University of Southampton

Faculty of Medicine

Human Development and Health

Characterisation of the genomic landscape in splenic marginal zone lymphoma

by

Carolina Jaramillo Oquendo

Thesis for the degree of Doctor of Philosophy

September 2021

University of Southampton

Abstract

Faculty of Medicine

Human Development and Health

Doctor of Philosophy

Characterisation of the genomic landscape of splenic marginal zone lymphoma

by

Carolina Jaramillo Oquendo

Background: Somatic gene mutations can alter protein function, drive carcinogenesis and aid in the risk-adapted stratification of cancer patients. Splenic Marginal Zone Lymphoma (SMZL) is an indolent B-cell lymphoma comprising less than 2% of lymphoid neoplasms. Approximately 70% of patients develop a progressive disease requiring treatment whilst 30% of these will ultimately transform to a more aggressive lymphoma. There are currently no biomarkers recommended for establishing diagnosis, assessing prognosis, or determining the choice of therapy. This is in part due to superficial understanding of the molecular pathogenesis and heterogeneity of the disease.

Aims: The main aim of this study is to construct a detailed characterisation of the genetic landscape of SMZL through the identification of somatic variants in unmatched tumour samples in the largest SMZL cohort to date and explore their clinical significance by integrating relevant clinical data. In conjunction to the analysis of somatic variants, an important part of this project also centres around the bioinformatics processing and optimisation of pipelines to obtain the best sequencing results.

Methods: Tumour samples were sequenced using an amplicon-based approach consisting of 57 target genes. Paired end reads were aligned using BWA-mem to the hg38 reference genome and LocatIt was used to merge duplicate reads using unique molecular identifiers. Afterward, GATK's haplotype caller was used for variant calling and Annovar software for annotation. Variants were filtered using an unsupervised machine learning algorithm and validated *in-silico* using a genome viewer. Subsequently, variants were filtered once more to reduce likely germline variants. Finally, additional clinical and genetic data was integrated with the curated variant list to correlate genomic results with clinical outcomes.

Results: In concordance with the literature *NOTCH2* [13%], *TP53* [12%] and *KLF2* [12%] were found to be recurrently mutated among SMZL patients. As well as validating previous observations, key findings within this work included: 1) Genes *KMT2D* and *CCND3* were found mutated in a much higher number of cases than was expected; 2) *KLF2* and *CCND3* harbour

mutation hotspots which require functional validation but are predicted to affect protein function; 3) Evidence of somatic hypermutation (SHM) was found in the majority of cases, (only 8% showed no evidence of SHM); 4) Deletions of 7q were associated to *IGHV1-2*04* usage, *KLF2* and *NOTCH2* mutations, short telomeres, and low levels of SHM; 5) Identification of two potential genomic subgroups, one group characterised by 7q deletions, *KLF2* and *NOTCH2* mutations and *IGHV1-2*04* usage and a second group characterised by *MYD88* mutations and mutated *IGHV* genes and; 6) Identification of telomere length and gains of 3q and 8q as new potential prognostic factors.

Conclusion: This project collects the largest cohort of SMZL cases assessed to date imparting clarity to the genetic landscape of this cancer. The data supports distinct sub-groups of SMZL driven by *IGHV* usage and consistent genomic lesions. Additional studies across multiple discovery and validation cohorts, as well as prospective clinical trials are required to validate results, particularly disease outcomes.

Table of Contents

Table of Tables

Table of Figures

List of Accompanying Materials

Supplementary Tables

Supplementary Figures

Research Thesis: Declaration of Authorship

Print name: Carolina Jaramillo Oquendo

Title of thesis: Characterisation of the genomic landscape of splenic marginal zone lymphoma

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Jaramillo Oquendo, C., Parker, H., Oscier, D. *et al.* Systematic Review of Somatic Mutations in Splenic Marginal Zone Lymphoma. *Sci Rep* **9,** 10444 (2019). https://doi.org/10.1038/s41598-019-46906-1

   Oquendo CJ, Parker H, Oscier D, Ennis S, Gibson J, Strefford JC. The (epi)genomic landscape of splenic marginal zone lymphoma, biological implications, clinical utility, and future questions. J Transl Genet Genom 2021;5:89-111. https://dx.doi.org/10.20517/jtgg.2021.04

Signature: ..... Date: .....

Acknowledgements

I would like to thank my supervisors Professor Sarah Ennis, Dr Jane Gibson and Professor Jon Strefford for all of the support, guidance and time they have given me over the past four years. I have learnt so much during my PhD and could not have asked for better mentors. I would also like to thank all of my colleagues in both the Genomic Informatics and Cancer Genomics group for all their help and support specially Dr Helen Parker and Dr Dean Bryant for always helping me out and answering all my questions.

I would also like to acknowledge all the patients and clinicians who contributed clinical material and information for this study and my funder Colciencias.

Throughout this project I have had amazing support from friends and family, and I would have not survived had it not been for them:

Sam: Thank you for always being there for me, making me food whenever I was too busy, always cheering me on, and for the infinite hugs, I definitely needed them.

Lara, Clare and Imogen: The best PhD buddies! You guys made my PhD so much more enjoyable and it would not have been the same without you.

Mamá y hermano: Este proyecto se los dedico a ustedes. Espero poder volvernos a ver pronto para poder celebrar este nuevo logro. Gracias por su amor incondicional y su apoyo desde la distancia. Todo lo que hago es por ustedes y espero que estén orgullosos.

Definitions and Abbreviations

Abbreviation	Definition
AID	Activation induced-cytidine deaminase
APC	Anaphase promoting complex
BAM	Binary sequence Aligment/Map format
BCR	B-cell receptor
BWA	Burrows wheeler alignment
C	Constant genes
CBL-MZ	Clonal B-cell lymphocytosis of marginal zone origin
Cdc25	Cell division cycle 25
Cdks	Cyclin dependent protein kinases
CHG	Comparative genomic hybridisation
CLL	Chronic lymphocytic leukaemia
COSMIC	Catalogue of somatic mutations in cancer
CSR	Class switch recombination
D	Diversity genes
ddATP	2',3'-dideoxyadenosine triphosphate
ddCTP	2',3'-dideoxycytidine triphosphate
ddGTP	2',3'-dideoxyguanosine triphosphate
ddNTP	2',3'-dideoxynucleotide triphosphate
ddTTP	2',3'-dideoxythymidine triphosphate
DISCOVER	Discrete independence statistic controlling for observations with varying event rates
DLBCL	Diffuse large B-cell lymphoma
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
ENMZL	Extranodal marginal zone lymphoma
FDR	False discovery rate
FISH	Fluorescence in-situ hybridisation
FITC	Flourescein-5-thiocyanate
FL	Follicular lymphoma
FO	follicular B-cells
G0 phase	Non-proliferative state within the cell cycle
G1 phase	First growth phase of the cell cycle
G2 phase	Second growth phase of the cell cycle
GC	Germinal centres
HCLv	Hairy cell leukaemia variant
HCV	Hepatitis C
HGMD	Human Gene Mutation Database
HTS	High throughput sequencing
ID	Inhibitory domain
Ig	Immunoglobulin
IGHV	immunoglobulin heavy chain variable region
IGV	The Integrative Genomics Viewer

Abbreviation	Definition
J	Joining genes
LPL	Lymphoplasmacytic lymphoma
M phase	Last phase of the cell cycle (mitosis)
MANE	Matched annotation between NCBI and EBI
MCL	Mantel cell lymphoma
ML	Machine learning
MSC	Mutation significance cut-off
MZ	Marginal zone
MZL	Marginal zone lymphoma
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next generation sequencing
NHL	Non-Hodgkins lymphoma
NICD	Intracellular domain of the notch protein
NICD	Notch intracellular domain
NMZL	Nodal marginal zone lymphoma
NRD	Non-catalytic region domain
PB	Peripheral blood
PC1	Principal component 1
PC2	Principal component 2
PCA	Principal component analysis
PCR	Polymerase chain reaction
PEST	sequence that is rich in proline (P), glutamic acid (E), serine (S), and threonine (T)
PRC2	Polycomb repressor complex 2
QD	Quality normalised by depth
RB	Retinoblastoma proteins
S phase	DNA replication phase of the cell cycle
SAC	Spindle assembly checkpoint
SAM	Sequence Aligment/Map format
SDRPL	Splenic diffuse red pulp small B-cell lymphoma
SMZL	Splenic marginal zone lymphoma
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SNV	Single nucleotide variants
SWI/SNF	SWItch/Sucrose Non-Fermentable
TGF-β	Transforming growth factor–β
TIR	Toll/interleukin-1 receptor
TLR	Toll like receptors
TRIAL	tumour necrosis factor-related apoptosis-inducing ligand
TSL	Transcript support level
UMAP	Uniform manifold approximation and projection dimension reduction
UMB	Unique molecular barcode
V	Variable genes

Abbreviation	Definition
VAF	Variant allele frequency
VCF	Variant calling file format
VEP	Variant effect predictor
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization
WM	Waldenström's macroglobulinemia

Chapter 1 **Introduction**

1.1 Cell development and regulation

In general terms, cancer can be defined as an uncontrolled cell proliferation caused by changes in the DNA sequence as well as changes in gene expression. These changes lead to a population of cancerous cells that can invade other tissues and metastasize causing significant morbidity and death[1]. In theory, all cell types can become cancerous, so there are as many types of cancers as there are cell types in the human body.

Understanding normal cell development and the control mechanisms of cell growth and proliferation is crucial in comprehending cancer development. The main objective of the cell cycle is to pass down genetic information from a mother cell to two identical daughter cells. The cell cycle begins with a growth phase (G_1 phase), where cells begin to prepare for division by increasing in size and monitoring their environment for the presence of growth factors and mitogens. This initial growth phase is then followed by DNA replication (S phase). On occasion, some cells, such as normal liver cells, do not enter the S phase and instead fall into a non-proliferative state called G_0. G_0 is a temporary withdrawal from the cell cycle but cells can be stimulated to go into S phase if needed. If the cells continue into DNA replication, another growth stage (G_2 phase) will follow. Once cells are completely prepared to divide, they enter the shortest stage, mitosis (M phase). During mitosis, cells almost double their size and go through a nuclear division followed by a cytoplasmic division[2] .

To avoid uncontrolled cell growth and proliferation, cells rely on various molecular mechanisms to control these processes. One such mechanism is the regulatory phosphorylation of key protein and protein complexes in the cell cycle, which in turn act as checkpoints to determine whether the cell will continue onto the next stage. The phosphorylation reactions are carried out by kinases while the dephosphorylation by phosphatases. The main instigators of cell activity are the kinases, which themselves depend on another protein, cyclin, to become active. Hence, the protein complexes formed by kinases are known as cyclin-dependent protein kinases or Cdks[2]. **Figure 1-1** illustrates a summary of the cell cycle including the major checkpoints in its regulation.

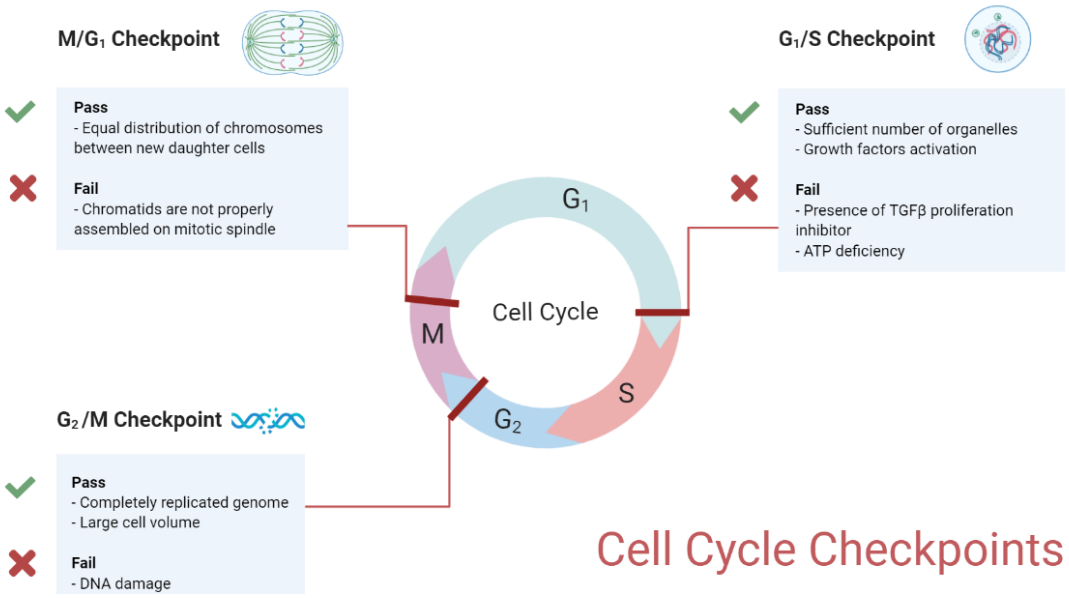


Figure 1-1. The cell cycle and its major regulatory checkpoints.The arrows show the direction in which the cycle progresses. The red bars represent the three major regulatory checkpoints. Figure created in bioRender.com.

The cell cycle has three major checkpoints shown in **Figure 1-1**. In the G1/S phase checkpoint, Cdk inhibitor proteins, such as transforming growth factor–β (TGF-β), block entry to S phase by blocking the assembly or activity of the Cdk complexes needed. DNA damage can also prevent the progression from G_1 to S phase[3]. Similarly, during the G_2/M checkpoint, if DNA is damaged or DNA replication is incomplete, the cell inhibits the cell division cycle 25 (Cdc25) phosphatase required to activate the mitosis cyclin dependent kinases (M-Cdks)[4]. In the last major checkpoint, the M/G_1 checkpoint or spindle assembly checkpoint (SAC), chromosome segregation is delayed until all kinetochores are attached to microtubules[5]. In this last checkpoint, the cell inhibits the activation of the anaphase-promoting complex (APC), which tags the cyclins in the M-Cdks with proteasomes (ubiquitin) that in turn break down the cyclins and inactivates the M-Cdks, causing the cell to exit mitosis[2]. Other cycle controls include transcription regulators such as p53, which initiates the transcription gene for Cdk inhibitor protein p21. The protein p53 is activated when there is DNA damage, orchestrating a variety of DNA damage response mechanisms or if the damage is too great it initiates programmed cell death or apoptosis[6]. The cell also has means of regulating cell growth and apoptosis through extra-cellular signals such as survival factors (suppress apoptosis), mitogens (stimulate cell division) and growth factors[2]. Regardless of the cells' tight control on proliferation, they can bypass one of its checkpoints and start proliferating in an uncontrolled manner, consequently, leading to cancer development.

1.2 Cancer development

1.2.1 Hallmarks of cancer

All cells in an organism are descendants of a progenitor cell that goes through consequent cell divisions, whether through mitosis in somatic cells or meiosis in germ cells. Throughout its lifespan, a cell can acquire sporadic DNA changes (variants) that eventually accumulate and propagate. Depending on the effect of the DNA change, this accumulation of acquired variants can lead to cancer development. On a molecular level, Hanahan and Weinberg[7] postulate that normal cells acquire certain hallmark capabilities as they progressively evolve into a malignant state:

Sustain proliferative signalling: This is probably the most distinctive trait of cancer, which is the ability of cancer cells to deregulate normal growth-promoting signals and proliferate without any limitation.

Evade growth suppressors: Cancer cells can avoid programs that negatively regulate cell proliferation, usually by deregulation of tumour suppressing genes such as those that truncate the function of the TP53 and RB proteins (tumour suppressor proteins).

Resist cell death: Programmed cell death or apoptosis is a vital component of various cell processes as a homeostatic mechanism. In tumours that transform into high-grade malignancies, apoptosis is attenuated, often triggered by insufficient survival factor signalling or hyperactive signalling from oncoproteins.

Enable replicative mortality: Unlike normal cells, which have a limited number of growth and cell division cycles, cancer cells can evade this threshold and acquire unlimited replicative potential.

Induce angiogenesis: For any cell to grow and develop, normal or cancerous, it must have enough sustenance (e.g. nutrients and oxygen) to continue to thrive. To have a constant intake of nutrients, cancer cells induce normally quiescent vasculature to sprout new vessels that aid their development.

Activate invasion and metastasis: This refers to the ability of cells to alter genes that encode cell-cell and cell-extracellular matrix interactions. This triggers an invasion-metastasis cascade allowing cancer cells to expand to other tissues.

1.2.2 Genomic variation in cancer cells

In cancerous cells, changes in the DNA can be classified as either somatic or germline. Germline changes are those that are present in all cells of the individual originating via the germ cells from previous generations and are inherited. On the other hand, somatic changes are those that have been acquired during a person's lifetime. Although many genomic studies focus solely on somatic variation, in some cases, germline variants may lead to an increased risk of developing cancer[8–10] and it might be pertinent to study these mutations as well.

The origin of somatic variation depends on each cell, but these can have an intrinsic origin, such as errors in DNA repair, or an extrinsic/epigenetic origin such as exposure to mutagenic agents. **Figure 1-2** illustrates the major DNA alterations that can arise within cancer genomes. The changes in the DNA may confer either a gain or loss of function in genes that usually encode proteins that stimulate cell division or inhibit cell differentiation and stop cell death.

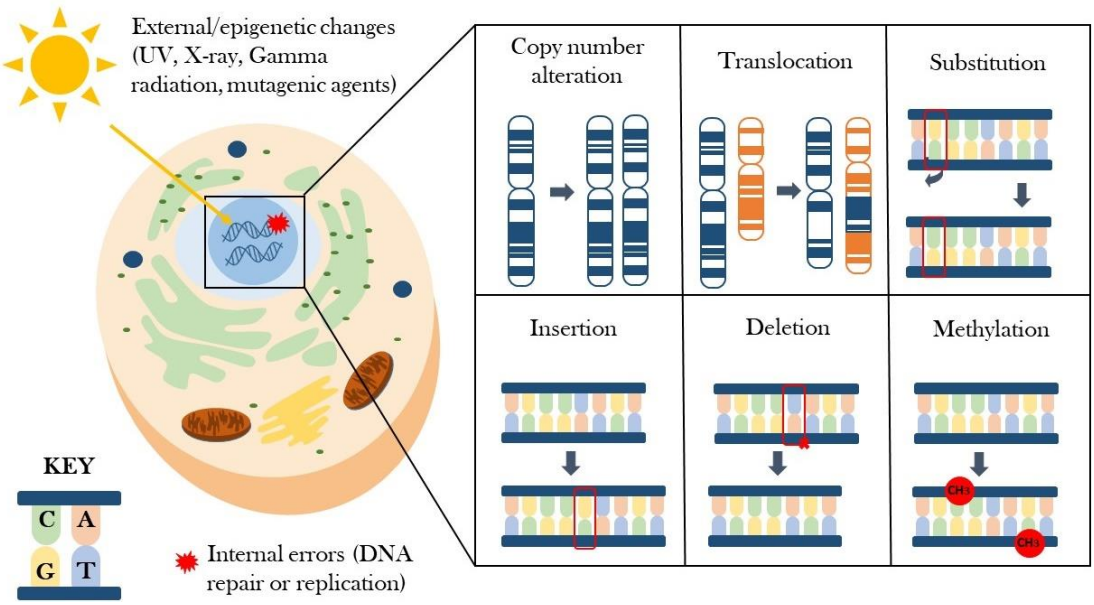


Figure 1-2. Major DNA alterations found within cancer genomes. This figure illustrates the six major DNA changes that can develop in cancer cells (copy number alterations, translocation, substitutions, insertions, deletions and methylation). These changes may arise as a result of external or internal factors such as exposure to different mutagenic sources, such as UV radiation, or errors in DNA repair or replication.

Changes in the DNA of cancer cells typically happen within two classes of genes, oncogenes or tumour suppressor genes. Oncogenes are those whose overexpression can cause cells to develop into cancer cells[3]. These tend to need one copy of the mutated gene to drive the cell into a malignant state and are often juxtaposed with enhancer elements, most likely as a consequence of gene fusions. Gene fusions result from chromosomal rearrangements (translocations, inversions, deletions and amplifications) where a chimeric protein is formed or there is a deregulation of genes due to proximity of a novel promoter or enhancer region[11].

Tumour-suppressor genes are those that in normal conditions will inhibit cancerous behaviour and whose inactivation drives the cell towards cancerous transformation by losing their function. Usually, both copies of the tumour suppressing gene need to be lost for cancer to develop, hence the DNA change will act in a recessive manner[2]. Tumour suppressor genes are further classified into gatekeeper and caretaker genes. Gatekeeper genes are those which regulate cell division, death/lifespan, while caretaker gens are those in charge of maintaining genetic stability or DNA repair[12].

1.2.3 Drivers of cancer

Cancer development is characterised by the accumulation of somatic variation within a cell, however, not all changes will lead to oncogenesis. Somatic variants can be classified as a driver or passenger, according to their consequences. Driver mutations or variants are those that confer a survival advantage to the cell and therefore "drive" the cell into cancer development, while passengers are all of the other variants that do not give the cell a growth advantage[8]. The number of drivers present in each cancer differs according to the type of cancer and affected genes. Exposure to mutagens can also affect the number of drivers, such as smokers versus non-smokers in lung cancer[13]. Cancers with mutations affecting DNA repair genes are also likely to have a higher number of driver mutations[14,15]. However, saying that a variant will either confer a survival advantage or it will not, simplifies a very complex interaction between the tumour and its environment. The effects of driver mutations likely lie on a continuum[16] and it is the different synergy between variants and many other factors that may provide a selective advantage. Furthermore, being able to identify which somatic variation will actually provide a selective advantage remains a difficult task[17].

1.3 Mature B-cell malignancies

This project focuses on splenic marginal zone lymphoma (SMZL) a mature B-cell malignancy. Mature B-cell malignancies are a heterogeneous group of diseases often with a germinal centre origin (more details in section **1.3.1**) that arise during different stages of B-cell differentiation[18]. There are more than 40 types of mature B cell lymphomas and according to Cancer Research UK, Non-Hodgkins lymphomas (NHLs) represent the sixth most common cancer in the UK. NHLs are a group of lymphomas that arise from lymphocytes all of which have a varied prognosis, therapy and goals of therapy[19]. Other leukemic variants such as chronic lymphocytic leukaemia (CLL) are also included in the mature B-cell classification since they derive from mature B-cells[18]. The specific lymphoma subtype reflects the stage of B-cell development in which the lymphoma originated and is the major basis for its classification. In general, these malignancies can also be

classified into low, intermediate, or high-grade lymphoma or leukaemia. These grades reflect the rate of growth, where a low-grade lymphoma is slow developing (chronic) while high-grade is considered aggressive[20].

1.3.1 B-cells and B-cell receptors

B-cells are a type of lymphocyte in the immune system that synthesise immunoglobulin (Ig) and display it on the cell surface as a receptor (BCR). Each B-cell possesses a unique receptor and the diversity of these cells is such, that B-cells can recognise all foreign molecules or substances (antigens) in our environment. BCRs are encoded by several genes, which rearrange during the early stages of B-cell development. Each B-cell will have a particular BCR rearrangement with greater than 26 million potential BCR binding combinations[21]. This creates millions of unique receptors, in turn creating millions of B-cells, all with different specificities[22].

Immunoglobulins or BCRs are "Y" shaped proteins composed of two identical light chains (κ and λ) connected to two identical heavy chains by a disulphide bond. Clusters of genes encoding the light chains are located on chromosome 2 and 22 and clusters of genes encoding the heavy chain on chromosome 14. The cluster of light chain genes includes a series of variable (V), joining (J), and constant (C) genes. While the cluster of heavy chain genes includes the same as the light chain (V, J and C) as well as diversity genes (D)[21]. The function of the BCR is to recognise and bind to antigens via the variable regions exposed on the cell surface and activate the B-cell leading to clonal expansion and antibody production[23].

1.3.2 B-cell development

As mentioned earlier, B-cell malignancies arise during various stages of B-cell differentiation, which in turn reflect the stages of immunoglobulin heavy and light chain rearrangement and surface expression that cell has gone through. B-cell development begins in the bone marrow where progenitor B-cells go through an Immunoglobulin gene rearrangement and develop into naive B-cells via pre-B-cells and immature B-cells (**Figure 1-3**). Before leaving the bone marrow, immature B-cells are tested for autoreactivity, and those that have no strong reactivity to self-antigens are allowed to mature. Cells then leave the bone marrow and migrate to the spleen and other lymphoid tissues. Within the peripheral lymphoid tissue, B-cells are present in loose aggregates (primary follicles) or in well-defined proliferating foci called germinal centres (GC) consisting of a dark and light zone[22]. In the peripheral lymphoid tissue naive cells will mature into follicular (FO) or mature B-cells. If the mature B-cell encounters an antigen that fits it surface Ig receptors it will then go through a rapid proliferation to mature into antibody-secreting plasma cells and memory B-cells[19] in the GC.

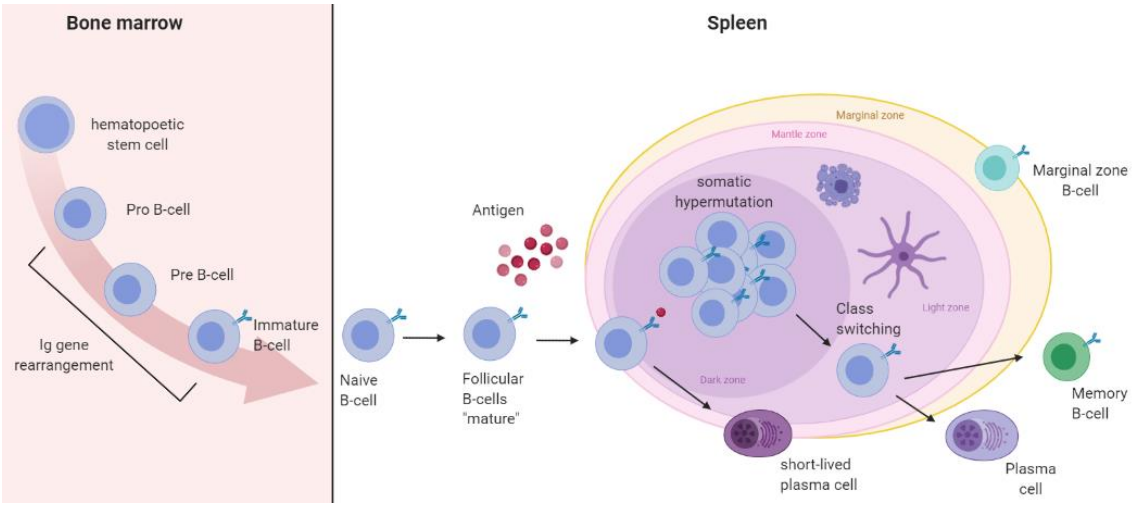


Figure 1-3. Overview of B-Cell differentiation. B-cell development begins in the bone marrow where progenitor B-cells go through an Ig gene rearrangement and develop into naive B-cells via pre-B-cells and immature B-cells. In the peripheral lymphoid tissue, follicular B-cells will encounter an antigen that fits their surface Ig receptors and it will mature these into antibody-secreting plasma cells and memory B-cells. The cells undergo rapid proliferation in the germinal centre, where those cells that have mutations resulting in better binding to the epitope are stimulated to proliferate and dominate the immune response[21]. Figure created in BioRender.com.

In an immune response, surface receptors play a key role in the activation of leukocytes, where the effectiveness of the interaction between the receptor (immunoglobulin) and ligand (antigen) will depend on the receptor's affinity for this ligand[22]. In mature B-cells, upon subsequent epitope (part of the antigen to which the receptor attaches) exposure, cells undergo extensive proliferation, somatic hypermutation, immunoglobulin isotype switching and antigen-affinity driven selection[24]. The early stages of the GC reaction are carried out in the dark zone where B-cells will undergo rapid proliferation and accumulate small point DNA mutations in the Ig heavy and light chain variable region genes to change their affinity to the antigen, this is known as somatic hypermutation[21]. Cells that have mutations resulting in better binding are stimulated to proliferate and dominate the immune response. Afterwards, cells will migrate into the light zone where they can encounter one of three fates: apoptosis if BCR affinity is too low, re-entre the dark zone for further proliferation and somatic hypermutation, or exit the GC and differentiate into plasma or memory B cells. This selection process is called affinity maturation or antigen-affinity driven selection and is followed by an immunoglobulin isotype switch or class switch recombination (CSR). In CSR changes to the constant region of the heavy chain locus of the BCR are made. This switches the class of the BCR (e.g. IgM to IgA) allowing the cell to interact with different effector molecules without changing the affinity of the BCR.

1.3.3 Splenic marginal zone B-cells

When talking about B-cell differentiation, a population of B-cells often excluded from this discussion are marginal zone (MZ) B-cells. In humans, MZ B-cells are mostly found in the marginal

zone of the spleen, but can also be found circulating in the blood[24,25]. However, MZ B-cells are not the only cells present in the splenic MZ. The splenic MZ also contains macrophages, dendritic cells, granulocytes and even passing memory cells[26] making the process of purifying and studying these cells difficult.

MZ B-cells are at the boundary between the innate and adaptive immune system and are viewed as a population of cells capable of inducing antibody response to both T dependent and independent antigens. This notion is reinforced by their location, since they are located in a place highly exposed to antigens and can, therefore, respond quickly to blood-borne pathogens[25,26]. There are many unknows surrounding the development of MZ B-cells, but their diverse function could imply the existence of MZ B-cell subgroups which develop separately. Some researchers believe that MZ B-cells could be the result of IgM memory B-cells that have exited the germinal centre (GC) reaction before isotype switch, evidenced by their phenotype and the somatic mutations in the immunoglobulin genes found in those cells[26–28]. Others think that before immature naïve cells develop into mature naïve cells (follicular B-cells) there is a transition step where these cells will differentiate into either MZ B-cells or follicular B-cells promoted by BCR and NOTCH signalling [24]. Descartoire et al. gave evidence of a NOTCH2 dependent MZ B-cell precursor in murine models, which favours the idea of a separate MZ B-cell lineage[29]. Furthermore, results from Weller et al. support the notion that MZ B-cells develop and mutate during the first years of life without being engaged in either a T-dependent or independent immune response[30].

1.4 Clinical phenotype of splenic marginal zone lymphoma (SMZL)

The World Health Organization (WHO) classification of tumours of the hematopoietic and lymphoid tissues defines three marginal zone lymphoma (MZL) entities, splenic marginal zone lymphoma (SMZL), nodal MZL (NMZL) and extranodal MZL (ENMZL)[31]. In addition, a number of provisional entities are emerging; these include splenic diffuse red pulp lymphoma (SDRPL), hairy cell leukaemia-variant (HCLv) and clonal B-cell lymphocytosis of MZ origin (CBL-MZ), the latter of which is clonally related to SMZL in a proportion of cases[32–35]. SMZL is a rare, low grade lymphoma involving the spleen, bone marrow and peripheral blood (PB) that comprises less than 2% of lymphoid neoplasms. Patients present with abdominal discomfort, splenomegaly, anaemia, villous lymphocytes or incidentally due to abnormal blood count. Median age of diagnosis is 65 and patients exhibit a 10-year median survival time[36,37]. Whilst a significant proportion of patients will exhibit a more indolent disease course, approximately 70% will require treatment, 30% of those will develop aggressive symptoms, whilst 5-15% will have a disease that transforms to diffuse large b-cell lymphoma (DLBCL) with dismal survival[38]. Treatment options include splenectomy, chemotherapy, immunotherapy (anti-CD20 monoclonal antibody rituximab), or immunotherapy

combined with chemotherapy[39,40]. Some SMZL cases are associated with hepatitis C (HCV) infection, and regression can be achieved with antiviral therapy[39]. Diagnosis of SMZL can be established through a combination of lymphocyte morphology and flow cytometry, bone marrow biopsy and immunohistochemistry[19,37]. Unfortunately, several other mature B-cell tumours, have overlapping clinicopathological and immunophenotypic features with SMZL, making for a challenging accurate diagnosis. In some cases it can be difficult to distinguish SMZL from HCLv and SDRPL without the use of spleen histology[38,40]. Despite a large number of studies, mechanistic understanding behind unfavourable outcome and transformation is still unknown[41] and treatment outcomes are variable with few prognostic markers that can aid in targeted treatment[40].

SMZL shares the CD27+IgM+IgD+ immunophenotype of human marginal zone B-cells as well as a similar somatic hypermutation status in the immunoglobulin heavy chain variable region (*IGHV)* genes[19,28]. According to the WHO classification, the normal SMZL counterpart is a B-cell of unknown differentiation stage and it suggests that SMZL could be a post germinal centre neoplasm, as approximately 50% of SMZL cases display *IGHV* gene somatic hypermutation[19]. Some studies have suggested that the other 50% which lack the *IGHV* gene somatic hypermutations may have a pre-germinal centre origin[42,43]. **Figure 1-4** shows different types of B-cell malignancies that can arise during B-cell development, keeping in mind that in SMZL the cell of origin is still under debate[44].

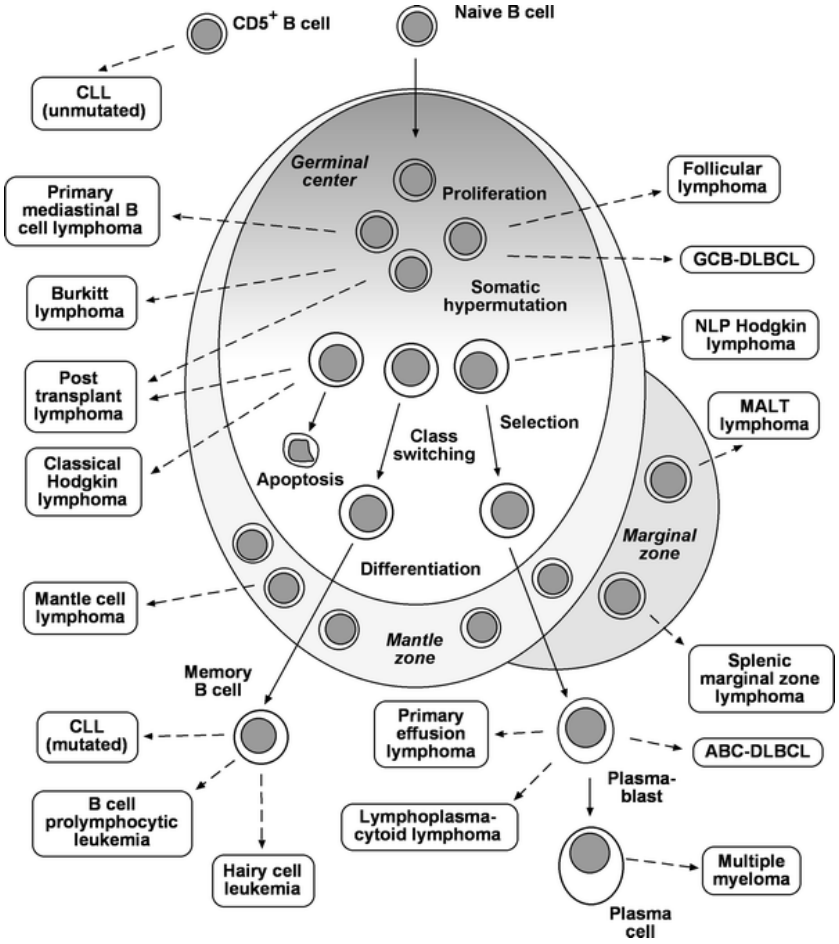


Figure 1-4. Origin of different B-cell malignancies. This figure shows the stages of B-cell development after a cell has left the bone marrow and the different types of lymphomas and leukaemia that can arise from each stage. Although SMZL is shown to be a post-germinal centre neoplasms studies suggest that approximately half of the cases have a naïve pre-germinal centre origin. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Methods in Molecular Biology. Origin and Pathogenesis of B Cell Lymphomas, Seifert M, Scholtysik R, Küppers R. Copyright © 2019.

Whilst SMZL is still a relatively under-studied malignancy, several works published over the last decade or so have begun to unravel the intrinsic molecular defects present in these cells, and extrinsic cellular mechanisms that reflect micro-environmental and antigenic interactions (**Figure 1-5**). Karyotype banding, fluorescence in situ hybridization (FISH) and comparative genomic hybridization arrays were the foundation of early SMZL studies[45–48] and much more recently high throughput sequencing (HTS) has been applied to try and de-convolute the genetic landscape of the disease[49–54]. To understand the impact that the new sequencing technology has had in the characterisation of SMZL, the next section will give a historical overview of genomic technologies and in later chapters how these have aided in the discovery of frequently mutated genes and affected pathways in SMZL.

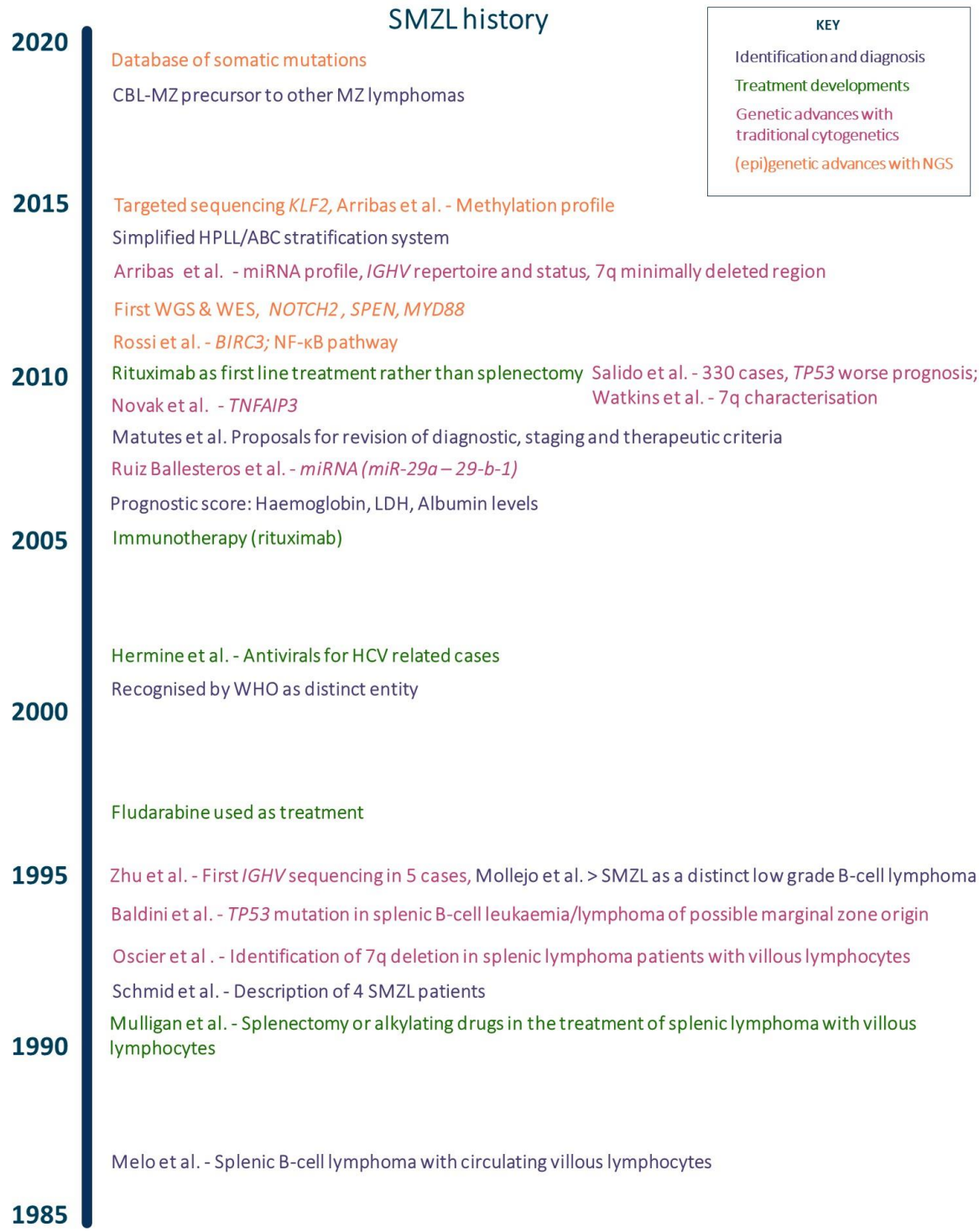


Figure 1-5. Timeline highlighting historical milestones in the understanding of SMZL. Figure by Jaramillo Oquendo et al[55] licenced under CC BY 4.0.

1.5 Historical overview of genomic technologies

1.5.1 Karyotyping

Before the development of high-throughput sequencing, karyotyping was the traditional way of analysing chromosomes. A karyotype (**Figure 1-6**) describes an individual's chromosome constitution and in principle, it can be obtained from all tissues that contain mitoses. To

karyotype, a cell culture is obtained from different tissues, usually blood as it is the most convenient. Cells are stimulated into cell division and are grown in culture medium for 72 hours. Before cells are harvested, a drug with colchicine-like effect, usually colcemid, is added to prevent spindle formation and arrest the cells in prometaphase or metaphase. To obtain chromosomes spread in one plane, cells are fixed on a slide, where it is then air-dried and stained. For better identification of each chromosome, there are different banding methods which result in distinctive banding patterns for each chromosome[56]. G banding is the most common method. It involves a trypsin treatment to digest GC rich regions, followed by Giemsa staining. This will produce a light and dark staining pattern where the regions digested by the trypsin will appear pale[57]. Karyotyping is a whole-genome approach able to detect large structural changes, such as translocations, deletions or insertions, as well as aneuploidies. Unfortunately, the resolution of this approach is dependent on the visible bands, which contain approximately 5-10 x 10^6 basepairs[56].

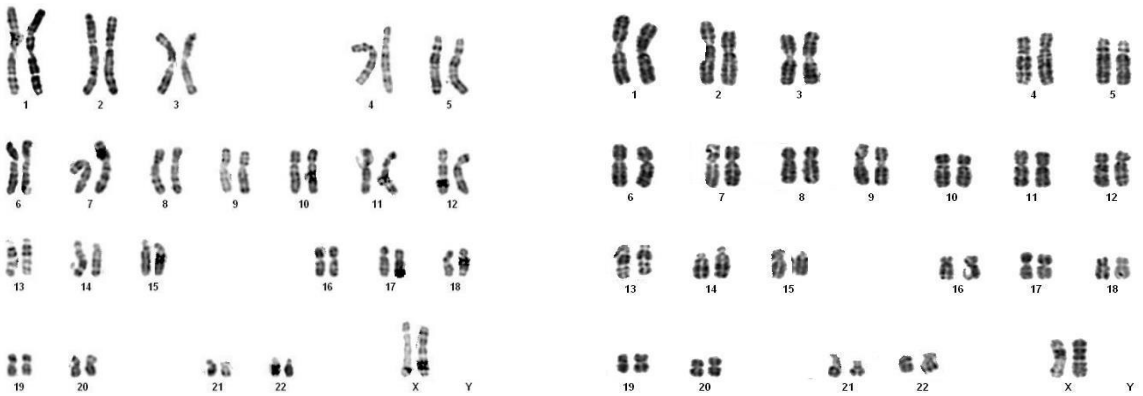


Figure 1-6. Representative human karyotype. Karyotypes obtained from metaphase cells displaying R banding.

1.5.2 Fluorescence *in situ* hybridization (FISH)

In the late 1980s, Pinkel et al. described a method of targeting certain chromosome sequences by hybridising metaphase spreads and interphase nuclei using fluorescently labelled DNA probes to detect the presence of complementary nucleic acid sequences (target sequences)[58]. This approach termed fluorescence in situ hybridisation (FISH) consists of five main steps: 1) DNA probes are prepared and labelled; 2) Metaphase chromosomes or interphase nuclei are prepared; 3) Denaturation of both probes and sample DNA; 4) In-situ hybridisation of the probes and sample and; 5) Fluorescent dye detection via ultraviolet light excitement of a fluorochrome, such as fluorescein-5-thiocyanate (FITC) or rhodamine[56]. **Figure 1-7** provides an overview of FISH protocol.

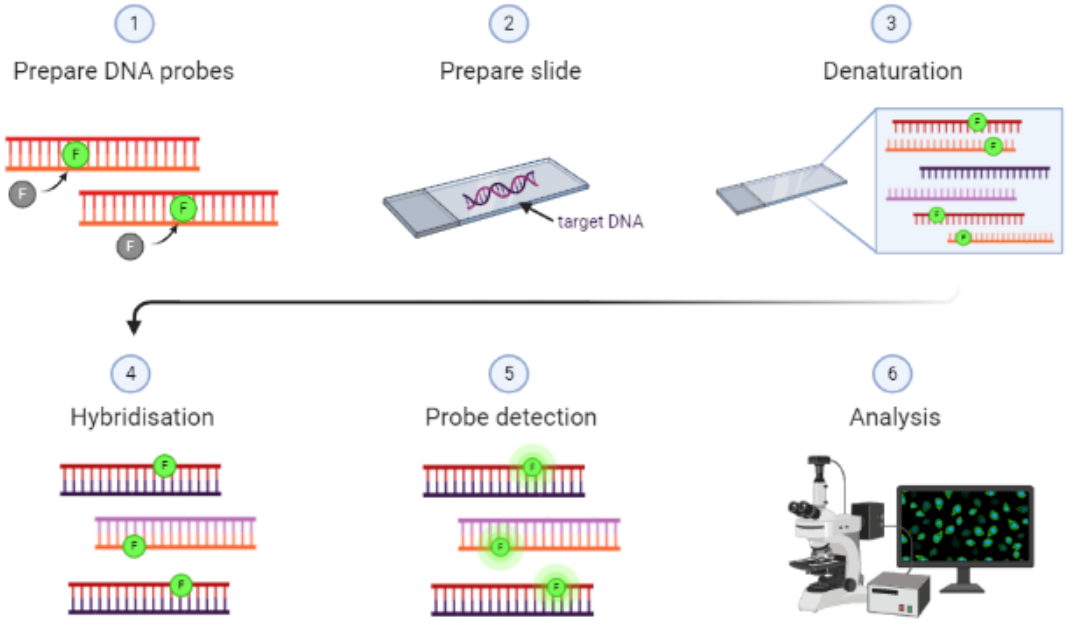


Figure 1-7. Overview of FISH protocol. The sample is in the form of a metaphase spread where previously prepared probes are added. Subsequently, there is denaturation of both the probes and DNA so these can hybridise. Once the hybridisation is complete, samples are washed and visualized under a microscope. Probes will light up if they are present in the sample. Figure created in BionRender.com

There are several types of probes all with different targets; these can be repetitive sequences, an entire chromosome, or unique sequences. This technique has a higher resolution than chromosome banding for the regions screened (gene level resolution), allowing a numerical representation of the region as well as involvement in translocations at all stages of the cell cycle[56]. Like karyotyping FISH helps determine structural changes (translocation, insertions, deletions) as well as aneuploidies but with a higher resolution (e.g. sub telomeric rearrangements, sub microscopic copy number variations, microdeletions/microduplications). However, FISH is not a whole genome approach and its disadvantages are that it can only detect known genetic aberrations and analysis is restricted to targeted regions. Both chromosome banding and FISH capture only a proportion of genetic variation in the genome as they do not capture changes at a nucleotide level.

1.5.3 Comparative genomic hybridisation arrays

Comparative genomic hybridisation (CGH) was developed to detect copy number alterations in solid tumours by comparing the DNA of a malignant cell to that of a normal cell. In the methods described by Kallioniemi et al. biotinylated total tumour DNA and digoxigenin-labelled normal genomic reference DNA are hybridised to normal metaphase spreads[59]. Once hybridised tumour DNA is detected with green-fluorescing fluorescein isothiocyanate (FITC)-avidin and normal DNA is detected with red-fluorescing rhodamine antidigoxinenin. The ratio of green to red is measured to quantify the abundance of the targeted sequences, where a high green to red ratio shows and

amplification and vice versa a high red to green ratio show deletions or chromosomal loss. A software then integrates the green and red fluorescence intensities to orthogonal strips on a chromosome axis to identify where the events are occurring[59]. However, like FISH and Karyotyping the resolution of CGHs are around 5-10 Mb.

1.5.4 Sanger Sequencing

In 1977 Sanger described a new method for determining nucleotide sequences in DNA using inhibitors to terminate replication of a template strand at one of the four bases (A, C, T or G)[60]. This method involved using a 2',3'- dideoxynucleotide (ddNTP) instead of deoxinucleotide (dNTP) during DNA synthesis, since it has an inhibitory effect on DNA polymerase. What this means is that while the DNA polymerase is synthesizing the new strand, it will terminate when it needs to incorporate the ddNTPs, as they have no 3' – hydroxyl group inhibiting further extension of the DNA chain. Four different reaction mixtures are made, each with a different ddNTP (ddATP, ddTTP, ddCTP, ddGTP), along with dNTPs, template strands (unknown DNA), DNA polymerase, and primers, where they go through various replication cycles. This creates different sized fragments each terminating where a ddNTP was incorporated. The reactions (four mixtures) are run in parallel on a gel to obtain a banding pattern that shows the distributions of the ddNTPs in the new DNA (**Figure 1-8**). This sequencing method became the stepping-stone in which subsequent sequencing methods are based upon. Today, Sanger sequencing is still used to confirm or validate variants, especially in repetitive and GC rich regions.

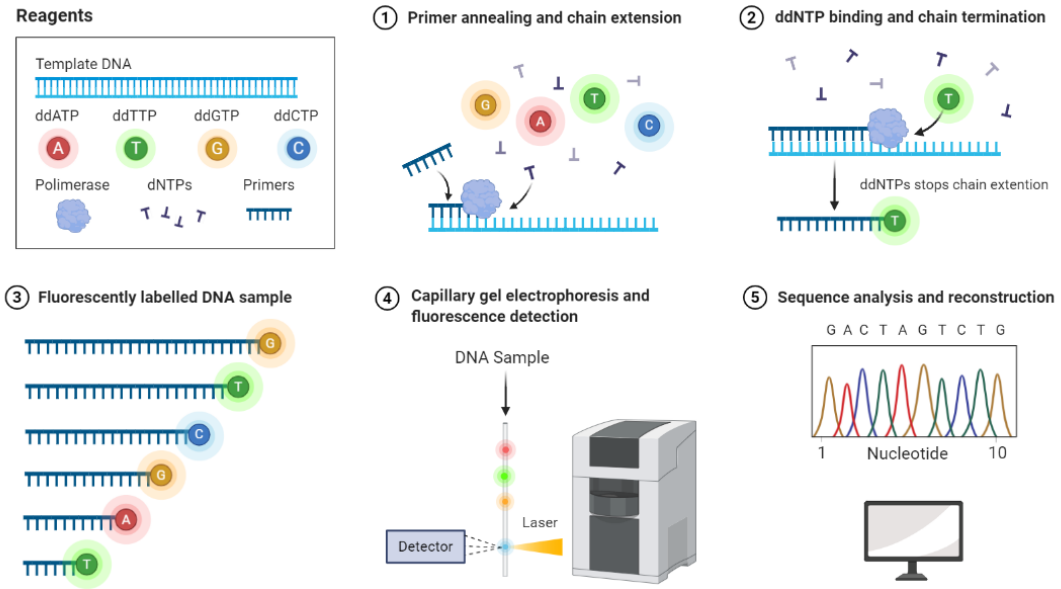


Figure 1-8. Sanger sequencing. Single stranded DNA with the unknown sequence is added to four different mixtures each containing a different ddNTP, the remaining dNTPs, primers and DNA polymerase. The mixtures will go through various replication cycles where they will create different sized fragments. Each mixture is run on a gel to detect each fragment. Once the gel is run, the sequence can be analysed. The resulting sequence will in turn be the complementary sequence of the unknown fragment. Figure created in BioRender.com.

1.5.5 High throughput sequencing

Since the development of Sanger sequencing, considered a 'first generation' technology, there have been major advances in the way sequencing is performed. Second generation sequencing often called next generation sequencing (NGS), allowed for the generation of millions of short reads (DNA sequence of fragments approximately 100 base pairs in length), at a fraction of the time and cost of Sanger sequencing[61]. NGS promised the identification of all genomic alterations, such as single nucleotide variants (SNV), insertions, deletions, copy number changes, and structural variations with a single method[62].

Illumina dominates the market of second-generation sequencing, with their sequencing by synthesis method. The sequencing by synthesis method is by far the most used, as it sequences millions of fragments in parallel by simultaneously identifying a DNA base while incorporating it into a nucleic acid chain[63]. It involves four basic steps: library preparation, cluster generation, sequencing, and data analysis (**Figure 1-9**). The library preparation is achieved by fragmenting the DNA or cDNA sample followed by specialized adapter ligation to both ends of the fragment (5' and 3'). Then, the library is inserted into a flow cell that has a lawn of complementary surface-bound oligos, where the adapters will attach. Priming occurs as the opposite end of a ligated fragment bends over and "bridges" to another complementary oligo on the surface. Repeated denaturation and extension cycles (similar to PCR) results in localized amplification of single molecules into millions of unique, clonal clusters across the flow cell. The 3' ends are blocked to prevent unwanted priming. Subsequently, sequencing begins with sequential cycles of DNA synthesis where DNA polymerase incorporates fluorescently labelled deoxyribonucleotide triphosphates (dNTPs) modified with a 3' block (reversible terminator) to avoid addition of more than one nucleotide per cycle. During each cycle, each nucleotide is identified by fluorophore excitation and the reversible terminator removed to continue the next cycle. When the sequencing is finished, the data is aligned to a reference genome, where it can be analysed to identify variations[63]

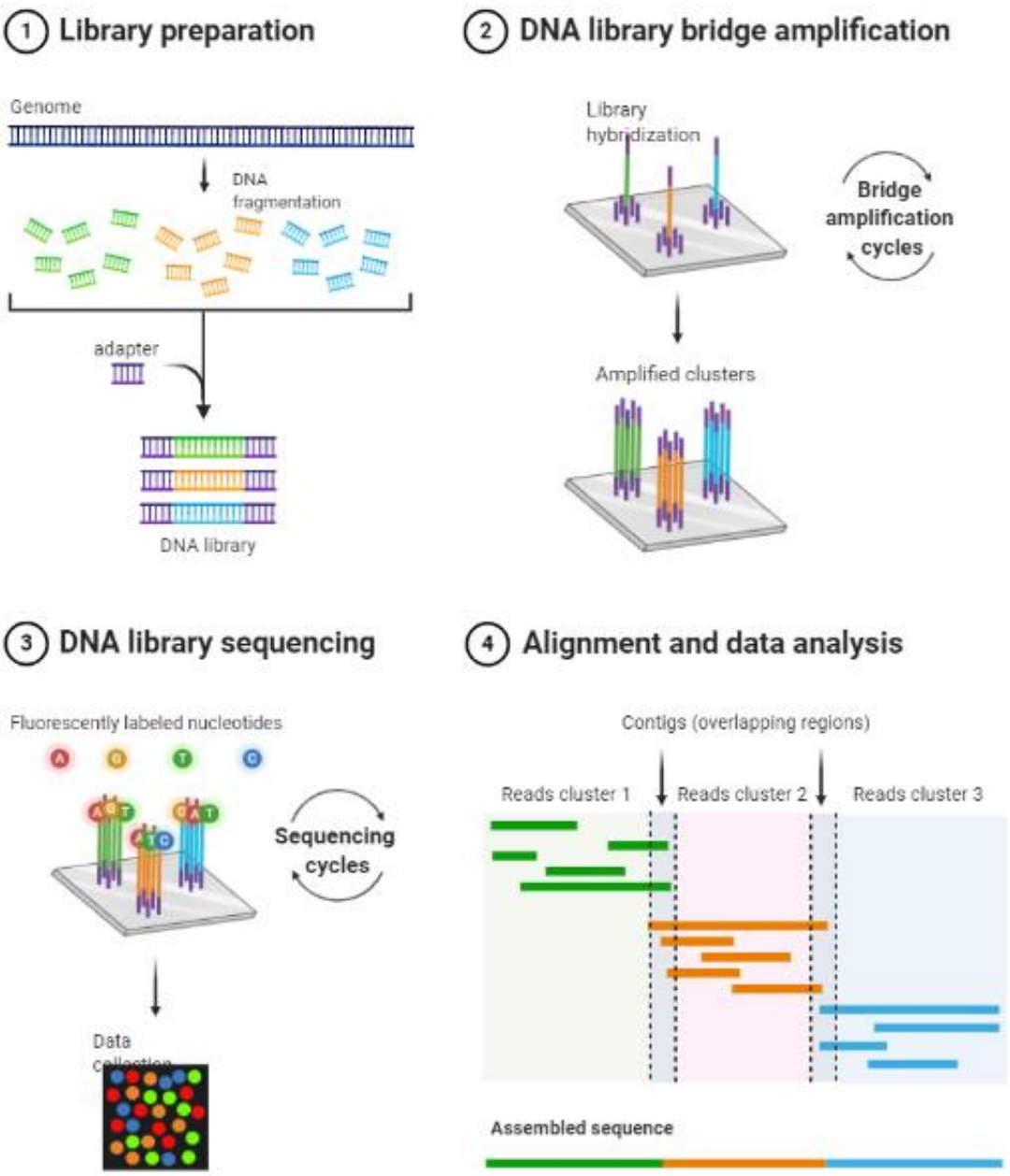


Figure 1-9. Illumina sequencing. 1. Library preparation. Genomic DNA is fragmented into small pieces followed by adapter ligation. 2. DNA bridge amplification. Library is inserted into flow cell where the adapters will attach. Priming occurs as the opposite end of a ligated fragment bends over and "bridges" to another complementary oligo on the surface. Repeated denaturation and extension cycles amplify the fragments and create clonal clusters across the flow cell. 3. Sequencing. Clusters are sequenced and during each cycle, each nucleotide is identified by fluorophore excitation and the reversible terminator removed to continue the next cycle. 4. Sequenced data is aligned to a reference genome and ready for further analysis. Figure created in BioRender.com.

The sequencing by synthesis method has three main advantages over first generation technologies. First, the DNA polymerase will not terminate the synthesis of the new DNA strand when it incorporates the modified dNTPs but rather stop momentarily, emit the fluorescent signal and continue with synthesis after the reversible terminator is removed[63]; Therefore, sequencing output is directly detected without the need for electrophoresis[64]. Second, it produces thousand-to-millions of sequencing reactions in parallel instead of hundreds, making it more cost and time effective. Lastly, second generation sequencing also allows for paired-end sequencing, which

involves sequencing both ends of the DNA fragment and then aligning them into paired reads to make a more accurate read alignment. Since the distance between the reads is known, paired end sequencing can be used to better detect insertions and deletions and map more precisely to highly repetitive regions.

Although, NGS is a huge advance in the way DNA is processed and sequenced it is not without its issues. Its main limitation is the considerably shorter read length (~100bp) compared to first generation sequencing (~800bp). Although paired end reads are helpful in this aspect, it means that: 1) highly repetitive regions are challenging to map and therefore uninformative and 2) large structural variation is difficult to identify. A second limitation is the lower read quality per base. A high read depth (number of DNA fragments that cover a specific region of the target) compensates for the lower quality in this type of sequencing. However, obtaining high read depth can be costly and Sanger sequencing remains a good alternative to identify high confidence variants.

The latest sequencing technology is referred to as third-generation sequencing, which generates long reads of over 10,000 base pairs in length[65]. The advantages of third-generation sequencing include improved analysis of structural variation and GC rich and repetitive regions of the genome. It also allows for a uniform coverage of the genome, as it is not as sensitive to GC content, and for a long-range characterisation of methylation patterns. Although third generation sequencing fills in the gaps that are left with second generation sequencing, this technology is still under development. This project does not make use of third generation technology; therefore, the next two sections will focus entirely on second generation (short read) sequencing.

1.5.6 Whole genome sequencing as a discovery approach

Whole genome sequencing (WGS), as the name implies, provides insight into the entire genome and it allows for an unbiased approach to finding targets of disease and identification of most genomic alterations. WGS is helpful when affected genes are unknown, when we want to interrogate non-coding regions of the genome or when we want to identify structural or copy number alterations.

In cancer studies, WGS is limiting in terms of depth and is mostly used as a discovery approach, followed by targeted sequencing to further examine genes of interest at higher depths. Sequencing an entire genome at an accessible price was only possible after the development of NGS. However, the cost of sequencing an entire genome at very high read depths can be prohibitive, considering that a genome with a read depth between 30X-50X costs around $1,000[66]

to sequence. Although 30X might be a reasonable read depth, to get a better understanding of the sub-clonal diversity in a tumour, deep sequencing (depth >100X) is necessary.

Furthermore, high throughput sequencing produces millions of reads in one sequencing run, resulting in vast amounts of data that need storing. Raw data from a single genome takes up approximately 200 GB of storage, equivalent to an average laptop's hard drive[67]. This can become a limitation as the amount of computational processing power and storage needs to be considerably large to process a single genome, let alone multiple samples.

1.5.7 Targeted Sequencing

To balance the cost and read depth required to study tumours, targeted sequencing is used in conjunction to WGS. Targeted sequencing approaches focuses on a subset of genes or regions of interest to interrogate, that have known or suspected associations with the disease or phenotype under study[68]. The broadest panel is whole exome sequencing (WES), which involves the capture of fragmented genomic DNA that collectively cover all exonic or protein coding regions. This represents approximately 1% of the entire genome but it contains nearly 85% of known disease related variants in Mendelian loci[69]. This is a cost-effective alternative to WGS and like WGS, is often used as a discovery approach.

Targeted gene panels can be customised to include as many or as few genes a study requires. The advantages of using targeted panels include: 1) assessing multiple genes across many samples in parallel; 2) time and costs associated with running multiple separate assays are reduced; 3) data set is smaller (~10 GB) and more manageable compared to WGS and; 4) high read depths (500–1000× or higher) are more cost effective.

1.6 Aims of research

The development of high-throughput sequencing technologies has allowed researchers to catalogue the mutational landscape of human cancers at an unprecedented rate. This has given researchers insight into how mutations can alter protein function, drive carcinogenesis and aid in the risk stratification of patients. However, SMZL is often precluded from large international studies resulting in an incomplete catalogue of tumour associated genomic lesions and mutational processes. The main aim of this study is to construct a detailed characterisation of the genetic landscape of SMZL through the identification of somatic variants in unmatched tumour samples in the largest SMZL cohort assessed to date. In conjunction with the analysis of somatic variants, an important part of this project also centres around the bioinformatics processing and optimisation of pipelines to obtain the best sequencing results. The first chapters will focus on establishing the

best methods to process sequencing data, while the later sections will focus on the analysis of the sequencing results. The different parts of this study with its respective objectives are detailed below:

Main objective: Construct a detailed characterisation of the genetic landscape of SMZL through the identification of somatic variants in tumour only SMZL samples.

Systematic literature review: Create, annotate and filter previously identified SMZL variants to establish a high-quality up-to-date database.

- Compile a list of variants from studies that have used NGS on SMZL tumours.
- Refine the catalogue of somatic mutations in SMZL, resulting in a high quality, annotated, up-to-date database to facilitate further studies.
- Add final list of variants to the bioinformatics pipeline.
- Determine if a systematic approach will yield any new insights into pathways targeted in SMZL or into gaps in the area.
- Make a critical analysis on the recurrent variants and or pathways targeted in SMZL.

Bioinformatics pipeline: Learn and develop a panel of genomic computational tools for the analysis of genomic datasets.

- Optimise bioinformatics pipeline to process tumour only SMZL samples to reduce or filter out false positives.
- Ensure sensitivity for known true positives.
- Compare different tools for variants calling and annotation.
- Ensure the quality of samples and data is adequate.

Filtering strategies: Develop and apply a filtering strategy to reduce the number of spurious calls present in the data set.

- Extract additional information from recalibrated BAM files.
- Develop a machine learning model using quality and sequencing information to cluster/classify variants into true variants or artefacts.
- Apply machine learning model to SMZL samples.
- Validate variants through inspection in a genomics viewer.
- Exclude germline variation using *in-silico* predictive scores and databases of known germline variation.

Analysis of NGS sequencing results: Establish a biologically relevant list of somatic mutations within the SMZL cohort.

- Establish recurrently mutated variants, genes, and pathways.

- Establish somatic interactions between genes.
- Characterise in detail the presentation of the disease correlating phenotypes to clinical phenotypes.
- Correlate genomic results with clinical outcomes.

Chapter 2 Systematic literature review of somatic mutations in splenic marginal zone lymphoma

2.1 Synopsis

This chapter interrogates the published literature to create a systematic literature review compiling all the somatic variants previously identified in splenic marginal zone lymphoma (SMZL). The collated variants are then used to create an up-to-date annotated database of published somatic mutations. This chapter takes a comprehensive look at previous SMZL studies to identify their strengths and weaknesses and identifies any gaps that could be addressed in subsequent sections.

Carolina Jaramillo Oquendo performed the systematic literature review (search, study selection, and data extraction), compiled the database of somatic variants and analysed the results. Dr. Helen Parker was the second investigator who performed the search and study selection. Prof Sarah Ennis, Prof Jon Strefford and Dr Jane Gibson acted as main supervisors overseeing the review and provided guidance in the analysis and interpretation of the data. This chapter was published in Scientific Reports in 2019[70].

2.2 Introduction

The World Health Organization (WHO) classifies splenic marginal zone lymphoma (SMZL) as a rare, low grade lymphoma comprising less than 2% of lymphoid neoplasms. At the genomic level, SMZL remains relatively understudied, with only six studies[49–54] undertaking a genome-wide approach to unravel the disease landscape, and only one study[49] employing whole genome sequencing (limited to six cases without matched germline DNA). However, the few unbiased approaches have been useful in the discovery of some of the affected pathways and key genes, such as *NOTCH2*, *KLF2* and *TP53* which are supported by extensive genomic analysis and functional work[49–51,54,71,72]. Furthermore, the targeted studies that followed or that were performed in conjunction with whole genome (WGS) or whole exome (WES) sequencing were key in establishing the recurrence of those genes in SMZL. A panel of less prevalent mutations have also been reported targeting key biological pathways, though their prevalence is uncertain, and their importance remains opaque.

A limited number of patients; the heterogeneous nature of the disease; and variable experimental and bioinformatic approaches add to the complexity of unravelling the genomic landscape of SMZL. Lack of tissue samples limit study sizes and most cases are analysed through re-sequencing

of targeted genes hypothesised to be relevant in SMZL, likely excluding genes in other pathways. Due to the rare nature of SMZL, this is a lymphoma that is poorly annotated in publicly accessible databases such as COSMIC[11]. Not only does COSMIC have limited entries on SMZL, but the way the data is organised makes it nearly impossible to separate SMZL from other marginal zone lymphomas. Additionally, not all published SMZL studies are found in COSMIC and only 2 KLF2 mutations identified in SMZL are included. Furthermore, there is a low number of unbiased studies with lack of clarity of what overlap exists between them[54].

Prior to a comprehensive analysis of an SMZL cohort, curation of a full catalogue of published somatic mutations in SMZL was needed. This would identify the strengths and weakness of the data available as well as any gaps in the study of the disease. A systematic approach was therefore necessary, and a systematic literature review was performed compiling all the variants published to date with two main aims. The first aim was to determine if this approach would yield any new insights into pathways targeted in SMZL or into gaps in the area. The second aim was to refine the catalogue of somatic mutations in SMZL, resulting in a high quality, annotated, up-to-date database to facilitate further studies.

2.3 Methodology

2.3.1 Search strategies and study selection

Two independent investigators undertook the literature search in January 2019 using PubMed (https://www.ncbi.nlm.nih.gov/pubmed) and Ovid (http://ovidsp.ovid.com) as the primary search engines according to the PRISMA-P\Preferred Reporting Items for Systematic Review and Meta-analysis Protocol[73]. Keywords used included: “Splenic Marginal Zone Lymphoma”, “SMZL”, “Marginal Zone”, “genetics”, “sequencing” and “mutation”. Data collection was performed by both investigators before any further steps. Duplicate manuscripts were removed to begin screening of the title and abstracts for those that would be used in the full-text review. Manuscript titles were screened to include records that sequenced SMZL and/or other similar mature B-cell lymphomas and exclude those that performed analysis of cases by methods other than high throughput sequencing (HTS) or Sanger sequencing. Manuscript abstracts were reviewed to include only those that sequenced confirmed SMZL cases and exclude those that were not peer-reviewed journal articles (conference abstracts). The full-text manuscripts and supplementary data were evaluated and selected for inclusion if the study reported a full list of variants with appropriate sample and mapping details. Manuscripts describing the analysis of both paired and unpaired samples were accepted. The search was limited to studies written in English. The filtering stages and inclusion and exclusion criteria are shown in **Figure 2-1**.

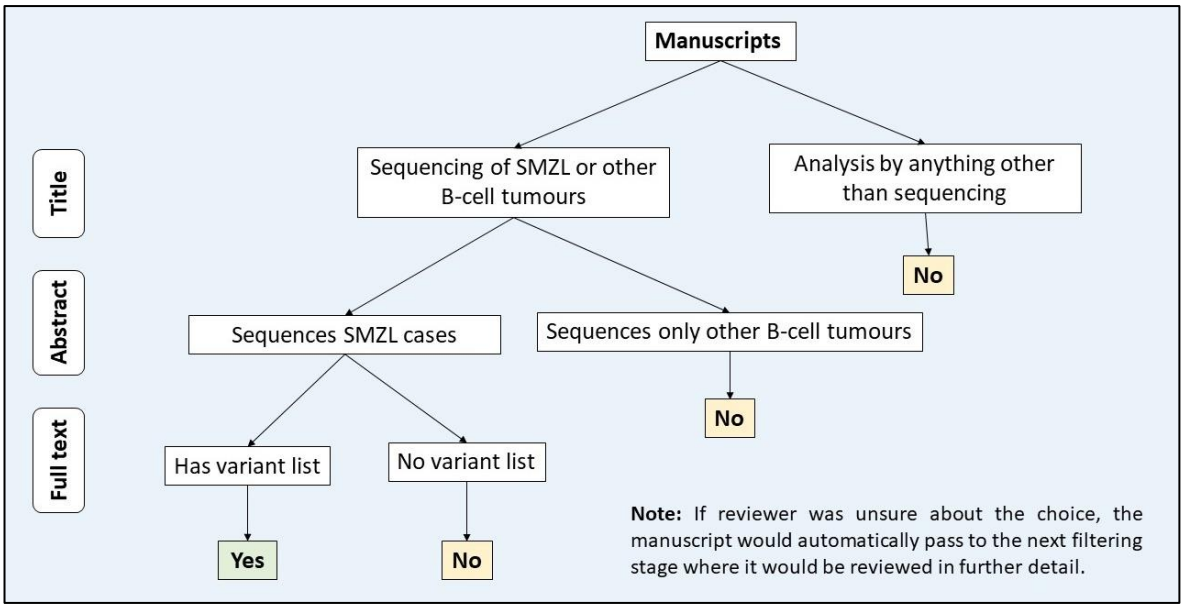


Figure 2-1. Decision tree of manuscript selection for systematic literature review. Manuscripts went through a title selection, followed by an abstract and full text review. The inclusion and exclusion criteria are shown for each of the three steps. If at any point the reviewer was unsure, the manuscript would automatically pass to the next filtering stage to be reviewed in further detail.

2.3.2 Data extraction

Genomic information was extracted from both main manuscripts and supplementary material, where the final list of variants was assembled in an Excel document. The missing base pair location and reference/alternate allele information was completed using the hg19 assembly of Ensembl Variant Effect Predictor (VEP)[74] using the mutated gene and protein or coding sequence change reported. It should be noted that for each variant, VEP outputs all possible effects of the nucleotide change in all possible transcripts. To overcome inconsistencies, three transcript tags (transcript support level, APRIS and GENCODE Basic) present in the VEP annotation were inspected to identify the highest quality and most relevant transcript to be used as well as references in the literature.

Before the list of variants was processed, there was manual curation to exclude the following:

1. Variants with an incompatible format. Variants with a file format that did not allow merging of the data easily into Excel and would mean manual imputation of large amounts of data. For example, in cases where a variant list is over 1000 variants, manual imputation is not feasible as this would likely introduce errors.

2. Duplicate variants from the same sample. These were a result of studies, likely from the same research group, which reported variants from previously published cases. Identification of duplicates was based on sample ID and variant characteristics

(sequencing depth and VAF)[50–52,54,71,75]. These variants were flagged and included only once.

3. Variants lacking information necessary to remap and annotate (location and reference/alternate allele).

Once the list was populated and filtered, it was remapped from hg19 to the hg38 genome assembly with the NCBI remap tool (https://www.ncbi.nlm.nih.gov/genome/tools/remap). Finally, the remapped list of variants was annotated using the Annovar software v.2016Feb01[76] adding a gene-based annotation to identify functional effects, frequency of the variant in specific databases, and scores that predict how mutations affect protein function (see **section 5.3.2** for detailed description of databases and versions used in Annovar). Additional information found in the manuscripts and supplementary files was also added which included: the variant allele frequency (VAF), depth, confirmed somatic status (lack of variant in matched germline DNA), sequencing method by which somatic status was confirmed and sequencing approach. Preceding data analysis, the list was filtered to retain only those variants that were likely to be somatic mutations or likely drivers of disease. Variants that were filtered out were those that: a) fell within UTRs and intergenic regions; b) synonymous variants and; c) variants that had a frequency greater than 1% in databases of normal germline variation (The Genome Aggregation Database[77], 1000 Genomes Project[78], NHLBI GO Exome Sequencing Project[79], Exome Aggregation Consortium[77]). The final database was comprised of all the remaining variants and these were the focus of subsequent analysis. Within the final list, there was a subset of variants that originated from unbiased approaches (WGS and WES). This WGS/WES subset was looked at separately. **Figure 2-2** summarises all the steps from data collation to the final variant list.

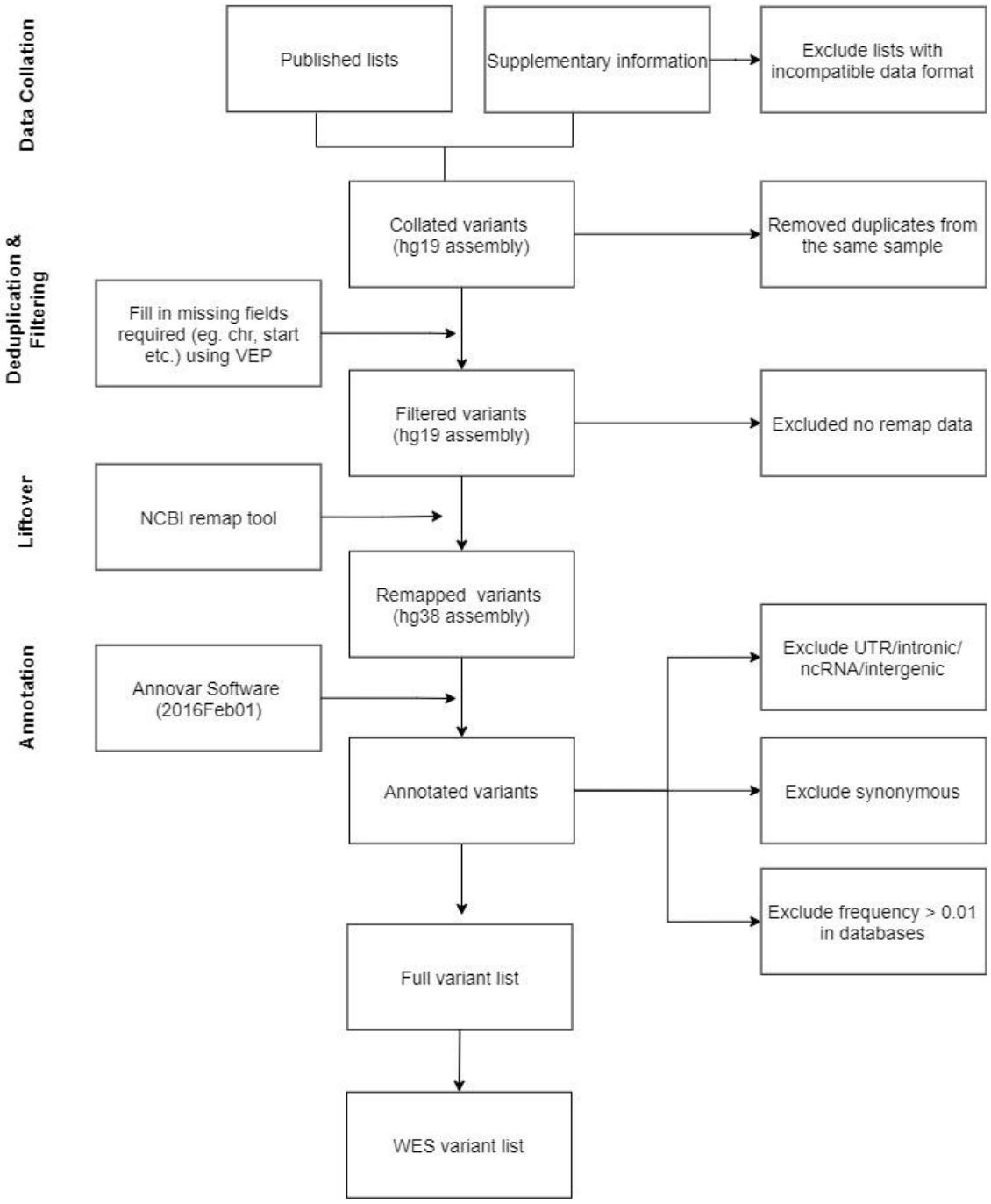


Figure 2-2. Flowchart of database compilation and variant filtering. The flowchart begins at the data collation step, where all variants from the published manuscripts and supplementary material were collated into a single list. This is followed by a manual curation step where duplicate samples were included only once and missing fields required for remapping filled in. Variants with not enough data were excluded and all remaining variants were remapped to the hg38 reference genome. Once remapped, variants were annotated using Annovar software. After annotation variants were filtered once more to enrich for somatic variants.

2.3.3 Data visualisation and analysis

The WGS/WES subset was analysed first as an unbiased cohort. However, variants from WGS were excluded since the file format did not allow merging of the data easily into Excel and effectively only WES samples remained. Subsequently, all collected variants were compared across studies.

To define the putative frequency of recurrently mutated genes, it was assumed that all genes were screened in all studies. However, for the twenty-one genes with the greatest cumulative number of variants, the number of assessed cases was ascertained individually. This was the only available approach, as the total number of genes analysed is not consistently or accurately reported across the targeted re-sequencing studies. This simplification, however, is likely to underestimate the prevalence of mutations in some genes.

The final annotated variant list was used as input into R packages maftools[80] and GenVisR[81] to proceed with the data visualisation and analysis. The Discrete Independence Statistic Controlling for Observations with Varying Event Rates (DISCOVER)[82] algorithm was used to test for co-occurrence and mutual exclusivity between genes in the SMZL database. This independent test accounts for the overall alteration rates of each individual tumour by creating a background matrix, which is how tumour specific alteration rates are incorporated by the test. The background matrix is created with simple binary mutation matrix of ***m*n*** dimension where ***m*** is the number of genes and ***n*** the number of cases to get a genome wide view of each tumour. Of the 14 published SMZL studies, six that did not assess more than 100 genes were excluded from this analysis to reduce bias, leaving 240 patients from eight published studies to be assessed.

2.4 Results

2.4.1 Study selection and characteristics

After collating and removing duplicate manuscripts, 625 unique manuscripts were kept of 1342 initially identified (n=793 in PubMed, n=549 in Ovid). 584 manuscripts were discarded after title review and 70 manuscripts were discarded after abstract review. The full texts of the remaining 41 manuscripts were carefully examined to ensure they met the inclusion criteria (**Figure 2-1**). Twenty-six records were excluded, as they did not report a list of variants (**Figure 2-3**). The remaining studies (n=14), to be included in the analysis, were split into three categories: 1) discovery; 2) confirmation/extension and; 3) comparison. **Table 2-1** lists the manuscripts selected and provides a general overview of the methods and samples used in each one. In terms of unbiased discovery approaches, there was only a single WGS study and five studies that

implemented WES, which subsequently confirmed variants using targeted or Sanger sequencing[49,51–54]. The nine extension/confirmation studies were hypothesis based, targeting pathways identified in discovery cohorts or validation of recurrently mutated genes[49,50,53,54,71,72,83–85]. Three studies sequenced SMZL cases and compared these to other B-cell lymphomas[86–88]. Lack of matched germline allowed only a fraction of the variants (25%) in the studies to be confirmed as somatic. Somatic status and method of confirmation are indicated in the final list of variants and on **Table 2-1**.

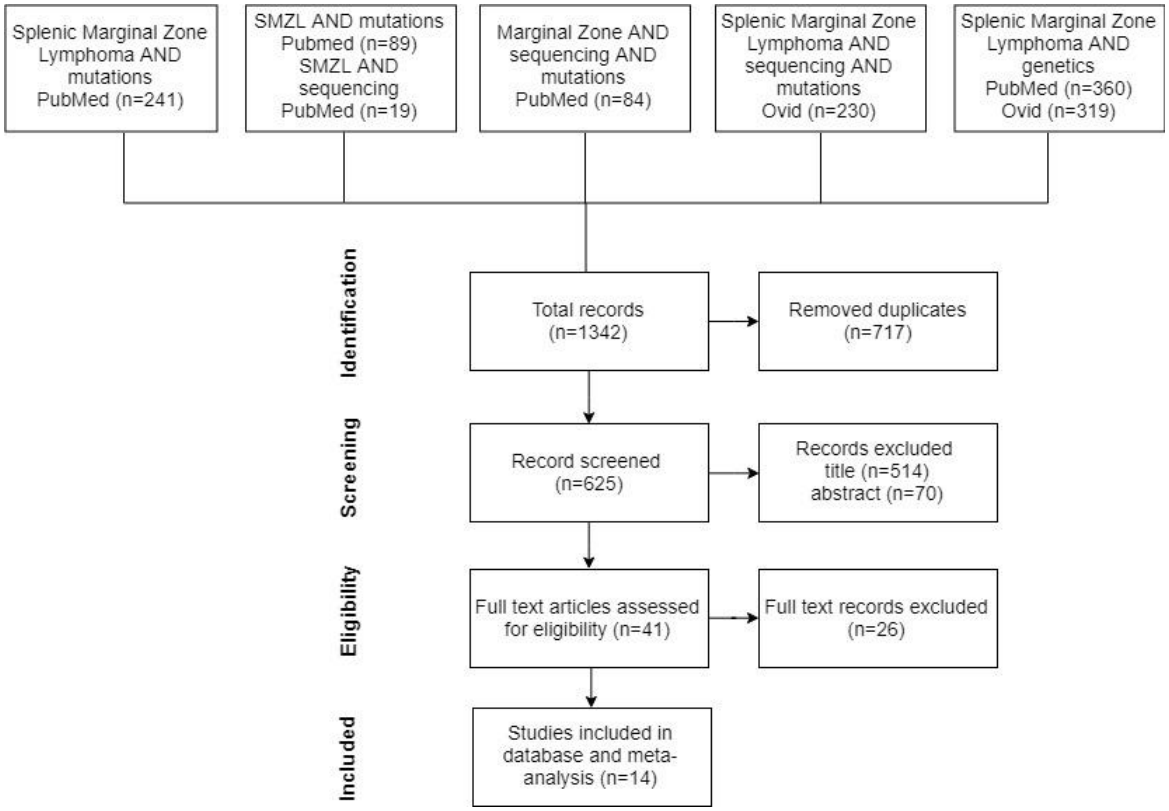


Figure 2-3. Flowchart of manuscript selection and filtering. The figure goes through the search strategy, starting with the combination of search terms used in the databases. Numbers denote amount of records or manuscripts at each step. Once all entries were compiled into a single list, duplicate manuscripts were removed and those remaining were reviewed to identify those that would be used in the full text review. The number of manuscripts excluded and those that were kept are stated in each of the steps. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

Table 2-1. Detailed characterisation of studies included in the SMZL database. Studies are listed in chronological order with a general overview of the methods and samples used in each one. Extended information on bioinformatics tools used in each study can be found in **Supplementary Table 1**.

Study	Methods Used M = Sequencing method C = Capture/chemistry S = Sequencing V = Validation	Samples TO = Tissue origin MG = Matched germline, D = Diagnosis
Rossi et al. (2011) PMID:21881048 *Extension / confirmation (n=101)*	**M:** Targeted (21 genes) **C:** PCR **S:** Sanger **V:** Comparison to matched normal	**TO:** Not specified **MG:** Saliva (n=18) **D:** WHO classification and SMZL Working Party criteria
Rossi et al. (2012) PMID: 22891273 *Discovery (n=8)* *Extension / confirmation (n=117)*	**M:** WES **C:** SureSelectXT Human Exon Capture 50Mb Kit (Agilent Technologies) **S:** HiSeq2000 (Illumina) - paired end 2x100 bp read option **V:** Sanger **M:** Targeted (61 genes) **C:** PCR **S:** Sanger **V:** Candidate confirmed on both strands	**TO:** Frozen spleen biopsies of newly diagnosed, previously untreated patients. **MG:** Saliva or peripheral blood granulocytes (n=48) **D:** Spleen histology and confirmed by centralized pathological revision. All cases in discovery and screening panel lacked t(11;18) and t(14;18). All cases lacked BRAF p.V600E mutation
Yan et al. (2012) PMID: 22102703 *Extension / confirmation (n=57)*	**M:** Targeted (6 genes) **C:** PCR **S:** Sanger **V:** Candidate confirmed by at least two independent PCR.	**TO:** Frozen tissue (n=23) and FFPE tissue (n=34) from spleen **D:** Histological assessment of spleen according to WHO classification. Analysis of micro dissected normal cells.
Kiel et al. (2012) PMID:22891276 *Discovery (n=6)* *Extension / confirmation (n=93)*	**M:** WGS **C:** QIAamp DNA extraction kit (QIAGEN) **S:** not specified **M:** Targeted (NOTCH2) **C:** PCR **S:** Sanger	**TO:** Frozen tumour tissue **MG:** NA **D:** Reviewed independently by three haematopatologists according to WHO classification criteria.

Study	Methods Used M = Sequencing method C = Capture/chemistry S = Sequencing V = Validation	Samples TO = Tissue origin MG = Matched germline, D = Diagnosis
Parry et al. (2013) PMID:24349473 *Discovery (n=7)*	**M:** WES **C:** SureSelectXT Human Exon Capture 51Mb V4, 50Mb V3 Kit (Agilent Technologies) **S:** HiSeq (Illumina) **V:** Sanger	**TO:** Spleen biopsies (n=5) and peripheral blood (n=2). Tissue CD19+ purified cells. **MG:** Saliva (n=7) **D:** Met criteria established my Matutes *et al*. 5/7 splenectomy with histology typical of SMZL, chromosomal aberrations targeting 7q and IGHV-2*04 usage.
Martinez et al. (2014) PMID:24296945 *Discovery (n=15)* *Extension / confirmation (n=16)*	**M:** WES **C:** SureSelectXT Human Exon Capture 50Mb Kit (Agilent Technologies) **S:** HiSeq2000 (Illumina) - paired end 76 bp read option **V:** 454 Roche and Sanger **M:** Targeted (NOTCH2) **C:** PCR **S:** Sanger	**TO:** WES - Isolated CD19 cells from peripheral blood (n=10) and freshly frozen biopsies n=5). FFPE tissue (n=16). All samples taken before therapy. **MG:** Oral mucosa (n=13) and granulocytes (n=2). **D:** Reviewed independently by three haematopatologist according to WHO classification.
Parry et al. (2015) PMID:25779943 *Extension / confirmation (n=175)*	**M:** Targeted (768 genes) **C:** HaloPlex Target Enrichment System (Agilent Technologies) **S:** not specified **V:** Sanger **M:** Targeted (NOTCH2) **C:** PCR **S:** Sanger	**TO:** Peripheral blood (n=135), bone marrow (n=22), spleen (n=17), or lymph nodes (n=1). **MG:** Buccal cells or sorted T-cell (n=25) **D:** Met criteria established my Matutes *et al*.
Piva et al. (2015) PMID: 25283840 *Extension / confirmation (n=96)*	**M:** Targeted (KLF2) **C:** Repli-g Mini Kit (QIAGEN) & PCR **S:** ABI PRISM 3100 Genetic Analyzer (Applied Biosystems) & Genome Sequencer Junior instrument (454 Life Sciences)	**TO:** All samples obtained at diagnosis from the involved site. **MG:** Saliva or blood granulocytes **D:** All cases lacked the t(11;18) and the t(14;18) translocations, and the p.V600E BRAF mutation.

Study	Methods Used M = Sequencing method C = Capture/chemistry S = Sequencing V = Validation	Samples TO = Tissue origin MG = Matched germline, D = Diagnosis
	V: Candidate confirmed by at least two independent PCR.	
Peveling-Oberhag, et al. (2015) PMID:26498442 *Discovery (n=2)* *Extension / confirmation (n=24)*	**M:** WES **C:** SureSelectXT Human Exon Capture 50Mb Kit (Agilent Technologies) **S:** SOLiD4 Platform (Life Technologies) **V:** Comparison to matched normal and Sanger **M:** Targeted (10 genes) **C:** PCR PyroMark PCR kit **S:** PyroMark Q24 (QIAGEN) **M:** Targeted (NOTCH2 & SMYD) **C:** PCR **S:** Sanger	**TO:** WES - Splenic tissue (n=2), Fresh frozen tissue (n=8), FFPE tissue (n=16). **MG:** Not specified **D:** Morphological, cytochemical and immunophenotypic methods according to 2008 WHO classification. All cases were CD5-, CD10-, Bcl-6-, CD23-, with typical pattern of white pulp involvement.
Clipson et al. (2015) PMID:25428260 *Discovery (n=16)* *Extension / confirmation (n=96)*	**M:** WES **C:** SureSelectXT Human Exon Capture 50Mb Kit (Agilent Technologies) **S:** HiSeq2000 (Illumina) - paired end 76 bp read option **V:** Sanger **M:** Targeted (KLF2) **C:** PCR **S:** Sanger **V:** Candidate confirmed by at least two independent PCR.	**TO:** Fresh frozen lymphoma tissue (n=77), Leukaemic peripheral blood (n=3), FFPE tissue (n=25). **MG:** non-neoplastic FFPE tissue (n=1), Buccal swap or non-involved peripheral blood (n=2) **D:** WHO classification
Spina et al. (2016) PMID:27335277 *Comparison (n=32)*	**M:** Targeted (504 genes) **C:** SeqCap EZ choice libraries (NimbleGen System) **S:** MiSeq Analyzer (Illumina) - paired end 2x250 bp read option	**TO:** Fresh frozen spleen (n=39), Cell lines VL51, SSK-41 and KARPAS-1718 (n=3). **MG:** Saliva or peripheral blood granulocytes (n=14) confirmed tumour free by PCR

Study	Methods Used M = Sequencing method C = Capture/chemistry S = Sequencing V = Validation	Samples TO = Tissue origin MG = Matched germline, D = Diagnosis
	V: Sanger. Candidate confirmed by at least two independent PCR	**D:** Confirmed by pathological revision of spleen histology, lack of clinical evidence of extra nodal or nodal disease, lacked cyclin D1 expression, t(11;14) and t(14;18) translocations and BRAF p.V600E mutation. Harboured 7q deletion (305). Preferentially utilized the IGHV1-2*04 allele (24%)
Campos-Martin et al. (2017) PMID:28522570 *Extension / confirmation (n=84)*	**M:** Targeted (NOTCH2 & KLF2) **C:** PCR **S:** Sanger	**TO:** Not specified **MG:** NA **D:** According to WHO classification and Matutes *et al*.
Jallades et al. (2017) PMID:28751561 *Comparison (n=46)*	**M:** Targeted (109 genes) **C:** Agilent enrichment method with biotinylated oligonucleotide probes **S:** HiSeq2000 (Illumina) - paired end 76 bp read option	**TO:** Spleen **MG:** NA **D:** According to WHO classification and other published studies.
Pillonel et al. (2018) PMID:29556019 *Comparison (n=12)*	**M:** Targeted (146 genes) **C:** IonTorrent AmpliSeq HTS Lymphoma panel **S:** IonTorrent S5XL	**TO:** Not specified **MG:** NA **D:** According to 2007 WHO classification.

2.4.2 Database collation

All variants reported by the selected studies were collated into a single list for a total of 3529 variants derived from 508 patient cases. Before annotation, the variant list was manually curated to exclude the following:

1. WGS variants from the supplementary list of the Kiel et al. (2012) manuscript. The supplementary file had a PDF format, which did not allow for the merging of the data in Excel. For this set of variants, manual imputation was not feasible as there were over 1000 variants.

2. Duplicate variants (n=256) originating from the same patient cases.

3. Variants (n=21) lacking information necessary to remap and annotate (location and reference/alternate allele).

Post-annotation, the following variants were excluded: a) 195 UTR/intergenic variants; b) 201 synonymous variants and; c) 39 variants with a frequency greater than 1% in known databases (**Figure 2-4**). The majority of the synonymous variants were reported by the Martinez study[52] (n=195), as the authors did not remove synonymous variants from the published list in their supplementary data. After variant filtering the resulting list contained 2817 variants, termed 'full variant list', with a subset of 1009 variants resulting from unbiased studies. This list from unbiased studies was meant to contain both WGS and WES, however, since the Kiel et al. data (the only WGS) was not included due to format issues, the list only included WES studies and is referred to as 'WES variant list'. In the final 2817 list of variants, 568/2817 could be annotated with a COSMIC ID (duplicate variants from different cases were included). **Figure 2-4** summarises the filtering criteria and the number of variants removed with each filter.

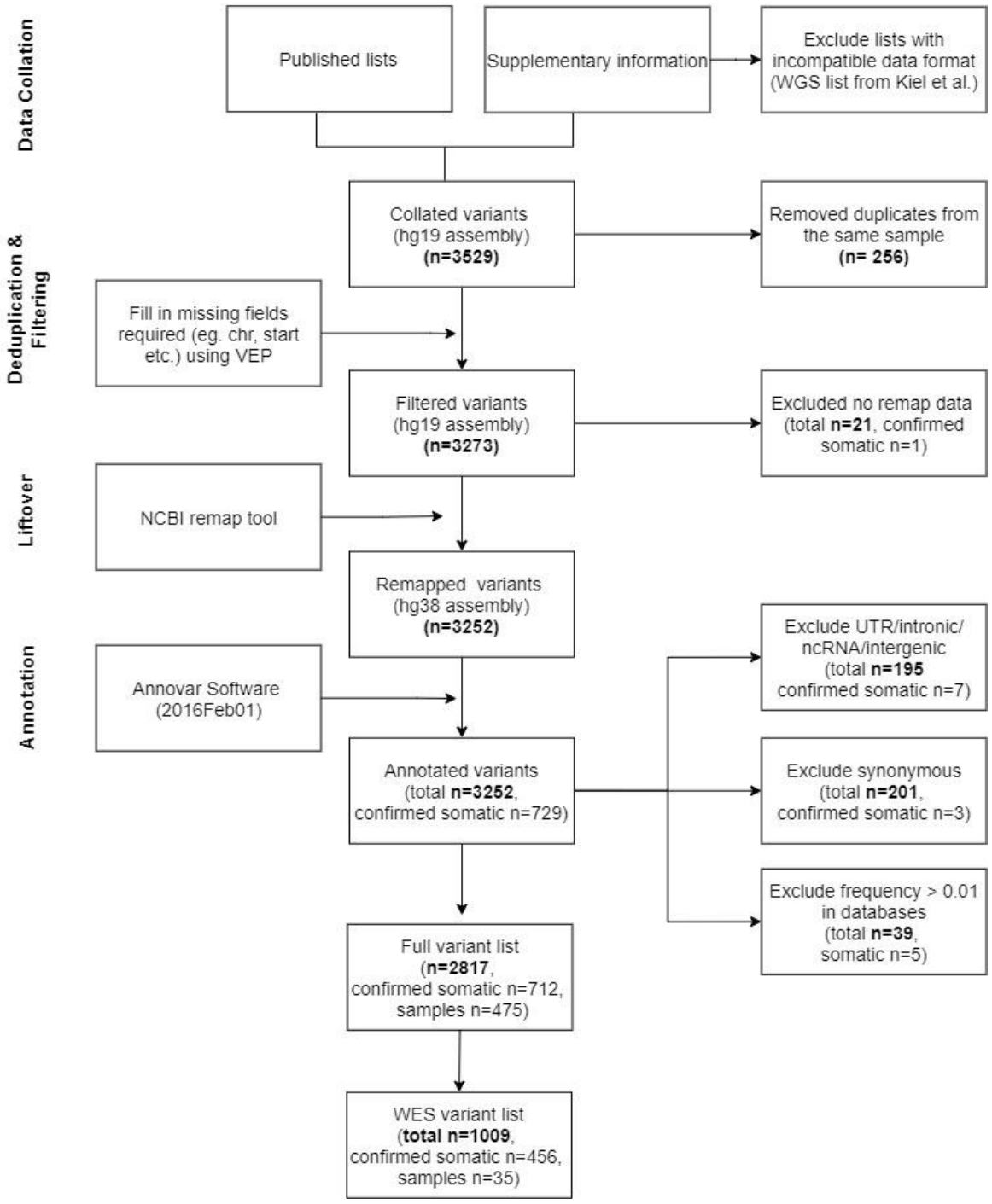


Figure 2-4. Flowchart of database compilation and variant filtering with results. The flowchart begins at the data collation step, where all the lists of variants from the published manuscripts and supplementary information were collated into a single list. Subsequent filtering strategies and data manipulation tools are described. Numbers in bold denote number of variants at each step. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

The final number of cases, variants, genes and confirmed somatic variants included in the final database list, as well as the WES subset, are detailed in **Table 2-2** and **Table 2-3** respectively. The final variant list was comprised of 2817 variants from all 14 studies. The Parry et al (2015), Martinez et al (2014) and Rossi et al (2012) studies contributed the highest number of variants accounting for around 73% of the total. 46% of the variants came from the Parry study, 17% of the variants from the Martinez study and 11% from the Rossi study. 711/2817 variants were

confirmed somatic variants where 37% came from the work by Rossi et al (2012). The breakdown of contributions to the database per paper is also detailed in tables **Table 2-2** and **Table 2-3**.

Table 2-2. SMZL database input. Breakdown of the number of cases, variants, genes and confirmed somatic variants included in the database from each study.

Source	samples	variants	genes	validated somatic
Campos-Martin, 2017	22	27	2	7
Clipson et al, 2015	76	271	141	48
Jallades et al, 2017	28	50	8	0
Kiel et al, 2012	25	26	1	0
Martinez et al, 2014	19	481	462	63
Parry et al, 2013	7	169	158	169
Parry et al, 2015	172	1283	425	73
Peveling-Oberhag et al, 2015	2	21	20	21
Pillonel et al, 2018	12	29	22	0
Piva et al, 2015	19	22	1	14
Rossi et al, 2011	17	17	4	9
Rossi et al, 2012	83	303	200	266
Spina et al, 2016	12	101	82	34
Yan et al, 2012	14	17	3	7
Total	508	2817	-	711

Table 2-3. SMZL database WES subset. Breakdown of the number of cases, variants, genes and confirmed somatic variants included in the WES subset from each study.

Source	samples	variants	genes	confirmed somatic
Clipson et al, 2015	3	142	135	2
Martinez et al, 2014	15	476	461	63
Parry et al, 2013	7	169	158	169
Peveling-Oberhag et al, 2015	2	21	20	21
Rossi et al, 2012	8	201	190	201
Total	35	1009	-	456

For variants with depth and VAF data available, the VAF distribution of the confirmed somatic variants was compared to that of variants not validated as somatic. The distribution of the validated variants had two major peaks, at a VAF of around 15% and 30%, and a tail with a very small peak at around 90% (**Figure 2-5**). The distribution of the non-validated variants also had two big peaks; however, the peaks were much more separate at a VAF of around 10% and 50%. These results indicate potential germline variation in the non-validated variants considering the high density of variants with ~ 0.5 VAF. The distribution for both validated and non-validated variants was skewed to the right (**Figure 2-5**).

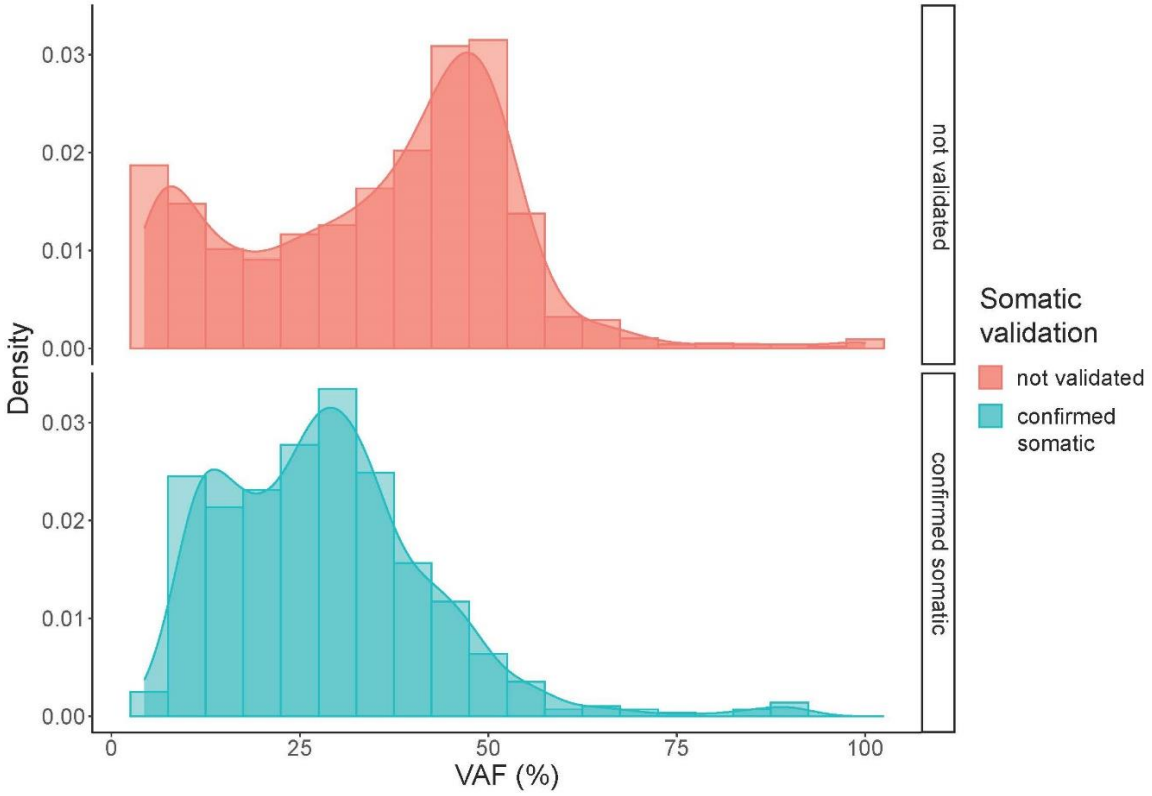


Figure 2-5. VAF distribution in validated and non-validated somatic variants in the SMZL database. The top (red/pink) distribution shows the non-validated variants, while the bottom (blue/green) shows the distribution of the VAF for the confirmed somatic variants. Both distributions are skewed right. Only variants that had depth and VAF data (n=2432) are shown here.

2.4.3 Recurrently mutated genes in WES subset

Overall, 35 unique samples were sequenced using WES across five different studies all with matched germline DNA, accounting for 1009 variants in the final variant list. **Figure 2-6** displays the number of overlapping genes across the five WES studies, with limited concordance between all five (there was no single gene harbouring somatic mutations across all studies). This is to be expected given the small sample size. There were however three genes (*AMOTL1, FAT4* and *USH2A)* mutated across three studies[50–52]. *SPEN* was the most frequently mutated gene, with five variants across four samples. Out of the five variants identified in *SPEN* three were confirmed somatic. *FAT4*, *MYD88, NOTCH2*, and *TNFAIP3* all followed *SPEN* with four variants each.

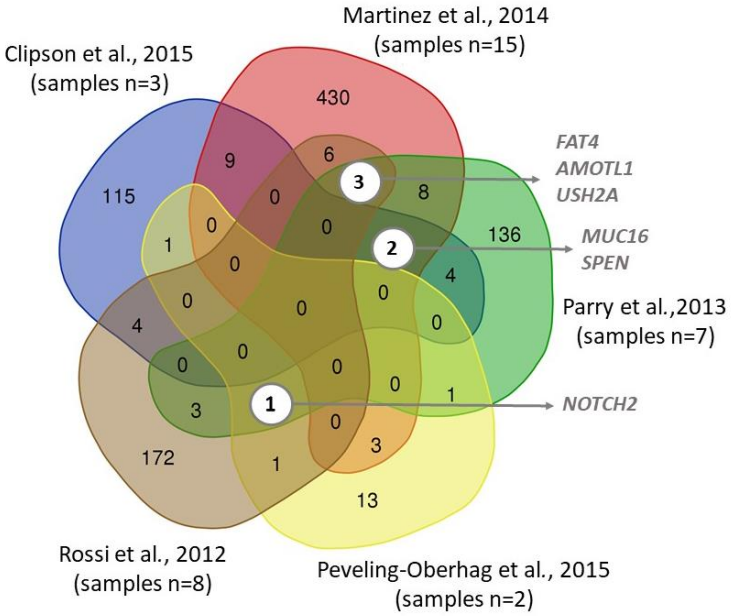


Figure 2-6. Venn diagram of gene overlap in WES studies. The figure shows the common and unique genes reported in each study with no overlap between all five. Where there was an overlap of genes identified by more than three studies there are white circles to emphasise the number and name of genes found. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

2.4.4 Recurrently mutated genes in the full dataset

Next, the full annotated list of 2817 variants was analysed. 711/2817 variants were confirmed somatic (**Table 2-2**), across 1239 genes. For future reference, the full list of variants obtained from this review will be referred to as SMZLrefDB. The number of variants per gene in the SMZLrefDB is represented as a Wordcloud in **Figure 2-7**, in which the font size reflects the prevalence of variants in each gene. A summary of the entire database can be seen in **Figure 2-8**.

Figure 2-7. Wordcloud of gene symbols present in SMZLrefDB. The size of each gene symbol is proportional to the number of mutations in each gene (range: 1-123 mutations). *NOTCH2* (n=123) and *KLF2* (n=121) had the highest number of mutations, followed by *TNFAIP3* (n=75), *TP53* (n=60) and *MYD88* (n=43). Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

Most variants found in the SMZLrefDB were missense mutations (n=2126), followed by frameshift deletions (n=238) and stopgain or nonsense mutations (n=237). The majority were single nucleotide changes (n=2444), followed by deletions (n=256) and a very small number of insertions (n=90). C to T substitutions were the most common base change (**Figure 2-8**).

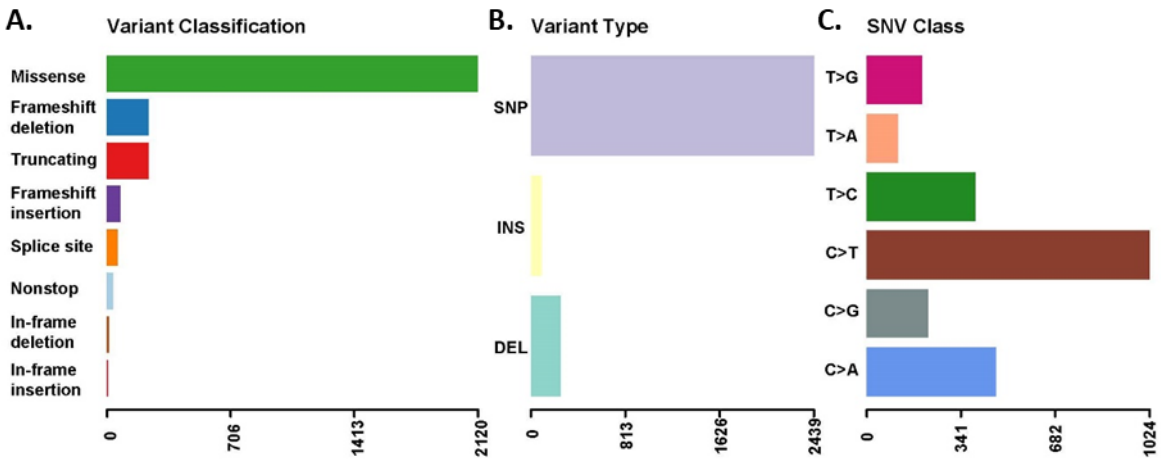


Figure 2-8. Summary of SMZL variants in the SMZLrefDB (n=2817). **A.** Number of variants across the SMZLrefDB classified by function. **B.** Breakdown of variant type. Single nucleotide polymorphisms (SNP) are in purple, insertions (INS) in yellow and deletions (DEL) in green. **C.** Breakdown of nucleotide substitution. X-axis for all three figures show the number of variants within the cohort.

Figure 2-9 expands on the twenty-one genes with the highest number of variants. The bar graph displays the mutational frequency (%) of each gene, ordered from high to low. Each gene (represented by a bar) is further annotated with the types of mutations present within it. The number of assessed cases in each gene is displayed in **Supplementary Table 2**.

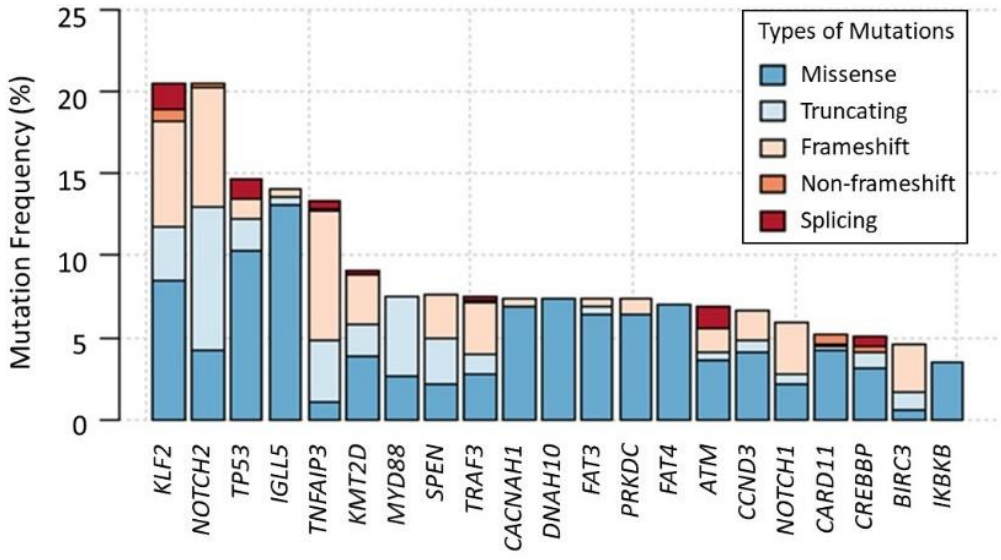


Figure 2-9. Mutation frequency (%) of the top 21 mutated genes. The graph displays the frequency of mutations (# mutated cases/ total # cases) in each gene as well as an overview of the type of mutations (missense, nonsense, frameshift and splicing). Genes are listed in descending order of most frequently mutated to least. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

The three genes with the highest mutational frequency were, as expected, *KLF2, NOTCH2* and *TP53*. *KLF2* (21%), had the greatest mutational frequency with mutations that included missense [n=50], frameshifts [n=38], truncating [n=19], splicing [n=9] and non-frameshift [n=5]. *KLF2* mutations were found throughout the entire protein (**Figure 2-10**). *KLF2* harboured a recurrent mutation (p.Q24X) found in 10 [1.7%, n=10/589] reported cases. This variant is not found in the COSMIC database but is predicted to be pathogenic with a CADD Phred score of 35[89].

Following *KLF2*, *NOTCH2* had the second-highest mutational frequency with the majority of its mutations having a truncating effect [n =53], followed by frameshifts [n=44], then nonsynonymous [n=25], and a single non-frameshift, most of which cluster around exon 34 in the C-terminal PEST domain (**Figure 2-10**). A recurrent variant (p.R2400X) located in the PEST domain was present in 27 [4.5%, 27/602] reported cases and is found in the COSMIC database (COSV56682519) predicted to be deleterious with a CADD Phred score of 44[89].

TP53 (15%) and *IGLL5* (14%), the third and fourth most frequently mutated genes respectively, harboured mostly nonsynonymous mutations and few truncating mutations. Mutations in *TP53* were enriched within the DNA binding domain (**Figure 2-10**) where the most recurrent variant (p.R141H) was present in three of the assessed cases. 28 of the 52 variants were annotated within the IARC *TP53* database[90].

Lolliplots of other recurrently mutated genes (*KMT2D, MYD88, SPEN,* and *TRAF3*) are shown in **Figure 2-11**.

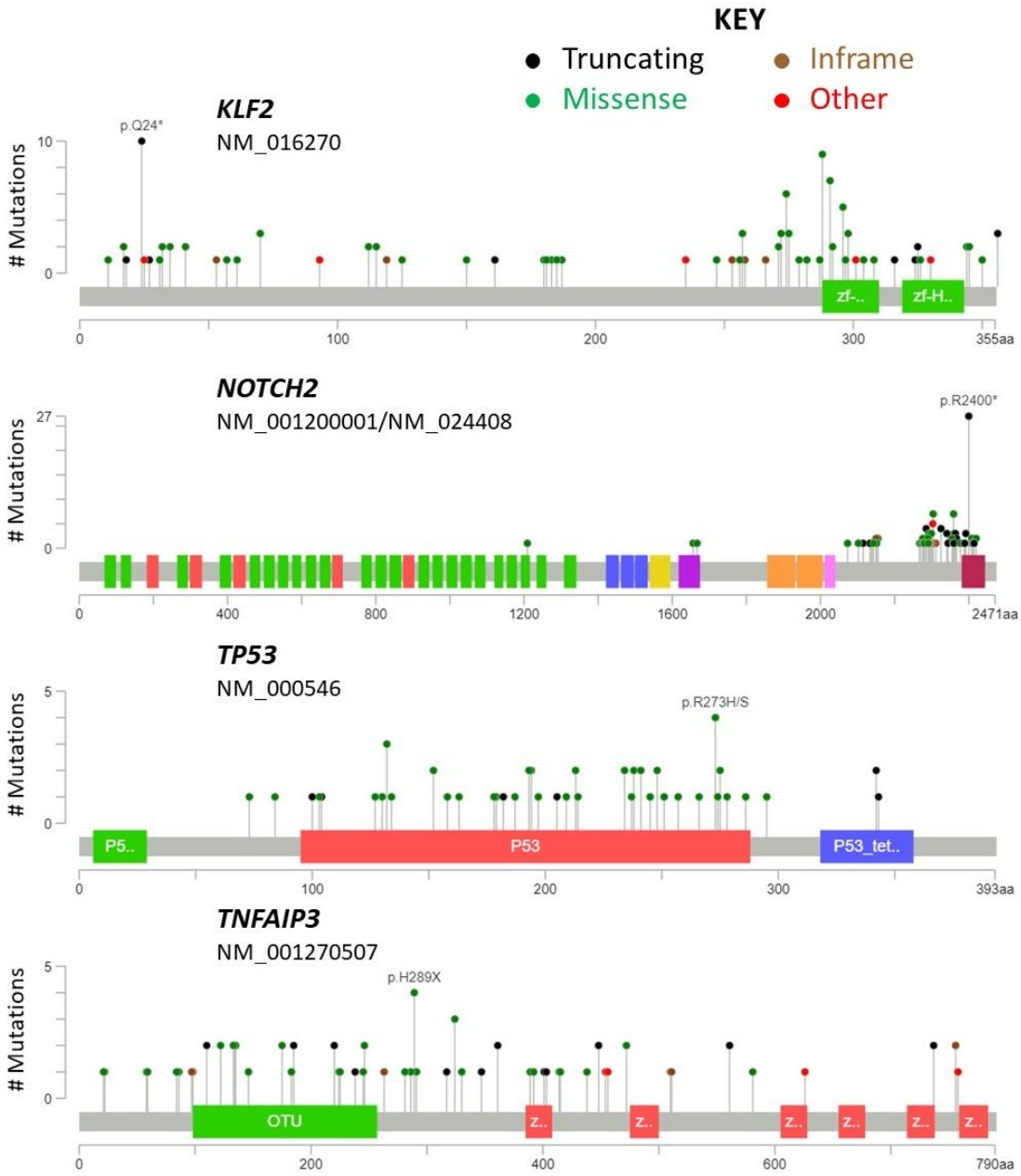


Figure 2-10. Lolliplot of *KLF2, NOTCH2, TP53* and *TNFAIP3*. Linear proteins representing each gene with their respective domains. The height is representative of the number of variants reported (The y-axis is not the same proportion for all figures), and circle colour identifies the type of mutation. The transcript used for each protein is stated under the gene name and the colours of the domains were randomly assigned. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

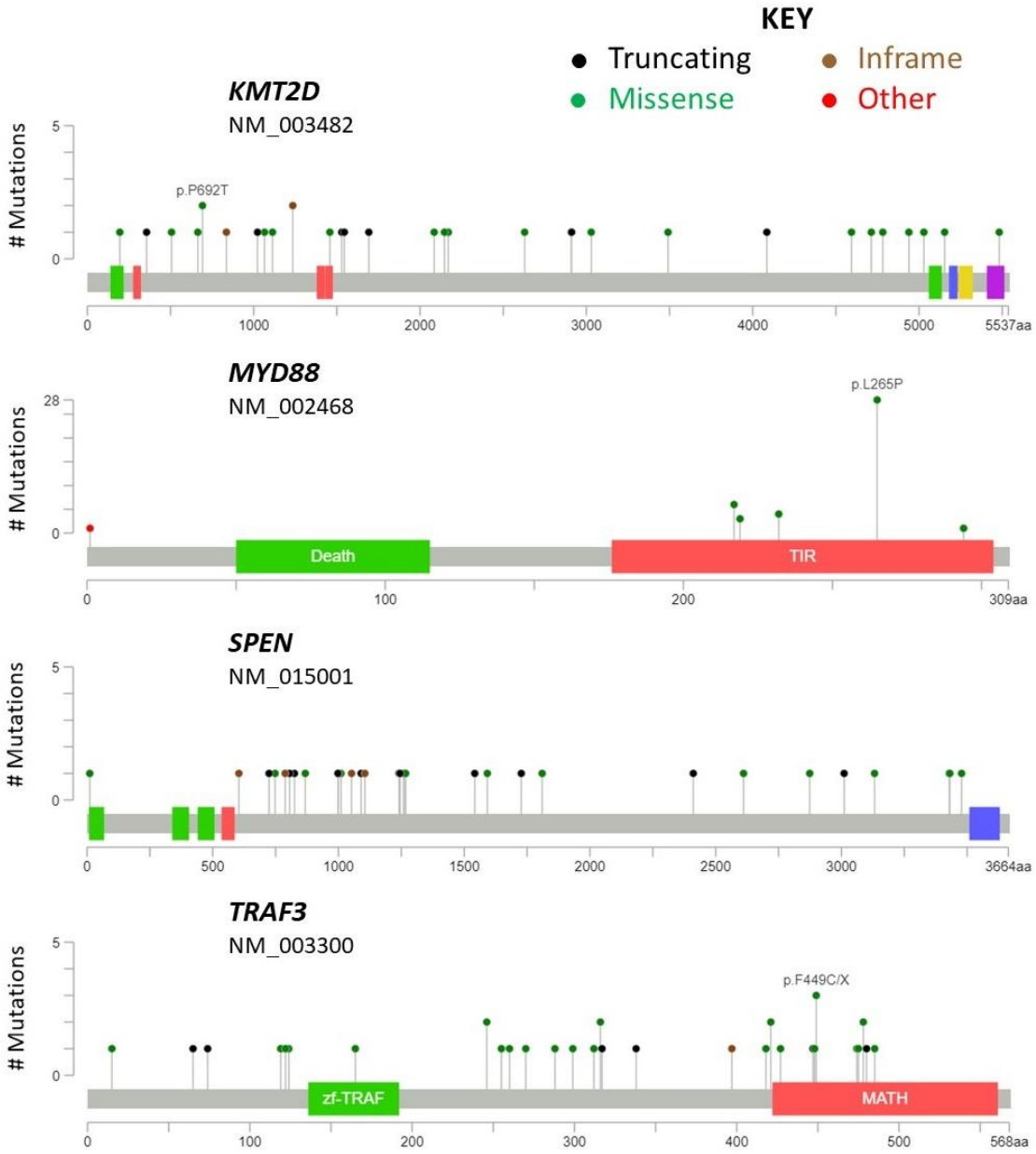


Figure 2-11. Lolliplot of *KMT2D, MYD88, SPEN,* and TRAF3. Linear protein representing each gene with their respective domains. The height is representative of the number of variants reported (The y-axis is not the same proportion for all figures), and circle colour identifies the type of mutation. The transcript used for each protein is stated under the gene name and the colours of the domains were randomly assigned. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

Other recurrently mutated genes included *TNFAIP3, KMT2D, MYD88, TRAF3, SPEN,* and *CCND3.* Mutations in *TNFAIP3* and *KMT2D* were not clustered events but rather distributed throughout the entire protein (**Figure 2-11**). The p.L265P *MYD88* variant accounted for 65% [28/43 mutations] of all *MYD88* variants. This variant is pathogenic according to ClinVar[91] and annotated in the COSMIC database (COSV57169334). The p.L265P *MYD88* variant is in the toll/interleukin-1 receptor homology (TIR) domain (**Figure 2-11**) and is recurrently mutated in several mature B-cell tumours[92]. Other recurrent *MYD88* mutations included p.V217F, pM232T, and p.S219C present in six, four, and three cases respectively. The latter (p.S219C), along with the p.L265P variant, has been identified in a recently recognized entity, termed clonal B-cell lymphocytosis of MZ origin

(CBL-MZ)[34,93], that can sometimes progress to SMZL[94]. The other two *MYD88* mutations (p.V217F, pM232T) have been identified in both chronic lymphocytic leukaemia (CLL) and Diffuse large B-cell lymphoma (DLBCL)[11]. *CCND3* mutations in the C-Terminal domain were mutated in 18 cases (7%). These mutations have been identified in a panel of mature B-cell neoplasms, and most recently in patients with splenic diffuse red pulp small B-cell lymphoma (SDRPL), another entity difficult to differentiate from SMZL[95]. In *IKBKB* a recurrent variant (p.K169E) predicted to be deleterious with a CADD Phred score of 29.7[89,96] was identified in nine cases [1.8%, 9/510].

2.4.5 Somatic interactions

Using the DISCOVER algorithm with and false discovery rate (FDR) of 1% there were three pairwise combinations that were significantly mutually exclusive:

1. *KLF2* and *IGLL5* (p<0.001)
2. *TP53* and *NOTCH2* (p=0.002)
3. *TP53* and *KLF2* (p=0.0045)

While there was evidence of mutual exclusivity, this algorithm did not pick up any co-occurring events. **Figure 2-12** shows the results of the DISCOVER test plotted on a heatmap.

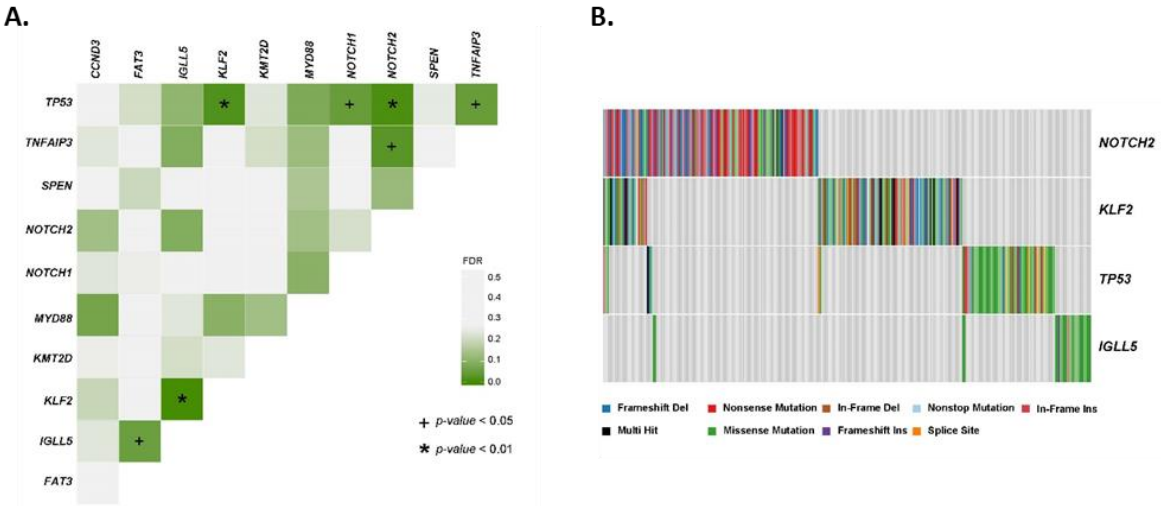


Figure 2-12. DISCOVER mutual exclusivity test results. **A**. Heat map displaying the corrected *p*-value for the gene pairs tested for mutual exclusivity in the DISCOVER algorithm. The plus sign (+) indicates correlations with a *p-value* < 0.05 and the asterisk (*) highlights those pairwise combinations with a *p-value* < 0.01. **B** Waterfall plot of mutations found in *KLF2, NOTCH2, TP53* and *IGLL5.* Each column represents a sample, and each row a gene. Each column is coloured according to the mutation type present in the sample and grey if no mutations are present. Figure by Jaramillo Oquendo et al[70] licenced under CC BY 4.0.

2.5 Discussion

Due to its rare nature, SMZL is not widely studied and has no entries in cancer databases such as The Cancer Genome Atlas (https://cancergenome.nih.gov/) or the International Cancer Genome Consortium (https://icgc.org/). The COSMIC database houses only 23% of all SMZL variants reported and does not include recent studies. To our knowledge, this is the first systematic review

of published SMZL genomic data, in this case, pooled from 14 studies. It fulfils an unmet need, as currently there are no other resources like it. Furthermore, unbiased genomic studies are limited and the available WGS or WES data fail to provide a complete detailed catalogue of somatic mutations. A systematic approach was the best option to begin to unravel the genomic landscape, providing confirmation of recurrently mutated genes and potentially highlighting genes that might not have been at the centre of SMZL studies due to power limitations in any single study.

Several limitations were encountered pertaining to the experimental/analytical design and the lack of required information in published data. The first limitation encountered was the number of unbiased studies. Only five of the fourteen studies performed WGS or WES and the remaining studies (validation and comparison) introduced bias by employing targeted panels of genes hypothesised to be implicated in SMZL biology or related mature B-cell malignancies. The second limitation was that the number of assessed genes was not the same across studies. Some genes will present more mutations simply because they were the target of many panels, likely biasing estimation of mutational frequency compared to less frequently assessed genes. The third limitation was the evolving nature of NGS experimental and analytical pipelines. Each study design employed specific parameters and conditions as shown in **Table 2-1** and further described in **Supplementary Table 1**, creating some difficulties in meaningfully combining certain variables and likely increasing discrepancies. Additionally, different approaches will have varied ability to detect mutations dependent on factors such as depth, GC content, variant allele frequency and tumour purity. A fourth limitation was the lack of matched germline tissue and absence of confirmed somatic status in some studies. As was shown in **Figure 2-5**, the distribution of VAF in the validated somatic variants against the non-validated variants differed between the two groups. The fact that there was a high density of variants with a VAF close to 50% in the non-validated group could indicate the presence of rare germline variation that is extremely difficult to filter without matched germline tissue. Although the VAF could have been used as a filter to eliminate potential rare germline variants, it was not done since not all studies published VAF data. This leads onto the final limitation which pertains to the quantity of information provided within the published manuscripts. Several manuscripts provided variant lists with detailed annotation, while others provided only high-level amino acid sequences. This meant that certain information such as chromosome start and allele information had to be inferred, increasing the variability of the data and possible errors.

Regardless of these limitations, this review created a valuable dataset and resource to understand what has already been done and what possible future action is required. Variants from WES studies were analysed separately to have an unbiased assessment of the somatic pathways and genes affected. Unfortunately, WGS and WES are often limited by sample size and this is the case

in SMZL with the largest cohort comprised of only 15 patient samples. According to the ICGC, 500 samples are needed to reliably detect genes that are somatically mutated in 3% of a tumour. For rarer tumours, they propose a two-tiered strategy to obtain comparable statistical power, processing a discovery (n=100) and a validation set (n=400). Both approaches may vary based on tumour heterogeneity, background mutation rate and sequencing depth. Considering the reduced sample size and lack of concordance between the five WES studies, it is clear that more unbiased genome-wide analysis is necessary.

Having established the limitations of the dataset, the full set of variants was then analysed for further insights into SMZL biology. The review validated the importance of *KLF2*, *NOTCH2* and *TP53* in SMZL pathogenesis. *NOTCH2* mutations targeted the C-terminal PEST domain necessary for the regulation of the intracellular domain (NICD)[97], and consequent transcriptional regulation. Two distinct clusters of mutations were found in *KLF2*, one consisted mostly of missense mutations flanking the ZF1 domain involved in DNA recognition, and the second in the activation domain. Several of these mutations, particularly those in the zinc finger domain have been shown to hinder the ability of *KLF2* to suppress NF-κb induction by upstream signalling pathways[54]. *TP53* was recurrently mutated, supporting the critical role this gene plays in cancer and more specifically in SMZL. As is observed in other mature B-cell tumours, mutations in *TP53* were clustered in the DNA binding domain, where they lead to protein dysfunction. This review also confirms the importance of genes that interact with the NF-κb pathway; with mutations in *TNFAIP3 (13%)*, MYD88 (8%), *TRAF3* (8%), *CARD11* (5%), *IKBKB* (4%), and *BIRC3* (4%). Most notable, *TRAF3*, which has not been considered a significant player in SMZL biology, was mutated across studies and warrants further study at the molecular and functional level.

KMT2D was an unexpected gene to find at such a high mutational frequency occurring in 9% of SMZL cases. *KMT2D* is also targeted by recurrent mutations in follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLBCL), where it functions as a tumour suppressor, promoting lymphomagenesis in murine models[98]. Another unexpected gene was *IGLL5*, where recurrent mutations have also been identified in CLL, linked to canonical activation induced-cytidine deaminase (AID) activity with a mutation pattern clustering around the transcription start site within the first intron[99]. AID induces clustered mutations in the immunoglobulin loci as well as some off-target regions, potentially an underlying cause of oncogenic mutations often seen in B-cell malignancies[100]. *IGLL5* is homologous to *IGLL1*, critical for B-cell development and hints at having biological importance in CLL[99]. In the pooled dataset, we found *IGLL5* mutations in 31/222 cases (14%). *IGLL5* mutations have also been reported in other marginal zone (nodal and extranodal) and lympho-plasmacytic lymphomas[88]. Although we do not have sufficient sequencing

information (WGS) to determine the mutational signature underpinning *IGLL5*, it is likely similar to the situation in CLL, a consequence of off-target AID activity.

The DISCOVER algorithm identified three pairs of genes (*KLF2* and *IGLL5*, *TP53* and *NOTCH2*, and *TP53* and *KLF2*) which were significantly mutually exclusive, suggesting possible disease subtypes in SMZL. It was not possible to do any further correlations with these results as the individual phenotypes were not published with the manuscripts.

2.6 Conclusion

The database created here represents a critical community resource as currently SMZL tumours are not included in the IGCG and TCGA, and only 23% of reported SMZL variants are included in COSMIC. Moreover, evidence is provided that the study of SMZL genomics requires expansive unbiased whole-genome mutational analysis to fully unravel the somatic landscape of the disease. This systematic review confirms the importance of *NOTCH2, KLF2* and *TP53,* and adds evidence to the importance of several other genes, such as *TNFAIP3*, *TRAF3*, and *KMT2D*, that will guide future molecular screening and functional experimentation and provides a resource for the interpretation of future genomic studies in SMZL.

Chapter 3 **Methodology**

3.1 Patient cohorts and sequencing of NGS libraries

3.1.1 Jaramillo cohort

Our primary cohort of 146 splenic marginal zone lymphoma samples, all meeting established diagnostic criteria[37], were obtained from 11 international collaborating centres (Spain, Greece, Italy, France, Germany, Sweden and the United Kingdom). Tumour DNA was extracted from peripheral blood [n=97], spleen cells [n=13] or bone marrow [n=2] however, for some cases samples were sent as DNA and the material type from which they came from is unknown [n= 34]. Informed consent was obtained from all patients in accordance with the Helsinki declaration and regional research ethics. Prior to DNA extraction (DNeasy blood and tissue kit, Qiagen), the CD19+/CD45- SMZL cells were purified using the EasySep Human B Cell enrichment kit without CD43 depletion (StemCell Technologies). Tumour purity of greater than 80% was confirmed with fluorescence-activated cell sorter (FACS) analysis.

Samples were analysed with a bespoke Agilent HaloPlex HS Target Enrichment system that enriched 383.74 kb of genomic DNA for 62 genes and genomic regions, designed with SureDesign (https://earray.chem.agilent.com/suredesign/). The gene design resulted in 98.95% *in silico* coverage of selected regions. 50 ng genomic DNA from each patient was digested using 16 restriction enzymes. The restriction digests were hybridised to the HaloPlex Probe Capture Library, with an Indexing Primer Cassette, which incorporates Illumina sequencing motifs and barcoding indices into the targeted fragments. The biotinylated target DNA-HaloPlex probe hybrids were captured and target-enriched using PCR amplification before purification using AMPure XP beads (Beckman Coulter). Subsequent to quantification, samples with different indices were pooled, in preparation for sequencing. A final concentration of 1.8 pM of enriched target DNA, with a 1% PhiX spike as an internal control, was sequenced, using 150 bp paired end sequencing on the Illumina NextSeq. Dr. Helen Parker designed the HaloPlex enrichment kits and performed all wet laboratory work described. David Oscier reviewed the cases and confirmed SMZL diagnosis.

Our primary cohort was sequenced in five batches throughout the span of two years (2017-2019). Within each batch, the SMZL samples [n=146] were sequenced alongside other B-cell malignancies which were also processed and used for development of the bioinformatic methods. 32 samples were sequenced more than once due to low quality or because they had been taken at different timepoints or from different tissues. For those sequenced multiple times, results from

different sequencing runs were not combined and the most appropriate run or sample was chosen according to the following criteria:

1. If samples came from the same patient, timepoint and tissue, the sample with the highest coverage was chosen. Details on how coverage was calculated can be found in **section 3.5.2**.

2. If samples came from the same patient but different timepoints, the first time point was chosen.

3. If samples came from the same patient but different tissues, the most informative sample was used (spleen > bone marrow > peripheral blood > other).

Table 3-1 details samples that were sequenced multiple times and whether they were used for analysis. Batch 2 was a particularly low-quality batch and samples within this batch [n=23] that had material available were re-sequenced as batch 4. Batch 4 had much higher quality than batch 2, hence all samples in batch 4 were used for analysis unless otherwise stated in **Table 3-1**

Table 3-1. Description of duplicate samples. Quality of sample is described by the mean target coverage and percentage of target bases that were covered at least 15x.

Duplicates	Sample ID	Batch	Mean target coverage	% Target bases > 15X	Tissue	Year of biopsy	Included in analysis
1	L104_14	1	194	89	-	-	Excluded
	1_S1	2	197	89	-	-	Excluded
	1_S1	4	316	91	-	-	Kept
2	13_MUT	3	91	87	-	-	Excluded
	1_MUT	3	169	90	-	-	Kept
3	L049_09_30	1	237	91	Blood	-	Excluded
	L049_09_31	1	609	92	Spleen	-	Kept
4	L098_13_S59	1	367	91	Skin	-	Kept
	L098_13_S60	1	95	86	Skin	-	Excluded
5	XXI_S38	5	111	82	-	-	Excluded
	30_MUT	3	175	75	-	-	Kept
6	Pangalis_35	3	137	85	-	-	Excluded
	21_S12	1	352	90	-	-	Kept

3.1.2 Parry cohort

The Parry cohort consisted of 175 SMZL patients, from 8 centres across Europe, all meeting established diagnostic criteria[37]. DNA was extracted from peripheral blood [n=135], bone marrow

[n=22], spleen [n=17], or lymph nodes [n=1]. Mantle cell lymphoma (MCL), Splenic/leukaemia unclassifiable (SLLU) and splenic diffused red pup lymphoma (SDRL) cases were excluded using FISH, conventional cytogenetics, and splenic histopathology. Transformation events were diagnosed histologically. Although results from this cohort have already been published[71], we applied new bioinformatics tools to update the results and it its integration was key in adding power to our analysis.

Similarly, for the Parry cohort samples were analysed with a bespoke HaloPlex Target Enrichment system (Agilent Technologies) that enriched 2.39 Mb of genomic DNA for the coding regions of 768 genes, designed with SureDesign (https://earray.chem.agilent.com/suredesign/). The gene design resulted in 98.58% *in silico* coverage of selected regions. Library preparation was performed using the BRAVO automated liquid handling system (Agilent) according to the manufacturer's instructions. 225 ng genomic DNA from each case was digested using eight restriction enzymes. The restriction digests were hybridised to the HaloPlex Probe Capture Library, with an Indexing Primer Cassette, which incorporates Illumina sequencing motifs and barcoding indices into the targeted fragments. The biotinylated target DNA-Haloplex probe hybrids were captured and target-enriched using PCR amplification before purification using AMPure XP beads (Beckman Coulter). Subsequent to quantification (Bioanalyser High Sensitivity DNA assay kit, Agilent) samples with different indices were pooled [n=22 per sequencing lane], in preparation for Illumina sequencing (HiSeq) of 100 bp paired end sequencing by collaborators at the University of Oxford High-Throughput Sequencing Centre.

3.1.3 CLL4 cohort

The CLL4 cohort is another previously published dataset[101]. However, unlike the Parry cohort, these samples belonged to chronic lymphocytic leukaemia patients and were used to determine the accuracy of our filtering strategy in **0.** This cohort was comprised of 499 patient samples taken at randomisation, diagnosed with iwCLL guidelines. Samples were analysed using a TruSeq Custom Amplicon kit (Illumina, San Diego, CA, USA) that enriched 250 or 50 ng of DNA according to manufacturer's instructions. Prepared libraries were taken forward for MiSeq sequencing, in maximum batch sizes of 20 per MiSEQ run. Dr. Stuart Blakemore conducted the DNA quantification, library preparation and sequencing with collaborators at the University of Oxford.

3.2 Targeted regions across cohorts

3.2.1 Targeted regions within the Jaramillo and Parry cohorts

For the Jaramillo cohort, libraries were prepared using two HaloPlex HS Target Enrichment kits (HS kit 1 and HS kit 2). Combined, the two kits consisted of 57 genes (

Figure 3-1), chosen due to their high mutation frequency in previous SMZL studies as well as other B-cell malignancies. 102 samples were run on HS kit 1 which consisted of 55 target genes. While 44 samples were run on HS kit 2, consisting of 51 target genes. HS kit 2 was an update on HS Kit 1, where poorly mapped regions were removed, and other genes of interest included (

Figure 3-1). Dr. Helen Parker was responsible for the redesign of HaloPlex HS kits and detailed kit design can be found in **Supplementary Table 3**. The HaloPlex kit (historical kit) used in the Parry cohort targeted 768 genes, however samples were processed so results would include only those regions found in both HS kit 1 and HS kit 2.

Figure 3-1 shows the overlap between HS kit 1 and HS kit 2 in the orange square. Genes in red font were those not targeted in the historical kit.

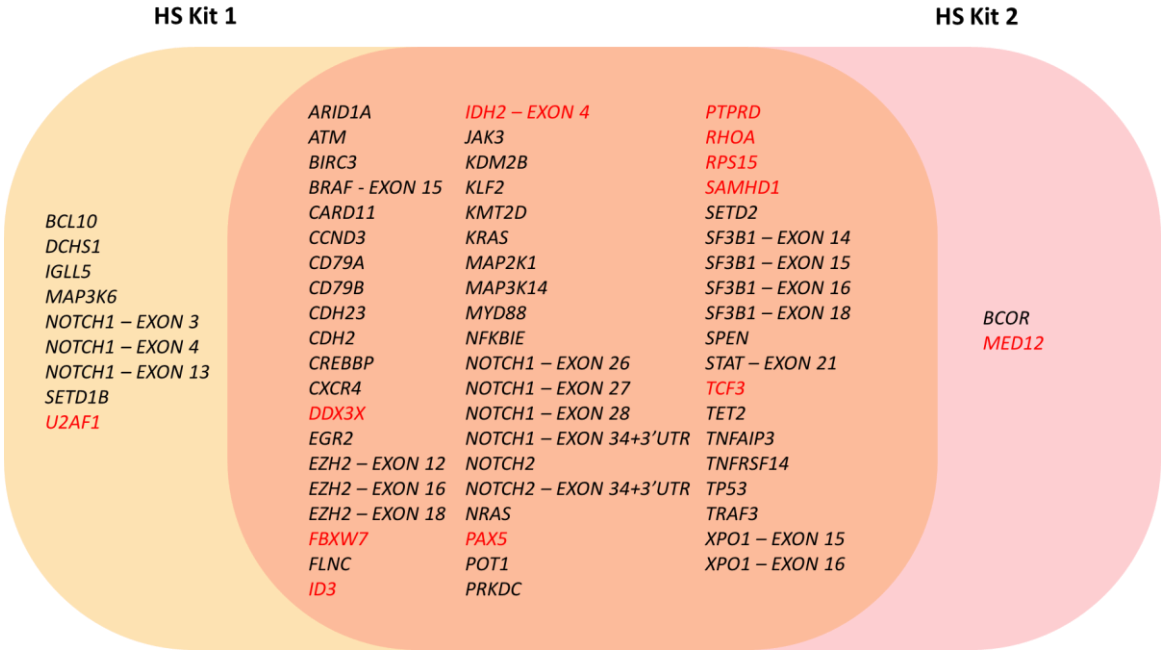


Figure 3-1. Overview of genes targeted by HaloPlex HS enrichment kits. Kit 1 was the first HaloPlex HS kit used in for the Jaramillo cohort but was redesigned (kit 2) to remove poorly mapped regions and include other genes of interest. All genes except those in red were targeted by the historical HaloPlex kit (without UMBs) which was used to sequence the Parry cohort.

3.2.2 HaloPlex HS vs HaloPlex

HaloPlex HS (Agilent) is an amplicon-based capture kit that introduces a unique molecular barcode (UMB) to each DNA fragment. The molecular barcodes are used to merge PCR duplicates and

create a high-quality consensus read. This is useful in downstream processes since this higher quality data allows for higher confidence during variant calling. HaloPlex is the previous version of the HaloPlex HS capture kit and does not add UMBs to the library. This meant that for the historical HaloPlex kit, PCR duplicates could not be merged to create a consensus read.

3.2.3 Targeted regions within the CLL4 cohort

Libraries for the CLL4 cohort were prepared with a TruSeq panel that targeted 20 genes including: *ATM, BIRC3, NOTCH1, SF3B1, TP53, MYD88, EGR2, NFKBIE, POT1, SAMHD1, BRAF, FBWW7, XPO1, CDH2, DDX3X, MED12, SETD2, RPS15, CTBP2, MGA, NBEAL2, KRAS, NRAS, HIST1H1E, ZFPM2*.

3.3 Overview of samples used throughout the project

Samples were sequenced at different time points and therefore used at different stages throughout the project. **Table 3-2** describes which samples were used in which process (i.e. bioinformatics pipeline development, analysis) and where they were used.

Table 3-2. Breakdown of samples used per chapter and process . Our primary cohort was sequenced in five batches throughout the span of two years (2017-2019). Within each batch, the SMZL samples [n=146] were sequenced alongside other B-cell malignancies which were also processed and used for development of the bioinformatic methods. Therefore, for the Jaramillo cohort, 'all samples' refers to the SMZL samples in addition to the other B-cell malignancies.

Chapters	Process	Samples used		
		Jaramillo cohort	Parry cohort	CLL4 cohort
Chapter 4 - Optimisation of bioinformatics pipeline to process targeted next generation sequencing data	PipelineV1	batch 1 & 2	-	-
	PipelineV2	batch 1 & 3	-	-
	PipelineV3	batch 1 & 4	-	-
	PipelineV4	batches 1 - 4	-	-
	PipelineV5	all samples	all samples	-
Chapter 5 - Preliminary results of next generation sequencing analysis of splenic marginal zone lymphoma patients	Quality control	all samples	all samples	-
	Analysis	146 (reviewed diagnosis)	-	-
Chapter 6 - Machine learning to distinguish true somatic variants from noise in tumour only NGS	Test set	batch 1	-	-
	Validation	-	-	batch chosen at random (miseq16-005)
	Run through model	all samples	all samples	-
Chapter 7 - Next generation sequencing analysis of splenic marginal zone lymphoma patients	Filtering strategy 1	146 (reviewed diagnosis)	all samples	-
	Filtering strategy 2	146 (reviewed diagnosis)	all samples	-
	Transcript selection	146 (reviewed diagnosis)	all samples	-
	Analysis	146 (reviewed diagnosis)	all samples	-
Chapter 8 -Integration of genomic results and clinical data of SMZL patients	Analysis	146 (reviewed diagnosis)	all samples	-

3.4 Bioinformatics pipeline

The bioinformatics pipeline in which the samples were processed had several iterations and changed throughout the project. **Chapter 4** discusses in detail the steps involved in the bioinformatics pipeline as well as all the changes made for optimisation.

The bioinformatics pipeline as well as the quality assessments described in the following sections were all run on the University's high-performance computing cluster IRIDIS4

3.5 Quality assessment of NGS data

3.5.1 FASTQ quality

FASTQ files are the raw sequencing files used as input for the bioinformatics pipeline (for further details see **section 4.2.1**). FASTQ sequence quality was assessed using FASTQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and merged with the MultiQC[102] tool. If samples were run in batches, each batch was assessed separately.

3.5.2 Coverage

Before data analysis, read quality metrics are assessed to make an informed decision on the downstream processing samples will require. Coverage refers to the number of reads that cover a specific base. Coverage statistics are examined to determine if there are enough bases covering a region to make confident calls and the abundance of poor-quality samples. Somatic variants may have very low variant allele frequencies (VAFs) and identification of these require higher read depth than germline variants for identification with high confidence. Read depth refers to the total number of usable reads or fragments from the sequencing machine. Using the binomial distribution, we calculated that a coverage of 30x is needed to identify variants with a VAF of 0.10 with 95% confidence.

Coverage statistics were calculated on a per-sample, per-gene, and per-region basis, using the BAM files, the regions BED file, and a reference file for gene annotation (hs38.fa) using GATK's DepthOfCoverage v3.7 tool. The BED files were modified to remove alternative contigs not present in the reference files (hs38.fa).

3.5.3 Percentage of similarity between samples

Coding variants from a sample are compared against the coding variants of other samples with an in-house script. The script calculates the percentage of variants a pair of samples have in common and outputs a matrix with the percentage of similarity across all samples being compared. Samples that share many coding variants (high percentage of similarity), could be related, from the same patient or could indicate a large number of artefacts (which have been identified as variants) from the sequencing and processing. If there are no duplicate or related samples within a cohort a large percentage of similarity could also indicate cross contamination of samples.

3.5.4 Conversion of BED files between reference genomes

BED files for all target enrichment kits were constructed using the hg19 reference genome and had to be converted to the hg38 reference for use in the DepthOfCoverage v3.7 tool. The BED files were therefore re-mapped from hg19 to hg38 using the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap).

3.5.5 Inspection of variants the Integrative Genomics Viewer (IGV)

The Integrative Genomics Viewer (IGV) [103] was used to validate variant *in-silico* and the recalibrated BAM files were used as input for visualisation. To determine if a variant was real or an artefact the standard operating procedure for identifying somatic variants published by Barnell et al. was followed[104]. A summary of the SOP can be seen in **Figure 3-2**.

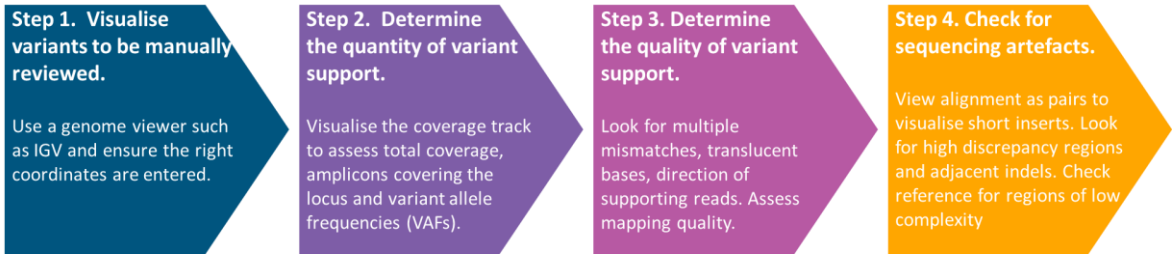


Figure 3-2. Step-by-step of somatic variant refinement via manual review.

Chapter 4 Optimisation of bioinformatics pipeline to process targeted next generation sequencing data

4.1 Synopsis

This chapter describes the steps required to process raw sequencing data to obtain a list of variants. An initial bioinformatics pipeline is described, followed by the modifications applied to optimise it for processing tumour only SMZL samples. The final pipeline described in this chapter was used to process all samples in subsequent sections.

The initial bioinformatics pipeline was developed by previous members of the Genomics Informatics group. Carolina Jaramillo Oquendo optimised the bioinformatics pipeline and this work was overseen by Dr. Jane Gibson. Prof Sarah Ennis, Prof Jon Strefford and Dr Jane Gibson acted as main supervisors and provided guidance in the analysis and interpretation of the data.

4.2 Bioinformatics pipeline overview

To analyse the ever-growing amount of high-throughput sequencing (HTS) data, computational tools and algorithms have been developed to aid our understanding of such complex information. In terms of analysis, each set of HTS data comes with its own challenges depending on what chemistry and what platform was used to sequence it. The processing or bioinformatics pipeline to analyse HTS data can be broken down into five stages: Raw data processing, mapping to reference, variant calling, annotation and filtering (**Figure 4-1**). Only after the raw data has gone through these steps can there be a meaningful analysis of the sequencing results.

Figure 4-1. Data processing workflow for HTS data. Raw sequencing data must be aligned and mapped back to a reference sequence. Following alignment differences between the reference and the sample are identified during variant calling. Once differences (variants) have been identified they are annotated with additional information that can be useful in the analysis. Often the annotated list of variants is further filtered to eliminate false positives.

4.2.1 Raw data processing

After sequencing libraries are prepared, a sequencing instrument will determine the nucleotide bases for each fragment, generating millions of short sequences or reads. Output from the

sequencing instrument is converted into FASTQ files where each read is encoded by four lines that contain read information, the nucleotide sequence and base quality scores.

Before analysis, FASTQ files (raw data) are checked to make sure the sequencing itself was done correctly. This quality control begins by looking at the base calling accuracy. Base calling accuracy is measured by the Phred quality score[105] (Q score) which calculates the probability of a base being miscalled by the sequencing instrument and it is determined by the formula:

$$q = 10 \times \log_{10} p$$

Where *p* is the estimated error probability for that base call. This means that a base call with a quality of 30 has a probability of 1/1000 of being incorrect or a 99.9% accuracy. These quality scores are subsequently used by the different algorithms to identify and exclude artefacts that may have been introduced along the sequencing process.

FASTQ files are assessed to ensure that the data obtained from sequencing is of good quality. FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is the most common tool to assess FASTQ quality, which can identify issues that come from sequencing itself or from the starting library. FASTQC software has several analysis modules which assess per base sequence quality, sequence content, GC content, sequence length distribution, duplicate sequences, and overrepresented sequences (adapters).

4.2.2 Alignment to a reference genome

After FASTQ quality assessment, reads are mapped to a reference genome. The construction of the human reference genome[106] began over 20 years ago and although the project has now completed, refinement and maintenance of this resource is always under constant development (https://www.ncbi.nlm.nih.gov/grc). Currently, the latest major assembly of the human reference genome available is the hg38 assembly and it will be the reference used throughout this project. Since the introduction of HTS technology, there have been at least 70 mapping tools developed, such as Bowtie[107], SOAP[108], Burrows-Wheeler Alignment (BWA)[109,110] and SHRiMP[111]. The tools essentially take a set of reads and aim to map or align them to a reference genome. This is not a simple computational task, as most reads will not have the exact sequence of the reference and there needs to be flexibility to allow alignments with mismatches (**Figure 4-2**).

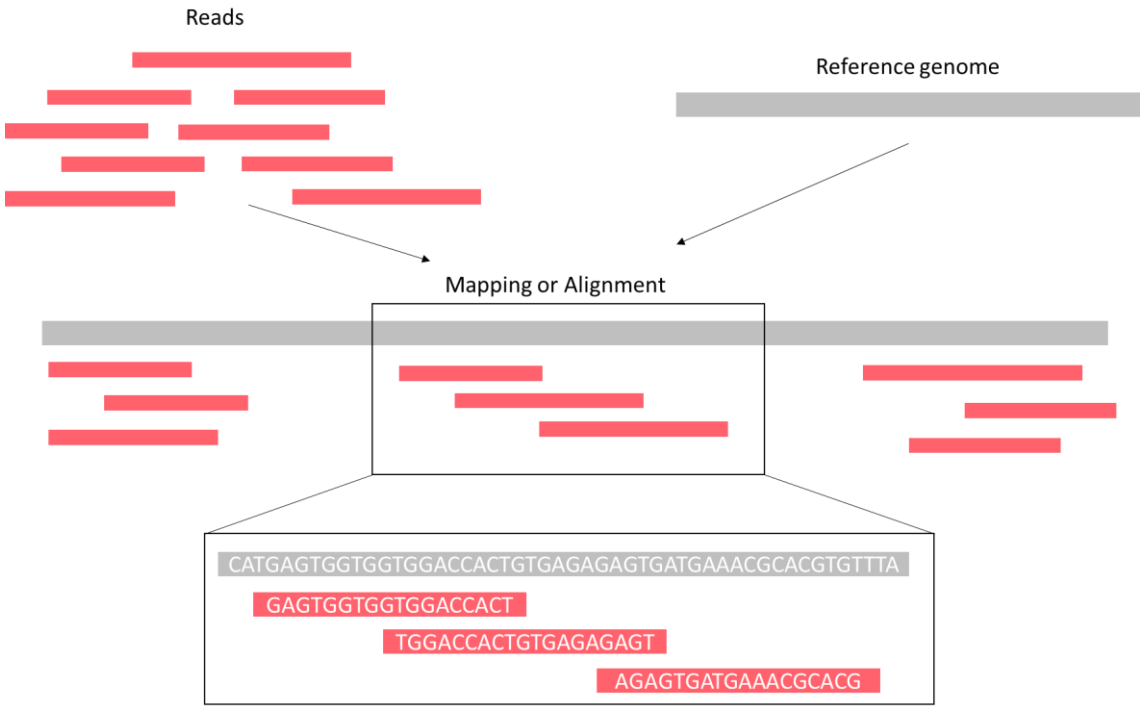


Figure 4-2. Read mapping process. The input is a set of reads and reference genome. The middle row represents the results of mapping. The square zooms in on three reads aligned to different positions of the reference.

For short read alignment, only a few of these tools are routinely used in the analysis of large datasets, mainly limited by the efficiency, in both time and space, of the different tools. This is why Bowtie and mappers that utilise the Burrows-Wheeler Alignment are among the most popular[112]. Alignment tools will output Sequence Alignment/Map format files (SAM), which store the sequencing data in tab-delimited ASCII columns (https://samtools.github.io/hts-specs/SAMv1.pdf). Due to their size SAM files are compressed into BAM (binary sequencing alignment/map) files which tend to be the files used as input for many of the genomic analysis tools.

4.2.3 Variant calling

After alignment, the next step is variant identification or variant calling. This step identifies where the aligned reads differ from the reference and outputs a file in variant calling format (VCF). VCF files contain meta-information lines, a header line, and data lines each containing information about a specific location in the genome. The format also has the ability to contain genotype information on samples for each position[113] as seen on **Figure 4-3**.

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=ClippingRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared a
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), f
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##reference=file:///temp/hgig/EXOME_DATA/cjoln15/HALOPLEX_PIPELINE_HS/completed_runs_SMZL_batch4/10_S10/hs38.fa
#CHROM  POS       ID  REF  ALT  QUAL      FILTER  INFO                              FORMAT          10_S10
chr1    2565143   .   G    C    71610.77  .       AC=2;AF=1.00;AN=2;DP=1607         GT:AD:DP:GQ:PL  1/1:0,1605:1605:99:71639,48
chr1    15873668  .   A    G    16167.77  .       AC=1;AF=0.500;AN=2;DP=962         GT:AD:DP:GQ:PL  0/1:484,475:959:99:16196,0,
chr1    15916468  .   G    T    18.59     .       AC=2;AF=1.00;AN=2;DP=1            GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr1    15921925  .   C    A    37.77     .       AC=1;AF=0.500;AN=2;DP=22          GT:AD:DP:GQ:PL  0/1:19,3:22:66:66,0,591
chr1    15929512  .   T    C    25532.77  .       AC=1;AF=0.500;AN=2;DP=1320        GT:AD:DP:GQ:PL  0/1:607,708:1315:99:25561,0,
chr1    15933318  .   A    G    26630.77  .       AC=1;AF=0.500;AN=2;DP=1655        GT:AD:DP:GQ:PL  0/1:914,737:1651:99:26659,0,
chr1    15938565  .   CTT  C,CT 3394.73   .       AC=1,1;AF=0.500,AN=0.500;DP=254   GT:AD:DP:GQ:PL  1/2:33,123,61:217:99:3432,4
chr1    15938927  .   A    G    12717.77  .       AC=1;AF=0.500;AN=2;DP=700         GT:AD:DP:GQ:PL  0/1:342,352:694:99:12746,0,
```

Figure 4-3. VCF file format. The lines that begin with a '#' contain meta-information and subsequent lines represent a position in the genome in which a variant has been identified.

As with alignment, there are many tools available for variant calling such as GATK Haplotype caller or Mutect2[114], Platypus[115], VarScan[116,117], SAMtools[118] and VarDict[119] but the choice varies depending on the type of sample and sequencing method. More on choosing an appropriate variant caller can be found in **section 4.4.5**.

4.2.4 Annotation of variants

VCF files will contain a list of variants annotated with their location as well as base quality and other sequencing metrics. However, further annotation is needed to aid in the identification of pathogenic variants. Annovar[76] is one of the most widely used annotation tools as it uses up-to-date databases to functionally annotate genetic variants detected from diverse genomes (including human genome, as well as mouse, worm, fly, yeast and many others). Annovar annotates variants against a range of publicly available datasets to add further context and identify the type of variant, amino acid and protein change and affected gene. It also annotates predictive scores to determine if a variant is deleterious and compares the variants against databases of known germline or somatic variation such as gnomAD[120] and COSMIC[11].

4.2.5 Filtering variants into a biologically relevant list

Variant calling generates a lengthy list of all 'variants' found in a sample. If a sample is processed using whole genome sequencing, the number of variants called will likely be in the millions. Likewise, in a targeted panel, the resulting number of variants will vary, depending on the size of the panel, from tens to thousands of variants. This raw list could include somatic variants, germline variants, variants called in off-target regions or they could be sequencing artefacts, especially in somatic samples without matched germline tissue. Filtering is an important step in obtaining a relevant list of somatic variants with the least number of false positives. Therefore,

following annotation variants are filtered to prioritise results into a biologically significant list. More details on filtering strategies are discussed in section **5.3.2 and section 6.2.**

4.3 Challenges of identifying somatic mutations in unmatched tumour tissue

The underlying biological processes that cancerous tissues undergo to become malignant is one of the factors that make identifying mutations in tumour tissue so complex. Tumour samples have a heterogeneous composition of cells, whereby a mutation may be present in all cells (fully clonal) or a subset (sub-clonal). This is problematic in terms of processing a sample, since this heterogeneity implies changing underlying assumptions in algorithms designed to call germline mutations. Germline mutations are expected to have 50% or 100% variant allele frequencies (VAF), while somatic mutations will have a much broader range of VAFs. When sequencing germline tissue, sequencing artefacts are filtered out based on the assumption that they have very low VAFs. However, since variants in somatic samples may also be present at very low frequencies, distinguishing between real low frequency sub-clonal variants and artefacts can be extremely challenging. Furthermore, while somatic variant callers are designed to identify somatic variants considering parameters such as tumour purity, ploidy and VAF among other factors, these are often designed to process matched normal-tumour pairs and are incompatible with unmatched tumour samples. This can be a problem as oftentimes tumour tissue is taken for clinical purposes, and matched germline tissue is not routinely collected alongside the tumour sample. Matched germline tissue is key not only for filtering out germline variation, but it is also helpful in excluding systematic errors in the bioinformatics pipeline. Systematic errors will appear both within the normal and tumour samples, and therefore if a variant shows evidence of being an artefact, we can use the germline to confirm its presence and to see if the same patterns or evidence are observed. Consequently, unmatched tumour HTS data will likely contain many false positives. **Table 4-1** compares the requirements for the successful identification of both germline and somatic mutations and why it is more complex to identify somatic mutations compared to germline.

Table 4-1. Requirements for successful identification of mutations in germline and tumour tissue.

Type of mutation	Characteristics	Requirements for successful identification
Germline mutation	50% or 100% variant allele frequency (VAF).	• Read depth 30x is enough to pick up with confidence. • False positives excluded using VAF. Not expecting anything other than 50% or 100%.
	Mutation is present in all cells.	• Sample purity is not an issue, and does not require other tissue to identify mutations.
Somatic mutation	1-100% variant allele frequency (MAF).	• High reads depths (>100x) needed to pick up low mutant allele frequencies (<10%) with confidence. • Requires more sophisticated variant caller. Mutations with low VAF could be sequencing artefact or real mutation.
	Mutation is present only in tumour cells.	• Tumour purity is important for precise VAF calculation. • Germline tissue also needed to exclude germline variation. • Lack of germline tissue is not always compatible with variant callers

To be able to process unmatched tumour samples and obtain optimal results from our sequencing data, having the appropriate tools and bioinformatics pipeline is key. This chapter aims to discuss the optimisation of a bioinformatics pipeline for unmatched SMZL tumour samples, sequenced using a targeted amplicon-based approach.

4.4 Materials and Methods

4.4.1 Samples

Batches 1- 4 of the Jaramillo cohort were used for the pipeline development/optimisation. Once the final pipeline (pipelineV5) had been established, all batches in the Jaramillo cohort as well as all samples from the Parry cohort were run through pipelineV5 (**Figure 4-4**).

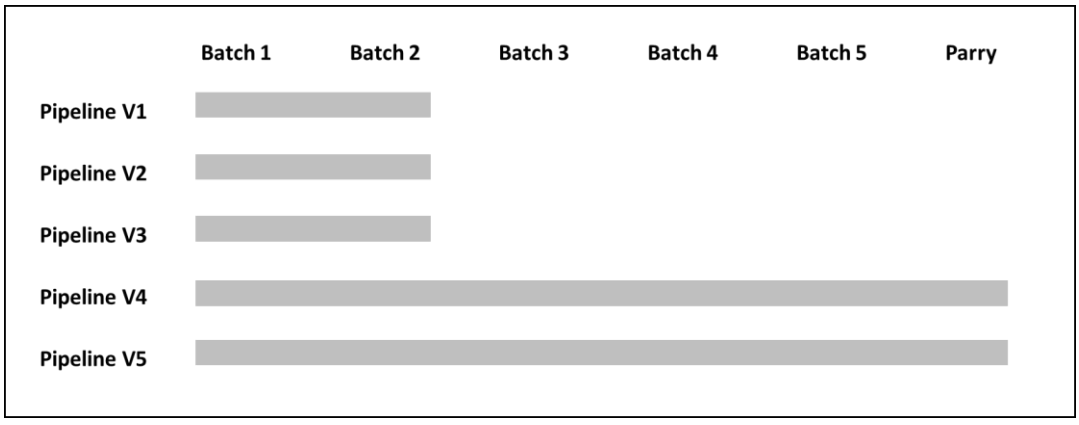


Figure 4-4. Breakdown of samples (batches) used for pipeline development. Batches 1 and 2 of the Jaramillo cohort were used for the development of the first three iterations of the bioinformatics pipeline. After the final pipeline was established all samples in both the Jaramillo and Parry cohort were processed.

4.4.2 PipelineV1 - Baseline pipeline

Due to availability of samples batch 1 (n=62) and 2 (n=54) were the first to be processed on a previously constructed pipeline for HaloPlex HS sequencing data. FASTQ files were run through FastQC for quality control and through SurecallTrimmer v3.5.1.46 to identify and remove adaptor sequences and trim low quality reads (qual < 20). SurecallTrimmer is part of the Agilent Genomics NextGen Toolkit (AGeNT) which provides adaptor trimming and duplicate read removal for HaloPlex HS data (Agilent). BWA-mem was used to align reads to the hg38 reference genome. Samtools v1.2.3[118] converted resulting SAM files into BAMs. Picard v1.97 (https://github.com/broadinstitute/picard) sorted and indexed BAMs. Duplicate reads were marked using LocatIt v3.5.1.46 and resulting BAMs were sorted and converted into FASTQs by Samtools and Picard respectively. LocatIt is also part of the Agilent Genomics NextGen Toolkit. Pear v1.97[121] merges the paired end reads (FASTQ) to increase the length of the reads and improve alignment. The outputs from Pear were assembled and unassembled FASTQs which were then run through BWA-mem and aligned to the hg38 reference genome. BAMs were converted to SAM format (Samtools) and sorted by Picard. Picard merged the assembled and unassembled BAMs and the resulting merged BAM was run through GATKs Haplotype caller v3.7 which called variants via local re-assembly of haplotypes. Following variant calling the resulting VCF files were annotated using Annovar software (v2016Feb01). **Figure 4-5** goes through the steps involved in the bioinformatics processing from raw FASTQ file to annotated variant list ready for prioritisation. This baseline pipeline is defined as pipelineV1.

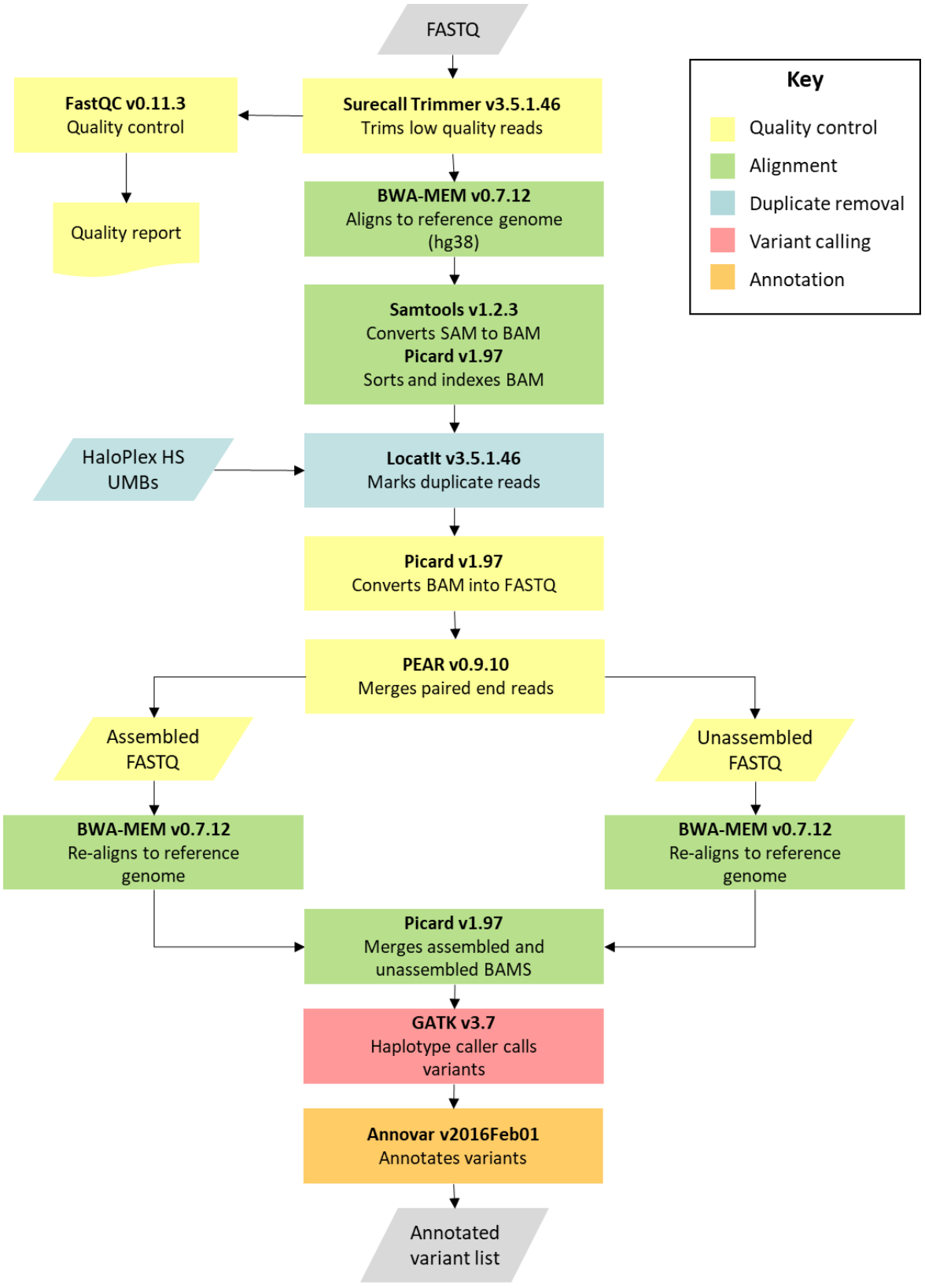


Figure 4-5. Flow diagram of steps involved in the bioinformatics pipeline before optimisation. Raw FASTQ files go through a quality control (yellow) before they are aligned to a reference genome (green) using BWA-mem. LocatIt marks duplicate reads using the unique molecular barcodes from the Haloplex HS enrichment kits (blue). Files are then converted into FASTQ format and paired end reads are merged with Pear. FASTQ files are aligned again using BWA-mem to the hg38 reference genome (green). GATK haplotype caller is used to call variants (pink) and Annovar to annotate with a range of public databases (orange).

4.4.3 PipelineV2 - Marking and merging duplicate reads

One of the main causes of false positives in amplicon-based approaches is the presence of PCR duplicates. The unique molecular barcodes (UMBs) introduced by the HaloPlex HS kit allow for the merging and reduction of these PCR duplicates. LocatIt is an Agilent software created to process the UMB information of HaloPlex HS. This software can mark or merge UMB duplicates and was chosen to handle the UMBs in the pipeline. The LocatIt command line in pipelineV1 is shown in **Box 4.1**.

Box 4.1. LocatIt command line

Java –Xmx19G – jar LocatIt –D –X [temp folder] –t [temp folder] –IB –OB –b [BED file] –o [output name] [input_BAM_file] [index_fastq_file]

-D	Marks duplicates	Java –Xmx19G – jar	Calls the software
-IB	input is SAM/BAM format	-OB	output is SAM/BAM format

The key input in this command line is the –D option, which only marks the duplicate reads. In the second version of the pipeline (pipelineV2), this option was removed, and the reads were marked and merged by the software. To evaluate the effect of changing the LocatIt option on the pipeline, the following analyses were made comparing pipelineV1 to V2:

1. Comparison of the number of variants called per sample against the total FASTQ size
2. Manual inspection of variants in a genome viewer (see **section 3.5.5**).
3. Comparison of the number of reads at different stages of the pipeline (initial reads, after LocatIt and mapped reads).
4. Comparison of percentage of variants shared by samples in pipelineV1 and V2 (see **section 3.5.3**).

4.4.4 PipelineV3 – Removal of adaptors left by SurecallTrimmer

SurecallTrimmer is an Agilent tool (https://www.agilent.com/en/download-agent-tool), which prior to alignment; processes read sequences to trim low-quality bases, removes adaptor sequences and mask enzyme footprints. Both pipelineV1 and pipelineV2 implement SurecallTrimmer prior to alignment.

PipelineV2 reduced the false positives being called by pipelineV1, however, upon inspection in IGV there seemed to be a high number of variants found only at the end of reads. These variants seemed to be likely leftover from adaptors and therefore in addition to SurecallTrimmer,

Cutadap[113] was used to trim the first and last three bases of all reads. The reads were trimmed after they had been processed through LocatIt and Pear (see **Figure 4-5**). The addition of Cutadapt created the third version of the pipeline (pipelineV3).

4.4.5 Variant caller comparison

At first glance, a somatic caller is the obvious choice to process SMZL samples, but as the matched 'normal' counterpart of the SMZL tumour samples was not provided, this created uncertainty as to which variant caller was most appropriate. Furthermore, samples were sequenced with a targeted approach and a variant caller aimed at amplicon specific data would be the most suitable. For our cohort of SMZL samples, the ideal variant caller should take into consideration the following data characteristics: 1) somatic samples; 2) unmatched samples; 3) amplicon-based approach.

The Broad institute have developed a step-by-step workflow for processing HTS data called 'The GATK Best Practices', comprised of several workflows tailored to specific applications depending on the type of data/variation and technology used[122]. In germline detection of variants, GATKs best practice workflow is the 'gold standard', unfortunately, for unmatched somatic samples there is no best practices. Although the 'GATK Best Practices' recommends a tumour/normal pair to be run through the somatic caller Mutect2, there is no consensus as to what the best workflow for tumour only data is.

Currently, only a single variant caller, Pisces[123], meets the criteria for the SMZL data available. There have been various studies[124–130] comparing the performance of variant callers in somatic samples, unfortunately, these benchmarking studies do not agree as to which approach is optimal and results are sometimes contradictory. Furthermore, most of the benchmarking studies[124,125,127,128,131] apply matched tumour samples and there are few that look at tumour only[129,130]. Choosing the best variant caller was not an obvious choice, therefore the germline 'gold standard' variant caller (GATK's haplotype caller) was compared to Pisces, a somatic caller designed for somatic, tumour only, amplicon-based data.

Sample 1_S1 was chosen to compare the variant callers as this sample was sequenced three times, once in batch 1, a second time with a new library in batch 2, and a third time where the library from batch 2 was re-sequenced in batch 4. This meant that there were three separate BAM files for this sample, which in theory should all have the same somatic mutations. Up to variant calling the samples were processed using pipelineV3. The recalibrated BAM files from each batch were run through both GATK haplotype caller[132] and Pisces[123] (variant calling tools). Once variants were called with each tool, the samples were run through Annovar resulting in six total annotated

files (**Figure 4-6**). In each run, the quality, variant allele frequency (VAF), depth, genotype quality and number of variants were compared.

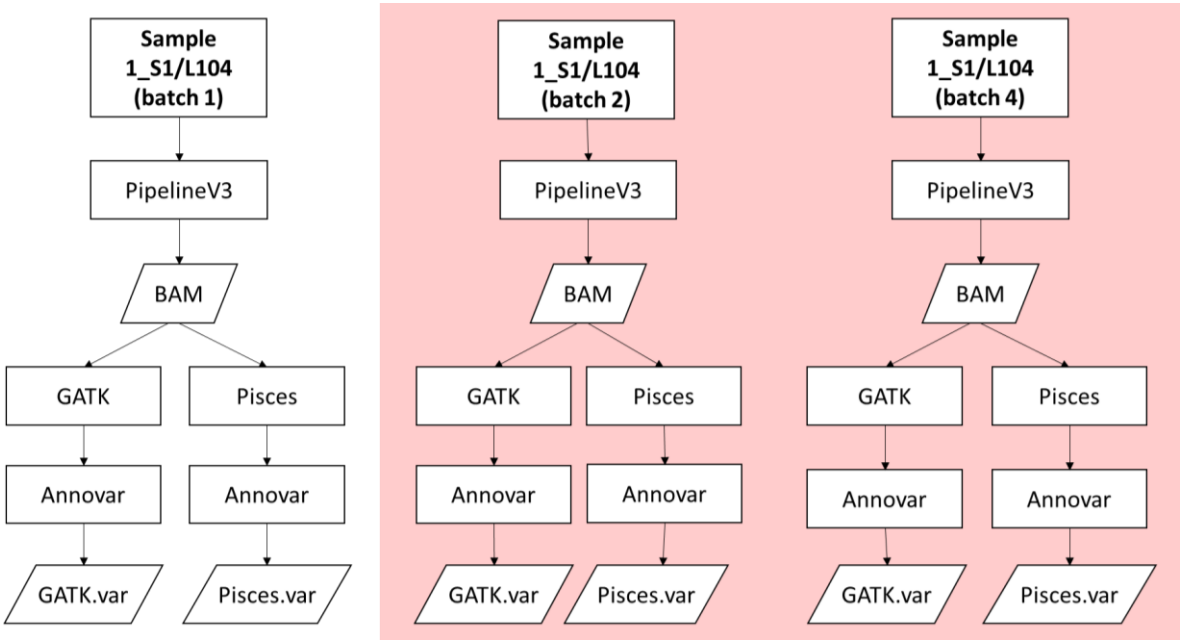


Figure 4-6. Summary of sample processing for variant caller comparison. Sample 1_S1/L104 was sequenced three times in batches 1, 2 and 4. The samples in the pink square were from the same library. In total two libraries were made, and one library was sequenced in two separate sequencing runs (batch 2 and batch 4).

After annotation variants were filtered and validated *in-silico* using IGV (as described in **section 3.5.5**). The filtering strategy for each caller is detailed below:

GATK

1. Include only exonic and or splicing variants
2. Exclude variants found with a frequency > 1% in databases of known germline variation.
3. Exclude variants adjacent to homopolymers (four or more identical consecutive bases)
4. Validation in IGV

Pisces

1. Include only exonic and or splicing variants
2. Exclude variants found with a frequency > 1% in databases of known germline variation.
3. Exclude variants adjacent to homopolymers (four or more identical consecutive bases)
4. Keep those with PASS flag
5. Validation in IGV

Pisces had an extra filter which added a PASS flag to any variant that was above a minimum threshold for depth (30), quality (30) and strand bias (-3.01). Detailed description of how each of the cut-offs are established can be found in the tools Github page (https://github.com/Illumina/Pisces/wiki).

4.4.6 PipelineV4 - Merging duplicate reads (LocatIt parameters)

Marking and merging of duplicates in the pipeline are dependent upon LocatIt using a BED file with all possible start/stops pairs to identify each unique amplicon. However, when the number of unique amplicons covering a specific locus was tracked using an in-house script and inspected in IGV, it became apparent that not all duplicates were being merged correctly. Therefore, the LocatIt command line (**Box 4.2**) was modified a second time and rather than merging the duplicates using a BED file, the program learned all the possible start/stop combinations as it was reading the data (-I option). The application of this modification led to pipelineV4.

Box 4.2. LocatIt command line

Java –Xmx19G – jar LocatIt –X [temp folder] –t [temp folder] –IB –OB –U –I –b [BED file] –o [output name] [input_BAM_file] [index_fastq_file]

-D	Marks duplicates	Java –Xmx19G – jar	Calls the software
-IB	input is SAM/BAM format	-OB	output is SAM/BAM format
-U	unsorted BAM/SAM output (faster)	-I	Incremental. Program learns all the possible start/stop combinations as it is reading the data.

4.4.7 PipelineV5 – Gap penalties

While running the Parry cohort through pipelineV4, it came to light that it was not calling some deletions that had been previously identified. This led to modification of parameters in the alignment tool. The baseline pipeline had been using a gap open penalty of 65 and gap extension penalty of 7 in BWA-mem. Increasing the gap open penalty will make gaps in the alignment less frequent, while increasing the extension penalty will make the gaps shorter. However, the recommended values for the gap open penalty and gap extension penalty are 6 and 1 respectively. Subsequently, pipelineV5 changed the gap open and extension penalties to the recommended values.

4.5 Results and discussion

4.5.1 Marking and merging duplicate reads

FASTQ size and number of variants called in both pipelineV1 and V2 were compared to assess how the number of variants varied across the two pipelines. A successful pipeline should identify a similar number of variants across samples, especially if they were sequenced in the same batch. Panel A of **Figure 4-7** shows plots of FASTQ size vs number of variants called in both pipelineV1. PipelineV1 calls a much higher number of variants in batch 2 (blue) compared to batch 1 (purple) and the coding variants in particular show a very pronounced difference between the two

batches. After inspection in IGV, it became clear that samples in batch 2 had more variants attributed to possible sequencing artefacts (PCR duplicates) that were not being removed compared to batch 1. After merging of duplicates, plots of FASTQ size vs number of variants called in pipelineV2 (panel C & D of **Figure 4-7**) showed a reduction in the variability in the number of variants called across samples as well as reduction of the batch effect seen in pipelineV1.

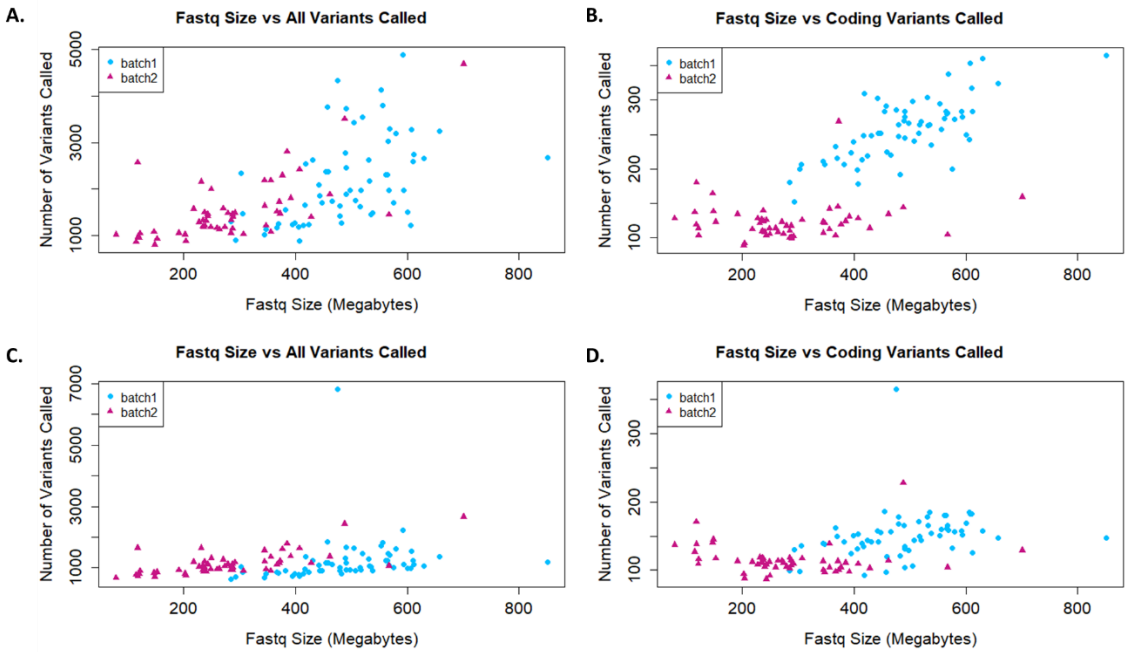


Figure 4-7. FASTQ size vs variants called (pipelineV1 & pipelineV2). **A.** All variants called per sample [n=116] against FASTQ size for pipelineV1. **B.** All coding variants called per sample [n=116] against FASTQ size for pipelineV1. **C.** All variants called per sample [n=116] against FASTQ size for pipelineV2. **D.** All coding variants called per sample [n=116] against FASTQ size for pipelineV2. In pipelineV2 duplicate reads were marked and merged rather than just marked.

The number of reads at different stages of the pipeline was also tracked to validate that LocatIt was merging the duplicate reads. The box and whisker plot in **Figure 4-8** illustrates the number of reads at three different time points in the pipeline (before LocatIt, after LocatIt, and mapped reads). **Figure 4-8** showed values were closer to the mean and had a much smaller range when duplicates were merged with locatIt.

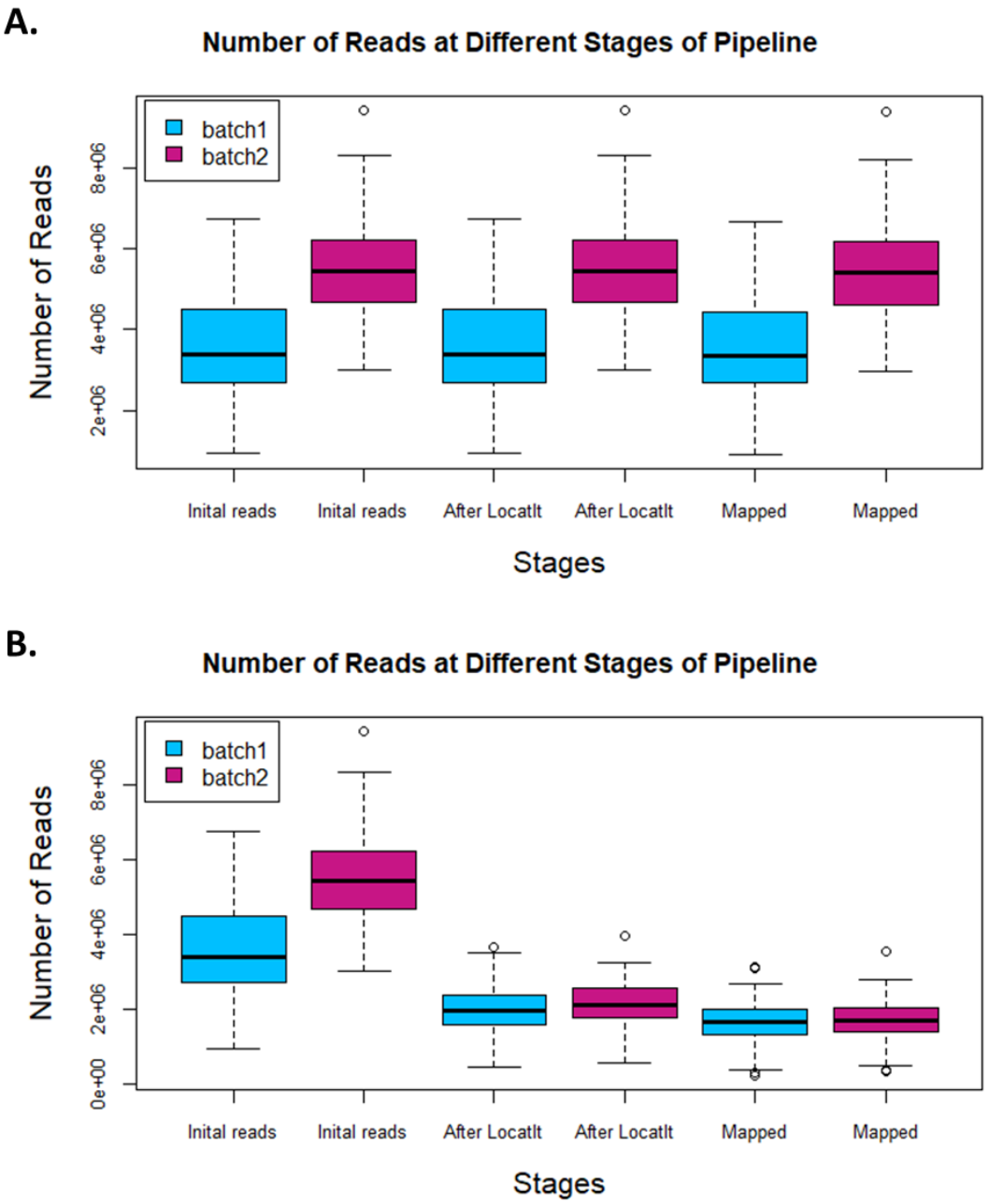


Figure 4-8. Number of reads at different stages of the pipeline. **A.** Number of variants *before* merging of duplicate reads in batch 1 [n=62] and batch 2 [n=54]. **B.** Number of variants *after* merging of duplicate reads in batch 1 [n=62] and batch 2 [n=54].

Figure 4-9 shows the histogram of percentage of similarity between samples. The histograms showed that the percentages tended to have a more normal distribution in pipelineV2 than they did in pipelineV1. Although this histogram imparts modest information, it is expected that the frequency of similarity between samples will have a normal distribution with a narrow peak, unless there are related samples, then a right tail is also expected.

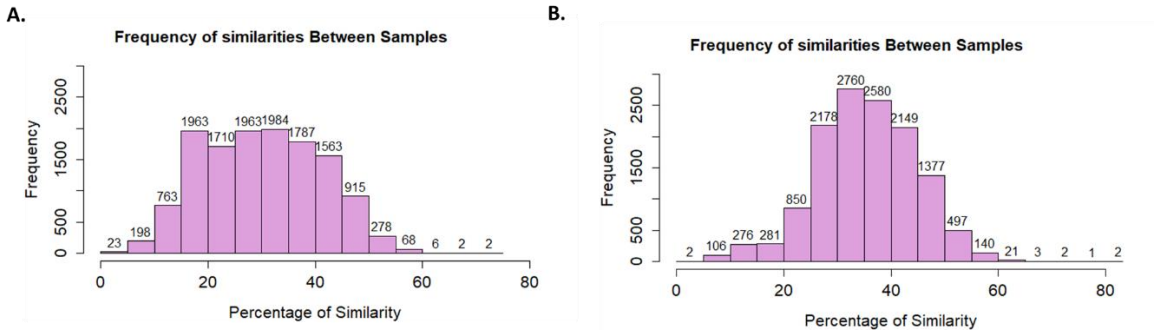


Figure 4-9. Percentage of shared variants across samples in batches 1 and 2. A. Percentage of shared variants across samples run through pipeline V1. B. Percentage of shared variants across samples run through pipeline V2.

The observations made in **Figure 4-7**, **Figure 4-8** and **Figure 4-9** provided evidence that pipelineV2 was successful at reducing noise and as a result, marking *and* merging of duplicate reads was deemed an essential part of the pipeline.

4.5.2 Variant caller comparison

After tailoring the pipeline to remove as many artefacts as possible (merging duplicated reads and removing leftover adapters), variant calling became the next point of focus. It was crucial to determine which would be the most appropriate variant caller as these were tumour-only samples and had been processed with a germline variant caller in pipelineV1, V2 and V3. **Table 4-2** shows the results of the variant caller comparison between GATK and Pisces using sample 1_S1 which had been sequenced in batches 1, 2 and 4. The top half of **Table 4-2** displays the range of values of different metrics available (quality scores are not on the same scale). The bottom half of the table shows the number of variants that remain after each filtering step.

Table 4-2. Variant caller comparison between GATK and Pisces. Results of variant calling with GATK and Pisces using a single sample that was sequenced multiple times in batch 1, 2 and 3. The top half of the table shows the minimum and maximum values for the base quality score, variant allele frequency (VAF), depth and genotype quality. The second half shows step by step, the number of variants remaining when each of the filters is applied to obtain the final list of variants. Pisces had an additional filter which added a PASS flag to any variant that was above a minimum threshold for depth (30), quality (30) and strand bias (-3.01).

Sample 1_S1	GATK			Pisces		
	batch 1	*batch 2*	*batch 3*	*batch 1*	*batch 2*	*batch 3*
Min quality	10	12	10	20	20	20
Max quality	25968	23230	43516	100	100	100
Min VAF	0.076	0.031	0.093	0.025	0.018	0.02
Max VAF	1	1	1	1	1	1
Min Depth	1	2	2	1	1	1
Max Depth	881	750	1227	880	1461	1239
Min GQ	1	2	2	0	0	0
Max GQ	99	99	99	99	99	99
Total # variants	462	753	485	1177	17361	1399
Only exonic and or splicing	63	85	53	233	7849	246
Exclude synonymous	31	54	27	149	6269	155
Exclude > 1% in database	13	36	9	131	6249	137
Exclude homopolymers	10	30	8	123	6165	130
Additional filter	-	-	-	63	1344	56
Validated variants (IGV)	3	3	3	3	3	3

The metrics of most interest were the minimum variant allele frequency (min VAF) and the number of true variants identified by each tool. Surprisingly, GATK was able to call variants with a VAF as low as 0.031, while Pisces was able to call variants with a VAF as low as 0.018. Pisces called all the variants that GATK called (100% overlap) but also called a significantly higher number than GATK in all three sequencing runs of sample 1_S1. Regardless of the number of variants called across each sample, the final number of likely somatic variants validated in IGV was always the same [n=3]. Pisces did not identify additional variants. Additionally, Pisces did not perform as well in terms of specificity. This was evidenced by the sample run in batch 2, that due to poorer data quality (many sequencing errors), Pisces called approximately ten times the number of variants than it did on the same sample in batch 1 and 4.

Although GATK was not designed to call somatic variants and extreme VAFs, it performed well and was able to call all three likely somatic variants validated in IGV. This is likely due to the high depth that characterises the dataset (for further details see **section 5.4.1**). It is possible that with cleaner data containing fewer sequencing errors or better DNA quality, Pisces could outperform GATK. However, for this dataset, GATK appears to have superior power to call variants with low variant allele frequency while excluding a large number of artefacts. The lengthy list of variants Pisces outputs is time consuming to sort through manually, especially if there are over 1000 variants per sample, leaving many opportunities to introduce error and bias. Furthermore, GATK

also includes extra quality metrics that Pisces lacks which could be useful in subsequent filtering strategies. Consequently, the ability to pick up true positives by both callers were the same, hence why GATK was chosen as the main variant caller.

4.5.3 Final optimised bioinformatics pipeline (pipelineV5)

After several iterations, the final pipeline optimised for the SMZL dataset was established (**Figure 4-10**). FASTQ files were run through FastQC for quality control and through SurecallTrimmer v4.0.1 to identify and remove adaptor sequences and trim low quality reads (qual < 20). BWA-mem was used to align reads to the hg38 reference genome using recommended default options. Samtools v1.2.3 converted SAM files into BAMs and Picard v2.8.3 sorted and indexed the BAMs. Duplicate reads were marked and merged using LocatIt v4.0.1 learning all possible start/stop combinations as it was reading the data rather than using the BED file. Resulting BAMs were sorted and converted into FASTQ by Samtools and Picard respectively. Pear v1.97 merged the paired end reads (FASTQs). The outputs from Pear were assembled and unassembled FASTQs which were then run through BWA-mem and aligned to the hg38 reference genome. BAMs were again converted to SAM format (Samtools) and sorted by Picard. Picard merged the assembled and unassembled BAMs and the resulting merged BAM was run through GATKs Haplotype caller (v3.7) which called variants via local re-assembly of haplotypes. The resulting VCF files were annotated using Annovar software (v2016Feb01) with the following databases: The Genome Aggregation Database[77], 1000 Genomes Project[78], NHLBI GO Exome Sequencing Project[79], Exome Aggregation Consortium[77], Kaviar, Haplotype consortium, dbsnf33a, ClinVar, COSMIC, nci60 and our own SMZL reference database (SMZLrefDB). Breakdown of what the databases contain, and additional annotation can be found **Supplementary Table 4**. **Figure 4-10** provides an overview of the final pipeline and modifications made to establish pipelineV5.

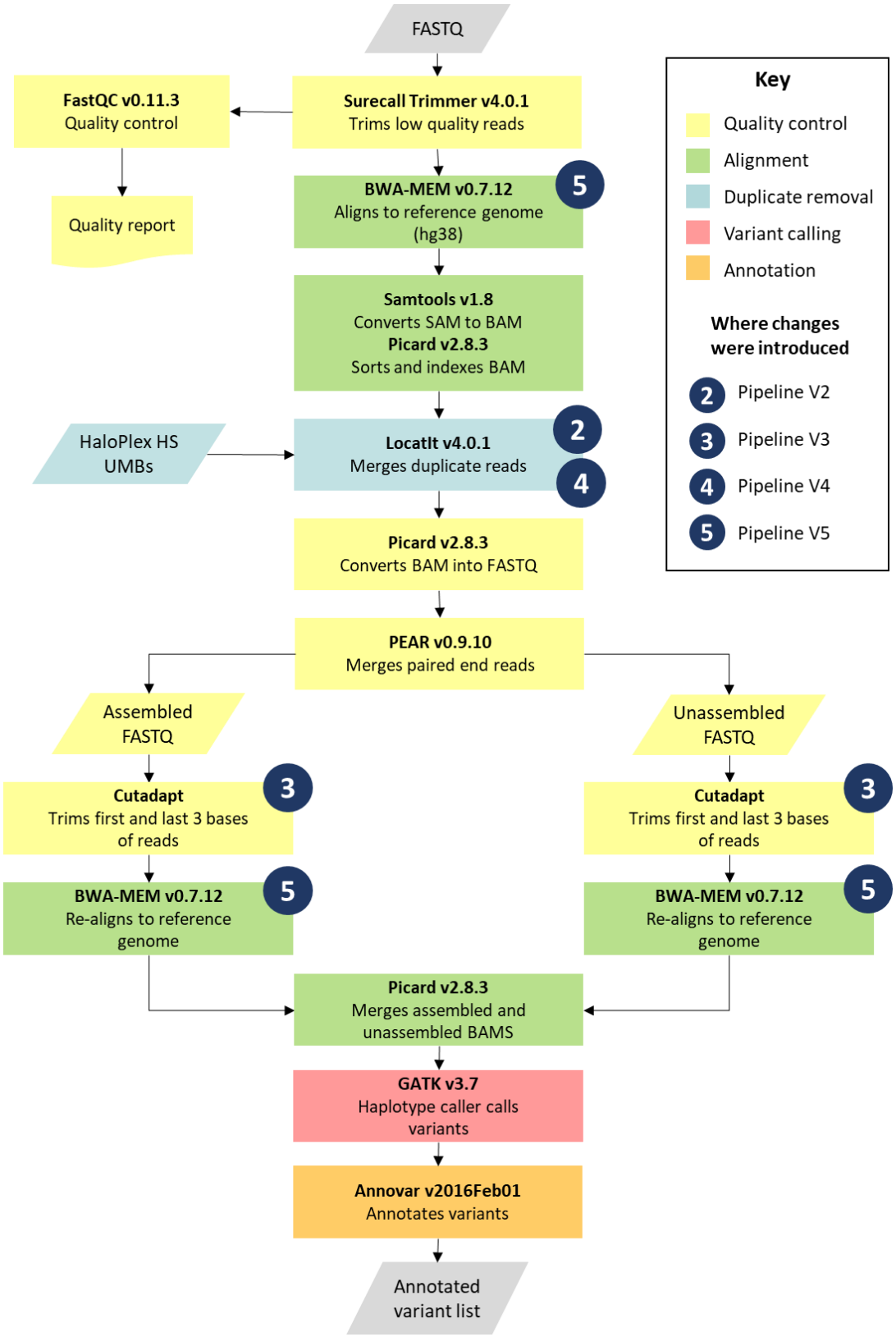


Figure 4-10. Flow chart of the steps involved in the bioinformatics pipeline after optimisation. Raw FASTQ files go through a quality control (yellow) before they are aligned to a reference genome (green) using BWA-mem. LocatIt marks and merges duplicate reads using the unique molecular barcodes from the Haloplex HS enrichment kits (blue). Files are then converted into FASTQ format and paired end reads are merged with Pear. Cutadapt trims the first and last three bases of reads for additional quality control (yellow) and FASTQ files are aligned again using BWA-mem to the hg38 reference genome (green). GATK haplotype caller is used to call variants (pink) and Annovar to annotate with a range of public databases (orange) to end with a list of fully annotated variants ready for prioritisation. Blue circles identify in which iteration of the pipeline modifications were made.

4.6 Conclusion

No bioinformatics pipeline is perfect and new tools or updated versions of the tools are being continuously released. However, at the time when this pipeline was established the latest versions of the tools were used and anyone using this pipeline is advised to check with the most current versions. Our final pipeline (pipeline V5) accounts for the unmatched nature of the data as well as the amplicon-based approach and it was modified to include the least number of false positives.

Chapter 5 Preliminary results of next generation sequencing analysis of splenic marginal zone lymphoma patients

5.1 Synopsis

This chapter focuses on the preliminary results obtained after processing the Jaramillo cohort through the final bioinformatics pipeline. Coverage across samples, genes, and regions is assessed and a brief overview of the variants called is given. This chapter provides the evidence and reasoning behind the filtering strategies developed in subsequent chapters prior to a meaningful analysis of the sequencing data.

Dr. Helen Parker was responsible for the transfer and demultiplexing of the sequencing data to university servers. Carolina Jaramillo Oquendo processed and analysed the data under the supervision of Prof Sarah Ennis, Prof Jonathan Strefford and Dr. Jane Gibson.

5.2 Introduction

Unlike more prevalent mature B-cell malignancies, splenic marginal zone lymphoma (SMZL) is currently precluded from large international sequencing projects, resulting in an incomplete catalogue of tumour associated genomic lesions and mutational processes, drawn from a paucity of published genomic data[49,50,85,51–54,71,72,83,84]. We performed a systematic literature review (**Chapter 2**) that compiled all the next generation sequencing data publicly available to refine the catalogue of somatic mutations in SMZL and identify the strengths and weaknesses in the study of the disease[70]. This would act as a reference which we aimed to build upon by generating and analysing our own sequencing data.

By sequencing the largest SMZL cohort to date we aimed to validate the recurrence of mutations within genes previously established as recurrently mutated such as *KLF2, NOTCH2* and *TP53* as well as mutations within NF-κB, marginal zone B-cell development, NOTCH, and cell cycle pathways. Furthermore, we wanted to use this cohort to gain insights into SMZL biology and establish how or if other molecular biomarkers and survival outcomes associated to these genetic mutations.

After sequencing files (FASTQ) are run through a bioinformatics pipeline the raw variant list resulting from the pipeline needs to be filtered to exclude unwanted variants and sequencing artefacts (for more details see **section 4.2.5**). The filtering strategies will depend upon the quality

of the data, the type of sample (i.e. germline or somatic, matched, or unmatched), and the biology of the disease. Following the optimisation of the bioinformatics pipeline in **Chapter 5** we now had the right tools available to process the sequencing data and begin the downstream filtering process. This chapter will aim to assess an initial filtering strategy and discuss preliminary sequencing results that will guide further filtering methods developed in subsequent chapters.

5.3 Materials and Methods

5.3.1 Samples

146 tumour only SMZL samples, all meeting established diagnostic criteria[37], were obtained from 11 international collaborating centres. Samples were analysed with a bespoke Agilent Haloplex HS Target Enrichment system that enriched 383.74kb of genomic DNA for 59 genes and genomic regions, designed with SureDesign (for further details on this cohort refer to **section 3.1.1**).

5.3.2 Bioinformatic processing and filtering of variants

146 SMZL tumour samples were run through pipelineV5 on IRIDIS4 (see **section 4.5.3**). Once completed, all annotated tab-delimited files were formatted to add a column with sample ID and variant allele frequency (VAF).

After variant calling and annotation, the identification of true somatic variants in unmatched samples requires a filtering strategy to exclude unwanted variants (germline and false positives). When a normal sample is not available for parallel sequencing comparison, germline variation can be excluded by identifying the frequency in which the mutation is found in databases of normal germline variation. If a variant is found at a high frequency in the general population, it is likely that the variant is in fact a germline mutation rather than somatic and so can be removed. This filtering strategy enriches the data for somatic mutations, but there is always risk that rarer germline variation not captured in the population databases persist within the data. Furthermore, many of the reference databases were constructed with sequencing data from studies in which most samples came from a population of European descent. For example, gnomAD includes ~60% European sequences and less than 10% from individuals of African ancestry[133]. Consequently, non-European germline variants are underrepresented in reference databases.

Therefore, within our cohort, the annotated files were filtered to enrich for somatic mutations and exclude variants that were not of interest. This entailed exclusion of variants that had a frequency greater than 1% in any population in databases of known germline variation (see **Table 5-1**).

Table 5-1. List of databases used to filter germline out variation. Annotation was added through Annovar software to the VCF files.

Database name (Annovar)	Description
AF	Genome Aggregation Data (gnomAD) in **ALL** populations v2.1.1[120].
AF_popmax	Genome Aggregation Data (gnomAD) v2.1.1. **Maximum allele frequency across populations** (excluding samples of Ashkenazi, Finnish, and indeterminate ancestry).
AF_male	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in **male** samples.
AF_female	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in **female** samples.
AF_raw	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in samples, **before removing low-confidence genotypes**.
AF_afr	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in samples of **African-American** ancestry.
AF_sas	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in samples of **South Asian** ancestry.
AF_amr	Genome Aggregation Data (gnomAD v2.1.1. Alternate allele frequency in samples of **Latino** ancestry.
AF_eas	Genome Aggregation Data (gnomAD v2.1.1. Alternate allele frequency in samples of **East Asian** ancestry.
AF_nfe	Genome Aggregation Data (gnomAD v2.1.1. Alternate allele frequency in samples of **non-Finnish European** ancestry.
AF_fin	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in samples of **Finnish** ancestry.
AF_asj	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in samples of **Ashkenazi Jewish** ancestry.
AF_oth	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency in individuals that did not unambiguously cluster with the major populations (i.e. afr, ami, amr, asj, eas, fin, mid, nfe, sas) in a principal component analysis (PCA).
AF_othnon_topmed_AF_popmax	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency only in samples that are **not present in the Trans-Omics for Precision Medicine** (TOPMed)- BRAVO release.
non_neuro_AF_popmax	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency only in samples from individuals who were **not ascertained for having a neurological condition** in a neurological case/control study.
non_cancer_AF_popmax	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency only in samples from individuals who were **not ascertained for having cancer** in a cancer study.
controls_AF_popmax	Genome Aggregation Data (gnomAD) v2.1.1. Alternate allele frequency from cases that did NOT come from common disease case/control studies.

Database name (Annovar)	Description
1000g2015aug_all	Alternative allele frequency data in 1000 Genomes Project for autosomes in **ALL population**s. Based on 201508 collection v5b (based on 201305 alignment)[78].
esp6500siv2_all	Alternative allele frequency in **ALL subjects** in the NHLBI-ESP project with 6500 exomes, including the indel calls and the chrY calls (http://evs.gs.washington.edu/EVS/)
ExAC_ALL	Exome Aggregation Consortium Data. v0.3 nonTCGA data **ALL** individuals. Version 0.3. Left normalization done[77].
ExAC_AFR	Exome Aggregation Consortium Data. v0.3 nonTCGA data **AFRICAN** individuals. Version 0.3. Left normalization done.
ExAC_AMR	Exome Aggregation Consortium Data. v0.3 nonTCGA data **AMERICAN** individuals. version 0.3. Left normalization done.
ExAC_EAS	Exome Aggregation Consortium Data. v0.3 nonTCGA data **EASTERN ASIAN** individuals. version 0.3. Left normalization done.
ExAC_FIN	Exome Aggregation Consortium Data. v0.3 nonTCGA data **FINNISH** individuals. version 0.3. Left normalization done.
ExAC_NFE	Exome Aggregation Consortium Data. v0.3 nonTCGA data **NON-FINNISH EUROPEAN** individuals. version 0.3. Left normalization done.
ExAC_OTH	Exome Aggregation Consortium Data. v0.3 nonTCGA data **OTHER** individuals. version 0.3. Left normalization done.
ExAC_SAS	Exome Aggregation Consortium Data. v0.3 nonTCGA data **SOUTH ASIAN** individuals. version 0.3. Left normalization done.

Furthermore, variants with a depth less than 30x were excluded as this is the minimum depth required to identify variants with a VAF of 0.10 with 95% confidence. Variants that were not in target regions such as intronic and intergenic regions (with the exception of *NOTCH1* and *NOTCH2* 3'-UTRs and *PAX* non-coding variants) were also excluded.

The *NOTCH1* and *NOTCH2* 3'-UTRs were included in the amplicon design since these two genes are very similar in structure and function and both contain a PEST domain coded for in exon 34. In chronic lymphocytic leukaemia (CLL), *NOTCH1* 3'-UTR mutations result in a splicing event that removes the PEST domain, which impairs the degradation of the NOTCH intracellular domain (NICD), resulting in the constitutive activation of downstream signalling[134]. Based on their similarity and presence of recurrent truncating mutations within the *NOTCH2* PEST domain in SMZL, it was hypothesised that the *NOTCH2* 3'UTR could also be a target of mutations. *PAX5* non-coding regions were included as recurrent mutations in enhancer non-coding regions in CLL have been associated to the altered expression of the gene[134]. **Table 5-2** lists the criteria for exclusion and the reasoning behind it.

Table 5-2. Exclusion criteria to enrich for somatic variants after variant calling and annotation. The column on the left describes which variants and thresholds were used to filter out variants from the raw list. The column on the right details reason for their exclusion.

Criteria for exclusion	Reason
Intronic or intergenic variants (except *NOTCH1* and *NOTCH2* 3'UTRs and *PAX* non-coding)	Not targets of study
Variants present with a frequency > 1% in any population in databases of known germline variation	Likely germline variation (SNPs)
Variants with a total depth < 30	Low confidence

After filtering, variants were validated *in-silico* using IGV as described in **section 3.5.5**

5.3.3 Quality assessment

Coverage was calculated per-sample, per-gene and per-region as described in section 3.5.

5.3.4 Analysis of NGS data

The annotated and filtered variant list collating the results of the 146 samples (Jaramillo cohort) was used as input into R package maftools[80] to proceed with the data visualisation and analysis. The UCSC genome browser (http://genome-euro.ucsc.edu/index.html) was used to visualise mappability and percentage of GC across regions of the genome and to conduct a BLAT search of sequences. BLAT is a tool designed to quickly find sequences of 95% and greater similarity of length of 25 bases or more.

The systematic literature review described in **Chapter 2** generated a database (SMZLrefDB) of published variants in SMZL which became a valuable reference set when annotating the results from our cohort. To aid with variant analysis, the SMZLrefDB was compared to our cohort to validate pipeline results.

5.4 Results

5.4.1 Quality assessment - Coverage

Per-sample: The mean coverage across target regions in all 146 SMZL samples was 305x, with a minimum of 18x and a maximum of 1107x. The mean target coverage for batches 1, 2, 3, 4, and 5 were 355x, 190x, 156x, 638x, and 169x respectively. **Figure 5-1** compares coverage against the number of variants called before any filtering, as well as how the coverage differed across batches. Results shown in **Figure 5-1** were obtained using the results (**Supplementary Table 5**) of the per-samples script.

The coverage across samples was generally above 100x which is ideal in somatic samples as it will allow for higher confidence for identifying variants with VAFs as low as 0.10. **Figure 5-1** also made it clear that some sequencing runs were more successful than others. We expect to call roughly the same number of variants across all samples regardless of the sequencing run, however batch 3 called a lower number of variants [n=418 variants] compared to the mean number of variants called across all batches [n= 565 variants]. This is likely due to the low coverage in this batch. On the other hand, batch 2 called a much higher number of variants [n=838 variants] than other batches pointing to a potentially high number of false positives.

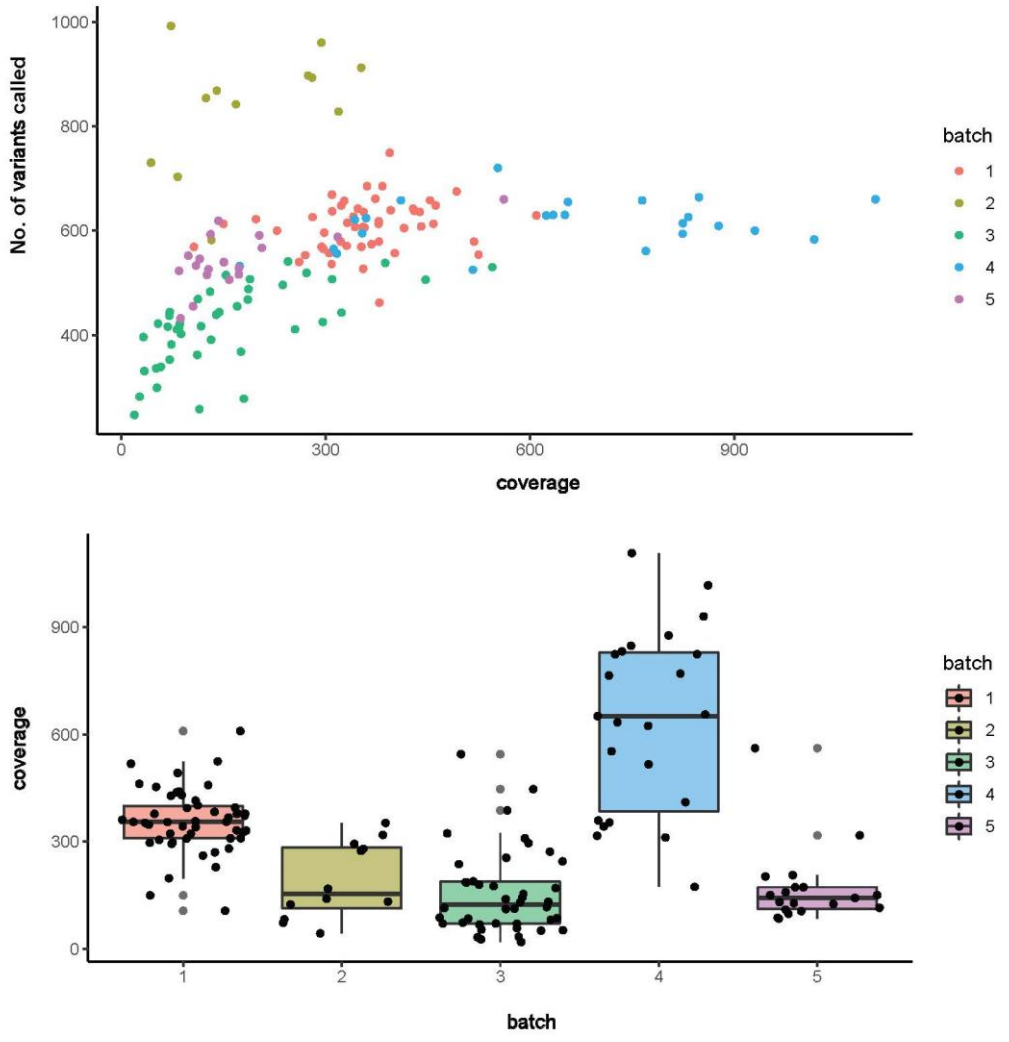


Figure 5-1. Coverage across 146 SMZL samples. The scatterplot (top) shows coverage against number of variants called and are coloured by batches. The boxplot (bottom) shows the distribution of coverage across the five batches. The mean, median and range of coverage across target genes was 305x, 206x, and 18x-1107x respectively.

Per-gene: The per-gene coverage, shown in .

Table 5-3 identified *U2AF1* as an outlier with a mean target coverage of 7x. This was identified initially with batches 1 and 2, but subsequent batches showed consistent results. *U2AF1* was examined using the Integrative Genomics Viewer (IGV) to better visualise the low coverage.

Table 5-3. Per-gene coverage across the Jaramillo cohort.Table displays the mean target coverage for each gene in all five batches. Coverage was calculated using DepthOfCoverage tool.

Gene	Mean	Batch 1	Batch 2	Batch 3 (kit 1)	Batch 3 (kit 2)	Batch 4	Batch 5
ARID1A	290.86	348.18	273.50	216.58	131.91	623.28	151.73
ATM	292.18	299.92	260.95	268.98	137.81	648.29	137.09
BCL10	390.93	312.88	272.66	305.53	-	672.63	-
BCOR	132.82	-	-	-	138.8042	-	126.833846
BIRC3	304.21	318.80	274.63	265.44	149.26	664.77	152.37
BRAF	379.34	380.96	320.76	391.68	188.31	796.03	198.31
CARD11	316.06	392.29	296.90	235.17	140.15	668.96	162.90
CCND3	305.24	369.29	300.80	218.90	140.28	642.59	159.56
CD79A	222.93	287.82	216.67	146.42	115.16	465.79	105.72
CD79B	296.02	384.08	276.83	220.39	135.43	613.24	146.14
CDH23	349.65	440.34	323.77	259.48	170.35	728.05	175.90
CHD2	360.10	393.77	334.69	291.18	154.94	800.33	185.68
CREBBP	326.44	390.23	304.48	242.12	145.29	702.42	174.09
CXCR4	477.26	571.06	417.82	390.75	244.12	1009.92	229.90
DCHS1	411.73	416.79	302.26	242.98	-	684.88	-
DDX3X	260.09	286.19	233.91	196.16	148.74	551.38	144.15
EGR2	347.87	409.08	339.77	223.95	172.55	756.29	185.56
EZH2	333.18	356.55	306.13	292.95	149.48	730.10	163.87
FBXW7	334.02	349.53	301.15	295.01	156.87	741.99	159.54
FLNC	284.12	366.98	270.43	196.35	154.77	584.70	131.49
ID3	313.12	369.82	299.18	211.07	160.89	677.37	160.39
IDH2	544.21	709.59	507.04	396.44	221.88	1164.31	265.97
IGLL5	393.11	361.80	274.21	302.14	-	634.30	-
JAK3	340.78	419.57	313.15	257.45	172.74	712.32	169.44
KDM2B	387.63	489.08	347.74	308.74	219.44	777.23	183.53
KLF2	215.63	255.84	186.03	173.54	148.84	440.75	88.82
KMT2D	363.07	477.56	351.28	241.56	175.54	748.32	184.13
KRAS	317.34	338.78	288.13	292.05	152.02	691.43	141.62
MAP2K1	343.15	386.18	307.75	296.22	166.14	737.37	165.27
MAP3K14	377.92	481.96	343.79	291.54	180.57	785.45	184.18
MAP3K6	438.39	445.55	319.37	266.75	-	721.88	-
MED12	136.06	-	-	-	150.859	-	121.264231
MYD88	362.32	435.81	348.01	245.06	173.66	784.54	186.82
NFKBIE	260.15	321.51	248.11	170.19	140.83	548.32	131.95
NOTCH1	276.32	334.60	245.99	207.16	175.10	569.38	125.70
NOTCH2	351.90	421.75	323.96	275.07	145.96	756.17	188.52
NRAS	328.50	330.62	286.96	303.75	175.50	710.61	163.58
POT1	314.67	324.68	289.60	297.43	129.74	708.57	138.01
PRKDC	353.66	385.85	317.84	316.30	160.62	768.27	173.06
PTPRD	355.50	389.46	327.10	305.54	149.78	794.69	166.42
RHOA	330.42	361.15	330.13	224.04	139.13	763.86	164.23
RPS15	288.69	357.62	257.99	208.73	186.67	591.41	129.71
SAMHD1	301.18	335.92	280.93	249.41	133.31	652.78	154.73
SETD1B	353.29	368.87	263.24	211.67	-	569.38	-
SETD2	341.08	372.51	310.48	276.77	161.85	758.71	166.14
SF3B1	349.53	367.36	305.09	352.37	152.25	749.60	170.51

Gene	Mean	Batch 1	Batch 2	Batch 3 (kit 1)	Batch 3 (kit 2)	Batch 4	Batch 5
SPEN	363.99	437.34	338.75	272.65	166.38	783.83	185.00
STAT3	292.61	340.69	298.39	191.79	95.96	672.88	155.97
TCF3	254.40	304.94	240.18	174.80	148.58	533.23	124.65
TET2	334.79	360.97	305.35	289.79	151.53	731.05	170.06
TNFAIP3	375.17	437.65	345.52	303.87	170.06	804.16	189.76
TNFRSF14	381.56	452.67	329.36	302.28	281.61	755.83	167.64
TP53	279.59	348.89	274.27	190.13	116.10	606.00	142.18
TRAF3	323.17	383.30	291.55	262.18	162.26	683.96	155.76
U2AF1	7.74	4.54	7.08	2.84	-	16.52	-
XPO1	262.64	245.75	227.31	239.75	116.79	606.61	139.65

Visualisation of *U2AF1* in IGV compared to *TP53* in a sample chosen at random can be seen in **Figure 5-2**. A BLAT search of the *U2AF1* gene sequence showed a 99% similarity with gene *U2AF1L5*, meaning the reads mapped equally well to both *U2AF1* and *U2AFL15*. Due to the low mappability of this gene it was excluded from the second HaloPlex HS enrichment kit (HS kit 2).

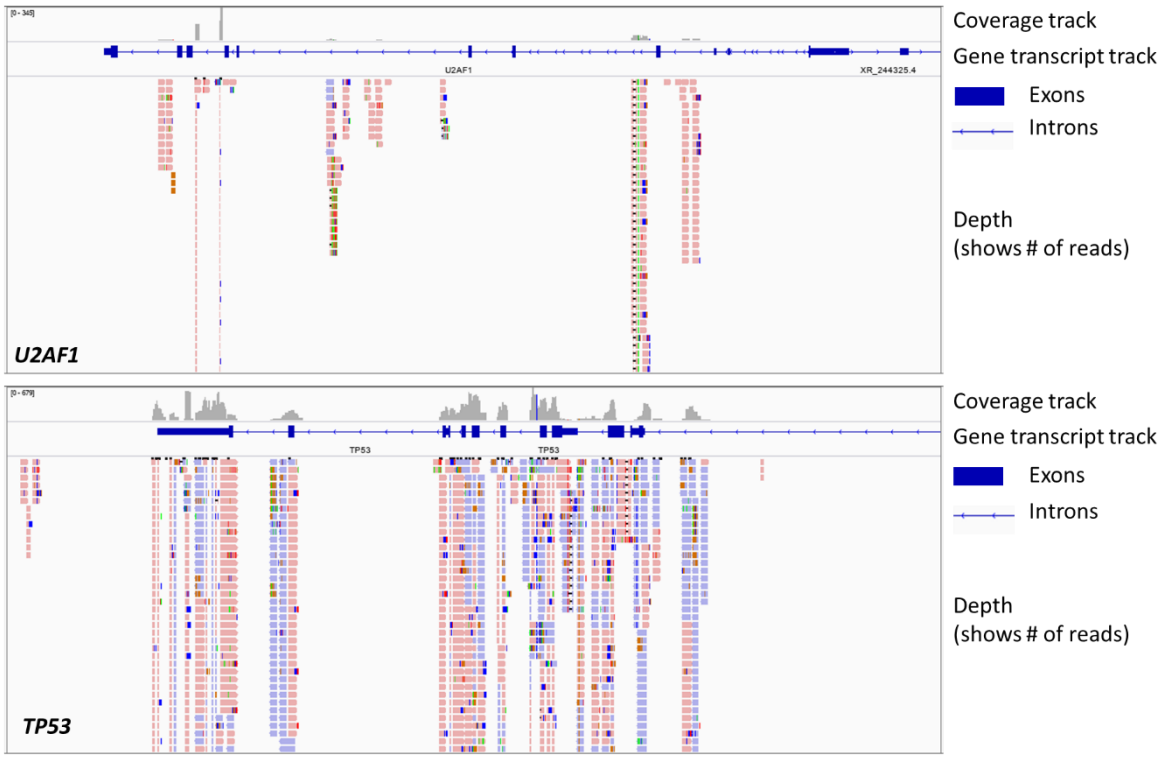


Figure 5-2. IGV view of *U2AF1* and *TP53* in sample 2_S1. The figure shows the reads across gene *U2AF1* (top) mostly mapping to intronic regions and with very low coverage. *TP53* is also shown (bottom) for comparison showing good coverage across the entire gene.

Per-region: Results of the per-region coverage allowed the identification of amplicons that had low coverage. **Table 5-4** lists the targeted regions that had less than 30x coverage and any observations as to why they had such low coverage. Unsurprisingly all the targeted regions or amplicons in *U2AF1* were flagged, as well as amplicons in eight other genes (*NOTCH2, RHOA, POT1, FLNC, ATM, TRAF3, TP53*, and *TCF3*). For the most part such low coverage was explained by low mappability (< 1 UMAP score across amplicon) and a high GC content (> 60% GC content

across amplicon), particularly when the reads fell within intronic regions. UMAP score (0 -1) represents the probability a randomly selected 24-mer, which overlaps with a given position, is uniquely mappable.

The per-region analysis also identified a low coverage region in *POT1* (chr7:124870901-124871051) that has a low GC content (mean GC content across amplicon: 32%) and good mappability (mean UMAP across amplicon: 1). In this case the reason for the low coverage could be attributed to deletions of this gene, as it falls within the 7q.31.33 arm which has been shown to be deleted in SMZL patients.

Table 5-4. Targeted regions with less than 30x coverage across Jaramillo cohort.

Chr	Start	End	Base pairs	Gene	Coverage	Observations
chr1	120029896	120029997	101	NOTCH2	0.00	Exon 2 - highly repetitive region
chr1	120051474	120051789	315	NOTCH2	0.14	Intronic region
chr1	120054110	120054235	125	NOTCH2	17.52	Intronic region
chr3	49360918	49361076	158	RHOA	15.37	Intronic region
chr7	124870901	124871051	150	POT1	20.32	7q31.33 - good mappability
chr7	128857998	128858227	229	FLNC	27.46	7q32.1 - high GC content & limited mappability data
chr11	108347269	108347375	106	ATM	9.69	Exon 59 - 11q22.3 - good mappability
chr14	102777466	102777685	219	TRAF3	5.30	14q32.32 - high GC content
chr17	7666076	7666254	178	TP53	6.51	Intronic region
chr19	1652290	1652615	325	TCF3	18.86	Exon 1 - high GC content
chr21	6484613	6486240	1627	U2AF1	2.00	All regions map equally to *U2AF1L5*
chr21	6486312	6488069	1757	U2AF1	7.29	All regions map equally to *U2AF1L5*
chr21	6489281	6496940	7659	U2AF1	3.15	All regions map equally to *U2AF1L5*
chr21	6499119	6499275	156	U2AF1	0.01	All regions map equally to *U2AF1L5*
chr21	43092946	43094573	1627	U2AF1	3.36	All regions map equally to *U2AF1L5*
chr21	43094645	43096400	1755	U2AF1	10.29	All regions map equally to *U2AF1L5*
chr21	43097612	43105262	7650	U2AF1	3.56	All regions map equally to *U2AF1L5*
chr21	43107441	43107597	156	U2AF1	0.00	All regions map equally to *U2AF1L5*

5.4.2 Analysis of NGS data

The raw list of variants in the Jaramillo cohort consisted of 97,524 mutations across 59 genes in 146 patients. After filtering (off-target mutations, those with a frequency higher than 1% in population databases and those with a depth < 30), the list consisted of 2,800 variants, averaging 19 mutations per individual sample. Of the 2,800 total variants 474 had a COSMIC annotation and 60 were previously reported in SMZL (SMZLrefDB). The five genes with the highest number of individuals with mutations in each gene were *KMT2D (79%), FLNC (62%), CREBBP (55%), SETD1B (39%) and JAK3 (37%).* Although all five genes were present in the SMZLrefDB, none were found in more than 6% of patients (**Figure 5-3**).

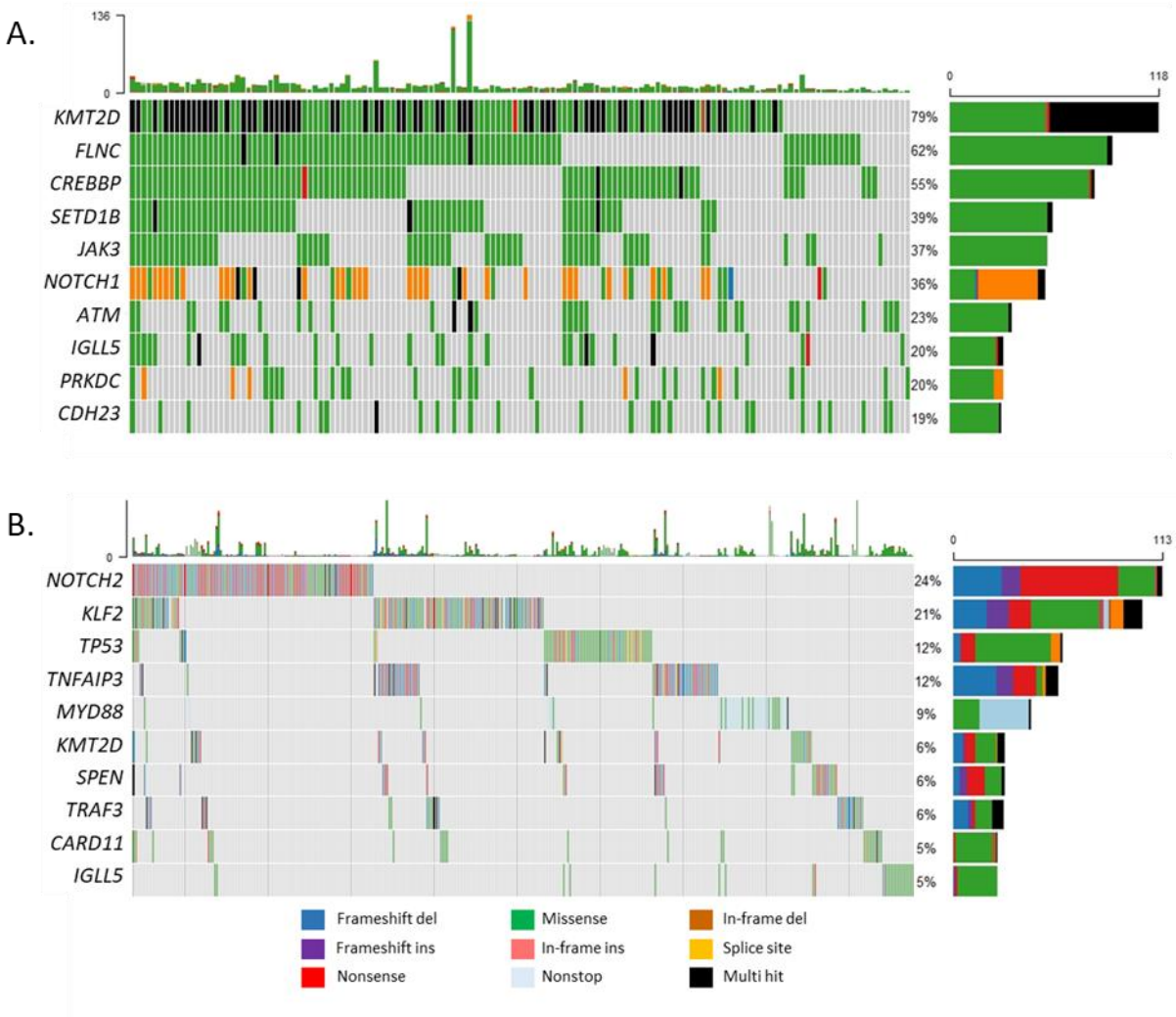


Figure 5-3. Waterfall plot of unfiltered preliminary results (Jaramillo cohort) compared to SMZL database (SMZLrefDB). A. Top 10 frequently mutated genes in preliminary results [n=147]. B. Top 10 frequently mutated genes in SMZL database compiled in the systematic review [n=508].

Most *KMT2D* mutations seemed likely false positives as not only did *KMT2D* have a high recurrence of mutations across patients, but these were multiple hits, meaning there was more than one mutation per patient. This warranted further inspection of *KMT2D* in IGV and the USCS Genome Browser. The three most recurrent mutations (p.774_836del, p.S1107R, and p.I3906Q) were reviewed in IGV, where it was clear that the location of these variants coincided with areas of low coverage that likely introduced bias in the variant calling (**Figure 5-4**). The location of variant p.774_836del (chr12:49051175-49051363) showed few reads were actually mapping back to this region compared to the adjacent reads. A BLAT search of the sequence surrounding the variant (chr12:49051175-49051363) indicated that this region had 100% identity with another region in chromosome 10. Inspection of variant p.S1107R showed a dip in coverage at this loci and surrounding regions (chr12:49,050,247-49,050,292), mostly due to high GC content (UCSC GC percent track not shown in figure). The last variant p.L3906Q, was located in a highly repetitive region (chr12:49,032,779-49,033,023) with low coverage, high number of mismatches across reads and variable mappability.

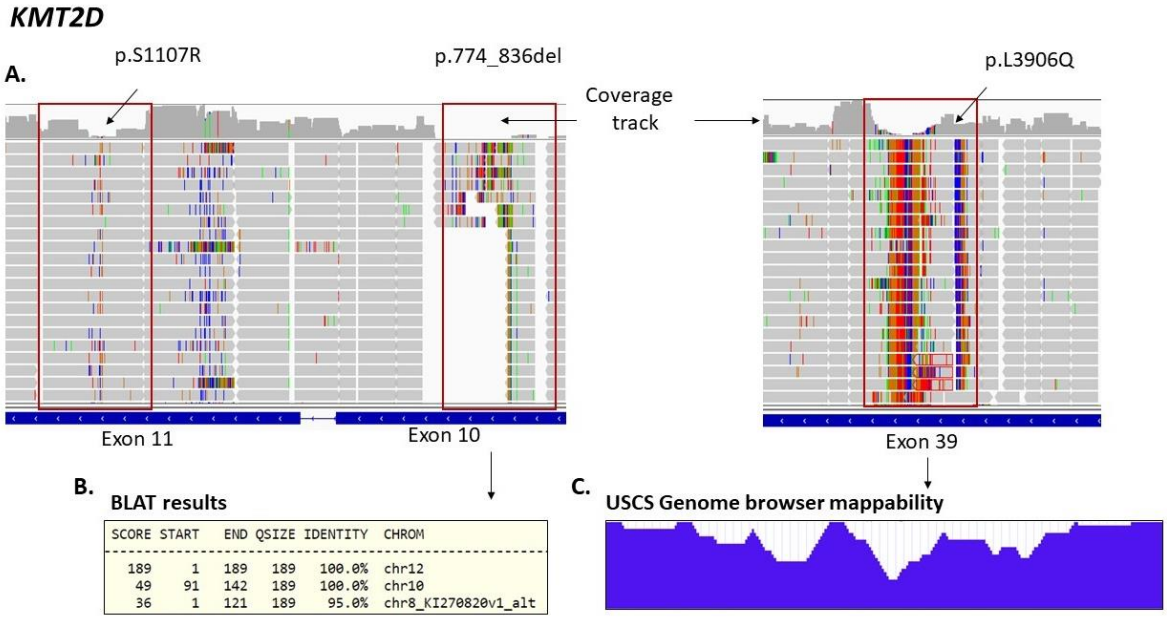


Figure 5-4. Visualisation of recurrent variant in *KMT2D*. **A.** IGV track. The figure shows a screenshot of IGV for the three most recurrent variants in KMT2D (p.S110R, p.744_836del and p.L3906Q). The red boxes surround the regions where the mutations are found. **B.** BLAT results. Displays three main hits for the region (chr12: 49051175-49051363) in which the p.744_836del is found. **C.** Mappability track. Displays multi read mappability for 24-mers of the repetitive region in exon 39 (chr12:49,032,776-49,033,025). These tracks represent the probability that a randomly selected k-mer, which overlaps with a given position, is uniquely mappable. Probabilities are represented in blue where the bottom of the box is 0 and the top 1.

Another indicator of potential false positives was the distribution of transition and transversions in our cohort (**Figure 5-5**). A transition is the interchange of a purine base for another purine base (A to G) or a pyrimidine base for another pyrimidine base (C to T). While a transversion, is the interchange of a purine base (A or G) for a pyrimidine base (C or T) or vice versa. Our SMZL cohort had an overabundance of T > G transversions where more than 50% of the mutations were of this nature. This was not in line with what was present in the SMZLrefDB database in which mutations with a T > G transversion made up less than 5% of the total number of mutations (**Figure 5-5**). Mutations in *KMT2D* alone made up more than 50% of the total number of mutations in this cohort; and of these mutations approximately 60% were T > G transversions. Further inspection of other genes frequently mutated genes showed that the distribution of transitions and transversions was biased towards those genes (*KMT2D, FLNC* and *CREBBP)* that harboured large number of mutations. Excluding *KMT2D, FLNC* and *CREBBP* from the analyses, the distribution of transitions and transversions became more consistent with the distribution in the SMZLrefDB database. This was suggestive that, although there were numerous artefacts that required cleaning, true mutations were also present in the data.

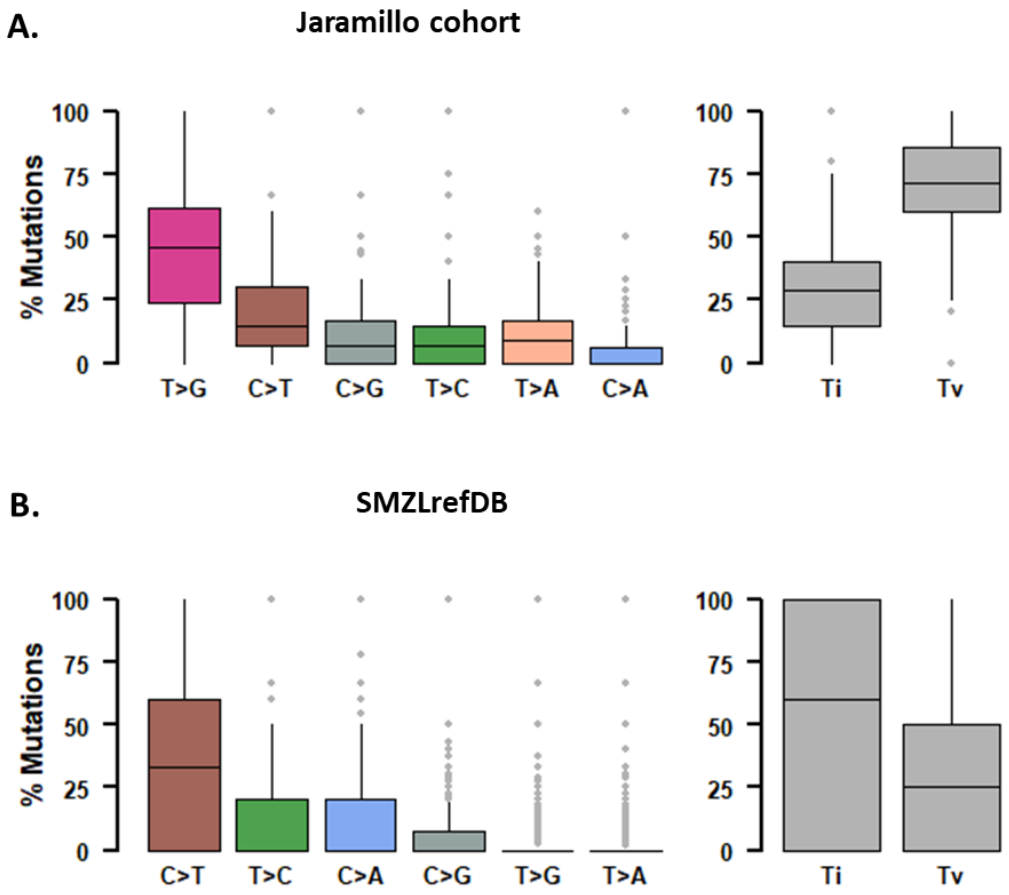


Figure 5-5. Distribution of transition and transversions across SMZLrefDB database and Jaramillo cohort. A. Distribution of transitions and transversions in Jaramillo cohort. The most frequent base change was T>G. B. Distribution of transitions and transversions in SMZLrefDB database. The most frequent base change was C>T.

5.5 Discussion

One of the aims of this project was to create a detailed characterisation of the mutational landscape of SMZL. The first step was the creation of a database (SMZLrefDB) to summarise known genes targeted by somatic mutations but more importantly to establish a point of reference to validate results from our own pipeline and cohort. A direct comparison between result of the SMZLrefDB and our cohort is not ideal since the SMZLrefDB was compiled from a heterogeneous set of unbiased WES and targeted studies, using different chemistries and different bioinformatics pipelines. Samples from our cohort were processed with HaloPlex HS (this chemistry had not been used previously in SMZL) and targeted only 57 genes. This meant a true comparison could only be made for the overlapping targeted genes and disregarding the different downstream processing each sample had. Furthermore, the number of mutations per sample vary depending if the sample was processed through WGS, WES or targeted sequencing. Despite these differences, the SMZLrefDB pointed to specific genes that we expected to find in SMZL patients (*NOTCH2, KLF2, TP53, TNFAIP3, MYD88, KMT2D, TRAF3, CARD11, SPEN, IGLL5*) at a frequency of

at least 5% and identified recurrent variants that were used as positive controls when interrogating variants called in our SMZL cases. The frequencies of the mutated genes within the SMZLrefDB provided an approximation of what was to be expected within our own cohort, and hence why the high recurrence of mutations within *KMT2D, FLNC, CREBBP, JAK3* and *SETD1B* was the first indicator that a substantial fraction of the variants may in fact be artefacts.

The SMZLrefDB also gave us an estimate of the expected distributions of transitions and tranversions, which upon comparison also provided more evidence of false positives. Our cohort had an overabundance of mutations with a T>C nucleotide change, which was not in line with expectations given by the SMZLrefDB. This could be explained by the fact that mutations in *KMT2D* alone made up more than 50% of the total number of mutations and 60% of these were T>G transversions. Consequently, the overall distribution of transitions and transversions were likely biased toward recurrent genes harbouring large numbers of mutations. If this bias was accounted for by excluding those genes, the signatures within our cohort looked more similar to those in the SMZLrefDB. This indicated that although there were numerous artefacts that required future scrutiny and cleaning, true mutations were likely present within the data. This was also evidenced by the 60 variants identified, that had been reported in previous studies.

The lack of matched germline tissue was another obstacle in the identification of true somatic variants as there is no sure way of knowing whether a mutation is germline or somatic. The filtering strategy to enrich for somatic variants was limited in that rare germline variants might still be present in the data and there is the question of whether excluding germline variation potentially excludes variants that impact somatic events. Our approach of excluding variants which were present with a frequency greater than 1% in public databases was conservative. In future filtering strategies the cut-offs could be lower, so rather than 1% frequency using 0.01% and possibly using the variant allele frequency to exclude likely germline variants.

5.6 Conclusion

Preliminary sequencing results were promising in that variants that had been previously reported to be present in SMZL cases were found within our cohort. However, due the mounting evidence of large numbers of false positives it was decided that a detailed analysis of the NGS data was not going to be representative of SMZL biology. Evidence included individuals with high number of mutations within *KMT2D* (79%), *FLNC* (62%), *CREBBP* (55%), *SETD1B* (39%) and *JAK3* (37%) all genes that although have been identified in SMZL patients before, their frequency was never greater than 6%. The high number of multiple hits (more than one variant in one patient) in *KMT2D* led to the identification of a highly repetitive region that was introducing artefacts in the variant calling. And the distribution T>C nucleotide changes, which was not in line with what we

had established in the reference database (SMZLrefDB). This led to the conclusion that we needed to develop a more sophisticated filtering strategy to reduce the noise without compromising sensitivity. Consequently, investigation of the genomic results will be discussed in later chapters once the data has undergone considerable curation.

Chapter 6 Machine learning to distinguish true somatic variants from noise in tumour only NGS

6.1 Synopsis

This chapter will discuss the development of an unsupervised machine learning model that tries to automate variant refinement of tumour only samples. The aim was to reduce the time and potential errors that could be introduced when validating hundreds of variants by manual review. The model will be applied to all batches of the Jaramillo cohort as well as an additional validation chronic lymphocytic leukaemia (CLL) cohort.

Dr. Jane Gibson wrote the Jamp.sh script which was used to extract additional information from the BAM files. Carolina Jaramillo Oquendo prepared the sequencing data, assessed all the quality metrics for input into the model, processed and analysed the results under the supervision of Prof Sarah Ennis, Prof Jonathan Strefford and Dr. Jane Gibson. Dr. Helen Parker and Carolina Jaramillo Oquendo independently reviewed variants in IGV with input from Prof Jonathan Strefford on ambiguous calls.

6.2 Introduction

As outlined in **section 4.1**, identification of somatic mutations in unmatched tumour samples is a computationally complex process. The lack of germline material results in a large number of false positives within the variant call files as was evidenced by the preliminary results shown in **Chapter 5.** Filtering of false positives can be carried out through time-consuming *in silico* manual validation (see **section 3.5.5**), additional sequencing or both. Additional sequencing is not always ideal as it is costly, requires material that could be scarce, and Sanger sequencing for example does not pick up variants with very low VAFs (< 12%). *In silico* manual validation relies upon labour intensive visual assessment of the quality and quantity of variant support and is prone to human error and bias. Manually reviewing or re-sequencing a limited number of samples is a feasible task, but if the same is to be done with hundreds of patient samples, all with hundreds of variants each, this becomes impractical, costly, and a more sophisticated filtering strategy is necessary.

6.3 Machine learning applied to unmatched somatic variant filtering

Machine learning (ML) is a branch of artificial intelligence, interested in the development of tools (computer algorithms) to make sense of complex data[135]. It involves the use of this complex data to build statistical models for predicting and estimating an output based on one or more inputs[136].

Most statistical methods used in ML can be classified as supervised or unsupervised. In supervised learning, for each observation (input) there is an associated response (output) and the idea is to fit a model that relates the response to the observation, with the aim of accurately predicting a response for future observations[136]. Unsupervised learning does not have an output or associated response for each observation, and instead the algorithms try to find complex patterns in the data.

Application of ML is growing in popularity for medical applications and this is certainly the case for cancer genomics. Most applications of ML in cancer genomics have been focused on the creation of new variant callers[137–140] and only three have been designed to identify somatic variants after variant calling[137,141,142]. A recent deep learning approach by Ainscough and colleagues created a ML model (DeepSVR) to automate somatic variant refinement[142]. The authors pointed out that manual review is often carried out as a final step after automated processing, but it is time-consuming, costly, poorly standardized, and non-reproducible. The model was built using 440 sample pairs (266 hematologic malignancies and 174 solid tumours) produced by whole genome, exome, or custom capture sequencing and was comprised of 41,000 variants from 21 studies. The work by Ainscough et al. showed that it was possible to automate somatic variant refinement as it accurately predicted somatic variants which were also confirmed by orthogonal validation[142]. The model constructed by Ainscough et al[142] was very thorough, using over 70 features per variant. Application of this method to the SMZL samples would have been ideal, however, it could only be applied to matched tumour-normal samples. Mahadeo and colleagues developed an optimised tumour-only variant refinement strategy, however, this work focused on eliminating germline variation rather than false positive variants calls[143]. The most appropriate model constructed thus far that could be potentially applied to our data has been the supervised approach taken by Wu et. al who developed a random forest classifier to distinguish sequencing artefacts from true-positives[137]. However, the model was constructed with training data aligned to the hg19 reference genome and the application to other data sets required two negative control BAMs.

6.4 Aims

Preliminary genomic results from the splenic marginal zone lymphoma cohort (**Chapter 5**) showed that the data required a more sophisticated variant filtering or refinement strategy as there was mounting evidence that many of the variants were false positives. Therefore, the aim was to develop an unsupervised machine learning model that identified true and false positive variants in NGS data from tumour only samples. An unsupervised approach was taken since developing a robust classifier requires a large, labelled dataset, ideally with orthogonal validation. However, raw data for such datasets is not easily available.

6.5 Materials and Methods

6.5.1 Samples

To develop the machine learning model, all samples from batch 1 [n=64] in the Jaramillo cohort were used as the test set and all other samples in the Jaramillo cohort (batches 2-5) were used for validation. A second independent validation set comprising of 20 CLL samples from the CLL4 cohort were also used. Detailed description of the samples can be found in **section 3.1.**

6.5.2 Data preparation

Batch 1 was used as the test set to develop the machine learning model. For all variants in all samples of batch 1 an in-house script (Jamp.sh) written by Dr. Jane Gibson was used to extract additional quality metrics from the BAM files (sequence alignment data). The Jamp.sh script outputs a tab delimited file where each row represents a variant and columns include information about the reads covering that loci. Metrics extracted from the BAM files are described below:

- Total number of amplicons covering a base. A high number of amplicons covering a base will give more confidence to determine if the call is real or not.
- Number of mismatches per base pair across all reads covering a specific locus. If a variant is found within a read that has many mismatches to the reference this could indicate poor sequencing quality, especially if the reference reads also have many mismatches.
- Number of reads, which have some softclipping in reference, alternate and other alleles at a specific locus. Softclipping occurs when either side of a read does not match well to the reference and only part of the read is actually aligned while the rest is ignored.

After samples from batch 1 were run through the Jamp.sh script, the tab-delimited output was concatenated to the annotated variant list resulting from pipelineV5. The variant list was then filtered to enrich for somatic variants and exclude off target reads (intronic/intergenic) and low depth variants (depth < 30x) as detailed in **section 5.3.2**. After filtering, variants [n=1,361] were validated *in silico* by myself and Dr. Helen Parker using the IGV genome viewer. Review of variants in IGV was done independently by each reviewer and each variant was labelled TRUE or FALSE according to the criteria provided in the SOP described in **section 3.5.5**. Any discrepancies between labels assigned by Dr. Parker and myself were resolved by Professor Jonathan Strefford.

6.5.3 Feature selection and clustering

K-means is one of the best known unsupervised approaches to cluster data, which partitions the observations intro a pre-specified number of clusters[136], using a set of features for each

observation. K-means clustering was performed on our data using R following feature selection detailed below.

Feature selection: All available quality metrics [n=22] were considered for their potential value in the unsupervised clustering model. The list and description of all metrics available are listed in **Supplementary Table 6**. Assessment was performed manually where quality metrics were excluded for the following reasons:

- Another version of the metric that was normalised was available (e.g. quality by depth rather than quality alone).
- GATK haplotype caller did not output values for metric (e.g. Haplotype score).
- An updated version of the metric was available (e.g. strand odds ratio rather than fisher's exact test for strand bias).
- Metrics that biased germline variants or could possibly split the data into germline clusters (e.g. maximum likelihood expectation for the allele counts).
- Metrics that only accounted for either alternate or reference allele rather than both (e.g. mapping quality).

This manual selection resulted in the use of ten metrics described in **Table 6-1**.

Table 6-1 Model features selected for unsupervised clustering analysis.Features used in the clustering model, with a description of each metric and where each metric was obtained.

Feature	Description	Source
Depth (DP)	Number of reads covering base location	GATK haplotype caller (vcf)
MQRankSum	Rank sum test for mapping qualities of REF vs ALT reads	GATK haplotype caller (vcf)
BaseQRankSum	Rank sum test of REF vs ALT base quality scores	GATK haplotype caller (vcf)
ReadPosRankSum	Rank sum test for relative positioning of REF vs ALT alleles within reads	GATK haplotype caller (vcf)
Strand Odds Ratio (SOR)	Strand bias estimated by the symmetric odds ratio test. Determines if there is strand bias between forward and reverse strands for REF or ALT alleles	GATK haplotype caller (vcf)
Quality by depth (QD)	Quality score normalised by read count	GATK haplotype caller (vcf)
Number of amplicons	Total number of amplicons covering base	Jamp.sh (BAM)
Sum of per base mismatches	Count of per base mismatches in reads containing REF, ALT and other alleles	Jamp.sh (BAM)
Sum of softclipped reads	Sum of softclipped reads in base pair location	Jamp.sh (BAM)
Variant allele frequency (VAF)	Frequency of variant allele in base location	GATK haplotype caller (vcf)

Correlation between features was measured to ensure all ten features were independent using Spearman's rank-order correlation. Once features were selected, variants with no values in any one of the features were excluded and reviewed independently. This often happened with deletions or insertions, where the metric was a rank sum test that takes into account alternate and reference reads. For example, if we want to calculate the BaseQRankSum of a deletion, where it compares the base quality scores of the alternate and reference alleles. Because it is a deletion there are physically no alternate alleles (with a base quality score) and hence the test cannot be performed.

Features were scaled using the scale function in R (z-score) and the optimal number of clusters was determined using the average silhouette and elbow method.

Principal component analysis (PCA): PCA was applied to the scaled features. PCA was performed to understand which features were driving the cluster separation or which features had the greatest impact on variance between the clusters. PCA was run in R using the *prcomp* command on the scaled data. Each variant was drawn on a biplot where the x-axis represented the first principal component (PC1) and the y-axis the second principal component (PC2). Variants were coloured according to k-means result (clustering) and annotated with the manual validation labels where available.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP): UMAP is a dimensionality reduction technique which was used in conjunction with the PCA plot to visualise clustering results. UMAP is an algorithm based in Riemannian geometry and algebraic topology that constructs a high dimensional graph representation of the data and then optimises a low-dimensional graph trying to preserve the local structure of the data[144,145]. In the UMAP plots variants were coloured according to the k-means result (clustering) and annotated with the manual validation labels. UMAP unlike PCA allowed the visualisation of the variation of all features in two dimensions.

Heatmap: Scaled data was visualised on a heatmap. Variants were ordered and annotated according to the manual validation labels, clustering results or both where available.

A complete workflow of the model development can be seen in **Figure 6-1.**

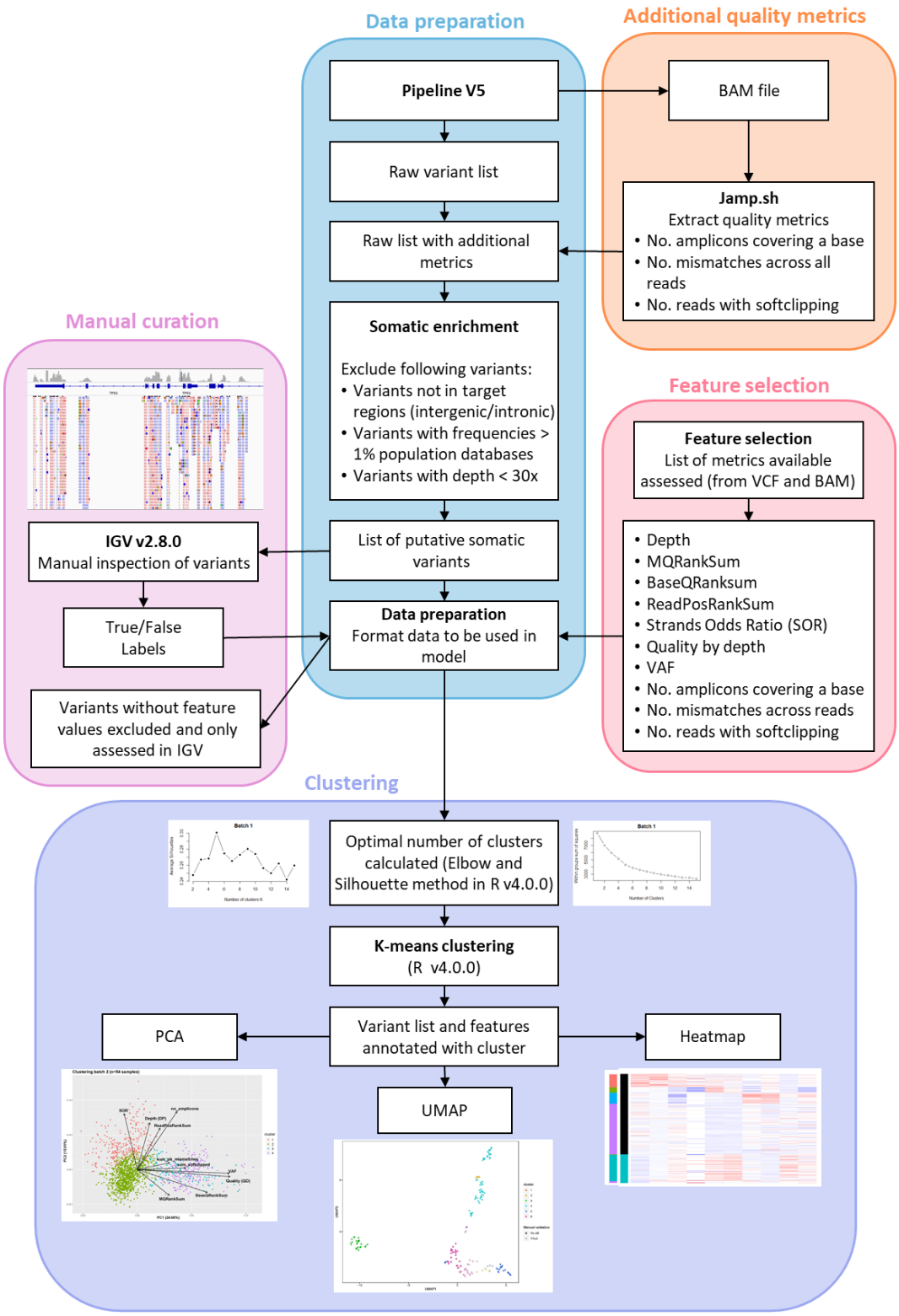


Figure 6-1. ML model development workflow . Development of the clustering model begins by processing all samples through pipeline V5. The BAM files (intermediate files from the pipeline) were run through the Jamp.sh script to extract additional sequencing metrics. Subsequently the raw variant list with the additional quality metrics were validated in IGV and labelled TRUE or FALSE. This manual validation step of all variants was done only for the test (batch1) and validation (CLL) set. At this point all quality metrics available were assessed and those that would be used in the model were extracted. The data was formatted and scaled, and the optimal number of clusters was identified using the elbow and silhouette method. Once the optimal number of clusters was established the data was run through the k-means algorithm and the variants were annotated with their assigned cluster. PCA, UMAP and heatmaps were used to determined which were the TRUE and the FALSE cluster or clusters.

6.5.4 Batches 2-5

The ML model was first implemented on batch 1 and subsequently on batches 2-5 of the Jaramillo cohort after evaluation of performance indicated high sensitivity and specificity. In order to inform whether the ML approach could be applied to the combined set of samples from batches 2-5 or should be implemented once for each batch, the Kruskal-Wallis test was used to determine if the distribution of the ML model features was statistically different between batches. Results of the Kruskal-Wallis test can detect significant differences between any batch pair that could suggest binning would be inappropriate. Pairwise comparison using the Wilcoxon rank sum test was used to determine which batch pairs, if any, demonstrated significant differences in the distribution of the ten selected features (**Figure 6-2**).

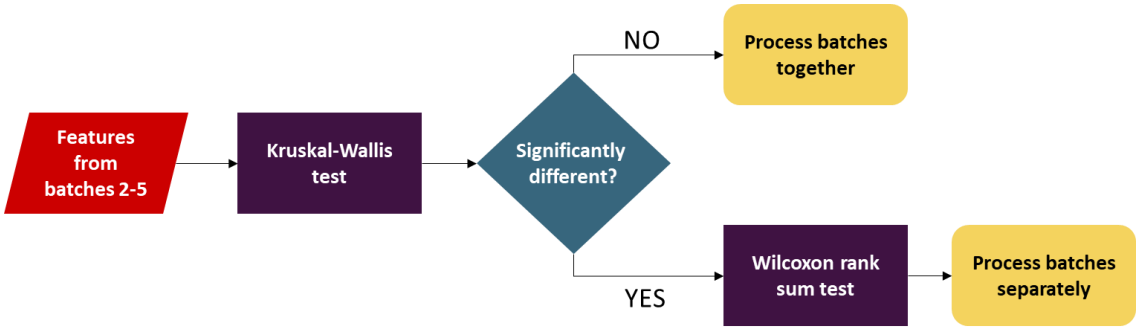


Figure 6-2. Decision tree to determine how batches 2-5 will be run.Each feature was used as input into the Kruskal-Wallis test in R and it was compared across all batches. If any of the batches differed, then the Wilcoxon rank sum test was performed per feature per pair of batches i.e. VAF in batch 2 against batch 3 then VAF in batch 2 against batch 4 and so on.

Manual validation in IGV of all variants in batch 1 was performed to ensure the model was working. In batches 2-5 manual review of variants was only performed in a small subset to confirm correct identification of true and false clusters.

6.5.5 Integration of ML model results to create a filtering strategy for unmatched NGS data

After validation of the model it was then integrated into the analysis of the 146 SMZL cases. The ML model was used on all 146 SMZL samples in the Jaramillo cohort as a triage tool. After clustering, all variants were grouped into TRUE and FALSE clusters and these labels were then used to triage variants into three categories: high, medium, and low confidence variants. High confidence variants were those that fell within the TRUE cluster. Medium confidence variants were those that fell within a FALSE cluster but met one of the following criteria: 1) Variant was found in genes: *KLF2, NOTCH2, NOCTH1* or *TP53;* 2) Variant was an insertion or deletion; 3). Variant was a stopgain or stoploss; 4) Variant had a quality greater than the median batch quality. Low confidence variants were all other variants in the FALSE clusters. High and medium

confidence variants were reviewed in IGV while low confidence variants were excluded from the analysis. **Figure 6-3** illustrates this process.

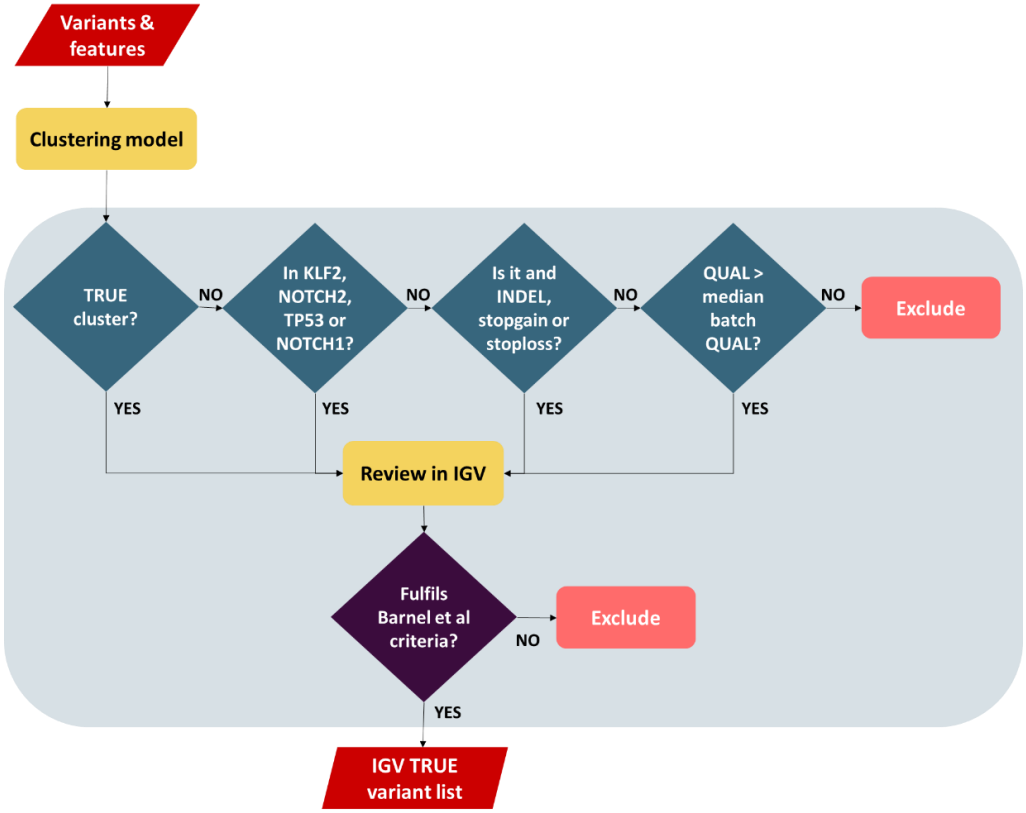


Figure 6-3. Flow diagram of filtering strategy to reduce false positives. After running the variants through the clustering model, these were grouped into TRUE and FALSE clusters. Variants were reviewed in IGV (using criteria given by Barnel et al.[104]) if they were considered high and medium confidence. High confidence variants were all variants that fell within the TRUE cluster. If variants did not fall within the TRUE cluster but they were: 1) Present in genes *KLF2, NOTCH2, NOCTH1* or *TP53;* 2) If variants were insertions, deletions, stop gains or stop losses or; 3) If the quality of the variant was greater than the mean batch quality; they were considered medium confidence variants. All other variants in a FALSE cluster were not reviewed in IGV and not included in the analysis.

Once variants were reviewed in IGV and true variants confirmed, the variant list was used as input for maftools in R to begin analysis of results.

6.6 Results

6.6.1 Feature selection

Spearman's rank order test was used to assess correlation between the available quality metrics. Results of the Spearman's test indicated that there was a significant positive correlation between VAF and quality normalised by depth (QD), r_s=0.89, p <0.001. These two features were expected to have high correlation since both are metrics that have been divided by the total depth. No other features showed high correlation between them, confirming that features or metrics were independent. Results of the Spearman correlation test are shown in **Figure 6-4** for all features across all batches.

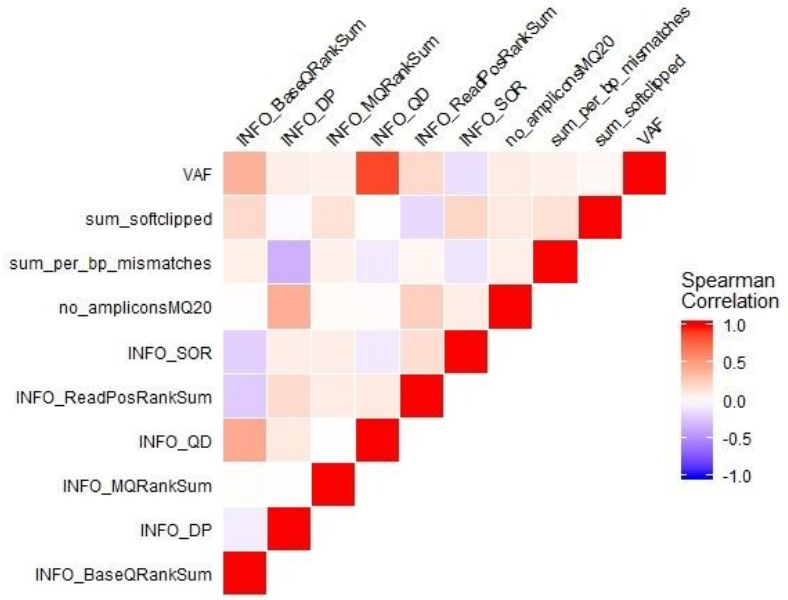


Figure 6-4 .Heatmap of Spearman correlation between features. A value of 0 (white) indicates there is no association between the features. A value close to 1 (red) indicates a positive association between features while a value close to -1 (blue) indicates a negative association between the features.

6.6.2 Batch 1 (test set)

PipelineV5 called 37,567 variants in all 62 samples (including non-SMZLs). After filtering, 1361 variants were left for clustering. **Figure 6-5** shows number of variants included (blue) and filtered out (red) during each step before input into model.

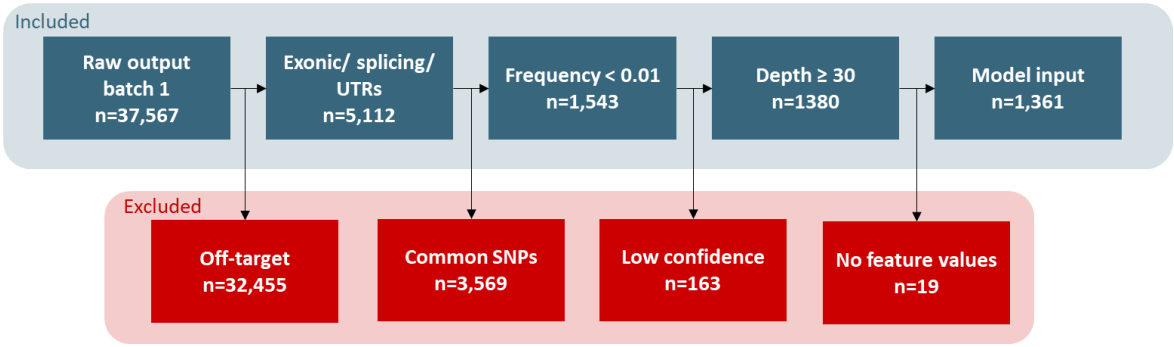


Figure 6-5. Filtering workflow for Batch 1 before input into ML model. The boxes in blue display the number of variants left after each filtering step, while the red boxes display how many variants were discarded.

The elbow and silhouette method both identified the optimal number of clusters as three. Clustering results (**Figure 6-6.A**) show cluster 3 (blue) contains most variants manually labelled as true, while most false variants were grouped into clusters 1 and 2. The UMAP shows discrepancies between manual validation labels and clusters produced by both k-means and by the UMAP.

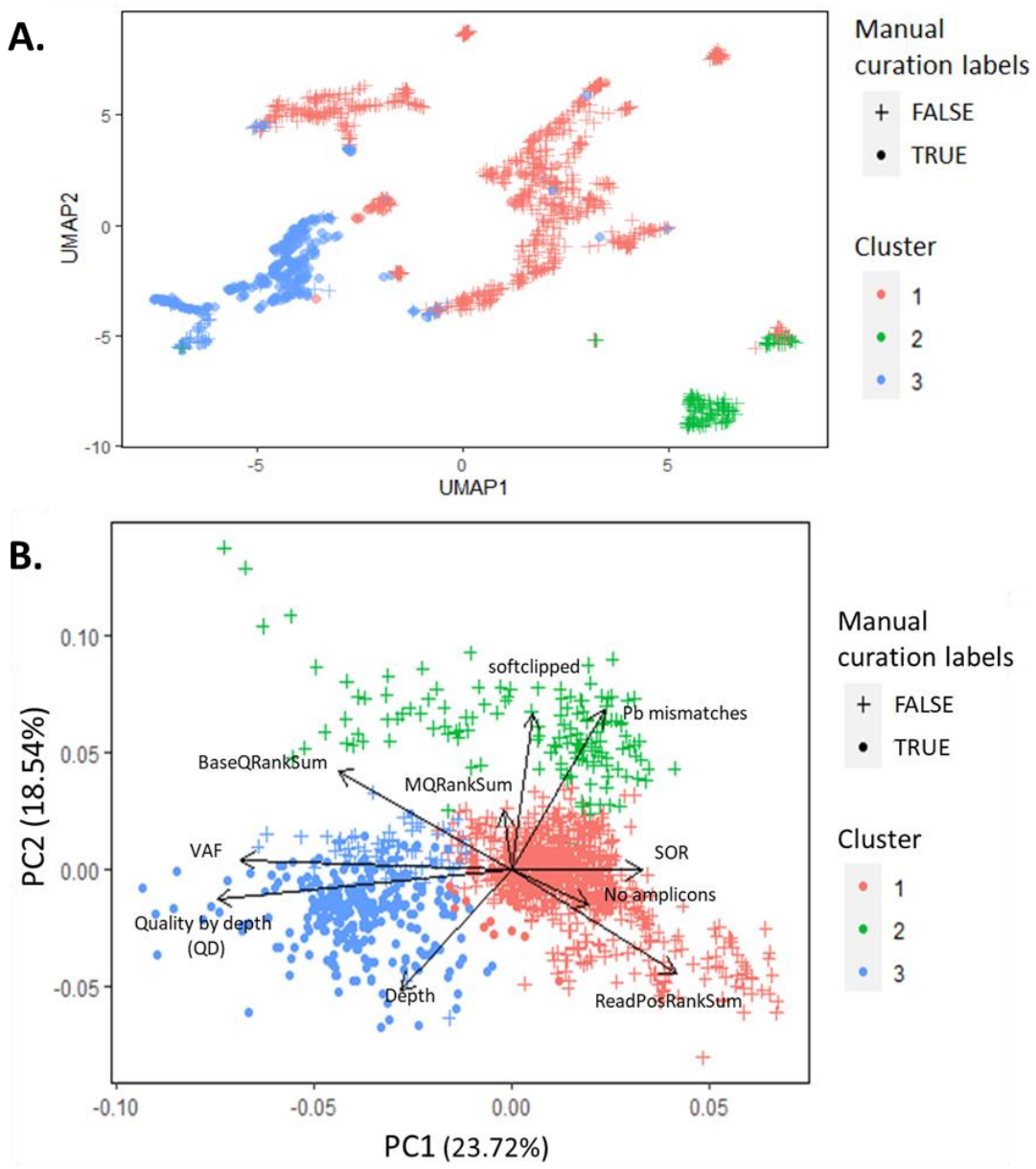


Figure 6-6. Clustering results for 1361 variants identified in 62 individual tumours from batch 1. **A.** UMAP plot of clustering results for test set. Each point represents a variant annotated with the manual validation labels (false variants represented by a "+" and true by a "•") and k-means clustering results. **B.** PCA plot of test set with loadings vectors. Each point represents a variant annotated according to clustering results (colour) and manual validation label (shape). Loading vectors for each feature which inform which features are driving the separation between clusters. The length of each vector shows the magnitude of the effect each feature has on the variance.

Principal component analysis (panel B of **Figure 6-6**) identified VAF and quality by depth (QD) as the features with the greatest effect on variance between true and false clusters. While the softclipped and per base mismatches had the greatest effect on the variance across false sub-groups. Panel B of **Figure 6-6** shows the loading vectors for each feature illustrating which ones were driving the separation between clusters. Vector length corresponded to the magnitude of the effect each feature had on the variance (the longer the arrow, the higher the effect and vice versa).

Scaled features were plotted on a heatmap and annotated according to cluster and manual validation labels. The heatmap (**Figure 6-7**) supported PCA results showing high values of VAF and quality by depth (QD) in true variants, which separated them from the false clusters. This pattern can be used to identify the "true" cluster in other data sets run through the model. Other features were more variable between true and false clusters. However, in the test set, high values of per base mismatches and softclipped reads separated false variants into two clusters.

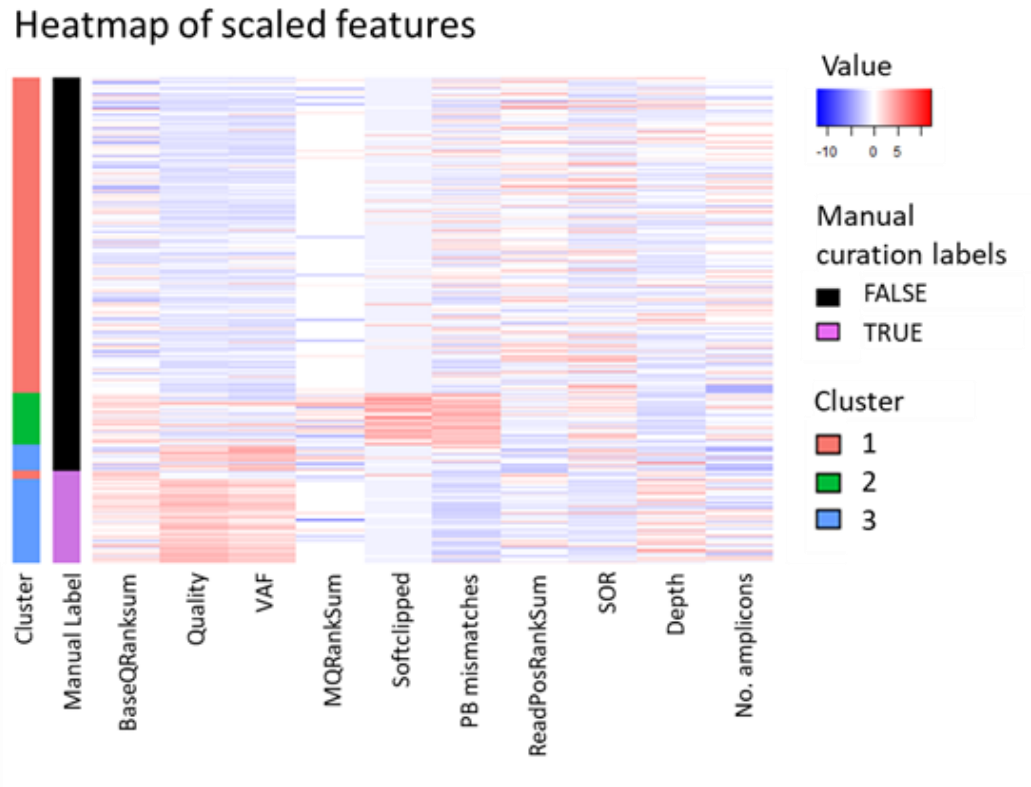


Figure 6-7. Heatmap of scaled features for batch 1 (test set). The first column represents cluster assignment (colour as in the UMAP plot), the second manual curation labels (black = FALSE, turquoise = TRUE) and the remaining columns show the ten features in used in the model. Each row represents a single variant.

The ML model accurately predicted 278/303 of the variants manually validated as true and 975/1058 of the false variants in batch 1 (**Table 6-2**). This resulted in a sensitivity and specificity of 92% (**Table 6-3**).

Table 6-2 Confusion matrix for test set (batch 1).

		truth	
		TRUE	**FALSE**
pred.	**TRUE**	278	25
	FALSE	83	975

Table 6-3 Statistics for test set (batch 1).

Stat	Value
Accuracy	0.92
Kappa	0.79
Sensitivity	0.92
Specificity	0.92

6.6.3 Overview of complete data set

Results of the Kruskal-Wallis test confirmed that the distribution of the features was in fact significantly different across all batches (**Table 6-4**).

Table 6-4 Results of Kruskal-Wallis test between five batches [n=4281 observations].

Kruskal-Wallis test		
Feature	**Chi-square**	**p value**
BaseQRankSum	510.29	$p < 0.001$
Depth	441.84	$p < 0.001$
MQRankSum	47.07	$p < 0.001$
QD	414.46	$p < 0.001$
ReadPosRanksum	233.62	$p < 0.001$
SOR	150.67	$p < 0.001$
No amplicons	374.52	$p < 0.001$
Sum pb mismatches	2057.1	$p < 0.001$
Sum softclipped	277.94	$p < 0.001$
VAF	791.98	$p < 0.001$

The Wilcoxon rank sum test gave more granularity as to which batches showed a difference in distribution across each of the features. It identified batch 2 as the batch that differed most from all other batches as the majority of the pairwise comparisons (35/40) were significantly different (**Figure 6-8**). The Wilcoxon rank sum test also indicated that there was a significant difference between the depth in all batches except between batch pair 3 and 4. Therefore, each batch was run through the ML model separately to avoid introducing additional noise or bias.

Wilcoxon rank sum test										
batch pairs	**BaseQRanksum**	**Depth (DP)**	**MQRanksum**	**Quality (QD)**	**ReadPosRankSum**	**SOR**	**No amplicons**	**PB mismatches**	**No softclipped**	**VAF**
1-2										
1-3										
1-4										
1-5										
2-3										
2-4										
2-5										
3-4										
3-5										
4-5										

p< 0.05 p ≥ 0.05

Figure 6-8. Pairwise comparison of features between batches. Values in green represent a significant difference between the pairs ($p < 0.05$) while orange no significant difference ($p \geq 0.05$).

Figure 6-9 illustrates the differences between the distributions of each feature across all five batches. Batch 2 is an outlier with respect to VAF and quality by depth (QD), the two features with the greatest effect on variance between false and true clusters, which were noticibly lower for batch 2 than the rest of the batches. BaseQRankSum distribution was similar across all batches with the exception of batch 2 which was notacibly lower. Batch 4 showed a higher mean depth

than other batches as well as a higher number of softclipped reads. The high number of softclipped reads could be linked to the higher depth as there will simply be more fragments within this batch.

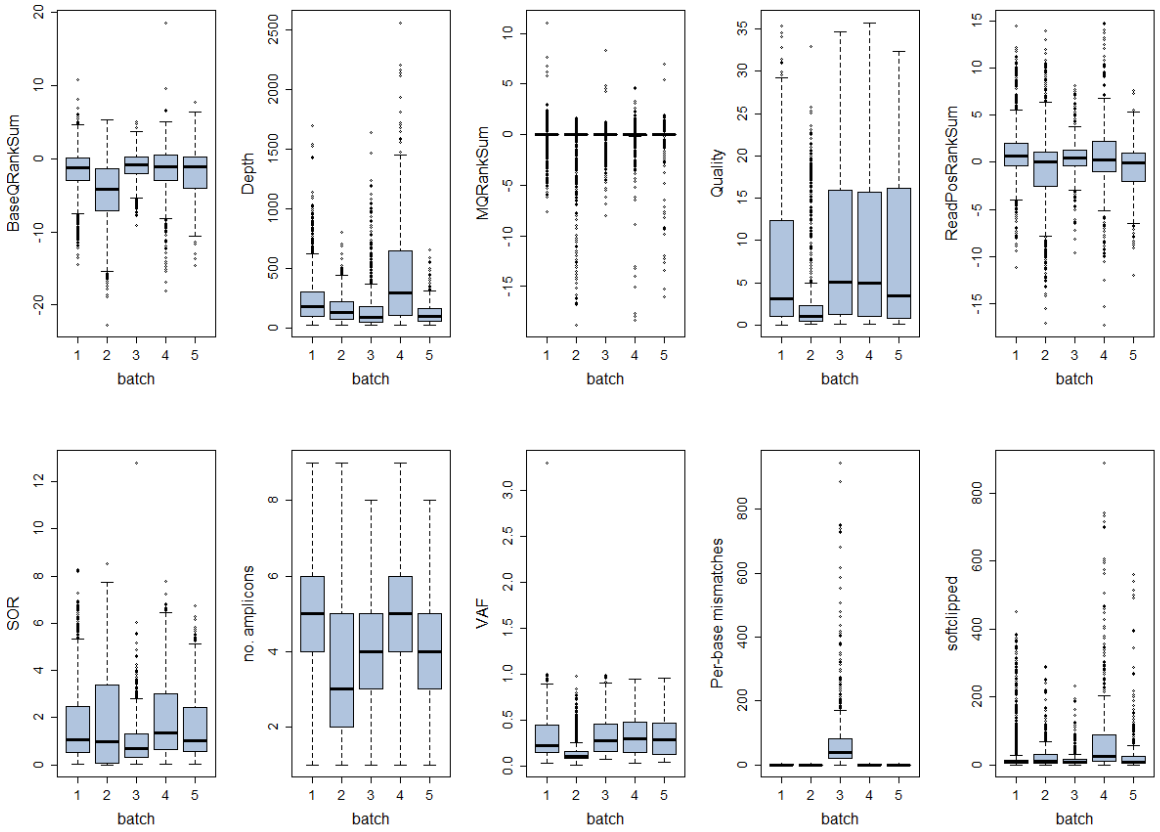


Figure 6-9. Feature distribution for batches 1-5 in the Jaramillo cohort. Each group of boxplots represents a feature (y-axis) and each box represents a batch. Boxplots illustrate the distributions of the features before clustering.

The distribution of values in the ten features across all batches suggests that each batch is subject to its own type of noise and or sequencing artefacts.

6.6.4 ML modelling data for individual batches

The results of the model (agreement, sensitivity, and specificity) in the test set (batch 1) were sufficiently high and deemed an acceptable approach to apply on all the samples. Unlike the test set, all variants in batches 2-5 were not reviewed in IGV before clustering as this was done to ensure the model was working properly.

Batches 2-5 were run separately through the ML model. **Figure 6-10** details the number of samples and variants in each batch after annotation using pipelineV5 and subsequently the number of variants used as input for clustering. Batch 2 had a high number of variants per sample (48 variants/sample) compared to other batches (10, 14, and 12 variants per sample in batches 3, 4, and 5 respectively).

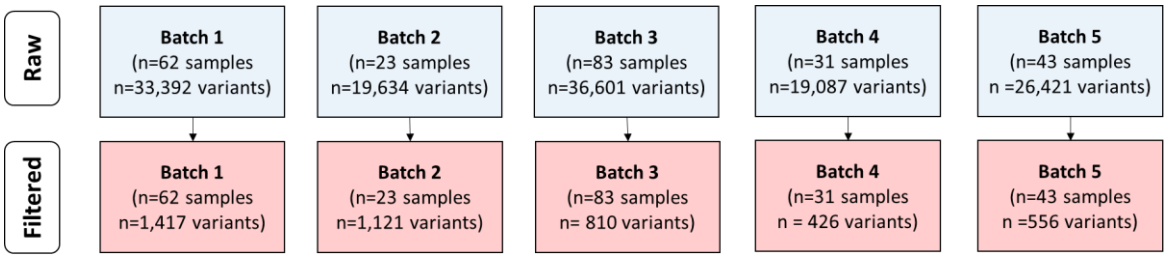


Figure 6-10 Flow diagram detailing sample and variant number in all batches run through ML model. The blue boxes show the raw number of variants in each batch before filtering. The pink boxes show the number of variants that were used for input in ML model. Batches include the 146 SMZL cases and additional samples that were sequenced concurrently.

The average silhouette method suggested 2, 5, 6 and 4 as the optimal number of clusters for batches 2, 3, 4 and 5 respectively. Results from the elbow method were ambiguous and therefore only the average silhouette method was used. Clustering results are shown in **Figure 6-11** .

A heatmap of the scaled features was created for all batches, annotated with clustering results and compared to the heatmap created for batch 1 (**Figure 6-12**). True clusters in batches 2-5 were identified as "true" using the patterns observed in batch 1, where VAF and quality had higher values (red) compared to other features. Additional validation in IGV of a small subset of variants [n=10 variants per cluster] was also used to consolidate which clusters belonged to true and false positives. Batch 2 contained two clusters where cluster 2 (blue/green) represented true variants and cluster 1 (red) false variants. In batch 3 the heatmap identified cluster 3 (green) as the "true" cluster and all others (1, 2, 4 and 5) as false. For batch 4 the heatmap identified cluster 6 (purple/pink) as the "true" cluster and clusters 1, 2, 3, 4 and 5 as false. Lastly, in batch 5 the heatmap identified cluster 1 (red) as the "true" cluster and clusters 2, 3 and 4 as false.

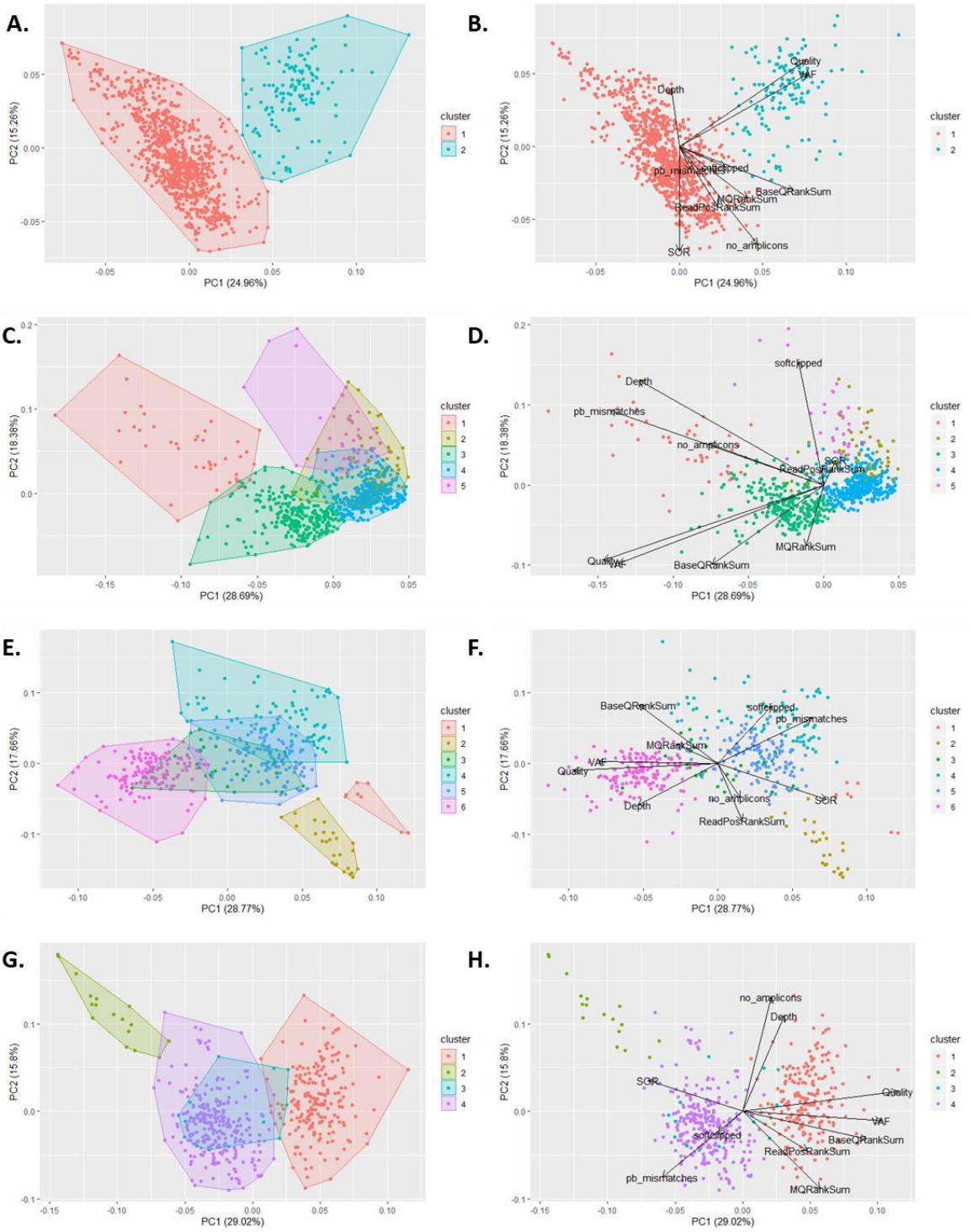


Figure 6-11. PCA of clustering results for batches 2-5. **A**. Batch 2: PCA results. **B.** Batch 2: PCA results with loading vectors. **C.** Batch 3: PCA results **D.** Batch 3: PCA results with loading vectors. **E.** Batch 4: PCA results. **F**. Batch 4: PCA results with loading vectors. **G**. Batch 5: PCA results. **H.** Batch 5: PCA results with loading vectors. Each point represents a variant and each variant is coloured according to clustering results. The plots with loading vectors illustrate which features are driving the separation between clusters. The length of each vector shows the magnitude of the effect each feature has on the variance.

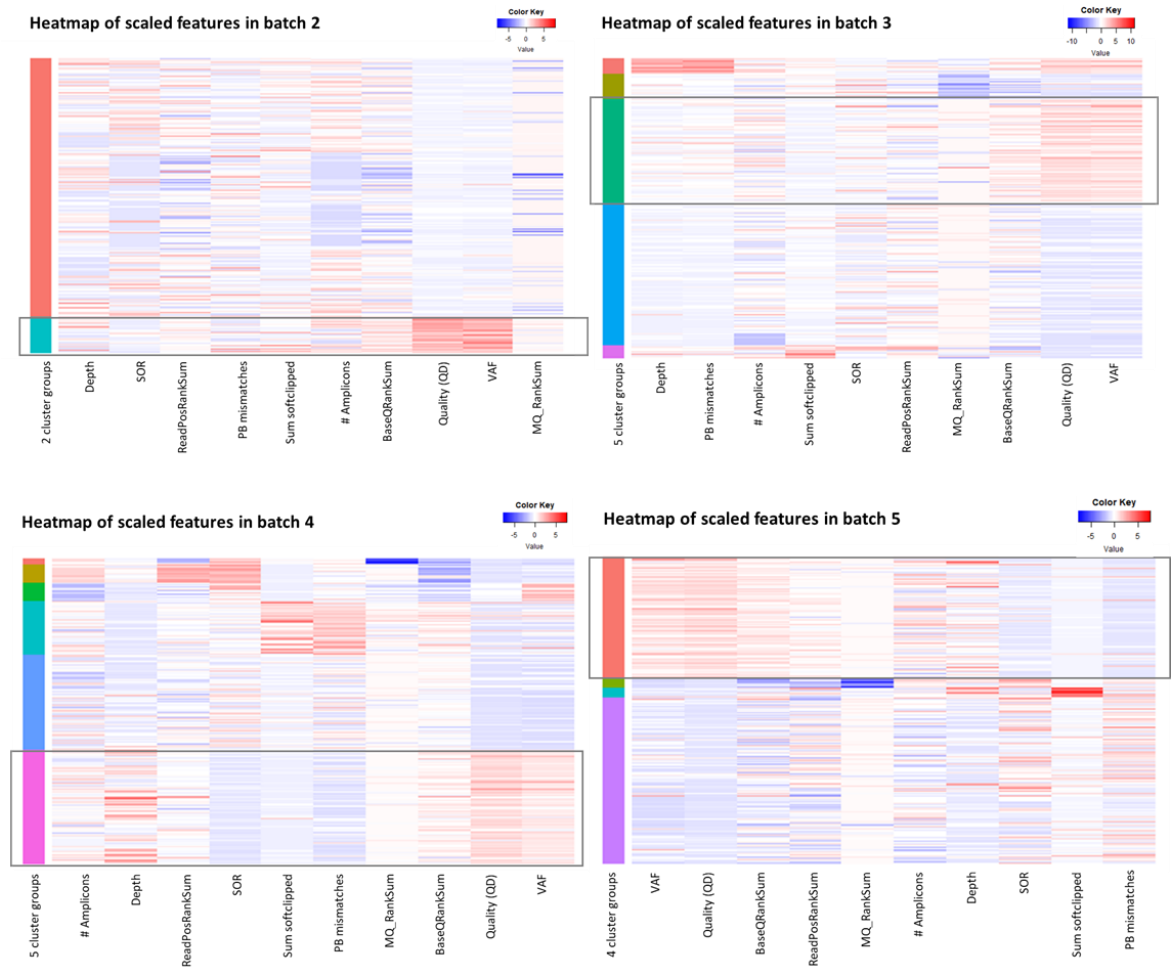


Figure 6-12. Heatmaps of scaled features for batches 2-5. Each row represents a variant, first column represents cluster assignment and other columns values for each of the features.

6.6.5 Validation cohort (CLL4 cohort - Truseq platform)

To validate the results of the clustering model, an independent sequencing cohort was obtained and processed through the bioinformatics pipeline. This cohort was comprised of CLL samples that were prepared using a TruSeq Custom Amplicon kit and sequenced on a MiSeq system at the University of Oxford. The TruSeq kit targeted 20 genes and samples were sequenced in several batches (more details on this cohort can be found in **section 3.1.3**). The Kruskal-Wallis test was used to determine if the distribution of the ML model features was statistically different between batches and would determine if it was possible to process all batches together or if the clustering would have to be done by batch. The Kruskal Wallis test revealed that the distribution of the ten features was significantly different across the 25 CLL batches. This meant that samples could not be grouped together and clustered at once. Consequently, one batch (miseq16-005) was chosen at random and run through the clustering model for validation. The batch was comprised of 175 variants 37 of which were validated true by either IGV or previous published work. The optimal number of clusters was 6 and results of clustering are shown in the PCA in **Figure 6-13** and in the UMAP plot in **Figure 6-14**. The majority of the true variants appeared to group within the sixth

cluster (purple), unfortunately, the clustering did not work as well as in the Haloplex HS data with a specificity of 79% and a sensitivity of 70% (**Table 6-6**).

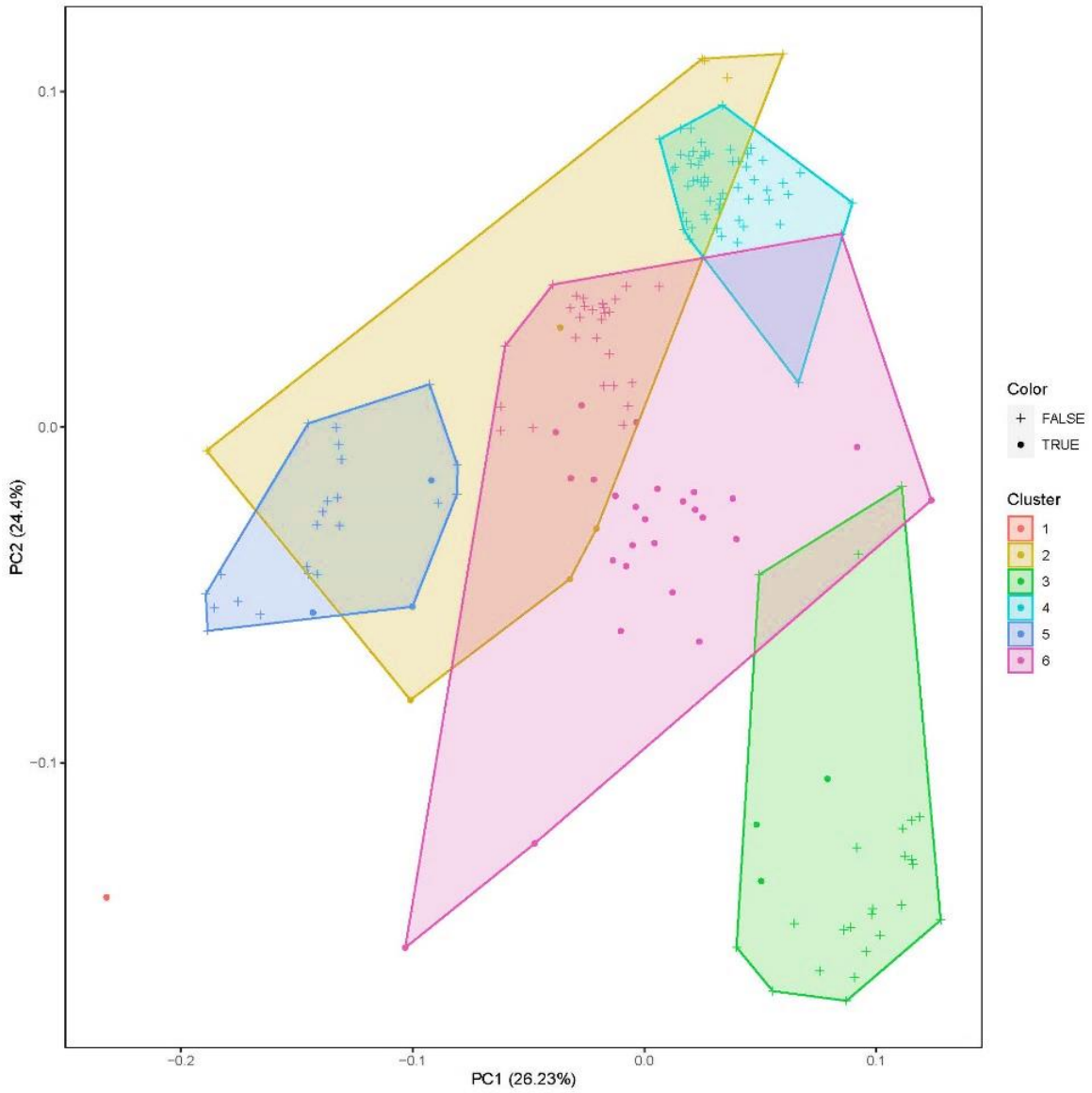


Figure 6-13. PCA of clustering results for CLL4 validation batch. Each point represents a variant annotated according to clustering results (colour) and in silico validation label (shape).

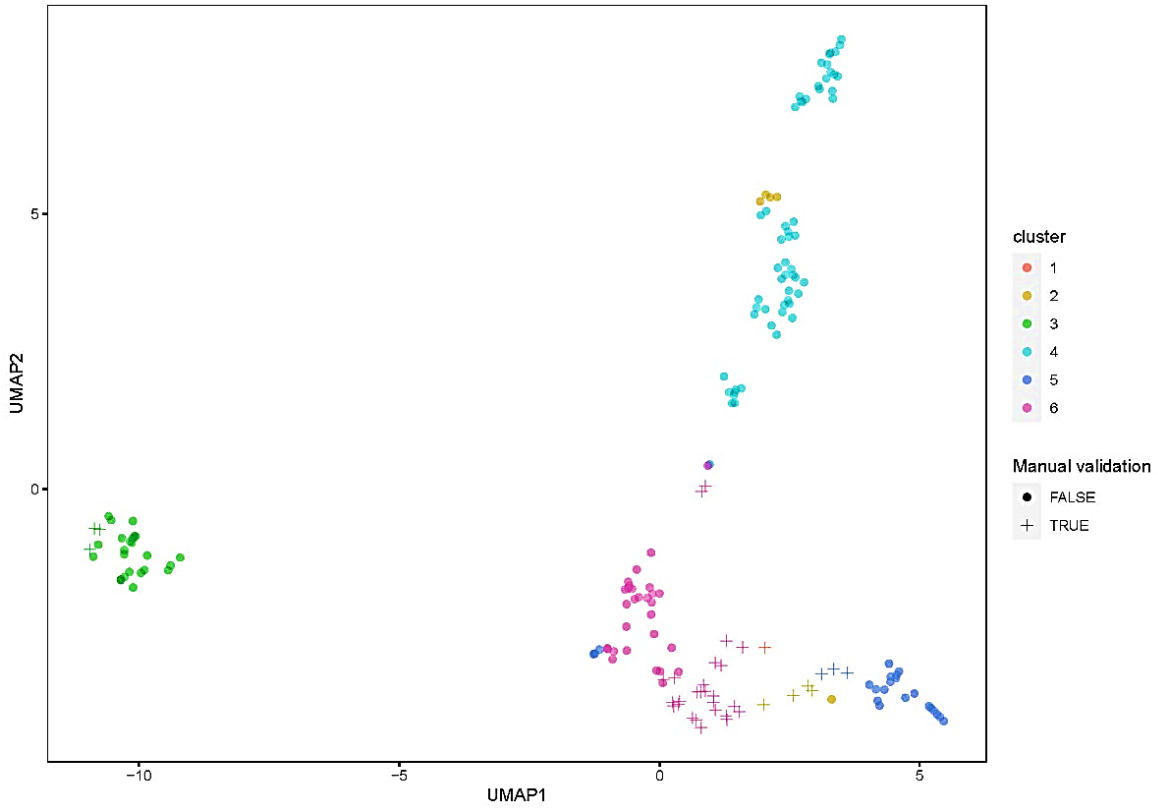


Figure 6-14. UMAP of clustering results for CLL4 validation batch. Each point represents a variant annotated with the manual validation labels (false variants represented by a "+" and true by a "•") and k-means clustering results.

Table 6-5 Confusion matrix for validation batch.

		truth	
		TRUE	**FALSE**
pred.	**TRUE**	26	11
	FALSE	29	106

Table 6-6 Statistics for validation batch.

Stat	Value
Accuracy	0.77
Kappa	0.41
Sensitivity	0.70
Specificity	0.79

6.6.6 Genomic landscape in filtered results

After successfully applying the ML model to all batches of the Jaramillo cohort and validating the results in an independent variant set, the ML model was integrated into the genomic analysis as a triage tool. This section will give a brief overview of the genomic results after integration of the model to evidence the effect it had on the sequencing data.

We compared the ten most frequently mutated genes in the 146 SMZL patients before and after clustering. The waterfall plots shown on **Figure 6-15** displays the ten most frequently mutated genes in the 146 SMZL patients before (panel A) and after (panel B) exclusion of false positives identified by the ML model. The waterfall plot in panel B was created after validation of variants in IGV and panel C (SMZLrefDB) is shown as reference.

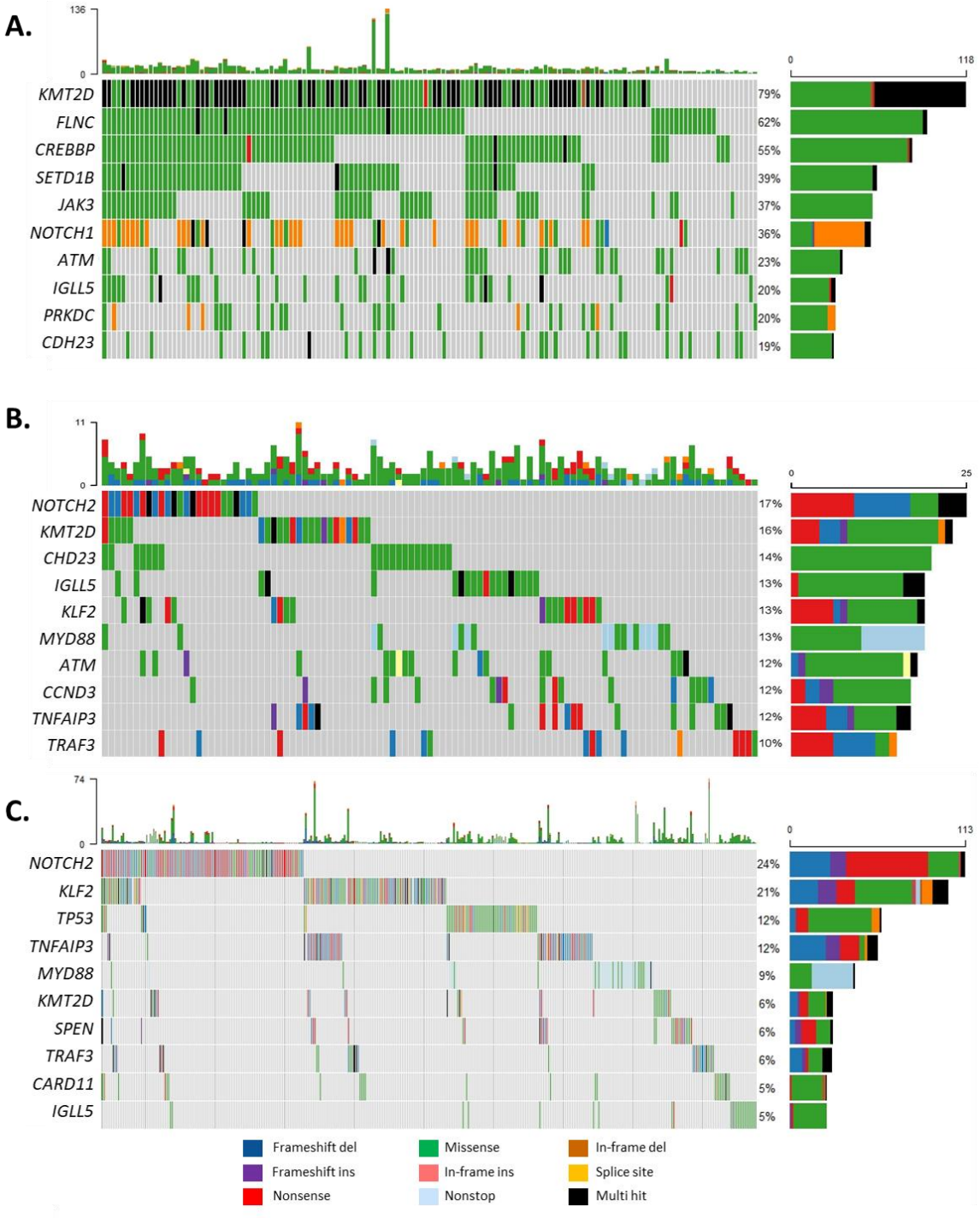


Figure 6-15. Waterfall plots comparing genomic results before and after use of the ML model. **A.** Top 10 frequently mutated genes in Jaramillo cohort [n=146] before processing with ML model. **B.** Top 10 frequently mutated genes in Jaramillo cohort [n-146] after application of ML model. **C.** Top 10 frequently mutated genes in SMZL dataset compiled in the systematic review [n=508]. Barplot on the top of each waterfall plot shows the number of variants per samples, while the barplot on the right displays the number of samples harbouring mutations in each gene.

The most obvious difference between the data before and after exclusion of false positives was the reduction of missense mutations across the ten most frequently mutated genes. In *KMT2D*, although there was still a high number of mutations, the multiple hits across the majority patients was no longer there. Another notable difference was the removal of the *NOTCH1* splicing variants (orange), although it is not obvious as *NOTCH1* was no longer in the ten most mutated genes after

cleaning. Furthermore, the gene list in panel B of **Figure 6-15** reflected more closely the genes previously identified as recurrently mutated in our systematic literature review (SMZLrefDB). The frequency (number of patients across cohort harbouring mutations/ total number of patients) of *NOTCH2, KLF2, TNFAIP3, MYD88* and *TRAF3*, was much closer to previously observed frequencies after exclusion of artefacts. These results further validated our methods showing a much more coherent list of variants which could now be assessed in more detail.

6.7 Discussion

Manual validation of variants is a key component when analysing sequencing data from any type of tissue, especially when additional sequencing is not possible due to either cost or availability of materials. However, it is a labour-intensive job prone to human error and bias. This is especially true when thousands of variants need to be reviewed (a problem amplified in tumour only samples) and if perhaps the quality of the sequencing is not optimal. There are times when manual validation reviewers might not agree whether a variant is true or false, and they might even consider a variant true purely because it is often seen in a specific disease, potentially biasing the selection of known mutations when the quality of a variant is below average. Therefore, developing an objective and efficient method to identify true variants from noise in tumour only data was important, not only in terms of reducing the time needed to validate variants, but also to have a more systematic and unbiased approach to filtering out false positives.

To fulfil this unmet need we developed an unsupervised machine learning model that can be used to identify true and false positive variants in tumour only amplicon data. The model uses ten metrics or features obtained from GATKs haplotype caller (BaseQRankSum, Depth, MQRankSum, QD, ReadPosRankSum, SOR and VAF) and the BAM file (No. of amplicons covering the loci, sum of per base mismatches in reads covering the loci, and sum of softclipped reads covering the loci). The number of optimal clusters varied by batch; therefore, each sequencing batch was processed separately. The model's sensitivity and specificity were 92% and had an agreement (kappa) of 79% to the manual validation, with the caveat that the manual validation labels may have some errors. We aimed for the highest sensitivity as retaining false positives in the data is not ideal but preferable to exclusion of true variants. Filtering out false positives from a reduced pool of variants is less arduous and time consuming compared to identifying a true variant hidden amongst hundreds of false positives. In the test set, twenty-five variants manually validated true in IGV were labelled as false by the model. Upon further examination, discrepancies were seen on variants with lower average depth and VAF compared to other true variants. This set of variants, not included in the true cluster, was comprised of seven deletions and 18 single nucleotide variants (SNVs). Two of these twenty-five were splicing variants and eight were stopgain SNVs.

Reasons for the enrichment of stopgains is unknown and could be due to the biology of the disease. Consequently, all stopgains and stoplosses in our cohort were assessed in IGV regardless of cluster.

The sensitivity and specificity of the model was over 20% lower in the validation cohort. The most likely cause being the enrichment kit used. The Truseq design did not allow for many overlapping amplicons covering the targeted regions and on average each variant was targeted by 2 amplicons compared to 4 in the Haloplex HS design. This will affect the calculation of some of the features by the variant caller. Features like ReadPosRankSum for example look at how many more alternate or reference alleles there are towards the end of a read and if we only have one amplicon, then the value for ReadPosRanksum might be inflated. This will cause the algorithm to cluster said variant with false positives since the end of a read tends to be of lesser quality. Although the model was not as successful in this platform it might be suitable for more random library designs with several overlapping probes or amplicons. It would be ideal to test this model in additional datasets perhaps exome or other capture kits to better understand generalisability, however that is beyond the current scope of this work.

The model developed herein, like any other clustering technique, has its limitations that should be considered before being used. The first limitation is that this model is not as effective on indels as it is on SNVs. Perhaps because indels are more difficult to align and will likely have lower quality than SNVs. It is advised that indels be looked at independently and not included in the clustering.

The second limitation is that this model could potentially miscall variants with low depth and variant allele frequencies (< 0.2) as these will have a lower number of supporting alternate reads leading to lower quality (calculated by variant caller) and perhaps be biased in the way they were captured during the library preparation. This is specially the case for genes with a high GC content However, if only poor-quality data is available for such genes (e.g. *NOTCH1, NOTCH2* and *KLF2*) because of their high GC content they will inherently have less false positives due to less reads mapping to them and manual validation is more feasible.

The aim of this model was to distinguish between true versus false positive variants in tumour only samples, and not to distinguish somatic versus germline. Although the data was filtered to enrich for somatic variants before clustering, there is a possibility that germline variants were still present in the data and cancer specific filters might need to be used downstream. In the test set, synonymous variants were included which make up 36% of the variants. Synonymous variants were included, as they have the potential to affect splicing, however, it might skew the model since they could also be rare germline variants.

Something else that should be considered is that different sequencing batches will have different types of noise and therefore each sequencing batch will need to be clustered separately, or batch effect should be corrected for. It is not known exactly why each batch differs when sequenced, but this could be due to technical errors during sequencing, differences in the reagents when inside the sequencing instrument, preparation of the libraries or the samples themselves. This is supported by the random forest model created by Wu and colleagues where batch effect was the third most important feature in their model[137]. It might be cumbersome to cluster each batch separately, but it does have an advantage. Supervised models are robust when they are trained on large numbers of the same type of data, however if the training data has a low variance and high bias the model is likely to underfit the data and vice versa if there is high variance and low bias it could potentially overfit the data and miss patterns within it[146]. This variance-bias balance might be difficult to calibrate in a supervised approach when accounting for batch effects. This was exemplified by the fact that it was thought the data might cluster into two groups: false positives and true variants; but this was not always the case. Therefore, an unsupervised approach allowed the identification of different types of noise within the data.

With an unsupervised approach not only can batch effects be accounted for across features, but samples can be processed without the need for large training datasets relying on data collected elsewhere.

6.8 Conclusion

In summary, the unsupervised machine learning model clustered suspected true somatic variants separately from suspected false positives with high sensitivity and specificity, allowing for a more efficient way of filtering out false positive variant calls in unmatched tumour samples. This demonstrates that it is possible to automate unmatched somatic variant filtering with an unsupervised ML model, reducing time when manually curating variants and even reducing bias that could be introduced by a manual reviewer. Although the results of the model were not used as a hard filter, they provided excellent additional annotation that was used to triage variants into high and low confidence and to obtain a variant list that was in line with previous published studies.

Chapter 7 Next generation sequencing analysis of splenic marginal zone lymphoma patients

7.1 Synopsis

In this chapter, 57 genes are interrogated across a cohort of 321 SMZL patients (Jaramillo and Parry cohort) with no matched germline tissue. We report a list of putative somatic variants and those within frequently mutated genes are examined in detail.

David Oscier performed the diagnosis of samples and provided clinical guidance. Dr. Helen Parker performed the design of the gene panels and performed the library preparation, sequencing of samples, and was responsible for the transfer and demultiplexing of the sequencing data to university servers. Carolina Jaramillo Oquendo ran the samples through the bioinformatics pipeline and the unsupervised machine learning model, followed by the interpretation of the NGS data. Prof Sarah Ennis, Prof Jonathan Strefford and Dr. Jane Gibson acted as supervisors overseeing the processing, analysis, and interpretation of the data.

7.2 Introduction

As mentioned in **section 1.4** there are three separate clinicopathological marginal zone (MZ) lymphoma entities, splenic marginal zone lymphoma (SMZL), nodal MZL (NMZL) and extranodal (EMZL) as defined by the World Health Organisation (WHO)[31]. Additional provisional entities such as splenic diffuse red pulp lymphoma (SDRPL), hairy cell leukaemia-variant (HCL-v) and clonal B-cell lymphocytosis of marginal zone origin (CBL-MZ) are emerging, the latter clonally related to SMZL[32–34,147]. SMZLs make up approximately 20% of MZLs and NMZLs comprise less than 10% of cases while EMZLs comprise around 60% of cases that occur at any extranodal sites[148].

The Diagnosis of SMZL relies on a combination of clinical features along with assessment of lymphocyte morphology and immunophenotype, bone marrow histology and immunohistochemistry. In the absence of splenic histology, the differential diagnosis of SMZL from common low-grade B-cell disorders such as chronic lymphocytic leukaemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL) is usually straightforward but SMZL lacks a disease-specific immunophenotype and the distinction between SMZL and some cases of HCL-v, SDRPL and lymphoplasmacytic lymphoma (LPL) may be more difficult. With the lack of biomarkers for differential diagnosis, genomics has the potential to aid in the identification of disease specific gene mutations. However, SMZL and marginal zone lymphomas tend to be excluded from large international sequencing consortia, resulting in limited genomic data for these cancers.

The systematic literature review (**Chapter 2)** revealed that only six genome-wide studies have been conducted on only 35 patients. Kiel and colleagues were the only ones to employ WGS but were limited to six cases without matched germline DNA[49]. Five WES studies have been carried out on discovery cases, and subsequently targeted relevant genes in additional samples. To date, none of these studies have reported mutational signatures nor mechanisms, such as kataegis and chromothripsis[149]. Even the somatic mutational burden remains disputed, with a range of somatic mutations per patients between 9 and 82 (mean 25)[50,51,54]. This limited agreement is likely due to the low patient numbers, and experimental and computational differences in WGS/WES processing, but could also allude to disease heterogeneity and statistical power insufficient to catalogue the complete mutational landscape of the disease[54,70]. Targeted re-sequencing approaches have helped elucidate recurrently mutated genes, but these studies have often included only small numbers of matched germ-line material for analysis.

Whilst we currently have only a limited picture of the somatic landscape of SMZL, several recurrently mutated genes have been identified, that are preferentially within physiologically important cellular processes, such as MZ B-cell maturation and migration, and cell cycle control (**Figure 7-1**). The three most important genes identified by the systematic review were *KLF2, NOTCH2* and *TP53*. The review also confirmed the importance of genes that interact with the NF-κB pathway such as *TNFAIP3*, MYD88, *TRAF3*, *CARD11*, *IKBKB*, and *BIRC3* and brought forward two genes, *TRAF3* and *KMT2D,* which had not been considered significant players in SMZL biology. Unfortunately, within our systematic literature review an in-depth analysis of the variants and their interactions was hindered by the experimental/analytical design and the lack of required information in published data.

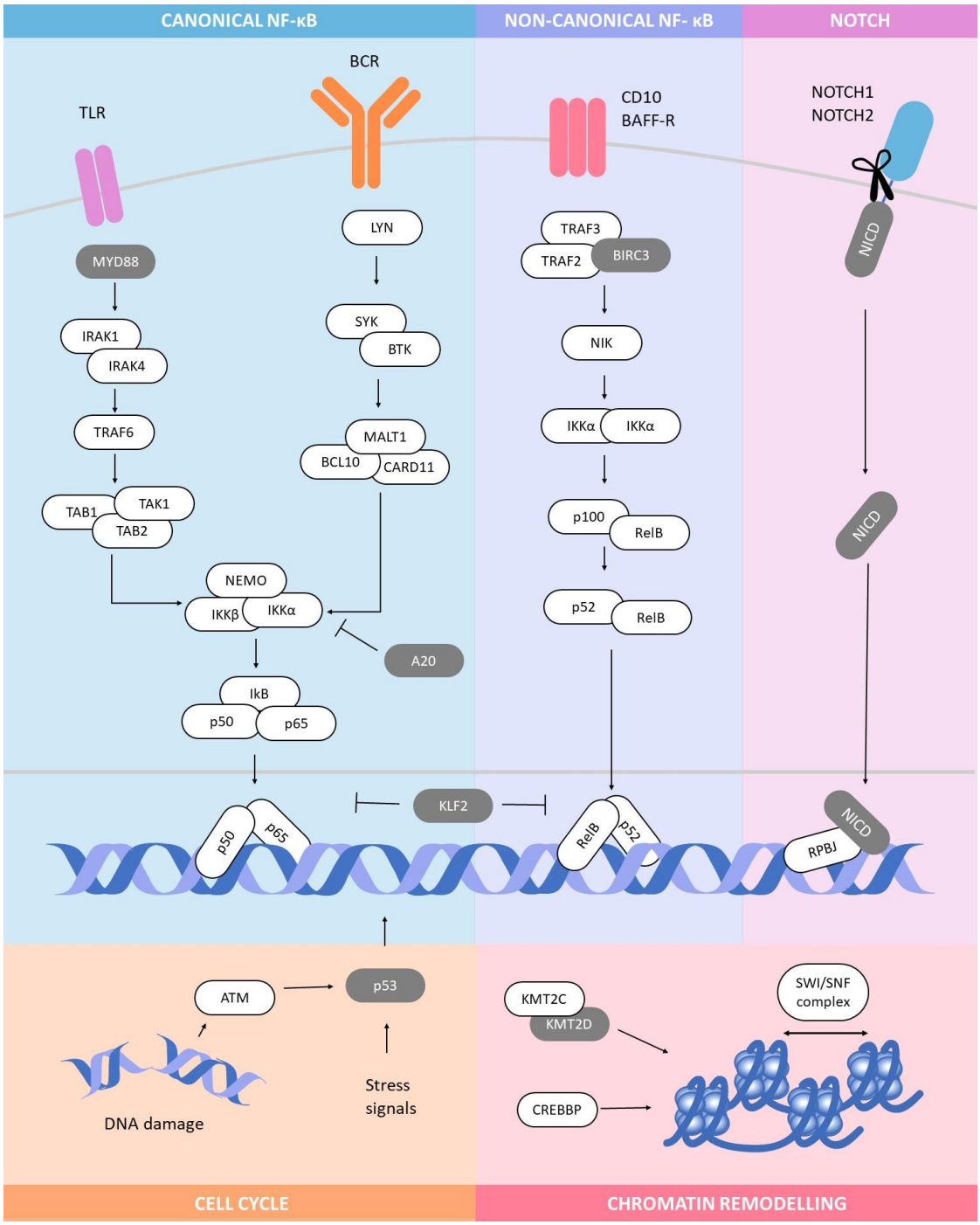


Figure 7-1. Main pathways targeted by somatic mutations in SMZL. Recurrently mutated genes in SMZL preferentially target physiologically important pathways, including canonical and non-canonical NF-kB activation (through BCR, TLR and BAFF-R signalling), Notch signalling, chromatin remodelling and cell cycle control. Genes encoding proteins in grey have a mutational frequency greater than 10% within SMZL cohorts. Figure by Jaramillo Oquendo et al[55] licenced under CC BY 4.0.

This chapter presents the results of targeted sequencing across 321 SMZL patients. The results shown here fulfil our main objective of constructing a detailed characterisation of the genetic landscape of SMZL through the identification of somatic variants in tumour only SMZL samples. It aims to impart more clarity on the mutational frequency of genes relevant in B-cell malignancies across a large SMZL cohort and allow a detailed look at the somatic interactions between these genes.

7.3 Materials and Methods

7.3.1 Cohorts

Jaramillo cohort: 146 tumour only SMZL samples, all meeting established diagnostic criteria[37], were obtained from 11 international collaborating centres. peripheral blood [n=97], spleen cells [n=13] or bone marrow [n=2] however, for some cases samples were sent as DNA and the material type from which they came from is unknown [n= 34]. Samples were analysed with a bespoke Agilent Haloplex HS Target Enrichment system that enriched 383.74kb of genomic DNA for 59 genes and genomic regions, designed with SureDesign (for further details on this cohort refer to **section 3.1.1**).

Parry cohort: 175 tumour only SMZL samples all meeting established diagnostic criteria[37], were obtained from 8 centres across Europe. DNA was extracted from peripheral blood [n=135], bone marrow [n=22], spleen [n=17], or lymph nodes [n=1). Samples were analysed with a bespoke Haloplex Target Enrichment system (Agilent Technologies) that enriched 2.39Mb of genomic DNA for the coding regions of 768 genes, designed with SureDesign (for more details see **section 3.1.2**).

Our approach to integrating the data was first to analyse the new cohort (Jaramillo) and see how the results compared to those of the Parry cohort. Then both cohorts were combined and analysed together to increase the power of our analysis.

7.3.2 Haloplex sequencing and bioinformatics pipeline

The Jaramillo cohort was sequenced in five batches using 150 bp paired end sequencing on a Nextseq system (Illumina). The mean target coverage was 305x (range 18-1107). The parry cohort was also sequenced in batches (16-47 samples) using 100 bp paired end sequencing on an Illumina HiSeq2000. The mean target coverage was 238x (range 97-546). Coverage across the Jaramillo and Parry cohorts can be found in **Supplementary Table 5**.

All samples (Jaramillo and Parry cohorts) were processed through pipelineV5 (see **section 4.5.3**). This consisted of aligning raw FASTQ reads to the reference genome (hg38) using BWA-mem, followed by variant calling using GATK haplotype caller and annotated using Annovar software. For detailed description of the bioinformatics pipeline refer to **Chapter 4**.

7.3.3 Exclusion of false positives and likely germline variants

After annotation, variants were filtered to enrich for somatic mutations and exclude those that were not of interest as described in section **5.3.2**. This initial filter excluded intronic and intergenic

variants, variants with a frequency less than 1% in databases of known germline variation, and variants with a total depth less than 30. However, preliminary results (**Chapter 5**) showed these initial filters still left a large number of false positives within the dataset. This led to the development and use of the unsupervised machine learning model to help identify spurious calls (**0**). The results from the machine learning model were used as a triage tool to categorise variants intro high, medium, and low confidence. High and medium confidence variants were validated in IGV and all others were excluded from the analysis (For further details see **section 6.5.5**).

A second filtering strategy was developed to exclude germline variation. To do this *in-silico* predictive scores (CADD phred score [89] and spliceAI[150]) were used to determine the likelihood of a variant being pathogenic. Alongside the CADD score the frequency of the variant in databases of known germline variation and the variant allele frequency (VAF) was also used.

First, a cut-off value for the CADD score was required to filter out "benign" or "non-functional" variants. However, the authors of CADD recommend integrating the scores with other evidence and not using CADD as the sole piece of evidence, therefore, a consensus approach was taken. Variants were also annotated using the mutation significance cut-off[151] (MSC) for both the Human gene Mutation Database (HGMD)[152] and ClinVar[91] with a 90% confidence interval. The MSC of a gene is defined as the lower limit of the confidence interval (90%, 95% or 99%) for the CADD score of all its high-quality mutations described as pathogenic in HGMD or ClinVar. Therefore, if a variant had a CADD score greater than or equal to the MSC cut-offs for either HGMD or Clinvar or if the CADD score was ≥ 15 it was kept. Furthermore, if any variant had a spliceAI score > 0.5 it was also included in the analysis. Synonymous variants were then filtered out and this list went onto the next stage of the filtering strategy. These steps are illustrated in blue in **Figure 7-2**. Variants had been filtered using the frequency in databases of known germline variation before clustering, however the preliminary data showed possible germline variants still present. Considering some of these databases might have some contamination a more nuanced approach was necessary and the variant allele frequency as well as the number of times the variant was found in our cohort was considered.

A heterozygous germline variant has two alleles *a* and *b*. If the variant is sequenced, then the probability of calling or identifying either allele *a* or *b* can be modelled by the binomial distribution. We can find a confidence interval based on the number of experiments, or in this case the depth, to determine if the observed variant allele frequency (VAF) falls within this distribution. If the VAF does fall within the confidence interval, then there is a high likelihood that the variant could be a germline variant. Using the variant depth, the Clopper–Pearson method was used for calculating a confidence interval of 95% of the binomial distribution per variant.

Consequently, for the remaining variants, if any variant had a frequency in any the databases of known germline variation (see **section 5.3.2**) greater than 0.001 and they were present in two or more samples they were excluded. Or if the variant was present at a frequency > 0.001 in databases of known germline variation and the VAF of that variant fell within the 95% confidence interval of the binomial distribution for that depth it was also excluded. These steps are illustrated in purple in **Figure 7-2**.

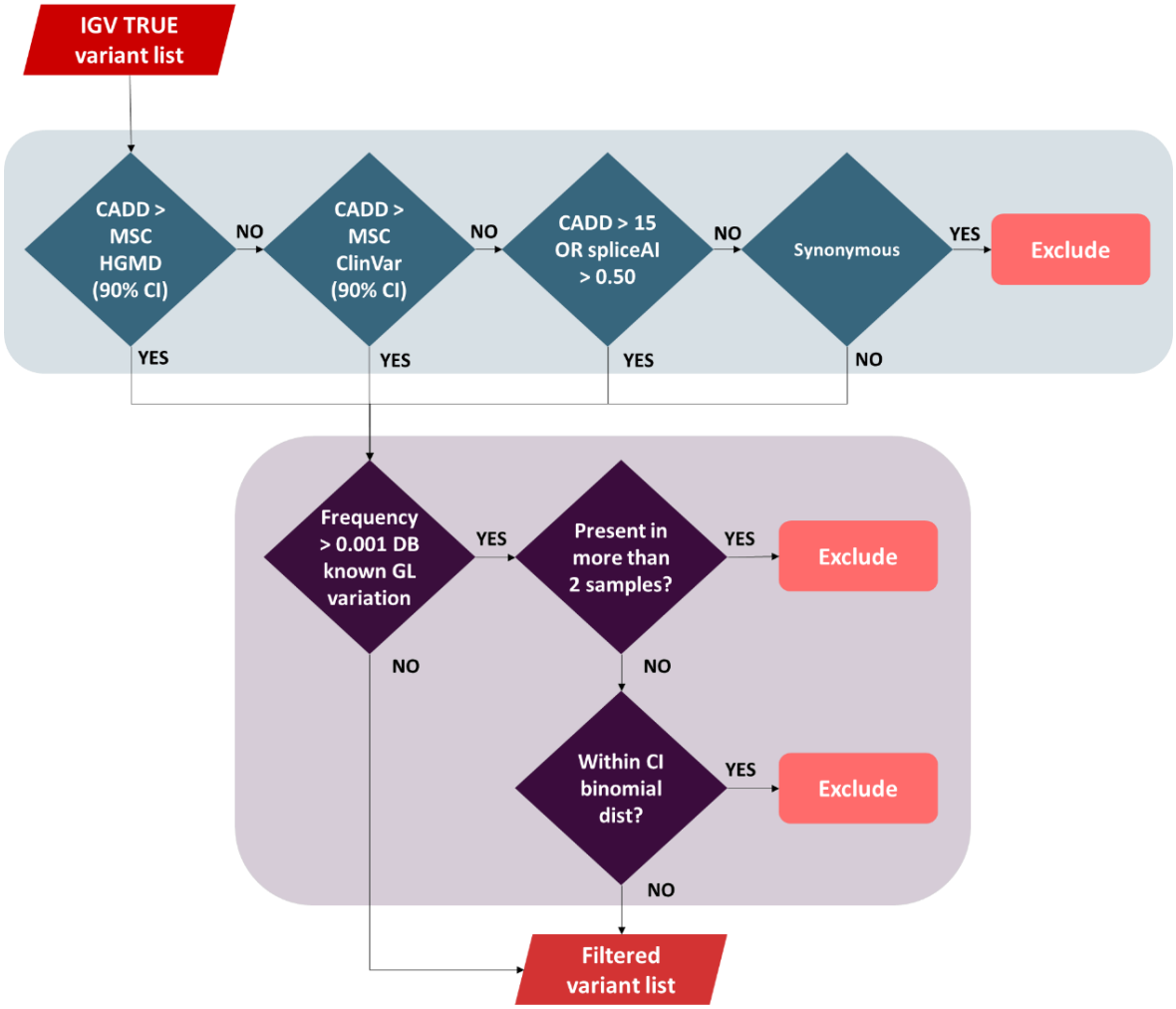


Figure 7-2. Flow diagram of filtering strategy to exclude germline variants. First CADD scores were compared against the mutation significance cut-off (MSC) in either the Human gene Mutation Database (HGMD) or ClinVar and proceeded to the next step if the CADD score was greater than MSC cut-off. If variants had a CADD score greater than 15 or a spliceAI score greater than 0.5 they also proceeded to the next step. Synonymous variants that did not meet any of cut-offs were excluded. The second step in the filtering process involved looking at the frequency of the variant in databases of known germline variation. If variants are found at frequencies greater than 0.001 AND present in more than two samples OR fall within the confidence interval for the binomial distribution of their respective depth, they are excluded.

7.3.4 Transcript selection

To correctly visualise variants, it was important that all variants within a gene were described using the same transcript. Transcript flags can be used to identify the highest quality or most

relevant transcripts and were obtained using Emsembl[153]. Ensembl uses five flags to annotate genes: MANE select, Transcript support level (TSL), APRIS, GENCODE Basic, 5' and 3' incomplete. A transcript was chosen using two main flags MANE and TSL. A full description of these flags taken from the Ensembl page[154] can be detailed below:

MANE (Matched Annotation between NCBI and EBI) Select

To determine the MANE Select transcript, Ensembl and NCBI independently identify which transcript we believe is the most biologically relevant. Where these match, the transcripts are labelled as MANE in both databases. The transcripts are absolutely identical in both databases, having matching splicing structure, sequence which matches the reference genome, 5' and 3' UTRs and start and end.

Transcript support level

The Transcript Support Level (TSL) is a method to highlight the well-supported and poorly-supported transcript models for users. The method relies on the primary data that can support full-length transcript structure: mRNA and EST alignments supplied by UCSC and Ensembl.

Transcripts with a MANE flag were chosen and if the transcript did not have the MANE flag then the transcript with the highest support level (TSL) was used. Once the main transcripts were selected the variant list was ready to be analysed.

Summary of the post processing steps to obtain the final variant list can be seen in **Figure 7-3**.

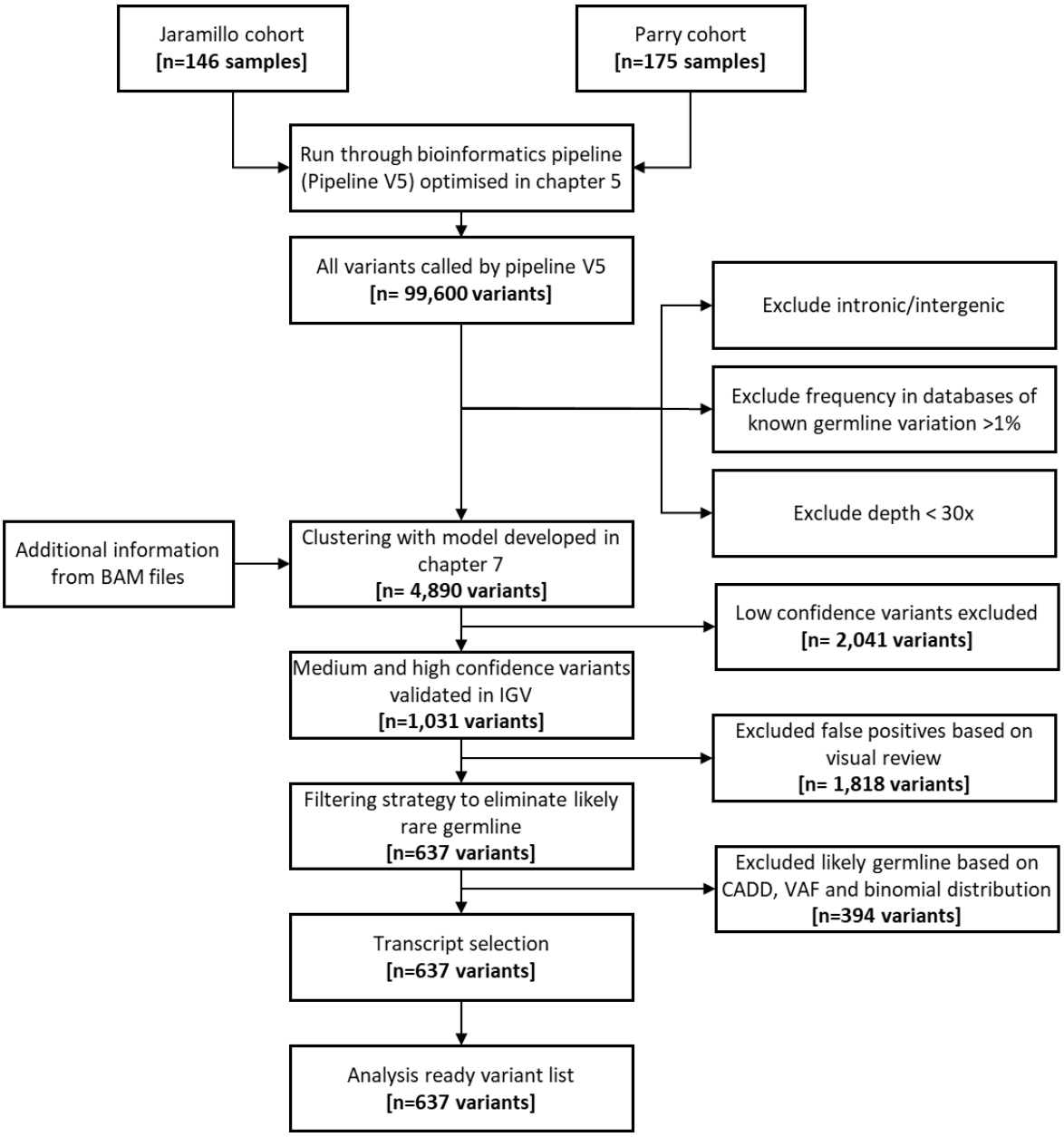


Figure 7-3. Flow diagram of filtering strategies to obtain final variant list . Samples from the Jaramillo and Parry cohort were run through the optimised bioinformatics pipeline (pipelineV5) before they began an extensive filtering process. The raw variant list that results from the pipeline was filtered to exclude variants that were not in targeted regions (i.e. introns/exons), those with a frequency less than 1% in databases of known germline variation, such as gnomAD and EXAC, and those with a depth less than 30x. Once variants go through this initial filtering they were clustered into groups, which determined if they were classed as low, medium, or high confidence variants. High and medium confidence variants were validated in IGV and all other variants excluded. After validation variants went through a second filtering strategy which used scores to predict deleteriousness as well as the variant allele frequency (VAF) and depth to exclude potential rare germline variation. Subsequently, all variants were annotated with the main transcript (MANE) and were then ready to be analysed.

7.3.5 Data visualisation and analysis

The final annotated variant list was used as input into R packages maftools and GenVisR for data visualisation. Cbioportal was used to visualise mutations overlaid on a linear protein (lolliplot)

representing each gene with its respective domains. To detect mutually exclusive or co-occurring set of genes a pairwise Fisher's Exact test was performed in R.

7.4 Results

For this section, unless stated, results refer to the combined Jaramillo-Parry cohort. After clustering, 1031 variants were validated using IGV and these were then filtered leaving a total of 633 variants ready for analysis [n=314/633 from the Parry cohort and n=319/633 from the Jaramillo cohort]. The 637 variants came from 261 out of the 321 assessed unique individuals and were found in 45 out of the 59 assessed genes. The complete list of variants can be found in **Supplementary Table 7**.

Most variants were missense mutations [n=358], followed by frameshift indels [n=134] and stopgain mutations [n=102]. With lesser frequency, splicing [n=29], in-frame indels [n=9], and stoploss [n=1] were also present. The majority were single nucleotide changes [n=488], followed by deletions [n=118] and a small number of insertions [n=27]. C to T substitutions were the most common base change, which was in line with results from the systematic review (**Figure 7-4**).

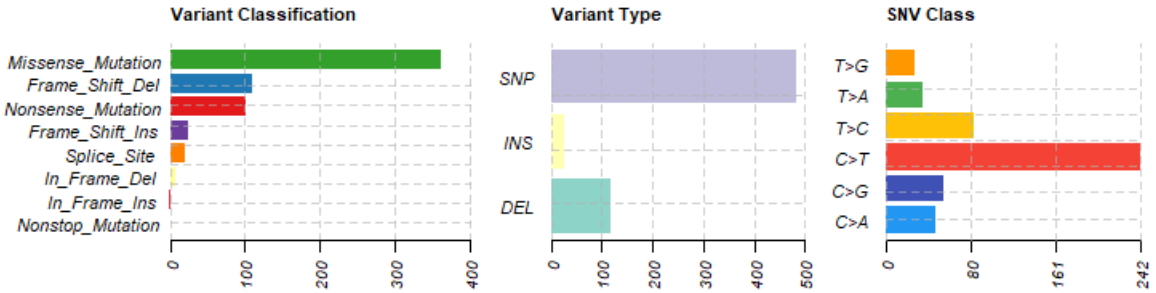


Figure 7-4. Variant summary in Jaramillo-Parry cohort. A. Break down of variant classification in the Jaramillo-Parry cohort. B. Bar chart of variant type. Single nucleotide polymorphisms (SNP) are in purple, insertions (INS) in yellow and deletions (DEL) in green. C. Breakdown of nucleotide substitution. X-axis for all three figures show the number of variants within the cohort.

7.4.1 Recurrently mutated genes

In concordance with the literature *NOTCH2* [13%], *TP53* [12%] and *KLF2* [12%] (**Figure 7-5**) harboured high mutational frequencies in the combined Jaramillo-Parry cohort. However, it was *KMT2D* [13%] which had the greatest mutational frequency with mutations in 43/321 patients. The Jaramillo cohort was plotted separately (**Figure 7-6**), giving similar results.

To compare the results of the combined Jaramillo-Parry cohort against the Jaramillo only cohort **Table 7-1** shows the rank, frequency of mutations and number of affected patients of the ten most frequently mutated genes in both the Parry cohort and the Jaramillo cohort.

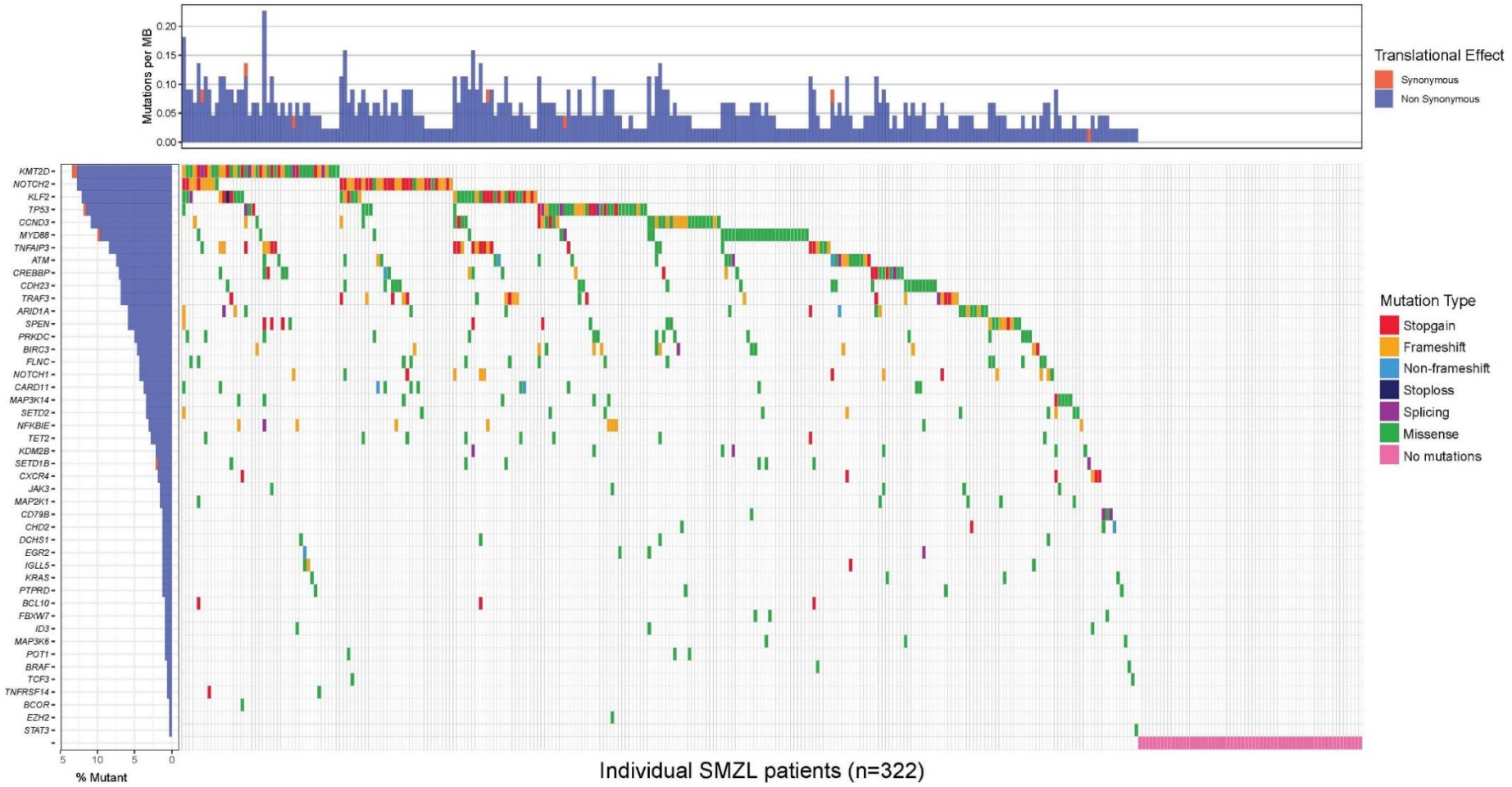


Figure 7-5. Waterfall plot of all mutations found in Jaramillo-Parry cohort. Each column represents unique SMZL patients and each row represent a gene. The mutation burden is calculated as mutations in sample/coverage space∗1,000,000 where the coverage space is the theoretical coverage space of the exome reagent "SeqCap EZ Human Exome Library v2.0". This will underestimate the mutation burden in genes where all exons were not targeted, however It was kept as this data was targeted using three different capture kits.

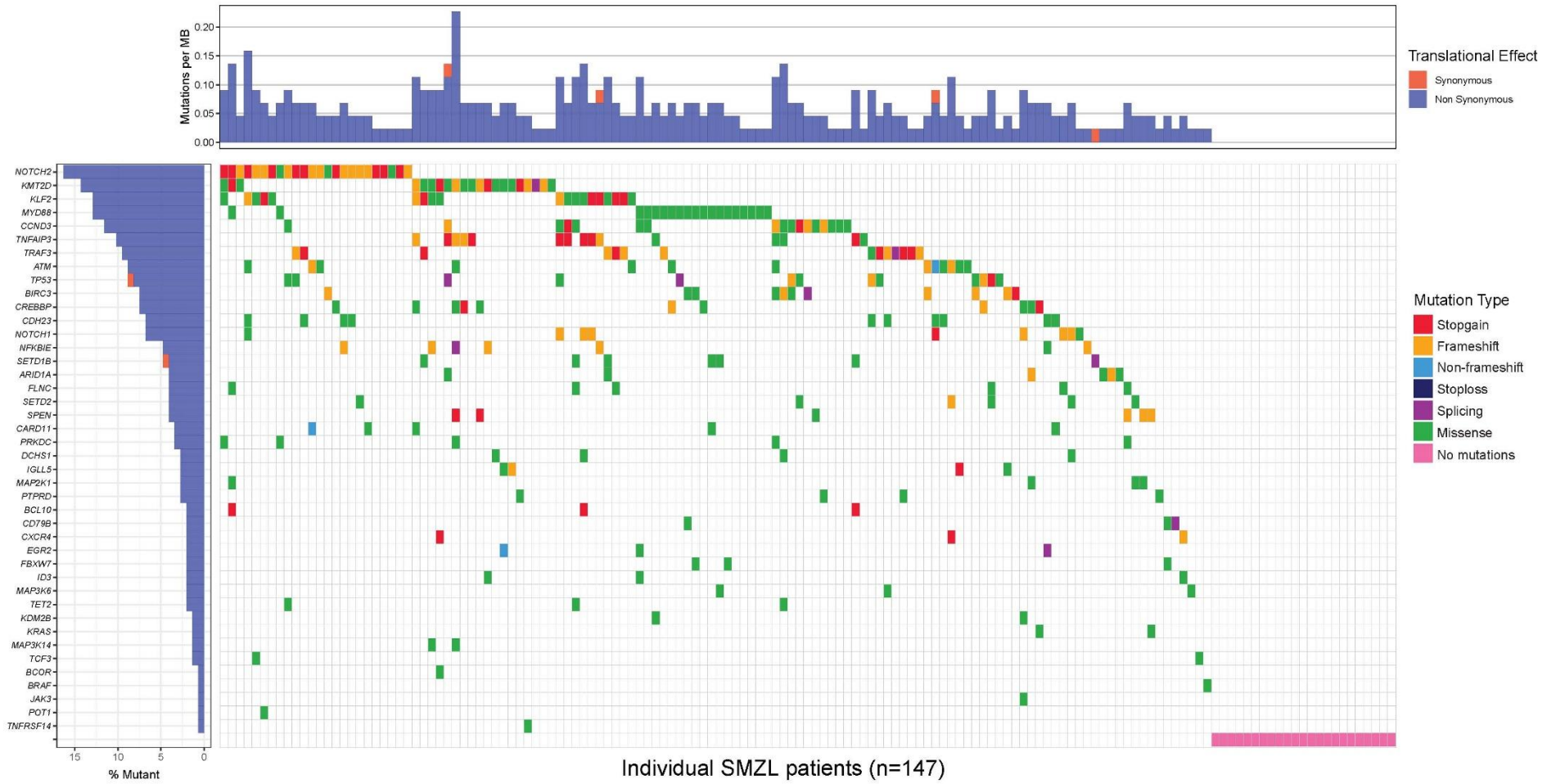


Figure 7-6. Waterfall plot of Jaramillo cohort. Each column represents unique SMZL patients and each row represent a gene. The mutation burden is calculated as mutations in sample/coverage space∗1,000,000 where the coverage space is the theoretical coverage space of the exome reagent "SeqCap EZ Human Exome Library v2.0". This will underestimate the mutation burden in genes where all exons were not targeted, however It was kept as this data was targeted using three different capture kits.

Table 7-1. Rank of the 15 most frequently mutated genes in across the Jaramillo and Parry cohorts.

Rank (most frequently mutated genes)	Jaramillo cohort [n=146]		Parry cohort [n=175]	
	Gene (frequency in cohort %)	Number of patients with mutations	Gene (frequency in cohort %)	Number of patients with mutations
1	*NOTCH2* (16.4%)	24	*TP53* (14.2%)	25
2	*KMT2D* (13.7%)	21	*KMT2D* (12.6%)	22
3	*KLF2* (13.0%)	19	*KLF2* (11.4%)	20
4	*MYD88* (13.0%)	19	*CCND3* (10.3%)	18
5	*CCND3* (11.6%)	17	*NOTCH2* (9.7%)	17
6	*TNFAIP3* (10.3%)	15	*ARID1A* (7.4%)	13
7	*TRAF3* (9.5%)	14	*CDH23* (7.4%)	13
8	*ATM* (8.9%)	13	*MYD88* (7.4%)	13
9	*TP53* (8.9%)	13	*SPEN* (7.4%)	13
10	*CDH23* (8.2%)	12	*CREBBP* (6.8%)	12
11	*BIRC3* (7.5%)	11	*TNFAIP3* (6.8%)	12
12	*CREBBP* (7.5%)	11	*ATM* (6.3%)	11
13	*NOTCH1* (6.8%)	10	*PRKDC* (6.3%)	11
14	*NFKBIE* (4.7%)	7	*MAP3K14* (5.2%)	9
15	*SETD1B* (4.7%)	7	*TRAF3* (4.5%)	8

KMT2D as mentioned before, had the greatest mutational frequency with 43/321 affected samples (13.4%). Variants were distributed throughout the protein (**Figure 7-7**) with no obvious clustering pattern or recurrent hot-spot mutations. Most mutation in *KMT2D* were missense [n=23], then frameshift [n=11], stopgain [n=9] and splicing [n=4]. Two of the splicing mutations (p.L2061L and p.P2036P) were classified as synonymous by Annovar but splicing by spliceAI.

KMT2D was closely followed by *NOTCH2* with a mutational frequency of 12.8% [41/321 samples]. *NOTCH2*, had a similar number of stopgain [n=19] and frameshift [n=18] mutations, and a small number [n=6] of missense mutations. Most of the variants in *NOTCH2* clustered around the end of the C-terminal PEST domain (**Figure 7-7**) and it is in this domain where recurrent variant p.P2303fs was identified in three patients.

KLF2 followed with a mutational frequency of 12.1% [39/321 samples]. Most of its variants were missense [n=22], followed by stopgain [n=11], frameshift [n=8], non-frameshift [n=1], stoploss [n=1] and splicing [n=1]. Variants in *KLF2* formed two distinct clusters, one at the start (first 50 amino acids) of the linear protein and one at the end (last 100 amino acids). There were two recurrent variants in this gene, p.Q24X and p.H288Y/D, found in seven and six patients respectively. Neither one of these recurrent variants is found in COSMIC or in any of the databases of known germline variation.

TP53 had a mutational frequency of 11.5% [37/321 samples] but most of its variants were missense [n=26]. *TP53* also had stopgain [n=6], frameshift [n=5] and splicing [n=4] mutations. One of the splicing variants p.E224E was classified as synonymous by Annovar but was re-classified as splicing by spliceAI. This splicing variant (p.E224E) is found in COSMIC (COSV52681634) and classed as likely pathogenic by ClinVar. Most mutations in *TP53* fell within the within the DNA binding domain where 6/37 cases were found both mutated and deleted.

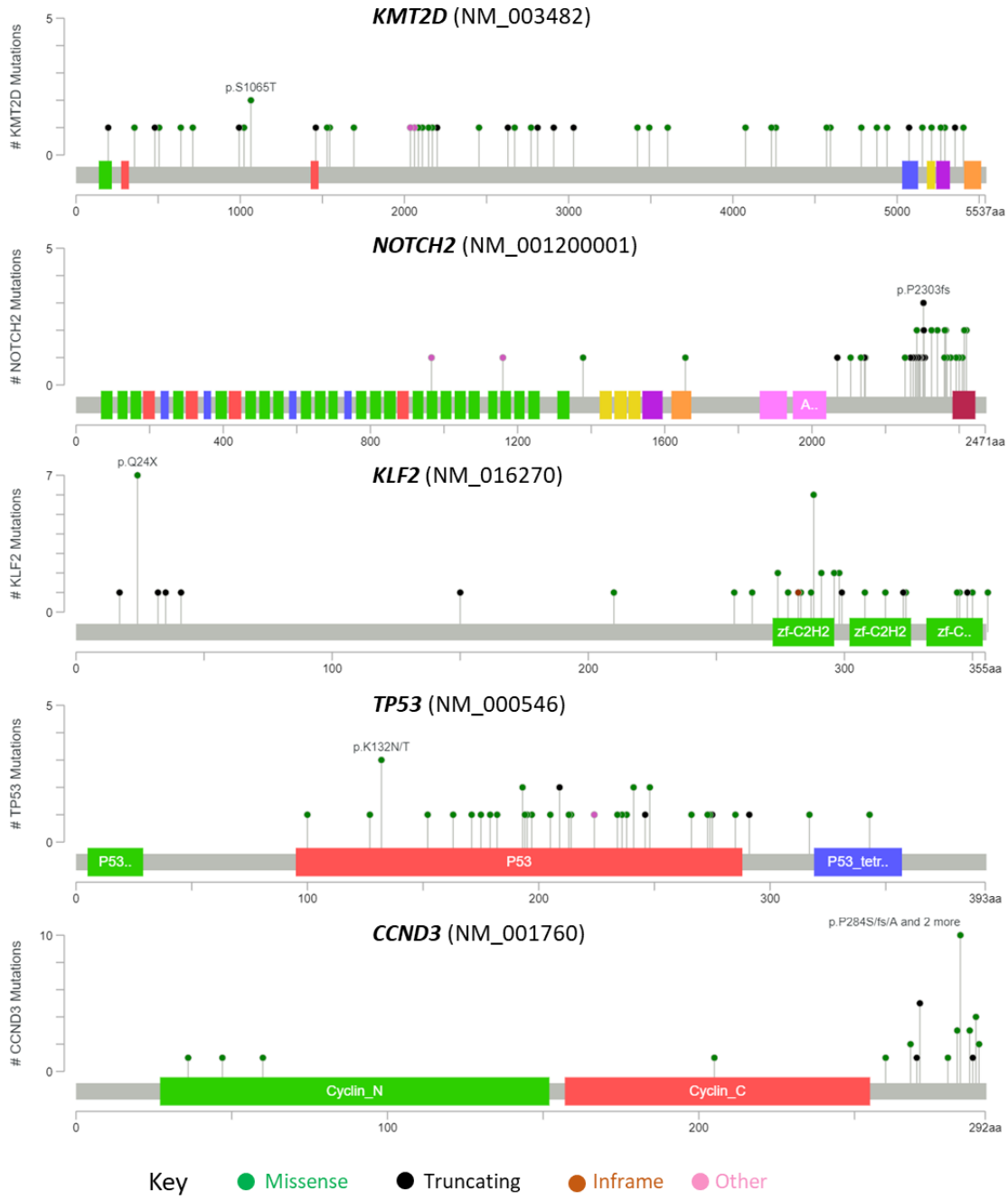


Figure 7-7. Lolliplot of the five most mutated genes in the Jaramillo-Parry cohort. Each lolliplot illustrates a linear protein representing each gene with its respective domains. The height is representative of the number of variants reported (Y-axis differs per gene) and circle colours identify mutation type. The transcript used for each protein is found next to the gene name and the colours of the domains were randomly assigned by the tool. Drawn using cBioPortal[155].

The fifth most recurrently mutated gene was *CCND3* with a mutational frequency of 10.9% [35/321 samples]. Most variants within this gene were missense [n=22], followed by frameshift [n=12] and stopgain [n=3]. Variants in *CCND3* clustered in exon 5 and there were two loci containing 40% of the variants. The first locus was on amino acid 284 where there were seven amino acid changes identified within ten patients. The seven different nucleotide changes found are described in **Table 7-2**. The second locus was on amino acid 271 where the variant p.R271fs was identified in five patients.

Table 7-2. Recurrent mutation in gene *CCND3* found in the Jaramillo-Parry cohort. Locations based on hg38 reference genome.

chr	start	end	ref	alt	AA change	Type of mutation	No. Patients
Chr 6	41935956	41935969	GTGACATCT GTAGG	-	p.P284fs	Frameshift	1
Chr 6	41935965	41935968	GTAG	-	p.P284fs	Frameshift	1
Chr 6	41935967	41935967	-	GGAGTGCT GGTCTGGC TGGGCTT	p.P284fs	Frameshift	1
Chr 6	41935968	41935968	G	T	p.P284H	Missense	1
Chr 6	41935969	41935969	G	C	p.P284A	Missense	1
Chr 6	41935969	41935969	G	A	p.P284S	Missense	4
Chr 6	41935969	41935969	G	T	p.P284T	Missense	1

MYD88 had a mutational frequency of 10.0% [32/321 samples] within the combined Jaramillo-Parry cohort. In the waterfall plot (**Figure 7-5**) it is easy to distinguish this gene as it is perhaps one of the few genes with exclusively missense mutations (green). **Figure 7-8** shows a clear recurrent variant (p.L252P) found in 16 patients. Although variant p.L252P is labelled as a variant of uncertain significance in ClinVar, it is found in COSMIC (ICOSV57169334) and has a high CADD Phred score of 31. All other mutations within *MYD88* cluster within the toll/interleukin-1 receptor homology (TIR) domain. *TNFAIP3, CREBBP, ATM* and *CDH23* all have similar mutational frequencies of 8.4%, 7.5%, 7.1%, and 6.8% respectively. Mutations in these genes did not cluster within any particular region or domain.

Figure 7-8. Lolliplot of the six to ten most mutated genes in the Jaramillo-Parry cohort. Each lolliplot illustrates a linear protein representing each gene with its respective domains. The height is representative of the number of variants reported (Y-axis differs per gene) and circle colours identify mutation type. The transcript used for each protein is found next to the gene name and the colours of the domains were randomly assigned by the tool. Drawn using cBioPortal[155].

7.4.2 Variant allele frequency across genes

The variant allele frequency (VAF) of all mutations was plotted against depth (**Figure 7-9**). Only 24/637 variants had a VAF ≤ 0.12 perhaps highlighting the limitations of our data and bioinformatics processing (unmatched tumour data processed through a germline variant caller). There is also a noticeable high density of variants around 0.50 VAF which could be potential rare germline variants that were not excluded.

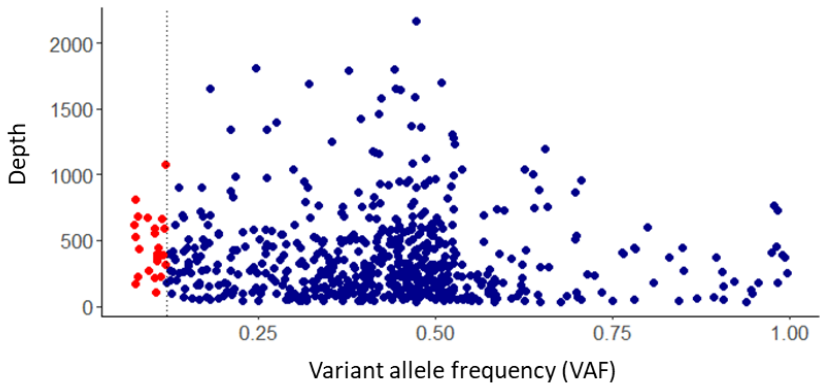


Figure 7-9. Variant allele frequency vs depth of all variants in the Jaramillo-Parry cohort. Each dot represents a single variant. Blue dots have a VAf > 12% whole red dots a VAF ≤ 12%. Minimum read depth is 30.

The VAF distribution was plotted per gene for the 20 most frequently mutated genes (**Figure 7-10**). For *ATM, KLF2, KMT2D, MYD88* and *NOTCH2* the VAF distribution was very similar all within a range of 0.10 and 0.75. While *TP53, TNFAIP3* and *CCND3* had variants with VAFs up to 1. *CDH23* seemed the most problematic since most of its variants were between 0.40 and 0.60 VAF.

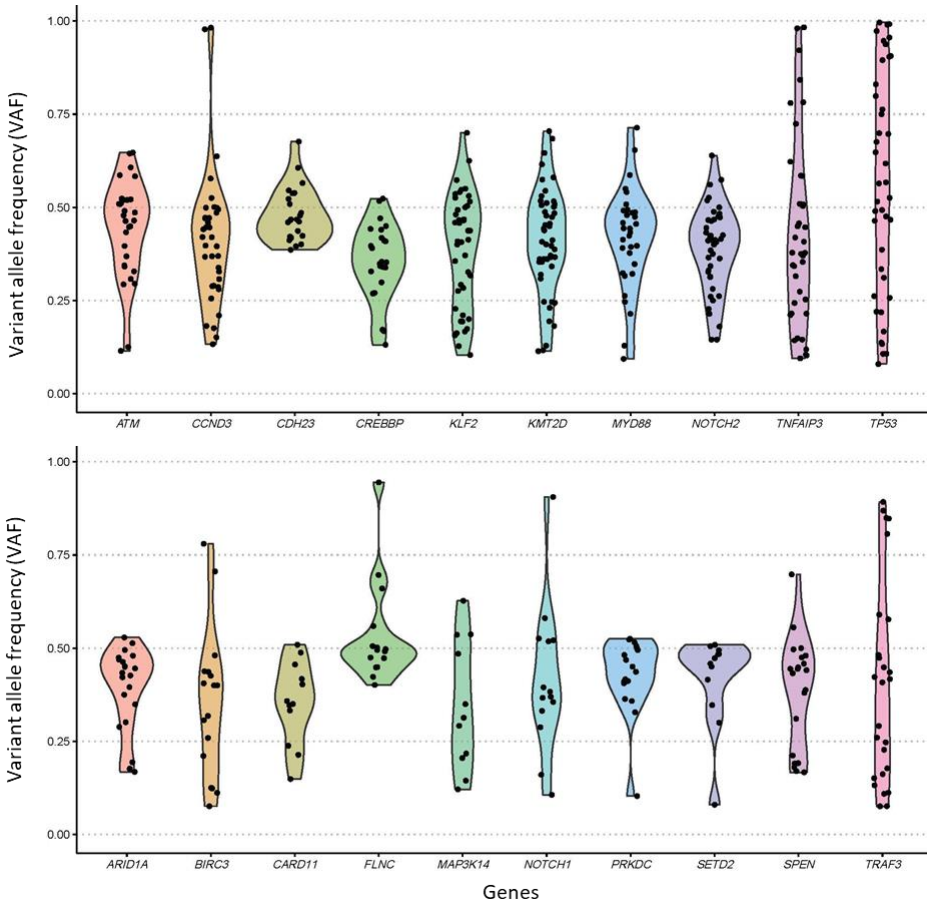


Figure 7-10. VAF distribution across the 20 most mutated genes in the combined Jaramillo-Parry cohort.

7.4.3 Associations between genes

Results of the Fisher's Exact test to determine associations between genes showed many co-occurring genes but no mutually exclusive interactions, likely due to the number of affected

samples (**Figure 7-11**). Most notably *KLF2* and *TNFAIP3* were the only pair to retain significance after Bonferroni correction ($p < 0.001$). Out of the 39 cases that had a *KLF2* mutation, 28% [n=11] also harboured *TNFAIP3* mutations. *KLF2* mutations also co-occurred with mutations in *KMT2D* [n=10], *TRAF3* [n=7], *CARD11* [n=4], and *SETD1B* [n=3]. Interestingly *KLF2* did not show a significant association with *NOTCH2* even though 9 cases had co-occurring mutations, considering this is an association that has been reported in the literature. However, *NOTCH2* did co-occur with *KMT2D* [n=10] and *CARD11* [n=5] mutations. Another interesting gene was *CCND3* which co-occurred along *TP53* [n=9], *TET2* [n=4] and *BIRC3* [n=6] mutations, these interactions though were mostly within the Jaramillo cohort (**Figure 7-12**).

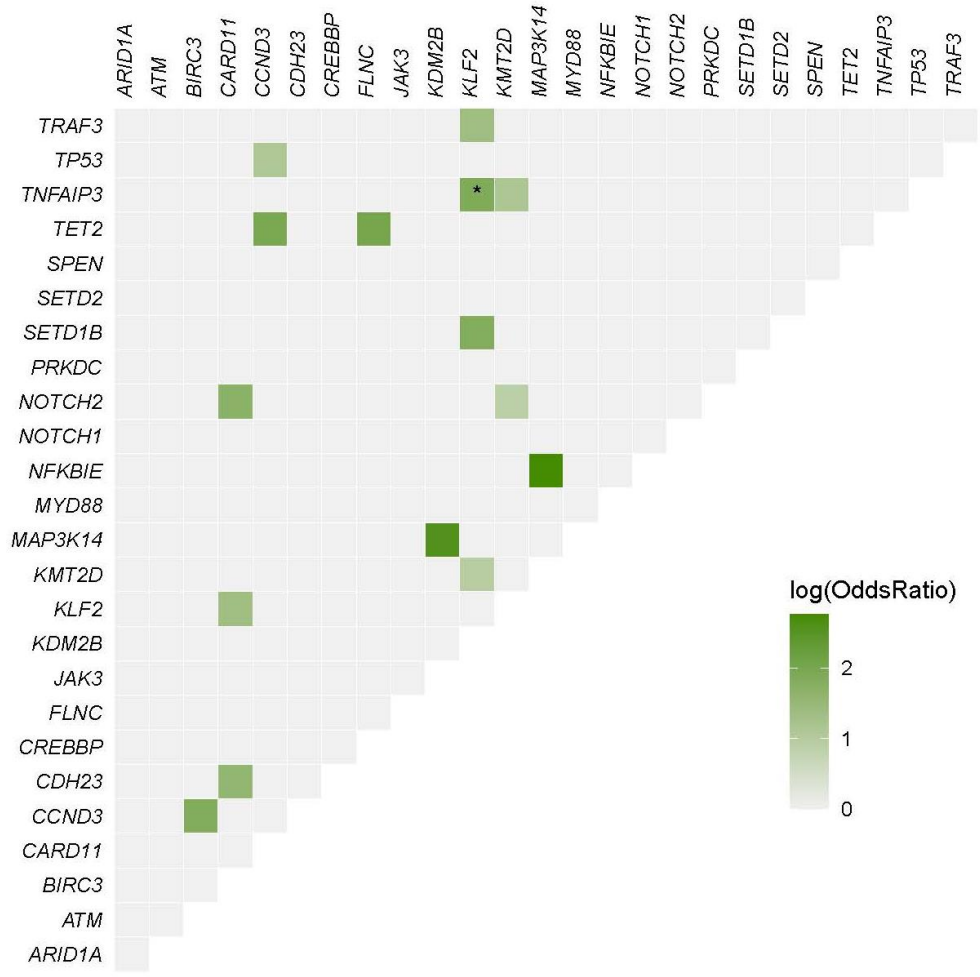


Figure 7-11. Results of Fisher's Exact test in combined Jaramillo-Parry cohort. Figures only show co-occurring interactions between genes (green), mutually exclusive interactions were not identified with the test. Only interactions with a p-value < 0.05 were coloured. * indicates significance after correction (p-value < 0.001).

Separate associations tests were carried out for the Jaramillo and Parry cohort. Both showed the same *KLF2* and *TNFAIP3* interactions, however without retaining significance after correction (**Figure 7-12**). The Jaramillo cohort had six cases with co-occurring *KLF2* and *TNFAIP3* mutations and the Parry cohort had five. The two cohorts did not share any other co-occurring associations.

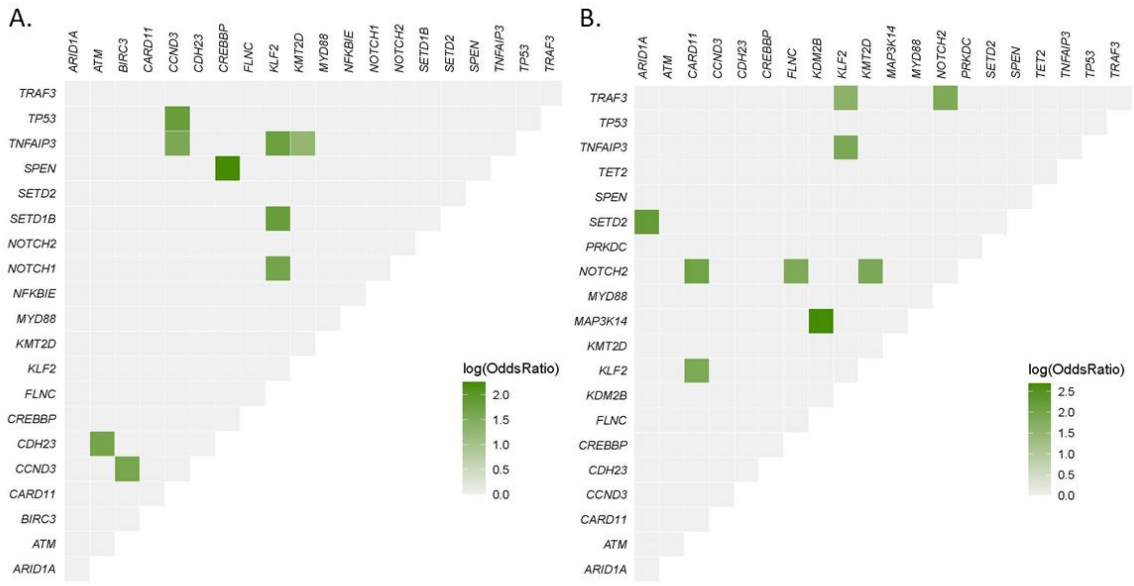


Figure 7-12. Results of Fishers exact test in Jaramillo and Parry cohorts. **A.** Somatic interactions within the Jaramillo cohort. **B.** Somatic interactions within the Parry cohort. Figures only show co-occurring interactions between genes, mutually exclusivity interactions were not identified with the test. Only interactions with a p-value < 0.05 were coloured and none retained their significance after correction.

7.5 Discussion

7.5.1 Filtering strategies

One of the biggest tasks in this project was to find a way to process tumour only SMZL samples with the least number of false positives and likely germline variants. The application of the machine learning model to triage variants into low, medium, and high confidence variants aided in the exclusion of roughly 50% of false positives which did not have to be manually reviewed in IGV. However, after validation of the remaining variants in IGV, assessment of the functional impact showed some evidence of potential germline variant still present in the data. This included variants that had high frequency within the cohort, variants that were classified as benign or likely benign by ClinVar and variants that were present in databases of known germline variation. A good filtering strategy was therefore key in obtaining the most biologically relevant list, but it had to be done in a systematic way so as not bias the results. Initially this involved exclusion of variants found with a frequency greater than 1% in databases of known germline variation and using a CADD cut-off of 20. However, the authors of CADD advice against the use of a single universal cut-off value for two main reasons: 1. CADD is a continuous score and binarizing it would result in the loss of information and; 2. The cut-off value would depend on analysis of specific factors, like the severity of the phenotype, whether the variant is dominant or recessive and the time available for curation and wet lab follow up of the variants[89]. The authors pointed to two

methods that did use CADD scores with hard cut-offs; GAVIN[156] and MSC[151]. After consideration the MSC was integrated into the analysis and was used along the CADD score for a consensus and gene specific approach.

After using the CADD scores and MSC cut-offs to exclude likely benign variants, results continued to show evidence of potential rare germline variation (i.e. VAF close to 0.50, found in multiple samples, likely benign in ClinVar). This led to the use of the VAF and the binomial distribution to exclude variants, but this was done with caution since VAFs were not corrected for tumour purity and somatic variants may also have a VAF close to 0.50. To address this issue only those variants that were present at frequencies greater than 0.01% in the databases of germline variation were assessed in this way. We cannot guarantee that the final variant list is comprised of only somatic variants, however, we are confident that it is representative of the disease. We found recurrently mutated genes previously implicated in MZ development (*KLF2, NOTCH2),* genes targeting NF -κB signalling (*MYD88, TNFAIP3, BIRC3, CARD11*), cell cycle (*TP53*, *CCND3, ATM*) and epigenetic modifiers (*KMT2D, CREBBP, ARID1A*). Most of the aberrant genes and pathways had been previously identified and were briefly discussed in our systematic literature review in **Chapter 2**.

7.5.2 Mutations targeting MZ B-cell development

KLF2 belongs to the family of Kruppel-like transcription factors, a subfamily of the zinc-finger class of DNA binding transcriptional regulators[157]. Kruppel like transcription factors play a key role in diverse biological processes, including cell growth and differentiation, embryogenesis and tumorigenesis[158]. KLF2 directly binds to promoters regulating gene expression genes involved in cycle control, cell homing and NF-κB signalling[54]. In murine systems, loss of KLF2 drives the germinal cells to a MZ-like phenotype and preclusion of migration to the splenic MZ[159–161], thereby preventing germinal centre B-cell responses to antigens in the MZ. Nuclear localization of the KLF2 protein and consequent DNA binding require three C-terminal highly conserved zinc finger domains and two nuclear localization sequences, respectively. Mutations in this gene were enriched within the highly conserved zinc finger domain regions as well as the activation domain. Mutations in the zinc finger domains consisted mostly of missense mutations, while those in the activation domain were mostly frameshift and stopgain. Studies have shown that missense substitutions within the nuclear localization sequences of *KLF2* or within the highly conserved regions of the zinc finger domains truncates KLF2 function and hinders the ability of KLF2 to suppress NF-κB signalling pathways, key to marginal zone B-cell development[26,54]. Within the activation domain there is one variant (p.Q24X) that has been reported in several studies suggesting a mutation hotspot[54,71,72,87]. Although no functional evidence on this specific variant is available, it is very likely that due to its position on the first exon it would result in a truncated and

non-functional protein. Furthermore, this variant has a CADD Phred score of 36 indicating that it is predicted to be the 0.1% most deleterious substitutions you can do to the human genome.

In murine models, and to a lesser extent in humans also, NOTCH2 plays a key role in MZ B-cell maturation and MZ retention[162–165]. NOTCH2 is a cell-surface receptor belonging to a family of evolutionarily conserved trans-membrane proteins. Notch pathways regulate cell proliferation, cell fate, differentiation, and cell death[166]. When a ligand binds to the extracellular domain of a Notch receptor, it initiates a cascade of proteolytic cleavages that lead to the detachment of the notch intracellular domain (NICD), which then moves into the nucleus to interact with target transcription factors[166,167]. In our cohort *NOTCH2* mutations target the C-terminal PEST domain on exon 34, necessary for the regulation of the intracellular domain and consequent transcriptional regulation. Based on the location of variants in exon 34, mutations are predicted to either eliminate or truncate the PEST domain, predicted to prolong the half-life of NICD2 and therefore increase NOTCH2 expression[168]. In a recent study of NOTCH2 activation, Shanmugan and colleagues propose that in the case of SMZL, NOTCH2 mutations are not initiating events, but that the sustained signalling provided by these mutations provides a selective advantage in tumours that have already established themselves in a ligand-rich microenvironment[169]. This is supported by evidence in mice where those with activating Notch2 mutations in mature B-cells displayed expansion of the marginal zone but did not develop lymphoma[170]. Expression of the Notch intracellular domain 2 (NICD2) can be detected in SMZL cases and is a common feature of both NOTCH2 wild-type and mutated SMZLs[169], similar to prior findings with NOTCH1 in CLL[171], suggesting that Notch activation is a general feature of SMZL tumour cells. The work by Shanmugan and colleagues showed higher frequency of NICD2+ cells in mutated versus wild-type tumours and higher in the marginal zones of the white pulp. It is yet to be determined if enhanced NICD2 expression in wild-type tumours is explained fully by mutations in other Notch regulators, such as NOTCH1 (~5%) and SPEN (~5%)[50,71], structural or copy number aberrations, or by the enrichment of NOTCH2 in the normal counterpart of SMZL.

7.5.3 Mutations targeting NF-κB pathway

NF-κB signalling plays an essential role in MZ B-cell development and differentiation[172]. When normal B lymphocytes respond to antigens, NF-κB signalling is activated, reprogramming cells to favour cell cycle progression, survival, cytokine secretion and inflammation[173,174]. NF-κB activation, through either the canonical or non-canonical pathway, is transient in normal cells and depends on external stimuli including ligands for the BCR and for the Toll-like receptors[173], while termination of signalling is dependent on negative feedback mechanisms including re-accumulation of IκBα and induction of TNFAIP3(A20)[174]. Canonical activation of NF-κB signalling is

essential for marginal zone B-cell development and differentiation[172] and it is a pathway commonly affected by genetic lesions in SMZL.

Results from our approach identified recurrently mutated genes belonging to the NF-κB pathway, as well as upstream pathways connected to NF-κB activation[173]. Most notably *TNFAIP3* (A20), a negative regulator of NF-κB signalling, was found mutated in 8% of cases. Other negative regulators that harboured mutations included *TRAF3* (7%) and *BIRC3* (5%). Both *TRAF3* and *BIRC3* are part of the regulatory system that negatively regulates *MAP3K14* (3%), a central activator of noncanonical signalling and another target of mutations in SMZL[84].

TNFAIP3 (A20) acts as a tumour suppressor since it brakes canonical NF-kB activation[175–178]. Within our cohort 31/39 *TNFAIP3* variants were inactivating mutations (17 frameshift and 14 stopgain), where 2/31 were biallelic mutations and the rest monoallelic. Here TNFAIP3 may function in a haplo-insufficient manner since previous studies showed that both biallelic and monoallelic inactivation of TNFAIP3 can induce NF-kB activation[179]. It is also possible that some of the monoallelic mutations in our cohort are accompanied by deletion of the locus, which will be discussed in the next chapter.

The majority of *TRAF3* mutation (22/26) were also inactivating mutations (13 frameshift and 9 stopgain). These mutations are predicted to truncate or eliminate the C-terminal MATH domain, required for MAP3K14 docking, necessary for BIRC3 degradation[84]. Like *TRAF3, BIRC3* had mostly inactivating mutations (9 frameshift and 1 stopgain) predicted to eliminate or truncate the RING domain, necessary for ubiquitin-mediated proteasomal degradation of MAP3K14[180]. In the study by Rossi and colleagues SMZL primary cells with monoallelic TRAF3 inactivation showed NF-κB localization, MAP3K14 accumulation, and active NFKB2 (p52) processing, while those with BIRC3 monoallelic inactivation displayed constitutive NF-κB activation, MAP3K14 accumulation, and active NFKB2 (p52) processing[84].

MAP3K14 (NIK) harboured mostly synonymous variants and one stopgain variant. 7/11 *MAP3K14* variants fell within the TRAF3 binding domain. This domain mediates the interaction with TRAF3 which holds MAPK314 inactive in the TRAF2/3/cIAPs (cellular inhibitor of apoptosis protein 1 and 2) complex[181]. 3/11 variants fell within the non-catalytic region domain (NRD), which allows protein binding to IKKα and p100[181] and might have a different effect (downstream signalling activation) then those that fall within the TRAF3 binding domain.

Activating mutations in positive regulators of NF-κB signalling were also found within our cohort, mainly in *CARD11* (4%) and *MYD88* (10%). *CARD11* harboured mostly missense mutations [n=10] and two frameshift deletions. 8/12 variants were within the coiled-coil domain which interacts

with the proteins inhibitory domain (ID) keeping CARD11 inactive in the absence of antigen receptor engagement[182]. Studies have shown that mutations both within the ID and coiled-coil domain confer a gain of function phenotype[182,183].

Toll-like receptor (TLR) signalling plays a key role in SMZL biology, as cellular proliferation is driven by TLR activation. MYD88 is an adaptor protein essential for proper TLR signal transduction which has several structural domains including a death domain responsible for oligomerization and interactions with IRAK1-4, that together lead to activation of NF-κB. The toll/interleukin-1 receptor homology (TIR) domain, at the proteins C-terminus, is responsible for the activation of downstream signalling. Within this cohort, *MYD88* harbours a recurrent variant (p.L265P) in the toll/interleukin-1 receptor homology (TIR) domain. There is mounting evidence suggesting that this mutation is an oncogenic driver, considering its high mutation frequency in some entities such as Waldenström's macroglobulinemia (WM) where up to 90% of patients harbour this mutation[184].

7.5.4 Mutations targeting epigenetic regulators

Chromatin remodelling enzymes are dynamic modulators of cell identity and regulate B-cell differentiation and proliferation through recognition of a spectrum of specific biochemical marks on histone proteins and DNA; thereby modifying chromatin accessibility to transcription machinery proteins[185].

KMT2D also known as *MLL2* belongs to a family of histone lysine methyltransferases that modifies lysine-4 of histone 3 (H3K4), and has an established tumour suppressor role in DLBCL and FL[186–188]. Loss of function of KMT2D can lead to the altered abundance and distribution of H3K27me3. This is because the polycomb repressor complex 2 (PRC2), a group of proteins that regulate chromatin compaction and gene expression, is unable to methylate H3K27 if H3K4 is trimethylated on the same histone tail[189]. *KMT2D* being the most recurrently mutated gene in the combined cohort was surprising since it has never been reported at such high frequencies in SMZL. Within the combined Jaramillo-Parry cohort 20/47 variants were inactivating mutations (11 frameshift and 9 nonsense). In DLBCL and FL the majority of mutations within this gene are nonsense or frameshift events, and the nonsense mutations affect the C-terminal portion of the gene[190]. **Figure 7-7** shows that *KMT2D* mutations are distributed across the protein and missense mutations do not cluster clearly at any point. It is difficult to establish how these nonsense mutations could affect the expression of the protein and it is plausible that many of the nonsense mutations could be rare germline variants, a possible explanation for the discrepancy between the observed and reported mutational frequency. This gene was particularly difficult to analyse since it is a large gene (55 exons) with highly repetitive regions that were hard to map and therefore permissive of many

false positives. However, our filtering strategy was extremely thorough particularly in eliminating false positives. Another potential cause of the discrepancy between observed and reported mutations frequencies, could be that even though *KMT2D* has been shown to be mutated in SMZL before, it has not been the main target of most studies.

CREBBP is a lysine acetyltransferase involved in the co-activation of many different transcriptional factors. It does this by acetylating histones at regulatory elements altering their charge and therefore loosening their associations with DNA making it more accessible to transcription factors. Mutant *CREBBP* are deficient in acetylating BCL6 and p53. BCL6 is necessary for GC development and acetylation of BCL6 leads to inactivation of its transcriptional repressor function, while acetylation of p53 is necessary for its transcriptional activity[191]. 10/25 *CREBBP* mutations were inactivating (3 frameshift and 6 nonsense), there was a single non-frameshift deletion and splicing mutation and 14 missense mutations. In FL most missense mutations were observed within the KAT domain[190], however in our cohort the missense mutations did not cluster in a particular locus or domain (**Figure 7-8**).

ARID1A (BAF250a) is a component of SWI/SNF (SWItch/Sucrose Non-Fermentable) family of evolutionary conserved, multi-subunit chromatin remodelling complexes, found mutated in various cancers[192]. SWI/SNF regulates DNA accessibility to other proteins involved in replication and repair, allowing the activation or suppression of gene transcription[185]. Mutations inactivating ARID1A results in the loss of both the caretaker and gatekeeper function in cells[193]. Within our combined cohort 7/18 variants were inactivating (6 frameshift and 1 nonsense), the rest were missense mutations. Like *KMT2D* and *CREBBP,* missense mutations in *ARID1A* did not cluster around any specific locus or domain, making it difficult to predict the functional effect of these mutations.

Inactivating somatic mutation in *KMT2D* and *CREBBP* truncate core epigenetic mechanisms that drive GC-derived lymphomas. It will be interesting to investigate the relationship between these mutations and any epigenetic fingerprint associated with the cell of origin in SMZL.

7.5.5 Mutations targeting cell cycle control

TP53 is one of the main SMZL associated genes implicated in cell cycle control, along with *CCND3* and *ATM*. However, the lack of germline material, make it difficult to confirm whether *ATM* mutations are truly somatic and should be looked at with caution. This is reinforced by the observed VAFs within *ATM* mutants being very close to 0.50. Most *TP53* mutations fell within the DNA binding domain, attenuating, or eliminating its function as tumour suppressor, since mutant proteins lose the ability to activate canonical p53 target genes. Only three variants fell outside the

DNA binding domain (1 frameshift and 2 stopgain). These three variants would likely eliminate or truncate the tetramerization domain, critical for protein-protein interactions. Mutant p53 leads to uncontrolled cell proliferation and permissive accumulation of genomic mutations that may culminate in tumour growth[194].

CCND3 was another surprising gene, since the mutational frequency in our cohort was much higher than in published literature. CCND3 (Cyclin D3) is a D-type cyclin which regulates the progression of cells into the G_1 phase of the cell cycle[195] (see **section 1.1**). D-type cyclins (D1, D2, D3) assemble with cyclin-dependent kinases CDK4 and CDK6 and phosphorylate retinoblastoma proteins (RBs) leading to their degradation and convert cells from a G1-specific to an S-phase-specific transcriptional mode[195]. In B-cells CCND3 is required for proliferative expansion of pre B-cells and in GC B-cells it is required for GC development[196,197]. Mutations in CCND3 will pre-dispose B-cells to acquire further somatic mutations or chromosomal re-arrangement which could then lead to tumour formation[198]. Studies have shown CCND3 to work in synergy with other factors to induce lymphomagenesis[199,200]. In our cohort 33/37 variants fell within the C-terminal PEST domain of CCND3, where 15 of the 33 variants were either frameshift or nonsense. Mutations within the PEST domain of *CCND3*, like those in *NOTCH2*, will stabilize the protein and promote cell proliferation[201].

7.6 Conclusions

These results confirm the importance of *NOTCH2, KLF2* and *TP53* in SMZL, and validate our findings from the systematic literature review. We provide further evidence that *KMT2D* and *CCND3* are a frequently mutated genes in this cancer and are found at much higher frequencies than previously observed in smaller cohorts. Recurrent variants in *KLF2* and *CCND3* also point to possible new mutation hotspots within these genes that will require functional validation. Furthermore, we validate the importance of NF-κB regulation in the disease considering a large percentage of samples harbour mutations within this pathway.

Chapter 8 Integration of genomic results and clinical data of SMZL patients

8.1 Synopsis

This chapter will focus on the integration of the genomic results from Chapter 7 to other molecular biomarkers, such as *IGHV* gene usage and mutations, telomere length, copy number alterations as well as clinical outcomes.

David Oscier reviewed patient diagnoses and provided clinical guidance. Dr. Helen Parker collated all the clinical data from collaborating centres and performed the wet laboratory work including the telomere length assays. Dr. Dean Bryant processed the methylation data to obtain copy number status, plotted the copy number profiles and generated the text output containing the segment locations. Carolina Jaramillo Oquendo manually curated copy number calls, and these were validated by Dr. Helen Parker. Carolina Jaramillo Oquendo processed and analysed the data under the supervision of Prof Sarah Ennis, Prof Jonathan Strefford and Dr. Jane Gibson.

8.2 Introduction

8.2.1 Genomic alterations in SMZL

Conventional chromosomal banding has shown that approximately 75% of SMZL cases display an abnormal karyotype, with 50% exhibiting a complex karyotype (defined as three or more cytogenetic aberrations). Structural chromosomal aberrations (75%) are more common than trisomy/monosomy events (25%), and gains are more prevalent than losses[47]. The most frequent cytogenetically visible aberrations are gain of 3q (20-30%) and 12q (20%) and deletions of 7q (30-40%), less frequently 1q, 6q, 8q and 14q are targeted (**Table 8-1**). Roughly 10% of SMZL cases show evidence of translocations involving the immunoglobulin heavy chain genes at 14q32. These translocations remain relatively under-studied, but include the t(14;19)(q32;q13), t(6;14)(p21;q32), t(9;14)(p13;q32) and t(1;14)(q21;q32) which target the genes *BCL3, CCND3, PAX5* and *BCL9/MUC1*, respectively[47]. Furthermore, SMZL lacks recurrent chromosome translocations and subsequent gene fusions that are typical in other lymphomas such as the t(14;18), t(11;18) and t(1;14) which affect *BCL2, BIRC3/MALT1* and *CCND1* genes respectively[86].

Table 8-1. Summary of recurrent chromosomal aberrations in SMZL. Table reprinted from Jaramillo Oquendo et al[55] licenced under CC BY 4.0.

Chromosome	Frequency	Target genes / Clinico-biological associations
7q-	14% - 44%	Unknown target gene(s) High frequency in SMZL compared to other MZL[45,202]. Associated with unmutated *IGHV* genes, *IGHV1-2*04* usage, and *KLF2* and *NOTCH2* somatic mutations[42,47,71].
17p-	5% - 32%	*TP53* gene Associated with worse prognosis in univariate analysis[47].
6q-	8% - 24%	A fraction of cases include deletion of *TNFAIP3* (negative regulator of NF-κB)
8p-	4% - 15%	Associated with poor outcome in MZLs. No link with outcome in SMZL unless co-existing with deletion of 17p [202].
13q-	5% - 18%	SMZL cases with this lesion showed a genetic profile consistent with SMZL diagnosis[202]
14q-	3% - 10%	Linked to inferior prognosis but detected in the context of a complex karyotype[47]
+3/3q+	15% - 34%	Two gained regions, one included BCL6 located at 3q27 Associated with complex karyotypes Tend to occur in cases without del 7q 3q gains have been associated with gains at 1q and 17q22-q25.3, del 6q23.2-q24.1 (*TNFAIP3*) and del 6q25[203]
+12/12q+	8% - 25%	Trisomy 12 & use of VH3 family variable gene segment were found significantly associated with worse OS in univariate analysis but lost significance in multivariate analysis[45] Associated with gain of chromosome 3[47]
+18/18q+	8% - 23%	Mutually exclusive to 7q deletions[45] Associated with gain of chromosome 3[47]
8q+	2% - 20%	Gains of 8q that include *MYC* gene locus were associated with poor clinical outcomes[204]

8.2.2 Immunoglobulin genes

Seminal immunogenetic studies have been performed on a myriad of mature B-cell neoplasms, including SMZL, where they have identified key features of the B-cell receptor immunoglobulin repertoire indicating that clonal B-cell selection by antigen/superantigens is an important feature of SMZL pathophysiology. Investigation of large SMZL cohorts has demonstrated bias usage of immunoglobulin heavy chain genes, namely enrichment of the *IGHV1-02* (30%), *IGHV4-34* (11%) and *IGHV3-23* (9%) genes[205,206]. The majority of *IGHV1-02* cases are *IGHV1-02*04* which is striking given that this allele is considerably less frequent in other B-cell tumours[205,206]. Additionally, evidence of somatic hypermutation (SHM) is seen in the majority of *IGHV1-02*04* SMZL cases (~95%) suggesting exposure to antigen is both important in progenitor tumour cell selection but also relevant to ongoing evolution[207].

8.2.3 Clinical utility of molecular lesions

Diagnosis of SMZL can be established through a combination of lymphocyte morphology and flow cytometry, bone marrow biopsy and immunohistochemistry[19,37]. Unfortunately, several mature B-cell tumours, such as splenic diffuse red pulp lymphoma (SDRPL), have overlapping clinicopathological and immunophenotypic features with SMZL. Therefore, in a minority of cases it is difficult to diagnose SMZL in the absence of spleen histology[38,40]. Additionally, there are no recommended biomarkers to differentiate between similar lymphomas in established international guidelines[40].

Whilst SMZL is a slow growing lymphoma, approximately 70% of patients develop a progressive disease requiring treatment, where approximately 30% of these will ultimately transform to a more aggressive lymphoma[38]. Molecular lesions, such as *IGHV* status, *NOTCH2* and *KLF2* mutation, *TP53* abnormalities and aberrant promoter methylation have been associated with poor outcomes in SMZL patients[47,71,204], but like with diagnosis, none have been used or included in clinical prognostic models and are not recommended in international guidelines[31,40].

8.2.4 Aims

The aim of this chapter was to explore and determine the clinical significance of somatically acquired genetic mutations within the Jaramillo-Parry cohort by integrating our sequencing data results with survival outcomes, copy number alterations and other molecular biomarkers. It also aimed to identify if there were any potential disease subgroups defined by the genomics of the disease.

8.3 Materials and Methods

8.3.1 Patients and samples

For the Jaramillo [n=146] and Parry [n=175] cohorts, targeted sequencing results from Chapter 7 were integrated with the clinical data available and unless otherwise stated, results shown are a combination of the two cohorts (termed 'Jaramillo-Parry' cohort). **Table 8-2** describes the cohort of combined SMZL patients (Jaramillo-Parry) as well as a breakdown per cohort. As mentioned in chapter 2, samples [n=321] were obtained from different international centres. All samples met established diagnostic criteria and for the Jaramillo cohort a further validation by expert haematologist David Oscier was made. DNA was extracted from peripheral blood, spleen, bone marrow, skin and lymph nodes. Detailed description of each cohort can be found in methods **section 3.1.**

Table 8-2. Patient characteristics.Description of the combined Jaramillo-Parry cohort as well as a breakdown of the Jaramillo and Parry cohort.

	Variable	Description	Jaramillo-Parry	Jaramillo	Parry	p value *
Cohort characteristics	Number of patients	SMZL diagnosis	321 (100%)	146 (100%)	175 (100%)	-
	Age at diagnosis	Median (range)	69 years (36-90)	67 years (36-88)	70 years (37-90)	0.162 ϕ
	Gender	Female	166 (52%)	76 (52%)	90 (51%)	1
		Male	153 (48%)	70 (48%)	83 (49%)	
Clinical	Follow up time	Median (range)	5.0 years (0-27)	4.85 years (0-23)	5.23 years (0-27)	0.241 ϕ
	Status	Dead	62 (22%)	17 (14%)	45 (27%)	0.013
		Alive	225 (78%)	102 (86%)	123 (73%)	
	Treatment	Treated (includes splenectomy)	213 (71%)	85 (67%)	128 (74%)	0.236
		Untreated	85 (29%)	41 (33%)	44 (26%)	
	Splenectomy	Yes	103 (40%)	47 (52%)	56 (34%)	0.007
		No	151 (60%)	43 (48%)	108 (66%)	
	Transformation to large cell lymphoma	Yes	31 (19%)	10 (16%)	21 (18%)	0.836
		No	143 (82%)	52 (84%)	91 (82%)	
	Time to first treatment	Median (range)	0.26 years (0-22)	0.43 years (0-14)	0.16 years (0-22)	0.025 ϕ
	Event free survival (EFS) status	Event (tranformation, 2nd treatment or death)	92 (45%)	28 (36%)	64 (51%)	0.042
		No event	111 (55%)	50 (64%)	61 (49%)	
IGHV genes	IGHV repertoire	IGHV1-2*04	38 (15%)	22 (18%)	16 (13%)	0.292
		not IGHV1-2*04	207 (85%)	99 (82%)	108 (87%)	
	IGHV mutation status (2 groups)	mutated (<98% GI)	139 (67%)	85 (76%)	54 (56%)	0.003
		unmutated (≥98% GI)	69 (33%)	27 (24%)	42 (44%)	
	IGHV mutation status (3 groups)	mutated (<97% GI)	117 (57%)	72 (65%)	45 (47%)	0.011 χ
		borderline (97-99% GI)	72 (35%)	34 (31%)	38 (40%)	
		unmutated (100% GI)	18 (8%)	5 (4%)	13 (13%)	
Chromosomal abnormalities	Genomic complexity (data generated from 450K or EPIC array)	Complex (3+)	56 (36%)	50 (36%)	6 (35%)	1.000
		Simple (< 3)	98 (64%)	87 (64%)	11 (65%)	
	Del 7q status ψ	del 7q	57 (24%)	38 (26%)	19 (21%)	0.354
		Normal	178 (76%)	106 (74%)	72 (79%)	
	TP53 status (del 17p or tp53 mutation) ψ	Aberrant	50 (16%)	19 (13%)	31 (18%)	0.281
		Normal	271 (84%)	127 (87%)	144 (82%)	

*** Fisher exact 2-sided test unless otherwise stated**
ϕ Mann Whitney U test
χ Chi square test
ψ Data generated rom either FISH, karyotype and or 450K & EPIC array data

Due to need for multi-centre studies predicated on the scarcity of samples, collecting a cohort of patients with a homogenous treatment was very challenging. Furthermore it is important to note that the Parry cohort was older than the Jaramillo cohort (**Figure 8-1**) and had longer follow up time (**Figure 8-2**).

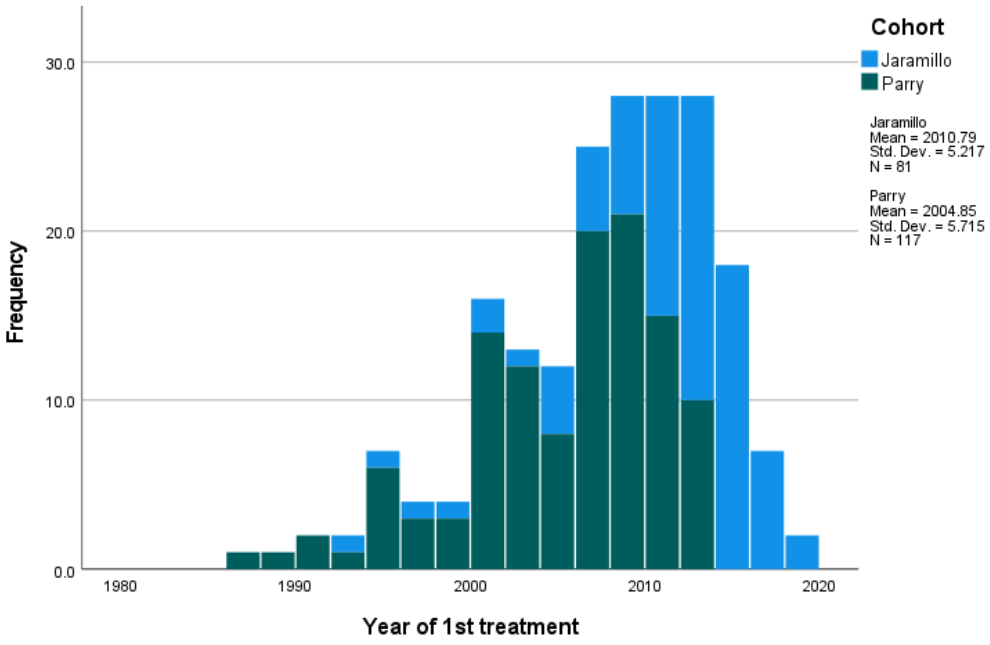


Figure 8-1. Histogram of year of 1st treatment across Jaramillo and Parry cohorts.

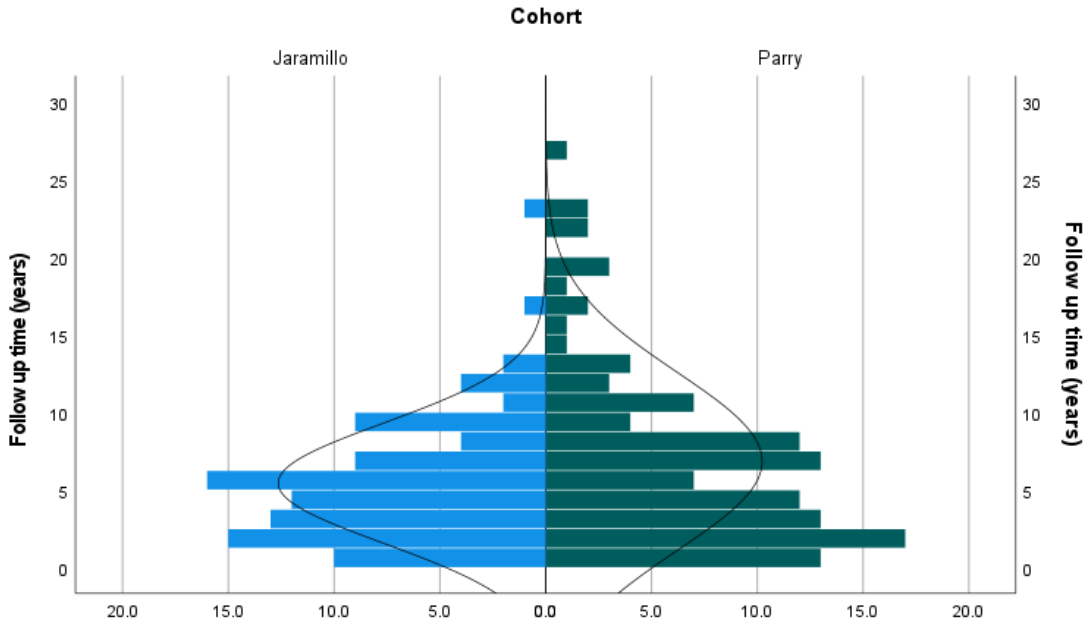


Figure 8-2. Histogram of follow up time across the Jaramillo and Parry cohort.

Most patients in the Jaramillo-Parry cohort, where data was available [n=293], were either untreated [n=84] or had a splenectomy [n=75]. This was followed by the use of rituximab [n=39], combination of chemotherapy and rituximab [n=35], chemotherapy [n=34], and other combinations [n=60]. **Figure 8-3** shows the percentage of patients that received each treatment per cohort. These figures were for first-line treatments.

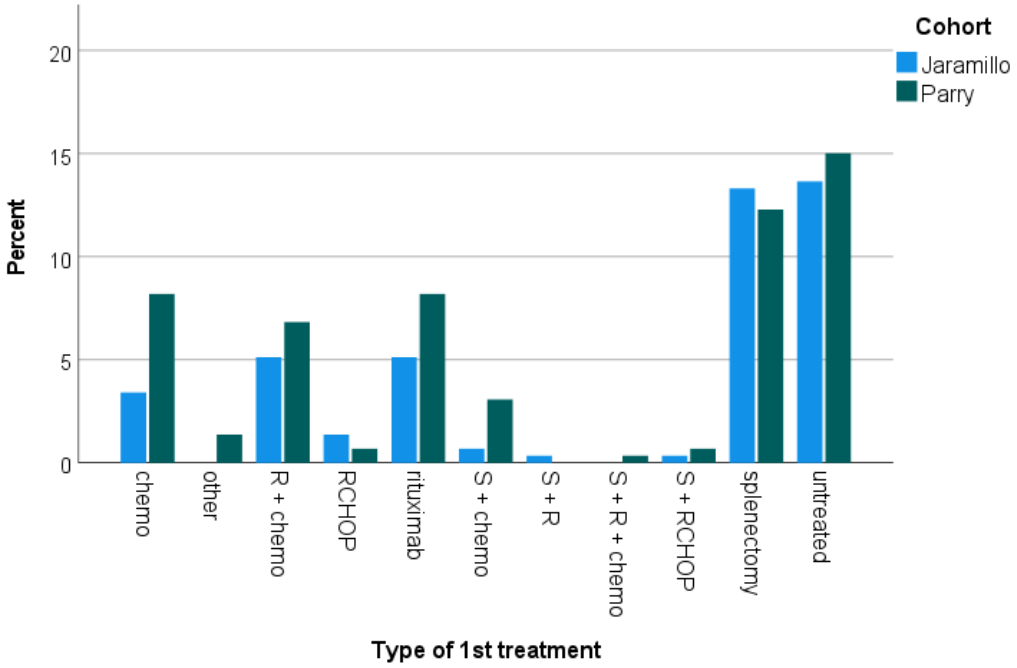


Figure 8-3. Types of first treatment compared across the Jaramillo and Parry cohorts.

First-line treatment across the two cohorts was heterogenous with a mixture of chemotherapy, splenectomy, and immunotherapy (rituximab). The main therapies for patients before the year 2000 were splenectomy and chemotherapy. However, after rituximab was introduced for the treatment of SMZL in the mid-2000s it became a first line treatment, either on its own or in combination with other therapies[40,208]. In addition to a therapeutic intervention, splenectomy is often used in the differential diagnosis of SMZL, which will impact on the natural history of the disease post splenectomy. This is because splenectomy may improve splenomegaly (enlarged spleen) related symptoms and improve cytopenias (reduction in the number of mature blood cells). Studies in small cohorts have shown that approximately 90% of patients respond well to splenectomy and some may not require further therapy[208]. In a more recent study of 227 SMZL patients, splenectomy up front was associated with a favourable outcome[209]. Therefore, splenectomy was classified as a treatment type regardless of the clinical indication. **Figure 8-4** shows the types of first treatments plotted against the year of first treatment across patients in the combined Jaramillo-Parry cohort. The figure shows a clear shift in the way patients were treated after 2005, where rituximab monotherapy and rituximab combined with chemotherapy became the main choice of first-line therapy for patients that required treatment.

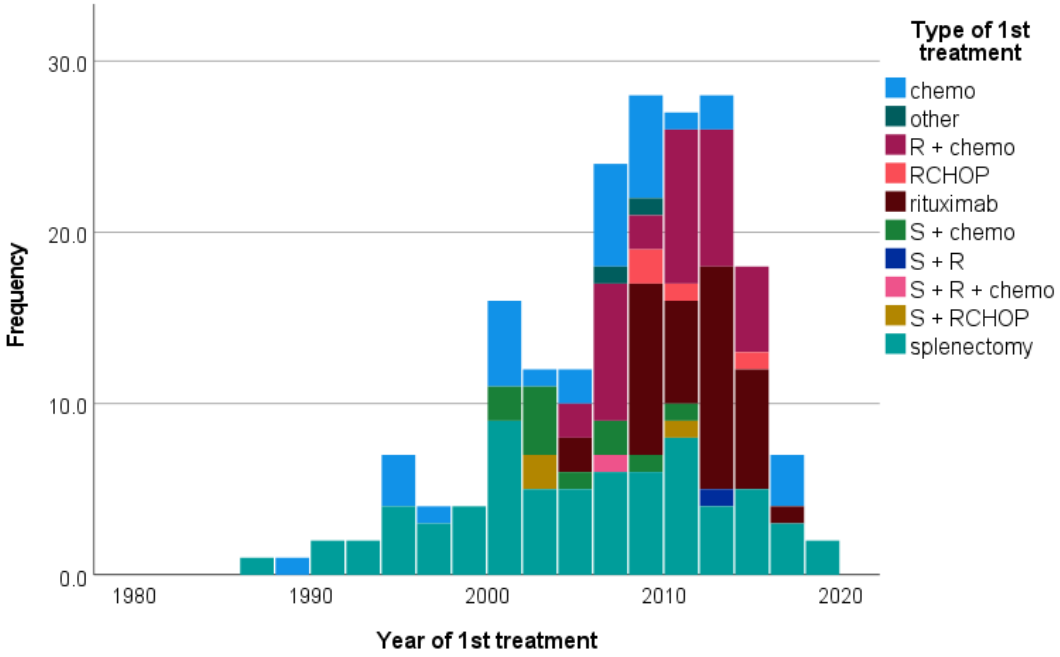


Figure 8-4. Histogram of type of 1st treatment plotted against year of 1st treatment across Jaramillo and Parry cohort. Chemo: chemotherapy, R: rituximab, S: splenectomy, RCHOP: rituximab, cyclophosphamide, doxorubicin, vincristine, prednisolone.

Pairwise Fisher's Exact test was used to ensure the two cohorts were comparable from a therapeutic perspective (results shown in **Table 8-3**). The tests showed no significant differences between the distribution of treatments across the two cohorts.

Table 8-3. Comparison of first-line treatments across Jaramillo and Parry cohort. A total of 293 patients had treatment data available.

Treatment	Jaramillo	Parry	p value*
chemo	10	24	0.0981
other	117	142	
R+chemo	15	20	1.0000
other	112	146	
RCHOP	4	2	0.4085
other	123	164	
rituximab	15	24	0.6035
other	112	142	
S+chemo	2	9	0.1217
other	125	157	
splenectomy	39	36	0.1047
other	88	130	
S + R/R+chemo/RCHOP	2	3	1.0000
other	125	163	
untreated	40	44	0.3640
treated	87	122	

*** Fisher exact 2-sided test**

8.3.2 Copy number aberrations (CNAs)

Samples [n=155] were processed using the Infinium Human Methylation 450 BeadChip [n=103] and Illumina Infinium Methylation EPIC BeadChip [n=52], according to manufacturer's instructions, at the Genomics and Proteomics Core Facility of the DKFZ (Heidelberg, Germany). Data from methylation arrays was used by Dr. Dean Bryant as input to conumee package[210] to obtain copy number status. Before input into conumee, raw data was processed through minfi[211] to create a Methylset object (object containing only methylated and unmethylated signals). In conumme the intensity values of methylated and unmethylated probes were combined and these were then normalised using a set of normal controls [n=10 samples for 450K, n=8 samples for EPIC]. The controls were a mix of copy neutral normal B-cells. Breakdown of B-cells used can be found in **Table 8-4**. Secondly, each sample was compared (linear regression) to the set of corresponding controls resulting in the log2 ratio of probe intensities for the query sample versus the controls. Thirdly, neighbouring probes were combined within predefined genomic bins and intensity values were corrected so that the copy-neutral state was zero. Subsequently, the genome was segmented into regions of the same copy-number state and results were plotted, per chromosome per sample. Text output was also generated containing the start and end of the segments identified and log ratios. These computationally assigned copy number calls were manually curated and identified by myself and Dr. Helen Parker. This process involved visual inspection of the copy number plots and determining if there was false segmentation, missing breakpoints and or incorrect copy number status. No specific thresholds for log ratios or size were set due to the variability of the data. Ideograms of the manually curated copy number calls from the conumee package were constructed in Rstudio using the package karyoploteR[212].

Table 8-4. Breakdown of B-cells used as controls for 450K and EPIC methylation arrays.

450K array controls [n=10 samples]	EPIC array controls [n=8 samples]
• Non classed switched memory B-cells • Marginal zone B-cells	• Classed switched memory B-cells • Non classed switched memory B-cells • Naïve B-cells • Mixed B-cells

8.3.3 Principal component analysis

105 samples were used for principal component analysis (PCA) to see if cases with certain genomic and other molecular features would cluster together in different groups. Since data for all samples across all features was required to run the PCA, this limited our analysis to those that had sufficient data across all features. Genes with mutations in more than 5% of samples (*MYD88, BIRC3, TRAF3, KLF2, NOTCH2, CCND3* and *TNFAIP3*) and CNAs found in more than 5% of samples

(trisomy 3, trisomy 12, trisomy 18, gain 3q, gain 8q, gain 12q, del 1p, del 6q, del 7q and del 8p) were included. Other features included in the PCA were telomere length, *IGHV* identity, *TP53* aberrations, and copy number count (genomic complexity). Samples without telomere length, percentage of *IGHV* identity and CNA data were excluded from the PCA.

PCA was run in R using the *prcomp* command on the scaled data. Each sample was drawn on a biplot where the x-axis represented the first principal component (PC1) and the y-axis the second principal component (PC2).

8.3.4 Telomere length (TL)

Telomere length (TL) relative to a standard reference sample (K562 cell line DNA) was determined in 140 Samples by Dr. Helen Parker using monochrome multiplex PCR (MMQPCR). Absolute TL in kb was extrapolated, using linear regression, from 82 CLL cases with single telomere length analysis (STELA)[213,214] data. The equation below was used to convert telomere length units (TLU) from MMQ-PCR to telomere size in kb:

$$ESTELA\ [kb] = 1.626 + (0.4952 * MMQPCR\ [TLU])$$

This equation was derived from the intercept and coefficient obtained by regressing STELA values on MMQ-PCR estimates. Description of the MMQPCR and STELA method as described by Strefford et al.[215] can be found below:

MMQPCR: DNA samples were analysed in triplicate in 96 well plates, alongside six serial dilutions of a reference DNA sample. PCR Amplification and measurement of SYBR green fluorescence was performed on LightCycler®480 real time PCR system (Roche). Raw fluorescence data for each sample and standard dilution was split into two sets of data points relating to the acquisition temperatures for the two targets (Alb and Tel) and the second derivative maximum cycle threshold (Ct) values were calculated for each of the two data sets using PCR miner. Linear regression of the Alb Ct and Tel Ct values for the standard DNA was used to generate separate standard curve equations for Alb and Tel in each run. Alb Ct and Tel Ct values for each of the test samples were applied to the respective standard curves to give Alb and Tel concentrations relative to the standard. Division of the telomere concentration by the Alb concentration gave the mean MMQPCR telomere length (MQTL) for each sample, relative to the standard, in telomere length units (TLU).

STELA: Multiplexed PCR products from specific XpYp telomeric sequences were resolved and detected using gel electrophoresis and Southern hybridization, respectively. The

molecular weight of all DNA fragments, including control fragments of known sizes, were calculated using the Phoretix 1D quantifier (Nonlinear Dynamics).

Recursive partitioning was used to identify cut-off values, based on percentiles, for telomere groups with the greatest prognostic power as previously described[215]. The hazards ratio for overall survival (OS) and progression free survival (PFS) was calculated between the cases with telomere length (TL) values below a set percentile or cut-off versus those with values equal to and above the cut-off (percentile). This was done for all percentiles with an increment of 5 until the 95th percentile. Results are shown in **Table 8-5**. OS is defined as time from diagnosis to death or date of last follow up for survivors. Progression free survival is defined as time from diagnosis to time to event, defined as second treatment, transformation, or death.

Table 8-5. Recursive partitioning based on telomere length (TL) to establish cut off values with maximum prognostic power. The hazards ratio for OS and PFS is calculated between the cases with TL values below the cut-off versus those with values above the cut-off using a Cox proportional hazards model. This is done for all percentiles with an increment of 5 until the 95th percentile.

		OS			PFS		
percentile	**cut-offs**	**Hazards ratio**	**95 % confidence interval**	**p value**	**Hazards ratio**	**95% confidence interval**	**p value**
10	2.588	4.679	1.283-17.068	0.019	4.271	1.579-11.554	0.004
15	2.702	5.702	2.017-16.117	0.001	2.905	1.267-6.661	0.012
20	2.784	4.285	1.595-11.511	0.004	2.49	1.123-5.524	0.025
25	2.849	2.968	1.050-8.391	0.04	2.257	1.015-5.018	0.046
30	2.930	3.272	1.168-9.169	0.024	2.076	0.948-4.543	0.068
35	2.991	2.405	0.848-6.825	0.099	1.883	0.872-4.069	0.107
40	3.069	2.2671	0.929-7.676	0.068	1.782	0.839-3.381	0.139
45	3.114	2.808	0.954-8.343	0.063	1.646	0.766-3.538	0.201
50	3.218	3.321	1.046-10.537	0.042	1.605	0.742-3.469	0.229
55	3.318	2.642	0.836-8.344	0.098	1.343	0.622-2.903	0.453
60	3.406	1.876	0.587-6.000	0.289	1.107	0.502-2.440	0.802
65	3.553	2.473	0.679-8.743	0.172	1.33	0.579-3.056	0.502
70	3.719	1.678	0.456-6.182	0.436	1.231	0.509-2.976	0.654
75	3.896	1.207	0.329-4.427	0.776	0.918	0.380-2.216	0.849
80	4.213	0.946	9.258-3.464	0.933	0.844	0.334-2.133	0.72
85	4.342	0.857	0.189-3.886	0.841	0.703	0.239-2.069	0.522
90	4.870	1.119	0.144-8.692	0.914	1.202	0.280-5.169	0.805
95	5.774	0.452	0.058-3.525	0.448	21.955	0.004-117987	0.481

The hazards ratio was plotted for both OS and PFS at each percentile. Using the values obtained a categorical variable was generated with three groups: short TL (<50 percentile), intermediate TL (50-75 percentile) and long TL (>75 percentile). The 50th percentile was chosen as the cut-off for short TL since more than half of the values below this threshold were significant according to the Cox proportional hazards model. For the long telomeres the 75th percentile was chosen as this was the cut-off when the HZ values dropped below 1 (**Figure 8-5**). Telomere length ranged from

2.34–3.20 kb (median: 2.86 kb) in the short group, 3.23–3.89 kb (median: 3.49 kb) in the intermediate group, and 3.89–7.57 kb (median: 4.68 kb) in the long group.

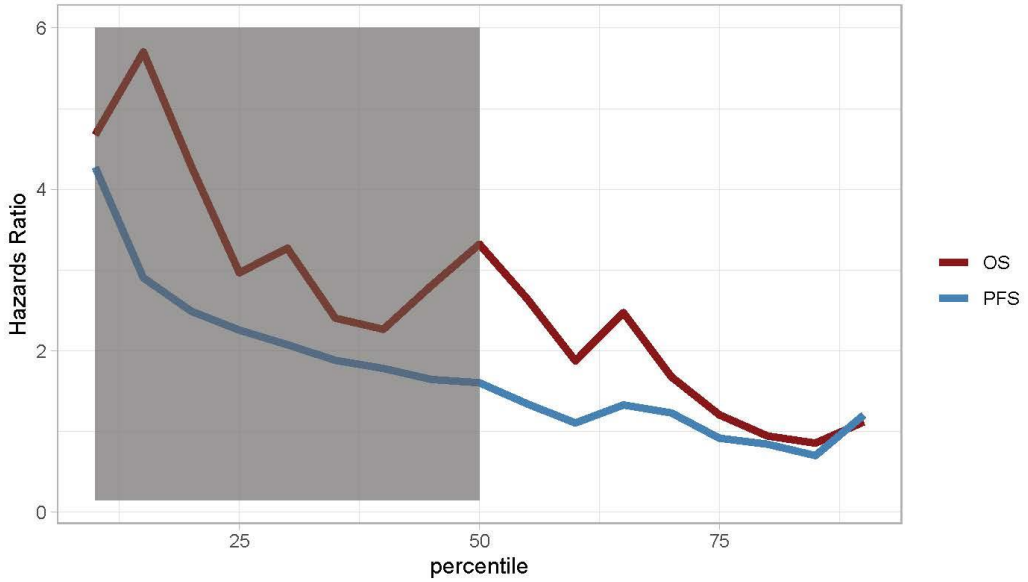


Figure 8-5. Hazards ratio for overall survival (OS) and progression free survival (PFS) between cases with telomere length (TL) values below a set cut-off (percentile) versus those with values equal to and above the cut-off (percentile).

8.3.5 Statistical analysis

Statistical tests were performed using SPSS (v27) and Rstudio. Mann Whitney U-test (two variables) and Kruskal Wallis test (more than two variables) was used to test for differences between independent variables including telomere length across groups and *IGHV* identity to germline in transformed vs non transformed patients. Chi square test was used to test for association between categorical variables (i.e. transformation status across groups). Fisher's exact test was used to look for significant ($p < 0.05$) associations between genetic and immunogenetic features. Bonferroni used for multiple-testing correction. Univariate survival analysis was performed by Kaplan-Meier and Cox regression analysis. Multivariate analysis, accounting for confounding factors, was performed using Cox proportional hazard analysis. The Cox proportional hazard analysis was performed using a backwards stepwise approach, which runs the model a number of times and each time or step the weakest correlated variable is removed. Overall survival (OS) was defined as time from diagnosis to death or data of last follow up (date) for survivors. Time to first treatment (TTFT) was defined as time from diagnosis to time to first treatment including splenectomy.

8.4 Results

8.4.1 Recurrent copy number alterations (CNAs)

The methylation arrays (450K and EPIC arrays) allowed for the identification of copy number alterations within our cohort. This section will detail the recurrent copy number alterations (CNAs) found and in subsequent sections how these CNAs interact with other genetic abnormalities and the clinico-and biological features of these patients.

Among the 154 SMZL cases with methylation array data, 129 cases had copy number alterations. The most frequent copy number alterations were 7q deletions [24.6%, n=37] followed by trisomy 12 [24.0%, n=36] and gain of 3q [15.3%, n=23]. **Figure 8-6** shows all the CNAs identified across chromosomes 1-22 and the percentage of affected samples in the most recurrent CNAs. Along with trisomy 12, trisomy 3 [8.6%, n=13] and 18 [8.0%, n=12] were the most recurrent trisomies. The most common gains were found in 3q [15%, n=23], 12q [8.0%, n=12] and 8q [6.0%, n=9]. While the most common deletions were 7q [25%, n=37], 8p [8.6%, n=13], 13q [9.3%, n=14], 6q [8.6%, n=13], 1q [7.3%, n=11] and 17p [8.0%, n=12]. 24.7% of cases [n=38] showed a single CNA, 22.7% of cases [n=35] carried two alterations and 36.4% of cases [n=56] had three or more CNAs and were considered to have complex genomes.

Figure 8-6. Summary of copy number alterations (CNAs) from 450K and EPIC array data. Each chromosome is numbered with its corresponding ideogram from the p-arm (left) to the q-arm (right). Genes shown were those targeted by the HaloPlex gene panels. Trisomies are coloured light green, gains in dark green and losses in red. Bar graph on the bottom right side shows the percentage of affected samples among the most common CNAs.

Deletions in the long arm of chromosome 7 are common in SMZL and this was reflected within our cohort. The minimally deleted region was difficult to determine as there were two patients (L060_09 and 92568) with deletions that did not overlap but were very close together and depending on which one was included the MDR would shift slightly. Since the CNAs were obtained using methylation arrays rather than gold standard high-density SNP arrays or WGS the real breakpoints might not be optimally resolved. Profiles of chromosome 7 for patient L060_09 and 92568 are shown in **Figure 8-7**.

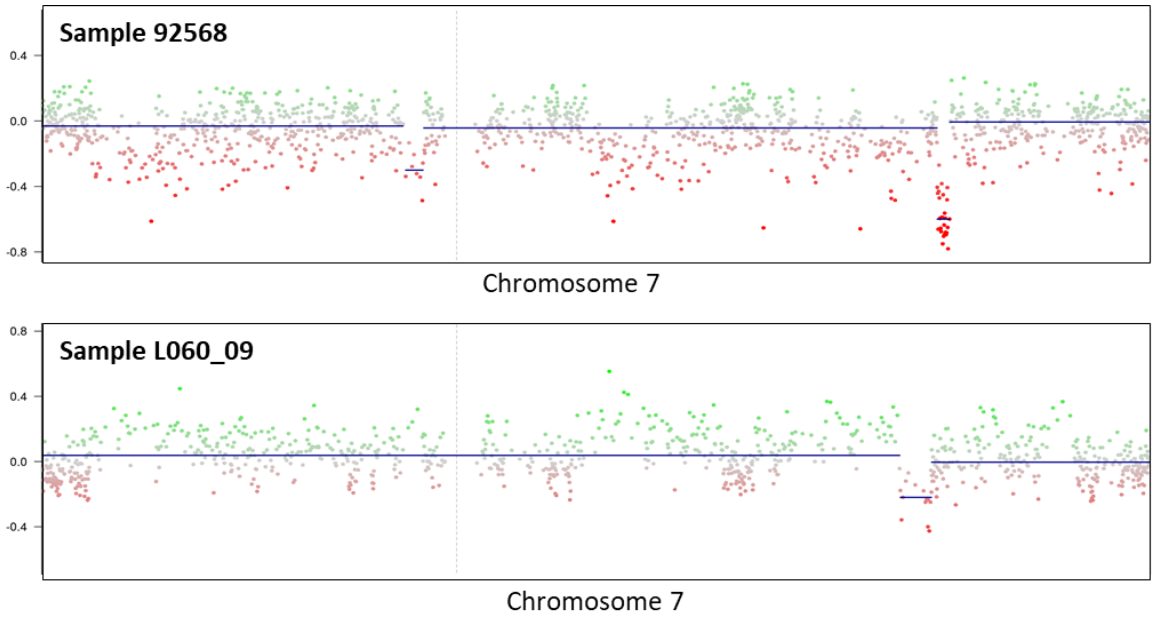


Figure 8-7. Deletion profiles of chromosome 7 for sample 92568 (top) and L060_09 (bottom) obtained from methylation arrays. Red dots deviating to the bottom, from a normal copy number (0) indicate a deletion.

To validate the deletions identified by the methylation arrays in patients L060_09 and 92568, the mean target coverage for all genes in sample L060_09 and 92568 was ascertained, normalised, and compared to a sample with no 7q deletions (L076). The results are shown in **Figure 8-8**, where the first boxplot shows the mean coverage of targeted genes in patient L060_09. *POT1*, a gene that fell within the 7q deleted region in sample L060 had much lower coverage in comparison to other genes. *MYD88, KMT2D,* and *KDM2B* were all genes that fell within gained regions (chromosome 3 and 12) and had higher coverage in relation to other genes. The boxplot for patient L060_09 also showed that *DDX3X* had relatively lower coverage, which was expected as this patient is male and gene *DDX3X* falls within chromosome X.

Similarly, for patient 92568, *FLNC* was the gene that fell within the 7q MDR and it also showed lower coverage compared to other genes. In patient 92568, *TNFAIP3* also fell within a deleted region showing good agreement between coverage and deletion status. In the patient with no 7q deletion the mean target coverage for both *POT1* and *FLNC* was very close to the mean coverage

across all genes. Therefore, with methylation and coverage data it was decided that both deletions in patient L060_09 and 92568 had enough evidence supporting them and represented real losses.

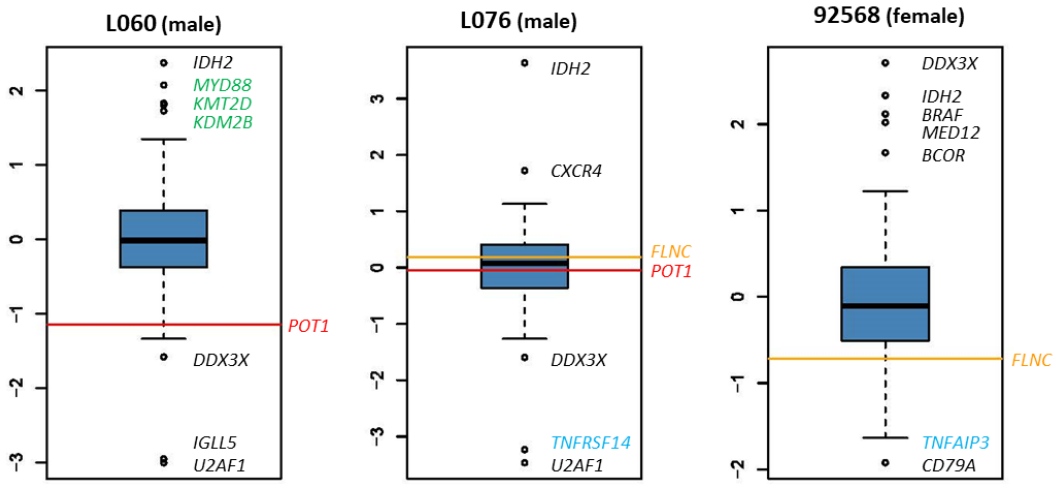


Figure 8-8. Mean target coverage (normalised) across all targeted genes in three patients (L060, L076 and 92568). Red lines show the normalised mean target coverage for *POT1* and orange lines the normalised mean target coverage for *FLNC*. *POT1* and *FLNC* were genes that fell within the deleted regions identified in patient L060_09 and 92568 respectively. Patient L076 is shown for comparison and does not have a 7q deletion. Genes in blue fell within deleted regions and genes in green within gained regions. *DDX3X* is located on chromosome X and is therefore expected to have much lower coverage in males than in females. *IDH2* was a gene with generally high coverage across all patients (mean depth: 544x), and similarly *U2AF1* was a gene that showed low coverage across all patients (mean depth: 8x).

This meant that two MDRs were identified in chromosome 7 rather than one. The MDR that identified in patient L060_09 was around 2,900 kb (chr7:124,200,000-127,100,000) which affected other 32 patients and deleted three genes (*GPR37, POT1, GRM8*). The MDR identified in patient 92568 was approximately 1,752 kb (chr7:128,575,000-130327262), also affecting 32 other patients and included genes *CALU, OPN1SW, CCDC136, FLNC, ATP6V1F, ATP6V1FNB, IRF5, TSPAN33, SMO, STRIP2, SMKR1, NRF1, UBE2H, ZC3HC1, KLHDC10, TMEM209, SSMEM1, CPA2,* and *CPA4*. **Figure 8-9** shows all the CNAs identified in chromosome 7 and zooms into the MDR regions.

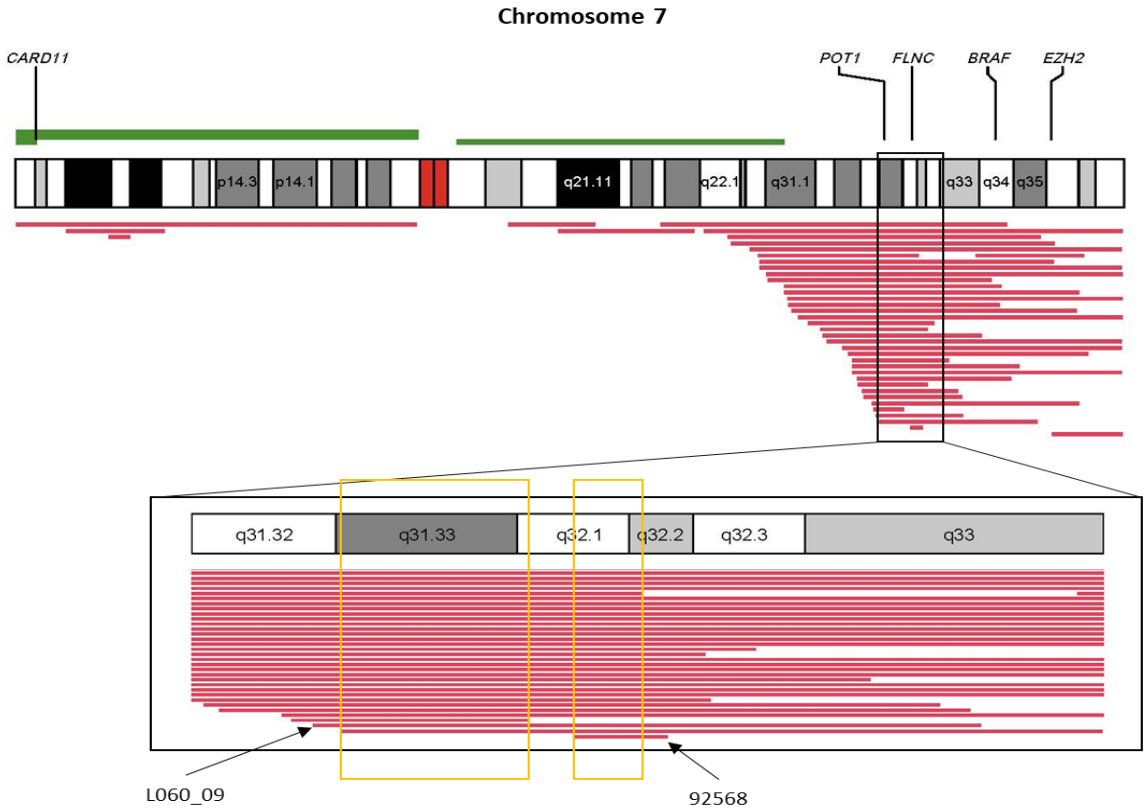


Figure 8-9. CNVs across chromosome 7.Black rectangle zooms in around bands 7q31.32-7q33. The orange box highlights the minimally deleted regions (MDRs). The genes shown were those targeted by the HaloPlex gene panel.

Comparison of the two newly identified MDRs to other MDRs identified in previous studies is shown in **Figure 8-10**. The MDR which included *POT1* overlapped with an MDR identified by two studies in the early 2000s (purple)[216,217]. The second MDR which included *FLNC* overlapped with all previously identified regions including the most recently published in 2012 by Watkins and colleagues[47,218]. Both *POT1* and *FLNC* were targeted by the gene panel, however, no cases with deletions in these genes had co-occurring mutations.

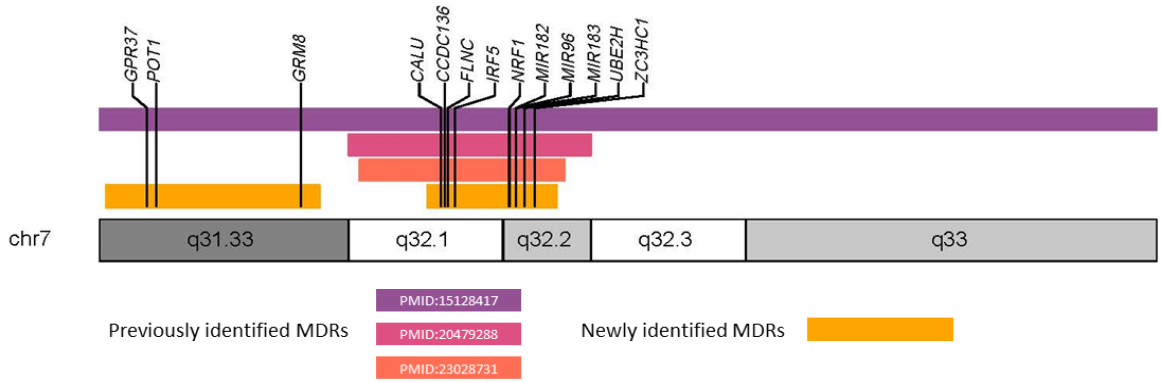
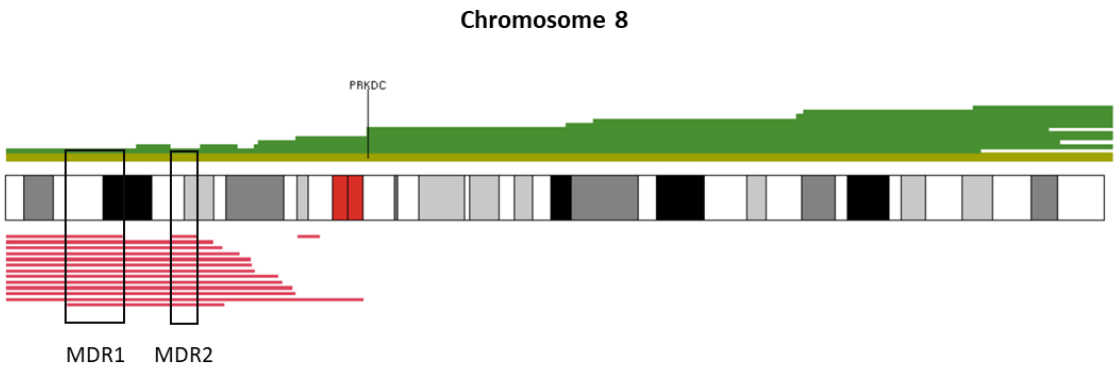


Figure 8-10. Minimally deleted regions (MDRs) identified in chromosome 7. Two MDRs (yellow/orange bars) were identified using copy number calls from 450K and EPIC arrays. Previously identified MDRs are also shown in purple[216,217], pink[47] and salmon[218]. Genes shown are those affected by the MDRs within the Jaramillo-Parry cohort.

In the short arm of chromosome 8 there were 13 patients that shared two minimally deleted regions of approximately 7,300 kb (chr8:8,175,001-15,475,000) and 3,635 kb (chr8:21,825,001-25,450,000). The first MDR includes 30 genes while the second MDR 40 genes (**Figure 8-11**). Putative targets of this deletion include the tumour necrosis factor receptor superfamily (TNFRSF) genes (*TNFRSF10A*, *TNFRSF10B, TNFRSF10C, TNFRSF10D*), *DOK2* and *BLK*.

A.



B.

Chromosome 8 -MDR1			Chromosome 8 - MDR2			
PRAG1	*SOX7*	*DEFB130A*	*DOK2*	*PHYHIP*	*RHOBTB2*	*NKX3-1*
CLDN23	*PINX1*	*CTSB*	*XPO7*	*SLC39A14*	*TNFRSF10B*	*NKX2-6*
MFHAS1	*XKR6*	*FDFT1*	*NPM2*	*HR*	*TNFRSF10C*	*STC1*
ERI1	*MTMR9*	*NEIL2*	*FGF17*	*POLR3D*	*TNFRSF10D*	*ADAM28*
PPP1R3B	*SLC35G5*	*USP17L7*	*FAM160B2*	*PIWIL2*	*R3HCC1*	*ADAM7*
TNKS	*FAM167A*	*LONRF1*	*NUDT18*	*PPP3CC*	*TNFRSF10A*	*NEFM*
MSRA	*BLK*	*TRMT9B*	*REEP4*	*C8orf58*	*CHMP7*	*NEFL*
PRSS55	*GATA4*	*DLC1*	*LGI3*	*BIN3*	*LOXL2*	*DOCK5*
RP1L1	*DEFB136*	*C8orf48*	*SFTPC*	*EGR3*	*ENTPD4*	*GNRH1*
C8orf74	*DEFB130B*	*SGCZ*	*BMPI*	*PEBP4*	*SLC25A37*	*KCTD9*

Figure 8-11. CNVs across chromosome 8 and putative target genes **A.** Black rectangles show the minimally deleted region (MDR) shared by 13 patients. *PRKDC* was the only gene in chromosome 8 targeted by the HaloPlex gene panels **B.** List of genes within the MDRs.

Chromosome 6 has a number of deletions in its long arm, however identification of the MDR was difficult since the breakpoints had several gaps (there is no way to tell if the gaps were due to the quality of the data, the distribution of the methylation probes or real deletions). Nevertheless, there were 12 patients which had deletions targeting *TNFAIP3* (**Figure 8-12**), three of which also harboured mutations in that gene.

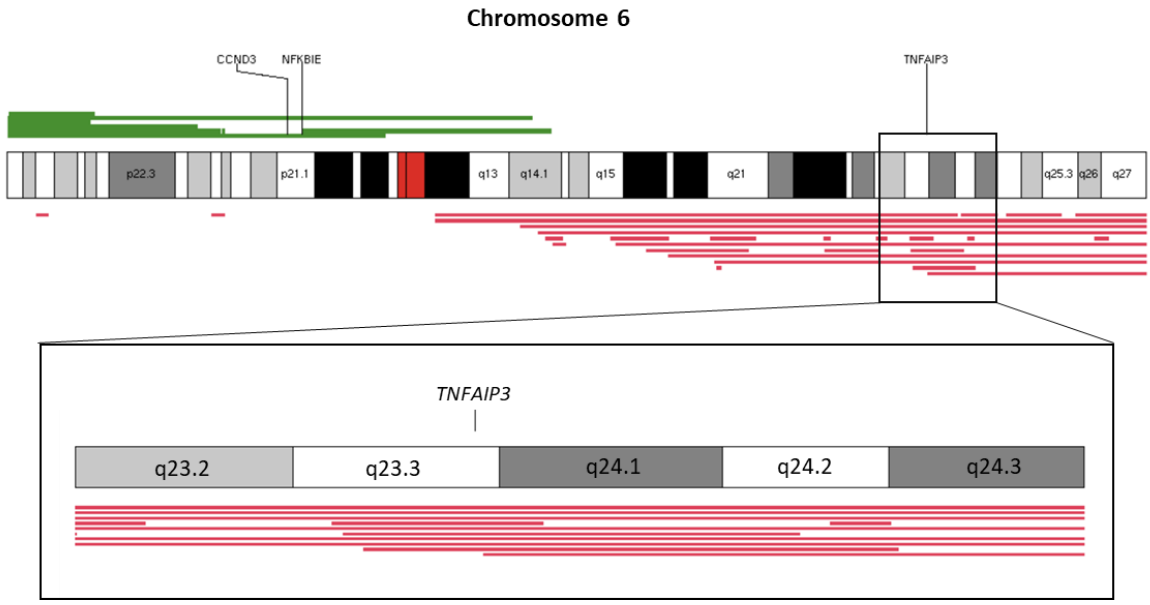


Figure 8-12. CNVs across chromosome 6. Black rectangle details bands 6q23.2 – 6q24.3 and the deletions identified in 12 patients targeting gene *TNFAIP3*.

Chromosome 17 was also a recurrent target of deletions. A minimally deleted region of approximately 850 kb (chr17:7225000-8075000) was identified in 12 patients (**Figure 8-13**). The target of this region is gene tumour suppressor gene *TP53* often deleted or mutated across different types of cancer.

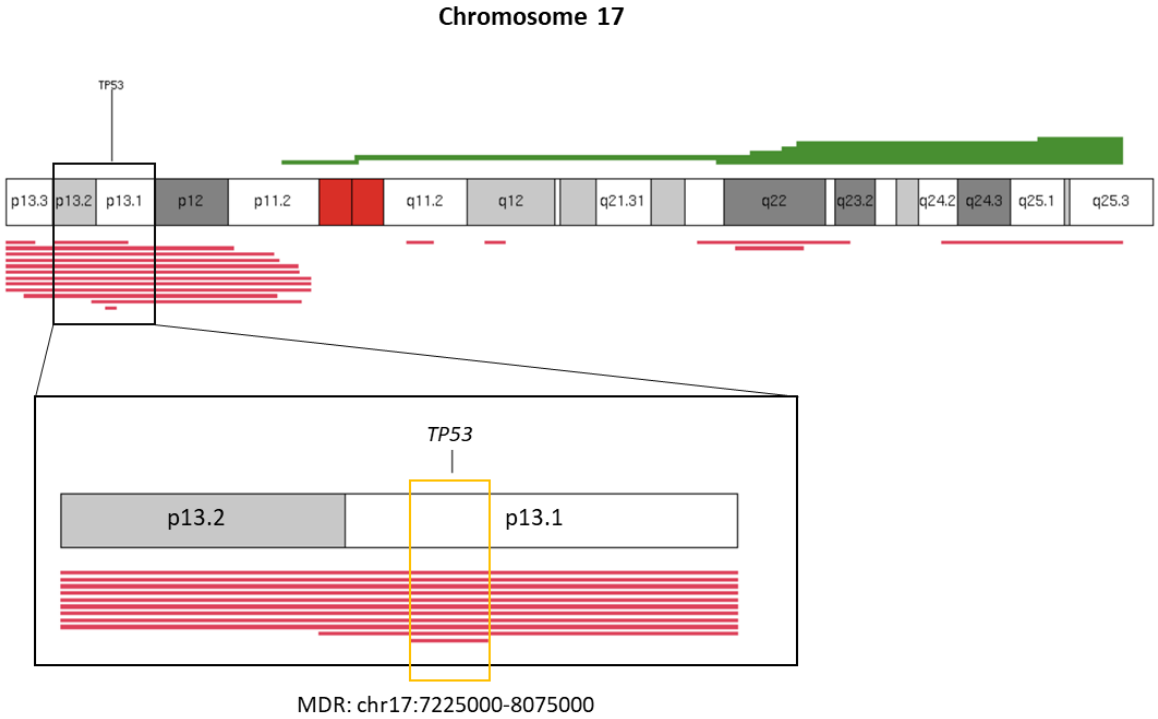


Figure 8-13. CNVs across chromosome 17. Black rectangle details bands 17p.13.2 – 17p13.1. The minimally deleted region (MDR) shared by 12 patients is shown by the orange rectangle (chr17:7225000-8075000) targeting tumour suppressor gene *TP53*.

In the long arm of chromosome 13 there was a small minimally deleted region of approximately 750 kb (chr13:50,675,001-51,425,000) identified in 13 patients (**Figure 8-14**). Although small, the region includes genes *DLEU1*, *DLEU7, RNASEH2B, FAM124A*, *SERPINE3* and *INTS6*.

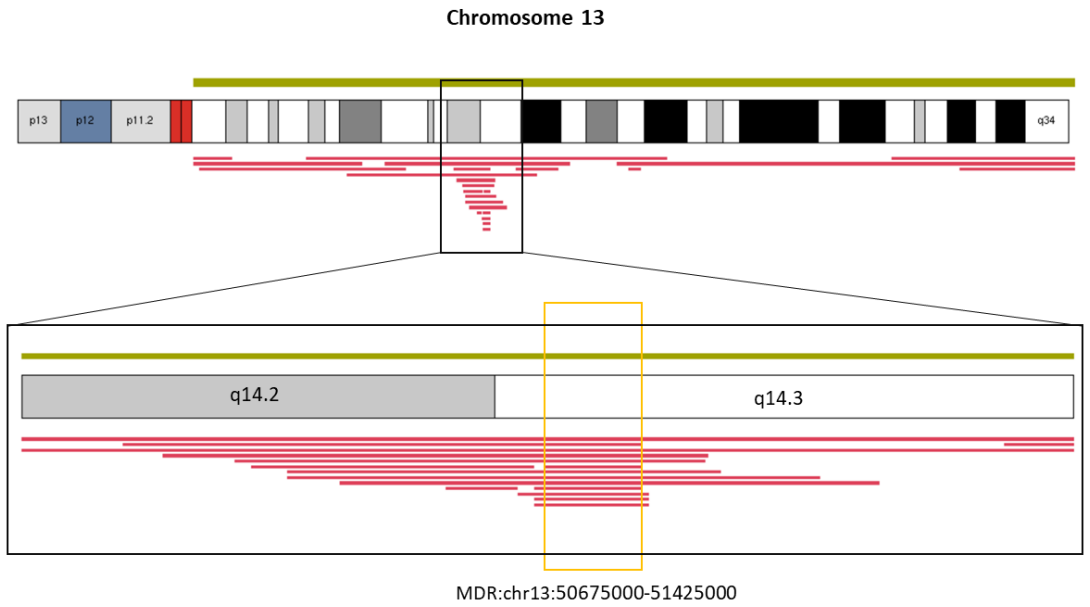


Figure 8-14. CNVs across chromosome 13. Black rectangle details bands 13q14.2 - 13q14.3. The minimally deleted region (MDR) shared by 13 patients is shown by orange rectangle.

Figure 8-15 shows the MDR in the short arm of chromosome 1 shared by eight patients. In this case, the deleted region was approximately 500 kb (chr1:26,625,001-27,125,000) which included genes *ARID1A, PIGV, ZDHHC18, SFN, GPN2, GPATCH3, NR0B2, NUDC, KDF1, TNRP1, TENT5B*, and *SLC9A1. ARID1A* was targeted by the gene panel, but no patients showed concurrent deletion and mutation.

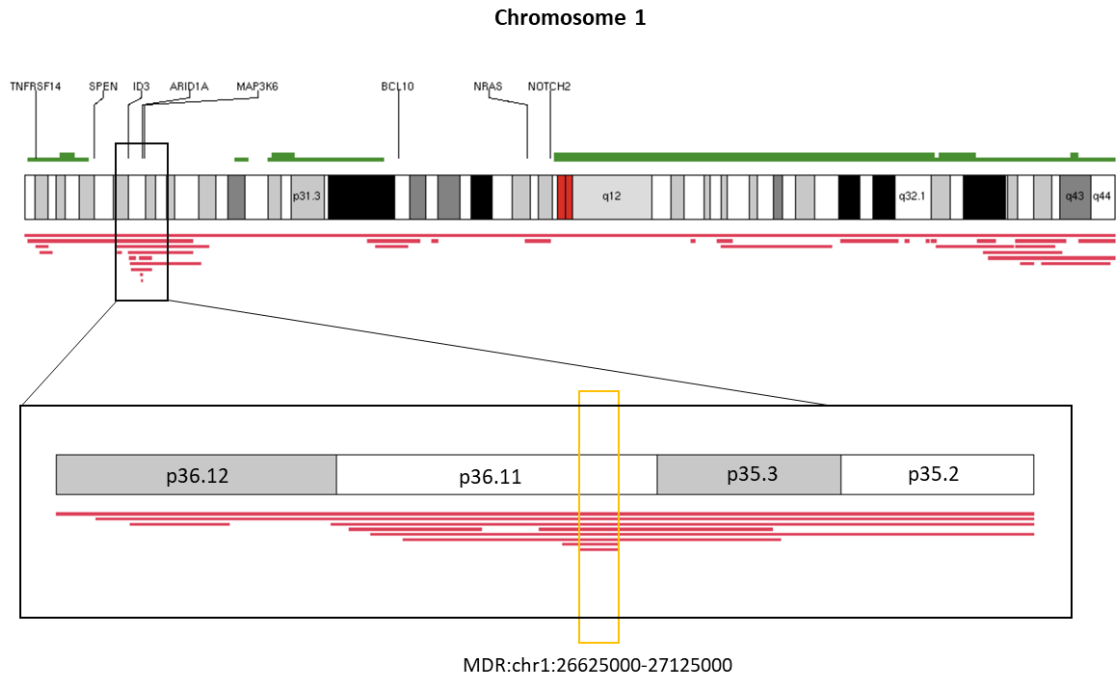


Figure 8-15. CNVs across chromosome 1. Black rectangle details bands 1p36.12 – 1p35.2. The minimally deleted region (MDR) shared by 8 patients is shown by the orange rectangle.

8.4.2 *IGHV* repertoire and somatic hypermutation status

207 patients had data on the mutational status of immunoglobulin heavy chain variable region (*IGHV*) genes. Cases ranged in homology from 47.90% to 100% and the mean and median were 95.60% and 96.18% respectively (**Figure 8-16**).

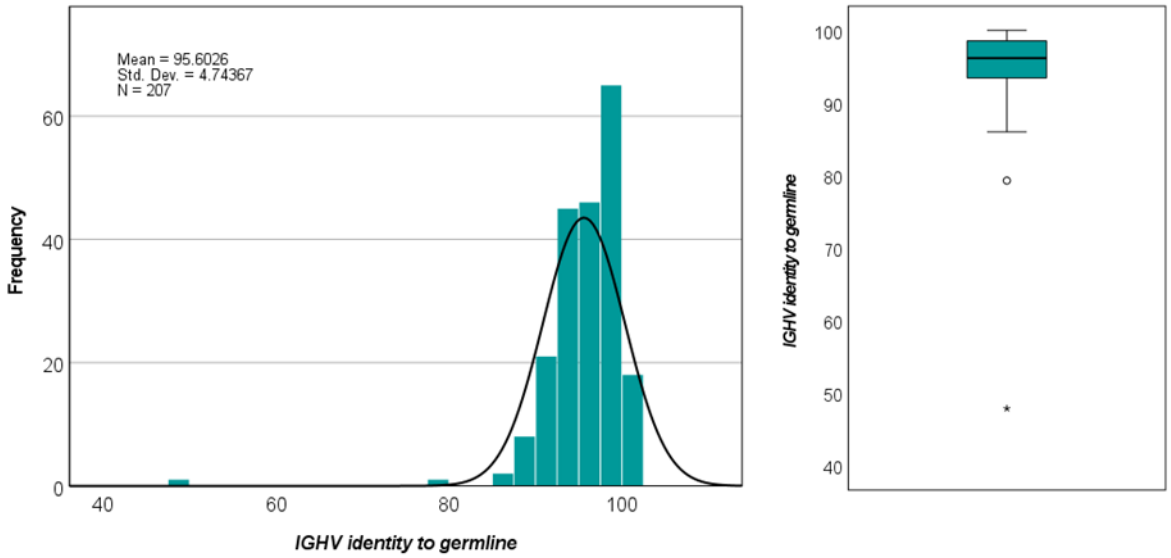


Figure 8-16. Distribution of percentage of *IGHV* identity to germline. The mean, median and range was 95.60%, 96.18%, 47.90% - 100% respectively and the 25th, 50th and 75th percentile were 93.40%, 96.18%, and 98.60% respectively.

In CLL, *IGHV* status is an important prognostic factor and a 98% cut off is used to dichotomise patients. Those patients with ≥98% *IGHV* identity to the germline sequence are considered unmutated, while sequences with <98% *IGHV* identity are considered mutated. Based on the CLL cut-offs, within the Jaramillo-Parry cohort 33% of sequences [n=69] were classified as unmutated (≥98%) and 66% of sequences [n=139] as mutated (< 98% GI). However, within the unmutated group there were sequences that were 100% identical to the germline. To be able to study this sub-group of truly unmutated samples separately, samples were also classified according to the thresholds used by Bikos et al.[205], whereby truly unmutated *IGHV* sequences were those with 100% identity to the germline [9%, n=18], sequences with 97-99.9% gene identity were classified as borderline/minimally mutated [35%, n=72] and sequences with gene identity <97% were classified as significantly mutated [56%, n=117]. **Figure 8-18** shows the number of samples in each cohort (Jaramillo and Parry) that fell within each of the categories in both the two and three group classification.

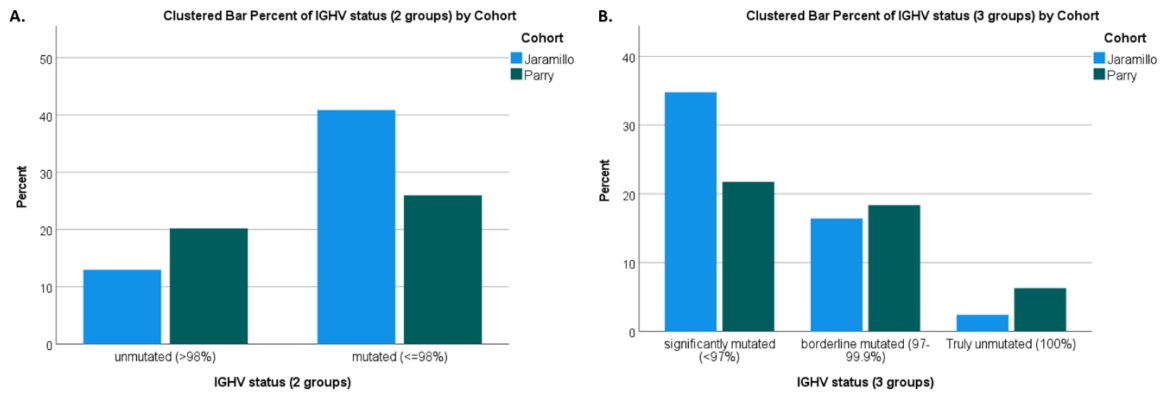


Figure 8-17. Somatic hypermutation within *IGHV* genes. **A.** Number of samples according to CLL classification of mutated (<98% identity) and unmutated (≥98% identity) *IGHV* genes **B.** Number of samples using Bikos et al. nomenclature. Truly unmutated samples had 100% identity to the germline, borderline/minimally mutated samples had sequences with 97-99.9% gene identity and significantly mutated sample had <97% gene identity to the germline.

The *IGHV* gene repertoire varied across patients, but the most frequent genes identified were by far *IGHV1-2*04* [15%, n=37], followed by *IGHV 3-23*01* [6%, n=13], *IGHV 3-23* [6%, n=12], *IGHV 3-30* [6%, n=12], *IGHV 4-34* [6%, n=12], and *IGHV 4-34*01* [6%, n=12]. **Figure 8-18** shows the most frequent genes identified in the Jaramillo-Parry cohort. The list of less prominent genes (<5% frequency) can be found in **Supplementary Table 8**.

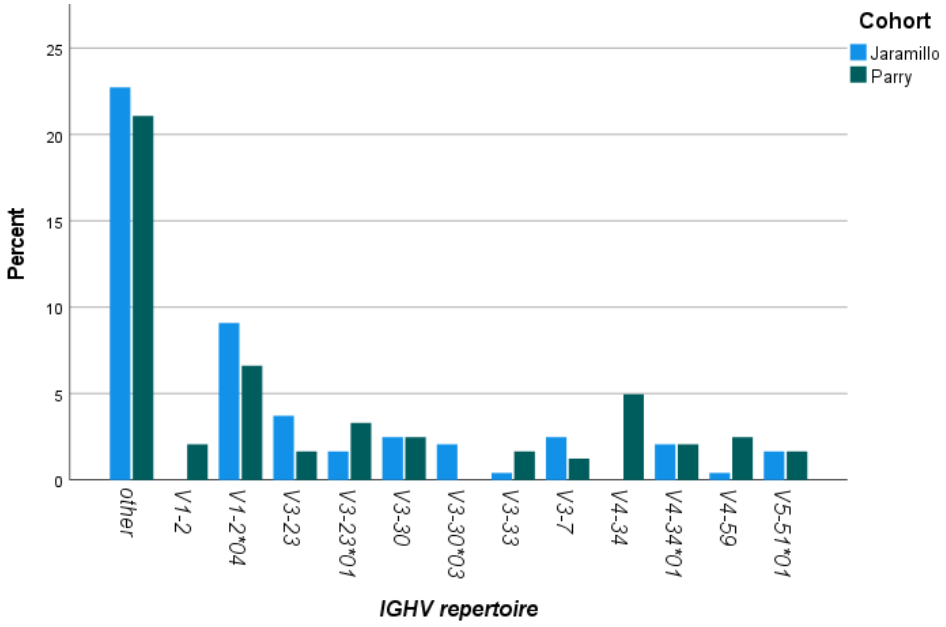


Figure 8-18. Most frequent *IGHV* genes (> 5% frequency) present in Jaramillo and Parry cohorts. The most frequent genes identified were *IGHV1-2*04* [15%, n=37], followed by *IGHV 3-23*01* [6%, n=13], *IGHV 3-23* [6%, n=12], *IGHV 3-30* [6%, n=12], *IGHV 4-34* [6%, n=12], and *IGHV 4-34*01* [6%, n=12].

The *IGHV1-2*04* gene represented 13% of all borderline mutated cases, but within the subgroup of *IGHV1-2*04* rearrangements, 73% of sequences [n=37] were borderline/minimally mutated (97-99.9% GI). In the subgroup with *IGHV3-23* rearrangements, 97% of sequences [n=12] were significantly mutated (<97% GI). Even within the *IGHV3-23*01* most sequences [69%, n=9,] were

also significantly mutated. Most truly unmutated cases (100% GI) where observed in the *IGHV4-34*01* subgroup [30%, n=4] (**Figure 8-19**).

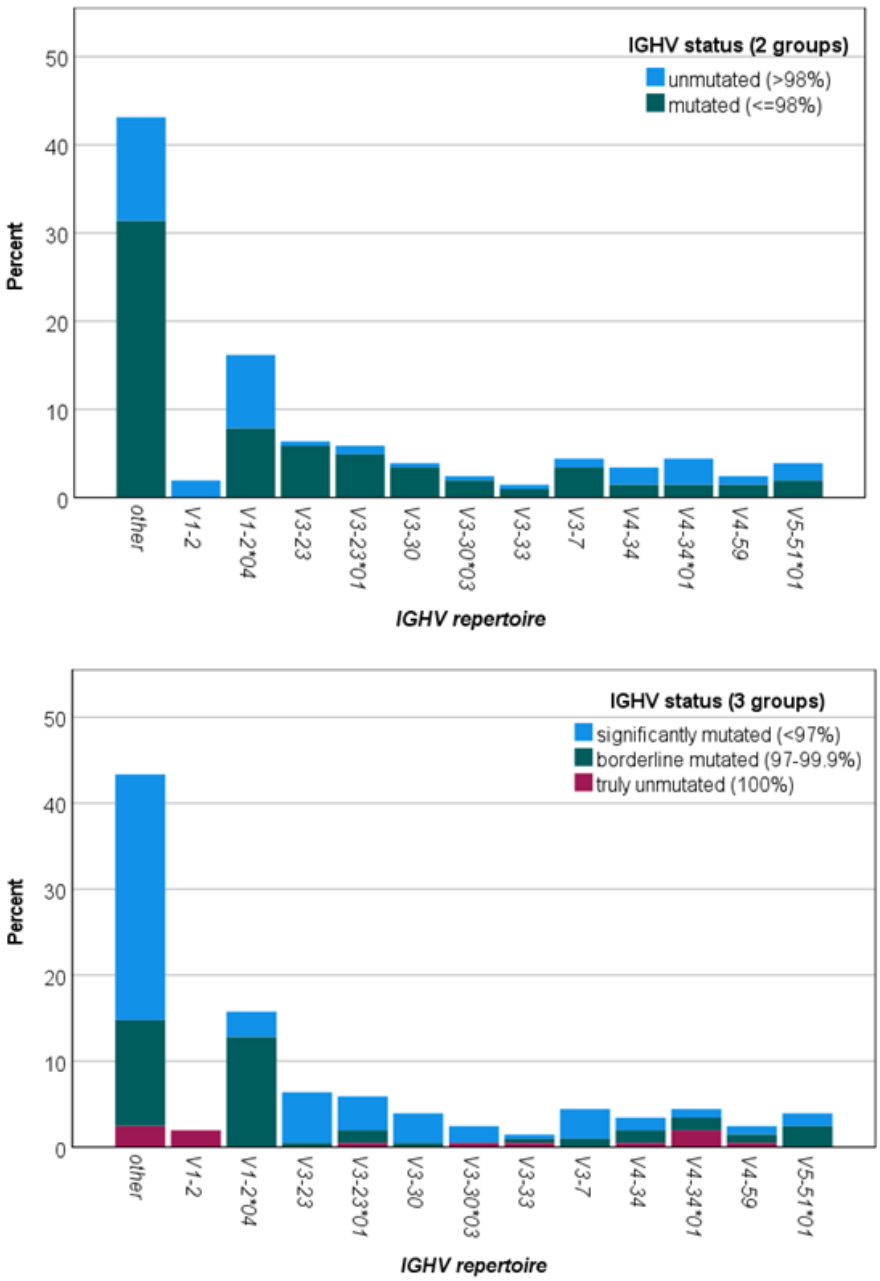


Figure 8-19. Breakdown of somatic hypermutation across *IGHV* genes within Jaramillo-Parry cohort. Top figure shows breakdown using conventional cut-offs while bottom figure shows cut-offs proposed by Bikos *et al*.

In the sections to follow the binary CLL cut-offs of mutated (<98% GI) and unmutated (≥98% GI) will be used to describe *IGHV* mutation status of the Jaramillo and Parry cohort unless otherwise stated.

8.4.3 Telomere length associates with key genomic features

A proportion of samples, mostly from the Jaramillo cohort, had telomere length data, determined by MMQPCR (**Figure 8-20**). This data was integrated with the genomic results to see if any

correlations could be made. Telomere length ranged from 2.384 kb to 7.568 kb. The mean and median telomere length were 3.52 kb and 3.21 kb respectively.

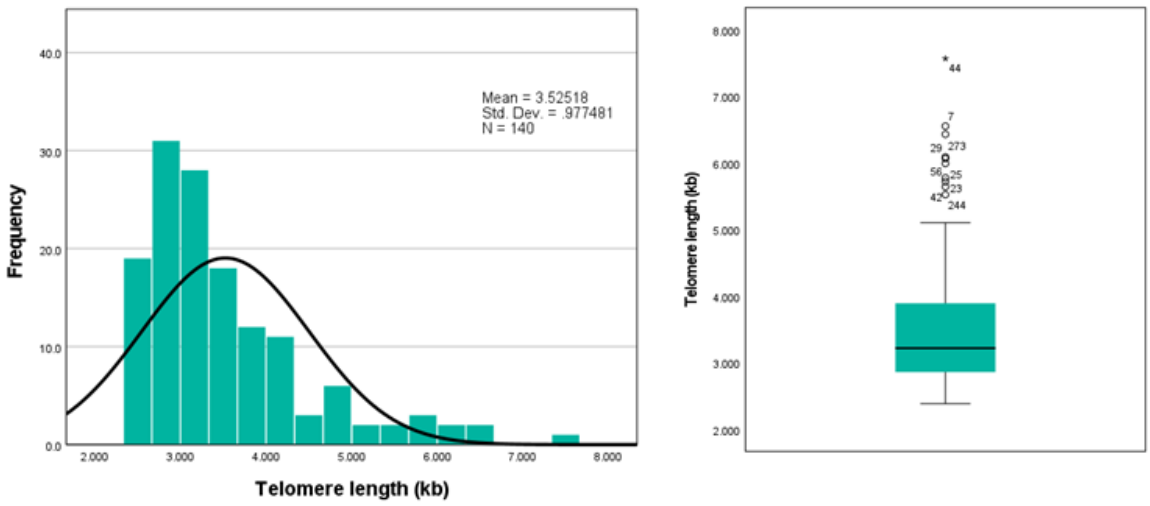


Figure 8-20. Distribution of telomere length. The mean, median and range of telomeric range was 3.52, 3.21, 2.384 - 7.568 respectively and the 25th, 50th and 75th percentile were 2.84, 3.21, and 3.89 respectively.

Patients with unmutated *IGHV* genes (≥ 98% GI) had significantly shorter telomeres that those with mutated *IGHV* genes (< 98% GI) (p=0.018). **Figure 8-21** shows the distribution of telomere length across the two subgroups. Interestingly, when comparing telomere length across the three subgroups proposed by Bikos et al. (significantly mutated, borderline mutated, and truly unmutated) telomere length was lower in the borderline mutated group (97%-99.9% GI) and not the truly unmutated (100% GI).

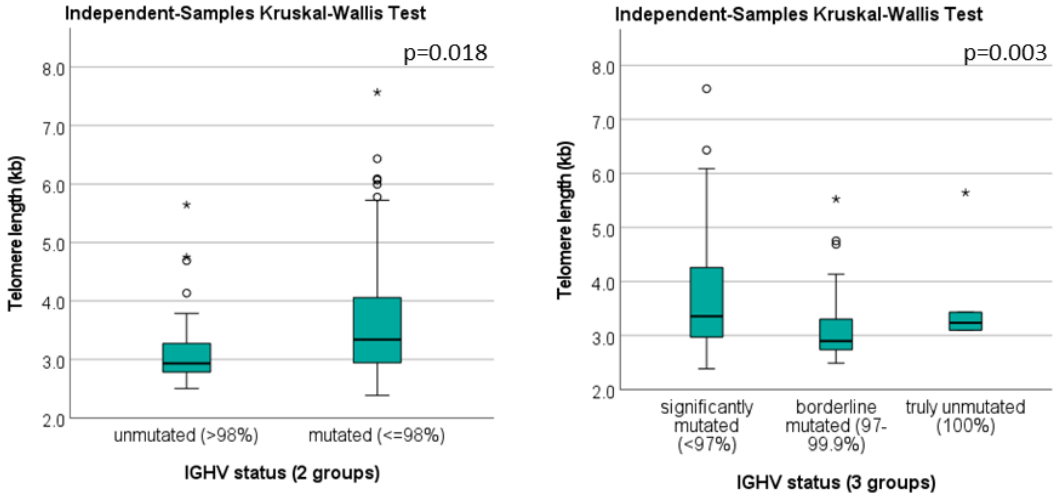


Figure 8-21. Distribution of telomere length (TL) across subgroups according to *IGHV* status. **Left:** TL distribution using conventional CLL cut-offs: Unmutated: those with 98% or more sequence identity to germline; mutated: those with less than 98% sequence identity to germline. **Right:** TL distribution across subgroups defined by Bikos et al. significantly mutated: <97% sequence identity to germline; borderline mutated: 97%-99.9% sequence identity to germline; truly unmutated: 100% sequence identity to germline.

The Kruskal-Wallis test showed a significant difference in the distribution of telomere length across *IGHV* status using both 2 and 3 groups. However, for the three subgroups it suggested a difference between at least one pair of groups but did not state which pair. Consequently, a pairwise comparison between truly unmutated, borderline mutated and significantly mutated was performed using the Dunn-Bonferroni approach (**Table 8-6**). The pairwise comparison showed that there was a significant difference in the distribution of the telomere length between borderline mutated and significantly mutated samples. While there was no significant difference the distribution of the telomere length between the truly unmutated group compared to the other two subgroups.

Table 8-6. Pairwise comparison of telomere length across three *IGHV* status subgroups . Results of the Dunn-Bonferroni tests on each pair of groups. The tests show that the distribution of telomere length in the borderline mutated group is significantly different to the distribution of the significantly mutated.

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
borderline mutated (97-99.9%)-significantly mutated (<97%)	22.124	6.643	3.330	.001	.003
borderline mutated (97-99.9%)-truly unmutated (100%)	-26.054	15.611	-1.669	.095	.285
significantly mutated (<97%)-truly unmutated (100%)	-3.930	15.160	-.259	.795	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Patients with *KLF2* mutations, 7q and 17p deletions showed significantly shorter telomere length than wild type (WT) patients. Samples characterised by genomic complexity, defined as having three or more CNAs (identified by methylation arrays), also had significantly lower telomere length than those with a simple genome (two or less CNAs). Lastly, patients with trisomy 3 showed longer telomeres than WT patients (**Figure 8-22**).

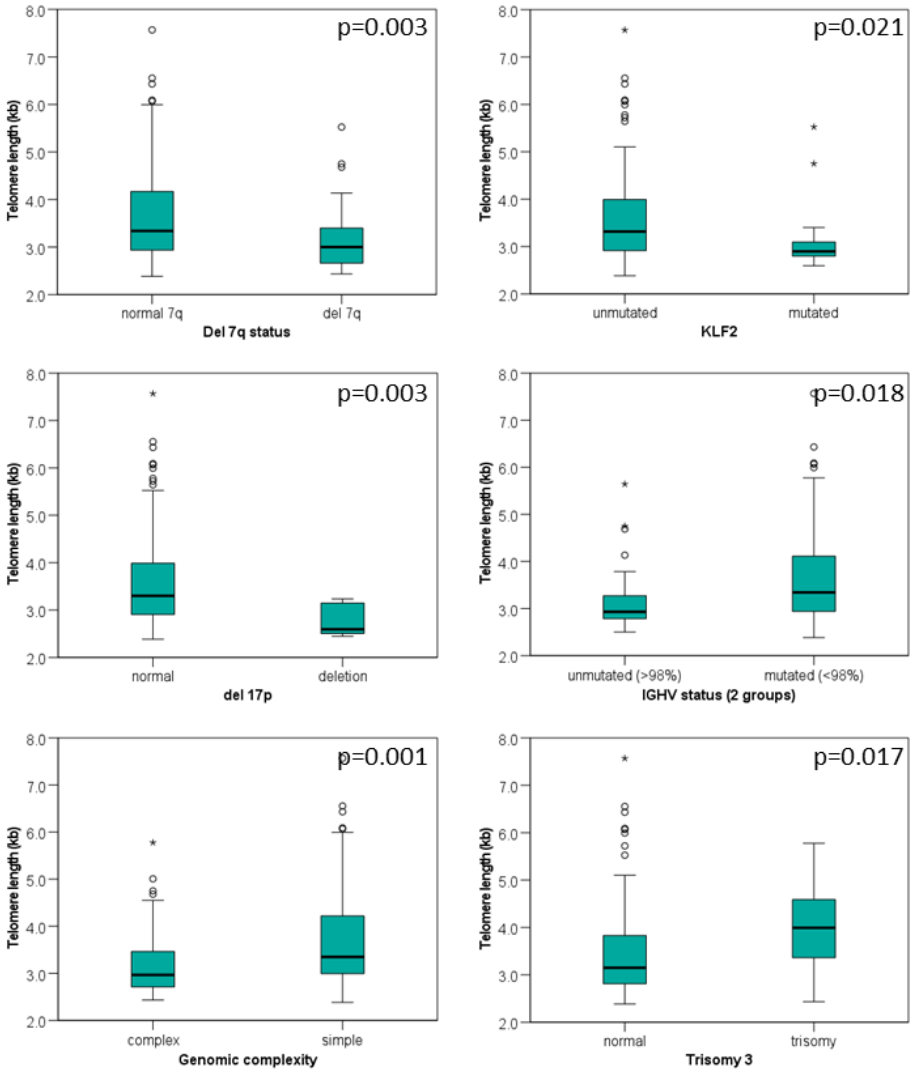


Figure 8-22. Distribution of telomere length across relevant genomic abnormalities. p-values obtained using Mann-Whitney U-test.

8.4.4 Genomic aberrations associate with clinically relevant biomarkers

To determine if there were any significant interactions or associations between genes and other molecular biomarkers pairwise Fisher's exact tests were carried out among the most recurrent genetic aberrations. **Figure 8-23** shows the results of the pairwise associations among significantly mutated genes, genetic and immunogenetic features across the combined Jaramillo-Parry cohort. Mutual exclusivity is shown in pink, while co-occurring relationships are shown in green. Those that remained significant after Bonferroni correction are marked by an asterisk (*). Number in brackets details number of affected cases. The number of cases included in each pairwise test varied per variable as values for all variables were not available for all cases. A total of 1332 pairwise comparisons were tested, where 73 were significant before correction and only 14 remained significant after Bonferroni correction. Most significant interactions were co-occurring interactions.

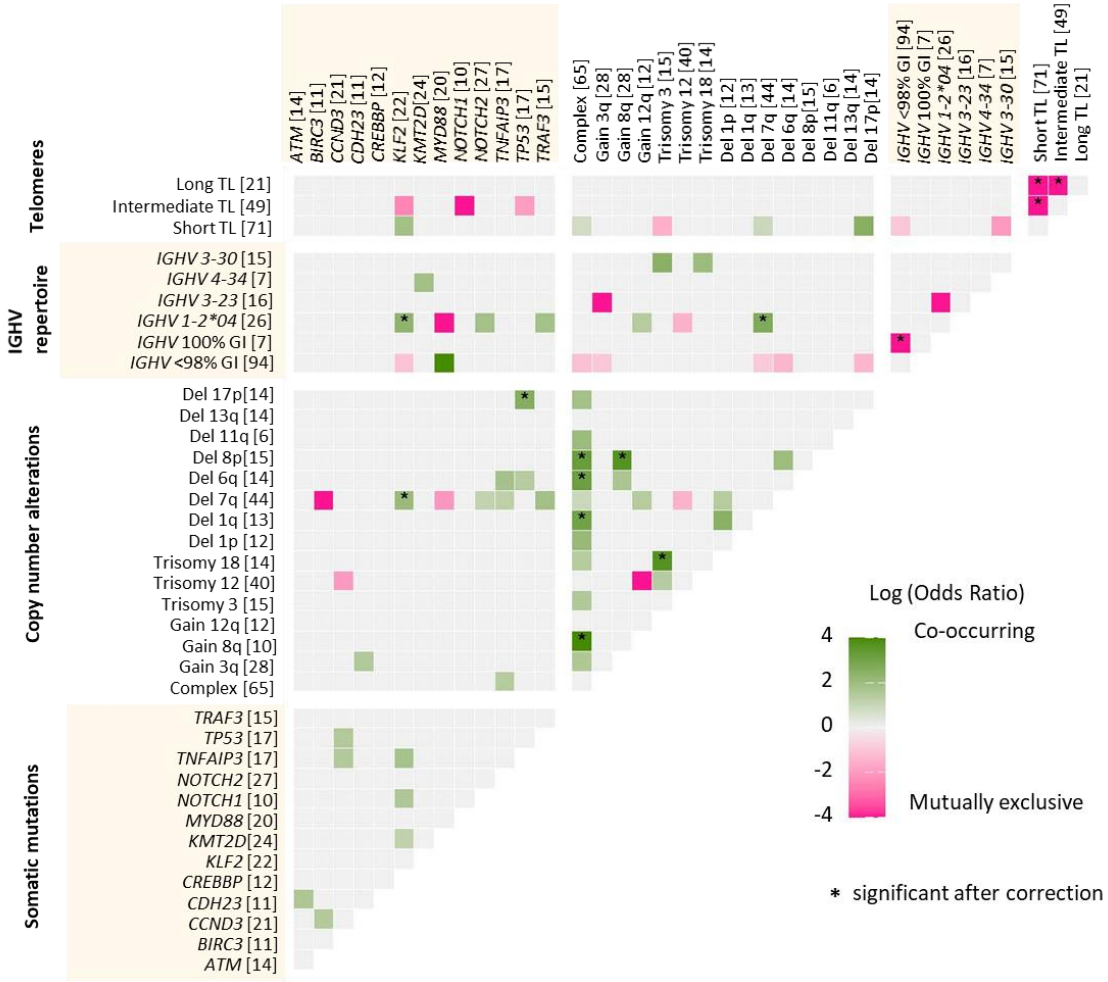


Figure 8-23. Interactions between genomic and clinical features. Interactions were detected by pair-wise Fisher's exact test. Co-occurrence is shown in gradients of green while mutual exclusivity in gradients of pink. Those with a p-value less than 0.05 are not shown and are coloured grey. Starred values are those that remain significant after Bonferroni correction. Number in brackets details number of affected cases. Total number of samples included in each test varied per variable as values for all variables were not available for all cases.

7q deletions were associated with short telomere length (p=0.023), *IGHV1-02*04* (p < 0.001), genomic complexity (p=0.011), gain of 12q (p=0.018), 1p deletion (p=0.018), as well as *KLF2* (p < 0.001), *NOTCH2* (p=0.015), *TNFAIP3* (p=0.019) and *TRAF3* (p<0.001) mutations. Furthermore, 7q deletions were mutually exclusive to *BIRC3* (p=0.036), *MYD88* (p=0.015), trisomy 12 (p=0.004), and mutated *IGHV* genes (p=0.04).

MYD88 has been previously reported as a mutually exclusive event and those observations are replicated here. The only significant association seen in *MYDD88* (**Figure 8-23**) is with mutated (<98% GI) *IGHV* genes (p=0.011).

The variable with the most associations was genomic complexity (3+ CNAs). Undoubtedly some of the significant associations with some of the CNAs such as deletions of 8p, 6q, and 1q and gain of 8q were present because those variables are the ones that define genomic complexity. Furthermore, cases with genomic complexity were associated with short telomere length

(p=0.026) and *TNFAIP3* mutations (p=0.008). **Figure 8-23** shows *TP53* mutations and 17p deletions as separate variables. In the figure only 17p deletions associated significantly (p=0.003) with genomic complexity. However, when combined *TP53* abnormalities (17p deletion and *TP53* mutations) still showed a significant positive association to genomic complexity, albeit with a lower p-value (p=0.034).

8.4.5 Genomic associations hint at potential disease subtypes

Principal component analysis of the most frequent aberrations showed a clear pattern in terms of what patients were "grouped" together (**Figure 8-24**). On the left of the PCA plot we find cases that are characterised by long telomeres, high levels of SHM, trisomies and somatic mutations in *MYD88* and *BIRC3*. While on the right of the PCA plot, cases were characterised by increasing copy number alterations, shorter telomeres, lower levels of SHM, gains of 8q, deletions of 7q, *KLF2* and *TNFAIP3* mutations.

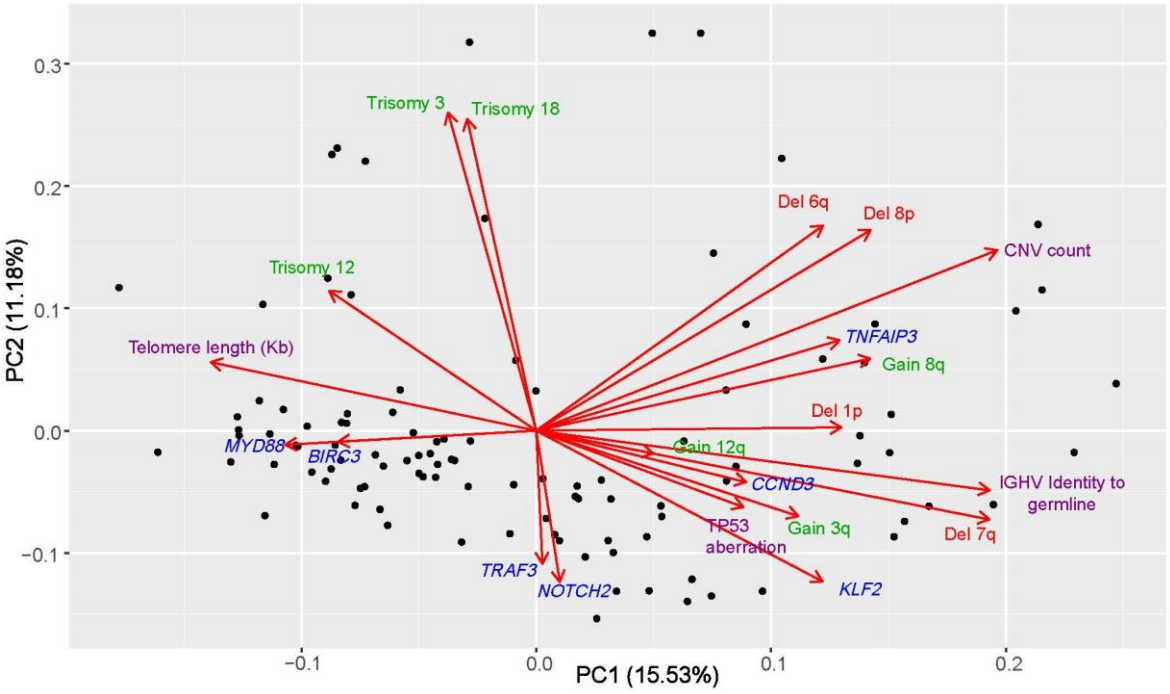


Figure 8-24. Principal component analysis of recurrent genomic aberrations and other molecular biomarkers. Telomere length, number of genomic aberrations, and *IGHV* identity to germline increases from left to right.

Results of the PCA and pairwise comparison hint at the existence of two potential disease sub types, one characterised by *MYD88* mutation and mutated *IGHV* genes, and a second group characterised by deletions of 7q, *IGHV1-2*04* gene usage, and *KLF2* mutations.

8.4.6 Transformation to a high-grade lymphoma is associated with genetic and immunogenetic features

Transformation to a high-grade lymphoma is associated with resistance to treatment and poor survival. Therefore, we looked at a small subset of patients that had transformed [n=31] to see if transformation was associated with any specific feature. In our cohort *TNFAIP3* mutations (p<0.001), *IGHV1-2*04* genes (p=0.013), females (p=0.003) and cases with *IGHV* identity closer to germline (p=0.001) were all associated with transformation (

Figure 8-25).

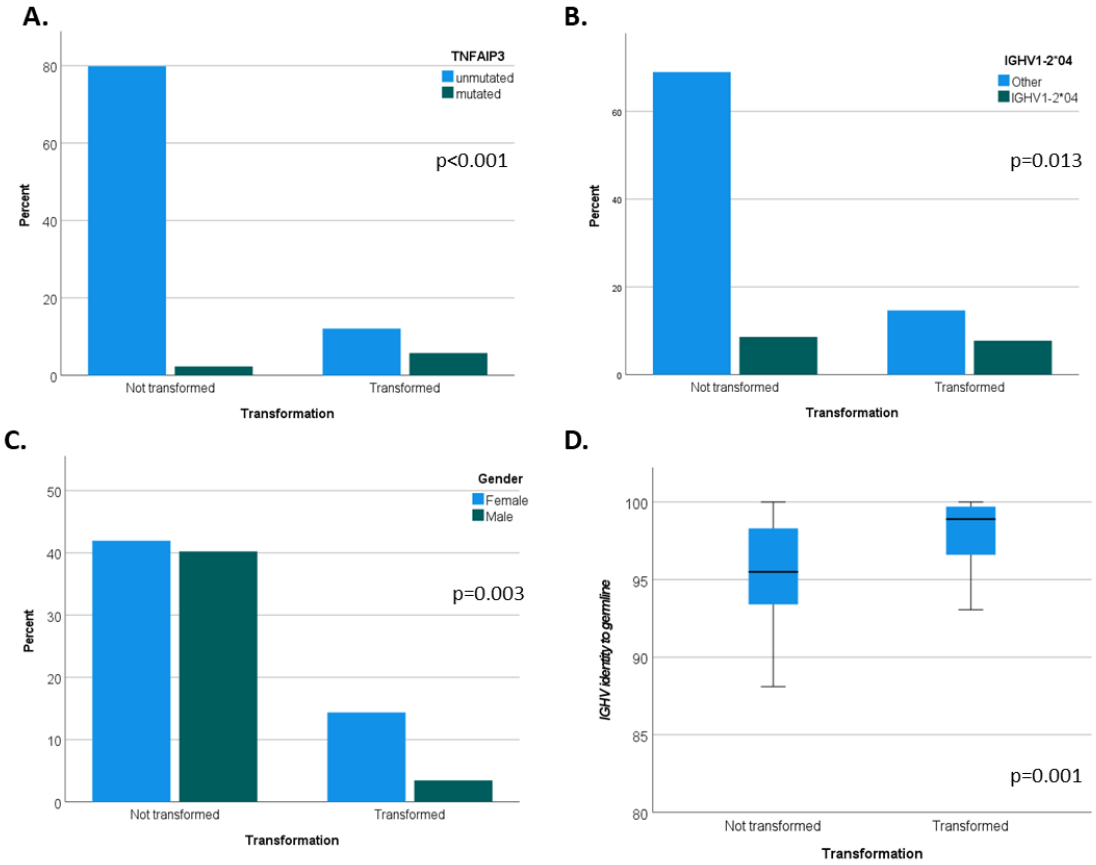


Figure 8-25. Comparison of transformed versus non transformed cases across genetic features. **A.** Number of patients with *TNFAIP3* mutations across transformed and non-transformed cases **B.** Number of patients with *IGHV1-2*04* genes across transformed and non-transformed cases **C.** Number of males and females across transformed and non-transformed cases **D.** Distribution of *IGHV* identity to germline across transformed and non-transformed cases.

8.4.7 Clinical significance of mutations, genetics and immunogenetics

While genomic data was available for all samples, clinical and immunogenetic data was sparse. Therefore, follow-up outcome data for overall survival (OS) was available for a maximum of 271 patients and varied depending on the feature. In univariate survival analysis 11 features were associated with shorter overall survival: *TP53* aberration (*TP53* deletion and or mutation)[HR 3.501, 95% CI 2.009-6.099, p<0.001], *TP53* mutation [HR 3.068, 95% CI 1.698-5.545, p<0.001],

TNFAIP3 mutation [HR 2.544, 95% CI 1.206-5.366, p=0.014], *MYD88* WT [HR 8.573, 95% CI 1.187-61.923, p=0.033], 17p deletion [HR 4.043, 95% CI 1.574-10.386, p=0.004], 7q deletion [HR 2.687, 95% CI 1.462-4.939, p=0.001], 1q deletion [HR 7.712, 95% CI 2.082-28.082, p=0.002], 6q deletion [HR 5.296, 95% CI 1.87-15.004, p=0.002], gain of 8q [HR 13.33, 95% CI 2.336-76.115, p=0.004], a genomic complexity (3+ CNAs) [HR 2.661, 95% CI 1.144-6.186, p=0.023], *IGHV* unmutated genes (< 98% GI) [HR 2.789, 95% CI 1.563-4.978, p=0.001] and age at diagnosis [HR 1.104, 95% CI 1.069-1.14, p<0.001]. It should be noted that 1q deletion, 6q deletion and gain of 8q all co-occur within a complex genome. **Table 8-7** lists the 12 significant features in univariate survival analysis for OS, the median survival time for each feature and subgroup, and the hazards ratio. Kaplan Meier curves for the 11 features can be seen in **Supplementary Figure 1** and **Supplementary Figure 2**.

Table 8-7. Univariate survival analysis for overall survival (OS).

Characteristic	Sub-group	No. of patients	No. of events	Median survival time (years)	Kaplan Meier p-value	Hazards Ratio (Cox proportional hazards)	Hazards Ratio confidence interval 95%	Cox proportional hazards p-value
TP53 aberration	normal TP53	236	41	16.49	p<0.001	3.501	2.009-6.099	p<0.001
	TP53 aberration	41	19	10.00				
TP53 mutation	unmutated	245	45	16.53	p<0.001	3.068	1.698-5.545	p<0.001
	mutated	32	15	10.10				
TNFAIP3 mutation	unmutated	258	52	16.00	0.011	2.440	1.206-5.366	0.014
	mutated	19	8	5.74				
MYD88 mutation	unmutated	247	59	15.20	0.010	0.117	0.016-0.843	0.033
	mutated	30	1	-				
Age	≤ 65 years	106	10	-	p<0.001	1.104*	1.069-1.14*	p<0.001*
	> 65 years	164	50	11.62				
Genomic complexity	complex (3+ CNAs)	41	10	10.11	0.014	2.661	1.44-6.186	0.023
	simple (< 3 CNAs)	87	13	-				
7q deletion	normal 7q	154	30	16.35	0.001	2.687	1.462-4.939	0.001
	del 7q	43	18	10.11				
Gain 8q	normal	122	21	15.14	p<0.001	13.333	2.336-76.115	0.004
	gain	6	2	-				
Del 17p	normal	120	17	-	0.002	4.043	1.574-10.386	0.004
	deletion	8	6	10.00				
Del 1q	normal	121	20	15.14	p<0.001	7.712	2.082-28.082	0.002
	deletion	7	3	-				
Del 6q	normal	117	18	15.14	0.001	5.296	1.87-15.004	0.002
	deletion	11	5	10.00				

*** tested as a continuous variable**

The impact of the variables found to be significant for OS in univariate analysis were further assessed using a multivariate Cox proportional hazard analysis, to account for potential confounding variables. Due to the large number of variables, two multivariate models were made using a backwards stepwise approach. The first model integrated all the mutation data, with disease outcomes accounting for age and gender. This was done for two reasons: 1) To verify that there was no interactions within genes that could be missed by only using significant variables from the univariate analysis and; 2) There was mutation data for all samples which allowed for more patients to be included in the model and therefore have more power. The second model took the significant results from the first model (mutation data) and integrated them with the other variables that were significant in the univariate model also accounting for age and gender.

OS Model 1: A backward stepwise regression was carried out starting with 22 variables (*TP53* status, *ARID1A, ATM, BIRC3, CARD11, CCND3, CDH23, CREBBP, FLNC, KLF2, KMT2D, MAP3K14, NOTCH1, NOTCH2, SETD2, SPEN, TNFAIP3, TRAF3, MYD88, PRKDC*, age at diagnosis, and gender), consisting of 270 cases, where 60 presented an event (death). 18 variables were removed (19th iteration) leaving *TP53* status, *TNFAIP3* mutations, *MYD88* mutations and age at diagnosis in the Cox proportional hazard analysis. All variables except *MYD88* mutations retained significance and these were then used as input for model 2.

OS Model 2: A backward stepwise regression was carried out starting with 11 variables (*TP53* status, *TNFAIP3*, age, gender, deletion of 7q, *IGHV* status, deletion of 6q, deletion of 1q, gain of 8q, genomic complexity, and gain of 3q), consisting of 92 cases, where 18 presented and event (death). 8 variables were removed (9th iteration) and the final Cox proportional hazard model showed that *TP53* status [HR, 4.85; 95% CI, 1.72-13.61], age at diagnosis [HR, 1.13; 95% CI, 1.06-1.221] and gain of 8q [HR, 18.47; 95% CI, 2.73-124.97] all had significant impact on OS (**Table 8-8**).

Table 8-8. Results of multivariate model for OS. Hazard ratio and p-value shown in the 7th and 6th column respectively.

Variable	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
P53 status	1.58	0.53	8.96	1.00	0.003	4.84	1.72	13.61
Age at diagnosis	0.13	0.04	12.85	1.00	p<0.001	1.14	1.06	1.22
Gain 8q	2.92	0.98	8.94	1.00	0.003	18.48	2.73	124.96

Similar to OS, follow-up outcome data for time to first treatment (TTFT) was available for a maximum of 278 patients. In univariate survival analysis 11 features were associated with reduced time to treatment: *TNFAIP3* mutation [HR 1.811, 95% CI 1.134-2.893, p=0.013], *KMT2D* mutation [HR 1.636, 95% CI 1.100-2.435, p=0.015], *TRAF3* mutation [HR 1.763, 95% CI 1.082-2.874, p=0.023], *NOTCH2* mutation [HR 1.818, 95% CI 1.255-2.697, p=0.003], *KLF2* mutation [HR 1.873, 95% CI 1.271-2.758, p=0.002], *ARID1A* mutation [HR 2.239, 95% CI 1.293-3.878, p=0.004], gender (females) [HR 1.692, 95% CI 1.270-2.253, p<0.001], *IGHV* status (unmutated) [HR 1.456, 95% CI 0.994-2.133, p=0.054], *IGHV1-2*04* genes [HR 2.533, 95% CI 1.637-3.917, p<0.001], telomere length [HR 0.741, 95% CI 0.557-0.987, p=0.04], and gain of 3q [HR 2.284, 95% CI 1.311-3.98, p=0.004]. **Table 8-9** lists the 11 significant features in univariate survival analysis for TTFT, the median survival time for each feature and subgroup, and the hazards ratio. Kaplan Meier curves for the 11 features can be seen in **Supplementary Figure 3** and **Supplementary Figure 4**.

Table 8-9. Univariate survival analysis for time to first treatment (TTFT). Patients that have died were censored.

Characteristic	Sub-group	No. of patients	No. of events	Median survival time (years)	Kaplan Meier p-value	Hazards Ratio (Cox proportional hazards)	Hazards Ratio confidence interval 95%	Cox proportional hazards p-value
TNFAIP3 mutation	unmutated	256	174	1.40	0.011	1.811	1.134-2.893	0.013
	mutated	22	20	0.40				
KMT2D mutation	unmutated	243	165	1.80	0.014	1.636	1.100-2.435	0.015
	mutated	35	29	0.19				
TRAF3 mutation	unmutated	258	176	1.40	0.021	1.763	1.082-2.874	0.023
	mutated	20	18	0.31				
NOTCH2 mutation	unmutated	243	164	1.70	0.002	1.818	1.255-2.697	0.003
	mutated	35	30	0.31				
KLF2 mutation	unmutated	243	163	1.80	0.001	1.873	1.271-2.758	0.002
	mutated	35	31	0.26				
ARID1A mutation	unmutated	261	180	1.30	0.003	2.239	1.293-3.878	0.004
	mutated	17	14	0.07				
Gender	female	141	108	0.57	p<0.001	1.692ϕ	1.270-2.253ϕ	p<0.001ϕ
	male	137	86	2.99				
IGHV status	unmutated (≥ 98% GI)	58	43	1.30	0.044	0.676	0.461-0.992	0.045
	mutated (<98% GI)	118	68	2.90				
*IGHV1-2*04*	IGHV1-2*04	180	110	2.73	p<0.001	2.533	1.637-3.917	p<0.001
	other	27	26	0.42				
Telomere length (kb)	short	54	41	0.80	0.113	0.741*	0.557-0.987*	0.04*
	intermediate	27	17	2.94				
	long	30	18	4.13				
Gain 3q	normal	113	73	2.70	0.003	2.284	1.293-3.878	0.004
	gain	18	16	0.36				

*** tested as a continuous variable**

ϕ Male used as reference for Cox regression

The impact of the variables found to be significant for TTFT in univariate analysis were also assessed using a multivariate Cox proportional hazard analysis, to account for potential confounding variables. Just like with OS two multivariate models were made using a backwards stepwise approach.

TTFT Model 1: A backward stepwise regression was carried out starting with 22 variables (*TP53* status, *ARID1A, ATM, BIRC3, CARD11, CCND3, CDH23, CREBBP, FLNC, KLF2, KMT2D, MAP3K14, NOTCH1, NOTCH2, SETD2, SPEN, TNFAIP3, TRAF3, MYD88, PRKDC*, age at diagnosis, and gender), consisting of 275 cases, where 192 presented an event (treated including splenectomy). 15 variables were removed (16th iteration) leaving *ARID1A* mutations*, CCND3* mutation*, NOTCH2* mutation*, TNFAIP3* mutation*, TRAF3* mutation, gender, and age at diagnosis in the Cox proportional hazards analysis. All variables except *CCND3* and *TRAF3* mutations retained significance and these were then used as input for model 2.

TTFT Model 2: A backward stepwise regression was carried out starting with 9 variables (*ARID1A, NOTCH2, TNFAIP3*, *IGHV* status, *IGHV1-2*04*, gain of 3q, telomere length, age at diagnosis and gender), consisting of 88 cases, where 55 presented an event (treatment including splenectomy), 9 variables were removed (7th iteration) and the final Cox proportional hazard model showed that gender [HR, 2.58; 95% CI, 1.46-4.57], gain of 3q [HR, 3.69; 95% CI, 1.91-7.13], and telomere length [HR, 0.68; 95% CI, 0.49-0.94] all had significant impact on TTFT (**Table 8-10**).

Table 8-10. Results of multivariate model for TTFT. Hazard ratio and p-value shown in the 7th and 6th column respectively

Variable	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gain 3q	1.17	0.33	12.79	1.00	0.000	3.23	1.70	6.15
Telomere length (kb)	-0.34	0.17	4.17	1.00	0.041	0.71	0.51	0.99
Gender	0.95	0.29	10.88	1.00	0.001	2.59	1.47	4.56

8.5 Discussion

Integration of somatically acquired mutations and other molecular biomarkers with disease outcomes is an important step in translating results from research into patient care. Having a granular understanding of the molecular drivers that underpin specific phenotypes and outcomes in SMZL will not only aid in the risk -adapted stratification of patients, it has the potential to steer prospective clinical trials to find novel therapies for these patients. Research over the last decade has begun to shed some light on the molecular pathogenesis of SMZL, but there are still many unanswered questions. The main aim of this chapter was to integrate somatic mutations with other molecular biomarkers as well as disease outcomes, to better characterise the disease and identify potential disease subgroups and risks factors associated with survival.

Somatic copy number profiles were obtained using methylation arrays. However, determining precise segment cut-offs using methylation arrays has its limitations. Firstly, the fact that CpG dinucleotides are not uniformly distributed across the genome[219], means coverage across the chromosomes will not be uniform. Secondly, identification of the start and end points of any CNA will vary depending on the number and quality of controls used. Lastly, outputs from the conumme package require extensive manual curation, potentially introducing error and or bias. However, even with these limitations, our results echoed what had already been found in previous SMZL cohorts adding confidence to our results.

In line with previous studies[45,47,71,202] deletions of the long arm of chromosome 7, trisomy 12 and gains of the long arm of chromosome 3 were the most recurrent CNAs within the cohort. Due to the high frequency of 7q deletions and the potential biological importance of this region in the pathogenesis of SMZL, researchers have attempted to identify the gene or genes targeted by it but results have only provided putative targets which require further study. Early investigations explored miRNA as potential targets of the 7q deletion. The first study led by by Ruiz-Ballesteros and colleagues measured miR-29a and miR-29b-1 expression (chosen due to their proximity to the 7q) in SMZL subgroups and other B-cell lymphomas[220]. They proposed miR-29a and miR-29b-1 as candidates for regulation of TCL1A, which had been found to be over-expressed in SMZL[221]. Subsequently, Watkins et al. showed a reduced expression of 7 miRNA (miR-593, miR-129, miR-

182, miR-96, miR-183, miR-335, miR-29a and miR-29b-1) in cases with 7q deletion[222]. The study by Watkins and colleagues also investigated DNA methylation profiles in a subset of cases in the context of 7q deletions where out of the 37 targeted genes, *CPA4, OPN1SW, NAG8, LRRC4, CPA5, CPA2, TSGA13, CPA1, C7orf45* and *NYD-SP18* consistently showed high levels of methylation but low gene expression[218]. More recently high throughput sequencing has been used to identify putative target genes that fall within the 7q deletion region, but the frequency of mutations within genes in this region remains low. Deep sequencing in four MZL-derived cell lines has identified four putative target genes (*IRF5, TMEM209, CALU* and *ZC3HC1*), however, results did not find any clear pathogenic mutations[204]. The study by Parry et al. also identified somatic mutations within the 7q MDR, but in isolated cases within genes *CUL1*, *FLNC* and *EZH2*[51,71].

Within this work, two MDRs were identified in chromosome 7. The first MDR was located on band 7q31.33 and was around 2,900 kb (chr7:124,200,000-127,100,000). The second MDR was located on band 7q32.2 and was approximately 1,752 kb (chr7:128,575,000-130327262). These two region fell within the 7q31.33-7q32.2 bands, aligning with previously identified MDRs[47,216,217,223,224]. Our gene panels only targeted two genes within these region (*POT1* and *FLNC*) but no samples showed co-occurring deletions and mutations. There could be pathogenic mutations within other genes in this locus, but this is impossible to determine without WGS or WES. Other possible targets include microRNAs, or even genes (somewhere else in the genome) regulated by elements within this locus. It was interesting that *POT1* fell within one of the MDRs, as it is a telomere protection gene. Inhibition of POT1 will result in telomere fragility, replication fork stalling, and genome instability[225,226]. In murine models, depletion of murine POT1a combined with p53 deficiency fuelled cancer progression in T-cell lymphomas[225]. Within our cohort 7 patients had both deletion of *POT1* and *TP53* aberrations. 6/7 patients had follow-up clinical data and of those 4/6 patients transformed into large B-cell lymphoma. Though this is a small subset of patients, this interaction might give us more clues to the potential drivers underpinning transformation and targets of the 7q deletion.

Although the role of the 7q deletion is still unclear, the results of this analysis have mirrored previous published work which have shown that deletion breakpoints at 7q are heterogeneous, with q21 the most proximal and q36 the most terminal breakpoints[47,48,216,227–230]. Deletions of 7q have been associated with unmutated *IGHV* genes*, IGHV1-2*04* usage and *KLF2* and *NOTCH2* mutations[42,47,71], observations which have all been validated here. Furthermore, this work has provided new insights as 7q deletions were also associated with short telomere length, genomic complexity, *TNFAIP3* and *TRAF3* mutations but mutually exclusive to trisomy 12, *MYD88* and *BIRC3* mutations.

Unfortunately, there is still no clear target of the 7q deletion and future studies will require the integration of several omic technologies as well as a genome wide approach to fully comprehend the role of the 7q deletion in SMZL. Nonetheless, this alteration could serve as a potential diagnostic marker. [41,45,204,221,231]Although deletion of 7q is not exclusive to SMZL, evidence suggests that SMZLs with 7q deletion combined with somatic mutations in *KLF2* and *NOTCH2* are highly specific for SMZL and could be used for differential diagnosis[54,71,232].

Trisomy 12 was the second most recurrent CNA found in 24% of patients. However, very little is known about its pathophysiology in SMZL. Not only has this aberration been reported at much lower frequencies (0-12%), there is conflicting evidence with regards to its prognostic impact[45,47,233]. Our results showed no impact on OS or TTFT and no significant associations with other genomic features. We did observe that patients with trisomy 12 do not tend to occur in those that have *IGHV1-2*04* genes and those with 7q deletions.

Recurrent gains of chromosome 3, 12 and 18 are common in marginal zone lymphomas (MZLs)[75]. However, gains of 3q are more frequent in SMZL than other MZLs[75,202]. In early studies it was thought that cases with gains of 3q were a separate cytogenetic subtype than those with 7q deletions[46], however, we did not find them to be mutually exclusive. As for the biological importance of this gain, potential gene dosage effect and differential expression of genes has been suggested, considering genes in this region are important in neoplastic transformation[234]. The work by Rinaldi and colleagues did in fact show overexpression of different genes, including *FOXP1, NFKBIZ*, and *BCL6* which might provide a survival advantage to neoplastic cells in all MZLs with gains of 3q[202]. In SMZL gains of 3q have been associated with genomic complexity[47], an observation validated here. Furthermore, within this work this aberration was identified as an independent risk factor in TTFT.

Deletion of 6q23-24 is a frequent abnormality in non-Hodgkin's Lymphoma[235,236]. In Waldenström macroglobulinemia and chronic lymphocytic leukaemia for example, patients with 6q deletions are more likely to display features associated with worse prognosis[237]. For chromosome 6, a minimally deleted region was not established since 37% [n=10] of segments were smaller than 5 Mb and did not overlap. However, region 6q23.3-q24.1 was a clear target of deletions, which included gene *TNFAIP3*. Additionally, 3/12 patients with deletions involving *TNFAIP3* also harboured somatic mutations (1 biallelic and 2 mono allelic frameshift deletions). As outlined in the previous chapter TNFAIP3 is a negative regulator of NF-κB signalling and acts as a tumour suppressor by halting canonical NF-κB activation (for further details on the role of TNFAIP3 refer to **section 7.5.3**). Within this work there is evidence that this gene is a potential driver of SMZL, considering the significant enrichment of mutations within transformed cases and results of

univariate survival analyses. In a study by Boonstra et al. of 13 SMZL cases using cytogenetic analysis and comparative genomic hybridization (CGH) two cases were found to have deletions in 6q where one of the patients went on to transform to large cell lymphoma[235]. This transformed patient however, showed deletion of 6q24 by array CGH which did not include *TNFAIP3*.

Another potential driver could be found within chromosome 8. Two large MDRs were identified within the short arm of chromosome 8 with putative targets including *TNFRSF10A*, *TNFRSF10B*, *DOK2* and *BLK*. *TNFRSF10A* and *TNFRSF10B* code for tumour necrosis factor-related apoptosis-inducing ligand (TRIAL) receptors which are important regulators of B-cell selection and germinal centre homeostasis[238]. Interestingly, across all B-cell, it is the GC cells that express the highest levels of all TRAIL receptors[239]. *DOK2* is a negative regulator of TLR/MYD88 signalling while *BLK* plays a role in B-cell receptor signalling and development[240]. Additionally, within our cohort it was found that gains of the long arm of chromosome 8 were associated with genomic complexity, loss of 8p and shorter overall survival. A potential target of this gain could be the *MYC* gene, frequently involved in human cacinogenesis[241,242]. *MYC* belongs to a family of transcription factors that bind to DNA in a non-specific manner[243]. It is a proto-oncogene, in charge of a multitude of cellular functions including cell cycle, cell growth and survival[244]. Unlike most proto-oncogenes which drive cells to a malignant state through somatic mutations, *MYC* drives transformation of cells via overexpression (gene amplification, chromosomal translocations or aberrant regulation of expression)[245]. The study by Fresquet et al. provides further evidence to the potential role of MYC in SMZL disease progression, where patients with gains of 8q (including the *MYC* locus) were associated with poor clinical outcomes[204].

Less common lesions affected chromosome 13 and chromosome 1. The MDR in chromosome 13 although adjacent to, did not include well known MDR in CLL comprised by the *DLEU2* gene, the *MIR15A/MIR16-1* cluster, and the first exon of the *DLEU1*[203], a locus which plays an important role in expansion of mature B-cells[246]. However, in ten patients the CLL MDR was deleted. In chromosome 1, the MDR included *ARID1A,* although with no concurrent somatic mutations. As mentioned in the previous chapter ARID1A is a component of the SWI/SNF chromatin remodelling complex, which regulates transcription via alteration of chromatin structure. Deletion of *ARID1A* could lead to the loss of both the caretaker and gatekeeper function in cells[193] (for further details on the role of ARID1A refer to **section 7.5.47.5.3**).

Immunoglobulin gene usage has been previously investigated in large SMZL cohorts and have shown strong repertoire bias in the immunoglobulin heavy chain genes, namely *IGHV1-02, IGHV4-34* and *IGHV3-23*[205,206]. By far the most frequent has been *IGHV1-02* where the majority use *the IGHV1-02*04* allele[205–207]. Our cohort validated previous observations, with most cases using the

*IGHV1-02*04* gene. Cases with *IGHV 1-2*04* genes were significantly associated with deletion of 7q, *KLF2* and *NOTCH2* mutations but mutually exclusive to *MYD88* mutations. Such a strong bias of the *IGHV* gene repertoire points to an antigen selection process in the pathogenesis of the disease. Results from this and previous studies[54,71], along with preliminary DNA methylation and transcriptomic studies[247,248], suggest cases with *IGHV1-02*04* represent a distinct patient sub-group, that is likely to emerge from a cell with a distinct ancestry and/or unique immune activation process followed by transformation, ongoing antigen exposure, with shared genomic lesions and poor survival[71].

The critical part that the BCR plays in SMZL is further emphasized by the presence of somatic hypermutations (SHM), as only 8% of assessed cases lacked evidence of SHM at the *IGHV* locus and might be considered truly unmutated. The remaining cases exhibited evidence of SHM at the *IGHV* locus, with 36% and 56% defined as borderline (97%-99.9% *IGHV* gene identity to germline) and significantly mutated (< 97% *IGHV* gene identity to germline), respectively. These percentages are all in line with previous work where roughly 12% of SMZL cases showed no evidence of SHM while 38% and 50% were defined as borderline and significantly mutated[205,206]. The low number of samples with truly unmutated genes adds evidence suggesting an antigen exposure in most SMZL cases.

Whilst levels of SHM represent a critically important prognostic and predictive biomarker in CLL[249,250], their clinical utility in SMZL is less established. In previous studies[202,251] that used SHM cut-offs established for CLL in SMZL patients, there was no difference in progression-free and overall survival between mutated and unmutated *IGHV* cases. However, it is likely that SHM levels are disease specific and better cut-offs need to be established within SMZL. Within our cohort *IGHV* status (using CLL cut offs) did have an impact on OS, but only in univariate analysis. Patients that transformed were significantly associated with *IGHV* identity closer to germline and had an enrichment of borderline/minimally mutated *IGHV* genes. This is slightly different from previous observations by Parry and colleagues where the complete absence of SHM (truly unmutated), provided independent prognostic information, with truly unmutated cases exhibiting reduced time to treatment[71]. Our analysis included all cases from the Parry study and differences between our observations are likely due to sample size and updated clinical information for some patients.

Another useful biomarker and predictor of survival in CLL that could potentially be used in SMZL is telomere length[215]. Telomeres play a key role in genome integrity, where in normal tissue, attrition of telomeres will lead to activation of senesce checkpoints, playing a tumour suppressing role[252]. Telomere attrition leads to uncapped chromosomes which will become unstable until they are capped[252]. This in turns leads to intra- and inter-chromosomal end fusions, the formation of

dicentric chromosomes with consequential breakage during anaphase, and genomic complexity, through the mechanisms of breakage-fusion-bridge formation[253]. The majority of human tumours exhibit eroded telomere length, compared to the corresponding normal tissue[254]. To date, preliminary telomere length analysis of SMZL has only been published in abstract form showing an enrichment of short telomeres in patients with progressive disease[255]. Our results validate these observations, where survival analysis showed telomere length had significant impact on time to first treatment. Similarly, cases with *KLF2* mutations, 7q and 17p deletions and genomic complexity (all features associated with progressive disease) all showed significantly shorter telomeres than WT patients. Interestingly patients with trisomy 3 had significantly longer telomeres than those without the trisomy, suggesting a more benign disease course for these patients.

The integration of the sequencing results with the additional molecular biomarkers and survival data available within our cohort suggest that there are at least two distinct molecular subgroups with distinct prognosis. This was first evidenced by the prevalence of certain CNAs. The prevalence of whole chromosome gains compared to the prevalence of gains of long arms, as is the case with chromosome 3 and 12, likely reflect different underlying mechanisms of lymphomagenesis[202]. In our case principal component analysis of the most recurrent abnormalities and other molecular biomarkers did associate cases with whole gain of chromosome 3, 12 and 18 separately than those with gains of 3q and or 8q. The second was co-occurrence of 7q deletions (the most recurrent CNA), in cases with *KLF2* mutations, *NOTCH2* mutations and *IGHV1-2*04* genes. This subset of cases was mutually exclusive to *MYD88* mutations, showed low levels of SMH and short telomeres. PCA also grouped these cases much closer together and to features that we associated with poor outcomes. On the other hand, there is the subset characterised by *MYD88* mutations, long telomeres, and high levels of SHM. These cases were mutually exclusive to *IGHV1-2*04* genes and showed better prognosis. In the PCA these were grouped with trisomy 12 and trisomy 3, the latter characterised by long telomeres likely to have a more benign course of disease.

8.6 Conclusions

The work outline herein has improved our understanding of the biological basis of SMZL. We validated previous observations including: 1) Recurrent CNAs (with deletion of 7q being the most frequent); 2) A highly restricted *IGHV* gene repertoire, including selective usage of the *IGHV1-2*04* allele in 15% of cases; 3) Evidence of SHM in the majority of cases, (only 8% showed no evidence of SHM) and; 4) Association of *IGHV1-2*04* to deletions of 7q, *KLF2* and *NOTCH2* mutations, and low levels of SHM. We provided additional evidence supporting the prognostic

impact of *TP53* aberrations, *NOTCH2* mutations and age of diagnosis and identified telomere length and gains of 3q and 8q as new potential prognostic factors. Furthermore, integration of different molecular markers suggested two distinct molecular subgroups with potential prognostic significance. Additional studies across multiple discovery and validation cohorts, as well as prospective clinical trials are required to validate results, particularly disease outcomes.

Chapter 9 Discussion and future directions

9.1 Discussion

9.1.1 Current state of play

Advances in next generation sequencing (NGS) technologies have transformed our understanding of prevalent mature b-cell tumours proving their utility in the clinical application for more sensitive diagnoses and targeted treatments. Chronic lymphocytic leukaemia (CLL) is a great example, where outcomes have improved over recent years due to the risk-adapted patient stratification and impact of novel therapies underpinned by a deep understanding of the biology of the disease[256,257]. In CLL several thousand patients have been examined either with whole genome sequencing (WGS), whole exome sequencing (WES) and or targeted sequencing. This has allowed the identification of recurrent coding and non-coding mutations targeting key biological processes, description of mechanisms such as chromothripsis and kataegis, and identification of mutations leading to therapy resistance[258–261]. Furthermore, understanding of the epigenetic mechanisms that play a role in CLL have provided additional biological insights as DNA methylation has allowed the classification of patients into three groups exhibiting different clinico-biological features[257,262]. The most accepted and validated genomic prognostic model for CLL patients is the Dohner model, based on the presence of deletions of 17p, 11p, 13q and trisomy 12, which has evolved to include mutational data[263,264]. Additionally, a recent study from the randomised UK LRF CLL4 trial supported the use of targeted resequencing to elucidate the prognostic impact of gene mutations[101].

Splenic marginal zone lymphoma (SMZL) is currently precluded for large international sequencing projects such as The Cancer Genome Atlas (TCGA), resulting in an incomplete catalogue of tumour associated genomic lesions for this rare cancer. There is lack of matched whole genome sequencing (WGS), as only Kiel *et al.*[49] has reported WGS but without matched-germline material. Whole exome sequencing (WES) studies of SMZL are also limited, with only five studies being reported to date on 35 discovery cases[70]. There are few studies assessing copy number (CNA) and structural aberrations of large cohorts[45,47,202] and targets of CNAs are mostly unknown, particularly targets of the deletion of the long arm of chromosome 7[45,204]. In SMZL there are no genomic prognostic models like the Dohner model in CLL. The recommended model is the HPLL score proposed by the Splenic Marginal Zone Lymphoma Study Group based upon three factors: 1) Haemoglobin concentration; 2) Platelet count and; 3) Lactate dehydrogenase (LDH) level and

extrahilar lymphadenopathy[40,265]. Furthermore, there is poor definition of biological disease subgroups and although SMZL is an indolent lymphoma, there is a need for biomarkers to help define subtypes that progress into an aggressive disease.

Most cancers are rarely characterised by a single mutation, however there are mutations found more frequently in specific cancers proving useful as diagnostic markers[266]. An example of this is the *BRAF* V600E variants in hairy cell leukaemia (HCL) where the diagnosis is based upon clinical and laboratory findings but the presence of this specific mutation is helpful in confirming diagnosis and is considered the causal genetic event[267]. Similarly, in Waldenström macroglobulinemia patients, although *MYD88* L265P mutations are not wholly unique to these patients, this mutation is present in approximately 90% of cases[268]. With regards to the prognostic significance of mutations in mature B-cell malignancies, the evaluation of mutations is constantly evolving as new genes are being discovered and methods for evaluating their clinical significance are also developing. An example of this can be observed in two widely studied lymphomas, diffuse large B-cell lymphoma (DLBCL) and chronic lymphocytic leukaemia (CLL). In CLL mutations in *NOTCH1, SF3B1, ATM, BIRC3* and *TP53* have been recognised as mutations of potential clinical relevance by the World Health Organisation in their latest update[31]. Moreover, the presence of *TP53* mutations or deletions of chromosome 17 encompassing *TP53*, impact choice of therapy as patients with these aberrations respond poorly to chemo(immuno)therapy but have improved clinical outcomes when treated with nonchemotherapeutic agents[269]. In DLBCL, a recent study which defined the landscape of 150 genetic drivers and integrated the results of mutational profiling, gene expression, CRISPR screens and clinical outcomes, developed a genomic risk model that outperformed current established methods: cell of origin, the International Prognostic Index comprising clinical variables, and dual *MYC* and *BCL2* expression[270]. The genomic risk model was able to identify those patients that would benefit from standard therapy and potentially inform sensitivity to currently available therapies[270]. In SMZL, however, the picture is not as clear. Similar informative genomic risk models of SMZL are not available to inform tumour profiling and optimal therapeutic intervention.

This project aimed to construct a detailed characterisation of the genomic landscape of SMZL through the identification of somatic variants in unmatched tumour samples across a panel of genes hypothesised to be of importance in SMZL and other mature B-cell malignancies. The cohort used represented the largest SMZL cohort assessed to date and enabled exploration of the clinical significance of genomic alterations by integrating relevant clinical data. In conjunction with the analysis of somatic variants, an important part of this project also centred around the

bioinformatics processing and optimisation of a bioinformatics pipeline for tumour only samples as no gold standard or best practice for processing this type of data had been established previously.

9.1.2 Considerations when sequencing and processing tumour only samples

Tumour heterogeneity in cancer samples makes the bioinformatics processing more complex than for germline samples. This heterogeneity requires many of the algorithms used in germline tissue to be modified to successfully identify variants found at low frequencies in tumour samples. This project highlighted the importance of choosing the right tools to process the data particularly when it came to choosing a variant caller. For unmatched tumour samples, GATKs haplotype caller proved the best option as it removed at least 60% of false positives without compromising sensitivity. However, a more robust analysis comparing GATKs haplotypecaller to somatic variant callers is desirable, with data that has been orthogonally validated. Moreover, there is still a need for a gold standard pipeline for processing unmatched tumour samples. Ideally, all cancer genomic studies should be conducted in parallel with matching germline tissue, however, in rarer cancers such as SMZL, obtaining even the tumour samples can prove difficult and necessitate the methods developed in this thesis.

Preliminary sequencing results investigated here identified a considerable burden of false positive variant calls that were not excluded with the initial filtering strategy. Due to the large number of samples and variants, additional sequencing was not feasible or cost effective. Manual review of sequencing data using the Integrative Genome viewer (IGV) represented the most accessible method for validating results, however this approach was highly time-consuming and could be subject to human error. Nevertheless, these validated data enabled the development of an unsupervised machine learning algorithm to automate the review of variants for digital stratification into true positives and false positives. The model demonstrated good performance particularly with libraries prepared using the Haloplex HS kits, compared to the TruSeq kits. We concluded that this was attributable to the amplicon design, as the Haloplex kits, despite being sequenced at lower depth, had more overlapping amplicons than the TruSeq design. Ultimately, the machine learning (ML) model was used as a triage tool rather than a strict filter as it required further fine-tuning to improve its sensitivity. The inclusion of the unique molecular identifiers also increased the confidence when identifying variants with lower depth, as the reads belonged to a unique biological molecule rather than a PCR duplicate. For future studies sequencing tumour only tissue, kits with unique molecular identifiers such as the HaploPlex HS kits represent a good

option for cleaner higher confidence data, as well as an amplicon design with as many overlapping amplicons as possible.

9.1.3 Genomic landscape of SMZL

Following application of the ML approach to filter noise sequencing data, a list of high-fidelity variants observed in SMZL tumours was available for analysis. Association with clinically relevant markers of patient outcome confirmed some of the key genomic signatures that had been previously implicated in SMZL. Mutations across a wide range of genes, for the most part, genes and pathways that had already been associated with SMZL were successfully identified[55]. Key pathways included marginal zone (MZ) B-cell development, NF-κB signalling, cell cycle control and epigenetic modifiers. Moreover, the prevalence of mutations within *NOTCH2, KLF2, KMT2D, TP53* and regulators of the NF-kB pathway (*MYD88, TNFAIP3, TRAF3*) were validated. The frequency of mutations in most genes was in line with what had been previously seen in published cohorts, with the exception of variant burden in *CCND3* and *KMT2D* which were higher than expected. Furthermore, despite our targeted approach, the cohort size allowed us to identify new variants and two new mutation hotspots, one in *KLF2* and another in *CCND3.*

In the *CCND3* gene 40% of variants were identified within a new gain-of-function mutation hotspot affecting the PEST domain. The mutation hotspot suggests a strong selection for these gain of functions variants that will increase the proliferative capacity of B-cells within the dark zone of the germinal centres[196,198]. Isolated mutations in *CCND3*, which lead to overexpression of cyclin D3, do not lead to a lymphoproliferative phenotype but they can in conjunction with other oncogenic factors (i.e. Burkitt's lymphoma where it co-occurs with *Myc-Igh* translocations)[198–200]. *CCND3* mutations have a high incidence in splenic diffuse red pulp lymphoma (SDRPL) and the presence of these mutations have even been suggested as a differential diagnostic marker between SMZL and SDRPL[35,271] . However, the high prevalence of *CCND3* in our data suggest that perhaps *CCND3* mutations are not a good molecular marker to distinguish these entities. The high prevalence of mutations within *CCND3* and *KMT2D* in the SMZL cohort also suggests a germinal centre origin. *CCND3* plays a key role in the GC centre affinity maturation reactions and both *CCND3* and *KMT2D* are frequently mutated within GC-derived lymphomas such as Burkitt's lymphoma, follicular lymphomas (FL) and diffuse large B-cell lymphoma (DLBCL)[196,272]. Similarly, loss of function *KMT2D* mutations are characteristic of GC derived lymphomas such as FL and DLBCL[186,273].

Findings detailed in **Chapter 7**, validated the importance of NF-κB dysregulation in SMZL tumours, a key target that could be useful in the development of new therapies. Currently there are five clinical trials[274–278] assessing the utility of Bruton's tyrosine kinase (BTK) and phosphoinositide 3-kinase (PI3K) inhibitors. BTK is a component of the BCR signalling pathway and PI3K has a role in the BCR-mediated downstream activation of NF-κB[279,280]. However, participation in the trails was not based upon molecular biomarkers but on patients who were refractory to, or relapsed after, one or more prior therapies, usually including an anti-CD20 antibody[55].

The genomic analysis in this thesis was limited by the lack of germline tissue, and therefore only putative somatic variants can be put forward. The nature of the targeted approach also meant that there was a percentage of samples [19%, 61/321 patients] in which no putative somatic mutations were identified. It is likely that the gene panels are missing potential targets, and the lack of mutations could be attributed to the approach rather than the biology. These results highlight the need for more objective whole genome approaches in the study of SMZL especially with matched germline tissue. Furthermore, in the case of missense mutations, even with the use of predictive scores such as CADD, acquiring further evidence of their effect on proteins was difficult, and in key genes such as *KLF2, CCND3, MYD88* and *KMT2D,* functional analyses are required to understand if and how these variations alter molecular pathways in SMZL.

Methylation arrays allowed the identification of CNAs within a subset of samples and validated results from previously published cases. As expected, deletion of 7q was the most recurrent CNA and the minimally deleted region (MDR) identified was in concordance with other reported MDRs. Statistical analyses showed that, 7q deletions had no prognostic significance, however 7q deletions were found to be associated with a number of other biological features, most notably *IGHV1-02*04* usage, *KLF2* and *NOTCH2* mutations, genomic complexity and short telomere length. Both NOTCH2 mutations and short telomere length were independent risk factors for time to first treatment. *NOTCH2* had been previously associated with adverse clinical outcome[49] while this is the first-time telomere length and its impact on survival have been assessed in SMZL. The results from our analysis show that as for CLL, short telomeres are associated with other poor prognostic clinical and genetic characteristics, translating to shorter survival for patients compared to those with longer telomeres[281]. Considering there are no established prognostic markers in SMZL, telomere length represents a viable candidate that could be used for risk stratification as well as monitoring of these patients. Moreover, the biological importance of other recurrent CNAs such as gain of 3q, 6q and 8q need to be explored further. Particularly patients with gains of 8q as this was found to be an independent risk factor for overall survival along with age and *TP53*

aberrations. As mentioned in **section 8.5** the long arm of chromosome 8 harbours the proto-oncogene *c-MYC* which could be a potential target of this gain. *MYC* tends to drive transformation through amplification[245], however functional analyses are required to better understand how gains of 8q observed in SMZL correlate with amplification of the *c-MYC* gene.

IGHV gene usage and somatic hypermutation (SHM) was another important biomarker examined in this study. In chronic lymphocytic leukaemia (CLL) SHM is a clinically relevant prognostic and predictive biomarker and it is implicated in defining the origin of the tumour[249,250]. According to the iwCLL guidelines[269] CLL patients with unmutated genes (98% or more sequence homology to germline) show inferior outcome compared to those with mutated genes. Furthermore, those with mutated *IGHV* genes, particularly in combination with other favourable prognostic factors show excellent outcome following chemoimmunotherapy with fludarabine, cyclophosphamide and rituximab[269]. In SMZL, SHM may have similar implications and could provide clues towards the identification of the cell of origin which is still under debate. Moreover, very few cases showed no evidence of SHM, indicating that most cells probably do go through antigen exposure. In this study low levels of SHM were associated with worse overall survival but only in univariate analysis. This was further validated by transformed cases which had *IGHV* identity closer to germline than those that did not. Although the CLL cut-offs of mutated (<98% GI) and unmutated (≥98% GI) *IGHV* genes were somewhat arbitrary, our work does provide evidence of two distinct SMZL subtypes partly discriminated on the basis of *IGHV* mutational status. Cases with mutated *IGHV* genes were associated with *MYD88* mutations, and both these markers were mutually exclusive of 7q deletions. These observations point to a possible utility of SHM in stratifying SMZL patients

9.2 Future directions

The work performed in this thesis has provided an in-depth look at the most frequently mutated genes, their somatic interactions and potential clinical utility across a targeted panel of genes. The current prognostic model for SMZL patients published by the European Society for Medical Oncology[40] does not include molecular biomarkers but does acknowledge the potential utility of markers such as *IGHV* mutational status, *NOTCH2* and *KLF2* mutations, *TP53* abnormalities and aberrant promoter methylation. Furthermore, the guidelines state that the model can aid in the discussion of treatment but is not validated as a tool to indicate treatment. We provide evidence of two potential genomic subgroups, one group characterised by 7q deletions, *KLF2* and *NOTCH2* mutations and *IGHV1-2*04* usage and a second group characterised by *MYD88* mutations and mutated *IGHV* genes, with the latter group associated with better clinical outcomes. To improve

the current prognostic model and potentially include biomarkers implicated here, more clinical studies are required with treatment naïve patients and uniform treatment strategies with long follow up times and rich clinical data. This will require international collaboration to create high-quality genomic datasets ideally with matched germline tissue.

Methylation arrays like the ones used in this project are not optimised for precise identification of CNA breakpoints or structural variation (SVs). Experiments that use high-density arrays or WGS would provide a more granular view of not only the CNAs and SVs but would also give insights into the types of genomic complexity that define these patients. Genomic complexity points to a dysregulation of cell cycle control or DNA damage response pathways associated with poor survival in SMZL[47] and other lymphomas[282–284], including in the context of targeted therapies[285,286]. It would be interesting to compare targeted pathways in complex versus simple genomes and identify if there are any established driver behind genomic complexity, such as *TP53* or *ATM* dysfunction, or if there are new drivers driving complexity in SMZL.

This project did not assess aberrant promoter methylation in SMZL; however, the genomic results presented herein are to be integrated as part of a larger international study, focussing on the DNA methylation profiles in SMZL. This will build on the work developed by Arribas et al.[247] where a subgroup of patients that showed high genome-wide promoter methylation (High-M) were associated with *NOTCH2* mutations, 7q deletions, *IGHV1-2* usage and inferior survival. This high-M group exhibited hypomethylation and high gene expression of genes involved in B-cell activation, NF-κB signalling and those encoding for components of the polycomb repressor complex 2 (PRC2). The aberrant methylation seems to play a role in the pathogenesis of the disease and the authors also provided evidence that treatment with demethylating agents could be useful in the treatment of High-M groups. The epigenetic study being carried out by a wider local collaborative team will aim to characterise the SMZL epigenome and compare it to other mature B-cell tumours and normal B-cells. This will help further define patient subgroups to improve differential diagnosis and provide insights into the cell or cells of origin.

Another ongoing local project that will aid in the understanding of SMZL is the whole genome sequencing (WGS) of matched SMZL tumour samples. This will be only the second WGS study on SMZL and the first to use matched tissue. As stated earlier, the targeted approach limits our ability to discover new mutated genes and other potential genomic abnormalities. The current lack of matched-germline whole genome approaches precludes a meaningful analysis, not only of the somatic variation, but of the underpinning mutational signatures, the non-coding mutational landscape, the structural alterations, and regions of chromothripsis and kataegis. WGS has the

potential to identify a wide array of genomic alterations which could be used in prognostic models and to further understand the molecular pathogenesis of SMZL. Determining mutational signatures within the disease could also prove useful in trying to determine the cell origin and in helping to further delineate the two genomic subgroups identified within this work.

In conclusion, the work outlined herein has improved our understanding of the biological basis of SMZL and the potential clinical implications with key findings being: 1) Genes *KMT2D* and *CCND3* were found mutated in a much higher number of cases than was expected; 2) *KLF2* and *CCND3* harbour mutation hotspots which require functional validation but are predicted to affect protein function; 3) Evidence of SHM was found in the majority of cases, (only 8% showed no evidence of SHM); 4) Deletions of 7q were associated to *IGHV1-2*04* usage, *KLF2* and *NOTCH2* mutations, short telomeres, and low levels of SHM; 5) Identification of two potential genomic subgroups, one group characterised by 7q deletions, *KLF2* and *NOTCH2* mutations and *IGHV1-2*04* usage and a second group characterised by *MYD88* mutations and mutated *IGHV* genes; 6) Validation of the prognostic impact of *TP53* aberrations, *NOTCH2* mutations and age of diagnosis and; 7) Identification of telomere length and gains of 3q and 8q as new potential prognostic factors. Most of these finding have a clear clinical utility, but ultimately there is still much work to do be done and a comprehensive multi-omic approach (genomics, transcriptomics, and epigenomics) will likely be the most successful in identifying patients for precision medicine (targeted treatments and clinical trials). Furthermore, the research community needs to continue to collaborate in collating all the resources and knowledge available to ultimately translate what we know about the molecular mechanisms on SMZL for direct patient benefit.

Supplementary materials

Supplementary tables

Supplementary Table 1. Bioinformatics approaches of studies included in the database. Studies are listed in chronological order.

Study	Bioinformatics	Data Quality
Rossi et al. (2011) PMID:21881048 *Extension / confirmation*	**1.** Database search to exclude SNPs. **2.** Variants present in matched germline excluded. **3.** Synonymous variants excluded. **4.** Recurrent variants excluded unless somatic origin confirmed. **5.** hg19 assembly	NA
Rossi et al. (2012) PMID: 22891273 *Discovery* *Extension / confirmation*	**1**. Aligned using BWA (v.0.5.0) with hg19 assembly. **2.** Confirmed somatic nonsynonymous tested in-silico using PolyPhen-2. **3.** Mutated genes verified for their presence in COSMIC and Cancer Gene Census database. **1.** Automated and/or manual curation. **2**. Mutation Surveyor Version 3.97. **3.** Synonymous mutations, previously reported germline polymorphisms and changed in matched normal removed. **4.** Database search to exclude SNPs.	* Mean depth 111x * At least 83% of target covered at 30x * 99% Phred score ≥ 20 * 96% Phred score ≥ 30
Yan et al. (2012) PMID: 22102703 *Extension / confirmation*	**1.** Database search to exclude SNPs. **2.** hg19 assembly	NA
Kiel et al. (2012) PMID:22891276 *Discovery* *Extension / confirmation*	**1.** Mapping and variants calling using CGAtools v1.3.0. **2.** Downstream analysis with custom PERL processing routines. **3.** Database search to exclude SNPs. **4.** hg19 assembly	* 97.6% genome coverage * 96.4% fully called exome coverage * median genomic sequencing depth > 80x in all samples
Parry et al. (2013) PMID:24349473 *Discovery*	**1.** PCR duplicates and reads mapping to multiple locations removed. **2.** Variants called using Varscan 2.3.3 using 'somaticFilter' command to remove false positives. **3.** Annotated using Annovar software v2012Jun21.	* Minimum depth 4 * Minimum VAF 0.1 * Mean depth 69x * Average of 82.2% of target sequences captured at 20x

	4. Database search to exclude SNPs. **5**. hg19assembly	
Martinez et al. (2014) PMID:24296945 *Discovery*	**1.** Reads trimmed until base quality was > 10. **2.** Mapped with Genome Multitool and BFAST. **3.** PCR duplicates and reads mapping to multiple locations removed. **4.** SAMtools and RAMSES used to call variants. Annovar used for annotation and snpEff for effect prediction. **5.** Database search to exclude SNPs. **6.** hg19 assembly	* Mean coverage 128.62x * Average of 90% of targeted bases captured at 15X
Parry et al. (2015) PMID:25779943 *Extension / confirmation*	**1.** Independent analysis of variant by two different entities **2.** hg19 assembly	* Mean depth 297x * Average of 85% of target bases captured at > 50x
Piva et al. (2015) PMID: 25283840 *Extension / confirmation*	**1.** Automated and/or manual curation. **2.** Compared to corresponding germline using the Mutation Surveyor Version 3.97 software package (SoftGenetics). **3.** Synonymous mutations, previously reported germline polymorphisms and changed in matched normal removed. **4.** hg19 assembly	NA
Peveling-Oberhag, et al. (2015) PMID:26498442 *Discovery* *Extension / confirmation*	**1.** Mapping and variant calling using Bioscope v1.2. **2.** Annotation using NGS-SNP using EMSEMBLE v61 database **3.** Intronic, UTR and synonymous mutations removed. **4.** Low quality SNPs filtered out (Bioscope mean quality). **5.** Manual review by two independent observers using integrative genomics viewer. **6.** hg19 assembly	* Minimum coverage 20x * Minimum VAF 0.1
Clipson et al. (2015) PMID:25428260 *Discovery* *Extension / confirmation*	**1.** Aligned using BWA algorithm (hg19 assembly). **2.** CaVEMan, Pindel and in-house algorithm used in variant calling. **3.** Post-processing filters applied to remove poor quality variants. **4.** Database search to exclude SNPs.	NA

Spina et al. (2016) PMID:27335277 *Comparison*	**1.** Aligned using BWA v.0.6.2 software (hg19 assembly) **2.** Variants calling using GATK **3.** Non-synonymous variants considered if: • Absent from dbSNP • Absent in normal DNA • represented in at least 2 forward and reverse reads • VAF > 0.1	* Mean depth 369x * Average of 92% of target bases captured at > 30x
Campos-Martin et al. (2017) PMID:28522570 *Extension / confirmation*	NA	NA
Jallades et al. (2017) PMID:28751561 *Comparison*	**1.** Image analysis and variant calling using CASAVA1.8.2. **2.** Annotation using IntegraGen in house pipeline. **3.** Database search to exclude SNPs. **4.** hg19 assembly	NA
Pillonel et al. (2018) PMID:29556019 *Comparison*	**1.** Variant calling using plug-in v5.2 IonTorrent software suite. **2.** Annotation using IonReporter Software. **3**. Filters applied: • phred-based quality > 50 • Strand bias ≤ 0.75 • Reads supporting variants • ≥ 10 • VAF > 0.05 • Include only synonymous • exonic and splicing **4.** Database search to exclude SNPs. **5.** Manual review using integrative genomics viewer **6**. GRCh37 assembly	* Mean depth 1400x

Supplementary Table 2. Assessed cases in 20 most mutated genes in SMZL database. Columns in green represent the number of cases across the 14 studies in which a specific gene (column 1) was assessed. The columns in orange, tally the total number of assessed cased and reports the number of mutations recorded for each gene to calculate the mutational frequency (number of mutated cases/ total number of assessed cases).

Genes	Rossi *et.al*, 2011	Rossi et.al, 2012	Yan et.al, 2012	Kiel et.al, 2012*	Parry et.al, 2013	Martinez et.al, 2014	Parry et.al, 2015	Pevelling et.al, 2015	Piva et.al, 2015	Clipson et.al, 2015	Spina et.al, 2016	Campos et.al, 2017	Jallades et.al, 2017	Pillonel et.al, 2018	# cases	# mut	Mut. Freq.
NOTCH2	0	117	0	93	7	31	175	2	0	3	32	84	46	12	602	123	0.20
KLF2	0	8	0	0	7	15	175	2	96	112	32	84	46	12	589	121	0.21
TNFAIP3	101	117	57	0	7	15	175	2	0	3	32	0	46	12	567	75	0.13
TP53	0	117	0	0	7	15	175	2	0	3	32	0	46	12	409	60	0.15
MYD88	101	117	57	0	7	15	175	2	0	3	32	0	46	12	567	43	0.08
TRAF3	101	117	0	0	7	15	175	2	0	3	-	0	46	12	478	36	0.08
KMT2D	0	117	0	0	7	15	175	2	0	3	32	0	0	12	363	33	0.09
IGLL5	0	8	0	0	7	15	175	2	0	3	-	0	0	12	222	31	0.14
SPEN	0	117	0	0	7	15	175	2	0	3	32	0	46	12	409	31	0.08
CARD11	101	117	57	0	7	15	175	2	0	3	32	0	0	12	521	27	0.05
NOTCH1	0	117	0	0	7	15	175	2	0	3	32	0	46	12	409	24	0.06
CCND3	0	8	0	0	7	15	175	2	0	3	-	0	46	12	268	18	0.07
IKBKB	101	117	0	0	7	15	175	2	0	3	32	0	46	12	510	18	0.04
BIRC3	101	8	0	0	7	15	175	2	0	3	-	0	46	12	369	17	0.05
ATM	0	8	0	0	7	15	175	2	0	3	-	0	0	12	222	15	0.07
CACNA1H	0	8	0	0	7	15	175	2	0	3	-	0	0	0	210	15	0.07
CREBBP	0	8	0	0	7	15	175	2	0	3	32	0	46	12	300	15	0.05
DNAH10	0	8	0	0	7	15	175	2	0	3	-	0	0	0	210	15	0.07
FAT3	0	8	0	0	7	15	175	2	0	3	-	0	0	0	210	15	0.07
FAT4	0	8	0	0	7	15	175	2	0	3	-	0	0	12	222	15	0.07
PRKDC	0	8	0	0	7	15	175	2	0	3	-	0	0	0	210	15	0.07

Supplementary Table 3. Summary of HaploPlex HS kits used for library preparation. Genes targeted by HaloPlex HS Enrichment kits, location (interval) in the genome (hg19) and the number of regions or amplicons that were used for each gene or region. The genes highlighted in pink are those that differed between kits.

Target gene	SMZL kit-17005-1495007299 (Kit 1)		B-cell kit v17005-1521455344 (kit 2)	
	Interval	Regions	Interval	Regions
ARID1A	chr1:27022512-27108611	21	chr1:27022512-27108611	21
ATM	chr11:108093200-108239839	64	chr11:108093200-108239839	64
BCOR	-	-	chrX:39909057-40036592	19
BCL10	chr1:85731449-85743781	4	-	-
BIRC3	chr11:102188171-102210145	12	chr11:102188171-102210145	12
BRAF_EX15	chr7:140453075-140453193	1	chr7:140453075-140453193	1
CARD11	chr7:2945700-3083589	26	chr7:2945700-3083589	26
CCND3	chr6:41902661-42018105	11	chr6:41902661-42018105	11
CD79A	chr19:42381180-42385449	4	chr19:42381180-42385449	4
CD79B	chr17:62006088-62009724	6	chr17:62006088-62009724	6
CDH23	chr10:73156681-73575714	75	chr10:73156681-73575714	75
CHD2	chr15:93426516-93571247	45	chr15:93426516-93571247	45
CREBBP	chr16:3775045-3930737	31	chr16:3775045-3930737	31
CXCR4	chr2:136871909-136875745	3	chr2:136871909-136875745	3
DCHS1	chr11:6642544-6677095	21	-	-
DDX3X	chrX:41192551-41223735	21	chrX:41192551-41223735	21
EGR2	chr10:64571746-64679670	10	chr10:64571746-64679670	10
EZH2EX12	chr7:148513776-148513870	1	chr7:148513776-148513870	1
EZH2EX16	chr7:148508717-148508812	1	chr7:148508717-148508812	1
EZH2EX18	chr7:148506402-148506482	1	chr7:148506402-148506482	1
FBXW7	chr4:153242400-153457263	15	chr4:153242400-153457263	15
FLNC	chr7:128470421-128499338	47	chr7:128470421-128499338	47
ID3	chr1:23884399-23886295	3	chr1:23884399-23886295	3
IDH2EX4	chr15:90631819-90631979	1	chr15:90631819-90631979	1
IGLL5	chr22:23229949-23238297	3	-	-
JAK3	chr19:17935579-17958890	24	chr19:17935579-17958890	24
KDM2B	chr12:121866890-122018930	26	chr12:121866890-122018930	26
KLF2	chr19:16435618-16438695	3	chr19:16435618-16438695	3
KMT2D	chr12:49412748-49453567	55	chr12:49412748-49453567	55
KRAS	chr12:25357713-25403880	7	chr12:25357713-25403880	7

Target gene	SMZL kit-17005-1495007299 (Kit 1)		B-cell kit v17005-1521455344 (kit 2)	
	Interval	Regions	Interval	Regions
MAP2K1	chr15:66679145-66784660	12	chr15:66679145-66784660	12
MAP3K14	chr17:43340475-43394440	16	chr17:43340475-43394440	16
MAP3K6	chr1:27681659-27693393	27	-	-
MED12	-	-	chrX:70338395-70362314	45
MYD88	chr3:38179959-38184523	4	chr3:38179959-38184523	4
NFKBIE	chr6:44225893-44233535	6	chr6:44225893-44233535	6
NOTCH1EX13	chr9:139408961-139409154	1	-	-
NOTCH1EX26	chr9:139399125-139399556	1	chr9:139399125-139399556	1
NOTCH1EX27	chr9:139397634-139397782	1	chr9:139397634-139397782	1
NOTCH1EX28	chr9:139396724-139396940	1	chr9:139396724-139396940	1
NOTCH1EX3	chr9:139418168-139418431	1	-	-
NOTCH1EX31	chr9:139395003-139395299	1	-	-
NOTCH1EX34_3'UTR	chr9:139388896-139392010	1	chr9:139388896-139392010	1
NOTCH1EX4	chr9:139417301-139417640	1	-	-
NOTCH2	chr1:120454166-120612327	37	chr1:120454166-120612327	37
NOTCH2EX34+UTR	chr1:120457806-120458167	1	chr1:120454176-120459317	1
NRAS	chr1:115247075-115259525	7	chr1:115247075-115259525	7
P53	chr17:7579838-7590856	2	chr17:7579838-7590856	2
PAX5NONCODING	chr9:37368940-37373593	1	chr9:37368940-37373593	1
POT1	chr7:124462430-124570047	23	chr7:124462430-124570047	23
PRKDC	chr8:48685659-48872753	86	chr8:48685659-48872753	86
PTPRD	chr9:8314236-10612733	53	chr9:8314236-10612733	53
RHOA	chr3:49396559-49450441	10	chr3:49396559-49450441	10
RPS15	chr19:1438348-1440593	1	chr19:1438348-1440593	1
SAMHD1	chr20:35518622-35580256	17	chr20:35518622-35580256	17
SETD1B	chr12:122242075-122270572	17	-	-
SETD2	chr3:47057888-47205477	25	chr3:47057888-47205477	25
SF3B1EX14	chr2:198267280-198267550	1	chr2:198267280-198267550	1
SF3B1EX15	chr2:198266709-198266854	1	chr2:198266709-198266854	1
SF3B1EX16	chr2:198266466-198266612	1	chr2:198266466-198266612	1
SF3B1EX18	chr2:198265439-198265660	1	chr2:198265439-198265660	1
SPEN	chr1:16174349-16266965	16	chr1:16174349-16266965	16

Target gene	SMZL kit-17005-1495007299 (Kit 1)		B-cell kit v17005-1521455344 (kit 2)	
	Interval	Regions	Interval	Regions
STAT3EX21	chr17:40474299-40474512	1	chr17:40474299-40474512	1
TCF3	chr19:1609279-1652614	19	chr19:1609279-1652614	19
TET2	chr4:106067022-106200983	13	chr4:106067022-106200983	13
TNFAIP3	chr6:138188315-138204461	9	chr6:138188315-138204461	9
TNFRSF14	chr1:2487068-2497071	6	chr1:2487068-2497071	6
TP53	chr17:7565087-7590878	14	chr17:7565087-7590878	14
TRAF3	chr14:103243803-103377847	15	chr14:103243803-103377847	15
U2AF1	chr21:44513055-44527707	4	-	-
XPO1EX15	chr2:61719460-61719616	1	chr2:61719460-61719616	1
XPO1EX16	chr2:61719170-61719333	1	chr2:61719170-61719333	1

Supplementary Table 4. List of annotations added to variants.

Header	Definition
Chr	Chromosome number (hg38)
Start	Position where variant starts (hg38)
End	Position where variant ends (hg38)
Ref	Reference allele
Alt	Alternative or mutant allele
Func.refGene	Regions (e.g., exonic, intronic, non-coding RNA)) that one variant hits according to RefGene
Gene.refGene	Gene name associated with variant according to RefGene
GeneDetail.refGene	Gene name, the transcript identifier and the sequence change in the corresponding transcript according to RefGene
ExonicFunc.refGene	Exonic variant function, e.g., nonsynonymous, synonymous, frameshift insertion according to RefGene
AAChange.refGene	Amino acid change according to RefGene e.g SAMD11:NM_152486:exon10:c.T1027C:p.W343R stands for gene name, Known RefSeq accession, region, cDNA level change, protein level change.
avsnp144	dbSNP144 with allelic splitting and left-normalization
gnomAD_genome_ALL	Genome Aggregation Data in **ALL** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_AFR	Genome Aggregation Data in **AFRICAN** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_AMR	Genome Aggregation Data in **AMERICAN** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_ASJ	Genome Aggregation Data in **ASHKENAZI JEW** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_EAS	Genome Aggregation Data in **EAST ASIAN** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_FIN	Genome Aggregation Data in **FINNISH** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_NFE	Genome Aggregation Data in **NON FINNISH** populations. gnomAD genome collection (v2.0.1)
gnomAD_genome_OTH	Genome Aggregation Data in **OTHER** populations. gnomAD genome collection (v2.0.1)
1000g2015aug_all	Alternative allele frequency data in 1000 Genomes Project for autosomes in **ALL** populations. Based on 201508 collection v5b (based on 201305 alignment)
1000g2015aug_afr	Alternative allele frequency data in 1000 Genomes Project for autosomes in **AFRICAN** population. Based on 201508 collection v5b (based on 201305 alignment)
1000g2015aug_amr	Alternative allele frequency data in 1000 Genomes Project for autosomes in **AD MIXED AMERICAN** population. Based on 201508 collection v5b (based on 201305 alignment)
1000g2015aug_sas	Alternative allele frequency data in 1000 Genomes Project for autosomes in **SOUTH ASIAN** population. Based on 201508 collection v5b (based on 201305 alignment)
1000g2015aug_eur	Alternative allele frequency data in 1000 Genomes Project for autosomes in **EUROPEAN** population. Based on 201508 collection v5b (based on 201305 alignment)

Header	Definition
1000g2015aug_eas	Alternative allele frequency data in 1000 Genomes Project for autosomes in **EAST ASIAN** population. Based on 201508 collection v5b (based on 201305 alignment)
esp6500siv2_all	Alternative allele frequency in **ALL** subjects in the NHLBI-ESP project with 6500 exomes, including the indel calls and the chrY calls
esp6500siv2_ea	Alternative allele frequency in **EROPEAN AMERICAN** subjects in the NHLBI-ESP project with 6500 exomes, including the indel calls and the chrY calls
ExAC_ALL	Exome Aggregation Consortium Data. v0.3 nonTCGA data **ALL** individuals. Version 0.3. Left normalization done.
ExAC_AFR	Exome Aggregation Consortium Data. v0.3 nonTCGA data **AFRICAN** individuals. Version 0.3. Left normalization done.
ExAC_AMR	Exome Aggregation Consortium Data. v0.3 nonTCGA data **AMERICAN** individuals. version 0.3. Left normalization done.
ExAC_EAS	Exome Aggregation Consortium Data. v0.3 nonTCGA data **EASTERN ASIAN** individuals. version 0.3. Left normalization done.
ExAC_FIN	Exome Aggregation Consortium Data. v0.3 nonTCGA data **FINNISH** individuals. version 0.3. Left normalization done.
ExAC_NFE	Exome Aggregation Consortium Data. v0.3 nonTCGA data **NON-FINNISH EUROPEAN** individuals. version 0.3. Left normalization done.
ExAC_OTH	Exome Aggregation Consortium Data. v0.3 nonTCGA data **OTHER** individuals. version 0.3. Left normalization done.
ExAC_SAS	Exome Aggregation Consortium Data. v0.3 nonTCGA data **SOUTH ASIAN** individuals. version 0.3. Left normalization done.
esp6500siv2_aa	Alternative allele frequency in **AFRICAN AMERICAN** subjects in the NHLBI-ESP project with 6500 exomes, including the indel calls and the chrY calls
cosmic70	ID in the Catalogue of Somatic Mutations in Cancer. COSMIC database version 68 on WGS data.
CLINSIG	Clinical significance : 0 - unknown, 1 - untested, 2 - non-pathogenic, 3 - probable-non-pathogenic, 4 - probable-pathogenic, 5 - pathogenic, 6 - drug-response, 7 - histocompatibility, 255 - other
CLNDBN	Variant disease name
CLNACC	Variant accession and versions
CLNDSDB	Variant disease database name
CLNDSDBID	Variant disease database ID
HRC_AF	Haplotype reference consortium: Non-reference allele frequency across all HRC.r1 cohorts (AC/AN)
HRC_AC	Haplotype reference consortium: Non-reference allele count across all HRC.r1 cohorts
HRC_AN	Haplotype reference consortium: Non-reference allele number across all HRC.r1 cohorts
HRC_non1000G_AF	Haplotype reference consortium: Non-reference allele frequency across all HRC.r1 cohorts excluding 1000G samples (AC_EXCLUDING_1000G/AN_EXCLUDING_1000G)
HRC_non1000G_AC	Haplotype reference consortium: Non-reference allele count across all HRC.r1 cohorts excluding 1000G samples
HRC_non1000G_AN	Haplotype reference consortium: Non-reference allele number across all HRC.r1 cohorts excluding 1000G samples
Kaviar_AF	170 million Known VARiants from 13K genomes and 64K exomes in 34 projects: Non-reference allele frequency across all cohorts (AC/AN)
Kaviar_AC	170 million Known VARiants from 13K genomes and 64K exomes in 34 projects: Non-reference allele count across all cohorts

Header	Definition
Kaviar_AN	170 million Known VARiants from 13K genomes and 64K exomes in 34 projects: Non-reference allele number across all cohorts
nci60	Human tumor cell line panel exome sequencing allele frequency data
SIFT_score	Score predicting whether an amino acid substitution affects protein function. Score ranges from 0.0 (deleterious) to 1.0 (tolerated). Scores from dbNSFP version 3.0a.
SIFT_pred	***D*****:** Deleterious (sift<=0.05); ***T*****:** tolerated (sift>0.05)
Polyphen2_HDIV_score	Score predicting possible impact of an amino acid substitution on the structure and function of a human protein. Compiled from all damaging alleles with known effects on the molecular function causing human Mendelian diseases. This score represents the probability that a substitution is damaging. score ranges from 0.0 (tolerated) to 1.0 (deleterious). Scores from dbNSFP version 3.0a.
Polyphen2_HDIV_pred	***D*****:** Probably damaging (pp2_hdiv≥0.957), ***P:*** Possibly damaging (0.453≤pp2_hdiv≤0.956), ***B:*** Benign (pp2_hdiv≤0.452)
Polyphen2_HVAR_score	Score predicting possible impact of an amino acid substitution on the structure and function of a human protein. consisted of all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging. This score represents the probability that a substitution is damaging. score ranges from 0.0 (tolerated) to 1.0 (deleterious). Scores from dbNSFP version 3.0a.
Polyphen2_HVAR_pred	***D:*** Probably damaging (pp2_hdiv≥0.957), ***P:*** Possibly damaging (0.453≤pp2_hdiv≤0.956), ***B:*** Benign (pp2_hdiv≤0.452)
LRT_score	Likelihood ratio test for significantly conserved amino acid positions within the human proteome. Scores from dbNSFP version 3.0a.
LRT_pred	***D:*** Deleterious; ***N:*** Neutral; ***U:*** Unknown Lower scores are more deleterious
MutationTaster_score	MutationTaster employs a Bayes classifier to eventually predict the disease potential of an alteration. The Bayes classifier is fed with the outcome of all tests and the features of the alterations and calculates probabilities for the alteration to be either a disease mutation or a harmless polymorphism. For this prediction, the frequencies of all single features for known disease mutations/polymorphisms were studied in a large training set composed of >390,000 known disease mutations from HGMD Professional and >6,800,000 harmless SNPs and Indel polymorphisms from the 1000 Genomes Project (TGP). Scores from dbNSFP version 3.0a.
MutationTaster_pred	***A:*** (""disease_causing_automatic""); ***D:*** (""disease_causing""); ***N:*** (""polymorphism [probably harmless]""); ***P:*** (""polymorphism_automatic[known to be harmless]" higher values are more deleterious"
MutationAssessor_score	This server predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense

Header	Definition
	polymorphisms. The functional impact is assessed based on evolutionary conservation of the affected amino acid in protein homologs. The method has been validated on a large set (60k) of disease associated (OMIM) and polymorphic variants. Scores from dbNSFP version 3.0a.
MutationAssessor_pred	***H:*** high; ***M:*** medium; ***L:*** low; ***N:*** neutral. H/M means functional and L/N means non-functional higher values are more deleterious
FATHMM_score	Score predicting the functional effects of protein missense mutations by combining sequence conservation within hidden Markov models (HMMs), representing the alignment of homologous sequences and conserved protein domains, with "pathogenicity weights", representing the overall tolerance of the protein/domain to mutations. Scores from dbNSFP version 3.0a.
FATHMM_pred	***D:*** Deleterious; ***T:*** Tolerated; lower values are more deleterious
PROVEAN_score	Score predicting whether an amino acid substitution or indel has an impact on the biological function of a protein.
PROVEAN_pred	***D:*** Deleterious; ***N:*** Neutral higher values are more deleterious
VEST3_score	Machine learning method that predicts the functional significance of missense mutations based on the probability that they are pathogenic. Higher values are more deleterious. Scores from dbNSFP version 3.0a.
CADD_raw	Scores the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome. Raw scores come straight from the model, and are interpretable as the extent to which the annotation profile for a given variant suggests that the variant is likely to be "observed" (negative values) vs "simulated" (positive values). These values have no absolute unit of meaning and are incomparable across distinct annotation combinations, training sets, or model parameters. However, raw values do have relative meaning, with higher values indicating that a variant is more likely to be simulated (or "not observed") and therefore more likely to have deleterious effects. Scores from dbNSFP version 3.0a.
CADD_phred	Normalised CADD scores. For example, reference genome single nucleotide variants at the 10th-% of CADD scores are assigned to CADD-10, top 1% to CADD-20, top 0.1% to CADD-30, etc. The results of this transformation are the "scaled" CADD scores. Scores from dbNSFP version 3.0a.
DANN_score	Deleterious Annotation of genetic variants using Neural Networks. Higher values are more deleterious. Scores from dbNSFP version 3.0a.
fathmm-MKL_coding_score	Predicting the effects of both coding and non-coding variants using nucleotide-based HMMs
fathmm-MKL_coding_pred	**D:** Deleterious ≥ 0.5 **T:** Tolerated < 0.5
MetaSVM_score	Score to integrate nine deleteriousness prediction scores and maximum minor allele frequency for more accurate and comprehensive evaluation

Header	Definition
	of deleteriousness of missense mutations using Support Vector Machine (SVM). Scores from dbNSFP version 3.0a.
MetaSVM_pred	**D:** Deleterious; **T:** Tolerated; higher scores are more deleterious
MetaLR_score	Score to integrate nine deleteriousness prediction scores and maximum minor allele frequency for more accurate and comprehensive evaluation of deleteriousness of missense mutations using Logistic Regression (LR). Scores from dbNSFP version 3.0a.
MetaLR_pred	**D:** Deleterious; **T:** Tolerated; higher scores are more deleterious
integrated_fitCons_score	Fitness consequences of functional annotation, integrates functional assays (such as ChIP-Seq) with selective pressure inferred using the INSIGHT method. The result is a score ρ in the range [0.0-1.0] that indicates the fraction of genomic positions evidencing a particular pattern (or "fingerprint") of functional assay results, that are under selective pressure. Higher scores are more deleterious. Scores from dbNSFP version 3.0a.
GERP++_RS	Identifies constrained elements in multiple alignments by quantifying substitution deficits. These deficits represent substitutions that would have occurred if the element were neutral DNA, but did not occur because the element has been under functional constraint. Thus, positive scores represent a substitution deficit (which would be expected for sites under selective constraint), while negative scores represent a substitution surplus. Higher scores are more deleterious. Scores from dbNSFP version 3.0a.
phyloP7way_vertebrate	Phylogentic p-values calculated from a LRT, score-based test, GERP test Use 7 species. Higher scores are more deleterious.
phyloP20way_mammalian	Phylogenetic hidden Markov model (phylo-HMM) Use 20 species. Higher scores are more deleterious.
phastCons7way_vertebrate	Identifies evolutionarily conserved elements in a multiple alignment, given a phylogenetic tree. A phylogenetic hidden Markov model (phylo-HMM) Use 7 species. Higher scores are more deleterious
phastCons20way_mammalian	Identifies evolutionarily conserved elements in a multiple alignment, given a phylogenetic tree. It is based on phylogenetic hidden Markov model (phylo-HMM). Higher scores are more deleterious.
SiPhy_29way_logOdds	Probablistic framework, HMM Use 29 species. Higher scores are more deleterious. Scores from dbNSFP version 3.0a.
Interpro_domain	One of a collection of amino acid sequences of identifiable features in known proteins that can be compared to unknown protein sequences. See InterPro at EMBL/EBI
dbscSNV_ADA_SCORE	These are ensemble scores, derived from the outputs of several machine learning algorithms. Both are scaled from 0 and 1, and higher values indicate a greater probability that the variant will alter the splicing of the gene. The developers suggest using 0.6 as a threshold value for dichotomous effects.
dbscSNV_RF_SCORE	
AAchange	Amino acid change
FuentesFalsePositive	Genes found in Fuentes false positive list. (Detecting false positive signals in exome sequencing, 2013)

Supplementary Table 5. Detailed batch information. The mean target coverage was calculated using the DepthOfCoverage tool. Number of variants called refers to the raw number of variants in the annotated files before any filtering.

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
1	2_S1	CBL-MZ	NO	HS kit 1	568	370.92	90.90
1	3_S2	CBL-MZ	NO	HS kit 1	622	426.27	91.40
1	4_S3	SMZL	YES	HS kit 1	579	517.63	91.60
1	6_S4	CBL-MZ	NO	HS kit 1	650	365.17	91.50
1	7_S5	CBL-MZ	NO	HS kit 1	622	488.56	91.80
1	8_S6	CBL-MZ	NO	HS kit 1	650	445.98	91.50
1	9_S7	SMZL	YES	HS kit 1	569	293.72	89.60
1	10_S8	SMZL	YES	HS kit 1	638	429.86	91.20
1	12_S9	SMZL	YES	HS kit 1	462	378.68	83.80
1	15_S10	CBL-MZ	NO	HS kit 1	659	413.45	91.50
1	18_S11	CBL-MZ	NO	HS kit 1	574	467.51	91.20
1	21_S12	SMZL	YES	HS kit 1	569	352.30	90.70
1	22_S13	SMZL	YES	HS kit 1	565	296.47	89.60
1	L37_S14	SMZL	YES	HS kit 1	600	228.24	91.00
1	L40_S15	CLL	NO	HS kit 1	556	417.25	90.60
1	4683_S16	SMZL	YES	HS kit 1	605	414.84	91.30
1	9641_S17	SMZL	YES	HS kit 1	618	377.51	91.00
1	11731_S18	SMZL	YES	HS kit 1	637	309.33	90.80
1	12600_S19	SMZL	YES	HS kit 1	553	269.86	90.30
1	12603_S20	SMZL	YES	HS kit 1	579	377.83	91.00
1	L009_02	SMZL	YES	HS kit 1	648	322.51	91.00
1	L011_03	SMZL	YES	HS kit 1	536	308.48	91.20
1	L012_04	SMZL	YES	HS kit 1	685	383.03	91.60
1	L017_06	SMZL	YES	HS kit 1	657	327.10	89.30
1	L018_06	SMZL	YES	HS kit 1	615	331.58	91.50
1	L019_06	SMZL	YES	HS kit 1	669	309.04	91.60
1	L022_06	SMZL	YES	HS kit 1	613	149.51	90.60
1	L023_07	SMZL	YES	HS kit 1	648	461.42	92.10
1	L024_07	SMZL	YES	HS kit 1	642	428.08	91.90
1	L025_07	SMZL	YES	HS kit 1	569	106.36	89.60
1	L027_07	SMZL	YES	HS kit 1	557	304.71	91.00
1	L029_07	SMZL	YES	HS kit 1	613	377.40	91.50
1	L031_07	SMZL	YES	HS kit 1	658	453.04	91.30
1	L034_08	SMZL	YES	HS kit 1	642	347.00	91.20
1	L036_08	SMZL	YES	HS kit 1	557	401.24	91.40
1	L037_08	SMZL	YES	HS kit 1	608	440.06	91.10
1	L038_08	SMZL	YES	HS kit 1	606	356.72	91.40
1	L043_08	SMZL	YES	HS kit 1	661	372.66	91.70
1	L044_08	SMZL	YES	HS kit 1	607	354.63	91.30
1	L048_09	SMZL	YES	HS kit 1	596	297.62	91.10
1	L049_09_S30	SMZL	NO	HS kit 1	616	237.68	90.70
1	L049_09_S31	SMZL	YES	HS kit 1	629	609.21	91.60

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
1	L050_09	unknown	NO	HS kit 1	623	370.92	91.30
1	L051_09	SMZL	YES	HS kit 1	627	340.72	91.30
1	L052_09	SMZL	YES	HS kit 1	540	260.77	90.40
1	L060_09	SMZL	YES	HS kit 1	636	355.36	91.00
1	L067_10	SMZL	YES	HS kit 1	613	458.00	91.20
1	L069_10	SMZL	YES	HS kit 1	607	342.82	90.80
1	L070_10	SMZL	YES	HS kit 1	636	437.95	91.60
1	L071_10	SMZL	YES	HS kit 1	675	492.06	91.50
1	L075_10	SMZL	YES	HS kit 1	554	524.28	91.50
1	L076_10	SMZL	YES	HS kit 1	749	393.74	91.50
1	L080_11	SMZL	YES	HS kit 1	639	395.29	91.30
1	L082_11	SMZL	YES	HS kit 1	622	197.36	90.70
1	L086_12	SMZL	YES	HS kit 1	626	280.62	90.00
1	L088_12	SMZL	YES	HS kit 1	571	330.56	90.60
1	L094_13	SMZL	YES	HS kit 1	579	321.79	90.60
1	L096_13	SMZL	YES	HS kit 1	527	354.83	90.80
1	L098_13_S59	SMZL	YES	HS kit 1	574	367.20	91.00
1	L098_13_S60	SMZL	NO	HS kit 1	509	95.28	86.20
1	L099_13	SMZL	YES	HS kit 1	685	360.54	91.00
1	L104_14	SMZL	NO	HS kit 1	530	194.16	88.90
2	11_S11	SMZL	NO	HS kit 1	802	345.12	91.50
2	12_S12	CBL-MZ	NO	HS kit 1	888	422.16	91.80
2	19_S19	SMZL	NO	HS kit 1	858	294.73	91.20
2	20_S20	SMZL	NO	HS kit 1	851	255.38	91.60
2	21_S21	SMZL	NO	HS kit 1	871	249.74	91.60
2	22_S22	unknown	NO	HS kit 1	859	337.86	92.20
2	23_S23	unknown	NO	HS kit 1	897	224.14	91.70
2	10_S10	SMZL	NO	HS kit 1	823	376.96	91.90
2	1_S1	SMZL	NO	HS kit 1	813	197.18	89.80
2	2_S2	SMZL	NO	HS kit 1	942	250.87	90.90
2	3_S3	SMZL	NO	HS kit 1	851	316.42	90.90
2	4_S4	SMZL	YES	HS kit 1	897	274.11	91.20
2	5_S5	SMZL	NO	HS kit 1	873	369.56	91.70
2	6_S6	SMZL	NO	HS kit 1	807	222.66	91.00
2	7_S7	SMZL	NO	HS kit 1	792	368.47	92.00
2	8_S8	SMZL	NO	HS kit 1	919	569.14	92.50
2	9_S9	SMZL	NO	HS kit 1	768	379.35	91.90
2	48_S48	SMZL	NO	HS kit 1	1042	286.18	92.00
2	24_S24	SMZL	NO	HS kit 1	927	243.82	91.20
2	25_S25	possible SMZL	NO	HS kit 1	905	294.63	91.60
2	26_S26	not SMZL	NO	HS kit 1	854	289.03	91.40
2	27_S27	SMZL	YES	HS kit 1	730	43.14	82.40
2	28_S28	SMZL	NO	HS kit 1	929	127.43	86.80
2	29_S29	SMZL	YES	HS kit 1	893	279.75	91.90

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
2	30_S30	SMZL	NO	HS kit 1	898	310.74	91.30
2	31_S31	SMZL	YES	HS kit 1	582	132.00	41.70
2	49_S49	SMZL	NO	HS kit 1	790	219.72	91.10
2	50_S50	SMZL	NO	HS kit 1	847	318.03	91.70
2	32_S32	SMZL	YES	HS kit 1	854	124.03	90.20
2	33_S33	SMZL	YES	HS kit 1	703	82.49	41.70
2	34_S34	SMZL	YES	HS kit 1	868	139.88	90.40
2	35_S35	SMZL	YES	HS kit 1	992	72.62	88.60
2	36_S36	SMZL	YES	HS kit 1	960	293.43	92.00
2	37_S37	SMZL	NO	HS kit 1	884	280.82	91.80
2	38_S38	SMZL	NO	HS kit 1	945	531.59	92.20
2	39_S39	SMZL	YES	HS kit 1	828	318.58	91.70
2	51_S51	SMZL	NO	HS kit 1	862	411.41	92.00
2	40_S40	SMZL	YES	HS kit 1	842	167.98	90.60
2	52_S52	SMZL	NO	HS kit 1	853	126.39	89.70
2	53_S53	SMZL	NO	HS kit 1	839	208.65	91.30
2	41_S41	SMZL	NO	HS kit 1	882	259.03	91.30
2	42_S42	CBL-MZ	NO	HS kit 1	914	286.56	91.70
2	43_S43	CBL-MZ	NO	HS kit 1	931	238.01	90.80
2	44_S44	CBL-MZ	NO	HS kit 1	881	404.46	92.10
2	45_S45	SMZL	YES	HS kit 1	912	352.00	91.70
2	54_S54	CBL-MZ	NO	HS kit 1	1033	286.25	91.90
2	46_S46	SMZL	NO	HS kit 1	855	602.65	92.40
2	47_S47	SMZL	NO	HS kit 1	926	349.12	91.40
2	13_S13	not SMZL	NO	HS kit 1	940	362.71	92.20
2	14_S14	not SMZL	NO	HS kit 1	839	259.86	91.60
2	15_S15	not SMZL	NO	HS kit 1	890	240.94	91.70
2	16_S16	not SMZL	NO	HS kit 1	892	300.57	91.80
2	17_S17	not SMZL	NO	HS kit 1	821	241.77	91.40
2	18_S18	not SMZL	NO	HS kit 1	852	296.32	91.90
3	774_5	not SMZL	NO	HS kit 1	328	26.04	64.00
3	Kalpadakis_11	not SMZL	NO	HS kit 1	549	493.51	89.60
3	Pangalis_32	unknown	NO	HS kit 1	442	79.13	82.20
3	Pangalis_35	SMZL	NO	HS kit 1	439	137.88	85.60
3	Pangalis_37	unknown	NO	HS kit 1	506	153.59	86.30
3	Kalpadakis_L17	SMZL	YES	HS kit 1	507	188.66	87.10
3	Kalpadakis_L21	SMZL	YES	HS kit 1	382	73.18	80.20
3	Kalpadakis_L34	CBL-MZ	NO	HS kit 1	493	129.71	85.50
3	3065GMG_S35	unknown	NO	HS kit 1	437	56.1	77.80
3	BA1033_S29	unknown	NO	HS kit 1	378	110.94	84.30
3	BG1121_S30	unknown	NO	HS kit 1	466	133.65	85.80
3	RC0438_S21	unknown	NO	HS kit 1	485	117.27	84.40
3	RBH_5274	not SMZL	NO	HS kit 1	486	173.5	86.40
3	RBH_5359	SMZL	YES	HS kit 1	469	112.11	85.60

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
3	P13445_S36	possible SMZL	NO	HS kit 1	543	147.39	86.40
3	P11977_S37	CBL-MZ	NO	HS kit 1	519	117.4	85.20
3	P13925_S47	possible SMZL	NO	HS kit 1	537	156.38	86.40
3	P7176_S41	CBL-MZ	NO	HS kit 1	505	124.67	85.60
3	P11669_S42	SMZL	YES	HS kit 1	422	53.62	77.70
3	PA0329_047	unknown	NO	HS kit 1	459	131.37	84.90
3	013_313	unknown	NO	HS kit 1	488	455.11	88.10
3	005_320	unknown	NO	HS kit 1	511	332.41	88.10
3	3_S24	unknown	NO	HS kit 1	644	3173.37	91.90
3	006_PtN3	unknown	NO	HS kit 1	444	144.83	85.80
3	9_S25	unknown	NO	HS kit 1	613	810.86	90.40
3	011_PGB	unknown	NO	HS kit 1	458	176.66	85.40
3	12_S33	unknown	NO	HS kit 1	578	928.32	91.00
3	016_00036	unknown	NO	HS kit 1	436	100.16	83.00
3	17_S28	unknown	NO	HS kit 1	333	36.94	72.50
3	018_AAA	unknown	NO	HS kit 1	487	242.51	88.40
3	130_S53	not SMZL	NO	HS kit 1	424	69.59	81.00
3	148_S16	not SMZL	NO	HS kit 1	452	89.28	82.60
3	164_S1	not SMZL	NO	HS kit 1	515	380.31	89.20
3	282_S38	not SMZL	NO	HS kit 1	422	101.64	82.80
3	443_S10	not SMZL	NO	HS kit 1	447	117.89	84.70
3	491_S11	SMZL	YES	HS kit 1	444	70.83	80.10
3	515_S2	not SMZL	NO	HS kit 1	400	619.93	83.20
3	587_S17	not SMZL	NO	HS kit 1	426	59.66	78.70
3	589_S39	not SMZL	NO	HS kit 1	385	48.96	76.00
3	606_S12	SMZL	YES	HS kit 1	439	139.04	84.90
3	617_S3	SMZL	YES	HS kit 1	396	32.15	70.00
3	622_S54	SMZL	YES	HS kit 1	278	179.62	69.20
3	623_S55	SMZL	YES	HS kit 1	506	446.52	89.50
3	638_S13	not SMZL	NO	HS kit 1	511	375.36	89.20
3	667_S14	possible SMZL	NO	HS kit 1	459	208.38	86.30
3	673_S4	SMZL	YES	HS kit 1	488	186.34	87.80
3	698_S15	not SMZL	NO	HS kit 1	421	78.6	82.30
3	717_S18	NA	NO	HS kit 1	462	247.47	87.10
3	726_S40	SMZL	YES	HS kit 1	437	70.28	80.80
3	753_S5	SMZL	YES	HS kit 1	541	244.5	88.30
3	773_S56	not SMZL	NO	HS kit 1	424	94.03	83.30
3	779_S19	not SMZL	NO	HS kit 1	348	37.46	70.90
3	785_S6	SMZL	YES	HS kit 1	515	153.3	86.60
3	793_S7	SMZL	YES	HS kit 1	496	236.73	87.30
3	836_S20	SMZL	YES	HS kit 1	353	70.75	77.40
3	868_S8	SMZL	YES	HS kit 1	519	271.65	88.40
3	875_S9	SMZL	YES	HS kit 1	414	84.4	80.60

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
3	1_MUT	SMZL	YES	HS kit 2	455	169.85	90.00
3	2_MUT	SMZL	YES	HS kit 2	362	111.17	87.80
3	3_MUT	SMZL	YES	HS kit 2	331	33.59	68.60
3	4_MUT	SMZL	YES	HS kit 2	538	387.2	93.60
3	5_MUT	SMZL	YES	HS kit 2	507	309.27	93.40
3	6_MUT	SMZL	YES	HS kit 2	411	254.71	89.20
3	7_MUT	SMZL	YES	HS kit 2	339	57.75	80.50
3	8_MUT	SMZL	YES	HS kit 2	483	130.09	88.40
3	9_MUT	SMZL	YES	HS kit 2	402	87.43	84.00
3	10_MUT	SMZL	YES	HS kit 2	422	85.67	86.80
3	11_MUT	SMZL	YES	HS kit 2	282	26.53	60.80
3	12_MUT	SMZL	YES	HS kit 2	416	68.03	84.20
3	13_MUT	SMZL	NO	HS kit 2	433	91.15	87.20
3	14_MUT	SMZL	YES	HS kit 2	444	143.61	90.30
3	15_MUT	SMZL	YES	HS kit 2	411	81.34	86.30
3	16_MUT	SMZL	YES	HS kit 2	336	50.79	80.00
3	17_MUT	SMZL	YES	HS kit 2	530	544.52	95.20
3	18_MUT	SMZL	YES	HS kit 2	417	116.79	90.00
3	19_MUT	SMZL	YES	HS kit 2	468	185.44	91.90
3	23_MUT	SMZL	YES	HS kit 2	443	322.89	89.80
3	25_MUT	SMZL	YES	HS kit 2	425	295.56	86.60
3	27_MUT	SMZL	YES	HS kit 2	258	114.28	61.20
3	30_MUT	SMZL	YES	HS kit 2	368	175.3	75.80
3	31_MUT	SMZL	YES	HS kit 2	391	131.44	77.40
3	33_MUT	SMZL	YES	HS kit 2	299	51.96	60.10
3	35_MUT	SMZL	YES	HS kit 2	247	18.84	51.90
4	11_S11	SMZL	NO	HS kit 1	580	742.53	92.10
4	12_S12	CBL-MZ	NO	HS kit 1	627	876.42	92.20
4	19_S13	SMZL	NO	HS kit 1	595	581.14	91.90
4	20_S14	SMZL	NO	HS kit 1	638	667.67	92.30
4	10_S10	SMZL	YES	HS kit 1	626	832.2	92.60
4	1_S1	SMZL	YES	HS kit 1	556	316.1	90.70
4	2_S2	SMZL	YES	HS kit 1	658	410.37	91.50
4	3_S3	SMZL	YES	HS kit 1	658	764.57	92.00
4	5_S5	SMZL	YES	HS kit 1	614	824.22	92.20
4	6_S6	SMZL	YES	HS kit 1	565	311.44	91.50
4	7_S7	SMZL	YES	HS kit 1	561	769.91	92.70
4	8_S8	SMZL	YES	HS kit 1	664	847.85	92.80
4	9_S9	SMZL	YES	HS kit 1	600	930.01	92.70
4	48_S42	SMZL	YES	HS kit 1	624	359.01	92.30
4	24_S18	SMZL	YES	HS kit 1	630	633.71	92.20
4	25_S19	possible SMZL	NO	HS kit 1	650	891.99	92.20
4	28_S22	SMZL	YES	HS kit 1	532	173.41	88.30
4	29_S23	SMZL	NO	HS kit 1	NA	NA	NA

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
4	30_S24	SMZL	YES	HS kit 1	594	824	92.00
4	49_S43	SMZL	YES	HS kit 1	525	515.9	91.90
4	50_S44	SMZL	YES	HS kit 1	655	655.53	92.10
4	35_S29	SMZL	NO	HS kit 1	NA	NA	NA
4	36_S30	SMZL	NO	HS kit 1	NA	NA	NA
4	37_S31	SMZL	YES	HS kit 1	629	623.61	92.20
4	38_S32	SMZL	YES	HS kit 1	660	1107.12	92.50
4	51_S45	SMZL	YES	HS kit 1	609	876.78	92.50
4	52_S46	SMZL	YES	HS kit 1	595	353.34	91.50
4	53_S47	SMZL	YES	HS kit 1	621	342.36	91.90
4	41_S35	SMZL	YES	HS kit 1	630	650.94	91.90
4	42_S36	CBL-MZ	NO	HS kit 1	622	571.6	92.20
4	43_S37	CBL-MZ	NO	HS kit 1	627	698.13	91.60
4	44_S38	CBL-MZ	NO	HS kit 1	NA	NA	NA
4	54_S48	CBL-MZ	NO	HS kit 1	670	568.84	92.40
4	46_S40	SMZL	YES	HS kit 1	583	1017.06	92.60
4	47_S41	SMZL	YES	HS kit 1	720	552.57	91.60
5	1994_S37	SMZL	YES	HS kit 2	546	114.69	92.00
5	1995_S36	SMZL	NO	HS kit 2	577	721	88.20
5	30668_S2	not SMZL	NO	HS kit 2	615	100.17	93.80
5	54764_S34	SMZL	YES	HS kit 2	588	317.32	95.60
5	51640_S3	SMZL	YES	HS kit 2	523	84.56	92.60
5	63721_S5	not SMZL	NO	HS kit 2	548	137.13	93.90
5	11717_S31	unknown	NO	HS kit 2	502	138.71	94.50
5	21562_S30	unknown	NO	HS kit 2	634	150.38	94.50
5	8737_S26	unknown	NO	HS kit 2	555	74.35	91.80
5	7875_S23	unknown	NO	HS kit 2	621	162.83	94.50
5	11215_S28	unknown	NO	HS kit 2	561	108.68	93.50
5	8260_S24	unknown	NO	HS kit 2	525	134.23	93.90
5	13664_S32	unknown	NO	HS kit 2	562	165.35	94.60
5	XXI_S38	SMZL	NO	HS kit 2	375	111.31	82.60
5	XXIX_S39	SMZL	YES	HS kit 2	432	86.71	76.00
5	H1558_S13	SMZL	YES	HS kit 2	539	150.3	94.50
5	H3829_S19	SMZL	YES	HS kit 2	567	206.11	94.90
5	H523_S21	SMZL	YES	HS kit 2	591	202.27	94.90
5	H1701_S14	SMZL	YES	HS kit 2	506	158.26	94.00
5	H2486_S18	SMZL	YES	HS kit 2	516	172.16	94.30
5	H2247_S16	SMZL	YES	HS kit 2	528	172.38	94.50
5	H2265_S17	SMZL	YES	HS kit 2	619	142.23	93.80
5	H1954_S15	SMZL	YES	HS kit 2	515	125.32	91.70
5	H4501_S20	SMZL	YES	HS kit 2	593	130.74	93.50
5	783_S12	SMZL	YES	HS kit 2	533	109.39	92.70
5	81389_S7	SMZL	YES	HS kit 2	455	105.13	93.60
5	87936_S8	SMZL	YES	HS kit 2	660	561.32	96.70

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
5	93164_S10	possible SMZL	NO	HS kit 2	585	153.85	95.00
5	92568_S9	SMZL	YES	HS kit 2	540	150.2	95.00
5	75629_S6	SMZL	YES	HS kit 2	552	97.91	93.10
5	7572_S1	SMZL	YES	HS kit 2	526	127.67	94.50
5	7001_S22	unknown	NO	HS kit 2	621	233.96	95.20
5	8443_S25	unknown	NO	HS kit 2	590	131.76	94.20
5	9824_S27	unknown	NO	HS kit 2	879	96.5	93.30
5	11313_S29	unknown	NO	HS kit 2	570	219.57	95.30
5	15740_S33	unknown	NO	HS kit 2	610	179.3	95.20
5	94234_S11	not SMZL	NO	HS kit 2	594	102.77	93.70
5	55276_S4	not SMZL	NO	HS kit 2	554	90.89	93.50
NA	WTCHG_87813_02	SMZL	YES	historical	96	221.09	93.40
NA	WTCHG_90118_04	possible SMZL	YES	historical	108	289.54	92.80
NA	WTCHG_91609_11	SMZL	YES	historical	118	177.21	88.40
NA	WTCHG_90118_06	SMZL	YES	historical	132	249.31	93.50
NA	WTCHG_90060_58	SMZL	YES	historical	109	216.6	89.60
NA	WTCHG_87813_18	SMZL	YES	historical	115	171.01	93.20
NA	WTCHG_87813_25	SMZL	YES	historical	101	155.27	92.60
NA	WTCHG_90060_55	SMZL	YES	historical	111	258.04	89.10
NA	WTCHG_88504_42	SMZL	YES	historical	122	260.84	94.00
NA	WTCHG_90118_13	possible SMZL	YES	historical	107	249.14	94.60
NA	WTCHG_90060_60	SMZL	YES	historical	101	305.57	94.70
NA	WTCHG_88505_53	possible SMZL	YES	historical	113	216.35	92.80
NA	WTCHG_90060_62	SMZL	YES	historical	128	223.66	90.40
NA	WTCHG_75645_43	SMZL	YES	historical	111	149.52	92.40
NA	WTCHG_90119_47	SMZL	YES	historical	117	208.12	93.20
NA	WTCHG_90119_66	SMZL	YES	historical	108	290.55	94.10
NA	WTCHG_90119_71	SMZL	YES	historical	116	308.26	94.90
NA	WTCHG_75645_02	SMZL	YES	historical	102	288.03	94.50
NA	WTCHG_75645_03	SMZL	YES	historical	131	300.22	94.60
NA	WTCHG_75645_04	SMZL	YES	historical	125	268.58	95.20
NA	WTCHG_75645_05	SMZL	YES	historical	107	375.37	95.10
NA	WTCHG_75645_06	SMZL	YES	historical	130	273.49	93.90
NA	WTCHG_75645_07	SMZL	YES	historical	86	354.06	89.20
NA	WTCHG_75645_08	SMZL	YES	historical	109	284.59	95.50
NA	WTCHG_75645_26	SMZL	YES	historical	117	252.53	94.80
NA	WTCHG_75645_28	SMZL	YES	historical	123	319.87	94.40
NA	WTCHG_75645_29	SMZL	YES	historical	113	282.4	95.00
NA	WTCHG_75645_30	SMZL	YES	historical	115	272.11	92.90
NA	WTCHG_75645_32	SMZL	YES	historical	108	290.19	94.30
NA	WTCHG_75645_42	SMZL	YES	historical	99	196.3	93.40
NA	WTCHG_75645_44	SMZL	YES	historical	95	203.54	94.20

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
NA	WTCHG_75645_56	SMZL	YES	historical	125	203.83	93.30
NA	WTCHG_75645_57	SMZL	YES	historical	119	161.39	93.00
NA	WTCHG_75645_58	SMZL	YES	historical	92	149.89	92.60
NA	WTCHG_75645_59	SMZL	YES	historical	103	236.16	94.10
NA	WTCHG_75645_60	SMZL	YES	historical	105	191.34	92.90
NA	WTCHG_75645_61	SMZL	YES	historical	100	182.63	77.20
NA	WTCHG_75645_62	SMZL	YES	historical	99	200.36	94.00
NA	WTCHG_76140_09	SMZL	YES	historical	104	197.72	93.60
NA	WTCHG_76140_10	SMZL	YES	historical	104	175.98	93.80
NA	WTCHG_76140_11	SMZL	YES	historical	90	221.49	93.50
NA	WTCHG_76140_12	SMZL	YES	historical	114	172.04	93.50
NA	WTCHG_76140_13	SMZL	YES	historical	112	184.73	93.90
NA	WTCHG_76140_14	SMZL	YES	historical	107	198.09	93.30
NA	WTCHG_76140_15	SMZL	YES	historical	125	204.13	93.70
NA	WTCHG_76140_16	SMZL	YES	historical	90	194.41	93.60
NA	WTCHG_76140_17	SMZL	YES	historical	109	199.62	94.70
NA	WTCHG_76140_18	SMZL	YES	historical	108	271.88	95.00
NA	WTCHG_76140_19	SMZL	YES	historical	89	258.62	94.90
NA	WTCHG_76140_20	SMZL	YES	historical	117	297.83	95.30
NA	WTCHG_76140_21	SMZL	YES	historical	99	245.61	94.70
NA	WTCHG_76140_22	SMZL	YES	historical	85	243.89	94.30
NA	WTCHG_76140_24	SMZL	YES	historical	97	223.21	93.80
NA	WTCHG_76140_33	SMZL	YES	historical	123	170.25	93.00
NA	WTCHG_76140_34	SMZL	YES	historical	90	138.87	69.00
NA	WTCHG_76140_35	SMZL	YES	historical	105	236.18	94.60
NA	WTCHG_76140_36	SMZL	YES	historical	102	179.67	93.60
NA	WTCHG_76140_37	SMZL	YES	historical	120	201.25	93.80
NA	WTCHG_76140_38	SMZL	YES	historical	91	182.38	93.60
NA	WTCHG_76140_39	SMZL	YES	historical	107	222.35	94.20
NA	WTCHG_76140_40	SMZL	YES	historical	118	148.98	92.60
NA	WTCHG_87813_03	SMZL	YES	historical	105	244.61	92.70
NA	WTCHG_87813_04	SMZL	YES	historical	103	179.74	92.30
NA	WTCHG_87813_05	SMZL	YES	historical	101	270.64	91.60
NA	WTCHG_87813_06	SMZL	YES	historical	111	300.45	93.80
NA	WTCHG_87813_08	SMZL	YES	historical	106	196.79	87.60
NA	WTCHG_87813_09	SMZL	YES	historical	98	226.2	93.30
NA	WTCHG_87813_13	SMZL	YES	historical	102	189.19	86.90
NA	WTCHG_87813_22	SMZL	YES	historical	118	215.54	93.00
NA	WTCHG_87813_24	SMZL	YES	historical	113	245.64	92.10
NA	WTCHG_87813_26	SMZL	YES	historical	124	268.31	94.10
NA	WTCHG_87814_46	SMZL	YES	historical	107	239.52	93.50
NA	WTCHG_87814_47	SMZL	YES	historical	109	294.14	95.70
NA	WTCHG_87814_48	SMZL	YES	historical	104	295.44	95.10
NA	WTCHG_88504_28	SMZL	YES	historical	108	219.72	93.40
NA	WTCHG_88504_31	SMZL	YES	historical	116	172.83	91.90

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
NA	WTCHG_88504_33	SMZL	YES	historical	108	235.91	94.10
NA	WTCHG_88504_36	SMZL	YES	historical	93	214.8	93.60
NA	WTCHG_88504_38	SMZL	YES	historical	121	242.56	95.20
NA	WTCHG_88504_41	SMZL	YES	historical	101	240.27	92.90
NA	WTCHG_88504_45	SMZL	YES	historical	102	335.17	94.40
NA	WTCHG_88504_47	SMZL	YES	historical	112	274.5	93.50
NA	WTCHG_88504_50	SMZL	YES	historical	127	268.51	94.30
NA	WTCHG_88505_51	SMZL	YES	historical	99	189.22	92.00
NA	WTCHG_88505_52	SMZL	YES	historical	115	245.24	93.70
NA	WTCHG_88505_54	SMZL	YES	historical	111	186.2	92.50
NA	WTCHG_88505_55	SMZL	YES	historical	114	235.97	94.40
NA	WTCHG_88505_56	SMZL	YES	historical	120	235.91	92.90
NA	WTCHG_88505_57	SMZL	YES	historical	110	241.95	93.60
NA	WTCHG_88505_58	SMZL	YES	historical	95	206.55	96.00
NA	WTCHG_88505_59	SMZL	YES	historical	114	222.5	93.80
NA	WTCHG_88505_61	SMZL	YES	historical	125	269.8	94.40
NA	WTCHG_88505_64	SMZL	YES	historical	111	235.99	95.60
NA	WTCHG_88505_65	SMZL	YES	historical	99	225.16	95.70
NA	WTCHG_88505_66	SMZL	YES	historical	107	258.83	94.00
NA	WTCHG_88505_67	SMZL	YES	historical	110	209.79	95.60
NA	WTCHG_88505_68	SMZL	YES	historical	125	237.34	95.40
NA	WTCHG_88505_69	SMZL	YES	historical	104	308.99	94.40
NA	WTCHG_88505_70	SMZL	YES	historical	107	194.77	94.80
NA	WTCHG_88505_71	SMZL	YES	historical	101	265.51	95.10
NA	WTCHG_88505_72	SMZL	YES	historical	101	302.83	95.10
NA	WTCHG_88505_73	SMZL	YES	historical	114	215.01	92.30
NA	WTCHG_88505_74	SMZL	YES	historical	108	288.87	96.30
NA	WTCHG_90060_49	SMZL	YES	historical	113	298.54	95.90
NA	WTCHG_90060_50	SMZL	YES	historical	115	245.17	94.40
NA	WTCHG_90060_51	SMZL	YES	historical	99	196.07	94.70
NA	WTCHG_90060_52	SMZL	YES	historical	91	246.42	96.30
NA	WTCHG_90060_53	SMZL	YES	historical	103	344.31	96.40
NA	WTCHG_90060_54	SMZL	YES	historical	91	227.37	94.20
NA	WTCHG_90060_57	SMZL	YES	historical	115	246.12	95.70
NA	WTCHG_90060_59	SMZL	YES	historical	102	230.68	81.90
NA	WTCHG_90060_61	SMZL	YES	historical	111	172.51	92.40
NA	WTCHG_90060_63	SMZL	YES	historical	116	250.6	96.10
NA	WTCHG_90060_65	SMZL	YES	historical	113	319.01	94.60
NA	WTCHG_90060_66	SMZL	YES	historical	111	176.01	92.10
NA	WTCHG_90117_75	SMZL	YES	historical	110	191.05	95.10
NA	WTCHG_90117_77	SMZL	YES	historical	119	191.02	93.00
NA	WTCHG_90117_78	SMZL	YES	historical	106	268.22	93.70
NA	WTCHG_90117_79	SMZL	YES	historical	118	248.24	93.20
NA	WTCHG_90117_80	SMZL	YES	historical	108	203.39	92.20
NA	WTCHG_90117_82	SMZL	YES	historical	119	291.32	95.40

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
NA	WTCHG_90117_84	SMZL	YES	historical	91	195.39	93.00
NA	WTCHG_90117_85	SMZL	YES	historical	108	198.78	92.10
NA	WTCHG_90117_86	SMZL	YES	historical	104	290.69	93.80
NA	WTCHG_90117_87	SMZL	YES	historical	125	220.78	95.70
NA	WTCHG_90117_88	SMZL	YES	historical	114	297.09	94.20
NA	WTCHG_90117_89	SMZL	YES	historical	105	202.7	92.00
NA	WTCHG_90117_90	SMZL	YES	historical	100	185.2	93.10
NA	WTCHG_90117_91	SMZL	YES	historical	114	193.07	94.70
NA	WTCHG_90117_92	SMZL	YES	historical	110	275.86	95.60
NA	WTCHG_90117_93	SMZL	YES	historical	111	368	94.10
NA	WTCHG_90117_94	SMZL	YES	historical	104	97.79	81.60
NA	WTCHG_90117_95	SMZL	YES	historical	108	244.72	92.90
NA	WTCHG_90117_96	SMZL	YES	historical	103	212.64	90.70
NA	WTCHG_90118_05	SMZL	YES	historical	102	243.84	93.30
NA	WTCHG_90118_07	SMZL	YES	historical	103	194.36	91.50
NA	WTCHG_90118_08	SMZL	YES	historical	113	245.85	92.70
NA	WTCHG_90118_16	SMZL	YES	historical	106	303.36	94.30
NA	WTCHG_90119_45	SMZL	YES	historical	117	276.47	94.50
NA	WTCHG_90119_46	SMZL	YES	historical	114	222.56	94.30
NA	WTCHG_90119_48	SMZL	YES	historical	112	255.39	94.10
NA	WTCHG_90119_49	SMZL	YES	historical	119	228.85	94.30
NA	WTCHG_90119_50	SMZL	YES	historical	115	252.23	94.90
NA	WTCHG_90119_51	SMZL	YES	historical	121	202.68	94.10
NA	WTCHG_90119_52	SMZL	YES	historical	108	233.19	95.90
NA	WTCHG_90119_53	SMZL	YES	historical	133	249.96	94.70
NA	WTCHG_90119_54	SMZL	YES	historical	101	224.01	94.00
NA	WTCHG_90119_55	SMZL	YES	historical	105	187.24	93.50
NA	WTCHG_90119_63	SMZL	YES	historical	107	272.18	94.30
NA	WTCHG_90119_64	SMZL	YES	historical	97	331.89	95.00
NA	WTCHG_90119_65	SMZL	YES	historical	97	325.93	94.80
NA	WTCHG_90119_67	SMZL	YES	historical	105	269.96	94.80
NA	WTCHG_90119_68	SMZL	YES	historical	97	282.87	91.10
NA	WTCHG_90119_69	SMZL	YES	historical	101	313.18	94.00
NA	WTCHG_90119_70	SMZL	YES	historical	107	452.25	85.70
NA	WTCHG_90119_72	SMZL	YES	historical	96	304.53	96.30
NA	WTCHG_91608_73	SMZL	YES	historical	110	262.46	94.30
NA	WTCHG_91608_74	SMZL	YES	historical	116	218.09	93.50
NA	WTCHG_91608_78	SMZL	YES	historical	88	198.83	93.40
NA	WTCHG_91608_79	SMZL	YES	historical	96	166.66	92.60
NA	WTCHG_91608_81	SMZL	YES	historical	92	176.75	92.90
NA	WTCHG_91608_82	SMZL	YES	historical	112	200.44	94.10
NA	WTCHG_91608_83	SMZL	YES	historical	104	164.31	92.90
NA	WTCHG_91608_91	SMZL	YES	historical	127	197.27	93.40
NA	WTCHG_91608_93	SMZL	YES	historical	92	276.92	94.10
NA	WTCHG_91608_95	SMZL	YES	historical	107	218.05	93.50

Sequence batch	Sample ID	Diagnosis	Included genomic analysis	Target enrichment kit	No. variants called	Mean target coverage	% bases > 15X
NA	WTCHG_91608_96	SMZL	YES	historical	104	256.28	89.70
NA	WTCHG_91609_14	SMZL	YES	historical	101	151.1	91.10
NA	WTCHG_91609_76	SMZL	YES	historical	102	247.23	96.50
NA	WTCHG_91609_84	SMZL	YES	historical	105	265.95	95.10
NA	WTCHG_91609_89	SMZL	YES	historical	126	335.11	96.70
NA	WTCHG_91609_90	SMZL	YES	historical	108	170.26	92.50
NA	WTCHG_91609_91	SMZL	YES	historical	107	546.42	96.10
NA	WTCHG_91609_94	SMZL	YES	historical	118	166.67	92.90
NA	WTCHG_91609_96	SMZL	YES	historical	117	268.04	94.80

Supplementary Table 6. List of quality metrics assessed for input into machine learning model.

Feature	Definition	Selected	Comments
QUAL	Variant confidence	no	QD the same score but normalised by depth
INFO_BaseQRankSum	Rank sum test of REF vs ALT base quality scores. Compares the base qualities of the data supporting the reference allele with those supporting the alternate allele.	**yes**	Ideal value is close to 0.
INFO_ClippingRankSum	Tests whether the data supporting the reference allele shows more or less base clipping (hard clips) than those supporting the alternate allele.	no	All values were 0
INFO_DS	Were any of the samples downsampled?	no	no values
INFO_ExcessHet	Phred-scaled p-value for exact test of excess heterozygosity.	no	no values
INFO_FS	Fisher's Exact Test to determine if there is strand bias between forward and reverse strands for the reference or alternate allele.	no	updated version SOR
INFO_HaplotypeScore	Consistency of the site with strictly two segregating haplotypes.	no	no values
INFO_InbreedingCoeff	Likelihood-based test for the inbreeding among samples.	no	no values
INFO_MLEAC	Maximum likelihood expectation (MLE) for the allele counts for each ALT allele	no	similar values across samples (binary). Biased towards germline.
INFO_MLEAF	Maximum likelihood expectation (MLE) for the allele frequency for each ALT allele	no	similar values across variants (binary). Biased towards germline.
INFO_MQ	Mapping quality	no	similar values across variants, MQRankSum takes into account
INFO_MQRankSum	Rank Sum Test for mapping qualities of REF versus ALT reads	**yes**	Takes into account the mapping quality of both the REF and ALT alleles.

Feature	Definition	Selected	Comments
INFO_QD	Variant confidence normalized by unfiltered depth of variant samples (QD).	**yes**	Variants in regions with deep coverage can have artificially inflated QUAL scores. Normalisation gives more objective picture
INFO_RAW_MQ	Raw mapping quality	no	no values
INFO_ReadPosRankSum	Rank Sum Test for relative positioning of REF versus ALT alleles within reads. Tests whether there is evidence of bias in the position of alleles within the reads that support them.	**yes**	
INFO_SOR	Determines if there is strand bias between forward and reverse strands for the reference or alternate allele.	**yes**	
frac_depth_MQ20	Fraction of reads with a mapping quality of >=20 (0-1	no	
Number of amplicons	Total number of amplicons covering base	**yes**	
Sum of per base mismatches	Count of per base mismatches in reads containing REF, ALT and other alleles	**yes**	
Sum of softclipped reads	Sum of softclipped reads in base pair location	**yes**	
GT	genotype	no	would divide groups into hom and het
VAF	Variant allele frequency	**yes**	

Supplementary Table 7. Complete list of variants identified in the Jaramillo-Parry cohort (n=321 patients).

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_76140_09	chr1	2556699	2556699	G	A	TNFRSF14	stopgain nonsynonymous	NM_003820	c.G35A	p.W12X
19_MUT	chr1	2558389	2558389	T	G	TNFRSF14	SNV	NM_003820	c.T225G	p.C75W
6_S6	chr1	15920926	15920926	T	A	SPEN	stopgain	NM_015001	c.T1692A	p.Y564X
WTCHG_76140_39	chr1	15922309	15922309	-	T	SPEN	frameshift insertion	NM_015001	c.1810_1811insT	p.Q604fs
92568_S9	chr1	15928434	15928434	C	T	SPEN	stopgain nonsynonymous	NM_015001	c.C2194T	p.R732X
WTCHG_90117_80	chr1	15928483	15928483	C	T	SPEN	SNV nonsynonymous	NM_015001	c.C2243T	p.P748L
WTCHG_87814_48	chr1	15928843	15928843	C	T	SPEN	SNV	NM_015001	c.C2603T	p.A868V
WTCHG_75645_02	chr1	15929232	15929232	C	T	SPEN	stopgain	NM_015001	c.C2992T	p.Q998X
WTCHG_87813_18	chr1	15929233	15929233	A	-	SPEN	frameshift deletion nonsynonymous	NM_015001	c.2993delA	p.Q998fs
WTCHG_90117_90	chr1	15929269	15929269	A	T	SPEN	SNV nonsynonymous	NM_015001	c.A3029T	p.E1010V
WTCHG_87813_24	chr1	15929398	15929398	A	G	SPEN	SNV	NM_015001	c.A3158G	p.K1053R
WTCHG_75645_61	chr1	15929961	15929961	C	T	SPEN	stopgain nonsynonymous	NM_015001	c.C3721T	p.R1241X
WTCHG_90119_70	chr1	15930040	15930040	G	A	SPEN	SNV	NM_015001	c.G3800A	p.G1267D
31_S31	chr1	15930092	15930092	A	-	SPEN	frameshift deletion	NM_015001	c.3852delA	p.V1284fs
16_MUT	chr1	15930708	15930708	A	-	SPEN	frameshift deletion nonsynonymous	NM_015001	c.4468delA	p.I1490fs
WTCHG_91608_82	chr1	15931015	15931015	G	T	SPEN	SNV	NM_015001	c.G4775T	p.R1592L
WTCHG_90060_61	chr1	15931419	15931419	C	T	SPEN	stopgain	NM_015001	c.C5179T	p.Q1727X
28_S22	chr1	15934646	15934647	AG	-	SPEN	frameshift deletion nonsynonymous	NM_015001	c.8406_8407del	p.S2802fs
45_S45	chr1	15934699	15934699	C	T	SPEN	SNV	NM_015001	c.C8459T	p.S2820L
WTCHG_91608_73	chr1	15935271	15935271	C	T	SPEN	stopgain	NM_015001	c.C9031T	p.R3011X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90060_61	chr1	15937422	15937425	CTTC	-	SPEN	frameshift deletion	NM_015001	c.10286_10289del	p.T3429fs
WTCHG_76140_36	chr1	15937570	15937570	C	G	SPEN	nonsynonymous SNV	NM_015001	c.C10434G	p.F3478L
WTCHG_76140_36	chr1	15937570	15937570	C	G	SPEN	nonsynonymous SNV	NM_015001	c.C10434G	p.F3478L
L023_07	chr1	23558965	23558965	G	A	ID3	nonsynonymous SNV	NM_002167	c.C355T	p.H119Y
30_MUT	chr1	23559237	23559237	G	A	ID3	nonsynonymous SNV	NM_002167	c.C190T	p.L64F
49_S43	chr1	23559362	23559362	A	G	ID3	SNV	NM_002167	c.T65C	p.L22P
WTCHG_90119_52	chr1	26696857	26696857	C	-	ARID1A	frameshift deletion	NM_006015	c.454delC	p.Q152fs
WTCHG_90118_05	chr1	26697129	26697129	-	GCG	ARID1A	nonframeshift insertion	NM_006015	c.726_727insGCG	p.G242delinsGA
WTCHG_75645_60	chr1	26697208	26697208	C	T	ARID1A	stopgain	NM_006015	c.C805T	p.Q269X
WTCHG_75645_60	chr1	26697478	26697478	C	T	ARID1A	nonsynonymous SNV	NM_006015	c.C1075T	p.H359Y
WTCHG_76140_24	chr1	26761426	26761426	G	A	ARID1A	SNV	NM_006015	c.G2204A	p.S735N
WTCHG_76140_21	chr1	26762217	26762217	C	-	ARID1A	frameshift deletion	NM_006015	c.2317delC	p.P773fs
50_S44	chr1	26762297	26762297	G	-	ARID1A	frameshift deletion	NM_006015	c.2397delG	p.Q799fs
WTCHG_87814_47	chr1	26763018	26763018	A	G	ARID1A	nonsynonymous SNV	NM_006015	c.A2465G	p.N822S
WTCHG_76140_34	chr1	26771150	26771150	C	A	ARID1A	nonsynonymous SNV	NM_006015	c.C3230A	p.A1077E
WTCHG_88505_61	chr1	26772529	26772529	A	G	ARID1A	nonsynonymous SNV	NM_006015	c.A3436G	p.T1146A
L018_06	chr1	26772541	26772541	A	T	ARID1A	SNV	NM_006015	c.A3448T	p.T1150S
WTCHG_88504_31	chr1	26772633	26772633	G	A	ARID1A	splicing	NM_006015	c.3539+1G>A	.
WTCHG_87813_18	chr1	26774631	26774631	T	-	ARID1A	frameshift deletion	NM_006015	c.4404delT	p.P1468fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90119_53	chr1	26779212	26779212	A	G	ARID1A	nonsynonymous SNV	NM_006015	c.A5314G	p.K1772E
WTCHG_88505_74	chr1	26779389	26779389	C	G	ARID1A	nonsynonymous SNV	NM_006015	c.C5491G	p.L1831V
33_MUT	chr1	26779439	26779439	-	G	ARID1A	frameshift insertion	NM_006015	c.5542dupG	p.G1847fs
H523_S21	chr1	26780010	26780010	G	A	ARID1A	nonsynonymous SNV	NM_006015	c.G6112A	p.D2038N
622_S54	chr1	26780081	26780081	G	C	ARID1A	nonsynonymous SNV	NM_006015	c.G6183C	p.L2061F
1_MUT	chr1	26780224	26780224	A	G	ARID1A	nonsynonymous SNV	NM_006015	c.A6326G	p.N2109S
WTCHG_76140_24	chr1	26780585	26780585	C	A	ARID1A	nonsynonymous SNV	NM_006015	c.C6687A	p.D2229E
WTCHG_90117_94	chr1	26780628	26780628	G	T	ARID1A	nonsynonymous SNV	NM_006015	c.G6730T	p.V2244L
9_S7	chr1	27359959	27359959	G	C	MAP3K6	nonsynonymous SNV	NM_004672	c.C2218G	p.P740A
40_S40	chr1	27364044	27364044	G	A	MAP3K6	nonsynonymous SNV	NM_004672	c.C737T	p.A246V
L071_10	chr1	27364354	27364354	G	A	MAP3K6	nonsynonymous SNV	NM_004672	c.C545T	p.T182M
L037_08	chr1	85267701	85267701	C	A	BCL10	stopgain	NM_003921	c.G628T	p.E210X
4683_S16	chr1	85267857	85267857	C	A	BCL10	stopgain	NM_003921	c.G472T	p.G158X
L096_13	chr1	85267867	85267867	G	T	BCL10	stopgain	NM_003921	c.C462A	p.Y154X
37_S31	chr1	119915464	119915464	C	A	NOTCH2	stopgain	NM_024408	c.G7258T	p.E2420X
WTCHG_90119_53	chr1	119915464	119915464	C	A	NOTCH2	stopgain	NM_024408	c.G7258T	p.E2420X
L067_10	chr1	119915478	119915481	AGGT	-	NOTCH2	frameshift deletion	NM_024408	c.7241_7244del	p.Y2414fs
WTCHG_75645_592_S2	chr1	119915480	119915480	G	T	NOTCH2	stopgain	NM_024408	c.C7242A	p.Y2414X
WTCHG_87813_18	chr1	119915497	119915497	G	A	NOTCH2	stopgain	NM_024408	c.C7225T	p.Q2409X
12_MUT	chr1	119915524	119915524	G	A	NOTCH2	stopgain	NM_024408	c.C7198T	p.R2400X
	chr1	119915546	119915546	A	T	NOTCH2	stopgain	NM_024408	c.T7176A	p.Y2392X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90119_66	chr1	119915593	119915593	G	A	NOTCH2	nonsynonymous SNV	NM_024408	c.C7129T	p.P2377S
WTCHG_87813_05	chr1	119915623	119915623	G	A	NOTCH2	stopgain	NM_024408	c.C7099T	p.Q2367X
31_MUT	chr1	119915632	119915632	G	A	NOTCH2	stopgain	NM_024408	c.C7090T	p.Q2364X
4683_S16	chr1	119915632	119915632	G	A	NOTCH2	stopgain	NM_024408	c.C7090T	p.Q2364X
38_S32	chr1	119915635	119915641	CGTCCTG	-	NOTCH2	frameshift deletion	NM_024408	c.7081_7087del	p.Q2361fs
WTCHG_90060_60	chr1	119915641	119915641	G	A	NOTCH2	stopgain	NM_024408	c.C7081T	p.Q2361X
H2265_S17	chr1	119915644	119915644	G	A	NOTCH2	stopgain	NM_024408	c.C7078T	p.Q2360X
15_MUT	chr1	119915701	119915701	G	A	NOTCH2	stopgain	NM_024408	c.C7021T	p.Q2341X
WTCHG_75645_06	chr1	119915701	119915701	G	A	NOTCH2	stopgain	NM_024408	c.C7021T	p.Q2341X
WTCHG_90117_77	chr1	119915749	119915749	G	A	NOTCH2	stopgain	NM_024408	c.C6973T	p.Q2325X
WTCHG_90119_45	chr1	119915749	119915749	G	A	NOTCH2	stopgain	NM_024408	c.C6973T	p.Q2325X
24_S18	chr1	119915802	119915803	AA	-	NOTCH2	frameshift deletion	NM_024408	c.6919_6920del	p.F2307fs
836_S20	chr1	119915811	119915814	ATGG	-	NOTCH2	frameshift deletion	NM_024408	c.6908_6911del	p.P2303fs
30_S24	chr1	119915812	119915812	-	G	NOTCH2	frameshift insertion	NM_024408	c.6909dupC	p.I2304fs
WTCHG_90118_08	chr1	119915812	119915812	-	G	NOTCH2	frameshift insertion	NM_024408	c.6909dupC	p.I2304fs
L017_06	chr1	119915813	119915813	G	-	NOTCH2	frameshift deletion	NM_024408	c.6909delC	p.P2303fs
L025_07	chr1	119915813	119915813	G	-	NOTCH2	frameshift deletion	NM_024408	c.6909delC	p.P2303fs
L075_10	chr1	119915813	119915816	GGGG	-	NOTCH2	frameshift deletion	NM_024408	c.6906_6909del	p.P2302fs
87936_S8	chr1	119915836	119915840	TGGTT	-	NOTCH2	frameshift deletion	NM_024408	c.6882_6886del	p.I2294fs
WTCHG_75645_03	chr1	119915848	119915848	-	C	NOTCH2	frameshift insertion	NM_024408	c.6873dupG	p.K2292fs
9641_S17	chr1	119915851	119915855	ACTTC	-	NOTCH2	frameshift deletion	NM_024408	c.6867_6871del	p.P2289fs
53_S47	chr1	119915869	119915869	G	-	NOTCH2	frameshift deletion	NM_024408	c.6853delC	p.Q2285fs
868_S8	chr1	119915869	119915869	G	A	NOTCH2	stopgain	NM_024408	c.C6853T	p.Q2285X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90117_96	chr1	119915879	119915880	GC	-	NOTCH2	frameshift deletion	NM_024408	c.6842_6843del	p.G2281fs
WTCHG_90060_55	chr1	119915896	119915897	CA	-	NOTCH2	frameshift deletion	NM_024408	c.6825_6826del	p.A2275fs
WTCHG_90119_63	chr1	119915915	119915919	ACCAA	-	NOTCH2	frameshift deletion	NM_024408	c.6803_6807del	p.F2268fs
L076_10	chr1	119915964	119915964	C	T	NOTCH2	stopgain	NM_024408	c.G6758A	p.W2253X
14_MUT	chr1	119916288	119916288	G	C	NOTCH2	stopgain	NM_024408	c.C6434G	p.S2145X
WTCHG_76140_09	chr1	119916295	119916295	C	-	NOTCH2	frameshift deletion	NM_024408	c.6427delG	p.E2143fs
WTCHG_75645_08	chr1	119916325	119916325	T	A	NOTCH2	stopgain	NM_024408	c.A6397T	p.K2133X
5_S5	chr1	119916408	119916408	C	T	NOTCH2	nonsynonymous SNV	NM_024408	c.G6314A	p.R2105Q
WTCHG_91608_93	chr1	119916505	119916515	CTGGAGGCTT	-	NOTCH2	frameshift deletion	NM_024408	c.6207_6217del	p.P2069fs
WTCHG_90119_66	chr1	119922671	119922671	T	C	NOTCH2	nonsynonymous SNV	NM_024408	c.A4967G	p.Q1656R
81389_S7	chr1	119925683	119925683	C	A	NOTCH2	nonsynonymous SNV	NM_024408	c.G4133T	p.C1378F
30_S24	chr1	119937325	119937325	T	C	NOTCH2	nonsynonymous SNV	NM_024408	c.A3479G	p.H1160R
35_MUT	chr1	119941611	119941611	C	G	NOTCH2	nonsynonymous SNV	NM_024408	c.G2896C	p.D966H
WTCHG_88505_53	chr10	62813488	62813488	G	T	EGR2	nonsynonymous SNV	NM_000399	c.C1150A	p.H384N
L40_S15	chr10	62813710	62813710	-	GGC	EGR2	nonframeshift insertion	NM_000399	c.927_928insGCC	p.Y310delinsAY
L023_07	chr10	62814442	62814442	C	T	EGR2	nonsynonymous SNV	NM_000399	c.G196A	p.G66R
783_S12	chr10	62816070	62816070	C	T	EGR2	splicing	.	.	.
868_S8	chr10	71566884	71566884	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G572A	p.R191Q

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
H1558_S13	chr10	71570868	71570868	A	G	CDH23	nonsynonymous SNV	NM_022124	c.A703G	p.I235V
WTCHG_90119_47	chr10	71615513	71615513	A	G	CDH23	nonsynonymous SNV	NM_022124	c.A842G	p.N281S
WTCHG_91608_83	chr10	71645903	71645903	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G1213A	p.V405I
9_S7	chr10	71677461	71677461	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C1520T	p.S507L
L067_10	chr10	71677461	71677461	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C1520T	p.S507L
WTCHG_87814_48	chr10	71677461	71677461	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C1520T	p.S507L
WTCHG_88504_36	chr10	71677613	71677613	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G1672A	p.V558M
WTCHG_88505_52	chr10	71677613	71677613	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G1672A	p.V558M
WTCHG_90117_82	chr10	71677613	71677613	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G1672A	p.V558M
WTCHG_88505_51	chr10	71687692	71687692	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G2032A	p.V678I
30_S24	chr10	71694233	71694233	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C2263T	p.H755Y
L076_10	chr10	71705116	71705116	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C2939T	p.T980M
3_MUT	chr10	71707005	71707005	A	T	CDH23	nonsynonymous SNV	NM_022124	c.A3062T	p.D1021V
WTCHG_87813_05	chr10	71709109	71709109	G	T	CDH23	nonsynonymous SNV	NM_022124	c.G3118T	p.D1040Y
L076_10	chr10	71712698	71712698	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C3254T	p.T1085I
2_MUT	chr10	71741829	71741829	A	T	CDH23	SNV	NM_022124	c.A4753T	p.I1585F

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_76140_16	chr10	71779307	71779307	C	A	CDH23	nonsynonymous SNV	NM_022124	c.C5228A	p.T1743N
WTCHG_75645_44	chr10	71779442	71779442	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C5363T	p.P1788L
WTCHG_90119_65	chr10	71785652	71785652	C	T	CDH23	nonsynonymous SNV	NM_022124	c.C5734T	p.R1912W
H1701_S14	chr10	71807518	71807518	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G8311A	p.G2771S
WTCHG_75645_07	chr10	71807518	71807518	G	A	CDH23	nonsynonymous SNV	NM_022124	c.G8311A	p.G2771S
24_S18	chr10	71811725	71811725	G	T	CDH23	nonsynonymous SNV	NM_022124	c.G9291T	p.K3097N
783_S12	chr10	71812003	71812003	A	T	CDH23	nonsynonymous SNV	NM_022124	c.A9368T	p.Y3123F
11731_S18	chr11	6626329	6626329	C	T	DCHS1	nonsynonymous SNV	NM_003737	c.G6416A	p.R2139Q
3_S3	chr11	6627067	6627067	C	T	DCHS1	nonsynonymous SNV	NM_003737	c.G5972A	p.R1991H
L096_13	chr11	6632813	6632813	G	A	DCHS1	nonsynonymous SNV	NM_003737	c.C2699T	p.T900M
L098_13_S59	chr11	6634216	6634216	G	A	DCHS1	SNV	NM_003737	c.C1888T	p.R630C
L088_12	chr11	102321901	102321901	A	G	BIRC3	splicing	.	.	.
L051_09	chr11	102324687	102324687	G	A	BIRC3	nonsynonymous SNV	NM_001165	c.G178A	p.V60M
WTCHG_88505_52	chr11	102324852	102324852	-	C	BIRC3	frameshift insertion	NM_001165	c.344dupC	p.S115fs
12603_S20	chr11	102331198	102331202	AAGG	-	BIRC3	frameshift deletion	NM_001165	c.1281_1285del	p.I427fs
WTCHG_91608_95	chr11	102331208	102331211	GAGA	-	BIRC3	frameshift deletion	NM_001165	c.1291_1294del	p.E431fs
L022_06	chr11	102331218	102331218	G	A	BIRC3	nonsynonymous SNV	NM_001165	c.G1301A	p.R434K
617_S3	chr11	102331235	102331235	G	T	BIRC3	stopgain	NM_001165	c.G1318T	p.E440X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
12_S9	chr11	102336116	102336116	C	T	BIRC3	nonsynonymous SNV	NM_001165	c.C1475T	p.A492V
3_S3	chr11	102336926	102336926	C	-	BIRC3	frameshift deletion	NM_001165	c.1639delC	p.Q547fs
6_MUT	chr11	102336926	102336926	C	-	BIRC3	frameshift deletion	NM_001165	c.1639delC	p.Q547fs
81389_S7	chr11	102336926	102336926	C	-	BIRC3	frameshift deletion	NM_001165	c.1639delC	p.Q547fs
WTCHG_91609_91	chr11	102336931	102336934	GCGG	-	BIRC3	frameshift deletion	NM_001165	c.1644_1647del	p.L548fs
L048_09	chr11	102336945	102336948	AAGA	-	BIRC3	frameshift deletion	NM_001165	c.1658_1661del	p.E553fs
WTCHG_88505_72	chr11	102336984	102336984	C	-	BIRC3	frameshift deletion	NM_001165	c.1697delC	p.S566fs
46_S40	chr11	102336990	102336990	T	A	BIRC3	nonsynonymous SNV	NM_001165	c.T1703A	p.V568E
3_S3	chr11	102337017	102337017	T	A	BIRC3	nonsynonymous SNV	NM_001165	c.T1730A	p.V577E
WTCHG_88505_52	chr11	102337083	102337083	T	C	BIRC3	SNV	NM_001165	c.T1796C	p.V599A
WTCHG_91609_90	chr11	108235832	108235832	T	G	ATM	stopgain	NM_000051	c.T494G	p.L165X
WTCHG_88504_38	chr11	108244072	108244072	A	C	ATM	nonsynonymous SNV	NM_000051	c.A616C	p.N206H
WTCHG_90060_49	chr11	108249082	108249082	T	-	ATM	frameshift deletion	NM_000051	c.1215delT	p.N405fs
1_S1	chr11	108250748	108250751	CTAA	-	ATM	frameshift deletion	NM_000051	c.1283_1286del	p.P428fs
92568_S9	chr11	108250840	108250840	C	T	ATM	nonsynonymous SNV	NM_000051	c.C1375T	p.L459F
WTCHG_91609_90	chr11	108251038	108251038	A	-	ATM	frameshift deletion	NM_000051	c.1573delA	p.K525fs
606_S12	chr11	108251956	108251956	T	C	ATM	nonsynonymous SNV	NM_000051	c.T1727C	p.I576T
2_MUT	chr11	108256340	108256340	G	A	ATM	synonymous SNV	NM_000051	c.G2250A	p.K750K
WTCHG_90118_05	chr11	108259077	108259077	T	G	ATM	splicing	NM_000051	c.2466+2T>G	.
46_S40	chr11	108271292	108271292	A	C	ATM	nonsynonymous SNV	NM_000051	c.A2963C	p.K988T

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_76140_15	chr11	108284405	108284405	G	A	ATM	nonsynonymous SNV	NM_000051	c.G3925A	p.A1309T
12603_S20	chr11	108289724	108289724	A	-	ATM	frameshift deletion	NM_000051	c.4359delA	p.I1453fs
WTCHG_88505_74	chr11	108289789	108289789	A	G	ATM	nonsynonymous SNV	NM_000051	c.A4424G	p.Y1475C
87936_S8	chr11	108301770	108301770	T	G	ATM	nonsynonymous SNV	NM_000051	c.T5300G	p.F1767C
WTCHG_90118_05	chr11	108301790	108301790	G	A	ATM	splicing	NM_000051	c.5319+1G>A	.
8_S8	chr11	108302901	108302901	G	C	ATM	nonsynonymous SNV	NM_000051	c.G5368C	p.D1790H
WTCHG_76140_37	chr11	108310255	108310255	C	G	ATM	nonsynonymous SNV	NM_000051	c.C5858G	p.T1953R
1_S1	chr11	108315872	108315872	A	C	ATM	nonsynonymous SNV	NM_000051	c.A6056C	p.Y2019S
L012_04	chr11	108326151	108326151	G	A	ATM	nonsynonymous SNV	NM_000051	c.G6901A	p.A2301T
L076_10	chr11	108327657	108327657	C	G	ATM	nonsynonymous SNV	NM_000051	c.C6988G	p.L2330V
L011_03	chr11	108329118	108329118	C	G	ATM	nonsynonymous SNV	NM_000051	c.C7187G	p.T2396S
WTCHG_90060_49	chr11	108329198	108329198	G	A	ATM	nonsynonymous SNV	NM_000051	c.G7267A	p.E2423K
WTCHG_91608_95	chr11	108330234	108330234	G	A	ATM	nonsynonymous SNV	NM_000051	c.G7328A	p.R2443Q
WTCHG_75645_56	chr11	108330381	108330381	T	G	ATM	SNV	NM_000051	c.T7475G	p.L2492R
WTCHG_90117_79	chr11	108332902	108332902	T	A	ATM	splicing	NM_000051	c.7927+2T>A	.
WTCHG_76140_37	chr11	108333941	108333943	TGT	-	ATM	nonframeshift deletion	NM_000051	c.7983_7985del	p.2661_2662del
3_MUT	chr11	108335109	108335109	G	T	ATM	nonsynonymous SNV	NM_000051	c.G8151T	p.K2717N
2_MUT	chr11	108345897	108345899	CTT	-	ATM	nonframeshift deletion	NM_000051	c.8573_8575del	p.2858_2859del

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_87813_0853_S47	chr11	108354869	108354869	G	A	ATM	nonsynonymous SNV	NM_000051	c.G8845A	p.V2949I
	chr11	108365415	108365415	-	A	ATM	frameshift insertion	NM_000051	c.9079dupA	p.L3026fs
16_MUT	chr12	25225663	25225663	G	C	KRAS	nonsynonymous SNV	NM_004985	c.C401G	p.A134G
L080_11	chr12	25245314	25245314	A	T	KRAS	nonsynonymous SNV	NM_004985	c.T71A	p.I24N
WTCHG_90117_86	chr12	25245350	25245350	C	A	KRAS	nonsynonymous SNV	NM_004985	c.G35T	p.G12V
H523_S21	chr12	49022717	49022717	G	A	KMT2D	SNV	NM_003482	c.C16211T	p.S5404F
WTCHG_88504_31	chr12	49022869	49022869	-	TGGGGCCT	KMT2D	frameshift insertion	NM_003482	c.16058_16059insCCCAGGCCCCA	p.H5353fs
23_MUT	chr12	49024860	49024860	C	T	KMT2D	nonsynonymous SNV	NM_003482	c.G15871A	p.E5291K
WTCHG_90060_55	chr12	49024934	49024934	C	T	KMT2D	nonsynonymous SNV	NM_003482	c.G15797A	p.R5266H
H523_S21	chr12	49026337	49026337	T	C	KMT2D	nonsynonymous SNV	NM_003482	c.A15629G	p.Y5210C
WTCHG_76140_09	chr12	49026505	49026505	C	T	KMT2D	SNV	NM_003482	c.G15461A	p.R5154Q
19_MUT	chr12	49026748	49026748	-	G	KMT2D	frameshift insertion	NM_003482	c.15217dupC	p.Q5073fs
WTCHG_88505_57	chr12	49027151	49027151	C	T	KMT2D	nonsynonymous SNV	NM_003482	c.G14815A	p.E4939K
L098_13_S59	chr12	49027817	49027817	T	G	KMT2D	nonsynonymous SNV	NM_003482	c.A14629C	p.S4877R
WTCHG_87814_46	chr12	49028866	49028866	C	A	KMT2D	SNV	NM_003482	c.G14344T	p.V4782L
WTCHG_88504_50	chr12	49029399	49029399	A	G	KMT2D	splicing	NM_003482	c.14075+2T>C	.
WTCHG_76140_38	chr12	49030660	49030660	C	G	KMT2D	nonsynonymous SNV	NM_003482	c.G13780C	p.A4594P

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
21_S12	chr12	49030729	49030729	C	T	KMT2D	nonsynonymous SNV	NM_003482	c.G13711A	p.A4571T
30_MUT	chr12	49031921	49031921	G	A	KMT2D	stopgain	NM_003482	c.C12784T	p.Q4262X
4683_S16	chr12	49032002	49032002	G	A	KMT2D	stopgain	NM_003482	c.C12703T	p.Q4235X
L40_S15	chr12	49032476	49032476	G	A	KMT2D	nonsynonymous SNV	NM_003482	c.C12229T	p.L4077F
Kalpadakis_L17	chr12	49033897	49033897	T	G	KMT2D	nonsynonymous SNV	NM_003482	c.A10808C	p.Q3603P
WTCHG_90119_66	chr12	49034446	49034446	G	A	KMT2D	nonsynonymous SNV	NM_003482	c.C10471T	p.R3491C
2_S2	chr12	49034911	49034911	T	C	KMT2D	SNV	NM_003482	c.A10256G	p.D3419G
WTCHG_90118_08	chr12	49038267	49038267	T	-	KMT2D	frameshift deletion	NM_003482	c.9089delA	p.N3030fs
WTCHG_90119_52	chr12	49038626	49038629	ACTT	-	KMT2D	frameshift deletion	NM_003482	c.8727_8730del	p.V2909fs
6_S6	chr12	49038924	49038925	TG	-	KMT2D	frameshift deletion	NM_003482	c.8431_8432del	p.Q2811fs
8_MUT	chr12	49039277	49039277	G	A	KMT2D	stopgain	NM_003482	c.C8311T	p.R2771X
22_S13	chr12	49039759	49039759	C	T	KMT2D	nonsynonymous SNV	NM_003482	c.G8011A	p.G2671S
WTCHG_87813_18	chr12	49039882	49039882	G	-	KMT2D	frameshift deletion	NM_003482	c.7888delC	p.H2630fs
38_S32	chr12	49040409	49040409	G	A	KMT2D	nonsynonymous SNV	NM_003482	c.C7361T	p.P2454L
39_S39	chr12	49041175	49041175	A	-	KMT2D	frameshift deletion	NM_003482	c.6595delT	p.Y2199fs
WTCHG_90117_96	chr12	49041262	49041262	G	C	KMT2D	nonsynonymous SNV	NM_003482	c.C6508G	p.Q2170E
WTCHG_75645_26	chr12	49041330	49041330	G	A	KMT2D	nonsynonymous SNV	NM_003482	c.C6440T	p.A2147V
H1954_S15	chr12	49041445	49041445	G	A	KMT2D	stopgain	NM_003482	c.C6325T	p.Q2109X
WTCHG_91609_91	chr12	49041517	49041517	T	G	KMT2D	nonsynonymous SNV	NM_003482	c.A6253C	p.N2085H
WTCHG_90119_63	chr12	49041917	49041917	C	T	KMT2D	synonymous SNV	NM_003482	c.G6183A	p.L2061L

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_75645_28	chr12	49042090	49042090	T	A	KMT2D	synonymous SNV	NM_003482	c.A6108T	p.P2036P
WTCHG_76140_17	chr12	49044410	49044410	A	C	KMT2D	stopgain	NM_003482	c.T5076G	p.Y1692X
35_S35	chr12	49046064	49046064	C	T	KMT2D	splicing	NM_003482	c.4693+1G>A	.
WTCHG_90117_96	chr12	49046125	49046125	G	A	KMT2D	stopgain	NM_003482	c.C4633T	p.Q1545X
WTCHG_76140_15	chr12	49046172	49046172	C	T	KMT2D	stopgain	NM_003482	c.G4586A	p.W1529X
WTCHG_76140_09	chr12	49046648	49046648	G	-	KMT2D	frameshift deletion nonsynonymous	NM_003482	c.4379delC	p.P1460fs
WTCHG_90117_90	chr12	49050395	49050395	A	T	KMT2D	SNV	NM_003482	c.T3193A	p.S1065T
WTCHG_91608_73	chr12	49050521	49050521	G	A	KMT2D	stopgain	NM_003482	c.C3067T	p.Q1023X
H2247_S16	chr12	49050604	49050610	GGCTCAG	-	KMT2D	frameshift deletion nonsynonymous	NM_003482	c.2978_2984del	p.P993fs
7572_S1	chr12	49051551	49051551	G	A	KMT2D	SNV nonsynonymous	NM_003482	c.C2132T	p.P711L
L086_12	chr12	49051771	49051771	G	C	KMT2D	SNV nonsynonymous	NM_003482	c.C1912G	p.P638A
WTCHG_90119_50	chr12	49052166	49052166	G	T	KMT2D	SNV	NM_003482	c.C1517A	p.P506Q
92568_S9	chr12	49052244	49052244	G	-	KMT2D	frameshift deletion	NM_003482	c.1439delC	p.P480fs
WTCHG_75645_07	chr12	49052961	49052961	G	A	KMT2D	stopgain	NM_003482	c.C1066T	p.Q356X
WTCHG_88505_57	chr12	49054063	49054063	G	-	KMT2D	frameshift deletion nonsynonymous	NM_003482	c.588delC	p.P196fs
10_S10	chr12	121439961	121439961	T	A	KDM2B	SNV nonsynonymous	NM_032590	c.A3725T	p.K1242M
WTCHG_76140_35	chr12	121443768	121443768	C	T	KDM2B	SNV	NM_032590	c.G2477A	p.G826D
WTCHG_90117_79	chr12	121453346	121453346	T	C	KDM2B	splicing nonsynonymous	NM_032590	c.1735-2A>G	.
WTCHG_76140_12	chr12	121494647	121494647	C	T	KDM2B	SNV nonsynonymous	NM_032590	c.G1666A	p.A556T
WTCHG_88505_72	chr12	121509779	121509779	C	T	KDM2B	SNV nonsynonymous	NM_032590	c.G1435A	p.E479K
17_MUT	chr12	121509784	121509784	G	A	KDM2B	SNV	NM_032590	c.C1430T	p.S477L

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90060_61	chr12	121548981	121548981	T	C	KDM2B	splicing	NM_032590	c.581-2A>G	.
L071_10	chr12	121805997	121805997	A	G	SETD1B	nonsynonymous SNV	NM_00135334 5	c.A436G	p.I146V
L043_08	chr12	121810423	121810423	A	G	SETD1B	nonsynonymous SNV	NM_00135334 5	c.A1478G	p.E493G
L037_08	chr12	121810599	121810599	C	T	SETD1B	nonsynonymous SNV	NM_00135334 5	c.C1654T	p.P552S
L094_13	chr12	121814163	121814163	C	G	SETD1B	nonsynonymous SNV	NM_00135334 5	c.C1948G	p.P650A
22_S13	chr12	121814665	121814665	A	G	SETD1B	nonsynonymous SNV	NM_00135334 5	c.A2450G	p.Y817C
L37_S14	chr12	121822731	121822731	C	T	SETD1B	synonymous SNV	NM_00135334 5	c.C4152T	p.G1384G
L018_06	chr12	121822943	121822943	G	A	SETD1B	nonsynonymous SNV	NM_00135334 5	c.G4364A	p.R1455H
WTCHG_90119_72	chr14	102870257	102870257	-	A	TRAF3	frameshift insertion	NM_145725	c.57dupA	p.L19fs
WTCHG_76140_24	chr14	102870421	102870421	G	T	TRAF3	stopgain	NM_145725	c.G220T	p.E74X
H2486_S18	chr14	102871966	102871966	A	-	TRAF3	frameshift deletion	NM_145725	c.295delA	p.K99fs
H3829_S19	chr14	102876356	102876356	A	G	TRAF3	splicing	NM_145725	c.403-2A>G	.
L018_06	chr14	102886210	102886210	C	-	TRAF3	frameshift deletion	NM_145725	c.592delC	p.P198fs
31_MUT	chr14	102897275	102897278	GAAT	-	TRAF3	frameshift deletion	NM_145725	c.834_837del	p.Q278fs
L018_06	chr14	102897280	102897280	A	-	TRAF3	frameshift deletion	NM_145725	c.839delA	p.E280fs
491_S11	chr14	102897297	102897298	AA	-	TRAF3	frameshift deletion	NM_145725	c.856_857del	p.K286fs
52_S46	chr14	102897369	102897369	C	T	TRAF3	stopgain	NM_145725	c.C928T	p.R310X
WTCHG_88505_59	chr14	102897375	102897378	AATG	-	TRAF3	frameshift deletion	NM_145725	c.934_937del	p.N312fs
25_MUT	chr14	102897399	102897399	C	T	TRAF3	stopgain	NM_145725	c.C958T	p.Q320X
WTCHG_76140_24	chr14	102897403	102897403	T	A	TRAF3	splicing	NM_145725	c.960+2T>A	.
WTCHG_88505_59	chr14	102903263	102903266	AGAC	-	TRAF3	frameshift deletion	NM_145725	c.969_972del	p.I323fs
WTCHG_76140_18	chr14	102903316	102903316	G	A	TRAF3	nonsynonymous SNV	NM_145725	c.G1022A	p.R341Q
WTCHG_75645_08	chr14	102903318	102903318	C	T	TRAF3	stopgain	NM_145725	c.C1024T	p.Q342X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
22_S13	chr14	102905245	102905245	C	T	TRAF3	stopgain	NM_145725	c.C1168T	p.Q390X
9_S7	chr14	102905250	102905253	GCTG	-	TRAF3	frameshift deletion	NM_145725	c.1173_1176del	p.M391fs
18_MUT	chr14	102905336	102905336	G	A	TRAF3	stopgain	NM_145725	c.G1259A	p.W420X
54764_S34	chr14	102905365	102905365	C	T	TRAF3	stopgain	NM_145725	c.C1288T	p.Q430X
WTCHG_75645_06	chr14	102905405	102905405	A	-	TRAF3	frameshift deletion	NM_145725	c.1328delA	p.Q443fs
L044_08	chr14	102905425	102905425	T	-	TRAF3	frameshift deletion	NM_145725	c.1348delT	p.F450fs
H1558_S13	chr14	102905461	102905461	G	A	TRAF3	nonsynonymous SNV	NM_145725	c.G1384A	p.G462R
WTCHG_90060_60	chr14	102905497	102905497	T	-	TRAF3	frameshift deletion	NM_145725	c.1420delT	p.F474fs
868_S8	chr14	102905551	102905551	C	T	TRAF3	stopgain	NM_145725	c.C1474T	p.Q492X
WTCHG_75645_06	chr14	102905590	102905590	C	T	TRAF3	stopgain	NM_145725	c.C1513T	p.R505X
WTCHG_90117_85	chr14	102905651	102905651	-	TG	TRAF3	frameshift insertion	NM_145725	c.1574_1575insTG	p.T525fs
L049_09_S31	chr15	66435105	66435105	T	A	MAP2K1	nonsynonymous SNV	NM_002755	c.T159A	p.F53L
WTCHG_76140_21	chr15	66435115	66435115	A	G	MAP2K1	nonsynonymous SNV	NM_002755	c.A169G	p.K57E
31_S31	chr15	66436825	66436825	C	T	MAP2K1	nonsynonymous SNV	NM_002755	c.C371T	p.P124L
4683_S16	chr15	66481793	66481793	G	A	MAP2K1	nonsynonymous SNV	NM_002755	c.G607A	p.E203K
50_S44	chr15	66481793	66481793	G	A	MAP2K1	SNV	NM_002755	c.G607A	p.E203K
WTCHG_76140_34	chr15	92949027	92949027	C	T	CHD2	stopgain	NM_001271	c.C1453T	p.R485X
WTCHG_91609_76	chr15	92967488	92967488	T	G	CHD2	nonsynonymous SNV	NM_001271	c.T2164G	p.S722A
WTCHG_90118_06	chr15	93020049	93020051	TGG	-	CHD2	nonframeshift deletion	NM_001271	c.4944_4946del	p.1648_1649del
WTCHG_88504_42	chr15	93024673	93024673	C	G	CHD2	nonsynonymous SNV	NM_001271	c.C5455G	p.P1819A
WTCHG_76140_24	chr16	3728446	3728446	G	A	CREBBP	stopgain	NM_004380	c.C6601T	p.Q2201X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_91609_94	chr16	3728602	3728602	C	T	CREBBP	nonsynonymous SNV	NM_004380	c.G6445A	p.V2149M
23_MUT	chr16	3729265	3729265	G	A	CREBBP	stopgain	NM_004380	c.C5782T	p.Q1928X
92568_S9	chr16	3729277	3729277	C	T	CREBBP	nonsynonymous SNV	NM_004380	c.G5770A	p.V1924M
WTCHG_87813_05	chr16	3731323	3731325	AGG	-	CREBBP	nonframeshift deletion	NM_004380	c.5039_5041del	p.1680_1681del
6_S6	chr16	3736667	3736667	T	A	CREBBP	nonsynonymous SNV	NM_004380	c.A4543T	p.I1515F
WTCHG_90060_59	chr16	3738604	3738604	T	C	CREBBP	nonsynonymous SNV	NM_004380	c.A4349G	p.Y1450C
WTCHG_90060_61	chr16	3738604	3738604	T	C	CREBBP	nonsynonymous SNV	NM_004380	c.A4349G	p.Y1450C
H2247_S16	chr16	3738650	3738650	C	T	CREBBP	nonsynonymous SNV	NM_004380	c.G4303A	p.D1435N
WTCHG_87814_46	chr16	3739596	3739596	C	T	CREBBP	SNV	NM_004380	c.G4262A	p.C1421Y
WTCHG_76140_20	chr16	3740454	3740454	G	A	CREBBP	stopgain	NM_004380	c.C4078T	p.R1360X
WTCHG_90119_55	chr16	3744942	3744942	A	C	CREBBP	nonsynonymous SNV	NM_004380	c.T3934G	p.L1312V
11_MUT	chr16	3749638	3749647	TAAGGTATCA	-	CREBBP	frameshift deletion	NM_004380	c.3816_3825del	p.N1272fs
WTCHG_90119_46	chr16	3757287	3757287	C	T	CREBBP	splicing	NM_004380	c.3698+1G>A	.
WTCHG_88504_33	chr16	3757813	3757813	C	T	CREBBP	nonsynonymous SNV	NM_004380	c.G3605A	p.R1202H
L080_11	chr16	3758897	3758897	A	C	CREBBP	stopgain	NM_004380	c.T3326G	p.L1109X
WTCHG_90119_47	chr16	3758916	3758916	G	A	CREBBP	stopgain	NM_004380	c.C3307T	p.R1103X
10_S10	chr16	3767881	3767881	G	A	CREBBP	nonsynonymous SNV	NM_004380	c.C3089T	p.S1030F
L080_11	chr16	3770719	3770719	G	A	CREBBP	stopgain	NM_004380	c.C2731T	p.Q911X

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
875_S9	chr16	3770751	3770751	G	A	CREBBP	nonsynonymous SNV	NM_004380	c.C2699T	p.T900I
14_MUT	chr16	3778119	3778119	G	A	CREBBP	nonsynonymous SNV	NM_004380	c.C2005T	p.R669W
50_S44	chr16	3780823	3780823	G	A	CREBBP	nonsynonymous SNV	NM_004380	c.C1732T	p.P578S
WTCHG_90119_46	chr16	3782764	3782764	G	A	CREBBP	SNV	NM_004380	c.C1493T	p.T498M
WTCHG_76140_33	chr16	3793570	3793570	G	-	CREBBP	frameshift deletion	NM_004380	c.1032delC	p.P344fs
8_S8	chr16	3810755	3810755	-	AC	CREBBP	frameshift insertion	NM_004380	c.822_823insGT	p.P275fs
WTCHG_88505_68	chr17	7670682	7670682	C	A	TP53	stopgain	NM_000546	c.G1027T	p.E343X
785_S6	chr17	7673579	7673579	G	A	TP53	stopgain	NM_000546	c.C949T	p.Q317X
51_S45	chr17	7673700	7673700	C	A	TP53	splicing	NM_000546	c.919+1G>T	.
WTCHG_88505_51	chr17	7673733	7673748	TGAGGCTCCCCTTTCT	-	TP53	frameshift deletion	NM_000546	c.872_887del	p.K291fs
L038_08	chr17	7673767	7673767	C	G	TP53	nonsynonymous SNV	NM_000546	c.G853C	p.E285Q
11_MUT	chr17	7673797	7673797	A	-	TP53	frameshift deletion	NM_000546	c.823delT	p.C275fs
H1558_S13	chr17	7673800	7673800	C	A	TP53	nonsynonymous SNV	NM_000546	c.G820T	p.V274F
L027_07	chr17	7673802	7673802	C	G	TP53	nonsynonymous SNV	NM_000546	c.G818C	p.R273P
WTCHG_75645_43	chr17	7673824	7673824	C	T	TP53	nonsynonymous SNV	NM_000546	c.G796A	p.G266R
WTCHG_75645_59	chr17	7674220	7674220	C	G	TP53	nonsynonymous SNV	NM_000546	c.G743C	p.R248P
WTCHG_91608_78	chr17	7674220	7674220	C	T	TP53	SNV	NM_000546	c.G743A	p.R248Q
H1558_S13	chr17	7674226	7674226	A	-	TP53	frameshift deletion	NM_000546	c.737delT	p.M246fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_88505_53	chr17	7674241	7674241	G	A	TP53	nonsynonymous SNV	NM_000546	c.C722T	p.S241F
31_MUT	chr17	7674242	7674242	A	G	TP53	nonsynonymous SNV	NM_000546	c.T721C	p.S241P
WTCHG_88505_69	chr17	7674250	7674250	C	T	TP53	nonsynonymous SNV	NM_000546	c.G713A	p.C238Y
L075_10	chr17	7674256	7674256	T	C	TP53	nonsynonymous SNV	NM_000546	c.A707G	p.Y236C
WTCHG_87813_04	chr17	7674262	7674262	T	C	TP53	SNV	NM_000546	c.A701G	p.Y234C
WTCHG_75645_02	chr17	7674858	7674858	C	T	TP53	splicing	NM_000546	c.672+1G>A	.
H523_S21	chr17	7674859	7674859	C	T	TP53	synonymous SNV	NM_000546	c.G672A	p.E224E
WTCHG_91608_96	chr17	7674890	7674890	T	C	TP53	nonsynonymous SNV	NM_000546	c.A641G	p.H214R
WTCHG_88505_57	chr17	7674894	7674894	G	A	TP53	stopgain	NM_000546	c.C637T	p.R213X
12_S9	chr17	7674904	7674905	TC	-	TP53	frameshift deletion	NM_000546	c.626_627del	p.R209fs
WTCHG_90060_50	chr17	7674904	7674905	TC	-	TP53	frameshift deletion	NM_000546	c.626_627del	p.R209fs
WTCHG_90119_70	chr17	7674916	7674916	A	C	TP53	stopgain	NM_000546	c.T615G	p.Y205X
WTCHG_91609_96	chr17	7674941	7674941	A	C	TP53	nonsynonymous SNV	NM_000546	c.T590G	p.V197G
WTCHG_88505_55	chr17	7674948	7674948	T	A	TP53	nonsynonymous SNV	NM_000546	c.A583T	p.I195F
WTCHG_88504_50	chr17	7674950	7674950	A	C	TP53	nonsynonymous SNV	NM_000546	c.T581G	p.L194R
6_MUT	chr17	7674953	7674953	T	C	TP53	nonsynonymous SNV	NM_000546	c.A578G	p.H193R
WTCHG_88505_55	chr17	7674953	7674953	T	C	TP53	SNV	NM_000546	c.A578G	p.H193R
WTCHG_76140_10	chr17	7674973	7674973	T	A	TP53	splicing	NM_000546	c.560-2A>T	.
WTCHG_91608_95	chr17	7675066	7675066	G	T	TP53	stopgain	NM_000546	c.C546A	p.C182X
WTCHG_91608_74	chr17	7675076	7675076	T	C	TP53	nonsynonymous SNV	NM_000546	c.A536G	p.H179R

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
L031_07	chr17	7675088	7675088	C	T	TP53	nonsynonymous SNV	NM_000546	c.G524A	p.R175H
25_MUT	chr17	7675100	7675100	T	C	TP53	nonsynonymous SNV	NM_000546	c.A512G	p.E171G
785_S6	chr17	7675124	7675124	T	C	TP53	nonsynonymous SNV	NM_000546	c.A488G	p.Y163C
WTCHG_87813_18	chr17	7675157	7675157	G	A	TP53	nonsynonymous SNV	NM_000546	c.C455T	p.P152L
WTCHG_87813_08	chr17	7675216	7675216	C	A	TP53	nonsynonymous SNV	NM_000546	c.G396T	p.K132N
WTCHG_90119_71	chr17	7675216	7675216	C	A	TP53	nonsynonymous SNV	NM_000546	c.G396T	p.K132N
WTCHG_90119_49	chr17	7675217	7675217	T	G	TP53	nonsynonymous SNV	NM_000546	c.A395C	p.K132T
WTCHG_90117_91	chr17	7675233	7675233	A	T	TP53	SNV	NM_000546	c.T379A	p.S127T
WTCHG_88505_72	chr17	7676071	7676071	G	A	TP53	stopgain	NM_000546	c.C298T	p.Q100X
WTCHG_90060_58	chr17	42322333	42322333	C	G	STAT3	nonsynonymous SNV	NM_139276	c.G2050C	p.G684R
92568_S9	chr17	45265203	45265203	C	T	MAP3K14	nonsynonymous SNV	NM_003954	c.G2639A	p.R880Q
WTCHG_91608_91	chr17	45266585	45266585	T	C	MAP3K14	nonsynonymous SNV	NM_003954	c.A2530G	p.M844V
WTCHG_88505_72	chr17	45267439	45267439	C	T	MAP3K14	nonsynonymous SNV	NM_003954	c.G2293A	p.V765I
WTCHG_91609_96	chr17	45271219	45271219	C	A	MAP3K14	nonsynonymous SNV	NM_003954	c.G1660T	p.D554Y
WTCHG_76140_14	chr17	45290495	45290495	G	C	MAP3K14	nonsynonymous SNV	NM_003954	c.C251G	p.A84G
21_S12	chr17	45290507	45290507	A	G	MAP3K14	nonsynonymous SNV	NM_003954	c.T239C	p.I80T
WTCHG_88504_47	chr17	45290507	45290507	A	G	MAP3K14	SNV	NM_003954	c.T239C	p.I80T

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90060_59	chr17	45290546	45290546	G	C	MAP3K14	nonsynonymous SNV	NM_003954	c.C200G	p.A67G
WTCHG_90060_60	chr17	45290546	45290546	G	C	MAP3K14	nonsynonymous SNV	NM_003954	c.C200G	p.A67G
WTCHG_90119_54	chr17	45290546	45290546	G	C	MAP3K14	nonsynonymous SNV	NM_003954	c.C200G	p.A67G
WTCHG_76140_12	chr17	45290640	45290640	G	A	MAP3K14	stopgain	NM_003954	c.C106T	p.Q36X
726_S40	chr17	63929438	63929438	T	C	CD79B	nonsynonymous SNV	NM_000626	c.A587G	p.Y196C
L051_09	chr17	63929438	63929438	T	G	CD79B	nonsynonymous SNV	NM_000626	c.A587C	p.Y196S
L052_09	chr17	63929476	63929476	C	T	CD79B	splicing	NM_000626	c.550-1G>A	.
WTCHG_91609_76	chr17	63929769	63929769	C	T	CD79B	splicing	NM_000626	c.549+1G>A	.
32_S32	chr19	1615471	1615471	G	A	TCF3	nonsynonymous SNV	NM_003200	c.C1636T	p.R546W
30_S24	chr19	1625666	1625666	G	A	TCF3	nonsynonymous SNV	NM_003200	c.C409T	p.P137S
WTCHG_76140_18	chr19	16324972	16324972	-	CA	KLF2	frameshift insertion	NM_016270	c.49_50insCA	p.P17fs
22_S13	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
491_S11	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
51640_S3	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
836_S20	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
WTCHG_88504_31	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
WTCHG_90060_65	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
WTCHG_90119_67	chr19	16324993	16324993	C	T	KLF2	stopgain	NM_016270	c.C70T	p.Q24X
WTCHG_90117_77	chr19	16325234	16325234	C	-	KLF2	frameshift deletion	NM_016270	c.94delC	p.P32fs
WTCHG_91609_14	chr19	16325244	16325250	GCGGCAC	-	KLF2	frameshift deletion	NM_016270	c.104_110del	p.G35fs
WTCHG_90060_53	chr19	16325261	16325267	CTCAACA	-	KLF2	frameshift deletion	NM_016270	c.121_127del	p.L41fs
WTCHG_90060_59	chr19	16325588	16325588	G	-	KLF2	frameshift deletion	NM_016270	c.448delG	p.E150fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
8_MUT	chr19	16325768	16325768	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C628T	p.H210Y
WTCHG_90119_52	chr19	16325909	16325909	C	G	KLF2	nonsynonymous SNV	NM_016270	c.C769G	p.P257A
10_MUT	chr19	16325932	16325932	G	A	KLF2	stopgain	NM_016270	c.G792A	p.W264X
H2265_S17	chr19	16325961	16325961	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G821A	p.C274Y
WTCHG_87813_25	chr19	16325961	16325961	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G821A	p.C274Y
5_MUT	chr19	16325973	16325973	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G833A	p.G278D
WTCHG_90117_77	chr19	16325984	16325989	ACCTAC	-	KLF2	nonframeshift deletion	NM_016270	c.844_849del	p.282_283del
L043_08	chr19	16325987	16325987	T	G	KLF2	nonsynonymous SNV	NM_016270	c.T847G	p.Y283D
52_S46	chr19	16326000	16326000	C	A	KLF2	stopgain	NM_016270	c.C860A	p.S287X
2_S2	chr19	16326002	16326002	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C862T	p.H288Y
21_S12	chr19	16326002	16326002	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C862T	p.H288Y
L076_10	chr19	16326002	16326002	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C862T	p.H288Y
L096_13	chr19	16326002	16326002	C	G	KLF2	nonsynonymous SNV	NM_016270	c.C862G	p.H288D
WTCHG_75645_06	chr19	16326002	16326002	C	G	KLF2	nonsynonymous SNV	NM_016270	c.C862G	p.H288D
WTCHG_90060_53	chr19	16326002	16326002	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C862T	p.H288Y
L011_03	chr19	16326011	16326011	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G871A	p.A291T
WTCHG_76140_33	chr19	16326012	16326012	C	T	KLF2	SNV	NM_016270	c.C872T	p.A291V

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
L018_06	chr19	16326026	16326026	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C886T	p.H296Y
WTCHG_90117_85	chr19	16326026	16326026	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C886T	p.H296Y
30_S24	chr19	16326032	16326032	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G892A	p.G298S
WTCHG_90060_55	chr19	16326855	16326855	G	-	KLF2	splicing	NM_016270	c.893-1G>-	.
WTCHG_87813_18	chr19	16326856	16326856	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G893A	p.G298D
L031_07	chr19	16326860	16326860	-	A	KLF2	frameshift insertion	NM_016270	c.898dupA	p.E299fs
WTCHG_75645_62	chr19	16326886	16326886	G	T	KLF2	nonsynonymous SNV	NM_016270	c.G923T	p.G308V
WTCHG_76140_37	chr19	16326910	16326910	C	G	KLF2	stopgain	NM_016270	c.C947G	p.S316X
L076_10	chr19	16326932	16326932	C	-	KLF2	frameshift deletion	NM_016270	c.969delC	p.Y323fs
WTCHG_90060_57	chr19	16326933	16326933	C	T	KLF2	stopgain	NM_016270	c.C970T	p.R324X
WTCHG_87813_25	chr19	16326994	16326994	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C1031T	p.S344F
WTCHG_90060_65	chr19	16326996	16326996	G	A	KLF2	nonsynonymous SNV	NM_016270	c.G1033A	p.D345N
H2247_S16	chr19	16327006	16327009	CGCT	-	KLF2	frameshift deletion	NM_016270	c.1043_1046del	p.A348fs
WTCHG_90060_61	chr19	16327011	16327011	C	T	KLF2	nonsynonymous SNV	NM_016270	c.C1048T	p.H350Y
WTCHG_75645_07	chr19	16327029	16327029	T	A	KLF2	stoploss	NM_016270	c.T1066A	p.X356K
WTCHG_90117_91	chr19	17830532	17830532	A	G	JAK3	nonsynonymous SNV	NM_000215	c.T3067C	p.Y1023H
10_S10	chr19	17832858	17832858	G	A	JAK3	nonsynonymous SNV	NM_000215	c.C2422T	p.L808F
WTCHG_90119_54	chr19	17837170	17837170	C	T	JAK3	nonsynonymous SNV	NM_000215	c.G1745A	p.R582Q
WTCHG_87814_47	chr19	17839584	17839584	C	T	JAK3	nonsynonymous SNV	NM_000215	c.G1334A	p.R445Q

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_91608_73	chr19	17840288	17840288	T	C	JAK3	nonsynonymous SNV	NM_000215	c.A1196G	p.Y399C
L094_13	chr19	41879574	41879574	C	A	CD79A	nonsynonymous SNV	NM_001783	c.C419A	p.T140N
WTCHG_76140_12	chr2	136114915	136114915	G	C	CXCR4	stopgain	NM_003467	c.C1013G	p.S338X
WTCHG_90119_48	chr2	136114922	136114922	C	A	CXCR4	stopgain	NM_003467	c.G1006T	p.G336X
8_MUT	chr2	136114928	136114928	G	A	CXCR4	stopgain	NM_003467	c.C1000T	p.R334X
WTCHG_88505_54	chr2	136114928	136114928	G	A	CXCR4	stopgain	NM_003467	c.C1000T	p.R334X
1_S1	chr2	136114953	136114953	-	CTAG ATC	CXCR4	stopgain	NM_003467	c.974_975insGATCTAG	p.S325_L326delinsRIX
49_S43	chr2	136114964	136114964	-	G	CXCR4	frameshift insertion	NM_003467	c.963dupC	p.R322fs
WTCHG_90119_48	chr2	136115453	136115453	C	T	CXCR4	nonsynonymous SNV	NM_003467	c.G475A	p.G159S
606_S12	chr22	22888063	22888063	A	T	IGLL5	stopgain	NM_001178126	c.A10T	p.K4X
L048_09	chr22	22888123	22888123	T	A	IGLL5	nonsynonymous SNV	NM_001178126	c.T70A	p.W24R
L086_12	chr22	22888130	22888130	T	C	IGLL5	nonsynonymous SNV	NM_001178126	c.T77C	p.L26P
L40_S15	chr22	22888183	22888183	G	T	IGLL5	nonsynonymous SNV	NM_001178126	c.G130T	p.A44S
L086_12	chr22	22888188	22888192	A	GCAA	IGLL5	frameshift deletion	NM_001178126	c.135_139del	p.P45fs
WTCHG_91608_78	chr3	38140527	38140527	C	T	MYD88	synonymous SNV	NM_002468	c.C603T	p.G201G
L044_08	chr3	38140534	38140534	G	T	MYD88	nonsynonymous SNV	NM_002468	c.G610T	p.V204F
WTCHG_76140_33	chr3	38140534	38140534	G	T	MYD88	nonsynonymous SNV	NM_002468	c.G610T	p.V204F
WTCHG_88505_56	chr3	38140534	38140534	G	T	MYD88	nonsynonymous SNV	NM_002468	c.G610T	p.V204F
17_MUT	chr3	38140541	38140541	C	G	MYD88	nonsynonymous SNV	NM_002468	c.C617G	p.S206C

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
5_S5	chr3	38140541	38140541	C	G	MYD88	nonsynonymous SNV	NM_002468	c.C617G	p.S206C
673_S4	chr3	38140541	38140541	C	G	MYD88	nonsynonymous SNV	NM_002468	c.C617G	p.S206C
L071_10	chr3	38140541	38140541	C	G	MYD88	nonsynonymous SNV	NM_002468	c.C617G	p.S206C
WTCHG_87813_03	chr3	38140541	38140541	C	G	MYD88	nonsynonymous SNV	NM_002468	c.C617G	p.S206C
WTCHG_90117_82	chr3	38140541	38140541	C	G	MYD88	nonsynonymous SNV	NM_002468	c.C617G	p.S206C
27_S27	chr3	38140544	38140544	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T620C	p.I207T
27_S27	chr3	38140566	38140566	G	C	MYD88	nonsynonymous SNV	NM_002468	c.G642C	p.K214N
47_S41	chr3	38140768	38140768	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T656C	p.M219T
L029_07	chr3	38140768	38140768	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T656C	p.M219T
WTCHG_88505_74	chr3	38140768	38140768	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T656C	p.M219T
4683_S16	chr3	38140801	38140801	G	A	MYD88	nonsynonymous SNV	NM_002468	c.G689A	p.S230N
L094_13	chr3	38140801	38140801	G	A	MYD88	nonsynonymous SNV	NM_002468	c.G689A	p.S230N
51_S45	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
793_S7	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
8_S8	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
875_S9	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
L019_06	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
L022_06	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
L023_07	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
L051_09	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
L099_13	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_75645_05	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_76140_17	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_76140_19	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_88504_28	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_90117_78	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_90117_79	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
WTCHG_91609_11	chr3	38141150	38141150	T	C	MYD88	nonsynonymous SNV	NM_002468	c.T755C	p.L252P
L049_09_S31	chr3	47017128	47017128	C	T	SETD2	SNV	NM_014159	c.G7660A	p.V2554I
1_S1	chr3	47088112	47088112	C	T	SETD2	splicing	NM_014159	c.5277+1G>A	.
1_S1	chr3	47097968	47097968	C	-	SETD2	frameshift deletion	NM_014159	c.5129delG	p.R1710fs
785_S6	chr3	47103422	47103422	A	T	SETD2	nonsynonymous SNV	NM_014159	c.T4841A	p.I1614K
WTCHG_76140_12	chr3	47113921	47113924	GTGA	-	SETD2	frameshift deletion	NM_014159	c.4667_4670del	p.L1556fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_87813_18	chr3	47120292	47120293	GG	-	SETD2	frameshift deletion	NM_014159	c.4343_4344del	p.S1448fs
11731_S18	chr3	47120299	47120299	C	T	SETD2	nonsynonymous SNV	NM_014159	c.G4337A	p.G1446E
L038_08	chr3	47120375	47120375	G	A	SETD2	nonsynonymous SNV	NM_014159	c.C4261T	p.L1421F
WTCHG_91609_11	chr3	47121842	47121842	C	T	SETD2	nonsynonymous SNV	NM_014159	c.G2794A	p.V932I
9641_S17	chr3	47122448	47122448	T	C	SETD2	nonsynonymous SNV	NM_014159	c.A2188G	p.K730E
WTCHG_90117_87	chr3	47123437	47123437	C	T	SETD2	nonsynonymous SNV	NM_014159	c.G1199A	p.R400Q
WTCHG_88505_61	chr3	47124488	47124488	C	T	SETD2	nonsynonymous SNV	NM_014159	c.G148A	p.A50T
WTCHG_75645_08	chr4	105234031	105234031	T	C	TET2	nonsynonymous SNV	NM_001127208	c.T89C	p.L30P
WTCHG_75645_62	chr4	105234952	105234952	A	C	TET2	nonsynonymous SNV	NM_001127208	c.A1010C	p.E337A
WTCHG_88504_45	chr4	105235321	105235321	C	T	TET2	nonsynonymous SNV	NM_001127208	c.C1379T	p.S460F
WTCHG_90117_96	chr4	105235928	105235928	C	G	TET2	nonsynonymous SNV	NM_001127208	c.C1986G	p.F662L
L075_10	chr4	105236572	105236572	A	G	TET2	SNV	NM_001127208	c.A2630G	p.D877G
WTCHG_75645_60	chr4	105237039	105237039	A	T	TET2	stopgain	NM_001127208	c.A3097T	p.K1033X
L043_08	chr4	105237181	105237181	C	T	TET2	nonsynonymous SNV	NM_001127208	c.C3239T	p.A1080V
3_S3	chr4	105259621	105259621	G	A	TET2	nonsynonymous SNV	NM_001127208	c.G3806A	p.R1269K
WTCHG_90119_49	chr4	105259628	105259628	C	G	TET2	nonsynonymous SNV	NM_001127208	c.C3813G	p.C1271W

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
L022_06	chr4	152326166	152326166	T	C	FBXW7	nonsynonymous SNV	NM_001349798	c.A1484G	p.H495R
726_S40	chr4	152326215	152326215	G	C	FBXW7	nonsynonymous SNV	NM_001349798	c.C1435G	p.R479G
47_S41	chr4	152350051	152350051	T	G	FBXW7	nonsynonymous SNV	NM_001349798	c.A575C	p.E192A
WTCHG_75645_06	chr6	41935945	41935951	GGTGTAT	-	CCND3	frameshift deletion	NM_001760	c.868_874del	p.I290fs
WTCHG_87813_02	chr6	41935950	41935950	A	T	CCND3	nonsynonymous SNV	NM_001760	c.T869A	p.I290K
WTCHG_88504_42	chr6	41935953	41935953	-	CT	CCND3	frameshift insertion	NM_001760	c.865_866insAG	p.A289fs
12_S9	chr6	41935954	41935954	C	G	CCND3	nonsynonymous SNV	NM_001760	c.G865C	p.A289P
WTCHG_90119_49	chr6	41935954	41935954	C	G	CCND3	nonsynonymous SNV	NM_001760	c.G865C	p.A289P
WTCHG_91609_91	chr6	41935954	41935954	C	G	CCND3	nonsynonymous SNV	NM_001760	c.G865C	p.A289P
WTCHG_75645_30	chr6	41935955	41935956	TGGTGACATA	-	CCND3	frameshift deletion	NM_001760	c.863_864del	p.T288fs
L088_12	chr6	41935956	41935969	GG	-	CCND3	frameshift deletion	NM_001760	c.850_863del	p.P284fs
27_MUT	chr6	41935959	41935959	A	T	CCND3	nonsynonymous SNV	NM_001760	c.T860A	p.V287D
3_S3	chr6	41935959	41935959	A	C	CCND3	nonsynonymous SNV	NM_001760	c.T860G	p.V287G
45_S45	chr6	41935959	41935959	A	C	CCND3	nonsynonymous SNV	NM_001760	c.T860G	p.V287G
WTCHG_90119_71	chr6	41935965	41935968	GTAG	-	CCND3	frameshift deletion	NM_001760	c.851_854del	p.P284fs
WTCHG_88505_70	chr6	41935967	41935967	-	GGAGTGC	CCND3	frameshift insertion	NM_001760	c.851_852insAAGCCCAGCCAGACCAGCACTCC	p.P284fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
					CTGG CTGG GCTT					
L031_07	chr6	41935968	41935968	G	T	CCND3	nonsynonymous SNV	NM_001760	c.C851A	p.P284H
623_S55	chr6	41935969	41935969	G	C	CCND3	nonsynonymous SNV	NM_001760	c.C850G	p.P284A
9_S9	chr6	41935969	41935969	G	A	CCND3	nonsynonymous SNV	NM_001760	c.C850T	p.P284S
L099_13	chr6	41935969	41935969	G	A	CCND3	nonsynonymous SNV	NM_001760	c.C850T	p.P284S
WTCHG_87813_22	chr6	41935969	41935969	G	A	CCND3	nonsynonymous SNV	NM_001760	c.C850T	p.P284S
WTCHG_87813_25	chr6	41935969	41935969	G	A	CCND3	nonsynonymous SNV	NM_001760	c.C850T	p.P284S
WTCHG_88504_41	chr6	41935969	41935969	G	T	CCND3	nonsynonymous SNV	NM_001760	c.C850A	p.P284T
L043_08	chr6	41935971	41935971	G	T	CCND3	nonsynonymous SNV	NM_001760	c.C848A	p.T283N
L023_07	chr6	41935972	41935972	T	C	CCND3	nonsynonymous SNV	NM_001760	c.A847G	p.T283A
WTCHG_90060_665_MUT	chr6	41935972	41935972	T	C	CCND3	nonsynonymous SNV	NM_001760	c.A847G	p.T283A
4_S3	chr6	41935981	41935981	G	A	CCND3	stopgain	NM_001760	c.C838T	p.Q280X
H523_S21	chr6	41936007	41936007	-	G	CCND3	frameshift insertion	NM_001760	c.811dupC	p.R271fs
WTCHG_87814_48	chr6	41936007	41936007	-	G	CCND3	frameshift insertion	NM_001760	c.811dupC	p.R271fs
WTCHG_90118_08	chr6	41936007	41936007	-	G	CCND3	frameshift insertion	NM_001760	c.811dupC	p.R271fs
WTCHG_75645_02	chr6	41936008	41936008	G	-	CCND3	frameshift deletion	NM_001760	c.811delC	p.R271fs
46_S40	chr6	41936008	41936009	GG	-	CCND3	frameshift deletion	NM_001760	c.810_811del	p.P270fs
L023_07	chr6	41936016	41936016	T	C	CCND3	nonsynonymous SNV	NM_001760	c.A803G	p.K268R

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
L038_08	chr6	41936017	41936017	T	A	CCND3	stopgain	NM_001760	c.A802T	p.K268X
WTCHG_91608_95	chr6	41936041	41936041	G	A	CCND3	stopgain	NM_001760	c.C778T	p.Q260X
L075_10	chr6	41936656	41936656	G	A	CCND3	nonsynonymous SNV	NM_001760	c.C614T	p.T205M
WTCHG_76140_20	chr6	41941472	41941472	G	C	CCND3	nonsynonymous SNV	NM_001760	c.C178G	p.L60V
WTCHG_76140_20	chr6	41941511	41941511	A	T	CCND3	nonsynonymous SNV	NM_001760	c.T139A	p.C47S
WTCHG_87813_09	chr6	41941544	41941544	C	T	CCND3	SNV	NM_001760	c.G106A	p.E36K
92568_S9	chr6	44260042	44260042	C	T	NFKBIE	splicing	NM_004556	c.1437+1G>A	.
WTCHG_91609_96	chr6	44260512	44260512	-	G	NFKBIE	frameshift insertion	NM_004556	c.1135dupC	p.L379fs
783_S12	chr6	44261697	44261697	A	T	NFKBIE	nonsynonymous SNV	NM_004556	c.T1037A	p.L346Q
10_MUT	chr6	44262560	44262560	C	T	NFKBIE	synonymous SNV	NM_004556	c.G885A	p.Q295Q
10_MUT	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
21_S12	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
30_MUT	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
L036_08	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
L067_10	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
WTCHG_88505_68	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
WTCHG_90117_91	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
WTCHG_91609_96	chr6	44265002	44265005	GTAA	-	NFKBIE	frameshift deletion	NM_004556	c.759_762del	p.T253fs
WTCHG_90118_13	chr6	137871291	137871291	C	-	TNFAIP3	frameshift deletion	NM_001270508	c.64delC	p.R22fs
WTCHG_87813_25	chr6	137871402	137871403	CA	-	TNFAIP3	frameshift deletion	NM_001270508	c.175_176del	p.Q59fs
7572_S1	chr6	137871517	137871518	CG	-	TNFAIP3	frameshift deletion	NM_001270508	c.290_291del	p.T97fs
WTCHG_88505_73	chr6	137871522	137871526	GGTA	-	TNFAIP3	frameshift deletion	NM_001270508	c.295_295del	p.G99fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
51640_S3	chr6	137874877	137874877	C	T	TNFAIP3	stopgain	NM_001270508	c.C328T	p.Q110X
WTCHG_87813_04	chr6	137874914	137874914	T	C	TNFAIP3	nonsynonymous SNV	NM_001270508	c.T365C	p.L122P
WTCHG_76140_18	chr6	137874948	137874949	AG	-	TNFAIP3	frameshift deletion	NM_001270508	c.399_400del	p.T133fs
WTCHG_87813_25	chr6	137874950	137874951	AC	-	TNFAIP3	frameshift deletion	NM_001270508	c.401_402del	p.D134fs
L031_07	chr6	137874976	137874976	C	T	TNFAIP3	stopgain	NM_001270508	c.C427T	p.Q143X
10_MUT	chr6	137874985	137874986	TC	-	TNFAIP3	frameshift deletion	NM_001270508	c.436_437del	p.S146fs
WTCHG_88504_31	chr6	137874985	137874986	TC	-	TNFAIP3	frameshift deletion	NM_001270508	c.436_437del	p.S146fs
WTCHG_90119_63	chr6	137875724	137875724	G	T	TNFAIP3	nonsynonymous SNV	NM_001270508	c.G523T	p.A175S
WTCHG_90119_63	chr6	137875725	137875725	C	T	TNFAIP3	nonsynonymous SNV	NM_001270508	c.C524T	p.A175V
WTCHG_87813_04	chr6	137876020	137876020	C	G	TNFAIP3	stopgain	NM_001270508	c.C659G	p.S220X
WTCHG_88504_31	chr6	137876098	137876101	ACCC	-	TNFAIP3	frameshift deletion	NM_001270508	c.737_740del	p.Y246fs
92568_S9	chr6	137876099	137876099	C	-	TNFAIP3	frameshift deletion	NM_001270508	c.738delC	p.Y246fs
H2247_S16	chr6	137876128	137876128	-	T	TNFAIP3	frameshift insertion	NM_001270508	c.768dupT	p.H256fs
L037_08	chr6	137877081	137877081	C	T	TNFAIP3	stopgain	NM_001270508	c.C811T	p.R271X
WTCHG_75645_60	chr6	137877126	137877139	TTAAAAGTTCACTT	-	TNFAIP3	frameshift deletion	NM_001270508	c.856_869del	p.L286fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
L096_13	chr6	137877180	137877180	A	T	TNFAIP3	stopgain	NM_001270508	c.A910T	p.K304X
5_MUT	chr6	137877221	137877221	G	A	TNFAIP3	stopgain	NM_001270508	c.G951A	p.W317X
WTCHG_91608_73	chr6	137877221	137877221	G	A	TNFAIP3	stopgain	NM_001270508	c.G951A	p.W317X
23_MUT	chr6	137877235	137877239	CTCAT	-	TNFAIP3	frameshift deletion	NM_001270508	c.965_969del	p.T322fs
L037_08	chr6	137877235	137877239	CTCAT	-	TNFAIP3	frameshift deletion	NM_001270508	c.965_969del	p.T322fs
WTCHG_91608_73	chr6	137877235	137877239	CTCAT	-	TNFAIP3	frameshift deletion	NM_001270508	c.965_969del	p.T322fs
WTCHG_90060_65	chr6	137878526	137878526	G	T	TNFAIP3	stopgain	NM_001270508	c.G1081T	p.E361X
L096_13	chr6	137878668	137878668	C	G	TNFAIP3	stopgain	NM_001270508	c.C1223G	p.S408X
WTCHG_90060_61	chr6	137878788	137878788	G	A	TNFAIP3	stopgain	NM_001270508	c.G1343A	p.W448X
WTCHG_75645_60	chr6	137878859	137878859	G	T	TNFAIP3	stopgain	NM_001270508	c.G1414T	p.E472X
WTCHG_90060_65	chr6	137878861	137878861	G	-	TNFAIP3	frameshift deletion	NM_001270508	c.1416delG	p.E472fs
46_S40	chr6	137878941	137878941	A	G	TNFAIP3	nonsynonymous SNV	NM_001270508	c.A1496G	p.H499R
7572_S1	chr6	137879126	137879126	C	T	TNFAIP3	stopgain	NM_001270508	c.C1681T	p.Q561X
WTCHG_90060_61	chr6	137879126	137879126	C	T	TNFAIP3	stopgain	NM_001270508	c.C1681T	p.Q561X
3_S3	chr6	137879181	137879181	G	A	TNFAIP3	nonsynonymous SNV	NM_001270508	c.G1736A	p.C579Y
WTCHG_90060_63	chr6	137880200	137880200	T	C	TNFAIP3	nonsynonymous SNV	NM_001270508	c.T2036C	p.I679T

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
17_MUT	chr6	137880241	137880241	G	C	TNFAIP3	nonsynonymous SNV	NM_001270508	c.G2077C	p.E693Q
H523_S21	chr6	137881137	137881137	G	T	TNFAIP3	stopgain	NM_001270508	c.G2191T	p.E731X
WTCHG_76140_18	chr6	137881213	137881213	-	C	TNFAIP3	frameshift insertion	NM_001270508	c.2268dupC	p.D756fs
12600_S19	chr6	137881228	137881228	G	A	TNFAIP3	nonsynonymous SNV	NM_001270508	c.G2282A	p.R761H
WTCHG_91608_83	chr7	2913386	2913386	G	A	CARD11	nonsynonymous SNV	NM_032415	c.C2920T	p.R974C
L025_07	chr7	2928689	2928689	G	A	CARD11	nonsynonymous SNV	NM_032415	c.C1663T	p.R555W
L094_13	chr7	2934479	2934479	G	A	CARD11	nonsynonymous SNV	NM_032415	c.C1492T	p.R498C
H2247_S16	chr7	2937968	2937968	T	C	CARD11	nonsynonymous SNV	NM_032415	c.A1082G	p.Y361C
WTCHG_87813_05	chr7	2937972	2937972	T	C	CARD11	nonsynonymous SNV	NM_032415	c.A1078G	p.M360V
WTCHG_90119_53	chr7	2938686	2938686	C	T	CARD11	nonsynonymous SNV	NM_032415	c.G1010A	p.R337Q
WTCHG_75645_62	chr7	2939861	2939861	A	G	CARD11	nonsynonymous SNV	NM_032415	c.T752C	p.L251P
WTCHG_90060_57	chr7	2939910	2939912	GGT	-	CARD11	nonframeshift deletion	NM_032415	c.701_703del	p.234_235del
53_S47	chr7	2939913	2939918	GCTTTA	-	CARD11	nonframeshift deletion	NM_032415	c.695_700del	p.232_234del
WTCHG_87813_04	chr7	2944324	2944324	T	C	CARD11	nonsynonymous SNV	NM_032415	c.A572G	p.N191S
H1701_S14	chr7	2944528	2944528	C	T	CARD11	nonsynonymous SNV	NM_032415	c.G368A	p.G123D
WTCHG_87813_18	chr7	2947734	2947734	C	T	CARD11	nonsynonymous SNV	NM_032415	c.G61A	p.A21T

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
836_S20	chr7	124824065	124824065	G	A	POT1	nonsynonymous SNV	NM_015450	c.C1802T	p.P601L
WTCHG_75645_30	chr7	124835294	124835294	G	A	POT1	nonsynonymous SNV	NM_015450	c.C1490T	p.T497I
WTCHG_87813_22	chr7	124851950	124851950	C	T	POT1	nonsynonymous SNV	NM_015450	c.G871A	p.D291N
L043_08	chr7	128835520	128835520	C	A	FLNC	nonsynonymous SNV	NM_001458	c.C547A	p.R183S
4683_S16	chr7	128838313	128838313	A	G	FLNC	nonsynonymous SNV	NM_001458	c.A1094G	p.E365G
WTCHG_87814_48	chr7	128838650	128838650	C	T	FLNC	nonsynonymous SNV	NM_001458	c.C1258T	p.R420W
WTCHG_90060_55	chr7	128841524	128841524	A	C	FLNC	nonsynonymous SNV	NM_001458	c.A2078C	p.D693A
L024_07	chr7	128842289	128842289	G	A	FLNC	nonsynonymous SNV	NM_001458	c.G2180A	p.R727H
WTCHG_76140_36	chr7	128844207	128844207	C	A	FLNC	nonsynonymous SNV	NM_001458	c.C3133A	p.H1045N
WTCHG_90060_60	chr7	128844253	128844253	C	T	FLNC	nonsynonymous SNV	NM_001458	c.C3179T	p.P1060L
WTCHG_88504_45	chr7	128844847	128844847	G	A	FLNC	nonsynonymous SNV	NM_001458	c.G3382A	p.E1128K
WTCHG_90119_69	chr7	128844965	128844965	G	A	FLNC	nonsynonymous SNV	NM_001458	c.G3500A	p.R1167H
28_S22	chr7	128845255	128845255	G	A	FLNC	nonsynonymous SNV	NM_001458	c.G3790A	p.G1264S
785_S6	chr7	128846428	128846428	G	C	FLNC	nonsynonymous SNV	NM_001458	c.G4092C	p.L1364F
WTCHG_88504_45	chr7	128848818	128848818	C	G	FLNC	nonsynonymous SNV	NM_001458	c.C4763G	p.A1588G
52_S46	chr7	128852998	128852998	G	A	FLNC	nonsynonymous SNV	NM_001458	c.G6175A	p.V2059M

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_91608_95	chr7	128856555	128856555	C	T	FLNC	nonsynonymous SNV	NM_001458	c.C7289T	p.A2430V
WTCHG_90119_53	chr7	128857145	128857145	C	A	FLNC	nonsynonymous SNV	NM_001458	c.C7589A	p.T2530N
L069_10	chr7	140753332	140753332	T	G	BRAF	nonsynonymous SNV	NM_004333	c.A1803C	p.K601N
WTCHG_88505_73	chr7	140753332	140753332	T	G	BRAF	nonsynonymous SNV	NM_004333	c.A1803C	p.K601N
WTCHG_90117_91	chr7	148811636	148811636	A	T	EZH2	nonsynonymous SNV	NM_004456	c.T1936A	p.Y646N
5_S5	chr8	47777835	47777835	G	A	PRKDC	nonsynonymous SNV	NM_006904	c.C11893T	p.R3965C
WTCHG_76140_33	chr8	47778616	47778616	G	A	PRKDC	nonsynonymous SNV	NM_006904	c.C11696T	p.A3899V
28_S22	chr8	47782211	47782211	C	T	PRKDC	nonsynonymous SNV	NM_006904	c.G11440A	p.D3814N
WTCHG_90117_92	chr8	47834219	47834219	T	C	PRKDC	nonsynonymous SNV	NM_006904	c.A8129G	p.D2710G
2_S2	chr8	47836496	47836496	C	T	PRKDC	nonsynonymous SNV	NM_006904	c.G7793A	p.R2598Q
92568_S9	chr8	47840115	47840115	C	T	PRKDC	nonsynonymous SNV	NM_006904	c.G7355A	p.R2452Q
WTCHG_88504_36	chr8	47862474	47862474	A	G	PRKDC	nonsynonymous SNV	NM_006904	c.T5818C	p.Y1940H
46_S40	chr8	47877748	47877748	G	T	PRKDC	nonsynonymous SNV	NM_006904	c.C5339A	p.S1780Y
92568_S9	chr8	47879606	47879606	A	T	PRKDC	nonsynonymous SNV	NM_006904	c.T5120A	p.L1707Q
WTCHG_90119_69	chr8	47888567	47888567	C	T	PRKDC	nonsynonymous SNV	NM_006904	c.G4364A	p.C1455Y
WTCHG_90117_96	chr8	47888650	47888650	G	C	PRKDC	nonsynonymous SNV	NM_006904	c.C4281G	p.S1427R

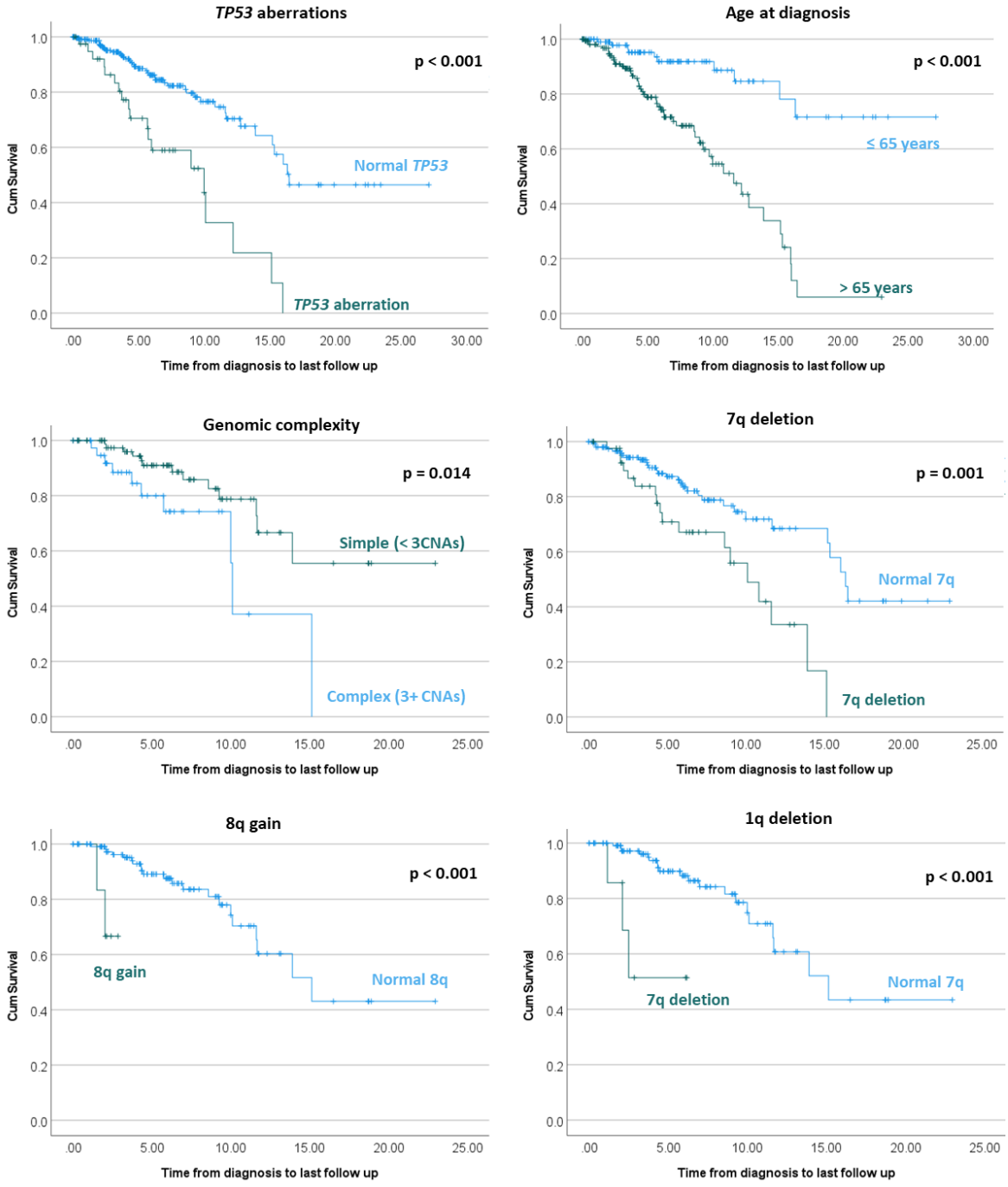
sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_76140_10	chr8	47902794	47902794	T	C	PRKDC	nonsynonymous SNV	NM_006904	c.A3044G	p.D1015G
WTCHG_76140_13	chr8	47914015	47914015	C	G	PRKDC	nonsynonymous SNV	NM_006904	c.G2667C	p.E889D
WTCHG_76140_20	chr8	47914015	47914015	C	G	PRKDC	nonsynonymous SNV	NM_006904	c.G2667C	p.E889D
WTCHG_88505_72	chr8	47914032	47914032	C	T	PRKDC	nonsynonymous SNV	NM_006904	c.G2650A	p.V884M
WTCHG_75645_30	chr8	47930738	47930738	G	A	PRKDC	nonsynonymous SNV	NM_006904	c.C1826T	p.A609V
WTCHG_88505_56	chr8	47943975	47943975	T	C	PRKDC	nonsynonymous SNV	NM_006904	c.A776G	p.Q259R
18_MUT	chr9	8404642	8404642	G	C	PTPRD	nonsynonymous SNV	NM_002839	c.C4105G	p.Q1369E
4_S3	chr9	8436641	8436641	A	G	PTPRD	nonsynonymous SNV	NM_002839	c.T4037C	p.I1346T
H1954_S15	chr9	8499772	8499772	G	A	PTPRD	nonsynonymous SNV	NM_002839	c.C2197T	p.R733C
41_S35	chr9	8528674	8528674	G	A	PTPRD	SNV	NM_002839	c.C458T	p.P153L
H3829_S19	chr9	136495693	136495693	T	C	NOTCH1	UTR3	NM_017617	c.*378A>G	.
L076_10	chr9	136496133	136496133	C	T	NOTCH1	nonsynonymous SNV	NM_017617	c.G7606A	p.V2536I
10_S10	chr9	136496197	136496198	AG	-	NOTCH1	frameshift deletion	NM_017617	c.7541_7542del	p.P2514fs
51640_S3	chr9	136496197	136496198	AG	-	NOTCH1	frameshift deletion	NM_017617	c.7541_7542del	p.P2514fs
L024_07	chr9	136496197	136496198	AG	-	NOTCH1	frameshift deletion	NM_017617	c.7541_7542del	p.P2514fs
L031_07	chr9	136496197	136496198	AG	-	NOTCH1	frameshift deletion	NM_017617	c.7541_7542del	p.P2514fs
L096_13	chr9	136496197	136496198	AG	-	NOTCH1	frameshift deletion	NM_017617	c.7541_7542del	p.P2514fs

sample ID	Chromosome	Start location	End location	Reference allele	Alternate allele	Gene	Function	Transcript	cDNA change	Amino acid change
WTCHG_90117_802_MUT	chr9	136496197	136496198	AG	-	NOTCH1	frameshift deletion	NM_017617	c.7541_7542del	p.P2514fs
WTCHG_88505_59	chr9	136496232	136496232	G	A	NOTCH1	stopgain	NM_017617	c.C7507T	p.Q2503X
WTCHG_75645_28	chr9	136496264	136496264	G	T	NOTCH1	stopgain	NM_017617	c.C7475A	p.S2492X
WTCHG_75645_08	chr9	136496307	136496310	CGGT	-	NOTCH1	frameshift deletion	NM_017617	c.7429_7432del	p.T2477fs
11731_S18	chr9	136496547	136496547	G	A	NOTCH1	stopgain	NM_017617	c.C7192T	p.Q2398X
7_S7	chr9	136496913	136496913	G	-	NOTCH1	frameshift deletion	NM_017617	c.6826delC	p.L2276fs
	chr9	136497455	136497455	C	T	NOTCH1	nonsynonymous SNV	NM_017617	c.G6284A	p.R2095H

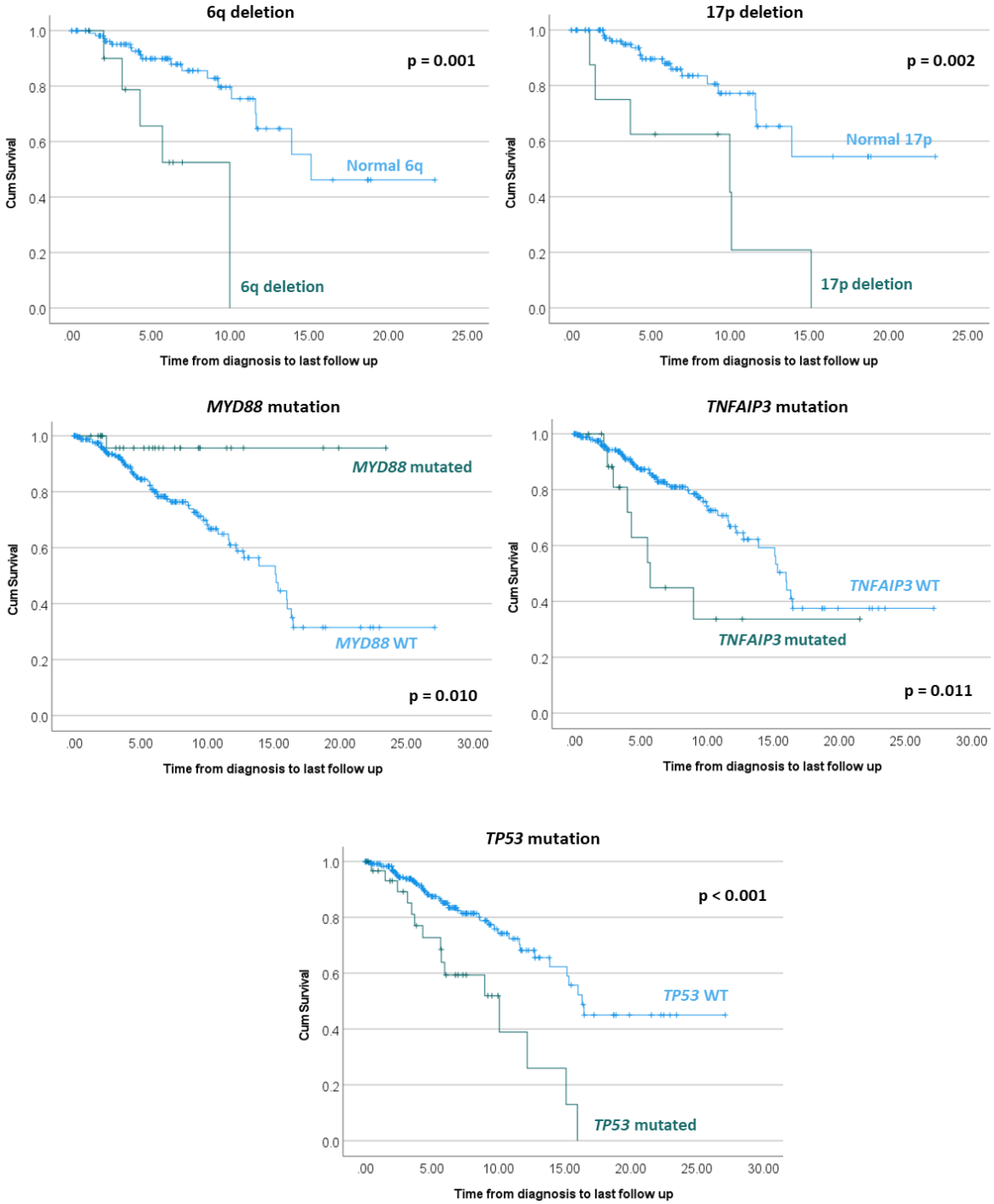
Supplementary Table 8. List of *IGHV* genes identified within the Jaramillo-Parry cohort.

***IGVH* gene repertoire**		
IGHV 3-15	*IGHV1-8*01*	*IGHV3-72*01*
*IGHV 3-15*01*	*IGHV2-5*10*	*IGHV3-73*
*IGHV 3-15*07*	*IGHV3-07*	*IGHV3-74*
*IGHV 3-21*01*	*IGHV3-11*01*	*IGHV3-74*01*
IGHV 3-23	*IGHV3-15*	*IGHV3-9*
*IGHV 3-23*01*	*IGHV3-15*01*	*IGHV4-30*2*
*IGHV 3-23*05*	*IGHV3-15*07*	*IGHV4-30-4*01*
*IGHV 3-30*02*	*IGHV3-21*	*IGHV4-31*
*IGHV 3-30*03*	*IGHV3-21*03*	*IGHV4-34*
*IGHV 3-33*01*	*IGHV3-23*	*IGHV4-34*01*
*IGHV 3-48*02*	*IGHV3-23*01*	*IGHV4-34*02*
IGHV 3-53	*IGHV3-30*	*IGHV4-34*03*
IGHV 3-64	*IGHV3-30*01*	*IGHV4-39*
*IGHV 3-7*01*	*IGHV3-30*02*	*IGHV4-39*01*
*IGHV 3-7*03*	*IGHV3-30*03*	*IGHV4-4*
IGHV 4-39	*IGHV3-30*07*	*IGHV4-4*02*
*IGHV 4-39*01*	*IGHV3-30-3*01*	*IGHV4-59*
*IGHV 5-51*01*	*IGHV3-33*	*IGHV4-59*01*
IGHV1-02	*IGHV3-33*01*	*IGHV4-59*02*
*IGHV1-02*04*	*IGHV3-43*	*IGHV4-59*03*
IGHV1-18	*IGHV3-48*	*IGHV4-59*08*
*IGHV1-18*01*	*IGHV3-48*02*	*IGHV4-61*
IGHV1-2	*IGHV3-48*03*	*IGHV4-61*01*
*IGHV1-2*02*	*IGHV3-49*	*IGHV4-61*02*
*IGHV1-2*04*	*IGHV3-49*04*	*IGHV5-51*
*IGHV1-68*01*	*IGHV3-53*	*IGHV5-51*01*
IGHV1-69	*IGHV3-53*01*	*IGHV5-51*03*
*IGHV1-69*01*	*IGHV3-64*	*IGHV6-1*
*IGHV1-69*06*	*IGHV3-7*	*IGHV6-1*01*
IGHV1-8	*IGHV3-7*01*	

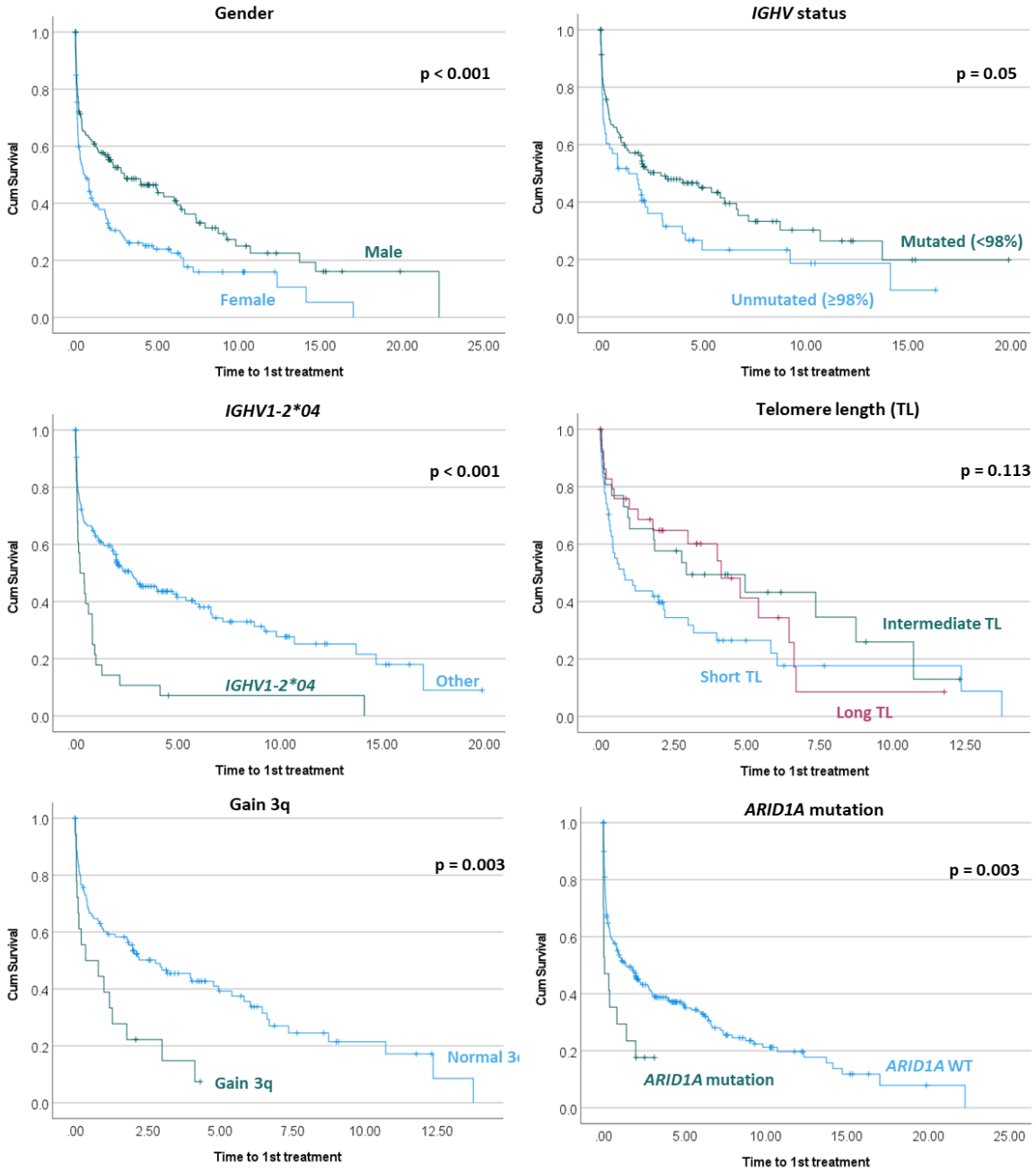
Supplementary figures



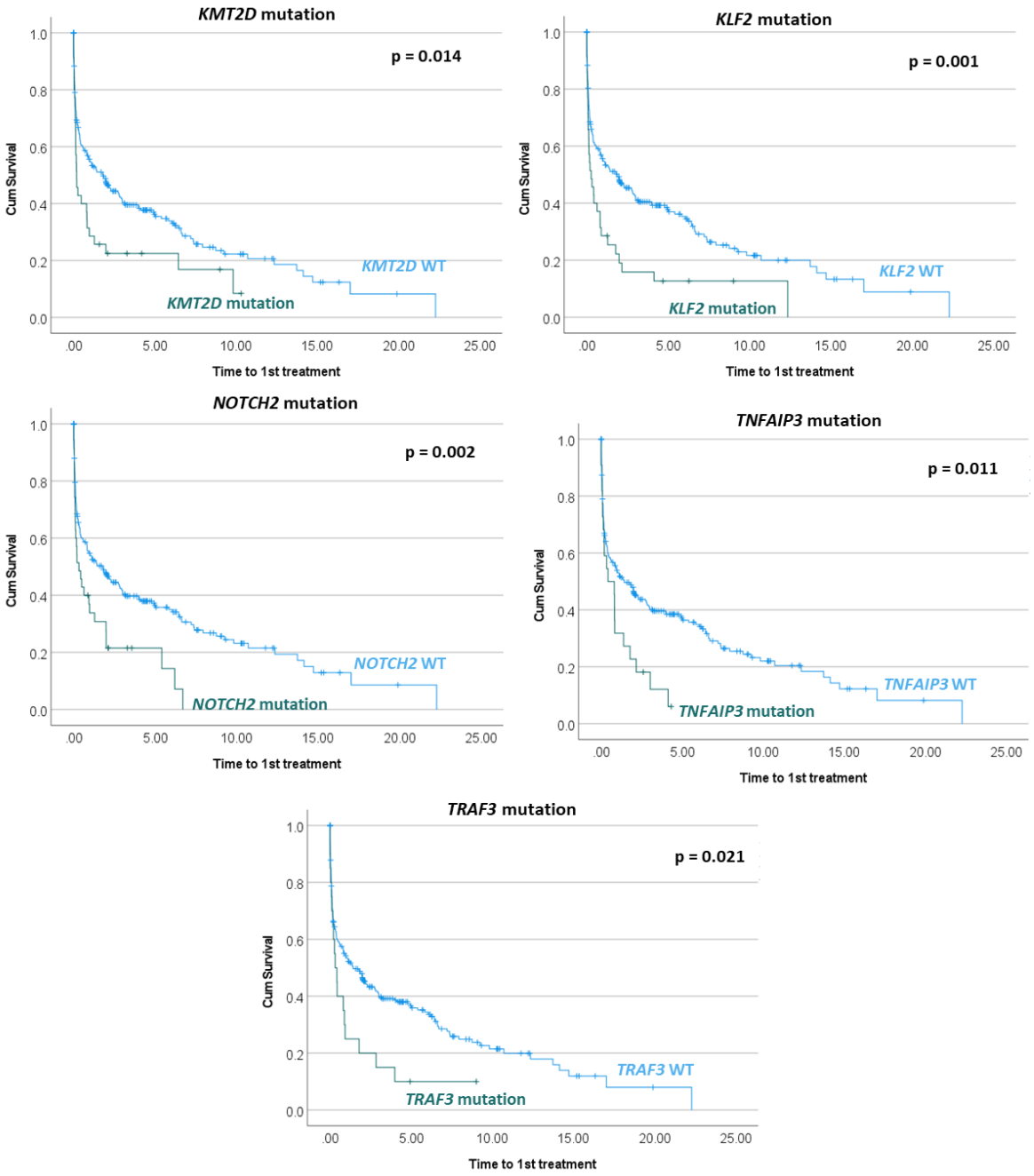
Supplementary Figure 1. Kaplan Meier curves for overall survival for *TP53* aberrations, age at diagnosis, genomic complexity, 7q deletion, 8q gain and 1q deletion.



Supplementary Figure 2. Kaplan Meier curves for overall survival for 6q deletion, 17p deletion, *MYD88* mutation, *TNFAIP3* mutation, and *TP53* mutation.



Supplementary Figure 3. Kaplan Meier curves for time to first treatment for gender, *IGHV* status, *IGHV1-2*04* status, telomere length, gain of 3q and *ARID1A* mutation.



Supplementary Figure 4. Kaplan Meier curves for time to first treatment *KMT2D* mutation, *KLF2* mutation, *NOTCH2* mutation, *TNFAIP3* mutation, and *TRAF3* mutation.

Bibliography

1. Ruddon, R. W. *Cancer Biology*. (Oxford University Press, 2007).

2. *Essential Cell Biology*. (Garland Science, Taylor & Francis Group, 2014).

3. Chapter 41 - G1 Phase and Regulation of Cell Proliferation. in *Cell Biology (Third Edition)* (eds. Pollard, T. D., Earnshaw, W. C., Lippincott-Schwartz, J. & Johnson, G. T.) 713–726 (Elsevier, 2017). doi:https://doi.org/10.1016/B978-0-323-34126-4.00041-4.

4. Chapter 43 - G2 Phase, Responses to DNA Damage, and Control of Entry Into Mitosis. in *Cell Biology (Third Edition)* (eds. Pollard, T. D., Earnshaw, W. C., Lippincott-Schwartz, J. & Johnson, G. T.) 743–754 (Elsevier, 2017). doi:https://doi.org/10.1016/B978-0-323-34126-4.00043-8.

5. Chapter 44 - Mitosis and Cytokinesis. in *Cell Biology (Third Edition)* (eds. Pollard, T. D., Earnshaw, W. C., Lippincott-Schwartz, J. & Johnson, G. T.) 755–778 (Elsevier, 2017). doi:https://doi.org/10.1016/B978-0-323-34126-4.00044-X.

6. Williams, A. B. & Schumacher, B. p53 in the DNA-Damage-Repair Process. *Cold Spring Harb. Perspect. Med.* **6**, a026070 (2016).

7. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

8. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

9. Näslund-Koch, C., Nordestgaard, B. G. & Bojesen, S. E. Increased Risk for Other Cancers in Addition to Breast Cancer for CHEK2 *1100delC Heterozygotes Estimated From the Copenhagen General Population Study. *J. Clin. Oncol.* **34**, 1208–1216 (2016).

10. Chatrath, A. *et al.* The pan-cancer landscape of prognostic germline variants in 10,582 patients. *Genome Med.* **12**, 15 (2020).

11. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

12. Tumor Suppressor Genes. in *Encyclopedia of Genetics* (eds. Brenner, S. & Miller, J. H.) 2081–2088 (Elsevier Science In, 2001).

13. Gou, L. Y., Niu, F. Y., Wu, Y. L. & Zhong, W. Z. Differences in driver genes between smoking-related and non-smoking-related lung cancer in the Chinese population. *Cancer* **121**, 3069–3079 (2015).

14. Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239-254.e6 (2018).

15. Chae, Y. K. *et al.* Genomic landscape of DNA repair genes in cancer. *Oncotarget* **7**, 23312–21 (2016).

16. Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Böckler, B. & Graham, T. A. The effects of mutational processes and selection on driver mutations across cancer types. *Nat. Commun.* **9**, 1857 (2018).

17. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science (80-.).* **339**, 1546–1558 (2013).

18. Olsen, M. M. & Zitella, L. J. *Hematologic Malignancies in Adults*. (Oncology Nursing Society, 2013).

19. Swerdlow, S. H. *et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. (International Agency for Research on Cancer, 2008).

20. Ninkovic, S. & Lambert, J. Non-Hodgkin lymphoma. *Medicine (Baltimore).* **45**, 297–304 (2017).

21. Doan, T., Melvold, R., Viselli, S. & Waltenbaugh, C. *Lippincott's Illustrated Reviews: Immunology*. (Lippincott Williams & Wilkins, 2013).

22. Lydyard, P., Whelan, A. & Fanger, M. *BIOS Instant Notes in Immunology*. (Taylor & Francis Group, 2011).

23. Murphy, K. M. *Janeway's Immunobiology*. (Taylor & Francis Group, 2011).

24. Kienzler, A.-K. & Eibel, H. Human B Cell Development and Tolerance. in *Encyclopedia of Immunobiology* vol. 1 105–121 (Elsevier, 2016).

25. Medina, K. L. *Overview of the immune system*. *Handbook of Clinical Neurology* vol. 133 (Elsevier B.V., 2016).

26. Cerutti, A., Cols, M. & Puga, I. Marginal zone B cells: Virtues of innate-like antibody-producing lymphocytes. *Nat. Rev. Immunol.* **13**, 118–132 (2013).

27. Du, M. Q. Pathogenesis of splenic marginal zone lymphoma. *Pathogenesis* **2**, 11–20 (2015).

28. Weill, J. C., Weller, S. & Reynaud, C. A. Human marginal zone B cells. *Annu. Rev. Immunol.* **27**, 267–285 (2009).

29. Descatoire, M. *et al.* Identification of a human splenic marginal zone B cell precursor with NOTCH2-dependent differentiation properties. *J. Exp. Med.* **211**, 987–1000 (2014).

30. Weller, S. *et al.* Somatic diversification in the absence of antigen-driven responses is the hallmark of the IgM+IgD+CD27+ B cell repertoire in infants. *J. Exp. Med.* **205**, 1331–1342 (2008).

31. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127**, 2375–2390 (2016).

32. Kanellis, G. *et al.* Splenic diffuse red pulp small B-cell lymphoma: revision of a series of cases reveals characteristic clinico-pathological features. *Haematologica* **95**, 1122–1129 (2010).

33. Traverse-Glehen, A. *et al.* Splenic diffuse red pulp small-B cell lymphoma: toward the emergence of a new lymphoma entity. *Discov. Med.* **13**, 253–265 (2012).

34. Xochelli, A. *et al.* Clonal B-cell lymphocytosis exhibiting immunophenotypic features consistent with a marginal-zone origin: is this a distinct entity? *Blood* **123**, 1199–1206 (2014).

35. Matutes, E. Diagnostic and therapeutic challenges in hairy cell leukemia-variant: where are we in 2021? *Expert Rev. Hematol.* **14**, 355–363 (2021).

36. Chacón, J. I. *et al.* Splenic marginal zone lymphoma: clinical characteristics and prognostic factors in a series of 60 patients. *Blood* **100**, 1648–1654 (2002).

37. Matutes, E. *et al.* Splenic marginal zone lymphoma proposals for a revision of diagnostic, staging and therapeutic criteria. *Leukemia* **22**, 487–495 (2008).

38. Arcaini, L., Rossi, D. & Paulli, M. Splenic marginal zone lymphoma: from genetics to management. *Blood* **127**, 2072–81 (2016).

39. Lumish, M. *et al.* How we treat mature B-cell neoplasms (indolent B-cell lymphomas). *J. Hematol. Oncol.* **14**, 5 (2021).

40. Zucca, E. *et al.* Marginal zone lymphomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **31**, 17–29 (2020).

41. Robinson, J. E. & Cutucache, C. E. Deciphering splenic marginal zone lymphoma pathogenesis: The proposed role of microRNA. *Oncotarget* vol. 9 30005–30022 (2018).

42. Algara, P. *et al.* Analysis of the IgVH somatic mutations in splenic marginal zone lymphoma defines a group of unmutated cases with frequent 7q deletion and adverse clinical course. *Blood* **99**, 1299–1304 (2002).

43. Bahler, D. W., Pindzola, J. A. & Swerdlow, S. H. Splenic Marginal Zone Lymphomas Appear to Originate from Different B Cell Types. *Am. J. Pathol.* **161**, 81–88 (2002).

44. Seifert, M., Scholtysik, R. & Küppers, R. Origin and Pathogenesis of B Cell Lymphomas. in (ed. Küppers, R.) 1–25 (Humana Press, 2013). doi:10.1007/978-1-62703-269-8_1.

45. Watkins, J. *et al.* Splenic marginal zone lymphoma: characterization of 7q deletion and its value in diagnosis. *J. Pathol.* **220**, 114–125 (2010).

46. Solé, F. *et al.* Splenic marginal zone B-cell lymphomas: two cytogenetic subtypes, one with gain of 3q and the other with loss of 7q. *Haematologica* **86**, 71–77 (2001).

47. Salido, M. *et al.* Cytogenetic aberrations and their prognostic value in a series of 330 splenic marginal zone B-cell lymphomas: a multicenter study of the Splenic B-Cell Lymphoma Group. *Blood* **116**, 1479–1488 (2010).

48. Oscier, D. G. *et al.* Cytogenetic studies in splenic lymphoma with villous lymphocytes. *Br. J. Haematol.* **85**, 487–491 (1993).

49. Kiel, M. J. *et al.* Whole-genome sequencing identifies recurrent somatic *NOTCH2* mutations in splenic marginal zone lymphoma. *J. Exp. Med.* **209**, 1553–1565 (2012).

50. Rossi, D. *et al.* The coding genome of splenic marginal zone lymphoma: activation of *NOTCH2* and other pathways regulating marginal zone development. *J. Exp. Med.* **209**, 1537–1551 (2012).

51. Parry, M. *et al.* Whole Exome Sequencing Identifies Novel Recurrently Mutated Genes in Patients with Splenic Marginal Zone Lymphoma. *PLoS One* **8**, e83244 (2013).

52. Martínez, N. *et al.* Whole-exome sequencing in splenic marginal zone lymphoma reveals mutations in genes involved in marginal zone differentiation. *Leukemia* **28**, 1334–1340 (2014).

53. Peveling-Oberhag, J. *et al.* Whole exome sequencing of microdissected splenic marginal zone lymphoma: a study to discover novel tumor-specific mutations. *BMC Cancer* **15**, 773 (2015).

54. Clipson, A. *et al.* KLF2 mutation is the most frequent somatic change in splenic marginal zone lymphoma and identifies a subset with distinct genotype. *Leukemia* **29**, 1177–1185 (2015).

55. Oquendo, C. J. *et al.* The (epi)genomic landscape of splenic marginal zone lymphoma, biological implications, clinical utility, and future questions. *J. Transl. Genet. Genomics* **5**, 89–111 (2021).

56. Vogel, F. & Motulsky, A. . *Human Genetic: Problems and Approaches*. (Springer).

57. Trent, R. . *Molecular medicine*. (1993).

58. Pinkel, D. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 9138–9142 (1988).

59. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (80-.).* **258**, 818–821 (1992).

60. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**, 5463–5467 (1977).

61. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

62. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).

63. Illumina Inc. *An introduction to Next-Generation Sequencing Technology*. (2017).

64. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–26 (2014).

65. Lee, H. *et al.* Third-generation sequencing and the future of genomics. (2016) doi:https://doi.org/10.1101/048603.

66. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). https://www.genome.gov/sequencingcostsdata/ (2018).

67. Genomics England. The 100,000 Genomes Project. https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/.

68. Illumina Inc. Targeted gene sequencing. https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/targeted-panels.html (2017).

69. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci.* **106**, 19096–19101 (2009).

70. Jaramillo Oquendo, C. *et al.* Systematic Review of Somatic Mutations in Splenic Marginal Zone Lymphoma. *Sci. Rep.* **9**, 1–9 (2019).

71. Parry, M. *et al.* Genetics and prognostication in splenic marginal zone lymphoma: Revelations from deep sequencing. *Clin. Cancer Res.* **21**, 4174–4183 (2015).

72. Piva, R. *et al.* The Krüppel-like factor 2 transcription factor gene is recurrently mutated in splenic marginal zone lymphoma. *Leukemia* **29**, 503–507 (2015).

73. Shamseer, L. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* **349**, g7647–g7647 (2015).

74. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

75. Spina, V. *et al.* The Genetics of Nodal Marginal Zone Lymphoma. *Blood* **128**, 1362–1374 (2016).

76. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).

77. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

78. Project, G. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

79. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). http://evs.gs.washington.edu/EVS/.

80. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).

81. Skidmore, Z. L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012–3014 (2016).

82. Canisius, S., Martens, J. W. M. & Wessels, L. F. A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.* **17**, 261 (2016).

83. Campos-Martín, Y. *et al.* Clinical and diagnostic relevance of NOTCH2 and KLF2 mutations in splenic marginal zone lymphoma. *Haematologica* **102**, e310–e312 (2017).

84. Rossi, D. *et al.* Alteration of BIRC3 and multiple other NF-κB pathway genes in splenic marginal zone lymphoma. *Blood* **118**, 4930–4934 (2011).

85. Yan, Q. *et al.* BCR and TLR signaling pathways are recurrently targeted by genetic changes in splenic marginal zone lymphomas. *Haematologica* **97**, 595–598 (2012).

86. Spina, V. & Rossi, D. Molecular pathogenesis of splenic and nodal marginal zone lymphoma. *Best Pract. Res. Clin. Haematol.* **30**, 5–12 (2017).

87. Jallades, L. *et al.* Exome sequencing identifies recurrent BCOR alterations and the absence of KLF2, TNFAIP3 and MYD88 mutations in splenic diffuse red pulp small B-cell lymphoma. *Haematologica* **102**, 1758–1766 (2017).

88. Pillonel, V. *et al.* High-throughput sequencing of nodal marginal zone lymphomas identifies recurrent BRAF mutations. *Leukemia* **32**, 2412–2426 (2018).

89. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

90. Bouaoun, L. *et al.* TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Hum. Mutat.* **37**, 865–876 (2016).

91. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2017).

92. Hunter, Z. R. *et al.* The genomic landscape of Waldenström macroglobulinemia is characterized by highly recurring MYD88 and WHIM-like CXCR4 mutations, and small somatic deletions associated with B-cell lymphomagenesis. *Blood* **123**, 1637–1646 (2014).

93. Xochelli, A., Oscier, D. & Stamatopoulos, K. Clonal B-cell lymphocytosis of marginal zone origin. *Best Pract. Res. Clin. Haematol.* **30**, 77–83 (2017).

94. Parker, H. *et al.* CBL-MZ is not a single biological entity: evidence from genomic analysis and prolonged clinical follow-up. *Blood Adv.* **2**, 1116–1119 (2018).

95. Curiel-Olmo, S. *et al.* Splenic diffuse red pulp small B-cell lymphoma displays increased expression of cyclin D3 and recurrent CCND3 mutations. *Blood* **129**, 1042–1045 (2017).

96. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

97. Wang, K. *et al.* PEST domain mutations in Notch receptors comprise an oncogenic driver segment in triple-negative breast cancer sensitive to a γ-secretase inhibitor. *Clin. Cancer Res.* **21**, 1487–1496 (2015).

98. Ortega-Molina, A. *et al.* The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nat. Med.* **21**, 1199–1208 (2015).

99. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 1–12 (2015).

100. Robbiani, D. F. *et al.* AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. *Cell* **135**, 1028–38 (2008).

101. Blakemore, S. J. *et al.* Clinical significance of TP53, BIRC3, ATM and MAPK-ERK genes in chronic lymphocytic leukaemia: data from the randomised UK LRF CLL4 trial. *Leukemia* **34**, 1760–1774 (2020).

102. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

103. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–6 (2011).

104. Barnell, E. K. *et al.* Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genet. Med.* **21**, 972–981 (2019).

105. Ewing, B., Hillier, L. D. & Wendl, M. C. Base-Calling of Automated Sequencer Traces Using Phred. *Genome Res.* **8**, 186–194 (1998).

106. Lander, S. *et al.* Correction: Initial sequencing and analysis of the human genome. *Nature* **412**, 565–566 (2001).

107. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

108. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).

109. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

110. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

111. Rumble, S. M. *et al.* SHRiMP: Accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**, 1–11 (2009).

112. Canzar, S. & Salzberg, S. L. Short Read Mapping: An Algorithmic Tour. *Proc. IEEE* **105**, 436–458 (2017).

113. Samtools. The Variant Call Format (VCF) Version 4.2 Specification. *Online Resour.* 1–28 (2015) doi:10.1016/j.ymeth.2012.07.021.

114. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

115. Lunter, G. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

116. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

117. Shen, D. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

118. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

119. Lai, Z. *et al.* VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, 1–11 (2016).

120. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

121. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

122. A. Van der Auwera, G. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11101–111033 (2014).

123. Lei, S. *et al.* Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data. *Bioinformatics* 1–3 (2018) doi:10.1093/bioinformatics/bty849.

124. Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 1–10 (2014).

125. Roberts, N. D. *et al.* A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* **29**, 2223–2230 (2013).

126. Hsu, Y.-C., Hsiao, Y.-T., Kao, T.-Y., Chang, J.-G. & Shieh, G. S. Detection of Somatic Mutations in Exome Sequencing of Tumor-only Samples. *Sci. Rep.* **7**, 15959 (2017).

127. He, L., Chou, K.-C., Cai, L., Zhang, Z. & Yuan, W. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci. Rep.* **6**, 1–9 (2016).

128. Krøigård, A. B., Thomassen, M., Lænkholm, A. V., Kruse, T. A. & Larsen, M. J. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One* **11**, 1–15 (2016).

129. Spencer, D. H. *et al.* Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data. *J. Mol. Diagnostics* **16**, 75–88 (2014).

130. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* **7**, 43169 (2017).

131. Wang, Q. *et al.* Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.* **5**, 91 (2013).

132. Geraldine A. Van der Auwera, B. D. O. *Genomics in the Cloud*. (O'Reilly Media, Inc., 2020).

133. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

134. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

135. Lantz, B. *Machine learning with R, 2nd edition*. *Packt Publishing Ltd* (2015).

136. Parang, K., Wiebe, L. I. & Knaus, E. E. Novel approaches for designing 5'-O-ester prodrugs of 3'-azido-2', 3'-dideoxythymidine (AZT). *Curr. Med. Chem.* **7**, 995–1039 (2000).

137. Wu, C. *et al.* Using Machine Learning to Identify True Somatic Variants from Next-Generation Sequencing. *Clin. Chem.* **66**, 239–246 (2020).

138. Transl, S. *et al. HHS Public Access A machine learning approach for somatic mutation discovery*. vol. 10 (2019).

139. Agajanian, S., Oluyemi, O. & Verkhivker, G. M. Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations. *Front. Mol. Biosci.* **6**, (2019).

140. Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).

141. Sun, J. X. *et al.* A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLOS Comput. Biol.* **14**, e1005965 (2018).

142. Ainscough, B. J. *et al.* A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.* **50**, 1735–1743 (2018).

143. Sukhai, M. A. *et al.* Somatic Tumor Variant Filtration Strategies to Optimize Tumor-Only Molecular Profiling Using Targeted Next-Generation Sequencing Panels. *J. Mol. Diagnostics* **21**, 261–273 (2019).

144. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

145. Coenen, A. & Pearce, A. Understanding UMAP. https://pair-code.github.io/understanding-umap/.

146. Mehta, P. *et al.* A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).

147. Merli, M. *et al.* Clonal B-Cell Lymphocytosis with Mariginal-Zone Features: Comparison with Overt Splenic Marginal-Zone Lymphomas in 77 Patients from a Monocentric Series. *Blood* **134**, 4017–4017 (2019).

148. Bertoni, F., Rossi, D., Raderer, M. & Zucca, E. Marginal Zone Lymphomas. *Cancer J.* **26**, 336–347 (2020).

149. Luijten, M. N. H., Lee, J. X. T. & Crasta, K. C. Mutational game changer: Chromothripsis and its emerging relevance to cancer. *Mutation Research - Reviews in Mutation Research* vol. 777 29–51 (2018).

150. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

151. Itan, Y. *et al.* The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat. Methods* **13**, 109–110 (2016).

152. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD ®): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).

153. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).

154. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

155. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–4 (2012).

156. van der Velde, K. J. *et al.* GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biol.* **18**, 1–10 (2017).

157. Feinberg, M. W., Lin, Z., Fisch, S. & Jain, M. K. An emerging role for Krüppel-like factors in vascular biology. *Trends Cardiovasc. Med.* **14**, 241–6 (2004).

158. Dang, D. T., Pevsner, J. & Yang, V. W. The biology of the mammalian Krüppel-like family of transcription factors. *Int. J. Biochem. Cell Biol.* **32**, 1103–1121 (2000).

159. Hart, G. T., Wang, X., Hogquist, K. A. & Jameson, S. C. Krüppel-like factor 2 (KLF2) regulates B-cell reactivity, subset differentiation, and trafficking molecule expression. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 716–21 (2011).

160. Hoek, K. L. *et al.* Follicular B Cell Trafficking within the Spleen Actively Restricts Humoral Immune Responses. *Immunity* **33**, 254–265 (2010).

161. Winkelmann, R. *et al.* B cell homeostasis and plasma cell homing controlled by Krüppel-like factor 2. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 710–5 (2011).

162. Arruga, F., Vaisitti, T. & Deaglio, S. The NOTCH Pathway and Its Mutations in Mature B Cell Malignancies. *Front. Oncol.* **8**, 1–18 (2018).

163. Moran, S. T. *et al.* Synergism between NF-κB1/p50 and Notch2 during the Development of Marginal Zone B Lymphocytes. *J. Immunol.* **179**, 195–200 (2007).

164. Saito, T. *et al.* Notch2 is preferentially expressed in mature B cells and indispensable for marginal zone B lineage development. *Immunity* **18**, 675–85 (2003).

165. Tanigaki, K. *et al.* Notch-RBP-J signaling is involved in cell fate determination of marginal zone B cells. *Nat. Immunol.* **3**, 443–50 (2002).

166. Kopan, R. Notch Signaling. *Cold Spring Harb. Perspect. Biol.* **4**, a011213–a011213 (2012).

167. Radtke, F., Wilson, A. & MacDonald, H. R. Notch signaling in T- and B-cell development. *Curr. Opin. Immunol.* **16**, 174–179 (2004).

168. Lee, S. *et al.* Gain-of-function mutations and copy number increases of Notch2 in diffuse large B-cell lymphoma. *Cancer Sci.* **100**, 920–926 (2009).

169. Shanmugam, V. *et al.* Notch activation is pervasive in SMZL and uncommon in DLBCL: implications for Notch signaling in B-cell tumors. *Blood Adv.* **5**, 71–83 (2021).

170. Hampel, F. *et al.* CD19-independent instruction of murine marginal zone B-cell development by constitutive Notch2 signaling. *Blood* **118**, 6321–31 (2011).

171. Fabbri, G. *et al.* Common nonmutational NOTCH1 activation in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E2911–E2919 (2017).

172. Pillai, S. & Cariappa, A. The follicular versus marginal zone B lymphocyte cell fate decision. *Nat. Rev. Immunol.* **9**, 767–77 (2009).

173. Spina, V. & Rossi, D. NF-κB deregulation in splenic marginal zone lymphoma. *Semin. Cancer Biol.* **39**, 61–67 (2016).

174. Lim, K.-H., Yang, Y. & Staudt, L. M. Pathogenetic importance and therapeutic implications of NF-κB in lymphoid malignancies. *Immunol. Rev.* **246**, 359–78 (2012).

175. Kato, M. *et al.* Frequent inactivation of A20 in B-cell lymphomas. *Nature* **459**, 712–716 (2009).

176. Zhang, F., Yang, L. & Li, Y. The role of A20 in the pathogenesis of lymphocytic malignancy. *Cancer Cell International* vol. 12 1–7 (2012).

177. Honma, K. *et al.* TNFAIP3/A20 functions as a novel tumor suppressor gene in several subtypes of non-Hodgkin lymphomas. *Blood* **114**, 2467–2475 (2009).

178. Escudero-Ibarz, L., Wang, M. & Du, M. Q. Significant functional difference between TNFAIP3 truncation and missense mutants. *Haematologica* vol. 101 e382–e384 (2016).

179. Honma, K. *et al.* TNFAIP3/A20 functions as a novel tumor suppressor gene in several subtypes of non-Hodgkin lymphomas. *Blood* **114**, 2467–2475 (2009).

180. Frazzi, R. BIRC3 and BIRC5: multi-faceted inhibitors in cancer. *Cell and Bioscience* vol. 11 8 (2021).

181. Pflug, K. M. & Sitcheran, R. Targeting NF-κB-inducing kinase (NIK) in immunity, inflammation, and cancer. *Int. J. Mol. Sci.* **21**, 1–19 (2020).

182. Lamason, R. L., McCully, R. R., Lew, S. M. & Pomerantz, J. L. Oncogenic CARD11 mutations induce hyperactive signaling by disrupting autoinhibition by the PKC-responsive inhibitory domain. *Biochemistry* **49**, 8240–8250 (2010).

183. Perham, R. N., Packman, L. C. & Radford, S. E. Supporting Online Material Oncogenic CARD11 Mutations in Human Diffuse Large B Cell Lymphoma. *Biochem. Biophys. Res. Commun* **54**, 451 (1987).

184. Weber, A. N. R. *et al.* Oncogenic MYD88 mutations in lymphoma: novel insights and therapeutic possibilities. *Cancer Immunology, Immunotherapy* vol. 67 1797–1807 (2018).

185. Lunning, M. A. & Green, M. R. Mutation of chromatin modifiers; an emerging hallmark of germinal center B-cell lymphomas. *Blood Cancer J.* **5**, e361 (2015).

186. Zhang, J. *et al.* Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nat. Med. 2015 2110* **21**, 1190–1198 (2015).

187. Green, M. R. *et al.* Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1116–E1125 (2015).

188. Okosun, J. *et al.* Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat. Genet.* **46**, 176–181 (2014).

189. Veneti, Z., Gkouskou, K. K. & Eliopoulos, A. G. Polycomb repressor complex 2 in genomic instability and cancer. *International Journal of Molecular Sciences* vol. 18 (2017).

190. Green, M. R. Chromatin modifying gene mutations in follicular lymphoma. *Blood* vol. 131 595–604 (2018).

191. Phan, R. T. & Dalla-Favera, R. The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. *Nature* **432**, 635–639 (2004).

192. Han, L. *et al.* Chromatin remodeling mediated by ARID1A is indispensable for normal hematopoiesis in mice. *Leukemia* **33**, 2291–2305 (2019).

193. Wu, R. C., Wang, T. L. & Shih, I. M. The emerging roles of ARID1A in tumor suppression. *Cancer Biology and Therapy* vol. 15 655–664 (2014).

194. Kim, M. P. & Lozano, G. Mutant p53 partners in crime. *Cell Death Differ.* **25**, 161–168 (2018).

195. Sherr, C. J. D-type cyclins. *Trends in Biochemical Sciences* vol. 20 187–190 (1995).

196. Ramezani-Rad, P., Chen, C., Zhu, Z. & Rickert, R. C. Cyclin D3 Governs Clonal Expansion of Dark Zone Germinal Center B Cells. *Cell Rep.* **33**, (2020).

197. Cato, M. H., Chintalapati, S. K., Yau, I. W., Omori, S. A. & Rickert, R. C. Cyclin D3 Is Selectively Required for Proliferative Expansion of Germinal Center B Cells. *Mol. Cell. Biol.* **31**, 127–137 (2011).

198. Pae, J. *et al.* Cyclin D3 drives inertial cell cycling in dark zone germinal center B cells. *J. Exp. Med.* **218**, (2021).

199. Caeser, R. *et al.* Genetic modification of primary human B cells to model high-grade lymphoma. *Nat. Commun.* **10**, 1–16 (2019).

200. Arita, K., Tsuzuki, S., Ohshima, K., Sugiyama, T. & Seto, M. Synergy of Myc, cell cycle regulators and the Akt pathway in the development of aggressive B-cell lymphoma in a mouse model. *Leukemia* vol. 28 2270–2272 (2014).

201. Schmitz, R. *et al.* Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490**, 116–120 (2012).

202. Rinaldi, A. *et al.* Genome-wide DNA profiling of marginal zone lymphomas identifies subtype-specific lesions with an impact on the clinical outcome. *Blood* **117**, 1595–1604 (2011).

203. Dal Bo, M. *et al.* 13q14 Deletion size and number of deleted cells both influence prognosis in chronic lymphocytic leukemia. *Genes, Chromosom. Cancer* **50**, 633–643 (2011).

204. Fresquet, V. *et al.* High-throughput sequencing analysis of the chromosome 7q32 deletion reveals IRF5 as a potential tumour suppressor in splenic marginal-zone lymphoma. *Br. J. Haematol.* **158**, 712–726 (2012).

205. Bikos, V. *et al.* Over 30% of patients with splenic marginal zone lymphoma express the same immunoglobulin heavy variable gene: ontogenetic implications. *Leukemia* **26**, 1638–46 (2012).

206. Xochelli, A. *et al.* Disease-biased and shared characteristics of the immunoglobulin gene repertoires in marginal zone B cell lymphoproliferations. *J. Pathol.* **247**, 416–421 (2019).

207. Bikos, V. *et al.* An Immunogenetic Signature of Ongoing Antigen Interactions in Splenic Marginal Zone Lymphoma Expressing IGHV1-2*04 Receptors. *Clin. Cancer Res.* **22**, 2032–2040 (2016).

208. Kalpadakis, C., Pangalis, G. A., Angelopoulou, M. K. & Vassilakopoulos, T. P. Treatment of splenic marginal zone lymphoma. *Best Pract. Res. Clin. Haematol.* **30**, 139–148 (2017).

209. Sima, A. *et al.* Superior outcome for splenectomised patients in a population-based study of splenic marginal zone lymphoma in Sweden. *Br. J. Haematol.* **194**, 568–579 (2021).

210. Hovestadt, V. & Zapatka, M. conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays.

211. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

212. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).

213. Lai, T. P. *et al.* A method for measuring the distribution of the shortest telomeres in cells and tissues. *Nat. Commun.* **8**, (2017).

214. Baird, D. M., Rowson, J., Wynford-Thomas, D. & Kipling, D. Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nat. Genet.* **33**, 203–207 (2003).

215. Strefford, J. C. *et al.* Telomere length predicts progression and overall survival in chronic lymphocytic leukemia: Data from the UK LRF CLL4 trial. *Leukemia* vol. 29 2411–2414 (2015).

216. Hernández, J. M. *et al.* Novel Genomic Imbalances in B-Cell Splenic Marginal Zone Lymphomas Revealed by Comparative Genomic Hybridization and Cytogenetics. *Am. J. Pathol.* **158**, 1843–1850 (2001).

217. Andersen, C. L. *et al.* A narrow deletion of 7q is common to HCL, and SMZL, but not CLL. *Eur. J. Haematol.* **72**, 390–402 (2004).

218. Watkins, A. J. *et al.* An integrated genomic and expression analysis of 7q deletion in splenic marginal zone lymphoma. *PLoS One* **7**, e44997 (2012).

219. Cho, S., Kim, H.-S., Zeiger, M. A., Umbricht, C. B. & Cope, L. M. Measuring DNA Copy Number Variation Using High-Density Methylation Microarrays. *J. Comput. Biol.* **26**, 295–304 (2019).

220. Ruiz-Ballesteros, E. *et al.* MicroRNA losses in the frequently deleted region of 7q in SMZL. *Leukemia* **21**, 2547–9 (2007).

221. Ruiz-Ballesteros, E. *et al.* Splenic marginal zone lymphoma: proposal of new diagnostic and prognostic markers identified after tissue and cDNA microarray analysis. *Blood* **106**, 1831–1838 (2005).

222. Basso, K. *et al.* Identification of the Human Mature B Cell miRNome. *Immunity* **30**, 744–752 (2009).

223. Gruszka-Westwood, A. M., Hamoudi, R., Osborne, L., Matutes, E. & Catovsky, D. Deletion mapping on the long arm of chromosome 7 in splenic lymphoma with villous lymphocytes. *Genes, Chromosom. Cancer* **36**, 57–69 (2003).

224. Vega, F. *et al.* Splenic marginal zone lymphomas are characterized by loss of interstitial regions of chromosome 7q, 7q31.32 and 7q36.2 that include the protection of telomere 1 (POT1) and sonic hedgehog (SHH) genes. *Br. J. Haematol.* **142**, 216–226 (2008).

225. Pinzaru, A. M. *et al.* Telomere Replication Stress Induced by POT1 Inactivation Accelerates Tumorigenesis. *Cell Rep.* **15**, 2170–2184 (2016).

226. Palm, W. & de Lange, T. How Shelterin Protects Mammalian Telomeres. *Annu. Rev. Genet.* **42**, 301–334 (2008).

227. Dierlamm, J. *et al.* Characteristic pattern of chromosomal gains and losses in marginal zone B cell lymphoma detected by comparative genomic hybridization. *Leukemia* **11**, 747–758 (1997).

228. Dierlamm, J. *et al.* Trisomy 3 in marginal zone B-cell lymphoma: a study based on cytogenetic analysis and fluorescence in situ hybridization. *Br. J. Haematol.* **93**, 242–249 (1996).

229. Solé, F. *et al.* Frequent involvement of chromosomes 1, 3, 7 and 8 in splenic marginal zone B-cell lymphoma. *Br. J. Haematol.* **98**, 446–449 (1997).

230. Troussard, X. *et al.* Genetic analysis of splenic lymphoma with villous lymphocytes: A Groupe Francais d'Hematologie Cellulaire (GFHC) study. *Br. J. Haematol.* **101**, 712–721 (1998).

231. Robledo, C. *et al.* Molecular Characterization of the Region 7q22.1 in Splenic Marginal Zone Lymphomas. *PLoS One* **6**, e24939 (2011).

232. Höllein, A. *et al.* Deletion 7q Is Associated with KLF2 and NOTCH2 Mutations and Is Strongly Correlated with Splenic Marginal Zone Lymphoma but Also Found in Lymphoplasmacytic Lymphoma and Hairy Cell Leukemia Variant. *Blood* **130**, 1465 (2017).

233. Yi, S. *et al. Del17p does not always significantly influence the survival of B-cell chronic lymphoproliferative disorders*. *Oncotarget* vol. 9 www.impactjournals.com/oncotarget (2018).

234. Gazzo, S. *et al.* Cytogenetic and molecular delineation of a region of chromosome 3q commonly gained in marginal zone B-cell lymphoma. *Haematologica* **88**, 31–8 (2003).

235. Boonstra, R. *et al.* Splenic marginal zone lymphomas presenting with splenomegaly and typical immunophenotype are characterized by allelic loss in 7q31-32. *Mod. Pathol. an Off. J. United States Can. Acad. Pathol. Inc* **16**, 1210–1217 (2003).

236. Taborelli, M. *et al.* Chromosome band 6q deletion pattern in malignant lymphomas. *Cancer Genet. Cytogenet.* **165**, 106–113 (2006).

237. Ocio, E. M. *et al.* 6q deletion in Waldenström macroglobulinemia is associated with features of adverse prognosis. *Br. J. Haematol.* **136**, 80–86 (2007).

238. Leveille, E. & Johnson, N. A. Genetic Events Inhibiting Apoptosis in Diffuse Large B Cell Lymphoma. *Cancers (Basel).* **13**, 2167 (2021).

239. Staniek, J. *et al.* TRAIL-R1 and TRAIL-R2 Mediate TRAIL-Dependent Apoptosis in Activated Primary Human B Lymphocytes. *Front. Immunol.* **0**, 951 (2019).

240. Jiménez, C. *et al.* Genomic evolution of ibrutinib-resistant clones in Waldenström macroglobulinaemia. *Br. J. Haematol.* **189**, 1165–1170 (2020).

241. Nguyen, L., Papenhausen, P. & Shao, H. The Role of c-MYC in B-Cell Lymphomas: Diagnostic and Molecular Aspects. *Genes (Basel).* **8**, 116 (2017).

242. Schaub, F. X. *et al.* Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas. *Cell Syst.* **6**, 282-300.e2 (2018).

243. Consortium, T. U. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

244. Casey, S. C., Baylot, V. & Felsher, D. W. The MYC oncogene is a global regulator of the immune response. *Blood* **131**, 2007–2015 (2018).

245. de Barrios, O., Meler, A. & Parra, M. MYC's Fine Line Between B Cell Development and Malignancy. *Cells* **9**, 523 (2020).

246. Klein, U. *et al.* The DLEU2/miR-15a/16-1 Cluster Controls B Cell Proliferation and Its Deletion Leads to Chronic Lymphocytic Leukemia. *Cancer Cell* **17**, 28–40 (2010).

247. Arribas, A. J. *et al.* DNA methylation profiling identifies two splenic marginal zone lymphoma subgroups with different clinical and genetic features. *Blood* **125**, 1922–1931 (2015).

248. Arribas, A. J. & Bertoni, F. Methylation patterns in marginal zone lymphoma. *Best Pract. Res. Clin. Haematol.* **30**, 24–31 (2017).

249. Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848–1854 (1999).

250. Damle, R. N. *et al.* Ig V Gene Mutation Status and CD38 Expression As Novel Prognostic Indicators in Chronic Lymphocytic Leukemia. *Blood* **94**, 1840–1847 (1999).

251. Arcaini, L. *et al.* Splenic marginal zone lymphoma: Clinical clustering of immunoglobulin heavy chain repertoires. *Blood Cells, Mol. Dis.* **42**, 286–291 (2009).

252. Desmaze, C., Soria, J.-C., Freulet-Marrière, M.-A., Mathieu, N. & Sabatier, L. Telomere-driven genomic instability in cancer cells. *Cancer Lett.* **194**, 173–182 (2003).

253. Gisselsson, D. *et al.* Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5357–5362 (2000).

254. Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).

255. Parker, A. E., Snell, K., Davis, Z. & Oscier, D. G. Telomere Length by Tel-PCR in B Cell Malignancies. *Blood* **106**, 2955–2955 (2005).

256. Strefford, J. C. The genomic landscape of chronic lymphocytic leukaemia: biological and clinical implications. *Br. J. Haematol.* **169**, 14–31 (2015).

257. Makewita, L. E. & Strefford, J. C. Molecular Genetics of Chronic Lymphocytic Leukaemia. in *eLS* 1–11 (Wiley, 2019). doi:10.1002/9780470015902.a0028437.

258. Fabbri, G. *et al.* Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation. *J. Exp. Med.* **208**, 1389–1401 (2011).

259. Rose-Zerilli, M. J. J. *et al.* Longitudinal copy number, whole exome and targeted deep sequencing of 'good risk' IGHV-mutated CLL patients with progressive disease. *Leukemia* **30**, 1301–1310 (2016).

260. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–30 (2015).

261. V, L. *et al.* Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. *Blood* **127**, 1007–1016 (2016).

262. AC, Q. *et al.* A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598–605 (2015).

263. Döhner, H. *et al.* Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **343**, 1910–1916 (2000).

264. Rossi, D. *et al.* Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. *Blood* **121**, 1403–1412 (2013).

265. Arcaini, L. *et al.* Splenic marginal zone lymphoma: A prognostic model for clinical use. *Blood* **107**, 4643–4649 (2006).

266. Sujobert, P. *et al.* The Need for a Consensus Next-generation Sequencing Panel for Mature Lymphoid Malignancies. *HemaSphere* **3**, e169 (2019).

267. Falini, B., Martelli, M. P. & Tiacci, E. BRAF V600E mutation in hairy cell leukemia: from bench to bedside. *Blood* **128**, 1918–1927 (2016).

268. Yu, X. *et al.* MYD88 L265P Mutation in Lymphoid Malignancies. *Cancer Res.* **78**, 2457–2462 (2018).

269. Hallek, M. *et al.* iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. *Blood* **131**, 2745–2760 (2018).

270. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494.e15 (2017).

271. Mollejo, M. & Piris, M. A. The complex pathology and differential diagnosis of splenic and nodal marginal zone lymphoma. *Ann. Lymphoma* **4**, 18–18 (2020).

272. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nat. Rev. Immunol.* **15**, 172–184 (2015).

273. Guidetti, F. *et al.* Molecular Subtypes of Splenic Marginal Zone Lymphoma (SMZL) Are Associated with Distinct Pathogenic Mechanisms and Outcomes - Interim Analysis of the IELSG46 Study. *Blood* **132**, 922–922 (2018).

274. Phillips, T. J. *et al.* Phase 2 Study Evaluating the Efficacy and Safety of Parsaclisib in Patients with Relapsed or Refractory Marginal Zone Lymphoma (CITADEL-204). *Blood* **136**, 27–28 (2020).

275. Opat, S. *et al.* Efficacy and Safety of Zanubrutinib in Patients with Relapsed/Refractory Marginal Zone Lymphoma: Initial Results of the MAGNOLIA (BGB-3111-214) Trial. *Blood* **136**, 28–30 (2020).

276. Zinzani, P. *et al.* UMBRALISIB MONOTHERAPY DEMONSTRATES EFFICACY AND SAFETY IN PATIENTS WITH RELAPSED/REFRACTORY MARGINAL ZONE LYMPHOMA: A MULTICENTER, OPEN-LABEL, REGISTRATION DIRECTED PHASE 2 STUDY. *Hematol. Oncol.* **37**, 182–183 (2019).

277. Noy, A. *et al.* Durable ibrutinib responses in relapsed/refractory marginal zone lymphoma: long-term follow-up and biomarker analysis. *Blood Adv.* **4**, 5773–5784 (2020).

278. Panayiotidis, P. *et al.* Efficacy and safety of copanlisib in patients with relapsed or refractory marginal zone lymphoma. *Blood Adv.* **5**, 823–828 (2021).

279. Phillips, T. J., Michot, J.-M. & Ribrag, V. Can Next-Generation PI3K Inhibitors Unlock the Full Potential of the Class in Patients With B-Cell Lymphoma? *Clin. Lymphoma Myeloma Leuk.* **21**, 8-20.e3 (2021).

280. Wen, T., Wang, J., Shi, Y., Qian, H. & Liu, P. Inhibitors targeting Bruton's tyrosine kinase in cancers: drug development advances. *Leukemia* **35**, 312–332 (2021).

281. Jebaraj, B. M. C. & Stilgenbauer, S. Telomere Dysfunction in Chronic Lymphocytic Leukemia. *Front. Oncol.* **10**, 3062 (2021).

282. Thompson, P. A. *et al.* Complex karyotype is a stronger predictor than del(17p) for an inferior outcome in relapsed or refractory chronic lymphocytic leukemia patients treated with ibrutinib-based regimens. *Cancer* **121**, 3612–3621 (2015).

283. Le Bris, Y. *et al.* Major prognostic value of complex karyotype in addition to *TP53* and *IGHV* mutational status in first-line chronic lymphocytic leukemia. *Hematol. Oncol.* **35**, 664–670 (2017).

284. Greenwell, I. B. *et al.* Complex karyotype in patients with mantle cell lymphoma predicts inferior survival and poor response to intensive induction therapy. *Cancer* **124**, 2306–2315 (2018).

285. Anderson, M. A. *et al.* Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood* **129**, 3362–3370 (2017).

286. Brown, J. R. *et al.* Extended follow-up and impact of high-risk prognostic factors from the phase 3 RESONATE study in patients with previously treated CLL/SLL. *Leukemia* **32**, 83–91 (2018).