The Genetics and Kinetics of *BCL2* Driven Lymphoid Malignancies

Philip Webster

A Thesis Submitted to Imperial College London for the Degree of Doctor of Philosophy

December 2014

Cancer Genomics Group

Medical Research Council, Clinical Sciences Centre

Imperial College London

Declaration of Originality

I declare that the work presented in this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Copyright Declaration

Signe  **Date:** 08th December, 2014

Abstract

Introduction: Non-Hodgkin Lymphoma (NHL) is rising in incidence. Treatment of this genetically heterogeneous disease has toxic side effects and significant numbers of relapsers / non-responders. BCL2, an anti-apoptotic protein, is commonly overexpressed in NHL as a result of the t(14;18) translocation. A number of BCL2 inhibitors have shown success in clinical trials but variable efficacy has meant that none have been licenced for use.

Methods: Retroviral insertional mutagenesis (RIM), using Moloney Murine Leukaemia Virus (MoMuLV) in transgenic mice overexpressing BCL2, was used to identify putative target genes deregulated alongside *BCL2* in lymphomagenesis. This project aimed to update MoMuLV integration site identification and sequencing, allowing quantification of integration site clonal abundance. Cohorts of mice were sacrificed at time points prior to disease onset in order to interrogate integration site kinetics. To test the oncogenic potential of candidate genes, C57BL/6 *Vav-BCL2* $p53^{+/-}$ mouse B cells were retrovirally transduced with genes of interest and transplanted into mice to study the speed of lymphoma onset.

Results & Conclusions: A novel, high throughput, quantitative library preparation and sequencing protocol compatible with an Illumina platform was validated. RIM screening in *BCL2* transgenic and wild-type mice identified different insertion sites profiles, detecting known oncogenes and tumour suppressor genes as well as novel candidate genes involved in pathways of lymphoid organ development, B-cell activation and differentiation. Study of insertion kinetics over time showed three patterns of clonal abundance and also allowed the study of specific gene deregulation prior to disease onset. Overexpression of *Cd86* slowed disease onset whilst *Ildr1* expedited disease onset

suggesting the former is a tumour suppressor gene and the latter an oncogene. Discovering genes mutated with *BCL2* in lymphoma may help to explain the lack of efficacy of BCL2 inhibitors and also identify novel therapeutic targets.

Acknowledgements

I am grateful for the opportunity and financial support provided by the MRC and Imperial College London who have made my fellowship, and this work, possible.

Throughout this project I have been conscious that mice, although imperative in answering the questions posed, are sentient animals. I believe that such research remains vital in forwarding our understanding and treatment of human diseases, and wanted to acknowledge that without them, this work would not have been possible.

First and foremost I would like to thank my supervisor, Anthony Uren, who must now be jaded from my never-ending pestering and questioning.

Thanks also to the rest of the Cancer Genomics group including Jo Dawes, Hamlata Dewchand, Barbara Iadarola, Bruce Bolt, Katalina Takacs and Jakub Kaczor for your practical help and making my time enjoyable.

I would like to express my gratitude to Laurence Game (the head of the sequencing facility), James Elliot (the head of the FACS facility) and Kikkeri Naresh (Professor of haematopathology) for your time and effort in assisting the development of new methods for my study.

Finally I would like to thank my wife, Louise, and my daughters, Emma and Becky – I don't know how you have put up with me.

Abbreviations

APC	Antigen presenting cell
ATL	Adult T-cell leukaemia/lymphoma
BCL2	B cell lymphoma 2
B-CLL	B-cell chronic lymphocytic leukaemia
BFP	Blue fluorescent protein
bp	Base pairs
CML	Chronic myeloid leukaemia
CMV	Cytomegalovirus
DLBCL	Diffuse large B cell lymphoma
DNase1	Deoxyribonuclease 1
FACS	Fluorescence-activated cell sorting
FL	Follicular lymphoma
GC	Germinal centre
GFP	Green fluorescent protein
Gluc	Gaussia luciferase
GO	Gene ontology
HTLV-1	Human T-lymphotropic virus type-1
IGH	Immunoglobulin heavy locus
i.p.	Intra-peritoneal
LTR	Long terminal repeat
MLV	Murine leukaemia virus
MoMuLV	Moloney murine leukaemia virus
MSCV	Murine stem cell virus
NGS	Next generation sequencing
NHL	Non-Hodgkin's lymphoma
NSG	NOD-scid IL2Rγ^{null} / NOD-scid-gamma
PBS	Phosphate-buffered saline
qPCR	Quantitative PCR
R-CHOP	Rituximab, cyclophosphamide, oncovin (vincristine), prednisolone
RCR	Replication competent retrovirus
WT	Wild-type

Table of Contents

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

1.1 Cancer

331,000 people in the UK were diagnosed with cancer in 2011. The incidence of all cancers has risen by 23% in males and 43% in females since the mid-1970s (http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/all-cancers-combined/). The development of cancer requires multiple genetic and epigenetic changes that can promote proliferative signalling, evade growth suppressors, promote cell survival, stimulate cell replication, induce angiogenesis and activate invasion and metastsasis (Hanahan & Weinberg, 2011). Identifying genes that are mutated or deregulated to promote cancer can help to unravel mechanisms of oncogenesis, aid diagnosis, indicate prognosis and also identify novel therapeutic targets. An example of this approach to drug development is the tyrosine kinase inhibitor, Imatinib (Gleevec / Glivec). The Philadelphia chromosome translocation BCR-ABL t(9;22) (q34;q11) is found in 90% of chronic myeloid leukaemia (CML). Imatinib inhibits the ABL kinase domain of the BCR-ABL fusion oncoprotein to effectively treat t(9;22) positive CML (Druker et al., 2001). Another example is PLX4032, a BRAF inhibitor, which facilitates complete or partial tumour regression in patients with metastatic melanoma who carry the V600E mutation in BRAF, that causes constitutive activation of BRAF and downstream MAP kinase pathway activation (Flaherty et al., 2010).

1.1.1 Lymphoma

Lymphoma, simply described, is a cancer of lymphocytes. Symptomatology is wide-ranging and non-specific, some patients getting few symptoms and some many.

Lymphoma can cause enlargement of any lymph node in the body, therefore causing local symptoms that can affect any organ system. As the disease progresses it can cause a wide range of systemic symptoms including fevers, night sweats, weight loss, fatigue and recurrent infections. Approximately 20 new cases of lymphoma per 100,000 people per year are diagnosed in the western world. There are histologically two main types; Hodgkin's lymphoma accounts for approximately 15% of diagnoses and generally has favourable outcomes. Eighty-five percent are Non-Hodgkin lymphoma (NHL) which is a particularly heterogenous group of malignancies with over 40 subtypes that derive from the different stages of B-cell development (Figure 1-1) (Lowry & Linch, 2013). NHL is the fifth most common cancer in the UK with 400 new diagnoses per year at Imperial College Healthcare NHS Trust which is a regional centre for the disease. Ninety-five percent are B-cell in origin. In the UK the age-standardised incidence rate has increased three-fold over the last 30 years, a figure that is comparable to other developed countries (http://info.cancerresearchuk.org/cancerstats/types/nhl/incidence/).

Follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLBCL) are the most common subtypes of B-cell lymphoma (20% and 30-40% respectively, (Küppers, 2005)). FL is a proliferation of malignant germinal centre B-cells (GCBs) with centroblasts and centrocytes being the predominant cell type. It often has an indolent course although cure rates are low and relapses are frequent (Kridel, Sehn, & Gascoyne, 2012). Survival rates vary depending on the 'Follicular Lymphoma International Prognostic Index' score which is divided into low, medium and high risk with overall five year survival being 91%, 51% and 36% respectively (Salles, 2007). Patients who are asymptomatic or have a low tumour burden may have no treatment, adopting a

'watch and wait' strategy. However, it has the potential to transform to the more aggressive, high grade DLBCL which has an overall five year survival spanning 73% to 26% for the low to high risk groups respectively (Sweetenham, 2005). Treatment of B-cell lymphoma often includes some or all components of rituximab, cyclophosphamide, doxorubicin (hydroxydaunorubicin), vincristine (oncovin) and prednisolone (collectively referred to as 'R-CHOP'). This may be followed by autologous or allogeneic stem cell transplantation. Whilst this combination is successful, it is not patient specific, there are still a significant proportion of relapsers / non-responders, and there are many associated severe side-effects. Cyclophosphamide may lead to infertility, pancreatitis, haemorrhagic cystitis, cardiotoxicity, hepatotoxicity and nephrotoxicity. Doxorubicin can cause diarrhoea, nephrotoxicity and cardiotoxicity. Vincristine is neurotoxic and may also cause intestinal necrosis, diarrhoea and eye disorders. Steroids have a multitude of side effects that include hypertension, impaired glucose tolerance, bruising, osteoporosis, infections and psychosis (http://www.bnf.org/). There is therefore a need to develop patient-personalised treatments as well as more targeted therapeutics.

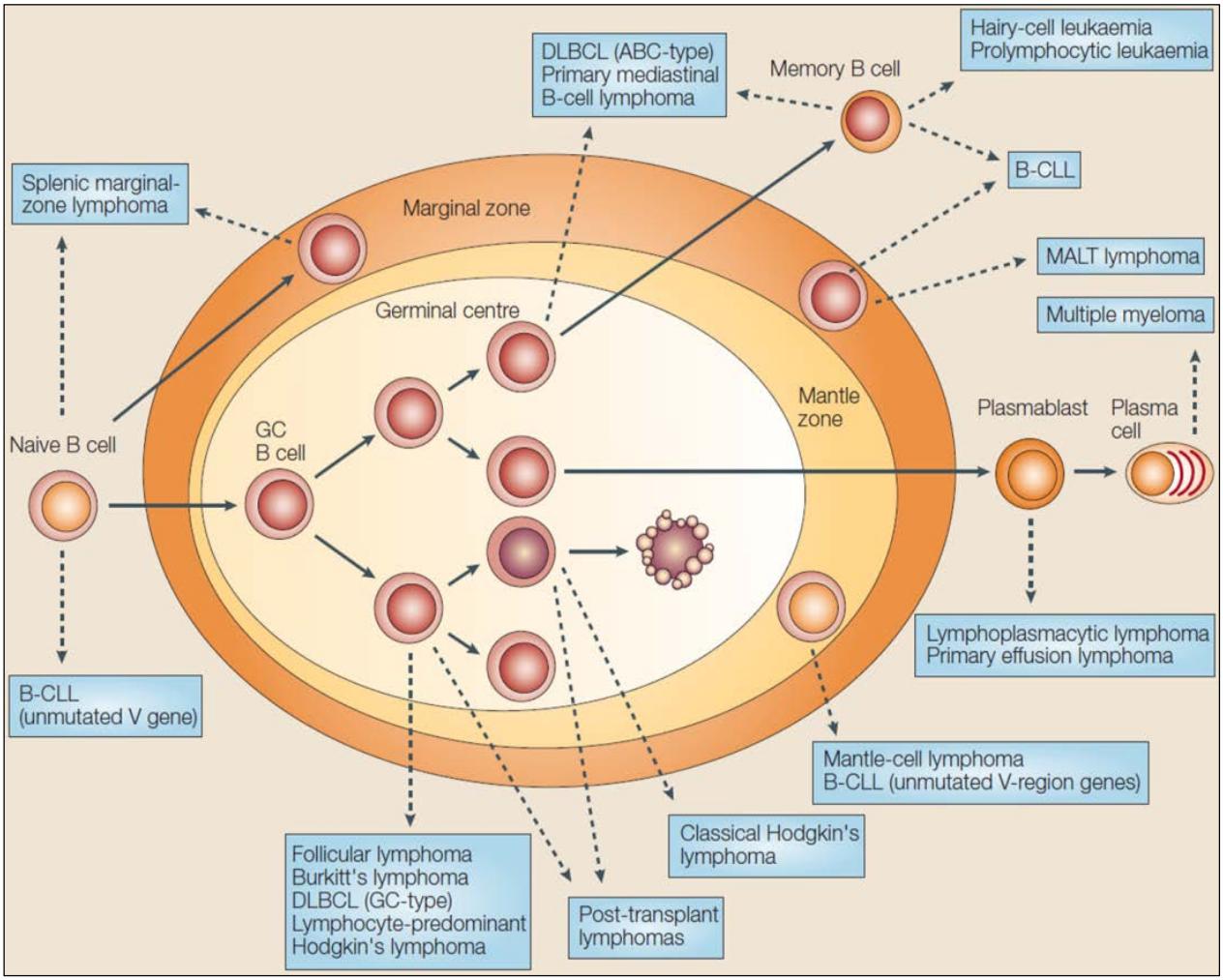


Figure 1-1 Types of lymphoma according to cellular origin
(From Küppers, 2005) Many leukaemias and lymphomas are derived from B-cells at various stages of development. Ninety-five of lymphomas are B-cell in origin. B-cells within, or that have passed through, germinal centres are the source of most lymphomas. Diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) are the commonest subtypes of non-Hodgkin lymphoma, which itself accounts for 80% of all lymphomas.

1.2 Lymphoma Genetics

NHL, Hodgkin lymphoma and other cancers show familial aggregation (Chatterjee et al., 2004). However, twin studies do not support the role of highly penetrant genes conferring risk in NHL. There has been a huge drive to research the complex genomic and epigenetic basis for lymphoma over recent years. Reciprocal chromosomal translocations between a proto-oncogene and immunoglobulin loci are found in many lymphomas. Normal B-cells differentiate from haematopoietic stem cells, involving

several stages of maturation that involve genetic recombination and mutation. Recombinase activating genes (RAG-1 and RAG-2) initiate DNA double-strand breaks during immunoglobulin and T cell receptor production, allowing VDJ gene segments (Variable / Diversity / Joining) to recombine in a multitude of different ways in order to generate antigen receptors capable of binding a vast number of possible antigens (J. H. Wang et al., 2009). However, this process can be a source of chromosomal translocations and mutations involving the immunoglobulin heavy (IgH) and light (IgL) chain genes and the T cell receptor genes. A subset of these can deregulate the expression of oncogenes. Further aberrations can occur during somatic hypermutation and class switching. Collectively these mutations and rearrangements can lead to the production of immortalised, proliferating, malignant B-cells.

The t(14;18) translocation which drives overexpression of *BCL2* is the hallmark of FL and is also found in DLBCL and a small proportion of chronic lymphocytic leukaemias (CLL). However, healthy individuals also carry this mutation (Limpens et al., 1995) suggesting that it is insufficient alone. Lymphomas also carry translocations which promote overexpression of *MYC* (t(8;14)), *BCL6* (t(3;14)) and *CDKN2A* (t(9;14)(p21;q32)) (Willis & Dyer, 2000). Over the past two decades lymphoma has been associated with a number of genetic polymorphisms affecting a diverse range of functions including DNA repair, one carbon metabolism, immune regulation, chemokines, oxidative stress, energy regulation, cell cycle regulation and hormone production. However, these were all identified by case-control association studies that were limited by population size and require larger collaborative studies (Skibola, Curry, & Nieters, 2007).

In the last five years, expression profiling and large-scale deep sequencing analysis of lymphoma has contributed enormously to our understanding of somatic mutations that drive lymphoma. DLBCL has 3 main subgroups with distinct genetic lesions. The groups reflect B cells at different stages of differentiation and are germinal-centre B-cell-like (GCB) DLBCL, activated B-cell-like (ABC) DLBCL and primary mediastinal B-cell lymphoma (PMBCL) (Alizadeh et al., 2000; Rosenwald et al., 2002). Pasqualucci *et al* made a significant impact within DLBCL. They performed massively parallel sequencing and also copy number analysis of tumour and matched normal DNA from 6 patients with *de* novo DLBCL. They reported mutations were mainly single nucleotide substitutions but also in-frame insertions or deletions, nonsense mutations, alterations in canonical splice sites and frameshift deletions. They showed that more than 30 clonal mutations occurred in each case, confirming many genes already implicated in lymphomagenesis and also 26 new genes, including those that methylate chromatin and those controlling immune recognition by T cells. Copy number analysis showed the highest number of lesions at chromosomes 1, 2, 3 and 6q (Pasqualucci et al., 2011). Some of the most commonly found mutated genes include *MLL2, CREBBP, TP53, MYOM2* and *TNFAIP3*.

Morin *et al* sequenced tumour and matched normal DNA from 13 cases of DLBCL and one FL along with RNA-seq from 113 NHLs, resequencing the original 14 cases to confirm 109 genes with somatic mutations. They found genes controlling histone modification were frequent targets for mutation including *MLL2, MEF1B, CREBBP* and *EP300* (Morin et al., 2011). In 2013 the same group performed whole genome sequencing of 40 DLBCL cases and 13 cell lines, combining the data with DNA copy number analysis and RNA-seq from an extended cohort of 96. They showed somatic

single nucleotide variants (SNVs) in *GNAI2* and *GNAI3*, loss of CDKN2A by chromothripsis and also commented on the relative temporal order of mutation acquisition based on calculations on integer copy number data. They also found that amplifications of *BCL2* occurred early, *REL* amplification started early but underwent continued increases over time and that driver mutations in *TP53*, *CARD11*, *MYD88*, and *CD79B* could all be acquired in later stages of tumourigenesis (Morin et al., 2013).

Cytogenetic studies and comparative genomic hybridization (CGH) arrays of FL have identified recurrent abnormalities including gains of chromosomes 7, 12, 18, and X, and losses of 6q and 1p (Cheung et al., 2009, 2010; d'Amore et al., 2008; Höglund et al., 2004; Horsman, Connors, Pantzar, & Gascoyne, 2001; Ross, Ouillette, Saddler, Shedden, & Malek, 2007). Bouska *et al* corroborated these findings in 277 FL and transformed FL lymphoma samples. They found that abnormalities associated with disease transformation are more likely to affect immune surveillance, B-cell transcription factors and both NF-κB and p53 pathways. They also found that abnormalities in chromosomes 6 and X are predictive of overall FL survival (Bouska et al., 2014).

Whilst our knowledge of lymphoma genetics has vastly improved in recent years, the heterogeneity of the disease means that there is a long way to go prior to understanding the full genetic landscape. Human sequencing studies are limited in their ability to identify genes driving the selection for large-scale copy number aberrations, aneuploidy or finding which genes are deregulated in expression by non-coding mutations / epigenetic deregulation.

1.3 Mouse Models of Cancer

Modelling human disorders in mice began more than a century ago and has been fundamental in our knowledge, understanding and treatment of human diseases, especially cancer (Lunardi, Nardella, Clohessy, & Pandolfi, 2014). Compared to humans, mice are small in size, (relatively) inexpensive to maintain, highly reproducible due to the use of inbred strains minimising genetic variation, and genetically modifiable. There is a strong genetic similarity between mice and humans, and 99% of mouse protein-coding genes share an equivalent homolog in humans. However, the percentage of sequence conservation between the two is approximately 5% indicating that conserved non-protein coding sequences are also under selective pressure (Guénet, 2005). In the initial years of studying cancer in mouse models, the development of spontaneous tumours in different inbred mouse strains were observed. Genetic manipulation of mice then revolutionised the field and the first stable and transmissible insertion of DNA into the mouse genome occurred in 1976 (Jaenisch, 1976). Transgenic mice have allowed the study of the *in vivo* gene function, genomics and epigenetics of diseases. There has been significant advancement in gene manipulation which has led to the development of a variety of genetically modified mice including transgenes under ubiquitous or tissue specific promoters, knock-outs, knock-ins, conditional alleles and regulatable alleles (Cheon & Orsulic, 2011). This has allowed study of loss or gain of gene function, either in specific tissues or ubiquitously, in an inducible or conditional system.

1.3.1 Insertional Mutagenesis

Somatic insertional mutagenesis in mice is used in forward genetics screens to identify mutations which promote specific cancers. Either retroviruses or transposons are

introduced into an animal and cause host gene mutation / deregulation (Copeland & Jenkins, 2010), and after a period of time that animal develops cancer. Integration sites of the virus or transposon are identified and those sites that occur in multiple independent tumours more frequently than expected by chance are termed common insertion sites (CIS). Genes identified at these sites may be implicated in oncogenesis and should be prioritised for further study.

Insertional mutagenesis can be performed using acute or slow transforming retroviruses, and also by using transposons. Acute transforming retroviruses cause polyclonal tumours within weeks by causing the overexpression of onco-proteins whereas slow transforming retroviruses form oligoclonal tumours after a longer latency by inserting next to host genome oncogenes or tumour suppressor genes and causing somatic mutations. Transposons are sequences of DNA that move independently around the genome. There are two classes including retrotransposons that move around by a 'copy and paste' mechanism and DNA transposons that move by a 'cut and paste' mechanism and rely on using the tranposase enzyme. DNA transposons are only active in plants and invertebrates and so this lead to the development of the *Sleeping Beauty* (SB) transposon in 1997 that consists of a tranposase and a transposon and is active in mammals (Ivics, Hackett, Plasterk, & Izsvák, 1997). PiggyBac (PB) is another transposon, from the cabbage looper moth, that is active in mammals (S. Ding et al., 2005). PB can move larger segments of DNA than SB. The main differences between virus and transposon mutagenesis are integration bias profiles. Murine leukaemia virus is biased towards the transcription start site of genes and CpG islands and are also inclined towards actively transcribed genes. PB has a strong prejudice for inserting at transcription start sites and at TTAA sequence motifs whereas SB enriches evenly along

the body of genes and inserts mostly at TA dinucleotide sites (de Jong et al., 2014). SB and to a lesser extent PB have a tendency to hop locally. As such transposon mutations that have undergone clonal selection may remobilise causing an enrichment of subclonal mutations nearby the original site.

Moloney Murine Leukaemia Virus (MoMuLV) is an ecotropic, slow transforming, RNA retrovirus and when inoculated in newborn mice, which have immature immune systems, they develop leukaemia / lymphoma. It infects cells through binding of virus envelope proteins to cell surface receptors. After the provirus inserts into the genome of the host cell, this cell produces viral envelope proteins which occupy the surface receptors to prevent reinfection. However recombination with endogenous viruses produces virus mutants that can reinfect cells via different receptors (Nethe, Berkhout, & van der Kuyl, 2005). This can happen several times over, leading to numerous mutations.

A subset of virus integrations will disrupt or deregulate the expression of nearby genes and those that are advantageous to cancer cause tumour outgrowth. The mutation of host genes is mediated by elements located in the long terminal repeats (LTRs), which contain three regions, at each end of the provirus (Rein, 2011). The U3 region has two parts: a promoter that contains sequences involved in recruiting basal transcription machinery, and an enhancer that contains binding sites for transcription factors. The R domain encodes 5' capping sequences and the poly adenylation (polyA) signals. The U5 region controls packaging of viral RNA into virions and also reverse transcription. Virus replication is restricted to environments that contain the transcription factors that bind to a virus LTR hence the tropism of MoMuLV for lymphocytes (see Figure 1-2).

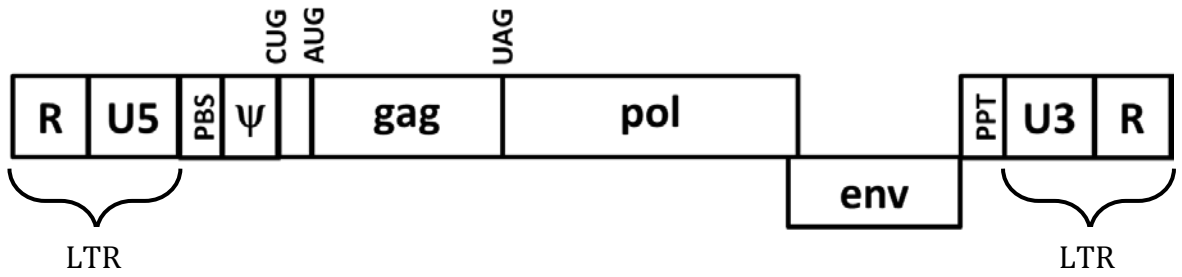


Figure 1-2 The Moloney Murine Leukaemia Virus Genome
(From Rein, 2011). The MoMuLV genome is 8332 nucleotides in length. The coding regions of the genome include Gag, Pol and Env which encode the viral capsid, synthesise viral DNA and facilitate entry of virus particles into host cells respectively. Pol and Env overlap by 58 bases. The noncoding regions are essential for virus function. The PBS ('primer binding site') is complementary to the host cell tRNA molecule, used for reverse transcription. During reverse transcription, the PPT ('polypurine tract') is resistant to degradation and acts as the primer for second strand synthesis of DNA. The ψ 'packaging signal' gives retroviral RNAs the priority to be encapsidated over other cellular mRNAs. The 'R' domains encode capping signals and polyA signals.

When MoMuLV inserts into the host genome it can mutate and deregulate genes in a number of ways (reviewed in A G Uren, Kool, Berns, & van Lohuizen, 2005). Enhancer insertions are common; they may target neighbouring genes or even those genes large distances away acting via chromatin loops, and usually cause upregulation of gene expression. They are usually found upstream of a gene in the antisense orientation or downstream in the sense orientation. Promoter insertions occur when the virus inserts into the promoter region of a target gene, in the same transcriptional orientation as the gene, placing it instead under the control of the virus promoter and usually causing

overexpression. Intragenic insertions can interfere with gene splicing, leading to translation of truncated or chimeric transcripts, or may even contain polyA signals that stop transcription of the target gene altogether, thereby inactivating it. By generating a cohort of mice and identifying the retroviral integrations, the relative frequency of these mutations indicates which are oncogenic (driver mutations) and which are coincidental (passenger mutations). Previous screens have identified many genes relevant to lymphoma including the known human oncogenes c-Myc, N-Myc, Notch1 and Flt 3 (Kool & Berns, 2009). Oncogenes are more commonly found than tumour suppressor genes, which usually requires loss of both alleles to be effective. Identification and sequencing of insertion sites has so far been performed using splinkerette PCR with shotgun sub cloning and 454 pyrosequencing.

The onset of cancer requires numerous mutations to occur. Retroviral insertional mutagenesis screens can be performed in animals with predisposing oncogenic lesions in their germ line. Virally induced somatic mutations accumulate, leading to lymphoma, and CISs in this case may identify gene mutations that are selected to cooperate with the predisposing lesion to promote cancer. This principle has been used in previous screens of Emu-Myc transgene driven lymphoma models and identified Pim1 and Bmi1 insertions as cooperating events with Myc in lymphomagenesis. Overexpression of Myc promotes cell proliferation to promote onset of cancer. However in mouse models bearing the Myc transgene, loss of a single copy of Bmi1 or both copies of Pim1 delays the onset of lymphoma (Jacobs et al., 1999; van der Lugt et al., 1995). These screens can therefore identify proteins, which when lost, delay the onset of lymphomagenesis. This confirms the role of established mutations in tumours and also identifies new mutations to search for in human tumours. This knowledge of gene interaction is valuable to

further our knowledge of molecular mechanisms in oncogenesis and may also identify new therapeutic drug targets.

1.4 BCL2

1.4.1 BCL2 in apoptosis

Altruistic cell death has an important physiological role in maintaining tissue homeostasis and avoiding disease throughout the lives of humans by ridding the body of unwanted cells. Apoptosis (or 'programmed cell death') has been recognised for over four decades (Kerr, Wyllie, & Currie, 1972). The process involves the shrinking of cell nuclei, mitochondria and cytoplasm which become encased in apoptotic bodies. Inappropriate apoptosis is implicated in many diseases: too little can lead to cancers and autoimmune diseases, too much can lead to ischaemia and degenerative diseases.

'B-cell lymphoma 2' (BCL2) is a 26KDa protein and is the prototypical member of a family of proteins that potentiate or inhibit the mitochondrial pathway of apoptosis, BCL2 itself being an inhibitor of apoptosis (Figure 1-3). There are 3 subclasses of proteins in this family that share units of sequence homology – BH 'BCL2 homology' domains. These BH domains consist of a helical bundle surrounding a core hydrophobic groove which is a surface for key family member interactions. BCL2, BCL-X_L, BCL-W, MCL1 and A1 are all anti-apoptotic. BAX, BAK and BOK have four BH domains in common and are pro-apoptotic effector proteins. As suggested by the name, BH3-only proteins share only one BH domain with the other family members. BIM and BID neutralise anti-apoptotic proteins and activate the pro-apoptotic effector proteins.

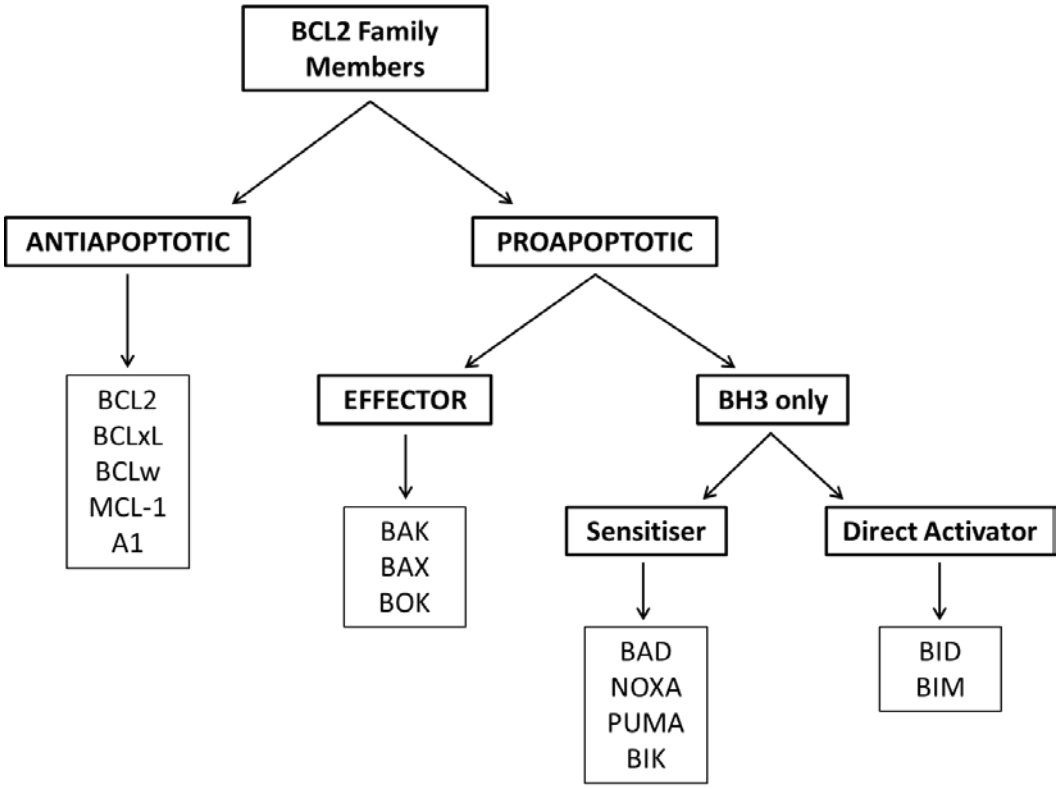


Figure 1-3 The BCL2 family of proteins

This intrinsic pathway of apoptosis is activated by a number of extracellular factors (UV light, chemotherapeutics, hypoxia etc) which alter BCL2 family member levels. Pro-apoptotic members (e.g. BAX) are up-regulated and promote mitochondrial outer membrane permeabilisation (MOMP) allowing leakage of cytochrome c into the cytosol which activates Apoptotic Protease Activating Factor 1 (APAF-1). A cascade of caspase activation concludes with effector caspase 3 activation leading to the apoptotic phenotype of membrane blebbing, cell shrinkage, nuclear fragmentation, chromatin condensation, and chromosomal DNA fragmentation (See Figure 1-4). Antiapoptotic BCL2 family members inhibit the ability of proapoptotic members to activate this cascade.

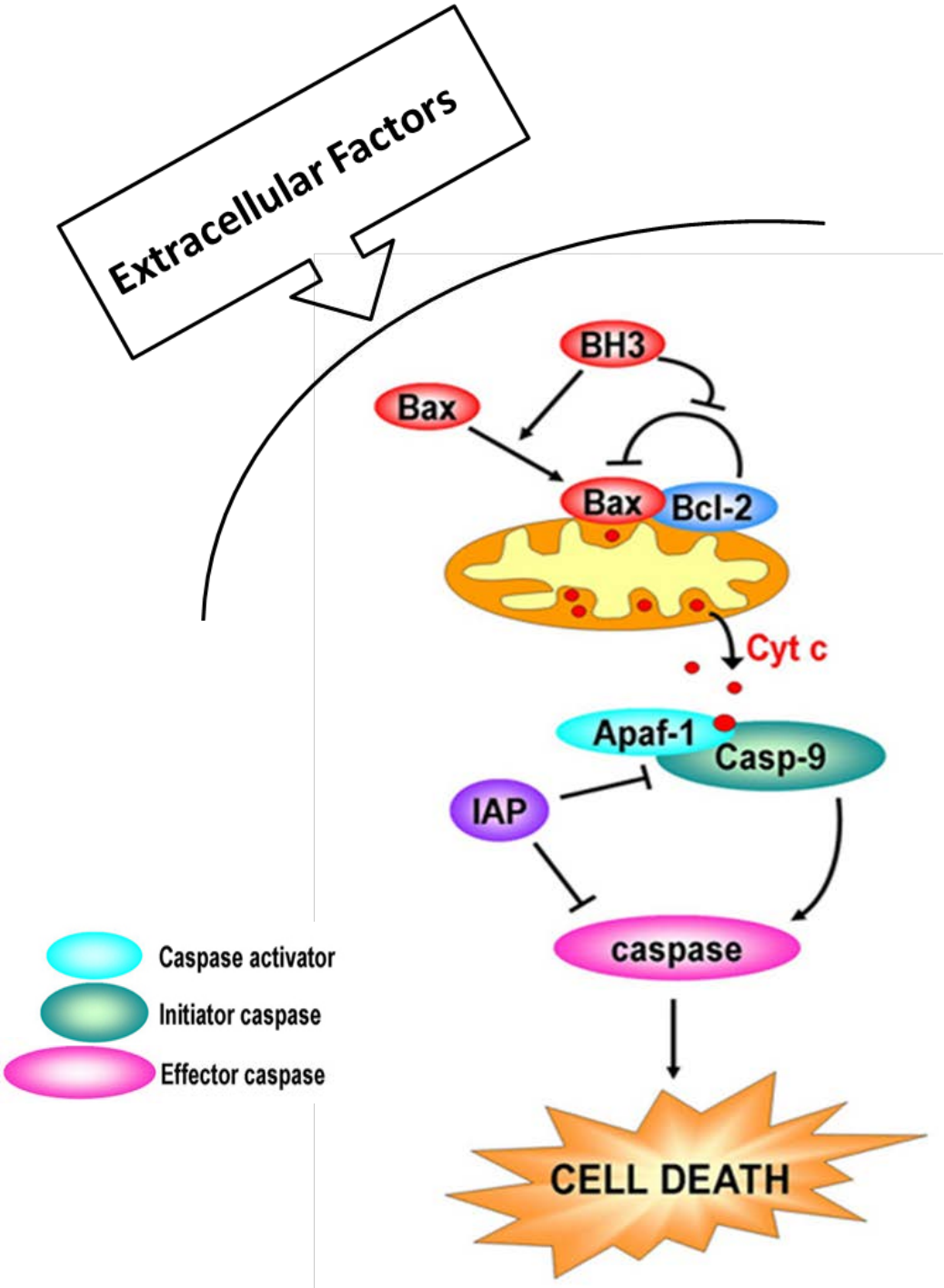


Figure 1-4 The Intrinsic Pathway of Apoptosis
(Adapted from Colin, Gaumer, Guenal, & Mignotte, 2009). A basic over view of the role of BCL2 family members in the intrinsic (or mitochondrial) pathway of apoptosis. Extracellular factors such as ultraviolet light, chemotherapy, hypoxia or toxins act at the surface of a cell. Propapoptotic members e.g. Bax are upregulated and induce mitochondrial outer membrane permeabilisation (MOMP) which allow leakage of cyctochrome c out of mitochondria into the cytosol which activates apoptotic protease activating factor 1 (APAF1) which in turn leads to a cascade of caspase activation and ultimately the effectors caspases which facilitate cell death. Apaf-1 = Apoptotic protease activating factor 1, Casp = Caspase, IAP = Inhibitor of apoptosis.

1.4.2 BCL2 in disease

The BCL2-IGH t(14;18)(q32;q21) translocation is found in 90% of FL and 35% of DLBCL (Küppers, 2005). This translocation places BCL2 under the control of the IGH promoter sequence causing overexpression of BCL2 thus down-regulating apoptosis and promoting cell survival. This mutation is thought to occur early in the disease process (Tsujimoto, Gorham, Cossman, Jaffe, & Croce, 1985) and also to confer a poorer prognosis and resistance to chemotherapy in DLBCL (Barrans, 2002). This translocation requires RAG recombinase which cleaves DNA in the IgH locus and at a number of possible breakpoints in *BCL2* (Raghavan, Swanson, Wu, Hsieh, & Lieber, 2004). However the *BCL2* translocation is found in approximately a quarter of healthy individuals (Schmitt et al., 2006), implying that other factors are necessary for lymphomagenesis.

Overexpression of BCL2 has been implicated in a number of other cancers including chronic lymphocytic leukaemia (CLL), acute lymphoblastic leukaemia (ALL), small cell lung cancer (SCLC) and prostate cancer (reviewed in Kluck, 2010). It is also implicated in the development of autoimmune diseases such as systemic lupus erythematosus in mouse models (A. Strasser et al., 1990).

There has been considerable effort to target therapy against BCL2. The anti-sense oligonucleotide G3139 'Genasense' targets the first six codons of *BCL2* mRNA (Kang & Reynolds, 2009). It was found to be a successful adjunctive therapy in patients with melanoma and relapsed / refractory CLL but failed to be effective in SCLC (Harazono, Nakajima, & Raz, 2013). Although it has not yet been FDA approved it is still under trial. GX-15-70 (Obatoclax) binds all anti-apoptotic BCL2 family members and overcomes resistance due to MCL1 expression. While showing promising results *in vitro*, it

displayed minimal clinical activity in phase II trials in patients with relapsed / refractory Hodgkin's lymphoma or relapsed SCLC (Oki et al., 2012; Paik et al., 2011). It also caused neurological side-effects (O'Brien et al., 2009). ABT-737 is small molecule inhibitor of BCL2 family members. Its structure mimics BH3 domains of proapoptotic family members, and in doing so inhibits BCL2, BCLW and BCLXL and was found to induce regression of solid tumours (Oltersdorf et al., 2005). ABT-737 was found to kill cells in a BAX/BAK dependent manner, although cells expressing the anti-apoptotic MCL-1 were resistant (van Delft et al., 2006). Bardwell *et* al found the BCL2 antagonist ABT-737 to be efficacious in treating animal models of arthritis and lupus (Bardwell et al., 2009). Drug delivery was also problematic leading to development of the orally bio-available ABT-263 (Navitoclax) which showed promise as an adjunct in phase I studies of lymphoma, myeloma and SCLC (Gandhi et al., 2011; Tse et al., 2008; Wilson, Hernandez-Ilizaliturri, Dunleavy, Little, & O'Connor, 2010; Wilson, O'Connor, et al., 2010). Both cause dose dependent thrombocytopenia and their therapeutic windows need clarifying. Recently ABT-263 was re-engineered to ABT-199 which is potent, orally bioavailable, BCL2 selective and has been shown to spare platelets. Early studies have been encouraging in patients with CLL (Souers et al., 2013). Whilst many of these drugs have shown potential in early trials, none have been approved and licenced for use. Lack of sufficient efficacy requires the identification and targeting of other genes that operate alongside *BCL2* in lymphomagenesis and so there still remains a need for novel therapeutic treatments of t(14;18) driven lymphoma.

1.4.3 *BCL2* mouse models

1.4.3.1 Eμ-*BCL2*

The malignant potential of *BCL2* has been studied using transgenic mouse models that places its control under elements of the IgH locus (McDonnell & Korsmeyer, 1991; McDonnell et al., 1989; A. Strasser et al., 1990; A. Strasser, Harris, & Cory, 1993). One of these models places the *BCL2* open reading frame under the control of a SV40 promoter and IgH locus Emu enhancer (a Strasser et al., 1991). *BCL2* overexpression is restricted to the B cell lineage of these mice, with no T cell expression. B-cell survival is therefore enhanced and so they have higher numbers of B lymphocytes and plasma cells. Transgenic mice showed up to 5-fold excess of B lymphocytes in spleen, lymph nodes and bone marrow and these mutant cells survived longer *in* vitro compared to wild-type cells. These mice had a low incidence of B-cell neoplasia which were pre-B, immunoblastic and plasmacytomas in origin.

This strain has previously been screened using MoMuLV. Thirty-five mice were infected and insertion characterisation done by southern blot analysis of insertion sites near a small number of known oncogenes. Insertions were found near *c-Myc*, *Pim-1* and *Pim-2* (Shinto et al., 1995).

1.4.3.2 VavP-*BCL2*

The *Vav* proto-oncogene is ubiquitously expressed in nucleated haematopoietic cells but is expressed in very few others (Katzav, Martin-Zanca, & Barbacid, 1989; Ogilvy, Metcalf, Gibson, et al., 1999). The promoter region of *Vav* was first exploited to induce overexpression of *BCL2* in a mouse model in 1999 (Ogilvy, Metcalf, Print, et al., 1999). In this case, BCL2 overexpression was not just restricted to B lineage cells but included all

nucleated cells throughout the haemopoietic compartment with a marked rise in both lymphoid and non-lymphoid cells. In this model, lymphocyte numbers were increased 15-fold due to an increase in both mature and immature B-lineage cells. Mice also developed an enlarged spleen with increased lymphoid follicles and expanded germinal centres. The thymuses of these mice showed a marked reduction in double-positive (CD4 / CD8) T cells which was balanced by rises of the three other CD4 / CD8 subsets. *VavP-BCL2* cells were more resistant to irradiation than control cells.

Egle *et al* studied the long term consequences of this transgene in C57BL/6 mice up to 18 months of age. Approximately 15-25% of these mice developed kidney disease in the form of a crescentic autoimmune glomerulonephritis at around 40 weeks of age. The remaining mice went onto develop lymphoid malignancies most similar to human FL by 18 months of age, with abnormal tissue rich in centroblasts and centrocytes (Egle, Harris, Bath, O'Reilly, & Cory, 2004). Whilst the tumours in this model represent human disease on a morphological level, at a molecular level the high expression of the *BCL2* transgene in the T cell compartment is an artificial situation that is not found in human FL. Both Eμ and Vav models highly expressed the transgene in B-cells so why does the latter develop spontaneous disease whilst the former does not? This group proposed that it was the five-fold increase in CD4 T-cell expression that lead to the Vav model getting FL, by increasing B-cell proliferation and also immunoglobulin class switching, where the Eμ model does not. *VavP-BCL2* mice have not been used in insertional mutagenesis screens but, given the closeness to human disease, would be ideal.

1.4.3.3 Strain: C57BL/6 *vs* BALB/c

Risser *et al* showed that BALB/c and (BALB/c x C57BL/6)F1 mice were more susceptible to developing lymphoma after infection with MoMuLV than the more

resistant C57BL/6 mice, proposing that different genes controlled susceptibility in each strain (Risser, Kaehler, & Lamph, 1985). BALB/c mice show increased numbers of B lymphocytes, Ig-secreting cells and serum Ig, as well as a prolonged antibody response to immunization compared to C57BL/6 mice (Pellegrini et al., 2007). In view of these findings, this project involves the use of an F1 cross between C57BL/6 and BALB/c strains in order to provide a faster model that will also skew the resulting lymphomas to the B-cell lineage.

1.5 Tumour Evolution / Clonality

The evolution of tumours from a single cell and the kinetics of mutation clonality that build to facilitate oncogenesis have been hot topics in the cancer research world for many years (Nowell, 1976). This has exploded in recent times and is set to continue into the future. Next generation sequencing has demonstrated that cancers share common clonal origins and different sub-clones are defined by different mutations that occur later as the cancer evolves leading to vast intra-tumour heterogeneity. These different sub-clones, harbouring different mutations, may contribute to resistance to chemotherapy drugs (Swanton, 2014).

This intra-tumour heterogeneity supports the need for single cell analysis of cancers. Navin *et al* performed next generation sequencing on single cells from two human breast cancer cases in order to quantify genomic copy number within a nucleus. Within one polyclonal tumour they found three distinct clonal subpopulations, presumed to represent sequential clonal expansions. Within a monoclonal primary tumour, a single clonal expansion formed the tumour and the metastasis. Both primary tumours had 'pseudodiploid cells' that did not travel to the site of metastasis. They concluded that the

tumours they studied grew by punctuated clonal expansions rather than gradual progression (Navin et al., 2011).

Ding *et al* sequenced the primary and relapsed tumours from eight patients with acute myeloid leukaemia (AML) along with matched normal skin tissue for each patient. They noticed two patterns of evolution. The first showed that the dominant sub-clone in the primary tumour gained extra mutations and evolved into the relapsed clone. The second showed that a minor sub-clone in the primary tumour, containing many but not all of the primary tumour mutations, survived, gained mutations and expanded at relapse. The first pattern was hypothesised to represent under-treatment, whereas the second may represent a tumour with a specific mutation conferring treatment resistance (L. Ding et al., 2012).

Okosun *et al* performed either whole-genome or whole-exome sequencing on 10 follicular lymphomas that went on to become transformed follicular lymphomas months or years later. They also noted two patterns of evolution from each case of FL (the initial diagnosis of FL was deemed 'early disease') to its paired transformed FL (when the initial FL transformed it was deemed 'late disease'). The pattern in 8 cases showed high similarity of mutation clonality between tumour pairs. The second pattern, in two tumours, showed only four nonsynonymous mutations were shared between tumour pairs. They identified recurrent 'early' driver mutations in chromatin regulator genes (CREBBP, EZH2 and MLL2) and a variety of possible mutations occurring at transformation including in EBF1 and NF-κB signalling (Okosun et al., 2014).

Even with thousands of sequenced cancer genomes the extent to which rare and/or sub-clonal mutations are oncogenic remains unclear, particularly for non-coding

mutations thought to represent the vast majority of somatic mutations. There is likely to be a great deal of contamination of cancer mutation data by statistical anomalies since mutation rates between tumours can vary by several orders of magnitude (between 0.1 and 10 mutations/Mb depending on the tumour type (Lawrence et al., 2013)). Even allowing for this variability, differences in mutation rates can't address the extent to which rare mutations below significance thresholds might still contribute to tumour formation.

This project addresses the issue of "what fraction of starting mutations are likely to be selected for and thus potentially be oncogenic" because we know the distribution of starting mutations in a way that can't be determined for human premalignant cells. By using mice, mutations at different stages of disease development can be studied, allowing interrogation of early events vs clonal expansion.

1.6 Hypotheses and Aims

My project is to perform a retroviral insertional mutagenesis screen in mice possessing a transgene construct causing overexpression of BCL2, using current high throughput sequencing technologies to reveal gene mutations that may cooperate with *BCL2* in lymphomagenesis. Discovery of genes that cooperate with *BCL2* in lymphomagenesis may help explain the mechanism behind the lack of efficacy of anti-BCL2 drugs and provide targets to direct adjuvant therapy to improve efficacy. These treatments may even have wider applicability as BCL2 is also overexpressed in other cancers (Yip & Reed, 2008) and autoimmune diseases (Bardwell et al., 2009) and may confer chemotherapeutic resistance (Kang & Reynolds, 2009).

1.6.1 Hypothesis

The *BCL2-IGH* t(14;18) translocation, which leads to overexpression of BCL2, cooperates with other mutated or deregulated genes to promote lymphomagenesis. Identification of these genes may explain the varied response to BCL2 inhibitors in clinical trials and direct future therapeutic drug targets.

1.6.2 Aims

- To design an up-to-date, cost effective, high-throughput, quantitative method of library preparation applicable to an Illumina platform for sequencing of retroviral insertion mutations.

- To detect the most commonly mutated or deregulated genes that promote lymphomagenesis in *BCL2* transgenic mouse models by insertional mutagenesis.

- To validate candidate genes as either oncogenes or tumour suppressor genes.

- To study the kinetics of mutation profiles as they occur over time, prior to disease, in mouse models of B-cell lymphoma.

- To study heterogeneity of mutation profiles within an organ and between different organs in the same diseased mouse.

CHAPTER 2 MATERIALS AND METHODS

2.1 Animals

Mice were housed and maintained in a controlled environment with *ad libitum* access to food and water. All procedures were performed in accordance with the UK Home Office Animals (Scientific Procedures) Act 1986, (project licence no. 70/7353, personal licence no. 70/23823).

Two *BCL2* transgenic mouse models were used in different genetic backgrounds. C57BL/6 mice heterozygous for a transgene construct consisting of human *BCL 2* cDNA in association with the Eμ immunoglobulin heavy chain enhancer and SV40 promoter in their B-cell lineage (The Jackson Laboratory, C.Cg-Tg(BCL2)22Wehi/J, http://jaxmice.jax.org/strain/002318.html) were bred with wild-type C57BL/6 and BALB/c mice (Charles River, UK) to produce C57BL/6 *Emu-BCL 2* and (BALB/c x C57BL/6) F1 *Emu-BCL 2* transgenic mice. C57BL/6 mice heterozygous for a transgene construct consisting of *BCL 2* cDNA under a Vav promoter (generated at Walter and Eliza Hall Institute of Medical Research and provided by Andreas Villunger, Medical University Innsbruck) were bred with wild type BALB/c mice to produce (BALB/c x C57BL/6) F1 *Vav-BCL 2* transgene.

Vav-BCL2 mice were bred with C57BL/6 $p53^{+/-}$ mice to generate C57BL/6 *Vav-BCL2* $p53^{+/-}$ mice. These mice were used for gene validation studies as described in section 2.7.3.

2.2 Genotyping

2.2.1 Breeding mice

Mice used to breed experimental mice were genotyped using DNA extracted from ear tissue. DNA was extracted using the Puregene® Cell and Tissue Kit (Qiagen; 158388) as per the manufacturer's instructions.

2.2.2 Experimental mice

Experimental mice were genotyped *post-hoc* using DNA from spleen tissue. DNA was extracted using either the AllPrep DNA/RNA 96 Kit (Qiagen; 80311) or the AllPrep DNA / RNA Mini Kit (Qiagen; 80204) as per the manufacturer's instructions. Disposable pestle and mortars were used to disrupt tissues and QIAshredders (Qiagen; 79656) were used to homogenise tissues.

2.2.3 Genotyping PCRs

The *Eμ-BCL2* transgene was detected using the following primers: 5'-TGGATCCAGGATAACGGAGG-3' (forward) and 5'-TGTTGACTTCACTTGTGGCC-3' (reverse). PCR amplification yielded a band of 170bp using the *Taq* DNA polymerase kit (Life Technologies; 18038026). A 25μl reaction contained 1μl DNA (50-100ng/μl), 2.5μl of 10x buffer, 2.5μl of 2mM dNTPs, 1.25μl of 50mM MgCl2, 2.5μl of each primers (10μM), 0.25μl of *Taq* DNA polymerase and 12.5μl of distilled H_2O. Cycling conditions were 95°C for 5min, followed by 29 cycles of 95°C for 30sec, 56°C for 30sec, and 72°C for 45sec, and finally 72°C for 10min.

The *Vav-BCL2* transgene was detected using the following primers: 5'-AGACATGATAAGATACATTGATGAG-3' (forwards) and 5'-CGAAGGGGTTCTCTAGTG-3' (reverse). PCR amplification yielded a band of 294bp using the Phusion Hot Start II High-Fidelity DNA Polymerase kit (Thermo Scientific, F549L). A 20µl reaction contained 1µl DNA (20ng/µl), 4µl of 5x buffer, 0.4µl of 10mM dNTPs, 1µl of both primers (10µM), 0.2µl of Phusion Hot Start II High-Fidelity DNA Polymerase and 12.4µl of distilled H_2O. Cycling conditions were 98°C for 30sec, followed by 31 cycles of 98°C for 5sec, 60°C for 10sec, and 72°C for 15sec, and finally 72°C for 10min.

The *p53*$^{+/-}$ transgene was detected using the following primers: 5'-AGCTAGCCACCATGGCTTGAGTAAGTCTGCA-3', 5'-TTACACATCCAGCCTCTGTGG-3' and 5'-CTTGGAGACATAGCCACACTG-3'. PCR amplification yielded a band of 270bp using the *Taq* DNA polymerase kit (Life Technologies; 18038026). A 25µl reaction contained 1µl DNA (50-100ng/µl), 2.5µl of 10x buffer, 2.5µl of 2mM dNTPs, 0.5µl of 50mM MgCl2, 0.5µl of the first primer and 1µl of the latter two primers (10µM), 0.25µl of *Taq* DNA polymerase and 15.8µl of distilled H_2O. Cycling conditions were 98°C for 3min, followed by 29 cycles of 95°C for 30sec, 61°C for 90sec, and 72°C for 45sec, and finally 72°C for 10min.

2.3 Insertional Mutagenesis

2.3.1 Moloney Murine Leukaemia Virus

2.3.1.1 MoMuLV Production

The pNCA plasmid containing the complete MoMuLV genome was transfected into 293T cells using polyethylenimine. A mixture of 40μg PEI and 10μg DNA in 1ml serum free media was added to a T75 flask of 293T cells (cultured to approximately 60% confluence) in a volume of 9ml of serum free media. This was incubated for 4 hours before replacing the culture media with DMEM containing 20% FCS and pen/strep 50U/50μg per ml. Supernatant containing virus was harvested at 4 days and then again after a further 4 days after passaging at 3:1.

2.3.1.2 MoMuLV Quantification

293T cells modified to possess the ecotropic MoMuLV receptor and also a Gaussia Luciferase (GLuc) reporter gene (a generous gift from the National Cancer Institute, Frederick, MD, USA) were used to quantify virus stocks. 50,000 cells were seeded in wells of tissue culture grade 6 well plates and maintained in DMEM with 10% FCS and pen/strep 50U/50μg per ml. When cells reached approximately 60% confluence all media was aspirated and replaced with 6 dilutions of MoMuLV containing supernatant (10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} and 10^{0} dilutions) with 900mls of fresh media. 20μl of supernatant was removed on day 3 and day 6 and GLuc levels measured on a luminometer in black 96 well plates.

2.3.2 Tumours

2.3.2.1 Tumour Induction & Collection

Newborn pups were injected via the intra-peritoneal (i.p.) route with 50µl MoMuLV stocks at 24-48 hours of age using a repeat dispensing syringe (Hamilton). After 6 weeks of age mice were weighed weekly and monitored three times per week for signs of illness. Mice exhibiting splenomegaly alone, tachypnoea alone, lymphadenopathy alone or 10% weight loss/gain with 2 features of hunched posture / piloerection / withdrawn behavior were sacrificed and organs (spleen, thymus, lymph nodes, bone marrow) were harvested and snap frozen in liquid nitrogen as soon as possible after death. All organs were harvested with sterile, autoclaved, DNA-Exitus Plus (AppliChem; A7089,2500RF) treated instruments to avoid DNA cross contamination between tissue samples. Single cell suspensions of spleen were prepared in all cases using the gentleMACS Dissociator (Miltenyi Biotec).

2.3.2.2 Tumour Time Course

Infected and control litters were sacrificed and organs harvested at predetermined time points prior to disease onset (2, 5, 9, 14, 28, 42, 56, 84 and 112 days). Where mice were 2 weeks old or less, only spleens and livers were harvested due to the technical difficulty of harvesting other organs.

2.3.2.3 Tumour Heterogeneity

A more detailed collection was performed on 12 animals in the (BALB/c x C57BL/6)F1 +/- *Vav-BCL2* cohort. This involved collection of spleen, thymus and bone marrow as well as separate collection of each lymph node from cervical, axillary, inguinal and mesenteric areas on the left and right.

2.4 Hi Throughput Insertion Site Sequencing

2.4.1 DNA Library Preparation

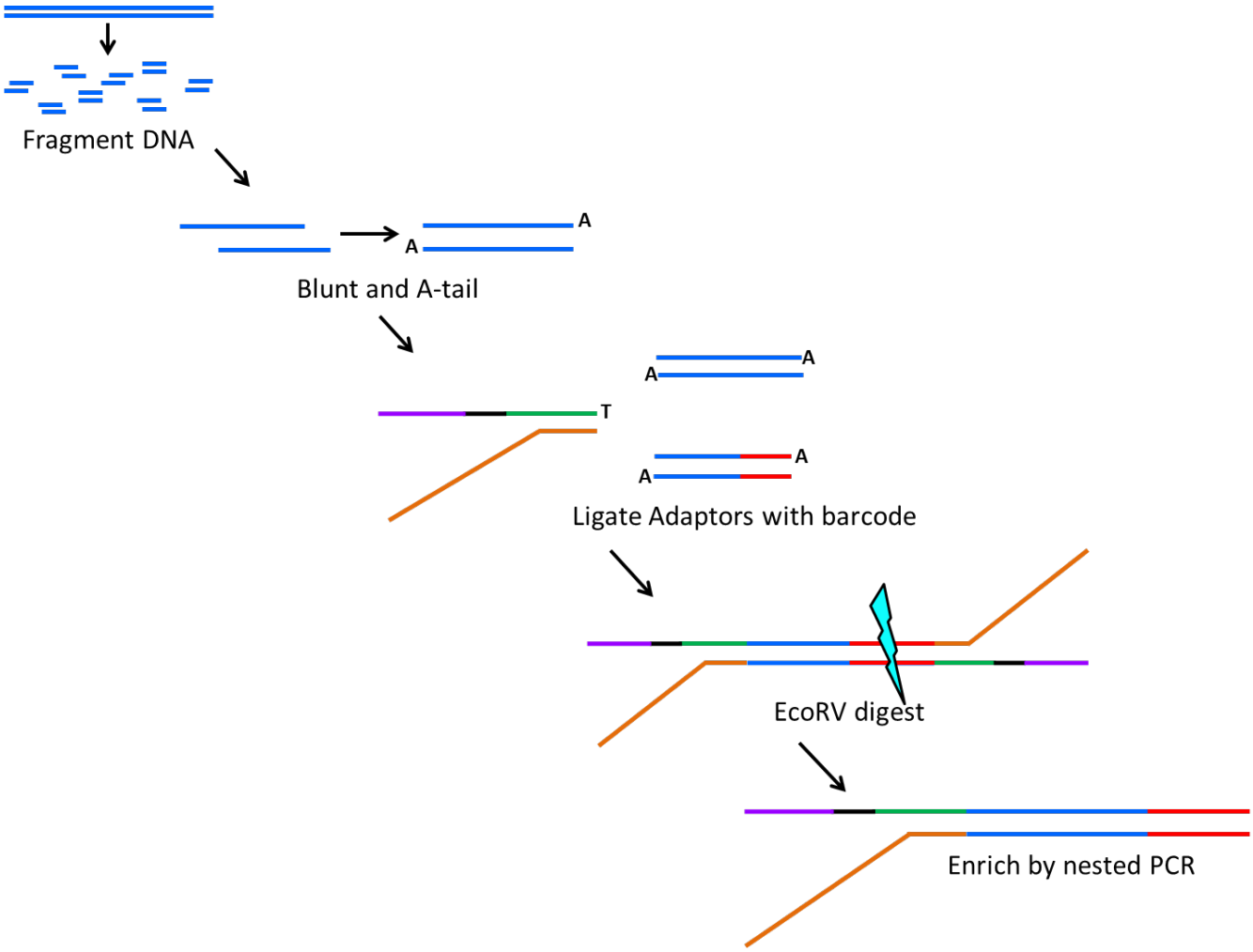


Figure 2-1 DNA library preparation protocol
DNA extracted from the spleens of mice used in the insertional mutagenesis screen was sheared by sonication. Overhanging ends were blunted and then A-tailed. Adaptors containing unique indices were ligated to these fragments and then these samples underwent EcoRV digestion overnight. Fragments containing MoMuLV LTR were enriched by nested PCR during which a second index was added to each fragment to allow a greater number of samples to be sequenced simultaneously. Performed high throughput in a 96 well plate.

2.4.1.1 DNA extraction and shearing

DNA from the spleens of all experimental mice used in the insertional mutagenesis screen were used for library preparation (see Figure 2-1). In a group of these mice DNA was also used from thymus, bone marrow and lymph nodes to look for insertion site heterogeneity across the immune system. DNA was extracted using either the AllPrep DNA/RNA 96 Kit (Qiagen; 80311) or the AllPrep DNA / RNA Mini Kit (Qiagen; 80204) as per the manufacturer's instructions. Disposable pestle and mortars were used to disrupt tissues and the QIAshredder (Qiagen; 79656) was used to homogenise tissues. 96 samples of DNA were processed simultaneously. DNA was quantified using the Qubit® dsDNA HS Assay Kit (Life Technologies; Q32854). Concentration was adjusted to give 55µl of 20ng/µl and was sheared in a Covaris 96 microTUBE™ Plate (LGC Genomics; 520078) on the Covaris E220 Sonicator with the E220 Intensifier (pn500141). A product size of 400bp was achieved with the following settings: Peak Incident Power 175watts, Duty Factor 10%, Cycles per Burst 200, Treatment Time 55sec. Product size was confirmed by running on a 2% agarose gel with ethidium bromide.

2.4.1.2 Blunting, A-tailing, Adaptor Ligation and Fragment Digestion

DNA fragments were blunted using NEBNext® End Repair Module (NEB; E6050L) as per the manufacturer's instructions but with a reaction volume modified for the volume of DNA. A 77µl reaction volume contained 52.5µl of DNA, 7.7µl of 10x reaction buffer, 4µl of End Repair Enzyme Mix and 12.8µl of H_2O. Blunted fragments were then cleaned using Agencourt AMPure XP magnetic beads (Beckman Coulter; A63880) in 96 well format on the Biomek® NXP Laboratory Automation Workstation (Beckman Coulter; A31839). Briefly, the 77µl volume of DNA was added to 90µl of beads and mixed. This reaction was incubated for 10 minutes and then placed on a 96 well magnet for 10

minutes. Supernatant was removed and then the DNA bound to the bead pellet was washed twice with fresh 80% ethanol. The bead pellet was air dried for 5 minutes and then removed from the magnet. The pellet was resuspended in 50µl of distilled water, incubated for 2 minutes and then placed back on the magnet for 5 minutes. 42µl of supernatant containing cleaned DNA was transferred to a new 96 well plate. These samples were then A-tailed using the NEBNext® dA-Tailing Module (NEB; E6053L) as per the manufacturer's instructions. Samples were again cleaned as described above with modified volumes (90µl beads, 50µl DNA sample, 36µl elution). Each of the 96 blunted, A-tailed DNA samples was ligated to a unique adaptor (see Table 2-1) using T4 DNA Ligase (NEB; M0202L) and then digested with EcoRV-HF® (NEB; R3195L) as per the manufacturer's instructions. Samples were then cleaned as described above (50µl beads, 50µl DNA sample, 100µl elution). These DNA fragments, that are now blunted, A-tailed and ligated to an adaptor were then size selected using Agencourt AMPure XP magnetic beads. Beads were added in a 3:5 ratio (bead volume:DNA volume) with the supernatant kept and beads discarded at this stage. Bead volume:DNA volume was then increased to a 1.1:1 ratio and with the supernatant discarded and the bead pellet being kept at this stage. The pellet was washed and eluted as described above.

2.4.1.3 Primary PCR

Fragments of mouse genome also containing MoMuLV insertions were enriched by nested PCR. Primary PCR was performed using the Phusion Hot Start II High-Fidelity DNA Polymerase kit (Thermo Scientific, F549L). A primer to the virus LTR (5'-GCGTTACTTAAGCTAGCTTGCCAAACCTAC -3') and to the index containing adaptor (5'-AATGATACGGCGACCACCGAGATCTACAC -3') were used in a 50µl reaction volume containing 28.5µl DNA, 10µl of 5x buffer, 1µl of 10mM dNTPs, 2.5µl of each primer

(10μM), 0.5μl of Phusion Hot Start II High-Fidelity DNA Polymerase and 5μl (0.1x final concentration) of SYBR® Green I nucleic acid gel stain, 10,000 × in DMSO (Sigma-Aldrich; S9430). qPCR cycling conditions were 98°C for 30sec, followed by 11 cycles of 98°C for 10sec, 66°C for 30sec, and 72°C for 30sec, and finally 72°C for 5min.

2.4.1.4 Secondary PCR

Cleaned primary PCR products were quantified using Qubit® and 50ng was used for enrichment by using the same adaptor primer and an LTR primer nested in relation to that LTR primer used in the primary PCR. This primer also contained a second index in order to create more unique index combinations enabling more samples to be included per MiSeq / HiSeq run (SeeTable 2-2). Reaction volumes and cycling information was the same as for primary PCR. Secondary PCR products were cleaned and then size selected as described above with a final elution volume of 30μl. Each of the 96 samples were quantified using Qubit® and 25ng of each sample used to compile a library.

2.4.1.5 Final Library Compilation

Each library of 96 samples was quantified for amplifiable fragments by qPCR using the KAPA Illumina SYBR Universal Lib Q. Kit (Anachem; KK4824) as per the manufacturer's instructions with DNA dilutions of 1/100, 1/1000, 1/10,000. This kit contains primers complementary to the Illumina sequences at each end of the PCR product used to bind to the flow cell. Equal amounts of each library of 96 were combined to give the final library used for sequencing. The development of these final steps including nested PCR and next generation sequencing (NGS) are discussed in detail in results chapter 4.

	Adaptor Sequence
1	AATGATACGGCGACCACCGAGATCTACACACGCACTCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
2	AATGATACGGCGACCACCGAGATCTACACTGTCGAGCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
3	AATGATACGGCGACCACCGAGATCTACACGAGTGCGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
4	AATGATACGGCGACCACCGAGATCTACACCTACAGTGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
5	AATGATACGGCGACCACCGAGATCTACACCGTGTCTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
6	AATGATACGGCGACCACCGAGATCTACACCTCGCGATATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
7	AATGATACGGCGACCACCGAGATCTACACTAGAGACACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
8	AATGATACGGCGACCACCGAGATCTACACGACACGCGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
9	AATGATACGGCGACCACCGAGATCTACACCGCATAGAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
10	AATGATACGGCGACCACCGAGATCTACACAGACGTATCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
11	AATGATACGGCGACCACCGAGATCTACACCACTACTATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
12	AATGATACGGCGACCACCGAGATCTACACGTATCTCTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
13	AATGATACGGCGACCACCGAGATCTACACTACGTCGTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
14	AATGATACGGCGACCACCGAGATCTACACTAGTACGTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
15	AATGATACGGCGACCACCGAGATCTACACGTACTAGACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
16	AATGATACGGCGACCACCGAGATCTACACGCTACGTAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
17	AATGATACGGCGACCACCGAGATCTACACGAGTAGTACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
18	AATGATACGGCGACCACCGAGATCTACACCTGTAGTCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
19	AATGATACGGCGACCACCGAGATCTACACCTAGTCTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
20	AATGATACGGCGACCACCGAGATCTACACCATACTCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
21	AATGATACGGCGACCACCGAGATCTACACCACGAGAGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
22	AATGATACGGCGACCACCGAGATCTACACCTCGTCTCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
23	AATGATACGGCGACCACCGAGATCTACACCGAGCGACGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
24	AATGATACGGCGACCACCGAGATCTACACACGCGTATGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
25	AATGATACGGCGACCACCGAGATCTACACATACTCGCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
26	AATGATACGGCGACCACCGAGATCTACACACATAGTAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
27	AATGATACGGCGACCACCGAGATCTACACACTGTACAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
28	AATGATACGGCGACCACCGAGATCTACACAGTATAGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
29	AATGATACGGCGACCACCGAGATCTACACAGACGACGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
30	AATGATACGGCGACCACCGAGATCTACACATAGCGTACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
31	AATGATACGGCGACCACCGAGATCTACACAGTACTCTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
32	AATGATACGGCGACCACCGAGATCTACACACGTAGCGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
33	AATGATACGGCGACCACCGAGATCTACACACGTCTACTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
34	AATGATACGGCGACCACCGAGATCTACACAGTCACGTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
35	AATGATACGGCGACCACCGAGATCTACACAGTGTGTGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
36	AATGATACGGCGACCACCGAGATCTACACATCACGTGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
37	AATGATACGGCGACCACCGAGATCTACACACGATCTGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
38	AATGATACGGCGACCACCGAGATCTACACAGACAGCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
39	AATGATACGGCGACCACCGAGATCTACACATCTACACTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
40	AATGATACGGCGACCACCGAGATCTACACACGTGATCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
41	AATGATACGGCGACCACCGAGATCTACACACTAGTGCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
42	AATGATACGGCGACCACCGAGATCTACACAGTCGCTAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
43	AATGATACGGCGACCACCGAGATCTACACATAGTATAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
44	AATGATACGGCGACCACCGAGATCTACACACATACGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
45	AATGATACGGCGACCACCGAGATCTACACACTACTCACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
46	AATGATACGGCGACCACCGAGATCTACACTATATACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
47	AATGATACGGCGACCACCGAGATCTACACTCGATCGCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
48	AATGATACGGCGACCACCGAGATCTACACTACTGCTAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
49	AATGATACGGCGACCACCGAGATCTACACTACGTGAGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
50	AATGATACGGCGACCACCGAGATCTACACTATGTATACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
51	AATGATACGGCGACCACCGAGATCTACACTCTCTCGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
52	AATGATACGGCGACCACCGAGATCTACACTCGTAGCACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

53	AATGATACGGCGACCACCGAGATCTACACTATACGATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
54	AATGATACGGCGACCACCGAGATCTACACTCGTACTGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
55	AATGATACGGCGACCACCGAGATCTACACTGTATACGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
56	AATGATACGGCGACCACCGAGATCTACACTGACTGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
57	AATGATACGGCGACCACCGAGATCTACACTCTGAGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
58	AATGATACGGCGACCACCGAGATCTACACTAGAGCGTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
59	AATGATACGGCGACCACCGAGATCTACACTACGCTATAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
60	AATGATACGGCGACCACCGAGATCTACACTGATGACGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
61	AATGATACGGCGACCACCGAGATCTACACTATGCGACTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
62	AATGATACGGCGACCACCGAGATCTACACTGTATATATAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
63	AATGATACGGCGACCACCGAGATCTACACTACTAGCATAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
64	AATGATACGGCGACCACCGAGATCTACACTCTCGCGTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
65	AATGATACGGCGACCACCGAGATCTACACTCACTATCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
66	AATGATACGGCGACCACCGAGATCTACACTACGCAGCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
67	AATGATACGGCGACCACCGAGATCTACACTGACGTCAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
68	AATGATACGGCGACCACCGAGATCTACACTACTGACTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
69	AATGATACGGCGACCACCGAGATCTACACTCACACTACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
70	AATGATACGGCGACCACCGAGATCTACACTCGTGTGACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
71	AATGATACGGCGACCACCGAGATCTACACTGCGACGACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
72	AATGATACGGCGACCACCGAGATCTACACGCGTATGTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
73	AATGATACGGCGACCACCGAGATCTACACGCACGACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
74	AATGATACGGCGACCACCGAGATCTACACGTCGTCATGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
75	AATGATACGGCGACCACCGAGATCTACACGAGCTGTCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
76	AATGATACGGCGACCACCGAGATCTACACGATGAGACGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
77	AATGATACGGCGACCACCGAGATCTACACGTAGATGAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
78	AATGATACGGCGACCACCGAGATCTACACGTGCGCGAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
79	AATGATACGGCGACCACCGAGATCTACACGTGACGCTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
80	AATGATACGGCGACCACCGAGATCTACACGCTAGTCGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
81	AATGATACGGCGACCACCGAGATCTACACGATCACTACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
82	AATGATACGGCGACCACCGAGATCTACACGTGTGTCACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
83	AATGATACGGCGACCACCGAGATCTACACGACATACACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
84	AATGATACGGCGACCACCGAGATCTACACGTCTATCTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
85	AATGATACGGCGACCACCGAGATCTACACGCGACTATATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
86	AATGATACGGCGACCACCGAGATCTACACGTCAGTAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
87	AATGATACGGCGACCACCGAGATCTACACGATCTACGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
88	AATGATACGGCGACCACCGAGATCTACACGCGACACGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
89	AATGATACGGCGACCACCGAGATCTACACGTAGAGTATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
90	AATGATACGGCGACCACCGAGATCTACACGATAGTGTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
91	AATGATACGGCGACCACCGAGATCTACACGCGCGTCTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
92	AATGATACGGCGACCACCGAGATCTACACGTCGCATACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
93	AATGATACGGCGACCACCGAGATCTACACGAGATCGACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
94	AATGATACGGCGACCACCGAGATCTACACGCAGTCGTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
95	AATGATACGGCGACCACCGAGATCTACACGAGTGACTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
96	AATGATACGGCGACCACCGAGATCTACACGAGCGTAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Table 2-1 96 unique adaptors ligated to DNA fragment prior to primary PCR

Primer Group	Primer
1	CAAGCAGAAGACGGCATACGAGATACGCTCGACAGCTAGCTTGCCAAACCTACAGGTGG
1	CAAGCAGAAGACGGCATACGAGATCTCGCGTGTCGCTAGCTTGCCAAACCTACAGGTGG
1	CAAGCAGAAGACGGCATACGAGATCGTAGACTAGGCTAGCTTGCCAAACCTACAGGTGG
1	CAAGCAGAAGACGGCATACGAGATCACGCTACGTGCTAGCTTGCCAAACCTACAGGTGG
2	CAAGCAGAAGACGGCATACGAGATACGAGTGCGTGCTAGCTTGCCAAACCTACAGGTGG
2	CAAGCAGAAGACGGCATACGAGATCGTGTCTCTAGCTAGCTTGCCAAACCTACAGGTGG
2	CAAGCAGAAGACGGCATACGAGATACGACTACAGGCTAGCTTGCCAAACCTACAGGTGG
2	CAAGCAGAAGACGGCATACGAGATATAGAGTACTGCTAGCTTGCCAAACCTACAGGTGG
3	CAAGCAGAAGACGGCATACGAGATAGACGCACTCGCTAGCTTGCCAAACCTACAGGTGG
3	CAAGCAGAAGACGGCATACGAGATTCTCTATGCGGCTAGCTTGCCAAACCTACAGGTGG
3	CAAGCAGAAGACGGCATACGAGATACGCGAGTATGCTAGCTTGCCAAACCTACAGGTGG
3	CAAGCAGAAGACGGCATACGAGATAGTCGAGAGAGCTAGCTTGCCAAACCTACAGGTGG
4	CAAGCAGAAGACGGCATACGAGATAGCACTGTAGGCTAGCTTGCCAAACCTACAGGTGG
4	CAAGCAGAAGACGGCATACGAGATTGATACGTCTGCTAGCTTGCCAAACCTACAGGTGG
4	CAAGCAGAAGACGGCATACGAGATACAGTATATAGCTAGCTTGCCAAACCTACAGGTGG
4	CAAGCAGAAGACGGCATACGAGATACACATACGCGCTAGCTTGCCAAACCTACAGGTGG
5	CAAGCAGAAGACGGCATACGAGATATCAGACACGGCTAGCTTGCCAAACCTACAGGTGG
5	CAAGCAGAAGACGGCATACGAGATCGAGAGATACGCTAGCTTGCCAAACCTACAGGTGG
5	CAAGCAGAAGACGGCATACGAGATACATACGCGTGCTAGCTTGCCAAACCTACAGGTGG
5	CAAGCAGAAGACGGCATACGAGATACTAGCAGTAGCTAGCTTGCCAAACCTACAGGTGG
6	CAAGCAGAAGACGGCATACGAGATATATCGCGAGGCTAGCTTGCCAAACCTACAGGTGG
6	CAAGCAGAAGACGGCATACGAGATATACGACGTAGCTAGCTTGCCAAACCTACAGGTGG
6	CAAGCAGAAGACGGCATACGAGATTCTACGTAGCGCTAGCTTGCCAAACCTACAGGTGG
6	CAAGCAGAAGACGGCATACGAGATTCTAGCGACTGCTAGCTTGCCAAACCTACAGGTGG
7	CAAGCAGAAGACGGCATACGAGATCATAGTAGTGGCTAGCTTGCCAAACCTACAGGTGG
7	CAAGCAGAAGACGGCATACGAGATCGTCTAGTACGCTAGCTTGCCAAACCTACAGGTGG
7	CAAGCAGAAGACGGCATACGAGATAGACTATACTGCTAGCTTGCCAAACCTACAGGTGG
7	CAAGCAGAAGACGGCATACGAGATAGTGCTACGAGCTAGCTTGCCAAACCTACAGGTGG
8	CAAGCAGAAGACGGCATACGAGATTACGAGTATGGCTAGCTTGCCAAACCTACAGGTGG
8	CAAGCAGAAGACGGCATACGAGATAGCGTCGTCTGCTAGCTTGCCAAACCTACAGGTGG
8	CAAGCAGAAGACGGCATACGAGATACGCGATCGAGCTAGCTTGCCAAACCTACAGGTGG
8	CAAGCAGAAGACGGCATACGAGATACATGACGACGCTAGCTTGCCAAACCTACAGGTGG

Table 2-2 Groups of 4 primers used for secondary PCR

2.4.2 HiSeq

The final library compilation was sequenced on the Illumina HiSeq. A dual-index, paired-end protocol was used, modified from a traditional Nextera protocol and is described in detail in Chapter 4. Illumina sequencing instruments generate *.bcl files as their primary sequencing output. CASAVA demultiplexes these files and also converts them to FASTQ files, one for each sample.

2.5 Bioinformatics

Paired reads were aligned against the C57BL/6 reference genome GRCm38. The reference position on the genome was identified with Burrows-Wheeler Aligner (BWA 0.6.1). The output of the alignment was a Sequence Alignment Map (SAM) file. The file was converted to BAM format, whose indexing allowed a fast retrieval of alignments overlapping a specific region without going through the whole alignment.

Collection of insertion positions and quality selection were performed: reads with an average PHRED-scaled quality score less than 30 (representing a 1 in 1000 probability of incorrect base call) were excluded. After this step, potential PCR duplicates were removed from the samples data i.e. reads with the same map position for read 1 and read 2 within a sample were assumed to be derived from the same original fragment of sheared DNA and grouped.

Reads were then grouped by viral LTR insertion position. Continuous runs of LTR insertions that were adjacent to each other at consecutive bases within a single library were assumed to result from errors in sequencing/mapping and grouped together. In

such groupings the LTR position with the highest frequency (named 'best base') was added to the sample output table (updating also with the chromosome, the start position, the end position, the orientation, and the multiplication of the number of bases by the contig depth information).

Common insertion sites are identified using a Gaussian kernel convolution (GKC) based method (http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.0020166) (de Ridder, Uren, Kool, Reinders, & Wessels, 2006) and target genes involved are determined for each insertion using Kernel Convolution Rules Based Mapping (KC-RBM) software (http://europepmc.org/abstract/med/21652642) (de Jong et al., 2011). Specifically the KC-RBM package identifies common insertion sites by plotting a Gaussian kernel at the site of each insertion. These kernels are summed together and the resulting density distribution (convolution) is scanned for peaks of density. After the estimation of the insertion density, the insertions are associated to the nearest peak are mapped, and merged into clusters. The mean locus are mapped to putative target genes, using four windows sizes (upstream-sense and antisense, downstream-sense and antisense, with respect to the transcription start site). When a cluster falls within a given window, the inserts in that cluster are associated to the gene. The output of KC-RBM program is directly saved into a local MySQL database, and automatically integrated with local copies of Ensembl, COSMIC and DrugBank.

2.6 Virus Kinetics

2.6.1 qPCR of experimental mice DNA

Cohorts of WT / transgenic, MoMuLV infected / uninfected control mice from all time points and also the survival cohort mice (defined as those mice that were culled after developing disease and shown in the Kaplan Meier curves in chapter 3 i.e. not time course mice) were used to look at relative MoMuLV copy number. DNA was quantified using the Qubit® dsDNA HS Assay Kit (Life Technologies; Q32854). qPCR was performed using primers that had been optimised to not detect endogenous retroviral sequences similar to MoMuLV. 5'-GTATGGGCAACTTCTGGCAAC-3' (forward) and 5'-GAGGGAGGTTAAAGGTTCTTCG-3' (reverse) amplified a 204bp region of MoMuLV in infected mice. *Gapdh* was used as a control gene using primers 5'-TGCACCACCAACTGCTTAG-3' (forwards) and 5'-GGATGCAGGGATGATGTTC-3' (reverse) to amplify a 175bp fragment.

The MESA Blue qPCR MasterMix Plus for SYBR® Assay No Rox kit (Eurogentec; RT-SY2X-03+NRWOUB) was used for amplification. DNA concentration was adjusted to 50ng/µl and 1µl put into a reaction volume of 20µl which also included 10µl of reaction buffer, 0.6µl of each primer (10µM stock concentration) and 7.8µl of H_2O. Cycling conditions were 95°C for 5min, followed by 39 cycles of 95°C for 15sec, 60°C for 60sec followed by a plate read, and then a melt curve from 65°C to 95°C incrementing of 0.5°C every 5sec.

2.6.2 qPCR of experimental mice cDNA

WT and transgenic, infected and uninfected control mice from all time points were used to look at relative MoMuLV expression levels. RNA samples were quantified by nanodrop spectrophotometer. RNA was treated with DNase1, Amplification Grade (Life Technologies; 18068-015) as per the manufacturer's instructions with the exception of using a 0.5µg input of RNA instead of 1µg. Treated RNA was then used to make cDNA using SuperScript II Reverse Transcriptase (Life Technologies; 18064) as per the manufacturer's instructions including the use of random primers and also performing the optional step of treating with Ribonuclease H (Life Technologies; 18021-071). cDNA was diluted 1/5 and then 1µl used in the same reaction and using the same primers, *Gapdh* control and cycling conditions as described in section 2.6.1.

2.7 Gene Validation - Candidate Gene Overexpression in Mice

2.7.1 Generation of cDNA of candidate genes

cDNAs of candidate genes were amplified from the RNA extracted from spleens of WT C57BL/6 mice as described in section 2.6.2. In cases where this was unsuccessful, cDNAs were amplified from ORFs purchased from Thermo Scientific Bio. PCR amplification was performed using the Phusion Hot Start II High-Fidelity DNA Polymerase kit (Thermo Scientific, F549L). Primers added on a Kozak consensus sequence and Sfi-1 restriction sites at each end of the gene (see Table 2-4). A 20µl reaction contained 1µl cDNA (100ng), 4µl of 5x buffer, 0.4µl of 10mM dNTPs, 1µl of each primer (10µM), 0.2µl of Phusion Hot Start II High-Fidelity DNA Polymerase and 12.4µl of distilled H_2O. Cycling conditions were 98°C for 30sec, followed by 15 cycles of 98°C for 10sec, 60°C for 30sec, and 72°C for 45sec, and finally 72°C for 5min. Samples were run on 1% agarose gel with ethidium bromide. Bands at the appropriate size were cut and purified using QIAquick Gel Extraction Kit (Qiagen; 28706).

Gene	Transcript ID	Length of cDNA (bp)	cDNA primers	Sequencing primers	Orientation
Vdac1	ENSMUST00000102758	852	5'-GTGGCCGTCAAGGCCCACCATGGCCGTGCCTCCCACATACG-3' 5'-GTGGCCCATGAGGCCTGTTATGCTTGAAATTCCAGTCCTAGGCCAAG-3'	 5´-GTAAAACGACGGCCAG-3´ 5´-CAGGAAACAGCTATGAC-3´	Forward Reverse Forward Reverse
Cd86	ENSMUST00000089620	930	5'-GTGGCCGTCAAGGCCCACCATGGACCCCAGATGCACCATGGG-3' 5'-GTGGCCCATGAGGCCTGTCACTCTGCATTTGGTTTTGCTGAAGCAATTTG-3'	 5´-GTAAAACGACGGCCAG-3´ 5´-CAGGAAACAGCTATGAC-3´	Forward Reverse Forward Reverse
Ildr1	ENSMUST00000089618	1482	5'-GTGGCCGTCAAGGCCCACCATGGGCTGCGGATTGCTCGCTG-3' 5'-GTGGCCCATGAGGCCTGCTAAATGACCACACTCCGTCCACTATG-3'	 5´-GTAAAACGACGGCCAG-3´ 5´-CAGGAAACAGCTATGAC-3´ 5'-CTGCCTGCGGATCTCAGAG-3' 5'-CTGGTGTGAGGGGTTGCTG-3'	Forward Reverse Forward Reverse Forward Reverse
Ildr1	ENSMUST00000119464	1551	5'-GTGGCCGTCAAGGCCCACCATGGGCTGCGGATTGCTCGCTG-3' 5'-GTGGCCCATGAGGCCTGCTATGGGAGCTCTCTCTCTCCTGG-3'	 5´-GTAAAACGACGGCCAG-3´ 5´-CAGGAAACAGCTATGAC-3´ 5'-CTGCCTGCGGATCTCAGAG-3' 5'-CTGGTGTGAGGGGTTGCTG-3'	Forward Reverse Forward Reverse Forward Reverse

Table 2-3 cDNA and sequencing primers used for candidate genes

2.7.2 Sub-cloning of candidate genes

Purified candidate gene cDNAs now with added Kozak and Sfi-1 sites were ligated into Zero Blunt® PCR Cloning Kit (Life Technologies; K2700-20) as per the manufacturer's instructions. These constructs were transformed into chemically competent cells and selected for with kanamycin in the agar plates and LB at 50ng/µl. Three colonies were picked per transformation and inoculated into LB with kanamycin for 14 hours. Plasmid DNA was isolated from competent cells using the QIAprep Spin Miniprep Kit (Qiagen; 27106). Plasmid DNA was sequenced (Table 2-4). Verified cDNAs were then re-transformed as described above and maxi-preps of plasmid DNA isolated using the endotoxin free HiSpeed Plasmid Maxi Kit (Qiagen; 12362). Candidate genes were then ligated into the murine stem cell virus (MSCV) retroviral expression system in preparation for transduction into mouse splenic B-cells (Figure 2-2). Two MSCV vectors, pMSCV-IRES-GFP (originally from the Vignali lab, University of Pittsburgh, USA (Holst, Vignali, Burton, & Vignali, 2006) and a kind gift from Dr. Istvan Bartok, Imperial College London, UK) and pMSCV-IRES-BFP, a modified version replacing GFP with BFP, were used in order that two genes could be co-transduced and simultaneously visualised. Briefly the pCR®-Blunt vector containing the candidate gene of interest and the MSCV retrovirus (modified to contain a Sfi1 cloning site) were digested with the Sfi1 restriction enzyme to give complementary ends (NEB; R0123L) as per the manufacturer's instructions. The digested MSCV retrovirus was treated with Antarctic Phosphatase (NEB; M0289L) as per the manufacturer's instructions in order to prevent re-ligation of the vector. The linearised DNA was run on a 0.8% agarose gel and then cut and gel purified using the QIAquick®Gel Extraction Kit (Qiagen; 28706). The gene of interest was then ligated into MSCV using T4 DNA Ligase (NEB; M0202L). This

construct was transformed into One Shot®Stbl3™ Chemically Competent E. coli (Life Technologies; C7373-03). The transformation process was repeated as described above in this section and maxi-preps of the DNA were again sequence verified.

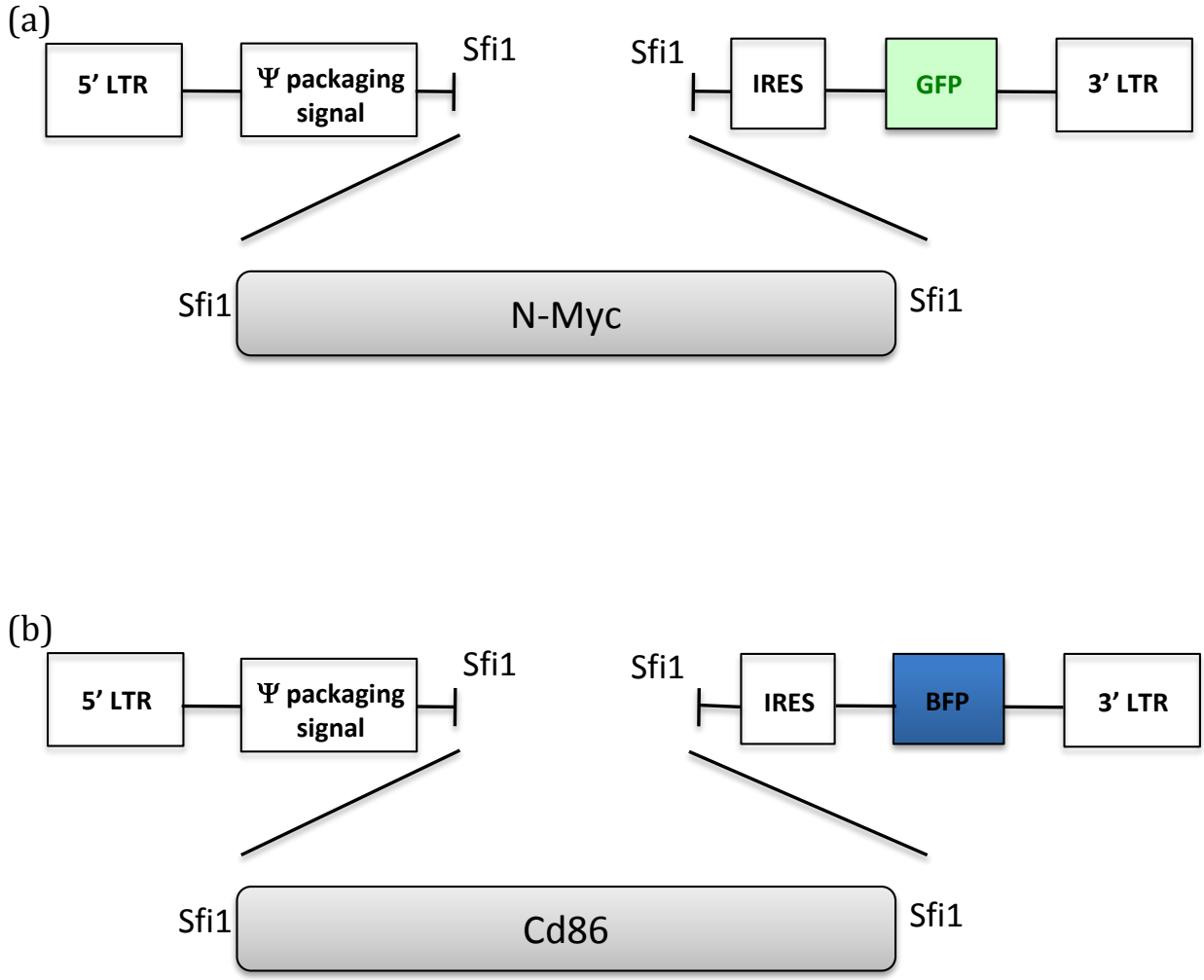


Figure 2-2 MSCV plasmid used to accept candidate genes of interest
MSCV-IRES-GFP (a) and MSCV-IRES-BFP (b) plasmids modified to contain Sfi1 sites to accept a candidate gene of interest from the insertional mutagenesis screen. cDNAs of candidate genes were amplified (adding Sfi1 sites and a Kozak sequence) from either WT C57BL/6 spleen tissue or from purchased ORFs and then sequence verified. cDNAs were then ligated into the above retroviral expression systems. Two plasmids with different reporters were used in order that 2 genes could be simultaneously transfected into packaging cells and then overexpressed in-vivo in mice.

2.7.3 Transduction of candidate genes into mouse B-cells

Host mouse B cells were retrovirally transduced with the MSCV vector and then transplanted by tail vein injection into recipient 8 week old mice matched for strain and sex. This facilitated the overexpression of candidate genes of interest *in-vivo* by the following method (Figure 2-3):

Day 1: 6x 2mls of Phoenix™ Eco packaging cells (originally from the Nolan Lab, Stanford University, USA) were plated out in a 6 well plate at a concentration of $1.2x10^6$/ml in IMDM (with 10% FCS and 2mM L-Glut) and incubated for 48hrs.

Day 3: Phoenix™ Eco packaging cells were transfected with the MSCV construct containing a candidate gene using Lipofectamine® 2000 Transfection Reagent (Life Technologies; 11668-019). Briefly for each of the 6 wells of packaging cells; 3µg MSCV construct and 1µg pCL-Eco (Imgenex; 10045P) helper vector was mixed into 250µl Opti-MEM media (Life Technologies; 31985062). 10µl lipofectamine was mixed into 250µl Opti-MEM media. Both were incubated at room temperature for 5 minutes before mixing the two together and incubating for 20 minutes at room temperature. This transfection mix was then added to each of the wells of the growing cells and rocked gently at 37°C overnight.

Day 4: The transfection mix was removed and the wells washed carefully with 1ml of IMDM (with 10% FCS). 2mls of IMDM (with 10% FCS) was added to each well of cells and left for 24hrs to produce virus. Also on day 4, the spleen of a Eµ-*BCL2* $p53^{+/-}$ C57BL/6 mouse was harvested and a single cell suspension in PBS (with 2% FCS) prepared using the gentleMACS Dissociator (Miltenyi Biotec). Splenic B-cells were

isolated using the EasySep™ Mouse B-cell Isolation Kit (Life Technologies; 19854) as per the manufacturer's instructions. 5x wells of $2x10^6$ B-cells in 1ml RPMI (with 10%FCS, 2mM L-Glut, 50μM β2-ME and pen/strep 50U/50μg per ml) were plated out in a 48-well tissue culture plate and stimulated with 20μg/ml of LPS for 24hrs, incubated at 37°C.

Day 5: Virus supernatant was collected from the Phoenix™ Eco packaging cells and polybrene added to a final concentration of 8μg/ml. In the case of two genes being transduced, the virus supernatants were mixed together at this point. At this stage, the optional extra step of harvesting the packaging cells and checking for GFP / BFP by flow cytometry to verify transfection efficiency was performed. All 5 wells of LPS-stimulated mouse B-cells were harvested in a 15ml falcon, and centrifuged (5mins, 1200rpm, room temperature). The pellet was resuspended in 10mls of virus supernatant / polybrene mix and 1ml/well plated into 10 wells of a 48 well plate. Spin infection was performed by spinning the plate for 90 minutes / 800g / room temperature. Supernatant was removed and fresh media added (RPMI with 10%FCS, 2mM L-Glut, 50μM β2-ME and pen/strep 50U/50μg per ml) with 20 μg /ml of LPS for 48hrs.

Day 7: Cells were fluorescence-activated cell sorted (FACS) into GFP / BFP positive/negative populations. Populations were centrifuged and resuspended in FACS buffer (PBS, 2% FCS, 1mM EDTA).

2.7.4 Introducing candidate genes into mice and generation of tumours

Each population of FACS sorted, transduced *Eμ-BCL2* $p53^{+/-}$mouse B-cells were injected intravenously via the tail vein into cohorts of five C57BL/6 mice. Each mouse received $1x10^5$ viable cells. In some cases where two genes were introduced, the cells were

sorted into four populations based on GFP/BFP expression levels: a. Gene 1 negative / Gene 2 negative cells, b. Gene 1 positive / Gene 2 negative cells, c. Gene 1 negative / Gene 2 positive cells and d. Gene 1 positive / Gene 2 positive cells. Three days prior to IV injection mice were transferred to sterile cages, with sterile food and drinking water containing Baytril. One day prior to injection, mice were exposed to 400cGy of irradiation. Immediately prior to injection, mice were pre-warmed in an incubator for 5 minutes to allow veno-dilatation. Post-transplantation mice were monitored three times per week for signs of illness. Any mouse exhibiting splenomegaly alone, tachypnoea alone, lymphadenopathy alone or 10% weight loss/gain with 2 features of hunched posture / piloerection / withdrawn behavior were sacrificed and organs (spleen, thymus, lymph nodes, bone marrow) were harvested and snap frozen in liquid nitrogen as soon as possible after death. Single cell suspensions of spleen were prepared in all cases using the gentleMACS Dissociator (Miltenyi Biotec).

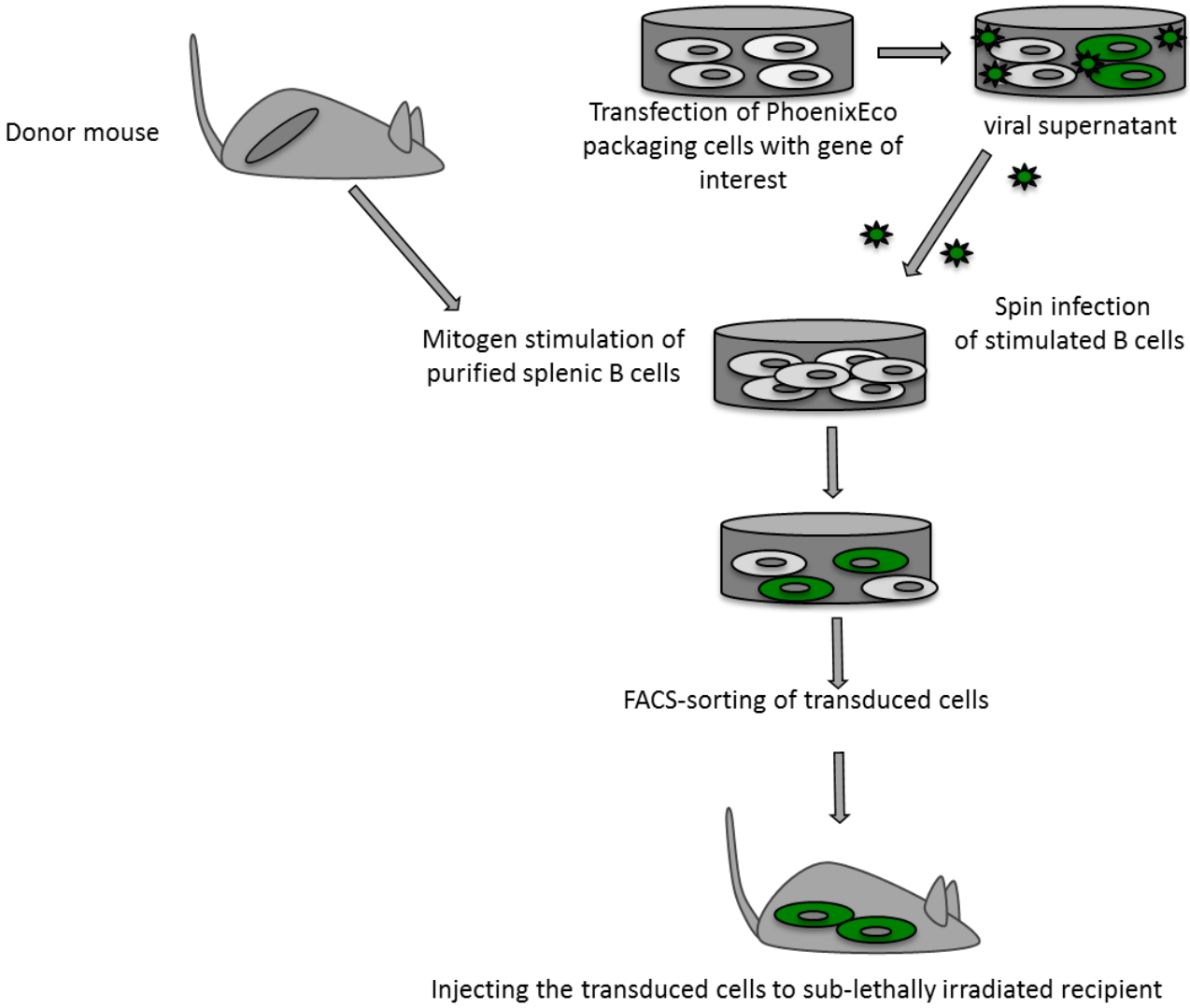


Figure 2-3 Candidate gene overexpression in mice
PhoenixEco packaging cells were plated out and cultured for 48hrs. They were then transfected with an MSCV plasmid containing a candidate gene and a reporter gene using Lipofectamine and incubated for 24hrs to produce virus. The spleen was then removed from a donor mouse, the B-cells isolated and then cultured with LPS stimulation for 24hrs. B-cells were then centrifuged and resuspended in virus supernatant and polybrene. Cells were spun for 90 minutes and then cultured for 24hrs with LPS stimulation. Cells were harvested, sorted by FACS and 100,000 cells injected by tail vein into host mice.

CHAPTER 3 RESULTS & DISCUSSION: ANIMALS

3.1 Tumour Generation and survival of mice

Mice developed different patterns of disease. A large number of the mice developed splenomegaly which was clinically detectable by abdominal palpation. Figure 3-1(a) shows a normal adult mouse spleen in a (BALB/c x C57BL/6)F1 mouse. Figure 3-1(b) shows an example of the splenomegaly found in diseased mice. The spleens of adult uninfected wild-type mice were approximately 0.1-0.2g. Those of uninfected transgenic mice were slightly larger. Infected mice had spleens that varied in mass up to 2.52g (Figure 3-10). Some mice developed an enlarged thymus and were tachypnoeic and displayed other signs of sickness including weight loss and being withdrawn. Figure 3-1(c) shows an example of an enlarged thymus that expanded to fill the chest cavity. Other mice developed lymphadenopathy which was occasionally detectable by palpation but often was only apparent at necropsy when they had developed other signs of disease. All the lymph node groups were examined including inguinal, mesenteric, axillary and cervical. Figure 3-1(d) shows an example of enlarged mesenteric lymph nodes attached to the bowel of a mouse. Mice developed a combination of some or all of the above signs. Those mice with an enlarged thymus (n=158) have a smaller mean spleen weight at death than those with a normal thymus (n=142) (Mann Whitney U $p<0.0001$) (See Figure 3-11). This may be explained by the more enlarged organ representing the primary site of disease, whilst the less enlarged organ represents metastasis. This data will be correlated with tumour characterisation by flow cytometry to distinguish tumour lineage.

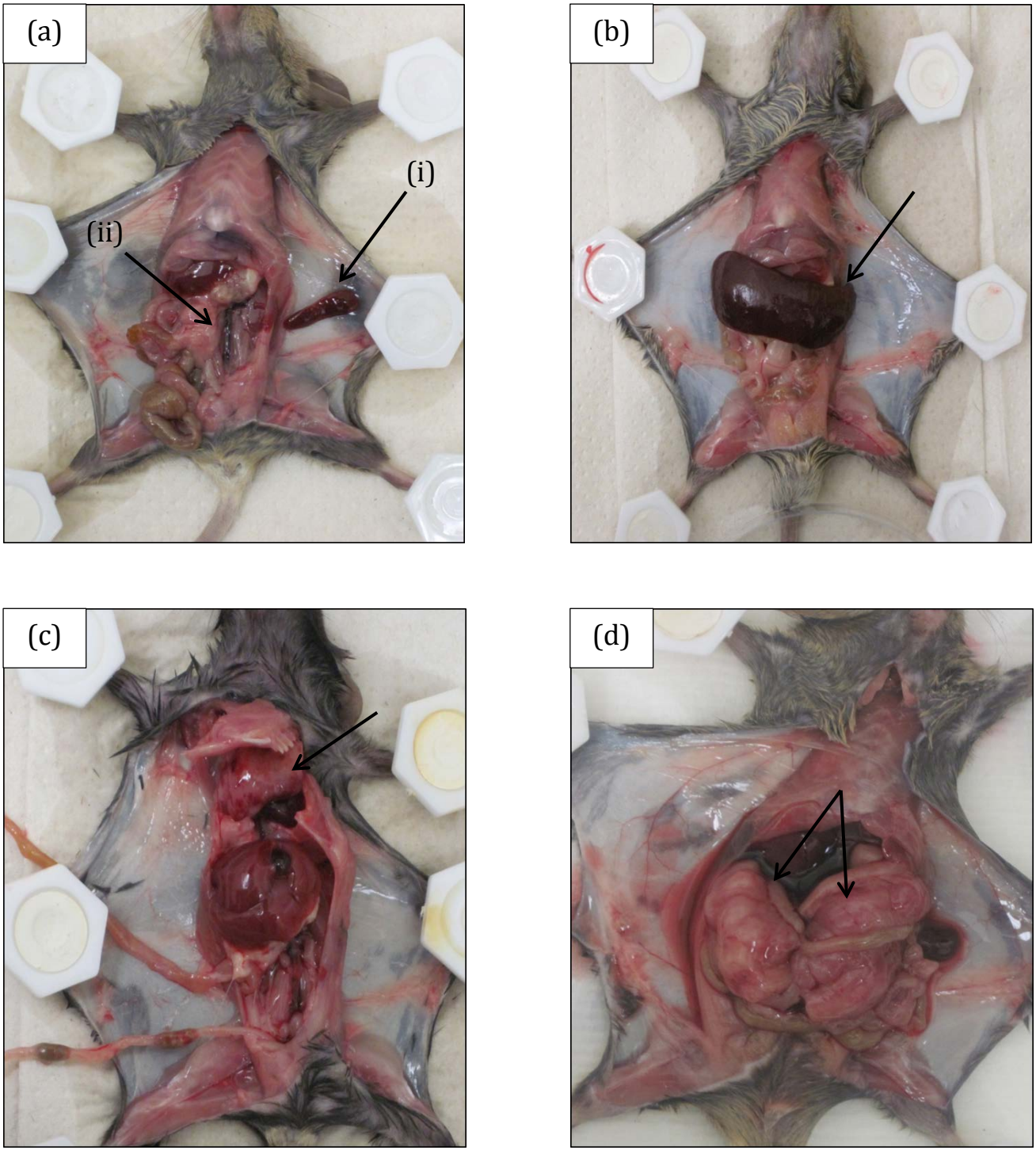


Figure 3-1 Organ sizes in healthy and diseased mice
All examples in (BALB/c x C57BL/6)F1 mice. Diseased mice developed varying degrees of all combinations of organomegaly shown above. (a) The normal spleen (i) and abdominal viscera (ii) of an adult, uninfected control mouse. (b) Marked splenomegaly in an MoMuLV infected mouse with lymphoma. This was easily palpable on abdominal examination. (c) The enlarged thymus of an MoMuLV infected mouse with lymphoma – this expanded to fill the chest cavity and caused tachypnoea. (d) Mesenteric lymphadenopathy in an MoMuLV infected mouse with lymphoma.

Two different transgenic mice were bred on two different background strains, which were infected with MoMuLV prepared in two different batches. This gave rise to three cohorts of mice. In both cases the transgene is heterozygous and so approximately 50% of offspring were wild-type within each cohort. Each cohort also included one or two litters of uninfected control mice that were injected with vehicle.

Cohort 1 included *Eμ-BCL2* transgenic C57BL/6 mice (n=46) and wild-type C57BL/6 mice (n=55) infected with virus prep 1. Uninfected control mice included Eμ-BCL2 transgenic C57BL/6 mice (n=5) and wild-type C57BL/6 mice (n=8). One infected wild-type mouse did not develop lymphoma by one year and so was censored from the data. All other infected mice developed lymphoma within one year and were included in the Kaplan Meier survival curves (see Figure 3-2). No uninfected control mice (transgenic or wild-type) developed lymphoma, however one uninfected transgenic mouse died at 50 days post-injection with vehicle of an unknown cause but with no evidence of lymphoma and so was censored from the data. In testing survival differences between these infected wild-type and transgenic mice, the Gehan-Breslow-Wilcoxon test showed transgenic mice developed disease significantly earlier than wild-types ($p=0.0163$). The log-rank Mantel Cox test did not reach significance ($p=0.1978$). This difference is likely to be due to the fact that the Gehan-Breslow-Wilcoxon test is more sensitive to differences in survival at earlier time points (Motulsky, 2013).

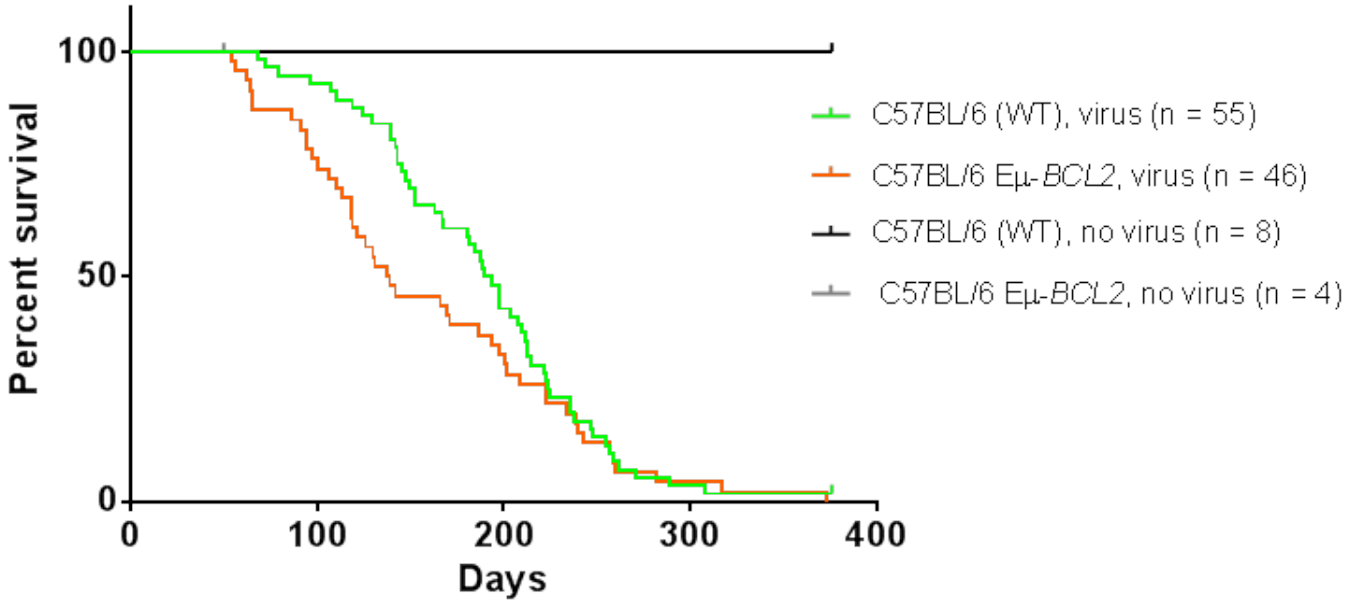


Figure 3-2 Kaplan Meier survival curves of cohort 1
Kaplan Meier survival curves of virus infected and uninfected control, (WT) and Eμ-*BCL2* transgenic C57BL/6 mice. Infected transgenic mice had a significantly shorter survival time than infected WT mice (Log-rank (Mantel-Cox) test p=0.1978, Gehan-Breslow-Wilcoxon test p=0.0163).

Cohort 2 included Eμ-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice (n=59) and wild-type (BALB/c x C57BL/6)F1 mice (n=67) also infected with virus prep 1. Uninfected control mice included Eμ-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice (n=6) and wild-type (BALB/c x C57BL/6)F1 mice (n=3). Two infected mice (one transgenic, one wild-type) had to be culled at 245 days post-infection for welfare reasons rather than lymphoma and so were censored. All other infected mice developed lymphoma within one year. Two uninfected mice, one transgenic and one wild-type, developed lymphoma at 318 and 331 days respectively. This finding is not unexpected since BALB/c mice have a higher incidence of lymphoma than C57BL/6, likely due to their increased numbers of circulating lymphocytes. Interestingly there was no difference in survival between infected wild-type and transgenic mice in this cohort, indicating that the addition of the Eμ-*BCL2* transgene to a background that includes BALB/c has no effect on the onset of lymphoma (see Figure 3-3).

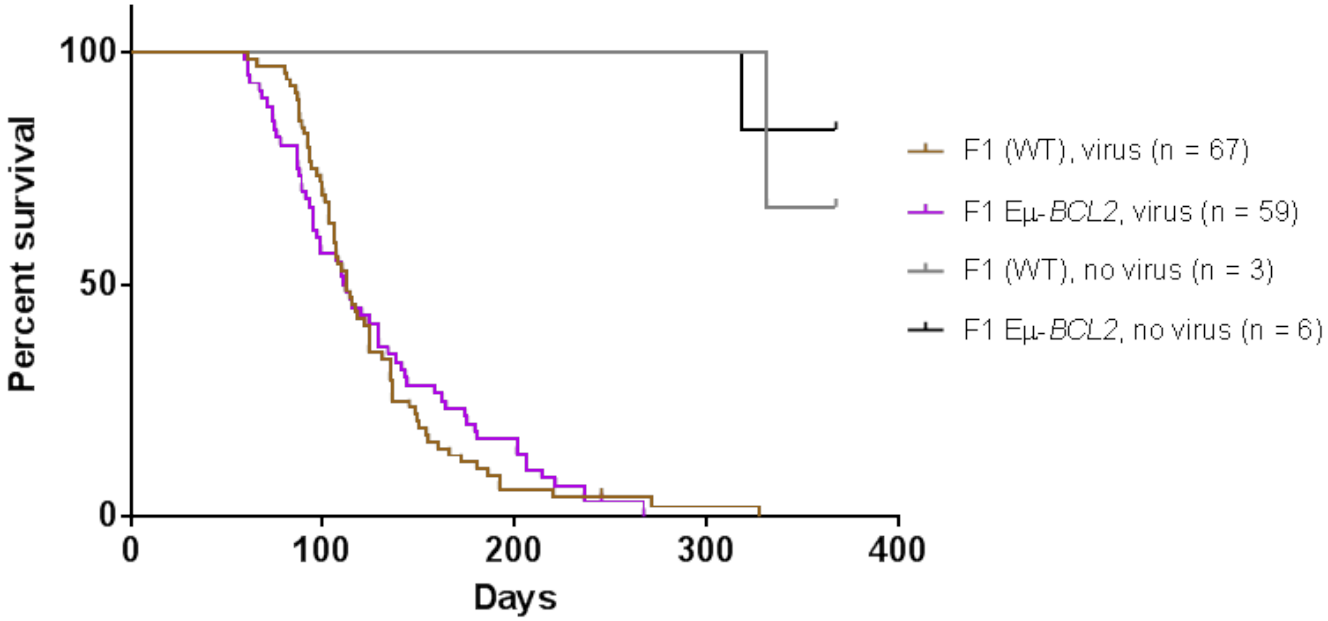


Figure 3-3 Kaplan Meier survival curves of cohort 2
Kaplan Meier survival curves of virus infected and uninfected control, wild-type (WT) and Eμ-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice. There was no significant difference in survival between infected WT and transgenic mice suggesting that the addition of the Eμ-*BCL2* transgene to mice with BALB/c in its background confers no extra promotion of oncogenesis (Log-rank (Mantel-Cox) test p=0.7848, Gehan-Breslow-Wilcoxon test p=0.6714).

Cohort 3 included VavP-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice (n=59) and wild-type (BALB/c x C57BL/6)F1 mice (n=66) infected with virus prep 2. Uninfected control mice included VavP-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice (n=6) and wild-type (BALB/c x C57BL/6)F1 mice (n=7). One wild-type mouse had to be culled after sustaining injuries from fighting. All infected mice developed lymphoma within one year and transgenic mice had a significantly reduced survival time compared to wild-type (Log-rank (Mantel-Cox) test $p<0.0001$, Gehan-Breslow-Wilcoxon test $p<0.0001$, see Figure 3-4). No uninfected wild-type control mice developed disease but as expected, all of the uninfected transgenic mice eventually developed disease but significantly slower than those infected transgenic mice. Figure 3-9 summarises the mean time to death of all three cohorts.

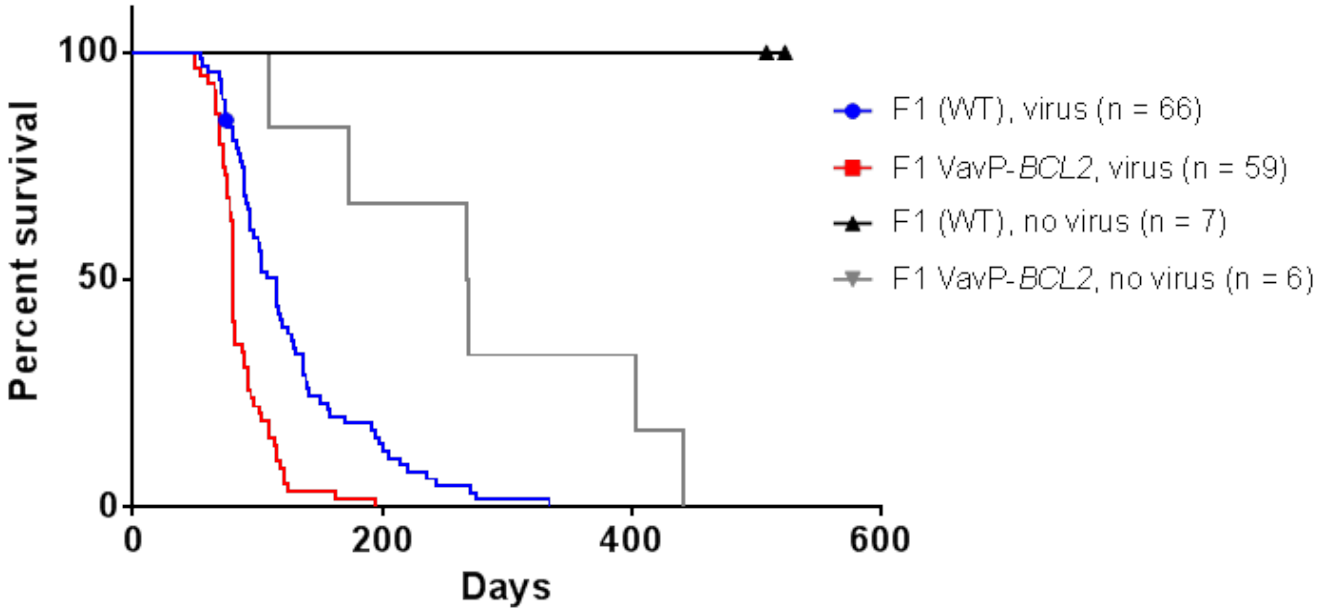


Figure 3-4 Kaplan Meier survival curves of cohort 3
Kaplan Meier survival curves of virus infected and uninfected control, wild-type (WT) and VavP-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice. Infected transgenic mice had a significantly shorter survival time than infected WT mice (Log-rank (Mantel-Cox) test $p<0.0001$, Gehan-Breslow-Wilcoxon test $p<0.0001$). As expected in this model, uninfected transgenic mice did eventually develop disease spontaneously, but significantly slower than infected WT mice (Log-rank (Mantel-Cox) test $p<0.0018$, Gehan-Breslow-Wilcoxon test $p<0.0096$).

Comparing Kaplan-Meier curves of the wild-type C57BL/6 mice from cohort 1 and the wild-type F1 mice from cohort 2 demonstrates that the background strain of the mouse does have an effect on survival, with virus infected F1 mice having a significantly shorter survival time than C57BL/6 (Log-rank (Mantel-Cox) test $p<0.0001$, Gehan-Breslow-Wilcoxon test $p<0.0001$, see Figure 3-5). The addition of BALB/c to C57BL/6 in the F1 experimental mice speeds up disease onset after viral infection likely due to increased numbers of B lymphocytes, Ig-secreting cells and serum Ig, as well as a prolonged antibody response to immunization compared to pure C57BL/6 mice. Studying survival between C57BL/6 and F1 mice, when both carry the Eμ-*BCL2* transgene shows no significant difference, suggesting that in the case of the F1 mice, the addition of Eμ-*BCL2* has no greater effect than the background strain itself in speeding up disease (Figure 3-6). This maybe because this transgene confers no greater B-cell survival / heightened immune response than the addition of BALB/c.

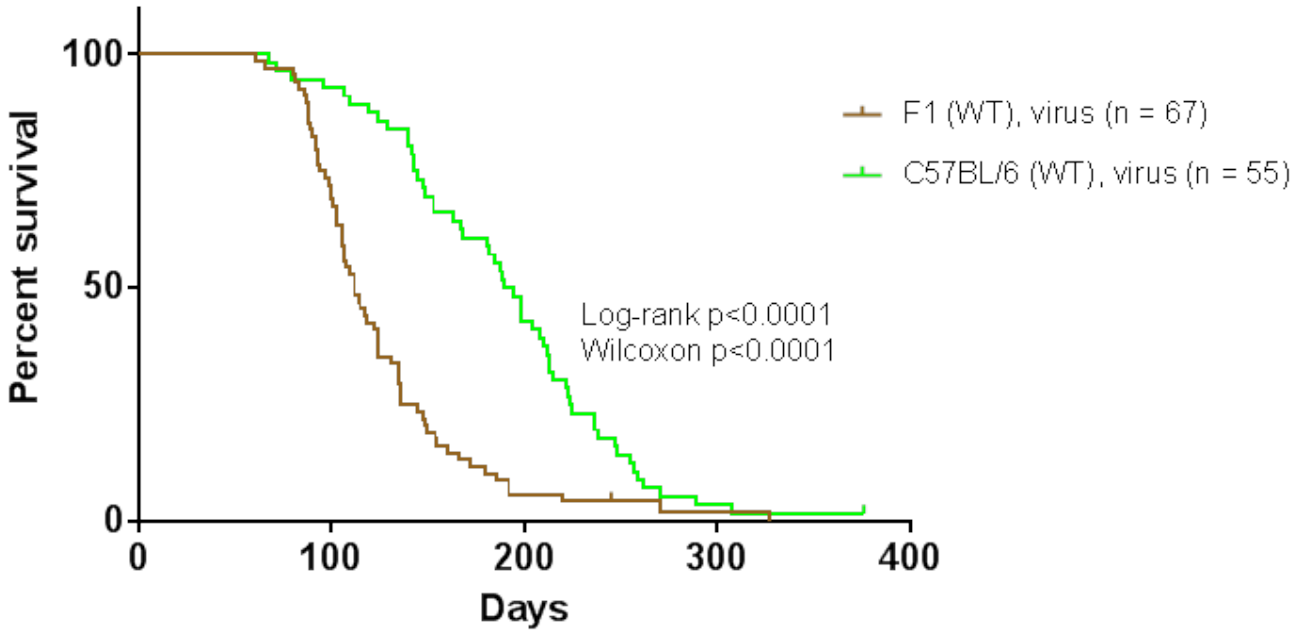


Figure 3-5 Kaplan Meier survival curves comparing strains
Survival of infected WT C57BL/6 (from cohort 1) *vs* infected WT (BALB/c x C57BL/6)F1 mice (from cohort 2). F1 mice had a significantly shorter survival time than C57BL/6 mice (Log-rank (Mantel-Cox) test p<0.0001, Gehan-Breslow-Wilcoxon test p<0.0001).

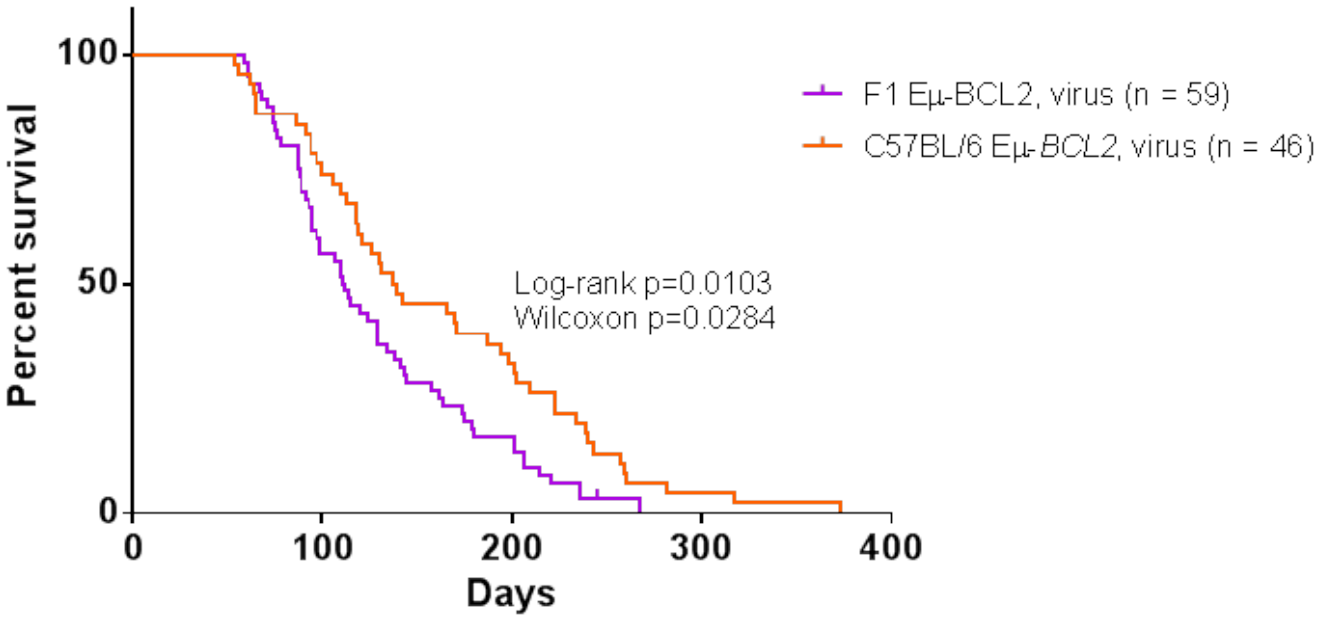


Figure 3-6 Kaplan Meier curves comparing the effect of Eμ-BCL2 on different strains
Eμ-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice have a significantly shorter survival time than Eμ-*BCL2* transgenic C57BL/6 mice (Log-rank (Mantel-Cox) test p=0.0103, Gehan-Breslow-Wilcoxon test p=0.0284).

After the first preparation of MoMuLV was used up on cohorts 1 and 2, a second batch had to be prepared. This raised concerns regarding whether the two batches of virus would have different titres and therefore cause problems comparing the mice in cohort 3 to those in the other cohorts. However, comparing the infected wild-type mice from cohorts 2 and 3 showed no difference in survival (Figure 3-7). The activity of the two preparations, virus expression levels and also relative DNA copy numbers are looked at in more detail in chapter 5. Having established that the two virus preparations are similar in effect allows comparisons between the different mouse cohorts, I wanted to examine whether the different transgenes caused any difference in survival. Figure 3-8 shows that infected VavP-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice had a significantly shorter survival time than their Eμ-*BCL2* counterparts, suggesting that the VavP-*BCL2* transgene is the more oncogenic.

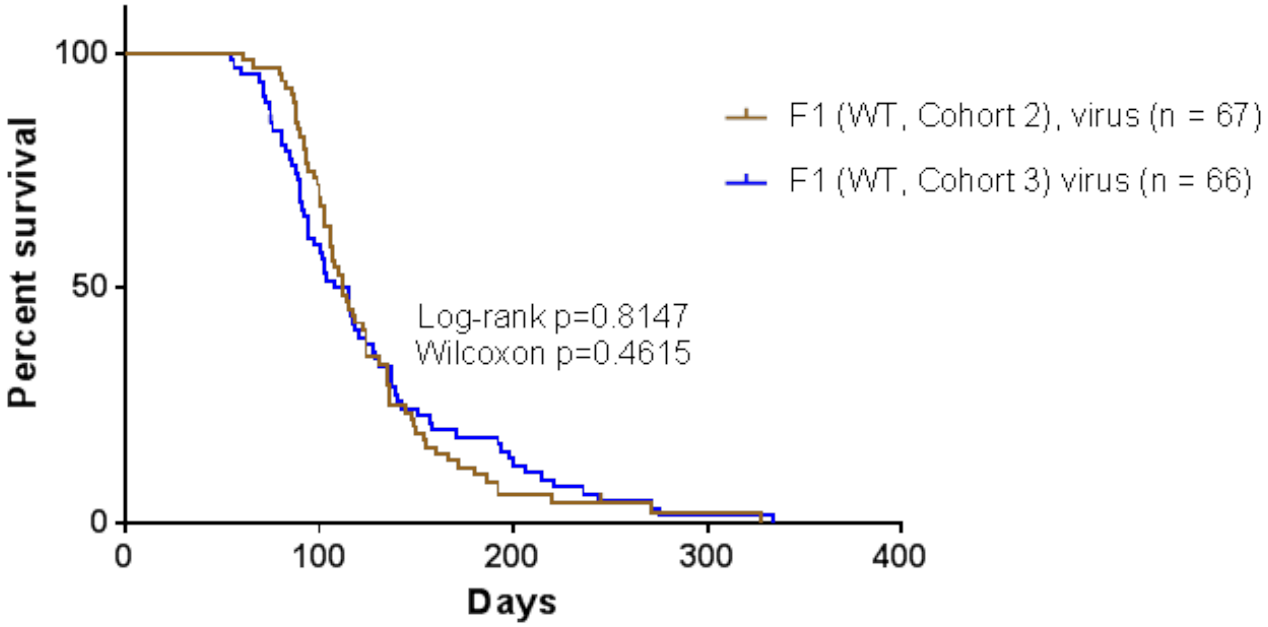


Figure 3-7 Kaplan Meier survival curves to study effect of two different virus preparations

Cohorts 2 and 3 received different batches of virus. Comparing the WT mice from these cohorts, that have the same background, shows that there was no difference in survival between despite the second preparation being more concentrated than the first (see).

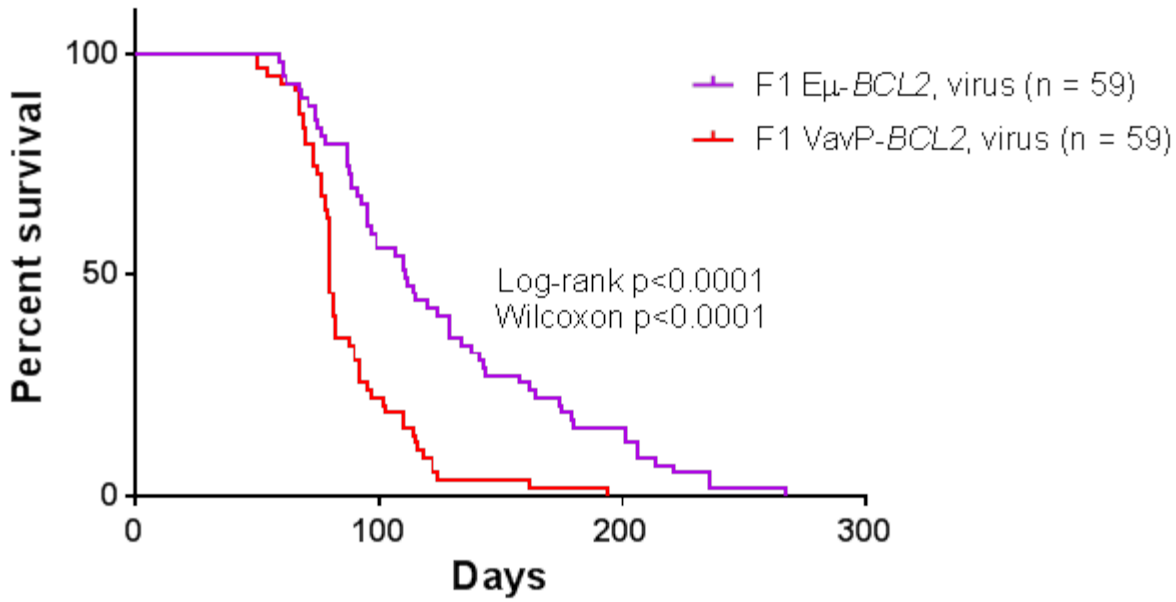


Figure 3-8 Kaplan Meier curves comparing Eμ-BCL2 and VavP-BCL2 transgenes
VavP-*BCL2* F1 mice had a significantly shorter survival time than Eμ-*BCL2* F1 mice (Log-rank (Mantel-Cox) test $p<0.0001$, Gehan-Breslow-Wilcoxon test $p<0.0001$), suggesting that the former is more oncogenic than the latter.

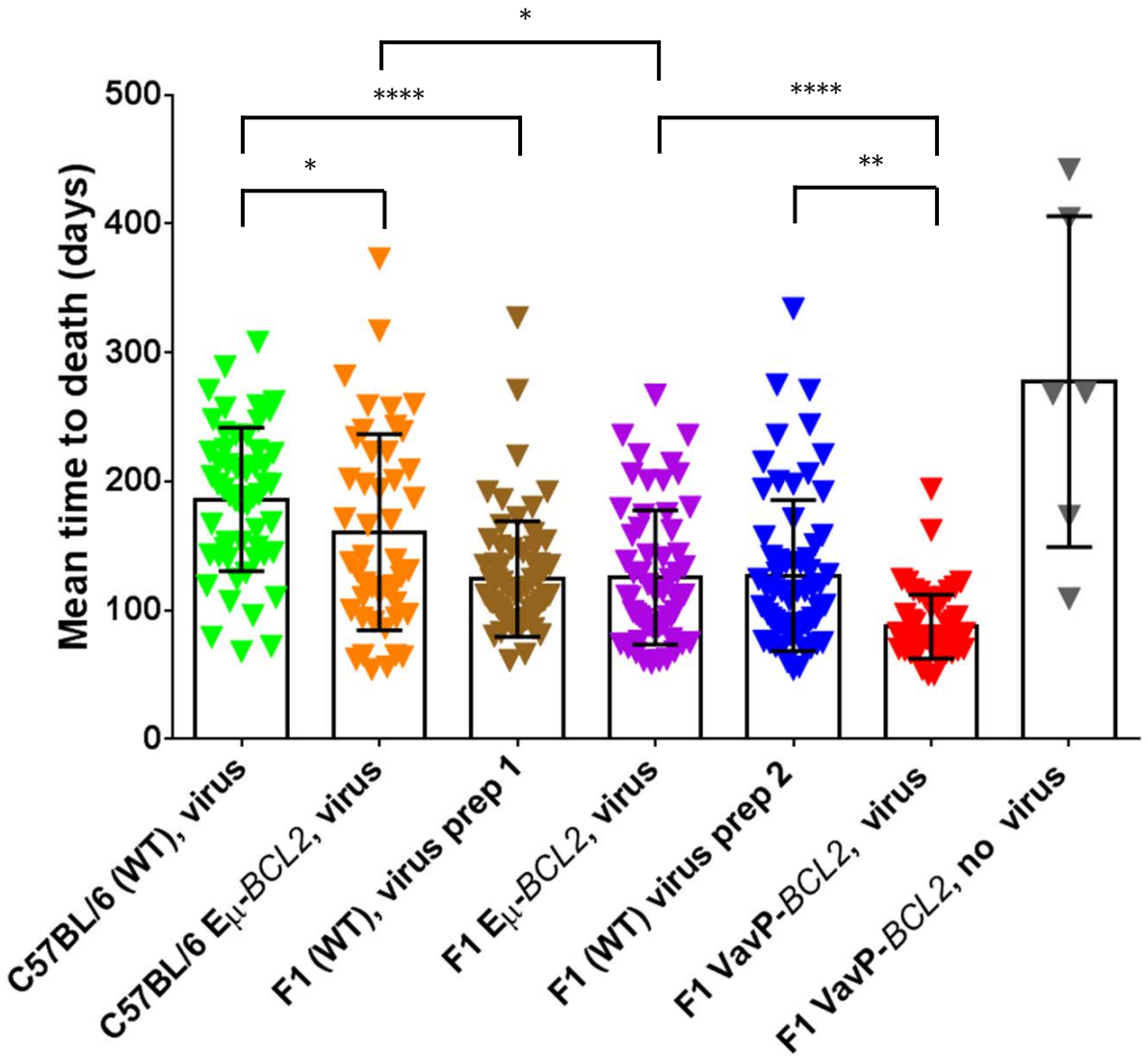


Figure 3-9 Mean survival time of mice

The mean survival time of the mice within each cohort is shown above. Bars represent the mean with standard deviation of the mean error bars. VavP-*BCL2* F1 uninfected control mice (n=6) are included as they developed lymphoma spontaneously (as expected) although slower than all infected groups of mice. No other uninfected controls developed disease and so are not displayed. Infected VavP-*BCL2* mice (n=59) had the shortest mean time to death. Surprisingly, WT F1 mice (virus prep 1, n=67) had a significantly shorter survival time than infected Eμ-*BCL2* C57BL/6 mice (n=46) ($p<0.05$) and a similar survival time to Eμ-*BCL2* F1 mice (n=59), suggesting that a background including BALB/c is more oncogenic than the Eμ-*BCL2* transgene on a C57BL/6 background after virus infection. * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$.

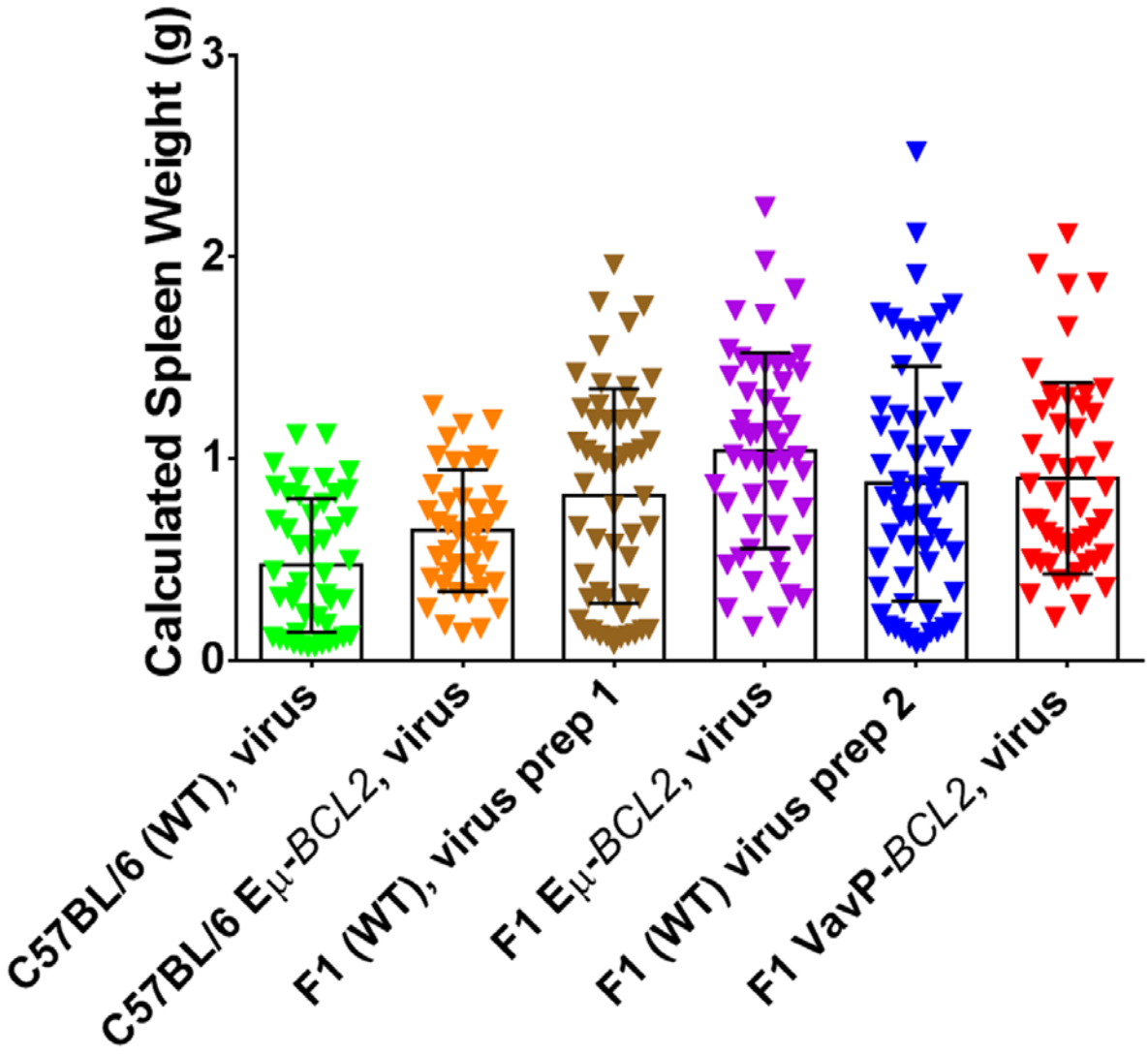


Figure 3-10 Calculated spleen weights of different mouse cohorts
Data bars represent mean spleen weight and error bars represent the standard deviation of the mean.

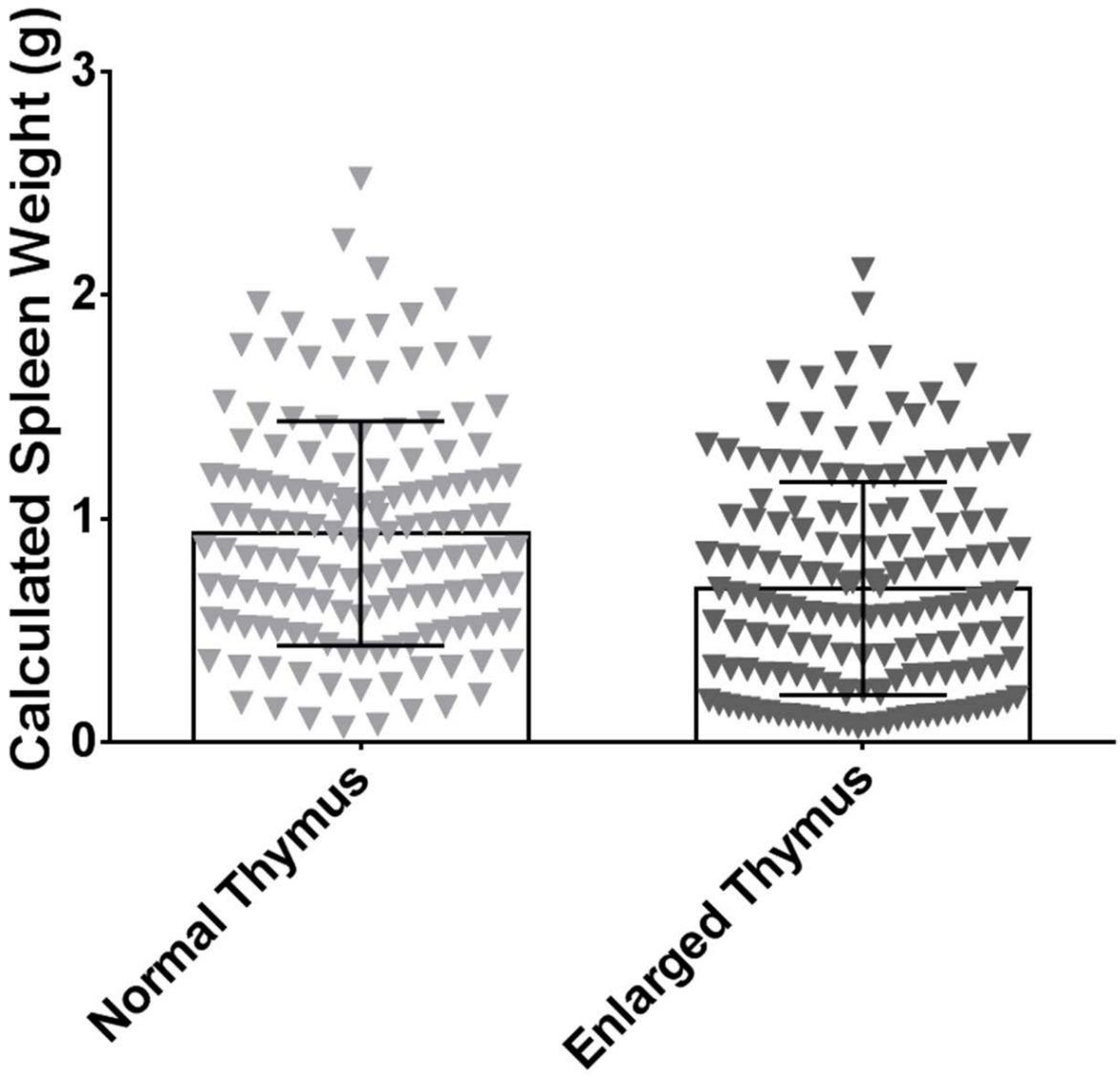


Figure 3-11 Spleen weights mice with and without an enlarged thymus
Those mice with an enlarged thymus (n=158) have a smaller mean spleen weight at death than those with a normal thymus (n=142) (Mann Whitney U $p<0.0001$). Data bars represent mean spleen weight and error bars represent the standard deviation of the mean.

CHAPTER 4 RESULTS & DISCUSSION: IDENTIFICATION, ENRICHMENT AND SEQUENCING OF MOMULV INSERTION SITES

My aim was to produce a robust, cost-effective, reproducible, high-throughput method for virus insertion site identification, enrichment and sequencing on a modern platform that would facilitate quantitative analysis of both highly clonal and subclonal populations. Extensive alterations and optimisation to previously described methods and to the current Illumina DNA library preparation protocols were required and I describe these differences in the following sections.

4.1 Read *vs* Fragment *vs* Insertion Site

After infecting mice with replicating MoMuLV, virus insertions are identified, enriched and then sequenced. Regardless of the method with which this is done, sequencing produces reads that are then mapped to the mouse genome to identify genes that may be implicated in the onset of disease. Reads that map to the same base on the mouse genome at the LTR end of the PCR product are grouped together as one virus insertion. Within each insertion, paired end reads that are of the exact same length are assumed to be PCR amplified copies of the same sheared DNA fragment (see Figure 4-1).

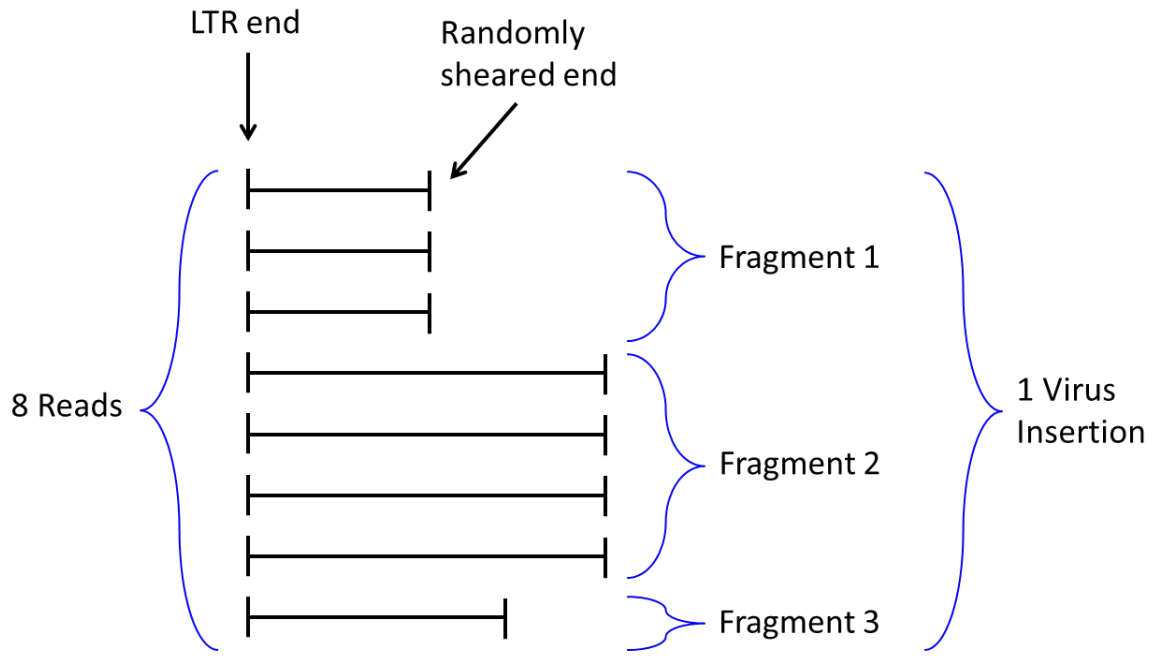


Figure 4-1 Read, Fragment and Insertion definitions
After sequencing the genome of an MoMuLV infected mouse, reads that map to the same base on the mouse genome at the LTR end of the PCR product are grouped together as one virus insertion. Within each insertion, paired end reads that are of the exact same length are assumed to be PCR amplified copies of the same sheared DNA fragment. The above example shows how 8 reads are grouped into 3 fragments that all indicate 1 insertion site.

4.2 The history of insertion site identification

The identification, enrichment and sequencing of virus insertion sites in retroviral insertional mutagenesis screening has progressed hugely in recent years. In the first screens, insertion sites were isolated and characterised by southern blot analysis and genomic library screening which is low throughput and very labour intensive (Hayward, Neel, & Astrin, 1981; Tsichlis, Strauss, & Hu, 1983). Viral sequences were used as probes on Lambda libraries constructed from tumour DNA. This was superseded by PCR based methods that enrich for the host DNA at the insertion site of the retrovirus by amplifying sequences that flank it. Two methods were used most commonly. The first was an inverse PCR that involved digesting DNA and then ligating it into circles and primers specific to the insert (pointing outwards) amplified the host DNA next to the insertion site. The second method was splinkerette PCR where DNA is digested using restriction enzymes and then ligated to linker DNA containing a non-complementary hairpin. In the first strand synthesis only the virus primer can bind since the linker primer is not complementary to the linker strand. The resulting first strand does have a section complementary to the linker primer which produces the complementary strand in the second round. Subsequent cycles of PCR amplify this product which contains the viral insertion site.

There were several similar alternatives to splinkerette PCR including vectorette PCR (Riley et al., 1990), T-linker PCR (Yuanxin et al., 2003), LAM-PCR (Schmidt et al., 2002) and boomerang DNA amplification (Hui, Wang, & Lo, 1998) all of which used similar linker strategies to avoid amplification of portions of host DNA that did not contain a virus insertion. These alternatives had various problems; most do not lend themselves easily to high-throughput methods and vectorette PCR is prone to 'end-repair priming'

of sequences unligated to linkers. Splinkerette PCR overcame this problem, as during the elongation step of the PCR the free 3' end of the splinkerette 'flips back' on itself to create a stable double-stranded DNA hairpin, and so became the predominant method used for this purpose. Both inverse PCR and splinkerette PCR are still used today (Baron et al., 2012; Fernald et al., 2013; Kool et al., 2010; Anthony G Uren et al., 2009).

To understand the importance of an individual insertion site in regard to its impact on lymphomagenesis, it is critical to distinguish clonal insertions that are present in many cells of a tumour / tumours from mutations that did not give rise to clonal expansion and are only present in a single cell, or few cells, of a tumour. In this thesis we use the terms 'clonality' and 'clonal abundance' to refer to relative abundance of an individual insertion within a given sample. It is desirable to quantify and rank the clonality of those insertions that drove tumour outgrowth, in order to give indications of which clonal insertions may represent the most important oncogenes.

Insertion site identification methods have long been hampered by the use of restriction enzyme digestion to fragment the DNA. The wide range in length of fragments introduces PCR bias as shorter fragments are amplified in preference to longer ones. Also, some fragments will be longer than sequencing platforms can amplify. Another difficulty is that as restriction sites are fixed (and thus the same) between samples, if an individual insertion is read many times during sequencing, this may represent multiple independent insertions at a given locus, or it could simply represent one insertion that for some reason was amplified by PCR more than others. These issues mean that the results are not quantitative and clonal abundance cannot be accurately assessed.

An alternative to restriction digest of DNA is shearing by sonication, which gives a controlled and reproducible size distribution of the resulting fragments, which eliminates the abovementioned caveats of restriction digestion. In addition, as the indexed adaptor is ligated to fragments of DNA that are sheared at random, the randomness of the breakpoints of shearing means that detection of fragments of different length at the same LTR/mouse genome junction will represent unique, independent DNA molecules within the original sample, rather than biased amplification of just one event. This sonication / fragment length approach to quantification of clonality has been previously used in a splinkerette PCR protocol (Koudijs et al., 2011).

4.3 The history of insertion site sequencing

The cells within these oligoclonal tumours generated by retroviruses (and transposons) usually contain multiple insertions, meaning that the amplification of these events has to be de-multiplexed in order to identify individual sites by sequencing. This can be performed by 'shotgun' cloning where amplified DNA containing virus insertions is transformed into competent cells, colonies picked and then DNA capillary sequenced. Whilst this method is robust, it is expensive and requires huge amounts of labour and the limited sampling (hundreds of colonies per PCR reaction) means that only the most abundant/clonal virus insertion sites are reliably identified.

More recently next generation sequencing platforms have allowed large numbers of samples to be processed in parallel with extensive coverage of each sample. By adding indices to each PCR reaction many samples can be pooled into a single run. This approach has been applied to the 454 pyrosequencing platform (C. A. Huser et al., 2014;

Klijn et al., 2013; Koudijs et al., 2011). A limitation of 454 is the relatively low number of reads containing the ligation point as 30% lack the splinkerette sequence at the 3' end. The Illumina HiSeq has not been used for insertion site sequencing to our knowledge.

4.4 Optimisation of my method

The DNA library preparation of infected mice involves fragmenting DNA, blunting the ends of the fragments, adding an A-tail to blunt the ends, adding an adaptor and then performing a restriction enzyme digest of a site that is both unique and located within the MoMuLV genome adjacent to the LTR. Each sample is then enriched by PCR for fragments containing the adaptor at one end and virus LTR at the other and then these fragments were sequenced. I made a number of adaptations and modifications to both the standard Illumina paired end sequencing protocol (http://support.illumina.com/content/dam/illumina-support/documents/myillumina/e5af4eb5-6742-40c8-bcb1-d8b350bcb964/paired-end_sampleprep_guide_1005063_e.pdf) and of other recently published insertional mutagenesis screens (Baron et al., 2012; C. A. Huser et al., 2014; Koudijs et al., 2011).

4.4.1 DNA fragmentation

I used sonication to fragment mouse DNA rather than restriction enzyme digestion for the reasons outlined in section 4.1.1.

4.4.2 DNA clean-up and size selection

Previous methods use column purification to clean-up DNA in between library preparation steps, and gel purification to size select prior to PCR enrichment. I used

solid phase reversible immobilisation (SPRI) beads for both clean-up and size selection. This allowed the processing of large numbers of samples simultaneously (using a 96 well magnet) and also allowed automation by robot (Biomek® NXP, Beckman Coulter, A31839). In order to increase the stringency of fragment size being sequenced, an additional size selection step was added, and so was performed both before nested PCR and also prior to sequencing.

4.4.3 DNA fragment enrichment and sequencing

The primary PCR first strand synthesis primer is a sequence found in the external U3 LTR of MoMULV that is also repeated in the internal U3 LTR (Figure 4-2). In order to prevent the unnecessary amplification of internal virus fragments an appropriate restriction site must be found to digest adaptor ligated DNA. I include an EcoRV digest step after adaptor ligation and prior to primary PCR that is not included in the Illumina protocol to facilitate this (Figure 4-3).

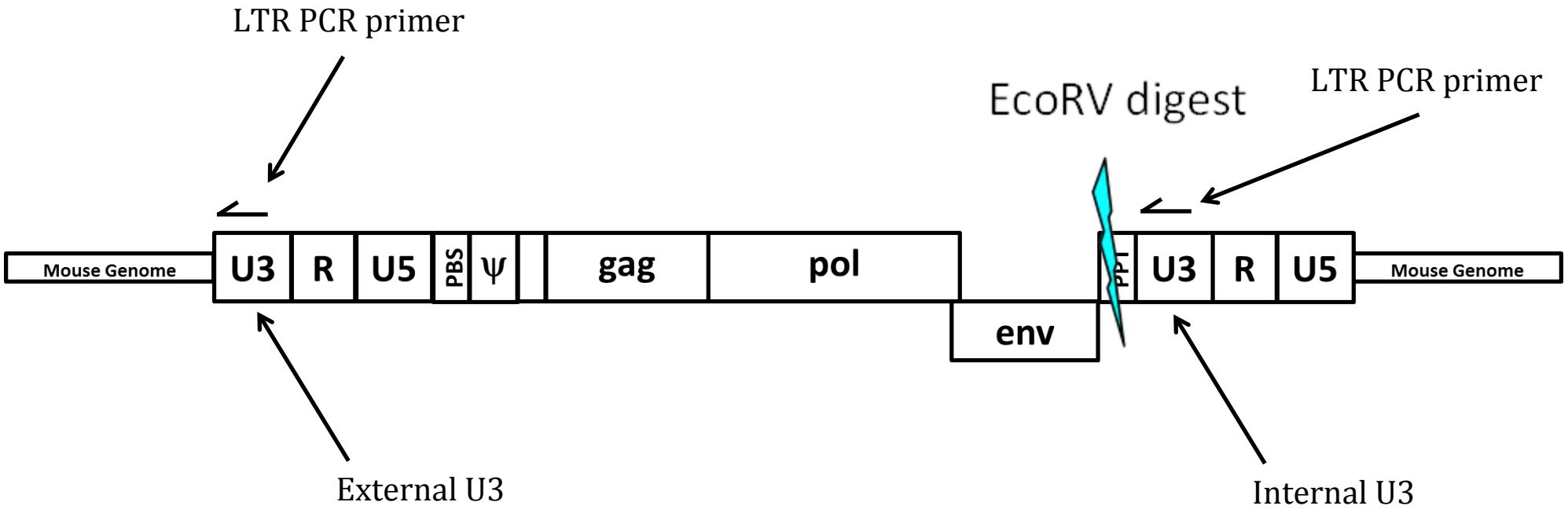


Figure 4-2 EcoRV digest of DNA fragments during library prep
The MoMuLV provirus inserts into the mouse genome after infection as shown above. DNA is fragmented at random by sonication and then ligated to an indexed adaptor during library preparation. This means that some fragments could contain only virus genome and would still be amplified by the primary PCR first strand synthesis primer that is complementary to both the external and internal U3 LTRs of MoMuLV. An EcoRV digest step is included prior to primary PCR to prevent inappropriate amplification of internal virus fragments.

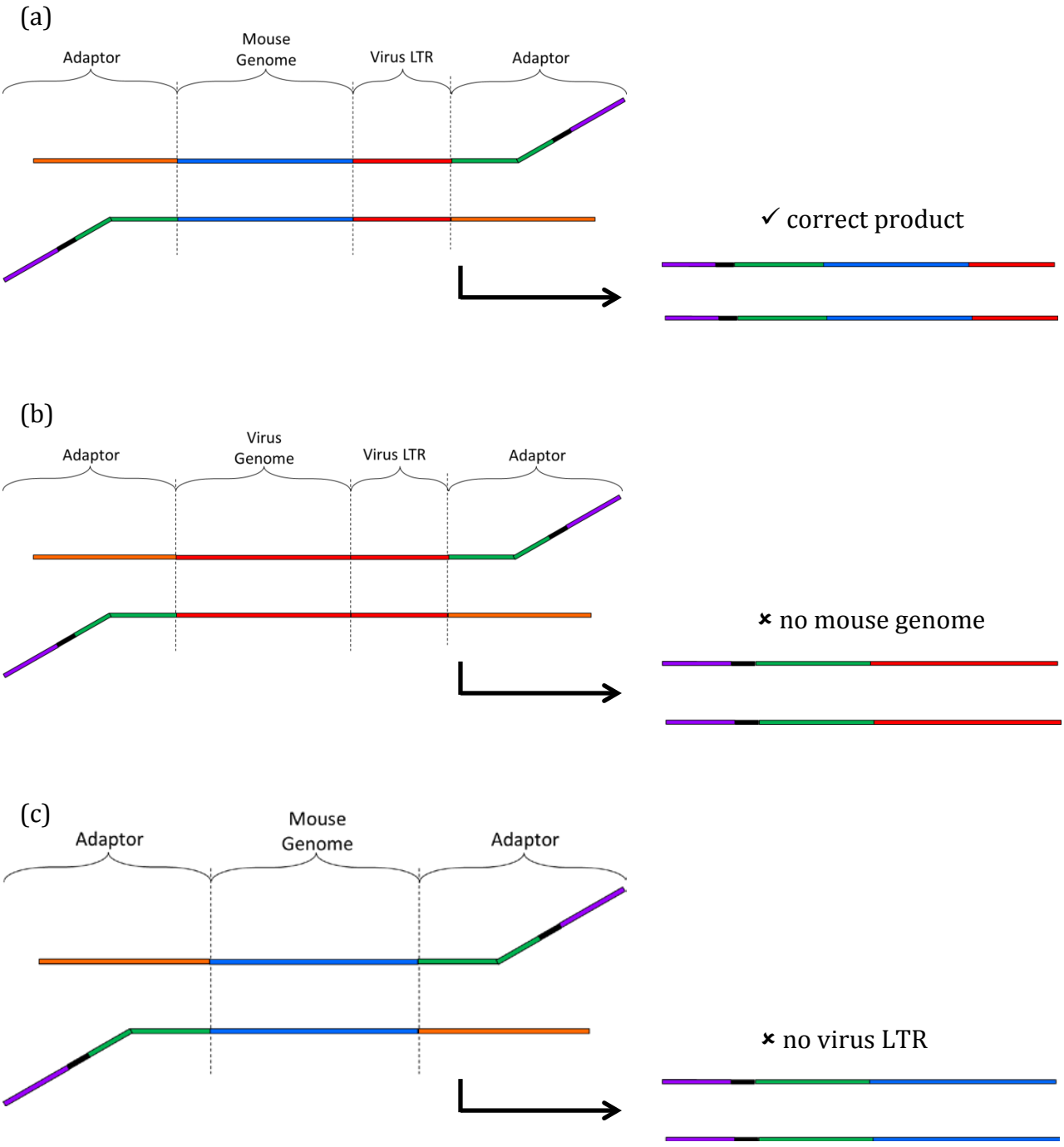


Figure 4-3 Illumina adaptor positioning / PCR strategy
The traditional Illumina library preparation protocol does not include an EcoRV digest. This is performed as the virus LTR primer binding sequence is repeated in the other LTR within a virus. The digest is performed to prevent undesirable amplification of fragments with only virus and no mouse genome, shown in (b). In addition the Illumina adaptor includes an index on the lower strand, meaning there is nothing to prevent inappropriate amplification of fragments containing only mouse genome and no virus insertion site, as in (c). So whilst some of the product will be correct, shown in (a), a significant proportion will be redundant.

The standard Illumina protocol whilst similar to insertion site cloning protocols has differences that required attention. After random shearing of DNA, the fragments of interest should contain both MoMuLV LTR and mouse genome. However, many fragments (in fact the majority) would contain mouse genome only with no LTR insertion. Using the Illumina protocol, which includes the index on the lower strand, would produce the correct product but would also allow the unnecessary PCR amplification of those DNA fragments with no LTR (Figure 4-4). To avoid this, I swapped the sequences of the adaptor lower and upper strands, which included repositioning the index onto the upper strand. In this way, only fragments containing a virus LTR sequence should be amplified. The final adaptor positioning and PCR strategy is shown in Figure 4-5.

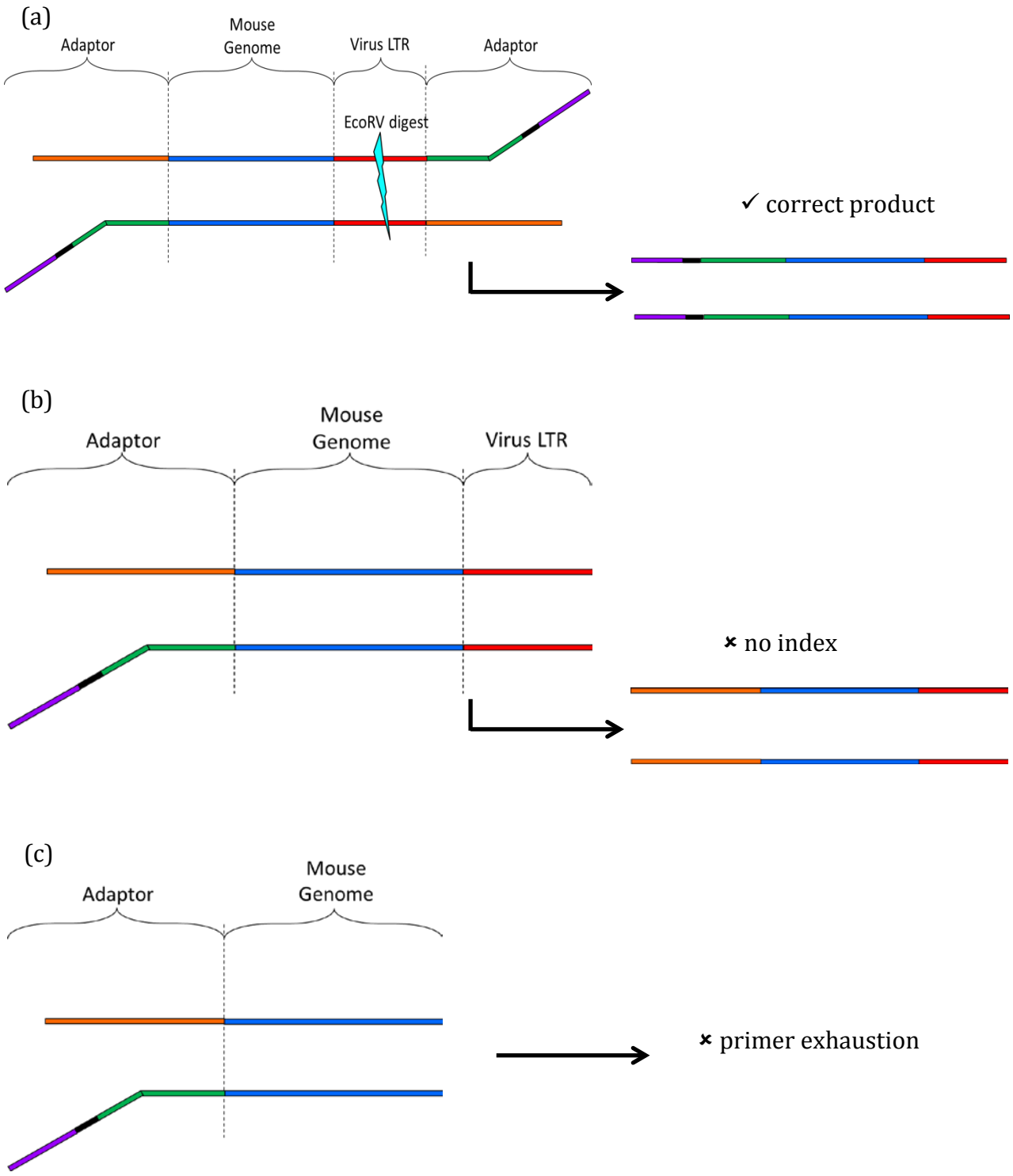


Figure 4-4 Initial adaptor positioning / PCR strategy

The initial strategy used after ligating adaptors to DNA fragments was also problematic. Whilst EcoRV digest eliminates the amplification of fragments with only internal virus portions and no mouse genome (as in Figure 4-3(b)), having the index on the lower strand for the primary PCR means that a proportion of the amplified product may contain no index which could not then be demultiplexed after reading on the HiSeq (b). Also, fragments that contain only mouse genome and no barcode, although would not be amplified, would exhaust one of the PCR primers during first strand synthesis (c). So again, whilst it is possible to amplify the correct product (a), a significant proportion of the product would be undesirable.

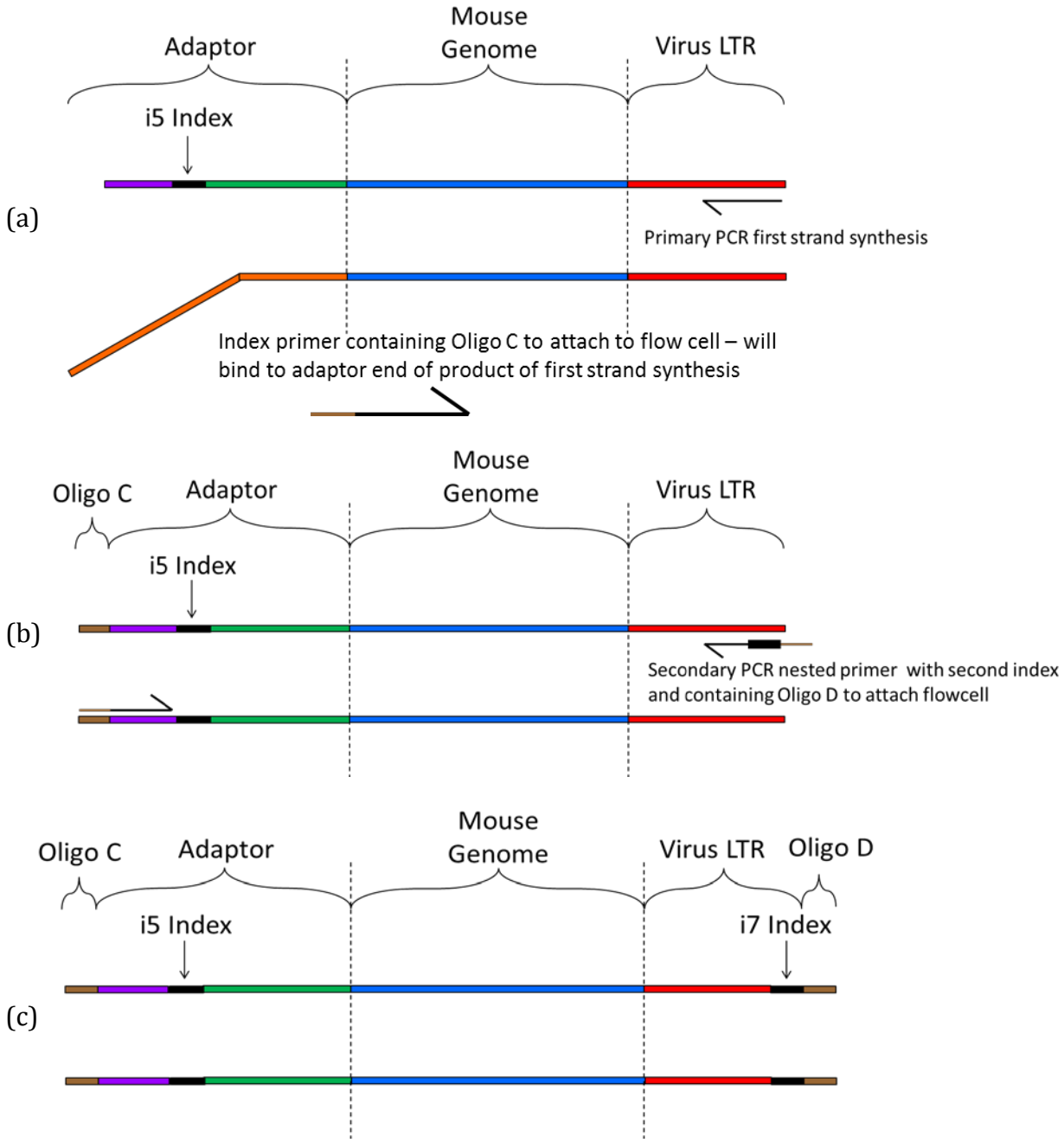


Figure 4-5 Final adaptor positioning / PCR strategy

The above figure shows the adaptor orientation and placement of indices used for nested PCR which is followed by dual-index paired-end sequencing on the Illumina HiSeq. (a) The upper and lower strands, including the position of the i5 adaptor are swapped compared to usual Illumina library preparation. First strand synthesis of the primary PCR starts at the virus LTR end of the fragment and the non-complementary hairpin (in orange) on the lower strand prevents binding of the second primer in first strand synthesis. This means that only fragments containing virus LTR are amplified, which would not be the case if the barcode was in the traditional orientation (see Figure 4-1 and Figure 4-2). During primary PCR, 'Oligo C' (a complementary oligo allowing attachment to the HiSeq flow cell) is added to the adaptor end. (b) The secondary PCR uses a virus LTR primer which is nested to that used in the primary PCR and also includes a second index and also 'Oligo D' (also to attach to the Illumina flowcell). (c) This shows the completed product which is read on the HiSeq. Also note that the positions of Oligo C and D are the opposite of the standard Illumina protocol in order that the HiSeq Read 1 is the adaptor end and not the virus end of the product.

The fact that many fragments contain only mouse genome also provides a problem at the sequencing stage. Although adjusting the position of the index as described above should prevent sequences without virus LTR being amplified, a proportion of the PCR products contain no viral LTR sequence (possibly due to similar endogenous murine retroviral sequences binding the PCR primers). For paired- end sequencing, each end of DNA attaches to the Illumina chip and is sequenced one after the other ('first read' and 'second read'). As only a subset can be sequenced by the LTR primer (the first read) this led to a high proportion of clusters not being sequenced on the Illumina flow cell whilst the second read could sequence many fragments (as every DNA fragment had a indexed adaptor attached regardless of whether it contained MoMuLV LTR). However only clusters that had a successful 'read one' signal could be recognized and subsequently sequenced in read two. This meant the second read was of poor quality and so de-multiplexing of the library was inefficient. I resolved this by reversing the reads (such that the LTR primer is used for the second read) which involved reversing the locations of the Illumina adaptor sequences, 'Oligo C' and 'Oligo D', in the final PCR product.

4.4.4 Increasing throughput

In order to run more samples simultaneously, Illumina indices were not used (at the time the project started only 24 were available from Illumina). Ninety-six 10bp indices were included in the ligated adapter (sequences were taken from the Fluidigm qPCR platform). Furthermore a dual indexing protocol for library preparation was devised to increase the chance that sample cross contamination might be recognised. The second index was added to the primers of the secondary PCR. Groups of four indices were incorporated into the design of the secondary PCR primer and applied to the samples in a 96 well plate in the form of a 'chequerboard' configuration which meant that any of

these four primers was never ligated to DNA in adjacent wells of a plate. This was useful to identify aerosol cross-contamination between samples after de-multiplexing. Due to the nature of the reading on NGS platforms, as much sequence variability in the first 3 bases of these 4 indices was integrated as possible.

Dual indexing is used in the Illumina Nextera protocol, however in view of the altered location of the Illumina named 'i5 index' in my samples (further from the flow cell than usual due to the swapping of upper and lower strands), the Nextera protocol was modified to include 7 dark cycles before reading the i5 index (seeFigure 4-6).

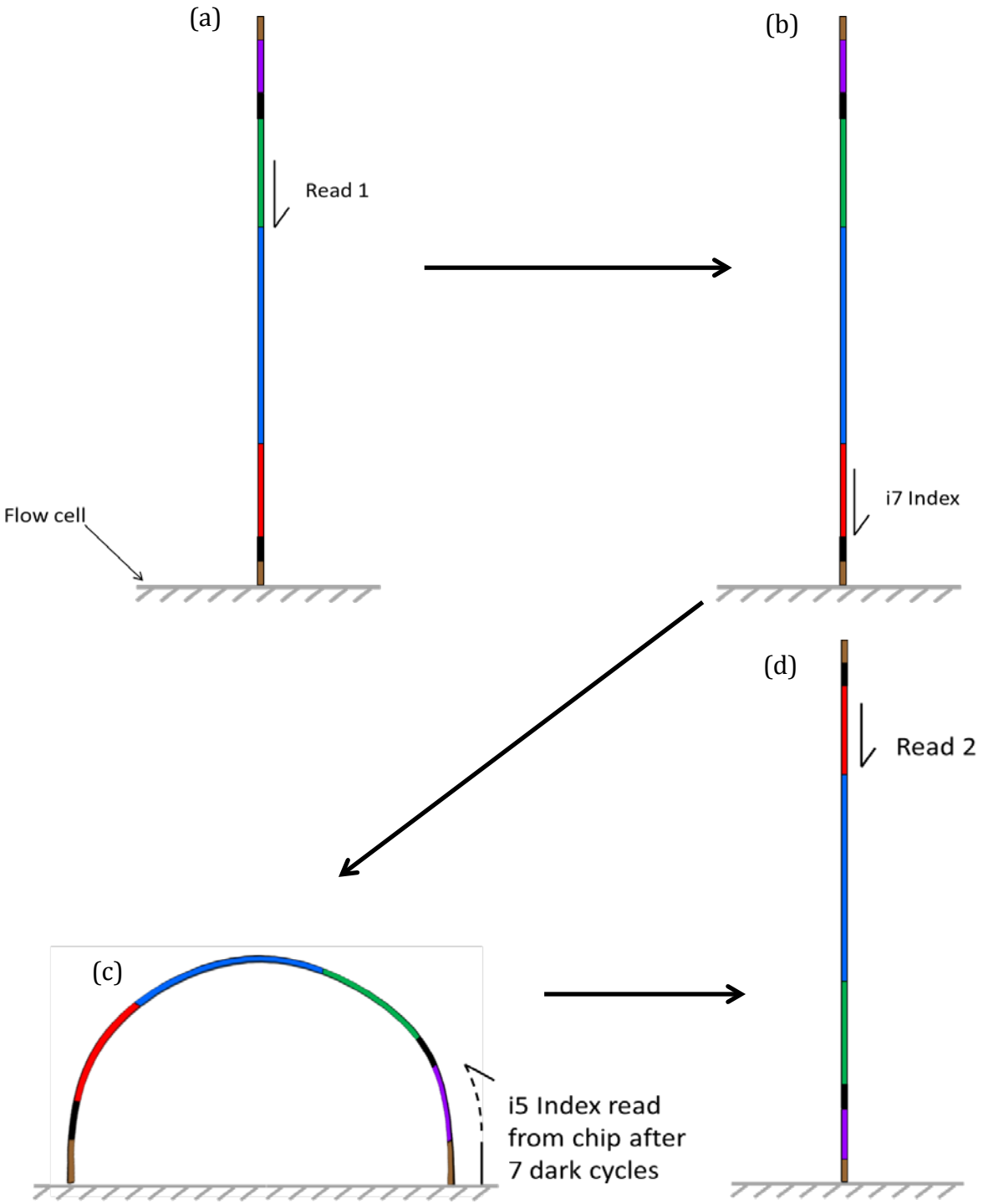


Figure 4-6 Sequencing on Illumina HiSeq

A dual index, paired-end sequencing protocol was designed for sequencing DNA libraries on the Illumina HiSeq. (a) The template binds by oligo D to the complementary oligo on the flow cell. The Read 1 sequencing primer anneals to this template strand during cluster generation. (b) The Read 1 product is removed and the i7 index sequencing primer anneals to the same template strand and reads. (c) The i7 product is removed and the other end of the template anneals to the flow cell via Oligo C. The i5 index is read by the primer grafted on the flow cell which attaches oligo C. Due to our movement of the indices and oligo C and D, 7 dark cycles are added at this point, after which the i5 index is read. (d) The index read product is removed and the original template strand is used to regenerate the complementary strand. The orginal strand is removed to allow binding of the Read 2 sequencing primer.

4.4.5 Improving PCR stringency

Having designed a working PCR / sequencing protocol, I wanted to further improve the stringency of the nested PCR, with respect to temperature and cycle number, to increase the stringency of primer binding and to prevent both amplification bias of valid products and the inappropriate amplification of endogenous sequences as much as possible. The C57BL/6 mouse genome has been fully sequenced. It is known to contain a number of endogenous retroviruses whose sequences can be similar, and occasionally identical, to portions of MoMuLV.

Whilst raising primer annealing temperature and reducing PCR cycle numbers would achieve this, that could be at the expense of the number of mappable reads as too few cycles could mean poor product amplification. Also, as previously mentioned, every fragment of sheared DNA will in theory ligate to an adaptor sequence (read 1) whereas only a proportion of these will contain virus LTR (read 2) and so I wanted to maximise the read 2:read 1 ratio to include as many fragments with virus insertions as possible.

Previous methods have used as many as 29 and 25 cycles for the primary and secondary PCRs respectively (Anthony G Uren et al., 2009). I wanted to test whether this was necessary and so I initially performed quantitative PCR on samples, where the primary PCR ranged from 6 to 30 cycles and the secondary PCR ranged from 6 to 16 cycles (Table 4-1). Figure 4-7(i) shows the amplification curves of samples that had 6 cycles of secondary, having had 15 to 25 cycles of primary PCR. It demonstrates that the secondary PCR was saturated within 6 cycles in all cases. Figure 4-7(ii) shows samples that had 16 cycles in the secondary PCR after 6 to 26 cycles of primary PCR. The secondary PCR plateaued by 11 cycles in all cases except those annotated ‘a.’ that had

only had 6 cycles in the primary PCR. Those samples annotated 'b.' had 11 cycles for primary PCR. Thus the number of cycles could be dramatically reduced in either the primary or the secondary PCR , although the total number of cycles required appears to be more than 20.

I then repeated a full DNA library preparation protocol subjecting samples to a variety of different primary and secondary PCR cycle numbers at two different primer annealing temperatures followed by sequencing on the Illumina MiSeq in order to investigate how read numbers were affected (Table 4-1). The ultimate aim was to use the lowest number of PCR cycles that would achieve the highest number of mappable reads and a high a proportion of successful read 2s (containing virus LTR read) as assessed by the read2:read1 ratio and the percentage of read 2 of the total reads. There was some variability between samples processed with the same cycle numbers, making it more difficult to determine the optimal cycle numbers. However, between the results of the qPCRs and MiSeq, compromise values were chosen and so I proceeded with 12 cycles for both the primary and secondary PCRs. Regarding annealing temperature, those samples processed at a primer annealing temperature of 66°C consistently showed a greater proportion of mappable paired end reads and also a higher fraction of successful read 2s than those processed at 64°C and so the higher temperature was chosen.

				Paired-End Reads			Single-End Reads			
ID	**1° PCR Cycles**	**2° PCR Cycles**	**Annealing Temp. (°C)**	**Total Reads**	**Aligned Reads**	**Aligned Reads (%)**	**Read 1**	**Read 2**	**Read 2 / Total Reads (%)**	**Read 2 / Read 1 Ratio**
1	6	16	64	584286	329436	56%	222571	50495	17.28	0.227
2	11	16	64	581932	372237	64%	204063	147768	50.79	0.724
3	11	11	64	292478	162843	56%	87199	69104	47.25	0.792
4	16	11	64	2139452	1157810	54%	618697	491857	45.98	0.795
5	16	6	64	182444	107478	59%	55602	48087	52.71	0.865
6	6	16	66	123068	75292	61%	37506	35363	57.47	0.943
7	11	16	66	617416	373332	60%	186853	177394	57.46	0.949
8	11	11	66	147476	93044	63%	45839	44419	60.24	0.969
9	16	11	66	415746	262145	63%	129697	125846	60.54	0.970
10	16	6	66	146788	88401	60%	43686	41980	57.20	0.961
11	6	16	64	92222	48506	53%	29585	14246	30.90	0.482
12	11	16	64	307858	183272	60%	100859	72376	47.02	0.718
13	11	11	64	43394	25188	58%	14295	9502	43.79	0.665
14	16	11	64	398008	247936	62%	138100	97074	48.78	0.703
15	16	6	64	413704	197624	48%	161241	15471	7.48	0.096
16	6	16	66	56622	30170	53%	15676	13185	46.57	0.841
17	11	16	66	421176	212307	50%	107995	96915	46.02	0.897
18	11	11	66	82432	48867	59%	24505	22799	55.32	0.930
19	16	11	66	374824	203538	54%	102309	95801	51.12	0.936
20	16	6	66	111088	66659	60%	39709	19925	35.87	0.502
23	16	16	64	4071794	2136295	52%	1186430	854965	41.99	0.721
24	16	16	66	4309220	1818739	42%	983188	710640	32.98	0.723
25	16	6	64	96906	46822	48%	25260	20079	41.44	0.795
26	16	6	64	64294	41570	65%	25599	13558	42.18	0.530
27	16	6	64	54392	29936	55%	18366	10126	37.23	0.551
28	16	6	66	39360	22189	56%	11553	10153	51.59	0.879
29	16	6	66	52576	30811	59%	15846	14271	54.29	0.901
30	16	6	66	148654	64582	43%	31946	29344	39.48	0.919
31	6	16	64	445246	315164	71%	165864	143567	64.49	0.866
32	6	16	64	92236	49541	54%	29060	17656	38.28	0.608
33	11	16	64	291170	155834	54%	96982	50480	34.67	0.521
34	11	16	64	1497838	681922	46%	413053	213091	28.45	0.516
35	11	16	66	465052	300792	65%	162582	130516	56.13	0.803
36	11	16	66	363046	202546	56%	110537	86414	47.60	0.782
37	16	11	64	817296	457964	56%	266312	173217	42.39	0.650
38	16	11	64	1114666	691608	62%	491411	110197	19.77	0.224
39	16	11	64	909696	564600	62%	359666	152937	33.62	0.425
40	16	11	66	1048928	693332	66%	399572	231510	44.14	0.579
41	16	11	66	546254	348786	64%	190646	138718	50.79	0.728
42	16	11	66	1240680	823504	66%	485644	247078	39.83	0.509
44	11	11	64	1330	579	44%	342	212	31.88	0.620
45	11	11	64	86888	47301	54%	28214	17180	39.55	0.609
46	11	11	66	104144	53307	51%	26739	25398	48.77	0.950
47	11	11	66	29850	19130	64%	10397	7777	52.11	0.748
48	11	11	66	3274	1188	36%	978	70	4.28	0.072

Table 4-1 Conditions used for optimisation of ligation-mediated PCR cycle number and primer annealing temperature

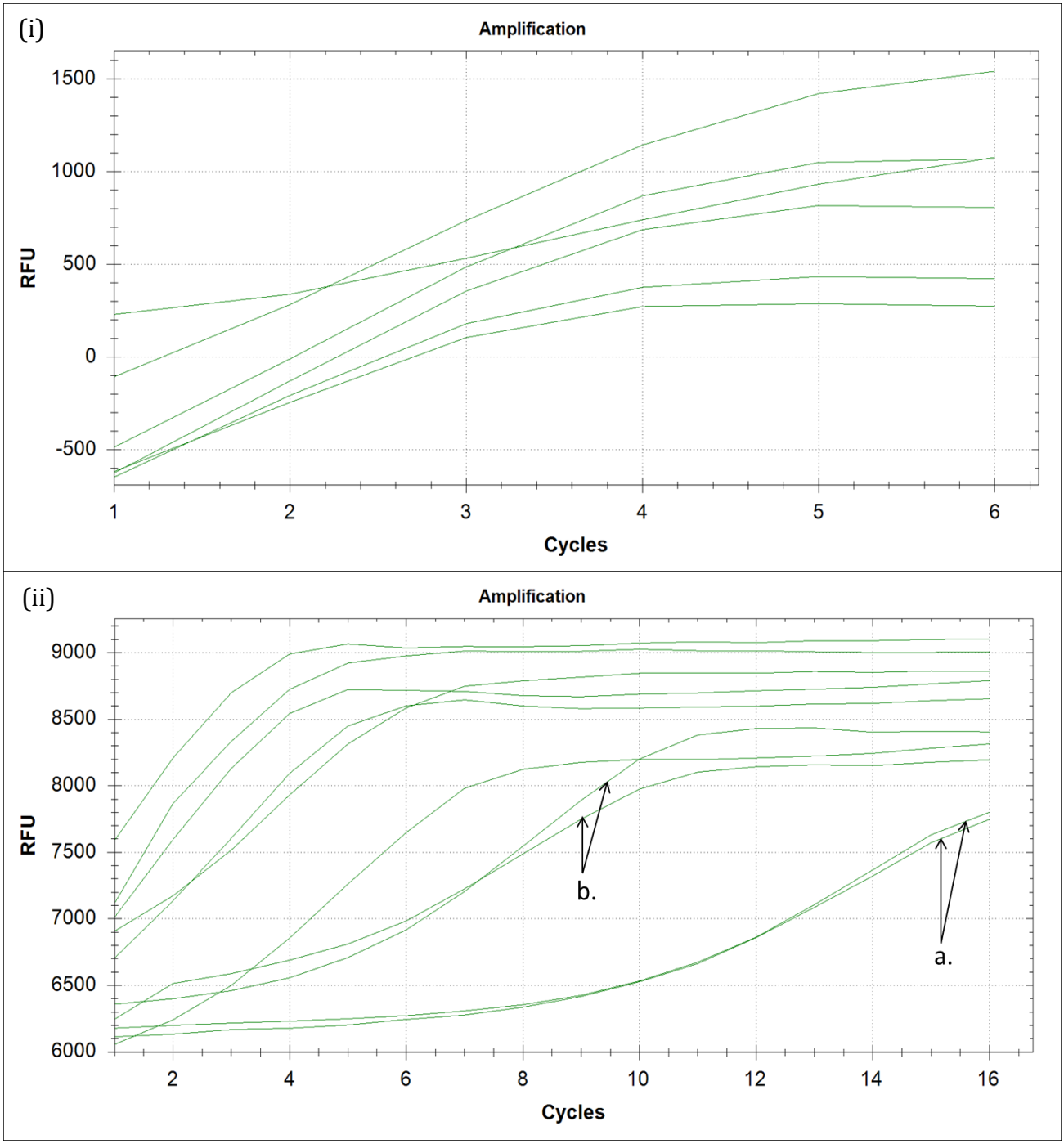


Figure 4-7 qPCR to determine optimal cycle number for nested PCR
In order to optimise the number of cycles for primary and secondary PCR required to most efficiently amplify fragments of infected mouse spleen DNA containing an MoMuLV insertion site, whilst also keeping the cycle number to a minimum in order to avoid amplification bias, and also to avoid amplifying similar endogenous retrovirus, samples were exposed to a variety of cycling number combinations. (i) The 6 cycle secondary PCR of samples undergoing 16-26 cycles in the primary PCR demonstrates that the PCR had reached plateau in all cases. (ii) The 16 cycle secondary PCR of samples undergoing 6 – 26 cycles in the primary PCR shows that the samples plateaued in all cases except a. where the primary was only 6 cycles. b. represents the lowest primary PCR cycle number (11 cycles) but still plateaued by 12 cycles of secondary PCR.

4.4.6 Assessing ability to quantify clonality

I performed a number of quality control checks to ensure that my ligation mediated (LM) PCR and sequencing protocol could be used for quantitative MoMuLV insertion site clonality analysis and to check that the protocol was reproducible.

The genomic DNA from 2 infected mice was mixed in varying ratios and the MoMuLV insertion sites were analysed using my ligation mediated PCR method followed by sequencing on the MiSeq. Mouse 1189 and 1208 had 9 and 11 clonal virus insertion sites respectively as assessed by examining unique fragment lengths and insertions at different bases (Figure 4-8, Figure 4-9). In all cases there was a strong linear correlation between the number of unique fragment lengths and the DNA mixing ratio i.e. the lower the concentration of one mouse DNA within the mixture, the lower the clonality of the insertion whilst the reverse is true for the second mouse DNA whose concentration was increasing. This illustrates that my method can be quantitative of insertion events within heterogeneous samples.

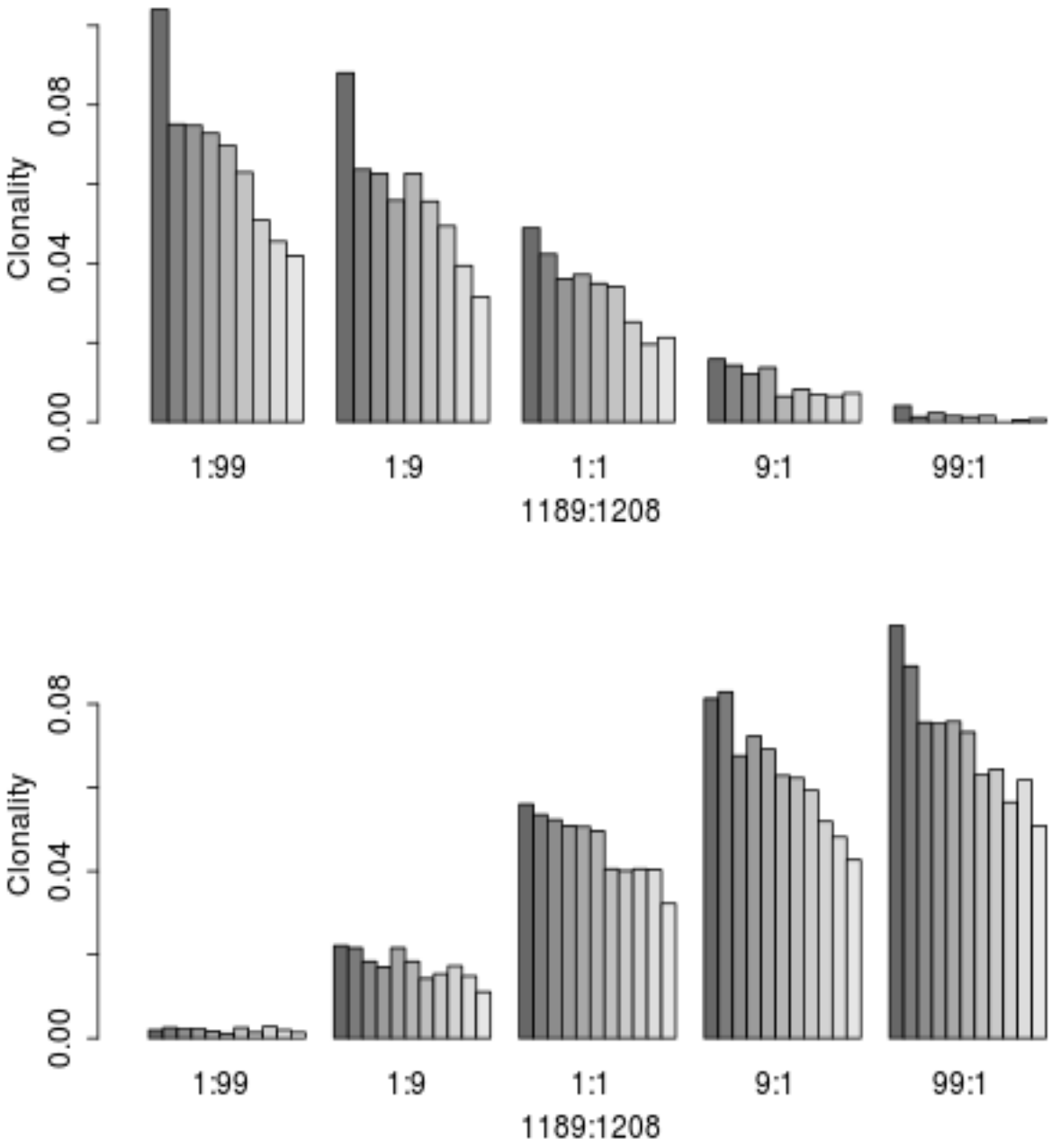


Figure 4-8 Dilution quality control demonstrating quantitative clonality analysis
Genomic DNA from mice 1189 and 1208 were mixed in different proportions as shown above and LM-PCR / sequencing performed to assess the ability to quantitate clonal insertions in a heterogeneous sample. The top row of graphs represents the 9 top clonal insertions in mouse 1189 and the bottom row shows the top 11 clonal insertions in mouse 1208 at those dilutions. As one mouse DNA becomes progressively less concentrated within a sample, the clonality of each insertion decreases. The opposite is true as that mouse DNA is more concentrated.

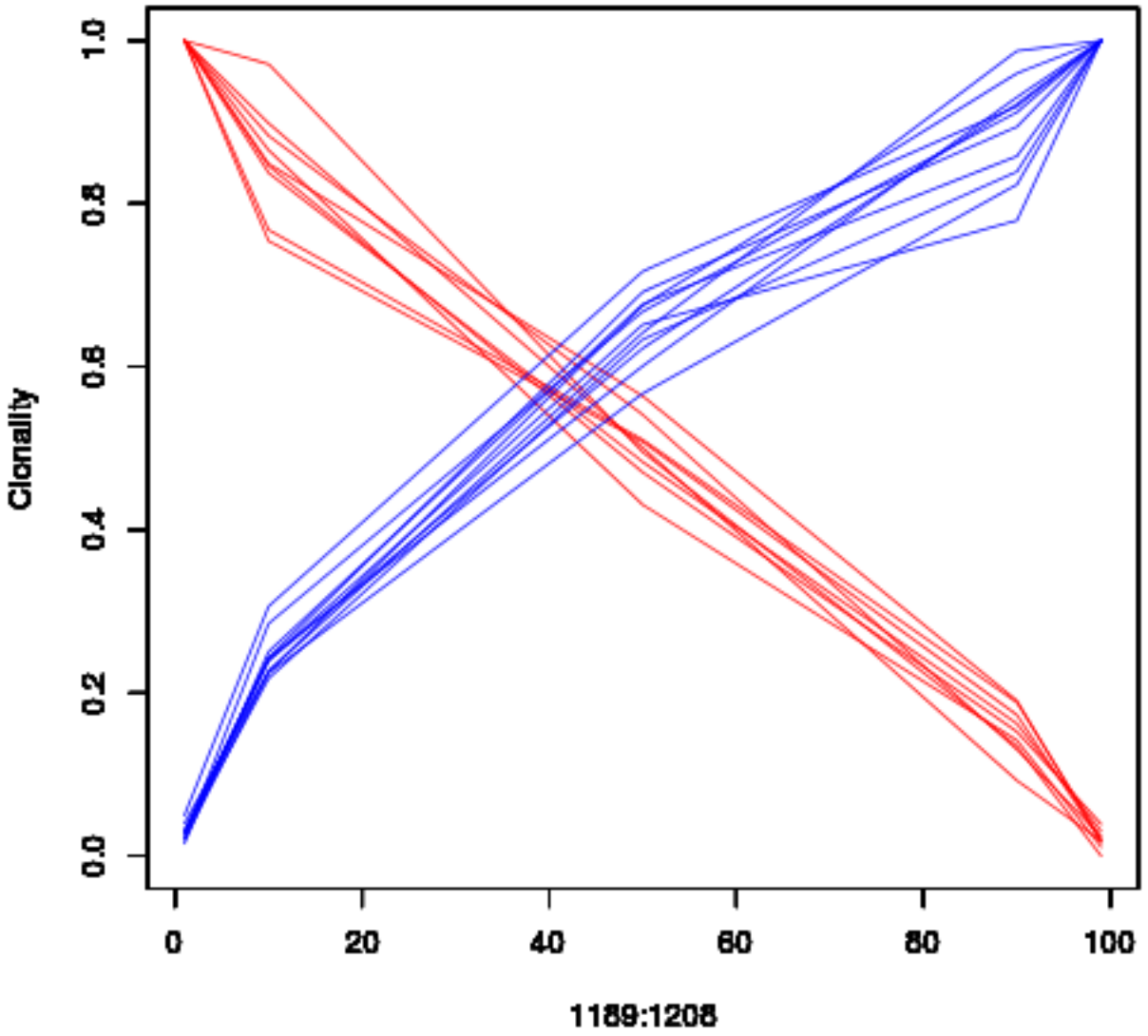
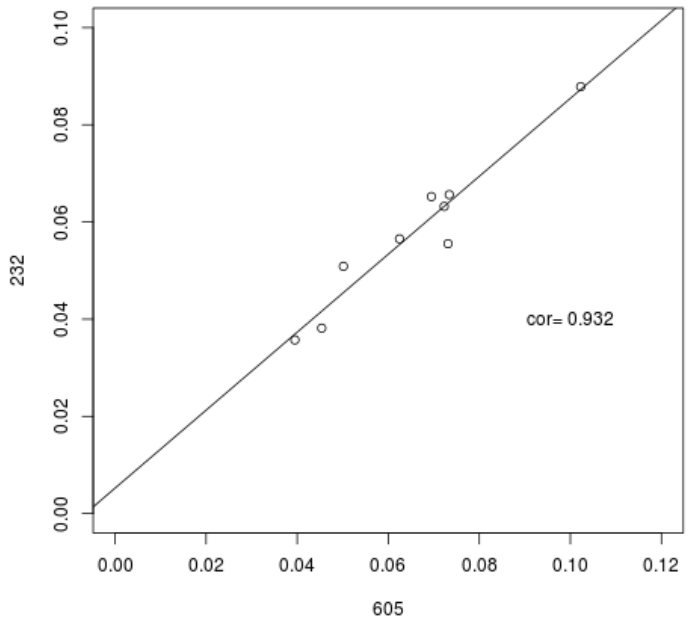


Figure 4-9 The linear relationship between DNA concentration and normalised clonality
The same data as in Figure 4-8 is shown here but this demonstrates the linear relationship that exists between DNA concentration and insertion site clonality. Each red and blue line shows the normalised clonality at different dilutions for each clonal insertion from mouse 1189 and 1208 respectively.

4.4.7 Reproducibility of the library prep / sequencing protocol

Having designed a working protocol to prepare and sequence DNA libraries of infected mice, I wanted to validate that it was reproducible. The samples containing the DNA of the two mice used in the dilution/clonality validation in section 4.4.6 can also be used to show reproducibility. For each mouse, comparing the raw clonality values of the most abundant virus inserts in the sample containing solely DNA from that mouse with the raw clonality values of those same inserts where the DNA from that same mouse was 99% of the input (ie. the 99:1 dilution) shows good reproducibility of the method (see Figure 4-10). Although there is a 1% difference in the DNA input from that mouse between these two samples, and so you would expect a 1% difference in raw clonality, this is within the error of DNA quantification combined with varying losses at each stage of DNA clean-up between library prep stages and so it is still appropriate to compare these two samples in the case of each mouse.

(a)



(b)

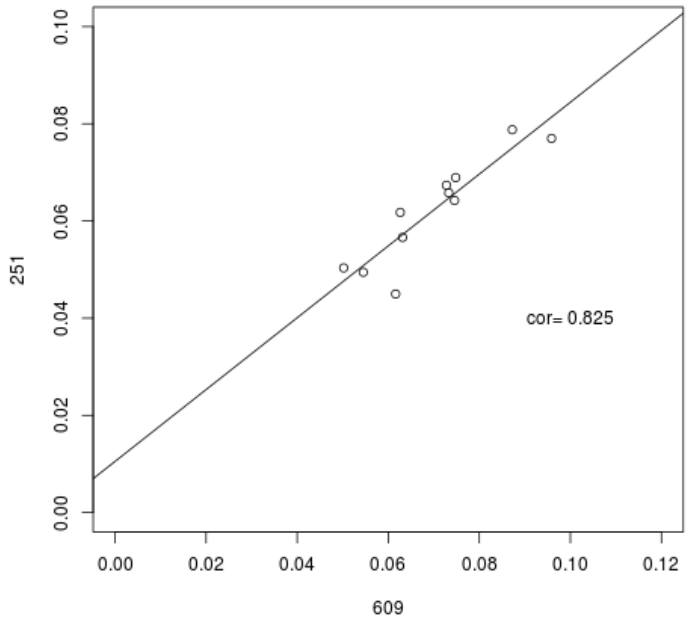


Figure 4-10 The reproducibility of the library prep and sequencing protocol
The raw clonality values of those insertions in the 2 samples used in the dilution experiment (Figure 4-6, Figure 4-7) also show the reproducibility of the ligation mediated PCR and sequencing protocol in separate samples. The points on graphs (a) and (b) represent the top 9 and 11 clonal insertion sites for mice 1189 and 1208 respectively. The y-axis represents the raw clonality values of these clonal insertions where the genomic DNA sequenced was derived solely from one mouse 1189, (a), and 1208 (b). The x axis represents the raw clonality values of a 99:1 dilution of the genomic DNA from 1189:1208 in (a) and 1208:1189 in (b). Both graphs show reproducible clonality values in these raw clonality values across the samples.

CHAPTER 5 RESULTS & DISCUSSION: SEQUENCING

The ligation mediated PCR of genomic DNA of infected and uninfected control mice was processed in seven batches of 96 samples. Each of these 672 samples (that included diseased mice, time course mice and quality control samples) had a unique combination of two indexs that allowed identification and demultiplexing after sequencing. After sequencing on the Illumina HiSeq, there were 39, 723,688 paired end reads / 2,207,189 total fragments mappable to the mouse genome. After grouping, there were 975,932 total MoMuLV insertion sites across the genome. After removal of contamination (recurrent PCR products identified in controls as well as MoMuLV infected samples) and data from replicate mice there were 20, 642,452 paired end reads / 1,256,544 fragments and 762,228 MoMuLV inserts.

Historically within the field of insertional mutagenesis, it has been difficult to predict which surrounding gene is affected by a given insert. In the case that there are several genes within the region of an insert it could be that any one gene, or a combination thereof, may be driving the onset of cancer. Furthermore, it may be that different gene / genes are deregulated depending on tumour lineage (Sauvageau et al., 2008). Manual mapping to putative targets can potentially introduce bias, preferring known cancer genes and thus overlooking possible novel genes. Nearest-gene mapping assigns the nearest gene to an insert, but does not aggregate insertion data across tumours and ignores insert orientation. From this list of 762,228 insertion sites I have identified common insertion / integration sites using GKC and KCRBM.

Table 5-1 shows the top 50 CISs, as determined by GKC, with a normalised clonality >0.05, identified using a 100,000bp window across the genome from those mice who developed disease (i.e. time course mice excluded). Gene names were subsequently assigned to each peak in an automated fashion by KC-RBM. Table 5-2 is a list of most commonly tagged genes

provided by using the KC-RBM which directly reports genes rather than insertions. KC-RBM uses a 20,000bp kernel for GKC by default. Insertions are annotated to genes (peaks are not reported by KCRBM).

It is worthwhile again noting that the gene names in these tables are automatic annotations from KC-RBM. Whilst this software uses predetermined rules to link an insertion site with a gene, it is possible that an alternative gene nearby may be the oncogene or tumour suppressor gene of interest. An example of this is the gene *Evi5* (ranked as the top CIS gene in both Table 5-1 and Table 5-2). Although this gene was identified when applying KC-RBM, just downstream of *Evi5* is *Gfi1*. Both of these genes have been shown to cooperate with human *BCL6* in a MoMuLV insertional mutagenesis screen in mice (Baron et al., 2014) and both have previously been implicated in haematological malignancies (X. Liao, Du, Morse, Jenkins, & Copeland, 1997; Xu & Kee, 2007). It is therefore possible that either or both genes are promoting lymphomagenesis in this screen, but *Gfi1* was not assigned.

Table 5-3 shows the gene ontology terms most frequently identified when the genes in Table 5-1 and Table 5-2 are input into the DAVID Bioinformatics Database (http://david.abcc.ncifcrf.gov/home.jsp). This program groups genes depending on biological processes and in this case recognizes many groups that are associated with transcription. This is consistent with a role in cancer, since a high proportion of genes mutated in cancer are thought to be transcriptional regulators (Futreal et al., 2004).

This work has identified a number of known proto-oncogenes like Myc, Notch1, Runx3 and Pvt1, tumour suppressor genes like Ikzf1, as well as a number of less well characterised / unknown candidate genes such as Pou2f2 (Oct2), Il2ra and Ubac2. Specific genes are discussed in detail in chapter 7.

Rank	Chromo-some	Base Position	Gene Name	Ensembl Gene ID	No. of insertions in wild-type mice	No. of insertions in *BCL2* transgenic mice	All other insertions in wild-type mice	All other insertions in *BCL2* transgenic mice
1	5	107739755	**Evi5**	ENSMUSG00000011831	86	79	7523	6639
2	17	47535909	**Taf8**	ENSMUSG00000023980	39	29	7570	6689
3	2	117396234	**Gm13982**	ENSMUSG00000085681	39	15	7570	6703
4	17	29480622	**AC163629.2**	ENSMUSG00000097125	32	18	7577	6700
5	3	30101462	**Mecom**	ENSMUSG00000027684	29	18	7580	6700
6	2	26498864	**Notch1**	ENSMUSG00000026923	31	12	7578	6706
7	7	25135338	**Pou2f2**	ENSMUSG00000008496	5	32	7604	6686
8	2	11626718	**Il2ra**	ENSMUSG00000026770	25	12	7584	6706
9	20	52750305	**Mir106a**	ENSMUSG00000065456	14	17	7595	6701
10	10	21182853	**Hbs1l, Ahi1**	ENSMUSG00000019977, ENSMUSG00000019986	14	15	7595	6703
11	11	11708428	**Ikzf1**	ENSMUSG00000018654	24	5	7585	6713
12	12	85612438	**Batf**	ENSMUSG00000034266	17	11	7592	6707
13	20	7872448	**Otud5, Hdac6**	ENSMUSG00000031154, ENSMUSG00000031161	14	12	7595	6706
14	11	100871963	**Stat3**	ENSMUSG00000004040	18	6	7591	6712
15	16	49806237	**Gm15518**	ENSMUSG00000087066	16	7	7593	6711
16	2	170197275	**AL844576.1**	ENSMUSG00000097514	17	6	7592	6712
17	8	122695703	**Gm20388**	ENSMUSG00000092329	12	10	7597	6708
18	19	4166927	**Rps6kb2**	ENSMUSG00000024830	12	9	7597	6709
19	13	28918766	**2610307P16Rik, Sox4**	ENSMUSG00000085936, ENSMUSG00000076431	16	5	7593	6713
20	11	23804521	**Rel, Gm12061**	ENSMUSG00000020275, ENSMUSG00000084769	10	11	7599	6707
21	11	68430221	**Ntn1**	ENSMUSG00000020902	14	7	7595	6711
22	5	148978142	**Katnal1**	ENSMUSG00000041298	11	9	7598	6709
23	13	30780305	**Exoc2, Dusp22**	ENSMUSG00000021357, ENSMUSG00000069255	9	10	7600	6708
24	14	115040940	**Gpc5, Mir18**	ENSMUSG00000022112, ENSMUSG00000065403	9	10	7600	6708

Table 5-1, continued on next page

25	15	74946659	**Cyp11b2, Ly6i**	ENSMUSG00000022589, ENSMUSG00000022586	10	9	7599	6709
26	11	98489843	**Ikzf3**	ENSMUSG00000018168	4	15	7605	6703
27	2	167813792	**Gm14319**	ENSMUSG00000085411	12	7	7597	6711
28	4	134107879	**Cd52, Sh3bgrl3**	ENSMUSG00000000682, ENSMUSG00000028843	11	7	7598	6711
29	4	149697114	**Slc25a33**	ENSMUSG00000028982	8	10	7601	6708
30	8	95014994	**Gpr56**	ENSMUSG00000031785	6	11	7603	6707
31	1	171913094	**Slamf6**	ENSMUSG00000015314	6	11	7603	6707
32	4	135086488	**Runx3**	ENSMUSG00000070691	10	6	7599	6712
33	11	99143950	**Ccr7**	ENSMUSG00000037944	8	8	7601	6710
34	11	87750777	**Supt4h1**	ENSMUSG00000020485	8	8	7601	6710
35	6	127217164	**Gm7308, Ccnd2**	ENSMUSG00000081113, ENSMUSG00000000184	10	5	7599	6713
36	4	32342369	**Bach2**	ENSMUSG00000040270	12	3	7597	6715
37	15	97767019	**Endou**	ENSMUSG00000022468	7	8	7602	6710
38	12	86855748	**2310044G17Rik**	ENSMUSG00000034157	4	11	7605	6707
39	10	60163321	**Anapc16**	ENSMUSG00000020107	8	7	7601	6711
40	4	136074882	**Gale, Tceb3**	ENSMUSG00000028671, ENSMUSG00000028668	10	5	7599	6713
41	2	127333979	**Astl**	ENSMUSG00000050468	6	9	7603	6709
42	2	152784061	**Bcl2l1**	ENSMUSG00000007659	10	5	7599	6713
43	16	92796986	**Runx1**	ENSMUSG00000022952	11	3	7598	6715
44	19	37500001	**Exoc6**	ENSMUSG00000053799	10	4	7599	6714
45	6	48688733	**Gimap7**	ENSMUSG00000043931	6	8	7603	6710
46	7	73601490	**Chd2, U6**	ENSMUSG00000078671, ENSMUSG00000065729	9	5	7600	6713
47	5	33694986	**Slbp**	ENSMUSG00000004642	9	5	7600	6713
48	11	119138138	**Ccdc40**	ENSMUSG00000039963	6	8	7603	6710
49	18	4318255	**Map3k8**	ENSMUSG00000024235	4	9	7605	6709
50	19	43615987	**Nkx2-3**	ENSMUSG00000044220	7	6	7602	6712

Table 5-1 Top 50 common insertion site genes from insertional mutagenesis screen derived by Gaussian kernal convolution
The top 50 MoMuLV common insertion site genes from infected mice in the 'most clonal' group (see Figure 6-8) based on normalised clonality (above 0.05) with a 100, 000bp window. Derived by 'CIMPL' which uses Gaussian kernel convolution values to rank the insertions.

Rank	Gene	Ensembl Gene ID	Total insertions	No. of insertions in *BCL2* transgenic mice	No. of insertions in wild-type mice	All other insertions in *BCL2* transgenic mice	All other insertions in wild-type mice
1	**Evi5**	ENSMUSG00000011831	182	87	95	6631	7514
2	**Gm20388**	ENSMUSG00000092329	90	39	51	6679	7558
3	**Taf8**	ENSMUSG00000023980	83	38	45	6680	7564
4	**Gm26885**	ENSMUSG00000097125	79	34	45	6684	7564
5	**Gm13982**	ENSMUSG00000085681	67	23	44	6695	7565
6	**Pvt1**	ENSMUSG00000097039	60	29	31	6689	7578
7	**Hbs1l**	ENSMUSG00000019977	59	28	31	6690	7578
8	**Notch1**	ENSMUSG00000026923	57	18	39	6700	7570
9	**Mecom**	ENSMUSG00000027684	57	22	35	6696	7574
10	**Pou2f2**	ENSMUSG00000008496	47	35	12	6683	7597
11	**Il2ra**	ENSMUSG00000026770	43	15	28	6703	7581
12	**Ahi1**	ENSMUSG00000019986	40	17	23	6701	7586
13	**Rps6kb2**	ENSMUSG00000024830	38	16	22	6702	7587
14	**Bach2**	ENSMUSG00000040270	38	9	29	6709	7580
15	**Gm17619**	ENSMUSG00000097514	38	11	27	6707	7582
16	**Ahdc1**	ENSMUSG00000037692	37	18	19	6700	7590
17	**2610307P16Rik**	ENSMUSG00000085936	37	6	31	6712	7578
18	**Anapc16**	ENSMUSG00000020107	36	18	18	6700	7591
19	**Ikzf1**	ENSMUSG00000018654	35	8	27	6710	7582
20	**Runx3**	ENSMUSG00000070691	34	14	20	6704	7589
21	**Stat3**	ENSMUSG00000004040	33	12	21	6706	7588
22	**Cnn2**	ENSMUSG00000004665	33	13	20	6705	7589
23	**Katnal1**	ENSMUSG00000041298	33	15	18	6703	7591
24	**Gm7308**	ENSMUSG00000081113	32	14	18	6704	7591
25	**Pde4c**	ENSMUSG00000031842	31	14	17	6704	7592
26	**Mir106a**	ENSMUSG00000065456	31	17	14	6701	7595
27	**Gm26745**	ENSMUSG00000097213	31	17	14	6701	7595
28	**Slc25a33**	ENSMUSG00000028982	30	15	15	6703	7594

Table 5-2, continued on next page

29	**Clec2e**	ENSMUSG00000030155	30	16	14	6702	7595
30	**Optc**	ENSMUSG00000010311	29	18	11	6700	7598
31	**Nfic**	ENSMUSG00000055053	29	14	15	6704	7594
32	**Gm14319**	ENSMUSG00000085411	29	11	18	6707	7591
33	**Ikzf3**	ENSMUSG00000018168	27	19	8	6699	7601
34	**Ntn1**	ENSMUSG00000020902	27	8	19	6710	7590
35	**Poll**	ENSMUSG00000025218	27	11	16	6707	7593
36	**Upk2**	ENSMUSG00000041523	27	14	13	6704	7596
37	**Brat1**	ENSMUSG00000000148	26	10	16	6708	7593
38	**Vps13d**	ENSMUSG00000020220	26	9	17	6709	7592
39	**Msh5**	ENSMUSG00000007035	25	14	11	6704	7598
40	**Exoc2**	ENSMUSG00000021357	25	12	13	6706	7596
41	**Gm15518**	ENSMUSG00000087066	25	9	16	6709	7593
42	**Myc**	ENSMUSG00000022346	24	10	14	6708	7595
43	**Ncl**	ENSMUSG00000026234	24	14	10	6704	7599
44	**Otud5**	ENSMUSG00000031154	24	13	11	6705	7598
45	**Ubac2**	ENSMUSG00000041765	24	11	13	6707	7596
46	**Gm22704**	ENSMUSG00000064961	24	14	10	6704	7599
47	**Ly86**	ENSMUSG00000021423	23	12	11	6706	7598
48	**Endou**	ENSMUSG00000022468	23	11	12	6707	7597
49	**Pitpnm2**	ENSMUSG00000029406	23	10	13	6708	7596
50	**Fgfr2**	ENSMUSG00000030849	23	11	12	6707	7597

Table 5-2 Top 50 common insertion site genes from insertional mutagenesis screen derived by Kernel Convolved Rules Based Mapping (KC-RBM)

Insertion site positions are determined by Gaussian Kernel Convolution and then KC-RBM software is used to determine the most likely implicated gene based on using four windows sizes (upstream-sense and antisense, downstream-sense and antisense, with respect to the transcription start site).

GO Term	Count	%	Genes	PValue	Benjamini corrected PValue	False Discovery Rate
GO:0045941~positive regulation of transcription	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	4.97E-08	3.84E-05	7.60E-05
GO:0010628~positive regulation of gene expression	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	6.69E-08	2.58E-05	1.02E-04
GO:0006350~transcription	22	3.648	IKZF3, BACH2, IKZF1, SMAD7, RREB1, TAF8, TCOF1, SOX4, STAT3, BATF, NOTCH1, FLI1, ETS1,	9.90E-08	2.55E-05	1.51E-04
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	1.08E-07	2.09E-05	1.65E-04
GO:0045893~positive regulation of transcription, DNA-dependent	12	1.990	NOTCH1, IKZF1, ETS1, ZMIZ1, TAF8, SOX4, RUNX1, MECOM, NFIC, MYC, STAT3, NKX2-3	1.25E-07	1.93E-05	1.91E-04
GO:0051254~positive regulation of RNA metabolic process	12	1.990	NOTCH1, IKZF1, ETS1, ZMIZ1, TAF8, SOX4, RUNX1, MECOM, NFIC, MYC, STAT3, NKX2-3	1.35E-07	1.73E-05	2.06E-04
GO:0051173~positive regulation of nitrogen compound metabolic process	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	1.51E-07	1.67E-05	2.31E-04
GO:0010557~positive regulation of macromolecule biosynthetic process	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	1.64E-07	1.59E-05	2.51E-04
GO:0031328~positive regulation of cellular biosynthetic process	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	2.55E-07	2.19E-05	3.89E-04
GO:0009891~positive regulation of biosynthetic process	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	2.81E-07	2.17E-05	4.29E-04
GO:0045944~positive regulation of transcription from RNA polymerase II	11	1.824	NOTCH1, IKZF1, ETS1, ZMIZ1, SOX4, RUNX1, MECOM, NFIC, MYC, STAT3, NKX2-3	3.09E-07	2.17E-05	4.72E-04
GO:0006355~regulation of transcription, DNA-dependent	19	3.151	BACH2, IKZF1, SMAD7, TAF8, SOX4, MECOM, STAT3, BATF, NOTCH1, FLI1, ETS1, ZMIZ1,	7.44E-07	4.79E-05	0.00113679
GO:0051252~regulation of RNA metabolic process	19	3.151	BACH2, IKZF1, SMAD7, TAF8, SOX4, MECOM, STAT3, BATF, NOTCH1, FLI1, ETS1, ZMIZ1,	9.37E-07	5.57E-05	0.00143178
GO:0010604~positive regulation of macromolecule metabolic process	13	2.156	IKZF1, TAF8, SOX4, MECOM, STAT3, NOTCH1, ETS1, ZMIZ1, POU2F2, NFIC, RUNX1, MYC, NKX2-3	1.10E-06	6.05E-05	0.00167629
GO:0045449~regulation of transcription	22	3.648	IKZF3, BACH2, IKZF1, SMAD7, RREB1, TAF8, SOX4, MECOM, STAT3, BATF, NOTCH1, FLI1, ETS1,	4.73E-06	2.44E-04	0.00722933
GO:0042127~regulation of cell proliferation	11	1.824	FGFR2, NOTCH1, IL2RA, CCND2, ZMIZ1, DUSP22, MECOM, MYC, NTN1, RUNX3, NKX2-3	1.20E-05	5.80E-04	0.01835547
GO:0006357~regulation of transcription from RNA polymerase II	11	1.824	NOTCH1, IKZF1, ETS1, ZMIZ1, SOX4, RUNX1, MECOM, NFIC, MYC, STAT3, NKX2-3	3.86E-05	0.0017499	0.05890728
GO:0008284~positive regulation of cell proliferation	7	1.161	FGFR2, NOTCH1, IL2RA, CCND2, ZMIZ1, MYC, NTN1	4.13E-04	0.01756999	0.62946459
GO:0051094~positive regulation of developmental process	6	0.995	FGFR2, NOTCH1, IKZF1, ETS1, RUNX1, NTN1	8.28E-04	0.03310586	1.25793111
GO:0007507~heart development	6	0.995	NOTCH1, SMAD7, ZMIZ1, SOX4, MECOM, NFATC1	9.96E-04	0.03774543	1.51138166

Table 5-3 Gene ontology of common insertion sites found in insertional mutagenesis screen

The common insertion sites from the insertional mutagenesis screen identified by both Gaussian Kernel Convolution and Kernel Convolved Rules Based Mapping were input to DAVID Bioinformatics Resources 6.7, using the 'GO Fat' database to functionally annotate groups of genes (Huang, Sherman, & Lempicki, 2009a, 2009b).

CHAPTER 6 RESULTS & DISCUSSION: INSERTION KINETICS & TIME COURSE

I wanted to investigate the clonal abundance of virus insertion sites in different samples and also to examine the presence of specific common insertion sites over time in the lead up to clinically detectable disease. As described in chapter 3 there were three different cohorts of mice involved in this study. In addition to those mice that were maintained until detectable disease, within each cohort infected and uninfected litters of mice were sacrificed at predetermined time points. These time points included days 2, 5, 9, 14, 28, 42, 56, 84 and 112 post injection with MoMuLV or vehicle. Table 6-1 summarises these groups.

As expected, 13 of the VavP-*BCL2* F1 mice from the later time points (day 84 and day 112) developed detectable disease prior to their predetermined date and so had to be sacrificed early. All other mice survived until their allotted time points with no signs of lymphoma. The clonality of insertions from these mice will be discussed in detail later in this chapter.

Time after injection	Cohort 1				Cohort 2				Cohort 3			
	Virus Infected		Uninfected Control		Virus Infected		Uninfected Control		Virus Infected		Uninfected Control	
	WT C57BL/6	Eμ-*BCL2* C57BL/6	WT C57BL/6	Eμ-*BCL2* C57BL/6	WT F1	Eμ-BCL2 F1	WT F1	Eμ-BCL2 F1	WT F1	VavP-*BCL2* F1	WT F1	VavP-*BCL2* F1
Day 2	3	4	3	1	2	3	5	1	1	3	1	3
Day 5	-	-	5	1	-	-	-	-	7	2	6	2
Day 9	1	3	7	1	-	-	5	7	0	6	4	2
Day 14	4	3	4	2	7	3	5	3	-	-	-	-
Day 28	0	7	3	4	4	4	5	5	2	3	5	2
Day 42	3	4	3	2	5	4	7	5	4	6	8	1
Day 56	2	4	2	4	5	4	5	5	5	5	2	4
Day 84	3	5	4	3	-	-	-	-	8	10	-	-
Day 112	4	3	4	3	-	-	-	-	8	2	4	6

Table 6-1 Time course mice used in the insertional mutagenesis screen

For each of the 3 cohorts of mice used in the insertional mutagenesis screen, in addition to those mice that were maintained until they developed detectable disease and used in the Kaplan Meier survival analysis, litters of mice were infected with either MoMuLV or vehicle control and sacrificed at predetermined time points after injection. The numbers of mice in each group are shown above. Within cohort 3, 13 infected and control transgenic mice in the day 84 and 112 groups developed disease prior to reaching their time points and had to be sacrificed early.

6.1 MoMuLV Quantification and Kinetics

As part of investigating the kinetics of virus insertion site accumulation, it is important to study virus quantification, especially considering that two separate preparations of virus were used. Quantification of replication competent retroviruses can be a lengthy process, and involves inoculating a permissive cell line which is passaged for 3 weeks and analysing the media for the presence of virus. As an alternative to this, the assay in my study uses DERSE (Detector of Exogenous Retroviral Sequence Elements) plasmids in a host cell line permissive of the retrovirus of interest (Aloia et al., 2013).

This method of virus quantification was devised to create a luciferase readout of reverse transcription and virus integration. The inGluc-MLV-DERSE plasmid consists of a Gaussia luciferase (Gluc) sequence oriented in a reverse direction with respect to flanking MoMuLV non-coding sequences. Within the *Gluc* sequence is an intron that is oriented in a forward direction and can be spliced by the host cell. In the absence of MoMuLV, only minus strand, spliced *Gluc* sequences are present in RNA in the cell. Once infected by MoMuLV, the DERSE cell packages the RNA containing the minus-strand Gluc sequence. In the next round of infection, reverse transcription of the encapsidated RNA produces a double-stranded DNA containing an uninterrupted *Gluc* gene that is coding and is integrated into the DNA of, and subsequently expressed by, the infected cell. Gluc is released into the cell culture media and can be quantified on a luminometer as a measure of virus activity (Figure 6-1).

For these assays we used 293mCAT1 cells (kindly provided by the lab of Dr. Alan Rein, National Cancer Institute, Frederick, MD, USA) which is a human cell line that has had the mCAT1 receptor of MuLV introduced so that it can be infected by MuLV (human

cells lack this receptor). The inGluc-MLV-DERSE reporter plasmid has also been introduced to these cells. Cells are plated and treated with a serial dilution of virus supernatant and luciferase signal is counted at a series of time points after infection.

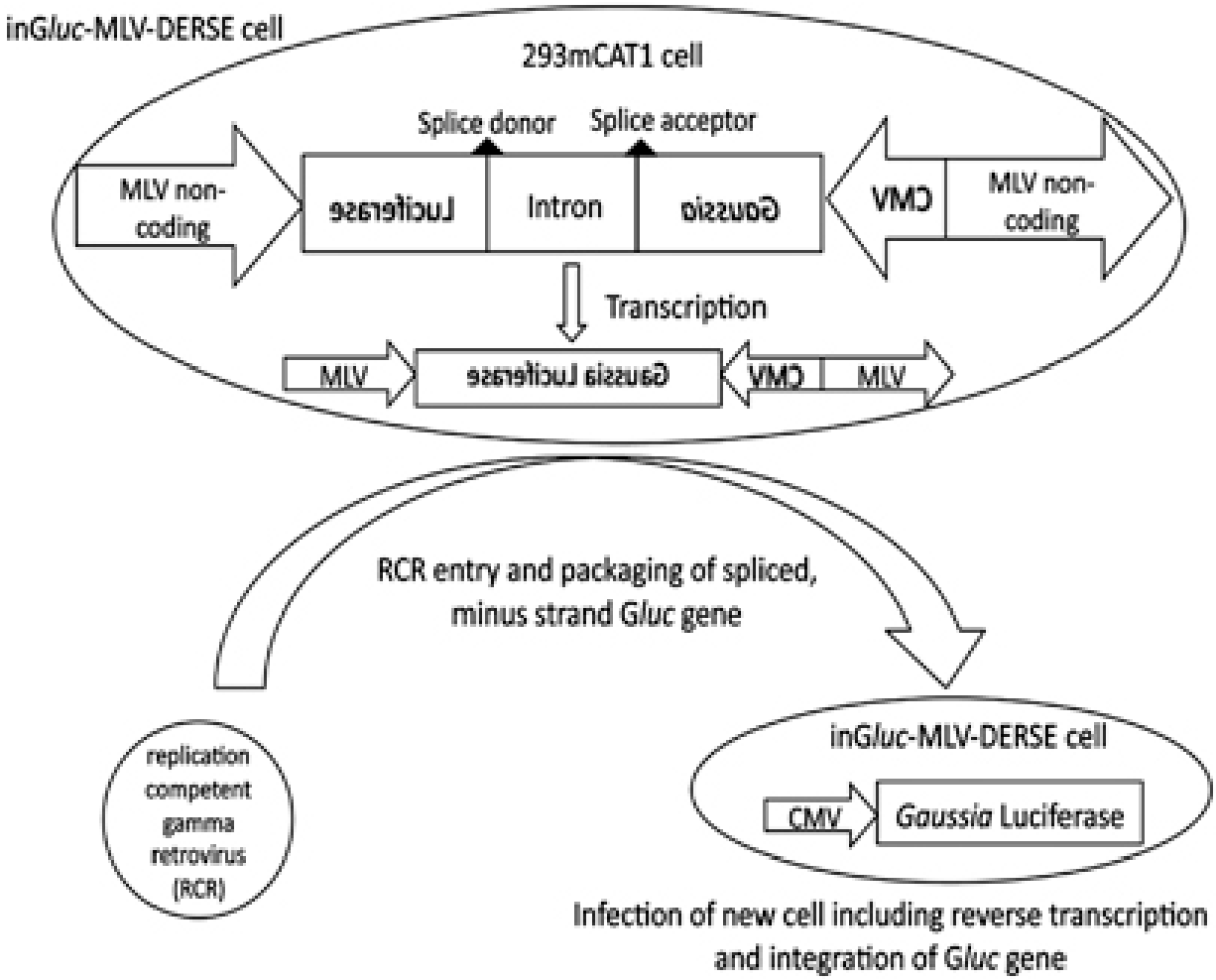


Figure 6-1 The inGluc-MLV-DERSE assay

From Aloia et al., 2013. This assay was used to quantify and compare the relative activities of the two MoMuLV virus preparations used in the insertional mutagenesis screen. 293mCAT1 cells that are permissive to infection with MoMuLV contain the inGluc-MLV-DERSE plasmid which contains a Gaussia luciferase (*Gluc*) sequence in reverse orientation to flanking MoMuLV non-coding sequences. An intron within the *Gluc* sequence can be spliced by the host cell after infection with MoMuLV. Copies of the virus with an intron removed are then capable of producing functional luciferase. The DERSE cell packages the RNA of the minus *Gluc* sequence after initial infection and the next round of infection produces double-stranded DNA with a coding *Gluc* sequence which is expressed by the cell and secreted into the supernatant to be measured by luminometer.

Figure 6-2 shows the results of the 'inGluc-MLV-DERSE assay' (MLV = murine leukaemia virus) to quantify MoMuLV. Five dilutions of the two virus preparations were used to infect inGluc-MLV-DERSE cells and supernatant harvested at five time points to quantify Gluc. In the exponential phase of virus replication, there is consistent trend of preparation 2 showing more activity than preparation 1 which indicates that it is of higher concentration. However, the assay eventually saturates, regardless of the dilution of the virus used which suggests that variation in injected virus titre in experimental mice is not a confounding factor.

Figure 6-3 and Figure 6-4 show the quantification of MoMuLV by qPCR in the cDNA (to compare virus expression levels) and DNA (to compare relative DNA copy number) of spleens from the time course mice. The values showed minimal variation between mice sacrificed at a single time point, showing reproducibility, and both figures show that virus expression and relative copy number initially rise in the first days and weeks after infection but peak and remain stable from around day 14 onwards until disease is detectable. Even when the different genotypes of mice at each time point are separated, the results remain the same (Figure 6-5(b) and (d)).

The one exception to this is at day 9 where there is a difference between mice from the B6 cohort (wild-type C57BL/6 Eμ-*BCL2*) and the F1 cohort (VavP-*BCL2*). We did not have enough samples to determine whether this difference is due to the presence of the VavP-*BCL2* transgene, the F1 background or the batch of virus used. Nonetheless by day 14 it appears the hematopoietic compartment is saturated in all cohorts.

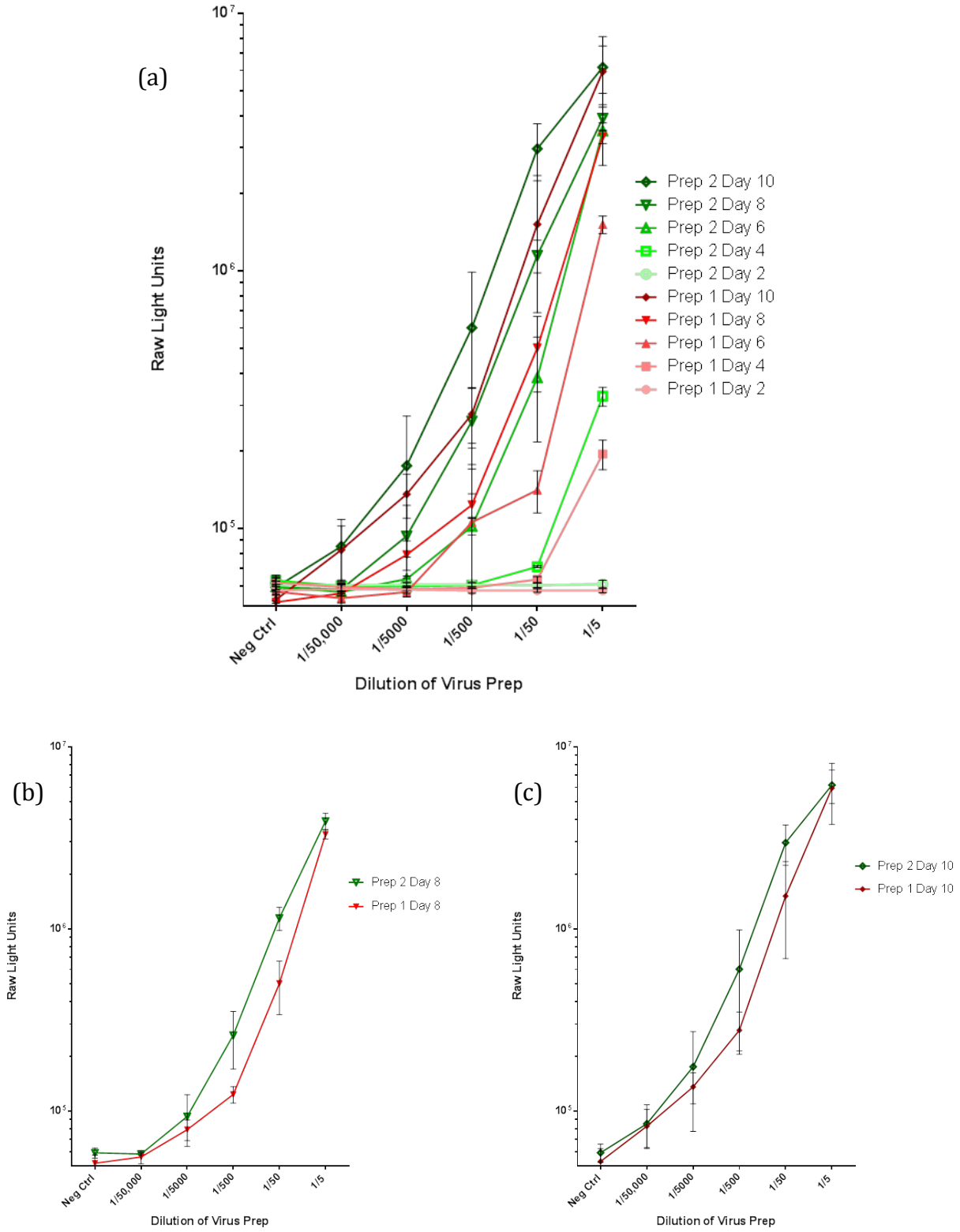


Figure 6-2 MoMuLV quantitation results from in*Gluc*-MLV-DERSE assay

The in*Gluc*-MLV-DERSE assay was used as described in Figure 5-1 to quantify the relative concentrations of the 2 virus preparations. 5 serial dilutions of each prep were used to infect MLV-DERSE cells and supernatant collected for each after 2, 4, 6, 8 and 10 days and Gaussia luciferase activity measured on a luminometer. (a) shows the results of all the above. (b) and (c) show the results of the various dilutions collected at days 8 and 10 respectively. They show that in the exponential phase of virus replication, there is a trend for preparation 2 to show more activity than preparation 1, although in both cases the assay saturates, suggesting that a peak virus concentration is reached regardless of titre.

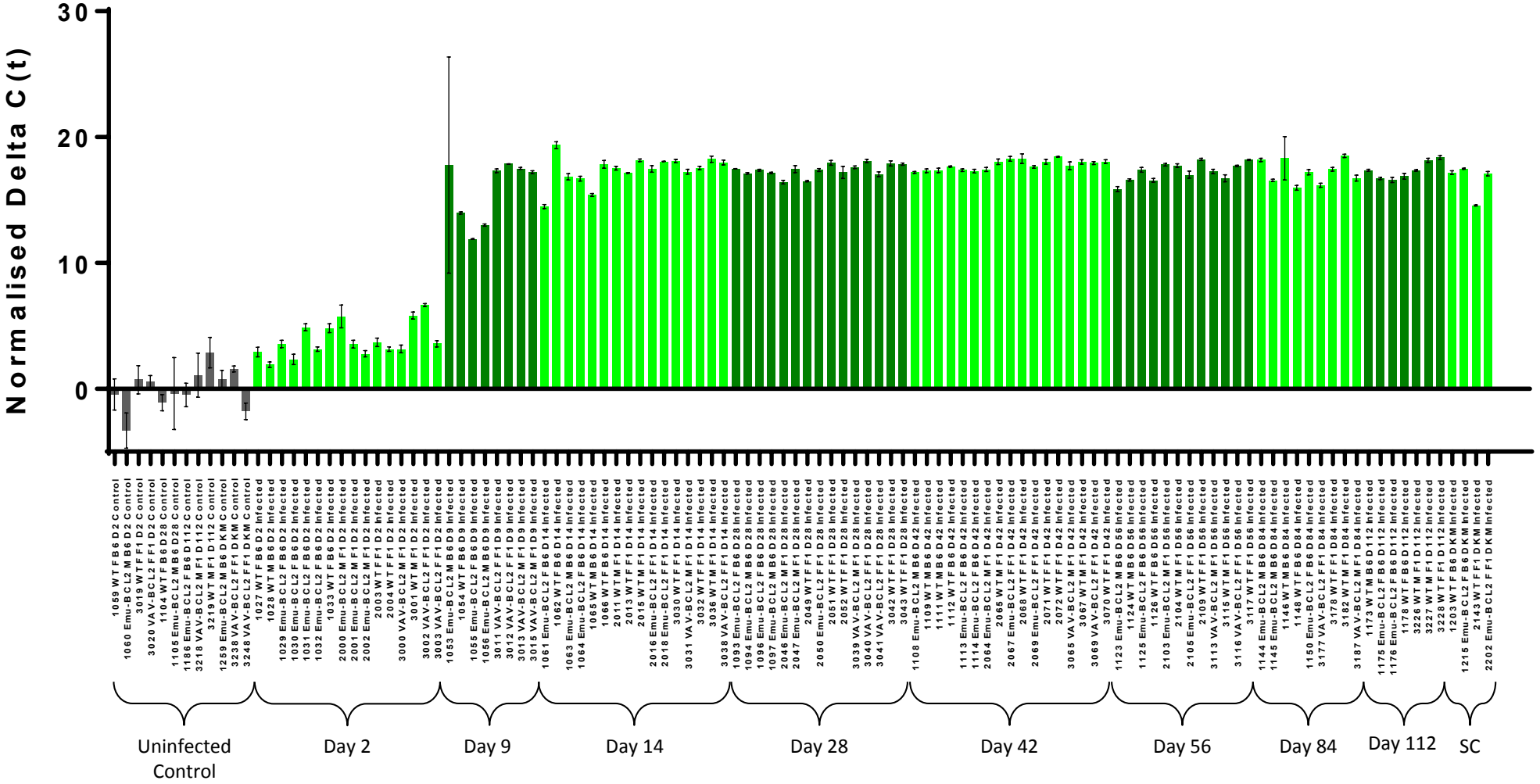


Figure 6-3 qPCR of cDNA from MoMuLV infected mice investigating virus expression levels over time
MoMuLV expression levels over time was studied by performing qPCR on cDNA from time course mice. Each bar represents one mouse. There was minimal inter-sample variation between mice at each time point. Virus expression levels rise over time after inoculation of a new born mouse pup but peak and remain stable from day 14 onwards. SC = survival cohort mice.

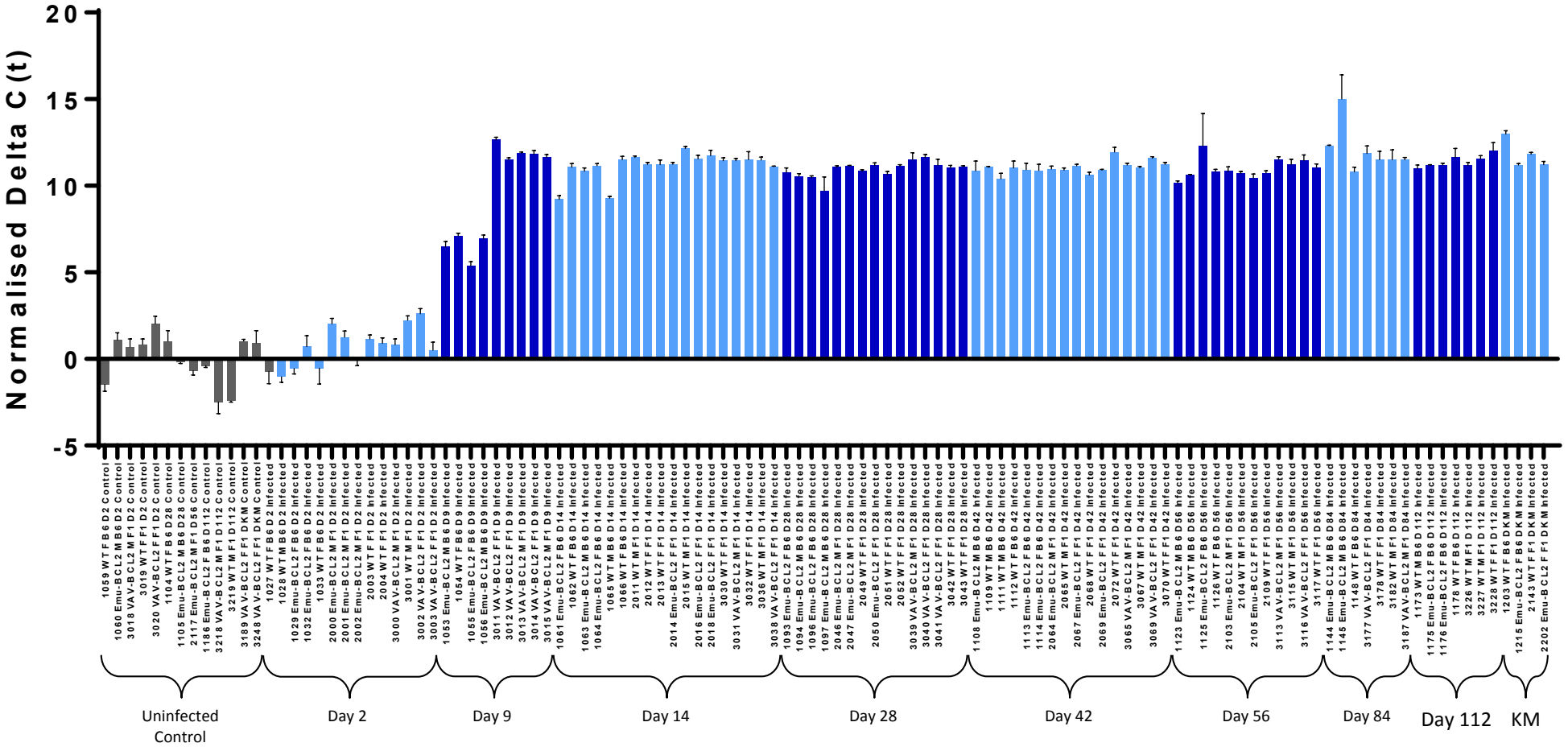
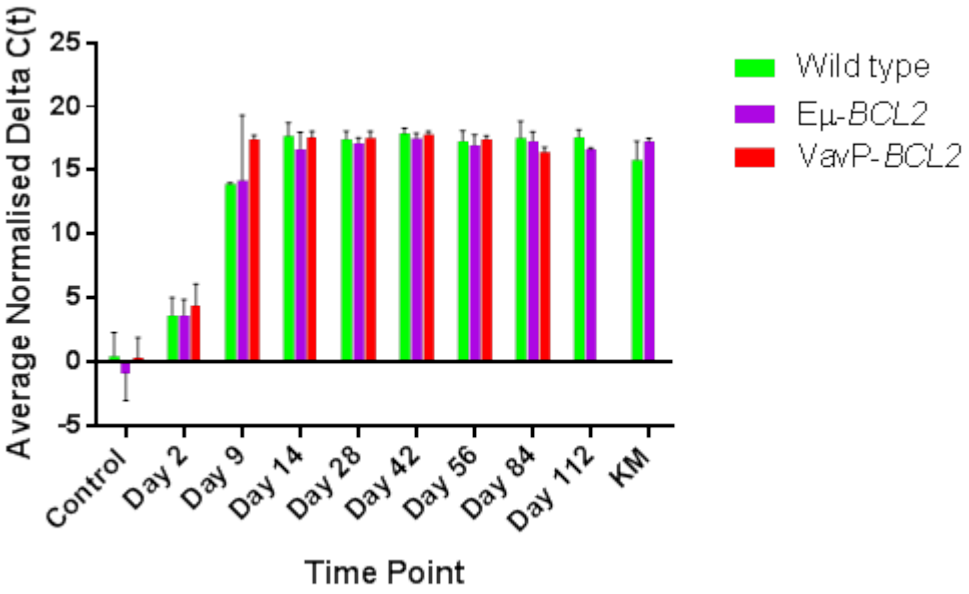


Figure 6-4 qPCR of DNA from MoMuLV infected mice investigating relative virus copy number over time

Relative MoMuLV copy number over time was studied by performing qPCR on DNA from time course mice. Each bar represents one mouse. There was minimal inter-sample variation between mice at each time point. Virus copy number rises over time after inoculation of a new born mouse pup but peaks and remains stable from day 14 onwards.SC = survival cohort mice.

(a)



(b)

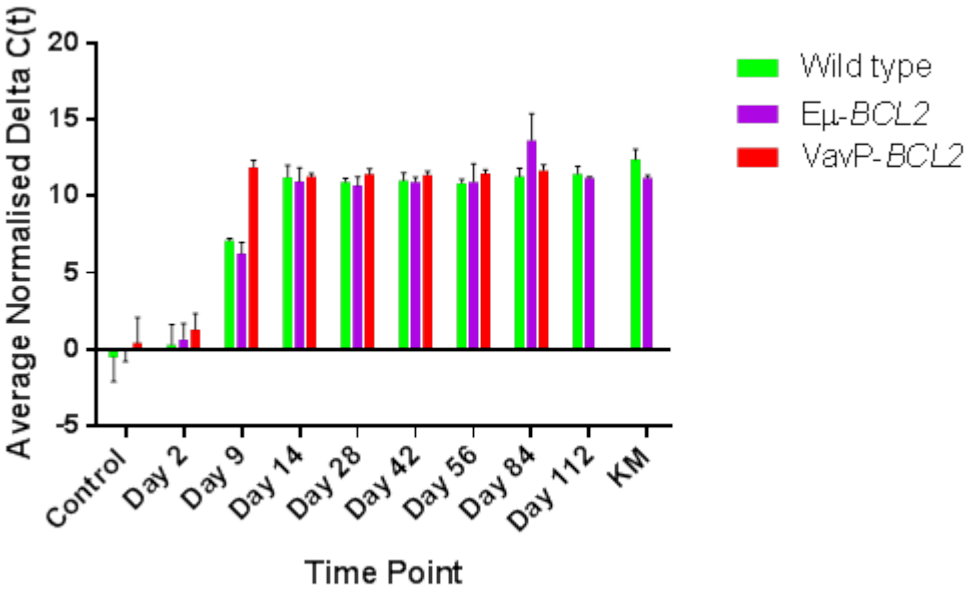


Figure 6-5 MoMuLV expression levels and relative copy number
After inoculation of new born mouse pups with MoMuLV: (a) and (b) show mean virus expression levels and DNA copy number respectively at each time point, separating the data for genotype. Both showing a rise over time and reaching a maximum at 14 days after infection.

6.2 Quantification of insert clonality profiles

The number of insertions identified in each sample varies between dozens to hundreds, each insertion having its own clonal abundance as determined by the total number of unique insertions in that sample. Insertion sites within a sample can then be ranked from the most clonally abundant to the least. Generally libraries from day 2 and day 5 mice had so few insertions that we did not analyse these further. Comparing the ranked clonality profiles of all samples shows that some tumours have a subset of highly clonal insertions (in some cases as few as one) with several much less clonal insertions. These were termed 'most clonal' samples. Other libraries have a flat distribution with all insertions having relatively similar, low clonality, termed 'least clonal'. A spectrum of profiles in between these extremes was also observed termed 'partially clonal' (Figure 6-6).

In comparing the relative clonality of insertions between samples it is worth noting that not all samples of tissue are derived from an equivalent proportion of lymphoma cells i.e. some samples may contain more non lymphoma cells than others. To compensate for this discrepancy between samples, in addition to using raw clonality values, we normalised all clonality values within a sample by setting the most abundant insertion value to 1. This also improves visualisation of libraries where all insertions are of low clonality.

As expected, early time course mice had lower clonality profiles than later time course mice and mice from the survival cohorts. It was also notable that the distribution of clonality between animals in the survival cohorts was quite variable, presumably

because some of these animals succumbed to symptoms that were unrelated to the mutations in the spleen tissue that had been analysed (e.g. an enlarged thymus).

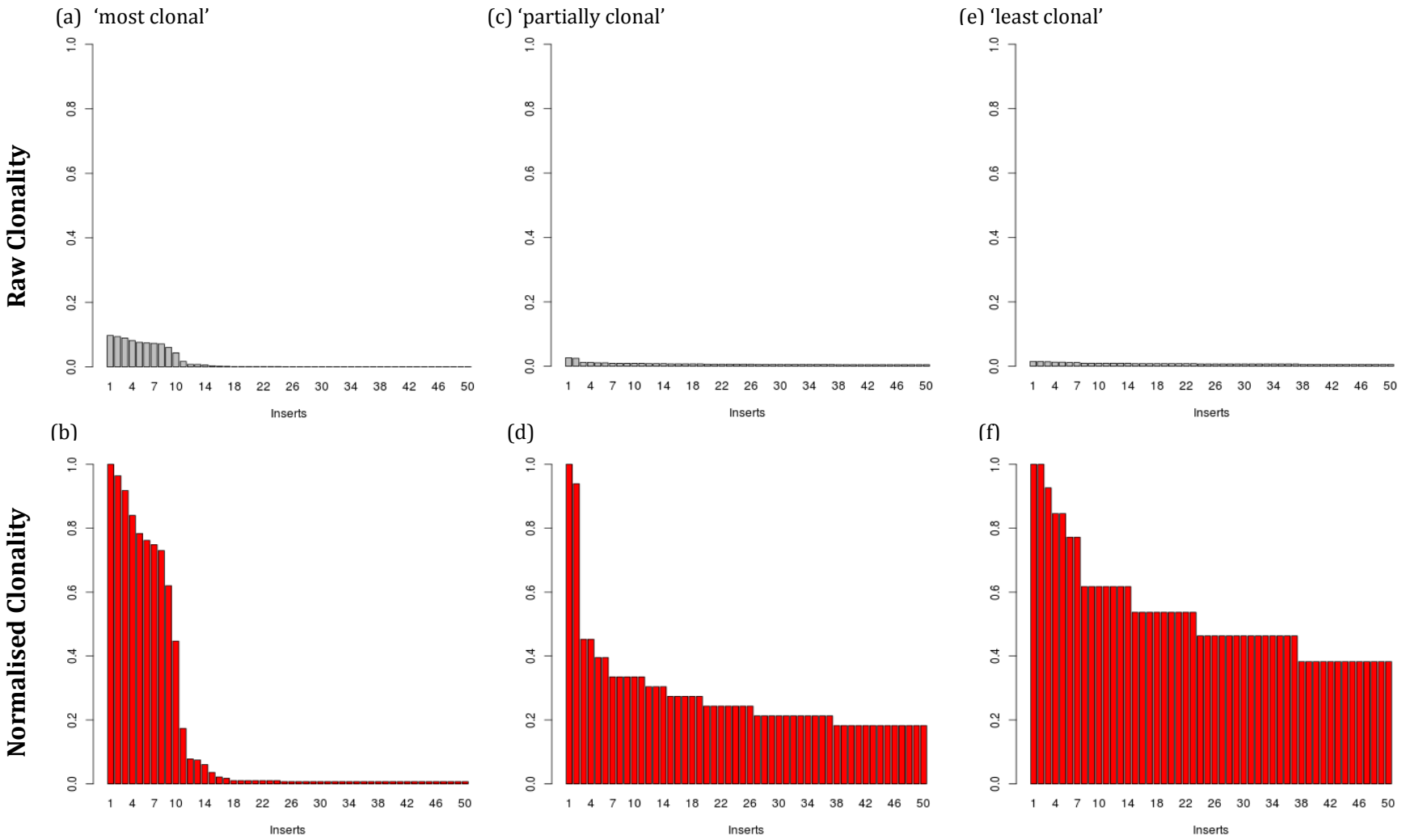


Figure 6-6 Examples of insertion site clonality profiles of processed mouse DNA

The clonal abundance of the top 50 MoMuLV insertion sites were studied for every diseased and time course mouse DNA. (a), (c) and (e) show the raw clonality values of the top 50 insertion sites in three different mice, with (b), (d) and (f) representing the normalised clonality values of the same three mice respectively. (a) shows a mouse with approximately 10 particularly clonally abundant insertions, possibly representing driver mutations, with the rest being subclonal. (c) shows a mouse with 2 just clonal insertions. (e) shows a mouse with no clonally abundant insertions; the most clonally abundant insert in this sample is little more than twice as clonal as the least, see (f).

6.3 Classification of Insertion Profiles

We attempted to classify these profiles by eye. Two independent blinded researchers were able to classify libraries into “low” and “high” clonality with approximately 80% agreement between individuals. We investigated various approaches to quantifying the extent to which the insertion profiles differ and classify them in a less subjective manner. We subtracted the area under the curves of two profiles (similar curves giving low values, different curves giving high values). We also used a method for analysing ordered series of data (“dynamic time warp”, http://dtw.r-forge.r-project.org/, http://www.jstatsoft.org/v31/i07/) which gives a score for how similar two ordered data series are. We compared the profiles using the top 50 and top 100 insertions of each sampleand compared the use of absolute clonality values vs normalised clonality values. Other measures of clonality can also be used. Shannon entropy is a measure of disorder of a series of data which can be used as a value for clonality. Entropy measures cannot be applied to data series of differing length although this can be addressed by truncating all datasets to a constant length (in our data – the top 50 insertions). Another measure that can be applied to data series of differing length is the Gini coefficient (Gini, 1914) and this has been applied as an estimator of oligoclonality in HTLV driven lymphomas (Gillet et al., 2011a).

We used dynamic time warp to createa distance matrix i.e. a grid of values that gives the pairwise scores of similarity/difference between each two samples. Using distance matrices from the above methods we clustered insertion profiles and found that dynamic time warp of the top 50 insertions using normalized clonality gave a classification of insertion profiles into low and high abundance that best matched the

classifications by eye. These clustered insertion profiles were then split into three groups; 'most', 'partial' and 'least' clonally abundant (Figure 6-7). This analysis was performed by Barbara Iadarola (bioinformatician). Table 6-2 shows the distribution of mice from different time points into these three clusters. I found that those mice used for survival analysis (i.e. those that were sacrificed for detectable disease) predominantly had between approximately 1 and 15 highly clonal insertions (the median number of highly clonal insertions was 8 (normalised clonality > 0.1)) and so the majority of survival cohort samples were in the group representing the 'most clonal' cluster. These highly clonal insertions are likely to be the driver mutations causing disease in these cases and should be prioritised for further study and validation. Those samples derived from mice sacrificed at early time points, prior to the onset of detectable disease, contained the lowest clonality insertions and also the least difference in the normalised clonality of their top 50 insertions and so fell in the 'least clonal' cluster. This is likely to be due to the fact that in younger mice, there was less time for the more clonal insertions to outgrow. The older the mouse was, the less likely it was to appear in the lowest clonality profile and the more likely it was to appear in the intermediate or higher clonality profiles. This clearly shows the change in mutation clonality over time from the earliest stages of disease initiation through to fulminant lymphoma. These results show that from a variety of early, low clonal abundance mutations, some are then selected for as being advantageous to lymphomagenesis and increase in clonal abundance over time until the onset of cancer.

Some mice that did develop disease, had low clonal abundance mutations and were in either the partially or least clonal groups. This could be for a number of reasons. Possibly their predominant disease was extra-splenic, or there was polyclonal

enlargement of the spleen with several 'mini-clones'. Possibly the spleen sample used for sequencing contained a high proportion of non-diseased tissue. Another reason could be that the specific gene mutations in these mice were so potent and synergistic that only low clonal abundance was required to cause disease. It could be that when the 'most efficient' combination of mutations co-occur that target the major cancer mechanisms (reducing apoptosis, increasing proliferation, evading anti-growth signals, sustaining angiogenesis etc), they are only required at low clonal abundance to lead to tumour growth. Further analysis is needed to correlate mutation profiles with organ sizes and tumour characterisation to establish this.

It is also worthwhile remembering that cells do not divide at equal rates (or mutate at equal rates). It is therefore expected that there will be variability in the rate that genes become clonal. The randomness of cell division and accumulating mutations means that clonal abundance of mutations would not necessarily demonstrate a clear correlation with time.

Establishing the different clonality profiles of the different mice at different ages allows the study of the most likely driver mutations within a tumour and also to look for the presence of these mutations at pre-clinical stages of disease development.

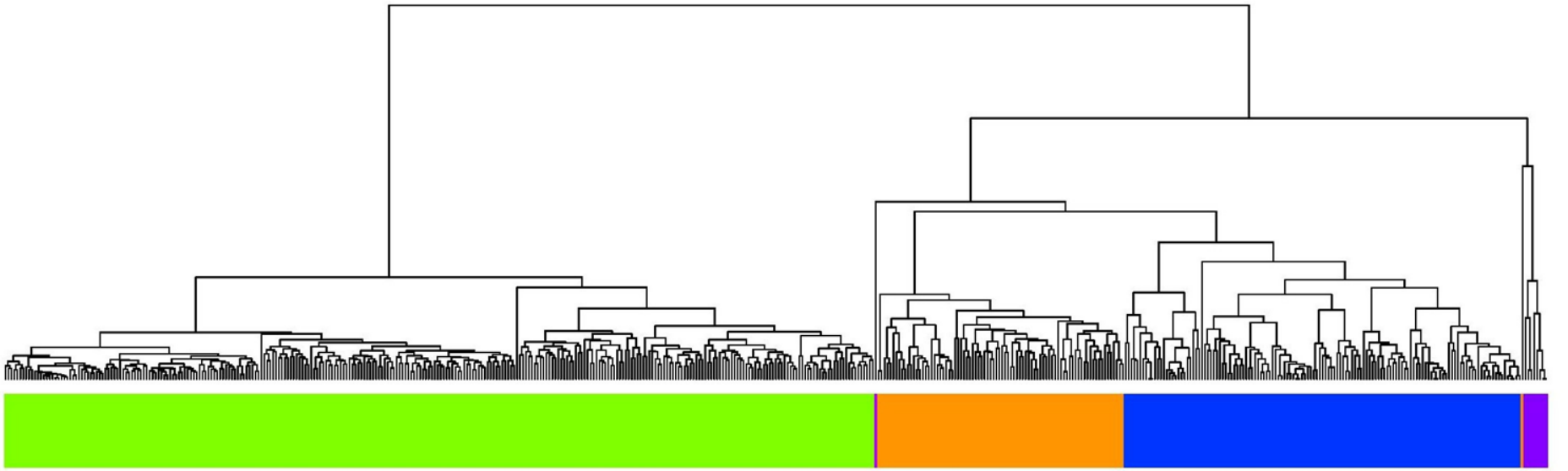
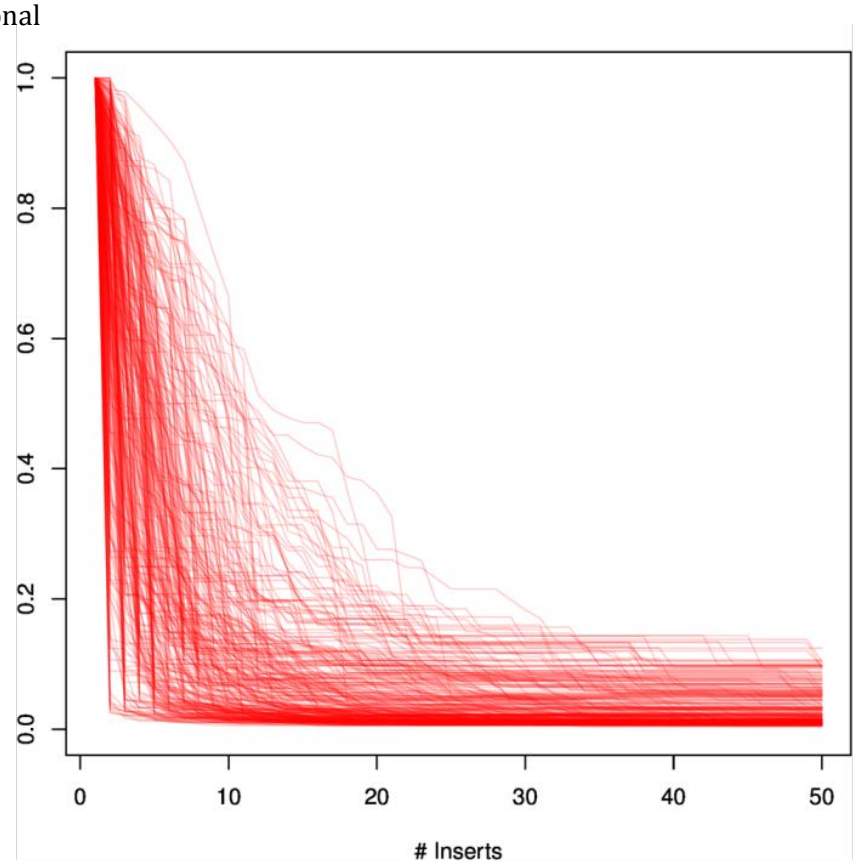


Figure 6-7 Dendogram grouping samples by normalised clonality
The normalised clonality profiles of each sample were clustered. All diseased and time course mice were included, The insertions are ranked by clonal abundance / clonality. To normalise clonality, for each sample the most clonal insertion was set as 1 and all other insertions adjusted relative to this. The difference or "distance" between samples was established by comparing the normalised clonality values of the top 50 insertions using the dynamic time warp algorithm. This resulted in 672 samples being subdivided into 3 main clusters as represented by the 3 large blocks of colour. Green represents those mice in the 'most clonal' cluster, orange represents those in the 'partially clonal' cluster and blue represents those in the 'least clonal' cluster (see Figure 6-6 and Figure 6-8).

(a) Most clonal



(b) Partially clonal

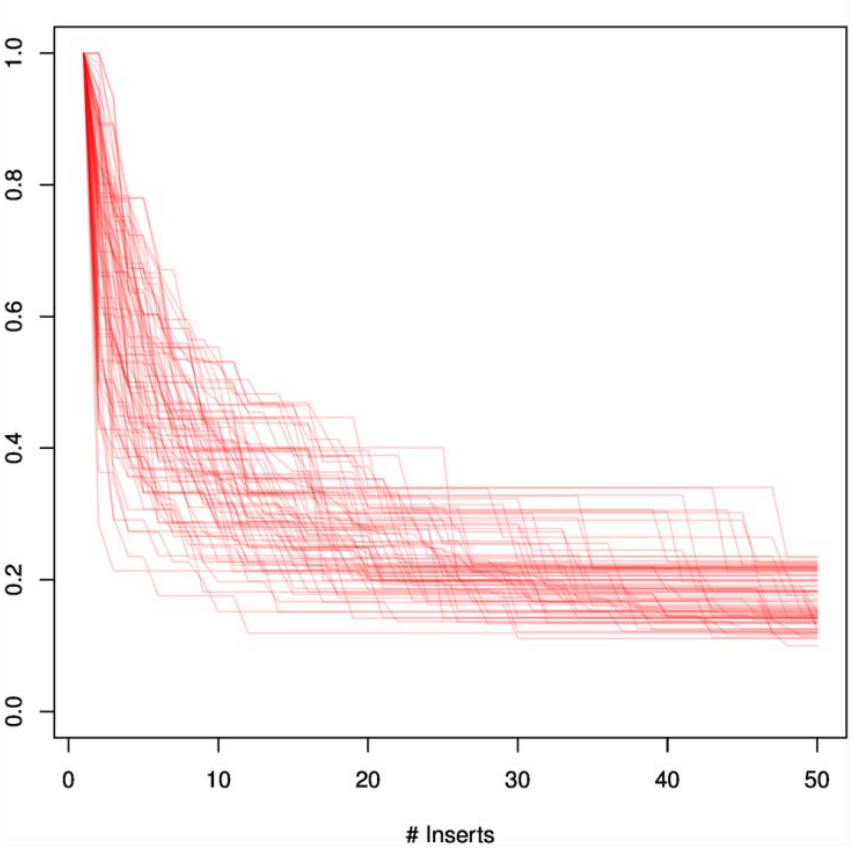


Figure 6-8 Clustering of samples based on clonality of insertions (continued overleaf)

(c) Least clonal

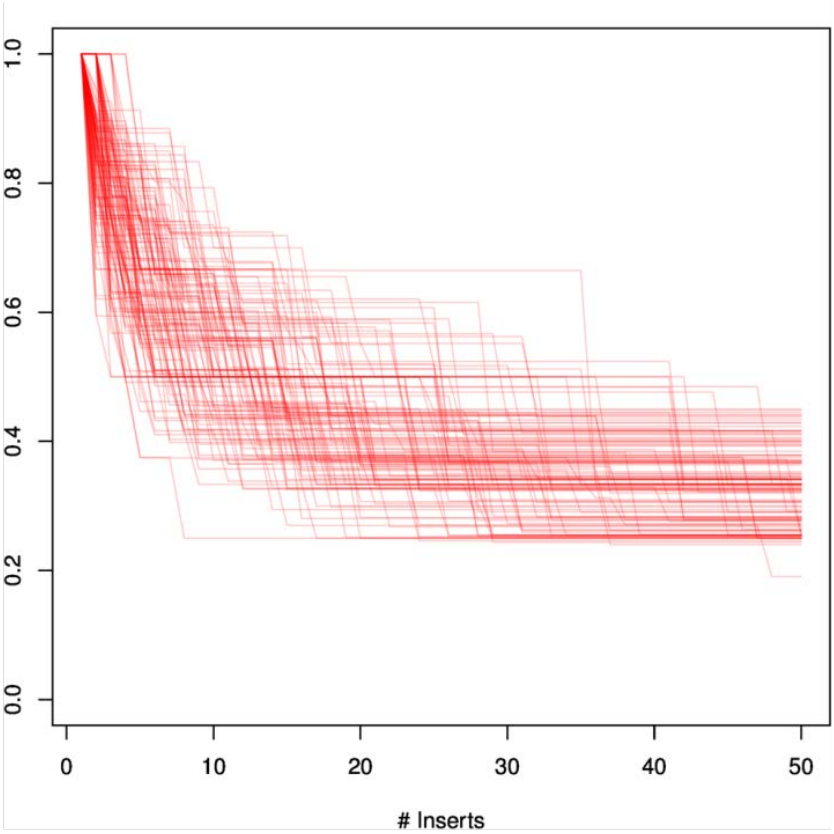


Figure 6-8 Clustering samples based on clonality of insertions
(a), (b) and (c) show plots of the normalised clonality distribution of the top 50 most clonal common insertion sites for every sample in each of the 3 main clusters from Figure 6-6. Each line represents a different mouse and points in the line are the ranked normalised clonality values for inserts within each sample. (a) shows mice where between 5-20 (approx.) insertions were very clonal compared to the remaining insertions. This cluster predominantly consists of mice that developed disease or those maintained until later time points (and so were very close to developing disease). This suggests that those very clonal insertions were driver mutations causing tumour outgrowth. (b) and (c) represent groups of mice with progressively reducing contrast in the relative clonality of insertions, where there was less time for a CIS / group of CISs to allow tumour outgrowth and as expected, mainly contain mice from earlier time points. Clustering the CISs from different mice in this way allows the study of relative insertion clonality over time and also gives clues as to which driver mutations to study in more detail. The number of samples in each cluster, and which time point they were from, is shown in Table 6-2.

Time Point	Wild-type or Transgenic	No. of samples in 'most clonal' cluster	No. of samples in 'partially clonal' cluster	No. of samples in 'least clonal' cluster
D2	Wild-type	1	0	0
D9	Wild-type	0	1	0
D14	Wild-type	0	5	18
D28	Wild-type	0	0	6
D42	Wild-type	1	2	7
D56	Wild-type	1	2	8
D84	Wild-type	4	0	6
D112	Wild-type	5	3	4
SC	Wild-type	154	23	17
D2	Transgenic	0	0	0
D9	Transgenic	0	2	5
D14	Transgenic	0	0	9
D28	Transgenic	0	3	11
D42	Transgenic	1	3	9
D56	Transgenic	1	5	7
D84	Transgenic	7	4	4
D112	Transgenic	1	2	1
SC	Transgenic	118	27	20
D2	Both combined	1	0	0
D9	Both combined	0	3	5
D14	Both combined	0	5	27
D28	Both combined	0	3	17
D42	Both combined	2	5	16
D56	Both combined	2	7	15
D84	Both combined	11	4	10
D112	Both combined	6	5	5
SC	Both combined	272	50	37

Table 6-2 Time point and genotype of mouse samples in each insertion profile cluster

Mice used in the time course experiment. 'Transgenic' refers to the combination of all three cohorts of mice with both *BCL2* transgenes. D = day, SC = Survival Cohort.

6.4 Organ heterogeneity of MoMuLV common insertion sites

In addition to studying the kinetics of mutation onset I also wanted to look at the distribution of mutation onset across different lymphoid organs. In humans, lymphoma can affect any organ / organs and with such deep characterisation of mutation profiles in this study I wanted to see if these profiles differed between different organs in the same mouse and if these patterns differed between different mice.

The lymphoid organs from 12 diseased VavP-*BCL2* transgenic (BALB/c x C57BL/6)F1 mice were all processed by ligation mediated PCR and sequenced on the Illumina HiSeq to look at insertion sites (Table 6-3). Five of these mice developed disease early (mean time to disease 75.2 days) and seven developed disease late (mean time to death 169.1 days). They show diverse presentations of organomegaly at time of death and represent both female and male, wild-type and transgenic mice. The top 50 insertions in all lymphoid organs were compared within each mouse (see Figure 6-9(a) – (l)). These 12 plots confirm the wide heterogeneity of lymphoma, showing diverse patterns of dissemination between tissues. (c) and (h) both show similar insertion profiles across all lymphoid organs. (a) and (e) show that all lymphoid organs can have mutations in the most clonally abundant insertions (approximately 5-10) but then the spleen has a host of other clonal insertions that are not seen in other organs. In addition, the lymph nodes in the right side of mouse (a) have clonal insertions that are similar to each other but not found on lymph nodes on the left or in other organs. (b), (d) and (k) show that different lymphoid organs can have distinctly different insertion profiles. (j) is an example of mouse that can have both splenomegaly and an enlarged thymus but have very few driver mutations.

These patterns may be useful to deduce the history of each lymphoma regarding the origin and where it metastasises. It also shows how polyclonal this disease can be. There is much more analysis that can be done with this and it would be interesting to look at larger numbers of biopsies from each tissue of a single animal and also single cell analysis.

These initial results could have significant implications for human disease. This organ heterogeneity may mean that sequencing of a single biopsy in order to risk stratify disease prognosis and deliver gene targeted therapy may be inadequate if different polyclones of disease exist in different organs. Analysis of 100 single cells in two human breast tumours (one monogenomic with a liver metastasis and one polygenomic tumour) showed distinct clonal populations of cells with one clone forming the metastasis. Both tumours showed subpopulations of genetically heterogeneous cells that did not metastasize (Navin et al., 2011).

Mouse ID	Wildtype or Transgenic?	Sex	Age at death (days)	Splenomegaly?	Thymus Enlargement?	Lymphadenopathy?
a	T	F	80	Yes	No	No
b	T	M	50	Yes	No	No
c	T	F	80	Yes	No	No
d	W	M	86	Yes	Yes	No
e	T	F	80	Yes	No	No
f	W	M	137	No	No	No
g	W	F	157	Yes	Yes	Yes
h	W	M	137	Yes	Yes	No
i	T	M	194	Yes	Yes	No
j	W	M	194	No	Yes	No
k	T	F	173	Yes	No	No
l	W	M	192	Yes	No	No

Table 6-3 Characteristics of mice used to study organ heterogeneity
T = Transgenic, W = Wild-type, F = Female, M = Male.

Figure 6-9a & b - Organ heterogeneity of common insertion sites
The raw clonality values of the 50 most abundant insertions across different lymphoid organs within a diseased mouse. Each plot represents an individual mouse. Each row represents an insertion. LN = lymph node.

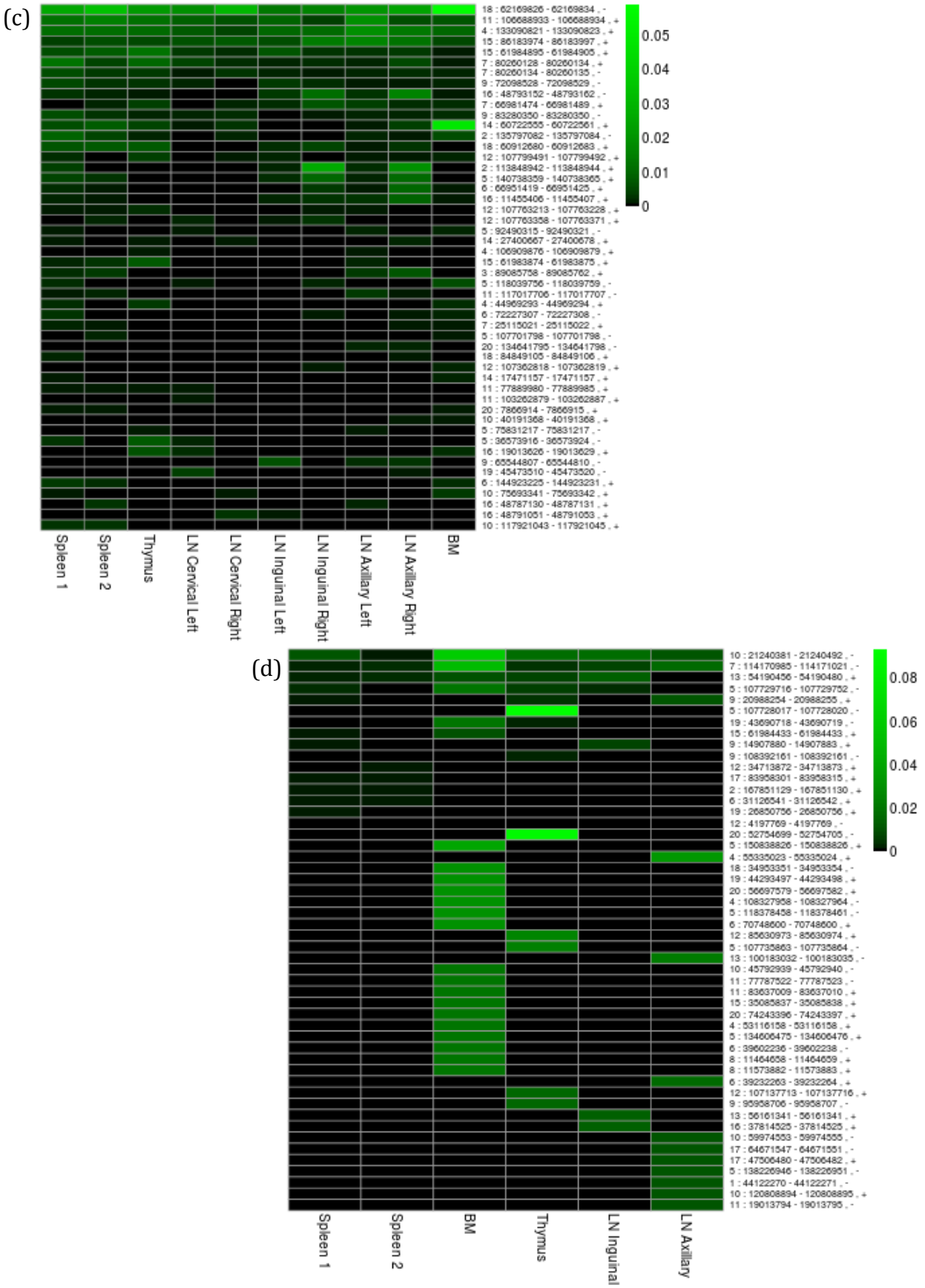
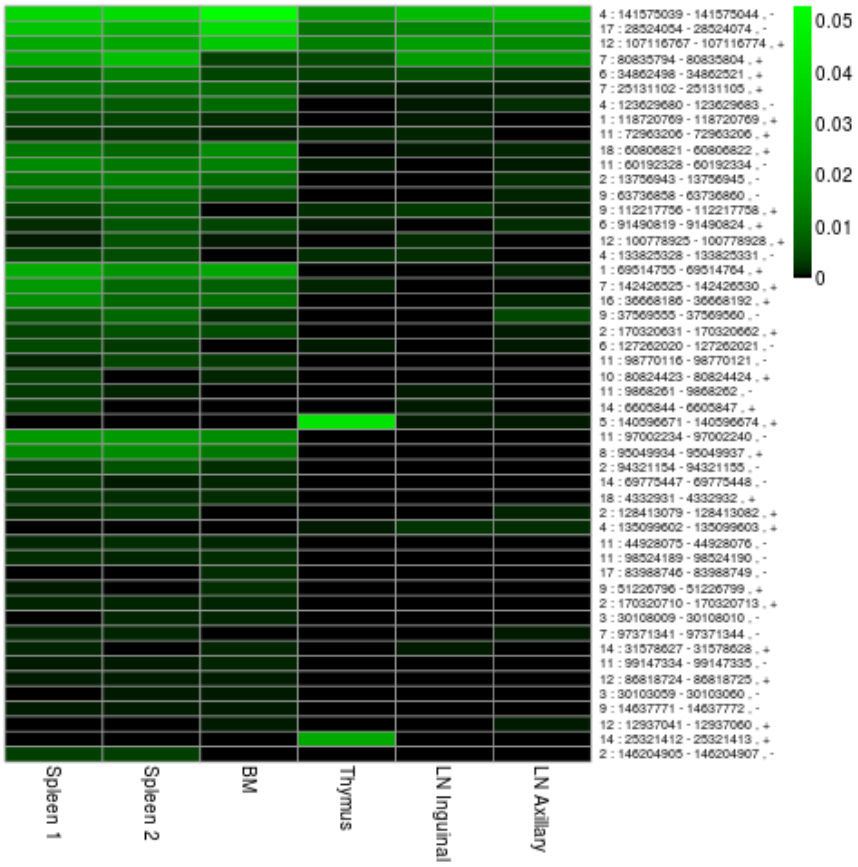


Figure 6-9c & d - Organ heterogeneity of common insertion sites
The raw clonality values of the 50 most abundant insertions across different lymphoid organs within a diseased mouse. Each plot represents an individual mouse. Each row represents an insertion. LN = lymph node.

(e)



(f)

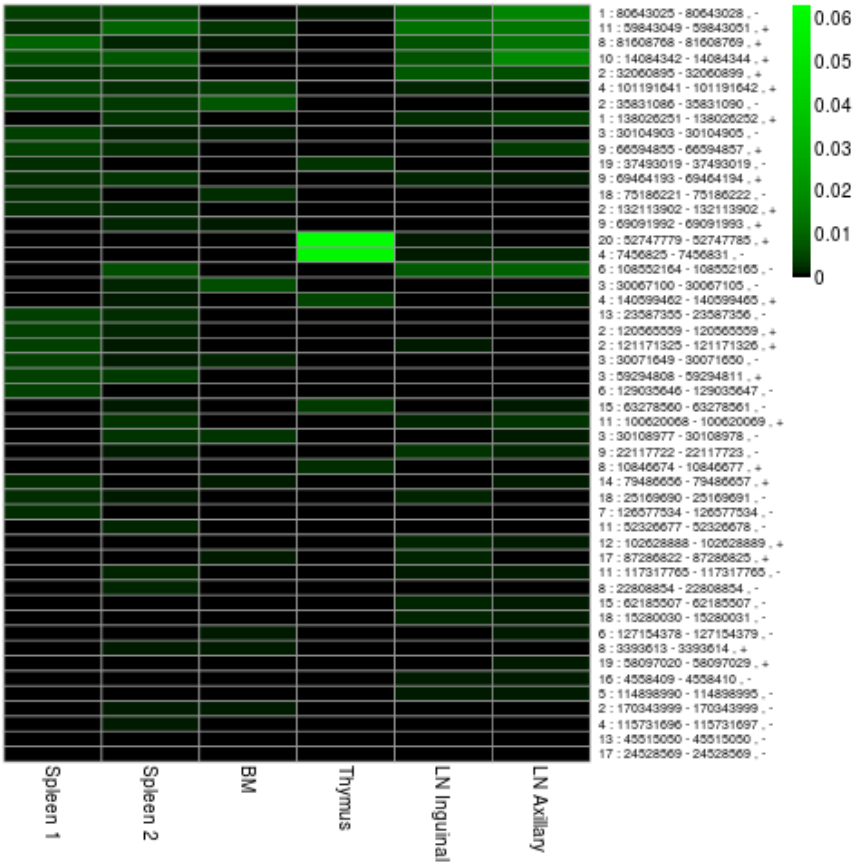
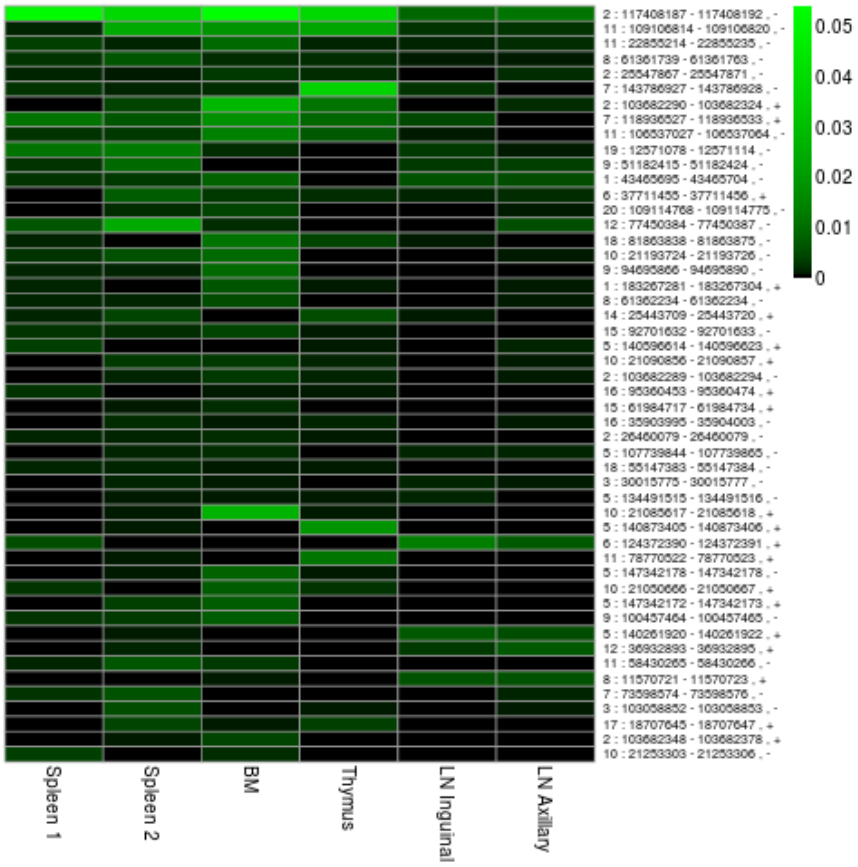


Figure 6-9e & f - Organ heterogeneity of common insertion sites
The raw clonality values of the 50 most abundant insertions across different lymphoid organs within a diseased mouse. Each plot represents an individual mouse. Each row represents an insertion. LN = lymph node.

(g)



(h)

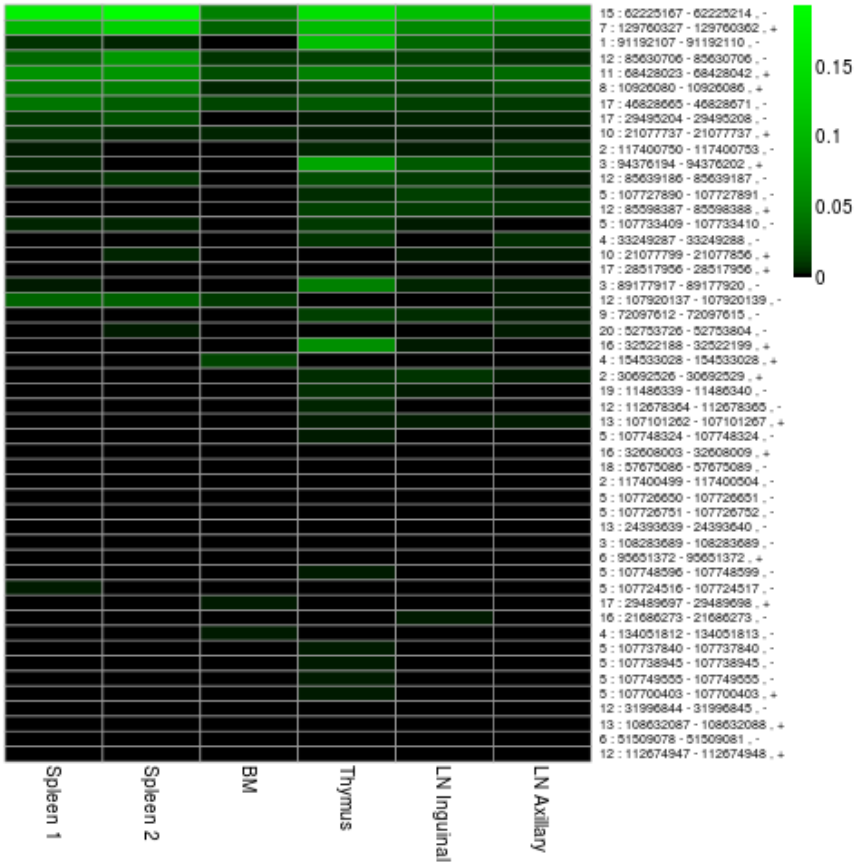
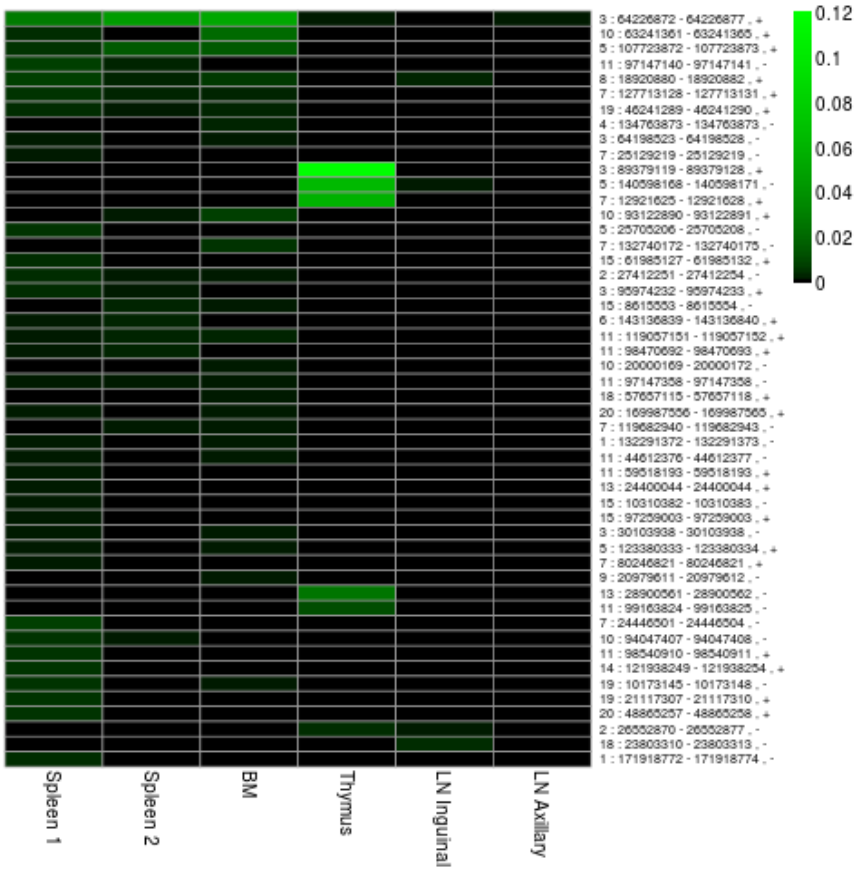


Figure 6-9g & h - Organ heterogeneity of common insertion sites
The raw clonality values of the 50 most abundant insertions across different lymphoid organs within a diseased mouse. Each plot represents an individual mouse. Each row represents an insertion. LN = lymph node.

(i)



(j)

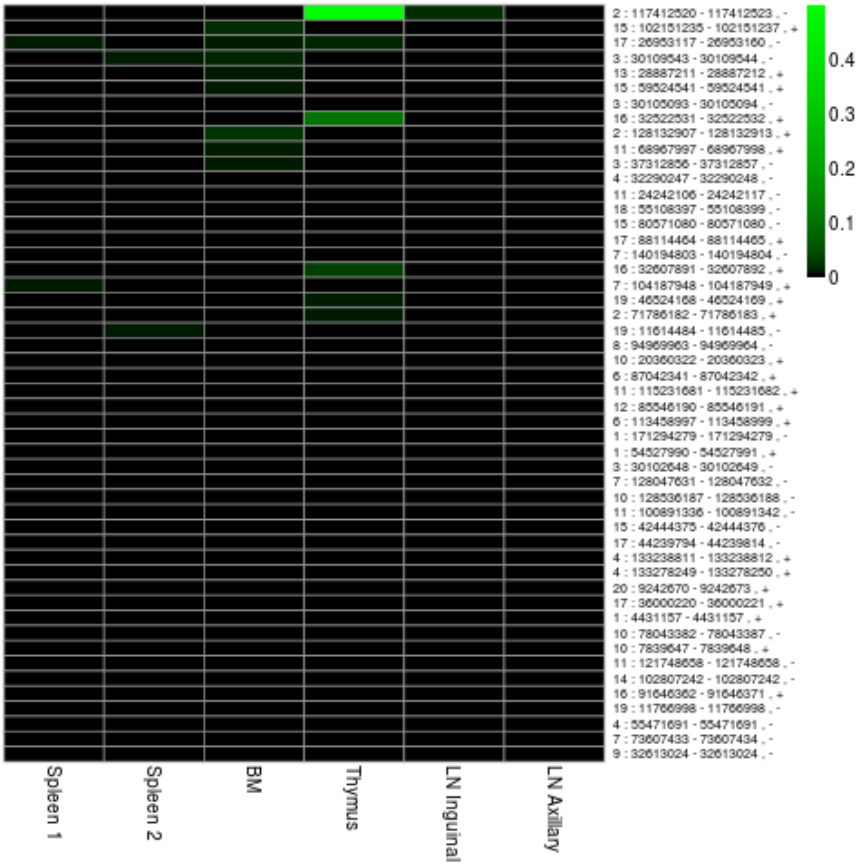
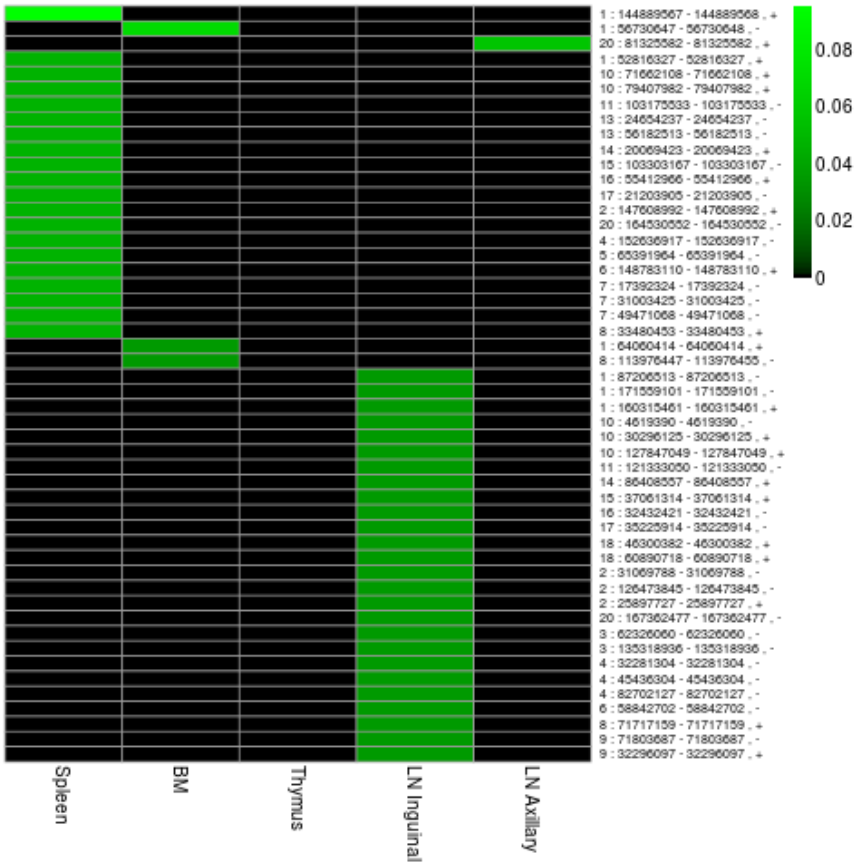


Figure 6-9i & j - Organ heterogeneity of common insertion sites
The raw clonality values of the 50 most abundant insertions across different lymphoid organs within a diseased mouse. Each plot represents an individual mouse. Each row represents an insertion. LN = lymph node.

(k)



(l)

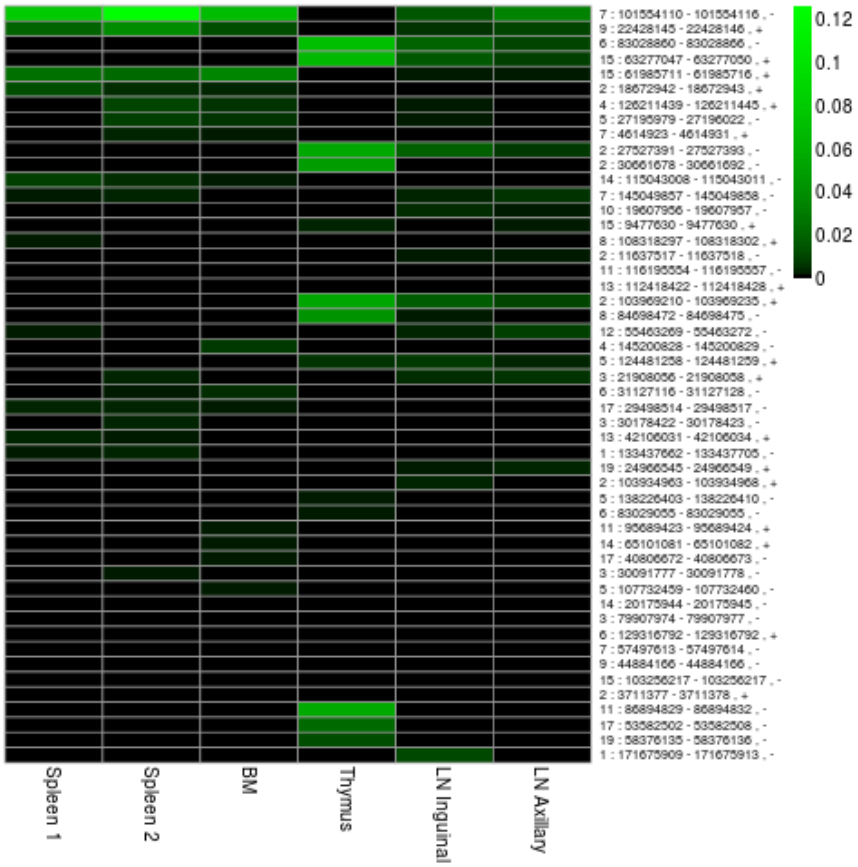


Figure 6-9k & l - Organ heterogeneity of common insertion sites
The raw clonality values of the 50 most abundant insertions across different lymphoid organs within a diseased mouse. Each plot represents an individual mouse. Each row represents an insertion. LN = lymph node.

CHAPTER 7 RESULTS & DISCUSSION: GENOTYPE SPECIFICTY & CANDIDATE GENE VALIDATION

7.1 *BCL2* co-occurring genes

Table 7-1 & Table 7-2 show lists of genes at the most clonally abundant common insertion sites found mutated significantly more frequently in transgenic mice than in wild-types (ie. *BCL2* co-occurring genes) as determined by GKC and KC-RBM respectively. I purport that these genes are oncogenes or tumour suppressor genes selected to act in synergy with *BCL2* and may need drug targeting alongside BCL2 therapies in t(14;18) driven lymphomas.

Gene ontology (GO) analysis of these lists identified biological processes that affect B-cell activation and differentiation, as well as those that influence transcription and lymphoid organ development (Table 7-3). This is likely to be due to the fact that the lymphomas were BCL2 dependent in the transgenic mice whereas in the WT mice different mutations predominated.

The most significant CIS in both tables is *Pou2f2* which is a known oncogene and discussed in more detail in section 7.1.1.

Rank	Chromo-some	Base Position	Gene Name	Ensembl Gene ID	No. of insertions in wild-type mice	No. of insertions in *BCL2* transgenic mice	All other insertions in wild-type mice	All other insertions in *BCL2* transgenic mice	p-value (two-tailed Fisher's Exact)
1	7	25135338	**Pou2f2**	ENSMUSG00000008496	5	32	7604	6686	0.0000
2	16	36689096	**Ildr1, Iqcb1**	ENSMUSG00000022900, ENSMUSG00000022837	0	13	7609	6705	0.0001
3	11	98489843	**Ikzf3**	ENSMUSG00000018168	4	15	7605	6703	0.0054
4	13	56159611	**H2afy, Tifab**	ENSMUSG00000015937, ENSMUSG00000049625	1	9	7608	6709	0.0081
5	20	134585268	**Btk**	ENSMUSG00000031264	0	6	7609	6712	0.0106
6	2	11242603	**Prkcq**	ENSMUSG00000026778	0	6	7609	6712	0.0106
7	11	102209465	**Hdac5**	ENSMUSG00000008855	0	6	7609	6712	0.0106
8	5	123064433	**Gm26745**	ENSMUSG00000097213	1	8	7608	6710	0.0156
9	10	62507791	**Supv3l1**	ENSMUSG00000020079	1	8	7608	6710	0.0156
10	19	41377270	**Pik3ap1**	ENSMUSG00000025017	0	5	7609	6713	0.0227
11	19	24677390	**Pgm5**	ENSMUSG00000041731	0	5	7609	6713	0.0227
12	17	71202339	**Gm26561**	ENSMUSG00000097625	0	5	7609	6713	0.0227
13	5	65501293	**Ube2k**	ENSMUSG00000029203	0	5	7609	6713	0.0227
14	2	132032001	**Rassf2**	ENSMUSG00000027339	0	5	7609	6713	0.0227
15	4	59332394	**Gm12596**	ENSMUSG00000086952	1	7	7608	6711	0.0298
16	1	181228848	**Nvl, Cnih3**	ENSMUSG00000026516, ENSMUSG00000026514	2	9	7607	6709	0.0303
17	20	113882000	**SNORA17**	ENSMUSG00000087765	0	4	7609	6714	0.0483
18	16	44999214	**Ccdc80**	ENSMUSG00000022665	0	4	7609	6714	0.0483
19	9	72416533	**CT954326.1**	ENSMUSG00000097211	0	4	7609	6714	0.0483
20	14	113303747	**Tpm3-rs7**	ENSMUSG00000058126	0	4	7609	6714	0.0483
21	14	10450561	**Fhit**	ENSMUSG00000060579	0	4	7609	6714	0.0483
22	1	58716420	**Cflar**	ENSMUSG00000026031	0	4	7609	6714	0.0483
23	8	117507392	**Plcg2**	ENSMUSG00000034330	0	4	7609	6714	0.0483
24	1	88237487	**Trpm8**	ENSMUSG00000036251	0	4	7609	6714	0.0483
25	5	23674626	**Srpk2**	ENSMUSG00000062604	0	4	7609	6714	0.0483
26	1	179830996	**Ahctf1, Cdc42bpa**	ENSMUSG00000026491, ENSMUSG00000026490	0	4	7609	6714	0.0483
27	1	138599254	**Nek7**	ENSMUSG00000026393	0	4	7609	6714	0.0483
28	11	44892142	**Ebf1, Gm12159**	ENSMUSG00000057098, ENSMUSG00000084773	0	4	7609	6714	0.0483
29	2	119710773	**Rtf1**	ENSMUSG00000027304	0	4	7609	6714	0.0483
30	2	23041828	**Apbb1ip, Yme1l1**	ENSMUSG00000026786, ENSMUSG00000026775	0	4	7609	6714	0.0483

Table 7-1 *BCL2* specific common insertion sites determined by Gaussian Kernel Convolution

Significant *BCL2* specific common insertion site genes from infected mice in the 'most clonal' group (see Figure 6-8) based on normalised clonality (above 0.05) with a 100, 000bp window. Derived by 'CIMPL' which uses Gaussian kernel convolution values to rank the insertions.

Rank	Gene Name	Ensembl Gene ID	No. of insertions in *BCL2* transgenic mice	No. of insertions in wild-type mice	All other insertions in *BCL2* transgenic mice	All other insertions in wild-type mice	2-tailed p-value
1	**Pou2f2**	ENSMUSG00000008496	35	12	6683	7597	0.0002
2	**Iqcb1**	ENSMUSG00000022837	9	1	6709	7608	0.0081
3	**Ube2j2**	ENSMUSG00000023286	10	2	6708	7607	0.0170
4	**Smad7**	ENSMUSG00000025880	10	2	6708	7607	0.0170
5	**Ikzf3**	ENSMUSG00000018168	19	8	6699	7601	0.0191
6	**Cep97**	ENSMUSG00000022604	11	3	6707	7606	0.0284
7	**Fcgr2b**	ENSMUSG00000026656	11	3	6707	7606	0.0284
8	**Etfb**	ENSMUSG00000004610	7	1	6711	7608	0.0298
9	**Cipc**	ENSMUSG00000034157	7	1	6711	7608	0.0298
10	**Mogat2**	ENSMUSG00000052396	7	1	6711	7608	0.0298
11	**Sh3bp1**	ENSMUSG00000022436	13	5	6705	7604	0.0347
12	**Supv3l1**	ENSMUSG00000020079	12	4	6706	7605	0.0414
13	**Lgmn**	ENSMUSG00000021190	10	3	6708	7606	0.0477
14	**Il21r**	ENSMUSG00000030745	10	3	6708	7606	0.0477

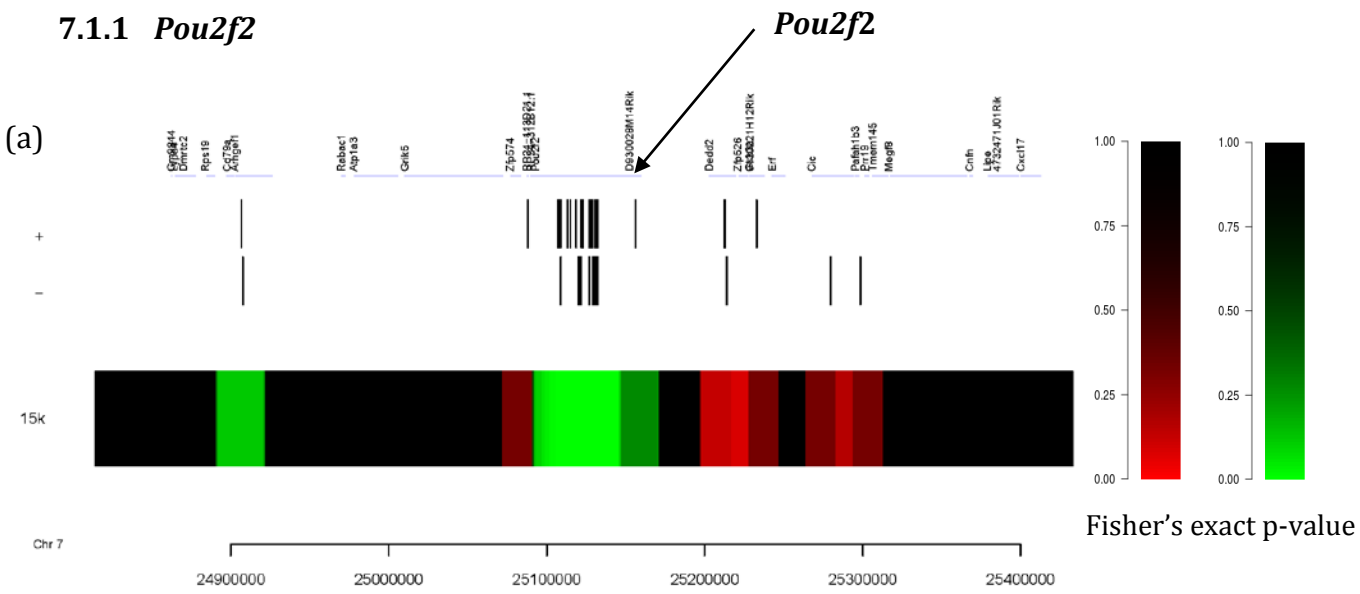
Table 7-2 *BCL2* specific common insertion sites determined by Kernel Convolved Rules Based Mapping (KC-RBM)

GO Term	Count	%	Genes	PValue	Benjamini corrected PValue	False Discovery Rate
GO:0030183~B-cell differentiation	3	7.692307692	HDAC5, PLCG2, POU2F2	0.0043	0.8514	5.9106
GO:0065003~macromolecular complex assembly	5	12.82051282	SRPK2, TRPM8, RTF1, AHCTF1, H2AFY	0.0055	0.7029	7.4638
GO:0043933~macromolecular complex subunit organization	5	12.82051282	SRPK2, TRPM8, RTF1, AHCTF1, H2AFY	0.0073	0.6605	9.8377
GO:0006468~protein amino acid phosphorylation	6	15.38461538	PRKCQ, SRPK2, SMAD7, CDC42BPA, NEK7, BTK	0.0106	0.6924	13.9864
GO:0045321~leukocyte activation	4	10.25641026	HDAC5, FCGR2B, PLCG2, POU2F2	0.0111	0.6274	14.5925
GO:0042113~B-cell activation	3	7.692307692	HDAC5, PLCG2, POU2F2	0.0119	0.5883	15.6445
GO:0002335~mature B-cell differentiation	2	5.128205128	PLCG2, POU2F2	0.0148	0.6120	19.0830
GO:0001775~cell activation	4	10.25641026	HDAC5, FCGR2B, PLCG2, POU2F2	0.0151	0.5703	19.4194
GO:0030097~hemopoiesis	4	10.25641026	HDAC5, PLCG2, POU2F2, AHCTF1	0.0160	0.5472	20.3776
GO:0016310~phosphorylation	6	15.38461538	PRKCQ, SRPK2, SMAD7, CDC42BPA, NEK7, BTK	0.0168	0.5275	21.3045
GO:0006508~proteolysis	7	17.94871795	PRKCQ, CFLAR, PGM5, UBE2K, LGMN, YME1L1, UBE2J2	0.0200	0.5564	24.8528
GO:0048534~hemopoietic or lymphoid organ development	4	10.25641026	HDAC5, PLCG2, POU2F2, AHCTF1	0.0215	0.5517	26.4835
GO:0002520~immune system development	4	10.25641026	HDAC5, PLCG2, POU2F2, AHCTF1	0.0244	0.5691	29.5080
GO:0030098~lymphocyte differentiation	3	7.692307692	HDAC5, PLCG2, POU2F2	0.0244	0.5430	29.5517
GO:0006796~phosphate metabolic process	6	15.38461538	PRKCQ, SRPK2, SMAD7, CDC42BPA, NEK7, BTK	0.0345	0.6450	39.1233
GO:0006793~phosphorus metabolic process	6	15.38461538	PRKCQ, SRPK2, SMAD7, CDC42BPA, NEK7, BTK	0.0345	0.6450	39.1233
GO:0002521~leukocyte differentiation	3	7.692307692	HDAC5, PLCG2, POU2F2	0.0366	0.6441	41.0321
GO:0007242~intracellular signaling cascade	6	15.38461538	PRKCQ, SRPK2, CNIH3, PLCG2, CDC42BPA, BTK	0.0422	0.6749	45.6874
GO:0046649~lymphocyte activation	3	7.692307692	HDAC5, PLCG2, POU2F2	0.0623	0.7946	59.7612
GO:0034622~cellular macromolecular complex assembly	3	7.692307692	SRPK2, AHCTF1, H2AFY	0.0778	0.8486	68.2186

Table 7-3 Gene ontology of *BCL2* specific common insertion sites found in insertional mutagenesis screen

The common insertion sites from the insertional mutagenesis screen identified by both Gaussian Kernel Convolution and Kernel Convolved Rules Based Mapping were input to DAVID Bioinformatics Resources 6.7, using the 'GO Fat' database to functionally annotate groups of genes (Huang et al., 2009a, 2009b).

7.1.1 *Pou2f2*



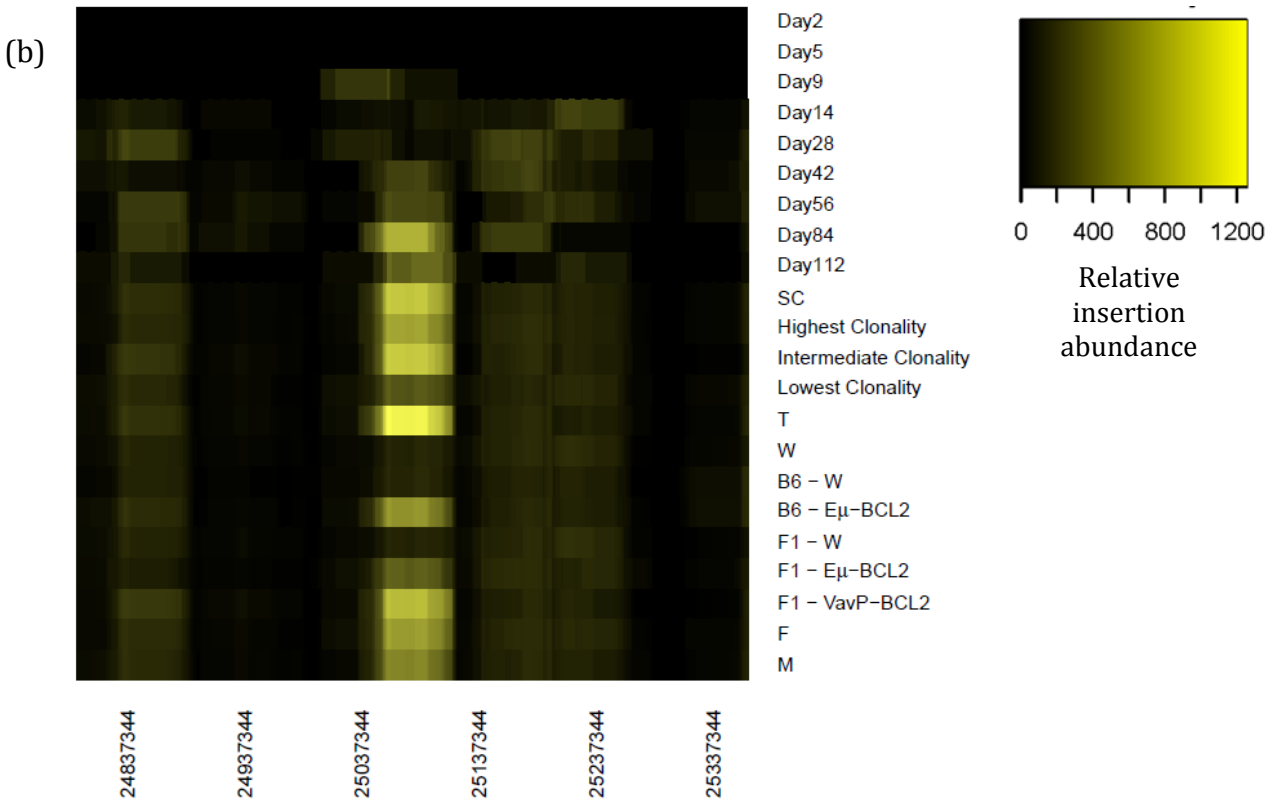


Figure 7-1 Genotype specificity and kinetics of Pou2f2 insertions
(a) A sliding window across the genome showing the MoMuLV insertions of all samples, measuring the genotype specificity by Fisher's Exact test. Green represents inserts biased towards transgenic mice and red towards wild-type mice. (b) The relative abundance of insertions at *Pou2f2* in different cohorts of mice. Calculated as: (number of inserts in a 50,000bp sliding window / total insertions in that cohort) x10^6. SC = survival cohort, F = female, M = male, T = transgenic, W = wild-type, B6 = C57BL/6, F1 = (BALB/c x C57BL/6) F1, + = forward strand, - = reverse strand.

Pou2f2 (also known as *Oct2*) is the most genotype specific CIS in this insertional mutagenesis screen occurring significantly more frequently in transgenic mice overexpressing BCL2 than in wild-type mice ($p=2.22x10^{-16}$, see Figure 7-1(a)). It belongs to a family of transcription factors containing the bipartite POU-homeodomain which contain POU-specific and POU-homeo subdomains that interact with DNA via helix-turn-helix motifs (Clerc, Corcoran, LeBowitz, Baltimore, & Sharp, 1988). *Oct2*$^{-/-}$ mice have functionally deficient B-cell compartments (Hasbold, Corcoran, Tarlinton, Tangye, & Hodgkin, 2004) and *Oct2* has been found to assist differentiation of activated B-cells into antibody secreting plasma cells. It has already been implicated in lymphoma and was found to directly regulate cell survival in t(14;18) driven lymphomas by inducing *BCL2* promoter activity (Heckman, Duan, Garcia, & Boxer, 2006). Missense mutations and mono-allelic frame shift mutations in *OCT2* were found in 8% of patients in a study of 114 FL cases although with disparate suggestions on its mechanism of lymphomagenesis as they concluded that most *Oct2* mutants were hypomorphic based on luciferase and cell line based assays (Li et al., 2014). It is therefore not surprising that *Oct2* has been found to be a significant MoMuLV CIS in this project. However, whilst this finding is not novel, identifying known human oncogenes validates this model in detecting new human candidate genes.

In my screen, insertions at *Pou2f2* occurred similarly between male and female mice and increased over time, being most abundant in those mice that developed disease (Figure 7-1(b)). This fact, combined with the literature, makes it very likely that *Pou2f2* is acting as a strong driver mutation, also selected to cooperate with *BCL2*, in this mouse model. The stronger signal for inserts in day 9 mice is likely to represent background noise due to the small number of mutations found in these mice.

7.1.2 *Ikzf3* (Aiolos)

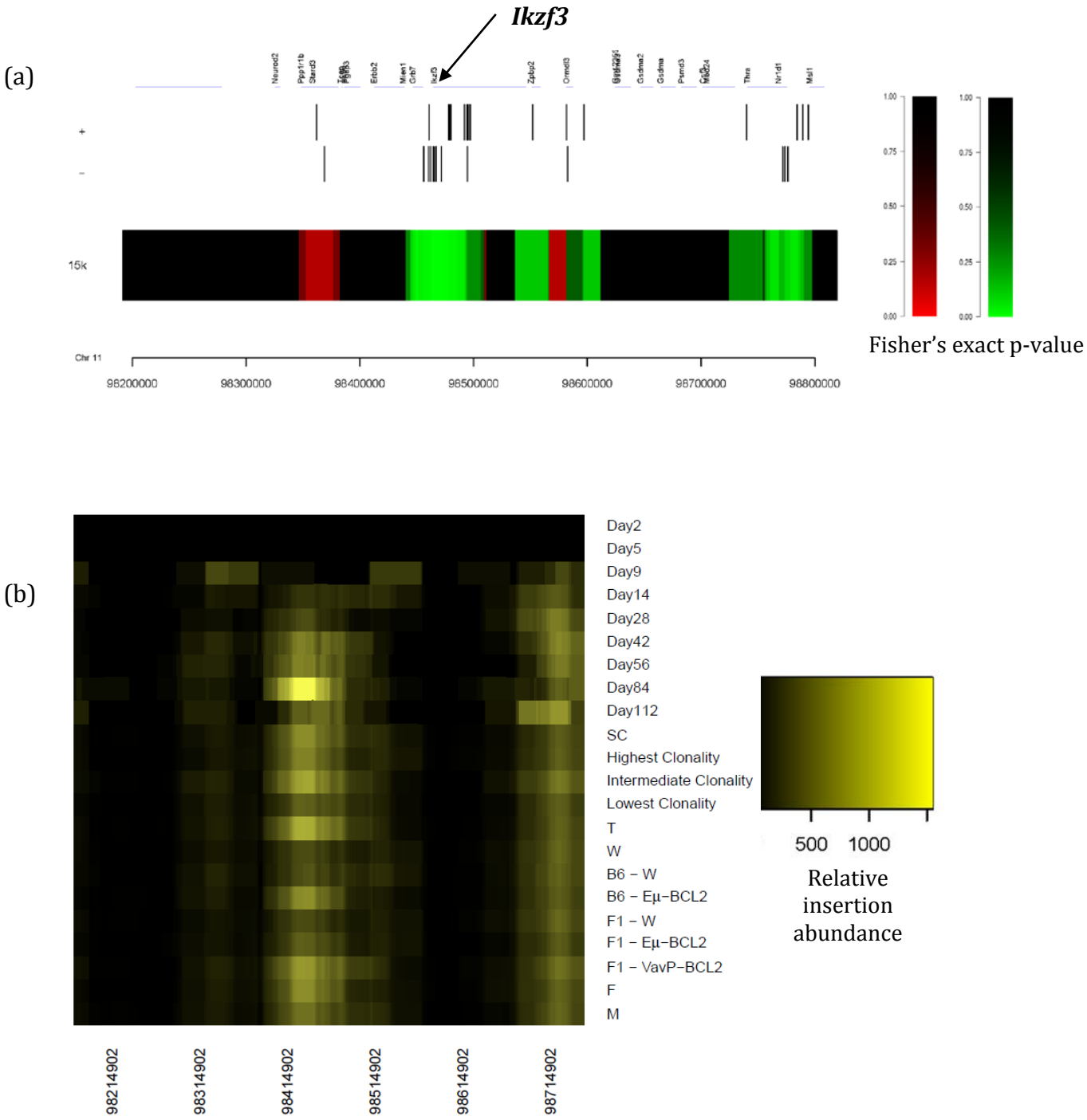


Figure 7-2 Genotype specificity and kinetics of Ikzf3 insertions
(a) A sliding window across the genome showing the MoMuLV insertions of all samples, measuring the genotype specificity by Fisher's Exact test. Green represents inserts biased towards transgenic mice and red towards wild-type mice. (b) The relative abundance of insertions at *Ikzf3* in different cohorts of mice. Calculated as: (number of inserts in a 50,000bp sliding window / total insertions in that cohort) x10^6. SC = survival cohort, F = female, M = male, T = transgenic, W = wild-type, B6 = C57BL/6, F1 = (BALB/c x C57BL/6) F1, + = forward strand, - = reverse strand.

Ikaros family member *Ikzf3* (Aiolos, Ikaros family zinc finger protein 3) encodes zinc-finger protein transcription factors that are an important regulator of lymphoid development and differentiation (Angelita Rebollo & Schmitt, 2003). It is also thought to play a role in apoptosis and has been found to regulate BCL2 family members including BCL2 via Aiolos-binding sites in the promoter (Romero, Martínez-A, Camonis, & Rebollo, 1999) and BCL-xL via IL-4 dependent mechanisms (A Rebollo, Ayllón, Fleischer, Martínez, & Zaballos, 2001). Loss of IKZF3 in mice leads to the development of B-cell lymphomas (J.-H. Wang et al., 1998) and was found to be occasionally mutated in human FL (Okosun et al., 2014). It has also been found to be mutated in chronic myeloid leukaemia (Menezes et al., 2013).

7.1.3 *Ebf1*

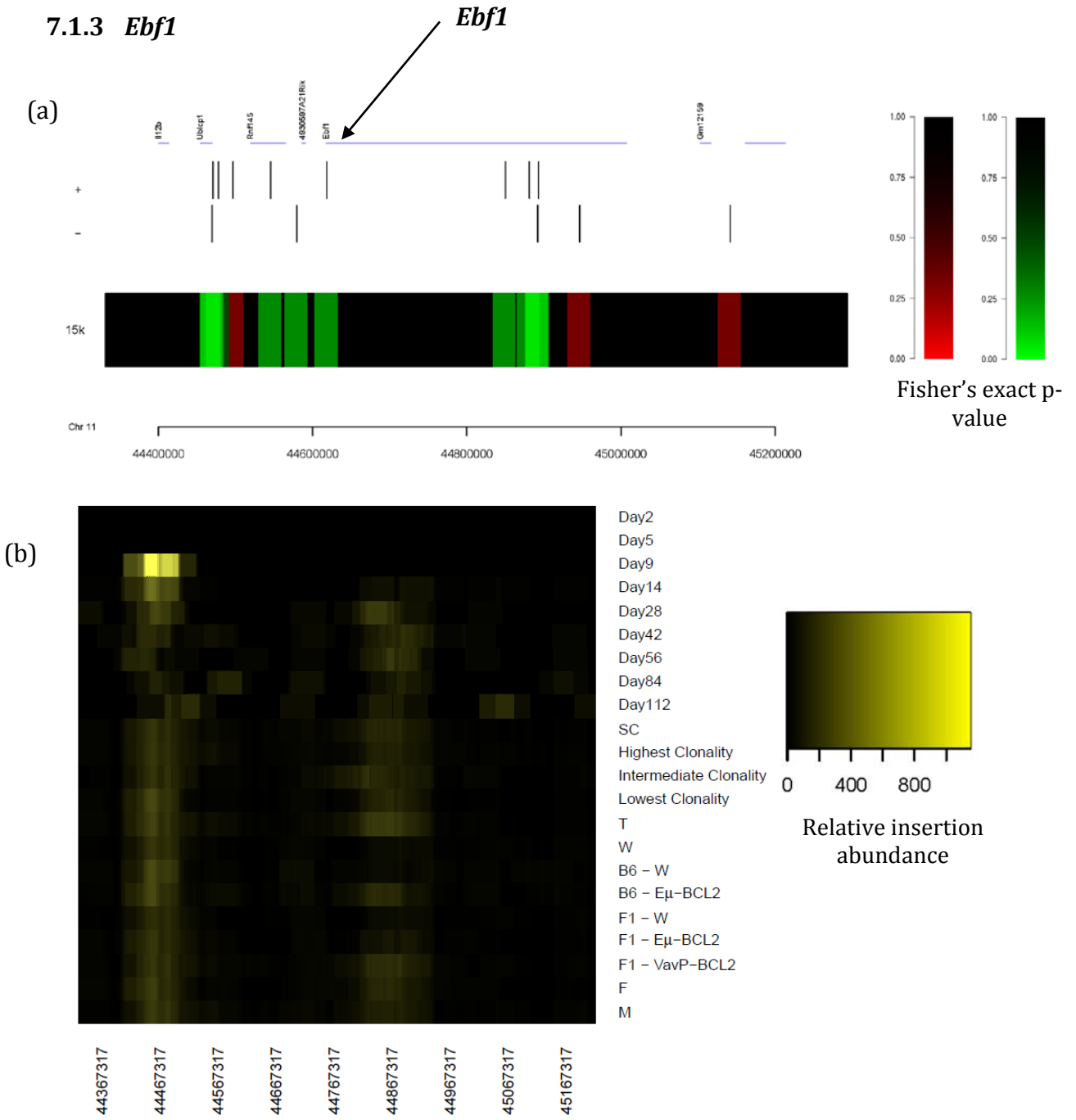


Figure 7-3 Genotype specificity and kinetics of Ebf1 insertions
(a) A sliding window across the genome showing the MoMuLV insertions of all samples, measuring the genotype specificity by Fisher's Exact test. Green represents inserts biased towards transgenic mice and red towards wild-type mice. (b) The relative abundance of insertions at *Ebf1* in different cohorts of mice. Calculated as: (number of inserts in a 50,000bp sliding window / total insertions in that cohort) x10^6. SC = survival cohort, F = female, M = male, T = transgenic, W = wild-type, B6 = C57BL/6, F1 = (BALB/c x C57BL/6) F1, + = forward strand, - = reverse strand.

Early B-cell factor 1 (*Ebf1*) is a DNA binding protein that activates transcription and is required in the early stages of B-cell development (Bain et al., 1994). It has been found to repress the activity of the histone acetyltransferases (HATs) *EP300* and *CREBBP* (Zhao, McCarrick-Walmsley, Akerblad, Sigvardsson, & Kadesch, 2003). HATs unwrap DNA from histones by acetylating histone tails, exposing the DNA to transcriptional machinery therefore playing a role in transcriptional activation (Bedford & Brindle, 2012). *EP300* and *CREBBP* act as tumour suppressors and inactivating mutations have been found to occur in 5% and 18% of DLBCL cases respectively (Morin et al., 2011; Pasqualucci et al., 2011). *EBF1* has been found to be translocated in human lymphomas (Bouamar et al., 2013). However, the mechanism of *EBF1* as an oncogene is unclear; whilst it could be assumed that down regulation would therefore promote the activity of the tumour suppressors *EP300* and *CREBBP*, loss of function mutations of *EBF1* has been found to play a role in Hodgkin lymphoma, B-progenitor acute lymphoblastic leukaemia and indeed FL (Bohle, Döring, Hansmann, & Küppers, 2013; D. Liao, 2009; Okosun et al., 2014).

Insertions at *Ebf1* were found significantly more frequently in transgenic mice than in wild-type litter mates (Figure 7-3(a)) although the level of significance reached was not as much as for other genes. In contrast to other candidate genes, the abundance of inserts at *Ebf1* is highest in early time course mice and appears to reduce over time (see Figure 7-3(b)). This may be because deregulated *Ebf1* is necessary for early clonal outgrowth of tumour but not in the later stages of disease onset.

7.1.4 *Cd86 & Ildr1*

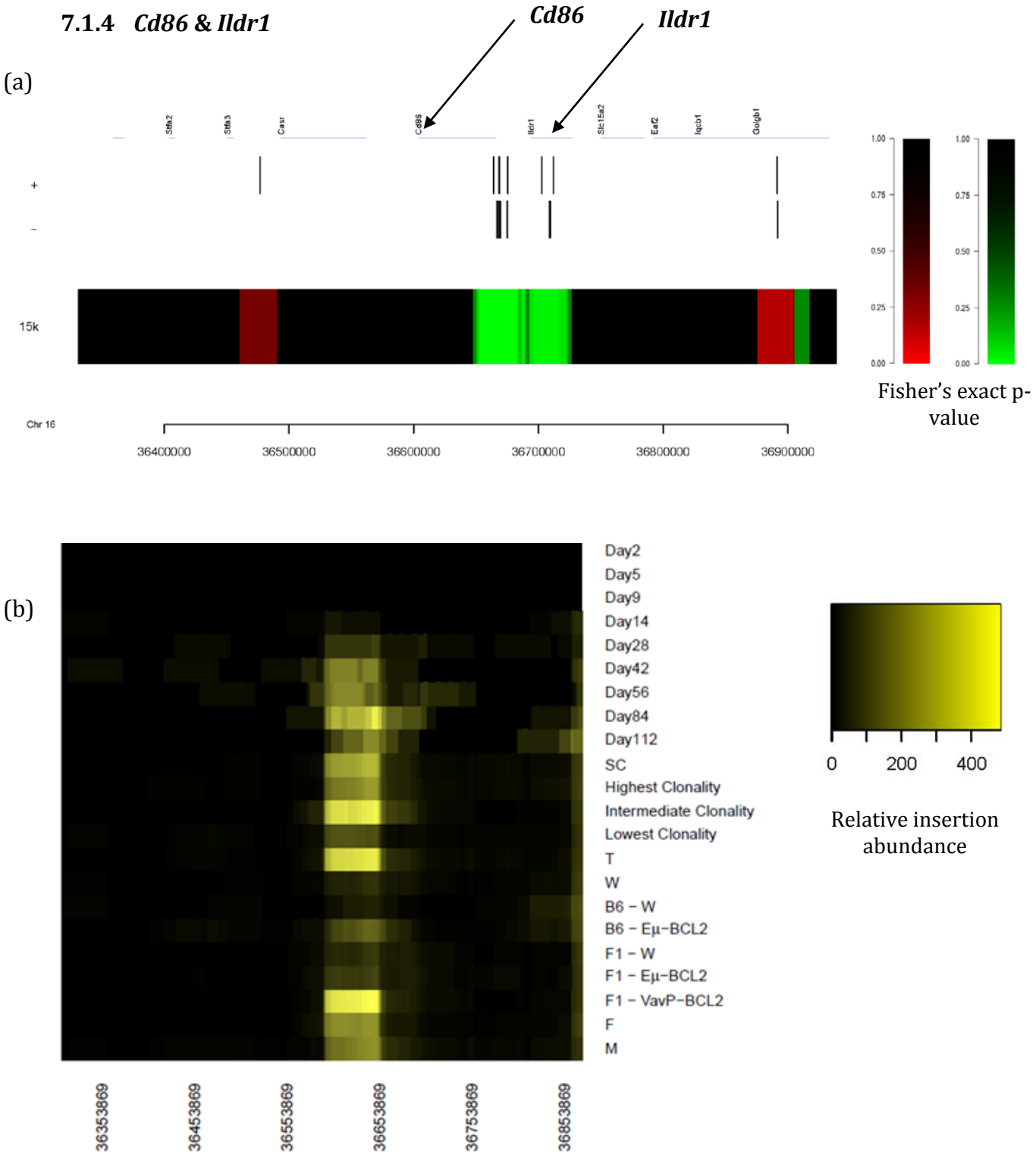


Figure 7-4 Genotype specificity and kinetics of Cd86 & Ildr1 insertions
(a) A sliding window across the genome showing the MoMuLV insertions of all samples, measuring the genotype specificity by Fisher's Exact test. Green represents inserts biased towards transgenic mice and red towards wild-type mice. (b) The relative abundance of insertions at *Cd86* & *Ildr1* in different cohorts of mice. Calculated as: (number of inserts in a 50,000bp sliding window / total insertions in that cohort) x10^6 SC = survival cohort, F = female, M = male, T = transgenic, W = wild-type, B6 = C57BL/6, F1 = (BALB/c x C57BL/6) F1, + = forward strand, - = reverse strand.

7.1.4.1 *Cd86*

T cell activation and functioning occurs by a two-signal model whereby the first signal is provided by the recognition of specific antigens by lymphocytes and the second is provided by additional 'co-stimulatory' signals between proteins on the surface of antigen presenting cells (APCs) and those on the surface of T cells (Sharpe & Freeman, 2002). Cluster of differentiation 86 (CD86), and also CD80, are proteins found on the surface of APCs including B-cells, macrophages and dendritic cells. They are crucial for controlling T cell activation within the adaptive immune response via their co-stimulatory effects on CD28 and co-inhibitory effects on CD152 (also known as CTLA4) which are proteins found on the surface of T cells (Sansom, Manzotti, & Zheng, 2003). Infection, stress and cellular damage activate APCs and induce the transcription, translation and transportation of CD80 and CD86 to the cell surface. CD28 is constitutively expressed on naïve T cells and stimulates T cell growth, differentiation, survival and function after ligation by CD86 and CD80. CD152 is induced following T cell activation and suppresses T cell responses (Chen & Flies, 2013). Activated $CD4^+$ T helper cells can then stimulate B-cells to produce immunoglobulins via increasing CD40 ligand expression and secretion of IL-4 (Finkelman et al., 1990; Stevens et al., 1988). In addition to activating T cells, CD86 has also been found to stimulate the activity of B-cells directly by increasing IgG production in anti-CD40/IL-4 primed human B-cells (Jeannin et al., 1997). It is also important in formation of germinal centres (Borriello, 1997). Interestingly, stimulation of CD86 was found to up-regulate Oct 2 (Podojil, Kin, & Sanders, 2004). CD80 has been found to be constitutively expressed on malignant B-cells (Munro et al., 1994) and both CD80 and CD86 may play a role in B-cell lymphoma (Suvas, Singh, Sahdev, Vohra, & Agrewala, 2002) although the mechanism is unclear. Galiximab is a chimeric anti-CD80 that showed promise in combination with rituximab

for treating FL (Czuczman et al., 2012). Most recently, a genome wide association study found associations of the CD86 locus with FL (Skibola et al., 2014) although this association is not significant. The role of *CD86* in lymphomagenesis is not clear, mutations have not been reported and inhibition of its action in-vivo has not been performed.

Abatacept is a recombinant fusion protein that consists of the extracellular domain of CTLA-4 (CD152) linked to the Fc portion of human IgG^1 that has been modified to avoid complement fixation. It competitively binds to CD80 and CD86, preventing their co-stimulatory action on CD28 and therefore inhibiting T cell activation. Intravenous Abatacept is licenced for treating patients with refractory rheumatoid arthritis. 8-year follow-up of the adverse events in abatacept treated patients have recently been published, showing that the standardised incidence ratio of lymphoma is higher, compared to the general population but comparable to other patients with rheumatoid arthritis, suggesting that decreasing CD80 and CD86 may be oncogenic. In fact, some malignancies (colorectal and breast) had lower incidence ratios (Weinblatt et al., 2013). Ipilimumab is an monoclonal antibody with anti-CTLA-4 activity which has been shown to have antitumour activity in patients with B-cell lymphoma (Ansell et al., 2009).

7.1.4.2 *Ildr1*

At the same viral insertion site for *Cd86* is the gene *Ildr1*. This gene encodes 4 splice variants and the shortest transcript has been implicated in the transformation of FL to high-grade DLBCL (Hauge, Patzke, Delabie, & Aasheim, 2004). More recently Ildr1 has been found to be overexpressed by the rare t(3;11)(q13;q14) translocation in myelodysplastic syndromes (Zagaria et al., 2012). The function of Ildr1 is poorly described to date and whilst Cd86 is highly likely to be involved in lymphomagenesis

based on what is already known, the interesting question is whether Ildr1 or some combination of both are also implicated, as the inserted MoMuLV could affect either gene.

In view of the CD86/Ildr1 locus being a very significant CIS in this screen, I decided to investigate the effects of these genes in-vivo in a number of ways. Firstly by retroviral transduction of these genes into mouse B cells that overexpress BCL2 and have p53 knocked out (mimicking the loss of 17p observed in haematologic malignancies) and then transplanting these cells into mice. The second is by abatacept treatment of human lymphoma cell lines subcutaneously transplanted into NOD-scid IL2Rγ^{null} (NSG) mice. NSG mice lack mature T cells, B-cells, functional NK cells and cytokine signalling and so are permissive to engrafting human cells and tissues.

In this insertional mutagenesis model either mutated *Cd86* or *Ildr1* (or a combination of both) could be responsible for lymphomagenesis. Interestingly, not only were insertions at this locus found significantly more in transgenic mice (Figure 7-4(a)) but more specifically in those transgenic mice with VavP promoter (Figure 7-4(b)). It may be that these mice express more BCL2 than those transgenic mice with Eμ promoter driven BCL2 and these deregulated genes rely on higher BCL2 expression to be selected for in clonal tumour development.

7.2 *BCL2* exclusive genes

Table 7-4 & Table 7-5 show lists of genes at the most clonally abundant common insertion sites found significantly more in wild-type mice than in the transgenic mice (ie. *BCL2* exclusive genes) as determined by GKC and KC-RBM respectively. In contrast to the *BCL2* specific genes, GO terms identified in association with these genes include biological processes that affect T cell activation and differentiation, as well as those that influence transcription. This is in contrast to the *BCL2* specific genes that identified GO terms in association with B cell processes. This is likely to be due to the fact that MoMuLV predominantly causes T-cell malignancies in wild-type mice.

Copz1 and *2610307P16Rik* are the top two genes on both lists. *2610307P16Rik* is not protein coding. *Copz1* knockdown has been shown to inhibit the growth of prostate cancer cells in-vitro (Shtutman et al., 2011) although there is no research on its role in lymphoma.

Rank	Chromo-some	Base Position	Gene Name	Ensembl Gene ID	No. of insertions in wild-type mice	No. of insertions in *BCL2* transgenic mice	All other insertions in wild-type mice	All other insertions in *BCL2* transgenic mice	p-value (two-tailed Fisher's Exact)
1	15	103295413	Copz1	ENSMUSG00000060992	13	0	7596	6718	0.0003
2	13	28548407	2610307P16Rik	ENSMUSG00000085936	11	0	7598	6718	0.0012
3	11	11708428	Ikzf1	ENSMUSG00000018654	24	5	7585	6713	0.0012
4	11	86881889	Dhx40	ENSMUSG00000018425	10	0	7599	6718	0.0023
5	5	140614769	Baat1	ENSMUSG00000000148	9	0	7600	6718	0.0044
6	1	34473955	Ptpn18	ENSMUSG00000026126	9	0	7600	6718	0.0044
7	2	117396234	Gm13982	ENSMUSG00000085681	39	15	7570	6703	0.0058
8	13	52553488	Diras2	ENSMUSG00000047842	10	1	7599	6717	0.0134
9	2	26498864	Notch1	ENSMUSG00000026923	31	12	7578	6706	0.0138
10	12	69478549	Gm9887	ENSMUSG00000052673	11	2	7598	6716	0.0257
11	15	62805492	Gm24810, SNORA17	ENSMUSG00000093058, ENSMUSG00000088897	6	0	7603	6718	0.0330
12	9	110643158	Kif9, Pth1r	ENSMUSG00000032489, ENSMUSG00000032492	6	0	7603	6718	0.0330
13	12	51858516	Hectd1	ENSMUSG00000035247	6	0	7603	6718	0.0330
14	15	84318788	1810041L15Rik	ENSMUSG00000062760	6	0	7603	6718	0.0330
15	2	92221927	Phf21a	ENSMUSG00000058318	6	0	7603	6718	0.0330
16	11	100871963	Stat3	ENSMUSG00000004040	18	6	7591	6712	0.0393
17	4	32342369	Bach2	ENSMUSG00000040270	12	3	7597	6715	0.0403
18	1	138178921	Ptprc	ENSMUSG00000026395	8	1	7601	6717	0.0423
19	13	28918766	2610307P16Rik, Sox4	ENSMUSG00000085936, ENSMUSG00000076431	16	5	7593	6713	0.0467

Table 7-4 Wild-type specific common insertion sites determined by Gaussian Kernel Convolution

Significant *BCL2* exclusive common insertion site genes from infected mice in the 'most clonal' group (see Figure 6-8) based on normalised clonality (above 0.05) with a 100, 000bp window. Derived by 'CIMPL' which uses Gaussian kernel convolution values to rank the insertions.

Rank	Gene	Ensembl Gene ID	No. of insertions in *BCL2* transgenic mice	No. of insertions in wild-type mice	All other insertions in *BCL2* transgenic mice	All other insertions in wild-type mice	2-tailed p-value
1	**2610307P16Rik**	ENSMUSG00000085936	6	31	6712	7578	0.0002
2	**Copz1**	ENSMUSG00000060992	1	14	6717	7595	0.0013
3	**Gm26660**	ENSMUSG00000097601	2	17	6716	7592	0.0018
4	**Bach2**	ENSMUSG00000040270	9	29	6709	7580	0.0050
5	**Ikzf1**	ENSMUSG00000018654	8	27	6710	7582	0.0058
6	**Notch1**	ENSMUSG00000026923	18	39	6700	7570	0.0232
7	**Arid5b**	ENSMUSG00000019947	1	9	6717	7600	0.0238
8	**Gm17619**	ENSMUSG00000097514	11	27	6707	7582	0.0332
9	**Scyl1**	ENSMUSG00000024941	4	15	6714	7594	0.0357
10	**Rgs1**	ENSMUSG00000026358	1	8	6717	7601	0.0423
11	**Phf21a**	ENSMUSG00000058318	1	8	6717	7601	0.0423
12	**Fcgrt**	ENSMUSG00000003420	2	10	6716	7599	0.0429
13	**Diras2**	ENSMUSG00000047842	2	10	6716	7599	0.0429
14	**Gm13982**	ENSMUSG00000085681	23	44	6695	7565	0.0489
15	**Iqgap1**	ENSMUSG00000030536	4	14	6714	7595	0.0557

Table 7-5 Wild-type specific common insertion sites determined by Kernel Convolved Rules Based Mapping
These insertions were found more frequently in the virus infected WT mice than virus infected transgenic mice

GO Term	Count	%	Genes	PValue	Benjamini corrected PValue	False Discovery Rate
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5	22.72727273	PTPRC, NOTCH1, IKZF1, SOX4, STAT3	0.003944865	0.892821173	5.62533958
GO:0051173~positive regulation of nitrogen compound metabolic process	5	22.72727273	PTPRC, NOTCH1, IKZF1, SOX4, STAT3	0.004406029	0.712764636	6.263347216
GO:0030217~T cell differentiation	3	13.63636364	PTPRC, IKZF1, SOX4	0.004457441	0.568876764	6.334225239
GO:0006357~regulation of transcription from RNA polymerase II promoter	5	22.72727273	NOTCH1, IKZF1, SOX4, PHF21A, STAT3	0.007699722	0.664386289	10.70456929
GO:0010604~positive regulation of macromolecule metabolic process	5	22.72727273	PTPRC, NOTCH1, IKZF1, SOX4, STAT3	0.008466457	0.617406676	11.70990993
GO:0030098~lymphocyte differentiation	3	13.63636364	PTPRC, IKZF1, SOX4	0.009778799	0.6036148	13.4062128
GO:0042110~T cell activation	3	13.63636364	PTPRC, IKZF1, SOX4	0.010110655	0.559668683	13.83032573
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	4	18.18181818	NOTCH1, IKZF1, SOX4, STAT3	0.011029711	0.543104744	14.99480164
GO:0002521~leukocyte differentiation	3	13.63636364	PTPRC, IKZF1, SOX4	0.014870066	0.609575222	19.70383596
GO:0045893~positive regulation of transcription, DNA-dependent	4	18.18181818	NOTCH1, IKZF1, SOX4, STAT3	0.016513481	0.609685015	21.64374611
GO:0051254~positive regulation of RNA metabolic process	4	18.18181818	NOTCH1, IKZF1, SOX4, STAT3	0.016832325	0.581855611	22.01501851
GO:0045597~positive regulation of cell differentiation	3	13.63636364	PTPRC, NOTCH1, IKZF1	0.022040254	0.649829595	27.8519906
GO:0006350~transcription	7	31.81818182	NOTCH1, BACH2, IKZF1, ARID5B, SOX4, PHF21A, STAT3	0.022459925	0.627408218	28.30417044
GO:0045941~positive regulation of transcription	4	18.18181818	NOTCH1, IKZF1, SOX4, STAT3	0.023432059	0.615920768	29.34148673
GO:0010628~positive regulation of gene expression	4	18.18181818	NOTCH1, IKZF1, SOX4, STAT3	0.025141313	0.616761242	31.13151136
GO:0046649~lymphocyte activation	3	13.63636364	PTPRC, IKZF1, SOX4	0.025942037	0.604723681	31.95545817
GO:0007242~intracellular signaling cascade	5	22.72727273	PTPRC, DIRAS2, PTH1R, IQGAP1, STAT3	0.029258826	0.627282844	35.27159991
GO:0010557~positive regulation of macromolecule biosynthetic process	4	18.18181818	NOTCH1, IKZF1, SOX4, STAT3	0.031122116	0.629318849	37.06783194
GO:0051094~positive regulation of developmental process	3	13.63636364	PTPRC, NOTCH1, IKZF1	0.032006491	0.619905903	37.90402649
GO:0045321~leukocyte activation	3	13.63636364	PTPRC, IKZF1, SOX4	0.033393109	0.616902506	39.19428719

Table 7-6 Gene ontology of *BCL2* exclusive (wild-type specific) common insertion sites found in insertional mutagenesis screen
The common insertion sites from the insertional mutagenesis screen identified by both Gaussian Kernel Convolution and Kernel Convolved Rules Based Mapping were input to DAVID Bioinformatics Resources 6.7, using the 'GO Fat' database to functionally annotate groups of genes (Huang et al., 2009a, 2009b).

7.2.1 *Ikzf1* (Ikaros)

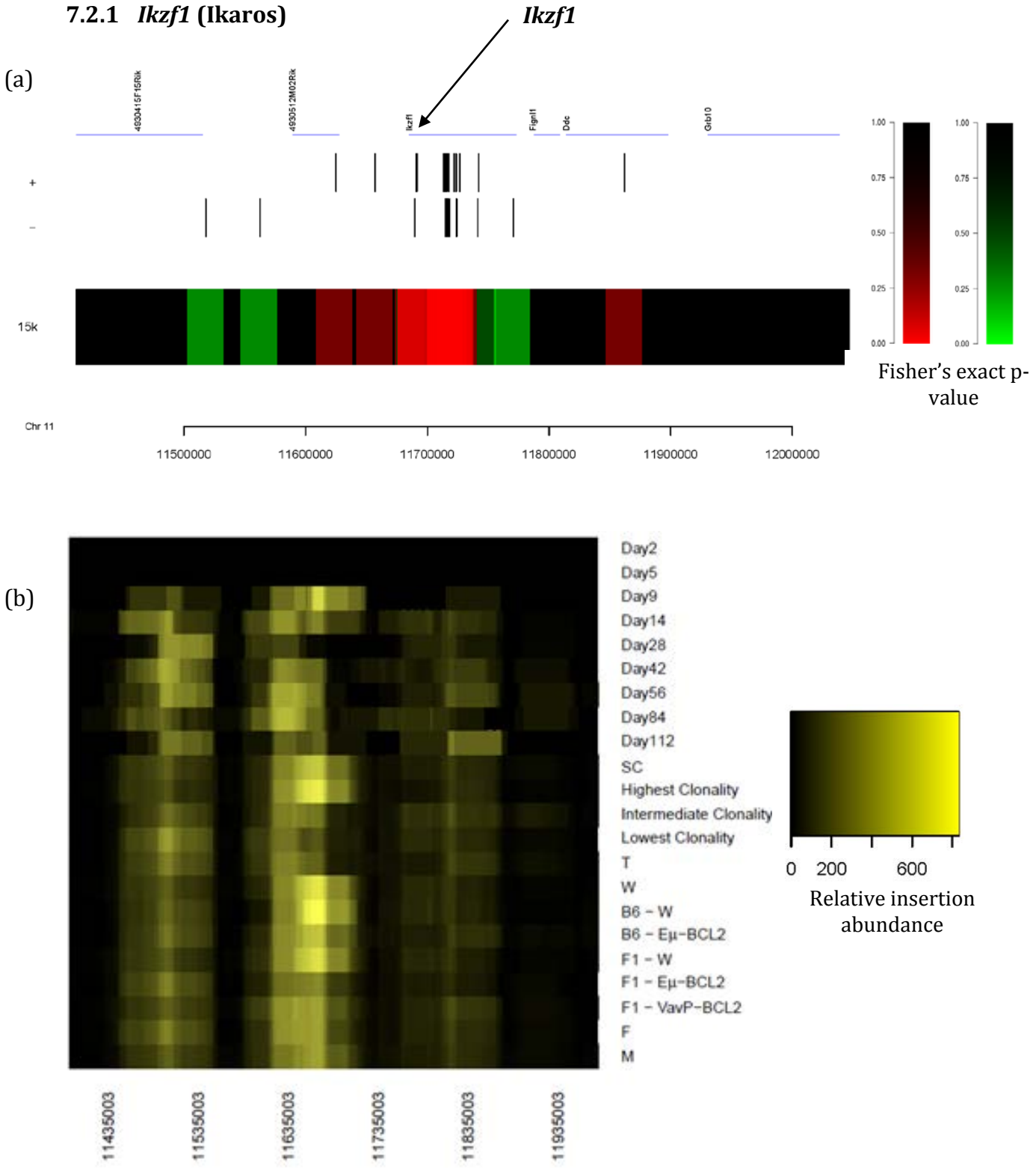


Figure 7-5 Genotype specificity and kinetics of Ikzf1insertions
(a) A sliding window across the genome showing the MoMuLV insertions of all samples, measuring the genotype specificity by Fisher's Exact test. Green represents inserts biased towards transgenic mice and red towards wild-type mice. (b) The relative abundance of insertions at *Ikzf1* in different cohorts of mice. Calculated as: (number of inserts in a 50,000bp sliding window / total insertions in that cohort) x10^6 SC = survival cohort, F = female, M = male, T = transgenic, W = wild-type, B6 = C57BL/6, F1 = (BALB/c x C57BL/6) F1, + = forward strand, - = reverse strand.

IKZF1 is the hallmark member of the Ikaros family of transcription factors involved in the determination of haemopoietic stem cell fate and lymphocyte development (John & Ward, 2011). Ikaros expression is mainly restricted to lymphopoietic tissues including spleen and thymus (Molnár et al., 1996). *Ikaros* knockout mice have a reduced capacity for progenitor cell self-renewal. They lack B-cells and their precursors, myeloid lineage cell differentiation is disrupted, they have severe anaemia but an increased platelet cell count is noted (John & Ward, 2011). Loss of function mutations in Ikaros are a common feature of B-cell acute lymphoblastic leukaemia (B-ALL) which suggests that it functions as a tumour suppressor (Mullighan et al., 2008). Lenalidomide is a drug used to treat both multiple myeloma and B-cell malignancies that has recently been found to act by selective ubiquitination and degradation of IKZF1 and IKZF3 leading to increased IL-2 production in T cells (Krönke et al., 2014).

7.3 Candidate Gene Validation

7.3.1 Lymphoma model used for validation

In order to study the effect of candidate gene overexpression *in-vivo*, an MSCV retrovirus was used to transduce mouse B-cells with different genes and then these cells were transplanted by tail vein injection into sublethally irradiated C57BL/6 mice. As one gene mutation alone is very unlikely to be oncogenic, a model that develops lymphoma spontaneously was established in order that the effect of a given candidate gene (which may be an oncogene or a tumour suppressor gene or indeed neither) on the time to lymphoma onset could be studied.

Eμ-*BCL2 p53*$^{+/-}$ transgenic C57BL/6 mouse B-cells were transduced with Mycn using an MSCV retrovirus. Cells were FACS sorted for high and low expression of Mycn and then injected intravenously by tail vein into C57BL/6 WT mice. Those mice that received cells with higher expression of Mycn developed lymphoma significantly faster than those that received cells with low Mycn expression (Figure 7-6).

Insertions at the loci correlating with both *Cd86* and *Ildr1* were found significantly more in transgenic mice, and although the literature would heavily implicate germline variation at this locus to be lymphomagenic, they have not previously been validated formerly as oncogenes or tumour suppressor genes. As potentially novel, but promising targets, they are therefore ideal candidate genes to test for their oncogenic potential.

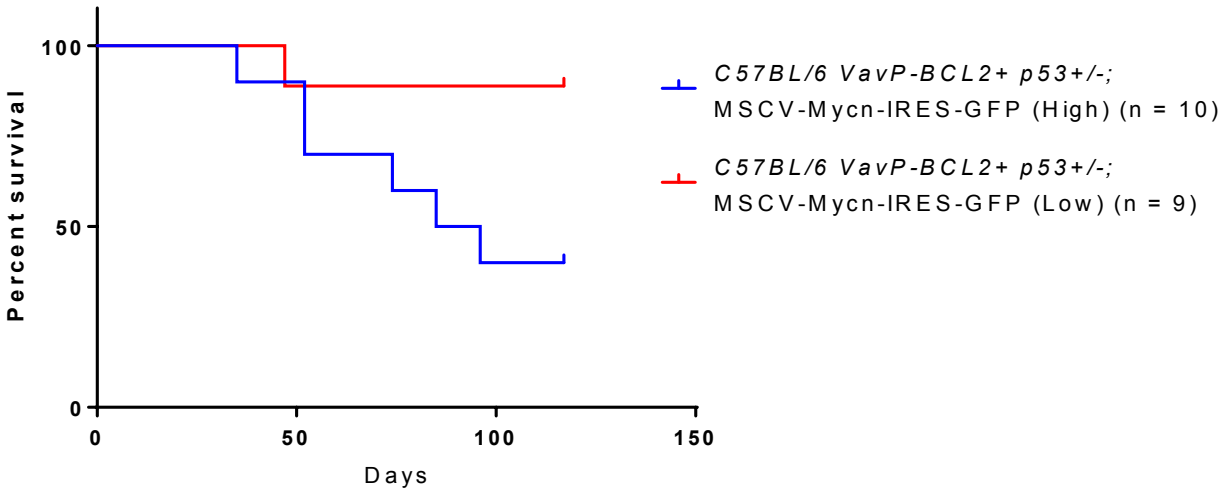


Figure 7-6 Kaplan Meier survival of C57BL/6 mice transplanted with Eμ-*BCL2 p53+/-* mouse B-cells overexpressing *Mycn*

Those mice transplanted with cells expressing higher levels of Mycn developed lymphoma significantly quicker than those transplanted with cells expressing low / no Mycn (Log-rank (Mantel-Cox) test p=0.0431, Gehan-Breslow-Wilcoxon test p=0.0593).

7.3.2 *Cd86* validation

Using the MSCV retrovirus model, *Cd86* and *Mycn* were both overexpressed in Eμ-*BCL2* *p53*$^{+/-}$ transgenic C57BL/6 mouse B-cells and then transplanted into C57BL/6 mice. Those mice overexpressing *Cd86* and *Mycn* survived, lymphoma free, significantly longer than those overexpressing *Mycn* alone (Figure 7-7). This would suggest that *Cd86* has some tumour suppressor gene functionality. The potential mechanism for this finding is not clear. Possibly a change in Cd86 level alters the balance of co-excitatory or co-inhibitory signals on T cell activation.

7.3.3 *Ildr1* validation

Using the MSCV retrovirus model, *Ildr1* and *Mycn* were both overexpressed in Eμ-*BCL2* *p53*$^{+/-}$ transgenic C57BL/6 mouse B-cells and then transplanted into C57BL/6 mice. Those mice overexpressing *Ildr1* and *Mycn* survived lymphoma free for a significantly shorter time than those overexpressing *Mycn* alone (Figure 7-8). This would suggest that *Ildr1* is an oncogene.

Further work needs to be done to corroborate these promising initial results.

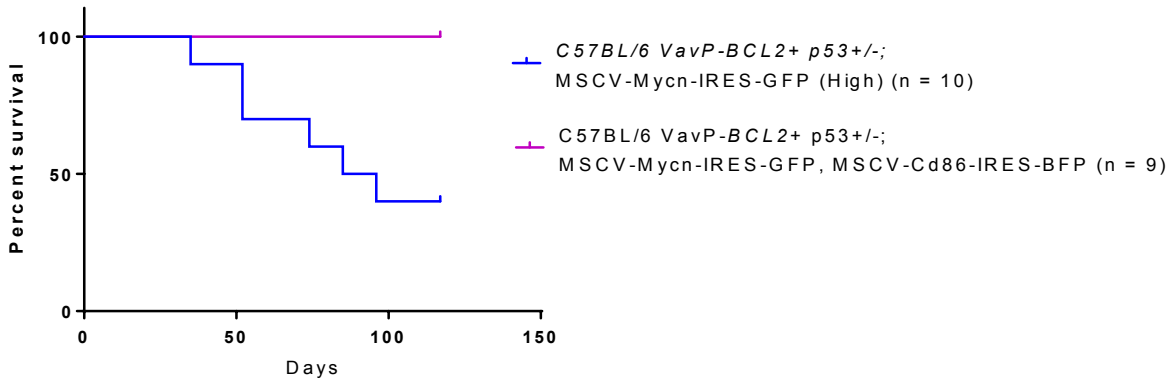


Figure 7-7 Kaplan Meier survival of C57BL/6 mice transplanted with *Eμ-BCL2 p53+/-* mouse B-cells overexpressing *Mycn* & *Cd86*
Those mice transplanted with cells overexpressing both *Mycn* and *Cd86* survived lymphoma free for significantly longer than those only overexpressing *Mycn* alone (Log-rank (Mantel-Cox) test p=0.0062, Gehan-Breslow-Wilcoxon test p=0.0073).

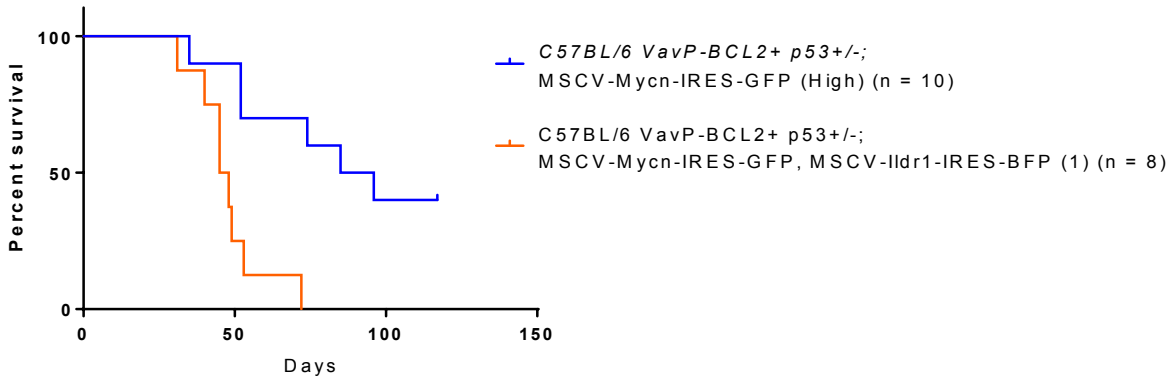


Figure 7-8 Kaplan Meier survival of C57BL/6 mice transplanted with Eμ-*BCL2 p53+/-* mouse B-cells overexpressing *Mycn* & *Ildr1*

Those mice transplanted with cells overexpressing both *Mycn* and *Ildr1* survived lymphoma free for a significantly shorter time than those only overexpressing *Mycn* alone (Log-rank (Mantel-Cox) test p=0.0015, Gehan-Breslow-Wilcoxon test p=0.0051).

CHAPTER 8 DISCUSSION

8.1 Insertional mutagenesis as a cancer model

Recent sequencing of human lymphomas has shown great success in identifying driver mutations of the exome, primarily because their significance can be readily estimated by comparing ratios of synonymous, non-synonymous and splicing mutations. However, the heterogeneity of this illness means there is more to cancer genomes than recurrent exonic mutations. For example, an exome sequencing study of 148 CLL patients identified 15 coding mutations per patient (Landau et al., 2013), however of these only 145 (<1 per exome) were found in recurrently mutated genes, suggesting that malignancy is also driven by a spectrum of rare coding mutations, non-coding mutations, large-scale copy number aberrations and epigenetic deregulation. Insertional mutagenesis screens in mouse models can play a complementary role in defining and validating rare and non-exonic driver mutations in human cancer, also expanding the set of putative therapeutic targets beyond currently identified cancer genes.

This thesis describes the use of retrovirally driven cancer in mice. However, there are many viruses that are well known to be oncogenic in humans, causing significant disease. Epstein-Barr virus (EBV) is a herpesvirus that is almost ubiquitous amongst humans. Acute infection may result in an infectious mononucleosis but latent infection is associated with a number of malignancies including post-transplant lymphoproliferative disorder (Hopwood, 2000) and Burkitt lymphoma (Epstein, Achong, & Barr, 1964; Shibata et al., 1993). Hepatocellular carcinoma is associated with

chronic hepatitis C virus infection (Davila, Morgan, Shaib, McGlynn, & El–Serag, 2004). Human Herpes Virus 8 is the aetiological agent in Kaposi's Sarcoma (Chang et al., 1994). Human papillomaviruses cause almost all invasive cervical cancer worldwide (Walboomers et al., 1999).

In the most part, viral proteins are thought to be oncogenic, however there may be a role for insertion mutations causing disease. Gene therapy for severe combined immunodeficiency (SCID)-X1 disease using Moloney retrovirus-mediated gene transfer of interleukin-2 receptor subunit gamma (*IL2RG*) into autologous CD34 bone marrow cells was initially very successful (Cavazzana-Calvo et al., 2000). However, three years after therapy two patients developed uncontrolled clonal proliferation of mature T cells. Both patients were found to have retrovirus integration in close proximity to the LMO2 proto-oncogene, leading to deregulated transcription and expression (Hacein-Bey-Abina et al., 2003). Notably the *Il2rg* and *Lmo2* genes were also found to be commutated in MoMuLV induced lymphomas of mice at rates higher than expected by chance, suggesting these genes cooperate in lymphomagenesis in both mice and humans (Davé et al., 2009).

Human T-lymphotropic virus type-1 (HTLV-1) is a retrovirus which is endemic in south Japan, South America, subtropical Africa and northern Iran. It infects and integrates into the genome of human CD4+ T lymphocytes causing the aggressive adult T-cell leukaemia/lymphoma (ATL) that carries a poor prognosis (Yoshida, Miyoshi, & Hinuma, 1982; Yoshida, 2005). HTLV-1 encodes many regulatory gene products that control its own transcription, the expression of host genes and the proliferation of the host cell (Matsuoka & Jeang, 2011). The importance of assessing HTLV-1 clonal abundance in interrogating the aetiology of ATL has recently been studied, showing that

populations with different levels of HTLV-1 clonality are all significant in disease pathogenesis (Cook et al., 2014).

In addition to virally driven malignancies, endogenous transposable elements in the human genome can facilitate mutagenic retrotranspositions that deregulate gene expression and this process is implicated in colorectal cancer, prostate cancer, ovarian cancer, multiple myeloma and glioblastoma (Lee et al., 2012).

8.2 Library prep / sequencing protocol

This study represents the most comprehensive MoMuLV insertional mutagenesis screen with the greatest depth of sequencing to date. Older methods including splinkerette PCR, restriction enzyme DNA fragmentation, shotgun subcloning and 454 pyrosequencing are still being used by other groups (Baron et al., 2012; C. a Huser et al., 2014; Klijn et al., 2013). Transgenic mice overexpressing BCL2 were previously screened and three proviral insertion sites were identified by southern blot analysis (Shinto et al., 1995). Most recently, deeper sequencing of a panel of 28 lymphomas revealed 12,485 insertion sites (C. a Huser et al., 2014). In contrast, even after highly stringent filtering to eliminate contaminating insertion site data, this screen identified 762,228 MoMuLV insertion sites. The methods used and depth of sequencing allows the study of insertion site clonal abundance and also the detailed study of subclonal populations. We have therefore produced a body of work that we plan to make available to the research community, via an online website, that can be mined.

The common insertion sites identified within this insertional mutagenesis screen warrant further study in order to validate which of these are either oncogenic or

tumour suppressive and if they are cooperating with *BCL2* in their oncogenicity. As discussed in section 7.3, some early promise has been shown with regard to both *CD86* and *IlDR1* in their ability to alter the onset of lymphoma and work in the immediate future on these is imperative. This work starts to look at combinations of mutations that could aid in the design of combinatorial treatments that target multiple strong driver mutations within a tumour. Not only would this be beneficial to the understanding and treatment of lymphoma and other malignancies, but also to autoimmune diseases which similarly represent over activity of lymphocytes and other immune cells. Rituximab, cyclophosphamide, corticosteroids and abatacept are all drugs used to treat systemic lupus erythematosus, rheumatoid arthritis and other immune complex / antibody mediated disorders, indicating that their shared biological mechanisms with lymphoma are significant.

Since the start of this study, groups studying viruses other than MoMuLV have developed Illumina based sequencing methods and used fragmentation by sonication to identify the clonal abundance of virus integration sites (Firouzi et al., 2014; Gillet et al., 2011b). In order to assess the clonal abundance and starting amount of DNA of single clones, one group developed a tag system where a sequence of random nucleotides was introduced during adaptor ligation, meaning every ligated fragment was unique (Firouzi et al., 2014). This meant that clone size could be experimentally measured rather than relying on statistical estimation.

8.3 Mutation kinetics / profiling

Many studies have looked into the kinetics of mutation onset in the development of disease by attempting to quantify the clonal abundance of mutations. Studies have

assumed that the most clonal mutations are early onset, whilst lower clonality mutations are late. Some human studies have used initial disease and disease relapse to represent early and late disease respectively. This study is the first, to our knowledge, to study mutation clonality and kinetics in an animal model prior to the onset of disease, which would be impossible in humans.

A recent MoMuLV insertional mutagenesis screen performed in transgenic mice expressing the two oncogenes *MYC* and *Runx2* analysed 28 tumours and identified 771 CISs (C. A. Huser et al., 2014). This group did a detailed analysis of clonality using restriction enzyme digestion to fragment DNA and 454-pyrosequencing. The first finding of this study was that a small set of genes in this transgenic model confers cell self-renewal, offering up a limited number of genes as potential drug targets. They also found that a larger pool of genes control the proliferation of malignant cell clones. The work in this thesis, in part, corroborates these findings in that several diseased mice had a small number of highly clonal mutations (those mice in cluster 1). However, a number of diseased mice had more than a 'small number' of highly clonal mutations (some up to approximately 20) and some diseased mice had many mutations of similar (and relatively low) clonal abundance suggesting that mutations are very heterogeneous. The studies of human lymphoma also support a widely heterogeneous disease profile and would suggest there is more than one genomic mechanism of developing lymphoma. It would also be interesting to look at the epigenetics and proteomics in disease pathogenesis, and also the impact of non-protein coding genes.

Understanding the kinetics of mutation clonal abundance in the lead up to cancer is useful for a number of reasons. It can help in our understanding of disease mechanism and aids in addressing a number of questions, e.g. How many mutations are required to

cause this disease?; How long are the mutations present before the onset of disease?; To what extent does identifying low clonality, pre-malignant mutations help in predicting onset of disease?; Which specific mutations come together to cause this disease? Understanding how clonally abundant a mutation has to be to allow disease onset, and which other mutations are required to facilitate disease may go some way in explaining why healthy people can live with known oncogenic mutations without developing disease. In addition to disease mechanism, monitoring mutation clonal abundance over time prior to the onset of clinically detectable disease could also have direct clinical applications in terms of predicting the likelihood and time to relapse after patients have gone into remission. This in turn could influence treatment strategies, as those patients predicted to relapse in the near future could be considered for pre-emptive / prophylactic treatment.

CHAPTER 9 CONCLUSIONS & FUTURE WORK

- A novel protocol for quantification of insertion sites in retroviral insertional mutagenesis that is cost effective, high-throughput and can be applied to an Illumina sequencing platform has been designed and validated. In addition to identifying known oncogenes, which validates this method, it has identified a number of new potential gene targets. This method not only improves the quality of retroviral insertional mutagenesis mouse models of cancer but could also be applied to human diseases.

- Novel putative oncogenic targets that occur in *BCL2* driven lymphoma have been identified, improving our understanding of the genomic landscape of this relatively common cancer and possibly facilitating new therapeutic drug discovery. Early promise has been shown in validating *Cd86* and *Ildr1* as genes that are deregulated in lymphoma and could be targeted for treatment. This was achieved through their overexpression in mouse models leading to the deceleration and acceleration of disease onset respectively. Further validation of these, and other potential candidates, is urgently required. A knockdown or knockout experiment would be the appropriate next step.

- A better characterisation of the kinetics of insertion site profiles over time and across different lymphoid organs has been established. Correlation of these profiles with organ size, tumour characterisation (by FACS) and time to disease onset, as well as the study of specific gene mutations over time, is now required.

- Correlation of the above findings in human lymphoma, using new and existing data sets, would be extremely valuable to further validate these findings.

Bibliography

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., … Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–11. doi:10.1038/35000501

Aloia, A. L., Duffy, L., Pak, V., Lee, K. E., Sanchez-Martinez, S., Derse, D., … Rein, A. (2013). A reporter system for replication-competent gammaretroviruses: the inGluc-MLV-DERSE assay. *Gene Therapy*, *20*(2), 169–76. doi:10.1038/gt.2012.18

Ansell, S. M., Hurvitz, S. A., Koenig, P. A., LaPlant, B. R., Kabat, B. F., Fernando, D., … Timmerman, J. M. (2009). Phase I study of ipilimumab, an anti-CTLA-4 monoclonal antibody, in patients with relapsed and refractory B-cell non-Hodgkin lymphoma. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *15*(20), 6446–53. doi:10.1158/1078-0432.CCR-09-1339

Bain, G., Maandag, E. C., Izon, D. J., Amsen, D., Kruisbeek, A. M., Weintraub, B. C., … van Roon, M. (1994). E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements. *Cell*, *79*(5), 885–92.

Bardwell, P. D., Gu, J., McCarthy, D., Wallace, C., Bryant, S., Goess, C., … Ghayur, T. (2009). The Bcl-2 family antagonist ABT-737 significantly inhibits multiple animal models of autoimmunity. *Journal of Immunology (Baltimore, Md. : 1950)*, *182*(12), 7482–9. doi:10.4049/jimmunol.0802813

Baron, B. W., Anastasi, J., Bies, J., Reddy, P. L., Joseph, L., Thirman, M. J., … Baron, J. M. (2014). GFI1B, EVI5, MYB--additional genes that cooperate with the human BCL6 gene to promote the development of lymphomas. *Blood Cells, Molecules & Diseases*, *52*(1), 68–75. doi:10.1016/j.bcmd.2013.07.003

Baron, B. W., Anastasi, J., Hyjek, E. M., Bies, J., Reddy, P. L., Dong, J., … Baron, J. M. (2012). PIM1 gene cooperates with human BCL6 gene to promote the development of lymphomas. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(15), 5735–9. doi:10.1073/pnas.1201168109

Barrans, S. L. (2002). Germinal center phenotype and bcl-2 expression combined with the International Prognostic Index improves patient risk stratification in diffuse large B-cell lymphoma. *Blood*, *99*(4), 1136–1143. doi:10.1182/blood.V99.4.1136

Bedford, D. C., & Brindle, P. K. (2012). Is histone acetylation the most important physiological function for CBP and p300? *Aging*, *4*(4), 247–55.

Bohle, V., Döring, C., Hansmann, M.-L., & Küppers, R. (2013). Role of early B-cell factor 1 (EBF1) in Hodgkin lymphoma. *Leukemia*, *27*(3), 671–9. doi:10.1038/leu.2012.280

BORRIELLO, F. (1997). B7-1 and B7-2 Have Overlapping, Critical Roles in Immunoglobulin Class Switching and Germinal Center Formation. *Immunity*, *6*(3), 303–313. doi:10.1016/S1074-7613(00)80333-7

Bouamar, H., Abbas, S., Lin, A.-P., Wang, L., Jiang, D., Holder, K. N., … Aguiar, R. C. T. (2013). A capture-sequencing strategy identifies IRF8, EBF1, and APRIL as novel IGH fusion partners in B-cell lymphoma. *Blood*, *122*(5), 726–33. doi:10.1182/blood-2013-04-495804

Bouska, A., McKeithan, T. W., Deffenbacher, K. E., Lachel, C., Wright, G. W., Iqbal, J., … Chan, W.-C. (2014). Genome-wide copy-number analyses reveal genomic abnormalities involved in transformation of follicular lymphoma. *Blood*, *123*(11), 1681–90. doi:10.1182/blood-2013-05-500595

Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., Yvon, E., Nusbaum, P., … Fischer, A. (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science (New York, N.Y.)*, *288*(5466), 669–72.

Chang, Y., Cesarman, E., Pessin, M. S., Lee, F., Culpepper, J., Knowles, D. M., & Moore, P. S. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science (New York, N.Y.)*, *266*(5192), 1865–9.

Chatterjee, N., Hartge, P., Cerhan, J. R., Cozen, W., Davis, S., Ishibe, N., … Severson, R. K. (2004). Risk of non-Hodgkin's lymphoma and family history of lymphatic, hematologic, and other cancers. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American*

Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, *13*(9), 1415–21.

Chen, L., & Flies, D. B. (2013). Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nature Reviews. Immunology*, *13*(4), 227–42. doi:10.1038/nri3405

Cheon, D.-J., & Orsulic, S. (2011). Mouse Models of Cancer.

Cheung, K.-J. J., Delaney, A., Ben-Neriah, S., Schein, J., Lee, T., Shah, S. P., … Horsman, D. E. (2010). High resolution analysis of follicular lymphoma genomes reveals somatic recurrent sites of copy-neutral loss of heterozygosity and copy number alterations that target single genes. *Genes, Chromosomes & Cancer*, *49*(8), 669–81. doi:10.1002/gcc.20780

Cheung, K.-J. J., Shah, S. P., Steidl, C., Johnson, N., Relander, T., Telenius, A., … Horsman, D. E. (2009). Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances. *Blood*, *113*(1), 137–48. doi:10.1182/blood-2008-02-140616

Clerc, R. G., Corcoran, L. M., LeBowitz, J. H., Baltimore, D., & Sharp, P. A. (1988). The B-cell-specific Oct-2 protein contains POU box- and homeo box-type domains. *Genes & Development*, *2*(12A), 1570–81.

Colin, J., Gaumer, S., Guenal, I., & Mignotte, B. (2009). Mitochondria, Bcl-2 family proteins and apoptosomes: of worms, flies and men. *Frontiers in Bioscience (Landmark Edition)*, *14*, 4127–37.

Cook, L. B., Melamed, A., Niederer, H., Valganon, M., Laydon, D., Foroni, L., … Bangham, C. R. M. (2014). The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma. *Blood*, *123*(25), 3925–31. doi:10.1182/blood-2014-02-553602

Czuczman, M. S., Leonard, J. P., Jung, S., Johnson, J. L., Hsi, E. D., Byrd, J. C., & Cheson, B. D. (2012). Phase II trial of galiximab (anti-CD80 monoclonal antibody) plus rituximab (CALGB 50402): Follicular Lymphoma International Prognostic Index (FLIPI) score is predictive of upfront immunotherapy responsiveness. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO*, *23*(9), 2356–62. doi:10.1093/annonc/mdr620

d'Amore, F., Chan, E., Iqbal, J., Geng, H., Young, K., Xiao, L., … Dave, B. J. (2008). Clonal evolution in t(14;18)-positive follicular lymphoma, evidence for multiple common pathways, and frequent parallel clonal evolution. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *14*(22), 7180–7. doi:10.1158/1078-0432.CCR-08-0752

Davé, U. P., Akagi, K., Tripathi, R., Cleveland, S. M., Thompson, M. A., Yi, M., … Copeland, N. G. (2009). Murine leukemias with retroviral insertions at Lmo2 are predictive of the leukemias induced in SCID-X1 patients following retroviral gene therapy. *PLoS Genetics*, *5*(5), e1000491. doi:10.1371/journal.pgen.1000491

Davila, J. A., Morgan, R. O., Shaib, Y., McGlynn, K. A., & El–Serag, H. B. (2004). Hepatitis C infection and the increasing incidence of hepatocellular carcinoma: A population-based study. *Gastroenterology*, *127*(5), 1372–1380. doi:10.1053/j.gastro.2004.07.020

De Jong, J., Akhtar, W., Badhai, J., Rust, A. G., Rad, R., Hilkens, J., … de Ridder, J. (2014). Chromatin landscapes of retroviral and transposon integration profiles. *PLoS Genetics*, *10*(4), e1004250. doi:10.1371/journal.pgen.1004250

De Jong, J., de Ridder, J., van der Weyden, L., Sun, N., van Uitert, M., Berns, A., … Wessels, L. F. A. (2011). Computational identification of insertional mutagenesis targets for cancer gene discovery. *Nucleic Acids Research*, *39*(15), e105. doi:10.1093/nar/gkr447

De Ridder, J., Uren, A., Kool, J., Reinders, M., & Wessels, L. (2006). Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Computational Biology*, *2*(12), e166. doi:10.1371/journal.pcbi.0020166

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., … DiPersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, *481*(7382), 506–10. doi:10.1038/nature10738

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., & Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell*, *122*(3), 473–83. doi:10.1016/j.cell.2005.07.013

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., … Sawyers, C. L. (2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England Journal of Medicine*, *344*(14), 1031–7. doi:10.1056/NEJM200104053441401

Egle, A., Harris, A. W., Bath, M. L., O'Reilly, L., & Cory, S. (2004). VavP-Bcl2 transgenic mice develop follicular lymphoma preceded by germinal center hyperplasia. *Blood*, *103*(6), 2276–83. doi:10.1182/blood-2003-07-2469

EPSTEIN, M. A., ACHONG, B. G., & BARR, Y. M. (1964). VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *Lancet*, *1*(7335), 702–3.

Fernald, A., Bergerson, R. J., Wang, J., McNerney, M. E., Karrison, T., Anastasi, J., … Beau, M. M. le. (2013). Retroviral Insertional Mutagenesis In Egr1+/- mice, Haploinsufficient For a Human Del(5q) Myeloid Leukemia Gene, Develop Myeloid Neoplasms With Proviral Insertions In Genes Syntenic To Human 5q. *Blood*, *122*(21), 1275.

Finkelman, F. D., Holmes, J., Katona, I. M., Urban, J. F., Beckmann, M. P., Park, L. S., … Paul, W. E. (1990). Lymphokine control of in vivo immunoglobulin isotype selection. *Annual Review of Immunology*, *8*, 303–33. doi:10.1146/annurev.iy.08.040190.001511

Firouzi, S., López, Y., Suzuki, Y., Nakai, K., Sugano, S., Yamochi, T., & Watanabe, T. (2014). Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites. *Genome Medicine*, *6*(6), 46. doi:10.1186/gm568

Flaherty, K. T., Puzanov, I., Kim, K. B., Ribas, A., McArthur, G. A., Sosman, J. A., … Chapman, P. B. (2010). Inhibition of mutated, activated BRAF in metastatic melanoma. *The New England Journal of Medicine*, *363*(9), 809–19. doi:10.1056/NEJMoa1002011

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., … Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, *4*(3), 177–83. doi:10.1038/nrc1299

Gandhi, L., Camidge, D. R., Ribeiro de Oliveira, M., Bonomi, P., Gandara, D., Khaira, D., … Rudin, C. M. (2011). Phase I study of Navitoclax (ABT-

263), a novel Bcl-2 family inhibitor, in patients with small-cell lung cancer and other solid tumors. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *29*(7), 909–16. doi:10.1200/JCO.2010.31.6208

Gillet, N. A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., … Bangham, C. R. M. (2011a). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood*, *117*(11), 3113–22. doi:10.1182/blood-2010-10-312926

Gillet, N. A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., … Bangham, C. R. M. (2011b). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood*, *117*(11), 3113–22. doi:10.1182/blood-2010-10-312926

Gini, C. (1914). *Sulla misura della concentrazione e della variabilita dei caratteri*. Venezia: Premiate officine grafiche C. Ferrari.

Guénet, J. L. (2005). The mouse genome. *Genome Research*, *15*(12), 1729–40. doi:10.1101/gr.3728305

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M. P., Wulffraat, N., Leboulch, P., … Cavazzana-Calvo, M. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science (New York, N.Y.)*, *302*(5644), 415–9. doi:10.1126/science.1088547

Harazono, Y., Nakajima, K., & Raz, A. (2013). Why anti-Bcl-2 clinical trials fail: a solution. *Cancer Metastasis Reviews*. doi:10.1007/s10555-013-9450-8

Hasbold, J., Corcoran, L. M., Tarlinton, D. M., Tangye, S. G., & Hodgkin, P. D. (2004). Evidence from the generation of immunoglobulin G-secreting cells that stochastic mechanisms regulate lymphocyte differentiation. *Nature Immunology*, *5*(1), 55–63. doi:10.1038/ni1016

Hauge, H., Patzke, S., Delabie, J., & Aasheim, H.-C. (2004). Characterization of a novel immunoglobulin-like domain containing receptor. *Biochemical and Biophysical Research Communications*, *323*(3), 970–8. doi:10.1016/j.bbrc.2004.08.188

Hayward, W. S., Neel, B. G., & Astrin, S. M. (1981). Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis. *Nature*, *290*(5806), 475–480. doi:10.1038/290475a0

Heckman, C. A., Duan, H., Garcia, P. B., & Boxer, L. M. (2006). Oct transcription factors mediate t(14;18) lymphoma cell survival by directly regulating bcl-2 expression. *Oncogene*, *25*(6), 888–98. doi:10.1038/sj.onc.1209127

Höglund, M., Sehn, L., Connors, J. M., Gascoyne, R. D., Siebert, R., Säll, T., … Horsman, D. E. (2004). Identification of cytogenetic subgroups and karyotypic pathways of clonal evolution in follicular lymphomas. *Genes, Chromosomes & Cancer*, *39*(3), 195–204. doi:10.1002/gcc.10314

Holst, J., Vignali, K. M., Burton, A. R., & Vignali, D. A. A. (2006). Rapid analysis of T-cell selection in vivo using T cell-receptor retrogenic mice. *Nature Methods*, *3*(3), 191–7. doi:10.1038/nmeth858

Hopwood, P. (2000). The role of EBV in post-transplant malignancies: a review. *Journal of Clinical Pathology*, *53*(4), 248–254. doi:10.1136/jcp.53.4.248

Horsman, D. E., Connors, J. M., Pantzar, T., & Gascoyne, R. D. (2001). Analysis of secondary chromosomal alterations in 165 cases of follicular lymphoma with t(14;18). *Genes, Chromosomes & Cancer*, *30*(4), 375–82. doi:10.1002/gcc.1103

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. doi:10.1038/nprot.2008.211

Hui, E. K.-W., Wang, P.-C., & Lo, S. J. (1998). Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. *Cellular and Molecular Life Sciences (CMLS)*, *54*(12), 1403–1411. doi:10.1007/s000180050262

Huser, C. a, Gilroy, K. L., de Ridder, J., Kilbey, A., Borland, G., Mackay, N., … Neil, J. C. (2014). Insertional mutagenesis and deep profiling reveals

gene hierarchies and a Myc/p53-dependent bottleneck in lymphomagenesis. *PLoS Genetics*, *10*(2), e1004167. doi:10.1371/journal.pgen.1004167

Huser, C. A., Gilroy, K. L., de Ridder, J., Kilbey, A., Borland, G., Mackay, N., … Neil, J. C. (2014). Insertional mutagenesis and deep profiling reveals gene hierarchies and a Myc/p53-dependent bottleneck in lymphomagenesis. *PLoS Genetics*, *10*(2), e1004167. doi:10.1371/journal.pgen.1004167

Ivics, Z., Hackett, P. B., Plasterk, R. H., & Izsvák, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, *91*(4), 501–10.

Jaenisch, R. (1976). Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(4), 1260–4.

Jeannin, P., Delneste, Y., Lecoanet-Henchoz, S., Gauchat, J.-F., Ellis, J., & Bonnefoy, J.-Y. (1997). CD86 (B7-2) on Human B Cells: A FUNCTIONAL ROLE IN PROLIFERATION AND SELECTIVE DIFFERENTIATION INTO IgE- AND IgG4-PRODUCING CELLS. *Journal of Biological Chemistry*, *272*(25), 15613–15619. doi:10.1074/jbc.272.25.15613

John, L. B., & Ward, A. C. (2011). The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. *Molecular Immunology*, *48*(9-10), 1272–8. doi:10.1016/j.molimm.2011.03.006

Katzav, S., Martin-Zanca, D., & Barbacid, M. (1989). vav, a novel human oncogene derived from a locus ubiquitously expressed in hematopoietic cells. *The EMBO Journal*, *8*(8), 2283–90.

Kerr, J. F., Wyllie, A. H., & Currie, A. R. (1972). Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer*, *26*(4), 239–57.

Klijn, C., Koudijs, M. J., Kool, J., ten Hoeve, J., Boer, M., de Moes, J., … Jonkers, J. (2013). Analysis of tumor heterogeneity and cancer gene networks using deep sequencing of MMTV-induced mouse mammary tumors. *PloS One*, *8*(5), e62113. doi:10.1371/journal.pone.0062113

Kluck, R. (2010). Bcl-2 family-regulated apoptosis in health and disease. *Cell Health and Cytoskeleton*, 9. doi:10.2147/CHC.S6228

Kool, J., Uren, A. G., Martins, C. P., Sie, D., de Ridder, J., Turner, G., … van Lohuizen, M. (2010). Insertional mutagenesis in mice deficient for p15Ink4b, p16Ink4a, p21Cip1, and p27Kip1 reveals cancer gene interactions and correlations with tumor phenotypes. *Cancer Research*, *70*(2), 520–31. doi:10.1158/0008-5472.CAN-09-2736

Koudijs, M. J., Klijn, C., van der Weyden, L., Kool, J., ten Hoeve, J., Sie, D., … Jonkers, J. (2011). High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. *Genome Research*, *21*(12), 2181–9. doi:10.1101/gr.112763.110

Kridel, R., Sehn, L. H., & Gascoyne, R. D. (2012). Pathogenesis of follicular lymphoma. *The Journal of Clinical Investigation*, *122*(10), 3424–31. doi:10.1172/JCI63186

Krönke, J., Udeshi, N. D., Narla, A., Grauman, P., Hurst, S. N., McConkey, M., … Ebert, B. L. (2014). Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science (New York, N.Y.)*, *343*(6168), 301–5. doi:10.1126/science.1244851

Küppers, R. (2005). Mechanisms of B-cell lymphoma pathogenesis. *Nature Reviews. Cancer*, *5*(4), 251–62. doi:10.1038/nrc1589

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., … Wu, C. J. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, *152*(4), 714–26. doi:10.1016/j.cell.2013.01.019

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–8. doi:10.1038/nature12213

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., … Park, P. J. (2012). Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)*, *337*(6097), 967–71. doi:10.1126/science.1222077

Li, H., Kaminski, M. S., Li, Y., Yildiz, M., Ouillette, P., Jones, S., … Malek, S. N. (2014). Mutations in linker histone genes HIST1H1 B, C, D, and E; OCT2 (POU2F2); IRF8; and ARID1A underlying the pathogenesis of follicular lymphoma. *Blood*, *123*(10), 1487–98. doi:10.1182/blood-2013-05-500264

Liao, D. (2009). Emerging roles of the EBF family of transcription factors in tumor suppression. *Molecular Cancer Research : MCR*, *7*(12), 1893–901. doi:10.1158/1541-7786.MCR-09-0229

Liao, X., Du, Y., Morse, H. C., Jenkins, N. A., & Copeland, N. G. (1997). Proviral integrations at the Evi5 locus disrupt a novel 90 kDa protein with homology to the Tre2 oncogene and cell-cycle regulatory proteins. *Oncogene*, *14*(9), 1023–9. doi:10.1038/sj.onc.1200929

Limpens, J., Stad, R., Vos, C., de Vlaam, C., de Jong, D., van Ommen, G. J., … Kluin, P. M. (1995). Lymphoma-associated translocation t(14;18) in blood B cells of normal individuals. *Blood*, *85*(9), 2528–36.

Lowry, L., & Linch, D. (2013). Non-Hodgkin's lymphoma. *Medicine*, *41*(5), 282–289. doi:10.1016/j.mpmed.2013.03.008

Lunardi, A., Nardella, C., Clohessy, J. G., & Pandolfi, P. P. (2014). Of model pets and cancer models: an introduction to mouse models of cancer. *Cold Spring Harbor Protocols*, *2014*(1), 17–31. doi:10.1101/pdb.top069757

Matsuoka, M., & Jeang, K.-T. (2011). Human T-cell leukemia virus type 1 (HTLV-1) and leukemic transformation: viral infectivity, Tax, HBZ and therapy. *Oncogene*, *30*(12), 1379–89. doi:10.1038/onc.2010.537

McDonnell, T. J., Deane, N., Platt, F. M., Nunez, G., Jaeger, U., McKearn, J. P., & Korsmeyer, S. J. (1989). bcl-2-immunoglobulin transgenic mice demonstrate extended B cell survival and follicular lymphoproliferation. *Cell*, *57*(1), 79–88.

McDonnell, T. J., & Korsmeyer, S. J. (1991). Progression from lymphoid hyperplasia to high-grade malignant lymphoma in mice transgenic for the t(14; 18). *Nature*, *349*(6306), 254–6. doi:10.1038/349254a0

Menezes, J., Salgado, R. N., Acquadro, F., Gómez-López, G., Carralero, M. C., Barroso, A., … Cigudosa, J. C. (2013). ASXL1, TP53 and IKZF3 mutations are present in the chronic phase and blast crisis of chronic myeloid leukemia. *Blood Cancer Journal*, *3*, e157. doi:10.1038/bcj.2013.54

Molnár, A., Wu, P., Largespada, D. A., Vortkamp, A., Scherer, S., Copeland, N. G., … Georgopoulos, K. (1996). The Ikaros gene encodes a family of lymphocyte-restricted zinc finger DNA binding proteins, highly

conserved in human and mouse. *Journal of Immunology (Baltimore, Md. : 1950)*, *156*(2), 585–92.

Morin, R. D., Mendez-Lago, M., Mungall, A. J., Goya, R., Mungall, K. L., Corbett, R. D., … Marra, M. A. (2011). Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, *476*(7360), 298–303. doi:10.1038/nature10351

Morin, R. D., Mungall, K., Pleasance, E., Mungall, A. J., Goya, R., Huff, R. D., … Marra, M. A. (2013). Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood*, *122*(7), 1256–65. doi:10.1182/blood-2013-02-483727

Motulsky, H. (2013). *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking* (p. 540). Oxford University Press.

Mullighan, C. G., Miller, C. B., Radtke, I., Phillips, L. A., Dalton, J., Ma, J., … Downing, J. R. (2008). BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature*, *453*(7191), 110–4. doi:10.1038/nature06866

Munro, J. M., Freedman, A. S., Aster, J. C., Gribben, J. G., Lee, N. C., Rhynhart, K. K., … Nadler, L. M. (1994). In vivo expression of the B7 costimulatory molecule by subsets of antigen-presenting cells and the malignant cells of Hodgkin's disease. *Blood*, *83*(3), 793–8.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., … Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*(7341), 90–4. doi:10.1038/nature09807

Nethe, M., Berkhout, B., & van der Kuyl, A. C. (2005). Retroviral superinfection resistance. *Retrovirology*, *2*(1), 52. doi:10.1186/1742-4690-2-52

Nowell, P. (1976). The clonal evolution of tumor cell populations. *Science*, *194*(4260), 23–28. doi:10.1126/science.959840

O'Brien, S. M., Claxton, D. F., Crump, M., Faderl, S., Kipps, T., Keating, M. J., … Cheson, B. D. (2009). Phase I study of obatoclax mesylate (GX15-070), a small molecule pan-Bcl-2 family antagonist, in patients with advanced chronic lymphocytic leukemia. *Blood*, *113*(2), 299–305. doi:10.1182/blood-2008-02-137943

Ogilvy, S., Metcalf, D., Gibson, L., Bath, M. L., Harris, A. W., & Adams, J. M. (1999). Promoter elements of vav drive transgene expression in vivo throughout the hematopoietic compartment. *Blood*, *94*(6), 1855–63.

Ogilvy, S., Metcalf, D., Print, C. G., Bath, M. L., Harris, A. W., & Adams, J. M. (1999). Constitutive Bcl-2 expression throughout the hematopoietic compartment affects multiple lineages and enhances progenitor cell survival. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(26), 14943–8.

Oki, Y., Copeland, A., Hagemeister, F., Fayad, L. E., Fanale, M., Romaguera, J., & Younes, A. (2012). Experience with obatoclax mesylate (GX15-070), a small molecule pan-Bcl-2 family antagonist in patients with relapsed or refractory classical Hodgkin lymphoma. *Blood*, *119*(9), 2171–2. doi:10.1182/blood-2011-11-391037

Okosun, J., Bödör, C., Wang, J., Araf, S., Yang, C.-Y., Pan, C., … Fitzgibbon, J. (2014). Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nature Genetics*, *46*(2), 176–81. doi:10.1038/ng.2856

Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., … Rosenberg, S. H. (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, *435*(7042), 677–81. doi:10.1038/nature03579

Paik, P. K., Rudin, C. M., Pietanza, M. C., Brown, A., Rizvi, N. A., Takebe, N., … Krug, L. M. (2011). A phase II study of obatoclax mesylate, a Bcl-2 antagonist, plus topotecan in relapsed small cell lung cancer. *Lung Cancer (Amsterdam, Netherlands)*, *74*(3), 481–5. doi:10.1016/j.lungcan.2011.05.005

Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., … Dalla-Favera, R. (2011). Analysis of the coding genome of diffuse large B-cell lymphoma. *Nature Genetics*, *43*(9), 830–837. doi:10.1038/ng.892

Pellegrini, A., Guiñazú, N., Aoki, M. P., Calero, I. C., Carrera-Silva, E. A., Girones, N., … Gea, S. (2007). Spleen B cells from BALB/c are more prone to activation than spleen B cells from C57BL/6 mice during a secondary immune response to cruzipain. *International Immunology*, *19*(12), 1395–402. doi:10.1093/intimm/dxm107

Podojil, J. R., Kin, N. W., & Sanders, V. M. (2004). CD86 and beta2-adrenergic receptor signaling pathways, respectively, increase Oct-2 and OCA-B Expression and binding to the 3'-IgH enhancer in B cells. *The Journal of Biological Chemistry*, *279*(22), 23394–404. doi:10.1074/jbc.M313096200

Raghavan, S. C., Swanson, P. C., Wu, X., Hsieh, C.-L., & Lieber, M. R. (2004). A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. *Nature*, *428*(6978), 88–93. doi:10.1038/nature02355

Rebollo, A., Ayllón, V., Fleischer, A., Martínez, C. A., & Zaballos, A. (2001). The association of Aiolos transcription factor and Bcl-xL is involved in the control of apoptosis. *Journal of Immunology (Baltimore, Md. : 1950)*, *167*(11), 6366–73.

Rebollo, A., & Schmitt, C. (2003). Ikaros, Aiolos and Helios: transcription regulators and lymphoid malignancies. *Immunology and Cell Biology*, *81*(3), 171–5. doi:10.1046/j.1440-1711.2003.01159.x

Rein, A. (2011). Murine leukemia viruses: objects and organisms. *Advances in Virology*, *2011*, 403419. doi:10.1155/2011/403419

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., … Markham, A. F. (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Research*, *18*(10), 2887–90.

Risser, R. E. X., Kaehler, D., & Lamph, W. W. (1985). Different genes control the susceptibility of mice to Moloney or Abelson murine leukemia Different Genes Control the Susceptibility of Mice to Moloney Abelson Murine Leukemia Viruses, *55*(3).

Romero, F., Martínez-A, C., Camonis, J., & Rebollo, A. (1999). Aiolos transcription factor controls cell death in T cells by regulating Bcl-2 expression and its cellular localization. *The EMBO Journal*, *18*(12), 3419–30. doi:10.1093/emboj/18.12.3419

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., … Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, *346*(25), 1937–47. doi:10.1056/NEJMoa012914

Ross, C. W., Ouillette, P. D., Saddler, C. M., Shedden, K. A., & Malek, S. N. (2007). Comprehensive analysis of copy number and allele status identifies multiple chromosome defects underlying follicular lymphoma pathogenesis. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *13*(16), 4777–85. doi:10.1158/1078-0432.CCR-07-0456

Sansom, D. M., Manzotti, C. N., & Zheng, Y. (2003). What's the difference between CD80 and CD86? *Trends in Immunology*, *24*(6), 313–318. doi:10.1016/S1471-4906(03)00111-X

Sauvageau, M., Miller, M., Lemieux, S., Lessard, J., Hébert, J., & Sauvageau, G. (2008). Quantitative expression profiling guided by common retroviral insertion sites reveals novel and cell type specific cancer genes in leukemia. *Blood*, *111*(2), 790–9. doi:10.1182/blood-2007-07-098236

Schmidt, M., Zickler, P., Hoffmann, G., Haas, S., Wissler, M., Muessig, A., … von Kalle, C. (2002). Polyclonal long-term repopulating stem cell clones in a primate model. *Blood*, *100*(8), 2737–43. doi:10.1182/blood-2002-02-0407

Schmitt, C., Balogh, B., Grundt, A., Buchholtz, C., Leo, A., Benner, A., … Leo, E. (2006). The bcl-2/IgH rearrangement in a population of 204 healthy individuals: occurrence, age and gender distribution, breakpoints, and detection method validity. *Leukemia Research*, *30*(6), 745–50. doi:10.1016/j.leukres.2005.10.001

Sharpe, A. H., & Freeman, G. J. (2002). The B7-CD28 superfamily. *Nature Reviews. Immunology*, *2*(2), 116–26. doi:10.1038/nri727

Shibata, D., Weiss, L. M., Hernandez, A. M., Nathwani, B. N., Bernstein, L., & Levine, A. M. (1993). Epstein-Barr virus-associated non-Hodgkin's lymphoma in patients infected with the human immunodeficiency virus. *Blood*, *81*(8), 2102–9.

Shinto, Y., Morimoto, M., Katsumata, M., Uchida, A., Aozasa, K., Okamoto, M., … Tsujimoto, Y. (1995). Moloney murine leukemia virus infection accelerates lymphomagenesis in E mu-bcl-2 transgenic mice. *Oncogene*, *11*(9), 1729–36.

Shtutman, M., Baig, M., Levina, E., Hurteau, G., Lim, C.-U., Broude, E., … Roninson, I. B. (2011). Tumor-specific silencing of COPZ2 gene encoding coatomer protein complex subunit ζ 2 renders tumor cells

dependent on its paralogous gene COPZ1. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(30), 12449–54. doi:10.1073/pnas.1103842108

Skibola, C. F., Berndt, S. I., Cerhan, J. R., Wang, Z., Vijai, J., Conde, L., … NHL GWAS Consortium. (2014). Abstract 5072: Meta-analysis of genome-wide association studies identifies novel susceptibility loci for follicular lymphoma. *Cancer Res.*, *74*(19_Supplement), 5072–. doi:10.1158/1538-7445.AM2014-5072

Skibola, C. F., Curry, J. D., & Nieters, A. (2007). Genetic susceptibility to lymphoma. *Haematologica*, *92*(7), 960–9.

Souers, A. J., Leverson, J. D., Boghaert, E. R., Ackler, S. L., Catron, N. D., Chen, J., … Elmore, S. W. (2013). ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine*, *19*(2), 202–8. doi:10.1038/nm.3048

Stevens, T. L., Bossie, A., Sanders, V. M., Fernandez-Botran, R., Coffman, R. L., Mosmann, T. R., & Vitetta, E. S. (1988). Regulation of antibody isotype secretion by subsets of antigen-specific helper T cells. *Nature*, *334*(6179), 255–8. doi:10.1038/334255a0

Strasser, a, Whittingham, S., Vaux, D. L., Bath, M. L., Adams, J. M., Cory, S., & Harris, a W. (1991). Enforced BCL2 expression in B-lymphoid cells prolongs antibody responses and elicits autoimmune disease. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(19), 8661–5.

Strasser, A., Harris, A. W., & Cory, S. (1993). E mu-bcl-2 transgene facilitates spontaneous transformation of early pre-B and immunoglobulin-secreting cells but not T cells. *Oncogene*, *8*(1), 1–9.

Strasser, A., Harris, A. W., Vaux, D. L., Webb, E., Bath, M. L., Adams, J. M., & Cory, S. (1990). Abnormalities of the immune system induced by dysregulated bcl-2 expression in transgenic mice. *Current Topics in Microbiology and Immunology*, *166*, 175–81.

Suvas, S., Singh, V., Sahdev, S., Vohra, H., & Agrewala, J. N. (2002). Distinct role of CD80 and CD86 in the regulation of the activation of B cell and B cell lymphoma. *The Journal of Biological Chemistry*, *277*(10), 7766–75. doi:10.1074/jbc.M105902200

Swanton, C. (2014). Cancer evolution: the final frontier of precision medicine? *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO*, *25*(3), 549–51. doi:10.1093/annonc/mdu005

Tse, C., Shoemaker, A. R., Adickes, J., Anderson, M. G., Chen, J., Jin, S., … Elmore, S. W. (2008). ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. *Cancer Research*, *68*(9), 3421–8. doi:10.1158/0008-5472.CAN-07-5836

Tsichlis, P. N., Strauss, P. G., & Hu, L. F. (1983). A common region for proviral DNA integration in MoMuLV-induced rat thymic lymphomas. *Nature*, *302*(5907), 445–449. doi:10.1038/302445a0

Tsujimoto, Y., Gorham, J., Cossman, J., Jaffe, E., & Croce, C. M. (1985). The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining. *Science (New York, N.Y.)*, *229*(4720), 1390–3.

Uren, A. G., Kool, J., Berns, A., & van Lohuizen, M. (2005). Retroviral insertional mutagenesis: past, present and future. *Oncogene*, *24*(52), 7656–72. doi:10.1038/sj.onc.1209043

Uren, A. G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A. H., Wilson, C. H., … Adams, D. J. (2009). A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nature Protocols*, *4*(5), 789–98. doi:10.1038/nprot.2009.64

Van Delft, M. F., Wei, A. H., Mason, K. D., Vandenberg, C. J., Chen, L., Czabotar, P. E., … Huang, D. C. S. (2006). The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. *Cancer Cell*, *10*(5), 389–99. doi:10.1016/j.ccr.2006.08.027

Walboomers, J. M., Jacobs, M. V, Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V, … Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *The Journal of Pathology*, *189*(1), 12–9. doi:10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F

Wang, J. H., Gostissa, M., Yan, C. T., Goff, P., Hickernell, T., Hansen, E., … Alt, F. W. (2009). Mechanisms promoting translocations in editing and

switching peripheral B cells. *Nature*, *460*(7252), 231–6. doi:10.1038/nature08159

Wang, J.-H., Avitahl, N., Cariappa, A., Friedrich, C., Ikeda, T., Renold, A., … Georgopoulos, K. (1998). Aiolos Regulates B Cell Activation and Maturation to Effector State. *Immunity*, *9*(4), 543–553. doi:10.1016/S1074-7613(00)80637-8

Weinblatt, M. E., Moreland, L. W., Westhovens, R., Cohen, R. B., Kelly, S. M., Khan, N., … Hochberg, M. C. (2013). Safety of abatacept administered intravenously in treatment of rheumatoid arthritis: integrated analyses of up to 8 years of treatment from the abatacept clinical trial program. *The Journal of Rheumatology*, *40*(6), 787–97. doi:10.3899/jrheum.120906

Willis, T. G., & Dyer, M. J. S. (2000). The role of immunoglobulin translocations in the pathogenesis of B-cell malignancies. *Blood*, *96*(3), 808–822.

Wilson, W. H., Hernandez-Ilizaliturri, F. J., Dunleavy, K., Little, R. F., & O'Connor, O. A. (2010). Novel disease targets and management approaches for diffuse large B-cell lymphoma. *Leukemia & Lymphoma*, *51 Suppl 1*, 1–10. doi:10.3109/10428194.2010.500045

Wilson, W. H., O'Connor, O. A., Czuczman, M. S., LaCasce, A. S., Gerecitano, J. F., Leonard, J. P., … Humerickhouse, R. A. (2010). Navitoclax, a targeted high-affinity inhibitor of BCL-2, in lymphoid malignancies: a phase 1 dose-escalation study of safety, pharmacokinetics, pharmacodynamics, and antitumour activity. *The Lancet Oncology*, *11*(12), 1149–59. doi:10.1016/S1470-2045(10)70261-8

Xu, W., & Kee, B. L. (2007). Growth factor independent 1B (Gfi1b) is an E2A target gene that modulates Gata3 in T-cell lymphomas. *Blood*, *109*(10), 4406–14. doi:10.1182/blood-2006-08-043331

Yoshida, M. (2005). Discovery of HTLV-1, the first human retrovirus, its unique regulatory mechanisms, and insights into pathogenesis. *Oncogene*, *24*(39), 5931–7. doi:10.1038/sj.onc.1208981

Yoshida, M., Miyoshi, I., & Hinuma, Y. (1982). Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(6), 2031–5.

Yuanxin, Y., Chengcai, A., Li, L., Jiayu, G., Guihong, T., & Zhangliang, C. (2003). T-linker-specific ligation PCR (T-linker PCR): an advanced PCR technique for chromosome walking or for isolation of tagged DNA ends. *Nucleic Acids Research*, *31*(12), e68.

Zagaria, A., Anelli, L., Coccaro, N., Casieri, P., Minervini, A., Buttiglione, V., … Albano, F. (2012). A new recurrent chromosomal translocation t(3;11)(q13;q14) in myelodysplastic syndromes associated with overexpression of the ILDR1 gene. *Leukemia Research*, *36*(7), 852–6. doi:10.1016/j.leukres.2012.01.026

Zhao, F., McCarrick-Walmsley, R., Akerblad, P., Sigvardsson, M., & Kadesch, T. (2003). Inhibition of p300/CBP by early B-cell factor. *Molecular and Cellular Biology*, *23*(11), 3837–46.