Deciphering the genomic landscape and evolution in multiple myeloma

PHUC HUU HOANG

Division of Genetics and Epidemiology
Division of Molecular Pathology
The Institute of Cancer Research
SM2 5NG

Submitted for the degree of Doctor of Philosophy in accordance with the regulations of the University of London
2020

Declaration

The work presented in this thesis is entirely my own work, except where stated in the ‘Statement of independent work attributable to candidate’ on page 7, and has not been submitted for a degree or comparable award to this or any other university or institution.

Abstract

Multiple myeloma (MM) is the second commonest haematological cancer in Western Countries, with most patients dying from progressive disease after relapse. Currently, the molecular mechanisms responsible for the initiation and evolution of MM are poorly understood. The work presented in this thesis aims to characterise novel coding and non-coding drivers, gain insight into the aetiological basis, and understand the genetics of MM evolution and relapse through integrated study of multiple next-generation sequencing datasets.

Firstly, using the CoMMpass dataset (>800 patients), multiple regulatory regions were identified as candidate non-coding drivers, including *cis*-regulatory elements (CREs) of *MYC* and a *PAX5* enhancer. Coding drivers in 40 genes, including 11 novel were identified. The study revealed that MM oncogenic pathways are targeted somatically through multiple novel mechanisms including coding and non-coding mutations; exemplified by *IRF4* and *PRDM1*, along with *BCL6* and *PAX5*, genes central to plasma cell differentiation. Secondly, coding and non-coding regions were dominated by distinct mutational processes with aging, DNA repair deficiency (DRD), and APOBEC/AID activity characterising MM. Mutational signatures showed subgroup specificity – APOBEC signatures with *MAF*-translocation t(14;16) and t(14;20) MM; DRD with t(4;14) and t(11;14); and aging with hyperdiploidy. Mutational signatures beyond that associated with APOBEC were independent of established prognostic markers and had relevance to predicting high-risk MM, providing a strong rationale for integration of mutational signatures to tailor therapy. Thirdly, analysis of high-coverage WGS dataset of primary and matched relapsed tumours from Myeloma XI trial validated several recurrently mutated CREs and discovered novel CRE targets (*e.g. BIRC2* and *IGLL5*). Relapsed patients were characterised by higher mutational burden, and associated with increased APOBEC/AID activity and DRD. Notably, further acquisition of high-risk large-scaled copy number variations at relapse was also observed, specifically enriched at pre-existing unstable genomic regions. Three major clonal evolutional patterns were identified at relapse: (i) no change in clonal composition; (ii) subclonal expansion; and (iii) emergence of new clones accompanied by decline of primary clones. Finally, defective transcription-

coupled DNA repairs was observed as predominant mutational process in MM mitochondrial DNA. Relapsed MM was characterised with global positive selection of non-synonymous mutations, most notably in genes encoding the NADH dehydrogenase complex (*MT-ND2*, *MT-ND4*, and *MT-ND5*).

Together, these findings provide increased insights into the complex genetic basis underlying MM and its progression to relapse, with potential to support the development of personalised and effective treatment strategies, and predictive biomarkers of therapeutic outcome.

Acknowledgments

Most importantly, I would like to thank, and am greatly indebted to, my supervisors Professor Richard Houlston and Dr Martin Kaiser for offering me this PhD opportunity, and providing me support and guidance throughout the journey. It has been a thoroughly enjoyable experience where I have learnt and developed tremendously, both professionally and personally.

Thank you also to all members of the ‘Molecular and Population Genetics’ and ‘Myeloma Molecular Therapy’ teams past and present for the wonderful support and friendship. Specifically, I would like to thank Sara, Alex, and Dan for helping me learn bioinformatics from scratch. Your encouragement and fruitful discussions have made the daunting transition to ‘dry lab’ a smooth and enjoyable journey.

I also would like to thank The Institute of Cancer Research, Cancer Research UK, Myeloma UK, and Bloodwise for funding my PhD. To our collaborators, thank you all for recruiting patients and providing the dataset to make this PhD possible.

Massive thanks to my mother and sisters, for always supporting me throughout all these years of education. Your love and belief in me have motivated me to overcome challenges and to achieve more than I had ever imagined. Lastly, I thank R. Quinn for being there every step of the PhD journey; it would not have been possible without your incredible support through the ups and downs.

Publications

Papers published either as a direct result from or through collaborative work during this thesis:

Hoang PH, Cornish AJ, Chubb D, Jackson G, Kaiser M*,* Houlston RS (2019). Impact of mitochondrial DNA mutations in multiple myeloma. *Under consideration for publication*.

Hoang PH, Cornish AJ, Sherborne AL, Chubb D, Kimber S, Jackson G, Morgan GJ, Kinnersley B, Kaiser M*,* Houlston RS (2019). An enhanced genetic model of multiple myeloma relapsed evolutionary dynamics. *Under consideration for publication*.

Cornish AJ**,** Chubb D, Frangou A, **Hoang PH**, Kaiser M*,* Wedge DC, Houlston RS (2019). Correcting reference bias from the Illumina Isaac aligner enables analysis of cancer genomes. *bioRxiv* 836171.

Hoang PH, Cornish AJ, Dobbins SE, Kaiser M, Houlston RS (2019). Mutational processes contributing to the development of multiple myeloma. *Blood Cancer Journal* **9**(8)**:** 60.

Cornish AJ, **Hoang PH**, Dobbins SE, Law PJ, Chubb D, Orlando G*,* Houlston RS (2019). Identification of recurrent noncoding mutations in B-cell lymphoma using capture Hi-C. *Blood Advances*; **3**(1)**:** 21-32.

Hoang PH, Houlston RS (2018). Multiple mechanisms can disrupt oncogenic pathways in multiple myeloma. *Oncotarget* **9**(88)**:** 35801-35802.

Hoang PH, Dobbins SE, Cornish AJ, Chubb D, Law PJ, Kaiser M*,* Houlston RS (2018). Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. *Leukemia* **32**: 2459–2470.

Li N Johnson DC, Weinhold N, Kimber S, Dobbins SE, Mitchell JS, Kinnersley B, Sud A, Law PJ, Orlando G, Scales M, Wardell CP, Forsti A, **Hoang PH**, Went M, Holroyd A, Hariri F, Pastinen T, Meissner T, Goldschmidt H, Hemminki K, Morgan GJ, Kaiser M, Houlston RS (2017). Genetic Predisposition to Multiple Myeloma at 5q15 Is Mediated by an ELL2 Enhancer Polymorphism.h *Cell Reports* **20**(11): 2556-2564.

Statement of independent work attributable to candidate

Chapter 1

This chapter is entirely my own work.

Chapter 2

This chapter is entirely my own work.

Chapter 3

All work is my own unless detailed below. Structural variant calling was conducted under the guidance of Sara Dobbins. Subtype association analysis was carried out under the guidance of Daniel Chubb. Mutational signature analysis was performed under the support of Alex Cornish.

Chapter 4

This chapter is entirely my own work.

Chapter 5 and 6

Myeloma XI trial samples were ascertained and collected by Richard Houlston, Martin Kaiser (both Institute of Cancer Research), Gareth Morgan (University of Arkansas for Medical Sciences), Graham Jackson (University of Leeds). Samples preparation for sequencing was carried out by Amy Sherborne and Scott Kimber. Quality control and pre-processing of data were performed by Daniel Chubb.

Chapter 7

This chapter is entirely my own work.

Table of contents

List of abbreviations

%	Percent
µmol	Micromole
A	Adenine
AD	Alternate depth
AID	Activation-induced deaminase
APOBEC	Apoliprotein B mRNA editing enzyme, catalytic polypeptide
ASC	Antibody-secreting cell
ASCAT	Allele-specific copy number analysis of tumours
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
ATP	Adenosine 5’-triphosphate
BAF	B allele frequency
BAM	Binary alignment map
bp	Base pair
BQSR	Base quality score calibration
BWA	Burrows-Wheeler aligner
C	Cytosine
CCF	Cancer cell fraction
CCRD	Carfilzomib, cyclophosphamide, lenalidomide, and dexamethasone
CHi-C	Capture Hi-C
ChIP-seq	Chromatin immunoprecipitation sequencing
Chr	Chromosome
CI	Confidence interval
CLL	Chronic lymphocytic leukemia
CNV	Copy number variant
CoMMpass	The Relating Clinical Outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile Study
COSMIC	Catalogue of somatic mutations in cancer

CRAB	Calcium levels, renal impairment, anaemia, bone lesions
CRE	*Cis*-regulatory element
CRUK	Cancer Research UK
CSR	Class switch recombination
CT	Computed tomography
CTD	Cyclophospamide, thalidomide, and dexamethasone
dbGaP	Database of genotype and phenotype
dbSNP	Database of short genetic variations
DNA	Deoxyribonucleic acid
DP	Total depth
DSB	Double strand break
EGA	European genome-phenome archive
EMM	Extramedullary myeloma
ENCODE	Encyclopedia of DNA elements
FDR	False discovery rate
FISH	Fluorescence *in situ* hybridisation
FLC	Free light chain
FN	False negative
FPKM	Fragments per kilobase of exons per million reads
FWER	Family wise error rate
g	Gram
G	Guanine
GATK	The genome analysis toolkit
GC	Germinal centre
GEP	Gene expression profiling
gnomAD	The genome aggregation database
GO	Gene ontology
GPCR	G-protein-coupled receptor
H3K27ac	Histone H3 lysine-27 acetylation
H3K27me3	Histone H3 lysine-27 trimethylation

H3K4me1	Histone H3 lysine-4 monomethylation
H3K4me3	Histone H3 lysine-4 trimethylation
HD	Hyperdiploidy
hg	Human genome
HR	Hazard ratio
HSP	Heavy strand promoter
IA	Interim analysis
ICGC	International cancer genome consortium
Ig	Immunoglobulin
IGH	Immunoglobulin heavy chain locus
IMiD	Immunomodulatory
IMWG	International Myeloma Working Group
indel	Insertion/deletion
ISS	International staging system
κ	Kappa
Kb	Kilobase
L	Litre
LogR	Log transform of read depth
LOH	Loss of heterozygosity
LSP	Light strand promoter
M protein	Monoclonal protein
Mb	Megabase
mg	Milligram
MGUS	Monoclonal gammopathy of undetermined significance
MM	Multiple myeloma
mmol	Millimole
MMRF	Multiple Myeloma Research Foundation
MRI	Magnetic resonance imaging
mRNA	MicroRNA
mRNA	Messenger RNA

MSeqDR	Mitochondrial disease sequence data resource
mSMART	Mayo stratification of myeloma and risk-adapted therapy
mtDNA	Mitochondrial DNA
MZ	Marginal-zone
NCBI	The national centre for biotechnology information
NF-κB	Nuclear factor kappa b
NGS	Next-generation sequencing
NIK	Nuclear factor kappa b-inducing kinase
NMF	Nonnegative matrix factorisation
O_H	Origin of heavy strand
O_L	Origin of light strand
OS	Overall survival
PAD	Bortezomib, doxorubicin, and dexamethasone
PCL	Plasma cell leukaemia
PCR	Polymerase chain reaction
PET-CT	Positron emission tomography–computed tomography
PFS	Progression free survival
R	Purine
RCD	Lenalidomide (Revlimid), cyclophosphamide, and dexamethasone
rCRS	Revised cambridge reference sequence
Repli-Seq	Replication sequencing
REV1	DNA repair protein REV1
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
RS	Rearrangement signature
SAM	Sequence alignment map
SBS	Single base substitution
Seq-FISH	Sequencing-based flourescence *in situ* hybridisation
sFLC	Serum free light chain

SHM	Somatic hypermutation
SMM	Smouldering multiple myeloma
SNV	Single nucleotide variant
SV	Structural variant
T	Thymine
TCGA	The cancer genome atlas
TP	True positive
tRNA	Transfer RNA
TSS	Transcription start site
UCSC	The University of California Santa Cruz
UK	United Kingdom
UTR	Untranslated region
UV	Ultraviolet
VAF	Variant allele frequency
VCF	Variant call format
W	Adenine or thymine
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WNT	Wingless/integrated
Y	Pyrimidine
β	Beta
λ	Lamda

List of figures

List of tables

CHAPTER 1 Introduction

1.1 Overview of multiple myeloma

Multiple myeloma (MM) is the second most common haematological malignancy in economically developed countries[12]. The disease is caused by an abnormal clonal expansion of plasma cells in the bone marrow[13]. Plasma cells are the final stage of B-cell differentiation, producing and releasing immunoglobulin (Ig). While MM prognosis has improved over the last 40 years with the advance of immunomodulatory agents and proteasome inhibitors, the disease remains essentially incurable and 10-year survival rate is about 30%, with most patients eventually dying from relapse[14].

1.1.1 The cellular origin of multiple myeloma

B-cells originate from pluripotent stem cells in the bone marrow in humans, with immature B-cells migrating from the bone marrow to the spleen where they exist as two main types of mature naïve B-cells – follicular B-cells and marginal zone (MZ) B-cells[15, 16]. Another type of mature naïve B-cells are B1-cells, present in the peritoneal and pleural cavities of the gut lamina propria and possesses self-renewing ability[17]. All three types of B-cells can differentiate into antibody-secreting cells (ASCs; plasmablasts and plasma cells) in response to antigenic stimulation (Figure 1.1).

B1-cells develop into ASCs when challenged with antigens, often from bacterial pathogens or viruses, and form part of the innate immune system[18]. Similarly, MZ B-cells contribute to the innate immunity by differentiating into ASCs upon exposure to polymeric epitopes of bacteria or viruses. ASCs developed from B1-cells and MZ B-cells are normally short-lived.

Follicular B-cells, as the most abundant mature B-cell subset, can generate ASCs in an early response like B1-cells and MZ B-cells when they encounter foreign antigens. With T-cells help, the follicular B-cells can also form a germinal centre (GC) within secondary lymphoid organs, such as the spleen and lymph nodes[15].

In the GC, the follicular B-cells undergo a clonal expansion, followed by somatic hypermutation (SHM) and class switch recombination (CSR) events[19]. SHM involves the Ig hypervariable domains of the heavy chain locus (*IGH*) undergoing affinity maturation to produce antibodies that are highly specific and avid for the antigens[20]. Functionality of the antibodies is further expanded during CSR, where the Ig constant regions undergo gene deletions to generate different Ig isotypes (IgA, IgG, and IgD)[20]. B-cells that bear high-affinity antibodies of various isotypes can differentiate into memory B-cells or ASCs, with some plasma cells becoming long-lived antibody response. Upon antigen rechallenge, the memory B-cells can differentiate into plasma cells rapidly and form secondary GC to generate higher-affinity antibodies[21].

Plasma cells within the bone marrow can undergo abnormal clonal expansion in the process of developing asymptomatic monoclonal gammopathy of undetermined significance (MGUS), which precedes symptomatic MM with a conversion rate of 1% per annum[22] (Figure 1.2). Smouldering multiple myeloma (SMM) is intermediary of MGUS and MM, with annual risk of 10% in first five years of progressing to MM, 3% per year in the subsequent five years and 1% per year thereafter[23]. Symptomatic MM is typified by the presence of monoclonal protein (M protein) in the blood or urine produced by the clonally-expanded plasma cells as well as the associated organ dysfunction[13]. During the development of the disease, clonal plasma cells can progress into plasma cell leukaemia (PCL) or extramedullary myeloma (EMM), migrating outside the bone marrow to the peripheral blood. Progression of the malignance is characterised by an accumulation of genetic aberrations. It is generally considered that multiple acquired genetic abnormalities disturb the intrinsic biological pathways of the plasma cells central to the development of MM[2, 24].

Figure 1.1: Key steps in normal B-cell differentiation. Upon antigen stimulation, mature naïve follicular B cells undergo B cell proliferation known as clonal expansion in germinal centres. Clonal expansion is followed by somatic hypermutation, with B cells bearing the highest affinity antibodies being preferentially selected. B-cells expressing high-antigen-affinity antibodies that have survived the germinal centre reaction ultimately differentiate into long-lived memory B-cells, antibody-secreting plasmablasts or plasma cells. Short-lived antibody-secreting plasmablasts and plasma cells can also develop from mature naïve marginal-zone B-cells and B1 cells. Adapted from Shapiro-Shelef *et al.*[11]

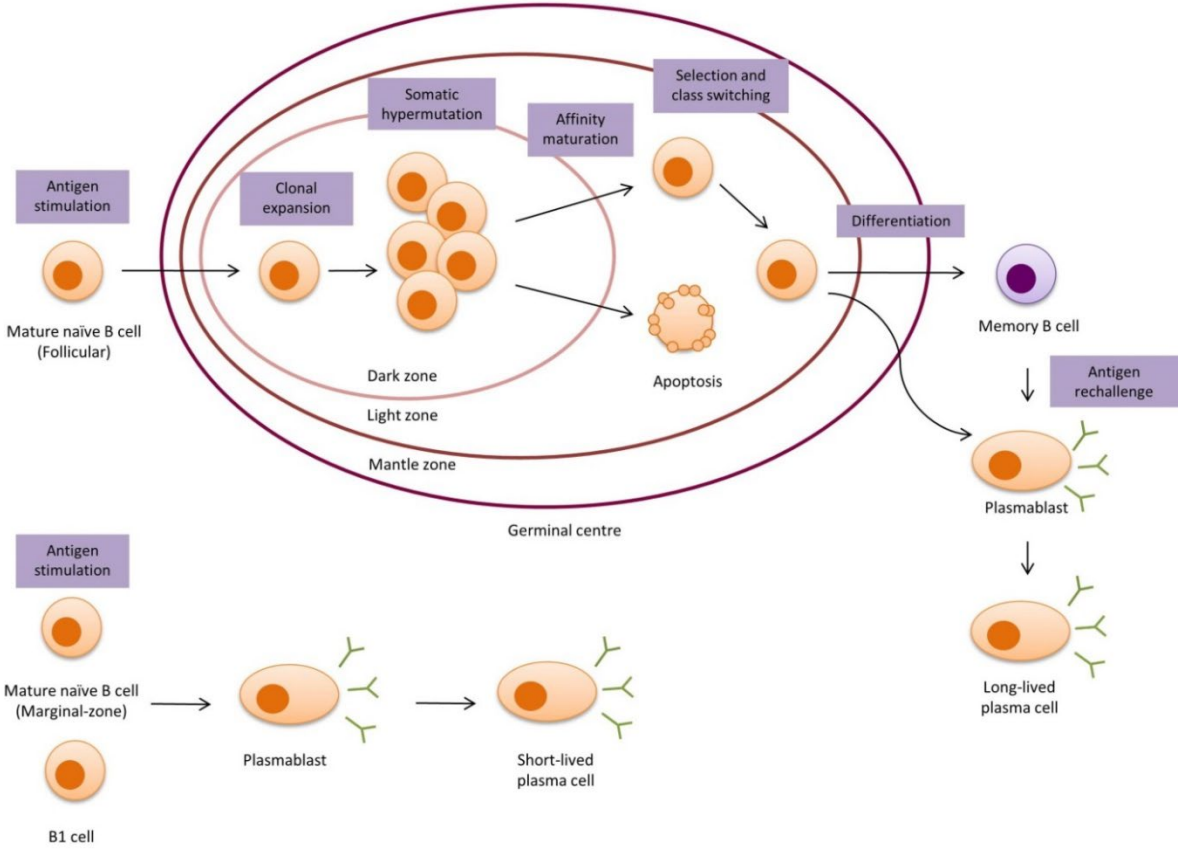
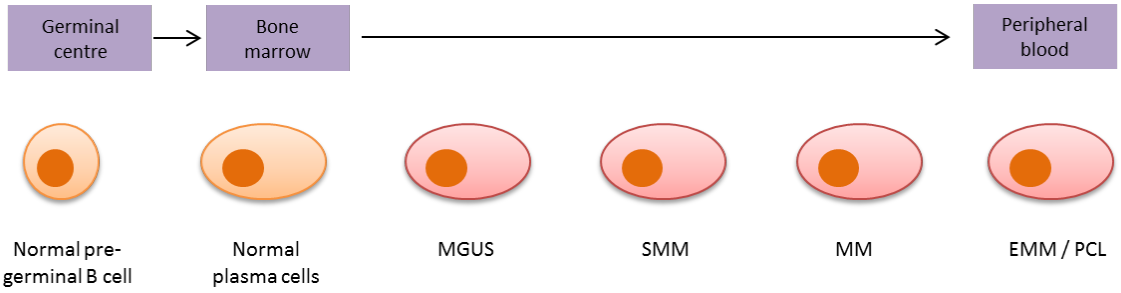


Figure 1.2: Initiation and progression in MM. MGUS, monoclonal gammopathy of undetermined significance; SMM, smouldering multiple myeloma; MM, multiple myeloma; EMM, extramedullary multiple myeloma; PCL, plasma cell leukaemia. Adapted from Morgan *et al.*[25]



1.1.2 The multiple myeloma genome

MM is characterised by the gain of genetic abnormalities from MGUS to symptomatic MM (Figure 1.3); these include hyperdiploidy (HD), chromosomal translocations, copy number changes, gene mutations, aberrant methylation, and microRNA deregulation[2, 24]. The primary genetic events can be broadly divided into HD and non-HD. HD is present in 55-60% of MM patients, involving trisomies of odd numbered chromosomes – 3, 5, 7, 9, 11, 15, 19, and 21.

Non-HD MM can be further subdivided based on translocations of the *IGH* locus at 14q32 with various recurrently observed genes[26-28] (Table 1.1). In normal B-cell differentiation, both CSR and SHM in the GC are mediated by double-strand DNA breaks (DSBs) with the expression of activation-induced deaminase (AID). Most AID-induced DSBs at the *IGH* locus are repaired locally, although DSBs can be joined to others occurring on different chromosomes, resulting in aberrant *IGH* chromosomal translocations detected in MM or MGUS plasma cells. Juxtaposition of genes next to the strongly transcriptionally active *IGH* enhancer tends to lead to their overexpression; for example, *FGFR3* (fibroblast growth factor receptor 3) and *MMSET* (myeloma SET domain protein) are overexpressed in t(4;14) MM[29]. The role of *FGFR3* in myelomagenesis remains to be established, although *FGFR3* overexpression in mice leads to tumour development, and targeting *FGFR3 in vitro* has shown to be cytotoxic in t(4;14)

MM cells[30, 31]. *MMSET* overexpression is thought to contribute to pathogenesis through epigenetic regulation and DNA repair[32, 33].

Figure 1.3: Pathogenesis of MM. The initial deregulated plasma cell in the bone marrow belongs to MGUS, which develops further genetic abnormalities in the progression to symptomatic MM, EMM/PCL. MGUS, monoclonal gammopathy of undetermined significance; SMM, smouldering multiple myeloma; MM, multiple myeloma; EMM, extramedullary multiple myeloma; PCL, plasma cell leukaemia. Adapted from Morgan *et al.*[25]

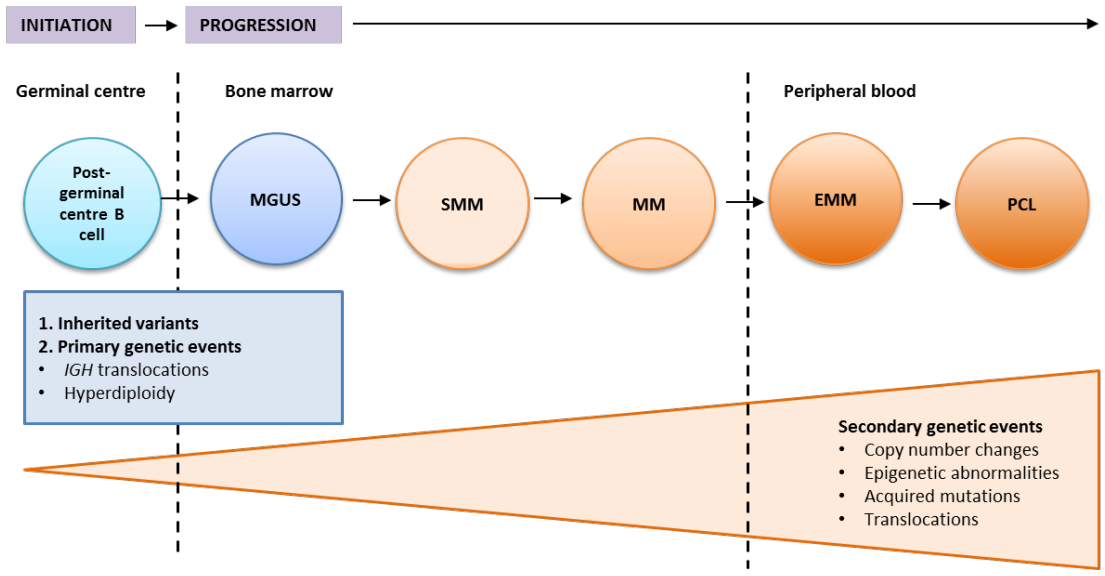


Table 1.1: The main primary chromosomal translocations in MM

Primary chromosomal	Frequency	Translocated gene partner
t(11;14)	15-20%	*CCND1*
t(4;14)	10-15%	*FGFR3, MMSET*
t(6;14)	2-5%	*CCND3*
t(14;16)	5%	*c-MAF*
t(14;20)	1-2%	*MAFB*

Secondary genetic events are implicated in the transition of MGUS to SMM and symptomatic MM, including the *MYC* aberrant expression from t(8;14), copy number changes, and mutations in RAS/MAPK signalling pathway (e.g. NRAS) (Figure 1.3)[25]. A key copy number variation is the gain of 1q21, present in > 40%

of SMM and MM cases[34]. Gain of 1q21 which implicates the oncogene *CKS1B*, shows a strong association with adverse patient prognosis[35-38].

Frequent deletions in MM are located at 1p (30%), 6q (33%), 8p (25%), 12p (15%), 13q (59%), 14q (39%), 16q (35%), 17p (7%), 20 (12%), and 22 (18%)[2, 35]. Loss of 1p is associated with poor prognosis in patients[39, 40]. Deletion of 1p12 and 1p32.3 are of particular interest, with *FAM46C*, *FAF1*, and *CDKN2C* located at these genomic regions. *FAM46C* and *FAF1* encode proteins in apoptosis regulation[41, 42], and *CDKN2C* is a key cell cycle suppressor[43, 44]. MM patients with 17p deletion, specifically 17p13, typically have an aggressive disease and poor outlook[45-47]. *TP53* is located at 17p13, a tumour suppressor gene with a role in cell cycle arrest, DNA repair, and apoptosis in response to DNA damage[48]. With the gain of genetic abnormalities and deregulation of signalling components, MM can further progress to PCL or EMM outside the bone marrow. It has been suggested that 17p deletion and the subsequent *TP53* dysfunction has a major impact on the development of PCL[49].

1.1.3 Diagnostic classification of multiple myeloma

Diagnostic classification of MM was established by the International Myeloma Working Group (IMWG) (Table 1.2)[50, 51]. Serum and urine M proteins are measured from patients by electrophoresis and immunofixation. The degree of CRAB symptoms is also evaluated in patients, which assess **C**alcium levels (hypercalcemia; serum calcium > 2.75 mmol/L), **R**enal impairment (serum creatinine > 177 μmol/L), **A**naemia (haemoglobin level < 100 g/L) and **B**one lesions (defined as ≥ 1 osteolytic lesions detected on skeletal radiography, computed tomography (CT), or positron emission tomography–computed tomography (PET-CT)).

Recently the IMWG has added the serum free light chain (sFLC) ratio to the diagnostic criteria of plasma cell disorders[52]. The normal serum free κ immunoglobulin light chain level is between 3.3-19.4 mg/L and that of free λ immunoglobulin light chain level 5.7-26.3 mg/L, with a normal κ/λ ratio of 0.26-1.65[53, 54]. Abnormal κ/λ ratio is a predictor of disease progression from MGUS, SMM to MM[55, 56], indicating that one FLC isotype is excessively produced and the

presence of clonal expansion of plasma cells. FLCs produced by clonal plasma cells are of the 'involved' FLC isotype, and a patient with involved to uninvolved ratio ≥ 100 and any of the myeloma-defining events is diagnosed with symptomatic MM (Table 1.2).

Table 1.2: International Myeloma Working Group diagnostic criteria of MM

Clinical stage	Diagnostic criteria
Monoclonal gammopathy of undetermined significance (MGUS)	• Serum M protein < 30 g/L, and • Clonal plasma cells < 10% in bone marrow, and • Absence of myeloma-related end-organ damage or tissue impairment or CRAB.
Asymptomatic / smouldering multiple myeloma (SMM)	• Serum M protein level ≥ 30 g/L or urinary M protein ≥ 500mg per 24 hours, and/or clonal plasma cells 10%-60% in bone marrow, and • Absence of myeloma defining-events (*i.e.* no myeloma-related end-organ damage or tissue impairment or CRAB, involved: uninvolved serum free light chain ratio < 100, no focal lesions identified by magnetic resonance imaging (MRI).
Symptomatic MM	• Clonal plasma cells ≥ 10% in bone marrow or biopsy-proven bony or extramedullary plasmacytoma, and any one of the following: • Clonal plasma cells in bone marrow ≥ 60%, or • Involved: uninvolved serum free light chain ratio ≥ 100 (providing involved FLC ≥ 100mg/L), or • Evidence of end-organ damage related to myeloma or CRAB, or • >1 MRI focal lesion.

1.1.4 Prognostic factors

An International Staging System (ISS) was established by the IMWG as prognostic factors for MM patient outcome, based on the serum levels of β_2-microglobulin and albumin[57]. Cytogenetic information from fluorescence *in situ* hybridisation (FISH) has also been used to risk-stratify myeloma patients[58, 59] (Table 1.3). Generally patients with *IGH* chromosomal translocations t(14;16), t(14;20), t(4;14), and 17p deletions are considered high risk. However, recent data has suggested patients with and without t(4;14) have similar survival outcomes in bortezomib-based initial therapy in conjunction with autologous stem cell transplantation and bortezomib maintenance[60].

Table 1.3: Cytogenetic risk-stratification of MM. Adapted from Bersagel *et al.*[59].

Standard risk	Intermediate risk	High risk
Hyperdiploidy t(11;14) t(6;14)	t(4;14)	t(14;16) t(14;20) 17p deletion

Unsupervised clustering of messenger RNA (mRNA) expression profiles have been used to categorise MM cells into molecular subgroups determined by their gene expression signatures[61, 62]. For example, *MAFB* and *c-MAF* overexpression from t(14;20) and t(14;16) respectively clustered as one subgroup designated 'MF', suggesting the over-expression of the *MAF* family results in deregulation of mutual downstream genes in MM. Different molecular subgroups have demonstrated differences in both event-free and overall survival[63]. Recently, mutational load has also been linked to a poorer outcome[1].

Risk-stratification of MM based on gene expression profiling (GEP), mutational load, and FISH analysis is increasingly being used to define patient treatment; for example, in the Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART)[64] and ongoing trials Total Therapy 4 and 5 conducted by the University of Arkansas[36].

1.1.5 Treatment strategies of multiple myeloma

Treatment for MM generally involves chemotherapy with or without radiotherapy[65, 66]. Patients who are younger (usually < 70 years) without co-morbidities are typically treated by high-dose therapy followed by an autologous stem cell transplantation and maintenance therapy. Older and/or less fit patients who are unsuitable for stem cell transplant, undergo chemotherapy treatment only.

Chemotherapy drugs used in the treatment of MM include the classical DNA damaging drugs such as alkylating agent melphalan and cyclophosphamide, and anthracycline agents such as doxorubicin[65]. Other drugs also include the immunomodulating agents, such as thalidomide and lenalidomide, proteasome inhibitors such as bortezomib and steroids such as dexamethasone and prednisolone. Examples of combinatorial therapies include cyclophosphamide, thalidomide, and dexamethasone (CTD) or bortezomib, doxorubicin, and dexamethasone (PAD)[67, 68], which rely on the synergistic effects of the therapy agents. Combinatorial treatments can be used as the induction treatment prior to high-dose therapy and stem cell transplantation, as an initial treatment for older and less fit patients, or at relapse.

With no curative therapies for MM and development of drug resistance in patients, relapsed MM after a period of remission is generally inevitable. A regimen of formerly administered chemotherapy drugs or novel agents with or without stem cell transplantation is given at relapse, depending on the patient's health at relapse (*e.g.* age, renal function, bone marrow function, presence of comorbidities), timing of relapse, and the efficacy and toxicity of the drugs used in prior therapy [66].

Next-generation proteasome inhibitors (*e.g.* carfilozomib, ixazomib) and immunomodulatory agents (pomalidomide) are emerging as effective therapies for relapsed MM patients[69, 70]. Other novel agents are also now in development, including monoclonal antibodies in immunotherapies (*e.g.* daratumumab, elotuzumab, indatuximab, SAR650984), repurposed alkylating agents, kinesin spindle protein inhibitors, histone deacetylase inhibitors, and inhibitors of key complexes in MM development and progression, namely cyclin-dependent

kinase, interleukin 6, Bruton's tyrosine kinase, B-cell lymphoma 2, protein kinase B, and phosphoinositide 3-kinase pathway components[71].

1.2 Somatic mutational characteristic of multiple myeloma

1.2.1 Somatic mutations in cancer

Somatic mutation is a DNA alteration occurring after conception. Somatic mutation is a universal feature of cancer, and considered to be a fundamental step in driving oncogenic growth[72]. Mutation types vary in size and complexity, from large-scale whole chromosomal gains/losses, through to complex structural changes (*e.g.* fusion genes) and single nucleotide variants (SNVs). Somatic mutation is not a process exclusive to cancer however and increasing evidence demonstrates that somatic mutation is also a common feature of normal tissue[73]. Many cancers develop as a consequence of abnormal cell proliferation due to accumulation of somatic mutations altering vital processes, including cell division and DNA damage. These mutations are known as 'driver mutations' as they provide proliferative advantage to some subpopulations of cells and drive their expansion and eventually tumourigenesis. In contrast, 'passenger mutations' provide no such fitness benefit. Given their central role, the study of driver genes has been of great interest across all tumour types. At their most impactful, the targeting of a single driver event can completely halt/control cancer growth, as exemplified by BCR-ABL fusion gene inhibition with imatinib, administration of which in chronic myeloid leukaemia contributed substantially to the dramatic increase in survival rates from around 40% to 89% (5-year)[74, 75]. It is worth noting however that the majority of subsequent efforts to inhibit targeted driver genes have been less successful, due to issues of intra-tumour heterogeneity and redundancy in the driver gene pathways, leading to targeted therapy resistance.

Somatic mutations can be classified as those increase cell survival/proliferation ('driver' or positively selected mutations), those provide no fitness advantage (neutrally selected), and those could result in cell death or senescence (negatively selected). One common approach to quantify the selection of mutations in cancer genomes is using the normalised ratio of non-synonymous to synonymous mutations or dN/dS[76-79]. The concept has been long used in

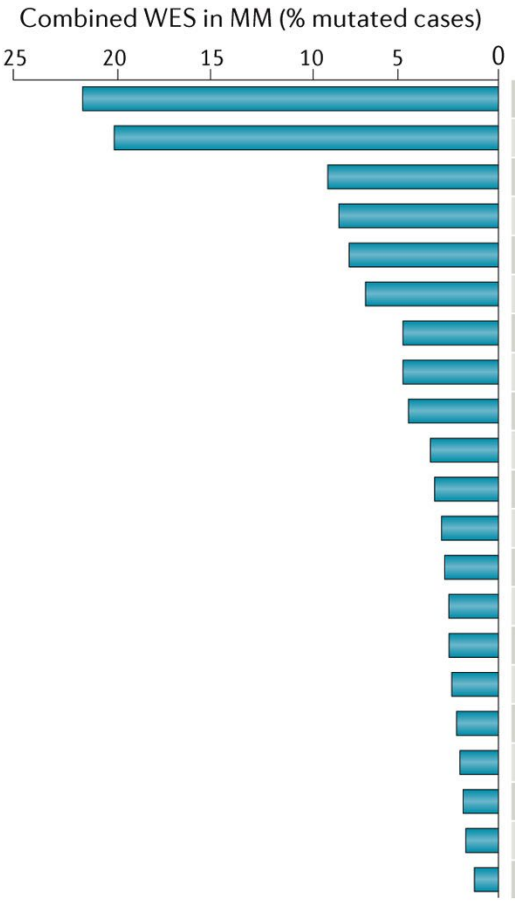
studying selection in species evolution, but a number of modifications are required to study somatic evolution[76]: (i) comprehensive models of context-dependent mutational processes, (ii) inclusion of other types of non-synonymous mutations including nonsense and splice site mutations as well as insertion/deletion (indel), (iii) stringent filtering of somatic mutations to avoid biases caused by common germline polymorphisms, (iv) taking into account of mutation rate variation across human cancer genome. Neutral mutations have dN/dS values approximately 1.0, while values of > 1.0 and < 1.0 represent positive and negative selection respectively.

The search for new driver genes intensified from around 2005 onwards, through large-scale international projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) (section 2.2.3). These colossal projects leveraged high-throughput sequencing technologies to comprehensively profile > 30 different tumour types across > 10,000 patients. The typical DNA sequencing approach for these, and other comparable studies, is whole-exome sequencing (WES) or whole-genome sequencing (WGS) of matched tumour and normal (germline) DNA. Normal germline variants can be extracted to identify true somatic changes observed only in tumour tissue. A large number of novel driver genes have been identified from these studies, and the results have been captured in large open-access databases, such as the Catalogue of Somatic Mutations in Cancer (COSMIC)[80], which currently lists > 700 genes for which mutations have been causally implicated in cancer (https://cancer.sanger.ac.uk/census, accessed 04/12/2019). However, when cross-referenced with other published cancer gene sequencing studies, a small number of about 100 driver genes were found to be recurrently and robustly established[81]. As well as identification of novel driver genes, the results from TCGA and IGCG have had a broader impact on cancer research, leading to the discovery of novel copy number variants (CNVs), non-coding driver mechanisms, clinicopathological-molecular associations, and databases of tumour specific mutation signatures[82]. The cumulative insights from somatic tumour sequencing studies have been fundamental to redefine diagnosis and prognosis as well as the development of multiple novel cancer therapies used in the clinic[82].

1.2.2 Established multiple myeloma driver genes

Existing knowledge on somatic mutations in MM was based primarily on two sources: (i) cytogenetic studies, which profiled major translocation status and copy number changes at a relatively low level of resolution (detecting whole arm deletions/gains), and (ii) targeted sequencing/WES studies, where only mutations in coding regions were assessed. The results from early cytogenetic studies are detailed in section 1.1.2. Existing compendium of driver genes were identified from three major cohorts being studied using WES[1-3, 5, 83, 84]. Several genes are recurrently mutated across these independent cohorts, thus considered as driver events in MM tumourigenesis. Among these, 16 genes were found to be mutated in significant proportion of patients in one of the three published WES studies[1-3, 5] (Figure 1.4)

Figure 1.4: Most frequent somatic mutations in patients with MM. Mutation frequencies were calculated by averaging the data from three whole-exome sequencing studies comprising a total of 733 patients[1, 3, 5]. MM, multiple myeloma; WES, whole-exome sequencing. Figure taken from Manier *et al.*[2]



	Walker *et al.* % (*n* = 463)	Lohr *et al.* % (*n* = 203)	Boli *et al.* % (*n* = 67)
KRAS	21*	23*	20*
NRAS	19*	20*	20*
FAM46C	6*	11*	10*
BRAF	7*	6*	12*
TP53	3*	8*	12*
DIS3	9*	11*	1
PRDM1	–	5*	–
SP140	–	4	6*
EGR1	4*	4	6
TRAF3	4*	5*	2
ATM	3	4	3
CCND1	2*	3	4
HISTH1E	3*	–	–
LTB	3*	1	4*
IRF4	3*	2	–
FGFR3	3*	2	–
RB1	2	3*	–
ACTG1	–	2*	–
CYLD	2*	2*	1
MAX	2*	1	–
ATR	1	1	2

*Mutations reaching significance

1.3 Mutational processes in multiple myeloma

1.3.1 Mutational signatures in cancer

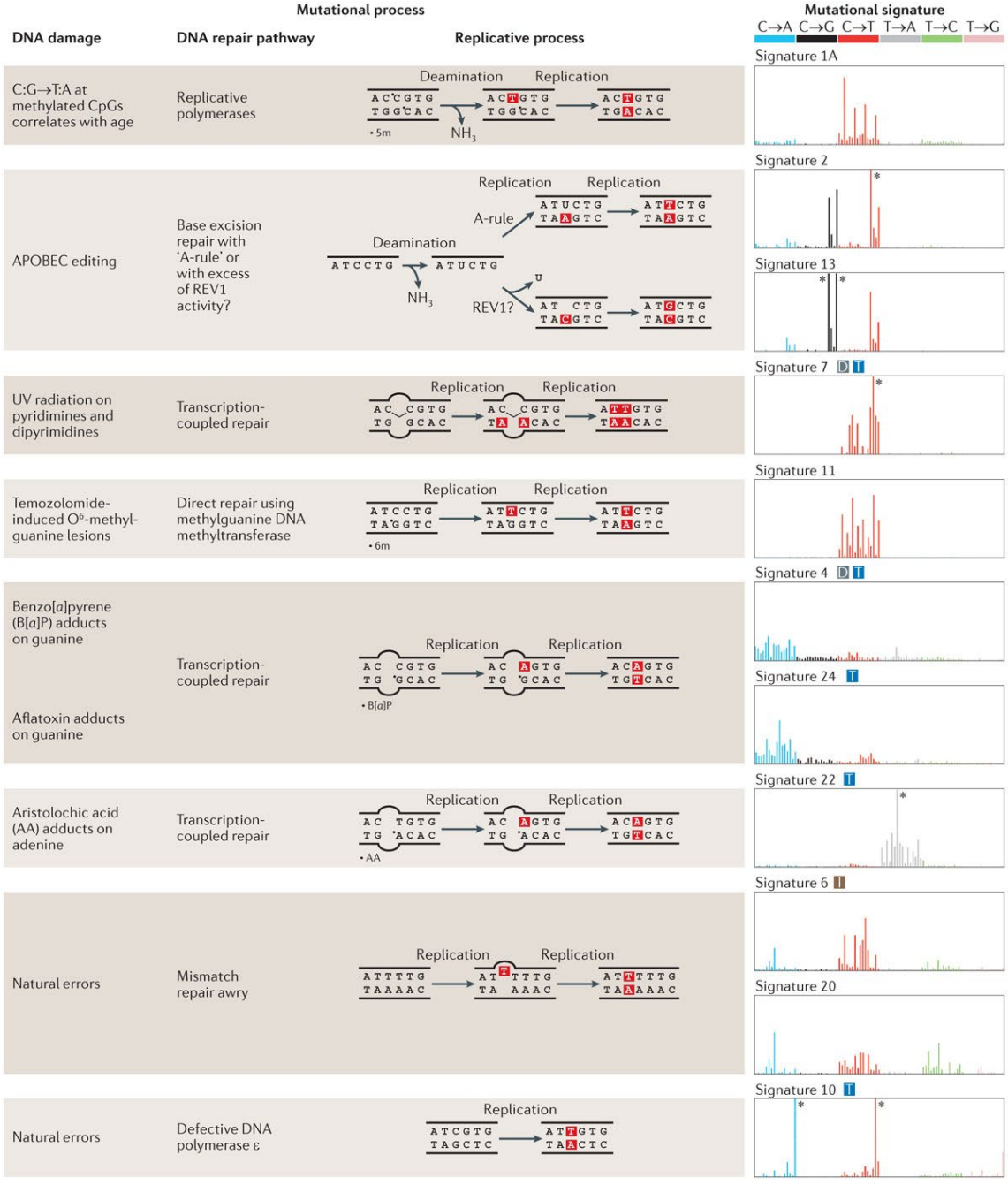
The somatic mutations we observe in a cancer are outcomes of multiple mutagenic processes that have been operative during the lifetime of a patient. Each of these processes will leave an imprint or 'mutational signature' defined by the type of base substitutions, indels, or structural variants (SVs) and therefore we could have single base substitution (SBS), small insertion and deletion, and rearrangement signatures (RS) respectively. For instance, mutations in smoking-related lung cancers are mostly G•C>A•T transversions[85], while the C•G>T•A transitions are associated with ultraviolet radiation exposure in skin cancers[86]. Recent research has shown that these mutational signatures are identifiable and quantifiable using mathematical models such as the nonnegative matrix factorisation (NMF)[87, 88]. By correlating these mutational signatures with endogenous and exogenous factors such as aging, smoking, UV radiation, DNA repair deficiency, these signatures can provide insight into the underlying mutational processes in cancer, as well as potential biomarkers or targets for treatment[10]. Mutational processes can either act continuously throughout lifetime of cancer cell (clock signatures)[89] or periodically, with some are influenced by the patient's lifestyle[90].

1.3.2 Framework to study mutational signatures

The first mutational signatures introduced were SBS, in which a signature is characterised by the type of specific base change and its direct 5' and 3' flanking bases. Given the six classes of base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) and 4 different flanking bases on the 5' end and 3' end (A, T, C, G), there are 96 distinguishable trinucleotide substitution. Since it is not possible to identify on which strand the mutation initially occurred, C>A is considered equivalent to G>T and both are counted as a C>A substitution. NMF computational framework decomposes distinguishable patterns of mutational signatures, which are characterised by different relative contribution of each trinucleotide mutation[87, 88]. Similarly, structural rearrangement signatures could also be extracted based on the NMF framework[91, 92]. SVs could be classified into subclasses by types of SVs (deletion, insertion, tandem duplication, and translocation), clustered versus non-clustered SVs, and sizes of SVs. Until recently, there are 30 different reference

SBS signatures extracted from a previous pan-cancer study[87] categorised in the COSMIC database (https://cancer.sanger.ac.uk/cosmic/signatures_v2, accessed on 4/12/19); however only some are associated with known aetiologies (Figure 1.5). In contrast, rearrangement signatures are much less well-defined.

Figure 1.5: Summary of some mutational signatures with known aetiologies, and the DNA damage and repair that constitute the mutational processes. Asterisk indicates instances where limits of the y-axes are exceeded. T, transcriptional strand bias. D, excess of dinucleotide mutations. I, association with insertions and deletions. APOBEC, apoliprotein B mRNA editing enzyme, catalytic polypeptide. REV1, DNA repair protein REV1. UV, ultraviolet. Figure taken from Helleday *et al.*[10]



1.3.3 Established mutational processes in multiple myeloma

Prior to the work described in this thesis, mutational signatures in MM were only examined in WES[84, 87], thus restricted to identification of the mutational processes primarily active in the coding regions. Therefore, there is a gap in knowledge that requires a more thorough interrogation of all mutational processes present in MM using large cohorts of WGS data. These early studies extracted two predominant mutational signatures in MM: (i) a generic signature found in many cancers enriched of C>T transitions in CpG context, and (ii) a signature enriched for C>G and C>T in TpCpA context attributed to apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) activity. The APOBEC mutational signature was seen in 3.8% of 463 patients and enriched for MAF-translocated MM t(14;16) and t(14;20)[84]. Patients with the APOBEC signatures were also associated with higher mutational burdens and poor prognosis[84]. However, a more comprehensive analysis, taking into account of all mutational signatures and established risk factors, is required to refine the roles of mutational signatures in predicting patients' prognosis.

1.4 Clonal heterogeneity and evolution

1.4.1 Overview of tumour heterogeneity and evolution

Most cancers arise through the accumulation of changes in genome and epigenome[93, 94]. A tumour cell with driver mutations are conferred with proliferative advantage over others, thus generating more daughter cells in a process called clonal expansion[95, 96]. As a consequence, tumours are composed of subpopulations of cells (subclones) that have distinguished mutations including SNVs, indels, CNVs, and SVs. Somatic mutations can be divided into (i) clonal mutations - those acquired before the complete selective sweep hence shared by all tumour cells, and (ii) subclonal mutations - those emerge after the 'most recent common ancestor' thus shared by a subpopulation of cells or subclones (Figure 1.6).

The emergence of next-generation sequencing (NGS) has revolutionised the ability to elucidate tumour heterogeneity at single nucleotide level and define

evolutionary trajectories. For the majority of available statistical methods, the first step for subclonal reconstruction is estimating genomic copy number profile and tumour purity using a number of different tools such as ASCAT[97], ABSOLUTE[98], and Sequenza[99]. This is followed by estimating the cancer cell fraction (CCF) of mutations[100]:

$$CCF = m \times \frac{VAF}{\rho}(\rho \times n_{tumour} + (1 - \rho) \times n_{normal})$$

where m is mutation multiplicity, VAF is variant allele frequency, ρ is tumour purity, n_{tumour} and n_{normal} are local copy numbers in tumour and normal genome respectively. VAF of mutation i ($V_{mut,i}$) can be calculated from read depths of variant ($r_{mut,i}$) and reference alleles ($r_{ref,i}$):

$$V_{mut,i} = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}}$$

Subsequently, to reconstruct subclonal architecture, a number of methods employ the fact that many mutations with similar CCF correspond to a cluster of clonal or subclonal mutations[100]. For such purpose, the Bayesian Dirichlet clustering process is used to cluster and infer posterior density of mutations based on their CCF (Figure 1.7). Since the algorithm does not require *a priori* number of subclones, it can both infer the number of clusters and assign mutations to each cluster identified[101].

Figure 1.6: Schematic diagram of phylogenetic tree reconstructing evolutional trajectory of a tumour. The thickness of branches indicates the proportion of tumour cells comprising that lineage. Each node of the tree represents a population of cells, with A is the founding clone or clonal population; while B, C, and D are subclones.

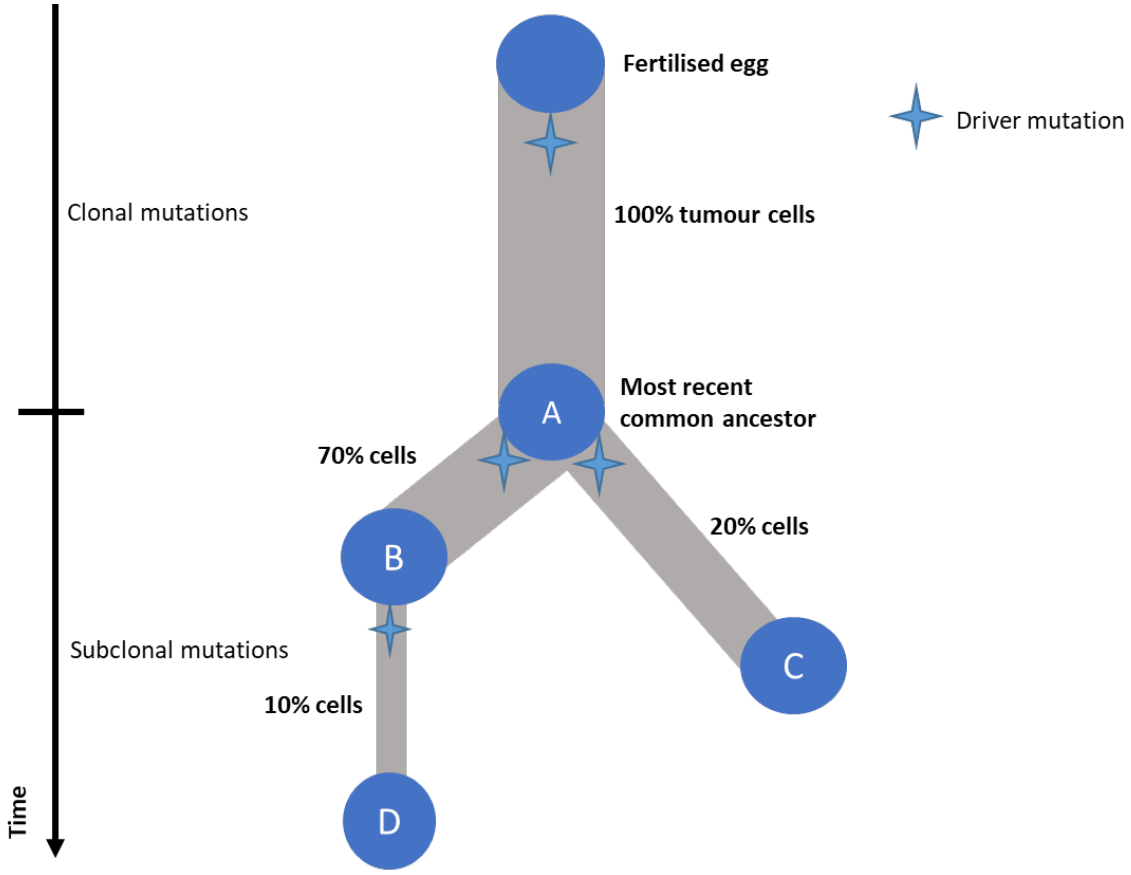
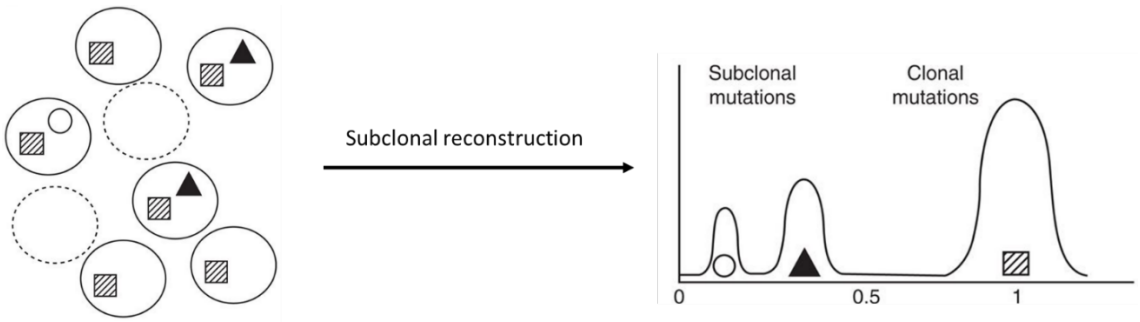


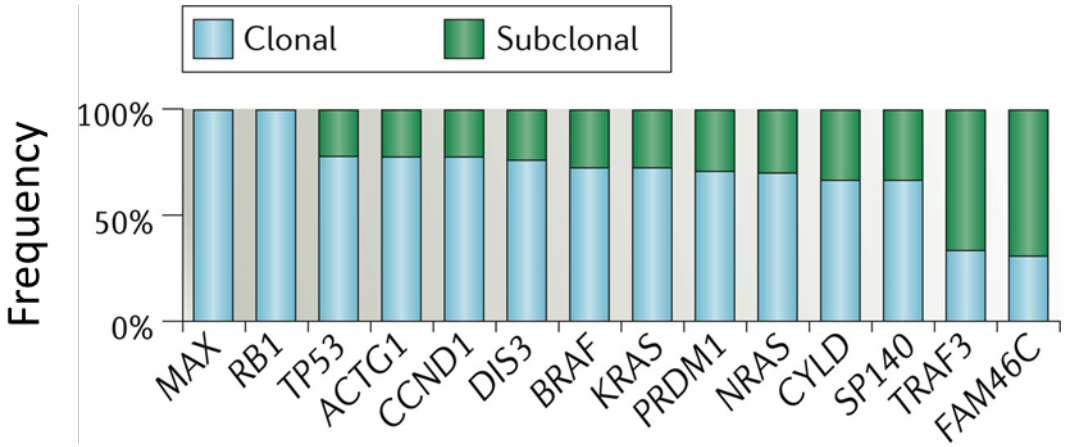
Figure 1.7: Subclonal architecture reconstruction in tumour. Tumours typically consist a mixture of tumour cells with various mutations (solid lines) and normal cells (dashed line). Some mutations are carried by all tumour cells (squares) while some are carried by a subset of tumour cells (circle and triangles). During subclonal reconstruction process, clustering algorithm can be applied to decipher the number of subclones and assign mutations to these clones, based on mutations cancer cell fractions. Adapted from Dentro *et al.*[100]



1.4.2 Tumour heterogeneity and evolution in multiple myeloma

Clonal heterogeneity and evolution in MM has previously been examined primarily using WES/targeted sequencing[5, 6, 8, 102, 103], low coverage sequencing[104], or FISH and/or array technology[102, 105]. These early studies suggest that MM tumours are highly heterogeneous, with an average of five detectable subclones per tumour[2]. In addition, some recurrently mutated genes were frequently found to be clonal (*e.g. MAX*, *RB1*, *TP53*), suggesting they are important for early event of tumour progression[2, 3, 5] (Figure 1.8).

Figure 1.8: Frequency of driver genes clonal and subclonal mutations. The results were based on 203 patients' whole-exome sequencing data. Figure adapted from Manier *et al*.[2]



MGUS and SMM have been reported to have very similar mutational profile to MM, in which all the predominant clones and initiating structural rearrangement were already present prior to MM stage[8, 106]. Two patterns of tumour progression from SMM to MM were observed: (i) the static progression model where subclonal architecture is conserved, and (ii) the spontaneous evolution model where subclonal composition is changed during the progress[106].

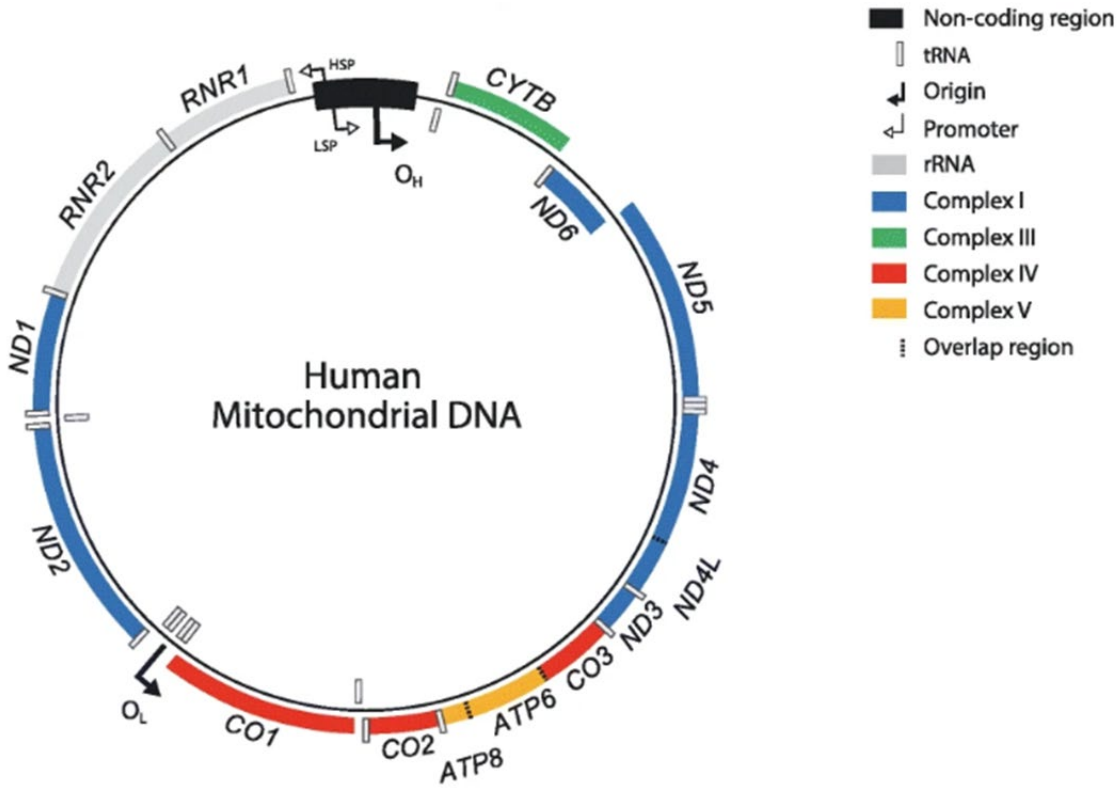
Previous studies examining clonal dynamics before (primary tumours) and after therapy (relapsed tumours) observed different patterns in limited number of WES data available[5, 107]: (i) stable tumour with no changes in subclonal heterogeneity; (ii) differential clonal response where relative proportion of subclone changes after treatment; (iii) linear evolution where new subclones emerge at relapse; and (iv) branching clonal shift where new clones emerge while other decline at

relapse. However, complexity of subclonal heterogeneity and evolutional patterns are dependent much on the depth of sequencing, as well as number of mutations and samples included. Therefore, there is a need for a larger cohort with higher coverage WGS data to accurately characterise the evolutional patterns in MM.

1.5 Mitochondrial DNA and cancer

Mitochondria are important cellular organelles with their major function being adenosine triphosphate (ATP) production. They have small (16.5 Kb) and circular genome, which are present at 100 - 10,000 copies per cell depending on cell type[108, 109]. The two strands in the human mitochondrial DNA (mtDNA) are often classified as heavy and light strand, with the heavy strand is enriched with guanine. The mtDNA encode 13 proteins, 2 ribosomal RNAs (rRNAs), and 22 transfer RNA (tRNAs)[110] (Figure 1.9). Proteins encoded by mtDNA are subunits forming respiratory chain complexes I, III, IV, and ATP synthase that are essential for energy production (Figure 1.9).

Figure 1.9: Annotated genetic composition of human mitochondrial DNA. HSP, heavy strand promoter. LSP, light strand promoter. O_H, origin of heavy strand. O_L, origin of light strand. Figure taken from Gammage *et al*.[111]



Mitochondria have long been considered important for tumour transformation and treatment response[112]. The majority of cancers have altered metabolism attributed to defective mitochondria[113] to adapt to unrestrained growth[114, 115], in part by switching from oxidative phosphorylation to glycolysis and increased uptake of glucose (*i.e.* the 'Warburg effect')[116]. In addition, mitochondria also have important roles in multiple key processes linked to tumourigenesis including regulation of apoptosis, cell cycle, cell growth, and signalling[117].

Recent evidences suggest the association of mitochondria with chemotherapies resistance and disease progression in MM[118, 119]. In addition, pre-clinical studies have further indicated promising outcomes for treatment targeting mitochondria in relapsed MM[120, 121]. Although recent studies have employed NGS to examine mtDNA mutations in various cancers[122-126], the functional implications and spectrum of mtDNA mutations in MM have not been well characterised, partly due to limited sample size and depth of WES[124]. Furthermore, any characteristics specific to MM mitochondria have been largely dismissed due to overwhelmingly dominant number of other cancer types included in previous pan-cancer studies[124]. Hence, there is an unfilled gap to comprehensively characterise and examine the impact of mutations in MM mtDNA through using a larger cohorts and high-depth sequencing data.

1.6 Study aims and scope of enquiry

The work detailed in this thesis aims to gain further insight into somatic mutational landscape in MM, making use of large NGS dataset. It is anticipated that research into the genetic basis of the plasma cell malignancy will lead to increased insight into MM biology and potentially identify novel therapeutic strategies.

Specifically:

- Chapter 3 reports on the identification of novel coding and non-coding drivers in MM, making use of The Relating Clinical Outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile Study (CoMMpass) dataset (interim analysis, IA9 release). Through integrated pathways analysis, multiple mechanisms disrupting key oncogenic pathways in MM could be identified.
- Chapter 4 reports on the analysis of CoMMpass dataset (IA10 release) to identify mutational processes contributing to development of MM, using mutational signatures analysis. Through integrating with patients' survival data, the use of mutational signatures for novel risk stratification is explored.
- Chapter 5 reports on the analysis of Myeloma XI trial, in which matched relapsed tumours are available. Given the high-coverage of the data, coding and non-coding drivers identified previously could be validated. In addition, the evolutionary trajectories at relapse are examined to shed light on the impacts of treatment on MM clonal evolution.
- Chapter 6 reports on spectrum and impacts of mtDNA mutations in MM, making use of both CoMMpass and Myeloma XI trial dataset. The pathogenic and prognostic implications of mtDNA mutations in MM are also examined.

CHAPTER 2 Material and Methods

2.1 Dataset

The analyses made use of two dataset: CoMMpass[4, 127] and Myeloma XI trial[128].

2.1.1 The Multiple Myeloma Research Foundation (MMRF) CoMMpass dataset

CoMMpass is an initiative launched in 2011 by The Multiple Myeloma Research Foundation (MMRF) (https://research.themmrf.org/). The aim of the study is to collect molecular and clinical data of 1000 patients with MM, creating the largest dataset for the disease. WGS raw fastq data of baseline newly diagnosed bone-marrow samples and their matched normal were downloaded from the database of Genotype and Phenotype (dbGaP, accession code phs000748.v4.p3). The IA9 and IA10 releases consist of WGS from 765 and 850 patients, respectively. MM tumour specimens were enriched from bone marrow aspirates by CD138 antibody conjugation yielding on average 99% CD138+ plasma tumour cell purity[129]**.** WES somatic variants, matched tumour RNA-seq (606 patients) processed by HTseq[130], CNVs, and sequencing-based fluorescence *in situ* hybridisation (Seq-FISH) data were obtained from the MMRF web portal (https://research.themmrf.org/).

Classifications of translocations in the CoMMpass dataset were based on Seq-FISH data[131]. Preliminary analysis from the MMRF CoMMass network suggests that Seq-FISH assay has similar specificity and greater sensitivity to clinical FISH[131]. Hyperdiploid was defined as amplification of 90% of the chromosome in at least two autosomal chromosomes.

2.1.2 Myeloma XI trial dataset

The Myeloma XI trial was a randomised, phase 3 design trial carried out at 110 National Health Service hospitals throughout the United Kingdom[128]. The trial featured two treatment groups – intensive (high-dose therapy and a stem cell transplant) and non-intensive groups. Bone marrow aspirates and blood samples were obtained from 80 patients with newly diagnosed MM and 25 matched

relapsed patients being treated according to the UK National Cancer Research Institute Myeloma XI trial protocol[132]. Tumour DNA was extracted from plasma cells selected and sorted using CD138 microbeads[35]. Germline DNA was derived from matched blood samples. WGS sequencing libraries were prepared using an Illumina SeqLab specific TruSeq Nano High Throughput library preparation kit (Illumina Inc, San Diego, CA 92122 USA) and sequenced using paired end on a HiSeqX instrument. Matched RNA-seq data was available for 54 of the 80 primary and 7 of the 25 relapsed tumours. RNA samples were prepared using NEB ultra II total RNA kit and sequenced paired end with the HiSeq 2500 system. Clinical data and informed consent was obtained from all patients. Ethical approval for the study was obtained by the Oxfordshire Research Ethics Committee (MREC 17/09/09, ISRCTN49407852).

Tumour *IGH*-translocation status was determined using multiplexed real-time polymerase chain reaction (PCR)[133], cross-referenced by expression of translocation target genes from RNA-seq and SVs called from WGS data (section 2.2.7.3). Hyperdiploid MM was defined as gain of any two of chromosomes five, nine and fifteen[35].

2.2 Bioinformatics analysis

2.2.1 R Software

All statistical analyses were carried out using the statistical software programme R[134] v3.5.0, unless otherwise stated. R is a publicly available software environment for statistical computing and data visualisation. Functions in R can be extended by the installation of packages, enabling wide range of statistical and bioinformatics methods to be applied[134]. All bar plots presented in this thesis generated by R have Whisker bar extend within $\pm$ 1.5 × interquartile range, unless otherwise stated.

2.2.2 Statistical significance assessment

The *P*-value, defined as the probability of obtaining a value that is at least as extreme as that of the actual sample by chance, was used to assess statistical

significance. If the *P*-value is smaller than a pre-set threshold then the null hypothesis of no association is rejected and the result is considered significant. For a single test $P < 0.05$ is deemed significant in order to control the family wise error rare (FWER; the probability of making even one type I error) at 0.05. The rate of type I error, achieving significant result by pure chance, increases when conducting multiple tests on the same dependent variable. A Bonferroni correction of the *P*-value can be applied to minimise false positives and keep FWER at 0.05. The corrected *P*-value is given by the equation $P = \alpha/n$, where α equates to the initially accepted level of significance (0.05) and n to the number of independent tests performed.

2.2.3 Databases

2.2.3.1 University of California Santa Cruz genome browser

The University of California Santa Cruz (UCSC) genome browser[135] (http://genome.ucsc.edu/) is a virtual map of the human genome, annotated with known genes, transcripts, polymorphic variation, repeated sequences, conservation, structural variation, and experimental data from external databases such as The encyclopedia of DNA elements (ENCODE, section 2.2.3.3). These features are mapped against their physical positions in the genome. Various bioinformatics tools and information are contained within the website and were utilised as follows:

- *Genome Browser* tool was used to query specific regions of DNA and visualise genes, introns, regulatory elements, and other features of the genomic location.
- *LiftOver* tool was used to convert genome coordinates between different genome assemblies (hg19 and hg38).
- *Table Browser* tool was used to download data associated with specific tracks in the genome browser. For example this tool was used to download genomic co-ordinates of simple repetitive regions for somatic variants filtering step.
- Cytoband definitions (hg19 and hg38) and cancer cell lines replication timing data (hg19) were downloaded.

2.2.3.2 National Centre for Biotechnology Information

The National centre for biotechnology information (NCBI) web server (http://www.ncbi.nlm.nih.gov/) hosts a multitude of databases and bioinformatics tools[136]. Specific tools used in this work are:

- *PubMed* for literature searches and citations.
- *RefSeq* to obtain reference sequences of chromosomes, genomic contigs, mRNAs, and proteins. These data can also be queried in UCSC. *RefGene* database, which specifies known human protein-coding and non-protein-coding genes was created from *RefSeq* using the UCSC database.
- *dbSNP* database of short genetic variations to query specific SNPs for position, allele and frequency information. Variant data from dbSNP were used to minimise false positives attributable to germline variation for somatic variants calling.
- *ClinVar* to query genetic variant pathogenicity.

2.2.3.3 The Encyclopedia of DNA Elements

The encyclopedia of DNA elements (ENCODE)[137] aims to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA level, as well as DNA regulatory elements. The ENCODE project integrates genome-wide experimental data for over 100 different cell types. The following data were used in this thesis:

- Mappability tracks which indicate how mappable a genome region in terms of short reads sequencing (75mers).
- Replication sequencing (Repli-Seq) for lymphoblast cell lines.

2.2.3.4 1000 Genomes project

The 1000 Genomes Project (http://www.1000genomes.org/) aims to provide a comprehensive catalogue of human genetic variation with frequencies > 1% through sequencing large numbers of individuals[138]. Combining data from all individuals allows for accurate imputation of variants not directly covered in this low coverage sequencing. Data from the pilot phase, phase one and phase three

of the project have been made publicly available. Variant data from 1000 Genomes project were used as part of human common single nucleotide polymorphism (SNP) reference (section 2.2.11.1).

2.2.3.5 The Genome Aggregation Database

The Genome Aggregation Database (gnomAD) (https://gnomad.broadinstitute.org/) is a resource developed by an international consortium to aggregate exome (> 100,000 exomes) and whole-genome (> 70,000 WGS) sequencing data from a wide variety of large-scale sequencing projects[139]. The database offers a comprehensive human genetic variation, including both single nucleotide and indels. It is currently the largest publicly available resource for genome-wide variant frequency data across different populations worldwide. Common SNPs and indels from gnomAD were used to remove potential false positives somatic mutations attributed to germline variants (section 2.2.12.1 and 5.2.2.1).

2.2.3.6 Ensembl genome browser

The Ensembl genome browser (http://www.ensembl.org) is a genome annotation database supported by the European bioinformatics institute[140]. Along with the ensembl biomart (http://www.ensembl.org/biomart/) it is of particular use for retrieval of gene information including genomic organisation of exons, introns and known regulatory domains, known transcripts, proteins, homologues and recorded variation within the gene sequence and also hosts the Variant Effect Predictor for annotation of variant effects[140].

2.2.3.7 Catalogue of somatic mutations in cancer

The Catalogue of somatic mutations in cancer (COSMIC) (https://www.cancer.sanger.ac.uk/cosmic) is a source of manually curated somatic mutation information in human cancers[80]. Variant data from COSMIC were used to minimise false positives for somatic variant calling (section 2.2.4.5).

2.2.3.8 BLUEPRINT

The BLUEPRINT portal (http://www.dcc.blueprint-epigenome.eu/) provides information of haemopoetic epigenomes, including RNA-seq and ChIP-seq of healthy and blood-based diseased cell lines and individuals[141]. ChIP-seq data of naïve B-cell were downloaded for the use of this thesis in chapter 3.

2.2.3.9 MITOMAP

The Mitomap[142] (https://www.mitomap.org/) is a human mitochondrial genome database, which contains information on mtDNA reference and published data on polymorphisms and mutations. Mitomaster tool as part of Mitomap can query and annotate specific mitochondrial variants. The Mitomap database was used for the analysis of mitochondria described in chapter 6.

2.2.4 Whole-genome sequencing analysis

The following programmes were used to analyse WGS data.

2.2.4.1 Description of file formats in next generation sequencing

FASTQ format

The FASTQ format is a text-based format for storing nucleotide NGS reads and their corresponding per-base quality scores[143]. Additional information relating to whether reads are single-end or paired-end is also stored. Base quality scores (Q) are Phred-based and related to the probability (p) of a base call being false by the equation: $Q = -10\ log_{10}\ p$. For example, a Q score of 10 corresponds to a 1 in 10 chance of an incorrect base call, whereas a Q score of 30 corresponds to a 1 in 1,000 chance.

Sequence alignment/map (SAM) format

The sequence alignment map (SAM) format is the most widely used file format for storing read alignments against reference sequences[144]. Details of aligned

and unaligned reads are stored along with associated mapping qualities. SAM files are typically stored in the binary form as binary alignment map (BAM) files.

Variant call format (VCF)

The variant call format (VCF) is a widely used specification for storing genetic sequence variations relative to a specified reference genome[145]. These files are typically generated by variant callers such as MuTect[146]. A variant in this format is defined as containing an allele (called the alternate allele) that is not the reference allele at that position. For a given genetic variant, the likely genotype is given along with a Phred-based genotype quality score, information about read depths for the reference and alternate alleles, genotype likelihoods as well as any additional meta-information such as variant annotation.

2.2.4.2 Sequencing quality check

All raw sequencing reads underwent quality control check with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), which performs a set of analyses to provide impressions of any problems in the sequencing data.

2.2.4.3 Sequence alignment

The Burrows-Wheeler aligner (BWA)[147] is a software package designed for mapping of single-end and paired-end sequencing of short reads against a large reference genome. The alignment of sequencing reads to human hg37 and hg38 reference genome was carried out by BWA v0.7.12.

2.2.4.4 Picard tools

Picard (http://broadinstitute.github.io/picard/) is a set of command line tools for working with NGS data in a reliable and efficient manner. In the WGS analysis pipeline, Picard v1.94 was used to filter duplicate reads arising during sample preparation (*e.g.* PCR library construction) and generate coverage metrics.

2.2.4.5 Genome Analysis Toolkit

The Genome Analysis Toolkit (GATK) is a widely used software package developed for use in analysis of high-throughput sequencing data[148, 149]. It was chosen for its ability to perform a wide range of analyses from local realignment, base score calibration, and variant calling. In the WGS sequence analysis pipeline, GATK v3.7 and v4.0 and was used to pre-process and call somatic variants in CoMMpass and Myeloma XI dataset respectively, according to GATK best practices (https://software.broadinstitute.org/gatk/best-practices/).

Base quality score recalibration

The base quality score recalibration (BQSR) package attempts to recalibrate base quality scores of sequence reads in a BAM file. The aim is for these quality scores to more truly reflect the probability of mismatching the reference genome through correcting for variation in quality with machine cycle and sequence context.

Coverage estimation

Coverage of was estimated using the DepthOfCoverage tool, restricting to genomic regions of interest (*e.g*. *cis*-regulatory regions).

MuTect

MuTect is a tool developed by the Broad institute to accurately and reliable identify somatic variants in cancer genome NGS data, and was chosen due to its low false positive rate[146]. The tool takes in matched tumour and normal tissues sequencing data, and outputs somatic mutations. Mutect v1.1.7 and v2.0 were used to call somatic variants on WGS data from CoMMpass and Myeloma XI respectively.

MuTect v1 starts by pre-processing aligned reads in tumour and normal sequencing data, omitting reads with low quality scores or with too many mismatches. Two Bayesian classifiers are then used to identify candidate somatic mutations – the first aims to detect whether the tumour is non-reference

at a given site and when this is found, the second classifier makes sure the normal does not carry the variant allele. Finally, post-processing of candidate somatic mutations is carried out to eliminate artifacts of next-generation sequencing, short read alignment, and hybrid capture. Mutect1 could only identify SNVs. Mutect v2 combines the original MuTect v1 with the assembly-based GATK HaplotypeCaller[150], enabling identification of somatic SNVs and indels[151].

2.2.4.6 Telomere length estimation

Telomerecat estimates average telomere length from WGS input, taking into account of aneuploidy as well as noise from the interstitial telomeric and sub-telomeric sequences[152].

2.2.5 Promoter capture Hi-C analysis

The HiCUP pipeline[153] v0.6.1 was used to process raw promoter capture Hi-C (CHi-C) sequencing reads, map di-tag positions against the reference human genome hg38, and remove duplicate reads. The pipeline was performed for three biological replicates of raw promoter CHi-C generated on naïve B-cells[154]. Statistically significant interactions were called using the CHiCAGO pipeline[155], with all three biological replicates processed in parallel to obtain a unique list of reproducible long-range contacts. Interactions with a –log(weighted *P*-value) $\geq$ 5 were considered significant, and only promoter-CRE interactions with linear distance $\leq$ 1Mb were considered for downstream analysis as previously advocated[156].

2.2.6 RNA-seq analysis

RNA samples were prepared using the NEB ultra II total RNA kit and sequenced paired end with the HiSeq 2500 system. Raw sequencing reads were quality checked with FastQC and trimmed for adapter with Trim Galore v.0.6.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Trimmed

reads were aligned to reference genome hg38 with HISAT2[157] v2.1.0. RNA read counts were then obtained using HTSeq[130] v.0.10.0 using default parameters.

2.2.7 General somatic genomic analysis

2.2.7.1 Somatic variant calling

The core WGS processing pipeline was followed, as described in section 2.2.4, with final BAM files generated. SNVs were then called using MuTect making use of data from dbSNP v147 and COSMIC noncoding variants v77[80] to minimize false positives attributable to germline variation. Variants were then filtered for potential DNA oxidation artefacts during sample preparation[158], and only retained if they had a minimum of one alternative read in each strand direction, a mean Phred base quality score > 26, a mean mapping quality ≥ 50, and an alignability score of 1.0 based on alignability of 75mers defined by the ENCODE/CRG GEM mappability tool[84, 159].

2.2.7.2 Significantly mutated coding genes

Two methods were used to identify significantly mutated coding genes: MutSigCV[160] v1.2 and dNdScv[76].

MutSigCV

MutSigCV analyses list of somatic mutations and identifies genes that are somatically mutated more often than would be expected by chance, given the background model[160]. The covariates incorporated in background mutation rate calculation include DNA replication time, chromatin state, and general level of transcription activity. Since the covariate file provided is only available for hg37, MutSigCV was used for CoMMpass data with default settings (Chapter 3). Prior to running MutSigCV, somatic variants were annotated with Oncotator[161]. Genes with $Q < 0.05$ were considered significantly mutated.

dNdScv

dNdScv detects cancer driver genes through a background mutation rate incorporating local (synonymous and nonsynonymous mutations in the gene) and global covariates (mutation rate variation across genes, epigenomic information), as well as sequence composition of each gene, and mutational signatures[76]. The method was used to detect somatic driver genes in Myeloma XI trial dataset (Chapter 5). The tool was also used to estimate dN/dS values (section 1.2.1) per gene and across genome (Chapter 6).

2.2.7.3 Somatic structural variants

Somatic SVs were identified on WGS data using MANTA[162] v1.2.0, LUMPY[163] v0.2.13, and/or DELLY[164] v0.7.9. All the software call translocations, inversions, deletion, and tandem duplications. These tools exhibited top performance in recent bench-marking study for SV calling[165].

2.2.7.4 Kataegis

Kataegis is a pattern of localised hypermutated regions seen in cancer genomes, mostly characterised by C>T substitution and co-localised with somatic structural rearrangements[87, 166]. Kataegis foci were identified using the KataegisPortal with default parameters (https://github.com/MeichunCai/KataegisPortal) and defined as having six or more consecutive SNVs with an average mutational distance ≤ 1 Kb, excluding immune hypermutated regions[127].

2.2.7.5 Chromoplexy

Chromoplexy is a phenomenon in which multiple genomic arrangements arising in an interdependent manner[167], disrupting multiple cancer genes co-ordinately within a single cell cycle and providing proliferative advantage to a (pre-) cancerous cell. Chromoplexy was detected using ChainFinder v1.0.1 with default parameters[167] and UCSC cytoband definitions.

2.2.7.6 Chromothripsis

Chromothripsis is a chromosome shattering phenomenon triggered by DSB, characterised by oscillation copy number states in a localised genomic region[168]. Chromothripsis was identified using ShatterSeek with default parameters[169].

2.2.8 Non-coding drivers analysis

2.2.8.1 Defining regulatory regions

Promoter regions were defined as intervals spanning 400 bp upstream and 250 bp downstream of the annotated transcription start site (TSS) from RefGene database[170] as *per* Rheinbay *et al*.[171] *Cis*-regulatory elements (CREs) were defined using publicly accessible promoter CHi-C data generated on naïve B-cells[154]. Only promoter-CRE interactions with linear distance ≤ 1 Mb[156] and only interactions with a CHiCAGO score ≥ 5 were considered statistically significant[155] (section 2.2.5).

2.2.8.2 Identification of recurrently mutated regulatory regions

Promoters and CREs were tested independently for recurrence of non-coding mutations based on the approach of Melton *et al*.[172] The statistical modelling of recurrent mutations assumes a Poisson binomial model, in which the mutation probability for each regulatory region in each tumour is determined by fitting a logistic regression model with glm R function to all data in CREs and promoters separately, taking into account the following factors at every nucleotide base[172]: tumour ID, mutational status, reference base pair (A/T versus G/C), replication timing, and coverage. Since replication timing influences mutational rate at each nucleotide base[173], replication timing at a base position was estimated as the average of replication timing data for hg37 (from Hela, K562, HEPG2, MCF7, and SKNSH cell lines)[173] and hg38 (two B-lymphocyte replicates downloaded from https://www.replicationdomain.com/). CRE regions that overlap with open reading frames (extended by 5 bp to account for splice sites), and 5' untranslated region (UTR) and 3' UTR as defined by Ensembl v73[174] were excluded from the

analysis. For promoters, mutations overlapping with open reading frames as defined by Ensembl v73[174] were excluded.

The mutation probability of each defined regulatory region is defined as:

$$\mathrm{P}(region\ is\ mutated\ ) = 1 - \prod_{k=1}^{s}(1 - p_k)$$

where s is the size of the regulatory region tested, k is the nucleotide position, p_k is the mutational probability at base k. The Poibin R package was used for approximation of Poisson binomial to estimate the empirical *P*-value for each CRE and promoter regions as *per* Melton *et al*.[172]

Mutations in each promoter and CRE region were tested for clustering based on the number of mutations occurring at the same nucleotide positions across all samples in the defined region, as recurrence of exact somatic mutations across different tumour samples implies particular SNVs have an impact on tumorigenesis. For each regulatory region containing at least three mutations[171], the mutation positions were permuted 10,000 times within the same length of the tested region under uniform distribution. The empirical clustering *P*-value for each tested region was calculated as the fraction of times that a set of permutated mutations having at least the same number of mutations occurring at the exact position as in the tested region.

The clustering *P*-value and background estimated *P*-value were combined, implementing the Fisher method within metap R package to derive combined *P*-values for recurrent mutation as *per* Rheinbay *et al*.[171] The Benjamini-Hochberg False Discovery Rate (FDR) procedure was used to adjust for multiple-hypothesis testing with significance thresholded at $Q < 0.05$.

2.2.8.3 Effect of regulatory region SNVs on gene expression

Promoter and CRE regions which were significantly mutated were examined for differential gene expression. Difference in gene expression between mutated and non-mutated tumours was tested using a negative binomial model[174], implemented in edgeR[175]. Samples with CNVs (including aneuploidy) at either the gene or the related regulatory regions were excluded[174]. Regulatory regions

were not tested if the CRE was mutated in fewer than three samples, after the removal of samples with overlapping CNVs. Where many mutated CREs were identified as interacting with a promoter, tumours harbouring mutations in more than one CRE fragment were excluded and only samples with no mutations in any of the recurrently mutated CREs were used for comparison. Regulatory regions interacting with multiple genes were tested multiple times. Only CREs interacting with protein-coding genes were evaluated. *P*-values obtained were adjusted by Benjamini-Hochberg FDR. Regions with fold change in gene expression ≥ 1.2 or ≤ 0.8, and threshold $Q < 0.1$ are reported.

2.2.8.4 Analysis of gene expression and CNVs at CREs

Focal deletions and amplifications by CNVs were defined as those with size < 3 Mb. Tumours with deleted or amplified CREs were defined as those overlapping CNVs and for each promoter-gene, CREs were excluded based on the following criteria (i) amplification or deletion of the target gene; (ii) observed <1% of total sample size. Gene expression between mutated and unmutated samples were compared using edgeR[175] using default parameters as *per* SNV analysis (Section 2.2.8.3).

2.2.9 Gene-set enrichment analysis

Gene ontology (GO) term enrichment analysis was performed to examine for the over-representation of sets of genes for specific GO annotations. To ensure that the analysis was not biased towards GO term annotations enriched amongst genes whose promoters interact with greater numbers of CREs, individual CRE-promoter interactions with the GO terms associated with the contacted genes were annotated, and the enrichment analysis was completed at the level of the CRE-promoter interaction for CREs and all TSS defined for a gene, rather than the gene level. Hence, all promoters and CRE-promoter interactions were used as the background set. Enrichment of GO term annotations obtained from GO.db[176] were tested using a hypergeometric test. The 37 GO terms spanning 10 previously defined cancer hallmarks[177] and in signalling pathways involved

MM, including NIK/NF-κB signalling, MAPK signalling, B-cell proliferation, and B-cell activation and differentiation were tested.

2.2.10 Analysis of mutational signatures

Analysis of mutational signatures were carried out using deconstructSigs[178] and Palimpsest[179].

2.2.10.1 deconstructSigs

Contribution of known mutational processes to a tumour can be determined using deconstructSigs. The method allows fitting of 96-trinucleotide SBS mutational catalogue (section 1.3.2) of tumours to a pre-selected mutational signatures framework. Assignment to the 30 COSMIC mutational signatures proposed by the Wellcome Trust Sanger Institute was performed using the R package deconstructSigs with default parameters[178].

2.2.10.2 Palimpsest

Palimpsest includes functions to extract SBS and SV signatures based on pre-defined framework (*e.g.* 30 COSMIC signatures) or *de novo* based on the NMF framework[87]. It also estimates the probability each individual mutation is due to a mutational signature, assisting identification of driver events origin.

2.2.10.3 Mutational contribution normalisation

Regional differences in trinucleotide composition were accounted for when comparing the contribution of mutational signatures between two genomic regions (regions X and Y). Such normalisation was conducted by changing the number of mutations from each mutational category in region X to that expected if the trinucleotide composition of region X was identical to the trinucleotide composition of region Y, assuming a constant rate of mutation at positions of each

trinucleotide context. The normalised number of mutations $U_{norm}^{C,X}$ of category C in region X was calculated as:

$$\boldsymbol{U_{norm}^{C,X}} = \boldsymbol{U^{C,X}} \frac{\boldsymbol{V^{C,Y}W^{X}}}{\boldsymbol{V^{C,X}W^{Y}}}$$

where $U^{C,X}$ is the number of mutations of category C observed in region X, $V^{C,X}$ is the number of positions at which a mutation of category C can occur in region X, and W^{X} is the size of region X (in base pairs). As $U_{norm}^{C,X}$ is not necessarily an integer, it is rounded to the closest integer before comparisons are completed. Mutation numbers were normalised within each tumour. Since small numbers of mutations may impact on normalisation, in each comparison the larger region was designated as region X, the smaller region designated as region Y.

2.2.11 Clonality analysis with Battenberg pipeline

Reconstruction of clonality was conducted using Battenberg v2.2.8 pipeline and DPClust[180]. There are several steps to the standard pipeline of running the workflow:

2.2.11.1 Allele-specific copy number analysis of tumours (ASCAT)

Battenberg uses ASCAT for allele specific copy number estimation[97]. Allele counting is performed with default settings by the alleleCount package v4.0.0.0 (https://github.com/cancerit/alleleCount), which outputs the reads and genotype of each position in a known SNP list. The genomic coordinates of the original SNP list (from 1000 Genomes project, section 2.2.3.4) in hg37 were converted to hg38 by UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The ASCAT algorithm uses the allele counts to generate normalised log transform of read depth (LogR) B allele frequencies (BAF) for both tumour and normal. BAF is a normalised measure of allelic intensity ratio of two alleles (A and B), with a BAF of 1 or 0 indicates complete absence of one allele (AA or BB) and a BAF of 0.5 indicates equal presence of both copies (AB). The ‘B allele’ is the non-reference allele observed in heterozygous SNP, which is also observed in most tumours. The allelic frequency of SNPs may change in tumour due to allele-

specific CNVs, loss of heterozygosity (LOH), or allelic imbalance. By comparing to matched normal BAF, such changes can provide evidence of gain or loss of germline copies in tumour. For instance, a 3-copy segment in a diploid genome would have BAF of 67% (ABB) or 33% (AAB) for pure tumour cell.

LogR is corrected for the GC content as genomic regions with extreme GC content are less amenable to hybridization, amplification and sequencing. Hence, these regions will appear to have lower coverage than regions of average GC content. LogR and BAF are then filtered and segmented using alle-specific piecewise constant fitting algorithm[97].

2.2.11.2 Calling clonal and subclonal copy number profiles

Battenberg algorithm takes output generated from ASCAT for subclonal copy number analysis[180]. Battenberg phases SNPs using IMPUTE2[181], which is implemented for hg37. To call CNVs, SNP positions were converted to hg37 before running Battenberg and the output segment positions were converted back to hg38.

2.2.11.3 Estimation of ploidy and tumour purity

As part of Battenberg pipeline, ASCAT plots the segmented logR/BAF to estimate the best solution for copy number of the whole sample (ploidy) and normal cell contamination (tumour purity)[182]. Tumour purity estimated by Battenberg was compared against and corrected using Ccube v1.0[183].

2.2.11.4 Assessing clonality

Clonality reconstruction was conducted with DPClust v2.2.8[180] using SNVs from autosomes and X chromosomes. Analysis of clonality was conducted using only SNVs in diploid regions, as miscalled copy number states can confound such analyses. Potential neutral tail mutations were identified using MOBSTER[184] and excluded prior to clustering procedure to minimise calling false positive clones. For each primary and relapse tumour pair, two-dimensional variant clustering

using a Bayesian Dirichlet process implemented in DPclust[5, 180] were performed. Only those clusters with ≥ 1% of total mutations and ≥ 100 SNVs were considered. Clonal SNVs were defined as those with a cancer cell fraction (CCF) ≥ 0.9[185].

2.2.12 Mitochondrial analysis

2.2.12.1 Mitochondrial variant calling

Mitochondrial somatic and germline variants from matched tumour-normal pairs were called using MuTect2 (v4.0.3.0)[158] according to best practices (Section 2.2.4.5), using gnomAD[139] file in hg38 provided as part of the GATK resource. Additional somatic variants called from 850 WGS tumour-normal pairs, generated as part of the MMRF CoMMpass Study (release IA10)[4, 127], were used to independently validate mutational spectrum and strand biases. Somatic variants were filtered for cross-sample contamination, oxidation artefact, alternative allele frequency ≥ 2.5%, base quality score ≥ 20, mapping quality score ≥ 20, and at least one alternative read in each strand direction[186]. Variants in known false positive regions based on revised Cambridge Reference Sequence (rCRS) (rCRS 302-315, rCRS 513 – 525, and 3105-3110)[124] were also excluded. Recurrent germline variants (present in > 10% of samples) were further removed[186] if they are not reported in MitoMap database[187]. All variants were annotated for functional and pathogenic implications using Mitochondrial Disease Sequence Data Resource (MSeqDR)[188]. Functional implication of tRNA variants was evaluated using MitoTIP[189], with likely pathogenic variants are those with MitoTIP score > 16.25 and > 75% quartile of pathogenicity score database.

2.2.12.2 Mitochondrial copy number and heteroplasmy estimation

Mitochondrial copy number were estimated using fastMitoCalc with default parameters[190], with tumour mitochondrial DNA copy number (CN_{tumour}) corrected for tumour ploidy (n_{tumour}) and tumour purity (ρ) using the following formula:

$$CN_{tumour} = \frac{mtDNA\ average\ coverage}{autosomal\ DNA\ average\ coverage}(\rho \times n_{tumour} + (1 - \rho) \times 2)$$

Tumour ploidy and purity were estimated by Battenberg[180], with purity compared and corrected using Ccube[183] (Section 2.2.11.3). When comparing variant allele frequency (VAFs) between shared primary and relapse mutations of patient i (VAF^i), VAF were normalised for purity as:

$$VAF^i_{relapse\ normalised} = \frac{r_{alt}}{(r_{alt} + r_{ref})} \times \frac{\rho^i_{primary}}{\rho^i_{relapse}}$$

where r_{alt} , r_{ref} are number of alternative reads and reference reads respectively.

2.2.12.3 Somatic mitochondrial transfer

Identification of mitochondria somatic nuclear transfer integration to nuclear genome was performed using MitoSeek[191]. To minimise false positives, only events supported by at least 5 reads[125] were considered and events with the same breakpoints present in ≥ 3 samples were excluded.

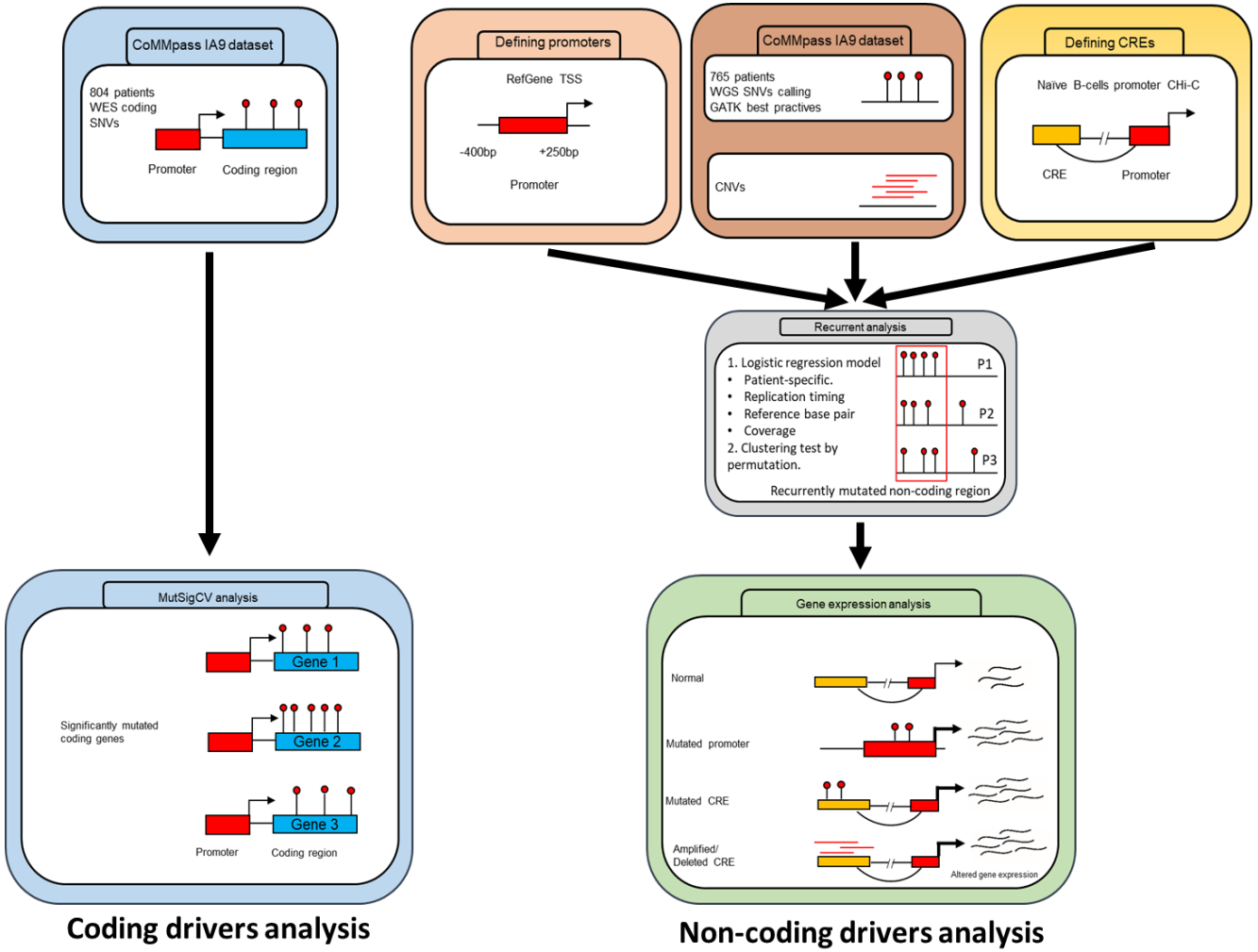
CHAPTER 3 Identification of novel coding and non-coding drivers from CoMMpass

3.1 Overview and rationale

Large-scale sequencing of MM exomes from recent studies[1-3, 5] have largely focussed on searching for driver mutations in the protein-coding components of the genome. With the increasing availability and affordability of WGS, there is an opportunity for the remaining 98% non-coding regions to be further systematically examined for driver mutations.

Mutation recurrence is an indication for positive selection in tumours, hence often used to define mutation drivers. However, the vast size of the non-coding genome presents a challenging statistical burden on robustly establishing recurrent mutations. CREs and promoters modulating gene expression represent a highly enriched subset of regulatory regions in the non-coding genome in which to search for driver mutations. Therefore, to both reduce the search space and segment the genome into functional blocks, information from promoter CHi-C in naïve B-cells[154] and TSS proximity were used in an analysis of WGS data. By linking these data to gene expression, recurrently mutated non-coding regulatory regions could be identified. Here I performed analysis on WES and WGS data of 804 and 765 MM tumour-normal pairs respectively generated by CoMMpass Study (IA9 release)[192] to search for novel coding and non-coding drivers (Figure 3.1) as well as pathways disrupted.

Figure 3.1: Overview of analysis workflow to identify coding and non-coding drivers. P, patient. TSS, transcription start site. CNV, copy number variant. SNV, single nucleotide variant.



3.2 Study design

3.2.1 Sequencing dataset

All data analysed in this chapter were generated as part of the MMRF CoMMpass Study (release IA9). WGS raw fastq data on 765 matched tumour-normal baseline newly diagnosed bone-marrow samples, WES somatic variants, matched tumour RNA-seq (606 of the 765 WGS patients) processed by HTseq, CNVs, and Seq-FISH data for karyotype classification were obtained as described in section 2.1.1. Processed promoter CHi-C data was obtained from Javierre *et al.*[154] Histone ChIP-seq sequencing data for H3K4me1, H3K27ac, H3K4me3, and H3K27me3 were downloaded from BLUEPRINT under accession number EGAD00001002466, sample S00XAQH1 (section 2.2.3.8). UCSC LiftOver tool was used to derive genome coordinates ChIP-seq coordinates in hg37. Replication timing data of five cancer cell lines Hela, K562, HEPG2, MCF7, and SKNSH cell lines were downloaded from the UCSC Genome Browser.

3.2.2 Statistical and bioinformatics analysis

Quality control, sequence alignment to hg37, and variant calling were performed using FastQC v.0.11.4/BWA v0.7.12/GATK/Mutect v1.1.7 software as described in section 2.2.4. Somatic SNVs were filtered further to minimise false positive as detailed in section 2.2.7.1.

3.2.2.1 Assessment of variant calling

Sensitivity and specificity to detect clonal mutations in the low-coverage WGS CoMMpass dataset were estimated by comparing called variants from WGS with those identified in the high-coverage WES data in IA9 dataset (alternate allele ratio > 0.2). SNVs detected from both WES and WGS were considered true positives. SNVs identified from WGS but not in WES were considered false positives. Variants detected by WES but not WGS were considered false negatives. Specificity and sensitivity were calculated for all patients with available matched WES and WGS data as follow:

$$\text{Sensitivity} = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$\text{Specificity} = \frac{True\ positives}{True\ negatives + False\ positives}$$

3.2.2.2 Significantly mutated coding genes

Somatic mutations from WES data of 805 patients were annotated using Oncotator[161] and applied MutSigCV[160] v1.2 adopting default settings (section 2.2.7.2). Genes with $Q < 0.05$ were considered significantly mutated.

3.2.2.3 Analysis of copy number variants

Deletions and amplifications were defined as abs(log_2ratio) ≥ 0.1613 based on circular binary segmentation defined copy number segments (Jonathan Keats, personal communication). A chromosome was considered amplified if at least 90% of the chromosome overlapped with an amplification. Cytoband definitions (hg19) were downloaded from UCSC. Gene exon boundaries were downloaded from RefSeq (hg19). Affected genes and cytobands were identified by overlaying CNVs using bedtools[193]. CNV plots were produced using the package karyoploteR[9].

3.2.2.4 Analysis of structural variants

BAM files were analysed and annotated using Illumina's MANTA[162] and NIRVANA[194] software with default settings, allowing identification of SVs falling within gene boundaries. To search for genes in the vicinity of breakpoints whose expression may be affected by SVs, the composite chromosome (as per SAMtools variant call format v4.1 specifications) was first assembled and then genes within 1 Mb of the breakpoints were identified using the RefGene database. The immunoglobulin loci *IGH*, *IGK* and *IGL* were defined to occur at 14q32.33, 2p11.2, and 22q11.22 respectively. SV plots were produced using Circos R package[195].

3.2.2.5 Non-coding drivers analysis

Non-coding drivers were identified as detailed in section 2.2.8.

3.2.2.6 Subgroup analysis

Subgroup analysis was restricted to the main groups for which there was reasonable power to detect a relationship. Specifically, the most frequent myeloma subtypes were included – HD, t(4:14), t(11:14) and t(14:16) - along with the t(8:14) *MYC* translocation subgroup. The enrichment by subgroup of (i) frequently mutated genes (defined by analysis in this thesis section 3.2.3 and previously published work[1, 3, 5]) and (ii) those CREs identified as recurrently mutated and differentially expressed were assessed based on Fisher's exact tests. Furthermore, to confirm the combined analysis had not missed any subgroup specific effects, coding and non-coding SNV analyses (section 3.2.3 and 3.2.6) were performed separately for each subgroup.

3.2.2.7 Gene-set enrichment analysis

Over-representation of sets of genes for specific GO annotations was performed as detailed in section 2.2.9.

3.2.2.8 Integrated pathway analysis

The Reactome tool[196] was used to evaluate pathways significantly altered by coding and non-coding drivers identified, with *Q* values < 0.05 being considered statistically significant.

3.2.2.9 Analysis of mutational signatures

All somatic variants from WES and WGS passing filtering were considered for mutational signature analysis. Assignment to the 30 mutational signatures proposed by the Wellcome Trust Sanger Institute was performed using the R package deconstructSigs with default parameters[178] (section 2.2.10.1). Non-coding variants disrupting CREs corresponding to *PAX5* were analysed. Associations between APOBEC mutations and MM translocation subgroups, as well as recurrently mutated genes and regulatory regions identified as statistically altering gene expression, were performed using Fisher's exact test. A $P < 0.05$ (one-sided) was considered statistically significant.

3.3 Results

The median age of patients at diagnosis was 64 years (range 31 – 93). The frequency of MM translocation subgroups in the CoMMpass series is similar to that reported in unselected patients[2] (Table 3.1). The median exonic mutation rate across all tumour samples was 1.95 mutations/Mb consistent with published literature[2, 87], with t(16;14) MM displaying the highest mutation rate[84] (P = 2.2 × 10^{-6}, Wilcoxon rank-sum test; Table 3.1). Whilst the low coverage WGS data (average 6-12×) was not primarily produced for mutational analysis, an estimated average sensitivity of 20% to detect clonal SNVs based on comparisons between paired WGS and WES (average 120–150×) data available for 734 samples. A global whole-genome comparison with previously published mutation rate in MM[2, 87] suggests up to 35% sensitivity. Given this limitation, the analysis is therefore expected to provide insights into mostly clonal mutation associated with early events underlying tumorigenesis[197].

3.3.1 Recurrently mutated non-coding regulatory regions

Quality control and filtering of WGS data resulted in a total of 71,573 SNVs across all tumours. Recurrently mutated regions were identified as those containing highly-clustered mutations and a greater number of mutations than that expected given the background mutation rate (section 2.2.8.2). To identify somatic mutations in the non-coding regulatory regions, I defined 28,629 regions associated with 23,635 genes as promoters[171]. Promoters associated with 34 target genes were identified as recurrently mutated (Q < 0.05). Using promoter CHi-C in naïve B-cells[154], I then defined 79,894 fragments containing putative CREs identifying 221,380 unique significant interactions with promoters. These CRE fragments (median size 2 Kb with median linear distance to respective interacting promoter of 300 Kb) constituted 15% of the genome and were enriched for ATAC-seq accessibility and regulatory histone marks[154]. Among the CRE regions, 114 recurrently mutated CRE regions interacting with the promoters of 271 genes were identified (Q < 0.05). These genes were over-represented for pathways associated with cell adhesion (P = 4.4 × 10^{-4}), inflammatory response (P = 5.6 × 10^{-4}), NIK/NF-κB signalling (P = 1.7 × 10^{-2}), regulation of B-cell

activation ($P = 3.6 \times 10^{-2}$), and B cell differentiation ($P = 4.7 \times 10^{-2}$), including *PAX5* and *BCL6* (Table 3.2)

Table 3.1: CoMMpass karyotype classification and average somatic mutations (release IA9). *, the data was taken from Manier *et al.*[2]. **, associations between the number of somatic mutations and MM karyotype were performed using a Wilcoxon rank-sum test comparing the distribution of mutations for each karyotype with all other samples.

MMRF CoMMpass karyotype classification				Somatic Mutation Counts				
Subgroups	**No. of samples**	**Percentage**	**Published literature***	**Mean WES**	**Median WES**	**Mean WGS**	**Median WGS**	**Mutation enrichment *P*-value****
t(11;14)	150	19.6%	15-20%	162	128	1322	1220	7.0E-03
t(4;14)	93	12.1%	15%	156	146	1609	1551	5.0E-03
t(14;16)	31	4.1%	5%	622	412	4200	2620	2.2E-06
t(6;14)	11	1.4%	1-2%	167	149	2002	1270	9.0E-01
t(14;20)	9	1.3%	1%	483	152	2933	2212	4.0E-02
MYC-translocation	109	14.2%	15-20%	179	156	1547	1288	5.9E-01
Hyperdiploidy	423	55.3%	50%	175	153	1461	1288	3.0E-02

Table 3.2: Significant gene-set enrichment for recurrently mutated *cis*-regulatory elements. Only significant gene ontology (GO) terms are shown ($P < 0.05$).

GO term ID	GO term name	Cancer hallmark category	Number of occurences of annotation in candidate set	Expected number of occurences of annotation in candidate set	Number of occurences of annotation in background set	*P*-value
GO:0007155	Cell adhesion	Activating invasion	25	12.270	251	4.38E-04
GO:0006954	Inflammatory response	Tumour-promoting inflammation	14	5.182	106	5.61E-04
GO:0038061	NIK/NF-kappaB signaling	Sustaining proliferative signaling	4	1.027	21	1.72E-02
GO:0050864	Regulation of B-cell activation	Sustaining proliferative signaling	4	1.271	26	3.56E-02
GO:0030183	B-cell differentiation	Sustaining proliferative signaling	3	0.831	17	4.72E-02

3.3.2 Effect of regulatory SNVs on gene expression

To identify non-coding driver mutations in regulatory regions, the expression levels of respective target genes in mutated and non-mutated tumours were compared. Tumours having copy number changes overlapping either the regulatory region or target gene were excluded from the analysis. Recurrent mutation of the *NBPF1* promoter was identified (20 tumours, $Q = 1.3 \times 10^{-15}$); these mutations were associated with increased *NBPF1* expression ($Q = 7.9 \times 10^{-4}$, 1.7-fold; Figure 3.2). *NBPF1* belongs to the neuroblastoma breakpoint family, members of which have been observed to be overexpressed in sarcomas[198] and non-small-cell lung cancer[199]. *NBPF1* is directly regulated by NF-κB[200], whose signalling pathway is recurrently affected in MM, suggesting the relevance of this novel candidate in MM development. Six recurrently mutated CREs associated with differential expression of their respective target genes were identified (*PAX5, ST6GAL1, CALCB, COBLL1, HOXB3*, and *ATP13A2*), four annotated by epigenetic marks indicative of active enhancers ($Q < 0.1$, Table 3.3, Figure 3.3). The *PAX5* CRE (71 clustered mutations across 55 tumours, 7% of all tumours) maps 3 Kb downstream of the *PAX5* chronic lymphocytic leukaemia (CLL) enhancer[201] (Figure 3.3g). The 4.6-fold reduced expression associated with CRE mutation is consistent with *PAX5* functioning as a tumour suppressor in MM, as in other B-cell malignancies[201-203]. This CRE forms part of a cluster of 12 recurrently mutated CRE fragments interacting with the *PAX5* promoter. While 28% (212/765) of tumours harboured mutations in at least one of these *PAX5* CREs, the mutations were not always associated with a significant change in *PAX5* expression. Five CREs, interacting with the *ST6GAL1* promoter, were recurrently mutated in 8% (64/765) of samples. Although the mutated CREs showed an overall consistent trend of association between mutation and upregulation of *ST6GAL1,* only one CRE was significantly associated with increased gene expression (3% of samples, $Q = 0.036$, 1.4-fold upregulation, Table 3.3, Figure 3.3). *ST6GAL1*, which primarily generates α2,6 linked sialic, is overexpressed in multiple cancers[204] and the increased expression may contribute to aberrant immunoglobulin-G glycosylation seen in MM development[205, 206].

Mutations of the *COBLL1* CRE were associated with increased gene expression (Table 3.3, Figure 3.3). *COBLL1* plays a role in NF-κB pathway activation, is

important for normal hematopoiesis[207], and is upregulated in MM[208]. Conversely mutations in the *HOXB3* CRE were associated with reduced expression (Table 3.3, Figure 3.3), consistent with *HOXB3* acting as a tumour suppressor in MM, as in acute myeloid leukemia[209].

By restricting analysis to subgroups of MM, a CRE interacting with the *TPRG1* promoter was identified as recurrently mutated, resulting in significant differential gene expression in HD and *MYC*-translocation MM (Table 3.4). Although mutated in only 2% of HD (9/423) and 3% (3/109) of *MYC*-translocation samples, these were associated with 6.3-fold and 3.6-fold upregulation in gene expression respectively (based on 4/118 and 3/34 tumours respectively; Table 3.4, Figure 3.4). Relative paucity of mutations in regulatory regions of *PAX5* in t(11:14) MM ($P = 2.7 \times 10^{-3}$, Table 3.5) was also identified. Intriguingly, since this subgroup is enriched for coding mutations in *IRF4,* it suggests complementary genomic alteration impacting on the plasma cell differentiation pathway in MM (Table 3.6).

Figure 3.2: Mutations in the promoter region affect gene expression of *NBPF1*. (n = 461 versus n = 14). **, $Q < 0.05$; mut, mutated.

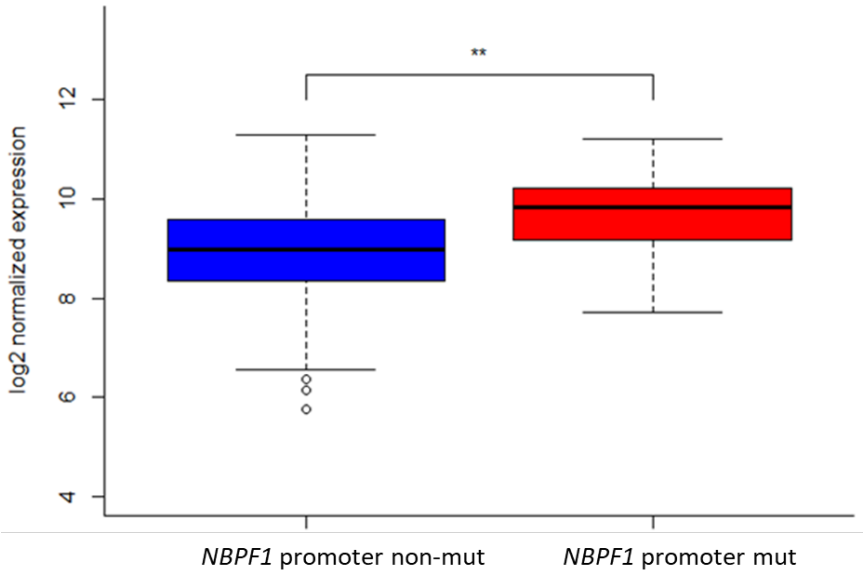


Table 3.1: CREs whose mutations are associated with altered expression of the contacted gene. ($Q < 0.1$)

Fragment	Size (bp)	Gene	Total number of mutations	Total number of mutated samples	Number of mutated samples in differential expression analysis	Number of unmutated samples in differential expression analysis	Mean $\log_2$ gene expression in mutated samples	Mean $\log_2$ gene expression in unmutated samples	Fold change	Differential expression Q-value
chr11:14579387-14583849	4462	*CALCB*	7	7	4	365	7.884	6.159	3.375	9.46E-08
chr2:165615060-165624028	8968	*COBLL1*	12	8	8	491	12.296	11.418	1.762	3.61E-02
chr17:46094139-46103073	8934	*HOXB3*	6	5	5	453	-0.857	4.233	0.037	3.61E-02
chr3:186739608-186745052	5444	*ST6GAL1*	32	25	15	315	14.363	13.893	1.440	3.61E-02
chr9:37375172-37395282	20110	*PAX5*	71	55	13	197	4.471	7.187	0.216	8.39E-02
chr1:16944603-16958779	14176	*ATP13A2*	23	21	14	461	9.594	9.272	1.249	8.83E-02

Table 3.2: CREs whose mutations are associated with altered expression of the contacted gene by subtypes. ($Q < 0.1$)

Subtype	Fragment	Size (bp)	Gene	Total number of mutations	Total number of mutated samples	Number of mutated samples in differential expression analysis	Number of unmutated samples in differential expression analysis	Mean $\log_2$ gene expression in mutated samples	Mean $\log_2$ gene expression in unmutated samples	Fold change	Differential expression Q-value
Hyperdiploid	chr3:187635970-187636359	389	*TPRG1*	9	9	4	114	5.234	2.568	6.347	1.75E-02
***MYC*-translocation**	chr3:186739608-186745052	5444	*ST6GAL1*	6	5	4	30	14.650	13.734	1.887	3.29E-02
	chr3:187635970-187636359	389	*TPRG1*	3	3	3	31	4.627	2.787	3.580	5.17E-02

Figure 3.3: SNVs at CREs affect gene expression in multiple myeloma. Mutations in the CRE significantly alter **(a)** *PAX5* (*n* = 197 versus *n* = 13), **(b)** *ST6GAL1* (*n* = 315 versus *n* = 15) expression. (**c**) *COBLL1* (*n* = 491 versus *n* = 8), (**d**) *HOXB3* (*n* = 453 versus *n* = 5), (**e**) *CALCB* (*n* = 365 versus *n* = 4) and (**f**) *ATP13A2* (*n* = 461 versus *n* = 14). *, *Q* < 0.1, **, *Q* < 0.05, ***, *Q* < 0.01 **(g)** Chromatin looping interactions between *PAX5* promoter and differentially expressed CRE. Also shown are the ChIP-seq signals and relative positions of SNVs. Mut, mutated.

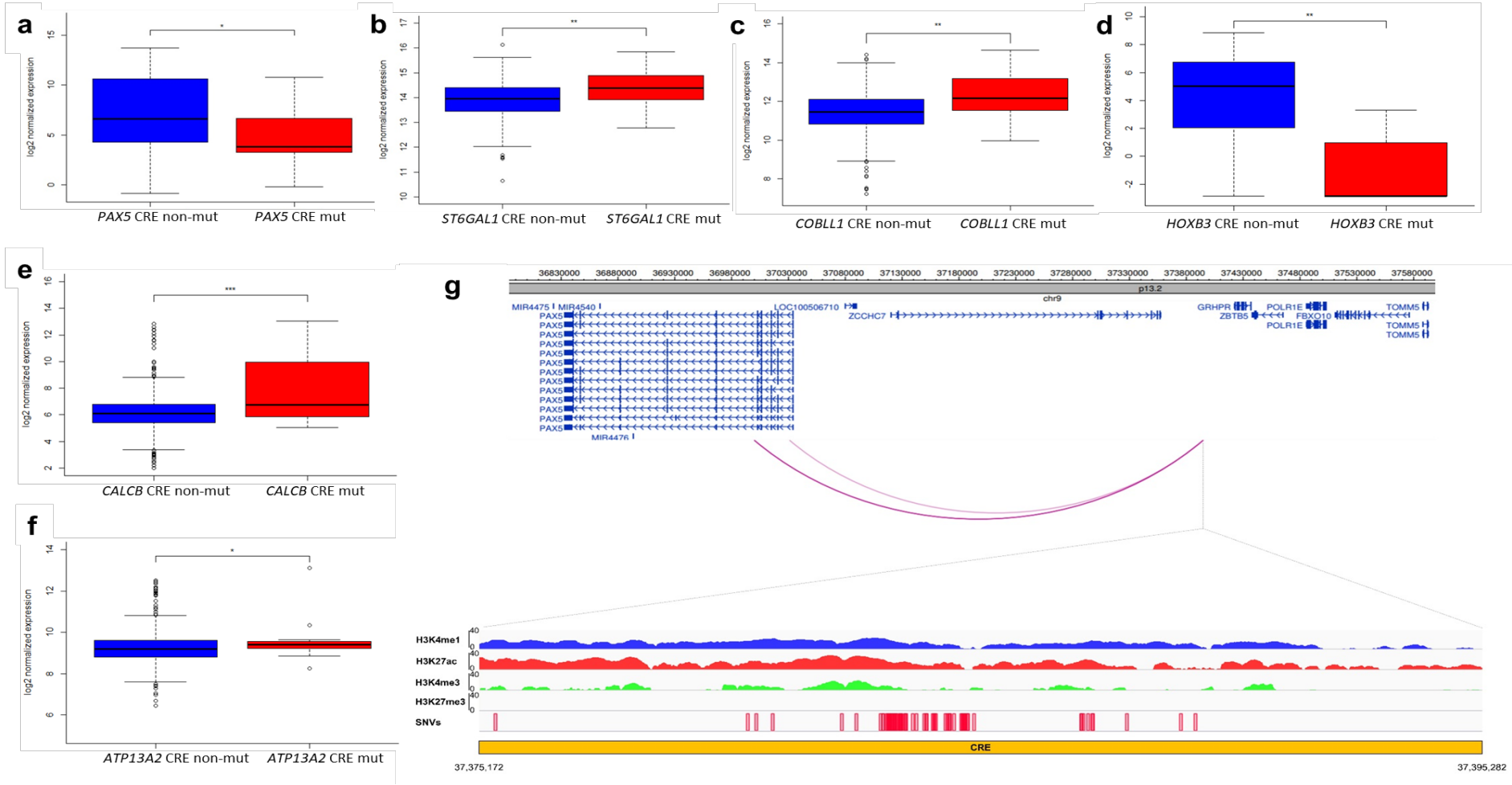


Figure 3.4: CRE mutations affect gene expression of *TPRG1*. (**a**) Hyperdiploid subtype (*n* = 114 versus *n* = 4) and (**b**) *MYC*-translocation subtype (*n* = 31 versus *n* = 3). *, *Q* < 0.1; **, *Q* < 0.05.

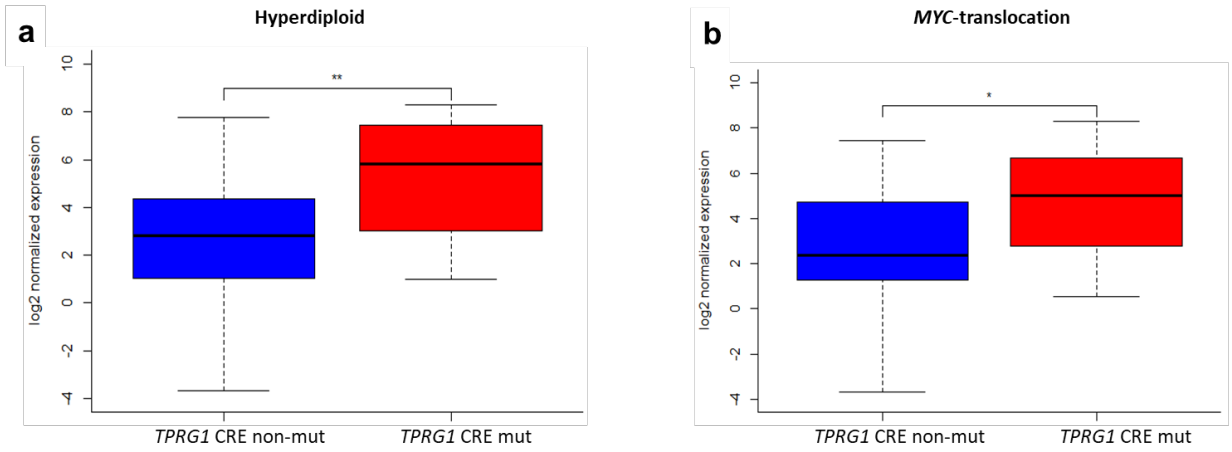


Table 3.3: Subtype analysis to identify associations between the main translocation subtypes and SNVs influencing non-coding CREs. Values in bold indicate statistical significance after adjustment for multiple testing. OR, odd ratio.

Gene	t(4;14)		t(11;14)		t(14;16)		*MYC* translocation		Hyperdiploidy	
	OR	*P*-value	OR	*P*-value	OR	*P*-value	OR	*P*-value	OR	*P*-value
HOXB3	1.767	4.87E-01	2.657	2.66E-01	0.00	1.00E+00	0.00	1.00E+00	0.18	1.67E-01
PAX5	1.864	8.83E-02	**0.212**	**2.65E-03**	1.42	4.78E-01	1.70	1.62E-01	1.40	3.08E-01
NBPF1	1.794	2.98E-01	1.329	5.75E-01	1.27	5.64E-01	0.00	5.81E-02	0.52	1.65E-01
ST6GAL1	0.631	7.57E-01	0.785	8.00E-01	3.70	6.66E-02	1.56	3.78E-01	0.78	5.43E-01
COBLL1	2.375	2.62E-01	0.562	1.00E+00	3.51	2.80E-01	0.00	6.12E-01	0.10	1.13E-02
CALCB	1.174	1.00E+00	1.589	6.34E-01	4.10	2.50E-01	0.97	1.00E+00	0.29	1.37E-01

Table 3.4: Subgroup analysis to identify associations between the major MM subgroups and significantly mutated genes. Values in bold indicate statistical significance after adjustment for multiple testing. Only genes with significant association with at least one subtype are shown. OR, odd ratio.

Gene	t(4;14)		t(11;14)		t(14;16)		*MYC* translocation		Hyperdiploidy	
	OR	*P*-value	OR	*P*-value	OR	*P*-value	OR	*P*-value	OR	*P*-value
MAX	1.894	2.83E-01	1.846	2.45E-01	1.343	5.45E-01	0.000	9.36E-02	0.038	**1.30E-06**
DIS3	2.487	5.55E-03	1.970	2.34E-02	4.184	2.84E-03	0.913	1.00E+00	0.285	**1.64E-06**
PRKD2	7.244	**1.01E-05**	1.896	2.01E-01	0.999	1.00E+00	0.497	5.62E-01	0.179	**2.04E-04**
IRF4	0.428	7.10E-01	10.169	**8.02E-06**	0.000	1.00E+00	0.358	4.91E-01	0.197	2.61E-03
CCND1	0.000	2.37E-01	inf	**1.21E-10**	0.000	1.00E+00	0.000	2.42E-01	0.126	2.67E-03
NRAS	0.138	**1.32E-06**	1.333	1.84E-01	0.389	1.69E-01	1.453	1.30E-01	1.780	2.68E-03

3.3.3 Copy number variants at CREs regulate gene expression

To examine the relationship between CNV at CREs and expression of interacting genes, CNVs that contained both the CRE and its respective target gene from the analysis were excluded. The *MYC* promoter showed both upstream and downstream interactions with 69 CREs; 24 were amplified across 51 tumours and these had significantly higher *MYC* expression ($Q < 0.05$, Table 3.7). These 24 CRE regions clustered within a 110 Kb region forming 10 non-contiguous regions 500 Kb downstream of *MYC* annotated by epigenetic marks indicative of active enhancers (*i.e.* overlapping with strong signals of H3K4me1, H3K27ac, and weak signals of repressive H3K27me3) (Figure 3.5a). Five CRE regions upstream of *MYC* interacting with *MYC* promoter were deleted in 10 tumours (distinct from the 51 tumours with CREs amplified) which were associated with higher *MYC* expression ($Q < 0.1$, Table 3.8). These CREs, clustered within a 13 Kb region, 850 Kb upstream of *MYC*, form two non-contiguous regions with weaker signals for H3K4me1, H3K4me3 and H3K27ac, and stronger signals for repressive mark H3K27me3, consistent with putative silencers of *MYC* (Figure 3.5a). Since *MYC* is translocated in 15-20% of newly diagnosed MM[2] (14% of CoMMpass samples, Table 3.1) the possibility that upregulation of *MYC* expression associated with CRE CNVs might be the consequence of translocation of *MYC* to proximal super-enhancers was examined. A broader set of 209 samples with putative *MYC* translocations (24% of total tumours) was defined and 51 samples with amplified CREs are indeed highly enriched for translocations (34/51, $P = 1.2 \times 10^{-11}$, Fisher's exact test), with the breakpoints mapping to the region of amplification. The deletions at CREs were not, however, enriched for translocations (1/10, $P = 0.9$) and in *MYC*-translocation negative cases the CNVs at *MYC* CREs were still associated with significantly increased *MYC* expression (Figure 3.5b, $P = 8.6 \times 10^{-3}$, 2.3-fold).

Six other novel candidate genes whose expression was significantly altered by CNVs at respective interacting CREs were identified: *PACS2*, *TEX22*, *KDM3B*, *RAB36*, *PLD4*, and *SP110* (Table 3.7, Table 3.8, Figure 3.6). While each of the respective CREs were annotated by epigenetic marks indicative of functional regulatory regions these genes reside close to regions of common structural variation, making interpretation of their specific relevance problematic.

Table 3.5: CREs whose amplification is associated with significantly altered gene expression. ($Q < 0.1$)

Fragment	Gene	Number of mutated samples	Number of unmutated samples	Log_2 fold-change in expression	Differential expression Q-value
chr14:106003753-106005907	*PACS2*	21	333	0.654	3.34E-12
chr14:106003753-106005907	*TEX22*	12	333	0.888	1.27E-02
chr8:129213699-129215844	*MYC*	18	375	0.996	1.27E-02
chr8:129215845-129218722	*MYC*	18	376	0.992	1.27E-02
chr8:129218723-129220126	*MYC*	18	376	0.992	1.27E-02
chr8:129220127-129222593	*MYC*	18	376	0.992	1.27E-02
chr8:129222594-129223807	*MYC*	19	375	0.986	1.27E-02
chr8:129223808-129225891	*MYC*	19	375	0.986	1.27E-02
chr8:129283627-129290175	*MYC*	35	359	0.757	1.27E-02
chr8:129290176-129291125	*MYC*	36	358	0.811	1.27E-02
chr8:129303419-129306797	*MYC*	41	354	0.783	1.27E-02
chr8:129241713-129243523	*MYC*	21	373	0.907	1.30E-02
chr8:129340049-129341634	*MYC*	30	365	0.844	1.30E-02
chr8:129341635-129344398	*MYC*	30	365	0.844	1.30E-02
chr8:129281412-129283626	*MYC*	33	361	0.737	1.61E-02
chr14:105814065-105821419	*PLD4*	8	352	2.332	1.61E-02
chr8:129256404-129257791	*MYC*	26	368	0.800	1.65E-02
chr8:129200591-129213698	*MYC*	16	377	0.979	1.89E-02
chr8:129273487-129276207	*MYC*	30	364	0.736	1.89E-02
chr8:129276532-129278298	*MYC*	30	364	0.736	1.89E-02
chr8:129278299-129278864	*MYC*	30	364	0.736	1.89E-02
chr8:129278865-129281411	*MYC*	30	364	0.736	1.89E-02
chr8:129314900-129319098	*MYC*	37	358	0.687	2.14E-02
chr8:129314402-129314899	*MYC*	36	359	0.659	3.08E-02
chr8:129306798-129308447	*MYC*	37	358	0.647	3.40E-02
chr8:129319128-129321850	*MYC*	35	360	0.666	3.40E-02
chr8:129324805-129327116	*MYC*	34	361	0.657	4.24E-02
chr5:138607458-138607716	*KDM3B*	9	285	0.261	4.27E-02

Table 3.6: CREs whose deletion is associated with significantly altered gene expression. ($Q < 0.1$)

Fragment	Gene	Number of mutated samples	Number of unmutated samples	Log_2 fold-change in expression	Differential expression Q-value
chr8:127886760-127889453	*MYC*	10	388	1.012	8.00E-02
chr8:127889454-127891696	*MYC*	10	388	1.012	8.00E-02
chr8:127891697-127895194	*MYC*	10	388	1.012	8.00E-02
chr8:127895195-127897477	*MYC*	10	388	1.012	8.00E-02
chr8:127897662-127899869	*MYC*	10	388	1.012	8.00E-02
chr2:231268251-231269730	*SP110*	7	462	0.578	8.00E-02
chr2:231269731-231271492	*SP110*	7	462	0.578	8.00E-02
chr2:231282634-231286088	*SP110*	7	462	0.578	8.00E-02
chr2:231286089-231290028	*SP110*	7	462	0.578	8.00E-02
chr2:231296729-231301548	*SP110*	7	462	0.578	8.00E-02
chr2:231301549-231311636	*SP110*	7	462	0.578	8.00E-02
chr2:231311637-231316348	*SP110*	7	462	0.578	8.00E-02
chr2:231316445-231318918	*SP110*	7	462	0.578	8.00E-02
chr2:231318919-231321183	*SP110*	7	462	0.578	8.00E-02
chr22:23300832-23302691	*RAB36*	133	236	0.360	8.00E-02

Figure 3.5: Copy number variations at *cis*-regulatory elements affect *MYC* gene expression. (a) Upper panel shows *MYC* gene expression may be regulated by CREs; CNVs at either the upstream putative silencers or downstream putative enhancers causing upregulation of *MYC*. Middle panel shows chromatin looping interactions between *MYC* promoter and CREs. Lower panel details ChIP-seq signals and relative positions of CNVs at these CREs in naïve B-cells. (b) CNV status at CREs and *MYC* expression. Difference in expression was assessed pairwise between samples with different CNVs status and the same translocation status. ***, $P < 0.01$. Trans, translocation. Del, deletion. From left to right n = 345, n = 9, respectively.

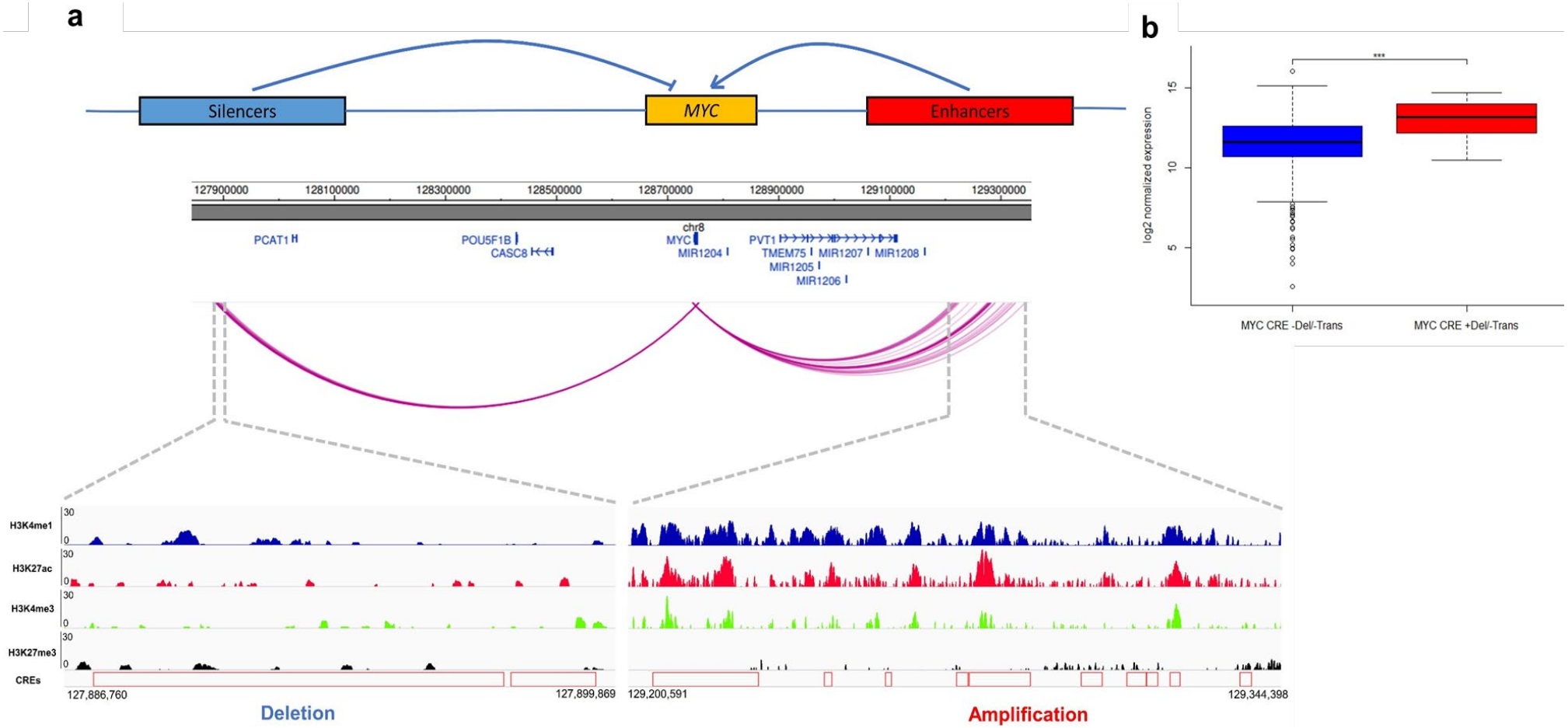
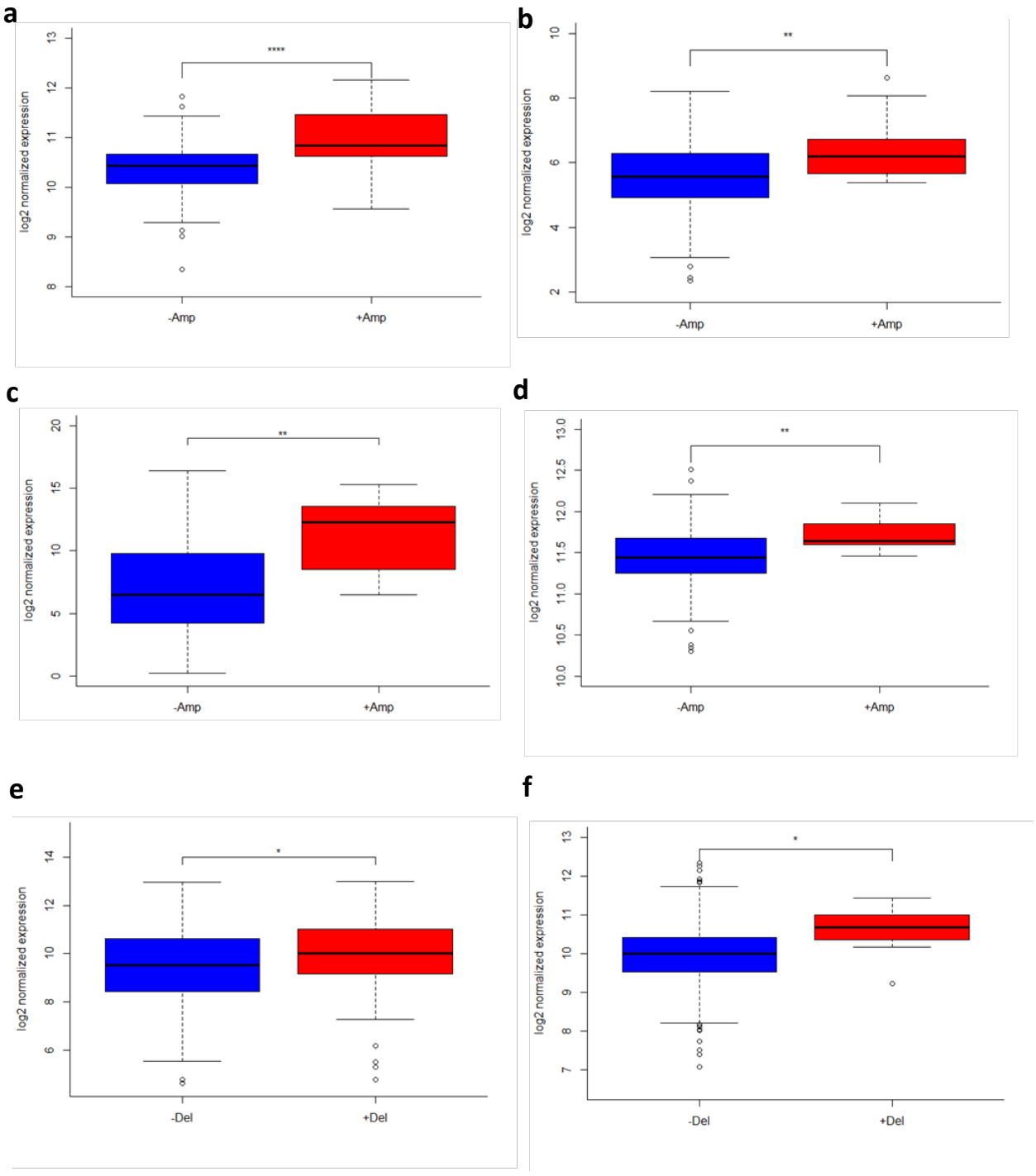


Figure 3.6: The effects of CNVs at CREs on gene expression in MM. Boxplots show differential gene expression between CNV unaffected (blue) versus CNV affected samples (red) at CREs interacting with promoters of (a) *PACS2* (n = 333 versus n = 21); (b) *TEX22* (n = 333 versus n = 12); (c) *PLD4* (n = 352 versus n = 8); (d) *KDM3B* (n = 285 versus n = 9); (e) *RAB36* (n = 236 versus n = 133); and (f) *SP110* (n = 462 versus n = 7). *, $Q < 0.1$; **, $Q < 0.05$; ****, $Q < 0.001$. Amp, amplification. Del, deletion.



3.3.4 Chromosomal copy number alterations

Multiple frequent copy number alterations were detected in MM tumours (Figure 3.7). Pre-eminently, gain of odd numbered chromosomes, characteristic of HD MM[2], was seen in 59% of the tumours, with chromosome 9, 15, and 19 most often amplified (83-86% HD, Table 3.9); concordant with published observations[2]. Deletion of chromosomal cytobands containing IG loci *IGK* (2p11.2), *IGH* (14q32.33) and *IGL* (22q11.22) were present in 95%, 98% and 57% of the tumours respectively (Figure 3.7), consistent with the rearrangements expected at IG loci during normal B-cell development[210]. Common deletions were also seen at 13q (63%), 14q (43%), 16q (38%) and 8p (38%). Despite the relatively low overall level of chromosome 8 amplification, 28% of the tumours exhibited amplification overlapping 8q24.21 that incorporates *MYC* (13%) and *PVT1* (16%)[211, 212].

Figure 3.7: Summary of amplifications and deletions in 725 MM samples. The proportion of samples with amplifications (cyan) and deletions (orange) overlapping each cytoband is plotted by karyoplotR[9]. The frequent deletions (orange peaks) at 2p11.2, 14q32.33 and 22q11.22 overlap with the immunoglobulin loci *IGK*, *IGH*, and *IGL* respectively.

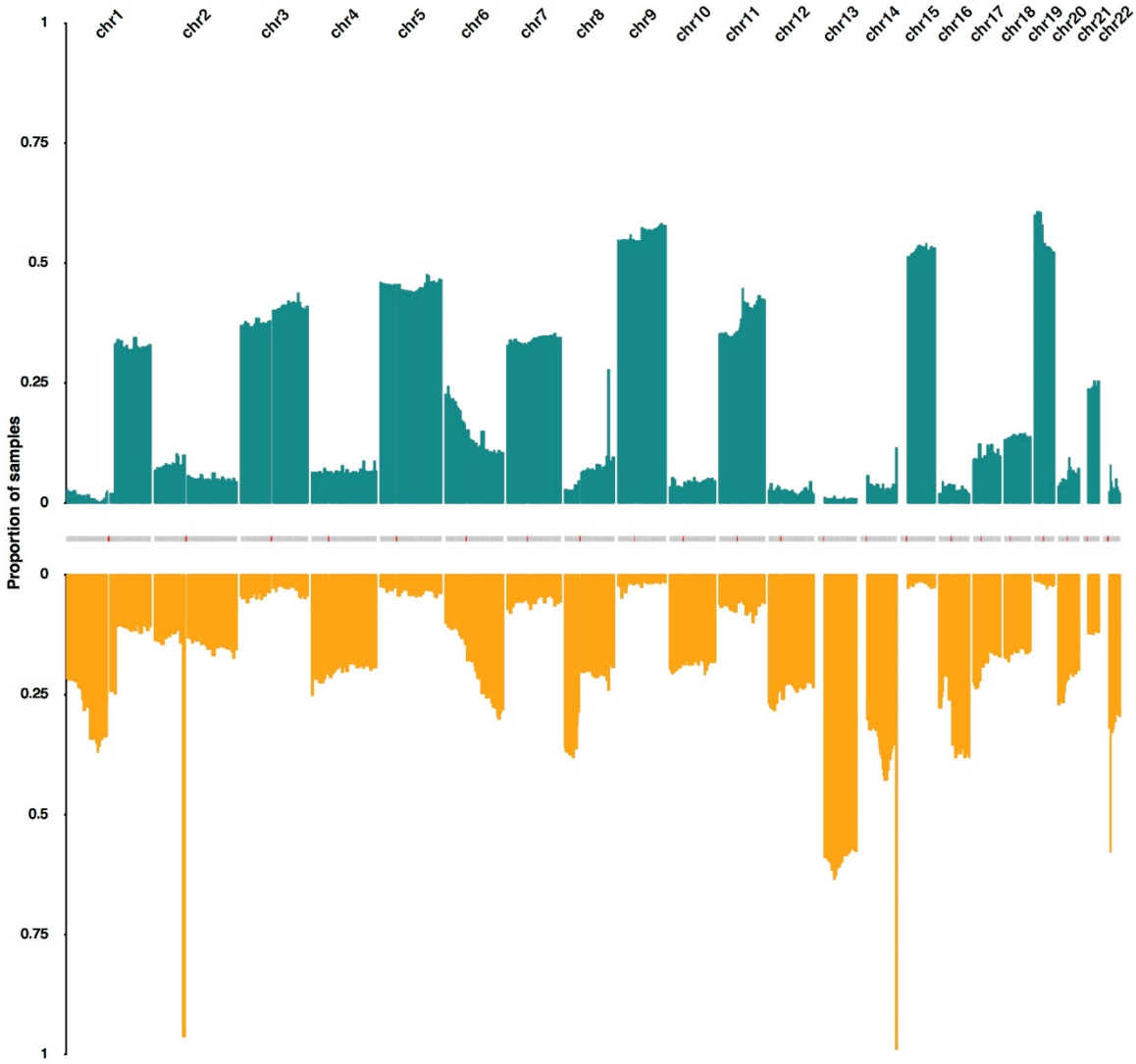


Table 3.7: Copy number alterations in 725 MM samples

Chromosome amplification and hyperdiploidy		
Chromosome	Number of samples	
	Amplified	Hyperdiploidy
chr1	5	5
chr2	41	41
chr3	270	261
chr4	40	40
chr5	319	317
chr6	81	81
chr7	240	239
chr8	14	14
chr9	381	366
chr10	19	19
chr11	256	255
chr12	7	7
chr13	5	5
chr14	17	16
chr15	361	355
chr16	2	2
chr17	50	49
chr18	90	86
chr19	373	365
chr20	19	19
chr21	123	119
chr22	6	6

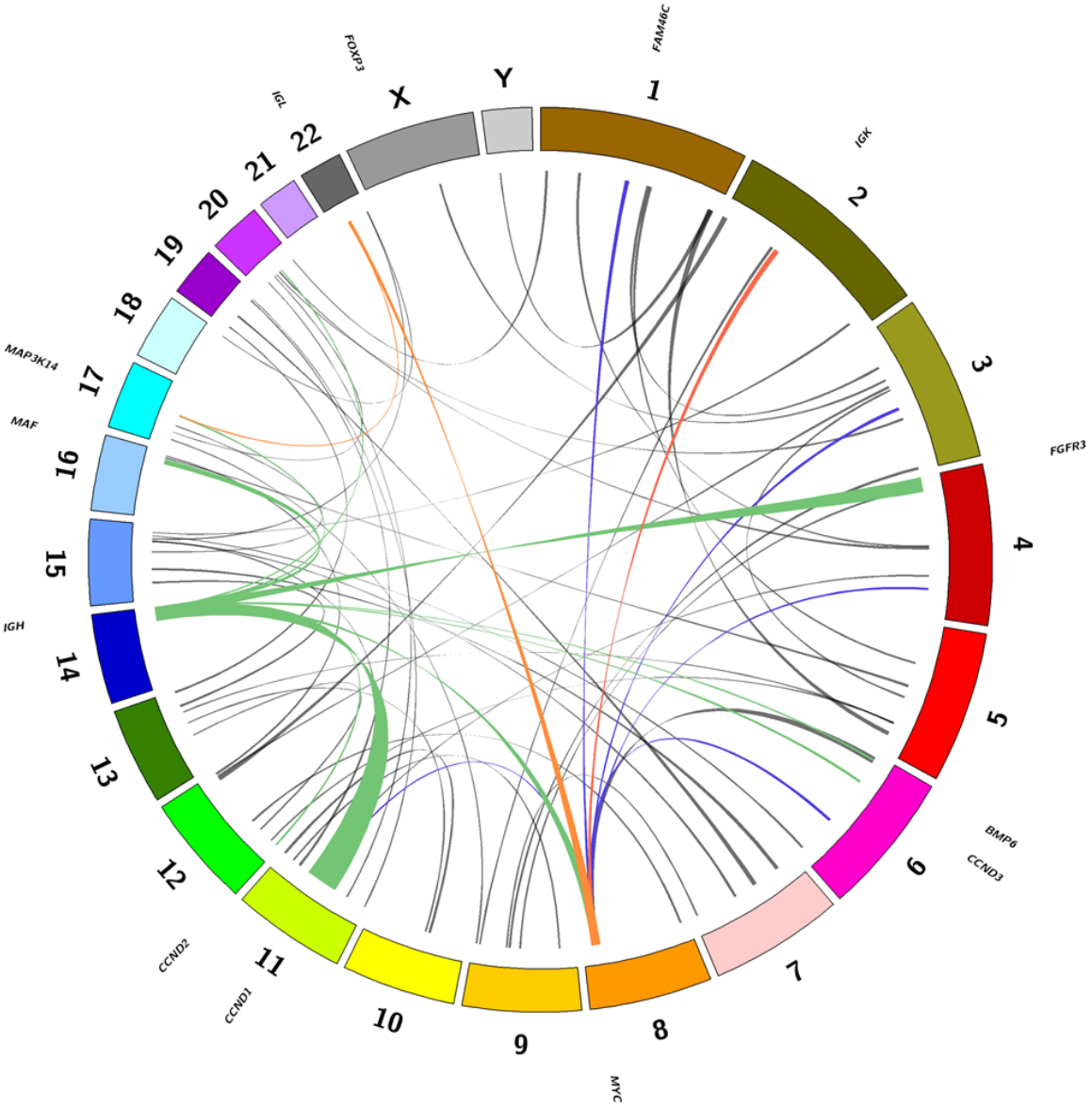
3.3.5 Structural variation

The median rate of SVs was 10 across tumours; four translocations (range 0-147) and six inversions (range 0-2,790). Considering SVs falling within gene boundaries, on average six genes were disrupted per tumour. SVs were also identified as affecting genes commonly mutated in MM[1, 3, 5] including *CYLD* with inversions disrupting the protein sequence in five samples (Table 3.10). Widening the definition of SVs to genes within a 1 Mb window of translocation breakpoints identified multiple recurrent rearrangements including *MYC*, *CCND1* and *FGFR3,* detected in 173 (23%), 124 (16%) and 46 (6%) of samples, respectively. *MYC* rearrangements involved a plethora of partner sites including *IGH* (32/765), *IGL* (32/765), *IGK* (11/765), and cytobands encompassing *BMP6* (21/765), *FAM46C* (9/765), *CCND1* (1/765) and *MAF* (1/765). Novel *MYC* translocations disrupting *CD96* (immune checkpoint receptor target) were identified in eight tumours and translocations intergenic to *PRDM1* and *FBXW7* in eight and five tumours, respectively. Restricting this analysis to translocations incorporating the *IGH*, *IGK* and *IGL* loci, common translocations were identified affecting 17q21.31, encompassing *MAP3K14*, in 16 tumours, and 10 tumours with translocations affecting 12p13.32, encompassing *CCND2* (Figure 3.8). Tumours with these translocations were associated with upregulation of *MAP3K14* (7.4-fold upregulation, $P = 5.05 \times 10^{-41}$), and *CCND2* (11.9-fold upregulation, $P = 7.5 \times 10^{-5}$).

Table 3.8: Structural variants affecting genes reported as recurrently mutated in MM. Gene lists were combined from Walker *et al.*[1], Lohr *et al.*[3], Bolli *et al.*[5], and CoMMpass study.

Gene	Chromosome location	Number of samples	
		Within gene boundary	Within 1Mb
KRAS	12p12.1	0	10
NRAS	1p13.2	1	18
FAM46C	1p12	6	32
BRAF	7q34	0	9
TP53	17p13.1	1	19
DIS3	13q22.1	0	6
PRDM1	6q21	1	23
SP140	2q37.1	2	14
EGR1	5q31.2	0	10
TRAF3	14q32.32	4	11
ATM	11q22.3	2	9
CCND1	11q13.3	3	124
HIST1H1E	6p22.2	0	4
LTB	6p21.33	0	14
IRF4	6p25.3	1	7
FGFR3	4p16.3	0	46
RB1	13q14.2	3	15
ACTG1	17q25.3	0	9
CYLD	16q12.1	5	14
MAX	14q23.3	2	17
ATR	3q23	0	7
SAMHD1	20q11.23	2	23
PRKD2	19q13.32	1	11
PTPN11	12q24.13	0	4
TGDS	13q32.1	0	3
DNAH5	5p15.2	2	4
MYH2	17p13.1	0	9
BMP2K	4q21.21	2	10
ZNF208	19p12	0	28
RPL10	Xq28	0	11
TBC1D29	17q11.2	0	12
FBXO4	5p13.1	0	5
RASA2	3q23	2	13
OR5M1	11q12.1	0	16
RPS3A	4q31.3	0	6
PTH2	19q13.33	0	18
BAX	19q13.33	0	22
C8orf86	8p11.22	0	12
CELA1	12q13.13	0	6
FCF1	14q24.3	0	10
FTL	19q13.33	0	22
OR9G1	11q12.1	0	15
TNFSF12	17p13.1	0	18
FAM154B	15q25.2	0	0
HIST1H4H	6p22.2	0	3
LEMD2	6p21.31	0	4
TRAF2	9q34.3	1	5
SGPP1	14q23.2	0	7
RPN1	3q21.3	1	7
PABPC1	8q22.3	6	11

Figure 3.8: Circos plot of common translocations (> 5 samples). *IGK* (chr2) *IGH* (chr14) and *IGL* (chr22) translocations are depicted in red, green and orange, respectively. *MYC* translocations in blue. The ribbon is centred on the cytoband implicated with the ribbon width proportional to the number of affected samples.



3.3.6 Significantly mutated protein-coding genes

To gain insight into mutations affecting the protein-coding regions, MutSigCV[160] was applied to variants identified from WES data. I identified 33 significantly mutated genes ($Q < 0.05$, Table 3.11). These were over-represented in pathways involved in sustaining proliferative signalling, activating invasion, evading growth suppressors, tumour-promoting inflammation, resisting cell death, enabling replicative immortality, and angiogenesis ($P < 0.05$, Table 3.12). While 16 of the 33 genes have previously been documented to be recurrently mutated in MM (*KRAS*, *NRAS, HIST1H1E*, *MAX*, *SP140*, *RASA2, FCF1*, *DIS3, BRAF, TP53*, *SAMHD1*, *TRAF3, PRKD2*, *TGDS*, *CYLD*, and *RB1;* Table 3.13)[1-3, 5, 192], 17 novel significantly mutated genes were identified. These included 12 genes previously reported as recurrently mutated, albeit not significantly (*PTPN11*, *DNAH5*, *MYH2*, *BMP2K*, *ZNF208*, *RPL10*, *FBXO4*, *OR5M1*, *PTH2*, *CELA1*, *OR9G1*, and *TNFSF12*)[1, 3, 5-8] and five novel genes (*TBC1D29*, *RPS3A*, *BAX*, *C8orf86*, and *FTL*) (Table 3.11).

Stratifying MM according to its major subgroups (HD, *MYC*-translocation, t(4;14), t(11;14), t(14;16)) allowed identification of additional drivers; *FAM154B*, *HIST1H4H*, *LEMD2* and *PABPC1* in HD; *RPN1* and *TRAF2* in *MYC*-translocation; *SGPP1* in t(11;14); and *TRAF2* in t(14;16) (Table 3.14). Furthermore, t(4;14) MM was identified as being enriched for *PRKD2* mutations (13% of subtype, $P = 1.0 \times 10^{-5}$) but having a paucity of *NRAS* mutations ($P = 1.3 \times 10^{-6}$); possibly reflecting dysregulation of the MAPK-signalling, a consequence of the translocation-mediated *FGFR3* overexpression (Table 3.6). As previously reported, t(11:14) MM was identified as associated with *CCND1* mutation[84] (10%, $P = 1.2 \times 10^{-10}$) and *IRF4* mutation (8%, $P = 8.0 \times 10^{-6}$). In contrast, mutations in *PRKD2* ($P = 2.0 \times 10^{-4}$), *MAX* ($P = 1.3 \times 10^{-6}$) and *DIS3* ($P = 1.6 \times 10^{-6}$) were infrequent in HD. Finally, somatic mutations in the following genes had low alternative allelic fraction - *RPS3A* (range 0.1 – 0.5), *TBC1D29* (range 0.1 – 0.5), *PABPC1* (range 0.1 – 0.4), and *TRAF2* (range 0.1 – 0.9), reflecting the heterogeneity of MM.

Table 3.9: Significantly mutated genes identified in 804 tumours from CoMMpass (IA9 dataset). ($Q < 0.05$).

Gene	Chromosome location	Start (bp)	End (bp)	No. non-silent mutations	*Q*- value
KRAS	12p12.1	25357723	25403870	221	2.22E-16
NRAS	1p13.2	115247090	115259515	195	2.22E-16
HIST1H1E	6p22.2	26156559	26157343	33	2.22E-16
MAX	14q23.3	65472892	65569413	26	2.22E-16
SP140	2q37.1	231067826	231223762	26	2.22E-16
TBC1D29	17q11.2	28884130	28890511	14	2.22E-16
RASA2	3q23	141205889	141334184	13	2.22E-16
RPL10	Xq28	153618315	153637504	13	2.22E-16
RPS3A	4q31.3	152020725	152025804	11	2.22E-16
C8orf86	8p11.22	38368352	38386180	6	2.22E-16
FBXO4	5p13.1	41925356	41941845	6	2.22E-16
OR5M1	11q12.1	56380031	56380978	6	2.22E-16
OR9G1	11q12.1	56467864	56468781	5	2.22E-16
PTH2	19q13.33	49925671	49926698	5	2.22E-16
CELA1	12q13.13	51722227	51740463	4	2.22E-16
FCF1	14q24.3	75179847	75203394	4	2.22E-16
FTL	19q13.33	49468558	49470135	3	2.22E-16
TNFSF12	17p13.1	7452208	7464925	3	2.22E-16
BAX	19q13.33	49458072	49465055	2	2.22E-16
DIS3	13q22.1	73329540	73356234	85	5.86E-12
BRAF	7q34	140419127	140624564	62	9.07E-12
TP53	17p13.1	7565097	7590856	46	4.47E-09
SAMHD1	20q11.23	35518632	35580246	19	8.29E-09
TRAF3	14q32.32	103243813	103377837	72	1.12E-06
PTPN11	12q24.13	112856155	112947717	19	4.72E-05
PRKD2	19q13.32	47177532	47220384	26	9.91E-05
TGDS	13q32.1	95226308	95248511	14	5.49E-04
CYLD	16q12.1	50775961	50835846	22	1.18E-03
MYH2	17p13.1	10424465	10453274	24	9.13E-03
DNAH5	5p15.2	13690440	13944652	46	1.07E-02
BMP2K	4q21.21	79697496	79837526	16	1.37E-02
RB1	13q14.2	48877887	49056122	15	1.82E-02
ZNF208	19p12	22115760	22193751	28	2.92E-02

Table 3.10: Gene-set enrichment analysis of significantly mutated genes. GO, gene ontology.

GO term ID	GO term	Cancer hallmark category	Number of occurences of annotation in candidate set	Expected number of occurences of annotation in candidate set	Number of occurences of annotation in background set	*P*-value
GO:0007166	Cell surface receptor signalling pathway	Sustaining proliferative signaling	16	4.793	2548	3.70E-06
GO:0070848	Response to growth factor	Sustaining proliferative signaling	7	1.653	879	1.05E-03
GO:0016477	Cell migration	Activating invasion	7	2.069	1100	3.79E-03
GO:0008283	Cell proliferation	Evading growth suppressors	9	3.422	1819	5.21E-03
GO:0045321	Leukocyte activation	Tumor-promoting inflammation	5	1.258	669	7.78E-03
GO:0002326	B-cell lineage commitment	Sustaining proliferative signaling	1	0.009	5	9.37E-03
GO:0012501	Programmed cell death	Resisting cell death	8	3.297	1753	1.40E-02
GO:0010941	Regulation of cell death	Resisting cell death	7	2.703	1437	1.57E-02
GO:0030183	B-cell differentiation	Sustaining proliferative signaling	2	0.203	108	1.75E-02
GO:0090399	Replicative senescence	Enabling replicative immortality	1	0.019	10	1.87E-02
GO:0001525	Angiogenesis	Angiogenesis	3	0.754	401	3.90E-02
GO:0007155	Cell adhesion	Activating invasion	6	2.573	1368	3.99E-02
GO:0060548	Negative regulation of cell death	Resisting cell death	4	1.621	862	7.67E-02
GO:0090398	Cellular senescence	Enabling replicative immortality	1	0.096	51	9.17E-02
GO:0042100	B-cell proliferation	Sustaining proliferative signaling	1	0.162	86	1.50E-01
GO:0032200	Telomere organization	Enabling replicative immortality	1	0.166	88	1.53E-01
GO:0007049	Cell cycle	Evading growth suppressors	5	3.002	1596	1.77E-01
GO:0000819	Sister chromatid segregation	Genome instability	1	0.211	112	1.91E-01
GO:0006091	Generation of precursor metabolites and energy	Disrupting cellular energetics	2	0.813	432	1.95E-01
GO:0000187	Activation of MAPK activity	Sustaining proliferative signaling	1	0.275	146	2.41E-01
GO:0006281	DNA repair	Genome instability	1	0.775	412	5.44E-01
GO:0006954	Inflammatory response	Tumor-promoting inflammation	1	1.117	594	6.79E-01
GO:0001910	Regulation of leukocyte mediated cytotoxicity	Avoiding immune destruction	0	0.090	48	1.00
GO:0002507	Tolerance induction	Avoiding immune destruction	0	0.049	26	1.00
GO:0002767	Immune response-inhibiting cell surface receptor signaling pathway	Avoiding immune destruction	0	0.009	5	1.00
GO:0007065	Sister chromatid cohesion	Genome instability	0	0.000	0	1.00
GO:0010695	Regulation of spindle pole body separation	Genome instability	0	0.000	0	1.00
GO:0010718	Positive regulation of epithelial to mesenchymal transition	Activating invasion	0	0.058	31	1.00
GO:0019882	Antigen processing and presentation	Avoiding immune destruction	0	0.408	217	1.00
GO:0030997	Regulation of centriole-centriole cohesion	Genome instability	0	0.006	3	1.00
GO:0031577	Spindle checkpoint	Genome instability	0	0.092	49	1.00
GO:0034330	Cell junction organization	Activating invasion	0	0.463	246	1.00
GO:0038061	NIK/NF-kappaB signaling	Sustaining proliferative signaling	0	0.188	100	1.00
GO:0046605	Regulation of centrosome cycle	Genome instability	0	0.060	32	1.00
GO:0051383	Kinetochore organization	Genome instability	0	0.024	13	1.00
GO:0051988	Regulation of attachment of spindle microtubules to kinetochore	Genome instability	0	0.019	10	1.00
GO:0090224	Regulation of spindle organization	Genome instability	0	0.038	20	1.00

Table 3.11: Significantly mutated genes in MM identified in different studies. Walker *et al.*[1], Lohr *et al.*[3], Bolli *et al.*[5], *Kortum et al.*[6], *Hofman et al.*[7], Walker *et al. 2012*[8], CoMMpass (this study). *, identified as significantly mutated in the study.

Gene	Walker *et al.* % (n = 463)	Lohr *et al.* %(n = 203)	Bolli *et al.* %(n = 67)	CoMMpass %(n = 804)	Other study
KRAS	21*	23*	25*	24*	
NRAS	19*	20*	25*	22*	
FAM46C	6*	11*	12*	9	
BRAF	7*	6*	15*	7*	
TP53	3*	8*	15*	5*	
DIS3	9*	11*	1	10*	
PRDM1	2	5*	0	2	
SP140	2	4	7*	3*	
EGR1	4*	4	7	4	
TRAF3	4*	5*	3	7*	
ATM	3	4	3	3	
CCND1	2*	3	4	2	
HIST1H1E	3*	0	0	4*	
LTB	3*	1	4*	3	
IRF4	3*	2	0	3	
FGFR3	3*	2	0	3	
RB1	2	3*	0	2*	
ACTG1	5	2*	0	3	
CYLD	2*	2*	3	2*	
MAX	2*	1	0	3*	
ATR	1	1	1	1	
SAMHD1	<1	2	1	2*	
PRKD2	2	3	4	3*	2/22 (Walker *et al*. 2012)
PTPN11	2	2	0	2*	2% (Kortum *et al*.)
TGDS	1	0	4	2*	
DNAH5	3	5	6	5*	3/22 (Walker *et al.* 2012)
MYH2	2	1	0	3*	
BMP2K	1	1	0	2*	1/22 (Walker *et al*. 2012)
ZNF208	1	3	4	3*	
RPL10	1	2	0	2*	2% (Hofman *et al*.)
TBC1D29	2	0	0	2*	
FBXO4	0	1	1	1*	
RASA2	1	3	3	1*	
OR5M1	0	1	0	1*	
RPS3A	0	0	0	1*	
PTH2	0	1	0	1*	
BAX	<1	0	0	<1*	
C8orf86	<1	0	0	<1*	
CELA1	0	<1	0	<1*	
FCF1	<1	0	0	<1*	
FTL	0	0	0	<1*	
OR9G1	<1	<1	0	<1*	
TNFSF12	0	<1	0	<1*	
TRAF2	1	2	0	2*	
FAM154B	<1	<1	0	<1*	
HIST1H4H	1	<1	0	<1*	
LEMD2	<1	0	0	<1*	
PABPC1	1	1	0	4*	
RPN1	0	0	0	<1*	
SGPP1	<1	1	0	1*	

Table 3.12: Significantly mutated genes identified through CoMMpass (IA9 dataset) by major subgroups. ($Q < 0.05$). *, genes that were not previously identified as significantly mutated in general analysis.

Subtype	Gene	Chromosome location	Start (bp)	End (bp)	Number of non-silent mutations	*Q*-value
Hyperdiploidy	*BRAF*	7q34	140419127	140624564	30	2.22E-16
	FAM154B*	15q25.2	82555151	82577271	2	2.22E-16
	HIST1H4H*	6p22.2	26281283	26285762	4	2.22E-16
	LEMD2*	6p21.31	33738979	33756913	2	2.22E-16
	NRAS	1p13.2	115247090	115259515	119	2.22E-16
	OR9G1	11q12.1	56467864	56468781	5	2.22E-16
	RASA2	3q23	141205889	141334184	9	2.22E-16
	RPL10	Xq28	153618315	153637504	12	2.22E-16
	RPS3A	4q31.3	152020725	152025804	9	2.22E-16
	TRAF3	14q32.32	103243813	103377837	28	2.22E-16
	KRAS	12p12.1	25357723	25403870	108	1.01E-11
	TP53	17p13.1	7565097	7590856	17	4.47E-06
	DIS3	13q22.1	73329540	73356234	23	5.48E-06
	PABPC1*	8q22.3	101698044	101735037	21	1.55E-03
	HIST1H1E	6p22.2	26156559	26157343	15	3.49E-03
***MYC*-translocation**	*C8orf86*	8p11.22	38368352	38386180	2	2.22E-16
	KRAS	12p12.1	25357723	25403870	36	2.22E-16
	NRAS	1p13.2	115247090	115259515	35	2.22E-16
	RPN1*	3q21.3	128338817	128399918	3	2.22E-16
	RPS3A	4q31.3	152020725	152025804	5	2.22E-16
	TRAF2*	9q34.3	139776364	139821059	2	2.22E-16
	BRAF	7q34	140419127	140624564	9	4.13E-03
	RPL10	Xq28	153618315	153637504	4	1.35E-02
	DIS3	13q22.1	73329540	73356234	10	1.53E-02
t(4;14)	*NRAS*	1p13.2	115247090	115259515	4	2.22E-16
	PRKD2	19q13.32	47177532	47220384	12	2.22E-16
	KRAS	12p12.1	25357723	25403870	19	5.58E-12
	DIS3	13q22.1	73329540	73356234	17	1.74E-04
	TRAF3	14q32.32	103243813	103377837	10	1.92E-03
	BRAF	7q34	140419127	140624564	6	2.01E-02
	TBC1D29	17q11.2	28884130	28890511	5	2.57E-02
	MAX	14q23.3	65472892	65569413	4	2.57E-02
t(11;14)	*HIST1H1E*	6p22.2	26156559	26157343	11	2.22E-16
	MAX	14q23.3	65472892	65569413	8	2.22E-16
	NRAS	1p13.2	115247090	115259515	47	2.22E-16
	SGPP1*	14q23.2	64150932	64194757	2	2.22E-16
	TP53	17p13.1	7565097	7590856	17	2.22E-16
	KRAS	12p12.1	25357723	25403870	49	2.44E-12
	DIS3	13q22.1	73329540	73356234	25	6.82E-09
	BRAF	7q34	140419127	140624564	12	1.08E-06
t(14;16)	***TRAF2****	9q34.3	139776364	139821059	6	1.10E-03

3.3.7 Pathways targeted by both coding and non-coding mutations

Pathways targeted by coding and non-coding mutations were identified using the Reactome pathway tool[196]. These included MAPK signalling, NF-κB signalling, cytokine signalling, GPCR signalling, transcriptional and post-translational expression regulation, hematopoietic development, DNA damage, and apoptosis ($Q < 0.05$, Appendix 1). Many of the genes in these pathways are targeted by both coding and non-coding drivers (Table 3.15, Figure 3.9), exemplified by *IRF4* and *PRDM1,* along with *BCL6* and *PAX5*, genes central to plasma cell differentiation[2].

3.3.8 Mutational signatures

To gain insight into the aetiological basis of MM mutations, mutational signatures were analysed[87]. Mutational signature 2 (C > T/G in TC dinucleotide motif), a consequence of the activity of the APOBEC family of cytidine deaminases[87], associated with poor prognosis[84, 87], was seen in 30% (230/765) of tumours (Appendix 2) and associated with coding mutations in *DNAH5* ($P = 8.8 \times 10^{-7}$), *SAMHD1* ($P = 7.2 \times 10^{-4}$), *TP53* ($P = 9.3 \times 10^{-3}$), and *BRAF* ($P = 3.7 \times 10^{-2}$). This signature was primarily enriched in *MAF* translocations t(14;16) (30/31, $P = 1.2 \times 10^{-15}$, mean mutational contribution 0.37) and t(14;20) (7/9, $P = 4.1 \times 10^{-3}$, mean mutational contribution 0.28) and to a lesser extent with t(4;14) (46/93, $P = 1.1 \times 10^{-5}$, mean mutational contribution 0.07).

Other mutational signatures previously reported in MM[5, 84, 87, 213] were also identified, including signature 1, 5, 9, and 13 in 18% (135/765), 73% (557/765), 96% (737/765), and 5% (36/765) of tumours, respectively (Appendix 2). Almost all samples (35/36) with signature 13 also exhibited signature 2, consistent with the published literature[87]. Mutational signatures not previously reported in MM included signatures 3, 8, 16, and 30 seen in more than 30% of tumours (Appendix 2). No additional signatures were identified when analysing the high coverage WES data. Signature 9 (T > G in WT motif with W = A or T), a consequence of activation-induced cytidine deaminase (AID) activity[87], is also a feature of chronic lymphocytic leukemia (CLL) and B-cell lymphomas. The fact that, despite its prevalence, this signature had not previously been identified in earlier large scale

analyses, agrees with the assertion that AID related mutations are enriched in non-coding regions and early mutation events[213]. Since signature 9 suggests AID off-target activity, the mutational patterns of somatic variants affecting the *PAX5* CREs, known off-targets of AID in B-cell malignancies[214], were examined. Somatic mutations in CREs interacting with *PAX5* promoters showed both canonical AID (C > T/G in WRCY motifs with R = purine, Y = pyrimidine, W = A or T) and non-canonical AID (A > C/G in WA motifs)[215] mutational signatures (Figure 3.10), in agreement with *PAX5* enhancers mutated by AID in mouse B-cells and diffuse large B-cell lymphoma[214].

Table 3.13: Summary of novel findings from the study. *, these genes reside close to regions of common structural variation, making interpretation of their specific relevance problematic.

Novel genes disrupted in coding regions		Novel genes disrupted by mutations in non-coding regions		
Genes disrupted by structural variants	Genes disrupted by SNVs and indels	Promoters disrupted by SNVs	CREs disrupted by SNVs	CREs disrupted by CNVs
CD96	*BAX*	*NBPF1*	*CALCB*	*MYC*
PRDM1	*C8orf86*		*COBLL1*	*PLD4**
FBXW7	*FAM154B*		*HOXB3*	*KDM3B**
MAP3K14	*FTL*		*ST6GAL1*	*SP110**
CCND2	*HIST1H4H*		*PAX5*	*RAB36**
	LEMD2		*ATP13A2*	*PACS2**
	PABPC1		*TPRG1*	*TEX22**
	RPN1			
	RPS3A			
	SGPP1			
	TBC1D29			

Figure 3.9: Several key pathways in MM are disrupted by a range of mechanisms. Figure adapted from Manier *et al.*[2] and Kumar *et al.*[216]

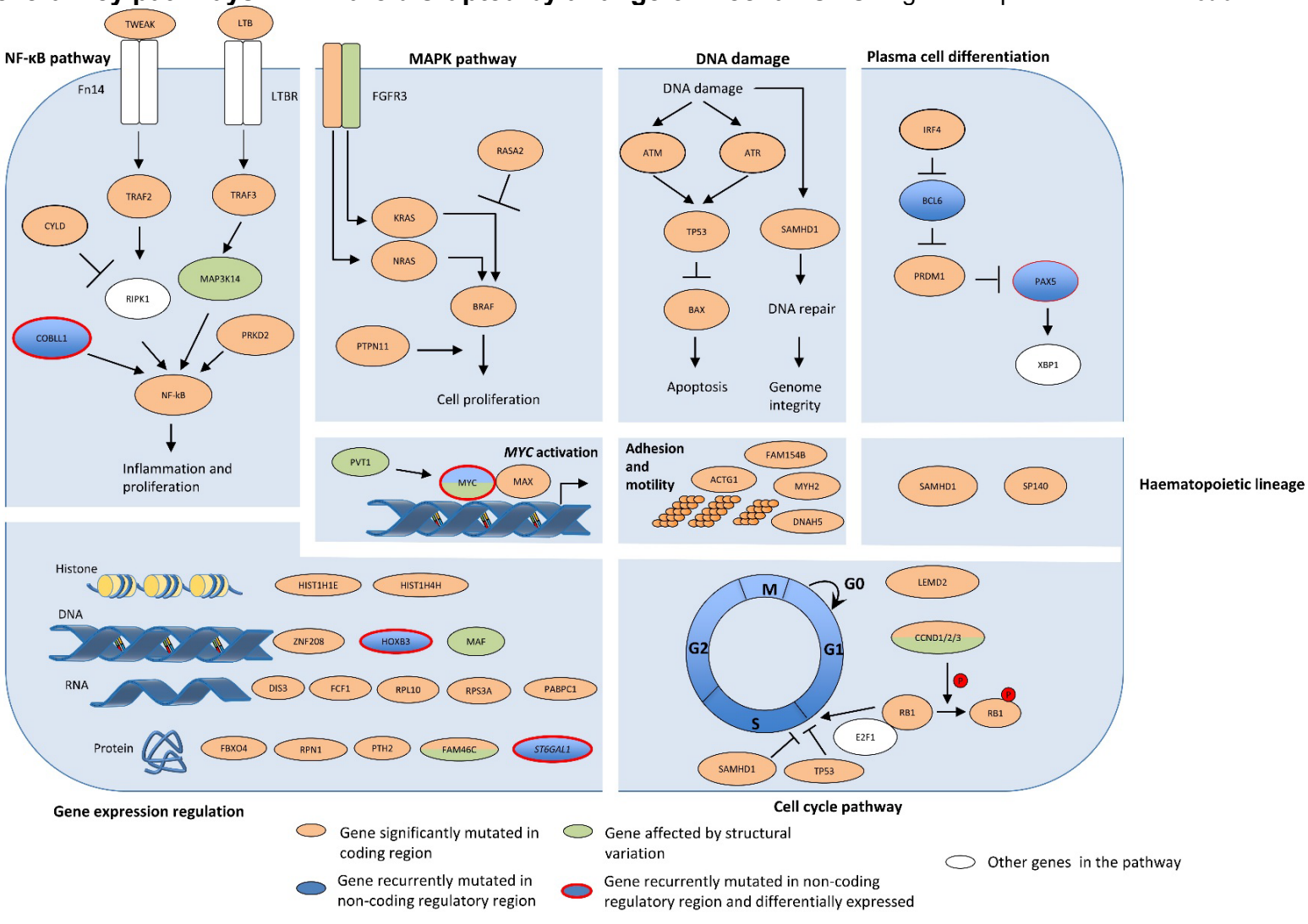
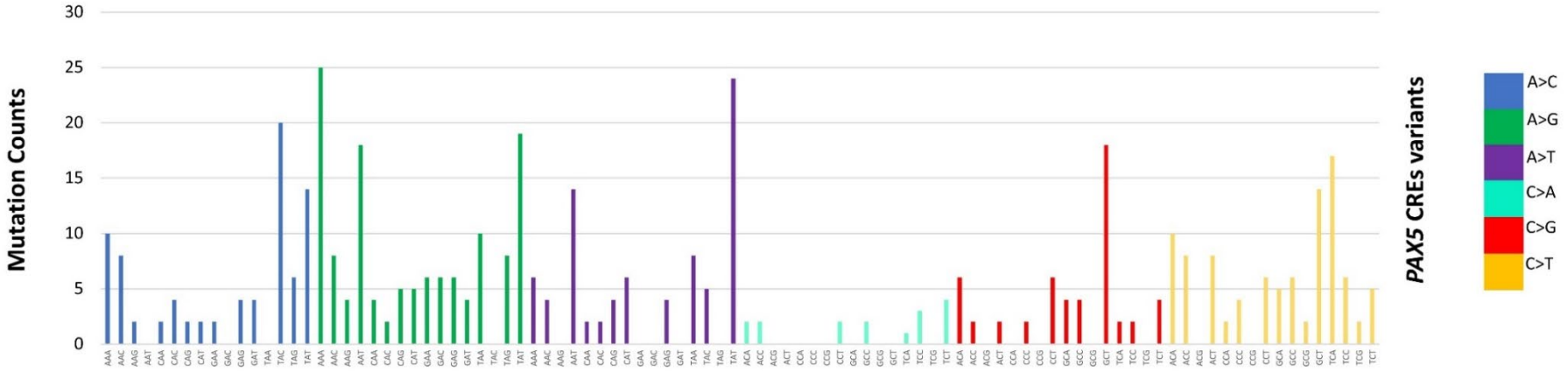


Figure 3.10: Mutational signatures in MM affecting *PAX5* CREs. Mutational patterns of somatic mutations in CREs interacting with *PAX5* promoters display both canonical (C > T/G in WRCY motifs with R = purine, Y = pyrimidine, W = A or T) and non-canonical (A > C/G in WA motifs) activation-induced cytidine deaminase (AID) signatures.



3.4 Discussion

This analysis has identified new coding and non-coding drivers as well as highlighting that pathways, key to the development of MM, can be targeted somatically through a range of mechanisms (Figure 3.9). Strikingly, although upregulation of *MYC* through gene amplification or translocation is well established in MM[2], it was demonstrated that *MYC* can be dysregulated by alternative mechanisms. These include CNVs altering *MYC* non-coding regulatory regions and specifically, the data implicates a region syntenic to the murine *Myc* enhancer cluster that has recently been reported to be essential for the maintenance of *MLL–AF9*-driven leukemia in mice[217].

The downregulation of tumour suppressors *PAX5*[201-203] and *HOXB3*[209] by CRE mutations in MM is entirely consistent with their decreased expression contributing to development and progression of MM as is the case with other B-cell malignancies. It has previously been demonstrated that disruption of the NF-κB pathway in MM can be the consequence of coding mutations and loss of genes. Here the study adds *TWEAK*, *TRAF2* and *PRKD2* to the list of genes disrupted via coding mutations, demonstrates *COBLL1* as dysregulated via mutations of a non-coding regulatory region, and identifies *MAP3K14* as upregulated via translocation to the IG loci[218].

Whilst utilizing WGS data facilitates the identification of signatures enriched in the non-coding genome it also, by nature of the low coverage data, focuses the analysis on early mutational processes. Accepting this limitation, I identified a number of mutational signatures previously unreported in MM, and strikingly the AID-attributed signature 9 being detectable in a high proportion of MM, a finding consistent with a contemporaneous report[213]. Although mutational patterns suggestive of AID activity have been documented in certain genes in MM such as *EGR1*[3] and *CCND1*[5], the findings suggest that off-target AID activity could be more widespread than previously envisaged. Moreover, as off-target AID activity is associated with genomic instability and chromosomal translocation in B-cells[219], it may be a major aetiological factor driving mutation of MM.

It should be, however, acknowledge that the present analysis has limitations. Firstly, a cellular model of naïve B-cells was used to map the CREs, which is

unlikely to fully and specifically recapitulate the spectrum of pathogenic SNVs and CNVs seen in MM. Secondly, the low coverage of CoMMpass WGS data means that the data have likely underestimated the somatic variants in the tumours, and increased noise to gene expression analysis. The sensitivity of the analysis is dependent on the clonal architecture of the samples, and it is likely that this analysis is limited to the identification of clonal, early drivers of MM. Thirdly, inevitably as CNVs are highly recurrent in MM[2], this has restricted the study power of the gene expression analysis as samples were excluded. Lastly, non-coding RNAs were not considered in gene expression analysis although many have been identified as recurrently mutated in their regulatory regions. Despite the restricted sensitivity, I have identified multiple targets of non-coding mutations, highlighting the importance of broadening the search for cancer drivers into the regulatory genome. Validation of the candidates that were identified in this study will be contingent on functional studies including, for example, CRISPR-mediated genome editing, *in vitro* reporter assays, and proliferation assays coupled with transcriptional profiling.

In conclusion, the findings provide integrated analysis of novel coding and non-coding drivers in MM, demonstrating the genetic complexity contributing to this malignancy. Thus by developing a more comprehensive picture of the underlying genetic basis of MM, I extend the list of genes and pathways for which novel therapeutic agents may be identified through network-based drug search methodologies[220, 221], offering the prospect of future individualized therapy in MM.

CHAPTER 4 Mutational processes contributing to the development of multiple myeloma

4.1 Overview and rationale

Cancers have variable numbers of somatic mutations that have accumulated during the life history of the tumours as a consequence of diverse cellular processes, including defective DNA replication or DNA repair, and exposure to endogenous or exogenous DNA-damaging agents[10, 87]. Each of these processes results in mutational signatures, which can serve as proxy for the cellular processes that have gone amiss. Mathematical deconvolution[88] of these mutational signatures in large pan-cancer series has revealed multiple distinct signatures[87], several of which are associated with known aetiologies, but many remain unexplained[87, 90, 222]. Hence studying the mutational signatures of cancers provides a mechanism for gaining insight into the aetiological basis of tumour development.

Whole-exome and whole-genome sequencing studies from my study (Chapter 3) and others have so far identified over 40 driver genes that are recurrently altered in MM[1-5]. However, the molecular mechanisms giving rise to these mutations are yet to be fully elucidated.

Here I report a comprehensive analysis of the mutation signatures of over 800 MM genomes. Major mutational signatures in MM reflective of three known principle mutational processes were identified: aging[87, 92, 223], DNA repair deficiency[87, 166, 223-227], and AID/APOBEC activity[87, 166, 201, 228]. These mutational signatures tend to show subgroup specificity and are reflective of the molecular mechanisms involved in tumorigenesis. Additionally, this study shows that information on mutational signatures beyond that associated with APOBEC has relevance to predicting patient prognosis and defining high-risk MM.

4.2 Study design

4.2.1 Samples and dataset

All data analysed in this chapter were generated as part of the MMRF CoMMpass Study release IA10. WGS data on 850 matched tumour-normal baseline newly diagnosed bone-marrow samples were downloaded from dbGaP as detailed in section 2.1.1**.** WES variants (detected by at least two out of three variant callers – MuTect, Seurat, and Strelka) from 874 samples, RNA-seq, CNV, clinical data, and Seq-FISH data (MMRF IA10 dataset) were downloaded from MMRF web portal (https://research.themmrf.org/) (section 2.1.1). WES and WGS data were available for 824 samples.

4.2.2 Statistical and bioinformatics analysis

Quality control, sequence alignment to hg37, and variant calling performed using FastQC v.0.11.4/BWA v0.7.12/GATK/Mutect v1.1.7 software as described in section 2.2.4. Somatic SNVs were filtered for oxidation artefacts[158] and by quality score as detailed in section 2.2.7.1. Mutations mapping to immune hypermutated regions (429 immunoglobulin and the major histocompatibility complex loci, each region extended by 50 Kb, as defined in Ensembl v73)[174], were excluded to avoid bias from mutation as a consequence of normal B-cell development.

4.2.2.1 Determination of myeloma karyotype

Translocation status of MM tumours was based on Seq-FISH[131] (section 2.1.1). HD was defined as amplification of 90% of the chromosome in at least two autosomes[4]. Prognostic chromosome-arm events (>1 Mb) were defined as deleted or amplified with $\text{abs}(\log_2\text{ratio}) \geq 0.1613$ occurring at 1p12, 1p32.3, 1q21.1, 1q23.3, and 17p13[2].

4.2.2.2 Mutational signatures

Characterisation of the 30 COSMIC mutational signatures and *de novo* extraction of signatures was performed using Palimpsest[92, 179] with default parameters (section 2.2.10.2). *De novo* mutational signatures were compared with 30 pre-defined COSMIC signatures by computing their cosine similarities[87]. A *de novo* mutational signature was assigned to a COSMIC signature if the cosine similarity was > 0.75 as previously advocated[92]. If multiple COSMIC signatures passed this threshold, then the most-similar COSMIC signature was assigned to the *de novo* signature. Proportion of COSMIC mutational signatures was compared between high-coverage WES clonal mutations (alternate allele ratio > 0.9) and low-coverage WGS mutations restricted to exome regions; as well as between CoMMpass exome and Walker *et al.*[1] exome mutations. Correlations were tested using Spearman's correlation. For those signatures with an apparent flat profile these were considered in concert, by combining the respective contributions of signatures 3, 5 and 8.

MANTA was used to identify somatic structural variants (SVs) from the WGS data adopting default settings[162] (section 2.2.7.3). The same statistical framework used for signature analysis of SVs implemented in Palimpsest[179] was applied to extract *de novo* rearrangement signatures (as previously described in sections 1.3.2 and 2.2.10.2)[92]. Correlations between SV signatures and major COSMIC pre-defined SNV signatures (>1% mutational contribution in WGS) were tested using Spearman's correlation. No significant correlation was seen after adjusting for multiple testing (*i.e.* $Q > 0.05$).

The relationship between mutational signatures and clinico-pathological parameters was examined confining the analysis to the major MM subgroups - HD, t(4;14), t(11;14), t(14;16), t(14;20) and t(8;14) *MYC*. Test of association between each signature and subgroups was based on a two-tailed Fisher's exact test using Benjamini-Hochberg FDR procedure to address multiple testing.

Contribution of each mutational signature to coding and non-coding regions was compared using WGS data. To calculate contribution of a mutational signature to a genomic region, first the probability that each mutation was due to the process underlying each signature was estimated and the cumulative probability of all

mutations in each region was calculated, as *per* Letouze *et al*.[92] After computing these probabilities, the regional differences in trinucleotide composition were normalised as detailed in section 2.2.10.3.

4.2.2.3 Replication timing and replication strand bias

Replication sequencing (Repli-seq) data generated by the ENCODE consortium for the lymphoblast cell lines with GM12878, GM06990, GM12801, GM12812, and GM12813 were used to define early and late-replicating regions; as well as leading and lagging DNA strands using Repli-Seq signal peaks from GM12801 as previously described[92, 223]. Mutation rates across deciles of replication timings were estimated globally using WGS data and for each signature, with each mutation assigned to a single signature by Palimpsest[92, 179]. The replication timing slope was estimated by linear regression model. To test the null hypothesis that the slope gradients equal zero, the replication timing deciles were permuted 10,000 times. Empirical *P*-values were calculated as the fraction of permutations with absolute slope values at least as great as the absolute slope value computed using the true replication timing deciles.

Analysis of mutational replication strand bias between leading and lagging strands was performed across all 30 COSMIC signatures as previously described[92], using WGS data. The Wilcoxon rank-sum test was used to determine significant difference of mutational contribution from each COSMIC signature between leading and lagging strands. Levels of asymmetry were considered significant if strand imbalances were > 30%[223] and $Q < 0.05$.

4.2.2.4 Transcriptional levels and strand bias

To correlate mutational processes with gene expression, RNA-seq data were normalised to FPKM (fragments per kilobase of exons per million reads)[92]. For each tumour, genes were partitioned into pentiles based on respective FPKM. Immunoglobulin-related genes and genes known to be highly upregulated in MM as a result of translocations (*CCND1*, *CCND3*, *FGFR3*, *MMSET*, *MAF*, *MAFB*, and *MYC*)[2] were excluded to mitigate against bias. Mutation rates of genes within

each of the 5 transcriptional level categories were estimated per tumour based on WES called mutations. Average alignability score for highly expressed genes was based on alignability of 75mers defined by the ENCODE/CRG GEM mappability tool[159]. Mutation rates were examined on transcribed and non-transcribed strands globally and for each signature as described previously[92] using Palimpsest[92, 179]. Wilcoxon rank-sum tests, corrected for multiple testing, were used to determine significant difference of mutational contribution from each COSMIC signature between transcribed and non-transcribed strands. Levels of asymmetry were again considered significant if strand imbalances were > 30%[223] and $Q < 0.05$.

4.2.2.5 Kataegis

Kataegis analysis was restricted to high-coverage WES data, where there was sufficient coverage to detect local hypermutation. Kataegis foci were defined and identified as detailed in section 2.2.7.4. Co-localization of kataegis and structural rearrangements was assessed based on the proportion of SV regions having kataegis foci residing within 10 Kb. To examine enrichment of a mutational signature at kataegis regions, mutational contribution of each signature was compared across all mutations at kataegis foci with other mutations in tumours with and without kataegis being detected using Wilcoxon rank-sum test, corrected for multiple testing and imposed a threshold of $Q < 0.05$.

4.2.2.6 Association of mutational signatures with the mutation of driver genes

For SNV mutational signatures, Wilcoxon rank-sum tests were used to compare contribution of each mutational signature in coding drivers[1, 3-5] and other exonic mutations, normalising for trinucleotide composition as described in section 2.2.10.3. For each somatic mutation, the probability that it was the result of each mutational process was estimated considering the tri-nucleotide context and the number of mutations attributed to each process in the respective tumour as per Letouze *et al*[92]. I then compared, for each driver gene and mutational signature,

the probability distribution in mutations affecting the driver gene as compared to all other mutations in tumours with and without the driver gene mutated using Wilcoxon rank-sum tests, imposing Benjamini-Hochberg correction for multiple testing. All driver genes identified in chapter 3 and previous studies[1, 3-5] were evaluated with $Q < 0.05$.

4.2.2.7 Association of signatures with clinical features

Multivariable Cox-regression was performed to adjust for covariates including age at diagnosis, sex, translocation status, and APOBEC mutational contribution (COSMIC signature 2 and 13). The ConsensusClusterPlus R package[229] was used to hierarchically cluster patients based on *de novo* SV and major COSMIC SNV signatures (> 1% contribution) extracted from WGS with default settings[91]. Fisher's exact test was used to test whether clusters were associated with MM subgroups or driver gene mutations, imposing Benjamini-Hochberg correction for multiple testing. The log-rank test was used to assess the differences in progression free survival (PFS) and overall survival (OS) between all cluster groups. To delineate clusters into low- and high-risk groups, pairwise comparisons in survival distributions were performed using the pairwise_survdiff function implemented in the survminer R package[61].

Multivariable Cox-regression was performed for each subgroup versus other subgroups, adjusting for age at diagnosis, sex, translocation status, APOBEC contribution, 1p deletion, 1q gain, 17p deletion, and *TP53* non-synonymous mutations.

4.3 Results

4.3.1 Genome sequencing of multiple myeloma

To examine the diversity of mutational signatures, I analysed overlapping WGS and WES data on 850 and 874 MM tumour-normal pairs respectively, generated by CoMMpass (IA10 release). The frequency of the MM major subgroups – HD, t(11;14), t(4;14), t(14;16), t(14;20) and t(8;14) *MYC*-translocation - is similar to other unselected series of patients that have been reported from CoMMpass IA9 dataset[2] (Chapter 3) (Table 4.1). The high-coverage WES data (120-150×, 136,074 SNVs) were used to analyse coding regions and the low-coverage WGS data (6-12×, 1,348,881 SNVs and 44,155 SVs) to provide genome-wide insights into clonal mutations associated with early processes underlying tumorigenesis[4, 197].

Table 4.1: CoMMpass IA10 karyotype classification (n = 814). Karyotypes data were only available for 814 samples. *, published literature was based on Manier *et al.*[2]

Subgroups	Number of samples	IA10 percentage	Published literature*
t(11;14)	160	19.7%	15-20%
t(4;14)	102	12.5%	15%
t(14;16)	32	3.9%	5%
t(6;14)	15	1.8%	1-2%
t(14;20)	10	1.2%	1%
t(8;14) *MYC*-translocation	120	14.7%	15-20%
Hyperdiploidy	469	57.6%	50%

4.3.2 Mutational signatures in multiple myeloma

Application of the NMF framework[92] (Figure 4.1) to extract *de novo* SNV mutational signatures did not identify any novel mutational signatures (Figure 4.2, Figure 4.3), consistent with a recent analysis on CoMMpass exome dataset[230]. Overall a total 9 of the 30 mutational signatures referenced by COSMIC were seen at >1% mutational contribution in the WGS data (Table 4.2) - signature 1 related to aging[87]; 2 and 13 to activity of the APOBEC family of cytidine deaminases[87, 166, 228]; 9 to polymerase η implicated with the activity of AID during somatic hypermutation[87, 201, 228]; signature 30 reflective of mismatch repair deficiency[227], and signature 16 which has as yet an unknown aetiology. I also extracted flat signatures, which cannot be unambiguously assigned to signatures 3, 5, or 8 in tumours but all are indicative of DNA repair deficiency (homologous recombination deficiency and nucleotide repair deficiency)[87, 166, 223-226].

However, five novel *de novo* structural RS were identified (Figure 4.4): RS1 (19% of SVs across samples) – characterised by non-clustered deletions, large-scale tandem duplications and inversions; RS2 (17%) – characterised by clustered translocations; RS3 (13%) – characterised by inversions; RS4 (21%) – characterised by non-clustered small-scale deletions and tandem duplications; RS5 (30%) – characterised by non-clustered translocations. The study therefore focussed on the 9 major SNV and 5 *de novo* SV mutational signatures for subsequent analyses.

Following on from this, the contributions of the 9 major COSMIC SNV mutational signatures in both WES and WGS dataset were examined. The signature profiles recovered from analysis of clonal WES and exome-restricted WGS data were highly correlated ($r = 1$, $P < 2.2 \times 10^{-16}$, Spearman's correlation, Figure 4.5). Hence, while the average sensitivity to detect clonal SNVs from the WGS data is 20-35%[4] (Chapter 3), these findings indicate the mutational signatures identified by WGS are valid and representative of early mutational processes in MM. A high concordance of mutational signature was also observed in WES data from CoMMpass and that reported by Walker *et al.*[1] ($r = 0.86$, $P = 0.014$, Spearman's correlation, Figure 4.6), reflecting the generalizability of the observations. No significant association between the major COSMIC SNV signatures and those associated with rearrangements was seen (Table 4.3).

Figure 4.1 Summary of mutational signatures extraction in the study. WES, Whole-exome sequencing. WGS, Whole-genome sequencing. SNV, single nucleotide variant. SV, structural variant. Figure adapted from Helleday *et al.*[10] :

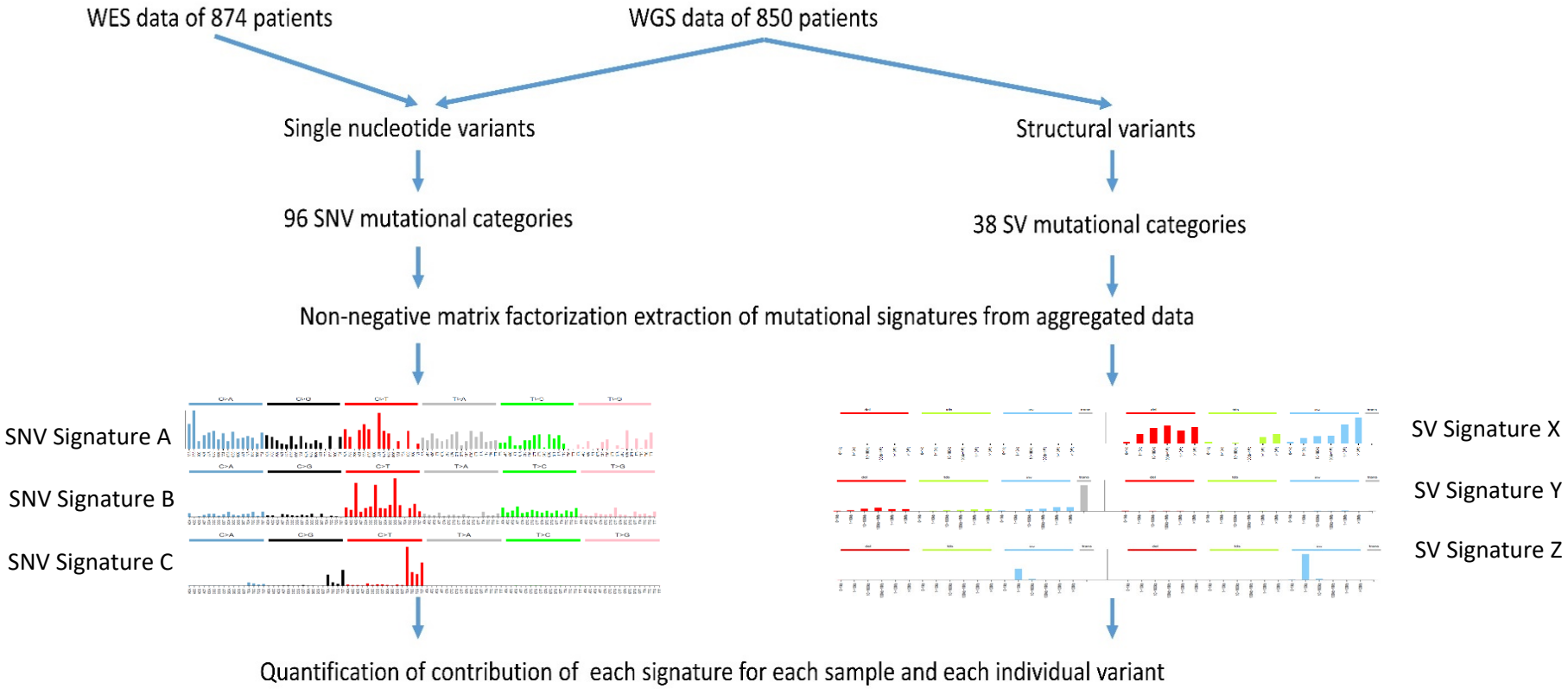


Figure 4.2: *De novo* extraction of WES single nucleotide variants signatures using non-negative matrix factorization algorithm. (a) Summary of three *de novo* mutational signatures extracted. (b) Cosine similarity heatmap. *De novo* extracted mutational signatures are compared against 30 COSMIC mutational signatures. The colour code (0 to 1) represents the resemblance between each pair of signatures. Signatures are grouped together by hierarchical clustering. Figures are generated using Palimpsest R package. NE, *de novo* exome signature.

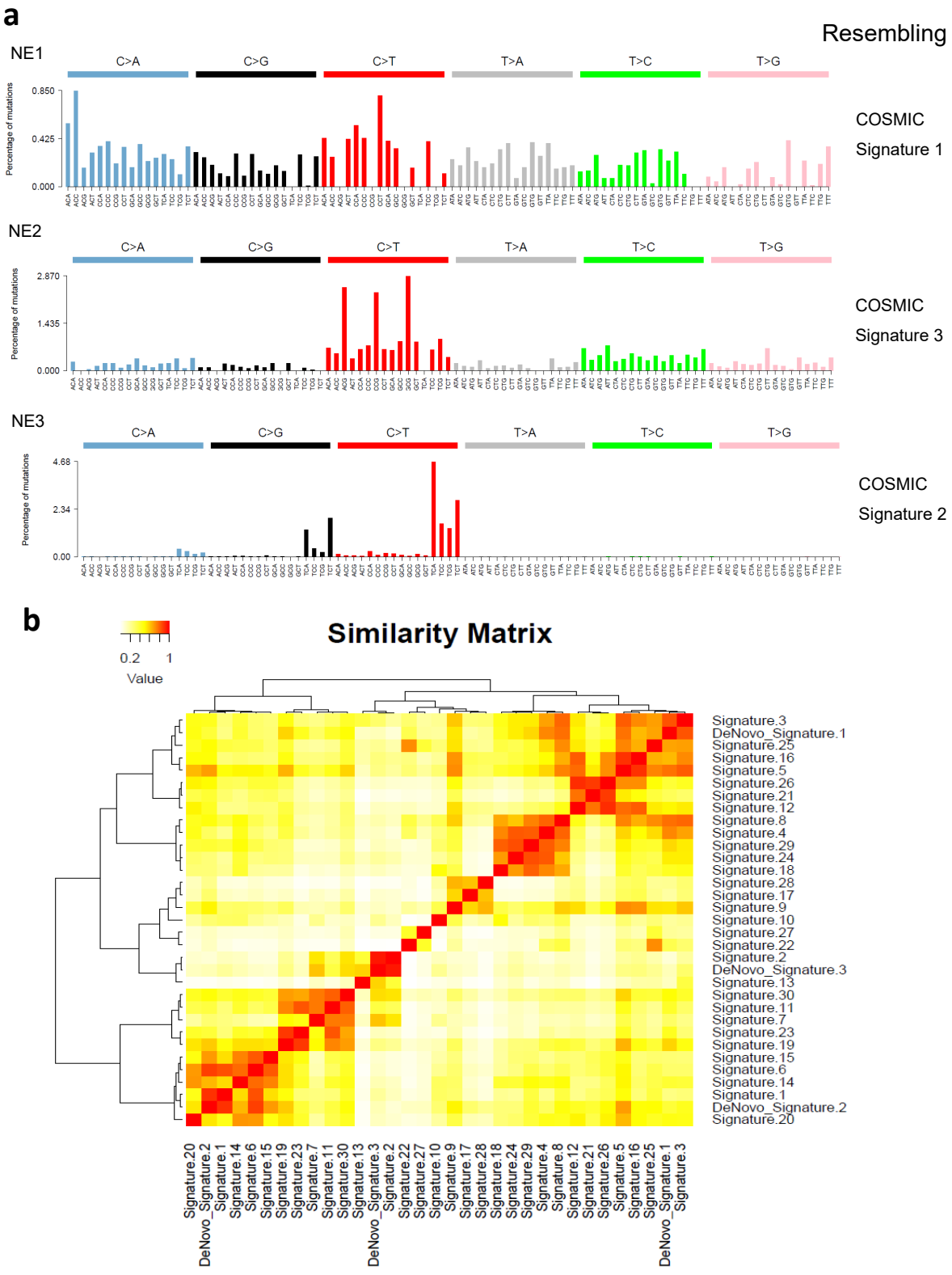


Figure 4.3: *De novo* extraction of WGS single nucleotide variants signatures using non-negative matrix factorization algorithm. (a) Summary of three de novo mutational signatures extracted. (b) Cosine similarity heatmap. *De novo* extracted mutational signatures are compared against 30 COSMIC mutational signatures. The colour code (0 to 1) represents the resemblance between each pair of signatures. Signatures are grouped together by hierarchical clustering. NG, de novo

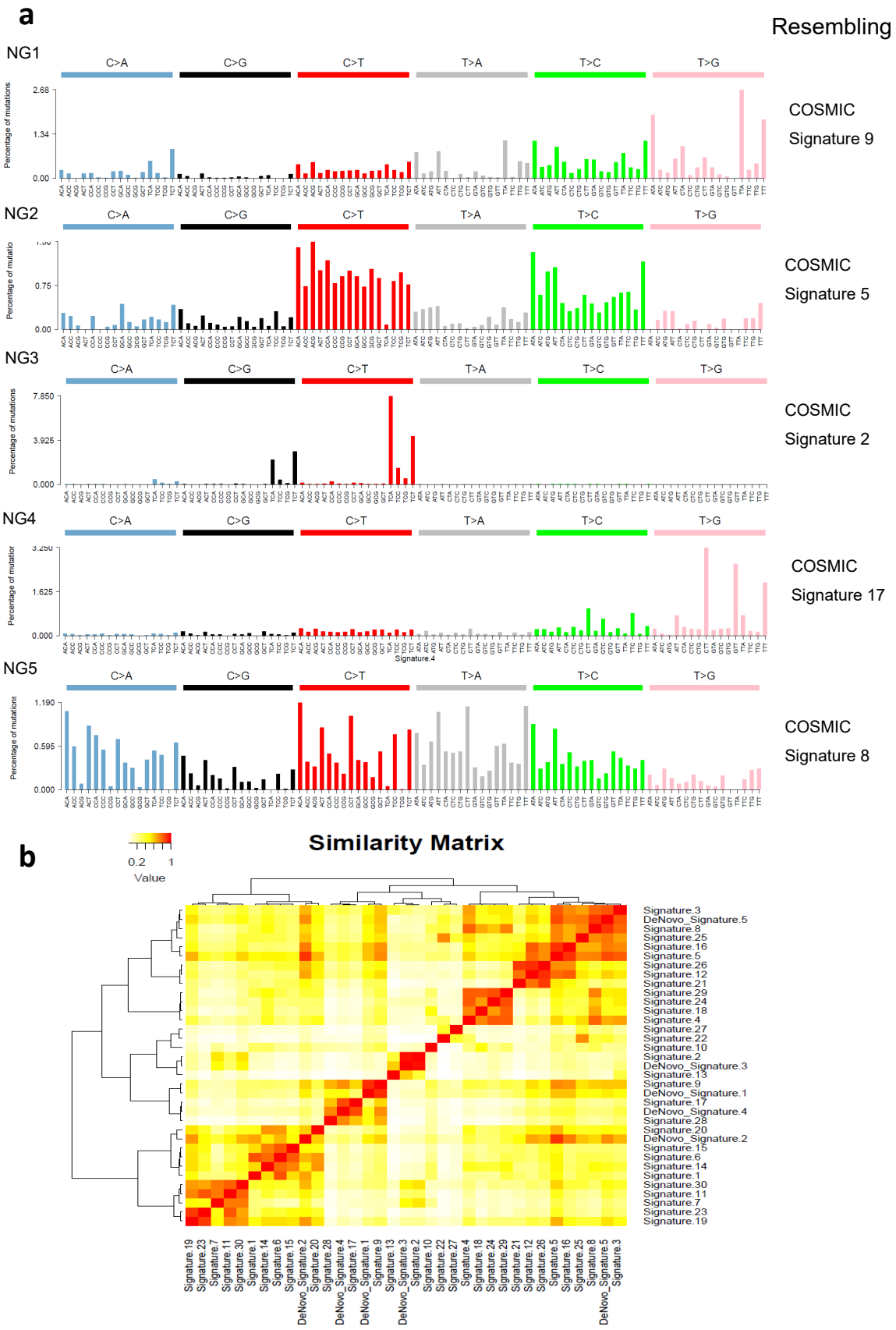
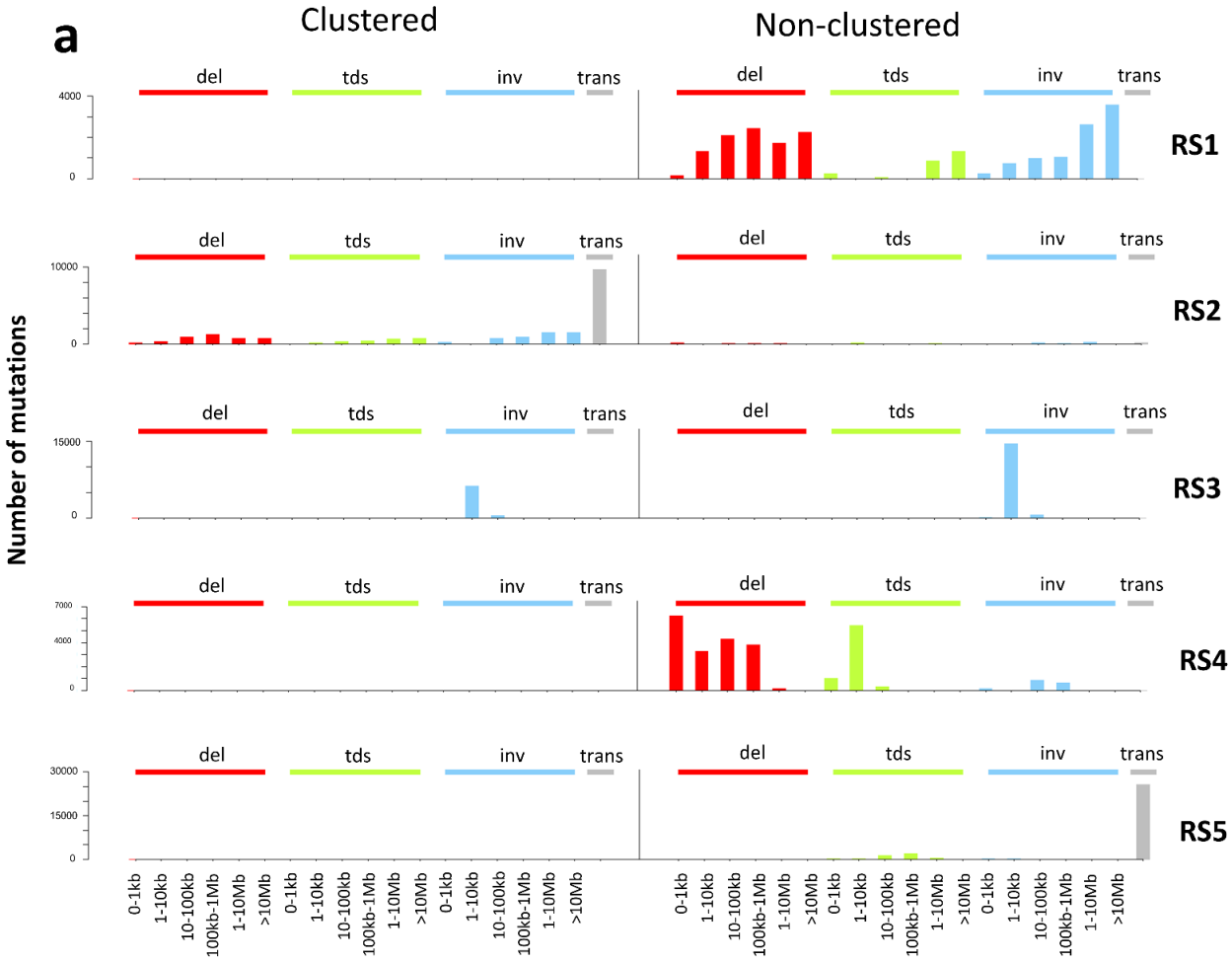


Table 4.2: COSMIC mutational contribution in WGS (n = 824). Both WGS and WES data were available for 824 tumours. In **bold**, these mutational signatures have > 1% mutational contributions.

COSMIC signatures	WGS contribution (%)
Signature 1	**2.199**
Signature 2	**4.803**
Signature 3	**7.015**
Signature 4	0.000
Signature 5	**7.782**
Signature 6	0.008
Signature 7	0.031
Signature 8	**9.654**
Signature 9	**45.975**
Signature 10	0.008
Signature 11	0.214
Signature 12	0.783
Signature 13	**1.000**
Signature 14	0.000
Signature 15	0.009
Signature 16	**11.466**
Signature 17	0.935
Signature 18	0.000
Signature 19	0.831
Signature 20	0.010
Signature 21	0.037
Signature 22	0.000
Signature 23	0.000
Signature 24	0.000
Signature 25	0.871
Signature 26	0.165
Signature 27	0.000
Signature 28	0.284
Signature 29	0.000
Signature 30	**5.918**

Figure 4.4: *De novo* structural rearrangements signatures. (a) Five *de novo* structural rearrangement signatures (RS) extracted in multiple myeloma. (b) Cumulative mutational contribution of the structural rearrangements across 850 WGS samples. Del, deletions; tds, tandem duplications; inv, inversions; trans, translocations.



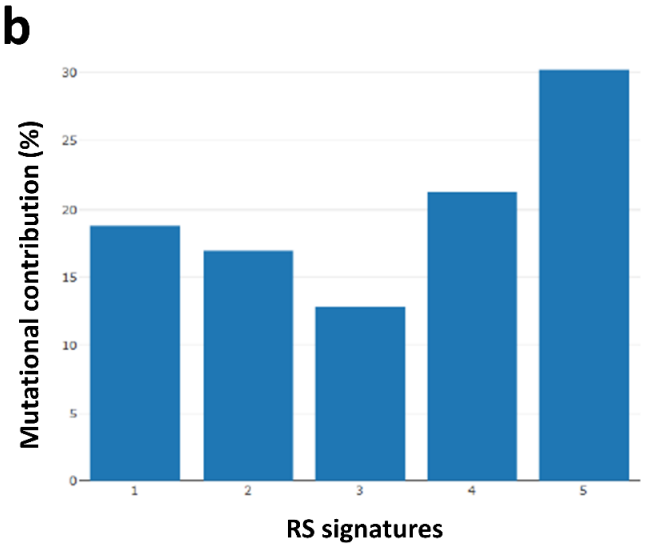


Figure 4.5: Concordance between clonal whole-exome and exome-restricted whole-genome single nucleotide variants mutational signatures (n = 525). (a) Cumulative mutational contributions of major COSMIC mutational signatures in clonal whole-exome sequencing (WES, blue) mutations and exome-restricted whole-genome sequencing (WGS, orange). (b) Scatter plot showing high concordance (Spearman's correlation) between mutational signatures identified in clonal WES mutations and exome-restricted WGS. Flat signatures include COSMIC signatures 3, 5, and 8. Sig, signature.

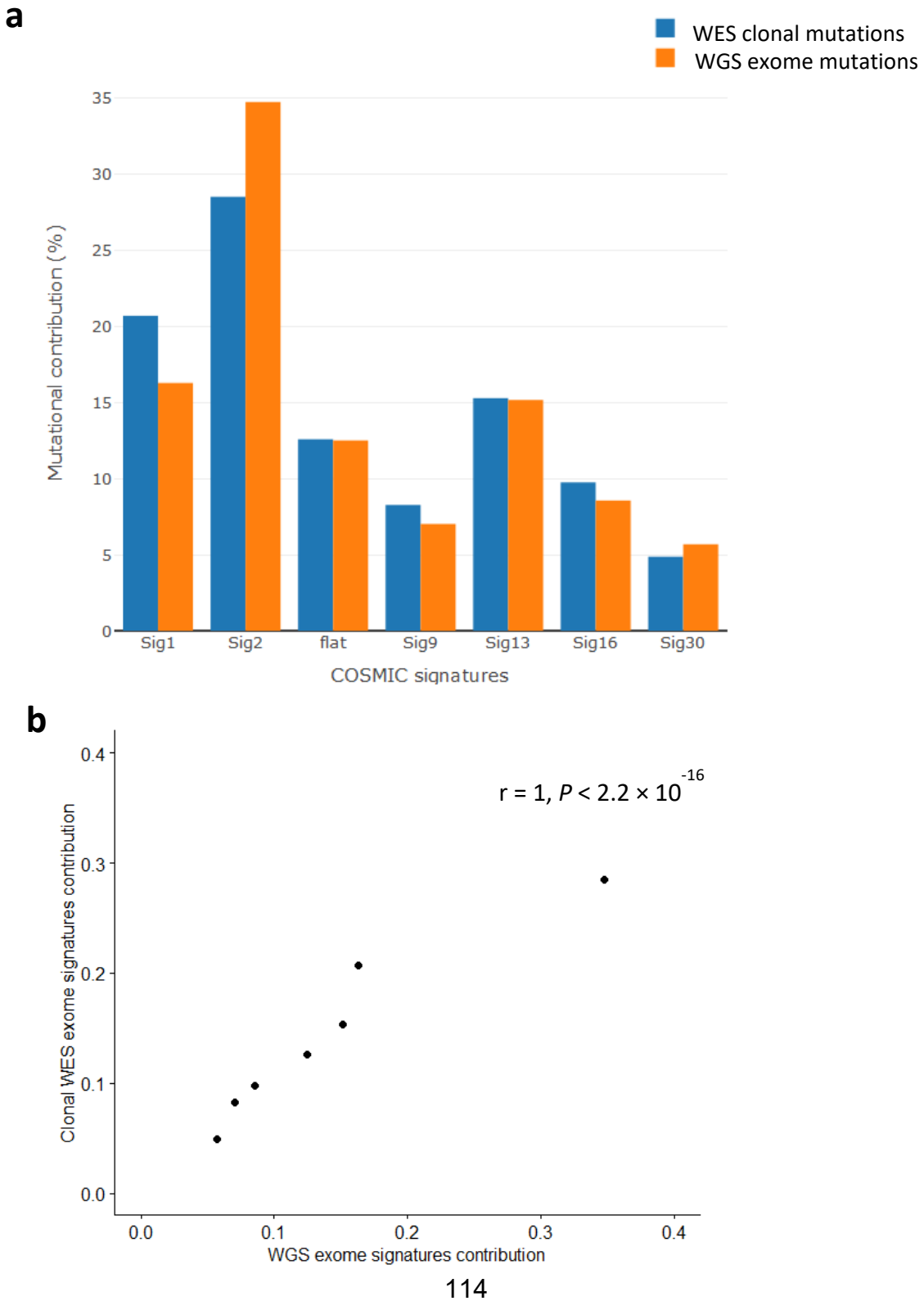


Figure 4.6: Concordance between CoMMpass and Walker *et al.*[2] exome single nucleotide variants mutational signatures. (a) Cumulative mutational contributions of major COSMIC mutational signatures in exome variants from CoMMpass (blue, n = 874) and Walker *et al.* exome study (orange, n = 463); and (b) Scatter plot showing high concordance (Spearman's correlation) between mutational signatures identified in CoMMpass and Walker's exome mutations. Flat signatures include COSMIC signatures 3, 5, and 8. Sig, signature.

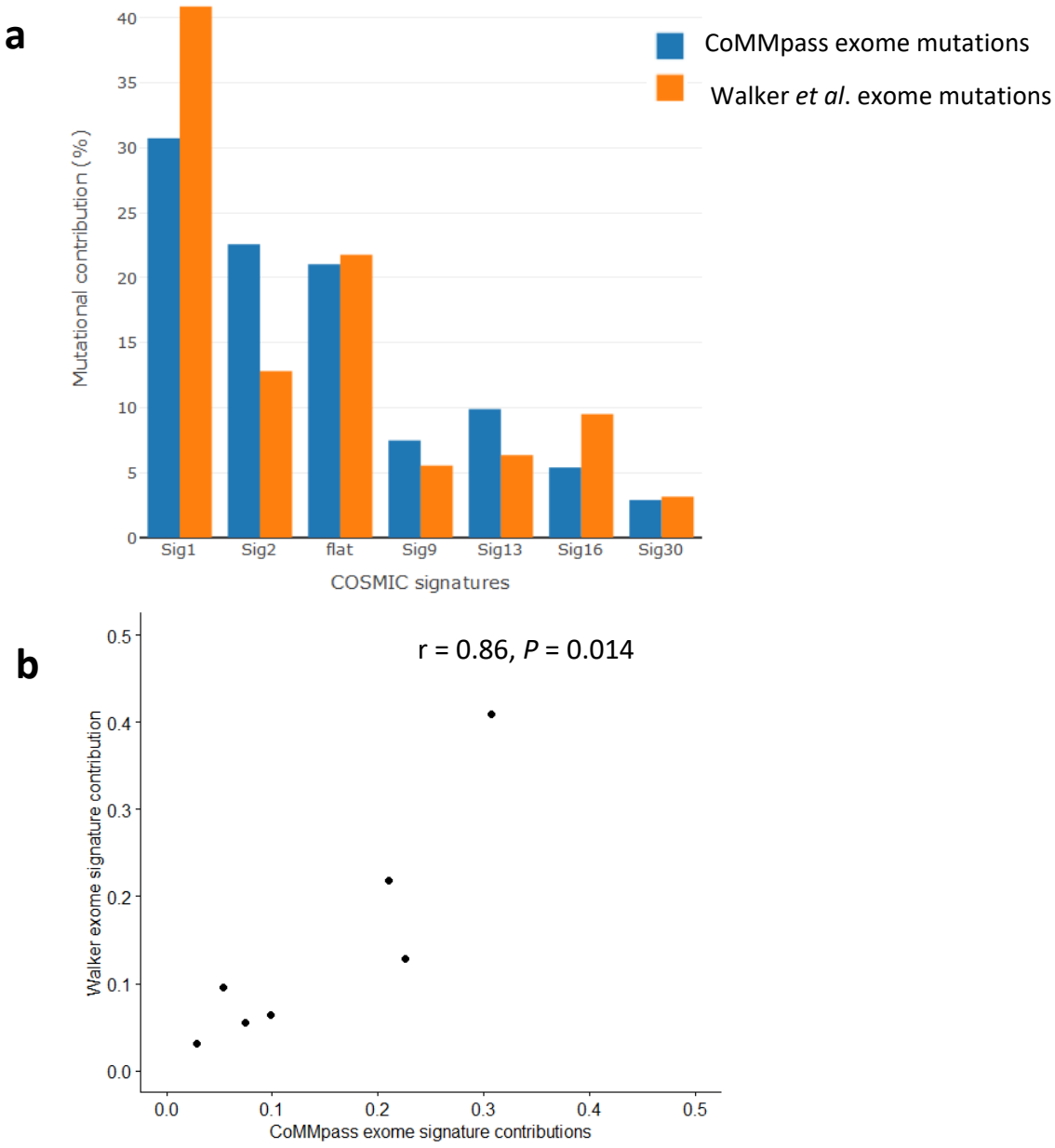


Table 4.3: Association between major COSMIC SNV and *de novo* SV signatures (Spearman's correlation *Q*-values).

Rearrangment signatures	Signature 1	Signature 2	Flat signatures	Signature 9	Signature 13	Signature 16	Signature 30
RS1	9.17E-01	6.26E-01	5.48E-01	6.26E-01	9.17E-01	9.83E-01	8.38E-01
RS2	9.83E-01	8.38E-01	9.83E-01	9.03E-01	5.48E-01	9.17E-01	6.58E-02
RS3	8.78E-01	8.39E-01	9.57E-01	8.39E-01	5.48E-01	8.39E-01	8.39E-01
RS4	9.57E-01	8.39E-01	5.48E-01	9.17E-01	5.48E-01	5.48E-01	7.41E-01
RS5	9.57E-01	5.48E-01	8.38E-01	5.48E-01	8.38E-01	8.89E-01	6.58E-02

4.3.3 Influence of DNA replication and transcription on mutational signatures

The impact of DNA replication and transcription on mutational signatures was broadly consistent with observations previously made in analyses of other cancers[92, 223, 231]. Specifically, an overall increased mutation rate in late-replicating regions was shown ($P < 1 \times 10^{-4}$) (Table 4.4**,** Figure 4.7a), with the exception of signature 13 having higher mutation rate in early-replicating regions ($P < 1 \times 10^{-4}$, Table 4.5, Figure 4.8), consistent with generalized replication time-dependent DNA damage mechanisms that operate in other cancers such as those of the breast[223] and liver[92]. The difference in how replication timing influences mutation rates in signatures 2 and 13, both of which are associated with APOBEC activity, suggests they are intrinsically different replication-linked mutational processes[223].

Similarly, as previously documented, strong replicative strand asymmetry (>30% imbalances)[223] was shown with respect to signatures 2 ($Q = 4.0 \times 10^{-16}$) and 13 ($Q = 4.0 \times 10^{-16}$) with higher mutation proportion in the lagging strand (Figure 4.7b). These findings are consistent with APOBEC activity primarily affecting lagging strands.

Overall, increased mutation rate was associated with increased transcription, suggesting the mutagenic role of the transcriptional process in MM (Figure 4.7c). This contrasts markedly to hepatocellular carcinoma[92], suggesting that transcription-associated mutagenesis may overwhelm transcription-coupled repair in MM[232]. Moreover, strikingly elevated mutation rates of both SNVs and indels were shown for highly expressed genes (Figure 4.7c). A number of these highly expressed genes (*i.e.* FPKM > 100), which are also frequently mutated, including *EGR1*[233], *XBP1*[234], *BTG2*[235], *DDX5*[236], and *NFKBIA*[5], have well-established roles in plasma cell differentiation and MM. The strong replicative, but weak transcriptional mutational asymmetry (Figure 4.7d) seen in MM is consistent with the mutual exclusivity trend of replicative and transcriptional asymmetries shown in many cancers[231].

Table 4.4: Mutation rate (SNV mutations/Mb) and DNA replication time

DNA replication time deciles	WGS mutation rate (mutations/Mb)	WES mutation rate (mutations/Mb)
1 (Earliest)	0.214	0.518
2	0.237	0.440
3	0.278	0.445
4	0.292	0.489
5	0.330	0.524
6	0.377	0.515
7	0.479	0.599
8	0.682	0.666
9	0.868	0.834
10 (Latest)	0.901	1.141
Slope (mutations/decile)	80	59
***P*-value**	<1.0E-4	<1.0E-4

Table 4.5: Major COSMIC mutational signatures and DNA replication time.
Flat signatures include COSMIC signatures 3, 5, and 8.

COSMIC signatures	WGS		WES	
	Slope (mutations/decile)	*Q*-values	Slope (mutations/decile)	*Q*-values
Signature 1	0.433	<1.0E-4	13.191	<1.0E-4
Signature 2	5.426	<1.0E-4	3.717	<1.0E-4
Flat signatures	12.887	<1.0E-4	15.548	<1.0E-4
Signature 9	52.430	<1.0E-4	7.037	<1.0E-4
Signature 13	0.012	<1.0E-4	-2.179	<1.0E-4
Signature 16	3.529	<1.0E-4	4.117	<1.0E-4
Signature 30	3.780	<1.0E-4	1.520	<1.0E-4

Figure 4.7: Relationship between replication and transcription in mutational processes. (a) Mutation rates across different DNA replication timing bins for SNVs. WGS mutation rate (blue) was estimated from low-coverage WGS data (6–12×). WES mutation rate (orange) was estimated from high-coverage WES data (120-150×) with variants called by at least two variant callers (b) Proportion of mutations on leading and lagging strands per signature based on WGS data. Asterisks indicate significant asymmetry ($Q < 0.05$ and strand imbalances >30%). (c) Relationship between transcriptional level and mutation rate. The range of number of genes across all samples included in each FPKM category (from low to high gene expression) are category 1: 4062 - 6800 (median 4209); category 2: 1323 - 4062 (median 3914); category 3: 4060 - 4062 (median 4061); category 4: 4060 - 4061 (median 4061); category 5: 4062. Error bars represent the 95% confidence intervals (d) Proportion of mutations on transcribed and non-transcribed strands across major signatures based on WES data. WGS, whole-genome sequencing; WES, whole-exome sequencing; SNVs, single nucleotide variants. FPKM, fragments per

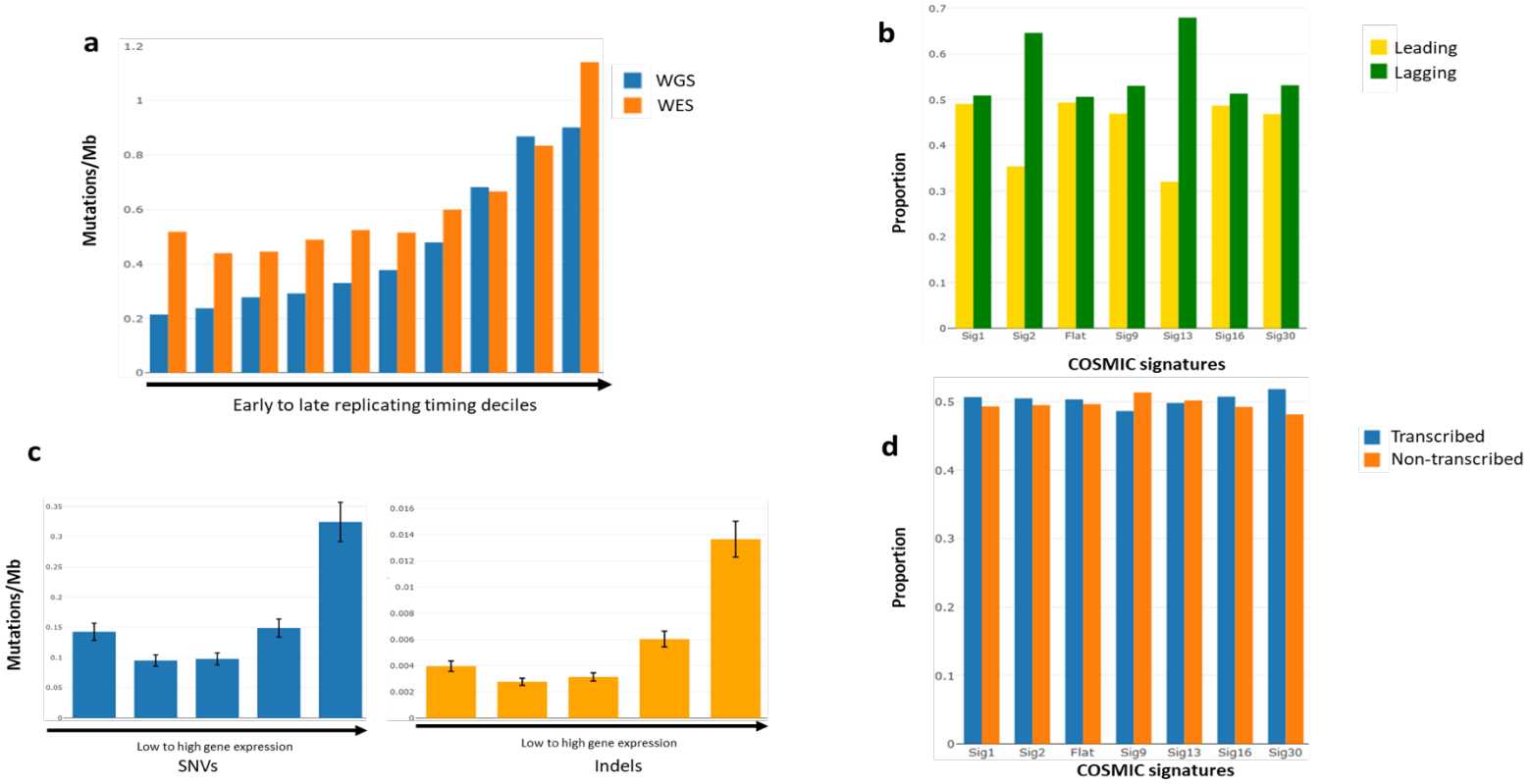
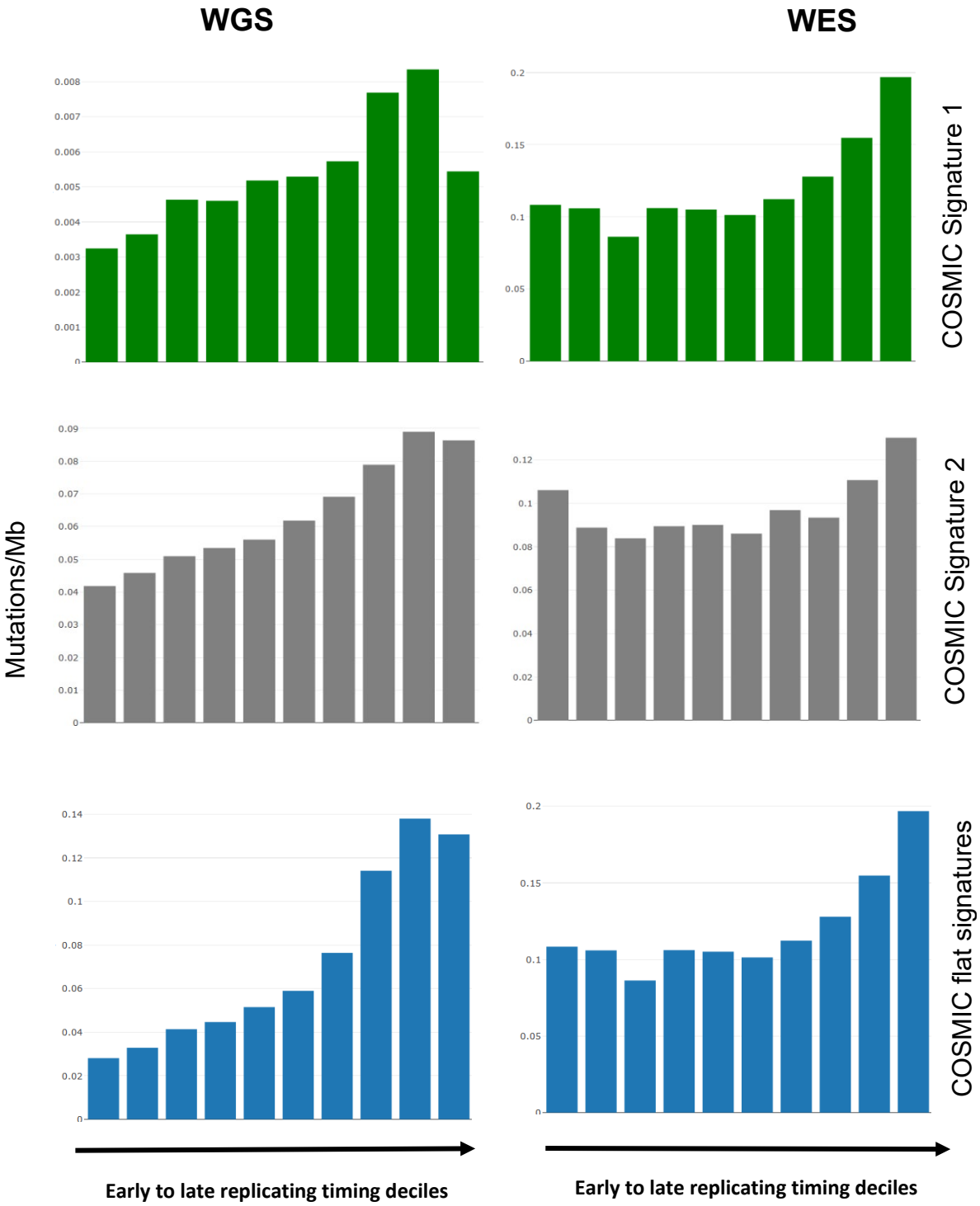
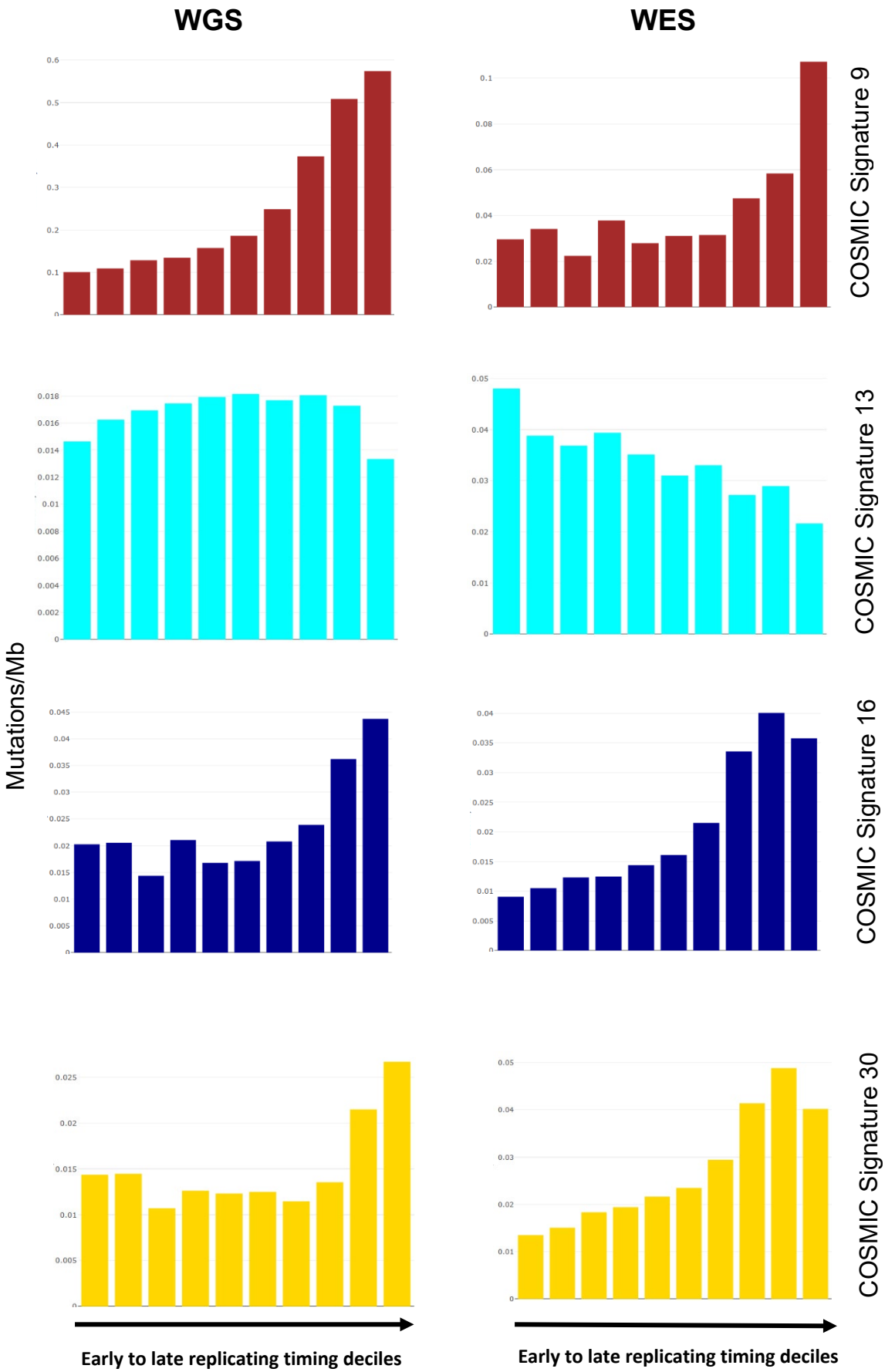


Figure 4.8: Correlation between DNA replication timing and SNV mutation rates per major COSMIC signatures. Flat signatures include COSMIC signatures 3, 5, and 8.




WGS
WES
COSMIC Signature 9
COSMIC Signature 13
COSMIC Signature 16
COSMIC Signature 30
Mutations/Mb
Early to late replicating timing deciles
Early to late replicating timing deciles

4.3.4 Mutational signatures in coding and non-coding regions

A significant difference in all mutational signatures within coding and non-coding regions was shown (Figure 4.9), implying different genomic regions are subject to specific mutational processes, consistent with earlier observations[213]. AID-attributed signature 9 predominates in non-coding regions, whereas exonic mutations are dominated by signatures 1, 2, 13 implicating aging and APOBEC signatures as important.

4.3.5 Relationship between mutational signatures and kataegis

Local hypermutated regions of tumour genomes, or kataegis, has been observed in MM[5, 237] and other B-cell malignancies[87]. I examined COSMIC mutational signatures contributing to kataegis (defined on the basis of average inter-mutation distance ≤ 1 Kb[88, 91]; Table 4.6), which were detected in 9% of samples (71/874). I did not observe significant and consistent enrichment of mutational contribution at kataegis foci compared to other mutations in tumours with and without kataegis detected (Table 4.7). I identified 70 genes disrupted by kataegis (Table 4.8), including *CCND1, CCND3, MAF*, and *FZD2* which are often affected by chromosomal rearrangements[2, 61]. Globally, 62% of kataegis foci co-localize with 5% of somatic structural arrangement sites (Figure 4.10), consistent with previous finding that most genomic rearrangements do not feature kataegis in nearby regions[87].

4.3.6 Mutational signatures and myeloma subgroups

Significant association between specific mutational signatures and MM subgroups was observed (Table 4.9). Signature 1 was enriched in HD MM ($Q = 3.2 \times 10^{-4}$) consistent with the correlation between age and frequency of HD[238] (Table 4.9). APOBEC-attributed signatures 2 and 13 were enriched in *MAF*-translocation subgroups - t(14;16) ($Q = 1.7 \times 10^{-15}$ and $Q = 3.5 \times 10^{-19}$ respectively), t(14;20) ($Q = 1.4 \times 10^{-3}$ and $Q = 6.4 \times 10^{-6}$ respectively) - and to a lesser extent in t(4;14) (only signature 2, $Q = 9.3 \times 10^{-6}$) consistent with previous reports[4, 84]. Flat COSMIC signatures, attributable to DNA repair deficiency, were

enriched in t(11;14) MM (Q = 3.3 × 10^{-4}). An enrichment of non-clustered deletions, large-scale tandem duplications, and inversions RS1 (Q = 3.8 × 10^{-6}); and clustered translocation RS2 (Q = 0.010) signatures was observed in t(4;14) MM (Table 4.10). Although speculative it is possible that the t(4;14) translocation, which leads to up-regulation of histone methyltransferase (MMSET), may affect genomic instability through some as yet undisclosed epigenetic mechanism.

The links between established prognostic mutational events (1p deletion, 1q gain, 17p deletion, and *TP53* mutations) with mutational signatures were further explored (Table 4.11). Associations between chromosome-arm events at 1p and 1q with COSMIC signatures 2, 13, and RS1 (Q < 0.05), and between *TP53* mutations tumours with RS1 (Q = 0.033) and RS2 (Q = 7.4 × 10^{-3}) raising the possibility of causal relationships.

Figure 4.9: Contribution of each single nucleotide variant mutational signature in coding (blue) and non-coding (orange) regions. Flat signatures include COSMIC signatures 3, 5, and 8. Sig, signature.

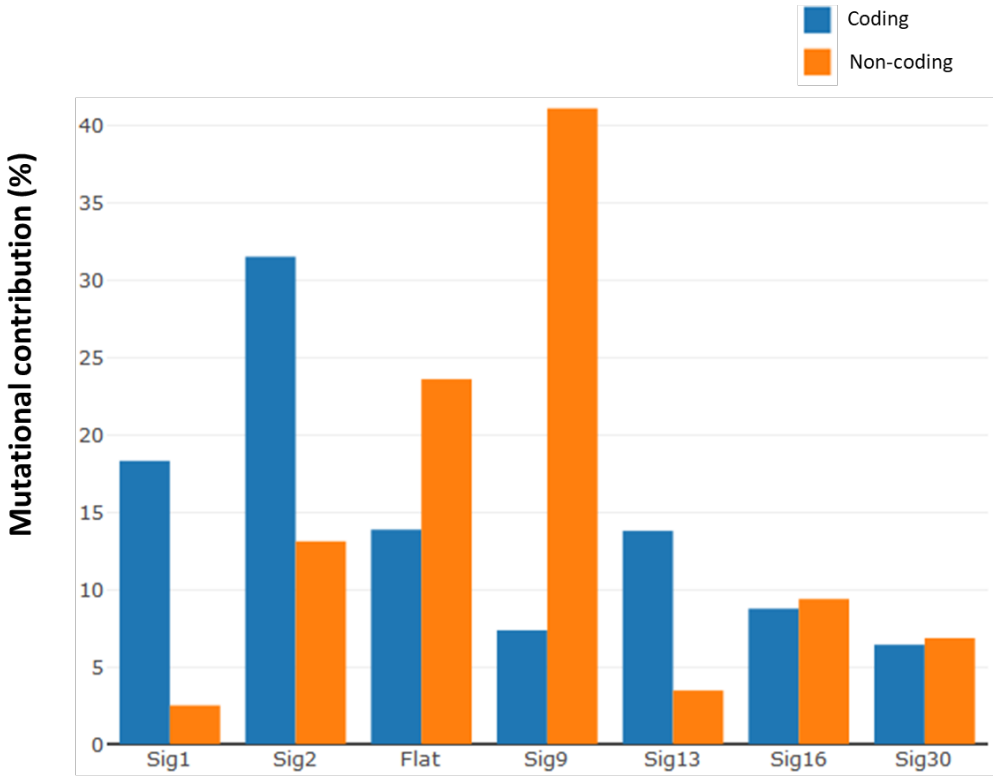


Table 4.6: Mutational contribution at exonic kataegis foci. Flat signatures include COSMIC signatures 3, 5, and 8.

COSMIC signatures	Mutational contribution (%)
Signature 1	19.635
Signature 2	27.892
Flat signatures	20.813
Signature 9	6.719
Signature 13	16.368
Signature 16	5.414
Signature 30	3.158

Table 4.7: Enrichment of mutational signatures at kataegis foci. COSMIC signatures contribution was compared at kataegis foci.

COSMIC Signature	Kataegis mutations mean	Other mutations in tumours without kataegis detected mean	*Q*-value	Other mutations in tumours with kataegis detected mean	*Q*-value
Signature 1	0.143	0.235	4.63E-15	0.151	7.09E-16
Signature 2	0.203	0.136	1.32E-19	0.325	4.32E-23
Flat signatures	0.151	0.164	8.53E-01	0.084	2.22E-36
Signature 9	0.049	0.058	8.53E-01	0.033	1.86E-12
Signature 13	0.119	0.062	5.21E-13	0.125	6.94E-22
Signature 16	0.039	0.042	3.33E-03	0.024	1.40E-27
Signature 30	0.023	0.021	8.53E-01	0.021	1.21E-01

Figure 4.10: Examples of kataegis plots (a) MMRF 1579: the kataegis focus on chromosome 1 and 22 detected co-localize with del 1p and an inversion on chromosome 12 respectively; (b) MMRF 2186: the kataegis foci on chromosome 11 co-localises with t(11;14) (q13;q32). Bolder arrows indicate regions with higher confidence being identified as kataegis.

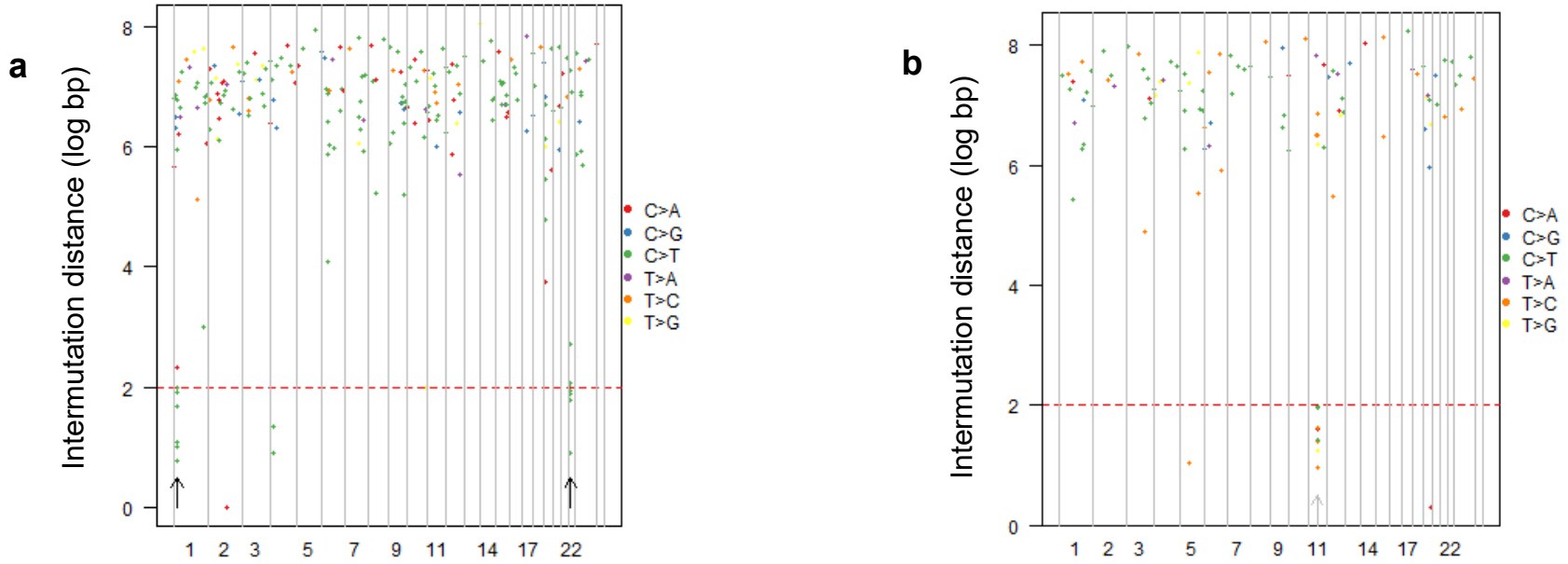


Table 4.8: Genes affected by kataegis and their frequency

Gene	Number of affected samples	Gene	Number of affected samples
BCL7A	3	*EGLN1*	1
CCND1	3	*WFDC9*	1
OR1S2	2	*ZNF292*	1
OR10G8	2	*SLC44A2*	1
NFKB2	2	*CLNK*	1
MAF	2	*RPS11*	1
CCND3	2	*hsa-mir-150*	1
WNT2	1	*C11orf74*	1
OR5AR1	1	*DTX1*	1
OSGIN2	1	*SHANK2*	1
RSPRY1	1	*PSD*	1
DEF8	1	*ZBTB39*	1
SPTLC2	1	*MERTK*	1
CLSTN3	1	*ZRANB3*	1
TTC40	1	*ERC1*	1
PRR14L	1	*AEN*	1
CREBRF	1	*COL1A1*	1
INPP4B	1	*PLEKHG1*	1
ZFP36L1	1	*FMNL1*	1
WNT5B	1	*FZD2*	1
RHCE	1	*SDK2*	1
IL17RA	1	*MYO1E*	1
ATG16L1	1	*FAM81A*	1
DOC2A	1	*BIRC3*	1
STAT5B	1	*VPS8*	1
RUNDC3A	1	*ARHGAP27*	1
BMP6	1	*CTDSP2*	1
MUC16	1	*SYBU*	1
TECPR2	1	*PPRC1*	1
NAV2	1	*ICK*	1
AKR1C1	1	*MAX*	1
AKR1C2	1	*NLRC5*	1
KIAA1456	1	*GAPVD1*	1
RNF150	1	*YIPF2*	1
TYMP	1	*C19orf52*	1

Table 4.9: Association of COSMIC mutational signatures in MM subgroups. (a) Summary statistics, in **bold**: significant values (*Q* < 0.05); and (b) summary of enrichment and the associated aetiologies. OR, odd ratio. Sig, Signature. *MYC*, t(8;14) *MYC*-translocation subgroup. HD, Hyperdiploid. NA, not available.

a

Subgroups	Sig1		Sig2		Flat signatures		Sig9		Sig13		Sig16		Sig30	
	Q- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- values	OR
HD	**3.24E-04**	**1.978**	**2.92E-05**	**0.476**	2.06E-01	0.734	2.01E-01	2.075	**1.25E-04**	**0.205**	4.15E-01	1.157	1.69E-01	1.285
t(11;14)	6.22E-01	1.147	**5.59E-05**	**0.357**	**3.32E-04**	**3.367**	2.24E-01	4.752	3.57E-01	0.516	3.57E-01	1.223	**3.80E-03**	**0.561**
t(4;14)	**3.75E-08**	**0.084**	**9.30E-06**	**2.944**	**3.26E-02**	**2.428**	2.31E-01	Inf	3.68E-01	0.412	2.10E-01	1.420	**1.88E-03**	**2.114**
t(14;16)	2.77E-01	0.452	**1.74E-15**	**51.011**	**1.69E-03**	**0.258**	**8.97E-08**	**0.037**	**3.45E-19**	**68.171**	5.37E-02	0.431	**2.99E-05**	**0.130**
t(14;20)	7.82E-01	1.394	**1.40E-03**	**12.116**	2.77E-01	0.438	5.07E-02	0.092	**6.37E-06**	**39.343**	7.92E-01	1.526	**2.49E-02**	**0.105**
MYC	1.62E-01	1.477	1.00E+00	0.974	2.13E-01	0.678	7.92E-01	1.570	1.00E+00	0.962	6.97E-01	0.897	2.01E-01	0.730

b

Subgroup	Signature enrichment	Suggested aetiologies
Hyperdiploid	Signature 1	Aging
t(11;14)	Flat signatures	Potentially DNA repair deficiency
t(4;14)	Signature 2, 30, and flat signatures	APOBEC and potentially DNA repair deficiency
t(14;16)	Signature 2 and 13	APOBEC
t(14;20)	Signature 2 and 13	APOBEC
MYC	NA	NA

Table 4.10: Association of myeloma subgroups and structural rearrangement signatures (RS)

Subgroups	RS1		RS2		RS3		RS4		RS5	
	Q-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR
HD	2.93E-01	1.31	7.56E-02	0.68	1.00E+00	0.99	4.70E-01	1.40	1.00E+00	1.01
t(11;14)	**3.14E-05**	**0.41**	8.16E-02	1.56	8.42E-01	0.83	2.06E-01	2.06	9.24E-01	1.27
t(4;14)	**3.76E-06**	**5.28**	**1.01E-02**	**2.04**	1.97E-01	1.59	6.10E-01	0.69	1.00E+00	1.11
t(14;16)	1.00E+00	1.00	8.16E-02	2.45	8.95E-01	1.31	8.16E-02	0.32	9.91E-01	0.77
t(14;20)	8.84E-01	0.58	1.00E+00	0.89	1.00E+00	0.71	9.24E-01	0.87	1.00E+00	Inf
MYC	9.43E-01	0.90	5.15E-01	1.30	8.84E-01	1.15	1.00E+00	1.08	7.44E-01	1.63

Table 4.11: Association of established poor prognostic markers and mutational signatures. (a): COSMIC signatures and (b) rearrangement signatures (RS). Sig, signature.

a

Prognostic events	Sig1		Sig2		Flat signatures		Sig9		Sig13		Sig16		Sig30	
	Q- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR
1p deletion	3.59E-01	0.75	**2.41E-06**	**2.50**	3.62E-01	1.33	3.62E-01	0.69	**1.83E-02**	**2.62**	3.59E-01	0.81	3.62E-01	1.17
1q gain	**5.46E-04**	**0.51**	**5.39E-18**	**4.24**	**4.60E-02**	**1.59**	7.70E-01	0.76	**4.60E-02**	**2.11**	9.00E-01	0.96	1.00E+00	1.00
17p deletion	4.69E-01	0.00	4.69E-01	2.98	6.61E-01	0.74	5.27E-01	0.30	5.27E-01	2.34	6.21E-01	0.66	5.27E-01	2.29
TP53 mutations	6.04E-01	1.43	7.44E-02	2.45	6.04E-01	2.09	8.57E-01	Inf	6.04E-01	1.96	8.63E-01	0.93	8.57E-01	0.86

b

Prognostic events	RS1		RS2		RS3		RS4		RS5	
	Q- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR	*Q*- value	OR
1p deletion	**1.96E-05**	**2.61**	1.26E-01	1.37	7.12E-01	1.12	6.30E-02	0.55	1.00E+00	1.04
1q gain	**1.47E-04**	**1.99**	1.18E-01	1.36	2.38E-01	1.27	2.39E-01	1.40	1.00E+00	1.02
17p deletion	2.10E-01	Inf	2.10E-01	3.21	9.04E-01	1.24	9.04E-01	Inf	1.00E+00	Inf
TP53 mutations	**3.26E-02**	**3.36**	**7.42E-03**	**3.12**	2.45E-01	0.56	**3.26E-02**	**0.36**	2.16E-01	0.46

4.3.7 Mutational signatures and driver genes

To identify the aetiological mutational processes underlying driver mutations in MM, mutational contribution in driver genes was compared to other exonic mutations. Overall, the same diversity of processes in driver mutations was seen as in other coding mutations, but with differences: lower contribution of signatures 2 and 13; and higher contribution of signatures 1, 9, 16, 30, and flat signatures in coding regions of driver genes, compared to other exonic mutations (Figure 4.11). Notably, an over-representation of signatures reflective of aging in *CCND1* and *DNAH5* mutations, and AID in *EGR1* mutations was observed (Table 4.12, Figure 4.11). In contrast, a relative under-representation of signatures 2 and 13 suggests APOBEC mutations are ubiquitous mutational processes and they do not specifically affect driver genes. Driver genes were replicated earlier than other coding genes ($P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) and I therefore assessed whether this difference could explain enrichment of the signatures. APOBEC signature 2 is enriched in late replicating regions (Figure 4.8), hence the tendency of driver genes to be replicated early may explain the lower frequency of signature 2 mutations associated with driver genes. Signatures 1, 9, 16, 30, and the flat signatures were also associated with late replicating regions (Figure 4.8) but conversely were more frequently associated with driver gene mutations. To test if the enrichment of mutational processes in driver genes were due to positive selection of certain mutations, I excluded all mutations that occurred at the exact same position in multiple tumours (46% of mutations) and repeated the analysis. Exclusion of recurrent mutations did not change the overall results, inferring that positive selection of specific mutations did not bias the analysis. No significant transcriptional strand bias across mutational signatures was observed (Figure 4.7d), suggesting that the differences in mutational contribution between driver genes and other exonic mutations are unlikely to be influenced by transcription.

Figure 4.11: Mutational signatures associated with driver genes. (a) Cumulative mutational contribution of mutational signatures across 50 MM driver genes[1, 3-5] (blue, 1679 mutations in total) and other exonic mutations (orange). (b**)** Normalised cumulative mutational contribution of signatures with top ten contribution for most frequently mutated MM driver genes (+) versus other mutations (-) in tumours with the corresponding driver gene being mutated: *KRAS* (n = 247), *NRAS* (n = 204), *DIS3* (n = 104), *TRAF3* (n = 83), *CCND1* (n = 78), *BRAF* (n = 70), *FAM46C* (n = 70), *EGR1* (n = 65), *TP53* (n = 52), *SP140* (n = 30), *PRDM1* (n = 26), *ATM* (n = 19); n, number of mutations. Flat signatures include COSMIC signatures 3, 5, and 8.

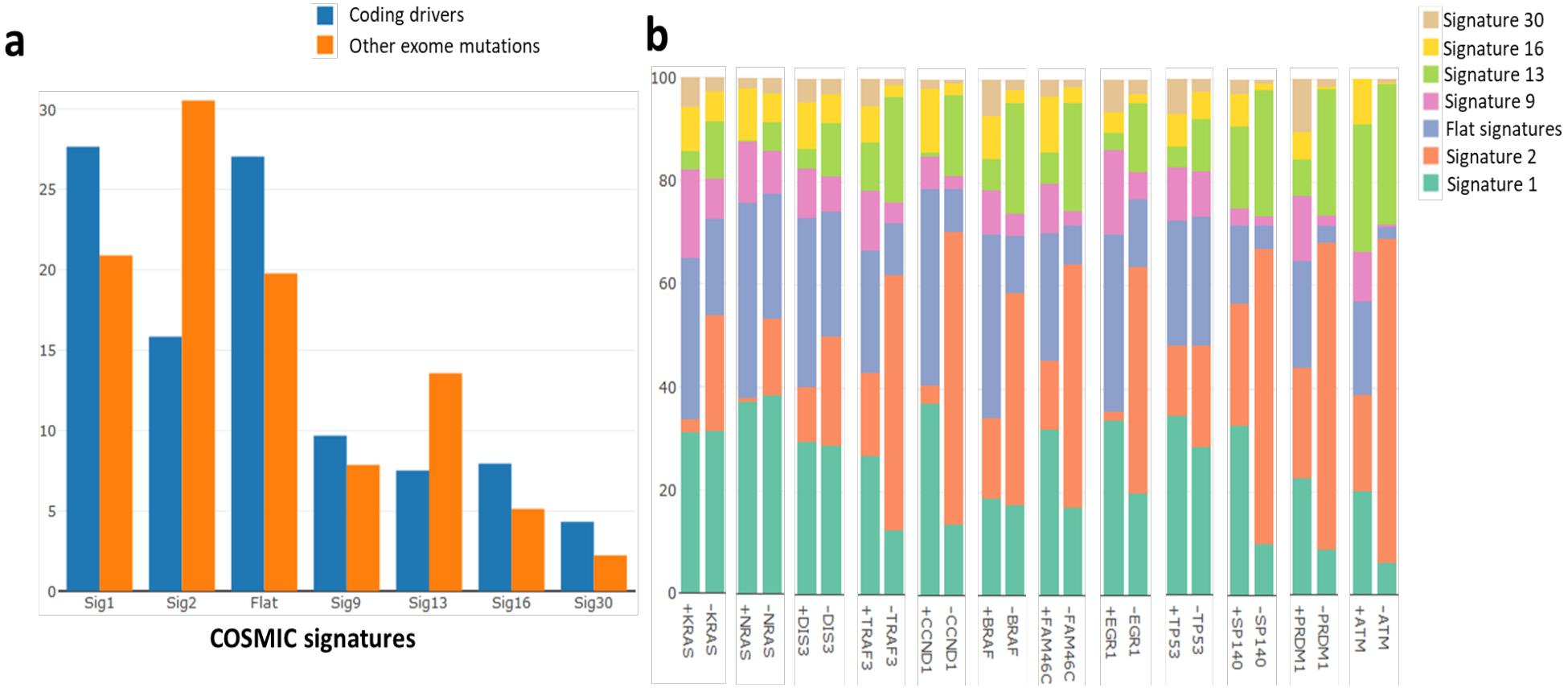


Table 4.12: Driver genes significantly preferentially targeted by certain mutational processes ($Q < 0.05$). wt, wild-type.

Gene	Signature	Mutated gene vs gene-mutated tumours			Mutated gene vs gene-wt tumours		
		Q-values	Mean mutated gene	Mean other mutations	*Q*-values	Mean mutated gene	Mean other mutations
CCND1	1	2.46E-19	0.224	0.079	3.86E-02	0.224	0.171
	2	7.23E-18	0.030	0.467	3.41E-07	0.030	0.204
	13	2.11E-16	0.006	0.153	2.00E-03	0.006	0.102
	16	5.23E-15	0.086	0.022	8.94E-04	0.086	0.019
DNAH5	1	9.73E-12	0.263	0.093	3.15E-02	0.263	0.180
	2	1.14E-11	0.088	0.422	2.79E-02	0.088	0.179
EGR1	2	1.24E-08	0.015	0.370	3.54E-04	0.015	0.212
	9	9.06E-04	0.093	0.063	1.12E-02	0.093	0.062
	13	1.11E-07	0.015	0.115	1.86E-03	0.015	0.106
FAM46C	13	5.91E-12	0.042	0.195	2.79E-02	0.042	0.087
HIST1H1E	13	7.57E-05	0.037	0.139	2.73E-02	0.037	0.104
MAX	16	5.09E-85	0.146	0.004	5.46E-03	0.146	0.044
TP53	2	3.59E-02	0.111	0.189	2.41E-02	0.111	0.233
	13	3.81E-03	0.054	0.096	2.77E-02	0.054	0.108

4.3.8 Prognostic impact of mutational signatures

The prognostic impact of mutational signatures was next investigated using the prospective data from CoMMpass. The APOBEC signature has previously been reported to be associated with a worse patient outcome[84, 230]. In this study after adjusting for age, sex, translocation status, chromosome-arm events, and *TP53* status no statistically significant association was shown suggesting that APOBEC status does not represent an independent biomarker of patient outcome; progression free survival (PFS: hazard ratio [HR] = 2.45, 95% confidence interval [CI] = 0.94 – 6.37, $P = 0.066$) and overall survival (OS: HR = 2.81, CI = 0.96 – 10.10, $P = 0.10$) (Table 4.13). I next explored whether incorporating information on major SNVs and SVs mutational signatures could further enhance the prediction of patient outcome after taking into account of the established prognostic factors. Unsupervised hierarchical clustering provided evidence for 7 distinct groups (A-G) associated with both PFS (log-rank $P = 3.4 \times 10^{-4}$) and OS (log-rank $P = 0.011$) (Figure 4.12, Figure 4.13, Table 4.14); with group C being enriched for hyperdiploid MM, group G is featuring tumours with 1p deletion, while group D being characterised by APOBEC mutation, enrichment for *MAF*-translocation subgroups, 1p deletion, and 1q gain (Table 4.15)**.** Post-hoc delineation allowed stratifications of patients in 7 groups into low- (A, B, C, and E) and high-risk groups (D, G, and F) (Table 4.16). Classification of MM based on mutational signatures captured by these 7 groups are independent prognosis factors. Notably, group F was independently associated with adverse prognosis (PFS: HR = 1.95, 95% CI = 1.35 – 2.81, $P = 3.3 \times 10^{-4}$; OS: HR = 1.47, 95% CI = 1.02 – 2.13, $P = 0.039$) (Table 4.17), despite not being associated with the high-risk features of APOBEC, t(14;16)/t(14;20), 1p/1q/17p chromosome-arm events or *TP53* mutation status; but was typified by non-clustered structural rearrangements (Figure 4.12a, Figure 4.13, Table 4.14).

Table 4.13: Multivariable Cox regression analysis of progression free and overall survival with APOBEC mutational contribution.
HR, hazard ratio.

Variates	Progression Free Survival				Overall survival			
	HR	Lower 95%	Upper 95%	*P*-value	HR	Lower 95%	Upper 95%	*P*-value
APOBEC mutation	2.45	0.94	6.37	6.62E-02	2.81	0.96	10.10	1.04E-01
Age	**1.04**	**1.02**	**1.05**	**4.18E-08**	**1.04**	**1.03**	**1.06**	**3.49E-06**
Male/Female	**1.57**	**1.18**	**2.10**	**2.03E-03**	**1.97**	**1.30**	**3.00**	**1.54E-03**
Hyperdiploidy	0.93	0.65	1.33	6.99E-01	1.09	0.68	1.76	7.12E-01
t(11;14)	1.23	0.79	1.92	3.49E-01	0.80	0.42	1.54	5.09E-01
t(4;14)	1.06	0.69	1.64	7.85E-01	0.93	0.51	1.70	8.17E-01
t(14;16)	0.67	0.28	1.64	3.83E-01	0.81	0.25	2.58	7.17E-01
t(6;14)	1.04	0.32	3.35	9.44E-01	1.29	0.30	5.44	7.33E-01
t(14;20)	0.87	0.26	2.90	8.15E-01	1.31	0.36	4.84	6.82E-01
***MYC*-translocation**	**1.65**	**1.17**	**2.33**	**3.93E-03**	1.39	0.87	2.24	1.69E-01
1p del	1.28	0.94	1.75	1.20E-01	**1.89**	**1.26**	**2.82**	**1.89E-03**
1q gain	**1.68**	**1.27**	**2.20**	**2.15E-04**	**1.59**	**1.09**	**2.32**	**1.64E-02**
17p del	0.62	0.15	2.51	5.00E-01	0.52	0.07	3.80	5.23E-01
***TP53* mutations**	**1.77**	**1.05**	**2.96**	**3.10E-02**	1.59	0.78	3.24	1.98E-01

Figure 4.12: Integrative clusters based on mutational signatures and patient prognosis. (a) Heatmap showing proportions of rearrangement signatures and major COSMIC signatures in unsupervised hierarchical clusters. Flat signatures include COSMIC signatures 3, 5, and 8. The lower panel shows distribution of translocations, prognostic chromosome-arm events, and *TP53* non-synonymous mutations across all samples. (b) Progression free survival and (c) overall survival across different cluster groups. The global *P*-values across all cluster groups were calculated to assess whether there is survival difference between groups.

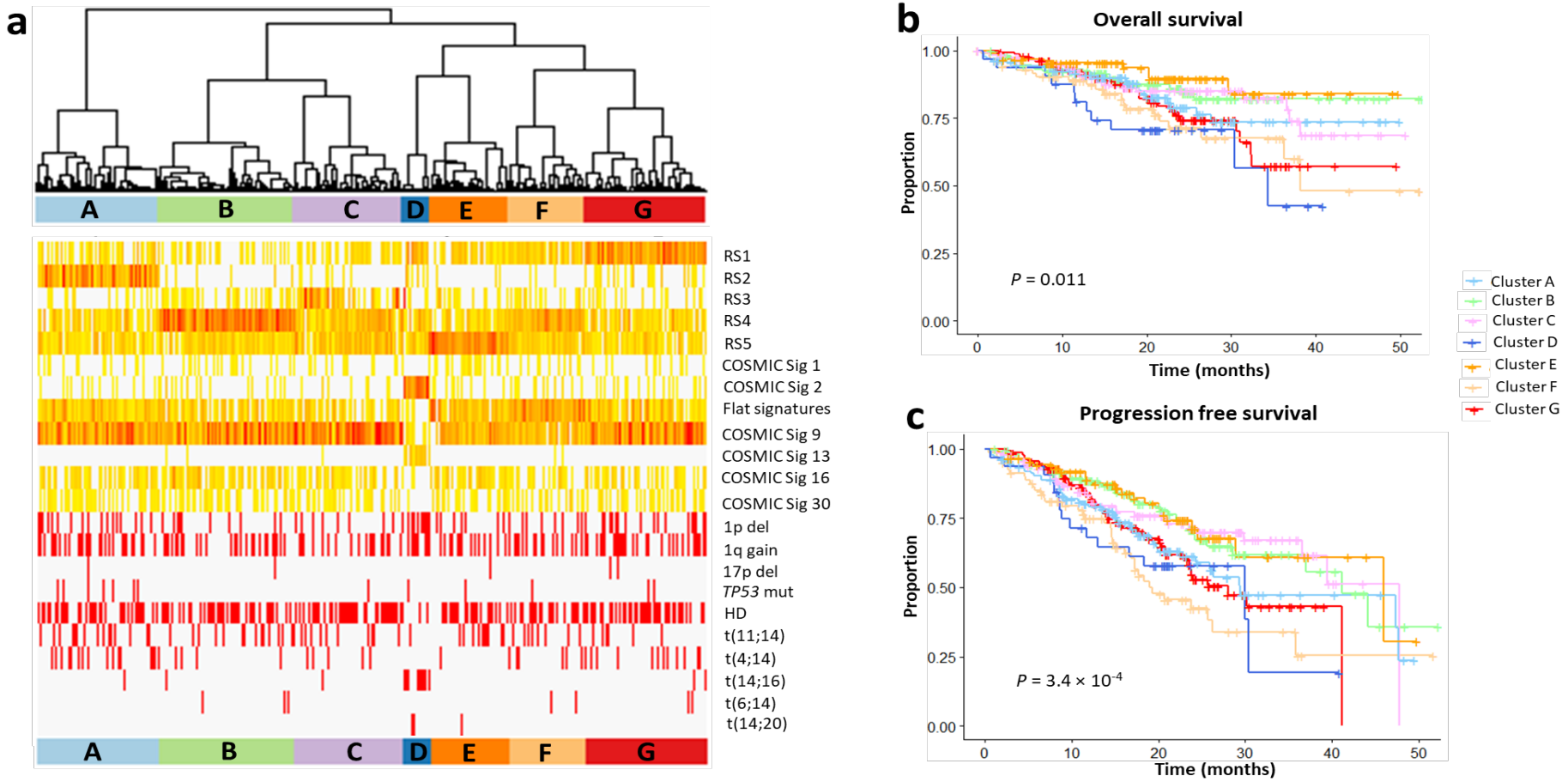


Figure 4.13: Contribution of mutational signatures in each of the unsupervised hierarchical clustered subgroups (A – G). (a) Structural rearrangements and (b) COSMIC single nucleotide variant signatures (>1% contribution across all subgroups). RS, structural rearrangement signatures. Sig, signature.

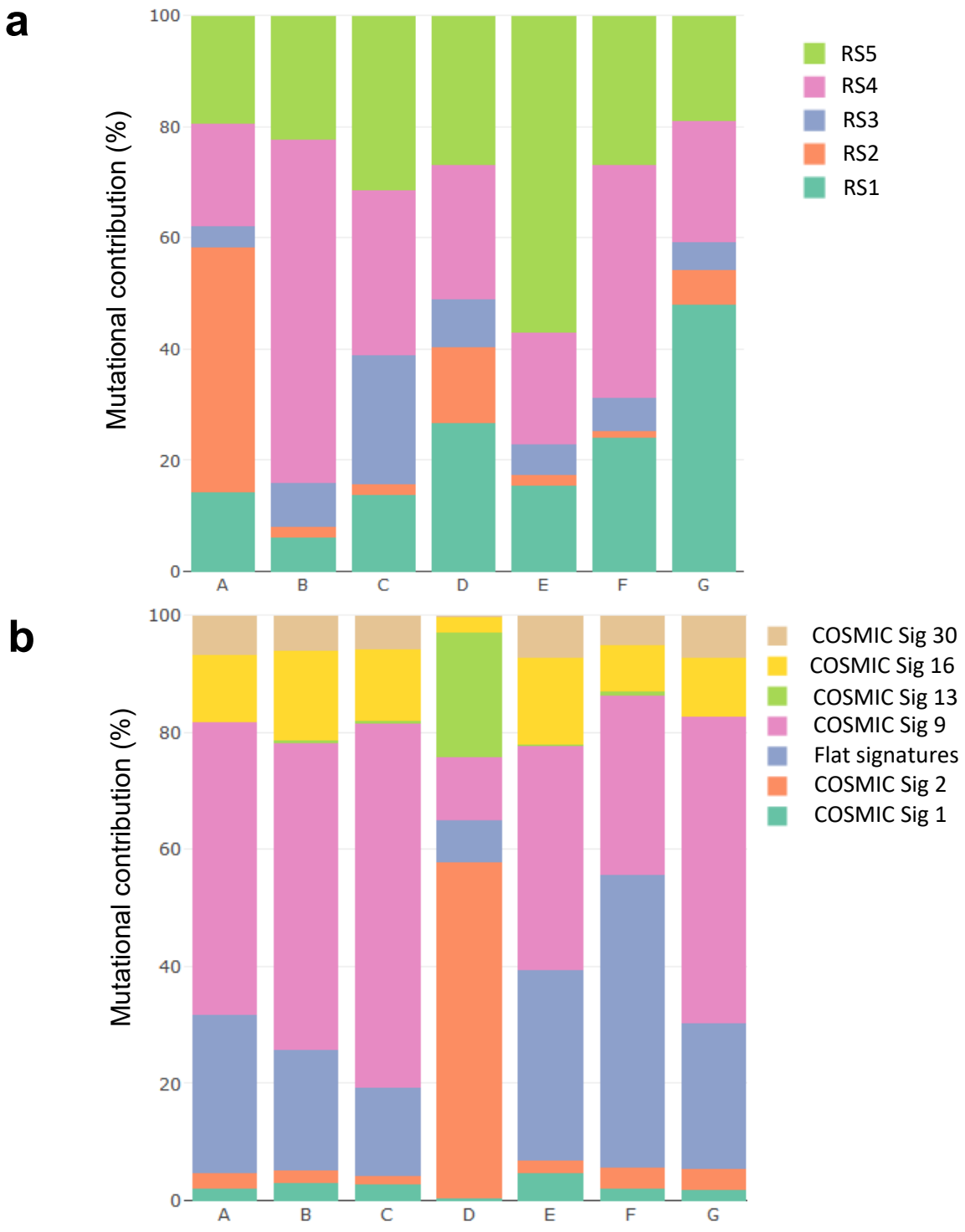


Table 4.14: Summary of characteristics of the seven cluster subgroups. SV, structural variant; SNV, single nucleotide variant.

Cluster	n	SV features	SNV features	Subgroup association	Known prognostic events
A	155	Clustered translocations		Enriched for t(11;14) and t(4;14)	*TP53* mutations
B	172	Non-clustered small-scaled deletions & tandem duplications			
C	138	Mixture of non-clustered SVs		Enriched for hyperdiploidy	
D	35	Mixture of non-clustered SVs	APOBEC mutations	Enriched for t(14;16) and t(14;20)	1p deletion and 1q gain
E	99	Non-clustered translocations			
F	97	Mixture of non-clustered SVs			
G	154	Large-scaled non-clustered deletions, tandem duplications, and inversions		Enriched for t(4;14)	1p deletion

Table 4.15: Association of myeloma subgroups and known prognostic events with unsupervised hierarchical clusters. OR, odd ratios. In bold, significant values.

Cluster	HD		t(11;14)		t(4;14)		t(14;16)		t(14;20)		*MYC*-translocation		1p deletion		1q gain		17p deletion		*TP53* mutations	
	Q-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR	*Q*-value	OR
A	9.08E-01	0.96	**2.69E-02**	**1.86**	**1.36E-02**	**2.20**	6.19E-01	0.48	4.76E-01	0.00	6.53E-01	1.21	8.94E-01	1.05	6.06E-01	1.20	1.00E+00	1.12	**4.58E-02**	**2.67**
B	3.58E-01	1.32	8.94E-01	1.05	1.48E-01	0.52	5.89E-02	0.13	4.76E-01	0.00	8.27E-01	1.11	2.42E-01	0.67	2.58E-01	0.73	8.26E-01	0.44	4.76E-01	0.48
C	**4.60E-02**	**1.66**	7.90E-01	0.86	6.34E-01	0.74	6.69E-01	0.53	6.34E-01	0.00	1.00E+00	0.99	1.05E-01	0.56	**2.14E-02**	**0.54**	6.34E-01	0.00	7.06E-01	0.63
D	**1.33E-02**	**0.29**	2.13E-01	0.27	6.53E-01	0.46	**5.27E-20**	**81.86**	**1.42E-05**	**44.23**	1.00E+00	0.83	**3.37E-02**	**2.82**	**1.08E-02**	**3.52**	6.19E-01	2.63	4.52E-01	2.22
E	8.51E-01	1.09	4.76E-01	1.39	1.44E-01	0.39	1.44E-01	0.00	7.85E-02	5.61	7.42E-01	1.19	5.09E-01	0.71	7.96E-01	0.90	6.19E-01	1.91	6.53E-01	0.44
F	9.52E-01	0.97	1.29E-01	1.78	7.82E-01	0.80	1.44E-01	0.00	7.82E-01	0.00	5.32E-01	0.65	8.94E-01	0.92	7.82E-01	1.14	7.82E-01	0.00	7.82E-01	1.27
G	6.34E-01	1.19	**2.14E-02**	**0.43**	**1.09E-02**	**2.35**	8.94E-01	1.10	4.76E-01	0.00	6.53E-01	1.22	**1.09E-02**	**1.98**	3.12E-01	1.34	2.58E-01	3.06	7.96E-01	0.72

Table 4.16: Multiple pair-wise comparisons between unsupervised hierarchical clusters using log-rank test (*P*-values). (a) Overall survival and (b) progression-free survival. In bold: significant values.

a

Clusters	A	B	C	D	E	F
B	3.52E-01	-	-	-	-	-
C	5.71E-01	7.10E-01	-	-	-	-
D	6.21E-02	**7.10E-03**	**1.88E-02**	-	-	-
E	9.69E-02	3.78E-01	2.66E-01	**2.30E-03**	-	-
F	1.42E-01	**1.98E-02**	**4.92E-02**	4.73E-01	**5.40E-03**	-
G	5.12E-01	9.79E-02	1.22E-01	1.86E-01	**2.41E-02**	5.20E-01

b

Clusters	A	B	C	D	E	F
B	5.51E-02	-	-	-	-	-
C	1.15E-01	8.09E-01	-	-	-	-
D	3.51E-01	**1.40E-02**	**3.30E-02**	-	-	-
E	6.92E-02	7.82E-01	6.61E-01	**2.01E-02**	-	-
F	**4.95E-02**	**9.50E-05**	**7.00E-04**	5.35E-01	**4.20E-04**	-
G	8.51E-01	**2.19E-02**	**4.78E-02**	3.49E-01	**2.90E-02**	6.01E-02

Table 4.17: Multivariable Cox regression analysis of progression free and overall survival for subgroup F versus other subgroups.
In bold, significant values. HR, hazard ratio.

Variates	Progression Free Survival				Overall survival			
	HR	Lower 95%	Upper 95%	*P*-values	HR	Lower 95%	Upper 95%	*P*-values
Subgroup F/Non-F	**1.95**	**1.35**	**2.81**	**3.32E-04**	**1.47**	**1.02**	**2.13**	**3.89E-02**
Age	**1.04**	**1.02**	**1.05**	**7.75E-08**	**1.03**	**1.02**	**1.05**	**8.07E-07**
Male/Female	**1.55**	**1.16**	**2.07**	**2.90E-03**	**1.52**	**1.14**	**2.02**	**4.59E-03**
Hyperdiploidy	0.93	0.65	1.32	6.66E-01	0.96	0.66	1.39	8.18E-01
t(11;14)	1.25	0.81	1.94	3.17E-01	1.17	0.74	1.83	5.05E-01
t(4;14)	1.09	0.70	1.68	7.08E-01	0.92	0.59	1.44	7.17E-01
t(14;16)	0.73	0.29	1.80	4.92E-01	0.75	0.29	1.96	5.63E-01
t(6;14)	1.02	0.32	3.26	9.75E-01	0.99	0.31	3.21	9.89E-01
t(14;20)	0.95	0.28	3.18	9.28E-01	0.80	0.23	2.83	7.31E-01
***MYC*-translocation**	**1.75**	**1.24**	**2.47**	**1.47E-03**	**1.59**	**1.12**	**2.25**	**9.18E-03**
APOBEC mutation	2.44	0.91	6.56	7.64E-02	2.24	0.77	6.52	1.37E-01
1p del	1.34	0.98	1.83	7.04E-02	**1.38**	**1.00**	**1.89**	**4.66E-02**
1q gain	**1.64**	**1.24**	**2.16**	**4.49E-04**	**1.60**	**1.21**	**2.11**	**9.22E-04**
17p del	0.68	0.17	2.78	5.91E-01	0.79	0.19	3.24	7.46E-01
***TP53* mutations**	**1.73**	**1.03**	**2.90**	**3.73E-02**	**1.97**	**1.17**	**3.33**	**1.12E-02**

4.4 Discussion

The analysis of over 800 myeloma genomes has afforded a global overview of the mutational processes in MM tumorigenesis. A major finding of this study is that a combination of signatures linked to aging, APOBEC/AID and indicative DNA repair deficiency - account for around 80% of mutations in MM. Despite the difficulty of assigning flat signatures (signatures 3, 5, and 8)[178, 239], their detection of such profiles in large patient series supports the role of defective DNA repair in MM. By utilizing both WES and WGS data, I was able to extract five novel structural rearrangement signatures and identify differential prevalent mutational processes in coding (aging and APOBEC) and non-coding regions (AID), consistent with a previous report[213]. The work supported previous findings[213] in implying an early role for AID in shaping the MM mutational landscape. I also identified new and validated previously reported subgroup associations with mutational signatures, allowing further categorization of MM beyond simple translocation status and providing additional insight in the aetiological processes implicated in tumorigenesis (Figure 4.14).

Mutations do not occur uniformly over the genome and local mutation rates are modulated by replication, transcription, and chromatin organisation[223]. An enrichment of somatic mutations in late replicating regions, as seen across several cancers[240], and highly expressed regions was observed. Previous analyses which have sought to establish the mutational profile of myeloma genomes have been based on data solely from exome sequencing projects. Here I sought to provide a more comprehensive analysis however, I acknowledge that the low coverage of CoMMpass WGS raises the possibility that the global mutation rate may have been underestimated. The strong replicative asymmetry observed is consistent with mutations in MM being predominantly associated with APOBEC-family of mutations[231]. In addition, I identified that coding drivers are likely to be originated from a number of mutational processes including aging and DNA repair deficiency. In contrast, while APOBEC enzymes appear to act more ubiquitously within coding regions, they do not specifically affect coding drivers.

The different MM translocation subgroups showed striking differences in their mutational signatures, reflective of the cellular processes driving respective clonal expansions (Figure 4.14). As previously reported, t(14;16) and t(14;20) MM

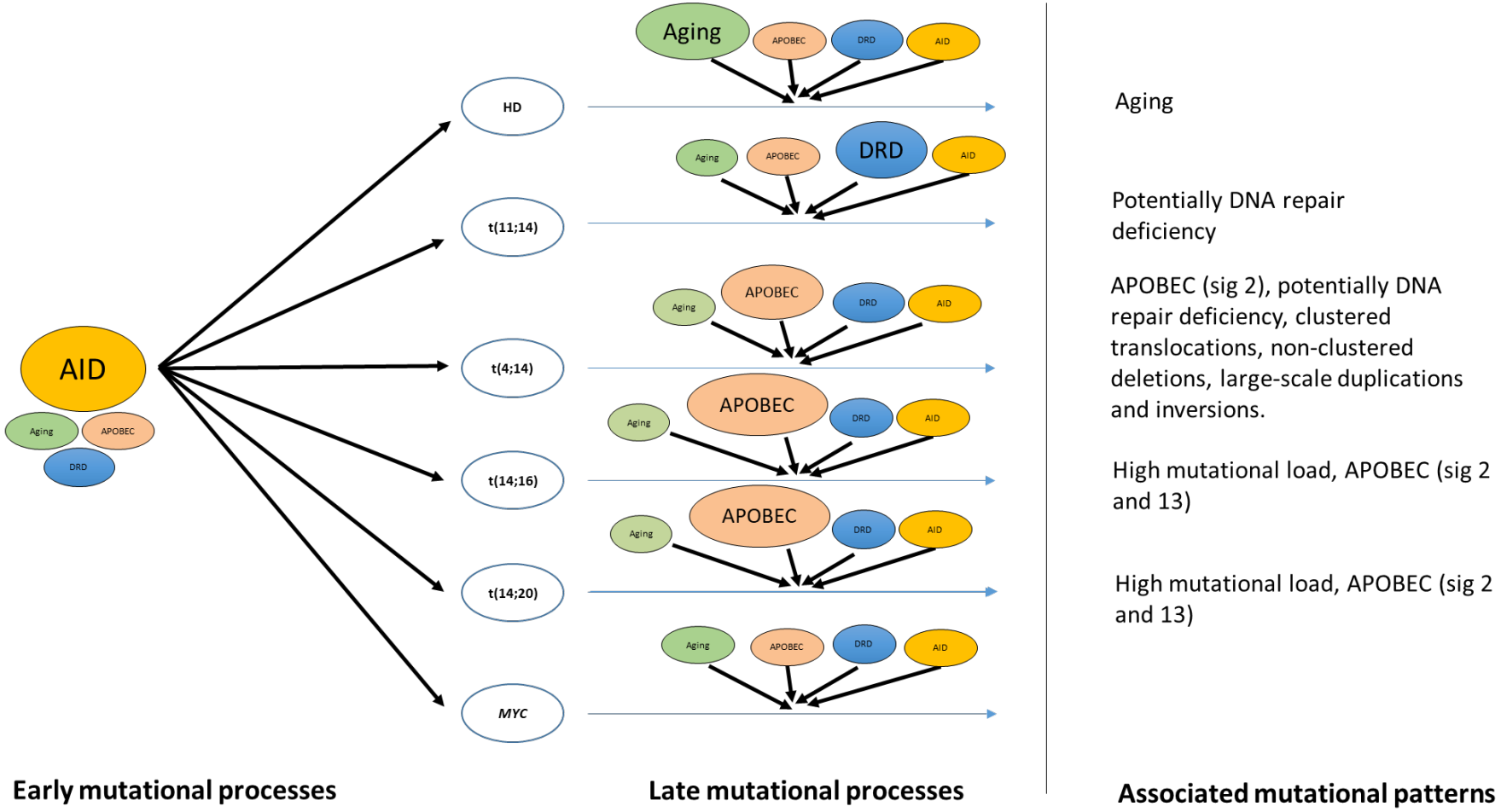
were enriched with APOBEC signatures 2 and 13[4, 84]. This is a consequence of the over-expression of APOBEC genes, specifically *APOBEC3A* and *APOBEC3B*, mediated through the over-expression of MAF transcription factors[84]. The t(4;14) subgroup was also enriched with APOBEC mutational patterns, although only for signature 2 and to a lesser extent as compared to *MAF*-translocation subgroups. Since signatures 2 and 13 are reflective of different mutational processes[223] I speculate that the mutational processes associated with t(4;14) are likely to be different from those with *MAF*-translocation subgroups. In contrast signatures indicative of homologous recombination and aging were associated with t(11;14) and HD respectively. DNA breaks unsuccessfully repaired due to defective DNA repair may facilitate the generation of chromosomal translocations[241]. Because of the flat structure of signatures 3, 5 and 8, robust insight into the aetiological contribution of homologous recombination deficiency to MM tumorigenesis requires assiduous signature fitting and adjustment for confounding covariates[239]. The molecular mechanisms responsible for initiating HD in MM are unknown. However, by inference from childhood acute lymphoblastic leukemia[242], it is likely it is a consequence of the simultaneous gain of chromosomes in a single abnormal cell division. Cells failing to execute programmed cell death in response to mitotic failure are likely to divide asymmetrically, resulting in generation of aneuploidy cells[243]. The association between aging with increased cell division errors[244] and decreased apoptosis[245], further supports a relationship between hyperdiploid MM and aging. Signatures defined by large-scale structural aberrations were associated to varying degrees with MM subgroups but clustered translocations and non-clustered deletions, large-scale tandem duplications and inversions showed a significant association in t(4;14) MM.

The APOBEC mutational signatures are inextricably linked to a high mutation load[4, 84] and the adverse t(14;16) and t(14;20) *MAF*-translocation subgroups. The study shows that molecular classification based solely on APOBEC signatures do not fully differentiate the underlying genomic complexity in MM relevant to predicting patient outcome. Hence while APOBEC activity is an adverse prognostic factor in MM[84, 230], using it as a sole classifier does not fully capture high-risk MM which with genetically unstable genome is typified by complex

structural variants. The findings support the need for considering other mutational signatures to refine prediction of patient prognosis.

This study does, however suggest that analysis of APOBEC activity together with other molecular features at diagnosis should allow for the identification of high-risk MM patients that may benefit from more intensive treatment. Collectively these data shed new light on the diversity of cellular processes generating somatic mutations in MM. Moreover, they provide a strong rationale for integration of mutational signatures data in conventional molecular profiling of patient tumours to tailor therapy.

Figure 4.14: Contribution of major mutational processes operative in MM. This model represents differential contribution of various identified mutational processes in myeloma. For early mutational processes, AID has the overall largest contribution to mutational processes across all subgroups represented by a larger oval. For late mutational processes, major mutational processes with known aetiologies associated with aging, APOBEC, DNA repair defects (DRD), and AID are depicted. Larger oval sizes indicate larger relative contribution of the mutational process.



CHAPTER 5 An enhanced genetic model of multiple myeloma evolutionary dynamics at relapse

5.1 Overview and rationale

Despite recent advances, MM is essentially an incurable malignancy, and most patients die from progressive disease after multiple relapses irrespective of treatment. Our limited knowledge of the molecular changes associated with relapse is a barrier to developing new therapeutic strategies to overcome drug resistance. Therefore, there is a need to understand the mutational spectrum, together with clonal dynamics and evolution from primary to relapsed tumours for future molecularly targeted therapy.

To advance our understanding of MM tumour evolution and the mutational mechanisms that shape their history, analysis was performed on WGS of 80 newly diagnosed MM tumour-normal pairs, of which 25 also had matched relapsed tumours from Myeloma XI trial patients[132], in this chapter. Through comprehensive characterisation and comparison between MM primary and relapsed genomes, I identified patterns of genetic alterations acquired at relapse, inferred the order of mutational events, and showed that relapse is associated with acquisition of new mutations and clonal selection, in part shaped by patient therapy. I also provided evidence for distinct patterns of clonal evolution, a finding that has important implications in guiding future therapy choices.

5.2 Study design

5.2.1 Samples and dataset

Myeloma XI trial dataset were obtained as detailed in section 2.1.2.

5.2.2 Statistical and bioinformatics analysis

5.2.2.1 Whole genome sequencing analysis

Quality control and sequence alignment to hg38 were performed using FastQC v.0.11.4/BWA v0.7.13/GATK v4.0.3.0 software as described in section 2.2.4. SNVs and indels were called using MuTect2 according to best practices[151], using gnomAD[139] file in GRCh38 provided as part of the GATK resource. Variants were filtered for cross-sample contamination, oxidation artefacts[158], quality score[4], and using a panel of normals generated from 80 germline samples. Variants with a germline population allele frequency > 0.1% gnomAD or in repetitive regions defined by UCSC were excluded. Somatic indels were excluded if they were supported by < 20% of tumour sample reads overlapping the position[92] or were located within 10 base pairs of a germline indel catalogued by gnomAD. Reconstruction of clonal and subclonal CNVs for primary and relapsed tumours was conducted using Battenberg[180] as described in section 2.2.11. Tumour purity estimated by Battenberg was compared against and corrected using Ccube[183] as detailed in section 2.2.11.3. Somatic SVs were identified taking a consensus approach, as implemented by The Pancancer Analysis of Whole Genomes[246] (PCAWG), considering only variants identified by at least two of SV callers MANTA[162] v1.2.0, LUMPY[163] v0.2.13, or DELLY[164] v0.7.9 (section 2.2.7.3). Chromothripsis was identified using ShatterSeek with default parameters[169]. Chromoplexy was detected using ChainFinder v1.0.1 with default parameters[167] and hg38 UCSC cytoband definitions (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/). Telomere length was estimated using Telomerecat[152] with default parameters. Kataegis foci were identified using the KataegisPortal with default parameters excluding immune hypermutated regions[127] (section 2.2.7.4).

5.2.2.2 Identifying driver mutations

Coding drivers were identified using dNdScv with default parameters[76]. Non-silent mutations in 87 established coding drivers (identified in chapter 3 and other studies)[4, 247], and all coding genes were compared in matched primary and relapsed tumours. To identify non-coding drivers, promoter and CREs were analysed as described in section 2.2.8. Promoters were defined as intervals spanning 400 bp upstream and 250 bp downstream of TSS from GENCODE (release 25)[248]. CREs were defined using promoter CHi-C data generated on naïve B-cells[154], with raw sequencing reads from EGA (accession code EGAS00001001911) were aligned to hg38 using HiCUP (v0.6.1)[153] and promoter-CRE interactions were called with CHiCAGO (v1.8)[155] (section 2.2.5). Only interactions with linear distance $\leq$ 1Mb and CHiCAGO score $\geq$ 5 were considered[4].

Recurrently mutated promoters and CREs were identified using a Poisson binomial model as previously described[4, 172] (section 2.2.8.2), taking into account tumour ID, trinucleotide context, and replication timing. Replication timing with hg38 coordinates was estimated as the average of two B-lymphocyte replicates (downloaded from https://www.replicationdomain.com). For those promoters and CREs mutated in $\geq$ 3 samples, the clustering of mutations was examined using a permutation approach considering the number of mutations occurring at the same nucleotide position as previously described[4] (section 2.2.8.2). For each promoter and CRE, a combined *P*-value from the mutational recurrence and clustering analyses were obtained using Fisher's method[4, 171]. Only CREs and promoters mutated in at least 3 tumours were reported. To test for the effects of focal CNV, focal deletion and amplification were defined from Battenberg output and size < 3 Mb. Gene expression was compared using edgeR[175] between mutated and unmutated samples, excluding those with CNV at the target gene[4]. Regulatory regions were only tested if they were mutated in at least two samples. *P*-values were adjusted for FDR thresholded at $Q < 0.05$.

5.2.2.3 Chronology of mutational events

The relative chronological timing of SNVs and CNVs was estimated independently for 80 primary tumours as previously described[249]. For SNVs only driver genes mutated in ≥ 4 samples were considered to allow reliable estimation of relative timing. For CNVs only large-scale autosomal events (≥ 3Mb) present in ≥ 8 samples were considered[249]. Cytobands were assigned based on UCSC hg38 definitions. One sample (8573) displayed hyperdiploid characteristics and this was excluded from the analysis. Each cytoband or driver gene was ordered by mean of CCF from highest to lowest. The Tukey's range test and a stepwise approach were used to test for difference between the means of consecutive cytoband/driver gene to establish distinct clonality groups[249].

5.2.2.4 Mapping evolutionary trajectories

Analysis of clonality was conducted using only SNVs in diploid regions, as miscalled copy number states can confound such analyses. Potential neutral tail mutations were identified using MOBSTER[184] and excluded prior to clustering procedure to minimise calling false positive clones. For each primary and relapse tumour pair, two-dimensional variant clustering was performed using a Bayesian Dirichlet process implemented in DPclust[5, 180] (section 2.2.11). Only those clusters with ≥ 1% of total mutations and ≥ 100 SNVs were considered. Muller plots were generated with Timescape R package (http://bioconductor.org/packages/release/bioc/html/timescape.html). Clonal SNVs were defined as those with a CCF ≥ 0.9[185]. For each cluster in primary tumour and matched relapse, the proportion of SNVs shared was calculated.

5.2.2.5 Mutational signatures

De novo extraction of signatures was performed on 80 primary and 25 relapse genomes separately using NMF implemented in Palimpsest R package[179]. *De novo* mutational signatures were compared and assigned to 30 COSMIC signatures[87] as detailed in section 2.2.10.2. Signature fitting was performed using deconstructSigs[178] (section 2.2.10.1) considering only those COSMIC signatures

extracted *de novo*, as previously recommended[239]. Novel signature M1 was primarily detected in only one tumour and therefore was not included it when fitting signatures[92]. In view of potential ambiguous assignment with respect to homologous recombination, the contributions of the flat profile signatures 3, 5 and 8[127, 178, 239] were combined as described in section 4.2.2.2. The Benjamini-Hochberg FDR procedure was used to adjust for multiple hypothesis testing with significance thresholded at $Q < 0.05$. Mutational signature proportions in paired primary and relapse samples were compared using the chi-squared test[180].

5.3 Results

5.3.1 Overview of primary tumours mutational landscape

WGS was carried out on 80 newly diagnosed MM tumour-normal pairs from the Myeloma XI trial, and matched relapsed tumours from 25 patients. The 80 patients had either t(4;14) (n = 38), t(11;14) (n = 38), or t(14;16) (n = 4) MM, with one patient carrying both t(4;14) translocation and trisomy of chromosomes 9 and 15 (Appendix 3). WGS resulted in a median of 38x coverage for normal samples (30 – 44x), 111× for primary tumours (82 – 155x), and 114x for the 25 relapsed tumours (102 – 156x) (Appendix 3). I began by surveying for important genetic alterations in the 80 primary MM tumours through considering the contribution of both protein-coding and non-coding SNVs and indels. As expected, significantly mutated genes ($Q < 0.05$) at presentation were *DIS3*, *KRAS*, *NRAS*, *FGFR3*, *MAX*, *CCND1*, *TP53*, *IGLL5*, *IRF4*, and *PRKD2* (Table 5.1). The promoters of 17 genes including *BCL6*, *CXCR4*, *BIRC3*, *MYO1E*, *CRIP1*, *FLT3LG*, and *DPP9* were also significantly mutated as well as 9 CREs interacting with genes including *PAX5*, *BCL6*, *ZCCHC7*, and *IFNGR1* (Table 5.2, Table 5.3, Figure 5.1). Focal deletions of CREs resulting in decreased *BIRC2* (25 fold, $Q = 2.4 \times 10^{-3}$) and *IGLL5* (414 fold, $Q = 1.1 \times 10^{-3}$) expression were also identified (Figure 5.1). Chromothripsis was only observed in 3 tumours (3.8%) (Figure 5.2) affecting chromosomes 1q, 3, 8, 11, and 12; whereas 78% (62/80) of tumours featured chromoplexy. The driver genes[4, 247] most commonly disrupted by chromoplexy were *SP140*, *SF3B1*, *IDH1*, and *DUSP2* (Table 5.4). Overall across the 80 tumours, high-risk subtypes MM t(4;14) and t(14;16) were associated with a higher number of chromoplexy events ($P = 3.9 \times 10^{-3}$, Wilcoxon rank-sum test) and shorter telomeres ($P = 9.2 \times 10^{-5}$) (Figure 5.3)

Table 5.1: Significantly mutated genes identified from 80 primary tumours. ($Q < 0.05$). n, number.

Gene	n synomynous	n missesnse	n nonsense	n splice site	n indel	*P*-value	*Q*-value
DIS3	0	17	0	1	0	2.22E-16	2.22E-16
KRAS	1	13	0	0	0	2.22E-16	2.22E-16
NRAS	0	15	0	0	0	2.22E-16	2.22E-16
FGFR3	0	13	0	0	0	1.18E-13	5.37E-10
MAX	0	6	0	0	0	1.52E-08	5.52E-05
CCND1	2	6	0	0	0	5.94E-07	1.80E-03
TP53	1	3	1	0	1	1.74E-06	4.53E-03
IGLL5	2	5	0	0	0	2.51E-06	5.71E-03
IRF4	1	6	0	0	0	4.05E-06	8.19E-03
PRKD2	1	7	0	0	0	1.33E-05	2.42E-02

Table 5.2: Recurrently mutated *cis*-regulatory elements from 80 primary tumours. ($Q < 0.05$)

Fragment	Size	Target gene	Number of mutations	Number of mutated samples	*Q*-value
chr9:37375175-37395285	20110	*PAX5; AL161781.2*	27	15	2.15E-11
chr9:37369119-37373681	4562	*PAX5; AL161781.3*	20	13	1.89E-10
chr9:37406897-37411656	4759	*PAX5; AL161781.4*	14	11	1.89E-10
chr9:37025270-37031362	6092	*ZCCHC7; AL512604.2*	14	8	1.02E-09
chr3:188747605-188754794	7189	*BCL6; LPP; LPP-AS2*	16	11	2.00E-09
chr3:187746361-187747275	914	*BCL6; AC022498.2; LPP; LPP-AS1; LPP-AS2; miR28*	6	4	3.92E-04
chr15:74772989-74775174	2185	*CSK; FAM219B; MPI; SEMA7A*	6	3	1.41E-02
chr17:68772663-68776750	4087	*ABCA10; ABCA5*	5	4	3.10E-02
chr6:137413091-137415723	2632	*IFNGR1*	4	3	3.10E-02

Table 5.3: Recurrently mutated promoters from 80 primary tumours. (Q < 0.05)

Fragment	Size	Target gene	Number of mutations	Number of mutated samples	Q-value
chr3:187745209-187745859	650	*BCL6*	9	7	1.43E-15
chr3:187745187-187745837	650	*BCL6*	11	9	1.77E-15
chr2:136117487-136118137	650	*CXCR4*	11	6	1.15E-12
chr3:187745222-187745872	650	*BCL6*	7	6	3.67E-12
chr11:102317163-102317813	650	*BIRC3*	6	4	3.87E-11
chr15:59372277-59372927	650	*MYO1E*	6	5	2.08E-09
chr15:59372293-59372943	650	*FAM81A*	6	5	2.08E-09
chr14:105487454-105488104	650	*CRIP1*	5	4	2.08E-09
chr14:105487821-105488471	650	*CRIP1*	4	3	9.43E-08
chr14:105487784-105488434	650	*TEDC1*	4	3	9.43E-08
chr14:105487812-105488462	650	*TEDC1*	4	3	9.43E-08
chr11:102317095-102317745	650	*BIRC3*	4	3	1.26E-07
chr11:102317102-102317752	650	*BIRC3*	4	3	1.26E-07
chr3:187745477-187746127	650	*BCL6*	4	3	3.28E-06
chr3:187745475-187746125	650	*BCL6*	4	3	3.56E-06
chr19:49473852-49474502	650	*FLT3LG*	3	3	3.75E-06
chr19:49473828-49474478	650	*FLT3LG*	3	3	3.75E-06
chr19:49473834-49474484	650	*FLT3LG*	3	3	3.75E-06
chr19:49473836-49474486	650	*FLT3LG*	3	3	3.75E-06
chr19:49473807-49474457	650	*FLT3LG*	3	3	3.75E-06
chr11:102317084-102317734	650	*BIRC3*	3	3	3.75E-06
chr19:4723532-4724182	650	*DPP9*	3	3	3.75E-06
chr19:4723500-4724150	650	*DPP9*	3	3	3.75E-06
chr19:4723613-4724263	650	*DPP9*	3	3	3.75E-06
chr19:4723547-4724197	650	*DPP9*	3	3	3.75E-06
chr19:4723580-4724230	650	*DPP9*	3	3	3.75E-06
chr19:4723570-4724220	650	*DPP9*	3	3	3.75E-06
chr19:4723592-4724242	650	*DPP9*	3	3	3.75E-06
chr19:4723563-4724213	650	*DPP9*	3	3	3.75E-06
chr19:4723585-4724235	650	*DPP9*	3	3	3.75E-06
chr19:4723556-4724206	650	*DPP9*	3	3	3.75E-06
chr5:159100222-159100872	650	*LINC02202*	3	3	4.53E-06
chr5:159100282-159100932	650	*LINC02202*	3	3	4.53E-06
chr11:132211340-132211990	650	*NTM*	3	3	6.93E-06
chr11:132211357-132212007	650	*NTM*	3	3	6.93E-06
chr11:132211399-132212049	650	*NTM*	3	3	7.03E-06
chr5:147906240-147906890	650	*C5orf46*	3	3	1.36E-05
chr5:147906288-147906938	650	*C5orf46*	3	3	1.36E-05
chr5:147906252-147906902	650	*C5orf46*	3	3	1.36E-05
chr4:177442159-177442809	650	*AGA*	4	4	1.09E-04
chr19:10230013-10230663	650	*MIR4322*	5	3	6.40E-04
chr14:94475735-94476385	650	*SERPINA9*	5	5	1.01E-03
chr4:177442253-177442903	650	*AGA*	3	3	1.50E-03
chr17:58331075-58331725	650	*MIR142*	4	4	1.14E-02
chr19:17776226-17776876	650	*FCHO1*	3	3	1.17E-02

Figure 5.1: Non-coding drivers identified in 80 primary tumours. Recurrently mutated promoters (a) and *cis*-regulatory elements (b). Plots annotated with the target genes of the most significantly mutated non-coding elements. Effect of CRE focal deletion on expression of (c) *BIRC2* (n = 2 vs n = 9) and (d) *IGLL5* (n = 2 vs n = 9). Boxplots show gene expression in tumours with and without copy number alterations. **: *Q* < 0.001. Del, deletion; CRE: *cis*-regulatory element.

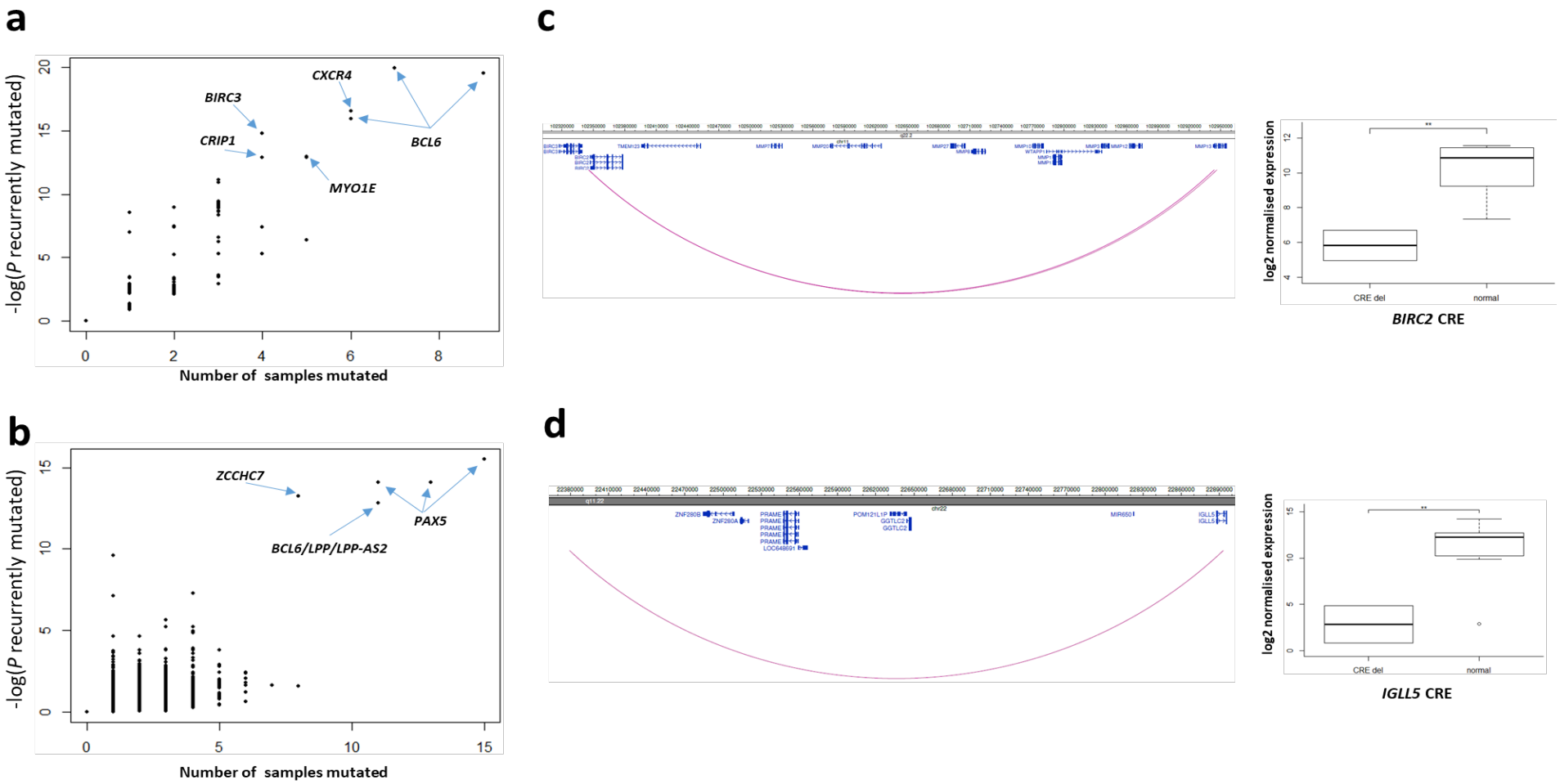


Figure 5.2: Chromothripsis events in primary tumours. Chromothripsis events detected in samples (a) 6016, (b) 9166, and (c) 7801. Each block of diagram represents chromothripsis event at individual chromosome. For each block, the top panel indicates genomic location of the chromothripsis event, the middle panel shows consensus structural variants, and the bottom panel shows total copy number calls for the genomic

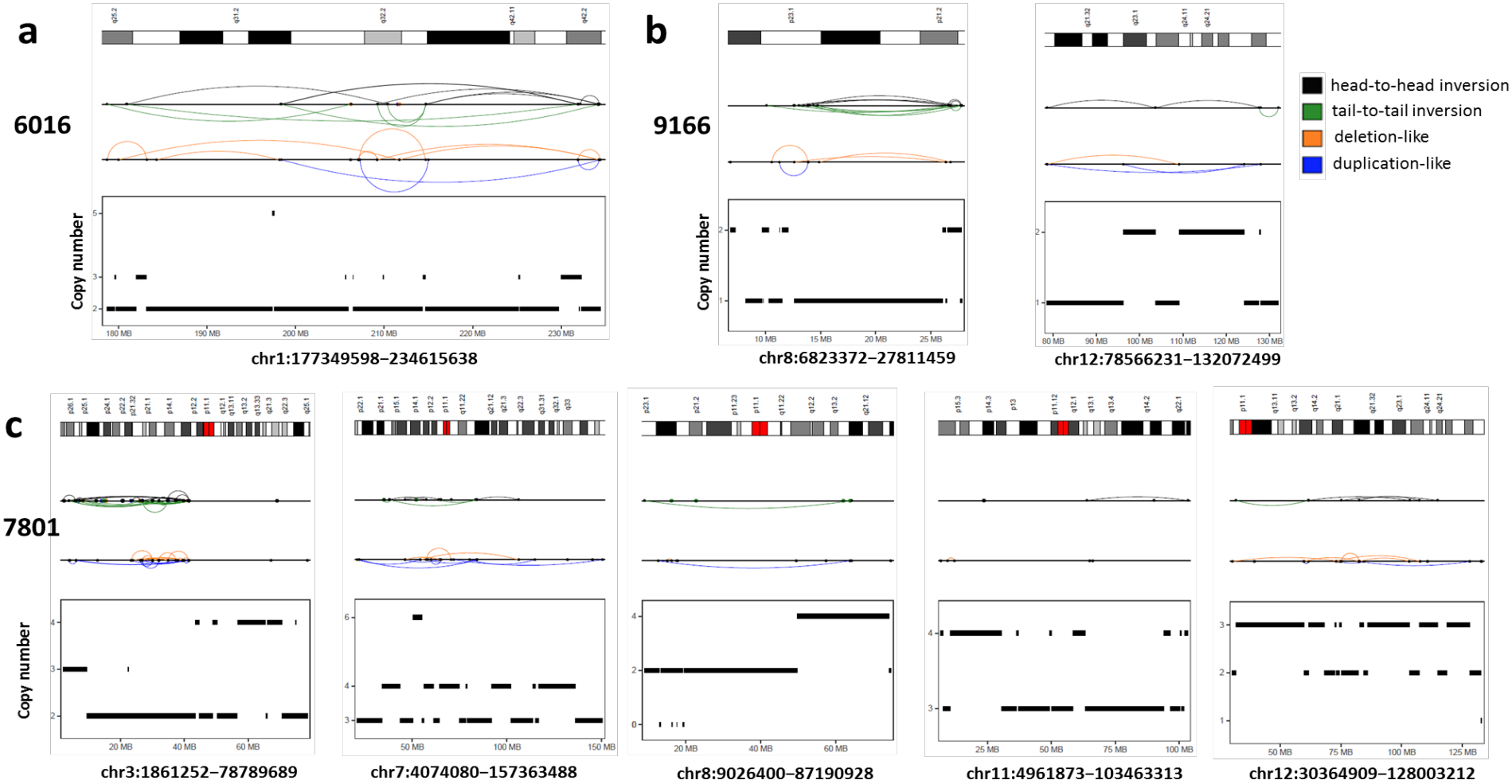
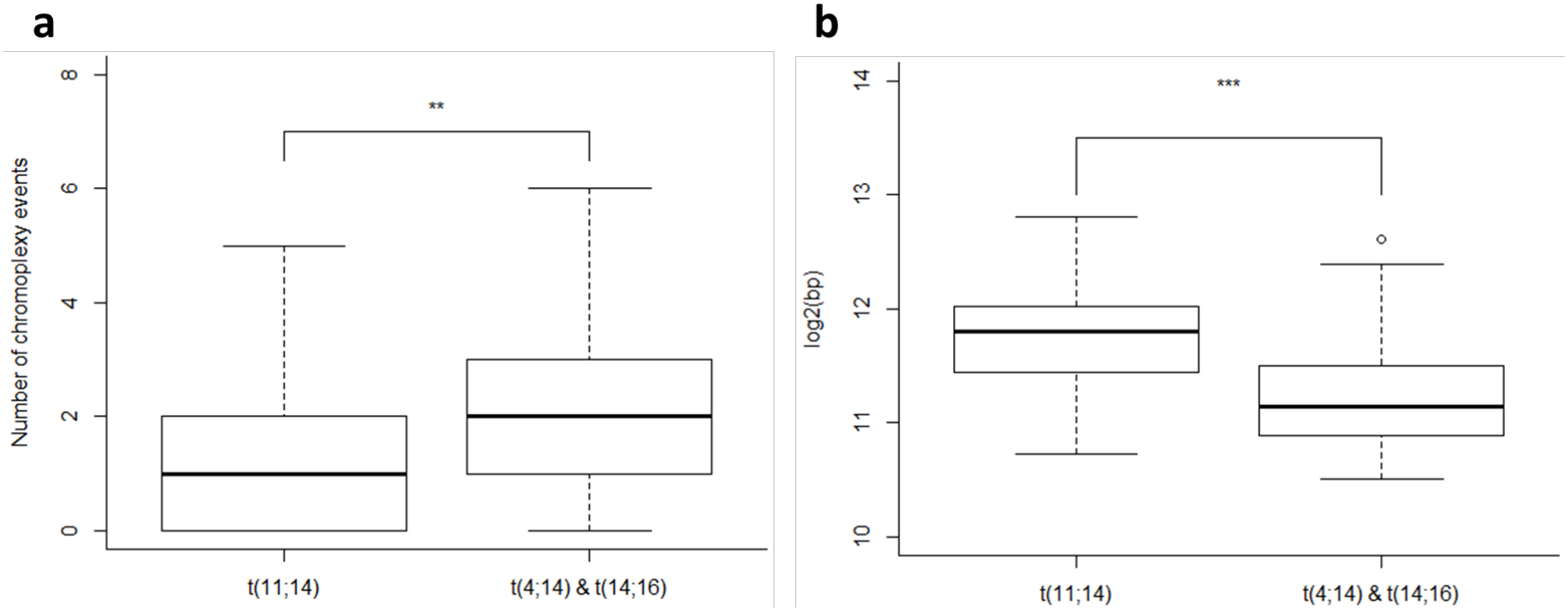


Table 5.4: Frequency of coding drivers disrupted by chromoplexy. SV, structural variant.

Driver gene	Number of samples affected by chromoplexy	Number of samples affected by non-chromoplexy SVs	Driver gene	Number of samples affected by chromoplexy	Number of samples affected by non-chromoplexy SVs
SP140	12	7	*PRKD2*	1	1
DUSP2	11	6	*DNAH5*	1	1
SF3B1	12	7	*BMP2K*	2	2
IDH1	12	7	*ZNF208*	0	0
NRAS	5	3	*RPL10*	0	0
DIS3	5	3	*FBXO4*	2	2
TRAF3	6	4	*RASA2*	3	3
MAX	5	3	*OR5M1*	0	0
TGDS	5	3	*PTH2*	1	1
TBC1D29	2	0	*BAX*	1	1
FCF1	5	3	*CELA1*	4	4
TRAF2	5	3	*FTL*	1	1
PABPC1	2	0	*OR9G1*	0	0
SGPP1	5	3	*TNFSF12*	1	1
UBR5	2	0	*FAM154B*	0	0
NF1	2	0	*HIST1H4H*	2	2
TET2	4	2	*LEMD2*	2	2
NFKBIA	4	2	*RPN1*	3	3
ZFP36L1	5	3	*HUWE1*	0	0
BRAF	1	0	*ZNF292*	4	4
RB1	3	2	*KLHL6*	2	2
ACTG1	3	2	*MLL3*	0	0
PTPN11	6	5	*ARID1A*	1	1
MYH2	2	1	*CREBBP*	0	0
RPS3A	3	2	*KMT2B*	0	0
C8orf86	2	1	*ATRX*	0	0
KMT2C	1	0	*SETD2*	2	2
EP300	3	2	*RFTN1*	0	0
XBP1	3	2	*DNMT3A*	1	1
NCOR1	2	1	*KDM5C*	0	0
C8orf34	1	0	*KDM6A*	0	0
KRAS	7	7	*ARID2*	4	4
FAM46C	0	0	*FUBP1*	2	2
TP53	1	1	*MAF*	1	1
PRDM1	4	4	*CDKN1B*	6	6
EGR1	3	3	*MAN2C1*	1	1
ATM	1	1	*NFKB2*	2	2
CCND1	1	1	*ABCF1*	2	2
LTB	2	2	*MAML2*	2	2
IRF4	0	0	*CDKN2C*	2	2
FGFR3	0	0	*MAFB*	1	1
CYLD	1	1	*PIK3CA*	2	2
ATR	3	3	*IDH2*	1	1
SAMHD1	1	1			

Figure 5.3: Comparison of (a) number of chromoplexy events and (b) telomere lengths between subtypes. Boxplots show (a) number of chromoplexy events and (b) $\log_2$ of telomere length base pairs (bp) of high-risk subtypes t(4;14) and t(14;16) versus lower-risk t(11;14). **: $P < 0.01$, ***: $P < 0.001$.



5.3.2 Chronology of mutational events in primary tumours

By integrating somatic mutations and copy number profiles, the relative timing of important molecular alterations in MM was inferred. Mutations of *CCND1*, *MAX*, *PRKD2*, *DIS3*, and *NRAS* were identified as early events whereas mutations of *KRAS*, *IRF4*, *FGFR3*, *TP53*, and *TET2* occurred as late events (Figure 5.4a). The most frequent large-scale CNVs were deletion of 13q (59%) or 1p (35%), and gain of 1q (46%). (Table 5.5, Appendix 4). Copy number neutral loss of heterozygosity (nLOH) at 13q was seen in 21% of tumours (Table 5.5). Aberrations of 13q was enriched in high-risk t(4;14) and t(14;16) MM ($P = 3.5 \times 10^{-5}$, OR = 16.2, Fisher's exact test). Chronological timing of major CNVs (present in $\geq$ 10% of total samples)[249] identified 21q gain, 22q nLOH, 19 gain, and 13q nLOH, and 1q nLOH as being early events (Figure 5.4b). In contrast to previous reports[250], 13q deletion was observed to be a subclonal event (Figure 5.4b). 1p deletion and 1q gain, which has been linked to patient prognosis were identified as occurring post 13q deletion (Figure 5.4b).

Figure 5.4: Chronology of (a) coding drivers and (b) major copy number events. Red dots denote mean of relative timing for each event with blue lines indicating 95% confidence intervals of the relative timing. Dotted red lines denote discrete clonality events. Frequency, number of tumours with each mutational event; Del, deletion; LOH, loss of heterozygosity.

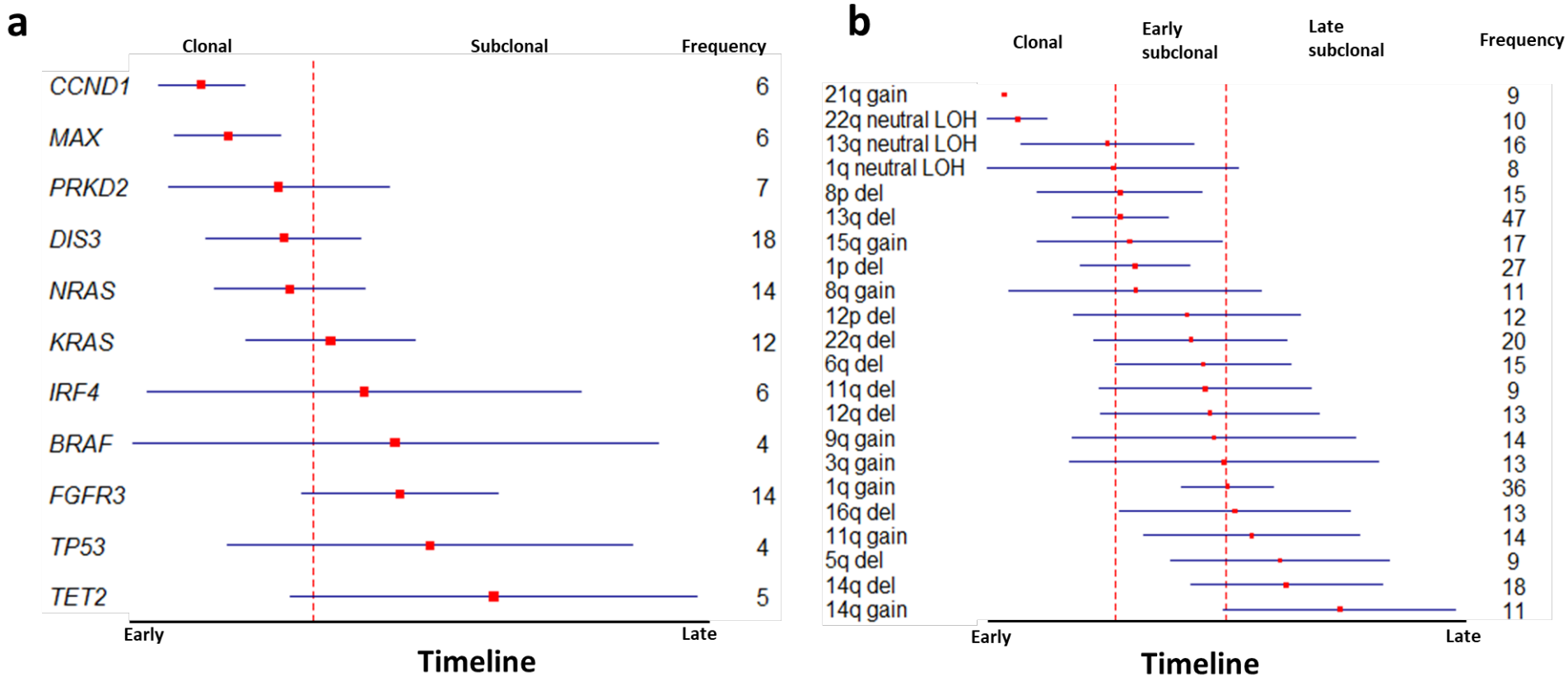


Table 5.5: Frequency of large-scale copy number alterations events in 80 primary tumours. Only events occur in at least 5 tumours are shown. LOH, loss of heterozygosity.

Chromosome arm events	No. of samples affected	Proportion(%)
13q deletion	47	59
1q gain	37	46
1p deletion	28	35
22q deletion	20	25
14q deletion	19	24
15q gain	18	23
13q neutral LOH	17	21
8p deletion	16	20
9q gain	15	19
6q deletion	15	19
11q gain	14	18
3q gain	14	18
12p deletion	13	16
16q deletion	13	16
12q deletion	13	16
14q gain	11	14
8q gain	11	14
22q neutral LOH	10	13
11q deletion	10	13
5q deletion	10	13
21q gain	9	11
1q neutral LOH	8	10
19 gain	8	10
3p gain	7	9
9p gain	7	9
2p deletion	7	9
18 gain	7	9
9 gain	7	9
4q gain	6	8
17p deletion	6	8
12q gain	6	8
2q deletion	6	8
1p neutral LOH	6	8
2p gain	6	8
20q gain	6	8
3 gain	6	8
1q deletion	5	6
6q gain	5	6
17q gain	5	6
4q deletion	5	6
6q neutral LOH	5	6
7p deletion	5	6
5p gain	5	6
1p gain	5	6
10q deletion	5	6
8q deletion	5	6

5.3.3 Mutational landscape of relapse

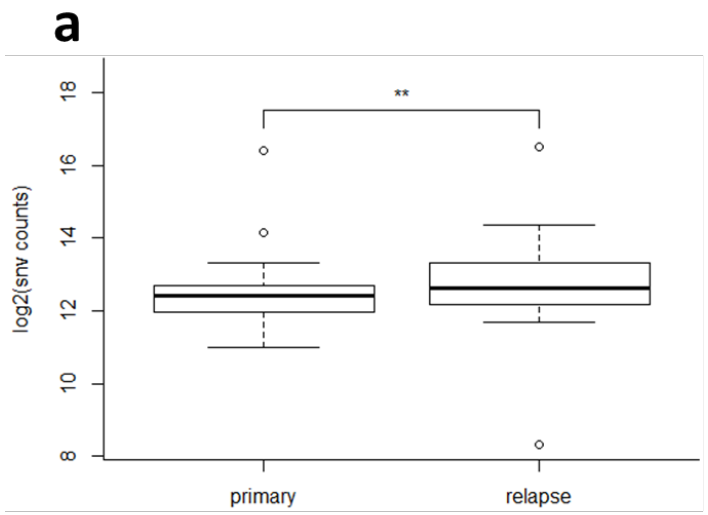
Following on from the analysis, the molecular features of MM relapse of the 25 primary-relapse pairs were investigated. Relapse was associated with a higher mutational burden than primary tumours (Figure 5.5a-b, $P < 0.01$, paired Wilcoxon rank-sum test). Varied proportions (9 - 99%) of SNVs and indels identified in primary tumours were not detectable at relapse (Figure 5.5c), suggesting eradication and heterogenous clonal dynamics of the respective clone. Despite the increased mutational burden, relapsed tumours did not exhibit significantly more kataegis (Figure 5.6, Table 5.6). Only one of the 25 relapsed tumours showed additional chromothripis (Figure 5.7). Although both primary and relapsed tumours had shorter telomeres compared to plasma cells ($P < 0.01$, paired Wilcoxon rank-sum test), relapse was associated with longer telomeres ($P = 3.4 \times 10^{-3}$) (Figure 5.8).

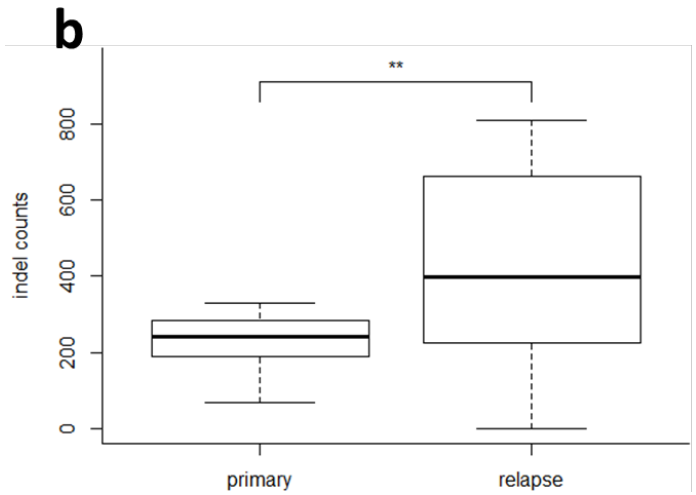
A translocation bringing the *IGH* loci in proximity to *MAP3K14* was gained at relapse in one tumour, which was associated with a six-fold upregulation of *MAP3K14* expression to primary tumour (Figure 5.9). Driver genes only mutated at relapse included *FAM46C*, *TRAF2*, *LTB*, *OR9G1*, *FAM154B*, *NF1*, *XBP1*, and *IDH2* (Figure 5.10). Other driver mutations acquired at relapse were those in *KRAS* and *NRAS* genes, detected in three and two tumours respectively. Extending the analysis to all coding genes, non-silent mutations frequently gained at relapse included those in *SYNE1*, *MTCL1*, *ABCA13*, *ADAMTS9*, and *ZNF521* (Table 5.7). As expected from tracking of driver mutations, the increase in CCF of *TET2*, *ZNF292*, *MYH2*, and *DNAH5* mutations implied selection of subclones (Figure 5.11). The promoters and CREs of an additional 16 genes were significantly mutated at relapse including genes with established roles in the biology of MM or other B-cell malignancies such as *XBP1*, *BCL7A*, and *BCL9* (Table 5.8, Table 5.9).

Relapse was associated with additional CNVs, notably for 13q and 17p deletions (Figure 5.12a, Appendix 5). In addition, subclonal 22q deletion at diagnosis emerged as clonal at relapse (Figure 5.13). Other relapsed CNV-associated changes, which occurred at pre-existing unstable genomic regions, include the progression of nLOH to LOH and LOH to complete deletion at 13q; as well as further copy number gains at 1q and 10p (Figure 5.12b-c, Figure 5.14). High-risk

t(4;14) and t(14;16) MM are associated with higher increased number of CNV events at relapse compared to t(11;14) (Appendix 5), consistent with previous observation[105].

Figure 5.5: Mutational burdens in primary versus relapse tumours. Boxplots show (a) $\log_2$ of point mutation counts and (b) indel counts in primary and matched relapsed tumours. (c) Proportions of shared, relapse-specific and primary-specific mutations across samples. **, $P < 0.01$.





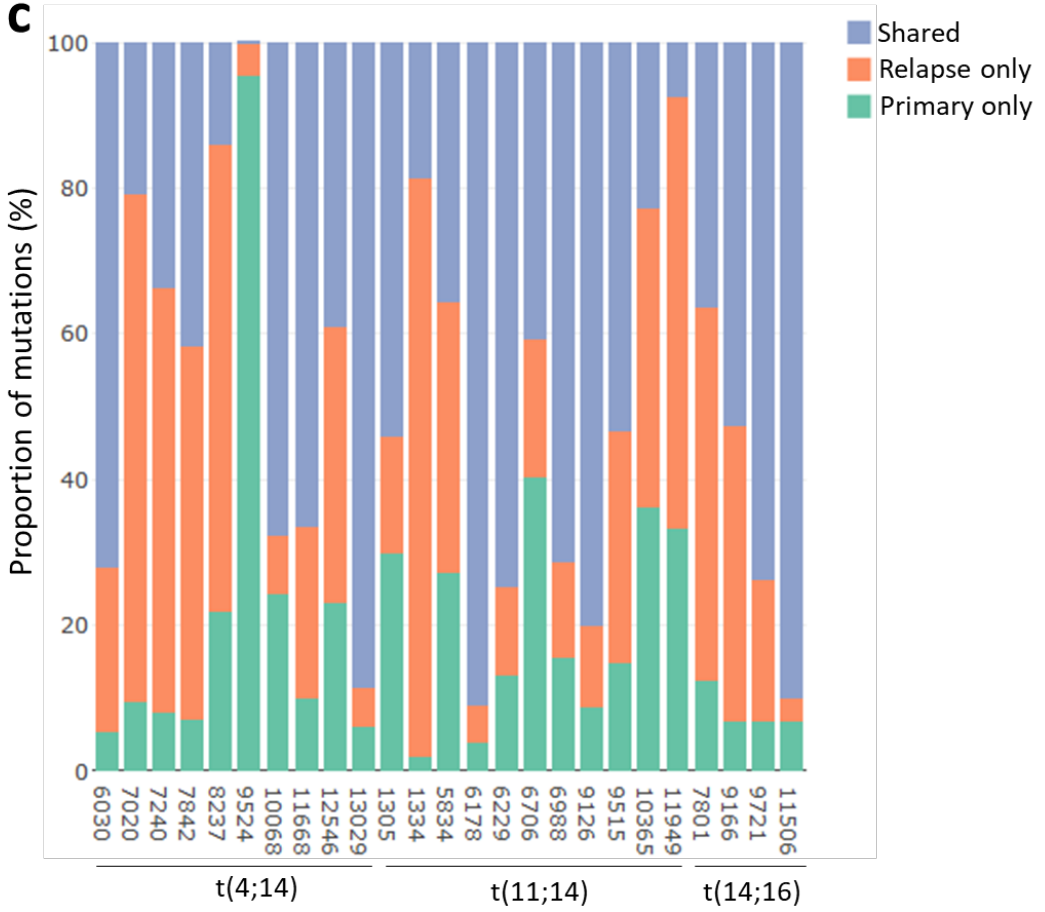


Figure 5.6: Kataegis events in primary versus relapse. (a) Circos plot summarising kataegis foci detected in 25 primary (inner circle) and their matched relapse tumours (outer circle). Each dot represents a kataegis event, positioned by distinct samples based on height and genomic location based on width of the circle. (b) Boxplot show number of kataegis events detected in per primary versus relapse tumours. ns, not significant.

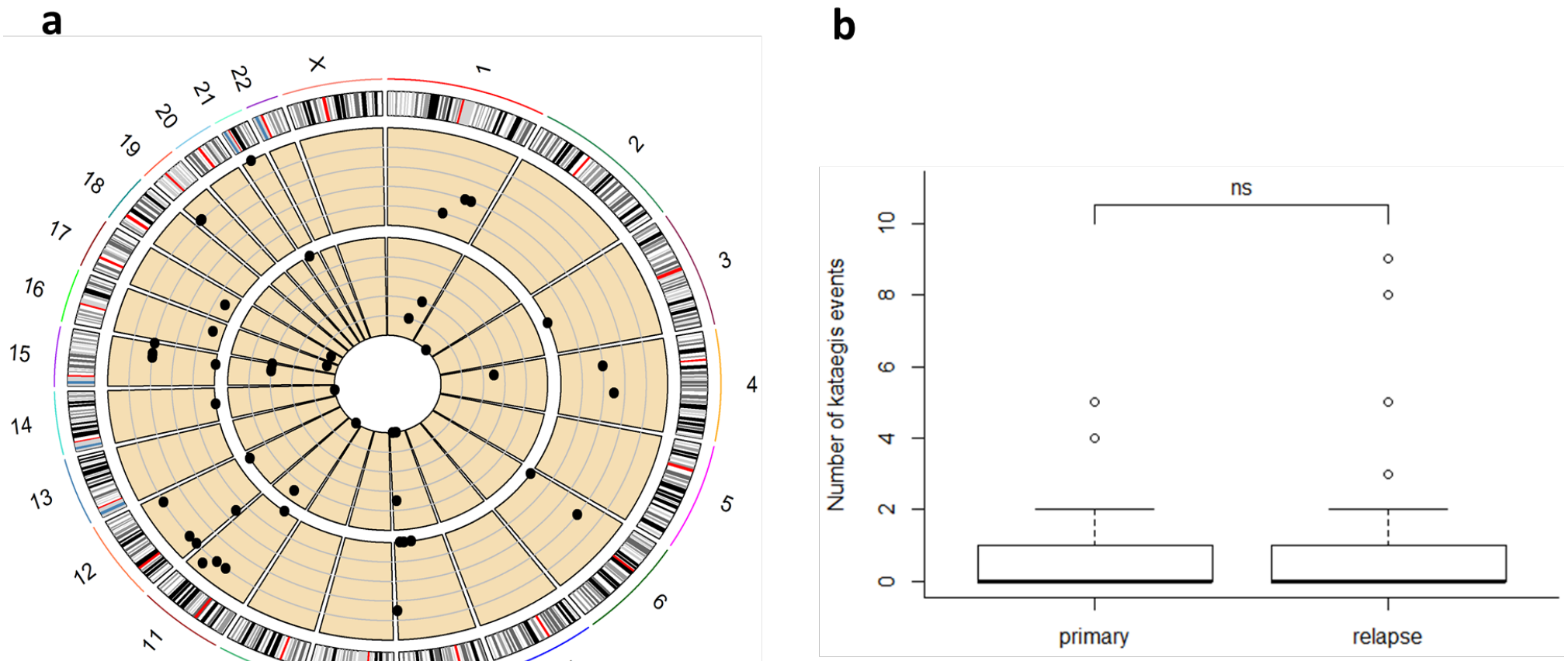


Table 5.6: Kataegis foci for 25 (a) primary and (b) matched relapsed tumours.

a) Primary tumours

Sample	Chromosome	Start	End	Chromosome arm	Length(bp)	No. of mutations
1305	chr4	46010942	46012272	4p	1330	6
6178	chr6	14925742	14926804	6p	1062	6
7240	chr15	67693030	67694400	15q	1370	8
7240	chr15	74771964	74776187	15q	4223	12
7240	chr15	77047614	77048970	15q	1356	7
7240	chr15	101347313	101349535	15q	2222	11
7842	chr8	83722033	83725013	8q	2980	8
9126	chr11	69638929	69641462	11q	2533	11
9721	chr12	110870086	110873251	12q	3165	8
9721	chr21	17796091	17797632	21q	1541	7
10365	chr2	154804979	154806162	2q	1183	6
10365	chr8	116656213	116659447	8q	3234	9
10365	chr8	128250536	128251990	8q	1454	7
10365	chr11	49488216	49489172	11p	956	6
10365	chr14	55618500	55622472	14q	3972	10
11506	chr16	56837018	56838397	16q	1379	6
11506	chr17	45288300	45290340	17q	2040	7
11668	chr1	188776810	188778426	1q	1616	7
13029	chr1	147661939	147664919	1q	2980	8

b) Relapsed tumours

Sample	Chromosome	Start	End	Chromosome arm	Length(bp)	No. of mutations
1305	chr4	46010942	46012271	4p	1330	6
7240	chr6	36250372	36254274	6p	3903	12
7240	chr15	67693030	67694624	15q	1595	10
7240	chr15	74771964	74776186	15q	4223	13
7240	chr15	77047614	77048969	15q	1356	8
7240	chr15	101347313	101349534	15q	2222	11
7842	chr8	144267366	144268884	8q	1519	6
7842	chr8	144272319	144275023	8q	2705	7
8237	chr11	92135792	92137526	11q	1735	6
8237	chr11	111009010	111011030	11q	2021	8
8237	chr12	4955664	4958055	12p	2392	9
8237	chr12	4979157	4980818	12p	1662	11
8237	chr12	5013623	5015268	12p	1646	7
8237	chr12	25465952	25466877	12p	926	6
8237	chr19	5010478	5012392	19p	1915	7
8237	chr19	5791420	5792816	19p	1397	6
8237	chr19	8131086	8132082	19p	997	7
9126	chr11	69638929	69641461	11q	2533	11
9721	chr21	17796091	17797631	21q	1541	7
10365	chr3	96980348	96981090	3q	743	6
10365	chr6	12531671	12533555	6p	1885	6
10365	chr8	94948159	94950081	8q	1923	8
10365	chr8	116656213	116659446	8q	3234	9
10365	chr8	127987546	127990685	8q	3140	6
10365	chr11	49487717	49489171	11p	1455	7
10365	chr14	55618500	55622471	14q	3972	10
10365	chr15	67132673	67137210	15q	4538	19
11506	chr16	56837018	56839208	16q	2191	7
11506	chr17	45288300	45290339	17q	2040	7
11668	chr1	188776810	188778335	1q	1526	6
12546	chr12	116285797	116286657	12q	861	6
13029	chr1	147661939	147664918	1q	2980	8
13029	chr1	204316137	204317777	1q	1641	6
13029	chr4	115478347	115479773	4q	1427	6

Figure 5.7: Additional chromothripsis events detected in relapsed tumour. Chromothripsis previously unidentified in primary detected in relapse tumour sample 7842. Each block of diagram represents chromothripsis event at individual chromosome. For each block, the top panel indicates genomic location of the chromothripsis, the middle panel shows consensus structural variants, and the bottom panel shows total copy number calls for the genomic region.

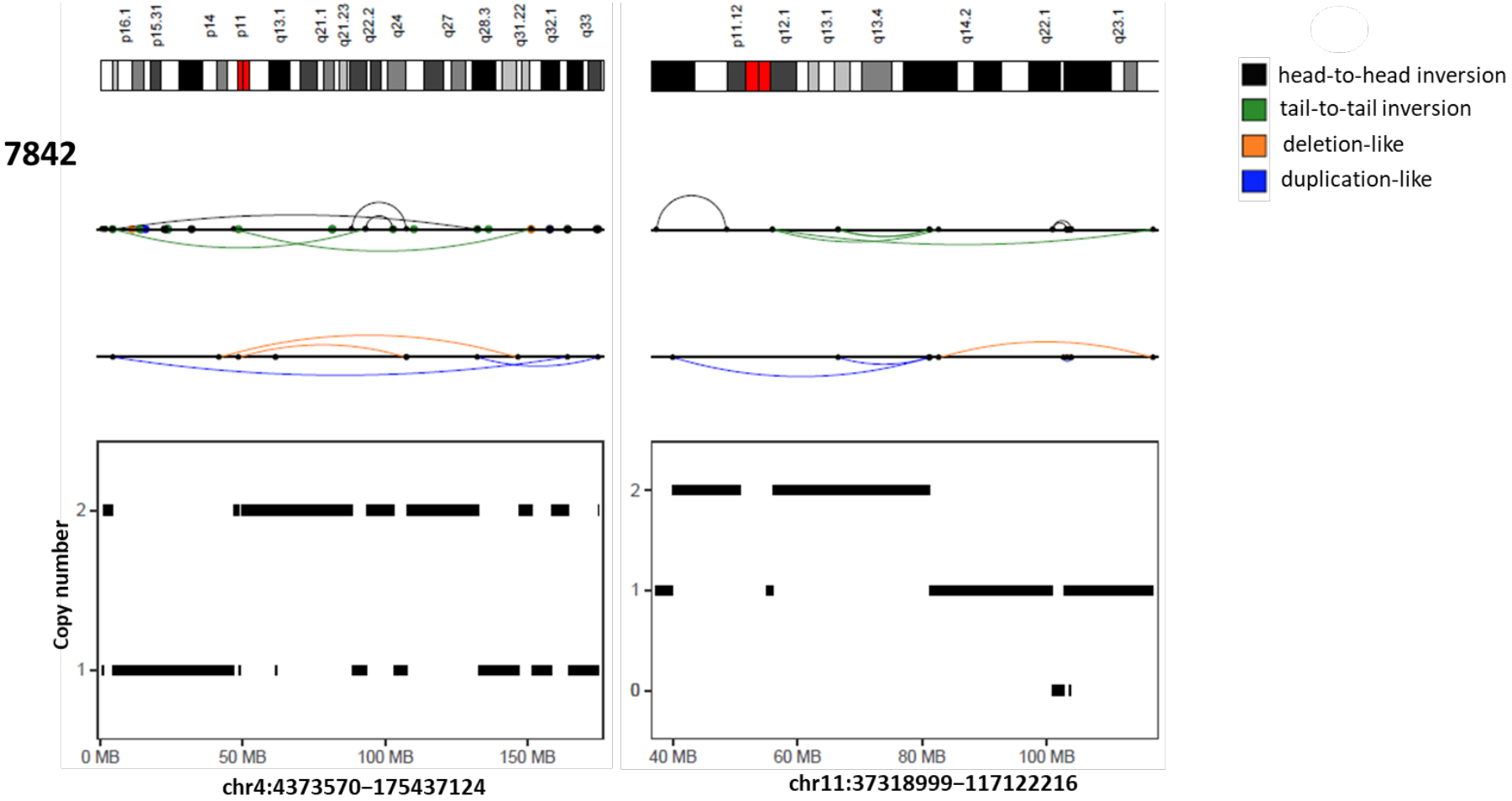


Figure 5.8: Telomere length comparison. Boxplots show $\log_2$ (base pair) of telomere lengths of 25 matched normal, primary, and relapse samples. **, $P < 0.01$; ***, $P < 0.001$.

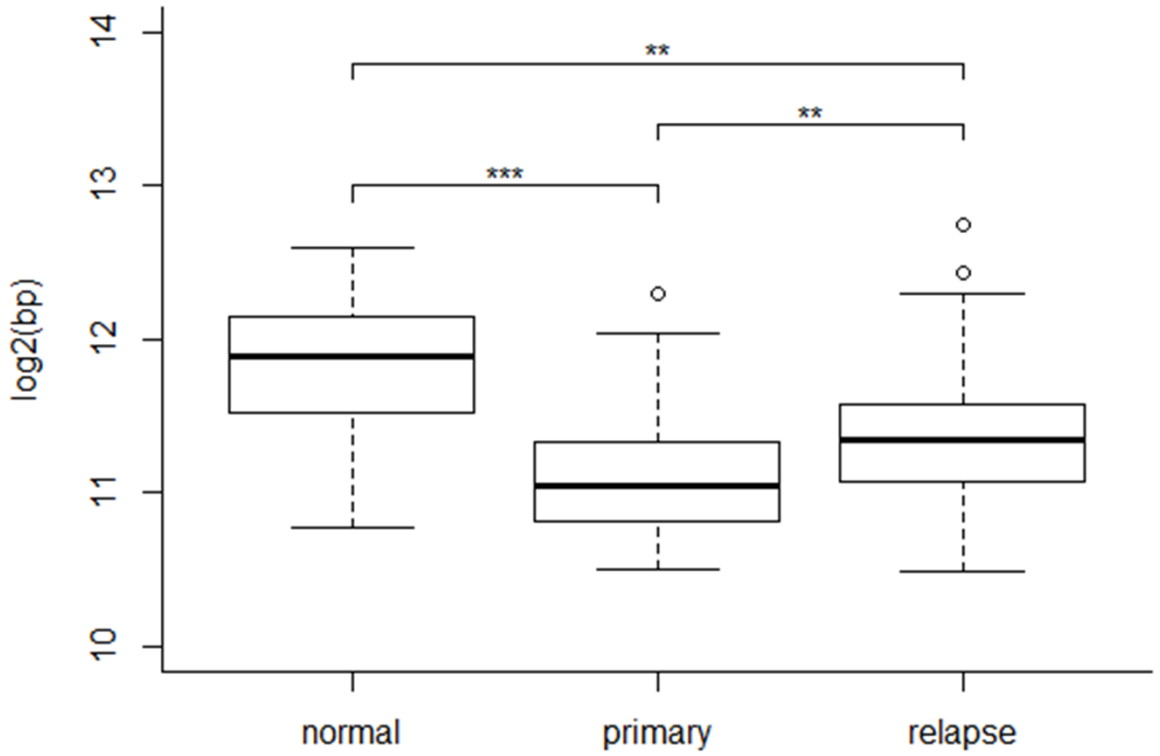


Figure 5.9: Acquisition of chromosomal translocation in proximity to *MAP3K14* at relapse in sample 8237. Upper panel shows relative location of chromosomal translocation to *MAP3K14*. Lower panels show IGV screenshots indicating *de novo* acquisition of chromosomal translocation (14q32;17q21) at relapse (right panel) not present in primary (left panel).

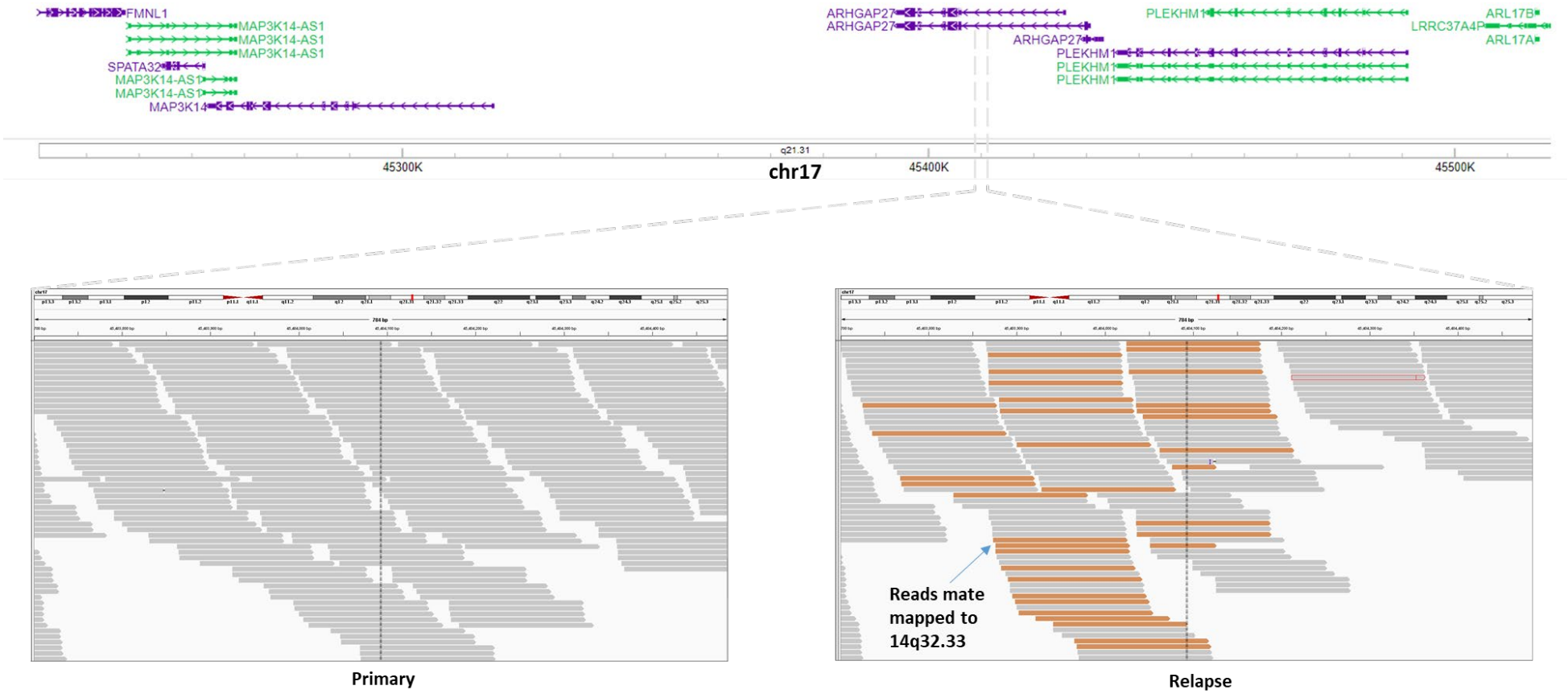


Figure 5.10: Non-silent single nucleotide variants and indels disrupting established driver genes, and established translocations, in primary and matched relapsed tumours.

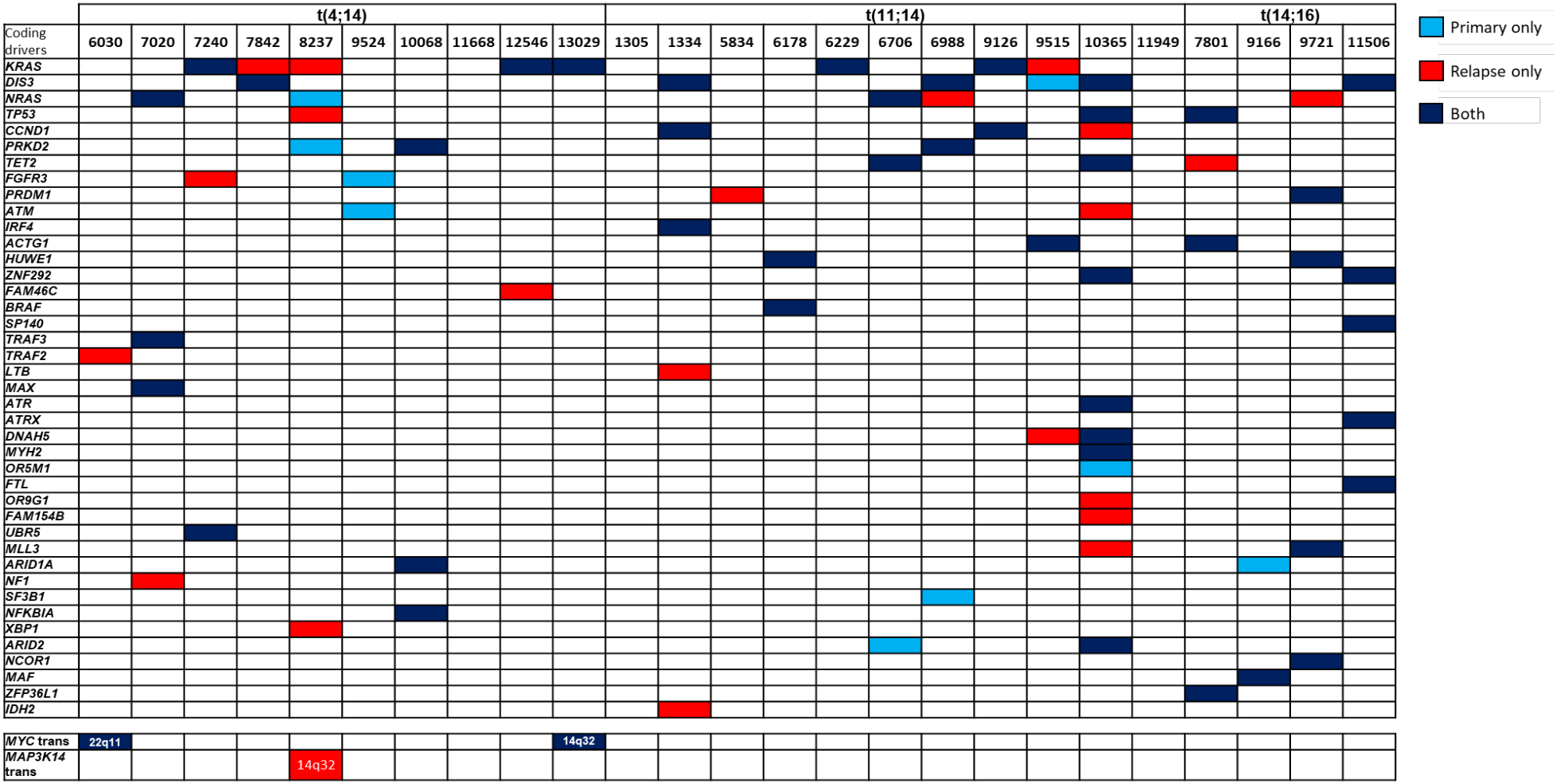


Table 5.7: Net increase in number of non-silent coding mutations in relapse.
Only genes additionally acquired in at least 2 tumours are shown.

Genes	No. of samples mutated in primary	No. of samples mutated in relapse	Net increase
SYNE1	0	5	5
MTCL1	0	3	3
KRAS	5	8	3
ABCA13	1	3	2
ADAMTS9	0	2	2
C1orf168	0	2	2
C1orf27	0	2	2
C2orf16	1	3	2
CCDC108	0	2	2
CD163L1	0	2	2
CENPF	0	2	2
CEP295	0	2	2
CHD6	0	2	2
CHD7	0	2	2
CPZ	0	2	2
CRYBG3	0	2	2
CUL3	0	2	2
EFCAB5	1	3	2
GALNT13	0	2	2
GLOD4	0	2	2
GPR75	0	2	2
HCN2	0	2	2
HEATR7A	0	2	2
HELZ	0	2	2
HERC2	0	2	2
IFNA21	0	2	2
IGLL5	0	2	2
ITGB3	0	2	2
JPH4	0	2	2
KIAA0947	0	2	2
LAMA1	0	2	2
LAMA2	0	2	2
LTK	0	2	2
MDN1	0	2	2
MS4A12	0	2	2
MYO18B	1	3	2
MYO9B	0	2	2
NID1	0	2	2
PTPN23	0	2	2
RAPGEF1	0	2	2
RGS3	1	3	2
SCG2	0	2	2
SEMA4D	0	2	2
SI	0	2	2
TARBP1	0	2	2
TNKS	0	2	2
TTN	1	3	2
UHRF1BP1L	0	2	2
ZNF460	0	2	2
ZNF521	0	2	2
RAGE	0	2	2

Figure 5.11: Cancer cell fractions (CCFs) of coding driver genes in primary and relapsed tumours. Each dot represents a non-silent mutation in a driver gene. Relationships between CCF of a driver gene mutation in primary and relapse are indicated by the lines linking them. Genes with a large increase in CCF at relapse (*i.e.* clonal expansion of subclones carrying the mutations) are annotated by symbol.

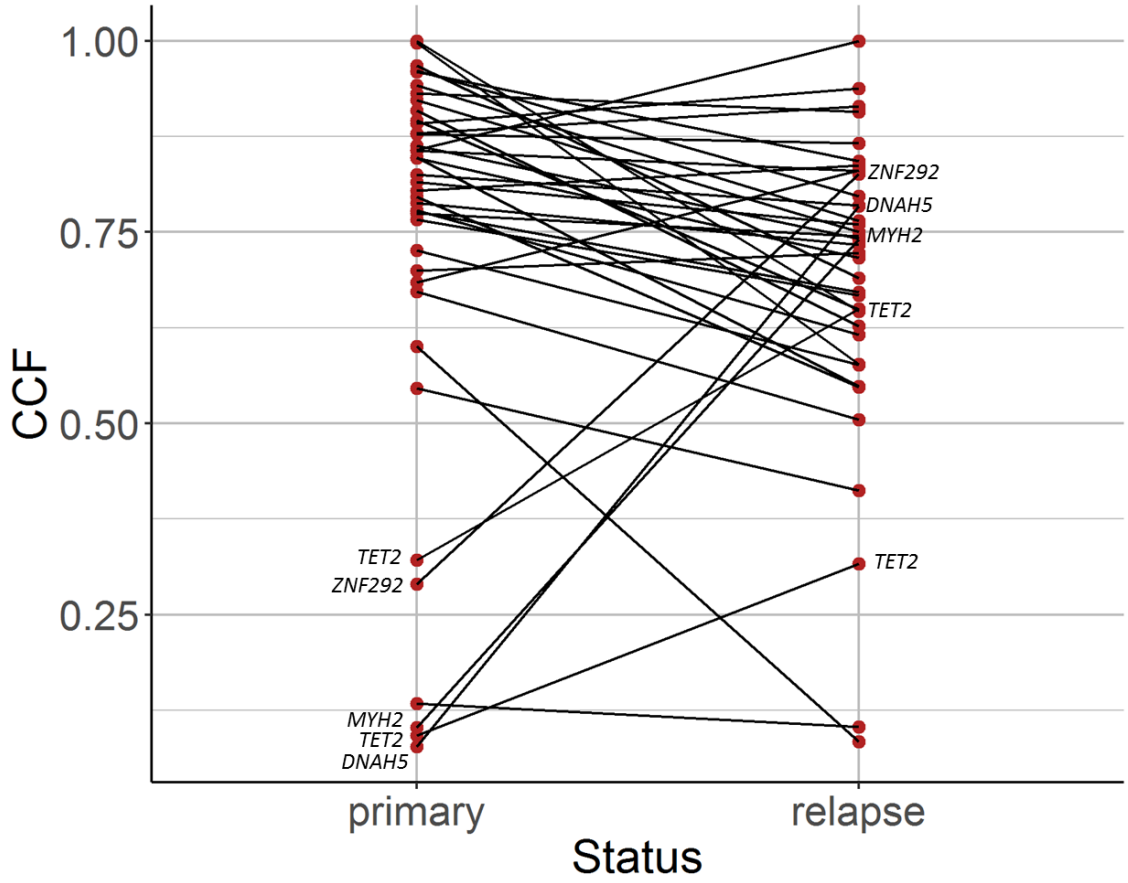


Table 5.8: Significantly mutated promoters in 25 relapsed tumours. (Q < 0.05). In **bold**, genes additionally significantly mutated at relapse.

Fragment	Size (bp)	Target gene	Number of mutations	Number of mutated samples	Q-value
chr2:136117487-136118137	650	*CXCR4*	7	3	4.80E-13
chr1:25820428-25821078	650	***MTFR1L***	5	3	3.79E-10
chr1:25820362-25821012	650	***MTFR1L***	5	3	3.72E-09
chr19:49473807-49474457	650	*FLT3LG*	3	3	3.73E-07
chr19:49473828-49474478	650	*FLT3LG*	3	3	3.73E-07
chr19:49473834-49474484	650	*FLT3LG*	3	3	3.73E-07
chr19:49473836-49474486	650	*FLT3LG*	3	3	3.73E-07
chr19:49473852-49474502	650	*FLT3LG*	3	3	3.73E-07
chr3:159988350-159989000	650	***IL12A***	3	3	3.80E-07
chr15:89334567-89335217	650	***POLG***	3	3	4.16E-07
chr15:89334597-89335247	650	***POLG***	3	3	4.16E-07
chr15:89334611-89335261	650	***POLG***	3	3	4.16E-07
chr22:28800319-28800969	650	***XBP1***	3	3	4.18E-07
chr22:28800322-28800972	650	***XBP1***	3	3	4.18E-07
chr22:28800347-28800997	650	***XBP1***	3	3	4.18E-07
chr3:161104890-161105540	650	***B3GALNT1***	3	3	8.03E-07
chr12:38316439-38317089	650	***ALG10B***	3	3	1.03E-06
chr12:38316362-38317012	650	***ALG10B***	3	3	1.11E-06
chr12:38316367-38317017	650	***ALG10B***	3	3	1.11E-06
chr22:40950947-40951597	650	***RBX1***	3	3	1.47E-05
chr22:40950959-40951609	650	***RBX1***	3	3	1.71E-05
chrX:12975258-12975908	650	***TMSB4X***	4	3	5.02E-05
chr15:59372293-59372943	650	*FAM81A*	3	3	7.13E-05
chr15:59372277-59372927	650	*MYO1E*	3	3	7.34E-05
chr16:29925986-29926636	650	***KCTD13***	3	3	6.33E-03
chr12:122021486-122022136	650	***BCL7A***	3	3	6.41E-03
chr16:29925962-29926612	650	***KCTD13***	3	3	6.55E-03
chr16:29925959-29926609	650	***KCTD13***	3	3	6.59E-03
chrX:17737049-17737699	650	***SCML1***	3	3	3.54E-02
chrX:17737068-17737718	650	***SCML1***	3	3	3.55E-02
chrX:17737069-17737719	650	***SCML1***	3	3	3.55E-02
chrX:17737151-17737801	650	***SCML1***	3	3	3.69E-02
chrX:17737325-17737975	650	***SCML1***	3	3	3.74E-02

Table 5.9: Recurrently mutated *cis*-regulatory elements in 25 relapsed tumours. (Q < 0.05). In **bold**, genes additionally significantly mutated at relapse.

Fragment	Size	Target gene	Number of mutations	Number of mutated samples	Q-value
chr15:74772989-74775174	2185	*CSK;FAM219B;MPI;SEMA7A*	7	3	1.47E-06
chr9:37406897-37411656	4759	*AL161781.2;PAX5*	7	5	1.31E-04
chr6:154713487-154721838	8351	***SCAF8***	8	3	2.12E-04
chr17:68772663-68776750	4087	***ABCA10;ABCA5***	5	5	7.14E-04
chr7:44631701-44638839	7138	***H2AFV;LINC01952***	6	3	1.98E-02

Figure 5.12: Copy number alterations associated with relapse. (a) Net change of CNV frequency in primary and matched relapse tumours; red and blue bars represent positive and negative changes respectively. (b) Copy number profiles of patients 7842, 9166 and 9515. In 7842 copy number neutral loss of heterozygosity (nLOH) at 13q becomes LOH at relapse. In 9166 LOH at 13q progresses to complete loss of 13q. In 9515 copy number gain at chromosome 10 and 11 progresses to additional chromosome gain. Thick and thin lines represent clonal and subclonal copy number states respectively. Yellow and blue lines denote total and minor copy number respectively. Blue arrows indicate regions with copy number change at relapse (copy number states > 5 not shown). (c) Patterns of copy number change across paired primary-relapse samples at 1q, 10p, and 13q. Lines indicate relationship between primary and matched relapse tumours, with width being proportional to event frequency. Only chromosome arms with CNVs are plotted, with a copy number of 2 corresponding to nLOH.

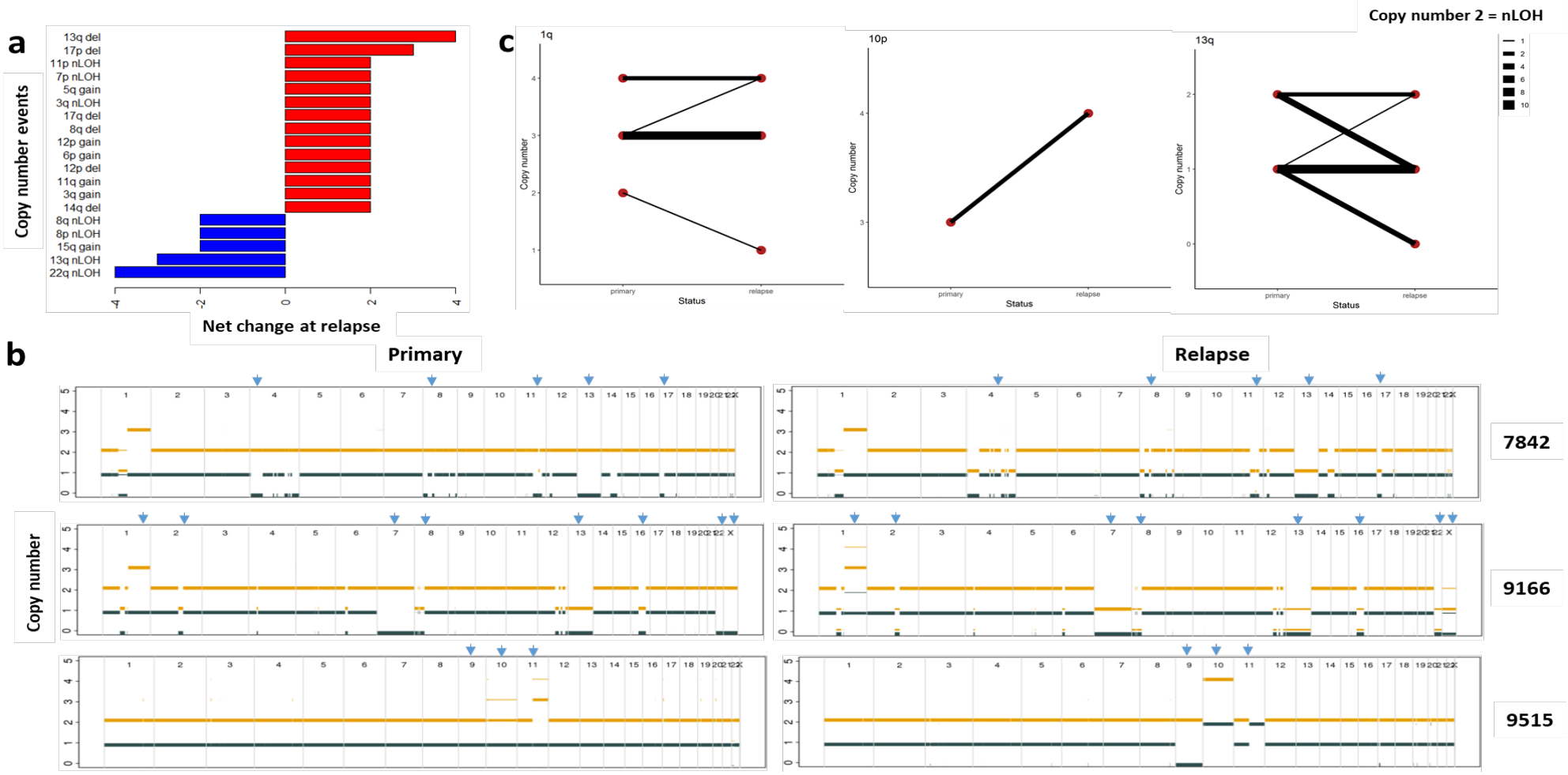

a
Copy number events
13q del
17p del
11p nLOH
7p nLOH
5q gain
3q nLOH
17q del
8q del
12p gain
6p gain
12p del
11q gain
3q gain
14q del
8q nLOH
8p nLOH
15q gain
13q nLOH
22q nLOH
Net change at relapse
c
Copy number 2 = nLOH
1q
10p
13q
Copy number
Status
primary
relapse
b
Primary
Relapse
Copy number
7842
9166
9515

Figure 5.13: Cancer cell fractions (CCF) of major chromosome arm events in primary and relapse. The number above each bar indicates the number of patients having the chromosome arm event. Only major chromosome arm events occurring in at least 4 primary and relapse tumours are considered. Del, deletion.

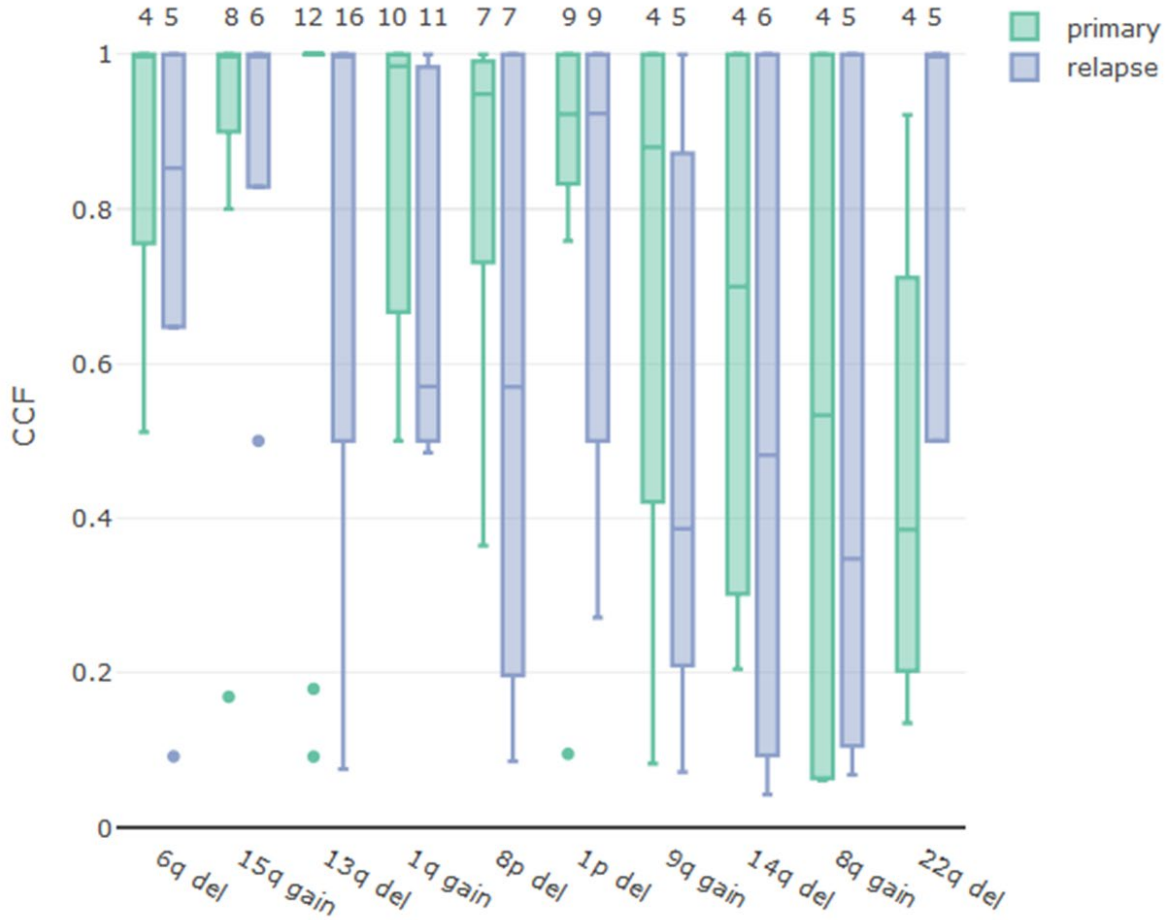
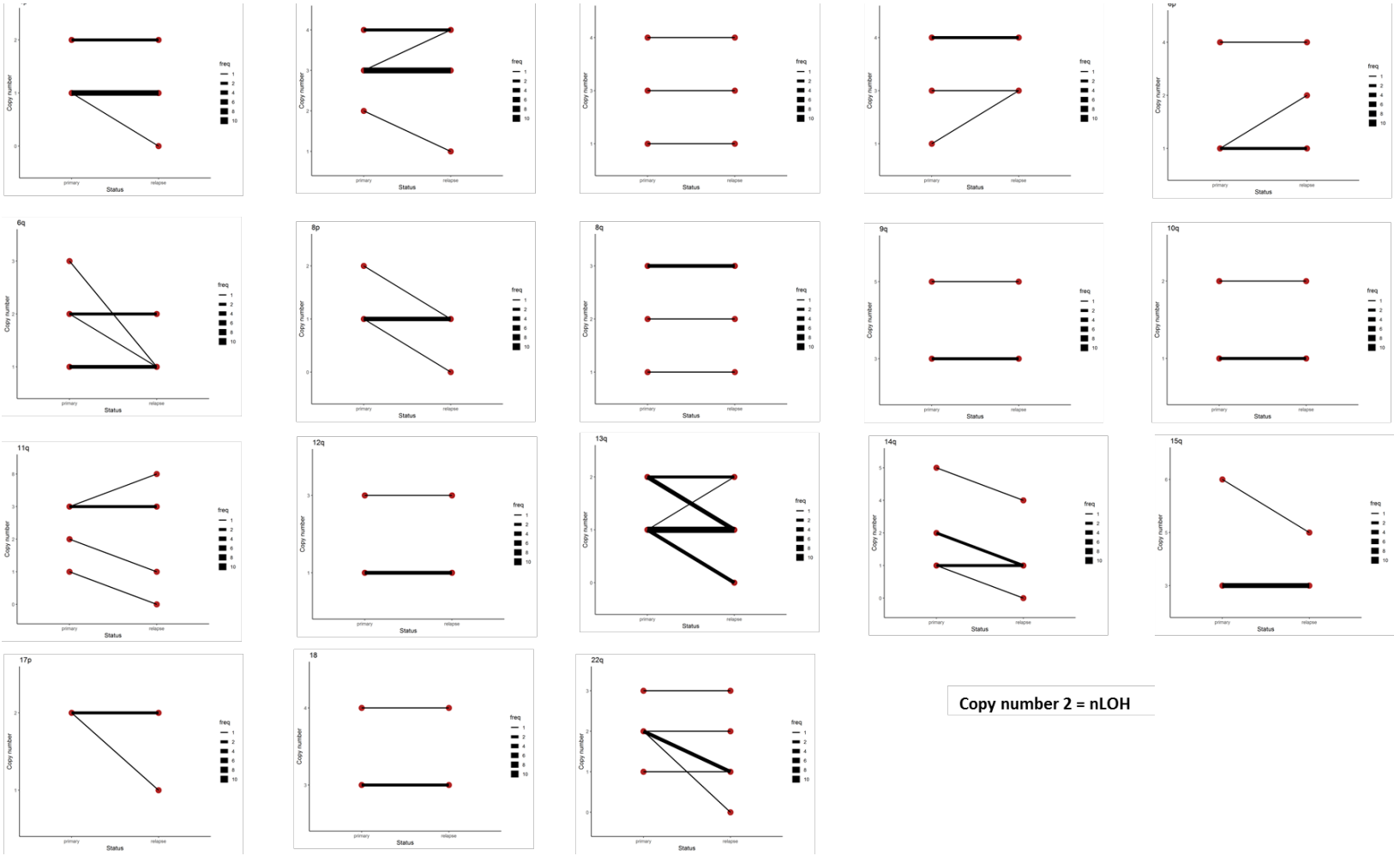


Figure 5.14: Patterns of major copy number changes in primary and relapsed tumours. Lines connecting dots indicate relationship between primary and matched relapse tumours. The intensity of lines is proportional to frequency (freq) of events. Only chromosomes or chromosome arms with copy number variations are plotted, thus copy number of 2 is copy number neutral loss of heterozygosity (LOH).



5.3.4 Mutational processes active at relapse

At diagnosis the major mutational signatures in tumours were those indicative of aging (COSMIC Signature 5), AID/APOBEC (COSMIC Signatures 2, 9, and 13) and DNA repair deficiency (COSMIC Signatures 3, 5, and 8) in MM[87, 127, 224-226, 251] (Figure 5.15, Figure 5.16). At relapse, the increased mutational burden was associated with increased APOBEC activity and DNA repair deficiency signatures (Figure 5.17). An increased C•G>G•C transversion rate in relapse-specific mutations was observed (Q = 0.015, paired Wilcoxon rank-sum tests) (Figure 5.18), a feature previously reported in relapsed acute myeloid leukaemia[252]. Additionally, a novel signature (M1) was identified at relapse, primarily in one patient, characterised by C•G>T•A mutations, which has been associated with alkylating agents[253], and thymidine mutations at specific contexts (Figure 5.19**,** Table 5.10).

5.3.5 Evolutionary trajectories of relapse

Three patterns of clonal evolution were apparent at relapse (Figure 5.20). In Pattern 1 (3/25 patients), the dominant clone in primary survives treatment and gains additional mutations at relapse (Figure 5.20a, Figure 5.21a). Tumours with Pattern 1 are characterised with no change in clonal composition of the dominant clones, suggesting that they were potentially unaffected by treatment. Pattern 2 (4/25 patients) is featured by subclonal expansion whereby a subclone in the primary survives treatment and expands to become the dominant clone at relapse (Figure 5.20b, Figure 5.21b). I suspect these clones might have mutations (*e.g.* *TET2*, *ZNF292*, *MYH2*, *DNAH5*, 6q deletion) giving them survival and selective advantage. Pattern 3 (18/25 patients) is characterised by the emergence of new clones at relapse, accompanied by the disappearance or decline of primary clones (Figure 5.20c, Figure 5.21c). One patient (sample 9524) had no clonal mutations shared between the primary and the relapse tumour (Figure 5.21c); however, this observation may reflect low tumour purity (Appendix 3). The three patterns of clonal evolution were not associated with therapy or molecular karyotype. It was, however, of note that time to relapse was shorter with Pattern 2 (median 11.6 versus 19.3 months, P = 0.019, Wilcoxon rank-sum test).

Figure 5.15: *De novo* extraction of WGS single nucleotide variants signatures using non-negative matrix factorization algorithm in 80 primary tumours. (a) Summary of five de novo mutational signatures extracted. (b) Cosine similarity heatmap. *De novo* extracted mutational signatures are compared against 30 COSMIC mutational signatures. The colour code (0 to 1) represents the resemblance between each pair of

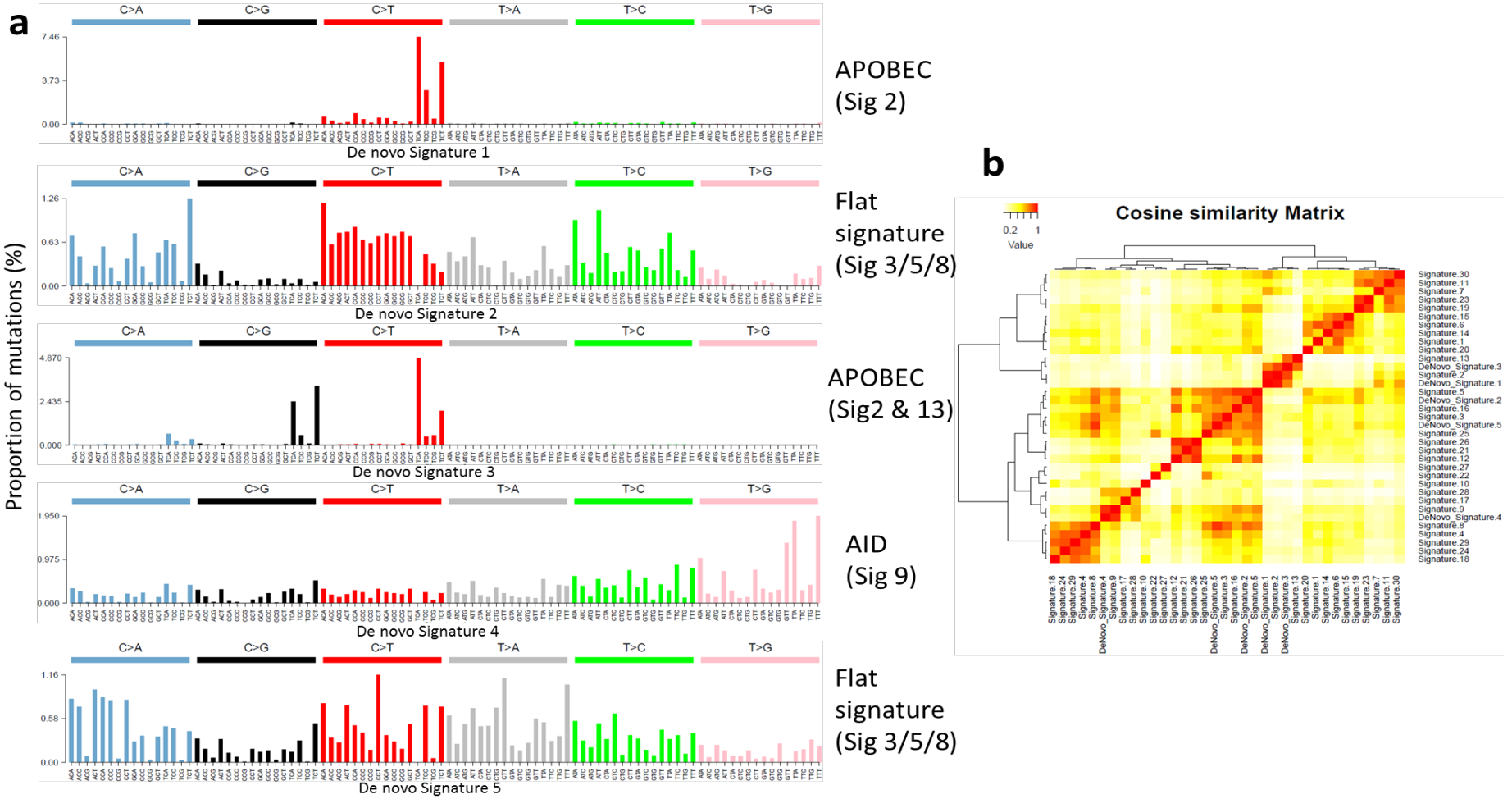


Figure 5.16: Mutational signatures contribution across 80 primary tumours. Mutational signatures contribution fitting from deconstructSig. Only major COSMIC mutational signatures extracted de novo were considered. APOBEC signature includes COSMIC signatures 2 and 13. Flat signature includes COSMIC signatures 3, 5, and 8. AID, activation-induced deaminase.

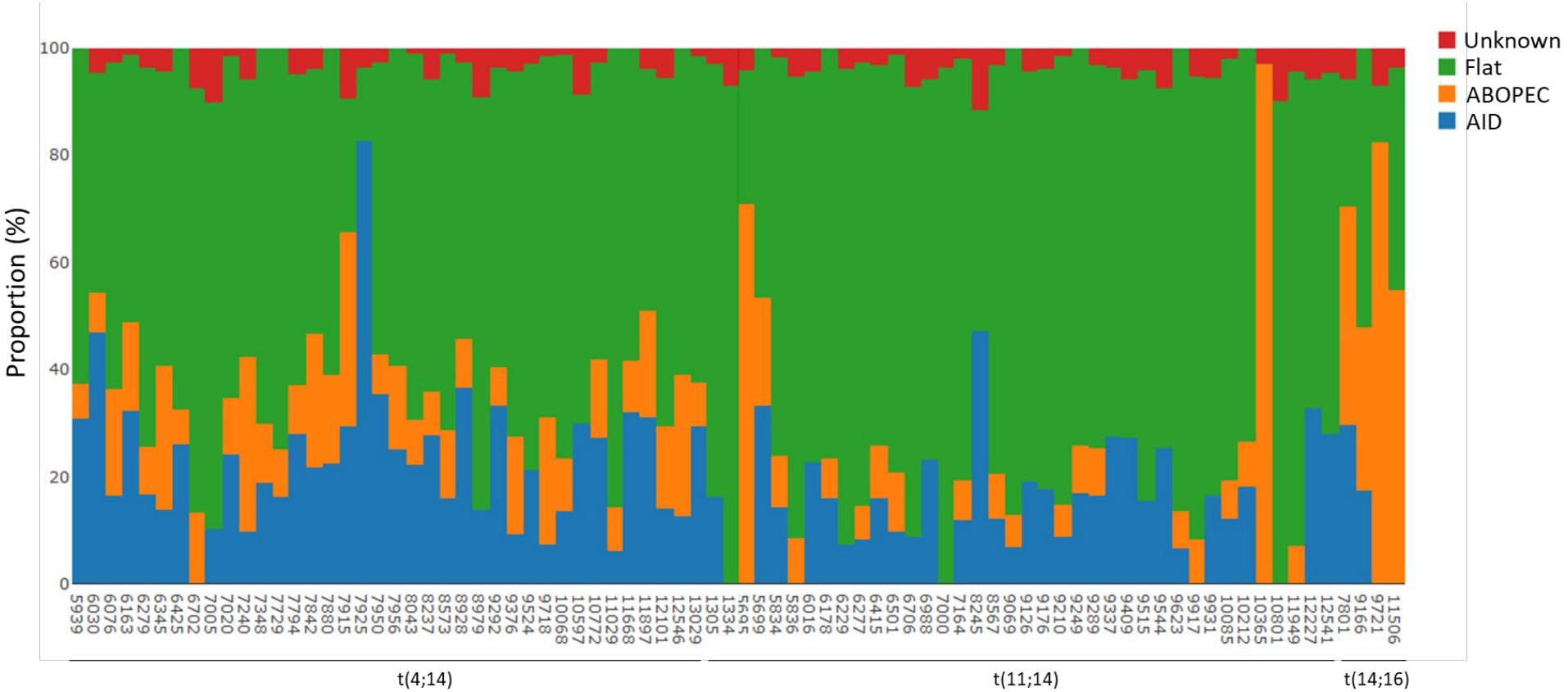
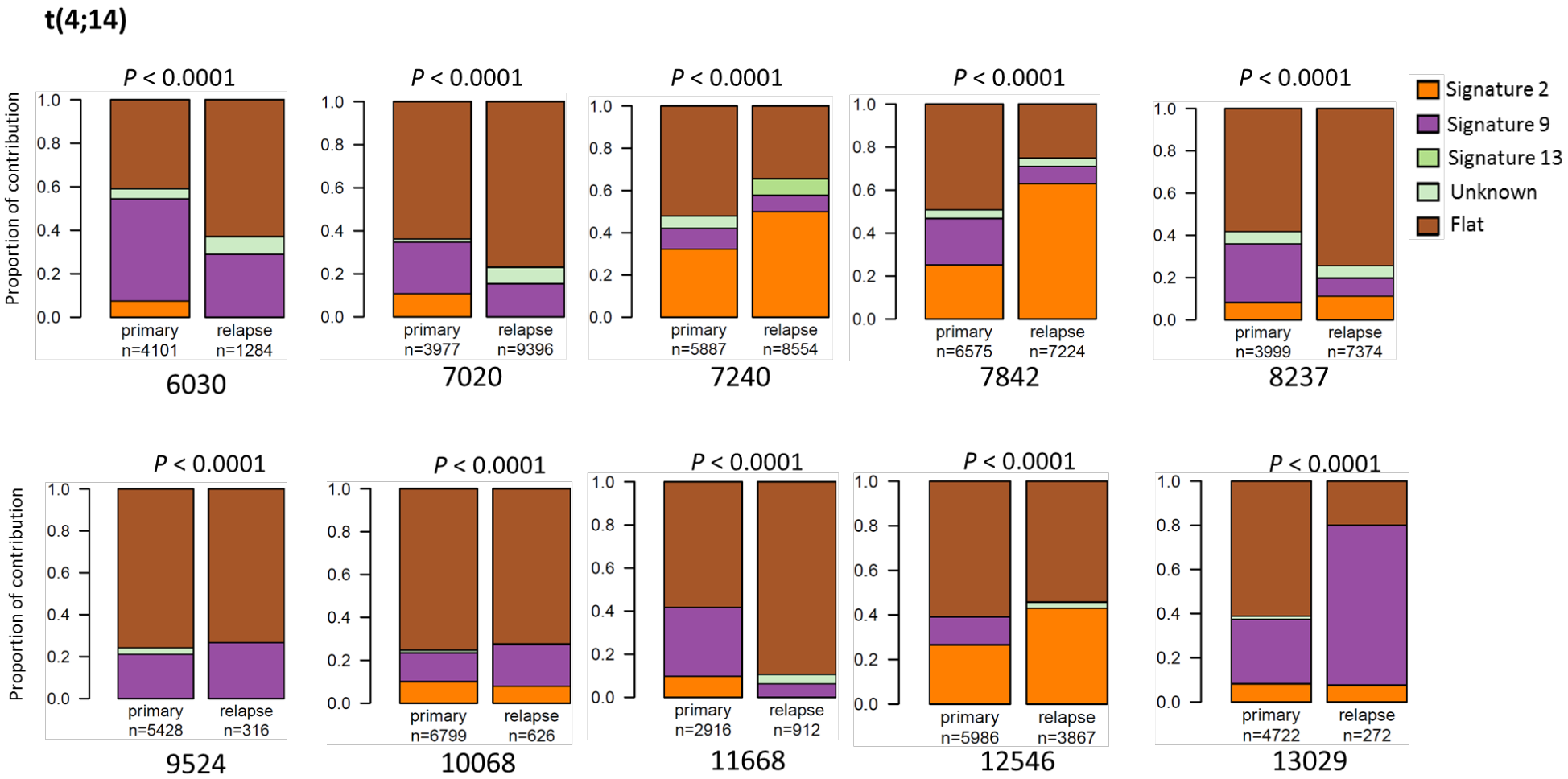
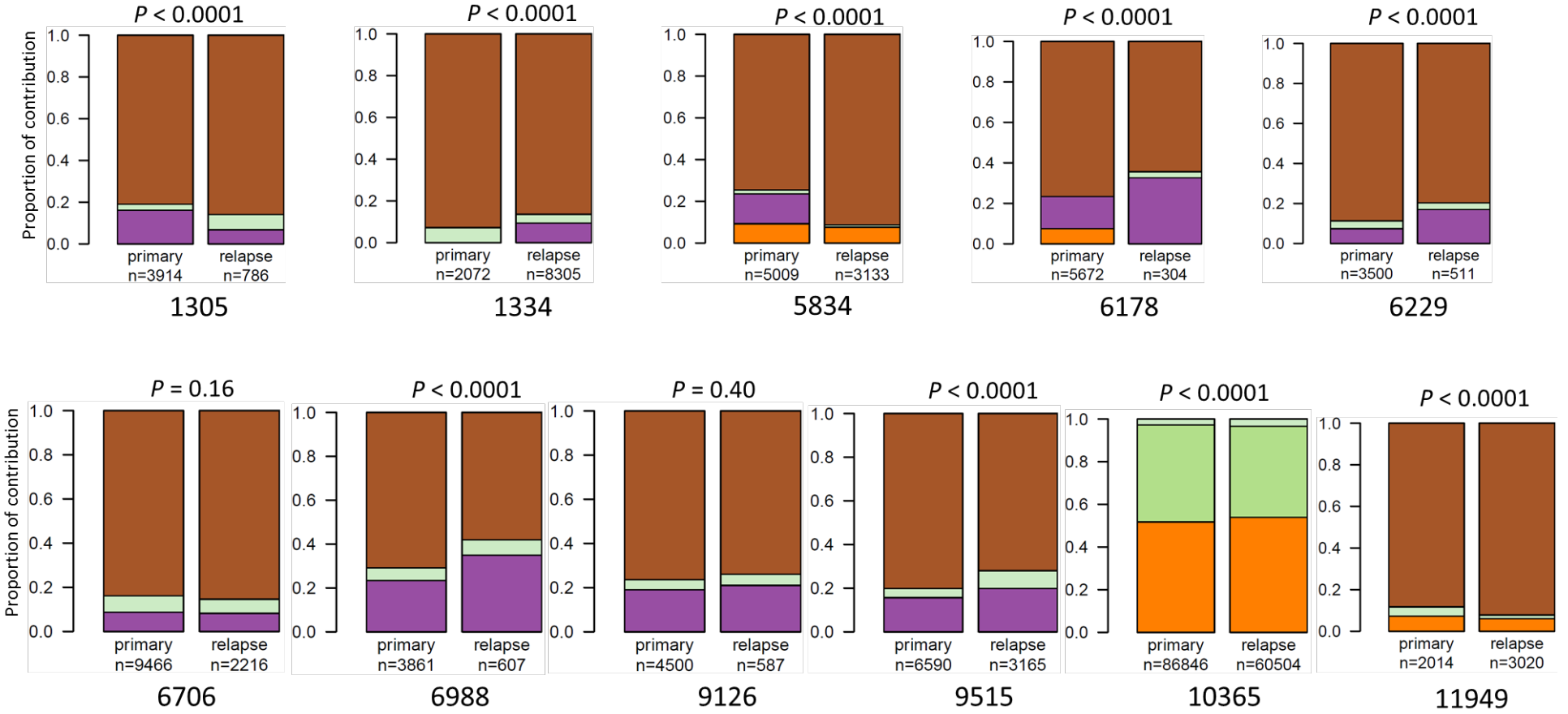


Figure 5.17: Mutation signatures contribution in primary versus relapsed tumours. Stacked bar charts showing comparisons of major mutational signatures between primary versus relapse-specific mutations. The *P*-values refer to the overall difference in distribution between primary and relapse-specific mutations (chi-squared test). n = number of mutations. Flat signatures include COSMIC signatures 3, 5, and 8.



t(11;14)


P < 0.0001
Proportion of contribution
1.0
0.8
0.6
0.4
0.2
0.0
primary n=3914
relapse n=786
1305
P < 0.0001
primary n=2072
relapse n=8305
1334
P < 0.0001
primary n=5009
relapse n=3133
5834
P < 0.0001
primary n=5672
relapse n=304
6178
P < 0.0001
primary n=3500
relapse n=511
6229
P = 0.16
primary n=9466
relapse n=2216
6706
P < 0.0001
primary n=3861
relapse n=607
6988
P = 0.40
primary n=4500
relapse n=587
9126
P < 0.0001
primary n=6590
relapse n=3165
9515
P < 0.0001
primary n=86846
relapse n=60504
10365
P < 0.0001
primary n=2014
relapse n=3020
11949

t(14;16)

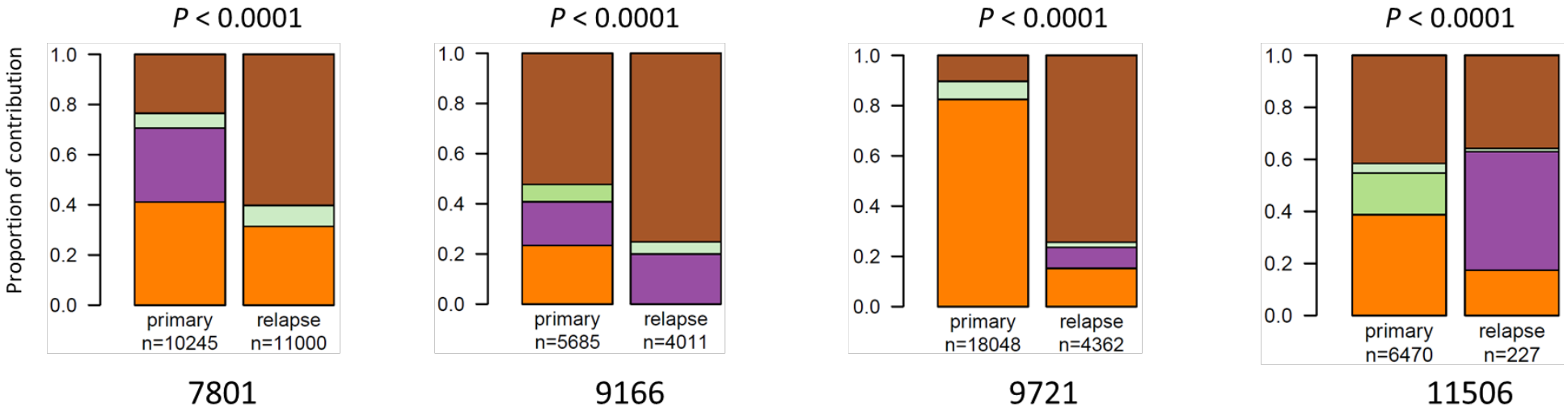

P < 0.0001
P < 0.0001
P < 0.0001
P < 0.0001
Proportion of contribution
1.0
0.8
0.6
0.4
0.2
0.0
primary
n=10245
relapse
n=11000
primary
n=5685
relapse
n=4011
primary
n=18048
relapse
n=4362
primary
n=6470
relapse
n=227
7801
9166
9721
11506

Figure 5.18: Mutation types in primary versus relapse-specific mutations. Boxplots show proportions of different mutation types in primary and relapse-specific mutations. *, $Q < 0.05$

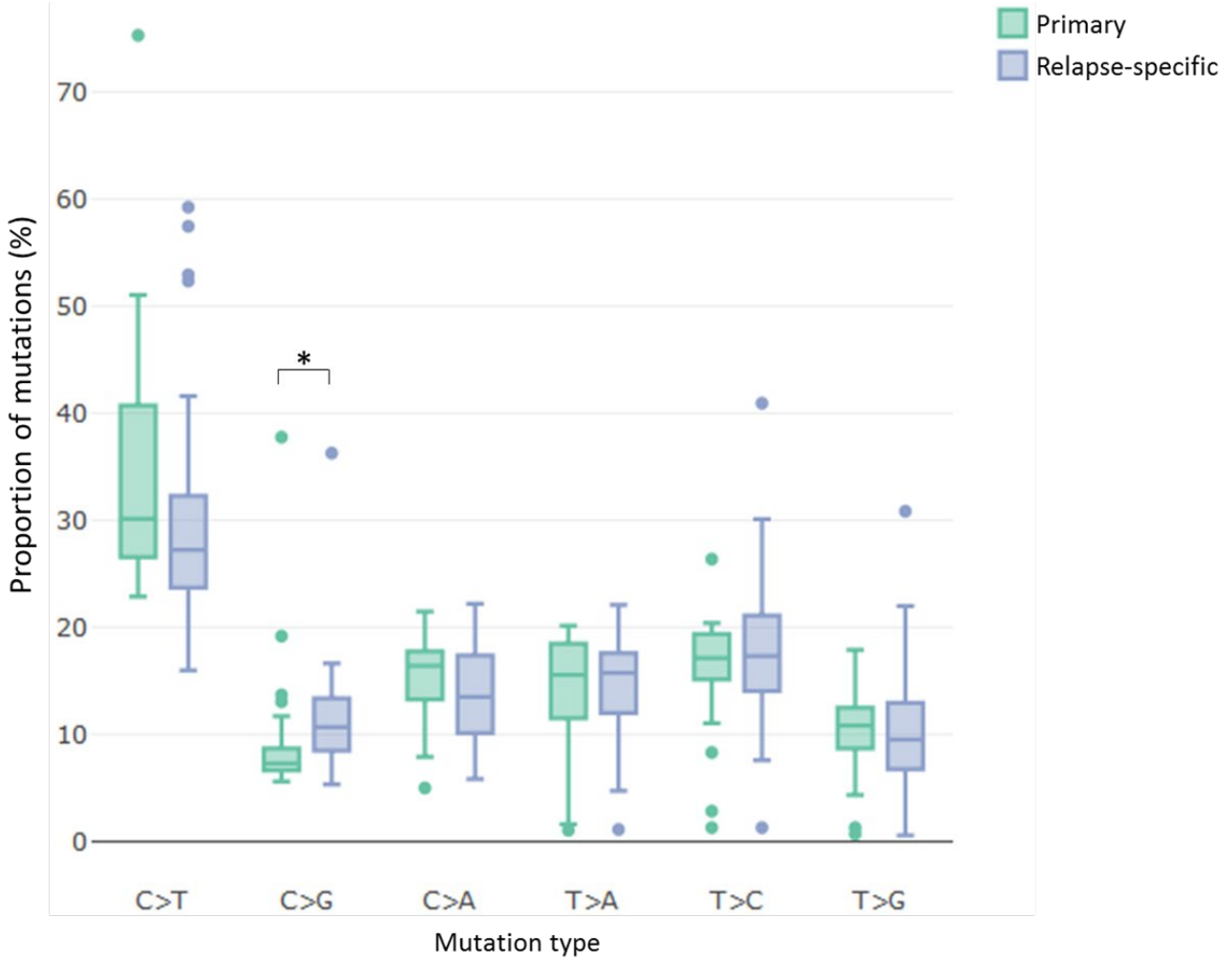


Figure 5.19: *De novo* extraction of WGS single nucleotide variants signatures using non-negative matrix factorization algorithm in 25 relapsed tumours. (a) Summary of four *de novo* mutational signatures extracted. (b) Cosine similarity heatmap. *De novo* extracted mutational signatures are compared against 30 COSMIC mutational signatures. The colour code (0 to 1) represents the resemblance between each pair of signatures.

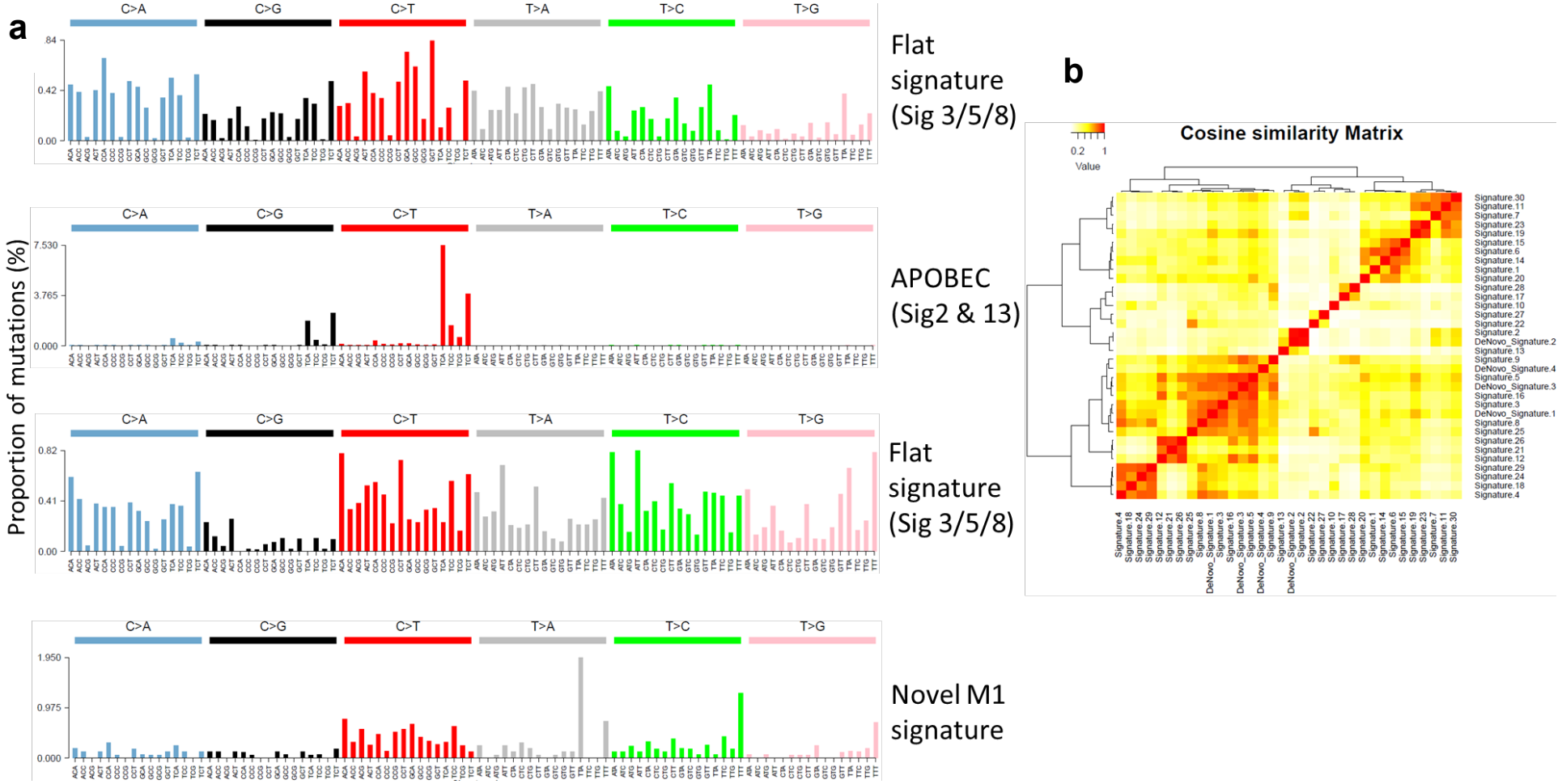


Table 5.10: Fitting of mutational signatures with M1 signature included in 25 relapsed tumours. Signature M1 is mostly confined to relapsed tumour 9524.

Samples	Signature M1	Signature 2	Flat signatures	Signature	Signature 13	Unknown
10068	0.09	0.09	0.69	0.13	0.00	0.00
10365	0.00	0.53	0.00	0.00	0.43	0.04
11506	0.07	0.39	0.35	0.00	0.16	0.02
11668	0.14	0.09	0.55	0.23	0.00	0.00
11949	0.10	0.00	0.83	0.00	0.00	0.08
12546	0.15	0.31	0.43	0.09	0.00	0.02
13029	0.00	0.08	0.53	0.29	0.00	0.09
1305	0.07	0.00	0.68	0.21	0.00	0.03
1334	0.11	0.00	0.78	0.07	0.00	0.04
5834	0.09	0.07	0.74	0.00	0.00	0.10
6030	0.13	0.06	0.34	0.39	0.00	0.08
6178	0.09	0.07	0.70	0.14	0.00	0.00
6229	0.12	0.00	0.79	0.00	0.00	0.09
6706	0.10	0.00	0.78	0.08	0.00	0.05
6988	0.13	0.00	0.57	0.25	0.00	0.05
7020	0.11	0.07	0.66	0.15	0.00	0.02
7240	0.06	0.41	0.36	0.08	0.00	0.08
7801	0.00	0.32	0.37	0.16	0.00	0.15
7842	0.00	0.45	0.25	0.14	0.00	0.16
8237	0.09	0.09	0.63	0.14	0.00	0.05
9126	0.15	0.00	0.66	0.16	0.00	0.03
9166	0.11	0.15	0.54	0.15	0.00	0.05
9515	0.09	0.00	0.70	0.16	0.00	0.05
9524	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
9721	0.00	0.69	0.24	0.00	0.00	0.07

Figure 5.20: Evolutionary trajectories of relapse. (a) Pattern 1 (3/25), dominant clone in primary survives treatment and gains additional mutations at relapse; (b) Pattern 2 (4/25), subclone in primary survives treatment and expands to become dominant clone at relapse; (c) Pattern 3 (18/25), eradication or decrease in frequency of one or more clones in primary and emergence of new clones not previously detected in primary. Left panel, two-dimensional density plots showing clustering of mutations by cancer cell fraction (CCF) in primary and relapse tumours. Darker red areas indicate location of a high posterior probability of a cluster. Clusters are annotated with coding driver mutations and major copy number alterations. Central panels, chromosomal copy-number profiles of primary (upper) and relapse (lower) tumours. Thick and thin lines represent clonal and sub-clonal copy number states respectively. Yellow and dark blue lines denote total and minor copy number alleles. Right panels, Muller plots of evolutionary trajectories. P, primary; R, relapse.

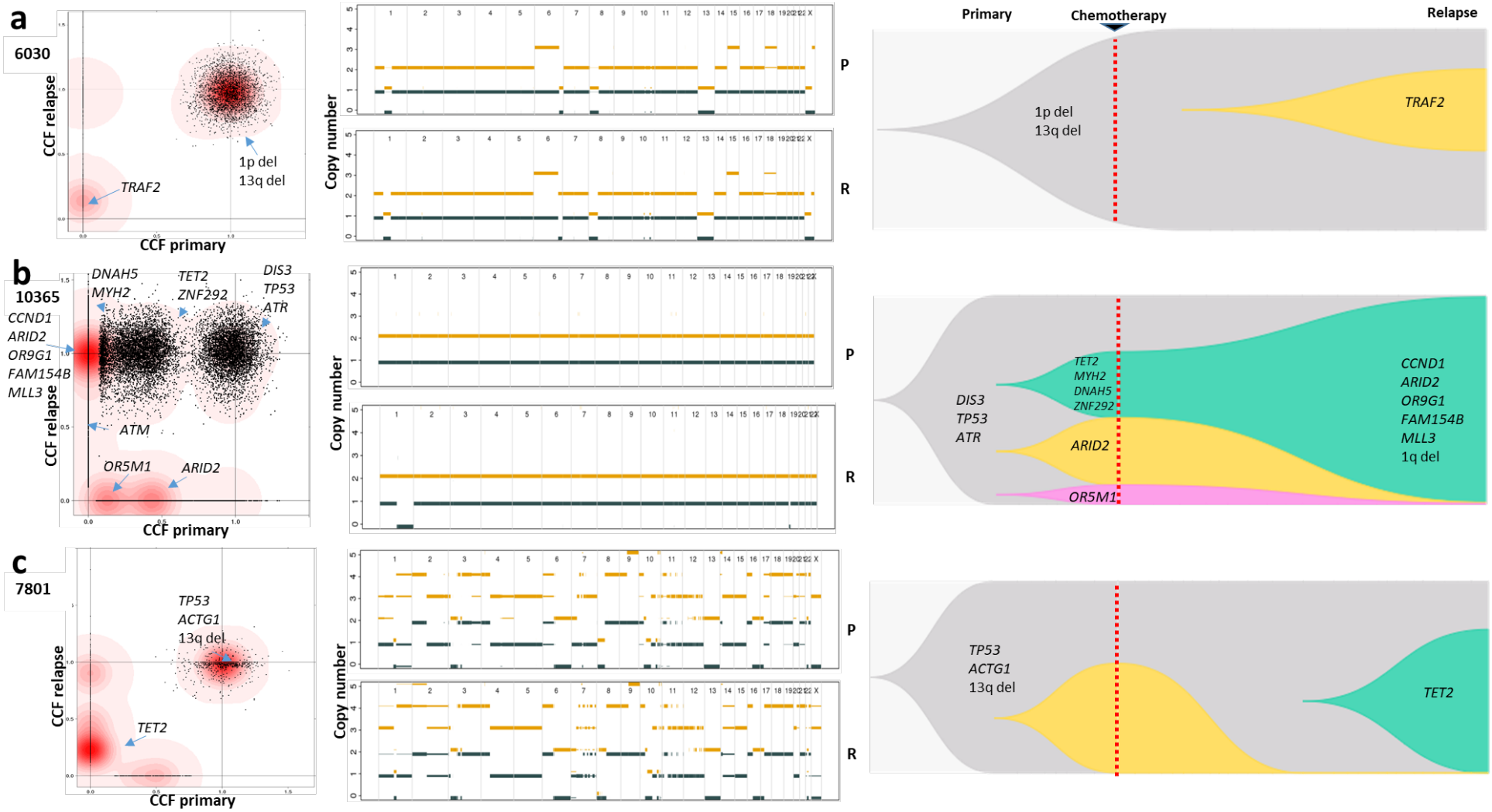
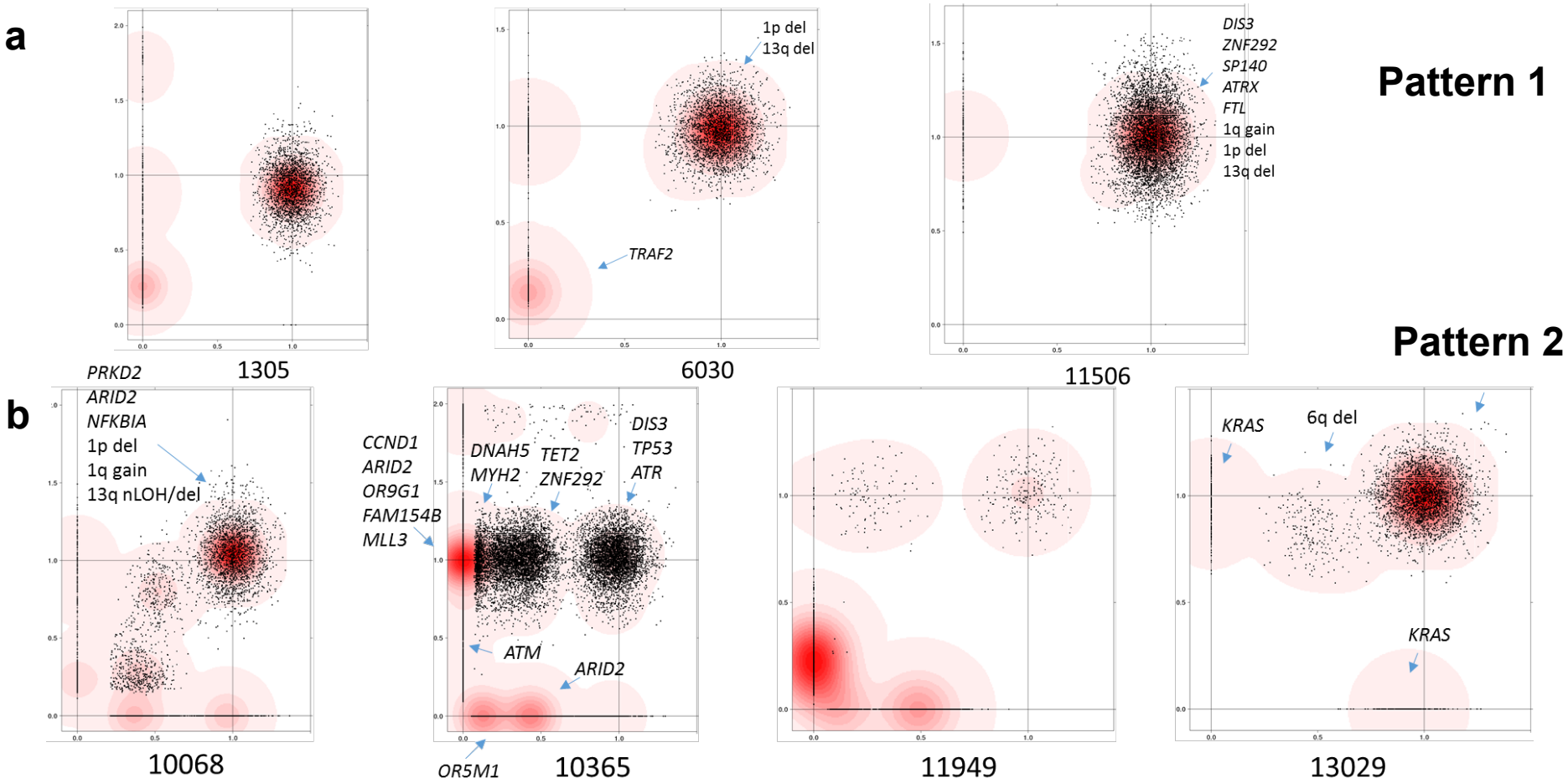
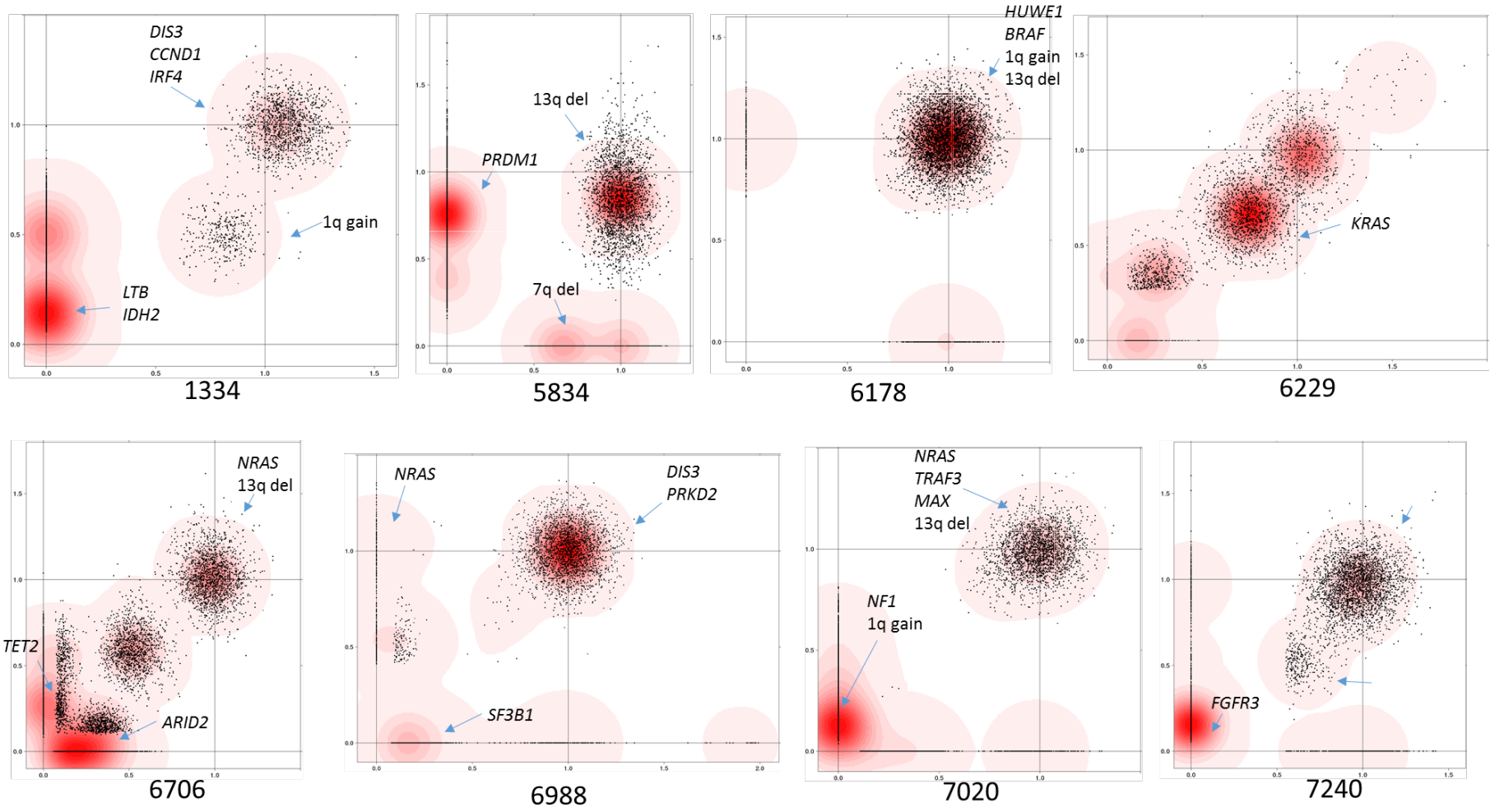

a
6030
CCF relapse
CCF primary
1p del
13q del
TRAF2
Copy number
P
R
Primary
Chemotherapy
Relapse
1p del
13q del
TRAF2
b
10365
DNAH5
MYH2
TET2
ZNF292
DIS3
TP53
ATR
CCND1
ARID2
OR9G1
FAM154B
MLL3
ATM
OR5M1
ARID2
CCF relapse
CCF primary
Copy number
P
R
TET2
MYH2
DNAH5
ZNF292
DIS3
TP53
ATR
ARID2
OR5M1
CCND1
ARID2
OR9G1
FAM154B
MLL3
1q del
c
7801
TP53
ACTG1
13q del
TET2
CCF relapse
CCF primary
Copy number
P
R
TP53
ACTG1
13q del
TET2

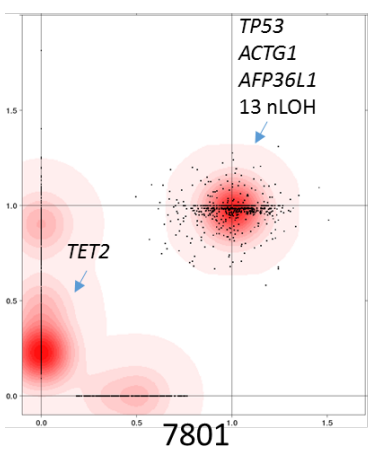
Figure 5.21: Evolutionary trajectories of relapse in 25 relapsed tumours. Two-dimensional density plots showing the clustering of mutations (black dots) by cancer cell fraction (CCF) in primary (x-axis) and relapsed tumours (y-axis). Darker red areas denote high posterior probability of a cluster (*i.e.* a clone). Clusters are annotated with coding driver mutations and major copy number alteration events. (a) Pattern 1: Dominant clone in primary gains additional mutations at relapse. (b) Pattern 2: A subclone survives and expands to become the dominant clone at relapse. (c) Pattern 3: Eradication or decline of one or more of primary clones and emergence of new clones not previously detected in primary. CCF, cancer cell fraction.



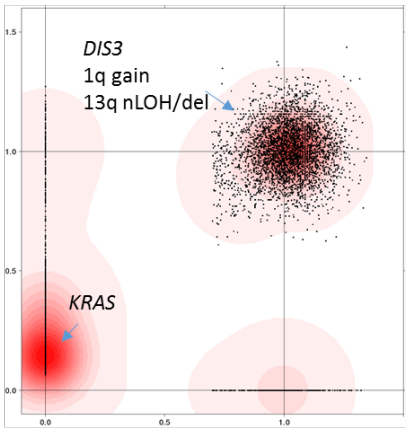
c **Pattern 3**


DIS3
CCND1
IRF4
1q gain
LTB
IDH2
1334
13q del
PRDM1
7q del
5834
HUWE1
BRAF
1q gain
13q del
6178
KRAS
6229
NRAS
13q del
TET2
ARID2
6706
NRAS
DIS3
PRKD2
SF3B1
6988
NRAS
TRAF3
MAX
13q del
NF1
1q gain
7020
FGFR3
7240

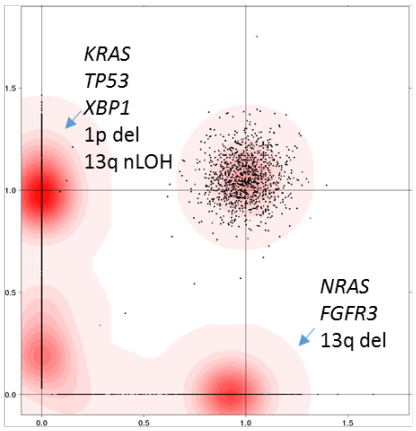
Pattern 3



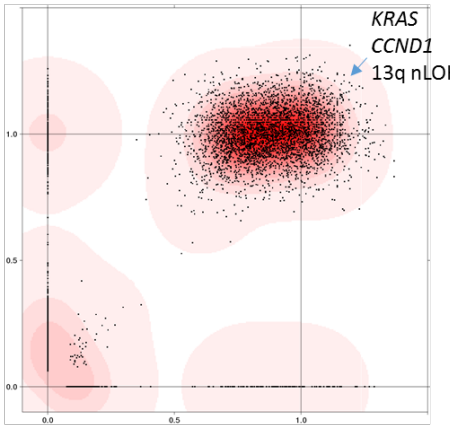
7801



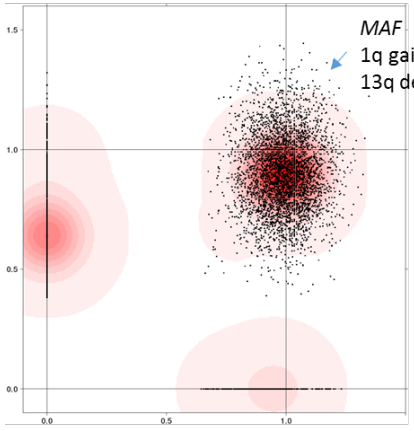
7842



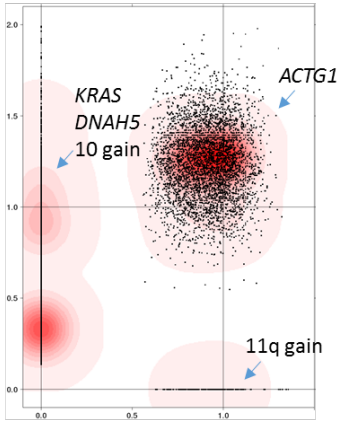
8237



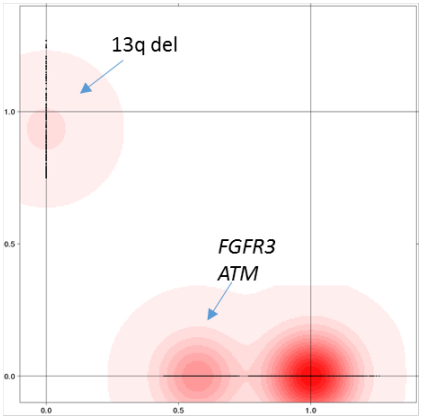
9126



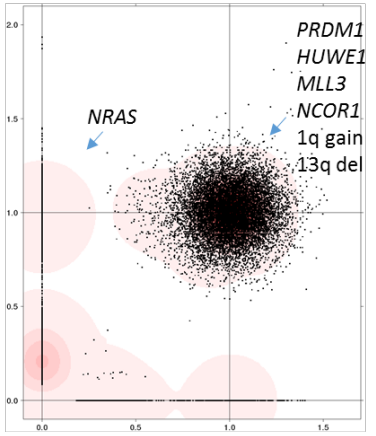
9166



9515

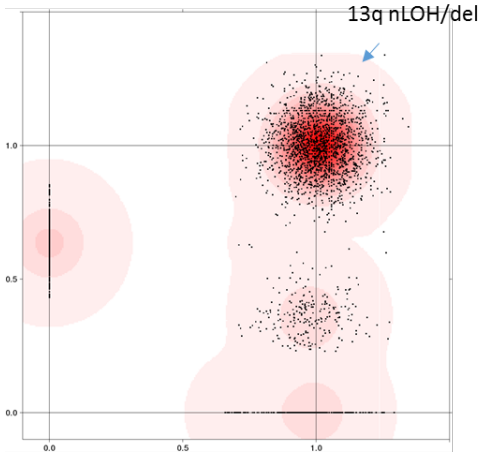


9524

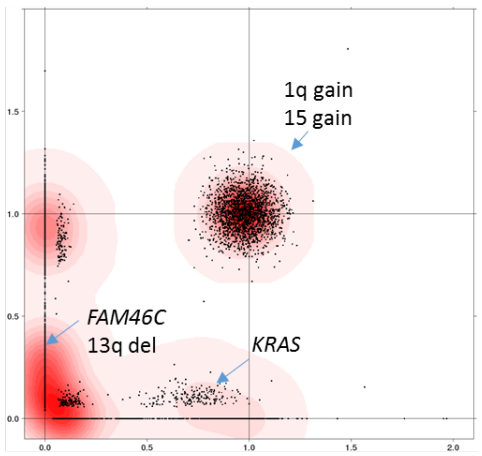


9721

Pattern 3



11668



12546

5.4 Discussion

Using high-depth WGS, this study provides for an enhanced genetic model of the development and progression of MM. This study expands upon previous findings which have been based on WES/targeted sequencing[5, 6, 8, 102, 103], low coverage sequencing[104], or fluorescence *in situ* hybridization and/or array technology[102, 105]. While the analysis was restricted to MM with initiating translocation, it provides clear evidence for a common origin of tumour subpopulations with many tumours being composed of at least one subclone, reflecting the clonal heterogeneity present in both primary and relapse MM.

In addition to known coding drivers, the study extends the number of potential non-coding drivers in MM, including those associated with *CXCR4*, *BIRC2*, *BIRC3*, and *IGLL5*. Non-coding regulatory regions were additionally disrupted at relapse, included those influencing expressions of *XBP1*, *RBX1,* and *SCML1*. Common pathways affected by coding and non-coding mutations arising in MM relapse included those associated with WNT-, MAPK- and NOTCH-signalling, base excision repair, cell cycle, telomere maintenance, and cellular senescence (Table 5.12). Notably, relapse was characterised by frequent CNVs, the most common being 13q and 17p deletion. Since the additional CNVs often occurred at unstable genomic regions, it suggests increased chromosome instability and chromothripsis are important means to escape therapy, analogous to that seen with chronic myeloid leukaemia in response to imatinib[254]. While 21q gain, 22q nLOH, 1q nLOH and mutation of *CCND1*, *MAX*, *PRKD2*, *DIS3*, and *NRAS* are early events; my findings suggest that 13q deletion is preceded by nLOH.

Overall, the mutational load was higher in relapse MM and aberrations previously linked to MM resurfaced in both primary pre-treatment and relapsed tumours in the cohort, including mutations in *RAS* genes, *DIS3*, *TP53*, *FGFR3*, and *PAX5* CRE mutations. As well as highlighting mutation of genes with established roles in MM, a number of frequently acquired *de novo* coding mutations was identified (*e.g*. *FAM46C*, *TRAF2*, *NF1*, *XBP1*, *SYNE1*, *MTCL1*, *ABCA13*, *ADAMTS9*, *ZNF521*), *de novo* translocation (*MAP3K14*) and pre-existing mutations (*e.g*. *TET2*, *ZNF292*, *MYH2*, *DNAH5* and 22q deletion) as potentially important in enhancing survival and chemo-resistance at relapse. *SYNE1* missense mutations have previously been reported in drug-resistant MM[255]. *MTCL1*

regulates microtubule organisation, whose disruption could lead to defect in cell division[256]. Mutations in cereblon (*CRBN*) and those associated with Cullin-RING E3 ubiquitin ligase complex have been reported as a feature of relapse MM with immunomodulatory (IMiD) therapy[6]. While all of the patients studied were treated with thalidomide or lenalidomide the emergence of mutations in these genes was not observed, consistent with a recent exome-based analysis[103]. The data are therefore consistent with the assertion that IMiD resistance is mediated through alternative mechanisms.

With high-depth WGS, I have been able to refine complex genomic evolution patterns at relapse in MM compared to previous study, which had relied on WES[5]. For instance, the 'branching evolution' model described previously often co-occurs with the 'differential clonal response' model as identified by the dataset. In addition, I did not find association between t(11;14) subtypes with 'no change/linear' models[5] and I suspect model reconstruction might have been confounded due to limited number of mutations of previous studies using WES[5]. In addition, with more refined evolutionary pattern classification, I observed an unprecedented association between patients with subclonal expansion patterns and significantly shorter time to relapse.

Higher proportion of C•G>G•C at relapse is associated with DNA damage by oxidative stresses[257], possibly due to oncogene activation and/or enhanced metabolism in relapsed MM[258]. The increased mutational burden in relapse was associated with increased APOBEC/AID and DNA repair deficiency. Chemotherapeutic agents potentially contribute to emergence of additional subclonal mutations at relapse, bearing DNA repair deficiency characteristics, through induction of DNA inter-strand cross-links causing stalling or incomplete resumption of DNA repair during regeneration of surviving tumour cells[259].

Inevitably, due to technical limitations, the ability to detect mutations in rare cells (mostly related to currently achievable levels of coverage with WGS) and spatial sampling constraints, the models potentially underestimate clonal heterogeneity in MM. However the loss of primary tumour clones was observed at relapse in 22 of 25 cases, suggesting that some subclones are eradicated by therapy (Figure 5.21). Nevertheless, treatment failed to eradicate the founding clones in many cases. The data also imply the acquisition of new mutations in the founding clone

or one of its subclones, which subsequently undergo selection and clonal expansion contributing to disease progression. It is likely that some mutations gained at relapse may alter the growth properties of MM cells, or confer resistance to additional chemotherapy.

Presently strategies to improve the poor cure rates of relapsed MM are limited. Here the study has demonstrated that relapsed MM harbour significantly more mutations than primary tumours and clonal selection of mutations occurs at relapse, which are accompanied by subclonal heterogeneity. MM cells routinely acquire a small number of additional mutations at relapse, and some of these mutations may contribute to clonal selection and chemotherapy resistance. Theoretically, these data provide a rationale for identifying disease-causing mutations for MM, which may be amenable to targeted therapies to avoid the use of cytotoxic drugs, many of which are mutagens. However, it remains to be determined whether the current arsenal of therapies directed against downstream effectors of mutated genes will be effective given that the MM genome in an individual patient is likely to be continuously evolving. Hence, it can be asserted that eradication of the founding clone and all of its subclones will be required to achieve complete cure.

Table 5.11: Summary of relapse-specific coding driver mutations, promoter mutations, CRE mutations, driver translocations, and copy number alterations identified in 25 primary tumour-relapse pairs grouped by subtype. CRE: *cis*-regulatory element.

Subtype	Coding drivers	Promoters	CREs	Driver translocations	Frequent copy number alterations
t(4;14)	*KRAS; TP53; FGFR3; FAM46C; TRAF2; NF1; XBP1*	*MTFRL1; FLT3LG; IL12A; POLG; XBP1; B3GALNT1; ALG10B*	*ABCA10; ABCA5*	*MAP3K14* t(17,14)(q21,q32)	13q deletion 17p deletion Further copy number changes at unstable genomic regions
t(11;14)	*PRDM1; LTB; IDH2; KRAS; NRAS; CCND1; ATM; DNAH5; OR9G1; FAM154B; MLL3*	*RBX1; FAM81A; POLG; KCTD13; SCML1*	*SCAF8*		
t(14;16)	*NRAS; TET2*	*MYO1E; ALG10B; TMSB4X; KCTD13; SCML1*			

CHAPTER 6 Impact of mitochondrial DNA mutations in multiple myeloma

6.1 Overview and rationale

Mitochondria have long been considered important for tumour transformation and treatment response[112]. The majority of cancers have altered metabolism[260] and increased uptake of glucose (*i.e.* the 'Warburg effect') attributed to defective mitochondria[113]. In addition, mitochondria are associated with multiple key processes linked to tumourigenesis including apoptosis, cell cycle, cell growth, and signalling[117].

Recent evidence indicates mitochondria dysfunction is important in defining chemotherapy resistance and disease progression in MM[118, 119]. In addition, pre-clinical studies have suggested agents targeting mitochondria in relapsed MM can improve patient outcome[120, 121]. Despite this, the spectrum of mtDNA mutations and their functional implications in MM have not been well characterised, partly due to limited sample size and WES depth[124]. Furthermore, any characteristics specific to MM mitochondria have largely been dismissed due to overwhelmingly dominant number of other cancer types included in previous pan-cancer studies[124]. By analysing WGS data from the Myeloma XI trial[128], the somatic mutation landscape, mutation selection at relapse, nuclear genome integration and copy number of MM mitochondria were characterised in this chapter.

6.2 Study design

6.2.1 Samples and dataset

Myeloma XI trial samples and dataset were obtained as detailed in section 2.1.2.

6.2.2 Statistical and bioinformatics analysis

Raw WGS sequencing data were quality checked using FastQC v.0.11.4 and aligned using the Burrows-Wheeler Alignment tool[261] BWA v0.7.13 to the human genome hg38 assembly and human mtDNA rCRS[262] using default parameters. Somatic and germline variants calling were performed as described in section 2.2.12.1. Mitochondrial copy number and heteroplasmy estimation were carried out as detailed in section 2.2.12.2. Identification of somatic mitochondrial transfer was performed as described in section 2.2.12.3.

6.2.2.1 Strand bias and mutational signatures analysis

Analysis of replication and transcriptional strand bias was performed as previously described[124]. Substitution rates for each of the 96 trinucleotide context on L and H strands were calculated and normalised for trinucleotide context[127] (section 2.2.10.3). To examine replication and transcriptional strand biases, I considered 12 substitution classes: 6 possible base substitution × 2 strands (H/L strand or transcribed/non-transcribed strand)[185]. I included all substitutions for replication bias analysis, while transcriptional strand bias was considered for substitutions residing in mtDNA genes (13 protein-coding, 22 tRNA, and 2 rRNA genes). The proportion test was used to determine significant difference in strand biases.

Signature fitting of all primary and relapse somatic mutations against 30 COSMIC signatures were carried out using deconstructSigs[178] with default settings (section 2.2.10.1).

6.2.2.2 dN/dS analysis

dN/dS values for somatic variants were calculated globally and across 13 mitochondrial coding genes using dNdScv R package with default parameters[76]. To minimise the effect of extreme replication bias[124], *MT-ND6* on H strand was excluded when estimating global dN/dS values. The Benjamini-Hochberg FDR procedure was used to adjust for multiple hypothesis testing with coding genes with significance thresholded at $Q < 0.05$.

6.3 Results

To investigate mtDNA somatic mutations in MM, WGS of the 80 matched tumour and normal blood of newly diagnosed patients, of which 25 also had matched relapsed tumours, were utilised[128]. Due to high cellular copy number of mtDNA genomes, far greater mtDNA genome coverage was obtained (normals: median 2149×, range 1015-7777×; primary tumours: median 7836×, range 2376-7938×; relapsed tumours: median 7826×, range 4678-7929×) compared to the nuclear genome (Table 6.1, Appendix 3).

Table 6.1: Mitochondrial coverage, purity, karyotype, and clinical information for all samples from Myeloma XI study

Sample ID	Normal	Primary	Relapse	Sample ID	Normal	Primary	Relapse
	mtDNA mean coverage				mtDNA mean coverage		
1305	3146.49	7911.41	7715.81	7005	2105.563	7753.149	NA
1334	4954.99	7667.44	7769.75	7164	1819.997	7868.564	NA
5834	2599.06	7901.01	7776.38	7348	3226.039	7764.438	NA
6030	2351.02	7932.07	7928.95	7729	2061.198	7887.94	NA
6178	2609.91	4736.47	7788.26	7794	2768.001	7877.458	NA
6229	2162.77	7907.74	7876.58	7880	1848.325	5041.794	NA
6706	2043.31	7920.90	7891.89	7915	7615.676	7758.537	NA
6988	1488.16	7810.58	7914.46	7925	7327.606	5046.893	NA
7020	1903.45	6763.04	7873.73	7950	2123.583	7412.19	NA
7240	3217.18	7901.02	4718.03	7956	5583.107	7762.829	NA
7801	2304.20	6295.85	4677.96	8043	3636.235	7795.991	NA
7842	3552.01	7773.31	7631.30	8245	2293.997	7873.637	NA
8237	3512.01	7636.62	7902.83	8567	2461.213	4546.482	NA
9126	2064.85	7921.43	7880.12	8573	1690.852	7494.349	NA
9166	2843.02	7922.27	7866.82	8928	1014.808	7897.757	NA
9515	2788.05	7937.64	7916.00	8979	1382.481	7907.568	NA
9524	1388.47	7919.95	7789.75	9069	1164.399	7822.628	NA
9721	5029.26	4354.25	7886.04	9176	1926.553	7830.108	NA
10068	2466.19	7909.97	7825.64	9210	2284.554	7899.118	NA
10365	7776.70	7373.59	7536.41	9249	1527.055	7864.805	NA
11506	2952.32	7901.00	7825.34	9289	1641.367	7905.844	NA
11668	3404.18	7898.25	7915.56	9292	1883.016	7822.762	NA
11949	2542.27	7879.31	7900.26	9337	1627.241	7919.883	NA
12546	3041.40	7235.25	7759.24	9376	1630.022	7841.582	NA
13029	2502.02	7850.08	7824.53	9409	2074.175	7919.635	NA
5695	1660.269	7699.335	NA	9544	1489.989	7820	NA
5699	1952.174	7816.793	NA	9623	2919.537	7630.801	NA
5836	1919.918	7898.204	NA	9718	1264.072	7650.709	NA
5939	2207.42	7907.82	NA	9917	1683.971	7745.494	NA
6016	1311.494	7850.447	NA	9931	1027.767	7732.399	NA
6076	2549.937	7883.836	NA	10085	2074.804	7885.826	NA
6163	1922.513	4035.634	NA	10212	1889.014	7784.58	NA
6277	1758.064	7905.633	NA	10597	1405.381	7756.545	NA
6279	2319.718	7867.252	NA	10772	7663.268	7843.713	NA
6345	1571.32	2375.531	NA	10801	3120.295	7824.747	NA
6415	2134.605	7914.818	NA	11029	2132.637	7867.625	NA
6425	2383.117	4245.959	NA	11897	1758.225	7742.244	NA
6501	3277.534	7866.631	NA	12101	1433.479	5210.757	NA
6702	1391.19	3516.607	NA	12227	2798.262	7922.649	NA
7000	6500.789	7898.802	NA	12541	2352.256	7577.769	NA

6.3.1 Somatic mitochondrial mutation landscape in multiple myeloma

I identified 210 mtDNA SNVs in 80 primary tumours (median 3 SNVs/tumour). These showed strong replicative strand bias, predominantly C>T on heavy strand and T>C on light strand (Figure 6.1), which was previously ascribed to replication-coupled process partly due to the lack of transcriptional strand bias observed[124]. Examining the sequence context of mutations revealed the contribution of defective transcription-coupled DNA repair COSMIC signatures 12 (16%), 21 (15%), 23 (11%), and 26 (48%) (Figure 6.2a). In concordance, transcriptional strand bias was observed across all genes (Figure 6.2b), with the strongest signal for C>T where transcribed strand is more frequently repaired[263]. The weaker transcriptional strand bias for T>C is likely due to the neutralising effects from COSMIC signatures with opposing transcriptional strand biases (Figure 6.3). Collectively, these findings are consistent with the contribution of transcription-coupled DNA repair defects in MM mtDNA.

I identified 14/210 (6%) somatic mutations as pathogenic (Table 6.2). A number of these variants occur in more than one patient and associated with established diseases[188] including m.4136A>G (Leper's optic atrophy), m.9185T>C (Charcot-Marie-Tooth disease, Leigh syndrome, complex V deficiency), m.15246G>A (development delay, hearing impairment, macrocephalus), and m.15287T>C (familial breast cancer). Since mitochondrial disease is rare in general population (around 1 in 5000)[264], it is likely these variants have a direct effect on gene function.

Figure 6.1: Mutational patterns by 96 trinucleotide context across 80 primary tumours from Myeloma XI trial. Substitution rate is normalised for trinucleotide context difference between mitochondrial light and heavy chains.

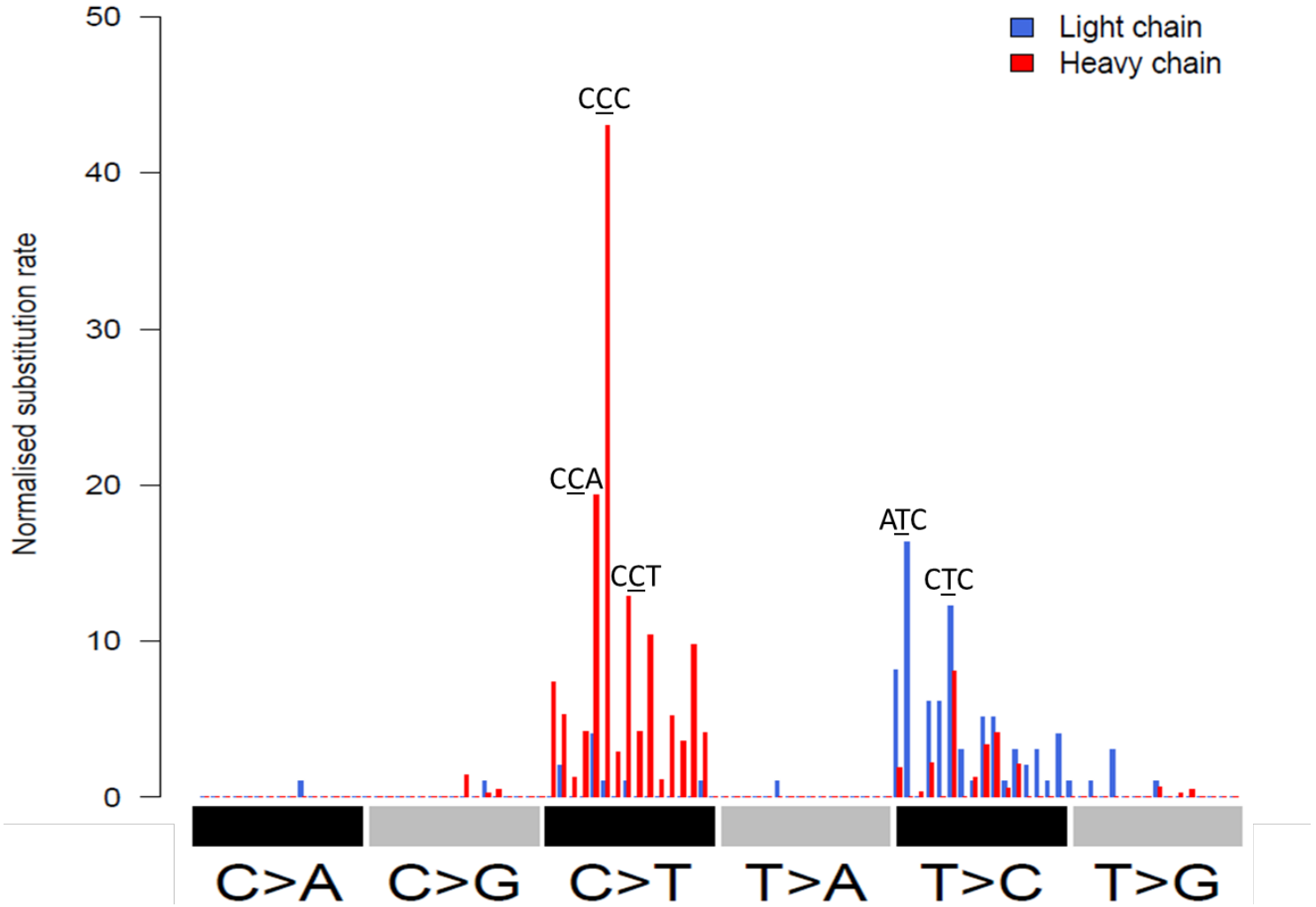


Figure 6.2: Mutational signatures in mitochondrial DNA of 80 primary tumours from Myeloma XI trial. (a): Contribution of COSMIC mutational signatures extracted by deconstructSigs[178]. (b) Transcriptional strand biases across all mitochondrial genes. Significant difference in strand bias was assessed by proportion tests. **, $P < 0.01$; ***, $P < 0.001$.

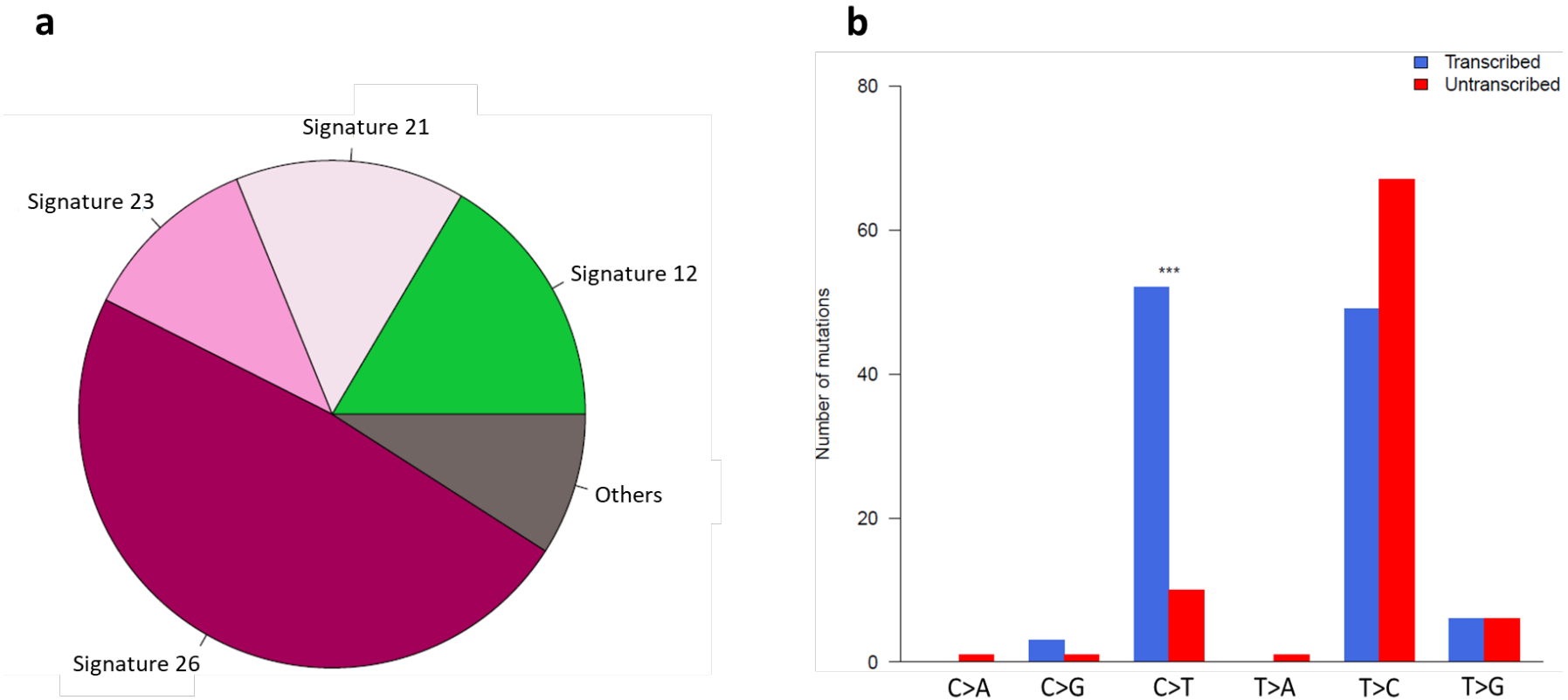


Figure 6.3: Transcriptional strand bias contributed by various COSMIC mutational signatures extracted in 80 Myeloma XI primary tumours. Left panel: Number of substitutions observed on transcribed and untranscribed strand. Significant strand bias difference was assessed by proportion tests. ***, *P* < 0.001. Right panel: Screenshots of from COSMIC website (https://cancer.sanger.ac.uk/cosmic/signatures/SBS/) indicating transcriptional strand bias of COSMIC mutational signatures extracted from this study. COSMIC single base substitution (SBS) signatures 21 and 26 have opposing transcriptional strand bias with signature 12 for T>C.

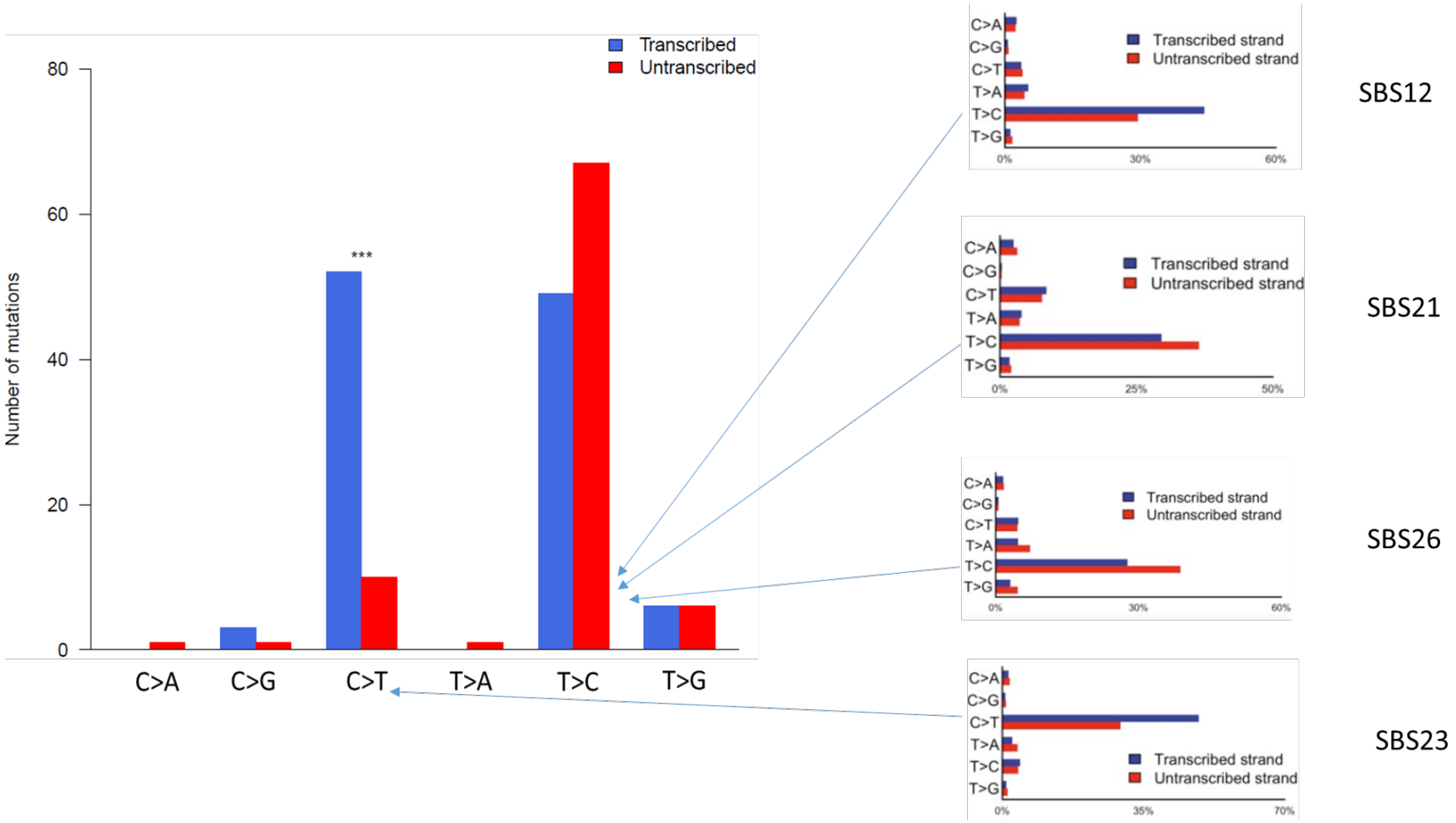


Table 6.2: Mitochondrial somatic variants in 80 patients from Myeloma XI trial associated with pathogenicity.

Mito variants	Clinical significance	Known disease associated
m.14319T>C	risk factor	Parkinson disease 6, autosomal recessive early-onset
m.14846G>A	Pathogenic	Exercise intolerance
m.15246G>A	Likely pathogenic	Developmental delay; Hearing impairment; Macrocephalus
m.15287T>C	Likely pathogenic	Familial cancer of breast
m.3946G>A	Pathogenic	Juvenile myopathy, encephalopathy, lactic acidosis and stroke
m.4136A>G	Pathogenic	Leber's optic atrophy
m.5591G>A	Pathogenic	Mitochondrial myopathy
m.5628T>C	Likely pathogenic	Ophthalmoplegia, deafness, gout
m.5637T>C	Likely pathogenic	
m.5703G>A	Pathogenic	Ophthalmoplegia
m.5703G>A	Pathogenic	Ophthalmoplegia, mitochondrial myopathy
m.5920G>A	Pathogenic	Recurrent myoglobinuria
m.9185T>C	Pathogenic	Charcot-Marie-Tooth disease, Leigh syndrome, Mitochondrial complex v (ATP synthase) deficiency
m.9379G>A	Pathogenic	Hepatic failure, early-onset, and neurologic disorder due to cytochrome C oxidase deficiency

6.3.2 Positive selection of mtDNA mutations is a feature of relapse

Significant difference in mtDNA somatic mutational burdens was not observed between MM subtypes, or between primary and relapsed tumours (Figure 6.4). Most germline variants are homoplasmic while somatic variants are more variable in their heteroplasmic level ($P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) (Figure 6.5). The majority of germline mutations are located outside protein-coding regions or synonymous mutations, with no loss-of-function (*i.e.* truncating) variants detected (Figure 6.6a). In contrast, somatic mutations are more enriched for missense and truncating variants ($P < 2.2 \times 10^{-16}$) (Figure 6.6a), suggesting germline and somatic variants are under different selection constraints. The most frequently disrupted mtDNA coding genes by non-synonymous somatic mutations include *MT-ND5* (29% of primary tumours), *MT-ND4* (24%), *MT-CO1* (20%), and *MT-ND1* (15%) (Table 6.3).

The dN/dS ratio shows no evidence of positive or negative selection for somatic mutations in primary tumours (dN/dS = 1.24, 95% CI: 0.76 – 2.03; $P = 0.39$) (Figure 6.6b), consistent with the observation that missense and truncating mutations do not have significantly different heteroplasmic levels compared to silent mutations (Figure 6.6c). However, non-synonymous mutations are positively selected at relapse (dN/dS = 3.01, 95% CI: 1.09 – 8.25; $P = 0.033$) (Figure 6.6b), in concordance with significant increase in homoplasmy of non-synonymous mutations at relapse (Figure 6.7). Notably, missense mutations in mitochondrial genes composing of the NADH dehydrogenase complex (*MT-ND2*, *MT-ND4*, and *MT-ND5*) have higher than expected rate of missense mutations (*i.e.* positively selected) at relapse ($Q < 0.05$) (Figure 6.6d) with non-synonymous mutations in *MT-ND5* and *MT-CO3* being most frequently acquired at relapse (Table 6.4), implying potential survival advantage rendered through disruption of these genes.

Figure 6.4: Mitochondrial mutational burdens (a) across multiple myeloma subtypes and (b) between primary and relapsed tumours. Significant difference was assessed using Wilcoxon rank-sum test. ns, not significant.

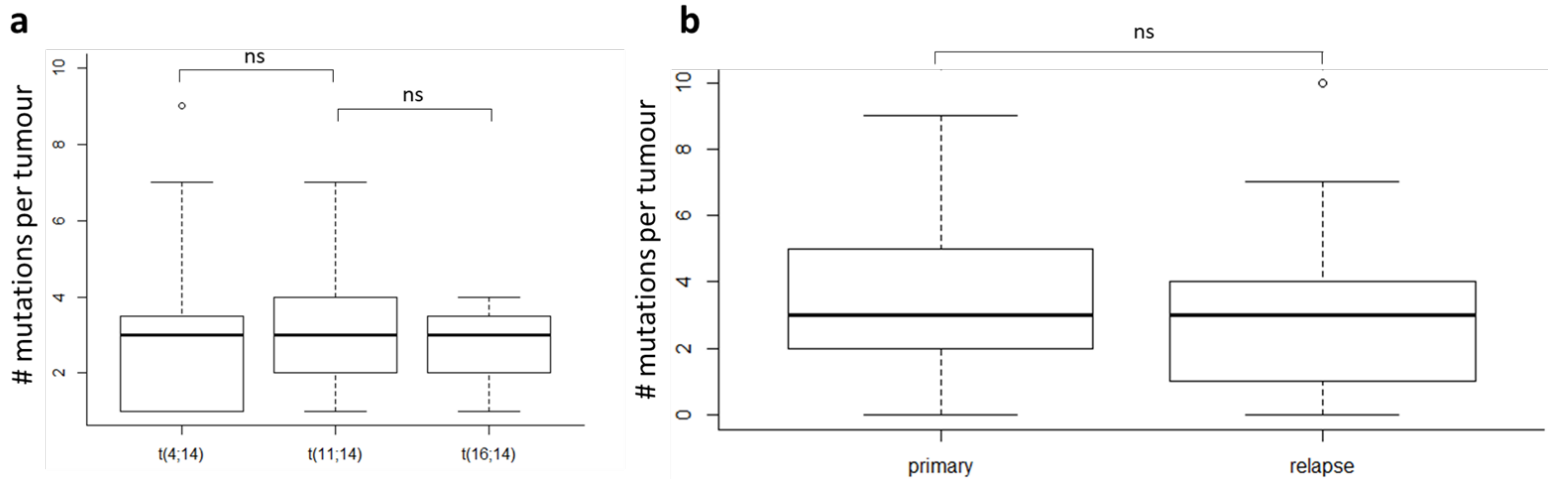


Figure 6.5: Heteroplasmic level comparison between mitochondrial germline (n = 2137) and somatic mutations (n = 223). Significant difference was assessed using Wilcoxon rank-sum test. ***, *P* < 0.01. VAF, variant allele frequency.

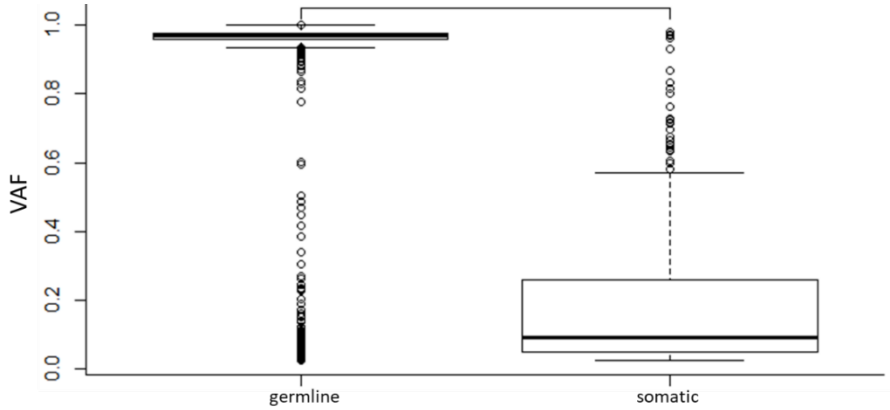


Figure 6.6: Selection of mtDNA somatic mutations in primary and relapse multiple myeloma tumours. (a): Proportion of mutation type in mitochondrial germline and somatic mutations. Difference on mutation type contribution was assessed by chi-squared test. (b) Global dN/dS ratio for all 80 primary tumours, 25 matched primary tumours, and 25 relapsed tumours. *, $P < 0.05$. Vertical lines depict 95% CI. (c) Heteroplasmic level comparison between silent (n = 26), missense (n = 102), and truncating mutations (n = 23) in 80 primary tumours. (d) Missense dN/dS ratio for *MT-ND2*, *MT-ND4*, and *MT-ND5* suggest positive selection of missense mutations in these genes at relapse. *, $Q < 0.05$; ***, $Q < 0.001$; Vertical lines depict 95% CI. LOF: loss of function (i.e. truncating mutations), VAF: variant allele frequency, ns: not significant.

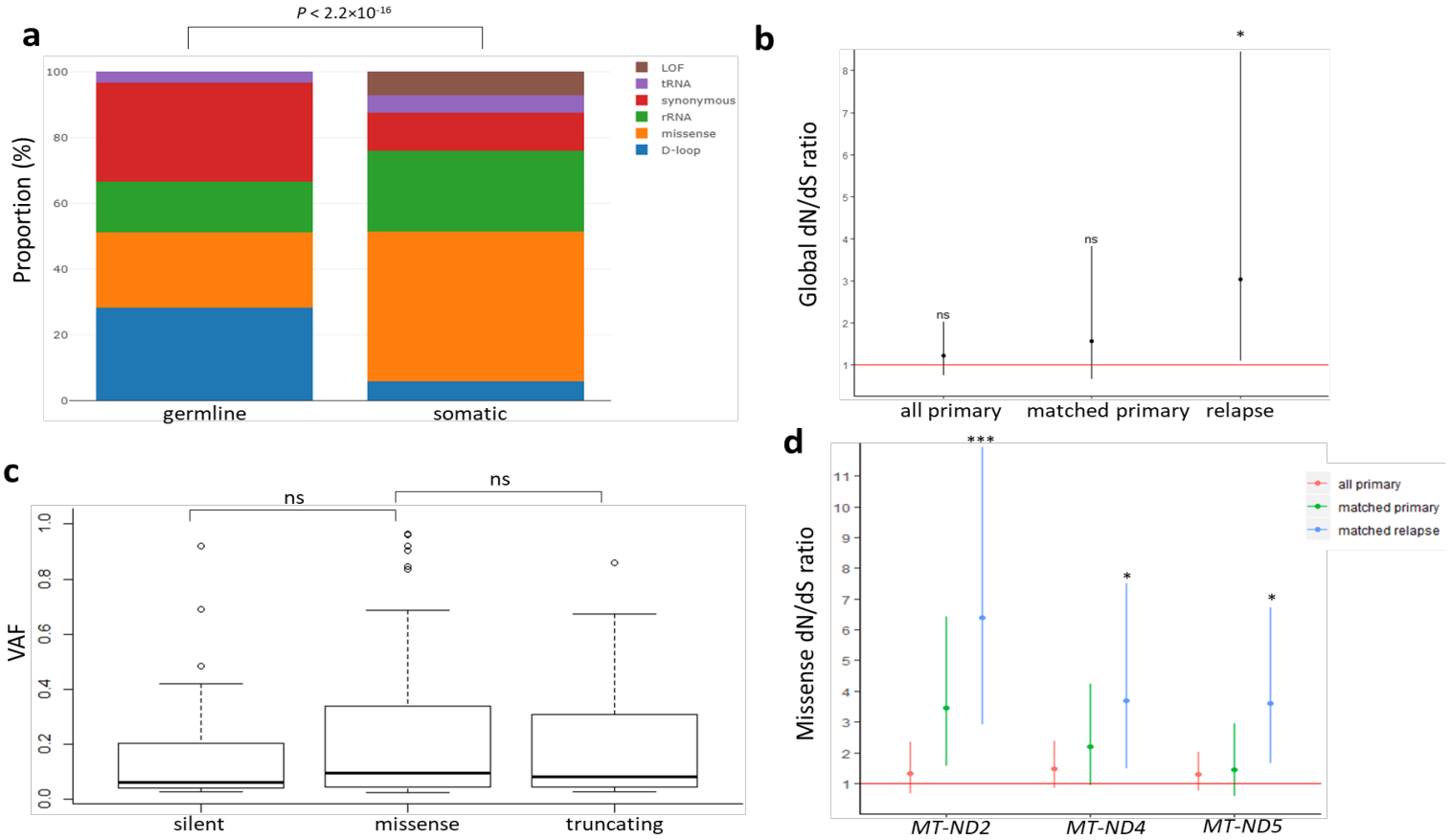


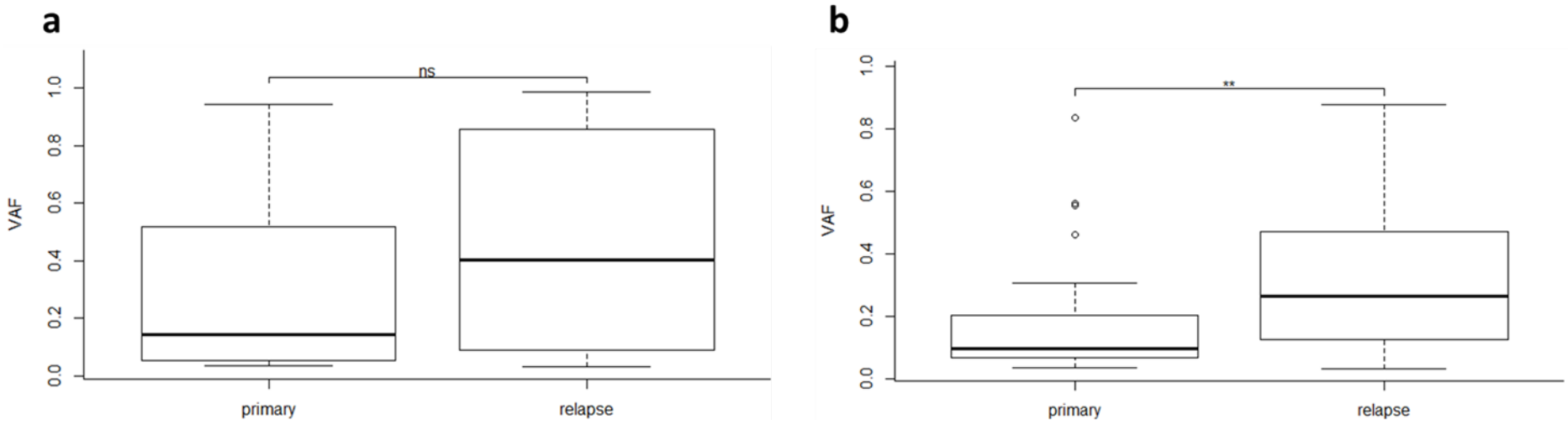
Table 6.3: Frequency of non-synonymous somatic mutations disrupting mtDNA coding gene in 80 primary tumours from Myeloma XI trial.

Gene	Non-synonymous mutations frequency	Proportion (%)
MT-ND5	23	29
MT-ND4	19	24
MT-CO1	16	20
MT-ND1	12	15
MT-CYB	11	14
MT-ND2	9	11
MT-CO3	7	9
MT-ATP6	5	6
MT-CO2	5	6
MT-ND6	4	5
MT-ATP8	2	3
MT-ND3	2	3

Table 6.4: Net increase of non-synonymous mutations disrupting mtDNA coding genes at relapse from Myeloma XI trial.

Gene	Primary frequency	Relapse frequency	Net increase frequency
MT-ND5	7	10	3
MT-CO3	2	4	2
MT-CO2	1	2	1
MT-ND2	6	7	1
MT-ND3	1	2	1
MT-ND6	1	2	1
MT-CYB	2	2	0
MT-ATP6	4	3	-1
MT-ND1	2	1	-1
MT-CO1	7	5	-2
MT-ND4	7	5	-2

Figure 6.7: Heteroplasmic level comparison between shared (a) silent mutations (n = 20) and (b) non-synonymous mutations (n = 47) in primary and matched relapsed tumours. Significant different was assessed using paired Wilcoxon rank-sum test. **: $P < 0.01$, ns: not significant. VAF, variant allele frequency.



6.3.3 mtDNA copy number and somatic transfer

The effects of mtDNA copy numbers in MM were next examined. No significant association was observed between mtDNA copy number of tumours and their matched normal, relapsed tumours versus primary tumours, or between high- and low-risk MM subtypes (Figure 6.8). The results therefore do not support pathogenic and prognostic contribution of mtDNA copy number in MM.

Somatic transfer of mtDNA to nuclear DNA was observed in 11/80 primary tumours and 6/25 relapsed tumours (Table 6.5). Transfer breakpoints disrupt open reading frames known oncogenes including *CENPP*, *FOXK1*, *MGAT5*, *ST8SIA1*, and *RAB4A*, suggesting their potential contribution in MM tumourigenesis.

Figure 6.8: Comparison of average mtDNA copy number between (a) normal and tumour, (b) primary and matched relapse tumours, and (c) high-risk [t(4;14) and t(16;14)] and low-risk [t(11;14)] multiple myeloma subtypes. Significant different was assessed using paired Wilcoxon rank-sum test. Ns, not significant.

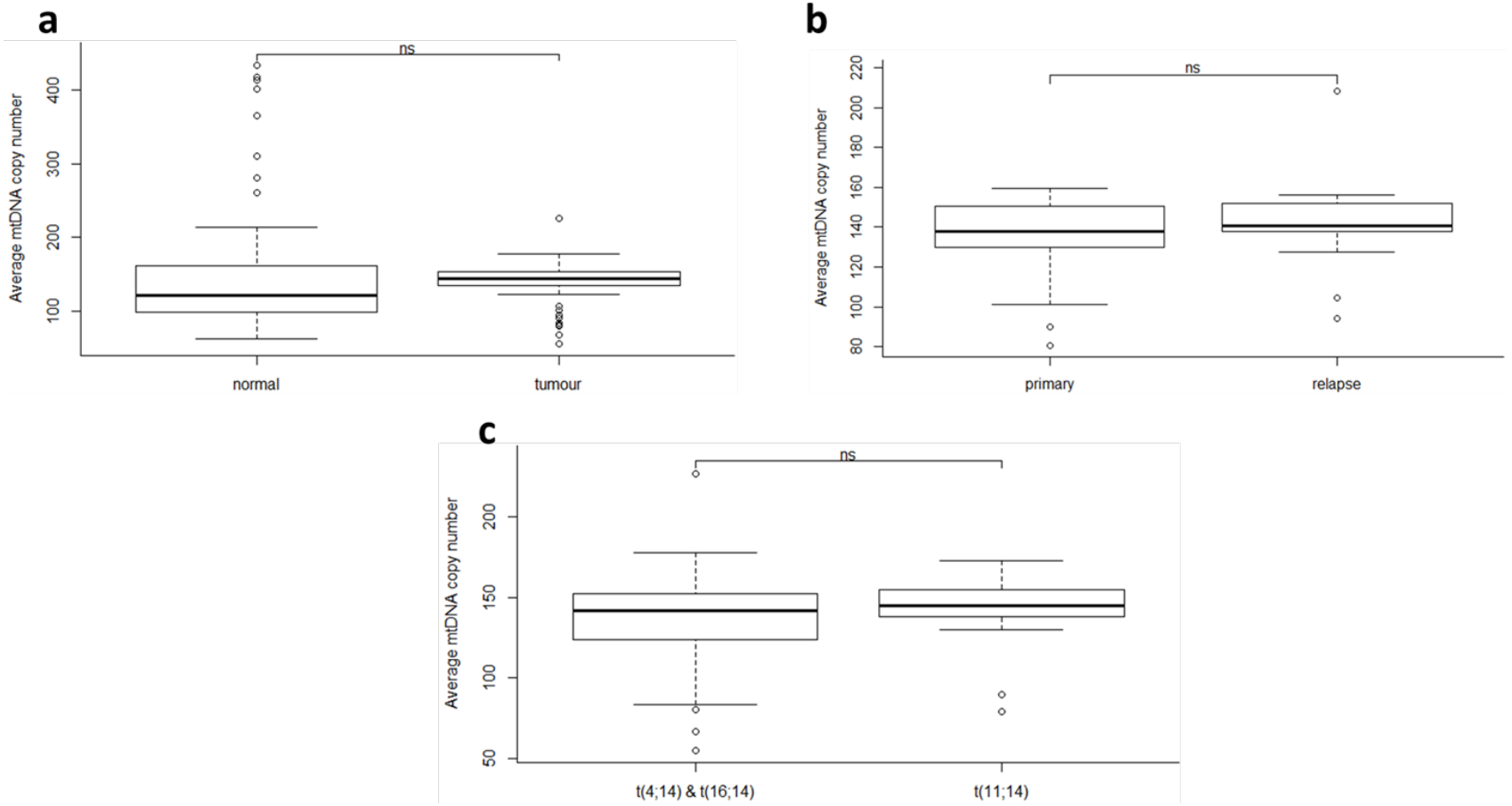


Table 6.5: Somatic nuclear transfer for (a) 80 primary tumours and (b) 25 relapsed tumours from Myeloma XI trial. Mito, mitochondria.

a

Sample	Mito position	Nuclear chromosome	Nuclear position	Mito gene	Nuclear genes disrupted
1305	15066	chr2	213419203	*CYTB*	*SPAG16*
6076	5427	chr4	75352750	*ND2*	*AC025244.1;AC025244.2*
6076	11723	chr7	4767594	*ND4*	*FOXK1*
6076	12835	chr4	29439843	*ND5*	
6076	14611	chr4	144032722	*ND6*	*AC139713.2*
7240	10303	chr4	33818874	*ND3*	*AC016687.3*
7240	12567	chr4	62999060	*ND5*	
7240	13790	chr3	157535410	*ND5*	
7915	10372	chr12	22293349	*ND3*	*ST8SIA1*
8043	11499	chr11	100144656	*ND4*	*CNTN5*
8043	14079	chr2	134151058	*ND5*	*MGAT5*
8237	16388	chr9	92344406		*CENPP*
9210	12835	chr4	29439433	*ND5*	
9524	568	chr2	32916230		*LINC00486*
9544	15178	chr5	144790245	*CYTB*	
10597	16199	chr15	90244106		*GDPGP1;AC091167.8*
12227	10879	chr8	86639771	*ND4*	*CNGB3*

b

Sample	Mito position	Nuclear chromosome	Nuclear position	Mito gene	Nuclear genes disrupted
6706	568	chr2	32916253		*LINC00486*
7801	12853	chr4	29439122	*ND5*	
9166	14777	chr2	33667485	*CYTB*	
9524	12406	chr10	20469830	*ND5*	
11949	9698	chr11	67779748	*COX3*	
12546	170	chr1	229287938		*RAB4A*
12546	12013	chr8	60096505	*ND4*	

6.4 Discussion

In this chapter, I present mtDNA mutational spectrum, the potential underlying mutational processes, and mechanisms in which they could contribute to MM development. I observed transcriptional strand bias of somatic mutations, suggesting transcription-coupled DNA repair defects as one of the main contributing mutational processes in MM mtDNA. This observation is consistent with the general observation of mitochondria having reduced DNA repair pathways[265-267]. As different defective transcription-coupled DNA repair processes have opposing transcriptional strand biases[263] and their contribution are varied across tumour types, the transcriptional strand bias might have been neutralised in previous pan-cancer study[124].

While mtDNA mutations are under strong negative selection in normal cells[186], there was no evidence supporting either negative or positive selection in primary MM. However, my results do support positive selection at relapse, potentially providing survival and resistance advantage for MM tumours. In consistent with this, significant dN/dS ratio was observed for missense mutations for genes comprising complex I (*MT-ND2*, *MT-ND4*, and *MT-ND5*); and mutations disrupting *MT-ND5* and *MT-CO3* (cytochrome c oxidase) are frequently acquired at relapse. Functional studies have suggested mutations impacting mitochondrial genes can recapitulate the Warburg effect and provide an alternative mechanism for tumour growth[268, 269]. Although mtDNA copy numbers do not have pathogenic or prognostic implication in MM, mitochondria-nuclear genome integration could potentially contribute to tumourigenesis through disruption of oncogenic genes (*e.g. CENPP*, *FOXK1*, *MGAT5*, *ST8SIA1*, *RAB4A*).

In summary, the findings provide evidence for altered metabolism through mitochondrial mutations disrupting electron transport chain, providing potential growth and resistance at relapse MM. Further studies are required to examine the clinical value of mitochondrial mutations as biomarkers and explore the therapeutic potential of targeting dysregulated metabolism in MM.

CHAPTER 7 General discussion, future work, and concluding remarks

7.1 Coding and non-coding drivers in multiple myeloma

The work presented in chapter 3 represents the first comprehensive study on non-coding drivers in MM using a large cohort from MMRF's CoMMpass study. Many of these targets have been validated subsequently in high-depth WGS data from Myeloma UK trial in chapter 6, including those associated with key genes in plasma cell differentiation pathway *PAX5* and *BCL6*.The lower coverage nature (8-12×) of the WGS dataset does, however, mean that many non-coding drivers identified are likely to arise during early tumourigenesis and with high mutational frequency.

While dysregulation of *MYC* through gene amplification and translocation mechanisms is well-established in MM[2] and various cancers[270], the work presented herein demonstrated novel alternative mechanisms including CNVs altering *MYC* non-coding regulatory regions (Chapter 3). It is therefore possible that many of the non-coding drivers in MM identified in chapter 3 and 6 could also be potential targets in other cancers.

From the integrative analysis of coding and non-coding drivers presented in chapter 3, I have highlighted several pathways key to MM, and that they can be targeted somatically through a range of mechanisms. This is notably exemplified by the plasma cell differentiation pathway, in which *IRF4* and *PRDM1* are frequently targeted in the coding regions, while *PAX5* and *BCL6* are primarily disrupted in the non-coding regulatory regions. The complementary genomic alteration impacting the same pathway was also demonstrated through the relative paucity of mutations in *PAX5* regulatory regions of in t(11:14) MM, but enrichment of *IRF4* coding mutations. Therefore, the findings from this study further highlight the importance to examine non-coding drivers in cancer to characterise targets for personalised treatment. In addition, it opens up potential opportunities for identifying novel therapeutic agents in MM through network-based drug search methodologies[220, 221].

In terms of future studies, it will be important to perform functional validations, including luciferase reporter assay and CRISPR-Cas9 knockout to confirm the *in vivo* regulatory roles of identified CREs. Furthermore, target genes could be knocked out or knocked in to analyse their effects of cell proliferation. Higher priority for functional validation would be given to more well-known targets implicated in MM and B-cell malignancies such as *MYC* and *PAX5*.

In addition, defining CREs through utilisation of patient-specific or MM cellular models, integrated with various ChIP-seq information, would enable us to fully and specifically recapitulate CREs spectrum relevant to MM. To take this forward at the Institute of Cancer research, promoter CHi-C are being performed on MM cellular models and ChIP-seq data are being collected from patient samples. Once the current methodology described in this thesis is established and validated with functional works, similar strategies could potentially be applied to identify non-coding mutation drivers on various types of cancers. Ideally, promoter CHi-C data could be generated from cancer cell lines while somatic mutations, RNA-seq, CNVs could be potentially be obtained from public dataset such as TCGA. Recent studies have also demonstrated an association between germline and somatic variants in various cancers[271-273]. It would be interesting to investigate in future studies whether such association exists in MM and which genes/pathways are complementarily affected by germline and somatic mechanisms.

7.2 Mutational processes in multiple myeloma

Prior to the work in this thesis, mutational signature analyses in MM were mostly restricted to WES data[2, 5, 84]. Therefore, through utilisation of large WGS dataset, my analysis has represented a comprehensive characterisation of mutational processes underlying MM development, with the contribution of varied processes in different MM subtypes and their implication in refining patient prognosis. The flat signatures (3, 5, and 8) account for > 20% of MM mutational contribution, suggesting DRD playing an important role in MM tumourigenesis. It will be important to develop algorithms to accurately resolve these flat COSMIC signatures as each could be extrinsically linked to deficiency in a specific DNA repair pathway. Successful differentiation of the flat signatures would, therefore,

provide further insight into pathway disrupted in MM and narrow down therapeutic targets.

Additionally, the work presented in chapter 4 suggests different MM subtypes are specifically associated with different mutational processes. However, further studies are required to elucidate the mechanistic insights on such association, *i.e.* whether the mutational processes are simply passive consequences or playing an active role in driving MM subtype differentiation.

Recently, mutational signatures extraction from large cohort of cancer WGS from PCAWG study has revealed novel signatures[263], with some are imprints of patients' therapies[274]. Many previously unknown COSMIC signatures have also been better established from functional studies[224-227, 275]. It is expected that future work will be involved with functional studies to fully elucidate and refine mutational signatures and their associated aetiologies. In addition, it will also be important to apply the extended COSMIC signatures framework on larger cohort of high-coverage WGS primary and relapsed tumours. Such efforts could lead to identification of novel mutational processes, especially those associated with later tumour development and specific treatment.

7.3 Tumour evolution at relapse

My study in chapter 6 on clonal evolution at relapse expands upon previous findings that have been based on WES/targeted sequencing[5, 6, 8, 102, 103], low coverage sequencing[104], or FISH/array technology[102, 105]. With larger cohort and high-depth WGS data from Myeloma XI trial[128], my data has afforded to identify frequently acquired coding and non-coding drivers, and refine complex genomic evolution patterns at relapse in MM. For instance, *CRBN* and those associated with Cullin-RING E3 ubiquitin ligase complex are unlikely a feature of relapse MM with immunomodulatory drugs therapy. Additionally, with more refined evolutionary pattern classification, an unprecedented association between subclonal expansion patterns and significantly shorter time to relapse was observed. To better understanding tumour evolution in MM, single-cell genomic sequencing methods[276] would be important to enable delineation of smaller subclones, spatial architecture of tumours[277], and differentiation of driver versus

passenger by quantifying fitness contribution of each individual mutation[2]. In addition, due to limited sample size and lack of WGS data[8, 106], there remains a knowledge gap in characterising genomic landscape present in the pre-malignant states MGUS and MM, and the mechanisms resulting in progression to MM across all disease subtypes.

The results from chapter 7 demonstrate the likely contributing roles of mitochondrial mutations in treatment-resistance and proliferation at MM relapse. Further functional studies, such as CRISPR-Cas9 knockout of mtDNA genes, will be required to fully establish the roles on mtDNA in MM.

7.4 Concluding remarks

Work carried out in this thesis has provided for a more comprehensive characterisation of the somatic mutations landscape across large MM cohorts and the aetiological mutational processes contributing to tumourigenesis. Firstly, the results presented have highlighted MM as a complex heterogenous malignancy with multiple oncogenic pathways are disrupted via various coding and non-coding somatic mutation mechanisms, as supported by data in chapter 3. Secondly, there are three principle mutational processes underlying MM tumourigenesis, namely AID/APOBEC, aging, and DRD. Intriguingly, although AID has large contribution in early mutational process, each MM subtype is predominantly associated with distinct mutational processes. In addition, incorporating mutational signatures information could potentially refine patient prognosis, beyond previously established risk factors. These are supported by data in chapter 4. Thirdly, the work presented in chapter 5 feature three distinct clonal evolutionary patterns at relapse, with one pattern is associated with worse prognosis. Relapsed MM is characterised with frequent acquisition of various coding and non-coding mutations, as well as CNVs at pre-existing unstable genomic regions. The findings from chapter 5 also suggest that the use of any targeted therapies would need to take into account of heterogenous clonal dynamics and the potential mutations acquired at relapse shaped in part by therapies. Fourthly, data detailed in chapter 6 do suggest the implication of targeting mitochondria specifically for relapsed MM.

While the results presented in this thesis are unlikely to constitute a complete model that explains MM tumourigenesis, they have provided a greater insight in describing and understanding the disease. Looking forward, further advancing the understanding of the genomic basis of MM offers clear opportunity for clinical benefits, in terms of identifying patients with high risk disease progression, devising kinder and more effective treatments using precision medicine, as well as developing strategies to overcome therapy-resistance. For instance, subgroup of patients with *BRAF* V600E activating mutation could potentially be treated with *BRAF* inhibitor vemurafenib[278]. However, *BRAF* inhibitors could activate the MAPK pathway resulting in treatment resistance if coexistent subclones harbour *KRAS*/*NRAS* mutations or wild-type *BRAF*[3, 279]. Such signalling interactions could be tackled by combining *BRAF* and *MEK* inhibitors[280, 281]. Therefore, a comprehensive molecular knowledge of clonal heterogeneity and evolution is crucial in aiding future targeted therapy approach in MM.

References

1. Walker, B.A. *et al.* Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma. *J Clin Oncol* **33**, 3911-3920 (2015).
2. Manier, S. *et al.* Genomic complexity of multiple myeloma and its clinical implications. *Nat Rev Clin Oncol* **14**, 100-113 (2017).
3. Lohr, J.G. *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25**, 91-101 (2014).
4. Hoang, P.H. *et al.* Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. *Leukemia* (2018).
5. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).
6. Kortum, K.M. *et al.* Targeted sequencing of refractory myeloma reveals a high incidence of mutations in CRBN and Ras pathway genes. *Blood* **128**, 1226-1233 (2016).
7. Hofman, I.J.F. *et al.* Low frequency mutations in ribosomal proteins RPL10 and RPL5 in multiple myeloma. *Haematologica* **102**, e317-e320 (2017).
8. Walker, B.A. *et al.* Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma. *Blood* **120**, 1077-1086 (2012).
9. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017).
10. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-598 (2014).
11. Shapiro-Shelef, M. & Calame, K. Regulation of plasma-cell development. *Nature reviews. Immunology* **5**, 230-242 (2005).
12. Group, T.I.M.W. Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group. *British Journal of Haematology* **121**, 749-757 (2003).
13. Kyle, R.A. & Rajkumar, S.V. Multiple Myeloma. *New England Journal of Medicine* **351**, 1860-1873 (2004).
14. Brenner, H., Gondos, A. & Pulte, D. Recent major improvement in long-term survival of younger patients with multiple myeloma. *Blood* **111**, 2521-2526 (2008).
15. Shapiro-Shelef, M. & Calame, K. Regulation of plasma-cell development. *Nature Reviews Immunology* **5**, 230-242 (2005).
16. Nutt, S.L., Hodgkin, P.D., Tarlinton, D.M. & Corcoran, L.M. The generation of antibody-secreting plasma cells. *Nat Rev Immunol* **15**, 160-171 (2015).
17. Fairfax, K.A., Kallies, A., Nutt, S.L. & Tarlinton, D.M. Plasma cell development: From B-cell subsets to long-term survival niches. *Seminars in Immunology* **20**, 49-58 (2008).
18. Martin, F., Oliver, A.M. & Kearney, J.F. Marginal Zone and B1 B Cells Unite in the Early Response against T-Independent Blood-Borne Particulate Antigens. *Immunity* **14**, 617-629 (2001).

19. LeBien, T.W. & Tedder, T.F. B lymphocytes: how they develop and function. *Blood* **112**, 1570-1580 (2008).
20. González, D. *et al.* Immunoglobulin gene rearrangements and the pathogenesis of multiple myeloma. *Blood* **110**, 3112-3121 (2007).
21. Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nature Reviews Immunology* **15**, 149-159 (2015).
22. Kyle, R.A. *et al.* A Long-Term Study of Prognosis in Monoclonal Gammopathy of Undetermined Significance. *New England Journal of Medicine* **346**, 564-569 (2002).
23. Kyle, R.A. *et al.* Clinical Course and Prognosis of Smoldering (Asymptomatic) Multiple Myeloma. *New England Journal of Medicine* **356**, 2582-2590 (2007).
24. Morgan, G.J., Walker, B.A. & Davies, F.E. The genetic architecture of multiple myeloma. *Nature Reviews Cancer* **12**, 335-348 (2012).
25. Morgan, G.J., Walker, B.A. & Davies, F.E. The genetic architecture of multiple myeloma. *Nature reviews. Cancer* **12**, 335-348 (2012).
26. Gould, J. *et al.* Plasma cell karyotype in multiple myeloma. *Blood* **71**, 453-456 (1988).
27. Sawyer, J.R., Waldron, J.A., Jagannath, S. & Barlogie, B. Cytogenetic findings in 200 patients with multiple myeloma. *Cancer Genet Cytogenet* **82**, 41-49 (1995).
28. Smadja, N.V. *et al.* Chromosomal analysis in multiple myeloma: cytogenetic evidence of two different diseases. *Leukemia* **12**, 960-969 (1998).
29. Kalff, A. & Spencer, A. The t(4;14) translocation and FGFR3 overexpression in multiple myeloma: prognostic implications and current clinical strategies. *Blood Cancer Journal* **2**, e89-e89 (2012).
30. Li, Z. *et al.* The myeloma-associated oncogene fibroblast growth factor receptor 3 is transforming in hematopoietic cells. *Blood* **97**, 2413-2419 (2001).
31. Qing, J. *et al.* Antibody-based targeting of FGFR3 in bladder carcinoma and t(4;14)-positive multiple myeloma in mice. *J Clin Invest* **119**, 1216-1229 (2009).
32. Martinez-Garcia, E. *et al.* The MMSET histone methyl transferase switches global histone methylation and alters gene expression in t(4;14) multiple myeloma cells. *Blood* **117**, 211-220 (2011).
33. Pei, H. *et al.* MMSET regulates histone H4K20 methylation and 53BP1 accumulation at DNA damage sites. *Nature* **470**, 124-128 (2011).
34. Hanamura, I. *et al.* Frequent gain of chromosome band 1q21 in plasma-cell dyscrasias detected by fluorescence in situ hybridization: incidence increases from MGUS to relapsed myeloma and is related to prognosis and disease progression following tandem stem-cell transplantation. *Blood* **108**, 1724-1732 (2006).
35. Walker, B.A. *et al.* A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood* **116**, e56-65 (2010).
36. Shaughnessy, J.D., Jr. *et al.* A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276-2284 (2007).
37. Boyd, K.D. *et al.* A novel prognostic model in myeloma based on co-segregating adverse FISH lesions and the ISS: analysis of patients treated in the MRC Myeloma IX trial. *Leukemia* **26**, 349-355 (2012).

38. Shaughnessy, J. Amplification and overexpression of CKS1B at chromosome band 1q21 is associated with reduced levels of p27Kip1 and an aggressive clinical course in multiple myeloma. *Hematology (Amsterdam, Netherlands)* **10 Suppl 1**, 117-126 (2005).
39. Boyd, K.D. *et al.* Mapping of chromosome 1p deletions in myeloma identifies FAM46C at 1p12 and CDKN2C at 1p32.3 as being genes in regions associated with adverse survival. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 7776-7784 (2011).
40. Chang, H. *et al.* Impact of genomic aberrations including chromosome 1 abnormalities on the outcome of patients with relapsed or refractory multiple myeloma treated with lenalidomide and dexamethasone. *Leukemia & lymphoma* **51**, 2084-2091 (2010).
41. Zhang, Q.Y., Yue, X.Q., Jiang, Y.P., Han, T. & Xin, H.L. FAM46C is critical for the anti-proliferation and pro-apoptotic effects of norcantharidin in hepatocellular carcinoma cells. *Scientific reports* **7**, 396 (2017).
42. Menges, C.W., Altomare, D.A. & Testa, J.R. FAS-Associated Factor 1 (FAF1): diverse functions and implications for oncogenesis. *Cell Cycle* **8**, 2528-2534 (2009).
43. Leone, P.E. *et al.* Deletions of CDKN2C in multiple myeloma: biological and clinical implications. *Clinical cancer research : an official journal of the American Association for Cancer Research* **14**, 6033-6041 (2008).
44. Abd El-Naby, A., Gawaly, A. & Elshweikh, S. CKS1B/CDKN2C (P18) amplification/deletion as prognostic markers in multiple myeloma patients. *The Egyptian Journal of Haematology* **41**, 87-93 (2016).
45. Fonseca, R. *et al.* Clinical and biologic implications of recurrent genomic aberrations in myeloma. *Blood* **101**, 4569-4575 (2003).
46. Avet-Loiseau, H. *et al.* Genetic abnormalities and survival in multiple myeloma: the experience of the Intergroupe Francophone du Myelome. *Blood* **109**, 3489-3495 (2007).
47. Drach, J. *et al.* Presence of a p53 gene deletion in patients with multiple myeloma predicts for short survival after conventional-dose chemotherapy. *Blood* **92**, 802-809 (1998).
48. Kastenhuber, E.R. & Lowe, S.W. Putting p53 in Context. *Cell* **170**, 1062-1078 (2017).
49. Tiedemann, R.E. *et al.* Genetic aberrations and survival in plasma cell leukemia. *Leukemia* **22**, 1044-1052 (2008).
50. Campo, E. *et al.* The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117**, 5019-5032 (2011).
51. IMWG Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group. *British journal of haematology* **121**, 749-757 (2003).
52. Rajkumar, S.V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *The Lancet. Oncology* **15**, e538-548 (2014).
53. Agarwal, A. & Ghobrial, I.M. Monoclonal gammopathy of undetermined significance and smoldering multiple myeloma: a review of the current understanding of epidemiology, biology, risk stratification, and management of myeloma precursor disease. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 985-994 (2013).

54. Jenner, E. Serum free light chains in clinical laboratory diagnostics. *Clinica chimica acta; international journal of clinical chemistry* **427**, 15-20 (2014).
55. Dispenzieri, A. *et al.* Immunoglobulin free light chain ratio is an independent risk factor for progression of smoldering (asymptomatic) multiple myeloma. *Blood* **111**, 785-789 (2008).
56. Rajkumar, S.V. *et al.* Serum free light chain ratio is an independent risk factor for progression in monoclonal gammopathy of undetermined significance. *Blood* **106**, 812-817 (2005).
57. Greipp, P.R. *et al.* International staging system for multiple myeloma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **23**, 3412-3420 (2005).
58. Zhou, Y., Barlogie, B. & Shaughnessy, J.D., Jr. The molecular characterization and clinical management of multiple myeloma in the post-genome era. *Leukemia* **23**, 1941-1956 (2009).
59. Bergsagel, P.L., Mateos, M.V., Gutierrez, N.C., Rajkumar, S.V. & San Miguel, J.F. Improving overall survival and overcoming adverse prognosis in the treatment of cytogenetically high-risk multiple myeloma. *Blood* **121**, 884-892 (2013).
60. Neben, K. *et al.* Administration of bortezomib before and after autologous stem cell transplantation improves outcome in multiple myeloma patients with deletion 17p. *Blood* **119**, 940-948 (2012).
61. Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020-2028 (2006).
62. Bergsagel, P.L. *et al.* Cyclin D dysregulation: an early and unifying pathogenic event in multiple myeloma. *Blood* **106**, 296-303 (2005).
63. Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020-2028 (2006).
64. Kumar, S.K. *et al.* Management of newly diagnosed symptomatic multiple myeloma: updated Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) consensus guidelines. *Mayo Clinic proceedings* **84**, 1095-1110 (2009).
65. Smith, D. & Yong, K. Multiple myeloma. *Bmj* **346**, f3863 (2013).
66. Bird, J.M. *et al.* Guidelines for the diagnosis and management of multiple myeloma 2011. *British journal of haematology* **154**, 32-75 (2011).
67. Oakervee, H.E. *et al.* PAD combination therapy (PS-341/bortezomib, doxorubicin and dexamethasone) for previously untreated patients with multiple myeloma. *British journal of haematology* **129**, 755-762 (2005).
68. Popat, R. *et al.* Bortezomib, doxorubicin and dexamethasone (PAD) front-line treatment of multiple myeloma: updated results after long-term follow-up. *British journal of haematology* **141**, 512-516 (2008).
69. Siegel, D.S. *et al.* A phase 2 study of single-agent carfilzomib (PX-171-003-A1) in patients with relapsed and refractory multiple myeloma. *Blood* **120**, 2817-2825 (2012).
70. Lacy, M.Q. *et al.* Pomalidomide (CC4047) plus low-dose dexamethasone as therapy for relapsed multiple myeloma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 5008-5014 (2009).
71. Naymagon, L. & Abdul-Hay, M. Novel agents in the treatment of multiple myeloma: a review about the future. *Journal of hematology & oncology* **9**, 52 (2016).
72. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719-724 (2009).

73. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
74. Druker, B.J. *et al.* Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *New England Journal of Medicine* **355**, 2408-2417 (2006).
75. Kantarjian, H. *et al.* Improved survival in chronic myeloid leukemia since the introduction of imatinib therapy: a single-institution historical experience. *Blood* **119**, 1981-1987 (2012).
76. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021 (2017).
77. Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R. & Easton, D.F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187-2198 (2006).
78. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
79. Yang, Z., Ro, S. & Rannala, B. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics* **165**, 695-705 (2003).
80. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-811 (2015).
81. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell* **156**, 1324-1335 (2014).
82. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**, A68-A77 (2015).
83. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).
84. Walker, B.A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat Commun* **6**, 6997 (2015).
85. Pfeifer, G.P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**, 7435-7451 (2002).
86. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
87. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
88. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-259 (2013).
89. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nature Genetics* **47**, 1402-1407 (2015).
90. Alexandrov, L.B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622 (2016).
91. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).
92. Letouze, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun* **8**, 1315 (2017).
93. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).

94. Tabin, C.J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300**, 143-149 (1982).
95. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306-313 (2012).
96. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558 (2013).
97. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915 (2010).
98. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology* **30**, 413-421 (2012).
99. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* **26**, 64-70 (2014).
100. Dentro, S.C., Wedge, D.C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb Perspect Med* **7,** 8 (2017).
101. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396-398 (2014).
102. Weinhold, N. *et al.* Clonal selection and double-hit events involving tumor suppressor genes underlie relapse in myeloma. *Blood* **128**, 1735-1744 (2016).
103. Corre, J. *et al.* Multiple myeloma clonal evolution in homogeneously treated patients. *Leukemia* **32**, 2636-2647 (2018).
104. Egan, J.B. *et al.* Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood* **120**, 1060-1066 (2012).
105. Keats, J.J. *et al.* Clonal competition with alternating dominance in multiple myeloma. *Blood* **120**, 1067-1076 (2012).
106. Bolli, N. *et al.* Genomic patterns of progression in smoldering multiple myeloma. *Nature Communications* **9**, 3363 (2018).
107. Melchor, L. *et al.* Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia* **28**, 1705-1715 (2014).
108. Clay Montier, L.L., Deng, J.J. & Bai, Y. Number matters: control of mammalian mitochondrial DNA copy number. *Journal of Genetics and Genomics* **36**, 125-131 (2009).
109. Wachsmuth, M., Hubner, A., Li, M., Madea, B. & Stoneking, M. Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number. *PLoS Genet* **12**, e1005939 (2016).
110. Gustafsson, C.M., Falkenberg, M. & Larsson, N.-G. Maintenance and Expression of Mammalian Mitochondrial DNA. *Annual Review of Biochemistry* **85**, 133-160 (2016).
111. Gammage, P.A. & Frezza, C. Mitochondrial DNA: the overlooked oncogenome? *BMC Biology* **17**, 53 (2019).
112. Porporato, P.E., Filigheddu, N., Pedro, J.M.B., Kroemer, G. & Galluzzi, L. Mitochondrial metabolism and cancer. *Cell Res* **28**, 265-280 (2018).
113. Vander Heiden, M.G., Cantley, L.C. & Thompson, C.B. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science* **324**, 1029-1033 (2009).
114. Cairns, R.A., Harris, I.S. & Mak, T.W. Regulation of cancer cell metabolism. *Nat Rev Cancer* **11**, 85-95 (2011).

115. Greaves, L.C. & Taylor, R.W. Mitochondrial DNA mutations in human disease. *IUBMB Life* **58**, 143-151 (2006).
116. Koppenol, W.H., Bounds, P.L. & Dang, C.V. Otto Warburg's contributions to current concepts of cancer metabolism. *Nat Rev Cancer* **11**, 325-337 (2011).
117. Hengartner, M.O. The biochemistry of apoptosis. *Nature* **407**, 770-776 (2000).
118. Song, I.S. *et al.* Mitochondrial modulation decreases the bortezomib-resistance in multiple myeloma cells. *International Journal of Cancer* **133**, 1357-1367 (2013).
119. Zhan, X. *et al.* Alteration of mitochondrial biogenesis promotes disease progression in multiple myeloma. *Oncotarget* **8** (2017).
120. Chanan-Khan, A.A., Borrello, I., Lee, K.P. & Reece, D.E. Development of target-specific treatments in multiple myeloma. *British Journal of Haematology* **151**, 3-15 (2010).
121. Bahlis, N.J. *et al.* Feasibility and Correlates of Arsenic Trioxide Combined with Ascorbic Acid-mediated Depletion of Intracellular Glutathione for the Treatment of Relapsed/Refractory Multiple Myeloma. *Clinical Cancer Research* **8**, 3658-3668 (2002).
122. Larman, T.C. *et al.* Spectrum of somatic mitochondrial mutations in five cancers. *Proc Natl Acad Sci U S A* **109**, 14087-14091 (2012).
123. Stewart, J.B. *et al.* Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers. *PLoS Genet* **11**, e1005333 (2015).
124. Ju, Y.S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3** (2014).
125. Ju, Y.S. *et al.* Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. *Genome Res* **25**, 814-824 (2015).
126. Reznik, E. *et al.* Mitochondrial DNA copy number variation across human cancers. *Elife* **5** (2016).
127. Hoang, P.H., Cornish, A.J., Dobbins, S.E., Kaiser, M. & Houlston, R.S. Mutational processes contributing to the development of multiple myeloma. *Blood Cancer Journal* **9**, 60 (2019).
128. Jackson, G.H. *et al.* Lenalidomide maintenance versus observation for patients with newly diagnosed multiple myeloma (Myeloma XI): a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol* **20**, 57-73 (2019).
129. Manojlovic, Z. *et al.* Comprehensive molecular profiling of 718 Multiple Myelomas reveals significant differences in mutation frequencies between African and European descent cases. *PLoS Genet* **13**, e1007087 (2017).
130. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
131. Miller, C. *et al.* A Comparison of Clinical FISH and Sequencing Based FISH Estimates in Multiple Myeloma: An Mmrf Commpass Analysis. *Blood* **128**, 374-374 (2016).
132. Shah, V. *et al.* Prediction of outcome in newly diagnosed myeloma: a meta-analysis of the molecular profiles of 1905 trial patients. *Leukemia* **32**, 102-110 (2018).
133. Kaiser, M.F. *et al.* A TC classification-based predictor for multiple myeloma using multiplexed real-time quantitative PCR. *Leukemia* **27**, 1754-1757 (2013).

134. R Development Core Team (R Foundation for Statistical Computing, Vienna, Austria; 2013).
135. Karolchik, D., Hinrichs, A.S. & Kent, W.J. The UCSC Genome Browser. *Curr Protoc Bioinformatics* **Chapter 1**, Unit1 4 (2012).
136. Coordinators, N.R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **41**, D8-D20 (2013).
137. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
138. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
139. Karczewski, K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210 (2019).
140. Flicek, P. *et al.* Ensembl 2014. *Nucleic acids research* **42**, D749-755 (2014).
141. Fernandez, J.M. *et al.* The BLUEPRINT Data Analysis Portal. *Cell Syst* **3**, 491-495 e495 (2016).
142. Wallace, D.C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harb Perspect Biol* **5**, a021220 (2013).
143. Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. & Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767-1771 (2010).
144. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
145. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
146. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013).
147. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
148. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
149. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **11**, 11 10 11-11 10 33 (2013).
150. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178 (2018).
151. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054 (2019).
152. Farmery, J.H.R., Smith, M.L., Diseases, N.B.-R. & Lynch, A.G. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep* **8**, 1300 (2018).
153. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
154. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e1319 (2016).
155. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**, 127 (2016).

156. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065 (2015).
157. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019).
158. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **41**, e67 (2013).
159. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012).
160. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
161. Ramos, A.H. *et al.* Oncotator: cancer variant annotation tool. *Hum Mutat* **36**, E2423-2429 (2015).
162. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
163. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84 (2014).
164. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
165. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).
166. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).
167. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677 (2013).
168. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226-1236 (2013).
169. Cortés-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *bioRxiv*, 333617 (2018).
170. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501-504 (2005).
171. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55-60 (2017).
172. Melton, C., Reuter, J.A., Spacek, D.V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet* **47**, 710-716 (2015).
173. Hansen, R.S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**, 139-144 (2010).
174. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165 (2014).
175. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
176. Carlson, M.R., Pages, H., Arora, S., Obenchain, V. & Morgan, M. Genomic Annotation Resources in R/Bioconductor. *Methods Mol Biol* **1418**, 67-90 (2016).

177. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).
178. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).
179. Shinde, J. *et al.* Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics* (2018).
180. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
181. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
182. Raine, K.M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current Protocols in Bioinformatics* **56**, 15.19.11-15.19.17 (2016).
183. Yuan, K., Macintyre, G., Liu, W. & Markowetz, F. Ccube: A fast and robust method for estimating cancer cell fractions. *bioRxiv*, 484402 (2018).
184. Caravagna, G. *et al.* Model-based tumor subclonal reconstruction. *bioRxiv*, 586560 (2019).
185. Adalsteinsson, V.A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).
186. Triska, P. *et al.* Landscape of Germline and Somatic Mitochondrial DNA Mutations in Pediatric Malignancies. *Cancer Res* **79**, 1318-1330 (2019).
187. Lott, M.T. *et al.* mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr Protoc Bioinformatics* **44**, 1 23 21-26 (2013).
188. Shen, L. *et al.* MSeqDR: A Centralized Knowledge Repository and Bioinformatics Web Resource to Facilitate Genomic Investigations in Mitochondrial Disease. *Hum Mutat* **37**, 540-548 (2016).
189. Sonney, S. *et al.* Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. *PLoS Comput Biol* **13**, e1005867 (2017).
190. Qian, Y. *et al.* fastMitoCalc: an ultra-fast program to estimate mitochondrial DNA copy number from whole-genome sequences. *Bioinformatics* **33**, 1399-1401 (2017).
191. Guo, Y., Li, J., Li, C.-I., Shyr, Y. & Samuels, D.C. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* **29**, 1210-1211 (2013).
192. Keats, J.J. *et al.* Molecular Predictors of Outcome and Drug Response in Multiple Myeloma: An Interim Analysis of the Mmrf CoMMpass Study. *Blood* **128**, 194-194 (2016).
193. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34 (2014).
194. Stromberg, M. *et al.* in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics 596-596 (ACM, Boston, Massachusetts, USA; 2017).
195. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).
196. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487 (2016).

197. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst* **1**, 210-223 (2015).
198. Meza-Zepeda, L.A. *et al.* Positional cloning identifies a novel cyclophilin as a candidate amplified oncogene in 1q21. *Oncogene* **21**, 2261-2269 (2002).
199. Petroziello, J. *et al.* Suppression subtractive hybridization and expression profiling identifies a unique set of genes overexpressed in non-small-cell lung cancer. *Oncogene* **23**, 7734-7745 (2004).
200. Zhou, F. *et al.* NBPF is a potential DNA-binding transcription factor that is directly regulated by NF-kappaB. *Int J Biochem Cell Biol* **45**, 2479-2490 (2013).
201. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519-524 (2015).
202. Dang, J. *et al.* PAX5 is a tumor suppressor in mouse mutagenesis models of acute lymphoblastic leukemia. *Blood* **125**, 3609-3617 (2015).
203. Shah, S. *et al.* A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat Genet* **45**, 1226-1231 (2013).
204. Lu, J. & Gu, J. Significance of beta-Galactoside alpha2,6 Sialyltranferase 1 in Cancers. *Molecules* **20**, 7509-7527 (2015).
205. Mittermayr, S. *et al.* Polyclonal Immunoglobulin G N-Glycosylation in the Pathogenesis of Plasma Cell Disorders. *J Proteome Res* **16**, 748-762 (2017).
206. Aurer, I. *et al.* Aberrant glycosylation of Igg heavy chain in multiple myeloma. *Coll Antropol* **31**, 247-251 (2007).
207. Han, S.H. *et al.* Cobll1 is linked to drug resistance and blastic transformation in chronic myeloid leukemia. *Leukemia* **31**, 1532-1539 (2017).
208. Broyl, A. *et al.* Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood* **116**, 2543-2553 (2010).
209. Lindblad, O. *et al.* The role of HOXB2 and HOXB3 in acute myeloid leukemia. *Biochem Biophys Res Commun* **467**, 742-747 (2015).
210. Max EE, F.S. *Immunoglobulins: molecular genetics*. (Philidelphia: Lippincott Williams & Wilkins, 2013).
211. Walker, B.A. *et al.* Translocations at 8q24 juxtapose MYC with genes that harbor superenhancers resulting in overexpression and poor prognosis in myeloma patients. *Blood Cancer J* **4**, e191 (2014).
212. Nagoshi, H. *et al.* Frequent PVT1 rearrangement and novel chimeric genes PVT1-NBEA and PVT1-WWOX occur in multiple myeloma with 8q24 abnormality. *Cancer Res* **72**, 4954-4962 (2012).
213. Maura, F. *et al.* Analysis of Mutational Signatures Suggest That Aid Has an Early and Driver Role in Multiple Myeloma. *Blood* **128**, 116-116 (2016).
214. Qian, J. *et al.* B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524-1537 (2014).
215. Roberts, S.A. & Gordenin, D.A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer* **14**, 786-800 (2014).
216. Kumar, S.K. *et al.* Multiple myeloma. *Nat Rev Dis Primers* **3**, 17046 (2017).
217. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515-520 (2018).

218. Lonial, S. *et al.* Interim Analysis of the Mmrf Commpass Trial: Identification of Novel Rearrangements Potentially Associated with Disease Initiation and Progression. *Blood* **124**, 722-722 (2014).
219. Keim, C., Kazadi, D., Rothschild, G. & Basu, U. Regulation of AID, the B-cell genome mutator. *Genes Dev* **27**, 1-17 (2013).
220. Patel, M.N., Halling-Brown, M.D., Tym, J.E., Workman, P. & Al-Lazikani, B. Objective assessment of cancer genes for drug discovery. *Nat Rev Drug Discov* **12**, 35-50 (2013).
221. Mitsopoulos, C., Schierz, A.C., Workman, P. & Al-Lazikani, B. Distinctive Behaviors of Druggable Proteins in Cellular Networks. *PLoS Comput Biol* **11**, e1004597 (2015).
222. Pfeifer, G.P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med* **2**, 54 (2010).
223. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat Commun* **7**, 11383 (2016).
224. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat Commun* **9**, 1744 (2018).
225. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606 (2016).
226. Jager, M. *et al.* Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res* (2019).
227. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234-238 (2017).
228. Di Noia, J.M. & Neuberger, M.S. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* **76**, 1-22 (2007).
229. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).
230. Maura, F. *et al.* Biological and prognostic impact of APOBEC-induced mutations in the spectrum of plasma cell dyscrasias and multiple myeloma cell lines. *Leukemia* **32**, 1044-1048 (2018).
231. Haradhvala, N.J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-549 (2016).
232. Park, C., Qian, W. & Zhang, J. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep* **13**, 1123-1129 (2012).
233. Chen, L. *et al.* Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma. *Blood* **115**, 61-70 (2010).
234. Reimold, A.M. *et al.* Plasma cell differentiation requires the transcription factor XBP-1. *Nature* **412**, 300-307 (2001).
235. Leone, E. *et al.* Targeting miR-21 inhibits in vitro and in vivo multiple myeloma cell growth. *Clin Cancer Res* **19**, 2096-2106 (2013).
236. Felix, R.S. *et al.* SAGE analysis highlights the importance of p53csv, ddx5, mapkapk2 and ranbp2 to multiple myeloma tumorigenesis. *Cancer Lett* **278**, 41-48 (2009).
237. Ashby, C.C. *et al.* Whole Genome Sequencing Reveals the Extent of Structural Variants in Multiple Myeloma and Identifies Recurrent Mutational Hotspots within the Non-Coding Regions. *Blood* **130**, 3032-3032 (2017).

238. Ross, F.M. *et al.* Age has a profound effect on the incidence and significance of chromosome abnormalities in myeloma. *Leukemia* **19**, 1634 (2005).
239. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat Commun* **10**, 2969 (2019).
240. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393-395 (2009).
241. Nambiar, M. & Raghavan, S.C. How does DNA break during chromosomal translocations? *Nucleic Acids Res* **39**, 5813-5825 (2011).
242. Onodera, N., McCabe, N.R. & Rubin, C.M. Formation of a hyperdiploid karyotype in childhood acute lymphoblastic leukemia. *Blood* **80**, 203-208 (1992).
243. Castedo, M. *et al.* Cell death by mitotic catastrophe: a molecular definition. *Oncogene* **23**, 2825-2837 (2004).
244. Ly, D.H., Lockhart, D.J., Lerner, R.A. & Schultz, P.G. Mitotic misregulation and human aging. *Science* **287**, 2486-2492 (2000).
245. Tower, J. Programmed cell death in aging. *Ageing Res Rev* **23**, 90-100 (2015).
246. Wala, J.A. *et al.* Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv*, 187609 (2017).
247. Walker, B.A. *et al.* Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. *Blood* **132**, 587-597 (2018).
248. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
249. Aktas Samur, A. *et al.* Deciphering the chronology of copy number alterations in Multiple Myeloma. *Blood Cancer J* **9**, 39 (2019).
250. Kaufmann, H. *et al.* Both IGH translocations and chromosome 13q deletions are early events in monoclonal gammopathy of undetermined significance and do not evolve during transition to multiple myeloma. *Leukemia* **18**, 1879-1882 (2004).
251. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet* **49**, 1476-1486 (2017).
252. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510 (2012).
253. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).
254. Hochhaus, A. *et al.* Molecular and chromosomal mechanisms of resistance to imatinib (STI571) therapy. *Leukemia* **16**, 2190-2196 (2002).
255. Ikeda, H. *et al.* Molecular diagnostics of a single drug-resistant multiple myeloma case using targeted next-generation sequencing. *Onco Targets Ther* **8**, 2805-2815 (2015).
256. Fanale, D. *et al.* Stabilizing versus destabilizing the microtubules: a double-edge sword for an effective cancer treatment option? *Anal Cell Pathol (Amst)* **2015**, 690916 (2015).
257. Kino, K. & Sugiyama, H. UVR-induced G-C to C-G transversions from oxidative DNA damage. *Mutat Res* **571**, 33-42 (2005).
258. Liou, G.Y. & Storz, P. Reactive oxygen species in cancer. *Free Radic Res* **44**, 479-496 (2010).
259. Kondo, N., Takahashi, A., Ono, K. & Ohnishi, T. DNA damage induced by alkylating agents and repair pathways. *J Nucleic Acids* **2010**, 543531 (2010).

260. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
261. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
262. Andrews, R.M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147 (1999).
263. Alexandrov, L.B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*, 322859 (2019).
264. Gorman, G.S. *et al.* Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann Neurol* **77**, 753-759 (2015).
265. Clayton, D.A., Doda, J.N. & Friedberg, E.C. The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. *Proc Natl Acad Sci U S A* **71**, 2777-2781 (1974).
266. Miyaki, M., Yatagai, K. & Ono, T. Strand breaks of mammalian mitochondrial DNA induced by carcinogens. *Chem Biol Interact* **17**, 321-329 (1977).
267. Alexeyev, M., Shokolenko, I., Wilson, G. & LeDoux, S. The maintenance of mitochondrial DNA integrity--critical analysis and update. *Cold Spring Harb Perspect Biol* **5**, a012641 (2013).
268. Park, J.S. *et al.* A heteroplasmic, not homoplasmic, mitochondrial DNA mutation promotes tumorigenesis via alteration in reactive oxygen species generation and apoptosis. *Human Molecular Genetics* **18**, 1578-1589 (2009).
269. Hofhaus, G. & Attardi, G. Efficient selection and characterization of mutants of a human cell line which are defective in mitochondrial DNA-encoded subunits of respiratory NADH dehydrogenase. *Mol Cell Biol* **15**, 964-974 (1995).
270. Kalkat, M. *et al.* MYC Deregulation in Primary Human Cancers. *Genes (Basel)* **8** (2017).
271. Middlebrooks, C.D. *et al.* Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat Genet* **48**, 1330-1338 (2016).
272. Bailey, S.D. *et al.* Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. *Nat Genet* **48**, 1260-1266 (2016).
273. Kanchi, K.L. *et al.* Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* **5**, 3156 (2014).
274. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nature Genetics* **51**, 1732-1740 (2019).
275. Kucab, J.E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e816 (2019).
276. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
277. Ren, X., Kang, B. & Zhang, Z. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biology* **19**, 211 (2018).
278. Andrulis, M. *et al.* Targeting the BRAF V600E Mutation in Multiple Myeloma. *Cancer Discovery* **3**, 862-869 (2013).

279. Poulikakos, P.I., Zhang, C., Bollag, G., Shokat, K.M. & Rosen, N. RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. *Nature* **464**, 427-430 (2010).
280. Mey, U.J.M., Renner, C. & von Moos, R. Vemurafenib in combination with cobimetinib in relapsed and refractory extramedullary multiple myeloma harboring the BRAF V600E mutation. *Hematological Oncology* **35**, 890-893 (2017).
281. Broman, K.K., Dossett, L.A., Sun, J., Eroglu, Z. & Zager, J.S. Update on BRAF and MEK inhibition for treatment of melanoma in metastatic, unresectable, and adjuvant settings. *Expert Opinion on Drug Safety* **18**, 381-392 (2019).

Appendix 1: Results of Reactome integrated pathway analysis (Chapter 3). ($Q < 0.05$)

Pathway name	*Q*-value	Pathway classification
Activation of IRF3/IRF7 mediated by TBK1/IKK epsilon	1.08E-02	Toll-like receptors cascade
Formation of Senescence-Associated Heterochromatin Foci (SAHF)	1.08E-02	DNA damage
MAPK family signaling cascades	1.08E-02	MAPK signalling pathway
Regulation of TP53 Expression	1.08E-02	TP53 regulation pathway
Signaling by FGFR1	1.08E-02	FGFR signalling pathway
Signaling by FGFR3	1.08E-02	FGFR signalling pathway
Signaling by FGFR4	1.08E-02	FGFR signalling pathway
Signaling by RAS mutants	1.08E-02	MAPK signalling pathway
Tie2 Signaling	1.08E-02	Signalling for vascular and hematopoietic development
TNF receptor superfamily (TNFSF) members mediating non-canonical NF-kB pathway	1.08E-02	Non-canonical NF-kB signaling pathway
Deubiquitination	1.10E-02	Post-translational protein modification
FRS-mediated FGFR1 signaling	1.10E-02	FGFR signalling pathway
FRS-mediated FGFR3 signaling	1.10E-02	FGFR signalling pathway
FRS-mediated FGFR4 signaling	1.10E-02	FGFR signalling pathway
RAF activation	1.10E-02	MAPK signalling pathway
RAS signaling downstream of NF1 loss-of-function variants	1.10E-02	MAPK signalling pathway
Regulation of RAS by GAPs	1.10E-02	MAPK signalling pathway
Ub-specific processing proteases	1.10E-02	Post-translational protein modification
Transcriptional regulation by RUNX3	1.15E-02	Transcriptional regulation
Cytokine Signaling in Immune system	1.18E-02	Cytokine signalling in immune system/Immune system
Downstream signaling of activated FGFR3	1.18E-02	FGFR signalling pathway
FRS-mediated FGFR2 signaling	1.18E-02	FGFR signalling pathway
TP53 Regulates Transcription of Genes Involved in Cytochrome C Release	1.18E-02	Apoptosis
Cyclin A:Cdk2-associated events at S phase entry	1.22E-02	Cell cycle
Cyclin E associated events during G1/S transition	1.22E-02	Cell cycle
DDX58/IFIH1-mediated induction of interferon-alpha/beta	1.22E-02	Immune system
Downstream signal transduction	1.22E-02	FGFR signalling pathway
Downstream signaling of activated FGFR4	1.22E-02	FGFR signalling pathway
GRB2 events in EGFR signaling	1.22E-02	MAPK signalling pathway
Oncogenic MAPK signaling	1.22E-02	MAPK signalling pathway
Signaling by FGFR2	1.22E-02	FGFR signalling pathway
Signalling to ERKs	1.22E-02	MAPK signalling pathway
SOS-mediated signalling	1.22E-02	MAPK signalling pathway
Activation of RAS in B cells	1.29E-02	Signalling in B cells
Binding of TCF/LEF:CTNNB1 to target gene promoters	1.29E-02	MYC regulation
Downstream signaling of activated FGFR1	1.29E-02	FGFR signalling pathway
Downstream signaling of activated FGFR2	1.29E-02	FGFR signalling pathway
Apoptosis	1.39E-02	Apoptosis
Calcitonin-like ligand receptors	1.39E-02	GPCR signalling pathway
EGFR Transactivation by Gastrin	1.39E-02	MAPK signalling pathway
MAP2K and MAPK activation	1.39E-02	RNA metabolism
Negative regulation of MAPK pathway	1.39E-02	MAPK signalling pathway
RUNX3 regulates WNT signaling	1.39E-02	Transcriptional regulation
SHC1 events in EGFR signaling	1.39E-02	MAPK signalling pathway
Signaling by high-kinase activity BRAF mutants	1.39E-02	MAPK signalling pathway
Signaling by moderate kinase activity BRAF mutants	1.39E-02	MAPK signalling pathway
TNFR2 non-canonical NF-kB pathway	1.39E-02	Non-canonical NF-kB signaling pathway
Diseases of signal transduction	1.42E-02	Signalling pathway
Paradoxical activation of RAF signaling by kinase inactive BRAF	1.43E-02	MAPK signalling pathway
Signaling by FGFR	1.49E-02	Receptor tyrosine kinase signalling pathways
MAPK1/MAPK3 signaling	1.53E-02	MAPK signalling pathway
Programmed Cell Death	1.53E-02	Apoptosis
Signaling by SCF-KIT	1.53E-02	SCF/KIT signalling pathway
Signaling by EGFR	1.59E-02	EGFR signalling pathway
TICAM1-dependent activation of IRF3/IRF7	1.66E-02	Toll-like receptors cascade
TNFR1-induced proapoptotic signaling	1.66E-02	Death receptor signalling
Interleukin-20 family signaling	1.66E-02	Cytokine signalling in immune system/Immune system
MET activates RAS signaling	1.66E-02	MAPK signalling pathway
SHC-related events triggered by IGF1R	1.66E-02	MAPK signalling pathway
Repression of WNT target genes	1.88E-02	MYC regulation
Signaling by FGFR3 fusions in cancer	1.88E-02	FGFR signalling pathway
IRS-mediated signalling	2.18E-02	Cytokine signalling in immune system/Immune system
IRS-related events triggered by IGF1R	2.26E-02	Insulin receptor signalling pathway

Pathway name	*Q*-value	Pathway classification
p38MAPK events	2.26E-02	MAPK signalling pathway
PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases	2.26E-02	MAPK signalling pathway
Signaling by FGFR4 in disease	2.26E-02	FGFR signalling pathway
Spry regulation of FGF signaling	2.26E-02	FGFR signalling pathway
DNA Damage/Telomere Stress Induced Senescence	2.38E-02	DNA damage
GRB2 events in ERBB2 signaling	2.38E-02	MAPK signalling pathway
IGF1R signaling cascade	2.38E-02	Insulin receptor signalling pathway
Insulin receptor signalling cascade	2.38E-02	Insulin receptor signalling pathway
SHC1 events in ERBB4 signaling	2.38E-02	MAPK signalling pathway
Signaling by BRAF and RAF fusions	2.38E-02	MAPK signalling pathway
Signaling by Interleukins	2.38E-02	Cytokine signalling in immune system/Immune system
Signaling by PDGF	2.38E-02	Receptor tyrosine kinase signalling pathways
Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)	2.38E-02	Insulin receptor signalling pathway
TP53 Regulates Transcription of Genes Involved in G2 Cell Cycle Arrest	2.38E-02	Cell cycle
Transcriptional regulation by RUNX2	2.46E-02	Transcriptional regulation
G1/S Transition	2.71E-02	Cell cycle
SHC-mediated cascade:FGFR3	3.20E-02	FGFR signalling pathway
Signalling to RAS	3.20E-02	MAPK signalling pathway
Transcription of E2F targets under negative control by DREAM complex	3.20E-02	Cell cycle
S Phase	3.33E-02	Cell cycle
Constitutive Signaling by EGFRvIII	3.45E-02	Sigalling pathway
SHC-mediated cascade:FGFR4	3.45E-02	FGFR signalling pathway
Signaling by EGFRvIII in Cancer	3.45E-02	EGFR signalling pathway
CD209 (DC-SIGN) signaling	3.63E-02	Immune system
Mitotic G1-G1/S phases	3.63E-02	Cell cycle
RAF/MAP kinase cascade	3.63E-02	MAPK signalling pathway
SHC-mediated cascade:FGFR1	3.63E-02	FGFR signalling pathway
Signaling by MET	3.63E-02	Receptor tyrosine kinase signalling pathways
TNFR1-induced NF-kB signaling pathway	3.63E-02	Death receptor signalling
TP53 Regulates Transcription of Cell Death Genes	3.63E-02	Apoptosis
VEGFR2 mediated cell proliferation	3.63E-02	Receptor tyrosine kinase signalling pathways
Degradation of beta-catenin by the destruction complex	3.64E-02	Signalling by WNT
NGF signalling via TRKA from the plasma membrane	3.64E-02	Receptor tyrosine kinase signalling pathways
Activation, translocation and oligomerization of BAX	3.77E-02	Apoptosis
Downstream signaling events of B Cell Receptor (BCR)	3.86E-02	Signalling in B cells
Signaling by Insulin receptor	3.98E-02	Cytokine signalling in immune system/Immune system
SHC-mediated cascade:FGFR2	4.10E-02	FGFR signalling pathway
Class B/2 (Secretin family receptors)	4.47E-02	GPCR signalling pathway
Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	4.47E-02	MAPK signalling pathway
DAP12 signaling	4.47E-02	MAPK signalling pathway
G0 and Early G1	4.47E-02	Cell cycle
Generic Transcription Pathway	4.47E-02	Transcriptional regulation
Major pathway of rRNA processing in the nucleolus and cytosol	4.47E-02	Insulin receptor signalling pathway
MAPK6/MAPK4 signaling	4.47E-02	MAPK signalling pathway
Negative regulation of FGFR1 signaling	4.47E-02	FGFR signalling pathway
Negative regulation of FGFR3 signaling	4.47E-02	FGFR signalling pathway
Negative regulation of FGFR4 signaling	4.47E-02	FGFR signalling pathway
Negative regulators of DDX58/IFIH1 signaling	4.47E-02	Immune system
Regulation of TNFR1 signaling	4.47E-02	Death receptor signalling
SHC1 events in ERBB2 signaling	4.47E-02	MAPK signalling pathway
Signaling by EGFR in Cancer	4.47E-02	EGFR signalling pathway
Signaling by Ligand-Responsive EGFR Variants in Cancer	4.47E-02	EGFR signalling pathway
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	4.47E-02	Transcriptional regulation
TRAF3 deficiency - HSE	4.47E-02	Disease associated with TLR signalling cascade
TRAF6 mediated IRF7 activation	4.47E-02	Immune system
MyD88-independent TLR4 cascade	4.66E-02	Toll-like receptors cascade
TRIF(TICAM1)-mediated TLR4 signaling	4.66E-02	Toll-like receptors cascade
Association of TriC/CCT with target proteins during biosynthesis	4.69E-02	Protein folding
Signaling by FGFR3 in disease	4.69E-02	FGFR signalling pathway
Signaling by FGFR3 point mutants in cancer	4.69E-02	FGFR signalling pathway
Negative regulation of FGFR2 signaling	4.92E-02	FGFR signalling pathway

Appendix 2: Contribution of each mutational signature proposed by the Wellcome Trust Sanger Institute per sample (Chapter 3). The file is included in the CD-R attached with this thesis.

Appendix 3: Coverage, purity, karyotype, and clinical information for all samples in Myeloma XI study (Chapter 5). CTD: cyclophosphamide, thalidomie, and dexamethasone; RCD: Lenalidomide (Revlimid), cyclophosphamide, and dexamethasone; CCRD: carfilzomib, cyclophosphamide, lenalidomide, and dexamethasone; Intensive pathway: treatment with high dose melphalan after induction. HD: hyperdiploid. NA: not available

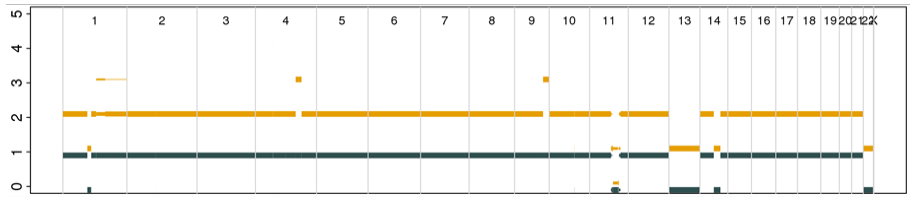
Sample ID	Normal	Primary	Relapse	Primary	Relapse	Karyotype	Gender	Age	Elapsed time (months)	Induction	Maintenance	Pathway
	Coverage			Purity								
1305	38.05	125.66	106.44	0.94	0.52	11;14	Male	51	38.34	CTD	No maintenance	Intensive
1334	41.38	117.96	118.87	0.90	0.99	11;14	Female	43	24.00	CTD	Missing	Intensive
5834	38.31	117.53	116.79	0.98	0.31	11;14	Female	69	29.93	CTDa	No maintenance	Non-intensive
6030	39.08	106.03	119.35	0.94	0.73	4;14	Female	36	19.75	CTD	No maintenance	Intensive
6178	39.92	110.32	103.75	0.98	0.82	11;14	Female	67	18.40	RCD	Missing	Intensive
6229	43.91	120.53	107.41	0.72	0.59	11;14	Male	74	9.23	CTDa	Missing	Non-intensive
6706	42.72	120.92	114.44	0.92	0.71	11;14	Male	59	25.43	RCD	No maintenance	Intensive
6988	37.54	121.95	107.64	0.80	0.83	11;14	Male	69	12.26	RCDa	No maintenance	Non-intensive
7020	38.96	111.10	122.49	0.93	0.92	4;14	Female	58	14.69	CTD	Missing	Intensive
7240	37.57	131.00	102.37	0.63	0.86	4;14	Male	55	11.30	RCD	Lenalidomide maintenance	Intensive
7801	39.48	116.50	106.97	0.54	1.00	14;16	Female	48	14.49	CTD	Missing	Intensive
7842	38.23	112.19	113.46	0.92	0.86	4;14	Male	66	17.64	CTD	No maintenance	Intensive
8237	36.54	102.98	110.72	0.91	0.86	4;14	Female	49	14.00	CTD	No maintenance	Intensive
9126	42.02	110.75	120.59	0.87	0.95	11;14	Male	64	16.23	CTDa	Missing	Non-intensive
9166	42.11	115.52	113.68	0.97	0.51	14;16	Female	68	27.24	CCRD	No maintenance	Intensive
9515	40.86	120.94	110.38	0.89	0.61	11;14	Male	68	26.15	RCDa	Lenalidomide maintenance	Non-intensive
9524	40.40	155.44	156.08	0.95	0.09	4;14	Male	51	33.81	RCDa	Lenalidomide maintenance	Non-intensive
9721	37.93	117.26	115.58	0.73	0.75	14;16	Male	64	29.44	CTD	Lenalidomide maintenance	Intensive
10068	37.23	112.08	108.68	0.92	0.59	4;14	Male	71	13.77	RCDa	Lenalidomide and vorinostat maintenance	Non-intensive
10365	40.43	114.99	114.13	0.93	0.85	11;14	Male	76	9.33	CTD	Missing	Intensive
11506	43.51	115.19	119.67	0.96	0.49	14;16	Male	77	11.83	CTDa	Lenalidomide maintenance	Non-intensive
11668	38.04	118.52	118.81	0.95	0.88	4;14	Male	49	19.29	RCDa	Missing	Non-intensive
11949	41.50	116.51	114.17	0.96	0.98	11;14	Male	76	14.65	CTD	Missing	Intensive
12546	39.82	153.13	153.71	0.99	0.88	4;14	Male	77	30.59	RCD	Missing	Intensive
13029	37.36	112.96	104.78	0.89	0.93	4;14	Male	62	6.90	CTD	Missing	Intensive
5695	35.75	101.44	NA	0.98	NA	11;14	Male	64	15.61	CTD	No maintenance	Intensive
5699	36.84	108.95	NA	0.94	NA	11;14	Female	68	6.24	CTD	Missing	Intensive
5836	35.21	112.77	NA	0.89	NA	11;14	Male	77	36.07	CTDa	No maintenance	Non-intensive
5939	40.32	109.19	NA	0.96	NA	4;14	Male	65	34.30	CTD	Missing	Intensive
6016	36.74	108.30	NA	0.92	NA	11;14	Female	55	71.39	RCD	Missing	Intensive
6076	38.51	103.46	NA	0.91	NA	4;14	Male	72	11.53	RCDa	Lenalidomide maintenance	Non-intensive
6163	34.25	103.88	NA	0.97	NA	4;14	Male	75	6.90	RCDa	Missing	Non-intensive
6277	38.07	121.51	NA	0.47	NA	11;14	Male	56	77.86	RCD	Lenalidomide maintenance	Intensive
6279	33.10	110.18	NA	0.89	NA	4;14	Male	62	21.91	RCD	Lenalidomide maintenance	Intensive
6345	32.48	94.20	NA	0.94	NA	4;14	Female	72	10.58	CTDa	Missing	Non-intensive
6415	36.76	105.25	NA	0.96	NA	11;14	Female	68	5.42	RCDa	Missing	Non-intensive
6425	37.31	104.87	NA	1.00	NA	4;14	Male	67	23.95	RCD	Lenalidomide and vorinostat maintenance	Intensive
6501	37.56	110.79	NA	0.92	NA	11;14	Female	51	13.11	RCD	Missing	Intensive
6702	30.68	107.23	NA	0.90	NA	4;14	Female	78	2.30	CTDa	Missing	Non-intensive
7000	37.47	109.85	NA	0.91	NA	11;14	Female	78	1.87	CTDa	Missing	Non-intensive

Sample ID	Normal	Primary	Relapse	Primary	Relapse	Karyotype	Gender	Age	Elapsed time (months)	Induction	Maintenance	Pathway
	Coverage			Purity								
7005	36.88	105.23	NA	0.98	NA	4;14	Male	74	8.38	CTDa	Missing	Non-intensive
7164	37.27	110.13	NA	0.99	NA	11;14	Female	80	0.72	RCDa	Missing	Non-intensive
7348	36.11	110.24	NA	0.95	NA	4;14	Male	67	10.65	RCDa	No maintenance	Non-intensive
7729	32.76	119.42	NA	0.89	NA	4;14	Male	65	37.72	RCD	Lenalidomide and vorinostat maintenance	Intensive
7794	32.11	117.88	NA	0.94	NA	4;14	Female	52	15.15	CTD	No maintenance	Intensive
7880	37.80	105.38	NA	1.00	NA	4;14	Female	82	6.11	RCDa	Missing	Non-intensive
7915	38.20	97.07	NA	0.93	NA	4;14	Male	59	40.71	CTD	Lenalidomide and vorinostat maintenance	Intensive
7925	36.21	81.62	NA	0.83	NA	4;14	Male	59	6.41	CTD	Missing	Intensive
7950	38.58	116.13	NA	0.87	NA	4;14	Male	49	33.05	CTD	Lenalidomide and vorinostat maintenance	Intensive
7956	38.67	116.29	NA	0.94	NA	4;14	Female	56	6.34	CTD	Missing	Intensive
8043	37.02	97.85	NA	0.93	NA	4;14	Female	81	8.51	CTDa	Missing	Non-intensive
8245	38.07	102.24	NA	0.86	NA	11;14	Female	63	55.85	RCD	Lenalidomide maintenance	Intensive
8567	37.32	123.34	NA	0.47	NA	11;14	Female	66	19.38	RCDa	Lenalidomide and vorinostat maintenance	Non-intensive
8573	38.19	111.54	NA	0.93	NA	4;14/HD	Female	82	10.22	CTDa	Missing	Non-intensive
8928	32.98	96.98	NA	0.97	NA	4;14	Male	52	7.36	CTD	Missing	Intensive
8979	36.54	117.56	NA	0.38	NA	4;14	Male	76	26.22	CTDa	Missing	Non-intensive
9069	37.30	97.76	NA	0.95	NA	11;14	Male	73	0.99	RCDa	Missing	Non-intensive
9176	38.27	103.83	NA	0.93	NA	11;14	Male	78	3.42	RCDa	Missing	Non-intensive
9210	37.48	109.25	NA	0.91	NA	11;14	Male	69	10.55	CTD	Missing	Intensive
9249	37.51	105.83	NA	0.94	NA	11;14	Male	58	54.60	RCD	Lenalidomide maintenance	Intensive
9289	36.46	103.20	NA	0.75	NA	11;14	Male	56	24.08	CTD	No maintenance	Intensive
9292	38.92	103.79	NA	0.98	NA	4;14	Female	74	3.71	CTDa	Missing	Non-intensive
9337	37.87	110.25	NA	0.49	NA	11;14	Female	71	26.05	CTDa	Missing	Non-intensive
9376	37.21	111.57	NA	0.78	NA	4;14	Female	64	48.00	RCD	Missing	Intensive
9409	37.54	112.13	NA	0.83	NA	11;14	Male	73	26.91	CTDa	Missing	Non-intensive
9544	38.36	111.21	NA	0.93	NA	11;14	Male	67	54.24	RCDa	No maintenance	Non-intensive
9623	38.53	118.15	NA	0.84	NA	11;14	Male	58	36.57	RCD	Lenalidomide maintenance	Intensive
9718	35.55	85.51	NA	0.95	NA	4;14	Male	66	8.18	RCDa	No maintenance	Non-intensive
9917	37.89	106.90	NA	0.95	NA	11;14	Male	76	0.00	CTDa	Missing	Non-intensive
9931	36.08	100.36	NA	0.86	NA	11;14	Female	55	15.74	RCD	Missing	Intensive
10085	37.48	113.93	NA	0.89	NA	11;14	Female	59	27.27	CCRD	Lenalidomide maintenance	Intensive
10212	37.06	104.96	NA	0.91	NA	11;14	Female	79	48.66	RCDa	Lenalidomide maintenance	Non-intensive
10597	30.59	114.68	NA	0.89	NA	4;14	Male	59	22.51	CCRD	No maintenance	Intensive
10772	39.40	113.42	NA	0.85	NA	4;14	Female	63	17.25	CCRD	Missing	Intensive
10801	37.37	111.23	NA	0.96	NA	11;14	Male	77	23.79	RCDa	Missing	Non-intensive
11029	38.80	111.50	NA	0.92	NA	4;14	Female	73	11.43	RCDa	Missing	Non-intensive
11897	40.44	90.88	NA	0.87	NA	4;14	Male	58	12.49	CCRD	Lenalidomide maintenance	Intensive
12101	34.41	85.24	NA	0.91	NA	4;14	Male	62	8.05	CCRD	Missing	Intensive
12227	36.99	88.91	NA	0.92	NA	11;14	Male	57	30.95	CCRD	No maintenance	Intensive
12541	30.07	99.29	NA	0.95	NA	11;14	Male	56	30.42	CTD	Missing	Intensive

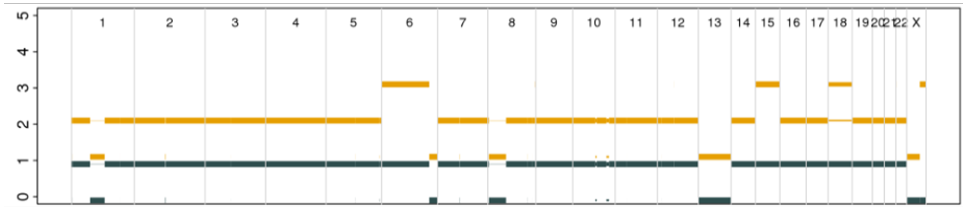
Appendix 4: Copy number plots for 80 primary tumours organised by karyotypes (Chapter 5). Clonal copy numbers are represented as solid line with higher intensity than subclonal copy number changes represented as thin line. Yellow: total copy number, dark blue: copy number of the minor allele. Copy number > 5 is not shown. Y-axis: copy number, x-axis: chromosomes.

t(4;14)

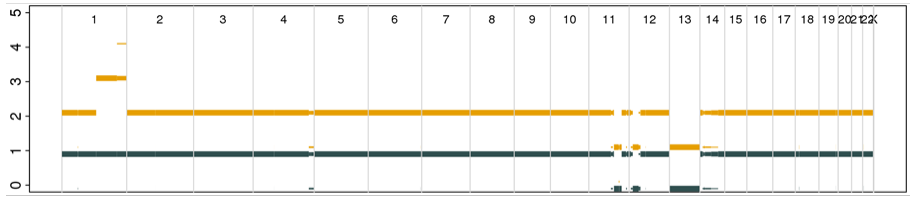
5939



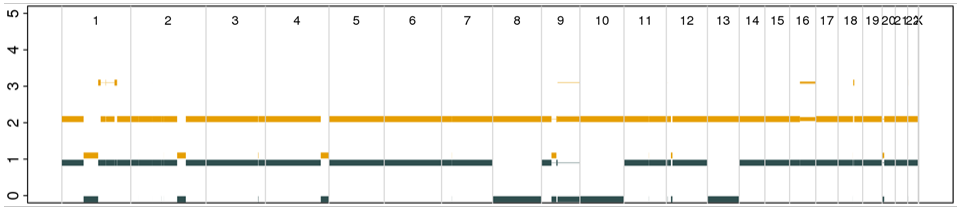
6030



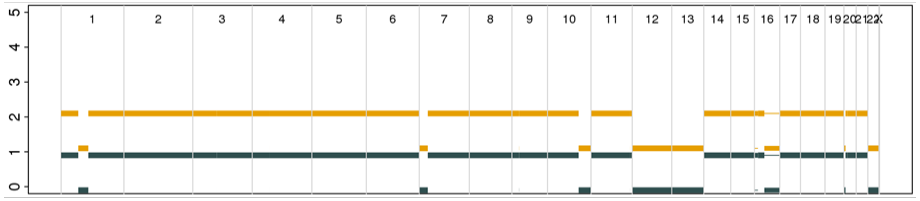
6076



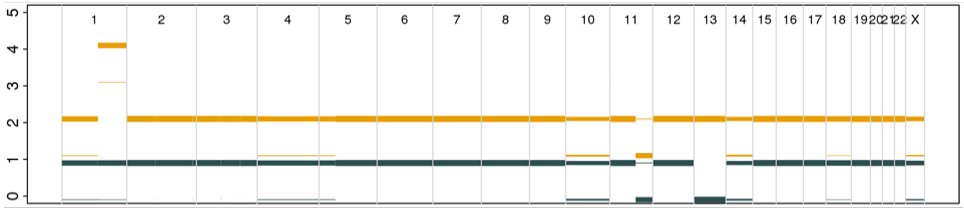
6163



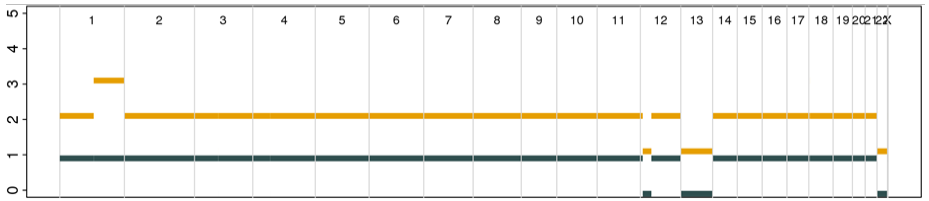
6279



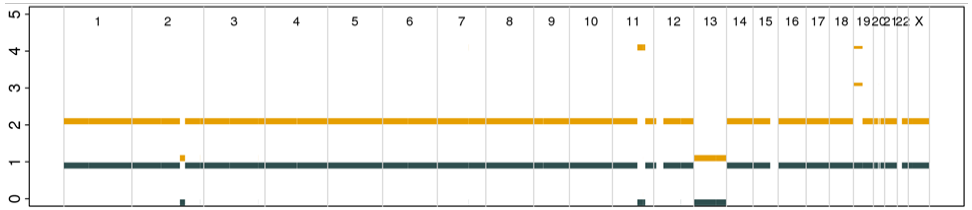
6345



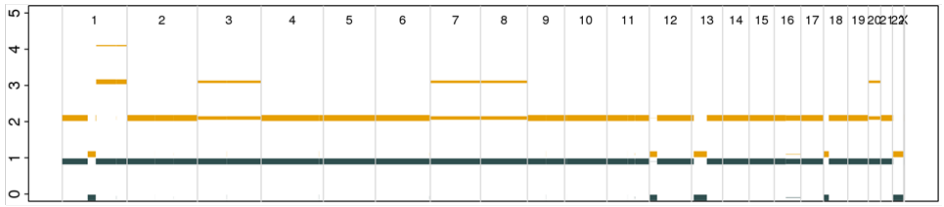
6425



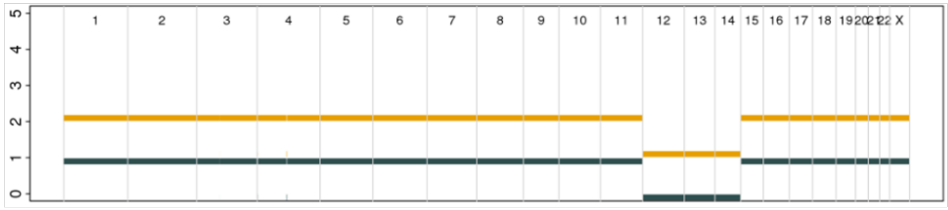
6702



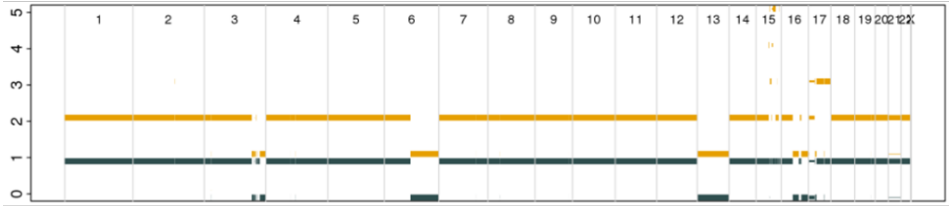
7005



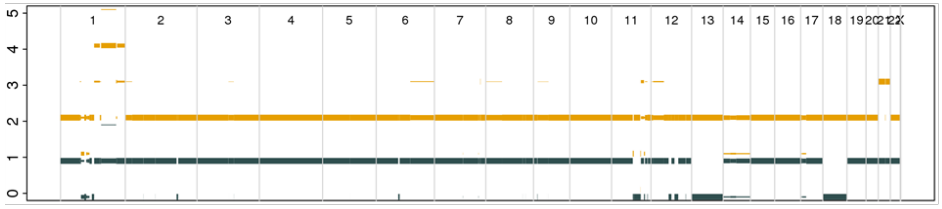
7020



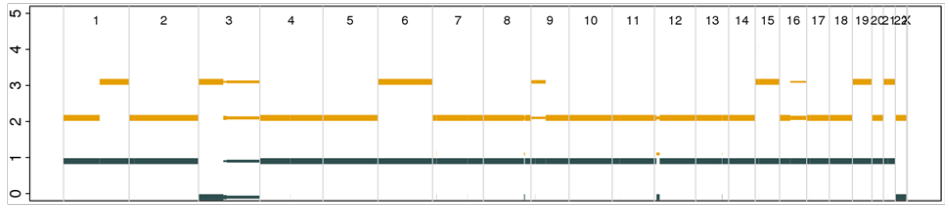
7240



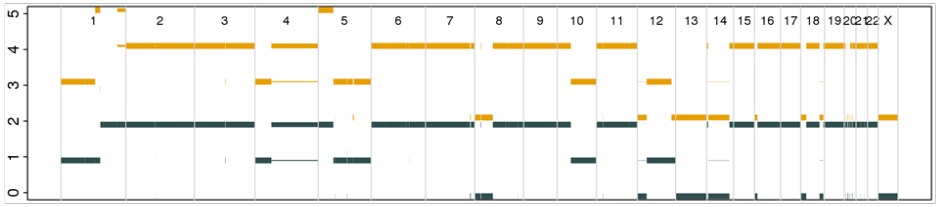
7348



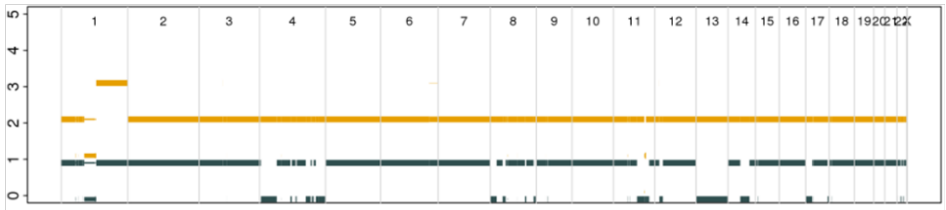
7729



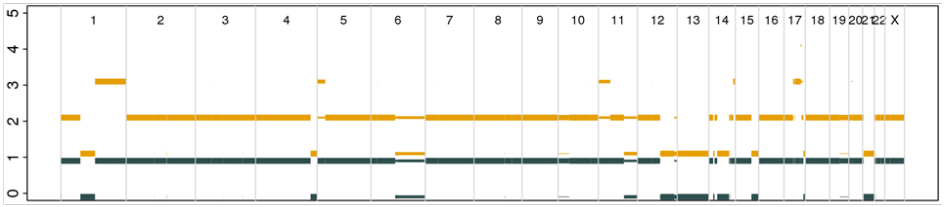
7794



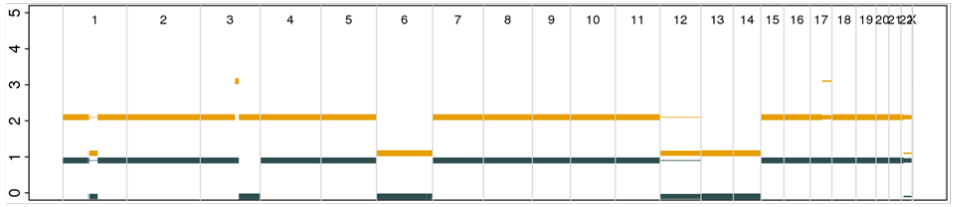
7842



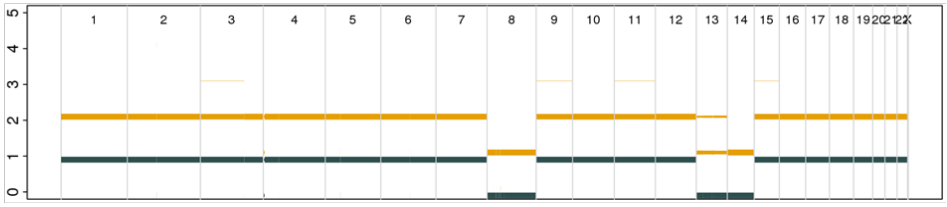
7880



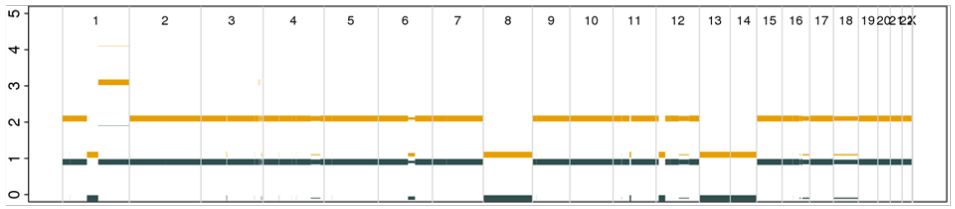
7915



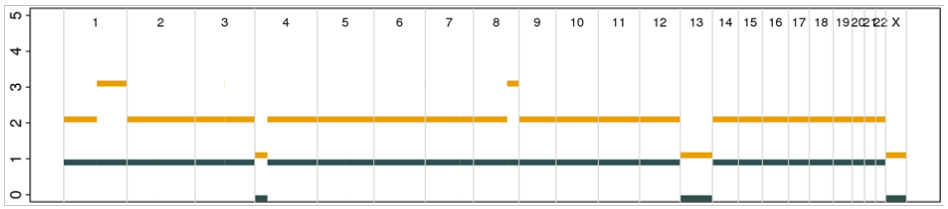
7925



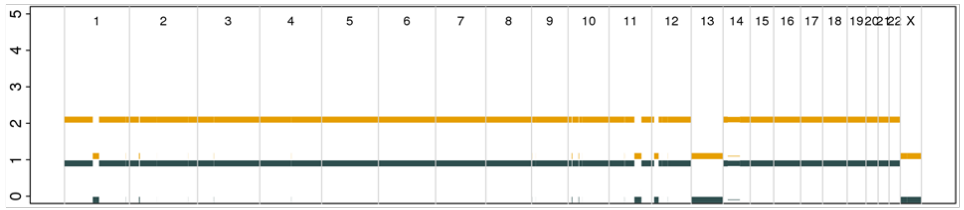
7950



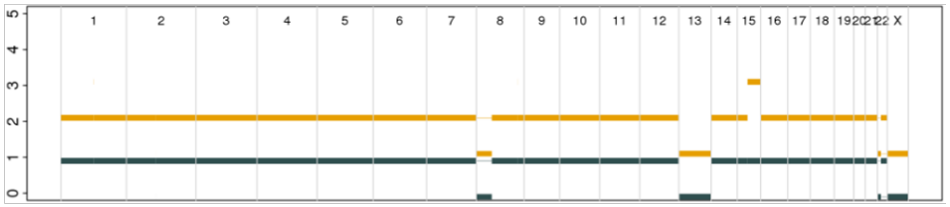
7956



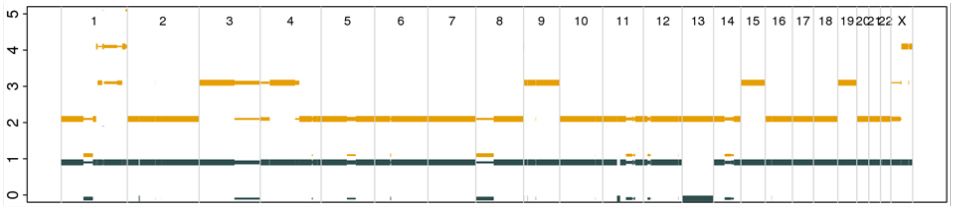
8043



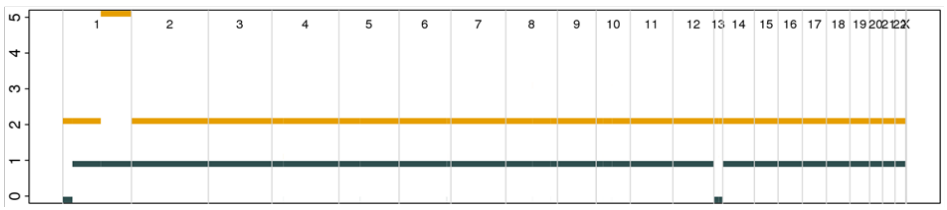
8237



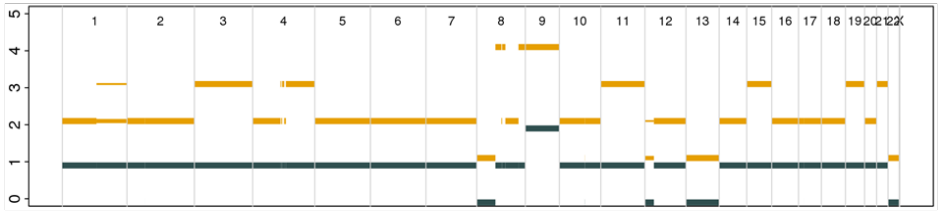
8573



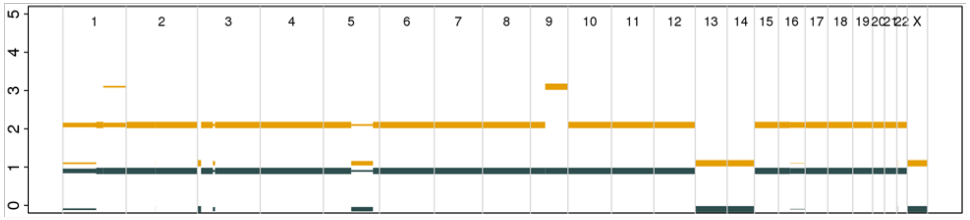
8928



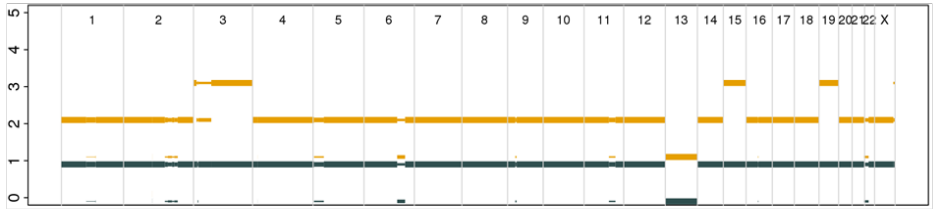
8979



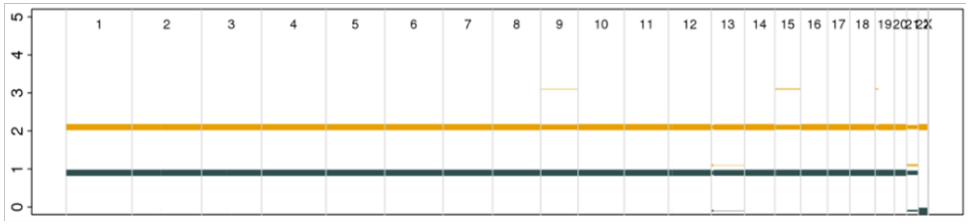
9292



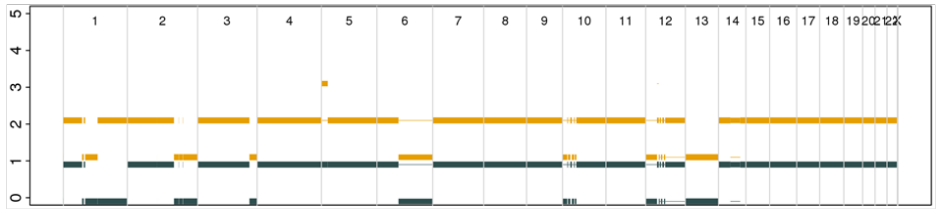
9376



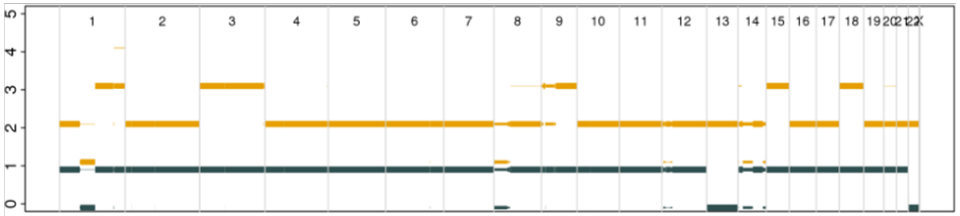
9524



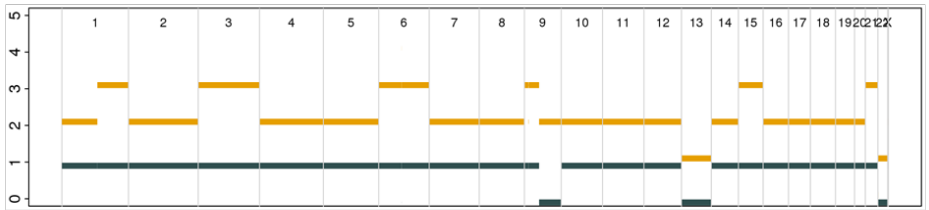
9718



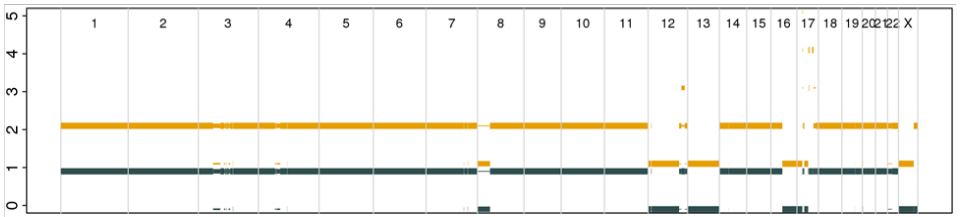
10068



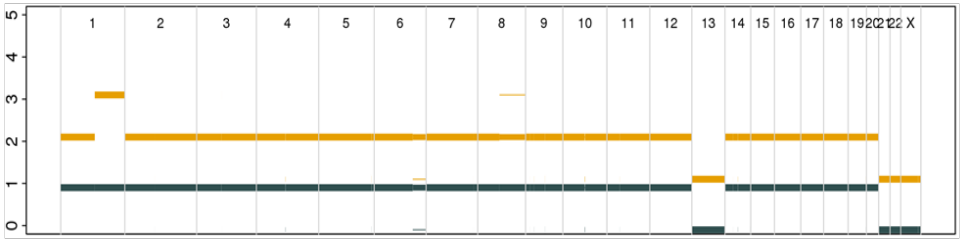
10597



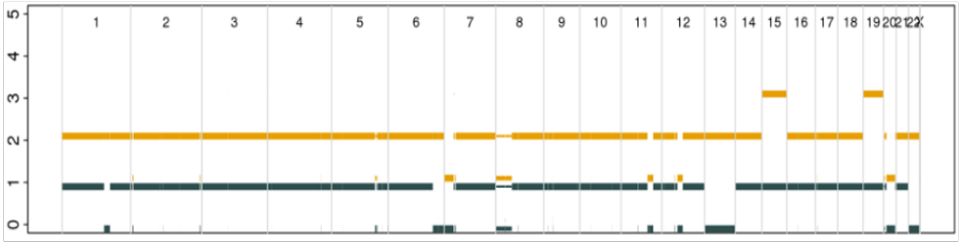
10772



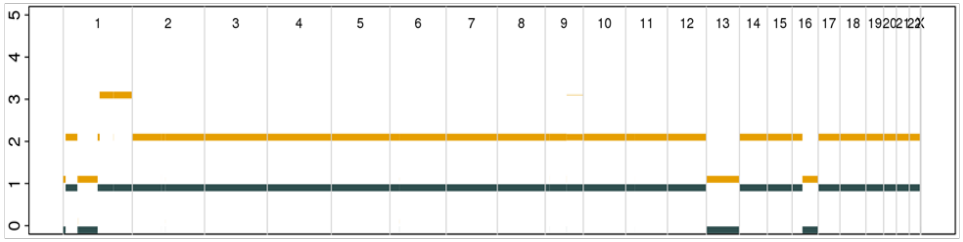
11029



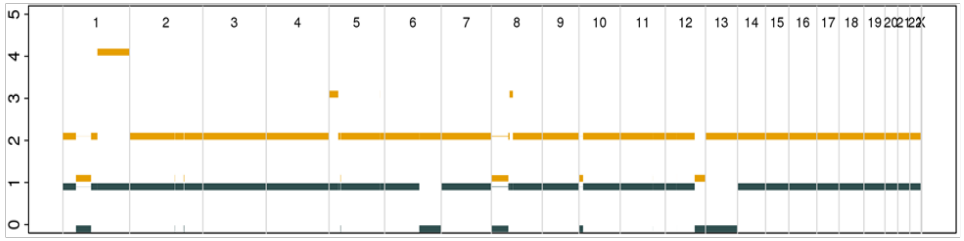
11668



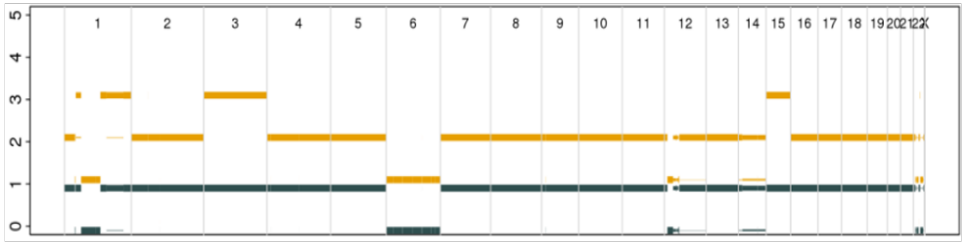
11897



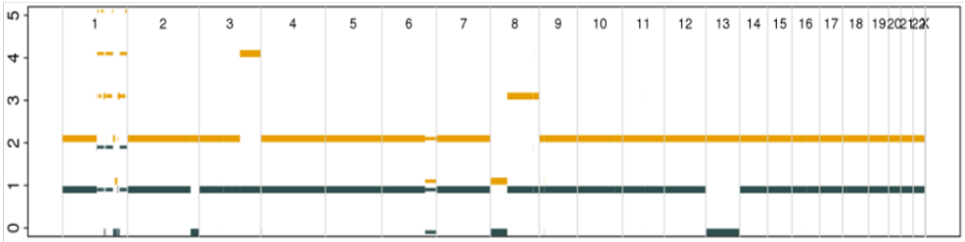
12101



12546

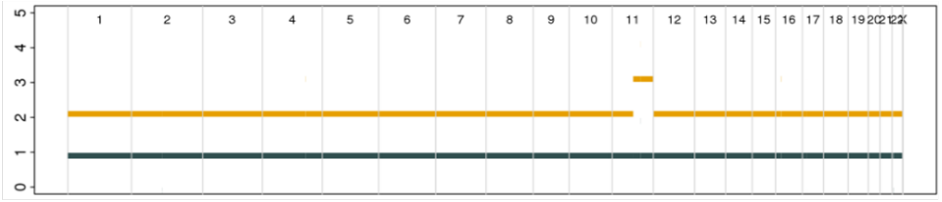


13029

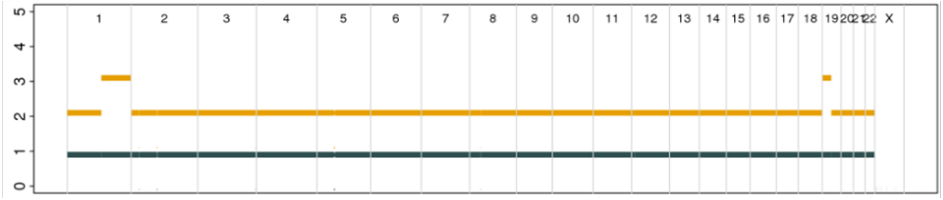


t(11;14)

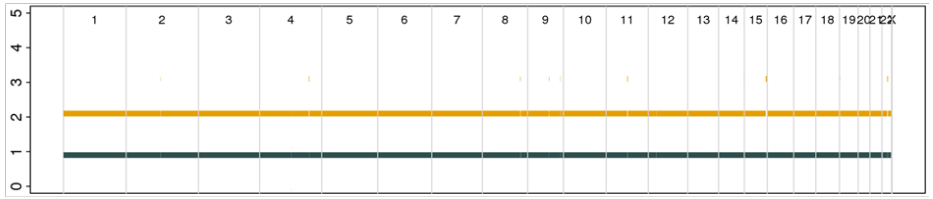
1305



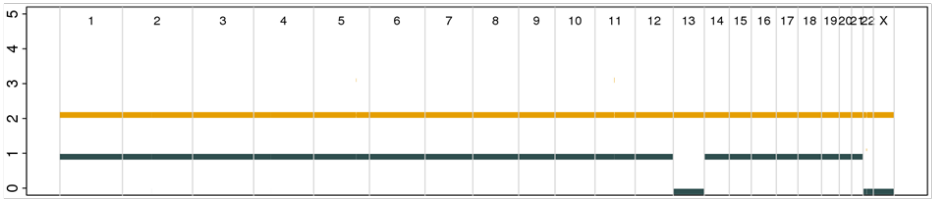
1334



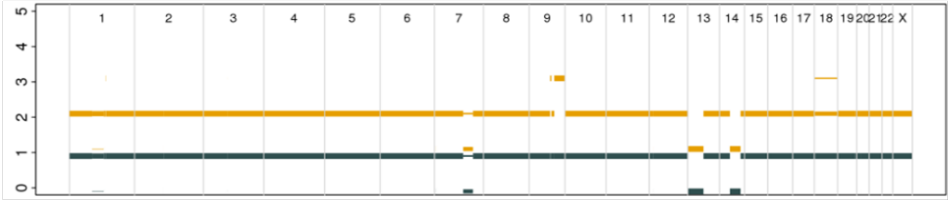
5695



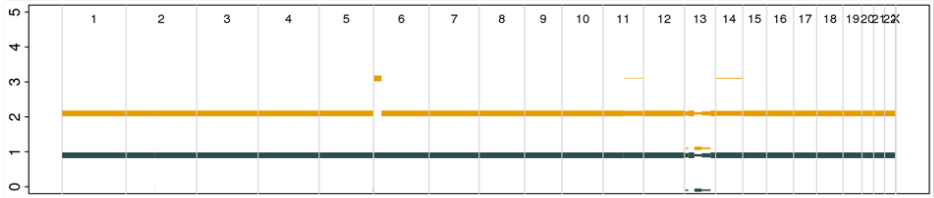
5699



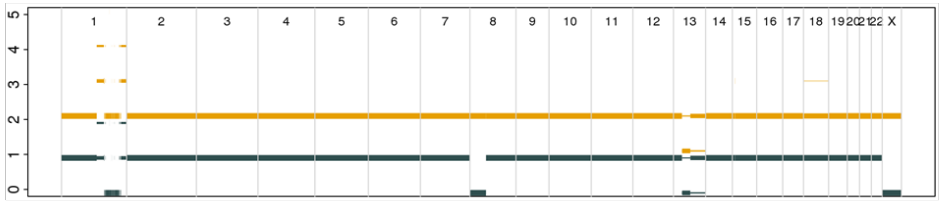
5834



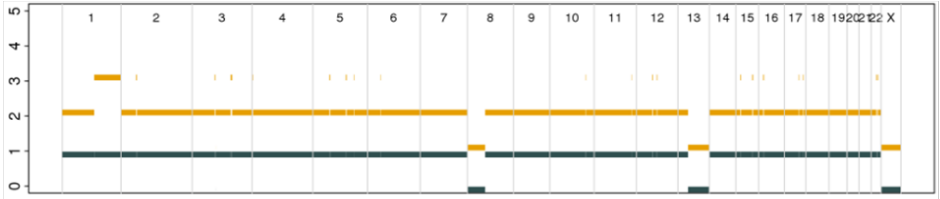
5836



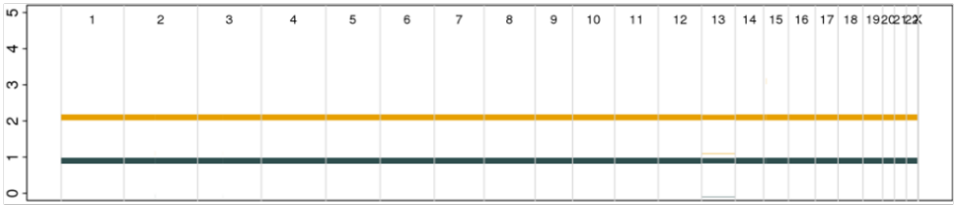
6016



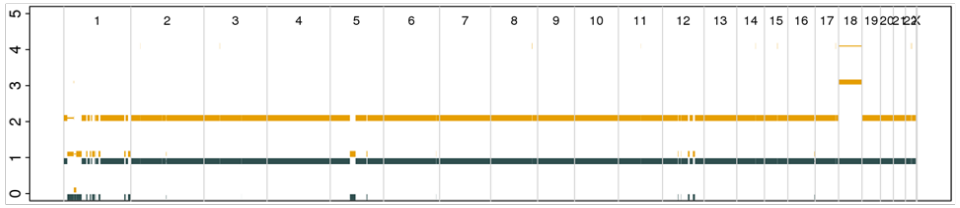
6178



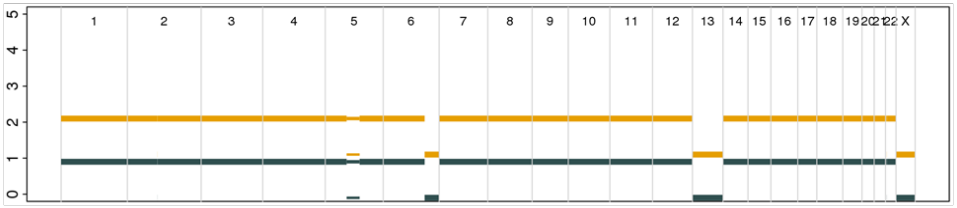
6229



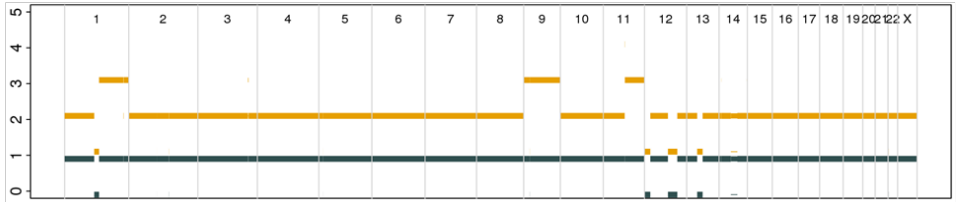
6277



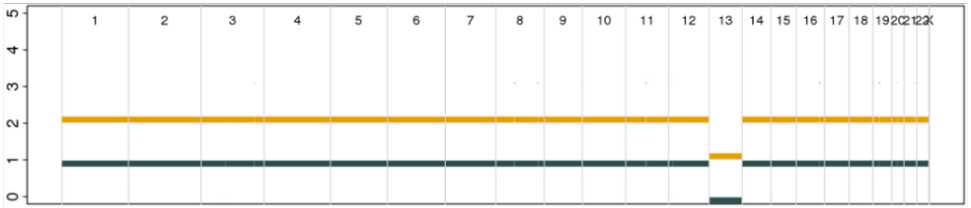
6415



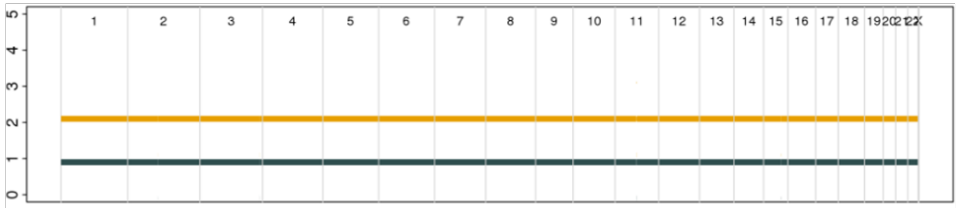
6501



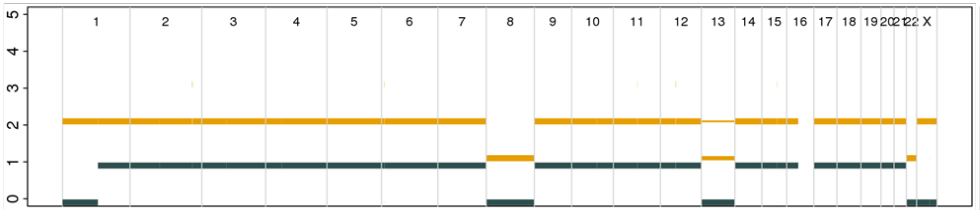
6706



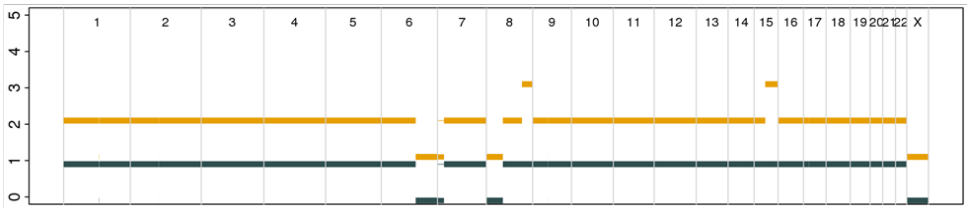
6988



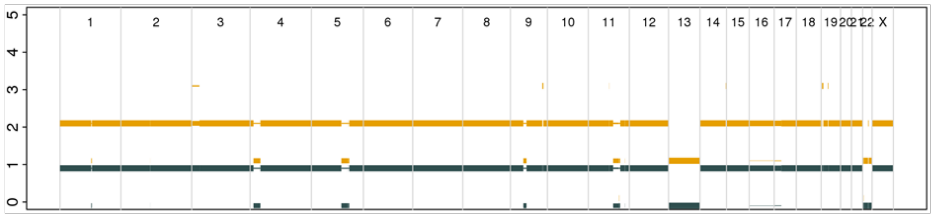
7000



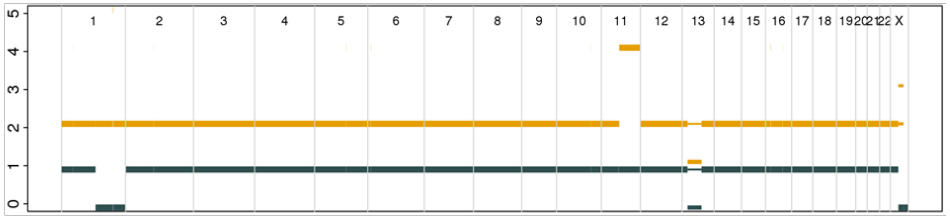
7164



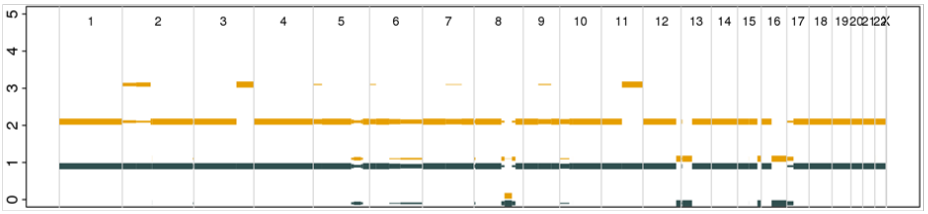
8245



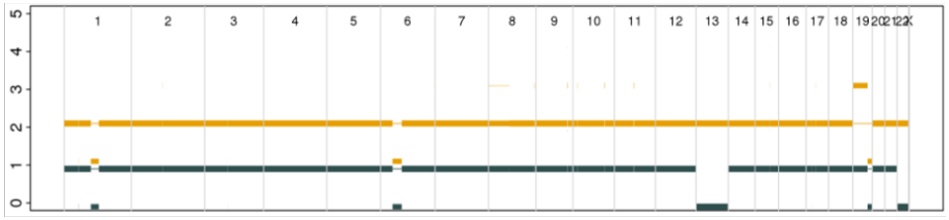
8567



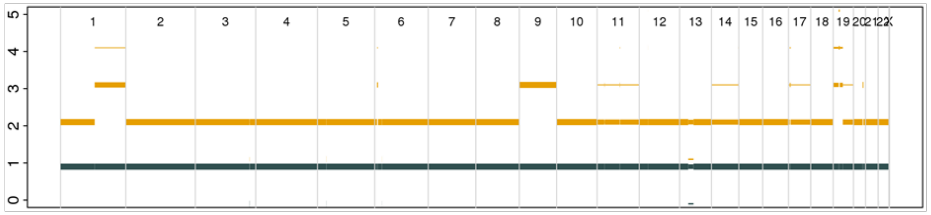
9069



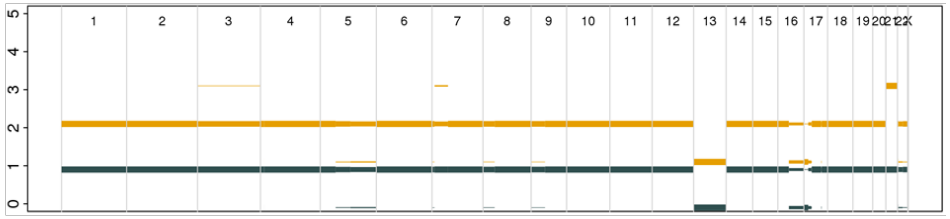
9126



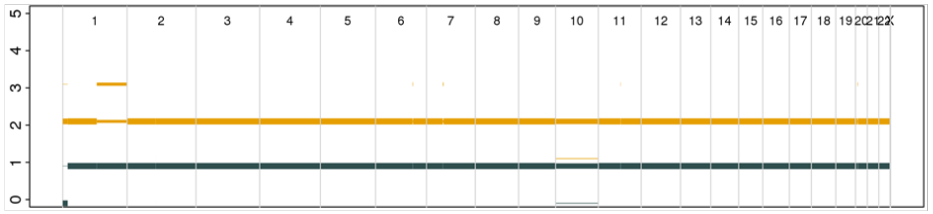
9176



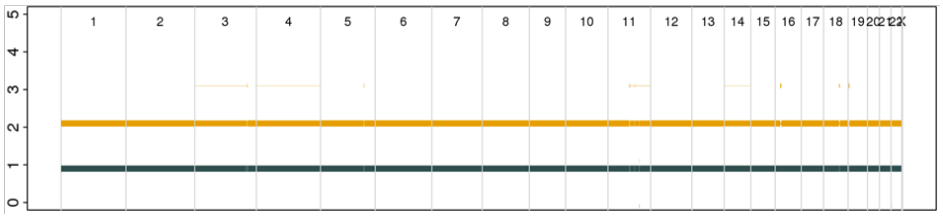
9210



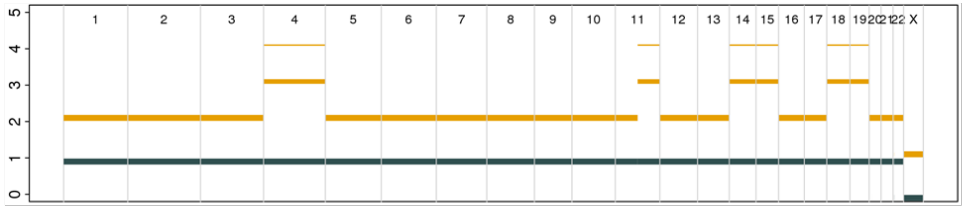
9249



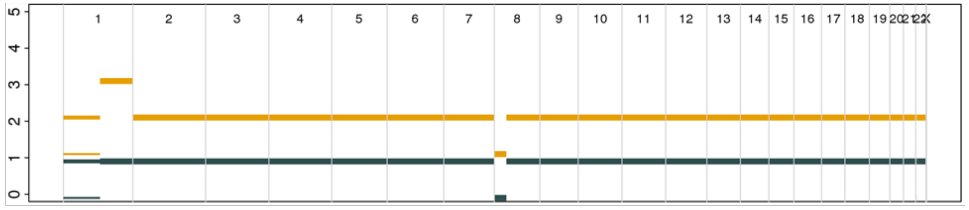
9289



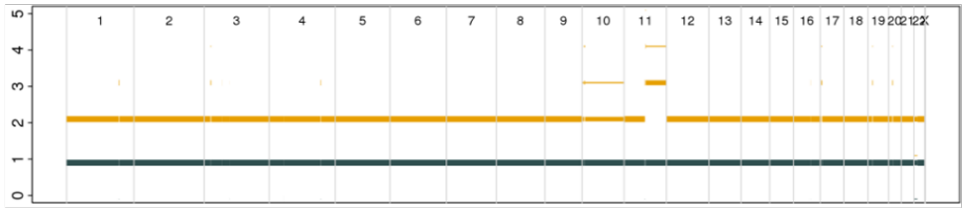
9337



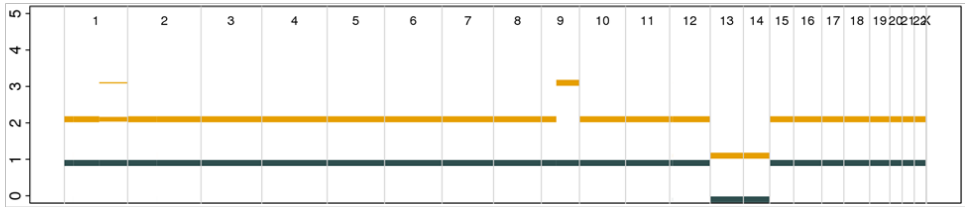
9409



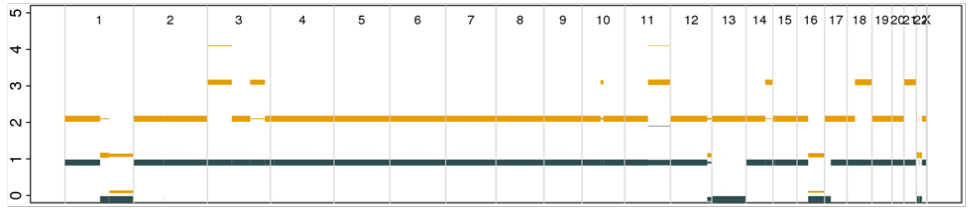
9515



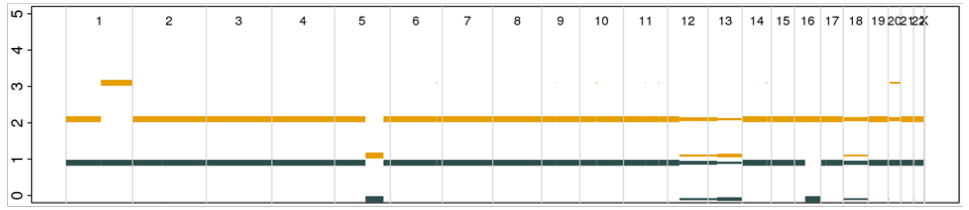
9544



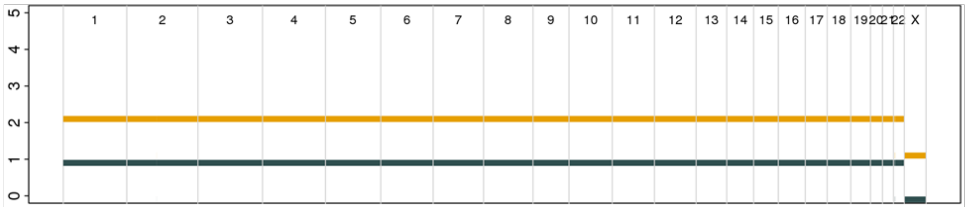
9623



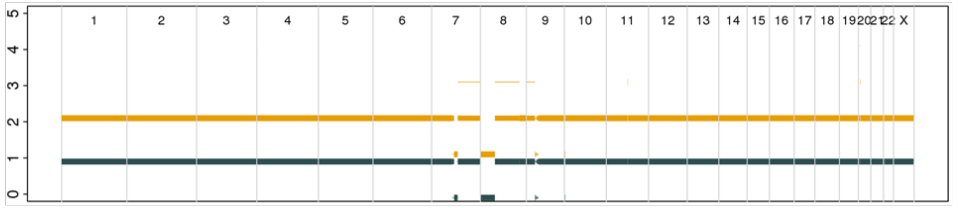
9917



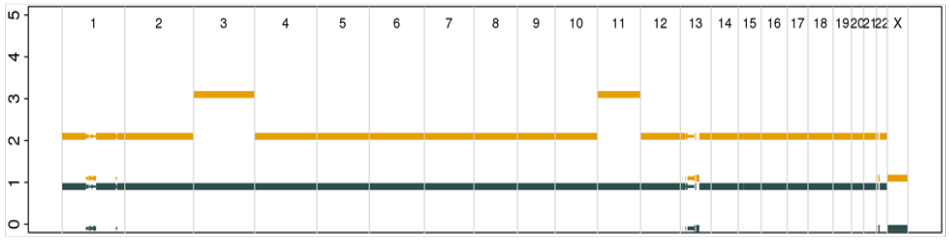
9931



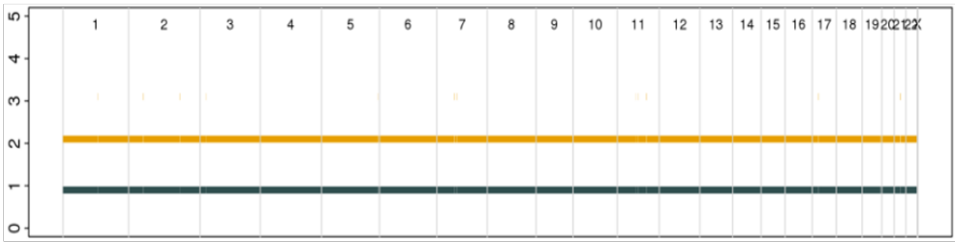
10085



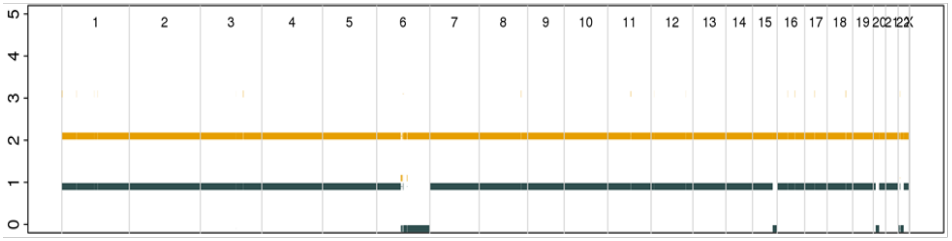
10212



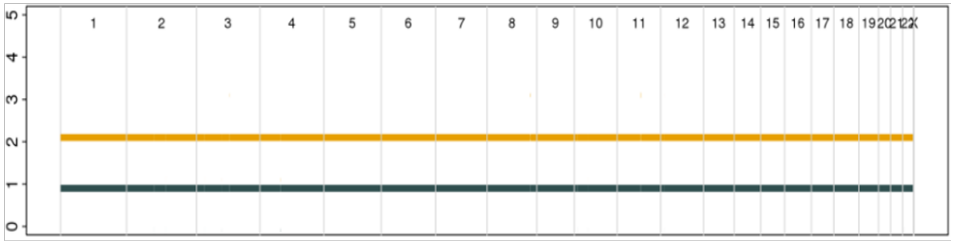
10365



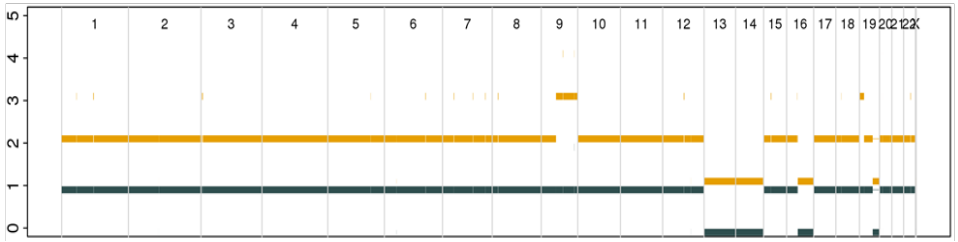
10801



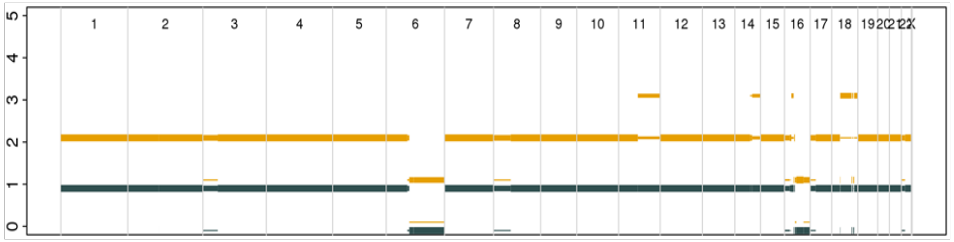
11949



12227

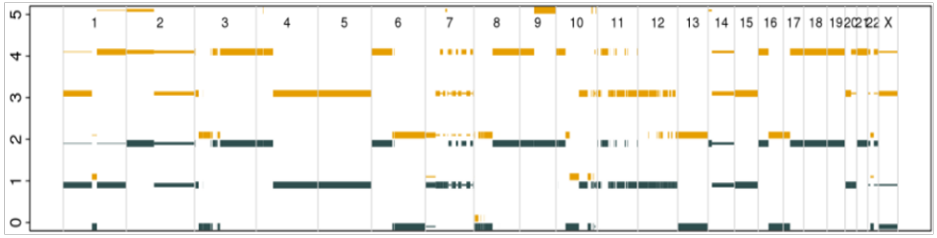


12541

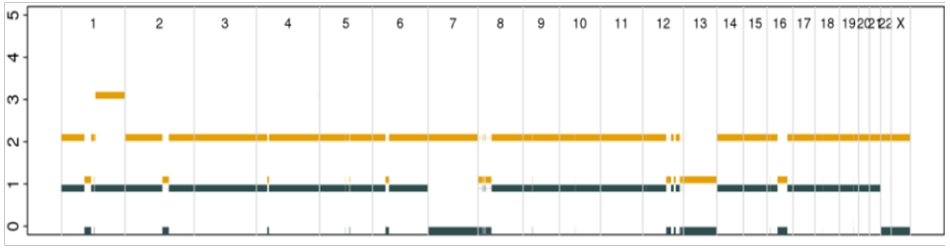


t(14;16)

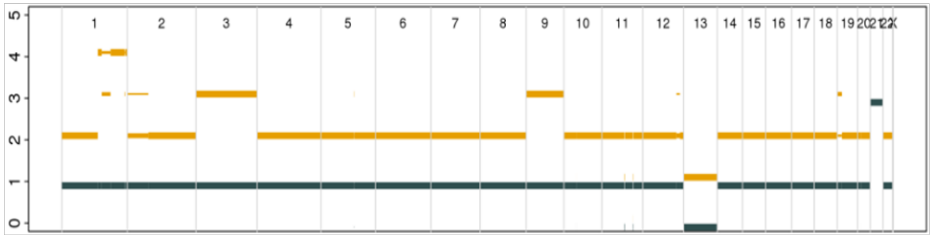
7801



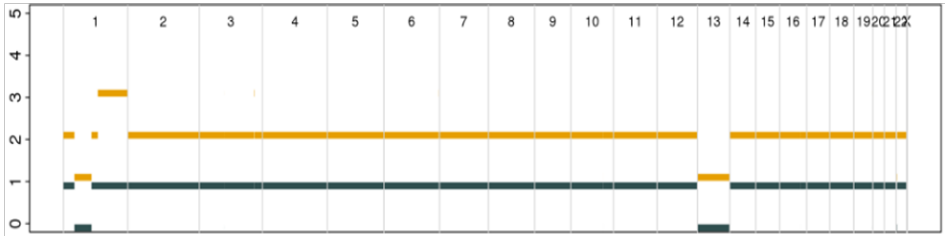
9166



9721



11506



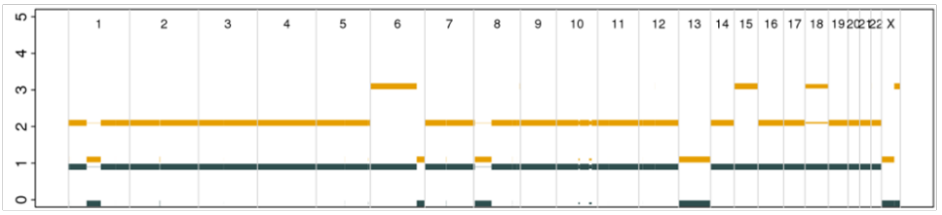
Appendix 5: Copy number plots for 25 matched primary (left) and relapsed (right) tumours organised by karyotypes (Chapter 5). Clonal copy numbers are represented as solid line with higher intensity than subclonal copy number changes represented as thin line. Yellow: total copy number, dark blue: copy number of the minor allele. Copy number > 5 is not shown. Y-axis: copy number, x-axis: chromosomes.

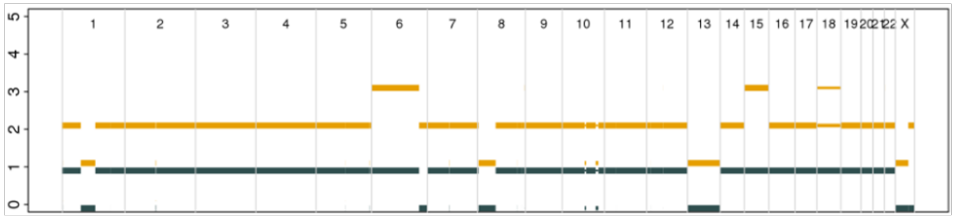
t(4;14)

Primary

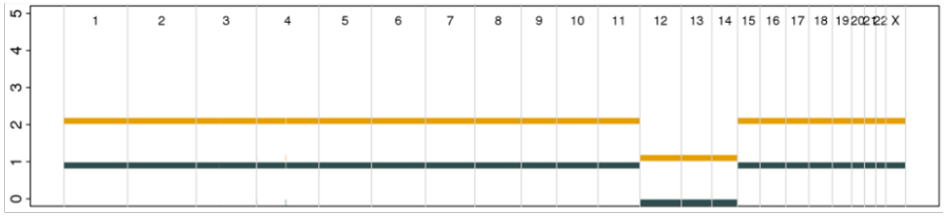
Relapse

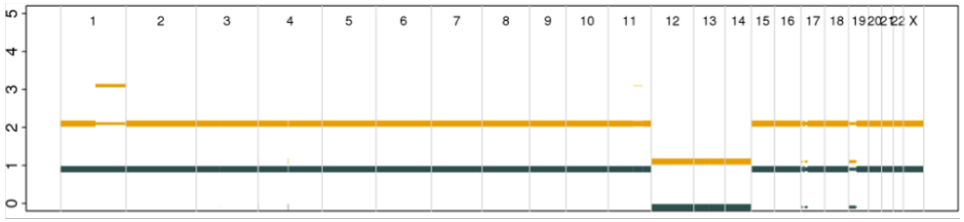
6030



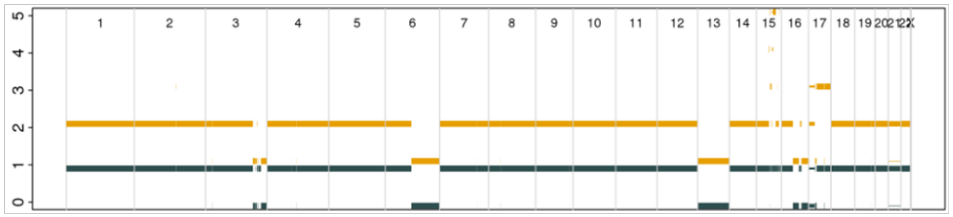


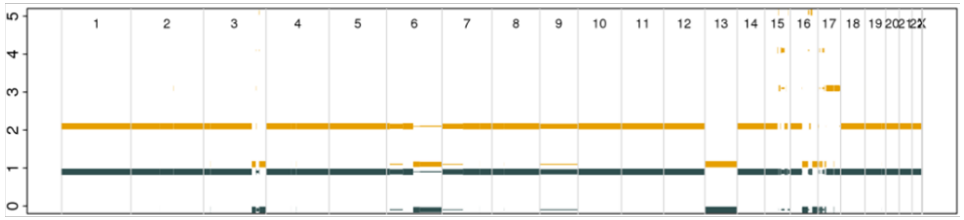
7020



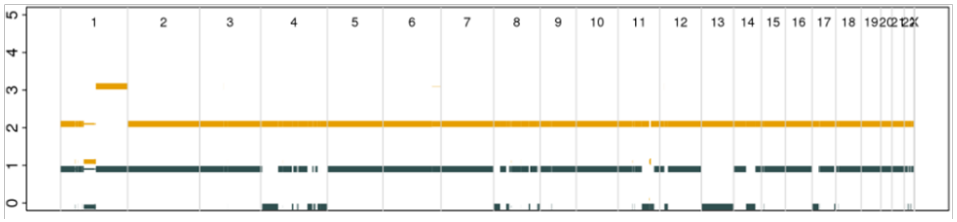


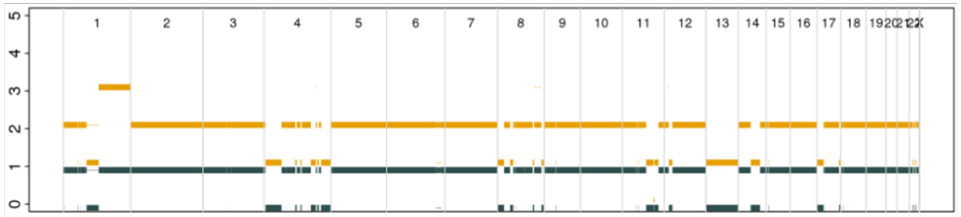
7240



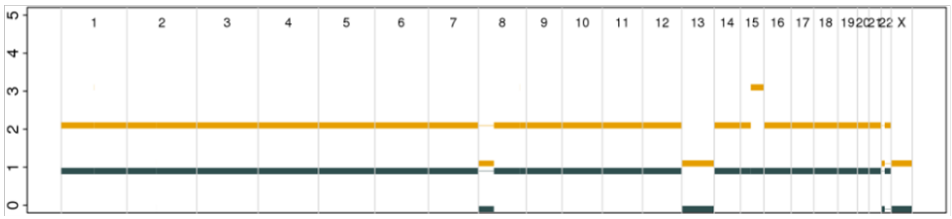


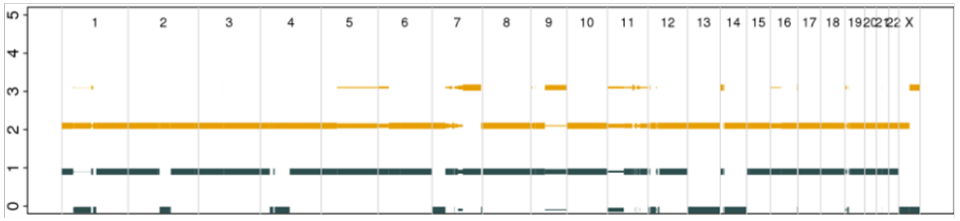
7842



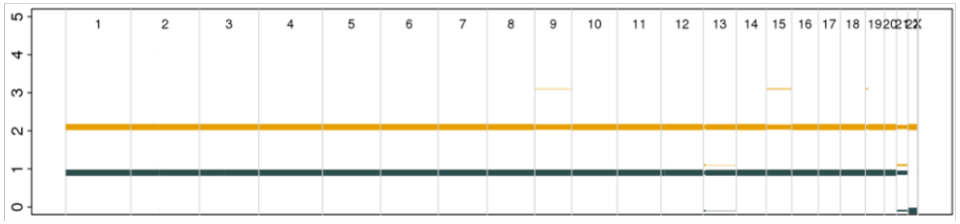


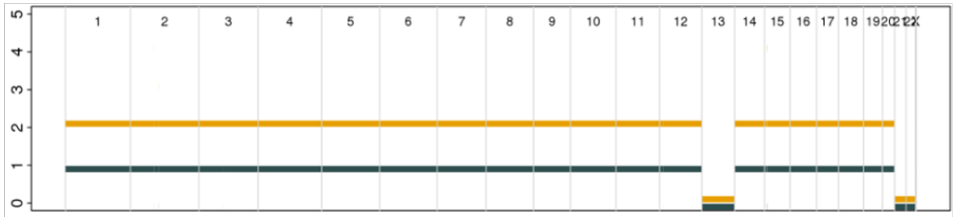
8237



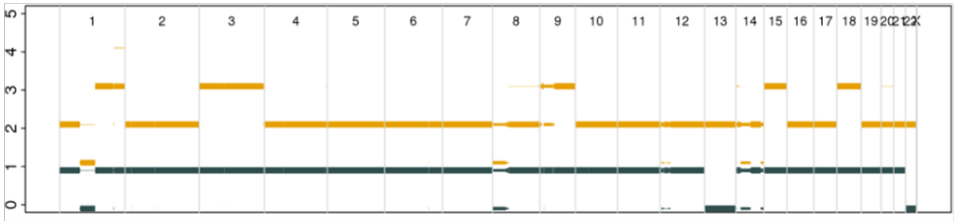


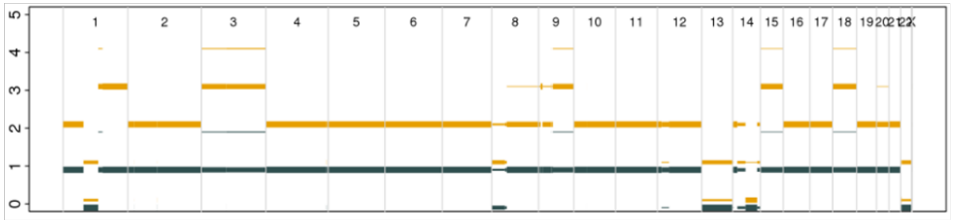
9524



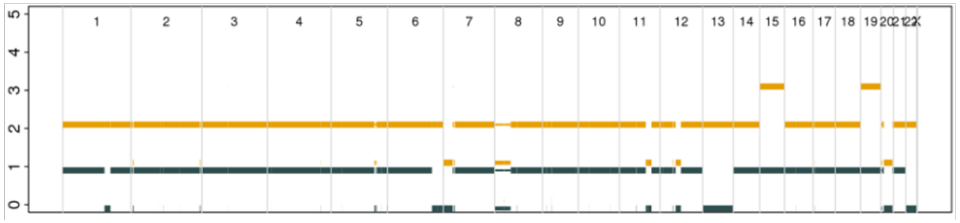


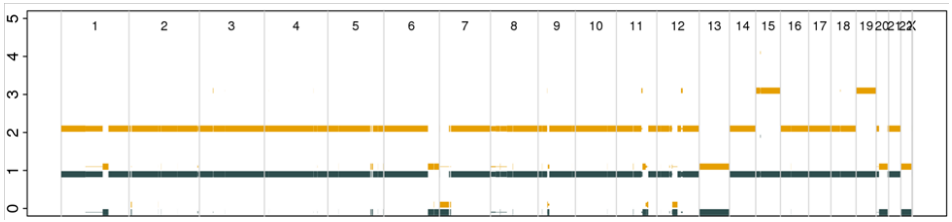
10068



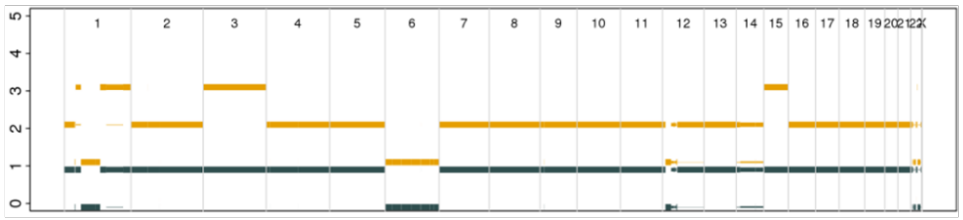


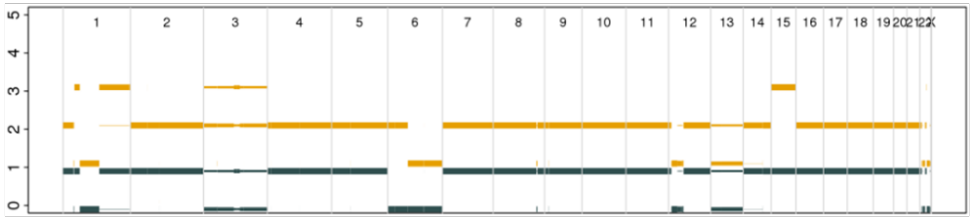
11668



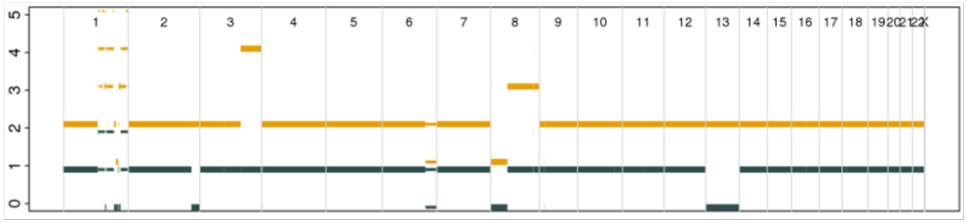


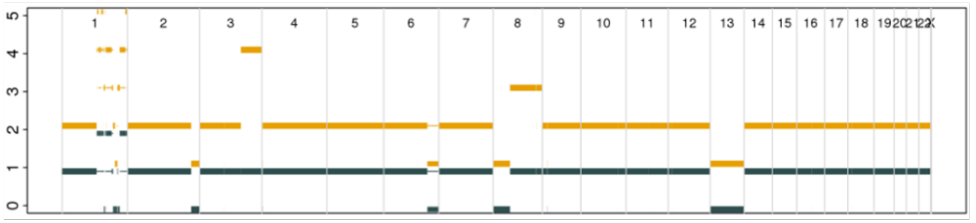
12546





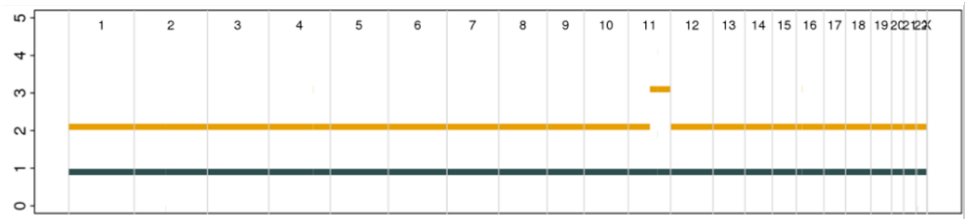
13029

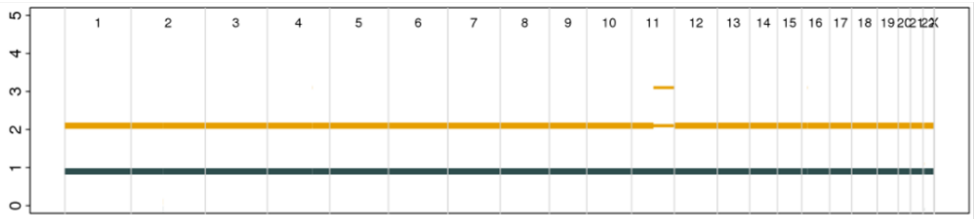




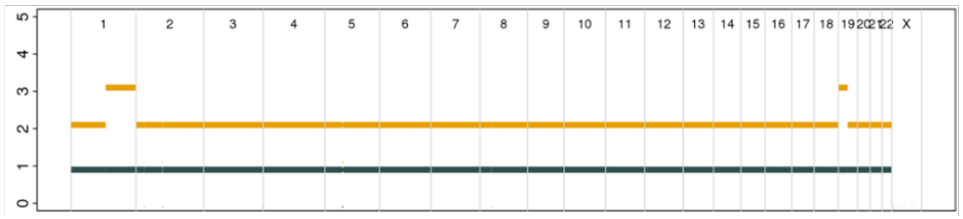
t(11;14)

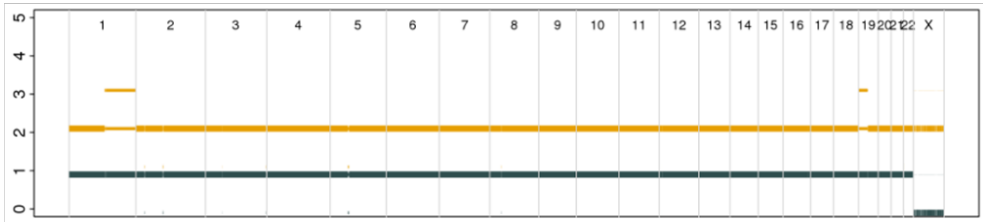
1305



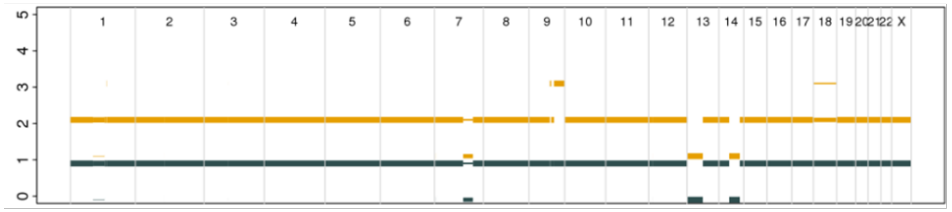


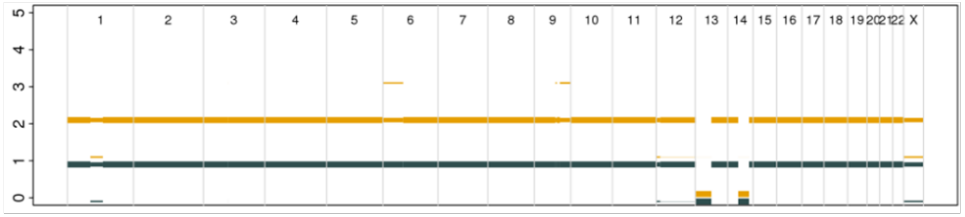
1334



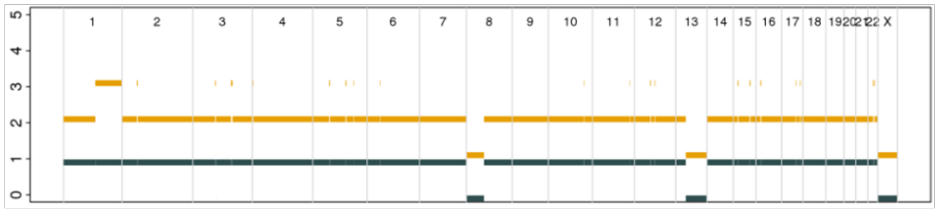


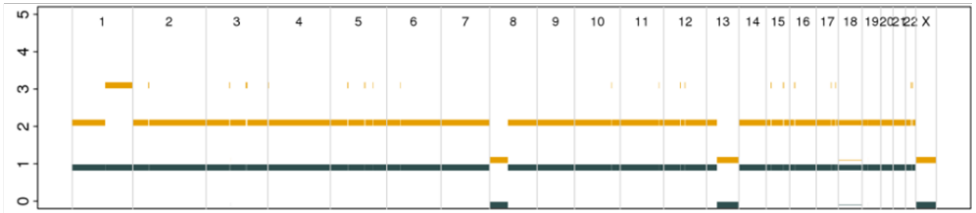
5834



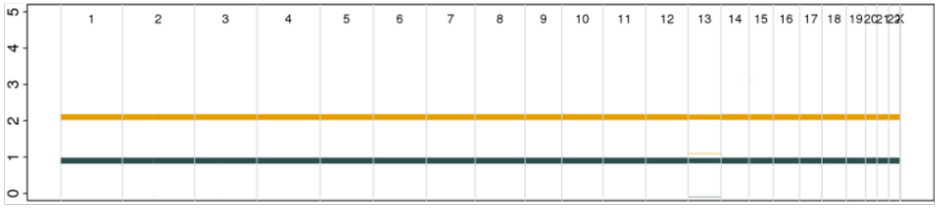


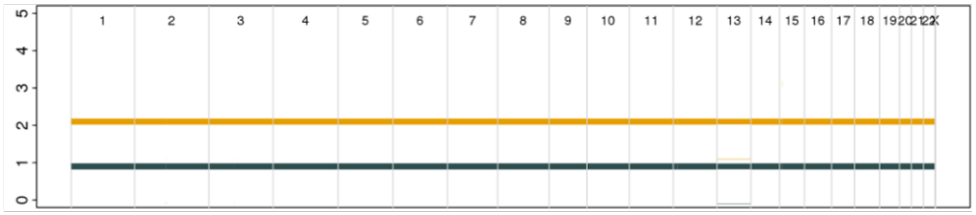
6178



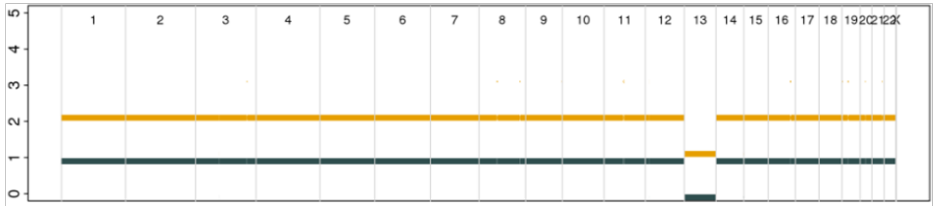


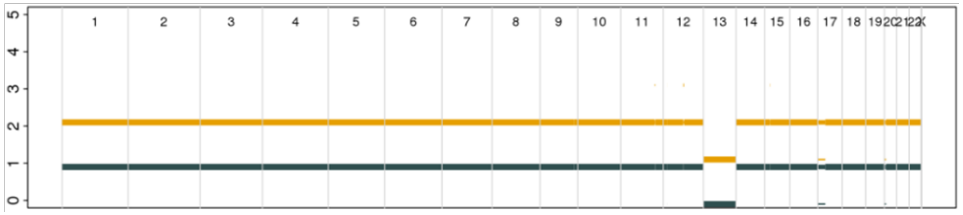
6229



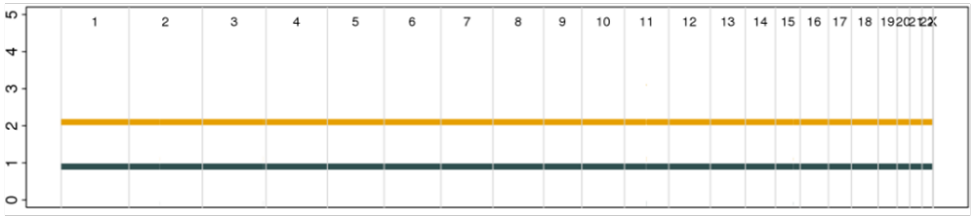


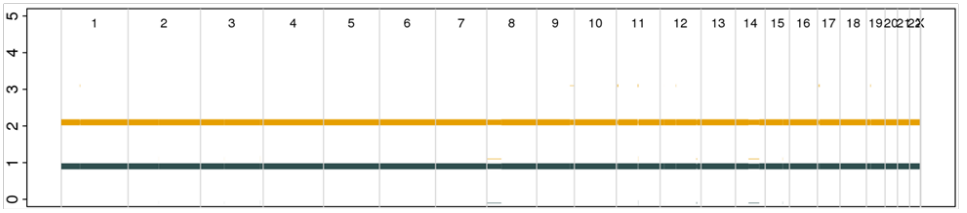
6706



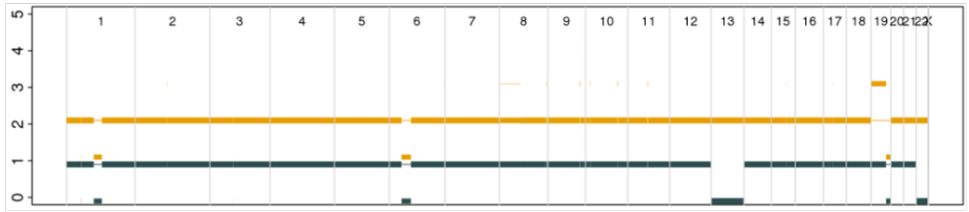


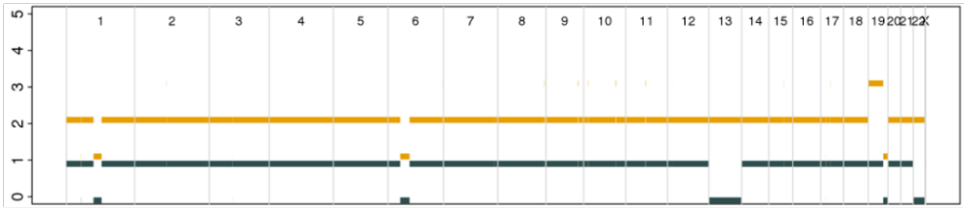
6988



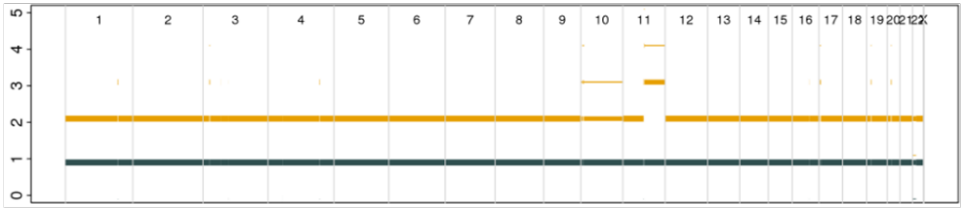


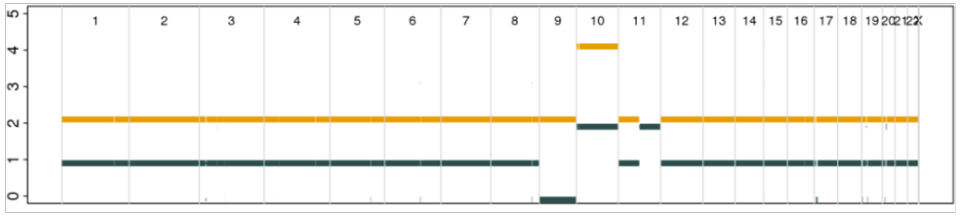
9126



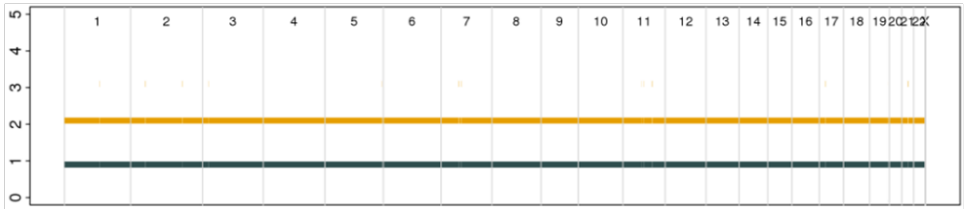


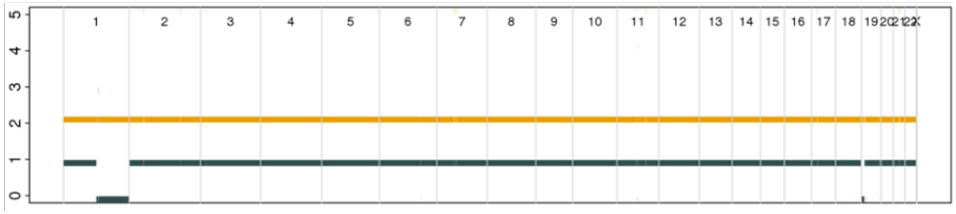
9515



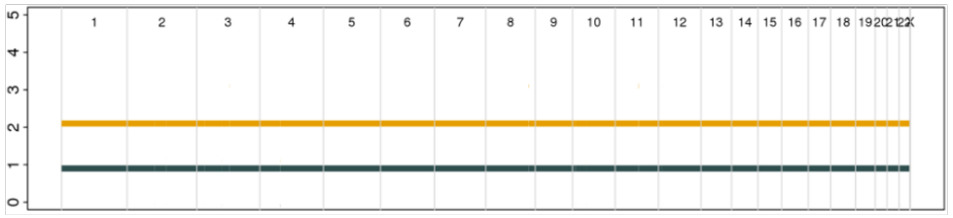


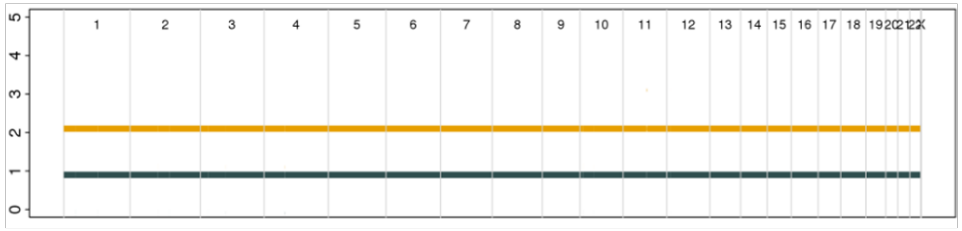
10365





11949

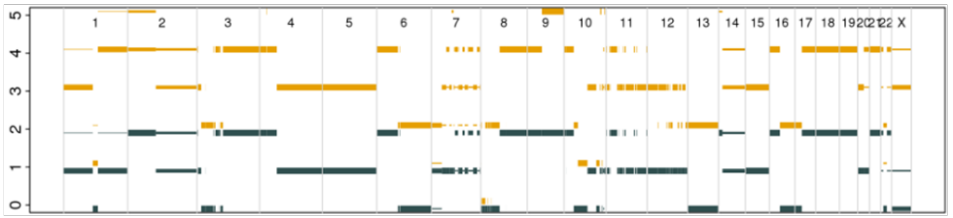


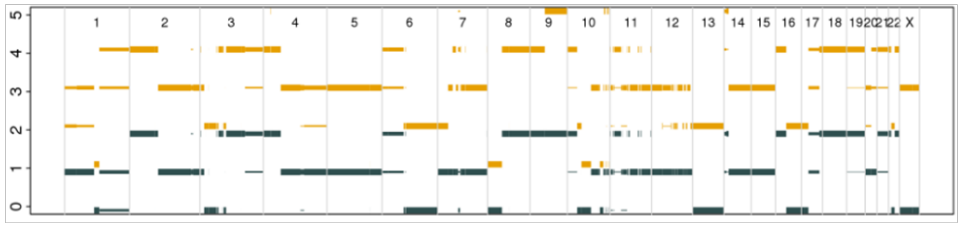


---

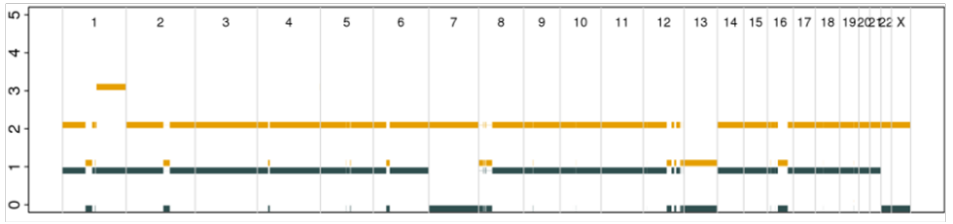
t(14;16)

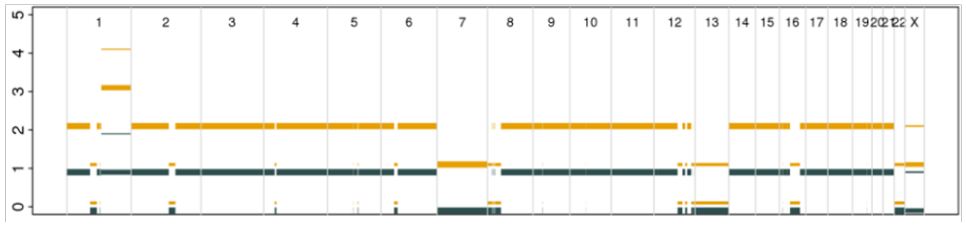
7801



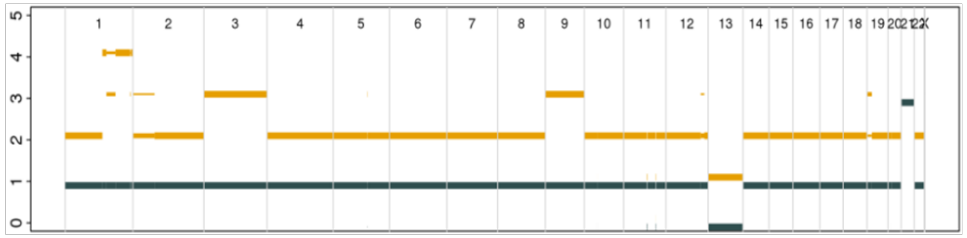


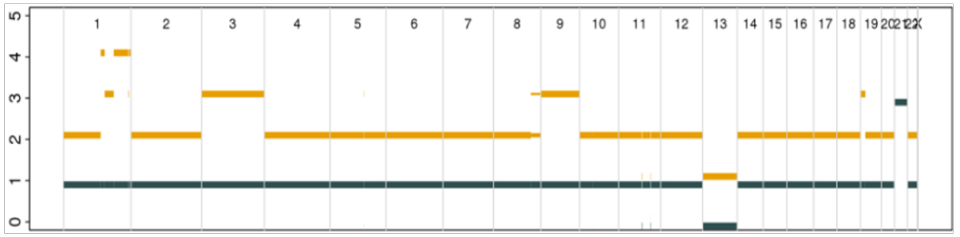
9166



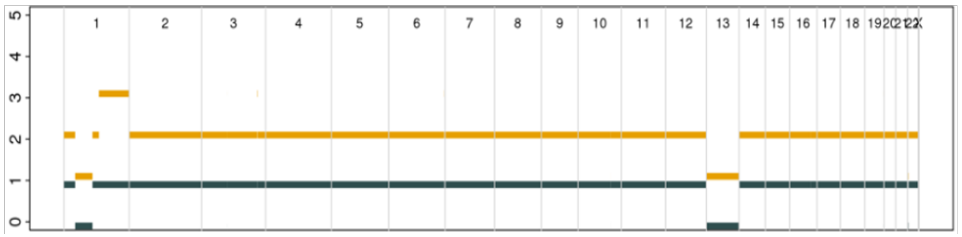


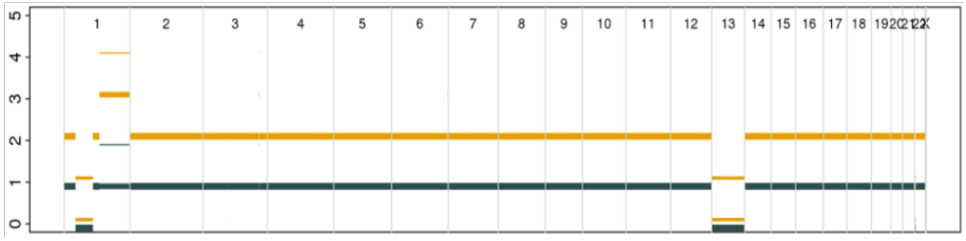
9721


0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22X


0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22X

11506


0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22X


0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22X