Virus Induced Cell-Cell Fusion as a Driver of Metastasis

G Maria Jakobsdottir

Merton College



A thesis presented for the degree of

Doctor of Philosophy

Medical Sciences Doctoral Training Centre

University of Oxford

Hilary 2021

Declaration

Declaration

I confirm that this thesis is my own work, and that the contributions of others to this work are acknowledged where appropriate.

Abstract

Viral infection is thought to be responsible for 15% of all cancers and has previously been linked to cancer progression through the virus induced cell fusion theory of metastasis. This theory proposes that fusion of different cells may endow the resulting hybrid cells with metastatic properties. I developed a number of measures that I applied to WGS data from 6,337 tumours to identify tumours in which cell fusion may have occurred. Based on these measures I identified 15 primary tumours, and 31 metastatic tumours with evidence of cell fusion, and further characterised their genomes. I also investigated cell fusion *in vitro* and generated a number of cell lines which underwent either virus induced cell fusion, or became tetraploid after a failure of cytokinesis, to study changes in karytoype and genome stability over time, and to elucidate whether differences could be observed resulting from the mechanism by which the genomes were duplicated. While the scope of this work was limited by experimental constraints, the conclusions were consistent with previous literature suggesting that whole genome duplication by any means results in an initial increase in heterogeneity, followed by adoption of a stable karyotype over time.

Viral infection has also been associated with activation of several members of the Apolipoprotein B mRNA editing enzyme catalytic polypeptide 3 (APOBEC3) family of cytidine deaminases that act on cytosines in single stranded DNA, in specific cellular contexts including the developing immune system and as a response to invading viral pathogens. The action of these enzymes on DNA leaves specific, well defined, mutational signatures characterised by C to T and C to G transitions in particular contexts (single base substitution signatures known as SBS2 and SBS13, attributed to the activity of APOBEC3A and APOBEC3B). SBS2 and SBS13 mutations are observed in most cancer types, and have previously been associated with sites of chromosomal rearrangements. In this thesis I show that the presence and number of SBS2 and SBS13 mutations in the genome is associated with increasing genomic instability, and I introduce novel measures of genomic instability in whole

genome sequencing (WGS) data. The relationship between viral infection, genomic instability, and APOBEC3 activity was also investigated but results remain inconclusive.

COVID-19 Impact Statement

In March 2020 the Wellcome Centre for Human Genetics was closed, with the exception of essential COVID-19 related research. For my project, this meant suddenly halting several ongoing cell culture based experiments, and precluded important follow-up experiments from being conducted.

In August 2020 I returned to limited on-site working under COVID restrictions. After consulting with my supervisors, we decided that I would focus my efforts on finishing essential experiments that had been started before the lockdown due to the limitations to on-site working, as well as time constraints with regards to the completion of this DPhil.

Were it not for the pandemic, the following three experiments or areas of research would have been conducted or expanded on:

(1) A proteomics screen for DNA damage responses in HCT116 cells infected with pseudoviral particles. Discussed further in Sections 3.5 and 6.2, analysis of proteomics data would have added an important extra dimension to my data and would have provided answers to some of the outstanding questions in Chapter 3. Plans were in place to conduct a mass spectrometry screen of protein data generated from an experiment analogous to the RNAseq experiment described in Chapter 3.

(2) Scale up of cell fusion and isolation of single cell clones. Discussed in section 5.5, in order to generate the cell lines required to address the questions posed in Chapter 5, my previous cell fusion and single cell isolation experiments had to be repeated and scaled up. This was not possible due to the time commitment required for these experiments. In lieu of this, I conducted a limited analysis of previously isolated clones.

(3) Further follow up of potential cell fusions identified in whole genome sequencing data. Detailed follow up and characterisation of tumours identified as potential cell fusions from WGS data in Chapter 4 would likely not have been feasible without the pandemic. However, under normal circumstances I would have concentrated more

of my efforts on contacting the organisations responsible for these samples to assess the potential for further follow-up and whether any other distinguishing features had been identified.

Acknowledgements

I would like to start by thanking my supervisors, Cath Green, David Wedge, and Sergi Padilla-Parra for taking on this unconventional project, and for seeing it through to completion in these unconventional times. Thank you for giving me the freedom to explore my own ideas while also providing support and guidance along the way. It has been a privilege to be part of the friendly and collaborative atmospheres that you have created both within and between your different labs.

In the Wedge group I would particularly like to thank Anna, Andreas, Kasia, and Naser for their help and guidance with the bioinformatics aspects of this project. I would also like to thank Maro and Irene from the Padilla-Parra group, Maciej in the Green group, James from the Cellular Imaging core, Ruddy from the Flow Cytometry Facility, and Cathy and Jordan in Ross Chapman's group for sharing their wet lab expertise with me. I also owe a massive debt of gratitude to Daniela, Julia, and Paulina at the Chromosome Dynamics core for the mammoth amount of work that they have done for this project. Finally, I would like to thank all of my colleagues, past and present, who took the time to try to understand a project that was outside their field of work, and who approached this project with intrigue, creativity, and an open mind.

I would also like to thank my friends in science, Alice S and Sotiris, for inspiring me to do a PhD. Thanks also to Maro, Mary, and Dalia, for inspiring me to finish that PhD and reminding me why I stay in science. I also owe a huge thank you to my friends outside science, who have been there to celebrate the good times and supported me through the bad times. In particular, thanks to Cherlyn, Alice L, and Ollie for providing some much-needed levity through thesis writing and lockdown 3.0. I cannot express how much your friendly faces on a screen and friendly words on the phone have meant to me. I also owe a thank you to my housemates, for weathering three lockdowns together.

I would like to extend a great big thank you to my partner, Ali. I'm not going

to say "I couldn't have done it without you", but I certainly wouldn't have wanted to. Thank you for the walks and weekend breakfasts. Thank you for reminding me to take breaks and to go outside every now and then. And thank you for always bringing a healthy dose of perspective to the table. Thank you for your continued love and support, it has been a joy and a privilege getting to spend this time with you!

Last, but not least, I would like to thank my family. Thank you to my parents, Jakob and Fjóla, for supporting me in all of my endeavours, and for teaching me to value my education. Thank you to my grandparents, Ardís, Guðrún, and Sigurður for teaching me the value of hard work. And thanks to my siblings, Siggi and Karen, for being my friends. Takk fyrir stuðninginn í gengnum tíðina. Ég hefði aldrei komist hingað né klárað þetta án ykkar.

Funding

This work was supported by the Wellcome Trust (Grant no. 203852/Z/16/A), and the University of Oxford Medical Sciences Doctoral Training Centre.

The research was also supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z with funding from the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Contents

Contents **ix**

List of Figures **1**

List of Tables **9**

List of Equations **12**

1 Introduction **14**

1.1 Whole Genome Duplication 15

1.1.1 The Effects of Whole Genome Duplication 15

1.1.1.1 Effects on Gene Expression 15

1.1.1.2 Genomic Instability 16

1.1.1.3 Tumour Heterogeneity and Evolution 18

1.1.1.4 Resistance to Therapy 19

1.1.1.5 Acquisition of Metastatic Properties 20

1.1.2 Paths to Polyploidy 21

1.1.2.1 Errors During the Cell Cycle 21

1.1.2.2 Cell Fusion 23

1.2 Cell Fusion 23

1.2.1 Mechanisms of Cell Fusion 23

1.2.1.1 General Principles and Mechanisms 23

1.2.1.2 Virus-Induced Cell Fusion 25

1.2.2 The Effects of Cell Fusion 30

1.2.2.1 Altered Transcription 30
1.2.2.2 Tumour Heterogeneity and Genomic Instability . . . 34
1.2.2.3 Resistance to Therapy 36
1.2.2.4 Acquisition of Metastatic Properties 36
1.2.2.5 Cell Fusion Theory of Metastasis 37
1.3 Cell Fusion in Cancer 37
1.3.1 Detection of Cell Fusions 39
1.3.1.1 Fluorescent Markers 39
1.3.1.2 Premature Chromosome Condensation 39
1.3.1.3 Gene Expression Profile 40
1.3.1.4 Cell Surface Markers 40
1.3.1.5 Presence of Y Chromosomes in Male to Female Bone Marrow Transplant Recipients 41
1.3.1.6 STR Profiling 42
1.3.2 Evidence of Cell Fusion in Human Cancers 42
1.3.3 Whole Genome Sequencing 44
1.4 APOBEC3s 45
1.4.1 Introduction 45
1.4.2 APOBEC3A 46
1.4.3 APOBEC3B 46
1.4.4 APOBEC3G 47
1.4.5 APOBEC Activity in Response to Viral Infections 48
1.4.6 Processing of Deaminated Cytosines 48
1.5 APOBEC3s in Cancer 50
1.5.1 The APOBEC3 Mutational Signatures 50
1.5.2 Role of APOBEC3s in Different Cancer Types 53
1.5.3 DNA Damage Responses Following APOBEC3 Activation . . 55
1.5.4 APOBEC3 Activation Following DNA Damage 56

1.5.5 Why not both? . . . 58
1.6 Aims of the Thesis . . . 58

2 APOBEC3 mutational signatures are associated with extensive and diverse genomic instability across multiple tumour types 61

2.1 Introduction . . . 61
2.2 Materials and Methods . . . 62
2.2.1 Data . . . 62
2.2.2 Mixed Effects Models . . . 63
2.2.3 Survival Analysis . . . 64
2.2.4 Genomic Instability . . . 64
2.2.5 Volcano Plot . . . 65
2.3 Models . . . 66
2.3.1 Mutation Burden . . . 66
2.3.2 Genomic Instability . . . 68
2.3.3 Genomic Instability and TP53 Mutation . . . 71
2.3.4 Virus Models . . . 75
2.3.5 CoxME Models . . . 77
2.4 Results . . . 79
2.4.1 Number of APOBEC3 Mutations Correlates With Total Mutation Burden . . . 79
2.4.2 Presence of APOBEC3 Mutations is Associated With Increased Genomic Instability . . . 82
2.4.3 The Number of APOBEC3 Mutations Predicts the Level of Genomic Instability Across Multiple Tumour Types . . . 84
2.4.4 Both Presence of APOBEC3 Mutations and TP53 Mutation Affect Genome Stability . . . 88
2.4.5 The Number of Non-Kataegis APOBEC3 Mutations is Associated With Increased Genomic Instability . . . 93

2.4.6 Associations Between Presence of Viral DNA and APOBEC3 Mutations . . . 96
2.5 Discussion . . . 99

3 Effect of in vitro infection with pseudoviral particles on APOBEC3A expression, DNA damage, and genomic instability 105

3.1 Introduction . . . 105
3.2 Materials . . . 106
3.2.1 Cell Cultures . . . 106
3.2.2 Pseudoviral Particles . . . 106
3.2.3 Antibodies Used for Immunofluorescence . . . 107
3.2.4 APOBEC Over-Expression Plasmids . . . 107
3.2.5 Data Availability . . . 107
3.3 Methods . . . 108
3.3.1 *In vitro* Infection and Fixation for Immunofluorescence . . . 108
3.3.2 Immunofluorescence . . . 108
3.3.3 Microscopy . . . 109
3.3.4 Analysis of Immunofluorescent Images . . . 109
3.3.5 RNAseq Set Up and Library Preparation . . . 109
3.3.6 RNAseq Data Analysis . . . 110
3.3.7 Transfection and Infection of Cells for Whole Genome Sequencing . . . 111
3.3.8 Single Cell Sorting Using Flow Cytometry . . . 111
3.3.9 DNA Extraction for Whole Genome Sequencing . . . 111
3.3.10 Analysis of Whole Genome Sequencing Data . . . 112
3.3.10.1 Pre-processing, Quality Control, and Variant Calling 112
3.3.10.2 Mutational Signatures . . . 113
3.3.10.3 Structural Variants . . . 113
3.3.10.4 Copy Number Variation . . . 113

3.4 Results . . . 114
3.4.1 In Vitro Infection of Cells With Pseudotyped HIV Particles Increases Nuclear APOBEC3A Levels and DNA Damage . . . 114
3.4.2 RNAseq Analysis of HCT116 Cells Infected With Pseudotyped HIV Over Time . . . 114
3.4.2.1 Differential Expression Analysis . . . 117
3.4.2.2 APOBEC3 Expression . . . 119
3.4.2.3 Expression of DNA Damage Response Genes . . . 122
3.4.3 Effect of *in vitro* Infection With Pseudoviral Particles and Transfection With A3A on Genome Stability Detected by Whole Genome Sequencing . . . 123
3.4.3.1 SNVs and INDELs . . . 123
3.4.3.2 Structural Variants . . . 127
3.4.3.3 Copy Number . . . 131
3.5 Discussion . . . 131

4 Identification of virus-induced cell-cell fusions in whole genome sequencing data 139

4.1 Introduction . . . 139
4.2 Materials . . . 140
4.2.1 PCAWG Data Set . . . 140
4.2.2 Hartwig Medical Foundation (HMF) Data Set . . . 140
4.3 Theory . . . 141
4.3.1 Variation in Mutation Density . . . 141
4.3.2 Proportion of Mutations With Multiplicity 1 . . . 143
4.3.3 Average Ploidy . . . 146
4.3.4 Presence of Viral DNA . . . 146
4.3.5 Presence of Mutational Signatures . . . 147
4.3.6 Timing Whole Genome Duplication . . . 147

4.3.7 Detecting Premature Chromosome Condensation in Sequencing Data . . . 148
4.4 Methods . . . 149
4.4.1 Processing of Samples in the Hartwig Medical Foundation Data Set . . . 149
4.4.2 Quantifying Variation in Mutation Density . . . 149
4.4.3 Calculating the Proportion of Mutations with Multiplicity 1 (PM_1) . . . 150
4.4.4 Average Ploidy . . . 150
4.4.5 Timing Whole Genome Duplication . . . 151
4.4.6 Identifying Samples Potentially Containing Fusions . . . 151
4.4.6.1 3^{rd} Quartile Cutoff . . . 151
4.4.6.2 Standard Deviation Based Cutoffs . . . 152
4.4.6.3 Precision and Recall Based Cutoffs . . . 152
4.4.6.4 Introducing Cutoff Thresholds by Eye . . . 152
4.5 Results . . . 154
4.5.1 Identifying Cell-Cell Fusions in the PCAWG Data Set . . . 154
4.5.1.1 Identifying Samples Potentially Containing Fusions . 154
4.5.1.2 Summary of Samples Potentially Containing Fusions 154
4.5.1.3 Timing of Fusions or Whole Genome Duplications . . 160
4.5.1.4 Premature Chromosome Condensation . . . 161
4.5.1.5 Genome Stability . . . 167
4.5.1.6 Mutational Signatures . . . 171
4.5.2 Identifying Cell-Cell Fusions in the Hartwig Medical Foundation Data Set of Metastatic Samples . . . 173
4.5.2.1 Identifying Samples Potentially Containing Fusions . 173
4.5.2.2 Summary of Samples Potentially Containing Fusions 175
4.5.2.3 Timing of Fusions or Whole Genome Duplications . . 176

4.5.2.4 Premature Chromosome Condensation 176
4.5.2.5 Genome Stability 182
4.5.2.6 Mutational Signatures 184
4.6 Discussion 187

5 In vitro induced cell fusion and its effect on genome stability 196
5.1 Introduction 196
5.2 Materials 198
5.2.1 Cell Cultures 198
5.2.2 Fluorescent Cytoplasmic Dyes 198
5.2.3 Pseudoviral Particles 198
5.3 Methods 198
5.3.1 Interruption of Cytokinesis Using Cytochalasin B 198
5.3.2 PEG-Induced Cell Fusion 198
5.3.3 Virus-Induced Cell Fusion 199
5.3.4 Isolation of Fused Cells Using Flow Cytometry 199
5.3.5 Metaphase Spread 200
5.3.6 Microscopic Analysis of Metaphase Spreads 201
5.3.7 Multicolor Fluorescent In Situ Hybridisation (FISH) 201
5.3.8 Analysis of Karyotype Over Time 201
5.3.9 Karyotype Stability 202
5.4 Results 202
5.4.1 Effects of Cell Fusion and Whole Genome Duplication on Karyotype Over Time 202
5.4.2 Changes in Genome Stability 207
5.5 Discussion 209

6 Conclusion 215
6.1 Results Summary 215

6.1.1 Chapter 2 . . . 215
6.1.2 Chapter 3 . . . 216
6.1.3 Chapter 4 . . . 216
6.1.4 Chapter 5 . . . 217
6.2 Future Work . . . 217
6.2.1 APOBECs, Viruses, and Genomic Instability . . . 217
6.2.1.1 Viral Infection and APOBEC3-Induced Genomic Instability in Cancer Genomes . . . 217
6.2.1.2 APOBEC3 as a Causative Agent of Genomic Instability . . . 218
6.2.1.3 A3A and DNA Damage in Cell Lines . . . 219
6.2.2 Cell Fusion in WGS Data . . . 220
6.2.2.1 Validation of Tumour Samples Identified as Potential Cell Fusions . . . 220
6.2.2.2 Validation of Methods Used to Identify Potential Cell Fusions in WGS Data . . . 221
6.2.2.3 Expanding the Analysis to Larger Data Sets . . . 222
6.2.3 Cell Fusion Experiments . . . 223
6.2.3.1 Further Fusion Experiments and Sequencing . . . 223
6.3 Concluding Remarks . . . 223

A Appendix: List of publications 225

B Appendix: Materials and Methods 226

B.1 Processing of samples in the HMF data set . . . 226
B.1.1 Pseudo-Battenberg files . . . 226
B.1.2 Pseudo-Locus files . . . 227
B.1.3 Pseudo-Allele Frequency files . . . 227
B.1.4 Pseudo-Cellularity files . . . 227

B.1.5 Gender Information 228
B.1.6 Filtering Gaps in the Reference Genome 228
B.1.7 Merging Adjacent Segments With the Same Copy Number . . 228

C Appendix: Results **230**
C.1 Differential Expression Analysis 230
C.1.1 All Genes 230
C.1.2 Hallmark DNA Repair Genes 243
C.2 Mutational signatures 248
C.2.1 PCAWG Tumour Types 248
C.2.2 Selected PCAWG Samples 259
C.2.3 Hartwig Tumour Types 276
C.2.4 Selected Hartwig Samples 287
C.3 Representative Karyotypes 318
C.3.1 Karyotypes at Harvest 0 318
C.3.2 Karyotypes at Harvest 1 321
C.3.3 Karyotypes at Harvest 2 324
C.3.4 Karyotypes at Harvest 3 327

D Appendix: Abbreviations **330**
D.1 Abbreviations used in text 330
D.2 Tumour Type Abbreviations 338

References **340**

List of Figures

1.1 Causes of Cytokinesis Failure 22
1.2 Stages of membrane fusion 24
1.3 Outcomes of cell fusion 31
1.4 Cell fusion theory of metastasis 38
1.5 HIV-1 restriction by APOBEC3 enzymes 49
1.6 DNA deamination model for mutation in cancer. 51
1.7 SBS2 Trinucleotide Context 52
1.8 SBS13 Trinucleotide Context 52
1.9 Schematic Illustration of Kataegis 57

2.1 Correlation between number of SBS2 and SBS13 mutations and non-SBS2 and SBS13 mutations. 80
2.2 The effect of SBS2 and SBS13 presence on genomic instability 83
2.3 The effect of SBS2 and SBS13 presence on genomic instability by tumour type 85
2.4 Overall survival stratified by presence of SBS2 and SBS13 mutations and p53 mutation status 93

3.1 HCT116 immunofluorescence 115
3.2 RPE-1 immunofluorescence 116
3.3 Principal component analysis of HIV infected HCT116 cells over time 117
3.4 Volcano plots for RNAseq of HIV infected HCT116 cells 118

3.5 Heatmap of most variable genes in HIV infected HCT116 cells over time . . . 120
3.6 Top enriched Gene Ontology terms . . . 121
3.7 Number of transcripts by APOBEC3 . . . 122
3.8 Volcano plot of DDR genes at each time point . . . 124
3.9 Heatmap of DDR expression . . . 125
3.10 Overview of whole genome sequencing of clones . . . 126
3.11 Clonal INDELs and SNVs detected in whole genome sequencing of A3A transfected and virally infected cells . . . 128
3.12 APOBEC3 mutations and mutational signatures in A3A transfected and virally infected cells . . . 129
3.13 Structural variants in A3A transfected and virally infected cells . . . 130
3.14 CNVs in A3A transfected and virally infected cels . . . 132

4.1 Differences in mutation density following loss of DNA . . . 142
4.2 Coefficient of variation of the mutation density . . . 144
4.3 Multiplicity of mutations and whole genome duplication . . . 145
4.4 Overview of fusion measures calculated for tumours in the PCAWG data set . . . 153
4.5 Thresholds applied to PCAWG fusion measures . . . 155
4.6 Overlap between sample sets identified by applying different thresholds to the PCAWG fusion measures . . . 156
4.7 Summary of samples identified as hybrids in the PCAWG data set . . . 157
4.8 Proportion of tetraploid PCAWG samples passing threshold, by tumour type . . . 158
4.9 Proportion of all PCAWG samples passing threshold, by tumour type 159
4.10 Proporiton of PCAWG samples that are tetraploid vs proportion of samples that pass the cutoff, by tumour type . . . 160
4.11 Timing WGD in the PCAWG data set . . . 161

4.12 Timing WGD in the PCAWG data set by tumour type 162
4.13 Chromothripsis and structural variants in potential cell fusions 163
4.14 Number of chromothripsis events by tumour type 164
4.15 Number of chromosomes affected by chromothripsis by tumour type . 165
4.16 Number of structural variant events by tumour type 166
4.17 Number of CN segments, INDELs, and PGA in selected PCAWG samples 167
4.18 Number of CN segments by tumour type - PCAWG 168
4.19 PGA by tumour type - PCAWG 169
4.20 INDELs by tumour type - PCAWG 170
4.21 Overview of fusion measures calculated for tumours in the HMF data set 173
4.22 Thresholds applied to the HMF fusion measures 174
4.23 Overlap between sample sets identified by applying different thresholds to the HMFfusion meausres 175
4.24 Summary of samples identified as hybrids in the HMF data set 177
4.25 Proportion of tetraploid HMF samples passing threshold, by tumour type 178
4.26 Proportion of all HMF samples passing threshold, by tumour type . . 178
4.27 Proportion of HMF samples that are tetraploid vs proportion of samples that pass the cutoff, by tumour type 179
4.28 Timing of WGD in the HMF data set 180
4.29 Timing WGD in the HMF data set by tumour type 181
4.30 Number of SVs, CN segements, and PGA HMF samples 182
4.31 Number of structural variants in WGD HMF samples, by tumour type 183
4.32 Hartwig PGA by tumour type 185
4.33 Hartwig CN segments by tumour type 186

5.1 Boxplot of total chromosome number over time 204

5.2 Ridgeline plot of total chromosome number over time 205
5.3 Chromosome counts for individual cells 206
5.4 Karyotype over time 208
5.5 Proportion of chromosomes altered by clone 210

C.1 Selected PCAWG BLCA-US mutational signatures 249
C.2 Selected PCAWG BOCA-UK mutational signatures 250
C.3 Selected PCAWG BRCA-AU mutational signatures 251
C.4 Selected PCAWG ESAD-UK mutational signatures 252
C.5 Selected PCAWG GACA-CN mutational signatures 253
C.6 Selected PCAWG MELA-AU mutational signatures 254
C.7 Selected PCAWG OV-AU mutational signatures 255
C.8 Selected PCAWG OV-US mutational signatures 256
C.9 Selected PCAWG PAEN-AU mutational signatures 257
C.10 Selected PCAWG PBCA-DE mutational signatures 258
C.11 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample BLCA - DO689 260
C.12 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample BOCA - DO52600 261
C.13 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample BRCA - DO6231 262
C.14 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample ESAD - DO50355 263
C.15 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample ESAD - DO50320 264
C.16 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample GACA - DO217822 265
C.17 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample MELA - DO220904 266

C.18 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample OV - DO46382 . . . 267
C.19 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample OV - DO28053 . . . 268
C.20 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample OV - DO29146 . . . 269
C.21 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample OV - DO28024 . . . 270
C.22 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample PAEN - DO51181 . . . 271
C.23 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample PBCA - DO35571 . . . 272
C.24 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample PBCA - DO35624 . . . 273
C.25 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample PBCA - DO35543 . . . 274
C.26 Mutational signatures in 1+1 and 2+1 segments in PCAWG sample DO36111 - PBCA . . . 275
C.27 Selected HMF-Bone mutational signatures . . . 277
C.28 Selected HMF-Breast mutational signatures . . . 278
C.29 Selected HMF-Colorectal mutational signatures . . . 279
C.30 Selected HMF-Lung mutational signatures . . . 280
C.31 Selected HMF-Lymphoid mutational signatures . . . 281
C.32 Selected HMF-Ovarian mutational signatures . . . 282
C.33 Selected HMF-Prostate mutational signatures . . . 283
C.34 Selected HMF-Testicular mutational signatures . . . 284
C.35 Selected HMF-CUP mutational signatures . . . 285
C.36 Selected HMF-Urinary mutational signatures . . . 286

C.37 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001044B 288
C.38 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001971A 289
C.39 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002775A 290
C.40 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002350A 291
C.41 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001041A 292
C.42 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002824A 293
C.43 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001145A 294
C.44 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001145B 295
C.45 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF000851B 296
C.46 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002841A 297
C.47 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF000567A 298
C.48 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001504A 299
C.49 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002569A 300
C.50 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001802A 301

C.51 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001498A . . . 302
C.52 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF004202A . . . 303
C.53 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF004159A . . . 304
C.54 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003844A . . . 305
C.55 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003158A . . . 306
C.56 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF000586A . . . 307
C.57 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003031A . . . 308
C.58 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003403A . . . 309
C.59 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003282A . . . 310
C.60 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002916A . . . 311
C.61 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF000945A . . . 312
C.62 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003747A . . . 313
C.63 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF002757A . . . 314
C.64 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF000227A . . . 315

C.65 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF001233A . . . 316
C.66 Mutational signatures in 1+1 and 2+1 segments in HMF sample HMF003110A . . . 317
C.67 M-B10 karyotype at harvest 0 . . . 318
C.68 C-F10 karyotype at harvest 0 . . . 318
C.69 C-F3 karyotype at harvest 0 . . . 319
C.70 C-G9 karyotype at harvest 0 . . . 319
C.71 V-F4 karyotype at harvest 0 . . . 320
C.72 M-B10 karyotype at harvest 1 . . . 321
C.73 C-F10 karyotype at harvest 1 . . . 321
C.74 C-F3 karyotype at harvest 1 . . . 322
C.75 C-G9 karyotype at harvest 1 . . . 322
C.76 V-F4 karyotype at harvest 1 . . . 323
C.77 M-B10 karyotype at harvest 2 . . . 324
C.78 C-F10 karyotype at harvest 2 . . . 324
C.79 C-F3 karyotype at harvest 2 . . . 325
C.80 C-G9 karyotype at harvest 2 . . . 325
C.81 V-F4 karyotype at harvest 2 . . . 326
C.82 M-B10 karyotype at harvest 3 . . . 327
C.83 C-F10 karyotype at harvest 3 . . . 327
C.84 C-F3 karyotype at harvest 3 . . . 328
C.85 C-G9 karyotype at harvest 3 . . . 328
C.86 V-F4 karyotype at harvest 3 . . . 329

List of Tables

1.1 Examples of fusogenic viruses that affect humans 26

2.1 Spearman correlation coefficients for the correlation between log number of APOBEC3 mutations and the log number of non-APOBEC3 mutations by tumour type 81

2.2 Mixed Effects Models for Predicting the Number of non-APOBEC3 mutations 82

2.3 Wilcoxon rank sum test, comparing the levels of each of the genomic instability measures between samples containing SBS2 and SBS13 mutations, and those not containing SBS2 and SBS13 mutations . . . 84

2.4 The number of samples belonging to each tumour type that either contain SBS2 and SBS13 mutations, or do not contain SBS2 and SBS13 mutations 86

2.5 Mixed effects models predicting the levels of six different measures of instability using age, the number of SBS2 and SBS13 mutations, accounting for the effects of tumour type as a random variable 87

2.6 Combined p-values using Fisher's combined probability test 89

2.7 Mixed Effects Models Predicting the levels of six different measures of instability using the log number of SBS2 and SBS13 mutations and TP53 mutation status, as well as accounting for the effects of tumour type as a random variable 90

2.8 Akaike Information Criteria (AIC) and p-values from ANOVAs comparing models with (equations 2.4-2.9) and without TP53 status (equations 2.10-2.15)for each GI measure. Full models are specified in Section 2.3. 91
2.9 Coxme model using only SBS2 and SBS13 presence to predict survival 92
2.10 CoxME model using SBS2 and SBS13 presence, p53 mutation status, and the interaction between them as predictors of survival 92
2.11 Mixed Effects Models Predicting the levels of six different measures of instability using the log number of SBS2 and SBS13 mutations excluding those attributed to kataegis, accounting for the effects of tumour type as a random variable 94
2.12 Mixed Effects Models Predicting the levels of six different measures of instability using the log number of SBS2 and SBS13 mutations excluding those attributed to kataegis, and TP53 mutation status, accounting for the effects of tumour type as a random variable 95
2.13 Mixed effects models using virus presence to predict number of SBS2 and SBS13 mutations 96
2.14 Mixed effects models using virus genome type to predict number of SBS2 and SBS13 mutations 97
2.15 Mixed effects models using virus genus to predict number of SBS2 and SBS13 mutations 98

3.1 Table of viruses used 106
3.2 Primary antibodies used for immunofluorescence 107
3.3 Secondary antibodies used for immunofluorescence 107

5.1 Modal chromosome numbers at each harvest 205

C.1 Differentially expressed genes 6h after infection 230
C.2 Differentially expressed genes 12h after infection 230

C.3 Differentially expressed genes 24h after infection 233

C.4 DDR genes and corresponding log2FC values and p-values 243

D.1 List of tumour type abbreviations used in PCAWG 339

List of Equations

2.1 Mixed effects model of the relationship between age and mutation burden 67

2.2 Mixed effects model of the relationship between APOBEC3 and mutation burden 67

2.3 Mixed effects model of the relationship between age, APOBEC3, and mutation burden 67

2.4 Mixed effects model of the relationship between APOBEC3 mutations and PGA 69

2.5 Mixed effects model of the relationship between APOBEC3 mutations and CNV segments 69

2.6 Mixed effects model of the relationship between APOBEC3 mutations and SVs 70

2.7 Mixed effects model of the relationship between APOBEC3 mutations and INDELs 70

2.8 Mixed effects model of the relationship between APOBEC3 mutations and ID8 INDELs 71

2.9 Mixed effects model of the relationship between APOBEC3 mutations and ID6 INDELs 71

2.10 Mixed effects model of the relationship between APOBEC3 mutations, TP53 mutation, and PGA 72

2.11 Mixed effects model of the relationship between APOBEC3 mutations, TP53 mutation, and CNV segments 73

2.12 Mixed effects model of the relationship between APOBEC3 mutations, TP53 mutation, and SVs 73
2.13 Mixed effects model of the relationship between APOBEC3 mutations, TP53 mutation, and INDELs 74
2.14 Mixed effects model of the relationship between APOBEC3 mutations, TP53 mutation, and ID8 INDELs 74
2.15 Mixed effects model of the relationship between APOBEC3 mutations, TP53 mutation, and ID6 INDELs 75
2.16 Mixed effects model of the relationship between presence of viral DNA and number of SBS2 and SBS13 mutations 76
2.17 Mixed effects model of the relationship between presence of different viral genome types and number of SBS2 and SBS13 mutations 77
2.18 Mixed effects model of the relationship between presence of specific viral genera and number of SBS2 and SBS13 mutations 77
2.19 CoxME model of the relationship between APOBEC3 mutations and survival 78
2.20 CoxME model of the relationship between APOBEC3 mutations, TP53 mutation, and survival 78
4.1 Timing for copy number 2+1 148
4.2 Timing for copy number 2+2 148
4.3 Timing for copy number 2+0 148
4.4 Coefficient of Variation 150
4.5 Calculating the average time of whole genome duplication in a tumour . 151
5.1 Equation for karyotype stability 202

Chapter 1

Introduction

Metastasis, or dissemination of a tumour from a primary site to a secondary site, is a hallmark of cancer, and is responsible for the majority of cancer related deaths for solid tumours (Hanahan and Weinberg, 2011; Dillekås et al., 2019). A common feature of the genomes of metastatic tumours, reviewed in Birkbak and McGranahan (2020) and Patel et al. (2020), is whole genome duplication (WGD). Genomes can become duplicated through a number of different processes, such as through failure of cytokinesis, or through cell fusion. It is not yet known whether the consequences of WGD caused by failure of cytokinesis differ from those of WGD caused by cell fusion. Further to this, it is also not known whether the effects of virus induced cell fusion differ from those of spontaneous cell fusion.

This chapter provides an introduction to the effects of WGD (Section 1.1) and cell fusion (Section 1.2) on the cancer genome, and discusses techniques that have previously been used to detect cell fusion (Section 1.3.1). Evidence for cell fusion occurring in human cancers is also summarised (Section 1.3.2). Finally, the APOBEC3 family of cytidine deaminases which form part of the host response to viral infection is introduced (Section 1.4), and its involvement in cancer is discussed.

1.1 Whole Genome Duplication

1.1.1 The Effects of Whole Genome Duplication

Whole genome duplication (WGD) is widely considered an important transformative event in the evolution of a cancer genome (Laughney et al., 2015; Watkins et al., 2020; López et al., 2020; Bielski et al., 2018). WGD has been estimated to be the second most common molecular aberration found in cancer genomes, after TP53 mutation, and is estimated to occur in 28.2-37% of primary tumours and 56% of metastatic tumours (Bielski et al., 2018; Zack et al., 2013; Priestley et al., 2019). It is thought that WGD may function as a stepping stone en route to aneuploidy, and that WGD can expedite diversification and evolution of the cancer genome, potentially facilitating the emergence of therapy resistance and metastatic properties. In this section I will discuss the effect of WGD on cancer genomes.

1.1.1.1 Effects on Gene Expression

A number of studies have sought to investigate the effect of WGD on gene expression, both *in vitro*, and by making use of large sets of RNA sequencing (RNAseq) data from studies of human cancer. *In vitro* studies of WGD in colorectal cancer cell lines (DLD-1 and RKO) have reported altered expression of genes associated with the cell cycle and mitosis, such as over-expression of RPM2 and MCM2, which are associated with DNA replication (Wangsa et al., 2018). These cells displayed an increase in expression of genes associated with energy metabolic pathways, as well as pathways associated with membrane biogenesis (Stingele et al., 2012). A separate study investigating WGD in HCT116 cells reported increased expression of factors that promote cell survival following stress, as well as reporting differential expression of factors involved in apoptosis inhibition and MAP kinases (Kuznetsova et al., 2015). Interestingly, analysis of cells derived from the colorectal cancer cell line HCT116 in which only a single chromosome was present in trisomy or tetra-

somy, reportedly showed a decrease in the expression of pathways associated with DNA repair, replication, transcription, and mRNA processing, suggesting that the effects of WGD on transcription may differ from the effects of single chromosome gains (Stingele et al., 2012). Similar results were observed for the untransformed epithelial cell line RPE-1 (Stingele et al., 2012). Wangsa et al. (2018) also report increased expression of oncogenes such as KRAS and myc in several WGD clones. The authors note that genes present on individual monosomic chromosomes are expressed at a similar level to their expression in cells disomic for the chromosome, with the exception of genes associated with the aforementioned pathways, which were found to be altered (Stingele et al., 2012). Analysis of the transcriptional effects of CNVs in human cancers has yielded similar conclusions (Bhattacharya et al., 2020). Bhattacharya et al. (2020) suggest that for most genes, particularly those involved in immune signalling, adaptive mechanisms such as methylation buffer the effects of CNVs on transcription, in contrast to genes involved in proliferation, which are sensitive to the effects of CNVs.

In concordance with the results of *in vitro* experiments, this suggests that WGD may specifically increase expression of genes involved in cell cycle progression and proliferation. One might expect this to translate to increased proliferation of tetraploid cells; however, several *in vitro* studies have demonstrated slower proliferation, cell cycle arrest, and replication errors in WGD cells; an effect that is attributed to replication stress and instability associated with the initial stages following a WGD event (Wangsa et al., 2018; Dewhurst et al., 2014; Kuznetsova et al., 2015).

1.1.1.2 Genomic Instability

Increased replication stress following WGD has been demonstrated in a number of cellular models. Wangsa et al. (2018) observed increased levels of the replication stress markers phospho-CHK1 and phospho-RPA32 in both WGD colorectal cancer cells and WGD non-transformed RPE-1 cells, with replication stress levels appearing

much higher in the cancer cells compared to the non-transformed cells. This was associated with S phase delay and subsequent growth impairment of the WGD cells (Wangsa et al., 2018). Galofré et al. (2020) also investigated the relationship between replication stress and WGD, demonstrating that administration of an ATR inhibitor reverses treatment resistance seen in WGD cells, as WGD cells are highly dependent on CHK1 due to their increased replication stress.

As a result of this increased replication stress, many studies have reported increases in genomic instability and errors during cell cycle progression in WGD cells. Replication stress is thought to cause genomic instability through premature centriole disengagement, which results in increased frequency of merotelic attachments and chromosome missegregation (Wilhelm et al., 2019). Chromosome missegregation by this mechanism has also been reported as a consequence of transient multipolar mitotic spindles caused by the presence of extra centrosomes following WGD (Ganem et al., 2009). Thus, increased levels of genomic instability (GI) are frequently observed following WGD.

Dewhurst et al. (2014) report increasing levels of GI in WGD HCT116 cells compared to their diploid counterparts, and found that GI increased with time spent in culture. Kuznetsova et al. (2015) reported similar results for HCT116 cells, but observed heterogeneity with respect to the induction of GI in WGD untransformed RPE-1 cells, some of which displayed increased GI whereas others did not. Higher levels of GI in WGD cells compared to diploid cells were also observed for derivatives of the DLD-1 and RKO colorectal cancer cell lines (Wangsa et al., 2018). All of the aforementioned studies also report increases in mitotic errors in WGD cells compared to their diploid counterparts, consistent with missegregation as the underlying mechanism (Kuznetsova et al., 2015; Dewhurst et al., 2014; Wangsa et al., 2018).

Further, increased numbers of structural variants have been observed both in *in vitro* models of WGD, as well as in genomes of human tumours that have undergone

WGD (Wangsa et al., 2018; Carter et al., 2012; Bakhoum et al., 2018), though in some cases normalising for the number of chromosomes removes this difference (Dewhurst et al., 2014).

In addition to causing chromosome missegregation and increased GI, it has been suggested that WGD mediates tolerance of these aberrations, in addition to deleterious mutations, increases in copy number alterations, and aneuploidy, which can lead to increased tumour heterogeneity and evolution of the cancer genome (Wangsa et al., 2018; López et al., 2020; Kuznetsova et al., 2015; Dewhurst et al., 2014; Galofré et al., 2020; Bielski et al., 2018; Laughney et al., 2015; Watkins et al., 2020).

1.1.1.3 Tumour Heterogeneity and Evolution

WGD is widely considered an important factor in promoting subclonal diversification of tumours; indeed, several *in vitro* studies have demonstrated increased variation in the karyotypes of WGD cells compared to their diploid counterparts (Dewhurst et al., 2014; Kuznetsova et al., 2015; Wangsa et al., 2018). Increased heterogeneity has also been observed in human tumour genomes, in which subclonal changes in copy number have been found to be enriched in WGD tumours, in addition to general increases in copy number alterations, most of which occur after the WGD event (Watkins et al., 2020; Carter et al., 2012; Bielski et al., 2018). Evolution of the genome of WGD cells has been reported *in vitro*, in that comparison of early and late passages of WGD cell lines revealed novel genomic alterations observed in late passage WGD cells compared to early passage WGD cells; a trend that was not observed in diploid cells (Wang et al., 2019; Dewhurst et al., 2014; Kuznetsova et al., 2015; Fujiwara et al., 2005). It has also been proposed that more subclonal selection occurs in WGD tumours, compared to tumours which have not been duplicated (Watkins et al., 2020).

After an initial period of diversification, it is thought that WGD tumours settle at a near triploid karyotype. This idea is supported both by modelling, and by

data from human cancers which suggests that tumours which have undergone WGD experience a loss of genetic material following this event, such that the average estimated ploidy of a WGD tumour settles around 2.75-3.3, in contrast to 1.99-2.1 for diploid tumours (Laughney et al., 2015; Zack et al., 2013; Bielski et al., 2018; Carter et al., 2012).

1.1.1.4 Resistance to Therapy

In addition to increasing genomic instability, altering transcription, and promoting intratumoural heterogeneity, WGD has also been implicated in the development of resistance to drug treatment. Kuznetsova et al. (2015) describe the emergence of multidrug resistance in their model of HCT116 and RPE-1 WGD. HCT116 based WGD clones were found to have acquired increased resistance to daunorubicin, doxorubicin, etoposide, cisplatin, docetaxel, paclitaxel, and vorinostat, whereas RPE1 based WGD clones showed increased resistance to 5-fluorouracil (5-FU), nutlin3a, pelitinib, daunorubicin, doxorubicin, and etoposide (Kuznetsova et al., 2015). The only compound to which HCT116 WGD clones were sensitised was 6-mercaptopurine (Kuznetsova et al., 2015). A similar multidrug resistance phenotype was described in three other WGD colorectal cancer cell lines (DLD-1, RKO, and SW837) by Galofré et al. (2020), who report increased resistance to oxaliplatin, FOLFOX, and 5-FU, even when concentrations of oxaliplatin and 5-FU were doubled to account for any effect that duplication of DNA content may have had. DLD-1 based WGD clones also had increased resistance to gemcitabine and irinotecan treatment, but had increased sensitivity to paclitaxel (Galofré et al., 2020). The authors also describe increased resistance to 5-FU and oxaliplatin in WGD RPE-1 cell lines, similar to the results reported by Kuznetsova et al. (2015). Three WGD colorectal cell lines, as well as WGD RPE-1 cells, were additionally found to have increased resistance to ionising radiation (Galofré et al., 2020).

1.1.1.5 Acquisition of Metastatic Properties

The observation that WGD is more common in metastatic cancers than primaries (Bielski et al., 2018; Bakhoum et al., 2018; Priestley et al., 2019), and the effects of WGD on heterogeneity and tolerance of GI has led to speculation that WGD may lead to acquisition of metastatic properties. WGD DLD-1 colorectal cancer cell lines display increased migratory capacity and wound healing, as well as invasive properties, especially in later passages, whereas these phenotypes are largely heterogeneous in WGD RKO colorectal cancer cell lines (Wangsa et al., 2018). Similarly, non-transformed RPE-1 cells gained the ability to grow in soft agar following WGD, suggesting that WGD had a transformative effect on these cells (Kuznetsova et al., 2015). In contrast, Fujiwara et al. (2005) did not observe anchorage independent growth in their WGD mouse mammary epithelial cells (MMECs), suggesting that cell type may be an important factor in governing whether WGD results in acquisition of metastatic properties. However, the WGD MMECS had gained the ability to form tumours when injected subcutaneously, while the diploid MMECs did not (Fujiwara et al., 2005). When MMECs were subjected to an initiation/promotion transformation assay, WGD clones demonstrated anchorage independent growth (Fujiwara et al., 2005). This suggests that selective pressures and the environment faced by WGD cells may be influential with regards to transformation of these cells.

In human tumours, cells at the invasive front of primary colorectal adenocarcinomas were frequently found to be aneuploid and near triploid or tetraploid unlike cells in the main tumour mass, suggesting that these likely WGD cells displayed invasive characteristics *in vivo* (Wangsa et al., 2018). Analysis of primary breast tumours and their metastases has also revealed WGD and aneuploidy along with higher levels of GI in metastases compared to primaries (Bakhoum et al., 2018). The relationship between aneuploidy, GI, and metastasis has also been established in lymph node metastases from locally advanced head and neck squamous cell carcinomas, in which chromosome missegregation was associated with the presence of

metastases (Bakhoum et al., 2018).

1.1.2 Paths to Polyploidy

There are many ways in which a cell may become polyploid or WGD, which can be split up into two main paths: errors occurring during the cell cycle; and cell fusion. In this section I will discuss each of these paths to polyploidy, particularly in the context of cancer, with an added focus on cell fusion in the next section (Section 1.2).

1.1.2.1 Errors During the Cell Cycle

The mammalian cell cycle, reviewed in Harper and Brooks (2005), consists of five main phases: G_0, G_1, S, G_2, and M. In short, G_0 represents a quiescent state, from which replicating cells enter the G_1 phase in which RNA and proteins are produced. After G_1, cells proceed to S phase, in which DNA is replicated, followed by G_2, in which cells prepare to undergo mitosis. After completing G_2, cells enter M phase, where they proceed through mitosis, which finishes with cytokinesis (division into two daughter cells).

The cell cycle and its constituent phases are tightly regulated and coordinated, but errors do occur, and three specific errors have been identified that result in tetraploidy. These include obstruction of the abscission checkpoint, altered expression of or mutation of genes regulating cytokinesis, and mitotic slippage (Figure 1.1, reviewed in Lens and Medema (2019)).

What the aforementioned pathways have in common is that they result in cells proceeding from the G_2 phase of the cell cycle to the G_1 phase without completing mitosis or cytokinesis. In the final stages of cytokinesis, a cleavage furrow is formed, including a contractile ring that is responsible for cleaving the two cells (Harper and Brooks, 2005). If the cleavage or abscission is physically blocked, as depicted in Figure 1.1a, the cleavage furrow regresses and the cell proceeds into the next cell

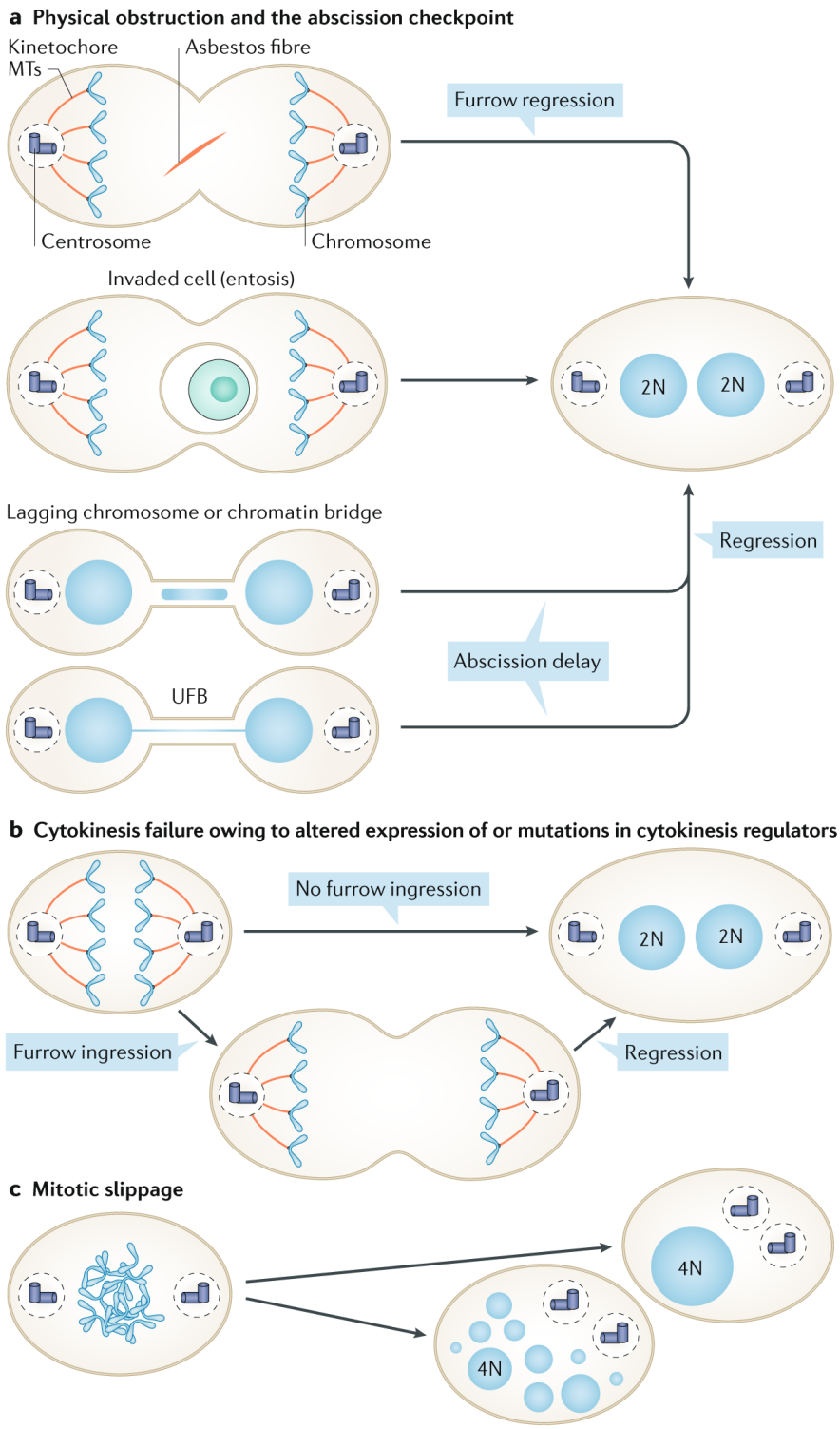


Figure 1.1: Schematic of the possible causes of cytokinesis failure. **a** Physical obstruction of abscission, for example by asebestos fibres, cell entosis, lagging chromatin fibres, and ultrafine bridges. **b** Altered expression of or mutation of genes regulating cytokinesis. **c** Mitotic slippage after a prolonged delay in mitosis. UFB = Ultrafine Bridges. Reprinted by permission from Springer Nature: Nature, Nature Reviews Cancer, Cytokinesis defects and cancer Lens and Medema (2019) Copyright 2019.

cycle, now with two nuclei and 4N DNA content (Lens and Medema, 2019). Cytokinesis can also be bypassed by mutation or altered expression of genes involved in cytokinesis, again resulting in a failure of abscission (Figure 1.1b). In contrast to failures of abscission, mitotic slippage results in WGD after a prolonged delay in mitosis causes cells to re-enter the cell cycle without proceeding through anaphase and cytokinesis (Figure 1.1c; Lens and Medema (2019)). For example, the persistent DNA damage response generated in response to telomere dysfunction has been shown to cause mitotic slippage and tetraploidisation (Davoli et al., 2010).

1.1.2.2 Cell Fusion

Cell fusion is the process by which two cells combine to become one, and can be thought of as the opposite of cell division. Much like WGD, cell fusion can come about in a number of ways with varied consequences for the cell. In the following sections (Sections 1.2 and 1.3) I discuss the various causes of cell fusion, the effects of cell fusion on the cancer genome, methods to identify cell fusion, and evidence for cell fusion in human cancers.

1.2 Cell Fusion

In this section I will begin by discussing the general mechanisms by which cell fusion occurs (reviewed in Brukman et al. (2019)), followed by a summary of the downstream effects of cell fusion in the context of cancer. I will use the term "spontaneous cell fusion" to refer to any fusion event that does not involve a virus.

1.2.1 Mechanisms of Cell Fusion

1.2.1.1 General Principles and Mechanisms

For cell fusion to occur the plasma membranes of the fusing cells must be brought into close contact, assisted by fusion proteins (fusogens), which can be present on

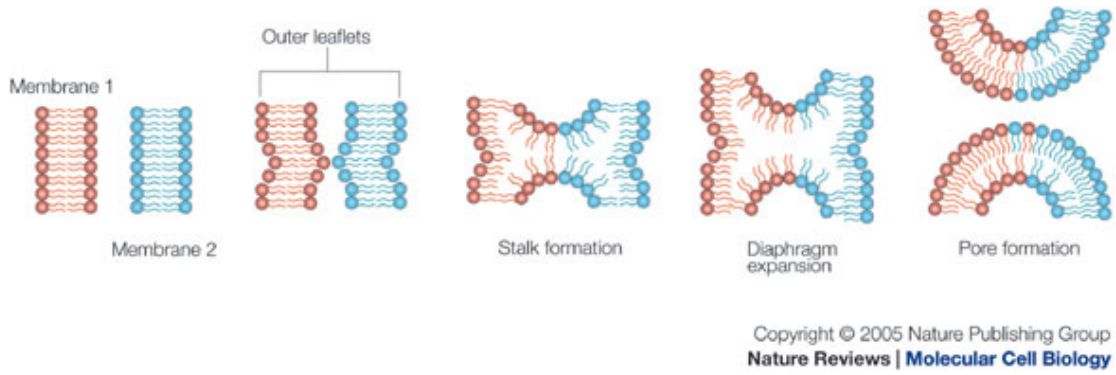


Figure 1.2: Stages of membrane fusion. A schematic representation of the the different stages of membrane fusion. This theoretical model is based on studies of virus-cell fusion. Phospholipids in the proximal layer of the plasma membranes approach each other, likely assisted by fusogens (not shown), and form a stalk structure allowing phospholipid mixing. The stalk expands to a diaphragm, which promotes merging of the distal phospholipid layer, and formation of the fusion pore. Reprinted by permission from Springer Nature: Nature, Nature Reviews Molecular Cell Biology, Biological implications of cell fusion Ogle et al. (2005). Copyright 2005.

one (unilateral fusion) or both of the fusing membranes (bilateral fusion) (Brukman et al., 2019). These proteins bring the two membranes into close proximity, causing deformations in the plasma membrane which eventually overcome the energy barrier required for lipid mixing (Brukman et al., 2019; Sapir et al., 2008). Once the barrier to fusion has been overcome, this process is thought to proceed through a defined set of stages beginning with hemifusion, in which the outer layers of the bilayers mix, while the inner layers remain intact (Figure 1.2; Brukman et al. (2019); Sapir et al. (2008); Chernomordik and Kozlov (2008)). Hemifusion is followed by merging of the internal layer of the lipid bilayer, creating a fusion pore, that can expand for full cell fusion (Brukman et al., 2019; Sapir et al., 2008; Chernomordik and Kozlov, 2008).

Cell fusion is a tightly regulated process that occurs naturally during the formation of myoblasts, placental trophoblasts, osteoclasts, and cells of the eye lens, as well as in stem cells, between macrophages, and of course between sperm and oocytes (Wassarman and Litscher, 2008; Wakelam, 1985; Bassnett et al., 2011; Malassiné

et al., 2011; Jansen et al., 2012; Helming and Gordon, 2009; Ying et al., 2002). Cell fusion has also been described following viral infection, and it has been estimated that 18 different genera that infect humans may contain fusogenic viruses (Duelli et al., 2007). Interestingly, Sycytin-1 and Syncytin-2, endogenous fusogens involved in fusion of placental trophoblasts, are thought to be derived from endogenous retroviruses (Mi et al., 2000). This highlights the importance of cell fusion and viral fusogens in diverse biological processes. In the next section I will discuss some examples of virus induced cell fusion in the body.

1.2.1.2 Virus-Induced Cell Fusion

Many viruses that infect humans have the potential to fuse cells together. Table 1.1 lists several examples of such viruses and supporting evidence. A classic example of virus induced cell-cell fusion playing a role in disease is the formation of polykaryocytes, or multinucleated cells, by varicella-zoster virus (VZV) which is the causative agent for chickenpox and shingles (reviewed in Oliver et al. (2020)). During the course of VZV infection cell fusion functions as a mechanism for the virus to spread throughout the body (Oliver et al., 2020).

Interestingly, the oncogenic Human Papillomavirus 16 (HPV16) has also been implicated as a cause of cell fusion in two studies (Hu et al., 2009; Hu and Ceresa, 2009). More recently, SARS-CoV-2, which causes COVID-19 and is responsible for the current global pandemic, has been found to fuse cells *in vitro*, and pathology reports have described the presence of syncytia in the lungs of autopsied COVID-19 patients (Musarrat et al., 2020; Xu et al., 2020).

This serves to highlight both the diversity of viruses that have the potential to cause cell fusion, as well as providing an indication of the frequency with which individuals may be exposed to virus-induced cell fusion.

Table 1.1: Examples of viruses that fuse human cells. Reprinted by permission from Springer Nature: Nature, Nature Reviews Cancer; Cell-to-cell fusion as a link between viruses and cancer Duelli and Lazebnik (2007). Copyright 2007. Adapted and updated by M. Jakobsdottir

Family	Example genera with evidence of human cell-cell fusion	*In vitro* Experiments	Human Data
Arenaviridae	Junin Virus	Castilla et al. (2005)*	
		Castilla et al. (2005)*	
		Castilla and Mersich (1996)*	
		York et al. (2004)*	
Bornaviridae	Borna Disease Virus	Gonzalez-Dunia et al. (1998)*	
		Furrer et al. (2004)*	
Bunyaviridae	La crosse bunyavirus	Plassmeyer et al. (2005)*	
		Gonzalez-Scarano et al. (1984)*	
	Hantavirus (HFRS)	Arikawa et al. (1985)*	
	Hantaan virus	Ogino et al. (2004)*	

Note: * pH altered to induce fusion

Continued on next page

Family	Example genera with evidence of human cell-cell fusion	*In vitro* Experiments	Human Data
Coronaviridae	SARS	Petit et al. (2005)	
		Li et al. (2003)	
	SARS-Cov-2	Musarrat et al. (2020)	Xu et al. (2020)
Filoviridae	Ebola virus	Bär et al. (2006)*	
	Marburg virus	Chan et al. (2001)*	
Flaviviridae	Hepatitis C virus	Takikawa et al. (2000)*	
		Kobayashi et al. (2006)*	
Hepadnaviridae	Hepatitis B virus	Tugizov et al. (1993)	
Herpesviridae	Herpes simplex virus	Jones and Geraghty (2004)	
	Varicella zoster	Cole and Grose (2003)	
		Oliver et al. (2020)	
	Cytomegalovirus	Kinzler and Compton (2005)	
	Epstein-Barr	Bayliss and Wolf (1980)	
		Wu et al. (2005)	
Orthomyxoviridae	Parainfluenzavirus	Greengard et al. (2000)	

Note: * pH altered to induce fusion

Continued on next page

Family	Example genera with evidence of human cell-cell fusion	*In vitro* Experiments	Human Data
	Influenzavirus	Chan et al. (2001)*	
		Varečková et al. (2003)*	
Papillomaviridae	Human Papillomavirus (HPV)	Hu et al. (2009)	
Paramyxoviridae	Hendra virus	Bossart et al. (2002)	
	Nipah virus	Bossart et al. (2002)	
	Measles virus	Herschke et al. (2007)	Lightwood et al. (1970)
Pneumovirinae	Human Respiratory Syncytial Virus	Branigan et al. (2006)	Neilson and Yunis (1990)
			Hall et al. (1986)
Poxviridae	Vaccinia virus	Senkevich and Moss (2005)*	
Reoviridae	Rotavirus	Kutsuzawa et al. (1982)	
Retroviridae	HIV	Bär et al. (2006)	Dargent et al. (2000)
	HERV		
		Kondo et al. (2015)	
Rhabdoviridae	VSV	White et al. (1981)*	
	Rabies	Mifune et al. (1982)*	

Note: * pH altered to induce fusion

Continued on next page

Family	Example genera with evidence of human cell-cell fusion	*In vitro* Experiments	Human Data
Togaviridae	Sindbis virus	Paredes et al. (2004)*	
		Elena Zaitseva and Chernomordik (2005)*	
	Semliki forest virus	White et al. (1981)*	

Note: * pH altered to induce fusion

1.2.2 The Effects of Cell Fusion

The consequences of cell fusion are manifold, ranging from the formation of large syncytia to formation of unique hybrid cells, which will be the focus of this section (Figure 1.3). In the initial phases, cell fusion can result in binucleate cells, termed heterokaryons, which can become synkaryons if the nuclei fuse together to become one, resulting in a combined, duplicated, genome. Conceptually, there are differences between whole genome duplication that proceeds through failure of cytokinesis compared to cell fusion, as cell fusion may involve fusion of different cell types with different epigenomes and transcriptional programmes, or cells in different stages of the cell cycle. Thus, cell fusion and formation of a synkaryon is expected to have different consequences for the genome than failure of cytokinesis.

Although studies of cell fusion in human tumours are rare, numerous studies have investigated the effect of cell fusion in cancer using animal models and *in vivo* systems. For the most part, these studies have consistently demonstrated an increase in metastatic potential in fused cells, as well as altered transcription, increased tumour heterogeneity, and in some cases resistance to therapy. In the following sections I will discuss the effects of cell fusion on the genome in the context of cancer.

1.2.2.1 Altered Transcription

It has been suggested based on computational modelling of cell fusion, that bringing together two cells under different transcriptional control might result in a hybrid transcriptional state, representing an intermediate between the two parental cells (Koulakov and Lazebnik, 2012). However, modelling also suggested the emergence of aberrant transcriptional states upon cell fusion, indicating that cell fusion may result in transcriptional profiles that bear no resemblance to the parental cells, or any known cell type (Koulakov and Lazebnik, 2012). While modelling transcriptional profiles may involve oversimplification, several studies have indeed reported altered

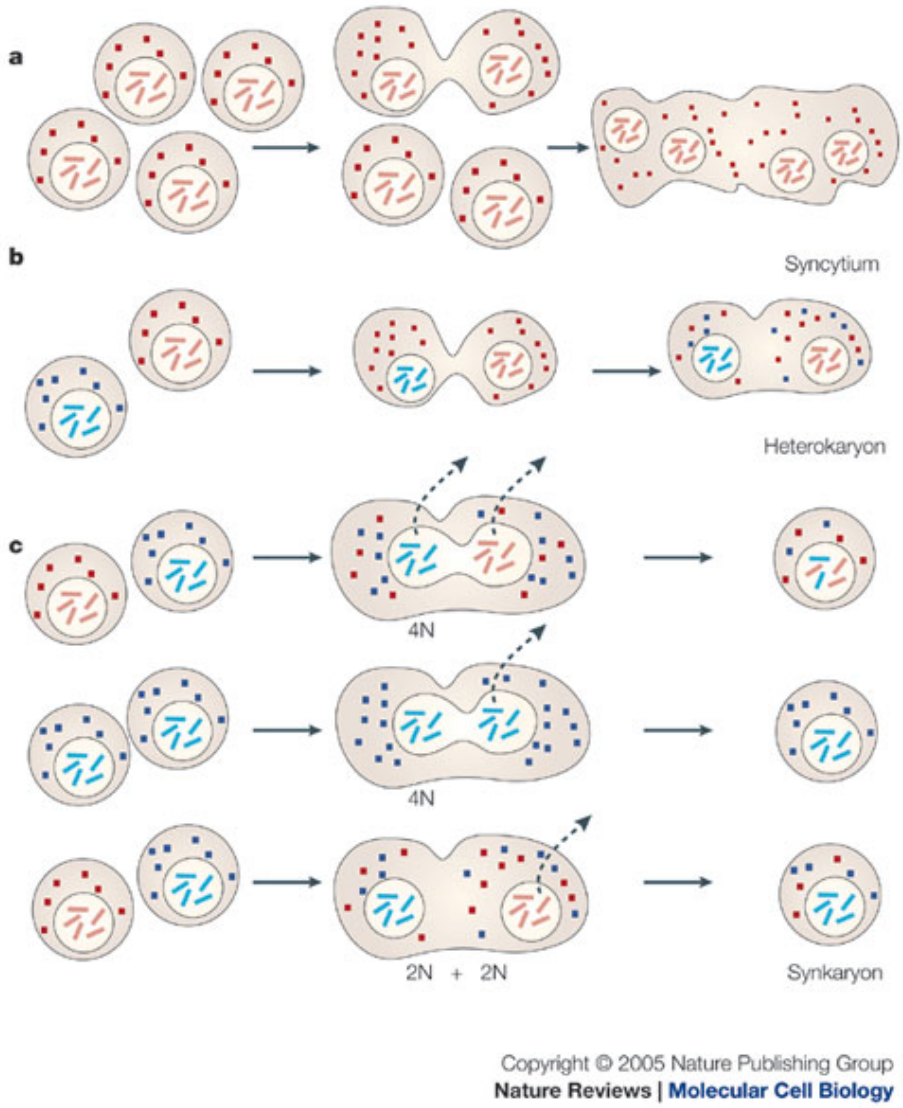


Figure 1.3: a. Cells of the same type fuse to form a syncytium, or a large cell with multiple nuclei. b. Two different cell types fuse together to form a heterokaryon, i.e. a fused cell with two distinct nuclei. c. Two cells fuse to form a synkaryon by allowing their nuclei to fuse, resulting in a cell with a single nucleus with 4N DNA content. After nuclear fusion, chromosomes can be lost or redistributed (arrows). If the nuclei do not fuse, a multinucleate cell may become a synkaryon by shedding a nucleus. Reprinted by permission from Springer Nature: Nature, Nature Reviews Molecular Cell Biology. Biological implications of cell fusion Ogle et al. (2005). Copyright 2005.

transcription following cell fusion.

Many studies have found that the transcription profiles of hybrids fall somewhere in between those of the parental cell lines, and that some hybrids can retain expression of lineage specific genes or pathways, while also presenting a transcriptional profile that is unlike that of either parent. Analysis of hybrids derived from fusion between breast cancer cells and breast epithelial cells demonstrated a transcriptional profile that was an intermediate of the two parents (Dittmar et al., 2011). Searles et al. (2018) report the expression of cell type restricted genes from both parent cells in hybrids formed after the fusion of mouse melanoma cells and mouse embryonic fibroblasts, though expression patterns differed between different clones. The authors suggest that cell fusion may play a role in diversification of transcriptional profiles, thereby contributing to tumour heterogeneity (Searles et al., 2018).

Similar results have been reported for hybrids formed between breast cancer cells and stromal cells in an *in vivo* mouse model (Chitwood et al., 2018). Hybrids were found to express genes that were critical to each parental cell type; however, the single cell RNAseq profiles of fused cells formed a cluster that was distinct from either parental cell cluster, again suggesting increased diversity in hybrids (Chitwood et al., 2018). Distinct clustering of fused cells compared to parental cells has also been reported for fusions between human myoblast cell lines (Merle et al., 2020). Interestingly, the authors were also able to separate early-passage hybrids from late-passage hybrids, suggesting that the transcriptional profile evolves with the tumour (Merle et al., 2020).

Miroshnychenko et al. (2021) also report hybrid single-cell transcriptional profiles that represent an intermediate between two profiles of the parental MDA-MB-231 and SUM159PT breast cancer cell lines. The authors report a shift in the transcriptional profiles between early and late passages, indicating the emergence of a distinct hybrid population following diversification, which increased over time (Miroshnychenko et al., 2021).

Freeman et al. (2016) report that single cell sequenced hybrids derived from mesenchymal stem cells (MSCs) and cardiomyocytes each displayed high levels of heterogeneity, with 39.3% closely resembling MSCs, 25% resembling cardiomycytes, and 35.7% displaying unique transcriptomes. These unique hybrids displayed increased expression of genes associated with ribosomal biogenesis, the cell cycle, DNA replication, and chromosome segregation (Freeman et al., 2016). In addition, two of the single cell hybrids had expression profiles that clustered with data from cancer cells, despite being the products of non-cancerous cells, suggesting a role for cell fusion in cancer initiation (Freeman et al., 2016). Finally, hybrids formed between macrophages and melanoma cells have been shown to retain expression of macrophage signatures, while also displaying an increase in the expression of pathways associated with metastatic spread and neoplasia (Gast et al., 2018).

Further to this, several other studies of cell fusion have also reported expression of genes that may be associated with acquisition of metastatic properties. Hybrids between breast cancer cells and macrophages were found to have increased expression of genes downstream from the Wnt/β-catenin pathway, as well as expression changes suggestive of epithelial mesenchymal transition (EMT) (Zhang et al., 2019). In the breast cancer-stromal cell fusions, hybrids presented increased expression of mitochondrial elements, which was postulated to be an adaptation to meet the increased energy demands of a tetraploid cell (Chitwood et al., 2018). These hybrids also displayed increases in MAPK signalling, cell proliferation pathways, cell junction assembly, and protein synthesis, hinting at the development of metastatic properties (Chitwood et al., 2018).

Additionally, polyploid giant cells, generated through fusion between two different ovarian cancer cell lines, demonstrated changes in expression of genes associated with cell adhesion, cell motility, and invasiveness (Yart et al., 2020). Fusion between immortalised myoblasts and transformed fibroblasts also revealed increased transcription of processes potentially involved in metastasis, such as genes associ-

ated with cell mobility, proliferation, lipid metabolism, angiogenesis and apoptosis (Delespaul et al., 2020). Further, a study in which two metastatic cell lines with different tropisms (bone metastatic SCP2 cells and lung metastatic LM2 cells both derived from MDA-MB-231 cells) found that fused cells expressed both metastasis signatures (Lu and Kang, 2009).

Taken together, these studies confirm that cell fusion between various different cell types can generate hybrids with transcriptional profiles that are distinct from the parental cells. These studies also demonstrate that this altered transcription may promote metastasis associated behaviour.

1.2.2.2 Tumour Heterogeneity and Genomic Instability

In the previous section I alluded to changes in transcriptional profiles reflecting increases in tumour heterogeneity as a result of cell fusion. Indeed, multiple studies have demonstrated increased intratumoural heterogeneity of tumours developed after injection of hybrid cells into mice, compared to parental cells. Following their observation of increased heterogeneity of transcription profiles, Searles et al. (2018) demonstrate the formation of hybrids *in vivo* in a mouse model, which presented with varied genome sizes and stability. Variation in ploidy, particularly over time, has also been observed in fusions involving many different cell types (Merle et al., 2020; He et al., 2015; Lartigue et al., 2020).

Delespaul et al. (2019) report increases in both genomic instability and CNVs in the genomes of clones isolated from hybrid grafts after a period of growth, postulating that these clones initially give rise to a very heterogeneous population from which a dominant clone emerges after a selection event. Interestingly, they show that grafting of cells isolated from hybrid tumours does not result in further evolution of the genome, suggesting that while a hybrid genome may initially lead to greater heterogeneity and rapid evolution of the tumour genome, hybrids eventually adopt a "stable" genome (Delespaul et al., 2019). It is noteworthy that the final

pattern of CNVs adopted by lung fibroblast fusions described by Delespaul et al. (2019) closely resembles those observed in human pleomorphic sarcomas.

In addition to heterogeneity of transcriptomic profiles, genomic instability, DNA content, and CNVs, increased heterogeneity of hybrids has also been described with regards to the appearance and behaviour of these cells. Yin et al. (2020) report high levels of heterogeneity within colonies of individual hybrids arising from fusion of rat neural stem cells and human prostate cancer cells. They also report differences in the ability of different clones to form tumours when injected into mice, and differences in the retention of PSA expression (Yin et al., 2020). Similarly, Wang et al. (2019) report heterogeneity of the growth rate and morphology of fused prostate stromal and mesenchymal cell lines, compared to their parent cells.

Miroshnychenko et al. (2021) propose that cell fusion may act as a mechanism of parasexual recombination in cancer, contributing to increased genomic diversification. The authors describe variable inheritance of SNPs from different parental cell lines, as well as heterogeneity in DNA content in hybrids derived from MDA-MB-231 breast cancer cells fused with either Hs578T or SUM159PT breast cancer cells. Interestingly, hybrids derived from fusion between SUM159PT and MCF10DCIS or MDA-MB-231 and MCF10DCIS appeared to result in more "stable" hybrids that maintained elevated DNA content, suggesting that the degree to which cell fusion promotes increased heterogeneity may be dependent on the combined stability of the parental genomes (Miroshnychenko et al., 2021).

Thus, several groups have demonstrated increased heterogeneity with regard to both the genome and the behaviour of fused cells. These studies have also demonstrated the evolution of the hybrid genome towards a stable karyotype, displaying a similar trend to the cancer genome evolution observed in WGD tumours (Section 1.1.1.3).

1.2.2.3 Resistance to Therapy

Resistance to, or increased tolerance of, chemotherapeutic agents has been noted for hybrid cells on several occasions. Specifically, increased resistance to paclitaxel was observed both *in vivo* and *in vitro* for mouse melanoma and embryonic fibroblast fusions, as well as increased resistance to doxorubicin *in vitro* (Searles et al., 2018). Similarly, mouse BMDC and mammary cell hybrids displayed increased resistance to doxorubicin, paclitaxel, etoposide, and 17-DMAG *in vitro*, thought to be mediated by increased expression of multidrug transporters in hybrid cells (Nagler et al., 2011). Miller et al. (1989) have also shown decreased resistance to methotrexate and melphalan in mouse mammary cell fusions *in vitro*.

1.2.2.4 Acquisition of Metastatic Properties

Given the changes in transcription, genome stability, and behaviour of hybrids, it is not surprising that studies have also reported acquisition of metastatic properties following cell fusion. Many studies have demonstrated the development of tumours in experimental animals following the injection of hybrid cells, or after *in vivo* cell fusion, for a diverse complement of cells (Delespaul et al., 2019; De Baetselier et al., 1981, 1984; Gast et al., 2018; Yin et al., 2020; Sun et al., 2019; Merle et al., 2020; Clawson et al., 2015; Goldenberg et al., 1974; Miroshnychenko et al., 2021; He et al., 2015; Lartigue et al., 2020). In addition, acquisition of metastatic properties, such as increased growth rate, ability to grow on gel matrices, and increased motility have been reported for hybrid cells grown in culture (Chitwood et al., 2018; Yart et al., 2020; Sun et al., 2019; Zhang et al., 2019; Merle et al., 2020; Gast et al., 2018; Miroshnychenko et al., 2021; He et al., 2015; Rachkovsky and Pawelek, 1999). However, a few studies have reported a decrease in metastatic potential after fusion, suggesting that this effect may also be dependent on the cell type involved (Melzer et al., 2019; Lizier et al., 2016).

Taken together, this brings us to the cell fusion theory of metastasis, discussed

in the next section (Section 1.2.2.5).

1.2.2.5 Cell Fusion Theory of Metastasis

The idea that cell fusion might play a role in tumour progression was first proposed by a German physician in 1911 (Aichel, 1911). Since then, a number of studies, such as the ones described in the previous sections, have examined the effects of cell fusion on cancer progression and the acquisition of metastatic properties. The cell fusion theory of metastasis holds that the fusion of two distinct cell types may endow the resulting hybrid cell with metastatic properties, either through inheritance of specific properties, such as motility, from the parent cells, or through the facilitation of cancer genome evolution (Figure 1.4; Duelli and Lazebnik (2007); Pawelek and Chakraborty (2008)).

Within the cell fusion theory of metastasis, virus induced cell fusion is of particular interest (Duelli and Lazebnik, 2007). While cell fusion occurs naturally in various tissues, virus induced cell fusion is perhaps less discriminate with regard to the timing of cell fusion, and the types of cells that are fused, provided they express the required cell surface receptors (Duelli and Lazebnik, 2007). In addition, several viruses have been identified as causative agents in cancer initiation (reviewed in Moore and Chang (2010)). Inhibition of host factors, such as p53, by viruses may facilitate tolerance of cell fusion and evolution of the cancer genome.

1.3 Cell Fusion in Cancer

In this section I will discuss some of the methods used to identify cell fusion, both in experimental systems and in humans. I will also focus on evidence of cell fusion in human cancers.

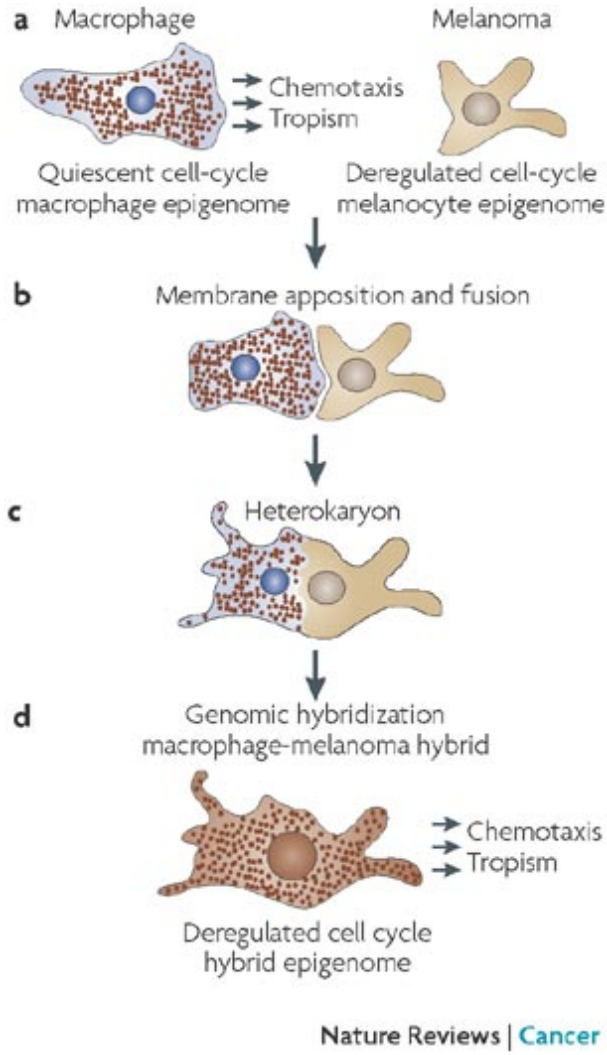


Figure 1.4: a. Two cells with different epigenomes, a melanoma cell and a macrophage, are attracted to each other. b. The plasma membranes of the two cells come into close contact, as the macrophage prepares to ingest the melanoma cell. c. Instead of ingesting the melanoma cell, the macrophage and the melanoma cell fuse, forming a heterokaryon with two distinct nuclei. d. The nuclei fuse to form a synkaryon with properties of both parental cells such as the macrophage's motility and the deregulated cell cycle from the melanoma cell. Reprinted by permission from Springer Nature: Nature, Nature Reviews Cancer. Fusion of tumour cells with bone marrow-derived cells: a unifying explanation for metastasis Pawelek and Chakraborty (2008). Copyright 2008.

1.3.1 Detection of Cell Fusions

1.3.1.1 Fluorescent Markers

A common approach for identifying the products of cell fusion in the lab relies on the detection of multiple fluorescent markers, originating from different parental cells, in the fused cell. This allows for identification of fused cells using fluorescence microscopy, as well as isolation of fused cells using fluorescence activated cell sorting (FACS). A number of different systems have been developed for this purpose. Wang et al. (2012) and others have used combinations of dyes retained in the cytoplasm or plasma membrane, such as red fluorescent DiD and green fluorescent DiO. Similar results have been achieved using expression of fluorescent proteins such as the combination of cells expressing CFP and dsRed described by Delespaul et al. (2020), or GFP and mCherry described by Miroshnychenko et al. (2021).

Another elegant approach is the use of Cre/LoxP system described by Searles et al. (2018). The authors fused together a cancer cell expressing Cre recombinase and a normal cell containing the fluorescent protein tdTomato preceded by a floxed stop codon, resulting in tdTomato expression upon fusion (Searles et al., 2018). The authors also demonstrated that cell fusion, rather than exchange of extracellular vesicles during co-culture, was required for this system to work (Searles et al., 2018).

1.3.1.2 Premature Chromosome Condensation

Premature chromosome condensation (PCC) occurs in interphase cells when they are fused with mitotic cells (Sperling and Rao, 1974). Upon fusion, the interphase cells are forced into undergoing mitosis prematurely, which allows their chromosomes to be visualised and differentiated from the chromosomes of the mitotic cell (Sperling and Rao, 1974). PCC following cell fusion has been used as a technique in cytogenetics to investigate the regulation of mitosis, the cell cycle, and chromosome composition, reviewed in Sperling and Rao (1974). It has also been used to iden-

tify fused cells in multiple cancer types, discussed in further detail in section 1.3.2 (Kovacs, 1985).

While this technique provides direct evidence for the occurrence of cell fusion in human tumours, it is of limited use as it can only detect cell fusions directly after they have occurred. Following the initial PCC caused by the fusion of two cells, the cell cycle is synchronised in the resulting hybrid cell (Johnson and Harris, 1969; Rao and Johnson, 1970).

1.3.1.3 Gene Expression Profile

Gene expression profiling has been proposed as another potential avenue for identification of fused cells. Evidence from *in vitro* and animal models of cell fusion suggests that fusions may be distinguished from parental cells in this way (Section 1.2.2.1).

Although expression profiling may be a powerful technique for differentiating fused cells from their un-fused parental counterparts, this may be impractical in the context of identifying cell fusion in patient tumours. In human tumours with evidence of cell fusion, fused cells only represent a very small fraction of the cells in a tumour (discussed in Section 1.3.2). Sequencing would thus have to be carried out at the single cell level to distinguish these cells. Alternatively, if a tumour consisted solely of cells derived from a fused cell, it might not be recognised as such, without the comparison of un-fused parental cells. Hence, this approach to identification of fused cells would only be practical in tumours in which both fused and un-fused cells are found at relatively high levels.

1.3.1.4 Cell Surface Markers

Further to gene expression profiling, the expression of certain surface markers as indicators of cell fusion has been used in many different contexts to identify fused cells. The macrophage marker CD163 has been suggested as a useful cell surface marker

to identify macrophage-breast cancer fusions (Shabo et al., 2015). In addition to CD163 other macrophage markers, such as CD45, CD68, CSFR1, CD66b, CD14, CD16, and CD11c have been proposed as indicators of macrophage-cancer cell fusions (Gast et al., 2018). Similarly, suspected macrophage-melanoma fusions have been identified through joint expression of melanocyte markers, such as MLANA and ALCAM, and macrophage markers, such as CD14, CD68, CD204, CD206, and CD163 (Clawson et al., 2015).

However, a potential problem with this approach is that one cannot rule out the possibility that alternative marker expression may be due to genomic instability (Weiler and Dittmar, 2019). It has also been pointed out that fused cells may not maintain consistent expression of markers derived from different cells over time (Weiler and Dittmar, 2019). This approach is further complicated by trogocytosis, the phenomenon by which plasma membrane proteins, and occasionally cytoplasmic components, are acquired from one cell by another through a process described as cell "nibbling" (Bettadapur et al., 2020). Whether trogocytosis plays a role in initiation of cell fusion has not yet been explored.

1.3.1.5 Presence of Y Chromosomes in Male to Female Bone Marrow Transplant Recipients

Cell fusions can potentially be identified through detection of marker chromosomes from different cells within the same cell. This approach has been used with some success in several instances in which the presence of a Y chromosome was detected in tumours which emerged following a sex mismatched bone marrow transplant (Yilmaz et al., 2005; Myerson and Parkin, 2019; Gast et al., 2018). These findings are discussed in more detail in Section 1.3.2.

The presence of different marker chromosomes has also been a useful tool in *in vitro* experiments, providing further confirmation of cell fusion, in addition to detection using fluorescent markers (Gast et al., 2018).

1.3.1.6 STR Profiling

Short Tandem Repeat (STR) profiling, a technique borrowed from the field of forensic genetics, has also been used to successfully identify cell fusion following bone marrow transplant (LaBerge et al., 2017; Lazova et al., 2013). STRs are short sequences of repetitive DNA that are hypervariable and passed on from parent to offspring (Tautz, 1989). If enough individual STR loci are considered, a unique profile can be built up for each individual, which makes it a particularly useful technique for determining the owner of DNA left at a crime scene or, in this case, the presence of two distinct genetic profiles in a single sample (Tautz, 1989).

1.3.2 Evidence of Cell Fusion in Human Cancers

Despite widespread speculation that cell fusion may play an important role in cancer initiation and progression, evidence for cell fusion in human tumours is scant. This is in part due to the difficulty of detecting fusion between cells that contain the same genome. To date, the most convincing evidence for cell fusion occurring in human tumours comes from the study of PCC, described in Section 1.3.1.2. PCC has been observed in leukemia in both adults and children, breast carcinoma, bladder carcinoma, ovarian cancer, cervical cancer, gastric cancer, colorectal carcinoma, malignant histiocytosis, rhabdomyosarcoma, and a paediatric tumour of the central nervous system (CNS) (Kovacs, 1985; Miles and Wolinska, 1973; Williams et al., 1976; Sreekantaiah et al., 1987; Reichmann and Levin, 1981; Atkin, 1979; Petković et al., 1988; Sainati et al., 1996; Petković et al., 1994). In these studies, the percentage of metaphases displaying PCC ranged from 1.4% to 100% (Kovacs, 1985; Williams et al., 1976).

Evidence of cell fusion in human tumours has also emerged from several studies of tumours arising after a bone marrow transplant. Notably, Myerson and Parkin (2019) provide convincing evidence of cell fusion in the form of hybrid cells in the livers of four female patients following a sex mismatched bone marrow transplant,

detected using conjoint immunohistochemistry *in situ* hybridisation. Three of the four patients presented with tetraploid cells containing a Y chromosome. Cell fusion was determined to be the most likely explanation, as male tetraploids (XXYY) and their derivatives (e.g. XYY or YY) were not detected, making transdifferentiation followed by tetraploidisation of donor cells unlikely, whereas derivatives of mixed male/female hybrids (e.g. XXXY or XXY) were detected in all of the samples (Myerson and Parkin, 2019). Presence of a Y chromosome and multiple X chromosomes has also been reported in epithelial cells of the intestine following sex-mismatched bone marrow transplant (Silk et al., 2013).

The idea of identifying different marker chromosomes as a sign of cell fusion has also been used in tumours from patients that did not receive a transplant of any sort. Andersen et al. (2007) demonstrate the presence of specific myeloma translocations in osteoclast nuclei from 8 multiple myeloma patients, and further confirm the inclusion of myeloma nuclei into osteoclasts *in vitro*. In another instance of potential cell fusion in tumours that did not involve a transplant, Clawson et al. (2015) isolated suspected macrophage-melanoma fusions from the peripheral blood of cutaneous melanoma patients. The cells stained positive for both melanocyte specific markers, macrophage markers, and epithelial markers, and over half of the populations studied contained inactivating BRAF mutations, common to melanomas (Clawson et al., 2015). Further to this, tissue specimens from these tumours also contained similar cells that stained positive for all three markers (Clawson et al., 2015).

In addition, studies by Lazova et al. (2013) and LaBerge et al. (2017) used STR analysis to detect the presence of both donor and patient DNA within the same tumour. Lazova et al. (2013) describe a melanoma-brain metastasis with a hybrid genome, detected using STR profiling. The authors identified alleles originating from both the donor and the patient in multiple regions of the same tumour. LaBerge et al. (2017) describe a patient who developed a malignant melanoma with spread to

the lymph nodes, 8 years after receiving a sex mismatched bone marrow transplant. The authors detected STR alleles from both the donor and the patient in microdissected regions of the tumour. While both studies took care to microdissect regions of the tumour free from leukocytes, fusion could not be determined definitively due to the nature of the STR assay, which requires bulk analysis of 500-2000 cells (Lazova et al., 2013). Thus, without analysis at the single cell level, fusion cannot be confirmed. In addition, it is possible that the tumours may have contained two populations of cells, one derived from the donor and one from the patient.

In a less convincing study, Yilmaz et al. (2005) present FISH analysis of a renal carcinoma in a female patient who received a bone marrow transplant from a male relative two years prior. The authors probed for chromosomes 4, 11, 17, and Y, and identified cells carrying a Y chromosome in the tumour, as well as a trisomy of chromosome 17. Chromosomal gains were not detected on chromosomes 4 and 11, suggesting that these cells are unlikely to have resulted from either cell fusion or whole genome duplication, as one would expect to see polysomies of multiple chromosomes in this event. The authors also note that gain of chromosome 17 is common in renal cell carcinomas. In this case, it is likely that the detection of the Y chromosome in the tumour was due to differentiation of the bone marrow transplant cells into cancer cells, or due to infiltration of bone marrow derived cells in the tumour.

1.3.3 Whole Genome Sequencing

To date, no study has attempted to identify the products of cell fusion in tumour whole genome sequencing (WGS) data, and previous approaches to identifying cell fusion in human tumours do not translate well into WGS data analysis. While researching ways in which virus induced cell fusion might be detected in WGS data, I came across the APOBEC3 family of cytidine deaminase enzymes, which play a role in the host response to viral infection and are associated with specific mutations

seen in cancer.

The suggestion that the presence of APOBEC3 mutations in cancer might be an indication of previous viral infection (Alexandrov et al., 2013) prompted me to investigate the relationship between viral infection and APOBEC3 activity in my data, in the hope that this might distinguish the products of virus induced cell fusion from spontaneous cell fusion. This then burgeoned into a study of the relationship between APOBEC3 mutations and genome stability.

The following Section (Section 1.4) provides an introduction to the APOBEC3 enzymes, and discusses their involvement as causative agents of DNA damage and mutations in cancer.

1.4 APOBEC3s

1.4.1 Introduction

The human Apo-lipoprotein B mRNA editing catalytic polypeptide 3 (APOBEC3) family of enzymes encompasses eight closely related cytidine deaminases termed APOBEC3A-H, found exclusively in mammals (Conticello et al., 2005). The APOBEC3 enzymes target single stranded DNA, in which they deaminate cytosines to uracils, resulting in characteristic C>T mutations at specific sequence contexts (Bishop et al., 2004). APOBEC3 enzymes originally rose to prominence after it was discovered that they can act as restriction factors in HIV infection; however, they have since been studied in great detail in both cancer genomics and in the genome stability field (Sheehy et al., 2002; Bishop et al., 2004). In the following sections, I will discuss the relationship between the activity of APOBEC3s and cancer, genome stability, and viral infection. I will mainly focus on APOBEC3A and APOBEC3B, due to their importance in cancer.

1.4.2 APOBEC3A

APOBEC3A (A3A) is deaminase proficient with a preference for deaminating cytosines in the 5'-TC sequence context (the edited C being underlined; Bishop et al. (2004); Shinohara et al. (2012)). A3A expression has been reported in peripheral blood mononuclear cells (PBMCs), such as monocytes, macrophages and dendritic cells (Koning et al., 2009; Refsland et al., 2010; Stenglein et al., 2010). Expression of A3A varies between different tissues, with particularly high levels observed in the lung and spleen, and moderate levels reported in adipose tissue (Refsland et al., 2010).

Endogenous A3A is primarily located in the cytoplasmic compartment of the cell (Land et al., 2013). However, A3A constructs that are over-expressed in cellular systems localise to both the nucleus and the cytoplasm (Land et al., 2013; Landry et al., 2011; Stewart et al., 2019; Mussil et al., 2013). During the cell cycle, A3A is excluded from condensed chromosomes between prophase and anaphase, but maintains a cell-wide localisation in telophase (Lackey et al., 2013). The deaminase activity of A3A does not change between interphase and mitosis (Lackey et al., 2013).

A3A is robustly induced after exposure to IFN-α, with both expression levels and protein levels increasing 30-100 fold (Koning et al., 2009; Refsland et al., 2010; Stenglein et al., 2010). However, A3A is not induced in response to T cell mitogens such as IL-2 and PHA (Refsland et al., 2010).

Over-expression of A3A is lethal in many cellular systems, and has been shown to cause cell cycle arrest (Lackey et al., 2013; Landry et al., 2011; Mussil et al., 2013; Burns et al., 2013).

1.4.3 APOBEC3B

Similar to A3A, APOBEC3B (A3B) displays a preference for deamination of cytosines at the 5'-TC sequence context (Bishop et al., 2004). A3B localises only to

the nucleus and is not found in the cytoplasm (Landry et al., 2011; Lackey et al., 2013; Leonard et al., 2013). However, A3B is excluded from mitotic chromosomes (Lackey et al., 2013). Despite this exclusion, the protein levels and deaminase activity of A3B remain constant throughout the cell cycle (Lackey et al., 2013).

A3B expression has been reported in B-cells, lymphoblastic cell lines and leukemia cell lines (Koning et al., 2009; Refsland et al., 2010; Takei et al., 2020). A3B expression has also been reported at moderate levels in the colon, lung, and spleen, and at lower levels throughout many other tissues (Refsland et al., 2010). Unlike A3A, A3B is not responsive to IFN-α stimulation, but is induced in response to IL-2 and PHA (Koning et al., 2009; Refsland et al., 2010).

1.4.4 APOBEC3G

Unlike A3A and A3B, APOBEC3G (A3G) acts mainly at the 5'-C$\underline{\text{C}}$ sequence context and localises to the cytoplasmic compartment (Bishop et al., 2004; Landry et al., 2011; Lackey et al., 2013). A3G does not appear to come into contact with genomic DNA at any point in the cell cycle, and for this reason it is not thought to contribute to mutations in cancer (Lackey et al., 2013).

Similar to A3A and A3B, moderate expression of A3G has been reported in the spleen and lungs; however, A3G is also expressed at moderate levels in the thymus, testes, and ovaries (Refsland et al., 2010; Koning et al., 2009). A3G expression is also detected in PBMCs, B cells, T cells, macrophages, and dendritic cells (Koning et al., 2009; Bishop et al., 2004).

A3G is perhaps best known for its role as a restriction factor for HIV. It has been shown to decrease activity of HIV strains that do not express the viral infectivity factor Vif, by deaminating viral sequences (Bishop et al., 2004).

1.4.5 APOBEC Activity in Response to Viral Infections

APOBEC3 enzymes act as part of the innate immune response to viral infection by deaminating cytosines to uracils in single stranded viral DNA, resulting in hypermutation of viral DNA (Harris et al., 2003). Figure 1.5 represents a schematic view of the actions of A3D, A3F, A3G, and A3H in response to Vif deficient HIV infection. Following HIV replication, the aforementioned APOBEC3s are incorporated into the virion and exert their deamination activities on nascent cDNA produced after entry into a new cell, restricting infection (Harris et al., 2003; Harris and Dudley, 2015). Vif, or the viral infectivity factor, is a HIV protein that targets A3D, A3F, A3G, and A3H for ubiquitination and degradation, preventing restriction of the virus (Goila-Gaur and Strebel, 2008; Harris and Dudley, 2015).

A3A is not sensitive to Vif activity, and is potently induced at both the protein and mRNA levels in response to the increased IFN-α levels as a result of viral infection (Goila-Gaur and Strebel, 2008; Koning et al., 2009; Refsland et al., 2010; Stenglein et al., 2010; Berger et al., 2011). In addition to restricting HIV infection, several APOBEC3 enzymes have also been shown to act on Human T-lymphotropic virus-1 (HTLV-1), Human endogenous retroviruses (HERV), Epstein-Barr virus (EBV), Torque Teno Virus (TTV), parvoviruses, Kaposi sarcoma virus (KSV), vaccinia virus, simian foamy virus (SFV), murine leukaemia virus (MLV), herpes simplex virus-1 (HSV-1), Human Papillomavirus, and Hepatitis B virus (HBV) (Sheehy et al., 2002; Bishop et al., 2004; Harris and Dudley, 2015; Willems and Gillet, 2015; Vartanian et al., 2008).

This highlights the potential for activation of several APOBEC3 enzymes following infection with a diverse complement of viruses.

1.4.6 Processing of Deaminated Cytosines

After cytidine deamination by APOBECs, the resulting uracils can be processed in a number of ways, resulting in a range of different outcomes for the DNA, summarised

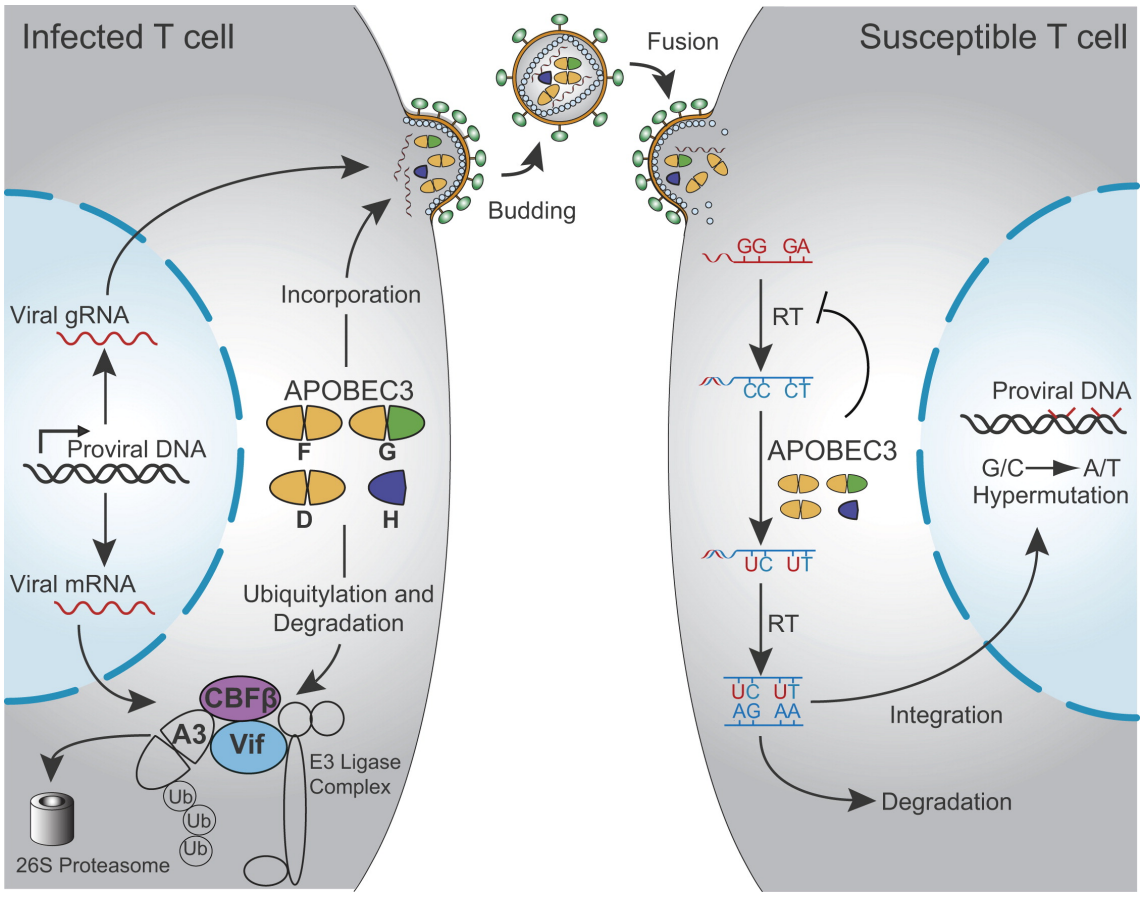


Figure 1.5: "Model for HIV-1 restriction by APOBEC3 enzymes and virus protection by Vif. HIV-1 infection of T cells allows expression of Vif, which recruits CBF-β to form an E3 ubiquitin ligase complex that degrades several different A3 proteins (A3D, A3G, A3F, and A3H). In the absence of Vif, these A3 proteins can package into virions. During subsequent infection of susceptible cells, these A3 proteins deaminate viral cDNA cytosines to uracils and directly impede reverse transcription. Viral cDNA uracils template the insertion of genomic strand adenines and result in hallmark G-to-A hypermutations." Reprinted from Virology, 479, Reuben S. Harris and Jaquelin P. Dudley, APOBECs and virus restriction, Pages No. 131-145, Copyright 2015, with permission from Elsevier. (Harris and Dudley, 2015)

in Figure 1.6. Uracils can be repaired in an error free manner, through base excision repair (BER) (Figure 1.6 A; reviewed in Krokan et al. (2014)). BER involves recognition and excision of the uracil by enzymes such as uracil DNA glycosylase (UNG/UDG), which is followed by strand incision (e.g. by APE1), nucelotide insertion by DNA polymerase-β, and finally, ligation of the DNA backbone (Krokan et al., 2014). In contrast, replication over the uracil results in characteristic C >T Transitions (Figure 1.6 B). Further, error prone mismatch repair (MMR) or translesion DNA synthesis (TLS) following excision of the uracil can result in the generation of transition or transversion mutations (Figure 1.6 C and D).

Abasic sites formed after excision of uracil by UDG can also result in the formation of strand breaks (Leonard et al., 2013). Studies in both yeast and human cell lines suggest that processing of uracils into abasic sites by UDG is a crucial step in the generation of DNA damage by APOBEC3s (Taylor et al., 2013; Landry et al., 2011). In the absence of UDG, the number of APOBEC3 mutations and clusters of APOBEC3 mutations (discussed in Section 1.5.4) is greatly reduced (Taylor et al., 2013; Stenglein et al., 2010).

1.5 APOBEC3s in Cancer

Although the APOBEC3 enzymes have well defined roles in the cell, they have come under investigation as potential sources of cancer initiation and progression due to their off-target effects on the host genome.

1.5.1 The APOBEC3 Mutational Signatures

Due to their well defined preferred sequence contexts, A3A and A3B are thought to be responsible for two single base substitution (SBS) signatures, SBS2 and SBS13, described in Alexandrov et al. (2013) and Alexandrov et al. (2020). SBS2 is mainly characterised by C>T mutations at the 5'-T<u>C</u>W trinucleotide context, where W

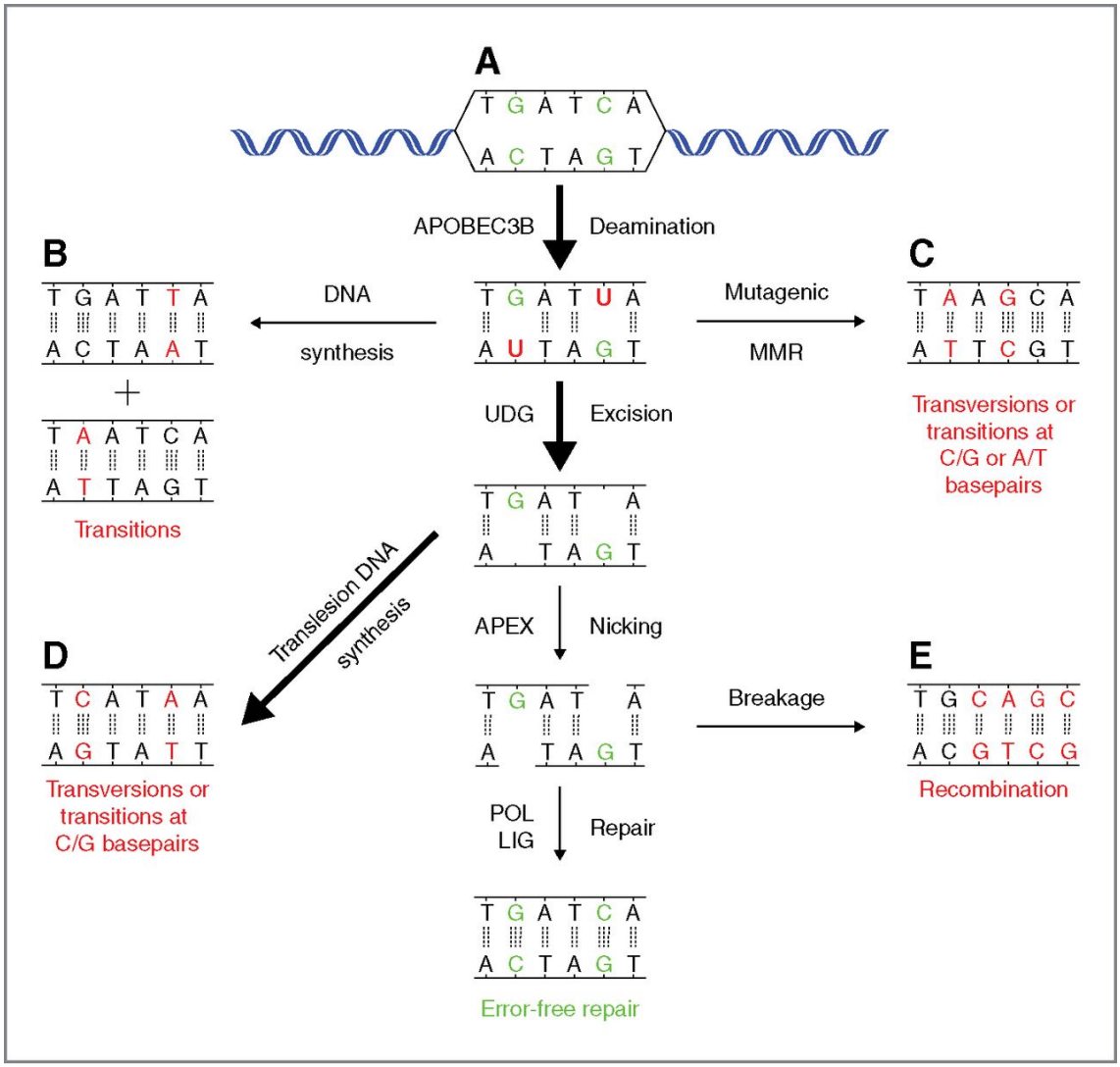


Figure 1.6: "DNA deamination model for mutation in cancer. APOBEC3B catalysed C-to-U deamination events in single-stranded DNA can be repaired error-free (A) or processed in an error-prone manner by DNA synthesis (B, D), mutagenic repair (C), or recombination (E)."(Leonard et al., 2013). Reprinted from Cancer Research, Copyright 2013, 73(24), pp. 7222-7231, Leonard et al. (2013), Up-regulation and Genomic Mutation Patterns in Serous Ovarian Carcinoma, with permission from AACR

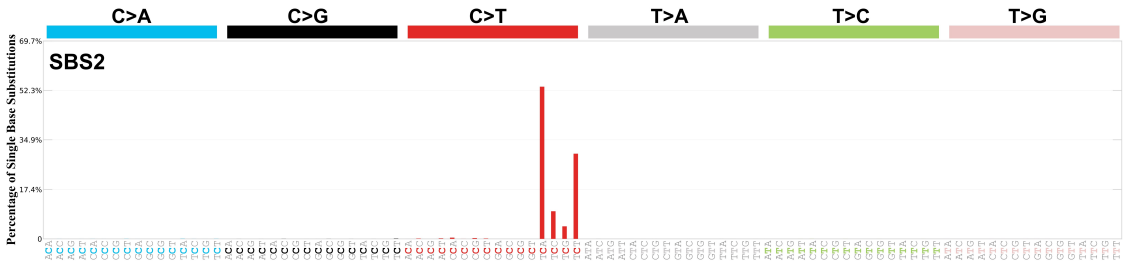


Figure 1.7: Trinucleotide context of the SBS2 mutation signature, attributed to APOBEC3A and APOBEC3B activity (Alexandrov et al., 2020). This figure was downloaded from the Catalogue Of Somatic Mutations In Cancer (COSMIC) website V3.2 March 2021 (Tate et al., 2019).

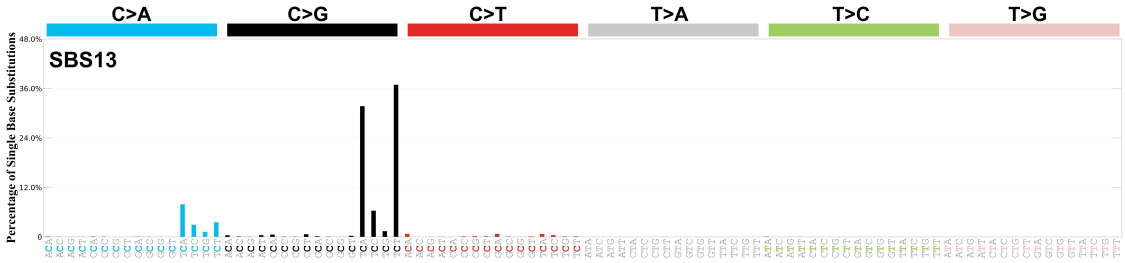


Figure 1.8: Trinucleotide context of the SBS13 mutation signature, attributed to APOBEC3A and APOBEC3B activity (Alexandrov et al., 2020). This figure was downloaded from the Catalogue Of Somatic Mutations In Cancer (COSMIC) website V3.2 March 2021 (Tate et al., 2019).

represents an adenine or a thymine. SBS2 is also characterised by a smaller contribution of C >T mutations at the 5'-TCS context, where S represents a guanine or a cytosine (Figure 1.7). Similarly, SBS13 is characterised by C>G mutations at the 5'-TCW trinucleotide context, with an additional, smaller, C>A component at the 5'-TCN context and C>G mutations at the 5'-TCS trinucleotide context, where N represents any nucleotide (Figure 1.8).

A3A expression has been associated with APOBEC3 signature mutations in several cancer types, including breast, head and neck, bladder, cervical, lung adenocarcinoma, and thyroid cancer (Chen et al., 2019; Middlebrooks et al., 2016). In addition, expression of A3B is associated with APOBEC3 mutations in stomach, pancreatic, and kidney cancers (Chen et al., 2019; Middlebrooks et al., 2016).

Expression of human A3A was recently shown to produce both SBS2 and SBS13 mutations in an *in vivo* mouse model of adenomatous polyposis coli (APC) and

liver cancer, demonstrating a causal link between APOBEC3 expression and the two mutational signatures (Law et al., 2020). These conclusions are also supported by recent work from Petljak et al. (2021), who demonstrate that A3A is the main APOBEC3 responsible for episodic bursts of SBS2 and SBS13 mutations in cancer cell lines, with A3B contributing to a lesser degree.

1.5.2 Role of APOBEC3s in Different Cancer Types

The carcinogenic potential of APOBEC3s has been highlighted in many different cancers including multiple myeloma, breast cancer, lung cancer, and urothelial carcinoma (Roberts et al., 2013; Walker et al., 2015; Roberts et al., 2013; Swanton et al., 2015; Taylor et al., 2013; Mullane et al., 2016; Glaser et al., 2018; Harris et al., 2020; Burns et al., 2013). mRNA expression levels of A3A and A3B have been found to be higher in bladder, breast, lung, kidney, prostate, and ovarian tumours compared to matched normals (Roberts et al., 2013; Leonard et al., 2013). In addition, levels of A3A and A3B correlate with mutation burden and increased numbers of APOBEC3 mutations (Roberts et al., 2013; Glaser et al., 2018; Burns et al., 2013).

High levels of APOBEC3 mutations have been linked with poor prognosis in multiple myeloma (Walker et al., 2015). Similarly, several APOBEC3s were found to be up-regulated in acute myeloid leukemia (AML), in which their up-regulation indicated worse survival (Takei et al., 2020). In contrast, high levels of APOBEC3 mutations were associated with better survival in urothelial carcinoma (Glaser et al., 2018). High APOBEC3 expression levels have also been associated with better overall survival specifically in cisplatin treated urothelial carcinoma (Mullane et al., 2016). It has been suggested that this may be due to the increased mutation burden associated with APOBEC expression (Mullane et al., 2016). Improved survival was also seen in bladder cancers with high numbers of APOBEC3 mutations, in particular in treated tumours with elevated A3A expression, as well as clear cell ovarian carcinomas (CCOCs) treated with cisplatin, in which A3B was over-expressed (Mid-

dlebrooks et al., 2016; Harris et al., 2020).

The presence of APOBEC3 mutational signatures has been associated with specific translocations found in multiple myeloma (Walker et al., 2015). In addition, clusters of cytosine and guanine mutations have been found to associate with chromosomal rearrangement breakpoints in colorectal cancer genomes (Roberts et al., 2013).

In vitro experiments involving repeated A3B exposure demonstrated an increase in copy number variations, which correlated with the number of cytosine mutations (Akre et al., 2016). However, a study on breast cancer genomes did not find any correlation between the number of copy number variation (CNV) breakpoints and enrichment for an APOBEC3 mutational signature, but expression of A3B was associated with higher numbers of CNVs in another (Roberts et al., 2013; Kanu et al., 2016).

Several studies have reported an association between APOBEC3 mutations and increased mutation burden. Walker et al. (2015) report an increase in mutation burden, and poor prognosis, for myeloma tumours containing an APOBEC3 mutational signature developed by the authors. A correlation between expression of A3A and A3B and mutation burden has also been shown in bladder cancer (Glaser et al., 2018). Similar results have been reported for A3B expression in 19 different tumour types by Burns et al. (2013), though the authors did not replicate this analysis with A3A expression. Additionally, in a separate study by Burns et al. (2013), the authors demonstrate an increase in C >T mutation load, and total mutation load, with increasing A3B expression in a panel of breast cancers.

These studies demonstrate the presence of A3A and A3B expression and mutations across a range of different cancers. The presence of A3A and A3B mutations and expression has implications for survival, mutation burden, translocations, and copy number variations.

1.5.3 DNA Damage Responses Following APOBEC3 Activation

Several studies have demonstrated formation of DNA strand breaks, DNA damage responses, and cell cycle arrest following overexpression of A3A (Landry et al., 2011; Land et al., 2013; Mussil et al., 2013; Taylor et al., 2013). This response is characterised by the formation of γH2AX foci and 53BP1 foci, indicative of double strand breaks, and RPA foci, indicative of single stranded breaks (Landry et al., 2011; Land et al., 2013; Mussil et al., 2013; Burns et al., 2013). Downstream mediators of DNA damage signalling are also activated following A3A expression, such as phosphorylation of checkpoint kinase 2, Nijmegen breakage syndrome 1, and ATM (Landry et al., 2011; Mussil et al., 2013). A3A induced DNA damage has also been detected using the terminal deoxynucleotidyl transferase dUTP nick end-labelling (TUNEL) assay (Landry et al., 2011). For DNA damage to occur, A3A must be present in the nucleus (Land et al., 2013).

Overexpression of A3B has been reported to cause DNA damage and to inhibit cell cycle progression in some systems (Burns et al., 2013; Lackey et al., 2013; Nikkilä et al., 2017; Taylor et al., 2013; Shimizu et al., 2018), but not others (Landry et al., 2011). In the context of A3B, this damage is characterised by ATR dependent γH2AX foci, 53BP1 foci, and RPA phosphorylation (Taylor et al., 2013; Burns et al., 2013; Nikkilä et al., 2017; Shimizu et al., 2018). A3B overexpression also caused DNA fragmentation in the comet assay and elevation of p53 levels (Burns et al., 2013; Nikkilä et al., 2017). A3B expression has also been associated with increased replication stress in breast cancer cell lines (Kanu et al., 2016). In addition, Takei et al. (2020) report the formation of double strand breaks (DSBs) following A3B expression, as do Shimizu et al. (2018).

These studies indicate that expression of A3A and A3B can cause DNA damage in cellular systems.

1.5.4 APOBEC3 Activation Following DNA Damage

Expression of A3B has been shown to increase in cells treated with either bleomycin, hydroxyurea, aphidicolin, gemcitabine or camptothecin, and in response to γ-irradiation, all of which contribute to the formation of DSBs or single stranded DNA (Middlebrooks et al., 2016; Shimizu et al., 2018; Kanu et al., 2016). Further, experiments in yeast have demonstrated that DSBs may be nucleating events for APOBEC3 catalysed DNA damage (Taylor et al., 2013).

Taken together, these observations suggest that A3B may generate characteristic APOBEC3 mutations in response to damage caused by other processes. Indeed, the role of APOBEC3 enzymes, and the presence of APOBEC3 mutations, is considered largely reactionary to other forms of DNA damage by many in the cancer genomics field.

Activity of the APOBEC3 enzymes at sites of DNA damage has been proposed as the causative factor in kataegis, which is thought to result from the action of APOBEC3 enzymes at single stranded DNA exposed during resection of DNA at DNA strand breaks (Figure 1.9, Taylor et al. (2013); Maciejowski et al. (2015)). Kataegis refers specifically to areas of localised hypermutation characterised by C>G or C>T mutations occurring on the same strand of a chromosome, and is defined as a stretch of six or more mutations with an intermutational distance <1,000bp (Nik-Zainal et al., 2012; Nik-Zainal and Morganella, 2017).

Kataegis has been observed in A3B overexpressing cells both proficient and deficient in p53 expression, as well as in yeast cells transformed with A3A or A3B (Nikkilä et al., 2017; Taylor et al., 2013). In yeast cells, kataegis patterns were produced more readily following a double strand break (Taylor et al., 2013). Kataegis has also been reported in breast cancer genomes, where they have been associated with the presence of genomic rearrangements (Nik-Zainal et al., 2012).

The results of the aforementioned studies suggest that, rather than causing strand breaks, the role of A3A and A3B in cancer is limited to causing muta-

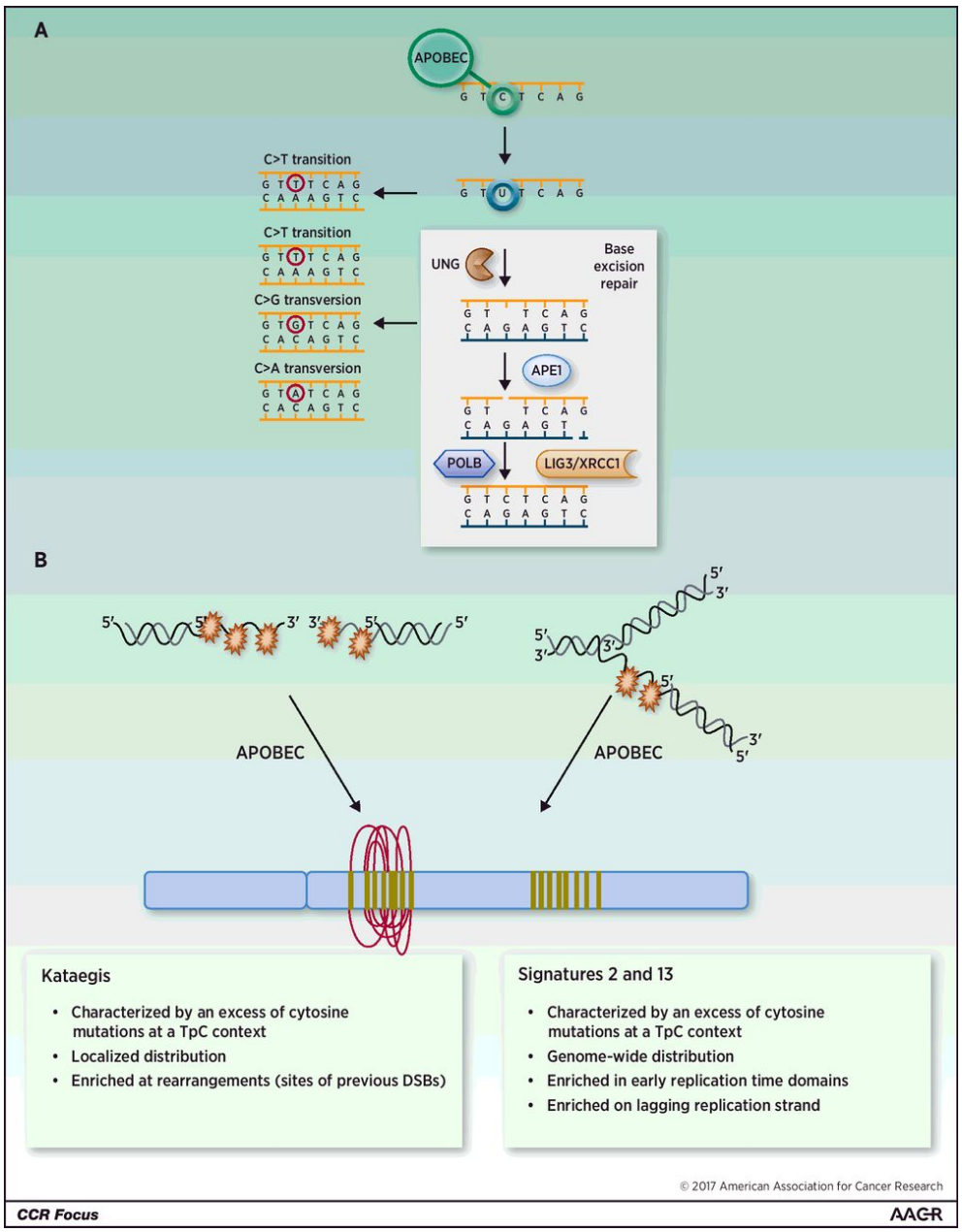


Figure 1.9: Schematic illustration of kataegis. A. APOBEC deaminates cytosines to form uracils in single stranded DNA resulting in characteristic C >X transition or transversion mutations if uncorrected. Uracil can be excised by uracil-N-glycosylase (UNG), resulting in an abasic site repaired by base excision repair (BER). B. Left - Clustered regions of hypermutation affecting the same DNA strand, termed kataegis, are generated following cytidine deamination by APOBEC3s on long stretches of single stranded DNA generated following a double strand break. Right - Individual mutations attributed to SBS2 and SBS13 are generated following APOBEC3 mediated cytidine deamination of single stranded DNA intermediates during DNA replication. Reprinted from Clinical Cancer Research, Copyright 2017, 23(11), pp. 2617-2629, Nik-Zainal and Morganella (2017), Mutational Signatures in Breast Cancer: The Problem at the DNA Level, with permission from AACR

tions through deamination of single stranded byproducts generated during repair of lesions produced by other processes.

1.5.5 Why not both?

The narrative used to describe the action and interaction between APOBEC3s and DNA varies considerably between the fields of genome stability and DNA damage research on the one hand, and cancer genomics on the other. The genome stability field has highlighted the potential for APOBEC3 catalysed cytidine deamination to cause strand breaks and DNA damage (Section 1.5.3), whereas pockets of the cancer genomics field have highlighted the role of APOBECs in causing mutations in response to DNA damage (Section 1.5.4).

More recently, the propensity for APOBEC3s to act as both reactionary and causative agents, with respect to DNA damage, has begun to be explored. The formation of secondary γH2AX foci 12-48 hours after γ irradiation in A3B expressing cells, has been described by Shimizu et al. (2018) in a scenario that suggests A3B may act at sites of DNA damage, causing further, persistent, DNA damage when the resulting uracils are processed. Interestingly, the authors suggest that A3B is consistently synthesised and degraded under normal conditions but stabilised in response to DNA damage, accounting for the increase in A3B levels (Shimizu et al., 2018).

1.6 Aims of the Thesis

The initial aims of this project were to investigate the effects of virus-induced cell-cell fusion on the genome, and to develop methods for identifying tumours in which cell-cell fusion may have occurred using whole genome sequencing data. As the project progressed, I became aware of the involvement of the APOBEC3 enzymes in both viral restriction and their association with genomic instability. Thus, the project

was expanded to investigate the relationship between viral infection, APOBEC3, cell fusion, and genome stability.

The overall hypothesis of this thesis was that virus induced cell fusion plays a role in tumourigenesis and tumour progression. I sought to test this hypothesis using experimental and computational methods by identifying a possible mechanism by which viral exposure could cause genomic instability through APOBEC3s (Chapters 2), assessing whether in vitro viral infection causes genomic instability (Chapter 3), developing novel computational methods to identify cell fusion in whole genome sequencing data from multiple different tumour types (Chapter 4), and by examining genomic changes following cell fusion and whole genome duplication with the aim of identifying a mechanism by which cell fusion promotes tumour progression (Chapter 5). The overarching theme or goal of this thesis is to better understand the ways in which viral infection can indirectly influence the cancer genome.

In Chapter 2 I investigate the relationship between the APOBEC3 mutational signature and genomic instability in whole genome sequencing data from 2,658 tumours. The aim of this chapter is to address the disconnect between different fields in how the interaction between APOBEC3 and genomic instability is viewed. Specifically, I investigate increases in genomic instability associated with the APOBEC3 signature. I also sought to investigate whether evidence of previous viral infection in tumours was associated with the presence of APOBEC3 mutational signatures. I present novel measures of genomic instability and present a pan-cancer analysis of the relationship between APOBEC3 and genomic instability using WGS data.

In Chapter 3 I approach the questions broached in Chapter 2 from the wet lab perspective. The aim was to investigate whether viral infection causes increased genomic instability, whether APOBEC3A, specifically, is associated with increased genomic instability, and whether viral infection is associated with increased levels of APOBEC3A.

In Chapter 4 I aim to identify tumours in which cell fusion may have occurred

by analysing whole genome sequencing data from two large consortia. Identification of cell fusion in WGS data has never been attempted before, existing methods for identifying cell fusions translate poorly for this medium, and no previous study has compared the genomes of hybrid cells to cells which have undergone WGD by another process. Thus, my approach relies on detecting the products of cell fusion between a normal cell and a tumour cell, based on retention of observable differences between these cell types. In this chapter I present several tumour samples which I believe may have originated from a cell fusion event.

Finally, in Chapter 5 I aimed to address differences between the genomes of cells which have undergone virus induced cell fusion, compared to those which have undergone fusion by another method or those which have undergone a failure of cytokinesis. I generated WGD cells through fusion or failure of cytokinesis and monitored the karyotype of these cells over time.

Chapter 2

APOBEC3 mutational signatures are associated with extensive and diverse genomic instability across multiple tumour types

A large portion of the material presented in the following chapter was submitted as a manuscript to Genome Biology. After significant revisions, the manuscript is due to be re-submitted to BMC biology. Some passages are reproduced verbatim from this manuscript.

2.1 Introduction

It has been suggested that APOBEC3 enzymes may play a more causative role in the generation of genomic instability by causing the formation of double strand breaks, either through the excision of uracils and cleavage of the abasic site on opposing strands, or through stalling of replication forks at single strand breaks (Burns et al., 2013). The role of AID (Activation Induced Deaminase), which is closely related to

the APOBEC3 family, in somatic antibody diversification, and its association with translocations in B cell tumours, lends credence to this model of APOBEC3 induced double strand breaks (Casellas et al., 2016).

Previous work has largely focused on APOBEC3 activity in breast cancer, and has often been limited to exome sequencing data. In this chapter, I provide evidence that APOBEC3 causes an increased mutation burden and genomic instability via generation of double strand breaks, through analysis of whole genome sequencing data from 2,451 samples across 39 tumour types in the Pan-Cancer Analysis of Whole Genomes Project (PCAWG) (Campbell et al., 2020).

2.2 Materials and Methods

Unless otherwise stated, analyses were carried out using R version 3.6.2 (R Core Team, 2018).

2.2.1 Data

In this study I analysed whole genome sequencing of 2,451 white listed primary tumour samples made available through the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium (Campbell et al., 2020). The full data set consists of 2,658 samples; however, I restricted my analysis to primary tumours included on PCAWG's white list. PCAWG data can be accessed through the ICGC at `http://dcc.icgc.org/pcawg/` . Access to controlled data was granted by the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO) for the ICGC portion of the PCAWG data, and by The Cancer Genome Atlas (TCGA) Data Access Committee for the TCGA portion of the data.

Analysis of the mutational signatures was carried out by the PCAWG Mutation Signatures and Processes working group (Alexandrov et al., 2020). For the analysis reported here I used signatures called using SigProfiler. I also made use of structural

variation data, which was made available through the PCAWG Structural Variation working group (Li et al., 2020). Clustered mutation data related to kataegis was provided by the Evolution and Heterogeneity working group (Campbell et al., 2020). Identification of viral DNA was carried out by the Pathogens working group (Zapatka et al., 2020).

2.2.2 Mixed Effects Models

Mixed effects models were created using version 1.1-23 of the `lme4` package and version 1.0.2.1 of `glmmTMB` (Bates et al., 2014; Brooks et al., 2017). The results of the linear and mixed effects models were presented using version 5.2.2 of the `Stargazer` package and version 1.37.5 of the `texreg` package (Hlavac, 2018; Leifeld, 2013).

I created three mixed effects models to account for the effect of tumour type on the relationship between the number of APOBEC3 mutations and age on the total number of non-APOBEC3 mutations (Equations 2.1-2.3). In addition, six mixed effects models were created to investigate the relationship between the number of APOBEC3 mutations, and the six measures of genomic instability that were investigated (Equations 2.4-2.9). A further six models were constructed to investigate the additional effect of TP53 mutation (Equations 2.10-2.15). Models in which I excluded mutations attributed to kataegis were constructed using the same formulas as equations 2.4-2.15. I also modelled the relationship between the presence of viral DNA, the type of viral genome present, and the virus genus present on the number of SBS2 and SBS13 mutations, accounting for tumour type in equations 2.16-2.18.

For mixed effects modelling of the relationship between number of APOBEC3 mutations and genomic instability I only consider samples which contain APOBEC3 mutations. For models which excluded mutations attributed to kataegis, the number of mutations located in kataegis clusters attributed to APOBEC3 were subtracted from the total number of SBS2 and SBS13 mutations; samples for which this pro-

duced a negative number of mutations were excluded from further analysis.

For each measure of genomic instability I formulated models with and without interaction terms between the dependent variables that were surveyed. I also built models based on different distributions for the independent variable (e.g. the normal distribution, negative binomial distribution, and the zero-inflated negative binomial distribution). I selected the optimum model for each measure by selecting the model with the lowest Akaike information criterion (AIC) and a p-value <0.05 when compared to other models using an ANOVA. In addition, I used version 0.3.3.0 of the `DHARMa` R package (Hartig, 2020) to assess model fit and whether the data were overdispersed or zero inflated.

2.2.3 Survival Analysis

Survival analysis and generation of Cox Proportional Hazard mixed effects models was carried out using the `survminer`, `survival`, and `coxme` packages for R (Kassambara et al., 2017; Therneau, 2015, 2020). The patient's overall survival was used as an endpoint. The CoxME models generated are described in detail in equations 2.3.5 and 2.3.5.

2.2.4 Genomic Instability

Genomic instability is characterised by a range of different changes at the chromosome level. Frequent changes include increased numbers of insertions, deletions, translocations, and structural variants (Bayani et al., 2007). I assessed the number of each of these changes using data provided by the PCAWG Structural Variation working group (Li et al., 2020).

Changes in ploidy have also been associated with genomic instability (Bayani et al., 2007). Changes in ploidy were assessed by investigating the proportion of the genome altered (PGA), which describes the proportion of the genome that deviates from copy number 2 or 4, for diploid and tetraploid tumours, respectively. I also

examined the number of copy number segments, which provides an insight into the number of copy number changes across the genome.

In addition, I assessed the number of insertions and deletions (INDELs) that are attributed to INDEL signatures ID6 and ID8. Both ID6 and ID8 have been attributed to error prone non-homologous end joining repair of double strand breaks (Alexandrov et al., 2020). Double strand breaks, when repaired incorrectly, can lead to translocations and genomic instability (Morgan et al., 1998). I reasoned that increased numbers of DNA breaks caused by increased APOBEC3 activity could also be detected as increased levels of ID6 and ID8, reflecting elevated DNA damage repair activity, as well as higher numbers of translocations and INDELs as outcome measures.

2.2.5 Volcano Plot

To aid with visualisation, and to prevent division by 0 when estimating effect sizes, a pseudocount of 1 was added to the medians of the genome instability measures calculated for tumours in each tumour type that either carry SBS2 and SBS13 mutations or do not carry these mutations, with the exception of PGA, before the ratio of the medians was taken. All statistical analysis was carried out on the raw data, without a pseudocount.

For the comparison of the effect of APOBEC3 mutation presence or absence on genomic instability within tumour types, my hypothesis for all of the comparisons is the same, that the presence of SBS2 and SBS13 mutations results in increased genomic instability. For this reason, I used Fisher's combined probability test, to combine the p-values for each hypothesis test within individual tumour types (Fisher, 1970).

2.3 Models

The models that follow are written in the `lme4` and `glmmTMB` formula syntax (a), as well as using mathematical notation (b). I have tried to be consistent in how I constructed the models; however, they do differ with regard to the distribution used to model the independent variable, and the structure of the fixed effects (e.g. some models include interaction terms between different fixed effects, whereas others do not). Ultimately, the model that reflected the data in the most appropriate way was selected. The criteria that were used for model selection consisted of identifying the model with the lowest Akaike Information Criterion value, as well as comparing different models using an ANOVA, and selecting the model that performed significantly better than the others.

I acknowledge that my models do not paint a complete picture of the cellular processes contributing to genomic instability, but they do provide a valuable insight into the effects of APOBEC3 on genomic instability and mutation burden.

2.3.1 Mutation Burden

Models 2.1-2.3 investigate whether the number of APOBEC3 mutations and age affect the number of non-APOBEC3 mutations, after taking into account the effects of tumour type. In these models, α represents the intercept, and a is used to indicate a random intercept for each tumour type. β is used to represent the slope, and ϵ is used to represent the residual error term. The term "TT" is used as an abbreviation for "Tumour Type".

$$\mathrm{a}_i \sim N(0, \sigma^2)$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$log(\text{non-SBS2/SBS13}) \sim \text{Age} + (1|\text{Tumour Type}) \tag{2.1a}$$

$$log(\text{non-SBS2/13}) = \alpha + \beta_{Age} \cdot \text{Age} + \text{a}_{TT} + \epsilon \tag{2.1b}$$

Model family: Linear Mixed Effects Model

R package: lme4

$$log(\text{non-SBS2/SBS13}) \sim log(\text{SBS2/SBS13}) + (1|\text{Tumour Type}) \tag{2.2a}$$

$$log(\text{non-SBS2/13}) = \alpha + \beta_{SBS} \cdot log(\text{SBS213}) + \text{a}_{TT} + \epsilon \tag{2.2b}$$

Model family: Linear Mixed Effects Model

R package: lme4

$$log(\text{non-SBS2/SBS13}) \sim \text{Age} + log(\text{SBS2/SBS13}) + (1|\text{Tumour Type}) \tag{2.3a}$$

$$log(\text{non-SBS2/13}) = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS213}) + \text{a}_{TT} + \epsilon \tag{2.3b}$$

Model family: Linear Mixed Effects Model

R package: lme4

2.3.2 Genomic Instability

Equations 2.4-2.9 model the effects of age and the number of APOBEC3 mutations on different measures of genomic instability, after taking the effects of tumour type into account.

Throughout, α represents the intercept, and a is used to indicate a random intercept. β is used to represent the slope, and ϵ is used to represent the residual error term. The random intercept a and the residual error term ϵ are normally distributed (N) with mean 0 and variance σ^2. The term "SBS2/13" is used to represent the number of mutations belonging to mutational signatures SBS2 and SBS13, "TT" refers to tumour type. η represents the linear predictor in models based on the negative binomial (NB) distribution.

In models 2.8 and 2.9, the number of ID8 and ID6 mutations is modelled based on the zero-inflated negative binomial (ZINB) distribution, which is used to model overdispersed count data with excess zeros (Yau et al., 2003). These models contain both an NB and a logistic (logit) component. In the logistic component of the ZINB models, ξ represents the logit linked linear predictor, Θ is used to represent the intercept, θ is used to indicate a random intercept for each tumour type, δ is used to represent the slope, and E denotes the residual error term.

$$\text{PGA} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + (1|\text{Tumour Type}) \tag{2.4a}$$

$$\text{PGA} = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS213}) + \text{a}_{TT} + \epsilon \tag{2.4b}$$

Model family: Linear Mixed Effects Model

R package: lme4

$$\text{CN segments} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + (1|\text{Tumour Type}) \quad (2.5a)$$

$$\text{CNA segments} \sim NB(\mu, \text{k}) \quad (2.5b)$$

$$log(\mu) = \eta$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \text{a}_{TT} + \epsilon$$

Model family: Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{SVs} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + \text{Age} \times log(\text{SBS2/SBS13 mutations}) + (1|\text{Tumour Type}) \quad (2.6a)$$

$$\text{SVs} \sim NB(\mu, \text{k}) \quad (2.6b)$$

$$log(\mu) = \eta$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/SBS13}) + \beta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \text{a}_{TT} + \epsilon$$

Model family: Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{INDELs} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + \text{Age} \times log(\text{SBS2/SBS13 mutations}) + (1|\text{Tumour Type}) \tag{2.7a}$$

$$\text{INDELs} \sim NB(\mu, \text{k}) \tag{2.7b}$$

$$log(\mu) = \eta$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/SBS13}) + \beta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \text{a}_{TT} + \epsilon$$

Model family: Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{ID8} \sim \text{Age} + log(\text{SBS2/SBS13}) + \text{Age} \times log(\text{SBS2/SBS13}) + (1|\text{Tumour Type}) \tag{2.8a}$$

$$\text{ID8} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1 - p \end{cases}$$

$$log(\mu) = \eta \tag{2.8b}$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \delta_{SBS} \cdot log(\text{SBS2/13}) + \delta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{ID6} \sim \text{Age} + log(\text{SBS2/SBS13}) + \text{Age} \times log(\text{SBS2/SBS13}) + (1|\text{Tumour Type}) \quad (2.9a)$$

$$\text{ID6} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1-p \end{cases}$$

$$log(\mu) = \eta \quad (2.9b)$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \delta_{SBS} \cdot log(\text{SBS2/13}) + \delta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

2.3.3 Genomic Instability and TP53 Mutation

Models 2.10-2.15 model the relationship between different measures of genomic instability and the number of APOBEC3 mutations, age, and TP53 mutation status, accounting for the effects of tumour type. Similar to the section above, α represents the intercept, and a is used to indicate a random intercept. β is used to represent

the slope, and ϵ is used to represent the residual error term. The term "SBS2/13" is used to represent the number of mutations belonging to mutational signatures SBS2 and SBS13, "TT" refers to tumour type, and "TP53" refers to TP53 mutation status. η represents the linear predictor in models based on the negative binomial (NB) distribution.

In models 2.14 and 2.15, the number of ID8 and ID6 mutations is modelled based on the zero-inflated negative binomial (ZINB) distribution. In the logistic component of the ZINB models, ξ represents the logit linked linear predictor, Θ is used to represent the intercept, θ is used to indicate a random intercept for each tumour type, δ is used to represent the slope, and E denotes the residual error term.

$$\text{PGA} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + \text{TP53 status} + (1|\text{Tumour Type}) \tag{2.10a}$$

$$\text{PGA} = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{TP53} \cdot \text{TP53} + \text{a}_{TT} + \epsilon \tag{2.10b}$$

Model family: Linear Mixed Effects Model

R package: lme4

$$\begin{aligned}\text{CNA segments} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + \text{TP53 status} \\ + log(\text{SBS2/SBS13 mutations}) \times \text{TP53 status} + (1|\text{Tumour Type})\end{aligned} \tag{2.11a}$$

$$\text{CN segments} \sim NB(\mu, \text{k}) \tag{2.11b}$$

$$log(\mu) = \eta$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{TP53} \cdot \text{TP53} + \beta_{SBS \times TP53} \cdot log(\text{SBS2/13}) \times \text{TP53} + \text{a}_{TT} + \epsilon$$

Model family: Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{SVs} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + \text{TP53 status} + \text{Age} \times log(\text{SBS2/SBS13}) \tag{2.12a}$$
$$+ log(\text{SBS2/SBS13 mutations}) \times \text{TP53 status} + (1|\text{Tumour Type})$$

$$\text{SVs} \sim NB(\mu, \text{k}) \tag{2.12b}$$

$$log(\mu) = \eta$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{TP53} \cdot \text{TP53}$$
$$+ \beta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \beta_{SBS \times TP53} \cdot log(\text{SBS2/13}) \times \text{TP53} + \text{a}_{TT} + \epsilon$$

Model family: Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{INDELs} \sim \text{Age} + log(\text{SBS2/SBS13 mutations}) + \text{TP53 status} \tag{2.13a}$$
$$+ log(\text{SBS2/SBS13 mutations}) \times \text{TP53 status} + (1|\text{Tumour Type})$$

$$\text{INDELs} \sim NB(\mu, \text{k}) \tag{2.13b}$$

$$log(\mu) = \eta$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{TP53} \cdot \text{TP53} + \beta_{SBS \times TP53} \cdot log(\text{SBS2/13}) \times \text{TP53} + \text{a}_{TT} + \epsilon$$

Model family: Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{ID8} \sim \text{Age} + log(\text{SBS2/SBS13}) + \text{Age} \times log(\text{SBS2/SBS13}) + \text{TP53} + (1|\text{Tumour Type}) \tag{2.14a}$$

$$\text{ID8} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1 - p \end{cases}$$

$$log(\mu) = \eta \tag{2.14b}$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{TP53} \cdot \text{TP53} + \beta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \delta_{SBS} \cdot log(\text{SBS2/13}) + \delta_{TP53} \cdot \text{TP53} + \delta_{Age \times SBS} \cdot \text{Age} \times log(\text{SBS2/13}) + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{ID6} \sim \text{Age} + log(\text{SBS2/SBS13}) + \text{TP53} + log(\text{SBS2/13}) \times \text{TP53} + (1|\text{Tumour Type}) \tag{2.15a}$$

$$\text{ID6} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1 - p \end{cases}$$

$$log(\mu) = \eta \tag{2.15b}$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{SBS} \cdot log(\text{SBS2/13}) + \beta_{TP53} \cdot \text{TP53} + \beta_{SBS \times TP53} \cdot log(\text{SBS2/13}) \times \text{TP53} + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \delta_{SBS} \cdot log(\text{SBS2/13}) + \delta_{TP53} \cdot \text{TP53} + \delta_{SBS \times TP53} \cdot log(\text{SBS2/13}) \times \text{TP53} + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

2.3.4 Virus Models

Models 2.16-2.18 investigate the relationship between the number of APOBEC3 mutations in a sample, and virus presence, the type of viral genome present, and the virus genus present, accounting for the effects of tumour type. The following models use the same nomenclature as above, with the exception of the extra term γ_n, which represents a categorical variable with $n > 2$ levels, the subscript of which indicates the nth level.

$$\text{SBS2/SBS13 mutations} \sim \text{Virus Presence} + \text{Age} + (1|\text{Tumour Type}) \tag{2.16a}$$

$$\text{SBS2/SBS13 mutations} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1 - p \end{cases}$$

$$log(\mu) = \eta \tag{2.16b}$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \beta_{Virus} \cdot \text{Virus Presence} + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \beta_{Virus} \cdot \text{Virus Presence} + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{SBS2/SBS13 mutations} \sim \text{Virus Genome Type} + \text{Age} + (1|\text{Tumour Type}) \quad (2.17a)$$

$$\text{SBS2/SBS13 mutations} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1-p \end{cases}$$

$$log(\mu) = \eta \quad (2.17b)$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \gamma_{Genome} + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \gamma_{Genome} + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

$$\text{SBS2/SBS13 mutations} \sim \text{Virus Genus} + \text{Age} + (1|\text{Tumour Type}) \tag{2.18a}$$

$$\text{SBS2/SBS13 mutations} \sim \begin{cases} 0, \text{ with probability } p \\ NB(\mu, \text{k}), \text{ with probability } 1-p \end{cases}$$

$$log(\mu) = \eta \tag{2.18b}$$

$$\eta = \alpha + \beta_{Age} \cdot \text{Age} + \gamma_{Genus} + \text{a}_{TT} + \epsilon$$

$$logit(p) = \xi$$

$$\xi = \Theta + \delta_{Age} \cdot \text{Age} + \gamma_{Genus} + \theta_{TT} + E$$

Model family: Zero Inflated Negative Binomial Mixed Effects Model

R package: glmmTMB

2.3.5 CoxME Models

The function fitted by the CoxME package is summarised in Therneau (2020) as:

$$\lambda(t) = \lambda_0(t) e^{X\beta + Zb}$$

$$b \sim G(0, \Sigma(\theta))$$

Here, t represents time, X represents the design matrix for fixed effects, and Z

represents the design matrix for the random effects. $\lambda(t)$ represents the function that is being fit and $\lambda_0(t)$ represents a baseline hazard function. β represents a vector of coefficients for the fixed effects, and b represents a vector of coefficients for the random effects. G refers to the Gaussian distribution.

$$\text{Survival(Survival Time, Censoring Status)} \sim \text{SBS2/SBS13 Presence+(1|Tumour Type)} \tag{2.19a}$$

$$\text{Hazard}(t) = \text{Baseline Hazard Function}(t) \cdot e^{X \cdot \text{SBS} + Z \cdot \text{TT}} \tag{2.19b}$$

$$TT \sim G(0, \Sigma(\theta))$$

$$\text{Survival(Survival Time, Censoring Status)} \sim \text{SBS2/SBS13 Presence} * \text{TP53 Mutation} + (1|\text{Tumour Type}) \tag{2.20a}$$

$$\text{Hazard}(t) = \text{Baseline Hazard Function}(t) \cdot e^{X \cdot (\text{SBS and TP53}) + Z \cdot \text{TT}} \tag{2.20b}$$

$$TT \sim G(0, \Sigma(\theta))$$

2.4 Results

2.4.1 Number of APOBEC3 Mutations Correlates With Total Mutation Burden

I investigated the relationship between the number of classical APOBEC3 mutations (SBS2 and SBS13) and total mutation burden, excluding mutations attributed to APOBEC3. Of the 2,451 primary tumours that were investigated, 741 (30.2%) were found to harbour mutations attributed to the APOBEC3 mutation signatures. Tumours carrying APOBEC3 mutations were found across 26 of the 39 tumour types included in the PCAWG data set (Figure 2.1 and Table 2.4), and had a significantly higher mutation burden than tumours that did not carry APOBEC3 mutations (One sided Wilcoxon Rank Sum Test $p = 1.49 \times 10^{-26}$). Further, the number of APOBEC3 mutations was significantly correlated with the total mutation burden for 14 of the 22 tumour types that were surveyed (63.6%) for which there were at least three samples available to calculate Spearman correlation from (Figure 2.1 and Table 2.1), as previously observed in oral squamous cell carcinomas (Gillison et al., 2019).

I constructed a series of mixed effects models to investigate the relationship between the number of APOBEC3 mutations and age on the total number of non-APOBEC3 mutations in tumours carrying APOBEC3 mutations. After taking into account the effect of tumour type, both age and the number of classical APOBEC3 mutations were significant predictors of the number of non-APOBEC3 SNVs (Mixed effects model, $p = 2.26 \times 10^{-3}$ and $p = 2.27 \times 10^{-49}$, respectively. Table 2.2).

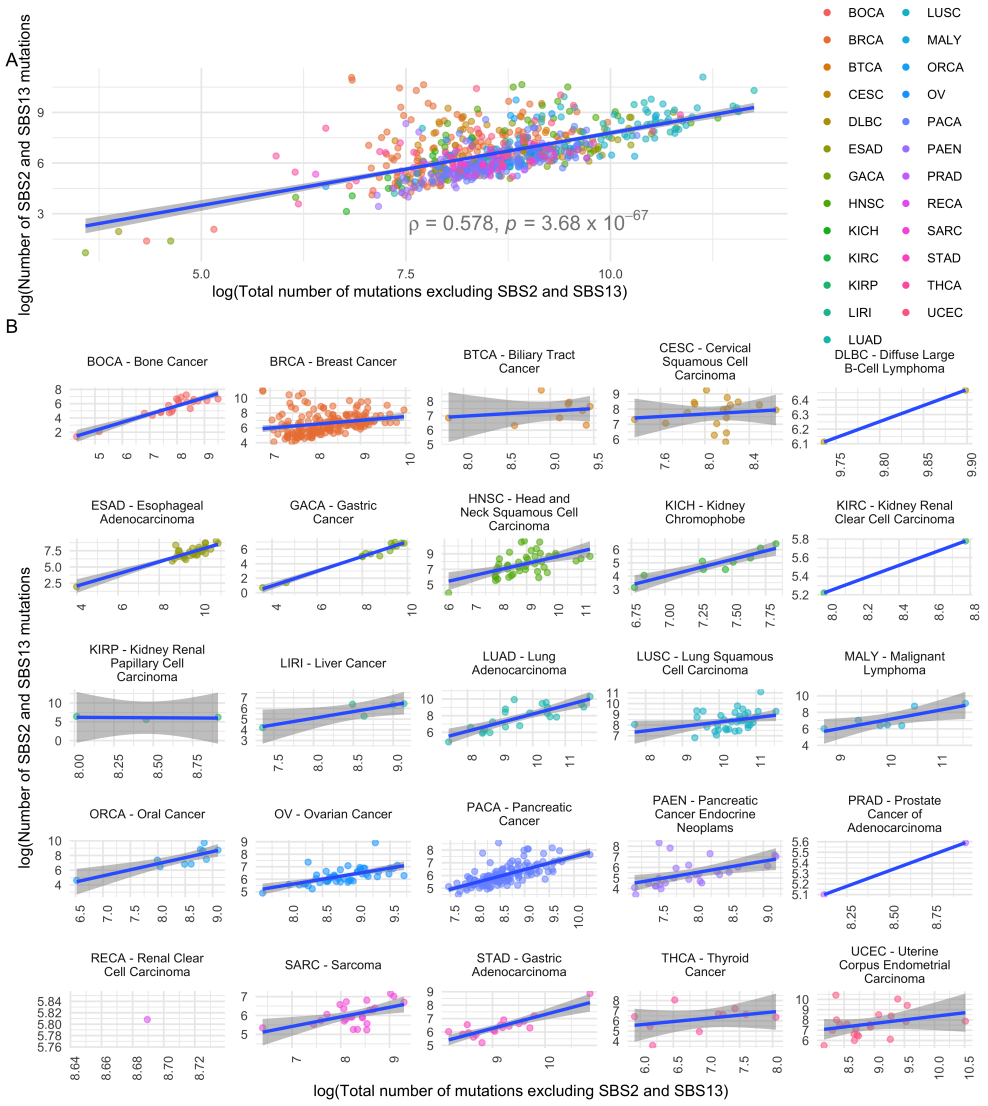


Figure 2.1: Correlation between number of SBS2 and SBS13 mutations and non-SBS2 and SBS13 mutations. Spearman correlation between the number of SBS2 and SBS13 SNVs and the total number of non-SBS2 and SBS13 SNVs for samples containing at least one SNV attributed to SBS2 and SBS13, coloured by tumour type and project code. The number of mutations was log transformed, using the natural logarithm. Shaded area represents the 95% confidence interval. Spearman's ρ and p-values for each of the correlations between the number of SBS2 and SBS13 and non-SBS2 and SBS13 SNVs by project code are presented in Table 2.1. (n = 741) Tumour type abbreviations are listed in Table D.1 in Appendix D.

Table 2.1: Spearman correlation coefficients for the correlation between log number of APOBEC3 mutations and the log number of non-APOBEC3 mutations by tumour type. p-values were corrected for multiple testing using the Benjamini-Hochberg procedure for controlling false discovery rate (FDR) at $\alpha = 0.05$. Tumour type abbreviations are listed in Table D.1 in Appendix D.

Project Code	n	ρ	p-value	Significance	Adjusted p-value
BLCA - Bladder Cancer	23	0.161	0.462	Not significant	0.535
BOCA - Bone Cancer	22	0.917	1.9×10^{-9}	Significant	1.05×10^{-8}
BRCA - Breast Cancer	185	0.247	6.87×10^{-4}	Significant	0.00151
BTCA - Biliary Tract Cancer	9	0.225	0.56	Not significant	0.616
CESC - Cervical Squamous Cell Carcinoma	18	0.132	0.602	Not significant	0.63
DLBC - Diffuse Large B-Cell Lymphoma	2				
ESAD - Esophageal Adenocarcinoma	32	0.881	2.8×10^{-11}	Significant	3.08×10^{-10}
GACA - Gastric Cancer	14	0.986	9.97×10^{-11}	Significant	7.31×10^{-10}
HNSC - Head and Neck Squamous Cell Carcinoma	43	0.535	2.21×10^{-4}	Significant	6.09×10^{-4}
KICH - Kidney Chromophobe	9	0.921	4.23×10^{-4}	Significant	0.00103
KIRC - Kidney Renal Clear Cell Carcinoma	2				
KIRP - Kidney Renal Papillary Cell Carcinoma	3	-0.247	0.841	Not significant	0.841
LIRI - Liver Cancer	5	0.883	0.0472	Not significant	0.0649
LUAD - Lung Adenocarcinoma	22	0.801	7.58×10^{-6}	Significant	2.78×10^{-5}
LUSC - Lung Squamous Cell Carcinoma	39	0.328	0.0413	Not significant	0.0606
MALY - Malignant Lymphoma	7	0.8	0.0308	Significant	0.0483
ORCA - Oral Cancer	10	0.84	0.00234	Significant	0.00468
OV - Ovarian Cancer	49	0.603	4.62×10^{-6}	Significant	2.03×10^{-5}
PACA - Pancreatic Cancer	145	0.676	1.08×10^{-20}	Significant	2.38×10^{-19}
PAEN - Pancreatic Cancer Endocrine Neoplasms	27	0.514	0.00615	Significant	0.0113
PRAD - Prostate Cancer of Adenocarcinoma	2				
RECA - Renal Clear Cell Carcinoma	1				
SARC - Sarcoma	24	0.528	0.00794	Significant	0.0134
STAD - Gastric Adenocarcinoma	18	0.834	1.66×10^{-5}	Significant	5.22×10^{-5}
THCA - Thyroid Cancer	10	0.354	0.316	Not significant	0.386
UCEC -Uterine Corpus Endometrial Carcinoma	20	0.299	0.201	Not significant	0.26

Table 2.2: Mixed Effects Models for Predicting the Number of non-APOBEC3 mutations. Models 1, 2, and 3 correspond to models 2.1, 2.2, and 2.3. Model 1 only considers the effects of age on the total mutation burden excluding APOBEC3 mutations. Model 2 considers only the effects of APOBEC3 mutations, and model 3 considers both the effects of age and the number of APOBEC3 mutations on the non-APOBEC3 mutation burden.

	Dependent variable:		
	log(Number of non-APOBEC3 mutations)		
	(1)	(2)	(3)
Age	Coefficient: 0.010 p = 3.76 x 10^{-6} *** 95% CI (0.006, 0.014)		Coefficient: 0.007 p = 2.26 x 10^{-3} ** 95% CI (0.003, 0.010)
log(Number of APOBEC3 mutations)		Coefficient: 0.342 p = 4.16 x 10^{-56} *** 95% CI (0.303, 0.381)	Coefficient: 0.349 p = 2.27 x 10^{-49} *** 95% CI (0.310, 0.388)
Constant	Coefficient: 8.12 p = 3.52 x 10^{-51} *** 95% CI (7.746, 8.499)	Coefficient: 6.415 p = 1.08 x 10^{-60} *** 95% CI (6.061, 6.770)	Coefficient: 5.996 p = 2.45 x 10^{-63} *** 95% CI (5.608, 6.384)
Observations	725	741	725
Log Likelihood	−831.379	−741.003	−703.723
Akaike Inf. Crit.	1670.759	1490.005	1417.445
Bayesian Inf. Crit.	1689.104	1508.437	1440.376
Note:			*p<0.05; **p<0.01; ***p<0.001

2.4.2 Presence of APOBEC3 Mutations is Associated With Increased Genomic Instability

It has previously been suggested that mutations assigned to signatures other than SBS2 and SBS13 may arise through further processing of deaminated cytosines by DNA repair enzymes, resulting in the generation of transitions, transversions, and double strand breaks (DSBs)(Burns et al., 2013). Errors in the repair of DSBs then result in mutations, as well as causing chromosomal rearrangements (Burns et al., 2013; Cannan and Pederson, 2016). Taking the number of mutations attributed to the SBS2 and SBS13 mutational signatures associated with APOBEC3 as an indicator of previous APOBEC3 activity, I investigated their effect on multiple measures of genomic instability.

I used the number of structural variants (SV), copy number (CN) segments, the percentage of the genome altered by copy number aberrations (PGA), and the num-

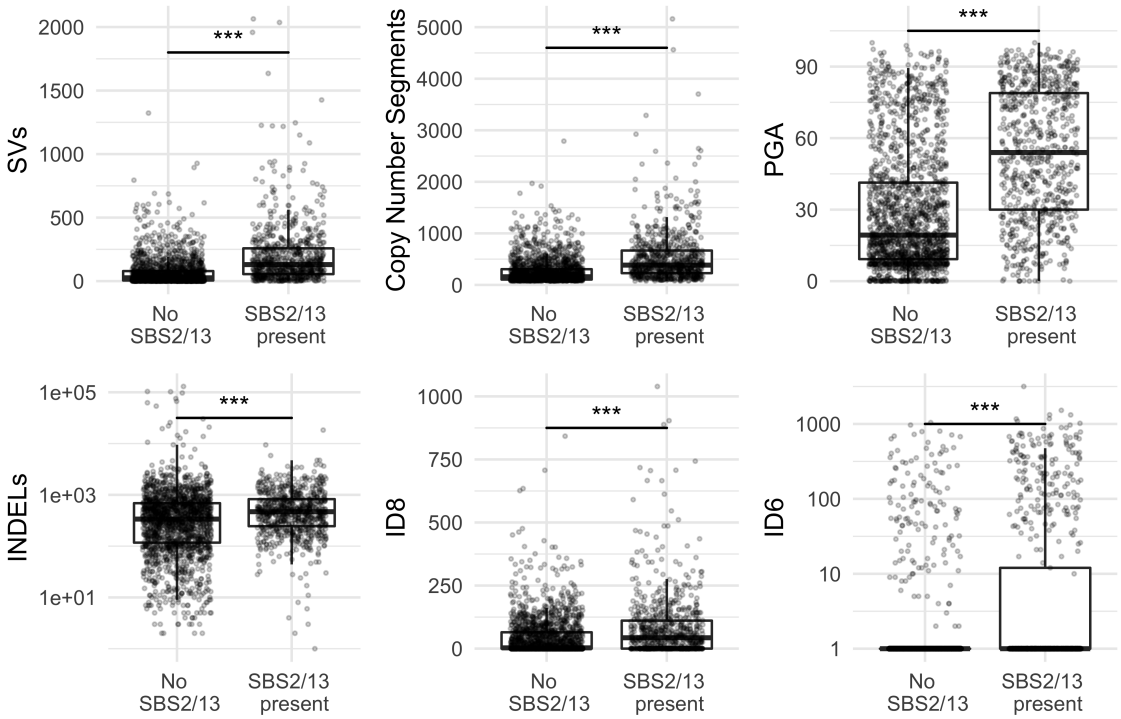


Figure 2.2: The effect of SBS2 and SBS13 presence on genomic instability Measures of genomic instability by presence of SBS2 and SBS13 related signatures. SV = Structural variant. PGA = Percentage of the Genome Altered. INDELs = Insertions and Deletions. ID8 = Insertion and deletion signature 8. ID6 = Insertion and deletion signature 6. (Wilcoxon Rank Sum test; * = p <0.05, ** = p <0.01, *** = p <0.001, n = 2,451 for INDELs, ID8, ID6, PGA and Copy Number segments. n = 2,427 for SVs.). Individual p-values are provided in Table 2.3.

ber of insertions and deletions as measures of genomic instability. I also examined the number of insertions and deletions (INDELs) attributed to INDEL signatures 6 and 8 (ID6 and ID8), which have been associated with non-homologous end joining (NHEJ) of double strand breaks (DSBs) (Alexandrov et al., 2020). For all six of the genome instability measures that were considered, samples carrying APOBEC3 mutations had significantly higher values than samples with no APOBEC3 mutations (Wilcoxon Rank Sum test p <0.001; Figure 2.2 and Table 2.3).

Table 2.3: Wilcoxon rank sum test, comparing the levels of each of the genomic instability measures between samples containing SBS2 and SBS13 mutations, and those not containing SBS2 and SBS13 mutations. Results are represented visually in Figure 2.2.

Measure	p-value	Significance	Adjusted p-value
SVs	8.71×10^{-100}	Significant	5.22×10^{-99}
Copy Number Segments	3.56×10^{-90}	Significant	1.07×10^{-89}
PGA	1.15×10^{-86}	Significant	2.30×10^{-86}
INDELs	7.71×10^{-15}	Significant	7.71×10^{-15}
ID8	4.01×10^{-20}	Significant	4.81×10^{-20}
ID6	7.74×10^{-34}	Significant	1.16×10^{-33}

2.4.3 The Number of APOBEC3 Mutations Predicts the Level of Genomic Instability Across Multiple Tumour Types

I constructed mixed effects models to investigate whether the number of APOBEC3 mutations could be used to predict the levels of the instability measures, taking both age and tumour type into account. The models show that tumours carrying APOBEC3 mutations are more genomically unstable and that the number of APOBEC3 mutations is associated with all measures of genomic instability, except the number of ID6 INDELs (Table 2.5). Age had a significant, predictive effect for the total number of INDELs and the number of structural variants ($p = 9.61 \times 10^{-6}$ and $p = 0.0151$, respectively).

Comparing the median values for each of the six measures within a given tumour type highlighted several tumour types in which the presence of APOBEC3 mutations had a strong effect on genomic instability (Figure 2.3). When individual measures of genomic instability are considered, 13 of the 24 tumour types (54.2%) had significant association between presence of APOBEC3 mutations and a measure of genomic instability. Specifically, higher levels of genomic instability were observed across mul-



Figure 2.3: The effect of SBS2 and SBS13 presence on genomic instability by tumour type Ratio of the median value of each measure of genomic instability for tumours containing SBS2 and SBS13 mutation to those that do not contain SBS2 and SBS13 mutations. p-values were derived from one-sided Wilcoxon rank sum tests, and the horizontal grey lines indicates an FDR of 0.05, which includes points that fall on the line. The number of samples in which SBS2 and SBS13 mutations are present and absent are reported for each tumour type in Table 2.4. Tumour type abbreviations are listed in Table D.1 in Appendix D.

Table 2.4: The number of samples belonging to each tumour type that either contain SBS2 and SBS13 mutations, or do not contain SBS2 and SBS13 mutations. Tumour type abbreviations are listed in Table D.1 in Appendix D.

Tumour type	Number of samples with SBS2 and SBS13 mutations	Number of samples without SBS2 and SBS13 mutations	Total number of samples
BLCA - Bladder Cancer	23	0	23
BOCA - Bone Cancer	22	39	61
BRCA - Breast Cancer	185	24	209
BTCA - Biliary Tract Cancer	9	3	12
CESC - Cervical Squamous Cell Carcinoma	18	2	20
CLLE - Chronic Lymphocytic Leukemia	0	90	90
CMDI - Chronic Myeloid Disorders	0	30	30
COAD - Colon Adenocarcinoma	0	37	37
DLBC - Diffuse Large B-Cell Lymphoma	2	5	7
EOPC - Early Onset Prostate Cancer	0	41	41
ESAD - Esophageal Adenocarcinoma	32	65	97
GACA - Gastric Cancer	14	18	32
GBM - Glioblastoma Multiforme	0	30	30
HNSC - Head and Neck Squamous Cell Carcinoma	43	0	43
KICH - Kidney Chromophobe	9	34	43
KIRC - Kidney Renal Clear Cell Carcinoma	2	35	37
KIRP - Kidney Renal Papillary Cell Carcinoma	3	29	32
LAML - Acute Myeloid Leukemia	0	8	8
LGG - Low Grade Glioma	0	18	18
LICA - Liver Cancer	0	5	5
LIHC - Liver Hepatocellular Carcinoma	0	53	53
LINC - Liver Cancer	0	28	28
LIRI - Liver Cancer	5	245	250
LUAD - Lung Adenocarcinoma	22	15	37
LUSC - Lung Squamous Cell Carcinoma	39	8	47
MALY - Malignant Lymphoma	7	93	100
ORCA - Oral Cancer	10	3	13
OV - Ovarian Cancer	49	52	101
PACA - Pancreatic Cancer	145	85	230
PAEN - Pancreatic Cancer Endocrine Neoplasms	27	54	81
PBCA - Pediatric Brain Cancer	0	230	230
PRAD - Prostate Cancer of Adenocarcinoma	2	151	153
READ - Rectum Adenocarcinoma	0	15	15
RECA - Renal Clear Cell Carcinoma	1	73	74
SARC - Sarcoma	24	10	34
SKCM - Skin Cutaneous Melanoma	0	2	2
STAD - Gastric Adenocarcinoma	18	18	36
THCA - Thyroid Cancer	10	38	48
UCEC - Uterine Corpus Endometrial Carcinoma	20	24	44

Table 2.5: Mixed effects models predicting the levels of six different measures of instability using age, the number of SBS2 and SBS13 mutations, accounting for the effects of tumour type as a random variable. The number of SBS2 and SBS13 mutations was log transformed using the natural logarithm. These models correspond to equations 2.4-2.9, detailed in Section 2.3. (SV = Structural Variant, CNV = Copy Number Variation, PGA = Proportion of the Genome Altered, INDELs = Insertions and Deletions, ID8 = INDEL signature 8, ID6 = INDEL signature 6, LMM = Linear Mixed Effects Model, NB = Negative Binomial, ZINB = Zero Inflated Negative Binomial)

	Dependent variable:						
		PGA	CN Segments	SVs	INDELs	ID8	ID6
	Model Type	LMM	NB	NB	NB	ZINB	ZINB
Count model: (Intercept)	Coefficient	3.62	4.42	0.397	1.36	0.514	5.46
	95% CI	$(-9.49, 16.7)$	$(4.01, 4.83)$	$(-1.26, 2.05)$	$(0.257, 2.46)$	$(-1.52, 2.55)$	$(1.30, 9.63)$
	p-value	0.589	$3.62\text{x}10^{-99***}$	0.638	0.0157^{*}	0.620	0.0102^{*}
Count model: Age	Coefficient	0.0634	-0.000362	0.0323	0.0393	0.0250	-0.0562
	95% CI	$(-0.0731, 0.200)$	$(-0.00370, 0.00298)$	$(0.00624, 0.0583)$	$(0.0219, 0.0567)$	$(-0.00703, 0.0570)$	$(-0.121, 0.0083)$
	p-value	0.363	0.832	0.0151^{*}	$9.61\text{x}10^{-6***}$	0.126	0.0877
Count model: log(SBS2/13)	Coefficient	6.50	0.247	0.684	0.725	0.589	0.0191
	95% CI	$(4.99, 8.01)$	$(0.206, 0.289)$	$(0.442, 0.927)$	$(0.562, 0.888)$	$(0.281, 0.897)$	$(-0.578; 0.616)$
	p-value	$2.51\text{x}10^{-16***}$	$5.66\text{x}10^{-32***}$	$3.14\text{x}10^{-8***}$	$3.43\text{x}10^{-18***}$	$1.77\text{x}10^{-4***}$	0.950
Count model: Age:log(SBS2/13)	Coefficient			-0.00515	-0.00552	-0.00372	0.00738
	95% CI			(-0.00904,-0.00126)	$(-0.00812, -0.00292)$	$(-0.00854, 0.00110)$	(-0.0019,0.0166)
	p-value			0.00942^{**}	$3.14\text{x}10^{-5***}$	0.130	0.118
Zero model: (Intercept)	Coefficient					-0.684	1.07
	95% CI					$(-0.839, -0.530)$	$(0.903, 1.24)$
	p-value					$4.20\text{x}10^{-18***}$	$3.85\text{x}10^{-36***}$
AIC		6738.4	10064.1	8701.2	10444.3	6357.3	3270.9
Log Likelihood		-3364.2	-5027.1	-4344.6	-5216.1	-3171.6	-1628.5
Num. obs.		725	725	717	725	725	725
Num. Tumour Types		26	26	26	26	26	26

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

tiple measures in tumours that contained APOBEC3 mutations than those that did not for both pancreatic cancer subtypes (Pancreatic Cancer Endocrine Neoplasms (PAEN) and Pancreatic Cancer (PACA)), Bone Cancer (BOCA), Kidney Renal Papillary Cell Carcinoma (KIRP), and Malignant Lymphoma (MALY). In addition, significant associations were observed for a single measure of genomic instability for Breast Cancer (BRCA), Lung Adenocarcinoma (LUAD), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Chromophobe (KICH), Gastric Adenocarcinoma (STAD), Uterine Corpus Endometrial Carcinoma (UCEC), Sarcoma (SARC), and Prostate Cancer of Adenocarcinoma (PRAD). When p-values were combined for all measures of genomic instability, a further 2 tumour types, Biliary Tract Cancer (BTCA) and Cervical Squamous Cell Carcinoma (CESC), showed significant association between presence of APOBEC3 mutations and GI (62.5%, Fisher's combined probability test with Benjamini-Hochberg correction for multiple testing p-value <0.05; Table 2.6).

2.4.4 Both Presence of APOBEC3 Mutations and TP53 Mutation Affect Genome Stability

Several studies have found that activity of APOBEC3 proteins is closely linked with p53 activity, with p53 acting as a negative regulator of APOBEC3B activity (Nikkilä et al., 2017; Periyasamy et al., 2017). In addition, APOBEC3 activity has been associated with mutations in the TP53 gene (Burns et al., 2013). To further investigate this link, I built new models adding the effects of TP53 alterations.

The proportion of tumours carrying missense or nonsense mutations in TP53 was significantly higher in tumours carrying APOBEC3 mutations (41.6%) than in tumours not carrying any APOBEC3 mutations (19.9%; one-sided Fisher exact test, $p = 9.91 \times 10^{-28}$). Tumours carrying missense or nonsense mutations in TP53 also had a higher number of APOBEC3 mutations (One-sided Wilcoxon Rank Sum test, $p = 4.84 \times 10^{-33}$), as well as a higher non-APOBEC3 mutation burden (One-sided

Table 2.6: Combined p-values using Fisher's combined probability test. p-values were combined for the different measures of genomic instability within each tumour type, and the resulting Fisher method p-values were corrected for multiple testing using the Benjamini-Hochberg procedure for controlling false discovery rate (FDR) at $\alpha = 0.05$. Tumour type abbreviations are listed in Table D.1 in Appendix D.

Tumour Type	Number of GI measures	Fisher method p-value	Significance	Adjusted p-value
BOCA - Bone Cancer	6	3.14×10^{-5}	Significant	9.42×10^{-5}
BRCA - Breast Cancer	6	7.77×10^{-8}	Significant	3.73×10^{-7}
BTCA - Biliary Tract Cancer	5	9.13×10^{-3}	Significant	0.0183
CESC - Cervical Squamous Cell Carcinoma	5	0.0151	Significant	0.0241
DLBC - Diffuse Large B-Cell Lymphoma	6	0.975	Not significant	0.975
ESAD - Esophageal Adenocarcinoma	6	0.298	Not significant	0.358
GACA - Gastric Cancer	6	0.202	Not significant	0.262
KICH - Kidney Chromophobe	6	0.0101	Significant	0.0186
KIRC - Kidney Renal Clear Cell Carcinoma	6	1.92×10^{-4}	Significant	4.19×10^{-4}
KIRP - Kidney Renal Papillary Cell Carcinoma	6	3.17×10^{-7}	Significant	1.27×10^{-6}
LIRI - Liver Cancer	6	0.0638	Not significant	0.0958
LUAD - Lung Adenocarcinoma	6	2.95×10^{-5}	Significant	9.423×10^{-5}
LUSC - Lung Squamous Cell Carcinoma	6	0.944	Not significant	0.975
MALY - Malignant Lymphoma	6	5.76×10^{-10}	Significant	4.61×10^{-9}
ORCA - Oral Cancer	6	0.922	Not significant	0.975
OV - Ovarian Cancer	6	0.206	Not significant	0.262
PACA - Pancreatic Cancer	6	3.47×10^{-13}	Significant	8.34×10^{-12}
PAEN - Pancreating Cancer Endocrine Neoplasms	6	9.75×10^{-11}	Significant	1.17×10^{-9}
PRAD - Prostate Cancer of Adenocarcinoma	6	7.29×10^{-5}	Significant	1.94×10^{-4}
RECA - Renal Clear Cell Carcinoma	6	0.207	Not significant	0.262
SARC - Sarcoma	5	1.03×10^{-4}	Significant	0.000247
STAD - Gastric Adenocarcinoma	6	0.0130	Significant	0.0223
THCA - Thyroid Cancer	5	0.545	Not significant	0.623
UCEC - Uterine Corpus Endometrial Carcinoma	6	3.67×10^{-9}	Significant	2.20×10^{-8}

Table 2.7: Mixed Effects Models Predicting the levels of six different measures of instability using the log number of SBS2 and SBS13 mutations and TP53 mutation status, as well as accounting for the effects of tumour type as a random variable. The number of SBS2 and SBS13 mutations was log transformed using the natural logarithm. These models correspond to equations 2.10-2.15, detailed in Section 2.3. (SV = Structural Variant, CNV = Copy Number Variation, PGA = Proportion of the Genome Altered, INDELs = Insertions and Deletions, ID8 = INDEL signature 8, ID6 = INDEL signature 6, LMM = Linear Mixed Effects Model, NB = Negative Binomial, ZINB = Zero Inflated Negative Binomial)

		Dependent variable:					
		PGA	CN Segments	SVs	INDELs	ID8	ID6
	Model Type	LMM	NB	NB	NB	ZINB	ZINB
Count model: (Intercept)	Coefficient	5.37	4.44	0.257	3.16	0.581	1.62
	95% CI	$(-7.32, 18.1)$	$(4.03, 4.84)$	$(-1.36, 1.87)$	$(2.67, 3.65)$	$(-1.44; 2.61)$	$(0.355, 2.88)$
	p-value ‘	0.408	$1.16\text{x}10^{-100***}$	0.755	$3.04\text{x}10^{-41***}$	0.574	0.0121^{*}
Count model: Age	Coefficient	0.0490	-0.000657	0.0281	0.00315	0.0235	-0.00354
	95% CI	$(-0.084, 0.182)$	$(-0.00398, 0.00267)$	$(0.00250, 0.0537)$	$(-0.00049, 0.00679)$	$(-0.00840, 0.0554)$	$(-0.0130, 0.00596)$
	p-value	0.472	0.699	0.0314^{*}	0.0903	0.149	0.465
Count model:	Coefficient	5.83	0.241	0.704	0.449	0.575	0.577
log(SBS2/13)	95% CI	$(4.34, 7.32)$	$(0.200, 0.282)$	$(0.468, 0.940)$	$(0.391, 0.507)$	$(0.268, 0.882)$	$(0.419, 0.735)$
	p-value	$7.55\text{x}10^{-14***}$	$1.20\text{x}10^{-30***}$	$5.00\text{x}10^{-9***}$	$3.03\text{x}10^{-45***}$	$2.44\text{x}10^{-4***}$	$8.07\text{x}10^{-13***}$
Count model: TP53	Coefficient	11.0	0.131	1.26	1.08	0.123	2.16
	95% CI	$(7.15, 14.9)$	$(0.039, 0.224)$	$(0.502, 2.02)$	$(0.502, 1.65)$	$(-0.024, 0.271)$	$(0.258, 4.07)$
	p-value	$3.70\text{x}10^{-8***}$	$5.27\text{x}10^{-3**}$	0.00114^{**}	$2.37\text{x}10^{-4***}$	0.102	0.0260^{*}
Count model:	Coefficient			-0.156	-0.158		-0.347
log(SBS2/13):TP53	95% CI			$(-0.268, -0.045)$	$(-0.241, -0.0742)$		$(-0.622, -0.0729)$
	p-value			0.00610^{**}	$2.17\text{x}10^{-4***}$		0.0131^{*}
Count model:	Coefficient			-0.00462		-0.00354	
Age:log(SBS2/13)	95% CI			(-0.00845,-0.000799)		$(-0.00834, 0.00126)$	
	p-value			0.0178^{*}		0.148	
Zero model: (Intercept)	Coefficient					-0.684	1.07
	95% CI					$(-0.839, -0.530)$	$(0.902, 1.24)$
	p-value					$4.19\text{x}10^{-18***}$	$4.23\text{x}10^{-36***}$
AIC		6710.3	10058.3	8687.9	10448.8	6356.6	3266.8
Log Likelihood		-3349.1	-5023.2	-4335.9	-5217.4	-3170.3	-1625.4
Num. obs.		725	725	717	725	725	725
Num. Tumour Types		26	26	26	26	26	26

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Wilcoxon Rank Sum test, p = 3.53 x 10^{-67}).

Adding the TP53 mutation status of the tumours to the mixed effects models generated in the previous section suggests that TP53 mutation is a significant predictor of the genomic instability measures, with the exception of the number of ID8 INDELs (Table 2.7). Importantly, the number of APOBEC3 mutations remained a highly significant predictor throughout, and also emerged as a significant predictor for the number of ID6 INDELs. For PGA, the number of copy number segments, the number of structural variants, and the number of ID6 INDELs, including TP53 in the model improved it significantly, but not for the total number of INDELs or ID8 INDELs (ANOVA p <0.05, Table 2.8). The effects of age on the measures of genomic instability remained non-significant, with the exception of the effects of age on the number of structural variants (Table 2.7).

Table 2.8: Akaike Information Criteria (AIC) and p-values from ANOVAs comparing models with (equations 2.4-2.9) and without TP53 status (equations 2.10-2.15)for each GI measure. Full models are specified in Section 2.3.

GI Measure	p-value	AIC without TP53	AIC with TP53
PGA	4.02 x 10^{-8}	6738	6710
Copy Number Segments	5.32 x 10^{-3}	10064	10058
Structural Variants	1.75 x 10^{-4}	8701	8687
INDELs	1.00	10444	10448
ID8	0.101	6357	6356
ID6	0.0133	3270	3266

I also investigated the effect of TP53 mutation and APOBEC3 mutations on overall survival by constructing Cox proportional hazards models combined with mixed effects models, taking the effects of tumour type into account (CoxME models). When presence of APOBEC3 mutations is considered alone it does not have a significant effect on survival (p = 0.129, Hazard Ratio = 1.18; Table 2.9).

However, when TP53 mutation status is included APOBEC3 mutations increase

Table 2.9: Coxme model using only SBS2 and SBS13 presence to predict survival (n = 1,492). This model corresponds to equation 2.3.5 in Section 2.3.

Random Effects	N	Standard Deviation	Variance		
Tumour Type	36	1.14	1.31		
Fixed Effects	Coefficient	Hazard Ratio	95% CI	z	p
SBS2/13 presence	0.162	1.18	(0.95,1.45)	1.52	0.127

Table 2.10: CoxME model using SBS2 and SBS13 presence, p53 mutation status, and the interaction between them as predictors of survival. This model corresponds to equation 2.3.5 in Section 2.3. (n = 1,492)

Random Effects	N	Standard Deviation	Variance		
Tumour Type	36	1.14	1.31		
Fixed Effects	Coefficient	Hazard Ratio	95% CI	z	p
SBS2/13 Presence	0.366	1.44	(1.08,1.92)	2.49	0.0128*
TP53 Mutation	0.384	1.47	(1.14,1.90)	2.95	0.00318**
SBS2/13 Presence *TP53 mutation	-0.362	0.697	(0.487,0.996)	-1.98	0.0477*

the hazard ratio when TP53 is unmutated, and TP53 mutation significantly increases the hazard ratio when APOBEC3 mutations are not present, negatively affecting survival in both cases (APOBEC3 mutation presence p = 0.0128, Hazard Ratio = 1.44; TP53 mutation p = 0.00318, Hazard Ratio = 1.47; 2.10). The interaction between APOBEC3 mutation presence and TP53 mutation was also significant (p = 0.0477, Hazard Ratio = 0.697), but had a hazard ratio below 1, suggesting that the effect of APOBEC3 mutation presence on survival is less pronounced when combined with TP53 mutation, and vice versa, indicating that the two effects are not additive (Figure 2.4).

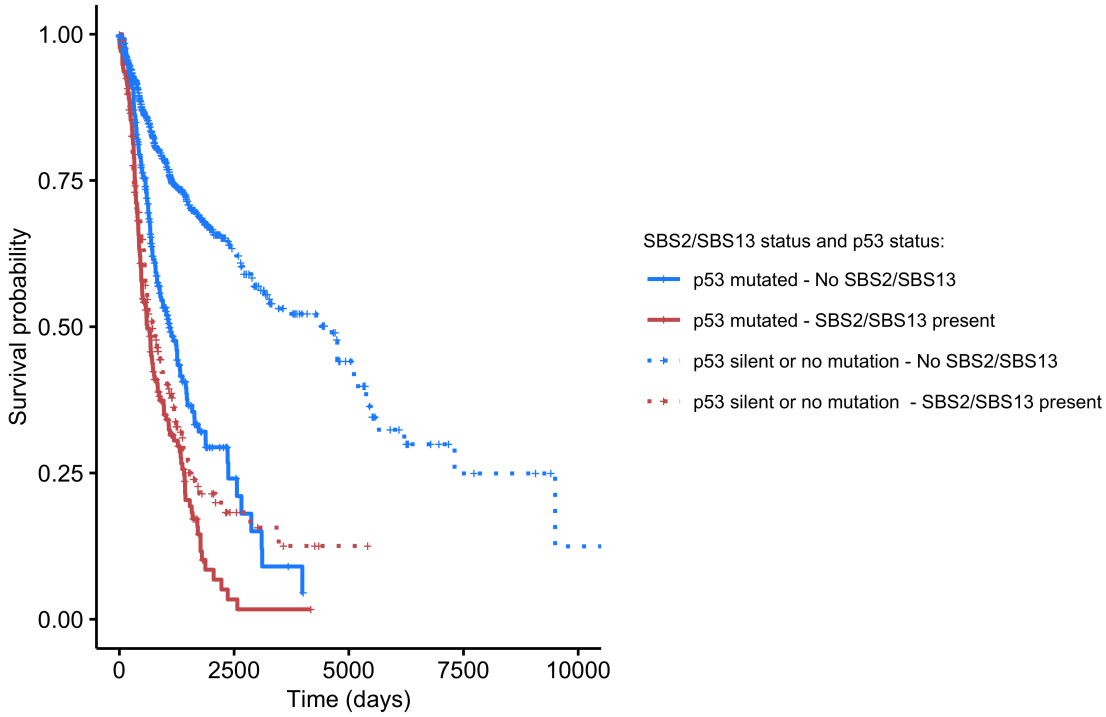


Figure 2.4: Overall survival stratified by presence of SBS2 and SBS13 mutations and p53 mutation status. Kaplan-Meier overall survival for all patients for which survival data was available. (n = 1,492)

2.4.5 The Number of Non-Kataegis APOBEC3 Mutations is Associated With Increased Genomic Instability

To address whether the results of the models could be attributed to processes such as kataegis, in which APOBEC3s act on single stranded DNA byproducts of DNA damage repair rather than causing strand breaks themselves, I reconstructed my models excluding SNVs attributed to kataegis events involving APOBEC3 mutations (described in (Campbell et al., 2020)). Excluding APOBEC3 mutations associated with kataegis did not appreciably alter my conclusions. I found that the number of APOBEC3 mutations, excluding those attributed to kataegis, remained a significant predictor for each of the measures of genomic instability when the effects of TP53 mutation were accounted for (Tables 2.11 and 2.12). This strongly suggests that APOBEC3s may play an active role in the generation of widespread and diverse genomic instability.

Table 2.11: Mixed Effects Models Predicting the levels of six different measures of instability using the log number of SBS2 and SBS13 mutations excluding those attributed to kataegis, accounting for the effects of tumour type as a random variable. These models correspond to equations 2.4-2.9, detailed in Section 2.3.

	Dependent variable:						
		PGA	CN Segments	SVs	INDELs	ID8	ID6
	Model Type	LMM	NB	NB	NB	ZINB	ZINB
Count model: (Intercept)	Coefficient	7.67	4.72	1.33	1.65	0.233	5.19
	95% CI	$(-5.05, 20.4)$	$(4.31, 5.13)$	$(-0.220, 2.88)$	$(0.613, 2.68)$	$(-1.54, 2.01)$	$(1.18, 9.21)$
	p-value	0.238	$3.38\text{x}10^{-114***}$	0.0926	$1.79\text{x}10^{-03**}$	0.796	0.0112^{*}
Count model: Age	Coefficient	0.0687	$-2.37x10^{-7}$	0.0265	0.0381	0.0319	-0.0522
	95% CI	$(-0.0684, 0.206)$	$(-0.00341, 0.00341)$	$(0.00257, 0.0503)$	$(0.0219, 0.0542)$	$(0.00426, 0.0596)$	$(-0.114, 0.00976)$
	p-value	0.326	1.00	0.0299^{*}	$3.95\text{x}10^{-6***}$	0.024^{*}	0.0987
Count model:	Coefficient	5.97	0.204	0.554	0.696	0.645	0.056
log(SBS2/13 - kataegis)	95% CI	$(4.52, 7.42)$	$(0.165, 0.244)$	$(0.324, 0.783)$	$(0.540, 0.852)$	$(0.370, 0.920)$	$(-0.523, 0.636)$
	p-value	$3.67\text{x}10^{-15***}$	$2.35\text{x}10^{-24***}$	$2.21\text{x}10^{-6***}$	$2.83\text{x}10^{-9***}$	$4.19\text{x}10^{-6***}$	0.849
Count model: Age:	Coefficient			-0.00428	-0.00545	-0.00488	0.00686
log(SBS2/13 - kataegis)	95% CI			(-0.00791,-0.000651)	$(-0.00790, -0.00299)$	$(-0.00914, -0.000621)$	(-0.0021,0.0158)
	p-value			0.0208^{*}	$1.41\text{x}10^{-5***}$	0.0247^{*}	0.134
Zero model: (Intercept)	Coefficient					-0.682	1.07
	95% CI					$(-0.837, -0.528)$	$(0.901, 1.23)$
	p-value					$5.40\text{x}10^{-18***}$	$5.42\text{x}10^{-36***}$
AIC		6735.3	10081.1	8722.8	10430.9	6339.6	3267.7
BIC		6758.2	10104.0	8750.2	10458.4	6371.7	3299.8
Log Likelihood		-3362.6	-5035.6	-4355.4	-5209.5	-3162.8	-1626.9
Num. obs.		724	724	716	724	724	724
Num. Tumour Types		26	26	26	26	26	26

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2.12: Mixed Effects Models Predicting the levels of six different measures of instability using the log number of SBS2 and SBS13 mutations excluding those attributed to kataegis, and TP53 mutation status, accounting for the effects of tumour type as a random variable. These models correspond to equations 2.10-2.15, detailed in Section 2.3.

		Dependent variable:					
		PGA	CN Segments	SVs	INDELs	ID8	ID6
	Model Type	LMM	NB	NB	NB	ZINB	ZINB
(Intercept)	Coefficient	8.97	4.73	1.23	3.42	0.313	1.60
	95% CI	$(-3.32, 21.3)$	$(4.32, 5.13)$	$(-0.292, 2.75)$	$(2.96, 3.89)$	$(-1.46, 2.09)$	$(0.362, 2.83)$
	p-value	0.154	$2.29\text{x}10^{-115***}$	0.113	$3.16\text{x}10^{-47***}$	0.729	0.0113^{*}
Count model: Age	Coefficient	0.0536	−0.000350	0.0214	0.00310	0.0301	−0.00366
	95% CI	$(-0.0803, 0.188)$	$(-0.00375, 0.00305)$	$(-0.00234, 0.0452)$	$(-0.000557, 0.00676)$	$(0.00234, 0.0578)$	$(-0.0131, 0.00574)$
	p-value	0.433	0.840	0.0772	0.0966	0.0335^{*}	0.445
Count model:	Coefficient	5.35	0.199	0.568	0.419	0.628	0.584
log(SBS2/13 - kataegis)	95% CI	$(3.93, 6.78)$	$(0.160, 0.239)$	$(0.343, 0.793)$	$(0.365, 0.473)$	$(0.353, 0.903)$	$(0.430, 0.737)$
	p-value	$6.78\text{x}10^{-13***}$	$1.73\text{x}10^{-23***}$	$7.64\text{x}10^{-7***}$	$8.25\text{x}10^{-52***}$	$7.49\text{x}10^{-6***}$	$9.32\text{x}10^{-14***}$
Count model: TP53	Coefficient	11.2	0.148	1.13	0.957	0.130	2.20
	95% CI	$(7.34, 15.1)$	$(0.0540, 0.242)$	$(0.449, 1.82)$	$(0.440, 1.47)$	$(-0.0172, 0.277)$	$(0.353, 4.05)$
	p-value	$2.27\text{x}10^{-8***}$	$2.02\text{x}10^{-3**}$	$1.17\text{x}10^{-3**}$	$2.85\text{x}10^{-4***}$	0.0835	0.0196^{*}
Count model: log(SBS2/13	Coefficient			−0.136	−0.141		−0.356
- kataegis):TP53	95% CI			$(-0.238, -0.0344)$	$(-0.218, -0.0650)$		$(-0.624, -0.0870)$
	p-value			$8.73\text{x}10^{-3**}$	$2.89\text{x}10^{-4***}$		$9.48\text{x}10^{-3**}$
Count model: Age:	Coefficient			-0.00365		−0.00464	
log(SBS2/13 - kataegis)	95% CI			(-0.00726,-0.0000348)		$(-0.00890, -0.000380)$	
	p-value			0.0478^{*}		0.0328^{*}	
Zero model: (Intercept)	Coefficient					−0.682	1.07
	95% CI					$(-0.837, -0.528)$	$(0.900, 1.23)$
	p-value					$5.39\text{x}10^{-18***}$	$6.03\text{x}10^{-36***}$
AIC		6706.2	10073.6	8708.5	10437.4	6338.6	3262.9
BIC		6733.7	10101.1	8745.0	10469.5	6375.2	3299.6
Log Likelihood		−3347.1	−5030.8	−4346.2	−5211.7	−3161.3	−1623.4
Num. obs.		678	724	716	724	724	724
Num. Tumour Types		26	26	26	26	26	26

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2.13: Mixed effects models predicting the number of SBS2 and SBS13 mutations based on virus presence and age, accounting for the effects of tumour type as a random variable. This model is represented by Equation 2.16.

		Dependent variable:
		SBS2 and SBS13 mutations
Count model: (Intercept)	Coefficient	1.91
	95% CI	$(-0.572, 4.40)$
	p-value	0.131
Count model: Virus Presence	Coefficient	0.0923
	95% CI	$(-0.184, 0.368)$
	p-value	0.512
Count model: Age	Coefficient	0.00790
	95% CI	$(0.000433, 0.0154)$
	p-value	0.0381^{*}
Zero model: (Intercept)	Coefficient	0.448
	95% CI	$(0.351, 0.545)$
	p-value	$1.69 \times 10^{-19***}$
AIC		13973.64
Log Likelihood		−6980.82
Num. obs.		2328
Num. Tumour Types		37

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

2.4.6 Associations Between Presence of Viral DNA and APOBEC3 Mutations

Activation of APOBEC3 enzymes following viral infection has been described for many viruses (Section 1.4.5). I sought to investigate whether the presence of viral DNA in a tumour might be used to predict the presence of APOBEC3 mutations, and by extension, whether viral infection could be established as a potential cause of APOBEC3 mutations and subsequent genomic instability in tumours. I constructed mixed effects models to predict the number of APOBEC3 mutations in a tumour using either virus presence, virus genome type, and virus genus as predictive factors, and accounting for the effects of tumour type.

I found that the age of the patient at diagnosis was a significant predictor for the number of APOBEC3 mutations in all models ($p < 0.05$, Tables 2.13, 2.14 and 2.15). Virus presence did not have a significant effect on the number of APOBEC3 mutations ($p = 0.512$, Table 2.13). Viruses with partially double stranded DNA genomes, characteristic of the Orthohepadnaviruses, were significant negative pre-

Table 2.14: Mixed effects models predicting the number of SBS2 and SBS13 mutations based on the virus genome present and age, accounting for the effects of tumour type as a random variable. Only viral genome types that were identified in more than 10 samples were included in this analysis. dsDNA = double stranded DNA, dsDNA partial = partially double stranded DNA, ssDNA = single stranded DNA, ssRNA(+) = positive sense single stranded RNA, ssRNA(-) = negative sense single stranded RNA. This model is represented by Equation 2.17.

		Dependent variable:
		SBS2 and SBS13 mutations
Count model: (Intercept)	Coefficient	1.79
	95% CI	$(-0.691, 4.28)$
	p-value	0.157
Count model: ssDNA	Coefficient	1.68
	95% CI	$(0.401, 2.96)$
	p-value	0.0101 *
Count model: ssRNA(+)	Coefficient	-0.503
	95% CI	$(-1.20, 0.190)$
	p-value	0.155
Count model: dsDNA	Coefficient	-0.0221
	95% CI	$(-0.320, 0.276)$
	p-value	0.884
Count model: dsDNA partial	Coefficient	-2.53
	95% CI	$(-4.03, -1.03)$
	p-value	9.31×10^{-4}***
Count model: Age	Coefficient	0.00938
	95% CI	$(0.00193, 0.0168)$
	p-value	0.0136 *
Zero model: (Intercept)	Coefficient	0.435
	95% CI	$(0.338, 0.532)$
	p-value	1.93×10^{-18}***
AIC		14213.59
Log Likelihood		-7097.80
Num. obs.		2353
Num. Tumour Types		37

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 2.15: Mixed effects models predicting the number of SBS2 and SBS13 mutations based on virus genus present and age, accounting for the effects of tumour type as a random variable. Only viruses identified in more than 10 samples were considered in this analysis. This model is represented by Equation 2.18.

		Dependent variable:
		SBS2 and SBS13 mutations
Count model: (Intercept)	Coefficient	1.85
	95% CI	$(-0.620, 4.33)$
	p-value	0.141
Count model: Alphapapillomavirus	Coefficient	0.357
	95% CI	$(-0.208, 0.923)$
	p-value	0.216
Count model: Cytomegalovirus	Coefficient	-0.552
	95% CI	$(-1.41, 0.303)$
	p-value	0.206
Count model: Gammaretrovirus	Coefficient	-0.151
	95% CI	$(-1.00, 0.698)$
	p-value	0.727
Count model: Lymphocryptovirus	Coefficient	-0.0633
	95% CI	$(-0.450, 0.323)$
	p-value	0.748
Count model: Orthohepadnavirus	Coefficient	-2.52
	95% CI	$(-4.02, -1.02)$
	p-value	$9.80 \text{ x } 10^{-4***}$
Count model: Roseolovirus	Coefficient	0.0374
	95% CI	$(-0.489, 0.564)$
	p-value	0.889
Count model: Age	Coefficient	0.00911
	95% CI	$(0.002, 0.17)$
	p-value	$0.0170\ ^{*}$
Zero model: (Intercept)	Coefficient	0.436
	95% CI	$(0.339, 0.534)$
	p-value	$2.33 \text{ x } 10^{-18***}$
AIC		14004.57
Log Likelihood		-6991.29
Num. obs.		2323
Num. Tumour Types		37

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

dictors of the number of APOBEC3 mutations ($p = 9.31 \times 10^{-4}$, Table 2.14). Additionally, single stranded DNA (ssDNA) genomes emerged as significant positive predictors of the number of APOBEC3 mutations in the genome level analysis ($p = 0.0101$, Table 2.14). However, the genera that made up the ssDNA group (unclassified Anelloviridae, Alphatorquevirus, Dependoparvovirus, and Erythroparvovirus), were individually found in fewer than 10 tumours, and were therefore excluded from the genus level analysis. At the genus level, the Orthohepadnavirus presence was the only genus that had a significant effect on the number of APOBEC3 mutations ($p = 9.80 \times 10^{-4}$, Table 2.15).

2.5 Discussion

I show, for the first time using whole genome sequencing data from 24 different tumour types, that increases in APOBEC3 signatures are associated not only with increased mutation burden, but also that the presence and amount of these mutations correlate with multiple measures of genomic instability across multiple different cancer types. I expand on previous work in the field, which primarily used mutation burden and kataegis as measures of genomic instability (see Burns et al. (2013) and Roberts et al. (2012)), and introduce six novel measures of genomic instability, two of which (INDEL signatures ID6 and ID8) have not been studied before. It has previously been suggested that the increase in base substitutions observed in cancers over-expressing APOBEC3B (A3B) may be due to A3B induced U/G mis-pairs being processed by repair enzymes, which may result in other patterns of mutations, as well as strand breaks and chromosomal rearrangements (Burns et al., 2013; Harris, 2015). My analysis of the relationship between APOBEC3 mutations and the measures of genomic instability strongly suggests that this is the case and that APOBEC3s play an active role in the generation of genomic instability.

I found higher levels of structural variants, copy number segments, and INDELs

in tumours carrying APOBEC3 mutations (Figure 2.2), all common outcomes of double strand break (DSB) repair (Currall et al., 2013). In addition, INDEL signatures ID6 and ID8, which have been proposed as indicators of non-homologous end-joining (NHEJ) repair of DSBs, are also present in higher numbers in tumours carrying APOBEC3 mutations (Alexandrov et al., 2020). While PGA may not be directly related to DSBs, it reveals samples in which relatively few but large copy number events have occurred, as a result of DSBs, which would not necessarily be reflected by the number of copy number segments. Tumours containing APOBEC3 mutations were also found to have higher levels of PGA. The observation that the number of APOBEC3 mutations served as a significant positive predictor for all of the measures of genomic instability, after accounting for variation between tumour types and the effect of TP53 mutation, suggests that the two are closely related.

It can be argued that higher levels of APOBEC3 mutations are a consequence, rather than a cause, of increasing genomic instability. The conventional view of the involvement of APOBEC3 in genomic instability paints APOBEC3 as reactionary to double strand breaks and other processes that result in the generation of single stranded DNA. Several groups have demonstrated the occurrence of clusters of classical APOBEC3 mutations in the vicinity of double strand breaks (Taylor et al., 2013; Roberts et al., 2012).

However, the immunoglobulin translocations caused by activation induced cytidine deaminase (AID) in B-cell tumours serve as a precedent for the generation of DSBs, and their downstream consequences, by cytidine deaminases (Casellas et al., 2016). AID, which is ancestral to the APOBEC3 enzymes (Conticello et al., 2005), deaminates cytosines in the switch region near the immunoglobulin locus. The resulting uracils are excised by uracil N glycosylase (UNG), resulting in an abasic site which is processed into a single strand break (reviewed in Xu et al. (2012)). These single strand breaks can then form double strand breaks, either through further processing of the site, or due to close proximity of multiple single strand breaks (Xu

et al., 2012). The resolution of the DSBs precipitated by AID in these regions, is the basis of class switch recombination (CSR) (Xu et al., 2012). In addition to its role in CSR, off-target activity of AID is known to result in translocations between IGH and various genes, most notably MYC, BCL1, BCL2, MALT1, E2A, and CRLF2 (Lieber, 2016). AID mediated translocations are thought to account for half of all human haematopoietic malignancies (Lieber, 2016).

Undoubtedly APOBEC3 can be activated in response to, and act on, the products of DNA damage. My results suggest that it can also be a contributing factor in DNA damage and genomic instability. Kataegis is associated with so-called 'opportunistic' action of APOBEC3s at single stranded DNA during repair of DNA strand breaks. When I excluded mutations attributed to kataegis from my analysis, the strong association between APOBEC3 mutations and genomic instability remained in place for five of the six measures of genomic instability that I investigated. Thus, my results support a model of APOBEC3 mediated mutagenesis resulting in genomic instability via double strand break formation, which I posit mirrors the effects of AID in B-cell tumours.

Associations between APOBEC3 signature prevalence and genomic instability were observed across multiple tumour types. Particularly strong correlations were seen for pancreatic cancer, pancreatic endocrine neoplasms, kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, malignant lymphoma, bone cancer, and uterine corpus endometrial carcinoma (Figure 2.3).

Although large studies of pancreatic cancer genomes have highlighted APOBEC3 activity as one of the main mutagenic processes in pancreatic cancer (Bailey et al., 2016; Chang et al., 2014; Scarpa et al., 2017), these findings have not been investigated in depth, and the role of APOBEC3 activity in pancreatic cancer appears not to have been studied in great detail. However, preliminary data suggest that APOBEC3A activity may result in widespread genomic instability through a non-deaminase dependent mechanism, in a mouse model of pancreatic cancer (Woermann

et al., 2019). Suggesting the possibility of novel therapeutics for pancreatic cancer.

The presence of APOBEC3 related mutations in kidney cancer has also not been studied in great detail. Although I observed significantly higher levels of genomic instability in both kidney renal clear cell carcinoma and kidney renal papillary cell carcinomas that carry APOBEC3 mutations, I would urge caution when interpreting these results, as they are based on relatively few positive samples (2 and 3 positive samples, respectively). Further work is required to completely understand the role that APOBEC3 mediated mutagenesis may play in kidney cancer.

Interestingly, bone cancer and APOBEC3 induced genomic instability have been linked through the presence of kataegis in 50-85% of osteosarcoma samples (Perry et al., 2014; Chen et al., 2014). In addition to kataegis, osteosarcomas frequently display high levels of genomic instability, in the form of structural rearrangements and copy number aberrations, as well as carrying mutations in TP53 (Perry et al., 2014; Chen et al., 2014). It would be interesting to see if any of these abnormalities may be linked to the activity of APOBEC3 enzymes.

Analysis of TP53 mutations in this data set lends further support to work by other groups, in which TP53 mutations are observed more frequently in tumours expressing high levels of APOBEC3B (Burns et al., 2013). TP53 mutation has previously been linked with aneuploidy and copy number variations (Ciriello et al., 2013), and in this study positively associated with the number of copy number segments, PGA, structural variants, INDELs, and ID6 INDELs. Despite the inclusion of TP53 status, the number of APOBEC3 mutations was consistently identified as a highly significant predictor for all six measures of genomic instability.

APOBEC3 activity is thought to occur in episodic bursts following virus and retrotransposon activity (Petljak et al., 2019), so I investigated whether the presence of APOBEC3 mutations could be attributed to previous viral infections, indicated by the presence of viral DNA in the tumour. In my analysis, virus presence did not have a significant effect on the number of APOBEC3 mutations, when the effects of age

and tumour type were accounted for. However, ssDNA viruses, when combined, were found to have a significant positive effect, and partially double stranded DNA viruses were found to have a significant negative effect. The Orthohepadnavirus genus, which is the only genus with a partially double stranded DNA genome, also had a significant negative effect on the number of APOBEC3 mutation, after adjusting for the effects of age and tumour type. While this may potentially be an interesting finding for further follow up, I would like to note that APOBEC3 mutations were only observed in 5 of 336 liver tumours in this data set, and are not commonly present in liver cancers studied by others (Alexandrov et al., 2013; Schulze et al., 2015). Thus, while the results were statistically significant, they may not be biologically meaningful, and may be attributed to a handful of outliers, with respect to the complement of mutational signatures observed in these samples.

The lack of association between any of the other virus genera identified in the data set was surprising. However, this data set was not designed for the purpose of detecting APOBEC3 related mutagenesis in tumours carrying evidence of a viral infection, and any viruses identified in this data set should be considered incidental findings. Another caveat of this analysis is that historical viral infections which have been cleared, and ongoing RNA virus infections, may not be identifiable in whole genome sequencing data due to the absence of viral genomes, and RNAse treatment of samples during DNA extraction, respectively. In addition to this, several viruses are known to express proteins that inhibit APOBEC3 activity or expression in a variety of ways, such as the viral infectivity factor (Vif), expressed by HIV (Goila-Gaur and Strebel, 2008). While I did not identify any further associations between specific virus genera and the presence of APOBEC3 mutations, this warrants further research, and may implicate viruses as causative factors in more cancers than previously thought.

I found that both the presence of APOBEC3 mutations, and missense or nonsense mutations in TP53 had a negative effect on survival. It has been suggested

that cancers with an APOBEC3 mutation component could be treated with DNA damaging drugs, resulting in synthetic lethality (Swanton et al., 2015). This is an interesting idea, and evidence from studies of urothelial carcinoma suggests that this may indeed improve treatment outcomes (Mullane et al., 2016; Glaser et al., 2018). Similarly, it has recently been reported that a subset of clear cell ovarian carcinoma (CCOC) patients over-expressing A3B had better survival outcomes when treated with platinum based drugs (Harris et al., 2020). It was theorised that the increased survival of the patients in this CCOC subset was due to A3B mediated DNA damage sensitising the tumour cells to further damage by platinum based drugs (Harris et al., 2020). This suggests that A3B activity and the presence of APOBEC3 related mutations may be used to inform treatment decisions and may also provide an insight into treatment outcomes (Harris et al., 2020; Mullane et al., 2016). My results suggest that this approach may be beneficial for patients with pancreatic cancer, kidney cancer, malignant lymphoma, bone cancer, and uterine corpus endometrial carcinoma, carrying APOBEC3 mutations.

Chapter 3

Effect of in vitro infection with pseudoviral particles on APOBEC3A expression, DNA damage, and genomic instability

3.1 Introduction

In this chapter I investigate whether viral infection of cells results in increased levels of APOBEC3A, and whether this is related to increased DNA damage and DNA damage signalling. I also investigate whether viral infection, or transfection of cells with A3A results in the same pattern of genomic instability observed in tumour whole genome sequencing data in Chapter 2.

I have focused my analysis on A3A for three main reasons, (1) A3A, but not A3B, is strongly induced at the protein and mRNA level after viral infection, due to stimulation of IFN-α, which is commonly released in response to viral infection (Koning et al., 2009; Refsland et al., 2010; Stenglein et al., 2010; Berger et al., 2011), (2) along with A3B, A3A is associated with the SBS2 and SBS13 mutational

signatures that are found in cancer genomes (Alexandrov et al., 2013; Chen et al., 2019; Middlebrooks et al., 2016), and (3) the effects of A3B on genome stability have already been studied (Akre et al., 2016), whereas the effects of A3A have not been studied in as much detail.

3.2 Materials

3.2.1 Cell Cultures

HCT116 colorectal cancer cells were grown in high glucose DMEM medium supplemented with 10% FBS, penicillin and streptomycin, and L-glutamine. Cells were kept in an incubator at 37°C with 5% CO_2. The HCT116 cells were kindly provided by Ross Chapman's group.

RPE-1 cells were grown in DMEM:F12 media with L-glutamine, supplemented with 10% FBS, and penicillin and streptomycin, under the same conditions as the HCT116 cells.

3.2.2 Pseudoviral Particles

Pseudotyped HIV particles were created by co-transfecting the pR8ΔEnv plasmid, which encodes the HIV genome with a deletion in the Env region, pcRev, Gag, and either HXB2, JRFL, or VSV-G as the envelope glycoprotein, into a Lenti-HEK producer cell line, described in Iliopoulou et al. (2018). These materials were shared by Greg Melikyan of Emory University and James Binley of the Torrey Pines Institute for Molecular Studies.

Table 3.1: Table of viruses used

Abbreviation	Genome	Envelope	Fluorescent component
GFP-VSVG	pR8ΔEnv	VSV-G	Gag-GFP
Unlabelled-VSVG	pR8ΔEnv	VSV-G	none

3.2.3 Antibodies Used for Immunofluorescence

The primary antibodies used in the experiments described in this thesis are listed in Table 3.2. Secondary antibodies are listed in Table 3.3.

Table 3.2: Primary antibodies used for immunofluorescence

Target	Target species	Antibody species	Source	Catalogue number	Concentration
γ H2AX	Human	Mouse	Millipore	05-636	1:500
APOBEC3A	Human	Goat	Biorbyt	orb19763	1:500

Table 3.3: Secondary antibodies used for immunofluorescence

Antibody species	Target species	Fluorophore	Source	Serial number	Concentration
Donkey	Mouse	Alexa Fluor Plus 488	Life Technologies Ltd.	A32766	1:1000
Donkey	Goat	Alexa Fluor Plus 555	Life Technologies Ltd.	A32816	1:1000

3.2.4 APOBEC Over-Expression Plasmids

A plasmid expressing GFP-tagged APOBEC3A (Cat. no. RG220995), as well as the corresponding vector containing only GFP (Cat. no. PS100010), were purchased from OriGene Technologies GmbH.

3.2.5 Data Availability

The RNAseq data generated in this chapter, described below, have been deposited in NCBI's Gene Expression Omnibus (GEO; Edgar et al. (2002)), and are accessible through GEO Series accession number GSE172164 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE172164).

The whole genome sequencing data produced as part of this chapter is in the process of being submitted to the Sequence Read Archive (SRA; Leinonen et al. (2010)).

3.3 Methods

3.3.1 *In vitro* Infection and Fixation for Immunofluorescence

Cells were seeded into 24-well plates containing glass coverslips and allowed to adhere for 24 hours. Virus suspensions were prepared by diluting the virus in media to reach the required multiplicity of infection (MOI), on the day of infection. MOI represents the number of viral particles per cell. Media was aspirated and replaced with either the viral suspension or fresh media. 24 hours later, cells were fixed with 4% PFA for 15 minutes at room temperature. Three biological replicates were performed on three different days for each experiment, and 20 fields of view were captured for each condition.

3.3.2 Immunofluorescence

First, cells were stained with Wheat Germ Agglutinin (WGA) fluorescently conjugated with the far red dye CF640R (Insight Biotechnology 29026-1), to label the plasma membrane. WGA was diluted in phosphate buffered saline (PBS) to a final concentration of 5μg/mL, and applied to cells for 10 minutes.

Next, coverslips were blocked for 15 minutes in a buffer containing 3% bovine serum albumin (BSA) and 0.1% triton (immunofluorescence buffer). Samples were then incubated in the same buffer containing primary antibodies at the concentrations listed in Table 3.2 for 1 hour, followed by three washes in immunofluorescence buffer. After washing, the samples were incubated in immunofluorescence buffer containing the secondary antibodies in Table 3.3 for 1 hour. Following another three washes in immunofluorescence buffer, the coverslips were stained with Hoechst 33258. Coverslips were mounted on slides using ProLong Gold Antifade Mountant (Life Technologies P10144), allowed to cure overnight, and sealed with clear nail varnish before imaging.

3.3.3 Microscopy

Imaging was conducted on a Leica Microsystems DMi8 S TIRF Microscope stand with automatic focus correction system. A HC PL APO 100x/1.47 TIRF lens was used for imaging. All images were captured using a Hamamatsu Orca Flash 4.0 V2 Camera (CMOS monochrome). Wide field fluorescence illumination was conducted using a Leica EL6000 fluorescence illuminator and filter cubes for GFP, RHOD and Y5 (CY5). A fluorescence filter wheel for fast switching with the following filters for excitation EFW: 395/25, 470/35, 575/30, dichroic: 420, 495, 595, emission EFW: 435/25, 520/45, 615/35, was used to capture Hoechst staining.

3.3.4 Analysis of Immunofluorescent Images

The intensity of nuclear A3A and γH2AX staining was quantified using a fully automated analysis pipeline developed in `CellProfiler` software (Kamentsky et al., 2011). In short, nuclei were identified using hoechst staining and the mean pixel intensity within the nucleus was measured for A3A and γH2AX staining.

3.3.5 RNAseq Set Up and Library Preparation

HCT116 cells were seeded into four 6-well plates and allowed to adhere overnight. Unlabelled, pseudotyped HIV expressing the VSV-G (Unlabelled-VSVG) virus was added to the media at MOI 1. Cells were harvested at 0 (uninfected), 6, 12, and 24 hours post-infection, and immediately placed in RNAlater solution (Applied Biosystems AM7020).

Total RNA was extracted using the PARIS kit supplied by Life Technologies (AM1921). Extracted RNA was treated with DNAse supplied by Life Technologies (AM1906), and checked for quality using RNA ScreenTapes on a 4200 TapeStation system (GS2991AA), both supplied by Agilent Technologies, before proceeding to cDNA synthesis. cDNA was synthesised using the NEBNext Ultra II Directional

RNA Library Prep with Sample Purification Beads, following NEB's Protocol for use with NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490). Sequencing libraries were prepared using the aforementioned kit and individual index primers for each sample (supplied by NEB). Individual libraries were checked for quality using Qubit fluorometry (Life Sciences), as well as on the 4200 TapeStation using the High Sensitivity D1000 ScreenTape assay. Libraries were sequenced using an Illumina NextSeq 500 sequencing machine.

3.3.6 RNAseq Data Analysis

Analysis of raw RNAseq data was carried out using version 3.0 of the `nf-core/rnaseq` pipeline developed by Ewels et al. (2020), within the `nextflow` framework (`Nextflow` version 20.12.0-edge, Di Tommaso et al. (2017)). In short, quality control was carried out using `FastQC` version 0.11.9, unique molecular identifiers (UMIs) were extracted with `UMI-tools` (Smith et al., 2017), and adaptors were trimmed with version 0.6.6 of `Trim Galore!`. Reads were aligned to hg19 (GRCH37) using version 2.6.1d of the `STAR` aligner (Magoč and Salzberg, 2011), and counted using version 1.4.0 of `Salmon` (Patro et al., 2017). Alignments were sorted and indexed using version 1.10 of `SAMtools` (Zhang et al., 2013), de-duplicated using `UMI-tools`, and duplicate reads were marked using version 2.23.9 of `picard MarkDuplicates`. Transcripts were assembled and quantified using `StringTie` version 2.1.4 (Hu et al., 2016), and bigWig coverage files were created using `BEDTools` version 2.29.2 (Ruepp et al., 2010). Quality control was carried out using version 3.0.1 of `RSeQC` (Wang et al., 2012), version 2.2.2-dev of `Qualimap` (Okonechnikov et al., 2016), version 1.18.0 of `dupRadar` (Xing et al., 2019), version 2.0.3 of `Preseq` (Daley and Smith, 2013), version 1.28.0 of `DESeq2` (Love et al., 2014), and aggregated using `MultiQC` (Ewels et al., 2016). Downstream analysis was conducted in `R` version 3.6.2 (R Core Team, 2018). Differential expression analysis was carried out using `DESeq2` version 1.26.0 (Love et al., 2014) and `goseq` version 1.38.0 (Young et al., 2010). Genes and transcripts

to which fewer than 10 reads were mapped were excluded from this analysis.

Plots were constructed using `ggplot2` version 3.3.3 (Wickham, 2016), `EnhancedVolcano` version 1.5.4 (Blighe et al., 2020), `gplots` version 3.1.1 (Warnes et al., 2020), `gridExtra` version 2.3 (Auguie, 2017), and `viridis` version 0.5.1 (Garnier, 2018).

3.3.7 Transfection and Infection of Cells for Whole Genome Sequencing

Transfection of RPE-1 cells with GFP and APOBEC3A was carried out using GeneJuice Transfection Reagent (Sigma-Aldrich 70967), following the protocol recommended by the manufacturer. RPE-1 cells were infected with unlabelled pseudotyped HIV expressing VSV-G at an MOI of 50.

3.3.8 Single Cell Sorting Using Flow Cytometry

All FACS experiments were carried out on a Sony SH800S Cell Sorter. Single cell suspensions were prepared by trypsinising cells and pipetting them up and down generously before loading the sample tubes into the machine. Cells were sorted 1.5 days after transfection and infection.

3.3.9 DNA Extraction for Whole Genome Sequencing

DNA extractions were carried out as soon as sufficient cells were available (22-32 days after sorting). Genomic DNA was extracted from cells using the Macherey-Nagel Nucleospin DNA Rapidlyse kit, supplied by Fisher Scientific (Product Code 15715422). Sample concentration was assayed using a NanoDrop One fluorometer. Further quality control and sample concentration analysis was carried out by Source BioScience, using both Qubit and NanoDrop fluorometers.

PCR free libraries were prepared by Source BioScience using the KAPA Hyper Prep Kit using Illumina TruSeq DNA adapters. Libraries were assessed for quality

using High sensitivity DNA Chips on an 2100 Agilent BioAnalyzer, and sequenced to a target depth of 30X, using 150bp paired end reads on a NovaSeq sequencer by Source BioScience.

3.3.10 Analysis of Whole Genome Sequencing Data

3.3.10.1 Pre-processing, Quality Control, and Variant Calling

Variants were called using a variant calling pipeline developed by Kedzierska (2021) using `Snakemake` (Köster and Rahmann, 2012). In short, quality control was assessed using `FastQC` version 0.11.8, followed by read trimming using `TrimGalore!` version 0.6.5. Reads were then aligned to GRCh38 using `BWA` version 0.7.17 (Li and Durbin, 2009), and duplicate alignments were marked using `picard MarkDuplicates` version 2.21. Remaining reads were then filtered using `SAMtools` version 1.9 and `BEDtools` version 2.26 (Quinlan and Hall, 2010). Finally, variants were called using GATK `Mutect2` version 4.1.4, then filtered using GATK `FilterMutectCalls` (Benjamin et al., 2019; McKenna et al., 2010), and annotated using the Ensembl variant effect predictor (`VEP`) (Hu et al., 2016).

Variants that failed any of the filters applied by `Mutect2` were discarded. The variants were filtered further, using the Ensembl variant effect predictor (`VEP`) (Hu et al., 2016), to exclude common variants with a minor allele frequency (MAF) >1%.

Variants were also called using `Octopus` in tumour only mode (Cooke et al., 2021). Variants called by `Octopus` were also filtered using `VEP` to remove variants with a population MAF >1%. I then selected for variants that were identified by both `Octopus` and `Mutect2`, and discarded any variants that were identified in more than one sample based on the position of the variant, leaving only unique variants for each sample.

Further to this, I limited my analysis to SNVs with a variant allele frequency (VAF) between 0.3-0.7. These SNVs represent clonal variants found in all of the cells in each clone, rather than variants which have accumulated while the clones

were being grown.

3.3.10.2 Mutational Signatures

The `mutSignatures` R package (version 2.1.1) was used to identify COSMIC signatures present in the whole genome sequenced clones (Fantini et al., 2020).

3.3.10.3 Structural Variants

Three structural variant callers, `DELLY` (Rausch et al., 2012), `Manta` (Chen et al., 2016), and `GRIDSS` (Cameron et al., 2017) were run using the `sv-callers` workflow (Kuzniar et al., 2020). A consensus call set based on SVs identified by at least two out of the three SV callers was generated using `SURVIVOR` (Jeffares et al., 2017). Individual SVs identified by the three different pipelines were considered overlapping, and subsequently merged, if they were of the same type (e.g. identified as translocations by both pipelines), if they were identified by at least two SV callers, and the coordinates of both their start and end sites were within 1kb of each other. Variants were filtered to remove SVs identified in all three control samples. I also removed SVs that overlapped with regions of poor mapping quality. I then removed variants identified in more than one sample, resulting in a list of unique structural variants for each sample.

3.3.10.4 Copy Number Variation

Copy number was assessed using an unpublished branch of the `Battenberg` copy number caller, modified to work with cell line data (Nik-Zainal et al., 2012).

3.4 Results

3.4.1 In Vitro Infection of Cells With Pseudotyped HIV Particles Increases Nuclear APOBEC3A Levels and DNA Damage

HCT116 and RPE-1 cells were infected with HIV pseudoviral particles carrying the Vesicular Stomatitis Virus glycoprotein (VSV-G). Cells were fixed 24 hours after infection, and labelled with antibodies against APOBEC3A (A3A) and γH2AX to determine whether viral infection led to an increase in nuclear A3A staining and DNA strand breaks assessed by γH2AX staining.

I found that infection with pseudotyped HIV significantly increased nuclear A3A staining in both cell lines at all MOIs, with the exception of MOI 10 for HCT116 and MOI 1 for RPE-1 (One-sided Wilcoxon Rank Sum Test, $p < 0.05$, Figures 3.1 and 3.2). I also saw a significant increase in the intensity of nuclear γH2AX staining at all MOIs in both cell lines, with the exception of MOI 50 in HCT116 cells (Figures 3.1 and 3.2, One-sided Wilcoxon Rank Sum Test, $p < 0.05$).

3.4.2 RNAseq Analysis of HCT116 Cells Infected With Pseudotyped HIV Over Time

In order to further characterise the cellular response to viral infection, I infected HCT116 cells with VSV-G pseudotyped HIV at an MOI of 1 and collected RNA from cells at 0 hours (uninfected, T0), 6 hours (T6), 12 hours (T12), and 24 hours (T24) after addition of virus to the medium, and carried out RNA sequencing. To visualise differences in gene expression across the different time points I performed principal components analysis (PCA), which demonstrated clear separation of the different time points with respect to their gene expression profiles, with the 24 hour time point appearing as the most distinct (Figure 3.3). This is consistent with previous

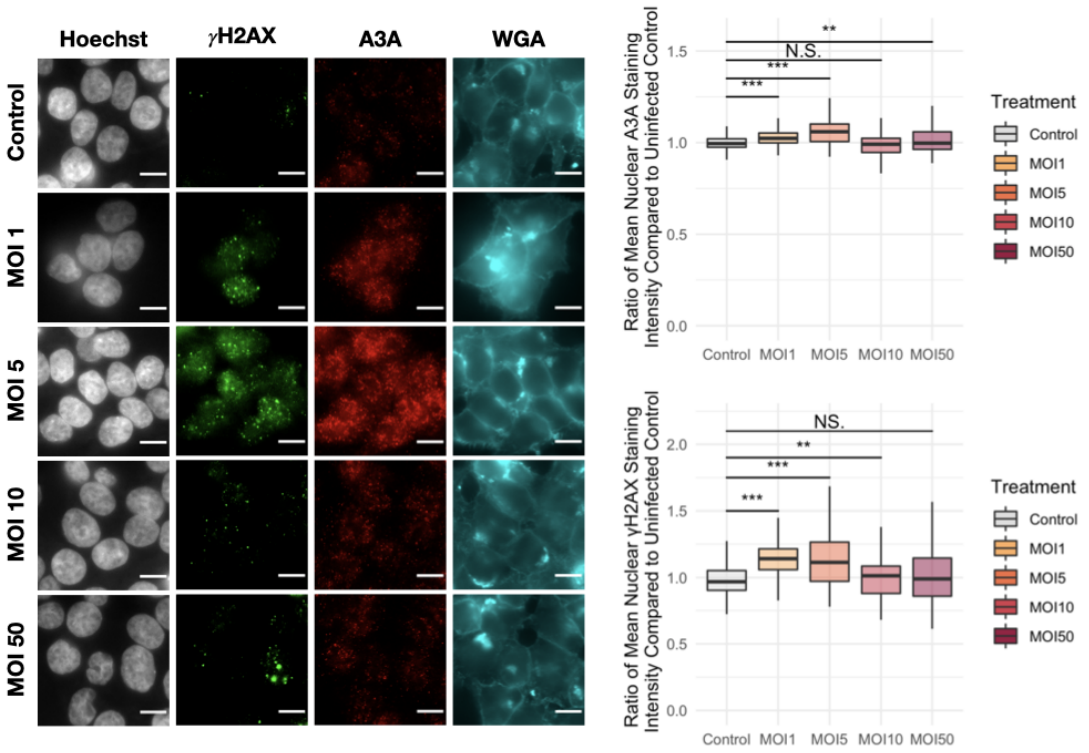


Figure 3.1: HIV infected HCT116 24 hours post-infection, stained with antibodies against A3A and γH2AX. Scale bar 10 μm. One-sided Wilcoxon rank sum test $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Single cell measurements of mean nuclear staining intensity were normalised to the uninfected control and measured across three biological replicates. Twenty fields of view were captured for each condition within each replicate, resulting in 649-1085 cells measured per condition per replicate. WGA = Wheat Germ Agglutinin.

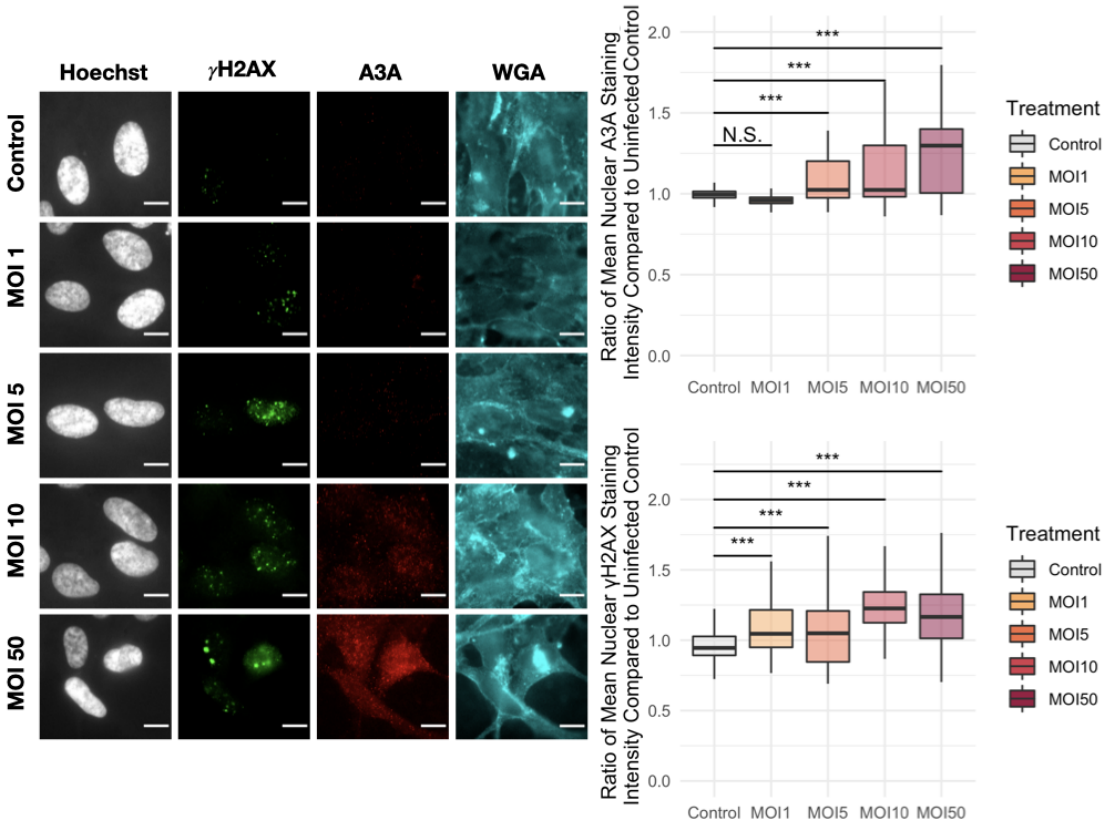


Figure 3.2: HIV infected RPE-1 24 hours post- infection, stained with antibodies against A3A and γH2AX. Scale bar 10 μm. One-sided Wilcoxon rank sum test $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Single cell measurements of mean nuclear staining intensity were normalised to the unifected control and measured across three biological replicates. Twenty fields of view were captured for each condition within each replicate, resulting in 160-260 cells measured per condition per replicate. WGA = Wheat Germ Agglutinin.

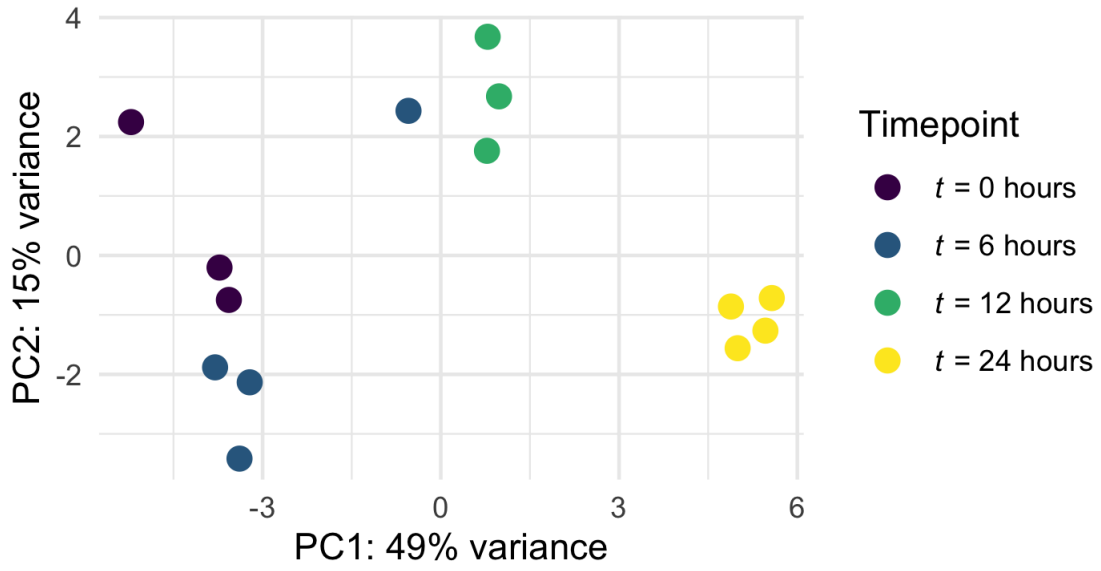


Figure 3.3: Principal component analysis of HIV infected HCT116 cells over time. PC = Principal Component.

reports of large changes in transcription 24 hours after HIV infection (Chang et al., 2011). The time point of 24 hours after infection is thought to coincide with peak viral replication or the late phase of viral replication, which involves release of viral particles (Chang et al., 2011; Mohammadi et al., 2013).

3.4.2.1 Differential Expression Analysis

Next, I investigated differential expression of genes at each time point relative to time point T0 (Figure 3.4). Six hours after addition of virus, KRT6A was the only gene that exceeded the threshold of an absolute log2 fold change (LFC) in expression >1 and an adjusted p-value <0.05 (Table C.1.1 in Appendix C). At 12 hours after addition of virus, 92 genes had exceeded the thresholds, and at 24 hours post addition of virus, 371 genes were significantly differentially expressed with an LFC >1 (Figure 3.4, Tables C.2 and C.3 in Appendix C).

Following differential expression analysis, I investigated the top 40 genes with

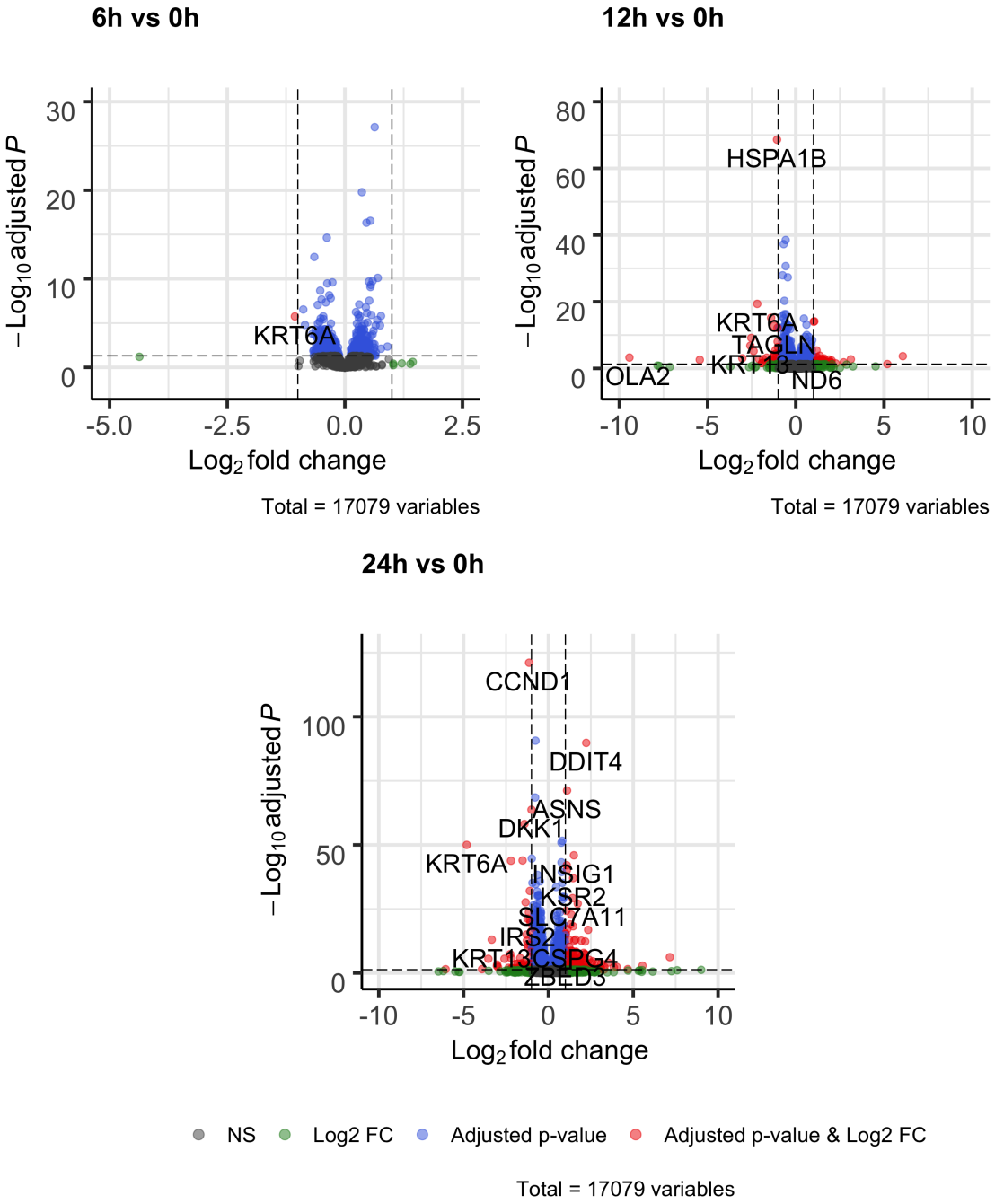


Figure 3.4: Volcano plots for RNAseq of HIV infected HCT116 cells. Grey points represent genes which are not significantly differentially expressed with respect to time point T0. Green points represent genes whose expression has changed by an absolute log2 fold change (LFC) of >1, but did not exceed our threshold of an adjusted p-value <0.05. Blue points represent genes with an adjusted p-value <0.05 and a LFC <1. Red points represent genes whose LFC exceeded 1 and p-value <0.05. Vertical dashed lines indicate LFC = 1 and the horizontal dashed line represents the adjusted p-value cutoff of p <0.05.

the highest variance across the different samples (Figure 3.5). Time point 24 appeared to be the most distinct from the other three, and was characterised by high expression of Mitochondrially Encoded ATP Synthase 8 (ATP8), DNA Damage Inducible Transcript 4 (DDIT4), MT-ND2 Pseudogene 28 (MTND2P28), Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 2 and 6 (ND2 and ND6), Insulin Induced Gene 1 (INSIG1), Asparagine Synthase (ASNS), Dehydrogenase/Reductase 3 (DHRS3), and H1.0 Linker Histone (H1-0), and low expression of ATP6V1G2-DDX39B, Ribosomal Protein S26 (RPS26), BolA Family Member 2 (BOLA2), CTD-2031P19.4, Brain Cytoplasmic RNA 1 (BCYRN1), Metastasis Associated Lung Adenocarcinoma Transcript 1 (MALAT1), Coagulation Factor III (F3), Cyclin D1 (CCND1), and Myosin Light Chain 9 (MYL9).

I also carried out gene ontology (GO) analysis to identify GO categories enriched at each time point (Figure 3.6). Ten GO categories were significantly enriched 6 hours after infection, 96 were enriched 12 hours after infection, and 229 were enriched 24 hours after infection. The "cytosol" and "membrane-bounded organelle" GO terms were significantly enriched at every time point, relative to uninfected. The most highly enriched GO terms occurring at 24 hours post infection, were "intracellular organelle part", "RNA binding", and "organelle part". Interestingly, several of the top GO terms identified as enriched at all time points were related to cell cycle changes (e.g. "mitotic cell cycle process", "mitotic cell cycle", and "cell cycle process", to name a few), indicating a clear relationship between HIV infection and alterations of the cell cycle.

3.4.2.2 APOBEC3 Expression

Surprisingly, there was no evidence of A3A or A3G expression. Fewer than 10 reads were mapped to A3A and A3G in each of the samples; thus they were excluded from the differential expression analysis. However, expression of A3B, A3C, A3D, A3F, and A3H was observed (Figure 3.7). None of the APOBECs were significantly

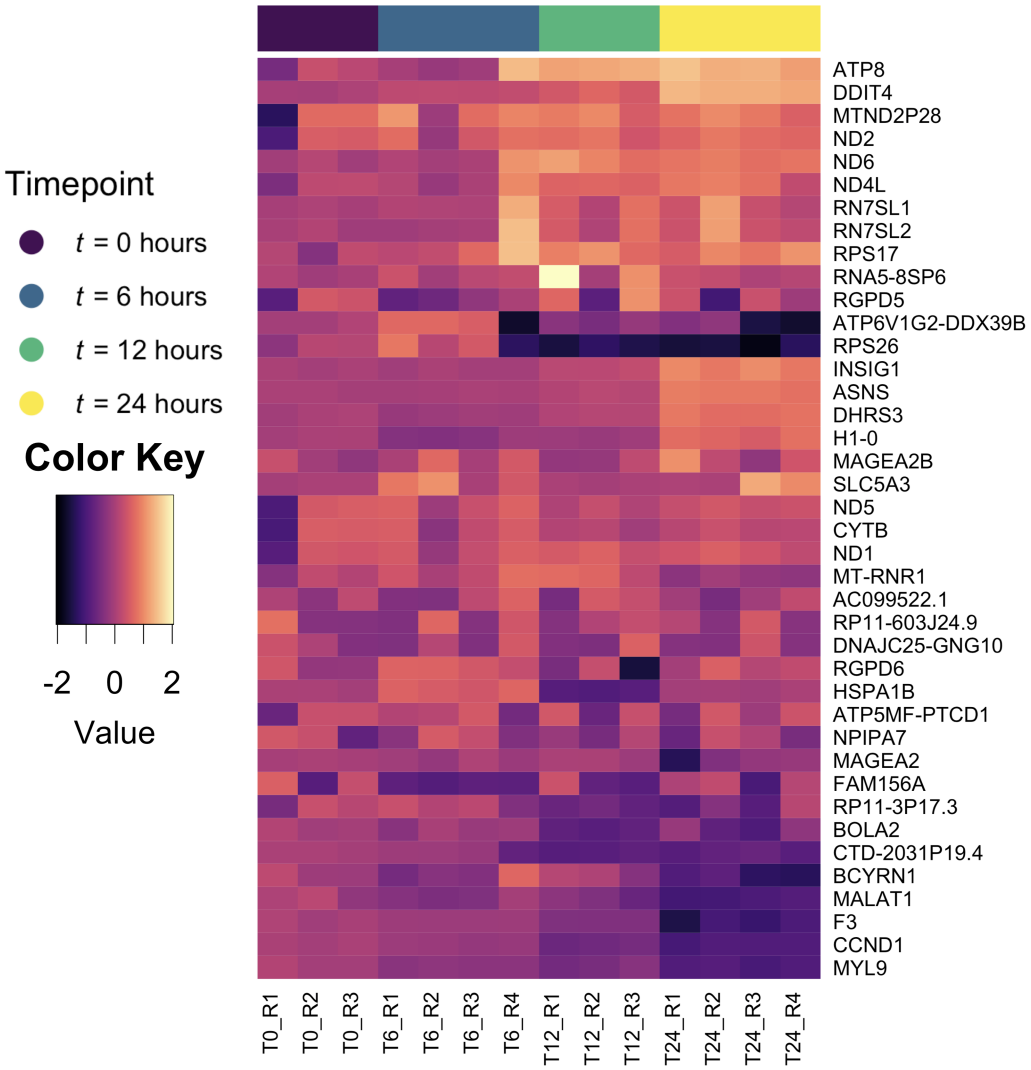


Figure 3.5: Heatmap of most variable genes in HIV infected HCT116 cells over time. Colours represent regularised-logarithm (rlog) transformed transcript counts centred around T0.

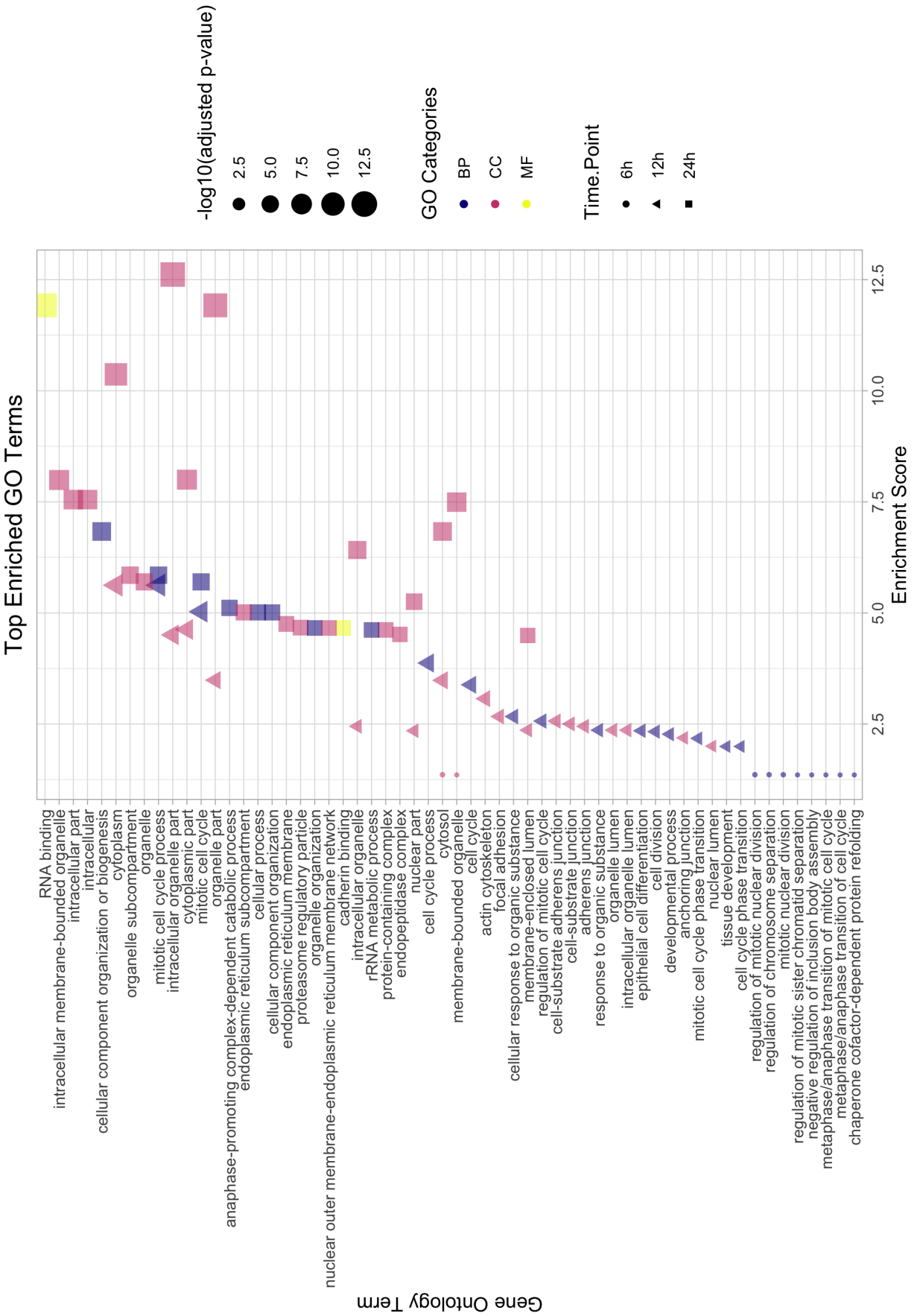


Figure 3.6: Top 30 Gene Ontology terms enriched at each time point. BP = Biological Process, CC = Cellular Component, MF = Molecular Function. p-values were adjusted using the Benjamini-Hochberg procedure.

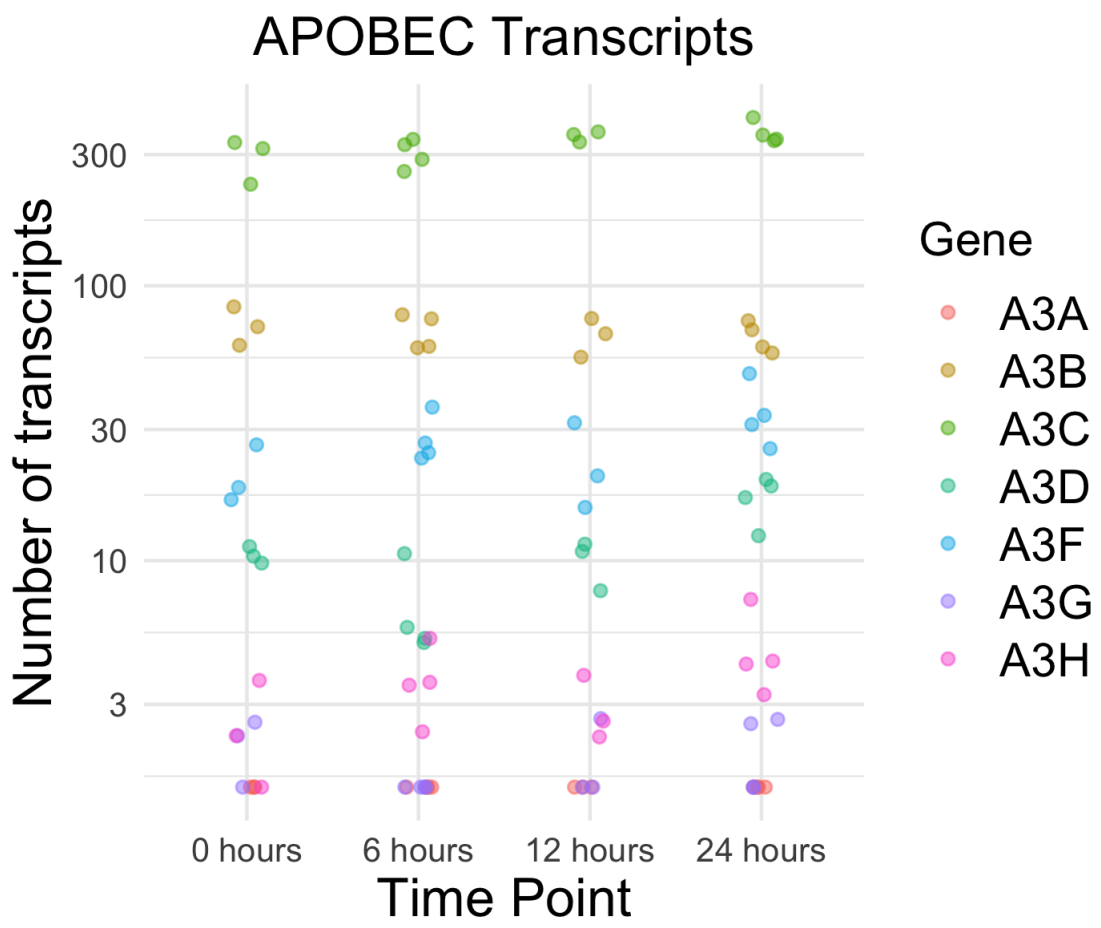


Figure 3.7: Number of transcripts mapped to each APOBEC3 gene at each time point.

differentially expressed at any time point (p >0.05).

3.4.2.3 Expression of DNA Damage Response Genes

To determine if activation of the DNA damage response (DDR) could be detected following infection, as I did with immunofluorescence, I investigated the expression of a list of DDR genes curated as part of the Molecular Signatures Database (MSigDB) Hallmark DNA repair gene set (Figure 3.8; the full list of genes along with their LFC values and adjusted p-values for each time point relative to control can be found in Table C.4).

A number of genes were significantly differentially expressed at various time points; however, only one of these, BOLA2, passed the LFC threshold as well.

When this data is represented as a heatmap, as in Figure 3.9, there appears to be a large downward shift in roughly half of the genes between 12 and 24 hours after infection.

3.4.3 Effect of *in vitro* Infection With Pseudoviral Particles and Transfection With A3A on Genome Stability Detected by Whole Genome Sequencing

To further investigate whether viral infection or A3A activity caused an increase in genomic instability, and to see whether I could validate some of my findings from Chapter 2, I carried out a number of sequencing experiments. RPE-1 cells were transfected with plasmids expressing either GFP or GFP-tagged A3A, and then single cell sorted and clonally expanded GFP positive cells for sequencing. I also infected RPE-1 cells with an excess of pseudoviral particles carrying VSV-G, single cell sorted them, and clonally expanded these cells, along with an uninfected control. Figure 3.10 presents a schematic overview of this experiment, along with an overview of the sequencing analysis. To limit heterogeneity derived from the parental cell line, RPE-1 cells were single cell sorted and clonally expanded prior to transfection and infection. The same parental RPE-1 clone was used as the origin for the 15 clones investigated. I sequenced three clones each, for cells that had been exposed to GFP, virus, or control, and six clones that had been exposed to A3A.

3.4.3.1 SNVs and INDELs

First, I quantified the number of clonal SNVs and small INDELs using the `Mutect2` and `Octopus` variant callers (Figure 3.11; Benjamin et al. (2019); Cooke et al. (2021)). While I did not observe any statistically significant differences in the number of SNVs and INDELs identified in each of the groups, there appeared to be a general upward trend in the number of variants identified, with control clones

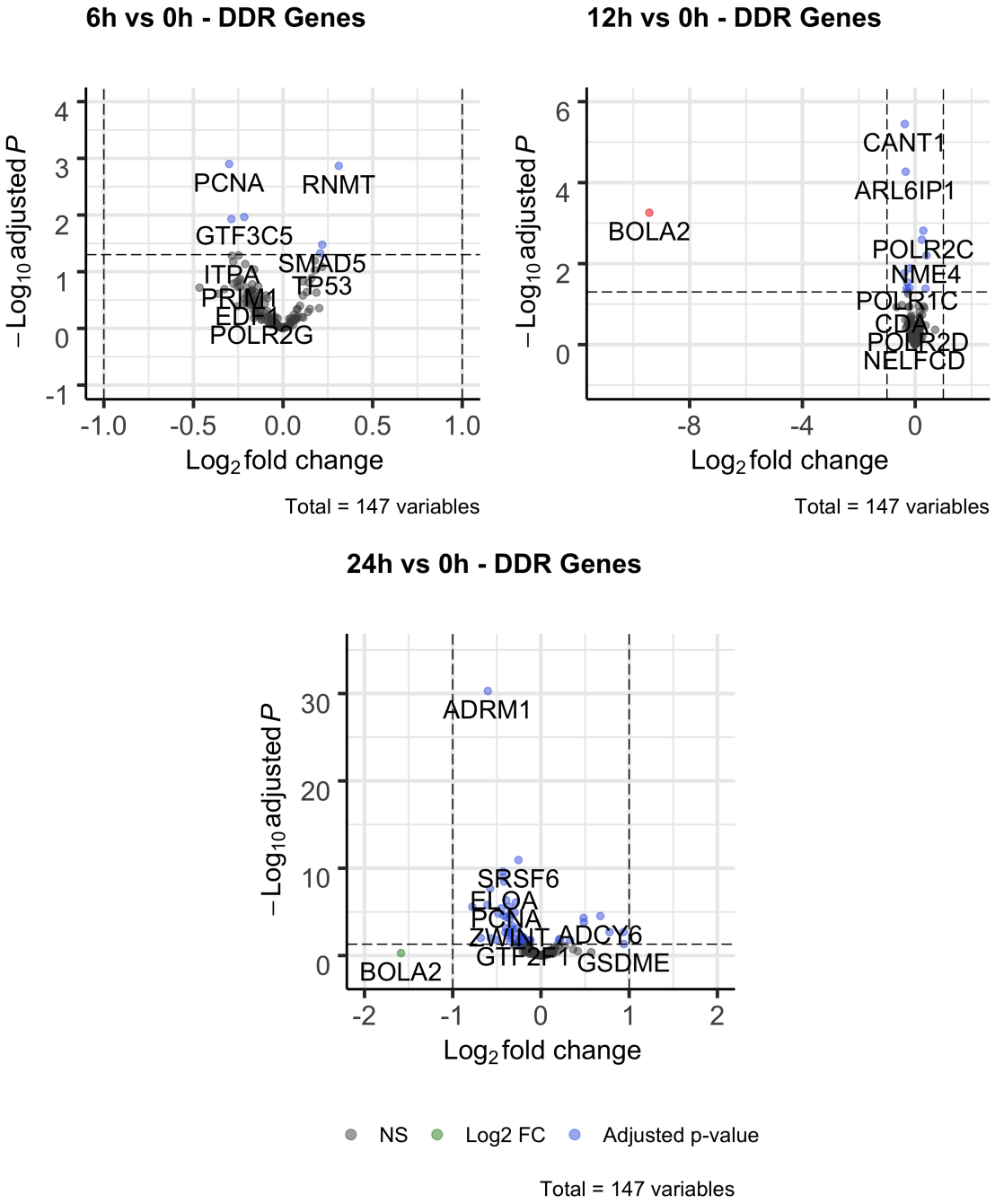


Figure 3.8: Volcano plots for expression of DDR genes at each time point. Grey points represent genes which are not significantly differentially expressed with respect to time point T0. Green points represent genes whose expression has changed by an absolute log2 fold change (LFC) of >1, but did not exceed our threshold of an adjusted p-value <0.05. Blue points represent genes with an adjusted p-value <0.05 and a LFC <1. Red points represent genes whose LFC exceeded 1 and p-value <0.05. Vertical dashed lines indicate LFC = 1 and the horizontal dashed line represents the adjusted p-value cutoff of p <0.05.

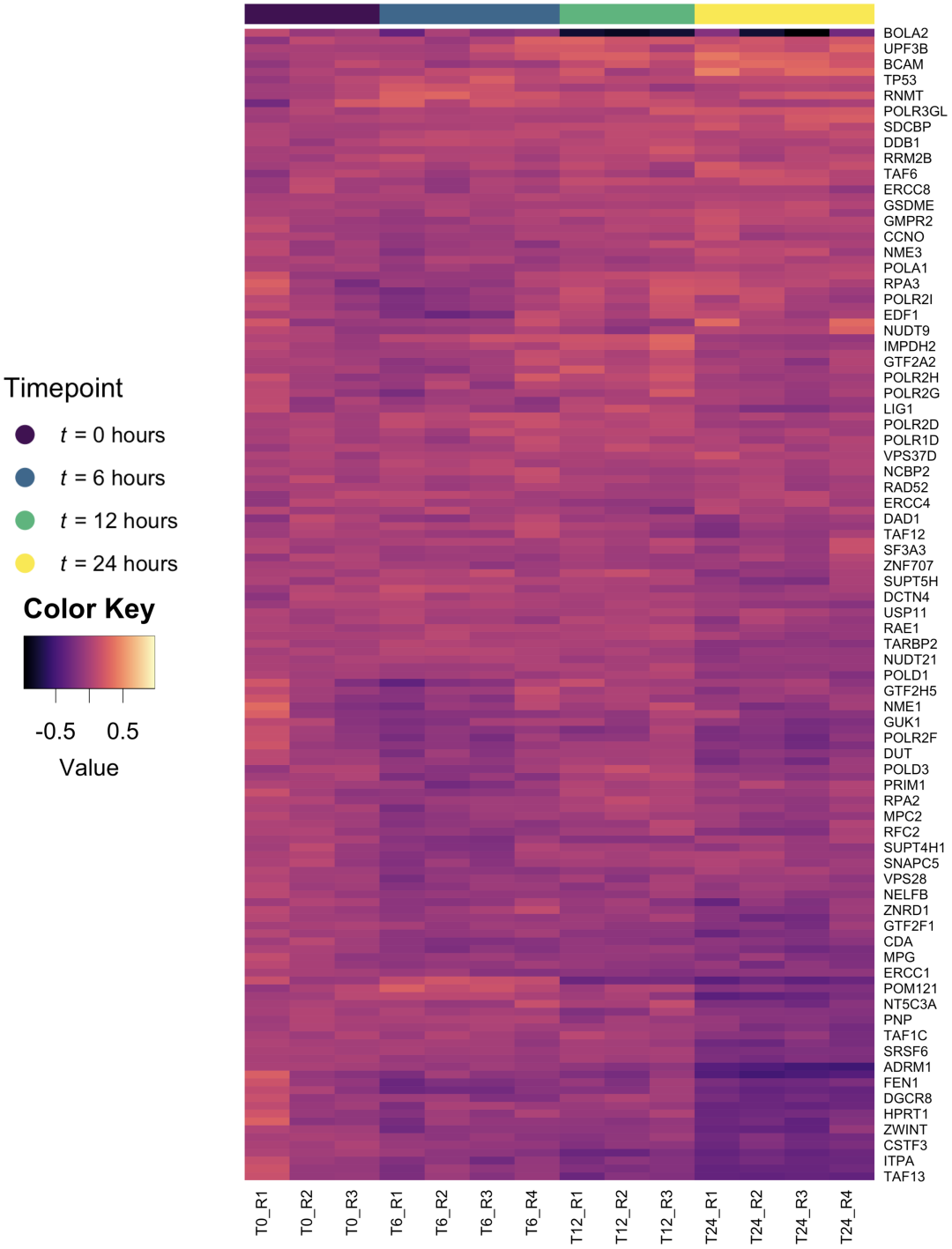


Figure 3.9: Heatmap of DDR gene expression by time point. Colours represent regularised-logarithm (rlog) transformed transcript counts centred around T0. Due to spatial limitations I was not able to label every gene surveyed. A full list of genes along with their LFC and adjusted p-values is provided in Table C.4, in Appendix C

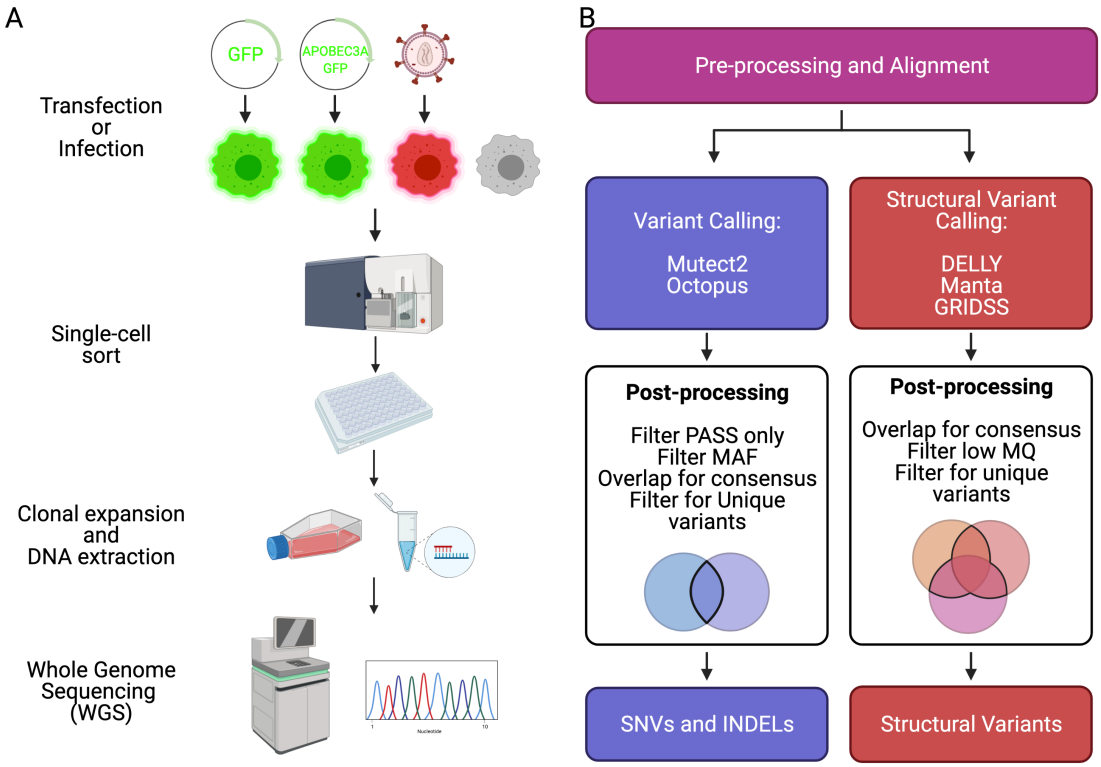


Figure 3.10: Whole genome sequencing of infected and transfected clones. A. Overview of wet lab workflow. B. Overview of analysis workflow. Image created with BioRender.com.

consistently having the lowest number, followed by GFP transfected clones, virally infected clones, and A3A transfected as the highest (Figure 3.11 B and C, two-sided Wilcoxon rank sum test p >0.05).

Next, I extracted COSMIC mutational signatures present in the clones (Figure 3.12). Similar to the results for the total mutation burden, I did not find a statistically significant difference in the number of SBS2 and SBS13 mutations identified in each group of clones (Figure 3.12 A, two-sided Wilcoxon rank sum test p >0.05). However, I observed a trend towards increased numbers of SBS2 and SBS13 mutations in clones that had been infected or transfected.

The majority of mutations identified were expected to fall into the SBS1 and SBS5 mutational signatures, which have been associated with ageing or cellular turnover, and are thought to display clock-like properties (Alexandrov et al., 2013, 2015). Half to a quarter of the mutations identified in each clone could be attributed to these signatures (Figure 3.12 B). Signatures SBS3, SBS12, and SBS16 also appeared frequently. SBS3 is associated with defective homologous recombination and is a relatively featureless signature, making it difficult to deconvolute (Alexandrov et al., 2013, 2020). The aetiology of SBS12 and SBS16 has not yet been proposed (Alexandrov et al., 2013).

3.4.3.2 Structural Variants

I also investigated the number of structural variants present in these clones, and found that there was a trend towards increasing numbers of structural variants in virally infected clones, though this did not reach statistical significance (Figure 3.13, two-sided Wilcoxon rank sum test p >0.05). The main types of SVs identified across all clones were deletions and inversions.

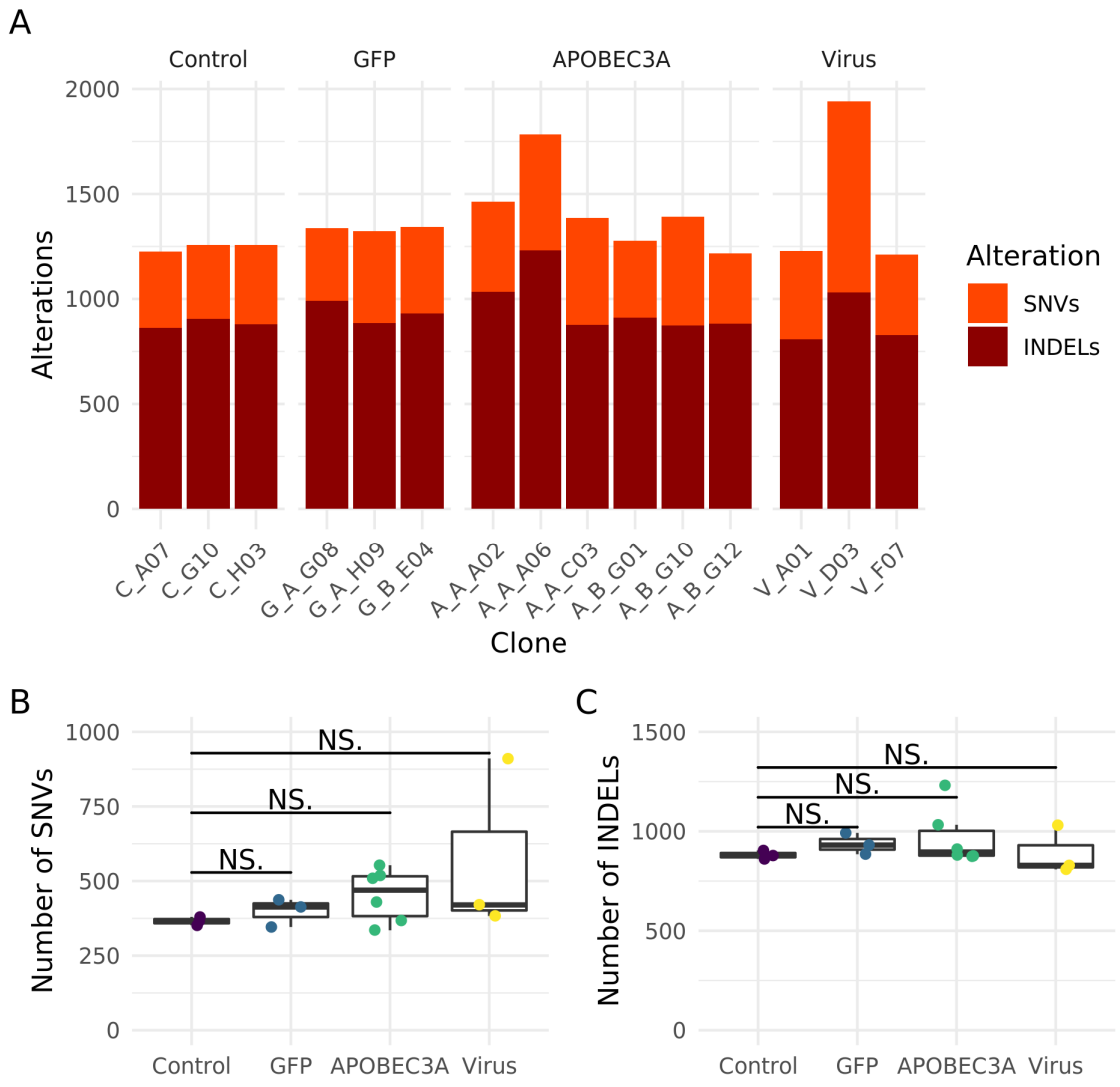


Figure 3.11: Number of clonal INDELs and SNVs identified in sequenced clones. A. Total number of SNVs and INDELs identified in individual clones that were either transfected with a GFP containing plasmid, transfected with a GFP and A3A containing plasmid, infected with VSV-G pseudotyped HIV particles, or untreated. B. Number of SNVs identified, split by treatment. C. Number of INDELs identified in each treatment group. Data points represent individual samples. Two-sided Wilcoxon rank sum test $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

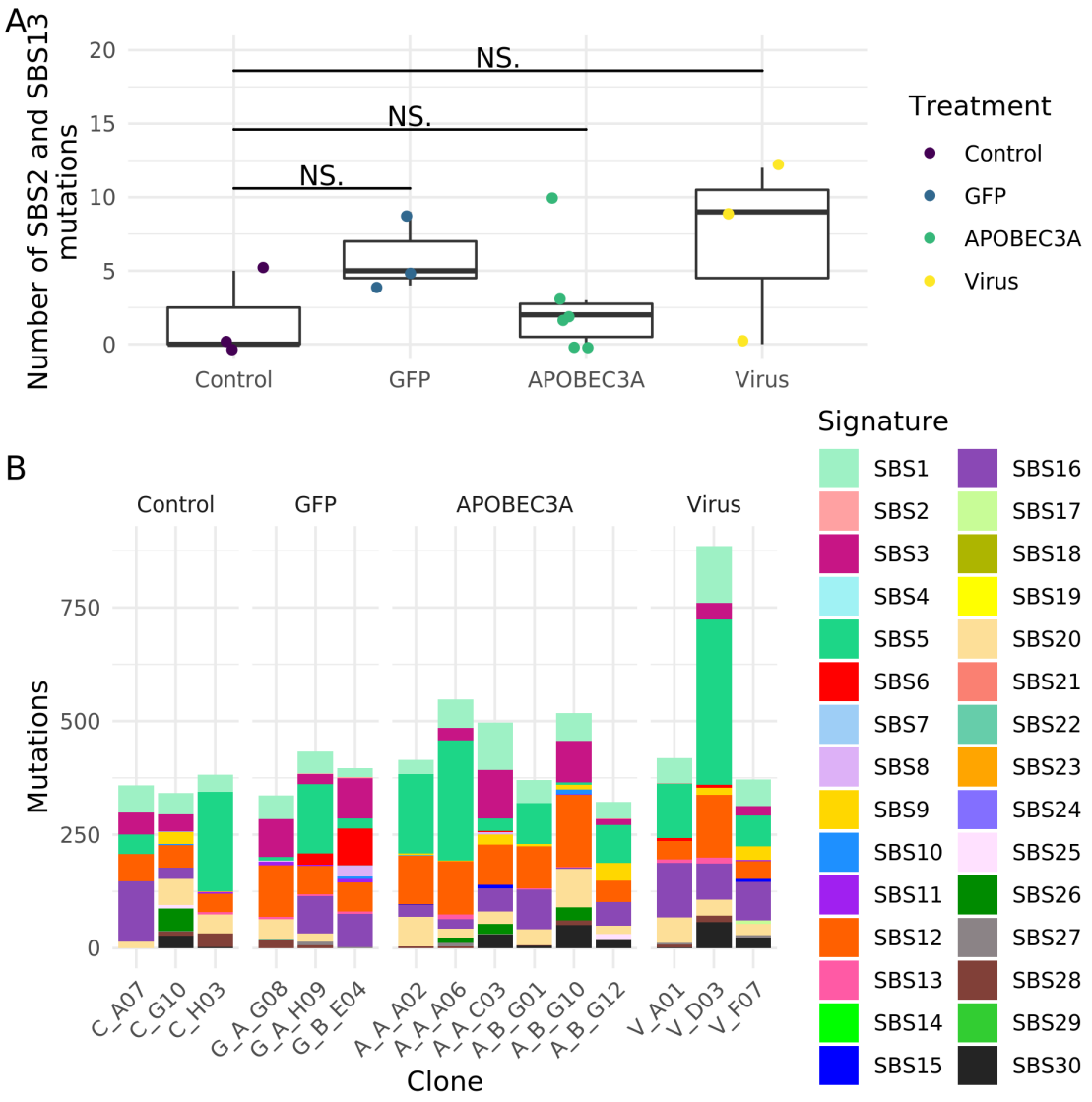


Figure 3.12: Mutational signatures of clonal mutations identified in whole genome sequenced clones. A. The number of SBS2 and SBS13 mutations identified in each sample (points), split by treatment. B. Complement of mutational signatures identified in each clone. Two-sided Wilcoxon rank sum test $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

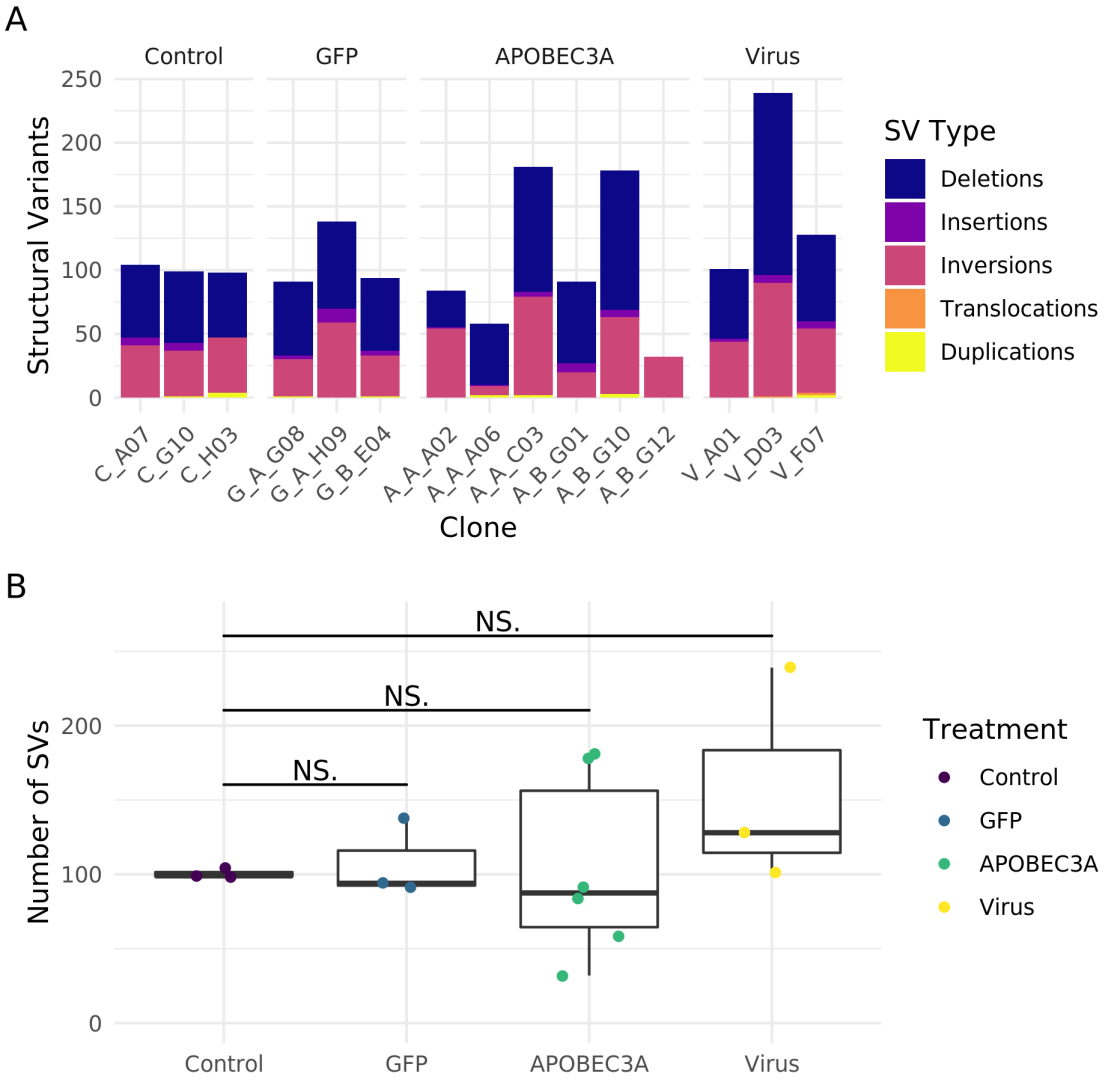


Figure 3.13: Number and type of structural variants identified in each clone (A). B. Total number of structural variants identified, split by treatment. Points represent individual clones. Two-sided Wilcoxon rank sum test $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

3.4.3.3 Copy Number

The number of CNVs in each clone was assessed using an unpublished version of the Battenberg copy number caller developed for use with cell line data. There was no significant difference in the number of copy number segments identified, the average length of each copy number segment, or the percentage of the genome altered by copy number changes between the four conditions that were tested (Figure 3.14).

The results of the copy number analysis were very consistent between different samples, suggesting that the transient transfection and infection did not affect CNV formation.

3.5 Discussion

In this chapter I found that infection of cells with pseudoviral particles led to a significant increase in nuclear A3A and γH2AX 24 hours after infection, quantified via immunofluorescent staining in two different cell lines. I followed this up with RNA sequencing of infected cells at 0, 6, 12, and 24 hours after infection to characterise the DNA damage response in the early stages of infection. However, I did not detect significant changes in expression of any APOBEC3 proteins, and the results suggest that expression of DDR genes may be mildly down-regulated after infection.

Given the increase in nuclear A3A staining in the immunofluorescence experiments, the lack of A3A expression in the RNAseq data was puzzling. It is possible that the results of the immunofluorescence experiments reflect shuttling of A3A from the cytoplasm to the nucleus, rather than an absolute increase in A3A at the protein or mRNA level. However, analysis of the ratio of A3A staining in the nucleus and the cytoplasm did not reveal significant changes following infection (data not shown). Given more time, analysis of content material might elucidate whether any changes in A3A are observed at the protein level rather than the mRNA level, and might also further verify (or contradict) the results of the immunofluorescence

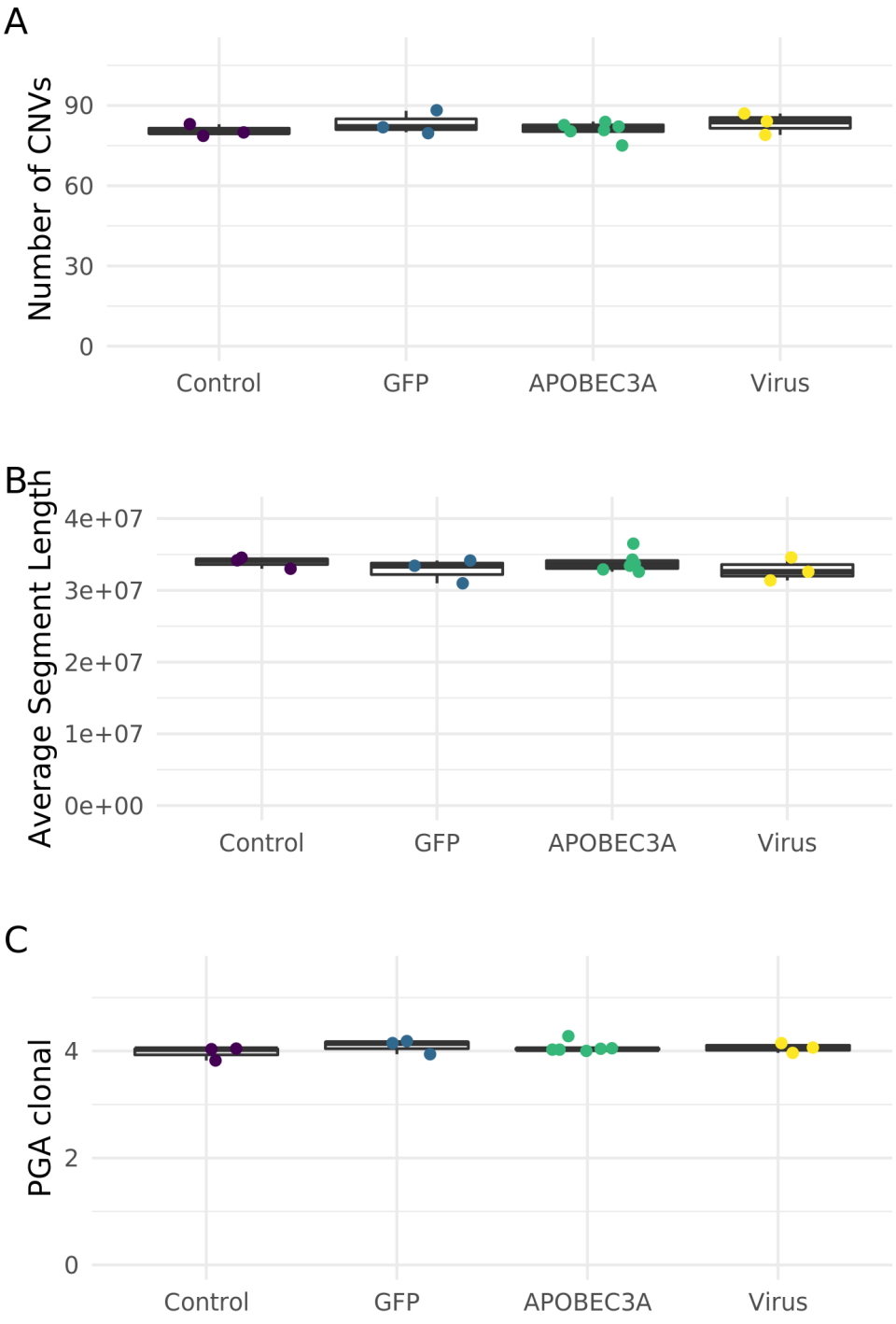


Figure 3.14: A. Number of copy number segments called in each clone. B. Average length of each copy number segment identified. C. Percentage of the genome clonally altered from a copy number state of 2. Points represent indiviudal clones. Comparisons between different groups were not significant and are not displayed. Two-sided Wilcoxon rank sum test p >0.05.

experiments.

It is of course possible that the A3A antibodies used may have cross reacted with other APOBECs, or even other proteins. Cross-reactivity has been described for commercial A3B antibodies, which often label A3A and A3G due to their high sequence similarity (Brown et al., 2019). Commercial A3B antibodies have also demonstrated inconsistent or non-specific labelling (Brown et al., 2019). I trialled two different commercial A3A antibodies which demonstrated very similar staining patterns (data not shown). Although the antibodies were validated by the provider, but given more time, they could have been validated in house by western blotting or further immunofluorescence experiments using knock out cell lines. I think it unlikely that the immunofluorescence data represent an increase in either A3B or A3G following infection, as A3B has not been reported to increase in expression following viral infection (Berger et al., 2011), and the viruses that were used express the viral infectivity factor (Vif), which has been shown to inhibit A3G expression by targeting it for ubiquitination and degradation (Goila-Gaur and Strebel, 2008). Thus, analysis of protein material may prove a crucial experiment.

Another factor that may, in part, explain the discordance between the results is the cell type used. Work by others has demonstrated dramatic increases in A3A mRNA in the early stages of HIV infection in monocytes (Berger et al., 2011), and cells of the myeloid lineage are known to express A3A at high levels (Refsland et al., 2010; Berger et al., 2011). In contrast, these experiments were carried out in retinal pigment epithelial cells (RPE-1) and colorectal cancer cells (HCT116). It is possible that A3A expression in these samples, whether increased or not following infection, remained below the threshold for detection by RNA sequencing, and it may have been more appropriate to carry out these experiments in a cell line known to express A3A.

Interestingly, Shimizu et al. (2018) have described consistent expression of A3B, which is stabilised at the protein level in response to DNA damage. While such

a scenario has not been described for A3A, it would be interesting to investigate whether we see changes in A3A or A3B at the protein level in these samples. While the literature suggests that A3B is not induced in response to IFN-α stimulation of viral infection (Refsland et al., 2010; Berger et al., 2011), it is possible that one might see its stabilisation at the protein level in response to the increased DNA damage defined by the increased γH2AX staining that was observed, as suggested by Shimizu et al. (2018). This is an interesting thought that requires further study, and might, once again, be addressed through analysis of protein samples.

A proteomics screen may also more accurately capture any changes in DNA damage signalling, which is more commonly reflected in post-translational modifications rather than changes in expression (Derks et al., 2014). Nevertheless, I did observe some changes in the expression of DDR genes in my RNAseq data. At 24 hours post infection I found that DDIT4, or DNA damage inducible transcript 4, also referred to as REDD1 and RTP801, was signficantly differentially expressed with an LFC of 2.2. DDIT4 is thought to be a target of p53, and is induced in response to DNA damage and cellular stress (Ellisen et al., 2002). This is one of the few indicators of the presence of DNA damage or cellular stress in the RNAseq data. Up-regulation of DDIT4 has previously been reported in HIV infected T-cells that are resistant to apoptosis (Reuschel et al., 2015; Yin et al., 2004). Thus up-regulation of DDIT4 in our cells may represent a survival response, following infection.

When I investigated differential expression of DDR genes, BOLA2 emerged as a gene that was significantly differentially expressed at 12 hours post infection with an LFC of -9.4, and an LFC of -1.6 at 24 hours post infection. BOLA2 is an Iron-Sulfur cluster chaperone (Frey et al., 2016). Interestingly, BOLA2 has also been found to be down-regulated in HPV-positive squamous cell carcinoma of the head and neck compared to those that are HPV-negative (Martinez et al., 2007).

I also carried out gene ontology (GO) analysis to determine the biological impact of the transcriptional changes that were observed. The results of the GO analysis

appear broadly consistent with extended results reported by Chang et al. (2011), with the exception of certain terms that are likely to be cell type specific, such as "T cell differentiation" and "Lymphocyte differentiation". Chang et al. (2011) also report similar patterns in terms of the number of genes that are differentially expressed at each time point, with 24 hours post-infection serving as the time point with the most differentially expressed genes. Trypsteen et al. (2016) report a similar trend, with the number of differentially expressed genes increasing over time. 24 hours post infection is thought to coincide with peak viral replication and the late phase of HIV infection (Chang et al., 2011; Mohammadi et al., 2013).

Many of the GO terms that were enriched in my data set referred to the cell cycle or cell cycle processing. I suspected that this may in part have been due to the nature of the experiment, which involved monitoring cells over an extended period of time during which some cells will have proceeded through the cell cycle. However, GO terms such as "Cell cycle" and "Chromosome organisation" and pathways associated with cell cycle regulation were also reported as significantly enriched by Chang et al. (2011) and Trypsteen et al. (2016) who both studied HIV infected CD4+ T-cell lines at several time points post-infection with a time-matched control, suggesting that these cell cycle effects may indeed be due to infection. Enrichment of GO terms associated with cell cycling is consistent with reports that HIV arrests the cell cycle in G2 during infection (Goh et al., 1998). This is also reflected in the presence of the "organelle" GO term as highly enriched at 24 hours post-infection. The "organelle" term has been found to be enriched for genes that are expressed at lower levels in the G0/G1 phase of the cell cycle, and at higher levels at G1/S and M phases (Liu et al., 2017). Interestingly, Trypsteen et al. (2016) suggest that enrichment of terms related to cell cycling may reflect responses to DNA damage as a result of viral infection and integration.

I also sought to assess the effects of viral infection and A3A expression on the genome. Similar to the results of the RNAseq analysis, the WGS analysis of A3A

transfected and HIV pseudovirus infected clones also provided some interesting but unclear results. Overall, there seemed to be a trend towards increasing numbers of SNVs, INDELs, and SVs in clones transfected with A3A or infected with virus; however, this trend was not statistically significant. This may in part be due to the way in which these experiments were conducted, and due to small sample size, which limited statistical power.

The intention was to investigate whether one could detect the occurrence of SBS2 and SBS13 mutations, as well as an increase in genomic instability, as a direct consequence of viral infection or transient A3A expression. In this experiment I extracted DNA from cells as soon as sufficient numbers of cells were available, which meant that cells were grown for 22-32 days, which may have resulted in initial differences in mutational signatures and genome stability being masked by accumulation of further mutations. Brody et al. (2018) have estimated that the mutation rate of RPE1 cells lies around 39 SNVs per cell division, which would account for an expected number of 858-1,248 mutations in these clones, in the absence of other mutational processes. I identified between 586 and 1,733 SNVs in my analysis, which is broadly consistent with what might be expected based on the findings of Brody et al. (2018). To mitigate the effect of mutation accumulation I focused my analysis on SNVs with a VAF between 0.3 and 0.7, as a measure of clonal mutations. This decreased the number of SNVs surveyed to 314-871 SNVs per clone, but may have still included SNVs that occurred very early in culture.

These results are also complicated by the fact that one cannot quantify the degree of A3A expression that the cells were exposed to before sorting. Toxicity of A3A over-expression has been reported by several groups (Land et al., 2013; Landry et al., 2011). It is possible that I may have inadvertently selected for cells that did not retain A3A activity after sorting, or selected cells that were only exposed to low levels of A3A expression, by selecting clones that proliferated well. In addition, cell cycle arrest is a common response to DNA damage (Landry et al., 2011), further

suggesting that the clones that were selected were those that had sustained less damage.

The cell line used for these experiments is also a factor to consider here. The rationale for using RPE-1 cells was to minimise the amount of "noise" mutations and alterations resulting from other mutational processes, in order to focus the analysis on SBS2 and SBS13 mutations. RPE-1 cells are chromosomally stable, non-transformed, and have a relatively low mutation rate and a flat mutation spectrum (Brody et al., 2018). It is possible that other cancer cell lines may have tolerated A3A expression better.

Alternative approaches to investigating the effects of APOBEC3 enzymes on the genome have involved growing cells known to carry APOBEC3 mutations for extended periods of time allowing accumulation of further APOBEC3 mutations, which have been proposed to occur in bursts of activity (Petljak et al., 2019). Others have designed a cell line with doxycycline inducible A3B, and subjected cells to 10 rounds of toxic A3B exposure, each lasting 7 days (Akre et al., 2016). Such a system would mimic the burst-like behaviour described by Petljak et al. (2019) while mitigating the toxic effects of APOBEC3 exposure, allowing for increased accumulation of mutations over time. While systems of this kind could help us establish a link between A3A expression and A3A associated mutational signatures and genome stability, they do not address whether transient A3A upregulation following viral infection results in the presence of SBS2 and SBS13 mutations and genomic instability.

In this study I investigated the effects of HIV based pseudoviral particles, as these were the tools available to me. In future, it would be interesting to assess the effects of viruses that are more common in the population. The relationship between HPV infection, which is responsible for more than half of all infection driven cancers in women, and APOBEC3 mutations has previously been studied (Plummer et al., 2016; Gillison et al., 2019). However, I am particularly interested in investigating

the effect of viruses that have not previously been linked to cancer, to see if they may indirectly contribute to cancer initiation or progression through APOBEC3 mediated mutagenesis.

Chapter 4

Identification of virus-induced cell-cell fusions in whole genome sequencing data

4.1 Introduction

Evidence for cell fusion in human cancers is limited and, to date, no group has attempted to identify cell fusions in whole genome sequencing data. It has been noted by Weiler and Dittmar (2019) that current approaches to identifying tumour cells originating from cell fusion, such as identification of specific cell surface markers or expression of marker genes (discussed further in Section 1.3.1), may not be suitable for use with human tumour biopsies, and that the identification of naturally occurring tumour cell fusions is fraught with difficulty. Weiler and Dittmar (2019) point out that the fusion of cells that do not express any lineage specific markers or two cells of the same type (e.g. homotypic fusion between two tumour cells) will result in cells that cannot be identified through marker expression profiling. Thus, the identification of such hybrids will rely on the use of genomic, epigenomic, or transcriptomic data. In this chapter I propose a number of measures derived from whole

genome sequencing data to identify tumour samples containing cells that may be derived from fusion between a tumour cell and a normal cell. I focus my approach on these tumour-normal hybrids due to the observable and quantifiable differences between their genomes.

4.2 Materials

4.2.1 PCAWG Data Set

The PanCancer Analysis of Whole Genomes (PCAWG) data set consist of 2,658 whole genome sequenced tumour and matched normal samples, split across 38 different tumour types (Campbell et al., 2020). Copy number calls generated by the Battenberg algorithm and information relating to the multiplicity of mutations in the PCAWG data set were shared by Dr. Stefan Dentro on behalf of the PCAWG Evolution and Heterogeneity working group. The majority of the PCAWG data can be accessed through the ICGC at http://dcc.icgc.org/pcawg/ . Access to controlled data was granted by the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO) for the ICGC portion of the PCAWG data, and by The Cancer Genome Atlas (TCGA) Data Access Committee for the TCGA portion of the data. Chromothripsis data used in this chapter was made available through the PCAWG Evolution and Heterogeneity working group, and published in Campbell et al. (2020).

4.2.2 Hartwig Medical Foundation (HMF) Data Set

The Hartwig Medical Foundation (HMF) has curated a database of 3,679 metastatic tumour and matched normal samples, as well as related clinical data. The HMF data set is the largest collection of whole genome sequencing data for metastatic solid tumours and is described in detail in Priestley et al. (2019). The data can be accessed by submitting a data access request to the HMF at https://www.

hartwigmedicalfoundation.nl/en/applying-for-data/. For this project the HMF provided both metadata and somatic data including gene level CNV calls, somatic CNV calls, tumour purity information, quality control data, somatic variant calls, somatic structural variant calls, and circos plots. A detailed description of the data types available from the HMF can be found in the HMF Data Access Request Guide (https://github.com/hartwigmedical/documentation/blob/master/data-access-req
md).

Mutational signature extraction was carried out by Dr. Andreas Gruber using SigProfiler (Alexandrov et al., 2013).

4.3 Theory

4.3.1 Variation in Mutation Density

Given the observation that cancer cells generally carry more mutations than healthy cells, it follows that cancer cells will have a higher mutation density compared to normal cells (Lee-Six et al., 2019; Moore et al., 2020). Therefore, one might expect the product of a cell fusion between a tumour cell and a healthy cell to display an intermediate level of mutation density, with regions derived from the normal cell having a relatively low mutation density, and regions derived from the cancer cell having a higher mutation density. In addition to this, it is accepted that cells often lose genetic material following a whole genome duplication event, and are thus likely to lose DNA following a cell fusion event (Figure 4.1; Laughney et al. (2015); Zack et al. (2013); Bielski et al. (2018); Carter et al. (2012)).

To identify tumours in which such a situation may have arisen, I examine the coefficient of variation of the mutation density in segments of differing copy number across the genome. I expect to see a high coefficient of variation of the mutation density in these tumours when looking at regions where DNA has been lost. Regions in which the DNA of the normal cells is retained and DNA from the cancer cell is

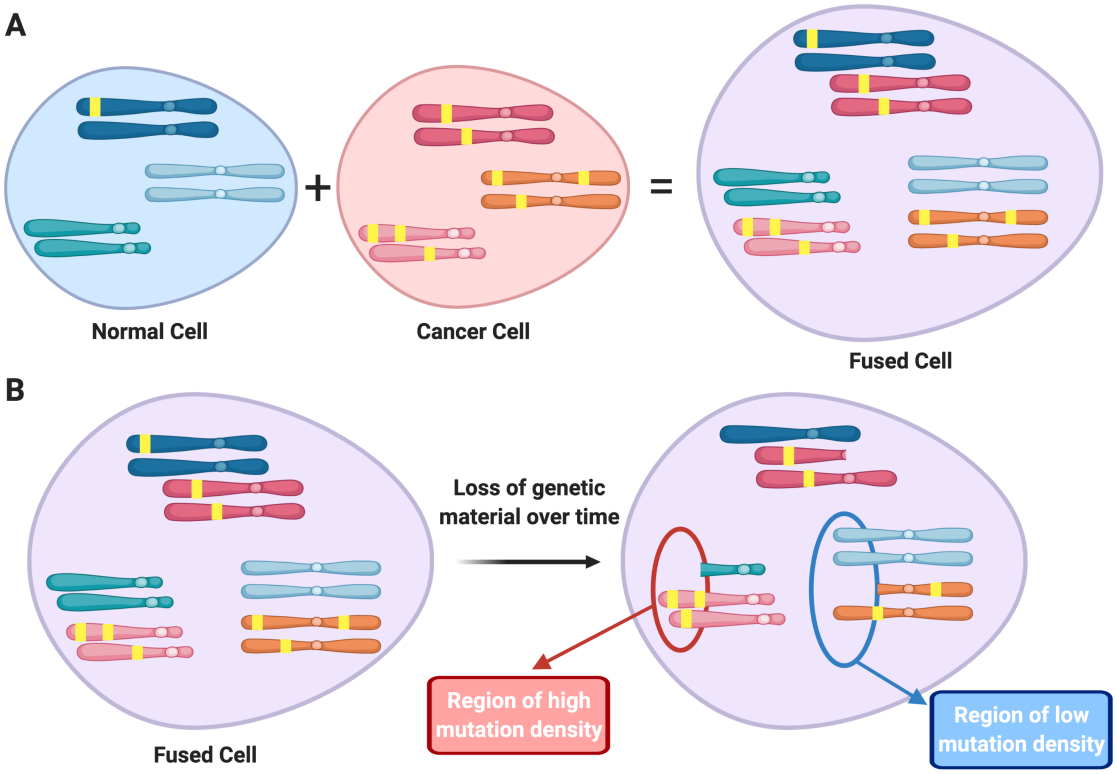


Figure 4.1: Differences in mutation density following loss of DNA. A) In the event of cell fusion between a normal cell and a cancer cell, the genome of the resulting hybrid cell contains chromosomes derived from both parental states. B) Upon loss of genetic material over time, the mutation density will vary based on the origin of the remaining genomic regions. Regions originating from the cancer cell will be expected to have a higher mutation density (red spotlight) than regions retained from the normal cell (blue spotlight). Yellow bars represent mutations. Image created with BioRender.com.

lost are expected to have a low mutation density, whereas regions containing only DNA derived from a tumour cell are expected to have higher mutation densities (Figure 4.1).

When these differences in density are examined across the genome, this trend should be revealed by the coefficient of variation of the mutation density. For this analysis I examine regions of copy number 1+1 and 2+1, as they are expected to have different combinations of parental DNA. For example, 1+1 regions may contain segments that either originate entirely from a cancer cell, entirely from a normal cell, or one segment from each, whereas 2+1 regions will have retained two copies from the cancer cell and one from the normal cell or vice versa (Figure 4.2).

If these changes in segment copy number are due to a whole genome duplication event or the fusion between two cancer cells with similar mutational density I expect the coefficient of variation of the mutation density to be low, as these segments will likely be very similar, if not identical in terms of mutation density.

4.3.2 Proportion of Mutations With Multiplicity 1

The multiplicity of mutations has previously been used to provide information relating to the timing of whole genome duplication events (Gerstung et al., 2020). It is assumed that mutations occurring before a whole genome duplication event will have a multiplicity of 2, as they will be multiplied in this duplication event, whereas mutations arising after the duplication will have a multiplicity of 1 (i.e. they will only be present on one chromosome). In the event of a cell fusion, I expect the multiplicity of the majority of (if not all) mutations to be 1, as it is unlikely that a cancer cell and a normal cell will carry the same mutations (Figure 4.3).

This should distinguish cells undergoing cell fusion from cells which have undergone a whole genome duplication event by another mechanism, with the exception of cells which have undergone very early whole genome duplication events followed by the accumulation of mutations.

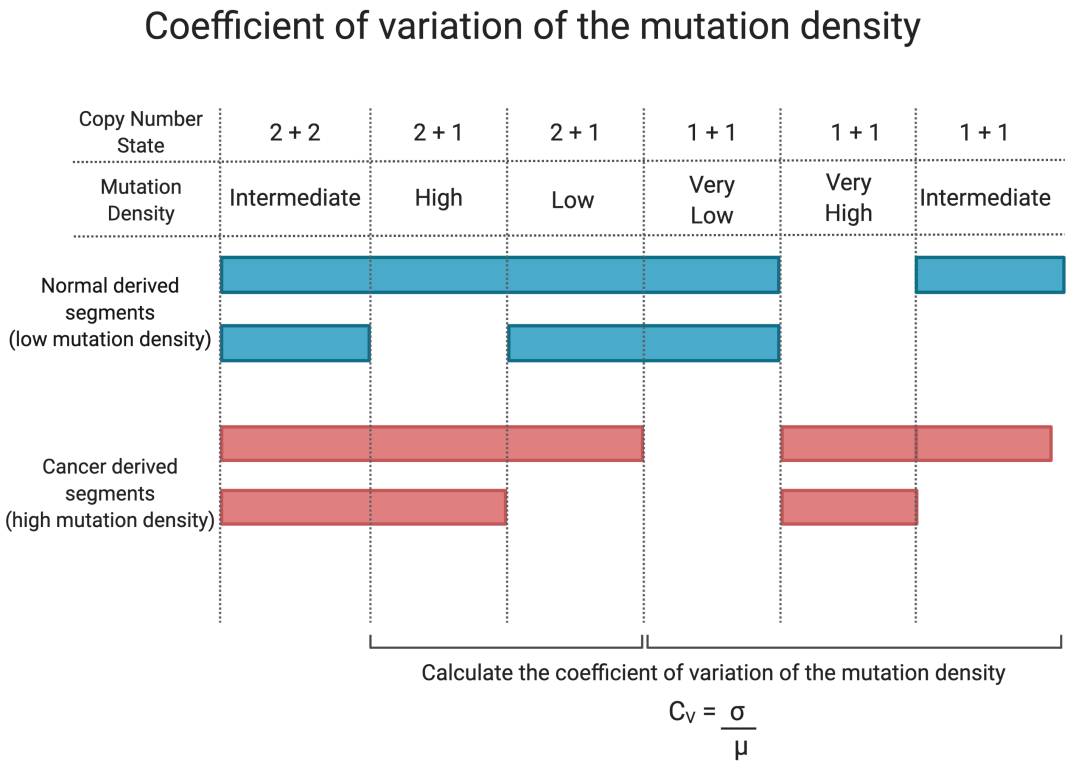


Figure 4.2: Coefficient of variation of the mutation density. Regions in a tetraploid cell originating from a cell fusion event between a normal cell and a cancer cell are expected to have variable mutation density dependent on the origin of the copy number regions that are lost or gained. I specifically focus my analysis on copy number states in which one or two chromosomes have been lost, i.e. copy numbers 2+1 and 1+1. Areas of the genome in which more segments originating from the normal cell (blue) are retained are expected to have a lower mutation density. In contrast, retention of regions originating from the cancer cell (red) are expected to have a higher mutation rate. When the coefficient of variation is calculated across all of these segments, tumours originating from a cell fusion between a normal cell and a tumour cell are expected to have a high coefficient of variation, indicating the presence of some regions with higher mutation density and some regions with lower mutation density. σ is defined as the standard deviation, and μ is defined as the mean. Image created with BioRender.com.

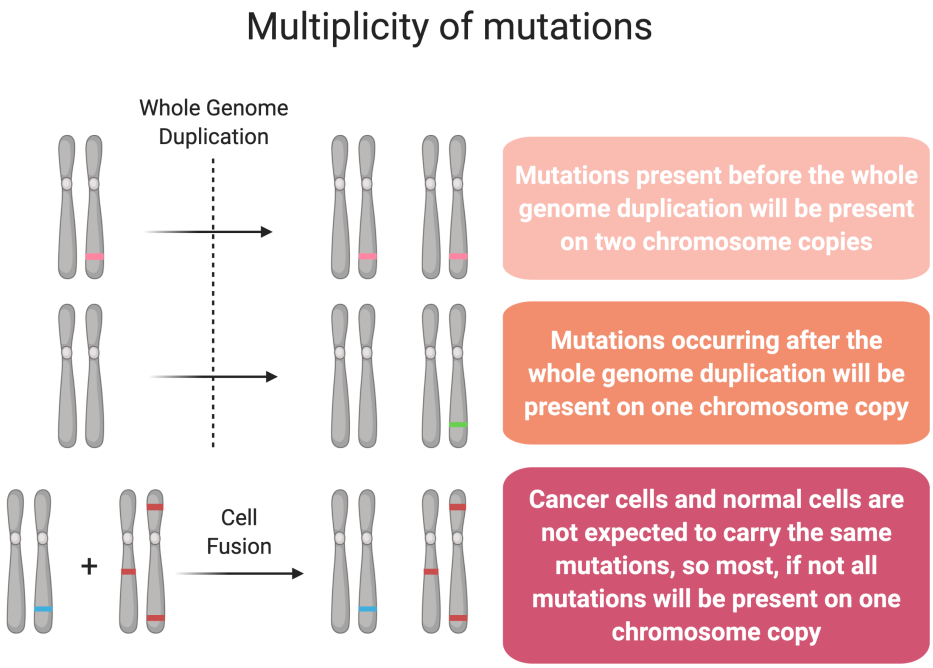


Figure 4.3: Multiplicity of mutations and whole genome doubling. Pink mutations represent mutations present before a WGD event, which will appear on two chromosome copies after WGD. Green mutations represent mutations arising after a WGD event; these will be present on one copy. Red mutations represent mutations originating from a tumour cell, and blue mutations represent mutations originating from a normal cell. Image created with BioRender.com.

4.3.3 Average Ploidy

Tumours which have undergone a whole genome duplication event can be identified by their increased ploidy. I expect this to be true for the product of a cell fusion as well. While these cells may initially occupy a tetraploid state, I expect the ploidy to be lower than n = 4 due to the loss of genetic material that tends to follow any duplication of the genome (see Sections 1.1.1 and 1.2.2.2).

4.3.4 Presence of Viral DNA

The presence of viral DNA in tumour whole genome sequencing data, when combined with these measures, also provides support for the hypothesis that the main population in the tumour has arisen from virus-induced cell-cell fusion. Many viruses integrate into the host genome following infection. However, it is difficult to time when the virus was present in the tumour, as the presence of its DNA may indicate a historic infection which involved integration of the viral genome, or the viral infection may simply have arisen shortly before the tumour was sampled. In addition to this, the potential fusogenicity of the viral species identified should also be considered. Only fusogenic viruses are expected to be involved in the fusion of two cells.

It should also be noted that RNA viruses which have not been reverse transcribed or integrated into the host genome may be missed in this analysis, as samples are treated with RNAse before DNA extraction and sequencing, removing the viral RNA genome. Thus, such viruses may be present in tumour samples but go undetected. Attribution of a fusion event to a virus is further complicated by the ability of some viruses to fuse cells without infecting them (fusion from without), leaving no trace of the virus' involvement (Duelli and Lazebnik, 2007).

4.3.5 Presence of Mutational Signatures

In mutation signature analysis, signatures 2 and 13 have been attributed to the activity of APOBEC3A and APOBEC3B enzymes (Section 1.5.1; Alexandrov et al. (2013)). I did not find any evidence of increased numbers of APOBEC mutations in samples containing viral DNA in the PCAWG data set in Chapter 2. Nevertheless, I investigate whether there is an increased level of APOBEC mutation signatures in samples that may have originated from a fusion event.

Mutational signatures may also enable us to distinguish segments originating from a normal cell and a tumour cell. Previous studies have demonstrated that mutations identified in healthy tissues, and organoids derived from adult stem cells, are primarily attributed to SBS1 and SBS5, and less frequently to SBS18 and SBS40 (Blokzijl et al., 2016; Lee-Six et al., 2019; Moore et al., 2020). Tumour tissue, on the other hand, is typically characterised by a much wider array of different mutational signatures, attributed to different mutational processes, and exposure to various mutagens (Alexandrov et al., 2013, 2020).

I postulate that segments with low mutation density, which I presume have arisen from normal cells, will mainly be characterised by signatures SBS1 and SBS5, with the possible addition of SBS18 and SBS40. Segments with higher mutation density, which I think may have originated from a tumour cell, are expected to carry mutations from other signatures in addition to the aforementioned signatures 1, 5, 18, and 40, which occur in healthy cells.

4.3.6 Timing Whole Genome Duplication

In order to further distinguish between fusion-containing tumours and tumours which have undergone a whole genome duplication by other means, I looked at the timing of the whole genome duplication (WGD) event. The timing of a WGD can be inferred from the ratio of the number of mutations present on two chromosomes (n_2), to the number mutations present on one chromosome (n_1) for copy

number states of 2 + 1, 2 + 2, and 2 + 0, using the following equations (Gerstung et al., 2020):

$$\text{Copy number } 2+1\text{: } T = \frac{3n_2}{2n_2 + n_1} \tag{4.1}$$

$$\text{Copy number } 2+2\text{: } T = \frac{2n_2}{2n_2 + n_1} \tag{4.2}$$

$$\text{Copy number } 2+0\text{: } T = \frac{2n_2}{2n_2 + n_1} \tag{4.3}$$

The value of T will fall between 0 and 1, with lower values indicating early WGD and values closer to 1 indicating late WGD (Gerstung et al., 2020). As discussed in Section 4.3.2, I expect the majority of mutations in fusion-containing tumours to be present on one chromosome copy, which would present as an early WGD event.

4.3.7 Detecting Premature Chromosome Condensation in Sequencing Data

Premature chromosome condensation (PCC) has been proposed as one of the mechanistic origins of chromothripsis as it is a catastrophic event that results in shattering and reassembly of chromosomes (Pantelias et al., 2019). Thus, I investigate whether the samples that I have identified as potential fusions contain higher numbers of chromothripsis events than the other samples in the PCAWG data set. I also investigate differences in the number of structural variants in each sample.

As chromothripsis, and PCC, have also been associated with replication of micronuclei (Pantelias et al., 2019), I reasoned that the effects of PCC resulting from cell fusion might be differentiated from PCC associated with micronuclei based on

the number of chromosomes affected by chromothripsis events. Micronuclei typically only contain a single chromosome or fragments of a chromosome (Leach and Jackson-Cook, 2001; Lindberg et al., 2008), whereas cell fusion is expected to affect a whole complement of chromosomes.

4.4 Methods

All analyses were carried out using R version 3.6.2, unless stated otherwise (R Core Team, 2018). Plots were generated using version 3.3.0-2 of the `ggplot2` package, version 0.6.0 of the `ggsignif` package, and version 1.4.0 of the `UpSetR` package (Wickham, 2016; Ahlmann-Eltze, 2019; Gehlenborg, 2016).

4.4.1 Processing of Samples in the Hartwig Medical Foundation Data Set

In order to be able to make fair comparisons between the analysis of tumours in the PCAWG and HMF data sets, a number of pre-processing steps were carried out to ensure that the structure of the HMF data was similar to that of the PCAWG data. The processing that was carried out is described in detail in Section B.1 of Appendix B. In short, data processed using the HMF's core analysis pipeline (Hartwig Medical Foundation, 2020) was reformatted to resemble the input and output files associated with the Battenberg and DPClust algorithms (Dentro et al., 2017; Nik-Zainal et al., 2012). These "pseudo-Battenberg" files were then used to determine the multiplicity of the mutations identified for each tumour, and subsequently used to calculate the fusion measures that were established.

4.4.2 Quantifying Variation in Mutation Density

Mutation density was calculated as the number of mutations in a copy number segment divided by the size of the copy number segment in base pairs (bp). Only

segments of copy numbers 1+1 and 2+1 were considered, as these are the segments in which I expect to observe variation in the mutation density, depending on the pattern of chromosomal losses discussed in Section 4.3.1 and illustrated in Figure 4.2. I also focused my analysis exclusively on segments larger than 10 Mb in size to ensure that the results were not influenced by smaller segments that are perhaps less likely to have mutations, and more likely to be influenced by local features such as chromatin state, which can influence mutation rate (Schuster-Böckler and Lehner, 2012). The coefficient of variation of the mutation density (CV_{MD}) was then calculated using the following equation:

$$\text{Coefficient of Variation} = \frac{\sigma}{\mu} \tag{4.4}$$

where σ is defined as the standard deviation, and μ is defined as the mean.

The coefficient of variation was calculated separately for 1+1 and 2+1 segments, as well as being calculated across all 1+1 and 2+1 segments together.

4.4.3 Calculating the Proportion of Mutations with Multiplicity 1 (PM_1)

The proportion of mutations of multiplicity 1 was calculated for 2+1 and 2+2 segments larger than 10 Mb by dividing the number of mutations found on 1 chromosome copy by the total number of mutations in that segment. This value was then averaged across all segments that were considered, both separately for 2+1 and 2+2 segments, as well as jointly across all 2+1 and 2+2 segments.

4.4.4 Average Ploidy

The average ploidy of each tumour was calculated as the copy number of each segment multiplied by the segment size, divided by the sum of the size of all of the

copy number segments in the tumour.

4.4.5 Timing Whole Genome Duplication

The timing of whole genome duplication was carried out using the equations proposed by Gerstung et al. (2020), and described in Section 4.3.6. I combined the timing estimates T for each of the different copy number states to calculate an average time of duplication, weighted by the number of mutations, for each tumour, as follows:

$$\text{Weighted T} = \frac{(T_{2+1} * n_{2+1}) + (T_{2+2} * n_{2+2}) + (T_{2+0} * n_{2+0})}{n_{2+1} + n_{2+2} + n_{2+0}} \tag{4.5}$$

where n represents the number of mutations used.

4.4.6 Identifying Samples Potentially Containing Fusions

In order to identify samples in which cell fusion between a cancer cell and a normal cell may have occurred, I tried multiple different thresholds and approaches to thresholding the measures described above.

4.4.6.1 3rd Quartile Cutoff

The first approach that was taken for thresholding the data was to introduce a cutoff point at the 3rd quartile of the values of the coefficient of variation of 1+1 and 2+1 segments (CV_{MD}) and the proportion of mutations of multiplicity 1 (PM_1) in 2+1 and 2+2 segments across the entire data set. I also used a threshold of 2.5 for ploidy in all of the analyses, as I expect that tumours originating from a cell fusion event will have average copy number somewhat above 2.

4.4.6.2 Standard Deviation Based Cutoffs

Following visual inspection of the data and its relation to the standard deviation, I decided to threshold PM_1 at 1 standard deviation above the mean and CV_{MD} at 2 standard deviations above the mean.

4.4.6.3 Precision and Recall Based Cutoffs

I tried two different precision and recall based thresholds. First, I used the presence of viral DNA in a tumour as the ground truth, to try to maximise the number of tumours containing viral DNA in the group of samples that I believe may contain cell fusions. Second, I used the classification of a tumour as metastatic or recurrent as the ground truth. Virus presence and a classification as metastatic or recurrent tumours were used for this purpose as these may be defining features of tumours resulting from cell fusion, and would be expected to be present at a higher proportion in tumours originating from cell fusion.

The optimum threshold for precision and recall based approaches was selected as the threshold with the highest F1 score. Precision and recall thresholds were calculated separately for PM_1, CV_{MD}, and ploidy. The three thresholds were then applied to the data, to select a subset of samples that exceeded all three thresholds.

4.4.6.4 Introducing Cutoff Thresholds by Eye

After visualising the distribution of PM_1 and CV_{MD}, thresholds were applied to the data by eye at places at which the distribution appeared to have a natural threshold.

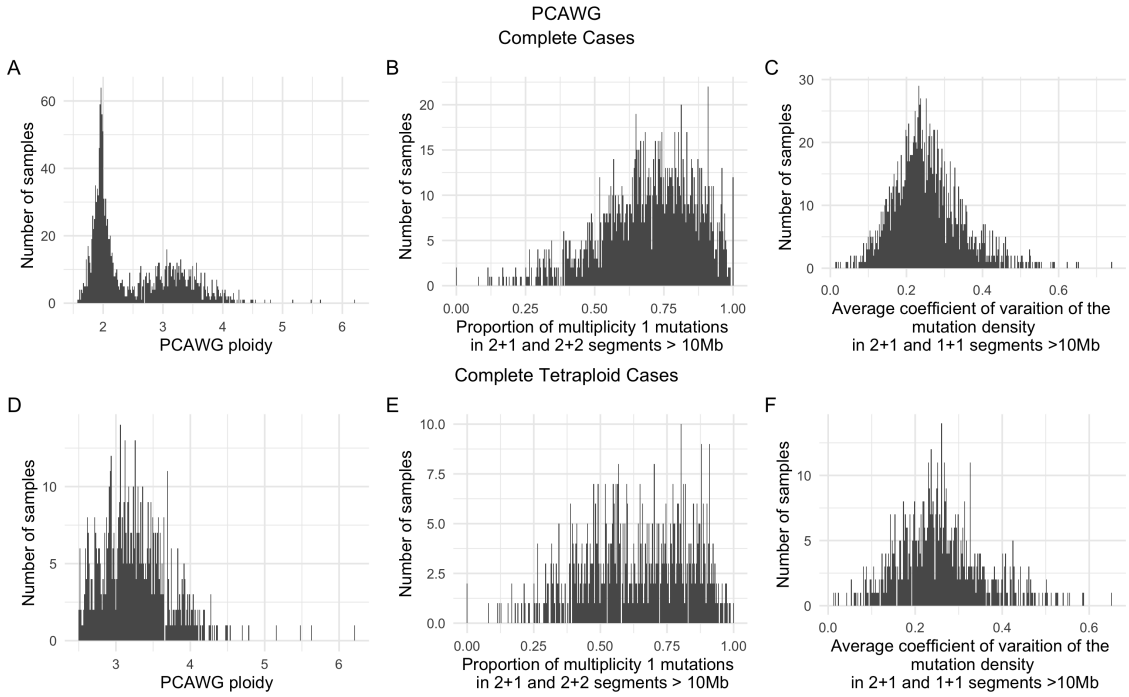


Figure 4.4: Overview of fusion measures calculated for tumours in the PCAWG data set. Panels A-C depict the distribution of the fusion measures calculated for all sample in the PCAWG data set for which I was able to calculate these measures. Panels D-F depict the fusion measures calculated for tetraploid tumours. (A-C n = 1,625, D-F n = 675)

4.5 Results

4.5.1 Identifying Cell-Cell Fusions in the PCAWG Data Set

4.5.1.1 Identifying Samples Potentially Containing Fusions

To identify samples in which a cell fusion event may have occurred, I applied a number of thresholds to the CV_{MD} and PM_1 with the aim of selecting outlier samples with particularly high values for these measures. The different thresholds that were used are depicted in Figure 4.5, and consisted of two precision and recall based thresholds, a threshold that was determined by eye, a threshold at the third quartile of the data, and a threshold involving a number of standard deviations (described in further detail in Section 4.4.6).

After applying various thresholds to the data individually, I investigated the overlap between the samples identified using each of the thresholds, and decided to further investigate samples that passed at least three different thresholds (Figure 4.6).

4.5.1.2 Summary of Samples Potentially Containing Fusions

16 samples within the PCAWG data set passed three or more thresholds designed to identify samples that were outliers with respect to their CV_{MD} and PM_1 values (Figure 4.7). Interestingly, Paediatric Brain Cancer (PBCA-DE) and Ovarian Cancer (OV-AU and OV-US) each accounted for a quarter of the samples that were identified. In addition, I identified two Esophageal Adenocarcinomas (ESAD-UK), one Breast Cancer (BRCA-US), one Bone Cancer (BOCA-UK), one Bladder Urothelial Carcinoma (BLCA-US), one Gastric Cancer (GACA-CN), one Melanoma (MELA-AU), and one Pancreatic Endocrine Neoplasm (PAEN-AU). The bladder cancer that was identified as potentially having undergone a fusion event was found to contain Alphapapillomavirus 7 or HPV18, and the Melanoma sample was identified as a metastatic sample. The age at diagnosis ranged from 9 to 80, with the average

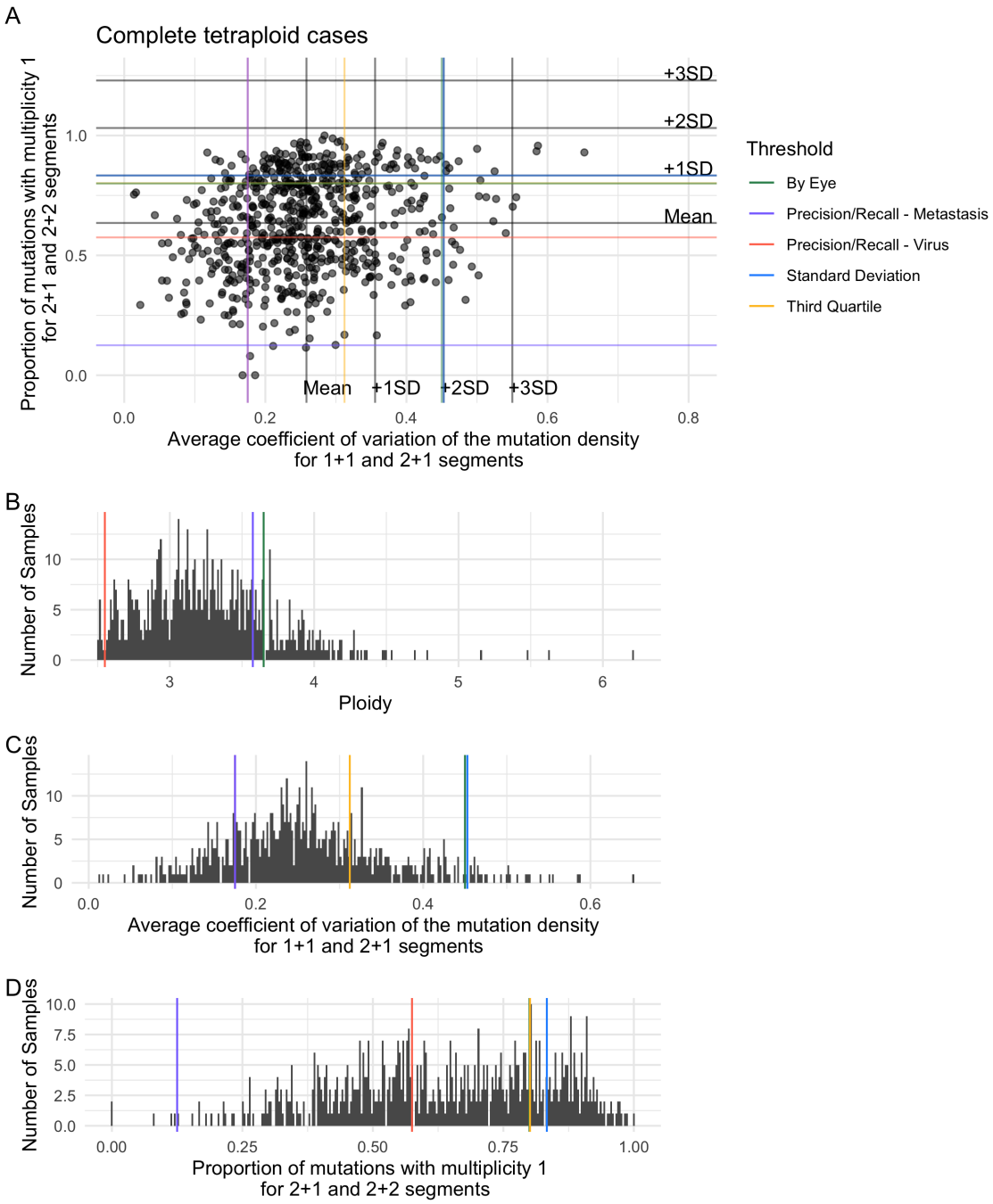


Figure 4.5: Thresholds applied to the PCAWG fusion measures for tetraploid samples. (n = 675)

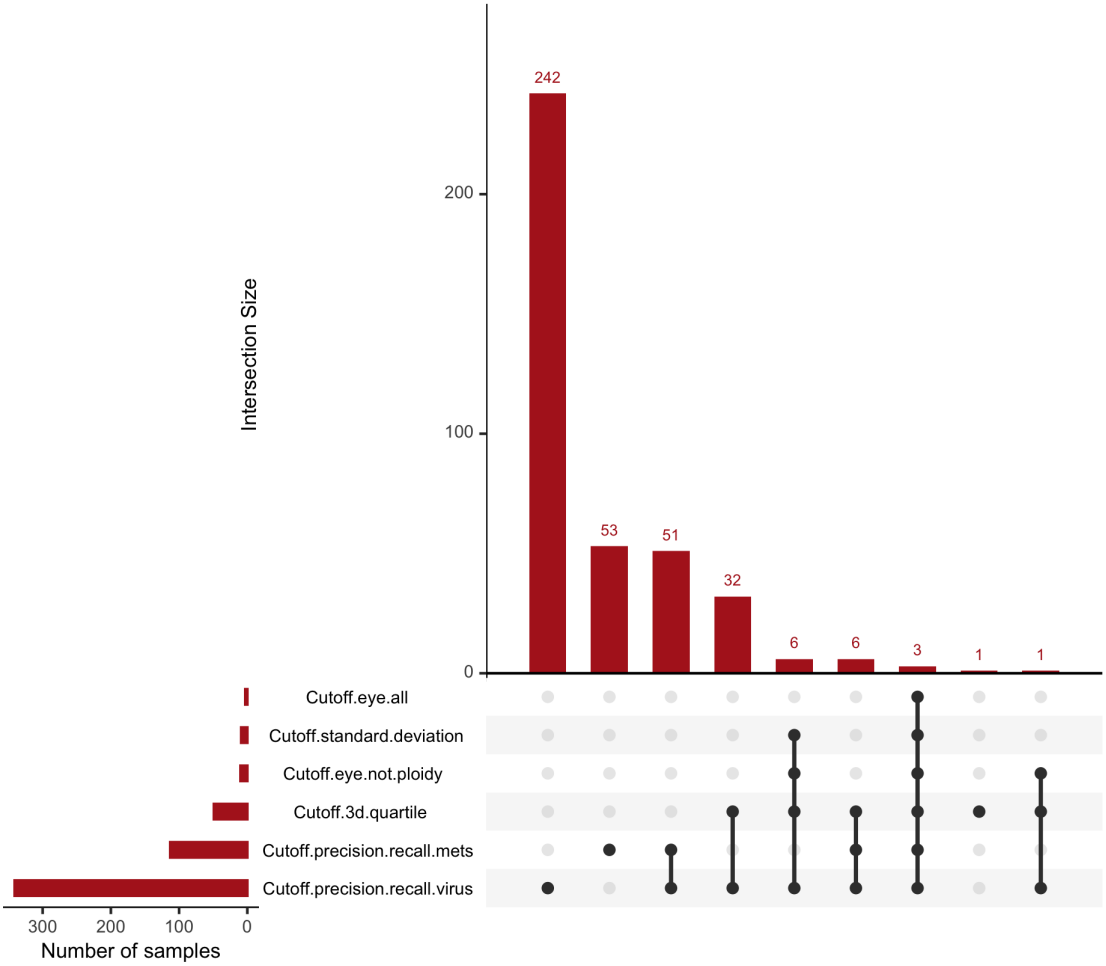


Figure 4.6: Overlap between sample sets identified by applying different thresholds to the PCAWG fusion measures.

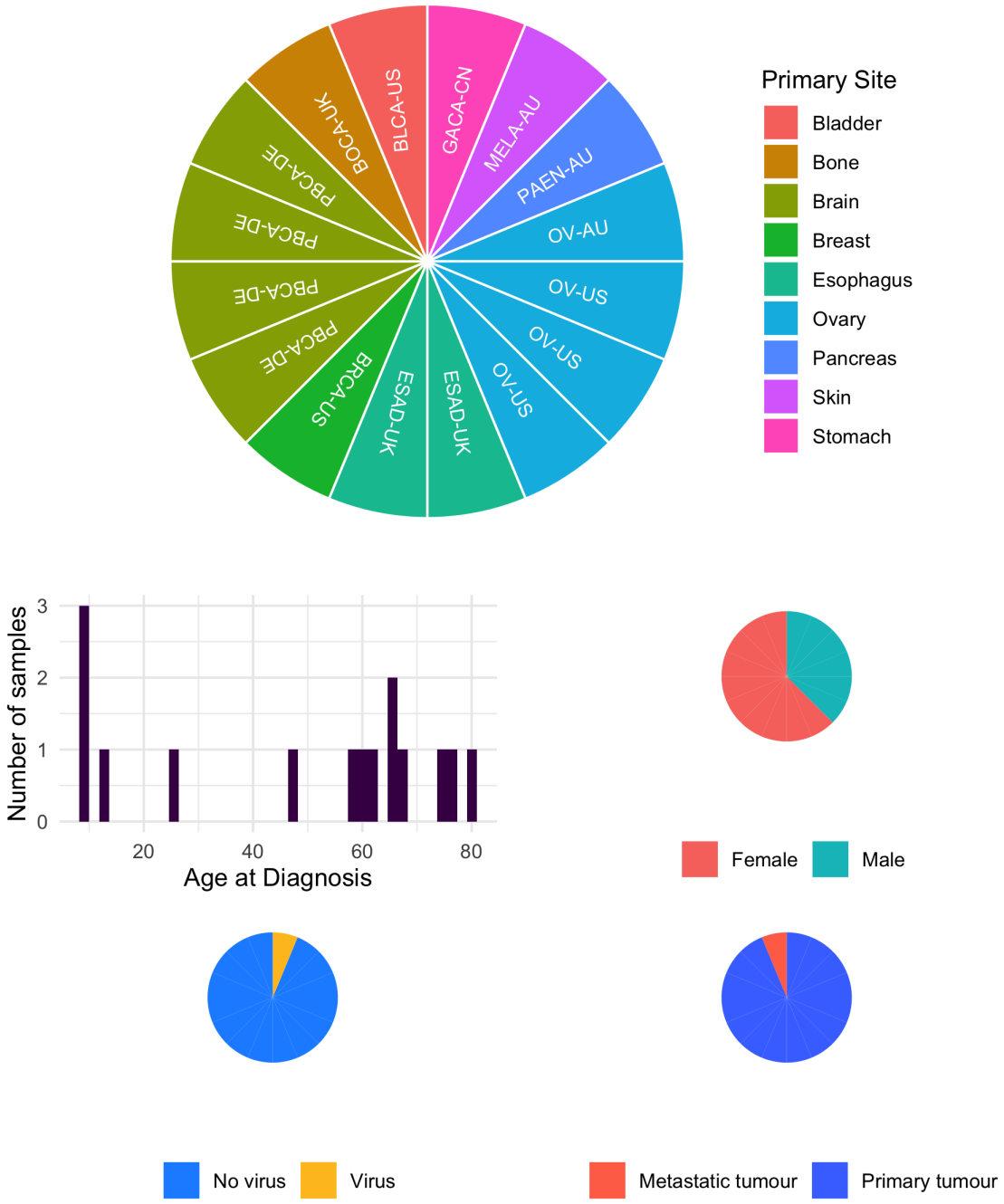


Figure 4.7: Summary of samples identified as potential hybrids in the PCAWG data set. The summary pie chart is coloured by the site of the primary tumour, and annotated with the PCAWG tumour type project code. A list of PCAWG tumour types and their abbreviations is provided in Table D.1 in Appendix D. (n = 16)

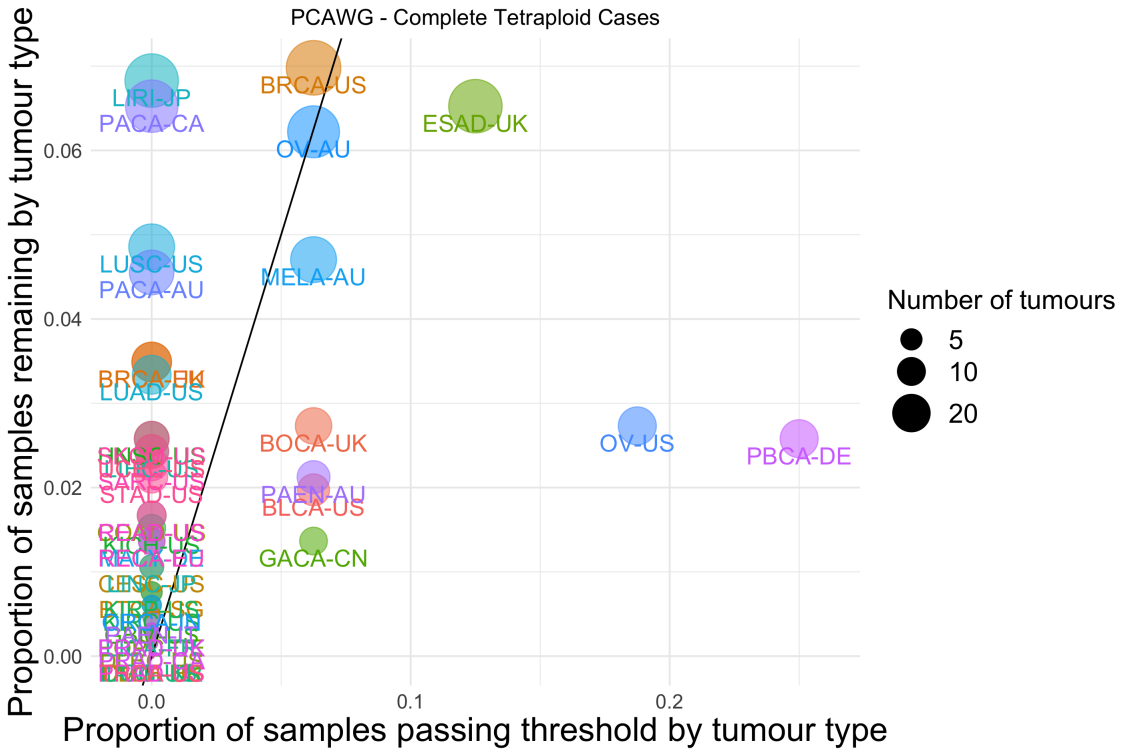


Figure 4.8: Proportion of tetraploid PCAWG samples passing our threshold, by tumour type. The diagonal line is plotted as y = x. Tumour type abbreviations are listed in Table D.1 in Appendix D.

age of diagnosis for paediatric cancers was 10.5 (±1.73 SD), and the average age at diagnosis for non-paediatric cancers was 62.2 (±15.0 SD). Age at diagnosis was not available for the Bone Cancer sample. Finally, ten samples were from female patients, and six samples were from male patients.

When the prevalence of each tumour type beyond the cutoff is plotted against the prevalence of that tumour type in the remaining tetraploid samples, it appears that certain tumour types are over-represented in the subset of samples (Figure 4.8). A similar trend emerges when one considers the prevalence of each tumour type among all remaining samples, including diploid tumours (Figure 4.9).

In both cases, OV-US, PBCA-DE, and ESAD-UK appear to be present at a higher proportion in the samples identified beyond the cutoffs than they are across the PCAWG data set, whereas tumour types such as Liver Cancer (LIRI-JP) and Pancreatic Cancer (PACA-CA and PACA-AU), which represent a large portion of

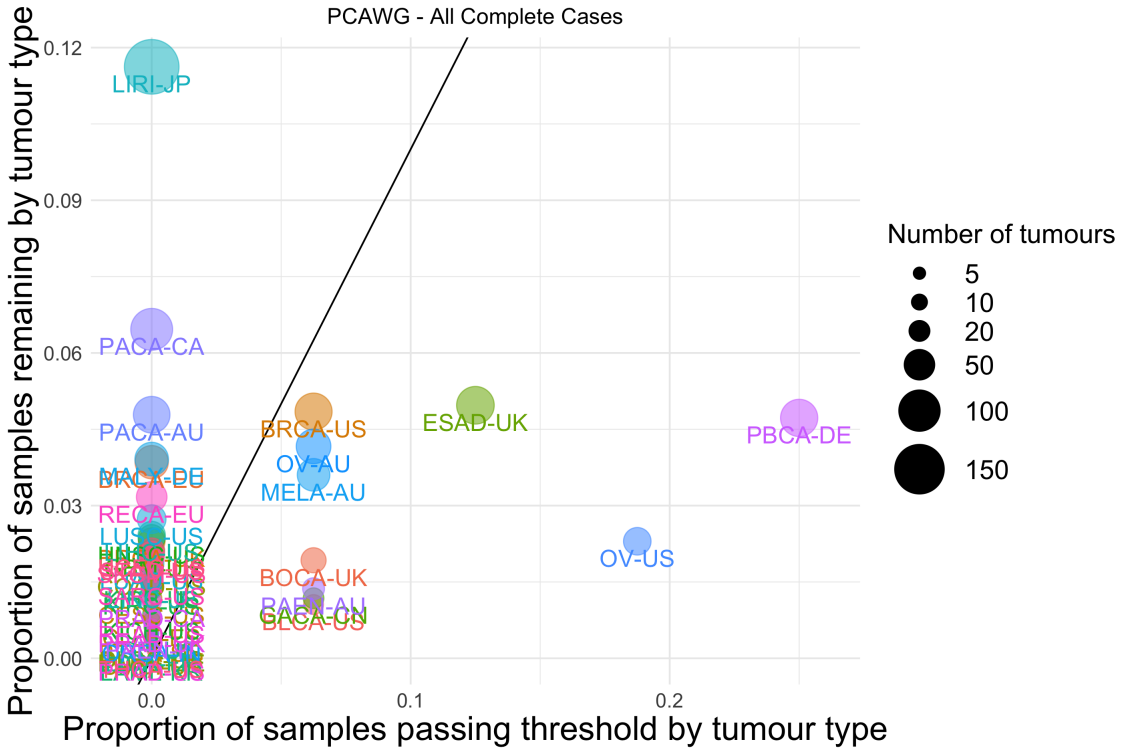


Figure 4.9: Proportion of all PCAWG samples passing threshold, by tumour type. The diagonal line is plotted as y = x. Tumour type abbreviations are listed in Table D.1 in Appendix D.

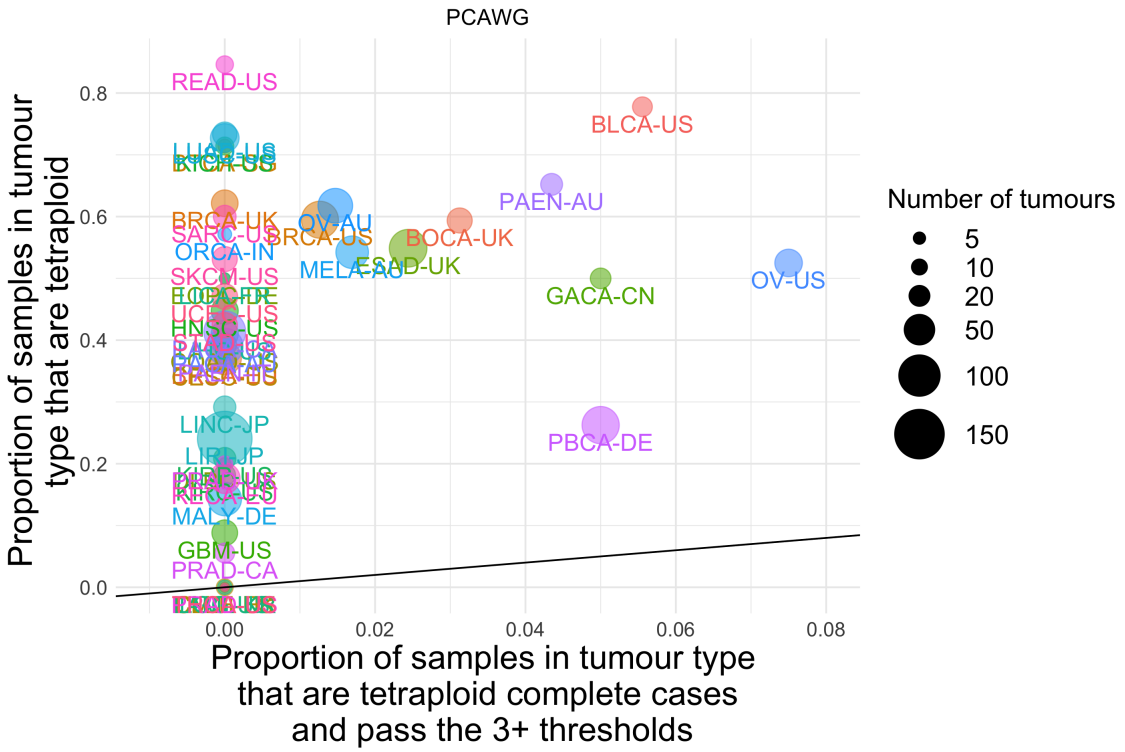


Figure 4.10: Proportion of PCAWG samples that are tetraploid vs proportion of samples that pass the cutoff, by tumour type. The diagonal line is plotted as y = x. Tumour type abbreviations are listed in Table D.1 in Appendix D.

the PCAWG data set in terms of sample numbers, are not identified in my fusion analysis. This suggests that the tumours that have been identified as frequently containing potential fusions have not been identified merely based on their prevalence in the data set.

Similarly, when proportion of samples that are tetraploid within each tumour type is compared to the proportion of samples belonging to each tumour type that pass the threshold, the data suggest that the samples passing the cutoff do not merely reflect the proportion of samples within a tumour type that are tetraploid (Figure 4.10).

4.5.1.3 Timing of Fusions or Whole Genome Duplications

The timing of each whole genome duplication event was estimated based on the fraction of mutations present on one chromosome copy or two, across segments

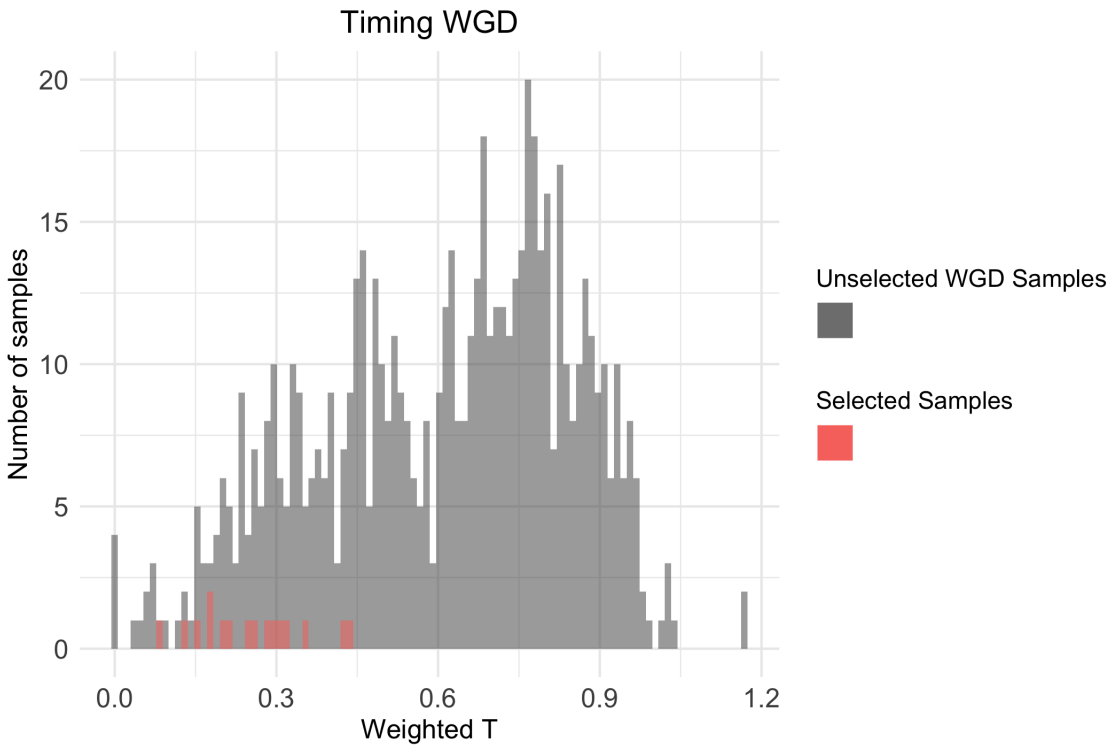


Figure 4.11: Estimated timing of whole genome duplication in whole genome duplicated samples in the PCAWG data set. Samples identified as potential cell fusions ('Selected' samples) are plotted in red and remaining WGD samples for which timing and fusion measures could be calculated are plotted in grey. (n = 675)

of different copy numbers, as described in Section 4.3.6. I found that all of the samples that were selected for further analysis clustered towards the lower end of the spectrum, suggesting either an early whole genome duplication event or a fusion (Figure 4.11).

Investigation of WGD timing within the individual tumour types revealed a similar pattern, with selected samples appearing as early events relative to the timing of WGD in other cancers of the same type (Figure 4.12).

4.5.1.4 Premature Chromosome Condensation

When samples identified as potentially originating from a cell fusion event ("Selected" samples) were compared to the remaining WGD samples, I did not find a significant difference in the number of chromothripsis events, chromosomes affected

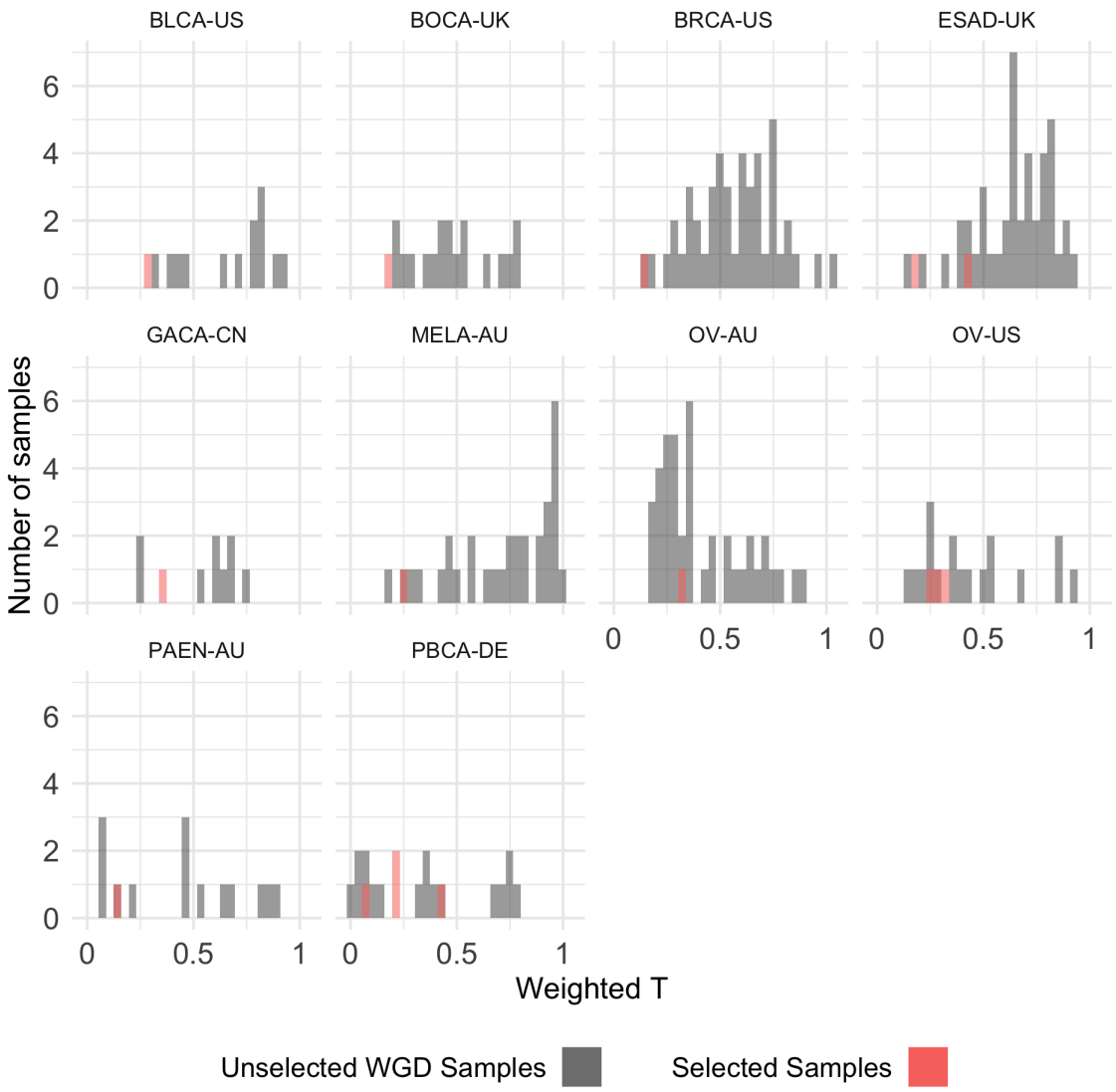


Figure 4.12: Estimated timing of whole genome duplication in whole genome duplicated samples in the PCAWG data set, split by tumour type. For clarity, only tumour types in which potential fusions were identified are presented. Samples identified as potential cell fusions ('Selected' samples) are plotted in red and the remaining WGD samples for which fusion measures ('Unselected' samples) could be calculated are plotted in grey. (n Selected = 1-4 n Unselected = 9-49)

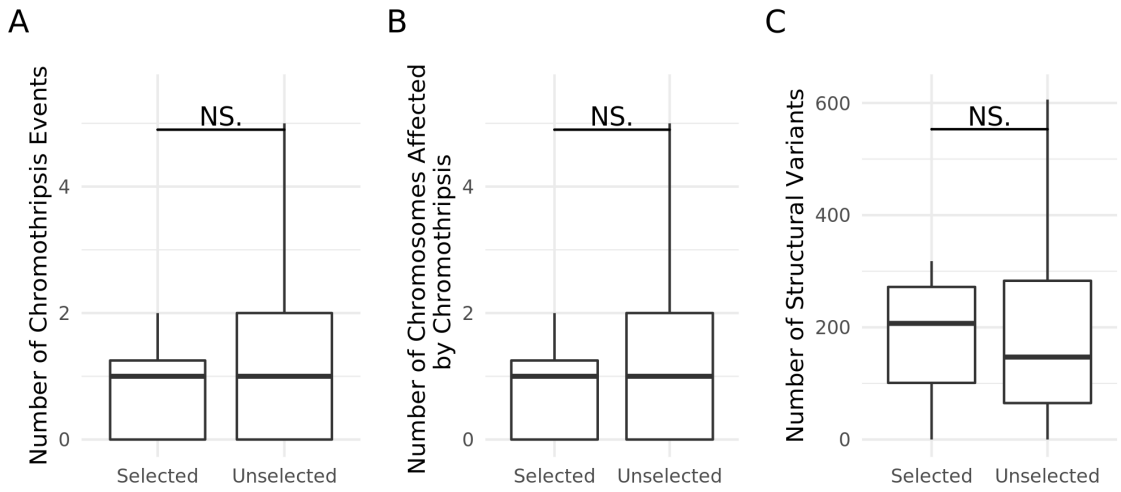


Figure 4.13: Number of chromothripsis events and structural variants in samples identified as potential cell fusions compared to other WGD samples, as well as the number of chromosomes affected by chromothripsis in each sample. (Wilcoxon Rank Sum test; * = p <0.05, ** = p <0.01, *** = p <0.001, n = 662)

by chromothripsis, or the number of structural variants (Wilcoxon rank sum tests p >0.05, Figure 4.13).

The frequency of chromothripsis events has been found to be very heterogeneous within the PCAWG data set (Cortés-Ciriano et al., 2020), so I also investigated the presence of chromothripsis and structural variants within individual tumour types (Figures 4.14-4.16). I only consider WGD samples for this analysis, as chromothripsis is more frequent in WGD samples, and WGD is one of the criteria that was used to identify potential fusions (Cortés-Ciriano et al., 2020).

The melanoma (MELA-AU) sample stands out from the remaining samples belonging to each tumour type due to the high number of chromothripsis events observed. The melanoma sample contained 23 events affecting 16 chromosomes (Figures 4.14 and 4.15). The number of structural variants, chromothripsis events, and chromosomes affected by chromothripsis appeared much higher in this selected sample than in the WGD melanoma samples that did not pass the thresholds.

Similar results were obtained from data published by Cortés-Ciriano et al. (2020), which suggested that 18 chromosomes in this sample were affected by chromothripsis

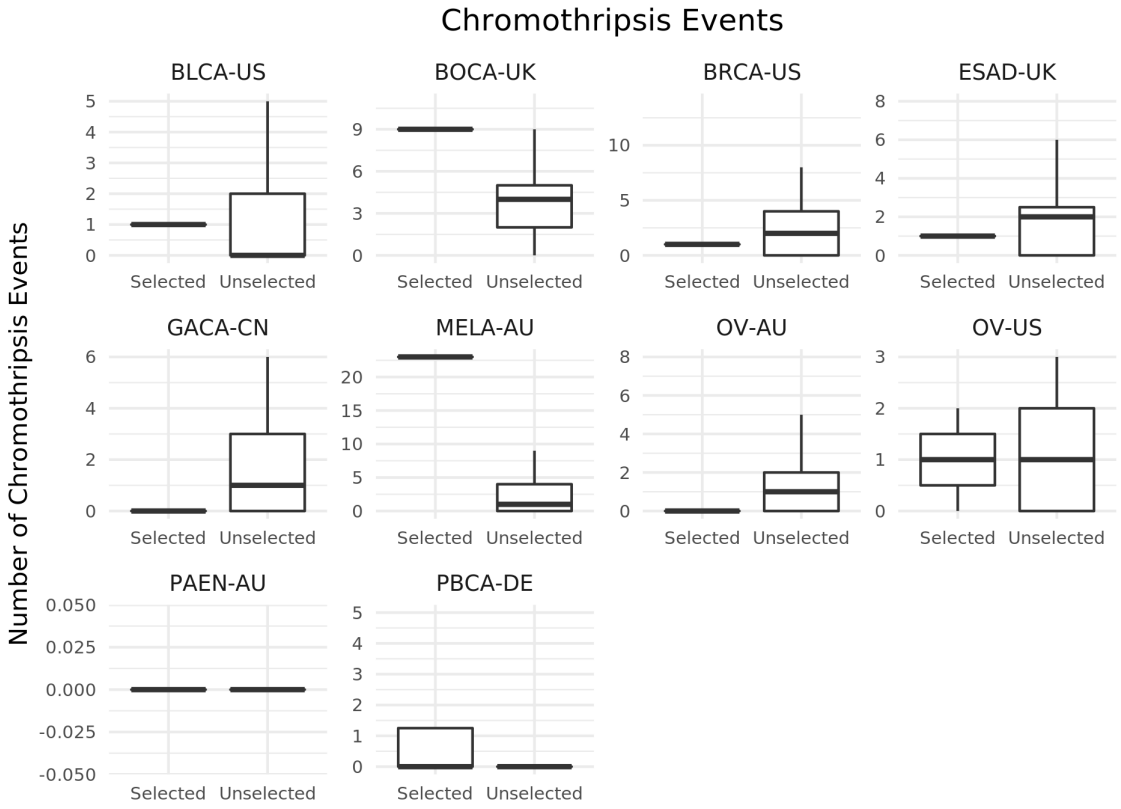


Figure 4.14: Number of chromothripsis events by tumour type. For clarity, only tumour types in which potential fusions were identified are displayed. Only whole genome duplicated tumours are considered. n = 1-4 for Selected samples and n = 9-46 for Unselected samples. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

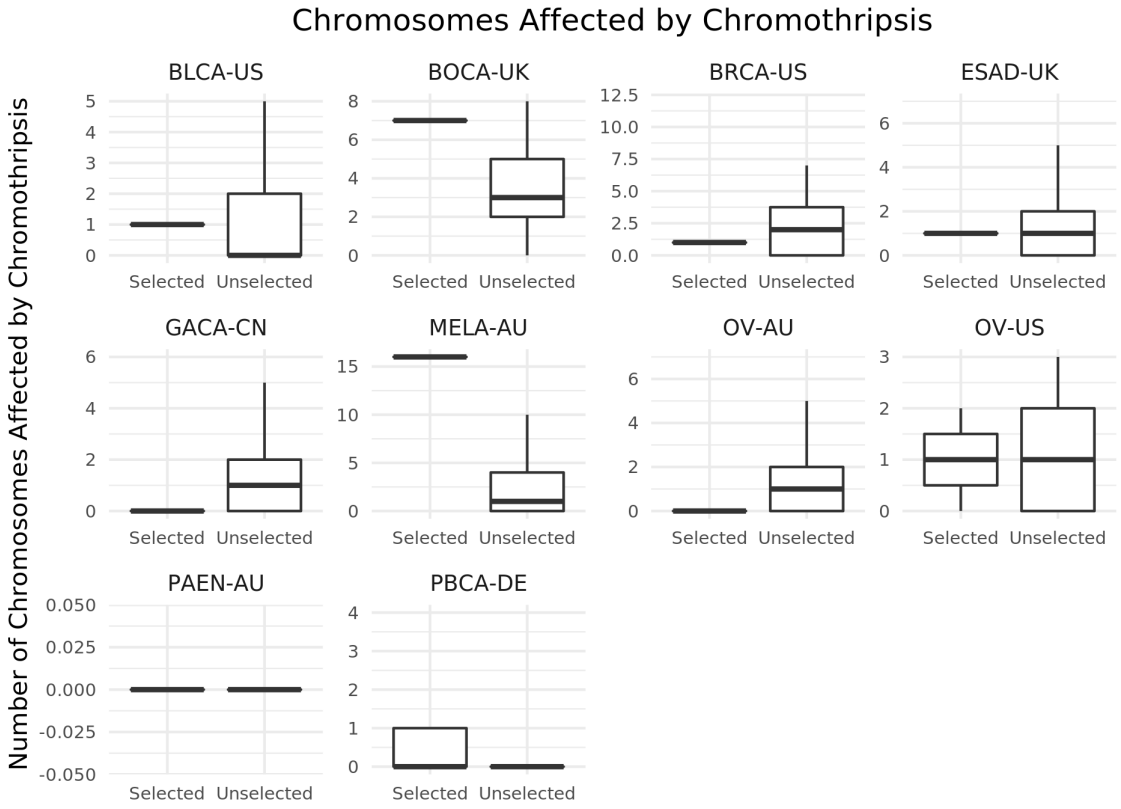


Figure 4.15: Number of chromosomes affected by chromothripsis by tumour type. For clarity, only tumour types in which potential fusions were identified are displayed. Only whole genome duplicated tumours are considered. n = 1-4 for Selected samples and n = 9-46 for Unselected samples. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

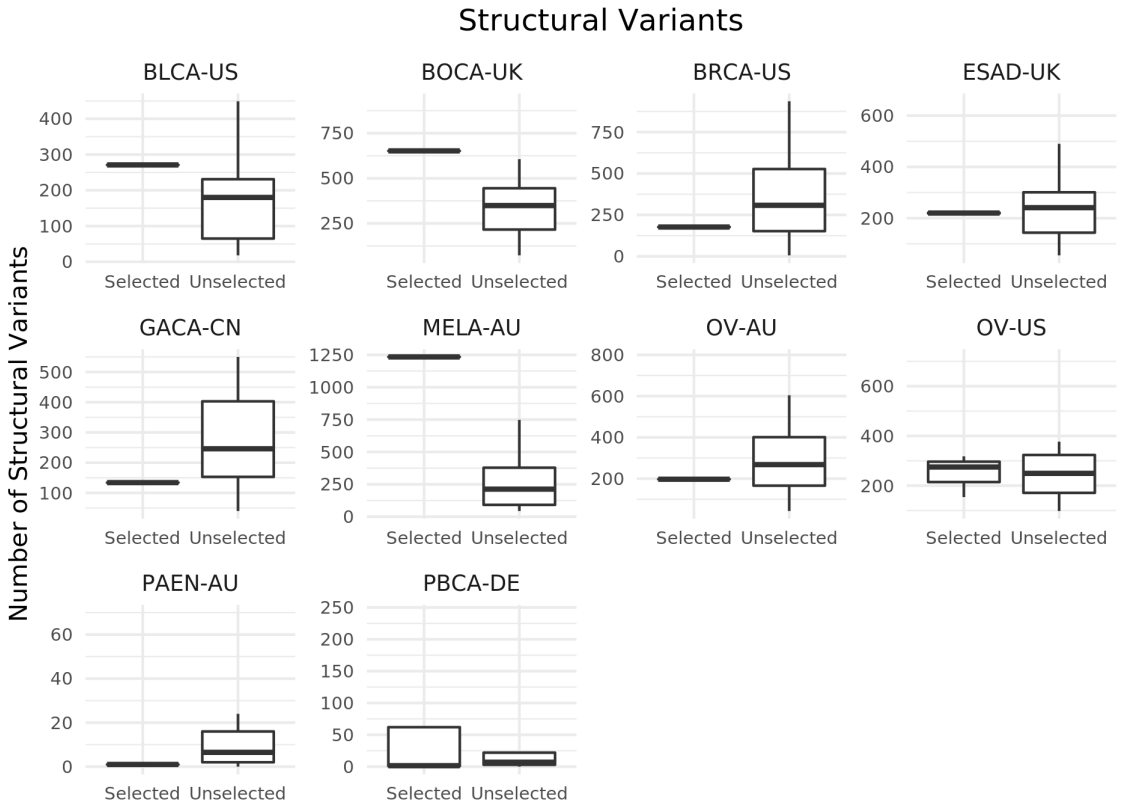


Figure 4.16: Number of structural variant events by tumour type. For clarity, only tumour types in which potential fusions were identified are displayed. Only whole genome duplicated tumours are considered. n = 1-4 for Selected samples and n = 9-46 for Unselected samples. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

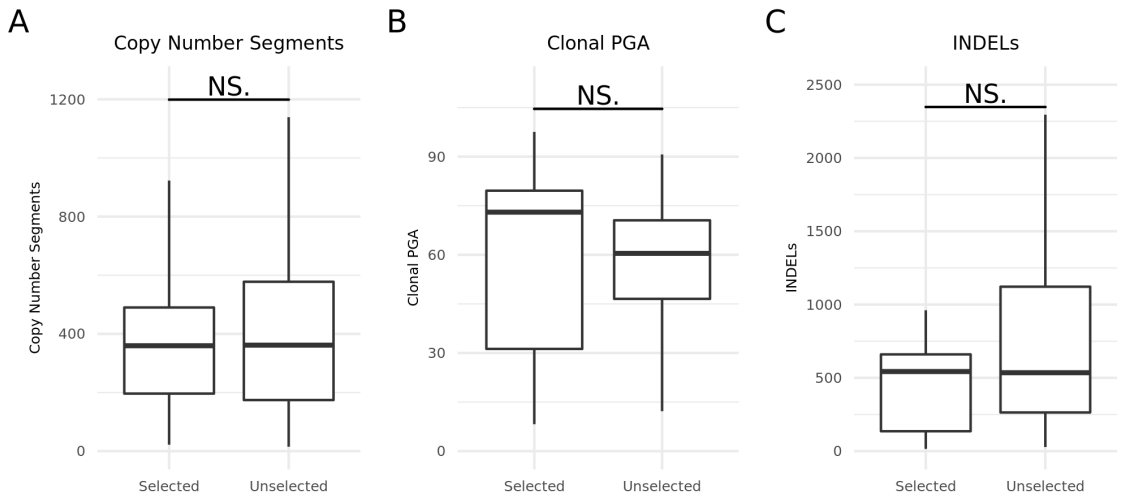


Figure 4.17: Number of copy number segments, INDELs, and PGA in samples identified as potential cell fusions compared to other WGD samples, for which fusion measure could be calculated. Wilcoxon rank sum test ; * = p <0.05, ** = p <0.01, *** = p <0.001, n = 675

(data not shown).

4.5.1.5 Genome Stability

Chromothripsis has previously been associated with an increase in CNVs (Zack et al., 2013), which prompted me to further investigate the genome stability of the selected samples, using some of the measures introduced in Chapter 2.

Overall, I did not find a statistically significant difference in the number of INDELs, number of CN segments, or PGA between samples identified as potential fusions compared to their WGD counterparts that did not exceed the thresholds (Figure 4.17; Wilcoxon rank sum test p >0.05). Comparisons within tumour types were also not significant for any of the three measures, although higher numbers of CN segments were observed in selected bladder cancers, bone cancers, ovarian cancers, and melanoma (Figure 4.18). For PGA, higher levels were observed in selected breast cancers, gastric cancers, melanoma, ovarian cancers, and pancreatic endocrine neoplasms (Figure 4.19). Only one tumour type, breast cancer, presented higher numbers of INDELs in selected samples compared to unselected (Figure 4.20).

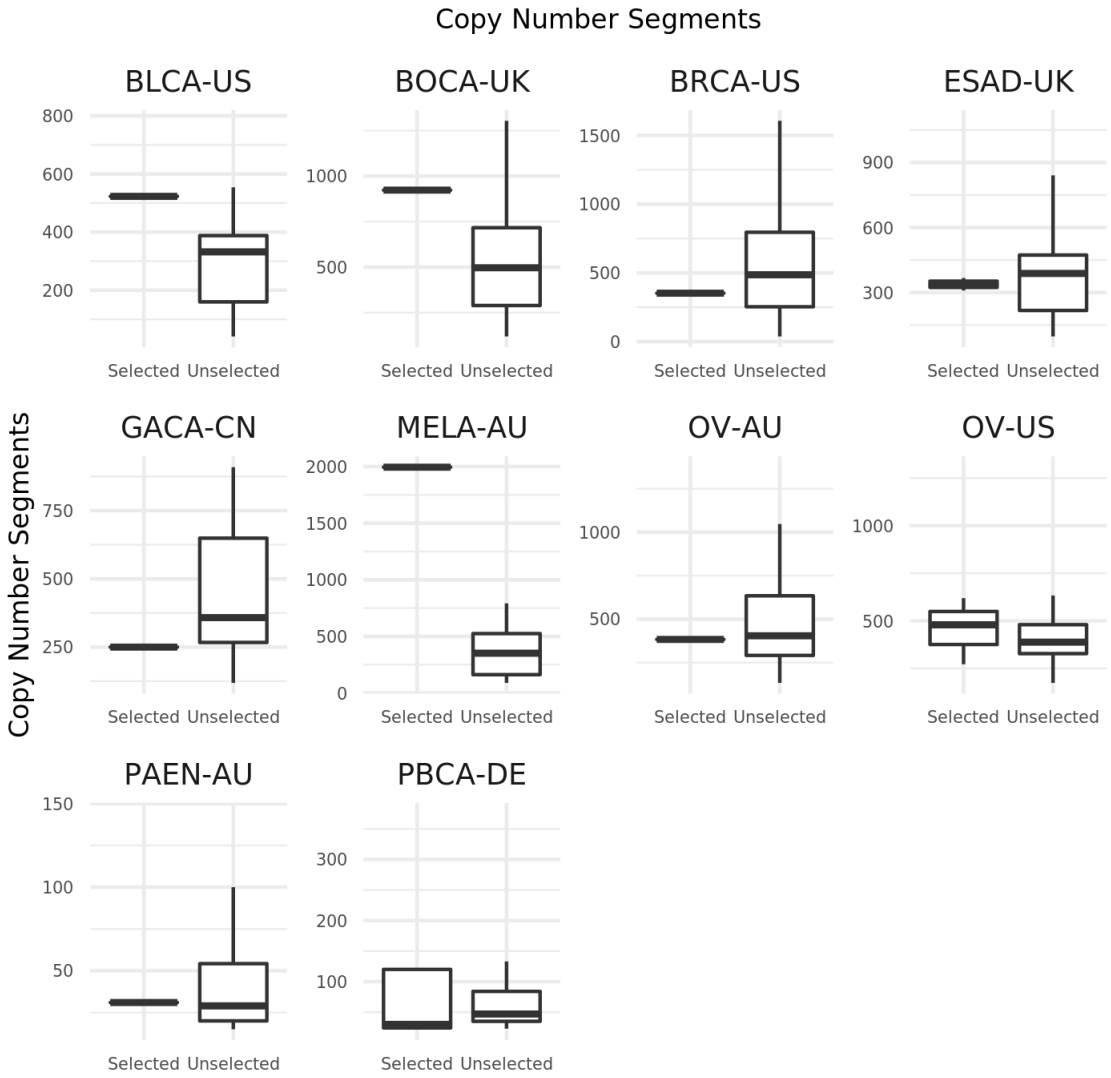


Figure 4.18: Number of copy number segments identified in each tumour stratified by tumour type. For clarity, only tumour types in which potential cell fusions were identified are displayed. Only WGD samples were considered. n = 1-4 for Selected samples and n = 9-46 for Unselected samples. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

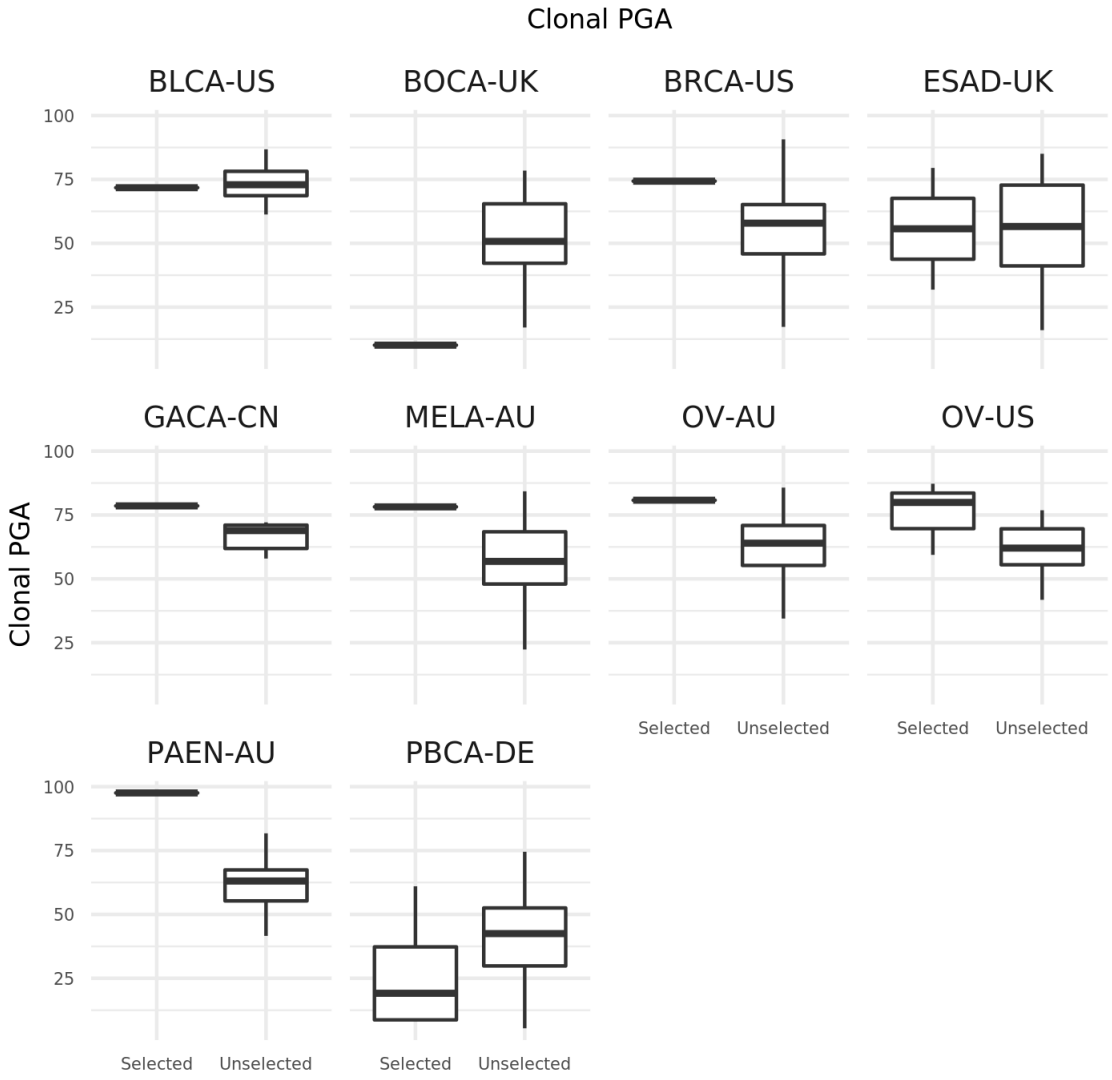


Figure 4.19: Percentage of the genome altered in samples identified as potential cell fusions compared to their WGD counterparts. For simplicity, only tumour types in which potential cell fusions were identified are represented. n = 1-4 for Selected samples and n = 9-46 for Unselected samples. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

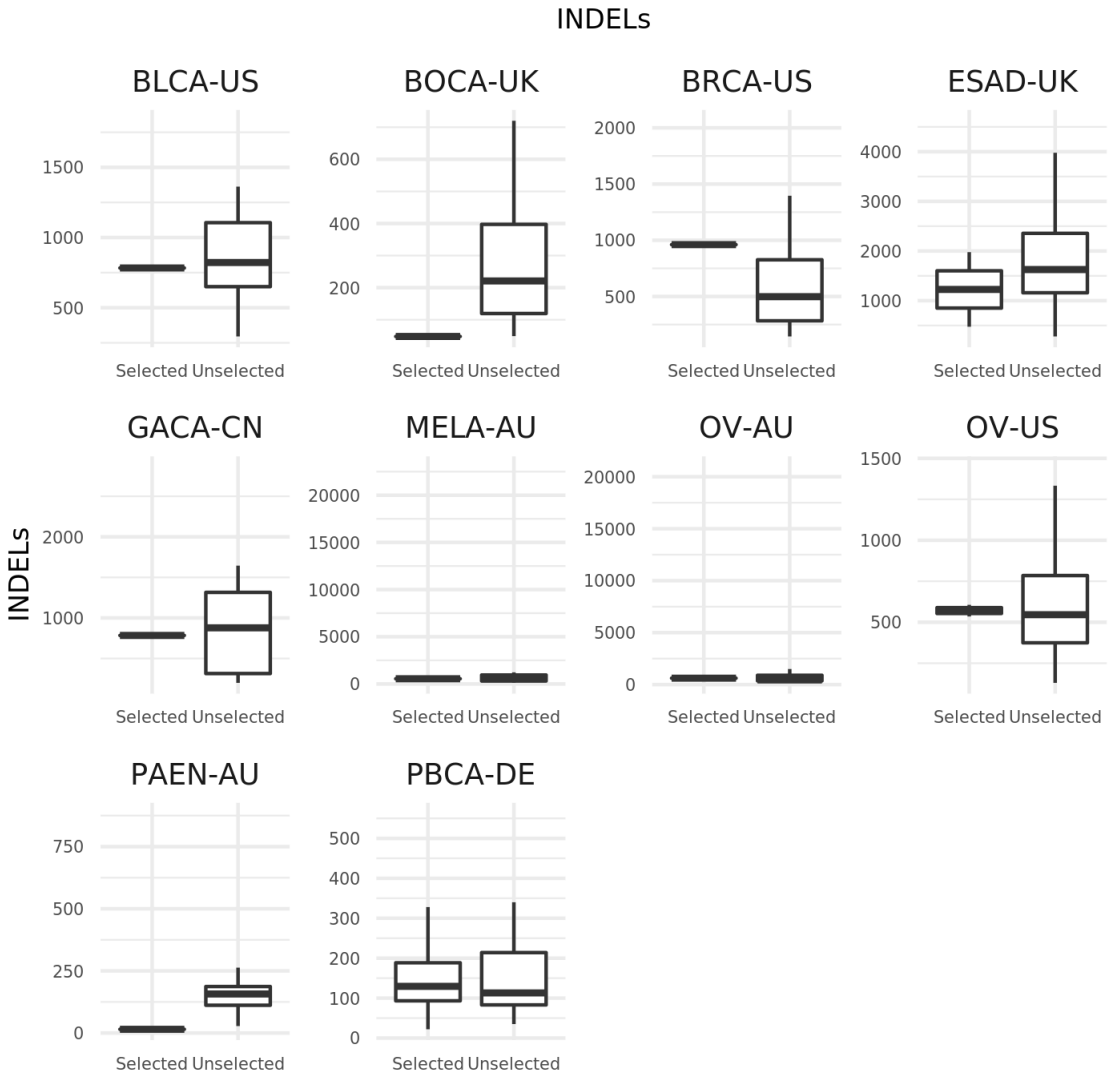


Figure 4.20: Number of INDELs observed in samples identified as potential cell fusions compared to their WGD counterparts. Only tumour types in which potential cell fusions were identified are represented. n = 1-4 for Selected samples and n = 9-46 for Unselected samples. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

4.5.1.6 Mutational Signatures

There was no significant difference in the number of SBS2 and SBS13 mutations observed in selected and unselected samples of the same tumour type (data not shown). Overall, I did not observe any consistent differences in the composition of mutational signatures present in selected samples, compared to tumours from the same tissue (Figures C.1-C.10 in Appendix C.2.1). However, several of the individual samples had unusual signatures present. The breast cancer sample that was identified (belonging to donor DO6231) stands out from the other breast cancer samples in this sub-set as the sample with the highest mutation burden, and due to the large number of SBS9 mutations present (Figure C.3 in Appendix C.2.1). SBS9 is attributed to polymerase η activity following AID induced cytidine deamination; it is mainly observed in malignant B-cell lymphomas and chronic lymphocytic leukemias, and is not commonly observed in breast cancers (Alexandrov et al., 2020, 2013). SBS9 mutations were also observed in two breast cancer samples that were not identified as potential fusions, though to a much smaller degree. I also found that roughly one fourth of the mutations in this breast cancer sample were attributed to SBS17a and SBS17b. SBS17a and SBS17b are frequently seen in esophageal cancer, stomach cancer, liver cancer, and B-cell lymphomas, and only infrequently observed in breast cancer (Alexandrov et al., 2020, 2013). I think that the presence of uncharacteristic mutational signatures in this sample may be an indicator of the involvement of a different cell type in this tumour, for example through heterotypic cell fusion.

The Australian ovarian cancer sample that was identified, belonging to donor DO46382, was also quite distinct from other ovarian cancers within that project (Figure C.7 in Appendix C.2.1). The mutations identified in this ovarian cancer were attributed to SBS1, SBS5, and SBS40, all of which have previously been identified in normal tissues or organoids derived from adult stem cells (Blokzijl et al., 2016; Lee-Six et al., 2019; Moore et al., 2020). No other mutational signatures were

identified in this sample. Further, SBS3, which is attributed to defective homologous recombination or mutations in BRCA1/2, was absent in this sample but present in nearly all of the other ovarian cancers surveyed, including those from the United States (Figures C.7 and C.8 in Appendix C.2.1).

The pancreatic cancer endocrine neoplasm (PAEN) DO51181 also stood out from other tumours of the same type (Figure C.9 in Appendix C.2.1). This sample had the lowest mutation burden of the PAEN samples that were investigated, and only contained mutations attributed to SBS1 and SBS5.

Given that the complement of mutational signatures identified in segments derived from normal cells is expected to differ from signatures identified in segments derived from tumour cells, assessing mutational signatures across the tumour in its entirety may not be the best approach to identify these differences. Thus, I investigated mutational signatures identified in the CN 1+1 and 2+1 regions that were identified in the analysis of the CV_{MD}, to see if I could identify variation in mutational signatures coinciding with variation in mutation density (Figures C.11-C.26 in Appendix C.2.2).

Contrary to my expectations, I observed almost no variation in mutational signatures identified in segments with high mutation density compared to low mutation density, with the majority of segments carrying mutations from all of the signatures identified in the tumour, at similar proportions. The potential exception to this observation is the breast cancer sample DO6231, in which the proportion of mutations attributed to SBS5 appeared to be higher in segments with low mutation density, and lower in segments with high mutation density, hinting that these segments may have been disproportionately affected by different mutational processes (Figure C.13 in Appendix C.2.2).

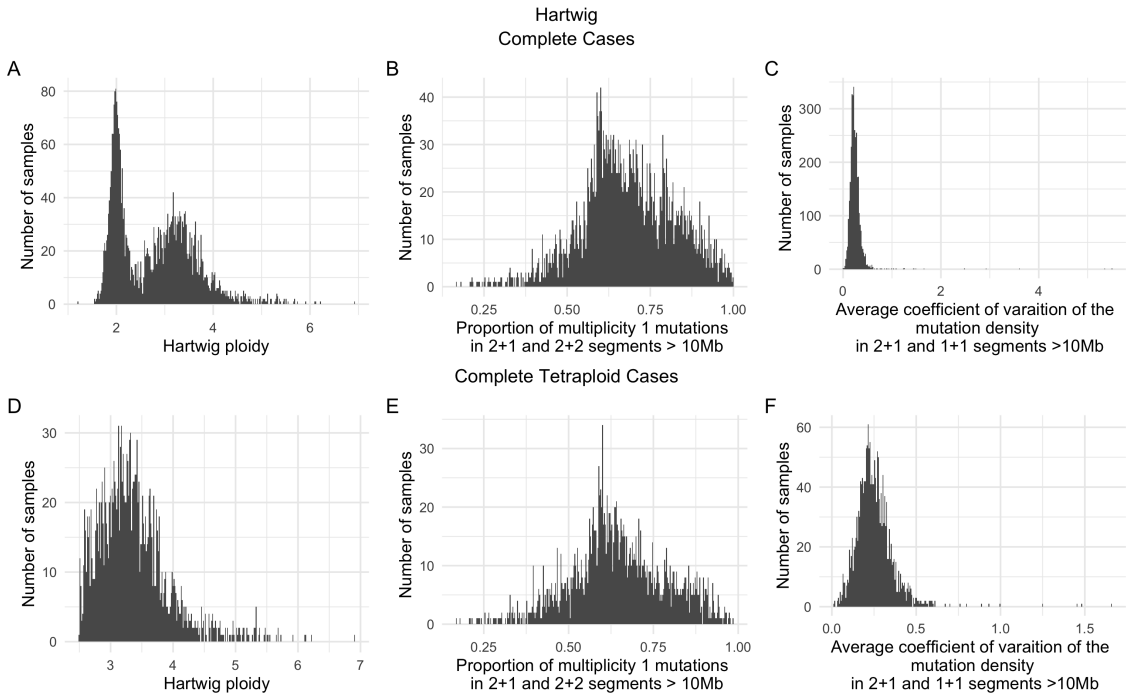


Figure 4.21: Overview of fusion measures calculated for tumours in the HMF data set. Panels A-C depict the distribution of the fusion measures calculated for all samples in the HMF data set for which I was able to calculate these measures. Panels D-F depict the fusion measures calculated for tetraploid tumours. (A-C n = 3,308, D-F n = 1,872

4.5.2 Identifying Cell-Cell Fusions in the Hartwig Medical Foundation Data Set of Metastatic Samples

Following my analysis of the PCAWG data set, I carried out a similar investigation on samples provided by the Hartwig Medical Foundation (HMF). The HMF data set consists of 3,679 whole genome sequenced metastatic tumours. This data set is of particular interest due to the connection between cell fusion and metastasis that has been proposed.

4.5.2.1 Identifying Samples Potentially Containing Fusions

Similar to the analysis carried out on the PCAWG data set, I began by surveying the PM_1 and CV_{MD} (Figure 4.21), and applied a series of thresholds to identify samples that might be outliers (Figure 4.22).

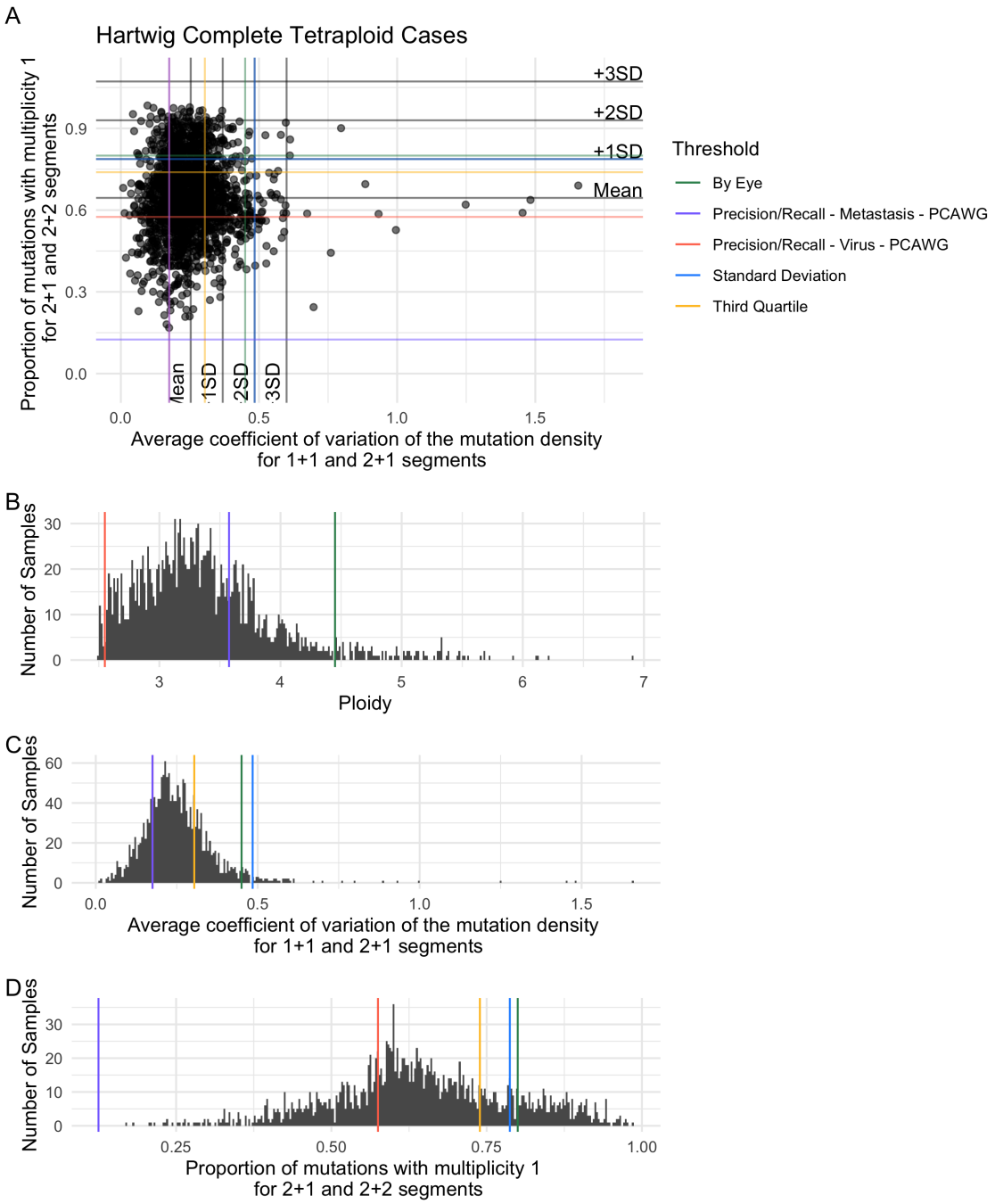


Figure 4.22: Thresholds applied to the HMF fusion measures for tetraploid samples. (n = 1,872)

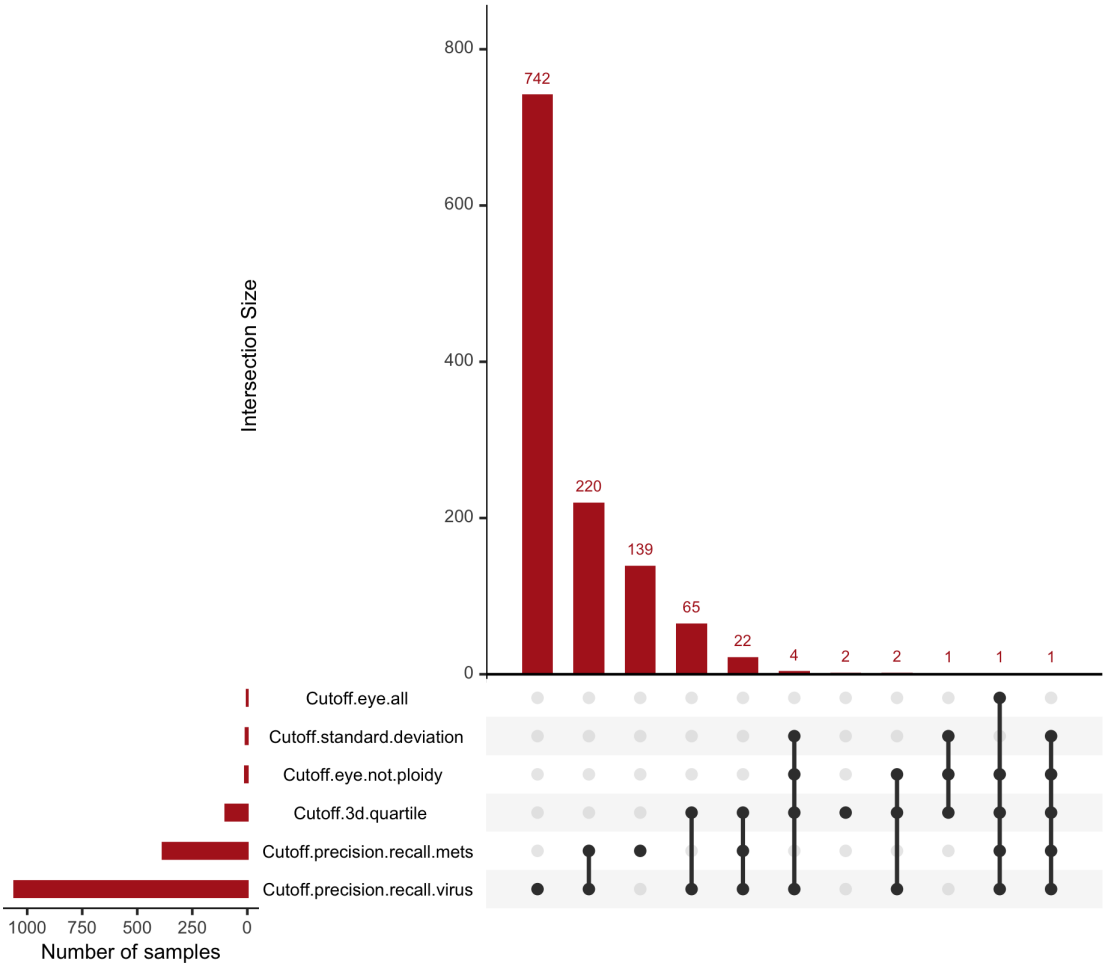


Figure 4.23: Overlap between sample sets identified by applying different thresholds to the HMF fusion measures.

4.5.2.2 Summary of Samples Potentially Containing Fusions

I focused my analysis on samples that passed three or more thresholds, resulting in a set of 30 samples identified for further investigation (Figure 4.23).

In contrast to the tumour types identified in the PCAWG data set cutoffs, the most abundant primary sites of the tumours identified in the HMF data set were breast (13/30, 43.3%), lung (3/30, 10.0%), ovaries (3/30, 9.68%), prostate (3/30, 10.0%), and the urothelial tract (3/30, 10.0%, Figure 4.24). Two of the breast cancer tumours identified were derived from the same individual. Primaries originating in the colorectum, lymphoid cancers, testicular cancer, and cancers of the bone and soft tissue appeared only once each (1/30, 3.33%). Interestingly, one of the tumours

that I identified was marked as a cancer of unknown primary (CUP). The mean age at biopsy was 61.1 years (±10.2 years SD).

Investigating the prevalence of each primary tumour location in the cutoff versus in the remaining samples, whether tetraploid or not (Figures 4.25 and 4.26), suggests that breast cancers are indeed identified more frequently in the cutoff analysis than would be expected based on their prevalence in the HMF data set. Ovarian cancers and cancers of the urothelial tract also appear to be identified more frequently than would be expected. In contrast, colorectal cancers, skin cancers, and esophageal cancers, which represent a similar number of samples as breast cancer and ovarian cancer, only passed the thresholds infrequently, if at all.

When the proportion of samples within each tumour type that are tetraploid is compared to the proportion of samples within each tumour type that pass three or more thresholds, the proportion of tumours that are tetraploid within each tumour type does not reflect the proportion of tumours that pass the thresholds (Figure 4.27).

4.5.2.3 Timing of Fusions or Whole Genome Duplications

Similar to the samples selected from the PCAWG data set, the timing estimates for the samples selected from the HMF data suggested that these tumours all appeared to have undergone WGD relatively early (Figure 4.28). WGD appeared to have occurred especially early compared to other tumours of the same type in tumours from the bone and soft tissue, colorectum, lung, and prostate (Figure 4.29).

4.5.2.4 Premature Chromosome Condensation

Chromothripsis data was not available for the HMF data set, so I investigated the number of structural variants instead. There was a significant increase in the number of SVs in the samples that were selected for further study, compared to other whole genome duplicated samples when compared across the entire HMF data set (Figure

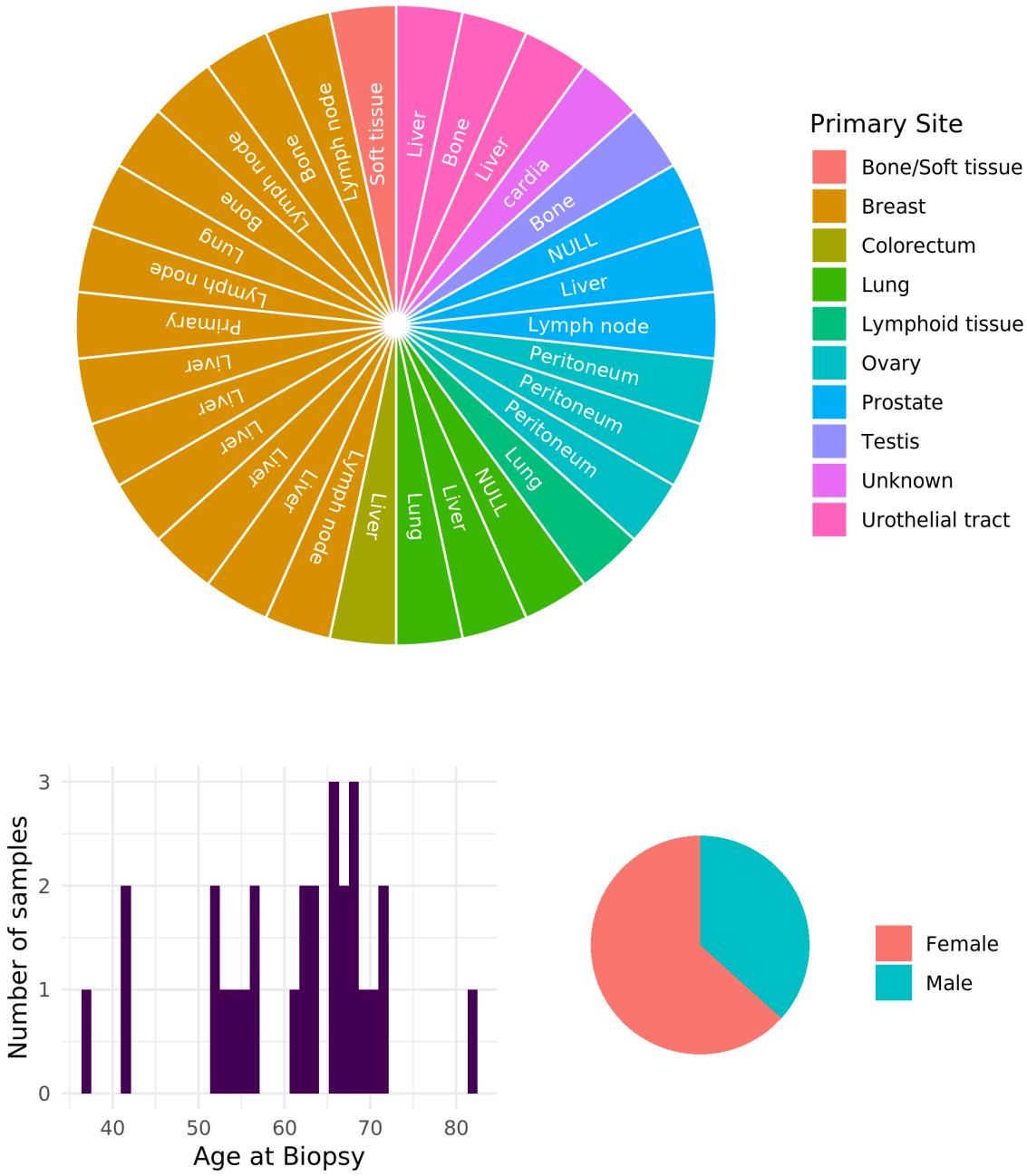


Figure 4.24: Summary of samples identified as potential hybrids in the HMF data set. The pie chart representing the tumour types of the identified samples is coloured by the site of the primary tumour, and annotated by the biopsy site at which the metastatic tumour was sampled. (n = 30)

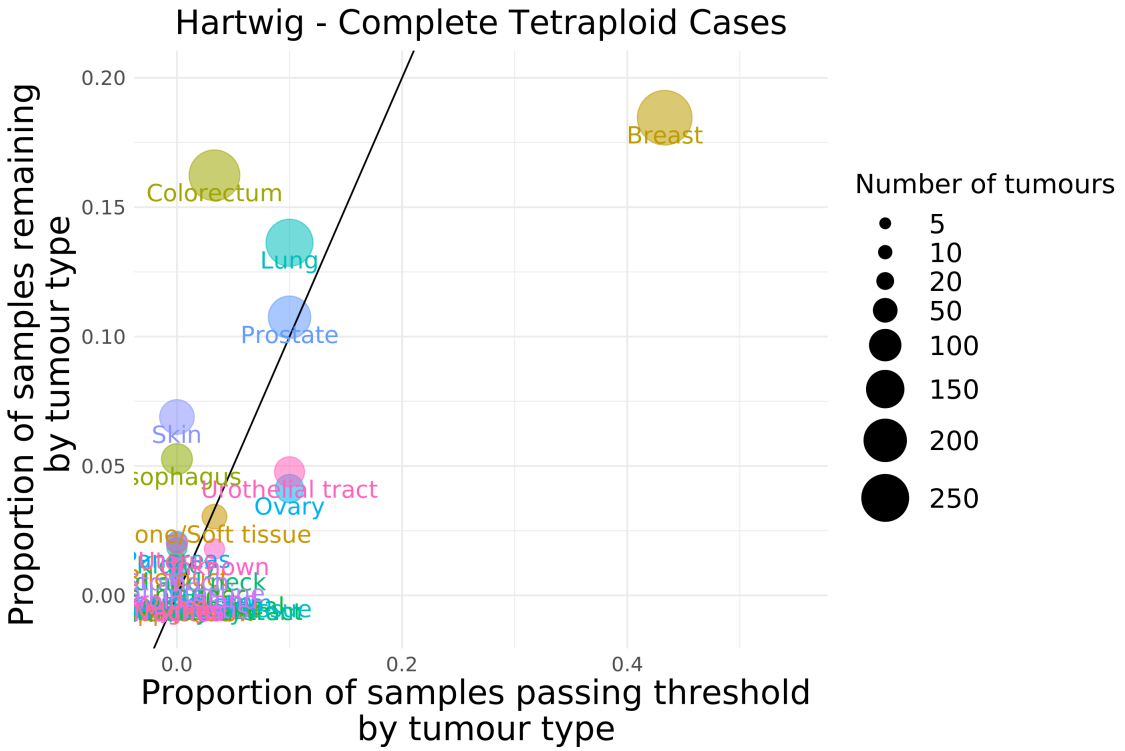


Figure 4.25: Proportion of tetraploid HMF samples passing threshold, by tumour type. The diagonal line is plotted as y = x.

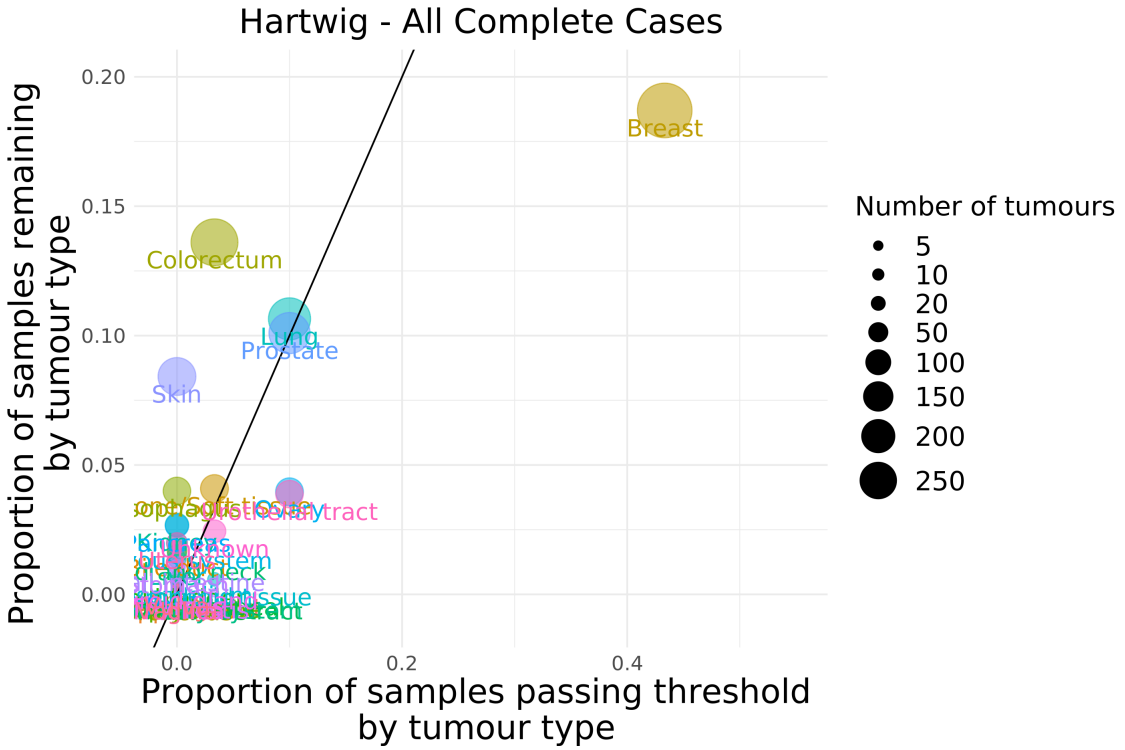


Figure 4.26: Proportion of all HMF samples passing threshold, by tumour type. The diagonal line is plotted as y = x.

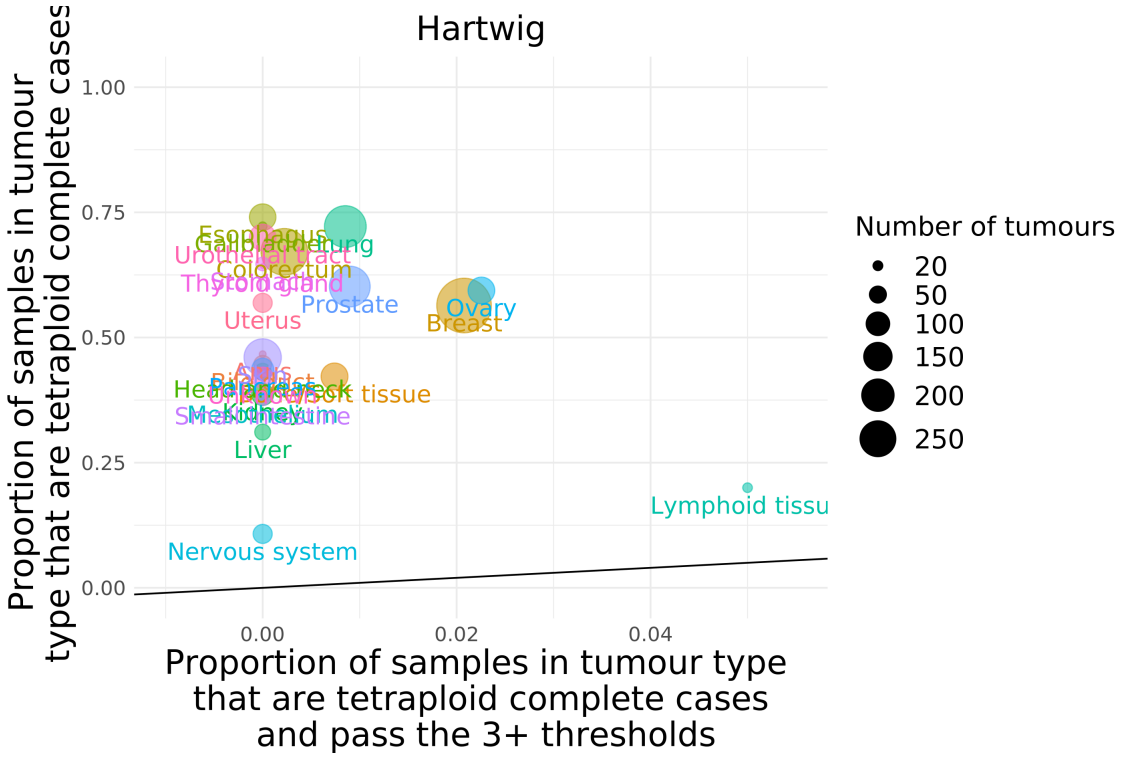


Figure 4.27: Proportion of HMF samples that are tetraploid vs proportion of samples that pass the cutoff, by tumour type. For clarity, only tumour types that contain more than 10 samples have been represented. The diagonal line is plotted as y = x.

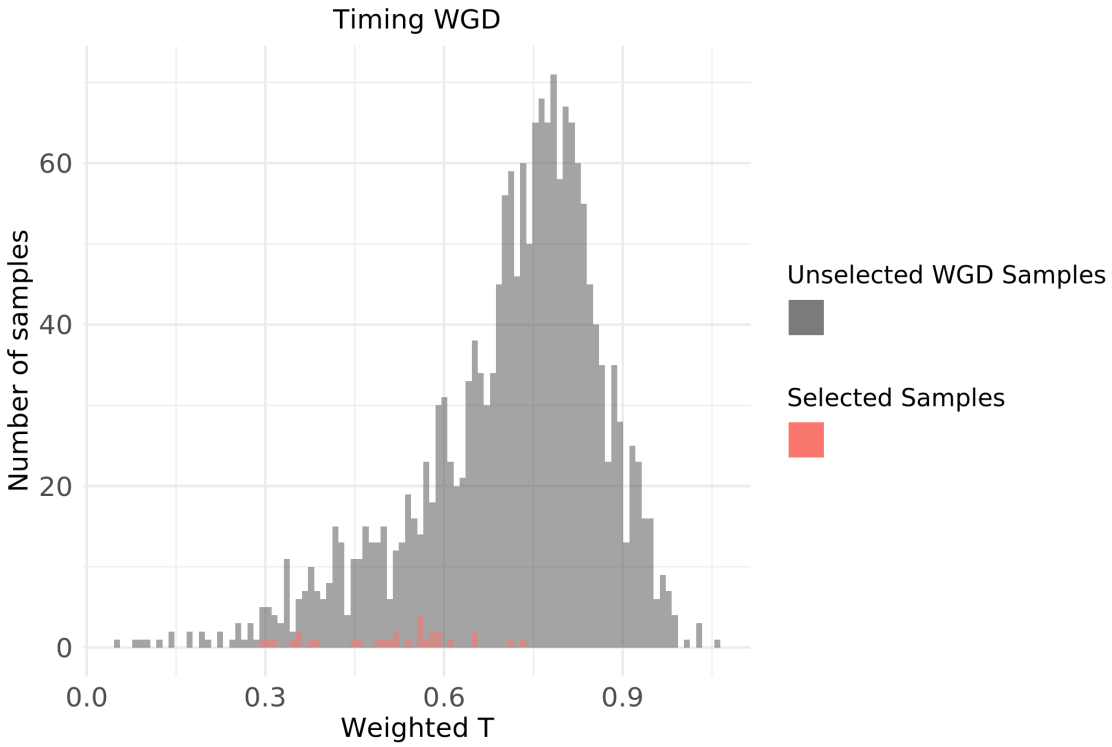


Figure 4.28: Estimated timing of whole genome duplication in whole genome duplicated samples in the HMF data set. Samples identified as potential cell fusions ('Selected' samples) are plotted in red and remaining WGD samples for which timing and fusion measures could be calculated are plotted in grey. (n = 1,872)

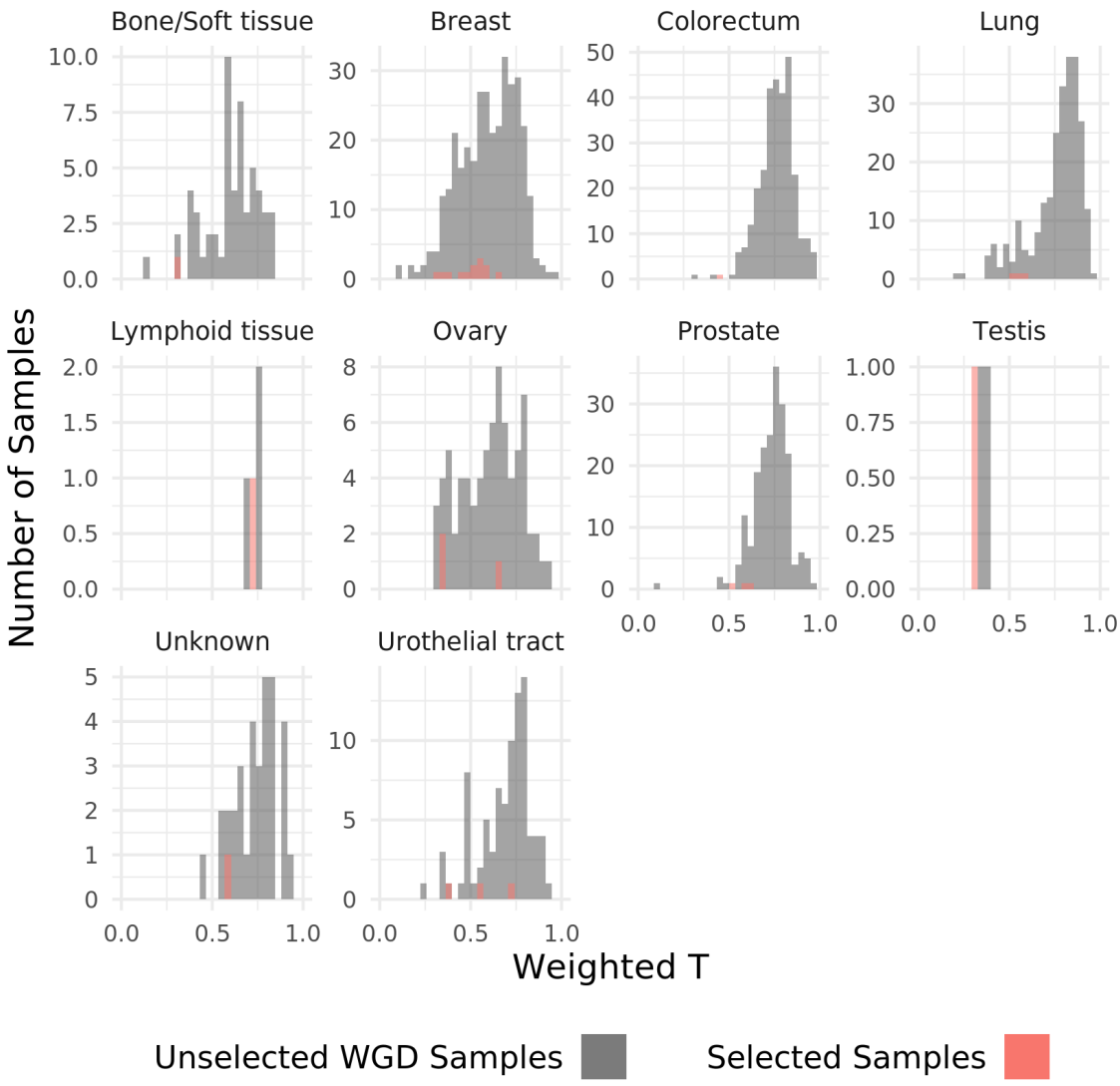


Figure 4.29: Estimated timing of whole genome duplication in whole genome duplicated samples in the HMF data set, split by tumour type. For clarity, only tumour types in which potential fusions were identified are presented. Samples identified as potential cell fusions ('Selected' samples) are plotted in red and remaining WGD samples for which fusion measures could be calculated are plotted in grey. (n Selected = 1-13 n Unselected = 3-340)

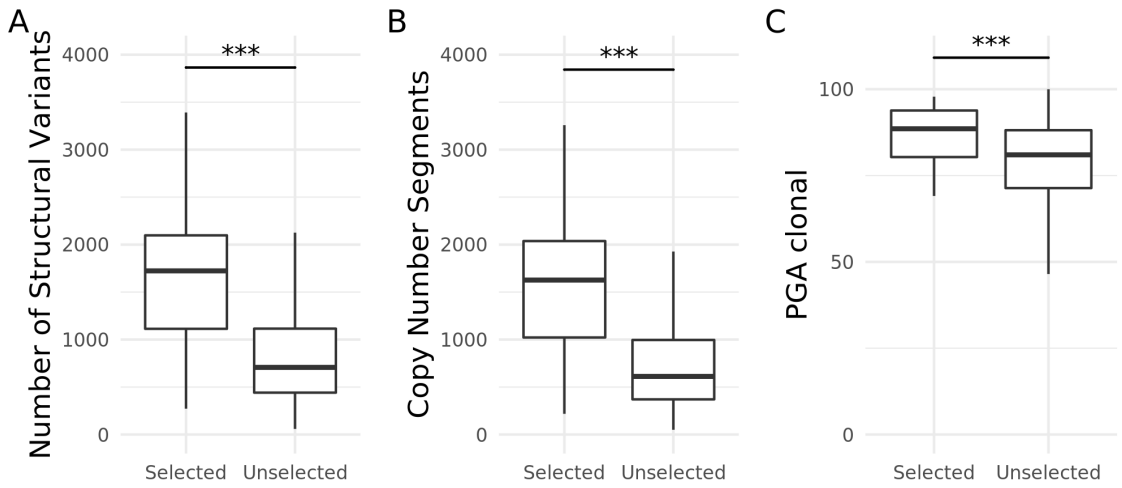


Figure 4.30: A. Number of structural variants in WGD samples in the HMF data set, for which fusion measures could be calculated. B. Number of copy number segments in WGD samples in the HMF data set, for which fusion measures could be calculated. C. Percentage of the genome altered (PGA) from copy number 4 in WGD samples in the HMF data set, for which fusion measures could be calculated. (Wilcoxon Rank Sum test; * = p <0.05, ** = p <0.01, *** = p <0.001, n = 1,872)

4.30 A).

Although other tumour type specific comparisons did not reach significance, there appeared to be a trend towards higher numbers of SVs in samples identified as potential fusions in nine out of the ten tumour types surveyed (Figure 4.31).

4.5.2.5 Genome Stability

When I investigated the stability of the genomes of samples that were identified as potential fusions, I found that there was an overall increase in the number of copy number segments, and percentage of the genome altered clonally (Figure 4.30.B and 4.30.C, Wilcoxon rank sum test p <0.05).

I also investigated these measures within each tumour type and found that, for most tumour types, there was a trend towards increasing levels of all measures (Figures 4.32, and 4.33). Comparisons within individual tumour types did not reach significance after multiple testing correction, likely due to small sample sizes (Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p

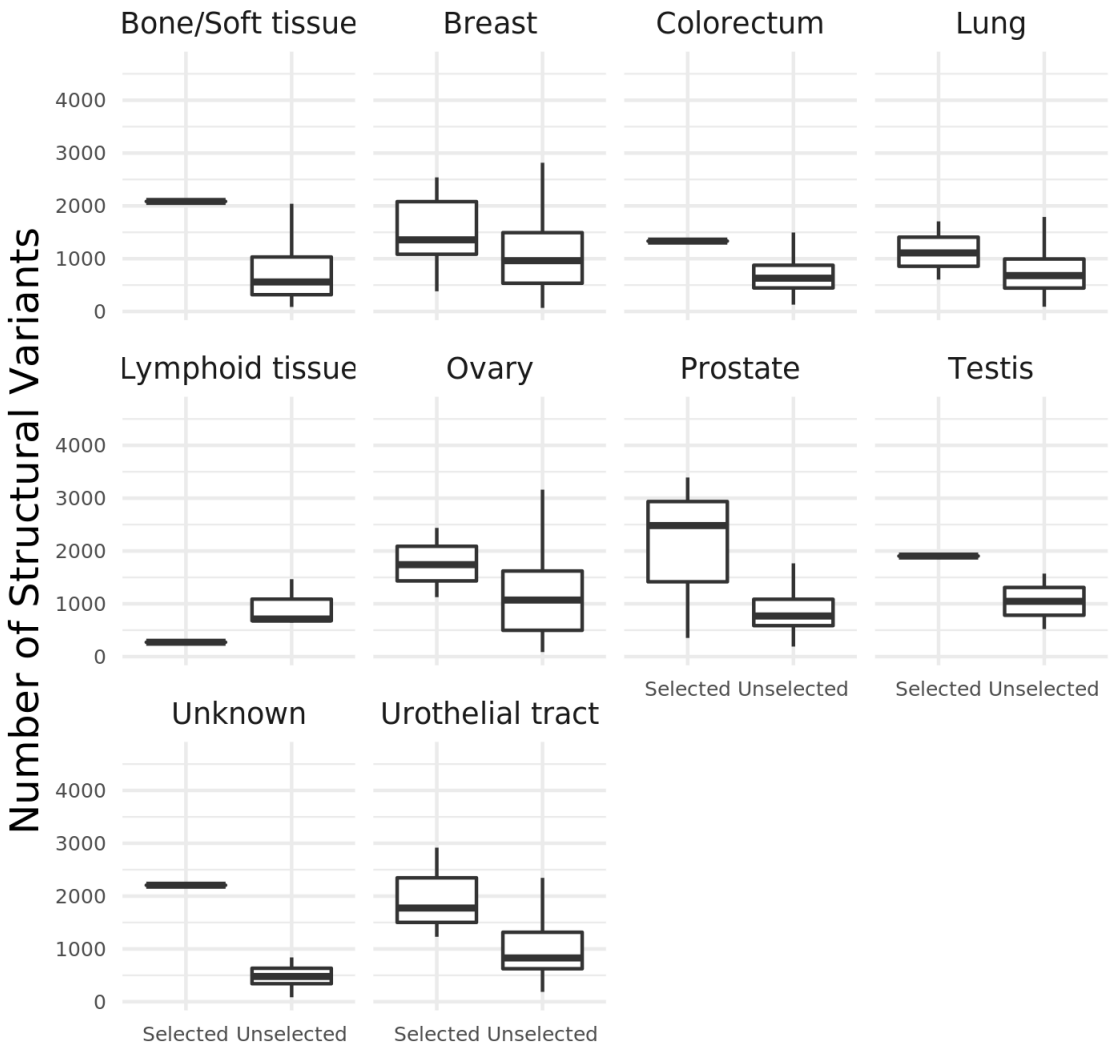


Figure 4.31: Number of structural variants in WGD samples in the HMF data set, split by tumour type. For clarity, only tumour types in which potential fusions were identified are presented. n Selected = 1-13, n Unselected = 3-340. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

>0.05).

4.5.2.6 Mutational Signatures

Similar to the results from the PCAWG data set, I did not observe consistent differences in the mutational signatures present in most of the selected samples, compared to other tumours of the same type, though there are a few interesting results (Figures C.27-C.36 in Appendix C.2.3). Breast cancer sample HMF001971A, which metastasised to the lung, is noteworthy for its high number of mutations compared to other breast cancers in our analysis (Figure C.38 in Appendix C.2.3). The mutations identified in HMF001971A were exclusively classified into signatures SBS5, SBS17a, and SBS17b. SBS17a and SBS17b were also observed in the breast cancer sample identified as a potential fusion in the PCAWG data set. As mentioned previously, SBS17a and SBS17b are not commonly observed in breast cancer, but have been observed in stomach, esophageal, and liver cancers, and B-cell lymphomas (Alexandrov et al., 2013, 2020).

The CUP sample that was identified, HMF002757A, is also rather interesting (Figure C.63). The metastatic tumour was identified as a neuroendocrine carcinoma, and all mutations identified were classified as SBS1, SBS5, SBS17a, and SBS17b. Comparing this sample to other CUPs is not particularly informative, as these represent a diverse group of tumours arising from various tissues. However, it is interesting to note the similarities between this sample and the aforementioned breast cancer identified in the HMF data set, in terms of the complement of mutational signatures identified.

I also investigated mutational signatures identified in the CN 1+1 and 2+1 regions that were identified in the analysis of the CV_{MD} (Figures C.37-C.66 in Appendix C.2.4). Similar to the results from the PCAWG data set, the proportion of mutations attributed to each mutational signature appeared to be quite consistent between different segments within the same tumour. As in the PCAWG data set,

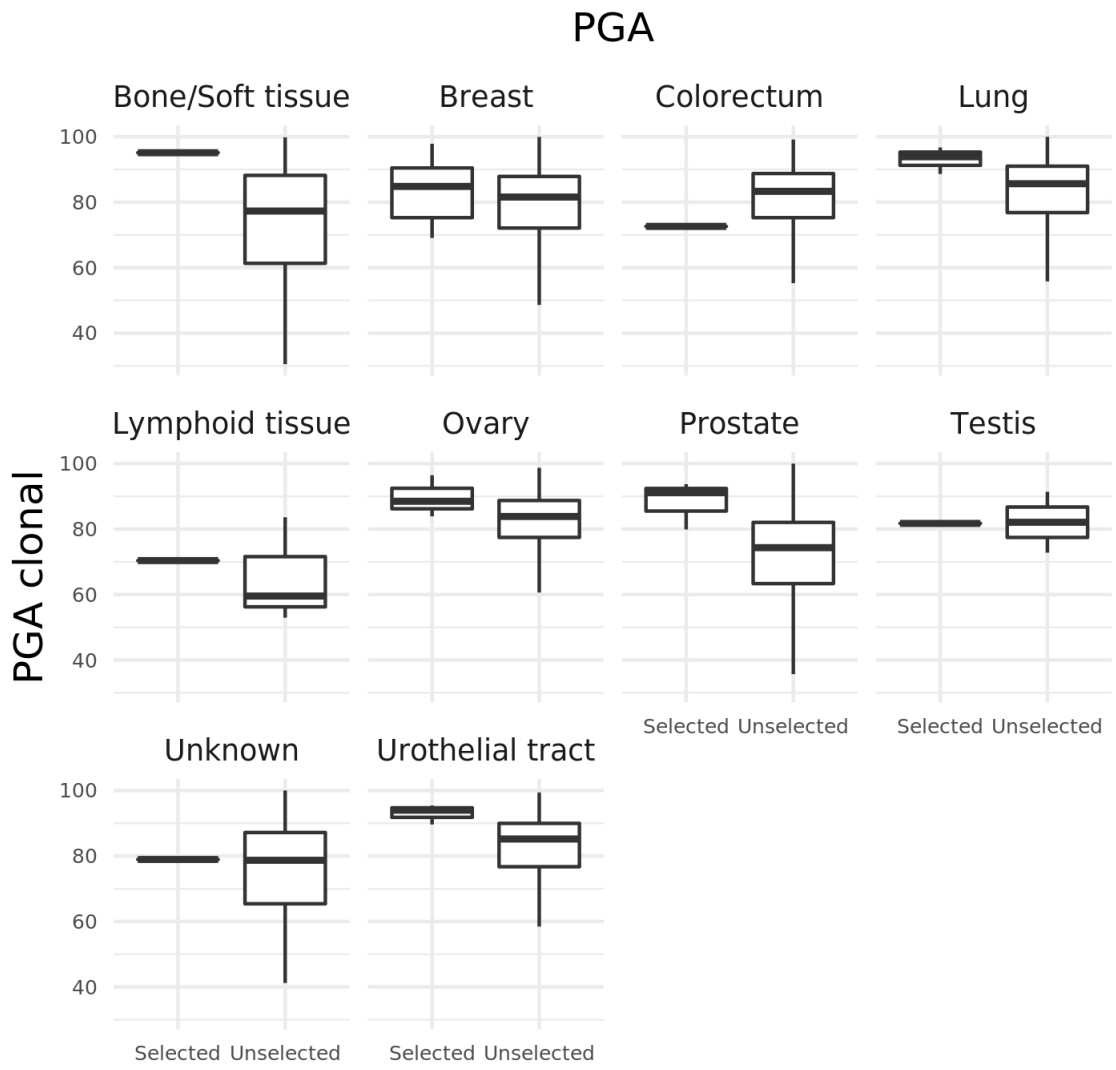


Figure 4.32: Percentage of the genome altered from a copy number of 4, split by tumour type. n Selected = 1-13, n Unselected = 3-340. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

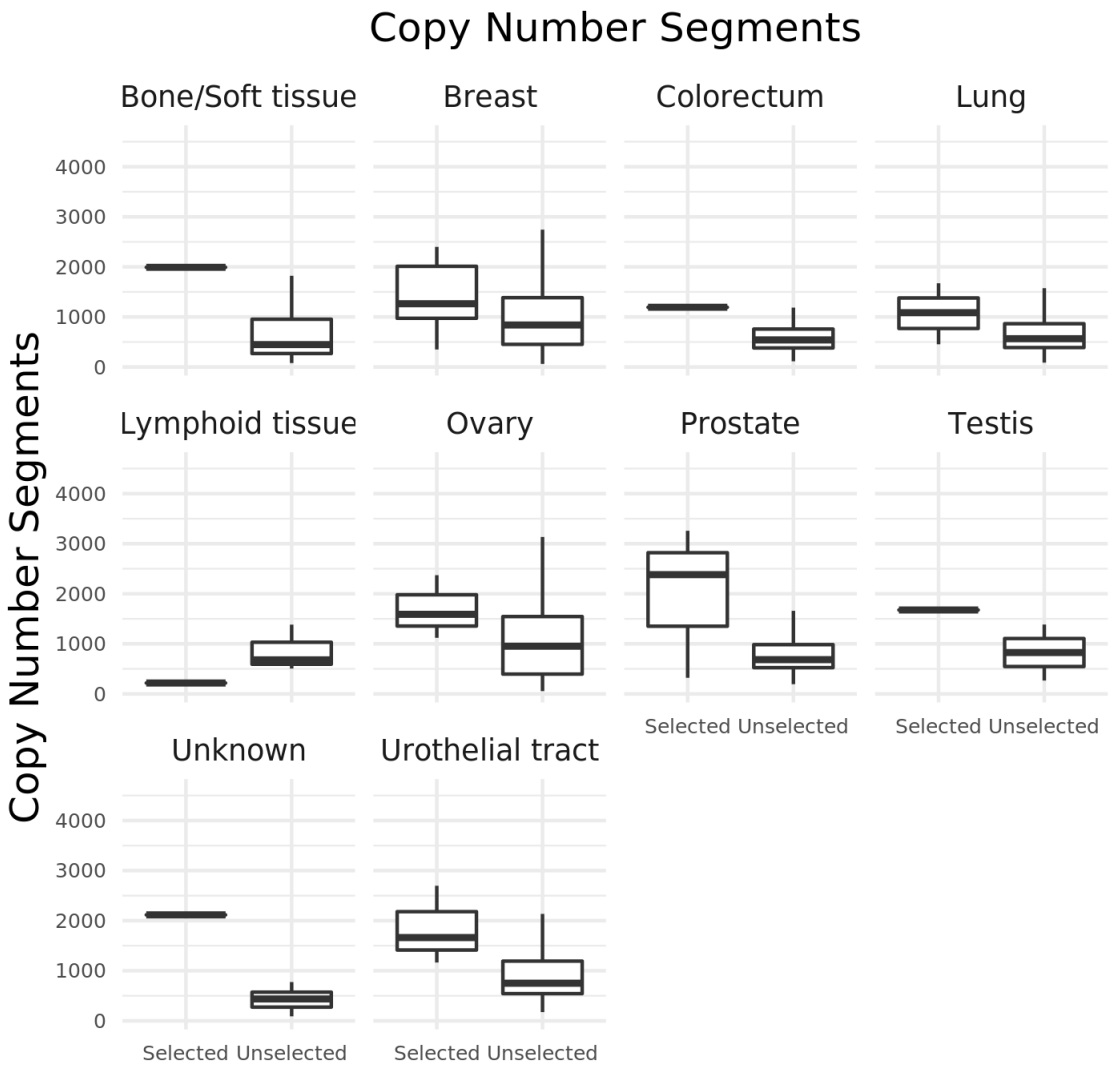


Figure 4.33: Number of copy number segments per genome, split by tumour type. n Selected = 1-13, n Unselected = 3-340. Individual comparisons were not significant after correction for multiple testing and are not displayed; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing p >0.05.

there was one possible exception to this trend, presented by the CUP identified (Figure C.63 in Appendix C.2.4). In this sample the proportion of mutations attributed to SBS5 appeared to decrease with increasing mutation density, suggesting that SBS5 may be the main cause of mutations in segments with low mutation density, but SBS17a and SBS17b were responsible for a higher proportion of mutations in segments with high mutation density.

4.6 Discussion

In this chapter I have presented a series of measures that may be used to identify products of cell fusion between a normal cell and a tumour cell. I applied these methods and measures to 6,337 whole genome sequenced tumours, and identified a total of 46 tumours from 15 different tumour types that may have originated from a cell fusion event. Premature chromosome condensation has been reported in five of these 15 tumour types (bladder, brain, breast, colorectal, and ovarian cancers), and cell fusion in human cancers has been reported for one further tumour type (melanoma; Kovacs (1985); Miles and Wolinska (1973); Atkin (1979); Reichmann and Levin (1981); Sainati et al. (1996); Lazova et al. (2013)).

In the PCAWG data set, the most commonly identified tumour types were ovarian cancer and paediatric brain cancer (4/16 each). I also identified three ovarian cancer samples in the HMF data set. In addition to observations of PCC in ovarian cancer patients, cell fusion has also been reported in mouse models of ovarian carcinoma and *in vitro* models of polyploid giant cancer cell formation in ovarian cancer (Ramakrishnan et al., 2013; Yart et al., 2020). Fusion between ovarian cancer cells and mesenchymal stem cells *in vitro* has also been reported, and was found to reduce chemosensitivity but decrease tumourigenicity (Melzer et al., 2018). Three ovarian cancers were identified as potential hybrids in the HMF data set, all of which had metastasised to the peritoneum, which is one of the most common sites for ovarian

cancer metastases (Figure 4.24; Lengyel (2010)). While the location of the ovarian metastases may not be unusual, the observation of PCC in ovarian tumours and cell fusion in models of ovarian cancer serves as a precedent for cell fusion occurring in this tumour type and supports the idea that cell fusion occurred in these tumours.

Unlike the ovarian cancers, no brain cancers, paediatric or otherwise, were identified in the HMF data set despite representing a quarter of the samples identified in the PCAWG data set. This difference can perhaps be explained by the differences in the composition of paediatric brain cancers found in the PCAWG data set and the central nervous system (CNS) tumours in the HMF data set. The paediatric brain cancers are mainly composed of medulloblastomas and piloastrocytomas, which are more commonly diagnosed in children than adults (Merchant et al., 2010). The metastatic CNS tumours in the HMF data set, on the other hand, are mostly gliomas and glioblastomas, which are brain tumours more commonly diagnosed in adults (Merchant et al., 2010).

PCC has been observed in a paediatric brain cancer (Sainati et al., 1996). Cell fusion has also been described in the brains of sex mismatched bone marrow recipients, resulting in purkinje neuron - BMDC hybrids identified by the presence of donor Y chromosomes and two X chromosomes (Weimann et al., 2003). Fusion between purkinje neurons and BMDCs has also been studied in animal models where purkinje neuron-BMDC fusions persisted as stable non-dividing binucleate heterokaryons whose nuclei assumed a purkinje neuron nuclear morphology (Weimann et al., 2003; Alvarez-Dolado et al., 2003). Although cell division was not observed in these hybrids, it begs the question whether other cell types in the brain can fuse with BMDCs, and whether the terminal nature of purkinje cells can be overcome to produce proliferating hybrids.

Interestingly, cell fusion has been suggested to occur in the brains of patients suffering from subacute sclerosing panencephalitis (SSPE), which is a complication of measles virus infection, occurring years after the initial infection (Paula-Barbosa

and Cruz, 1981; Esiri et al., 1982; Ludlow et al., 2015) Additionally, expression of measles virus fusion proteins from autopsies of SSPE patient brains have been found to cause cell fusion and syncytia formation *in vitro* (Cattaneo and Rose, 1993). It is tempting to speculate that fusion caused by measlesvirus may have played a role in these paediatric brain tumours. However, there was no evidence of viral DNA in these tumours. Without access to detailed medical records and tissue specimens, no connection can be made between measles virus infection and paediatric brain cancers in this case. The high frequency of paediatric brain cancers among the samples identified as potential cell fusions makes this an interesting area for further study.

Due to the ability of viruses to fuse cells I expected viral DNA to be identified in several samples that were selected for further study. However, viral DNA was only identified in one of the PCAWG tumours that was selected, and information on virus presence was not available for the HMF data set at the time of writing. Human papillomavirus 18 (HPV18) was identified in the bladder cancer sample that passed the thresholds. Cell fusion has been reported for the closely related HPV16 virus (Hu et al., 2009). However, no studies demonstrating cell fusion by HPV18 are available. Binucleate cells have been reported in keratinocyte raft cultures expressing the HPV18 E6 protein and in a HPV18 positive urethral sample. However, the mechanisms by which these binucleate cells were produced were not explored (Southern et al., 2001; Shigehara et al., 2014). The roles of HPV16 and HPV18 in cell fusion warrant further research. These two viruses are of particular interest as they express proteins that inactivate p53, which may facilitate tolerance of cell fusion or whole genome duplication (Thomas et al., 1999).

Of the 16 PCAWG samples identified, only the melanoma sample was identified as metastatic. This sample distinguished itself from other WGD melanomas based on the number of chromothripsis events and structural variants present in the sample. Chromothripsis events were found on 16-18 different chromosomes, suggestive

of a PCC event occurring after cell fusion. In addition, the timing of the WGD in this sample appeared as particularly early compared to other melanoma samples, further suggesting that this event may have been an important early event in the development of this tumour. It is perhaps not surprising that only one of the tumours selected displayed chromothripsis patterns suggestive of PCC, as PCC only occurs if a mitotic cell fuses with a cell in interphase, with different degrees of DNA distortion occurring depending on the specific phase of interphase that the fusing cell was in (Sperling and Rao, 1974). Considering the rarity of a viable cell fusion hybrid, a hybrid that happened to result from two cells fusing at different stages of the cell cycle resulting in PCC must be rarer still.

The melanoma sample identified in the PCAWG data set distinguished itself from other WGD melanomas in almost every measure. In addition to the large number of chromothripsis events observed in this sample, the sample also contained higher numbers of structural variants, copy number segments, and PGA compared to other WGD melanomas. Interestingly, the number of INDELs identified in this sample appeared to be lower than or comparable to other melanomas. This sample represents a strong candidate for further follow up studies. We have reached out to the individuals responsible for collection and storage of this specimen in the hope that they may be able to provide further clinical information. If sample material is available, I plan on carrying out long read sequencing, in order to definitively resolve whether I can identify haplotypes with different mutation densities at sites that I have identified in the WGS data. This would validate my approach of looking for segments with variation in mutation density.

Investigating DNA methylation in these samples might also be an interesting avenue for follow up analysis, in particular if sample material is also available from the primary tumour and the matched normal. Different tissues and cell types have been shown to have distinct methylation patterns (Gu et al., 2016; Farlik et al., 2016). Results from *in vitro* studies of cell fusion have suggested that methylation

patterns, rather than patterns of gene expression, may be more informative for differentiating fused cells from their parental counterparts (Su et al., 2015). Su et al. (2015) demonstrate that the epigenetic profiles of fused breast cancer cell lines form a distinct intermediary state between the parental breast cancer cells. Further, it has been noted that very few regions are differentially methylated upon whole genome duplication during megakaryocyte differentiation (Farlik et al., 2016). Taken together, this suggests that analysis of the epigenomes of these tumours may be an effective way to differentiate between whole genome duplication and cell fusion. With the advent of long read Tet-assisted pyridine borane sequencing (lrTAPS, Liu et al. (2020)), which allows for phasing of the epigenome, the use of methylation data for identification of cell fusion in human tumours may well lead to a breakthrough in the cell fusion field.

Unexpectedly, the melanoma sample appeared very similar to several other WGD melanoma samples in terms of the complement of mutational signatures present, and each of the CN 2+1 and 1+1 segments within this sample appeared to carry mutations attributed to most or all of the mutational signatures identified within the sample. This was a common trend across almost all of the samples that were investigated and did not match my expectations of observing mutational signatures associated with ordinary cellular processes in segments with lower mutation density, and mutational signatures associated with cancers in segments with higher mutation density. This unexpected trend in the distribution of mutational signatures across segments of different mutation density may be explained by the exposure of all copy number segments within the cell to ongoing mutational processes after cell fusion, regardless of their origin. In the initial stages after cell fusion, one would expect to mainly observe mutations attributed to SBS1 and SBS5 in segments derived from the normal cell, and other mutational signatures on the segments derived from the tumour cell. However, if the tumours that were selected are derived from a cell fusion event, they will have had the opportunity to acquire further cancer related

mutations over time. If these post-fusion mutations have accumulated evenly across the genome one would still expect to see differences in mutation density between segments derived from different cells; however, the difference in the complement of mutational signatures would be lost.

These results can also partly be attributed to the lack of resolution in the methods used, in that I do not have information about the individual chromosome copy on which each mutational signature was observed. A copy number state of 1+1 or 2+1 in which some segments are derived from a normal cell and some from a cancer cell, when considered as a whole, will likely appear similar to a segment in which only copies from the cancer cell are retained. To address this, future work would involve phasing the mutations in these segments to identify if mutational signatures associated with ordinary physiological processes are found primarily on one chromosome copy, and mutational signatures associated with cancer on the other.

Although the analysis of mutational signatures was largely unsuccessful in differentiating the samples that I had identified as potential fusions from other WGD samples, several interesting trends emerged. SBS17a and SBS17b were unexpectedly identified in breast cancer samples from both the PCAWG and HMF data sets, as well as in the CUP sample that was identified in the HMF data set. As previously stated, the presence of these signatures is uncommon in breast cancers but has been associated with stomach and esophageal cancers, and corresponding cell lines (Alexandrov et al., 2013, 2020; Petljak et al., 2019). SBS17a and SBS17b have also been associated with somatic retrotransposition (Petljak et al., 2019). It is possible that the unusual presence of SBS17a and SBS17b in these tumours reflects an alternative origin of these tumours, such as cell fusion. The presence of SBS9 in the PCAWG breast cancer sample, if true, also suggests an alternative origin due to its association with AID activity, or polymerase η repair of cytosines deaminated by AID (Alexandrov et al., 2013; Maura et al., 2019). AID expression

is ordinarily restricted to germinal centre B-cells and B-cell malignancies, though aberrant AID expression has been described following inflammation (Pasqualucci et al., 2004; Shimizu et al., 2012). This requires further study and I believe that these breast cancer samples may be interesting candidates for further follow up.

The presence of SBS17a and SBS17b in the CUP sample is perhaps less surprising, as the biopsy was taken from the cardia, which is the part of the stomach that meets the oesophagus (NCI, NCI). SBS17a and SBS17b are commonly found in stomach and esophageal cancers (Alexandrov et al., 2013), thus it is likely that the presence of these signatures can be attributed to the physiological location of the metastasis. Indeed, the distribution of mutational signatures in this sample appears very similar to that seen in the gastric and esophageal cancers in the PCAWG data set (Figures C.63, C.16, and C.4 in Appendix C). Although the primary tumour type from which this sample originated remains unknown, the tumour subtype specific information suggests that this is a neuroendocrine carcinoma (NET), and the tumour was previously listed as such, before being reclassified as a CUP. NETs represent an exceptionally diverse group of tumours whose primary sites may be found in the lungs, thymus, stomach, duodenum, ileum, cecum, appendix, colon, rectum, pancreas, and liver (Yao et al., 2008). However, classification or identification of the primary site is not possible for roughly 13% of NETs (Yao et al., 2008). Thus, this CUP remains an intriguing sample, as its origins are not clear cut and cell fusion remains a possible initiating event.

Finally, my analysis of mutational signatures also highlighted a pancreatic endocrine neoplasm (PAEN) that stood out from similar tumours in the PCAWG data set due to its low mutation burden and the attribution of all mutations in this tumour to signatures SBS1 and SBS5. Pancreatic endocrine neoplasms, sometimes referred to as pancreatic neuroendocrine tumours (PanNETs) are a subset of NETs that arise in the pancreas. This PAEN sample appeared very similar to other PAEN samples in the PCAWG data set across the different measures of genome stability,

with the exception of PGA, which was much higher in the sample that was identified as a potential fusion, reaching almost 100% (Figure 4.19). In contrast, the mutation burden and the number of INDELs identified in this sample appeared much lower than in other tumours of the same type, making this a rather unusual sample. The copy number profile of this sample is characterised by the loss of whole chromosomes rather than chromosome fragments, with respect to a tetraploid genome, reminiscent of a WGD event followed by chromosome loss (data not shown). The copy number profile, in combination with the mutational signatures observed, and the lack of other defining features, suggests that this tumour may have arisen from WGD or cell fusion of normal cells or early stage cancer cells, followed by loss of genetic material, making this another interesting candidate for further follow up.

To date, no other study has attempted to identify cell fusion in whole genome sequencing data, so I compare my results to lab based studies of cell fusion and its effect on the genome. While there appears to be a significant overlap in the effects of WGD and cell fusion on the genome, it is interesting to consider both the similarities and the differences between the genomes of the samples that were identified as potential cell fusions and other WGD tumours of the same type. Increased genomic instability has been reported as a consequence of both cell fusion and WGD compared to parental cells (Delespaul et al., 2019; Dewhurst et al., 2014; Kuznetsova et al., 2015). Instead of comparing the tumours that were identified as potential fusions with their diploid counterparts, I opted to compare them to their tetraploid counterparts. Unfortunately, the statistical power to detect differences between these two groups was limited by the small number of samples that were identified as potential fusions within each tumour type. However, I did observe a trend towards increased numbers of structural variants and copy number segments in potential fusions belonging to bladder cancer, bone cancer, ovarian cancer, and melanoma in the PCAWG data set. In the HMF data set I observed a similar trend for tumours from bone and soft tissue, breast, lung, ovary, prostate, and urothelial

tract. Interestingly, the presence of higher numbers of SVs and copy number segments did not necessarily coincide with higher levels of PGA. The bladder cancer that was identified appeared to have a comparable level of PGA to other tumours in the PCAWG data set, whereas the bone cancer had a dramatically lower PGA. PGA levels were higher in selected samples in all other tumour types in the PCAWG data set, with the exception of paediatric brain cancers, which had broadly similar numbers of copy number segments and SVs between groups. In the HMF data set there was a slight difference in this trend, in that all of the tumour types in which an increase in both the number of SVs and copy number segments was observed tended to also have higher levels of PGA. Thus, overall, there appeared to be higher levels of genomic instability in tumours identified as potential cell fusions.

While this may reflect incidental individual variation between selected samples and similar WGD tumours, it may also reflect tumour type specific differences in response to cell fusion. The results of lab based studies have suggested that there may be a great degree of heterogeneity with regards to the behaviour and phenotype of fused cells and whole genome duplicated cells (Yin et al., 2020; Kuznetsova et al., 2015). In future, I hope that by expanding this analysis to larger data sets, such as those curated by Genomics England, I may be able to identify enough potential fusions within each tumour type to allow for more rigorous assessment of the differences between cell fusion and whole genome duplication on the cancer genome.

Chapter 5

In vitro induced cell fusion and its effect on genome stability

5.1 Introduction

Whole genome duplication is considered a transformative event in the evolution of the cancer genome, and can occur in several different ways. Previous research into the effects of whole genome duplication and cell fusion on the transcription and behaviour of cancer cells has revealed a large overlap in the results of these studies (see Sections 1.1.1 and 1.2). However, it has been argued that the consequences of cell fusion may be worse than those of whole genome duplication by failure of cytokinesis, due to the potentially deleterious and unpredictable effects that bringing together different cell types or cells in different stages of the cell cycle may have, such as premature chromosome condensation (Koulakov and Lazebnik, 2012). Further to this, one might expect virus-induced cell fusion to be even more deleterious than spontaneous fusion due to the activation of APOBEC3 cytidine deaminases in response to viral infection, or inhibition of tumour suppressors by some viruses such as HPV (Scheffner et al., 1990). To date, no study has investigated whether the consequences of cell fusion are distinct from those of whole genome duplication by

other means. The different effects of viral cell fusion compared to spontaneous cell fusion have also not been explored.

The intentions of this work were to investigate whether cell fusion caused by a virus had different consequences for the cell than spontaneous fusion, or whole genome duplication through failure of cytokinesis. My aim was to generate whole genome duplicated cell lines through cell fusion using either polyethylene glycol (PEG) as a model of spontaneous cell fusion, or pseudoviral particles to emulate virus induced cell fusion, and to compare these to cell lines generated through Cytochalasin B induced failure of cytokinesis, as well as to the diploid cell line of origin. These cell lines would be monitored using a series of metastasis assays, long term karyotype analysis, and sequencing, to determine whether there are qualitative or quantitative differences in the effects of cell fusion on the cancer genome and metastatic properties, compared to the effects of WGD by other means.

A number of different approaches and cell lines were trialled, which included fusing different cell types, isolating cells using laser microdissection, generating cell lines that stably expressed fluorescent reporters, and generating cells expressing different halves of a bi-fluorescent complementation reporter. None of these approaches generated viable fused WGD cell lines in sufficient numbers for the quantitative assays that I had planned. However, I was able to isolate one confirmed WGD virally fused clone, 5+ confirmed WGD Cytochalasin B treated clones, and 5+ confirmed diploid co-cultured clones. This limited the panel of comparators and precluded my ability to compare the effects of whole genome duplication by different means. For this reason I present a limited analysis of five different clones in this chapter. I surveyed one virally fused clone, three Cytochalasin B treated clones, and one diploid clone derived from co-culture as a control, for a period of three months.

5.2 Materials

5.2.1 Cell Cultures

HCT116 cells were grown as described in Section 3.2.1 in Chapter 3.

5.2.2 Fluorescent Cytoplasmic Dyes

CellTracker Blue CMAC dye (Invitrogen C2110) and CellTracker Orange CMTMR dye (Invitrogen C2927) were used to label cells for fusion experiments.

5.2.3 Pseudoviral Particles

Pseudoviral particles were designed and produced as described in Section 3.2.2 in Chapter 3.

5.3 Methods

5.3.1 Interruption of Cytokinesis Using Cytochalasin B

HCT116 cells were seeded in a 10cm plate and allowed to adhere overnight. Cells were then treated with 8µg/mL Cytochalasin B for 24 hours. After 24 hours in Cytochalasin B, cells were provided with fresh media, and allowed to rest for 24 hours prior to FACS sorting.

5.3.2 PEG-Induced Cell Fusion

HCT116 cells were seeded in 10 cm plates and labelled with either CMAC blue or CellTracker Orange fluorescent dyes after being allowed to adhere overnight. The two populations of cells were mixed in equal ratios and centrifuged at 820 x g for 10 minutes, after which the supernatant was discarded. Cells were re-suspended in serum free medium and incubated at 37°C for 10 minutes. Polyethylene glycol

(PEG-1000) 50% in PBS was added drop-wise to the cells, and cells were then mixed for 1-1.5 minutes using a sterile loop. Serum free medium was then added to the mixture of cells and PEG, and cells were incubated at 37°C for 10 minutes. Cells were then centrifuged again for 10 minutes at 590 x g, after which they were re-suspended in complete medium and plated in a new 10cm plate. Cells were allowed to rest for 24 hours before FACS sorting.

5.3.3 Virus-Induced Cell Fusion

The protocol for producing virally fused cells was intended to mirror that used to produce cells fused with PEG. HCT116 cells were seeded in 10cm plates and allowed to adhere overnight before being labelled with CMAC blue or CellTracker Orange fluorescent dyes. The two populations of cells were mixed and centrifuged at 820 x g for 10 minutes. The supernatant was removed, cells were re-suspended in serum free medium, and incubated at 37°C for 10 minutes. Cells were centrifuged again at 820 x g for 10 minutes, after which the supernatant was discarded and replaced with serum free medium containing VSV-G pseudotyped HIV at a multiplicity of infection (MOI) of 50 viral particles per cell, at pH 5. Media at pH 5 was used to ensure cell fusion at the plasma membrane, as VSV-G mediated cell fusion is ordinarily initiated at low pH in lysosomes (Matlin et al., 1982). The cell and virus combination was mixed with a sterile loop for 1-1.5 minutes, after which additional serum free medium was added to the mixture. The cells were then incubated at 37°C for 10 minutes, before centrifugation at 590 x g for 10 minutes. Cells were re-suspended in complete medium, plated in 10 cm plates, and allowed to rest for 24 hours before FACS sorting.

5.3.4 Isolation of Fused Cells Using Flow Cytometry

FACS sorting was carried out using a Sony SH800 Cell Sorter. Populations of PEG-fused and virus-fused cells were identified based on the presence of both fluorescent

markers, and single-cell sorted into 96-well plates. A population of cells that had been fluorescently labelled and co-cultured but not subjected to PEG or virus treatment was also single cell sorted using the same parameters used to sort fused cells. Whole genome duplicated Cytochalasin B treated cells were identified based on DNA content determined by Hoechst staining. Cytochalasin B treated cells were selected for 8N DNA content (WGD cells undergoing mitosis), and single cell sorted into a 96-well plate.

For each experimental condition, a population of cells was also bulk sorted into 24-well plates, and single-cell sorted two weeks later for 8N DNA content, determined by Hoechst staining.

5.3.5 Metaphase Spread

To confirm the ploidy of the single cell sorted clones, metaphase spreads were carried out to assess the number of chromosomes in each cell. Cells were seeded in T25 flasks and allowed to adhere overnight after which they reached approximately 60% confluency. Cells were then incubated with 30ng/mL colcemid for 3 hours. Cells and medium were collected from the flask after trypsinisation and centrifuged at 300 x g for 10 minutes. The supernatant was removed, cells were re-suspended in hypotonic solution (75mM KCl) at 37°C, and incubated at 37°C for 15 minutes. Cells were then centrifuged at 300 x g for 10 minutes, after which the supernatant was removed and cells were re-suspended in cold Carnoy fixative solution (three parts methanol to one part glacial acetic acid). Cells were fixed at room temperature for 10 minutes, then centrifuged at 300 x g for 10 minutes, after which the supernatant was removed and cells were re-suspended in a low volume and dropped onto microscopy slides. Preparations were stained and mounted using VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories H1200).

Clones derived from PEG-induced fusion, virus-induced fusion, and Cytochalasin B treatment that were determined to be diploid using this method were discarded.

5.3.6 Microscopic Analysis of Metaphase Spreads

For initial metaphase spreads used to determine whether single cell sorted cells were WGD or not, metaphase spreads were imaged using a Leica DMi8 FALCON microscope, using a 100x oil objective or a 63x objective. Metaphase chromosomes were counted manually.

For long term tracking of the karyotype of selected clones, chromosome harvests and counting were carried out at monthly intervals by Julia Jankowska, Paulina Bryjka, and Daniela Moralli of the Chromosome Dynamics Core at the Wellcome Centre for Human Genetics.

Three independent chromosome harvests were carried out for each cell line at each time point. A minimum of 20 cells were counted for each replicate.

5.3.7 Multicolor Fluorescent In Situ Hybridisation (FISH)

Multicolor fluorescent in situ hybridisation (M-FISH) was also carried out at monthly intervals by Julia Jankowska, Paulina Bryjka, and Daniela Moralli of the Chromosome Dynamics Core at the Wellcome Centre for Human Genetics. M-FISH was performed using an M-FISH probe kit 24XCyte (Zeiss MetaSystems) according to the manufacturer instructions. Images were acquired with an Olympus BX60 microscope for epifluorescence equipped with a JAI CVM4 camera (Leica Biosystems), and analysed using the Cytovision software (Leica Biosystems).

A minimum of 10 cells were counted for each cell line at each harvest.

5.3.8 Analysis of Karyotype Over Time

All analyses were carried out using R version 3.6.2, unless stated otherwise (R Core Team, 2018). Plots were generated using version 3.3.0-2 of the `ggplot2` package, version 0.6.0 of the `ggsignif` package, version 0.5.3 of the `ggridges` package, version 0.4.0 of `ggpubr`, version 0.5.1 of `viridis`, and version 1.1-2 of `RColorBrewer`

(Wickham, 2016; Ahlmann-Eltze, 2019; Wilke, 2021; Kassambara, 2020; Garnier, 2018; Neuwirth, 2014).

Bimodality was assessed, and each mode calculated, using version 1.1.9 of the `BimodalIndex` package (Coombes, 2019). Distributions were treated as unimodal if the Bimodal Index (BI) was below 1.1 or if the modes of the two peaks were within 10 chromosomes of each other.

5.3.9 Karyotype Stability

To assess the stability of the karyotype, I calculated the proportion of chromosomes affected by de novo alterations (PCDA), for each cell assessed by M-FISH, using the following formula:

$$\text{PCDA} = \frac{\text{Total number of de novo alterations}}{\text{Total number of chromosomes}} \tag{5.1}$$

Higher PCDA values indicate a more unstable karyotype. De novo alterations were defined as non-marker chromosome alterations.

5.4 Results

5.4.1 Effects of Cell Fusion and Whole Genome Duplication on Karyotype Over Time

Metaphase chromosome counting was carried out for each of the five cell lines at monthly intervals to monitor the total number of chromosomes over time (Harvests 0-3). The clones that I investigated were a WGD virally fused clone (V-F4), three WGD Cytochalasin B treated clones (C-F10, C-F3, and C-G9), and a diploid clone (M-B10) derived from co-culture, as a control.

I found that there was a significant increase in the total number of chromosomes

in clones C-F10, C-F3, and V-F4 after 1 month of culture (harvest 0 to harvest 1), while clones M-B10 and C-G9 did not differ significantly in terms of the number of chromosomes present at harvest 0 and harvest 1 (Figure 5.1, Wilcoxon Rank sum test with Benjamini-Hochberg correction for multiple testing $p < 0.05$). Despite appearing as an increase in chromosome number relative to harvest 0, the median number of chromosomes remained below the 90 chromosomes expected from whole genome duplication or fusion of two 45X HCT116 cells. Thus, rather than representing a gain of individual chromosomes, a more biologically parsimonious explanation for this observation is that it represents fluctuation in prevalence of subclones which have lost different numbers of chromosomes. This variation may also in part be due to technical variation in chromosome harvest conditions.

After two months of culture (harvest 0 vs harvest 2), C-F10 and C-F3 were no longer significantly different from their counterparts at harvest 0, suggesting that the cells had lost chromosomes after the increase in the number of chromosomes that occurred between harvest 0 and harvest 1(Figure 5.1, Wilcoxon Rank sum test with Benjamini-Hochberg correction for multiple testing $p < 0.05$). Clones C-G9 and V-F4 appeared to have lost a significant number of chromosomes between harvest 0 and harvest 2; this loss of chromosomes was also significant between harvests 1 and 2 (Figure 5.1, Wilcoxon Rank sum test with Benjamini-Hochberg correction for multiple testing $p < 0.05$). Clones C-F3 and C-G9 continued to lose chromosomes between harvests 2 and 3, whereas the number of chromosomes in clone C-F10 remained constant between harvests 2 and 3, and clone V-F4 gained chromosomes between harvests 2 and 3 (Figure 5.1, Wilcoxon Rank sum test with Benjamini-Hochberg correction for multiple testing $p < 0.05$).

Overall, clones C-F3 and C-G9 lost a significant amount of chromosomes after three months of harvest, following a brief increase in the number of chromosomes after 1 month of harvest (Figure 5.1). After three months of culture C-F10 and V-F4 did not differ significantly in terms of the number of chromosomes, when compared

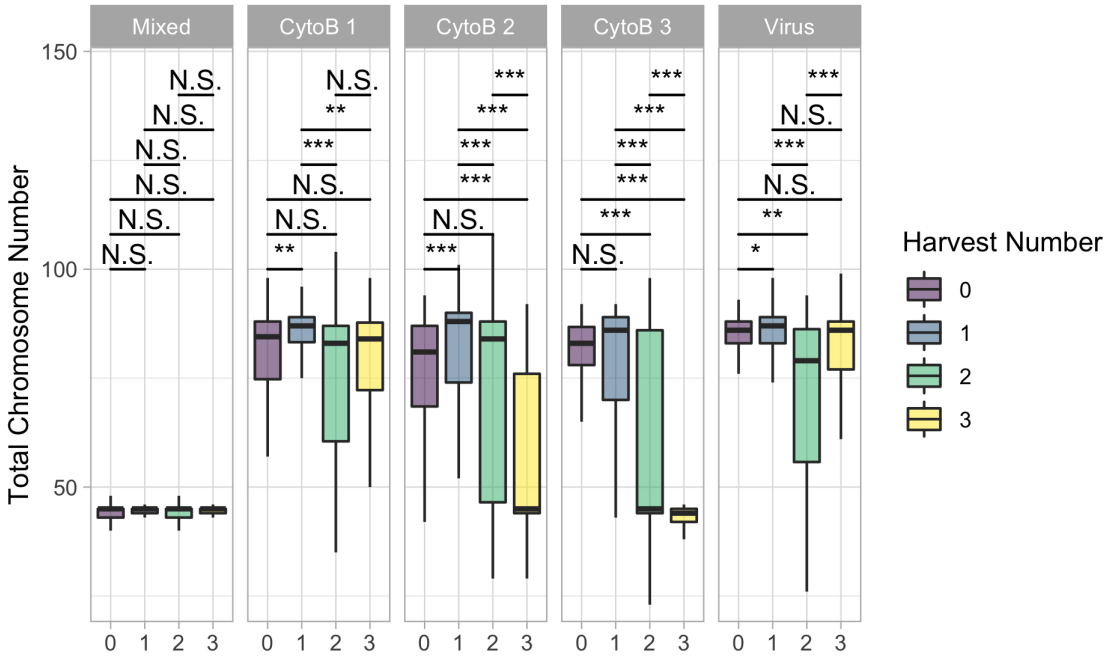


Figure 5.1: Total number of chromosomes in each clone compared across different harvests. Chromosomes were harvested at intervals of 1 month. Three independent chromosome harvests were carried out for each clone at each harvest. Two-sided Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

to harvest 0; however, both clones underwent significant fluctuations in chromosome number in between the two time points (Figure 5.1). The number of chromosomes in control clone M-B10 did not undergo significant changes at any point during culture (Figure 5.1).

A large range of chromosome numbers was observed at each time point in the WGD clones, which likely reflects the presence of subclones with varying numbers of chromosomes (Figure 5.1). To investigate this further, I studied the distribution of chromosome counts between different time points (Figures 5.2 and 5.3). Interestingly, clones C-G9 and C-F3 appeared to have a population of diploid cells that increased in frequency between harvests, and appeared to dominate the final harvest for clone C-G9 (Figure 5.2). Both clones C-F10 and V-F4 appeared to have populations of cells that were near diploid or near triploid at later harvests; however, the majority of the cells surveyed remained WGD (Figures 5.2 and 5.3). The dis-

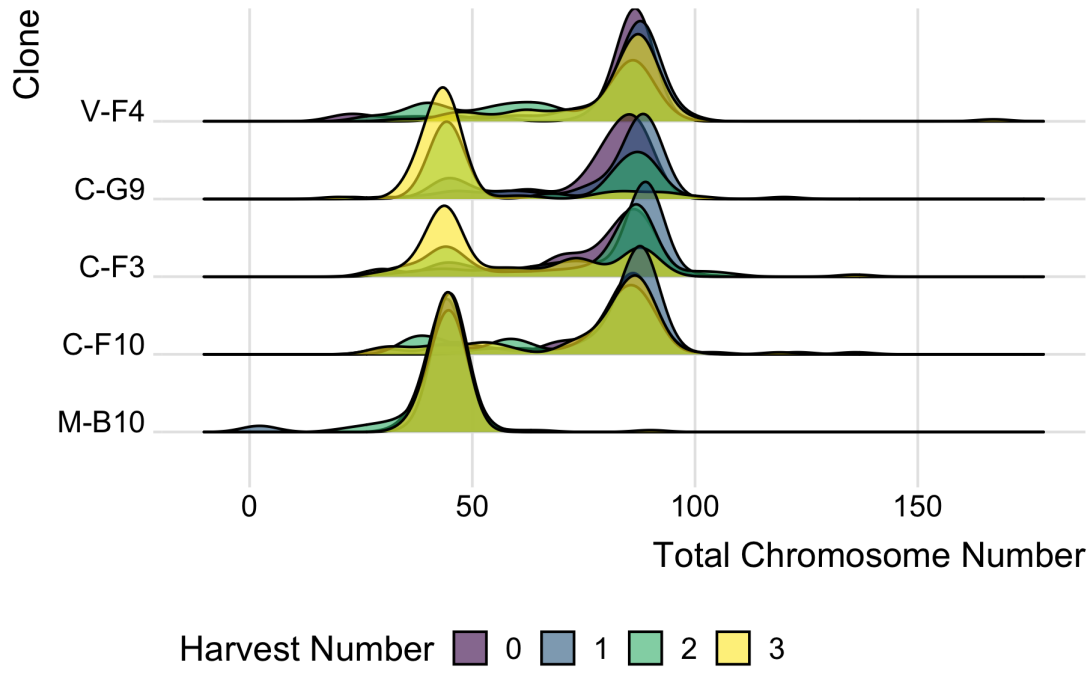


Figure 5.2: Distribution of the total number of chromosomes in each clone across different harvests. Chromosomes were harvested at intervals of 1 month.

tribution of chromosomes in clone M-B10 was consistently centred around diploid (Figure 5.2).

I calculated the modal chromosome number at each peak for bimodally distributed samples and compared these to control (Table 5.1). Interestingly, for all of the samples that had a diploid population, the mode of this population was 45 chromosomes at the final harvest.

	Harvest 0		Harvest 1		Harvest 2		Harvest 3	
	Mode 1	Mode 2	Mode 1	Mode 2	Mode 1	Mode 2	Mode 1	Mode 2
M-B10	45		45		45		45	
C-F10	44	85	59	86	47	85	45	85
C-F3	44	82	50	86	44	86	45	84
C-G9	53	84	49	87	44	86	45	
V-F4	37	86	54	86	47	82		87

Table 5.1: Modal chromosome numbers for each clone at each harvest. Mode 1 represents the modal chromosome number for near diploid cells, and Mode 2 represents the modal chromosome number for WGD cells. Modes estimated from bimodally distributed samples are given to the nearest whole chromosome number.

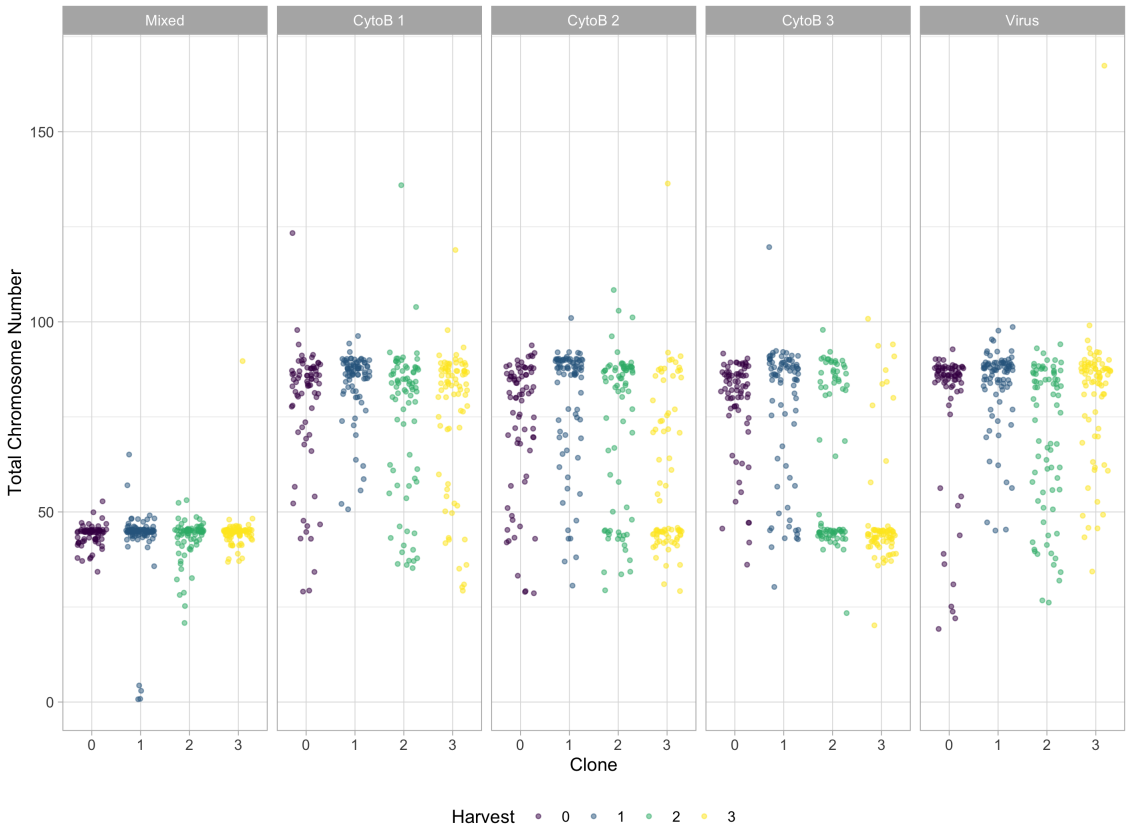


Figure 5.3: Chromosome counts for individual clones at each harvest.

5.4.2 Changes in Genome Stability

To characterise changes in the karyotype of the clones at the chromosome level, M-FISH chromosome counting was carried out at monthly intervals, in addition to the chromosome counting of metaphase spreads. Individual chromosome counts, and chromosomal alterations for each clone at each time point are represented in Figure 5.4. Images of representative karyotypes for each clone at each harvest are presented in Appendix C.3.

The most common alterations, which were observed in almost every cell surveyed were a translocation between chromosomes 16 and 8 (t16/8), a translocation between chromosomes 18 and 17 (t18/17), and the presence of a derivative chromosome 10 (der(10)). These three alterations constitute marker chromosomes consistent with previous reports of the HCT116 karyotype (Simões-Sousa et al., 2018; Kleivi et al., 2004). Interestingly, in the WGD clones der(10) was only observed at low frequency at harvest 0; however, it was present in almost every cell surveyed at subsequent harvests (Figure 5.4). No other alterations appeared as frequently as these three marker chromosomes. A derivative chromosome 4 (der4) was observed at least once in all of the WGD clones. The presence of der4 has also previously been described in HCT116 cells (Knutsen et al., 2010). Alterations of chromosomes 1 and 2 were also observed in all of the WGD clones, with C-F10 displaying a particularly high frequency of alterations at harvests 0, 2, and 3. I also noted frequent alterations affecting chromosome 8 in clone V-F4, which consisted of additional translocations and derivatives co-occurring with the t16/8 marker translocation.

While I did not observe a consistent pattern in the types of alterations or the specific chromosomes affected by alterations in the WGD clones, Figure 5.4 highlights the increased variability in both chromosome number and presence of alterations within the WGD clones when compared to the control clone M-B10. This also highlights the heterogeneity observed within each clone over time.

When I compared the number of chromosomes affected by alterations, excluding



Figure 5.4: Heatmap of karyotype with alterations represented as symbols. Some alterations are present on more than one chromosome copy. Representative karyotypes for each clone at each harvest are provided in Appendix C.3.

marker chromosomes t16/8, t18/17 and der10, I found that there was a significant increase in the number of chromosomes carrying alterations compared to control at harvests 0 and 2 (Figure 5.5 A; Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing $p < 0.01$). There did not appear to be a significant difference in the proportion of chromosomes affected within each clone between different harvests for C-F10, C-F3, and V-F4 (Figure 5.5 B). In contrast, significant fluctuations in the proportion of chromosomes affected by de novo alterations were observed for M-B10, and a significant decrease in de novo alterations was seen in C-G9, which became near diploid (Figure 5.5 B).

5.5 Discussion

In this chapter I isolated and clonally expanded a number of single cells that had undergone WGD either through failure of cytokinesis or through virus-induced cell-cell fusion. These cells were karyotyped at monthly intervals with the aim of investigating the effects of WGD on the evolution of the cancer genome, and to investigate qualitative and quantitative changes in genome content, attributed to the manner in which WGD was achieved.

Unlike the diploid control clone which maintained a stable chromosome number, all of the WGD clones underwent significant fluctuation in the total number of chromosomes present at each harvest. During culture, dominant diploid subclones emerged in clones C-G9 and C-F3. The diploid subclones appeared to be present at all harvests, increasing in frequency over time. Interestingly, diploid subclones also appeared to be present in clones V-F4 and C-F10 at all harvests, but the majority of the cells surveyed maintained a near tetraploid number of chromosomes. I also observed an increase in the proportion of chromosomes affected by alterations in the WGD clones compared to control, which fluctuated between different harvests, suggesting increased heterogeneity and ongoing subclonal diversification and selec-

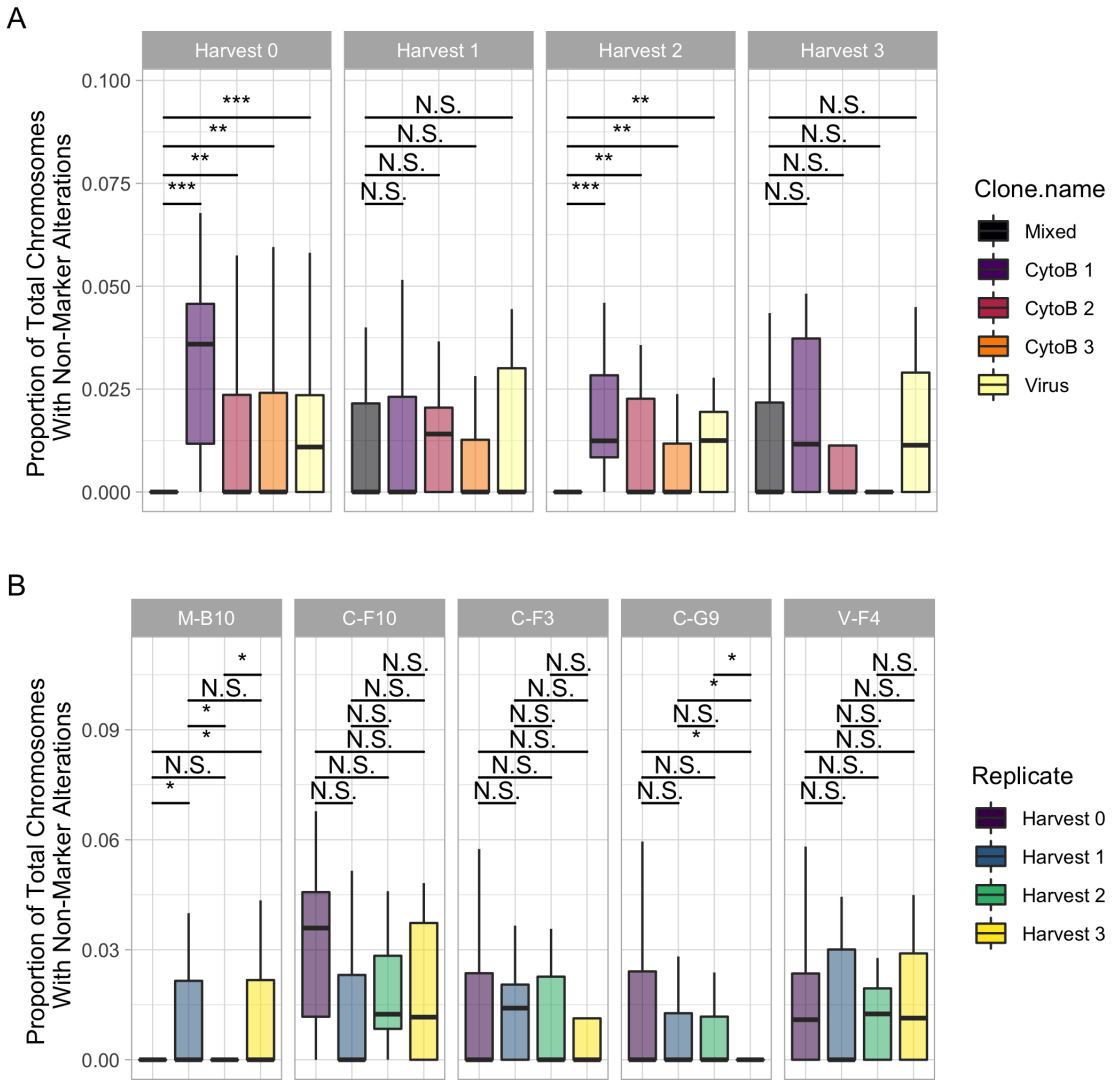


Figure 5.5: Proportion of chromosomes affected by non-marker alterations. Two-sided Wilcoxon rank sum test with Benjamini-Hochberg correction for multiple testing. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

tion in the WGD clones. These results are consistent with the literature, which suggests that whole genome duplication, or cell fusion, gives rise to an initial increase in subclonal heterogeneity, after which cells with a proliferative advantage are maintained and cells with unfavourable characteristics are eliminated, until a stable karyotype is reached (Merle et al., 2020; Delespaul et al., 2019; Dewhurst et al., 2014; Miroshnychenko et al., 2021; Laughney et al., 2015).

Data from human cancers suggests that tumours that have undergone WGD tend to settle at a near triploid state (Laughney et al., 2015; Zack et al., 2013; Bielski et al., 2018; Carter et al., 2012). Unexpectedly, the dominant subclones that emerged during this analysis were either near diploid or near tetraploid, though some near triploid cells were observed in all of the WGD clones and harvests. This difference may in part be due to the different environmental pressures faced by tumours *in vivo* compared to cell lines grown *in vitro*. This may also be related to the stability of HCT116 karyotype, as other groups have reported lower rates of ploidy reduction in cell fusions derived from stable parental cells (Miroshnychenko et al., 2021). However, clones C-G9 and R-C-F10 appear to have undergone ploidy reduction, resulting in the emergence of a diploid subclone, which calls this theory into question. Long term culture of naturally occurring near tetraploid HCT116 cells was carried out by Dewhurst et al. (2014), who did not report reversion of near tetraploid cells to diploid, suggesting that this may be something specific to this study. The authors did, however, observe a consistent loss of genetic material over time, with several formerly near tetraploid clones approaching a triploid karyotype over a length of time comparable to our experiments (Dewhurst et al., 2014). It is possible that the near-tetraploid peak in the distribution of chromosome counts may shift towards a near triploid karyotype with more time in culture.

A possible explanation for the presence of a diploid population in these clones is that it arose from contamination with the diploid M-B10 clone during culture. While I think that this is unlikely due to good aseptic technique being observed,

this cannot be ruled out. As no distinguishing alterations reached fixation in these clones, they cannot be differentiated based on karyotype, with the possible exception of potential loss of heterozygosity of known HCT116 variants. Ordinarily, der(10), t16/8, and t18/17 are only found on one chromosome copy each, accompanied by an unaffected copy of chromosomes 10, 16, and 18. If the diploid population that was seen originated from a reductive mitosis occurring after WGD, and chromosomes were randomly split between two cells, one might observe loss of heterozygosity if two of the same marker chromosomes or two non-marker chromosomes were to segregate together, which would confirm the WGD origin of the clone. At harvest 3 I observed five cells that contained two full copies of chromosome 10 each, as opposed to one full copy and one der(10). One of these five cells had also lost the t16/8 marker chromosome and was homozygous for two full copies of chromosomes 16 and 8. The absence of der(10) in and of itself does not confirm the origin of the diploid population, as this was not present in all of the C-F10 cells previously surveyed, nor was it present in all of the M-B10 cells surveyed. The absence of the t16/8 translocation represents a more interesting finding, as it was present in every single M-B10 cell surveyed. However, this loss of heterozygosity was only observed in a single cell, and is not indicative of the entire diploid population in C-G9. Thus, the origin of this diploid population remains a mystery. If the diploid population did indeed originate from a ploidy reduction event, the persistence of the HCT116 marker chromosomes may reflect a survival advantage conferred by their presence, as homozygosity at these sites may be disadvantageous for the cell.

As opposed to the contamination theory, a more parsimonious explanation in my opinion is that these cells underwent ploidy reduction which resulted in a near diploid subclone, that perhaps had a faster growth rate than the remaining near tetraploid cells. Zhou et al. (2015) have described the occurrence of multipolar mitoses following cell fusion, which leads to unequal distribution of DNA between daughter cells. The authors also describe near tetraploid and near triploid fused

clones becoming near diploid after 10 passages, setting a precedent for the proposed theory (Zhou et al., 2015).

It is interesting to note that reversion from WGD to diploid was only observed in cells that became WGD through a failure of cytokinesis in this study (Clones C-G9 and C-F3). However, ploidy reduction is not limited to fused cells in the literature, and is a common occurence in hepatocytes, which periodically undergo successive rounds of polyploidisation followed by ploidy reduction (Duncan et al., 2010).

Zhou et al. (2015) report diploid or near diploid chromosome numbers for a number of cells in very early passages after FACS sorting. The authors present these cells as fusion derived diploid clones (Zhou et al., 2015). However, an alternative explanation is that the authors sorted doublets, which was acknowledged as a frequent occurrence by Miroshnychenko et al. (2021), who validated their FACS sorted fusions using alternative methods. While selecting clones for follow up I ensured that clones derived from cell fusion and Cytochalasin B treatment contained a near tetraploid complement of chromosomes, and discarded clones which appeared to be diploid, based on chromosome counting. It remains possible that I excluded clones that underwent cell fusion followed by rapid ploidy reduction in very early cell divisions from my analysis. If ploidy reduction immediately, or very soon, after cell fusion is a common event, this suggests that cell fusion may be a more widespread phenomenon than is currently appreciated, and further complicates detection of cell fusion *in vivo*.

In addition to differences in ploidy over time, I also observed differences in the pattern of alterations identified in the different clones. Interestingly, the two samples with the most alterations were C-F10, and V-F4, which appeared to follow a similar trajectory in terms of total chromosome number. Alterations in C-F10 primarily occurred on chromosomes 1 and 2, and less so on chromosome 4. In contrast, the majority of alterations affecting V-F4 occurred on chromosome 8. It is possible that the increase in alterations affecting chromosome 8 may be due to the activa-

tion of a common fragile site (CFS). Chromosome 8 harbours an olfactory receptor gene cluster that has previously been associated with rearrangements occurring on chromosome 8 (Giglio et al., 2001).

Previous studies of cell fusion and cloning of near tetraploid cells have reported low survival and cloning efficiency for these cell types. Dewhurst et al. (2014) report 6% cloning efficiency for HCT116 based near tetraploid cells, compared to 63% for diploid cells. Several papers have cited a figure of up to 1% survival of fused cells (Miller et al. (1988) and Wakeling et al. (1994) in Searles et al. (2018) and Duelli and Lazebnik (2007)). Similar figures (roughly 1.1-2.8% clonogenicity) were reported by Miroshnychenko et al. (2021) after selection of fused cells using dual antibiotic resistance. However, cloning efficiencies as high as 19% have also been reported for fused cells sorted into conditioned medium, suggesting that the mechanism of selection and treatment of cells post-sort may play an important role in the survival of fused cells (Zhou et al., 2015). Though it is worth bearing in mind that some of the cells sorted by Zhou et al. (2015) may have constituted doublets, which have been estimated to represent 70-80% of cells sorted based on the presence of two fluorescent markers by Miroshnychenko et al. (2021), suggesting that the percentage of viable fusions achieved may be closer to 3.8-5.7% of cells.

In my experiments, I routinely sorted single cells into a single 96 well plate for each condition. While this was sufficient for isolation of diploid cells from co-culture, and WGD cells generated through a failure of cytokinesis, this is far too few cells to sort with the aim of isolating rare surviving cell fusions. Were further cell fusions to be conducted, I would focus my efforts on scaling up isolation of fused clones, as well as working to increase viability of cells post-fusion.

Chapter 6

Conclusion

6.1 Results Summary

6.1.1 Chapter 2

In Chapter 2 my aim was to investigate the relationship between the APOBEC mutational signature and genomic instability in whole genome sequencing data from the PCAWG consortium, and whether the APOBEC3 mutational signature could be linked to viral infection. I found that APOBEC3 mutations were associated with a higher mutation burden, in addition to higher numbers of INDELs, structural variants, copy number segments, ID6 INDELs, ID8 INDELs, and a higher percentage of the genome altered (PGA), suggesting a relationship between APOBEC3 mutations and increasing genomic instability. The association between APOBEC3 mutations and increased genomic instability remained when I accounted for the effects of age, TP53 mutations, and tumour type, and when I removed mutations associated with Kataegis. I also found that the presence of APOBEC3 mutations was associated with worse survival outcomes.

Further, I investigated whether the presence of APOBEC3 mutations could be linked to the presence of viral DNA in these tumours. I found that presence of single stranded DNA viruses was associated with increased numbers of APOBEC3

mutations; however, I was not able to investigate this relationship further due to small sample numbers. I also observed a negative association between the number of APOBEC3 mutations and the presence of Orthohepadnavirus DNA, which I believe to be an artefact of inherently low numbers of APOBEC mutations in liver cancers.

6.1.2 Chapter 3

In Chapter 3 I investigated whether viral infection is associated with increased A3A levels and genomic instability experimentally. I observed increased nuclear A3A staining and increased nuclear γH2AX staining after infection with pseudoviral particles. RNAseq analysis of infected cells unexpectedly revealed a decrease in the expression of several genes involved in DNA repair and the DNA damage response at the mRNA level 24 hours after infection. Further to this, no A3A expression was detected following infection.

To address whether A3A overexpression or viral infection resulted in similar patterns of genomic instability as those observed in whole genome sequencing data in Chapter 2, I whole genome sequenced virally infected cells, cells transfected with an A3A overexpression plasmid, as well as suitable controls. I did not see a significant difference in the number of mutations, INDELs, or structural variants in these cells compared to control. The relationship between A3A expression, viral infection, and genome stability was not clear cut in this analysis, but I believe this may be due to the nature of my experiments, as discussed in Sections 3 and 6.2.1.3.

6.1.3 Chapter 4

In Chapter 4 I developed a set of measures designed to identify tumours arising from fusion between a normal cell and a tumour cell based on whole genome sequencing data. These measures were then applied to 2,658 primary tumours in the PCAWG data set and 3,679 metastatic tumours from the HMF data set. I identified 16 samples in the PCAWG data set and 30 samples in the HMF data set that I believe

may have arisen as a result of a cell fusion event. Several tumours that were identified had additional characteristics suggesting that they may have undergone a cell fusion event or that they may have originated from a different cell type. A metastatic melanoma sample in the PCAWG data set emerged as a particularly interesting sample with several characteristics that suggest it may have arisen from cell fusion, and was the strongest candidate for further follow up.

6.1.4 Chapter 5

In Chapter 5 I presented a limited, analysis of karyotype evolution for several whole genome duplicated clones derived from a failure of cytokinesis, and a single virally fused clone. The scope of these experiments was limited by the number of viable clones that I was able to isolate.

My results were consistent with literature reporting increased heterogeneity in WGD clones. Over three months of culture, two of the WGD clones reverted to a near diploid karyotype, whereas the other two WGD clones remained near tetraploid but had more frequent chromosomal alterations.

Further work is required to address the aims of this chapter, which were to compare the effects of virus induced cell fusion, spontaneous cell fusion, and WGD by failure of cytokinesis on the genome.

6.2 Future Work

6.2.1 APOBECs, Viruses, and Genomic Instability

6.2.1.1 Viral Infection and APOBEC3-Induced Genomic Instability in Cancer Genomes

One of the questions that I had hoped to address in Chapter 2 was whether I could identify a relationship between evidence of prior viral infection in a tumour, and the

presence of SBS2 and SBS13 mutations. Due to the nature of the PCAWG data set, which was not created with the aim of studying relationships between viral infection and cancer, and therefore contained very few tumours with viral DNA, I was not able to answer this question. As part of my future work, I would like to study this relationship further in either a larger data set, or in a set of patients known to have been exposed to a virus. Gillison et al. (2019) have previously described an increase in SBS2 and SBS13 in HPV infected oral squamous cell carcinomas (OSCCs) compared to uninfected OSCCs, and correlated this with an overall increase in mutation burden. I think it would be interesting to determine if the same trend is seen with other viruses, which have not been as well characterised with regards to their involvement in cancer. However, I am particularly interested in viruses that have not previously been linked with cancer. Perhaps with the advent of large scale sequencing projects such as the 100,000 genomes project, enough tumours with incidental virus exposure will be identified to allow for a pan-virus pan-cancer analysis of the relationship between viral infection and APOBEC3 mutations in cancer.

6.2.1.2 APOBEC3 as a Causative Agent of Genomic Instability

As part of my future work, I would also like to further characterise the relationship between APOBEC3 mutations and genomic instability in whole genome sequencing data. Whether APOBEC3 is responsible for causing double strand breaks, which result in structural variants and INDELs when repaired, or whether APOBEC3 is reactionary and acts on single stranded DNA generated following a double strand break is still open to debate. I intend to examine the sequence context at which SVs and INDELs occur in tumours carrying SBS2 and SBS13 mutations, to investigate whether their break ends coincide with regions of the reference genome containing the 5'-T$\underline{\text{C}}$ sequence motif favoured by A3A and A3B (Bishop et al., 2004). I am particularly interested in sites at which preferred A3A and A3B sequence motifs are

closely apposed on opposing strands, or present at high density. Current estimation of INDEL signatures is based on the presence of specific numbers of homopolymers and sequence repeats and does not take the exact nucleotide context at which the INDEL occurs into account (Alexandrov et al., 2020). I think this analysis may be crucial in determining whether the alterations that are seen in whole genome sequencing data can be attributed to the activity of APOBEC3 enzymes.

6.2.1.3 A3A and DNA Damage in Cell Lines

My lab-based investigation of the relationship between A3A, viral infection, and genome stability also raised a number of unanswered questions that I would like to follow up. To begin with, I would like to analyse protein data using mass spectrometry from an experiment analogous to the one in which RNAseq data was extracted, to determine if we see any alterations in A3A content or post-translational modification of DNA damage response factors in response to infection. I hope that analysing the protein content may help bridge the gap between the results of my immunofluorescent staining and my RNAseq experiments.

As I discussed in Section 3.5, I think that further experiments investigating the effects of A3A expression and viral infection on genome stability might benefit from being modelled after experiments described by Akre et al. (2016), in which a doxycycline inducible A3B expression system was built, and Petljak et al. (2019), in which cell lines were clonally selected and grown for extended periods of time. In the context of my experiments, this would involve creating an inducible A3A expression system on the one hand, and exposing cells to sequential rounds of A3A expression, followed by clonal isolation, expansion, and sequencing. In parallel, un-induced cells could be exposed to sequential rounds of viral infection, to mimic chronic infection, and subsequently clonally isolated, expanded, and sequenced. This would allow the effect of ongoing mutational processes to be detected, and should improve the signal to noise ratio between passenger mutations and mutations caused by chemically

induced A3A expression and virally induced processes. I think these experiments would also benefit from the inclusion of a larger number of samples.

6.2.2 Cell Fusion in WGS Data

6.2.2.1 Validation of Tumour Samples Identified as Potential Cell Fusions

In Chapter 4 I highlighted several samples that were identified as potential cell fusions based on characteristics identified in whole genome sequencing data. A natural next step would be to investigate these samples in more detail to determine whether they have other characteristics that might support or disprove that a cell fusion event occurred. Previous characterisation of cell fusion in human tumours has relied on the detection of donor DNA or PCC in patient derived material (see Sections 1.3.2 and 1.3.1.2). Neither of these approaches is likely to be informative when trying to identify a tumour in which cell fusion was an initiating event or an early event, as PCC is only visible immediately after cell fusion, and detection of donor material relies on the tumours having arisen in patients receiving sex mismatched transplants, which is not the case in my data sets. Instead I propose a review of available tissue specimens by trained pathologists, to determine if the selected tumours have any unusual or unexpected features. The rationale for this approach is the reported phenotypic heterogeneity observed in cells derived from cell fusion (Yin et al., 2020; Wang et al., 2019). Both Yin et al. (2020) and Wang et al. (2019) report the emergence of clones that resemble the parental cells, but also observed clones with unique cell shapes and different growth rates. I expect that cell fusion may give rise to hybrids with distinct properties, and may appear different from known cancer subtypes. Unexpected histological features, such as unusual cell shape and arrangement, have previously been described in animal models in which cell fusion derived hybrid cells were transplanted, providing a precedent for this approach (Duelli et al., 2007).

Another approach would be to stain for the cell type specific proteins of suspected fusion partners. Expression of the macrophage specific CD163 has been suggested as a potential marker of macrophage-breast cancer fusion and would be an interesting protein to stain for in remaining tumour material from the breast cancers that were identified (Shabo et al., 2015). It would also be of interest to stain the melanoma sample that I identified for CD163, as macrophage-melanoma hybrids have previously been described (Lazova et al., 2013; Clawson et al., 2015).

Finally, sequencing samples from other studies in which cell fusion has been confirmed or identified, such as the studies outlined in Section 1.3.2, or even from mouse studies of tumours derived from cell fusion, such as those carried out by Duelli et al. (2007), would be an interesting approach for validating the methods that I have proposed and perhaps providing more insight into how these cells might be identified.

6.2.2.2 Validation of Methods Used to Identify Potential Cell Fusions in WGS Data

The methods that were designed to identify tumours derived from the fusion of a normal cell with a tumour cell have not yet been validated experimentally. I propose two approaches to address this. First, to determine whether the measure of variable mutation density has detected the existence of segments of low mutation density, high mutation density, and some combination of the two, I would like to sequence some of the samples that were identified using PacBio SMRT long read sequencing. Long read sequencing allows for direct haplotyping of samples and would clarify whether four haplotypes are present, two of which have a low number of mutations, and two of which have a high number of mutations. Further to this, haplotyping would also allow us to determine whether different mutational signatures are observed on different haplotypes.

As long read sequencing is a costly venture, and further sequencing of sample

material will not be possible for all samples due to limitations of consent, particularly for samples hosted by the TCGA, I propose statistical phasing of available data as a starting point. Several algorithms are available for phasing of mutations identified in WGS data (Choi et al., 2018). This is an exciting approach, but the results may be clouded by potential recombination between segments originating from different cells. Work by others has suggested that cell fusion may function as a mechanism for parasexual recombination in cancer cells, allowing for recombination between chromosomes derived from different parental cells (Miroshnychenko et al., 2021).

The second approach that I propose for validating these methods for the detection of cell fusion between a normal cell and a cancer cell is to sequence experimentally fused cells, similar to the ones I tried to generate in Chapter 5. In order to replicate the scenario of a normal cell fusing with a tumour cell, I would seek to do this experiment using an untransformed cell line, such as RPE-1 cells, and perhaps an ovarian cancer or a breast cancer cell line, as these were the most commonly identified tumour types in my analysis. Another interesting approach would be to use THP-1 cells as the normal cell fusion partner. THP-1 cells are a monocytic cell line that can be differentiated into macrophages. Given the fusogenic nature of macrophages, and speculation that macrophages may act as fusion partners for tumour cells, this would be a very interesting avenue to explore.

6.2.2.3 Expanding the Analysis to Larger Data Sets

In my analysis of the PCAWG and HMF data sets I found that only very few tumours were identified as potentially resulting from a cell fusion event. While this was expected due to the rarity of cell fusion and the low survival of fused cells, it severely limited the statistical power for comparing these tumours to their WGD counterparts. To address this, I would like to expand my analysis to larger data sets, such as the 100,000 Genomes Project data set curated by Genomics England. The 100,000 Genomes Project plans to whole genome sequence tumours from 25,000

individuals with common cancers (Peplow, 2016).

6.2.3 Cell Fusion Experiments

6.2.3.1 Further Fusion Experiments and Sequencing

In Chapter 5 I analysed the karyotype of clones derived from cell fusion, failure of cytokinesis, and co-culture over several months. My analysis was limited by the number of whole genome duplicated and experimentally fused clones that I was able to obtain. This limited the analysis that I was able to carry out, and precluded me from teasing out the different effects that whole genome duplication, cell fusion, and virus induced cell fusion have on the cancer genome. I still believe that this is an interesting question to investigate, and this has not yet been addressed by other groups working on cell fusion. Building on my previous cell fusion work, I would focus future efforts on scaling up isolation of fused cells, and work on increasing the viability of isolated cells, in order to be able to isolate enough viable clones to carry out quantitative characterisation of the differences between conditions. In addition, expanding this analysis to involve combinations of different cell types could inform my approach to identifying tumours in which cell fusion has occurred.

6.3 Concluding Remarks

The idea that cell fusion accompanied by aneuploidy might contribute to cancer initiation and progression was first proposed over a century ago by German gynaecologist and anthropologist Otto Aichel (Aichel, 1911). It is not known how Aichel arrived at this theory (Pawelek et al., 2020), but since its proposal a myriad studies have investigated the occurrence of cell fusion both *in vitro* and *in vivo*, and evidence for the occurrence of cell fusion in human cancers is mounting. However, many questions remain unanswered: how large is the contribution of cell fusion to cancer initiation and progression? How can we identify tumours originating from

cell fusion events? Are the consequences of cell fusion different from those of whole genome duplication by other means, and is there a difference between the effects of spontaneous cell fusion and virus induced cell fusion?

In this thesis I have touched on a number of these topics, and while I have no definitive answers to the questions proposed, I believe this work is an important contribution to cell fusion research. This thesis represents the first foray into identifying cell fusion in human tumours based on whole genome sequencing data. In this work I have developed a number of methods to identify the products of cell fusion in WGS data, and I have also identified several tumours with evidence of cell fusion. Further analysis of these specimens will provide a valuable insight into the prevalence, consequences, and perhaps also the causes of cell fusion in cancer.

This thesis has also provided an important insight into the contribution of APOBEC3 enzymes to genomic instability in tumour genomes. Using a panel of novel and established measures of genomic instability, I demonstrated a relationship between the presence and number of APOBEC3 mutations in the genome and the different measures of genomic instability, across different tumour types. This is the first study of its kind to use whole genome sequencing data to investigate this relationship. Further, the relationship between APOBEC3 mutations and genomic instability held when other factors such as age, TP53 mutation status, and tumour type were accounted for, and when mutations attributed to APOBEC3 activity occurring at sites of DNA damage (kataegis) were excluded. While further study is required, these results support a model in which APOBEC3 activity plays an important causative role in the generation of genomic instability.

I hope that the results of this work will further our understanding of the different processes that contribute to cancer initiation and progression, and that it may inform approaches to cancer treatment in the future.

Appendix A

Appendix: List of publications

Iliopoulou Maro, Nolan Rory, Alvarez Luis, Watanabe Yasunori, Coomer Charles A, ***Jakobsdottir G Maria****, Bowden Thomas A, Padilla-Parra Sergi.* A dynamic three-step mechanism drives the HIV-1 pre-fusion reaction. Nature structural molecular biology. 2018. 25, 9. 814–822.

Jakobsdottir G Maria*, Iliopoulou Maro, Nolan Rory, Alvarez Luis, Compton Alex A, Padilla-Parra Sergi.* On the whereabouts of HIV-1 cellular entry and its fusion ports. Trends in molecular medicine. 2017. 23, 10. 932–944.

Nolan Rory, Alvarez Luis AJ, Elegheert Jonathan, Iliopoulou Maro, ***Maria Jakobsdottir G****, Rodriguez-Mu noz Marina, Radu Aricescu A, Padilla-Parra Sergi.* nandb—number and brightness in R with a novel automatic detrending algorithm. Bioinformatics. 2017.

Appendix B

Appendix: Materials and Methods

B.1 Processing of samples in the HMF data set

Raw data was processed by the HMF using their in-house analysis pipeline (Hartwig Medical Foundation, 2020).

In order to be able to analyse the HMF data in the same way that the PCAWG pilot analysis was carried out, the Purity, Copy Number, and Somatic Variant VCF files provided by the HMF were converted to a series of pseudo- or PCAWG-like files, described in sections B.1.1-B.1.5. Files had to be in very specific formats to be compatible with the `runGetDirichletProcessInfo` function from version 1.0.8 of the `dpclust3p` R package (Dentro and Wedge, 2020), which was used to calculate the multiplicity of mutations.

B.1.1 Pseudo-Battenberg files

Because our analysis focuses exclusively on segments that are clonal within the tumour, the first step in this process was to identify clonal segments in the HMF Copy Number files. The HMF defines diploid regions as regions with a minor and major allele copy number of $\pm$ 0.2. We decided to define clonal segments in the HMF data as segments with minor and major alleles within $\pm$ 0.2 of a whole number.

To create copy number files similar to the files created by Battenberg, clonal segments were selected as described above, and assigned a Battenberg *frac1* value of 1. For these segments, the *minorAllelePloidy* and *majorAllelePloidy* values provided by the HMF were rounded to their nearest whole number, and assigned to the Battenberg values of *nMin1_A* and *nMaj1_A*, respectively. The *chromosome*, *start*, and *end* values provided by the HMF were used as the Battenberg values of *chr*, *startpos*, and *endpos*.

The Battenberg values of *BAF*, *pval*, *LogR*, and *ntot* were assigned 0, as they are not used in this analysis, and *nMaj2_A*, *nMin2_A*, and *frac2_A* were assigned NA for the same reason.

B.1.2 Pseudo-Locus files

To create a pseudo-locus file, the *CHROM*, *POS*, *REF*, and *ALT* values from the HMF Somatic Variant VCF files were extracted and renamed as *Chromosome*, *Position*, *Ref*, and *Alt*.

B.1.3 Pseudo-Allele Frequency files

Information contained in the HMF Somatic Variant VCF files also formed the basis of the pseudo-allele frequency files. The *Chromosome* and *Position* fields were carried over directly. The *AD*, or "Allelic depths for the ref and alt alleles in the order listed", value was then extracted from the final column, and used as the number of reads corresponsind to the reference and alternative alleles listed.

B.1.4 Pseudo-Cellularity files

Pseudo-cellularity files, providing informationo n the purity of the tumour were created by extracting the tumour purity from the HMF Purity file and assigning this to the *[rho,FRAC_GENOME]* row and column. All other values were left blank.

B.1.5 Gender Information

Information relating to the patient's gender was extracted from the Purity file.

B.1.6 Filtering Gaps in the Reference Genome

Copy number data generated by the HMF contains copy number information for some regions of the genome that are not covered by the reference genome. For example, the PURPLE copy number caller will extend copy number information from the long arm of certain chromosomes to the short arm, in cases where no copy number information is available for the short arm, and on chromosomes 13,14,15,21, or 22 (Hartwig Medical Foundation, 2020). This results in regions which artificially have zero mutations, which skews our analysis of the coefficient of variation of the mutation density. For this reason, we filter out regions in the copy number files for which the reference genome contains gaps. A list of gaps in the GrCh37.p13 was downloaded from the Genome Reference Consortium at the NCBI on May 14th 2020. This filtration step leads to the removal of segments that fit within a gap region, truncation of segments that partially overlap a gap region, and fragmentation of segments that fully overlap a reference genome gap.

B.1.7 Merging Adjacent Segments With the Same Copy Number

As a feature of the Battenberg copy number caller, adjacent regions with the same copy number profile are merged. To replicate this behaviour, after filtering out regions of the reference genome that are not mapped, we merge each segment to the nearest segment on the same chromosome, provided they have the same copy number arrangement. However, for calculations of mutation density we calculate the effective segment size as the sum of the size of the segments that make up the merged segment, to prevent potential bias introduced by merging over unmapped

regions.

Appendix C

Appendix: Results

C.1 Differential Expression Analysis

C.1.1 All Genes

Table C.1: Differentially expressed genes 6h after infection with an adjusted p-value <0.05 and an absolute log2 fold change >1. log2FC represents the log2 fold change in expression compared to 0h post infection, lfcSE refers to the standard error of the log2 fold change, stat represents the Wald test statistic, p-values represent raw p-values from the Wald test, and and p-adjusted refers to the multiple testing adjusted Wald test p-value.

	baseMean	log2FChange	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000205420	69.3	-1.06	0.182	-5.83	5.41e-09	1.74e-06	KRT6A

Table C.2: Differentially expressed genes 12h after infection with an adjusted p-value <0.05 and an absolute log2 fold change >1. log2FC represents the log2 fold change in expression compared to 0h post infection, lfcSE refers to the standard error of the log2 fold change, stat represents the Wald test statistic, p-values represent raw p-values from the Wald test, and and p-adjusted refers to the multiple testing adjusted Wald test p-value. Genes are listed from highest significant p-value to lowest.

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000204388	8.66e+03	-1.06	0.0587	-18.1	1.88e-73	2.4e-69	HSPA1B
ENSG00000205420	69.3	-2.19	0.22	-9.94	2.73e-23	4.35e-20	KRT6A

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000172530	220	-1.4	0.157	-8.89	6.08e-19	6.46e-16	BANP
ENSG00000120738	408	1.05	0.122	8.59	8.65e-18	7.37e-15	EGR1
ENSG00000168209	361	1.01	0.119	8.54	1.33e-17	9.96e-15	DDIT4
ENSG00000149591	141	-1.23	0.151	-8.17	3.19e-16	1.77e-13	TAGLN
ENSG00000185950	180	-1.04	0.13	-7.99	1.3e-15	6.38e-13	IRS2
ENSG00000228532	51.9	-2.52	0.359	-7.02	2.21e-12	7.25e-10	SUMO2P21
ENSG00000059378	138	-1.03	0.156	-6.62	3.67e-11	9.36e-09	PARP12
ENSG00000171401	18.5	-2.58	0.418	-6.16	7.31e-10	1.51e-07	KRT13
ENSG00000135480	147	-1.02	0.171	-5.97	2.38e-09	3.88e-07	KRT7
ENSG00000188818	91.1	-1.16	0.21	-5.51	3.69e-08	4.1e-06	ZDHHC11
ENSG00000173546	76.8	1.18	0.214	5.5	3.75e-08	4.12e-06	CSPG4
ENSG00000264281	114	-2.39	0.442	-5.42	6.09e-08	6.19e-06	RPL41
ENSG00000129646	60.6	-1.14	0.227	-5.01	5.42e-07	3.95e-05	QRICH2
ENSG00000204389	119	-1.19	0.243	-4.9	9.52e-07	6.2e-05	HSPA1A
ENSG00000164744	39.6	1.36	0.289	4.69	2.69e-06	0.000151	SUN3
ENSG00000129451	61.9	-1.08	0.232	-4.64	3.41e-06	0.000187	KLK10
ENSG00000255782	14.6	6.06	1.31	4.61	3.98e-06	0.000213	Z83840.1
ENSG00000248112	36.2	-1.28	0.284	-4.51	6.35e-06	0.000314	RP11-78C3.1
ENSG00000101265	47.5	1.2	0.27	4.44	8.89e-06	0.000412	RASSF2
ENSG00000183336	67.4	-9.43	2.16	-4.37	1.27e-05	0.000553	BOLA2
ENSG00000258388	49.6	-1.65	0.38	-4.34	1.44e-05	0.000607	PPT2-EGFL8
ENSG00000182685	52.9	-1.02	0.238	-4.28	1.91e-05	0.000764	BRICD5
ENSG00000196656	290	-3.03	0.718	-4.22	2.41e-05	0.000918	AC004057.1
ENSG00000196220	28	1.6	0.382	4.2	2.7e-05	0.001	SRGAP3
ENSG00000138336	37.6	1.25	0.3	4.17	3.03e-05	0.00109	TET1
ENSG00000234009	51.2	1.43	0.351	4.08	4.58e-05	0.00154	RPL5P34
ENSG00000254996	39.7	-1.86	0.457	-4.07	4.79e-05	0.00159	ANKHD1-EIF4EBP3
ENSG00000234231	38.3	1.51	0.372	4.06	4.85e-05	0.00161	LOC730268
ENSG00000257524	18.6	-2.08	0.515	-4.04	5.41e-05	0.00175	ST6GALNAC6
ENSG00000251705	128	3.12	0.777	4.01	6.02e-05	0.00186	RNA5-8SP6
ENSG00000269693	27.1	1.88	0.475	3.96	7.52e-05	0.00225	CTD-2192J16.20
ENSG00000128849	32.7	1.48	0.378	3.92	8.72e-05	0.00246	CGNL1
ENSG00000249839	10	-5.44	1.39	-3.91	9.23e-05	0.00257	AC011330.5
ENSG00000205426	24	-1.37	0.352	-3.9	9.82e-05	0.00267	KRT81
ENSG00000229994	13.8	2.03	0.538	3.76	0.000167	0.00398	RPL5P4
ENSG00000242687	20.1	-1.47	0.396	-3.72	0.000203	0.00463	AC004893.11
ENSG00000249936	30.7	-1.38	0.373	-3.71	0.000204	0.00465	RAC1P2
ENSG00000141622	19.8	1.64	0.444	3.69	0.000223	0.00498	RNF165

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000121207	27.6	1.36	0.371	3.67	0.000245	0.00531	LRAT
ENSG00000184898	24.3	1.25	0.341	3.66	0.000255	0.0055	RBM43
ENSG00000204314	36	1.1	0.302	3.65	0.00026	0.00558	PRRT1
ENSG00000106852	24.2	1.36	0.378	3.6	0.000317	0.00652	LHX6
ENSG00000267858	22.4	-1.29	0.359	-3.6	0.000317	0.00652	MZF1-AS1
ENSG00000198695	2.35e+03	1.19	0.332	3.57	0.000351	0.00702	ND6
ENSG00000254837	26.6	1.3	0.368	3.53	0.000415	0.00784	LOC100287896
ENSG00000111981	43.4	1.08	0.306	3.52	0.000428	0.00804	ULBP1
ENSG00000226084	22.4	1.57	0.448	3.51	0.000452	0.00836	RP4-706A16.3
ENSG00000124507	17.7	1.61	0.46	3.5	0.000466	0.00857	PACSIN1
ENSG00000213860	69.4	-1.01	0.292	-3.47	0.000513	0.00916	RPL21P75
ENSG00000234664	46.9	1.04	0.302	3.46	0.000532	0.00939	HMGN2P5
ENSG00000108306	30.9	1.17	0.339	3.46	0.000545	0.00956	FBXL20
ENSG00000125650	29.8	-1.09	0.317	-3.46	0.000549	0.0096	PSPN
ENSG00000143847	14.9	1.94	0.561	3.46	0.000549	0.0096	PPFIA4
ENSG00000272578	41.1	-1.23	0.357	-3.45	0.000558	0.0097	AP000347.2
ENSG00000113327	17	1.52	0.447	3.4	0.000672	0.0113	GABRG2
ENSG00000198754	27.6	1.08	0.321	3.37	0.000759	0.0125	OXCT2
ENSG00000154319	30.5	1.03	0.307	3.36	0.000769	0.0126	FAM167A
ENSG00000253741	19.5	1.26	0.378	3.34	0.000823	0.0132	LNCOC1
ENSG00000139278	12	-1.67	0.499	-3.34	0.000826	0.0132	GLIPR1
ENSG00000152076	11.3	1.91	0.573	3.33	0.000875	0.0137	CCDC74B
ENSG00000144785	12.1	2.72	0.824	3.29	0.00099	0.0148	RP11-977G19.10
ENSG00000203836	12.6	2.01	0.615	3.27	0.00106	0.0154	NBPF24
ENSG00000211454	11	1.78	0.549	3.24	0.00122	0.0169	AKR7L
ENSG00000155016	21.7	1.19	0.376	3.17	0.00154	0.0201	CYP2U1
ENSG00000171954	30.1	1.02	0.323	3.16	0.00158	0.0204	CYP4F22
ENSG00000198959	31.4	-1.05	0.334	-3.15	0.00163	0.0209	TGM2
ENSG00000119698	24.9	1.18	0.379	3.13	0.00177	0.0219	PPP4R4
ENSG00000248334	12.7	1.69	0.542	3.11	0.00187	0.0229	WHAMMP2
ENSG00000213442	41.6	-1.92	0.624	-3.09	0.00203	0.0246	RPL18AP3
ENSG00000269378	19.5	-1.91	0.627	-3.06	0.00224	0.0264	ITGB1P1
ENSG00000128266	28.1	1.1	0.364	3.03	0.00246	0.0282	GNAZ
ENSG00000188760	16.7	-1.26	0.416	-3.03	0.00248	0.0283	TMEM198
ENSG00000147174	15.1	-1.29	0.428	-3.02	0.0025	0.0284	GCNA
ENSG00000230715	19	1.32	0.438	3.01	0.00257	0.0291	RP11-274B21.4
ENSG00000173295	20.1	1.14	0.379	3.01	0.00259	0.0292	FAM86B3P
ENSG00000266402	22.4	1.41	0.469	3.01	0.00263	0.0295	SNHG25

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000170579	9.95	2.27	0.758	2.99	0.00279	0.0308	DLGAP1
ENSG00000228253	7.21e+03	1.43	0.483	2.96	0.00308	0.0328	ATP8
ENSG00000157021	25	1.22	0.42	2.9	0.00376	0.038	FAM92A1P1
ENSG00000145358	17.4	1.21	0.422	2.87	0.00412	0.0404	DDIT4L
ENSG00000116983	27.4	1	0.35	2.86	0.00425	0.041	HPCAL4
ENSG00000255198	56	1	0.351	2.86	0.00427	0.0412	SNHG9
ENSG00000149654	10.9	1.52	0.539	2.83	0.00469	0.044	CDH22
ENSG00000224114	139	1.76	0.624	2.83	0.00468	0.044	RP11-343H5.4
ENSG00000266972	15.4	1.44	0.516	2.8	0.00508	0.0466	LRRC37A9P
ENSG00000168913	14.2	1.46	0.523	2.8	0.0051	0.0467	ENHO
ENSG00000183208	18	-1.09	0.392	-2.79	0.00524	0.0475	GDPGP1
ENSG00000250988	22	1.11	0.399	2.79	0.00532	0.048	SNHG21
ENSG00000272617	21.8	5.19	1.87	2.78	0.00547	0.0488	RP11-343C2.12
ENSG00000203867	19	1.12	0.405	2.77	0.00555	0.0493	RBM20

Table C.3: Differentially expressed genes 24h after infection with an adjusted p-value <0.05 and an absolute log2 fold change >1. log2FC represents the log2 fold change in expression compared to 0h post infection, lfcSE refers to the standard error of the log2 fold change, stat represents the Wald test statistic, p-values represent raw p-values from the Wald test, and and p-adjusted refers to the multiple testing adjusted Wald test p-value. Genes are listed from highest significant p-value to lowest.

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000110092	1.05e+04	-1.14	0.0478	-23.9	5.34e-126	7.52e-122	CCND1
ENSG00000168209	361	2.22	0.108	20.6	3.41e-94	1.6e-90	DDIT4
ENSG00000070669	907	1.1	0.0597	18.4	1.56e-75	5.49e-72	ASNS
ENSG00000107984	4.35e+03	-1	0.0575	-17.4	7.86e-68	1.84e-64	DKK1
ENSG00000101335	624	-1.41	0.0849	-16.6	3.43e-62	6.89e-59	MYL9
ENSG00000205420	69.3	-4.82	0.312	-15.5	7.18e-54	1.01e-50	KRT6A
ENSG00000186480	332	1.5	0.101	14.8	8.8e-50	1.13e-46	INSIG1
ENSG00000117525	1.57e+03	-1.52	0.105	-14.5	1.19e-47	1.29e-44	F3
ENSG00000149591	141	-2.21	0.152	-14.5	1.52e-47	1.53e-44	TAGLN
ENSG00000019186	673	1.08	0.0761	14.2	9.76e-46	8.59e-43	CYP24A1
ENSG00000164292	733	1.09	0.0783	13.9	8.21e-44	6.8e-41	RHOBTB3
ENSG00000171435	274	1.45	0.109	13.3	1.29e-40	9.08e-38	KSR2
ENSG00000058085	390	-1.1	0.0884	-12.4	1.7e-35	9.19e-33	LAMC2

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000151012	462	1.42	0.12	11.9	1.47e-32	6.09e-30	SLC7A11
ENSG00000142871	283	-1.35	0.117	-11.6	6.69e-31	2.55e-28	CCN1
ENSG00000162496	135	1.73	0.151	11.5	2.21e-30	8.21e-28	DHRS3
ENSG00000163435	337	1.05	0.0964	10.9	1.76e-27	5.63e-25	ELF3
ENSG00000158062	275	-1.15	0.107	-10.7	6.97e-27	2.13e-24	UBXN11
ENSG00000198919	240	1.34	0.128	10.5	6.25e-26	1.6e-23	DZIP3
ENSG00000185950	180	-1.23	0.121	-10.2	2.58e-24	5.95e-22	IRS2
ENSG00000137801	347	-1.09	0.109	-10.1	9.07e-24	1.91e-21	THBS1
ENSG00000059145	328	-1.06	0.109	-9.81	1.06e-22	1.91e-20	UNKL
ENSG00000005020	163	1.29	0.135	9.62	6.57e-22	1.09e-19	SKAP2
ENSG00000164684	132	1.46	0.154	9.44	3.86e-21	5.79e-19	ZNF704
ENSG00000009950	220	1.12	0.122	9.16	5.19e-20	7.02e-18	MLXIPL
ENSG00000128965	54.1	2.34	0.258	9.08	1.12e-19	1.48e-17	CHAC1
ENSG00000136542	217	-1	0.112	-8.96	3.23e-19	4.13e-17	GALNT5
ENSG00000127946	292	1	0.115	8.73	2.52e-18	3.03e-16	HIP1
ENSG00000079215	228	1.03	0.119	8.65	5.03e-18	5.86e-16	SLC1A3
ENSG00000167767	178	-1.29	0.151	-8.59	8.94e-18	1.01e-15	KRT80
ENSG00000173546	76.8	1.6	0.199	8.02	1.04e-15	9.25e-14	CSPG4
ENSG00000171401	18.5	-3.34	0.417	-8.02	1.07e-15	9.38e-14	KRT13
ENSG00000100889	118	1.21	0.151	8.01	1.18e-15	1.03e-13	PCK2
ENSG00000006468	92.4	1.52	0.192	7.91	2.52e-15	2.14e-13	ETV1
ENSG00000135480	147	-1.27	0.161	-7.91	2.66e-15	2.24e-13	KRT7
ENSG00000235823	46.3	1.94	0.245	7.9	2.81e-15	2.33e-13	OLMALINC
ENSG00000111981	43.4	2.17	0.278	7.81	5.81e-15	4.68e-13	ULBP1
ENSG00000251562	4.02e+03	-1.24	0.167	-7.43	1.11e-13	7.25e-12	MALAT1
ENSG00000059378	138	-1.06	0.144	-7.36	1.86e-13	1.17e-11	PARP12
ENSG00000135365	115	1.22	0.171	7.13	9.92e-13	5.7e-11	PHF21A
ENSG00000212724	106	-1.21	0.171	-7.1	1.29e-12	7.28e-11	KRTAP2-3
ENSG00000144730	65.2	1.35	0.204	6.6	4.17e-11	1.82e-09	IL17RD
ENSG00000157851	86.4	1.11	0.169	6.58	4.68e-11	2.01e-09	DPYSL5
ENSG00000007402	53.4	1.41	0.217	6.51	7.77e-11	3.19e-09	CACNA2D2
ENSG00000101265	47.5	1.62	0.252	6.45	1.15e-10	4.55e-09	RASSF2
ENSG00000188290	306	-1.02	0.159	-6.43	1.3e-10	5.05e-09	HES4
ENSG00000168916	41.2	1.65	0.261	6.32	2.62e-10	9.72e-09	ZNF608
ENSG00000128266	28.1	2.08	0.331	6.27	3.66e-10	1.32e-08	GNAZ
ENSG00000138336	37.6	1.73	0.279	6.2	5.53e-10	1.95e-08	TET1
ENSG00000117266	50.1	1.44	0.235	6.1	1.03e-09	3.44e-08	CDK18
ENSG00000118523	25.1	-2.25	0.371	-6.07	1.24e-09	4.08e-08	CCN2

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000128849	32.7	2.14	0.352	6.07	1.29e-09	4.21e-08	CGNL1
ENSG00000196220	28	2.16	0.357	6.06	1.38e-09	4.49e-08	SRGAP3
ENSG00000156869	73.3	1.13	0.189	6.01	1.89e-09	5.99e-08	FRRS1
ENSG00000185880	62	1.15	0.196	5.85	5e-09	1.43e-07	TRIM69
ENSG00000150593	91.9	1.07	0.185	5.8	6.52e-09	1.81e-07	PDCD4
ENSG00000134013	72.5	1.12	0.194	5.79	6.84e-09	1.87e-07	LOXL2
ENSG00000213931	93.3	1	0.177	5.66	1.48e-08	3.76e-07	HBE1
ENSG00000095539	68.2	1.18	0.209	5.65	1.56e-08	3.94e-07	SEMA4G
ENSG00000128165	24.1	1.93	0.342	5.65	1.61e-08	4.06e-07	ADM2
ENSG00000109654	145	1	0.178	5.62	1.87e-08	4.63e-07	TRIM2
ENSG00000186594	79.8	-1.08	0.192	-5.62	1.91e-08	4.72e-07	MIR22HG
ENSG00000116574	54.9	1.29	0.23	5.61	2.06e-08	5.06e-07	RHOU
ENSG00000111266	74.5	1.07	0.192	5.58	2.38e-08	5.71e-07	DUSP16
ENSG00000255782	14.6	7.16	1.29	5.56	2.69e-08	6.38e-07	Z83840.1
ENSG00000132196	66.3	1.03	0.186	5.53	3.25e-08	7.51e-07	HSD17B7
ENSG00000154529	69.1	1.15	0.209	5.49	3.94e-08	8.91e-07	CNTNAP3B
ENSG00000177570	70.9	1.05	0.191	5.47	4.55e-08	1.01e-06	SAMD12
ENSG00000138764	24.2	2.01	0.369	5.44	5.43e-08	1.18e-06	CCNG2
ENSG00000228532	51.9	-1.66	0.308	-5.4	6.77e-08	1.43e-06	RP1-241P17.4
ENSG00000132846	131	1	0.187	5.36	8.15e-08	1.68e-06	ZBED3
ENSG00000117016	45.5	1.28	0.239	5.36	8.21e-08	1.69e-06	RIMS3
ENSG00000264281	114	-2.19	0.409	-5.34	9.08e-08	1.85e-06	ENSG00000264281
ENSG00000140044	47.7	1.26	0.239	5.28	1.33e-07	2.58e-06	JDP2
ENSG00000166833	28.5	1.59	0.302	5.26	1.46e-07	2.77e-06	NAV2
ENSG00000196656	290	-3.54	0.674	-5.26	1.46e-07	2.78e-06	AC004057.1
ENSG00000257524	18.6	-2.6	0.494	-5.25	1.49e-07	2.82e-06	ST6GALNAC6
ENSG00000186814	70.6	1.09	0.209	5.23	1.71e-07	3.19e-06	ZSCAN30
ENSG00000234231	38.3	1.82	0.35	5.19	2.05e-07	3.75e-06	LOC730268
ENSG00000165730	35.8	1.46	0.281	5.18	2.19e-07	3.96e-06	STOX1
ENSG00000108306	30.9	1.6	0.316	5.07	3.99e-07	6.78e-06	FBXL20
ENSG00000266972	15.4	2.4	0.474	5.05	4.31e-07	7.3e-06	LRRC37A9P
ENSG00000146350	65.9	1.07	0.213	5.01	5.41e-07	8.98e-06	TBC1D32
ENSG00000168280	63	1.04	0.208	4.99	6.08e-07	1e-05	KIF5C
ENSG00000182621	38.7	1.36	0.274	4.97	6.77e-07	1.11e-05	PLCB1
ENSG00000185838	53.3	1.09	0.219	4.96	7.16e-07	1.16e-05	GNB1L
ENSG00000126016	36.1	1.35	0.273	4.94	7.9e-07	1.26e-05	AMOT
ENSG00000099954	36.3	1.31	0.265	4.93	8.33e-07	1.33e-05	CECR2
ENSG00000124507	17.7	2.12	0.431	4.92	8.48e-07	1.35e-05	PACSIN1

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000261824	61.4	1.03	0.211	4.89	9.99e-07	1.55e-05	LINC00662
ENSG00000146278	37	1.29	0.265	4.89	1.01e-06	1.57e-05	PNRC1
ENSG00000121207	27.6	1.7	0.348	4.88	1.06e-06	1.63e-05	LRAT
ENSG00000141376	74.1	1.01	0.209	4.84	1.29e-06	1.96e-05	BCAS3
ENSG00000170500	13.5	2.61	0.549	4.76	1.89e-06	2.73e-05	LONRF2
ENSG00000205464	43.3	1.13	0.239	4.72	2.36e-06	3.32e-05	ATP6AP1L
ENSG00000116983	27.4	1.51	0.324	4.67	3.05e-06	4.15e-05	HPCAL4
ENSG00000109771	23.1	1.61	0.346	4.66	3.15e-06	4.28e-05	LRP2BP
ENSG00000254837	26.6	1.59	0.345	4.61	3.94e-06	5.17e-05	LOC100287896
ENSG00000176809	22.2	1.99	0.431	4.61	4.03e-06	5.28e-05	LRRC37A3
ENSG00000180573	26.9	1.68	0.364	4.6	4.27e-06	5.56e-05	H2AC6
ENSG00000247828	29.2	1.41	0.307	4.58	4.71e-06	6.07e-05	TMEM161B-AS1
ENSG00000196542	13.5	2.21	0.488	4.54	5.72e-06	7.15e-05	SPTSSB
ENSG00000105289	39	1.46	0.323	4.52	6.05e-06	7.5e-05	TJP3
ENSG00000017373	26.9	1.44	0.318	4.51	6.45e-06	7.9e-05	SRCIN1
ENSG00000205426	24	-1.45	0.323	-4.5	6.96e-06	8.44e-05	KRT81
ENSG00000144785	12.1	3.52	0.784	4.49	7.04e-06	8.53e-05	RP11-977G19.10
ENSG00000164744	39.6	1.23	0.276	4.45	8.76e-06	0.000103	SUN3
ENSG00000165388	47.8	1.1	0.25	4.42	9.89e-06	0.000113	ZNF488
ENSG00000261685	10.6	2.56	0.584	4.38	1.17e-05	0.000131	RP11-401P9.4
ENSG00000213918	36.7	1.11	0.255	4.37	1.27e-05	0.00014	DNASE1
ENSG00000115902	38.2	1.08	0.247	4.36	1.29e-05	0.000142	SLC1A4
ENSG00000248175	17.9	-1.58	0.363	-4.36	1.33e-05	0.000146	CTC-428G20.3
ENSG00000116852	42.5	1.09	0.25	4.35	1.35e-05	0.000148	KIF21B
ENSG00000170579	9.95	3.11	0.716	4.35	1.36e-05	0.000149	DLGAP1
ENSG00000135678	24.8	1.43	0.33	4.34	1.4e-05	0.000152	CPM
ENSG00000198929	31.6	1.26	0.29	4.34	1.42e-05	0.000155	NOS1AP
ENSG00000169330	46.3	1.04	0.239	4.33	1.46e-05	0.000158	MINAR1
ENSG00000142661	46	1.01	0.233	4.32	1.55e-05	0.000167	MYOM3
ENSG00000100167	73	1.04	0.24	4.32	1.56e-05	0.000168	SEPTIN3
ENSG00000134769	22.9	1.54	0.358	4.32	1.57e-05	0.000168	DTNA
ENSG00000248112	36.2	-1.09	0.254	-4.3	1.71e-05	0.000181	RP11-78C3.1
ENSG00000114796	33	1.17	0.274	4.27	1.96e-05	0.000203	KLHL24
ENSG00000115423	8.5	2.86	0.669	4.27	1.97e-05	0.000203	DNAH6
ENSG00000254996	39.7	-1.81	0.424	-4.27	1.99e-05	0.000205	ANKHD1-EIF4EBP3
ENSG00000118263	33.2	1.31	0.308	4.26	2.03e-05	0.000209	KLF7
ENSG00000178445	22.7	1.56	0.369	4.24	2.26e-05	0.00023	GLDC
ENSG00000179178	39.9	-1.03	0.244	-4.22	2.47e-05	0.000248	TMEM125

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000108813	43.2	1.02	0.242	4.2	2.66e-05	0.000264	DLX4
ENSG00000205246	58.6	-1.11	0.264	-4.2	2.69e-05	0.000266	RPSAP58
ENSG00000144583	16.3	-1.93	0.465	-4.16	3.16e-05	0.000308	MARCHF4
ENSG00000266714	40.5	-1.06	0.257	-4.14	3.45e-05	0.000332	MYO15B
ENSG00000249936	30.7	-1.42	0.343	-4.13	3.64e-05	0.000347	RAC1P2
ENSG00000196684	49.7	1.01	0.246	4.12	3.81e-05	0.00036	HSH2D
ENSG00000167103	27.4	1.34	0.328	4.09	4.31e-05	0.000399	PIP5KL1
ENSG00000236824	731	-1.63	0.402	-4.05	5.15e-05	0.000467	BCYRN1
ENSG00000143248	32.8	1.17	0.289	4.05	5.2e-05	0.00047	RGS5
ENSG00000138792	27.3	1.21	0.299	4.05	5.2e-05	0.00047	ENPEP
ENSG00000172986	22.5	1.41	0.349	4.04	5.35e-05	0.000482	GXYLT2
ENSG00000124374	36.7	1.13	0.279	4.04	5.36e-05	0.000482	PAIP2B
ENSG00000213260	6.68	-3.02	0.752	-4.02	5.76e-05	0.000515	YWHAZP5
ENSG00000259673	35.7	1.09	0.272	4.01	6.1e-05	0.00054	IQCH-AS1
ENSG00000049283	41.4	1.02	0.256	4	6.22e-05	0.000549	EPN3
ENSG00000178878	14	2.03	0.508	4	6.41e-05	0.000561	APOLD1
ENSG00000167964	30.9	1.14	0.288	3.96	7.38e-05	0.000634	RAB26
ENSG00000259328	41.1	1.14	0.288	3.95	7.93e-05	0.000675	RP11-152F13.7
ENSG00000175591	25.3	1.29	0.329	3.92	8.8e-05	0.000733	P2RY2
ENSG00000204682	24.3	1.28	0.329	3.89	0.0001	0.000816	MIR1915HG
ENSG00000188760	16.7	-1.51	0.389	-3.89	0.000102	0.000827	TMEM198
ENSG00000099251	11.3	1.87	0.481	3.88	0.000105	0.00085	HSD17B7P2
ENSG00000095637	15.5	1.76	0.453	3.87	0.000107	0.000865	SORBS1
ENSG00000183196	43.7	1.09	0.283	3.85	0.000118	0.000941	CHST6
ENSG00000243232	21.3	1.35	0.353	3.81	0.000137	0.00107	PCDHAC2
ENSG00000125864	35.4	1.14	0.298	3.81	0.000138	0.00107	BFSP1
ENSG00000138613	16.2	1.63	0.427	3.81	0.000142	0.00109	APH1B
ENSG00000182185	23.7	1.32	0.347	3.8	0.000146	0.00112	RAD51B
ENSG00000065618	40	-1.09	0.292	-3.75	0.000176	0.00132	COL17A1
ENSG00000172915	31.1	1.13	0.301	3.75	0.000177	0.00133	NBEA
ENSG00000259399	4.37	5.54	1.49	3.73	0.000192	0.00142	TGIF2-RAB5IF
ENSG00000105808	20	1.42	0.382	3.73	0.000195	0.00143	RASA4
ENSG00000063180	37.5	1.02	0.275	3.72	0.000197	0.00144	CA11
ENSG00000188211	28.5	1.15	0.308	3.72	0.000198	0.00145	NCR3LG1
ENSG00000068831	18.4	1.4	0.378	3.71	0.000208	0.00151	RASGRP2
ENSG00000130518	26.2	1.23	0.331	3.71	0.00021	0.00153	IQCN
ENSG00000114631	31.9	-1.05	0.284	-3.7	0.000213	0.00154	PODXL2
ENSG00000071242	23.6	1.4	0.381	3.68	0.000235	0.00167	RPS6KA2

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000112139	11.6	1.83	0.498	3.67	0.000242	0.00171	MDGA1
ENSG00000104967	17.5	1.46	0.4	3.66	0.00025	0.00176	NOVA2
ENSG00000113578	5.21	-3.04	0.832	-3.65	0.00026	0.00182	FGF1
ENSG00000234009	51.2	1.22	0.333	3.65	0.000261	0.00182	RPL5P34
ENSG00000174721	25.8	1.21	0.332	3.64	0.000268	0.00186	FGFBP3
ENSG00000105088	31.9	1.08	0.296	3.64	0.000276	0.00191	OLFM2
ENSG00000143847	14.9	1.93	0.536	3.61	0.000307	0.00209	PPFIA4
ENSG00000119698	24.9	1.29	0.357	3.61	0.00031	0.0021	PPP4R4
ENSG00000137103	11.8	1.97	0.547	3.6	0.000322	0.00217	TMEM8B
ENSG00000002746	16.4	1.63	0.455	3.59	0.000329	0.00221	HECW1
ENSG00000221995	22.2	1.37	0.382	3.59	0.000333	0.00224	TIAF1
ENSG00000143494	26.4	1.09	0.305	3.57	0.00036	0.00238	VASH2
ENSG00000145362	16.1	1.51	0.423	3.56	0.000368	0.00242	ANK2
ENSG00000115525	23.5	1.21	0.34	3.55	0.000384	0.00251	ST3GAL5
ENSG00000106714	34.8	1.14	0.321	3.55	0.000387	0.00253	CNTNAP3
ENSG00000185482	14	1.61	0.456	3.53	0.000418	0.0027	STAC3
ENSG00000155016	21.7	1.24	0.355	3.49	0.000487	0.00307	CYP2U1
ENSG00000153132	20	1.3	0.374	3.49	0.000491	0.00308	CLGN
ENSG00000005187	21.8	1.19	0.342	3.48	0.000494	0.0031	ACSM3
ENSG00000106852	24.2	1.25	0.361	3.48	0.000505	0.00316	LHX6
ENSG00000070182	24.8	1.19	0.341	3.47	0.000517	0.00322	SPTB
ENSG00000126882	25.4	1.11	0.323	3.45	0.000568	0.0035	FAM78A
ENSG00000197565	21.5	1.36	0.397	3.43	0.0006	0.00366	COL4A6
ENSG00000188177	12.4	1.58	0.461	3.42	0.000632	0.00382	ZC3H6
ENSG00000205269	17.9	1.33	0.392	3.4	0.000677	0.00404	TMEM170B
ENSG00000107249	23.3	-1.09	0.32	-3.4	0.000679	0.00405	GLIS3
ENSG00000019505	13.2	1.61	0.477	3.38	0.000734	0.00434	SYT13
ENSG00000259726	5.45	3.3	0.978	3.37	0.000738	0.00435	CSPG4P11
ENSG00000145358	17.4	1.34	0.398	3.37	0.000754	0.00444	DDIT4L
ENSG00000145703	4.82	4.02	1.2	3.37	0.000759	0.00446	IQGAP2
ENSG00000228253	7.21e+03	1.52	0.452	3.35	0.000795	0.00464	ATP8
ENSG00000091137	8.06	2.29	0.684	3.35	0.000813	0.00473	SLC26A4
ENSG00000141622	19.8	1.42	0.426	3.34	0.000835	0.00484	RNF165
ENSG00000092068	21.4	1.13	0.338	3.34	0.000845	0.00488	SLC7A8
ENSG00000182310	12.3	1.53	0.461	3.32	0.000901	0.00516	SPACA6
ENSG00000006611	22	1.22	0.368	3.32	0.000909	0.0052	USH1C
ENSG00000120049	17.7	1.32	0.398	3.31	0.000935	0.00532	KCNIP2
ENSG00000187123	25.2	1.04	0.316	3.3	0.000964	0.00547	LYPD6

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000082196	9.33	1.9	0.578	3.29	0.000984	0.00557	C1QTNF3
ENSG00000151470	22.6	1.07	0.328	3.28	0.00105	0.00587	C4orf33
ENSG00000186081	3.85	-2.97	0.908	-3.27	0.00109	0.00607	KRT5
ENSG00000186417	4.51	3.86	1.18	3.26	0.00111	0.00618	GLDN
ENSG00000226381	18.3	1.32	0.407	3.26	0.00113	0.00627	RP11-119F19.2
ENSG00000162999	12.9	1.52	0.47	3.23	0.00123	0.00673	DUSP19
ENSG00000165801	22.3	1.12	0.346	3.23	0.00123	0.00675	ARHGEF40
ENSG00000069712	14.3	1.4	0.434	3.22	0.00127	0.00693	KIAA1107
ENSG00000138380	14.1	1.48	0.461	3.21	0.00131	0.00712	CARF
ENSG00000139278	12	-1.41	0.438	-3.21	0.00132	0.00717	GLIPR1
ENSG00000106006	18	1.22	0.379	3.21	0.00134	0.00723	HOXA6
ENSG00000243302	47.6	1.05	0.326	3.2	0.00136	0.0073	RP11-274B21.2
ENSG00000115266	16.7	1.23	0.385	3.2	0.00139	0.00747	APC2
ENSG00000167037	22	1.06	0.333	3.19	0.00141	0.00755	SGSM1
ENSG00000203867	19	1.22	0.382	3.19	0.00142	0.00756	RBM20
ENSG00000254353	15.1	1.5	0.472	3.18	0.00145	0.00772	RP11-195E2.4
ENSG00000255090	11.9	1.62	0.508	3.18	0.00147	0.00782	RP11-820L6.1
ENSG00000110455	21	1.05	0.33	3.18	0.00149	0.0079	ACCS
ENSG00000204128	19.5	1.18	0.373	3.17	0.00155	0.00813	C2orf72
ENSG00000174514	7.6	1.97	0.622	3.17	0.00155	0.00813	MFSD4A
ENSG00000250988	22	1.19	0.377	3.15	0.00162	0.00844	SNHG21
ENSG00000258077	15.4	1.25	0.397	3.14	0.00168	0.00867	RP11-114H23.1
ENSG00000102934	6.13	2.35	0.75	3.13	0.00175	0.00897	PLLP
ENSG00000156966	26.5	1.15	0.37	3.11	0.00188	0.00952	B3GNT7
ENSG00000181143	4.74	-2.34	0.755	-3.1	0.00191	0.00964	MUC16
ENSG00000117472	18.5	-1.11	0.358	-3.1	0.00196	0.00984	TSPAN1
ENSG00000125434	15.1	1.29	0.416	3.09	0.00197	0.0099	SLC25A35
ENSG00000235363	24.1	1.44	0.468	3.07	0.00215	0.0107	SNRPGP10
ENSG00000223940	8.1	-1.73	0.563	-3.06	0.00218	0.0108	KRT8P8
ENSG00000168913	14.2	1.52	0.496	3.06	0.00218	0.0108	ENHO
ENSG00000056558	14.8	1.29	0.421	3.06	0.00219	0.0108	TRAF1
ENSG00000136010	5.1	2.48	0.81	3.06	0.00221	0.0109	ALDH1L2
ENSG00000231305	7.03	2.25	0.734	3.06	0.00221	0.0109	RP11-723O4.2
ENSG00000100968	7.62	1.81	0.59	3.06	0.00223	0.011	NFATC4
ENSG00000136250	4.35	3.1	1.02	3.05	0.00228	0.0111	AOAH
ENSG00000229994	13.8	1.58	0.522	3.03	0.00241	0.0117	RPL5P4
ENSG00000267896	18	1.21	0.4	3.02	0.00249	0.012	AC018766.4
ENSG00000273000	28.1	-2.14	0.709	-3.02	0.00253	0.0121	KB-1572G7.2

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000230715	19	1.26	0.416	3.02	0.00253	0.0121	RP11-274B21.4
ENSG00000235706	6.62	2.15	0.713	3.02	0.00255	0.0122	DICER1-AS1
ENSG00000119686	8.3	1.97	0.654	3.01	0.00259	0.0123	FLVCR2
ENSG00000124107	7.3	-1.85	0.617	-3	0.00266	0.0126	SLPI
ENSG00000266037	7.49	2.93	0.975	3	0.0027	0.0127	RN7SL3
ENSG00000133808	11.7	1.52	0.506	3	0.00272	0.0128	MICALCL
ENSG00000261716	17.6	1.14	0.379	3	0.00274	0.0129	H2BC20P
ENSG00000266401	13.3	1.48	0.494	2.99	0.00279	0.0131	LOC105371967
ENSG00000253366	3.68	4.37	1.47	2.99	0.00283	0.0132	RP11-589F5.4
ENSG00000105991	22.8	1.09	0.366	2.98	0.00288	0.0134	HOXA1
ENSG00000163032	11.1	1.49	0.499	2.98	0.0029	0.0134	VSNL1
ENSG00000180818	11	1.43	0.483	2.97	0.003	0.0138	HOXC10
ENSG00000197415	20.2	1.11	0.373	2.97	0.003	0.0138	VEPH1
ENSG00000103089	18.4	1.16	0.392	2.97	0.00302	0.0139	FA2H
ENSG00000224114	139	1.73	0.585	2.96	0.00304	0.014	RP11-343H5.4
ENSG00000172794	10.9	1.46	0.495	2.96	0.00307	0.0141	RAB37
ENSG00000187240	16	1.3	0.441	2.96	0.00309	0.0141	DYNC2H1
ENSG00000167895	18.3	1.26	0.428	2.96	0.00312	0.0142	TMC8
ENSG00000237883	24.7	1.09	0.372	2.93	0.00337	0.0152	DGUOK-AS1
ENSG00000149212	12.8	1.3	0.445	2.93	0.00343	0.0154	SESN3
ENSG00000131737	6.76	-2.1	0.72	-2.92	0.0035	0.0157	KRT34
ENSG00000148346	8.33	-1.67	0.574	-2.91	0.00363	0.0161	LCN2
ENSG00000257913	16	1.13	0.39	2.91	0.00365	0.0162	DDN-AS1
ENSG00000155066	17.4	1.17	0.405	2.9	0.00376	0.0166	PROM2
ENSG00000196787	7.44	2.13	0.735	2.89	0.00384	0.0169	H2AC11
ENSG00000197576	7.8	1.85	0.639	2.89	0.00388	0.0171	HOXA4
ENSG00000141854	10.6	1.47	0.512	2.88	0.00399	0.0174	MISP3
ENSG00000259366	4.34	2.91	1.01	2.88	0.00403	0.0176	CTD-2647L4.4
ENSG00000166396	7.05	-2.2	0.767	-2.87	0.00405	0.0176	SERPINB7
ENSG00000085563	19.4	1.23	0.427	2.87	0.00411	0.0179	ABCB1
ENSG00000260196	7.5	1.96	0.684	2.86	0.0042	0.0182	RP1-239B22.5
ENSG00000157303	9.52	1.76	0.614	2.86	0.00424	0.0183	SUSD3
ENSG00000250802	9.72	1.59	0.555	2.86	0.00424	0.0183	ZBED3-AS1
ENSG00000271576	14.2	1.21	0.427	2.84	0.00454	0.0194	RP11-486G15.2
ENSG00000197249	5.77	-1.9	0.671	-2.84	0.00455	0.0194	SERPINA1
ENSG00000172554	12.4	1.34	0.472	2.83	0.00462	0.0196	SNTG2
ENSG00000163689	18.6	1.07	0.378	2.83	0.00464	0.0197	C3orf67
ENSG00000167754	8.68	-1.43	0.506	-2.83	0.00466	0.0198	KLK5

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000188015	17.8	-1.09	0.387	-2.82	0.00473	0.02	S100A3
ENSG00000198750	9.31	1.74	0.617	2.82	0.0048	0.0202	GATSL2
ENSG00000230513	10.9	1.63	0.579	2.82	0.00485	0.0204	THAP7-AS1
ENSG00000270195	7.18	1.89	0.674	2.81	0.00495	0.0208	RP11-572O17.1
ENSG00000162148	10	1.5	0.535	2.8	0.00506	0.0211	PPP1R32
ENSG00000234684	15	1.15	0.41	2.8	0.00507	0.0212	SDCBP2-AS1
ENSG00000149346	23.9	1.01	0.362	2.8	0.00513	0.0214	SLX4IP
ENSG00000149654	10.9	1.44	0.516	2.8	0.00515	0.0214	CDH22
ENSG00000243708	27.1	-1.19	0.425	-2.8	0.00516	0.0215	PLA2G4B
ENSG00000153162	11	1.6	0.572	2.79	0.00521	0.0216	BMP6
ENSG00000129675	4.56	2.67	0.959	2.79	0.00535	0.0221	ARHGEF6
ENSG00000169435	11.9	1.37	0.492	2.78	0.00539	0.0223	RASSF6
ENSG00000174125	6.42	1.93	0.699	2.77	0.00563	0.0231	TLR1
ENSG00000157021	25	1.1	0.399	2.77	0.00566	0.0232	FAM92A1P1
ENSG00000103942	5.96	1.98	0.715	2.76	0.00573	0.0234	HOMER2
ENSG00000182774	2.05e+03	1.04	0.378	2.76	0.00576	0.0235	RPS17
ENSG00000266402	22.4	1.22	0.446	2.74	0.00614	0.0248	SNHG25
ENSG00000185477	3.83	2.6	0.949	2.74	0.00618	0.0249	GPRIN3
ENSG00000188859	9.58	1.38	0.504	2.74	0.00621	0.025	FAM78B
ENSG00000175785	15	1.16	0.424	2.73	0.00627	0.0252	PRIMA1
ENSG00000169860	11.1	1.39	0.508	2.73	0.00634	0.0255	P2RY1
ENSG00000257335	6.01	2.05	0.751	2.73	0.00639	0.0257	MGAM
ENSG00000182472	4.9	2.38	0.874	2.73	0.00641	0.0257	CAPN12
ENSG00000248334	12.7	1.43	0.523	2.72	0.00643	0.0258	WHAMMP2
ENSG00000256616	3.99	2.9	1.06	2.72	0.00646	0.0259	RP11-815J4.6
ENSG00000136630	13.4	1.23	0.453	2.72	0.0066	0.0264	HLX
ENSG00000174564	9.76	1.47	0.541	2.71	0.00665	0.0265	IL20RB
ENSG00000132003	7.82	1.55	0.572	2.71	0.00666	0.0265	ZSWIM4
ENSG00000182179	4.28	2.93	1.09	2.7	0.007	0.0276	UBA7
ENSG00000174370	13.7	1.24	0.46	2.7	0.007	0.0277	C11orf45
ENSG00000168994	16.6	1.09	0.406	2.69	0.00715	0.0281	PXDC1
ENSG00000159905	10.8	1.39	0.519	2.68	0.00728	0.0285	ZNF221
ENSG00000225813	6.36	2.22	0.827	2.68	0.00734	0.0287	AC009299.4
ENSG00000173727	7.61	1.6	0.601	2.66	0.00786	0.0304	AP000769.1
ENSG00000213462	18.1	1.01	0.38	2.66	0.0079	0.0305	ERV3-1
ENSG00000151572	5.57	2.29	0.866	2.65	0.00812	0.0312	ANO4
ENSG00000272617	21.8	4.67	1.77	2.65	0.00813	0.0312	RP11-343C2.12
ENSG00000226221	8.12	-3.92	1.48	-2.65	0.00814	0.0312	AC022431.1

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000119138	14.7	1.1	0.416	2.64	0.00822	0.0315	KLF9
ENSG00000227921	9.56	1.84	0.699	2.64	0.00839	0.032	AL353791.1
ENSG00000158373	14.6	1.08	0.411	2.63	0.00845	0.0322	H2BC5
ENSG00000167807	11.4	-6.06	2.31	-2.63	0.00856	0.0326	CTD-2369P2.10
ENSG00000241732	6.42	1.82	0.692	2.63	0.00856	0.0326	RP11-38P22.2
ENSG00000066248	13.7	1.14	0.433	2.63	0.00861	0.0327	NGEF
ENSG00000272841	8.91	-1.6	0.609	-2.62	0.00867	0.0329	RP3-428L16.2
ENSG00000180998	10.1	1.51	0.575	2.62	0.00874	0.0331	GPR137C
ENSG00000188266	11.2	1.27	0.483	2.62	0.00876	0.0332	HYKK
ENSG00000006283	10.2	1.19	0.453	2.62	0.00876	0.0332	CACNA1G
ENSG00000184384	4.83	2.22	0.846	2.62	0.00879	0.0333	MAML2
ENSG00000213442	41.6	-1.5	0.574	-2.62	0.0088	0.0333	RPL18AP3
ENSG00000196268	18.7	1.13	0.433	2.61	0.00898	0.0338	ZNF493
ENSG00000106333	16.1	1.06	0.407	2.61	0.00899	0.0338	PCOLCE
ENSG00000260128	6.55	2.48	0.95	2.61	0.00903	0.034	ULK4P2
ENSG00000214922	10.5	1.33	0.512	2.59	0.00954	0.0355	HLA-F-AS1
ENSG00000128573	4.43	2.22	0.856	2.59	0.0096	0.0356	FOXP2
ENSG00000161082	15.2	1.08	0.418	2.59	0.00968	0.0359	CELF5
ENSG00000204060	13.2	1.14	0.442	2.58	0.00986	0.0365	FOXO6
ENSG00000146856	5.06	1.97	0.764	2.58	0.00996	0.0368	AGBL3
ENSG00000270800	164	1	0.392	2.56	0.0105	0.0385	RPS10-NUDT3
ENSG00000235499	3.93	3.07	1.2	2.56	0.0106	0.0387	AC073046.25
ENSG00000225031	12.8	-1.16	0.454	-2.56	0.0106	0.0387	EIF4BP7
ENSG00000216775	6.66	1.72	0.674	2.55	0.0106	0.0389	LOC730101
ENSG00000167550	8.56	1.49	0.583	2.55	0.0108	0.0392	RHEBL1
ENSG00000239467	4.46	2.73	1.07	2.55	0.0109	0.0395	ERICH2
ENSG00000069020	14.2	1.04	0.41	2.55	0.0109	0.0396	MAST4
ENSG00000036448	15.9	1.09	0.43	2.54	0.011	0.0397	MYOM2
ENSG00000253552	4.38	2.29	0.9	2.54	0.011	0.0398	HOXA-AS2
ENSG00000080947	14.2	1.07	0.422	2.53	0.0113	0.0407	CROCCP3
ENSG00000189136	10	1.38	0.544	2.53	0.0114	0.0409	UBE2Q2P1
ENSG00000136367	5.9	1.71	0.674	2.53	0.0114	0.041	ZFHX2
ENSG00000166257	5.43	1.96	0.777	2.53	0.0115	0.0413	SCN3B
ENSG00000248405	25.1	-1.07	0.423	-2.52	0.0116	0.0417	PRR5-ARHGAP8
ENSG00000163633	11.9	1.16	0.463	2.51	0.012	0.0426	C4orf36
ENSG00000260655	6.35	1.73	0.688	2.51	0.0121	0.0428	CTA-250D10.23
ENSG00000188641	6.39	1.76	0.702	2.51	0.0122	0.0433	DPYD
ENSG00000257497	7.02	1.73	0.69	2.5	0.0123	0.0434	RP11-585P4.5

Continued on next page

	baseMean	log2FC	lfcSE	stat	p-value	p-adjusted	Gene
ENSG00000129595	18.5	1.04	0.417	2.5	0.0125	0.044	EPB41L4A
ENSG00000273274	7.22	1.57	0.63	2.5	0.0125	0.0441	ZBTB8B
ENSG00000260269	3.72	3.28	1.31	2.49	0.0126	0.0444	CTD-2323K18.1
ENSG00000268175	6.63	2.01	0.809	2.49	0.0128	0.045	AC117395.1
ENSG00000196338	5.39	1.81	0.728	2.49	0.0129	0.0452	NLGN3
ENSG00000261655	4.27	2.15	0.865	2.48	0.013	0.0455	CTD-3064M3.3
ENSG00000136883	4.21	2.14	0.864	2.48	0.0131	0.0456	KIF12
ENSG00000260461	14.9	1	0.404	2.48	0.0132	0.0458	RP11-541N10.3
ENSG00000185745	7.59	1.49	0.602	2.47	0.0135	0.0467	IFIT1
ENSG00000163082	9.68	1.32	0.536	2.47	0.0136	0.047	SGPP2
ENSG00000105519	3.88	2.35	0.954	2.47	0.0137	0.0473	CAPS
ENSG00000164778	5.14	1.94	0.789	2.46	0.0139	0.0478	EN2
ENSG00000204771	28.5	2.56	1.04	2.46	0.014	0.0483	CTD-3232M19.2
ENSG00000213760	4.65	2.06	0.839	2.46	0.0141	0.0484	ATP6V1G2
ENSG00000215472	14	1.59	0.649	2.45	0.0143	0.049	RPL17-C18orf32
ENSG00000144452	6.2	2.01	0.824	2.44	0.0145	0.0495	ABCA12
ENSG00000222009	10.3	1.23	0.502	2.44	0.0146	0.0497	BTBD19

C.1.2 Hallmark DNA Repair Genes

Table C.4: DDR genes curated as part of the Molecular Signatures Database (MSigDB) Hallmark DNA repair gene set. log2FC represents the log2 fold change in expression compared to 0h post infection, lfcSE refers to the standard error of the log2 fold change, and p-adj refers to the multiple testing adjusted Wald test p-value.

	6h			12h			24h		
Gene	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj
AAAS	0.0372	0.0759	0.832	0.0748	0.0806	0.627	-0.115	0.0761	0.262
ADA	0.0548	0.136	0.866	0.322	0.142	0.127	0.122	0.134	0.532
ADCY6	0.202	0.145	0.44	0.163	0.155	0.571	0.673	0.142	2.97e-05
ADRM1	-0.0873	0.0487	0.291	-0.175	0.0522	0.0131	-0.601	0.0497	5.09e-31
AGO4	0.193	0.271		0.264	0.286	0.63	0.936	0.258	0.00199
AK1	-0.255	0.118	0.184	0.0374	0.124	0.899	-0.101	0.117	0.553
AK3	0.0802	0.0826	0.626	0.113	0.0879	0.466	0.0107	0.0826	0.939
ALYREF	-0.164	0.0928	0.301	-0.23	0.0993	0.118	-0.576	0.0931	2.19e-08
APRT	-0.167	0.0989	0.328	0.103	0.105	0.601	0.0306	0.0982	0.852

Continued on next page

	6h			12h			24h		
Gene	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj
ARL6IP1	0.177	0.0633	0.0637	-0.337	0.0682	5.36e-05	-0.427	0.0638	1.03e-09
BCAM	-0.0675	0.115	0.792	0.247	0.121	0.183	0.487	0.112	0.000153
BCAP31	-0.174	0.103	0.329	-0.18	0.11	0.314	-0.368	0.103	0.00233
BOLA2	-0.374	1.7		-9.43	2.16	0.000553	-1.59	1.7	0.52
BRF2	-0.153	0.143	0.58	-0.159	0.153	0.573	-0.778	0.148	2.67e-06
CANT1	-0.161	0.0618	0.0915	-0.369	0.0667	3.56e-06	-0.431	0.0624	2.39e-10
CCNO	-0.207	0.226		-0.00137	0.238	0.998	0.0543	0.222	0.886
CDA	-0.627	0.188		-0.474	0.199	0.106	-0.57	0.185	0.0103
CETN2	-0.0924	0.122	0.72	0.0728	0.128	0.793	0.0851	0.12	0.638
CLP1	-0.177	0.111	0.364	-0.0894	0.118	0.709	-0.404	0.112	0.00221
CMPK2	0.467	0.502		0.709	0.522	0.431	0.57	0.494	0.412
COX17	-0.362	0.187	0.245	0.0422	0.197	0.929	-0.126	0.185	0.653
CSTF3	-0.13	0.0677	0.25	-0.209	0.0726	0.0397	-0.36	0.0682	2.57e-06
DAD1	-0.0632	0.0975	0.766	-0.0726	0.104	0.737	-0.0998	0.0975	0.473
DCTN4	0.0382	0.0704	0.809	0.0129	0.0752	0.946	-0.0628	0.0706	0.543
DDB1	0.111	0.0481	0.149	0.121	0.0513	0.11	0.0391	0.0482	0.584
DDB2	-0.129	0.122	0.587	0.296	0.126	0.114	-0.0356	0.12	0.859
DGCR8	-0.173	0.0836	0.209	-0.02	0.0889	0.926	-0.393	0.0841	4.01e-05
DGUOK	-0.0132	0.097	0.96	0.14	0.102	0.429	0.129	0.0958	0.327
DUT	-0.16	0.08	0.226	-0.0178	0.0848	0.932	-0.215	0.08	0.0279
EDF1	-0.197	0.109	0.285	0.0963	0.116	0.67	0.0329	0.108	0.855
EIF1B	0.0258	0.105	0.924	0.015	0.112	0.959	-0.0214	0.105	0.905
ELL	-0.0327	0.114	0.912	0.0447	0.121	0.874	-0.126	0.114	0.431
ELOA	0.0945	0.0635	0.398	-0.0627	0.0681	0.63	-0.418	0.0644	3.52e-09
ERCC1	-0.18	0.0895	0.225	-0.268	0.0962	0.048	-0.218	0.089	0.0494
ERCC2	-0.147	0.0848	0.315	0.0827	0.0892	0.628	0.0682	0.0835	0.58
ERCC3	0.0782	0.0585	0.459	-0.0103	0.0627	0.949	-0.131	0.0589	0.0783
ERCC4	-0.00534	0.175	0.991	-0.351	0.191	0.244	0.0866	0.173	0.753
ERCC5	0.00211	0.215	0.997	-0.163	0.231	0.733	0.365	0.213	0.194
ERCC8	-0.0716	0.158	0.848	0.0608	0.167	0.876	-0.0406	0.157	0.879
FEN1	-0.288	0.0808	0.0118	-0.133	0.0861	0.353	-0.33	0.0808	0.000404
GMPR2	-0.0677	0.0873	0.713	0.0194	0.0927	0.932	0.136	0.0862	0.237
GPX4	-0.174	0.113	0.379	0.161	0.12	0.439	0.0854	0.112	0.611
GSDME	-0.179	0.416		-0.0178	0.436	0.986	0.939	0.382	0.0479
GTF2A2	-0.0251	0.1	0.923	0.103	0.106	0.607	-0.069	0.1	0.649
GTF2B	-0.206	0.107	0.25	-0.128	0.114	0.536	-0.128	0.106	0.385
GTF2F1	-0.0663	0.0639	0.595	-0.133	0.0684	0.208	-0.2	0.0641	0.00923

Continued on next page

	6h			12h			24h		
Gene	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj
GTF2H1	0.178	0.0688	0.0939	0.0994	0.0735	0.434	0.000191	0.0691	0.999
GTF2H3	0.0641	0.0721	0.663	0.125	0.0768	0.317	-0.129	0.0725	0.175
GTF2H5	-0.0793	0.151	0.815	-0.00911	0.16	0.982	-0.185	0.151	0.379
GTF3C5	-0.216	0.0599	0.0108	-0.106	0.0636	0.304	-0.297	0.06	1.19e-05
GUK1	-0.149	0.107	0.44	-0.062	0.114	0.803	-0.277	0.107	0.0371
HPRT1	-0.129	0.0992	0.478	-0.0783	0.106	0.717	-0.395	0.1	0.000668
IMPDH2	0.00745	0.0569	0.961	0.237	0.0606	0.00257	-0.0676	0.057	0.397
ITPA	-0.285	0.0984	0.0522	-0.302	0.105	0.0406	-0.485	0.0991	1.55e-05
LIG1	-0.0807	0.0806	0.611	0.145	0.0854	0.294	-0.219	0.0809	0.0268
MPC2	-0.159	0.0924	0.318	-0.0329	0.0977	0.887	-0.165	0.0919	0.169
MPG	-0.256	0.175	0.409	-0.29	0.187	0.351	-0.51	0.177	0.0171
MRPL40	0.0365	0.107	0.891	0.162	0.113	0.4	0.0289	0.106	0.872
NCBP2	0.111	0.0589	0.263	-0.0521	0.0634	0.675	-0.0108	0.0591	0.915
NELFB	-0.138	0.0638	0.181	-0.114	0.068	0.301	-0.147	0.0637	0.0662
NELFCD	-0.0967	0.0512	0.261	-0.0232	0.0544	0.852	-0.284	0.0515	8.2e-07
NELFE	-0.248	0.0853	0.0514	-0.248	0.0911	0.0554	-0.0228	0.0837	0.871
NFX1	-0.0206	0.0663	0.904	-0.0436	0.0708	0.773	-0.264	0.0669	0.000683
NME1	-0.121	0.152	0.704	0.0179	0.162	0.967	-0.175	0.152	0.412
NME3	-0.166	0.124	0.458	-0.0372	0.131	0.905	0.149	0.121	0.375
NME4	0.103	0.107	0.626	0.404	0.112	0.00626	0.483	0.104	4.86e-05
NPR2	0.032	0.139	0.929	-0.248	0.151	0.314	0.0427	0.138	0.853
NT5C	-0.258	0.112	0.15	-0.0405	0.119	0.885	-0.377	0.113	0.00471
NT5C3A	-0.00583	0.101	0.983	0.0689	0.108	0.763	-0.365	0.103	0.00259
NUDT21	0.0342	0.0404	0.681	-0.0419	0.0433	0.607	-0.118	0.0406	0.0165
NUDT9	-0.119	0.147	0.697	-0.136	0.157	0.656	0.193	0.145	0.333
PCNA	-0.301	0.0697	0.00126	-0.15	0.0744	0.189	-0.393	0.0698	4.61e-07
PDE4B	0.0456	0.0775	0.791	0.0593	0.0826	0.726	0.194	0.0768	0.0418
PNP	0.0412	0.0633	0.766	-0.0981	0.068	0.395	-0.203	0.0639	0.00772
POLA1	-0.102	0.0642	0.366	-0.0144	0.068	0.931	0.0649	0.0631	0.47
POLA2	-0.278	0.1	0.0665	0.0866	0.104	0.671	-0.19	0.099	0.139
POLB	-0.0422	0.127	0.896	0.144	0.134	0.561	-0.102	0.127	0.587
POLD1	-0.0564	0.0626	0.657	0.0295	0.0664	0.846	-0.188	0.0628	0.013
POLD3	-0.192	0.0977	0.238	0.155	0.102	0.362	-0.178	0.0971	0.159
POLD4	-1.02	0.276		-0.659	0.285	0.118	-0.0695	0.25	0.869
POLE4	-0.466	0.219	0.191	-0.335	0.232	0.394	-0.682	0.221	0.00999
POLH	0.15	0.111	0.455	-0.027	0.119	0.925	-0.114	0.112	0.472
POLL	-0.207	0.127	0.349	-0.249	0.136	0.244	-0.348	0.127	0.0246

Continued on next page

	6h			12h			24h		
Gene	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj
POLR1C	-0.175	0.0865	0.22	-0.28	0.0927	0.0287	-0.452	0.0872	3.94e-06
POLR1D	0.0135	0.0805	0.95	0.0437	0.0857	0.816	0.0108	0.0803	0.937
POLR2A	0.186	0.0943	0.233	0.145	0.101	0.395	-0.0163	0.0944	0.92
POLR2C	0.159	0.0667	0.13	0.288	0.0707	0.00153	-0.119	0.0674	0.178
POLR2D	0.142	0.063	0.156	0.113	0.0672	0.3	-0.0794	0.0634	0.367
POLR2E	0.0612	0.0561	0.57	0.00323	0.06	0.983	0.027	0.056	0.763
POLR2F	-0.245	0.106	0.143	-0.13	0.112	0.521	-0.32	0.106	0.0118
POLR2G	-0.119	0.118	0.608	0.0905	0.125	0.724	-0.0383	0.117	0.845
POLR2H	0.0311	0.102	0.907	0.165	0.109	0.363	-0.0886	0.102	0.555
POLR2I	-0.205	0.138	0.398	0.228	0.144	0.335	0.00767	0.136	0.974
POLR2J	0.113	0.122	0.647	0.368	0.129	0.0416	0.272	0.121	0.0757
POLR2K	-0.137	0.0965	0.427	-0.122	0.103	0.512	-0.246	0.0968	0.0397
POLR3C	-0.177	0.0881	0.225	-0.18	0.094	0.22	-0.425	0.0889	2.55e-05
POLR3GL	0.179	0.224		0.37	0.234	0.337	0.778	0.215	0.00202
POM121	0.207	0.0701	0.0472	0.014	0.0752	0.94	-0.227	0.0708	0.00719
PRIM1	-0.239	0.101	0.135	0.0236	0.106	0.927	-0.0115	0.0994	0.946
RAD51	-0.188	0.106	0.296	0.0627	0.111	0.794	-0.11	0.105	0.459
RAD52	-0.00817	0.112	0.978	-0.0481	0.12	0.862	-0.0324	0.111	0.861
RAE1	0.05	0.0655	0.718	0.0684	0.0697	0.602	-0.104	0.0657	0.236
RALA	0.011	0.0794	0.959	0.114	0.0843	0.435	-0.0244	0.0793	0.853
RBX1	-0.157	0.131	0.526	-0.04	0.139	0.904	-0.136	0.131	0.466
REV3L	0.0759	0.0911	0.686	-0.04	0.0977	0.859	0.0696	0.0908	0.608
RFC2	-0.171	0.0772	0.167	-0.00297	0.0818	0.987	-0.197	0.0771	0.0389
RFC3	0.133	0.0665	0.229	0.108	0.0709	0.363	0.106	0.0662	0.233
RFC4	-0.0813	0.0923	0.666	0.163	0.0978	0.303	0.109	0.0917	0.395
RFC5	-0.0379	0.0758	0.827	0.157	0.0802	0.206	0.152	0.0752	0.116
RNMT	0.311	0.0724	0.00136	0.149	0.0776	0.219	0.219	0.0726	0.0122
RPA2	-0.0884	0.0698	0.493	0.0794	0.0738	0.558	-0.133	0.0697	0.141
RPA3	-0.163	0.182	0.66	0.2	0.192	0.573	0.141	0.18	0.6
RRM2B	0.0719	0.0788	0.652	0.111	0.0839	0.447	0.0527	0.0787	0.66
SAC3D1	-0.327	0.156	0.203	-0.124	0.166	0.713	-0.208	0.156	0.331
SDCBP	0.0497	0.0704	0.742	0.16	0.0745	0.157	0.206	0.0695	0.0141
SEC61A1	0.0418	0.0476	0.666	0.00865	0.0509	0.947	-0.135	0.0478	0.0204
SF3A3	-0.0665	0.0725	0.649	0.000466	0.0774	0.998	-0.0252	0.0725	0.835
SMAD5	0.218	0.0699	0.0334	-0.0404	0.0754	0.805	0.0587	0.0702	0.57
SNAPC4	-0.0437	0.0711	0.781	-0.0473	0.0759	0.769	-0.348	0.072	2.03e-05
SNAPC5	-0.212	0.12	0.3	-0.00649	0.126	0.983	-0.0567	0.118	0.764

Continued on next page

	6h			12h			24h		
Gene	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj	log2FC	lfcSE	p-adj
SRSF6	-0.0301	0.0345	0.667	-0.0101	0.0368	0.908	-0.255	0.0347	1.16e-11
SSRP1	0.0184	0.0633	0.911	0.0418	0.0676	0.771	-0.00443	0.0633	0.967
STX3	-0.135	0.0834	0.356	-0.209	0.0894	0.116	-0.323	0.0839	0.000941
SUPT4H1	-0.168	0.0868	0.245	-0.0358	0.0922	0.866	-0.126	0.0865	0.281
SUPT5H	0.0445	0.0641	0.747	0.0269	0.0685	0.865	-0.164	0.0644	0.0395
SURF1	-0.277	0.124	0.163	-0.0477	0.13	0.876	-0.145	0.122	0.396
TAF10	-0.191	0.0956	0.229	0.0111	0.101	0.967	-0.0268	0.0944	0.865
TAF12	-0.0123	0.109	0.966	-0.095	0.117	0.68	-0.176	0.109	0.23
TAF13	-0.239	0.111	0.185	-0.386	0.12	0.0175	-0.606	0.113	1.55e-06
TAF1C	-0.0389	0.0695	0.802	0.03	0.0739	0.86	-0.21	0.0698	0.0122
TAF6	-0.0589	0.0946	0.777	-0.0251	0.101	0.917	0.206	0.0935	0.0812
TARBP2	0.0651	0.0888	0.728	-0.0182	0.0952	0.938	-0.25	0.09	0.0225
TK2	-0.0647	0.15	0.855	-0.119	0.161	0.717	0.122	0.147	0.573
TMED2	0.0803	0.0597	0.457	-0.012	0.0639	0.939	-0.135	0.0599	0.0737
TP53	0.218	0.082	0.0822	0.141	0.0876	0.325	0.156	0.082	0.143
TSG101	-0.104	0.0875	0.529	-0.0778	0.0932	0.669	-0.0202	0.0867	0.891
TYMS	-0.0425	0.0441	0.628	0.0364	0.0469	0.7	-0.125	0.0442	0.0202
UMPS	-0.132	0.0761	0.315	-0.0112	0.0808	0.958	-0.264	0.0764	0.00337
UPF3B	0.14	0.113	0.512	0.204	0.121	0.295	0.33	0.113	0.0153
USP11	0.00928	0.0691	0.96	0.0268	0.0735	0.876	-0.00128	0.0688	0.992
VPS28	-0.205	0.0963	0.191	-0.0996	0.102	0.605	-0.127	0.0954	0.331
VPS37B	-0.0103	0.0765	0.96	-0.152	0.0822	0.241	-0.227	0.077	0.0147
VPS37D	0.28	0.311		0.2	0.332	0.776	0.418	0.307	0.319
XPC	-0.146	0.0994	0.408	0.0364	0.105	0.883	-0.104	0.0989	0.457
ZNF707	-0.147	0.142	0.596	-0.128	0.151	0.662	-0.161	0.141	0.417
ZNRD1	-0.027	0.132	0.94	-0.157	0.142	0.545	-0.302	0.133	0.0731
ZWINT	-0.211	0.0776	0.0732	-0.146	0.0828	0.267	-0.355	0.0778	6.38e-05

C.2 Mutational signatures

C.2.1 PCAWG Tumour Types

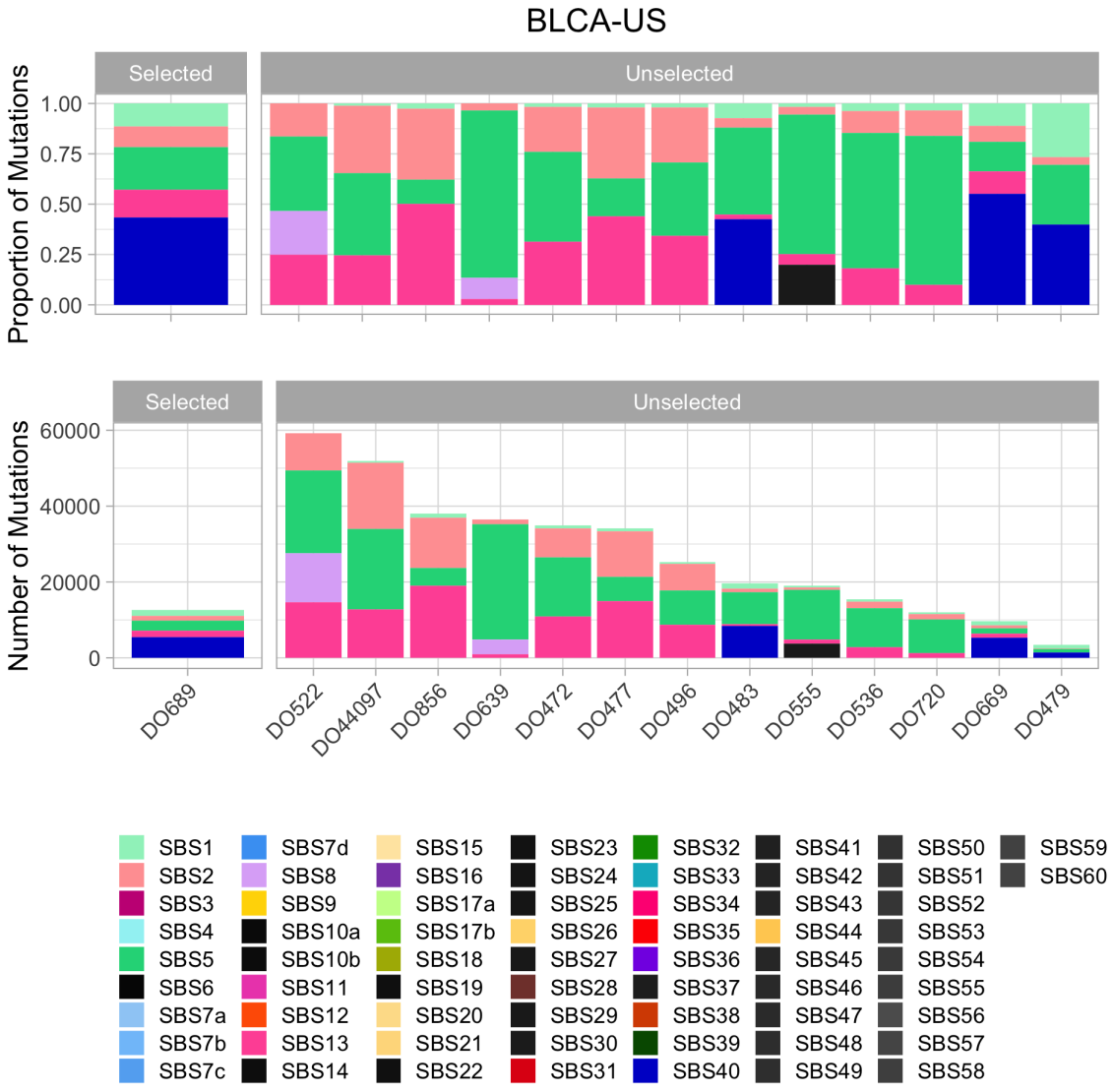


Figure C.1: Distribution of mutational signatures in selected and unselected bladder cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

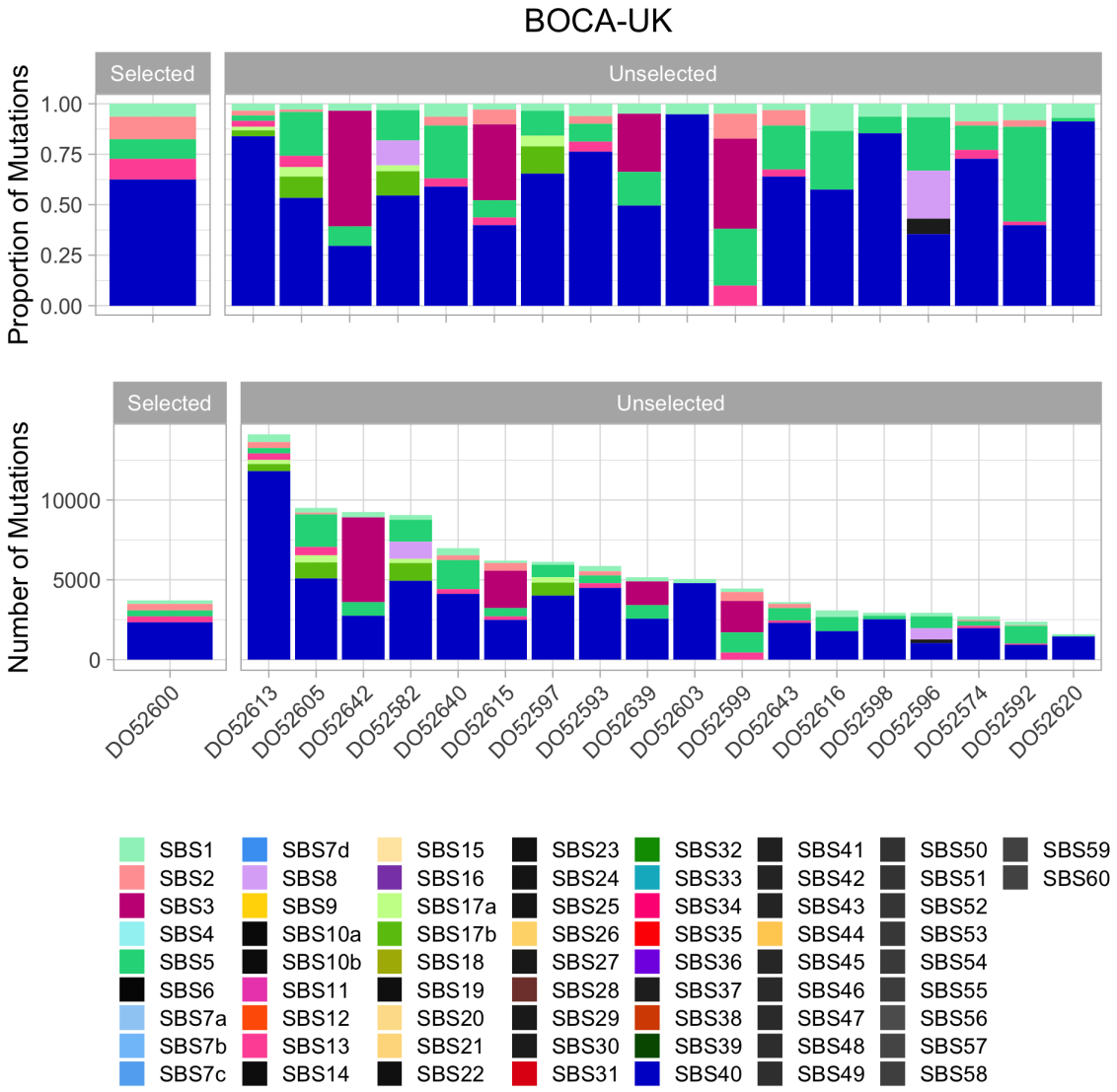


Figure C.2: Distribution of mutational signatures in selected and unselected bone cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

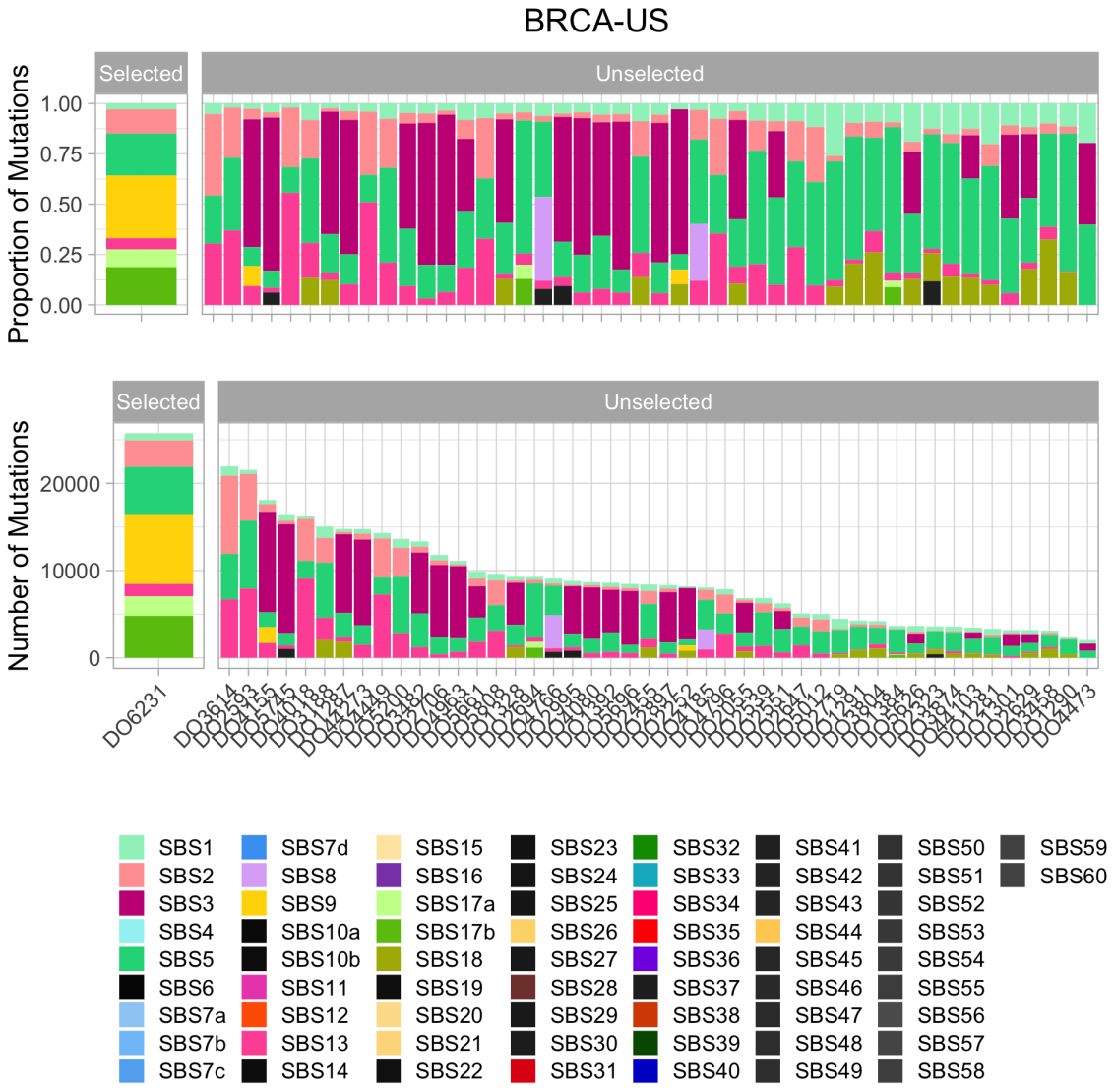


Figure C.3: Distribution of mutational signatures in selected and unselected breast cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

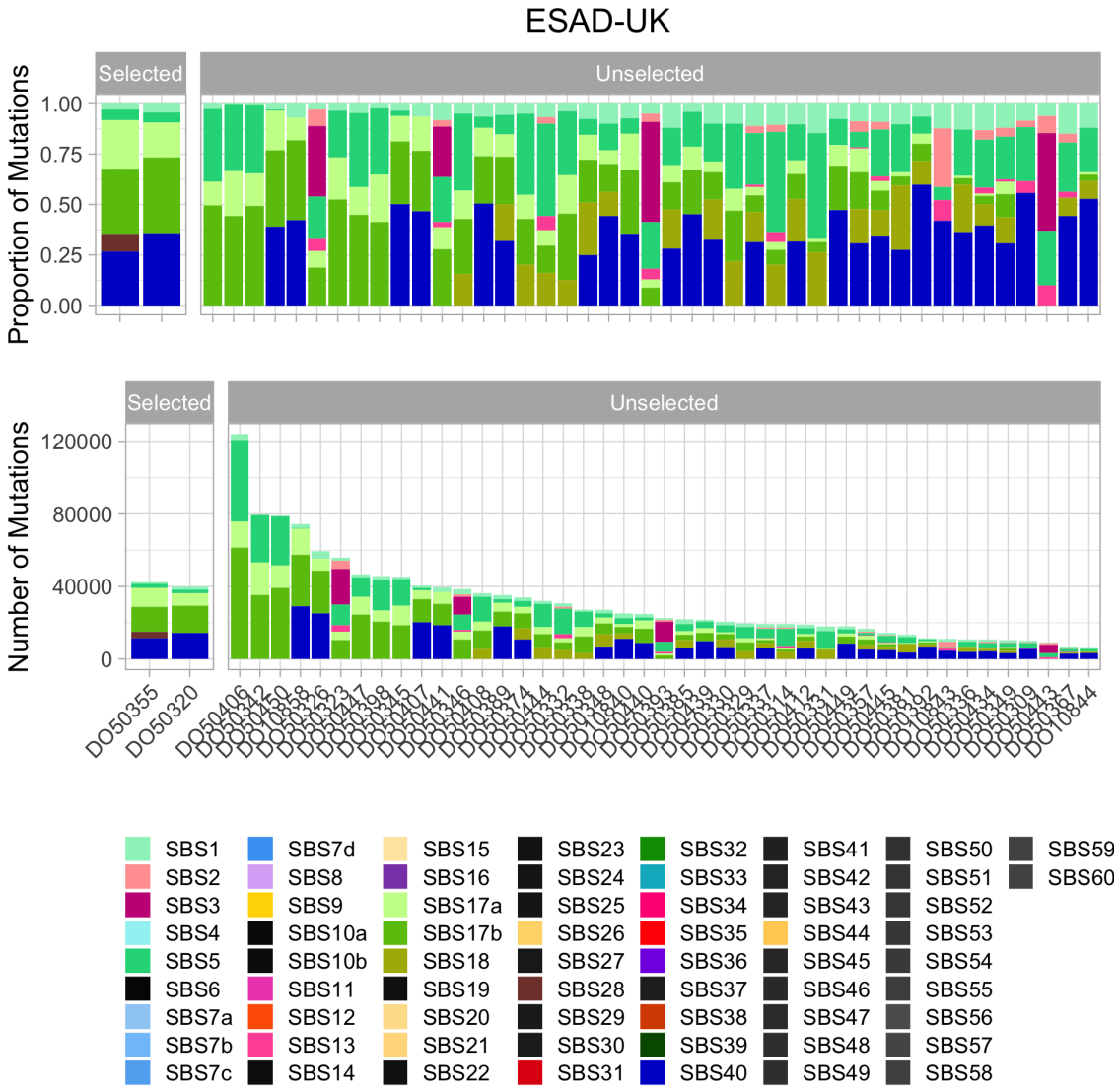


Figure C.4: Distribution of mutational signatures in selected and unselected esophageal cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

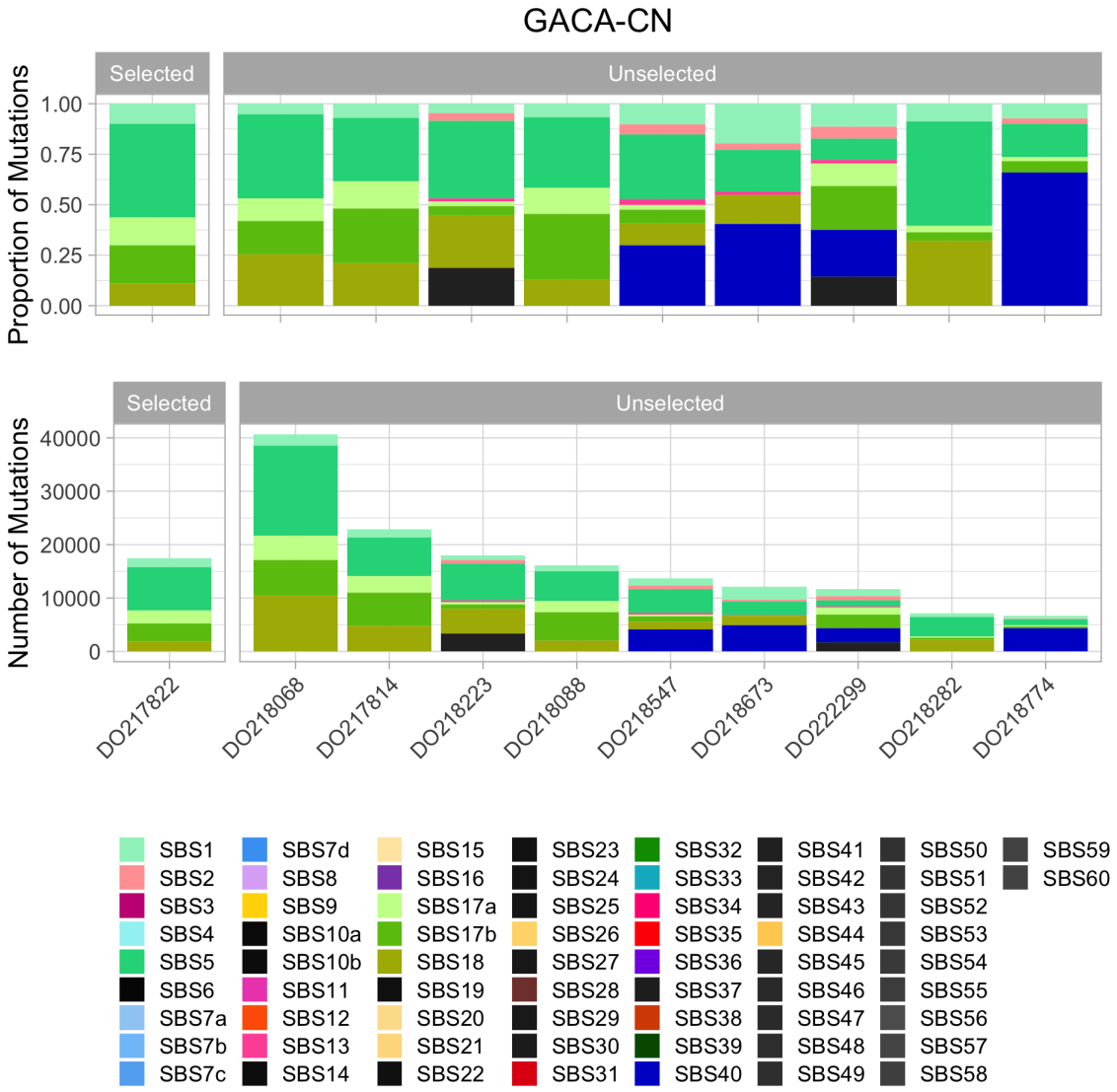


Figure C.5: Distribution of mutational signatures in selected and unselected gastric cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

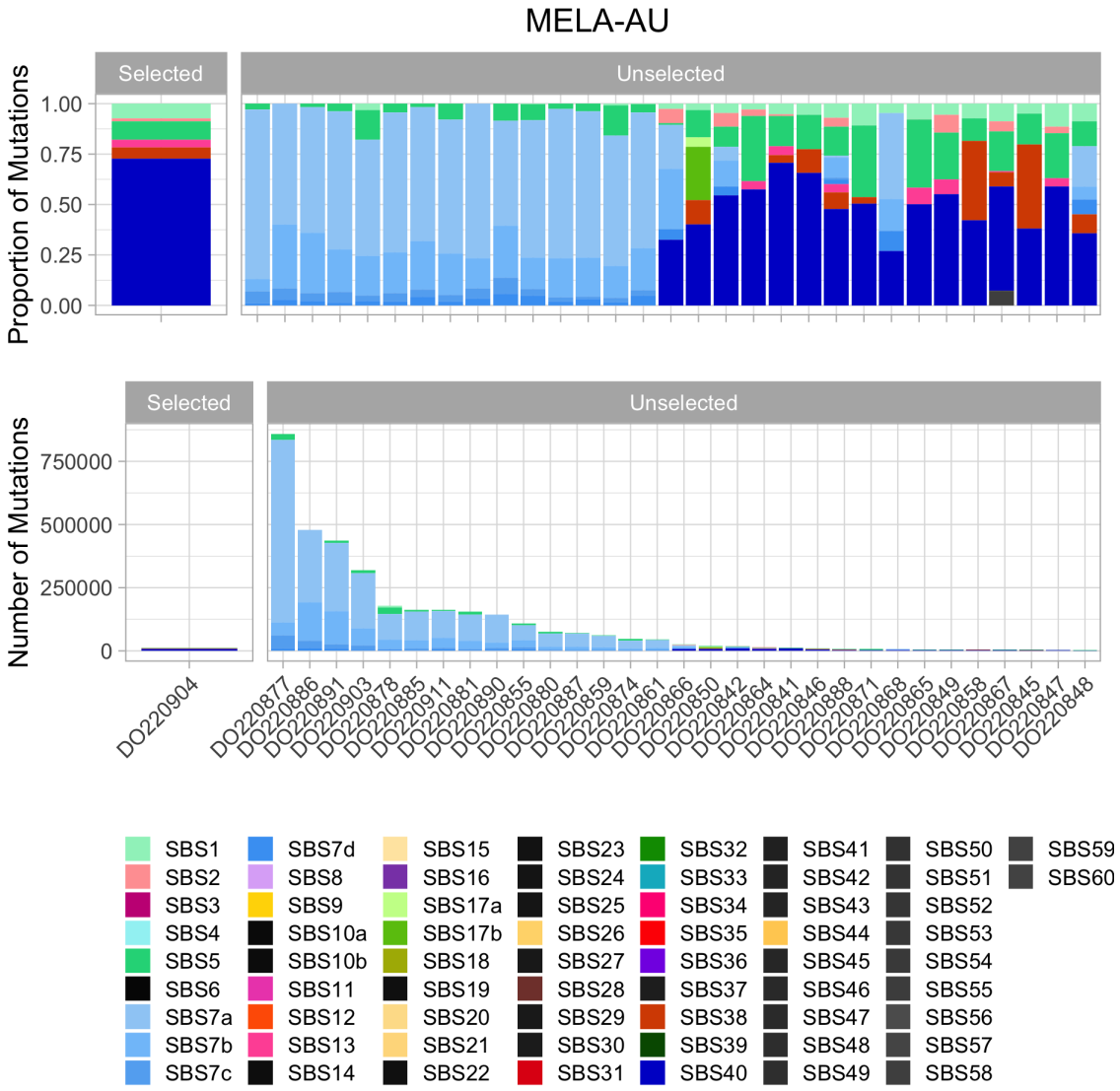


Figure C.6: Distribution of mutational signatures in selected and unselected melanoma cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

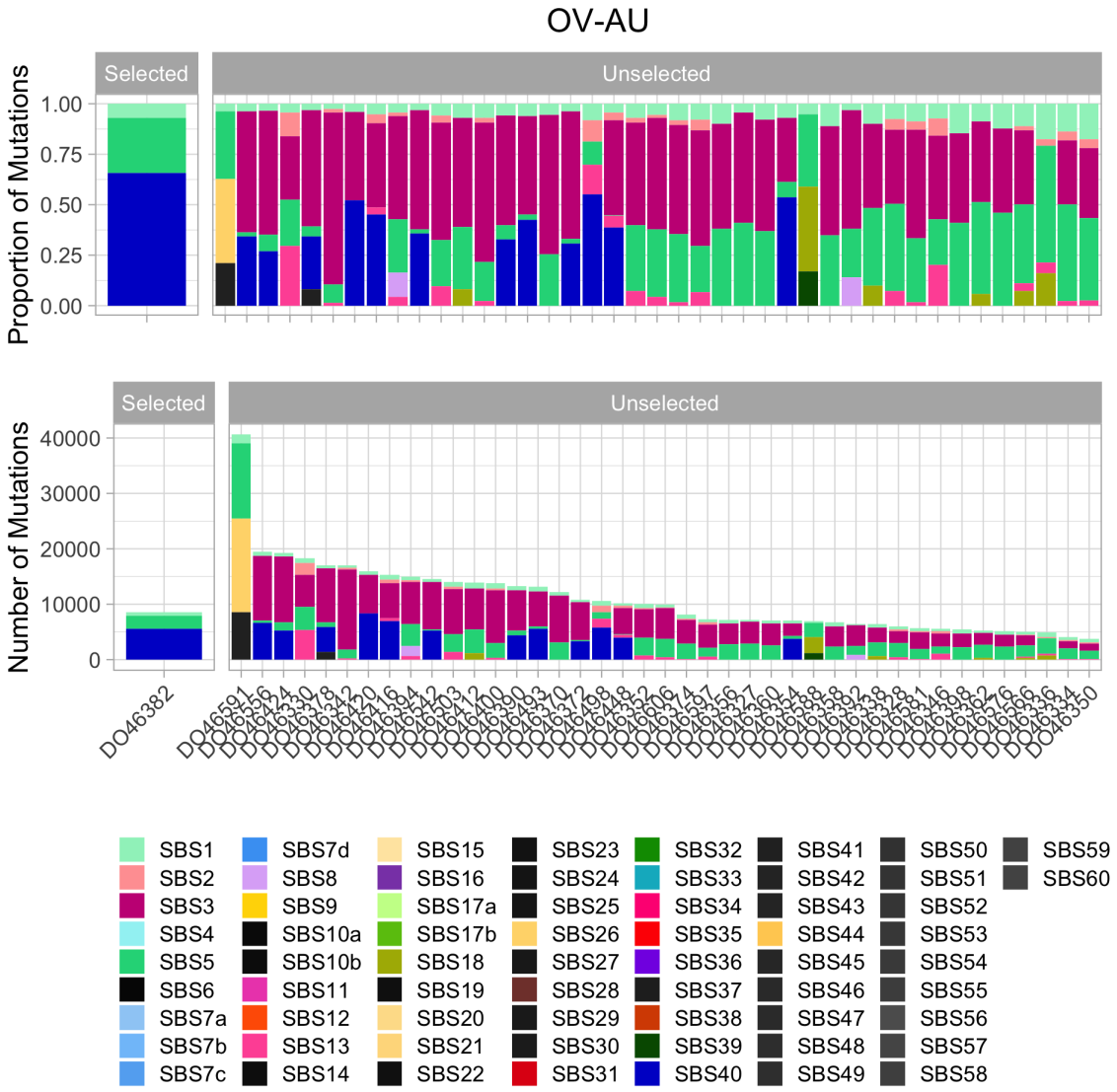


Figure C.7: Distribution of mutational signatures in selected and unselected ovarian cancers from Australia in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

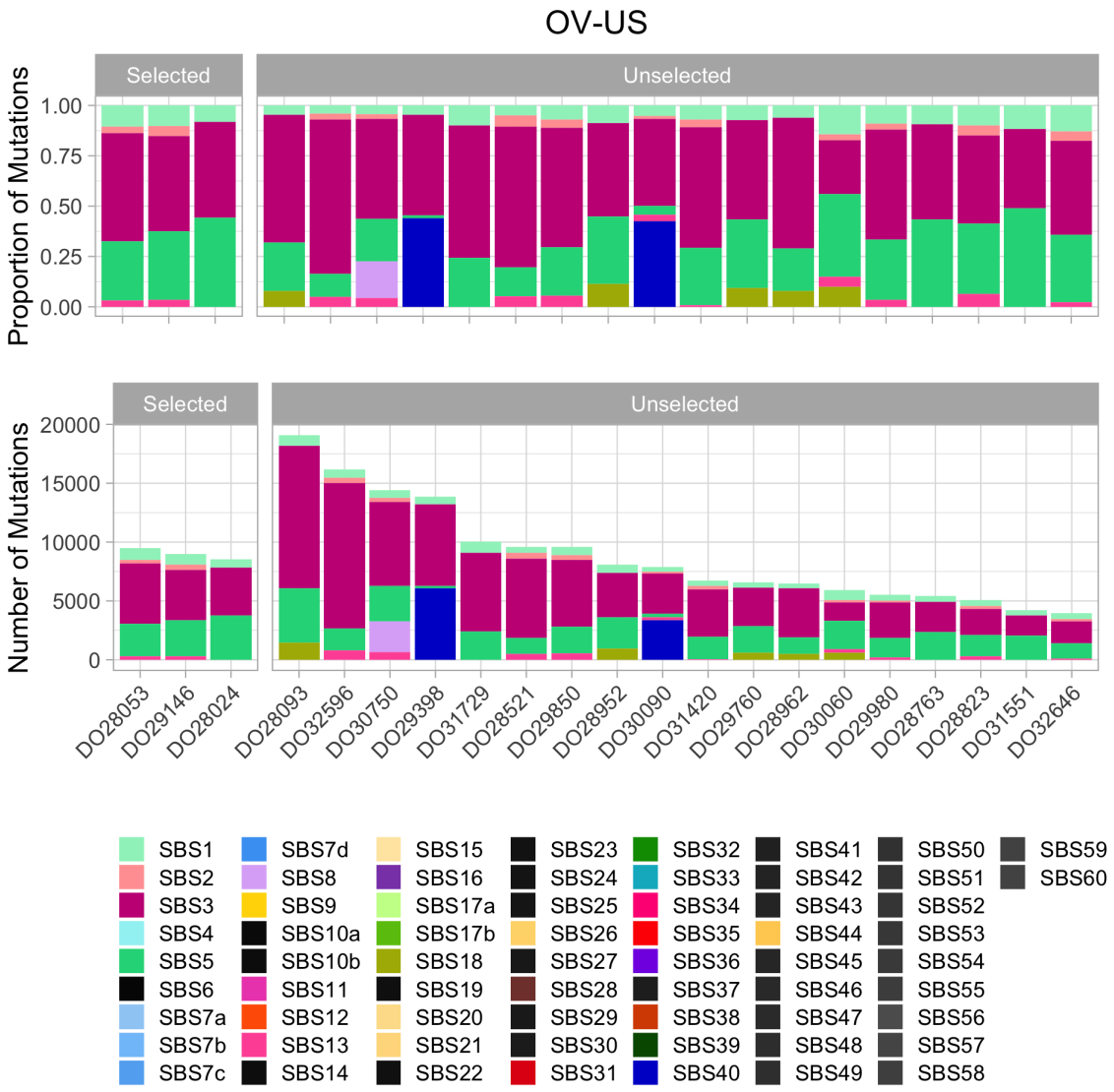


Figure C.8: Distribution of mutational signatures in selected and unselected ovarian cancers from the United States of America in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

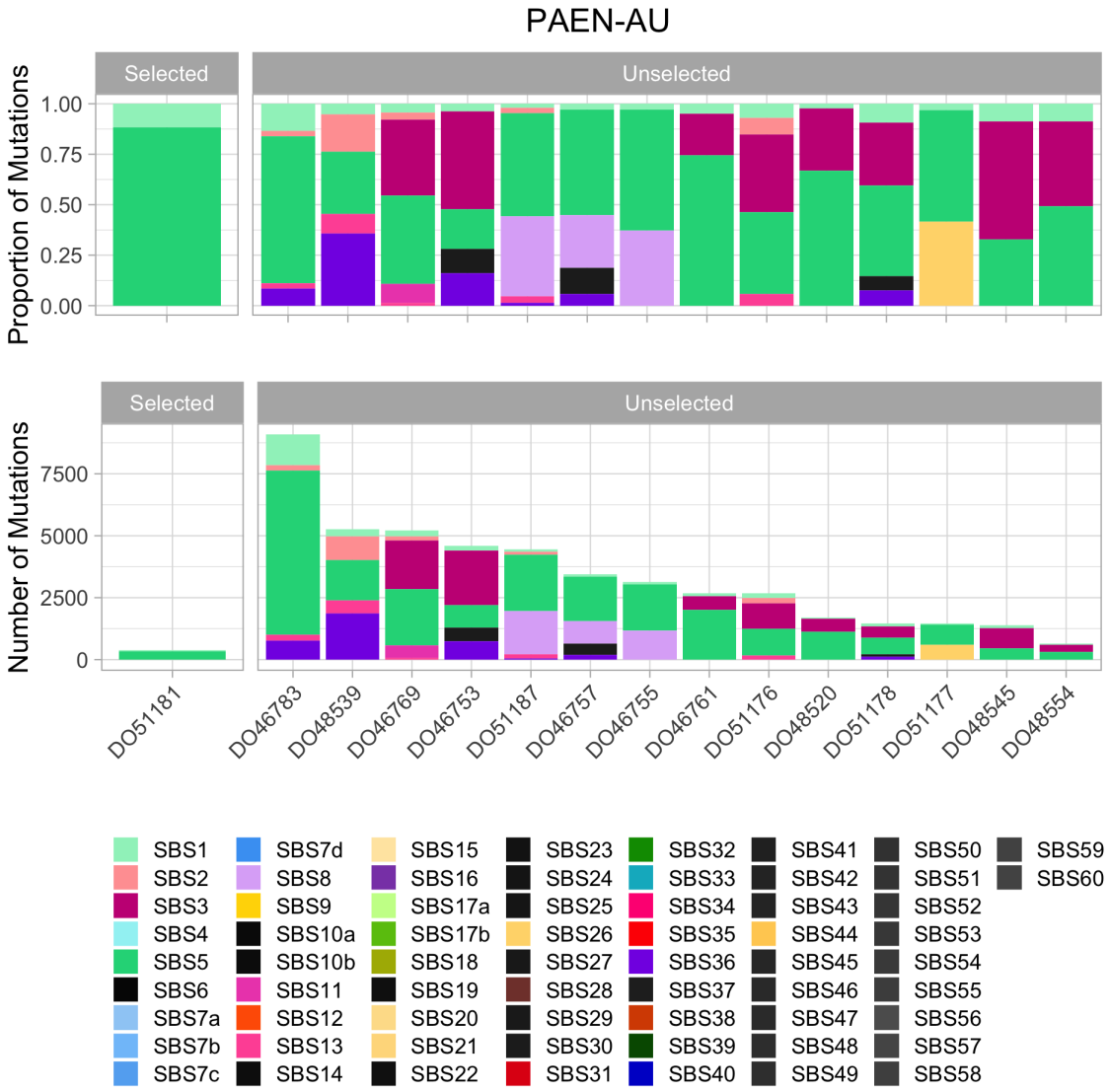


Figure C.9: Distribution of mutational signatures in selected and unselected pancreatic neuroendocrine neoplasms in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

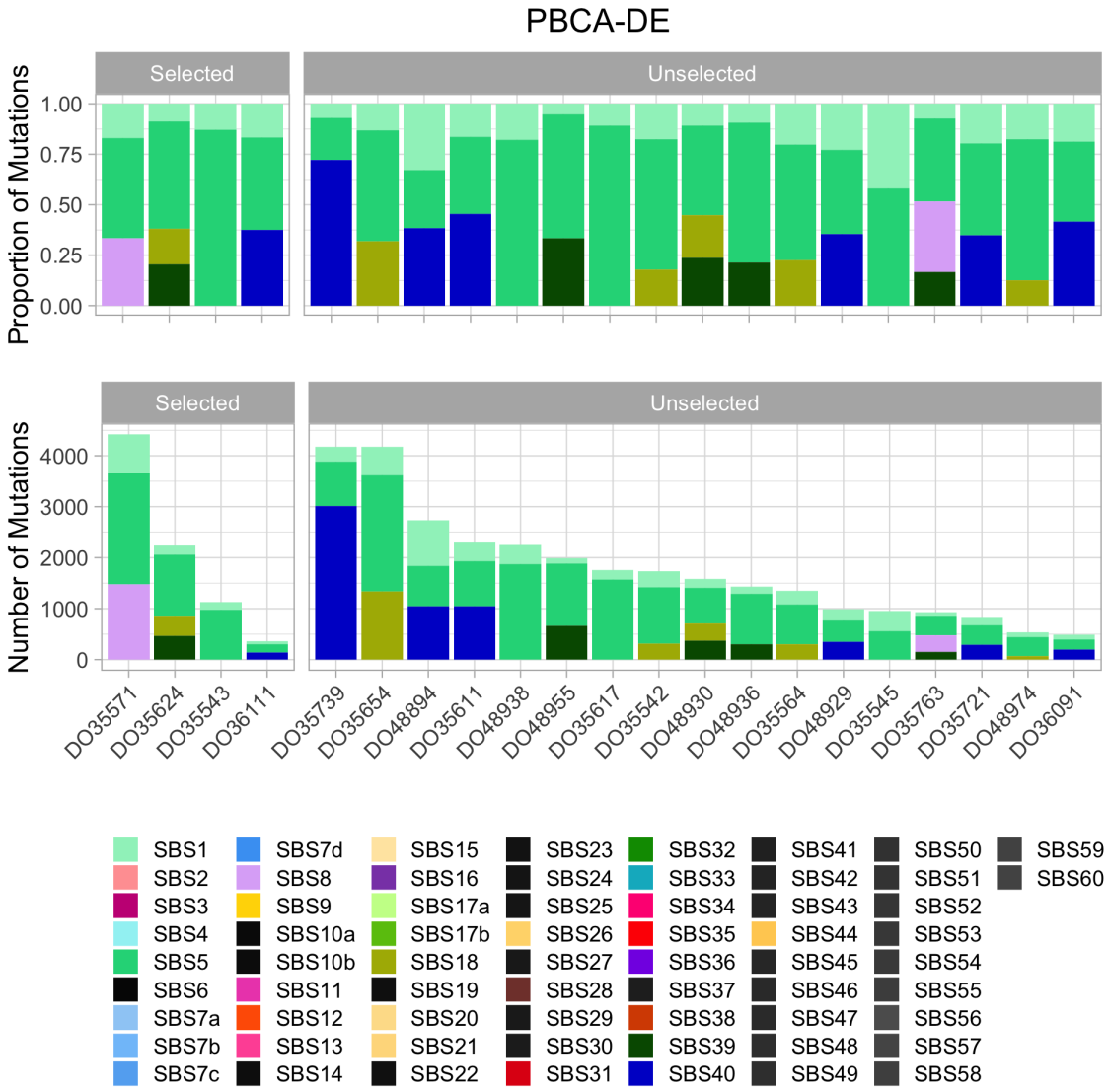


Figure C.10: Distribution of mutational signatures in selected and unselected paediatric brain cancers in the PCAWG data set. Only tetraploid samples for which fusion measures could be calculated were considered.

C.2.2 Selected PCAWG Samples

In the following figures, opacity and bar width has been adjusted to aid visualisation of overlapping or very close segments.

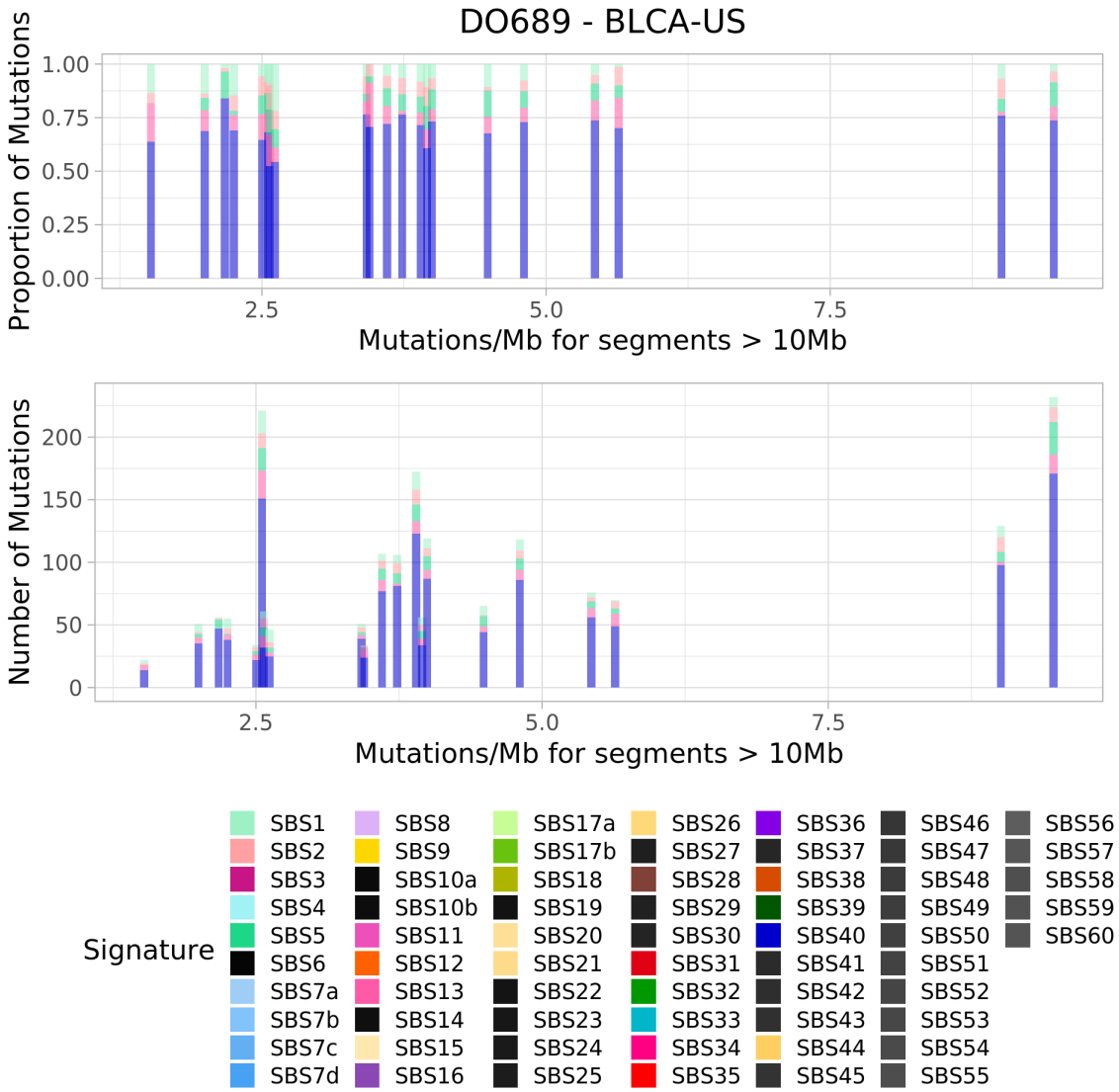


Figure C.11: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample BLCA - DO689.

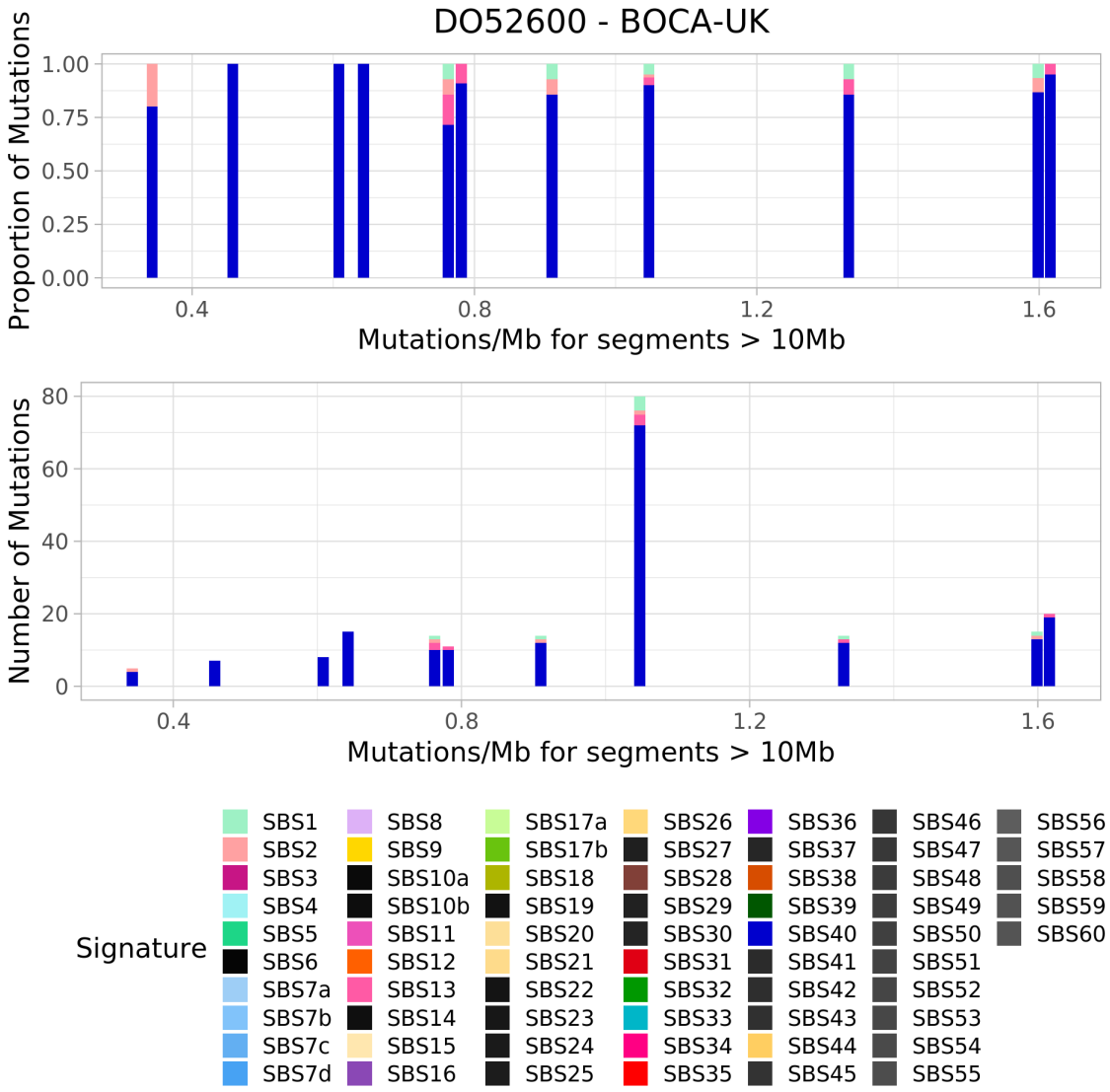


Figure C.12: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample BOCA - DO52600.

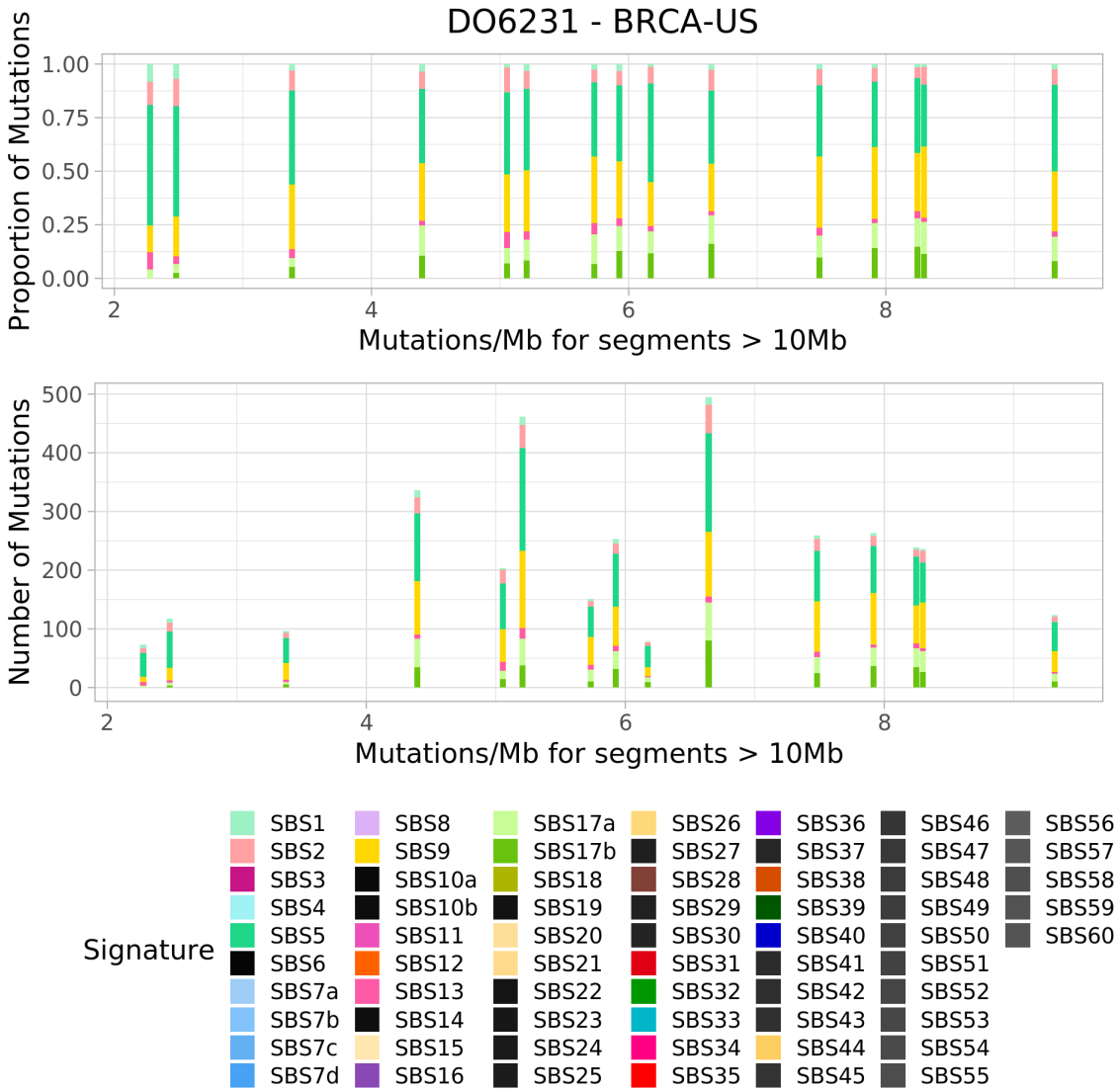


Figure C.13: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample BRCA - DO6231.

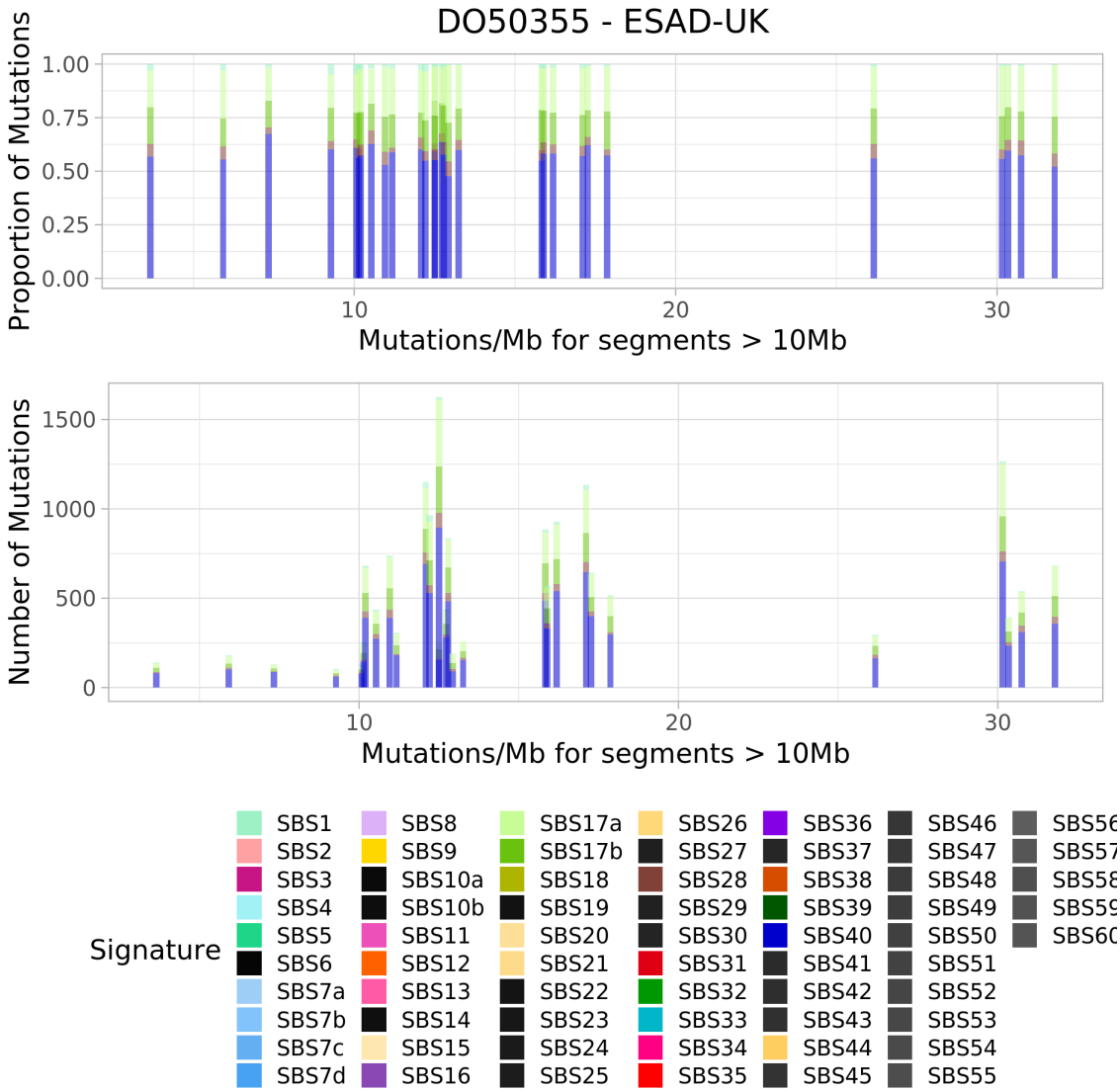


Figure C.14: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample ESAD - DO50355.

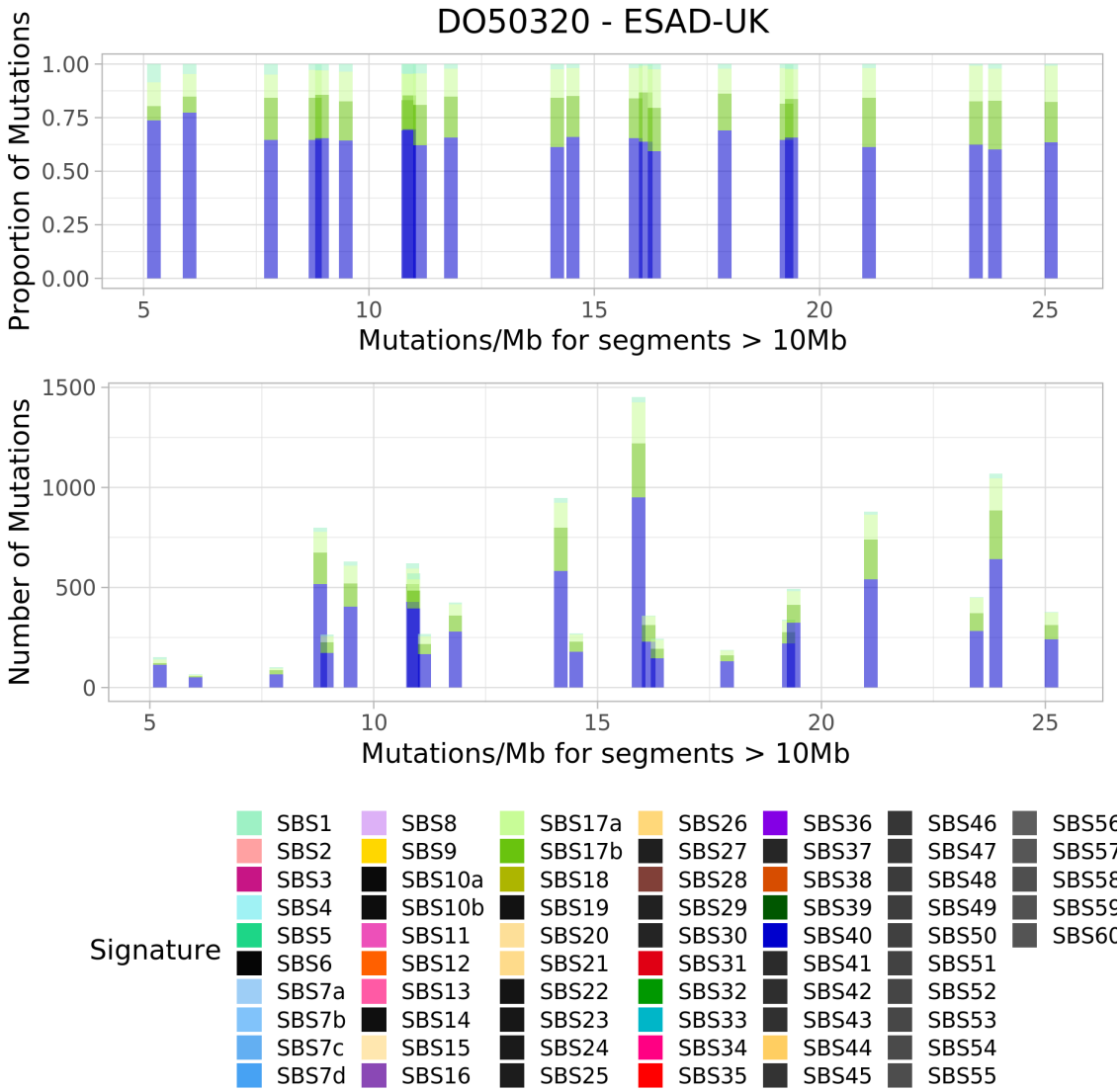


Figure C.15: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample ESAD - DO50320.

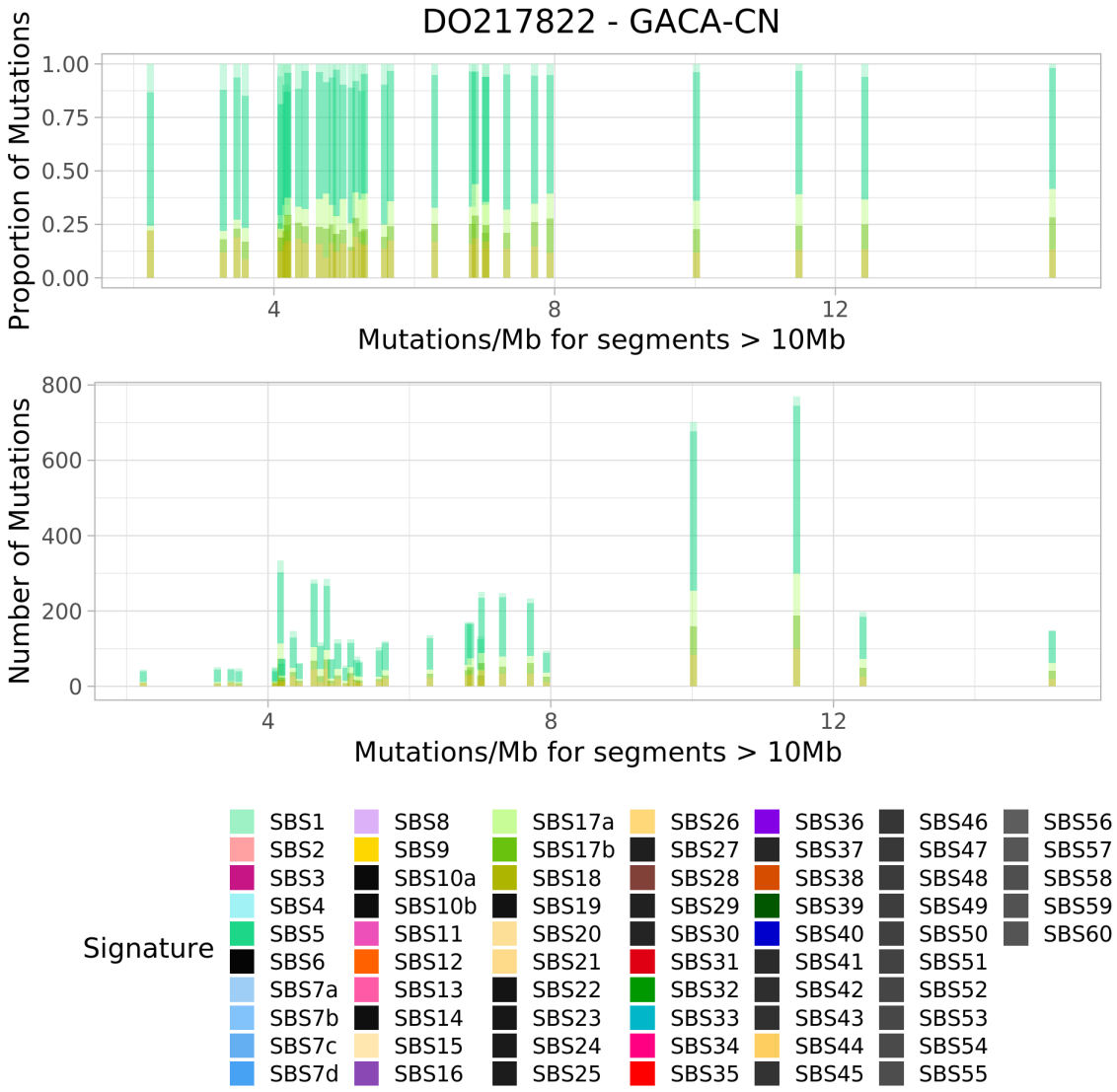


Figure C.16: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample GACA - DO217822.

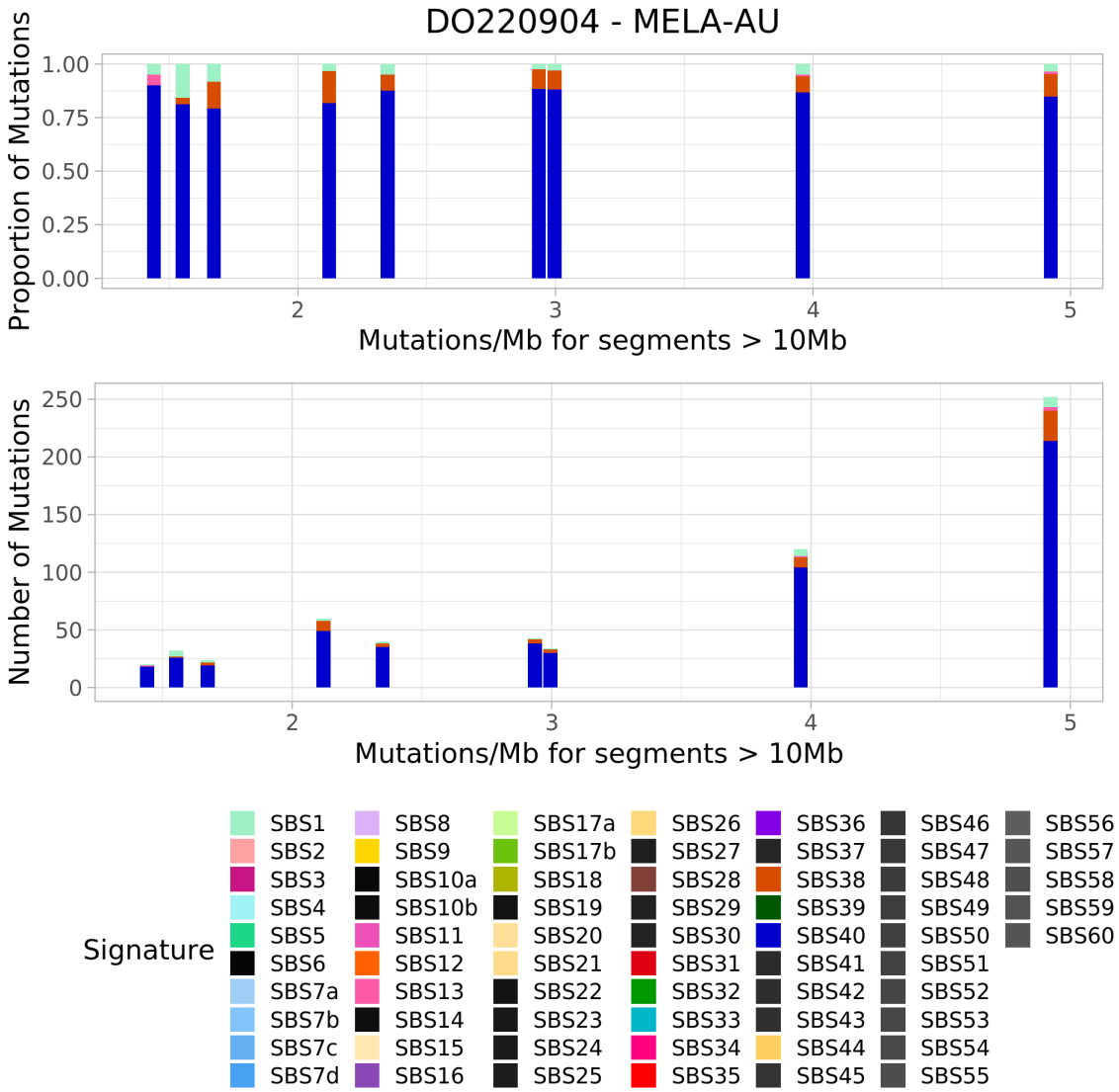


Figure C.17: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample MELA - DO220904.

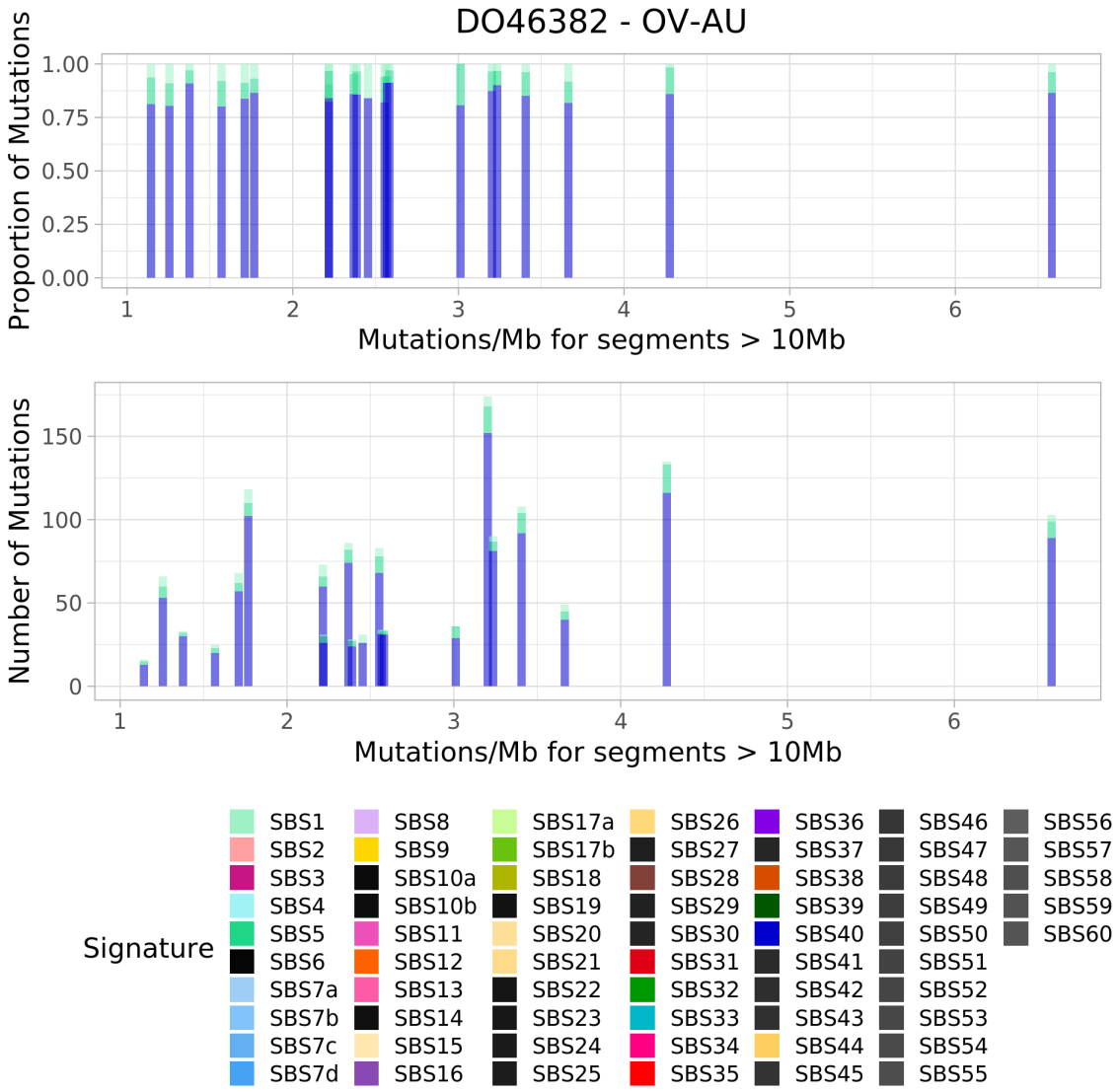


Figure C.18: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample OV - DO46382.

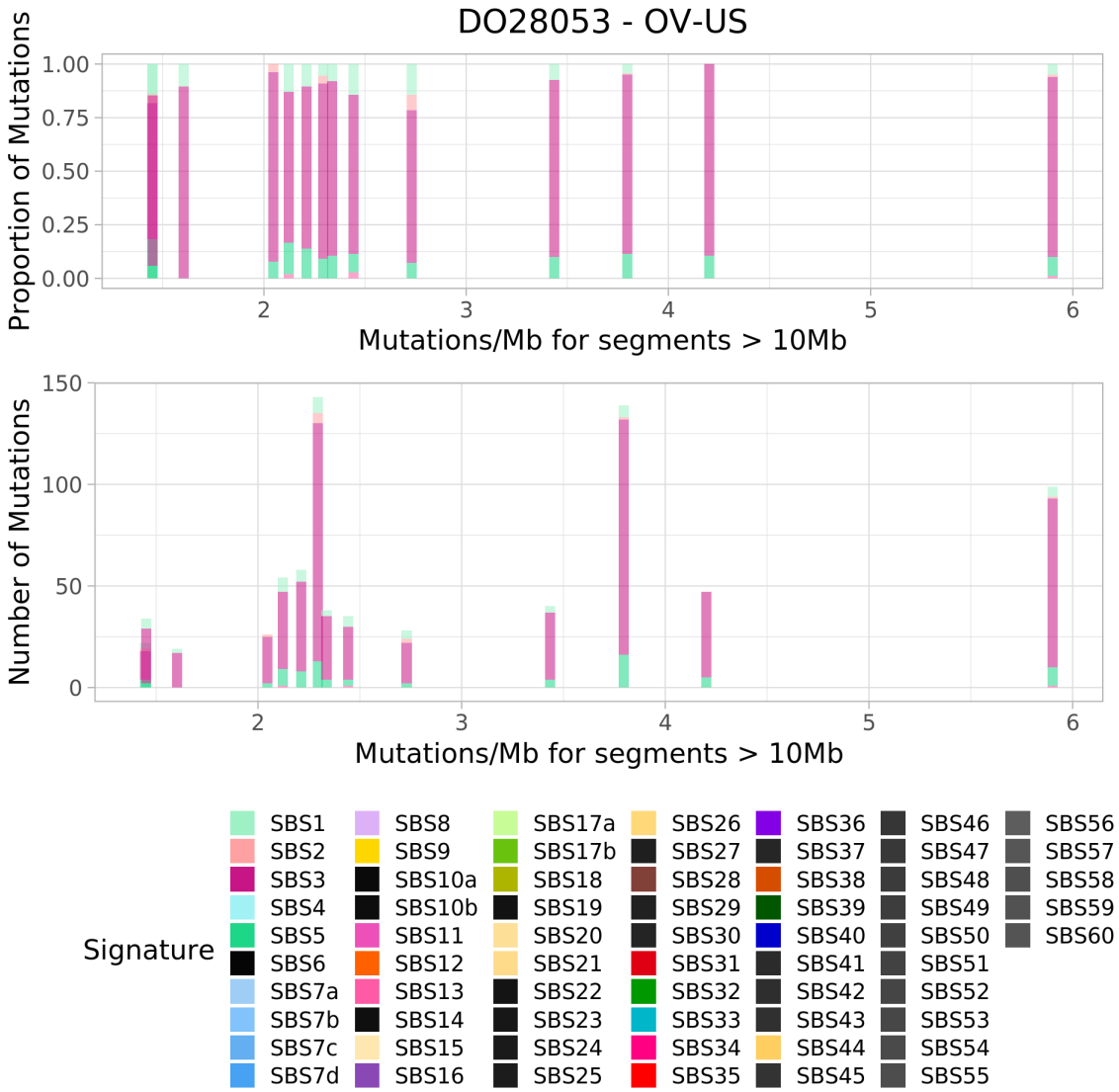


Figure C.19: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample OV - DO28053.

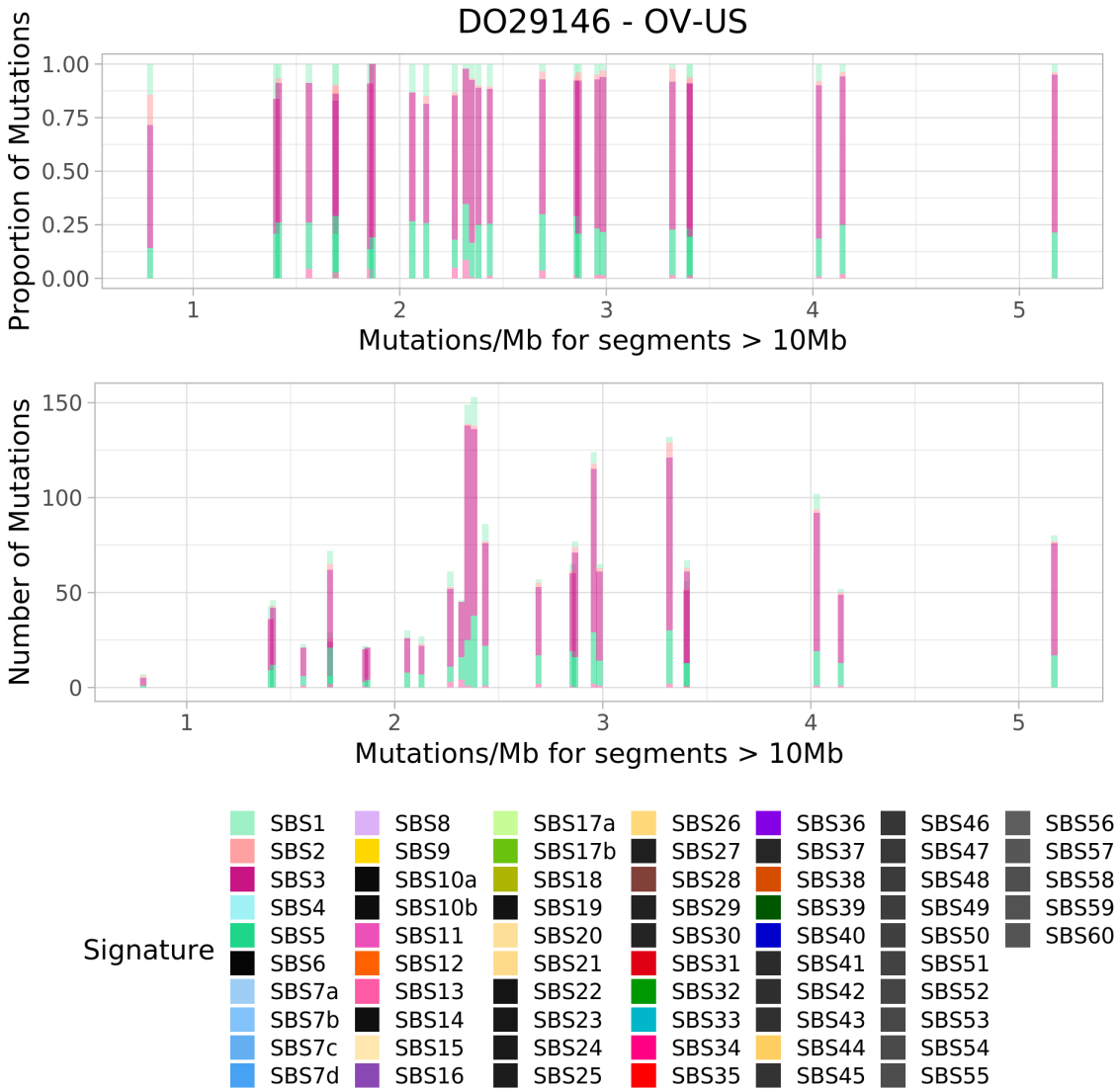


Figure C.20: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample OV - DO29146.

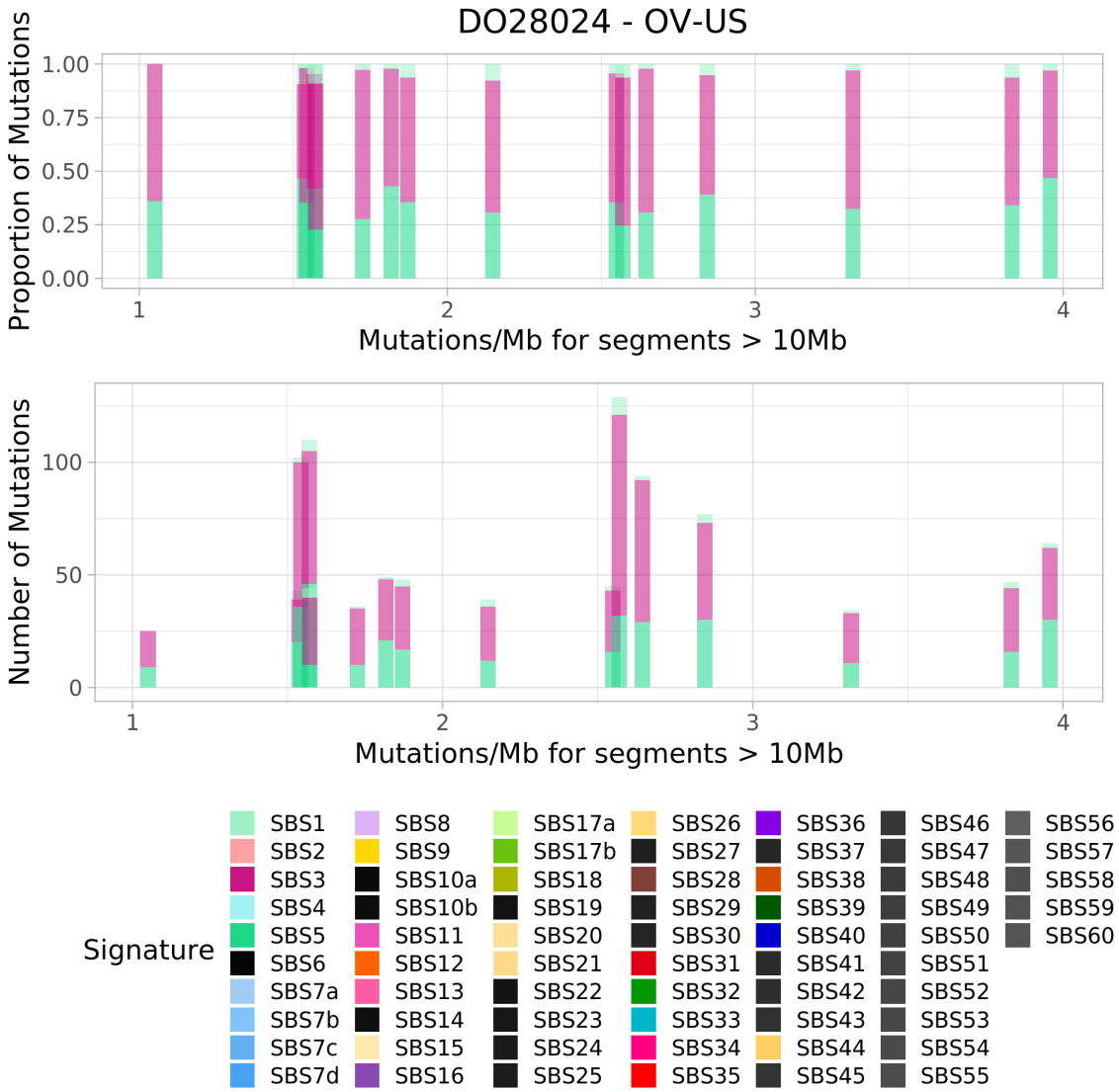


Figure C.21: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample OV - DO28024.

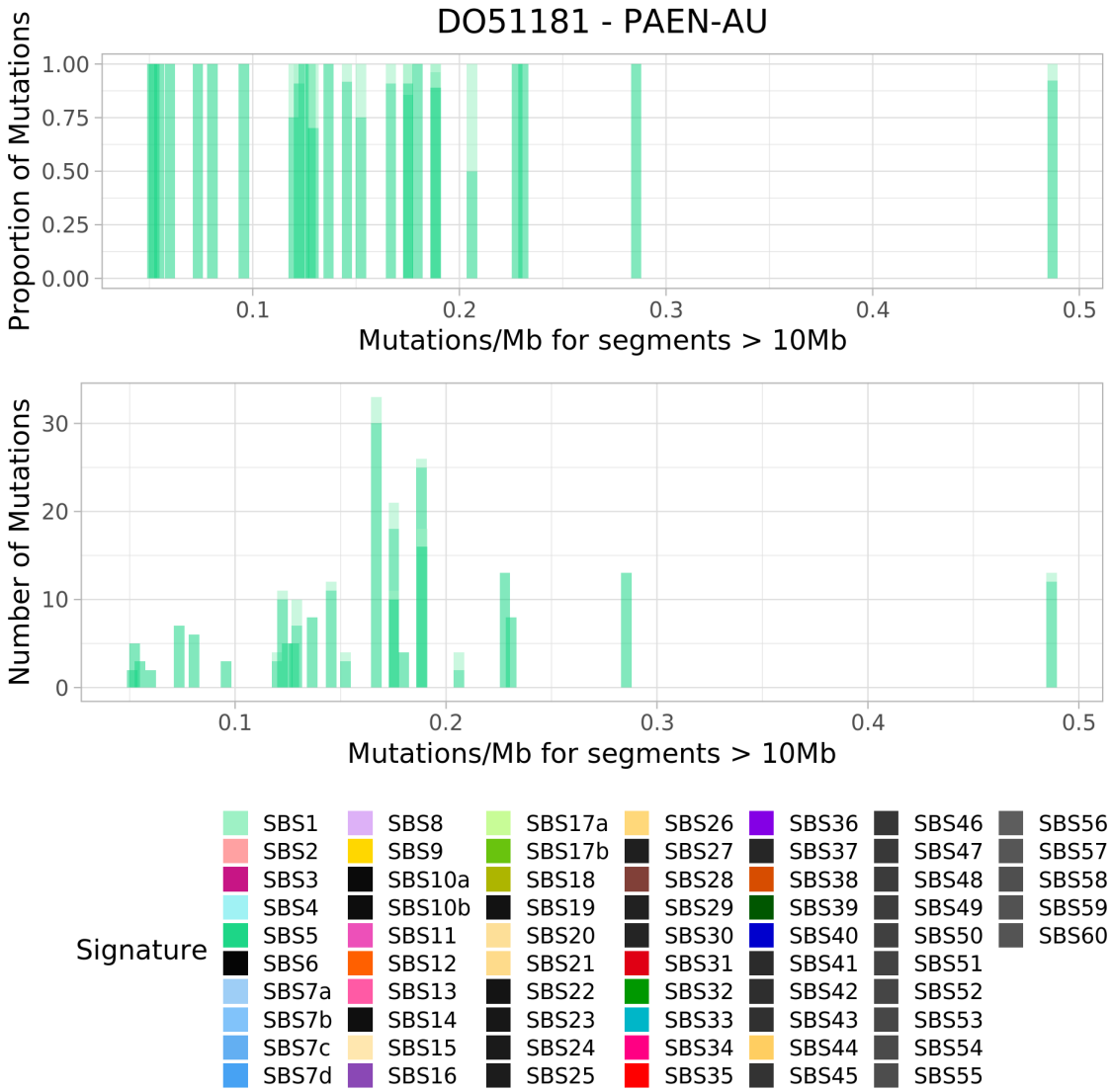


Figure C.22: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample PAEN - DO51181.

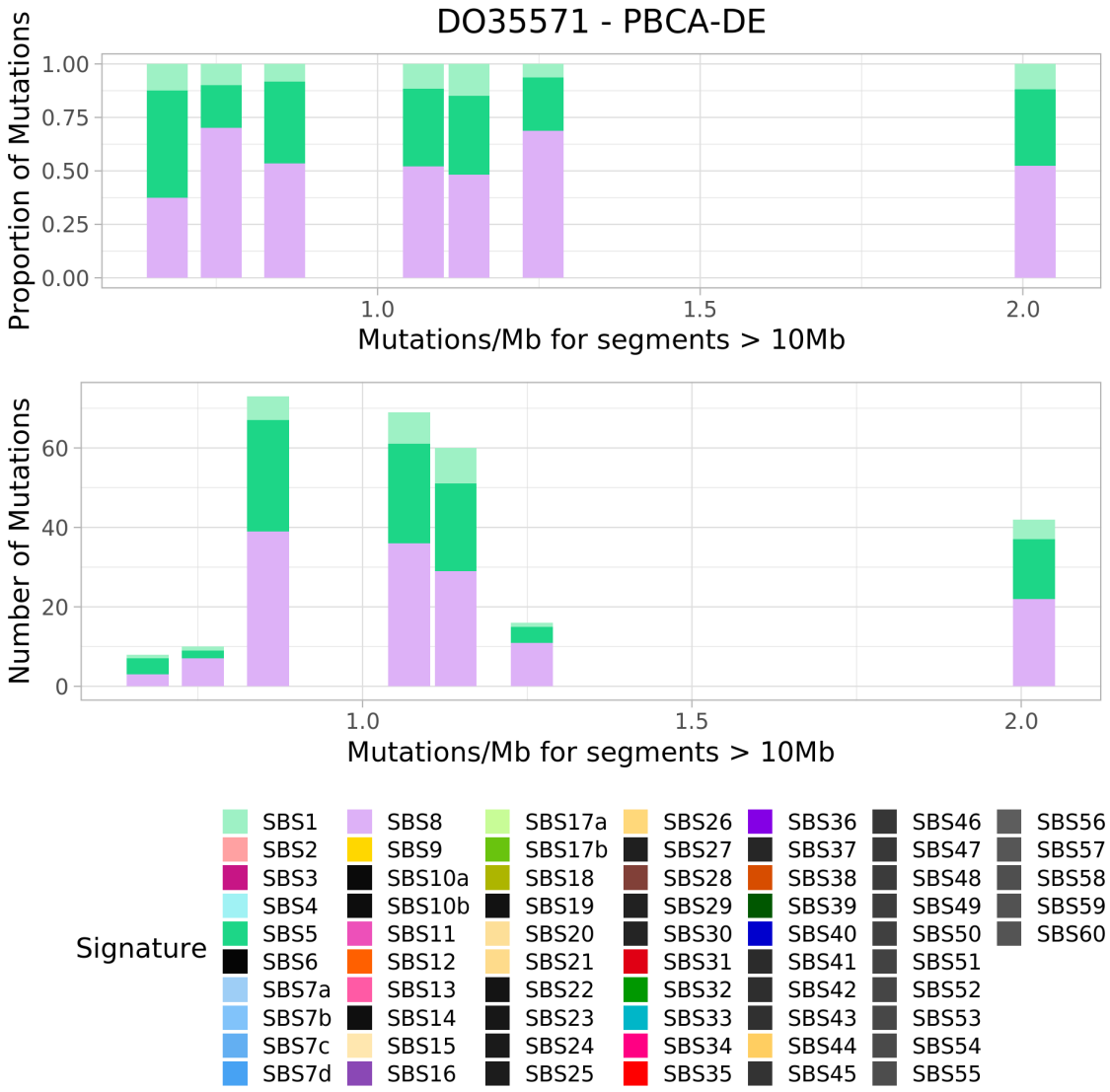


Figure C.23: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample PBCA - DO35571.

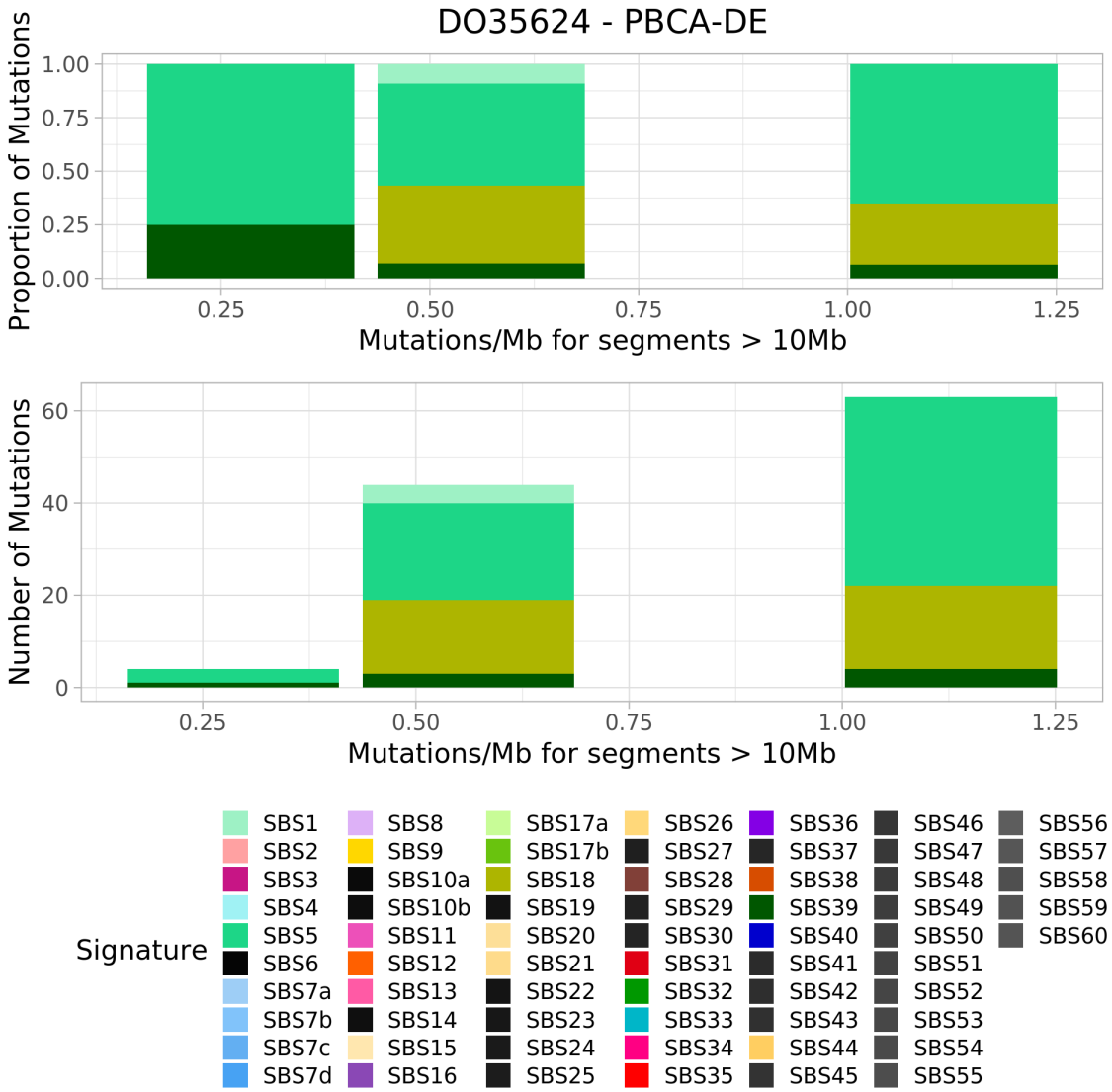


Figure C.24: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample PBCA - DO35624.

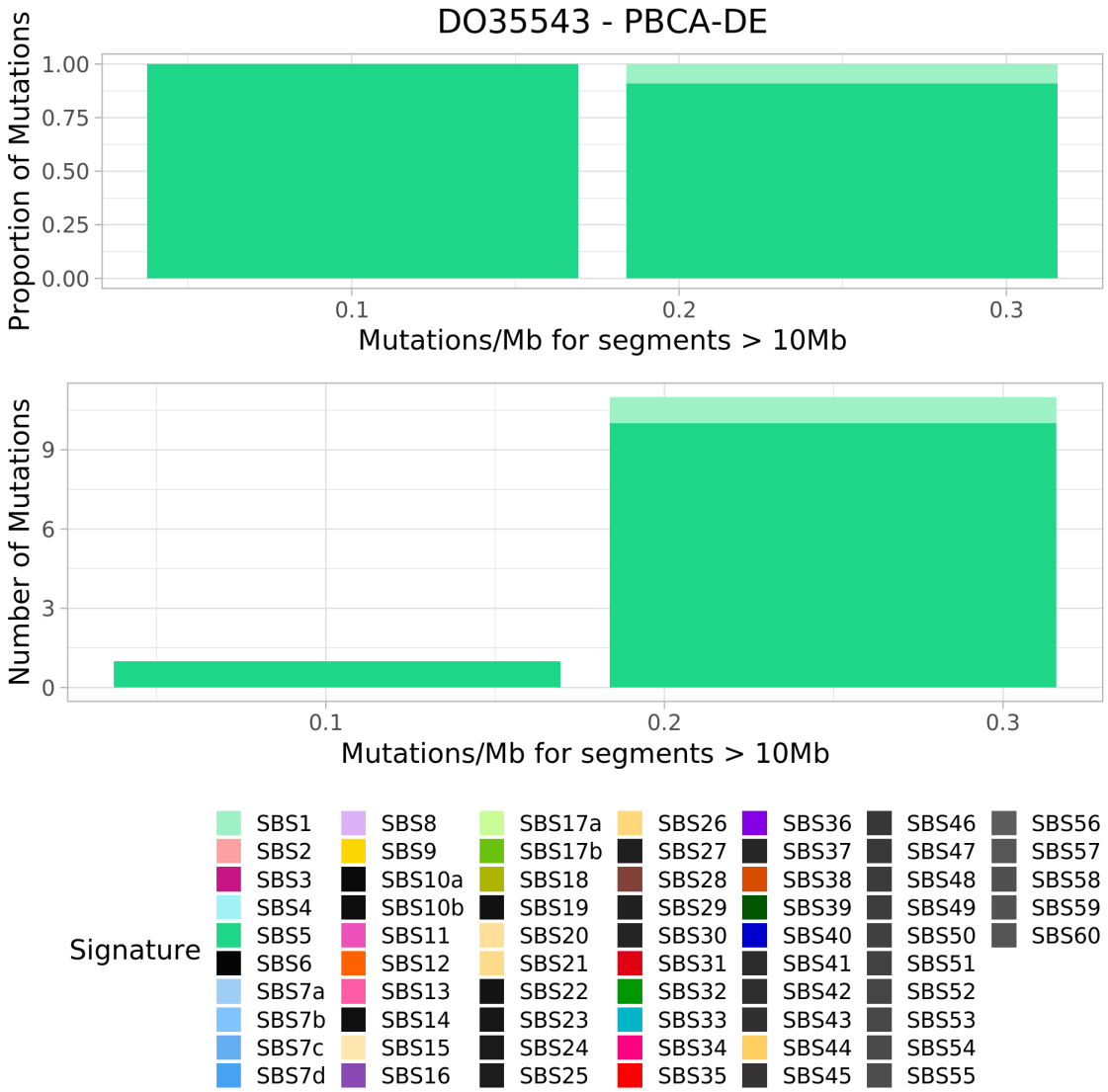


Figure C.25: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample PBCA - DO35543.

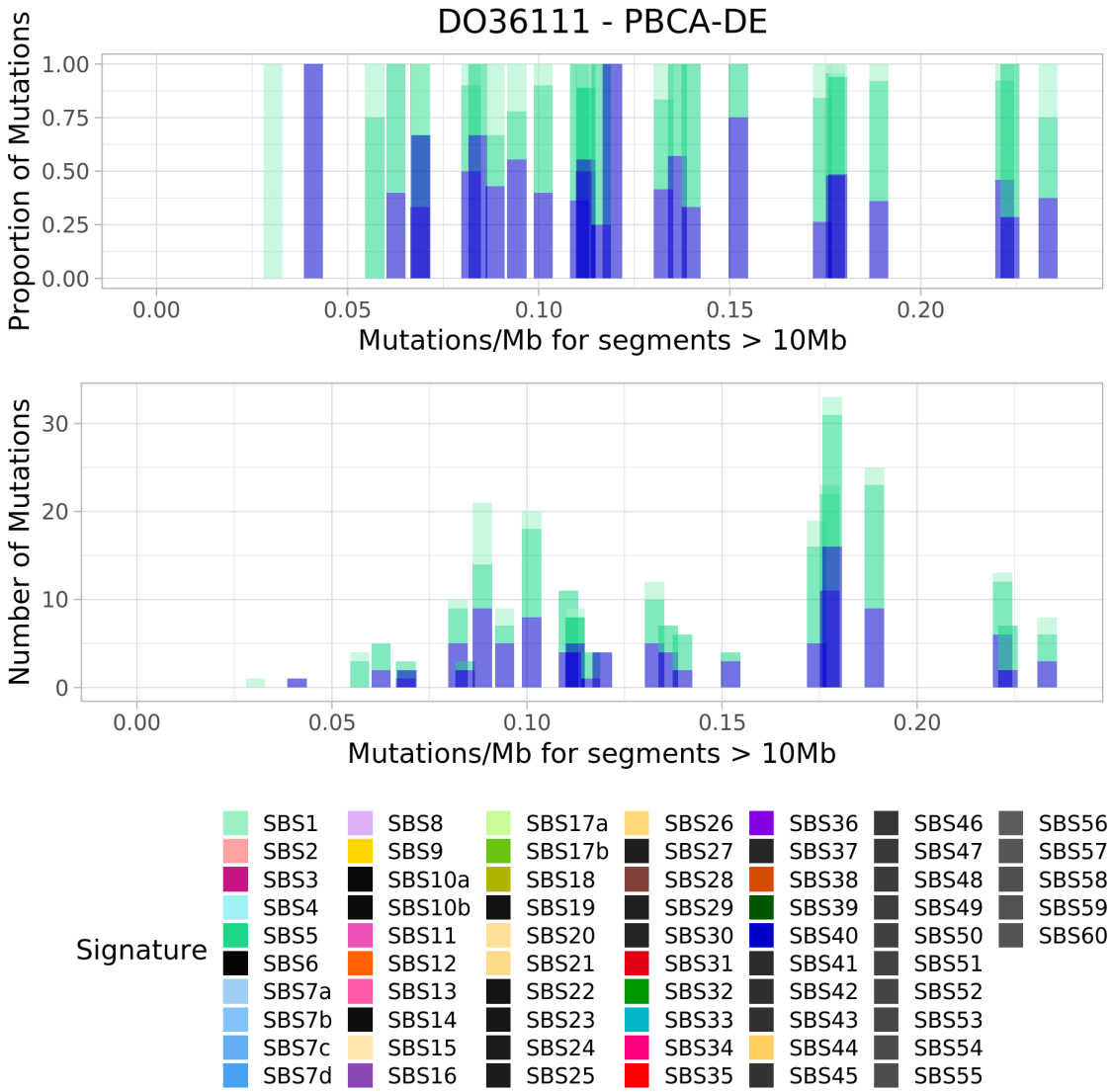


Figure C.26: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in PCAWG sample DO36111 - PBCA.

C.2.3 Hartwig Tumour Types

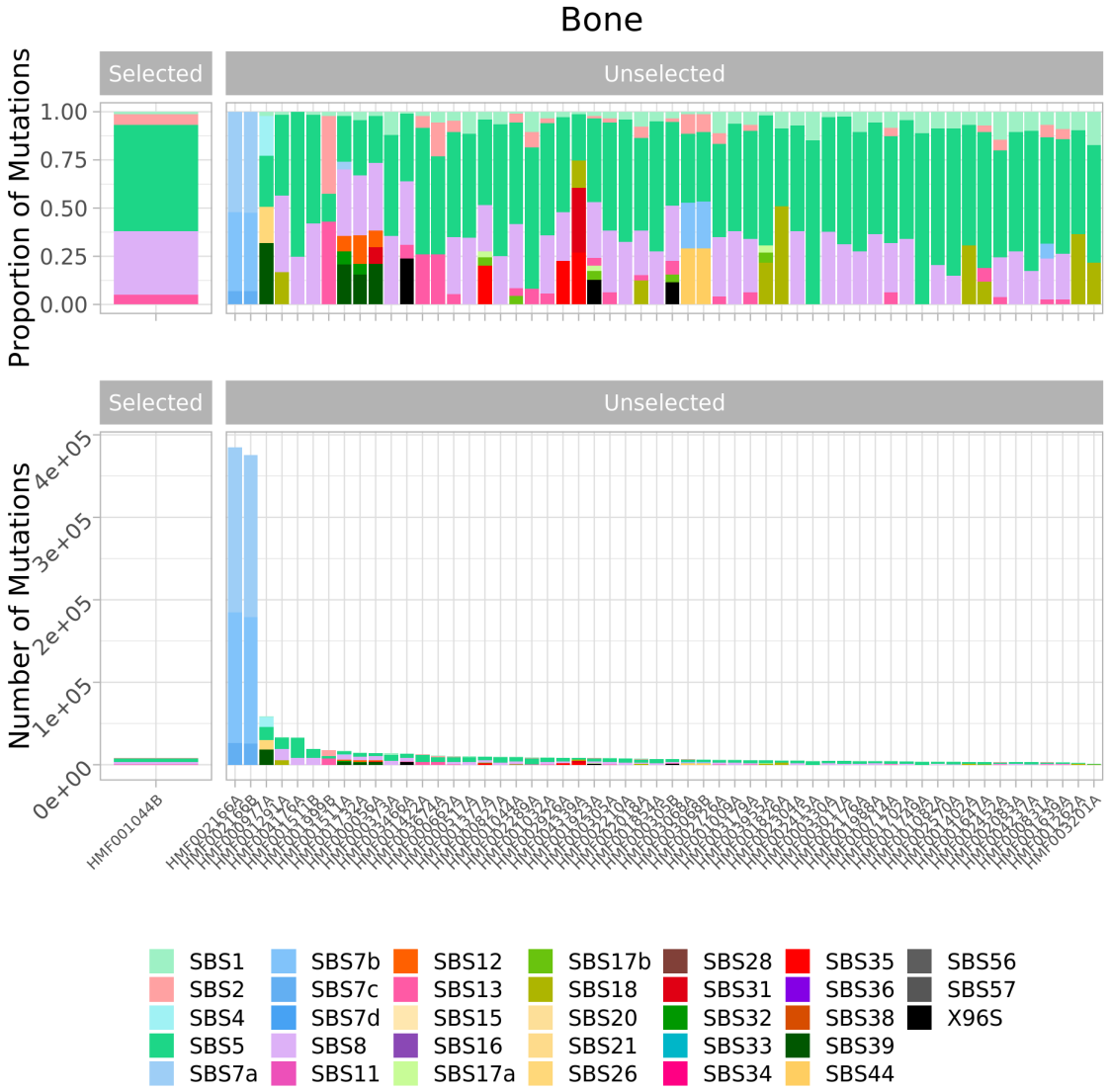


Figure C.27: Distribution of mutational signatures in selected and unselected bone cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

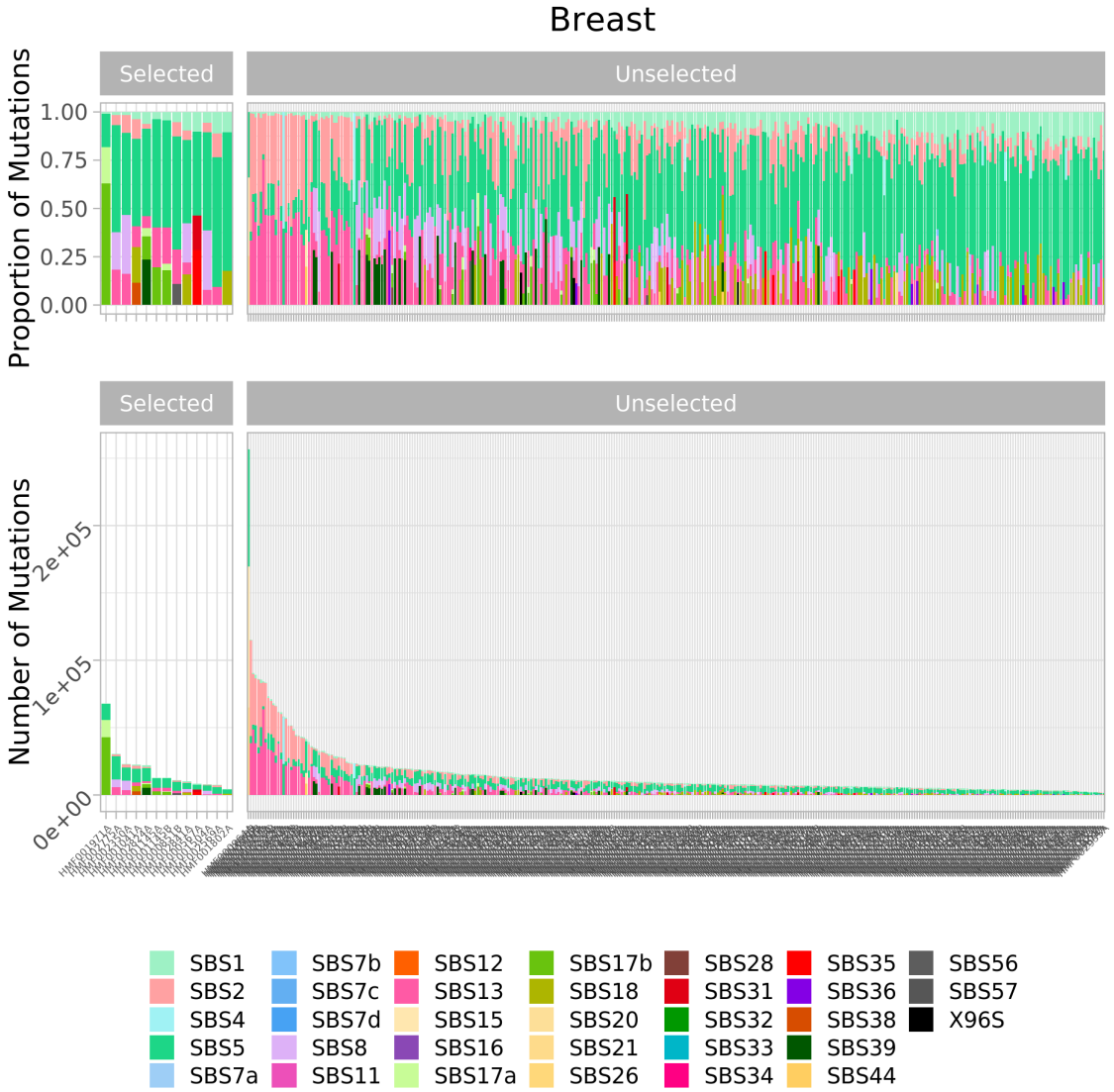


Figure C.28: Distribution of mutational signatures in selected and unselected breast cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

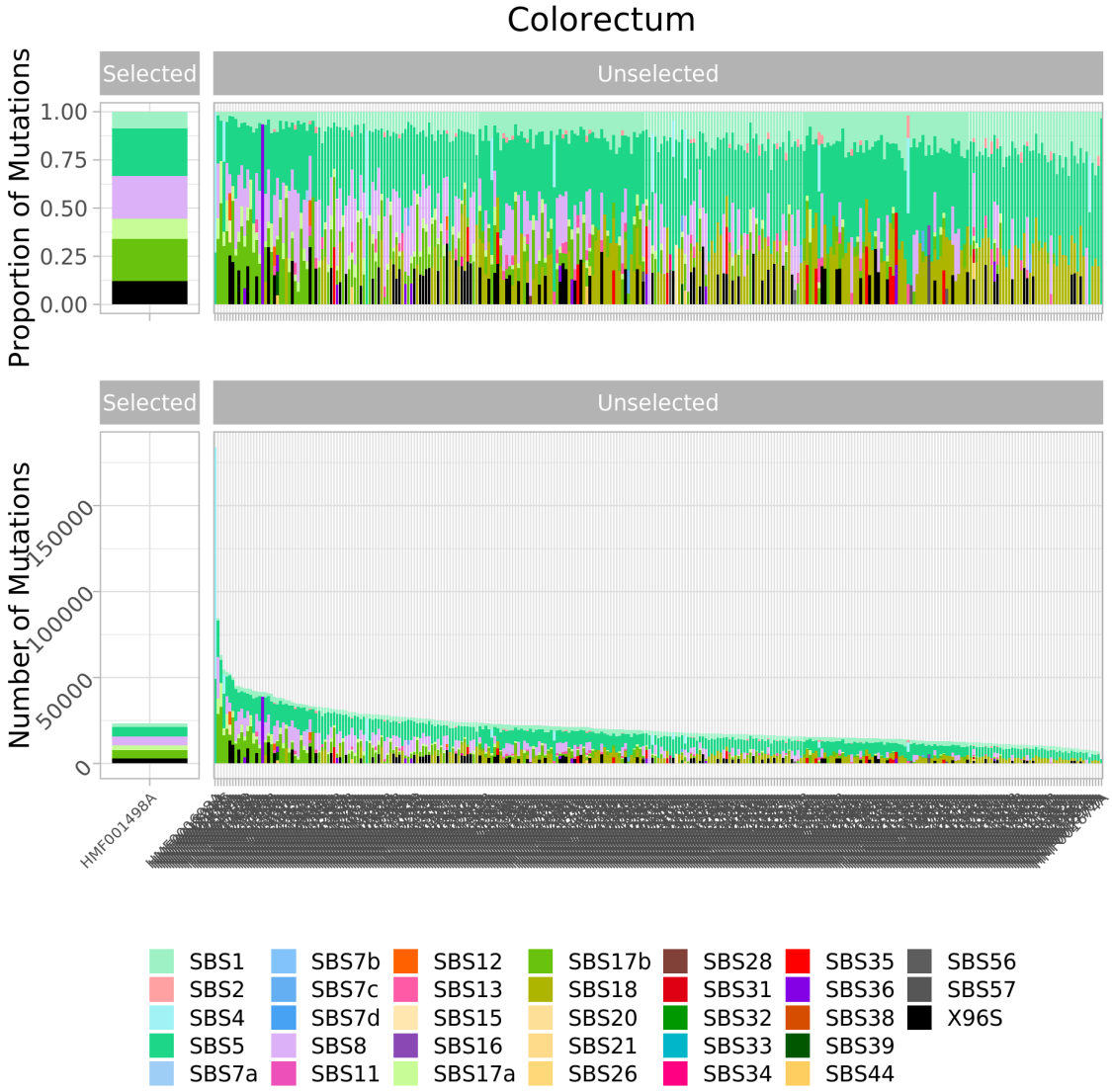


Figure C.29: Distribution of mutational signatures in selected and unselected colorectal cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

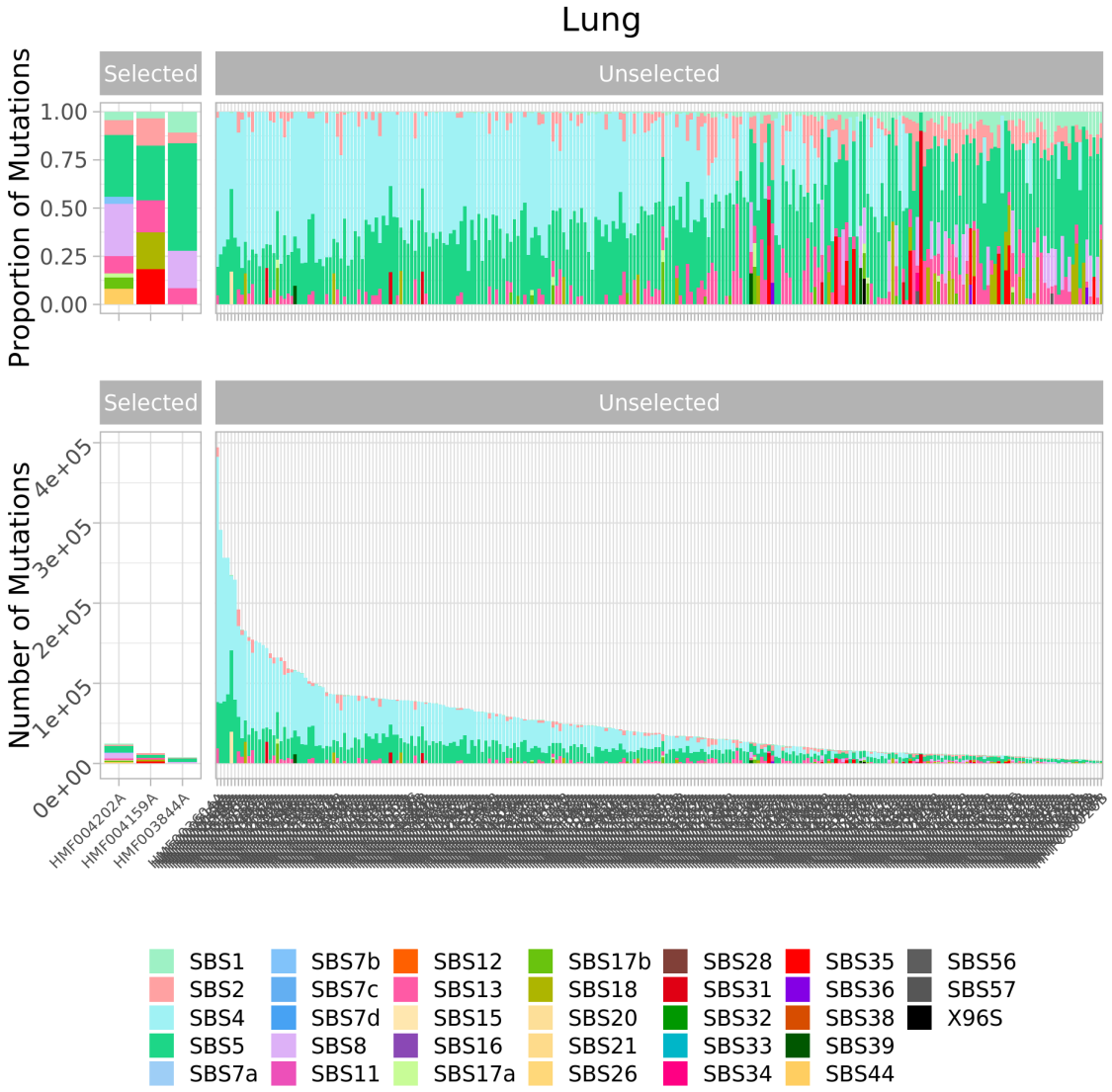


Figure C.30: Distribution of mutational signatures in selected and unselected lung cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

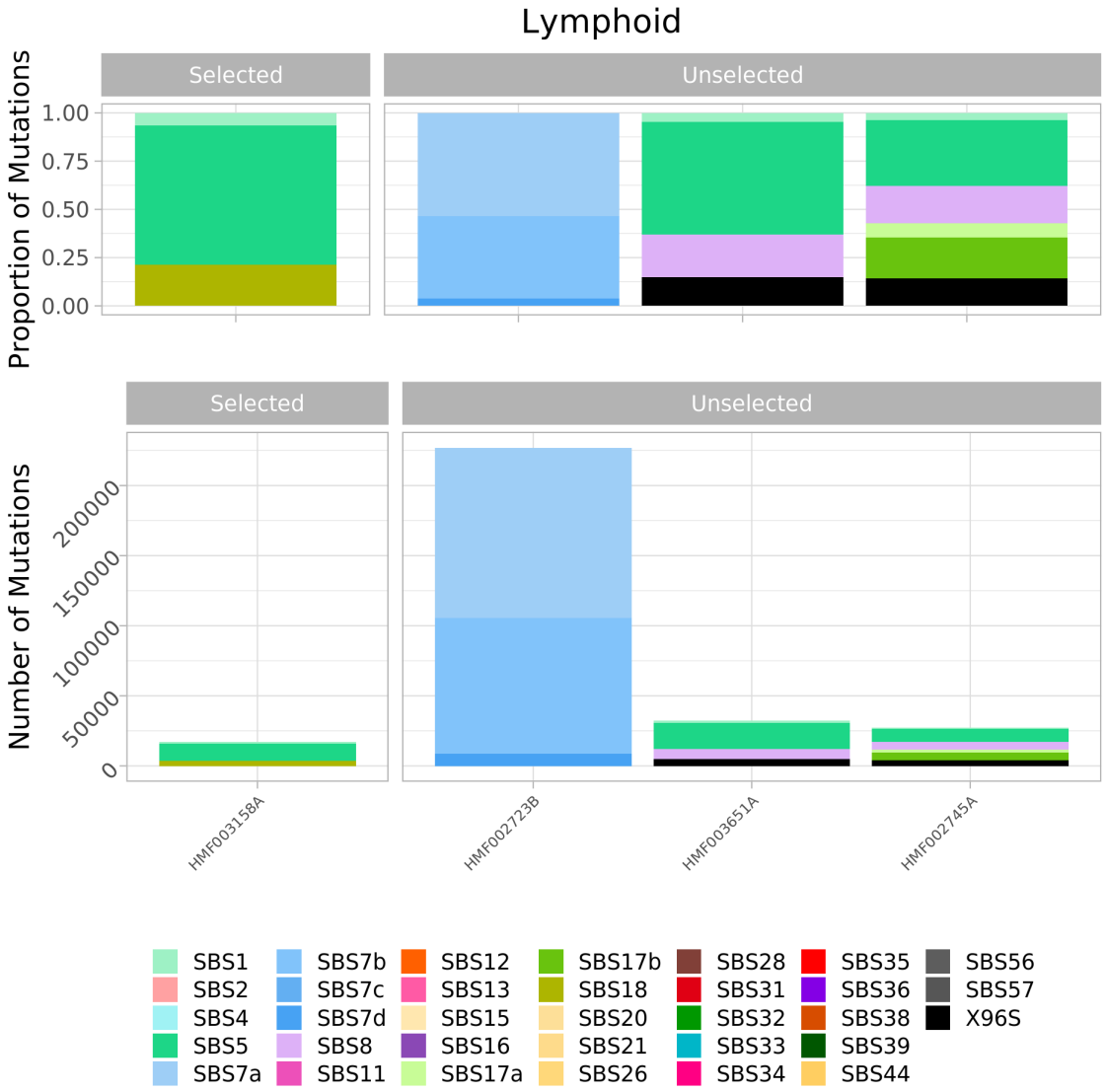


Figure C.31: Distribution of mutational signatures in selected and unselected lymphoid cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

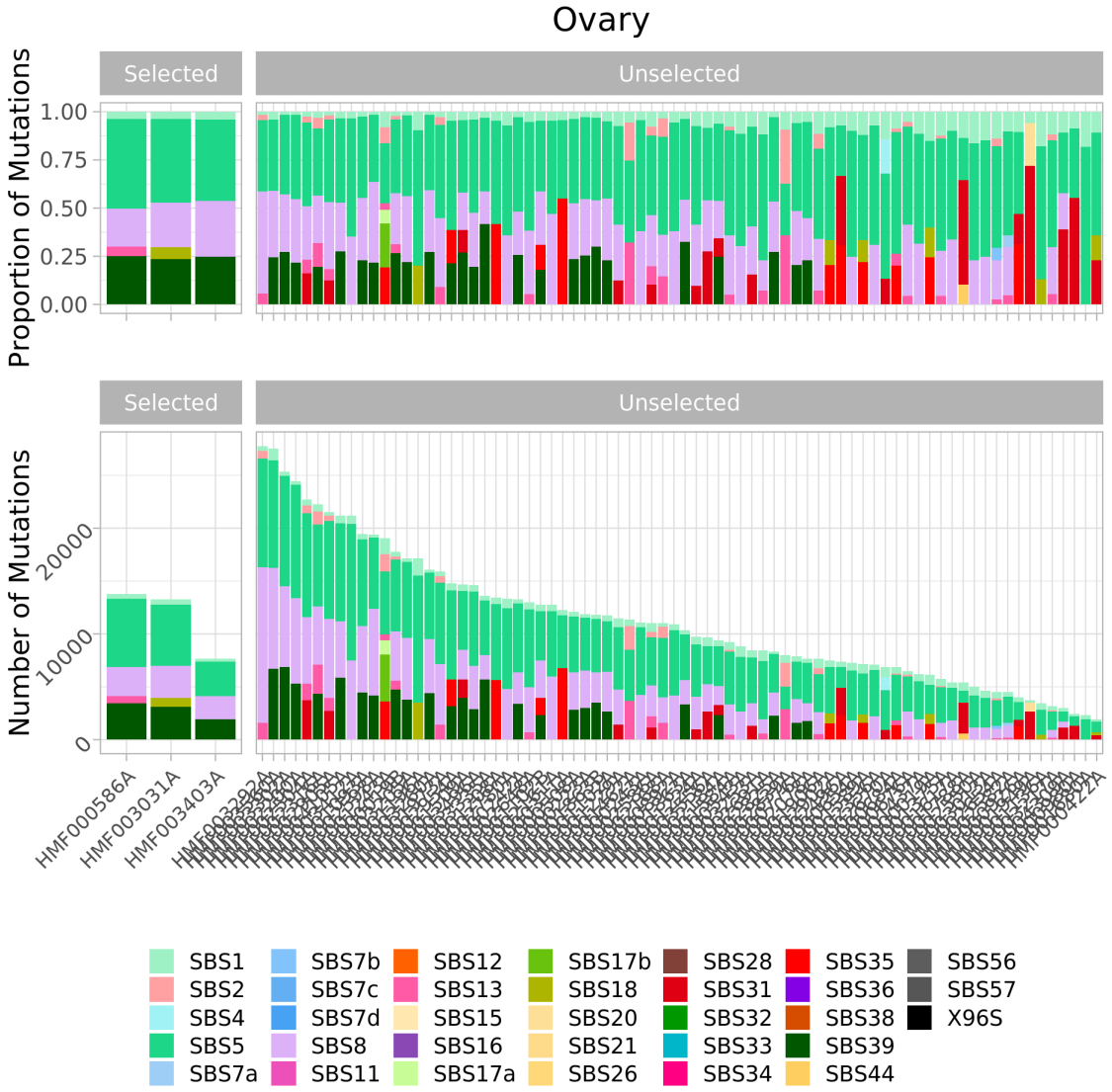


Figure C.32: Distribution of mutational signatures in selected and unselected ovarian cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

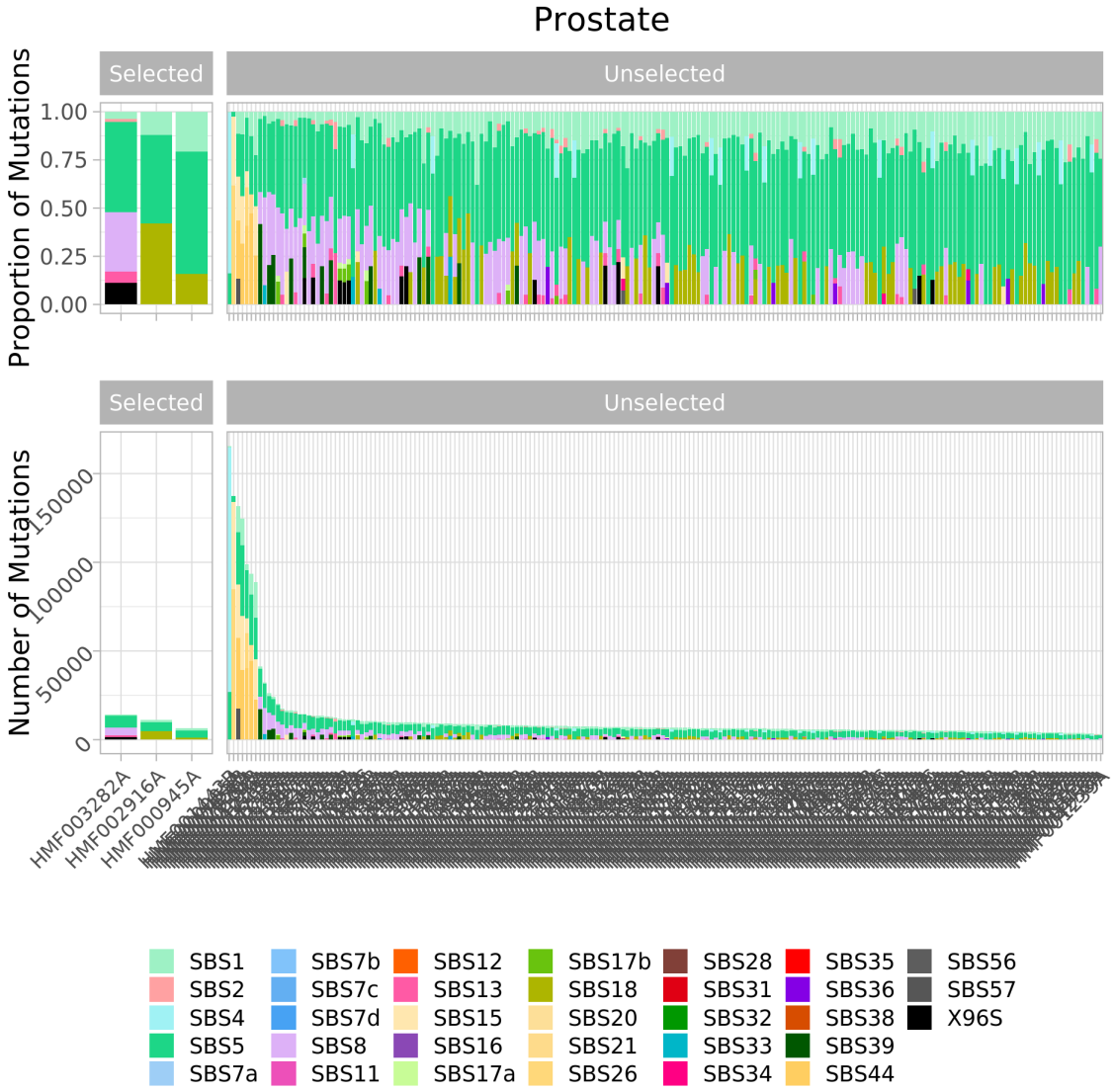


Figure C.33: Distribution of mutational signatures in selected and unselected prostate cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

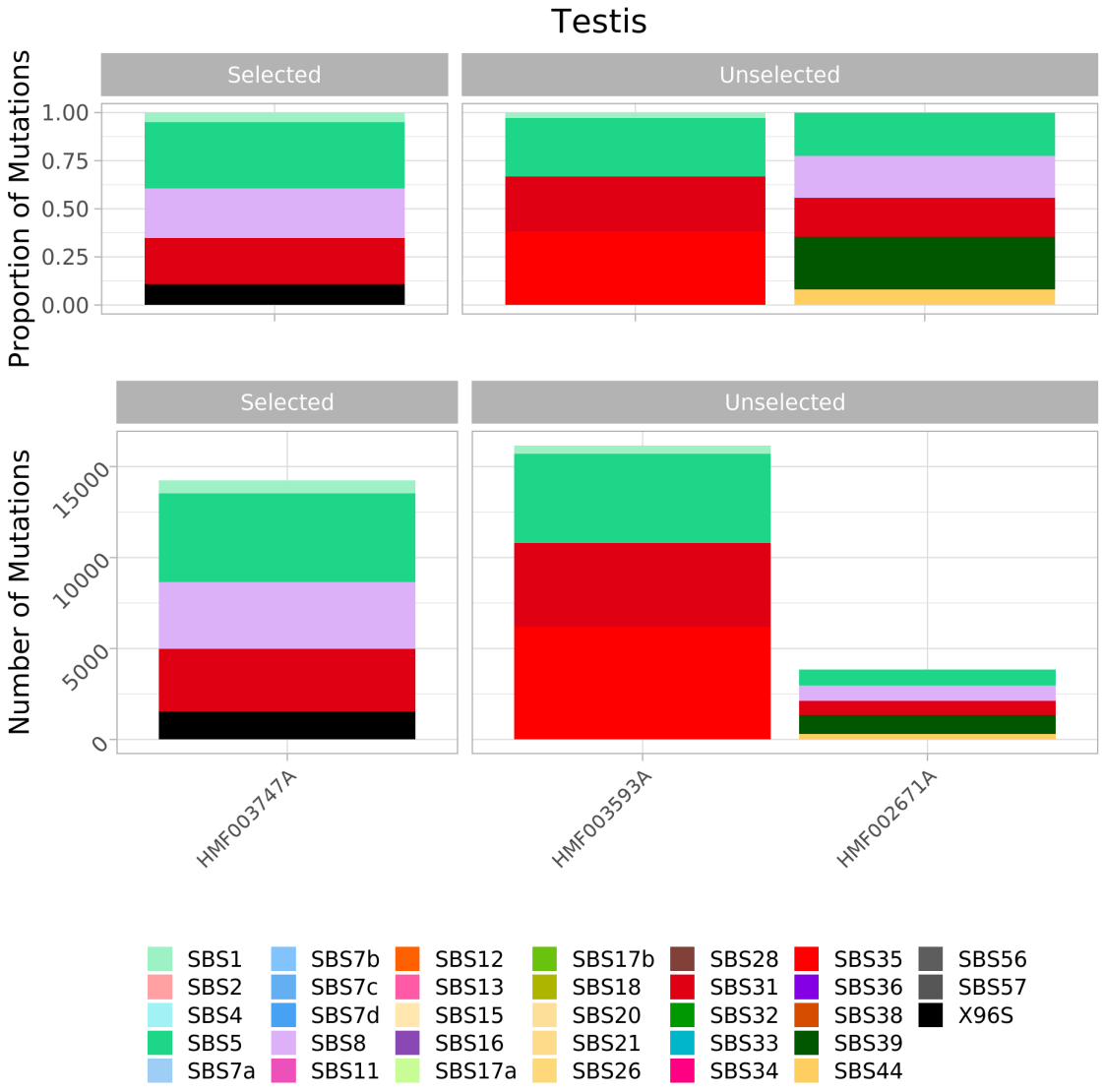


Figure C.34: Distribution of mutational signatures in selected and unselected testicular cancers in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

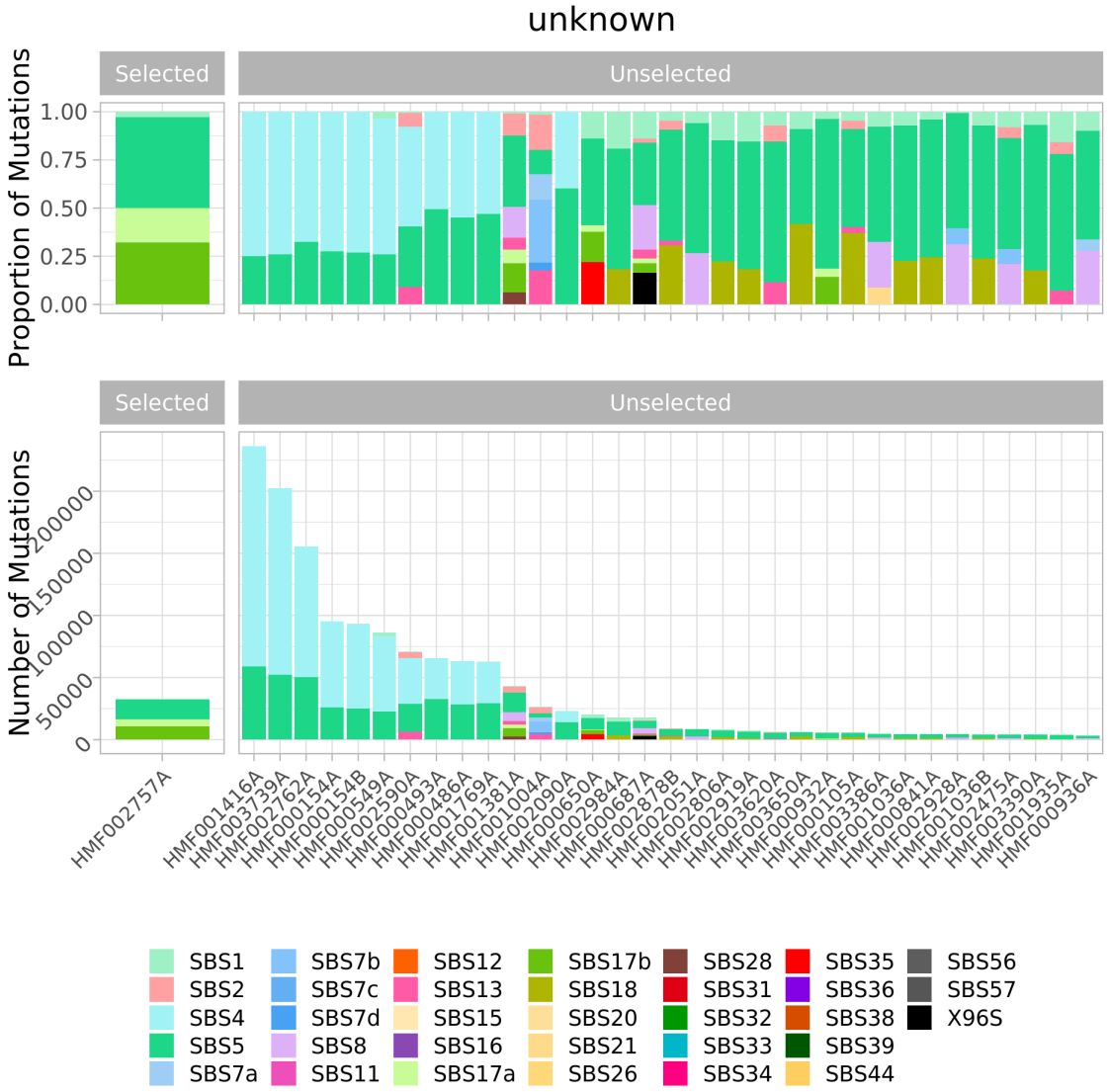


Figure C.35: Distribution of mutational signatures in selected and unselected cancers of unknown primary in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

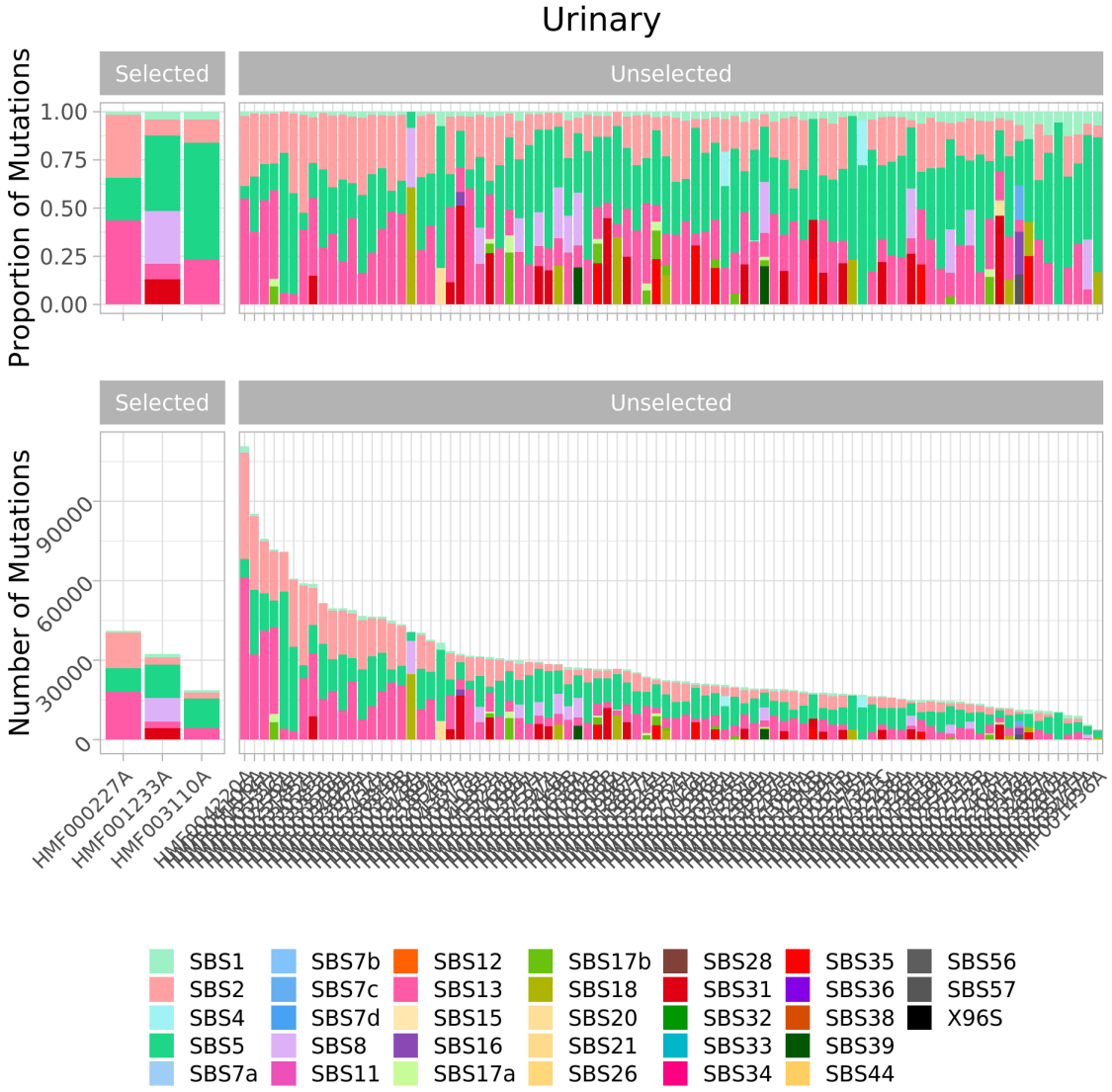


Figure C.36: Distribution of mutational signatures in selected and unselected cancers of the urinary tract in the HMF data set. Only tetraploid samples for which fusion measures could be calculated were considered.

C.2.4 Selected Hartwig Samples

In the following figures, opacity and bar width has been adjusted to aid visualisation of overlapping or very close segments.

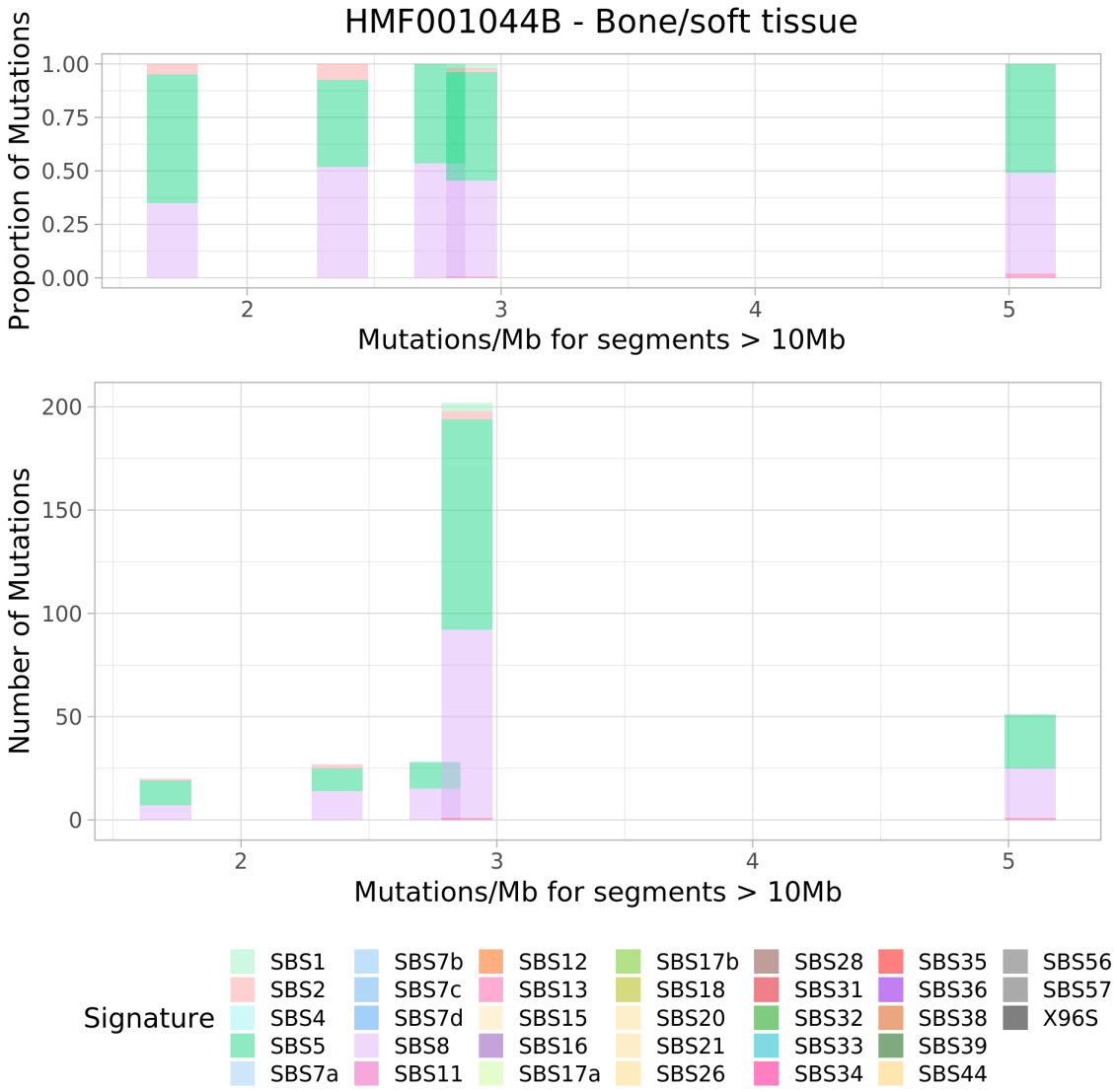


Figure C.37: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001044B.

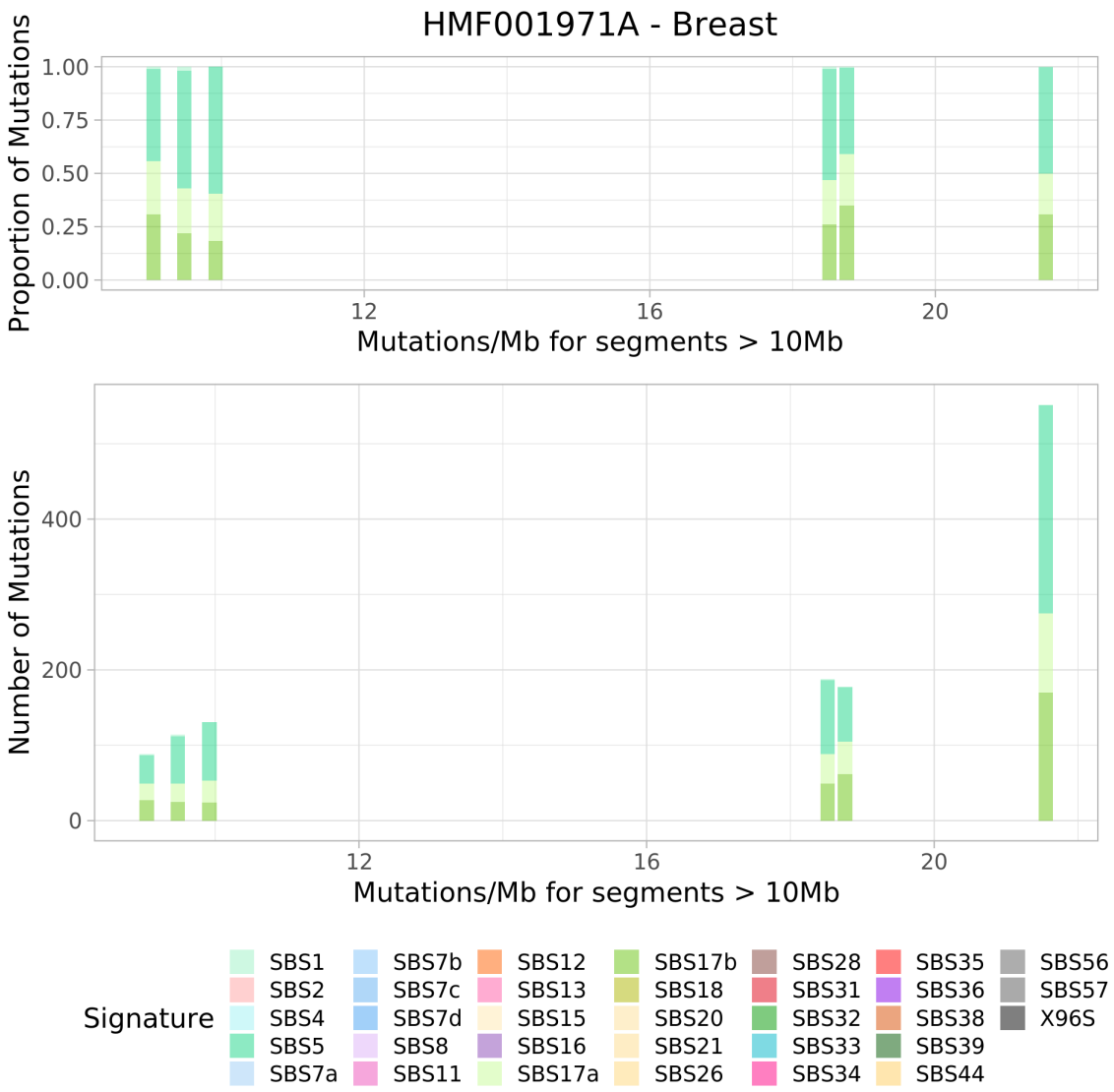


Figure C.38: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001971A.

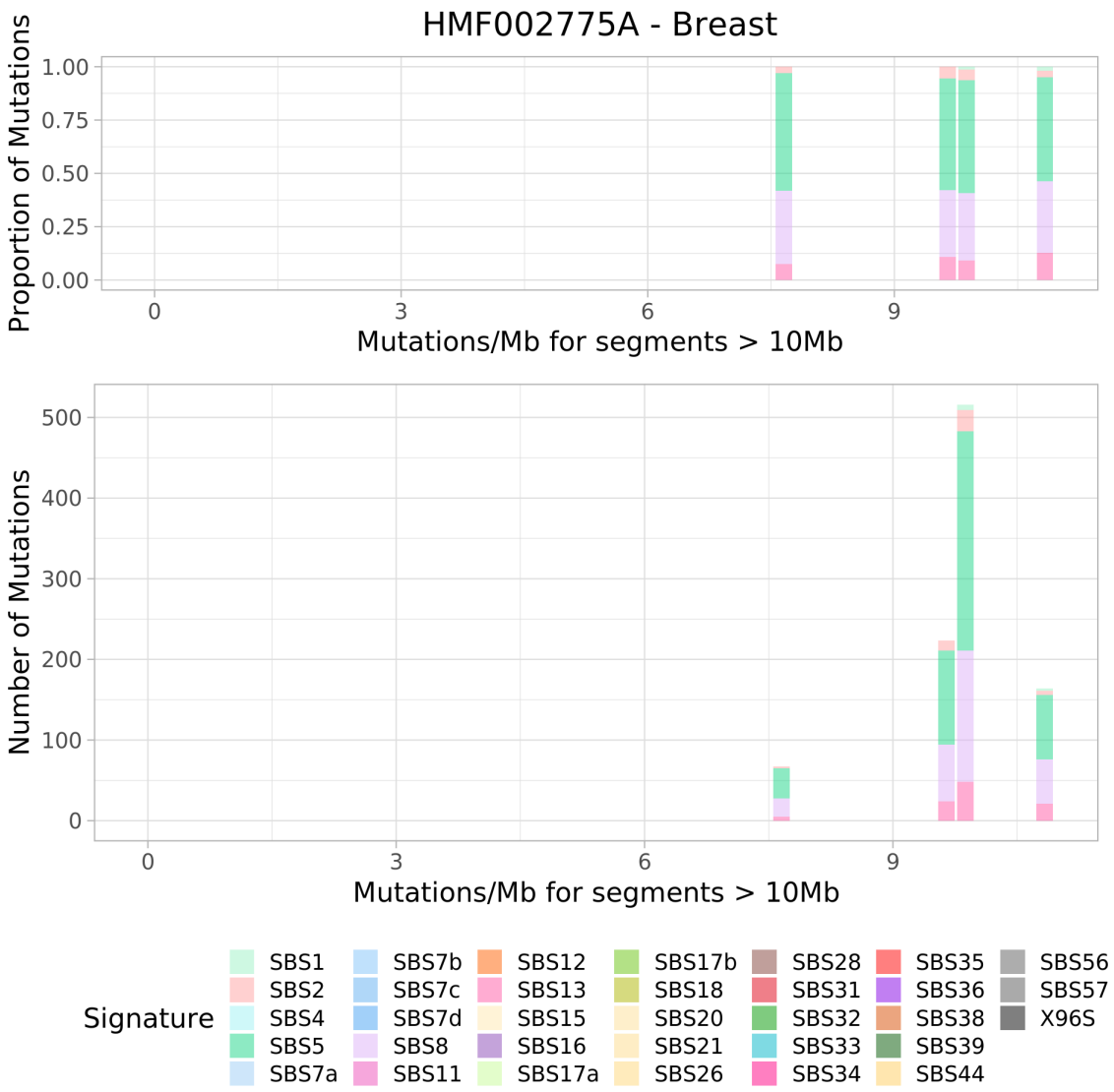


Figure C.39: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002775A.

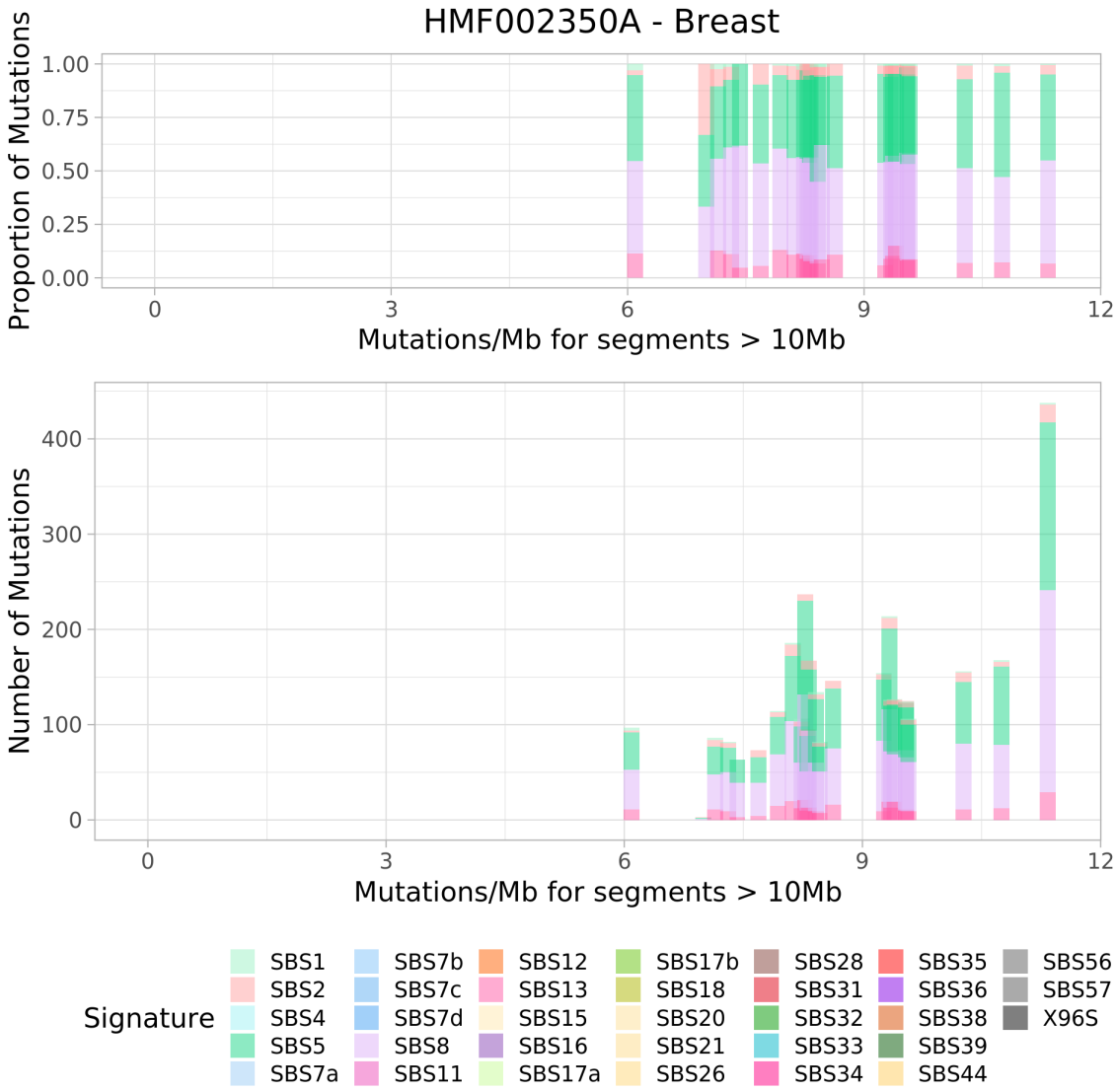


Figure C.40: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002350A.

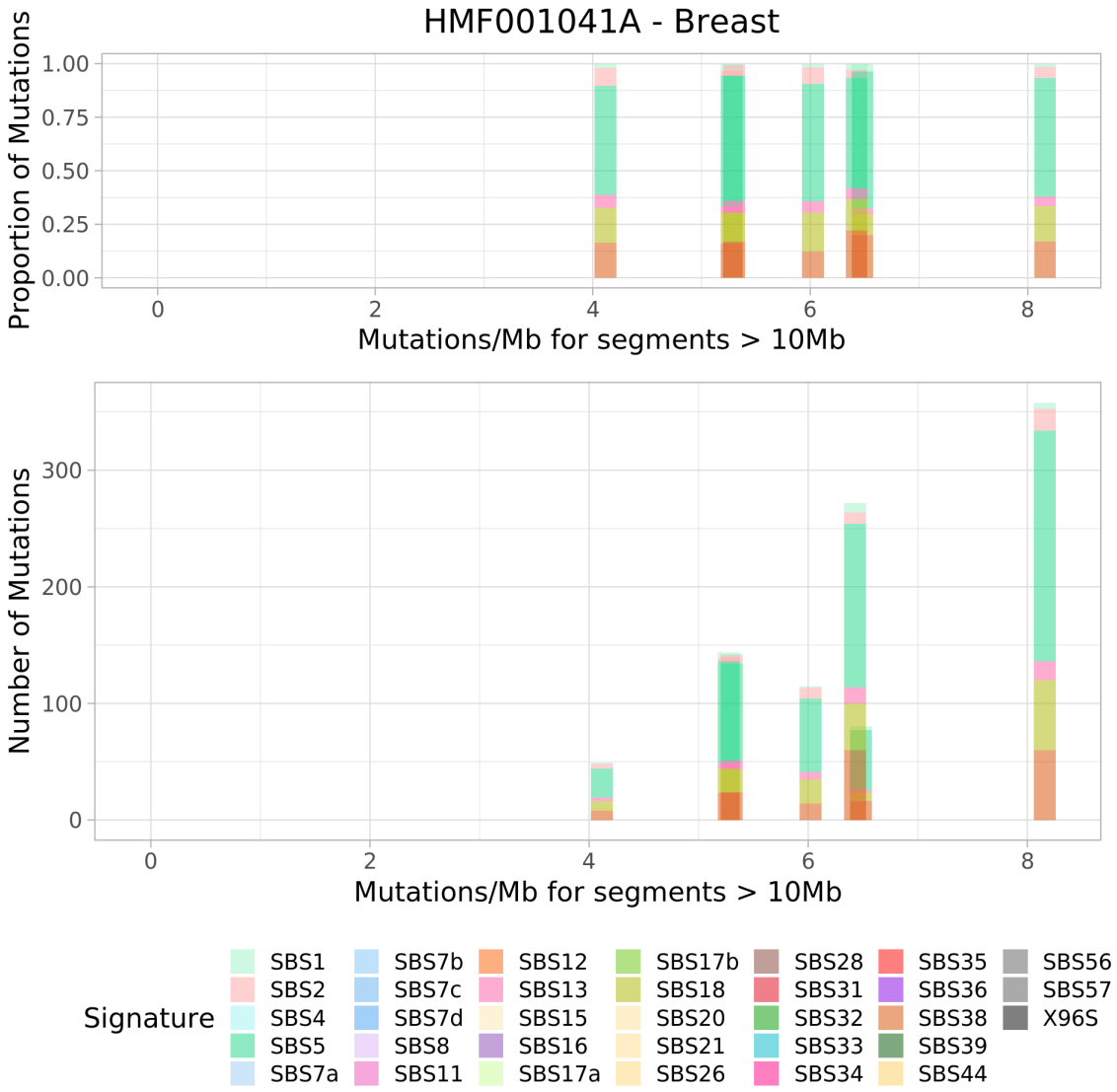


Figure C.41: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001041A.



Figure C.42: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002824A.

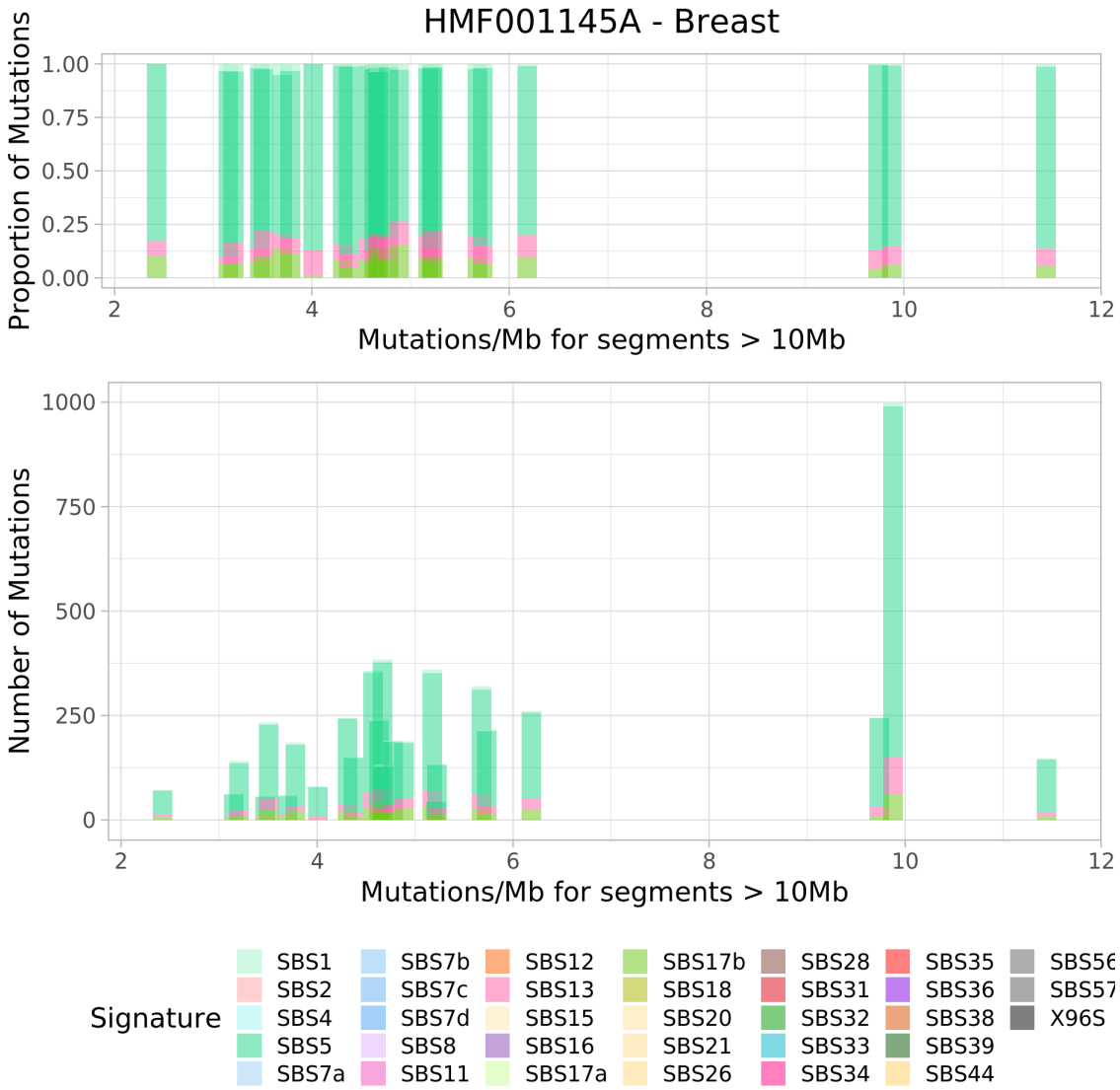


Figure C.43: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001145A.



Figure C.44: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001145B.



Figure C.45: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF000851B.

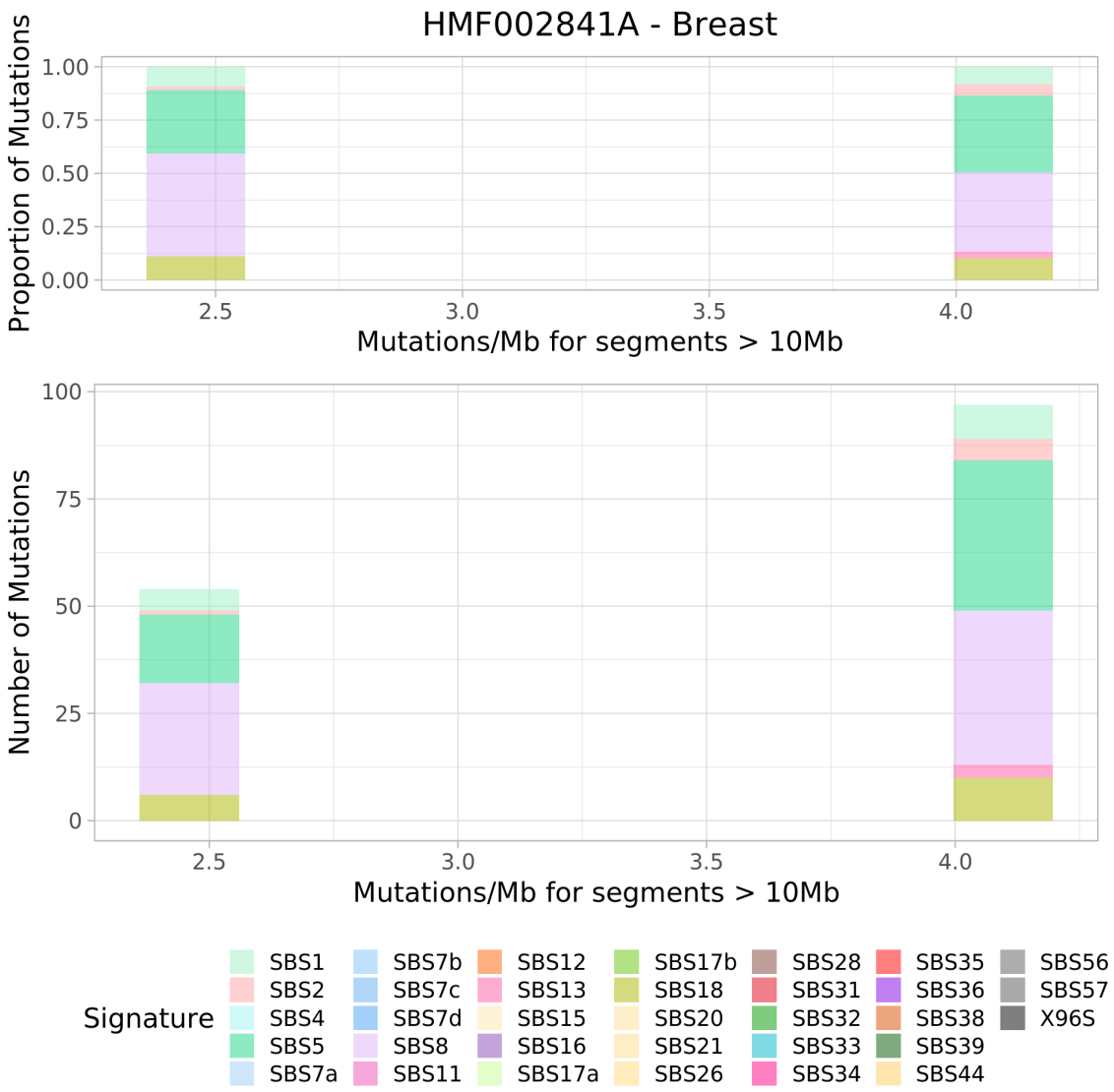


Figure C.46: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002841A.



Figure C.47: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF000567A.

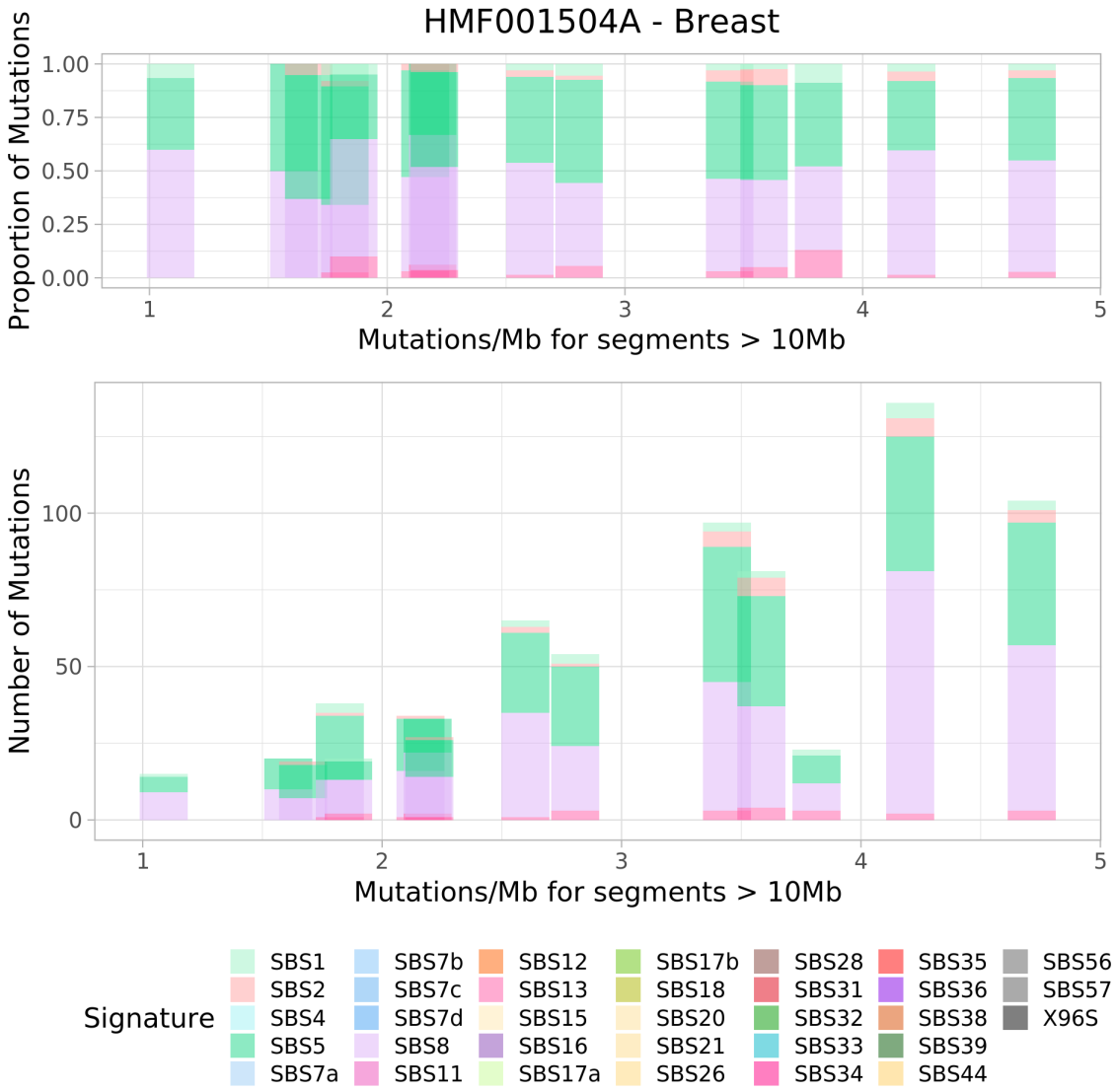


Figure C.48: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001504A.

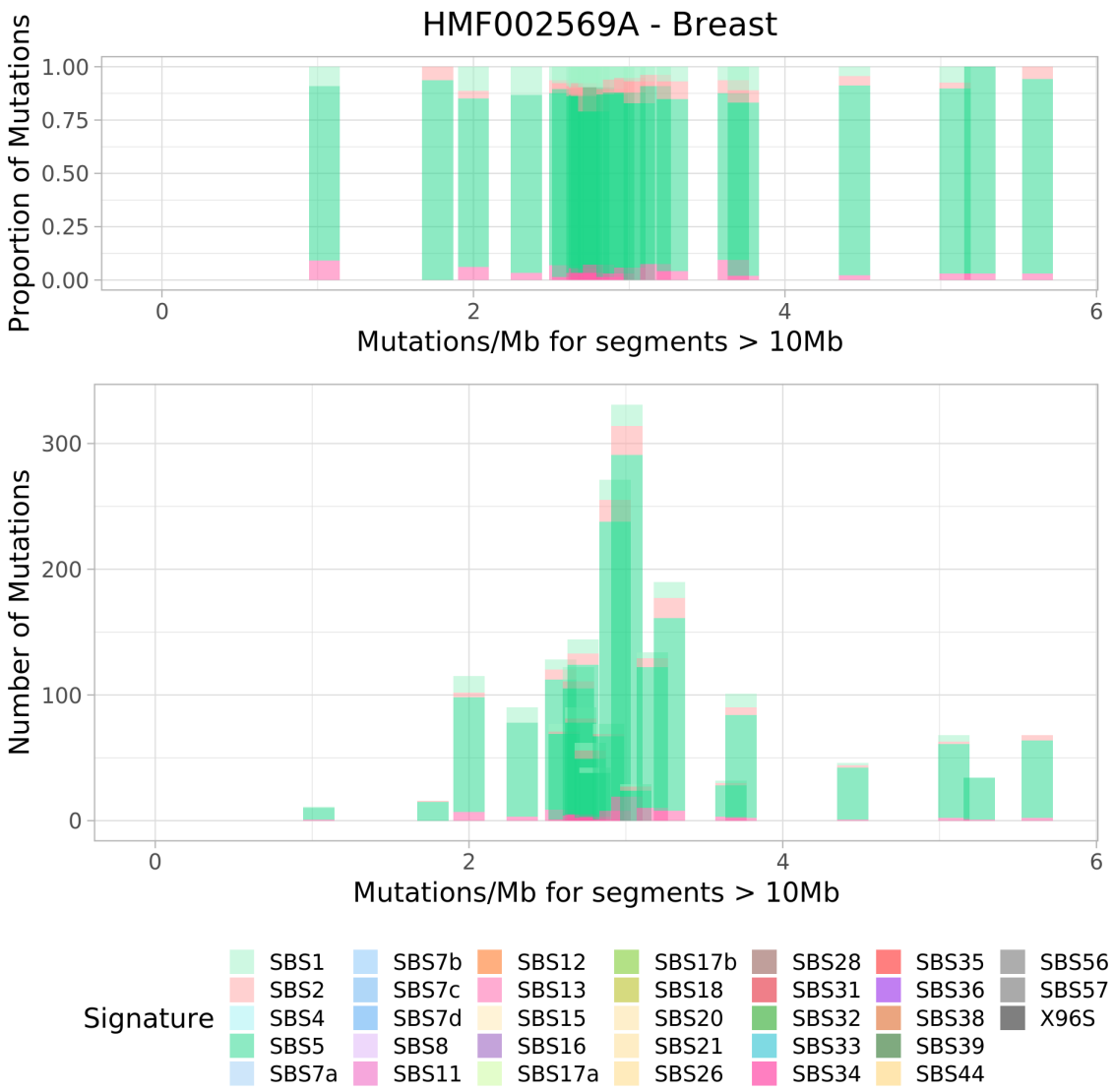


Figure C.49: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002569A.

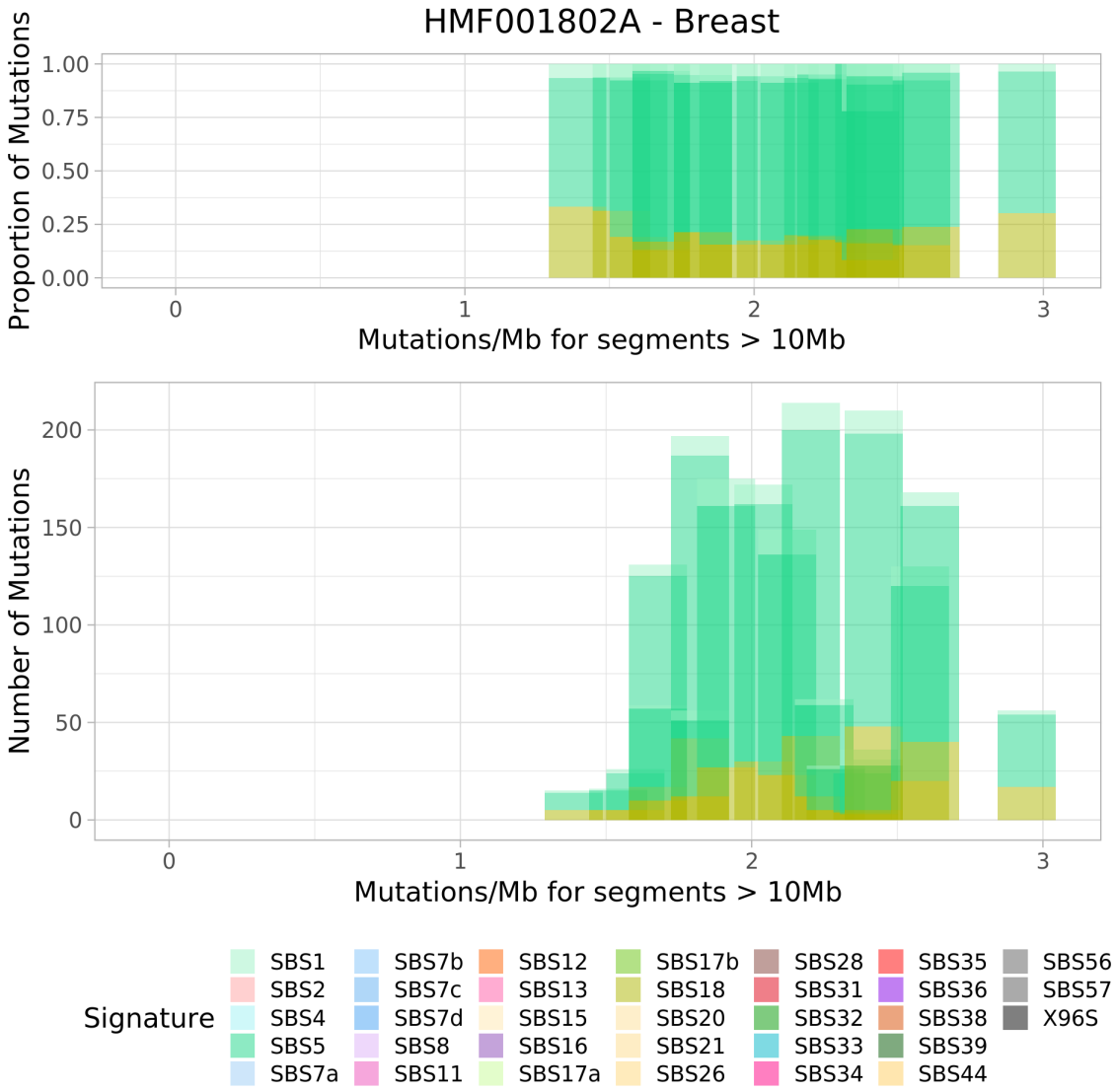


Figure C.50: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001802A.

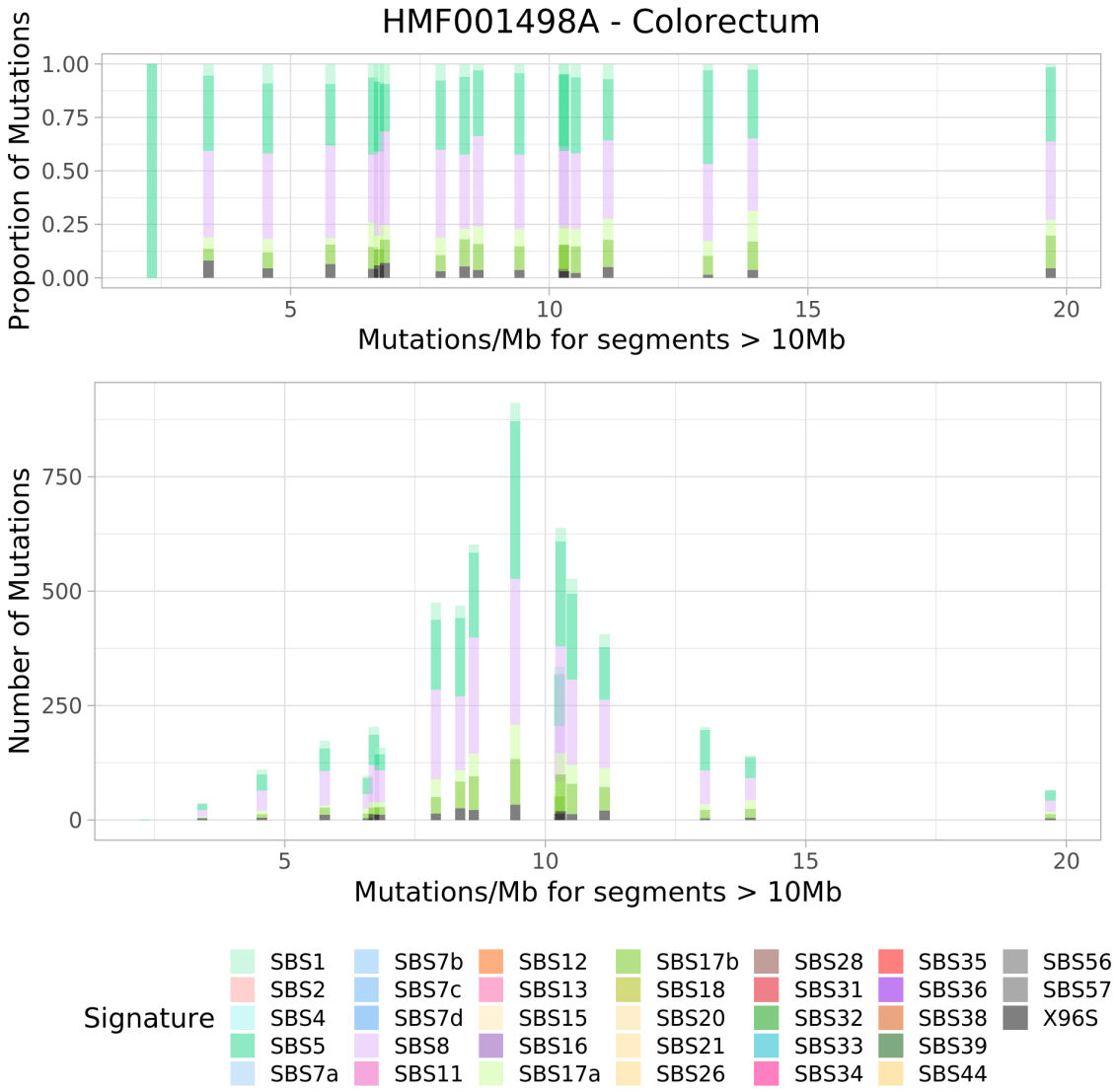


Figure C.51: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001498A.

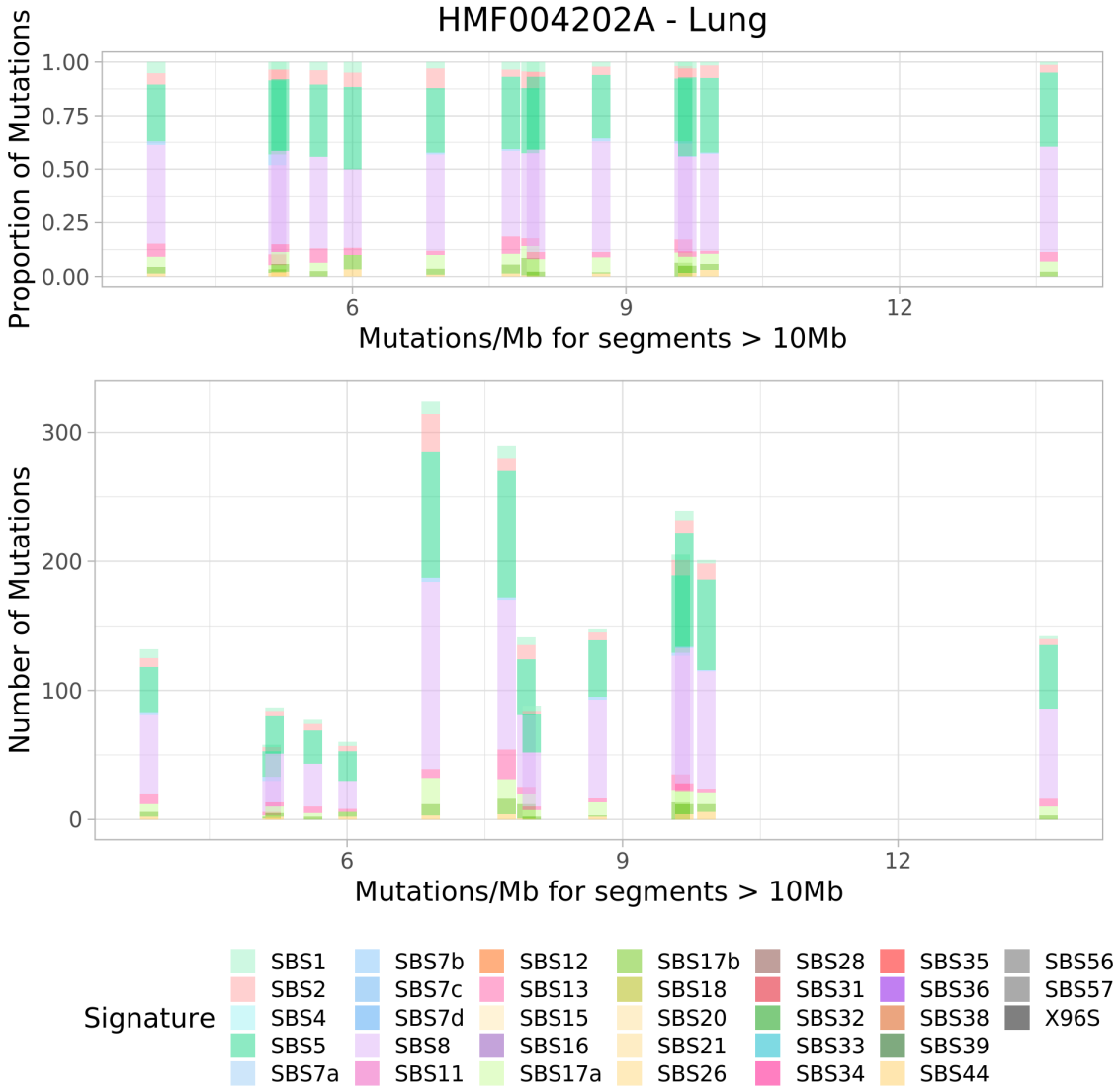


Figure C.52: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF004202A.

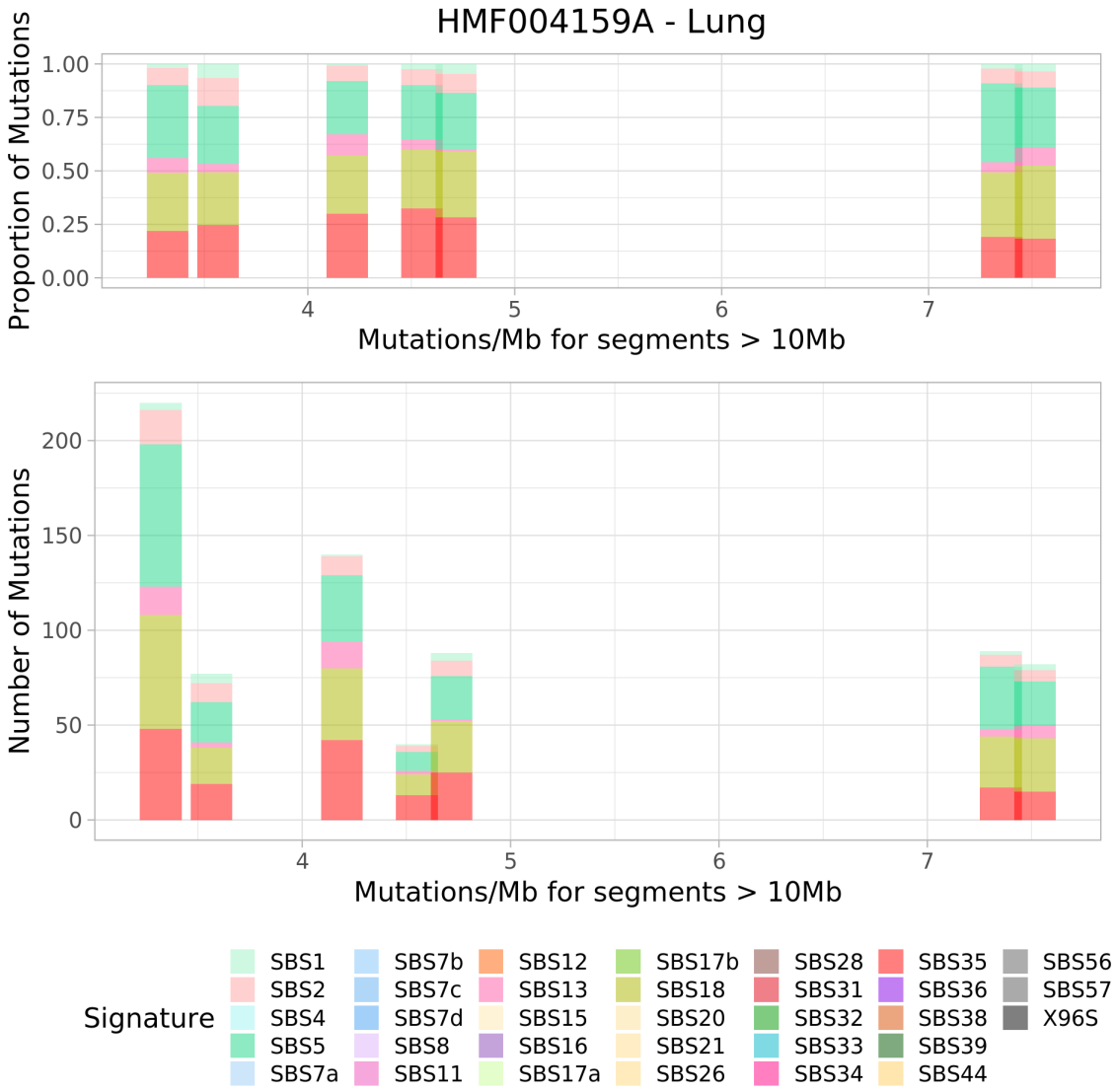


Figure C.53: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF004159A.

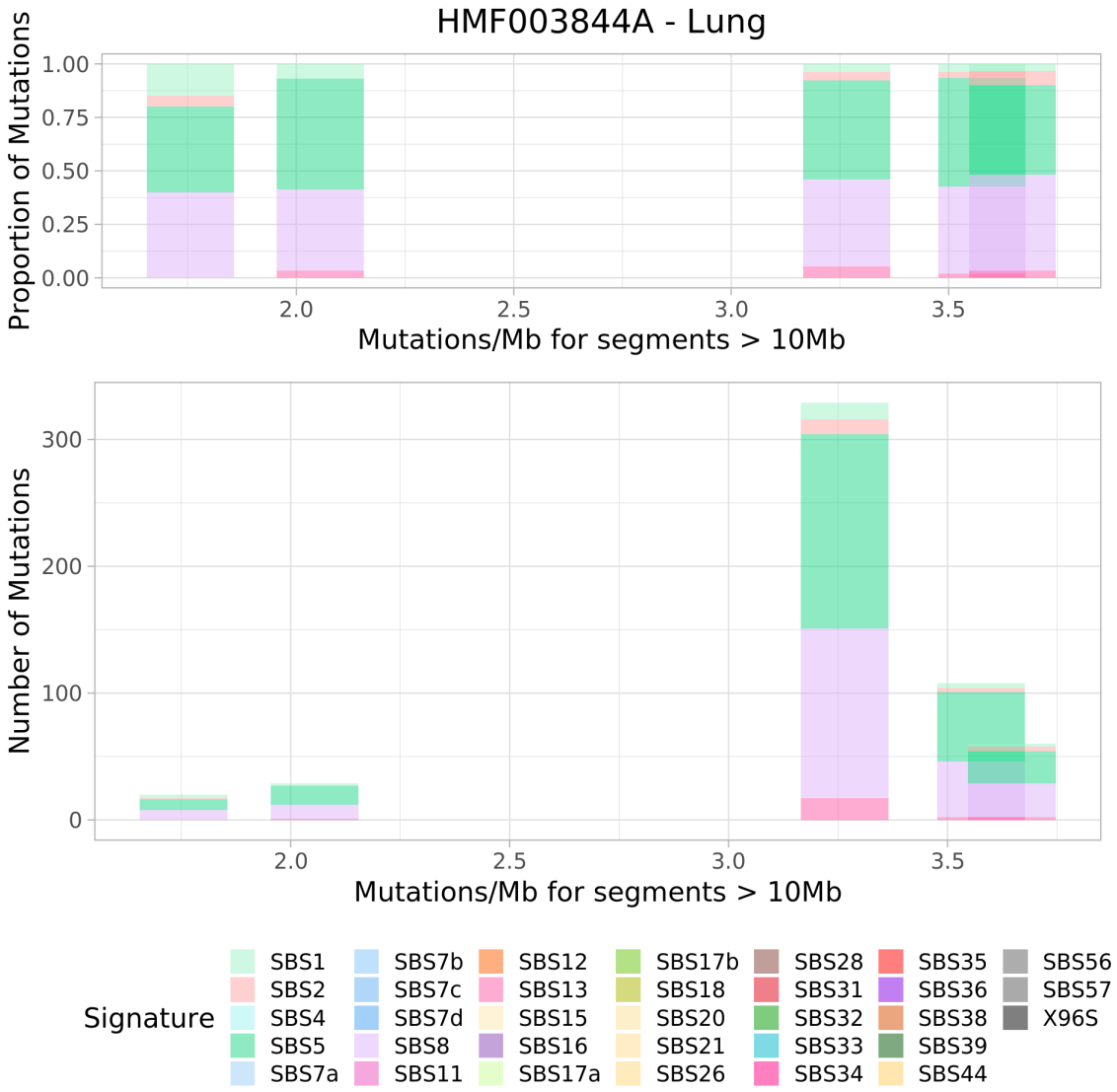


Figure C.54: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003844A.

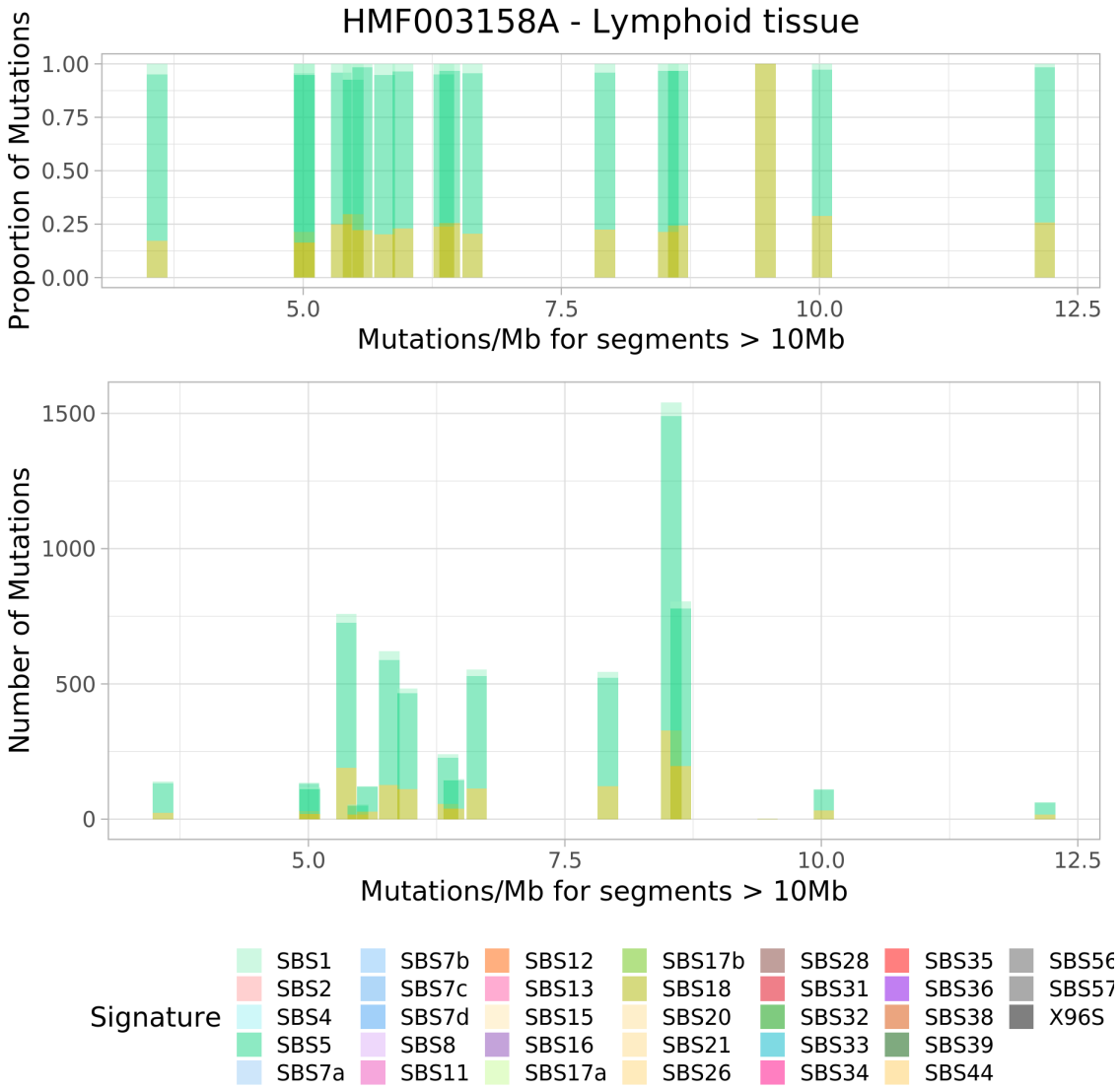


Figure C.55: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003158A.

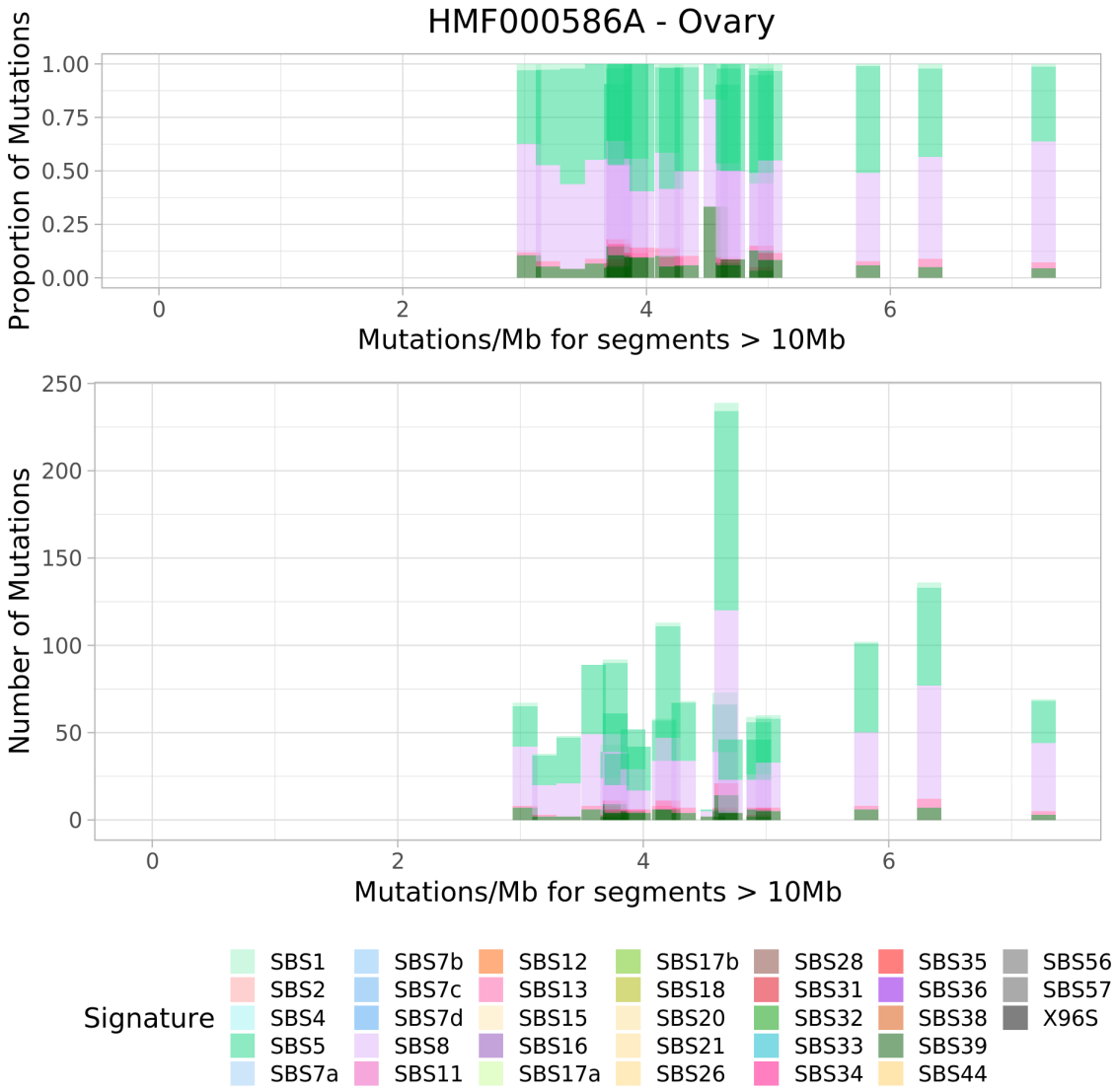


Figure C.56: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF000586A.

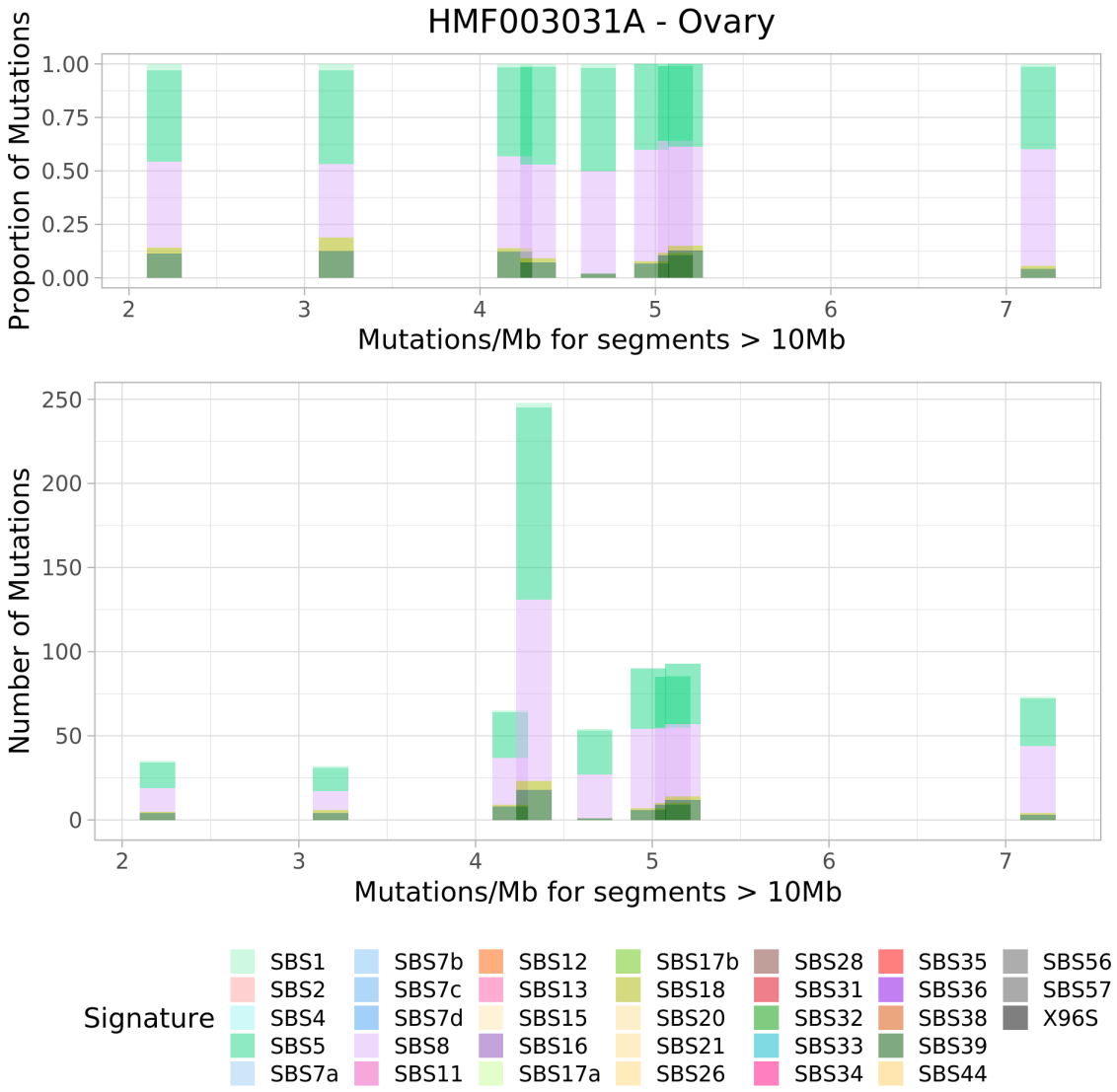


Figure C.57: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003031A.

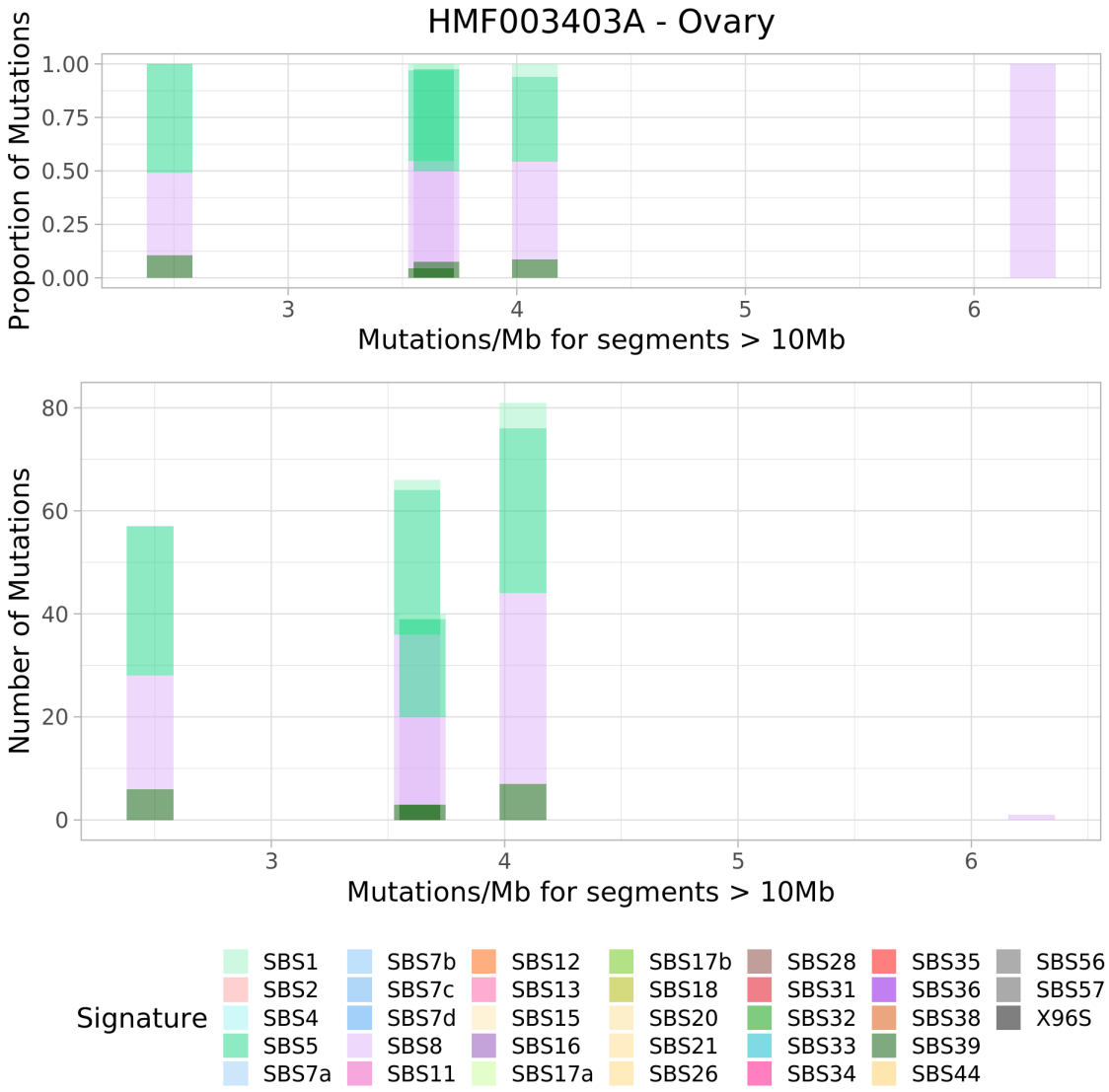


Figure C.58: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003403A.

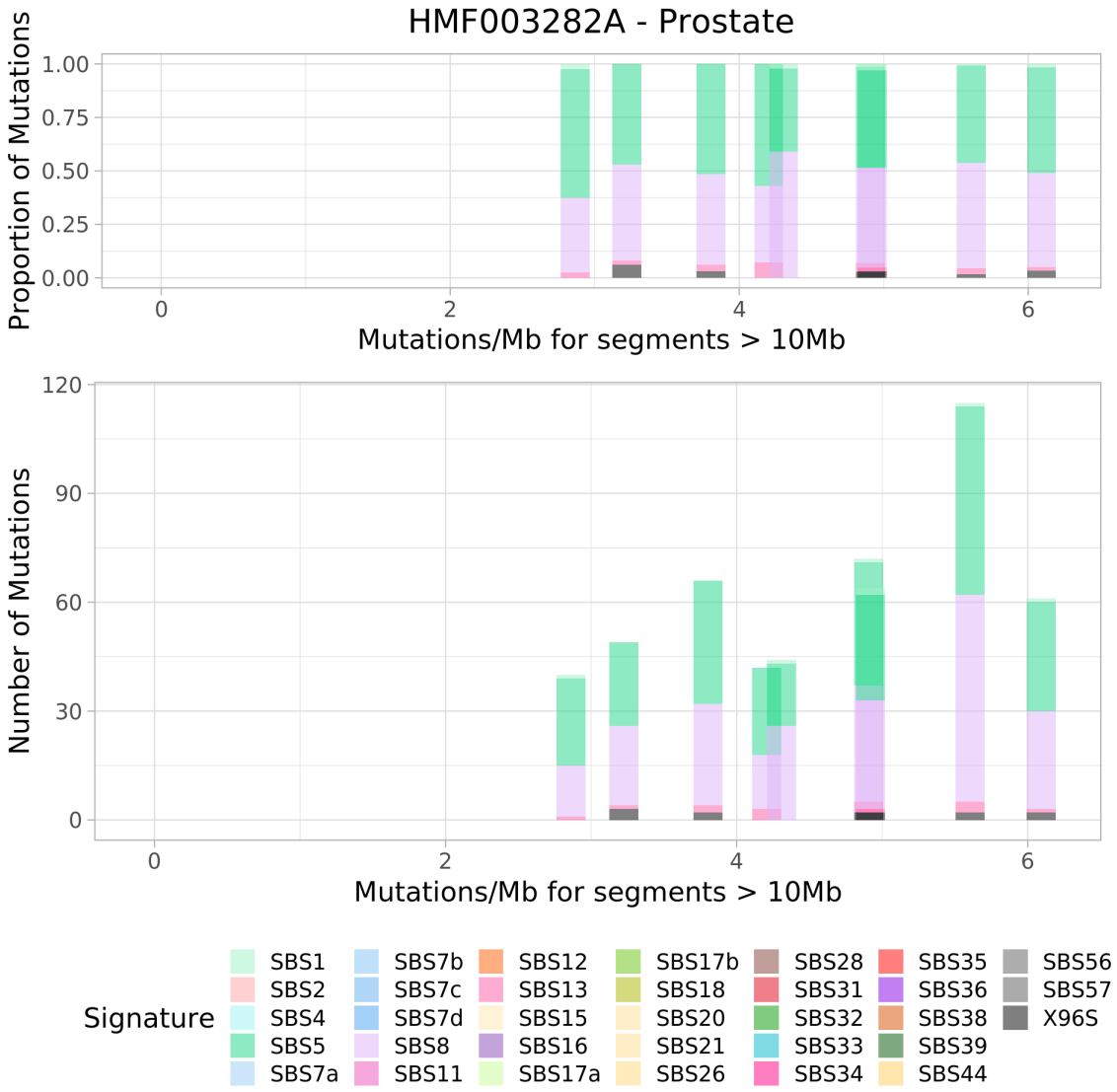


Figure C.59: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003282A.

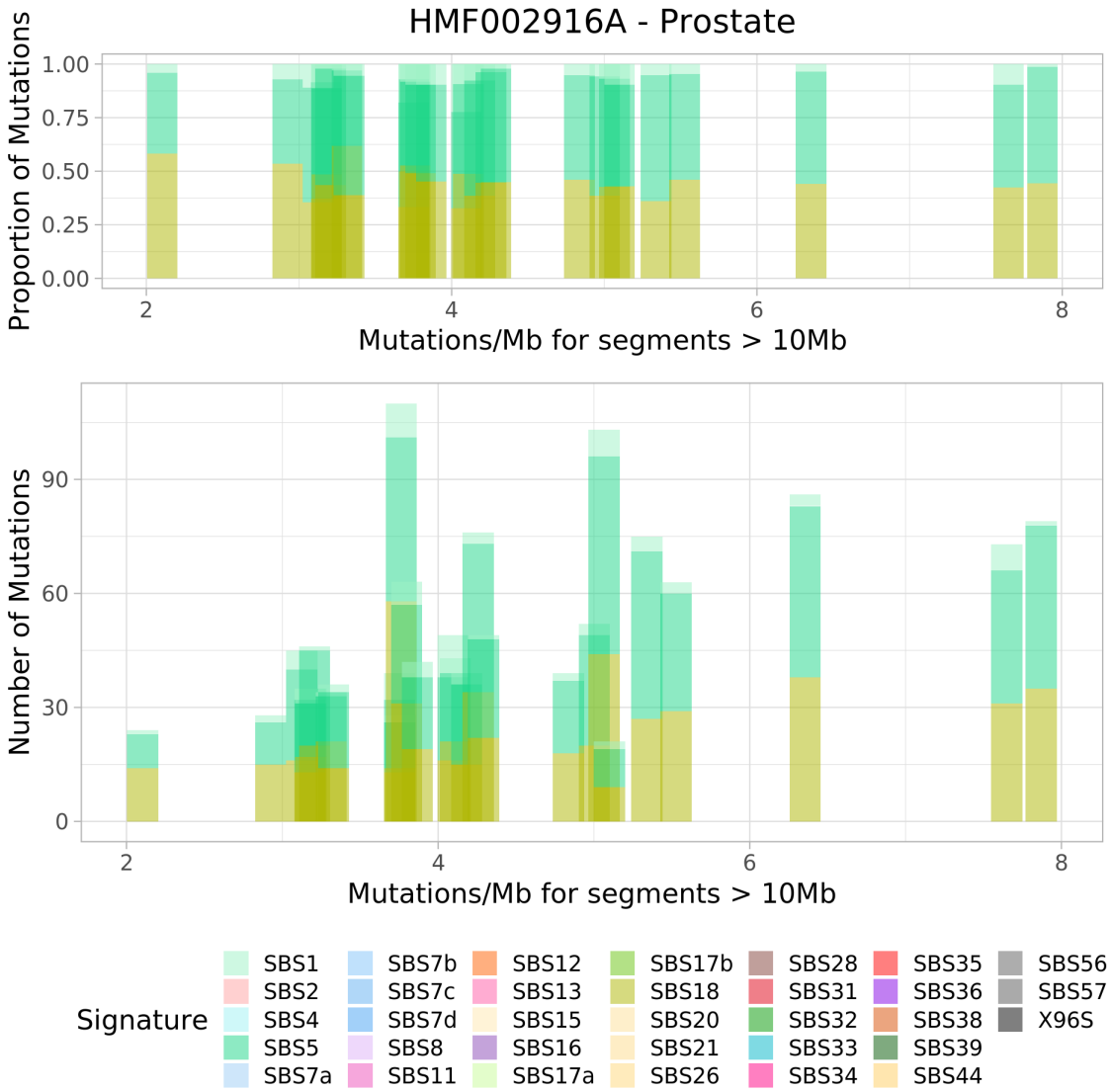


Figure C.60: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002916A.

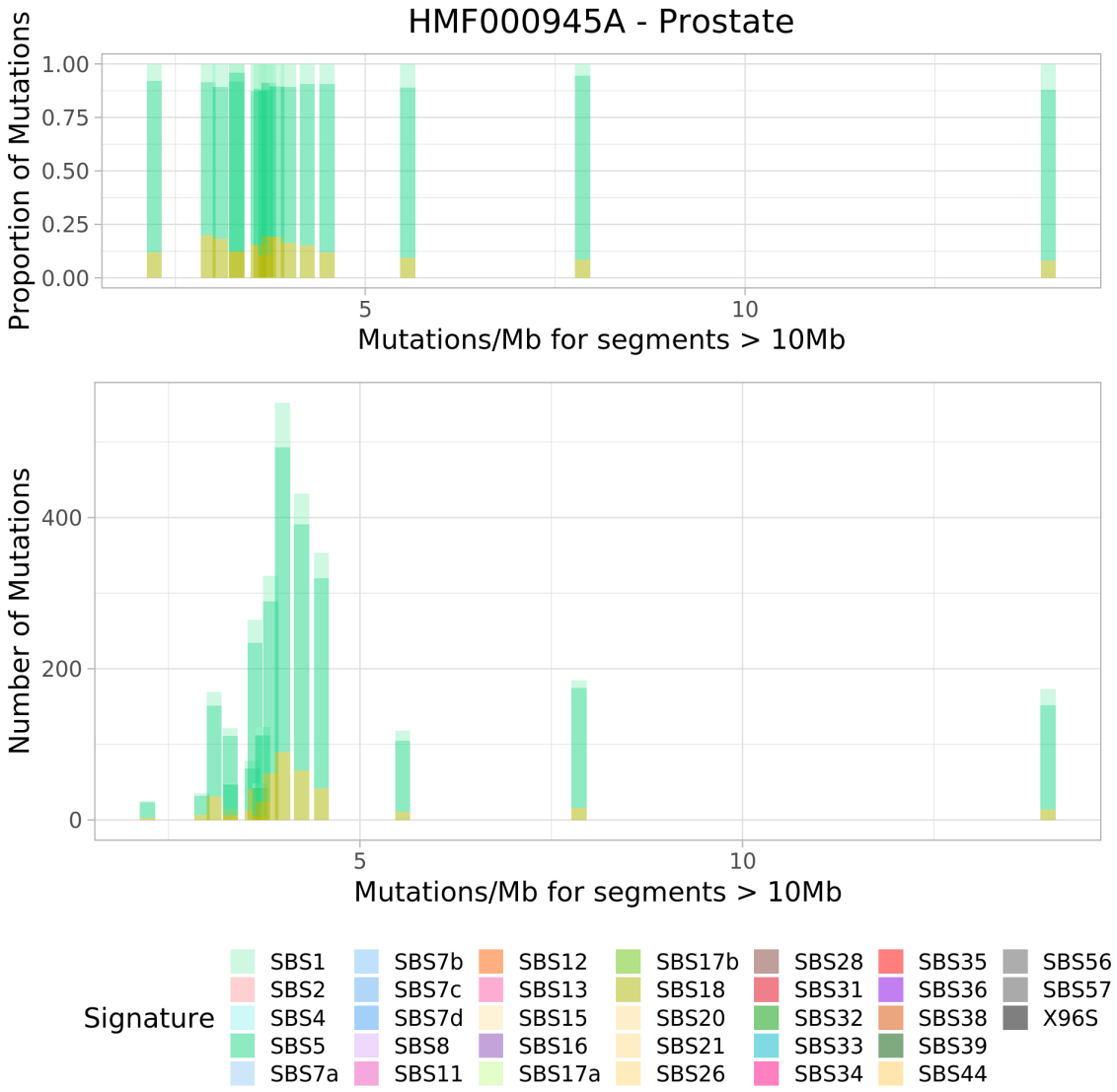


Figure C.61: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF000945A.

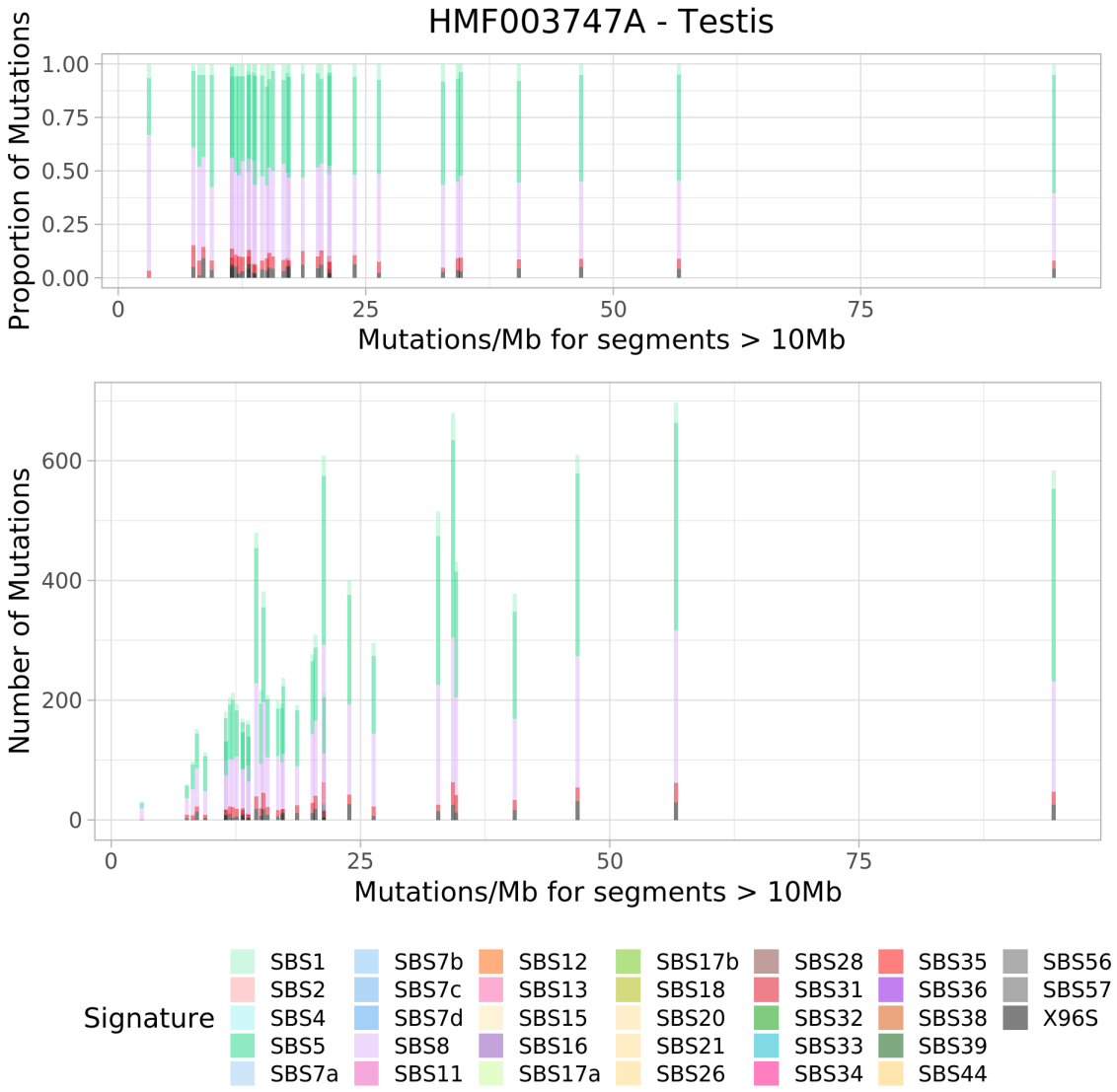


Figure C.62: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003747A.

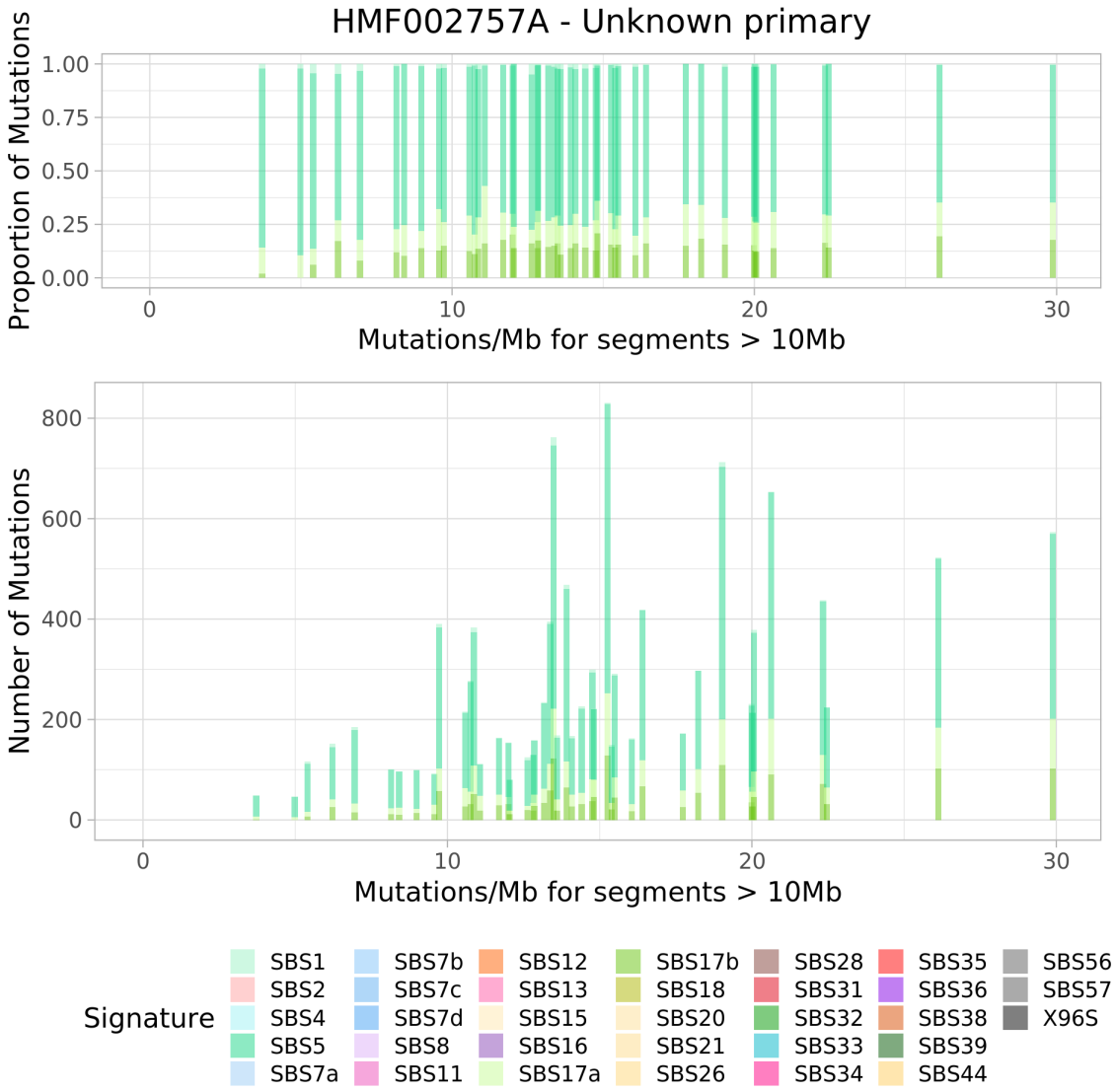


Figure C.63: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF002757A.



Figure C.64: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF000227A.

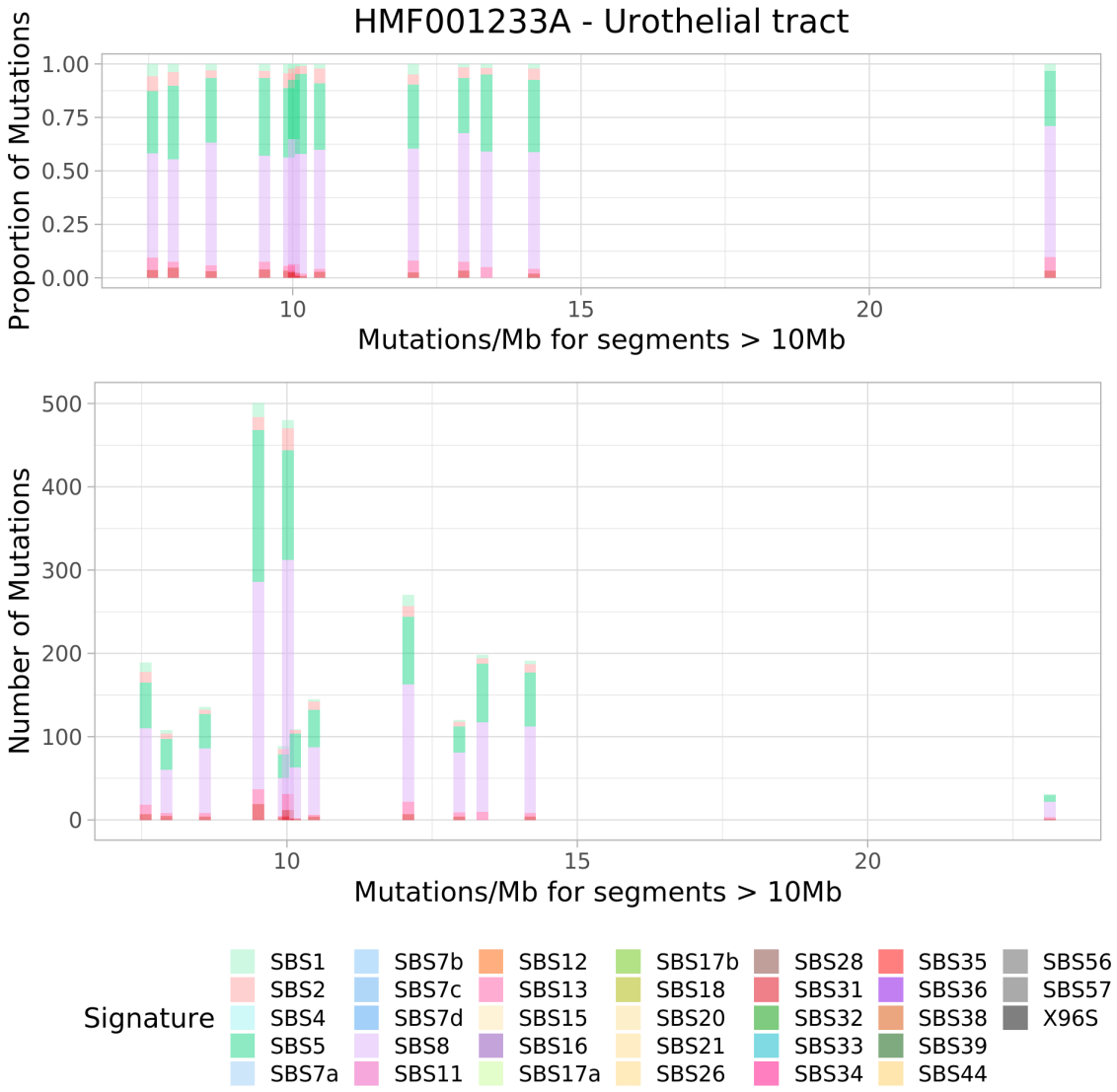


Figure C.65: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF001233A.

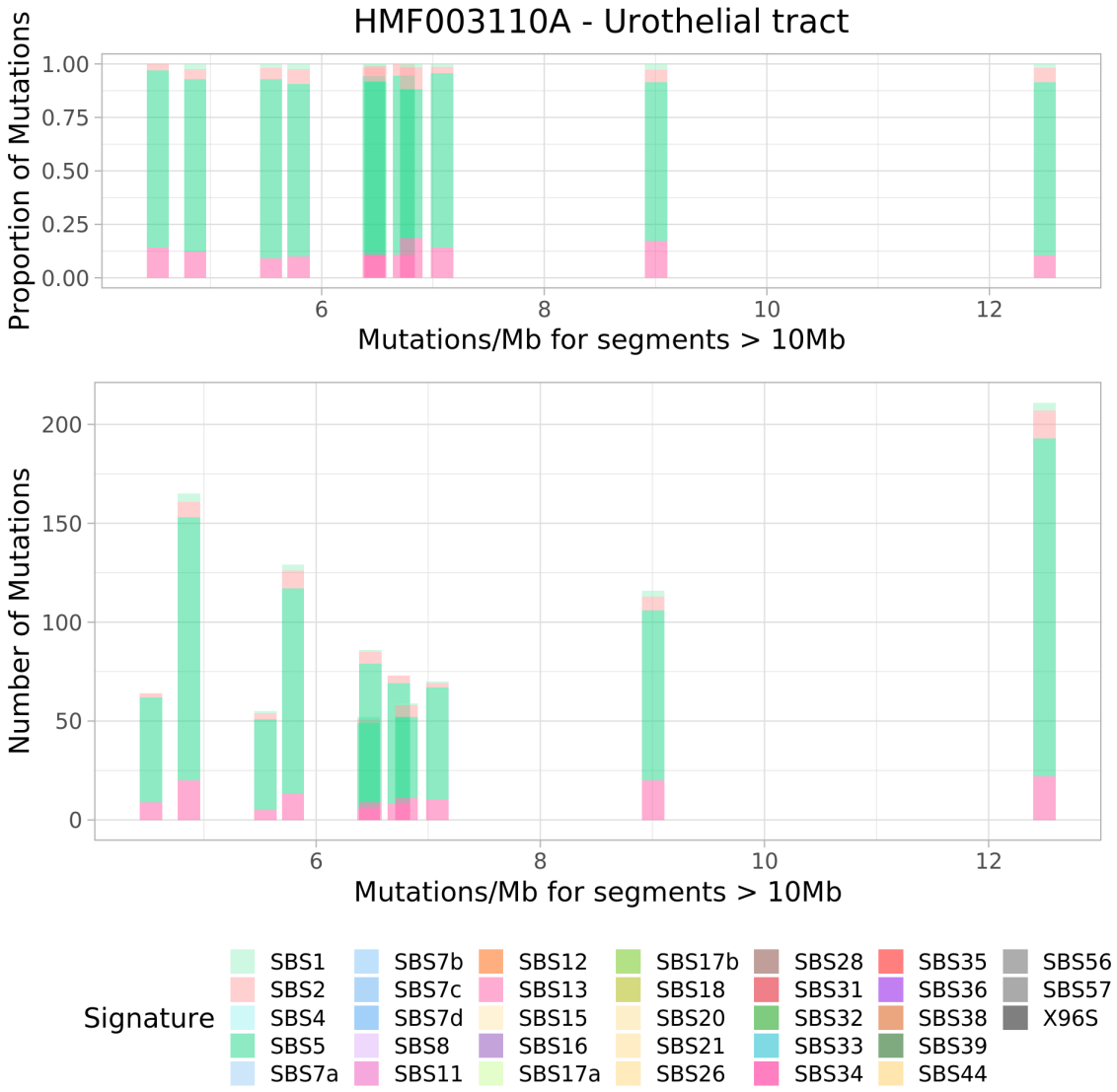


Figure C.66: Distribution of mutational signatures across clonal segments larger than 10Mb with a copy number state of 1+1 or 2+1 in Hartwig sample HMF003110A.

C.3 Representative Karyotypes

C.3.1 Karyotypes at Harvest 0

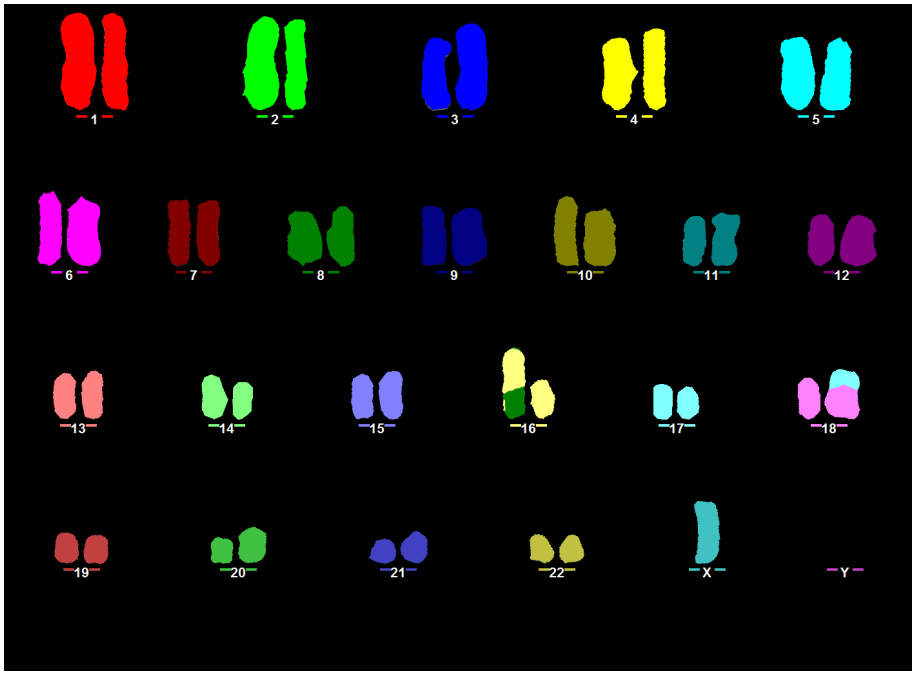


Figure C.67: M-B10 karyotype at harvest 0.

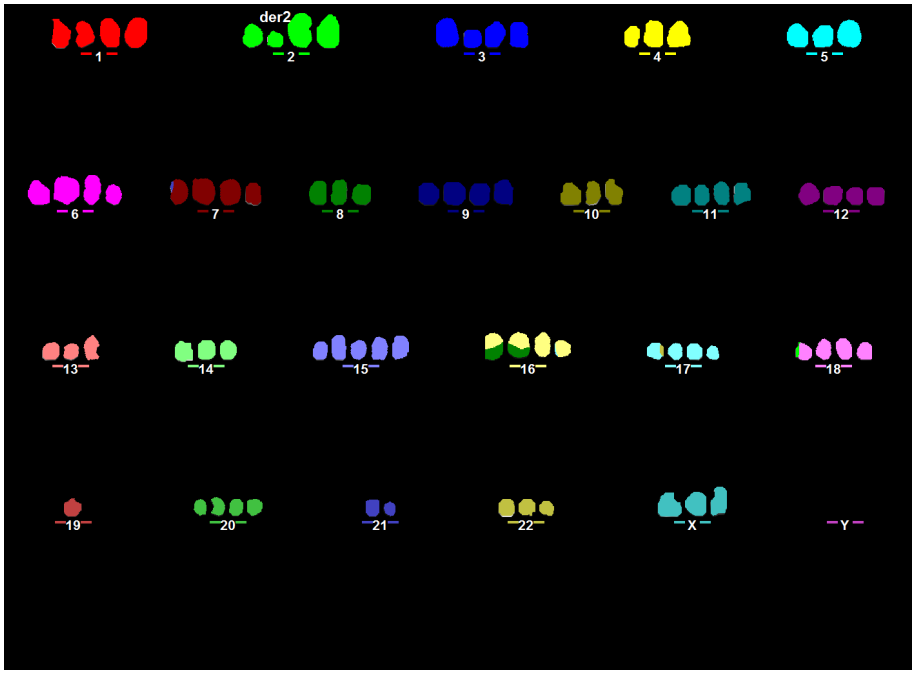


Figure C.68: C-F10 karyotype at harvest 0.

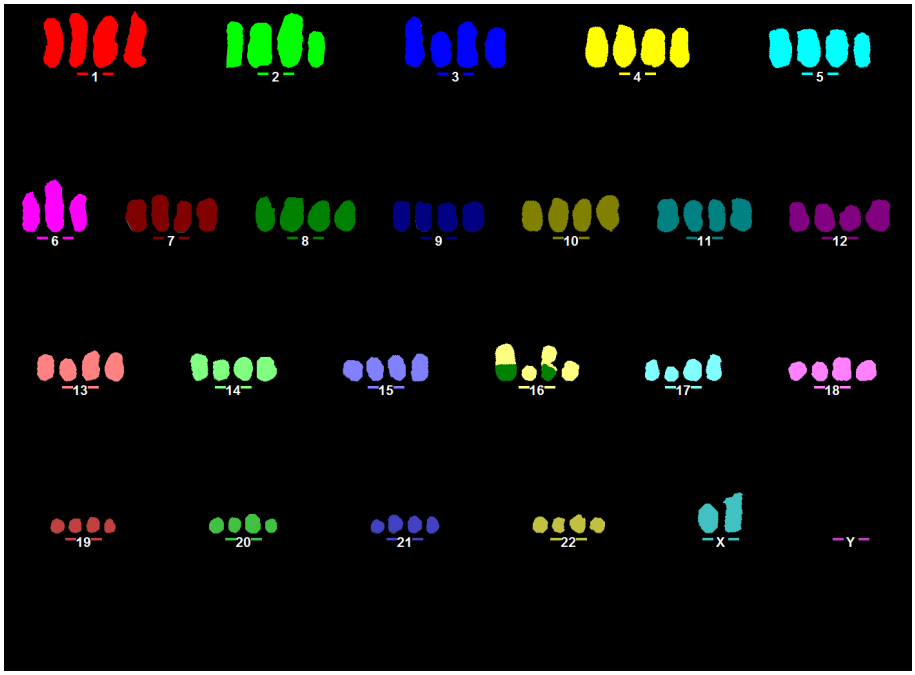


Figure C.69: C-F3 karyotype at harvest 0.

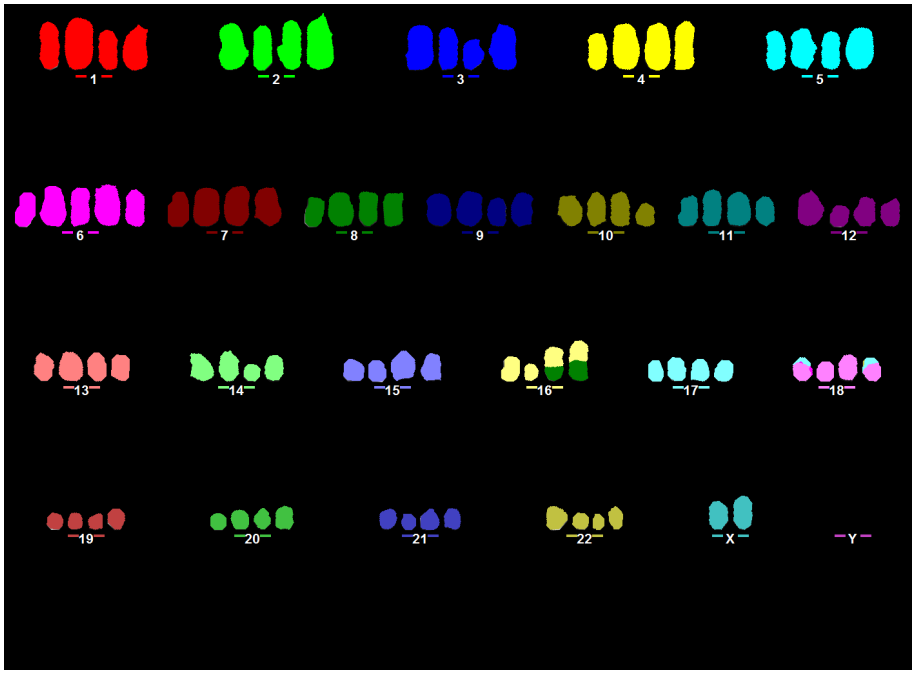


Figure C.70: C-G9 karyotype at harvest 0.

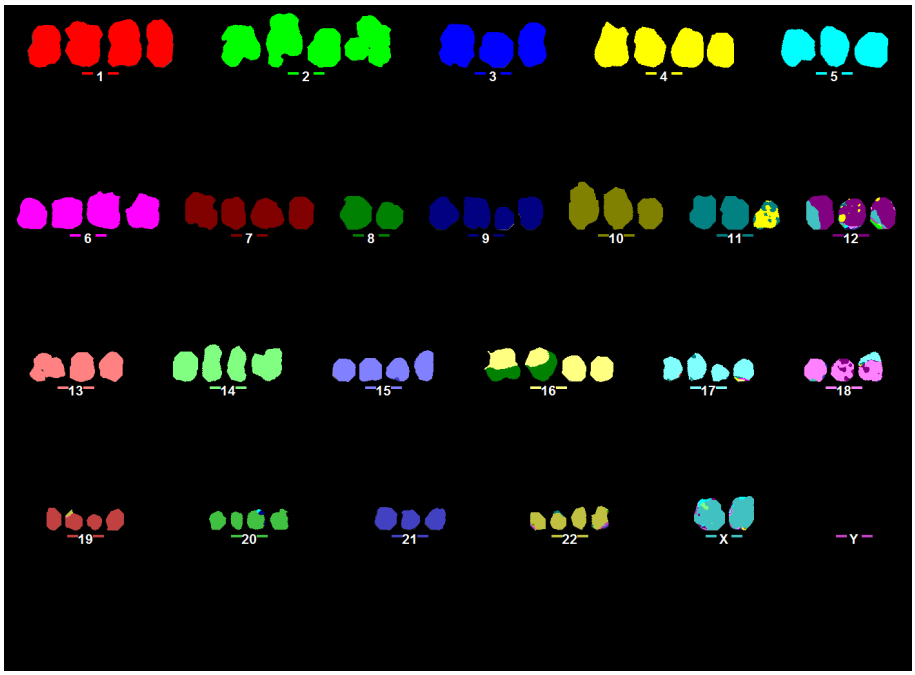


Figure C.71: V-F4 karyotype at harvest 0.

C.3.2 Karyotypes at Harvest 1

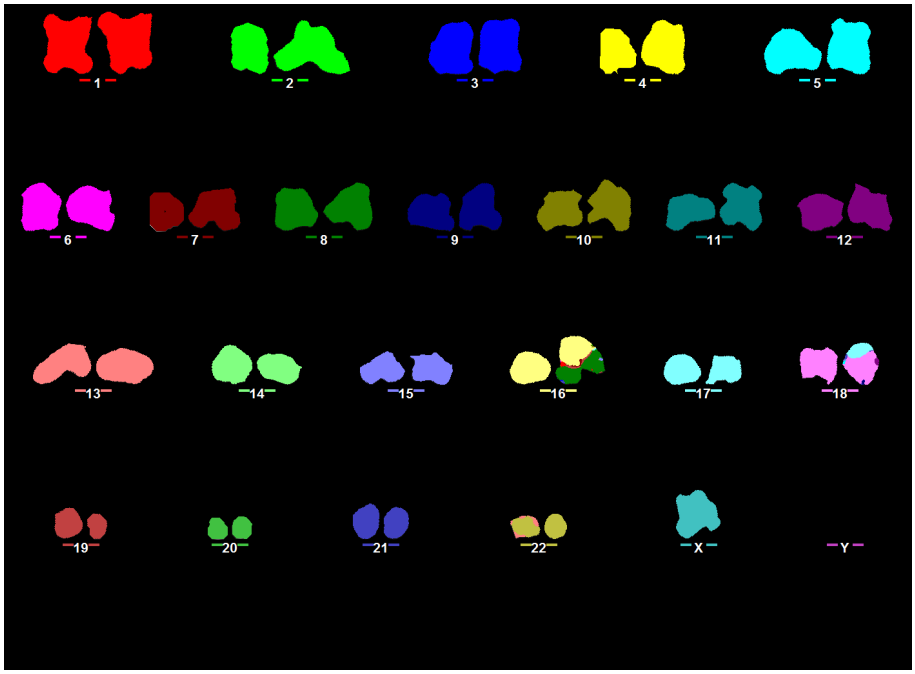


Figure C.72: M-B10 karyotype at harvest 1.

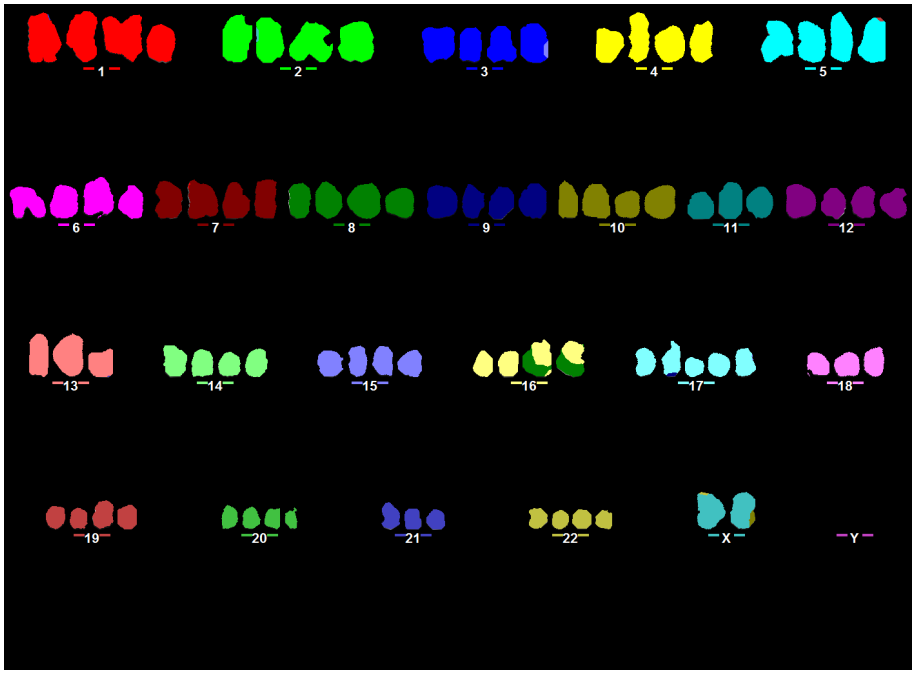


Figure C.73: C-F10 karyotype at harvest 1.

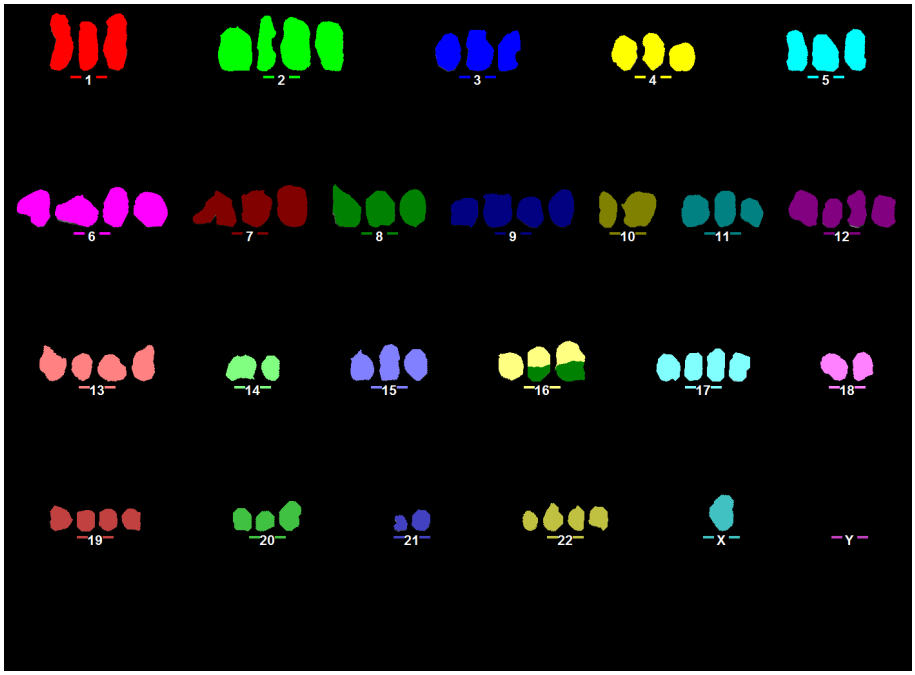


Figure C.74: C-F3 karyotype at harvest 1.

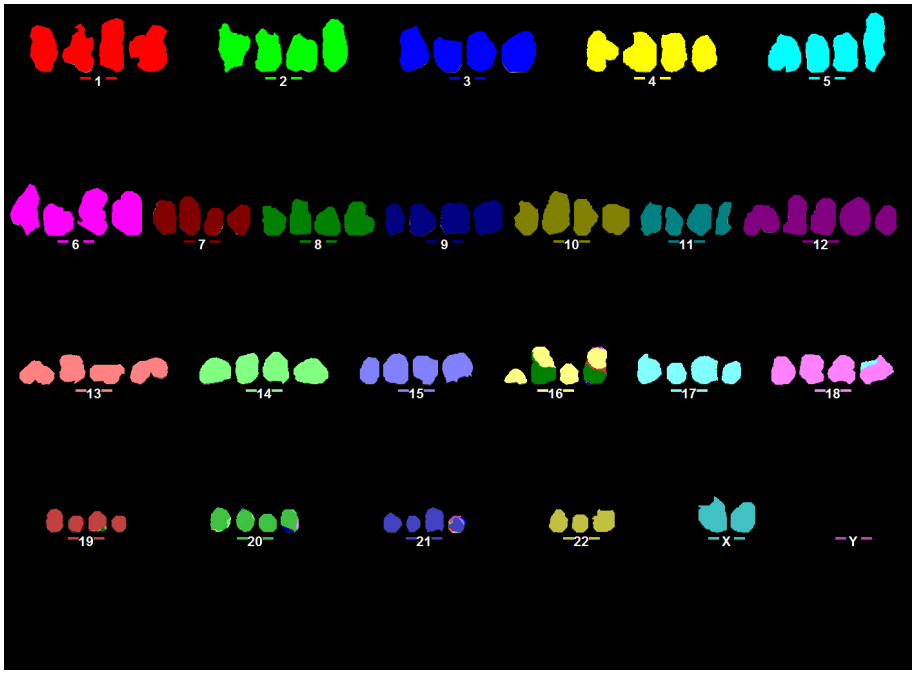


Figure C.75: C-G9 karyotype at harvest 1.

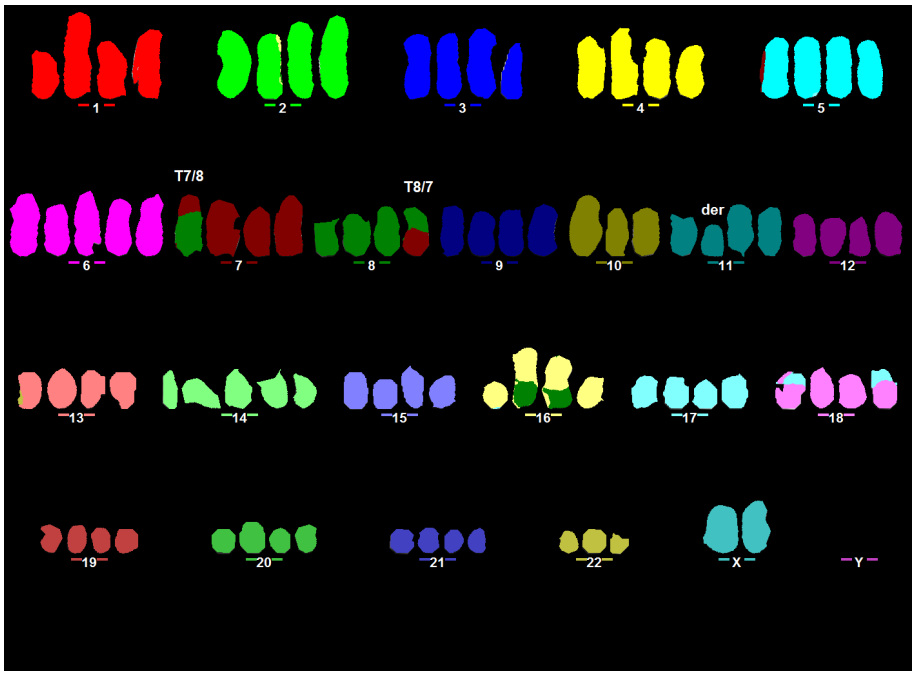


Figure C.76: V-F4 karyotype at harvest 1.

C.3.3 Karyotypes at Harvest 2

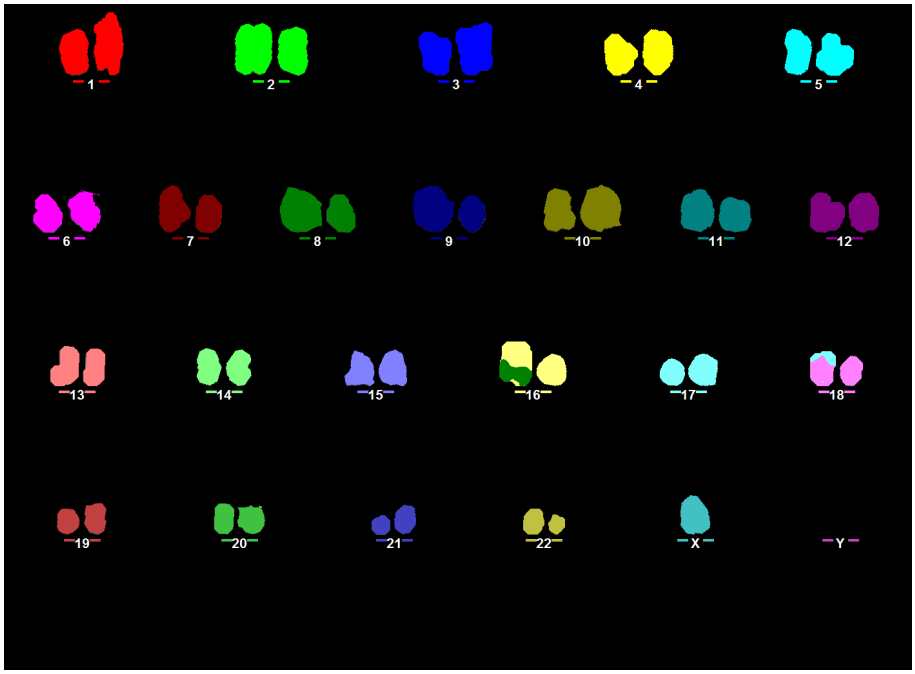


Figure C.77: B-10 karyotype at harvest 2.

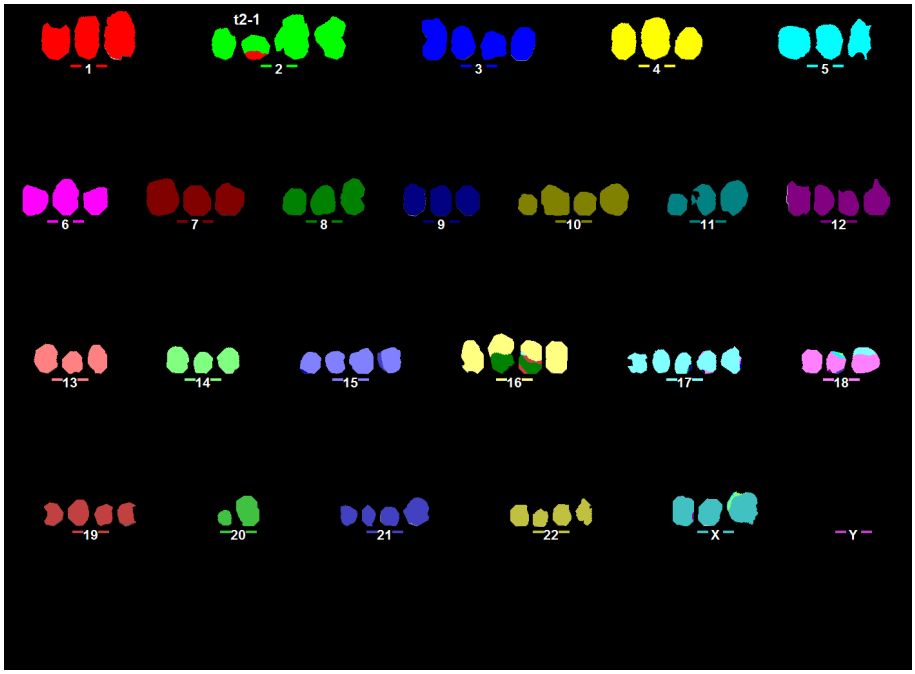


Figure C.78: C-F10 karyotype at harvest 2.

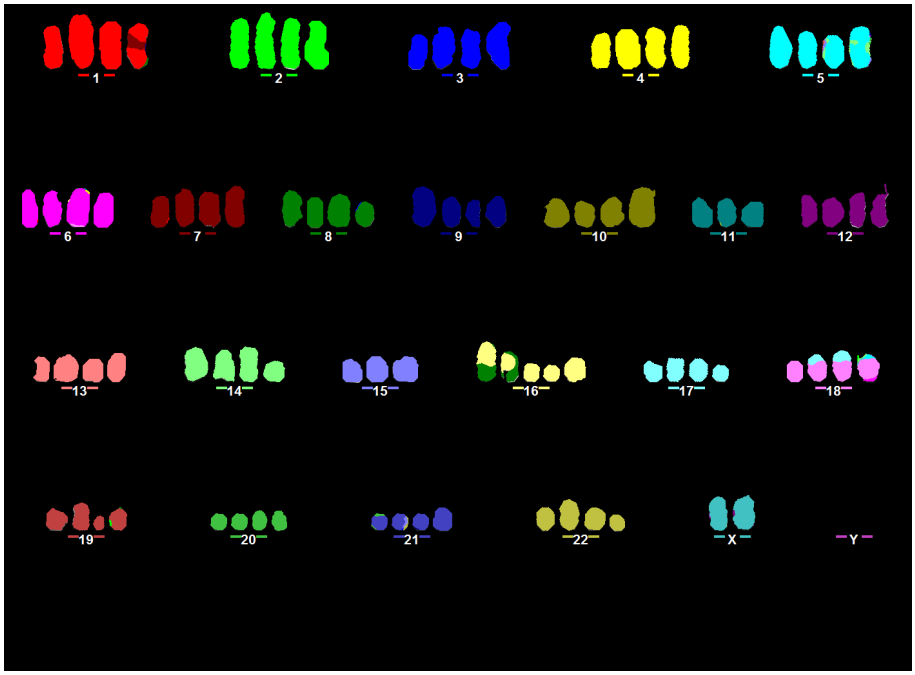


Figure C.79: C-F3 karyotype at harvest 2.

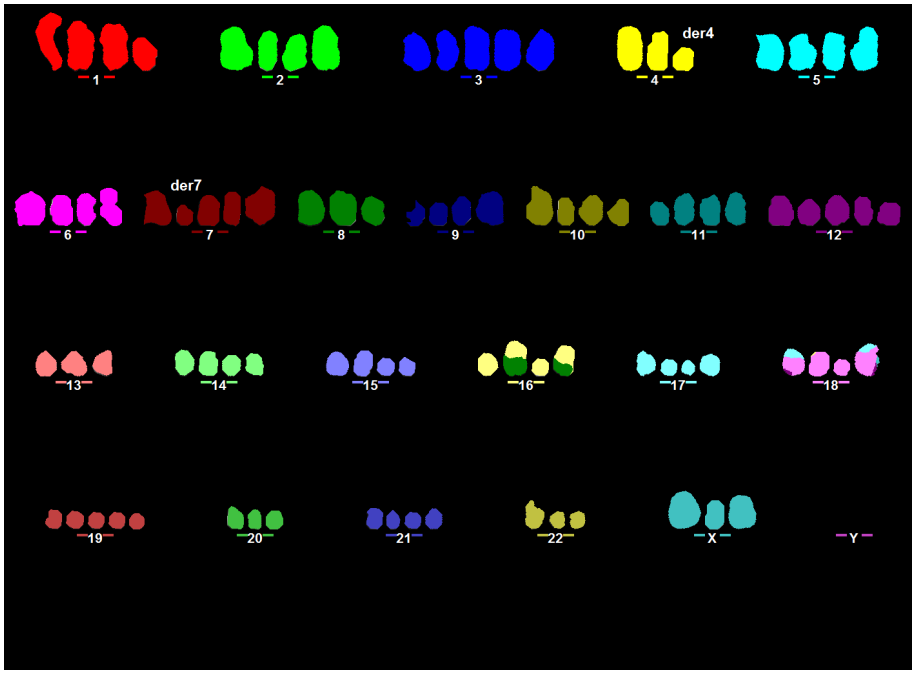


Figure C.80: C-G9 karyotype at harvest 2.

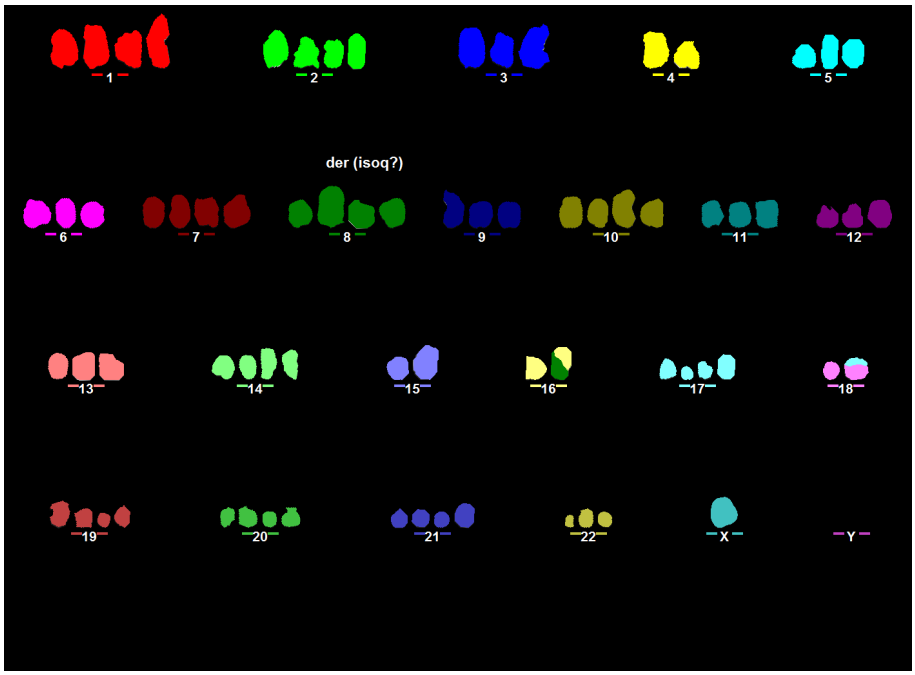


Figure C.81: V-F4 karyotype at harvest 2.

C.3.4 Karyotypes at Harvest 3

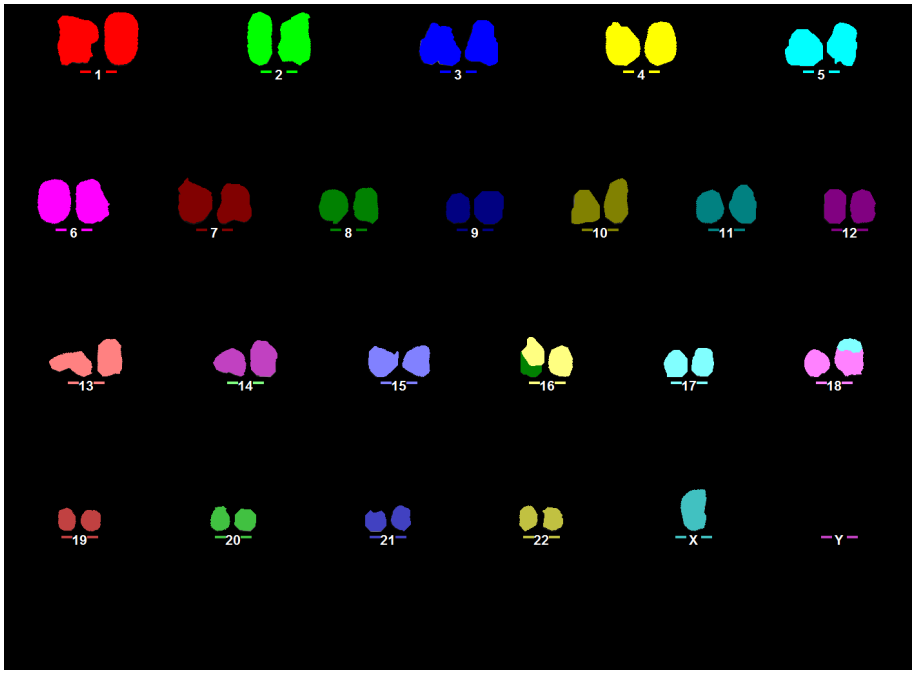


Figure C.82: M-B10 karyotype at harvest 3.

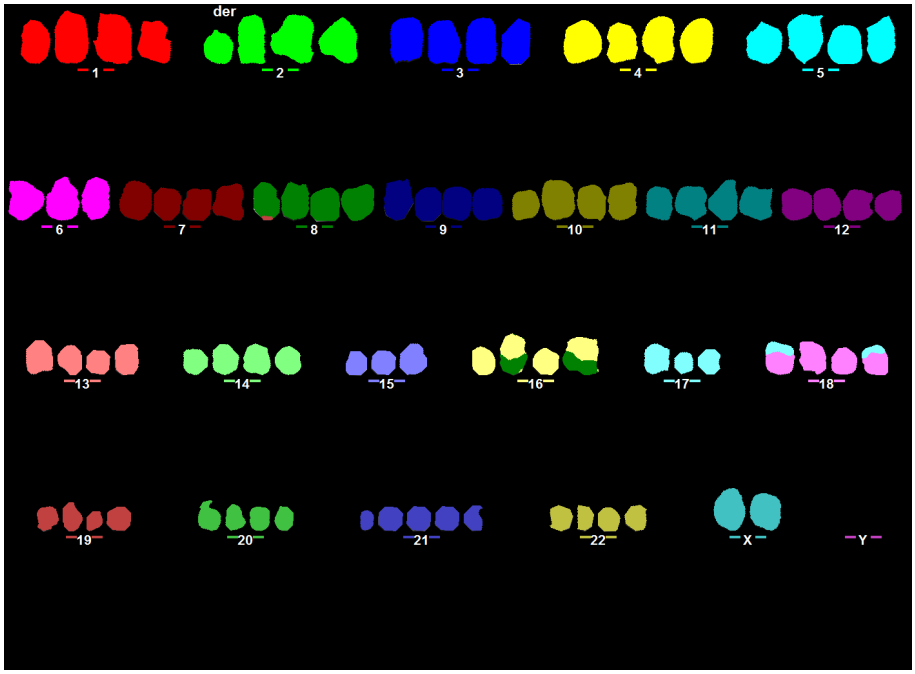


Figure C.83: C-F10 karyotype at harvest 3.

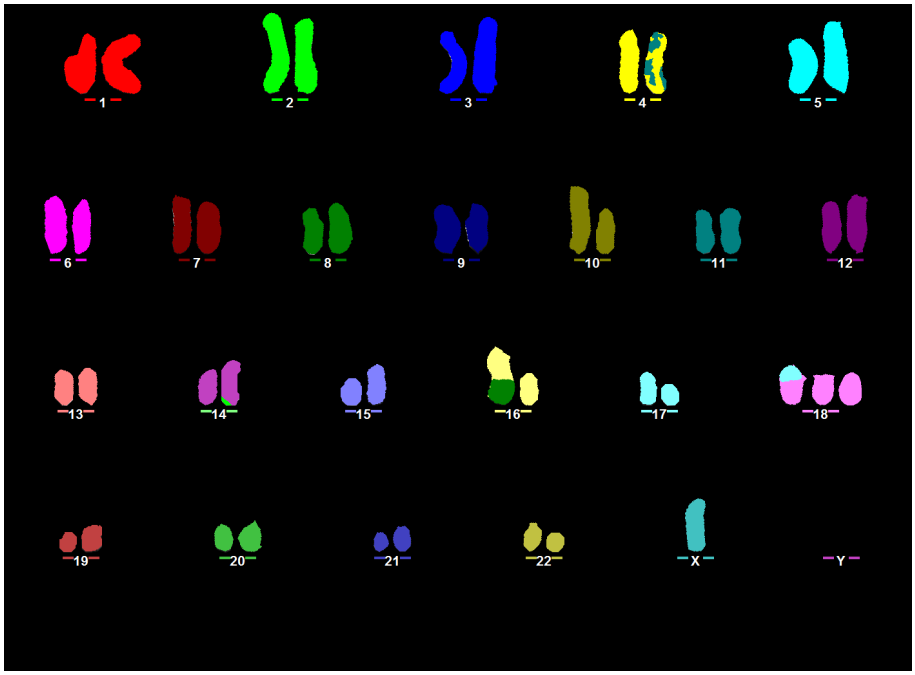


Figure C.84: C-F3 karyotype at harvest 3.

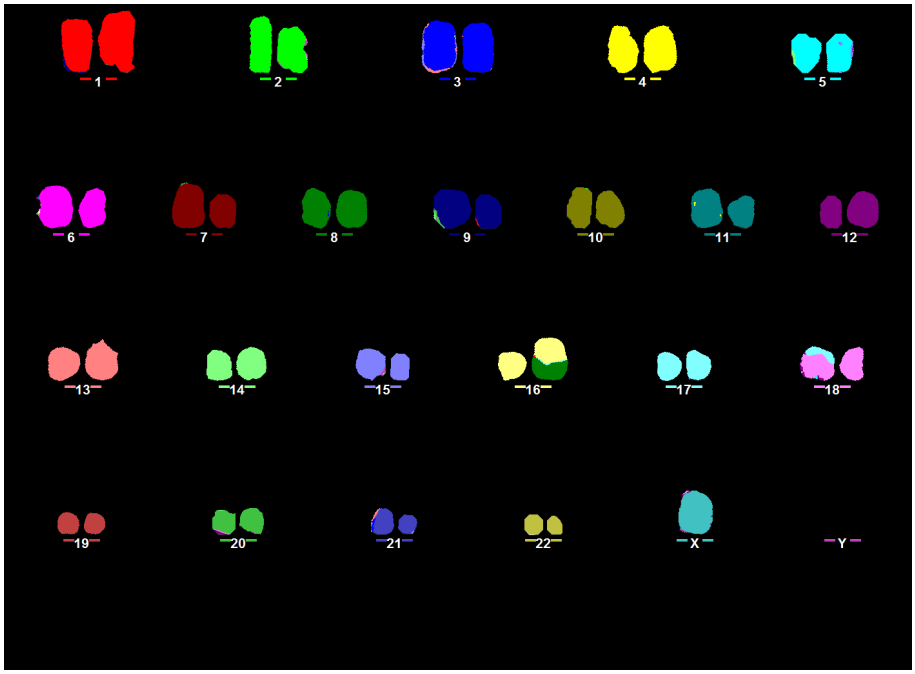


Figure C.85: C-G9 karyotype at harvest 3.

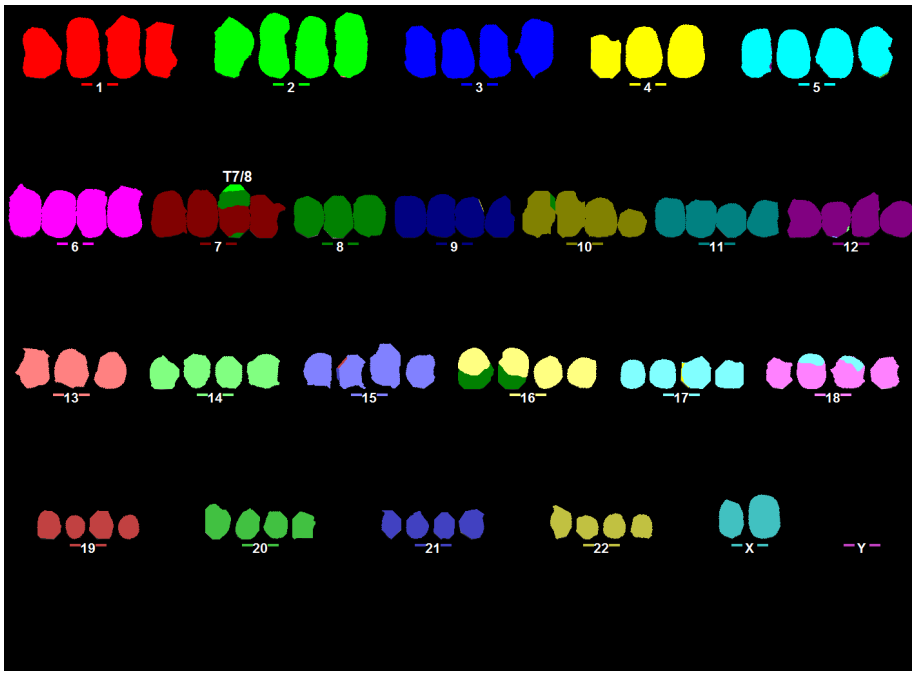


Figure C.86: V-F4 karyotype at harvest 3.

Appendix D

Appendix: Abbreviations

D.1 Abbreviations used in text

A

A3A: APOBEC3A

A3B: APOBEC3B

A3G: APOBEC3G

AIC: Akaike Information Criterion

AID: Activation Induced cytidine Deaminase

AML: Acute myeloid leukemia

ANOVA: Analysis of variance

APC: Adenomatous polyposis coli

APOBEC: Apolipoprotein B mRNA editing enzyme catalytic polypeptide

B

BER: Base excision repair

BIC: Bayesian Information Criterion

BMDC: Bone marrow derived cell

bp: base pairs

BSA: Bovine serum albumin

C

CCOC: Clear Cell Ovarian Carcinoma

cDNA: Complementary DNA

CFP: Cyan Fluorescent Protein

CFS: Common Fragile Site

CI: Confidence Interval

CN: Copy Number

CNS: Central nervous system

CNV: Copy Number Variation

COSMIC: Catalogue of Somatic Mutations in Cancer

CSR: Class switch recombination

CUP: Cancer of Unknown Primary

CV_{MD}: Coefficient of variation of the mutation density

D

DACO: Data Access Compliance Office

DDR: DNA damage response

DMEM: Dulbecco's Modified Eagle Medium

DNA: Deoxyribonucleic acid

DSB: Double strand break

dsDNA: Double stranded DNA

E

EBV: Epstein-Barr virus

EMT: Epithelial mesenchymal transition

F

FACS: Fluorescent Activated Cell Sorting

FBS: Fetal Bovine Serum

FISH: Fluorescent in situ hybridisation

FU: Fluorouracil

G

GEL: Genomics England

GEO: Gene Expression Omnibus

GFP: Green Fluorescent Protein

GI: Genomic Instability

GO: Gene Ontology

H

HBV: Hepatitis B Virus

HCV: Hepatitis C Virus

HERV: Human Endogenous Retrovirus

HIV: Human Immunodeficiency Virus

HMF: Hartwig Medical Foundation

HPV16/18: Human Papillomavirus 16 and/or 18

HSV: Herpes Simplex Virus

HTLV: Human T-Lymphotripic Virus

I

ICGC: International Cancer Genome Consortium

ID: Insertion and Deletion signature

IFN: Interferon

INDEL: Insertion and/or deletion

J

K

KSV: Kaposi sarcoma virus

L

LFC: log2 Fold Change

LMM: Linear mixed effects model

M

MAF: Minor Allele Frequency

MLV: Murine Leukemia Virus

M-FISH: Multicolor Fluorescent In Situ Hybridisation

MLV: Murine leukemia virus

MMEC: Mouse mammary epithelial cells

MMR: Mismatch repair

MOI: Multiplicity of infection

MQ: Mapping quality

mRNA: messenger RNA

MSC: Mesenchymal stem cell

MSigDB: Molecular Signatures Database

MT: Microtubule

N

NB: Negative Binomial

NHEJ: Non-homologous end joining

NET: NeuroEndocrine Tumour/neoplasm/carcinoma

O

OSCC: Oral squamous cell carcinoma

P

PBMC: Peripheral blood mononuclear cell

PBS: Phosphate buffered saline

PC: Principal component

PCA: Principal component analysis

PCC: Premature chromosome condensation

PCAWG: Pan-Cancer Analysis of Whole Genomes

PEG: Polyethylene glycol

PFA: Paraformaldehyde

PGA: Proportion/Percentage of the Genome Altered

PM_1: Proportion of mutations with multiplicity 1

PSA: Prostate specific antigen

Q

R

rlog: regularised logarithm

RNA: Ribonucleic acid

RNAseq: Ribonucleic acid sequencing

S

SBS: Single base substitution

SFV: Simian foamy virus

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variation

SRA: Sequence Read Archive

ssDNA: Single stranded DNA

SSPE: Subacute sclerosing panencephalitis

ssRNA: Single stranded RNA

STR: Short tandem repeat

SV: Structural variant

T

TCGA: The Cancer Genome Atlas

TLS: Translesion DNA synthesis

TT: Tumour type

TTV: Torque Teno Virus

U

UDG: Uracil DNA glycosylase

UFB: Ultrafine bridge

UMI: Unique molecular identifier

UNG: Uracil-N-glycosylas

V

VAF: Variant Allele Frequency

Vif: Viral infectivity factor

VSV-G: Vesicular Stomatitis Virus Glycoprotein

VZV: Varicella-zoster virus

W

WGA: Wheat Germ Agglutinin

WGD: Whole Genome Duplicate/Duplicated/Duplication

WGS: Whole Genome Sequencing

X

Y

Z

ZINB: Zero Inflated Negative Binomial

D.2 Tumour Type Abbreviations

Table D.1: List of tumour type abbreviations used in the PCAWG project

Abbreviation	Tumour Type or Project Name
BLCA	Bladder Urothelial Carcinoma
BOCA	Bone Cancer
BRCA	Breast Cancer
BTCA	Biliary Tract Cancer
CESC	Cervical Squamous Cell Carcinoma
CLLE	Chronic Lymphocytic Leukemia
CMDI	Chronic Myeloid Disorders
COAD	Colon Adenocarcinoma
DLBC	Diffuse Large B-Cell Lymphoma
EOPC	Early Onset Prostate Cancer
ESAD	Esophageal Adenocarcinoma
GACA	Gastric Cancer
GBM	Glioblastoma multiforme
HNSC	Head and Neck Squamous Cell Carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LICA	Liver Cancer
LIHC	Liver Hepatocellular carcinoma
LINC	Liver Cancer
LIRI	Liver Cancer
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MALY	Malignant Lymphoma
ORCA	Oral Cancer
OV	Ovarian Cancer
PACA	Pancreatic Cancer
PAEN	Pancreatic Cancer Endocrine Neoplasms
PBCA	Pediatric Brain Cancer
PRAD	Prostate Cancer of Adenocarcinoma
READ	Rectum Adenocarcinoma
RECA	Renal Clear Cell Carcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous melanoma
STAD	Gastric Adenocarcinoma
THCA	Thyroid Cancer
UCEC	Uterine Corpus Endometrial Carcinoma

References

Ahlmann-Eltze, C. (2019). *ggsignif: Significance Brackets for 'ggplot2'*. R package version 0.6.0.

Aichel, O. (1911). *Über Zellverschmelzung mit qualitativ abnormer Chromosomenverteilung als Ursache der Geschwulstbildung*. Number 13. W. Engelmann.

Akre, M. K., G. J. Starrett, J. S. Quist, N. A. Temiz, M. A. Carpenter, A. N. Tutt, A. Grigoriadis, and R. S. Harris (2016). Mutation processes in 293-based clones overexpressing the DNA cytosine deaminase APOBEC3B. *PloS One 11*(5).

Alexandrov, L. B., P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics 47*(12), 1402.

Alexandrov, L. B., J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, et al. (2020). The repertoire of mutational signatures in human cancer. *Nature 578*(7793), 94–101.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, et al. (2013). Signatures of mutational processes in human cancer. *Nature 500*(7463), 415.

Alvarez-Dolado, M., R. Pardal, J. M. Garcia-Verdugo, J. R. Fike, H. O. Lee, K. Pfeffer, C. Lois, S. J. Morrison, and A. Alvarez-Buylla (2003). Fusion of bone-

marrow-derived cells with Purkinje neurons, cardiomyocytes and hepatocytes. *Nature 425*(6961), 968–973.

Andersen, T. L., P. Boissy, T. Sondergaard, K. Kupisiewicz, T. Plesner, T. Rasmussen, J. Haaber, S. Kølvraa, and J.-M. Delaissé (2007). Osteoclast nuclei of myeloma patients show chromosome translocations specific for the myeloma cell clone: a new type of cancer–host partnership? *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland 211*(1), 10–17.

Arikawa, J., I. Takashima, and N. Hashimoto (1985). Cell fusion by haemorrhagic fever with renal syndrome (HFRS) viruses and its application for titration of virus infectivity and neutralizing antibody. *Archives of Virology 86*(3-4), 303–313.

Atkin, N. B. (1979). Premature chromosome condensation in carcinoma of the bladder: presumptive evidence for fusion of normal and malignant cells. *Cytogenetic and Genome Research 23*(3), 217–219.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

Bailey, P., D. K. Chang, K. Nones, A. L. Johns, A.-M. Patch, M.-C. Gingras, D. K. Miller, A. N. Christ, T. J. Bruxner, M. C. Quinn, et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature 531*(7592), 47–52.

Bakhoum, S. F., B. Ngo, A. M. Laughney, J.-A. Cavallo, C. J. Murphy, P. Ly, P. Shah, R. K. Sriram, T. B. Watkins, N. K. Taunk, et al. (2018). Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature 553*(7689), 467–472.

Bär, S., A. Takada, Y. Kawaoka, and M. Alizon (2006). Detection of cell-cell fusion mediated by Ebola virus glycoproteins. *Journal of Virology 80*(6), 2815–2822.

Bassnett, S., Y. Shi, and G. F. Vrensen (2011). Biological glass: structural determinants of eye lens transparency. *Philosophical Transactions of the Royal Society B: Biological Sciences 366*(1568), 1250–1264.

Bates, D., M. Maechler, B. Bolker, S. Walker, et al. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version 1*(7), 1–23.

Bayani, J., S. Selvarajah, G. Maire, B. Vukovic, K. Al-Romaih, M. Zielenska, and J. A. Squire (2007). Genomic mechanisms and measurement of structural and numerical instability in cancer cells. In *Seminars in cancer biology*, Volume 17, pp. 5–18. Elsevier.

Bayliss, G. J. and H. Wolf (1980). Epstein–Barr virus-induced cell fusion. *Nature 287*(5778), 164–165.

Benjamin, D., T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein (2019). Calling somatic SNVs and INDELs with mutect2. *BioRxiv*, 861054.

Berger, G., S. Durand, G. Fargier, X.-N. Nguyen, S. Cordeil, S. Bouaziz, D. Muriaux, J.-L. Darlix, and A. Cimarelli (2011). APOBEC3A is a specific inhibitor of the early phases of HIV-1 infection in myeloid cells. *PLoS Pathogens 7*(9), e1002221.

Bettadapur, A., H. W. Miller, and K. S. Ralston (2020). Biting off what can be chewed: trogocytosis in health, infection, and disease. *Infection and immunity 88*(7), e00930–19.

Bhattacharya, A., R. D. Bense, C. G. Urzúa-Traslaviña, E. G. de Vries, M. A. van Vugt, and R. S. Fehrmann (2020). Transcriptional effects of copy number alterations in a large set of human cancers. *Nature Communications 11*(1), 1–12.

Bielski, C. M., A. Zehir, A. V. Penson, M. T. Donoghue, W. Chatila, J. Armenia, M. T. Chang, A. M. Schram, P. Jonsson, C. Bandlamudi, et al. (2018). Genome

doubling shapes the evolution and prognosis of advanced cancers. *Nature Genetics 50*(8), 1189–1195.

Birkbak, N. J. and N. McGranahan (2020). Cancer genome evolutionary trajectories in metastasis. *Cancer Cell 37*(1), 8–19.

Bishop, K. N., R. K. Holmes, A. M. Sheehy, N. O. Davidson, S.-J. Cho, and M. H. Malim (2004). Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Current biology 14*(15), 1392–1396.

Blighe, K., S. Rana, and M. Lewis (2020). *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling*. R package version 1.5.4.

Blokzijl, F., J. De Ligt, M. Jager, V. Sasselli, S. Roerink, N. Sasaki, M. Huch, S. Boymans, E. Kuijk, P. Prins, et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature 538*(7624), 260–264.

Bossart, K. N., L.-F. Wang, M. N. Flora, K. B. Chua, S. K. Lam, B. T. Eaton, and C. C. Broder (2002). Membrane fusion tropism and heterotypic functional activities of the Nipah virus and Hendra virus envelope glycoproteins. *Journal of Virology 76*(22), 11186–11198.

Branigan, P. J., N. D. Day, C. Liu, L. L. Gutshall, J. A. Melero, R. T. Sarisky, and A. M. Del Vecchio (2006). The cytoplasmic domain of the F protein of human respiratory syncytial virus is not required for cell fusion. *Journal of general virology 87*(2), 395–398.

Brody, Y., R. J. Kimmerling, Y. E. Maruvka, D. Benjamin, J. J. Elacqua, N. J. Haradhvala, J. Kim, K. W. Mouw, K. Frangaj, A. Koren, et al. (2018). Quantification of somatic mutation flow across individual cell division events by lineage sequencing. *Genome Research 28*(12), 1901–1918.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*(2), 378–400.

Brown, W. L., E. K. Law, P. P. Argyris, M. A. Carpenter, R. Levin-Klein, A. N. Ranum, A. M. Molan, C. L. Forster, B. D. Anderson, L. Lackey, et al. (2019). A rabbit monoclonal antibody against the antiviral and cancer genomic DNA mutating enzyme APOBEC3B. *Antibodies 8*(3), 47.

Brukman, N. G., B. Uygur, B. Podbilewicz, and L. V. Chernomordik (2019). How cells fuse. *Journal of Cell Biology 218*(5), 1436–1451.

Burns, M. B., L. Lackey, M. A. Carpenter, A. Rathore, A. M. Land, B. Leonard, E. W. Refsland, D. Kotandeniya, N. Tretyakova, J. B. Nikas, D. Yee, N. A. Temiz, D. E. Donohue, R. M. McDougle, W. L. Brown, E. K. Law, and R. S. Harris (2013). APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature 494*(7437), 366–370.

Burns, M. B., N. A. Temiz, and R. S. Harris (2013). Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nature Genetics 45*(9), 977.

Cameron, D. L., J. Schröder, J. S. Penington, H. Do, R. Molania, A. Dobrovic, T. P. Speed, and A. T. Papenfuss (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research 27*(12), 2050–2060.

Campbell, P. J., G. Getz, J. O. Korbel, J. M. Stuart, J. L. Jennings, L. D. Stein, M. D. Perry, H. K. Nahal-Bose, B. F. F. Ouellette, C. H. Li, E. Rheinbay, G. P. Nielsen, D. C. Sgroi, C.-L. Wu, W. C. Faquin, V. Deshpande, P. C. Boutros, A. J. Lazar, K. A. Hoadley, D. N. Louis, L. J. Dursi, C. K. Yung, M. H. Bailey, G. Saksena, K. M. Raine, I. Buchhalter, K. Kleinheinz, M. Schlesner, J. Zhang, W. Wang,

D. A. Wheeler, L. Ding, J. T. Simpson, B. D. O'Connor, S. Yakneen, K. Ellrott, N. Miyoshi, A. P. Butler, R. Royo, S. I. Shorser, M. Vazquez, T. Rausch, G. Tiao, S. M. Waszak, B. Rodriguez-Martin, S. Shringarpure, D.-Y. Wu, G. M. Demidov, O. Delaneau, S. Hayashi, S. Imoto, N. Habermann, A. V. Segre, E. Garrison, A. Cafferkey, E. G. Alvarez, J. Heredia-Genestar, F. Muyas, O. Drechsel, A. L. Bruzos, J. Temes, J. Zamora, A. Baez-Ortega, H.-L. Kim, R. J. Mashl, K. Ye, A. DiBiase, K.-l. Huang, I. Letunic, M. D. McLellan, S. J. Newhouse, T. Shmaya, S. Kumar, D. C. Wedge, M. H. Wright, V. D. Yellapantula, M. Gerstein, E. Khurana, T. Marques-Bonet, A. Navarro, C. D. Bustamante, R. Siebert, H. Nakagawa, D. F. Easton, S. Ossowski, J. M. C. Tubio, F. M. De La Vega, X. Estivill, D. Yuen, G. L. Mihaiescu, L. Omberg, V. Ferretti, R. Sabarinathan, O. Pich, A. Gonzalez-Perez, A. Taylor-Weiner, M. W. Fittall, J. Demeulemeester, M. Tarabichi, N. D. Roberts, P. Van Loo, I. Cortés-Ciriano, L. Urban, P. Park, B. Zhu, E. Pitkänen, Y. Li, N. Saini, L. J. Klimczak, J. Weischenfeldt, N. Sidiropoulos, L. B. Alexandrov, R. Rabionet, G. Escaramis, M. Bosio, A. Z. Holik, H. Susak, A. Prasad, S. Erkek, C. Calabrese, B. Raeder, E. Harrington, S. Mayes, D. Turner, S. Juul, S. A. Roberts, L. Song, R. Koster, L. Mirabello, X. Hua, T. J. Tanskanen, M. Tojo, J. Chen, L. A. Aaltonen, G. Rätsch, R. F. Schwarz, A. J. Butte, A. Brazma, S. J. Chanock, N. Chatterjee, O. Stegle, O. Harismendy, G. S. Bova, D. A. Gordenin, D. Haan, L. Sieverling, L. Feuerbach, D. Chalmers, Y. Joly, B. Knoppers, F. Molnár-Gábor, M. Phillips, A. Thorogood, D. Townend, M. Goldman, N. A. Fonseca, Q. Xiang, B. Craft, E. Piñeiro-Yáñez, A. Muñoz, R. Petryszak, A. Füllgrabe, F. Al-Shahrour, M. Keays, D. Haussler, J. Weinstein, W. Huber, A. Valencia, I. Papatheodorou, J. Zhu, Y. Fan, D. Torrents, M. Bieg, K. Chen, Z. Chong, K. Cibulskis, R. Eils, R. S. Fulton, J. L. Gelpi, S. Gonzalez, I. G. Gut, F. Hach, M. Heinold, T. Hu, V. Huang, B. Hutter, N. Jäger, J. Jung, Y. Kumar, C. Lalansingh, I. Leshchiner, D. Livitz, E. Z. Ma, Y. E. Maruvka, A. Milovanovic, M. M. Nielsen, N. Paramasivam, J. S. Pedersen, M. Puiggròs, S. C. Sahinalp, I. Sarrafi,

C. Stewart, M. D. Stobbe, J. A. Wala, J. Wang, M. Wendl, J. Werner, Z. Wu, H. Xue, T. N. Yamaguchi, V. Yellapantula, B. N. Davis-Dusenbery, R. L. Grossman, Y. Kim, M. C. Heinold, J. Hinton, D. R. Jones, A. Menzies, L. Stebbings, J. M. Hess, M. Rosenberg, A. J. Dunford, M. Gupta, M. Imielinski, M. Meyerson, R. Beroukhim, J. Reimand, P. Dhingra, F. Favero, S. Dentro, J. Wintersinger, V. Rudneva, J. W. Park, E. P. Hong, S. G. Heo, A. Kahles, K.-V. Lehmann, C. M. Soulette, Y. Shiraishi, F. Liu, Y. He, D. Demircioğlu, N. R. Davidson, L. Greger, S. Li, D. Liu, S. G. Stark, F. Zhang, S. B. Amin, P. Bailey, A. Chateigner, M. Frenkel-Morgenstern, Y. Hou, M. R. Huska, H. Kilpinen, F. C. Lamaze, C. Li, X. Li, X. Li, X. Liu, M. G. Marin, J. Markowski, T. Nandi, A. I. Ojesina, Q. Pan-Hammarström, P. J. Park, C. S. Pedamallu, H. Su, P. Tan, B. T. Teh, J. Wang, H. Xiong, C. Ye, C. Yung, X. Zhang, L. Zheng, S. Zhu, P. Awadalla, C. J. Creighton, K. Wu, H. Yang, J. Göke, Z. Zhang, A. N. Brooks, M. W. Fittall, I. Martincorena, C. Rubio-Perez, M. Juul, S. Schumacher, O. Shapira, D. Tamborero, L. Mularoni, H. Hornshøj, J. Deu-Pons, F. Muiños, J. Bertl, Q. Guo, and T. I. P.-C. A. of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature 578*(7793), 82–93.

Cannan, W. J. and D. S. Pederson (2016). Mechanisms and consequences of double-strand DNA break formation in chromatin. *Journal of cellular physiology 231*(1), 3–14.

Carter, S. L., K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology 30*(5), 413–421.

Casellas, R., U. Basu, W. T. Yewdell, J. Chaudhuri, D. F. Robbiani, and J. M. Di Noia (2016). Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. *Nature Reviews Immunology 16*(3), 164–176.

Castilla, V., M. Contigiani, and S. Mersich (2005). Inhibition of cell fusion in junin

virus-infected cells by sera from Argentine hemorrhagic fever patients. *Journal of Clinical Virology 32*(4), 286–288.

Castilla, V., M. Larzábal, N. A. Sgalippa, M. B. Wachsman, and C. E. Coto (2005). Antiviral mode of action of a synthetic brassinosteroid against Junin virus replication. *Antiviral research 68*(2), 88–95.

Castilla, V. and S. Mersich (1996). Low-pH-induced fusion of Vero cells infected with Junin virus. *Archives of Virology 141*(7), 1307–1317.

Cattaneo, R. and J. Rose (1993). Cell fusion by the envelope glycoproteins of persistent measles viruses which caused lethal human brain disease. *Journal of Virology 67*(3), 1493–1502.

Chan, S. Y., C. J. Empig, F. J. Welte, R. F. Speck, A. Schmaljohn, J. F. Kreisberg, and M. A. Goldsmith (2001). Folate receptor-α is a cofactor for cellular entry by Marburg and Ebola viruses. *Cell 106*(1), 117–126.

Chang, D. K., S. M. Grimmond, and A. V. Biankin (2014). Pancreatic cancer genomics. *Current opinion in genetics & development 24*, 74–81.

Chang, S. T., P. Sova, X. Peng, J. Weiss, G. L. Law, R. E. Palermo, and M. G. Katze (2011). Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell line. *MBio 2*(5).

Chen, X., A. Bahrami, A. Pappo, J. Easton, J. Dalton, E. Hedlund, D. Ellison, S. Shurtleff, G. Wu, L. Wei, et al. (2014). Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Reports 7*(1), 104–112.

Chen, X., O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, and C. T. Saunders (2016). Manta: rapid detection of

structural variants and INDELs for germline and cancer sequencing applications. *Bioinformatics 32*(8), 1220–1222.

Chen, Z., W. Wen, J. Bao, K. L. Kuhs, Q. Cai, J. Long, X.-o. Shu, W. Zheng, and X. Guo (2019). Integrative genomic analyses of APOBEC-mutational signature, expression and germline deletion of APOBEC3 genes, and immunogenicity in multiple cancer types. *BMC medical genomics 12*(1), 1–13.

Chernomordik, L. V. and M. M. Kozlov (2008). Mechanics of membrane fusion. *Nature structural & molecular biology 15*(7), 675–683.

Chitwood, C. A., C. Dietzsch, G. Jacobs, T. McArdle, B. T. Freeman, A. Banga, F. K. Noubissi, and B. M. Ogle (2018). Breast tumor cell hybrids form spontaneously in vivo and contribute to breast tumor metastases. *APL bioengineering 2*(3), 031907.

Choi, Y., A. P. Chan, E. Kirkness, A. Telenti, and N. J. Schork (2018). Comparison of phasing strategies for whole human genomes. *PLoS Genetics 14*(4), e1007308.

Ciriello, G., M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics 45*(10), 1127–1133.

Clawson, G. A., G. L. Matters, P. Xin, Y. Imamura-Kawasawa, Z. Du, D. M. Thiboutot, K. F. Helm, R. I. Neves, and T. Abraham (2015). Macrophage-tumor cell fusions from peripheral blood of melanoma patients. *PLoS One 10*(8), e0134320.

Cole, N. L. and C. Grose (2003). Membrane fusion mediated by herpesvirus glycoproteins: the paradigm of varicella-zoster virus. *Reviews in medical virology 13*(4), 207–222.

Conticello, S. G., C. J. Thomas, S. K. Petersen-Mahrt, and M. S. Neuberger (2005).

Evolution of the AID/APOBEC family of polynucleotide (deoxy) cytidine deaminases. *Molecular biology and evolution 22*(2), 367–377.

Cooke, D. P., D. C. Wedge, and G. Lunter (2021). A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology*, 1–8.

Coombes, K. R. (2019). *BimodalIndex: The Bimodality Index*. R package version 1.1.9.

Cortés-Ciriano, I., J. J.-K. Lee, R. Xi, D. Jain, Y. L. Jung, L. Yang, D. Gordenin, L. J. Klimczak, C.-Z. Zhang, D. S. Pellman, et al. (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics 52*(3), 331–341.

Currall, B. B., C. Chiangmai, M. E. Talkowski, and C. C. Morton (2013). Mechanisms for structural variation in the human genome. *Current genetic medicine reports 1*(2), 81–90.

Daley, T. and A. D. Smith (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods 10*(4), 325–327.

Dargent, J.-L., L. Lespagnard, A. Kornreich, P. Hermans, N. Clumeck, and A. Verhest (2000). HIV-associated multinucleated giant cells in lymphoid tissue of the Waldeyer's ring: A detailed study. *Modern Pathology 13*(12), 1293–1299.

Davoli, T., E. L. Denchi, and T. de Lange (2010). Persistent telomere damage induces bypass of mitosis and tetraploidy. *Cell 141*(1), 81–93.

De Baetselier, P., E. Gorelik, Z. Eshhar, Y. Ron, S. Katzav, M. Feldman, and S. Segal (1981). Metastatic properties conferred on nonmetastatic tumors by hybridization of spleen B-lymphocytes with plasmacytoma cells. *Journal of the National Cancer Institute 67*(5), 1079–1087.

De Baetselier, P., E. Roos, L. Brys, L. Remels, and M. Feldman (1984). Generation of invasive and metastatic variants of a non-metastatic T-cell lymphoma by in vivo fusion with normal host cells. *International journal of cancer 34*(5), 731–738.

Delespaul, L., C. Gélabert, T. Lesluyes, S. Le Guellec, G. Pérot, L. Leroy, J. Baud, C. Merle, L. Lartigue, and F. Chibon (2020). Cell–cell fusion of mesenchymal cells with distinct differentiations triggers genomic and transcriptomic remodelling toward tumour aggressiveness. *Scientific Reports 10*(1), 1–12.

Delespaul, L., C. Merle, T. Lesluyes, P. Lagarde, S. Le Guellec, G. Pérot, J. Baud, M. Carlotti, C. Danet, M. Fèvre, et al. (2019). Fusion-mediated chromosomal instability promotes aneuploidy patterns that resemble human tumors. *Oncogene*, 1.

Dentro, S. and D. Wedge (2020). *dpclust3p: DPClust pre-processing*. R package version 1.0.8.

Dentro, S. C., D. C. Wedge, and P. Van Loo (2017). Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor perspectives in medicine 7*(8), a026625.

Derks, K. W., J. H. Hoeijmakers, and J. Pothof (2014). The DNA damage response: the omics era and its impact. *DNA repair 19*, 214–220.

Dewhurst, S. M., N. McGranahan, R. A. Burrell, A. J. Rowan, E. Grönroos, D. Endesfelder, T. Joshi, D. Mouradov, P. Gibbs, R. L. Ward, et al. (2014). Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer discovery 4*(2), 175–185.

Di Tommaso, P., M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology 35*(4), 316–319.

Dillekås, H., M. S. Rogers, and O. Straume (2019). Are 90% of deaths from cancer caused by metastases? *Cancer medicine 8*(12), 5574–5576.

Dittmar, T., S. Schwitalla, J. Seidel, S. Haverkampf, G. Reith, S. Meyer-Staeckling, B. H. Brandt, B. Niggemann, and K. S. Zänker (2011). Characterization of hybrid cells derived from spontaneous fusion events between breast epithelial cells exhibiting stem-like characteristics and breast cancer cells. *Clinical & experimental metastasis 28*(1), 75–90.

Duelli, D. and Y. Lazebnik (2007). Cell-to-cell fusion as a link between viruses and cancer. *Nature Reviews Cancer 7*(12), 968–976.

Duelli, D. M., H. M. Padilla-Nash, D. Berman, K. M. Murphy, T. Ried, and Y. Lazebnik (2007). A virus causes cancer by inducing massive chromosomal instability through cell fusion. *Current Biology 17*(5), 431–437.

Duncan, A. W., M. H. Taylor, R. D. Hickey, A. E. H. Newell, M. L. Lenzi, S. B. Olson, M. J. Finegold, and M. Grompe (2010). The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature 467*(7316), 707–710.

Edgar, R., M. Domrachev, and A. E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research 30*(1), 207–210.

Elena Zaitseva, Aditya Mittal, D. E. G. and L. V. Chernomordik (2005). Class II fusion protein of alphaviruses drives membrane fusion through the same pathway as class I proteins. *The Journal of Cell Biology 169*(1), 167–177.

Ellisen, L. W., K. D. Ramsayer, C. M. Johannessen, A. Yang, H. Beppu, K. Minda, J. D. Oliner, F. McKeon, and D. A. Haber (2002). REDD1, a developmentally regulated transcriptional target of p63 and p53, links p63 to regulation of reactive oxygen species. *Molecular Cell 10*(5), 995–1005.

Esiri, M., D. Oppenheimer, B. Brownell, and M. Haire (1982). Distribution of measles antigen and immunoglobulin-containing cells in the CNS in subacute sclerosing panencephalitis (SSPE) and atypical measles encephalitis. *Journal of the neurological sciences 53*(1), 29–43.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics 32*(19), 3047–3048.

Ewels, P. A., A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology 38*(3), 276–278.

Fantini, D., V. Vidimar, Y. Yu, S. Condello, and J. Meeks (2020). *MutSignatures: an R package for extraction and analysis of cancer mutational signatures*. R package version 2.1.1.

Farlik, M., F. Halbritter, F. Müller, F. A. Choudry, P. Ebert, J. Klughammer, S. Farrow, A. Santoro, V. Ciaurro, A. Mathur, et al. (2016). Dna methylation dynamics of human hematopoietic stem cell differentiation. *Cell stem cell 19*(6), 808–822.

Fisher, R. A. (1970). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Freeman, B. T., J. P. Jung, and B. M. Ogle (2016). Single-cell RNA-seq reveals activation of unique gene groups as a consequence of stem cell-parenchymal cell fusion. *Scientific Reports 6*(1), 1–9.

Frey, A. G., D. J. Palenchar, J. D. Wildemann, and C. C. Philpott (2016). A glutaredoxin· BolA complex serves as an iron-sulfur cluster chaperone for the cytosolic cluster assembly machinery. *Journal of Biological Chemistry 291*(43), 22344–22356.

Fujiwara, T., M. Bandi, M. Nitta, E. V. Ivanova, R. T. Bronson, and D. Pellman (2005). Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature 437*(7061), 1043–1047.

Furrer, E., O. Planz, and L. Stitz (2004). Inhibition of Borna disease virus-mediated cell fusion by monoclonal antibodies directed against the viral glycoprotein. *Intervirology 47*(2), 108–113.

Galofré, C., Ö. Gönül Geyik, E. Asensio, D. Wangsa, D. Hirsch, C. Parra, J. Saez, M. Mollà, Z. Yüce, A. Castells, et al. (2020). Tetraploidy-associated genetic heterogeneity confers chemo-radiotherapy resistance to colorectal cancer cells. *Cancers 12*(5), 1118.

Ganem, N. J., S. A. Godinho, and D. Pellman (2009). A mechanism linking extra centrosomes to chromosomal instability. *Nature 460*(7252), 278–282.

Garnier, S. (2018). *viridis: Default Color Maps from 'matplotlib'*. R package version 0.5.1.

Gast, C. E., A. D. Silk, L. Zarour, L. Riegler, J. G. Burkhart, K. T. Gustafson, M. S. Parappilly, M. Roh-Johnson, J. R. Goodman, B. Olson, et al. (2018). Cell fusion potentiates tumor heterogeneity and reveals circulating hybrid cells that correlate with stage and survival. *Science advances 4*(9), eaat7828.

Gehlenborg, N. (2016). *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. R package version 1.4.0.

Gerstung, M., C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu, M. Tarabichi, A. Deshwar, J. Wintersinger, K. Kleinheinz, I. Vázquez-García, K. Haase, L. Jerman, S. Sengupta, G. Macintyre, S. Malikic, N. Donmez, D. G. Livitz, M. Cmero, J. Demeulemeester, S. Schumacher, Y. Fan, X. Yao, J. Lee, M. Schlesner, P. C. Boutros, D. D. Bowtell,

H. Zhu, G. Getz, M. Imielinski, R. Beroukhim, S. C. Sahinalp, Y. Ji, M. Peifer, F. Markowetz, V. Mustonen, K. Yuan, W. Wang, Q. D. Morris, S. C. Dentro, A. G. Deshwar, D. J. Adams, P. C. Boutros, D. D. Bowtell, P. J. Campbell, S. Cao, E. L. Christie, Y. Cun, K. J. Dawson, R. M. Drews, R. Eils, M. Fittall, D. W. Garsed, G. Ha, H. Lee-Six, D. G. Livitz, I. Martincorena, T. J. Mitchell, L. Oesper, M. Peto, B. J. Raphael, S. C. Sahinalp, A. Salcedo, R. Shi, S. J. Shin, O. Spiro, L. D. Stein, S. Vembu, D. A. Wheeler, T.-P. Yang, Q. D. Morris, P. T. Spellman, D. C. Wedge, P. Van Loo, P. T. Spellman, D. C. Wedge, P. E. . H. W. Group, and P. Consortium (2020). The evolutionary history of 2,658 cancers. *Nature 578*(7793), 122–128.

Giglio, S., K. W. Broman, N. Matsumoto, V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, L. Voullaire, D. Larizza, R. Giorda, et al. (2001). Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *The American Journal of Human Genetics 68*(4), 874–883.

Gillison, M. L., K. Akagi, W. Xiao, B. Jiang, R. K. Pickard, J. Li, B. J. Swanson, A. D. Agrawal, M. Zucker, B. Stache-Crain, et al. (2019). Human papillomavirus and the landscape of secondary genetic alterations in oral cancers. *Genome Research 29*(1), 1–17.

Glaser, A. P., D. Fantini, Y. Wang, Y. Yu, K. J. Rimar, J. R. Podojil, S. D. Miller, and J. J. Meeks (2018). APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. *Oncotarget 9*(4), 4537.

Goh, W. C., M. E. Rogel, C. M. Kinsey, S. F. Michael, P. N. Fultz, M. A. Nowak, B. H. Hahn, and M. Emerman (1998). HIV-1 Vpr increases viral expression by manipulation of the cell cycle: a mechanism for selection of Vpr in vivo. *Nature medicine 4*(1), 65–71.

Goila-Gaur, R. and K. Strebel (2008). HIV-1 Vif, APOBEC, and intrinsic immunity. *Retrovirology 5*(1), 51.

Goldenberg, D. M., R. A. Pavia, and M. C. Tsao (1974). In vivo hybridisation of human tumour and normal hamster cells. *Nature 250*(5468), 649–651.

Gonzalez-Dunia, D., B. Cubitt, and J. C. de la Torre (1998). Mechanism of Borna disease virus entry into cells. *Journal of Virology 72*(1), 783–788.

Gonzalez-Scarano, F., N. Pobjecky, and N. Nathanson (1984). La Crosse bunyavirus can mediate pH-dependent fusion from without. *Virology 132*(1), 222–225.

Greengard, O., N. Poltoratskaia, E. Leikina, J. Zimmerberg, and A. Moscona (2000). The anti-influenza virus agent 4-GU-DANA (zanamivir) inhibits cell fusion mediated by human parainfluenza virus and influenza virus HA. *Journal of Virology 74*(23), 11108–11114.

Gu, J., M. Stevens, X. Xing, D. Li, B. Zhang, J. E. Payton, E. M. Oltz, J. N. Jarvis, K. Jiang, T. Cicero, et al. (2016). Mapping of variable dna methylation across multiple cell types defines a dynamic regulatory landscape of the human genome. *G3: Genes, Genomes, Genetics 6*(4), 973–986.

Hall, C. B., K. R. Powell, N. E. MacDonald, C. L. Gala, M. E. Menegus, S. C. Suffin, and H. J. Cohen (1986). Respiratory syncytial viral infection in children with compromised immune function. *New England Journal of Medicine 315*(2), 77–81.

Hanahan, D. and R. A. Weinberg (2011). Hallmarks of cancer: the next generation. *Cell 144*(5), 646–674.

Harper, J. V. and G. Brooks (2005). The mammalian cell cycle. In *Cell Cycle Control*, pp. 113–153. Springer.

Harris, R. S. (2015). Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer. *Breast Cancer Research 17*(1), 8.

Harris, R. S., K. N. Bishop, A. M. Sheehy, H. M. Craig, S. K. Petersen-Mahrt, I. N. Watt, M. S. Neuberger, and M. H. Malim (2003). DNA deamination mediates innate immunity to retroviral infection. *Cell 113*(6), 803–809.

Harris, R. S. and J. P. Dudley (2015). APOBECs and virus restriction. *Virology 479*, 131–145.

Harris, R. S., A. A. Serebrenik, P. Argyris, M. C. Jarvis, W. L. Brown, M. Bazzaro, R. I. Vogel, B. K. Erickson, S.-H. Lee, K. M. Goergen, et al. (2020). The DNA cytosine deaminase APOBEC3B is a molecular determinant of platinum responsiveness in clear cell ovarian cancer. *Clinical Cancer Research*.

Hartig, F. (2020). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.3.3.0.

Hartwig Medical Foundation, . (2020, March). hartwigmedical/pipeline5. Version 5.1.123.

He, X., B. Li, Y. Shao, N. Zhao, Y. Hsu, Z. Zhang, and L. Zhu (2015). Cell fusion between gastric epithelial cells and mesenchymal stem cells results in epithelial-to-mesenchymal transition and malignant transformation. *BMC cancer 15*(1), 1–11.

Helming, L. and S. Gordon (2009). Molecular mediators of macrophage fusion. *Trends in cell biology 19*(10), 514–522.

Herschke, F., S. Plumet, T. Duhen, O. Azocar, J. Druelle, D. Laine, T. F. Wild, C. Rabourdin-Combe, D. Gerlier, and H. Valentin (2007). Cell-cell fusion induced by measles virus amplifies the type I interferon response. *Journal of Virology 81*(23), 12859–12871.

Hlavac, M. (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). R package version 5.2.2.

Hu, L. and B. P. Ceresa (2009). Characterization of the plasma membrane localization and orientation of HPV16 E5 for cell–cell fusion. *Virology 393*(1), 135–143.

Hu, L., K. Plafker, V. Vorozhko, R. E. Zuna, M. H. Hanigan, G. J. Gorbsky, S. M. Plafker, P. C. Angeletti, and B. P. Ceresa (2009). Human papillomavirus 16 E5 induces bi-nucleated cell formation by cell–cell fusion. *Virology 384*(1), 125–134.

Hu, Y., K. Huang, Q. An, G. Du, G. Hu, J. Xue, X. Zhu, C.-Y. Wang, Z. Xue, and G. Fan (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology 17*(1), 1–11.

Iliopoulou, M., R. Nolan, L. Alvarez, Y. Watanabe, C. A. Coomer, G. M. Jakobsdottir, T. A. Bowden, and S. Padilla-Parra (2018). A dynamic three-step mechanism drives the HIV-1 pre-fusion reaction. *Nature Structural & Molecular Biology 25*(9), 814–822.

Jansen, I. D., J. A. Vermeer, V. Bloemen, J. Stap, and V. Everts (2012). Osteoclast fusion and fission. *Calcified tissue international 90*(6), 515–522.

Jeffares, D. C., C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, F. Balloux, C. Dessimoz, J. Bähler, and F. J. Sedlazeck (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications 8*(1), 1–11.

Johnson, R. and H. Harris (1969). DNA synthesis and mitosis in fused cells: I. HeLa homokaryons. *Journal of Cell Science 5*(3), 603–624.

Jones, N. A. and R. J. Geraghty (2004). Fusion activity of lipid-anchored envelope glycoproteins of herpes simplex virus type 1. *Virology 324*(1), 213–228.

Kamentsky, L., T. R. Jones, A. Fraser, M.-A. Bray, D. J. Logan, K. L. Madden, V. Ljosa, C. Rueden, K. W. Eliceiri, and A. E. Carpenter (2011). Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. *Bioinformatics 27*(8), 1179–1180.

Kanu, N., M. A. Cerone, G. Goh, L.-P. Zalmas, J. Bartkova, M. Dietzen, N. McGranahan, R. Rogers, E. K. Law, I. Gromova, et al. (2016). DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biology 17*(1), 185.

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0.

Kassambara, A., M. Kosinski, P. Biecek, and S. Fabian (2017). Package 'survminer'.

Kedzierska, K. (2021). *WES Pipeline*.

Kinzler, E. R. and T. Compton (2005). Characterization of human cytomegalovirus glycoprotein-induced cell-cell fusion. *Journal of Virology 79*(12), 7827–7837.

Kleivi, K., M. R. Teixeira, M. Eknæs, C. B. Diep, K. S. Jakobsen, R. Hamelin, and R. A. Lothe (2004). Genome signatures of colon carcinoma cell lines. *Cancer genetics and cytogenetics 155*(2), 119–131.

Knutsen, T., H. M. Padilla-Nash, D. Wangsa, L. Barenboim-Stapleton, J. Camps, N. McNeil, M. J. Difilippantonio, and T. Ried (2010). Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines. *Genes, Chromosomes and Cancer 49*(3), 204–223.

Kobayashi, M., M. C. Bennett, T. Bercot, and I. R. Singh (2006). Functional analysis of hepatitis C virus envelope proteins, using a cell-cell fusion assay. *Journal of Virology 80*(4), 1817–1825.

Kondo, N., M. Marin, J. H. Kim, T. M. Desai, and G. B. Melikyan (2015). Distinct requirements for HIV-cell fusion and HIV-mediated cell-cell fusion. *Journal of Biological Chemistry 290*(10), 6558–6573.

Koning, F. A., E. N. Newman, E.-Y. Kim, K. J. Kunstman, S. M. Wolinsky, and M. H. Malim (2009). Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. *Journal of Virology 83*(18), 9474–9485.

Köster, J. and S. Rahmann (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics 28*(19), 2520–2522.

Koulakov, A. A. and Y. Lazebnik (2012). The problem of colliding networks and its relation to cell fusion and cancer. *Biophysical journal 103*(9), 2011–2020.

Kovacs, G. (1985). Premature chromosome condensation: evidence for in vivo cell fusion in human malignant tumours. *International journal of cancer 36*(6), 637–641.

Krokan, H. E., P. Sætrom, P. A. Aas, H. S. Pettersen, B. Kavli, and G. Slupphaug (2014). Error-free versus mutagenic processing of genomic uracil—relevance to cancer. *DNA repair 19*, 38–47.

Kutsuzawa, T., T. Konno, H. Suzuki, A. Kapikian, T. Ebina, and N. Ishida (1982). Isolation of human rotavirus subgroups 1 and 2 in cell culture. *Journal of clinical microbiology 16*(4), 727–730.

Kuznetsova, A. Y., K. Seget, G. K. Moeller, M. S. de Pagter, J. A. de Roos, M. Dürrbaum, C. Kuffer, S. Müller, G. J. Zaman, W. P. Kloosterman, et al. (2015). Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell cycle 14*(17), 2810–2820.

Kuzniar, A., J. Maassen, S. Verhoeven, L. Santuari, C. Shneider, W. P. Kloosterman, and J. de Ridder (2020). sv-callers: a highly portable parallel workflow for structural variant detection in whole-genome sequence data. *PeerJ 8*, e8214.

LaBerge, G. S., E. Duvall, Z. Grasmick, K. Haedicke, and J. Pawelek (2017). A melanoma lymph node metastasis with a donor-patient hybrid genome following bone marrow transplantation: A second case of leucocyte-tumor cell hybridization in cancer metastasis. *PLoS One 12*(2).

Lackey, L., E. K. Law, W. L. Brown, and R. S. Harris (2013). Subcellular localization of the APOBEC3 proteins during mitosis and implications for genomic DNA deamination. *Cell Cycle 12*(5), 762–772.

Land, A. M., E. K. Law, M. A. Carpenter, L. Lackey, W. L. Brown, and R. S. Harris (2013). Endogenous APOBEC3A DNA cytosine deaminase is cytoplasmic and nongenotoxic. *Journal of Biological Chemistry 288*(24), 17253–17260.

Landry, S., I. Narvaiza, D. C. Linfesty, and M. D. Weitzman (2011). APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO reports 12*(5), 444–450.

Lartigue, L., C. Merle, P. Lagarde, L. Delespaul, T. Lesluyes, S. Le Guellec, G. Pérot, L. Leroy, J.-M. Coindre, and F. Chibon (2020). Genome remodeling upon mesenchymal tumor cell fusion contributes to tumor progression and metastatic spread. *Oncogene 39*(21), 4198–4211.

Laughney, A. M., S. Elizalde, G. Genovese, and S. F. Bakhoum (2015). Dynamics of tumor heterogeneity derived from clonal karyotypic evolution. *Cell Reports 12*(5), 809–820.

Law, E. K., R. Levin-Klein, M. C. Jarvis, H. Kim, P. P. Argyris, M. A. Carpenter, G. J. Starrett, N. A. Temiz, L. K. Larson, C. Durfee, et al. (2020). APOBEC3A

catalyzes mutation and drives carcinogenesis in vivo. *Journal of Experimental Medicine 217*(12).

Lazova, R., G. S. LaBerge, E. Duvall, N. Spoelstra, V. Klump, M. Sznol, D. Cooper, R. A. Spritz, J. T. Chang, and J. M. Pawelek (2013). A melanoma brain metastasis with a donor-patient hybrid genome following bone marrow transplantation: first evidence for fusion in human cancer. *PloS One 8*(6), e66731.

Leach, N. T. and C. Jackson-Cook (2001). The application of spectral karyotyping (SKY) and fluorescent in situ hybridization (FISH) technology to determine the chromosomal content (s) of micronuclei. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis 495*(1-2), 11–19.

Lee-Six, H., S. Olafsson, P. Ellis, R. J. Osborne, M. A. Sanders, L. Moore, N. Georgakopoulos, F. Torrente, A. Noorani, M. Goddard, et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature 574*(7779), 532–537.

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software 55*(8), 1–24.

Leinonen, R., H. Sugawara, M. Shumway, and I. N. S. D. Collaboration (2010). The sequence read archive. *Nucleic acids research 39*(suppl_1), D19–D21.

Lengyel, E. (2010). Ovarian cancer development and metastasis. *The American journal of pathology 177*(3), 1053–1064.

Lens, S. M. and R. H. Medema (2019). Cytokinesis defects and cancer. *Nature Reviews Cancer 19*(1), 32–45.

Leonard, B., S. N. Hart, M. B. Burns, M. A. Carpenter, N. A. Temiz, A. Rathore, R. I. Vogel, J. B. Nikas, E. K. Law, W. L. Brown, et al. (2013). APOBEC3B upregulation and genomic mutation patterns in serous ovarian carcinoma. *Cancer Research 73*(24), 7222–7231.

Li, H. and R. Durbin (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics 25*(14), 1754–1760.

Li, W., M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L. Sullivan, K. Luzuriaga, T. C. Greenough, et al. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature 426*(6965), 450–454.

Li, Y., N. D. Roberts, J. A. Wala, O. Shapira, S. E. Schumacher, K. Kumar, E. Khurana, S. Waszak, J. O. Korbel, J. E. Haber, et al. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature 578*(7793), 112–121.

Lieber, M. R. (2016). Mechanisms of human lymphoid chromosomal translocations. *Nature Reviews Cancer 16*(6), 387.

Lightwood, R., R. Nolan, M. Franco, and A. White (1970). Epithelial giant cells in measles as an aid in diagnosis. *The Journal of Pediatrics 77*(1), 59–64.

Lindberg, H. K., G. C.-M. Falck, H. Järventaus, and H. Norppa (2008). Characterization of chromosomes and chromosomal fragments in human lymphocyte micronuclei by telomeric and centromeric FISH. *Mutagenesis 23*(5), 371–376.

Liu, Y., S. Chen, S. Wang, F. Soares, M. Fischer, F. Meng, Z. Du, C. Lin, C. Meyer, J. A. DeCaprio, et al. (2017). Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Sciences 114*(13), 3473–3478.

Liu, Y., J. Cheng, P. Siejka-Zielińska, C. Weldon, H. Roberts, M. Lopopolo, A. Magri, V. D'Arienzo, J. M. Harris, J. A. McKeating, et al. (2020). Accurate targeted long-read dna methylation and hydroxymethylation sequencing with taps. *Genome biology 21*(1), 1–9.

Lizier, M., A. Anselmo, S. Mantero, F. Ficara, M. Paulis, P. Vezzoni, F. Lucchini, and G. Pacchiana (2016). Fusion between cancer cells and macrophages occurs in a

murine model of spontaneous neu+ breast cancer without increasing its metastatic potential. *Oncotarget 7*(38), 60793.

López, S., E. L. Lim, S. Horswell, K. Haase, A. Huebner, M. Dietzen, T. P. Mourikis, T. B. Watkins, A. Rowan, S. M. Dewhurst, et al. (2020). Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nature Genetics 52*(3), 283–293.

Love, M. I., W. Huber, and S. Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology 15*(12), 1–21.

Lu, X. and Y. Kang (2009). Efficient acquisition of dual metastasis organotropism to bone and lung through stable spontaneous fusion between MDA-MB-231 variants. *Proceedings of the National Academy of Sciences 106*(23), 9385–9390.

Ludlow, M., S. McQuaid, D. Milner, R. L. de Swart, and W. P. Duprex (2015). Pathological consequences of systemic measles virus infection. *The Journal of pathology 235*(2), 253–265.

Maciejowski, J., Y. Li, N. Bosco, P. J. Campbell, and T. de Lange (2015). Chromothripsis and kataegis induced by telomere crisis. *Cell 163*(7), 1641–1654.

Magoč, T. and S. L. Salzberg (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics 27*(21), 2957–2963.

Malassiné, A., G. Pidoux, P. Gerbaud, J. L. Frendo, and D. Evain-Brion (2011). Human trophoblast in trisomy 21: a model for cell–cell fusion dynamic investigation. In *Cell Fusion in Health and Disease*, pp. 103–112. Springer.

Martinez, I., J. Wang, K. F. Hobson, R. L. Ferris, and S. A. Khan (2007). Identification of differentially expressed genes in HPV-positive and HPV-negative oropharyngeal squamous cell carcinomas. *European journal of cancer 43*(2), 415–432.

Matlin, K. S., H. Reggio, A. Helenius, and K. Simons (1982). Pathway of vesicular stomatitis virus entry leading to infection. *Journal of molecular biology 156*(3), 609–631.

Maura, F., A. Degasperi, F. Nadeu, D. Leongamornlert, H. Davies, L. Moore, R. Royo, B. Ziccheddu, X. S. Puente, H. Avet-Loiseau, et al. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications 10*(1), 1–12.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research 20*(9), 1297–1303.

Melzer, C., J. von der Ohe, and R. Hass (2018). MSC stimulate ovarian tumor growth during intercellular communication but reduce tumorigenicity after fusion with ovarian cancer cells. *Cell Communication and Signaling 16*(1), 1–9.

Melzer, C., J. von der Ohe, and R. Hass (2019). In vivo cell fusion between mesenchymal stroma/stem-like cells and breast cancer cells. *Cancers 11*(2), 185.

Merchant, T. E., I. F. Pollack, and J. S. Loeffler (2010). Brain tumors across the age spectrum: biology, therapy, and late effects. In *Seminars in radiation oncology*, Volume 20, pp. 58–66. Elsevier.

Merle, C., N. Thébault, S. LeGuellec, J. Baud, G. Pérot, T. Lesluyes, L. Delespaul, L. Lartigue, and F. Chibon (2020). Tetraploidization of immortalized myoblasts induced by cell fusion drives myogenic sarcoma development with DMD deletion. *Cancers 12*(5), 1281.

Mi, S., X. Lee, X.-p. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X.-Y. Tang, P. Edouard, S. Howes, et al. (2000). Syncytin is a captive retroviral

envelope protein involved in human placental morphogenesis. *Nature 403*(6771), 785–789.

Middlebrooks, C. D., A. R. Banday, K. Matsuda, K.-I. Udquim, O. O. Onabajo, A. Paquin, J. D. Figueroa, B. Zhu, S. Koutros, M. Kubo, et al. (2016). Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics 48*(11), 1330.

Mifune, K., M. Ohuchi, and K. Mannen (1982). Hemolysis and cell fusion by rhabdoviruses. *FEBS letters 137*(2), 293–297.

Miles, C. P. and W. Wolinska (1973). A comparative analysis of chromosomes and diagnostic cytology in effusions from 58 cancer patients. *Cancer 32*(6), 1458–1469.

Miller, F. R., D. McInerney, C. Rogers, and B. E. Miller (1988). Spontaneous fusion between metastatic mammary tumor subpopulations. *Journal of cellular biochemistry 36*(2), 129–136.

Miller, F. R., A. N. Mohamed, and D. McEachern (1989). Production of a more aggressive tumor cell variant by spontaneous fusion of two mouse tumor subpopulations. *Cancer Research 49*(15), 4316–4321.

Miroshnychenko, D., E. Baratchart, M. C. Ferrall-Fairbanks, R. Vander Velde, M. A. Laurie, M. M. Bui, A. C. Tan, P. M. Altrock, D. Basanta, and A. Marusyk (2021). Spontaneous cell fusions as a mechanism of parasexual recombination in tumour cell populations. *Nature Ecology & Evolution 5*(3), 379–391.

Mohammadi, P., S. Desfarges, I. Bartha, B. Joos, N. Zangger, M. Munoz, H. F. Günthard, N. Beerenwinkel, A. Telenti, and A. Ciuffi (2013). 24 hours in the life of hiv-1 in a t cell line. *PLoS Pathog 9*(1), e1003161.

Moore, L., D. Leongamornlert, T. H. Coorens, M. A. Sanders, P. Ellis, S. C. Dentro,

K. J. Dawson, T. Butler, R. Rahbari, T. J. Mitchell, et al. (2020). The mutational landscape of normal human endometrial epithelium. *Nature 580*(7805), 640–646.

Moore, P. S. and Y. Chang (2010). Why do viruses cause cancer? highlights of the first century of human tumour virology. *Nature reviews cancer 10*(12), 878–889.

Morgan, W. F., J. Corcoran, A. Hartmann, M. I. Kaplan, C. L. Limoli, and B. Ponnaiya (1998). DNA double-strand breaks, chromosomal rearrangements, and genomic instability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 404*(1-2), 125–128.

Mullane, S. A., L. Werner, J. Rosenberg, S. Signoretti, M. Callea, T. K. Choueiri, G. J. Freeman, and J. Bellmunt (2016). Correlation of APOBEC mRNA expression with overall survival and PD-L1 expression in urothelial carcinoma. *Scientific Reports 6*, 27702.

Musarrat, F., V. Chouljenko, A. Dahal, R. Nabi, T. Chouljenko, S. D. Jois, and K. G. Kousoulas (2020). The anti-HIV drug nelfinavir mesylate (Viracept) is a potent inhibitor of cell fusion caused by the SARSCoV-2 spike (S) glycoprotein warranting further evaluation as an antiviral against COVID-19 infections. *Journal of medical Virology 92*(10), 2087–2095.

Mussil, B., R. Suspène, M.-M. Aynaud, A. Gauvrit, J.-P. Vartanian, and S. Wain-Hobson (2013). Human APOBEC3A isoforms translocate to the nucleus and induce DNA double strand breaks leading to cell stress and death. *PloS One 8*(8), e73641.

Myerson, D. and R. K. Parkin (2019). Donor-derived hepatocytes in human hematopoietic cell transplant recipients: evidence of fusion. *Virchows Archiv 474*(3), 365–374.

Nagler, C., C. Hardt, K. S. Zänker, and T. Dittmar (2011). Co-cultivation of murine

BMDCs with 67NR mouse mammary carcinoma cells give rise to highly drug resistant cells. *Cancer cell international 11*(1), 1–13.

NCI. National Cancer Institute Dictionaries. In *NCI Dictionary of Cancer Terms*.

Neilson, K. A. and E. J. Yunis (1990). Demonstration of respiratory syncytial virus in an autopsy series. *Pediatric pathology 10*(4), 491–502.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Nik-Zainal, S., L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell 149*(5), 979–993.

Nik-Zainal, S. and S. Morganella (2017). Mutational signatures in breast cancer: the problem at the DNA level.

Nik-Zainal, S., P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, et al. (2012). The life history of 21 breast cancers. *Cell 149*(5), 994–1007.

Nikkilä, J., R. Kumar, J. Campbell, I. Brandsma, H. N. Pemberton, F. Wallberg, K. Nagy, I. Scheer, B. G. Vertessy, A. A. Serebrenik, et al. (2017). Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. *British journal of cancer 117*(1), 113–123.

Ogino, M., K. Yoshimatsu, H. Ebihara, K. Araki, B.-H. Lee, M. Okumura, and J. Arikawa (2004). Cell fusion activities of Hantaan virus envelope glycoproteins. *Journal of Virology 78*(19), 10776–10782.

Ogle, B. M., M. Cascalho, and J. L. Platt (2005). Biological implications of cell fusion. *Nature reviews Molecular cell biology 6*(7), 567–575.

Okonechnikov, K., A. Conesa, and F. García-Alcalde (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics 32*(2), 292–294.

Oliver, S. L., M. Zhou, and A. M. Arvin (2020). Varicella-zoster virus: molecular controls of cell fusion-dependent pathogenesis. *Biochemical Society Transactions 48*(6), 2415–2435.

Pantelias, A., I. Karachristou, A. G. Georgakilas, and G. I. Terzoudi (2019). Interphase cytogenetic analysis of micronucleated and multinucleated cells supports the premature chromosome condensation hypothesis as the mechanistic origin of chromothripsis. *Cancers 11*(8), 1123.

Paredes, A. M., D. Ferreira, M. Horton, A. Saad, H. Tsuruta, R. Johnston, W. Klimstra, K. Ryman, R. Hernandez, W. Chiu, et al. (2004). Conformational changes in Sindbis virions resulting from exposure to low pH and interactions with cells suggest that cell penetration may occur at the cell surface in the absence of membrane fusion. *Virology 324*(2), 373–386.

Pasqualucci, L., R. Guglielmino, J. Houldsworth, J. Mohr, S. Aoufouchi, R. Polakiewicz, R. Chaganti, and R. Dalla-Favera (2004). Expression of the AID protein in normal and neoplastic B cells. *Blood 104*(10), 3318–3325.

Patel, S. A., P. Rodrigues, L. Wesolowski, and S. Vanharanta (2020). Genomic control of metastasis. *British journal of cancer*, 1–10.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods 14*(4), 417–419.

Paula-Barbosa, M. and C. Cruz (1981). Nerve cell fusion in a case of subacute sclerosing panencephalitis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 9*(4), 400–403.

Pawelek, J. M. and A. K. Chakraborty (2008). Fusion of tumour cells with bone marrow-derived cells: a unifying explanation for metastasis. *Nature Reviews Cancer 8*(5), 377–386.

Pawelek, J. M., Z. Grasmick, E. Duvall, and G. S. LaBerge (2020). Leucocyte-tumor cell hybridization can initiate cancer metastasis. *Journal of Cancer Immunology 2*(1).

Peplow, M. (2016). The 100 000 genomes project. *BMJ 353*.

Periyasamy, M., A. K. Singh, C. Gemma, C. Kranjec, R. Farzan, D. A. Leach, N. Navaratnam, H. L. Pálinkás, B. G. Vértessy, T. R. Fenton, et al. (2017). p53 controls expression of the DNA deaminase APOBEC3B to limit its potential mutagenic activity in cancer cells. *Nucleic acids research 45*(19), 11056–11069.

Perry, J. A., A. Kiezun, P. Tonzi, E. M. Van Allen, S. L. Carter, S. C. Baca, G. S. Cowley, A. S. Bhatt, E. Rheinbay, C. S. Pedamallu, et al. (2014). Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma. *Proceedings of the National Academy of Sciences 111*(51), E5564–E5573.

Petit, C. M., J. M. Melancon, V. N. Chouljenko, R. Colgrove, M. Farzan, D. M. Knipe, and K. Kousoulas (2005). Genetic analysis of the SARS-coronavirus spike glycoprotein functional domains involved in cell-surface expression and cell-to-cell fusion. *Virology 341*(2), 215–230.

Petković, I., M. Dominis, M. Nakić, and M. Cepulić (1994). Embryonal rhabdomyosarocoma with 92, XXYY,+ dmin and evidence of spontaneous cell fusion. *Cancer genetics and cytogenetics 76*(2), 125–128.

Petković, I., M. Nakić, A. Tiefenbach, J. Konja, M. Kaštelan, L. Rajić, and R. Feminić-Kes (1988). Premature chromosome condensation in children with acute lym-

phocytic leukemia (L1) and malignant histiocytosis. *Cancer genetics and cytogenetics 35*(1), 37–40.

Petljak, M., L. B. Alexandrov, J. S. Brammeld, S. Price, D. C. Wedge, S. Grossmann, K. J. Dawson, Y. S. Ju, F. Iorio, J. M. Tubio, et al. (2019). Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell 176*(6), 1282–1294.

Petljak, M., K. Chu, A. Dananberg, E. N. Bergstrom, P. von Morgen, L. B. Alexandrov, M. R. Stratton, and J. Maciejowski (2021). The APOBEC3A deaminase drives episodic mutagenesis in cancer cells. *bioRxiv*.

Plassmeyer, M. L., S. S. Soldan, K. M. Stachelek, J. Martín-García, and F. González-Scarano (2005). California serogroup Gc (G1) glycoprotein is the principal determinant of pH-dependent cell fusion and entry. *Virology 338*(1), 121–132.

Plummer, M., C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. *The Lancet Global Health 4*(9), e609–e616.

Priestley, P., J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature 575*(7781), 210–216.

Quinlan, A. R. and I. M. Hall (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics 26*(6), 841–842.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rachkovsky, M. and J. Pawelek (1999). Acquired melanocyte stimulating hormone-inducible chemotaxis following macrophage fusion with cloudman s91 melanoma

cells. *Cell Growth and Differentiation-Publication American Association for Cancer Research 10*(7), 517–524.

Ramakrishnan, M., S. R. Mathur, and A. Mukhopadhyay (2013). Fusion-derived epithelial cancer cells express hematopoietic markers and contribute to stem cell and migratory phenotype in ovarian carcinoma. *Cancer Research 73*(17), 5360–5370.

Rao, P. N. and R. T. Johnson (1970). Mammalian cell fusion: studies on the regulation of DNA synthesis and mitosis. *Nature 225*(5228), 159–164.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics 28*(18), i333–i339.

Refsland, E. W., M. D. Stenglein, K. Shindo, J. S. Albin, W. L. Brown, and R. S. Harris (2010). Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic acids research 38*(13), 4274–4284.

Reichmann, A. and B. Levin (1981). Premature chromosome condensation in human large bowel cancer. *Cancer genetics and cytogenetics 3*(3), 221–225.

Reuschel, E. L., J. Wang, D. K. Shivers, K. Muthumani, D. B. Weiner, Z. Ma, and T. H. Finkel (2015). REDD1 is essential for optimal T cell proliferation and survival. *PloS One 10*(8), e0136323.

Roberts, S. A., M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena, et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics 45*(9), 970.

Roberts, S. A., J. Sterling, C. Thompson, S. Harris, D. Mav, R. Shah, L. J. Klimczak, G. V. Kryukov, E. Malc, P. A. Mieczkowski, et al. (2012). Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Molecular Cell 46*(4), 424–435.

Ruepp, A., B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H.-W. Mewes (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids research 38*(suppl_1), D497–D501.

Sainati, L., S. Bolcato, A. Montaldi, P. Celli, M. Stella, A. Leszl, L. Silvestro, G. Perilongo, L. C. di Montezemolo, and G. Basso (1996). Cytogenetics of pediatric central nervous system tumors. *Cancer genetics and cytogenetics 91*(1), 13–27.

Sapir, A., O. Avinoam, B. Podbilewicz, and L. V. Chernomordik (2008). Viral and developmental cell fusion mechanisms: conservation and divergence. *Developmental cell 14*(1), 11–21.

Scarpa, A., D. K. Chang, K. Nones, V. Corbo, A.-M. Patch, P. Bailey, R. T. Lawlor, A. L. Johns, D. K. Miller, A. Mafficini, et al. (2017). Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature 543*(7643), 65–71.

Scheffner, M., B. A. Werness, J. M. Huibregtse, A. J. Levine, and P. M. Howley (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell 63*(6), 1129–1136.

Schulze, K., S. Imbeaud, E. Letouzé, L. B. Alexandrov, J. Calderaro, S. Rebouissou, G. Couchy, C. Meiller, J. Shinde, F. Soysouvanh, et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nature Genetics 47*(5), 505–511.

Schuster-Böckler, B. and B. Lehner (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature 488*(7412), 504–507.

Searles, S. C., E. K. Santosa, and J. D. Bui (2018). Cell-cell fusion as a mechanism of DNA exchange in cancer. *Oncotarget 9*(5), 6156.

Senkevich, T. G. and B. Moss (2005). Vaccinia virus H2 protein is an essential component of a complex involved in virus entry and cell-cell fusion. *Journal of Virology 79*(8), 4744–4754.

Shabo, I., K. Midtbö, H. Andersson, E. Åkerlund, H. Olsson, P. Wegman, C. Gunnarsson, and A. Lindström (2015). Macrophage traits in cancer cells are induced by macrophage-cancer cell fusion and cannot be explained by cellular interaction. *BMC cancer 15*(1), 1–11.

Sheehy, A. M., N. C. Gaddis, J. D. Choi, and M. H. Malim (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature 418*(6898), 646.

Shigehara, K., T. Sasagawa, S. Kawaguchi, K. Nakashima, T. Nakashima, M. Shimamura, K. Furubayashi, and M. Namiki (2014). Cytological evaluation using liquid-based cytology in the male urogenital tract infected with human papillomavirus. *Diagnostic cytopathology 42*(6), 491–497.

Shimizu, A., H. Fujimori, Y. Minakawa, Y. Matsuno, M. Hyodo, Y. Murakami, and K.-i. Yoshioka (2018). Onset of deaminase APOBEC3B induction in response to DNA double-strand breaks. *Biochemistry and biophysics reports 16*, 115–121.

Shimizu, T., H. Marusawa, Y. Endo, and T. Chiba (2012). Inflammation-mediated genomic instability: roles of activation-induced cytidine deaminase in carcinogenesis. *Cancer science 103*(7), 1201–1206.

Shinohara, M., K. Io, K. Shindo, M. Matsui, T. Sakamoto, K. Tada, M. Kobayashi, N. Kadowaki, and A. Takaori-Kondo (2012). APOBEC3B can impair genomic stability by inducing base substitutions in genomic DNA in human cells. *Scientific Reports 2*, 806.

Silk, A. D., C. E. Gast, P. S. Davies, F. D. Fakhari, G. E. Vanderbeek, M. Mori, and M. H. Wong (2013). Fusion between hematopoietic and epithelial cells in adult human intestine. *PLoS One 8*(1), e55572.

Simões-Sousa, S., S. Littler, S. L. Thompson, P. Minshall, H. Whalley, B. Bakker, K. Belkot, D. Moralli, D. Bronder, A. Tighe, et al. (2018). The p38α stress kinase suppresses aneuploidy tolerance by inhibiting Hif-1α. *Cell Reports 25*(3), 749–760.

Smith, T., A. Heger, and I. Sudbery (2017). UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Research 27*(3), 491–499.

Southern, S. A., F. Noya, C. Meyers, T. R. Broker, L. T. Chow, and C. S. Herrington (2001). Tetrasomy is induced by human papillomavirus type 18 E7 gene expression in keratinocyte raft cultures. *Cancer Research 61*(12), 4858–4863.

Sperling, K. and P. N. Rao (1974). The phenomenon of premature chromosome condensation: its relevance to basic and applied research. *Humangenetik 23*(4), 235–258.

Sreekantaiah, C., M. K. Bhargava, and N. J. Shetty (1987). Premature chromosome condensation in human cervical carcinoma. *Cancer genetics and cytogenetics 24*(2), 263–269.

Stenglein, M. D., M. B. Burns, M. Li, J. Lengyel, and R. S. Harris (2010). APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nature structural & molecular biology 17*(2), 222.

Stewart, J. A., T. C. Holland, and A. S. Bhagwat (2019). Human herpes simplex virus-1 depletes APOBEC3A from nuclei. *Virology 537*, 104–109.

Stingele, S., G. Stoehr, K. Peplowska, J. Cox, M. Mann, and Z. Storchova (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Molecular systems biology 8*(1), 608.

Su, Y., A. Subedee, N. Bloushtain-Qimron, V. Savova, M. Krzystanek, L. Li, A. Marusyk, D. P. Tabassum, A. Zak, M. J. Flacker, et al. (2015). Somatic cell fusions reveal extensive heterogeneity in basal-like breast cancer. *Cell Reports 11*(10), 1549–1563.

Sun, C., X. Dai, D. Zhao, H. Wang, X. Rong, Q. Huang, and Q. Lan (2019). Mesenchymal stem cells promote glioma neovascularization in vivo by fusing with cancer stem cells. *BMC cancer 19*(1), 1–12.

Swanton, C., N. McGranahan, G. J. Starrett, and R. S. Harris (2015). APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity. *Cancer discovery 5*(7), 704–712.

Takei, H., H. Fukuda, G. Pan, H. Yamazaki, T. Matsumoto, Y. Kazuma, M. Fujii, S. Nakayama, I. S. Kobayashi, K. Shindo, et al. (2020). Alternative splicing of APOBEC3D generates functional diversity and its role as a DNA mutator. *International Journal of Hematology*, 1–14.

Takikawa, S., K. Ishii, H. Aizaki, T. Suzuki, H. Asakura, Y. Matsuura, and T. Miyamura (2000). Cell fusion activity of hepatitis C virus envelope proteins. *Journal of Virology 74*(11), 5066–5074.

Tate, J. G., S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research 47*(D1), D941–D947.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic acids research 17*(16), 6463–6471.

Taylor, B. J., S. Nik-Zainal, Y. L. Wu, L. A. Stebbings, K. Raine, P. J. Campbell, C. Rada, M. R. Stratton, and M. S. Neuberger (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *elife 2*, e00534.

Therneau, T. (2020). Mixed effects Cox models. *R package*.

Therneau, T. M. (2015). A package for survival analysis in R; 2015. version 2.38. *R package*.

Thomas, M., D. Pim, and L. Banks (1999). The role of the E6-p53 interaction in the molecular pathogenesis of HPV. *Oncogene 18*(53), 7690–7700.

Trypsteen, W., P. Mohammadi, C. Van Hecke, P. Mestdagh, S. Lefever, Y. Saeys, P. De Bleser, J. Vandesompele, A. Ciuffi, L. Vandekerckhove, et al. (2016). Differential expression of lncRNAs during the HIV replication cycle: an underestimated layer in the HIV-host interplay. *Scientific Reports 6*(1), 1–12.

Tugizov, S. M., I. Savchenkova, I. Grabovskaya, N. Makarova, T. Eraizer, E. Revazova, and A. Kushch (1993). Changes in expression of surface and core antigens of hepatitis B virus in different mutant clones of hepatoma PLC-PRF-5 cells. *Virus research 30*(2), 189–203.

Varečková, E., V. Mucha, S. Wharton, and F. Kostolanskỳ (2003). Inhibition of fusion activity of influenza A haemagglutinin mediated by HA2-specific monoclonal antibodies. *Archives of virology 148*(3), 469–486.

Vartanian, J.-P., D. Guétard, M. Henry, and S. Wain-Hobson (2008). Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science 320*(5873), 230–233.

Wakelam, M. (1985). The fusion of myoblasts. *Biochemical Journal 228*(1), 1.

Wakeling, W. F., J. Greetham, and D. C. Bennett (1994). Efficient spontaneous fusion between some co-cultured cells, especially murine melanoma cells. *Cell biology international 18*(3), 207–210.

Walker, B. A., C. P. Wardell, A. Murison, E. M. Boyle, N. M. Dahir, P. Z. Proszek, L. Melchor, C. Pawlyn, M. F. Kaiser, D. C. Johnson, et al. (2015). APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nature Communications 6*, 6997.

Wang, L., S. Wang, and W. Li (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics 28*(16), 2184–2185.

Wang, R., M. S. Lewis, J. Lyu, H. E. Zhau, S. J. Pandol, and L. W. Chung (2019). Cancer-stromal cell fusion as revealed by fluorescence protein tracking. *The Prostate*.

Wang, Y., H. Fan, B. Zhou, Z. Ju, L. Yu, L. Guo, J. Han, and S. Lu (2012). Fusion of human umbilical cord mesenchymal stem cells with esophageal carcinoma cells inhibits the tumorigenicity of esophageal carcinoma cells. *International journal of oncology 40*(2), 370–377.

Wangsa, D., I. Quintanilla, K. Torabi, M. Vila-Casadesús, A. Ercilla, G. Klus, Z. Yuce, C. Galofré, M. Cuatrecasas, J. Jose Lozano, et al. (2018). Near-tetraploid cancer cells show chromosome instability triggered by replication stress and exhibit enhanced invasiveness. *The FASEB Journal 32*(7), 3502–3517.

Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. Huber, A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables (2020). *gplots: Various R Programming Tools for Plotting Data*. R package version 3.1.1.

Wassarman, P. M. and E. S. Litscher (2008). Mammalian fertilization is dependent on multiple membrane fusion events. In *Cell Fusion*, pp. 99–113. Springer.

Watkins, T. B., E. L. Lim, M. Petkovic, S. Elizalde, N. J. Birkbak, G. A. Wilson, D. A. Moore, E. Grönroos, A. Rowan, S. M. Dewhurst, et al. (2020). Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, 1–7.

Weiler, J. and T. Dittmar (2019). Cell fusion in human cancer: the dark matter hypothesis. *Cells 8*(2), 132.

Weimann, J. M., C. A. Charlton, T. R. Brazelton, R. C. Hackman, and H. M. Blau (2003). Contribution of transplanted bone marrow cells to Purkinje neurons in human adult brains. *Proceedings of the National Academy of Sciences 100*(4), 2088–2093.

Weimann, J. M., C. B. Johansson, A. Trejo, and H. M. Blau (2003). Stable reprogrammed heterokaryons form spontaneously in Purkinje neurons after bone marrow transplant. *Nature Cell Biology 5*(11), 959–966.

White, J., K. Matlin, and A. Helenius (1981). Cell fusion by Semliki Forest, influenza, and vesicular stomatitis viruses. *The Journal of cell biology 89*(3), 674–679.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilhelm, T., A.-M. Olziersky, D. Harry, F. De Sousa, H. Vassal, A. Eskat, and P. Meraldi (2019). Mild replication stress causes chromosome mis-segregation via premature centriole disengagement. *Nature Communications 10*(1), 1–14.

Wilke, C. O. (2021). *ggridges: Ridgeline Plots in 'ggplot2'*. R package version 0.5.3.

Willems, L. and N. Gillet (2015). APOBEC3 interference during replication of viral genomes. *Viruses 7*(6), 2999–3018.

Williams, D., C. Scott, and T. Beck (1976, 04). Premature chromosome condensation in human leukemia. *Blood 47*(4), 687–693.

Woermann, S. M., R. Cowan, S. M. Ross, and A. D. Rhim (2019). Abstract PR03: A novel deaminase independent function of APOBEC3A catalyzes widespread chromosomal instability to drive an aggressive metastatic phenotype in pancreatic cancer.

Wu, L., C. M. Borza, and L. M. Hutt-Fletcher (2005). Mutations of Epstein-Barr virus gH that are differentially able to support fusion with B cells or epithelial cells. *Journal of Virology 79*(17), 10923–10930.

Xing, F., Y. Xie, X. Shi, P. Chen, Z. Zhang, and L. Yang (2019). Towards pixel-to-pixel deep nucleus detection in microscopy images. *BMC Bioinformatics 20*(1), 1–16.

Xu, Z., L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, S. Liu, P. Zhao, H. Liu, L. Zhu, et al. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet respiratory medicine 8*(4), 420–422.

Xu, Z., H. Zan, E. J. Pone, T. Mai, and P. Casali (2012). Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nature Reviews Immunology 12*(7), 517–531.

Yao, J. C., M. Hassan, A. Phan, C. Dagohoy, C. Leary, J. E. Mares, E. K. Abdalla, J. B. Fleming, J.-N. Vauthey, A. Rashid, et al. (2008). One hundred years after "carcinoid": epidemiology of and prognostic factors for neuroendocrine tumors in 35,825 cases in the United States. *Journal of clinical oncology 26*(18), 3063–3072.

Yart, L., D. Bastida-Ruiz, et al. (2020). Linking endoplasmic reticulum stress to polyploidy in ovarian cancer cells. *bioRxiv*.

Yau, K. K., K. Wang, and A. H. Lee (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences 45*(4), 437–452.

Yilmaz, Y., R. Lazova, M. Qumsiyeh, D. Cooper, and J. Pawelek (2005). Donor y chromosome in renal carcinoma cells of a female BMT recipient: Visualization of putative BMT–tumor hybrids by fish. *Bone marrow transplantation 35*(10), 1021–1024.

Yin, J., M. F. Chen, and T. H. Finkel (2004). Differential gene expression during HIV-1 infection analyzed by suppression subtractive hybridization. *AIDS 18*(4), 587–596.

Yin, L., P. Hu, X. Shi, W. Qian, H. E. Zhau, S. J. Pandol, M. S. Lewis, L. W. Chung, and R. Wang (2020). Cancer cell's neuroendocrine feature can be acquired through cell-cell fusion during cancer-neural stem cell interaction. *Scientific Reports 10*(1), 1–13.

Ying, Q.-L., J. Nichols, E. P. Evans, and A. G. Smith (2002). Changing potency by spontaneous fusion. *Nature 416*(6880), 545–548.

York, J., V. Romanowski, M. Lu, and J. H. Nunberg (2004). The signal peptide of the Junin arenavirus envelope glycoprotein is myristoylated and forms an essential subunit of the mature G1-G2 complex. *Journal of Virology 78*(19), 10783–10792.

Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology 11*, R14.

Zack, T. I., S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics 45*(10), 1134–1140.

Zapatka, M., I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, M. Alawi, N. Desai, H. Sültmann, H. Moch, C. S. Cooper, et al. (2020). The landscape of viral associations in human cancers. *Nature Genetics 52*(3), 320–330.

Zhang, L., J. Zhang, J. Yang, D. Ying, Y. l. Lau, and W. Yang (2013). PriVar: a toolkit for prioritizing SNVs and INDELs from next-generation sequencing data. *Bioinformatics 29*(1), 124–125.

Zhang, L.-N., Y.-H. Huang, and L. Zhao (2019). Fusion of macrophages promotes breast cancer cell proliferation, migration and invasion through activating epithelial-mesenchymal transition and wnt/β-catenin signaling pathway. *Archives of biochemistry and biophysics 676*, 108137.

Zhang, L.-N., C.-F. Kong, D. Zhao, X.-L. Cong, S.-S. Wang, L. Ma, and Y.-H. Huang (2019). Fusion with mesenchymal stem cells differentially affects tumorigenic and metastatic abilities of lung cancer cells. *Journal of cellular physiology 234*(4), 3570–3582.

Zhou, X., K. Merchak, W. Lee, J. P. Grande, M. Cascalho, and J. L. Platt (2015). Cell fusion connects oncogenesis with tumor evolution. *The American journal of pathology 185*(7), 2049–2060.