A Study of Genomic Instability in Chronic Lymphocytic Leukaemia (CLL)

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by

Faris Jamal M Tayeb

March, 2018

Abstract of thesis

Chronic Lymphocytic Leukaemia (CLL) is a malignancy of mature B cells. The median age at diagnosis is 70-years old, mostly seen in Western Societies. Half of the patients show an indolent phenotype and watchful waiting is the recommended approach for their management. However, once treated, patients are heterogeneous in terms of response and relapse. Regarding prognosis, genetic testing, alongside current clinical staging, is important to guide treatment and prognosis. Deletion of the short arm of chromosome 17 (17p) is one of the worst prognostic markers for CLL and usually involves loss of heterozygosity (LOH) and mutations in the TP53 gene. P53 is one of the cell-cycle regulators that activates senescence, cell-cycle arrest, DNA repair or apoptosis as part of the DNA damage response (DDR). In some severe cases of CLL with inactivated P53, the DDR is affected in favour of CLL survival. However, there is a little knowledge regarding the relationship between TP53 and mutations affecting other DNA repair genes and whether these could lead to a synergistic effect. It was, therefore, the aim of this study to address this important question.

Genomic DNA was extracted from blood CLL cells of 10 patients, all of whom were bearing mutation(s) in TP53 as identified with Sanger sequencing and had progressive disease at the time of sampling. 194 known human DNA maintenance genes were identified and biotinylated-cRNA probes designed (Agilent SureSelect) to enrich DNA from their exons (2786 regions - total size = 500kb) for sequencing using an Ion Torrent Personal Genome Machine (PGM). In terms of coverage, about 99.92% of targeted regions were successfully enriched with 297x average coverage depth. Using the Torrent

Variant Caller (TVC), 365 candidate missense variants in 113 genes were identified from the samples. 268 were single nucleotide variants (SNVs), and 97 were previously unknown or novel (0.002 variants per 1kbp per patient). 90% sensitivity was achieved whereas 60% specificity resulted from a high rate of false positives (FP) found as homopolymer indels. Each of two out of the 10 samples (20%) had separate POLE novel missense mutations, which were validated by Sanger and Whole Genome Sequencing (WGS). This was further investigated with an expanded cohort of patients divided according to TP53 status into TP53 wild-type (n=28) and TP53 mutated (n=31). The results showed no further POLE mutation in the cohort (3.39%) and confirmed the independent role of TP53 pathogenesis in CLL.

Whole genome Sequencing (WGS) to a lesser depth was also applied to the same primary cohort of ten CLL samples. Coverage analysis demonstrated there to be a 98.5% average base coverage and 29.7x average coverage depth. Data analysis found an average of 250 novel missense variants (2.5×10^{-5} variants per 1kbp per sample). The data also confirmed the 17p deletions and mutations. Genotyping data shows that many genes could be affected, involving signal transduction and immune response pathways that may participate in B cell development and CLL pathogenesis affected novel cells, supporting the possibility that oncogenes may initiate CLL carcinogenesis prior to TP53 mutation and chromosomal instability.

Taken together, these results show that there are mutations in DNA repair genes but they are not common, at least for the samples examined. This suggests the independent role of P53 in deactivating DNA repairing mechanisms. Further validation should be applied

using a larger cohort. Furthermore, NGS proved to be a comprehensive tool for examining

a group of genes or even genomes in a robust manner for characterising CLL.

Acknowledgments

First, I would like to acknowledge my PhD supervisors: Professor Andrew Pettitt, Professor David Ross Sibson, and Dr Ke Lin for taking care of me throughout my study. I thank them for guiding me through the research plan, practical aspects, results interpretation, and thesis writing. Their precious help made me confident and able to perform this work.

In addition, I would like to thank all members of the haematology group for their assistance and support during my PhD journey. I'm particularly thanking Dr Gillian Johnson for her support in providing genomic DNA, FISH and FASAY results of CLL206 trial samples, Dr Melanie Oats for providing trial samples, Ms Katie Bullock for running PGM sequencing, Dr Alex Bee for teaching me about operating a Bioanalyzer, and Dr Lihui Wang for being overall supportive and helpful. Many thanks to Dr Xuan Liu and Dr Sam Haldenby from the University Centre of Genome Research for their Bioinformatics assistance. My gratitude also belongs to senior researchers, Dr Kathy Till and Dr Jack Zhuang, as well as the administrative team, including Ms Jackie Henderson and Ms Debra Horne of the Department of Molecular and Clinical Cancer Medicine for always being friendly and helpful.

I would also like to mention my colleagues in the Pettitt's group, including Dr Omar Alishlash, Dr Umair Khan, Dr Ishaque Mohamed, Mrs Waad Almalki, Dr Chan Su, Mr Khalid Naheet, Dr Jina Eagle, and Dr Sozan Karim. From Joseph and Nagash's group, gratitude is merited by Ms Alzahra Alshayeb, Mr Jehad Alhumoud, Mr Anil Mondru, Dr Andrew Duckworth, Dr Jemma Blocksidge, and Dr Mark Glen. From the Richard group, Dr Claire

4

Declaration

I hereby declare that all of the data presented in this thesis is the result of my own work except the DNA extraction, FASAY and FISH assays of CLL206 samples which had been performed by Dr Gillian Johnson. TP53 mutation status of 24 (out of 49 CLL samples) included in the last result's chapter was identified in a separate project contributed by Ms Nichola Rockliffe. The Ion Torrent-sequencing reading was performed by Ms Katie Bullock from Good Clinical Laboratory Practice (GCLP) laboratories, University of Liverpool. Whole Genome Sequencing (WGS) was carried out by the Beijing Genomic Institute (BGI), Hong Kong. WGS data analysis was conducted in collaboration with Dr Xuan Liu at the Centre of Genome Research (CGR), University of Liverpool.

Presentations arising from this work

1. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. NGS Study for Screening Aberrations in DNA Maintenance Genes and Genomic Instability in Chronic Lymphocytic Leukaemia (CLL) (Poster presentation). ITM research day, University of Liverpool, on 06 July 2017

2. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. NGS Study for Screening Aberrations in DNA Maintenance Genes and Genomic Instability in Chronic Lymphocytic Leukaemia (CLL) (Poster presentation). North West Cancer Research Scientific Symposium, poster presentation on 28 April 2017

3. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. A Study of Genomic Instability in Chronic Lymphocytic Leukaemia (CLL) (Poster presentation). Saudi Student Conference (SSC9) at Birmingham, poster presentation on 13 February 2016

4. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. A Study of Genomic Instability in Chronic Lymphocytic Leukaemia (CLL); Deep Sequencing results (Poster presentation) Saudi Student Conference 8 (SSC8) at London, poster presentation on 31 January 2015

5. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. A Study of Genomic Instability in Chronic Lymphocytic Leukaemia (CLL); Deep Sequencing results (Poster presentation) North West Cancer Research Scientific Symposium, poster presentation on 20 March 2014

6. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. A Study of Genomic Instability in Chronic Lymphocytic Leukaemia (CLL); Deep Sequencing initial results (Poster presentation) Saudi Student Conference 7 (SSC7) at Edinburgh, poster presentation on 28 November 2013

7. Faris Tayeb, Ke Lin, D. Ross Sibson and Andy Pettitt. A Study of Genomic Instability in Chronic Lymphocytic Leukaemia (CLL); Deep Sequencing initial results. Poster day at University of Liverpool, between 8-29 April 2013

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
Ab	Antibody
ABT-199	Venetoclax
ACIDA	Activation-Induced Cytidin Deaminase
ADCC	Antibody-Dependent Cellular Cytotoxicity
AID	Activation-Induced Cytidine Deaminase
ANK	Ankyrin
AP	Apurinic or Apyrimidinic Sites
AS	Allele Specific
ATM	Ataxia Telangiectasia Mutated
ATR	Ataxia-Telangiectasia and Rad3-Related
ATRIP	ATR-Interacting Protein
Av.	Average
BWA	Burrows-Wheeler Aligner
FASAY	Functional Analysis of Separated Alleles in Yeast
SSAHA	Sequence Search and Alignment by Hashing Algorithm.
53BP1	p53-Binding Protein 1
6TG	6-Thioguanine
8-oxoG	8-Oxoguanine
B	Billion
BAM	Binary Alignment File
BAM	Binary Alignment Map
BCL2	B-Cell Lymphoma2
BCL3	B-Cell Lymphoma 3
BCR	B- Cell Receptor
BED	Browser Extensible Data
BER	Base Excision Repair
BIRC3	Baculoviral IAP Repeat-Containing 3
BLM	Bloom Syndrome Helicase
BME	β-Mercaptoethanol
bp	Base Pair
BRCA1/2	Breast Cancer Susceptibility 1/2
BRIP1	BRCA 1-Interacting Protein 1
BSA	Bovine Serum Albumin
BTK	Bruton Tyrosine Kinase
BWA	Burrow Wheeler Aligner
C	Cytosine
c.DNA	Complementary DNA
Ca++	Calcium
c-AID	Canonical-AID
CAL101	Idelalisib
CDC	Complement-Dependent Cell Cytotoxicity
CDC25	Cell Division Cycle 25
CDK	Cyclin-Dependent Kinase

CGH	Comparative Genomic Hybridisation
CHD2	Chromodomain Helicase DNA Binding Protein 2
CHOP	Cyclophosphamide, Doxorubicin (Hydroxydaunomycin), Vincristine (Oncovin) and Prednisolone
chr.	Chromosome
CIN	Chromosomal instability
CLL	Chronic Lymphocytic Leukaemia
c-MYC	Avian Myelocytomatosis Virus Oncogene Cellular Homolog
LOH	Loss of Heterozygosity
CNVs	Copy-Number Variants
COSMIC-65	Catalogue of Somatic Mutations in Cancer
Cov.	Coverage
CpG	Cytosine Guanine (Linear Dinucleotide)
CR	Complete Remission
CSA	Cockayne syndrome WD Repeat Protein A
CSR	Class Switch Recombination
CTLp	CtBP-Interacting Protein
CXCR4	CXC Receptor 4
d	Day
DBMs	Domain-Binding Motifs
dbSNP	Database of Single Nucleotide Polymorphism
DD	Death Domain
ddNTP	Dideoxynucleotide
DDR	DNA Damage Response
del	Deletion
DGGE	Denaturing Gradient Gel Electrophoresis
DHLPC	Denaturing High-Performance Liquid Chromatography
DLBCL	Diffuse Large B-Cell Lymphoma
DLEU 1 /2	Deleted in Leukaemia 1 /2
DNA-PKcs	DNA-Dependent Protein Kinase Catalytic Subunit
dNTPs	Deoxynucleotides
DSBs	DNA Double-Strand Breaks
EXO1	Exonuclease 1
FANC	Fanconi Anaemia
FANCDE	Fanconi Anaemia group E
FC	Fludarabine and Cyclophosphamide
FCR	Fludarabine, Cyclophosphamide and Rituximab
FDCs	Follicular Dentritic Cells
FEN1	Flap Endonuclease 1
FISH	Fluorescence *In Situ* Hybridization
Flu	Fludarabine
GATK	Genome Alignment Tool Kit
Gbp	Gigabase Pair
GC	Germinal Centre
gDNA	Genomic DNA

GPS2	G-Protein Pathway Suppressor 2
GRCh37	Genome Reference Consortium Human Build 37
H+	Hydrogen Ion
hg19	Homo Sapiens (human) Genome Assembly 19
HNPCC	Hereditary Non-Polyposis Colon Cancer
HP	Homopolymer Regions
hr.	Hours
HRR	Homologous Recombination Repair
HTRA	HTR Protease A
IARC	International Agency for Research on Cancer
IBM SPSS	International Business Machines Statistical Package for Social Sciences
ICLs	Intrastrand Crosslinks
IDT	Integrated DNA Technologies
Ig	Immunoglobin
IGHV	Immunoglobin Heavy Chain Variable Genes
IGV	Integrative Genomic Viewer
InDels	Insertions/Deletions
ING3	Inhibitor of Growth Family Member 3
IR	ionizing Radiation
IRF4	Interferon Regulatory Factor 4
ISFET	Ion-Sensitive Field-Effect Transistors
ISP	Ion Sphere Particles
ITGA6	Integrin Subunit Alpha 6
IWCLL	International Workshop on Chronic Lymphocytic Leukaemia
Kbp	Kilo Base Pair
KLHL6	Kelch-Like Family Member 6
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
L chain	Light Chain
LDH	Lactate Dehydrogenase
LDT	Lymphocyte Doubling Time
LIG3	Ligase 3
LIG4	DNA Ligase 4
LN	Lymph Node
LOD	Limit of Detection
LOH	Loss of Heterozygosity
LRP1B	LDL Receptor-Related Protein1 B
M	Million
M	Male
MAP2K1	Mitogen-Activated Protein Kinase 1
MAPK	Mitogen Activated Protein Kinase
Mbp	Mega Base Pair
MDM2	Murine Double Minute 2
MDR	Minimal Deleted Region
mg	Milligram
MGA	MAX Dimerisation Protein
MGMT	O6-Methylguanine DNA Alkyltransferase

M-IGHV	Mutated IGHV
min.	Minutes
MIP	Molecular Inversion Probe
miR-15a	Micro RNA 15a
miR-16-1	Micro RNA 16-1
ml	Millilitre
MMR	Mismatch Repair
MNCs	Mononuclear Cells
mRNA	Messenger RNA
MTUS1	Microtubule Associated Tumour Suppressor 1
Mu	Mutated
MW	Molecular Weight
MYCN	V-Myc-Avian Myelocytomatosis Viral Oncogene Neuroblastoma
MYD88	Myeloid Differentiation Primary Response 88
MZ	Marginal Zone
MΩ	Milli Q
N/A	Not Available.
NA	Not Available
NaOH	Sodium Hydroxide
NBS1/NBN	Nijmegen Breakage Syndrome Protein 1
NCBI	National Centre for Biotechnology Information
NDRG1	N-Myc Downstream Regulated 1
NEIL	Nei-Like Protein
NER	Nucleotide Excision Repair
NES	Nuclear Export Signals
NFkB1	Nuclear Factor Kappa B Subunit 1
NFkBIE	NFkB Inhibitor Epsilon
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End-Joining
nM	Nanomolar
NOTCH1	Human NOTCH1 Gene
NR	No Response
NRAS	Neuroblastoma RAS Viral Oncogene Homolog
NT5E	5- Nucleotidase Ecto
ºC	Celsius
OGG1	8-Oxoguanine DNA Glycosylase
OMIM	Online Mendelian Inheritance in Man
OS	Overall Survival
p53	Tumour Suppressor Protein 53
PALB2	Partner and Localizer of BRCA2
PARP1	Poly (ADP-Ribose) Polymerase 1
PARP2	Poly (ADP-Ribose) Polymerase 2
PBS	Phosphate Buffered Saline
PCNA	Proliferating cell nuclear antigen
PCNA	proliferating cell nuclear antigen
PCR	Polymerase Chain Reaction

PD	Progressive Disease
pg	Picogram
PGM	Ion Torrent Personal Genome Machine
Ph+	Philadelphia-Positive
PI3K	Phosphatidyl-Inositol-3 Kinase
PIKK	PI3K-Related Protein Kinase
PIM1	Pim-1 Proto-Oncogene, Serine-Threonine-Kinase
PLCγ2	Phospholipase Cγ2
PLEKHG5	Pleckstrin Homology and Rho GEF Domain Containing G5
pM	Pico Molar
PNKP	Polynucleotide Kinase Phosphatase
Pol D	DNA Polymerase D
PolB	Polymerase B
POLE	DNA Polymerase Epsilon
PolyPhen 2	Polymorphisim Phenotyping
POT1	Protection of Telomere 1
PR	Partial Response
PreB	Precursor B-cells
ProB	Progenitor B-cells
Pro-B-Cells	Progenitor B-Lymphocytes
Q20	Phred Score 20
QC	Quality Control
R	Purine
RAG	Recombinase-Activating Endonuclease
RB1	Retino-Balstoma-1 Gene
RE	Restriction Enzyme Buffer
RECQL4	RecQ Protein-Like 4
Ref	Reference
RFC	Replication Factor C
RFLP	Restriction Fragment Length polymorphism
RGPD4	RANBP2-Like And GRIP Domain Containing 4
RIPK1	Receptor Interacting Serine/ Threonine Kinase 1
RIT	Rituximab
RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
RPA	Replication Protein A
rpm	Rounds Per Minute
RPS15	Ribosomal Protein S15
Rv	Reverse
SAM	Sequence Alignment Map
SAM	Sterile Alpha Motif
SAM	Sequence Alignment Map
SAMHD1	SAM Domain and HD Domain 1
SD	Stable Disease

SD	Standard Deviation
Sec.	Seconds
SF3B1	Splicing Factor 3 B Subunit 1
SHM	Somatic Hypermutation
SIFT	Scale Invariant Feature Transform (Sorting Intolerant from Tolerant)
sIg	Surface Immunoglobulin
SMARCA2	SWI/SNF Matrix-Associated Regulator of Chromatin A2
SMZ	Splenic Marginal Zone
SNP	Single Nucleotide Polymorphism
SNVs	Single Nucleotide Variants
SSAHA	Sequence Search and Alignment by Hashing Algorithm
SSBR	Single-Strand Break Repair
SSBs	Single-Strand Breaks
SVs	Structural Variants
T	Thymine
TAD	Transactivation domain
TAE	Tris-Acetate-EDTA
TBE	Tris-Borate-EDTA
Tbp	Tera Base Pair
TD	T-Cell-Dependent Immune Response
TDP1	Tyrosyl-DNA Phosphodiesterase 1
TFS	Treatment-Free Survival
Th	CD4+ T Helper Cells
TI	T-Cell-Independent Immune Response
Ti	Transition
TLS	Translesion Synthesis
Tm	Melting Temperate
TMAP	Torrent-Mapping Alignment Program
TNFAIP3	TNF Alpha-Induced Protein 3
Tv	Transversion
TVC	Torrent Variant Caller
Tx	Therapy
U	Uracil
UK	United Kingdom
UM-IGHV	Unmutated IGHV
UPD	Uni-Parental Disomy
USCS	The University of California, Santa Cruz Genome Database
UV	Ultraviolet
V	Voltage
v	Version
VAF	Variant Allele Frequency
Var	Variant
VCF	Variant Call Format
VDJ	Variable-Diversity-Joining
VQ	Variant Quality Score
W	Thymine or Adenine

WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
WRN	Werner Syndrome Helicase
Wt	Wild Type
XLF	XRCC4-Like Factor
XP	Xeroderma Pigmentosum
XPC	Xeroderma Pigmentosum Group C-Complementing Protein
XPO1	Exportin 1
Y	Pyrimidine
ZAP-70	Zeta-Chain-Associated Protein Kinase 70
ZMYM3	Zinc Finger MYM-Type Containing 3
β2-M	β2-Microglobulin
μg	Microgram
μl	Microlitre

1. Chapter 1. Literature Review

1.1. Introduction to the Clinico-Biology of Chronic Lymphocytic Leukaemia

Chronic lymphocytic leukaemia (CLL) is defined as a monoclonal B-cell malignancy, where B lymphocytes accumulate in the blood, bone marrow and lymphoid tissues. CLL is considered the most common leukaemia of adults in Western countries, accounting for up to 40% of all adult leukaemias (1). The disease's annual incidence is 3-4 per 100,000 and its likelihood increases with age with a median age at diagnosis of 71 years (2). The pathogenesis of CLL is unclear; environmental factors seemingly contribute to disease development. In a small proportion of patients (5-10%), familial predisposition is implicated (3) but most cases of CLL are of sporadic origin.

CLL can be a relatively indolent disease but its natural history is highly variable, and patient survival can range from 2 to 20 years after initial diagnosis (4). Although CLL is incurable, current treatment regimens can result in disease control and minimal residual disease. CLL's systemic nature means that treatment is in the form of adjuvant chemotherapy, which more recently also includes immunotherapy. However, some patients are treatment resistant and for others, there is an elevated chance of relapse to a more aggressive form of the disease (5).

Although a number of the factors that determine treatment response are understood, for example, defects in the pathways that lead to apoptosis, these are not sufficient in isolation to explain the highly variable nature of CLL. One "hallmark" of cancer is genetic instability. The purpose of the research project reported here was to investigate whether

there is evidence that genetic instability could contribute to the behaviour of the most aggressive forms of CLL.

1.2. Aim of the Thesis

The aim of this work is to target enriched DNA maintenance genes from a panel of aggressive forms of CLL with inactivated p53 and use Next-Generation Sequencing (NGS) at high depths to screen for mutations. In addition, a parallel study using low-pass Whole-Genome Sequencing (WGS) would enable incidences of global patterns of genetic alterations to be identified and compared to the maintenance genes.

1.3. Normal B-Cell Maturation

Normal B-cell differentiation is a complex and tightly regulated process. B-cell lymphocytes are differentiated from common lymphoid progenitor cells (CLPs), which are derived from haematopoietic stem cells (HSCs). Independent of antigens, these cells then pass through multiple stages in the following order: progenitor B-cells (proB), precursor B-cells (preB), immature B-cells and mature B-cells. They develop further into memory and plasma cells in secondary lymph tissues, such as spleen and lymph nodes, where through initial challenges by foreign antigens they recognise leads to their clonal selection through activation and proliferation or in the case of disease, by auto-antigens. These end-stage B-cells are then released into circulation where they respond to stimulation by the same antigens to which they originally encountered that caused stimulation (Figure 1-1).

Antibody diversity is achieved during B-cell maturation by antigen receptor gene diversification. Initially, the recombinase-activating gene (RAG) induces V(D)J

recombination at pre-B-stage in bone marrow (6, 7). Further remodelling occurs in the

germinal centres of lymph tissues via somatic hypermutation and class-switch

recombination (CSR) whereby the constant region of one immunoglobulin gene (IG) is

substituted with the constant region of another, thus modifying its effector function.

Central to this process are DNA double-strand breaks (DSBs), intermediate created by

activation-induced cytidine deaminase (AID) and subsequent repair of the distally-severed

ends (8, 9).
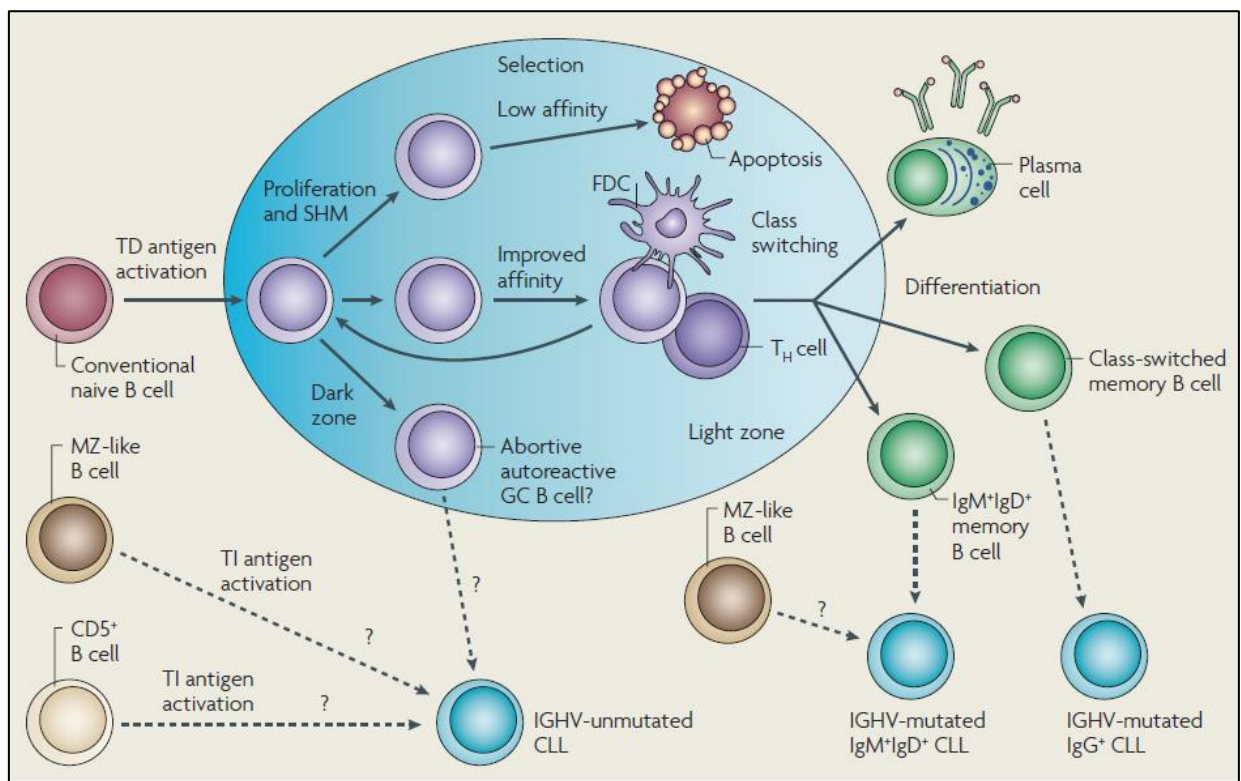
Figure 1- 1: Normal B-Cell Interactions in Germinal Centres (GC) and CLL Development. B-cells enter the GC, which is established in B-cell follicles after a T-cell-dependent (TD) immune response, leading to massive clonal expansion. In parallel, somatic hypermutation (SHM) takes place in the GC dark zone, resulting in a very high rate of induced mutations within immunoglobulin (Ig) variable region genes (10). B-cells then migrate to the GC light zone where cell-cell interactions occur with CD4+ T-helper (Th) cells and follicular dendritic cells (FDCs). Following this interaction, B-cells acquire a B-cell receptor (BCR); cells with high-affinity receptors are processed by class switch recombination (CSR) to different cell types, while low-affinity cells undergo apoptosis. CSR affects Ig heavy-chain constant region genes. Usually, B-cells developed previously in the GC repeatedly migrate between dark and light zones (11). B-cells with high affinity

undergo multiple rounds of multiplication, mutation, selection and class switching into memory B-cells and plasma cells and exit the GC. The figure is taken with permission from Zenz T, Mertens D, Kuppers R, Dohner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. Nature Reviews Cancer. 2010;10(1):37-50 (12).

1.3.1. B-Cell Precursors and the Development of CLL

Current knowledge supports post-GC memory cells, which are T cell-dependent (TD) and antigen-activated as well as the likely source for *IGHV*-mutated CLL (see Figure 1- 1**)**. Some argue that the antigenic specificities of CLL could originate from B-cells in both TD and T-cell-independent (TI) antigen activation (13-15). On the other hand, *IGHV*-unmutated CLL is likely derived from antigen-activated B-cells, but it is still unclear whether they are from conventional naïve B-cells, CD5+ B-cells, or marginal zone (MZ)-like B-cells.

1.4. Clinical Characterisation of CLL

The International Workshop on Chronic Lymphocytic Leukaemia (IWCLL) 2008 Guidelines (16) base the clinical diagnosis of CLL on the presence of 5000 clonal matured B-cells per microliter of peripheral blood as confirmed by flow cytometry. The detection of characteristic surface proteins for markers of CLL cells, which include CD5, B-cell antigens CD19 and CD23 and low levels of surface immunoglobulin (sIg), CD20 and CD79b (16-18) confirms the diagnosis.

Clinical presentation of CLL is variable. At the time of initial diagnosis, there are >25% of patients that are asymptomatic while others present features of advanced disease, including anaemia, infection and massive lymphadenopathy (19). Patients with the indolent disease have a lifespan similar to age- and sex-matched healthy controls while

others suffer from aggressive, therapy-resistant disease or one that transforms into a more aggressive disease (Richter Syndrome (RS)) and die within a few years of diagnosis (20, 21).

There are two clinical staging systems for CLL: modified Rai and Binet. The Rai system defines three risk categories, low-risk, intermediate- and high-risk, based on the number of lymphocytes, lymph node enlargement and features of marrow involvement (22). The Binet staging system employs similar parameters, but the number of involved anatomical lymph node areas is also considered (23). These staging systems are simple and useful as guidance for starting CLL treatment (20) but fail to specifically predict disease progression and response to therapy, which usually requires molecular testing. Therefore, they are not helpful for tailoring specific therapies to individual patients (24).

Regarding RS, and according to the World Health Organisation (WHO 2008), RS is defined as a pattern of aggressive lymphoma which is largely related to CLL with 0.5% incidence rate per annum. Using CD markers, CD30, CD15 and CD20, RS falls into two cellular types; diffuse large B-cell Lymphoma (DLBCL) or Hodgkin lymphoma (HL) (21). In general, many RS colonies have a relationship with CLL clones in the patient as detected after examining IGHV status (25). RS diagnosis is confirmed by the lymph node biopsy that is initially screened by Positron Emission Tomography (PET)- Computed Tomography (CT) (21). Furthermore, various prognostic markers are helpful in confirming the diagnosis, such as TP53, NOTCH1, MYC and cell cycle CDKN2A. RS usually has a severe prognosis owing to the fast infiltration of the cancer cells and chemo resistance (21).

1.5.   CLL Treatment

1.5.1.   Indications for treatment

Asymptomatic or early-stage patients are usually monitored without therapy whereas chemotherapy is applied with the goal of eliminating excess B-cells in progressive or symptomatic disease cases. Evidence that treatment prolongs survival in non-progressive disease is lacking (26), moreover chemotherapy may cause co-morbidity and can increase risks of secondary tumours (16). Patients having intermediate- or high-risk groups in the Rai system or at Binet Stage B and C do indeed require therapy. It can be in a form of a chemoimmunotherapy (CIT) regimen, novel agent and/or maintenance therapy, including bone marrow transplantation as an alternative option for high-risk patients (27). Regarding therapy outcomes, patients can achieve remission, minimal residual disease or disease control. Response rates up to 95% are possible with complete response in up to 43% of cases and progression-free survival of more than 30 months. Treatments need to be balanced for patient fitness and side effects, in particular neutropenia and thrombocytopenia. Relapsed CLL is typically responsive to additional rounds of treatment but may become resistant thereafter.

Given that CLL is an excess of B-cells, chemotherapy typically targets proliferation. Most therapies induce DNA damage, either through alkylating agents or incorporation of purine analogues, in an effort to trigger apoptosis.

1.5.2.   Front-line treatment

Alkylating agents, such as cyclophosphamide, chlorambucil and bendamustine, have been used in first-line therapy regimens. Their mechanism of action is to alkylate DNA, RNA and

proteins of tumour cells, leading DNA crosslink formation on guanine nucleobases, DNA damage and subsequent activation of apoptosis (28). They have been utilised in combination with immunotherapy regimens to improve survival and prolong the response duration. Moreover, in a German CLL 11 trial, treated treatment-naïve patients with chlorambucil with anti-CD20 antibody provide longer ORR and overall survival (OS) than clorabmbucil alone (29). However, the majority of patients eventually develop resistance. There are two mechanisms for drug resistance in CLL; the first is alteration to the P53-dependent apoptotic pathways, such as genetic alterations to TP53 and ATM (30, 31)– (see section 1.11 TP53 Defects in CLL, page63), while the second is alteration to metabolism and transport of alkylating agents (26).

Likewise, a purine analogue is another DNA damaging agent that is used in CLL treatment, such as fludarabine, cladribine and pentostatin. It initiates elimination of CLL cells by damaging DNA via purine incorporation (32), leading to apoptosis via activation of p53 (33) (see Section 1.9, P53-Dependent Apoptosis, page 54). In combination with alkylating agents, they are associated with longer overall survival, but with more adverse effects on patients (34, 35). Results from a UK CLL 4 trial showed that patients treated with treatment regimens including fludarabine plus cyclophosphamide (FC) had better overall response rate (ORR) than patients treated with fludarabine alone (34). Both F and C have a synergistic effect - they induce DNA damage by targeting purine nucleobases with different DNA damage mechanisms (36). Therefore, defects in the P53 pathway, either p53 itself or its upstream activator, ATM, are associated with resistance to these chemotherapies (32).

Monoclonal antibodies targeting the B-cell specific antigen CD20 have been introduced in the treatment of CLL in a treatment regimen combined with traditional chemotherapy. They exert their effects via complement-dependent cell cytotoxicity (CDC) and antibody-dependent cellular cytotoxicity (ADCC) and therefore, they are not reliant on the p53 pathway (37). Such antibodies include rituximab, ofatumumab (38) and obinutuzumab (GA-101) (39). The latter two antibodies have a different orientation and more stable binding to their targets, thereby achieving a better response than the first.

1.5.3. Second-Line Treatments

As drug resistance has emerged in CLL, especially to chemotherapy and DNA damaging agents, second-line treatments are currently recommended in high-risk CLL cases. One of the common agents used as an initial therapy is the monoclonal antibody, ibrutinib; a Bruton gamma-globulinaemia tyrosine kinase (BTK) inhibitor which inhibits BCR signalling pathways and therefore has been preferably employed in patients with aberrant p53 functioning and providing significantly longer Progression-Free Survival (PFS) and OS as a monotherapy in comparison to chlorambucil (40), or ofatumumab (30). It has been demonstrated to rapidly reduce lymadepathy accompanied by lymphocytosis owing to chemokine receptor inhibition by the agent that prevents lymphocyte migration to lymph nodes but not as a sign of disease progression (41). However, certain patients have acquired mutations in BTK and phospholipase Cγ2 (PLCG2), leading to ibrutinib resistance and adverse clinical outcomes (30, 42).

Recently, lenalidomide (a thalidomide analogue), which is an immune-modulatory agent that acts to block the induction of cancer cell-induced T-cell tolerance and angiogenesis,

has been approved for treating different haematological malignances. In CLL, lenalidomide works either as a monotherapy or as an adjuvant with anti-CD20 agent (43). Lenalidomide functions through activating p21 pathways that supress cell-cycle kinases and proliferation of CLL cells (187).

Furthermore, idelalisib is a new agent for relapsed CLL patients. A combination of PI3K-signalling inhibitors, such as idelalisib (CAL101) and rituximab, has been utilised in relapsed/refractory CLL patients with 17p deletion or TP53 mutations to elongate progression-free survival compared to rituximab monotherapy (44).

Another treatment option is the BCL2 inhibitor, ABT-199 (Venetoclax), which acts by mimicking pro-apoptotic BH3 family proteins to suppress BCL2 anti-apoptotic proteins. It has been shown to be effective in patients with aberrant p53 mutations and chemo resistance (45).

Several drugs have been released into the markets but are then discontinued based on severe side effects. Alemtuzumab is an example of such a treatment - it acts independently of the P53 pathway. Alemtuzumab binds the protein, CD52, found on mature lymphocytes and targets them for destruction by the immune system (26). In the past, for patients with fludarabine resistance and defective p53 functioning, alemtuzumab had been employed with a combination of high dose steroids to mitigate adverse reactions (26, 46). In addition, lenalidomide as an adjuvant agent has been included in treatment regimens for patients on alemtuzumab and dexamethasone and resulted in longer PFS .
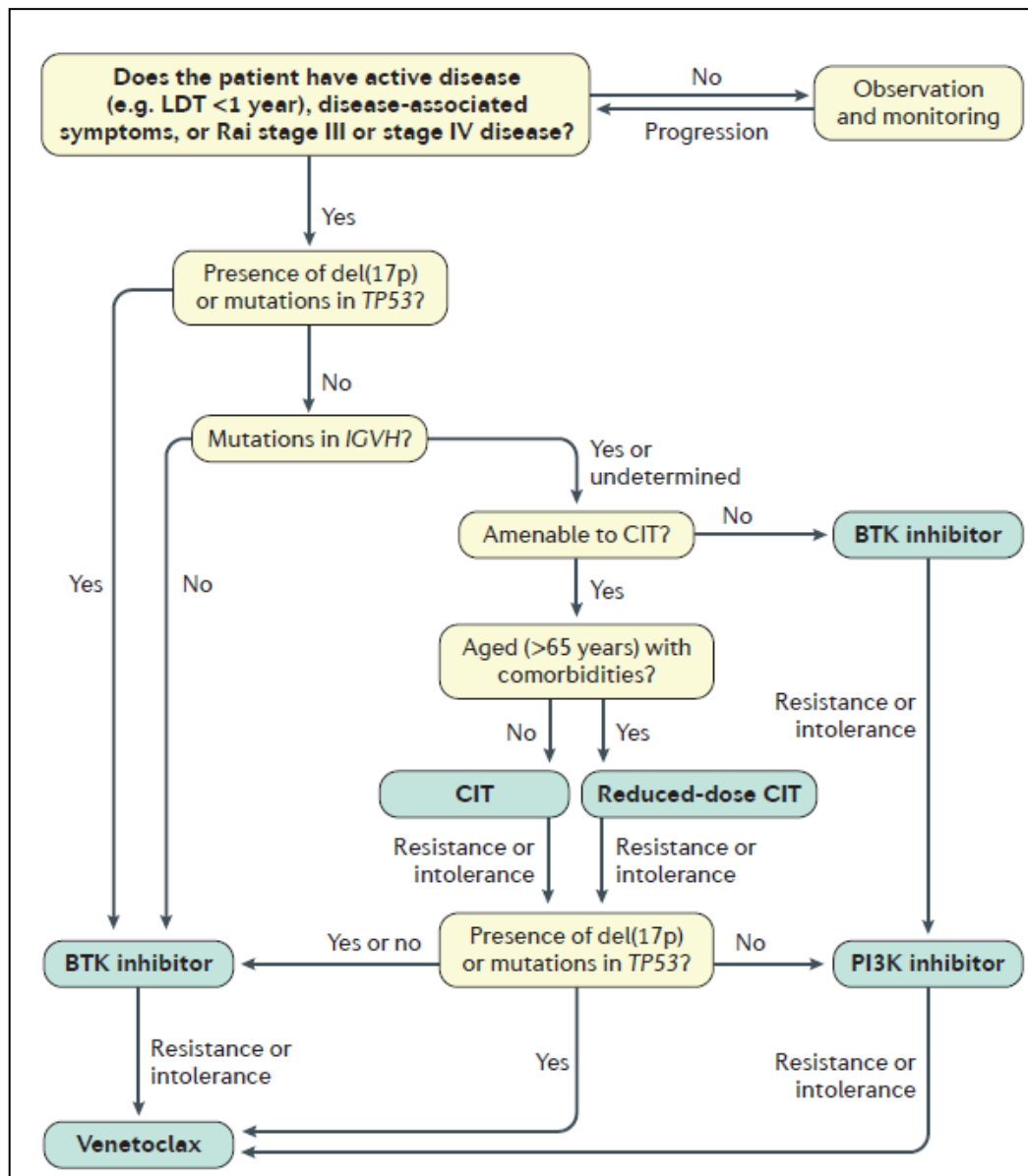
Figure 1- 2: CLL Treatment Strategy. In progressive CLL cases, a BTK inhibitor is used for TP53 deleted/mutated cases or as a primary treatment in general. In wild-type TP53 cases, a chemoimmunotherapy (CIT) regimen is recommended unless the patient is unfit for therapy or has non-mutated IGVH. Furthermore, PI3K inhibitor and Venetoclax (BCL2 inhibitor) could be second to BTK and/or CIT inhibitors in case of resistance or intolerance. The figure is adapted with permission from Kipps TJ, Stevenson FK, Wu CJ, Croce CM, Packham G, Wierda W, et al. Chronic lymphocytic leukaemia (vol 3, 16096, 2017). Nature Reviews Disease Primers. 2017;3:1.

1.5.4. NCRI CLL206 and NCRI CLL210 Trials

In this research project, many CLL samples were obtained from the NCRI CLL206 and

CLL210 trials. NCRI CLL206 was the National Cancer Research Institute CLL206 trial,

established as a UK multicentre trial using the induction regimen of alemtuzumab plus high-dose methylprednisolone for high-risk CLL with 17p deletions. The second phase showed that the regimen improves median PFS to 11.8 months (46, 47). NCRI CLL210 was also a UK multicentre trial and was established to improve upon the results of the CLL206 trial with the drug regimen comprised of alemtuzumab, dexamethasone and lenalidomide followed by randomisation to lenalidomide maintenance. The regimen appeared to be effective and the addition of lenalidomide had an acceptable safety profile with 29.3 PFS (48).

Despite steady and significant progress improving outcomes for CLL patients through the optimisation of chemotherapy regimens, treatment resistance is a major problem and effective solutions require the response of CLL cells to be taken into account. Cell-cycle checkpoints, DNA repair and the propensity for cell death all have a bearing on the treatment response for CLL. Chemotherapy for CLL in terms of damaging DNA invokes the DNA Damage response (DDR) to affect repair. Time to successfully accomplish the repair processes is obtained by integration of checkpoint signalling and cell-cycle control. This prevents permanent DNA damage that would otherwise occur during replication and mitosis. Cells whose DNA damage exceeds the threshold for satisfactory repair undergo cell death typically by programmed cell death or apoptosis.

1.6.   DNA Damage Response and Replicative Stress

DNA naturally incurs damage on a constant basis by multiple influences. Oxidative DNA damage happens at least 10,000 times per cell per day in humans as a result of free radicals produced by endogenous metabolism and cellular processes (47). Protective

mechanisms, known as the DNA damage response (DDR), have therefore evolved to manage the consequences either to fix such damage or remove the affected cells. The DDR, therefore, protects the organism against genomic instability leading to cancer development (49). In addition to DNA damage, DNA replication stress, whereby the DDR and other cellular responses result in a collapse in ordered DNA replication, for example through stalled or failed replication forks and accumulation of unpaired single-stranded DNA (50, 51).

1.7. DNA Repair Pathways

1.7.1. Base Excision Repair (BER) and Single-Strand Break Repair

1.8. Single-strand breaks (SSBs) are one form of frequent damage to genomic DNA. The action of reactive oxygen species (ROS) and base excision enzymes acting at the sites of DNA damage caused by spontaneous deamination, ROS oxidation, or alkylation can result in SSBs. ROS induces 8-oxoguanine (8-oxoG) and 5-hydroxycytosine (5hmC), which if left unchanged leads to transverse mutations with adenine and thymine, respectively, when they are copied during replication (52, 53). High levels of spontaneous DNA deamination are also observed that give rise to apurinic/apyrimidinic (AP) sites , which similarly will yield incorporation of incorrect bases if not repaired in advance of replication (52). Although SSBs superficially appear to be a less significant problem than abasic sites, the former are often left with modified 3' or 5' ends that cannot support normal replication. Failure of topoisomerase I, for example, results in an SSB with the enzyme attached to the 3' end of the break point.

Where present, 'abasic' sites (AP sites) are formed after removing damaged bases using BER glycosylases and BER endonucleases. Depending on the enzyme, a single-stranded nick is generated, if not already present. BER and single-strand break repair (SSBR) use overlapping components and have similar repair actions involving four main steps: detection, end processing, end filling and ligation. Lesions are detected primarily by a member of the poly (ADP-ribose) polymerase (PARP) family (54). A large number of possible enzymes then correct damage to 3' and/ or 5' termini depending on the lesion (55). The BER pathway is divided into short- or long-patch BER that either repair one or up to 13 nucleotide bases, respectively. The pathway relies on glycosylases, endonucleases, DNA ligases, DNA polymerases, poly (ADP-ribose) polymerase 1 (PARP1) and PARP2 to mediate the process (Figure 1.2). PARP1 is essential in BER and also participates in DSB repair (54). El-Kamisly et al. (2003) suggested that PARP1 activates DSB repair and recruits XRCC1 and other proteins during the BER process (55). Moreover, BER can repair DNA damage triggered by ionizing radiation (IR), topoisomerase I failure (56) or DNA methylating agents (57, 58).
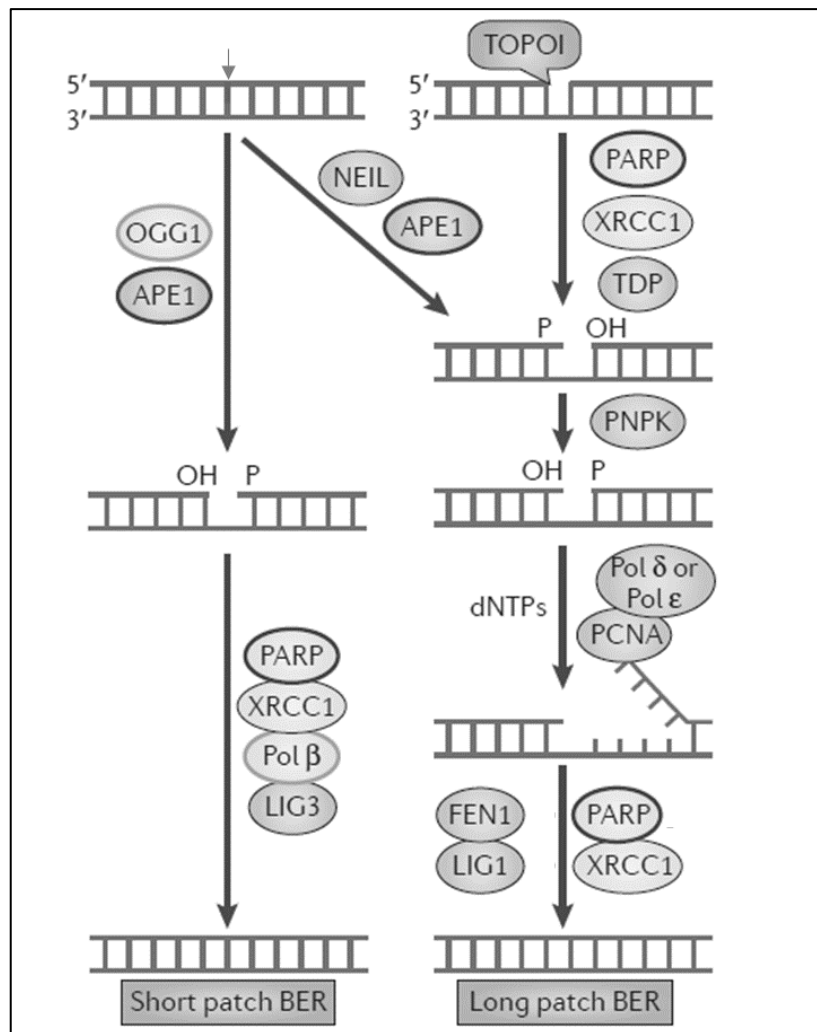
Figure 1-3. DNA Repair by Base Excision Repair (BER). The affected base is removed by OGG1 or by a NEIL family protein. The AP site is hydrolysed by APE1 and repaired according to the lesion size by either short- or long-patch BERs. In long-patch repair, PNPK is sometimes necessary to modify the broken ends followed by FEN1, POLD, POLE and LIG1 to complete patch repair Subsequently, FEN1 is required for long-patch repair. In short- (single) patch repair, POLB and LIG3 are recruited to nucleotides to replace and re-join the AP site. In addition, PARP1 and XRCC1 are involved to provide a scaffold for both long- and short-patch repairs (55). The figure is adapted with approval from Curtin et at. (2012) (59). Abbreviations: OGG1, 8-oxoguanine DNA glycosylase; NEIL, Nei-like protein; AP, apurinic or apyrimidinic; PNKP, Polynucleotide ki (60)nase phosphatase; PCNA, Proliferating cell nuclear antigen; FEN1, flap endonuclease 1; PolB, polymerase-B; LIG3, ligase 3; PARP1, Poly (ADP-ribose) polymerase 1; and TDP1, tyrosyl-DNA phosphodiesterase 1.

1.8.1. Nucleotide Excision Repair (NER)

Carcinogens, such as environmental and endogenous DNA damaging agents can introduce

DNA bulky adducts that stall replication forks and contribute to replication stress and

genomic instability (61). Examples include ultraviolet (UV) radiation and chemical mutagens found in tobacco, primarily polycyclic aromatic hydrocarbons (PAH) and the nicotine-derived nitrosamines. The cellular response to chemotherapy drugs mirrors many of the consequences of these agents.

DNA adducts distort the helix and in response, cells activate the NER pathway, which removes them. Interstrand and intrastrand crosslinks (ICLs) also invoke NER, which in this case requires ERCC1 and xeroderma pigmentosum (XP) proteins (Figure 1-4).

1.8.2. Translesional DNA Synthesis (TLS)

Several DNA polymerase classes, for example, Pol X, Y and A can synthesise DNA across and beyond the lesion site (known as translesion synthesis (TLS)), leading to the possibility of cell survival when the replication fork has otherwise stalled (61). However, many such polymerases have no DNA proofreading capability, thus error rates can be high. Commonly, a preferred, default nucleotide is inserted, for example, deoxyadenosine, regardless of the affected base (61). As a result, TLS is considered a DNA damage tolerance process in addition to a DNA repair mechanism(61).
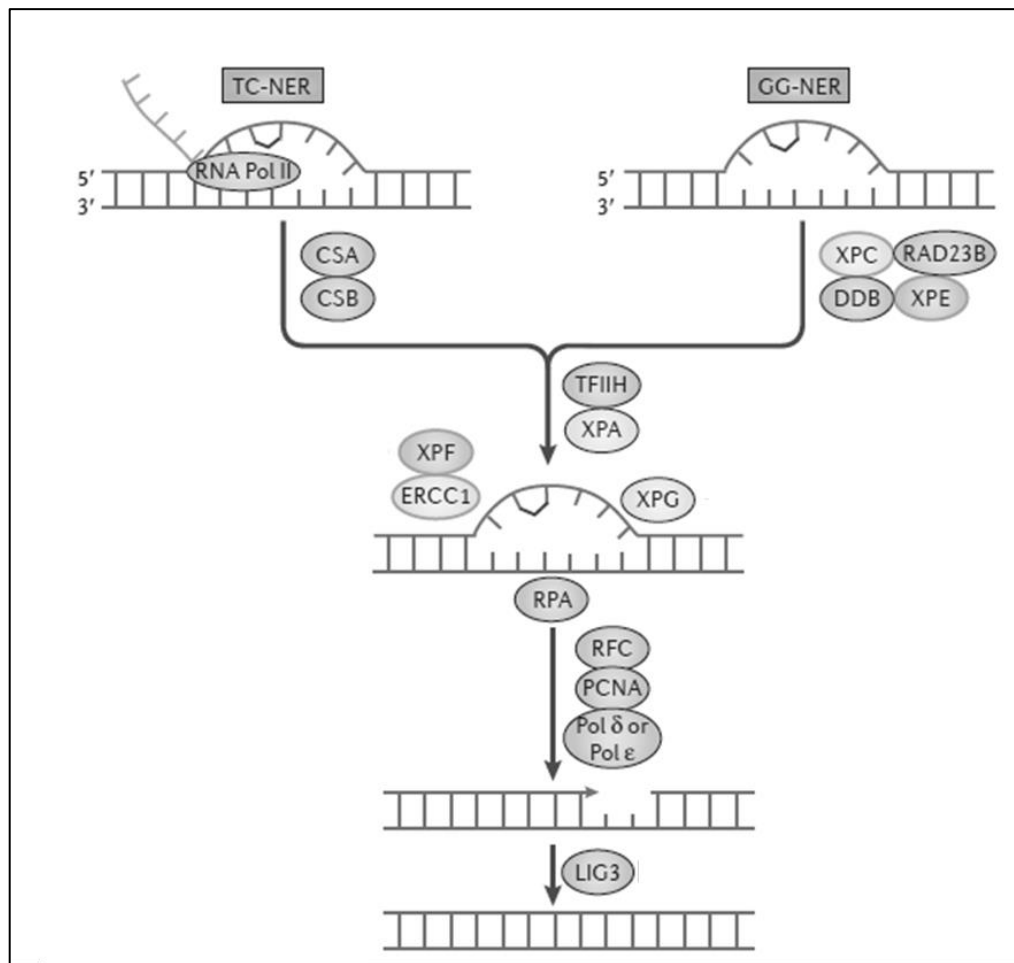
Figure 1-4. Nucleotide Excision Repair (NER). There are two common initiation routes of NER: transcription-coupled nucleotide (TC-NER) and global genome NER (GG-NER) (62). TC-NER recruits CSA and CSB, whereas GG-NER involves XPC and RAD23B. Both routes subsequently recruit the same proteins including XPA, RPA and TFIIH. The downstream steps include nucleotide excision by XPG, ERCC1 and XPF. Thereafter, the damaged nucleotides are replaced and ligated by POLD, POLE and LIG3. The figure is adapted with permission from Curtin et al. (2012) (59). Abbreviations: CSA, Cockayne syndrome WD repeat protein A; XPC, Xeroderma pigmentosum group C-complementing protein; RPA, replication protein A; Pol D, DNA polymerase-D; LIG3, ligase 3; PCNA, proliferating cell nuclear antigen; ERCC: ERCC Excision Repair 1, Endonuclease Non-Catalytic Subunit; DDB; Damage Specific DNA Binding protein; and RFC, replication factor C.

1.8.3. Mismatch Repair (MMR)

Despite its high fidelity, replication of genomic DNA can incorporate the wrong

nucleotides at a low frequency, leaving mismatches. MMR targets such replication errors

(Figure 1-5).

MMR must distinguish leading from lagging strand synthesis in order to identify the base that is in error. How this occurs in eukaryotes is unclear but PCNA, which acts as a clamp for DNA and a scaffold for DNA-modifying enzymes may have a role (59). Recent work has shown that Replication Factor C (RFC) depends on nicks to load the replication-sliding clamp PCNA. This takes place in an orientation-specific manner with one face of the doughnut-shape protein pointing towards the 3'-end at the nick. MMR proteins incorporated into the complex include MutLalpha to one strand in the presence of a mismatch and MutSalpha or MutSbeta (63). MutLalpha is an endonuclease and introduces strand breaks in the presence of a mismatch and these facilitate endonuclease that removes mismatched DNA and allows correct replacement by replication. Ligases and DNA polymerases are therefore amongst other essential proteins found in the complex (64).

1.8.4. Non-Homologous End-Joining (NHEJ)

Double-strand breaks (DSBs) may also occur in DNA and are difficult lesions to repair. It is estimated that each cell has about 50 DSBs daily. Most of them arise endogenously by the activity of ROS (65). NHEJ repairs DSBs by ligating lesions with a minimal processing, it works during all phases of the cell cycle, most predominantly in the G_0 and G_1 phases (66), yet it lacks proofreading abilities. NHEJ is responsible for about 85% of DSB fast repairs, which are induced by IR (67, 68). Essentially, three proteins form the NHEJ complex, namely KU70 (XRCC6), KU80 (XRCC5) and the catalytic subunit of DNA-dependant protein kinase (DNA-PKcs). These proteins form aggregates with ligase 4 (LIG4), Artemis and XRCC4-XLF (67). DSBs are then detected by ATM and MRN complexes, which include

MRE11, Nijmegen breakage syndrome protein 1 (NBS1) and RAD50. They can also

function as an early repair event in NHEJ.
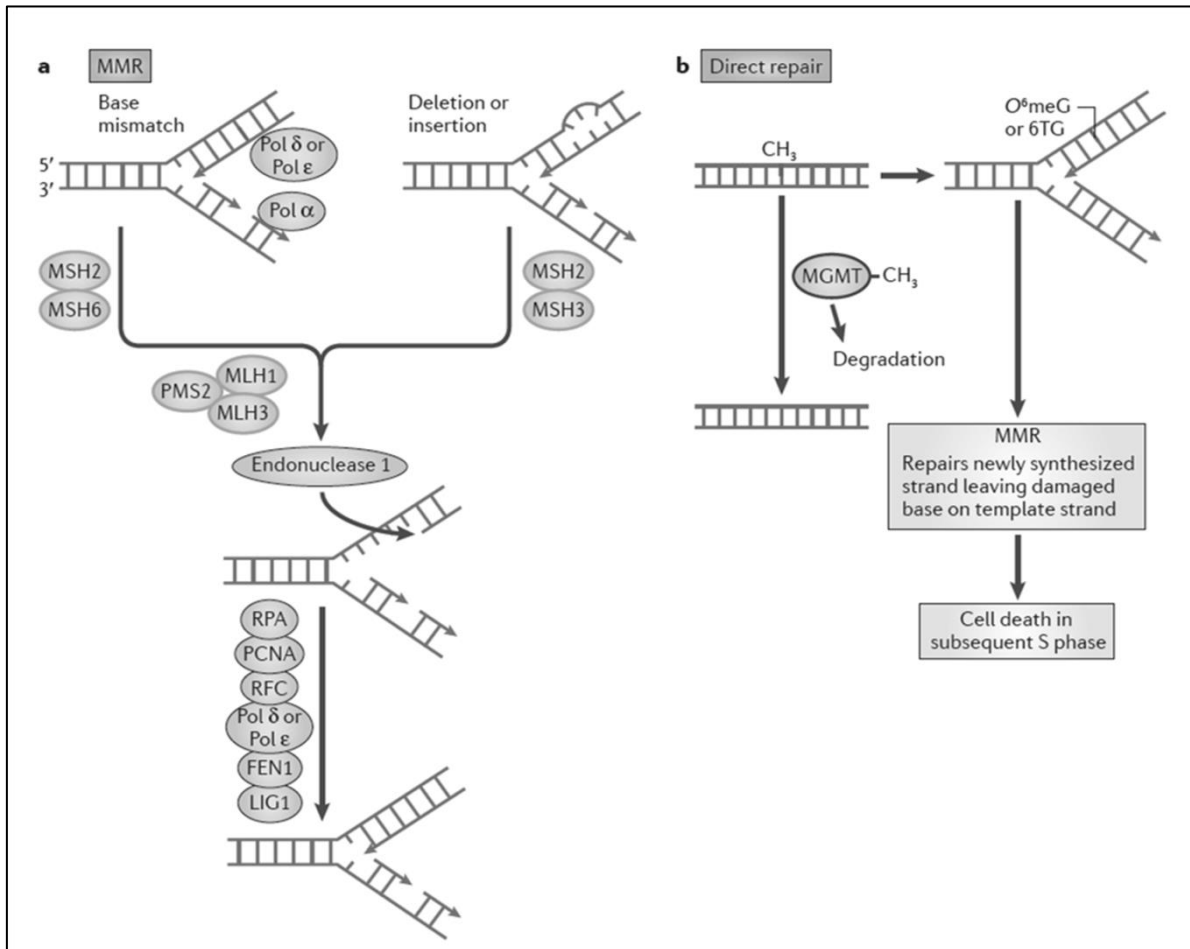
Figure 1-5. Mismatch Repair System. a - The repair is triggered by MSH2-MSH6 or MSH2-MSH3 complexes at a mismatched nucleotide or an indel, respectively. Thereafter, the repair continues interacting with a protein cluster of MultLalpha and MultLbeta (MLH1, PMS2 and MLH3 complex) and Endonuclease 1. The process then continues with PRA, PCNA, RFC, POLD, POLE and FEN1 to excise and resynthesize the affected regions. MMR is a strand-specific repair mechanism and therefore, is crucial for replication error repairs. b - When O^6meG and 6TG cause DSBs by mistakenly repairing the normal strand, apoptosis is initiated by MMR signalling to the ATR-CHK1 complex, LIG1 and MGMT(69). The figure is adapted with permission from Curtin at al. (2012) (59). Abbreviations; RPA, replication protein A; PCNA, proliferating cell nuclear antigen; RFC, replication factor C; Pol D, DNA polymerase-D; FEN1, flap endonuclease 1; O^6meG, O^6-methylguanine; 6TG, 6-thioguanine; ATR, ataxia-telangiectasia and Rad3-related; LIG1, DNA ligase 1; and MGMT, O^6-methylguanine DNA alkyl transferase.
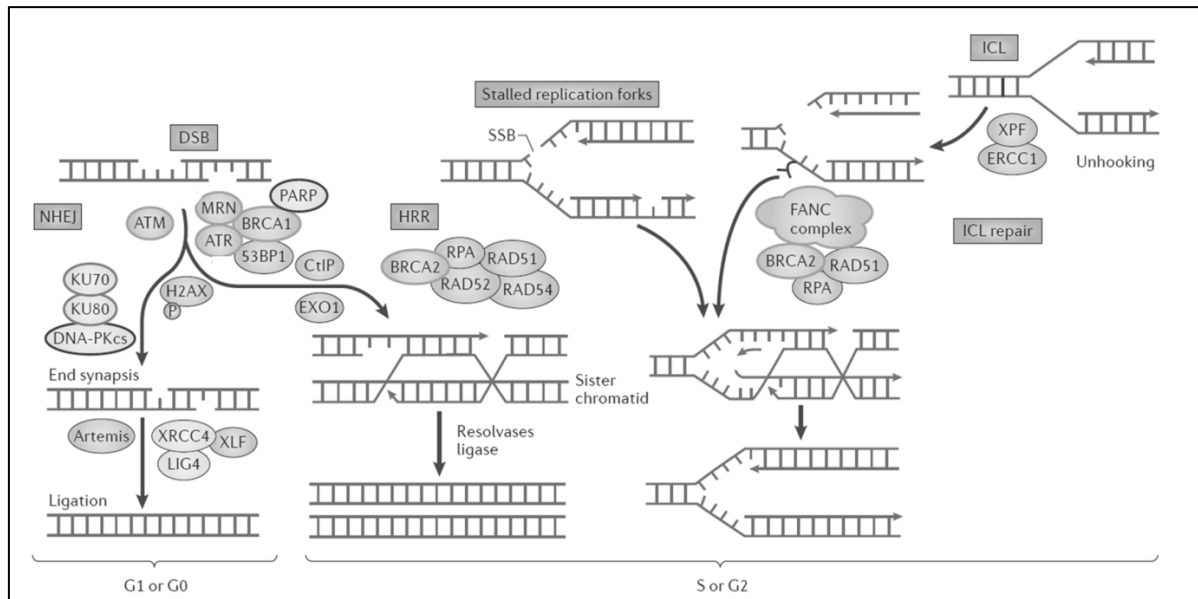
Figure 1-6. DNA Double-Stand Breaks (DSBs) by Non-Homologous End-Joining Repair (NHEJ) and Homologous Recombination Repair (HRR). At DSB sites, many DNA repair proteins are gathered, including MRN nuclease complex (consisting of MRE11, RAD50 and NBS1). In addition, KU70-KU80 complex, histone H2AX and DNA-PKcs are recruited to pair the lesion ends to the normal strand (70). Ligation is accomplished by involving Artemis, LIG4 and XLF (67). In HRR, the repair work of MRN proteins is facilitated by BRCA, PARP1, CTLp and EXO1. These proteins are joined together to resect the severed ends (71, 72). Next, MRN recruits and activates ATM, which subsequently activates MRE11, NBS1, EXO1 and CTLp. ATM also phosphorylates H2AX, which aids in recruiting 53BP1 and BRCA1 (73). The RPA protein is then attached to the overhanging strand to inhibit degradation and the ATR-ATRIP complex is recruited to phosphorylate CHK1 for S and G_2 cell cycle arrest (not presented). Subsequently, ATM and ATR recruit BRCA1 to activate ligation via E3 ubiquitin ligases. In addition, ATR activates RPA2 and CHK1, which phosphorylate RAD51. RAD51 then displaces RPA to create a nucleoprotein filament and a Holliday junction by targeting the complementary DNA duplex (74, 75). Thereafter, the lesion site in the targeted strand is extended and re-joined by the action of DNA polymerase and therefore, the polymerase forms crossover and non-crossover repairs. During stalled replication events, ATR is activated rather than ATM (76), followed by cell-cycle signalling through the ATR-CHK1 complex, RPA, BRCA1, FANCN and BRCA2, which are crucial in HRR. The figure is adapted with permission from Curtin at al. (2012) (59). Abbreviations; NBS1, Nijmegen breakage syndrome 1; DNA-PKcs, DNA-dependent protein kinase catalytic subunit; LIG4, DNA ligase 4; XLF, XRCC4-like factor; PARP1, poly (ADP) ribose polymerase 1; CTLp, CtBP-interacting protein; EXO1, exonuclease 1; ATM, ataxia-telangiectasia mutated; 53BP1, p53 binding protein 1; RPA, replication protein A; ATR, ataxia-telangiectasia and Rad3-related; ATRIP, ATR-interacting protein; FANC, Fanconi anaemia; and ICLs, interstrand crosslinks.

1.8.5. Homologous Recombination Repair (HRR)

HRR is a complex pathway that involves many proteins and acts against DSBs during the S and G_2 cell-cycle phases (66). The severed ends of a DSB are resected, followed by introducing the single strands into the sister chromatid to accurately initiate the resynthesis of the broken DNA (Figure 1-4). The two chromatids are attached together by a cohesion complex prior to mitosis, which makes them accessible during the repair process (66). HRR is a high-fidelity repair mechanism and therefore considered the most crucial repair process for DSBs. It works on stalled replication forks, single-ended DSBs and ICLs (in conjunction with NER and Fanconi anaemia pathways) (77, 78).

1.8.6. Cell-Cycle Checkpoints

DNA repair processes are integrated with checkpoint signalling and cell-cycle control to reduce the risk of permanent DNA damage that could become fixed during replication and mitosis. DSBs, for example, are repaired by either EJ or HR pathways. The former is active during interphase but is inhibited during mitosis, whereas HR is restricted to the S and G2 phases of the cell cycle when suitable templates in the form of sister chromatids are available. ATM and ATR, members of Phosphatidylinositol 3-kinase-related kinases (PIKK) family, have vital roles coordinating DNA damage to the DNA repair pathways and cell-cycle checkpoints (79, 80). Hundreds of target proteins are phosphorylated in an ATM- or ATR-dependent manner. These include other protein kinases, in particular, Chk1, Chk2 and MK2, which are activated and cause a second wave of phosphorylation. DSBs activate ATM whereas DSBs and a variety of other types DNA damage activate ATR.

P53 protein accumulation is essential for complete G1 arrest and is activated by ATM through phosphorylation of Ser 15 (81). However, in the absence of ATM, ATR is able to compensate. The nature of DNA damage determines the Ser 15 phosphorylation dependency with IR, primarily acting through ATM and ATR when other forms of DNA damage or genotoxic stress predominate (81). Accumulation of p53 protein is regulated by indirect ATM pathways, including the phosphorylation of Chk2, which in turn phosphorylates Ser 20 on p53, facilitated by casein kinase I phosphorylation of Ser 18. ATM also phosphorylates Ser 395 of the p53 antagonist, protein MDM2 (82). ATR additionally reinforces p53 activation by phosphorylating Chk1, which also acts on Ser 20 of p53 (81).

Unlike the G1 to S transition, S phase is not completely blocked by DNA damage but slowed. In response to IR, the ATM-dependent pathway functions as an S-phase inhibitor. Cdc25A is a protein tyrosine kinase, and G1 to S phase transition requires Cdc25A to activate cyclin A/cdk2 complexes. However, ATM phosphorylation of Chk2 allows the latter to target Cdc25A for ubiquitin-dependent proteasomal degradation by phosphorylation at its Ser 123 (83), slowing S phase as a result. ATM initiates HRR by recruiting BRCA2 to the DSBs, and it also activates BRCA1 and initiates NHEJ by recruiting p53-binding protein 1 (53BP1). Therefore, ATM regulates the cell cycle through its antagonistic functions (84).

ATR and Chk1 are required for S phase slowing when other forms of genotoxic stress affect DNA, in particular when replication forks are affected or stalled (85). The damage is signalled to ATR through proteins, including Rad1 and 9 as well as Hus1 that forms a trimeric complex on DNA at the site of damage where failed replication leads to the

accumulation of ssDNA and is left unchecked, specifically SSBs and DSBs (86).

Phosphorylation leads to the recruitment of additional proteins, HRR and ICL

intermediaries, such as BRCA2, RAD51, Fanconi anaemia group E (FANCDE) and FANCD2.

Cell-cycle consequences following ATM/ATR stimulation by NER junctions at resected

DSBs and at stalled replication forks is depicted in Figure Figure 1-7(76).

G2 to mitosis checkpoints are relatively insensitive compared to the G1 and S phase

checkpoints (87). Similar to G1, the G2/M checkpoint is initially activated by ATM but ATR

leads to a sustained block. ATM/ATR activates Chk1/2, which phosphorylates CDC25.

Phosphorylated CDC25 is sequestered in the cytoplasm and cyclinB-CDK1 is kept inactive

by Wee1, preventing entry into mitosis (88). Other kinases, in particular, the polo-like

kinases, are also regulatory (89). Maintenance of the G2 checkpoint is p53 dependent,

this being responsible for the transcription of cell-cycle inhibitors, including p21[Cip1],

GADD45 and 14-3-3sigma proteins, which act co-operatively. GADD45 dissociates CDK1

from cyclin B, 14-3-3sigma sequesters CDK1 to the cytoplasm and p21[Cip1] binds cyclinB-

CDK1 (90). Through p21-dependent suppression of CDK1 by p53, the pRB tumour

suppressor is activated, sequestering E2Fs and reducing transcription of G2/M target

genes required for cell-cycle progression, in particular, the anaphase-promoting complex

genes, APC/C, to further cause G2/M arrest (91).
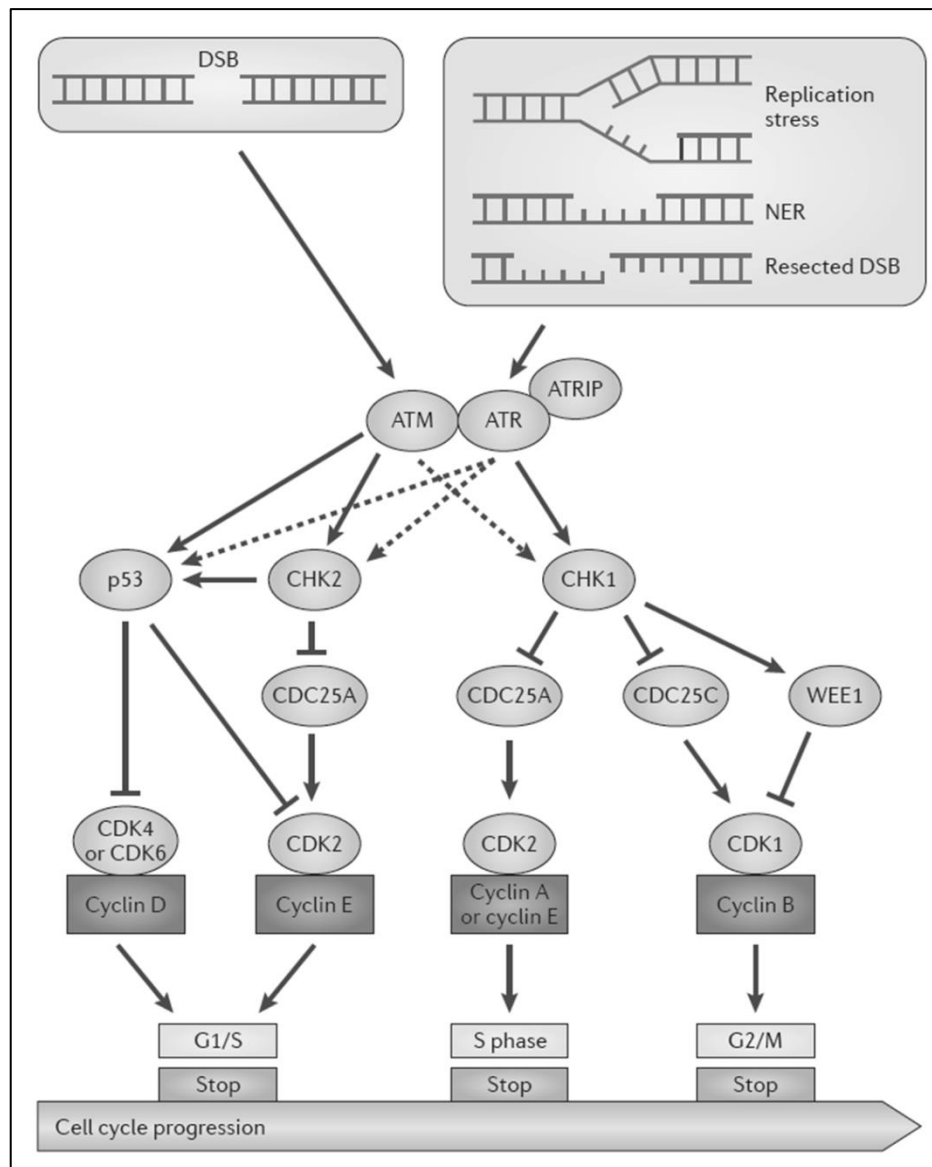
Figure 1-7. Cell-Cycle Checkpoints and DNA Damage Signalling. When DSBs arise, ATM is activated, triggering the G_1 checkpoint, resulting in activation of CHK2 and p53. In a similar manner, ATR is activated by defective single- and double-stranded DNA events, such as stalled replication forks, resected DSBs and NER intermediates. By activating ATR, subsequent phosphorylation occurs on CHK1, WEE and CD25C; these proteins work as S phase and G_2 checkpoints, which suppress cell-cycle progression by deactivating CDK activity (79). Generally, there is coordination between ATM and ATR to facilitate substrate sharing. Dashed arrows indicate other targets. The figure is adapted with permission from Curtin et al. (2012) (59). Abbreviations; ATM, Ataxia-telangiectasia mutated; DSBs, DNA double-strand breaks; ATR, Ataxia-telangiectasia and Rad3-related; NER, nucleotide excision repair; CDC25, cell division cycle 25; CDK, cyclin-dependent kinase; CHK Checkpoint Kinase: and ATRIP, ATR-interacting protein.

1.9. P53-Dependent Apoptosis

Multiple stress signals and all DDR pathways are integrated by p53, which in turn regulates a wide variety of antiproliferative responses that include cell-cycle arrest and apoptosis (92). Cell fate is determined by a balance of pro- and anti-survival signals from a wide variety of networks. There are two apoptotic pathways - the intrinsic or mitochondrial and the extrinsic or death receptor pathway. There is also a perforin/granzyme pathway that mediates T-cell cytotoxicity (93). p53 initiates apoptosis by transcribing pro-apoptotic proteins, including BAX, BID, PUMA and NOXA, which permeabilise the mitochondrial membrane and release additional pro-apoptotic factors. Although it is clear that ATM and downstream effectors like Chk1/2 are important for phosphorylating p53, leading to its activation by stabilisation, it is not clear precisely how p53 is central to apoptosis on a cell-by-cell basis (94). DNA repair proteins like Brca1 can also induce apoptosis by signalling to p53 (94). Shortly after induction of p53, apoptosis can be avoided. PARP localises to DNA damage and also signals to other repair proteins by synthesising poly(ADP-ribose) (PAR) chains. A rapid, transient parylation of nuclear proteins followed by the caspase-3 destruction of PARP is required for apoptosis to take place (95). Inactivation of PARP ensures adequate levels of NAD and ATP, which are required for later stages of apoptosis (96). Clearly, this represents only a few of the many fine-tuned factors that contribute to ultimate cell fate. P53/ATM are highly important for apoptosis of CLL.

1.9.1. TP53 Structure and Function

The human TP53 gene is located on the short arm of chromosome 17 (17p13.1) and has a length of 20 kb. It is composed of 10 coding exons and one non-coding exon, the first exon, 8-10 kb distance from exon 2 (97-99). TP53 encodes transcription factor p53 protein, which is composed of 393 amino acids. p53 protein has several functional domains: two transcription activation domains (1: amino acids 1-42) (100); (2: amino acids 43-63) (101, 102), a proline-rich domain (spanning amino acids 64-91) (103), a DNA-binding domain (within amino acids 100-300) (104), a nuclear localization signal (amino acids 316-325) (105), a tetramerisation domain (106), a nuclear export signal domain (107) (spanning residues 334-356) and a C-terminal basic domain (amino acids 364-393) (108). Primarily, four domains are essential for p53 function and protein-protein interactions: the transcriptional activation domain 1, DNA-binding domain, the tetramerisation domain and the C-terminal basic domain (109).

p53 essential protein domains affect DNA binding by p53 and thus, the activity of the p53 protein (108). The C-terminal domain is regulatory and controls p53 DNA-binding activity in response to p53 phosphorylation (108) and acetylation (110), which are all affected by its deletion (111). Association of anti-p53 antibodies (112) or peptides are able to enhance DNA-binding activity from the domain (113). The C-terminus of p53 is indispensable for transactivation or transrepression of the protein (114). In a normal physiological state, trace amounts of p53 protein are present because of its short half-life. Activation and stabilisation of the protein by upstream kinases, e.g., ATM- and ATR-induced conformational changes, and release from negative controls like those mediated

by its ubiquitin ligase MDM2, which targets it for proteasomal destruction and therefore p53 half-life is risen (109).

1.10. DNA Maintenance Gene Defects, Genomic Instability and Human Cancers

DNA maintenance or caretaker genes are responsible for the maintenance of genomic stability (115). Essentially, they include DNA repair and cell-cycle checkpoint genes, termed "classical caretaker genes" as described in Sections 1.7 DNA Repair Pathways) to 1.8.6 Cell-Cycle Checkpoints (116). Collectively, these act to detect DNA damage, provide an opportunity to influence its repair, complete the repair or program the cell for death if sufficient deviation from normality is reached. Defects in these genes, whether acquired or inherited, lead to genomic instability and the type of instability is characteristic of the nature of the "caretaker" gene affected. Genomic instability is defined as a cancer signature that is characterised by accumulation of DNA mutations and other lesions within cancer cells. Generally, there are different types of genomic instability. For example, chromosomal instability (CIN) is the most frequent form that is caused by high CNAs at the chromosome level. Another example is microsatellite instability (MSI), which increases the likelihood of lesions at genomic tandem repeats (117). However, all types of genomic instability are observed, ranging from single nucleotide changes to macroscopic alterations in DNA copy number and organisation (118). Defects in the caretaker genes can be inherited or acquired.

Many different kind of hereditary cancers are known. Hereditary non-polyposis colon cancer (HNPCC) is linked to mutations in DNA mismatch repair genes, leading to a MSI

phenotype (119, 120). Another hereditary neoplasm affecting the fidelity of single nucleotides is MYH polyposis, which causes an increase in the ratio of G-C to T-A transversions and results from biallelic germline mutations in the MYH gene. In hereditary cancers that carry the CIN genotype, genomic instability can be directly affected by mutations in NHR or NHEJ DNA repair genes; examples include HR mutations in the breast cancer susceptibility 1 and 2 (*BRCA1* and *BRCA2)* genes*, partner and localizer of BRCA2 (*PALB2)*, BRCA 1-interacting protein 1 (*BRIP1*), Werner syndrome helicase (*WRN*), Nijmegen breakage syndrome protein 1 (*NBS1* or *NBN*), Bloom syndrome helicase (*BLM*), RecQ protein-like 4 (RECQL4) and other genes belonging to the Fanconi anaemia pathway. These genes are associated with either DSBs or DNA interstrand linking that exacerbates the development of cancers, such as leukaemias and lymphomas (121, 122). Low levels of DNA telomerase are also associated with cancer development as the ends of chromosomes cannot be maintained, leading to exposed double-stranded ends that may undergo multiple rounds of breakage along with end re-joining cycles, accumulating damage and chromosomal rearrangements as they proceed (123). An analysis of nearly five million mutational events from over 7000 different cancers found at least 20 different types of mutational signatures, some of which are found in most cancer types and others in only a single type(124). The APOBEC family of cytidine deaminases are the most prevalent type and the origin of certain forms of damage have yet to be discovered. Hypermutation, localised to small, discrete genomic regions, was also found to occur. DNA strand bias was an additional feature, illustrating the importance of transcription for the repair process. Signatures of mutagens, for example, bulky adducts as found in tobacco smoke, were also prevalent.

Many of the genes that are involved in hereditary forms of cancer are also known to be altered by spontaneous mutations or DNA methylation in sporadic cancers. Sporadic cancers can be broadly classified into two types, those with predominantly multiple, single-base alterations and those with CIN (118). DNA maintenance, cell-cycle and chromosome organisation pathways are associated with each. NGS (See Section 1.14.3 Massively Parallel (Next-Generation) Sequencing, page 69) and other post-genomic technologies have allowed detailed analysis of cancer genomes and associated macromolecules to be performed. A comparison of over 3000 different cancers from 12 different, common types showed a mutation (M) class and a copy number (C) class (125). The M class was associated with defects in control of G1, the TP53 pathway and DNA DSB repair whereas the C class was associated with enhanced RAS signalling, Wnt signalling, PI3K/PTEN/AKT signalling, DNA mismatch repair and gene defects for the chromosomal organisation (SWI/SNF) (125).

The presence of germline mutations in maintenance genes supports the mutator hypothesis, which states that genomic instability occurs in precancerous lesions and cancers are able to arise because of their increased mutation rate. However, in sporadic neoplasms, genomic instability is rarely exhibited despite the presence of CIN in the majority of cancers (126). Two potential models for CIN in sporadic cancers are speculated - the first is the mutator hypothesis and the second is the replication stress (see Section 1.8.1 Nucleotide Excision Repair (NER), page 44) model for cancer development (50, 51, 127). CIN is initiated by oncogene-induced DNA replication stress, which is DNA damage that usually disrupts replication machinery, and is followed by DNA DSBs and genomic instability. Replication stress has been shown to accumulate in

precancerous lesions and tumours when compared to normal cells, supporting a role in contributing to genomic instability (127). It has been estimated that an approximately 100-fold increase in 8-oxoG is observed in cancer relative to normal tissues as a result of cancer-induced inflammation, which generates more ROS. (128). This could only become significant once cancer has developed, making it less likely that it is the initial cause of genomic instability.

DSBs are part of the essential mechanism to achieve antibody diversity (See Section 1.3

Normal B-Cell Maturation, page 32) and provide an opportunity for genotoxic damage, which could be considered to contribute significantly to the development of haematological malignancies, including CLL. CLL shares a mutational signature with other malignant B-cell lymphomas, essentially T>G transversions at ApTpN and TpTpN trinucleotides, and this signature is restricted to cancers that have undergone somatic immunoglobulin gene hypermutation (IGHV-mutated) associated with cytidine deaminase (AID) activation (124). The signature, however, lacks the known mutational features of AID and has therefore been proposed to be because of an error-prone polymerase (polymerase η) involved in processing AID-induced cytidine deamination (129).

The existence of characteristic mutational signatures and their association with underlying defects in pathways for cell-cycle control or DNA maintenance has supplied an opportunity for targeted therapies.
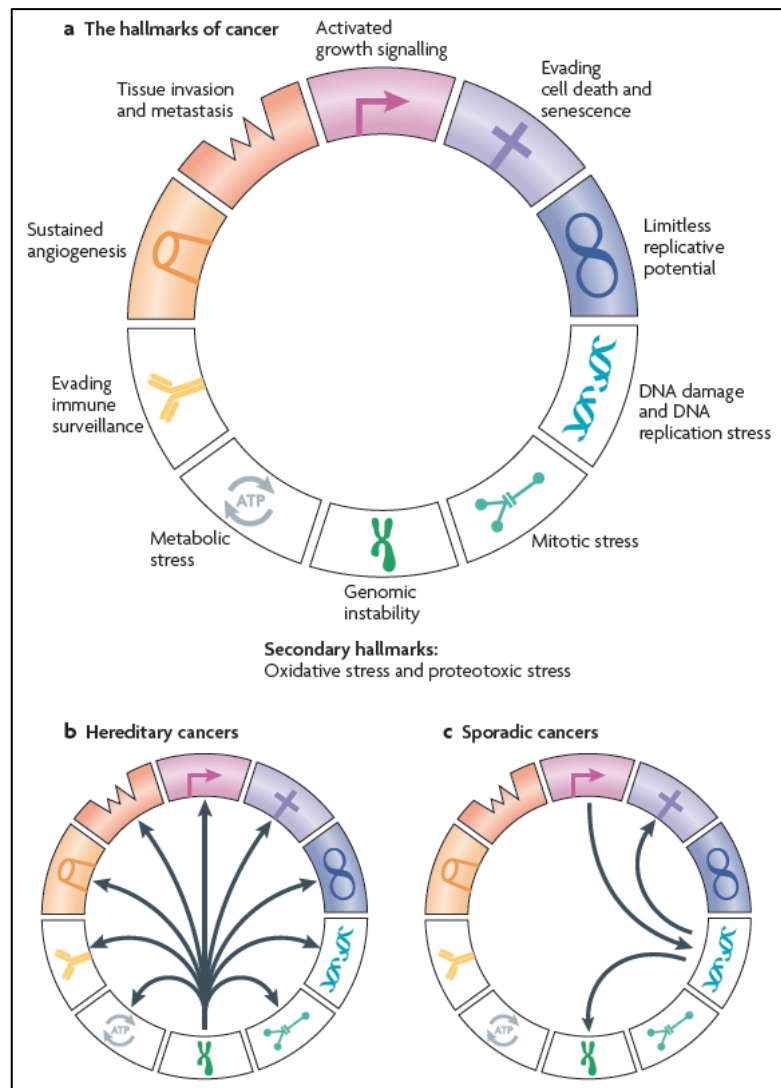
Figure 1-8. Genomic Instability as a Cancer Hallmark. a - Hallmarks of cancer include genomic instability. In this figure, activated growth-signalling hallmarks include self-sufficiency and insensitivity to anti-growth signals. The secondary hallmarks, such as oxidative stress and proteotoxic stress, are separate from the cancer hallmarks circle. b - In hereditary cancers, genomic instability is probably the initiation event, which later facilitates the initiation of other hallmarks. c - The temporal order of hallmark establishment in sporadic cancers. It is initiated by deregulation of growth-regulating genes. This causes DNA damage and DNA replication stress followed by genomic instability and the selective pressure of p53 inactivation and its tumour suppressor role. Subsequently, cancer cells become capable of evading cell death. Based on genomic instability, other hallmarks can be activated as seen in the hereditary cancers model. Figure is taken with permission from Negrini et al. (2010) (130, 131).

1.10.1. Therapeutic Targeting of Genomic Instability

In cancer, if one DNA repair pathway is deactivated, the other pathway(s) may work in favour of cancer survival. An example of this mechanism that may compromise chemotherapy efficacy is the up-regulation of DNA repair pathways which render cancer chemo-refractory. However, these pathways can also be targeted by inhibitors as a potential cancer therapy (59).

Germline defects in the Xeroderma pigmentosum genes are associated with hereditary skin cancer development and UV sensitivity (62). NER is defective and therefore platinum therapy is a worthwhile option because the ICL repair capacity is already reduced (132, 133).

MIN associated with MMR defects can increase mutation rates up to 1000-fold (134). Several DDR genes contain microsatellite sequences, for example ataxia-telangiectasia (ATM) and MER11, and therefore are susceptible to damage in high-MSI cancers. It is not uncommon for MMR defects to be associated with sensitivity to some DNA-damaging agents (135, 136) and resistance to others, such as temozolomide (TMZ), platinum agents and certain nucleotide analogues (69, 137).

Germline aberrations in the associated NHEJ repair proteins are found in cancers such as breast cancer (with KU70 SNPs) (138) and glioma (with SNPs in DNA-PKcs) (139). Many anti-tumour agents target NHEJ.A portion of them initiate DSBs directly by IR and topoisomerase II inhibitors, and indirectly via stalled replication forks induced by single-strand lesions. Topoisomerase II inhibitors produce a persistent DSB-associated protein (140), whereas IR induces an average of one DSB per 25 SSBs.

Other agents have been developed to target DNA-PKcs and inhibit PI3K as it is a member of the PIKK enzyme family, which includes ATM, ataxia-telangiectasia and ATR as well as mTOR. The agents, LY294002 and wortmannin, affect the strand-rejoining repair for DSB and have adjuvant cytotoxicity with DSB-inducing agents (141, 142). Higher levels of DNA-PKcs are seen in CLL cells, suggestive of poor prognosis and the DNA-PKcs inhibitor NU7026 has been shown using *in vitro* models to increase the sensitivity of CLL cells to topoisomerase II inhibitors (143).

Many of the proteins for HRR are tumour suppressors, including BRCA1, BRCA2 and ATM. Cancers with defective HRR are very sensitive to crosslinking agents, such as cisplatin and carboplatin, as well as therapeutically-induced DSBs for example by IR and topoisomerase I poison. Furthermore, tumours with high rates of HRR defects may potentiate the efficacy of cytotoxic therapy, hence providing a rationale to combine HRR inhibitors with conventional chemotherapy to sensitize tumours (for tumours with functional HRR); known by the term synthetic lethality. In Philadelphia-positive (Ph[+]) haematological malignancies, the proto-oncogene, *ABL1*, is important for activating phosphorylation of RAD51 as an important step in HRR. Imatinib, the BCR-ABL1 TK inhibitor, targets Ph[+] tumours and consequently sensitizes the cells to IR and DNA crosslinking agents (144, 145).

ATM is also a tumour suppressor. ATM-associated defects include mutations, polymorphisms, deletion and epigenetic silencing (10). Microsatellite repeats are a common frameshift indel in *ATR*, resulting in truncation of the protein in cancers with MSI (146). During the DDR response to DNA-damaging agents, checkpoint activation is a common feature where ATM and ATR have multiple protein targets (see section 1.8.6Cell-

Cycle Checkpoints, page 50 (Figure 1-7). Therefore, checkpoint inhibitors have been used to enhance the efficacy of many DNA-damaging agents (80, 147). An example of these is KU55933, which inhibits ATM activation during IR and sensitises tumour cells to topoisomerase inhibitors and IR (148).

1.11. TP53 Defects in CLL

Allelic loss and/or mutation of TP53 can occur in a broad range of malignancies; about half of human cancers are affected by TP53 somatic mutations and regarded as a cornerstone in tumourigenesis (149, 150). The frequency of TP53 mutations varies between different cancers - they account for approximately 10% of haematological malignancies (151). Yet, in other human neoplasms, such as colorectal (152), ovarian (153) and head and neck cancers, such an incidence is as high as 50-70% (154). During tumour progression, loss of heterozygosity (LOH) usually occurs after somatic mutation (155). This suggests that there is a selective pressure by Loss of Heterozygosity (LOH) to inactivate the wild-type allele (156, 157).

Although hot spots have been documented, mutations can occur anywhere within the coding region of TP53. Missense mutations are the most common form of the mutations in this gene; the resulting amino acid change hampers either DNA-binding ability or the conformational structure of p53, nevertheless the full-length protein can be formed (158, 159). In most cases, the mutations in one allele are associated with the loss of the wild-type allele, resulting in full loss of wild-type p53's functions (159), (160). In addition, mutated p53 may gain oncogenic properties (161).

In CLL, roughly 90% of mutations are localised to the DNA-binding domain in exons 5-8 of this gene (162, 163). The mutations can be found in 5-10% of therapy-naïve patients (164), while an incidence rate as high as 25% has been reported among chemotherapy-resistant patients (165, 166). With regards to missense mutations, transition changes are common, particularly at known CLL hotspots, including codons 175, 179, 220, 248, 273, 281 and 209 (162). Despite being accompanied by unmutated *IGHV*, TP53 mutation and/or deletion is independently strongly associated with disease severity and outcome in CLL (167). It has also been shown that positions of the p53 mutation are significant. Mutations in the DNA binding motif are associated with an extremely poor prognosis (167). At the chromosomal level, deletion of TP53 from 17p takes place in 7% of untreated patients (168). The percentage rises to 50% for refractory patients (169) and 80% in cases with inactivated p53 mutations(166). Cells acquiring this deletion have inhibited p53 tumour suppressor function and therefore, resistance to therapy-induced apoptosis and DNA-damaging agents (109). Notably, both deletions and mutations have the same biological and clinical impact (162), although TP53 deletions alone can be associated with prognostic features, including accelerated disease progression and chemotherapy resistance (170).

Heterogeneities in clinical course and response to treatment for CLL are linked to endogenous factors, including chromosomal instabilities, nucleotide mutations and subsequent clonal accumulation of these aberrations (171). To date, the most important biomarker for predicting prognosis and choice of therapy is the existence of TP53 mutations and/or chromosome 17p deletions (172). Accordingly, when DNA damaging agents are not suitable because of inactivated p53, alternative p53-independent

treatments are offered, such as monoclonal antibodies that target the antigenic surfaces of cancer cells (19) (see Section 1.5.3 Second-Line Treatment, page 38).

1.12. Clonal Evolution in CLL

It has been noted that CLL patients with faster clonal evolution are likely to have an aggressive clinical course. This is consistent with the concept that CLL subclones are expanding in such patients. In addition, this supports the notion that when CLL symptoms appear and absolute lymphocyte counts rise, distinct chromosomal aberrations are detectable in certain subclones (173). This was detected in the past using conventional FISH and other cytogenetic techniques of low resolution; for an example, in a study by Stilgenbauer et al. (2007) where clonal evolution was detected in roughly 43% of CLL patients, FISH was the main methodology and from progression to relapse periods, they were able to detect the appearance of biomarkers for a poor outcome, such as 17p and 11q deletions (174).

It is speculated that during CLL clonal evolution, various clones/subclones more rapidly multiply than others, and these cells likely bear the genomic DNA anomalies associated with aggressive prognostic features. Therefore, in order to inhibit the progressive evolution of CLL, a therapeutic plan to target these clones is desirable. In this regard, two main models are proposed: a linear evolution where the mutant clone increases in number with no alteration in the initial genetic composition, or branching evolution where complex interactions alter the initial genetic composition of different progeny (175).

As the most rapidly expanding leukemic subclone, from which permanent DNA abnormalities arise could lead to more aggressive disease, preferentially targeting these subclones might prevent successive clonal evolution and disease deterioration.

1.13. Methods for Detecting Genomic Instability

1.13.1. Cytogenetic Approaches to CLL Prognostication

In CLL, cytogenetics is highly relevant for defining prognosis (176). Giemsa banding (or G-banding) produces visible Giemsa-stained metaphase karyotypes from metaphase chromosomes. Using photographic representations, large-scale chromosomal defects can be detected throughout human chromosomes (177). Unlike other haematologic malignancies, few CLL studies have been conducted using chromosome-banding analysis, largely owing to lower proliferation activity of CLL cells in *in-vitro* culture, even with the availability of B-cell mitogens. (178, 179). Alternatively, FISH and comparative genomic hybridization (CGH) are cytogenetic approaches that do not require proliferating cells. FISH was developed in the 1980s. The approach employs fluorescent nucleic acid probes to detect complementary DNA or RNA sequences of interest without requiring metaphase proliferation. Therefore, it is more sensitive and specific compared to chromosome banding analysis and other conventional cytogenetic approaches. FISH can employ multi-colour probes and is useful in identifying different prognostic subgroups in haematological cancers (180). Interphase FISH analysis has been successfully applied for detecting recurrent chromosomal indels in CLL. It detects more than 80% of common CLL genomic aberrations, such as 17p and 11q deletions (which are associated with the worst prognosis), trisomy 12 (associated with intermediate poor prognosis) and 13q (associated

with good prognosis and favourable outcomes) (181). Thus, it is an important tool for distinguishing high-risk patients. However, the number and size of chromosome regions recognised by FISH probes are limited, and quantifying results sometimes are challenging. Generally, FISH can detect up to >5% of large chromosomal aberrations (182).

Another approach uses SNP arrays or array-comparative genome hybridisation (aCGH), where labelled genomic sample DNA is hybridised to arrays of oligonucleotide probes representing the entire genome and the signal strength from each data point is used to infer the relative copy number of the genome at that point in the genome of the sample (183). It can, therefore, measure the copy number changes throughout the whole genome and it improves in terms of resolution and coverage over the limitations of FISH. For example, Xu *et al.* observed that 50% of patients who were found to have a normal karyotype using FISH were found to have subclonal genetic legions associated with poor prognosis by SNP arrays, below the limit of detection for FISH (184). SNP arrays are preferred over aCGH because they can determine other types of copy number anomalies, including double deletions, uniparental disomy (UPD) and additional complex anomalies that aCGH cannot. It has been seen that having multiple lesions is a sign of clonal progression and is associated with reduced median overall survival of 22 months (185).

1.14. Identification of Single-Nucleotide Variants (SNVs) and Small Indels

1.14.1. Detection of Aberrations Using Physical Properties of DNA

Several methods have been developed to detect small aberrations, such as point mutations, small insertions and deletions. Choosing the proper method depends on the aberration status (known or unknown) and number and length of aberrations. Single-

strand conformational polymorphism (SSCP) is one method utilised to identify unknown point mutations that makes use of DNA variations resulting in a change to the conformational structure of denatured DNA fragments (of up to 200 bp in length) that can be detected when the mobility of normal and mutated DNA are compared by gel electrophoresis (186). Similarly, differential mobility in denaturing high-performance liquid chromatography (DHPLC) of heteroduplexes formed by mutant DNA annealed to wild-type DNA can establish point mutations compared to homoduplexes of wild-type alleles (187). Denaturing gradient gel electrophoresis (DGGE) makes use of the melting behaviour of mutant DNA for the detection of mutant alleles. This technique can detect aberrations in DNA fragments up to 1 kbp, with 95% sensitivity, however it is labour intensive and only suitable for known target abnormalities (187). None of the methods described so far can detect exact changes in the DNA sequence, which need to be confirmed by other methods, e.g., Sanger sequencing or NGS.

1.14.2. Sanger Sequencing

Sanger sequencing was developed by Frederick Sanger 40 years ago. It is a DNA polymerase-dependant approach that involves complementary DNA synthesis using deoxynucleotides (dNTPs) and unnatural dideoxynucleotide (ddNTP) terminators (188). One terminator is available for each possible base and the position of their incorporation is determined from the length of the resultant fragment polymerised. Knowing all of the termination positions for each of the possible bases allows the overall sequence to be inferred. Laser detection of fluorescent dye terminators (189, 190) and on fragments separated by capillary electrophoresis (191, 192) has allowed the process to be automated (193, 194). The accuracy of each base is calculated by a Phred score after the

corresponding base-calling algorithm that interprets the shape and complexity of the electropherograms are produced for each position. The standard Phred quality of 20 relates to a 99% accuracy (195, 196). Small background peaks produce mixed fluorescent signals and therefore the method is not sufficiently sensitive to identify mutations with variant frequencies below 10 (197). Electrophoretic resolution and the requirement for PCR amplification to prepare the sequencing template limits target size, to less than 900 bp, and it is laborious for large target regions.

1.14.3. Massively Parallel (Next-Generation) Sequencing

High-throughput sequencing, or NGS, was developed to increase throughput and decrease costs. Currently, a single sample or multiple samples are sequenced in parallel with the capability for millions of nucleotides reads. In addition, in comparison to Sanger sequencing, NGS can produce high base depth, which can detect somatic variants with as low as 0.1% allele frequency (198). Modifications to the approach allow variants as low as 1 in ~1000000 to be discerned with confidence but for specific sites of interest (199).

Various NGS platforms are commercially available, including large-scale machines, such as Illumina HiSeq and Illumina HiSeq x10, which can generate gigabases (Gb) and terabases of data per run, allowing for the whole human genome to be sequenced over several days for multiple individuals (200). Lower throughput platforms, such as the Roche 454, Illumina Miseq, Ion Torrent PGM and Ion Proton, have a lower capacity but faster turn-around times, therefore are more suited to specific questions, especially involving clinical samples and tests. All of the platforms are capable of producing megabases to gigabases of output at a very low cost compared to the Sanger method (201). Table 1-1 is a

summary of the different NGS platforms. The sequencing workflow has multiple stages, including DNA extraction, target DNA enrichment when required, template preparation, sequencing reactions, raw data production and bioinformatics analysis. Bioinformatics analysis is time-consuming and challenging as a large amount of data has multiple steps and needs considerable computing power for comprehensive analysis (201). All of the platforms yield short reads of data of up to several hundred bases per template. This is adequate for *de novo* sequencing. The reads are aligned to a reference genome and differences between the reads and the reference identify variants. Error rates are also high and therefore target sequences have to be read to high depth (many times) to have high confidence in variant calls (202). The quality of the reference genome for comparison is also important and is being constantly refined, in particular with regard to polymorphic variants of all types from single nucleotides to large structural alterations and copy number differences (203).

1.14.4. Whole-Genome Shotgun Sequencing (WGS)

One haploid copy of the human genome comprises approximately 3.1 billion nucleotides, but millions of these are polymorphic in the population and any given individual may have presented only a few percent of the total polymorphisms known (204). Sequencing random fragments from the genome and then comparing the reads to the reference genome has a variety of uses, depending on the depth of coverage. Low pass coverage of up to 10 fold is adequate to detect structural variants (205). Yet, only 10-fold depth of coverage is required to identify single-nucleotide variants, the random distribution of reads across the genome means that many regions are underrepresented, so only a small fraction of the genome would offer useful information, and this would be different by

chance for each genome compared. Most whole-genome studies, therefore, perform

sequencing to a depth of 30-50-fold coverage (206). Alternatively, target enrichment is

performed so that the same total number of sequence reads can be focused on regions of

interest at higher depth for improved variant calling.

Table 1-1. Summary of high-throughput sequencing. Data are adapted from Goodwin *et al.* (2016) (207) (208).

Specificity	Roche 454 GS Series	SOLiD 5500 Series	Illumina HiSeq 2500 Series	Illumina MiSeq v3	Ion Torrent PGM	Ion Proton	PacBio RS II
Template Amplification	Emulsion PCR	Emulsion PCR	Bridge amplification	Bridge amplification	Emulsion PCR	Emulsion PCR	No PCR
Sequencing Chemistry	Pyrosequencing	Oligonucleotide ligation	Sequencing by synthesis	Sequencing by synthesis	Sequencing by synthesis	Sequencing by synthesis	Single molecule real- time sequencing
Read Length (bp)	400-1000	50-75 (SE)	36 (SE) 50-250 (PE)	36 (SE) 25-300 (PE)	200-400 (SE)	200 (SE)	20 Kb
Throughput	35- 700 Mb	80- 320 GB	9 -72 Gb (SE) 25- 500 Gb (PE)	540 - 610 Mb (SE) 0.75- 15 Gb (PE)	0.03- 2 Gb	Up to 10 Gb	0.5- 1Gb
Reads	0.1M-1 M	0.7 - 1.4 B	0.3 - 2 B (SE) 0.6 - 4 B (PE)	12-15 M (SE) 24- 50 M (PE)	0.4- 5.5 M	60- 80 M	0.06 M
Runtime	23 - 10 h	10 - 6 d	6 d- 7 h	56 - 4 h	3.7- 7.3 h	2- 4 h	4 h
Error Profile	1% indel	≤0.1%, AT bias	0.1% substitution	0.1%, Substitution	1% indel	1% indel	13% single pass, ≤1% circular consensus read, Indel
Cost per Gb (US$, approx.)	9.5- 40 K (US$)	70- 130 US$	30- 230 US$	110- 1000 US$	300- 800 US$	80 US$	1000 US$

Abbreviations: B, Billion; AT, Adenine and thymine; bp, base pair; Gb, Gigabase; d, day; h, hours; K, Kilo, M; Million.

1.14.5. Whole-Exome Sequencing (WES)

WES is a common type of sequencing that utilises target enrichment (209). This is because any non-synonymous variants established can easily be understood in terms of functional consequences (209). A human exome contains both coding and non-coding regions, totalling about 30 Mbp and 180,000 exons (roughly 1.1-1.4% of the human genome) (210). Another advantage is that ~85% of the known disease-associated, polymorphic variants lie within exomes (211). Usually, the average coverage depth achieved by WES is 100-160x.

1.14.6. Targeted (Deep) DNA Sequencing

Targeted DNA sequencing is similar to exome sequencing except that the studies are typically even more focused, for example, only on the genes that are responsible for a particular pathway or known, expected types of variation of interest. Targeted deep sequencing further overcomes coverage limitations for variant calling (212).

Multiple types of target enrichment approaches are in general use. They vary in sensitivity, specificity, reproducibility and coverage uniformity. Other factors should also be considered before establishing an enrichment system, such as DNA input, manual labour and cost (213).

1.15. Target Enrichment

Target enrichment methods fall into one of two types - direct amplification or hybrid-select/capture. Multiple PCR primers can be designed to allow the simultaneous amplification of multiple different targets from every single reaction in a Multiplex PCR

(214). This has the benefit of being straightforward but as increasing numbers of targets are present, differences in their efficiency result in a number of them becoming underrepresented. Emulsion or droplet PCR has been developed to overcome this limitation. Millions of independent PCR reaction vessels are created either by the direct creation of oil/water emulsions or mechanically using automated microfluidics (215). The former is quicker but the latter has the advantage such that it can be more reliably formatted for many different targets (216). However, in both cases, long amplicons are difficult to produce because of the small size of the droplets, therefore large amounts of input DNA are required (217).

In the hybrid-select/capture system, nucleic acid probes complementary to the target are produced so that they can hybridise together. The probe has properties that allow it to be selected or captured and when this occurs, the target is co-purified (209). Probes can be synthesised on a surface or bead for direct capture or the probes can have a ligand attached so that it can be indirectly captured. Biotin ligands that can be captured by streptavidin-coated paramagnetic beads are one example of the latter (217). Non-captured DNA needs to be washed away as off-target DNA can reduce the efficiency of the process. High levels of input DNA are necessary (213). Indirect capture methods have the benefit that in-solution hybridisation can be used, which supplies a higher specificity of enrichment. In the case of the Agilent SureSelect hybrid capture, one example of such systems are RNA probes used so the excess probe can be removed by RNAses, lowering the background (218). SureSelect has the capacity to cover up to 24 Mbp of targets (219). Figure 1-9 portrays a schematic representation of an in-solution hybrid capture system. This was the system employed in this thesis and will be detailed in Chapter 2. Haloplex

also takes advantage of molecular techniques to achieve enrichment. In this case, probes

are designed to incorporate a restriction endonuclease site at the ends of the fragments

of interest. Selected targets can then be circularised and methods for enriching circular

DNA molecules are used to remove the unwanted background (213). The method is more

convenient than SureSelect but the total size of the regions to be targeted is smaller and

dependent on not being confounded by the presence of the restriction sites used.
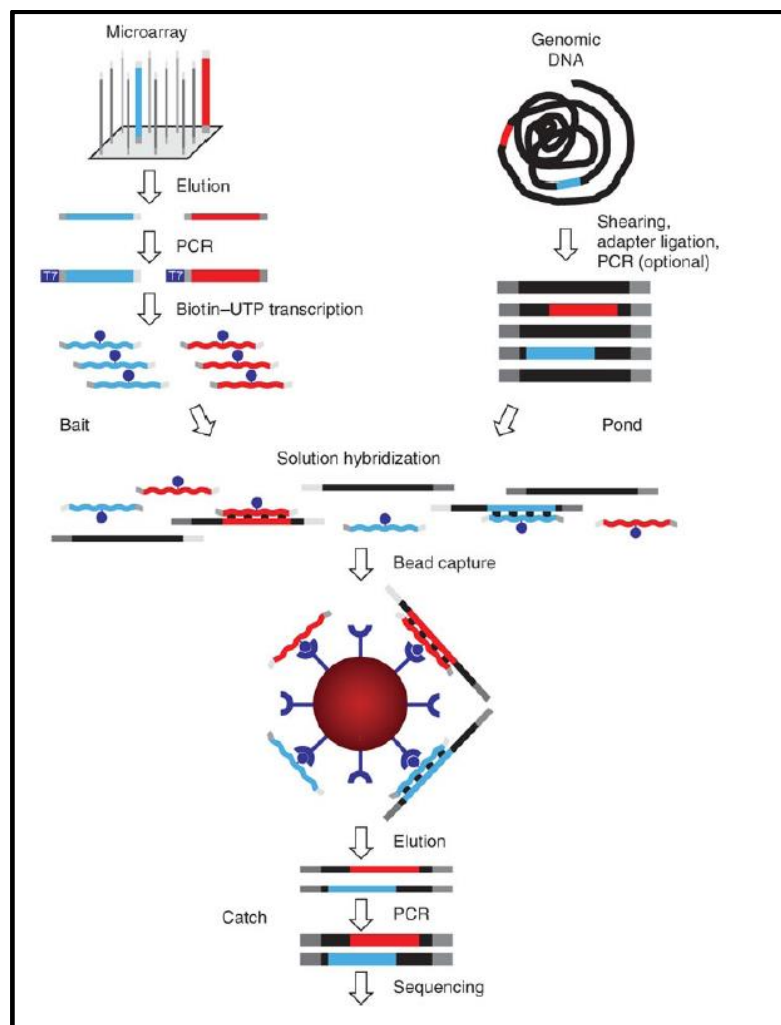
Figure 1-9. Schematic Representation of In-Solution Hybridisation. The diagram shows target enrichment steps that involve RNA-biotinylated probes (baits) (top left) hybridising the genomic material (top right). The hybrid library is then bound to streptavidin beads (as shown in the bead capture step) prior to washing out non-targeted materials. The figure is adapted with permission from (220).

Once enrichment has been performed, the captured targets need to be formatted for deep sequencing. Typically, DNA adapters are be ligated to DNA molecules. For the purpose of multiple samples, multiple, sequence-barcoded adapters are available, so each sample can be uniquely labelled and then identified after mixing (221). PCR amplification in readiness for the sequencing reactions can then be carried out using primers specific to the adapters.

1.16. Analysis of NGS Data

NGS sequence data analysis comprises signal production for sequences from each template, trimming of adapters and low-quality sequence, mapping of reads to the reference genome and annotation of variants (222). Data from the primary stage is filed and quality scored so that quality controls can be implemented at subsequent stages (223). Similarly, during the mapping stage, the best-aligned reads are determined and recalibrated locally around the indels using the genome alignment tool kit (GATK) (224). GATK uses a Sequence Alignment Map (SAM), which can be compressed to a Binary Alignment Map (BAM) (225, 226). An Integrative Genomic Viewer (IGV) is then employed to visualize the coverage and sequence of the aligned reads (227). Variants are identified after comparison of sequencing reads with a reference genome and tabulated variants are presented in a Variant Call Format (VCF) file, which specifies basic information on each identified variant (226, 228). If germline sequence information is also available from the same individual, acquired mutations can be identified. In the absence of germline material, population databases of germline polymorphisms are used, however, contamination with unreported polymorphisms can take place. Databases for variants are dbSNP and the 1000 Genomes Project (see Section1.17.3 Refinement of the Human

Reference Genome, page 78) for germline polymorphisms and COSMIC for somatic

mutations in cancer (229). Variant effect prediction (VEP) tool can be used to predict the

functional changes of the resultant proteins (Netto, 2015).

1.17. Ion Torrent PGM Next-Generation Sequencing

1.17.1. Sequencing Using Ion Torrent PGM

The Ion Torrent PGM is a small benchtop sequencer and the fastest and cheapest NGS

platform. The PGM uses pH-sensitive chips with thousands to millions of microwells, and

each one can trap a bead of clonally amplified DNA templates and detect/quantify the

incorporation of nucleotides by a DNA polymerase on the templates because each

instance releases a proton as part of the reaction (230). This change in pH is detected by

the underlying pH sensor and subsequently converted to electrical voltages. Each base is

utilised in turn in a repeating cycle. In this way, voltage is only detected if a nucleotide can

be incorporated and the peak voltage height is proportional to the number of nucleotides

incorporated. However, long HPs may be misinterpreted, which results in Hp errors (231).

DNA is attached to the surface of soft beads called Ion Sphere Particles (ISP) in

preparation for amplification by emulsion droplet PCR (232). Sequence output is

dependent on reading length and chip capacity (Table 1-1) (233).

1.17.2. Ion Torrent PGM Data Analysis

The raw voltage signals are processed into a linear sequence data file in a FASTQ format

using Torrent Base Caller. This file can be exported to various external pipelines for

downstream analysis.

Alignment of the sequence reads to a reference genome yields a BAM file and the Torrent

Variant Calling (TVC) plugin produces a Variant Call Format (VCF) file, which contains the

list of identified variants, their chromosomal location and quality metrics. This file can be

exported to various open-access variant effect predictor programmes or to Ion Reporter™

software, either manually or automatically using the Ion Reporter™ Uploader plugin. This

software is efficient, allowing an organised workflow for the annotation and prediction of

the effects of the identified variants. Moreover, it can be used to compare multiple

samples.

1.17.3. Refinement of the Human Reference Genome

Although a "completed "version of the Human Genome was published in 2003 this

marked a practical endpoint for a particular phase of the project and considerable work

was still required to correct misassemblies, errors and filling gaps (234). The genome was

also based on DNA from multiple individuals. To this day, the Reference Genome remains

a work in progress and regular "patches" are released. This has become essential as the

extent of polymorphic variation, in particular in the form of large-scale insertions,

deletions and rearrangements became apparent through the availability of a reference

against which comparisons could be made (234) (235). There are places in the genome

where completely different segments of DNA are found between one individual and the

next. These were originally considered to be errors or gaps, but it is now acknowledged

that they are previously unrecognised polymorphic variation. This variation is amongst

the information stored in relation to the reference genome. Significant efforts have been

made identifying simple polymorphic variants, in particular, SNPs, because millions are

considered to exist, and they are an invaluable resource for understanding the genetic

structure of human populations, their migration, evolution and phenotypic associations.

The HapMap project first identified one million SNPs from four generic populations worldwide (236). Since that time, there has been more than a 10-fold increase in the number of known SNPs, largely through the 1000 Genomes project. This was enabled by the availability of Next Generation Sequencers and target-enrichment techniques.

The 1000 Genome Study was carried out between 2008 and 2015 with the goal of discovering all human genetic variation with a frequency of at least 1% in the population. It focused sequencing on family trios of two parents and one child. The final data set was based on 2,504 individuals from 26 populations. Low coverage and WES were available for all of these individuals and 24 individuals were sequenced to high coverage (237). In 2015, the complete sequencing set had been reconstructed using WGS for 2,504 individuals from 26 different populations (238, 239). As a result, this project acted as a significant resource for conducting research on human variants across several populations (240). In addition, the approach can identify genetic anomalies in inherited diseases (241).

1.18. Next-Generation Sequencing of CLL, Clonal Evolution and Clinical Course

The availability of NGS has led to the more detailed characterisation of the genomes of CLL and their clonal evolution. An early study found recurrent mutations for notch1 (NOTCH1), exportin 1 (XPO1), myeloid differentiation primary response gene 88 (MYD88) and kelch-like 6 (KLHL6) (242). Although the recurrences were found in a screen of 363 patients, only four patients were surveyed initially, 46 somatic mutations that were predicted to affect gene function were identified. It was not clear whether the four

recurrently mutated genes actually drove the disease because an initial survey of more

cases could have implicated more genes. However, MYD88 and KLHL6, in contrast to

NOTCH1 and XPO1, were associated with mutated immunoglobulins. NOTCH1 mutated

patients were at a more advanced stage when diagnosed and had shorter overall survival,

suggestive of the more aggressive disease. A second study identified other genes to be

mutated in the notch signalling pathway, including FBXW7 and also ZMYM3, MAPK1, and

DDX3X as other novel potential drivers (243). Splicing factor SF3B1 was of particular

interest with mutations occurring in 15% of patients and in association with 11q deletions

and a poor prognosis. SF3B1 was also implicated in other studies (244).Comparisons of

multiple samples taken at different time points from the same patients permitted the

clonal evolution of CLL to be assessed. In one study, in a variety of cases, this included

later samples taken after chemotherapy. Ten of 12 versus 1/6 cases with and without

chemotherapy, respectively, demonstrated evidence of clonal evolution. In particular,

clones with mutations in genes (TP53 and SF3B1) were considered drivers expanded with

time (245). Specific examination of the Immunoglobulin Ig heavy and kappa chain (IGH

and IGK) loci in 25 out of 31 patients determined changes in these loci that were

consistent with clonal evolution in patients who underwent treatment (246, 247).

Examination of 17p (TP53) and 11q (ATM) also observed losses, consistent with their

known association with poor prognosis CLL. Out of 168 CLL patients who were wild-type

for these loci by FISH, eight and five patients acquired a 17p deletion and 11q deletion,

respectively (248). In eight of these cases, a TP53/ATM mutation was observed in 4-50%

(median=9%) of the baseline sample and was associated with high-risk chromosomal

clonal evolution (p=0.02). In a study involving 59 patients from a larger clinical trial of 538

cases, 44 recurrently mutated genes and 11 recurrent copy number changes comparing pre-treatment with relapsed samples after treatment found a high frequency (57/59 cases) of clonal evolution (245). The relapsed clone was already present in 18/59 of the original cases and was consistent with the presence or acquisition of 'driver' mutations.

1.19. The Study Hypothesis and Aims

NGS of CLL samples has resulted in significant progress describing the pathways that drive CLL and the clonal evolution that leads to aggressive disease. However, the underlying cause of genetic instability and its association with aggressive disease has not been fully addressed. TP53 deletion/mutation in CLL is associated with clonal instability (e.g., increased risk of Richter transformation) (20), which is in line with the known role of wild-type TP53 as a mediator of apoptosis, cell-cycle arrest and DNA repair in response to cellular stress (114). However, it is possible that clonal instability owing to P53 inactivation may be amplified because of the acquisition of secondary mutations in genes other than TP53 that are involved in the DNA repair 'snow ball effect'. Whether such a mechanism exists and if so, to what extent it contributes to clonal instability in cases of CLL with P53 inactivation has not previously been investigated. It is the objective of this thesis to address this important question (Figure 1-10).
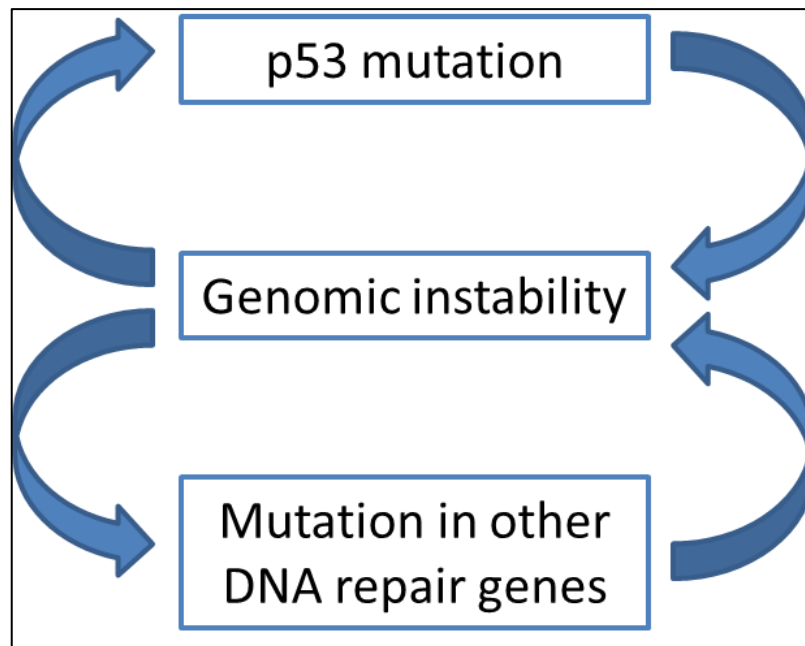
Figure 1-10. The Study Hypothesis. Generally, the normal function of p53 as a DNA repair protein is to mediate cell-cycle arrest and DNA repair activation. In addition, TP53 mutated/deleted cases are related to clonal instability in CLL. Taken together, the study hypothesis states that progressive CLL, such as inactivated p53 cases, could accumulate mutation burden in other DNA repair genes, which may increase the possibility of clonal instability, such as Richter Syndrome.

1.19.1. Hypothesis

Defects in genes responsible for the maintenance of DNA lead to genomic instability that allows more rapid progression of CLL to aggressive disease.

1.19.2. Approach

The hypothesis predicts that cases of CLL, pre-selected for having an advanced form of the disease, will have a high proportion of genetic abnormalities and these will be found in association with mutations in genes for DNA maintenance.

2. Chapter 2: Materials and Methods

2.1. Solutions, Reagents and Materials

1.5 ml LoBind tubes (AG Eppendorf, Hamburg, Germany)

PCR tubes (AG Eppendorf, Hamburg, Germany)

Sterilised pipette tips with aerosol filters (StarLab, Milton Keynes, UK)

Powder-free gloves

Nuclease-free water (Sigma-Aldrich, Dorset, UK)

TE buffer (pH 8.0) (Thermofisher Scientific, Loughborough, UK)

QIAquick PCR Purification Kit (Qiagen, Manchester, UK)

Wizard® SV Gel and PCR Clean-Up System (Promega, Southampton, UK)

Qubit® dsDNA HS Assay Kit (ThermoFisher Scientific, Loughborough, UK)

High Sensitivity DNA Kit (Agilent Genomics, Shrewsbury, UK)

Herculase II Fusion DNA Polymerase (Agilent Genomics, Shrewsbury, UK)

Agencourt AMPure XP beads (Beckman Coulter Genomics, High Wycombe, UK)

E-Gel SizeSelect 2% Agarose Gel (ThermoFisher Scientific, Loughborough, UK)

Ion Xpress Plus Fragment Library Kit (ThermoFisher Scientific, Loughborough, UK)

KAPA Adapter Kit for Ion Torrent Platform 8 libraries (KAPA Biosystems, Wilmington, USA)

Chapter 2: Materials and Methods

Dynabeads MyOne Streptavidin T1 (ThermoFisher Scientific, Loughborough, UK)

Ion PGM Sequencing 200 Kit v2 (ThermoFisher Scientific, Loughborough, UK)

2.2.   Equipment

Pipettes - P10, P20, P100, and P1000 (Gilson, Dunstable, UK)

Speed Vac Plus Vacuum Concentrator (Savant, New York, USA)

Vortex mixer

Water bath (Grant, Devizes, UK)

Minispin Microcentrifuge (AG Eppendorf, Hamburg, Germany)

Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, Loughborough, UK)

NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Loughborough, UK)

lid-heated thermal cycler (Mastercycler AG Eppendorf, Hamburg, Germany)

INGenius Imaging System (Syngene, Cambridge, UK)

2100 Bioanalyzer (Agilent technologies, Shrewsbury, UK)

E-Gel Safe Imager Real-Time Transilluminator (E-Gel iBase and E-Gel Safe imager combo kit (ThermoFisher Scientific, Loughborough, UK)

Magnetic stand (DynaMag™-PCR Magnet ThermoFisher Scientific, Loughborough, UK)

Ion OneTouch 2 system (ThermoFisher Scientific, Loughborough, UK)

2.3.   Clinical Samples

2.3.1.   CLL Samples for Targeted Sequencing (Chapters 4 and 5)

To meet the aim of the study, 10 CLL cases with severe prognosis were selected from the CLL206 trial (46, 47)- see Section 1.5.4. NCRI CLL206 and NCRI CLL210 Trials . Samples of peripheral blood were collected following informed consent from the patients and from trial committees. Lymphoprep (STEMCELL Technologies, Cat # 07851) was used to separate mononuclear cells (MNCs) from the peripheral blood (see section 2.4 Separation of Mononuclear Cells (MNCs)). MNC samples were then cryopreserved according to the protocol of the University of Liverpool Leukaemia Biobank. Samples were processed and gDNA was extracted by Dr Gillian Johonson.

2.3.2.   CLL Samples for Clinical Validation (Chapter 6)

49 cases of CLL patients were obtained from the CLL210 trial (see Section 1.5.4. NCRI CLL206 and NCRI CLL210 Trials and the local Liverpool blood biobank (249) according to TP53 mutation status and their blood MNCs prepared as in the following section.

Details of the samples, including the clinical and biological characteristics, are found in Chapter 6.

2.4.   Separation of Mononuclear Cells (MNCs)

MNCs were isolated from peripheral blood as instructed by the standard operating procedure (SOP) of the biobank. Heparinised blood was then transferred into a Lymphoprep TM (Axis-Shield PoC AS, Norway, d = 1.077)-containing tube (50ml falcon tube) and centrifuged for 30 minutes at 800xg. Afterwards, the MNC layer was collected

and this was followed by washing then resuspension in a mixture of sterile RPMI-1640 medium containing 10% fetal calf serum (1:1 v/v), providing 100% viability of MNCs (250). The cells were then mixed slowly with the same volume of mixture on ice. Afterwards, a RPMI-1640 mixture containing 20% dimethyl sulphoxide (DMSO) (Sigma-Aldrich, Dorset, UK) was prepared and added slowly to cells with 20 million cells per ml as the expected cell count. Samples were then transferred into 1 ml cryovials and stored in the -80°C freezer.

2.5. DNA Purification

2.5.1. Genomic DNA purification

The All Prep QIA Extraction Kit (Qiagen, Manchester, UK) was used according to the manufacturer's recommendations to purify macromolecules, including genomic DNA from MNCs.

Frozen samples were thawed and RLT plus reagents were prepared by adding 10 μl β-mercaptoethanol (β-ME) to every 1 ml of RLT. Samples were washed twice with Phosphate Buffered Saline (PBS) solution and recovered between washes by centrifugation at 4000 rpm in a benchtop minicentrifuge (Eppendorf, Hamburg, Germany) for 1 minute. 600 μl of RLT plus was added to each sample. The mixture was vortexed then transferred to a QIA Shredder column in a 2 ml microcentrifuge tube. The column was centrifuged at the maximum speed for 3 minutes. The eluate was then transferred into all prep DNA spin columns and centrifuged for 30 seconds at 10,000 rpm. The DNA column was then transferred into a new 2 ml collection tube, 500 μl of AW1 was added to the mixture and the tube centrifuged for 15 seconds at 10,000 rpm. 700 μl of AW2 was

added to the column which was then centrifuged at 14,000rpm for 2 minutes. The column

was placed into a new 1.5 ml tube and 50 µl of elution buffer added. After 1 minute, the

column was microcentrifuged at 10,000 rpm for 1 minute and the column subsequently

being discarded. The tube containing the extracted gDNA solution was stored at -20°C.

2.5.2. DNA Clean-Up

PCR products or samples of manipulated DNA were purified using either a QIA Quick

Column purification kit (Qiagen, Manchester, UK) or the similar Wizard® SV Gel and PCR

Clean-Up System (Promega, Southampton, UK).

2.5.2.1. QIAquick Column Purification Kit

Buffer PE was prepared according to the kit guide by adding five volumes of ethanol. Five

volumes of buffer PB was added to each volume of a DNA sample and it was verified that

the pH was 7.5. The sample was then placed into a QIAquick column and centrifuged for

60 seconds at 13,000 rpm. Samples were washed by adding 750 µl buffer PE to the

column and centrifuging for 60 seconds at 13,000 rpm. A further centrifugation step

without the addition of buffer was carried out to remove any residual wash buffer. DNA

was eluted by adding 50 µl buffer EB (10 mM TrisCl, pH 8.5) and collected following

centrifugation.

2.5.2.2. Wizard® SV Gel and PCR Clean-Up System

The Wizard® SV Gel and PCR Clean-Up System (Promega, Southampton, UK) can purify

manipulated DNA from molecular genetic reactions or DNA from excised agarose gels.

Agarose gel containing DNA was placed into the 2ml tube. 10 µl of membrane-binding

solution was then added per 10 mg of gel slice and the mixture incubated at 50-65°C in a

water bath for ~10 minutes until the gel was dissolved. Membrane wash solution was

prepared by adding 95% ethanol to the stock solution. The melted solution was

transferred to a SV minicolumn and placed in a collection tube and incubated for 1

minute, after which the tube was then centrifuged at 14,000 rpm for 1 minute, the

supernatant discarded, and the column transferred to a fresh 1.5 ml tube. The column

was washed by adding 700 μl of membrane-washing solution and centrifuged for 1

minute at 14,000 rpm. Flow-through was discarded and the washing step repeated using

500 μl of membrane-washing solution. The column was placed into a new 1.5 ml tube and

50 μl nuclease-free water was then added to the column, which was incubated for 1

minute at room temperature before collection by centrifugation at 16,000rpm for 1

minute. The eluted DNA was stored at 4°C. Gel-free DNA was similarly purified except

equal volumes of the DNA solution and membrane-binding buffer was mixed in the first

step.

2.5.2.3. DNA Purification Using Agencourt AMPure XP Beads

AMPure beads were incubated at room temperature for 30 minutes and then well-mixed

prior to the addition of the sample. 90 μl of the AMPure beads per sample were added to

a 1.5 ml Lobind tube and sample (~50 μl) was added and mixed by vortexing before

incubating for 5 minutes at room temperature. Beads were separated using the magnetic

stand and the supernatant was discarded. 500 μl of 70% ethanol was added and the

sample incubated for 1 minute to allow the beads to settle, after which the supernatant

was removed. Next, the washing with ethanol step was repeated. Samples were

maintained at room temperature for a maximum of 5 minutes to allow residual ethanol to

completely evaporate. 20 µl of TE buffer (pH 8.0) was added to the sample and mixed with the beads before placement onto the magnetic stand for 3 minutes, then removing the supernatant (~20 µl) containing the purified DNA for use.

2.6. Analysis of Nucleic Acids

2.6.1. Nanodrop Spectroscopy

Nucleic acid amounts and purity were estimated by spectroscopically determining A_{260}, A_{280} and A_{320} values in 1-2 µl samples using a Nanodrop ND1000 Spectrophotometer (Labtech, Heathfield, UK). A_{260} of 1 for a 1 cm path length was assumed to correspond to 50 ug/ml of double-stranded DNA, 30 ug/ml of single-stranded DNA or 40 ug/ml of RNA. An $A_{260/280}$ ratio of ~1.8 was accepted as "pure" DNA and ~2.0 as "pure" RNA. "Pure" nucleic acid also required an $A_{260/280}$ of ~2.0.

2.6.2. Qubit® 2.0 Fluorimetry

A fluorescence-based assay was used to quantify nucleic acids and proteins between 5 pg-500 ng in 1-20 µl solutions (HS DNA Qubit Kit) (253, 254).

Working solution was prepared by adding 1:200 HS DNA reagents to the working solution and vortexed to mix. Each reaction was prepared by adding 1-20 µl of sample and prepared working solution to 200 µl in qubit 0.5 ml tubes. A range of control samples covering a range DNAs and known amounts was set up in parallel. Mixtures were incubated for 2 minutes and then measured using the Qubit.

2.6.3. Nucleic Acid Assessment by Agarose Gel Electrophoresis

DNA samples were routinely analysed for size between 150 and 20 kbs and integrity by agarose gel electrophoresis using 12x14 cm gels between 0.9-2.4% agarose in 0.5x TBE buffer at 120 volts for ~45 minutes (255). Agarose powder (Sigma-Aldrich, Dorset, UK) was added to 0.5x TBE buffer (2mM Tris-borate and 2mM EDTA), and the mixture was heated in a microwave for ~2 minutes until the agarose was fully dissolved. 5 µl Midori Green Advance DNA stain (Geneflow, Lichfield, UK) was added per 100 ml agarose to stain the nucleic acids. The cooled solution was poured into an electrophoresis tray and a sample comb placed into the gel that was then allowed to solidify. Samples were mixed with 6x loading dye at 5:1 (v/v) before loading into wells. 100-1000 bp DNA ladder was run as controls (New England Biolabs, Hitchin, UK). Gels were immersed in 0.5x TBE in an electrophoresis tank for running and visualised using the INGenius Imaging System (Syngene, Cambridge, UK) upon completion.

2.6.4. Analytical DNA Quality and Quantity assessment

A Bioanalyzer High Sensitivity DNA assay (Agilent technologies, Shrewsbury, UK) was utilised to assess DNA for quality and quantity. The Bioanalyzer chip and DNA marker ladder were prepared according to the manufacturer's protocol. One µl of samples were loaded with DNA input ranges between 5-500 pg/µl. The chip was loaded into the 2100 Bioanalyzer (Agilent technologies, Shrewsbury, UK) and the run was started within 5 minutes of preparation using the settings recommended by the manufacturer.

2.7. Molecular Genetic Techniques

2.7.1. DNA Amplification

The Polymerase Chain Reaction (PCR) was made use of to amplify the DNA. The GoTaq®

Flexi DNA Polymerase kit or Herculase II Fusion DNA polymerase were employed. The

total PCR reaction volume per 50 µl, contained the components in Table 2- 1.

Table 2- 1: PCR Reaction Reagents per 50 µl PCR Reaction

PCR Reagent	Volume (µl)
Water	32.75
5x Buffer	10
25 mM MgCl	3
10 mM dNTPs	1
Taq polymerase	0.25 (units)
Forward and reverse Primers (20µM each)*	2
DNA	3 (1 to 100 ng)

 *Details of primers are discussed separately.

PCR was performed in an Eppendorf Mastercycler 5333 (AG Eppendorf, Hamburg,

Germany) using the optimised programme as follows:

94°C for 3 minutes (Initial denaturation)

94°C for 30 seconds (Denaturation)

60 or 65°C for 30 seconds (Annealing)

72°C for 30 seconds (Extension)

Repeat steps 2-4 for 32 cycles using (Go Taq) or 10 cycles using (Herculase II) polymerase reagents

72°C for 5 minutes (Final extension)

END

2.7.2. PCR Primer Design

Primers were designed according to Oligo Analyzer Version 3.1 (Integrated DNA Technologies (IDT, Surrey, UK).

2.7.3. Sanger Fluorescent Dideoxynucleotide Sequencing

PCR products were custom sequenced by separately using their corresponding forward and reverse PCR primers (Genewiz® Sanger Service, Takeley, UK).

2.7.4. DNA Fragmentation

DNA was mixed with Ion Shear Plus 2.5 µl of 10x reaction buffer and nuclease-free water to 25 µl. Then, 1 µl of Ion Shear Plus enzyme mix was added. The reaction mixture was incubated in a water bath at 37°C for 18 minutes. Five µl of Ion Shear Plus stop buffer was added and the sample was stored on ice.

2.7.5. DNA Adapter Ligation

A 75 µl reaction master mix was prepared per sample containing: 7.5 µl 10x ligase buffer; 7.5 µl Ion Xpress P1 adapter; 2 µl dNTP mix; 31 µl nuclease-free water; 4 µl DNA ligase; and 8 µl nick repair polymerase. 10 µl of one of the barcode adaptors (KAPA Biosystems,

Boston, USA) was added per sample. The ligation reaction was then incubated in a thermal cycler at 25°C for 15 minutes followed by 72°C for 5 minutes and held at 4°C.

2.7.6. DNA Capture Hybridisation

750 ng of genomic DNA per 3.4 µl of sample was required. Samples with a DNA concentration of < 221 ng/µl were therefore concentrated in a vacuum concentrator at ≤ 45°C. Regarding SureSelect hybridization buffer, 49 µl per sample was prepared by mixing 25 µl Hyb#1, 1 µl Hyb#2, 10 µl Hyb#3 and 13 µl Hyb#4, 40 µl was used. 2 µl of the SureSelect capture library was added to 5 µl of 10% RNase block per sample. Additionally, 5.6 µl of SureSelect block mix was prepared per sample using 2.5 µl Block#1, 2.5 µl Block#2, and 0.6 µl Block# 3. 3.4 µl of the 221 ng/µl each sample was added to a 0.5 µl PCR tube and 5.6 µl of SureSelect Block mix added before heating at 95°C for 5 minutes and lower the temperature to 65°C in a thermal cycler with the lid set to 105°C. 40 µl of hybridization buffer was added to a 0.5 µl PCR tube per sample and transferred to the thermal cycler maintained at 65°C, at which point the tube was incubated for 5 minutes. Seven µl of SureSelect capture library mix per sample was added to a 0.5 µl PCR tube and placed into the thermal cycler at 65°C for 2 minutes. The tubes were maintained in the thermal cycler and each prepped sample was transferred to a SureSelect Capture Library tube, with the contents then being mixed by pipetting 10 times and tube capped in the thermal cycler to allow hybridisation to proceed for 24 hours.

2.7.6.1. Magnetic Bead Preparation

SureSelect Wash 2 was prewarmed at 65°C in a water bath. 50 µl of resuspended Dynabeads MyOne Streptavidin T1 beads (Thermofisher Scientific, Loughborough, UK)

were added to a 1.5 ml LoBind tube and vortexed. The tube was then placed into the

Dynal magnetic separator (Thermofisher Scientific, Loughborough, UK) and the

supernatant removed. This latter step was repeated twice, each time re-suspending the

recovered beads in 200 µl of SureSelect binding buffer.

2.7.6.2. Hybrid Capture

The hybridized sample was directly added to the bead solution, and the tube was then

inverted three to five times to mix and incubate for 30 minutes at room temperature. A

magnetic separator was used to separate the beads and their captured DNA from the

buffer and the supernatant with uncaptured DNA was removed. Beads were washed by

adding 500 µl of SureSelect Wash 1 and mixed by vortexing before incubation for 15

minutes at room temperature with occasional mixing by vortex. Beads were recovered as

before and 500 µl of the pre-warmed SureSelect Wash 2 was added, mixed by vortexing

and incubated for 10 minutes at 65°C with occasional mixing by inversion for further

washing. Recovery of the beads was performed as before, and a second washing step was

repeated for a total of three times. Lastly, 30 µl of nuclease-free water was added to the

beads, with the targeted DNA enrichment retained.

Target enrichment was performed by cRNA hybrid capture according to the SureSelect TE

System (Agilent technologies, Shrewsbury, UK). This included cRNA probes according to

SureSelect protocol, which captures regions totalling from 0.5 Mb up to 2.9 Mb. The steps

required involve selection of cDNA probes, DNA sample preparation for hybridisation,

hybridisation capture and post-hybridisation amplification (256). Selection of

appropriately sized fragments was performed at intermediate stages. Figure 2.1 outlines

the relationships between the steps with the Sure Select TE System.
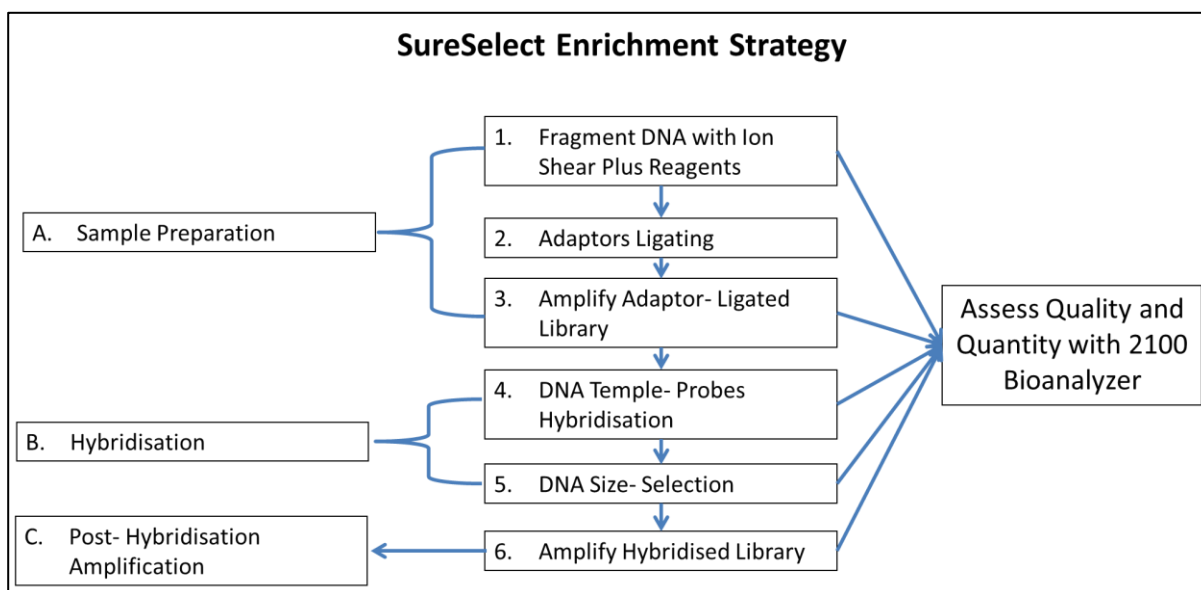
Figure 2- 1: The Steps of SureSelect Target Enrichment System (Agilent). Additional details are found in the manufacturer's protocol; SureSelect Target Enrichment System for Sequencing on Ion Proton (257).

2.8. cRNA Probe Design

Agilent SureDesign online software was employed to design biotinylated cRNA-

hybridisation probes. The Human Genome was selected from its database of genomes

and pre-designed primers retrieved for human genes from the human reference genome,

hg19. In addition, SureDesign provided stringency options to increase the likelihood of

coverage of difficult regions, such as those with high GC content, and also to reduce the

possibility of amplifying off-target regions. An average of two probes was selected per

exon but this was adjusted according to the aforementioned and also exon size and

masking for repeat sequences requiring the number of probes to be boosted in certain

regions or their density to be increased. The masking from least to most used the masker

tools: RepeatMasker, WindowMasker and the Duke Uniqueness 35 track. For the work

herein, SureDesign was chosen to design primers for enriching exons of 194 DNA maintenance genes and 2x probe coverage was the density with maximum boosting along with low masking options.

2.9.   Sample Pooling for Multiplexed Sequencing

Barcoded samples were mixed for multiplexing according to a "rule of thumb" formula:

$$\frac{V(f) \times C(f)}{\# \times C(i)}$$

where:

V(f) is the final volume of the pool;

 C(f) is the expected final concentration of all the multiplexed samples in the pool, which was 26 nM;

is the number of samples to be pooled; and

C(i) is the initial concentration of each barcoded sample.

2.10. Size-Selection of Samples

Size-selected DNA samples were prepared by agarose gel electrophoresis using the integrated E-Gel system and E-Gel SizeSelect 2% agarose gels.

Cassettes were inserted into the E-Gel iBase Power System and 20 µl of the sample, including loading buffer, was added per well. Additionally, 10 µl of 50 bp DNA ladder (26 ng/µl) was loaded into the middle lane (Lane M). Twenty-five µl of nuclease-free water was added to unused wells and the amber filter unit was placed prior to running. Samples

were electrophoresed until the required band (230 bp or depending on application) in the marker was reached on the reference lane. Twenty µl of each sample was then collected from each collection well and the electrophoresis was ceased.

2.11. Template Preparation for Ion PGM sequencing

For template preparation, the Ion OneTouch 2 System was used along with the Ion PGM Template OT2 200 Kit v2 and Ion PGM Sequencing 200 Kit v2 (Agilent technologies, Shrewsbury, UK). Template preparation, including the Ion PGM sequencer run, was performed as a service provided by the GCLP Laboratory, University of Liverpool, Liverpool, UK. For more information regarding the template and sequencing preparation, please refer to the manufacturer manuals: Ion PGM™ Template OT2 200 Kit (Publication Part Number MAN0007221 Rev. 4.0) and Ion PGM™ Sequencing 200 Kit v2 (Publication Number MAN0007360 Revision 1.0).

2.12. Ion Torrent PGM Sequencing Data Processing

An Ion Torrent Suite pipeline (illustrated in Figure 2-2) was used for analysing the Ion PGM sequencing data, where the FASTQ file (text-based sequences and quality scores) of the raw data was processed and then aligned using a TMAP processor to the human genome reference, hg19, as a Bam QC file (binary format). The Bam QC file was processed for variant calling using Torrent Variant Caller (TVCv4.2), generating the candidate variants in VCF QC format (variant representation). The variants were genotyped using Ion Reporter IR 4.2 and a final quality control report was produced for each PGM sequencing run.

The variant selection strategy was based on Ion reporter genotyping, dbSNP and COSMIC as a technical control as no paired-samples were available (Figure 2-3). Variants were divided into synonymous and non-synonymous variants and indels according to their consequences for amino acid coding. They were further divided into germline or somatic aberrations as estimated by comparison with those collected in dbSNP (258, 259) and COSMIC (260, 261).
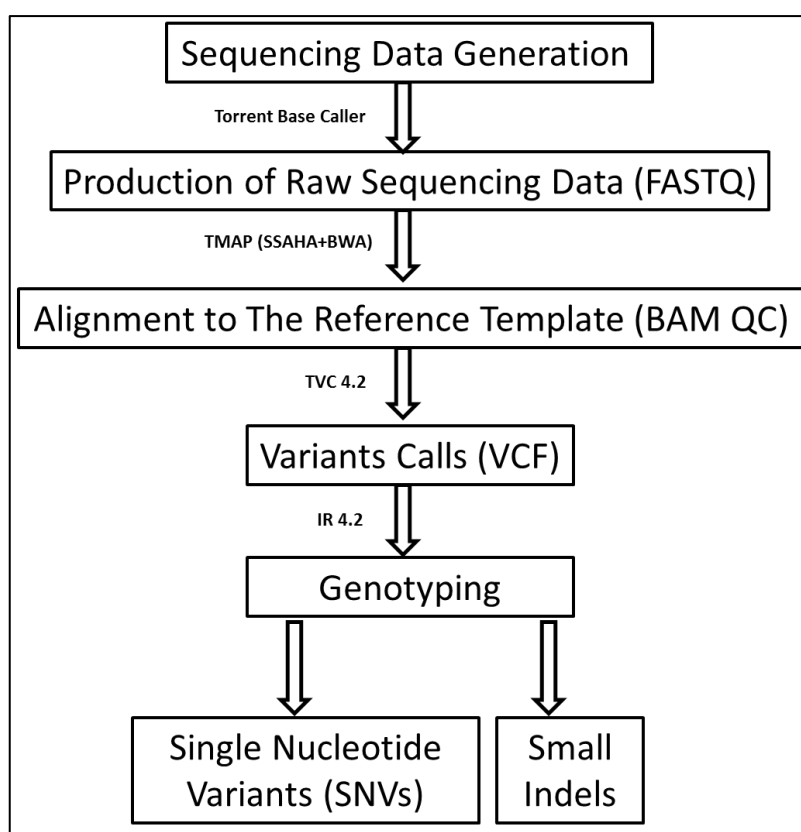
Figure 2-2: Sequencing Analysis Pipeline of Ion Suite. The raw data was generated as FASTQ using Torrent Base Caller and then the data were aligned to a reference template (in our case, human genome reference 19 (hg19)) using TMAP processor in BAM QC file format. Variant calling was applied using Torrent Variant Caller 4.2 in VCF format. For genotyping, Ion reporter (IR 4.2) on-line software was applied and different genotype variants identified by comparison using COSMIC, dbSNP Polyphen and sift. Abbreviations: TMAP - torrent mapping alignment program, BWA - Burrows-Wheeler Aligner, SSAHA - Sequence Search and Alignment by Hashing Algorithm.
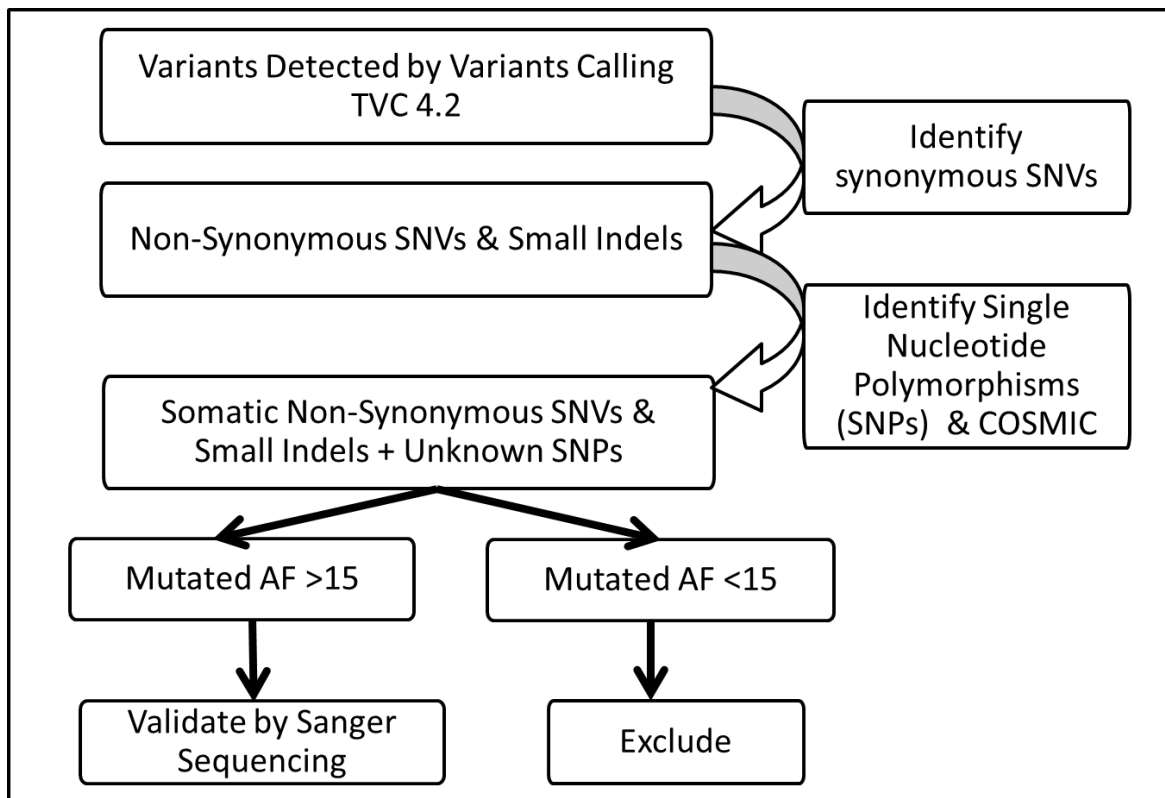
Figure 2-3: Genotyping of Variants Detected by Ion PGM Sequencing and Identification by Ion Reporter 4.2

2.13. Whole Genome Sequencing

Bi-directional WGS was performed using the Illumina reversible terminator chemistry by the Beijing Genomic Institute (BGI) in Hong Kong and data analysed at the Centre of Genomic Research (CGR) in Liverpool, UK.

Raw sequence data from base calling was processed using the Illumina bioinformatics pipeline. Adapter sequences and low-quality bases were removed from the raw data. Resultant sequence reads were aligned by Burrows-Wheeler Aligner (BWA) to the hg19/hg38 reference sequences and converted into the BAM file format. Mate-pairing information, read groups and PCR duplicate information were added to the Bam files. Refined BAM files were then used to call variants by GATK, which detected SNPs, small insertions/deletions, CNVs, SNVs and somatic indels. Somatic CNVs were detected by

BreakDancer. BreakDancer features of two algorithms (262). The first one (BreakDancer Max) identifies five types of structural variants that includes deletions, insertions, inversions, and intra/inter-chromosomal translocations. The second algorithm (BreakDancer Mini) is applied to the detection of small indels (10-10 bp) that are outside the scope of BreakDancer Max (Figure 2- 4). After filtering for high confidence, variants were annotated using SnpEff.

CNA detection was performed using control-FREEC using algorithms developed for WGS (263). Control-FREEC uses input aligned reads to generate normalized copy number and B-allele frequency (BAF) profiles. Both profiles are segmented, and the genotype status of each segment is ascribed using both copy number and allelic frequencies; genomic alterations are then annotated.
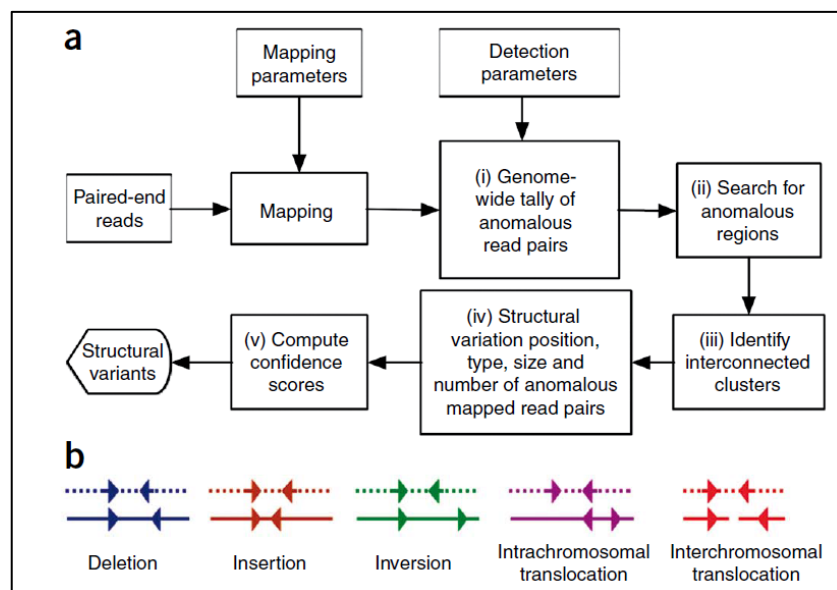
Figure 2- 4 BreakDancer Algorithm. A. BreakDancer workflow. B. Anomalous reads detected by BreakDancer. Arrows represent the read pair orientation. Dotted lines represent the analysed genome. Solid line is the reference genome. Used with permission from Chen et al. (262).

2.14. Variant Grouping and Coverage Analysis

SNPeff was employed to classify variants into known SNP variants and unknown (novel) variants. Repeat Masker was applied to classify variants into repetitive variants (264), including homopolymer, simple sequence or non-repetitive regions.

2.15. Statistical Analysis

Simple statistical analyses were carried out with IBM Statistical Package for the Social Sciences (SPSS) v22 (IBM, Chicago, USA). Data visualising, such as histograms, line charts and pie charts, were used. A measure of data spread was also applied, like mean (X), standard deviation (SD) and range. A *P*-value of 0.05 or less was considered statistically significant. Regarding genome-wide significance, **P** ≤ 5 × 10(-8) was considered significant and replicated (265). Details on tests are presented in the relevant result chapters.

3. Chapter 3: Development of Targeted NGS for the Identification of Mutations in DNA Maintenance Genes

3.1. Introduction

DNA maintenance genes are defined as any gene that produces a protein to protect the genomic integrity of the cell or causes the cell to undergo apoptosis as a self-defence mechanism to prevent carcinogenesis and other-gene related disease. Multiple repair mechanisms for different DNA maintenance proteins, including those related to the repair of single-stranded DNA, double-stranded DNA, cell-cycle checkpoints and apoptosis-induction are known and were discussed in the Introduction (See Chapter 1). Our hypothesis highlights the possible involvement of aberration of DNA maintenance genes in clonal and genomic instabilities in CLL, especially those with P53 inactivation. Publications support the notion that P53 inactivation plays a direct role in clonal instability in CLL, and this is in keeping with the wild-type P53 repair response being involved in cell-cycle arrest, induction of apoptosis and the DNA damage response (see Introduction, Sections 1.9 to 1.11). This thesis examines the possible involvement of mutations in other DNA maintenance genes involved in clonal instability by making use of high-throughput NGS to determine the genomic integrity of these genes in high-risk CLL cases. It was therefore necessary to identify a panel of important candidate genes that could contribute to genomic instability and then develop sequencing strategies that would allow them to be compared across multiple CLL cases. Target enrichment strategies based on hybrid capture (see Section 1.15) that could produce enriched DNA ready for NGS (see Section 1.17.1) was therefore developed and validated, which is the subject of this chapter.

3.2. Results

3.2.1. Candidate DNA Maintenance Genes

A total of 194 DNA maintenance genes were identified from the literature (see Introduction) and included in this study. The goal was to produce a canonical set so that all possibilities could be screened for mutations by sequencing. Their combined exon sequences totalled 499 kbp (see Table 3-1 for list and details) which summarises the output report from the SureDesign in terms of the number of genes targeted and the repair process in which they were involved. Genes included those for repair of single- and double-strand damage (59, 61, 78, 266-275), NER, BER and MMR, HR and NHEJ, other enzymes part of the repair process, damage response and cell-cycle control.

Table 3-1. Human DNA Maintenance Genes Chosen for This Study (61, 266, 276).

DNA repair gene	Genes	No. of genes
BER	MBD4, MPG, MUTYH (MYH), NEIL1, NEIL2, NEIL3, NTHL1 (NTH1), OGG1, SMUG1, TDG ,UNG	11
Other BER and strand break-joining factors	APEX1 (APE1), APEX2, APLF (C2ORF13), LIG3, PNKP, XRCC1	6
PARP enzymes that bind to DNA	PARP1 (ADPRT), PARP2 (ADPRTL2), PARP3 (ADPRTL3)	3
Direct reversal of damage	ALKBH2 (ABH2), ALKBH3 (DEPC1), MGMT	3
Repair of DNA-topoisomerase crosslinks	TDP1, TDP2 (TTRAP)	2
MMR	MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, PMS1, PMS2, PMS2P3	10
NER	CETN2, DDB1, DDB2 (XPE), RAD23A, RAD23B, RPA1, RPA2, RPA3, XPA, XPC -TFIIH: CCNH, CDK7, ERCC1, ERCC2 (XPD), ERCC3 (XPB), ERCC4 (XPF), ERCC5 (XPG), GTF2H1, GTF2H2, GTF2H3, GTF2H4, GTF2H5 (TTDA), LIG1, MNAT1 -NER-related: ERCC6 (CSB), ERCC8 (CSA), MMS19L (MMS19), XAB2 (HCNP)	29
HR	BRCA1, DMC1, EME1 (MMS4L), EME2, GEN1, GIYD1 (SLX1A), GIYD2 (SLX1B), MRE11A, MUS81, NBN (NBS1), RAD50, RAD51, RAD51L1 (RAD51B), RAD51L3 (RAD51D), RAD52, RAD54B, RAD54L, RBBP8 (CtIP), SHFM1 (DSS1), XRCC2, XRCC3	21
Fanconi anaemia	BRCA2 (FANCD1), BRIP1 (FANCJ), BTBD12 (SLX4), (FANCP), FAAP24 (C19orf40), FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG (XRCC9), FANCI (KIAA1794), FANCL, FANCM, PALB2 (FANCN),	17

	RAD51C (FANCO)	
NHEJ	DCLRE1C (Artemis), LIG4, NHEJ1 (XLF, Cernunnos), PRKDC, XRCC4, XRCC5 (Ku80), XRCC6 (Ku70)	7
Modulation of nucleotide pools	DUT, NUDT1 (MTH1), RRM2B (p53R2)	3
DNA polymerases (catalytic subunits)	MAD2L2 (REV7), PCNA, POLB, POLD1, POLE, POLG, POLH, POLI (RAD30B), POLK (DINB1), POLL, POLM, POLN (POL4P), POLQ, REV1L (REV1), REV3L (POLZ)	15
Editing and processing nucleases	APTX (aprataxin), EXO1 (HEX1), FAN1 (MTMR15), FEN1 (DNase IV), FLJ35220 (ENDOV), SPO11, TREX1 (DNase III), TREX2	8
Ubiquitination and modification	HLTF (SMARCA3), RAD18, RNF168, RNF4, RNF8, SHPRH, UBE2A (RAD6A), UBE2B (RAD6B), UBE2N (UBC13), UBE2V2 (MMS2), SPRTN (C1ORF124)	11
Chromatin structure and modification	CHAF1A (CAF1), H2AFX (H2AX), SETMAR (METNASE)	3
Genes defective in diseases associated with sensitivity to DNA-damaging agents	ATM, BLM, RECQL4, TTDN1 (C7orf11), WRN	5
Other identified genes with known or suspected DNA repair function	DCLRE1A (SNM1), DCLRE1B (SNM1B), HELQ (HEL308), OBFC2B (SSB1), RDM1 (RAD52B), RECQL (RECQ1), RECQL5, RPA4, PRPF19 (PSO4)	9
Other conserved DNA damage response genes	ATR, ATRIP, CHEK1, CHEK2, CLK2, HUS1, MDC1, PER1, RAD1, RAD17 (RAD24), RAD9A, TOPBP1, TP53, TP53BP1 (53BP1), RIF1	15
Cell-cycle control	CDK1, CDK2, CCND1, CCND2, CCND3, CCNE1, CCNE2, CCNA1, CCNA2, CCNB1, CCNB2, CCNB3, CDK4, CDK5, CDK6,	15
Cytidine deaminase	AICDA	1
Total	--	194 genes
Total size	--	499.047 kbp

3.2.2. cRNA Probe Design

A panel of cRNA probes for the exonic sequence of the selected DNA maintenance genes

was determined using the approach described in Materials and Methods section 2.8 using

SureDesign. Increasing the average number of probes per target increases stringency at

the expense of reducing the total possible number of target bases. Another aspect to

stringency is the possibility of off-target inclusion through sequence similarity or

repetitive sequences. Stringency was therefore modelled to determine an optimum and

included the possibility of boosting the number of probes for targets for which coverage

was low. Table 3-2 lists the effects of decreasing stringency from high to no masking. Low

stringency masking was predicted to require 15,166 probes and provide 680.6 Kbp of

target bases for an average of two-fold coverage.

Table 3-2: Output of stringency test modelling for design of cRNA probes targeting the
panel of DNA maintenance genes.

Stringency masking	High	Moderate	Low	No masking
% Coverage	98.94	99.18	99.82	100
Off-target number of bases	N/A	3.38Kb	8.58Kb	10.58Kb
Total number of probes	15,010	15,064	15,166	15,184
Total number of base pairs (Kbp)	672.023	657.4	680.6	682.6

Design results for different masker stringencies (maximum boost and 2X region coverage).

Low masking stringency was selected when designing the probes.

Given that over 99.8% of the target sequences were expected to be included, the cRNA

probes corresponding to the design were obtained. The relative coverage of the

SureDesign at low stringency masking in terms of the number of exons, probes, size of the

target region and total region targeted is shown in appendix table 1 for each of the

classes of DNA maintenance genes included.

3.2.3.   Clinical Samples

The 10 cases of CLL of the CLL206 trial were as described in the Materials and Methods

Section 2.3.1. Trial data concerning TP53 status as determined by FISH for the short arm

of chromosome 17 (17p) and FASAY (46, 47) were available. Screening for TP53 mutations

was performed by Sanger sequencing.

3.2.4. Genomic DNA extraction from CLL samples

The AllPrep DNA/RNA Mini Kit (see Materials and Methods section 2.5.1) was used to extract gDNA from samples for each case. Table 3-3 summarises the gDNA sample information and status of the original cases. The FISH results showed that the samples had a 17p deletion and FASAY demonstrated the negative functional activity of P53, confirming they were positive for TP53 mutations, and clonal TP53 mutations alleles were confirmed by Sanger sequencing. TP53 gene mutations in samples CLL03, CLL06 and CLL08-CLL10 were located within the TP53 DBMs. Pretreatment details available for certain cases revealed that many patients had received intensive treatment regimens with DNA-damaging agents, such as fludarabine and cyclophosphamide. Table 3-3: Samples Information for gDNA of the 10 Chosen CLL Samples Including a List of Clonal TP53 Mutations detected by Sanger Sequencing.

Sample no.	Code no.	FASAY	FISH	Mutation on codon	Pre-treatment	End treatment	DNA amount	DNA quality	Lymphocyte count
2631	Liv_01	100%	96%	Arg 213 Pro	N/A	CR	8.5ug/85µl	1.87	188
2681	Liv_02	96.20%	91%	Tyr 220 Cys	N/A	CR	12.9ug/85µl	1.91	247
2600*	Liv_03	91.40%	N/A	Arg 273 His	N/A	CR	21.6ug/86µl	1.87	71.3
2766	Liv_04	77%	N/A	Phe 109 Val	Chlo x12 FC x6	N/D	6.1ug/55µl	1.87	13.1
2640	Liv_05	91.10%	100%	Ile 162 Asn	FC x1 FCR x3 Met-Pred x6	SD	9ug/85µl	1.88	31.6
2657*	Liv_06	95%	90%	Met 246 Val	N/A	N/A	11.4ug/85µl	1.91	172.6
2642	Liv_07	100%	N/A	Tyr163 Cys	Chlo x2 Flu x2	PR	19.6ug/93µl	1.84	88.2
2554*	Liv_08	95%	N/A	Cys 238 Tyr	Chlo x2 Flu x6 CHop x6 Ritn x4	PD	6.9ug/80µl	1.89	12.7
2621*	Liv_9	100%	90%	Arg 248 Trp	Chlo/Pred x4	N/A	9.5ug/91µl	1.86	N/A
2550*	Liv_10	100%	72%	Try 175 His	Chlo x4 FC x3 Flu x6 Autograft x1	CR	23.1ug/84µl	1.86	53.2

* Samples with TP53 mutations inside the DBM. gDNA extraction and FASAY were performed by Dr Gillian Johnson. Genomic Integrity of Purified CLL DNA

The size of the gDNA yielded from each case was estimated by agarose gel

electrophoresis as described in Materials and Methods, section 2.5.1, Figure 3-1. In each

case, a tight band of DNA was observed close in size to the 23.1 kbp marker, indicating

that it was off from acceptably high molecular weight with undetectable degradation.

Nanodrop spectroscopy (see Materials and Methods, section 2.6.1) showed that purity

was also robust with an average $A_{260/280}$ ratio of 1.88 (277).
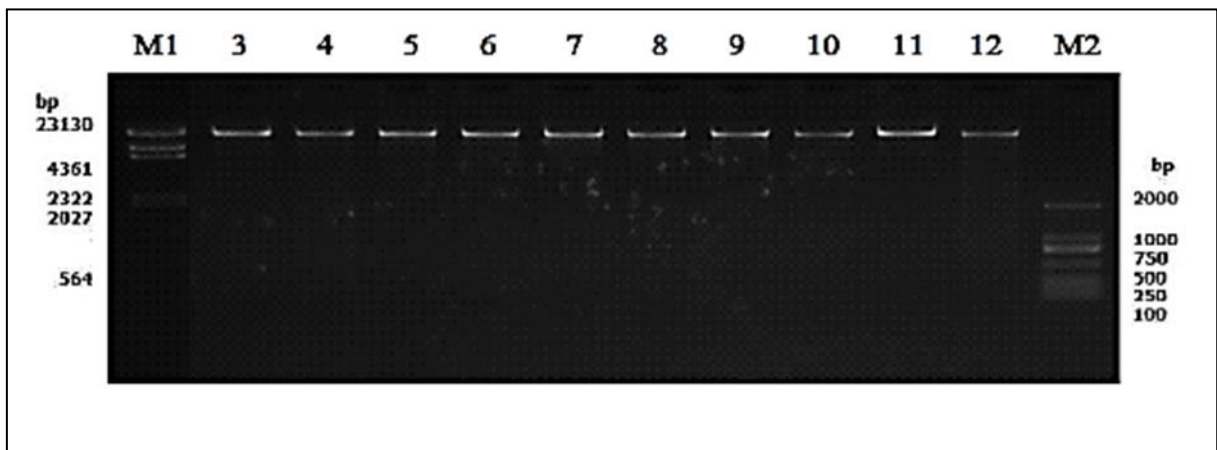
Figure 3-1: 1% Agarose Gel of CLL gDNA. Lanes 3-12: CLL01-CLL10 samples, respectively. M1 is λDNA-Hind III digest (Takara, Kusatsu, Japan) and M2 is D2000 ladder (Tiangen Biotech, Beijing China). Nanodrop (OD 260/280: 1.8-1.9).

3.2.5. Shearing of CLL gDNA

NGS requires sample DNA to be clonally amplified with a size range of ~200 bps. The

target DNA is therefore randomly fragmented, and adapters ligated to each end so that

PCR can proceed using primers corresponding to the adapter sequences. DNA

fragmentation was therefore performed enzymatically as described in Materials and

Methods, section 2.7.4, using Ion Shear, which cleaves DNA pseudo-randomly and repairs

the ends ready for ligation. A control gDNA sample was subjected to different periods of

shearing (5, 10, 15 and 18 minutes) for determining the optimum required to produce the

required average DNA fragment size. The sizes of the resulting DNA fragments were then

measured with a Bioanalyzer as described in Materials and Methods, section 2.6.4. As

expected, the longer the digestion, the smaller the size of DNA (Figure 3- 2). The 15-

minute digestion produced an average size of DNA fragments of 280 bp, which was

selected as optimum for the downstream adaptor ligation.
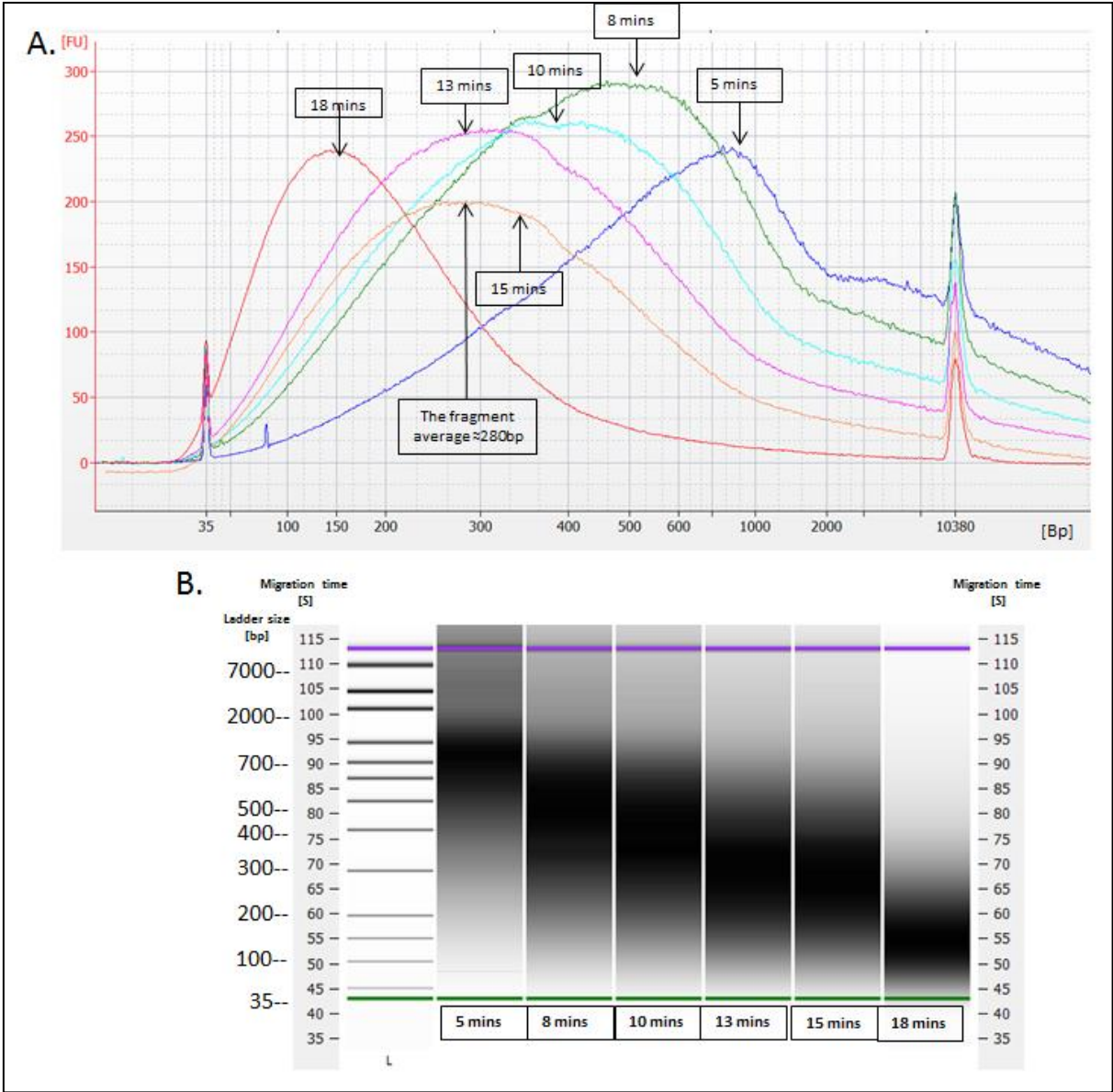
Figure 3- 2: Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of Ion-Sheared
Genomic Control DNA. A. Electropherogram comparison of gDNA sizes following different
fragmentation time reactions of 5, 10, 15, and 18 minutes. B. The corresponding gel

representation. Bioanalyzer 2100 HS DNA ladder marker was included. The average fragment of choice was 260bp (arrowed).

Having established the optimum incubation time for producing fragments of the required optimum size using control gDNA, fragmentation was repeated using the gDNA from the CLL cases. Resultant fragments in each sample were analysed for average size using the Bioanalyzer as in the previous (see Figure 3-3).
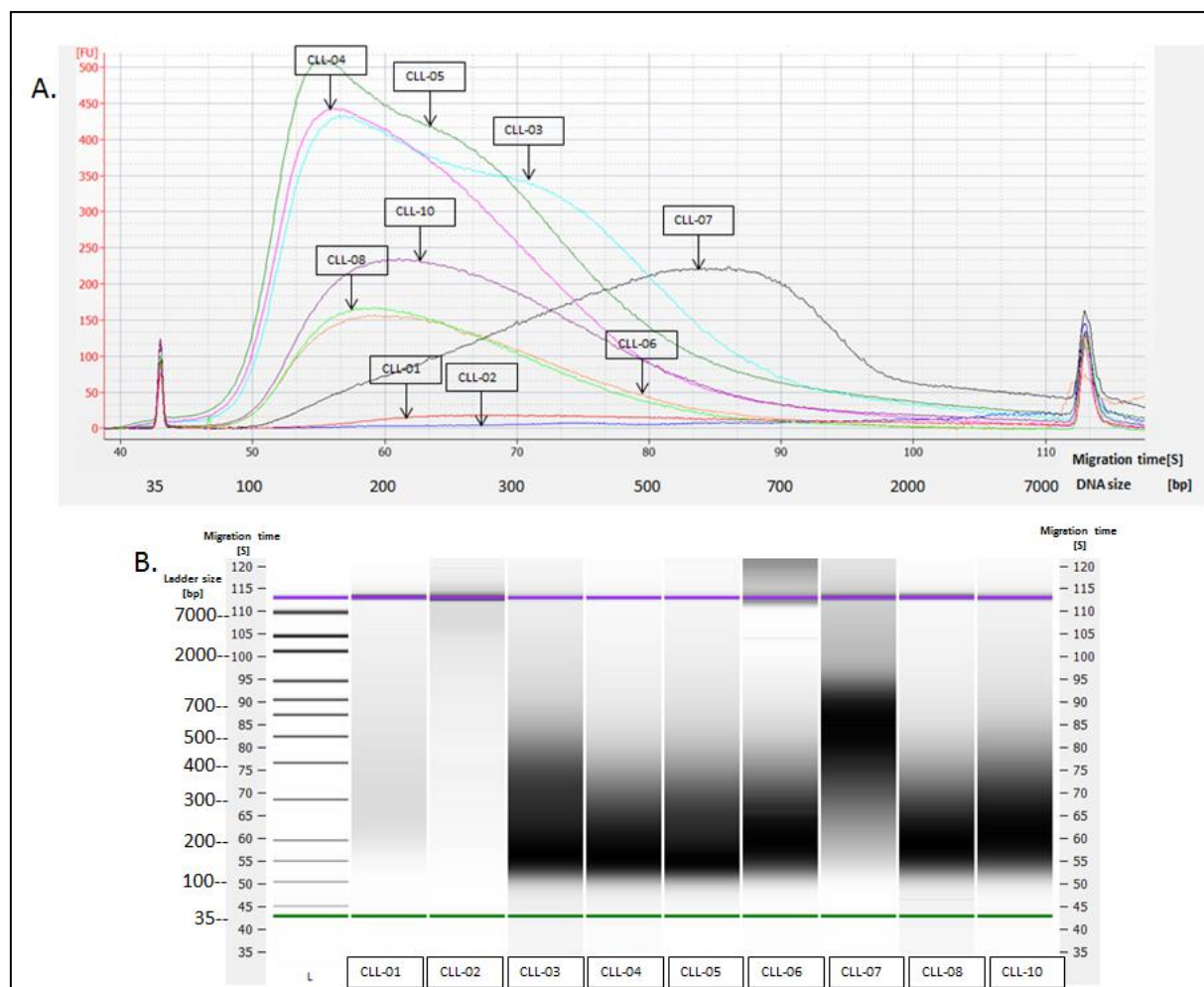
Figure 3-3: Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of 10 Samples of Ion-Sheared CLL gDNA. A. Electropherograms for samples CLL-01, CLL-02, CLL-03, CLL-04, CLL-05, CLL-06, CLL-07, CLL-08 and CLL10. B. Corresponding gel representation.

Significant variation was observed both in the amounts of DNA per lane and the average sizes across lanes. Variations in amounts are to be expected because the Bioanalyzer uses electrophoretic loading, which is highly sensitive to starting conditions and can lead to varying amounts being loaded to different lanes of the chip. It had been previously established that the starting amounts of gDNA were the same in each case. CLL-02 and CLL-07 showed signs of having been fragmented the least. Again, loading could have been an issue resulting in drag in these lanes and the high molecular weight material in CLL-06 despite the preponderance of small fragments in this sample is consistent with this possibility. However, poor mixing of Ion Shear could have produced a similar effect. It was considered safe to proceed with these samples because the molarity of the small fragments in all cases would have been very high and this was expected to be sufficient to support the clonal amplification required for sequencing. Adaptering was therefore performed.

3.2.6. DNA Adapter Ligation

Adapter ligation to the control sample of gDNA was performed as described in Materials and Methods, section 2.7.5. It was important to achieve a high ratio of adapter ends to gDNA fragment ends so that self-ligation of gDNA was avoided because it was outcompeted by the adapters, otherwise chimeric gDNA fragments could have resulted. The sizes of DNA fragments in pre- and post-ligated samples were therefore compared using the Bioanalyzer as earlier (see Figure 3-4). Adaptor-ligated DNA had a larger DNA size than non-adaptor-ligated DNA corresponding to a shift of ~150 bps as would be expected from ligation of the 76-86 bp adaptors (for both DNA strands) to each end of the fragmented DNA. Two different adapters are used to introduce directionality into a

subset of the ligated fragments, those receiving one type of adapter on one of their ends and the opposite adapter on their other end. This is necessary to ensure that the sequencing priming step is unidirectional on each template. Both images also show two peaks of free adaptors in adaptor-ligated DNA samples.
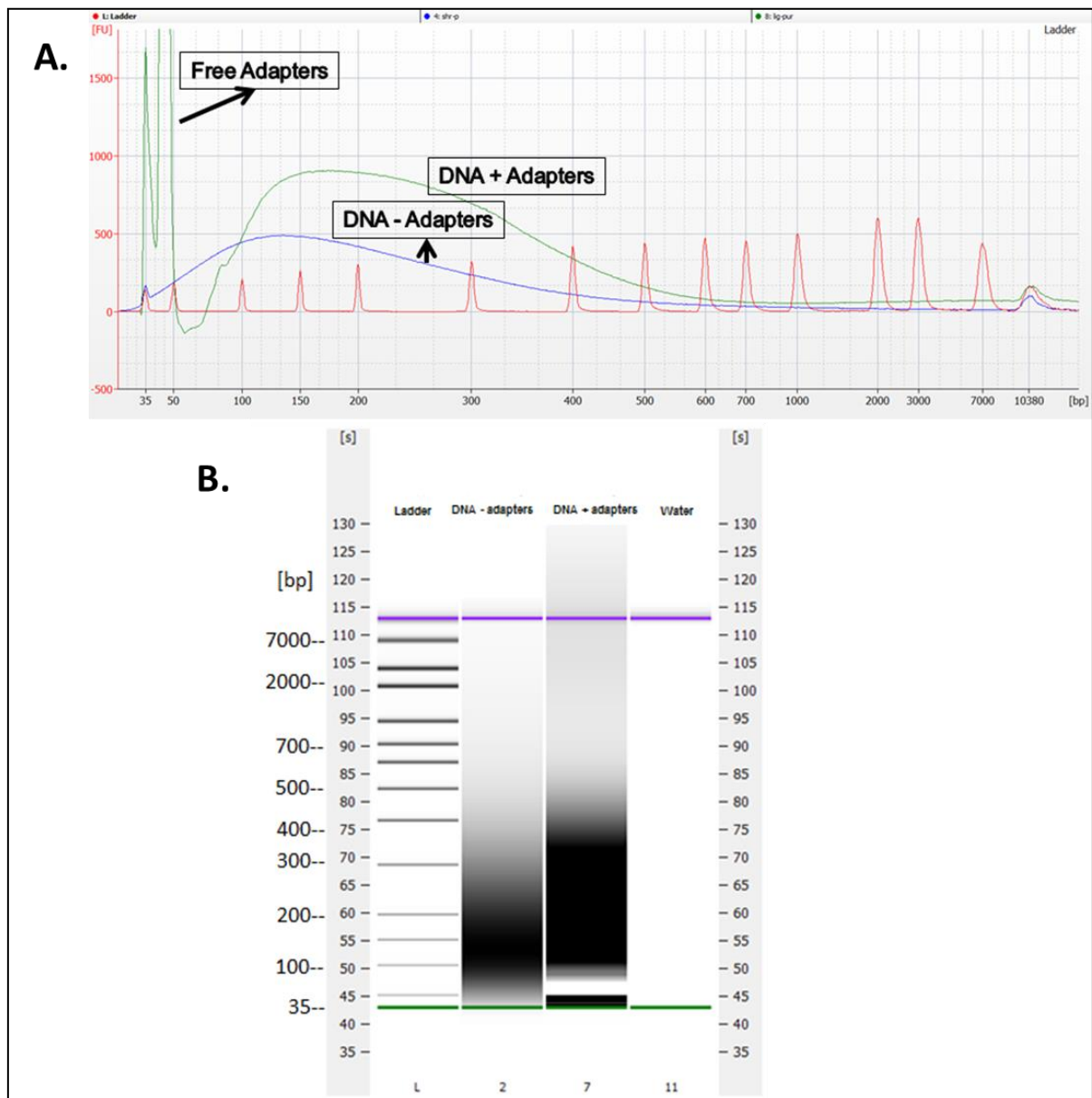
Figure 3-4: Bioanalyzer 2100 and Bioanalyzer HS-DNA chip size analysis of adapter ligation to fragmented control gDNA. A. Electropherogram of Bioanalyzer HS DNA ladder (red peaks), sheared gDNA – ligated adapters and sheared gDNA + ligated adapters. B. Corresponding gel representation.

Adapter ligation was repeated for the CLL gDNA samples (data not shown) and all samples were employed for prehybridization amplification.

3.2.7. Pre-Hybridisation Amplification

Further confirmation that adapter ligation had taken place successfully was achieved using PCR to amplify the adapted fragments. In addition, this provides a greater concentration of targets for target selection by hybridisation. PCRs were performed using a representative adaptered gDNA as described in Materials and Methods, section 2.7.1, using Herculase II as the DNA polymerase and forward and reverse primers corresponding to the two different types of the adapter. Controls included the use of a single primer. The products of the PCR reactions were analysed by the Bioanalyzer and also agarose gel electrophoresis as described in Materials and Methods, section 2.6.3. Figure 3-6 depicts the PCR efficacy of using both forward and reverse PCR primers and adapter-ligated DNA versus single primer as a negative control as determined by the Bioanalyzer. The significant product was only produced by the adapter-ligated DNA sample as expected. The PCR yield was better for the PCR using both PCR primers when compared to a single primer.

PCR	DNA	DNA+ Adaptor
No primers (prior PCR)		
Forward primer	--	
Reverse primer	--	
Forward + Reverse primers		
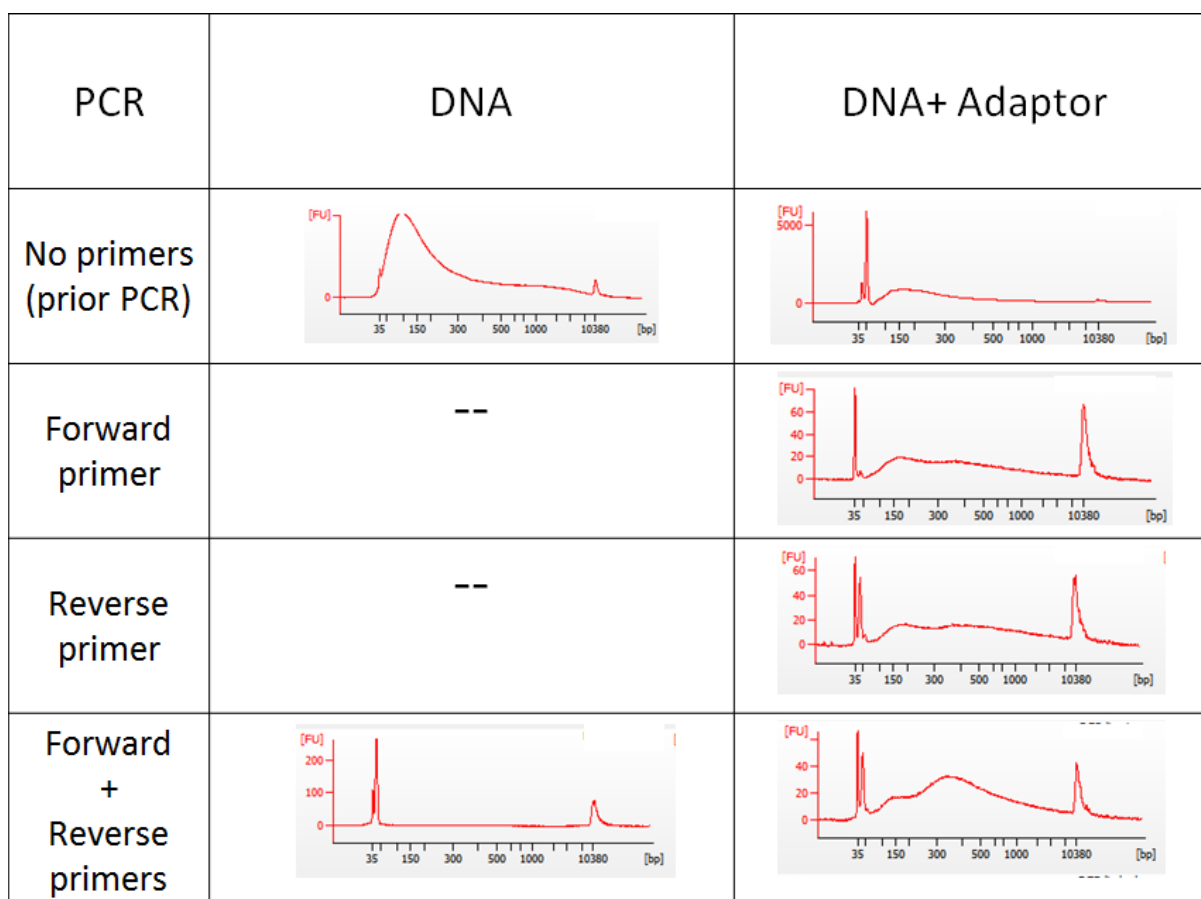
Figure 3-5: Electropherograms from Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of PCR Amplification of Adaptered gDNA Fragments by the Forward Adapter (F), Reverse Adapter (R) or both F and R Adapters.

The effect of size selecting the adaptered gDNA fragments prior to PCR was also determined. Size selection was performed using eGels as described in Materials and Methods, section 2.10. Three size classes of ~100, ~200 and ~300 bps were obtained and used for the PCRs, which were analysed by agarose gel electrophoresis as in Materials and Methods, section 2.10 (see Figure 3-6). The result also specifically depicts PCR yield when adapter-ligated DNA and two PCR primers, versus a single primer, were utilised. A faint PCR product was seen in the agarose gel when the forward primer was used. This is expected because fragments that receive the same adapter at each end are normally excluded from PCR because their ends become complementary as a result, causing the

single-stranded DNA to hairpin and self-anneal without a primer. However, break-through

can occur.
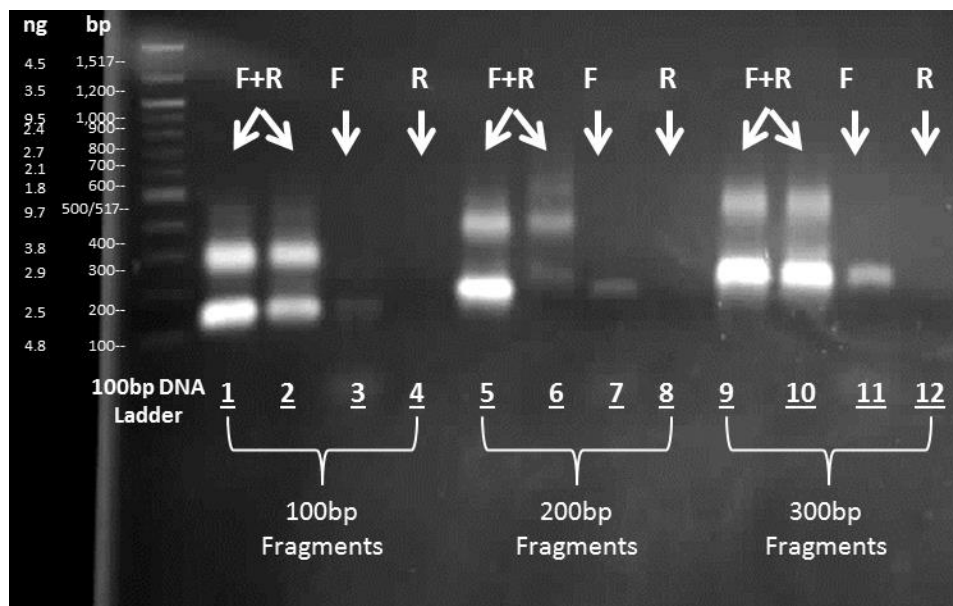
Figure 3-6: Agarose Gel Electrophoresis of Size-Selected Adaptered gDNA Fragments Amplified by PCR. Three different sized adaptered gDNA fragments were used (~100 bp, ~200 bp and ~300 bp). Four different PCR reactions each were performed: primer forward plus reverse primers, marked F+R on two lanes, forward-only primer, marked (F) and reverse-only primer, marked (R). 100 bp DNA ladder was employed as a marker. The agarose gel is 1.8% in TBE buffer.

PCR amplification was repeated for the adaptered CLL gDNA fragments, which were

analysed using the Bioanalyzer as in the previous (see Figure 3-7). Fragment sizes ranged

from 200-700 bp after PCR as expected after the DNA adapter ligation. Some samples

exhibited weaker amplification yielding 150-600 pg per µl. The target amount for the

cRNA target selection by hybridisation is 1 ug/3.4 µl. It was therefore necessary to use the

vacuum drier as described in Materials and Methods, section 2.7.6, to concentrate

samples with concentrations that were below the amount required. These could then be

used for targeted capture of DNA maintenance gene regions by cRNA.
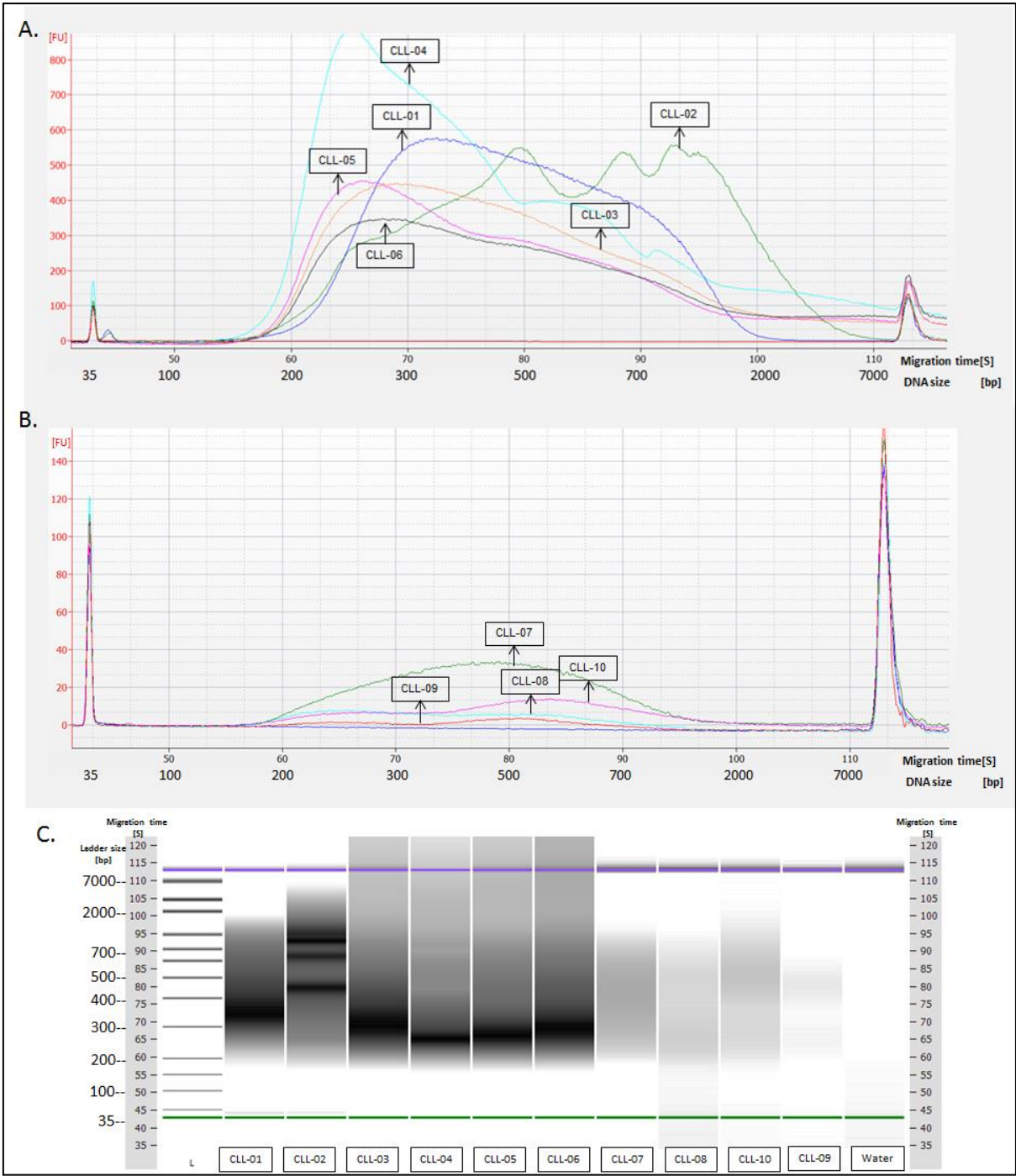
Figure 3-7: Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of PCR Amplification Products from Adaptered CLL gDNA Fragment PCRs. A. Electropherograms of cases CLL-01 to CLL-06. B. Cases CLL-07 to CLL-10 (1/10 dilution). C. Corresponding gel representation.

3.2.8. cRNA to CLL gDNA Hybridisation and Capture

cRNA probes were hybridised to target sequences in the amplified CLL gDNA and then captured using biotinylated beads according to Materials and Methods, sections 2.7.6 and 2.7.6.2, respectively. The Bioanalyzer was made use of as described for verification, post-hybridisation, the integrity of the DNA for CLL-03 as a representative sample (see Figure 3-8). The 200-500 bp size distribution of the DNA was found to be correlated with that in the pre-hybridisation analysis, showing that it had not been degraded during hybridisation and target capture was therefore performed.
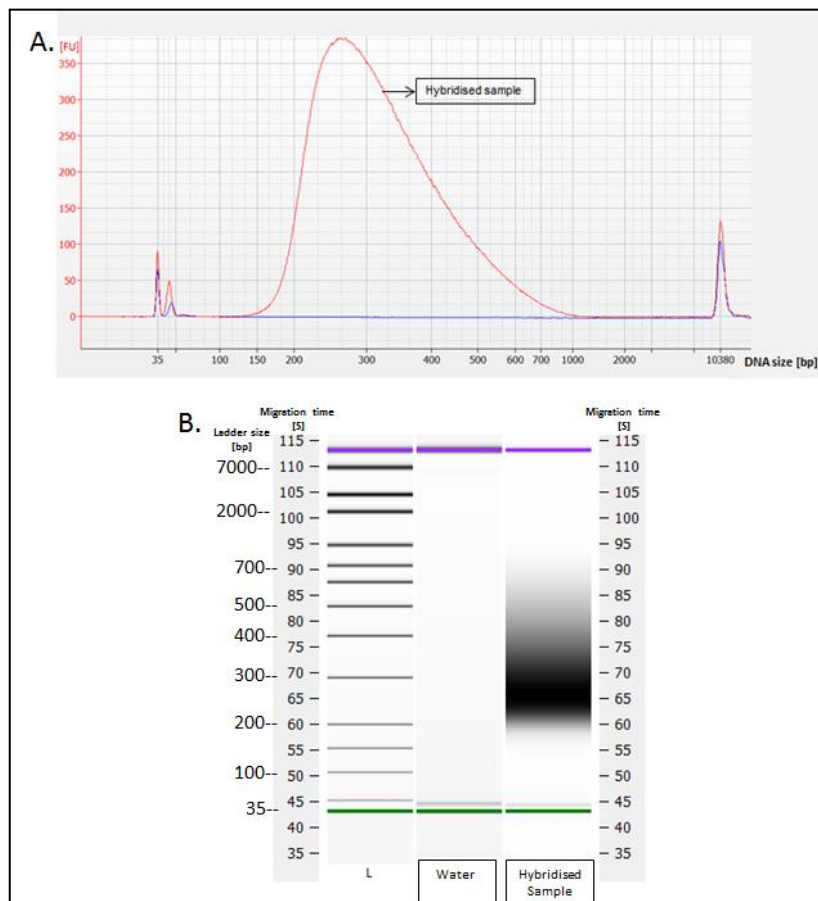
Figure 3-8: Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of cRNA Hybridised gDNA Sample, CLL-03. A. Electropherograms of cRNA library (L) with Bioanalyzer HS DNA ladder, negative control (Water) and hybridised sample. B. Corresponding gel representation.

3.2.9. Size Selection and PCR Amplification of cRNA/CLL gDNA Hybrids

After target capture, the DNA in each reaction was PCR amplified as before and then size-selected using eGels as described in Materials and Methods, section 2.10. Samples of the purified material were analysed using the Bioanalyzer. Figure 3-9 portrays the amplified DNA after size-selection for sample CLL-06. A narrow peak at 318 bp was seen, just as expected.
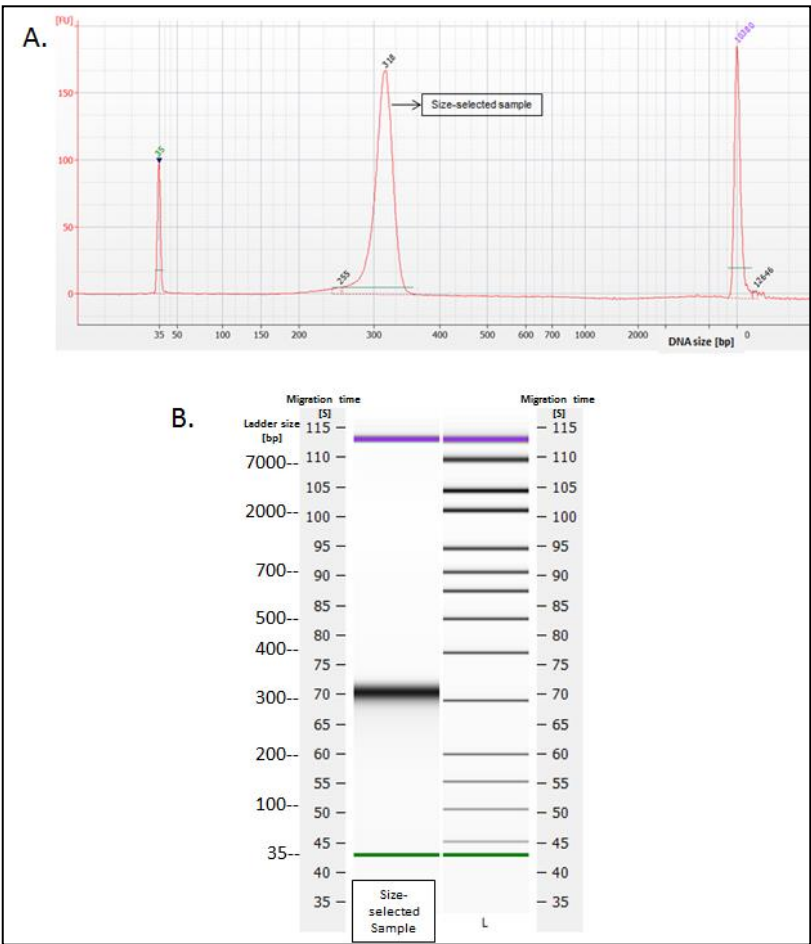
Figure 3-9: Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of Post-Hybridised, Size-Selected and PCR-Amplified Sample CLL-06. A. Electropherogram of size-selected and PCR-amplified DNA sample at 318 bp; Bioanalyzer HS DNA ladder is included. B. Corresponding gel representations.

Figure 3-10 shows the Bioanalyzer analysis for all 10 CLL samples. A narrow fragment size range between 200-350 bp was obtained and this was at the acceptable upper limit for

PGM sequencing. Differences in size distribution between samples were expected

because of small variations in timing, naturally occurring when the samples were

collected. The material produced in each case was therefore subjected to NGS.
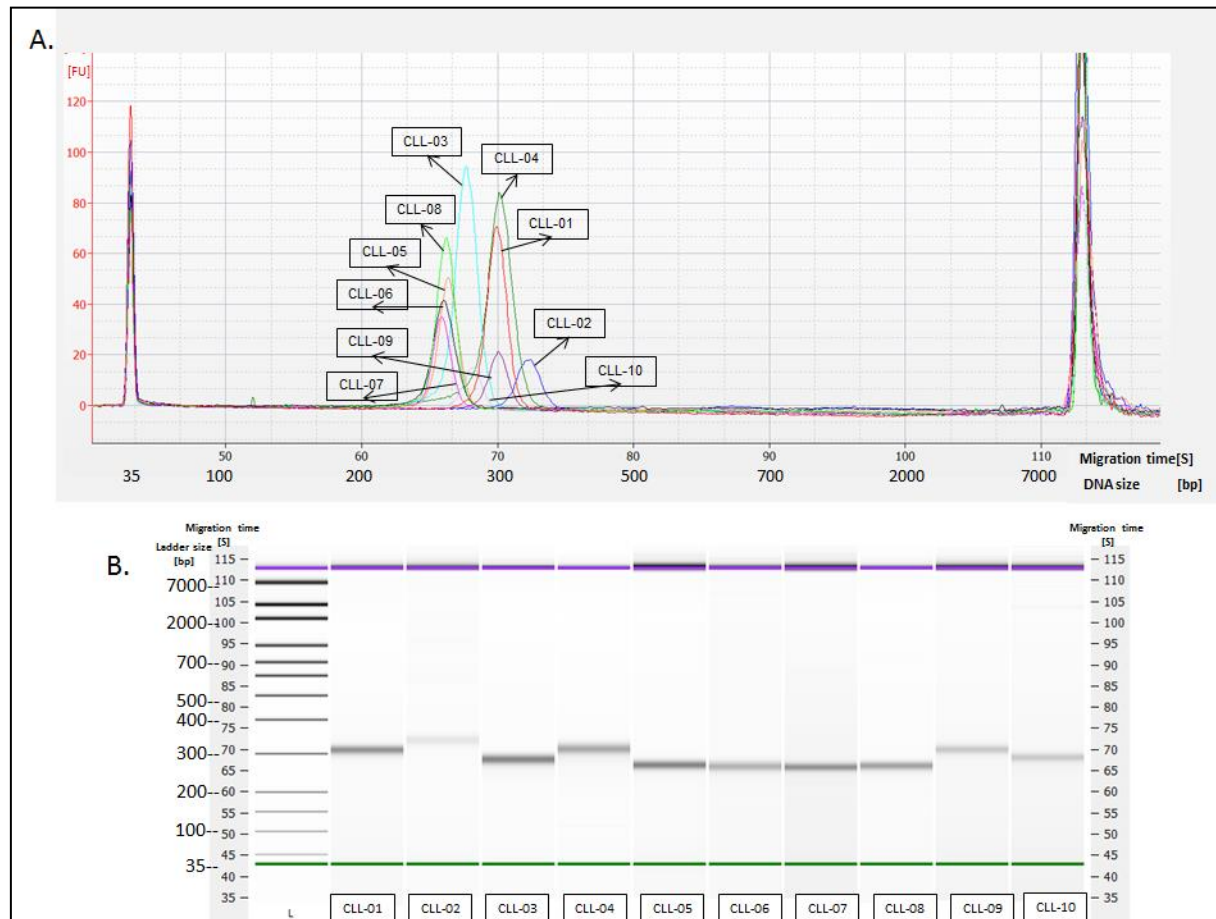
Figure 3-10: Bioanalyzer 2100 and Bioanalyzer HS-DNA Chip Size Analysis of Post-Hybridised, Size-Selected and PCR-Amplified Samples, CLL-01 to CLL-10. A. Electropherograms. B. Corresponding gel representations.

3.2.10. Initial NGS Analysis of cRNA Maintenance Gene Sequences in CLL Cases

PGM Sequencing was performed for the cRNA enriched CLL DNA samples using Ion Chip

316v2 as described in Materials and Methods, section 2.11. Sequence data was aligned to

hg19. An overview of the sequence data for all of the samples is seen in Table 3-4.

Table 3-4: Coverage Analysis for Ion Torrent PGM Sequencing Using Ion chips 316v2 to Read cRNA Enriched gDNA CLL Samples CLL01 to CLL10. Analysis includes the number of mapped reads (defined as the number of reads that were mapped to the full reference human genome, hg19), the percentage of reads on-target (defined as the percentage of mapped reads that were aligned over a target region), the base coverage depth (defined as the average base coverage depth over all bases targeted in the reference) and the uniformity of base coverage (defined as the percentage of target bases covered by at least 20% of the average base read depth) were utilised for sequencing. STDEV refers to standard deviation.

Sample	CLL-01	CLL-02	CLL-03	CLL-04	CLL-05	CLL-06	CLL-07	CLL-08	CLL-09	CLL-10	Mean	STDEV
Number of mapped reads (million reads)	3.8	3.1	1.9	1.75	2	1.79	2.92	3	3	3.3	2.7	0.7
Percent reads on - target (%)	49.5%	54.19%	43.7%	41.6%	26.4%	21%	45.6%	63.8%	64.5%	67%	47.7%	14.8%
On- target Base coverage Depth (x)	435	400	195	152.7	61.9	75.7	286	464.9	412.4	490.2	297.4	156.2
Uniformity of base coverage (%)	97.4	97.2	97.9	98.4	97.5	98.4	58.7	95.4	95.4	95.1	93.1	11.5

The mean number of reads mapped to hg19 was 2.7 m ± 0.7 m reads. This indicated

robust utilisation of the chip capacity, which is 2-3 m reads. Figure 3-11 portrays a heat

map distribution representing the amount of data produced from across the chip. On-

target coverage averaged 47.7% ± 14.8% of the total mapped bases. This was significantly

higher than the expected number of off-target bases predicted from the low-stringency

probe design utilised. Technical issues, like, for example, washing steps or adsorption,

may have been an issue. However, the mean base on-target coverage depth was 297.4 ±

156.2 times, with values ranging from 62-465 for all samples, including those with weaker

PCR amplification. This could detect alternative allele frequencies (AF) ≥5%,(278-281),

which was sufficient for the overall purposes of the investigation (282). Additionally,

samples CLL01-06 were multiplexed and ran pair-wise and in duplicate to increase

coverage depth.
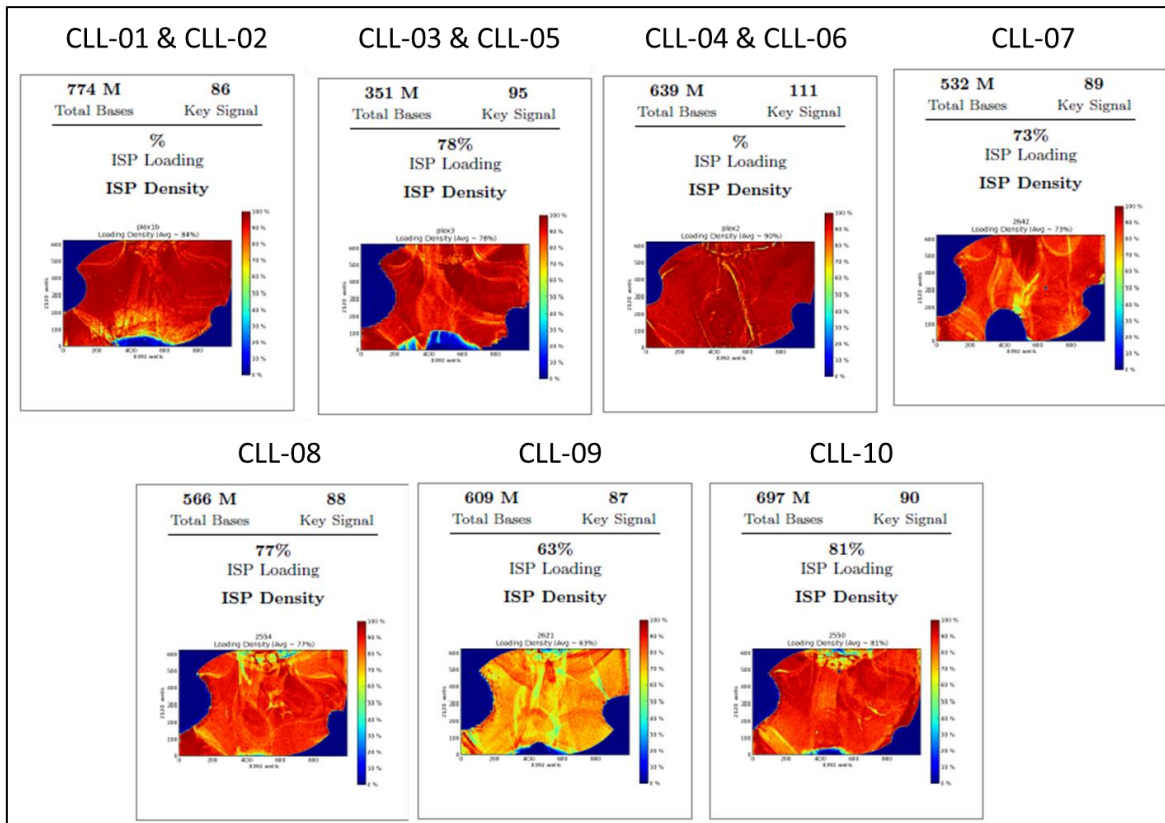
Figure 3-11: PGM Run Summary. Representative heat maps showing density distribution of data generating positive Ion Sphere Particles (ISPs) during PGM sequencing run of cRNA-enriched gDNA CLL samples. ISP is defined as streptavidin-coupled Dynabeads® that were used for enriching DNA templates for clonal amplification.

3.2.11. Data Validation

As the approach was new to the laboratory and the cRNA was of a novel probe design, it was appropriate to determine their reliability. Samples CLL-01, CLL-02, CLL-03 and CLL-05 were therefore repeated. Samples were multiplexed in pairs and then sequenced twice using 316v2 chips and the results are summarised in Figure 3- 12, which compares the read depths for AFs detected per sample in each run. The number of reads per variant per run plotted as a straight line at a 45° slope passing through the origin clustering on 50% and 100% allele frequency were as expected for reproducibility between runs. Most of the expected variants are SNPs and either homo- or heterozygotes
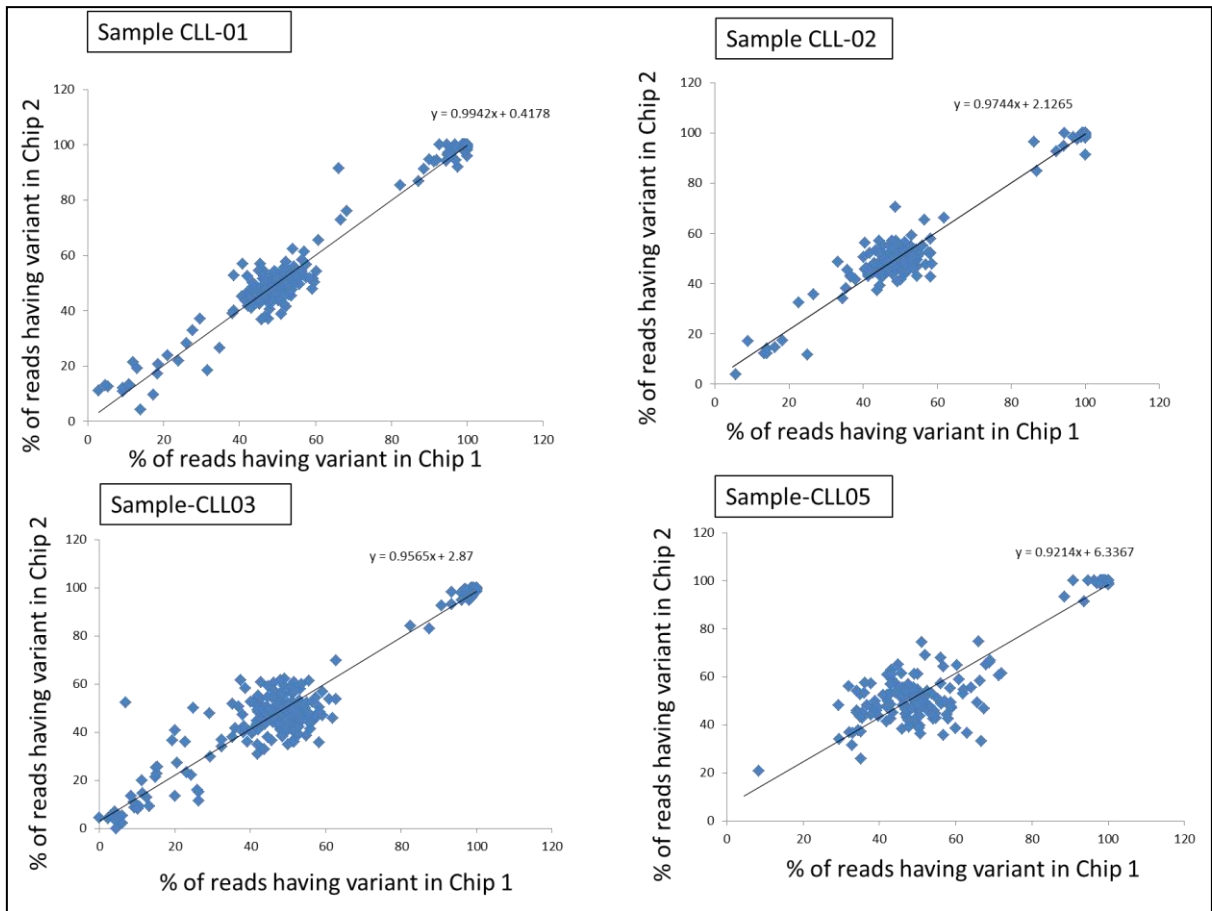
Figure 3- 12: Variant Call Reproducibility of Ion Torrent PGM Analysis of cRNA-Targeted CLL gDNA.

Plots A-D represent the linear correlation of variant calls on duplicate sequencing runs for samples CLL-01, CLL-02, CLL-03 and CLL-05, respectively. TVC 4.2 online software was used for variant calls. These plots specifically represent the percentage of reads having variants (allele frequency). The x-axis is for the chip 1 variants and Y-axis for chip 2 variants.

3.3. Discussion and Conclusion

Ion Torrent PGM was utilised to sequence the coding regions of 194 DNA maintenance genes enriched by cRNA capture from CLL gDNA samples. This involved an investigation using a PGM instrument in-house and involved identifying a canonical set of DNA

maintenance genes, designing corresponding cRNA probes, preparing samples, target-enriching the samples, sequencing the enriched targets and post-sequencing analysis.

The Ion Torrent was expected to have a number of advantages. The average read capacity was well-matched for this study; turn-around time was rapid, sequencing costs were feasible and post-sequencing data processing could be performed locally. However, the Ion Torrent PGM has an inherent weakness - it is unable to accurately calculate HP length and this may cause false positive errors, particularly for indels   (283). This problem arose from HPs of ≥ 4 repeated nucleotides, especially with G or C repetitions and increasing localisation away from the beginning of sequencing reads (283-285). This will be discussed in Chapter 4.

With respect to methodology development, it was crucial to develop and optimise the enrichment process to ensure the satisfactory yield of targeted DNA. The methodology development largely involved optimising DNA shearing, DNA purification, PCR amplification conditions and target enrichment. The Bioanalyzer results across the enrichment process demonstrated that samples were enriched in the specific DNA size range of approximately 300 bp. Most gDNA was successfully sheared into the range of 100-500 bp (average size: ~200 bp) (Figure 3-3). Sheared DNA was ligated to adapters (Figure 3-4), successfully amplified (Figure 3-6), hybridised (Figure 3-8) and size-selected (Figure 3-10). During sample processing, continuous quality control using Bioanalyzer across the enrichment process demonstrated that samples were successfully enriched for the specific DNA size range of around 300 bp - the 300 bp is sequenceable in PGM and longer reads improve sequencing accuracy (286, 287).

Concerning the custom cRNA panel targeting 194 DNA maintenance genes, 71% of 702,619 bp probes were on-target regions and 29% was off-target. This was anticipated because the design algorithm increased the size and number of the off-targets probes using adjacent regions to indirectly increase the representation of on-target. Experimentally, targeted regions were successfully enriched, achieving 99.7% high coverage of the total 2,786 exons for on-target coverage (Table 3-4). The average number of sequencing reads was 2.7 ± 0.7 m, which is within the upper limit of the Ion chips 316v2. Coverage analysis revealed unexpectedly lower than average on-target reads compared to off-targets (47.7% ± 14.8%). It had been expected that the design of the probe would compromise about 30% of the total reads for on-target, but there was a higher number of off-targets detected than expected owing to the low stringency of the probe design that increased PCR amplification bias and reduced hybridisation effectiveness (282, 288). Improvements could possibly be made by increasing the selected design option away from the lowest stringency masking to improve the proportion of on-target enrichment (289, 290), but this might have reduced the proportion of targets detected. Technical improvements, for example, rose numbers of washes or reduction of adsorption opportunities may have diminished the number of off-target reads. Overall, the design was satisfactory for the intended purpose, namely to detect acquired variants that may contribute to genetic instability in CLL.

The CLL samples had TP53 mutations/deletions and had been treated with multiple therapies, including DNA-damaging agents. Patients with TP53 mutations are expected to be chemo-refractory (30-32) as seen in the end treatment of various patients. Further analysis of the sequence data in this regard is presented in Chapter 4.

4. Chapter 4: Application of a Targeted NGS Method to Identify Mutations in DNA Maintenance Genes of CLL Samples with a Mutated/Deleted TP53 Gene

4.1. Introduction

In the Results of Chapter 3, a high-throughput methodology was successfully established for target enrichment of literature-selected DNA maintenance genes for NGS and applied to a small cohort of CLL samples having a TP53 mutation/deletion. An increasing number of publications report the identification of recurrent mutations for CLL by NGS (see Introduction, section 1.18). The study reported here is distinct because it focused on a cohort of samples with multiple drug resistance, severe clinical phenotypes and TP53 mutations/deletions in patients. Inactivated P53 protein is associated with clonal instability, such as increased risk of Richter transformation. This is in keeping with the known role of wild-type P53 as a mediator of apoptosis, cell-cycle arrest and DNA repair in response to cellular stress. The main aim of this chapter was to analyse the profile of mutations and polymorphisms in the DNA maintenance genes of the samples, as it is possible that clonal instability because of P53 inactivation may be amplified based on the acquisition of secondary mutations in genes (other than TP53) that are involved in the response to DNA damage.

The objective of this chapter was to refine candidate variants in the DNA maintenance genes identified by PGM sequencing and focus on non-synonymous somatic aberrations. The refined list of candidate variants would need to be validated by Sanger Sequencing, which remains the gold standard for candidate variants with ≥ 10% allele frequency (207).

Affected genes with validated aberrations would then be biologically correlated with CLL

genomic instability, identified as large-scale or extensive genomic alterations by WGS

with the same cohort of CLL samples (see
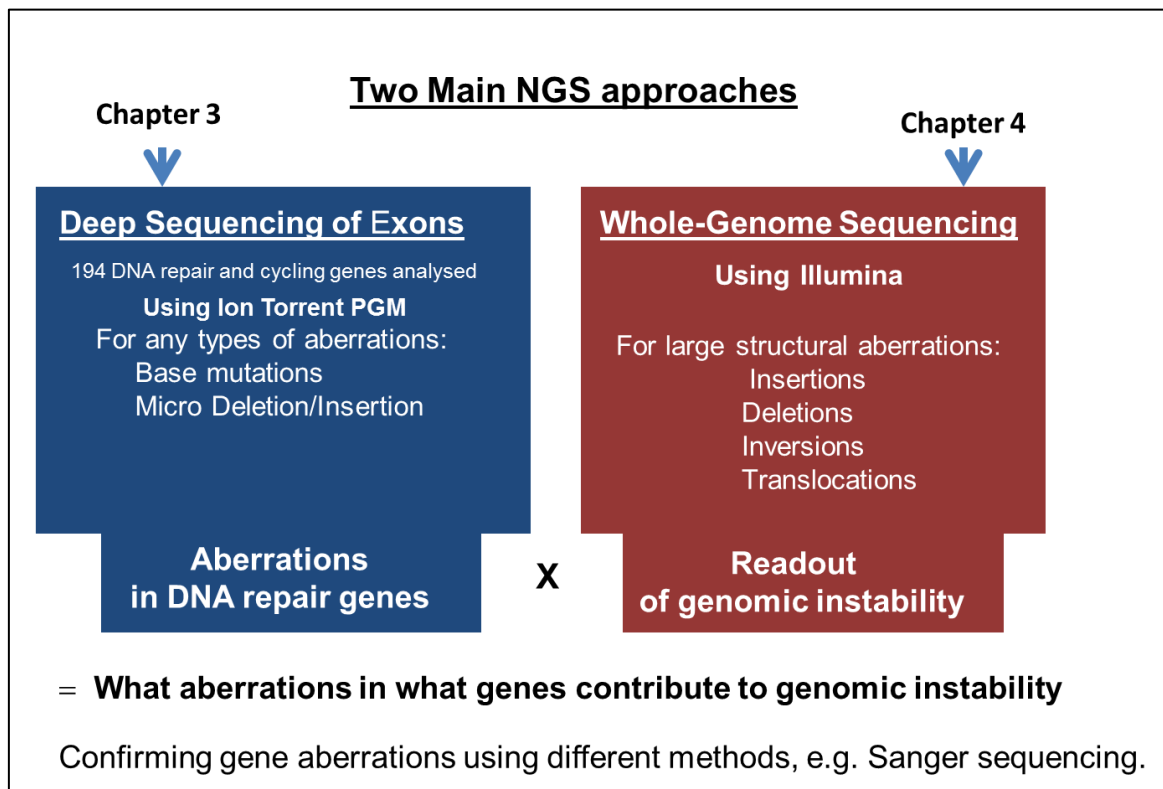
Figure 4-1).
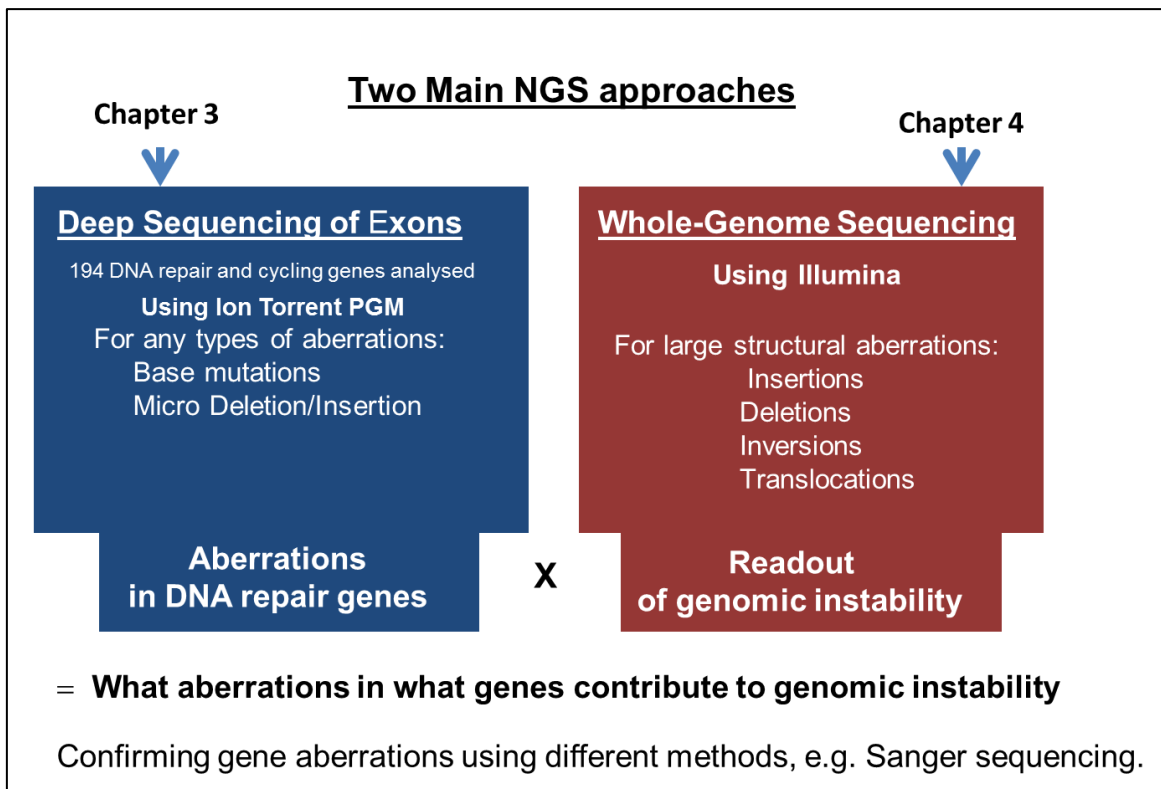
Figure 4-1: Relationship Between the Two NGS Approaches Used in the Study. Both approaches are applied to a small cohort of 10 CLL samples with TP53 either mutated or deleted. Deep sequencing of the canonical set of maintenance gene exons identifies somatic alterations that may contribute to genomic instability. WGS detects widespread or large structural alterations within the CLL genome of the same samples. Correlation of the maintenance gene mutations with the global alterations tests the hypothesis whether the former could be responsible for the latter.

4.1.1. TVC 4.2 Stringency and Candidate Variant Calling

Candidate variants were called from the CLL sequence data by TVCv4.2. The great

strength of NGS is its high capacity but sequences are not read precisely, and the data has

to be modelled to classify sequence differences from the reference genome as a variant.

Each base is represented on multiple clonal amplicons, where it is read and each time

there is a chance of an error occurring. Variants are therefore called according to the

number of reads that differ from expectation, leaving open the possibility of false

positives (FPs) and false negatives (FNs). The approach is therefore a screening approach,

hence the need for validation of candidates by Sanger sequencing. Different parameters

were therefore varied to control the stringency of TVC and determine the effect on the number and nature of the variants called. These included minimum allele frequency (proportion of reads that were variant), minimum quality (Phred score), minimum coverage and minimum coverage on both strands. These changed the sensitivity and accuracy of variant calling. High accuracy comes at the expense of low sensitivity. The high-stringency option was used in this project in order to minimise the number of FP results. Once variants had been identified, they were classified according to type, SNP or somatic candidates and the data was employed to examine the validity of the approach. Sanger sequencing of discovered candidate variants was used as the gold standard to determine whether variants were actually present in the CLL cases and permitted the efficiency and effectiveness of the variant selection criteria to be determined.

4.1.2. Filtering Strategy for the Classification of Candidate Variants

After variants were called and genotyped using TVCv4.2 and IRv4.2, respectively, they were checked for their occurrence in terms of both the dbSNP and 1000 genome databases. If the variants were reported to be germline origination, they were marked as SNPs, and if the variants were not reported, they were then checked against the COSMIC database. If they had not been reported as somatic, they were compared to DNA from buccal cells of the corresponding CLL patient to establish whether they were present and therefore of germline origin. Figure 4-2 shows the complete variant classification strategy.
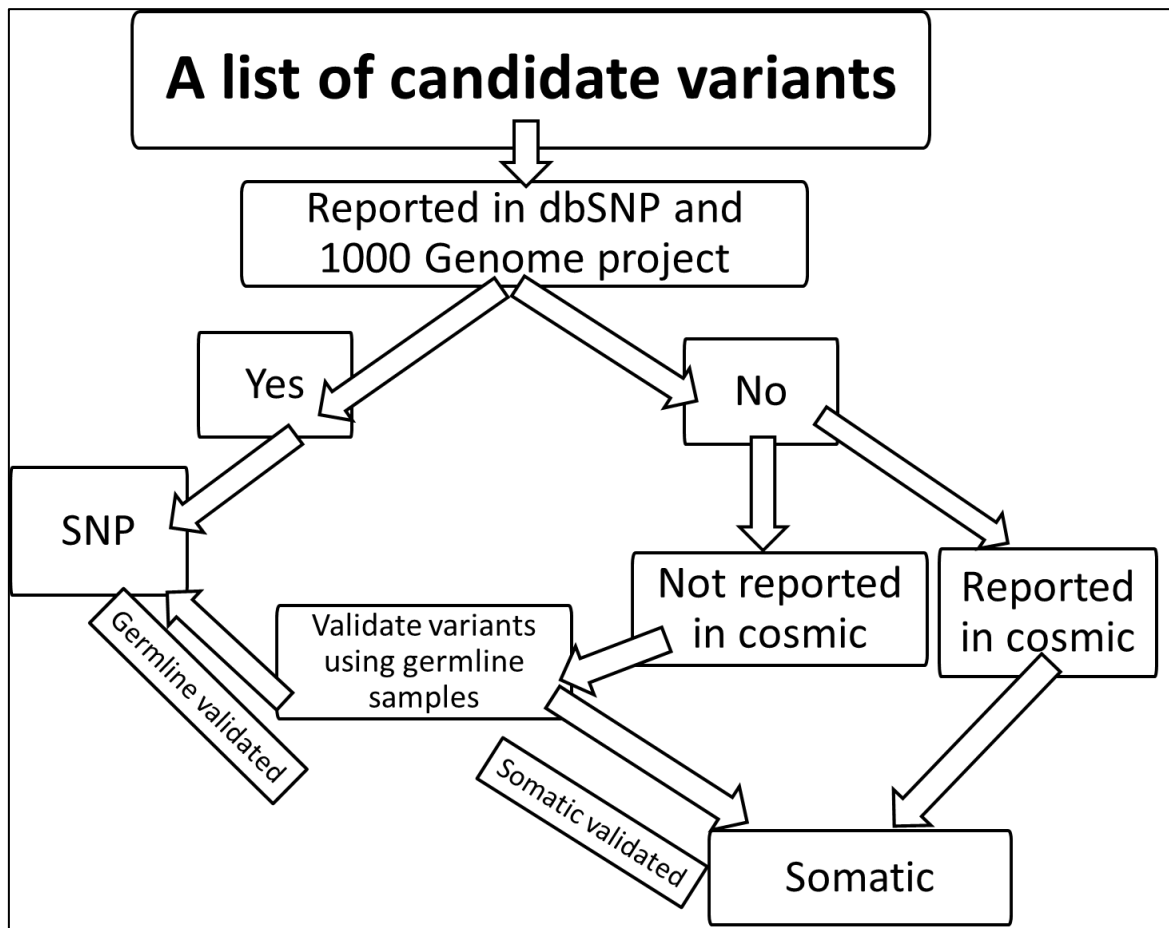
Figure 4-2: Flow Chart Outlining the Classification Strategy for Candidate Somatic Variants found in the NGS Data.

4.1.3. Investigation of the Validity of Variant Calling

Four QC methods were used to investigate the validity of variant calling. The first and

second QC rounds utilised known polymorphic SNPs found in the data and compared

their density versus the normal population (QC1) and within samples (QC2). Plots of

detected SNPs according to their genomic coordinates were also compared across

samples to qualitatively demonstrate that gross cross-contamination between samples

had not taken place. Ti)-to-Tv ratios on the targeted regions of samples was determined

for QC3, and the expected value was ~2.8 (291, 292). QC4 compared the TP53 mutational

status as determined by NGS with the TP53 mutations that had already been detected by

FASAY and Sanger sequencing.

4.2. Results

4.2.1. Types and Proportion of Variants Identified using High-Stringency Calling

A total of 2812 variants were detected in the 10 CLL samples by TVCv4.2 (average of 280

variants per sample), of which 1302 variants were non-synonymous missense variants

(130 variants average per sample) and eight variants were nonsense variants. Indels

comprised 76 frameshift insertions or deletions and there were 18 non-frame indels.

Synonymous variants were the largest proportion with 1407 in total (140 variants average

per sample; see
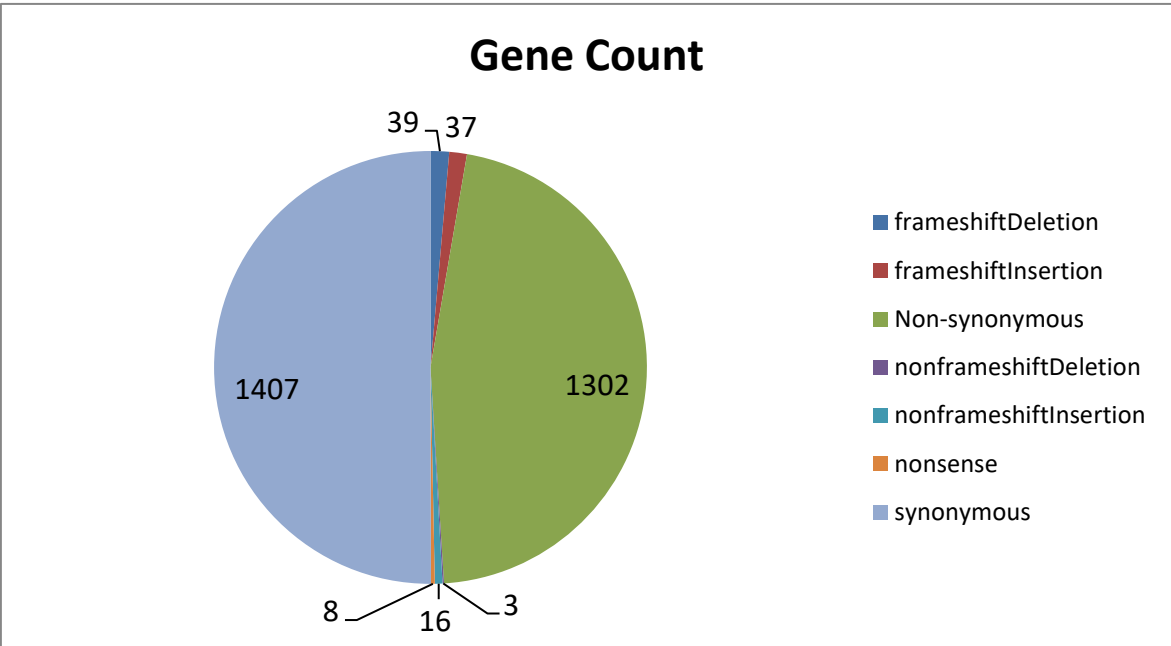
Figure 4- 3). The majority of the variants (2634) were present in dbSNP 135 (released on

Nov 2014) and were considered to be germline variants. Forty-two variants were

candidate somatic non-synonymous variants; the remaining 33 variants were

synonymous.
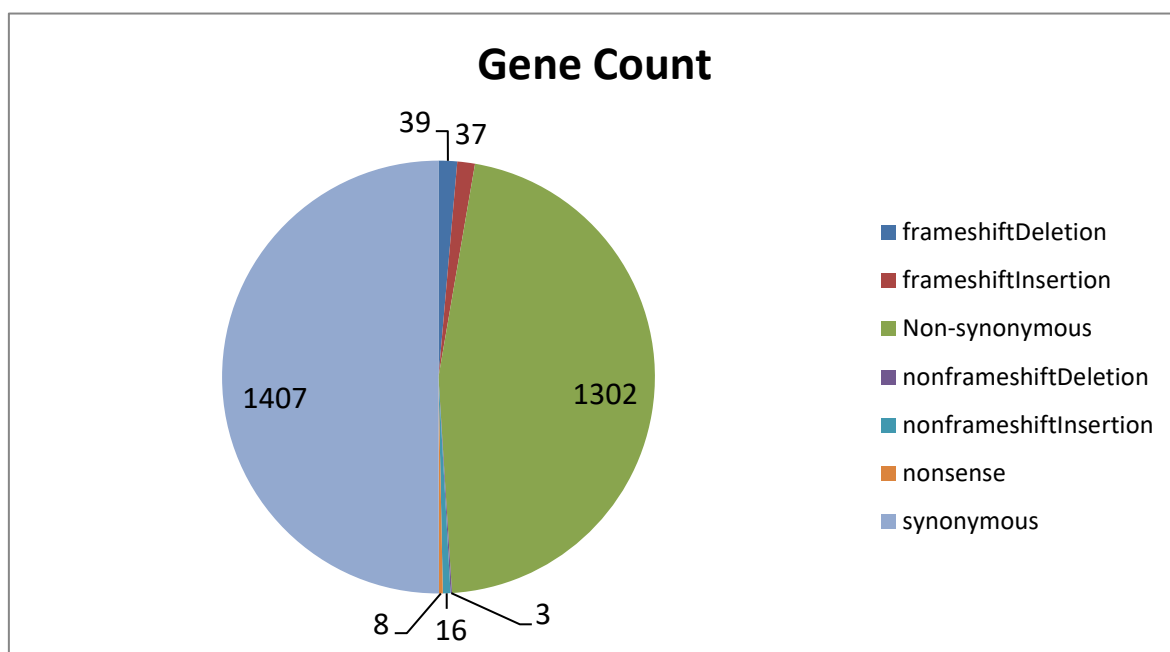
Figure 4- 3: Pie Chart of the Types and Proportions of Variants Identified in 10 Samples Within the Target Regions. Variant types include non-synonymous, nonsense, synonymous, frameshift deletion, frameshift insertion, non-frameshift deletion and non-frameshift insertion. The total number of variants detected was 2812. Missense and synonymous variants accounted for the highest number of detected variants, 1302, and 1407 variants, respectively. TVCv4.2 was used for variant calling.

4.2.2. Sanger Sequencing Validation of the Candidate Variants

The candidate non-synonymous somatic variants were validated by Sanger sequencing.

Primers for the somatic variants were designed by Integrated DNA Technologies (IDT) -

see Table 4- 1, PCR was optimised using GoTaq G2 Flexi DNA Polymerase (Promega,

Southampton, UK) and the PCR products were purified by Wizard® SV Gel and PCR Clean-

Up System (Promega , Southampton, UK)- see Materials and Methods section 2.5.2.2

Wizard® SV Gel and PCR Clean-Up System). Genewiz® Sanger service was employed to

sequence PCR products with both corresponding PCR primers.

Table 4- 1: Primers by Gene Name for PCR Amplification and Validation Testing of Candidate Somatic Variants

Sequence name	Accession number	Bases	Primer sequence	GC content (%)	Tm (MgCl) (Celsius)	Tm Temperature for PCR (Celsius)	PCR Product size (bp)
TP53 _For	NC_000017.11	20	5'GAT ACG GCC AGG CAT TGA AG 3'	55	60	60	314
TP53_Rev		23	3'GCA ATG GAT GAT TTG ATG CTG TC5'	43	61		
RECQL4_For	NC_000008.11	19	5'GCA CAT GTC TGC GCA GCT C 3'	63	62	60	455
RECQL4_Rev		20	3'TAC AGC GAG CCT TCA TGC AG5'	55	60		
REV3L_For	NC_000006.12	21	5'GAT TAC AGA CAT GAG CCA GTG 3'	48	60	58	389
REV3L_Rev		22	3'AGA GTA AAT AGG AGA AAG GGA G5'	41	58		
ERCC6_For	NC_000010.11	19	5'CTG TTC CTT GGC CTC ACT C 3'	58	60	58	394
ERCC6_Rev		21	3'ATC TGG ACC AGA AGA GTT GTC5'	48	60		
XPC_For	NC_000003.12	19	5'CCT TTG GCA CTT GGC CTG C 3'	63	62	60	301
XPC_Rev		21	3'GTT GAT CAC TGT CTG AGC TGG5'	52	61		
ATM_For	NC_000011.10	21	5'CTA GGA TTA GTG AGT AGG AGG 3'	48	60	58	301
ATM_Rev		21	3'CAC AAG GTG AGG TTC TAA	48	60		

Sequence name	Accession number	Bases	Primer sequence	GC content (%)	Tm (MgCl) (Celsius)	Tm Temperature for PCR (Celsius)	PCR Product size (bp)
			TCC5'				
BLM_for	NC_000015.10	24	5'TGT ATC TTC TTA TCA GGG AGT AAG 3'	38	60	58	503
BLM_Rev		22	3'GTA TCT CCA GTG TCA AGC ATA G5'	45	60		
HLTF_for	NC_000003.12	25	5'ACT GAA AGA ACA CTC TAA TAA TCT G 3'	32	59	58	556
HLTF_rev		22	3'CTA GCT AGT CCA GAT CAC ATA C5'	45	60		
MGMT_For	NC_000010.11	21	5'CGA CCA GCC TCT TAC CTA TAC 3'	52	61	60	373
MGMT_rev		21	3'ACA CAG GGA AGC TGC AAA TGC5'	52	61		
MSH4_for	NC_000001.11	25	5'ACT CTT TGA CTT ATT GCC TAT AAT G 3'	32	59	58	382
MSH4_rev		23	3'CAT ATG CTG TTT CCT TAA ATG GC5'	39	59		
POLE_for	NC_000012.12	18	5'TGC GAC TGG CTG GCA CTG 3'	67	61	60	393
POLE_rev		20	3'GTG TCC ACT CAT CTA CCA CC5'	55	60		
ERCC3_for	NC_000002.12	19	5'GCA TGC TTA CCA CCC AGA G 3'	58	60	58	317
ERCC3_rev		19	3'GCA GGT GGC TCT	58	60		

Sequence name	Accession number	Bases	Primer sequence	GC content (%)	Tm (MgCl) (Celsius)	Tm Temperature for PCR (Celsius)	PCR Product size (bp)
			TAG CTA G5'				
POLI_for	NC_000018.10	20	5'CTC CAC GAT TCC TTG GCA TG 3'	55	60	58	324
POLI_rev		22	3'CTT CTC ATT TAC ACC CAA GGA G5'	45	60		
RAD54L_for_2	NC_000001.11	19	5'CTA GGT TGC ACT GCC GAC G 3'	63	62	60	340
RAD54L_rev_2		21	3'GCA AAC ATC ATG CAG CCC TTC5'	52	61		
XRCC3_for	NC_000014.9	19	5'TGT GTC TGA ACC AGG CTC C 3'	58	60	58	359
XRCC3_rev		20	3'GCT TGC CTG CTT CCT GTT TC5'	55	60		
RAD54B_for	NC_000008.11	21	5'GAT TTG CTT CAC TGA GCT AGC 3'	48	60	60	469
RAD54B_rev		26	3'GAC CTT ACT ACT TAG ACA TTA AAG TC5'	35	62		
ERCC2_for	NC_000019.10	18	5'GCA AAC CGC TGT GGG CAG 3'	67	61	60	398
ERCC2_rev		20	3'TGA GTA GCT CTG TCT CCC AG5'	55	60		
TDG-Fp-1	NC_000012.12	23	5'AGA TGC CAA GTA ATA CTG TGT CG 3'	43	61	61	900
TDG-R-1		21	3'TGC CAT GTA TCA GGT CTC	52	61		

Sequence name	Accession number	Bases	Primer sequence	GC con-tent (%)	Tm (MgCl) (Celsius)	Tm Temperature for PCR (Celsius)	PCR Product size (bp)
			CAC5'				
POLD1-Fp-1	NC_000019.10	20	5'ACG ACC GCA TGG ACT GCA AG 3'	60	63	61	942
POLD-2-Rp-1		21	3'TGA CCT CCG ACT TCA TGT AGG5'	52	61		
DMC1-Fp-1	NC_000081.6	21	5'CAG GGA CCA AGT CTA TGT GTC 3'	52	61	61	589
DMC-1-Rp-1		23	3'CTC ACC TCA CTC CTT AGT TTA TG5'	43	61		
DCLRE1A-Fp-1	NC_000010.11	22	5'TAC TTC GGA GCA GGT GTA CTA G 3'	50	62	61	1115
DCLRE1A-Rp-1		22	3'TCA GAG TGT CCT GAT GGT CTT C5'	50	62		
BRCA1-Fp-1	NC_000017.11	22	5'AGT ACA CCA AGA CTC CCT CAT C 3'	50	62	61	619
BRCA1-Rp-1		21	3'TGG CAG GCA ACA TGA ATC CAG5'	52	61		
RAD54L-Rp-2	NC_000001.11	20	5'CAT GAA GGC GGA AGG TCT CA 3'	55	61	61	780
RAD54L-Fp-2		22	3'CAT GTG GTT GTT GAC CCT ATT C5'	45	60		
FANCA-Fp-1	NC_000016.10	22	5'AGA AGG CTC CAT GCG TCT AAT G 3'	50	62	61	652
FANCA-Rp-1		22	3'CAT GTC AGG TGA GTC CTG TTT C5'	50	62		

Sequence name	Accession number	Bases	Primer sequence	GC content (%)	Tm (MgCl) (Celsius)	Tm Temperature for PCR (Celsius)	PCR Product size (bp)
PARP1-Fp-1	NC_000001.11	22	5'GAA GAT GCT GTT ATG AGG GAG A 3'	45	60	61	532
PARP1-rp-1		22	3'AGA TGG TCT TCT GGT CGT TTC C5'	50	62		
PER1-Fp-2	NC_000017.11	20	5'GAG GGA GAG CTG AGT AAG AG 3'	55	61	61	750
PER1-Rp-2		20	3'GCT GGG AGG AAG GAC ATT TC5'	55	61		
FANCD2-Fp-2	NC_000003.12	22	5'GAT GCT TGA AGA GGG TTG CTA C 3'	50	62	61	674
FANCD2-Rp-2		21	3'CAG GGA AGA GGC CAG TAT TTC5'	52	61		

The previous table provides the forward and reverse primers of each set. The variants' location, number of bases, primer sequence, GC content, Tm and the size of PCR products are included.

Of the 42 non-synonymous aberrations tested, 15 were also found by Sanger sequencing and therefore classified as true positives (TPs). Only one variant, found in POLE, was also determined by Sanger sequencing in the buccal sample DNA and was therefore considered to be germline. A summary of the variants found after testing for validity is presented in Figure 4- 4.
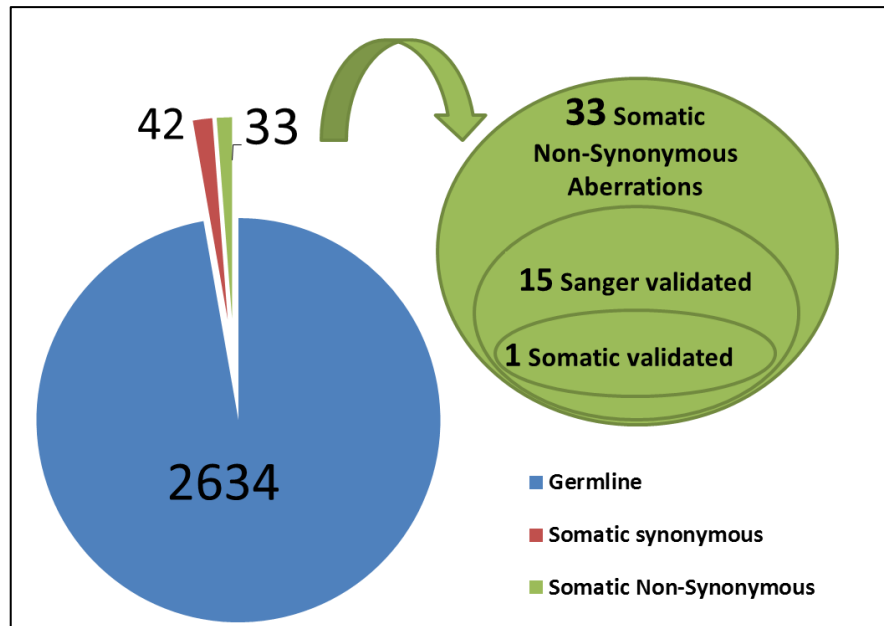
Figure 4- 4: Pie Chart Showing the Proportion of Somatic and Non-Somatic Aberrations Identified Within the Target Regions of 10 CLL Samples After Validation.

4.2.3. Sensitivity and Precision of the TVC 4.2 Stringency Set Used for Variant Calling.

Most variants were expected to be polymorphisms and therefore allele frequency in terms of reading coverage was expected to be bimodal at 50% and 100%, with such distribution having already been observed in the development of the approach (see Chapter 3, section 3.2.11, Figure 3- 12). However, low-frequency subclones harbouring key mutations could be significant for disease progression (see Introduction, section 1.18). It was therefore useful to make use of the validated versus non-validated variants to investigate the precision of variant calling. Two parameters were calculated based on the current results - sensitivity, which is the ability to testify how correct the TP variants are with the formula (TP/(TP+FN)) while accuracy and precision are the qualities of being correct, which is calculated by ((TP+TN)/(TP+TN+FP+FN)). The outcome is found in Table 4- 2.

Parameter	Generic-PGM-Somatic-HIGH Stringency	
	SNP	INDEL
Minimum allele frequency	0.02 (2%)	0.05 (5%)
Minimum quality	10	10
Minimum coverage	20	20
Minimum coverage on either strand	3	3
Maximum strand bias	0.9	0.85
Minimum relative read quality	8.5	-
TP	19	
FP	14	
FN	2	
Sensitivity	90.4%	
Precision (Accuracy)	51.3%	

Table 4- 2: The TVCv4.2 Stringency Set Used for Variant Calling and Resultant Sensitivity and Precision.

The chosen high-stringency default set is shown with 2% and 5% minimum allele

frequencies for SNP and indels, respectively. Ten was the minimum Phred quality score,

20 was the minimum coverage required on either side of a detected variant and 3

instances of the variant were required in each strand.

Even with the high-stringency settings, sensitivity for detection of variants was only 90.4%

and accuracy was 51.3%. This was considered to be a reasonable compromise in terms of

the aims of the project because a 90% chance of detecting any variants would have

offered a superb chance of detecting variants in association with particular forms of genomic instability across the samples used. Roughly 50% FPs effectively doubled the amount of validation required to find TPs. However, if base coverage had been uneven, this could have influenced the apparent sensitivity and precision and was therefore examined.

4.2.4. Base Coverage of TP53

Coverage of TP53 was assessed according to each CLL case. The covered bases were assessed depending on whether the known mutation was expected to be within the DBM or outside it. CLL-02, CLL-04, CLL-05, CLL-07 and CLL-10 were assessed inside the DBM and CLL-01, CLL-03, CLL-06, CLL-08 and CLL-09 outside. The results are found in Table 4- 3. According to different levels of possible fold coverage, from 1x to 100x, the percentage of bases that were covered to a depth of at least 50-fold ranges from 97.95-84.58% (average coverage: 91.3%), excluding sample 2642 (CLL-07) which had 60% coverage. Five samples had an average of at least 90% (93.97-97.115%) of their bases covered at least 100-fold.

Table 4- 3: Base Coverage for TP53 Either Within or Outside of the DBM of 10 CLL Samples According to the Expected Location of Their Known Mutation

Target base coverage	Inside DBM					Outside DBM				
	2550 CLL10	2640 CLL05	2642 CLL07	2681 CLL02	2766 CLL04	2554 CLL08	2621 CLL09	2631 CLL01	2657 CLL06	2600 CLL03
1X	99.92	99.97	99.87	99.81	99.85	99.9	99.91	99.81	99.79	99.97
10X	99.6	99.065	90.52	99.235	99.4	99.41	99.61	99.3	99.1	99.49
20X	99.16	97.745	74.34	98.89	98.92	98.89	99.09	98.935	97.94	98.92
30X	98.8	95.63	66.5	98.57	98.43	98.4	98.66	98.665	95.79	98.235
40X	98.35	92.68	62.67	98.28	97.64	97.96	98.15	98.465	91.65	97.395
50X	97.95	88.915	60	98.03	96.43	97.47	97.58	98.265	84.58	96.31
100X	95.16	58.425	53.03	96.545	81.47	93.97	94.28	97.115	22.22	85.47
Uniformity of base coverage	95.15	95.99	58.65	97.08	98.4	95.43	95.37	97.585	98.49	97.78

Samples are divided into groups according to the expected location of their TP53 mutation either inside the DBM or outside of it. The percentage of bases covered at the stated fold coverage is: 1x, 10x, 20x, 30x, 40x, 50x and 100x. Uniformity of base coverage is also seen. The uniformity of base coverage is defined as the percentage of target bases covered by at least 20% the average base-read depth.

Loss of heterozygosity (LOH) may have made a difference to the overall coverage of bases at the TP53 locus but there did not appear to be a relationship to the FISH results, suggesting that the PCR and hybrid selection as well as a total number of reads per case were overriding for coverage. Saturation of cRNA probes by the target may have offset any differences in relative allelic representation. Apart from CLL-07, uniformity of base coverage was over 95%, providing confidence that screening for novel variants would be successful.

4.2.5. Coverage of Known SNPs

Most of the variants were SNPs with bimodal allele AF ranges of 90-100% and 40-60%.

The average coverage was 200 reads with a quality score ≥ 10 as selected by TVCv4.2.

Coverage analysis graphs were used to show the pattern of allele coverage relative to the

total coverage and variant frequency (see Figure 4-5 for example using SNPs found in CLL-

04). Allele calls in this case were being made on reading depths of at least 100, offering a

high confidence that even minor variants would be detected.
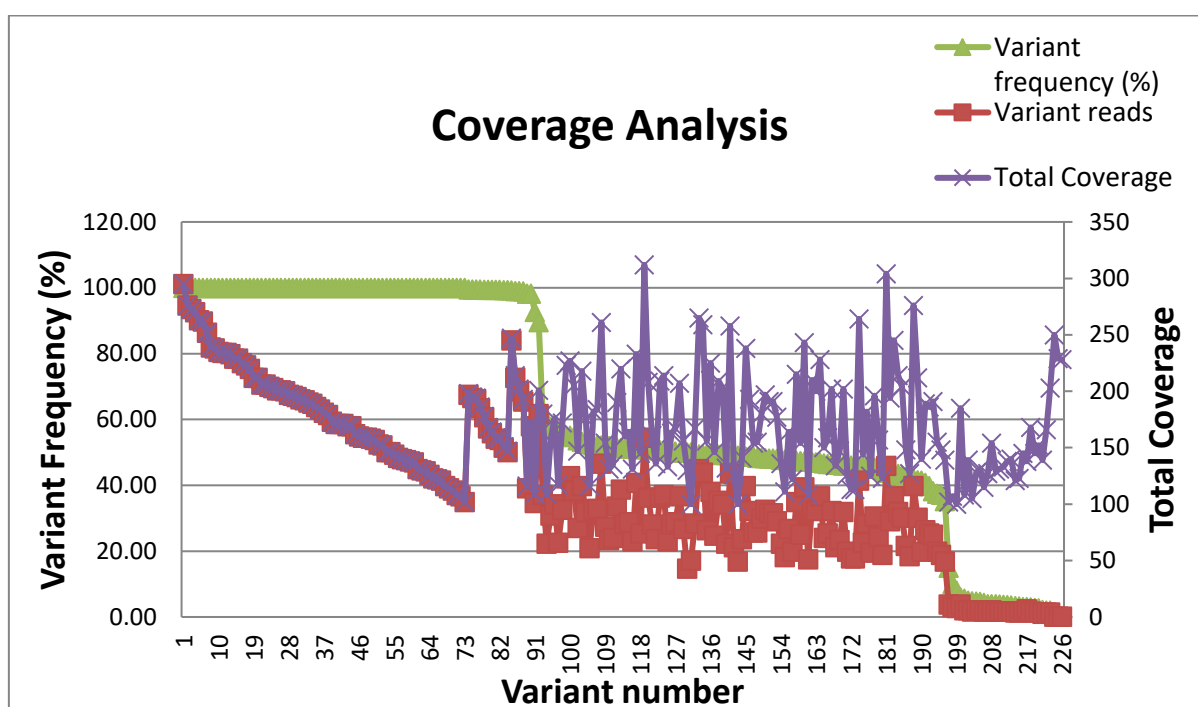
Figure 4-5: Coverage Analysis. Coverage versus variant proportion in relation to total coverage reads for SNPs found in sample CLL-04. 217 total variants are shown. The X-axis represents the variant number; the primary Y-axis represents variant frequency while the secondary Y-axis is base coverage (including variant coverage).

4.2.6. SNP Density Compared to the Normal Population (QC1) and Within Samples (QC2)

In order to further validate the sequencing data, SNP densities and patterns were

analysed. SNP densities in exonic regions of the 10 CLL samples had an average range of

Chapter 4: Application of a Targeted NGS Method …

248 SNPs per 499kbp (0.5 SNPs per 1 kbp; Table 4-4), which was as expected within the

normal population (293). This suggests that SNPs found in the CLL cases were

representative of the general population and that the methodology was not introducing

any forms of systematic bias. In addition, each sample had a very different pattern of

known SNPs, although a few variants were shared among the 10 samples (Figure 4-6).

This supports the representative nature of the SNPs found and also rules out any gross

cross-contamination between samples.

Table 4-4: Density of SNPs in Exonic Regions for CLL Samples 1-10.

Sample#	SNPs/499 kbp	SNPs/1 kbp
1	251	0.50
2	246	0.49
3	254	0.50
4	233	0.46
5	234	0.46
6	257	0.51
7	221	0.44
8	252	0.50
9	259	0.51
10	271	0.54
Mean	248	0.5

The number of SNP variants per case and density per 100 bases as detected by Ion

Reporter (IR 5.0).
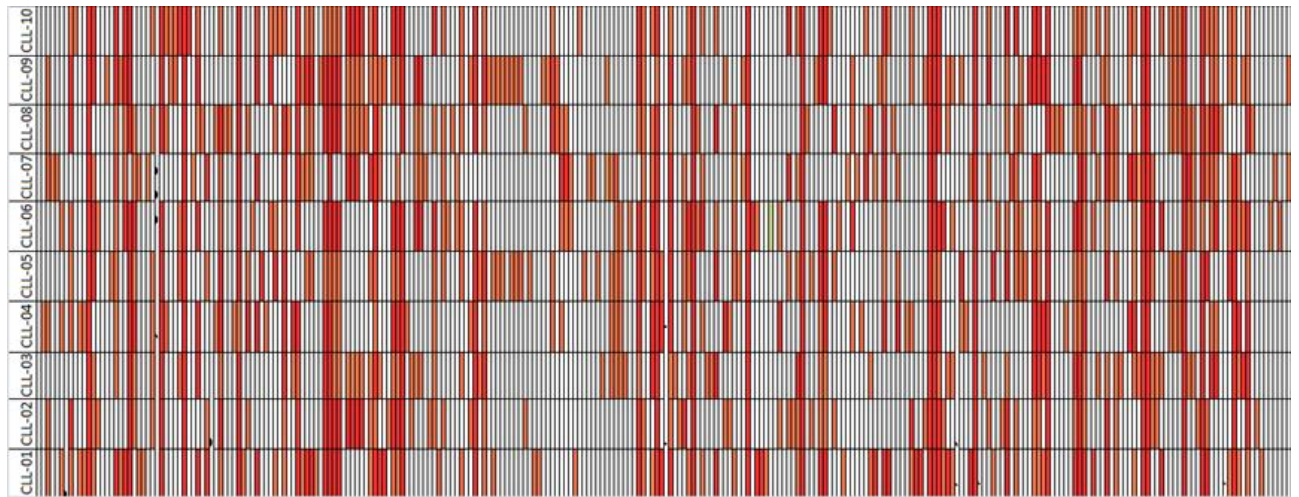
Figure 4-6: SNPs found in the 10 CLL Samples according to genomic position.

Red bars denote SNPs with ~100% VAF while the orange bars denote VAF of around 50%.

4.2.7. QC3 Ratio of Transition (Ti) and Transversion (Tv) Variants

The Ti-to-Tv ratio was calculated for the detected variants as a reliability measure for

variant calling (294, 295). Germline variants (SNPs) were classified as transitions or

transversions and the range of Ti/Tv was found to be an average of 2.4. The most

frequent variants were transitions of G:C>A:T with an average of 10 variants per change

per sample. This suggested that the active hypermutation role of activation-induced

cytidine deaminase (ACIDA) protein had been occurring in the CLL cells and this was as

expected (296-298). The second most frequent change was A:T>C:G with a range of (1 –

2.5) variants per sample. This may be related to activation of error-prone polymerase eta

(299, 300). These were more prominent in CLL samples with TP53-DBM mutations (301).

Overall, the results suggest that mechanisms of mutagenesis that could enhance genomic

variation and support progression may have been prevalent in the samples.

Table 4-5: Ti versus Tv for the Germline and Novel Variants

-		Transitions				-	Transversions								-	-
Nucleotide change	-	A>G	G>A	C>T	T>C	Total Ti	A>C	A>T	G>C	G>T	C>A	C>G	T>A	T>G	Total Tv	Ti/Tv
Germline variants	X	5.6	14.1	12.6	7.5	39.7	1.0	1.1	4.7	1.0	0.9	4.3	1.5	2.5	16.9	2.4
	SD	1.4	2.8	3.1	2.7	10.0	1.0	0.7	1.0	1.0	0.8	1.0	0.5	0.8	6.8	1.5
Novel variants	X	37.8	38.1	33.1	37.5	146.5	5.6	6.7	7.9	8.0	4.7	6.0	2.8	6.5	48.2	3.0
	SD	3.5	2.2	1.9	3.1	10.7	0.9	0.9	0.7	0.9	1.0	1.5	0.7	0.5	7.1	1.5

Table 4.5 lists 10 CLL samples and their means (X) and standard deviations (SD) of each group along with the Ti/Tv ratio of each sample. Definitions: Ti is the nucleotide change that interchanges into two ring purines (adenine (A) to guanine (G)) or one-ring pyrimidines (cytosine (C) to thymine (T)). Therefore, changes are within the same group that leads to four scenarios: A > G, G > A, C > T, and T > C. For Tv, the change of nucleotide causes interchange of nucleotide from purine to pyrimidine and vice versa, which results in eight scenarios: A > C, C > A, A > T, T > A, G > C, C > G, G > T and T > G. X is the numerical expression of the middle or central value in a certain data set. SD is the expression of how much the values in a group set differ from X.

QC4 comparison of the TP53 mutational status as determined by NGS with the TP53 mutations that had already been detected by FASAY and Sanger sequencing in the CLL cases.

Three previously identified TP53 mutations were used as further validation, two variants with > 90% read depth were validated using Sanger sequencing (Figure 4-7). One variant with 2% read frequency failed validation owing to the sensitivity of Sanger sequencing being limited to 10-20% of sample DNA. Identification of the expected TP53 mutations supported the efficacy of the overall approach for variant detection.

Figure 4-7 depicts the validation of three different TP53 variants on samples CLL-01 and CLL-10 using Sanger sequencing. Panel A depicts the Sanger validation of CLL-01 samples for a four-amino acid insertion at p.273 (90% AF). Panel B. portrays variant p.Arg 175 His (2% AF) failed to be validated by Sanger sequencing. Panel C shows the Sanger validation for p.Tyr 234 Asp (60% AF) of sample CLL-10.

Order number	Chr	Position	codon(3'-5')	coding	Protein change	AF (%)	Quality	Type	Original coverage	Allele cov	Allele cov+	Allele cov-	Strand bias
A	chr17	7577120	Ins CAAAAATACCAC	c.819_820ins GTGGTATTTTTG	p.Arg273_Val274 insTrpTyrPheCys	83.1	1686.61	INS	148	123	54	69	0.5309
B	Chr17	7578406	CGC-CAC	c.524 G>A	p.Arg 175 His	2	-	SNV	192	4	2	2	0.5
C	Chr17	7577581	TAC-GAC	c.699 T>G	p.Tyr 234 Asp	60.7	658.34	SNV	107	65	12	51	0.683

Figure 4-7: **Validating TP53 mutations by Sanger Sequencing.** Three TP53 variants from CLL01 (A) and CLL10 (B and C) were validated. A. CLL-01 with TP53 somatic Insertions at 273, including the Sanger sequencing result. B & C. CLL-10 with two somatic variants. Panel B depicts the PGM validation for 2% mutation at p.Tyr 234 Asp but failed validation with Sanger sequencing as it is below the sensitivity limit. Panel C shows for PGM the Sanger validation for mutation p.Arg 175 His.

4.2.8. Summary of Variants Validated

Figure 4- presents the validation testing of 16 non-synonymous aberrations detected by

Ion PGM, WGS and Sanger Sequencing used in the validation as detailed in Table 4- 6.

There was limited germline material available for validation. Germline materials were

available for the samples: CLL-02, CLL-06, CLL-08 and CLL-10. As a result, certain variants

were germline-validated, including the following genes; *BLM, RAD54B, XPC, ERCC2,*

ERCC6, HLFT, MSH4 and POLE. The findings indicated POLE p.Ala661Thr was somatically

validated with 50% AF (Figure 4-). In total, there were two POLE mutations and one

RAD53L that were regarded as somatic mutations in three CLL samples (30% of cohort

samples).

Table 4- 6**: Validated Non-Synonymous Variants Using Ion Reporter v4.2 with dbSNP Released in November 2014**. A total of 26 aberrations were validated by Sanger sequencing or WGS. Within the table is genotyping information that includes COSMIC, clinvar, dbsnp, maf, 5000 exomes, polyphen and sift. Ion Reporter v4.2 was employed for the genotyping. Abbreviations: COSMIC: Catalogue of Somatic Mutations in Cancer; Clinvar is open archives of reports associated with human variations and phenotypes. Sift is a tool which predicts the severity of amino acid substitutions based on the conservation degree of amino acids; maf is the minor allele frequency, which is reported in a given population; and the polyphen score predicts the impact of a substituted amino acid on the structure and function of the human protein.

#	Gene	# locus	Sample affected	Coding	(wt) codon	Affected codon	Affected protein	Validated by (in addition to PGM)	Germ-line validation	AF (%)	Cosmic	dbSNP (2014)	1K Genome (Stage 3) (2015)	maf	5000 Exomes	polyphen	sift
1	ATM	chr11:108151786	CLL03	c.3467C>T	ACG	ATG	p.Thr1156Met	Sanger+WGS	-	49.1	-	-	rs759951393	-	-	0.001	0.94
2	BLM	chr15:91346924	CLL02	c.3532C>A	CAA	AAA	p.Gln1178Lys	Sanger+WGS	Yes	52	-	-	-	-	-	0.002	0.16
3	RECQL4	chr8:145738669	CLL04 & CLL07	c.2395G>A	GTG	ATG	p.Val799Met	sanger (CLL07), WGS	-	36.8; 39.3	-	rs34293591	-	0.008	AMAF=0.0054:EMAF=0.0226:GMAF=0.0171	1	0
4	RAD54B	chr8:95423478	CLL08	c.367_370delGTTA	GTT	AAT	p.Val123fs	Sanger+WGS	Yes	42.9	-	-	rs766417386	-	-	-	-
5	RAD54L	chr1:46739401	CLL04	c.1592A>C	AAG	ACG	p.Lys531-Thr	WGS	-	46.7	-	-	-	-	-	0.12	0
6	XPC	chr3:14200382	CLL08	c.1001C>A	CCT	CAT	p.Pro334-His	Sanger+WGS	Yes	50.6	-	rs74737358	-	0.004	AMAF=0.0268:EMAF=0.001:GMAF=0.0088	-	-
7	ERCC2	chr19:45855803	CLL06	c.2005_2006delAG	AGG	GGG	p.Arg669fs	Sanger+WGS	Yes	52.1	-	-	rs757535186	-	-	-	-
8	ERCC3	chr2:128051252	CLL03	c.71_72insGGATGAAGAGGA	GAT	GAG	p.Glu23_-Asp24insGluAspGluGlu	Sanger+WGS	-	32.4	-	-	rs777778660	-	-	-	-

#	Gene	# locus	Sample affected	Coding	(wt) codon	Affected codon	Affected protein	Validated by (in addition to PGM)	Germ-line validation	AF (%)	Cosmic	dbSNP (2014)	1K Genome (Stage 3) (2015)	maf	5000 Exomes	poly-phen	sift
9	ERCC6	chr10:50678470	CLL02	c.3536A>G	TAT	TGT	p.Tyr1179Cys	Sanger+WGS	Yes	47.1	-	-	rs536757172	-	-	-	-
10	HLTF	chr3:148757858	CLL06	c.2462G>A	AGT	AAT	p.Ser821Asn	Sanger+WGS	Yes	54.3	-	-	-	-	-	-	-
11	MGMT	chr10:131334629	CLL03	c.206C>T	ACG	ATG	p.Thr69Met	Sanger+WGS	-	51.2	-	-	rs763636757	-	-	0.884	0.09
12	MSH4	chr1:76343968	CLL08	c.1505G>A	AGA	AAA	p.Arg502Lys	Sanger+WGS	Yes	49.3	-	-	rs767713032	-	-	0.918	0.03
13	POLE	chr12:133219559	CLL07	c.4572_4573insC	AGC	CAG	p.Ser1525fs	WGS	No	48.4	-	-	-	-	-	-	-
14	POLE	chr12:133245266	CLL10	c.1981G>A	GCA	ACA	p.Ala661Thr	Sanger+WGS	-	47.9	-	-	-	-	-	0.645	0.05
15	REV3L	chr6:111636520	CLL01	c.8416G>A	GAA	AAA	p.Glu2806Lys	Sanger+WGS	-	47.1	-	-	rs756226303	-	-	0.753	0.15

#	Gene	WT	Allele
1	ATM	ACG	ATG

1

Somatic RS

#	Gene	WT	Allele
2	BLM	CAA	AAA

2

Somatic FS

Germline FS

#	Gene	WT	Allele
3	RECQL4	GTG	ATG

3

G T G
A

Somatic FS

#	Gene	WT	Allele
4	RAD54B	GTT	FS del AAT

4

FS del AAT

Somatic FS

Germline FS

FS del AAT

#	Gene	WT	Allele
5	RAD54L	AAG	ACG

5

Somatic FR

RAD54L

#	Gene	WT	Allele
6	XPC	CCT	CAT

6

Somatic FS

Germline FS

#	Gene	WT	Allele
7	ERCC2	AGG	Fs del GGG

7

Somatic
RS

Germline
RS

#	Gene	WT	Allele
8	ERCC3	GAT	FS Ins GAG

8

Somatic
RS

#	Gene	WT	Allele
9	ERCC6	TAT	TGT

9

Somatic
FS

Germline
FS

#	Gene	WT	Allele
10	HLTF	AGT	AAT

10

Somatic
RS

Germline
RS

#	Gene	WT	Allele
11	MGMT	ACG	ATG

11

Somatic
FS

#	Gene	WT	Allele
12	MSH4	AGA	AAA

12

Somatic
FS

Germline
FS

Figure 4-7: Corresponding Variants Visualised for Variants in Table 4- 6. Somatic refers to variant validation using CLL samples, whereas Germline refers to variant validation with Germline DNA from a buccal swab of the same patient. Abbreviations; RS: reverse strand; and FR: forward strand. Details of variants, including allele coverage, are included in Figure 4-6.

4.3. Discussion

This chapter covers the application of NGS to targeted genomic regions for the identification of somatic variants in DNA maintenance genes using an in-house Ion Torrent PGM. The targeted genes consisted of 194 DNA maintenance genes (61, 266, 276). The cohort included 10 CLL cases with P53 inactivation. More specifically, the chapter sought to address the susceptibility of DNA maintenance gene(s) to somatic aberrations in CLL, which could play a role in CLL clonal instability and post-CLL Richter syndrome.

For variant calling, the default high-stringency TVCv4.2 was utilised, resulting in a variant sensitivity of 90.4% and accuracy of 51.3%. The reason for the low accuracy was the high number of FP variants based on the inherited inaccurate flow calls. These could cause overcalling of short HPs and undercalling on long HPs (283). The minimum AF was 20% with a minimum Phred quality score of 10, meaning the probability of an incorrect base call is 1 in 10.

Validation checking and comparison of SNP density and patterns across samples confirmed the representative nature of the SNPs found and excluded cross-contamination between different samples. Ti/Tv ratios suggested mutational processes were active in the CLL samples, in particular transition changes affected by AICDA (296-298) and polymerase eta (301).

In respect to the mutation profile, Ion Reporter v4.2 was used to genotype variants and differentiate SNPs from total variants. Figure 4- 4 shows 42 variants are non-synonymous and 26 were validated by Sanger sequencing. Twelve variants failed validation based on

the inherent inaccuracy of the sequencing flow calls. However, the germline validation of

certain samples revealed that many of the candidate variants were germline validated.

This was also seen in the updated versions of dbSNPs and the release of phase two and

three of the 1000 Genome project (238). As a result, many potential somatic variants

were filtered out as they were confirmed germline. Nevertheless, three non-synonymous

somatic mutations were identified in two DNA polymerase genes with the same function -

POLI and POLE in three CLL samples. In particular, the POLE gene was recurrently mutated

in different samples. As such, POLE was selected as a candidate to be subsequently

screened in a larger and independent cohort of CLL samples in Chapter 6 as well as

studying the genomic instability of the 10 CLL samples in Chapter 5.

5. Chapter 5: Investigation of Genome Integrity of CLL Having Compromised TP53

5.1. Introduction

This chapter concerns WGS of CLL already known to have TP53 alterations to investigate genomic instability and clonal transformation of this group. Types of DNA repair and the characteristic lesions that result when repair is defective are discussed in the Introduction. The persistence of repair defects in cancer are expected to contribute to genomic instability and therefore the presence of characteristic signatures changes resulting from the defect are predicted to indicate the prevailing type within any given case. Investigation of the most aggressive forms of CLL was expected to offer the best chance of detecting the signatures as they had P53 pathway defects and the greatest opportunity for accumulating damage. Forms of genomic instability included deletions, duplications, insertions and translocations (302, 303). Generally, damage can cover a wide range from one or a few nucleotides to large-scale inter- and intrachromosomal structural or copy number alterations. Consequently, the heterogeneity of millions of nucleotides could be affected, either thousands of small changes or millions of contiguous bases. Many variants that have arisen in the germline during the course of human evolution have become polymorphic in the population. This includes large-scale structural variants (SV).

In a normal population, germline SVs with a size range of one kb to several Mb may account for 4.8-9.5% of the total gDNA (304). In addition, common SVs can manifest somatically. In CLL, for example, approximately 80% of patients have between one and four common SVs. The four common SVs in CLL are deletions in 13q14.3, 11q, 17p and

trisomy 12 (181), which occur in 50%, 18%, 16% and 7% of cases, respectively (12).

Chromosome deletion 13q (del(13q)) is associated with a strong prognosis. The 13q-deleted region includes the DLEU2-mir-15-16 cluster, which regulates the expression of proteins that participate in cell-cycle progression or inhibit apoptosis. In contrast, 17p and 11q deletions severely affect clinical outcomes of CLL (181), 17p including the locus for TP53 and 11q, which is the locus for ATM. A novel form of genomic instability was first observed in CLL, specifically chromothripsis. This has since been detected by SNP-array analysis in approximately 2% of patients and it is mainly seen in CLL with mutated TP53 and unmutated IGHV status, and is associated with the poorest CLL prognosis (305).

This chapter aims to investigate whether there is evidence of genomic structural instability in the 10 cases found herein and if observed, establish whether it is associated with alterations to DNA maintenance genes. The data also provided the opportunity to look for other changes in genes not already sequenced and compare sequence data to that obtained by exome sequencing of cRNA-enriched targets.

5.2.   Results

WGS and sequence data processing were performed as described in the Materials and Methods, sections 2.13 to 2.14.

5.2.1. Sequencing Output and Alignment

After aligning data in the Bam file, sequencing outputs were generated such that they

included sequencing coverage. The alignment statistics of the 10 CLL samples showed

that the average proportion of reads that could be aligned was 96%; most of the

remainder were PCR duplications. Moreover, the average base coverage was about 30x

and the average genome coverage having 10x base coverage was 99.4 ± 0.4% (see Table

5-1).

Table 5-1: Sequencing Outputs of 10 WGS Experiments. Many parameters were
calculated, including mapped reads, mapped bases, unique rate, duplicate rate, mismatch
rate, average sequencing depth, coverage and coverage at least 10x.

Sequencing outputs	Average (10x WGS)	Std dev
Mapped reads	1,031,800,287	14,414,021
Mapped bases(bp)	91,705,276,908	1,245,616,942
Unique rate (%)	95.77%	0.29%
Duplicate rate (%)	5.79%	0.82%
Mismatch rate (%)	0.40%	0.06%
Average sequencing depth	30	0.4
Coverage	99.44%	0.36%
Coverage at least 10X (%)	96.70%	0.18%

5.2.2. Variant Coverage Analysis and Variant Grouping

Variants were detected using SNPeff and further grouped to unknown and known

variants, then sub-grouped into repetitive and non-repetitive variants. Coverage graphs

for representative Liv_01 were made for each of the variant groups; known SNPs

(8,706,597 variants), unknown repetitive variants, somatic variants (188,544 variants) and

somatic missense variants (
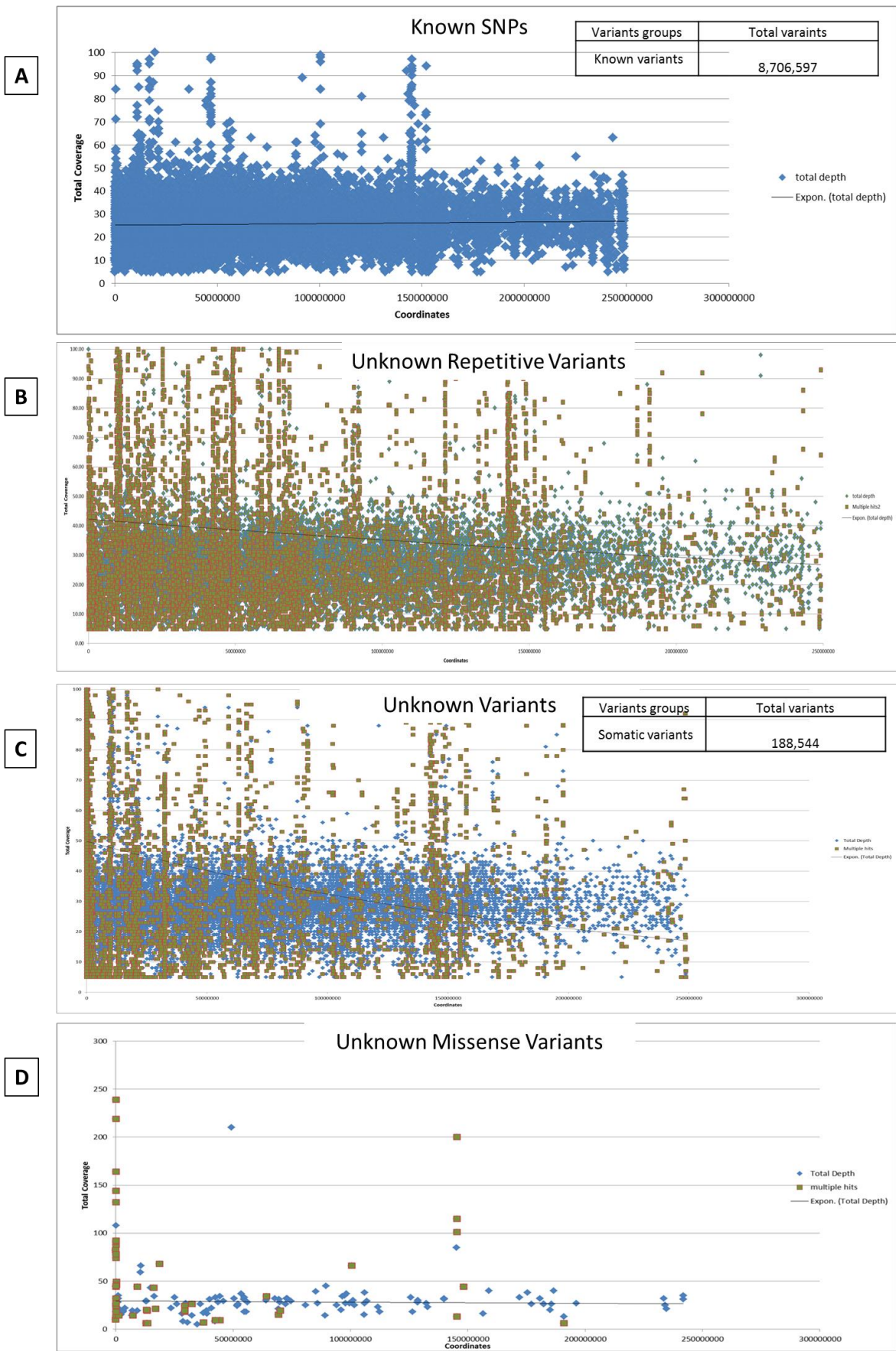
Figure 5-1). Notably, the repetitive group as seen in Panel B of Figure 5-1, as expected,

shows a pattern of outlier variants which could be recognised by excessive coverage of over 40 reads. Repetitive sequences are highly prone to misalignment because of a large number of possibilities for simple sequence errors to result in new FP hits. It is critical to distinguish between true variants and FP variants, but this is challenging with such a large number of potential artefactual alignments. One criterion are that they should be unique to a particular sample because it is highly unlikely that identical somatic mutations will occur in more than one of a small number of samples, excepting exceptional 'driver'-like mutations in KRAS or BRAF (306, 307). Most of the unknown variants in repetitive sequences are common between samples (dark-yellow points) in comparison to single-hit variants (Blue points) and therefore unlikely to be true positives.
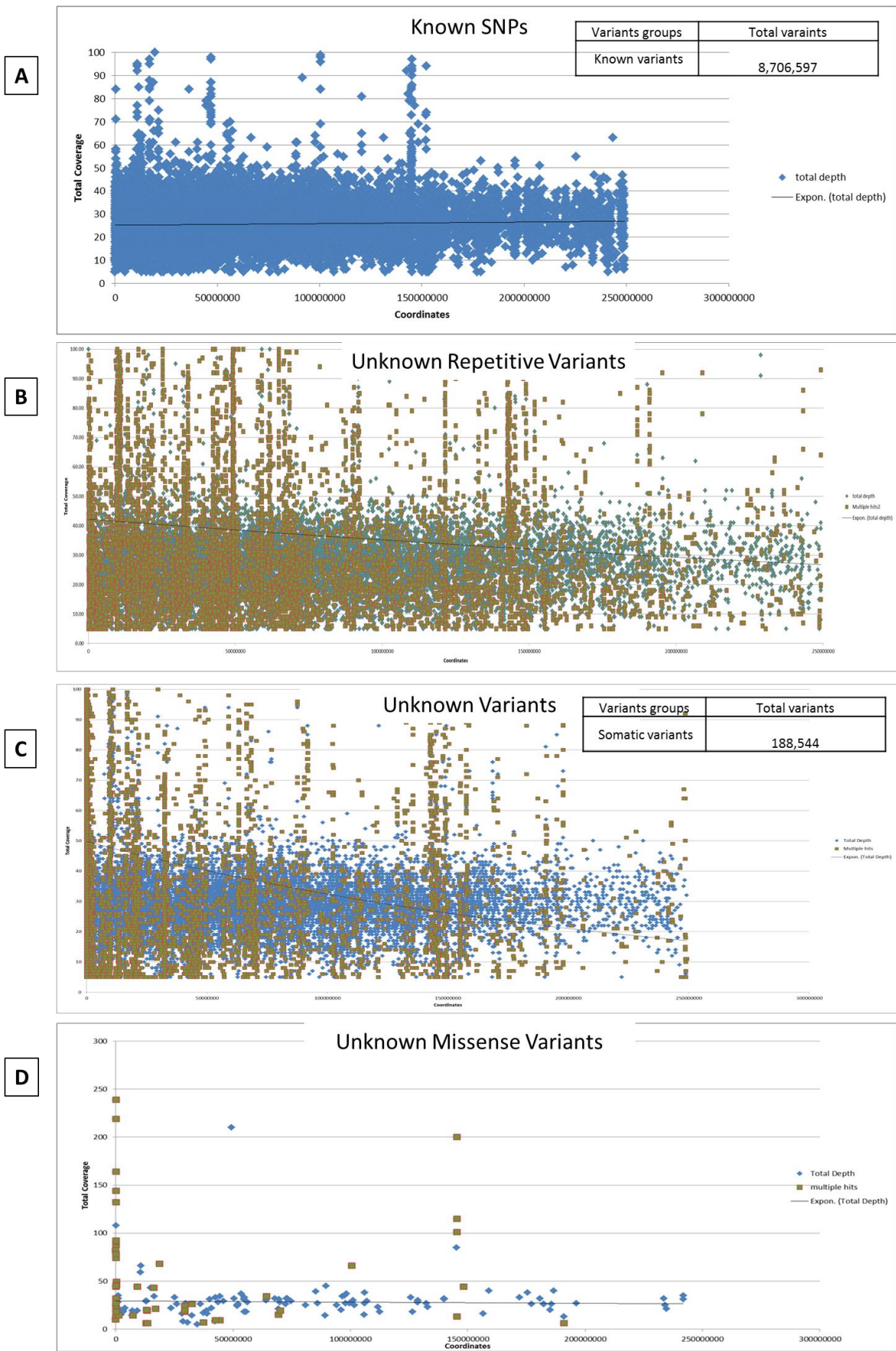
Figure 5-1: Coverage Analysis of Representative Sample, Liv_01. Four charts containing different variants groups; known SNPs, unknown repetitive variants, Unknown variants

and Unknown missense variants. X-axis represents genome coordinates and Y-axis signifies variants coverage (depth). Dark-green and blue points represent common and single-hit-only variants, respectively. A number of total variants is included for both known and somatic variants.

Selection of somatic variants was based on the following criteria: variants that were

unknown in either dbSNP or the 1000 Genome; variants that were outside of genome

sequences (identified using the CG-rich masker tool); variants that were unique to a given

sample; and then the missense variants that had total coverage ≤40 reads (Figure 5- 2).

The primary goal was to produce a set of variants with a minimum number of FPs.
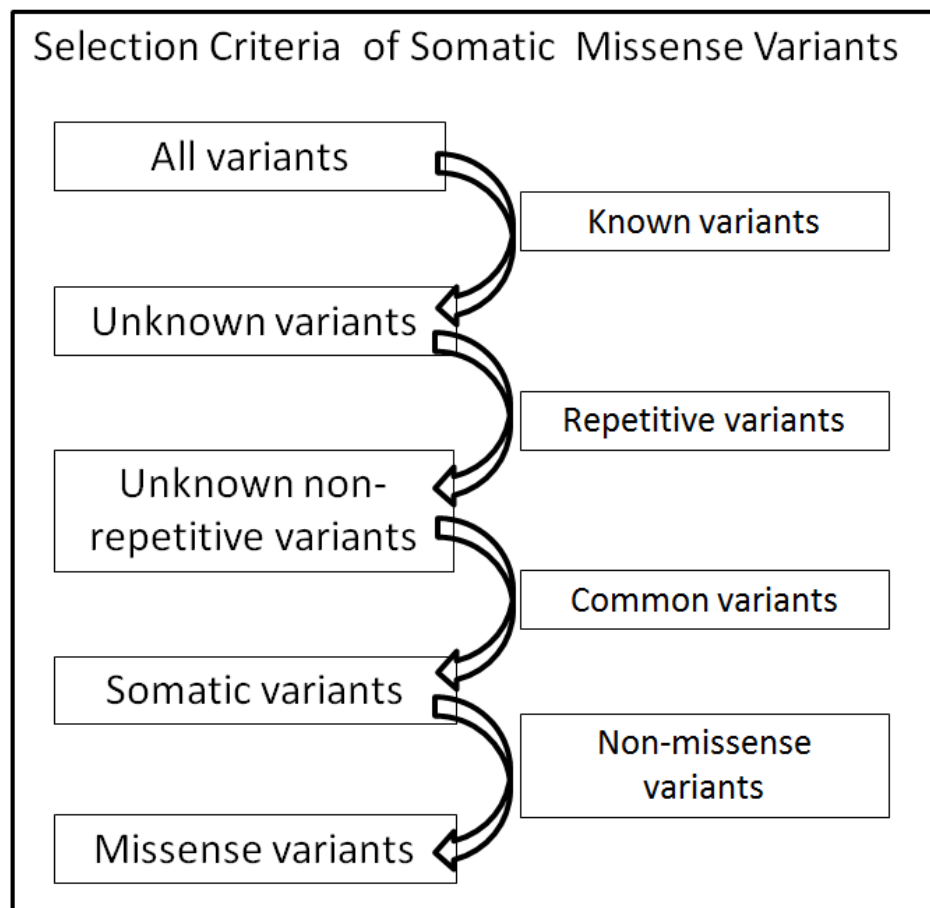
Figure 5- 2: Selection Criteria for Somatic Variants. Somatic missense variants were multistep recovered by identifying unknown variants, non-repetitive variants and novel (single-hit) variants. SNPeff software was used for grouping variants.

5.2.3. Somatic Variant Analysis

After variant selection, there were 78 affected genes carrying (8x-2x) multiple variants among 10 samples (Table 5-2). Among them, 8x hits were detected in Mitochondrial Ribosomal Protein S27 (MRPS27), 7x hits in IGHV3-13 and 5x hits in TP53, IQ Motif-Containing GTPase-Activating Protein 2 (IQGAP2) and Family with Sequence Similarity 186 Member A (FAM186A) (Panel A in Table 5-2). There are two DNA repair genes also noted, which were TP53 and POLE. Two IGHV genes were also detected, namely IGHV3-13, and IGHV1-3. Panel B in Table 5-2 shows the variant details of the most affected genes. The minimum AF observed was 25%, and both allele and reference bases were reliable, having average quality scores of 30. Most of the variants were missense variants. Two genes carried frameshift lesions with a high expectation of deleterious consequences: G-Protein Pathway Suppressor 2 (GPS2) and RANBP2-Like and GRIP Domain-Containing 4 (RGPD4). Most of the missense variants were protein-coding variants, and a number of variants are regarded as sites for amino acid modifications, either phosphoserine, for example, in IQGAP2, or acetyllysine in MRPS27. In TP53, there were five variants out of nine that were detected before applying filtering criteria.

5.2.4. Pathway Analysis

Although the objective was primarily to identify mutations in single maintenance genes and associate these with a genome-wide mutational signature, the possibility that maintenance pathways could have been impaired by combinations of variants was explored. The variants were therefore mapped protein-interactions pathways with the Reactome software (308, 309); see Figure 5-3). Using the set of missense variants as input, three interconnected pathways were predicted to be predominantly affected -

immune system, signal transduction and metabolism. The total number of genes involved

was 184, 126 and 117, respectively (Panels A and B in Table 5-2). Significant changes were

centred on sub-pathways, like, for instance, cytokine signalling and adaptive and innate

immunity. Examples of signal transduction are signalling by GPCR, interleukins,

interferons, FCERI and NGF. Metabolism of proteins and lipids were examples of

metabolism. Other pathways affected included membrane trafficking, cell cycle, gene

expression and post-protein modifications. In fact, the range of possibilities was so broad

it was impossible to make any direct connection to genome instability and that the

propensity for multiple pathways being important could be safely ignored.

Sample #	Chr	Pos	Wt N	Allele N	Variants - type 1	Prediction	Gene	Variants - type 2	N change	Protein change	Protein pos1	GQ	ABQ	Total coverage	Allele freq	AD D	AD F	AD R	RD D	RD F	RD R	PVAL	RBQ
Liv_02	5	75,749,486	T	G	sequence_feature	MODERATE	IQGAP2	amino_acid_modification:Phosphoserine	c.47-7909T>G			30	31	26	35%	9	3	6	17	7	10	8.493	31
Liv_02	17	7,216,899	G	A	stop_gained	HIGH	GPS2	transcript	c.622C>T	p.Gln208*	923/1379	56	31	21	67%	14	10	4	7	4	3	0.022	33
Liv_03	2	108,489,203	CA	C	frameshift_variant	HIGH	RGPD4	transcript	c.4745delA	p.Asn1582fs	4827/5383	55	32	30	50%	15	10	5	15	12	3	0.029	30
Liv_03	5	75,701,464	A	C	sequence_feature	MODERATE	IQGAP2	amino_acid_modification:Phosphoserine	c.46+2048A>C	-	-	23	32	24	29%	7	3	4	17	7	10	47.006	32
Liv_03	5	71,555,606	T	A	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.282-21651A>T			52	32	27	52%	14	9	5	13	10	3	0.062	32
Liv_03	5	71,563,662	GA	G	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.281+27695delT	-	-	39	32	25	44%	11	3	8	14	7	7	1.193	31
Liv_04	2	108,477,648	A	G	splice_acceptor_variant&intron_variant	HIGH	RGPD4	transcript	c.1921-2A>G	-	-	62	32	27	59%	16	6	10	11	5	6	0.006	32
Liv_04	12	50,745,822	T	G	missense_variant	MODERATE	FAM186A	transcript	c.4793A>C	p.Glu1598Ala	4793/7127	22	32	30	23%	7	1	6	23	19	4	52.713	30
Liv_04	17	7,579,362	A	C	missense_variant	MODERATE	TP53	transcript	c.325T>G	p.Phe109Val	515/2579	52	32	10	100%	10	0	10	0	0	0	0.054	0
Liv_05	17	7,578,445	A	T	missense_variant	MODERATE	TP53	transcript	c.485T>A	p.Ile162Asn	675/2579	58	31	11	100%	11	7	4	0	0	0	0.014	0
Liv_06	12	50,746,486	T	G	missense_variant	MODERATE	FAM186A	transcript	c.4129A>C	p.Thr1377Pro	4129/7127	23	21	21	33%	7	6	1	14	8	6	43.101	30
Liv_06	17	7,216,703	CTGAG	C	frameshift_variant	HIGH	GPS2	transcript	c.716_719delCTCA	p.Thr239fs	1020/1379	36	31	9	89%	8	5	3	1	1	0	2.057	36
Liv_06	17	7,577,545	T	C	missense_variant	MODERATE	TP53	transcript	c.736A>G	p.Met246Val	926/2579	53	28	12	92%	11	4	7	1	1	0	0.048	31
Liv_07	5	75,753,541	G	A	sequence_feature	MODERATE	IQGAP2	amino_acid_modification:Phosphoserine	c.47-3854G>A	-	-	53	33	34	44%	15	4	11	19	10	9	0.042	31
Liv_07	5	71,548,105	A	G	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.282-14150T>C			69	31	31	58%	18	9	9	13	8	5	0.001	33
Liv_07	5	71,563,816	C	T	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.281+27542G>A	-	-	39	33	23	48%	11	6	5	12	6	6	1.014	32
Liv_07	5	71,534,204	CAG	C	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.282-251_282-250delCT	-	-	42	21	29	41%	12	8	4	17	12	5	0.582	31
Liv_07	5	71,534,203	A	ATTT	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.282-249_282-248insAAA	-	-	42	29	30	40%	12	8	4	18	12	6	0.618	32

Sample #	Chr	Pos	Wt N	Allele N	Variants - type 1	Prediction	Gene	Variants - type 2	N change	Protein change	Protein pos1	GQ	ABQ	Total coverage	Allele freq	AD	ADF	ADR	RD	RDF	RDR	PVAL	RBQ
Liv_07	12	133,219,561	T	TG	frameshift_variant	HIGH	POLE	transcript	c.4572dupC	p.Ser1525fs	4616/7840	57	33	36	44%	16	10	6	20	12	8	0.018	33
Liv_08	12	50,746,740	T	G	missense_variant	MODERATE	FAM186A	transcript	c.3875A>C	p.Asn1292Thr	3875/7127	22	22	33	21%	7	6	1	26	7	19	54.855	28
Liv_08	12	50,746,164	A	G	missense_variant	MODERATE	FAM186A	transcript	c.4451T>C	p.Ile1484Thr	4451/7127	22	20	40	18%	7	3	4	33	17	16	58.688	28
Liv_08	17	7,216,149	C	CA	frameshift_variant	HIGH	GPS2	transcript	c.909dupT	p.Ala304fs	1210/1379	66	30	19	79%	15	7	8	4	2	2	0.003	30
Liv_08	17	7,577,568	C	T	missense_variant	MODERATE	TP53	transcript	c.713G>A	p.Cys238Tyr	903/2579	55	32	14	86%	12	7	5	2	2	0	0.03	26
Liv_09	5	75,705,917	A	G	sequence_feature	MODERATE	IQGAP2	amino_acid_modification:Phosphoserine	c.46+6501A>G	-	-	86	30	36	61%	22	7	15	14	6	8	0	33
Liv_09	5	71,567,881	G	A	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.281+23477C>T	-	-	49	32	23	57%	13	7	6	10	6	4	0.112	31
Liv_10	5	75,756,307	G	C	sequence_feature	MODERATE	IQGAP2	amino_acid_modification:Phosphoserine	c.47-1088G>C	-	-	33	29	17	53%	9	5	4	8	5	3	4.635	32
Liv_10	5	71,543,159	A	G	sequence_feature	MODERATE	MRPS27	amino_acid_modification:N6-acetyllysine	c.282-9204T>C	-	-	38	31	29	38%	11	3	8	18	9	9	1.52	32
Liv_10	12	50,746,243	T	G	missense_variant	MODERATE	FAM186A	transcript	c.4372A>C	p.Thr1458Pro	4372/7127	26	21	29	28%	8	7	1	21	5	16	22.392	28
Liv_10	12	133,245,266	C	T	missense_variant	MODERATE	POLE	transcript	c.1981G>A	p.Ala661Thr	2025/7840	33	31	32	31%	10	6	4	22	12	10	0.426	32
Liv_10	17	7,577,581	A	C	missense_variant	MODERATE	TP53	transcript	c.700T>G	p.Tyr234Asp	890/2579	41	31	14	71%	10	8	2	4	0	4	0.763	35

Table 5-2: Candidate Variants. Panel A lists the genes having multiples hits among samples. Underlined genes are the affected DNA maintenance genes. Panel B shows details of the most affected genes. Abbreviation; GT: genotype; GQ: genotype quality; ABQ: average quality of variant-supporting bases; AD: depth of variant-supporting bases; ADF: depth of variant-supporting bases on forward strand; ADR: depth of variant-supporting bases on reverse strand; RDF: depth of reference-supporting bases in forward strand; RDR: depth of variant-supporting bases on reverse strand; PVAL: P-value from Fisher's Exact Test; and RBQ: average quality of reference-supporting bases.

A

Pathway identifier	Pathway name	#Entities found	#Entities total	#Reactions found	#Reactions total
R-HSA-168256	Immune System	184	2765	376	1456
R-HSA-162582	Signal Transduction	126	3002	486	1922
R-HSA-1430728	Metabolism	117	3461	281	2039

B

C

Affected pathways	Correspondent protien
Immune Response + Signal Transduction	ABI2, ADAM10, ADCY4, AKAP9, AKT1S1, AP2B1, APOB*, CAMK2B, CCR5, CENPE, DYNC1I1, DYNC1I2, EGR1, ERBB3, FGFR2, FLG, FRK, IL17RD, IL2RB, IL6R, IQGAP1, IQGAP2, IRS2, ITGA2B, KSR1, LAP3, MIB2, MMP9, MYH9, NCKAP1L, PDE1C, PPP2CB, PPP2R5D, PSMD1, RASA2, RASAL2, RPN1, SMURF1, SPAG1, SPRED1, SPTB, TEP1, TNRC6A, TP53, UBE2M, VCL, WASF2

*Underline lines indicate proteins affected by multiple variants

Figure 5-3: Protein Interaction Analysis by Reactome. Panel A describes the main affected protein pathways by gene mutations that are also found in Panel B. Green-to-yellow colour bar indicating the most affected sub-pathways. Panel C presents the most common affected proteins that are involved in both immune response and signal transduction pathways. Underlines indicate proteins affected by multiple mutations.

5.2.5. **Nucleotide substitutions**

Nucleotide substitution frequencies determined for the different variants groups (Figure 5-4). Twelve types of nucleotide substitutions were considered: four interchangeable substitutions - C=T and A=G as Ti, and eight substitutions of A=T, C=G, A=C and G=T as Tv. In the group of known polymorphic variants, the most frequent nucleotide changes were G > A, C > T, T > C and A > G at 18%, 17%, 16% and 16%, respectively (Figure 5-4, Panel A). The Ti/Tv ratio was 2.0, which was lower than the expected value of ~2.0, and different samples in the bar chart exhibited a very similar percentage of substitutions, suggesting the possibility of a common driver. That the deviation was observed in the polymorphic variants indicated a technical explanation.

In the group of unknown variants, the most frequent nucleotide changes were C > T, G > A, A > G, T > C and G > T at 24%, 20%, 11%, 8% and 8%, respectively, with a Ti/Tv ratio of 1.0 (Figure 5-4, Panel B). Moreover, the bar chart illustrates no substantial difference in the nucleotide substitution changes between samples. This suggests there is a mutational mechanism active and that it is common across samples.
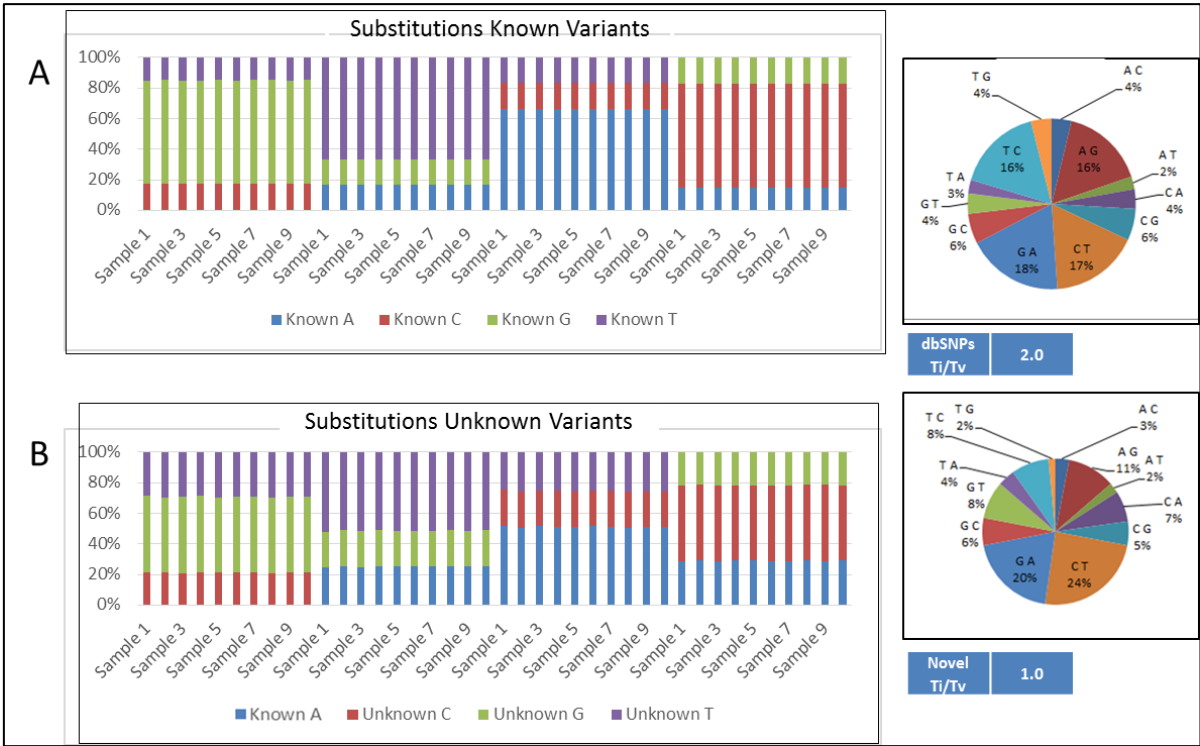
Figure 5-4: Nucleotide Substitutions. Panel A portrays the substitutions of the known variants between the 10 samples and as an accumulative pie chart. Similarly, Panel B depicts the substitutions of unknown variants.

5.2.6. Copy Number Alterations (CNA)

ControlFreec was used for detecting large CNAs and comparing the sequencing results to the FISH analysis already performed (Figure 5-5). A single copy deletion was observed at 17p in all 10 CLL patients (Figure 5-5, Panel A). In Figure 5-5, Panel B, therein is detailed information for large CNAs across samples, including 17p deletions. The majority of 17p (about 25.1 Mbp) was deleted in 10 samples, with an average of 17 Mbps and a range of 17p: 0-17,000,000 bp. Patient gender could also be confirmed relative to the clinical database by detecting single-copy differences in either the X or Y chromosomes in seven male and three female patients. Other CLL-related CNAs were also detected; 13q was deleted in the range of (13q: 41,500,000- 71,200,000 bp) in Liv_07 and Liv_09. Moreover,

most of 11q was deleted in the range of (11q: 67,100,000 – 117,900,000M bp) in Liv_05

and Liv_09.

The frequency of large CNAs differed between samples with Liv_04, Liv_01 and Liv_05

being the most affected with 6x-4x affected chromosomes, respectively (Figure 5-5, Panel

C), and the least was 2x chromosomes in Liv_02, Liv_03, Liv_06, Liv_07, Liv_08 and Liv_10.

Chromosome 5 was the second most affected chromosome (after chromosome 17) across

samples, affecting four cases in the range of (5q: 58,100,000- 170,900,000bp) for Liv_01

and Liv_05; and (5p: 2,100,000- 46,400,000bp) for Liv_04 and Liv_08. Chromosome 8 was

affected in three samples; single copy deletions for Liv_01 and Liv_04 in the range of

(8p:150,000-59,600,000 bp); and 3x copy multiplication for Liv_05 in the range (8q:

83,950,000- 146,364,022 bp). Chromosome 15 was affected in two samples; single copy

deletion for Liv_01 and 3x multiplication copy for Liv_02 over the range of

(15q:22,650,000-102,400,000 bp). Overall, these results suggest that there is a substantial

extent of CNA for all CLL samples studied and variability in the nature of the changes

found, with some similarities between small numbers of samples. This would suggest

defects in maintenance genes that foster gross chromosomal changes to be tolerated

rather than defects in a DNA repair mechanism that facilitate an accumulation of

replication errors.

A Chromosome 17 indel

B

Sample #	Chr	Start (bp)	End (bp)	Copy Number	Note	Lesion Size (bp)
9	3	111,100,000	146,350,000	1	loss	35,250,000
1	5	85,000,000	125,000,000	1	loss	40,000,000
1	5	177,150,000	177,350,000	1	loss	200,000
4	5	2,450,000	21,450,000	1	loss	19,000,000
5	5	58,100,000	99,250,000	1	loss	41,150,000
8	5	2,750,000	46,400,000	1	loss	43,650,000
3	6	93,200,000	170,950,000	1	loss	77,750,000
1	8	150,000	45,000,000	1	loss	44,850,000
4	8	150,000	59,600,000	1	loss	59,450,000
5	8	83,950,000	146,364,022	3	gain	62,414,022
4	10	98,800,000	126,450,000	1	loss	27,650,000
5	11	89,800,000	117,050,000	1	loss	27,250,000
9	11	67,150,000	117,900,000	1	loss	50,750,000
10	12	200,000	7,750,000	3	gain	7,550,000
7	13	51,500,000	65,550,000	1	loss	14,050,000
9	13	41,550,000	71,200,000	1	loss	29,650,000
1	15	22,650,000	44,050,000	1	loss	21,400,000
2	15	63,900,000	102,400,000	3	gain	38,500,000
1	17	4,100,000	18,950,000	1	loss	14,850,000
2	17	-	18,950,000	1	loss	18,950,000
3	17	-	18,300,000	1	loss	18,300,000
4	17	4,100,000	18,950,000	1	loss	14,850,000
5	17	-	19,400,000	1	loss	19,400,000
5	17	44,800,000	68,100,000	3	gain	23,300,000
6	17	-	18,350,000	1	loss	18,350,000
7	17	550,000	18,300,000	1	loss	17,750,000
8	17	-	18,300,000	1	loss	18,300,000
9	17	4,650,000	16,700,000	1	loss	12,050,000
10	17	-	18,300,000	1	loss	18,300,000
4	18	-	51,300,000	1	loss	51,300,000
1	19	7,000,000	8,900,000	1	loss	1,900,000
1	19	24,500,000	24,600,000	3	gain	100,000
6	20	-	20,050,000	1	loss	20,050,000

C

Sample #	Liv_01	Liv_02	Liv_03	Liv_04	Liv_05	Liv_06	Liv_07	Liv_08	Liv_09	Liv_10	No. CNA/ chromosome
TP53 dbm	In	Out	In	Out	Out	In	Out	In	In	In	-
Gender	M	M	M	F	M	F	M	M	M	F	-
chr3	-	-	-	-	-	-	-	-	Del	-	1
chr5	Del	-	-	Del	Del	-	-	Del	-	-	4
chr6	-	-	Del	-	-	-	-	-	-	-	1
chr8	Del	-	-	Del	Mult	-	-	-	-	-	3
chr9	-	-	-	Del	-	-	-	-	-	-	1
chr10	-	-	-	Del	-	-	-	-	-	-	1
chr11	-	-	-	-	Del	-	-	-	Del	-	2
chr12	-	-	-	-	-	-	-	-	-	Mult	1
Chr13	-	-	-	-	-	-	Del	-	Del	-	2
chr15	Del	Mult	-	-	-	-	-	-	-	-	2
chr17	Del	Del	Del	Del	Del-Mult	Del	Del	Del	Del	Del	10
chr18	-	-	-	Del	-	-	-	-	-	-	1
chr19	Del-Mult	-	-	-	-	-	-	-	-	-	1
chr20	-	-	-	-	-	Del	-	-	-	-	1
No. CNA/ sample	5	2	2	6	4	2	2	2	4	2	31
chrX	Del	Del	Del	-	Del	-	Del	Del	Del	-	-
chrY	Double Del	Del	Del	Double Del	Del	Double Del	Del	Del	Del	Double Del	-

.

Figure 5-5: Copy Number Alterations (CNAs) by ControlFREEC. Panel A depicts a ControlFreeq graph of chromosome 17, including the 17p Deletion. Green dots represent diploid copies, blue dots deleted copies and red dots other gained copies. In Panel B, detailed analysis of large CNVs in 10 CLL samples is portrayed. Related to Panel B, Panel C provides a summary of CNVs among different samples. Abbreviations - Del: Deletion; Mult: Multiplication; M: Male; and F: Female

5.2.7. Translocations and False Positive Results

The BreakDancer tool was employed to detect translocations, including those affecting

multiple samples. The primary analysis shows that there was an average of 54

intrachromosomal translocations (ITXs) along with 138 interchromosomal translocations

(CTXs) and indels (Table 5-3). A number of interchromosomal CTXs were common to all

samples and this was considered to be highly unlikely given the known biology of CLL

which recognises only the 17p, 11q, 13q and trisomy 12 as frequently occurring

chromosomal alterations apparent in CLL. The novel translocations were therefore

further investigated. A representative example wax translocation t(8;11) (Figure 5-6);

Panel A). The average size of the lesion is 470 bp with a high read coverage of 50-200

reads. Sample Liv_05 showed two t(8;11) translocations at chromosome locus

11:38,812,243 with either 8:52,731,606 or 8:52,730,634 and lesion sizes of 79bp and

583bp, respectively (Figure 5-6, Panel B). This was highly suggestive of errors introduced

by misalignment of reads to a repetitive sequence, which would lead to both the high

read depths and multiple destination points for the candidate translocation. The BLAST

alignment tool (1) was therefore applied to reads corresponding to the 11:38812235

lesion site (Figure 5-6, Panel C). The alignment results demonstrated low complexity

regions and multiple alignments with a high score within the breakpoint site. This

confirmed the detection artefact and allowed common translocations to be assigned a

status as FPs.

Table 5-3: Number of Translocations per Sample. There were on average 54 CTX lesions per sample, eight of which are common. ITX lesions were more profound with 138 lesions per sample.

Samples (Liv_01-Liv_10)	No. CTXs	No. ITXs
Average Somatic	54	138
Average Common	8	-

Panel A

sample	#Chr1	Pos1	coverage	affected gene	lesion difference (1)	#Chr2	Pos2	coverage	affected gene	lesion difference (2)	Size	Score	num_Reads	average	allele frequency
7	8	52,731,777	113	PCMTD1	-412	11	38,812,282	54	none	0	-479	99	86	83.5	103.0
5	8	52,731,606	147	PCMTD1	-583	11	38,812,243	62	none	-39	-478	99	36	104.5	34.4
1	8	52,731,430	177	PCMTD1	-759	11	38,812,235	80	none	-47	-479	99	51	128.5	39.7
2	8	52,731,429	149	PCMTD1	-760	11	38,812,185	46	none	-97	-481	99	22	97.5	22.6
10	8	52,731,423	176	PCMTD1	-766	11	38,812,264	74	none	-18	-490	99	32	125	25.6
8	8	52,731,423	166	PCMTD1	-766	11	38,812,187	59	none	-95	-481	99	35	112.5	31.1
9	8	52,731,422	166	PCMTD1	-767	11	38,812,253	65	none	-29	-480	99	31	115.5	26.8
4	8	52,731,419	128	PCMTD1	-770	11	38,812,231	80	none	-51	-479	99	40	104	38.5
6	8	52,731,419	197	PCMTD1	-770	11	38,812,070	69	none	-212	-474	99	37	133	27.8
10	8	52,730,713	207	PCMTD1	0	11	38,812,264	74	none	-18	-490	99	39	140.5	27.8
8	8	52,730,697	198	PCMTD1	-16	11	38,812,187	59	none	-95	-481	99	39	128.5	30.4
9	8	52,730,696	204	PCMTD1	-17	11	38,812,253	65	none	-29	-480	99	53	134.5	39.4
5	8	52,730,634	177	PCMTD1	-79	11	38,812,243	62	none	-39	-478	99	43	119.5	36.0
2	8	52,730,566	158	PCMTD1	-147	11	38,812,185	46	none	-97	-481	99	28	102	27.5
1	8	52,730,510	177	PCMTD1	-203	11	38,812,235	80	none	-47	-479	99	56	128.5	43.6
3	8	52,730,498	162	PCMTD1	-215	11	38,811,555	82	none	-727	-472	99	49	122	40.2
4	8	52,730,487	135	PCMTD1	-226	11	38,812,231	80	none	-51	-479	99	54	107.5	50.2
6	8	52,730,442	203	PCMTD1	-271	11	38,812,070	69	none	-212	-474	99	43	136	31.6

Figure 5-6: FP Translocations. Panel A provides details of T(8;11) as an example of common translocations, where sample CLL-05 is pointed out as a representative sample. Panel B describes translocations at t(8;11), where the blue segments represent an affected segment and the red bar is the locus breakpoint. Based on BLAST, Panel C shows the alignment results at the lesioned site; red lines indicate the highly repetitive sequence of the lesion.

5.3.   Discussions and Conclusions

This chapter seeks to investigate the possibility of chromosomal instability CINs in P53-inactivated CLL. This group of patients was selected because of its severe prognosis such that it is refractory to chemotherapy and undergoing clonal evolution. Whether such impacts are also caused by defects in other genes is not clear. To understand this, a genes panel and WGS were developed to screen CINs and somatic mutations in 194 DNA maintenance genes of a cohort of 10 CLL patients.

After data alignment, the WGS coverage results yielded the expected coverage with approximately 30x average reads of 90 billion reads per sample, eliciting a ≥ 20% AF detection limit. In terms of genome coverage, 96.7% genome coverage was found to have 10x read depth. This was also expected as the aim was low-pass sequencing. Different genome regions also vary in sequence complexities, affecting their relative proportions during PCR amplification. However, there was still sufficient data to detect a large proportion of the variants present.

In terms of validating the data, 17p deleted cases were confirmed. Expected TP53 mutations were also established as well as two POLE variants identified by exome sequencing as reported in the previous results chapter, specifically section 4.2.8.

A strategy was developed for identifying candidate somatic alterations without having the possibility of comparing to germline controls. Employing dbSNP and COSMIC as technical conrols, variant grouping minimized FPs. At the optimisation stage, the investigation of coverage provided evidence that variants at repetitive regions may produce significant mismatches (310), which were observed as very high coverage regions (>40% AF), and

quite a few were common between cases (311); these could be safely ignored as true

variants. Overall, there were 188,544 variants added as somatic to the study data set,

while 60% of TP53 mutations were detected and the remaining 40% undetected variants

were skipped because variants had already been identified within the known group. This

limitation was expected using tumour-only data. The data also detected 2x POLE somatic

variants which had been discerned earlier by deep exon sequencing.

Nucleotide substitution, the actual SNP Ti/Tv of 2 was at a lower than the expected level

of 2.1, and high rate of novel Ti/Tv of 1 (expected ~1) (312). This is consistent with the

activity of AICDA proteins in hypermutation processes during B-cell maturation.

The advantage of WGS is the possibility for variant detection across the genome,

including the exome sequence. This enabled genes with multiple hits to be detected.

Chromosome 14 mutations, especially in the IGHV genes, were noticeable in specific

samples (Liv_05, Liv_07 and Liv_10), reflecting their mutated IGHV status. Chromosome 5

was also observed to acquire multiple hits in IQGAP2 and MRP27. Moreover,

Chromosome 12 showed hits in FAM186A and POLE. Chromosome 1 had multiple hits for

the neuroblastoma breakpoint family members (NBPF1, NBPF9, NBPF10, NBPF14 and

NBPF20). Chromosome 16 was affected by different Nuclear Pore Complex-Interacting

Protein Family Members (NPIPB 5, NPIPB 6 and NPIPB 11). On chromosome 19, Zinc

Finger Proteins (ZNF98, ZNF418 and ZNF560) were affected in multiple samples, as well.

In summary, there are many interesting genes which may possibly relate to CLL

progression, though these genes would need to be analysed in the larger cohort for their

significance for CLL to be determined and tumour and germline samples would be required for validation of somatic origin.

The reactome offered insight into the protein interactions that may be affected by the variants observed. The immune system and signal transduction were highly affected; many relevant proteins were involved in both pathways. These pathways could be possibly related to B-cell development and CLL. This data supports the importance of immune response and signal transduction in CLL progression (309).

ControlFREEC was provided with a comprehensive copy number analysis. The results were validated by confirming 17p deletions in all samples as well as by gender confirmation with XY chromosomes. In addition, other CLL deletions were observed, such as 13p and 11q deletions. There are also other common CNAs that are implicated, such as chromosomes 5, 8, and 10 which affected four (40%), three (30%) and two (20%) of total patients, respectively. The common CINs were observed on chromosomes 5 and 8, affecting both arms. In addition, several samples had more indels than others, including CLL-04, CLL-5 and CLL-09 featuring six, four and four indels, respectively. With this study cohort, chromosome 17 was significantly impacted by LOH at the 17p arm and TP53 somatic mutations on the other strand, where inactivated P53 could suppress its function in regulating the cell cycle. This confirms TP53's prominent role in CLL pathogenesis and clonal evolution. Regarding other CINs, this data supports the notion that acquired CNAs are elevated in progressive CLL (313). However, the results would necessitate further validation by a clearly defined plan through a larger cohort, germline control and serial sampling (289).

Regarding translocations, initial analysis via BreakDancer showed that recurrent translocations resides within low complex regions. Observations suggested false findings owing to algorithm bias and mismatch alignment on those regions. This reflects the challenges in defining translocations and suggests using a germline control as an initial validation strategy. In parallel, different validation approaches can be utilised, such as multiplex FISH, CGH or SNPs microarrays (312).

A large number of CNAs were observed, and these were amongst the most variable features between different CLL samples, raising the possibility that they are one of the key drivers of differences between CLL cases in terms of drivers of progression and poor outcomes. Their very variability and large scale make it technically challenging to validate this possibility, and therefore this was noted but not pursued. It would useful to compare instances of CNA across a large cohort in comparison to TP53 pathway alterations to determine any possible relationship.

After combining WGS and targeted-deep sequencing, POLE emerged as one affected DNA repair gene by somatic aberrance in 20% of the cohort. Therefore, the next chapter concerns validating POLE in a larger cohort of patients and determining whether POLE is aberrant in CLL.

6. Chapter 6: Investigation of the Significance of DNA Maintenance Gene Mutations in Chronic Lymphocytic Leukaemia (CLL)

6.1. Introduction

Following the finding of a bias in the types of substitutions present in our CLL cases and multiple mutations in POLE, this chapter addresses: (1) whether inactivating mutations in POLE is related to inactivated TP53 mutations in CLL; and (2) whether they could be related to CLL progression.

6.1.1. POLE and Cancers

POLE comprises four subunits and the POLE gene encodes the catalytic and exonuclease subunit. It has an important function in leading and lagging strand synthesis during DNA replication and also DNA maintenance (314-317). S-phase checkpoints are linked to the replication apparatus by its non-catalytic carboxyl terminus domain (318). It is active in both BER and BER and may also be relevant to recombinational repair. POLE is located on chromosome 12 at 12q24.33 and has 52 exons. The structure of its protein has three main functional domains - the exonuclease domain, multifunctional domain and catalytic subunit domain (Figure 6-2) (319).

Germline POLE exonuclease domain mutations (EDMs) are predisposed to colorectal cancer (320). In addition, they are found in 5-10% of sporadic colorectal cancers where they are associated with a hypermutator, microsatellite-stable phenotype (321). EDMs are also found in 7% of endometrial cancers and were associated with a high frequency of base substitutions, especially G:C>T:A Tv (322). Enhanced mutation rates with an excess of C>A Tv have also been reported for endometrial cancer having EDMs (236). This strongly supports a

potential role for EDMs in POLE causing a hypermutator phenotype and contributing to genomic instability (319, 323). However, endometrial cancer cases with the hypermutation phenotype had a better prognosis and this is possibly caused by enhanced immunogenicity as a result compared to other cases (324). The COSMIC database reports hotspot somatic mutations in colorectal and endometrial cancers; the hotspots reside within POLE functional domains, predominantly in the exonuclease domain (325-327).

6.1.2. Rationale of the Study

Two somatic variants were found in POLE in two separate samples of the 10 CLL samples studied. One variant was a non-synonymous somatic mutation resulting in p.Ala661Thr and the other variant was a frameshift insertion causing p.Ser1525fs. Both mutations reside within the functional domains of POLE proteins. This indicates that POLE could be mutated in TP53-inactivated CLL cases and contribute to their genomic instability. The purpose of the investigation was therefore to determine whether POLE mutations are associated with inactivated p53 in a larger cohort of CLL patients. Sanger sequencing was used to detect mutations in the potential POLE hotspots. The goal was also to determine whether such mutations were associated with the disease status of CLL.

6.2. Results

6.2.1. Clinical and Molecular Characteristics of the CLL Cohort

The validation cohort contained a total of 49 samples; 24 samples with TP53 mutated (and/or 17p deleted) and 25 wild-type samples. Regarding TP53 mutated cases, 20/24 (83%) cases were TP53 mutated/17p deleted. Furthermore, 62% of the tested samples were IGHV hypermutated . Twenty-four of 49 patients had been treated with a cytotoxic drug, including

DNA-damaging agents, nine of them (~38%) had TP53 aberrations. Regarding other CLL-associated genomic aberrations, a 13q deletion was evident in (82%), an 11q deletion in (10%) and trisomy 12 in (19%). The ultimate survival ranged between 91-4790 days (1718 days on average). Details of the samples, including the clinical and biological characteristics, are found in Table 6-1. 24 samples were from the CLL210 trial and had been initially screened by FASAY (GCLP Laboratory), and then by Sanger sequencing using optimised primers (Table 6-2). The remaining 25 samples were from the local biobank and had been FASAY and Sanger sequenced for TP53 by Dr Gillian Johnson.

Table 6-1: Clinical and Molecular Data for the CLL Validation Cohort. Details which include age at sampling, gender, pre-treatment WBC and lymphocyte counts, IGVH mutation percentage, 17p deletion and TP53 mutation status, Binet and Rai stages, treatment history, OS and other genomic aberration statuses, such as deletion of 11q, 13q and trisomy 12. Abbreviations – F: Female; M: Male; Mut: Mutated; Cyclo - Cyclophosphamide

Sample #	Age at sampling	Gender	Lymphocyte count pre-treatment	VH%	17p del	TP53 sequence	Treatment history	Del11q22.3	Del 13q14	Tri 12
1	54	M	252.9	N/A	Deleted	Mut	None	N/A	N/A	N/A
2	67	M	146.5	N/A	Deleted	Mut	None	N/A	N/A	N/A
3	58	F	92.7	N/A	Deleted	Mut	None	N/A	N/A	N/A
4	59	M	234.7	N/A	Deleted	Mut	None	N/A	N/A	N/A
5	56	F	40.1	N/A	Deleted	Mut	None	N/A	N/A	N/A
6	64	M	95.5	N/A	Deleted	Mut	1. Alemtuzumab & methylprednisolone, 10 cycles, CR. 2. Rituximab & Bendamustine, 8 cycles, PD	N/A	N/A	N/A
7	60	M	170.8	N/A	Deleted	Mut	1. Chlorambucil, 24 cycles. 2. Fludarabine/ cyclophosphamide, 6 cycles.	N/A	N/A	N/A
8	69	F	290.3	N/A	Deleted	Mut	None	N/A	N/A	N/A
9	68	M	116.5	N/A	Deleted	Mut	1. Fludarabine, 6 cycles, CR. 2. fludarabine & cyclophosphamide, CR. 3. fludarabine + cyclophosphamide + rituximab, 4 cycles, CR	N/A	N/A	N/A
10	65	F	6.3	N/A	Deleted	Mut	None	N/A	N/A	N/A
11	42	M	1.9	N/A	Deleted	Mut	1. Alemtuzumab with high dose methylprednisolone, 7 cycles, PR. 2. fludarabine/ cyclophosphamide/ alemtuzumab conditions for reduced intensity allograft, 1 cycle, CR	N/A	N/A	N/A
12	60	F	8.5	N/A	Deleted	Mut	bendamustine, rituximab, prednisolone, 1 cycle, PR	N/A	N/A	N/A
13	77	M	393.4	N/A	Deleted	Mut	None	N/A	N/A	N/A
14	67	M	51.3	N/A	Deleted	Mut	1. chlorambucil, 6 cycles, CR. 2. fludarabine and	N/A	N/A	N/A
15	N/A	M	88	1.7	Deleted	Mut	Fludarabine, CR	Deleted	normal	Normal
16	N/A	N/A	8.9	N/A	Deleted	Mut	None	N/A	Deleted	N/A
17	N/A	N/A	64.7	4.76	Deleted	Mut	None	Normal	Deleted	Normal
18	N/A	M	N/A	4.47	Deleted	Mut	None	Normal	Deleted	Normal
19	N/A	N/A	N/A	0	Normal	Mut	None	N/A	N/A	N/A
20	N/A	N/A	51.6	0	Normal	Mut	Flu 2 cycles	Deleted	Normal	Normal
21	71	F	159.3	N/A	Deleted	Wt	None	N/A	N/A	N/A
22	52	M	40.1	N/A	Deleted	Wt	Dexamethasone, 2 cycles, PR	N/A	N/A	N/A
23	69	M	235.7	N/A	Deleted	Wt	None	N/A	N/A	N/A
24	50	M	23.7	N/A	Deleted	Wt	None	N/A	N/A	N/A
25	N/A	N/A	N/A	N/A	Normal	Synonymous mutation	None	N/A	N/A	N/A
26	72	M	116.4	N/A	Normal	Wt	1. Chlorambucil, 8 cycles, CR. 2. FC, 8 cycles, PR. 3. RCHOP, 6 cycles, SD	N/A	N/A	N/A
27	63	M	140.4	N/A	Normal	Wt	Rituximab, fludarabine, cyclophosphamide, 6 cycles, PR	N/A	N/A	N/A
28	72	F	189.8	N/A	Normal	Wt	Fludarabine & cyclophosphamide, 6 cycles, PR	N/A	N/A	N/A
29	65	M	1.8	N/A	Normal	Wt	rituximab / dexamethasone, 8 cycles, PR	N/A	N/A	N/A
30	58	M	3.1	N/A	Normal	Wt	1. rituximab, cyclophosphamide, fludarabine, 6 cycles, CR. 2. rituximab, cyclophosphamide, fludarabine, 5 cycles, PD	N/A	N/A	N/A
31	N/A	N/A	10.3	7.07	Normal	Wt	None	Normal	Deleted	Normal
32	N/A	N/A	8.8	0.34	Normal	Wt	None	Normal	Deleted	Yes
33	N/A	N/A	N/A	10.18	Normal	Wt	None	Normal	Deleted	Normal

Sample #	Age at sampling	Gender	Lymphocyte count pre-treatment	VH%	17p del	TP53 sequence	Treatment history	Del11q22.3	Del 13q14	Tri 12
34	N/A	M	N/A	0	Normal	Wt	CHOPx8,	Normal	normal	Yes
35	N/A	N/A	N/A	N/A	Normal	Wt	None	N/A	N/A	N/A
36	N/A	F	8.4	N/A	Normal	Wt	None	Normal	Deleted	Normal
37	N/A	N/A	N/A	N/A	Normal	Wt	None	Normal	Deleted	Normal
38	N/A	N/A	22.4	N/A	Normal	Wt	None	Normal	Deleted	Normal
39	N/A	N/A	120	N/A	Normal	Wt	None	Normal	Normal	Yes
40	N/A	M	13.4	N/A	Normal	Wt	Fludarabine+Cyclo (CR) 2-Cyclo (CR, relapsed).	Normal	Deleted	Normal
41	N/A	N/A	26.6	14.01	Normal	Wt	None	N/A	N/A	N/A
42	N/A	M	87.5	6.12	Normal	Wt	None	Normal	Deleted	Yes
43	N/A	F	23.1	N/A	Normal	Wt	None	Normal	Deleted	Normal
44	N/A	F	20.7	N/A	Normal	Wt	None	Normal	Deleted	Normal
45	N/A	F	9.2	5.35	Normal	Wt	None	Normal	Deleted	Normal
46	N/A	F	13.8	N/A	Normal	Wt	None	Normal	Deleted	Normal
47	N/A	F	N/A	N/A	Normal	Wt	None	Normal	Deleted	Normal
48	N/A	N/A	29.6	N/A	Normal	Wt	None	Normal	Deleted	Normal
49	N/A	N/A	N/A	9.22	Normal	Wt	None	Normal	Deleted	Normal

6.2.2. DNA Extraction

gDNA was extracted from MNCs of patient blood samples as described in Materials and Methods, section 2.5.1.

6.2.3. Sanger Sequencing of TP53 and POLE

TP53 sequencing data was not originally available for 24 of 49 CLL samples. The same PCR and sequencing approaches as originally used were therefore applied to determine their mutation status in TP53 coding exons 2-11.

PCR was performed for TP53 and POLE using optimised primers listed in Table 6-2 and Table 6- 3, respectively, as described in Materials and Methods, section 2.7.1, based on Promega GoTaq DNA polymerase.

The TP53 and POLE genomic structure, including the coding regions, are shown in Figure 6- 1 and Figure 6-2. POLE is depicted with the positions of mutations recorded in COSMIC (328). Four mutational hotspot regions were selected for POLE screening plus two other positions.

These included P.P286R, P.V411L and P.A456P at exons 9, 13 and 14, respectively (325).

Primers were synthesised by ITD Ltd., and the PCRs were optimised using control human

DNA before utility with CLL sample DNA (not shown).

PCRs from the CLL samples were purified and sequenced by the Sanger method as described

in the Materials and Methods, sections 2.5.2.2 and 2.7.3. Purified material was assessed for

amount and quality by agarose gel electrophoresis and Qubit fluorimetry (not shown) as

described in Materials and Methods, section 2.6.3 and 2.6.2, respectively
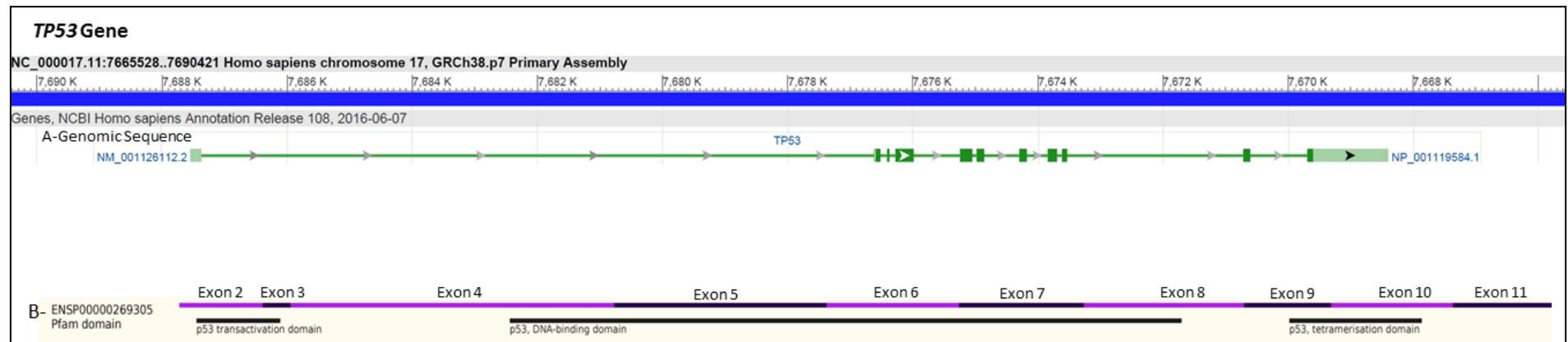
Figure 6- 1: TP53 Genomic Structure, Including the TP53 Protein Coding Sequence and its Functional Domain. A. TP53 genomic sequence and location on chromosome 17 (blue line). The green line represents the transcript, NM 001126112.2, including the exons of interest, namely 2-11. B. Representation of the protein-coding line, ENSP00000269305 (black and purple lines) featuring the protein functional domains, Pfam (the lines underneath the protein-coding line), which shows the three main domains: p53 transcriptional domain, the DNA-binding domain and the tetramerisation domain. Panel A is taken from online databases of Ensemble 89, while Panel B is from Gene [TP53]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – 2017 Jan 30. Available from: https://www.ncbi.nlm.nih.gov/gene/
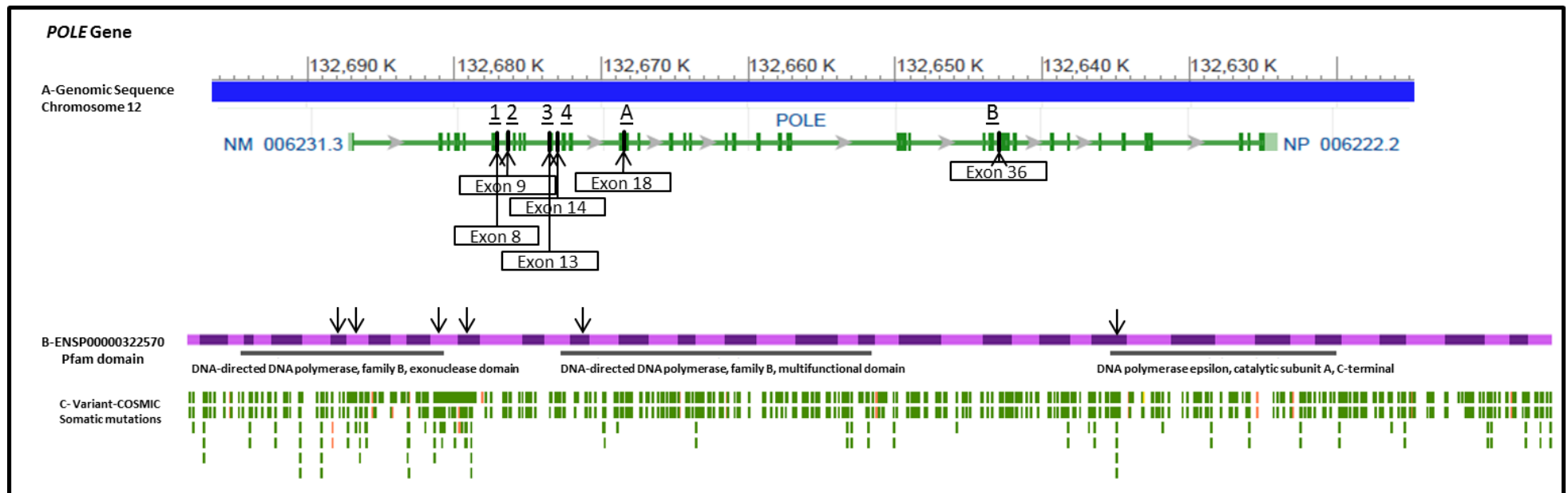
Figure 6-2: POLE Structure Containing the POLE Genomic Sequence and its Protein Coding Sequence. A. POLE genomic sequence and location on chromosome 12 (blue line). The green line represents POLE sequence (accession #: NM 005231.3), including the exons of interest; hot spots are marked in order from 1 to 4. A and B indicates variants A and B, respectively. B. POLE protein coding sequence with (accession #: ENSP000000322570 (in dark-light purple)) featuring the protein functional domains, Pfam, which shows three main domains; the exonuclease domain, the multifunctional domain and the catalytic domain. Arrows show locations of the affected exons. C. Variant spots alongside the protein-coding line using the COSMIC database. The green spots represent the somatic variants and the orange spots represent the somatic indels from different cancers. Panels A and C are taken from online databases of Ensemble 89 and the COSMIC v81, respectively. Panel B is adapted from Gene [POLE]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – 2017 Jan 30. Available from: https://www.ncbi.nlm.nih.gov/gene/.

Table 6-2: TP53 Primer Sets. DNA sequence was Provided by NCBI with BLAST-Primer Software Used for Designing Primers.

Primer set (size of product)	Primer sequence (5' – 3')	TP53-targeted exon(s)
TP1 (385bP)	Fp: CAG GGT TGG AAG TGT CTC AT Rp: GAA AAG AGC AGT CAG AGG AC	2-3
TP2 (358bp)	Fp: GTC CTC TGA CTG CTC TTT TC Rp: GCC AGG CAT TGA AGT CTC AT	4
TP3 (550bp)	Fp: TCT TTG CTG CCG TCT TCC AG Rp: CAG CAG GAG AAA GCC CCC	5-6
TP4 (322bp)	Fp: CCT CAT CTT GGG CCT GTG TT Rp: GTC CCA AAG CCA GAG AAA AG	7
TP5 (509bp)	Fp: GGG AGT AGA TGG AGC CTG GT Rp: TGT CTT TGA GGC ATC ACT GC	8-9
TP6 (318bp)	Fp: TGC ATG TTG CTT TTG TAC CGT C - Rp: TCA GCT GCC TTT GAC CAT GA	10
TP7 (262bp)	Fp: CCT TCA AAG CAT TGG TCA GG Rp: GCA AGC AAG GGT TCA AAG AC	11

Table 6- 3: POLE In-House-Optimised Primer Sets

Primers set (size of product)	Primer sequence (5' – 3')	POLE-targeted exon
POLE-hs1 (289 bp)	Fp: GTC AGA TTC ACT CTC CAG CAC Rp: CAG GGT TGG GTC GCT GC	8
POLE-hs2 (405bp)	Fp: TAC AGC TGG AGG TCG GAA C Rp: GTC TTA GGG TCC TTC TCC C	9
POLE-hs3 (364bp)	Fp: GAC CGG CAC AGG ACA AAA C Rp: GCT GCA TGT TAG AAT CAT CCT G	13
POLE-hs4 (238bp)	Fp:TGA GGA GGC CAG GGT GCC GA Rp:AGG CCA GGC TTT GCT TTC TGT G	14
POLE-VA(323bp)	Fp: TGG TAT TCG CTG CGA CTG G Rp: ACT TCC CGT GTC AGA GTC G	18
POLE-VB(279bp)	Fp: TGT AGG CGA GCA GGA ATC G Rp: TCG TGA TTG AAT GGC AGT GC	36

6.2.4. TP53 Screening

After optimisation, the TP53 primers amplified a single band from genomic DNA (Figure 6-3). PCR was then applied to the CLL samples. Agarose gel electrophoresis showed only faint bands for primers TP3 and TP5 (see Figure 6-3, Panels C and E), although the DNA yield was sufficient for Sanger sequencing. Primers TP1, covering exons 2 and 3, including the transactivation domain, showed two bands, one at 385 bp and a second at 258 bp (Figure 6-3, Panel A). The primer sets TP2-TP5 covered exons 4-8, which constitute the p53 DBM, and yielded bands at 358 bp, 550 bp, 322 bp and 509 bp, respectively (Figure 6-3, Panels B-E). Exons 9-11, which include the tetramerisation domain, were covered by the primer sets TP5, which elicited PCR products of 509 bp, 318 bp and 262 bp, respectively (Figure 6-3, Panels E-G).

The same PCR primers were also used for Sanger sequencing. The results confirmed coverage of the 10 coding exons 2-11 as expected (Figure 6-4). Strong electropherogram signals were apparent throughout the sequences. Several sequences, such as exons 7 and 10, had delayed signals in certain nucleotide positions, possibly caused by overly concentrated input DNA. In addition, various sequences had low background noise, but the sequence could still be verified by bi-directional sequencing with either forward or reverse primers.

From 25 samples examined, wild-type sequence alleles were detected in 10 samples, but the remaining 21 samples had 23 somatic mutations with samples 8 and 9 each having two different mutations (Figure 6-5 and Table 6-4).
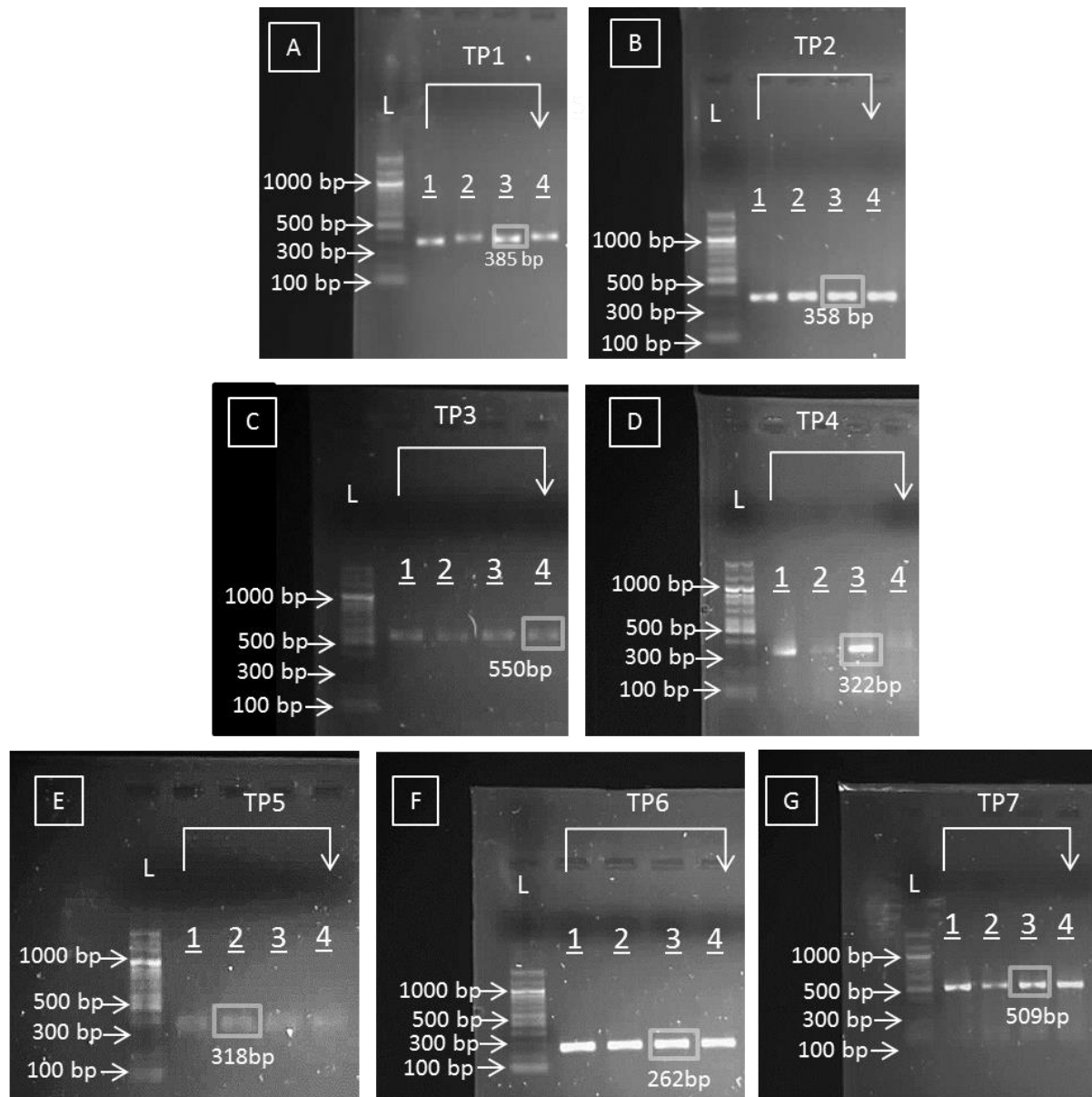
Figure 6-3: PCR Amplification for TP53 Screening. Panels A-G show agarose gel electrophoreses of PCR products amplified using TP53 primer sets TP1-TP7, respectively. Four representative samples with the 100-bp DNA ladder control (L).
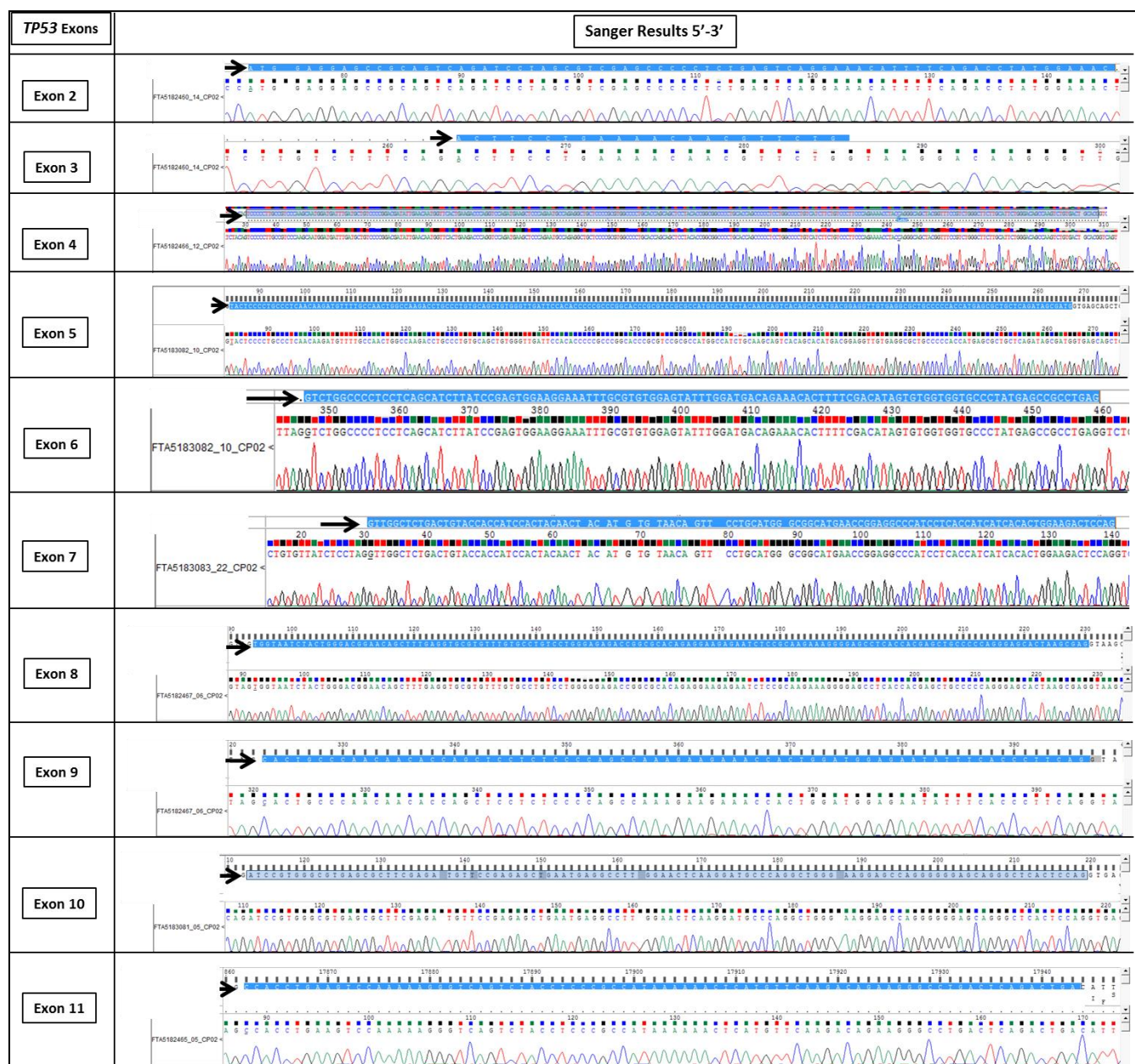
Figure 6-4: Sanger sequencing results of TP53 PCR products. The electropherograms portray the TP53 coding sequences (exons 2-11 in the 5'-3' direction) amplified using gDNA from the representative patient samples, 05, 06, 10, 12, 14 and 22 in terms of PCR and analysed by Sanger sequencing. The sequence of each exon is indicated by black arrows and highlighted in blue at top of each panel. The same results were achieved using 24 other CLL samples - samples 1-24.

6.2.5.   Refined TP53 Status in the CLL Cohort of Patients Studied

To address whether any mutations identified in POLE were associated with TP53 alterations, two groups of subjects (n=49) were selected based on their TP53 status. All of the samples already elicited information regarding 17p deletions (as detected by FISH). Twenty-five samples had the TP53 somatic mutation and/or deletion (as a result of a 17p deletion); of which 18 having both aberrations, four having only a deletion and three with just a mutation (Table 6-4). The remaining half (n=24) had neither type of TP53 genomic modification (Table 6-4).

Table 6-4: TP53 Mutation and/or 17p Deletion Status in the 25 Newly TP53-Sequenced Samples of the Cohort.

Sample #	17p deletion	TP53 gene	Nucleotide change	AF(%)	Codon change	Protein change
1	Y	Mut	C>T	60	TCC--TTC	Ser 241 Phe
2	Y	Mut	A Fs Del	20	CAA--CAT fs Del	Gln 52 His Fs Del
3	Y	Mut	A>G	80	GAA—GGA	Glu 286 Gly
4	Y	Mut	A>G	70	AGA—ACA	Arg 280 Thr
5	Y	Mut	G>T	20	GTC—TTC	Val 157 Phe
6	Y	Mut	G>C	80	CGT—CAT	Arg 273 His
7	Y	Mut	A>G	40	TAC—TGG	Tyr 163 Cys
8	Y	Mut	C>T, G Fs Del	20, 40	CAG—TAG, AAG--AAA FS del	Gln 104 x Lys 139 Lys Fs Del
9	Y	Mut	G>T, G>A	15, 40	TGT—TTT, GAA--CAA	Cys 277 Phe Glu 258 Gln
10	Y	Mut	A>G	75	TAT--TGT	Tyr 205 Cys
11	Y	Mut	T>G	95	TTC--TGC	Phe 341 Cys

Sample #	17p deletion	TP53 gene	Nucleotide change	AF(%)	Codon change	Protein change
12	Y	Mut	G>T	40	TGT--TTT	Cys 238 Phe
13	Y	Mut	C>T	90	TGC--TCC	Cys 242 Ser
14	Y	Mut	C>T	50	CGG--TGG	Arg 282 Trp
15	Y	Mut	-	-	-	-
16	Y	Mut	-	-	-	-
17	Y	Mut	-	-	-	-
18	Y	Mut	-	-	-	-
19	N	Mut	-	-	-	-
20	N	Mut	-	-	-	-
21	Y	Wt	-	-	-	-
22	Y	Wt	-	-	-	-
23	Y	Wt	-	-	-	-
24	Y	Wt	-	-	-	-
25	N	Mut	G>C	70	AGA--AGG	Arg 280 Arg

Y: with 17p deletion,   N: without 17p deletion;   Mut: mutant;   Wt: wild-type
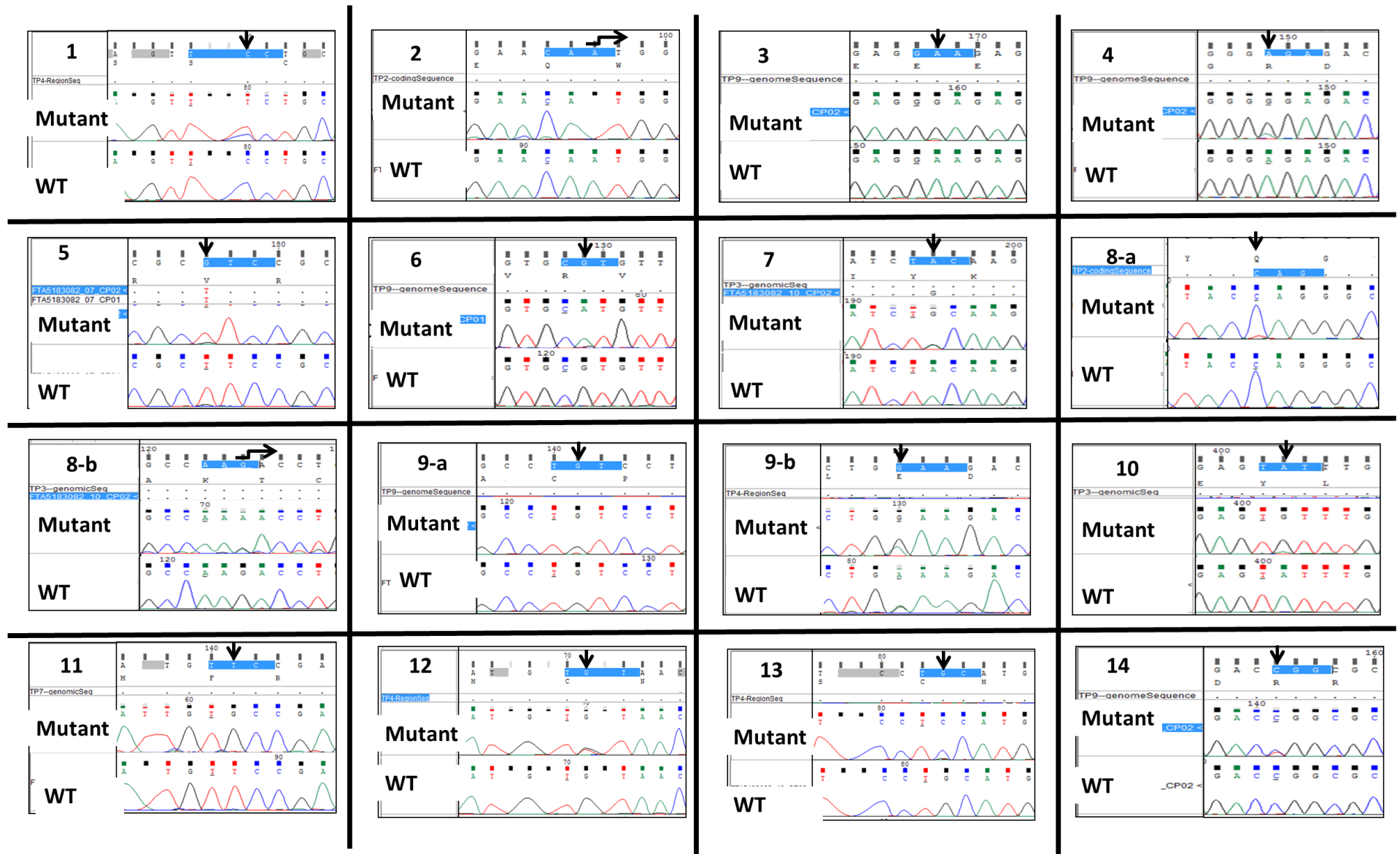
Figure 6-5: Sanger Sequencing of 25 Newly Analysed Samples Possessing six Mutations Identified in 14 of the 24 CLL Samples Previously Lacking TP53 Sequence Data. The affected nucleotide was marked by a black arrow with both mutated and wild-type (WT) sequences presented, and the affected codon is highlighted in blue. The results were analysed using ChromasPro v2.1.4.

6.2.6. POLE Sequencing

Agarose gel electrophoresis of PCR reactions using the six primer sets designed and optimised for POLE cohorts with the 49 CLL samples exhibited successful amplification with single PCR products as expected at 289bp, 273bp, 315bp, 238bp, 323bp and 279bp corresponding to the hotspot regions 1-3 and variants A-B, respectively (Figure 6-6). A band of presumed primer dimers at >100 bp were noted with different intensities in the hotspots primers 1-3.

The PCR products were sequenced with the original primers and reliable quality data matching the targeted regions was produced (Figure 6-7). The hotspot-3 electropherograms demonstrated peak tailing, possibly because of the high concentrations of DNA affecting gel mobility. The only variants detected were three known SNPs - rs5744751 , rs4883555 and rs5744798 – which affected 12%, 39% and 82% of the total samples, respectively (Figure 6-8). The first SNP is considered to be clinically benign and the third not to be significant. However, the first does fall into non-coding (ncRNA) and even synonymous SNPs are now known to cause pathology by altering gene expression (330). Moreover, the final SNP has been associated with changes in gene expression. Although the results rule out any possibility of POLE somatic mutations having any contributory role to the genomic instability of the cohort studied or association with TP53 status, the possibility that polymorphisms affect gene expression remain open. Our original hypothesis, however, was therefore not directly upheld.
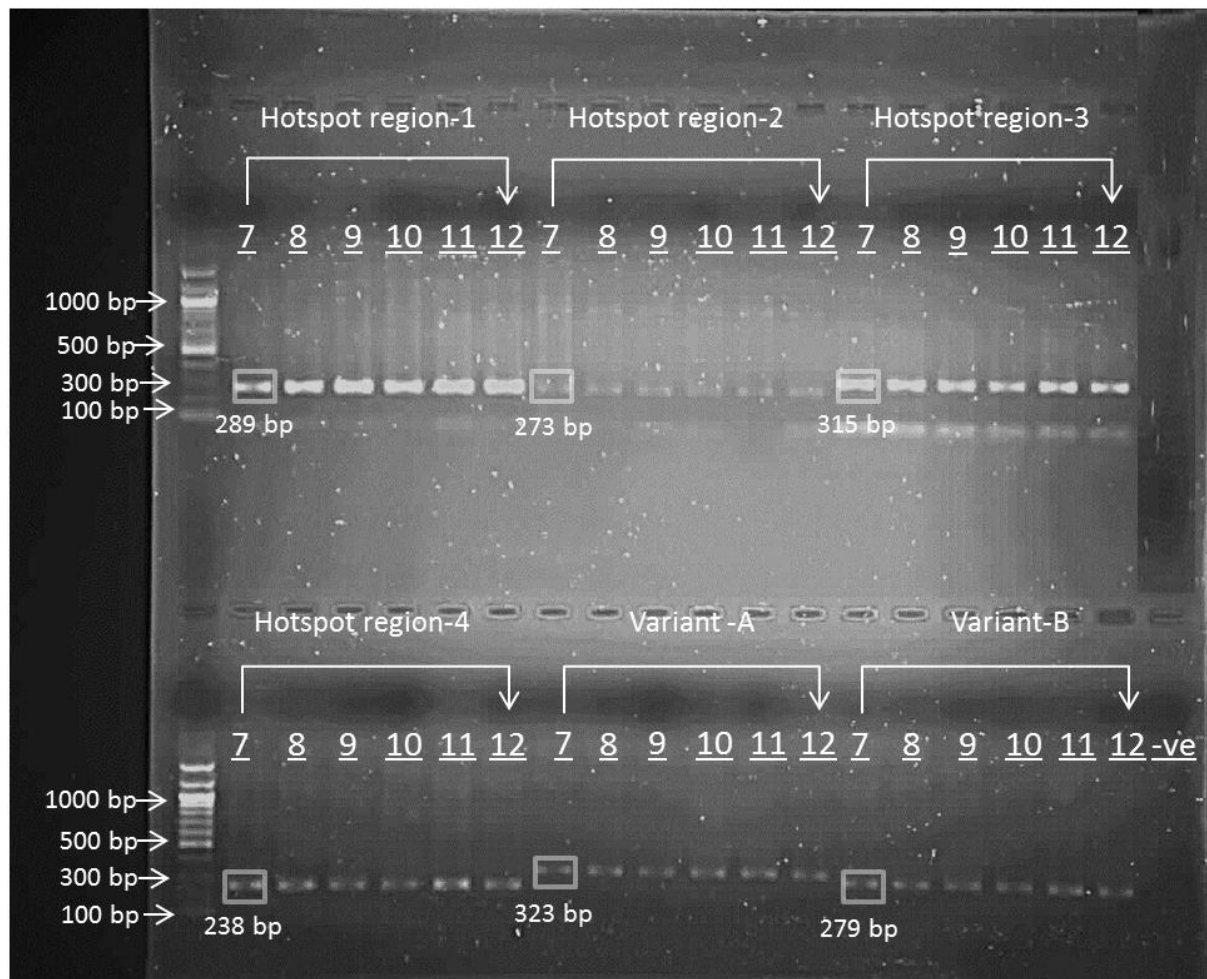
Figure 6-6: POLE Amplification Using PCR. The hotspot regions 1-4 and two other DNA fragments of POLE where the two somatic non-synonymous mutations, A and B, were identified in Chapter 3 were amplified in CLL samples 7-12 with six PCRs. The expected product size is marked below the band of the first sample for each PCR. A tube containing no gDNA was set as a negative PCR control (-ve).
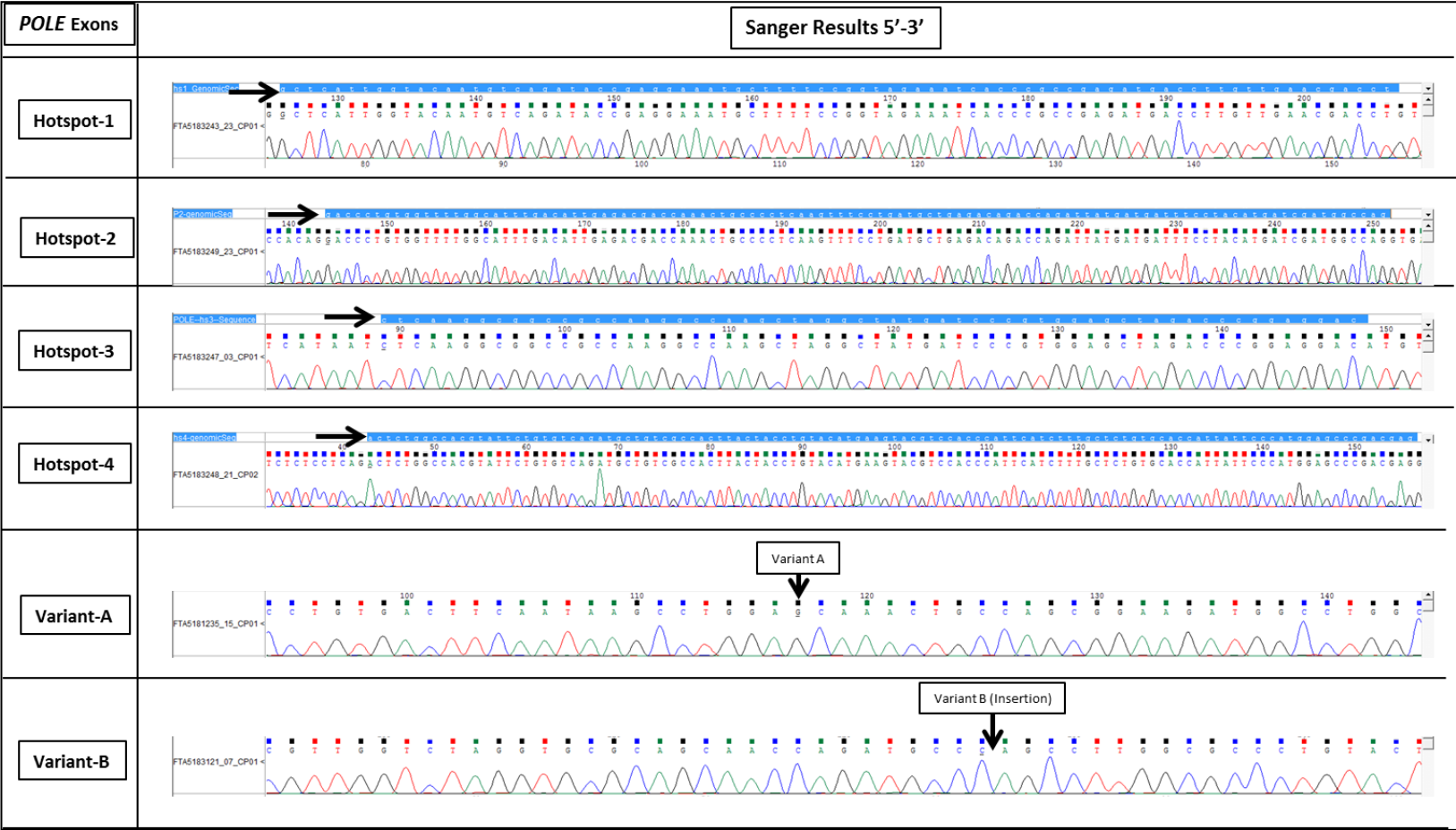
Figure 6-7: Sanger Sequencing Results for POLE Sequence. The 5'-3' sequence direction for primers hotspots 1-4 and variants A-B, respectively, are displayed. Targeted sequence hotspots 1-4 are marked by black arrows and highlighted in blue. Variant loci of A-B are also indicated by black arrows.
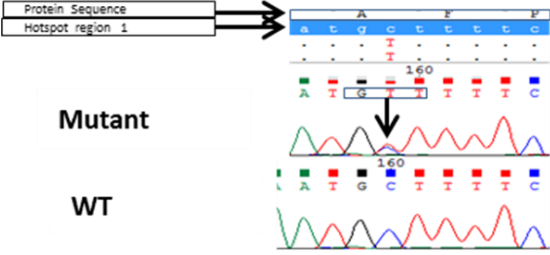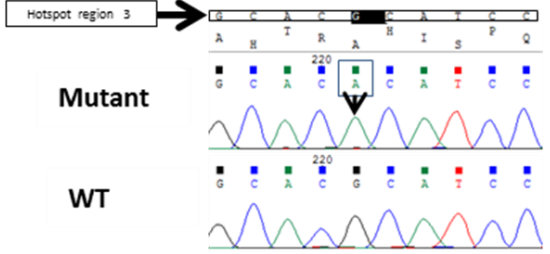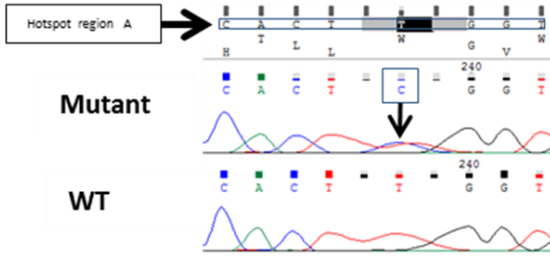
SNP #	Sanger Results 5'-3'	SNP details				
		Nucleotide change	Affected samples #	Protien change	Affected region	dbSNP
1		NC_000012.12:g.132677409G>A	4, 13, 14, 18, 34, 42	p.Ala 252 Val (GCT– GTT)	Exon 8	rs5744751
2		NC_000012.12:g.132673532C>T	2, 4, 6, 9, 11, 12, 14, 15, 19, 20, 22, 23, 24, 31, 33, 38, 42, 44, 47	Non-coding	Intergenic 13-14	rs4883555
3		NC_000012.12:g.132668563A>G	all except 1, 3, 4, 5, 6, 28, 29, 14, 25	Non-coding	Intergenic 18-19	rs5744798

Figure 6-8: POLE Variants Detected by Sanger Sequencing. Included are variant details, such as the Sanger sequence in the 5'-3' direction, nucleotide change, protein change and dbSNP reference.

6.3. Discussion and Conclusion

The investigations described in this chapter sought to detect somatic POLE variants in a

larger cohort of CLL cases and determine their association with TP53 alterations. Sanger

sequencing was utilised because it is the gold standard for validating variants with greater

than 10% AF with a 20% acceptable Phred score (197). PCR amplification and Sanger

sequencing were successful, covering 11 exonic regions of TP53 and six hotspot regions

from POLE. Sequencing was performed using both forward and reverse primers to

confirm the sequences in each strand.

Table 6-5: Combined TP53 and POLE mutational status in CLL.

TP53 mutation	Number of cases (n=59)	TP53/POLE Mutations tested by	Samples with POLE aberrations
A - Wild-type CLL	20	Sanger	0/20
B - TP53 mutant (primary cohort)	10	Sanger/PGM	2/10 (20%)
C - TP53 mutant (clinical validation cohort)	29	Sanger	0/29

Table 6-5 summarises 49 CLL screening cases and the original 10 cases screened for

alterations in their DNA maintenance genes, and all are grouped according to TP53

mutational status specifically with POLE mutations described.

Twenty (90%) of the 23 TP53 (17p) deleted cases were also mutants for TP53. This

supports the literature reporting that TP53 mutations are likely to coincide with deletions

(Figure 6-4), although both could be used individually to reflect on worse prognosis (331).

this was also supported by other prognostic markers such as unmutated IGHV status and

advance Binet stage. Furthermore, most of the somatic non-synonymous mutations were within exons 4 to 8, and would have affected the TP53 DNA-binding domain.

Concerning mutations in POLE, 20% of the primary CLL cohort were mutants. Seeing that 50% of the CLL samples that progressed acquired TP53 mutation (169), it is estimated that (0.2x0.5 = 0.01) 1% of progressed CLL would have POLE mutations. The absence of POLE mutations in the 49 samples of the clinical validation cohort may have been a chance occurrence or reflect a lack of importance for POLE in CLL (332) when compared to colon, rectal and endometrial cancers (333, 334). The three common SNPs detected in POLE are in accordance with their global AF as recorded in dbSNP and MAF (335). Therefore, the Sanger results suggest that the gene is mostly stable during the course of CLL, regardless of the TP53 mutation status.

 In conclusion, POLE seems to be a stable gene during progressive CLL, including those with inactivated p53. There is no positive relationship between POLE mutations and CLL progression or TP53 mutations/deletions in the tested cases. It could be that CNVs are drivers for CLL progression as deactivated p53 tolerates more DSBs and, hence, compromises the genomic integrity of CLL cells (127, 336, 337). Therefore, TP53 aberrations appear to be sufficient for CLL progression and altered TP53 subclones may dominate under the selective pressure of chemotherapy (338).

7.   CHAPTER 7: General Discussion

7.1.   Introduction and Rationale

CLL is a cancer of mature B lymphocytes, which is characterised by lymphocyte accumulations in blood, bone marrow and secondary lymphoid tissues. The average age at diagnosis is 70-years old, and it is usually indolent at diagnosis with a watchful waiting strategy applied. When symptoms appear, the treatment plan is begun based on RAI or BINET staging along with FISH results. Treatment outcomes are heterogeneous among patients with some having treatment resistance developed and a shorter OS. In addition, clonal instability and chromothripsis are seen in patients with TP53 aberrations, often resulting from deletion of one allele from 17p and mutation in another allele. Such LOH causes inactivation of this tumour suppressor, which directly contributes to CLL progression and subclonal evolution. This is largely based on the loss of function of the TP53 protein in regulation of apoptosis, DNA repair and cell-cycle checks (12, 339).

Evidence for abnormal DNA repair in CLL has also emerged. For example, distributions of certain single nucleotide polymorphisms in DNA repair genes, ERCC2 and XRCC1, are different between normal control and CLL cells, particularly those with the unfavourable cytogenetic aberrations (i.e., del 17p13 or del 11q22-23) (340). Moreover, increased activities in multiple DNA repair pathways have been documented in CLL. This seems to involve the upregulated DNA-PK and Rad51 activities as a response to ongoing DSB events (341, 342). Therefore, the abnormality in TP53 and other DNA repair genes may cause not only resistance to DNA-damaging chemotherapy, but also increase tolerance and burden of genomic aberrations, both of which contribute to poor clinical outcomes in CLL.

Clearly, understanding the molecular basis of the genomic instability in CLL, especially the roles of abnormalities of DNA repair, is among the most attractive research activities. However, investigation in this field has long been hampered, largely because of the resting status of circulating CLL cells, which made it particularly difficult in genetic studies using conventional methods. Specifically, it is unclear whether there is any underlying damage to the DNA repair genes, the nature of which could help explain unique clinical phenotypes. This may now be addressed through the advent of recently developed NGS, a powerful technique that can identify any form of genomic aberrations in target genes or in the whole genome of resting cells. This has therefore paved the way for investigating mechanisms for genomic instability in CLL (217, 220, 343-345).

7.2. Thesis Aim and Objectives

This thesis aims to address whether somatic lesions in DNA maintenance genes are involved in CLL clonal instability in a cohort of inactivated TP53 patients. For this reason, the NGS approach was established to enrich exon regions of 194 DNA maintenance genes (Chapters 3 and 4). Furthermore, genomic structural integrity was also determined via WGS (Chapter 5). Missense lesions occurring recurrently in POLE encouraged further testing of the gene in a larger cohort of patients (Chapter 6).

7.3. Summary of Results

In Chapter 3, the DNA repair panel of 194 genes was successfully established. The probed design utilises the maximum capacity of SureSelect in-solution hybridisation (346). With the SureDesign tool (347), the panel was set to increase enriching targeted regions by

choosing maximum boosting and minimum CG masker options. The total enriched regions

were 700 Kb, including 70% of targeted regions.

Regarding PGM coverage analysis, the results show that about 300x was the actual

coverage depth, which was 55% less than the expected results. This is largely off-target

from low complexity regions, which were elevated as a result of stringency reduction of

the designed probes. Lowering the CG masker option caused high off-target enrichment

to similar genomic regions which compromised sequencing outcomes. It is therefore

important to balance the probe design setting and try to avoid or reduce low complexity

regions to harvest the maximum benefits of sequencing capacity.

The Chapter 4 results confirmed the TP53 mutations and a recurrently altered DNA repair

gene, POLE. After optimising the NGS protocol, 10 CLL samples were in-house sequenced.

The results show 90% sensitivity of the variant caller and 50% specificity. The lower

overall specificity reflects the high impact of PGM HP errors, which could be minimised by

increasing coverage reads followed by rising stringency of the variant caller (285, 348).

After that, somatic variants were further validated by comparing findings to the germline

cells of patients using Sanger sequencing and WGS. The results confirmed that POLE had

somatic mutations in two of the 10 (20%) samples. These mutations affected POLE at

P.Ala 661 Thr (CLL10) and P.Ser 1525 Fs (CLL07) with a medium-high missense score of 0.6

sift and 0.05 polyphen. Generally, inactivated POLE could disrupt different DNA repair

pathways, such as BER, DNA damage bypass and DNA DSB repair via HR. The TP53 protein

is also interconnected to the DSB response by ATM activation (349, 350). Further

investigation of the genomic integrity was carried out with WGS and the same cohort of samples (Chapter 5).

The WGS was screened to investigate the structural integrity of the CLL genome and whether possible chromosomal instability was involved in the same cohort of patients. The coverage analysis demonstrated that 30x is the coverage depth with ≥20 AF being the detection limit. The results show acceptable uniformity despite the challenge of low complexity regions and misalignment, which reduced the coverage. CNA analysis confirmed 17p deletion of the samples. In addition, there are 1-3x single copy deletions between samples. These lesions were not common, but they do compromise the genomic integrity of CLL cells and could be a contributor to progression and poor outcomes. This would need to be determined in a larger series where it would also be possible to establish their possible association with TP53 pathway defects.

Common translocations are a feature in certain haematological malignancies, such as inter-chromosomal translocation causing oncogene ABL-BCR in CML. However, it was difficult to validate them in this study, especially after the bias issue of the BreakDancer algorithm in identifying translocations. This suggests future work in terms of data validation either by germline control or with different cytogenetic approaches.

Employing WGS data, there were two major pathways affected by mutations - immune response and signal transduction pathways. It was also noted that DNA repair and the cell cycle were the least affected by mutations other than TP53, which indicated the independent role of TP53 as a DNA repair gene in CLL pathogenesis.

CHAPTER 7: General Discussion

WGS-validated POLE mutations were discerned with deep sequencing and Sanger sequencing. Apart from TP53, there were no other LOH events within the mutated region. This may be reflective of genome stability despite the major molecular defects caused by TP53 loss (351). As a result, POLE was assigned as a candidate gene for conducting clinical validation in a larger cohort of CLL patients (Chapter 5).

In Chapter 6, using Sanger sequencing, CLL samples from an expanded cohort of 59 patients were screened to investigate the distribution of POLE mutations and its correlation with TP53 mutations. For this purpose, patients with different disease statuses were selected. There were six POLE regions examined; two potential loci and four hotspot regions which were recorded from other cancers (329). TP53 mutational status was examined in order to classify samples. It was important to examine recurrent mutation, if any, and whether it is TP53 dependent. Sanger sequencing is the gold standard - it can detect>10% AF, which could be considered a limiting factor of the approach in addition to background noise. In a combination of data of the 59 patients studied in Chapters 3, 4 and 5, the frequency of TP53 mutation was 52.54% (31/59), but that of POLE mutations was only 3.39% (2/59).

Although there is no statistical difference in this frequency between groups with and without TP53 mutations, as analysed with the data from the entire cohort (n=59) or only the validation cohort (n=49), both POLE-mutant cases were in the TP53-mutation group, rendering the mutation incidence at 6.45% and 9.52% in the entire and validation cohorts, respectively.

These findings therefore provide evidence that mutations in DNA repair genes indeed exist in CLL, although not as common as in other human cancers. A large study may be required to further explore their relationship with TP53 mutations in this disease.

7.4.   Strength and Weaknesses of the Whole Thesis

Strengths:

NGS provides the advantage of big data to generate and examine a research hypothesis in a comprehensive manner (352). A panel of 194 DNA maintenance genes was selected to examine their protein-coding regions in CLL (61, 276, 353, 354). WGS was also leveraged to provide a larger picture of the genome and whether there were structural abnormalities related to genomic instability.

One of the advantages was employing the CLL 206 and 210 trial samples for in-depth correlation of NGS to patient's response and treatment history (46, 47, 249). The primary cohort had 10 patients with inactivated TP53. Trial data stated that patients had multiple treatments with resistance to DNA-damaging agents. Therefore, it was suitable to examine their DNA repair genes.

 Deep sequencing utilising the in-house Ion Torrent PGM facility providing an in-depth understanding of pre- and post-sequencing processes. Sample processing was optimised for DNA quality, DNA shearing and PCR amplification. Post-sequencing involves coverage analysis, optimising variant detection and genotyping assessment.

After deep sequencing of cancer samples, germline validation was applied via Sanger sequencing. This was an important step prior to clinical validation. Novel variants of 40-

60% are likely SNPs and therefore, it was very significant to understand the sporadic origin of the variants.

Clinical validation in a large and separate cohort of patients was proposed to examine whether or not POLE mutations are dependent on TP53 mutations and are associated with disease status. This was important at this level to clearly define any affected DNA repair genes.

In addition to CNAs, WGS was utilised to interpret WES. It was valuable to look at other affected genes and correlate their protein pathways and genomic instability. Many non-coding RNAs were also examined, providing more depth to the study by examining post-translation regulators.

This study defines WGS data free of germline control for detecting CNAs and somatic mutations. CNAs were detected using ControlFREEC by comparing sample data to the hg38 reference sequence (263). Regarding somatic variants, multi-level filtration was employed to reduce alignment bias and false findings. Somatic variants were defined as those novel variants that resided within highly complex regions and single-hit variants among samples (355).

Weaknesses:

 Like any study, a number of weaknesses are observed; using NGS requires complex data analysis to be employed. Deep sequencing took place via Ion Torrent PGM with Ion suite for data alignment, providing clear data, coverage analysis, variants annotation and genotyping. However, with WGS, owing to data scale and technical facility, it was

necessary for bioinformatics support to primarily analyse data, including clear data

processing and variant annotations. This adds more technical and financial challenges

(356).

The primary data of NGS was lacking a germline control to exclude SNPs and other

germline aberrations. In addition, this step could be used in identifying false findings by

comparing cancer/germline NGS datasets - this would save time and money by increasing

specificity of the approach with lesser use of other approaches for validation. In addition,

many of the WGS somatic aberrations could be much easier to validate utilising germline

controls - this is applied to CNAs, translocations and SNVs.

Sequential samples were not used in this study. It would be important for observing CLL

evolution from the genomic instability prospective. Many studies suggested CLL evolution

by introducing more mutations in different genes which were useful for screening (357-

359).

Regarding the use of a gene panel, a large list of genes increases the number of probes

required and total enriched size, reducing the coverage depth and sensitivity to a

minimum of 8% AF (total coverage: 250 reads; minimum allele coverage: 20 reads). It also

increases the opportunity for off-target enrichment. Likewise, WGS offers a low-pass

sequence of a genome, resulting in reducing the sensitivity of variant detection to 20% AF

(coverage depth: 30x; minimum allele coverage: six reads).

PGM sequencing produces a modicum of inherited bias when interpreting HP sequences

and indels. Variant detection usually discerns HP errors as novel indels. As a result,

specificity of the variants was reduced and many indels turned out to be FPs, which also was a waste of time in terms of validating those variants (283).

7.5.   Conflicts Stated and Explained; Speculations About What The Results Might Mean

DNA repair genes are defective in many cancers. Endometrium cancers have defects in TP53, POLE and other DNA repair genes (324, 335, 360). In CLL, other DNA repair genes have little evidence of being involved in such a mechanism despite TP53 defects. In addition, the data show no common CINs observed other than 17p deleted. This may reflect the tolerated status of the CLL genome, at least for this cohort of patients.

POLE was found mutated in 20% of inactivated TP53 CLL. On the contrary, the inactivated POLE gene was confirmed with a robust prognosis in endometrial cancer (324).

In addition, the chromosome 13q14 deletion was considered for its favourable prognosis. However, two cases (20%) from the worst prognosis cohort acquired this deletion.

Many chromosomal instabilities were observed with WGS. The majority of them are single-copy deletions - these were considered LOH where heterogeneous chromosomes had been compromised by dominating SNPs and missense variants in protein-coding regions. Regarding translocations and CLL, no common translocations were observed compared to other malignancies, such as CML, a common translocation BCR-ABL fused oncogene within the Philadelphia chromosome. The frequency of large CNAs varied between samples most affected having six affected chromosomes and the least was two affected chromosomes in Liv_02, Liv_03, Liv_06, Liv_07, Liv_08 and Liv_10. Chromosome 5 was the second most affected chromosome. Polymorphic CNAs are extensive

throughout the genome, but paradoxically, the pathological consequences of CNAs are well-known, with Down's Syndrome being the classic example. It would therefore be interesting to investigate the prevalence and nature of CNAs in a larger cohort where the association with TP53 defects and outcomes could be assessed.

7.6. How to Advance The Knowledge Base Format Relating to the Existing Literature

This study supports the notion that oncogene activation is the initiation event in CLL, followed by TP53 inactivation and chromosomal instability. Furthermore, WGS suggests a stable CLL genome of the patients despite LOH events that may accumulate mutations in the CLL genome at later stages (117, 127, 361, 362). However, there is less evidence that other DNA repair genes are recurrently mutated compared to colorectal cancer. In addition, CLL seems to be more defective with respect to genes that are involved in immune response and signal transduction pathways. These pathways could be linked to B-cell development and DSB events.
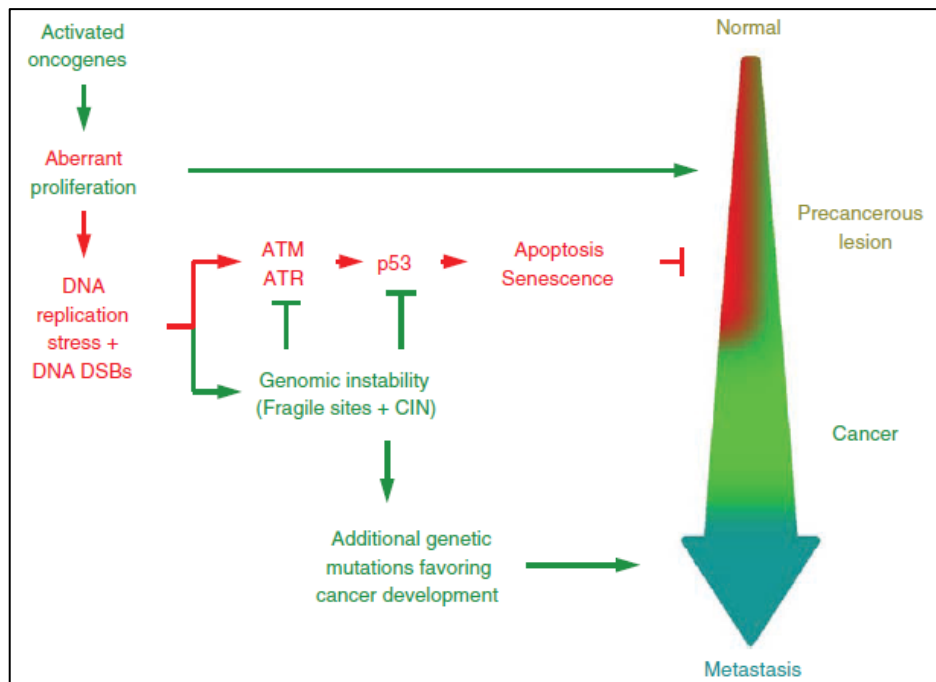
Figure 7- 1: Oncogene-Induced Model for Cancer Progression. Replication stress is caused by oncogenes followed by genomic instability and cancer development. This figure is adopted with permission from Halazonetis TD, Gorgoulis VG, Bartek J. An oncogene-induced DNA damage model for cancer development. Science. 2008;319(5868):1352-5.

7.7.   Future work

The clinical validation for the POLE gene could be expanded using more samples. In addition, ultra-deep sequencing could be implemented to detect tiny clones of mutations. Sequential samples could also be tested to look at the progress of clonal evolution within the disease course.

WGS is rich with genome data, and so it can be used retrospectively. This includes mutations which significantly affect protein pathways, such as immune response and signal transduction pathways. In parallel, low-pass sequencing employing germline samples should be applied to confirm SNPs and germline indels and CNVs.

Regarding probes enrichment for targeted-NGS, deep sequencing should be carefully established, especially for designing probes. Probes should be specifically designed and

tested to provide efficient enrichment and expected coverage depths. This could be carried out by targeting specific regions, avoiding CG-rich regions and maximising read length. Furthermore, in-house analysis pipelines should be established to reduce bioinformatics cost and providing flexibility for optimising and setting sequencing analysis.

7.8. Conclusions

Apart from TP53, this study shows there are no common mutations in DNA repair genes, thereby requiring further validation using a larger CLL cohort. It is also confirmed that the independent role of inactivated TP53 is as a molecular culprit for cancer survival and chemo resistance (363, 364). NGS proved to be advantageous as a comprehensive tool for detecting different molecular alterations. In addition, the 1000 Genome project would be a great asset for expanding sample numbers and increasing the statistical power of the discovery cohort with paired-sample WGS data from screened CLL samples.

References

1.   Fabbri G, Dalla-Favera R. The molecular pathogenesis of chronic lymphocytic leukaemia. Nat Rev Cancer. 2016;16(3):145-62.

2.   Demir HA, Bayhan T, Uner A, Kurtulan O, Karakus E, Emir S, et al. Chronic Lymphocytic Leukemia in a Child: A Challenging Diagnosis in Pediatric Oncology Practice. Pediatric Blood & Cancer. 2014;61(5):933-5.

3.   Goldin LR, Bjorkholm M, Kristinsson SY, Turesson I, Landgren O. Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. Haematologica-the Hematology Journal. 2009;94(5):647-53.

4.   Wierda WG, O'Brien S, Wang XM, Faderl S, Ferrajoli A, Do KA, et al. Prognostic nomogram and index for overall survival in previously untreated patients with chronic lymphocytic leukemia. Blood. 2007;109(11):4679-85.

5.   Kipps TJ, Stevenson FK, Wu CJ, Croce CM, Packham G, Wierda WG, et al. Chronic lymphocytic leukaemia. Nat Rev Dis Primers. 2017;3:1-19.

6.   Fugmann SD, Lee AI, Shockett PE, Villey IJ, Schatz DG. The rag proteins and V(D)J recombination: Complexes, ends, and transposition. Annual Review of Immunology. 2000;18:495-527.

7.   Tonegawa S. SOMATIC GENERATION OF ANTIBODY DIVERSITY. Nature. 1983;302(5909):575-81.

8.   Chaudhuri J, Tian M, Khuong C, Chua K, Pinaud E, Alt FW. Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature. 2003;422(6933):726-30.

9.   Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 2000;102(5):553-63.

10.   Klein U, Dalla-Favera R. Germinal centres: role in B-cell physiology and malignancy. Nature Reviews Immunology. 2008;8(1):22-33.

11.   Allen CDC, Okada T, Tang HL, Cyster JG. Imaging of germinal center selection events during affinity maturation. Science. 2007;315(5811):528-31.

12.   Zenz T, Mertens D, Kuppers R, Dohner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. Nature Reviews Cancer. 2010;10(1):37-50.

13.   Myhrinder AL, Hellqvist E, Sidorova E, Soderberg A, Baxendale H, Dahle C, et al. A new perspective: molecular motifs on oxidized LDL, apoptotic cells, and bacteria are targets for chronic lymphocytic leukemia antibodies. Blood. 2008;111(7):3838-48.

14.   Catera R, Silverman GJ, Hatzi K, Seiler T, Didier S, Zhang L, et al. Chronic Lymphocytic Leukemia Cells Recognize Conserved Epitopes Associated with Apoptosis and Oxidation. Molecular Medicine. 2008;14(11-12):665-74.

References

15.   Chu CC, Catera R, Hatzi K, Yan XJ, Zhang L, Wang XB, et al. Chronic lymphocytic leukemia antibodies with a common stereotypic rearrangement recognize nonmuscle myosin heavy chain IIA. Blood. 2008;112(13):5122-9.

16.   Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Dohner H, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. Blood. 2008;111(12):5446-56.

17.   Hallek M. Chronic lymphocytic leukemia: 2015 Update on diagnosis, risk stratification, and treatment. American Journal of Hematology. 2015;90(5):446-60.

18.   Ginaldi L, De Martinis M, Matutes E, Farahat N, Morilla R, Catovsky D. Levels of expression of CD19 and CD20 in chronic B cell leukaemias. Journal of Clinical Pathology. 1998;51(5):364-9.

19.   Gribben JG. How I treat CLL up front. Blood. 2009;115(2):187-97.

20.   Rozman C, Montserrat E. Chronic Lymphocytic Leukemia. New England Journal of Medicine. 1995;333(16):1052-7.

21.   Rossi D, Gaidano G. Richter syndrome: pathogenesis and management. Seminars in Oncology. 2016;43(2):311-9.

22.   Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN, Pasternack BS. CLINICAL STAGING OF CHRONIC LYMPHOCYTIC LEUKEMIA. Blood. 1975;46(2):219-34.

23.   Binet JL, Auquier A, Dighiero G, Chastang C, Piguet H, Goasguen J, et al. A NEW PROGNOSTIC CLASSIFICATION OF CHRONIC LYMPHOCYTIC-LEUKEMIA DERIVED FROM A MULTIVARIATE SURVIVAL ANALYSIS. Cancer. 1981;48(1):198-206.

24.   Van Bockstaele F, Verhasselt B, Philippé J. Prognostic markers in chronic lymphocytic leukemia: A comprehensive review. Blood Reviews. 2009;23(1):25-47.

25.   Aydin S, Rossi D, Bergui L, D'Arena G, Ferrero E, Bonello L, et al. CD38 gene polymorphism and chronic lymphocytic leukemia: a role in transformation to Richter syndrome? Blood. 2008;111(12):5646-53.

26.   Desai S, Pinilla-Ibarz J. Front-Line Therapy for Chronic Lymphocytic Leukemia. Cancer Control. 2012;19(1):26-36.

27.   Dreger P, Schetelig J, Andersen N, Corradini P, van Gelder M, Gribben J, et al. Managing high-risk CLL during transition to a new treatment era: stem cell transplantation or novel agents? Blood. 2014;124(26):3841-9.

28.   Panasci L, Paiement JP, Christodouplopoulos G, Belenkov A, Malapetsa A, Aloyz R. Chlorambucil drug resistance in chronic lymphocytic leukemia: The emerging role of DNA repair. Clinical Cancer Research. 2001;7(3):454-61.

29.   Goede V, Fischer K, Busch R, Engelke A, Eichhorst B, Wendtner CM, et al. Obinutuzumab plus Chlorambucil in Patients with CLL and Coexisting Conditions. New England Journal of Medicine. 2014;370(12):1101-10.

References

30.   Woyach JA, Johnson AJ. Targeted therapies in CLL: mechanisms of resistance and strategies for management. Blood. 2015;126(4):471-7.

31.   Johnson GG, Sherrington PD, Carter A, Lin K, Liloglou T, Field JK, et al. A Novel Type of p53 Pathway Dysfunction in Chronic Lymphocytic Leukemia Resulting from Two Interacting Single Nucleotide Polymorphisms within the p21 Gene. Cancer Research. 2009;69(12):5210-7.

32.   Pettitt AR. Mechanism of action of purine analogues in chronic lymphocytic leukaemia. British Journal of Haematology. 2003;121(5):692-702.

33.   Pettitt AR. p53 dysfunction in B-cell chronic lymphocytic leukemia: inactivation of ATM as an alternative to TP53 mutation. Blood. 2001;98(3):814-22.

34.   Catovsky D, Richards S, Matutes E, Oscier D, Dyer MJS, Bezares RF, et al. Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomised controlled trial. The Lancet. 2007;370(9583):230-9.

35.   Flinn IW, Neuberg DS, Grever MR, Dewald GW, Bennett JM, Paietta EM, et al. Phase III Trial of Fludarabine Plus Cyclophosphamide Compared With Fludarabine for Patients With Previously Untreated Chronic Lymphocytic Leukemia: US Intergroup Trial E2997. Journal of Clinical Oncology. 2007;25(7):793-8.

36.   Bosch F, Ferrer A, Lopez-Guillermo A, Gine E, Bellosillo B, Villamor N, et al. Fludarabine, cyclophosphamide and mitoxantrone in the treatment of resistant or relapsed chronic lymphocytic leukaemia. Br J Haematol. 2002;119(4):976-84.

37.   Laurenti L, Innocenti I, Autore F, Sica S, Efremov D. New developments in the management of chronic lymphocytic leukemia: role of ofatumumab. OncoTargets and Therapy. 2016:421.

38.   van Oers MHJ, Kuliczkowski K, Smolej L, Petrini M, Offner F, Grosicki S, et al. Ofatumumab maintenance versus observation in relapsed chronic lymphocytic leukaemia (PROLONG): an open-label, multicentre, randomised phase 3 study. Lancet Oncology. 2015;16(13):1370-9.

39.   Owen CJ, Stewart DA. Obinutuzumab for the treatment of patients with previously untreated chronic lymphocytic leukemia: overview and perspective. Therapeutic Advances in Hematology. 2015;6(4):161-70.

40.   Damm F, Mylonas E, Cosson A, Yoshida K, Della Valle V, Mouly E, et al. Acquired Initiating Mutations in Early Hematopoietic Cells of CLL Patients. Cancer Discovery. 2014;4(9):1088-101.

41.   Woyach JA, Furman RR, Liu T-M, Ozer HG, Zapatka M, Ruppert AS, et al. Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib. New England Journal of Medicine. 2014;370(24):2286-94.

42.   Maddocks KJ, Ruppert AS, Lozanski G, Heerema NA, Zhao W, Abruzzo L, et al. Etiology of Ibrutinib Therapy Discontinuation and Outcomes in Patients With Chronic Lymphocytic Leukemia. JAMA oncology. 2015;1(1):80-7.

43.   James DF, Werner L, Brown JR, Wierda WG, Barrientos JC, Castro JE, et al. Lenalidomide and rituximab for the initial treatment of patients with chronic lymphocytic leukemia: a

References

multicenter clinical-translational study from the chronic lymphocytic leukemia research consortium. J Clin Oncol. 2014;32(19):2067-73.

44.   Furman RR, Sharman JP, Coutre SE, Cheson BD, Pagel JM, Hillmen P, et al. Idelalisib and Rituximab in Relapsed Chronic Lymphocytic Leukemia. New England Journal of Medicine. 2014;370(11):997-1007.

45.   Roberts AW, Davids MS, Pagel JM, Kahl BS, Puvvada SD, Gerecitano JF, et al. Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. New England Journal of Medicine. 2016;374(4):311-22.

46.   Pettitt AR, Jackson R, Carruthers S, Dodd J, Dodd S, Oates M, et al. Alemtuzumab in Combination With Methylprednisolone Is a Highly Effective Induction Regimen for Patients With Chronic Lymphocytic Leukemia and Deletion of TP53: Final Results of the National Cancer Research Institute CLL206 Trial. Journal of Clinical Oncology. 2012;30(14):1647-55.

47.   Pettitt R, Matutes E, Dearden C, Oscier D, Carruthers S, Dodd J, et al. Results of the Phase Ii Ncri Cll206 Trial of Alemtuzumab in Combination with High-Dose Methylprednisolone for High-Risk (17p-) Cll. Haematologica-the Hematology Journal. 2009;94:138-9.

48.   Pettitt AR, Polydoros F, Dodd J, Oates M, Lin K, Kalakonda N, et al. Final Results of the Ncri Cll210 Trial of Alemtuzumab, Dexamethasone and Lenalidomide in Patients with High-Risk Cll (Original Protocol). Haematologica. 2016;101:726.

49.   Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.

50.   Bartkova J, Horejsi Z, Koed K, Kramer A, Tort F, Zieger K, et al. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. Nature. 2005;434(7035):864-70.

51.   Gorgoulis VG, Vassiliou LVF, Karakaidos P, Zacharatos P, Kotsinas A, Liloglou T, et al. Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. Nature. 2005;434(7035):907-13.

52.   Lindahl T. INSTABILITY AND DECAY OF THE PRIMARY STRUCTURE OF DNA. Nature. 1993;362(6422):709-15.

53.   van Loon B, Markkanen E, Huebscher U. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. DNA Repair. 2010;9(6):604-16.

54.   Schreiber V, Dantzer F, Ame J-C, de Murcia G. Poly(ADP-ribose): novel functions for an old molecule. Nature Reviews Molecular Cell Biology. 2006;7(7):517-28.

55.   El-Khamisy SF, Masutani M, Suzuki H, Caldecott KW. A requirement for PARP-1 for the assembly or stability of XRCC1 nuclear foci at sites of oxidative DNA damage. Nucleic Acids Research. 2003;31(19):5526-33.

56.   Plo I, Liao ZY, Barcelo JM, Kohlhagen G, Caldecott KW, Weinfeld M, et al. Association of XRCC1 and tyrosyl DNA phosphodiesterase (Tdp1) for the repair of topoisomerase I-mediated DNA lesions. DNA Repair. 2003;2(10):1087-100.

References

57.   Larsen E, Meza TJ, Kleppa L, Klungland A. Organ and cell specificity of base excision repair mutants in mice. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis. 2007;614(1-2):56-68.

58.   Sweasy JB, Lang TM, DiMaio D. Is base excision repair a tumor suppressor mechanism? Cell Cycle. 2006;5(3):250-9.

59.   Curtin NJ. DNA repair dysregulation from cancer driver to therapeutic target. Nature Reviews Cancer. 2012;12(12):801-17.

60.   Kipps TJ, Stevenson FK, Wu CJ, Croce CM, Packham G, Wierda WG, et al. Chronic lymphocytic leukaemia. Nat Rev Dis Primers. 2017;3:16096.

61.   Lange SS, Takata K-i, Wood RD. DNA polymerases and cancer. Nature Reviews Cancer. 2011;11(2):96-110.

62.   Andressoo JO, Hoeijmakers JHJ, de Waard H. Nucleotide excision repair and its connection with cancer and ageing. In: Nigg EA, editor. Genome Instability in Cancer Development. Advances in Experimental Medicine and Biology. 5702005. p. 45-81.

63.   Pluciennik A, Dzantiev L, Iyer RR, Constantin N, Kadyrov FA, Modrich P. PCNA function in the activation and strand direction of MutL alpha endonuclease in mismatch repair. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(37):16066-71.

64.   Li GM. Mechanisms and functions of DNA mismatch repair. Cell Research. 2008;18(1):85-98.

65.   Vilenchik MM, Knudson AG. Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(22):12871-6.

66.   Shrivastav M, De Haro LP, Nickoloff JA. Regulation of DNA double-strand break repair pathway choice. Cell Research. 2008;18(1):134-47.

67.   Mahaney BL, Meek K, Lees-Miller SP. Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. Biochemical Journal. 2009;417:639-50.

68.   Beucher A, Birraux J, Tchouandong L, Barton O, Shibata A, Conrad S, et al. ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2. Embo Journal. 2009;28(21):3413-27.

69.   Karran P, Bignami M. DNA-DAMAGE TOLERANCE, MISMATCH REPAIR AND GENOME INSTABILITY. Bioessays. 1994;16(11):833-9.

70.   Almeida KH, Sobol RW. A unified view of base excision repair: Lesion-dependent protein complexes regulated by post-translational modification. DNA Repair. 2007;6(6):695-711.

71.   Kurimasa A, Kumano S, Boubnov NV, Story MD, Tung CS, Peterson SR, et al. Requirement for the kinase activity of human DNA-dependent protein kinase catalytic subunit in DNA strand break rejoining. Molecular and Cellular Biology. 1999;19(5):3877-84.

References

72.   Zhong Q, Chen CF, Li S, Chen YM, Wang CC, Xiao J, et al. Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. Science. 1999;285(5428):747-50.

73.   Haince J-F, McDonald D, Rodrigue A, Dery U, Masson J-Y, Hendzel MJ, et al. PARP1-dependent kinetics of recruitment of MRE11 and NBS1 proteins to multiple DNA damage sites. Journal of Biological Chemistry. 2008;283(2):1197-208.

74.   Stiff T, O'Driscoll M, Rief N, Iwabuchi K, Lobrich M, Jeggo PA. ATM and DNA-PK function redundantly to phosphorylate H2AX after exposure to ionizing radiation. Cancer Research. 2004;64(7):2390-6.

75.   Jensen RB, Carreira A, Kowalczykowski SC. Purified human BRCA2 stimulates RAD51-mediated recombination. Nature. 2010;467(7316):678-83.

76.   Wang HY, Wang HC, Powell SN, Iliakis G, Wang Y. ATR affecting cell radiosensitivity is dependent on homologous recombination repair but independent of nonhomologous end joining. Cancer Research. 2004;64(19):7139-43.

77.   Saleh-Gohari N, Bryant HE, Schultz N, Parker KA, Cassel TN, Helleday T. Spontaneous homologous recombination is induced by collapsed replication forks that are caused by endogenous DNA single-strand breaks. Molecular and Cellular Biology. 2005;25(16):7158-69.

78.   Deans AJ, West SC. DNA interstrand crosslink repair and cancer. Nature Reviews Cancer. 2011;11(7):467-80.

79.   Derheimer FA, Kastan MB. Multiple roles of ATM in monitoring and maintaining DNA integrity. Febs Letters. 2010;584(17):3675-81.

80.   Kastan MB, Bartek J. Cell-cycle checkpoints and cancer. Nature. 2004;432(7015):316-23.

81.   Shi H, Le Calvez F, Olivier M, Hainaut P. Patterns of TP53 mutations in human cancer: Interplay between mutagenesis, DNA repair and selection. 25 Years of P53 Research. 2005:293-319.

82.   Shi D, Gu W. Dual Roles of MDM2 in the Regulation of p53: Ubiquitination Dependent and Ubiquitination Independent Mechanisms of MDM2 Repression of p53 Activity. Genes Cancer. 2012;3(3-4):240-8.

83.   Lee JH, Paull TT. Activation and regulation of ATM kinase activity in response to DNA double-strand breaks. Oncogene. 2007;26(56):7741-8.

84.   Bouwman P, Aly A, Escandell JM, Pieterse M, Bartkova J, van der Gulden H, et al. 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. Nature Structural & Molecular Biology. 2010;17(6):688-95.

85.   Willis N, Rhind N. Regulation of DNA replication by the S-phase DNA damage checkpoint. Cell Division. 2009;4.

86.   Delacroix S, Wagner JM, Kobayashi M, Yamamoto K, Karnitz LM. The Rad9-Hus1-Rad1 (9-1-1) clamp activates checkpoint signaling via TopBP1. Genes & Development. 2007;21(12):1472-7.

References

87.	Deckbar D, Jeggo PA, Lobrich M. Understanding the limitations of radiation-induced cell cycle checkpoints. Crit Rev Biochem Mol Biol. 2011;46(4):271-83.

88.	Perry JA, Kornbluth S. Cdc25 and Wee1: analogous opposites? Cell Div. 2007;2:12.

89.	Elia AE, Rellos P, Haire LF, Chao JW, Ivins FJ, Hoepker K, et al. The molecular basis for phosphodependent substrate targeting and regulation of Plks by the Polo-box domain. Cell. 2003;115(1):83-95.

90.	Nakanishi M, Shimada M, Niida H. Genetic instability in cancer cells by impaired cell cycle checkpoints. Cancer Science. 2006;97(10):984-9.

91.	Blais A, van Oevelen CJ, Margueron R, Acosta-Alvear D, Dynlacht BD. Retinoblastoma tumor suppressor protein–dependent methylation of histone H3 lysine 27 is associated with irreversible cell cycle exit. J Cell Biol. 2007;179(7):1399-412.

92.	Rodier F, Campisi J, Bhaumik D. Two faces of p53: aging and tumor suppression. Nucleic Acids Res. 2007;35(22):7475-84.

93.	Voskoboinik I, Whisstock JC, Trapani JA. Perforin and granzymes: function, dysfunction and human pathology. Nature Reviews Immunology. 2015;15(6).

94.	Maréchal A, Zou L. DNA Damage Sensing by the ATM and ATR Kinases. Cold Spring Harb Perspect Biol. 2013;5(9).

95.	Boulares AH, Yakovlev AG, Ivanova V, Stoica BA, Wang G, Iyer S, et al. Role of Poly(ADP-ribose) Polymerase (PARP) Cleavage in Apoptosis. 1999.

96.	Alano CC, Garnier P, Ying W, Higashi Y, Kauppinen TM, Swanson RA. NAD+ depletion is necessary and sufficient for PARP-1 – mediated neuronal death. J Neurosci. 2010;30(8):2967-78.

97.	Benchimol S, Lamb P, Crawford LV, Sheer D, Shows TB, Bruns GAP, et al. TRANSFORMATION ASSOCIATED P53 PROTEIN IS ENCODED BY A GENE ON HUMAN CHROMOSOME-17. Somatic Cell and Molecular Genetics. 1985;11(5):505-9.

98.	Isobe M, Emanuel BS, Givol D, Oren M, Croce CM. LOCALIZATION OF GENE FOR HUMAN P53 TUMOR-ANTIGEN TO BAND 17P13. Nature. 1986;320(6057):84-5.

99.	McBride OW, Merry D, Givol D. THE GENE FOR HUMAN P53 CELLULAR TUMOR-ANTIGEN IS LOCATED ON CHROMOSOME-17 SHORT ARM (17P13). Proceedings of the National Academy of Sciences of the United States of America. 1986;83(1):130-4.

100.	Lin JY, Chen JD, Elenbaas B, Levine AJ. SEVERAL HYDROPHOBIC AMINO-ACIDS IN THE P53 AMINO-TERMINAL DOMAIN ARE REQUIRED FOR TRANSCRIPTIONAL ACTIVATION, BINDING TO MDM-2 AND THE ADENOVIRUS-5 E1B 55-KD PROTEIN. Genes & Development. 1994;8(10):1235-46.

101.	Zhu JH, Zhou WJ, Jiang JY, Chen XB. Identification of a novel p53 functional domain that is necessary for mediating apoptosis. Journal of Biological Chemistry. 1998;273(21):13030-6.

References

102.	Venot C, Maratrat M, Sierra V, Conseiller E, Debussche L. Definition of a p53 transactivation function-deficient mutant and characterization of two independent p53 transactivation subdomains. Oncogene. 1999;18(14):2405-10.

103.	Walker KK, Levine AJ. Identification of a novel p53 functional domain that is necessary for efficient growth suppression. Proceedings of the National Academy of Sciences of the United States of America. 1996;93(26):15335-40.

104.	Ko LJ, Prives C. p53: Puzzle and paradigm. Genes & Development. 1996;10(9):1054-72.

105.	Shaulsky G, Goldfinger N, Tosky MS, Levine AJ, Rotter V. NUCLEAR-LOCALIZATION IS ESSENTIAL FOR THE ACTIVITY OF P53 PROTEIN. Oncogene. 1991;6(11):2055-65.

106.	Sturzbecher HW, Brain R, Addison C, Rudge K, Remm M, Grimaldi M, et al. A C-TERMINAL ALPHA-HELIX PLUS BASIC REGION MOTIF IS THE MAJOR STRUCTURAL DETERMINANT OF P53 TETRAMERIZATION. Oncogene. 1992;7(8):1513-23.

107.	Stommel JM, Marchenko ND, Jimenez GS, Moll UM, Hope TJ, Wahl GM. A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. Embo Journal. 1999;18(6):1660-72.

108.	Giaccia AJ, Kastan MB. The complexity of p53 modulation: emerging patterns from divergent signals. Genes & Development. 1998;12(19):2973-83.

109.	May P, May E. Twenty years of p53 research: structural and functional aspects of the p53 protein. Oncogene. 1999;18(53):7621-36.

110.	Gu W, Roeder RG. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. Cell. 1997;90(4):595-606.

111.	Hupp TR, Meek DW, Midgley CA, Lane DP. REGULATION OF THE SPECIFIC DNA-BINDING FUNCTION OF P53. Cell. 1992;71(5):875-86.

112.	Halazonetis TD, Davis LJ, Kandil AN. WILD-TYPE P53 ADOPTS A MUTANT-LIKE CONFORMATION WHEN BOUND TO DNA. Embo Journal. 1993;12(3):1021-8.

113.	Hupp TR, Sparks A, Lane DP. SMALL PEPTIDES ACTIVATE THE LATENT SEQUENCE-SPECIFIC DNA-BINDING FUNCTION OF P53. Cell. 1995;83(2):237-45.

114.	Zhu J, Zhang S, Jiang J, Chen X. Definition of the p53 Functional Domains Necessary for Inducing Apoptosis. Journal of Biological Chemistry. 2000;275(51):39927-34.

115.	Loeb LA. MUTATOR PHENOTYPE MAY BE REQUIRED FOR MULTISTAGE CARCINOGENESIS. Cancer Research. 1991;51(12):3075-9.

116.	Kinzler KW, Vogelstein B. Cancer-susceptibility genes - Gatekeepers and caretakers. Nature. 1997;386(6627):761-3.

117.	Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability - an evolving hallmark of cancer. Nature Reviews Molecular Cell Biology. 2010;11(3):220-8.

References

118.   Sadikovic B, Al-Romaih K, Squire J, Zielenska M. Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer. Curr Genomics. 2008;9(6):394-408.

119.   Fishel R, Lescoe MK, Rao MRS, Copeland NG, Jenkins NA, Garber J, et al. THE HUMAN MUTATOR GENE HOMOLOG MSH2 AND ITS ASSOCIATION WITH HEREDITARY NONPOLYPOSIS COLON-CANCER. Cell. 1993;75(5):1027-38.

120.   Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, et al. MUTATIONS OF A MUTS HOMOLOG IN HEREDITARY NONPOLYPOSIS COLORECTAL-CANCER. Cell. 1993;75(6):1215-25.

121.   Kennedy RD, D'Andrea AD. DNA repair pathways in clinical practice: Lessons from pediatric cancer susceptibility syndromes. Journal of Clinical Oncology. 2006;24(23):3799-808.

122.   Bachrati CZ, Hickson ID. RecQ helicases: suppressors of tumorigenesis and premature aging. Biochemical Journal. 2003;374:577-606.

123.   Shay JW, Wright WE. Role of telomeres and telomerase in cancer. Semin Cancer Biol. 2011;21(6):349-53.

124.   Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21.

125.   Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45(10):1127-33.

126.   Rajagopalan H, Lengauer C. Aneuploidy and cancer. Nature. 2004;432(7015):338-41.

127.   Halazonetis TD, Gorgoulis VG, Bartek J. An oncogene-induced DNA damage model for cancer development. Science. 2008;319(5868):1352-5.

128.   Wiseman H, Halliwell B. Damage to DNA by reactive oxygen and nitrogen species: Role in inflammatory disease and progression to cancer. Biochemical Journal. 1996;313:17-29.

129.   Delbos F, Aoufouchi S, Faili A, Weill JC, Reynaud CA. DNA polymerase η is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. J Exp Med. 2007;204(1):17-23.

130.   Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability - an evolving hallmark of cancer. Nature Reviews Molecular Cell Biology. 2010;11(3):220-8.

131.   Luo J, Solimini NL, Elledge SJ. Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. Cell. 2009;136(5):823-37.

132.   Koberle B, Grimaldi KA, Sunters A, Hartley JA, Kelland LR, Masters JRW. DNA repair capacity and cisplatin sensitivity of human testis tumour cells. International Journal of Cancer. 1997;70(5):551-5.

133.   Usanova S, Piee-Staffa A, Sied U, Thomale J, Schneider A, Kaina B, et al. Cisplatin sensitivity of testis tumour cells is due to deficiency in interstrand-crosslink repair and low ERCC1-XPF expression. Molecular Cancer. 2010;9.

References

134.   Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. Journal of the National Cancer Institute. 2004;96(4):261-8.

135.   Wu X, Xu Y, Chai W, Her C. Causal Link between Microsatellite Instability and hMRE11 Dysfunction in Human Cancers. Molecular Cancer Research. 2011;9(11):1443-8.

136.   Ham MF, Takakuwa T, Luo WJ, Liu AG, Horii A, Aozasa K. Impairment of double-strand breaks repair and aberrant splicing of ATM and MRE11 in leukemia-lymphoma cell lines with microsatellite instability. Cancer Science. 2006;97(3):226-34.

137.   Fordham SE, Matheson EC, Scott K, Irving JAE, Allan JM. DNA mismatch repair status affects cellular response to Ara-C and other anti-leukemic nucleoside analogs. Leukemia. 2011;25(6):1046-9.

138.   Willems P, Claes K, Baeyens A, Vandersickel V, Werbrouck J, De Ruyck K, et al. Polymorphisms in nonhomologous end-joining genes associated with breast cancer risk and chromosomal radiosensitivity. Genes Chromosomes & Cancer. 2008;47(2):137-48.

139.   Liu Y, Zhou K, Zhang H, Shugart YY, Chen L, Xu Z, et al. Polymorphisms of LIG4 and XRCC4 involved in the NHEJ pathway interact to modify risk of glioma. Human Mutation. 2008;29(3):381-9.

140.   McClendon AK, Osheroff N. DNA topoisomerase II, genotoxicity, and cancer. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis. 2007;623(1-2):83-97.

141.   Rosenzweig KE, Youmell MB, Palayoor ST, Price BD. Radiosensitization of human tumor cells by the phosphatidylinositol 3-kinase inhibitors Wortmannin and LY294002 correlates with inhibition of DNA-dependent protein kinase and prolonged G(2)-M delay. Clinical Cancer Research. 1997;3(7):1149-56.

142.   Boulton S, Kyle S, Durkacz BW. Mechanisms of enhancement of cytotoxicity in etoposide and ionising radiation-treated cells by the protein kinase inhibitor wortmannin. European Journal of Cancer. 2000;36(4):535-41.

143.   Willmore E, de Caux S, Sunter NJ, Tilby MJ, Jackson GH, Austin CA, et al. A novel DNA-dependent protein kinase inhibitor, NU7026, potentiates the cytotoxicity of topoisomerase II poisons used in the treatment of leukemia. Blood. 2004;103(12):4659-65.

144.   Aloyz R, Grzywacz K, Xu ZY, Loignon M, Alaoui-Jamali MA, Panasci L. Imatinib sensitizes CLL lymphocytes to chlorambucil. Leukemia. 2004;18(3):409-14.

145.   Choudhury A, Zhao H, Jalali F, Rashid SA, Ran J, Supiot S, et al. Targeting homologous recombination using imatinib results in enhanced tumor cell chemosensitivity and radiosensitivity. Molecular Cancer Therapeutics. 2009;8(1):203-13.

146.   Lewis KA, Mullany S, Thomas B, Chien J, Loewen R, Shridhar V, et al. Heterozygous ATR mutations in mismatch repair-deficient cancer cells have functional significance. Cancer Research. 2005;65(16):7091-5.

References

147.   Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, III, Hurov KE, Luo J, et al. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science. 2007;316(5828):1160-6.

148.   Hickson I, Yan Z, Richardson CJ, Green SJ, Martin NMB, Orr AI, et al. Identification and characterization of a novel and specific inhibitor of the ataxia-telangiectasia mutated kinase ATM. Cancer Research. 2004;64(24):9152-9.

149.   Hollstein M, Sidransky D, Vogelstein B, Harris CC. P53 Mutations in Human Cancers. Science. 1991;253(5015):49-53.

150.   Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. Nature. 2000;408(6810):307-10.

151.   Peller S, Rotter V. TP53 in hematological cancer: Low incidence of mutations with significant clinical relevance. Human Mutation. 2003;21(3):277-84.

152.   Iacopetta B. TP53 mutation in colorectal cancer. Human Mutation. 2003;21(3):271-6.

153.   Schuijer M, Berns EMJJ. TP53 and ovarian cancer. Human Mutation. 2003;21(3):285-91.

154.   Blons H, Laurent-Puig P. TP53 and head and neck neoplasms. Human Mutation. 2003;21(3):252-7.

155.   Miller C, Mohandas T, Wolf D, Prokocimer M, Rotter V, Koeffler HP. HUMAN P53 GENE LOCALIZED TO SHORT ARM OF CHROMOSOME-17. Nature. 1986;319(6056):783-4.

156.   Rivlin N, Brosh R, Oren M, Rotter V. Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. Genes & Cancer. 2011;2(4):466-74.

157.   Brosh R, Rotter V. When mutants gain new powers: news from the mutant p53 field. Nature Reviews Cancer. 2009;9(10):701-13.

158.   Bullock AN, Fersht AR. Rescuing the function of mutant p53. Nature Reviews Cancer. 2001;1(1):68-76.

159.   Kato S, Han SY, Liu W, Otsuka K, Shibata H, Kanamaru R, et al. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(14):8424-9.

160.   Sigal A, Rotter V. Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome. Cancer Research. 2000;60(24):6788-93.

161.   Bossi G, Lapi E, Strano S, Rinaldo C, Blandino G, Sacchi A. Mutant p53 gain of function: reduction of tumor malignancy of human cancer cell lines through abrogation of mutant p53 expression. Oncogene. 2006;25(2):304-9.

162.   Zenz T, Vollmer D, Trbusek M, Smardova J, Benner A, Soussi T, et al. TP53 mutation profile in chronic lymphocytic leukemia: evidence for a disease specific profile from a comprehensive analysis of 268 mutations. Leukemia. 2010;24(12):2072-9.

References

163.   Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, et al. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. Human Mutation. 2007;28(6):622-9.

164.   Zenz T, Eichhorst B, Busch R, Denzel T, Haebe S, Winkler D, et al. TP53 Mutation and Survival in Chronic Lymphocytic Leukemia. Journal of Clinical Oncology. 2010;28(29):4473-9.

165.   Jeromin S, Kern W, Schabath R, Alpermann T, Nadarajah N, Meggendorfer M, et al. Modulation of the Clonal Composition in Relapsed CLL: A Study Based on Targeted Deep-Sequencing of ATM, BIRC3, NOTCH1, POT1, SF3B1, SAMHD1 and TP53. Blood. 2015;126(23).

166.   Zenz T, Habe S, Denzel T, Mohr J, Winkler D, Buhler A, et al. Detailed analysis of p53 pathway defects in fludarabine-refractory chronic lymphocytic leukemia (CLL): dissecting the contribution of 17p deletion, TP53 mutation, p53-p21 dysfunction, and miR34a in a prospective clinical trial. Blood. 2009;114(13):2589-97.

167.   Trbusek M, Smardova J, Malcikova J, Sebejova L, Dobes P, Svitakova M, et al. Missense mutations located in structural p53 DNA-binding motifs are associated with extremely poor survival in chronic lymphocytic leukemia. J Clin Oncol. 2011;29(19):2703-8.

168.   Dohner H, Fischer K, Bentz M, Hansen K, Benner A, Cabot G, et al. P53 GENE DELETION PREDICTS FOR POOR SURVIVAL AND NONRESPONSE TO THERAPY WITH PURINE ANALOGS IN CHRONIC B-CELL LEUKEMIAS. Blood. 1995;85(6):1580-9.

169.   Zenz T, Gribben JG, Hallek M, Dohner H, Keating MJ, Stilgenbauer S. Risk categories and refractory CLL in the era of chemoimmunotherapy. Blood. 2012;119(18):4101-7.

170.   Dicker F, Herholz H, Schnittger S, Nakao A, Patten N, Wu L, et al. The detection of TP53 mutations in chronic lymphocytic leukemia independently predicts rapid disease progression and is highly correlated with a complex aberrant karyotype. Leukemia. 2008;23(1):117-24.

171.   Moreno C, Montserrat E. New prognostic markers in chronic lymphocytic leukemia. Blood Reviews. 2008;22(4):211-9.

172.   Walewska R, Oscier D. Prognostic markers in chronic lymphocytic leukemia. Advances in the Treatment of B-Cell Chronic Lymphocytic Leukemia: Future Medicine Ltd; 2012. p. 76-86.

173.   Damle RN, Calissano C, Chiorazzi N. Chronic lymphocytic leukaemia: a disease of activated monoclonal B cells. Best Practice & Research Clinical Haematology. 2010;23(1):33-45.

174.   Stilgenbauer S, Sander S, Bullinger L, Benner A, Leupolt E, Winkler D, et al. Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. Haematologica. 2007;92(9):1242-5.

175.   Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. Cell. 2013;152(4):714-26.

176.   Shanafelt TD, Call TG. Current approach to diagnosis and management of chronic lymphocytic leukemia. Mayo Clinic Proceedings. 2004;79(3):388-98.

177.   Speicher MR, Carter NP. The new cytogenetics: Blurring the boundaries with molecular biology. Nature Reviews Genetics. 2005;6(10):782-92.

References

178.   Oscier DG, Gardiner AC, Mould SJ, Glide S, Davis ZA, Ibbotson RE, et al. Multivariate analysis of prognostic factors in CLL: clinical stage, IGVH gene mutational status, and loss or mutation of the p53 gene are independent prognostic factors. Blood. 2002;100(4):1177-84.

179.   Juliusson G, Oscier DG, Fitchett M, Ross FM, Stockdill G, Mackie MJ, et al. Prognostic Subgroups in B-Cell Chronic Lymphocytic-Leukemia Defined by Specific Chromosomal-Abnormalities. New England Journal of Medicine. 1990;323(11):720-4.

180.   Wan TS, Ma ES. The role of FISH in hematologic cancer. http://dxdoiorg/102217/ijh129. 2012.

181.   Dohner H, Stilgenbauer S, Benner A, Leupolt E, Krober A, Bullinger L, et al. Genomic aberrations and survival in chronic lymphocytic leukemia. New England Journal of Medicine. 2000;343(26):1910-6.

182.   Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet. 2013;206(12):432-40.

183.   Aradhya S, Cherry AM. Array-based comparative genomic hybridization: clinical contexts for targeted and whole-genome designs. Genetics in Medicine. 2007;9(9).

184.   Xu X, Johnson EB, Leverton L, Arthur A, Watson Q, Chang FL, et al. The advantage of using SNP array in clinical testing for hematological malignancies--a comparative study of three genetic testing methods. Cancer Genet. 2013;206(9-10):317-26.

185.   Hagenkord JM, Monzon FA, Kash SF, Lilleberg S, Xie Q, Kant JA. Array-Based Karyotyping for Prognostic Assessment in Chronic Lymphocytic Leukemia. The Journal of Molecular Diagnostics. 2010;12(2):184-96.

186.   Konstantinos KV, Panagiotis P, Antonios VT, Agelos P, Argiris NV. PCR–SSCP: A Method for the Molecular Analysis of Genetic Diseases. Molecular Biotechnology. 2007;38(2):155-63.

187.   Brautigam S. DHPLC mutation analysis of phenylketonuria. Molecular Genetics and Metabolism. 2003;78(3):205-10.

188.   Sanger F, Nicklen S, Coulson AR. DNA SEQUENCING WITH CHAIN-TERMINATING INHIBITORS. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.

189.   Lee LG, Spurgeon SL, Heiner CR, Benson SC, Rosenblum BB, Menchen SM, et al. New energy transfer dyes for DNA sequencing. Nucleic Acids Research. 1997;25(14):2816-22.

190.   Metzker ML, Lu J, Gibbs RA. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. Science. 1996;271(5254):1420-2.

191.   Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. FLUORESCENCE DETECTION IN AUTOMATED DNA-SEQUENCE ANALYSIS. Nature. 1986;321(6071):674-9.

192.   Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, et al. A SYSTEM FOR RAPID DNA SEQUENCING WITH FLUORESCENT CHAIN-TERMINATING DIDEOXYNUCLEOTIDES. Science. 1987;238(4825):336-41.

References

193.	Takahashi S, Murakami K, Anazawa T, Kambara H. MULTIPLE SHEATH-FLOW GEL CAPILLARY-ARRAY ELECTROPHORESIS FOR MULTICOLOR FLUORESCENT DNA DETECTION. Analytical Chemistry. 1994;66(7):1021-6.

194.	Kheterpal I, Scherer JR, Clark SM, Radhakrishnan A, Ju JY, Ginther CL, et al. DNA sequencing using a four-color confocal fluorescence capillary array scanner. Electrophoresis. 1996;17(12):1852-9.

195.	Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research. 1998;8(3):175-85.

196.	Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research. 1998;8(3):186-94.

197.	Metzker ML. Emerging technologies in DNA sequencing. Genome Research. 2005;15(12):1767-76.

198.	Flaherty P, Natsoulis G, Muralidharan O, Winters M, Buenrostro J, Bell J, et al. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Research. 2012;40(1).

199.	ElSharawy A, Warner J, Olson J, Forster M, Schilhabel MB, Link DR, et al. Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next generation sequencing. BMC Genomics. 2012;13(1):500.

200.	van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in Genetics. 2014;30(9):418-26.

201.	Grada A, Weinbrecht K. Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology. 2013;133(8):1-4.

202.	Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:1-33.

203.	Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics. 2010;11(6):415-25.

204.	Karki R, Pandya D, Elston RC, Ferlini C. Defining "mutation" and "polymorphism" in the era of personal genomics. BMC Med Genomics. 2015;8.

205.	Shen Y, Song R, Pe'er I. Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association. Bioinformatics. 2017;27(14):1995-7.

206.	Telenti A, Pierce LCT, Biggs WH, Iulio Jd, Wong EHM, Fabani MM, et al. Deep sequencing of 10,000 human genomes. 2016.

207.	Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17(6):333-51.

208.	Ion PGM™ and Ion Proton™ System Chips 2017 [Available from: https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-

References

sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-ion-proton-system-chips.html.

209.   Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. G3 (Bethesda). 2015;5(8):1543-50.

210.   Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6.

211.   Christodoulou K, Wiskin AE, Gibson J, Tapper W, Willis C, Afzal NA, et al. Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes. Gut. 2013;62(7):977-84.

212.   Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nature Reviews Genetics. 2013;14(4):295-300.

213.   Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nature Methods. 2010;7(2):111-8.

214.   Sint D, Raso L, Traugott M. Advances in multiplex PCR: balancing primer efficiencies and improving detection success. Methods Ecol Evol. 2012;3(5):898-905.

215.   Shah RK, Shum HC, Rowat AC, Lee D, Agresti JJ, Utada AS, et al. Designer emulsions using microfluidics. Materials Today. 2008;11(4):18-27.

216.   Leamon JH, Link DR, Egholm M, Rothberg JM. Overview: methods and applications for droplet compartmentalization of biology. Nat Methods. 2006;3(7):541-3.

217.   Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111-8.

218.   Chen R, Im H, Snyder M. Whole-Exome Enrichment with the Agilent SureSelect Human All Exon Platform. Cold Spring Harb Protoc. 2015(7).

219.   Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Briefings in Functional Genomics. 2011;10(6):374-86.

220.   Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology. 2009;27(2):182-9.

221.   Narayan A, Carriero NJ, Gettinger SN, Kluytenaar J, Kozak KR, Yock TI, et al. Ultrasensitive Measurement of Hotspot Mutations in Tumor DNA in Blood Using Error-Suppressed Multiplexed Deep Sequencing. 2012.

222.   Genomic Applications in Pathology 2015. Available from: https://link.springer.com/book/10.1007%2F978-1-4939-0727-4.

223.   Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nature Biotechnology. 2012;30(11):1033-6.

References

224.   McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. 2010.

225.   Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2017;25(16):2078-9.

226.   Christoph Endrullat JG, Philipp Franke, Marcus Frohme. Standardization and quality management in next-generation sequencing - ScienceDirect. 2016.

227.   Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2012;14(2):178-92.

228.   Netto GJ. Genomic Applications in Pathology - Springer. 2015.

229.   Sabeti PC. A global reference for human genetic variation. 2015.

230.   Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.

231.   Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology. 2012;2012:1-11.

232.   Kohn AB, Moroz TP, Barnes JP, Netherton M, Moroz LL. Single-Cell Semiconductor Sequencing. Methods Mol Biol. 2013;1048:247-84.

233.   Wang YL, Wen ZJ, Shen JW, Cheng WW, Li J, Qin XL, et al. Comparison of the performance of Ion Torrent chips in noninvasive prenatal trisomy detection. Journal of Human Genetics. 2014;59(7):393-6.

234.   Hood L, Rowen L. The Human Genome Project: big science transforms biology and medicine. Genome Med. 52013. p. 79.

235.   Consortium. IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.

236.   Secko D. Phase I of HapMap Complete 2005 [Available from: https://www.the-scientist.com/?articles.view/articleNo/23483/title/Phase-I-of-HapMap-Complete/.

237.   Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

238.   Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-73.

239.   Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.

240.   Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82-90.

References

241.   Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, et al. Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. Science Translational Medicine. 2012;4(154):1-27.

242.   Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 2011;475(7354):101-5.

243.   Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, et al. SF3B1and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. New England Journal of Medicine. 2011;365(26):2497-506.

244.   Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nature Genetics. 2011;44(1):47-52.

245.   Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, et al. Mutations driving CLL and their evolution in progression and relapse. Nature. 2015;526(7574):525-30.

246.   Brown JR, Byrd JC, Coutre SE, Benson DM, Flinn IW, Wagner-Johnston ND, et al. Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110 delta, for relapsed/refractory chronic lymphocytic leukemia. Blood. 2014;123(22):3390-7.

247.   Plevova K, Francova HS, Burckova K, Brychtova Y, Doubek M, Pavlova S, et al. Multiple productive immunoglobulin heavy chain gene rearrangements in chronic lymphocytic leukemia are mostly derived from independent clones. Haematologica. 2014;99(2):329-38.

248.   Hurtado AM, Hamedi C, Przychodzen BP, Anton AI, Garcia-Malo MD, Soler G, et al. Anticipating High-Risk Chromosome Clonal Evolution in Chronic Lymphocytic Leukemia: A Next Generation Sequencing and Longitudinal FISH Combined Study. 2014.

249.   Pettitt AR, Polydoros F, Dodd J, Oates M, Lin K, Kalakonda N, et al. Final Results of the Ncri Cll210 Trial of Alemtuzumab, Dexamethasone and Lenalidomide in Patients with High-Risk Cll (Original Protocol). Haematologica. 2016;101:726-.

250.   Grievink HW, Luisman T, Kluft C, Moerland M, Malone KE. Comparison of Three Isolation Techniques for Human Peripheral Blood Mononuclear Cells: Cell Recovery and Viability, Population Composition, and Cell Functionality. Biopreservation and Biobanking. 2016;14(5):410-5.

251.   Rigoutsos I, Lee SK, Nam SY, Anfossi S, Pasculli B, Pichler M, et al. N-BLR, a primate-specific non-coding transcript leads to colorectal cancer invasion and migration. Genome Biology. 2017;18.

252.   Lin K, Glenn MA, Harris RJ, Duckworth AD, Dennett S, Cawley JC, et al. c-Abl expression in chronic lymphocytic leukemia cells: Clinical and therapeutic implications. Cancer Research. 2006;66(15):7801-9.

253.   O'Neill M, McPartlin J, Arthure K, Riedel S, McMillan ND, editors. Comparison of the TLDA with the Nanodrop and the reference Qubit system. 16th Conference in the Biennial Sensors and their Applications; 2011 2011; Univ Coll Cork, Tyndall Natl Inst, Cork, IRELAND2011.

References

254. Simbolo M, Gottardi M, Corbo V, Fassan M, Mafficini A, Malpeli G, et al. DNA qualification workflow for next generation sequencing of histopathological samples. PLoS One. 2013;8(6):1-8.

255. Smith DR. Agarose gel electrophoresis. Methods Mol Biol. 1996;58:17-21.

256. Custom SureSelect 2017 [Available from: http://www.genomics.agilent.com/en/SureSelect-DNA-Target-Enrichment-Baits-/Custom-SureSelect-/?cid=AG-PT-124&tabId=AG-PR-1309.

257. Genomics A. SureSelect Target Enrichment System for Sequencing on Ion Proton, Version C0 2016 [C0:[Available from: https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwi7hauqzdTXAhWELcAKHd2TAagQFggoMAA&url=https%3A%2F%2Fwww.agilent.com%2Fcs%2Flibrary%2Fusermanuals%2Fpublic%2FG7530-90005.pdf&usg=AOvVaw3dSQLEusy5V3KUYCB2k4tM.

258. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409(6822):928-33.

259. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11.

260. Pan Y, Yan C, Hu Y, Fan Y, Pan Q, Wan Q, et al. Distribution bias analysis of germline and somatic single-nucleotide variations that impact protein functional site and neighboring amino acids. Sci Rep. 2017;7.

261. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):777-83.

262. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods. 2009;6(9):677-81.

263. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423-5.

264. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2004;Chapter 4:1-14.

265. Panagiotou OA, Ioannidis JPA, Project G-WS. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. International Journal of Epidemiology. 2012;41(1):273-86.

266. Wood RD, Mitchell M, Sgouros J, Lindahl T. Human DNA repair genes. Science. 2001;291(5507):1284-9.

267. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis. 2005;577(1-2):275-83.

268. Fu D, Calvo JA, Samson LD. Balancing repair and tolerance of DNA damage caused by alkylating agents. Nature Reviews Cancer. 2012;12(2):104-20.

References

269.   Thompson LH. Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: The molecular choreography. Mutation Research-Reviews in Mutation Research. 2012;751(2):158-246.

270.   DiGiovanna JJ, Kraemer KH. Shining a Light on Xeroderma Pigmentosum. Journal of Investigative Dermatology. 2012;132(3):785-96.

271.   Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nature Reviews Genetics. 2014;15(9):585-98.

272.   Gaillard H, Garcia-Muse T, Aguilera A. Replication stress and cancer. Nature Reviews Cancer. 2015;15(5):276-89.

273.   Sharma S, Hicks JK, Chute CL, Brennan JR, Ahn JY, Glover TW, et al. REV1 and polymerase zeta facilitate homologous recombination repair. Nucleic Acids Research. 2012;40(2):682-91.

274.   Bernstein C, Bernstein H, Payne CM, Garewal H. DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis. Mutation Research-Reviews in Mutation Research. 2002;511(2):145-78.

275.   Roos WP, Thomas AD, Kaina B. DNA damage and the balance between survival and death in cancer biology. Nature Reviews Cancer. 2016;16(1):20-33.

276.   Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis. 2005;577(1-2):275-83.

277.   Gallagher SR, Desjardins PR. Quantitation of DNA and RNA with absorption and fluorescence spectroscopy. Curr Protoc Protein Sci. 2008;Appendix 3:1-21.

278.   Ihle MA, Fassunke J, Konig K, Grunewald I, Schlaak M, Kreuzberg N, et al. Comparison of high resolution melting analysis, pyrosequencing, next generation sequencing and immunohistochemistry to conventional Sanger sequencing for the detection of p.V600E and non-p.V600E BRAF mutations. BMC Cancer. 2014;14:13.

279.   Cottrell CE, Al-Kateb H, Bredemeyer AJ, Duncavage EJ, Spencer DH, Abel HJ, et al. Validation of a next-generation sequencing assay for clinical molecular oncology. J Mol Diagn. 2014;16(1):89-105.

280.   Johnson DB, Dahlman KH, Knol J, Gilbert J, Puzanov I, Means-Powell J, et al. Enabling a Genetically Informed Approach to Cancer Medicine: A Retrospective Evaluation of the Impact of Comprehensive Tumor Profiling Using a Targeted Next-Generation Sequencing Panel. Oncologist. 2014;19(6):616-22.

281.   Fichtenholtz A, Parker A, Donahue A, Maillet A, Boshoff C, Vietz C, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nature Biotechnology. 2013;31(11):1023.

282.   Son D-S, Park D, Park G, Ryu GH, Jeon H-J, Chung J, et al. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. Scientific Reports. 2016;6:26732.

References

283.   Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. Plos Computational Biology. 2013;9(4):18.

284.   Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Briefings in Bioinformatics. 2016;17(1):154-79.

285.   Feng WX, Zhao S, Xue DK, Song FF, Li ZW, Chen DJ, et al. Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. Bmc Genomics. 2016;17.

286.   Liu L, Li YH, Li SL, Hu N, He YM, Pong R, et al. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology. 2012.

287.   Liu L, Li YH, Li SL, Hu N, He YM, Pong R, et al. Comparison of Next-Generation Sequencing Systems. Advances in Biofuel Production: Algae and Aquatic Plants. 2014:279-303.

288.   Butte AJ, Euskirchen G, Lam HYK, Karczewski KJ, Clark MJ, Snyder M, et al. Performance comparison of exome DNA sequencing technologies. Nature Biotechnology. 2011;29(10):908.

289.   Hollinger A, Russ C, Nusbaum C, Jaffe DB, Costello M, Ross MG, et al. Characterizing and measuring bias in sequence data. Genome Biology. 2013;14(5).

290.   Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):1-14.

291.   Guo Y, Li J, Li CI, Long JR, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. Bmc Genomics. 2012;13.

292.   Guo Y, Ye F, Sheng QH, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Briefings in Bioinformatics. 2014;15(6):879-89.

293.   Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 2003;312:207-13.

294.   Wang G, Peng B, Leal S. Variant Association Tools for Quality Control and Analysis of Large-Scale Sequence and Genotyping Array Data. Am J Hum Genet. 2014;94(5):770-83.

295.   Bainbridge MN, Wang M, Wu YQ, Newsham I, Muzny DM, Jefferies JL, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biology. 2011;12(7).

296.   McCarthy H, Wierda WG, Barron LL, Cromwell CC, Wang J, Coombes KR, et al. High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor-prognosis chronic lymphocytic leukemia. Blood. 2003;101(12):4903-8.

297.   Kasar S, Kim J, Improgo R, Tiao G, Polak P, Haradhvala N, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat Commun. 2015;6.

References

298.   Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-+.

299.   Zhao Y, Gregory MT, Biertumpfel C, Hua YJ, Hanaoka F, Yang W. Mechanism of somatic hypermutation at the WA motif by human DNA polymerase eta. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(20):8146-51.

300.   Faili A, Aoufouchi S, Weller S, Vuillier F, Stary A, Sarasin A, et al. DNA Polymerase η Is Involved in Hypermutation Occurring during Immunoglobulin Class Switch Recombination. J Exp Med. 2004;199(2):265-70.

301.   Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell. 2017;170(3):534-47.

302.   Abbas T, Keaton MA, Dutta A. Genomic Instability in Cancer. 2013.

303.   Bacolla A, Tainer JA, Vasquez KM, Cooper DN. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res. 2016;44(12):5673-88.

304.   Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nature Reviews Genetics. 2015;16(3):172-83.

305.   Edelmann J, Holzmann K, Miller F, Winkler D, Buhler A, Zenz T, et al. High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. Blood. 2012;120(24):4783-94.

306.   Janku F, Lee JJ, Tsimberidou AM, Hong DS, Naing A, Falchook GS, et al. PIK3CA Mutations Frequently Coexist with RAS and BRAF Mutations in Patients with Advanced Cancers. Plos One. 2011;6(7).

307.   McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. 2013.

308.   Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42(Database issue):472-7.

309.   Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8(3):1-13.

310.   Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363-76.

311.   Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biology. 2013;14(5).

312.   Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics. 2012;13(Suppl 8):1-8.

313.   Knight SJL, Yau C, Clifford R, Timbs AT, Akha ES, Dreau HM, et al. Quantification of subclonal distributions of recurrent genomic aberrations in paired pre-treatment and relapse samples from patients with B-cell chronic lymphocytic leukemia. Leukemia. 2012;26(7):1564-75.

References

314.   Henninger EE, Pursell ZF. DNA polymerase epsilon and its roles in genome stability. IUBMB Life. 2014;66(5):339-51.

315.   Chilkova O, Stenlund P, Isoz I, Stith CM, Grabowski P, Lundström EB, et al. The eukaryotic leading and lagging strand DNA polymerases are loaded onto primer-ends via separate mechanisms but have comparable processivity in the presence of PCNA. Nucleic Acids Res. 2007;35(19):6588-97.

316.   Burgers PMJ. Polymerase Dynamics at the Eukaryotic DNA Replication Fork. 2009.

317.   Moser BA, Subramanian L, Chang YT, Noguchi C, Noguchi E, Nakamura TM. Differential arrival of leading and lagging strand DNA polymerases at fission yeast telomeres. 2009.

318.   Lou H, Komata M, Katou Y, Guan Z, Reis CC, Budd M, et al. Mrc1 and DNA polymerase ε function together in linking DNA replication and the S phase checkpoint. Mol Cell. 2008;32(1):106-17.

319.   Henninger EE, Pursell ZF. DNA Polymerase epsilon and Its Roles in Genome Stability. Iubmb Life. 2014;66(5):339-51.

320.   Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013;45(2):136-44.

321.   Briggs S, Tomlinson I. Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. J Pathol. 2013;230(2):148-53.

322.   Church DN, Briggs SE, Palles C, Domingo E, Kearsey SJ, Grimes JM, et al. DNA polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. Hum Mol Genet. 2013;22(14):2820-8.

323.   Rayner E, van Gool IC, Palles C, Kearsey SE, Bosse T, Tomlinson I, et al. A panoply of errors: polymerase proofreading domain mutations in cancer. Nature Reviews Cancer. 2016;16(2):71-81.

324.   Bellone S, Centritto F, Black J, Schwab C, English D, Cocco E, et al. Polymerase epsilon (POLE) ultra-mutated tumors induce robust tumor-specific CD4+T cell responses in endometrial cancer patients. Gynecologic Oncology. 2015;138(1):11-7.

325.   Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nature Biotechnology. 2016;34(2):155-63.

326.   Getz G. Integrated genomic characterization of endometrial carcinoma (vol 497, pg 67, 2013). Nature. 2013;500(7461):1.

327.   Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7.

328.   Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):777-83.

References

329.   Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nature Biotechnology. 2016;34(2):155-+.

330.   Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. 2013.

331.   Zenz T, Haebe S, Denzel T, Mohr J, Winkler D, Buehler A, et al. Detailed analysis of p53 pathway defects in fludarabine-refractory chronic lymphocytic leukemia (CLL): dissecting the contribution of 17p deletion, TP53 mutation, p53-p21 dysfunction, and miR34a in a prospective clinical trial. Blood. 2009;114(13):2589-97.

332.   Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 2015;526(7574):519-24.

333.   Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7.

334.   Getz G. Integrated genomic characterization of endometrial carcinoma (vol 497, pg 67, 2013). Nature. 2013;500(7461).

335.   Kane DP, Shcherbakova PV. A Common Cancer-Associated DNA Polymerase epsilon Mutation Causes an Exceptionally Strong Mutator Phenotype, Indicating Fidelity Defects Distinct from Loss of Proofreading. Cancer Research. 2014;74(7):1895-901.

336.   Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155(4):948-62.

337.   Lazarian G, Tausch E, Eclache V, Sebaa A, Bianchi V, Letestu R, et al. TP53 mutations are early events in chronic lymphocytic leukemia disease progression and precede evolution to complex karyotypes. Int J Cancer. 2016;139(8):1759-63.

338.   Malcikova J, Stano-Kozubik K, Tichy B, Kantorova B, Pavlova S, Tom N, et al. Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia. Leukemia. 2015;29(4):877-85.

339.   Landau DA, Tausch E, Taylor-Weiner AN, Bottcher S, Bahlo J, Stewart C, et al. Subclonal Driver Mutations Predict Shorter Progression Free Survival in Chronic Lymphocytic Leukemia Following First-Line Chemo(immuno) Therapy: Results from the CLL8 Trial. Blood. 2014;124(21).

340.   Ganster C, Neesen J, Zehetmayer S, Jager U, Esterbauer H, Mannhalter C, et al. DNA Repair Polymorphisms Associated with Cytogenetic Subgroups in B-Cell Chronic Lymphocytic Leukemia. Genes Chromosomes & Cancer. 2009;48(9):760-7.

341.   Muller C, Calsou P, Salles B. The activity of the DNA-dependent protein kinase (DNA-PK) complex is determinant in the cellular response to nitrogen mustards. Biochimie. 2000;82(1):25-8.

342.   Bello VE, Aloyz RS, Christodoulopoulos G, Panasci LC. Homologous recombinational repair vis-a-vis chlorambucil resistance in chronic lymphocytic leukemia. Biochemical Pharmacology. 2002;63(9):1585-8.

References

343. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

344. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu YT, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research. 2009;19(9):1527-41.

345. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6.

346. SureSelect - How it Works 2017 [Available from: https://www.genomics.agilent.com/article.jsp?pageId=3083.

347. Agilent SureDesign 2017 [Available from: https://earray.chem.agilent.com/suredesign/index.htm?sessiontimeout=true.

348. Smith AD, Xuan ZY, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. Bmc Bioinformatics. 2008;9.

349. Ciccia A, Elledge SJ. The DNA Damage Response: Making It Safe to Play with Knives. Molecular Cell. 2010;40(2):179-204.

350. Zhou BBS, Elledge SJ. The DNA damage response: putting checkpoints in perspective (Reprinted from Nature, vol 408, pg 433-439, 2001). Nature Reviews Molecular Cell Biology. 2001:25-31.

351. Dearth LR, Qian H, Wang T, Baroni TE, Zeng J, Chen SW, et al. Inactive full-length p53 mutants lacking dominant wild-type p53 inhibition highlight loss of heterozygosity as an important aspect of p53 status in human cancers. Carcinogenesis. 2007;28(2):289-98.

352. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. Bmc Genomics. 2012;13.

353. Wood RD, Mitchell M, Sgouros J, Lindahl T. Human DNA repair genes. Science. 2001;291(5507):1284-+.

354. Human DNA repair genes 2017 [Available from: https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html.

355. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends in Genetics. 2008;24(3):142-9.

356. Chrystoja CC, Diamandis EP. Whole Genome Sequencing as a Diagnostic Test: Challenges and Opportunities. Clinical Chemistry. 2014;60(5):724-33.

357. Landau Dan A, Carter Scott L, Stojanov P, McKenna A, Stevenson K, Lawrence Michael S, et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. Cell. 2013;152(4):714-26.

References

358.   Ouillette P, Malek S. Acquired genomic copy number aberrations in CLL. Adv Exp Med Biol. 2013;792:47-86.

359.   Ouillette P, Saiya-Cork K, Seymour E, Li C, Shedden K, Malek SN. Clonal Evolution, Genomic Drivers, and Effects of Therapy in Chronic Lymphocytic Leukemia. Clinical Cancer Research. 2013;19(11):2893-904.

360.   Getz G, Gabriel SB, Cibulskis K, Lander E, Sivachenko A, Sougnez C, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67-73.

361.   Bartkova J, Rezaei N, Liontos M, Karakaidos P, Kletsas D, Issaeva N, et al. Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. Nature. 2006;444(7119):633-7.

362.   Di Micco R, Fumagalli M, Cicalese A, Piccinin S, Gasparini P, Luise C, et al. Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. Nature. 2006;444(7119):638-42.

363.   Dicker F, Herholz H, Schnittger S, Nakao A, Patten N, Wu L, et al. The detection of TP53 mutations in chronic lymphocytic leukemia independently predicts rapid disease progression and is highly correlated with a complex aberrant karyotype. Leukemia. 2009;23(1):117-24.

364.   Zenz T, Kroeber A, Scherer K, Haebe S, Buehler A, Benner A, et al. Monoallelic TP53 inactivation is associated with poor prognosis in chronic lymphocytic leukemia: results from a detailed genetic characterization with long-term follow-up. Blood. 2008;112(8):3322-9.

Appendices

Appendix table 1: Details of the Designed Probes for Human DNA Maintenance Genes Using Suredesign (Agilent); I and II are Subtypes to NER.

Row labels	No. of target exons	No. of probe	No. of probes repli-cation	Size of target regions (bp)	Size of se-quence regions (bp)	Per-centage of Cover-age	Ex-ons of high cov-er-age	Ex-ons of lo w cov-er-age
BER	94	297	781	16716	23190	-	94	0
MBD4	8	34	43	1957	2492	100	8	0
MPG	5	17	82	1021	1320	100	5	0
MUTYH	17	42	109	2046	3378	100	17	0
NEIL1	10	30	103	1631	2375	100	10	0
NEIL2	4	17	26	1084	1260	100	4	0
NEIL3	10	33	54	2018	2580	100	10	0
NTHL1	6	19	103	1059	1460	100	6	0
OGG1	12	36	79	2098	2880	100	12	0
SMUG1	4	20	44	1151	1440	100	4	0
TDG	11	29	53	1464	2385	100	11	0
UNG	7	20	85	1187	1620	100	7	0
Cell-cycle control	125	396	675	21994	31107	-	125	0
CCNA1	9	28	31	1578	2220	100	9	0
CCNA2	8	25	42	1459	1980	100	8	0
CCNB1	9	28	44	1659	2220	100	9	0
CCNB2	9	27	42	1377	2148	100	9	0
CCNB3	11	76	76	4408	5220	100	11	0
CCND1	5	15	55	988	1200	100	5	0
CCND2	5	15	36	970	1200	100	5	0
CCND3	6	18	75	1151	1440	100	6	0
CCNE1	11	29	79	1471	2368	100	11	0
CCNE2	12	32	38	1554	2563	100	12	0
CDK1	7	19	21	1050	1560	100	7	0
CDK2	7	20	21	1037	1620	100	7	0
CDK4	7	19	30	1052	1560	100	7	0
CDK5	12	24	40	1119	2128	100	12	0
CDK6	7	21	45	1121	1680	100	7	0
Chromatin structure and	20	97	277	6136	7179	-	20	0

Appendices

Row labels	No. of target exons	No. of probe	No. of probes repli-cation	Size of target regions (bp)	Size of se-quence regions (bp)	Per-centage of Cover-age	Ex-ons of high cov-er-age	Ex-ons of lo w cov er-age
modification								
CHAF1A	15	56	148	3274	4260	100	15	0
H2AFX	1	7	70	452	480	100	1	0
SETMAR	4	34	59	2410	2439	100	4	0
Cytidine deaminase	5	14	22	697	1140	-	5	0
AICDA	5	14	22	697	1140	100	5	0
Direct reversal of damage	18	50	94	2870	4080	-	18	0
ALKBH2	3	12	17	846	900	100	3	0
ALKBH3	10	24	24	1207	2040	100	10	0
MGMT	5	14	53	817	1140	100	5	0
DNA polymerases (catalytic subunits)	302	1048	2078	58805	80534	-	302	0
MAD2L2	7	20	53	1170	1620	100	7	0
PCNA	6	16	34	906	1292	100	6	0
POLB	17	37	46	1623	3220	100	17	0
POLD1	26	73	427	3847	5771	100	26	0
POLE	52	147	299	8169	11824	100	52	0
POLG	22	72	226	4160	5624	100	22	0
POLH	10	41	41	2342	3060	100	10	0
POLI	10	40	74	2423	3000	100	10	0
POLK	16	57	69	3100	4380	100	16	0
POLL	9	33	52	2092	2452	100	9	0
POLM	12	38	173	2232	2997	100	12	0
POLN	26	69	109	3439	5660	100	26	0
POLQ	31	148	183	8804	10740	100	31	0
REV1	24	81	85	4384	6300	100	24	0
REV3L	34	176	207	10114	12594	100	34	0
Editing and processing nucleases	64	243	486	13254	18302	-	64	0
APTX	10	28	31	1283	2162	100	10	0
ENDOV	10	25	113	1141	2100	100	10	0
EXO1	13	47	49	2801	3600	100	13	0

Appendices

Row labels	No. of target exons	No. of probe	No. of probes repli-cation	Size of target regions (bp)	Size of se-quence regions (bp)	Per-centage of Cover-age	Ex-ons of high cov-er-age	Ex-on s of lo w cov er-age
FAN1	13	58	68	3339	4260	100	13	0
FEN1	1	19	23	1163	1200	100	1	0
SPO11	13	32	57	1451	2700	100	13	0
TREX1	1	18	56	1130	1140	100	1	0
TREX2	3	16	89	946	1140	100	3	0
Fanconi anaemia	309	1122	1764	62148	85570	-	309	0
BRCA2	26	182	211	10777	12436	100	26	0
BRIP1	19	71	76	4130	5400	100	19	0
BTBD12	14	93	291	5785	6420	100	14	0
C19ORF40	4	13	16	728	1010	100	4	0
C1ORF86	11	31	142	1794	2520	100	11	0
FANCA	44	108	190	5374	9023	100	44	0
FANCB	8	45	53	2740	3180	100	8	0
FANCC	15	38	47	2127	3180	100	15	0
FANCD2	45	118	122	5479	9780	100	45	0
FANCE	10	34	109	1811	2640	100	10	0
FANCF	1	19	67	1145	1200	100	1	0
FANCG	14	40	54	2149	3239	100	14	0
FANCI	37	93	95	4731	7710	100	37	0
FANCL	15	33	39	1492	2843	100	15	0
FANCM	23	110	139	6607	7980	100	23	0
PALB2	13	66	80	3821	4729	100	13	0
RAD51C	10	28	33	1458	2280	100	10	0
Genes defective in diseases associated with sensitivity to DNA-damaging agents	140	431	870	24692	34070	-	140	0
ATM	62	181	205	10411	14526	100	62	0
BLM	21	78	86	4674	5940	100	21	0
C7ORF11	2	8	62	580	600	100	2	0
RECQL4	21	71	410	4048	5384	100	21	0
WRN	34	93	107	4979	7620	100	34	0
HR	273	842	1561	45930	66365	-	272	1
BRCA1	24	112	116	6184	8160	100	24	0

Appendices

Row labels	No. of target exons	No. of probe	No. of probes repli-cation	Size of target regions (bp)	Size of se-quence regions (bp)	Per-centage of Cover-age	Ex-ons of high cov-er-age	Ex-on s of lo w cov-er-age
DMC1	13	29	31	1287	2520	100	13	0
EME1	8	33	36	1912	2441	100	8	0
EME2	8	24	206	1495	1899	100	8	0
GEN1	13	53	61	2987	3925	100	13	0
GIYD1	6	18	131	948	1389	100	6	0
GIYD2	6	18	131	948	1389	100	6	0
MRE11A	21	52	58	2583	4336	100	21	0
MUS81	16	39	140	1976	3191	100	16	0
NBN	18	52	77	2668	4140	100	18	0
RAD50	27	79	92	4751	6327	100	27	0
RAD51	10	24	24	1355	1980	90	9	1
RAD51B	12	31	33	1429	2580	100	12	0
RAD51D	11	26	50	1386	2220	100	11	0
RAD52	12	34	55	1824	2750	100	12	0
RAD54B	16	62	67	3953	4680	100	16	0
RAD54L	18	50	55	2604	4048	100	18	0
RBBP8	18	59	65	3076	4610	100	18	0
SHFM1	6	12	13	480	1080	100	6	0
XRCC2	3	16	27	903	1140	100	3	0
XRCC3	7	19	93	1181	1560	100	7	0
MMR	168	582	802	32296	44841	-	168	0
MLH1	20	52	52	2711	4320	100	20	0
MLH3	13	83	86	4696	5734	100	13	0
MSH2	18	55	80	3344	4380	100	18	0
MSH3	24	72	121	3894	5760	100	24	0
MSH4	20	57	85	3211	4620	100	20	0
MSH5	25	63	90	3149	5188	100	25	0
MSH6	10	70	123	4318	4800	100	10	0
PMS1	16	62	82	3397	4639	100	16	0
PMS2	15	51	66	2889	3960	100	15	0
PMS2L3	7	17	17	687	1440	100	7	0
Modulation of nucleotide pools	21	56	189	3180	4560	-	20	1
DUT	7	17	103	915	1440	100	7	0
NUDT1	4	11	26	620	900	100	4	0
RRM2B	10	28	60	1645	2220	91	9	1

Appendices

Row labels	No. of target exons	No. of probe	No. of probes repli-cation	Size of target regions (bp)	Size of se-quence regions (bp)	Per-centage of Cover-age	Ex-ons of high cov-er-age	Ex-ons of low cov-er-age
NHEJ	152	474	584	26563	37459	-	152	0
DCLRE1C	15	49	59	2528	3840	100	15	0
LIG4	1	45	47	2756	2760	100	1	0
NHEJ1	8	22	22	1218	1800	100	8	0
PRKDC	87	246	336	14163	19879	100	87	0
XRCC4	8	23	24	1183	1860	100	8	0
XRCC5	21	53	57	2645	4440	100	21	0
XRCC6	12	36	39	2070	2880	100	12	0
NER	115	322	552	16878	26137	-	115	0
CETN2	5	13	27	619	1080	100	5	0
DDB1	27	75	87	3963	6061	100	27	0
DDB2	10	27	34	1484	2220	100	10	0
RAD23A	9	24	67	1272	1962	100	9	0
RAD23B	12	31	49	1527	2580	100	12	0
RPA1	17	44	82	2191	3660	100	17	0
RPA2	9	26	43	1291	2100	100	9	0
RPA3	4	9	10	446	780	100	4	0
XPA	6	18	59	942	1440	100	6	0
XPC	16	55	94	3143	4254	100	16	0
I-NER-related	100	327	659	17696	25396	-	100	0
ERCC6	23	118	125	6949	8460	100	23	0
ERCC8	13	32	37	1492	2700	100	13	0
MMS19	32	80	108	3917	6592	100	32	0
UVSSA	13	42	158	2390	3300	100	13	0
XAB2	19	55	231	2948	4344	100	19	0
II-TFIIH	203	557	1002	29276	45270	-	203	0
CCNH	9	24	26	1209	1980	100	9	0
CDK7	13	30	38	1352	2580	100	13	0
ERCC1	10	24	50	1235	2040	100	10	0
ERCC2	24	59	225	2816	4827	100	24	0
ERCC3	15	46	87	2857	3660	100	15	0
ERCC4	12	50	69	3011	3720	100	12	0
ERCC5	25	93	100	5500	7080	100	25	0
GTF2H1	14	36	38	2032	3000	100	14	0
GTF2H2	15	35	44	1488	3000	100	15	0
GTF2H3	13	29	44	1187	2462	100	13	0

Appendices

Row labels	No. of target exons	No. of probe	No. of probes replication	Size of target regions (bp)	Size of sequence regions (bp)	Percentage of Coverage	Exons of high coverage	Exons of lo w coverage
GTF2H4	14	34	64	1715	2771	100	14	0
GTF2H5	2	5	5	256	420	100	2	0
LIG1	29	70	186	3528	5930	100	29	0
MNAT1	8	22	26	1090	1800	100	8	0
Other BER and strand break-joining factors	75	223	486	12199	17446	-	75	0
APEX1	4	16	17	1037	1200	100	4	0
APEX2	6	28	38	1677	2040	100	6	0
APLF	11	32	66	1795	2580	100	11	0
LIG3	20	61	70	3484	4860	100	20	0
PNKP	16	39	203	1886	3091	100	16	0
XRCC1	18	47	92	2320	3675	100	18	0
Other conserved DNA damage response genes	300	1000	1505	56916	77108	-	298	2
ATR	49	151	173	9120	11990	100	49	0
ATRIP	14	47	107	2695	3660	100	14	0
CHEK1	13	36	48	1963	2922	100	13	0
CHEK2	20	47	54	2326	3808	97	19	1
CLK2	12	31	45	1740	2571	100	12	0
HUS1	11	26	34	1168	2170	100	11	0
MDC1	14	111	135	6550	7467	100	14	0
PER1	23	78	277	4548	5967	100	23	0
RAD1	6	17	18	1027	1380	100	6	0
RAD17	18	48	57	2465	3902	100	18	0
RAD9A	11	28	111	1402	2176	100	11	0
RIF1	37	149	175	8417	11113	100	37	0
TOPBP1	27	86	94	5109	6770	100	27	0
TP53	14	33	57	1697	2632	94	13	1
TP53BP1	31	112	120	6689	8580	100	31	0
Other identified genes with known or suspected DNA repair function	97	337	501	18975	25940	-	97	0
DCLRE1A	9	56	58	3303	3900	100	9	0

Row labels	No. of target exons	No. of probe	No. of probes repli-cation	Size of target regions (bp)	Size of se-quence regions (bp)	Per-centage of Cover-age	Ex-ons of high cov-er-age	Ex-ons of lo w cov-er-age
DCLRE1B	4	27	33	1679	1860	100	4	0
HELQ	18	63	72	3666	4860	100	18	0
OBFC2B	6	14	22	756	1188	100	6	0
PRPF19	16	38	61	1835	3221	100	16	0
RDM1	7	19	22	1066	1560	100	7	0
RECQL	14	39	45	2230	3163	100	14	0
RECQL5	22	68	175	3634	5348	100	22	0
RPA4	1	13	13	806	840	100	1	0
PARP enzymes that bind to DNA	52	142	218	7782	11547	-	51	1
PARP1	25	69	88	3828	5580	98	24	1
PARP2	16	41	55	2106	3390	100	16	0
PARP3	11	32	75	1848	2577	100	11	0
Repair of DNA-topoisomerase crosslinks	25	70	96	3642	5700	-	25	0
TDP1	17	46	47	2226	3780	100	17	0
TDP2	8	24	49	1416	1920	100	8	0
Ubiquitination and modification	128	372	533	20402	29678	-	125	3
HLTF	28	70	92	3710	5854	100	28	0
RAD18	14	35	51	1872	2940	100	14	0
RNF168	6	31	31	1836	2220	100	6	0
RNF4	10	21	23	889	1719	96	9	1
RNF8	11	37	67	1898	2845	100	11	0
SHPRH	32	104	112	6060	8160	100	32	0
SPRTN	5	26	33	1605	1860	100	5	0
UBE2A	6	13	39	579	1140	100	6	0
UBE2B	7	13	24	675	1140	86	6	1
UBE2N	4	11	26	539	900	100	4	0
UBE2V2	5	11	35	739	900	76	4	1
Total (194 Genes)	2786	9002	15735	499047	702619	-	2778	8