baseline CN). 49

Figure 3.6: Genomic regions with recurrent focal amplifications (left) and deletions (right) in sensitive (top) and refractory (bottom) tumors. The G score is proportional to −log(probability|background). Cytobands of the most significant peaks are annotated. 50

Figure 3.7: Scatter plot showing the proportion of CNV-RNA/protein cascade genes in each pathway in sensitive (x axis) vs. refractory (y axis) tumors. Pathway size indicates the number of genes in the pathway. 51

Figure 3.8: Associations between refractoriness and chromosome losses, gains, and LOH. At the chromosome level, only Chr17-LOH is associated with refractoriness (FDR < 0.1). 52

Figure 3.9: Association between Chr17-LOH and refractoriness. Among the 15% of samples without Ch17-LOH, more than 75% are refractory, whereas most samples with Chr17-LOH are sensitive (p value = 0.0046, Fisher's Exact test). 52

Figure 3.10: MSK-IMPACT cohort HGSOC samples stratified by Chr17-LOH status. The group enriched with Chr17-LOH samples shows a better survival than the group without Chr17-LOH (p = 0.018, based on Cox regression model). 53

Figure 3.11: Distributions of genomic features and treatment response. 54

Figure 3.12: Boxplot showing the TP53 RNA abundances stratified by the mutation status of TP53 gene in the FFPE Discovery cohort. Wilcoxon rank sum tests were performed to derive the p-values. 55

Figure 3.13: Boxplot showing the TP53 protein abundances stratified by the mutation status of TP53 gene in the FFPE Discovery cohort. Wilcoxon rank sum tests were performed to derive the p-values. 55

Figure 3.14: Samples with Chr17-LOH have significantly higher p53 protein abundance (p = 0.011, Wilcoxon Rank Sum test) vs. samples without Chr17-LOH. 56

Figure 3.15: Samples with Chr17-LOH have significantly lower TP53-WT activity scores (p = 0.006, Wilcoxon Rank Sum test) vs. samples without Chr17-LOH. 57

Figure 3.16: p53 mutant activity score calculated using the RNA data, stratified by Chr17-LOH status. Samples with Chr17-LOH have significantly higher p53-mutant signature scores in the FFPE Discovery cohort (p=0.0063, based on Wilcox Rank test). 58

Figure 3.17: Boxplots showing RNA levels of MDM2, gene related to p53 pathway activity, stratified by Chr17-LOH status in the FFPE Discovery cohort. 58

Figure 3.18: Boxplots showing RNA levels of CDC20, gene related to p53 pathway activity, stratified by Chr17-LOH status in the FFPE Discovery cohort. 59

Figure 3.19: Refractory samples have higher TP53-WT activity scores (p = 0.0002, Wilcoxon Rank Sum test), vs. sensitive tumors. 60

Figure 3.20: p53 mutant activity score calculated using RNA data is significantly lower (p=0.016, based on t-test) in refractory samples compared to sensitive ones in the FFPE Discovery cohort. 60

Figure 3.21: Scatter plot showing the correlation of the RNA-derived TP53-WT activity score and p53-mutant activity score for the FFPE Discovery cohort. Colors indicate treatment responses and shapes indicate the Chr17-LOH status. The sample sizes in the different categories are 31 (R with chr17LOH Yes) + 53 (S with chr17LOH Yes), 11 (R with chr17LOH No), 1 (S with chr17LOH No), 6 (R with chr17LOH Unknown), 4 (S with chr17LOH Unknown). 61

Figure 3.22: Bar plot showing the fold change of select pathways, which were significantly associated (FDR<0.1) with Chr17-LOH status (left) and/or BRCA1/2 mutation (right) respectively, based on the protein data of the FFPE Discovery cohort. 62

Figure 3.23: Results of the multivariate logistic regression model with refractory status as the response variable and genomic features as predictor variables. 63

Figure 3.24: Results of the multivariate logistic regression model with refractory status as the response variable and genomic features as predictor variables. 64

Figure 3.25: Results of the multivariate logistic regression model with refractory status as the response variable and genomic features as predictor variables. 65

Figure 3.26: Frequency of events (gain, loss, and LOH) in the FFPE discovery cohort, where samples are stratified based on their Chr17-LOH status. While the similar pattern of gains and losses is observed in both groups, the group without Chr17-LOH have less LOH throughout the genome. 65

Figure 3.27: Forest plot of a multivariate logistic regression model predicting refractoriness based on HRD and genetic variables. 66

Figure 3.28: ROC curves showing the prediction performance of models based on (1) BRCA1/2 mutation status + clinical variables (patient age and tumor location) and (2) BRCA1/2 mutation status + clinical variables + Chr17-LOH status. 67

Figure 3.29: Boxplots showing the association of the WGS quality control metrics among tumors stratified by tumor phenotype, sample storage time, and source sites. The sample size stratified by the response is 65 S + 55 R, the sample size stratified by sample age (split by median) is 62 in new and 58 in old group, and the sample size stratified by source sites are 39 FHCRC, 36 Mayo and 45 UAB. 72

Figure 3.30: Average copy number (log ratio [LR]) of Chr17 in the FFPE discovery cohort stratified by Chr17-LOH status. Samples without Chr17-LOH have significantly higher LR values. The dotted (top) line shows the threshold (LR > 0.046) used to define a group of tumors enriched for samples without Chr17-LOH; the dashed (bottom) line shows the threshold (LR < −0.11) used to define a group of tumors enriched for samples with Chr17-LOH. These thresholds were used for the analysis of the MSK-IMPACT cohort.18 Wilcoxon rank sum test was performed to derive the p value. 76

Figure 3.31: Boxplots showing the distributions of the RNAseq quality control metrics among tumors stratified by treatment response, sample storage time, and source sites. The sample size stratified by the response is 58 S + 48 R, the sample size stratified by sample age (split by median) is 53 in each group, and the sample size stratified by source sites are 19 FHCRC, 41 Mayo and 46 UAB. 79

Figure 4.1: Overview of the cohort by number of patients and samples obtained from each tissue. 87

Figure 4.2: QC measures for RNA (top) and DNA (bottom) data..... 88

Figure 4.3: Purity (based on NGS) of tumors stratified by patient (F=3.49, Degree of freedom=25). Each dot represents one tumor. ANOVA P-value indicates that purity is significantly associated with the patient. 89

Figure 4.4: Purity (based on NGS) of tumors stratified by tissue (F=3.19, Degree of freedom=10). Each dot represents one tumor. ANOVA P-value indicates that purity is significantly associated with the tissue. 89

Figure 4.5: Alteration landscape of the cohort. Genes recurrently mutated in PCa and mutated at least once in this cohort are included. TMB indicates tumor mutation burden based on non-synonymous mutations. LOH is only shown when it is accompanied by a mutation. 90

Figure 4.6: Proportion of patients with alterations in genes recurrently altered in PCa (alterations shown in Figure 4.5). Among CNV alterations, only homozygous deletions and amplifications has been considered. 91

Figure 4.7: Proportion of mutations classified as present and clonal in all sites as a function of the number of sites evaluated..... 92

Figure 4.8: Proportion of mutations classified as present and clonal in all sites as a function of the number of sites evaluated. patients with at least four sites sequence are considered. 93

Figure 4.9: Average proportion of AR and TP53 mutations detected as present and clonal in all the evaluated sites. Only patients with at least 4 sequenced tumors are included. 94

Figure 4.10: CCF of mutations in PCa recurrently mutated genes, compared to all other detected mutations. 95

Figure 4.11: CCF of mutations in PCa recurrently mutated genes, compared to all other detected mutations. 96

Figure 4.12: Schematic representation of mutation classification used in this study. The left panel represents 3 metastatic sites from a single patient with tumor cells in each site. Mutations present in each cell are indicated by colors other than blue. The right panel lists mutations alongside their classifications according to their presence in metastatic sites and the cells within them (i.e., clonality). Each pink circle in the right panel corresponds to one of the metastatic sites. A fully hatched circle indicates a mutation that is present in all the cells within the site (i.e., clonal), a partially hatched circle signifies a mutation that is present in some but not all of the cells within the site (i.e., subclonal) and a clear circle shows the absence of any mutation. For example, the second mutation (Clonal-Shared) showed by a gray color is present in all cells of metastatic site 2, and only in cell 2 of metastatic site 3, making it present in more than one site (i.e., shared) and clonal in at least one site. 97

Figure 4.13: Illustration of mutation classification for one patient (WA_32). Private mutations appear always on the edge of all plots, while Truncal mutations are always in the middle. Shared and subclonal mutations have variable positions..... 98

Figure 4.14: The proportion of mutations in each class by each patient 99

Figure 4.15: The proportion of mutations in each class by each recurrently mutated gene in mCRPC 100

Figure 4.16: Fraction of mutations identified in each individual sample relative to all mutations for that patient in all sites, stratified by the number of sites profiled for a patient. Red dashed lines represent the average proportion of mutations detected..... 102

Figure 4.17: Concordance of mutations for all pairs of samples from each patient. Each y-axis value represents the proportion of mutations detected in both of a pair of samples (i.e., concordance). For a patient with n sites, there are n choose 2 pairs. 102

Figure 4.18: Detection rate of alterations for each gene. In this analysis, only patients with at least 4 sequenced tumors are assessed. The detection rate when evaluating four sites is set as a baseline of 1 for each gene. The plot then shows the relative detection rates when fewer than four sites are evaluated..... 103

Figure 4.19: Impact of profiling additional tumors on cohort-level observed prevalence of alteration in PCa recurrently mutated genes. Only patients with 3 sequenced tumors are assessed. 104

Figure 4.20: Schematic of the 'species accumulation model' (Methods) extrapolating the number of additional mutations that will be detected when another site is sequenced. The model predicts one step ahead, using all previously observed (solid) and predicted (semi-transparent) new mutations. 105

Figure 4.21: Validation of predictions from the model in panel (G). The x-axis shows the observed number of total mutations in four samples. Each y-axis box displays the interquartile range (IQR) of the corresponding prediction based on the data from the first three sites. 105

Figure 4.22: Results for the model in Figure 4.21. Observed (solid) and predicted (semi-transparent) number of new mutations per additional site profiled. Each curve represents an individual patient. WA_16 an MSI case is highlighted. 106

Figure 4.23: Number of private mutations stratified by patient. Each dot represents number of private mutations in one tumor 107

Figure 4.24: Proportion of mutations detected as increasing numbers of sites are evaluated. Each dot represents percent of mutations detected in a group of samples evaluated. 107

Figure 4.25: Heterogeneity in tumor ploidy across metastatic sites (top). Proportion of patients with truncal, heterogenous, or no observed WGD events (bottom) 108

Figure 4.26: Per patient and site genome-wide copy number alteration landscape. Colors indicate copy-number alterations (red - gain, blue - loss, green - LOH, gray - diploid). Cases and samples ordered by average ploidy. 109

Figure 4.27: Heterogeneity of homologous recombination deficiency (HRD) scores derived from CNV data. Large-scale state transitions (LST) scores (top), and proportion of genome harboring LOH (bottom). 110

Figure 4.28: Cohort-wide frequency of copy-number gains (red) and losses (blue). Shading indicates truncal (dark) and heterogenous observations (per-patient). Select loci recurrently altered in PCa are highlighted 111

Figure 4.29: Proportion of recurrently truncal gains (left) or losses (right) within regions of the genome stratified by alteration recurrence (Methods). 112

Figure 4.30: Proportion of patients with losses (left) and gains (right) on recurrently deleted and amplified regions in mCRPC. Shading indicates truncal (dark) and heterogenous observations (per-patient) 112

Figure 4.31: Heterogeneity of AR copy number and mutation status across metastatic sites. Presence of AR mutations indicated in (blue). Absolute copy estimates are capped at 35 113

Figure 4.32: AR locus visualization of normalized log2(tumor / normal) coverage ratios across samples from three index cases (black - diploid, red - gain, blue - loss). 114

Figure 4.33: Increase in the detection rate of copy number losses as a function of the number of sites sampled. Select loci recurrently altered in PCa are highlighted 115

Figure 4.34: Increase in the detection rate of copy number losses as a function of the number of sites sampled. Select loci recurrently altered in PCa are highlighted 115

Figure 4.35: Analytical workflow for clonal reconstruction, phylogenetic inference, and analysis of metastatic seeding patterns. Clones are identified by clustering mutations by their CCFs using PhylogicNDT. Clonal phylogenetic relationships (trees) are inferred using CONIPHER. Each clone is represented by a node with slices indicating which sites a clone is present in. Each edge indicates ancestral-descendent relationships between clones. Clones are classified into Trunk, Branch, Leaf based on their position in the tree. Phylogenetic relationships between metastatic clones are used to delineate mono- vs poly-phyletic seeding routes for each patient (tree). 116

Figure 4.36: Number of clones detected in each sample stratified by tissue site. Significance based on 117

Figure 4.37: Number of clones detected in each sample stratified by TMB, wGII and purity of the 118

Figure 4.38: Distribution of CCFs (density) within clone types stratified by position in the phylogenetic tree (Trunk, Branch, Leaf). 119

Figure 4.39: Proportion of clone types by patient. This plot excludes the clones not placed in the trees 120

Figure 4.40: Proportion of clone types by recurrently mutated gene. This plot excludes the clones not 121

Figure 4.41: Total number of detected clones as a function of the number of sites profiled. Each colored curve corresponds to an individual patient. 123

Figure 4.42: Number of patients density (Y-axis) compared to average number of leaf clones they harbor. Light green shows the predicted number and dark green shows the actual number of leaf clones. P-value is based on Wilcoxon rank sum test and indicates that the two distributions are not significantly different. 123

Figure 4.43: Proportion of clones detected as increasing numbers of sites are evaluated. 124

Figure 4.44: Proportion of clones detected as truncal (Y-axis) per different number of evaluated sites (X-axis) 125

Figure 4.45: Classification of clonal phylogenetic trees based on ancestral relationships between seeding clones. Bolded edges indicate that the descendant clone seeded a new site ... 126

Figure 4.46:Classification of clonal phylogenetic trees based on ancestral relationships between seeding clones. Bolded edges indicate that the descendant clone seeded a new site ... 126

Figure 4.47: Juxtaposition of clonal phylogenetic trees (left) and clonal migration diagrams (right). Clonal migration diagram shows a parsimonious reconstruction of evolutionary (top to bottom) and migratory (left to right) relationships. Arrows indicate evolution (solid) or migration (dashed). Each node represents a site with sections colored by clonal composition. Clone colors are matched between trees and diagrams. 127

Figure 4.48: Juxtaposition of clonal phylogenetic trees (left) and clonal migration diagrams (right). clonal migration diagram shows a parsimonious reconstruction of evolutionary (top to bottom) and migratory (left to right) relationships. Arrows indicate evolution (solid) or migration (dashed). Each node represents a site with sections colored by clonal composition. Clone colors are matched between trees and diagrams. 127

Figure 4.49: Overlap of mutations detected in each sequenced site of WA_24. Red color gradient indicates the fraction of all mutations in each intersection. 129

Figure 4.50: Tally of mutations classified according to clonality and distribution 130

Figure 4.51: Genome-wide copy-number profiles of normalized log2(tumor / normal) coverage ratios across samples from WA_24 (black - diploid, red - gain, blue - loss, green - LOH). Two representative regions are selected to illustrate LOH segments shared across all tumor samples. 131

Figure 4.52: Integrated summary of genetic drivers, gene expression, shared mutations, and notable LOH regions in WA_24 samples. 132

Figure 4.53: Results of permutation testing to compare the observed number of genes within regions of shared LOH in WA_24 (red dashed lines) with a distribution based on a null model (histogram). 133

Figure 4.54: Morphological heterogeneity and spectrum in WA_24 at different spatial sites with available FFPE slides. A-B: Mixed classical acinar and organoid pattern in the lung. C-D: Small cell neuroendocrine carcinoma-like pattern in the liver. E-F: Small cell neuroendocrine carcinoma-like pattern in the prostate. 134

Figure 4.55: Expression of marker genes for molecular subtypes mCRPC as defined by Labrecque et al.) 135

Figure 4.56: Expression of marker genes for molecular subtypes mCRPC as defined by Labrecque et al.) 136

Figure 5.1:Standard deviation of MHC region calculated by MHCnvex has a significant SD reduction compared to the same value calculated by standard CNV approach CNVEX 164

Figure 5.2: Number of segments present in the 15Mb window of genome and MHC region.... 166

Figure 5.3: CNV agreement between MHCnvex and a general CNV calling tool for different settings of WGS and WXS. 167

Figure 5.4: Samples with inconsistence CNV calls in panel D have significantly lower purity compared to the consistent ones 168

Figure 5.5: Top panel shows the difference in allelic expression of MHC class-I genes. Across all cohorts, samples detected by MHCnvex with MHC class-I LOH have significantly higher difference in expression of their alleles compared to normal samples or tumor samples detected to have balanced allelic number. Bottom panel shows that these different groups across all cohorts have comparable purity and therefor, in this analysis, purity would not affect the difference in expression 168

Figure 5.6: MHCnvex accepts input from an HLA haplotyping and a CNV analysis tool. Then works in 7 steps to call CNV and LOH for genes in MHC class-I and class-II. Finally, it outputs allelic copy number of each gene, LOH call for each gene, and a visualization of the loci. 3 examples of the MHC region have been shown 170

Abstract

Genomic instability is a hallmark of cancer, driving tumor evolution and enabling the acquisition of additional cancer hallmarks such as evasion and ability to metastasis. While genomic instability facilitates tumor adaptation and progression, it also makes tumors susceptible to therapies targeting DNA repair pathways (*e.g.,* PARP inhibitors) and immune checkpoints inhibitors. Reliable measurement of genomic instability is essential for predicting patient responses to specific treatments, determining prognosis, and discovering novel therapeutic targets. Despite its clinical significance, there is a scarcity of robust methods to quantify the degree of genomic instability in tumors, leading to a limited number of biomarkers predicting genomic instability and hence patients' responses to treatment in current clinical use.

In this dissertation, I present novel computational models to measure genomic instability in tumors. Utilizing these methods, I identify and validate new prognostic biomarkers in Lung adenocarcinoma based on the expression of a specific gene, making them practical biomarkers for predicting genomic instability in clinical applications. Furthermore, I will present a genomic alteration in high grade serous ovarian carcinoma which can be used as a predictive biomarker for patient response to platinum-based treatment. Additionally, I explore the downstream effects of genomic instability, particularly the resultant genomic heterogeneity in prostate cancer. In this dissertation, I aim to enhance the understanding of genomic instability and its implications in cancer biology by providing new computational methods to measure this phenomenon, ultimately contributing to biomarker development and improved clinical outcomes for patients.

I first introduce two new measurements of genomic instability for chromosomal changes called Genome Contiguity Index and breakage intensity and clustering. Then using these methods I identify *IGF2BP3* as a prognostic biomarker for genome instability in Lung adenocarcinoma.

Next, I will show loss of heterozygosity (LOH) of chromosome 17 is associated with a patient's response to platinum-based treatment in high grade serous Ovarian carcinoma.

Furthermore, I will investigate the underlying biology of chromosome 17 LOH and identify the presence of *TP53* functional copy to partially account for the association of chromosome 17 LOH and patient response.

In the next chapter, I will investigate the downstream effect of genomic instability and resulting genomic heterogeneity in prostate cancer (PCa). I will quantify the extent of genomic heterogeneity at mutation and CNV level, with a focus on pathogenic alterations in PCa patients. I will also investigate the effect of this heterogeneity on detection of alterations using a single sample, which is common in many of the current studies in the field.

Lastly, I will introduce another method for detection of copy number variations (CNV) changes, specifically in Human leukocyte antigens (HLA) genes. Loci of HLA genes has the highest polymorphism in the genome and hence, it requires special methods to identify CNVs in this region. In this chapter, we show that by utilizing dynamic reference genome selection (as opposed to standard reference genome), we can significantly improve our CNV calls for HLA genes.

Chapter 1 Introduction

1.1 Motivation

Genomic instability (GI), characterized by chromosomal rearrangements and genome mutations, is a hallmark of cancer [1,2]. High genome instability in cancer patients has been associated with the worst outcome [3] and increased genomic heterogeneity [4]. Furthermore, the presence of certain types of GI in cancer patients has been used in clinical settings as a predictive biomarker for patient response to poly ADP ribose polymerase (PARP) inhibitors. Despite the importance of GI in cancer treatment and biomarker development, GI is understudied and appropriate computational methods to measure GI in the genome are lacking. In this thesis, I will lay out (I) a tool developed to detect copy number variation (CNV) in tumors, which is essential in exploring chromosomal instability (CIN) which is a type of GI, (II) two novel methods developed to quantify and characterize GI with prognostic values, (III) incorporating developed tools and novels to identify prognostic and predictive biomarkers in high grade Serous Ovarian Carcinoma (HGSOC) and Lung Adenocarcinoma (LUAD), and (IV) characterization of downstream GI induced heterogeneity in metastatic castration-resistant Prostate cancer (mCRPC).

1.2 Genomics of Cancer

1.2.1 Cancer initiation and progression

Cancer arises as the result of defects in regulatory mechanisms within a cell which control cell proliferation. This defect results in uncontrolled cell proliferation and formation of lesions which can lead to multiple system failure, anemia [5] and patient death.

Defects in regulatory mechanisms of a cell can happen through multiple routes including genome alterations [6], epigenome alterations [7] and post-translational alterations [8]. Alterations in the cell genome are the most common route for tumors to achieve overproliferation. These alterations can be substitutions, deletions, insertions, amplifications and rearrangements [6]. Changes in the genome of the cell can lead to phenotypic changes such as RNA overexpression, fusion transcript and altered splicing [9], resulting in dysfunctional proteins.

Although changes in the genome of a cell can lead to alterations in protein sequence, not all the changes in protein sequence will lead to cancer. Alterations which can lead to cancer progression are called 'Driver alterations'[10] and otherwise, they are called 'Passenger alterations'. While driver alterations happen to different genes depending on the cancer type and the cell of origin, majority of driver alterations happens in two groups of genes: (I) Oncogenes, and (II) Tumor suppressor genes (TSGs). Alterations in oncogenes are generally considered to be a gain-of-function (GOF) which promotes cell growth, division, and survival. On the other hand, alterations in TSGs are generally loss-of-function (LOF) which disable the processes in the cell which regulate the cell growth, DNA replication and cell cycle checkpoint activation [11].

Thanks to over 20 years of research on cancer genomics, the majority of driver alterations in cancers are documented [10,12]. Furthermore, identification of driver events in cancer has enabled the discovery of drugs which exploit the deficiencies caused by the alterations such as PARP inhibitors for BRCA deficient tumors [13].

1.3 Detection of alterations in tumors

The initial sequencing of the human genome, done through the Human Genome Project (HPG), took over a decade to complete and cost over two billion dollars [14]. This effort enabled scientists to utilize the now known sequence of the human genome and develop technologies to

sequence patients' tumors much faster and cheaper, removing a bottleneck in the analysis of thousands of tumors. Next-generation sequencing (NGS) was introduced between 2004-2006 [15] as the next method to Sanger sequencing [16] (*i.e.,* First generation sequencing*)* which initiated the HGP [17]. Introduction of NGS led to significant increase in the number of sequenced genomes. NGS was designed based on two hallmarks, high throughput and single molecule sequencing [18]. Initial NGS platforms such as Illumina or Ion Torrent platforms sequenced short sequences of genome (*i.e.,* short-reads) in parallel and resulted in millions of short sequences of genome which needed to be assembled together. That gave rise to computational aligners which could utilize the first known sequence of the human genome (*i.e.,* HGP) and align the short reads onto the genome.

1.3.1 Alignment

Computational alignment methods align reads, extracted by an NGS platform, against a reference genome [19]. This process can happen by brute force but it is impractical since for each tumor, hundred millions of reads are generated and they all need to be aligned to the genome. Due to this problem, different bioinformatics algorithms have been developed which can map the reads to the genome while dynamically considering the possibility of a substitution, insertion or deletion within the read [20]. These capabilities alongside improvements in accuracy and speed of alignment algorithms over the past 20 years, makes alignment a standard and reliable step in genomics analysis.

1.3.2 Sequence altering variations detection

After the alignment is done, the next step is to detect sequence altering changes in the genome of the tumor cells. It is a challenging task to do so due to multiple reasons: (I) Sequencing errors as well as alignment errors can lead to reads which have a different sequence for a base in

the genome compared to the matched normal genome, but in reality there has not been any alteration in the tumor, (II) In a bulk sample, for each DNA base we get reads from both the tumor cells as well as normal cells (e.g., Immune cells present in the tumor microenvironment). This will lead to uncertainty in calling the alterations since not all the reads aligning to a base support the alterations. This problem can be exacerbated by the fact that mutations can happen subclonally (i.e, not present in all the tumor cells), and furthermore, they can appear on only some of the alleles within a tumor, which means even some of the reads originated from tumor cells will not support a real alteration.

To solve these problems of uncertainty in calling alterations, multiple tools such as Mutect2, VarScan, VarDict, Strelka2, etc.[21–24] have been developed. Majority of these tools incorporate likelihood based statistical models to call possible alterations within the sequence of a tumor. To increase the sensitivity of calls, pipelines such as GATK [25] and TPO [26] first use somatic variant calling tools to call every possible alterations with high sensitivity (which will include false positives such as erroneous calls and germline alterations). Then they use another step of either likelihood based or heuristic algorithm to further remove any erroneous or germline alteration call.

1.3.3 Detection of Copy number variations

Another type of cancer driving alterations which can be detected using NGS aligned reads are CNVs. For any CNV analysis, our end goal is to detect gain and losses in each genomic loci. To achieve this goal, the absolute copy number of tumors for each loci of the genome needs to be calculated. This calculation, as will be explained in the next section, depends on many factors such as read coverage of basepairs, B-allele frequency of germline variants, segmentation of the genome, purity (*i.e.,* proportion of tumor cells in the sequenced sample) and ploidy of the tumor.

Similar to detecting sequence altering variations, detecting CNVs using aligned reads comes with multiple challenges including: (I) NGS read counts are agnostic to the ploidy of the tumor, *i.e.,* given the same sequencing depth, the number of reads generated for a tumor with ploidy of two, is similar to the number of reads generated for a tumor with ploidy of four. Hence, it is not straightforward to call the ploidy of tumors using the read counts alone. (II) To determine the base ploidy of the tumor, we need to first know the purity of tumors. But similarly, to identify the purity, we need to know the ploidy of the tumor. So these two values need to be determined together, which will introduce scalability challenges and requires special methods to determine the values. To make matters worse, many algorithms which attempt to determine purity and ploidy simultaneously, end up with multiple solutions (different pair values of purity and ploidy), which then requires manual curation of an expert to select the best solution.

Multiple tools have been developed to determine CNVs including Absolute [27], ASCAT [28], PureCN [29] and CNVEX [26,30]. All of these tools use likelihood modeling to determine the purity and ploidy of tumors, and then incorporate those values along with other related values (explained in next section) to determine the absolute copy numbers of each loci. Using the absolute copy number of each loci and the base copy number of a tumor, one can determine the variations in copy number of genes.

1.4 Chromosomal instability in cancer

1.4.1 Copy number variations

CNVs are defined as alterations in the genome where the number of copies for a segment of the genome is changed. These changes can be duplication of a region which leads to additional copies of a segment (*i.e.,* gains) or it can be a deletion of a segment, leading to loss of one or sometimes both copies of a genome segment. In some cases, most often in short segments affecting

only one gene, the genome can acquire tens of additional copies, in which case these events are called amplifications and they are usually clinically significant [31,32]. CNVs in cancer can be seen in one of the 3 roles: (i) Driver CNVs that help initiating and progressing the cancer such as chromosome3 p-arm deletion in Kidney cancer or MYC amplification in Prostate cancer, (ii) Passenger CNVs which can appear as the result of high CIN but not providing benefits to the survival or overproliferation of the cell, and (iii) CNVs that appear as a response to treatment by providing survival benefits which will lead to acquired resistance for cancer. One well known example of CNV as a response to treatment is AR amplifications in Prostate cancer in response to androgen deprivation therapy (ADT), which results in overexpression of androgen, rendering ADT based treatments ineffective.

While driver CNVs occur recurrently in patients of the same cancer, there are usually few of them in each tumor and hence, the driver CNV itself does not contribute much to the overall pattern of CNVs in the genome. On the other hand, passenger CNVs occur on a pseudo-random positions on the genome, most often as a result of some DDR defect in the genome. These CNVs make up the majority of the CNVs seen on the genome (hence, form the CNV pattern of the genome) and depending on the deficient mechanism that gave rise to them, they can form different patterns on the genome. For instance, Prostate cells with bi-allelic deficient *CDK12* are known to form small but numerous CNVs [33], as opposed to tumors with *BRCA1/2* deficiencies which generate large scale imbalances in the genome [34]. By identifying these patterns and extent of them, we can identify the underlying deficient mechanism in the cell and hence, target the possible vulnerabilities caused by the underlying deficiencies.

1.4.2 Chromosomal Instability

Genome instability is defined as the increased tendency of a cell's genome to undergo changes like the single nuclei variation, gain, loss, or rearrangement of chromosomes or chromosome sections [35]. In this thesis, I focus on a type of genome instability called Chromosomal instability (CIN) which is characterized by high rates of CNV and aneuploidy in the genome [36]. Almost all tumors harbor some degree of CIN in their genome [37]. If the CIN is very low in a tumor, then the tumor can not evolve and it can not acquire cancer hallmarks such as the ability to metastasize or drug resistance. On the other hand, if CIN is too high, the cell's DNA will acquire too many alterations which eventually leads to defection in vital pathways within the cell and cause cell death. The degree of CIN in each tumor is dependent on the cell or origin (*i.e.,* cancer type) as well as the cell biology including the cell's genomics and epigenomics state. The amount of instability in tumors can be a biomarker for clinical response to certain treatments. For example, high-grade serous ovarian carcinoma (HGSOC) has a very high CIN and hence, it is prone to platinum-based treatments such as Cisplatin (cite) which will exacerbate the instability in tumors so much so that tumor cells can not survive. Furthermore, not only the amount of CIN can have clinical significance, but the pattern of CIN can also be used as a biomarker for response to certain treatments. For example, Deficiency in Homologous Recombination pathway (i.e., HRD) can lead to certain patterns in the CNVs of tumors [34], such as large scale Loss of heterozygosity (LOH) or losses of telomeres as a result of large deletions. Myriad Genetics has utilized a score to quantify these events and designed a GI based biomarker assay to detect patients with HRD. This has led to their assay receiving FDA approval for identifying patients who are likely to respond to PARP inhibitors [38], as patients with HRD are likely to respond to PARP inhibitors, but otherwise, they have shown no benefit from PARPi treatments.

In this thesis, I will introduce novel methods to measure CIN and further CIN based biomarkers which can open new paths in the study of genome instability and highlight its utility in translational research.

1.5 Genomic Heterogeneity

One of the downstream of genome instability (mutational as well as CNV instability) is genomic heterogeneity which leads to a heterogeneous mix of cells within a tumor with varying genomic alterations. For example, a tumor may initially have a homogeneous cell population but, due to genome instability, later develop subclonal populations with different copy number alterations and structural variations. The resulting genomic heterogeneity fuels tumor adaptation, allowing selection for advantageous alterations that can lead to tumor progression, metastasis, and therapy resistance. This will cause challenges for clinicians, as different tumors within a patient, might have different responses to treatments due to their difference in genetic alterations. Hence, it is essential to know the extent of heterogeneity in each cancer as well as heterogeneity of actionable alterations.

In this thesis, I extensively characterized the genomic heterogeneity of mCRPC and the heterogeneity of pathogenic genes within mCRPC.

1.6 References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144: 646–674.
2. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022;12: 31–46.
3. Li GX, Chen L, Hsiao Y, Mannan R, Zhang Y, Luo J, et al. Comprehensive proteogenomic characterization of rare kidney tumors. Cell Rep Med. 2024;5: 101547.
4. Bakhoum SF, Landau DA. Chromosomal instability as a driver of tumor heterogeneity and evolution. Cold Spring Harb Perspect Med. 2017;7. doi:10.1101/cshperspect.a029611
5. Spivak JL. The anaemia of cancer: death by a thousand cuts. Nat Rev Cancer. 2005;5: 543–555.

6. Chakravarthi BVSK, Nepal S, Varambally S. Genomic and epigenomic alterations in cancer. Am J Pathol. 2016;186: 1724–1735.
7. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128: 683–692.
8. Sharma B, Prabhakaran V, Desai A, Bajpai J, Verma R, Swain P. Post-translational modifications (PTMs), from a cancer perspective: An overview. Oncogen (Westerville). 2019. doi:10.35702/ONC.10012
9. PCAWG Transcriptome Core Group, Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578: 129–136.
10. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. Trends Mol Med. 2023;29: 554–566.
11. Lee EYHP, Muller WJ. Oncogenes and Tumor Suppressor Genes. Cold Spring Harbor Perspectives in Biology. 2010;2: a003236–a003236.
12. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019;575: 210–216.
13. González-Martín A, Pothuri B, Vergote I, DePont Christensen R, Graybill W, Mirza MR, et al. Niraparib in patients with newly diagnosed advanced ovarian cancer. N Engl J Med. 2019;381: 2391–2402.
14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.
15. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif). 2013;6: 287–303.
16. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265: 687–695.
17. Park ST, Kim J. Trends in next-generation sequencing and a New Era for whole genome sequencing. Int Neurourol J. 2016;20: S76–83.
18. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021;82: 801–811.
19. Kim J, Ji M, Yi G. A review on sequence alignment algorithms for short reads based on next-generation sequencing. IEEE Access. 2020;8: 189811–189822.
20. Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, et al. Technology dictates algorithms: recent developments in read alignment. Genome Biol. 2021;22: 249.
21. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22: 568–576.
22. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44: e108.
23. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15: 591–594.

24. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019. p. 861054. doi:10.1101/861054
25. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.
26. Chowdhury S, Kennedy JJ, Ivey RG, Murillo OD, Hosseini N, Song X, et al. Proteogenomic analysis of chemo-refractory high-grade serous ovarian cancer. Cell. 2023;186: 3476–3498.e35.
27. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30: 413–421.
28. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107: 16910–16915.
29. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11: 13.
30. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2103240118
31. Linja MJ, Savinainen KJ, Saramäki OR, Tammela TLJ, Vessella RL, Visakorpi T. Amplification and Overexpression of androgen receptor gene in hormone-refractory prostate Cancer1. Cancer Res. 2001;61: 3550–3555.
32. Gao Y, Vallentgoed WR, French PJ. Finding the right way to target EGFR in glioblastomas; Lessons from lung adenocarcinomas. Cancers (Basel). 2018;10: 489.
33. Wu Y-M, Cieślik M, Lonigro RJ, Vats P, Reimers MA, Cao X, et al. Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. Cell. 2018;173: 1770–1782.e14.
34. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017;23: 517–525.
35. Drews RM, Hernando B, Tarabichi M, Haase K, Lesluyes T, Smith PS, et al. A pan-cancer compendium of chromosomal instability. Nature. 2022; 1–8.
36. Thompson SL, Bakhoum SF, Compton DA. Mechanisms of chromosomal instability. Curr Biol. 2010;20: R285–95.
37. Geigl JB, Obenauf AC, Schwarzbraun T, Speicher MR. Defining “chromosomal instability.” Trends Genet. 2008;24: 64–69.
38. Myriad Genetics Receives FDA Approval of BRACAnalysis® CDx as a Companion Diagnostic for Lynparza® in Early Breast Cancer. In: Myriad Genetics [Internet]. [cited

19 Nov 2024]. Available: https://investor.myriad.com/news-releases/news-release-detail/22906/

Chapter 2 Mathematical Modeling of Genome Contiguity and Breakage Patterns Characterize Chromosomal Instability

Portions of this chapter is currently under review at Cancer Cell.

2.1 Introduction

Measuring CIN has been of interest to scientists for the past 40 years. One of the breakthroughs in studying CIN was cytokinesis-block assay proposed by Fenech and Morley [1] in 1985, where they essentially provided a method to differentiate in vitro cells which have completed nuclear division from those which have not divided. This enabled scientists to study effects of erroneous cell-cycles during cell division which included mis segregation of chromosomes and hence, chromosomal instability.

Further advancement in assays during 1990s, enabled the development of chromosome specific probes for centromeric, pre centromeric and telomeric regions of human [2,3] and mouse [4,5] cells. Using these methods, scientists were able to identify mechanisms that can give rise to CIN in cells and discover CIN driving effects in multiple diseases including cancer and birth defects [6].

In the 2000s, by the rise of NGS, methods to quantify CIN using sequencing data also emerged. Genome Instability Index(GII) [7] was one of the first methods which became highly adapted by other researchers. This method measured the proportion of genome which was altered relative to the baseline copy number of the genome (*e.g.,* if the whole genome is duplicated, the baseline is 4, otherwise it is 2). Later, this method was further refined by Burrel et al. [8] and called weighted genome instability index (wGII). wGII is essentially weighting the instability score

according to the length of the chromosomes to account for variations in chromosome length, so that each chromosome contributes equally to the score.

While measuring a portion of the genome with aberrant copy number can reflect CIN to some extent [9–11], it fails to capture CIN in complex chromosomal aberrations. One example of such shortcoming in capturing instability is when the aberrant portion of the genome is similar between two tumors, but one of them has few large aberrations, while the other tumor has many breakpoints and small segments. In this case, according to wGII, both genomes have similar degree of instability, while the genome with more segments has undergone more errors in division and hence, more unstable.

Given the limitations of wGII, which is the most widely used measure of CIN in the field, there is a need for a novel, more advanced measure of CIN which can capture the difference of CIN in genomes with complex alterations as well.

In this chapter, I will present a cohort of Lung adenocarcinoma (LUAD) patients and use two novel methods of calculating CIN to analyse LUAD chromosomal instability. Further, I will utilize CIN measures to identify prognostic biomarkers to predict instability in LUAD patients.

2.1.1 I CPC-CPTAC LUAD cohort

To investigate the chromosomal instability in LUAD, a cohort of 406 tumors and 388 normal adjacent tissues (NATs) were analyzed. This cohort consisted of 202 already published tumors [12,13] and 204 newly sequenced tumors, through the collaboration of the U.S. National Cancer Institute (NCI)'s Clinical Proteomic Tumor Analysis Consortium (CPTAC) and Taiwan Cancer Moonshot team under the International Cancer Proteogenomics Consortium (TW-ICPC).

2.2 Results

2.2.1 Chromosomal instability in Lung Adenocarcinoma

Each tumor harbors some degree of CIN depending on their biology and cell of origin. LUAD is commonly considered to have high instability with complex chromosomal aberrations[14]. While on average that is true in our cohort, we saw a large variance in the number of breakpoints and alterations in each tumor (**Figure 2.1**).

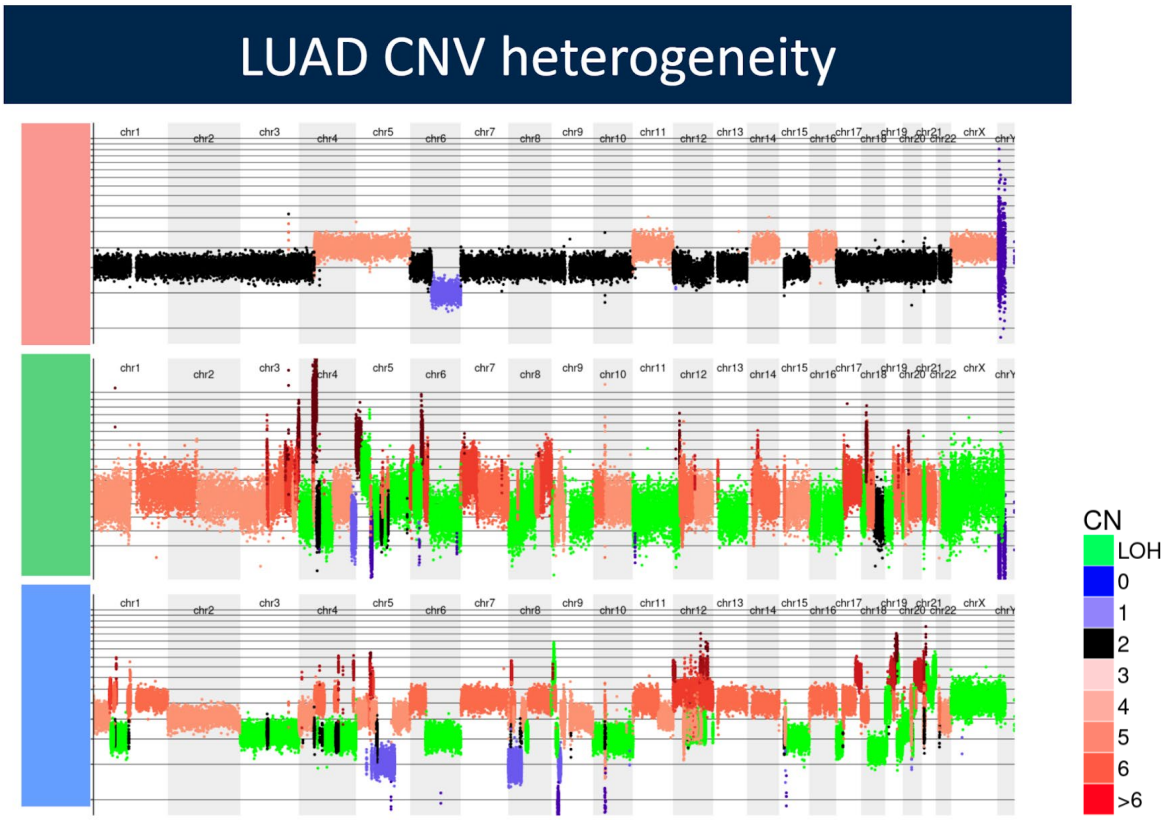


Figure 2.1: Copy number profile of 3 tumors from CPTAC-ICPC cohort. This figure highlights the CIN heterogeneity among LUAD tumors. The top figure shows a very stable tumor with few breakpoints, the middle figure highlights a very highly shattered genome with the majority of short segments, and the bottom figure depicts a tumor that is in general stable, but has been shattered locally in some regions.

This observation opted us to investigate the association between instability and cancer aggressiveness. We initially used a common measure of CIN, wGII, to characterize the instability of tumors in our cohort. We observed a binomial distribution of wGIIs in our cohort (**Figure 2.2**). Next, to separate patients into two groups of stable and unstable based on wGII, we evaluated thresholds incrementally from the minimum to maximum wGII values, beginning with a single patient classified as "stable" and iteratively reassigning patients by transferring the individual with

the lowest wGII from the "unstable" group to the "stable" group and running cox-regression each iteration.

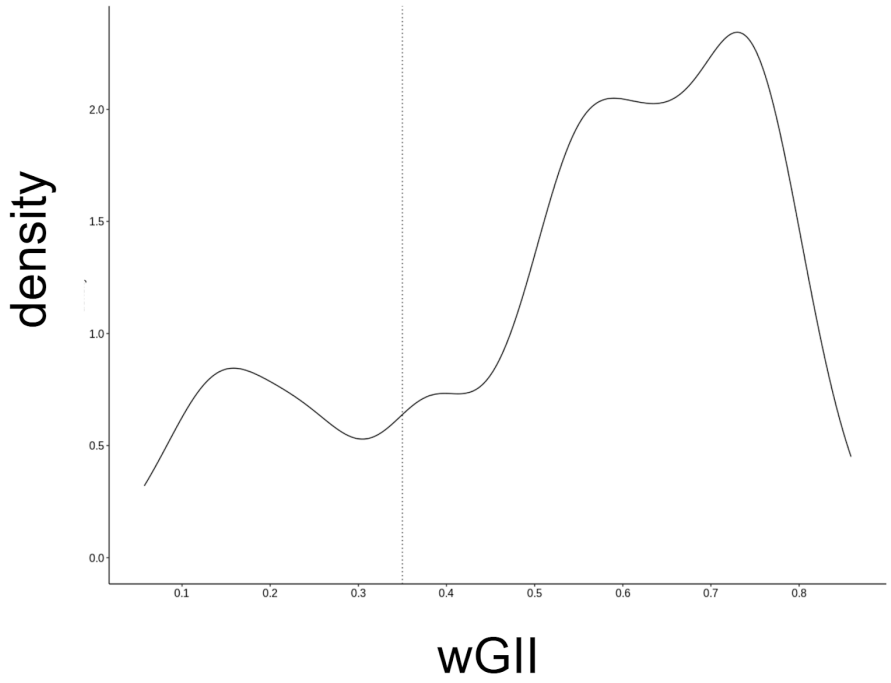


Figure 2.2: wGII distribution of tumors in CPTAC cohort. The dotted line represents the threshold which leads to best separation of patients based on their overall .

This approach identified a threshold of 0.35 as achieving the best separation in patient overall survival outcomes. However, even with this optimal threshold, wGII did not yield a statistically significant distinction in overall survival between the two groups (P=0.29, **Figure 2.3**). This observation highlighted the limitations of wGII, where quantifying genomic imbalances (aneuploidy) fails to predict tumor aggressiveness (and hence overall survival).

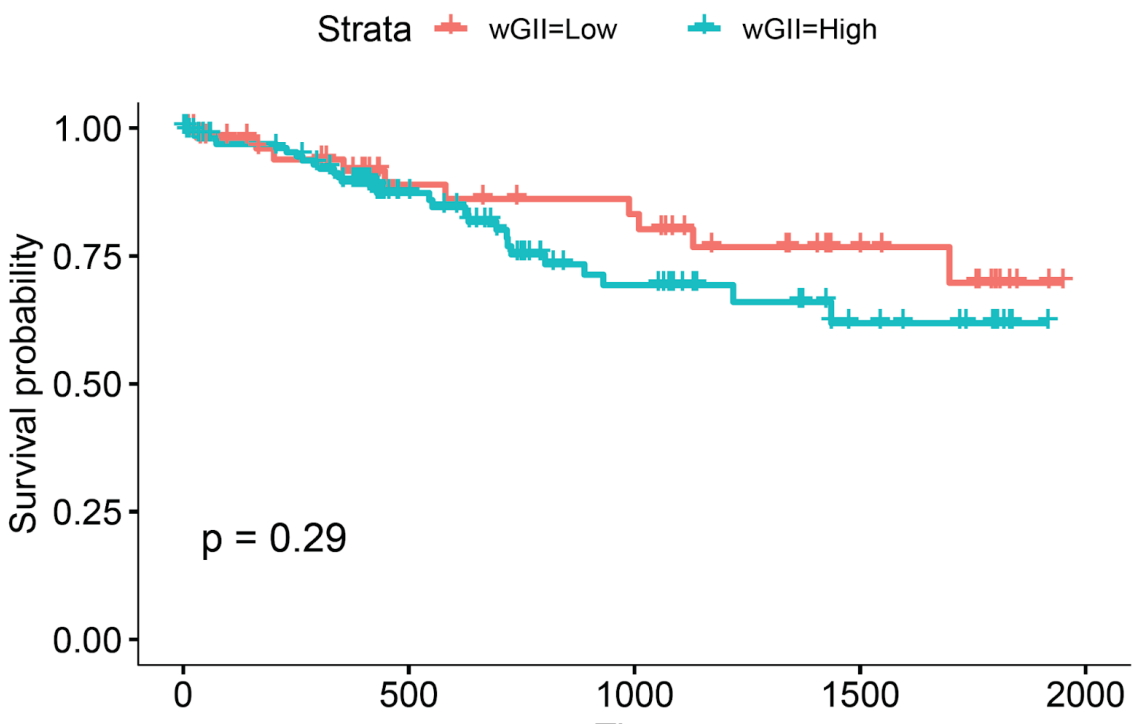


Figure 2.3: Overall survival of CPTAC patients grouped according to their wGII with threshold of wGII=0.35

2.2.2 Prognostic role of genome contiguity

Following the failure of wGII (genomic imbalances) in predicting patients overall survival, we hypothesized that the reason wGII fails to quantify aggressiveness is because it fails to consider the degree of fragmentation (contiguity). The degree of fragmentation in a genome can be reflected by the segment length distribution, with a higher level of fragmentation generally corresponding to shorter average segment lengths. Therefore, assessing segment length distribution may provide a more realistic indicator of CIN and tumor aggressiveness. We used the 75th percentile of segment length (*i.e.,* genome contiguity score) for each tumor as a proxy for the presence of large segments. Tumors with low 75th percentile have lower segments in general which will suggest more fragmentation of the genome and hence, higher CIN within the genome.

We grouped patients based on their corresponding genome contiguity score (**Methods**). We found that grouping patients based on their genome contiguity score is associated with overall survival ($p < 0.02$, **Figure 2.4**).

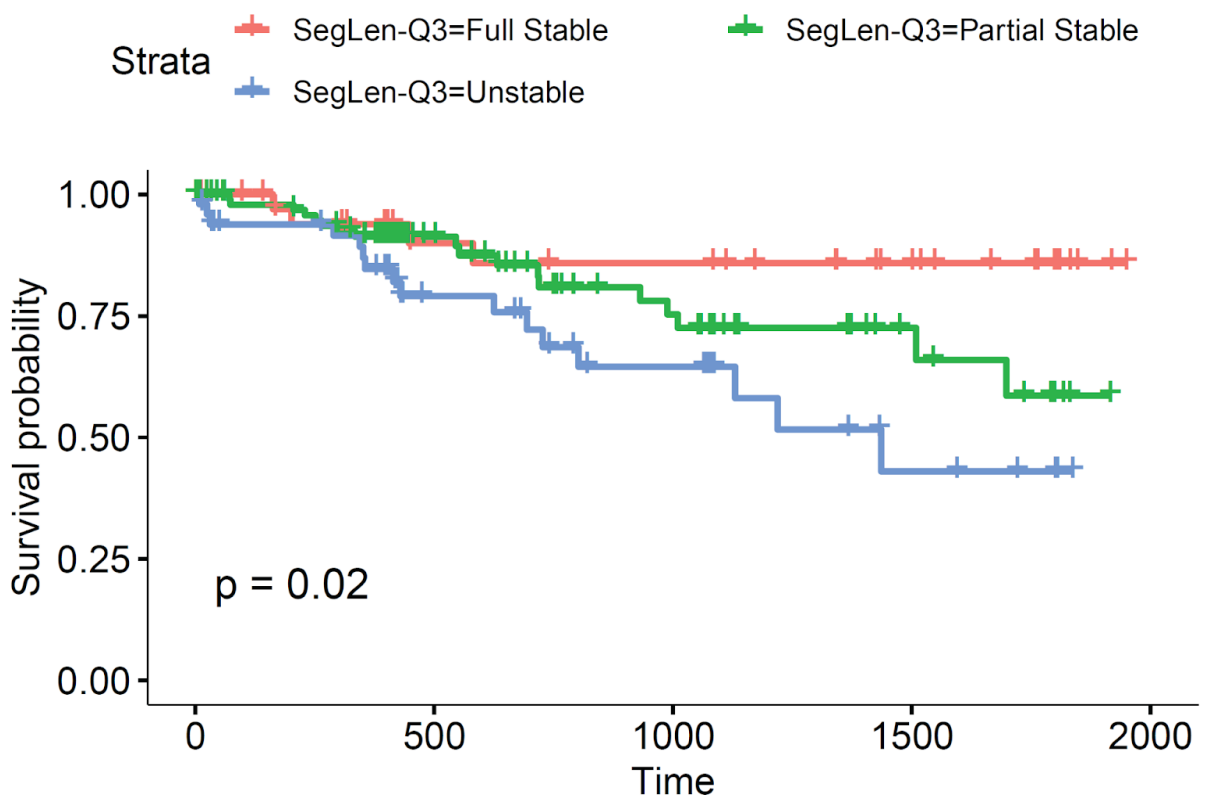


Figure 2.4: Overall survival of CPTAC patients grouped according to their contiguity score.

We further validated the prognostic power of contiguity score in an independent cohort of LUAD patients (APOLLO-LUAD[15]) and found a significant association in that cohort as well (P=0.024, **Figure 2.5**). We also showed the prognostic value of contiguity score in cohorts with different cancers including CPTAC-HNSCC[16] (P=0.034) and CPTAC-CCRCC[17] (P=0.0042, **Figure 2.6**).

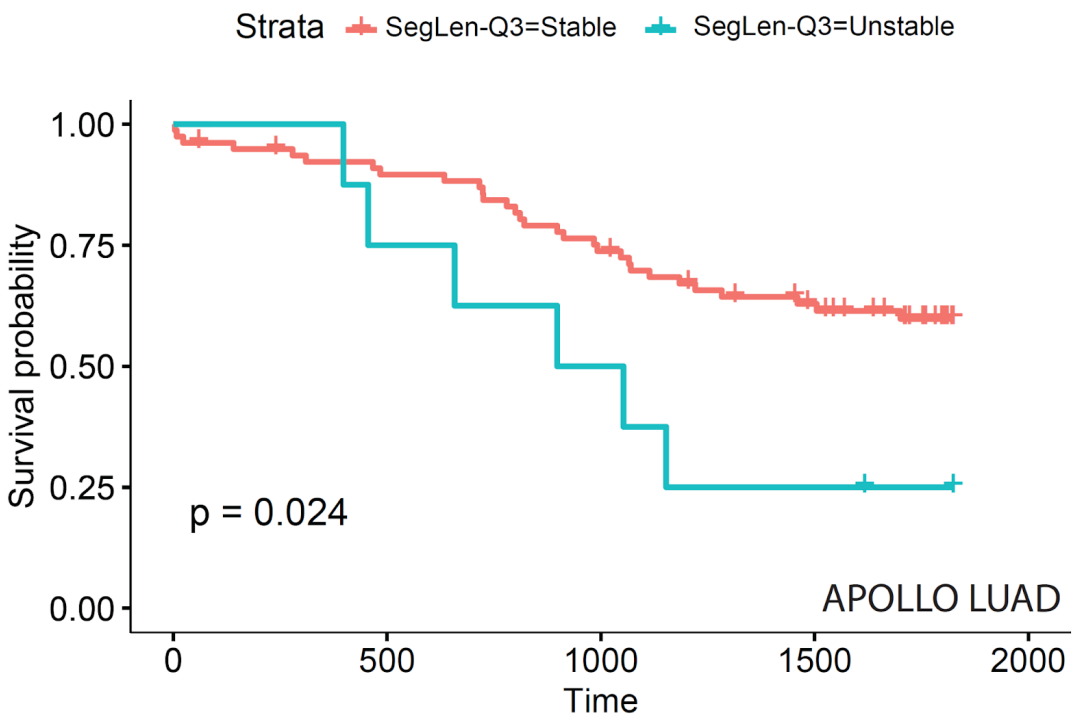


Figure 2.5: Overall survival of APOLLO-LUAD patients grouped according to their contiguity score.

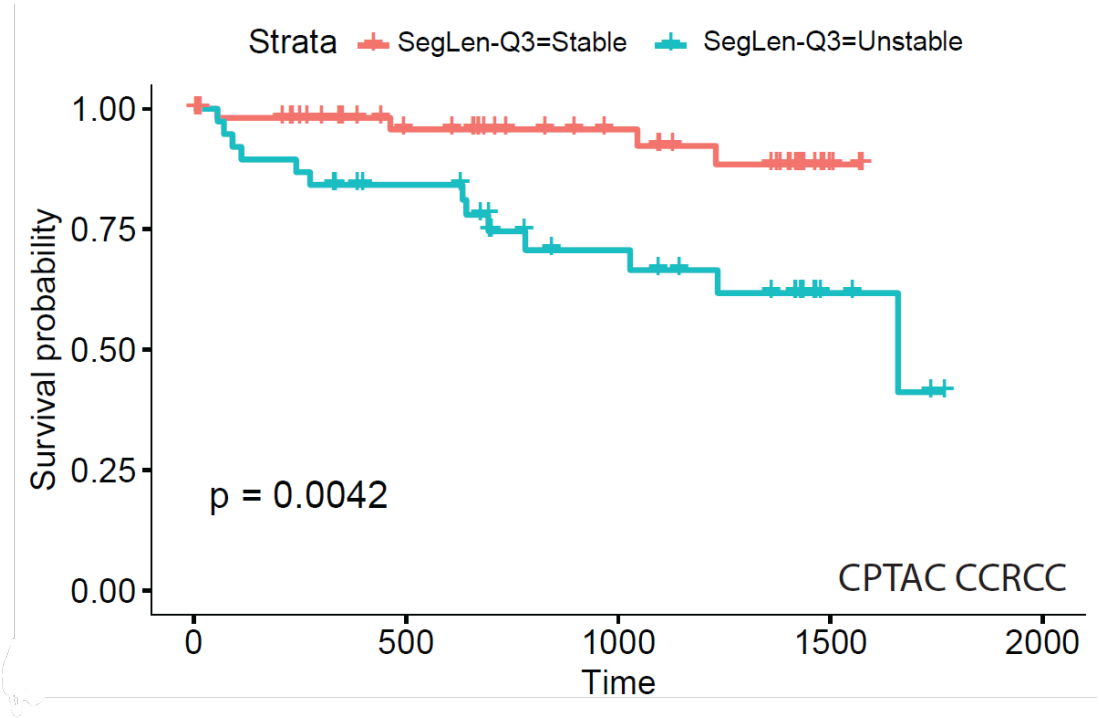


Figure 2.6: Overall survival of CPTAC-CCRCC patients grouped according to their contiguity score.

2.2.3 Genetics underpinning genome contiguity

To identify the underlying genetic associations with high genome contiguity, we stratified tumors into 3 groups of Fully stable, Partially stable and Unstable (**Methods**). We compared the regions of recurrent amplification in each group. We observed the expected increase in structural variants (SVs) (**Figure 2.7**) along with focal amplifications of *TERT*, *NKX2-1*, and *MYC loci* in the highly fragmented tumors (**Figure 2.8**).

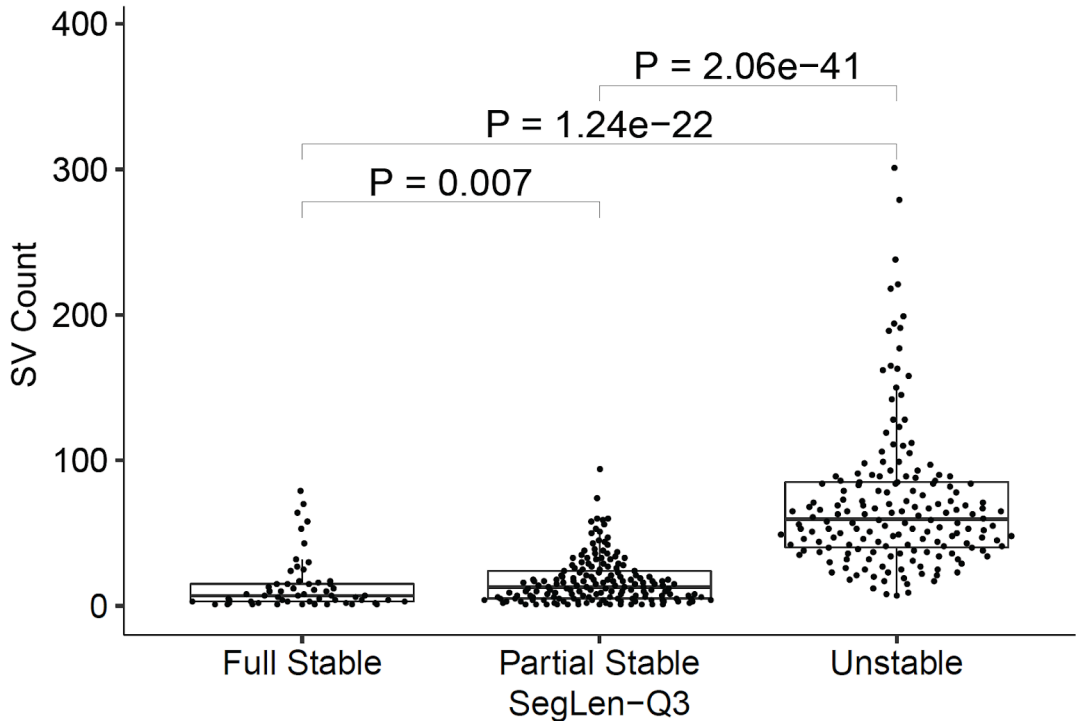


Figure 2.7: Number of detected SVs per each contiguity group, where as expected, the most fragmented group (Unstable) had significantly higher number of SVs

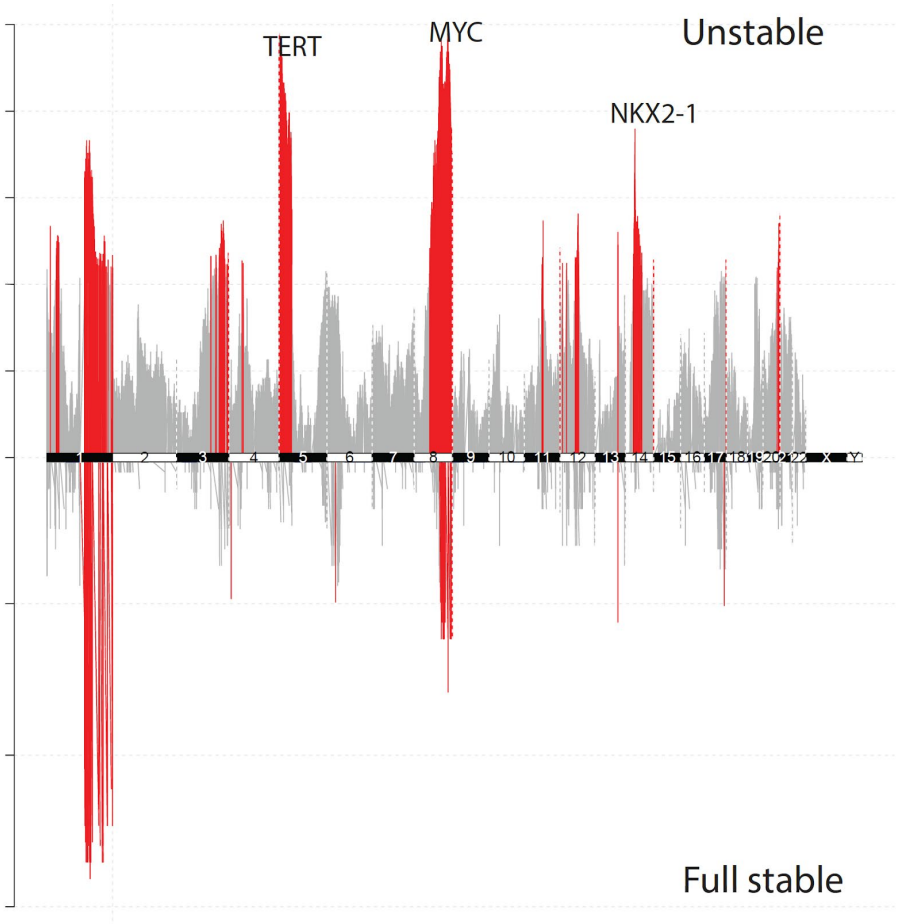


Figure 2.8: Regions of recurrent amplification in each contiguity group. As expected, a higher rate of recurrent amplification is observed in the fragmented group (Top) versus the full stable group (bottom), as well as certain loci of known oncogenes such as *TERT*, *MYC* and *NKX2-1*.

TERT, a telomerase activating gene, is expressed during development in embryonic stem cells and later silenced as cells mature [18]. Many cancers including Melanoma and LUAD activate *TERT* as they advance, given the ability of TERT in stabilizing chromosomes and enhancing proliferation [19]. In our study, we first confirmed that the target of amplification in amplified regions observed in **Figure 2.8** is the *TERT* gene itself. We evaluated the minimal region of amplification among all the samples with amplification on chromosome 5 short arm. We found the majority of amplifications in this region overlaps the *TERT* gene, suggesting *TERT* as the target of these amplifications (**Figure 2.9**).

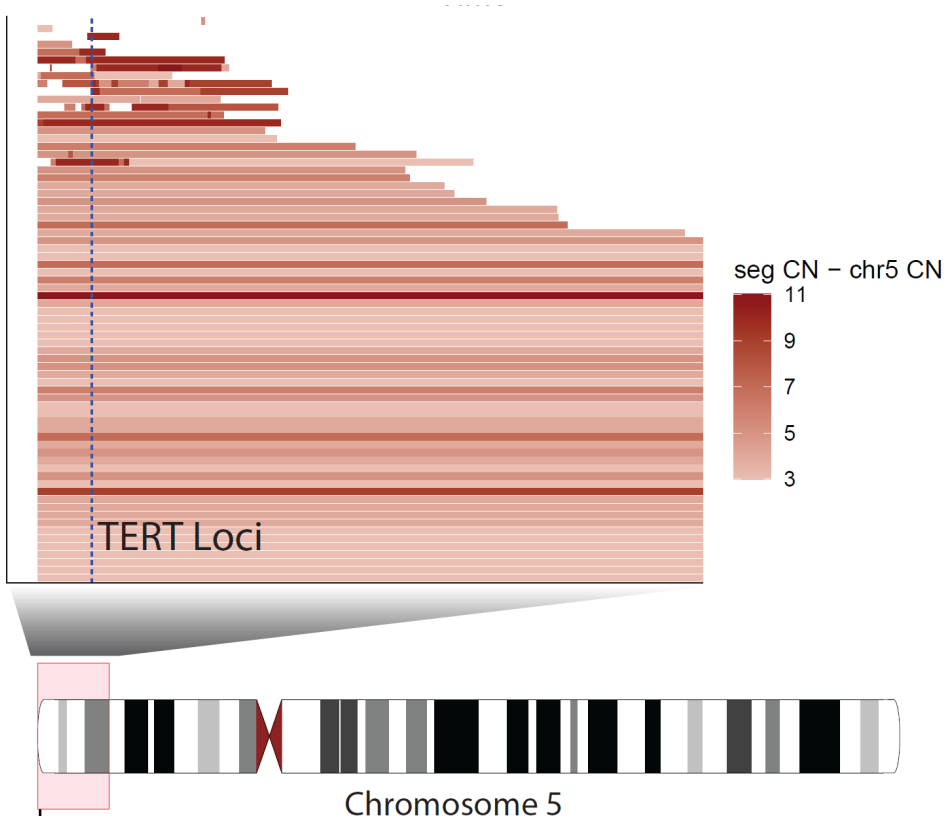


Figure 2.9: Amplified segments on chromosome 5 short arm. Each horizontal line shows an amplified segment. Each row of x-axis represents one patient, and y-axis represents coordinates along chromosome 5. Dotted line shows the *TERT* loci.

We also confirmed that *TERT* amplification leads to increased cognate expression (**Figure 2.10**) and downstream activity (**Methods**, **Figure 2.11**). We also observed *TERT* association with with OS (P=0.029, **Figure 2.12**), consistent with the role of *TERT* in sustaining proliferation of chromosomally unstable tumors with telomere attrition[20].

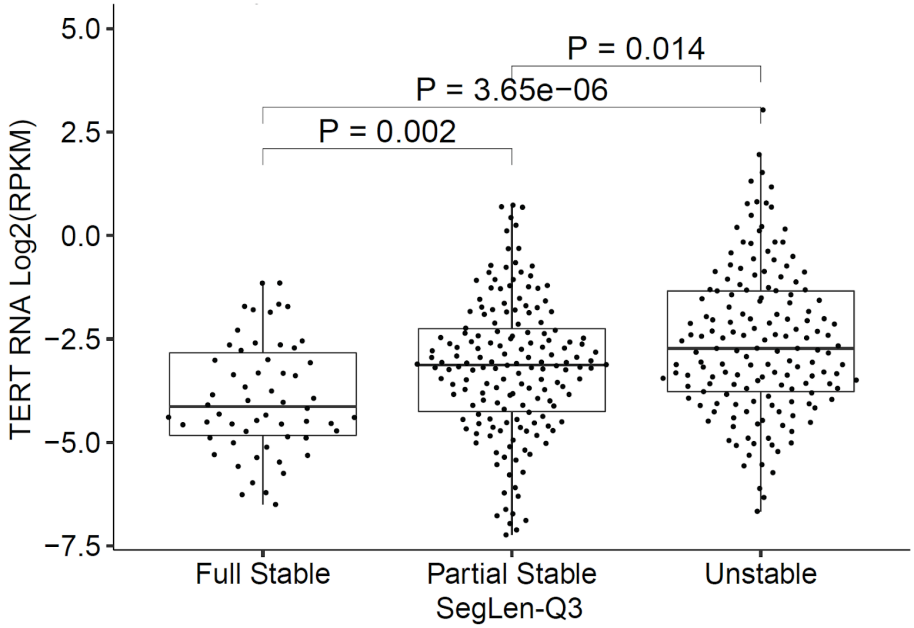


Figure 2.10: RNA expression of TERT gene stratified by genome contiguity groups. Each dot represents TERT expression for one sample

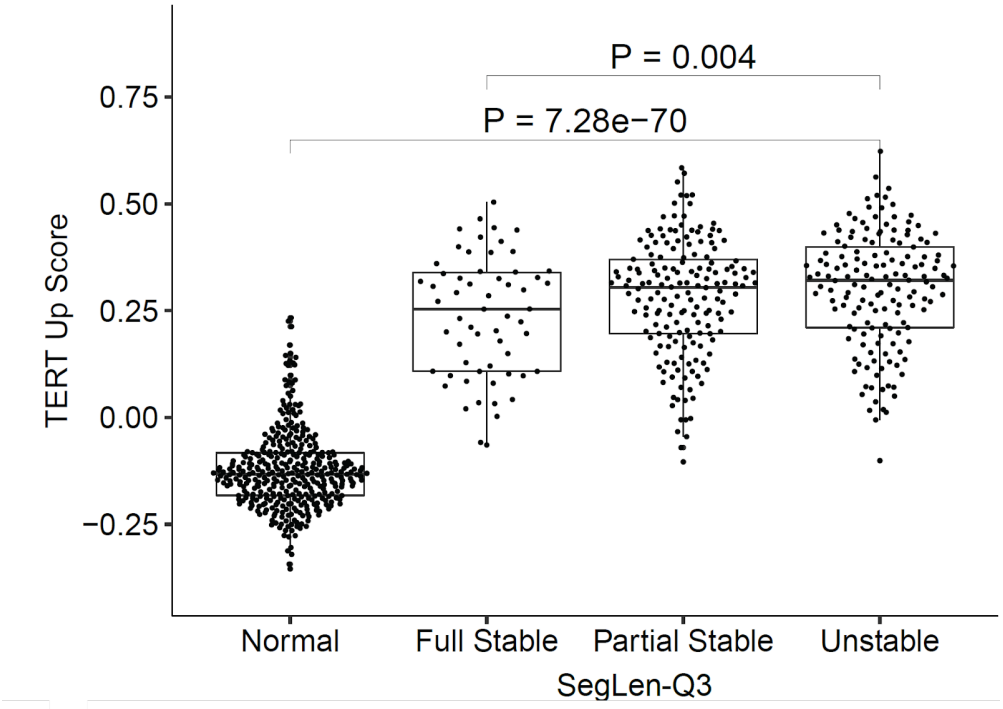


Figure 2.11: TERT downstream activity stratified by genome contiguity groups. As the activity of TERT increases, the downstream activity will increase too.

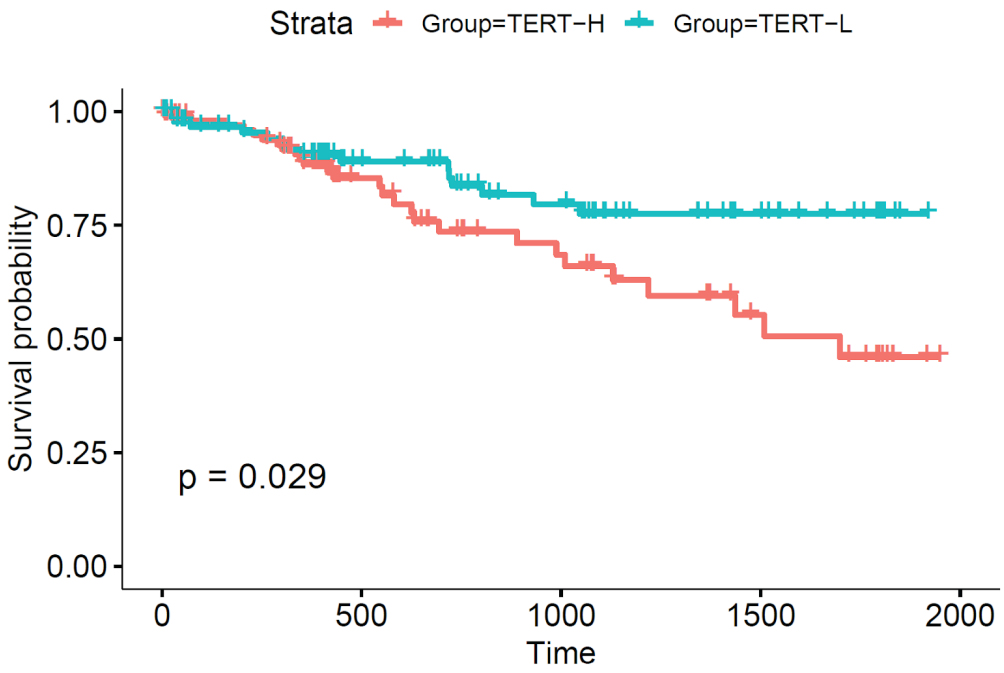


Figure 2.12: Association between TERT expression and overall survival of LUAD patients. Higher TERT is associated with the worst outcome.

Overall, our observations of genome contiguity is consistent with the genetics underpinning of samples with higher fragmentation. In this group, we saw more recurrent amplification, recurrent amplification in known oncogenes such as *MYC* and *TERT*, and finally, significantly higher activity of *TERT* which all are signs of a more advanced disease with worse outcome.

2.2.4 Breakage Intensity Clustering - a novel measure of CIN identifies NKX2-1-dependent and -independent paths to aggressive disease

After establishing the utility of genome fragmentation in predicting overall survival with underlying genetic alterations supporting the progression of cancer in fragmented tumors, we decided to break down the genome contiguity score to its components in order to increase sensitivity of our method. We reasoned that length of genome segments, the factor we measured for genome contiguity score, is dependent on two main factors: number of breakpoints and position of breakpoints. As the number of breakpoints increases, on average we can expect the length of

the segments to be shorter as well (**Figure 2.13** top panel). At the same time, if all the breakpoints are clustered together, we will still be left with many large segments (**Figure 2.13** bottom panel).

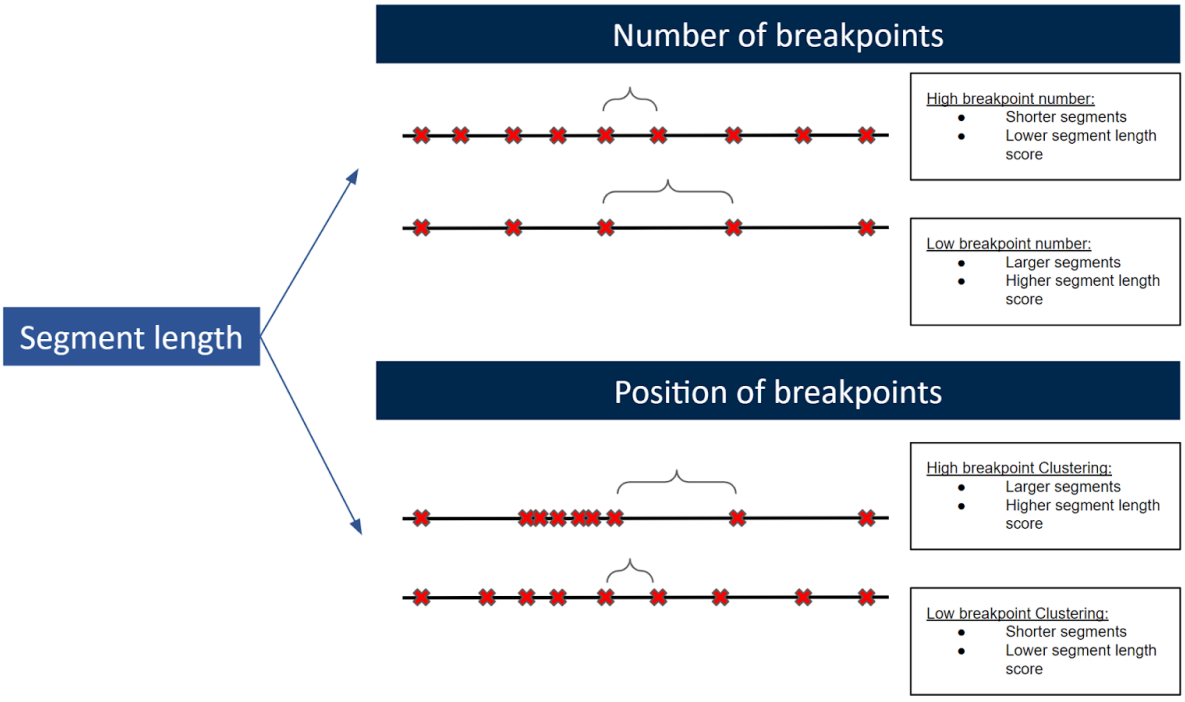


Figure 2.13: Factors impacting length of a segment in genome

Accordingly, we independently estimated the intensity of breakpoints by just counting the number of breakpoints in the genome. For the degree of clustering, we used a spatial point processing method called Ripley's K function [21,22] (**Methods**, **Figure 2.14**). Using these scores, we developed the Breakage Intensity Clustering (BIC) metric.

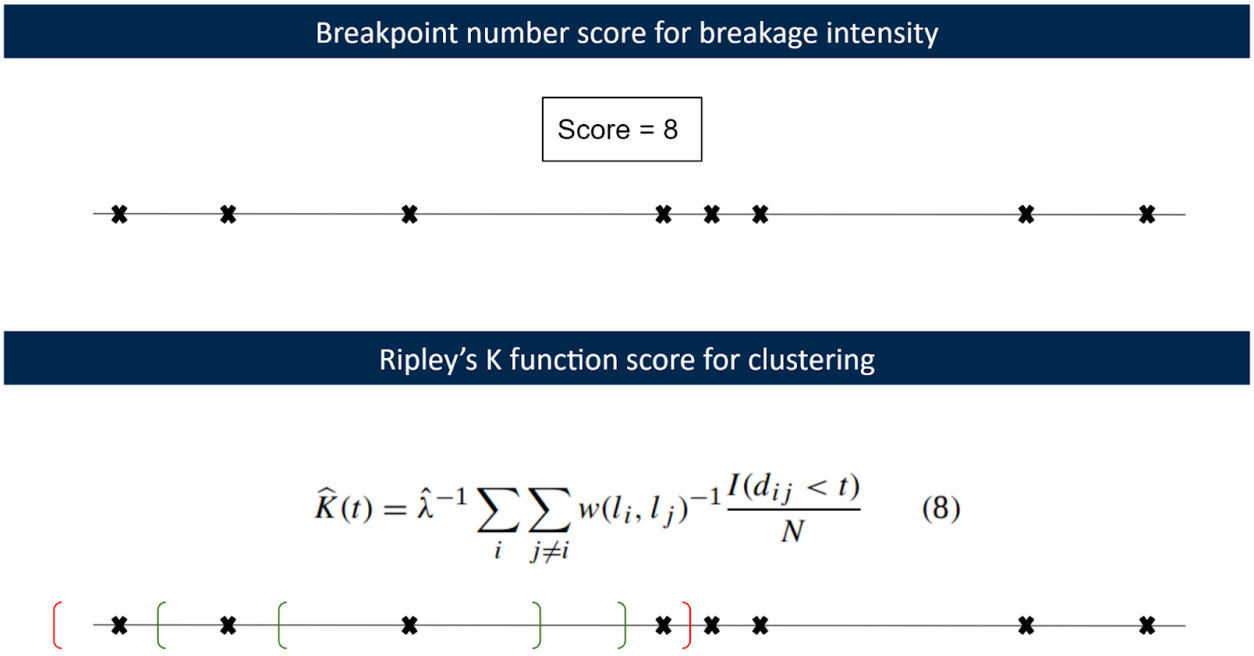


Figure 2.14: Methods used to calculate the positional and intensity aspects of breakpoints, which are impacting length of segments

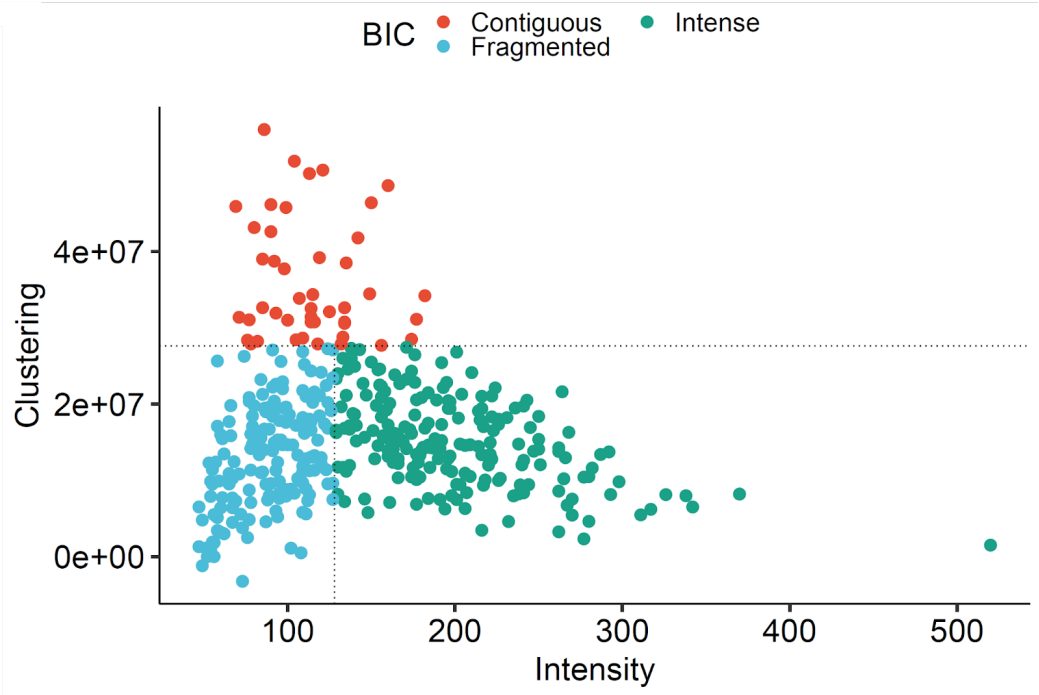


Figure 2.15: Classification of tumors according to their breakpoint intensity (x-axis) and clustering score (Y-axis) to derive BIC groups. Tumors are categorized into three groups: tumors with highly clustered breakpoints (Contiguous), tumors with low clustering and intensity of breakpoints (Fragmented), and tumors with low clustering and high intensity of breakpoints (Intense)

Classification of tumors along these two parameters (**Figure 2.15**) identified patient strata with exceptionally good (BIC-contiguous), typical (BIC-fragmented), and poor (BIC-intense) survival (**Figure 2.16**)

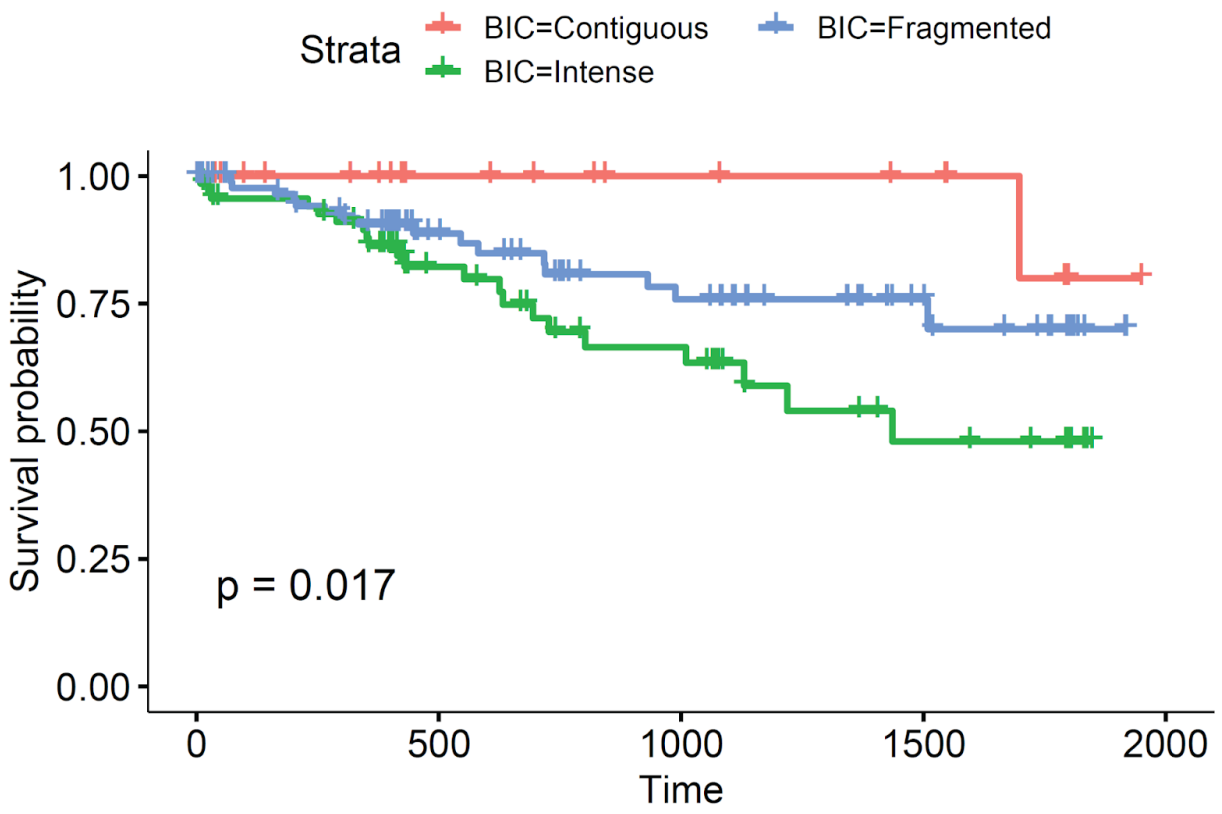


Figure 2.16: Kaplan-Meier survival curves showing the overall survival of patients, stratified by BIC. Tumors with high clustering show exceptionally good survival independent of their breakpoint intensity. In groups with low clustering, high intensity of breakpoints results in worse outcomes.

This metric, similar to genome contiguity score, tracks well with SV burden (**Figure 2.17**) and uniquely identifies tumors where significant breakage is localized to small genomic regions**.**

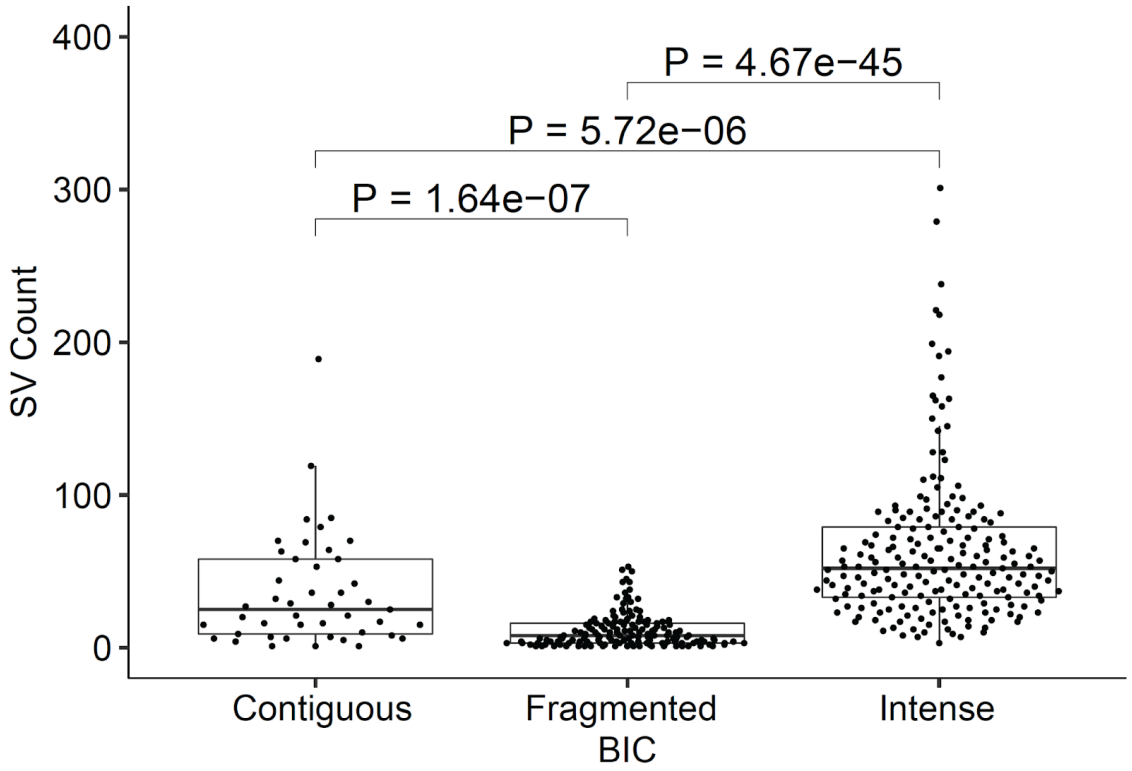


Figure 2.17: SV counts for each sample in BIC groups. P values are based on the Wilcoxon Rank Sum test.

Next, we investigated the *NKX2-1* gene, a lineage specific gene with a complex role in lung adenocarcinoma [23,24]. We noted that focal NKX2-1 amplification was largely restricted to BIC-Intense tumors (10%, P=0.008, **Figure 2.18**). These amplifications were associated with a significant increase in *NKX2-1* RNA (P<1.5e-5) and Protein (P=0.0037) expression (**Figure 2.19**). We further observed complete transcriptional silencing of *NKX2-1* (null expression) in a subset of BIC-intense and BIC-fragmented tumors (**Figure 2.18**). Unexpectedly, both *NKX2-1* amplification (p<0.0014) and silencing (p<2e-5) were associated with higher proliferation (**Figure 2.20**). While tumors with *NKX2-1* amplification had retained their cell identity, tumors that lost their *NKX2-1* expression also lost their alveolar cell identity as measured by inference of transcriptional programs (Method from Zhang et al. [25]) (**Figure 2.21**).

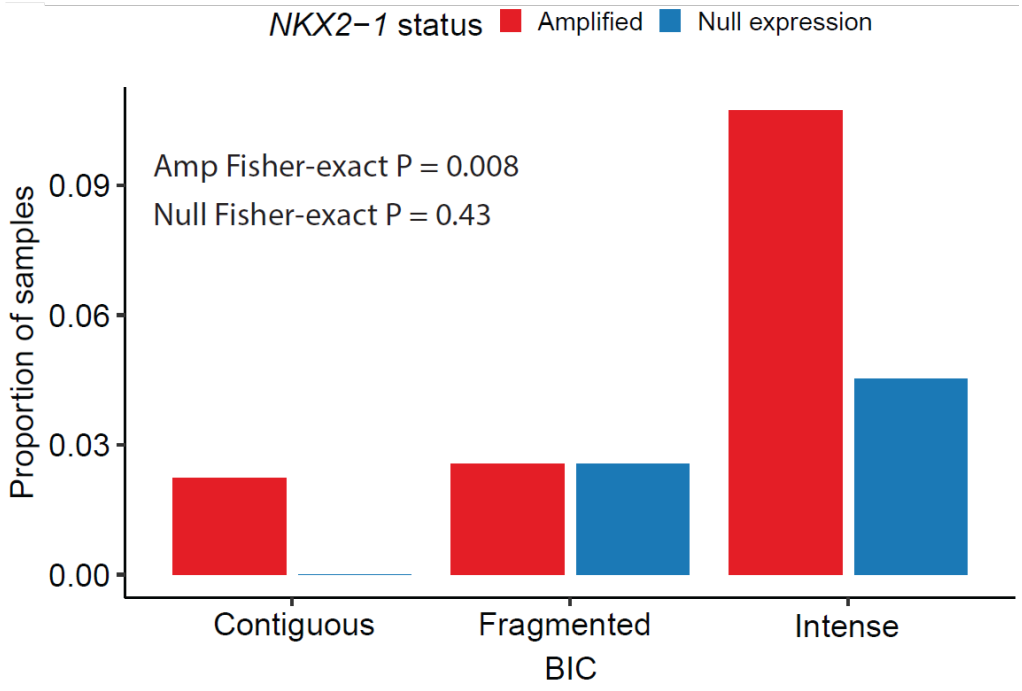


Figure 2.18: The proportion of tumors with NKX2-1 amplification (red) and null expression (blue). Tumors with amplified NKX2-1 show significant enrichment in the BIC-intense group.

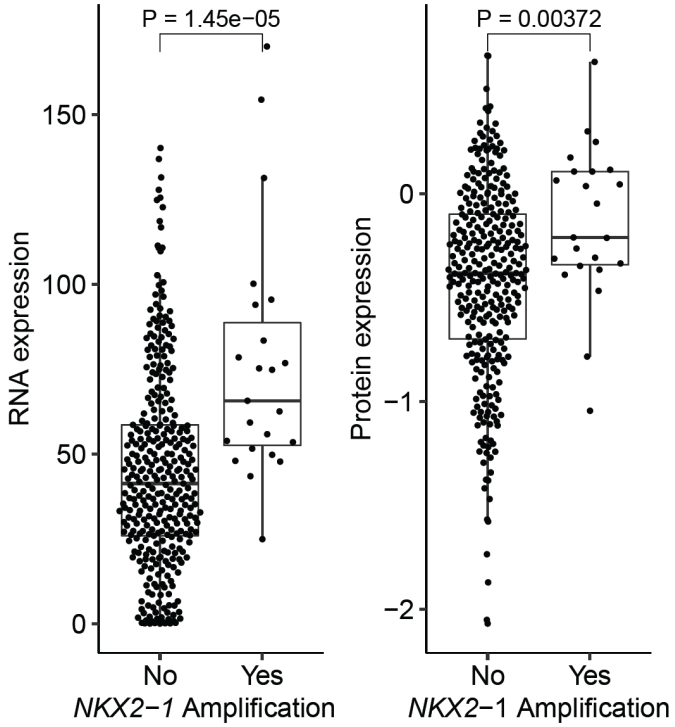


Figure 2.19: RNA and protein expression for each sample (represented by a point) stratified by *NKX2-1* amplification status.

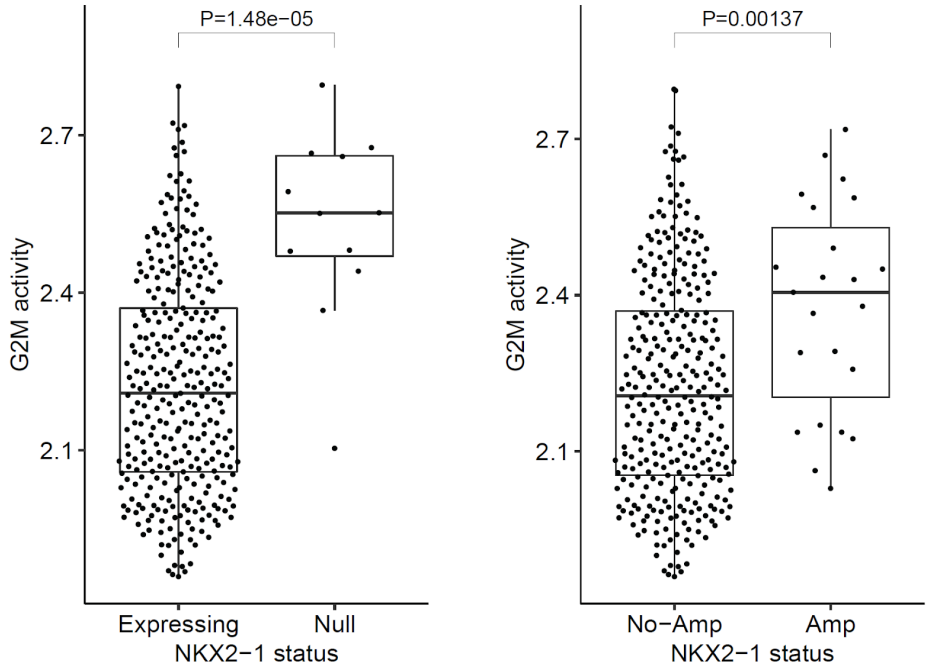


Figure 2.20: Proliferative activity of tumors stratified by NKX2-1 amplification (left) and expression (right) status. Proliferative activity is measured by ssGSEA analysis of RNA expression for the HALLMARK_G2M_CHECKPOINT pathway. P values are based on the Wilcoxon Rank Sum test.

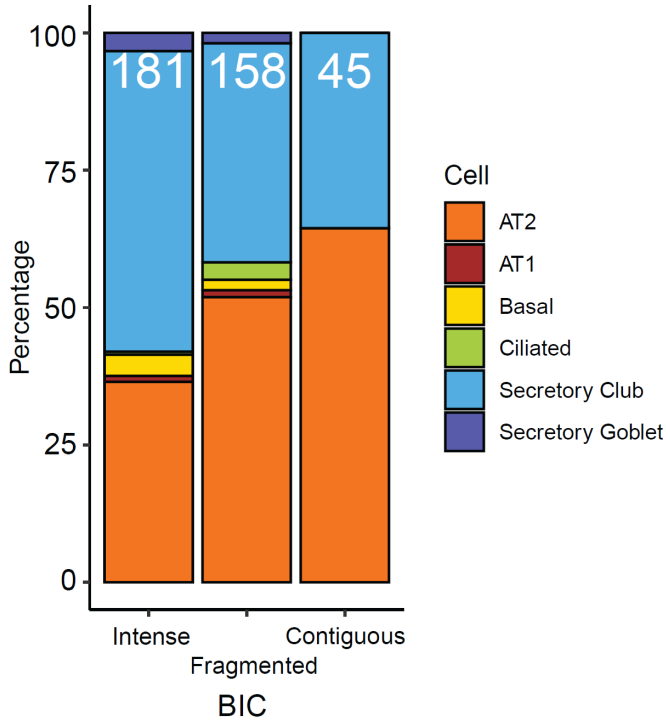


Figure 2.21: Predicted identity of LUAD tumors stratified by BIC status

2.2.5 IGF2BP3 is a robust biomarker of genome fragmentation

Recent therapies targeting dependencies of chromosomally unstable tumors [26,27] provide strong rationale for the development of robust instability biomarkers. Genomic diagnostic methods based on WGS are challenging in the clinical setting due to cost, sample availability, and turnaround time. After an unsuccessful attempt to train deep-learning models to identify instability based on H&E images [28], we pursued proteomic biomarkers as a clinically accessible alternative. By comparing fragmented and contiguous cancer genomes (**Figure 2.22**) we identified bona fide oncogene *IGF2BP3* [29] as a lead candidate biomarker associated with genome instability (FC: 1.50, adj. p<0.009).

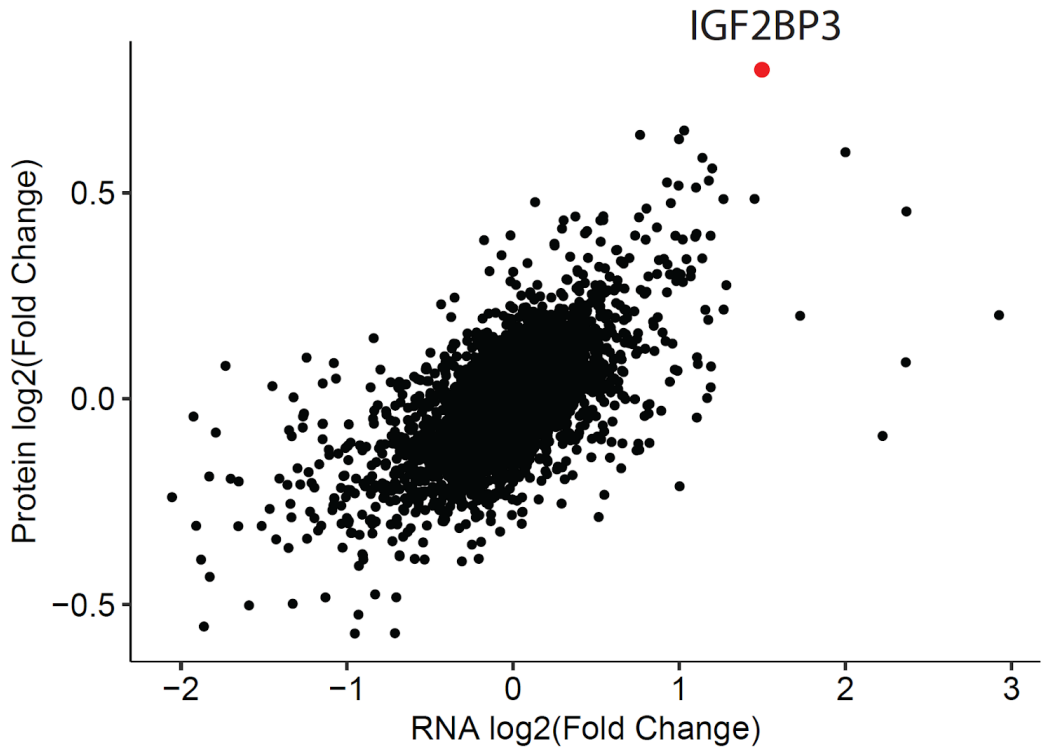


Figure 2.22: Log2 fold changes of genes from differential expression analysis comparing BIC-contiguous to BIC-intense tumors at both RNA (x-axis) and protein (y-axis) levels, showing IGF2BP3 to be highly differentially expressed in the BIC-intense group

We optimized an immunohistochemistry (IHC) assay to detect *IGF2BP3*, which showed varying staining patterns across the range of genome instability levels (**Figure 2.23**) with full stable samples lacking its expression, fragmented samples showing complete or pockets of *IGF2BP3* expression, and finally, majority of Intense samples showed a complete expression of *IGF2BP3*.

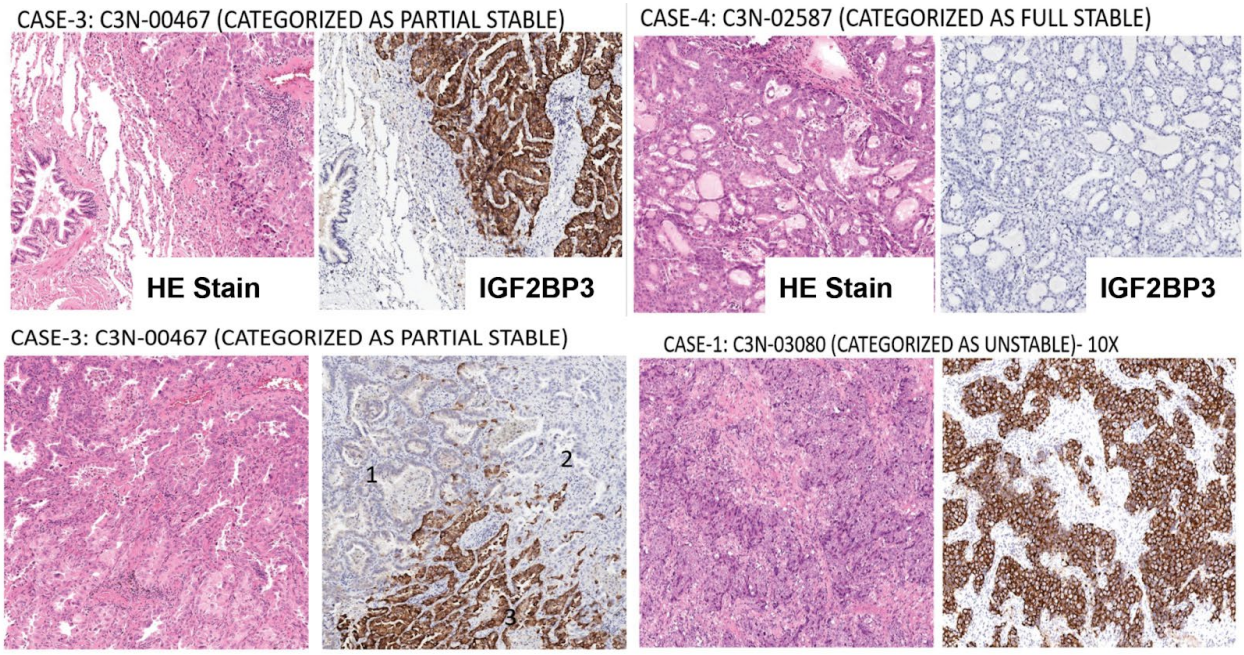


Figure 2.23: Corresponding IGF2BP3 IHC and H&E staining shows heterogeneous protein expression of IGF2BP3 in case-3 (partial stable LUAD), uniformly strong expression in case-1 (unstable), and complete abrogation of IGF2BP3 expression in case-4 (full stable). (IHC; H&E; X=200um)

To confirm these associations, we leveraged a tissue microarray (78 LUAD, 92 LSCC, 69 metastatic LUADs and 60 matched NATs) [30], and examined IGF2BP3 staining through LUAD progression. As expected, expression was absent in normal lung, elevated in a subset of primary LUAD and LSCC tumors, and almost universally high in metastatic LUAD tumors (**Figure 2.24**). Primary LSCC, which are generally more unstable than adenocarcinomas, showed significantly higher IGF2BP3 expression than primary LUAD (p=0.04).

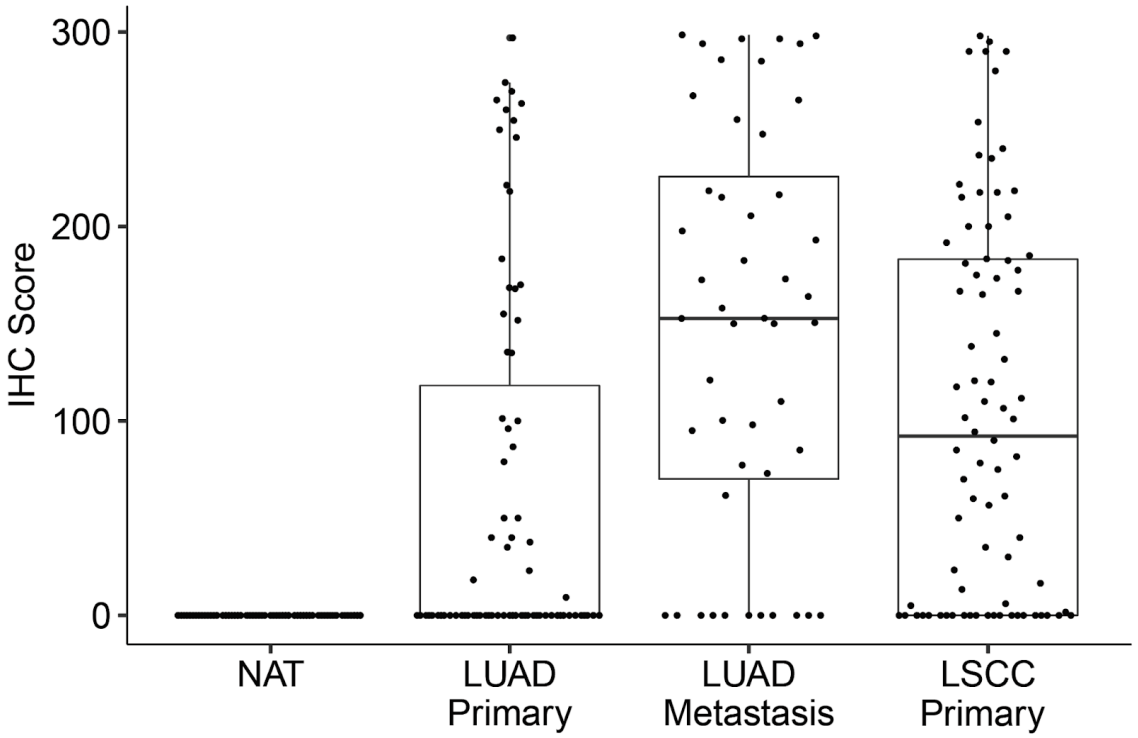


Figure 2.24: Independent validation in 3 patient TMAs reveals higher IGF2BP3 expression in metastatic LUAD compared to primary LUAD and primary LSCC. Also noteworthy is the absence of IGF2BP3 in benign lung parenchyma cores.

Chromosomally unstable tumors evade anti-tumor immunity through a plethora of mechanisms[31]. We therefore examined whether *IGF2BP3* could serve as a predictive biomarker for immune checkpoint inhibitor (ICI) therapy targeting PDL-1 (MPDL3280A, atezolizumab). In an independent cohort[32], tumors with high *IGF2BP3* expression displayed worse outcomes on ICI (**Figure 2.25**, p=0.013), but not against docetaxel (**Figure 2.26**, p=0.2), despite unstable tumors (BIC-fragmented and intense) having levels of PD-L1 expression comparable to or higher than stable (BIC-contiguous) ones (**Figure 2.27**).

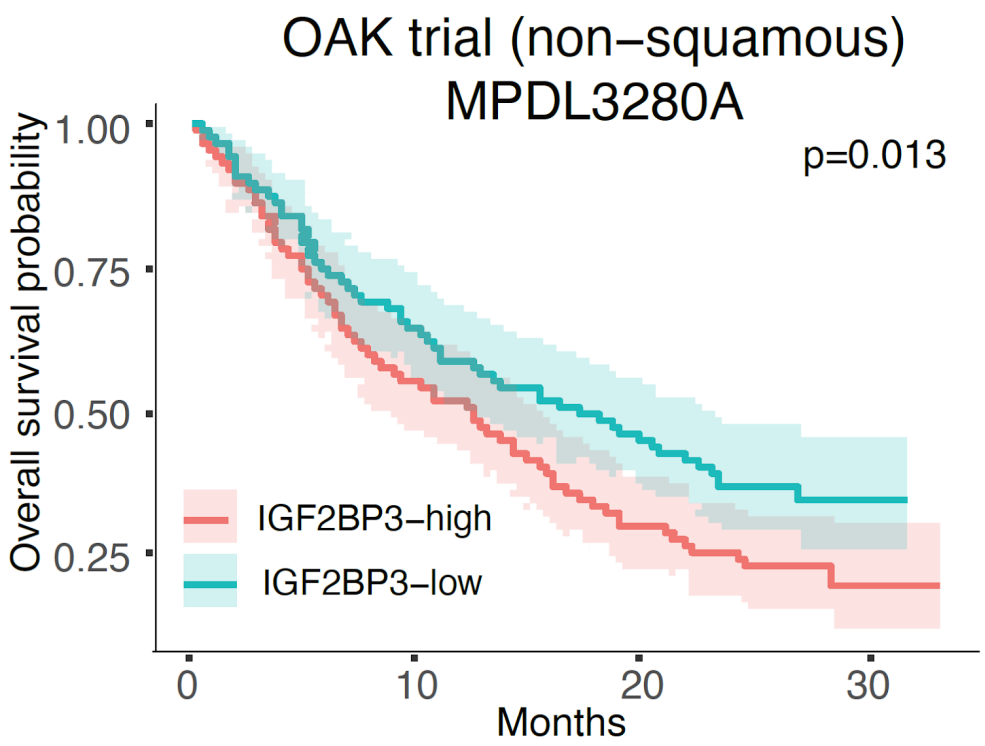


Figure 2.25: Kaplan-Meier survival curves showing the overall survival of non-squamous patients treated with MPDL3280A stratified by IGF2BP3 RNA expression in the OAK trial.

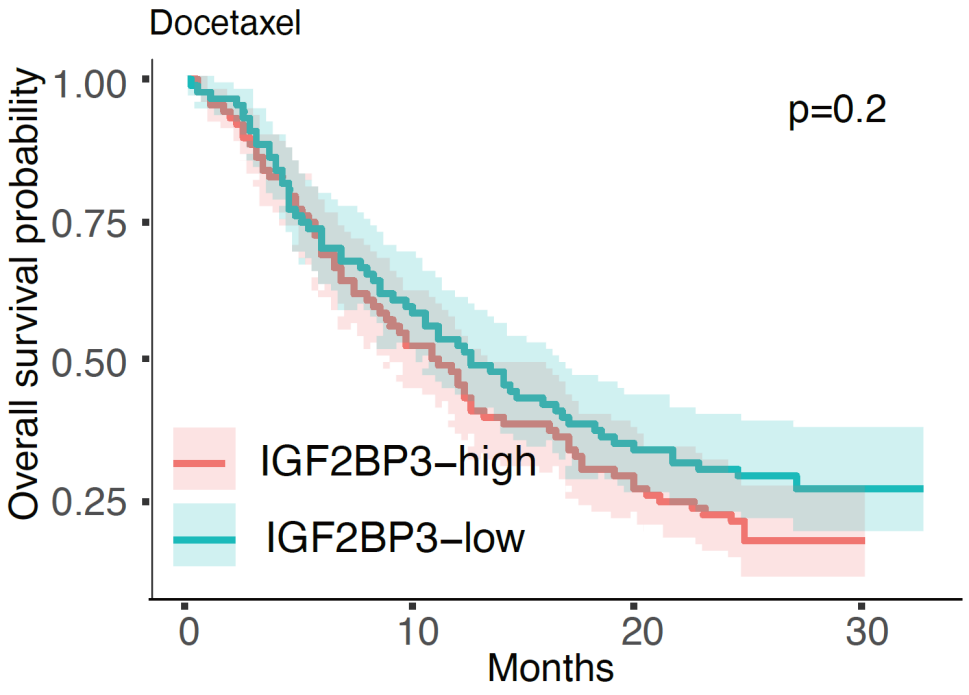


Figure 2.26: Kaplan-Meier survival curves showing overall survival of non-squamous patients treated with docetaxel stratified by IGF2BP3 RNA expression in the OAK trial.

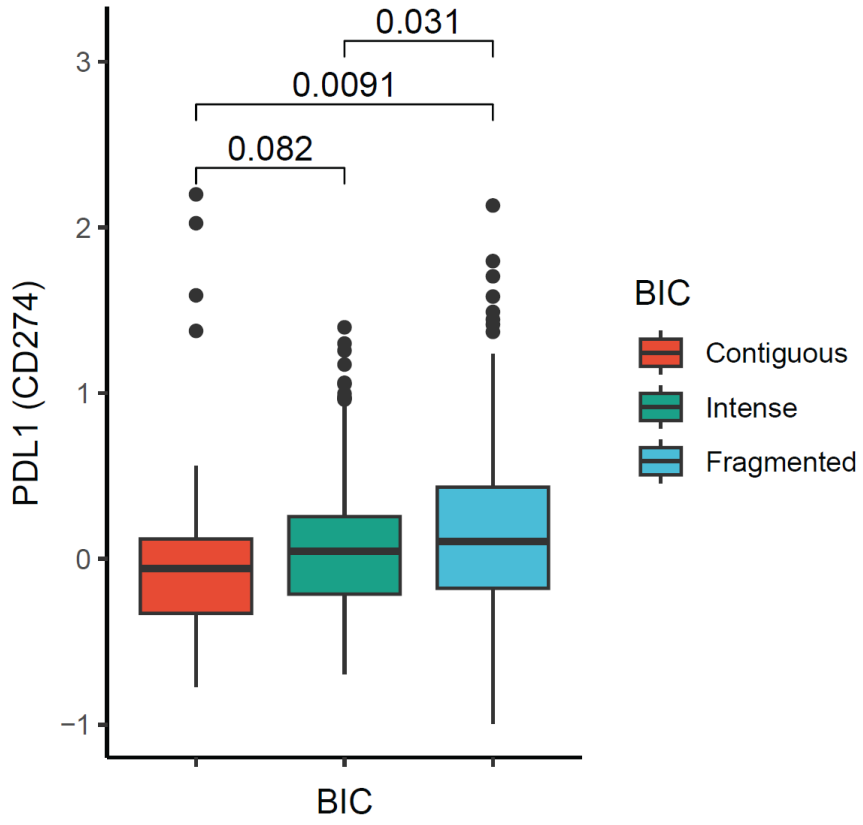


Figure 2.27: Boxplot showing PD-L1 (CD274) protein expression stratified by BIC status.

The high expression of *IGF2BP3* associated with genome instability appears to be triggered by loss of promoter methylation (**Figure 2.28**). BIC-intense tumors were only modestly associated with clinical stage or smoking status (**Figure 2.29**), but they were characterized by increased frequency of *TP53* and *EGFR* mutations, whole-genome doubling (WGD), *TERT* amplifications, and they had fewer KRAS mutations (**Figure 2.30**).

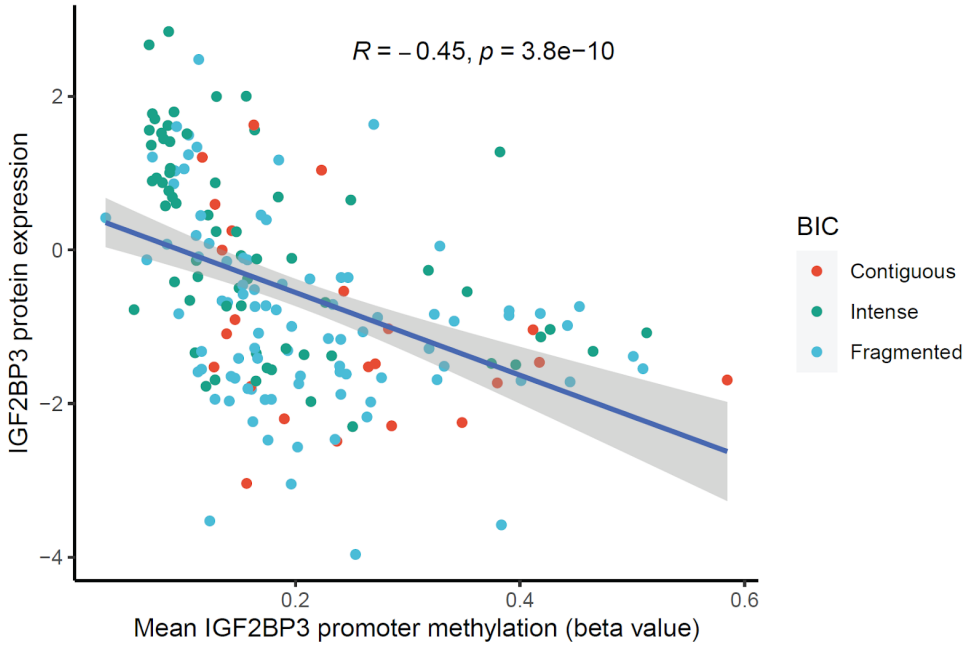


Figure 2.28: Scatter plot showing the correlation of IGF2BP3 protein expression and average IGF2BP3 promoter methylation

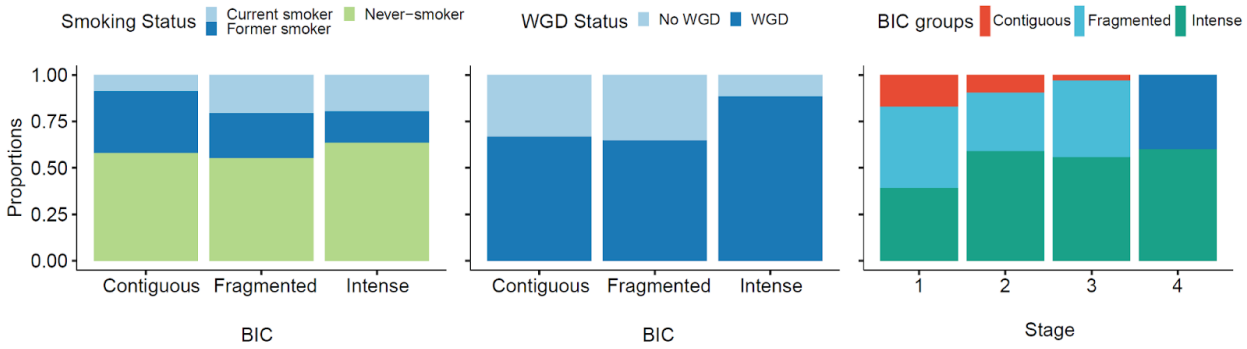


Figure 2.29: Association of BIC groups with smoking status (left), tumor WGD status (middle), and tumor stage (right).

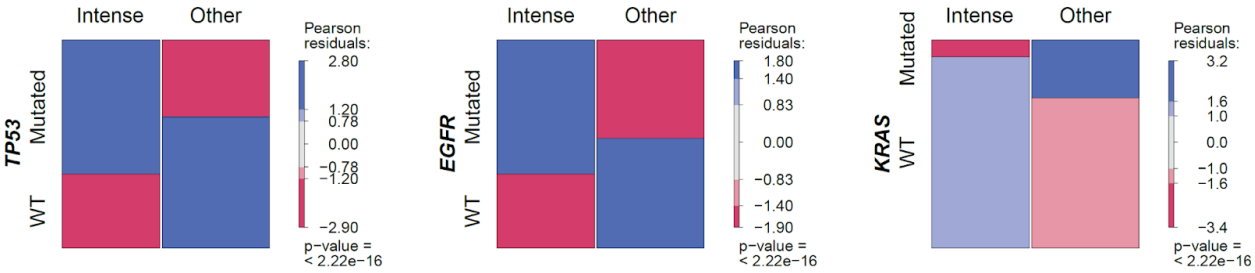


Figure 2.30: Association of the most aggressive, BIC-intense group, with mutations in TP53 (left), EGFR (middle) and KRAS (right). Each column shows all samples in each instability group. The block size in each column shows the proportion of patients with mutated (top) and WT (bottom) status.

2.3 Discussion

Lung cancer is a pressing global medical problem, demanding urgent international action due to its continued prevalence, lethality, and rapidly shifting epidemiology. While successful public health policies promoting smoking cessation have decreased incidence rates especially in males across many countries, rising rates of Lung cancer among never-smokers in and females globally and the growing impacts of environmental pollution continued to make lung cancer one of the main causes of cancer mortality [33].

In the presented chapter, the product of the international and interdisciplinary CPTAC/ICPC program, we analyzed the largest and most comprehensive molecular data resource for LUAD to date, a proteogenomic compendium representing the demographic, etiological, clinical, and molecular complexity of LUAD, addressing key aspects of chromosomal instability, from identification of a method to capture CIN, to derived biomarkers and underlying causes of instability in LUAD.

In this chapter, we focus on CIN, a promising genomic characteristic with potential to reveal therapeutic targets for previously considered undruggable conditions [26]. We have presented several methodological and analytical innovations for the study of CIN in LUAD, and showed the utility of these innovative methods in other cancer types. The heterogeneity of genomic instability led us to develop a novel synthetic metric designated “breakage intensity clustering”, a novel and uniquely prognostic measure of chromosomal instability that characterizes the positional distribution of genomic breakpoints and identifies highly-shattered genomes associated with poor outcomes. We established this measure to be associated with underlying genomic events such as higher rate of NKX2-1 amplification, as well as transcriptional silencing of NKX2-1. These results seem to be contradictory as amplification of a gene usually results in overexpression of that gene. One explanation for this observation can be the dual-role of NXK2-1 in lung cancers, as both an

oncogene and tumor suppressor gene as explained by Yamaguchi et al. [34]. We found while both amplification and silencing of NKX2-1 leads to more aggressive tumors evident by upregulation of G2M pathway, one major difference of the two paths to aggressiveness, is that amplified NKX2-1 samples maintain their cell identity, while the samples with lost expression of NKX2-1 seems to lose their lineage identity and have higher squamous cell markers.

Another underlying genetic event associated with instability in LUAD was the amplification of *TERT* in unstable tumors. *TERT* is a silenced gene in epithelial cells in adults [35]. It has been mentioned previously that TERT might be activated to maintain the length of telomeres in tumor cells which are frequently going through division, a process that can reduce the length of telomeres [19]. On the other hand, presence of telomeres is essential in maintaining genome stability, hence it is possible that when the genome becomes unstable, TERT's activation becomes essential to keep the genome stable to a viable degree [36]. It also points toward a possible vulnerability to unstable tumors that can be used against them in a clinical aspect.

Given the possible clinical use of instability for identifying tumor vulnerabilities, it is essential to have validated biomarkers for tumor instability. We identified *IGF2BP3* as a novel prognostic proteomic biomarker of chromosomally unstable, unfavorable lung cancer, and a panel of biomarkers that might facilitate the development of effective chemopreventive and monitoring strategies for environmental carcinogens that could be used to broaden lung cancer screening to other high risk populations left out of current CT screening guidelines.

Like other studies of this type, the main chief purpose of the study presented in this chapter is to be hypothesis-generating and to provide a resource to the community of basic science and clinical investigators. Additional studies will be needed to validate biological findings, and further investigation into the functional roles of identified biomarkers and therapeutic targets, such as

IGF2BP3. Further, we hope other studies will adapt the CIN calculating methods presented in this chapter and establish their utility at a pan-cancer level and not just LUAD.

2.4 Methods

2.4.1 Identifying regions of recurrent amplification

To identify significantly focally amplified or deleted as well as arm level copy number alterations (CNA) we applied the Genomic Identification of Significant Targets in Cancer (GISTIC2.0) algorithm[37]. Events with q value <0.25 were deemed significant.

2.4.2 Instability measure according to CNA segment length

To calculate the Genome contiguity measure of CIN, the lengths of CNA segments were determined for each tumor sample. The 75th percentile of the segment lengths was selected as the CIN measure for each tumor, as this percentile effectively captures the population of 'large segments' present within a tumor sample. Alternative percentiles, such as the median and 25th percentile, were also evaluated but did not demonstrate any significant biological relevance. In particular, the median segment length was deemed unsuitable due to the observation of binomial segment length distributions in numerous LUAD tumors. Consequently, the median size fails to accurately reflect the presence of either large or short segments within these tumor samples. The 75th percentile, in contrast, provides a more robust and biologically meaningful measure of CIN by focusing on the larger segments that are likely to have a greater impact on tumor biology and hence, was selected for the CIN measure.

2.4.3 Breakage Intensity-Clustering (BIC)

To further break down the instability measure based on segment length, we developed BIC, a CIN measure based on properties of ‘CNV segment breakpoints’ (called breakpoints hereinafter).

This measure focuses on the influence of breakpoints on segment length distribution through their positional distribution (clustering or dispersion) and occurrence rate. If breakpoints occur at a high frequency, the 75th percentile length of the segments (SegLen-Q3) will decrease. Conversely, if breakpoints are clustered together, the distribution of segment lengths will be skewed, affecting the SegLen-Q3 measure differently than if breakpoints were uniformly distributed along the genome

We measured the rate of breakpoints by counting the number of detected breakpoints per tumor. To measure the breakpoints clustering, we used an adapted version of the K-ripley [21] score.

K-ripley score (Kscore) can be defined for any radius of t, where Kscore will reflect the expected degree of clustering at a radius t of a randomly chosen breakpoint on the genome. We defined one dimensional Kscore as:

$$K(t) = \lambda^{-1} \sum_{i} \sum_{i \neq j} W^{-1}(l_i, l_j) I(d_{ij} < t)$$

Where λ is the rate of breakpoints, I is the identity function and d_{ij} is the distance between the ith and jth breakpoints. W function is an 'edge-correction' function. When analyzing a specific radius t, it is crucial to consider cases where the distance of breakpoint i to the start of the genome is less than radius t. In such instances, the score may underestimate the number of breakpoints at the radius t. This underestimation occurs because, for the distance beyond the edge of the genome, no breakpoints are counted, although theoretically, breakpoints could have been present if the ith point was located at a distance greater than t from the edge of the genome. To address this issue, we introduced the W function, which is defined as 1 when t does not extend beyond the edge of the genome. In cases where t extends beyond the edge of the genome, W represents the proportion of radius t that falls within the genome length. By employing this approach, we assume that the length

extending beyond the genome edge has the same expected number of points as the region within the genome.

In this study, we calculated K_{tumor} for each tumor at the radius of 5Mb. If the radius is too small we will be unable to capture focal clusters and if it is too large we will have to constantly correct for the radius going over the edge. We selected 5Mb as we empirically find it to best represent the focal clustering of breakpoints without going over the edge of chromosomes. For each tumor, we also calculated K_{null} which is Kscore for a random distribution of points (from a homogeneous Gaussian process) with the same value as the tumor. Our final measure of clustering is defined as $ClustScore = K_{tumor} - K_{null}$.

Finally, we used both ClustScore and Intensity score of breakpoints to group tumors into 3 main BIC groups: (I) 'BIC-contiguous' for tumors with high ClustScore, (II) 'BIC-fragmented' for any tumor with low ClustScore and low Intensity score and, (III) 'BIC-intense' for any tumor with low clustering but high Intensity of breakpoints.

2.5 References

1. Fenech M, Morley AA. Measurement of micronuclei in lymphocytes. Mutat Res. 1985;147: 29–36.
2. Becker P, Scherthan H, Zankl H. Use of a centromere-specific DNA probe (p82H) in nonisotopic in situ hybridization for classification of micronuclei. Genes Chromosomes Cancer. 1990;2: 59–62.
3. Eastmond DA. Induction of micronuclei and aneuploidy by the quinone-forming agents benzene and o-phenylphenol. Toxicol Lett. 1993;67: 105–118.
4. Boei JJ, Balajee AS, de Boer P, Rens W, Aten JA, Mullenders LH, et al. Construction of mouse chromosome-specific DNA libraries and their use for the detection of X-ray-induced aberrations. Int J Radiat Biol. 1994;65: 583–590.
5. Breneman JW, Ramsey MJ, Lee DA, Eveleth GG, Minkler JL, Tucker JD. The development of chromosome-specific composite DNA probes for the mouse and their application to chromosome painting. Chromosoma. 1993;102: 591–598.
6. Godek KM, Compton DA. Quantitative methods to measure aneuploidy and chromosomal instability. Methods Cell Biol. 2018;144: 15–32.

7. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, et al. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol. 2007;8: R215.
8. Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller M-C, Shaikh N, et al. Replication stress links structural and numerical cancer chromosomal instability. Nature. 2013;494: 492–496.
9. Dewhurst SM, McGranahan N, Burrell RA, Rowan AJ, Grönroos E, Endesfelder D, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov. 2014;4: 175–185.
10. Bakhoum SF, Ngo B, Laughney AM, Cavallo J-A, Murphy CJ, Ly P, et al. Chromosomal instability drives metastasis through a cytosolic DNA response. Nature. 2018;553: 467–472.
11. Correa AF, Ruth KJ, Al-Saleem T, Pei J, Dulaimi E, Kister D, et al. Overall tumor genomic instability: an important predictor of recurrence-free survival in patients with localized clear cell renal cell carcinoma. Cancer Biol Ther. 2020;21: 424–431.
12. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell. 2020;182: 200–225.e35.
13. Chen Y-J, Roumeliotis TI, Chang Y-H, Chen C-T, Han C-L, Lin M-H, et al. Proteogenomics of non-smoking lung cancer in East Asia delineates molecular signatures of pathogenesis and progression. Cell. 2020;182: 226–244.e17.
14. Li S, Wang W, Yu H, Zhang S, Bi W, Sun S, et al. Characterization of genomic instability-related genes predicts survival and therapeutic response in lung adenocarcinoma. BMC Cancer. 2023;23: 1115.
15. Soltis AR, Bateman NW, Liu J, Nguyen T, Franks TJ, Zhang X, et al. Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. Cell Rep Med. 2022;3: 100819.
16. Huang C, Chen L, Savage SR, Eguez RV, Dou Y, Li Y, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell. 2021;39: 361–379.e16.
17. Li Y, Lih T-SM, Dhanasekaran SM, Mannan R, Chen L, Cieslik M, et al. Histopathologic and proteogenomic heterogeneity reveals features of clear cell renal cell carcinoma aggressiveness. Cancer Cell. 2023;41: 139–163.e17.
18. Fernandez-Garcia I, Ortiz-de-Solorzano C, Montuenga LM. Telomeres and telomerase in lung cancer. J Thorac Oncol. 2008;3: 1085–1088.
19. Colebatch AJ, Dobrovic A, Cooper WA. TERT gene: its function and dysregulation in cancer. J Clin Pathol. 2019;72: 281–284.
20. Chiba K, Lorbeer FK, Shain AH, McSwiggen DT, Schruf E, Oh A, et al. Mutations in the promoter of the telomerase gene TERT contribute to tumorigenesis by a two-step mechanism. Science. 2017;357: 1416–1420.

21. Ripley BD. The Second-Order Analysis of Stationary Point Processes. J Appl Probab. 1976;13: 255–266.
22. Dixon PM. Ripley's K function. [cited 20 Jun 2022]. Available: https://www3.nd.edu/~mhaenggi/ee87021/Dixon-K-Function.pdf
23. Pulice JL, Meyerson M. Dosage amplification dictates oncogenic regulation by the NKX2-1 lineage factor in lung adenocarcinoma. bioRxivorg. 2023; 2023.10.26.563996.
24. Heylen E, Verstraete P, Van Aerschot L, Geeraerts SL, Venken T, Timcheva K, et al. Transcription factor NKX2-1 drives serine and glycine synthesis addiction in cancer. Br J Cancer. 2023;128: 1862–1878.
25. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2103240118
26. Bakhoum SF. Targeting the undruggable. Science. 2023;380: 47.
27. Payton M, Belmontes B, Hanestad K, Moriguchi J, Chen K, McCarter JD, et al. Small-molecule inhibition of kinesin KIF18A reveals a mitotic vulnerability enriched in chromosomally unstable cancers. Nat Cancer. 2024;5: 66–84.
28. Wang JM, Hong R, Demicco EG, Tan J, Lazcano R, Moreira AL, et al. Deep learning integrates histopathology and proteogenomics at a pan-cancer level. Cell Rep Med. 2023;4: 101173.
29. Panebianco F, Kelly LM, Liu P, Zhong S, Dacic S, Wang X, et al. THADA fusion is a mechanism of IGF2BP3 activation and IGF1R signaling in thyroid cancer. Proc Natl Acad Sci U S A. 2017;114: 2307–2312.
30. Rodriguez E, Chen L, Ao M-H, Geddes S, Gabrielson E, Askin F, et al. Expression of transcript factors SALL4 and OCT4 in a subset of non-small cell lung carcinomas (NSCLC). Transl Respir Med. 2014;2: 10.
31. Tijhuis AE, Johnson SC, McClelland SE. The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity. Mol Cytogenet. 2019;12: 17.
32. Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet. 2017;389: 255–265.
33. American Lung Association. Lung Cancer Trends Brief: Mortality. [cited 4 Nov 2024]. Available: https://www.lung.org/research/trends-in-lung-disease/lung-cancer-trends-brief/lung-cancer-mortality-(1)
34. Yamaguchi T, Hosono Y, Yanagisawa K, Takahashi T. NKX2-1/TTF-1: an enigmatic oncogene that functions as a double-edged sword for cancer cell survival and progression. Cancer Cell. 2013;23: 718–723.

35. Lee DD, Leão R, Komosa M, Gallo M, Zhang CH, Lipman T, et al. DNA hypermethylation within TERT promoter upregulates TERT expression in cancer. J Clin Invest. 2019;129: 223–229.
36. Maciejowski J, de Lange T. Telomeres in cancer: tumour suppression and genome instability. Nat Rev Mol Cell Biol. 2017;18: 175–186.
37. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12: R41.

Chapter 3 Proteogenomic Analysis of Chemo-Refractory High-Grade Serous Ovarian Cancer

Portions of this chapter have been published at Chowdhury et al [34].

3.1 Introduction

Epithelial ovarian cancer accounts for >185,000 deaths/year worldwide [1]. The most common subtype, high-grade serous ovarian cancer (HGSOC), accounts for 60% of deaths. Despite improvements in surgical and chemotherapeutic approaches, HGSOC mortality has not changed in decades [2]. The 5-year survival rate remains ~30% for the majority of patients [1,2].

Standard of care involves surgical debulking combined with adjuvant or neoadjuvant chemotherapy with carbo- or cisplatin in combination with a taxane [1,3]. At diagnosis, HGSOC is among the most chemo-sensitive of all epithelial malignancies, with initial response rates of ~85%, presumably related to DNA repair defects [4]. Platinum is thought primarily to drive the response rate, due to the lower single-agent response rate for taxanes [5].

10%–20% of HGSOC patients have treatment-refractory disease at diagnosis, fail to respond to initial therapy, and have a dismal prognosis [6]. The poor response to subsequent therapy and median overall survival of ~12 months for these patients has not changed in 40 years [7,8]. Even for the ~85% of tumors that respond to initial chemotherapy, most relapse as resistant disease [9], defined as recurrence within 6 months of completing platinum-based chemotherapy [10]. This chemo-resistant phenotype, which can occur following a single or multiple rounds of therapy, is also associated with poor responses to subsequent therapies [11] and limited survival. Patients whose ovarian cancer has responded with a partial or complete remission are often treated

with PARP inhibitors as maintenance therapy, with the greatest benefit in relapse-free survival observed in cancers with BRCA1 or BRCA2 mutations [12]. These regimens are supported by level 1 evidence from randomized phase III trials [13–16] but unfortunately do not apply to patients with platinum-refractory HGSOC.

Despite >30 years of literature studying platinum resistance in cancer [17], there currently is no way to distinguish refractory from sensitive HGSOCs prior to therapy. Consequently, patients with refractory disease experience the toxicity of platinum-based chemotherapy without benefit. Due to their rapid progression, they are commonly excluded from participating in clinical trials. Accordingly, there is no ongoing clinical research that could identify effective therapeutic agents for these patients or provide insights into molecular mechanisms of refractory disease.

To address this unmet clinical need, we performed proteogenomic analysis of treatment-naive HGSOCs (chemo-sensitive and chemo-refractory) and identified predictive chromosomal events associated with initial response to platinum based treatment in HGSOC. Additionally, we have looked into the underpinning genetic mechanisms of our biomarker which gives biological insight into the general paths to refractoriness in HGSOC.

3.2 Results

3.2.1 Study cohort and proteogenomic profiling

We performed our analysis on pre-treatment biopsies collected from three HGSOC patient cohorts: FFPE (formalin-fixed paraffin-embedded) discovery (n = 158), FFPE validation (n = 20), and frozen validation (n = 64). Biopsies were collected at the time of primary debulking surgery (prior to chemotherapy) and were reasonably balanced between chemo-refractory and sensitive tumors (**Figure 3.1**). Refractory cancers were defined as those that progressed or had stable disease within 6 cycles of initial platinum/taxane therapy after initial debulking surgery. Sensitive tumors

were defined as those that responded to initial platinum/taxane therapy and did not progress within 2 years.

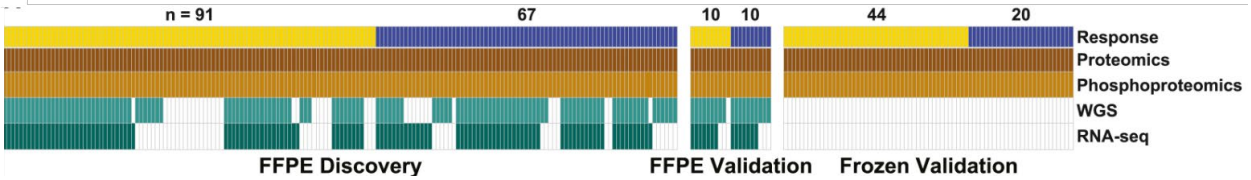


Figure 3.1: Availability of the proteomic, phospho-proteomic, WGS, and RNA-seq data across the cohorts.

The FFPE discovery cohort consisted of 158 biopsies (91 sensitive and 67 refractory) (**Figure 3.1**) sourced from 3 academic centers and representing a wide range of storage times (**Figure 3.2**). 8,800 proteins and 2,648 phosphosites were quantified and observed in >50% of the biospecimens in either sensitive or refractory groups and were included in downstream analyses. Whole-genome sequencing (WGS) and RNA-seq were performed on the FFPE discovery cohort samples. Using mutation and copy-number (CN) variation (CNV) information derived from the WGS data (**Methods**), we delineated sets of likely somatic or germline pathogenic variants ("mutations") and CN aberrations. After imposing strict quality control (QC) (**Methods**), genomic data were available for mutations (n = 120 tumors), CNV (n = 120 tumors), loss of heterozygosity (LOH) (n = 118 tumors), and RNA-seq-based gene expression (n = 106 tumors) (**Figure 3.1**). Similar data was generated for the validation cohort of 10 refractory and 10 sensitive patients, yielding DNA CNV data for 19 out of 20 patients, while for RNA, only 14 out of 20 patients passed QC filters. In addition to the cohorts mentioned above, another cohort, Memorial Sloan Kettering (MSK)-IMPACT [18] was used to validate some of the results.

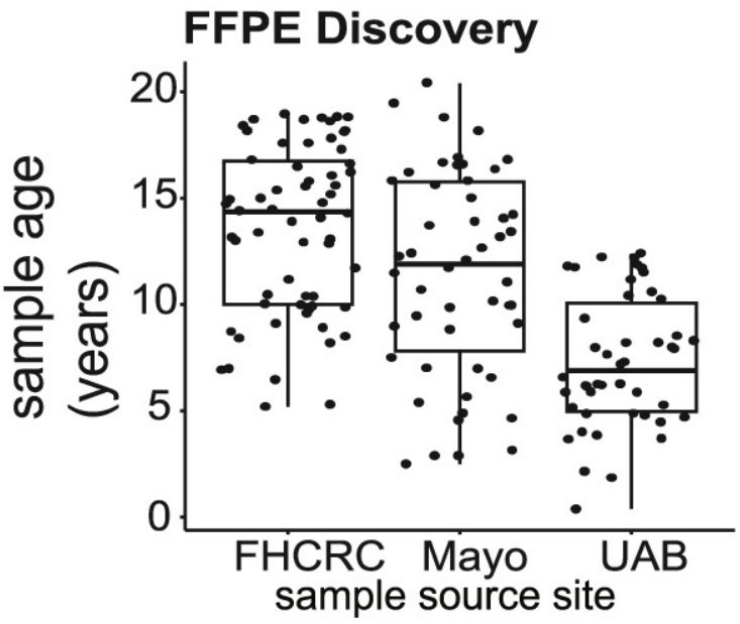


Figure 3.2: Boxplots showing the distribution of sample storage times across three biospecimen source sites: Fred Hutchinson Cancer Research Center (FHCRC), Mayo Clinic, and University of Alabama at Birmingham (UAB). The sample sizes for the three source sites are 62 FHCRC, 50 Mayo and 46 UAB.

3.2.2 Genomic alterations associated with chemo-refractoriness

Consistent with prior reports [19], we found few recurrently mutated genes (FFPE discovery cohort) (**Figure 3.3**). Tumors harbored 1–7 (median = 2) mutations in known tumor suppressor genes, with mutations in DNA damage response genes TP53 (67.5%), BRCA1 (15.8%), BRCA2 (15%), and CDK12 (10%) being among the most common. As expected, mutations in BRCA1 or BRCA2 were more frequent in sensitive (S) vs. refractory (R) tumors ($p = 0.001$) (**Figure 3.4**). This is consistent with previous studies suggesting better survival of patients with BRCA1/2 mutations in response to Platinum based therapy [20]. No other significant associations between simple mutations (including TP53) and response were observed.

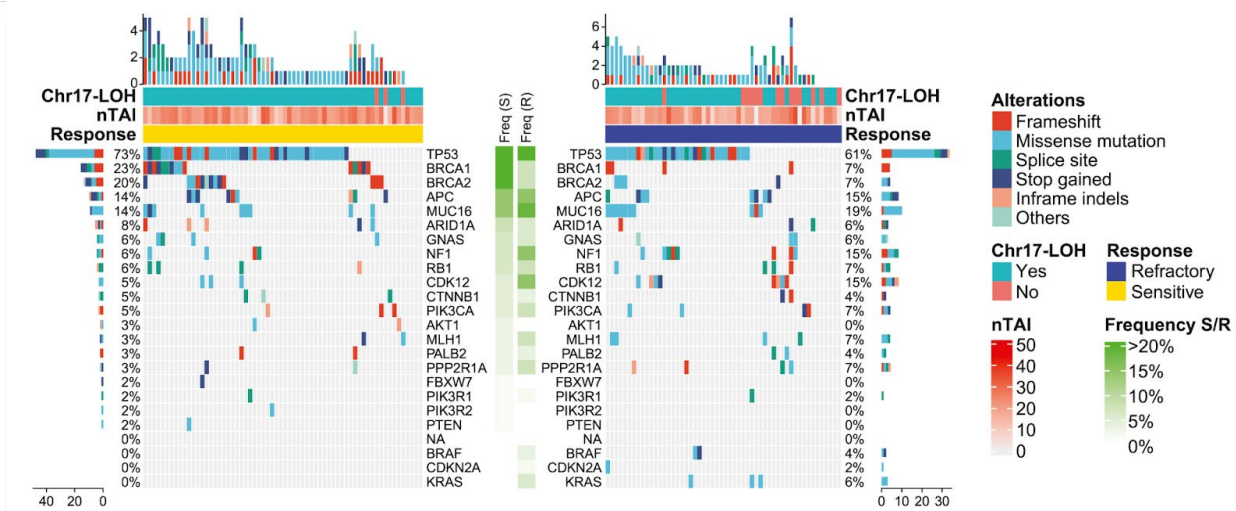


Figure 3.3: OncoPrint of likely pathogenic genetic variants in sensitive and refractory tumors. Genes are ordered by mutation frequency in sensitive tumors. Chr17-LOH binary variable indicates whether more than 25% of Chr17 has LOH. nTAI denotes the number of chromosome arms with telomeric allelic imbalance.

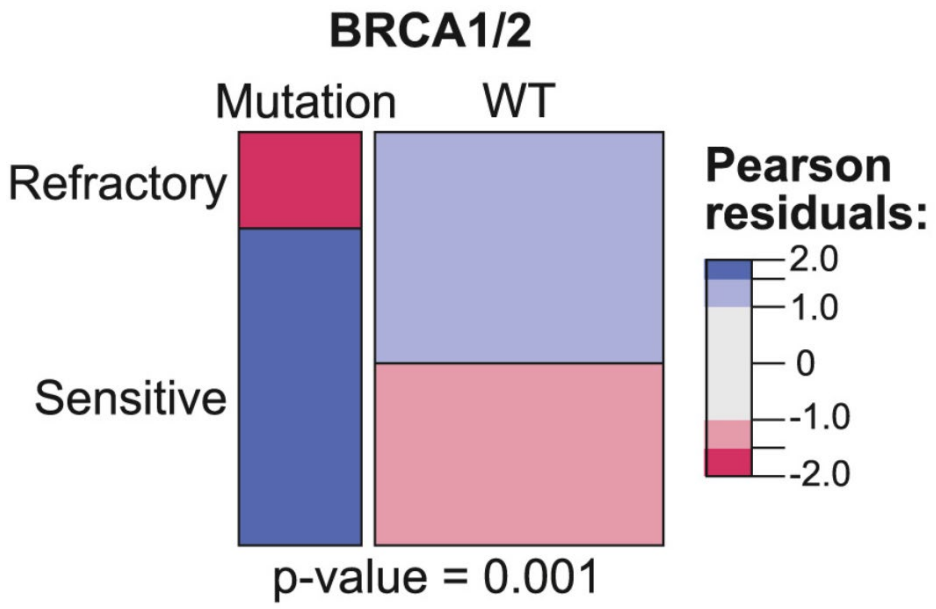


Figure 3.4: Association between mutations in BRCA1/2 and tumor response. The p value is based on the chi-square test.

We observed a lower frequency of TP53 mutations compared with previous studies [19,21], likely because we used an allele frequency cutoff of 11% for calling mutations, which was

considerably higher than that used in The Cancer Genome Atlas (TCGA) [19], resulting in a lower mutation call. We chose this cutoff to focus primarily on clonal mutation (driver) events likely to contribute to tumor biology and clinical behavior.

Because HGSOC is characterized by extensive genomic instability exemplified by segmental and chromosome-level alterations, or even whole-genome duplication, we performed absolute CN analysis (**Methods**). Both sensitive and refractory tumors were remarkably unstable, with similar levels of aneuploidy and polyploidy, as well as broadly similar frequencies and patterns of gains and losses (**Figure 3.5**). Further comparison of local genomic alterations revealed similar patterns of focal chromosomal gains and losses (**Figure 3.6**).

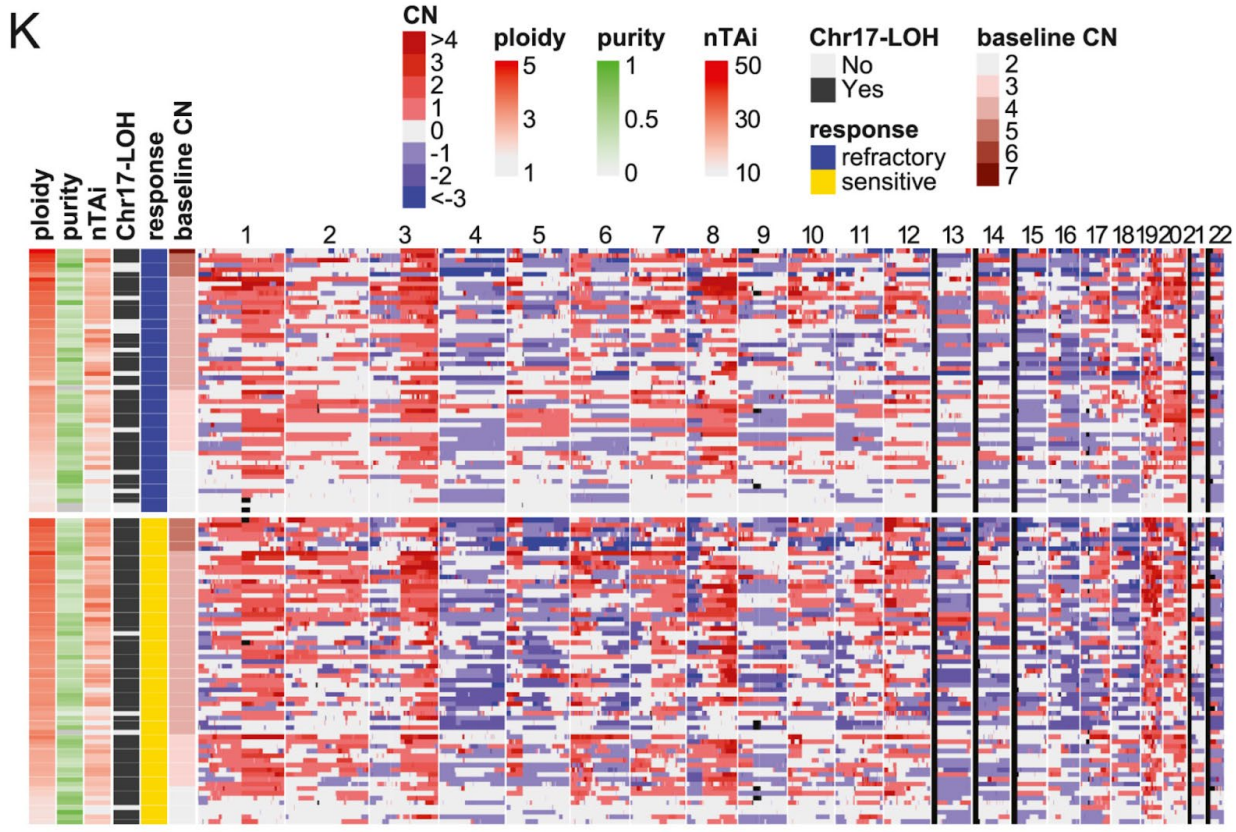


Figure 3.5: Sample-wise plot showing copy-number aberrations in sensitive and refractory tumors. Copy-number (CN) aberrations are quantified as changes relative to the modal ploidy (baseline CN).

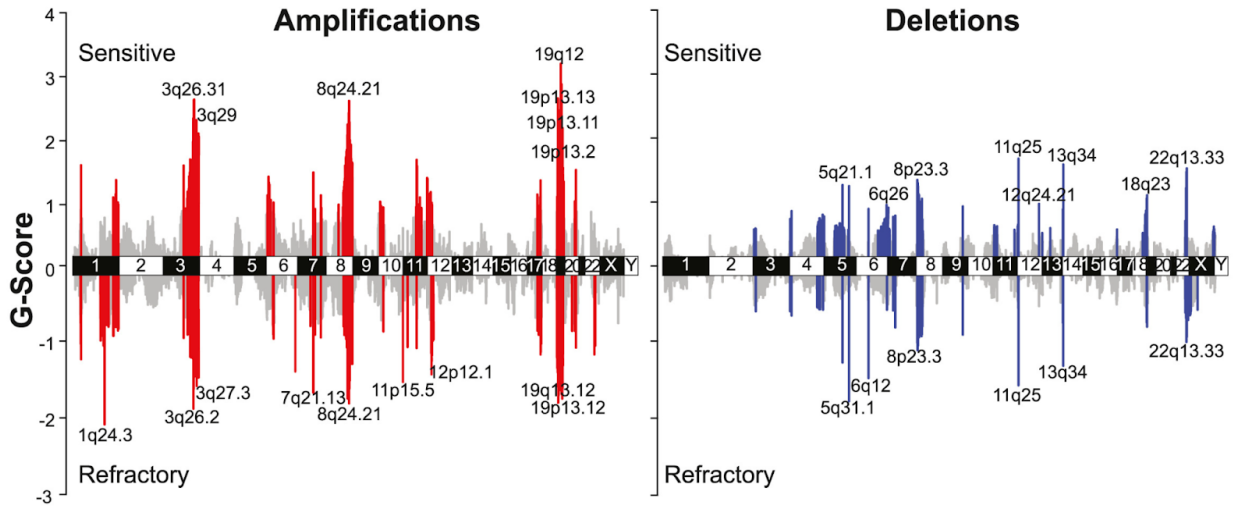


Figure 3.6: Genomic regions with recurrent focal amplifications (left) and deletions (right) in sensitive (top) and refractory (bottom) tumors. The G score is proportional to −log(probability|background). Cytobands of the most significant peaks are annotated.

3.2.3 Chr17-LOH is associated with chemo-sensitivity

Consistent with previous studies [22], we noted significant LOH for most chromosomes (Figure S2B), with chromosome 17 LOH (Chr17-LOH) being the most common (~80% of cases). This was expected as LOH is often considered as a second hit to mutations, and there are many TSG genes on chromosome 17, such as *TP53*, *BRCA1*, *NF1* and *CDK12*, therefore an LOH on a TSG rich chromosome can act as a way for tumors to inactivate the genes.

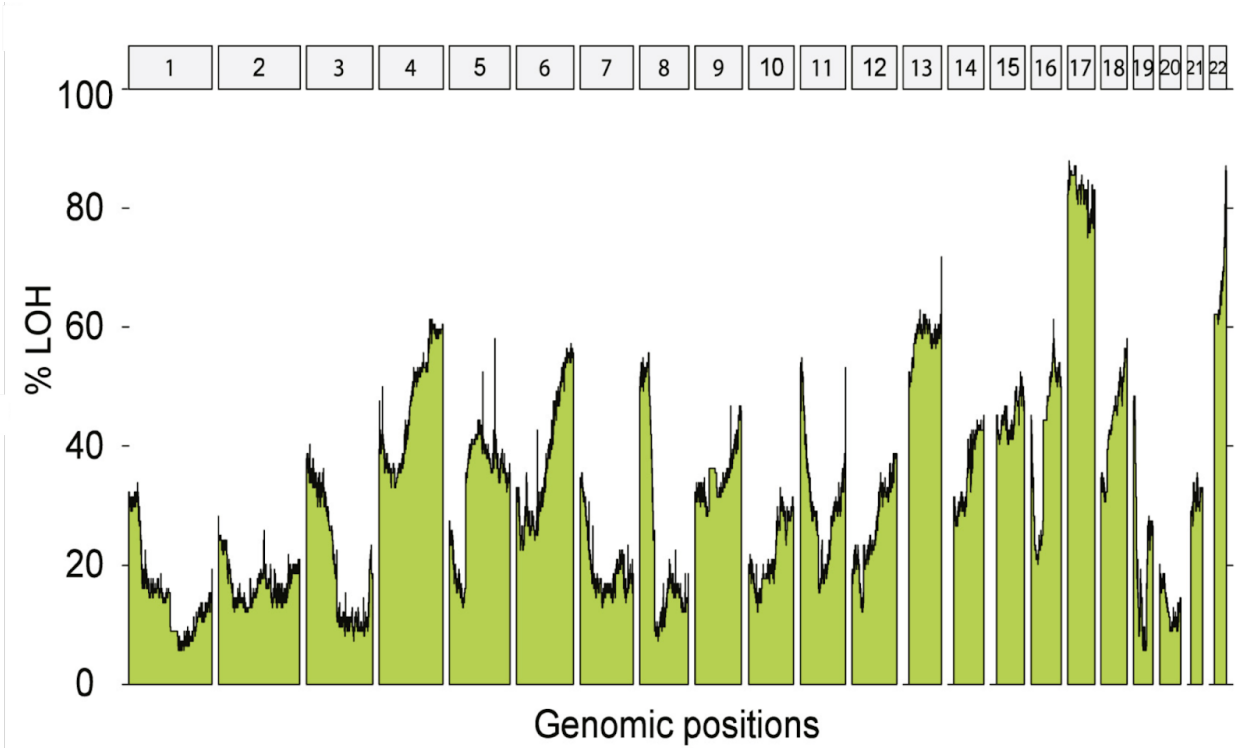


Figure 3.7: Scatter plot showing the proportion of CNV-RNA/protein cascade genes in each pathway in sensitive (x axis) vs. refractory (y axis) tumors. Pathway size indicates the number of genes in the pathway.

Among all LOH events, only Chr17-LOH was significantly (FDR < 0.1) associated with chemo-sensitivity (**Figures 3.8** and **Figure 3.9**). No other chromosome-level gain, loss, or LOH event showed association with treatment response (**Figure 3.8**).

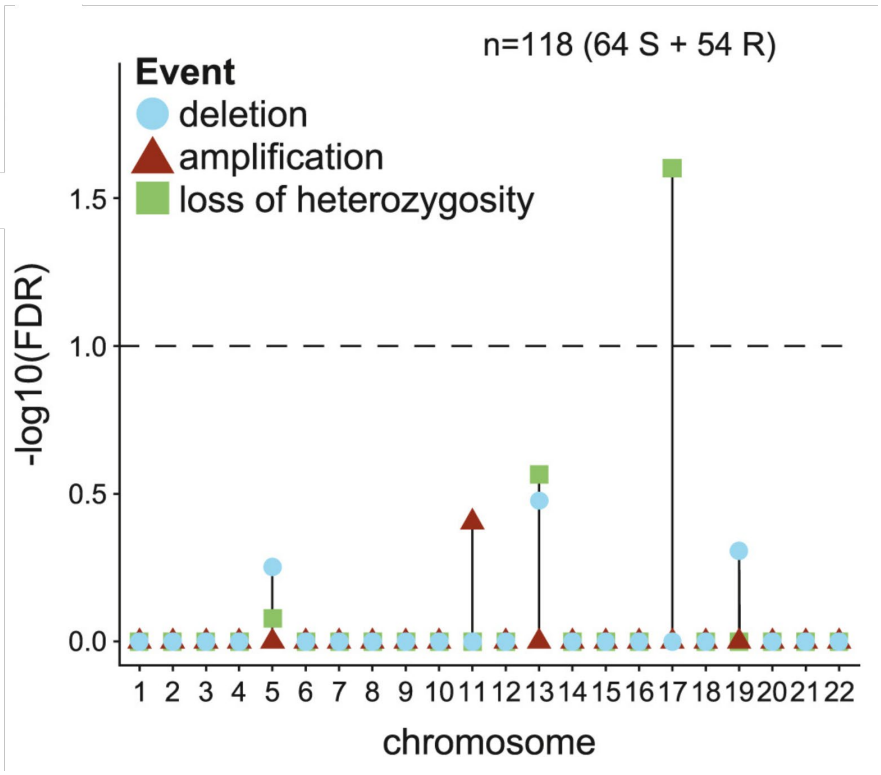


Figure 3.8: Associations between refractoriness and chromosome losses, gains, and LOH. At the chromosome level, only Chr17-LOH is associated with refractoriness (FDR < 0.1).

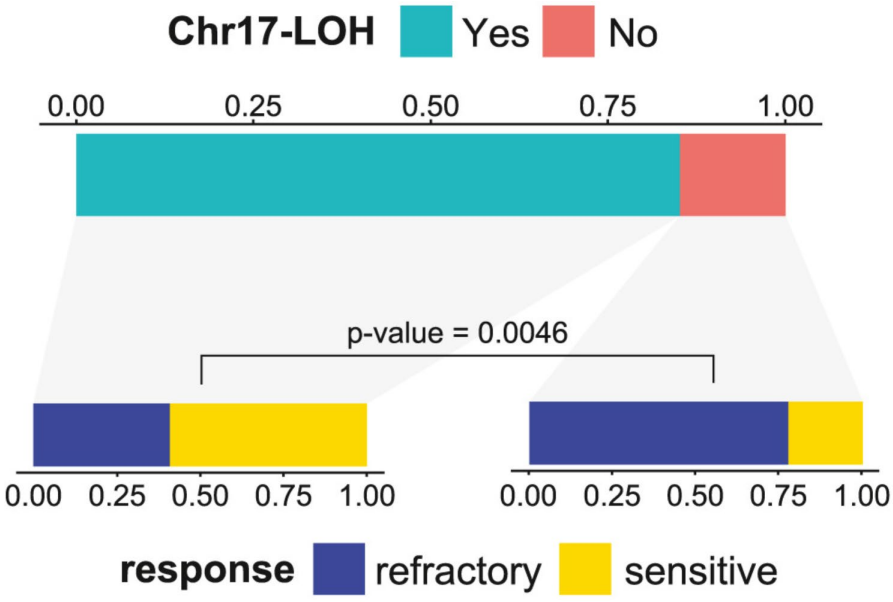


Figure 3.9: Association between Chr17-LOH and refractoriness. Among the 15% of samples without Ch17-LOH, more than 75% are refractory, whereas most samples with Chr17-LOH are sensitive (p value = 0.0046, Fisher's Exact test).

75% of tumors heterozygous for Chr17 (i.e., without Chr17-LOH) were refractory, compared with only ~40% of those with Chr17-LOH (**Figure 3.9**). Consistent with these results, we observed in an independent cohort, MSK-IMPACT that tumors enriched for Chr17-LOH have significantly higher overall survival (**Figure 3.10**, **Methods**), as expected for chemo-responsive disease.

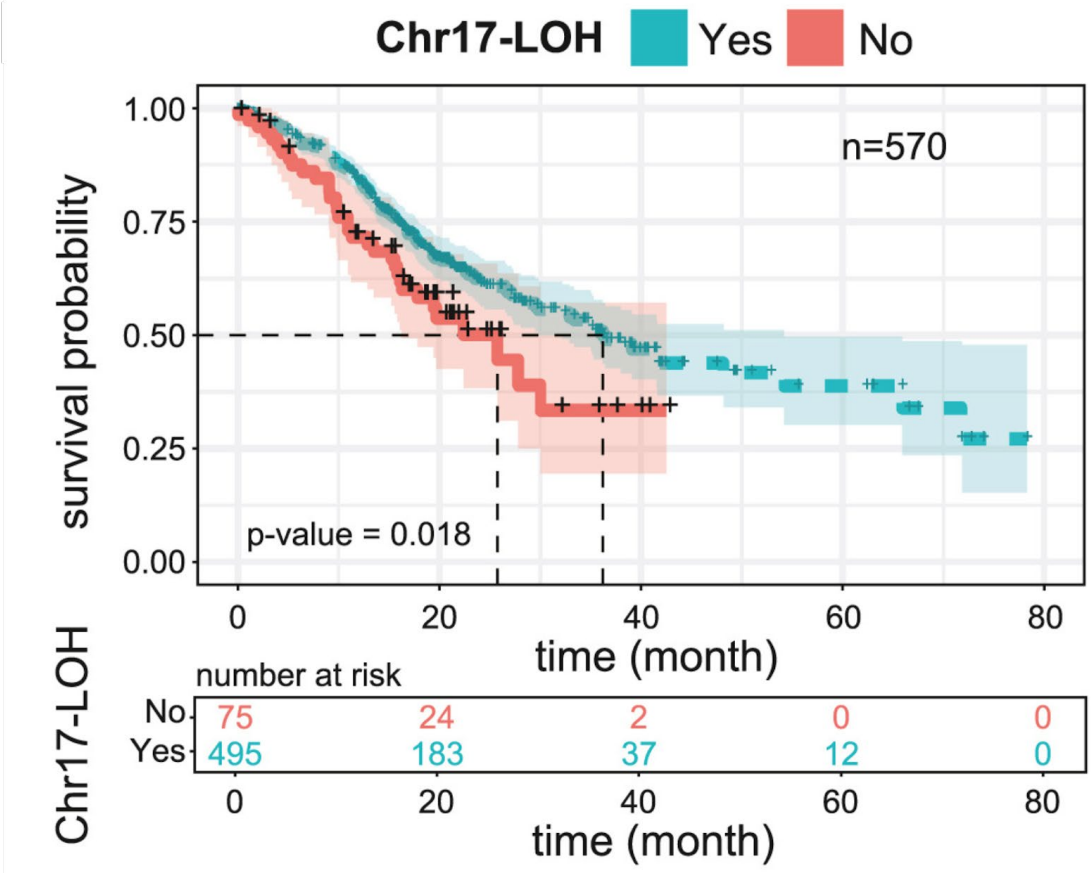


Figure 3.10: MSK-IMPACT cohort HGSOC samples stratified by Chr17-LOH status. The group enriched with Chr17-LOH samples shows a better survival than the group without Chr17-LOH (p = 0.018, based on Cox regression model).

3.2.4 TP53 inactivation is associated with chemo-sensitivity

Chr17 harbors multiple tumor suppressor genes, including *TP53*. Bi-allelic inactivation of *TP53* in cancers typically occurs through missense mutations accompanied by LOH [23]. Indeed,

we find that mutations in *TP53* are significantly associated with Chr17-LOH (odds ratio [OR]: 14.6, p < 6.05e−06) (**Figure 3.11**).

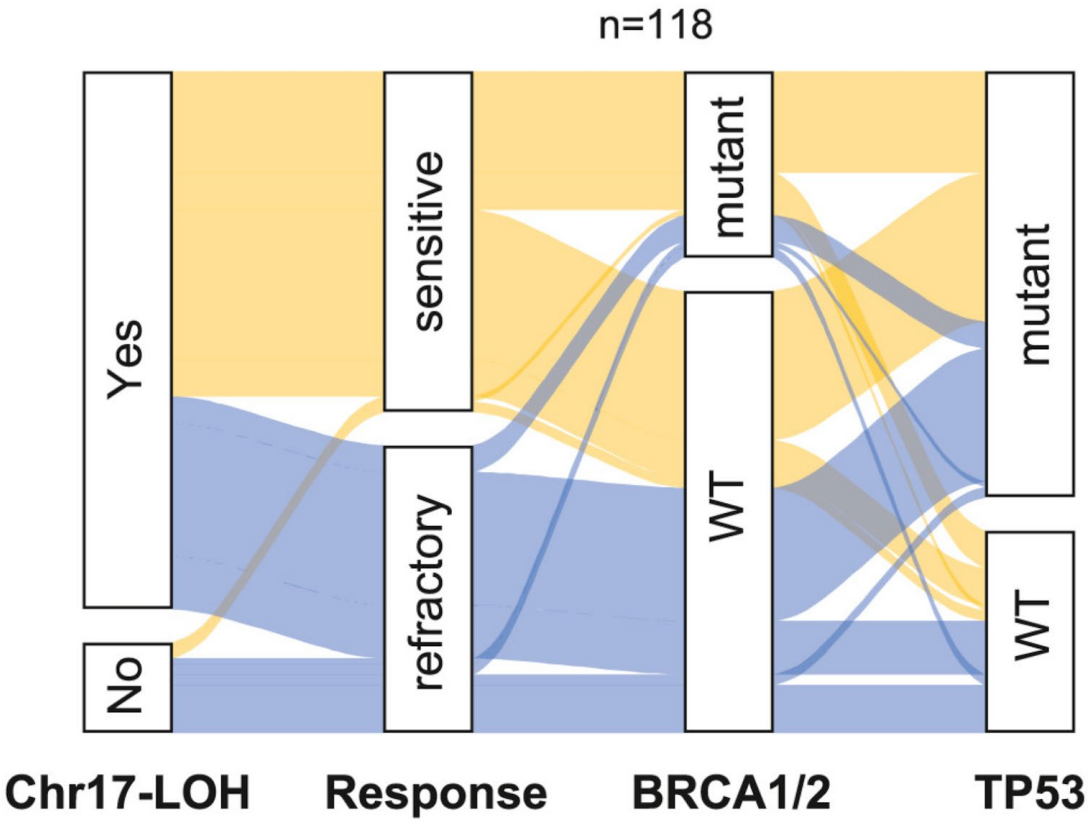


Figure 3.11: Distributions of genomic features and treatment response.

We observed significantly higher *TP53* RNA expression (p = 0.002) and p53 protein abundance (p = 0.001) in tumors with *TP53* missense mutations vs. wild-type (WT) (consistent with the known stabilization of p53 by missense mutations [24]), and lower TP53 RNA expression (p = 0.024) and p53 protein abundance (p = 0.004) in tumors with TP53-truncating mutations (**Figures 3.12** and **3.13**).

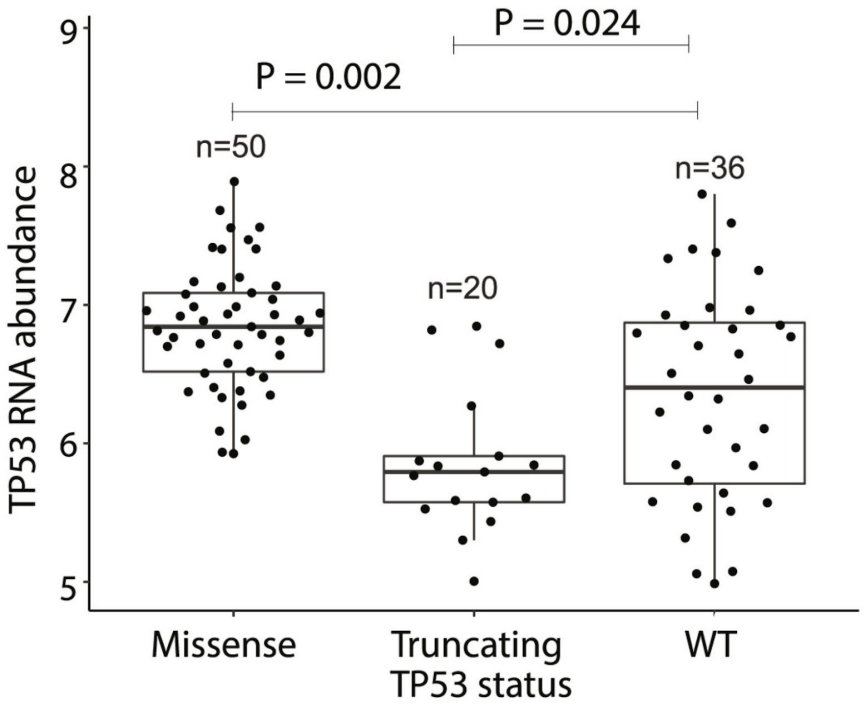


Figure 3.12: Boxplot showing the TP53 RNA abundances stratified by the mutation status of TP53 gene in the FFPE Discovery cohort. Wilcoxon rank sum tests were performed to derive the p-values.

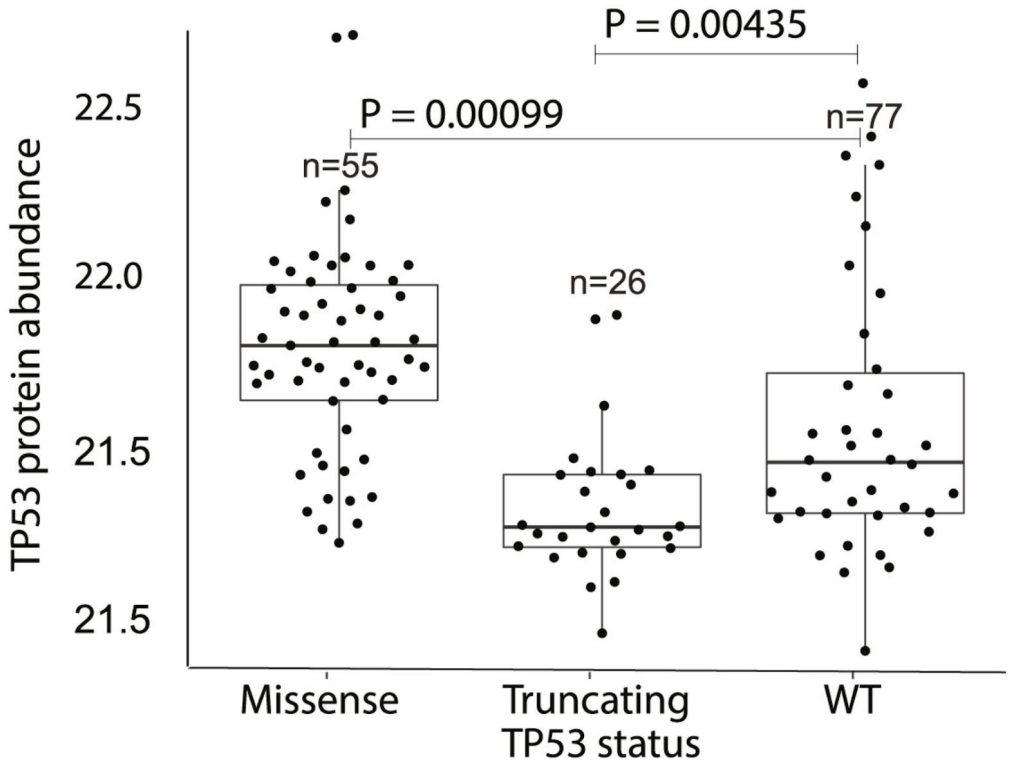


Figure 3.13: Boxplot showing the TP53 protein abundances stratified by the mutation status of TP53 gene in the FFPE Discovery cohort. Wilcoxon rank sum tests were performed to derive the p-values.

Additionally, we observed elevated p53 levels (p = 0.011) among cases with Chr17-LOH (**Figure 3.14**), in line with the enrichment of TP53 mutations in tumors with Chr17-LOH.

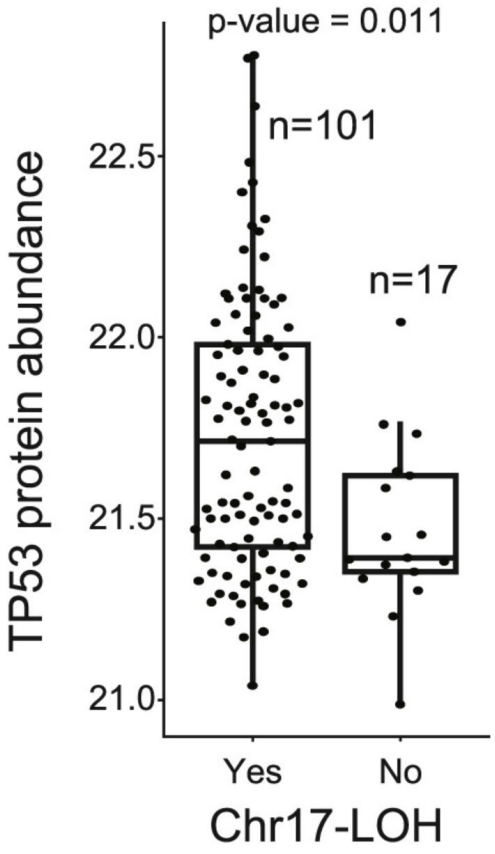


Figure 3.14: Samples with Chr17-LOH have significantly higher p53 protein abundance (p = 0.011, Wilcoxon Rank Sum test) vs. samples without Chr17-LOH.

We hypothesized that the association between Chr17-LOH and sensitivity is due to bi-allelic inactivation of TP53. We interrogated treatment responses stratified by genetic aberrations (*i.e., TP53* mutations and Chr17-LOH) and by transcriptional signatures of both "WT" and "mutant" p53. Missense mutations abrogate WT p53 transcriptional activity [25] and are associated with a mutant transcriptional signature [23]. A WT p53 signature based on validated, direct, transcriptional targets of WT p53 has also been reported [26]

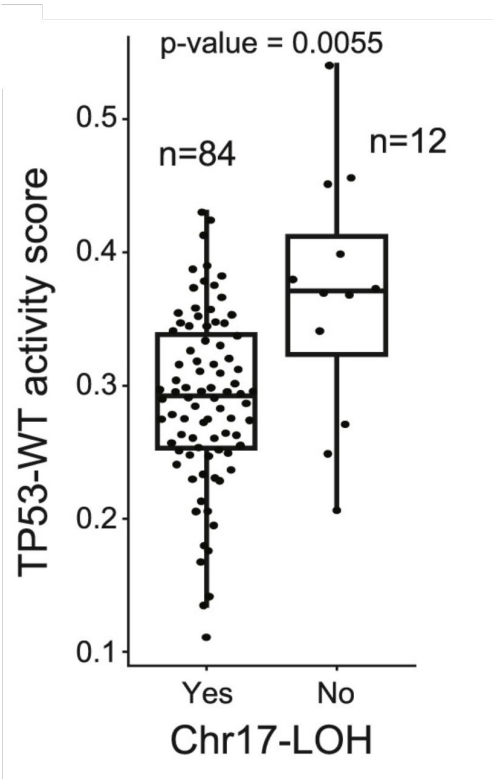


Figure 3.15: Samples with Chr17-LOH have significantly lower TP53-WT activity scores (p = 0.006, Wilcoxon Rank Sum test) vs. samples without Chr17-LOH.

As predicted, we found that Chr17-LOH was associated with a significantly lower WT p53 signature (**Figure 3.15**) and a significantly higher mutant p53 signature (**Figure 3.16**). The WT p53 signature tended to be higher in tumors without TP53 mutations, although the difference was not significant. Moreover, tumors without Chr17-LOH have significantly higher expression of MDM2, a gene activated by WT p53, and significantly lower expression of CDC20, a gene repressed by WT p53 (**Figures 3.17** and **3.18**). These results suggest that tumors without Chr17-LOH retain some p53 activity.

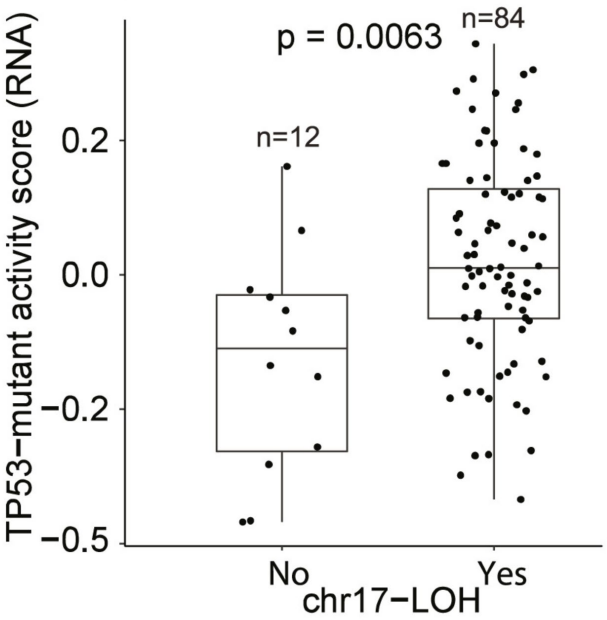


Figure 3.16: p53 mutant activity score calculated using the RNA data, stratified by Chr17-LOH status. Samples with Chr17-LOH have significantly higher p53-mutant signature scores in the FFPE Discovery cohort (p=0.0063, based on Wilcox Rank test).

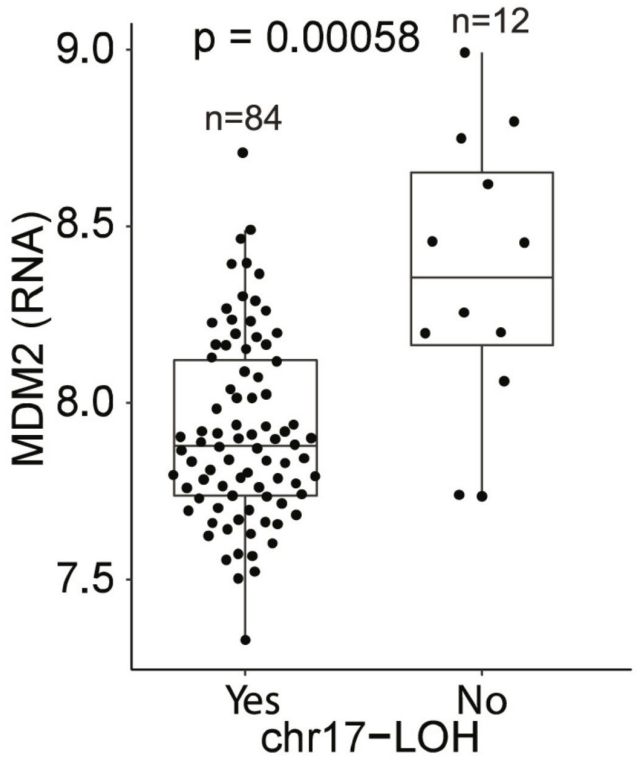


Figure 3.17: Boxplots showing RNA levels of MDM2, gene related to p53 pathway activity, stratified by Chr17-LOH status in the FFPE Discovery cohort.

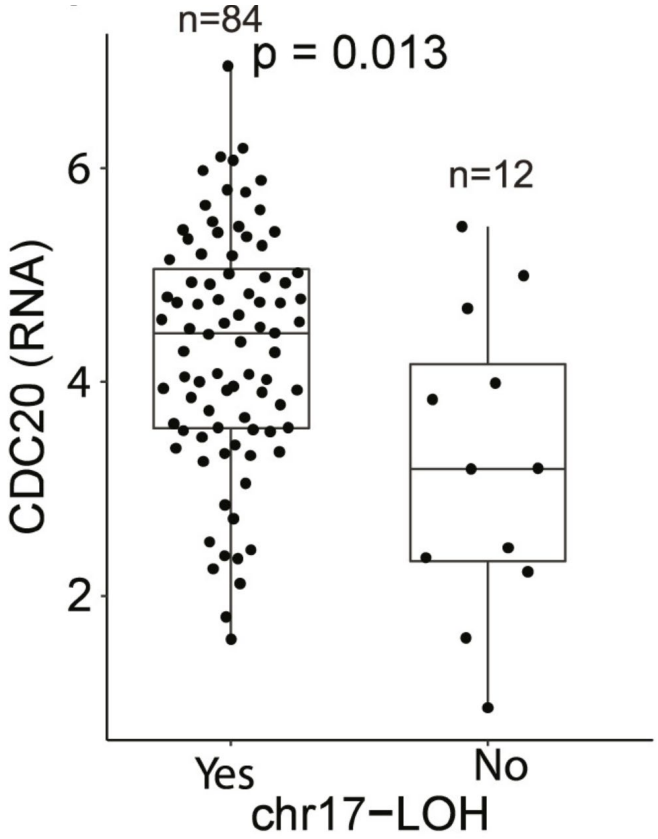


Figure 3.18: Boxplots showing RNA levels of CDC20, gene related to p53 pathway activity, stratified by Chr17-LOH status in the FFPE Discovery cohort.

Next, we asked whether WT or mutant p53 transcriptional signatures are associated with chemo-responsiveness. WT p53 activity was higher in refractory tumors (**Figure 3.19**), whereas the mutant p53 signature was higher in sensitive tumors (**Figure 3.20**).

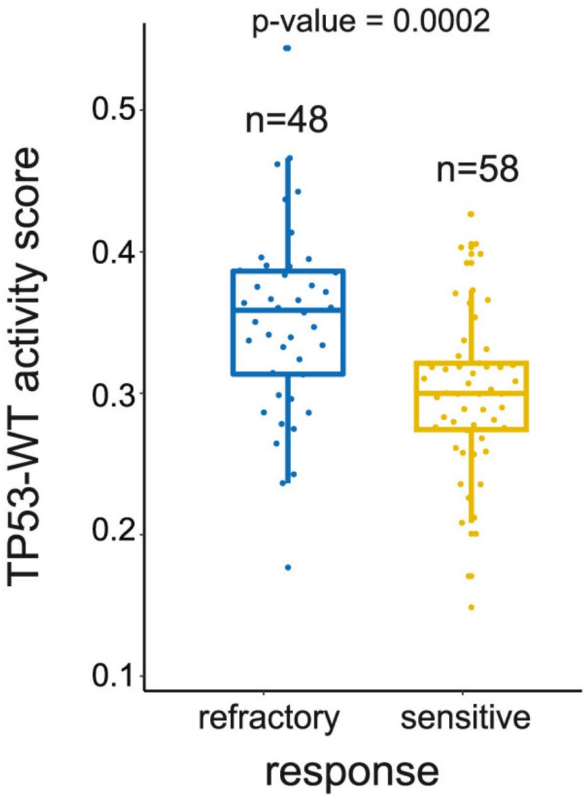


Figure 3.19: Refractory samples have higher TP53-WT activity scores (p = 0.0002, Wilcoxon Rank Sum test), vs. sensitive tumors.

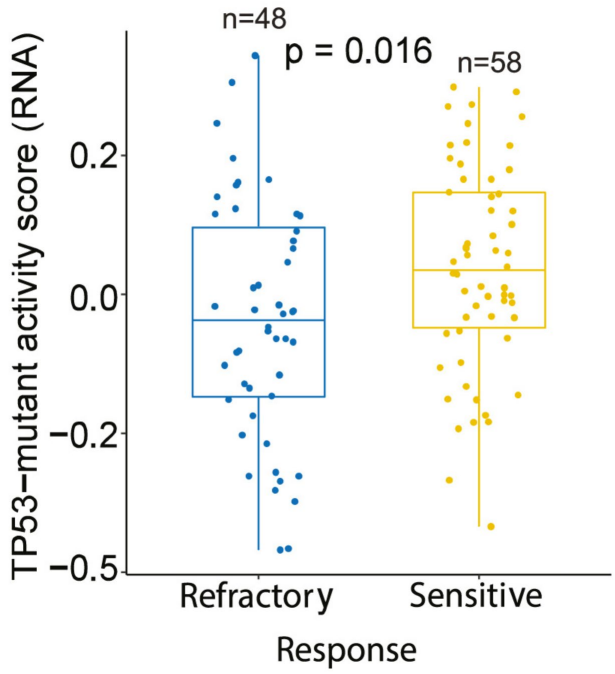


Figure 3.20: p53 mutant activity score calculated using RNA data is significantly lower (p=0.016, based on t-test) in refractory samples compared to sensitive ones in the FFPE Discovery cohort.

Tumors with both high WT TP53 activity and low mutant TP53 activity were predominately refractory (**Figure 3.21**).

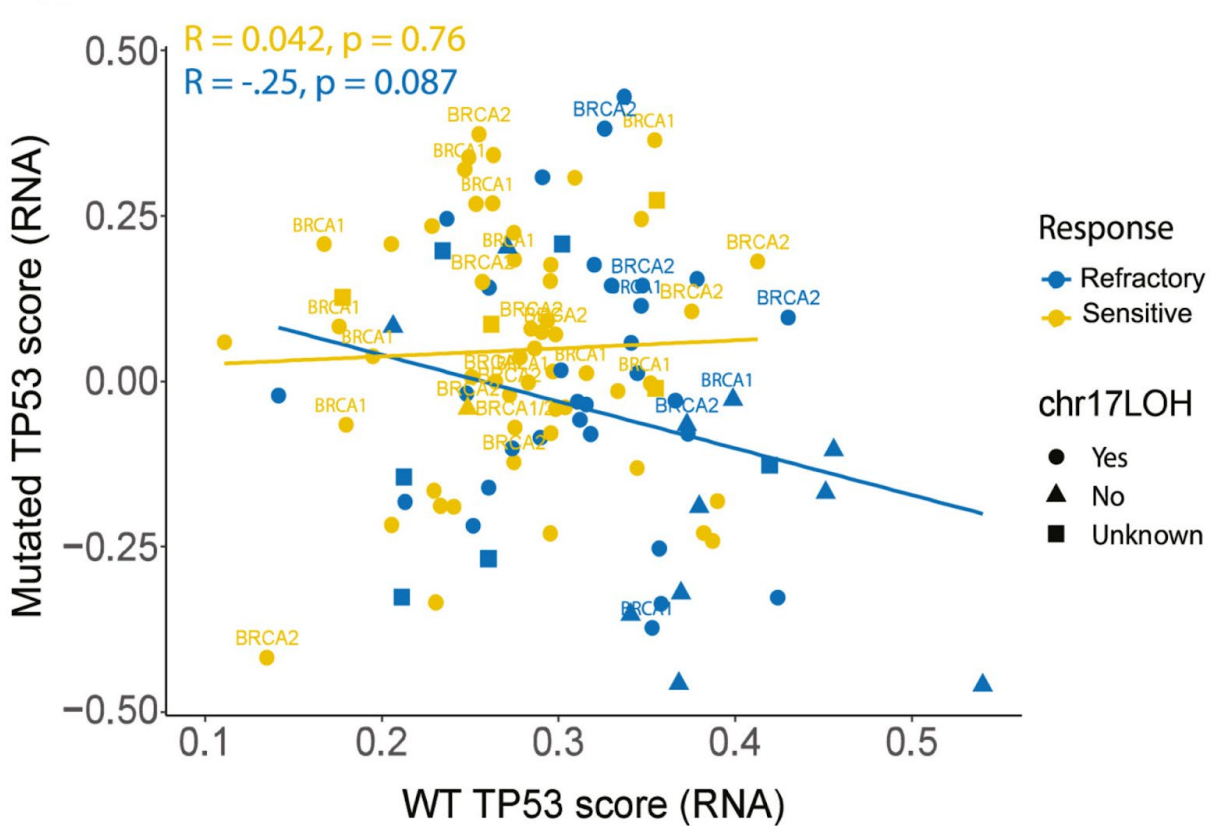


Figure 3.21: Scatter plot showing the correlation of the RNA-derived TP53-WT activity score and p53-mutant activity score for the FFPE Discovery cohort. Colors indicate treatment responses and shapes indicate the Chr17-LOH status. The sample sizes in the different categories are 31 (R with chr17LOH Yes) + 53 (S with chr17LOH Yes), 11 (R with chr17LOH No), 1 (S with chr17LOH No), 6 (R with chr17LOH Unknown), 4 (S with chr17LOH Unknown).

Figure 3.22 illustrates selected pathways associated with Chr17-LOH. E2F targets are upregulated in tumors with Chr17-LOH (FDR = 1.47e−20), consistent with their lower WT p53 activity (Figure 2H), since WT p53 (acting through p21/CDKN1A and RB1) represses E2F [27]. In addition, Chr17-LOH tumors show increased mTOR pathway expression (FDR = 3.1e−21), consistent with repression of the mTOR pathway by WT P53 [28] .

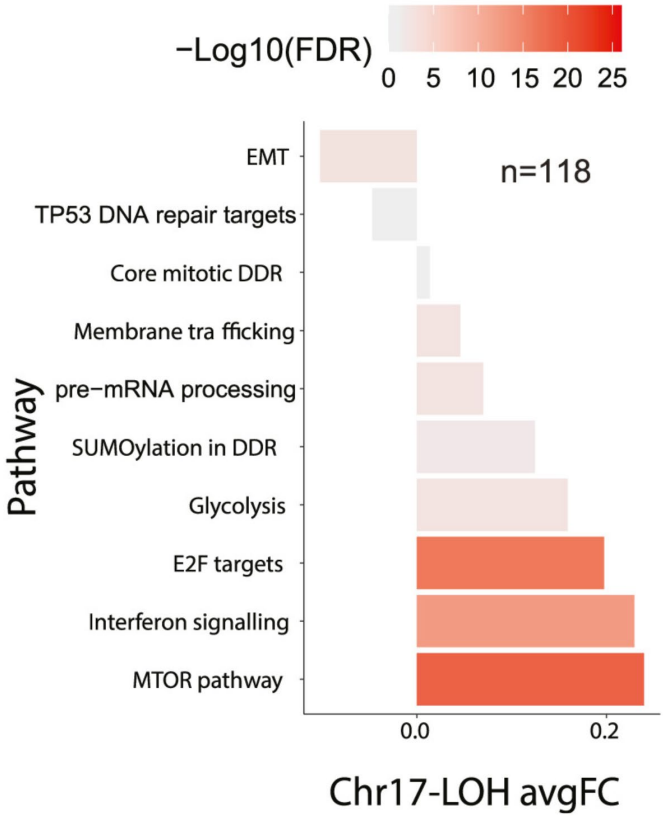


Figure 3.22: Bar plot showing the fold change of select pathways, which were significantly associated (FDR<0.1) with Chr17-LOH status (left) and/or BRCA1/2 mutation (right) respectively, based on the protein data of the FFPE Discovery cohort.

Using a multivariate regression model, we found that Chr17-LOH (OR: 15.76, $p < 0.026$), BRCA1/2 mutation (OR: 4.75, $p < 0.011$), and loss of WT TP53 activity (OR: 0.4, $p < 0.006$) remain independent predictors of response (**Figure 3.23**). These results suggest that WT p53 activity is associated with refractoriness and that tumors with Chr17-LOH lack WT p53 activity, likely due to bi-allelic inactivation of TP53. Conversely, tumors with monoallelic, perhaps dominant-negative TP53 mutations, may retain some p53 WT activity[24].

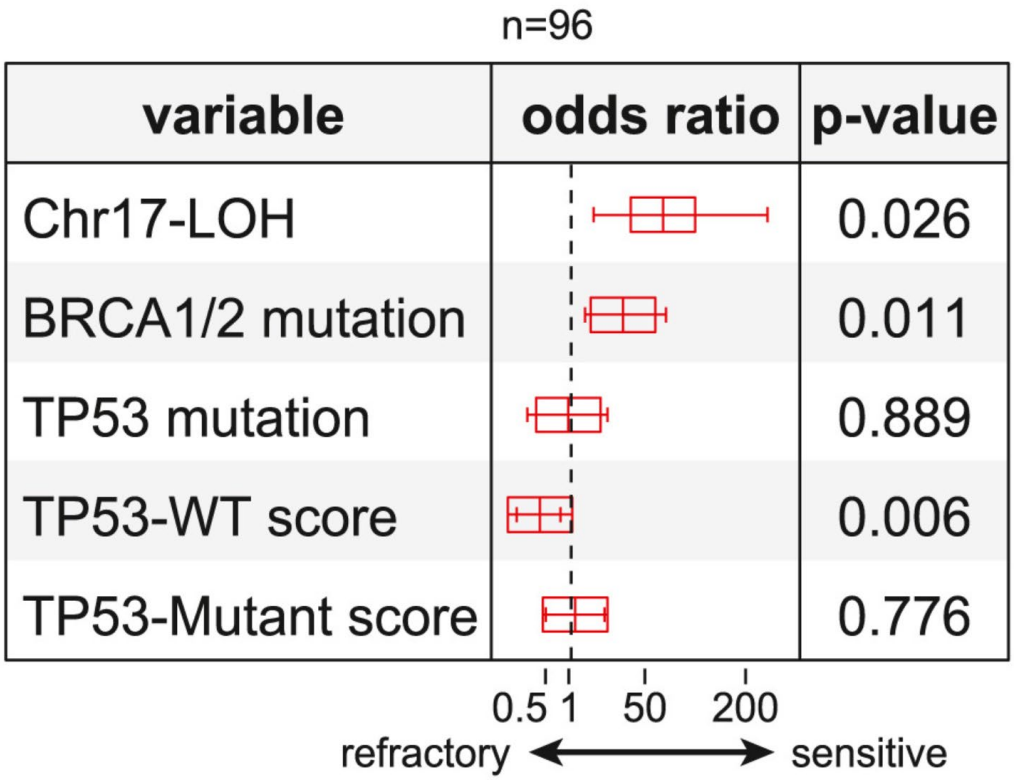


Figure 3.23: Results of the multivariate logistic regression model with refractory status as the response variable and genomic features as predictor variables.

3.2.5 Genomic scars associated with chemo-response

HGSOC genomes are characterized by "genomic scars" (i.e., summary scores reflecting patterns of genomic instability) [29]. In the clinical setting, the extent of number of telomeric allelic imbalances (nTAIs), weighted genomic integrity index (wGII), number of large-scale transitions (nLSTs), and number of LOH (nLOH) are used to identify homologous recombination-deficient (HRD) HGSOC as a Food and Drug Administration (FDA)-approved companion diagnostic for selecting patients who benefit from PARP inhibitors [30]. Since HRD is associated with chemo-sensitivity [31], we tested whether HRD scores differed between refractory and sensitive tumors (**Figure 3.24**).

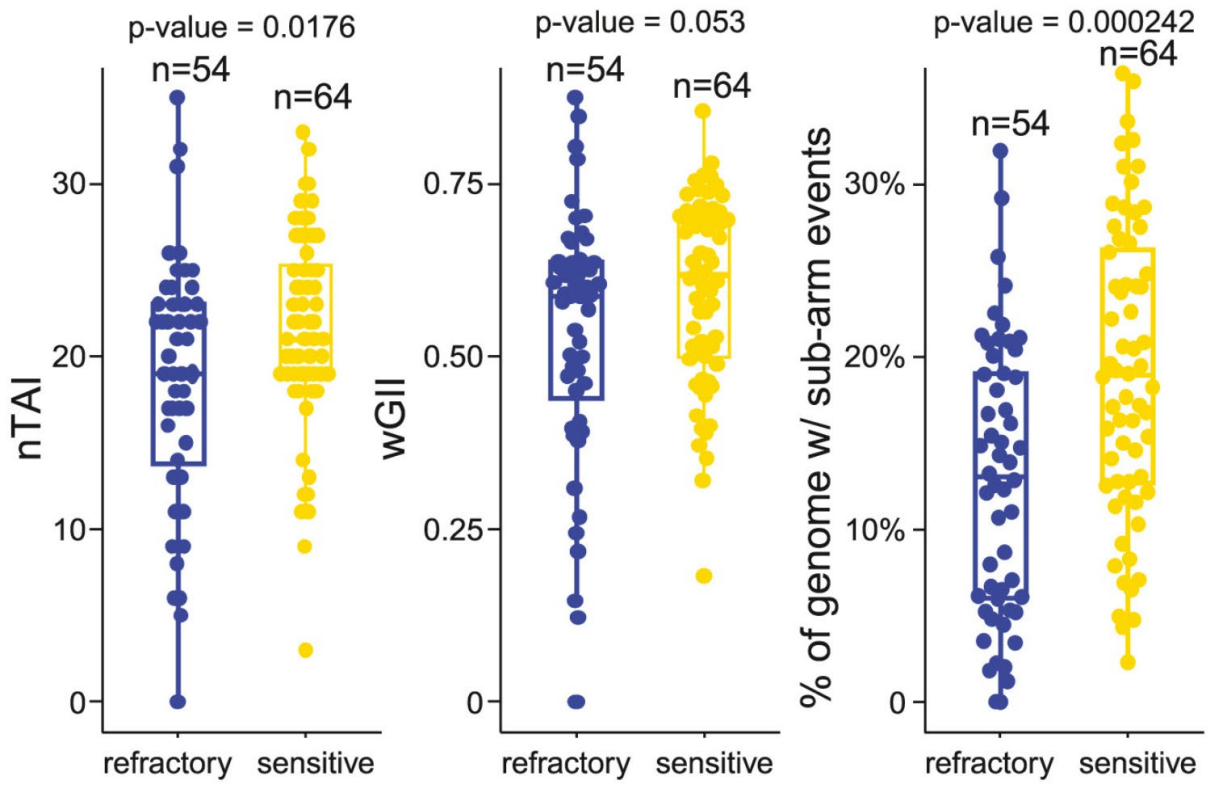


Figure 3.24: Results of the multivariate logistic regression model with refractory status as the response variable and genomic features as predictor variables.

We found elevated levels of "HRD scars" in sensitive tumors based on nTAI (p = 0.0176, **Figure 3.24** left) and wGII (p = 0.053053, **Figure 3.24** middle). Furthermore, chemo-sensitive tumors are associated with elevated frequencies of focal (short) gains and losses (p value = 0.0011, **Figure 3.24**, right), but not with LSTs.

Above, we reported an association between Chr17-LOH and treatment response. Interestingly, tumors lacking Chr17-LOH had lower levels of genomic instability and HRD scores (nTAI and wGII), as well as weaker genome-wide LOH patterns vs. Chr17-LOH tumors (**Figures 3.25** and **3.26**).

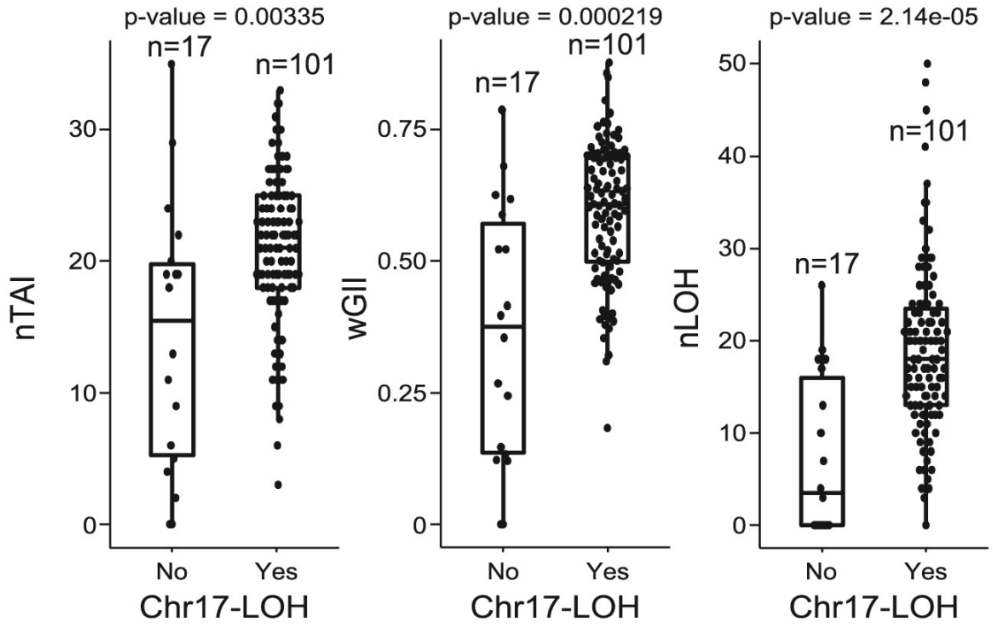


Figure 3.25: Results of the multivariate logistic regression model with refractory status as the response variable and genomic features as predictor variables.

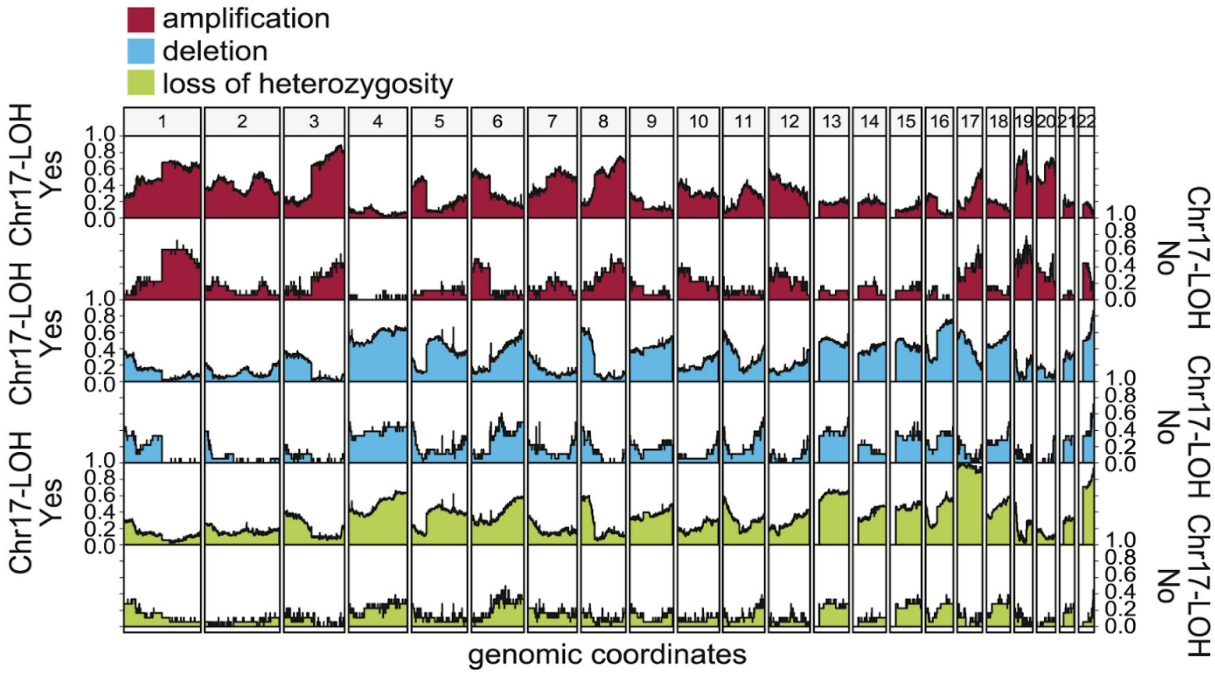


Figure 3.26: Frequency of events (gain, loss, and LOH) in the FFPE discovery cohort, where samples are stratified based on their Chr17-LOH status. While the similar pattern of gains and losses is observed in both groups, the group without Chr17-LOH have less LOH throughout the genome.

Conversely, the genome-wide frequencies and distribution of chromosomal gains and losses were similar between tumors with and without Chr17-LOH (**Figure 3.26**). To interrogate Chr17-LOH as a potential predictor of chemo-sensitivity in the context of the clinically established predictors (HRD score and BRCA1/2 mutation), we tested the association in a multivariate model (**Figure 3.27**) and found that only Chr17-LOH and BRCA1/2 mutation status were predictive of response, both associated with chemo-sensitivity.

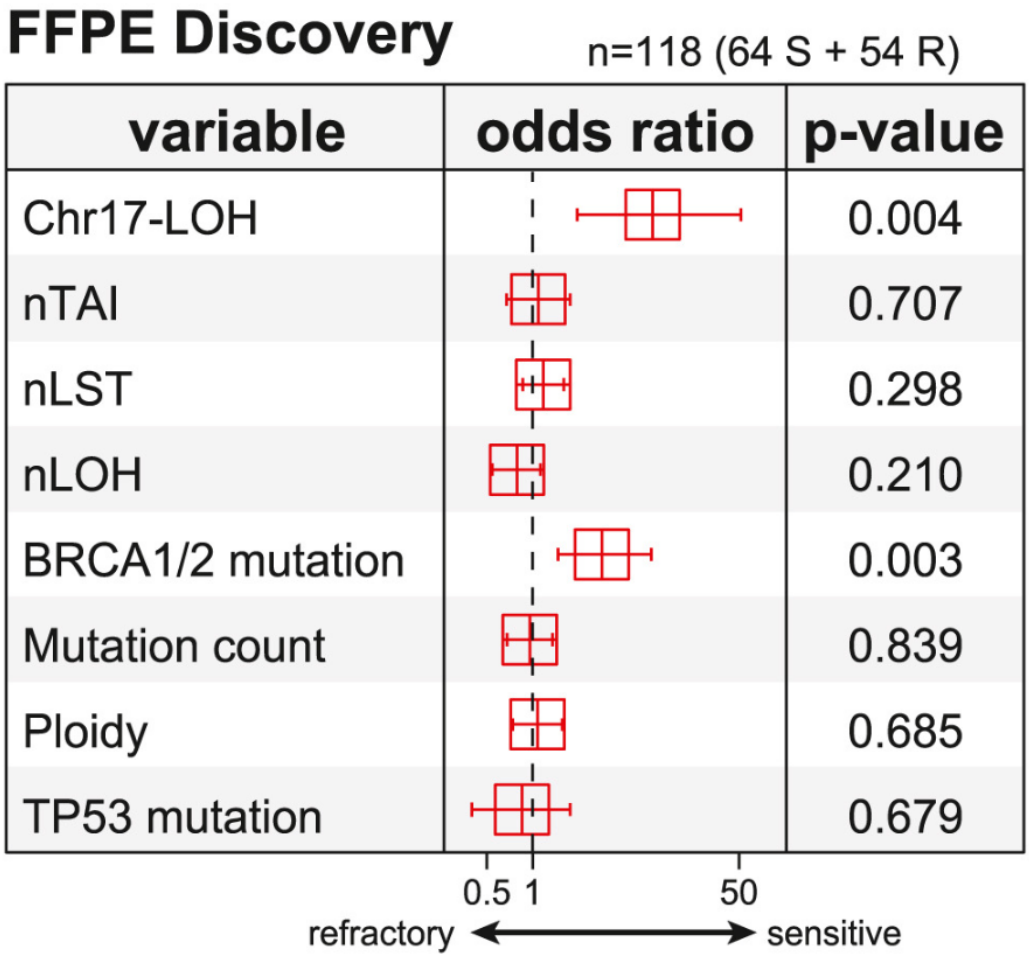


Figure 3.27: Forest plot of a multivariate logistic regression model predicting refractoriness based on HRD and genetic variables.

We also noted an improvement in predictive power for refractory patients when Chr17-LOH is added to a baseline model combining clinical covariates and BRCA1/2 status in two independent patient cohorts (**Figure 3.28**).

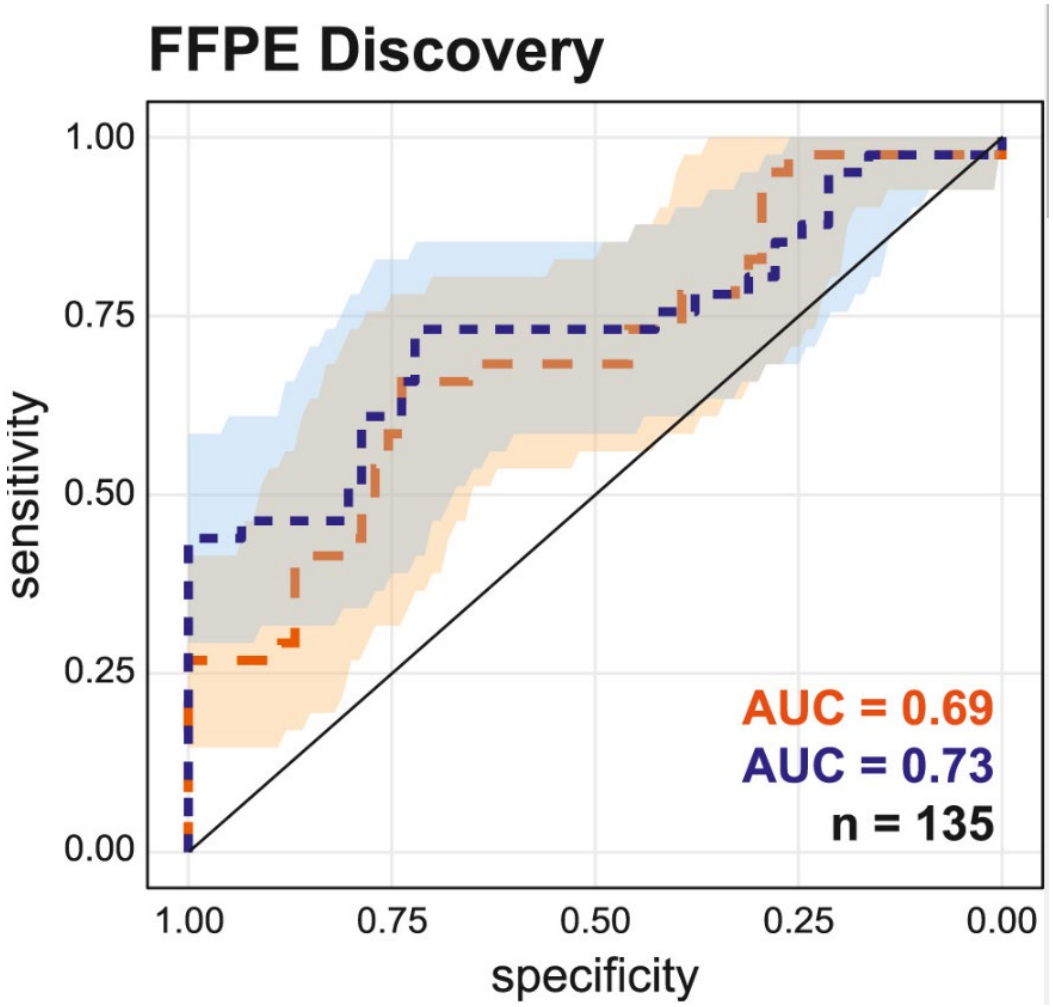


Figure 3.28: ROC curves showing the prediction performance of models based on (1) BRCA1/2 mutation status + clinical variables (patient age and tumor location) and (2) BRCA1/2 mutation status + clinical variables + Chr17-LOH status.

The area under the curve (AUC) of the receiver operating characteristic (ROC) of the resulting model is 0.73 with a 95% confidence interval (CI) of 0.63–0.84 (FFPE discovery cohort) (**Figure 3.28**). Despite these encouraging results, better prediction models are needed for clinical practice. This motivated us to explore the proteomic data to determine whether more effective prediction models could be built based on protein markers.

3.3 Discussion

Platinum based treatment is part of the standard first-line treatments for patients with high grade serous ovarian carcinoma [32]. Majority of these patients initially respond to treatment and later develop resistance through genomic and epigenomic alterations in their genome [33]. There is 10-20% of HGSOC which doesn't respond to treatment from the beginning without the need to alter their genomes [6]. That suggests there should be some differences in the biology of these tumors which makes them different from the majority of HGSOC in response to platinum treatment, and hence there is a potential for finding predictive biomarkers for response to platinum treatment by studying treatment-naive tumors.

Despite >3 decades of research on platinum responses in cancer [17], no predictive biomarker has been translated into clinical use. Predictors of refractory disease could enable a precision oncology approach and provide a means to select patients for clinical trials to identify and implement effective therapies.

In this study, we enriched our cohort with refractory patients to increase our power of finding differences between refractory and sensitive tumors. Furthermore, to increase the clinical utility of our findings, we used Formalin-Fixed Paraffin-Embedded (FFPE) samples for all subsequent sequencing.

Our finding of chr17-LOH association with sensitivity, although predictive, doesn't give much biological insight into how LOH of chr17 is causing sensitivity to treatment. To explore why, we leveraged technological advances (e.g., LC-MS/MS-based multiplex protein quantification, machine learning algorithms) to find underlying genetic events associated with platinum refractoriness in HGSOC.

According to our cohort and other published studies [19], many of TSGs which are frequently mutated in HGSOC (**Figure 3.3**) are located on chromosome 17, including *TP53*, *NF1*,

CDK12 and *BRCA1*. We started exploring the association of chr17LOH and treatment sensitivity by evaluating *TP53,* one of the most commonly mutated genes pan cancer [23]. Through two orthogonal gene signatures measuring p53 WT and mutant downstream activity, we confirmed samples without TP53 LOH, have significantly higher p53 WT activity and significantly lower p53 mutant signature. Also, almost all the HGSOC tumors harbor *TP53* mutation [19,23]. Given that LOH of chr17 can act as the second hit to inactivate *TP53,* we can hypothesis that tumors which maintain at least one functional copy of *TP53* (and hence have p53 WT activity) are able to show resistance to platinum treatment (and hence become refractory). Examining this hypothesis was out of scope for our study, but further confirmation of this study can be extremely important for biomarker development in HGSOC, and possibly other type of cancer with high rate of *TP53* mutations.

3.4 Methods

3.4.1 Patient selection and sequencing

As a part of the CPTAC-PTRC consortium, our team collected treatment-naive tissue specimens from Stage III or IV HGSOCs that underwent primary debulking followed by platinum/taxane adjuvant therapy were selected for the study and categorized with respect to response to adjuvant chemotherapy, as defined below:

Refractory tumors were defined as follows: After primary debulking: for R0 disease (no residual/microscopic disease after primary resection), radiographically detectable disease must have been present at the end of 6 cycles of initial platinum/taxane therapy; for R1 disease (radiographically detectable residual disease is present after primary resection), residual disease must either progress or stay stable (radiographically) after 6 cycles of initial platinum/taxane therapy.

Sensitive tumors were defined by R1 or R2 disease that received primary resection followed by platinum/taxane adjuvant therapy and had a progression-free survival of at least 2 years (no R0 disease).

Further information regarding sample collection and sequencing details can be found in Chowdhury et al [34].

3.4.2 WGS Data Processing and Alignment

All genomic data processing has been performed using the TPO workflow (https://github.com/mctp/tpo), which implements standardized pipelines for the analysis of DNA and RNA sequencing data.

DNA sequencing paired-end reads were trimmed from adapter sequences using BBMap's bbduk tool with parameters 'ktrim=r k=23 mink=11 hdist=1 ignorebadquality=t qin=33 tpe tbo' [35]. Sequencing data were aligned to the GRCh38 reference (GRCh38.d1.vd1), according to the functional equivalence standards [36] and BWA-mem [37] (version 0.7.17-r1188) settings '-Y -K 10000000'. The alignments were sorted using Sentieon tools sort, which is equivalent to samtools sort [38]. Alignment statistics have been collected using the algorithms MeanQualityByCycle, QualDistribution, GCBias, AlignmentStat, and InsertSizeMetricAlgo. Following sorting we followed GATK best-practices, including indel realignment, base quality score recalibration (using algorithmsRealigner and QualCal, respectively), settings '-k 'DBsnp all variants 20180418', and the 'known_indels' and 'Mills_indels' from the GATK resource bundle. The resulting alignments were genotyped at common heterozygous SNPs (DNAscope gVCF typing) and all samples from the same individual were verified to have matching genotypes 'tpo/refs/grch38/custom/genotype_positions_hg38.bed'. Duplicates were removed using Sentieon

LocusCollector and Dedup. Since the sequencing was performed on a Novaseq an optical duplicate threshold of 2500 was used to mark ExAmp duplicates '--optical_dup_pix_dist 2500'.

3.4.3 WGS Data QC

For the QC of WGS data, we derived multiple quality control measures, such as GC BIAS or duplication percentage, and confirmed that average qualities are not significantly different between the compared clinical patient subsets (**Figure 3.29**). Downstream copy-number analysis factored DNA data quality directly into the procedure of segmentation and inference to minimize the influence of data quality on sensitivity and recall.

3.4.4 WGS Data Variant/Mutation Calling

Tumor-only somatic variant calling has been performed using TNScope, which is a modified version of GATK3 MuTect2 algorithm, with the following settings '--max_fisher_pv_active 0.05 --min_tumor_allele_frac 0.0075 --min_init_tumor_lod 2.5 --assemble_mode 4 --trim_soft_clip' and '--dbsnp homo_sapiens_assembly38.dbsnp138.vcf.gz', which is included in the TPO resource bundle.

Specifically, the 'trim_soft_clip' parameter reduces the number FFPE-induced artifacts. Multi-nucleotide variants were merged based on phasing information using an inhouse script 'tpo/pipe/carat-anno/carat_mnv.R'. The resulting variants were first filtered to protein coding regions (based on Ensemble 97), annotated using VEP [39], and further annotated with vcfAnno[40], using the following sets of parameters: '--assembly GRCh38 --species homo_sapiens --cache_version 97 --format vcf --gene_phenotype --symbol --canonical --ccds --hgvs --biotype --tsl --uniprot --domains --appris --protein --variant_class --sift b, --polyphen b --no_stats --total_length --allele_number --no_escape --flag_pick_allele --pick_order canonical,tsl,biotype,rank,ccds,length --buffer_size 20000'. For vcfAnno multiple sources of

annotations were used including: GDC Panel-of-Normal v4136 based on MuTect2, 1000 Genomes phase3 variants, GnomAD exomes r301, GnomAD exomes r221, ClinVar release Dec. 2019, Cosmic v87, dbSNP release 20180418, UniProt release Dec. 2019

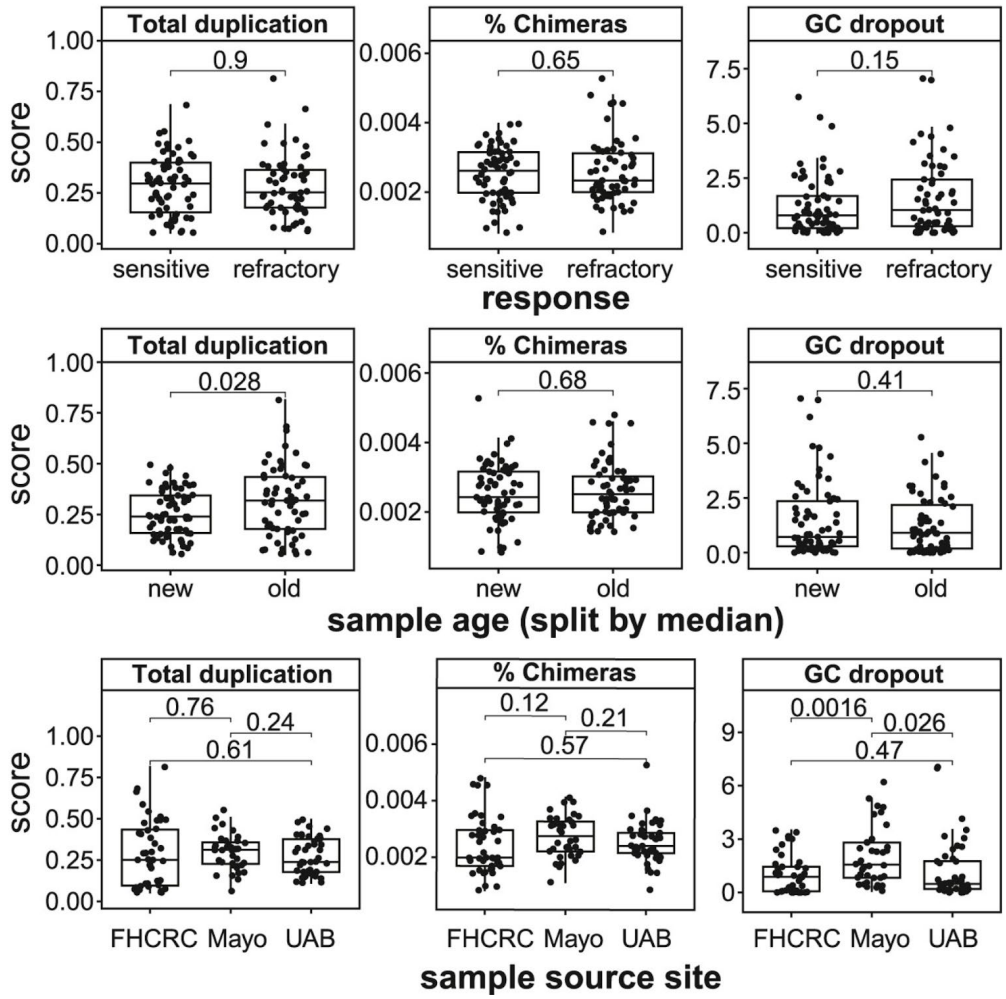


Figure 3.29: Boxplots showing the association of the WGS quality control metrics among tumors stratified by tumor phenotype, sample storage time, and source sites. The sample size stratified by the response is 65 S + 55 R, the sample size stratified by sample age (split by median) is 62 in new and 58 in old group, and the sample size stratified by source sites are 39 FHCRC, 36 Mayo and 45 UAB.

The resulting annotated variants were subsequently filtered to select for likely somatic and pathogenic variants based on automated rules, gene-specific threshold, and manual filters. We applied filters based on sequencing evidence (variant allele frequency, coverage, mutation likelihood TLOD, strand bias, allele depth, multi-allelic variants), and overlap in problematic

regions including regions with low-mappability and repetitive sequence as well as homopolymer repeats. In addition, all long indels (>20 bp) were removed, as these were significantly enriched in this data due to the low quality of the input DNA. All called variants in genes known to be recurrently mutated in HGSOC identified based on COSMIC and literature, were manually reviewed, according to a procedure following applicable recommendations in Barnell et al. [41].

3.4.5 CNV analysis

Copy-number analysis on the tumor DNA was performed by using whole-genome sequencing (WGS) coverage data and variant calls. To perform the analysis, we used CNVEX (https://github.com/mctp/cnvex), a comprehensive copy number analysis tool that has been used previously in our ccRCC studies [42]. CNVEX uses whole-genome aligned reads to estimate coverage within fixed genomic intervals and variant calls to compute B-allele frequencies (BAFs) at variant positions (as characterized above). Coverages were computed in 10 kb bins, and the resulting log coverage ratios between tumor and normal samples were adjusted for GC bias using weighted LOESS smoothing across mappable and non-blacklisted genomic intervals within the GC range 0.3-0.7, with a span of 0.5 (all target and configuration files are provided within the CNVEX repository). The adjusted log coverage-ratios (LR) and BAFs were jointly segmented by a custom algorithm based on Circular Binary Segmentation (CBS). Alternative probabilistic algorithms were implemented in CNVEX, including algorithms based on recursive binary segmentation (RBS) [43], and dynamic programming [44], as implemented in the R-package jointseg [45]. For the CBS-based algorithm, first LR and mirrored BAF were independently segmented using CBS (parameters alpha=0.01, trim=0.025) and all candidate breakpoints collected.

The resulting segmentation track was iteratively ‘‘pruned’’ by merging segments that had similar LR and BAFs, short lengths, enrichment in blacklisted regions, and a high variation in coverage among whole cohort germline samples. For the RBS- and DP-based algorithms, joint-break-points were ‘‘pruned’’ using a statistical model selection method [46]. For the final set of CNV segments, we chose the CBS-based results because they did not require specifying a prior number of expected segments (K) per chromosome arm, were robust to unequal variances between the LR and BAF tracks and provided empirically the best fit to the underlying data. Given the relatively high variance of the sequencing data, we noticed that some samples were over-segmented using the default settings. To adjust the segmentation algorithm to the variance of each sample individually, we performed an adaptive procedure of pruning. Default pruning: Perform pruning as described above for all samples. Strict pruning: Count the number of segments for each sample, take the samples with the top 20 percentile number of segments and perform pruning with stricter parameters. Relaxed pruning: Take the samples with the lowest 20 percentile number of segments and perform more loose pruning on them. The above procedure was repeated two times, effectively adjusting segmentation pruning parameters to the noise-level of each sample. The quality of the resulting segmentations was manually evaluated.

The resulting segmented copy-number profiles were then subject to joint inference of tumor purity and ploidy and absolute copy number states, implemented in CNVEX, which is most similar to the mathematical formalism of ABSOLUTE [22] and PureCN [47]. Briefly, the algorithm inputs the observed log-ratios (of 10 kb bins) and BAFs of individual SNPs. LRs and BAFs were assigned to their joint segments and their likelihood is determined given a particular purity, ploidy, absolute segment copy number, and number of minor alleles. To identify candidate combinations with a high likelihood, we followed a multi-step optimization procedure that

includes grid-search across purity-ploidy combinations, greedy optimization of absolute copy numbers, and maximum-likelihood inferences of minor allele counts. Following optimization, CNVEX ranked candidate solutions. Because the copy-number inference problem can have multiple equally likely solutions, further biological insights were necessary to choose the most parsimonious result. The solutions were reviewed by independent analysts following a set of guidelines. Solutions implying whole-genome duplication were supported by at least one large segment that cannot be explained by a low-ploidy solution, inferred purity was consistent with the variant-allele-frequencies of somatic mutations, and large homozygous segments were not allowed.

After selecting the best solution for purity/ploidy of a sample, we then used the method explained in PureCN133 to assign total absolute CN to each CNV segment using LR values. Purity and ploidy estimates also allowed us to use BAF values for calculating the absolute copy number for each allele. We then used the copy number of minor allele to detect LOH for each segment. Any segment that had a minor allele copy number of 0, was considered to harbor an LOH event.

To call samples with Chr17LOH, we used the same method mentioned above to call LOH for all the segments on chr17. Samples that had <25% of their chr17 covered with LOH segments were called chr17HET. Any other samples were placed in the chr17LOH group.

To validate the results of our chr17-LOH, CNV data from MSK-IMPACT (997 patients) was used. Because the LOH values were not provided for this cohort, we used the LogRatio values as a surrogate. In our own cohort, we found that Chr17-HET and Chr17-LOH have significantly different LR values (Wilcox test P=0.0023, **Figure 3.30**). We therefore used our own cohort of patients to define a LR threshold which separates patients into two groups of “Chr17-LOH enriched” and “Chr17-HET enriched”. For this purpose, we defined samples with chr17 average

LR > 0.046 as a surrogate for “Chr17-HET” (75 patients) group and samples with chr17 average LR < -0.114 as a surrogate for “Chr17-LOH” (495 patients) group. We then used this stratification to perform Cox proportional hazards survival analysis.

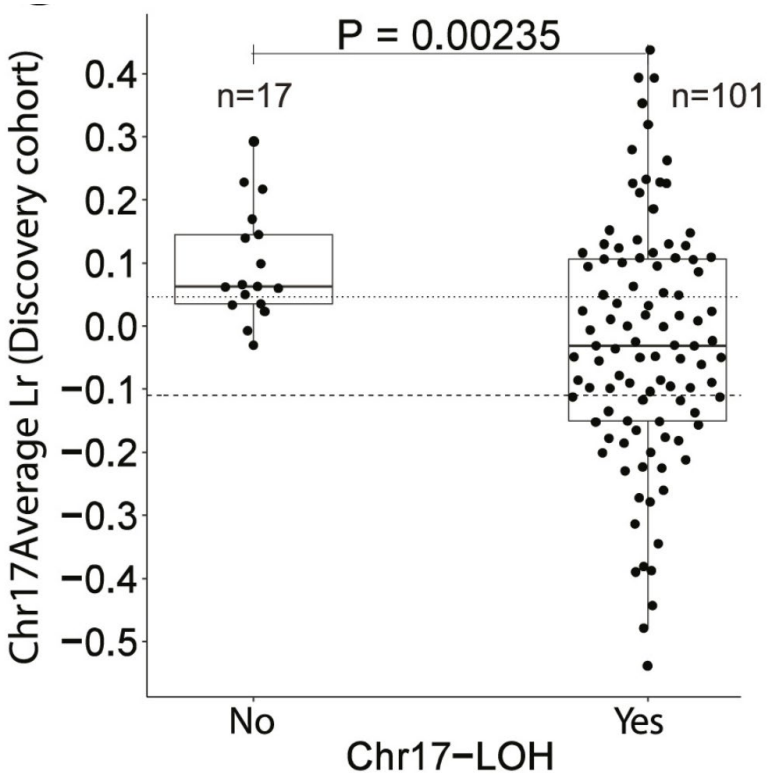


Figure 3.30: Average copy number (log ratio [LR]) of Chr17 in the FFPE discovery cohort stratified by Chr17-LOH status. Samples without Chr17-LOH have significantly higher LR values. The dotted (top) line shows the threshold (LR > 0.046) used to define a group of tumors enriched for samples without Chr17-LOH; the dashed (bottom) line shows the threshold (LR < −0.11) used to define a group of tumors enriched for samples with Chr17-LOH. These thresholds were used for the analysis of the MSK-IMPACT cohort.18 Wilcoxon rank sum test was performed to derive the p value.

3.4.6 Instability measures

For measuring overall instability, we used an already published measure of Weighted Genome Instability Index (wGII) [48], which was measured as the proportion of each chromosome that has a different copy number compared to the baseline copy number of the sample. Then the average of scores for each chromosome was calculated, weighted by the length of the chromosome such that each chromosome has the same contribution to the overall instability score.

To quantify the HRD for each sample, we reimplemented the measures described in Sztupinszki et al. [49] In short, allelic imbalance was defined as the unequal contribution of parental alleles to a region of the genome and contains LOH as well. For nLST we defined “large segment” as a segment with minimum length of 10 Mb, and to count one event of large segment transition, we allowed maximum distance of 3Mb between two “large segments” with allelic imbalances. For nTAI, we counted the allelic imbalances with minimum length of 5Mb that stretch to the telomeric end of each chromosome, which on GRCh38 we used the 10Kb of chromosome start/end as telomere. For nLOH, we counted any segment of minimum length of 15Mb which has Loss of heterozygosity. For all of these HRD measures, if the segment is affecting the whole arm of the chromosome (arm-level event), it is not counted toward the HRD measures.

3.4.7 RNA Sequencing Processing, Alignment and Quantification

The total RNA-seq stranded paired-end reads were trimmed of adapter sequences using BBMap’s bbduk2 tool with parameters ‘minlength=25 k=31 qskip=3 rieb=t tbo=t tpe=t’, the resulting reads were aligned to a GRCh38 reference including common human oncogenic viruses using STAR, using the following settings:

alignSJoverhangMin 8 \ alignSJDBoverhangMin 3 \ scoreGenomicLengthLog2scale 0 \ alignIntronMin 20 \ alignIntronMax 1000000 \ alignMatesGapMax 1000000 \ outStd "SAM" \ outReadsUnmapped "Fastx"

In addition, trimmed paired-end reads were merged into a synthetic single-end reads using the BBMap bbmerge tool, the resulting merged single-end and not-merged trimmed paired-end reads were aligned a second time with STAR using settings to increase the sensitivity of fusion detection (see below):

outFilterType Normal \ alignIntronMax 150000 \ alignMatesGapMax 150000 \ chimSegmentMin 10 \ chimJunctionOverhangMin 1 \ chimScoreSeparation 0 \ chimScoreJunctionNonGTAG 0 \ chimScoreDropMax 1000 \ chimScoreMin 1

The resulting RNA-seq BAM files were genotyped using DNAscope with the RNASplitReadsAtJunction algorithm and compared to the DNA-based genotypes.

Genotype RNA alignments were used subsequently for gene expression quantification (see below) as well as quality control. The alignments were input into CODAC, a component of TPO designed to detect gene fusions, as well as perform quality control, and re-alignment using minimap [50] and GMAP [51]. The following CODAC settings were used:

- Annotation: /tpo/refs/grch38/config/codac-grch38.97.rds

- Stringency: longread-balanced-stranded

Total RNA-seq were quantified using Kallisto[52]. First, RNA-seq FASTQ files were trimmed of adapter sequences as described above using BBMap bbduk2. Next, Kallisto was run using the default settings with '--rf-stranded' against the Ensembl 97 transcript database. Kallisto estimated counts were summarized (summed) at the gene level based. Both raw counts (optionally adjusted to total sequencing depth) and TPMs were used depending on the downstream analysis.

3.4.8 RNA-sequencing Data Quality Control

To mitigate challenges of analyzing genomic data from archival FFPE specimens, we performed extensive quality control of the RNA sequencing (RNA-seq) to identify low quality samples based on multiple individual control variables specific to the RNA, such as estimates of PCR duplication percentage, reverse transcription template-switching artifacts as well as 5' / 3' bias, transcriptome coverage and number of detected splice-junctions. Robust thresholds and cut-point for the control variables were established based on a University of Michigan institutional cohort

of FFPE and clinical samples[53]. While we did not detect any differences in quality measures of sensitive and refractory samples, we observed variation in duplication percentages and detected splice junction numbers across sample source sites (**Figure 3.31**). Thus, in the downstream analyses, as appropriate, sample source site has been used as a covariate in all association tests.

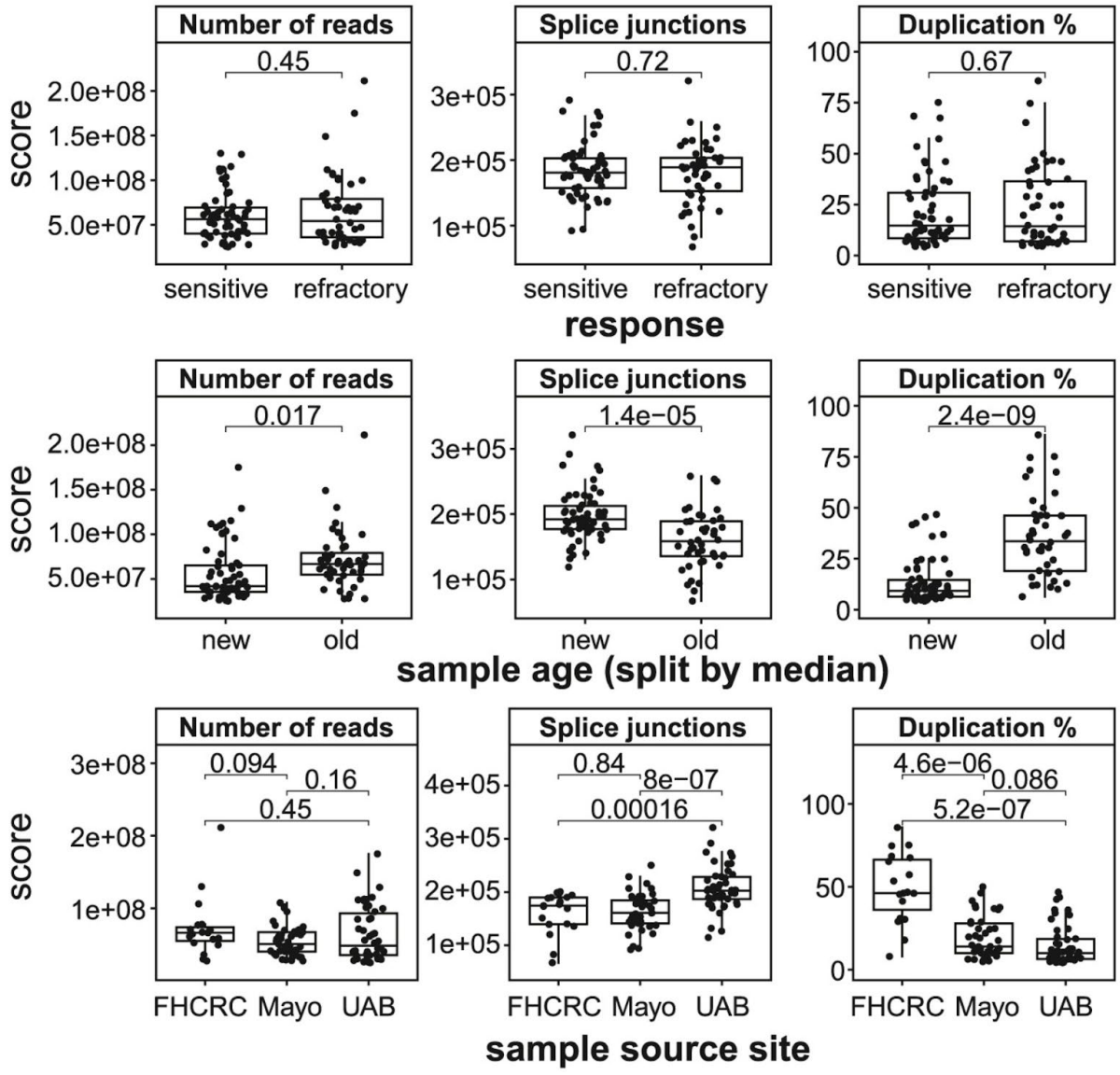


Figure 3.31: Boxplots showing the distributions of the RNAseq quality control metrics among tumors stratified by treatment response, sample storage time, and source sites. The sample size stratified by the response is 58 S + 48 R, the sample size stratified by sample age (split by median) is 53 in each group, and the sample size stratified by source sites are 19 FHCRC, 41 Mayo and 46 UAB.

3.4.9 TP53 Signatures

Previously published gene sets were used to develop the TP53 wild-type and mutant signatures. For the wild-type TP53 signature, we used the previously identified core TP53 transcriptional program based on Andrysik et al [26]. We focused on the key 31 genes with direct bounding and identified them in all three cell line experiments (HCT116, MCF7, SJSA). For the mutant TP53 signature, we used the previously published analysis [23]. The top 20 upregulated genes across all mutant TP53 cancers were selected for the signature. Using the GSVA package (method = "ggsea") in R [54] we obtained the ssGSEA scores for the TP53 wild-type and TP53 mutant gene sets for all samples in the FFPE Discovery RNA-seq dataset (n = 106). Additionally, we obtained the ssGSEA scores for the TP53 wild-type and TP53 mutant gene sets for all samples in the FFPE Discovery proteomics dataset (n = 158). We then performed a multiple linear regression model to test for the association of each TP53 signature with platinum response. Specifically, we considered the following model for testing the association: ssGSEA scores (WT or mutant) ~ platinum response + tumor location + source site + sample age + neoadjuvant treatment status + tumor purity.

3.5 References

1. Matulonis UA, Sood AK, Fallowfield L, Howitt BE, Sehouli J, Karlan BY. Ovarian cancer. Nat Rev Dis Primers. 2016;2: 16061.
2. Bast RC Jr, Matulonis UA, Sood AK, Ahmed AA, Amobi AE, Balkwill FR, et al. Critical questions in ovarian cancer research and treatment: Report of an American Association for Cancer Research Special Conference. Cancer. 2019;125: 1963–1972.
3. Vergote I, Coens C, Nankivell M, Kristensen GB, Parmar MKB, Ehlen T, et al. Neoadjuvant chemotherapy versus debulking surgery in advanced tubo-ovarian cancers: pooled analysis of individual patient data from the EORTC 55971 and CHORUS trials. Lancet Oncol. 2018;19: 1680–1687.
4. Cass I, Baldwin RL, Varkey T, Moslehi R, Narod SA, Karlan BY. Improved survival in women with BRCA-associated ovarian carcinoma. Cancer. 2003;97: 2187–2195.

5. Muggia FM. Sequential single agents as first-line chemotherapy for ovarian cancer: a strategy derived from the results of GOG-132. Int J Gynecol Cancer. 2003;13 Suppl 2: 156–162.
6. Cannistra SA. Cancer of the ovary. N Engl J Med. 2004;351: 2519–2529.
7. McGuire WP, Penson RT, Gore M, Herraez AC, Peterson P, Shahir A, et al. Randomized phase II study of the PDGFRα antibody olaratumab plus liposomal doxorubicin versus liposomal doxorubicin alone in patients with platinum-refractory or platinum-resistant advanced ovarian cancer. BMC Cancer. 2018;18: 1292.
8. Sayal K, Gounaris I, Basu B, Freeman S, Moyle P, Hosking K, et al. Epirubicin, cisplatin, and capecitabine for primary platinum-resistant or platinum-refractory epithelial ovarian cancer: Results of a retrospective, single-institution study. Int J Gynecol Cancer. 2015;25: 977–984.
9. McMullen M, Karakasis K, Madariaga A, Oza AM. Overcoming platinum and PARP-inhibitor resistance in ovarian cancer. Cancers (Basel). 2020;12: 1607.
10. Markman M, Rothman R, Hakes T, Reichman B, Hoskins W, Rubin S, et al. Second-line platinum therapy in patients with ovarian cancer previously treated with cisplatin. J Clin Oncol. 1991;9: 389–393.
11. Vaughan S, Coward JI, Bast RC Jr, Berchuck A, Berek JS, Brenton JD, et al. Rethinking ovarian cancer: recommendations for improving outcomes. Nat Rev Cancer. 2011;11: 719–725.
12. Wethington SL, Wahner-Hendrickson AE, Swisher EM, Kaufmann SH, Karlan BY, Fader AN, et al. PARP inhibitor maintenance for primary ovarian cancer - A missed opportunity for precision medicine. Gynecol Oncol. 2021;163: 11–13.
13. DiSilvestro P, Banerjee S, Colombo N, Scambia G, Kim B-G, Oaknin A, et al. Overall survival with maintenance olaparib at a 7-year follow-up in patients with newly diagnosed advanced ovarian cancer and a BRCA mutation: The SOLO1/GOG 3004 trial. J Clin Oncol. 2023;41: 609–617.
14. Ray-Coquard I, Pautier P, Pignata S, Pérol D, González-Martín A, Berger R, et al. Olaparib plus Bevacizumab as First-Line Maintenance in Ovarian Cancer. N Engl J Med. 2019;381: 2416–2428.
15. González-Martín A, Pothuri B, Vergote I, DePont Christensen R, Graybill W, Mirza MR, et al. Niraparib in patients with newly diagnosed advanced ovarian cancer. N Engl J Med. 2019;381: 2391–2402.
16. Moore K, Colombo N, Scambia G, Kim B-G, Oaknin A, Friedlander M, et al. Maintenance olaparib in patients with newly diagnosed advanced ovarian cancer. N Engl J Med. 2018;379: 2495–2505.
17. Huang D, Savage SR, Calinawan AP, Lin C, Zhang B, Wang P, et al. A highly annotated database of genes associated with platinum resistance in cancer. Oncogene. 2021;40: 6395–6405.

18. Nguyen B, Fong C, Luthra A, Smith SA, DiNatale RG, Nandakumar S, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell. 2022;185: 563–575.e11.
19. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474: 609–615.
20. Gallagher DJ, Konner JA, Bell-McGuinn KM, Bhatia J, Sabbatini P, Aghajanian CA, et al. Survival in epithelial ovarian cancer: a multivariate analysis incorporating BRCA mutation status and platinum sensitivity. Ann Oncol. 2011;22: 1127–1132.
21. Patch A-M, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S, et al. Whole-genome characterization of chemoresistant ovarian cancer. Nature. 2015;521: 489–494.
22. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30: 413–421.
23. Donehower LA, Soussi T, Korkut A, Liu Y, Schultz A, Cardenas M, et al. Integrated analysis of TP53 gene and pathway alterations in The Cancer Genome Atlas. Cell Rep. 2019;28: 1370–1384.e5.
24. Frum RA, Grossman SR. Mechanisms of mutant p53 stabilization in cancer. Subcell Biochem. 2014;85: 187–197.
25. Pfister NT, Prives C. Transcriptional regulation by wild-type and cancer-related mutant forms of p53. Cold Spring Harb Perspect Med. 2017;7. doi:10.1101/cshperspect.a026054
26. Andrysik Z, Galbraith MD, Guarnieri AL, Zaccara S, Sullivan KD, Pandey A, et al. Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. Genome Res. 2017;27: 1645–1657.
27. Polager S, Ginsberg D. p53 and E2f: partners in life and death. Nat Rev Cancer. 2009;9: 738–748.
28. Feng Z, Zhang H, Levine AJ, Jin S. The coordinate regulation of the p53 and mTOR pathways in cells. Proc Natl Acad Sci U S A. 2005;102: 8204–8209.
29. Watkins JA, Irshad S, Grigoriadis A, Tutt ANJ. Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. Breast Cancer Res. 2014;16: 211.
30. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, et al. Mutational processes shape the landscape of TP53 mutations in human cancer. Nat Genet. 2018;50: 1381–1387.
31. Pennington KP, Walsh T, Harrell MI, Lee MK, Pennil CC, Rendi MH, et al. Germline and somatic mutations in homologous recombination genes predict platinum response and survival in ovarian, fallopian tube, and peritoneal carcinomas. Clin Cancer Res. 2014;20: 764–775.

32. High-Grade Serous Carcinoma. In: Ovarian Cancer Research Alliance [Internet]. 27 Jul 2021 [cited 6 Nov 2024]. Available: https://ocrahope.org/news/high-grade-serous-carcinoma/
33. Christie EL, Bowtell DDL. Acquired chemotherapy resistance in ovarian cancer. Ann Oncol. 2017;28 Suppl 8: viii13–viii15.
34. Chowdhury S, Kennedy JJ, Ivey RG, Murillo OD, Hosseini N, Song X, et al. Proteogenomic analysis of chemo-refractory high-grade serous ovarian cancer. Cell. 2023;186: 3476–3498.e35.
35. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. 2014 [cited 8 May 2024]. Available: https://escholarship.org/uc/item/1h3515gn
36. Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun. 2018;9: 4038.
37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
39. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122.
40. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17: 118.
41. Barnell EK, Ronning P, Campbell KM, Krysiak K, Ainscough BJ, Sheta LM, et al. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. Genet Med. 2019;21: 972–981.
42. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell. 2019;179: 964–983.e31.
43. Gey S, Lebarbier E. Using CART to detect multiple change points in the mean for large sample. 2008. Available: https://hal.science/hal-00327146/document
44. Bellman R. On the approximation of curves by line segments using dynamic programming. Commun ACM. 1961;4: 284.
45. Pierre-Jean M, Rigaill G, Neuvial P. Performance evaluation of DNA copy number segmentation methods. Brief Bioinform. 2015;16: 600–615.
46. Lebarbier E. Detecting multiple change-points in the mean of Gaussian process by model selection. Signal Processing. 2005;85: 717–736.
47. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11: 13.

48. Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller M-C, Shaikh N, et al. Replication stress links structural and numerical cancer chromosomal instability. Nature. 2013;494: 492–496.
49. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, Favero F, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. NPJ Breast Cancer. 2018;4: 16.
50. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32: 2103–2110.
51. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. Methods Mol Biol. 2016;1418: 283–334.
52. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34: 525–527.
53. Cobain EF, Wu Y-M, Vats P, Chugh R, Worden F, Smith DC, et al. Assessment of clinical benefit of integrative genomic profiling in advanced solid tumors. JAMA Oncol. 2021;7: 525–533.
54. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14: 7.

Chapter 4 Divergent Clonal Evolution and Early Dissemination Promotes Genetic Heterogeneity of Metastases in Castration Resistant Prostate Cancer

Portions of this chapter is currently under review at Cancer Research.

4.1 Introduction

Every cancer harbors some degree of heterogeneity [1–3], which arises from genetic alterations in a subset of cells within a tumor (*i.e.* clones), resulting in functional and phenotypic differences among clones [1]. Prostate cancer (PCa) is known to be a notably genetically heterogeneous disease [4,5], which presents challenges in diagnosis and treatment [6]. Genetic heterogeneity across metastatic lesions can lead to resistance to treatment [2,7], and limit the utility of diagnostic approaches, biomarker assays and drug targets that rely on sequencing from a single site [8,9]. Underpinning this heterogeneity is the high clonal diversity [7] and multifocality [10] of metastatic and primary disease, respectively. For example, research by Lovf et al. [11] reports that within a prostatectomy, different foci frequently have distinct driver aberrations, with as many as 76% of PCa foci sharing no mutations. Kim et al. [12] described that even in patients where there is a common ancestor among different foci, a single primary tumor biopsy can harbor as few as 20% of the mutations detected across all foci.

While the genetic and histologic heterogeneity of primary and loco-regional prostate cancer is well-documented [4,9–11,13], most studies of metastatic castration resistant prostate cancer (mCRPC) have relied primarily on a single sample from each patient, which

underestimates the disease's heterogeneity [4,14]. Landmark multi-site sequencing studies of mCRPC have uncovered complex patterns and mechanisms of seeding [7,15] and significant *AR* genomic complexity [7,16], a striking finding considering that lethal prostate cancer (PCa) has passed through at least two evolutionary bottlenecks; acquisition of metastatic potential, and development of castration resistance. However, these studies are also limited by shallow sequencing depth, small number of patients, or narrow focus [7,15,16]. The breadth of clonal heterogeneity within mCRPC is inadequately described, thus limiting our understanding of the factors related to clonal selection and genetic drift. Understanding these factors may allow interventions to block the two most important events in the disease process, metastatic spread and resistance.

This chapter addresses this gap in knowledge by analyzing a cohort of 26 postmortem mCRPC patients with distant metastases enrolled in our rapid autopsy program (MLTP or Michigan Legacy Tissue Program); in this program autopsies are performed and tissues collected from patients with heavily treated mCRPC patients. Leveraging high-depth sequencing, and mathematical modeling, we characterized the extent of mutational, copy-number, and clonal heterogeneity. We quantitatively evaluate the impact of inter-tumor genetic heterogeneity on patient mutation detection rates. Furthermore, to uncover and understand pathways to metastasis between different tumor sites, we explored phylogenies of tumor clonal evolution, patterns of seeding and migration, and evolutionary timing of genetic alterations. Lastly, we characterized an instructive case of mCRPC with distinct differentiation phenotypes, very early clonal branching and metastatic dissemination. Altogether, this study highlights the independent evolution of metastatic lesions, and its implications for diagnostic and targeted therapy strategies.

4.2 Results

4.2.1 Multi-site sequencing of a rapid autopsy cohort of mCRPC

To understand and characterize intra-patient [2] heterogeneity in mCRPC we performed matched tumor-normal whole-exome DNA sequencing (WES) and tumor whole-transcriptome RNA sequencing (capture RNA-Seq [17]) on a rapid autopsy cohort of 26 patients (**Figure 4.1**) [18,19]; for a total of 93 tumors (median 3 metastatic sites per patient).

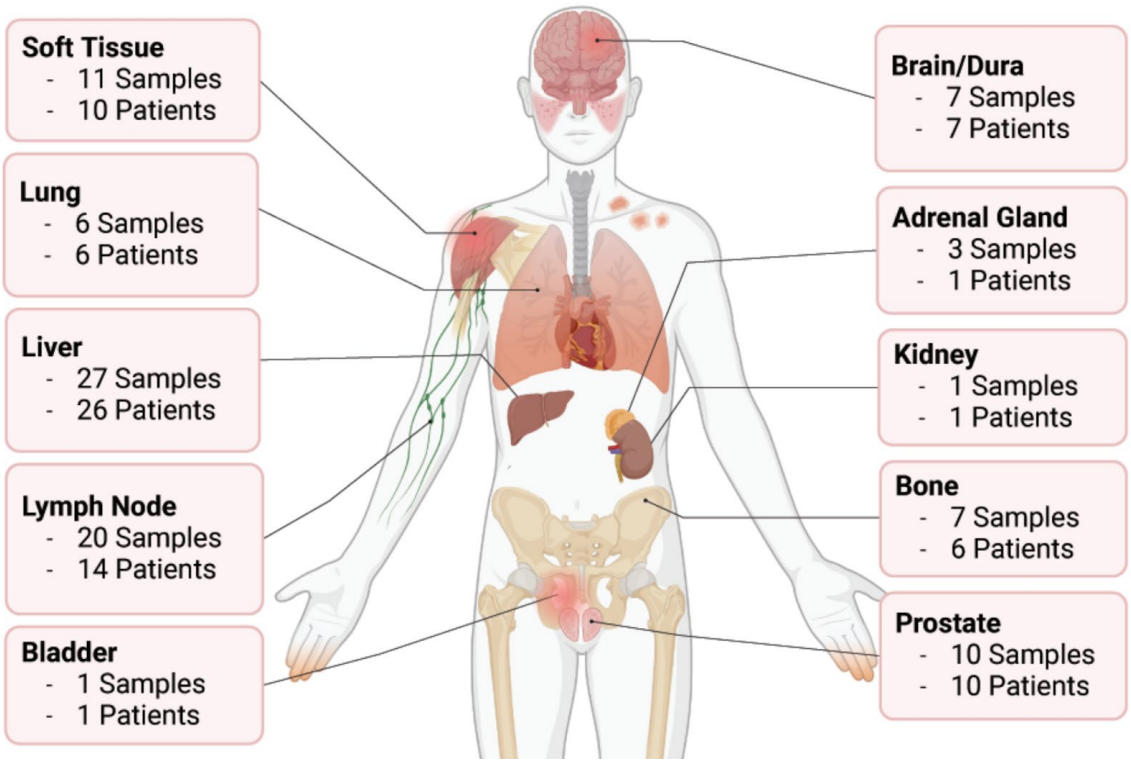


Figure 4.1: Overview of the cohort by number of patients and samples obtained from each tissue.

We profiled the matched tumor and normal samples to an average depth of 590X (95% CI [553X 627X]) and 520X (95% CI [360X 680X]), respectively. We performed extensive QC (**Figure 4.2**), focusing on library complexity and sequencing coverage, tumor purity and tumor-in-normal contamination [20], and excluded three samples with low tumor

content (purity<10%). Utilizing the WES data, we called both somatic and germline variants (**Methods**), inferred absolute copy number variants (CNV) [21] and tumor purity and ploidy.

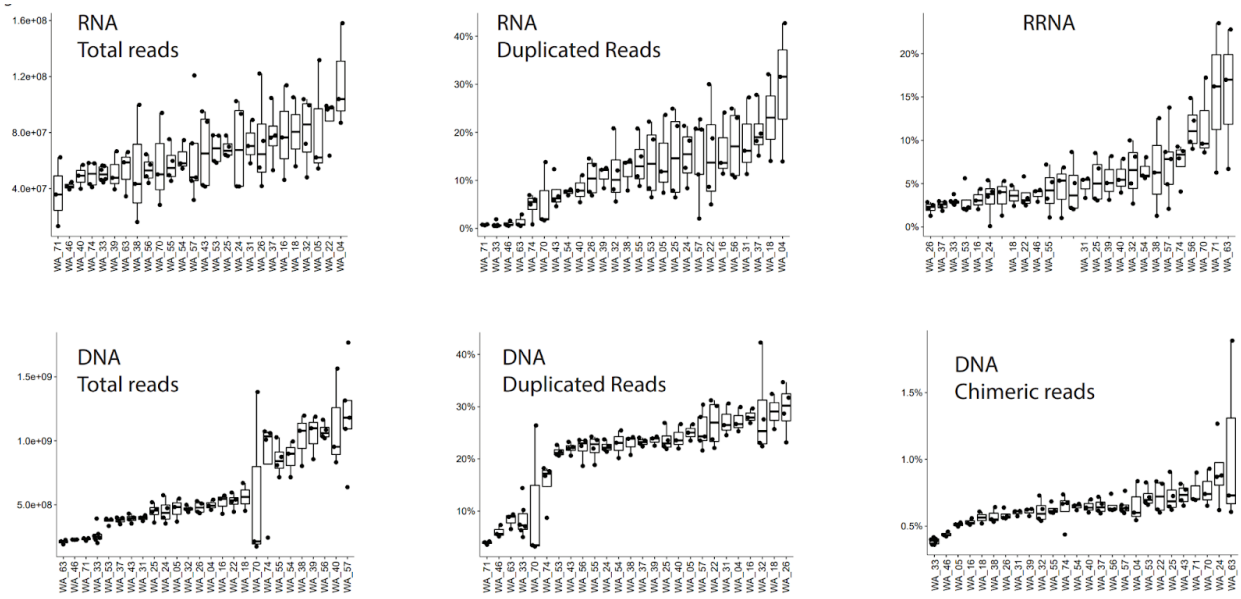


Figure 4.2: QC measures for RNA (top) and DNA (bottom) data

By analyzing independent tumor samples across multiple patients, we aimed to investigate the impact of shared cancer biology (*e.g.,* differences in intrinsic immunogenicity of each cancer) on tumor purity and the tumor microenvironment (**Methods**). We observed, that while tumor purities are highly variable, both within (average IQR=0.15) and across patients (IQR=0.33), they are strongly associated with both the patient (P=0.0011, **Figure 4.3**) and tissue site (P=0.0071, **Figure 4.4**), with lung and bone tissues having systematically the lowest purity.

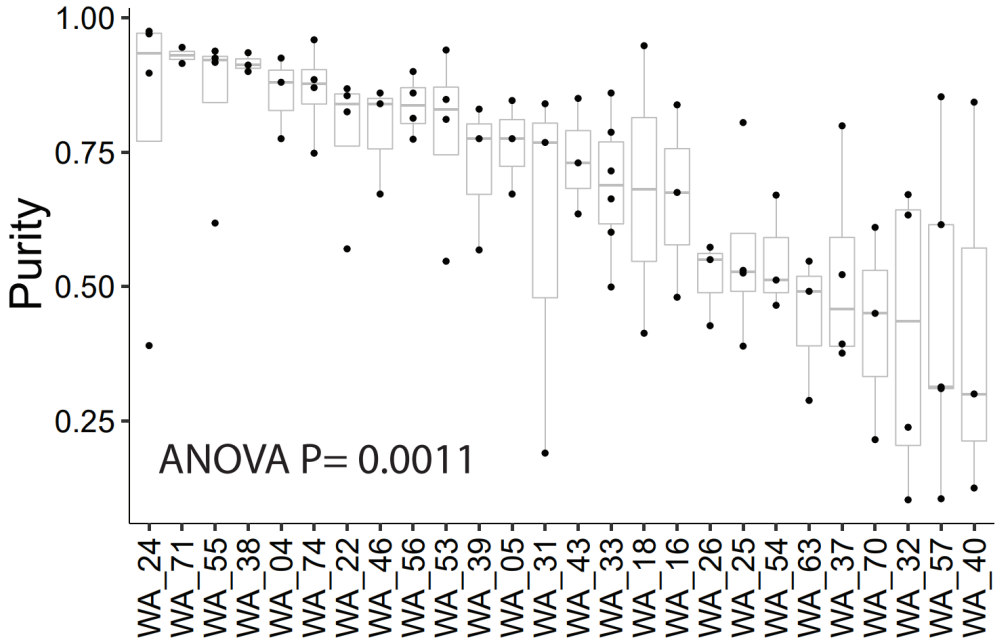


Figure 4.3: Purity (based on NGS) of tumors stratified by patient (F=3.49, Degree of freedom=25). Each dot represents one tumor. ANOVA P-value indicates that purity is significantly associated with the patient.

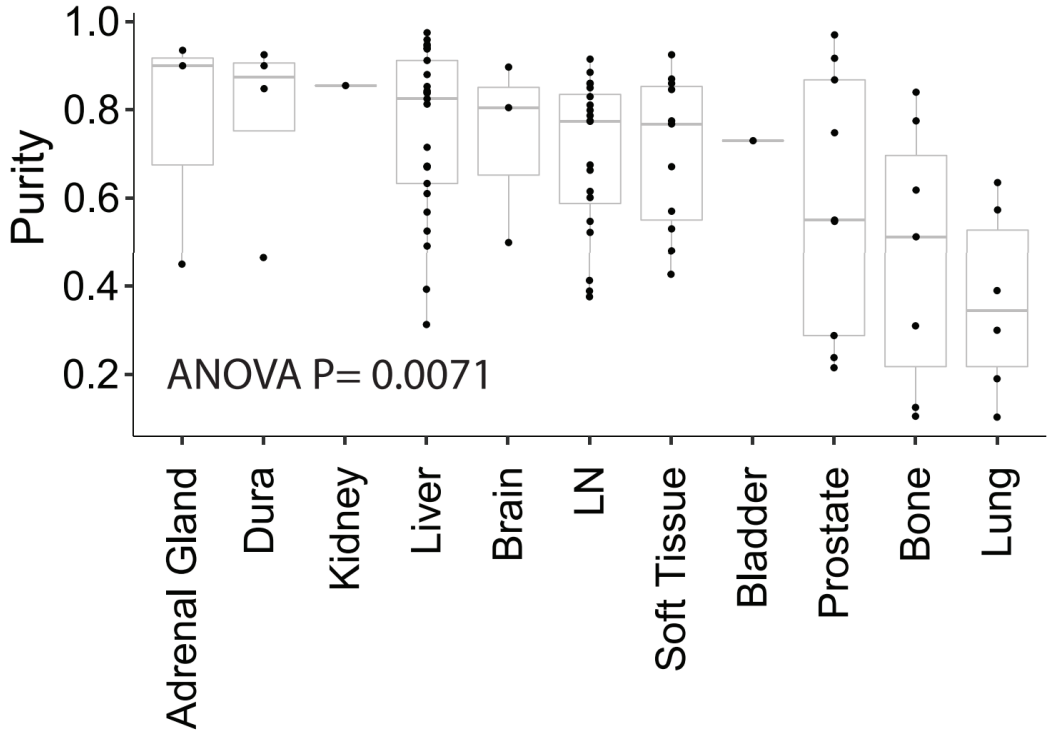


Figure 4.4: Purity (based on NGS) of tumors stratified by tissue (F=3.19, Degree of freedom=10). Each dot represents one tumor. ANOVA P-value indicates that purity is significantly associated with the tissue.

In agreement with prior studies [22], our cohort revealed a high incidence of non-synonymous genetic alterations in key genes associated with late-stage mCRPC, notably mutations in *TP53*, *AR*, *PTEN*, and *ETS* fusions (**Figure 4.5**).

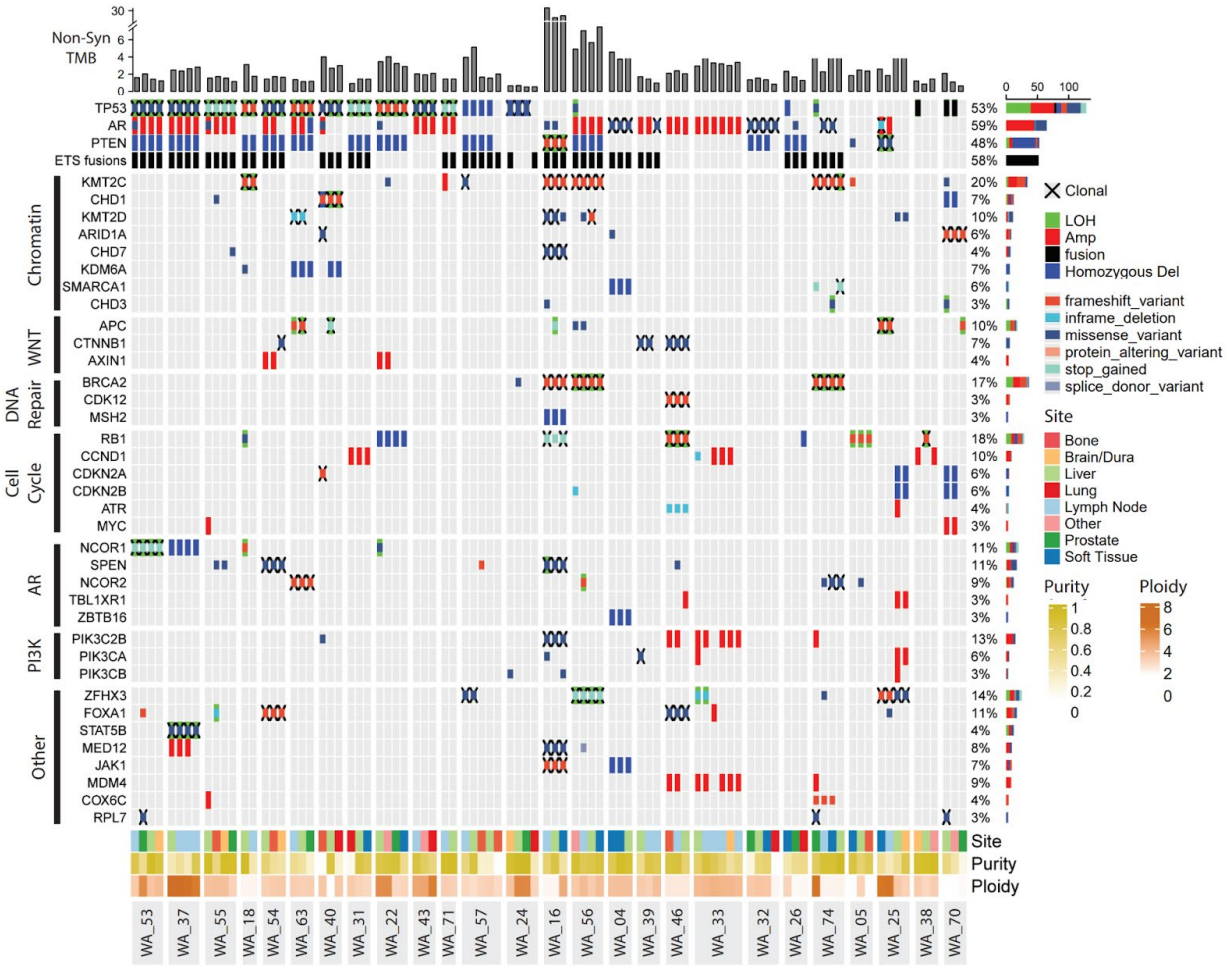


Figure 4.5: Alteration landscape of the cohort. Genes recurrently mutated in PCa and mutated at least once in this cohort are included. TMB indicates tumor mutation burden based on non-synonymous mutations. LOH is only shown when it is accompanied by a mutation.

ETS-fusions were identified using RNA sequencing (**Methods**) in 62% of our cohort, with a notable 58% of patients harboring these fusions across all of their metastatic sites (**Figure 4.6**).

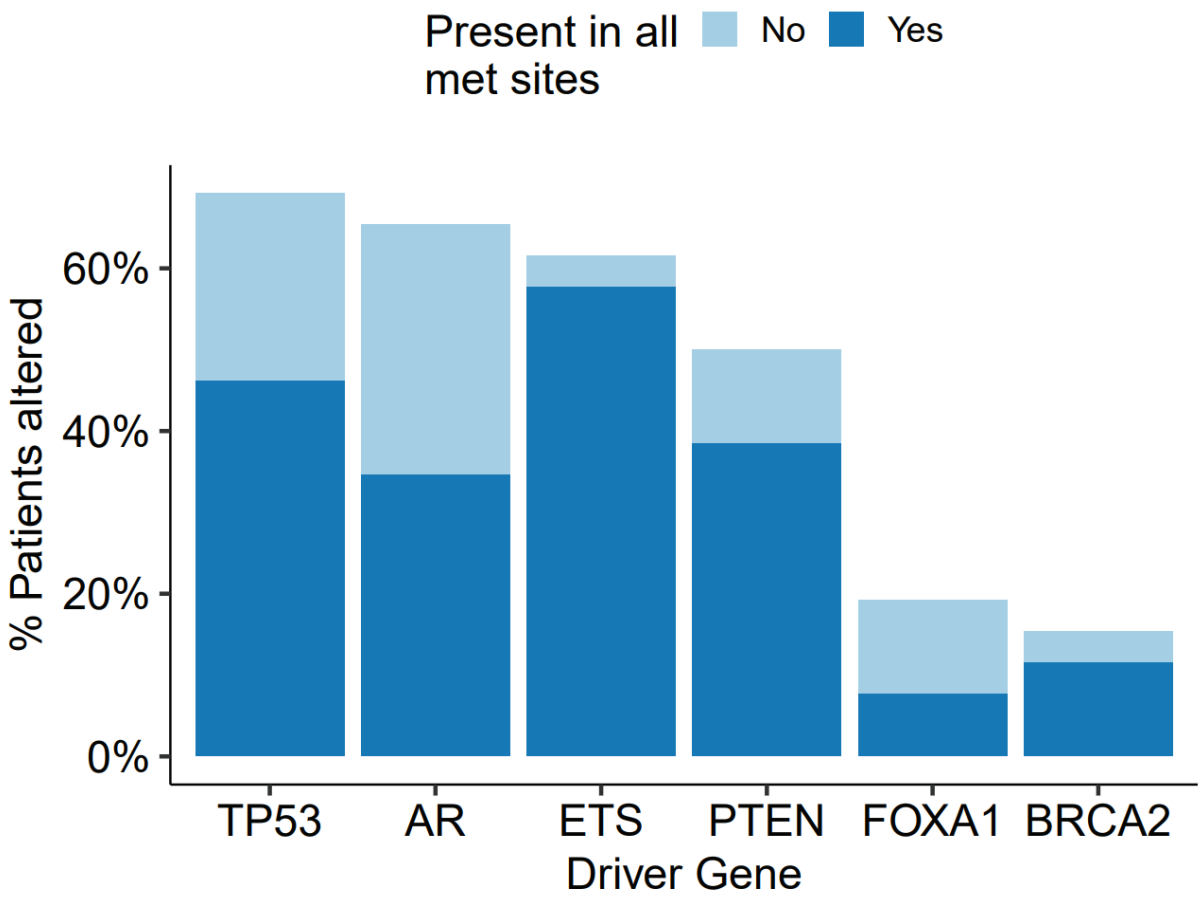


Figure 4.6: Proportion of patients with alterations in genes recurrently altered in PCa (alterations shown in Figure 4.5). Among CNV alterations, only homozygous deletions and amplifications has been considered.

BRCA2, an actionable target in mCRPC [23,24], was present in 15% of patients with 75% of them (3/4 patients) harboring BRCA2 mutations clonally in all sites and one patient harboring a likely passenger missense mutation. In contrast, *TP53* mutations were identified in 69% of patients, but they were present across all metastatic sites of only 46% of patients (**Figure 4.6**). A similar pattern was observed for *AR* alterations, where we found 65% of patients having *AR* alterations, but only 35% of them had the alterations present in all sites. Notably, two patients harbored a clonal hotspot *AR* mutation (Thr878Ala) and *AR* amplification, respectively; indicative of clonal sweep or re-seeding. While another patient had a subclonal hotspot mutation (His875Tyr), not observed in any of the sequenced metastatic sites.

4.2.2 Sequencing a single metastatic site overestimates clonality of mutations

"Illusion of clonality" [14,25] refers to the phenomenon where a mutation appears clonal if only a single sample is sequenced, but the mutation is either absent or subclonal in other metastatic sites. To examine the clonality of mutations, we quantified the CCF [26] for each mutation (**Methods**), the majority being likely passenger mutations, subsequently classifying them as clonal or subclonal based on their CCF values [27,28] (**Methods**). When a single sample is evaluated, 38% (95% CI [34% 45%]) of mutations appear clonal, however, if pairs of samples are jointly evaluated, the proportion of mutations which are present and clonal in both sites drops to 18% (95% CI [14% 23%]) (P=1.12e-5, **Figure 4.7**). To extrapolate this trend to widespread metastatic disease, we analyzed patients with at least four tumor sites sequenced. The rate of clonal mutations which are present in all evaluated sites decreased from 36% (95% CI [26% 45%]) with one evaluated sample to 15% (95% CI [6% 18%]) with two, and 11% (95% CI [3% 12%]) with four, showing that sampling more than two sites provides little benefit in identifying clonal mutations (**Figure 4.8**).

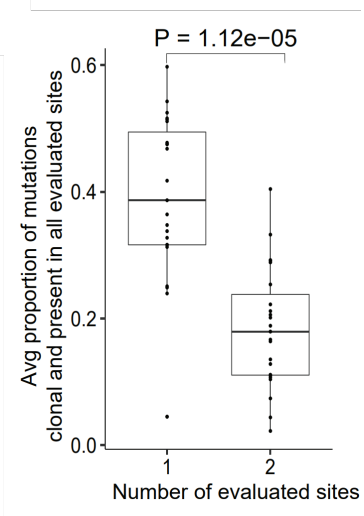


Figure 4.7: Proportion of mutations classified as present and clonal in all sites as a function of the number of sites evaluated.

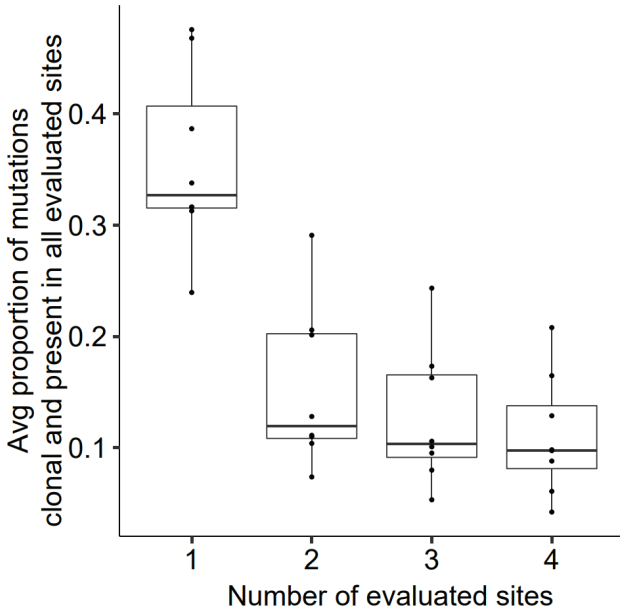


Figure 4.8: Proportion of mutations classified as present and clonal in all sites as a function of the number of sites evaluated. patients with at least four sites sequence are considered.

Next, we investigated the illusion of clonality for likely pathogenic mutations in genes recurrently altered in mCRPC among patients with 4 samples sequenced. For *TP53*, the proportion of ostensibly clonal mutations detected when a single site is sequenced was 58%. However, for 43% of patients all the mutations present across patient sites were clonal (**Figure 4.9**).

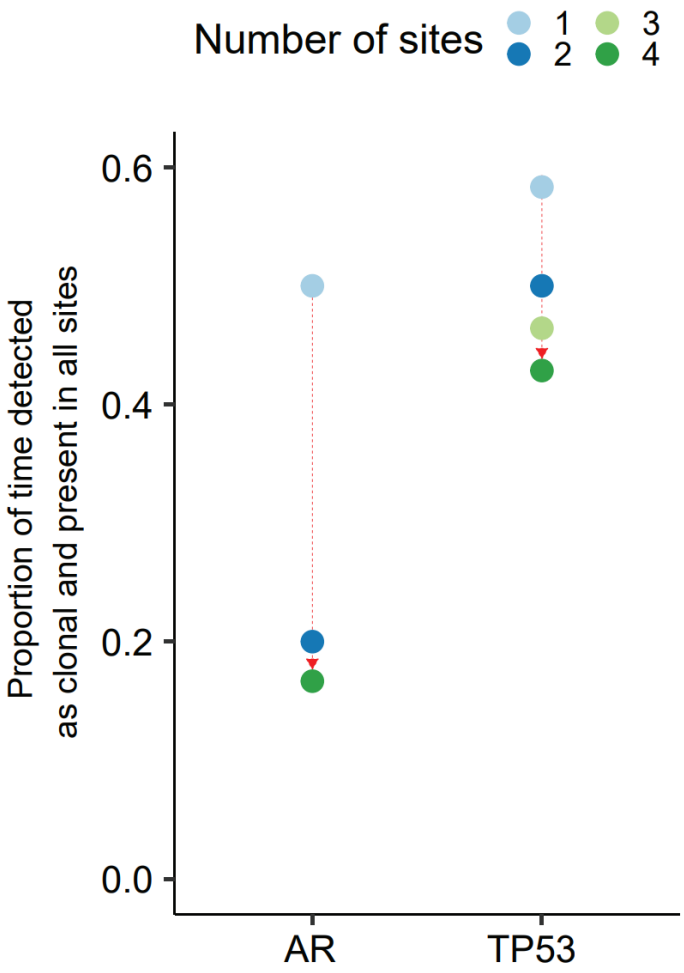


Figure 4.9: Average proportion of AR and TP53 mutations detected as present and clonal in all the evaluated sites. Only patients with at least 4 sequenced tumors are included.

More strikingly, while *AR* mutations were identified as clonal in 50% of patients with *AR* mutations present using a single sample, only 20% of patients harbored clonal *AR* mutations in all sites. We observed that likely pathogenic mutations had the highest average CCF, with the majority being clonal (CCF>0.85) (**Figure 4.10**).

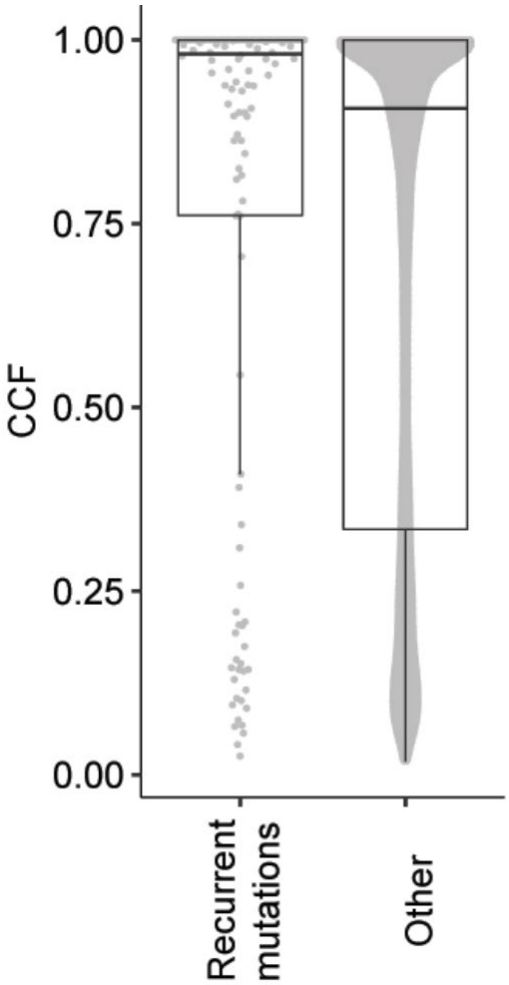


Figure 4.10: CCF of mutations in PCa recurrently mutated genes, compared to all other detected mutations.

CDK12, *PTEN*, and *TP53* hotspot and truncating mutations were also among the mutations with highest CCF while mutations in *AR*, *APC*, and *FOXA1* displayed a bimodal distribution of CCFs, suggesting their both clonal and subclonal prevalence within tumors (**Figure 4.11**). These patterns reinforce the observation that AR and Wnt pathway mutations are more common post-dissemination as a heterogenous resistance response to treatment [7], whereas *TP53* mutations may arise in the most recent common ancestor (MRCA) clone and likely preceding metastasis [29,30].

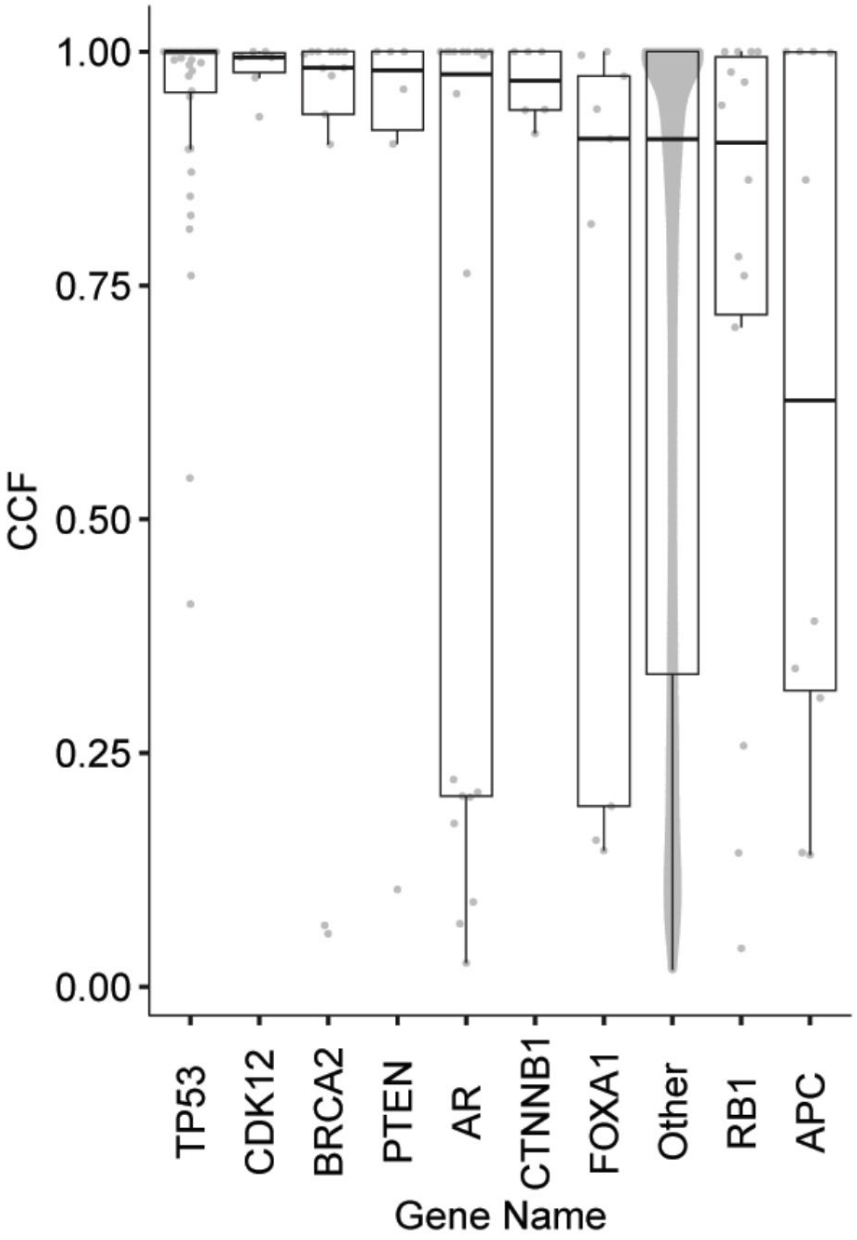


Figure 4.11: CCF of mutations in PCa recurrently mutated genes, compared to all other detected mutations.

4.2.3 Most mutations in mCRPC are subclonal and restricted to a single site

The clonality of a mutation provides information about the fitness and expansion of cells harboring that mutation; while presence of a mutation across different sites indicates whether the mutation appeared before or after the dissemination of clones from their common ancestor. Accordingly, we categorized mutations into five groups (Subclonal-Private,

Subclonal-Shared, Clonal-Private, Clonal-Shared, Truncal) based on their clonal status and distribution across metastatic sites (**Figure 4.12**, **4.13**, **Methods**).

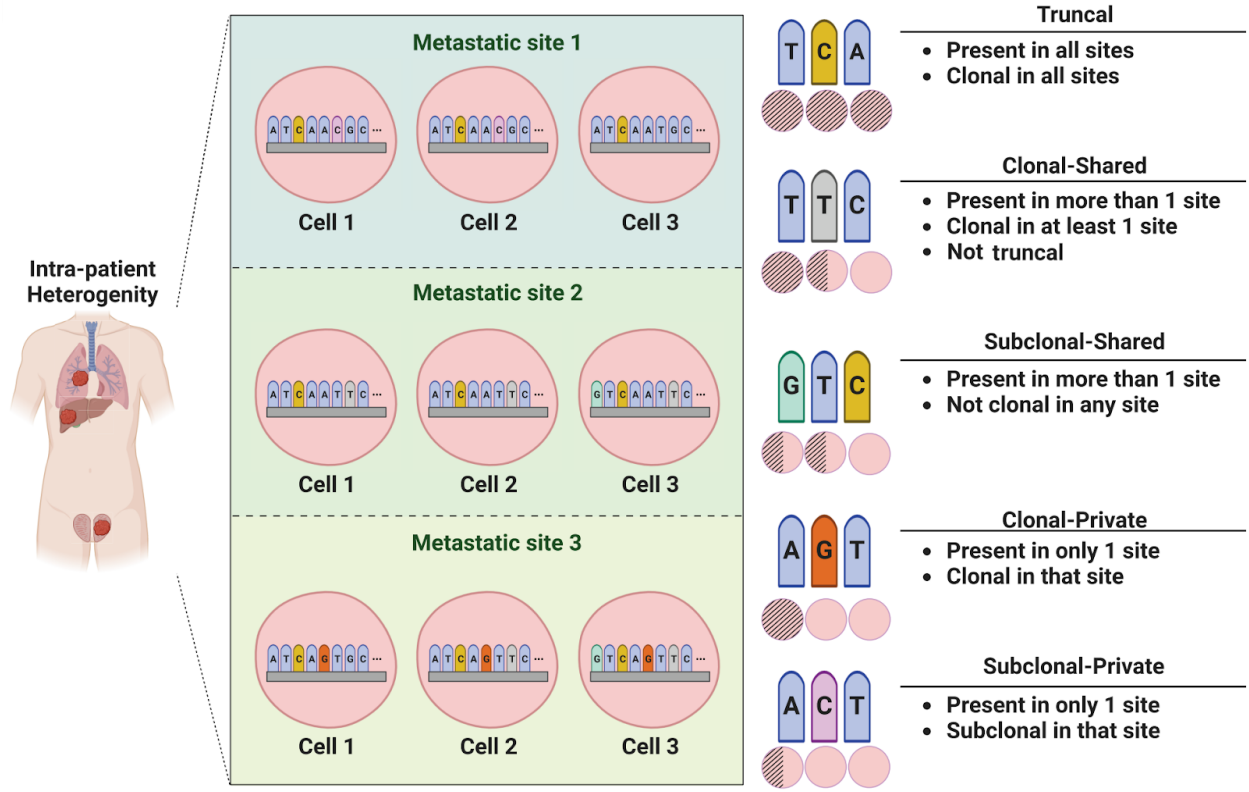


Figure 4.12: Schematic representation of mutation classification used in this study. The left panel represents 3 metastatic sites from a single patient with tumor cells in each site. Mutations present in each cell are indicated by colors other than blue. The right panel lists mutations alongside their classifications according to their presence in metastatic sites and the cells within them (i.e., clonality). Each pink circle in the right panel corresponds to one of the metastatic sites. A fully hatched circle indicates a mutation that is present in all the cells within the site (i.e., clonal), a partially hatched circle signifies a mutation that is present in some but not all of the cells within the site (i.e., subclonal) and a clear circle shows the absence of any mutation. For example, the second mutation (Clonal-Shared) showed by a gray color is present in all cells of metastatic site 2, and only in cell 2 of metastatic site 3, making it present in more than one site (i.e., shared) and clonal in at least one site.

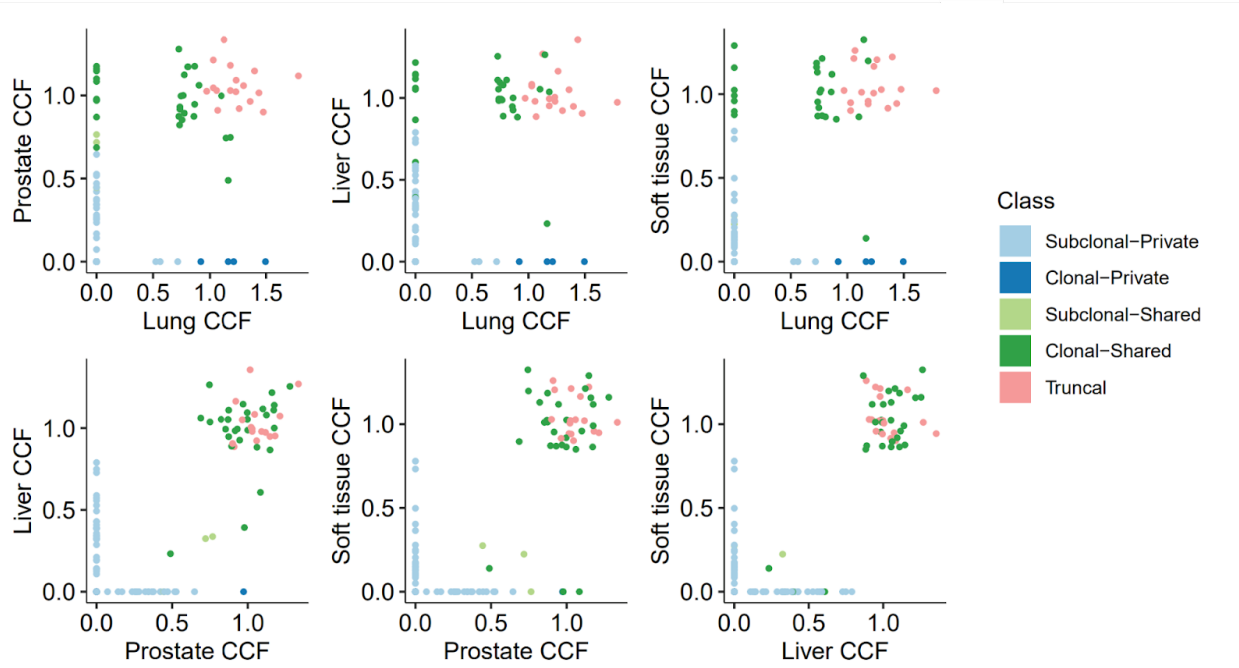


Figure 4.13: Illustration of mutation classification for one patient (WA_32). Private mutations appear always on the edge of all plots, while Truncal mutations are always in the middle. Shared and subclonal mutations have variable positions

We observed that Subclonal-Private mutations were the most common in our cohort, accounting for 55% of all the mutations, while Truncal mutations accounted for only 13%. This demonstrates that the majority of mutations in mCRPC occur after the dissemination of tumors from their common ancestor. While many of those mutations are associated with rapidly expanding clones (clonal sweeps), likely the majority are passengers. Mutations present in some, but not all sites i.e. Clonal-Shared and Subclonal-Shared were observed in patients at varying proportions of 15% and 3% (**Figure 4.14**), which is indicative of complex phylogenetic and dissemination patterns

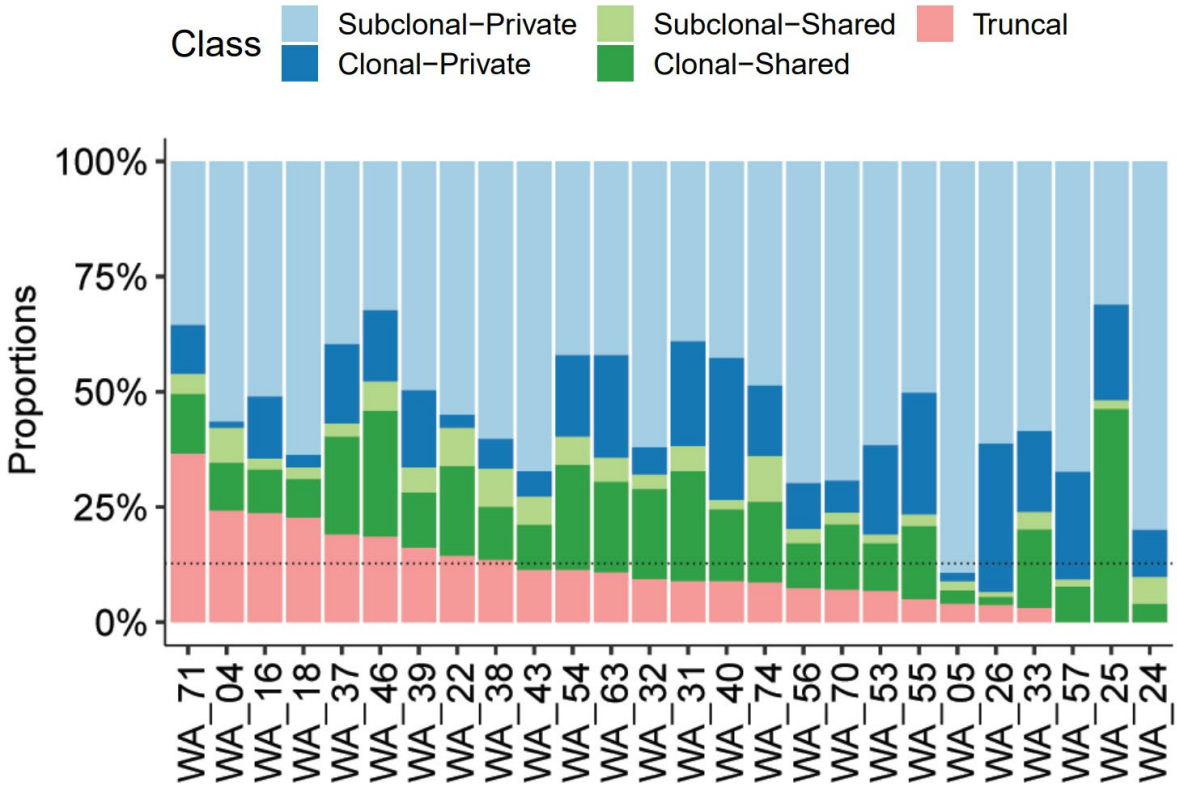


Figure 4.14: The proportion of mutations in each class by each patient

Clonal mutations observed in only one site, i.e. Clonal-Private mutations, which can be indicative of clonal sweeps [31,32], accounted for 14% of all mutations. Finally, we noted the near absence of Truncal mutations in some patients (**Figure 4.14**), e.g. WA_24 which we examine in detail later.

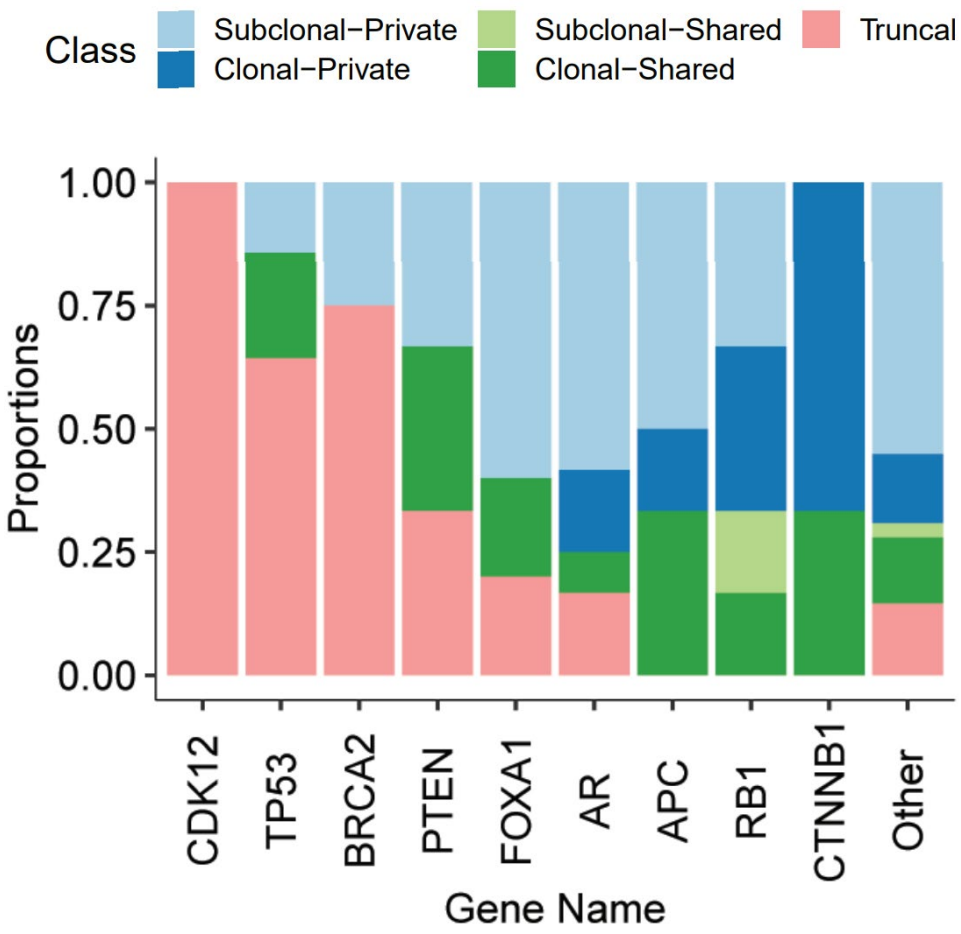


Figure 4.15: The proportion of mutations in each class by each recurrently mutated gene in mCRPC

We extended mutational classification to likely pathogenic mutations in genes commonly altered in mCRPC [22]. We observed that *CDK12* mutations were all Truncal (**Figure 4.15**). This was expected, as *CDK12* mutations have been reported to be early drivers of prostate cancer [33]. Moreover, we observed that mutations in *TP53* were Truncal in over half of the patients (**Figure 4.15**), indicating that, despite the tendency of pathogenic *TP53* mutations to emerge later in the tumor's evolution [22,34,35], a significant proportion occurred prior to tumor clonal dissemination from MRCA. Conversely, mutations in the AR gene, which similarly tend to arise later in the tumor's lifespan, were Truncal in fewer than

20% of cases (**Figure 4.15**). This suggests that AR mutations predominantly occur following tumor spread, consistent with previous reports [16,36].

4.2.4 Single-site sequencing underestimates patient and cohort-level mutation rates

Given the common clinical practice of characterizing metastatic tumors using a single biopsy [37,38], we evaluated what proportion of all mutations (including known pathogenic and likely passenger) present across multiple metastatic sites would be detected if only one sample is sequenced (i.e. per-sample detection rate, Methods) (**Figure 4.16**). As more sites are sequenced the detection rate decreases from 74% (95% CI [94% 54%]) for one additional site to 27% (95% CI [21% 33%]) for four or more. We noted significant variability of those estimates for both individual sites and across patients. Therefore, we explored mutational concordance of tumor pairs, i.e. the proportion of mutations shared between any pair of tumors for each patient. We observed concordance from as high as 60% to as low as 2% (**Figure 4.17**). These trends underscore that, for most mCRPC patients, analyzing a single metastatic tumor site would miss the majority of mutations.

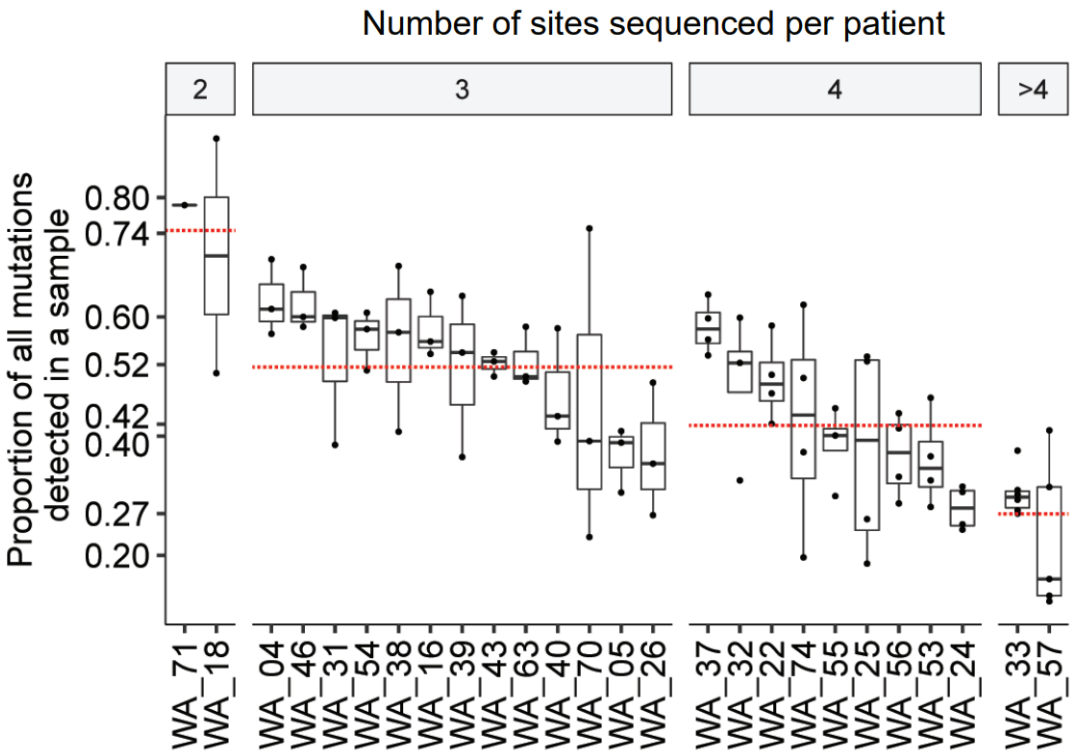


Figure 4.16: Fraction of mutations identified in each individual sample relative to all mutations for that patient in all sites, stratified by the number of sites profiled for a patient. Red dashed lines represent the average proportion of mutations detected.

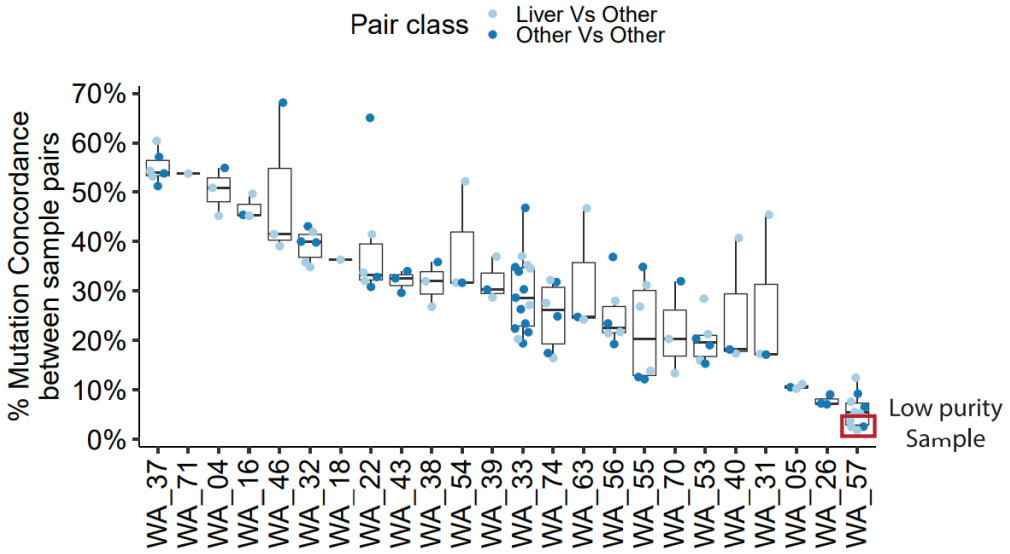


Figure 4.17: Concordance of mutations for all pairs of samples from each patient. Each y-axis value represents the proportion of mutations detected in both of a pair of samples (i.e., concordance). For a patient with n sites, there are n choose 2 pairs.

Next, we evaluated how profiling one compared to multiple sites per patient affects the chance of detecting an alteration (i.e. per-patient detection rate, **Methods**). For likely pathogenic genes in mCPRC, we observed an increased detection rate as the number of sampled sites increased (**Figure 4.18**).

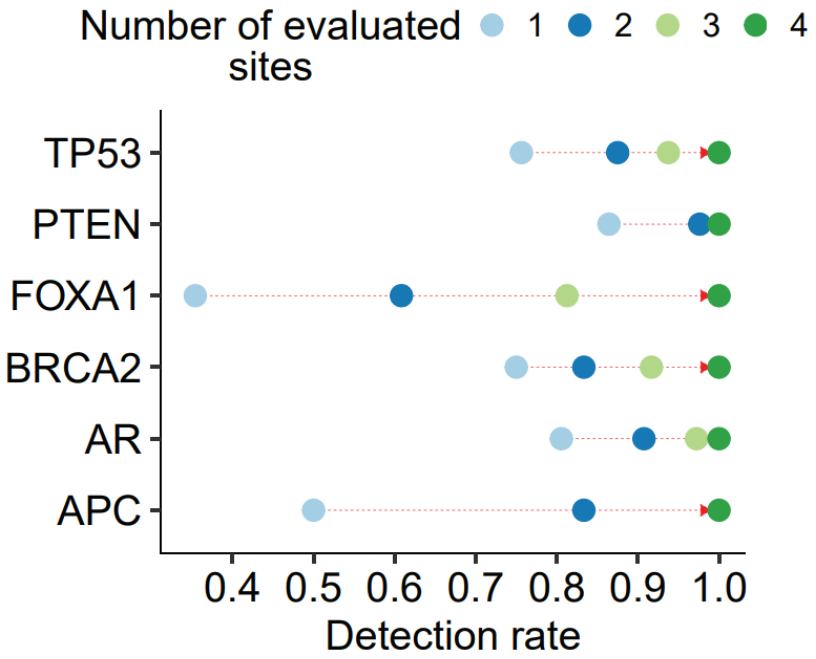


Figure 4.18: Detection rate of alterations for each gene. In this analysis, only patients with at least 4 sequenced tumors are assessed. The detection rate when evaluating four sites is set as a baseline of 1 for each gene. The plot then shows the relative detection rates when fewer than four sites are evaluated

The per-patient detection rate increased the most for FOXA1 and least for early alterations such as PTEN (**Figure 4.18**). Notably, TP53, BRCA2, and AR alterations remained highly detectable even from a single sample, while those in FOXA1, RB1, and APC benefited from sequencing multiple sites. Similarly, sampling a single-site results in an underestimation of cohort-wide alteration frequencies, most notably for APC and AR (**Figure 4.19**).

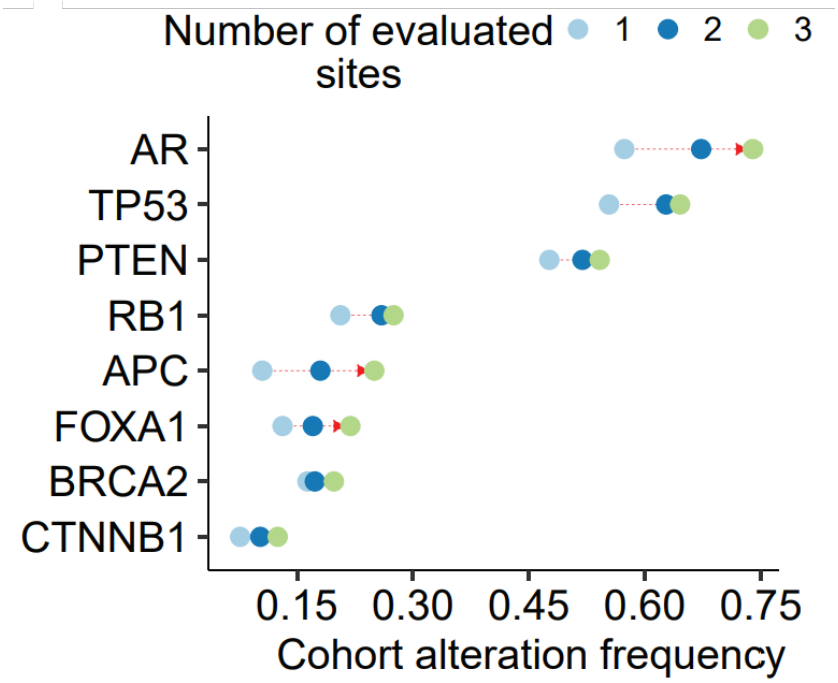


Figure 4.19: Impact of profiling additional tumors on cohort-level observed prevalence of alteration in PCa recurrently mutated genes. Only patients with 3 sequenced tumors are assessed.

Next, we employed a mathematical framework to model the expected yield of mutations with increasing numbers of sequenced samples (**Figure 4.20**, **Methods**).

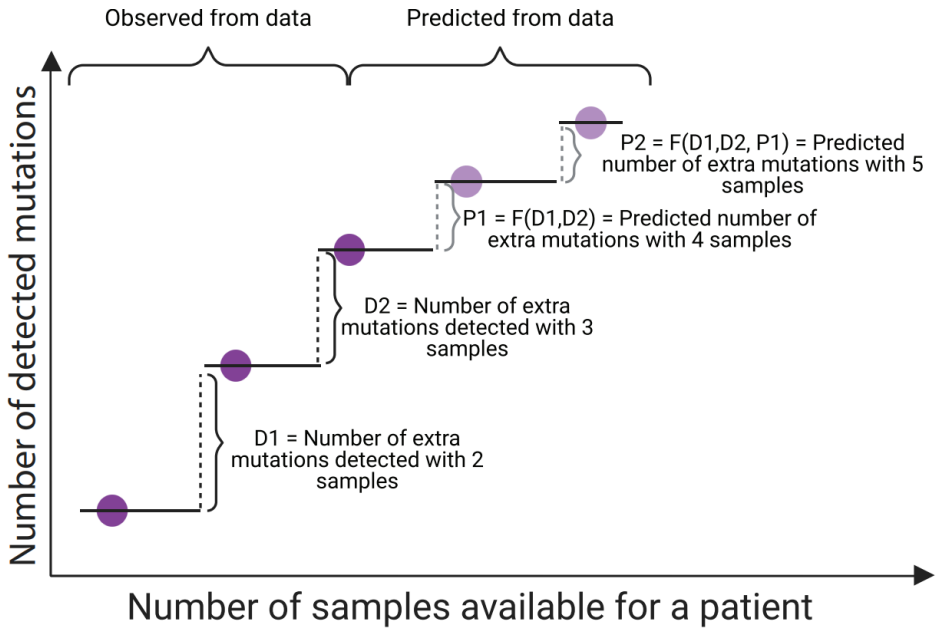


Figure 4.20: Schematic of the 'species accumulation model' (Methods) extrapolating the number of additional mutations that will be detected when another site is sequenced. The model predicts one step ahead, using all previously observed (solid) and predicted (semi-transparent) new mutations.

The model draws on the similarity between calling mutations in multiple tumors and sampling species at distinct ecological sites [39,40]. In this application, the model extrapolates the (unknown) lower bound of the total number of mutations per patient (**Methods**), which enables calculation of the proportion of detected mutations as a function of the number of sequenced sites (**Figure 4.20**). We evaluated the model's performance in predicting the number of new mutations in an unseen sample, and noted a high correlation between the predicted and actual count (R=0.99, P=2.5e-13) with cosine-similarity score of 0.999 (**Figure 4.21**).

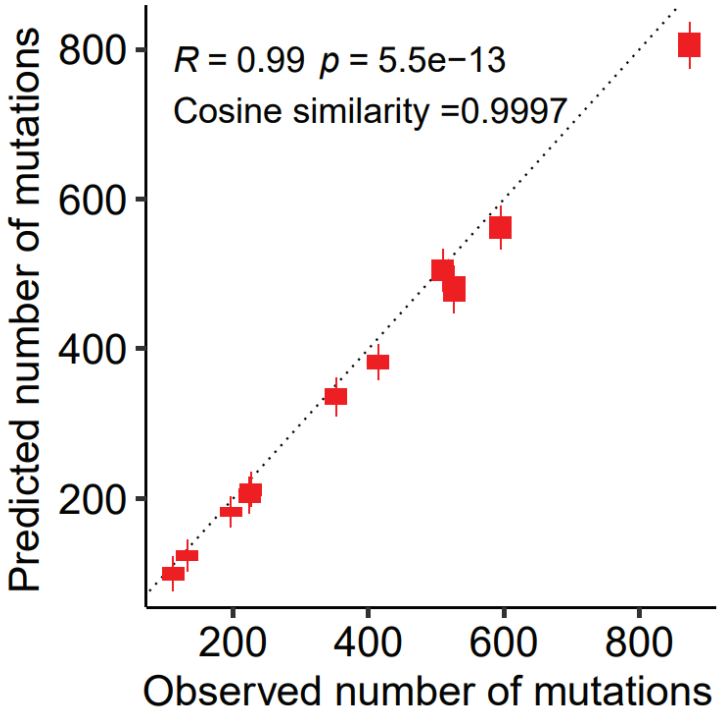


Figure 4.21: Validation of predictions from the model in panel (G). The x-axis shows the observed number of total mutations in four samples. Each y-axis box displays the interquartile range (IQR) of the corresponding prediction based on the data from the first three sites.

By extrapolating the model to a higher number of samples, we reach an asymptote, which can be used as a proxy for the lower bound of the total number of mutations harbored by each patient (**Figure 4.22**).

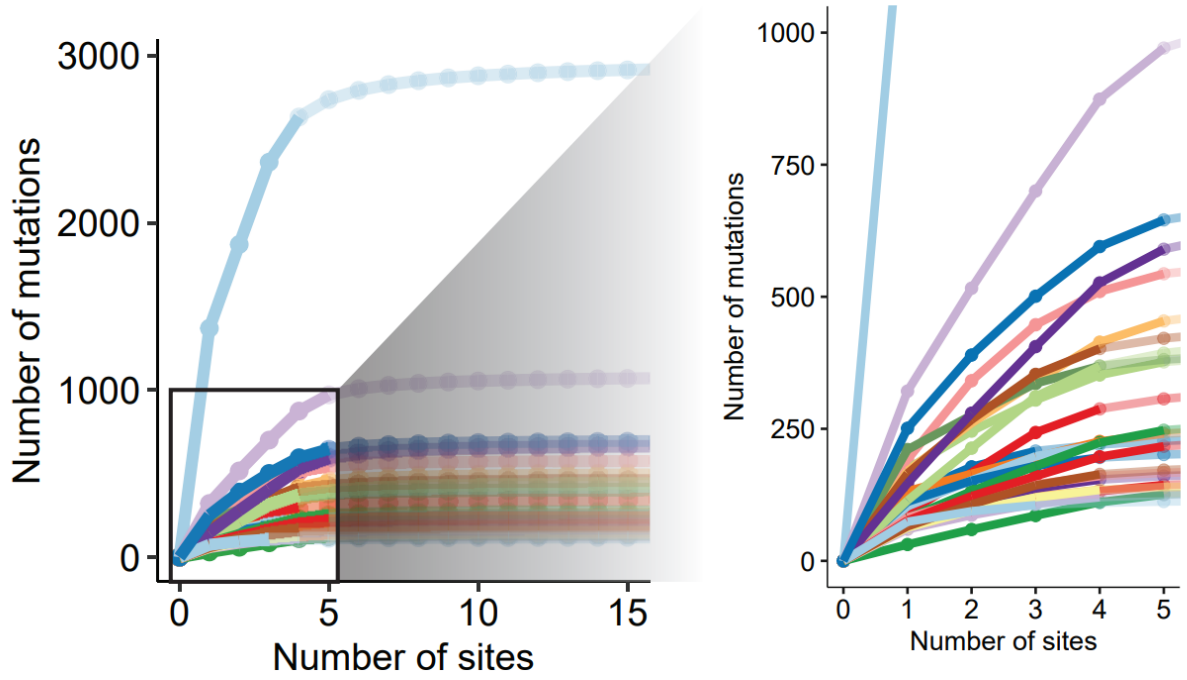


Figure 4.22: Results for the model in Figure 4.21. Observed (solid) and predicted (semi-transparent) number of new mutations per additional site profiled. Each curve represents an individual patient. WA_16 an MSI case is highlighted.

It is evident that this constitutes a lower bound, as each site will harbor a number of private mutations (average 71 mutations, 95% CI [52 91]) (**Figure 4.23**). We evaluated the proportion of detected mutations (relative to the asymptote) as a function of the number of sequenced sites, and found that sequencing a single sample on average uncovers only 40% (95% CI [35% 43%]) of a cancer's mutations (**Figure 4.24**), corroborating the previous results that, on average, less than half of a patient's mutations are captured using a single sample (**Figure 2D**). However, the yield increased to 82.2% (95% CI [80% 85%]) with the inclusion of four samples.

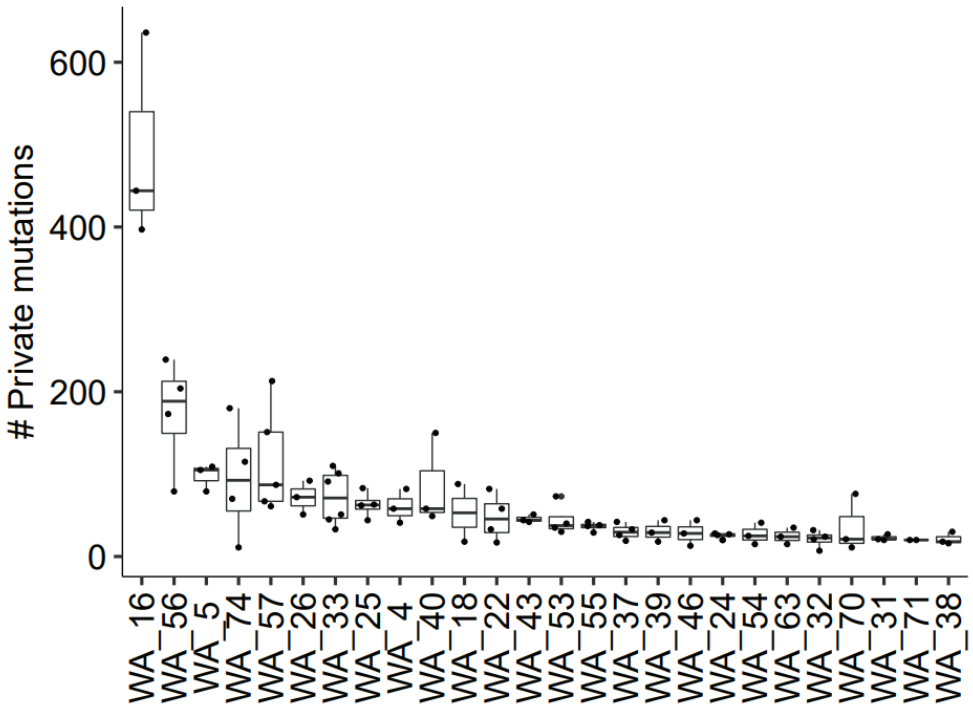


Figure 4.23: Number of private mutations stratified by patient. Each dot represents number of private mutations in one tumor

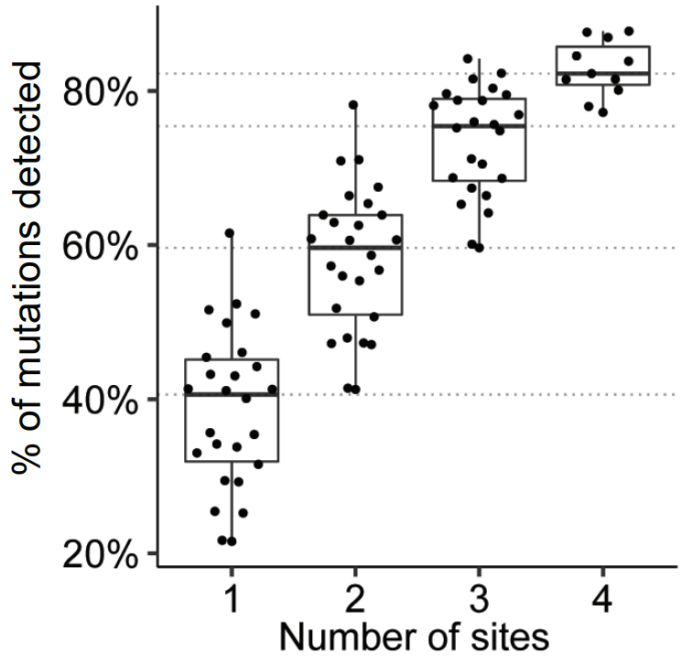


Figure 4.24: Proportion of mutations detected as increasing numbers of sites are evaluated. Each dot represents percent of mutations detected in a group of samples evaluated.

4.2.5 Most CNVs occur after metastasis but pathogenic alterations are largely truncal

We extended our analysis of inter-site mutational heterogeneity to copy-number alterations. First, we focused on whole-genome duplication (WGD) which is present in over 50% of metastatic cancers [41] and typically occurs late in disease progression [42]. We noted that 89% of patients had WGD in at least one of the sites (**Figure 4.25, 4.26**).

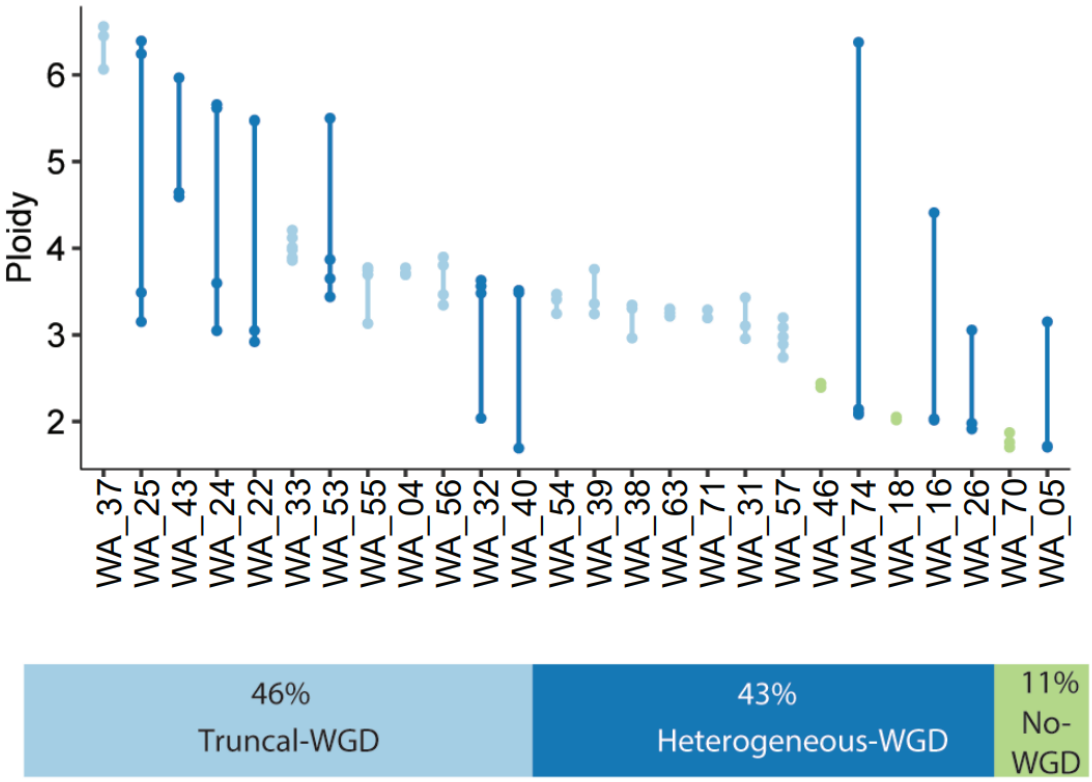


Figure 4.25: Heterogeneity in tumor ploidy across metastatic sites (top). Proportion of patients with truncal, heterogenous, or no observed WGD events (bottom)



Figure 4.26: Per patient and site genome-wide copy number alteration landscape. Colors indicate copy-number alterations (red - gain, blue - loss, green - LOH, gray - diploid). Cases and samples ordered by average ploidy.

However, for 43% of patients the baseline ploidy differed between metastatic sites, indicative of WGD occurrence after metastatic dissemination. Conversely, WGD was identical (i.e. likely truncal, Methods) in 46% of the patients. WGD is associated with an acceleration of chromosomal instability (CIN) [41,43], and hence we examined the inter-site variability of quantitative CIN measures (**Methods**). Despite considerable differences in baseline ploidy, measures of CIN such as %LOH, and LST [44]) were consistent across metastatic sites (**Figure 4.27**).

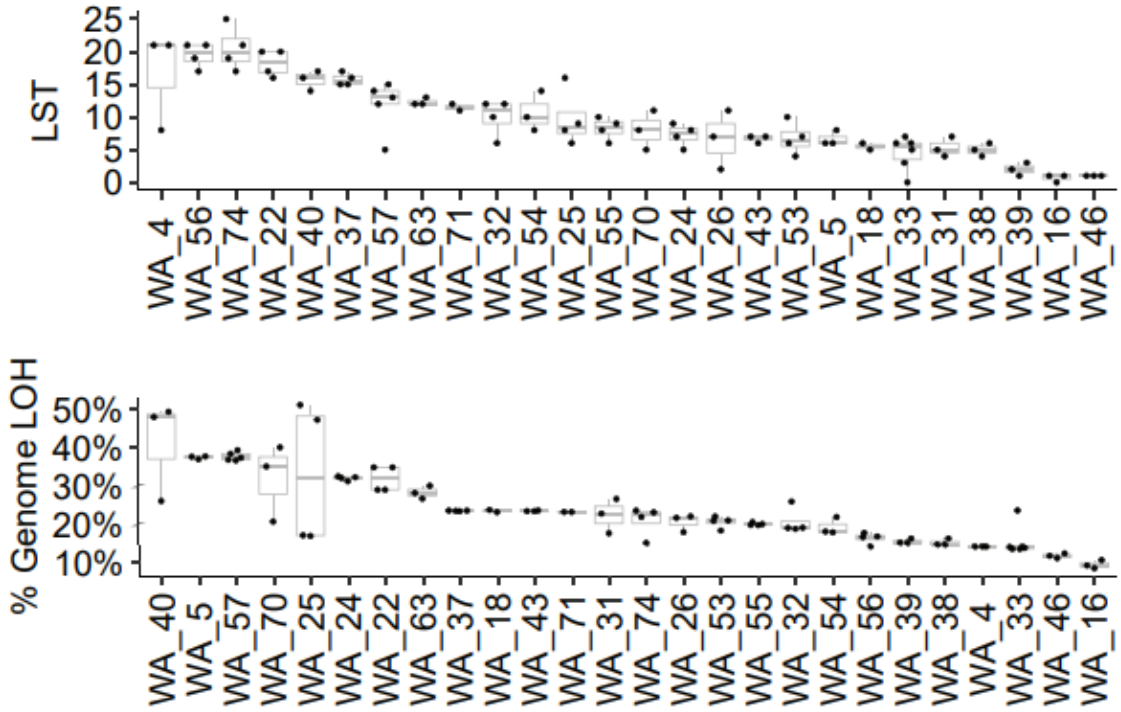


Figure 4.27: Heterogeneity of homologous recombination deficiency (HRD) scores derived from CNV data. Large-scale state transitions (LST) scores (top), and proportion of genome harboring LOH (bottom).

Strikingly, the same holds for tumor mutational burden (TMB), which is similar across metastatic sites despite the significant mutational heterogeneity. This latter finding suggests TMB results from a single site biopsy may still be useful for selecting patients for immune checkpoint inhibitor therapy [45] — a treatment that is approved for solid tumors exhibiting high TMB.

Next, we explored the heterogeneity of copy-number alterations, both genome-wide and for genes recurrently deleted or amplified in PCa [22]. Focusing on the truncal status, we categorized CNVs as truncal if gains or losses were present in all samples from a patient,

and heterogeneous otherwise (**Methods**). Genome-wide, only 37% of gains and 32% of losses were truncal (**Figure 4.28**).

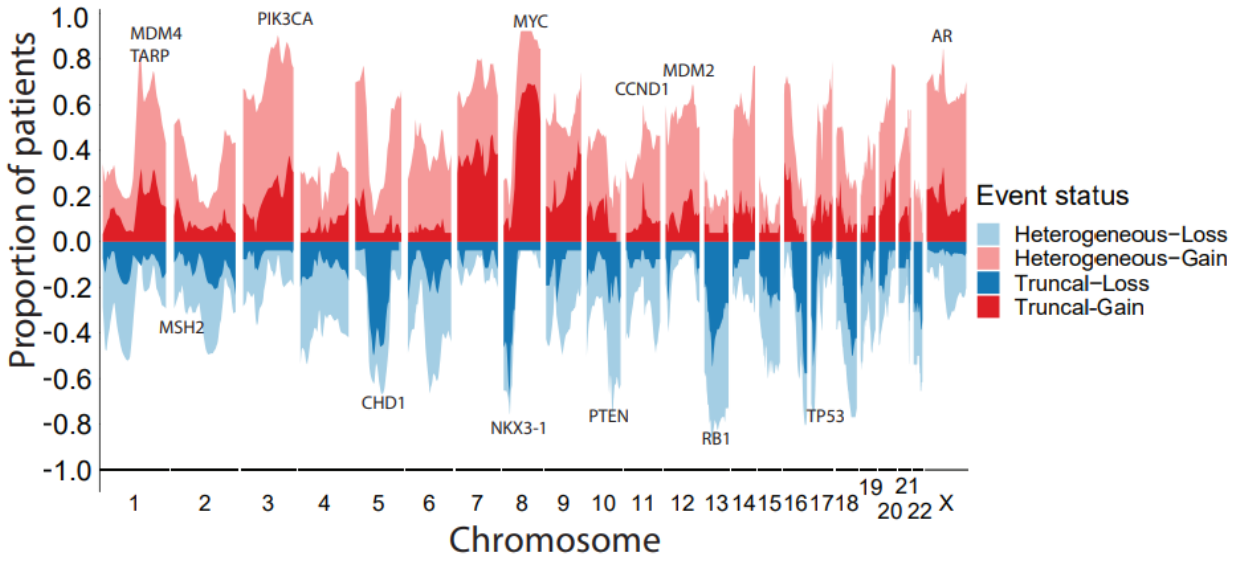


Figure 4.28: Cohort-wide frequency of copy-number gains (red) and losses (blue). Shading indicates truncal (dark) and heterogenous observations (per-patient). Select loci recurrently altered in PCa are highlighted

However, in regions with recurrent copy-number alterations (i.e., specific genomic regions that are consistently amplified or deleted across multiple cancer samples) (**Methods**) we noted a significantly higher degree of truncal gains and losses. Specifically, recurrent deletions, including genes such as *NKX3-1* and *TP53*, were truncal 58% of the time compared to 36% for all other genes (P<1e-16, **Figure 4.29**). Similarly, recurrent gains, including genes such as *MYC* and *CCND1*, exhibited truncal gains 34% of the time compared to 30% for other genes (P=0.01). The *NKX3-1* locus had the highest frequency of truncal losses, while the *MYC* locus showed the highest proportion of truncal gains (71%). *RB1* and *TP53* losses were also commonly truncal (61% and 68% respectively, **Figure 4.30**).

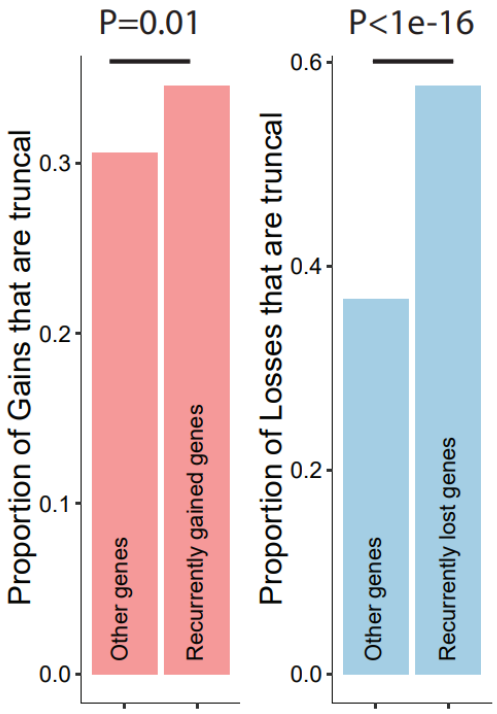


Figure 4.29: Proportion of recurrently truncal gains (left) or losses (right) within regions of the genome stratified by alteration recurrence (Methods).

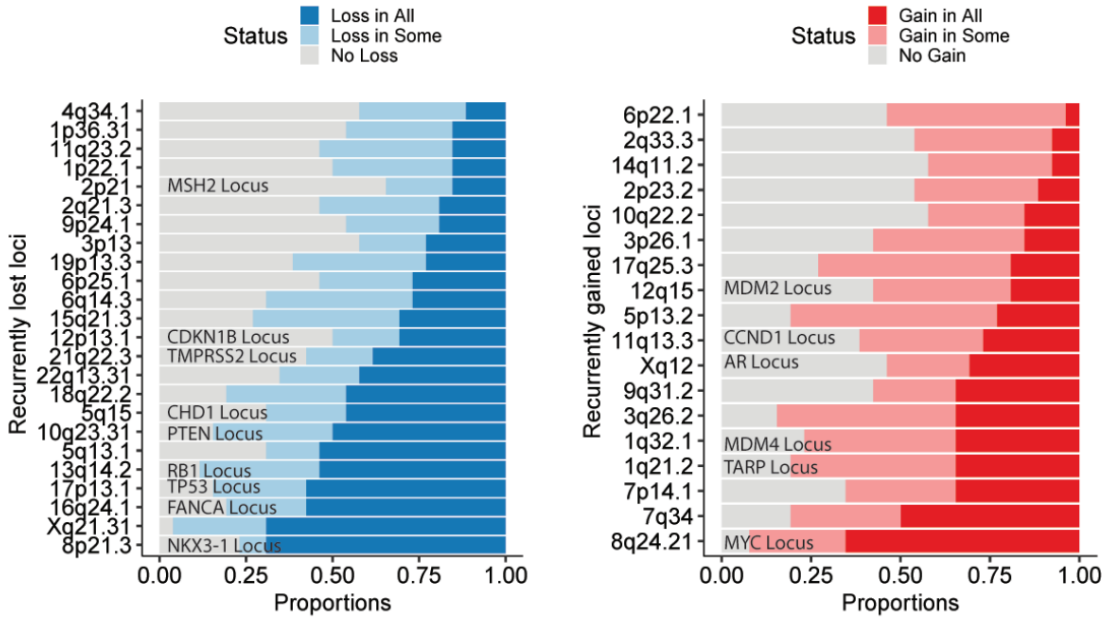


Figure 4.30: Proportion of patients with losses (left) and gains (right) on recurrently deleted and amplified regions in mCRPC. Shading indicates truncal (dark) and heterogenous observations (per-patient)

Moreover, while *PTEN* loss of one or more copies was truncal in 59% of patients, bi-allelic deletions were truncal in 82% of PTEN-deficient patients. Given the role of *PTEN* homozygous deletion in aggressiveness of PCa [46], these results suggest pathogenic *PTEN* alterations happen more frequently before metastasis.

Given the importance of AR alterations in ADT and ARSI resistance [47], and their late emergence, often post-dissemination [7], we sought to assess their heterogeneity in the context of lethal mCRPC. We observed pronounced AR mutational and copy-number heterogeneity (**Figure 4.31**, **4.32**).

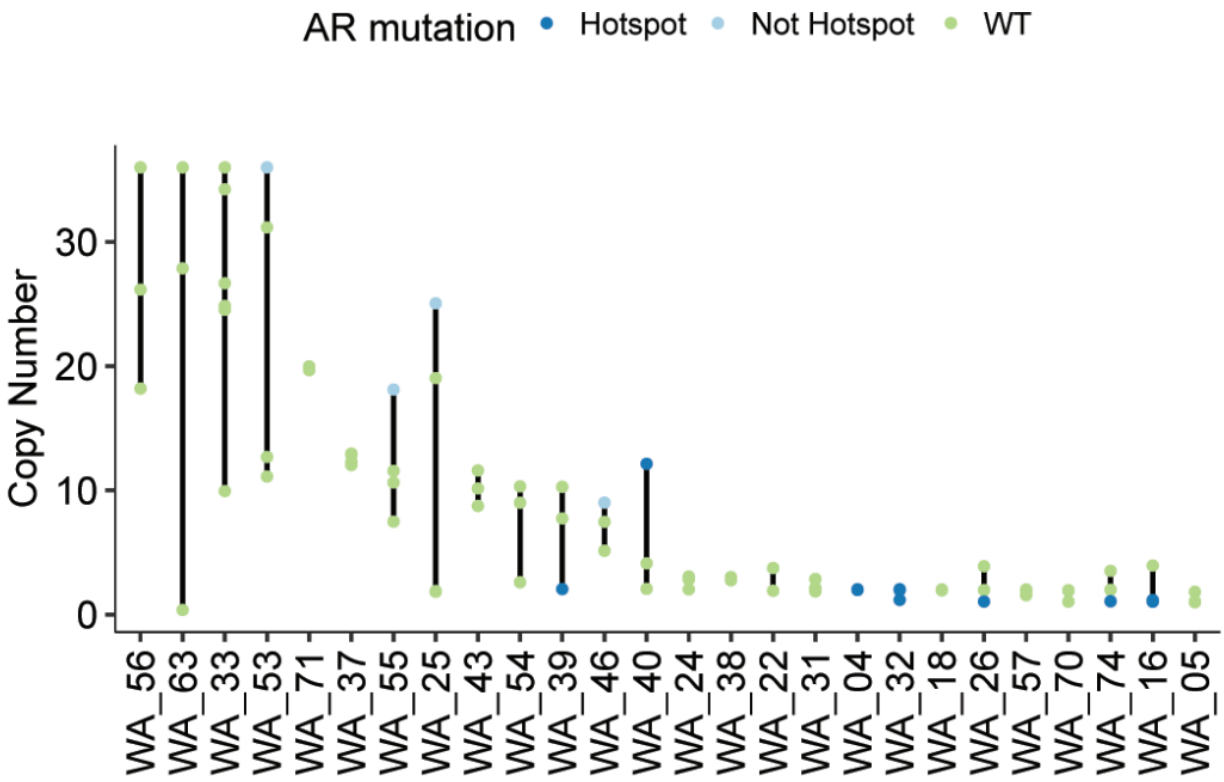


Figure 4.31: Heterogeneity of AR copy number and mutation status across metastatic sites. Presence of AR mutations indicated in (blue). Absolute copy estimates are capped at 35

Amplifications in the AR locus were truncal in 61% of the patients harboring AR amplifications, albeit with significant differences in amplification level [16].

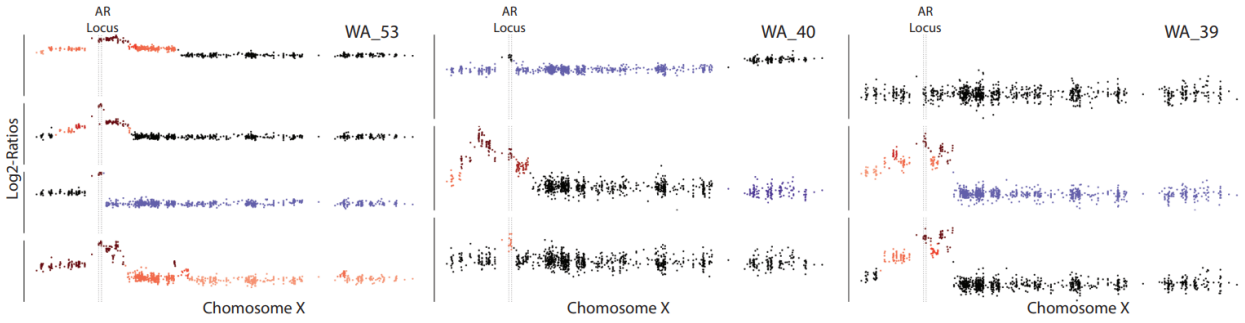


Figure 4.32: AR locus visualization of normalized log2(tumor / normal) coverage ratios across samples from three index cases (black - diploid, red - gain, blue - loss).

Notably, we found that samples with elevated AR gene copy numbers generally harbored fewer AR mutations, and interestingly, when present, these mutations were not in recurrent hotspots (**Figure 4.31**). Conversely, AR hotspot mutations occurred predominantly in samples with lower AR copy numbers; and only 1 of 8 hotspot mutations co-occurred with AR amplification (**Figure 4.31**). Patient WA_39 illustrates the divergent pathways to castration resistance, with two of three tumor sites harboring AR amplification, but a third site manifesting a hotspot mutation (**Figure 4.32**). Expression of AR target genes, such *KLK3*, remained high (average 1190.4 FPKM, 95% CI [969.2 1411.6]) in all of the samples.

Finally, mirroring previous analyses of mutations, we assessed the impact of tissue sampling on detecting likely pathogenic CNVs. Within recurrently altered loci we found alterations were detectable from a single-sample for the majority of patients, e.g. deletions of *RB1* (86%), *TP53* (83%), *PTEN* (78%) and amplifications of *MYC* (94%) and *AR* (90%). However, additional samples yielded a further average improvement of detections by 9% and 11% respectively for deletions and gains (**Figure 4.33, 4.34**).

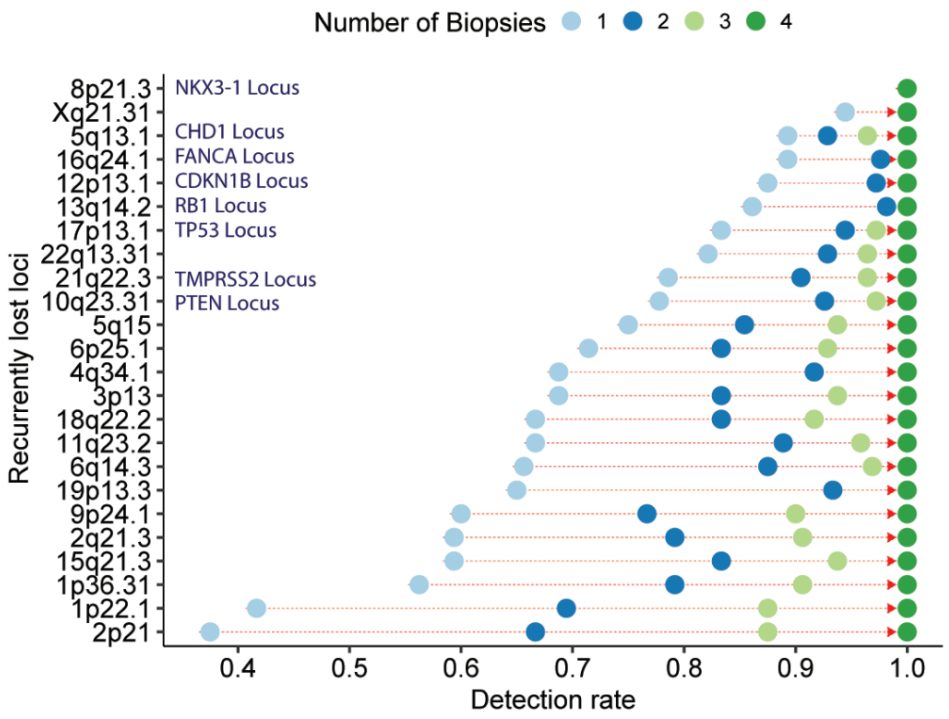


Figure 4.33: Increase in the detection rate of copy number losses as a function of the number of sites sampled. Select loci recurrently altered in PCa are highlighted

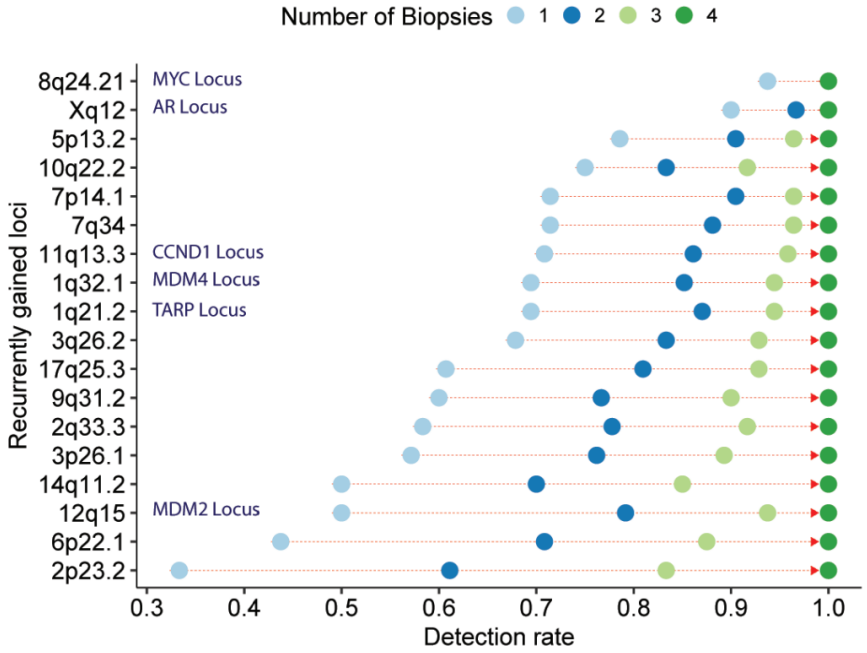


Figure 4.34: Increase in the detection rate of copy number losses as a function of the number of sites sampled. Select loci recurrently altered in PCa are highlighted

Although alterations of AR are highly detectable from a single sample, these alterations are often distinct and likely emerge independently, as previously reported [7,16].

4.2.6 The majority of tumor clones are detected in only a single site

To establish clonal phylogenies and tissue migration patterns we first identified clones by clustering variants according to their CCF. Next, using the identified clones, we inferred phylogenetic trees (**Methods**) for each patient [26] (**Figure 4.35)**.

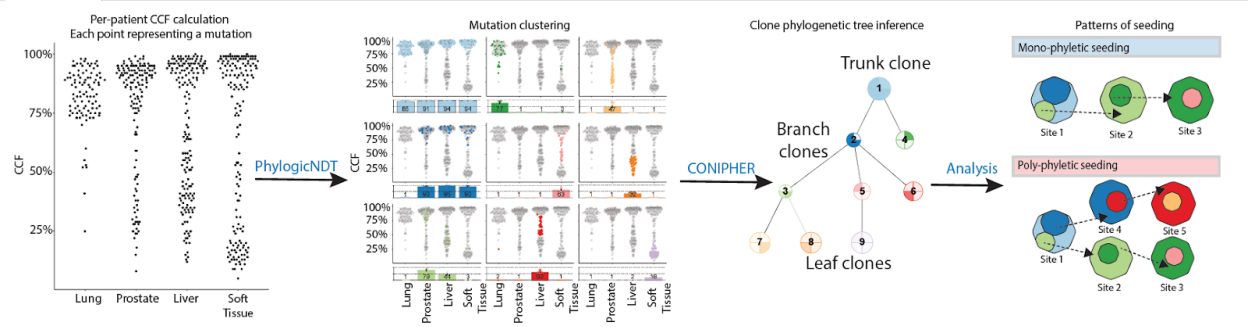


Figure 4.35: Analytical workflow for clonal reconstruction, phylogenetic inference, and analysis of metastatic seeding patterns. Clones are identified by clustering mutations by their CCFs using PhylogicNDT. Clonal phylogenetic relationships (trees) are inferred using CONIPHER. Each clone is represented by a node with slices indicating which sites a clone is present in. Each edge indicates ancestral-descendent relationships between clones. Clones are classified into Trunk, Branch, Leaf based on their position in the tree. Phylogenetic relationships between metastatic clones are used to delineate mono- vs poly-phyletic seeding routes for each patient (tree).

Rare clones which would induce cycles in the tree (e.g. due to clone migration) are omitted at this step based on the parsimony assumptions of the underlying algorithm CONIPHER [48]. We successfully inferred phylogenetic trees for 24 out of the 26 patients. We examined the association of metastatic site, purity, TMB, and wGII on clonal heterogeneity. We noted a significantly lower number of clones in tumors sampled from the lung (P=0.02) and prostate (P=0.04) compared to other tissues (**Figure 4.36**). TMB showed a significant correlation (P<2.2e-16, R=0.78) while wGII and purity showed no definitive trend (**Figure 4.37**).

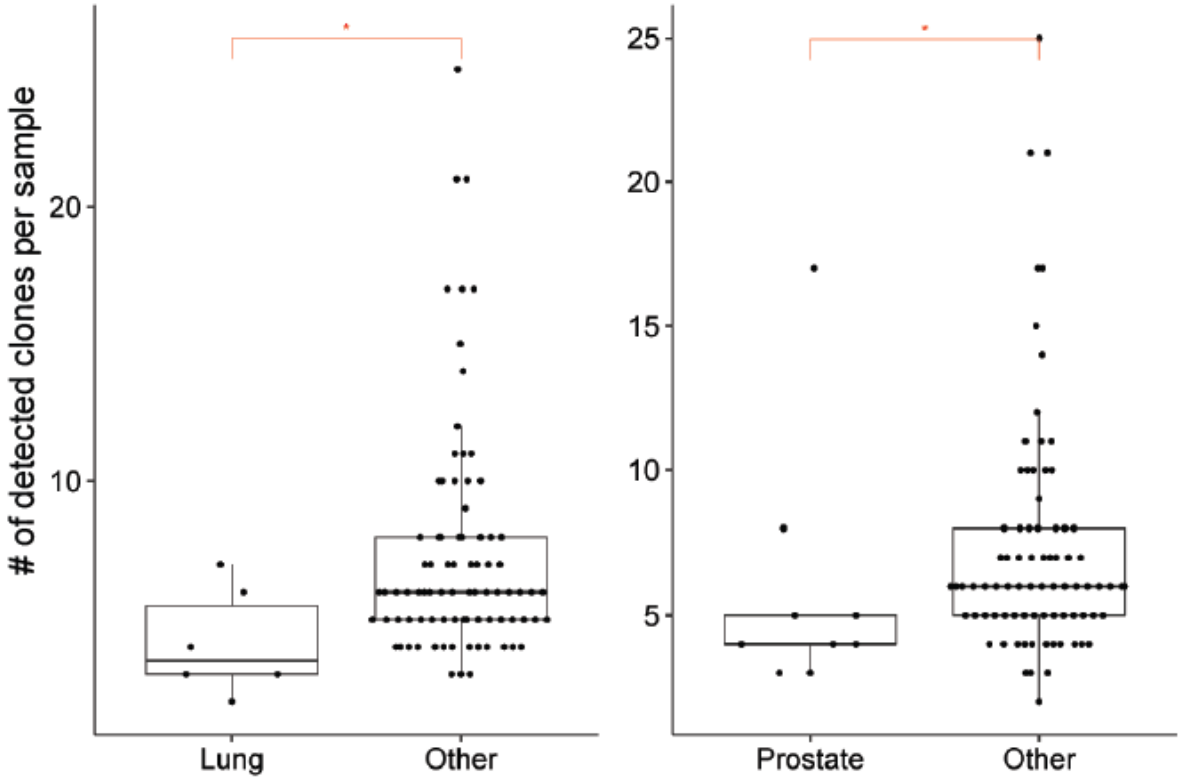


Figure 4.36: Number of clones detected in each sample stratified by tissue site. Significance based on Wilcoxon rank sum test.

This highlights the contribution of ongoing mutagenesis and organ tropism to tumor evolution. These data also suggest that metastatic dissemination to certain tissues may lead to enhanced fitness of dominant tumor clones (i.e., lung metastases), while other tissues may promote continued tumor evolution.

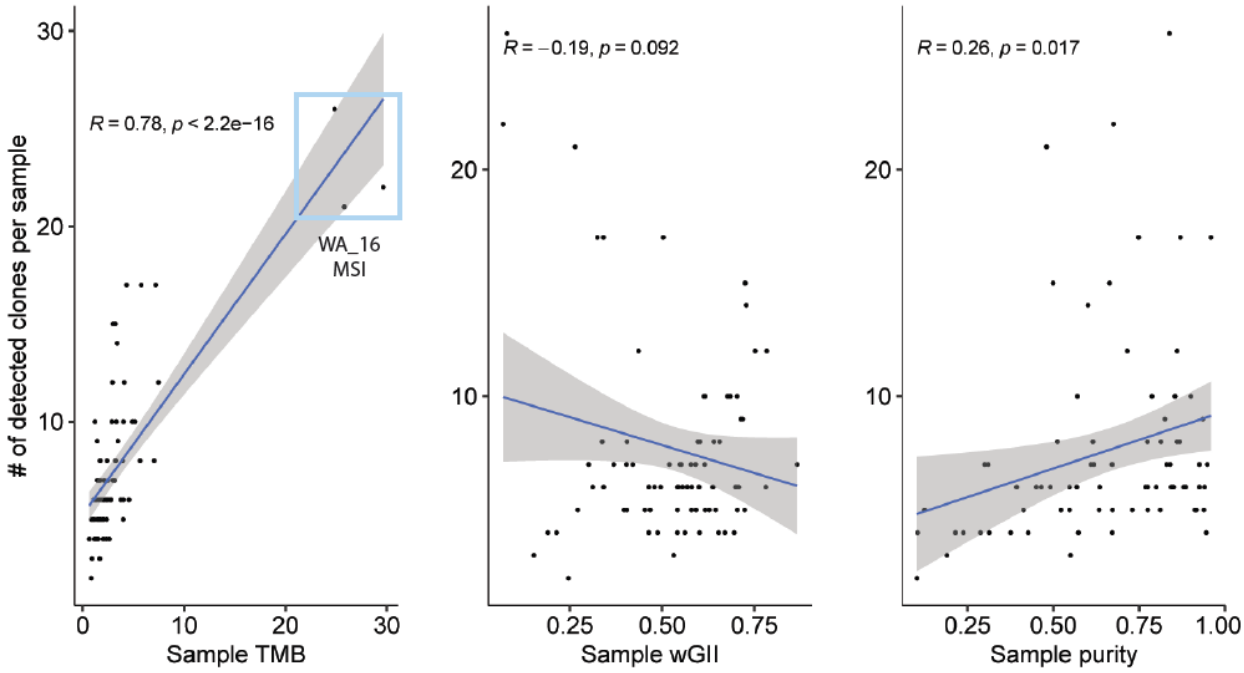


Figure 4.37: Number of clones detected in each sample stratified by TMB, wGII and purity of the samples. Significance based on Pearson correlation test.

Next, we classified clones into three categories based on their position on the phylogenetic tree: Trunk, Branch, and Leaf (**Methods**). Mutations from truncal clones are present and clonal across all metastatic sites and have the highest CCF (Average CCF=0.94, **Figure 4.38**). Branch clones migrate between tissues, being typically clonal at the destination site and either clonal or subclonal at the source sites (Average CCF=0.73, **Figure 4.38**). Leaf clones emerge post-dissemination from the MRCA, are restricted to a single site and primarily harbor subclonal mutations (Average CCF=0.48, **Figure 4.38**).

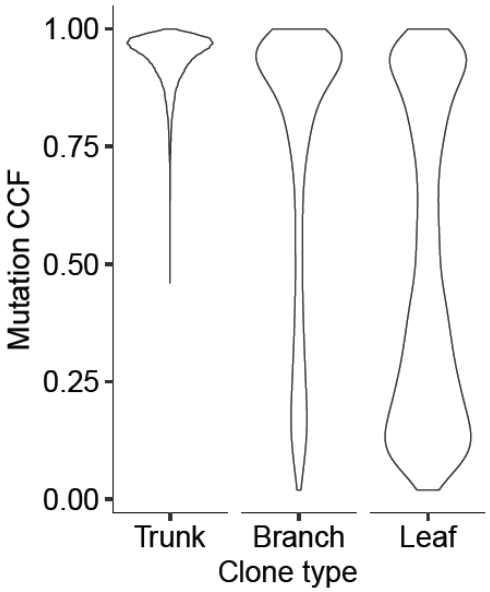


Figure 4.38: Distribution of CCFs (density) within clone types stratified by position in the phylogenetic tree (Trunk, Branch, Leaf).

Across all patients, the majority of clones were Leaf (Mean=61%, 95% CI [68.1% 53.9%]), followed by Branch (Mean=29%, 95% CI [22.2% 35.8%]) and Trunk (Mean=10%, 95% CI [11.1% 8.9%]) (**Figure 4.39**). The relatively small proportion of Trunk clones demonstrates that most clones in mCRPC patients evolved after dissemination from the MRCA. Additionally, the high proportion of Leaf clones in each site indicates that patients have many unique clones per site, leading to considerable heterogeneity across metastatic sites within the same patient.

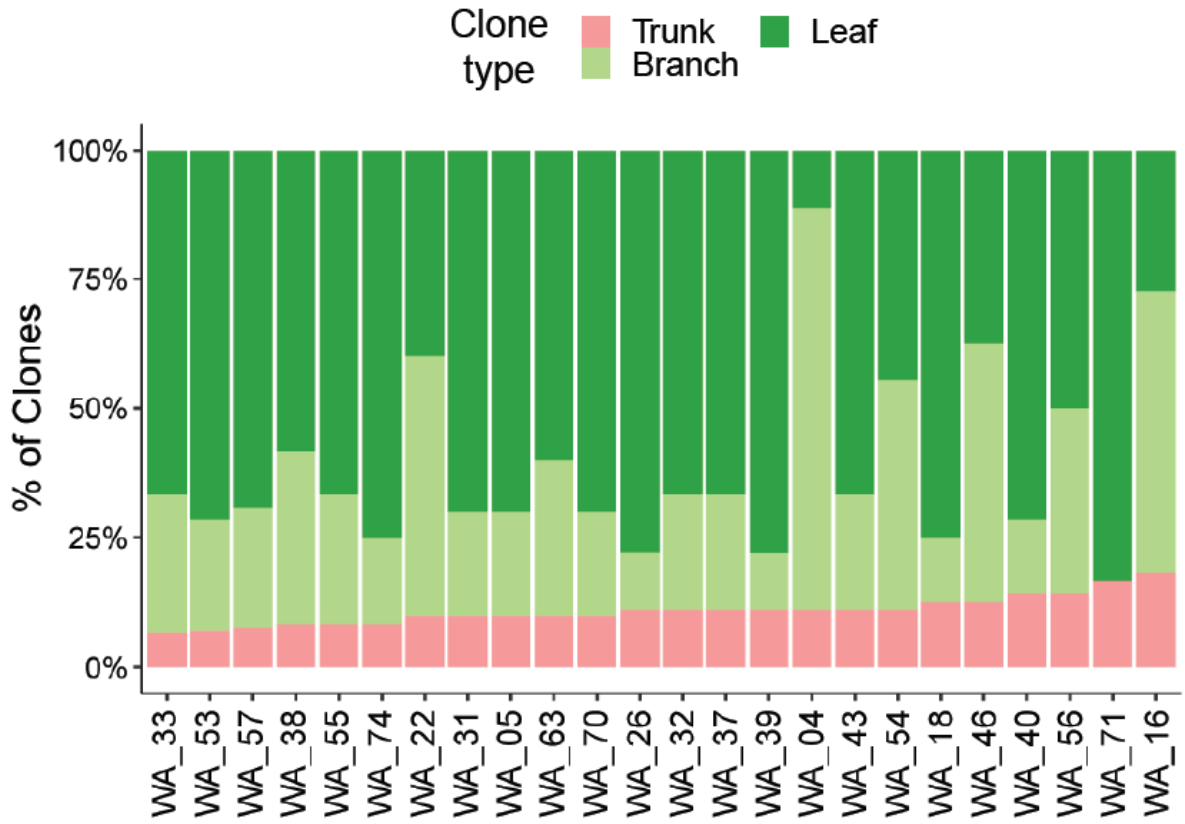


Figure 4.39: Proportion of clone types by patient. This plot excludes the clones not placed in the trees

4.2.7 Position of driver mutations within the phylogenetic tree is highly variable

Next, we focused on clones harboring pathogenic mutations and their position within the phylogenetic trees. Notably, for over 75% of the patients, *TP53* mutations appeared in the truncal clone (**Figure 4.40**).

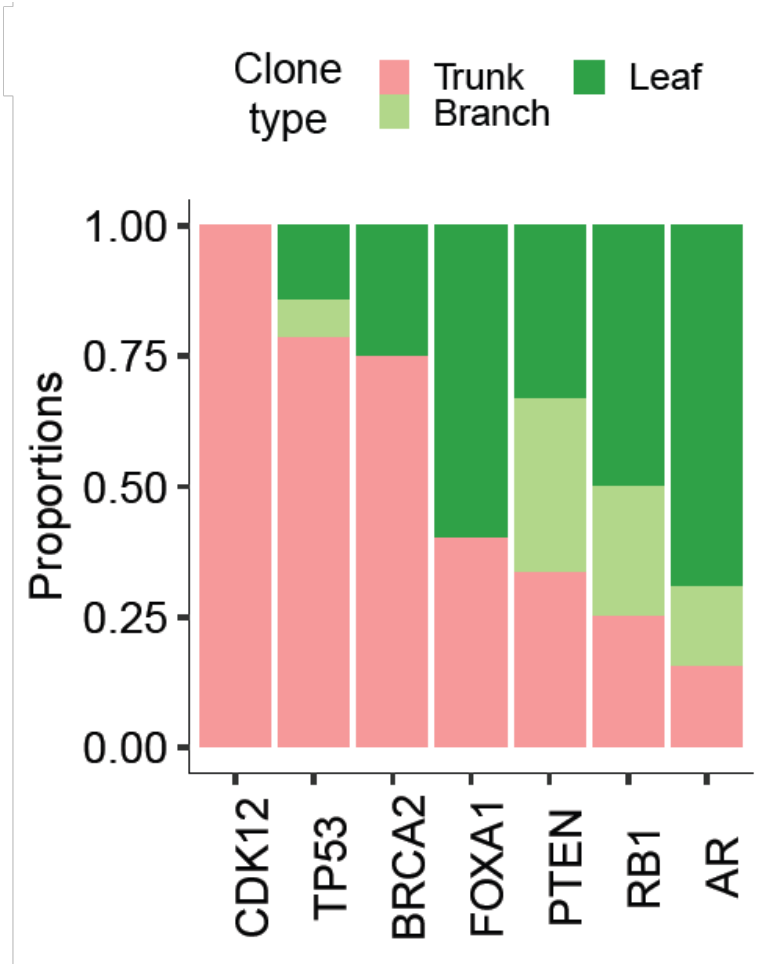


Figure 4.40: Proportion of clone types by recurrently mutated gene. This plot excludes the clones not placed in the trees

This finding underscores that the majority of these mutations occur prior to dissemination of clones from their MRCA, despite their appearance late in PCa progression [34]. Conversely, mutations in *AR* and *RB1* were truncal in 15% and 25% of cases, respectively, suggesting these mutations emerge with therapeutic pressure. *AR* mutations, in particular, were predominantly on the phylogenetic tree's leaves (**Figure 4.40**), indicating they were acquired independently in each site. *FOXA1* mutations exhibited a notable pattern; mirroring our previous data where *FOXA1* mutations showed a bimodal CCF distribution. Here *FOXA1* alterations were either positioned on the trunk or on the leaves of the

phylogenetic tree (**Figure 4.40**). Among the five patients in our cohort with *FOXA1* mutations, three possessed class 1 *FOXA1* mutations [49], with one situated on the tree's trunk and two on its leaves. Similarly, one class 2 *FOXA1* mutation was located on the leaf and another one the trunk of their respective phylogenetic tree.

4.2.8 Single-site sequencing significantly underestimates clonal heterogeneity

We investigate the effect of multi-site sampling on the detection of clonal heterogeneity, specifically assessing how many additional unique clones will be detected as more metastatic sites are sampled (**Methods**). Utilizing the same mathematical model as for mutations, we predicted the total number of clones present in a patient's tumor ecosystem (**Figure 4.41**).

To assess the accuracy of this prediction we compared the predicted and observed number of additional clones per site *i.e.* leaf clones which are only detectable if that site is sequenced. On average 1-2 clones are present in only a single site, with good agreement between the observation and prediction (**Figure 4.42**). Application of the model revealed that when one sample per patient is sequenced, only 34.6% (95% CI [31% 42%]) of all the tumor clones will be detected (**Figure 4.43**).

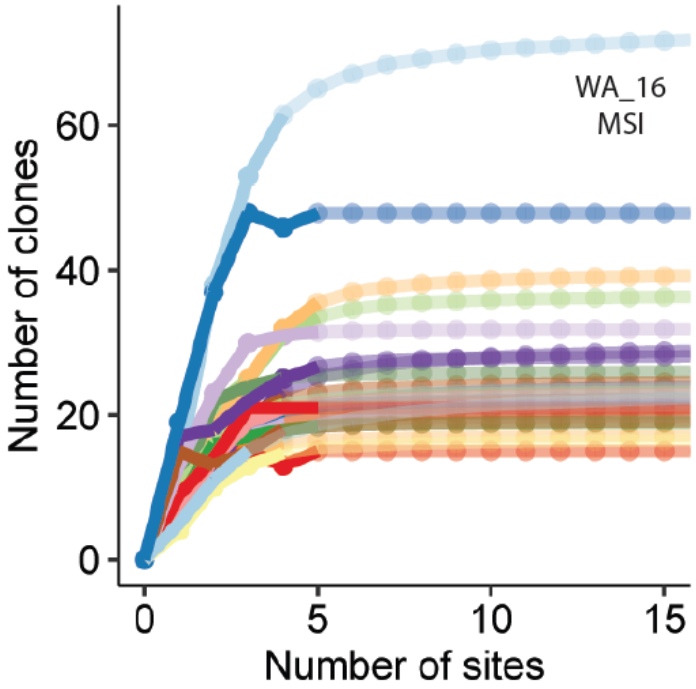


Figure 4.41: Total number of detected clones as a function of the number of sites profiled. Each colored curve corresponds to an individual patient.

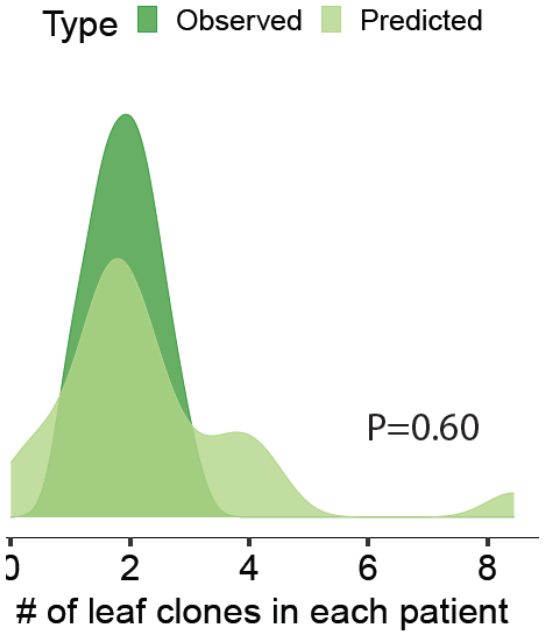


Figure 4.42: Number of patients density (Y-axis) compared to average number of leaf clones they harbor. Light green shows the predicted number and dark green shows the actual number of leaf clones. P-value is based on Wilcoxon rank sum test and indicates that the two distributions are not significantly different.

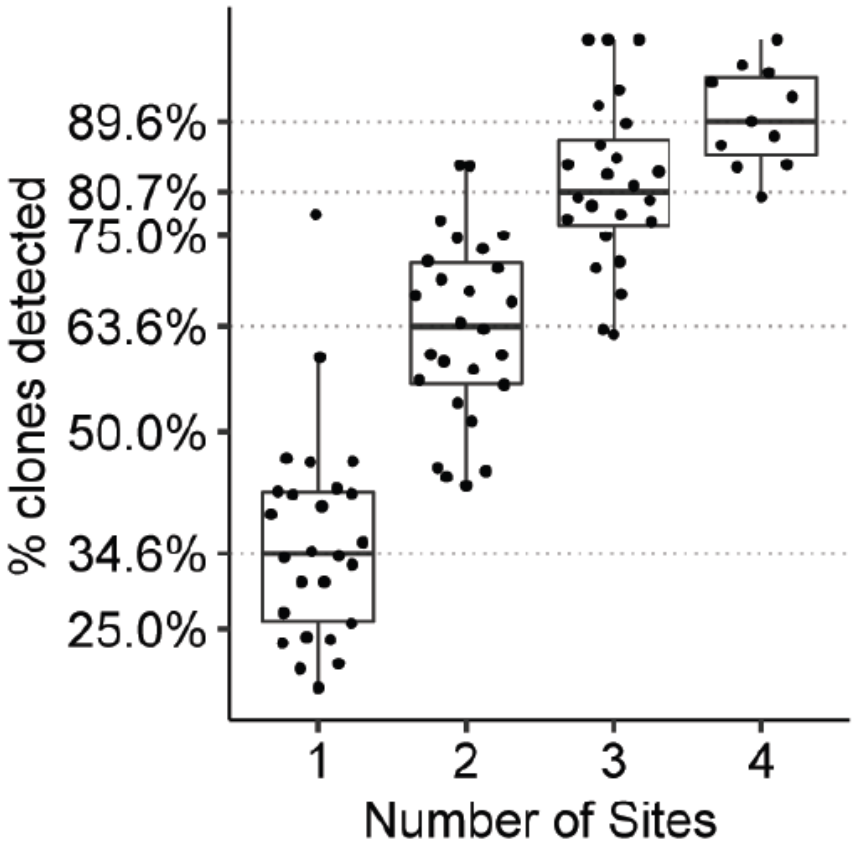


Figure 4.43: Proportion of clones detected as increasing numbers of sites are evaluated.

This number reaches 89.6% (95% CI [77% 86%]) when 4 samples are profiled. The determination of whether a clone is truncal is heavily influenced by sampling from multiple and diverse anatomical sites. Similar to the pattern previously described for mutations, we observed that based on the analysis of a single site 40% (95% CI [36% 44%]) of clones would be identified as truncal, compared to only 13% (95% CI [12% 14%]) and 6% (95% CI [4% 8%]) when two or more sites are examined, respectively (**Figure 4.44**). This highlights that dominant clones in one site are often subclonal (or absent) in others.

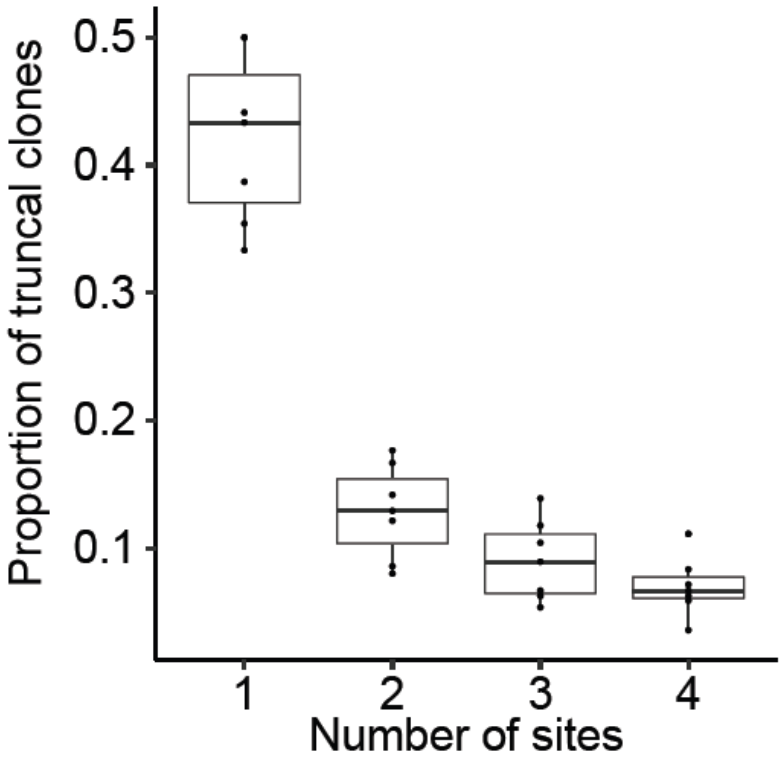


Figure 4.44: Proportion of clones detected as truncal (Y-axis) per different number of evaluated sites (X-axis)

4.2.9 High prevalence of polyclonal and polyphyletic metastatic seeding

Using the phylogenetic trees, we deduced the evolutionary paths of metastatic seeding, specifically the child-parent relationships of clones seeding or migrating between sites [50] (**Methods**). Accordingly, we categorized the phylogenetic trees into three groups. The first group comprised patients in which seeding clones had a child-parent relationship (*i.e*. monophyletic seeding) [50] (**Figure 4.45**). Monophyletic seeding was observed for the majority 71% (17/24) of patients, indicating that the capacity for metastasis had evolved once.

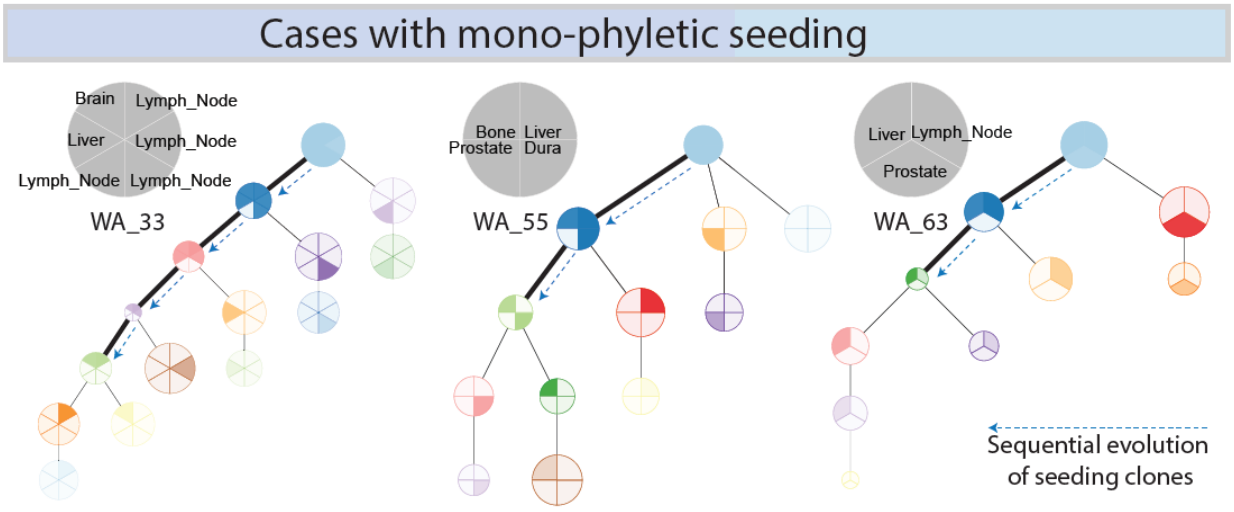


Figure 4.45: Classification of clonal phylogenetic trees based on ancestral relationships between seeding clones. Bolded edges indicate that the descendant clone seeded a new site

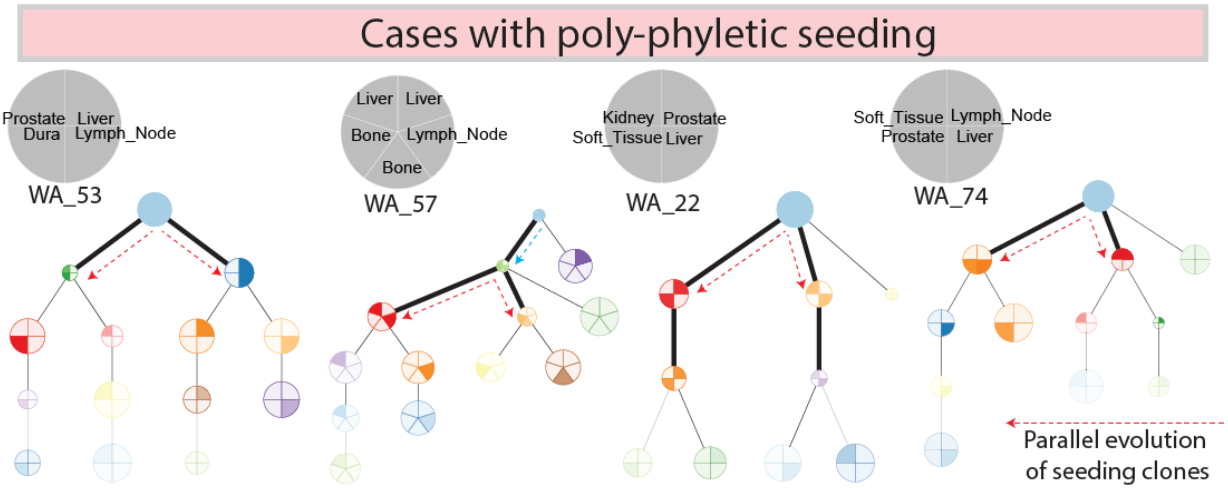


Figure 4.46:Classification of clonal phylogenetic trees based on ancestral relationships between seeding clones. Bolded edges indicate that the descendant clone seeded a new site

The second group, accounting for 17% (4/24) patients, had seeding clones located on parallel branches of the tree (**Figure 4.46**). This pattern of polyphyletic seeding, initially described in mCRPC by Gundem et al. [7], results in high genetic dissimilarity across

metastatic sites e.g. some polyphyletic clones may be more similar to clones in the primary tumor rather than clones in other metastatic sites [51].

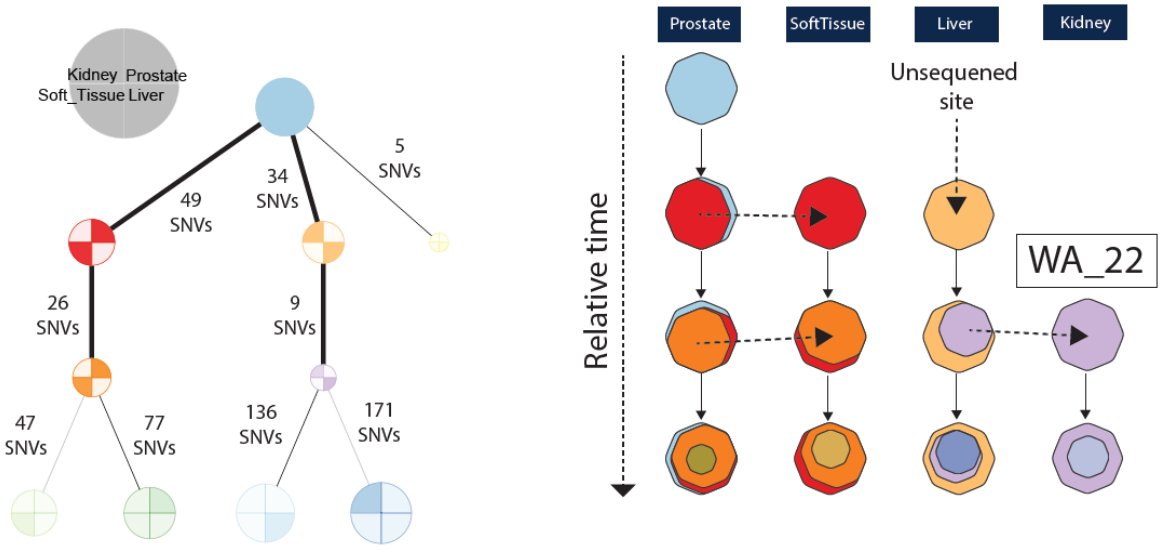


Figure 4.47: Juxtaposition of clonal phylogenetic trees (left) and clonal migration diagrams (right). Clonal migration diagram shows a parsimonious reconstruction of evolutionary (top to bottom) and migratory (left to right) relationships. Arrows indicate evolution (solid) or migration (dashed). Each node represents a site with sections colored by clonal composition. Clone colors are matched between trees and diagrams.

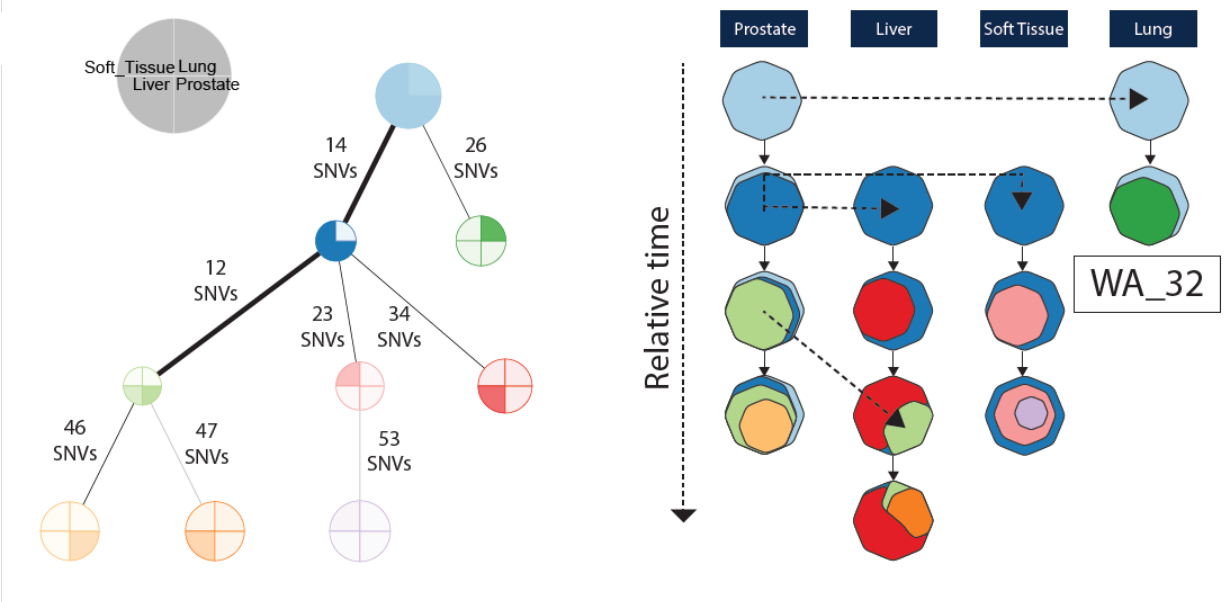


Figure 4.48: Juxtaposition of clonal phylogenetic trees (left) and clonal migration diagrams (right). clonal migration diagram shows a parsimonious reconstruction of evolutionary (top to bottom) and migratory (left to right)

relationships. Arrows indicate evolution (solid) or migration (dashed). Each node represents a site with sections colored by clonal composition. Clone colors are matched between trees and diagrams.

In the final group of three patients, a single clone seeded all the metastases. However, it is important to note that in this group, only two sites were sequenced for two of the patients.

Additionally, in 2 patients (WA_22 and WA_32), we observed a pattern of recurrent seeding and migration from the prostate tumor tissue to other metastatic sites. In patient WA_22, we observed initial seeding from the prostate to a soft tissue lesion (**Figure 4.47** left). The data are consistent with a parsimonious model [26,52] wherein the seeding clone continued to accrue an additional 26 mutations in the prostate evolving into a distinct clone. This clone migrated to the soft tissue site a second time (**Figure 4.47** right) and underwent further parallel evolution in both prostate and soft tissue, acquiring 77 and 47 distinct mutations, respectively. A similar pattern was observed in patient WA_32 (**Figure 4.48** left), where clones in the prostate repeatedly migrated to the liver. Following the initial metastasis to the liver, the liver-resident clone acquired 34 new mutations, while the progenitor clone in the prostate acquired 12 new mutations, forming a new clone. The newly evolved prostate clone migrated again to the liver, where it acquired 47 additional mutations, resulting in 3 clones (two in the liver and one in the prostate) which share a common ancestor but evolved separately (**Figure 4.48** right). Among the nine patients with available prostate tumor samples, recurrent seeding of metastatic sites from the prostate was observed in a minority of patients (2 / 9) which suggests an active role for prostate tumors in seeding of metastatic sites over the course of the disease.

4.2.10 Early metastatic dissemination in a case of neuroendocrine mCRPC

We identified an index case, WA_24, where no mutations were shared across all sites (**Figure 4.49**), raising questions regarding their phylogenetic relationships.

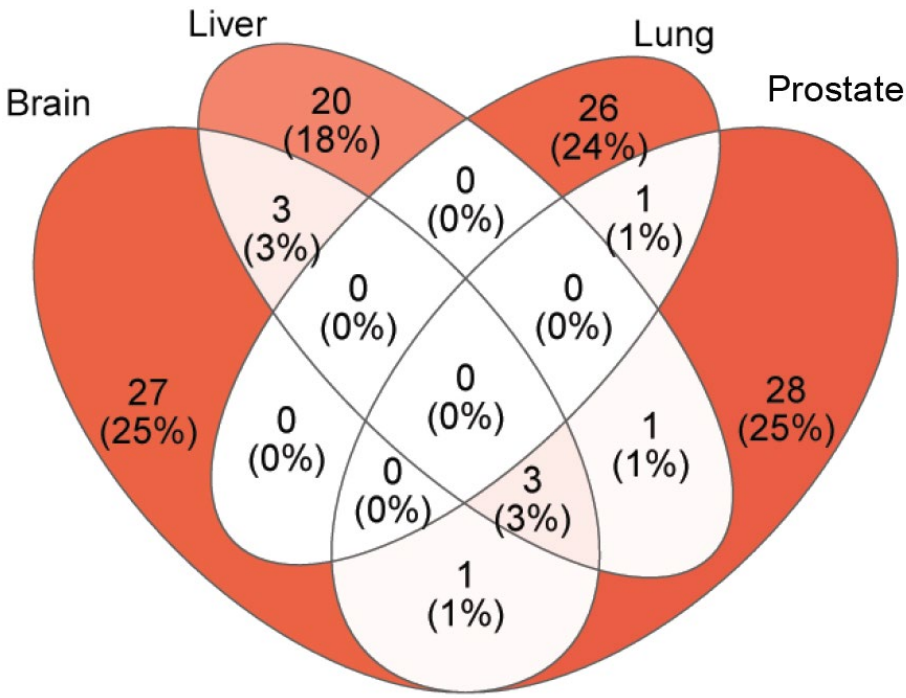


Figure 4.49: Overlap of mutations detected in each sequenced site of WA_24. Red color gradient indicates the fraction of all mutations in each intersection.

Among all non-synonymous mutations 92% (101/110) were subclonal and present in only a single site (*i.e.* subclonal-private, **Figure 4.50**), implying their emergence after seeding. The liver, brain, and prostate sites had mutations in common, but the lung site was mutationally distinct. Consistently, comparison of CNV profiles (**Methods**), also revealed higher similarity between the liver, brain, and prostate sites compared to the lung (**Figure 4.51**).

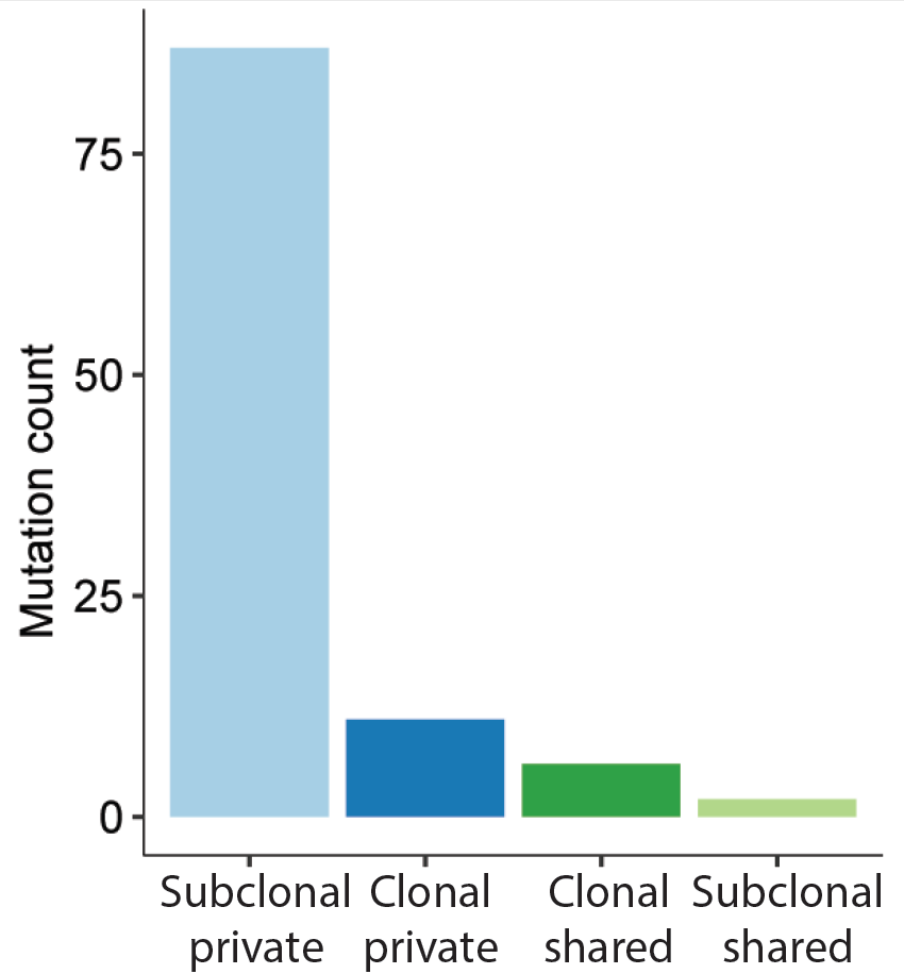


Figure 4.50: Tally of mutations classified according to clonality and distribution

Unlike mutations [53,54], copy number events can revert to the neutral state or diverge from the common ancestor [55], rendering them less reliable for inferring tumor phylogenies [56]. Conversely, LOH results in irreversible genetic loss [57] that can be used to establish phylogenetic relationships [16,56]. We noted extensive overlap of regions with LOH across the four tumors, including LOH of the p-arm of chromosome 12. Although the lung lesion displayed a somewhat divergent LOH pattern (**Figure 4.51**), the overlap of LOH regions was significant (*e.g.* 71% LOH overlap between lung and other sites, **Figure 4.52**).

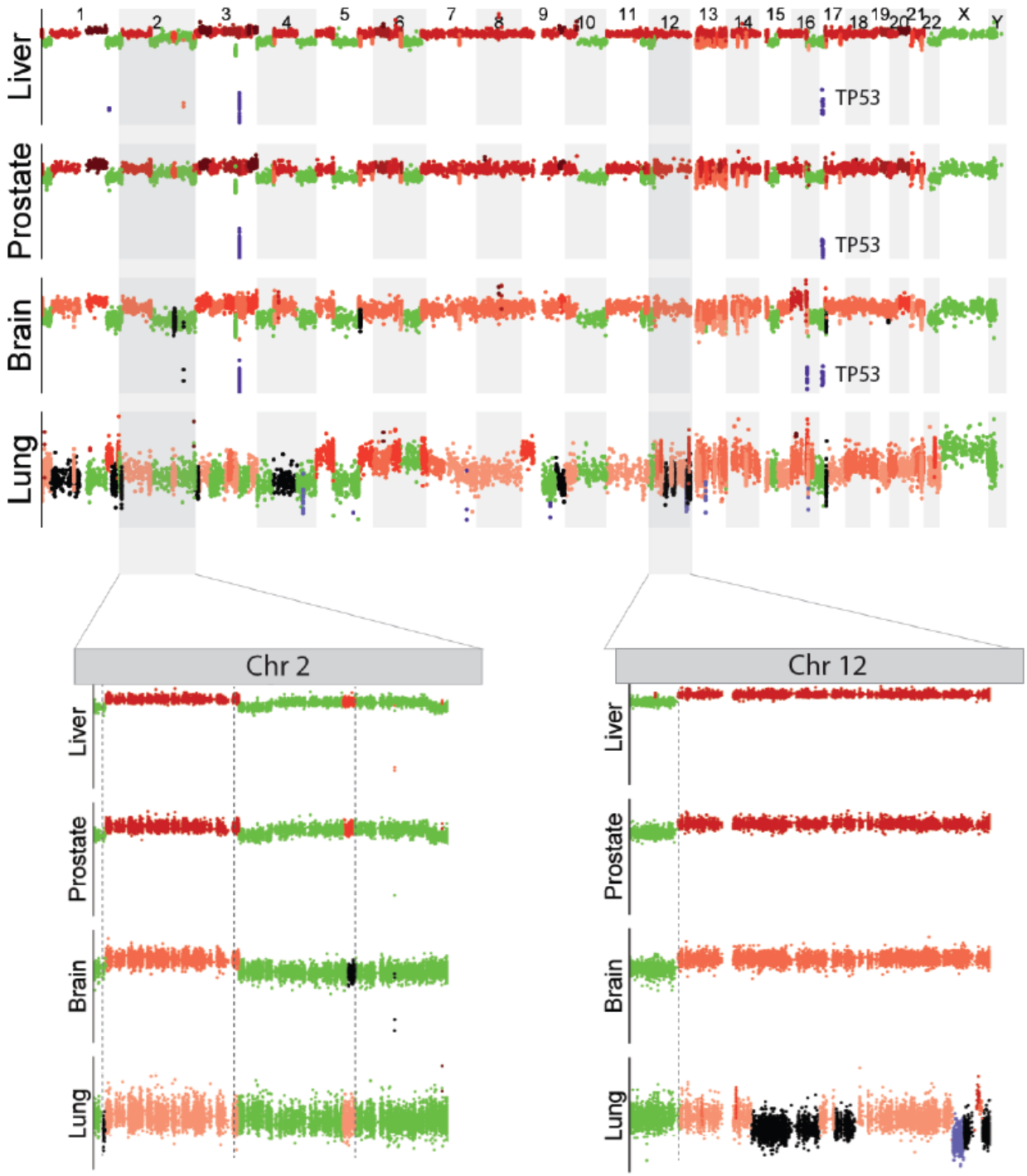


Figure 4.51: Genome-wide copy-number profiles of normalized log2(tumor / normal) coverage ratios across samples from WA_24 (black - diploid, red - gain, blue - loss, green - LOH). Two representative regions are selected to illustrate LOH segments shared across all tumor samples.

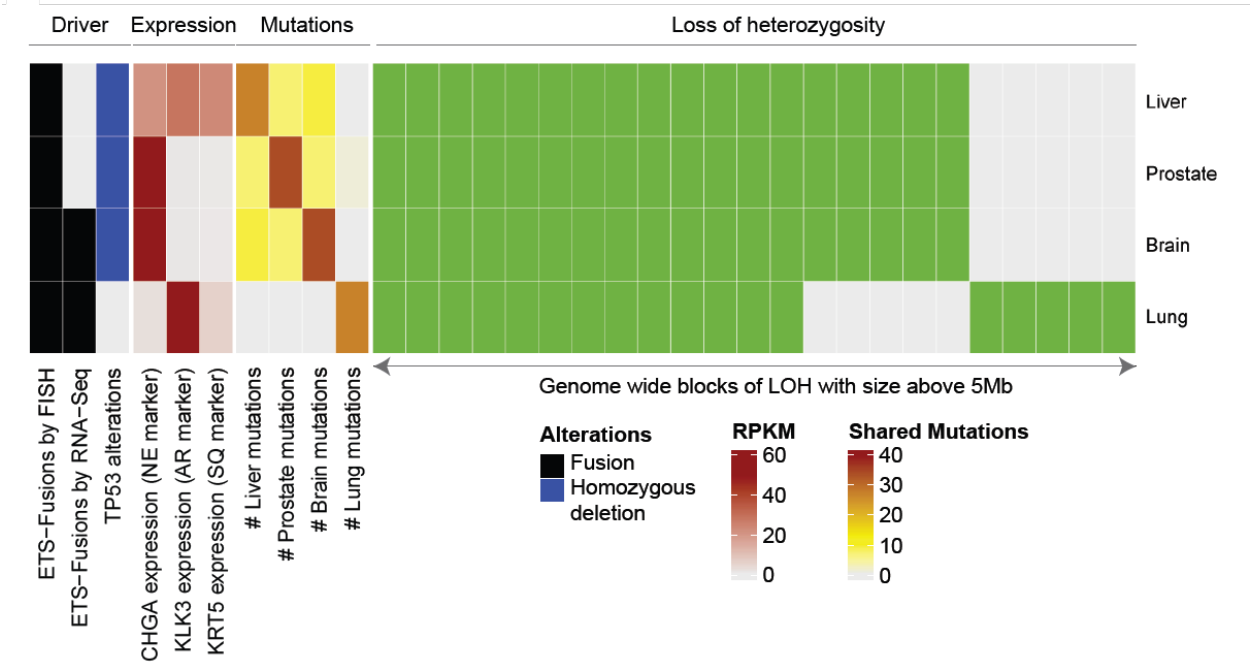


Figure 4.52: Integrated summary of genetic drivers, gene expression, shared mutations, and notable LOH regions in WA_24 samples.

To determine whether this could have occurred by chance we employed a permutation test [58] (**Methods**). The observed overlap significantly exceeded all random draws (P<1e-6, **Figure 4.53**), rejecting the null of coincidence.

Next, to better understand the oncogenesis of this tumor, we sought to identify genetic drivers present in all sites. Since no mutations were truncal, we examined the RNA-Seq for gene fusions (**Methods**) and identified an identical *TMPRSS2-ERG* alteration in only 2 out of 4 samples. However, Fluorescence in Situ Hybridization (FISH) experiments [59], indicated the presence of this fusion across all sites (**Figure 4.52**).

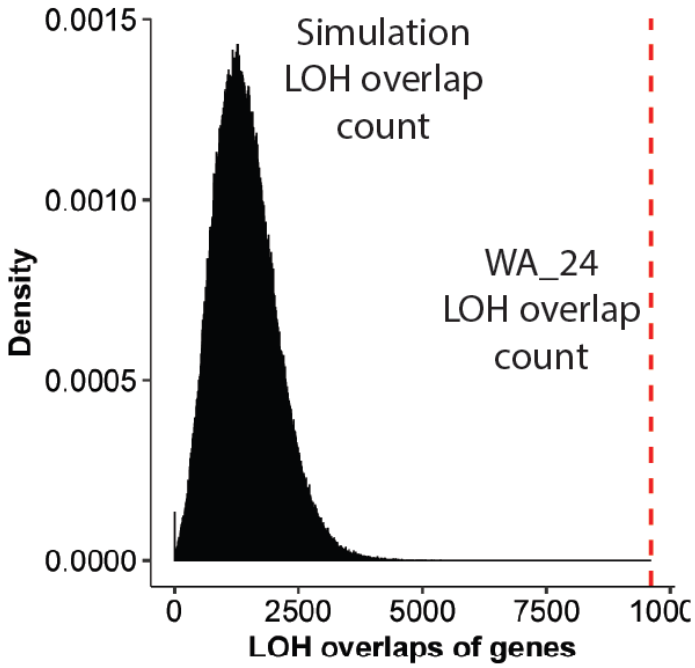


Figure 4.53: Results of permutation testing to compare the observed number of genes within regions of shared LOH in WA_24 (red dashed lines) with a distribution based on a null model (histogram).

By investigating this discrepancy, we found that 2 out of 4 samples in this patient had strong expression of neuroendocrine (NE) markers. Two of these three samples had minimal expression of the AR signaling marker KLK3, most consistent with neuroendocrine prostate cancer (NEPC) [60]. Loss of AR signaling explains the lack of AR-dependent *TMPRSS2-ERG* expression [61]. However, one of the three samples retained KLK3 expression—a pattern most consistent with amphicrine prostate cancer that harbors a hybrid cell state that is AR+NE+ [60]. (**Figure 4.54, 4.55**).

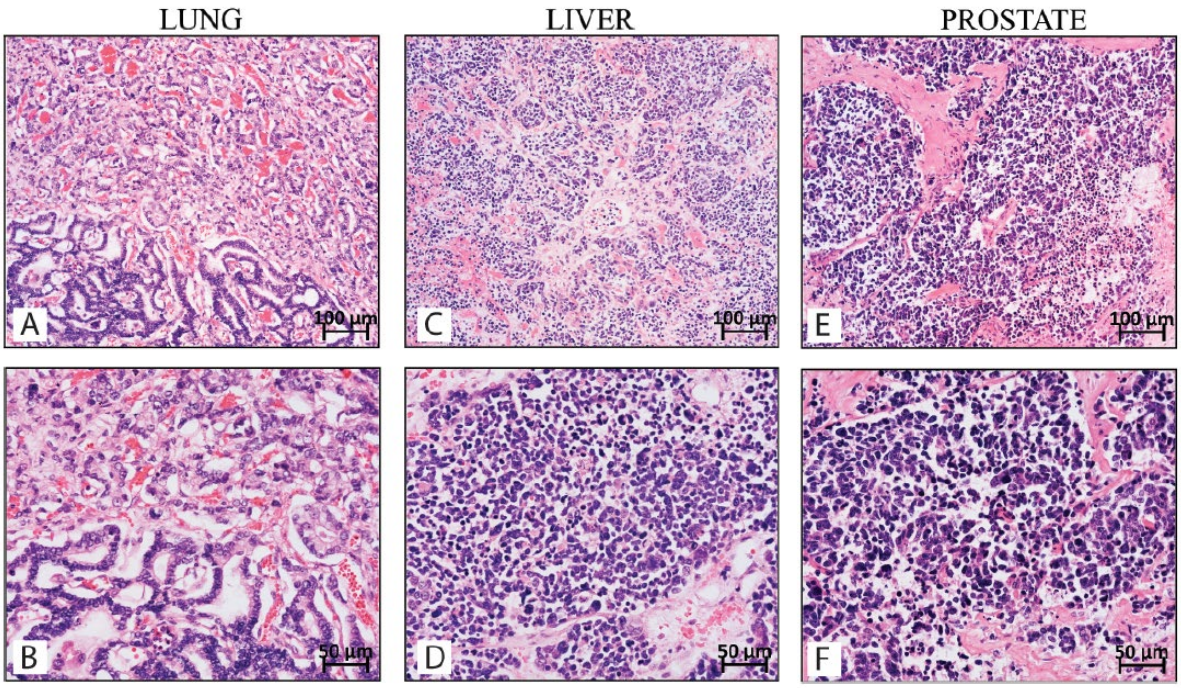


Figure 4.54: Morphological heterogeneity and spectrum in WA_24 at different spatial sites with available FFPE slides. A-B: Mixed classical acinar and organoid pattern in the lung. C-D: Small cell neuroendocrine carcinoma-like pattern in the liver. E-F: Small cell neuroendocrine carcinoma-like pattern in the prostate.

This observation explained the lack of *TMPRSS2-ERG* detection in RNA-Seq, as NEPC tumors downregulate TMPRSS2 expression [61]. Hence, ETS fusions might not always be detectable by RNA-Seq. Interestingly, the most genetically distinct lung site from this patient also showed the most strongly *AR*-positive gene expression pattern (**Figure 4.54, 4.55**). The three *AR* expression- and *AR* signaling-negative sites also harbored a TP53 homozygous deletion, an event often associated with progression to NEPC disease [62].

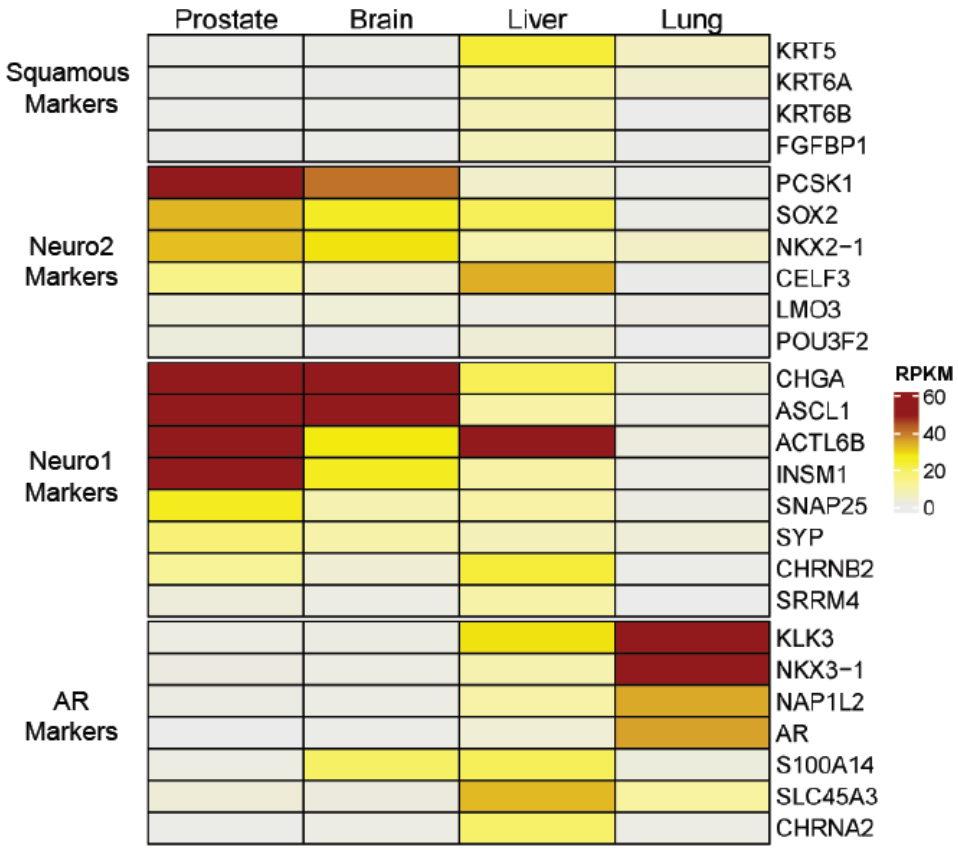


Figure 4.55: Expression of marker genes for molecular subtypes mCRPC as defined by Labrecque et al.).

Finally, to infer phylogenetic relationships between the tumors we integrated the irreversible genetic alterations (mutations, LOH events, ERG fusion, TP53 deletion) alongside the assessment of histological markers (**Figure 4.56**). The integrative phylogenetic model (**Figure 4.56**) suggests that the oncogenic process commenced with a cascade of structural alterations including the driver *TMPRSS2-ERG* fusion alongside multiple chromosomal gains and losses. The truncal *AR*-positive clone disseminated immediately to the lung prior to accumulating additional exonic mutations. Within the prostate the truncal clone acquired a *TP53* homozygous deletion, alongside additional LOH events and 3

passenger mutations, and metastasized to the liver exhibiting an amphicrine phenotype, and to the brain as NEPC. This model highlights that in rare instances of mCRPC evolution, acquisition of metastatic potential and lineage plasticity can occur very rapidly.

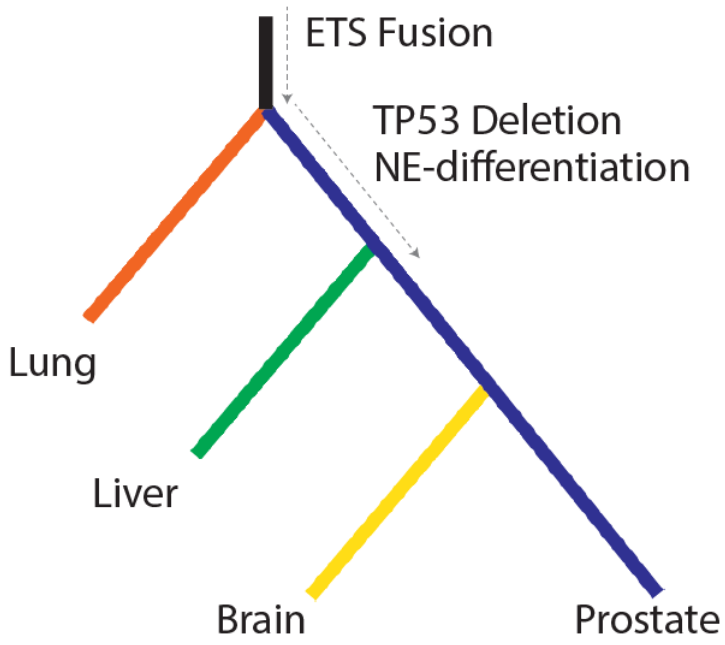


Figure 4.56: Expression of marker genes for molecular subtypes mCRPC as defined by Labrecque et al.).

4.3 Discussion

The genetic heterogeneity of primary PCa has been extensively characterized, with studies reporting multifocality and variable mutation patterns between different tumor foci [4,10,11,63]. Other studies have utilized cohorts of primary and local mHSPC to shed light on mechanisms and drivers of metastasis [64,65]. However, the heterogeneity of mCRPC remains relatively poorly understood.This is important because many providers begin to profile prostate tumors at the time of CRPC in order to identify targeted therapy options. Previous studies have uncovered seeding between metastatic sites and clonal evolution patterns [7], dissected the complexity of alterations within the AR locus [16], and shed light

on mechanisms and directions of metastasis in mCRPC [15]. However, these studies were limited by shallow sequencing depth (<=100X) [7,16], small cohort sizes (<=10 patients) [7,15,16] and focused scope [16].

To address these limitations, we sequenced a set of 93 tumors captured from 26 mCRPC patients in an effort to characterize the extent of genetic divergence among patient's metastatic tumors across mutational, transcriptional, and histological domains. High-depth sequencing facilitated the identification of copy-number alterations and subclonal mutations, crucial for identifying minor but genetically distinct cell populations. Mathematical modeling and phylogenetic reconstruction allowed us to characterize each patient's disease as a clonal ecosystem and evolutionary process, respectively. Together, these analyses quantify the extent of intra-patient tumor heterogeneity in mCRPC and its implications for diagnostic and therapeutic strategies.

Fundamentally, we identified striking levels of genetic heterogeneity for all of the patients. This heterogeneity is driven by both mutations and copy-number alterations, and is manifested at the level of individual mutations and tumor clones. For example, we found that on average 69% of mutations in each patient are present in only one site. A similar trend can be observed at the clone level, where on average 61% of all clones are leafs *i.e.* are present in only one site. We also demonstrate that profiling mutations based on a single sample is prone to erroneously classifying mutations as clonal. This phenomenon, known as "illusion of clonality" [14,25], is both biologically and clinically [14,66–68] significant as it is caused by metastasizing or rapidly expanding clones [67] and is stipulated to be a mechanism for targeted therapy resistance [8,66]. Illustratively, on average only 13% of mutations are truncal, while only on average 32% of mutations identified as truncal based on a single

sample are truncal in all the other sites. Identifying truncal clones and their mutations is crucial for guiding treatment as these mutations are the most suitable targets for durable response in patients as they are present in all cancer cells [69,70]. These observations highlight that by evaluating only one sample per patient we are significantly underpowered to characterize the tumor's clonal architecture. Similar, albeit not as extreme, trends have been reported in other cancers including non-small cell lung cancer [50] and esophageal squamous cell carcinoma [71].

Underpinning this profound intra-patient heterogeneity are characteristic patterns of clonal evolution. We estimate that the typical mCRPC tumor ecosystem consists of an average 7 (95% CI [6 8]) clones per metastatic site and 13.5 (95% CI [11 16]) clones in total. We note a significant correlation (R=0.78) between TMB and clonal heterogeneity, suggesting that persistent mutagenesis is driving the emergence of new clones. Seeding of metastases is typically polyclonal and often polyphyletic (17%) which manifests in diminished genetic similarity of metastatic lesions and signifies independent acquisition of metastatic potentials [50]. Notably, only 7% of cancer clones are truncal, illustrating that metastatic spread occurs early and is followed by parallel evolution within metastatic lesions; each metastatic site typically harboring two unique clones, which is an indicator of ongoing clonal sweeps [31,32].

To clarify the implications of these results for mCRPC genetic testing we examined the heterogeneity of driver alterations, and the impact of single-site testing on diagnostic yields. Overall, we found that alterations that initiate prostate cancer (*e.g. ERG*, *CDK12*, *BRCA2*, *NKX3-1*) or significantly contribute to cancer aggressiveness (*PTEN*, *TP53*), were largely truncal and benefited least from sampling multiple sites. Similarly, estimates of TMB

and CIN were consistent across multiple sites. As previously described [7,16], we found alterations of *AR* to often be distinct across sites, but, somewhat paradoxically, the presence of an *AR* alteration in one site was strongly correlated with the existence of equivalent alterations at other sites, indicative of convergent evolution. Conversely, detection rates of alterations in *RB1*, *FOXA1*, and the Wnt pathway, tended to improve with multi-site sequencing. Both *RB1* and Wnt have been reported to contribute to mCRPC aggressiveness [72–74]. Likewise, facets of aneuploidy, including genome doubling and copy-number alterations were also highly heterogeneous, indicative of ongoing chromosomal instability. For example, for 43% of patients baseline ploidy was variable across sites, suggesting that WGD had occurred after metastasis. Tumor purity has been reported to be associated with both intrinsic tumor biology (*e.g.* specific genomic alterations affecting tumor microenvironment [75]), organ tropism in metastasis *e.g.* [76], as well as technical challenges of tissue sampling and procurement [77,78]. We found that tumor purity is not merely a technical factor, but is biologically informative. Its impact on genomic readouts, such as variant allele frequencies or gene expression profiles, requires therefore especially careful modeling.

Finally, we expanded upon previous observations of multiclonal, metastatic-to-metastatic seeding in mCRPC [7] and identified biologically and clinically relevant examples of tumor phylogenies and tissue migration patterns. We noted a case (WA_24) of intrapatient, intralesional heterogeneity with early metastatic spread and a spectrum of neuroendocrine differentiation. All sites of WA_24 were clonally related, but did not share truncal mutations, indicating that the cancer, driven by a truncal *TMPRSS2-ERG* fusion, clonally diverged and disseminated prior to the acquisition of exonic mutations. This early

phylogenetic branching is further reflected in profound transcriptional and histological heterogeneity. Woodcock et al. [64] have reported a similar case (A34) wherein two early branching clones independently acquired metastatic potential. However, in that patient, the tumor had acquired multiple mutations before early dissemination. We also identified examples (WA_22, WA_32) of recurrent seeding of distal metastases from the prostate. Ongoing clonal evolution within the unresected prostate resulted in multiple waves of dissemination and re-seeding, suggesting an active role for prostate tumors in fueling metastatic disease. This finding along with studies showing that prostate radiotherapy in subsets of men with metastatic prostate cancer improves failure-free and overall survival [79,80] highlights the importance of considering primary therapy in patients with metastatic PCa. Together, these observations support the existence of multiple, but always related, clones within the primary PCa with metastatic potential [4].

The current study has several limitations. While using WES data, as opposed to whole genome sequencing, allows for higher sequencing depths and increases the sensitivity of detecting variants with low CCF, it detects fewer mutations which may result in the failure of detecting some clones. Increasing the number of sequenced samples per patient would further increase the ability to characterize patterns of metastatic dissemination and possibly improve analyses. Some of our results are based on mathematical modeling and use of inferential algorithms. While we cross-reference these results with descriptive data, inferences of processes that cannot be directly observed (*e.g.* tumor evolution), are always associated with uncertainty.

In summary, our study highlights the profound genetic heterogeneity of lethal PCa, and examines the critical contribution of early dissemination and independent evolution at

metastatic sites to its complex clonal architecture. We demonstrate that, except for a few truncal drivers, sampling multiple-sites increases diagnostic yields both at the patient and cohort levels. Our results echo prior recommendations against over-reliance on single-site biopsies for precision oncology in PCa [4]. The emergence of cell-free DNA profiling approaches from blood may augment single site metastatic biopsies to provide further clarity about tumor clonality within a given patient.

4.4 Methods

4.4.1 Sample collection

All tumor and matched normal specimens were obtained from the Michigan Legacy Tissue Program (MLTP). The rapid autopsy protocol has been described previously [59,81], and was conducted with the approval of the University of Michigan IRB and patient consent.

4.4.2 Preparation of whole exome and RNA sequencing libraries

Exome sequencing and exome capture RNA-seq was performed using standard protocols in our Clinical Laboratory Improvement Amendments-compliant (CLIA) sequencing laboratory [22]. Briefly, tumor's DNA and total RNA were purified from the same sample using the AllPrep DNA/RNA/miRNA kit (Qiagen). Matched normal genomics DNA from normal tissue were isolated using the DNeasy Blood & Tissue Kit (Qiagen). Exome-capture transcriptome platform [17] was used for RNA-seq. Exome libraries of matched pairs of tumor and normal DNA were prepared using the Agilent SureSelect Human All Exon v4 platform (Agilent) as previously described [82]. All the samples were sequenced on an HiSeq 2500 or Novaseq instrument (Illumina) in paired-end 150bp mode. The primary

base call files were converted into FASTQ sequence files using the bcl2fastq version 1.8.4 converter tool.

4.4.3 Comprehensive genomics analysis pipeline

All genomic data processing and analysis has been performed using the TPO workflow (https://github.com/mctp/tpo), which implements standardized pipelines for the analysis of DNA and RNA sequencing data, broadly following community best practices [83]. Relevant algorithmic details are explained below, a general overview of its functions has been provided in prior publications [84–86].

4.4.4 Read processing and data quality control

BBMap's bbduk [87] was used to perform trimming of DNA sequencing paired-end reads. Next, the data was aligned to the GRCH38 reference using BWA-mem [88]. Sentieon sort tool [89] was also utilized to sort the reads in the BAM files. Further quality control measures were extracted from the data according to GATK best-practices.

4.4.5 Somatic Variant Calling

For somatic variants, all tumor samples were matched with a normal tissue. TNscope [89] , an improved somatic caller based on GATK Mutect2 [90] was used to call the somatic variants and calculate variant allele frequency (VAF) according to its probabilistic model, taking realignment and mapping ambiguity into account. The following settings were used in TNScope:

--max_fisher_pv_active 0.05 --min_tumor_allele_frac 0.0075 --min_init_tumor_lod 2.5 --assemble_mode 4 --trim_soft_clip --normal_contamination_frac 0.25 --prune_factor 4

This setting was particularly selected to allow for variant calls with presented evidence in normal tissue as well as tumor tissue. Such an approach allowed us to later account for possible Tumor-In-Normal (TIN) contaminations (see below). The called variants were annotated using VEP [91] and vcfAnno [92]. All detected variants were further filtered (to retain somatic variants) according to a rule based pipeline. Filtering pipeline were filtering mutations based on sequencing evidence (variant allele frequency, coverage, mutation likelihood TLOD and NLOD, strand bias, allele depth, multi-allelic variants), overlap in problematic regions including regions with low-mappability and repetitive sequence, and homopolymer repeats. The resulting set of somatic variant calls included both on and off target mutations.

4.4.6 Somatic Variant Calling Quality Control

During manual review of variant calls for each sample, we identified 71 cases where a minority of likely somatic variants were misclassified as germline or noise due to possible tumor contamination in matched normal samples. To rescue these variant calls we used DeTIN [20], a tool designed to detect and quantify tumor-in-normal (TIN) contamination and reclassify variants where read evidence in the normal is consistent with contamination. We used DeTiN with the default settings of:

--mutation_prior=0.05, --aSCNA_threshold=0.1, --TiN_prior=0.5, --resolution=101, --ascna_probe_number_filter=200, --ascna_SNP_number_filter=20, --coverage_threshold=20, --SSNV_af_threshold=0.2, --aSCNA_variance_threshold=0.025

4.4.7 Absolute Copy Number Analysis

Copy-number analysis was performed by using whole-exome sequencing (WES) coverage data and variant calls based on the tumor DNA. CNVEX [85,86] was used to estimate CNVs. Briefly, CNVEX estimate coverage within fixed genomic intervals and variant calls to compute B-allele frequencies (BAFs) at variant positions. Coverage values are then normalized for GC-bias using LOESS smoothing across targeted regions within the GC range of 0.3 and 0.7, using the span=0.5. All the GC normalized coverages and BAFs are then jointly segmented using a custom algorithm based on Circular Binary Segmentation (CBS) [93]. The resulting segmented copy-number profiles were then subject to joint inference of tumor purity and ploidy and absolute copy number states, implemented in CNVEX, which is most similar to the mathematical formalism of ABSOLUTE [94] and PureCN [21]. Because the copy-number inference problem can have multiple equally likely solutions, further biological insights are necessary to choose the most parsimonious result. The solutions were reviewed by two independent field experts and the most likely solution was selected.

Using the absolute copy number states (CN) , we defined absolute baseline CN of a tumor as the weighted median CN of its segments (weighted based on the length of the segment). Next we defined CNV events as:

- Segment gain: Any segment with copy number higher than the baseline CN
- Segment amplification: Any segment with at least 7 copies and CN higher than 2 times of the baseline CN of the chromosome it is located on.
- Segment deletion: Any segment with copy number lower than the baseline CN
- Segment homozygous deletion: Any segment with absolute copy number of 0
- Segment LOH: Any segment with minor copy number (i.e, Copy number of the allele with lower CN) of 0

4.4.8 Whole Exome Sequencing (WES) Quality Control

Total number of reads, percent of duplicated reads, and percent of chimeric reads were calculated for WES data using the same criteria as Picard [95]. We confirmed that these quality measures did not affect the downstream analysis including estimation of purity and ploidy derived from copy number and germline variant calls. Other quality measures such as GC-bias [21] and read mappability were factored directly in the CNV and variant calling processes to minimize the effects of data quality on sensitivity of the calls.

4.4.9 RNA-seq expression quantification

For RNA-seq data analysis, including read-trimming, alignment, post-processing, and quantification, we used the same methods as previously described [86]. Briefly, adapter trimmed paired-end 150bp short reads were aligned using STAR [96] version 2.4.0j to the GRCh38 reference supplemented with human oncogenic viruses. To improve the sensitivity of detecting fusion calls, synthetic single end reads were generated using bbduk from the trimmed paired-end reads. These reads were also aligned using STAR with optimized settings. After alignment, capture RNA-seq data were quantified using Kallisto [97]. The paired-end and single-end RNA-seq alignments were used subsequently as an input to CODAC, a component of TPO designed to call fusions, perform additional QC and realign reads using Minimap [98] and GMAP [99].

4.4.10 Tumor mutational burden calls

Tumor mutational burden (TMB) was defined, as previously proposed [100], as the total number of nonsynonymous somatic mutations (see above) per megabase of targeted regions (Agilent V4 Exome). No additional variant allele frequency threshold was used.

Mutation types included in the calculation include missense, nonsense, frameshift and splice-site variants.

4.4.11 Cancer Cell Fraction (CCF) calculation

CCF of each mutation was calculated similar to the method explained in [101]. In short, CCF is a value between 0 to 1 which estimates the proportion of tumor cells harboring a particular somatic mutation. CCF were calculated in two steps, i) multiplicity (M) of the mutation were estimated (i.e. how many alleles in the tumor harbor the mutation), ii) normalizing VAF for M and tumor purity. To calculate M, the allele-count with maximum likelihood (L) was selected according to equation:

$$M_L = Beta(Ef, ADT, RDT)$$

where:

$$Ef = \frac{Allele_count_{mut}}{(Allele_count_{mut} + allele_count_{wt}) + allele_count_{normal}}$$

$$= \frac{P * M}{P * C + (1 - P) * nC}$$

In the equations above, ADT is the alternative read count in tumor (i.e. number of reads supporting the variant), RDT is the reference read count in tumor (i.e. number of reads supporting the reference at the variant position), P is the purity of the sample, C is the total absolute copy number of the locus harboring variant and nC is the germline copy number of locus harboring the variant 2 for autosomes, and in males 1 for chrX and 1 for chrY, in females 2 for chrX and 0 for chrY. Using the equation (1) we are able to pick the M that maximizes the beta likelihood.

$$\widehat{m} = argmax_{M \in \{1,\ldots,C_i^A\}} Beta(\frac{P * M}{P * C + (1 - P) * nC}, ADT, RDT)$$

After estimating M, CCF was calculated as:

$$CCF = \frac{AFT * nC * (1 - P) + C * P}{m * P}$$

Where AFT is the alt allele frequency in the tumor. Any mutation with CCF > 0.87 was considered to be clonal. This threshold was used as a marginally relaxed version of thresholds previously used by other publications [27,28,102,103]. We relaxed the threshold to better capture the left tail of normal distribution of CCFs around CCF=1.

4.4.12 Mutation classification according to presence and clonality in patients

We classified somatic mutations into five distinct categories. The first category, termed 'Truncal', encompasses mutations that are clonal and ubiquitously present across all metastatic tumors in a patient. These mutations are indicative of early oncogenic events, present in the common ancestral clone giving rise to all subsequent metastatic tumors.

The second category, 'Clonal-Shared', includes mutations that, while clonal in at least one site, are also detected in more than one location, but are not Truncal. These mutations arise after Truncal mutations and are not present in the common ancestor, with the associated clone most likely seeding at least one other metastatic site.

'Subclonal-Shared' mutations constitute the third category. These mutations are present in multiple sites but never reach clonal dominance, suggesting a pattern of migration between sites for their associated clones rather than seeding any metastatic site.

The final two categories are 'Private' mutations, unique to a single site. 'Clonal-Private' mutations are clonal within their specific site, likely representing either a seeded clone from another site where the source is not captured due to undersampling or a mutation

that originated and clonally expanded within the same site. 'Subclonal-Private' mutations, in contrast, either developed within their observed site without achieving clonal status or migrated from an unsequenced source site.

4.4.13 Mutational cluster detection

The emergence of a new tumor clone, marked by a distinct set of mutations, results in similar CCF values for these mutations, as they exist in the same subset of cancer cells. This pattern is detectable through clustering the CCFs for each patient, a process we conducted using the PhylogicNDT [104].

Using the calculated CCFs, PhylogicNDT Cluster module was used to cluster mutations based on their CCF among all the samples from the same patient. PhylogicNDT Cluster module was run using the settings:

-ni 500 -rb --Pi_k_mu 2 --Pi_k_r 10

The resulted clusters were iteratively pruned according to the the following criteria:

- Each cluster was required to be supported by at least 3 mutations.
- If there exists a cluster with mean CCF difference of <10% from another cluster in all samples of a patient, the two clusters are merged together and a new mean CCF for the merged cluster is calculated using the mean of all the mutations in the new cluster per each sample.

4.4.14 Categorizing clones according to their position in phylogenetic trees

We sought to categorize clones in each patient based on the location of mutation clusters within the phylogenetic trees into three distinct groups: Truncal, Branch, and Leaf.

"Truncal" clones were defined as any cluster located at the trunk of the tree. Mutations in these clones were present and clonal across all patient sites. Given that mutations in truncal clones exist ubiquitously across all samples and consistently exhibit clonal characteristics, they can be perceived as mutations that are present in the common ancestor of all the clones, and have likely occurred prior to the invasive state of the tumor which leads to distant tissues metastasis.

"Leaf" clones, by contrast, referred to clusters that were exclusive to one site. Such clones, private to a site, provide insight into the variability and potential adaptability of tumor cells.

Finally, a third group of clones labeled as "Branch" comprised all clusters that did not fall within the Truncal or Leaf classifications. These clusters would typically be found in more than one site which will place them in between, or along the branches of the tree, hence the aptly named category. Branch clones can be explained as clones which appeared after the dissemination of clones from their common ancestor, but also were involved in the metastasis process and therefore their presence in more than one site.

4.4.15 Measuring the extent of "illusion of clonality"

To measure the extent of the "illusion of clonality" we looked at each mutation within a patient separately. For each patient's mutations, we measured in what proportion of samples from that patient the mutation is clonal (PC):

$$P_{C,i} = \frac{N_{C,i}}{N_{A,i}}$$

In the equation above, $N_{C,i}$ correspond to the number of samples in which mutation i is present and clonal. $N_{A,i}$correspond to the number of samples in which mutation i is present (clonal or subclonal). If a mutation was not present in a sample, we didn't include that sample in the denominator of $P_{C,i}$, since an absent mutation would not cause the illusion of clonality. Next, we averaged over all the $P_{C,i}$for all the mutations to derive a clonality proportion at the patient level. Depending on the number of samples that are being evaluated for each patient, $P_{C,i}$values can be different. We repeated the same procedure for combinations of two, three and four samples from each patient. In the cases of more than one sample, if a mutation is clonal and present in all the samples, it will be called clonal, otherwise we consider it as subclonal. In the end, we have a clonality rate for each patient per number of evaluated samples.

4.4.16 Estimating event detection rates

To calculate the detection rates of mutations and CNVs per different number of samples available, we first recorded all the mutations and CNVs which are detected using all samples of each patient. Next, we permuted all the 2, 3, and 4 pairings of samples within each patient, and evaluated if the patient's mutations or CNVs are observed in the pairing. Depending on the mutation or CNV of interest, we calculated the proportion of 2, 3, and 4 pairing in which the alteration of interest is observed. Lastly, to derive a detection rate for each alteration independent of patients, we averaged over all the detection rates derived for each patient.

4.4.17 Unobserved mutation extrapolation model

To extrapolate the number of unobserved mutations in each patient in the unsequenced tumors, we used the model explained in [105]. Deng et al. proposes a non-parametric model that can robustly predict the number of unobserved clones for long range extrapolation, based on the number of mutations already observed in the sequenced tumors. The model aims to extrapolate the species accumulation curve [106,107] in order to compare the diversity of populations based on samples of differing sizes [108]. This class of models has previously been employed in other biological applications such as estimating immune receptor diversity [109], or predicting the unseen variants in the human genome [110].

We used the implementation of this method, function ds.rSAC.bootstrap() in preseqR package [105] to perform extrapolation using the r=1, mt=20 settings. We ran the model on all the patients separately using the available samples for each corresponding patient. From there, we found the predicted number of mutations and clones per each additional number of samples for the patient. The number in the model corresponding to 100 samples was used as the maximum number of mutations observable in each patient.

We validated the model on mutation data by selecting all the patients which had at least 4 sites available. We used 3 sites to train the model and then made a prediction of the number of total observable mutations for the case of 4 samples available. We then compared the predicted value with the actual observed value from the data to validate the model performance.

4.4.18 Reconstruction of clonal phylogenetic trees

Phylogenetic trees were built using the CONIPHER (version 0.1.0) [48]. In short, CONIPHER uses the pigeonhole principle and infinite sites hypothesis [111] to reconstruct

a phylogenetic tree based on CCF clusters of mutations i.e. clones. In this analysis, we input filtered synonymous and nonsynonymous mutations to the tool. In the process of reconstructing phylogenetic trees, CONIPHER removes any cluster that might be formed due to copy number deletions or LOH, doesn't fit the pigeonhole principle or might cause a cycle in the tree. To ensure no driver mutation has been removed, all the removed mutations were manually assessed according to a compiled list of mutated genes-of-interest (GOI) in prostate cancer same as in Figure 1C. 5 variants were detected to have been removed and were present in the list of GOI. Those mutations were merged back to the cluster with the closest mean CCF to the variant CCF.

4.4.19 LOH permutation model

We used a permutation test to quantitatively assess the statistical significance of the observed overlap in loss of heterozygosity (LOH) across the 4 samples from patient WA_24, a permutation test. This test measures the probability of observing concurrent LOH of genes in 4 tumors, if the LOHs are happening in random loci of each tumor (i.e, null hypothesis).

The first step of the test involves classifying segments according to their LOH status into 2 groups of LOH and noLOH. Next, in each tumor, LOH labels were uniformly shuffled among segments to simulate a randomly labeled genome. To make sure the rate of LOH in the simulated genome is equal or higher than the tumor (to achieve an upper bound of LOH overlaps), we augmented 20% more LOH labels in each tumor and removed any simulation that has less LOH in their genome compared to the tumors. This simulation was repeated a million times, and in each step, the number of overlapping genes with LOH in all 4 simulated samples were measured. The observed number for genes with LOH in all 4 samples of patient

WA_24 was then compared against this simulated distribution, yielding an empirical P-value.

4.5 References

1. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501: 328–337.
2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339: 1546–1558.
3. Menzies AM, Haydu LE, Carlino MS, Azer MWF, Carr PJA, Kefford RF, et al. Inter- and intra-patient heterogeneity of response and progression to targeted therapy in metastatic melanoma. PLoS One. 2014;9: e85004.
4. Warner EW, Van der Eecken K, Murtha AJ, Kwan EM, Herberts C, Sipola J, et al. Multiregion sampling of de novo metastatic prostate cancer reveals complex polyclonality and augments clinical genotyping. Nat Cancer. 2024;5: 114–130.
5. Singhal U, Nallandhighal S, Tosoian JJ, Hu K, Pham TM, Stangl-Kremser J, et al. Integrative multi-region molecular profiling of primary prostate cancer in men with synchronous lymph node metastasis. Nat Commun. 2024;15: 4341.
6. Haffner MC, Zwart W, Roudier MP, True LD, Nelson WG, Epstein JI, et al. Genomic and phenotypic heterogeneity in prostate cancer. Nat Rev Urol. 2021;18: 79–92.
7. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015;520: 353–357.
8. Devarakonda S, Govindan R. Clonal evolution: multiregion sequencing of esophageal adenocarcinoma before and after chemotherapy. Cancer Discov. 2015;5: 796–798.
9. Espiritu SMG, Liu LY, Rubanova Y, Bhandari V, Holgersen EM, Szyca LM, et al. The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. Cell. 2018;173: 1003–1013.e15.
10. Erickson A, Hayes A, Rajakumar T, Verrill C, Bryant RJ, Hamdy FC, et al. A Systematic Review of Prostate Cancer Heterogeneity: Understanding the Clonal Ancestry of Multifocal Disease. Eur Urol Oncol. 2021;4: 358–369.
11. Løvf M, Zhao S, Axcrona U, Johannessen B, Bakken AC, Carm KT, et al. Multifocal Primary Prostate Cancer Exhibits High Degree of Genomic Heterogeneity. Eur Urol. 2019;75: 498–505.
12. Kim T-M, Jung S-H, Baek I-P, Lee S-H, Choi Y-J, Lee J-Y, et al. Regional biases in mutation screening due to intratumoural heterogeneity of prostate cancer. J Pathol. 2014;233: 425–435.
13. Parry MA, Srivastava S, Ali A, Cannistraci A, Antonello J, Barros-Silva JD, et al. Genomic Evaluation of Multiparametric Magnetic Resonance Imaging-visible and -

nonvisible Lesions in Clinically Localised Prostate Cancer. Eur Urol Oncol. 2019;2: 1–11.
14. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cold Spring Harbor Laboratory. 2020. p. 312041. doi:10.1101/312041
15. Hong MKH, Macintyre G, Wedge DC, Van Loo P, Patel K, Lunke S, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. Nat Commun. 2015;6: 6605.
16. Hasan AMM, Cremaschi P, Wetterskog D, Jayaram A, Wong SQ, Williams S, et al. Copy number architectures define treatment-mediated selection of lethal prostate cancer clones. Nat Commun. 2023;14: 4823.
17. Cieslik M, Chugh R, Wu Y-M, Wu M, Brennan C, Lonigro R, et al. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. Genome Res. 2015;25: 1372–1381.
18. Shah RB, Mehra R, Chinnaiyan AM, Shen R, Ghosh D, Zhou M, et al. Androgen-independent prostate cancer is a heterogeneous group of diseases: lessons from a rapid autopsy program. Cancer Res. 2004;64: 9209–9216.
19. Udager AM, Shi Y, Tomlins SA, Alva A, Siddiqui J, Cao X, et al. Frequent discordance between ERG gene rearrangement and ERG protein expression in a rapid autopsy cohort of patients with lethal, metastatic, castration-resistant prostate cancer. Prostate. 2014;74: 1199–1208.
20. Taylor-Weiner A, Stewart C, Giordano T, Miller M, Rosenberg M, Macbeth A, et al. DeTiN: overcoming tumor-in-normal contamination. Nat Methods. 2018;15: 531–534.
21. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11: 13.
22. Robinson D, Van Allen EM, Wu Y-M, Schultz N, Lonigro RJ, Mosquera J-M, et al. Integrative clinical genomics of advanced prostate cancer. Cell. 2015;161: 1215–1228.
23. Center for Drug Evaluation, Research. FDA approves niraparib and abiraterone acetate plus prednisone for BRCA-mutated metastatic castration-resistant prostate cancer. In: U.S. Food and Drug Administration [Internet]. FDA; 11 Aug 2023 [cited 30 Jul 2024]. Available: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-niraparib-and-abiraterone-acetate-plus-prednisone-brca-mutated-metastatic-castration
24. Chi KN, Rathkopf D, Smith MR, Efstathiou E, Attard G, Olmos D, et al. Niraparib and abiraterone acetate for metastatic castration-resistant prostate cancer. J Clin Oncol. 2023;41: 3339–3351.

25. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346: 251–256.
26. Dentro SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harb Perspect Med. 2017;7. doi:10.1101/cshperspect.a026625
27. Abbosh C, Frankell AM, Harrison T, Kisistok J, Garnett A, Johnson L, et al. Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. Nature. 2023;616: 553–562.
28. Manier S, Park J, Capelletti M, Bustoros M, Freeman SS, Ha G, et al. Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. Nat Commun. 2018;9: 1691.
29. Hoang T, Sutera P, Nguyen T, Chang JH, Jagtap S, Song Y, et al. The impact of TP53 mutations and use of the TP53-mutation-reactivating agent APR-246 on metastatic castrate-sensitive prostate cancer. J Clin Orthod. 2023;41: e17077–e17077.
30. Kogan-Sakin I, Tabach Y, Buganim Y, Molchadsky A, Solomon H, Madar S, et al. Mutant p53(R175H) upregulates Twist1 expression and promotes epithelial-mesenchymal transition in immortalized prostate cells. Cell Death Differ. 2011;18: 271–281.
31. Venkatesan S, Swanton C. Tumor evolutionary principles: How intratumor heterogeneity influences cancer treatment and outcome. Am Soc Clin Oncol Educ Book. 2016;35: e141–9.
32. Nowell PC. The clonal evolution of tumor cell populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. Science. 1976;194: 23–28.
33. Wu Y-M, Cieślik M, Lonigro RJ, Vats P, Reimers MA, Cao X, et al. Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. Cell. 2018;173: 1770–1782.e14.
34. Zhang W, Dong Y, Sartor O, Zhang K. Deciphering the Increased Prevalence of TP53 Mutations in Metastatic Prostate Cancer. Cancer Inform. 2022;21: 11769351221087046.
35. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163: 1011–1025.
36. Heemers HV, Tindall DJ. Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex. Endocr Rev. 2007;28: 778–808.
37. Wyatt AW, Annala M, Aggarwal R, Beja K, Feng F, Youngren J, et al. Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer. J Natl Cancer Inst. 2017;109. doi:10.1093/jnci/djx118

38. Morrison GJ, Goldkorn A. Development and Application of Liquid Biopsies in Metastatic Prostate Cancer. Curr Oncol Rep. 2018;20: 35.
39. Chakraborty S, Arora A, Begg CB, Shen R. Using somatic variant richness to mine signals from rare variants in the cancer genome. Nat Commun. 2019;10: 5506.
40. Gravel S. Predicting Discovery Rates of Genomic Features. Genetics. 2014;197: 601–610.
41. Baker TM, Lai S, Lesluyes T, Yan H, Verfaillie A, Dentro S, et al. The history of chromosomal instability in genome doubled tumors. bioRxiv. 2023. p. 2023.10.22.563273. doi:10.1101/2023.10.22.563273
42. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578: 122–128.
43. López S, Lim EL, Horswell S, Haase K, Huebner A, Dietzen M, et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat Genet. 2020;52: 283–293.
44. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, Favero F, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. NPJ Breast Cancer. 2018;4: 16.
45. Sha D, Jin Z, Budczies J, Kluck K, Stenzinger A, Sinicrope FA. Tumor mutational burden as a predictive biomarker in solid tumors. Cancer Discov. 2020;10: 1808–1825.
46. Jamaspishvili T, Berman DM, Ross AE, Scher HI, De Marzo AM, Squire JA, et al. Clinical implications of PTEN loss in prostate cancer. Nat Rev Urol. 2018;15: 222–234.
47. Galletti G, Leach BI, Lam L, Tagawa ST. Mechanisms of resistance to systemic therapy in metastatic castration-resistant prostate cancer. Cancer Treat Rev. 2017;57: 16–27.
48. Grigoriadis K, Huebner A, Bunkum A, Colliver E, Frankell AM, Hill MS, et al. CONIPHER: a computational framework for scalable phylogenetic reconstruction with error correction. 2023 [cited 2 May 2023]. doi:10.21203/rs.3.pex-2158/v1
49. Parolia A, Cieslik M, Chu S-C, Xiao L, Ouchi T, Zhang Y, et al. Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. Nature. 2019;571: 413–418.
50. Al Bakir M, Huebner A, Martínez-Ruiz C, Grigoriadis K, Watkins TBK, Pich O, et al. The evolution of non-small cell lung cancer metastases in TRACERx. Nature. 2023;616: 534–542.
51. Birkbak NJ, McGranahan N. Cancer genome evolutionary trajectories in metastasis. Cancer Cell. 2020;37: 8–19.
52. El-Kebir M, Satas G, Oesper L, Raphael BJ. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Syst. 2016;3: 43–53.

53. Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics. 1969;61: 893–903.
54. Demeulemeester J, Dentro SC, Gerstung M, Van Loo P. Biallelic mutations in cancer genomes reveal local mutational determinants. Nat Genet. 2022;54: 128–133.
55. Watkins TBK, Lim EL, Petkovic M, Elizalde S, Birkbak NJ, Wilson GA, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020. doi:10.1038/s41586-020-2698-6
56. Petkovic M, Watkins TBK, Colliver EC, Laskina S, Swanton C, Haase K, et al. Whole-genome doubling-aware copy number phylogenies for cancer evolution with MEDICC2. Cold Spring Harbor Laboratory. 2021. p. 2021.02.28.433227. doi:10.1101/2021.02.28.433227
57. Bennett RJ, Forche A, Berman J. Rapid mechanisms for generating genome diversity: whole ploidy shifts, aneuploidy, and loss of heterozygosity. Cold Spring Harb Perspect Med. 2014;4. doi:10.1101/cshperspect.a019604
58. Good P. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media; 2013.
59. Mehra R, Tomlins SA, Yu J, Cao X, Wang L, Menon A, et al. Characterization of TMPRSS2-ETS gene aberrations in androgen-independent metastatic prostate cancer. Cancer Res. 2008;68: 3584–3590.
60. Labrecque MP, Coleman IM, Brown LG, True LD, Kollath L, Lakely B, et al. Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer. J Clin Invest. 2019;129: 4492–4505.
61. Beltran H, Rickman DS, Park K, Chae SS, Sboner A, MacDonald TY, et al. Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. Cancer Discov. 2011;1: 487–495.
62. Nyquist MD, Corella A, Coleman I, De Sarkar N, Kaipainen A, Ha G, et al. Combined TP53 and RB1 Loss Promotes Prostate Cancer Resistance to a Spectrum of Therapeutics and Confers Vulnerability to Replication Stress. Cell Rep. 2020;31: 107669.
63. Andreoiu M, Cheng L. Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. Hum Pathol. 2010;41: 781–793.
64. Woodcock DJ, Riabchenko E, Taavitsainen S, Kankainen M, Gundem G, Brewer DS, et al. Prostate cancer evolution from multilineage primary to single lineage metastases with implications for liquid biopsy. Nat Commun. 2020;11: 5070.
65. Nurminen A, Jaatinen S, Taavitsainen S, Högnäs G, Lesluyes T, Ansari-Pour N, et al. Cancer origin tracing and timing in two high-risk prostate cancers using multisample whole genome analysis: prospects for personalized medicine. Genome Med. 2023;15: 82.

66. Hao J-J, Lin D-C, Dinh HQ, Mayakonda A, Jiang Y-Y, Chang C, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nat Genet. 2016;48: 1500–1507.
67. Frankell AM, Dietzen M, Al Bakir M, Lim EL, Karasaki T, Ward S, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature. 2023;616: 525–533.
68. Burrell RA, Swanton C. Re-evaluating clonal dominance in cancer evolution. Trends Cancer. 2016;2: 263–276.
69. Willyard C. Cancer therapy: an evolved approach. Nature. 2016;532: 166–168.
70. Burrell RA, Swanton C. Tumour heterogeneity and the evolution of polyclonal drug resistance. Mol Oncol. 2014;8: 1095–1111.
71. Yan T, Cui H, Zhou Y, Yang B, Kong P, Zhang Y, et al. Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma. Nat Commun. 2019;10: 1670.
72. Abida W, Cyrta J, Heller G, Prandi D, Armenia J, Coleman I, et al. Genomic correlates of clinical outcome in advanced prostate cancer. Proc Natl Acad Sci U S A. 2019;116: 11428–11436.
73. Conteduca V, Mosca A, Brighi N, de Giorgi U, Rescigno P. New prognostic biomarkers in metastatic castration-resistant prostate cancer. Cells. 2021;10: 193.
74. Wang Y, Singhal U, Qiao Y, Kasputis T, Chung J-S, Zhao H, et al. Wnt Signaling Drives Prostate Cancer Bone Metastatic Tropism and Invasion. Transl Oncol. 2020;13: 100747.
75. Leong SP, Witz IP, Sagi-Assif O, Izraely S, Sleeman J, Piening B, et al. Cancer microenvironment and genomics: evolution in process. Clin Exp Metastasis. 2022;39: 85–99.
76. García-Mulero S, Alonso MH, Pardo J, Santos C, Sanjuan X, Salazar R, et al. Lung metastases share common immune features regardless of primary tumor origin. J Immunother Cancer. 2020;8. doi:10.1136/jitc-2019-000491
77. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6: 8971.
78. Zhang C, Cheng W, Ren X, Wang Z, Liu X, Li G, et al. Tumor Purity as an Underlying Key Factor in Glioma. Clin Cancer Res. 2017;23: 6279–6291.
79. Parker CC, James ND, Brawley CD, Clarke NW, Ali A, Amos CL, et al. Radiotherapy to the prostate for men with metastatic prostate cancer in the UK and Switzerland: Long-term results from the STAMPEDE randomised controlled trial. PLoS Med. 2022;19: e1003998.
80. Parker CC, James ND, Brawley CD, Clarke NW, Hoyle AP, Ali A, et al. Radiotherapy to the primary tumour for newly diagnosed, metastatic prostate cancer (STAMPEDE): a randomised controlled phase 3 trial. Lancet. 2018;392: 2353–2366.

81. Rubin MA, Putzi M, Mucci N, Smith DC, Wojno K, Korenchuk S, et al. Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. Clin Cancer Res. 2000;6: 1038–1045.
82. Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548: 297–303.
83. Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun. 2018;9: 4038.
84. Li GX, Chen L, Hsiao Y, Mannan R, Zhang Y, Luo J, et al. Comprehensive proteogenomic characterization of rare kidney tumors. Cell Rep Med. 2024;5: 101547.
85. Li Y, Lih T-SM, Dhanasekaran SM, Mannan R, Chen L, Cieslik M, et al. Histopathologic and proteogenomic heterogeneity reveals features of clear cell renal cell carcinoma aggressiveness. Cancer Cell. 2023;41: 139–163.e17.
86. Chowdhury S, Kennedy JJ, Ivey RG, Murillo OD, Hosseini N, Song X, et al. Proteogenomic analysis of chemo-refractory high-grade serous ovarian cancer. Cell. 2023;186: 3476–3498.e35.
87. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. 2014 [cited 8 May 2024]. Available: https://escholarship.org/uc/item/1h3515gn
88. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.
89. Freed D, Aldana R, Weber JA, Edwards JS. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. bioRxiv. 2017. p. 115717. doi:10.1101/115717
90. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019. p. 861054. doi:10.1101/861054
91. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122.
92. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17: 118.
93. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5: 557–572.
94. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30: 413–421.
95. GitHub - broadinstitute/picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. In: GitHub [Internet]. [cited 26 Jun 2024]. Available: https://github.com/broadinstitute/picard

96. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.
97. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34: 525–527.
98. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32: 2103–2110.
99. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. Methods Mol Biol. 2016;1418: 283–334.
100. Meléndez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, D'Haene N. Methods of measurement for tumor mutational burden in tumor tissue. Transl Lung Cancer Res. 2018;7: 661–667.
101. Cun Y, Yang T-P, Achter V, Lang U, Peifer M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. Nat Protoc. 2018;13: 1488–1501.
102. Mai Z, Liu Q, Wang X, Xie J, Yuan J, Zhong J, et al. Integration of Tumor Heterogeneity for Recurrence Prediction in Patients with Esophageal Squamous Cell Cancer. Cancers . 2021;13. doi:10.3390/cancers13236084
103. Diamond B, Yellapantula V, Rustad EH, Maclachlan KH, Mayerhoefer M, Kaiser M, et al. Positive selection as the unifying force for clonal evolution in multiple myeloma. Leukemia. 2021;35: 1511–1515.
104. Leshchiner I, Livitz D, Gainor JF, Rosebrock D, Spiro O, Martinez A, et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. Cold Spring Harbor Laboratory. 2019. p. 508127. doi:10.1101/508127
105. Deng C, Daley T, Smith AD. Applications of species accumulation curves in large-scale biological data analysis. Quant Biol. 2015;3: 135–144.
106. Good IJ, Toulmin GH. The number of new species, and the increase in population coverage, when a sample is increased. Biometrika. 1956;43: 45–63.
107. Colwell RK, Mao CX, Chang J. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology. 2004;85: 2717–2727.
108. Deng C, Daley T, Calabrese P, Ren J, Smith AD. Estimating the number of species to attain sufficient representation in a random sample. arXiv [stat.ME]. 2016. Available: http://arxiv.org/abs/1607.02804
109. Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Philos Trans R Soc Lond B Biol Sci. 2015;370. doi:10.1098/rstb.2014.0291
110. Ionita-Laza I, Lange C, M Laird N. Estimating the number of unseen variants in the human genome. Proc Natl Acad Sci U S A. 2009;106: 5008–5013.

111. Tarabichi M, Salcedo A, Deshwar AG, Ni Leathlobhair M, Wintersinger J, Wedge DC, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nat Methods. 2021;18: 144–155.

Chapter 5 MHCnvex: A Probabilistic Method For the Detection of MHC Class I and II Copy-Number Alterations

5.1 Introduction

Disrupting Immune surveillance is one of the main routes for malignant cells to survive in the body [1]. Tumor cells' major histocompatibility complex (MHC) is frequently altered to disrupt cell antigen representing pathway (ATP) [2] and therefore escape surveillance by the immune system. One of the main mechanisms by which tumor cells evade the immune system is the Loss of Heterozygosity (LOH) of MHC genes through copy number alteration. It is critical to understand this mechanism in the context of cancer immunology as it can help to reveal new therapeutic opportunities. However, High degree of polymorphism in the MHC locus has made it difficult for conventional copy number variation (CNV) inference methods to accurately profile copy numbers for genes in the MHC locus. One of the key challenges arises from the polymorphic nature of MHC region, is to correctly align sequencing reads to the genome. Inaccurate alignment can lead to inaccurate coverage estimates of the locus, thereby causing unreliable CNV calls. Despite the significance of identifying CNVs within the MHC region, there has been a dearth of methodologies developed to address this issue.

In this chapter, we present MHCnvex, a bioinformatics tool developed as an R package. It utilizes personalized germline references and likelihood modeling to effectively determine MHC class I and II CNV and LOH events with high precision. Unlike other tools [3,4] in the field which only look at the MHC region, MHCnvex integrates MHC loci into adjacent regions on chromosome 6 p-arm and performs a comprehensive CNV analysis of chromosome 6 p-arm. This

will reduce the inherent noise introduced into the data at sequencing and analysis steps, which will be diluted when analyzing larger regions of the genome. To the best of our knowledge, MHCnvex is the only tool that can detect LOH and absolute copy number of genes in both class 1 and class 2, providing a valuable tool for the scientific community. To widen the applicability of the tool, MHCnvex allows for the MHC analysis using WGS, WXS or both if available. To increase the utility of MHCnvex in a clinical setting, we have incorporated a "tumor only" method that allows the tool to operate in the absence of a matching normal by using a pool of non-matching normal samples from the same capture kit as tumor. This modification enhances the flexibility of MHCnvex, making it more suitable for use in a clinical setting where matched normal samples may not be readily available.

5.2 Results

5.2.1 MHCnvex significantly removes the coverage noise in the MHC region

The accurate calculation of read coverage for each locus is a crucial step in the analysis of copy number variations. Inaccurate coverage estimates can lead to over-segmentation of the genome, resulting in false CNV calls as the algorithm may interpret each segment as a separate CNV event. Therefore, it is imperative to accurately determine the read coverage of MHC locus to ensure the reliability of CNV analysis. In theory, coverage in a normal sample for all the patients should be similar as there are no CNV events, but the polymorphic nature of the MHC region can cause variabilities in the calculated coverage. We evaluated MHCnvex ability to reduce these variabilities compared to conventional CNV analysis tools using a measure of “Coverage variance” in WXS data. Variance in the coverage of non-polymorphic regions of chromosome 6 arm p was used as a baseline since these loci are not prone to errors involved with polymorphism. Our analysis showed that the variance in coverage of the MHC region obtained from the standard

approach was significantly higher than the variance expected in coverage of arm p of chromosome 6 (P-value < 1e-16, **Figure 5.1**).

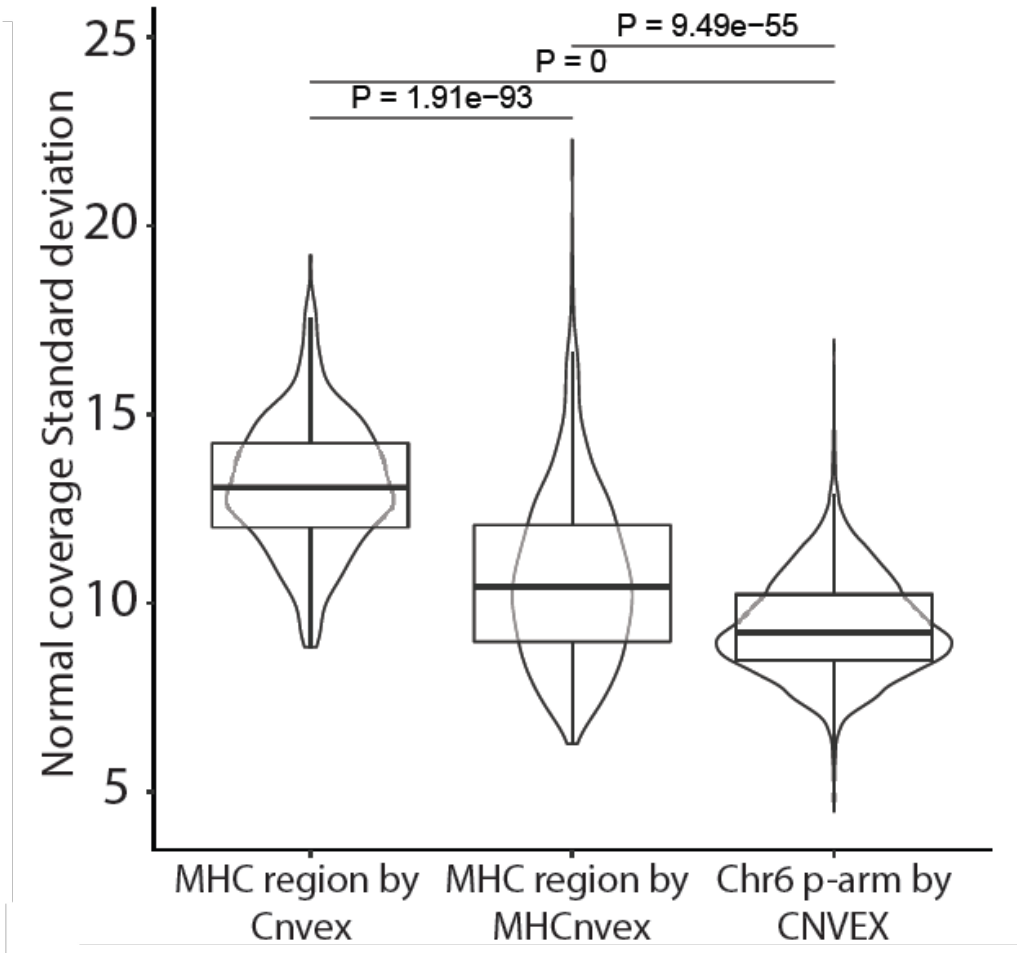


Figure 5.1:Standard deviation of MHC region calculated by MHCnvex has a significant SD reduction compared to the same value calculated by standard CNV approach CNVEX

When the same analysis was conducted using MHCnvex, we observed a significant reduction in variance compared to the standard approach for the MHC region (P-value < 1e-16). However, we still observed a higher variance rate for MHCnvex compared to the baseline which can be explained partly by the possible errors in data processing steps (i.e. Haplotyping of MHC region, pooling and library preparation according to the reference genome, etc.) prior to the

MHCnvex analysis. Overall, MHCnvex shows significantly less coverage variance among normal samples compared to the same region when coverage is calculated by standard approaches.

5.2.2 MHCnvex significantly reduces the number of segments in the MHC region

Another crucial step in the identification of CNV is the segmentation of loci based on their log-ratio and B allele frequency (BAF) values. However, regions with high levels of variability in coverage values are prone to over-segmentation and subsequently, the generation of faulty CNV calls. This variability is specially increased in the MHC region due to the polymorphism nature of the locus. To evaluate the ability of MHCnvex to avoid over-segmentation, we compared the number of segments produced by MHCnvex to those produced by standard approaches in the same region. A 15Mbp region on chromosome 6, comprising a 3Mbp MHC region and 6Mbp flanking regions on either side, was selected for this analysis. Additionally, we utilized the median number of segments for all other regions of the genome tiled into 15Mbp loci as a baseline.

MHCnvex showed a significant decrease in the number of segments compared to the standard approach (P-value in the figure, **Figure 5.2**), which is in the same direction as the baseline number of segments.

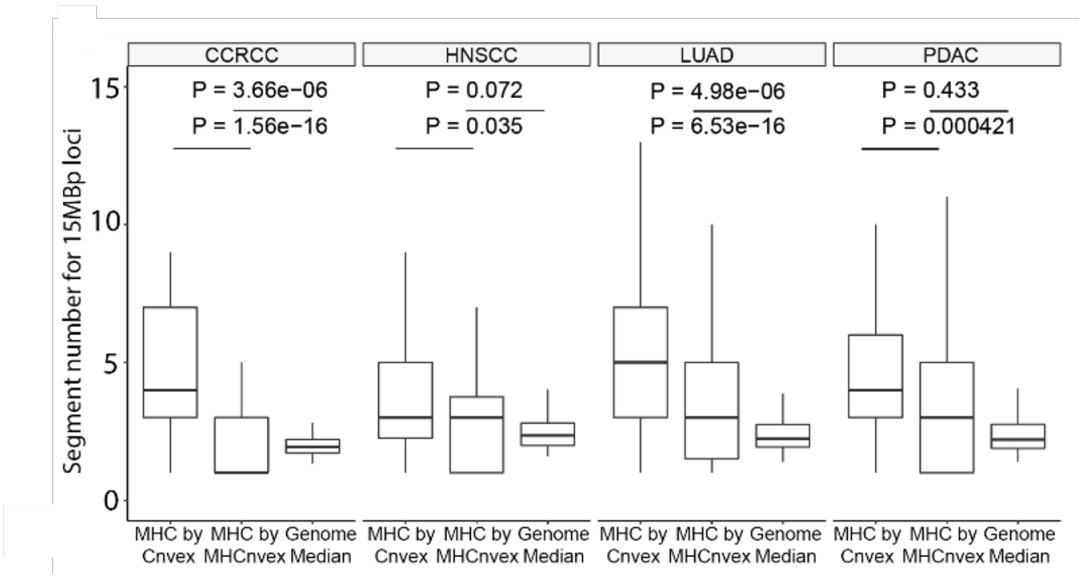


Figure 5.2: Number of segments present in the 15Mb window of genome and MHC region.

5.2.3 MHCnvex shows consistent results using WXS and WGS data

To test the concordance of different methods of CNV calling by MHCnvex, we ran MHCnvex with two different types of input: (i) WGS data, and (ii) WXS data. Copy number calls for flanking regions of MHC genes were also used to assess the concordance of MHCnvex inferred copy numbers with that of flanking regions. In terms of performance, MHCnvex showed a concordance of above 90% with copy numbers of flanking regions for both WGS and WXS settings. MHCnvex (WGS setting) showed a high concordance of 97% with CNVEX (using only WGS data) (**Figure 5.3**). Further, comparison of the MHCnvex WXS setting with CNVEX (ran with WGS data) showed a concordance of 92%. Comparison of the results from MHCnvex WXS and MHCnvex WGS settings also showed a high concordance of 91%. Some degree of inconsistency between WXS and WGS can be explained by the higher resolution of WXS setting in detecting focal events which WGS analysis doesn't have. Next, we compared samples showing inconsistency in CNV calls between the WXS and WGS setting of MHCnvex.

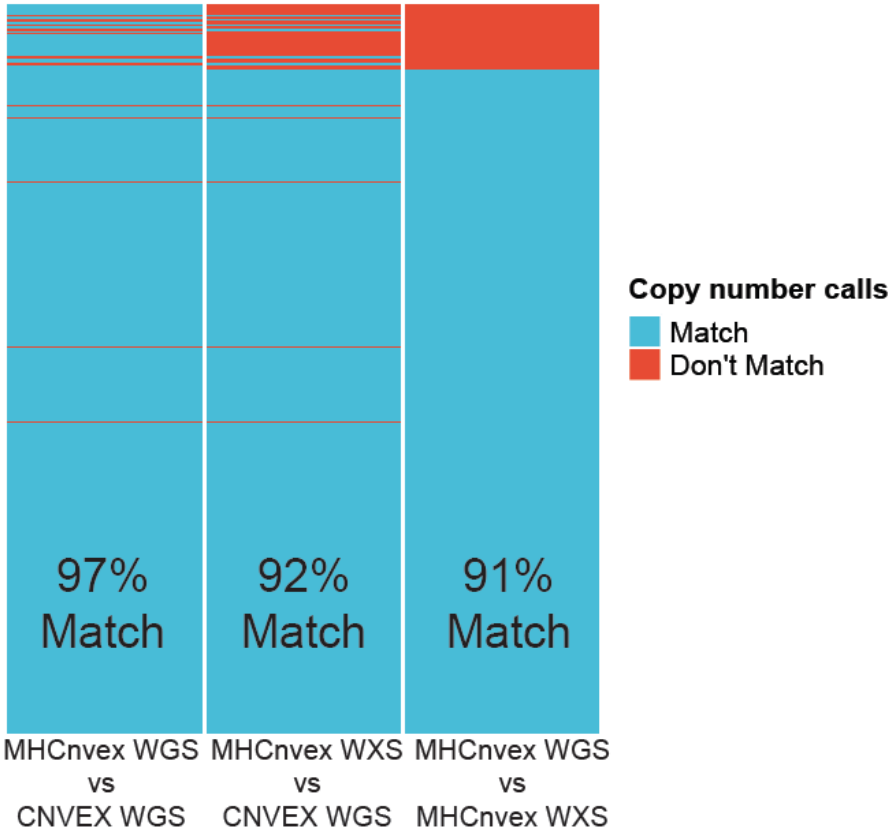


Figure 5.3: CNV agreement between MHCnvex and a general CNV calling tool for different settings of WGS and WXS.

Our analysis revealed that samples with mismatch had significantly lower purity (P-Value = 6.62e05, **Figure 5.3**). This findings are consistent with results from other MHC LOH calling tools, as they have also reported lower accuracy for tumors with low purity[3,5].

5.2.4 MHCnvex LOH calls correlate with changes in RNA expression

Next, we evaluated the accuracy of LOH calls using the RNA allelic-expression differences. Three categories of samples were analyzed: those with LOH detected by MHCnvex, those with equal copy numbers of each haplotype (referred to as "balanced"), and normal samples. Our study specifically focused on MHC class-I genes, as they are highly expressed in all cell types, allowing for the comparison of normal and tumor samples within each cohort. Our results showed

that, across all cohorts, the LOH group had a significantly higher difference in allelic RNA expression compared to the other two groups (P-value shown in figure, **Figure 5.4**).

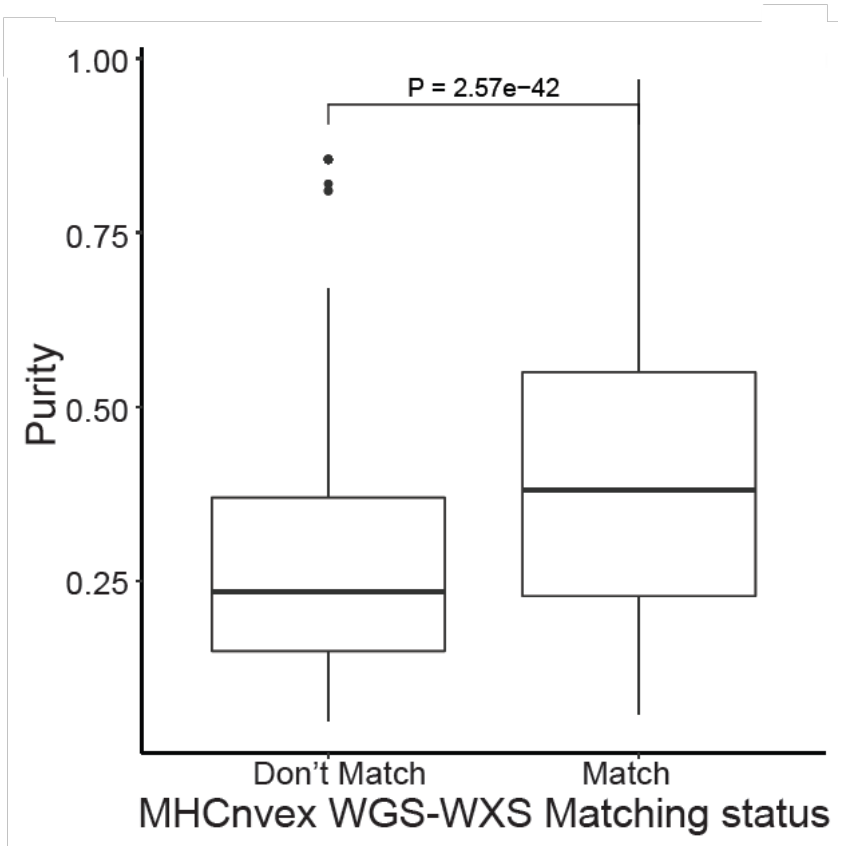


Figure 5.4: Samples with inconsistence CNV calls in panel D have significantly lower purity compared to the consistent ones

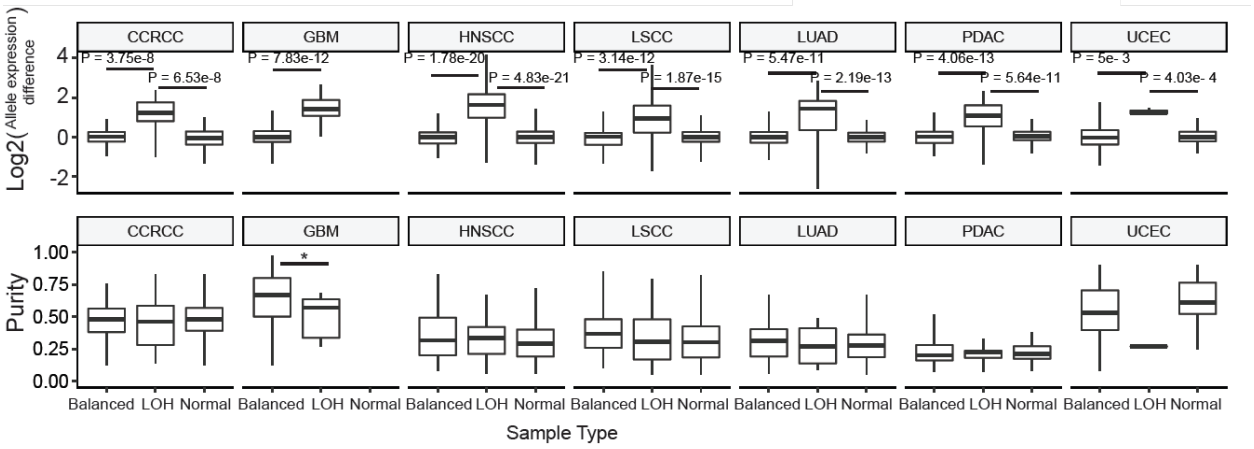


Figure 5.5: Top panel shows the difference in allelic expression of MHC class-I genes. Across all cohorts, samples detected by MHCnvex with MHC class-I LOH have significantly higher difference in expression of their alleles compared to normal samples or tumor samples detected to have balanced allelic number. Bottom panel shows that

these different groups across all cohorts have comparable purity and therefor, in this analysis, purity would not affect the difference in expression

Additionally, we found that the "balanced" group had similar average differences to the normal group, as expected given their equal allele copy numbers. The average purity of each group within each cohort was also similar, suggesting that purity is unlikely to be a confounding factor in the results (**Figure 5.5**).

5.3 Discussion

In this study, we present MHCnvex, an algorithm for the detection of copy number variation and loss of heterozygosity in the highly polymorphic MHC class I and II regions. MHCnvex is versatile and can be applied to both whole exome sequencing and whole genome sequencing data separately. It can also integrate them (if available) for a better LR and BAF calculations. To further enhance the clinical utility of MHCnvex, it is designed to work in instances without a matched normal for the tumor. One of the key features of MHC is that it takes advantage of the additional information from regions adjacent to the MHC in the process of segmentation, thus removing local biases of the data for the analysis. Our results show that MHCnvex has reduced noise caused by polymorphisms in the region, both at the coverage and segmentation level. Furthermore, we used allelic-expression differences in RNA-seq data to confirm that samples detected to have LOH indeed show significantly higher allelic-expression differences compared to the samples that are detected to have a balanced allelic copy number.

5.4 Methods

5.4.1 Data acquisition

In this study, whole exome sequencing (WXS), whole genome sequencing (WGS) and polyA RNA sequencing data were obtained for 712 patients from the CPTAC3 pan-cancer cohort. This data was processed using our in-house genomic analysis pipeline, TPO, for variant calling,

copy number analysis in genomic data and RNA quantification for RNA sequencing data. This cohort includes 7 different cancers with some of them (LSCC and LUSC) already reported to be enriched with HLA LOH events [3]. All the tumor samples have a matching normal except for the GBM cohort.

5.4.2 Overview of the algorithm

MHCnvex is an algorithm for copy number analysis of the MHC region, specifically MHC class I and II genes. The MHC locus is the most polymorphic region of the human genome [6] and this feature makes it challenging to accurately determine the read coverage and the copy number of MHC genes.

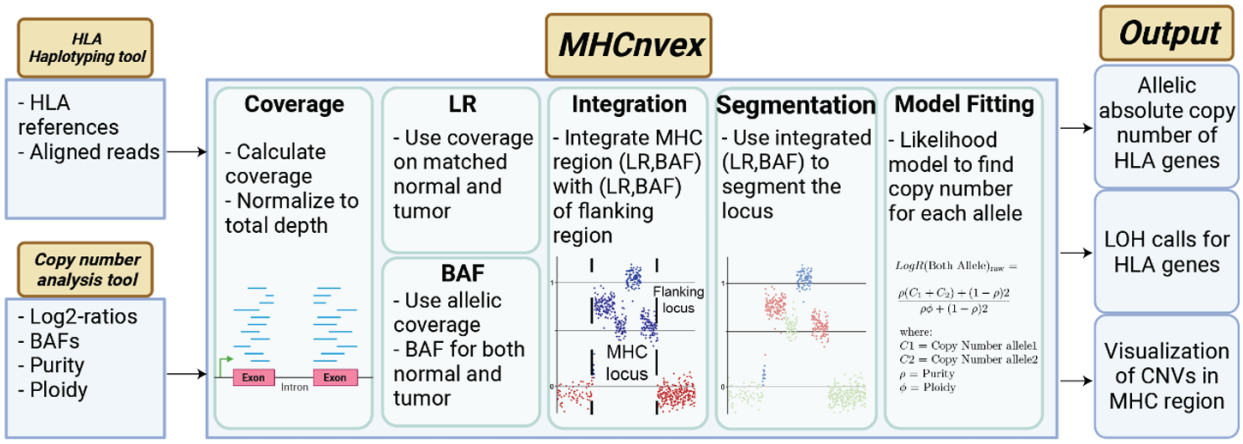


Figure 5.6: MHCnvex accepts input from an HLA haplotyping and a CNV analysis tool. Then works in 7 steps to call CNV and LOH for genes in MHC class-I and class-II. Finally, it outputs allelic copy number of each gene, LOH call for each gene, and a visualization of the loci. 3 examples of the MHC region have been shown

To solve this problem, MHCnvex incorporates personalized germline references for the MHC locus which can be inferred by haplotyping tools such as OptiType [7] or Hapster [8]. In order to calculate an accurate coverage, MHCnvex utilizes aligned reads to personalized gemline references as input. Furthermore, MHCnvex uses a segmental approach that leverages the fact that CNV events often occur in contiguous segments along the genome. To improve the segmentation accuracy, MHCnvex considers not only the MHC locus itself, but also incorporates adjacent

regions into its analysis. This unique approach provides a more comprehensive understanding of the MHC region's copy-number profile compared to other tools that only focus on the MHC locus alone. Recent years have seen the development of several tools aimed at addressing the challenge of detecting LOH in the MHC region, such as LOHHLA [3] and DASH [5]. However, to the best of our knowledge, none of these methods take into account the potential impact of integrated segmentation of MHC and flanking regions simultaneously on the detection of LOH for the MHC region. MHCnvex works in the 7 following steps (**Figure 5.6**):

1. Coverage calculation: To calculate the read coverage for each locus, MHCnvex counts reads mapped to each base pair using Mosdepth [9]. To ensure that the analyzed tumor and normal samples have comparable coverage, MHCnvex normalizes raw counts to the total sequencing depth.

2. MHC LR calculation: In order to calculate the LR of the MHC region, coverage of normal samples are used to detect local coverage peaks in MHC loci. Regions in normal samples with coverage peak can provide more reliable measurements as they have more reads mapped to them. By utilizing this approach, the LR calculation becomes independent of the capture panel used for sequencing. If a matched normal sample is not available, MHCnvex can use non matched normals to find the local peaks. Once the peaks for each gene are identified, adjacent base pairs (default 50 base pairs on each side) are pooled together to form one bin on the gene. Depending on the number of peaks per each gene, multiple bins per gene will form. This approach ensures that there are enough reads per bin to call reliable LR values and reduces noise in the segmentation step. By having the bins and their coverages in the normal and tumor samples, MHCnvex is then able to calculate the log2-ratio values as:

$$LR(bin_i) = log2(\frac{Tumor\ coverage\ on\ bin_i)}{Normal\ coverage\ on\ bin_i)}$$

3. MHC BAF calculation: BAF values for the MHC region are calculated by utilizing the coverage of each haplotype determined in the log2-ratio (LR) calculation step. Loci at which the two haplotypes are mismatched are identified as Single Nucleotide Polymorphisms (SNPs). For these mismatch regions, BAF is calculated as:

$$BAF(SNP_i) = \frac{SNP_i\ coverage\ on\ allele_A / SNP_i\ coverage\ on\ allele_B}{Total\ coverage\ of\ SNP_i}$$

4. Integration: At this step, LR and BAF values for the MHC locus and the regions adjacent to this locus (regions on chromosome 6 before and after MHC locus) are both available. MHCnvex then integrates these values to generate a continuous genomic sequence with their corresponding LR and BAF values. The range of the adjacent region for integration can be set by the user, with a default range being at the whole short arm of chromosome 6.

5. Segmentation: To segment the integrated MHC locus and flanking regions, a custom algorithm is employed based on Circular Binary Segmentation (CBS) to jointly analyze integrated log coverage ratios and BAFs directed toward identifying breakpoints in the tumor's genome. The algorithm proceeds in two stages: first, independent segmentation of LR and mirrored BAF using CBS with parameters alpha = 0.01 and trim = 0.025, followed by collection of all candidate breakpoints. In the second step, the resulting segmentation track is iteratively refined by merging segments that exhibit similar LR and BAF values, have short lengths, are located in blacklisted regions, and exhibit high variation in coverage among the whole cohort of normal samples. This approach is specifically designed to achieve high sensitivity in detecting breakpoints by initially collecting all possible breakpoints and then selectively removing those with low confidence levels.

6. Model fitting: MHCnvex performs the copy number state inference on the resulting segmented copy-number profiles, Using a custom model most similar to the mathematical formalism of ABSOLUTE [10] and PureCN [11]. Briefly, the algorithm uses the log-ratios (of

bins) and BAFs of individual SNPs and assigns them to their corresponding segments. Next, the likelihood of different possible copy numbers (C, ranging from 0 to 18) are calculated using a likelihood model. The likelihood model is based on the assumption that LR values are distributed normally around a segment described below:

$$r_i \sim N(log2\frac{Purity * C + (i - Purity) * 2}{Ploidy * Purity + (1 - Purity) * 2}, \sigma_i)$$

where r_i is the average variance of lr values, distributed around a segment. To calculate one likelihood value for each segment, the likelihood of all the bins on a segment are summed up together. Finally, for each segment, the C that yields the highest likelihood is picked.

After inferring the C using the LR values for each bin, MHCnvex uses BAFs in the equation below to determine the expected allele frequency for each SNP:

$$E(AF_i) = \frac{(Purity * M_i) + (1 - Purity)}{Purity * C_i + (1 - Purity) * 2}$$

Where Mi is the number of alleles harboring a specific SNP and Ci is the total copy number of the SNP locus. After inferring C through Eq. 3 and both 'theoretical expected' (Eq. 4) and observed value for AF (through sequencing) are determined, then a likelihood value can be alculated for different values of Mi according to this equation:

$$L(AF_i|Purity, C_i, M_i, DP_i) = Beta(E[AF_i]|DP_i * AF_i + 1, (1 - AF_i) + 1)$$

where DP_i is the total depth for SNP_i. By maximizing $L(AF_i|Purity, C_i, M_i, DP_i)$ which is parameterized by M, the multiplicity of each SNP and therefore number of each haplotype can be called.

7. Calling LOH: After having the allele specific copy number for each locus of the genome, MHCnvex can call the LOH for segments which have a minor copy number of 0. MHCnvex also returns the copy number of each HLA haplotype for genes in HLA class 1 and class 2.

5.4.3 WGS analysis

For WGS analysis, first MHCnvex calculates the BAF value of each MHC gene as mentioned above using personalized reference coverages. For the LR calculation in the WGS setting, MHCnvex uses larger bins compared to the WXS setting to ensure there are enough reads per each bin for accurate LR calculation. In the WGS setting, the minimum length of a bin for a reliable estimation of LR value exceeds the length of MHC genes. Hence, in order to get reliable LR values, it is not possible to only use reads mapped to personal references of MHC genes. To overcome this problem, MHCnvex follows the same approach mentioned above to calculate absolute copy numbers for regions adjacent to each gene (LR and BAF values for MHC genes are not used yet). If both sides of a gene had the same copy number value, that value is also assigned to the gene. If they have different copy numbers, then the side with mirrored BAF closest to the mirrored BAF of MHC gene is selected and its copy number is then assigned to the MHC gene. After having the absolute copy number of each gene, the calculated BAF value of the gene itself is used to estimate the allelic copy numbers.

5.4.4 Haplotyping and alignment

In this study, the MHC haplotyping is performed using Hapster, a tool that generates personalized reference sequences for each patient based on their unique MHC haplotypes. Reads are then mapped to these personalized references for both the tumor and its matching normal samples.

5.5 References

1. Maleno I, López-Nevot MA, Cabrera T, Salinero J, Garrido F. Multiple mechanisms generate HLA class I altered phenotypes in laryngeal carcinomas: high frequency of HLA haplotype loss associated with loss of heterozygosity in chromosome region 6p21. Cancer Immunol Immunother. 2002;51: 389–396.

2. Thibodeau J, Bourgeois-Daigneault M-C, Lapointe R. Targeting the MHC Class II antigen presentation pathway in cancer immunotherapy. Oncoimmunology. 2012;1: 908–916.
3. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, et al. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell. 2017;171: 1259–1271.e11.
4. X. D, S. E, C. D, A. J, A. L, A. N, et al. Somatic HLA Class I Loss Is a Widespread Mechanism of Immune Evasion Which Refi nes the Use of Tumor Mutational Burden as a Biomarker of Checkpoint Inhibitor Response. doi:10.1158/2159-8290.CD-20-0672
5. Pyke RM, Mellacheruvu D, Dea S, Abbott CW, Zhang SV, McDaniel L, et al. Validated machine learning algorithm with sub-clonal sensitivity reveals widespread pan-cancer human leukocyte antigen loss of heterozygosity. bioRxiv. 2021. p. 2021.05.20.445052. doi:10.1101/2021.05.20.445052
6. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PLoS One. 2011;6: e14643.
7. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30: 3310–3316.
8. Mumphrey MB, Hosseini N, Parolia A, Geng J, Zou W, Raghavan M, et al. Distinct mutational processes shape selection of MHC class I and class II mutations across primary and metastatic tumors. bioRxiv. 2023. p. 2023.01.22.523447. doi:10.1101/2023.01.22.523447
9. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34: 867–868.
10. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30: 413–421.
11. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11: 13.

Chapter 6 Concluding Remarks

Despite the importance of genomic instability and its potential for being utilized to improve cancer patients' care, there aren't many methods out there to measure it. Furthermore, although it has been studied, the translational benefits of this phenomenon are yet to be explored. In this thesis, I have presented two novel approaches to measure the CIN in tumors using NGS data. I have also incorporated different methods of detecting CIN to identify prognostic and predictive biomarkers in cancer. This highlights the potential of studying CIN in the context of cancer and its utility in the clinical context.

In chapter 2 of this thesis, we mentioned the high variance in the observed CIN of LUAD. We showed that previous measures of CIN, such as wGII are unable to reflect the true underlying CIN and hence, there is a need for new methods to measure CIN, which are more reflective of the underlying instability. We introduced two new measures, Genome Contiguity index and Breakage Intensity and Clustering. Both of these methods present a novel way of looking at CIN, by focusing on the positional aspects of breakpoints in the genome, rather than the portion of genome altered (which is the way wGII measures CIN). This approach showed a great improvement in capturing the underlying CIN, and enabled us to not only present it as a new prognostic measure, but also utilizing it to discover a new biomarker in LUAD, which reflects the instability of LUAD tumors. Furthermore, not only this biomarker showed to be prognostic in LUAD, we observed it to be higher in a more aggressive type of non small cell lung cancer, *i.e.,* Lung squamous cell carcinoma. These developed tools and observations had initiated multiple follow-up studies to investigate the

utility of our methods in other types of cancer, and to establish the pancancer value of our identified biomarker.

In chapter 3 of this thesis, we studied chromosomal instability in the context of HGSOC. HGSOC is known as one of the unstable tumors, due to frequent alterations in BRCA1 and BRCA2 genes. These genes are involved in the process of DNA repair and particularly in the homologous recombination pathway. Despite over 30 years of research on this disease, there are very few biomarkers to predict the patient response to the platinum based treatments, part of the first-line therapy for HGSOC patients. We first confirmed that BRCA1/2 mutations are indeed predictive of response in HGSOC. Furthermore, we showed that measures of CIN in HRD patients, previously used as predictive biomarkers of response to PARPi treatment, are also predictive of response to platinum based treatment as well. In addition to these biomarkers, we observed the strongest association of response is with the LOH of chromosome 17. Chromosome 17 harbors some of the major tumor suppressor genes, such as *BRCA1*, *TP53*, *NF1* and *CDK12*. We aimed to understand the underlying mechanisms for LOH of chromosome 17 being favorable for treatment. We discovered in samples without chr17-LOH, target pathways of *TP53* are more active and therefore, the activity of TP53 is related to the treatment response. Further research is needed to find the exact mechanisms underlying this phenomenon, but given that almost all the HGSOC tumors harbor a TP53 mutation, we can hypothesize that the presence of a functional copy of TP53 can determine the response to the treatment to some extent.

In chapter 4, we investigated the downstream effect of genome instability, particularly the resulting genetic heterogeneity of chromosomal instability, in mCRPC. Genetic heterogeneity of primary prostate cancer has been established previously through multiple studies. One of the main aspects of genetic heterogeneity in primary prostate cancer is the multi-foci nature of this cancer.

This phenomenon raises one question: does metastatic prostate cancer, the most lethal form of this cancer, arise from one focus or it can be due to metastasis of multiple foci. Over the past few year, it has been established that metastatic tumors within a patient share a common ancestor and hence there should be lower heterogeneity in metastatic prostate cancer. So in our study, we investigated the heterogeneity of metastatic prostate cancer both in terms of overall heterogeneity and heterogeneity of pathogenic alterations. We observed a gross amount of heterogeneity in genetic alterations both at SNV and CNV levels. Despite this heterogeneity, we observed that pathogenic alterations in drivers of prostate cancer were significantly more homogeneous, which points toward their early emergence. These observations can have clinical implications for target design in immunotherapy and other types of treatments.

Furthermore, we investigated the effect of multi-sampling on detecting alterations in patients. Given the previously observed high intra-patient heterogeneity, we expected one sample to not be representative of a generics of a patient's cancer. We were able to quantify the effect of this heterogeneity and show how much improvement we can get by increasing the number of samples from patients' metastatic lesions. These results will contribute to the treatment design and patient care for patients with mCRPC.

In chapter 5, I presented a tool to study chromosomal instability for HLA locus, the most polymorphic region of the genome. Given the high importance of the HLA gene in determining patients' response to immunotherapy treatments such as PD/L-1 blockers, we aimed to overcome the challenge of polymorphism in this region. Polymorphic regions are hard to study with current CNV analysis tools, as they assume that all the reads can be aligned to the standard human genome reference. This is not the case for HLA regions and hence we incorporated a dynamic reference selection algorithm to find the most suitable reference for this region and then performed the CNV

analysis. We designed our tool to be versatile and able to work with different types of sequencing such as WGS,WXS, and for utility in clinical settings, targeted sequencing.

In conclusion, in this thesis we showed the utility of studying genomic instability in finding prognostic and predictive biomarkers. Genomic instability is a powerful way of looking at genomics of a patient, because rather than a single event, it is consistent of multiple CNV or mutational processes, forming a phenotype for the underlying processes of the tumor. Furthermore, it can inform us about the possible vulnerabilities of the tumor, given that genomic instability arises from defects in normal cell processes. We saw an example of that in the second chapter, where HRD was both driving the tumors, and could be used as a biomarker for platinum based treatment. To advance this field, we introduced new measurements of instability in tumors using NGS data. Previous measures, such as wGII proved to have a very low resolution, hence, failing to capture the true degree of instability in cancers with heterogeneous degree of CNV events (such as LUAD). We introduced two new measures, one based on the length of the segment distribution and another based on the distribution and intensity of breakpoints. These measures, although more complicated, were able to capture the differences between a stable tumors with few large CNVs versus a cancer with many focal CNV, something that wGII fails to do.

I am hoping these results would shine light on the utility of genomic instability for biomarker design and precision medicine. With the advancement of genomics in recent year and completion of projects such as TCGA, PCAWG, MSK-IMPACT etc, most of the knowledge from the ‘direct’genomic analysis (i.e, looking at somatic mutations, CNVs, SVs etc) has been extracted and the field needs to move toward the study of indirect consequences of genomic alterations such as the study of genomic instability and heterogeneity among other things.