A Study of Molecular Synergy and Clonal Evolution in Haematopoietic Malignancies

Carolyn Suzanne Grove

Emmanuel College, Cambridge

August 2014

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of work undertaken in the laboratory of Dr George Vassiliou at the Wellcome Trust Sanger Institute. The dissertation is the result of my own work, except where specific reference is made to the work of others. Where data is the result of collaboration with others, this is clearly stated as such in the text. This work has not previously been submitted for any other degree or qualification. The text of this dissertation, excluding tables, figures and references, does not exceed 60 000 words.

Carolyn Grove
August 2014

Acknowledgements

I would like to thank my supervisor, George Vassiliou, who has been an enthusiastic and encouraging mentor, role model and friend. I consider it a privilege to have been his first PhD student and to be one of the early members of his laboratory, which is a dynamic, stimulating and pleasurable place to work.

I would also like to thank my co-supervisor Martin Turner, the other members of my thesis committee, Peter Campbell and Allan Bradley, and my 'unofficial' supervisor Roland Rad. All have given me considerable time, encouragement and support and I have benefited from their diverse expertise and willingness to help me to develop my skills.

This thesis work was completed with considerable support from other members of team 163 at the Wellcome Trust Sanger Institute (WTSI). I would particularly like to thank Jon Cooper, Meg Byrne and Sarah Paterson, all of whom generously assisted with the routine aspects of this work, in particular mouse genotyping, necropsies, blood analysis, tissue storage and DNA extraction. Jon taught me many laboratory skills and has been an enthusiastic helper and good friend. Meg looked after my mouse colonies during my period of maternity leave and Sarah kindly assisted me during two summer vacations, while also studying for her medical degree at Glasgow University. I would also like to thank the other members of team 163, all of whom provided me with useful insights and helped to generate such an enjoyable atmosphere in which to work.

There are several people who helped with particular aspects of this work, which are recognised throughout the text, but I would also like to acknowledge them here. Gary Hoffman, Iraad Bronner, Kosuke Yusa, Nicla Manes, Stephen Rice, Hannes Ponstingl, Ignacio Varela, David Wedge and Rachael Bashford-Rogers have all been tremendously supportive. Also, this work would not have been possible without the generous support of Leukaemia Lymphoma Research. I will always be grateful to this charity for the wonderful opportunity they have given me.

Finally, I would like to thank my family, in particular my husband Jonathan Francis, who has assisted in so many ways. This dissertation is much more visually attractive than it would have been without his help in formatting the many tables. It would not exist at all, without the strength, love and support that he has given me. I would also like to thank my mother-in-law Annette, who has been a source of inspiration in her resilience, especially over recent months. In addition I would like to thank Catherine and Malcolm, my sister- and father-in-law who supported me through so much and who shall always be remembered for their generosity of spirit. I also want to thank my parents, who have always been there for me and have willingly supported me in every endeavour. Finally, to Matthew, who came into the world during the course of this work and has kept a smile on my face, even during the toughest periods.

Abstract

Haematopoeitic malignancies evolve through the serial selection of cells with a growth advantage, in a multi-step process akin to natural selection. Transposon insertional mutagenesis (IM) is a powerful approach for the identification and validation of cancer driver mutations and compliments human sequencing efforts. This technology has not previously been applied to study tumour evolution, nor has the sub-clonal architecture of transposon driven tumours been carefully investigated.

In the first part of this work I have investigated the timing and pattern of acquisition of mutations in *NPM1*-mutant acute myeloid leukaemia (AML). *NPM1* mutations are found in around 30% of cases of AML and are thought to be critical events in leukaemogenesis. First, I present the detailed study of an informative human case of CMML evolving to AML and discuss the implications for clonal evolution and leukaemic transformation. Subsequently, I describe the investigation of an IM mouse model of *Npm1*-mutant AML in which the timing and order of acquisition of transposon integrations was characterised using pre-leukaemic blood samples. The driver status and co-occurrence of integrations was also investigated in serial transplant experiments. Transposon mobilisation continued throughout leukaemia evolution, but this data suggests that only a minority of integrations behave as 'driver' mutations in this context. Although some of these 'drivers' were detectable several weeks earlier, the onset of leukaemia was sudden and occurred without antecedent abnormalities in blood count parameters. Transplant experiments demonstrated that multiple distinct clones with different transposon integrations were present within the primary tumour cell population.

In the final part of this dissertation I present the findings of two mouse models in which *piggyBac* (*PB*) IM is targeted to the mature B cell compartment for cancer gene discovery. Both models were based on the published *Vk*MYC* mice, which were reported to develop highly penetrant plasma cell malignancies recapitulating the major features of human multiple myeloma. In one model, the *PB* transposase replaced the *MYC* transgene in the *Vk*MYC construct*. In the second, *MYC* and *PB* were co-expressed from the same cistron, in order to identify genes co-operating with *MYC* in oncogenesis. IM mice had a significantly reduced survival largely due to the development of mature B cell lymphomas; although plasma cell malignancies were not a feature. Mapping and common integration site analysis of transposon

insertions identified several recurrent integrations in known (e.g. *Bcl6*) and novel (e.g. *Rreb1*) lymphoma-associated genes.

Contents

Chapter 1: Introduction

1.1 Cancer as an evolutionary process

Peter Nowell was the first to describe cancer as an evolutionary process with parallels to Darwinian natural selection (Nowell, 1976). Complex organisms have evolved highly efficient systems to protect their cellular genomes from accumulating DNA mutations. However, such mechanisms are not impenetrable and cells slowly accumulate mutations over time even in the absence of identifiable exogenous mutagens. Carcinogenesis involves the serial selection of cells with a growth advantage, in a multi-step process akin to evolution by natural selection. Just like Darwinian evolution, the progression is not linear, but usually leads to the generation of multiple clades downstream of a single ancestor, the cell with a "cancer-initiating" mutation.

The change from a normal to a cancer cell requires acquisition of multiple somatic mutations which impart the malignant phenotype. The potential for limitless self-renewal is one of the hallmarks of cancer (Hanahan and Weinberg, 2000) although it is recognised that this capacity is often restricted to a sub-population of tumour cells; the cancer or leukaemia stem cells (CSC/LSC) (Lapidot et al., 1994). Individual cancer genomes are genetically heterogeneous. It follows that if LSCs drive sustained clonal expansion and disease progression, then these must also be genetically diverse. There is evidence that this is the case in acute lymphoblastic leukaemia (ALL). Transplantation of primary leukaemia cells into immune deficient mice revealed variable competitive regeneration of sub-clones in patterns reflecting the diversity within the primary tumour (Anderson et al., 2011; Notta et al., 2011).

Haematopoietic (HSC), like other normal stem cells are undifferentiated, long-lived cells capable of asymmetric division, facilitating both self-renewal and the generation of differentiated progeny in very large numbers. Haematopoiesis is normally polyclonal with contribution from a small proportion of all HSCs. During homeostasis normal peripheral blood is estimated to have contributions from approximately 1000 HSC (Catlin et al., 2011), but the majority of adult HSC are in a quiescent state (Arai et al., 2004; Li and Clevers, 2010). The signals that drive a G0 HSC to enter into cell cycle are not understood. It may be that this is a largely stochastic process (McKenzie et al.,

2006). On average human HSCs are thought to divide once every 40 weeks (Catlin et al., 2011), however blood cell production is a continuous process throughout life with an adult human producing an estimated 10^{11} cells daily (Beerman et al., 2010). Adult HSCs, like other tissue stem cells, are prime candidates for malignant transformation as they have inherent self-renewal capacity and persist throughout life. Nevertheless, the fact that some mutations can transform differentiated cells, suggests that HSCs may not be the unique source of LSCs (Cozzio et al., 2003; Huntly et al., 2004).

Typically hundreds to thousands of somatic mutations are identified in genomic DNA from most adult tumours. The mutations present in a cancer cell genome accumulate throughout the life of a patient and are the result of exposure to external mutagens, as well as cell-intrinsic mutational processes, such as errors in DNA replication or illegitimate action of DNA editing enzymes (Papaemmanuil et al., 2014). The median number of somatic mutations differs by more than 1000-fold between different types of human cancer (Alexandrov et al., 2013; Lawrence et al., 2013). It is estimated that about half of the variation in mutation frequencies can be explained by the difference in somatic mutation rates between tissues (Lawrence et al., 2013), however the number of somatic mutations can also vary by over 1000-fold between different cancers of the same subtype (Alexandrov et al., 2013; Lawrence et al., 2013). AML has one of the lowest number of mutations per case of any adult cancer studied to date (figure 1.1), yet the range varies by more than 100-fold between individual cases (Lawrence et al., 2013; TCGA_Research_Network, 2013).

The number of driver mutations that co-operate to induce a malignant phenotype is not well established and appears to differ between tumours. It is estimated that in common adult epithelial tumours there are 5-7 driver mutations, while in haematopoietic malignancies this number is thought to be lower (Stratton et al., 2009). Some of the difference is likely to be attributable to the pattern and intensity of the mutational processes underlying each cancer type rather than representing an intrinsic cellular characteristic. For example, a cancer arising through rare "background" stochastic mutations may be more likely to arise via a small number of powerful mutations, whilst one in which mutagenesis is avid may evolve through a larger number of weak mutations. If this were true one would usually expect the

former type to be rarer than the later and observations on the total number of mutations in different cancer types appear to broadly support this thesis (figure 1.1).

The binary classification of mutations into drivers and passengers is context dependent. Tumour sub-clones compete with each other and with normal cells for "real estate" and resources within the tissue microenvironment. Changes imposed on this ecosystem will alter the relative competitiveness of cells/clones. Mutations that in isolation have a neutral or even negative effect on long-term clonogenicity (passenger) may be "selected" if they co-occur with a fitness conferring mutation or are advantageous in the context of other mutations (epistatic effect). The highly variable number of passenger lesions both between and within sub-types of cancer reflects the dynamics of clonal evolution (Nik-Zainal et al., 2012; Welch et al., 2012). Factors that affect the number of passenger mutations in the final tumour include i) variation in the number of cellular divisions between the germline and the sequenced cancer cell ii) differences in susceptibility to somatic mutation iii) the fidelity of intrinsic DNA repair mechanisms and iv) differential exposure to mutagens. A major challenge for researchers is to distinguish the few driver mutations from the multitude of passengers within a cancer genome.

The explosion in cancer genomics has led to the identification of innumerable somatic mutations, most of which are functionally unexplored. As such their *driver* vs *passenger* status remains formally untested. For the time being, their recurrence rate within and between cancer types serves as a proxy for this status; i.e. genes mutated in cancer more often than expected by chance are considered to be *drivers*. This is very likely to be an oversimplification as "chance" is difficult to determine. For example some very large genes are recurrently mutated by virtue of their size and others by virtue of their chromatin organisation (Lawrence et al., 2013). Also, it is plausible that some presumed *drivers* function to accelerate mutagenesis rather than to impart improved fitness. It is highly likely that some true *drivers* have not been identified as such yet, because not enough cancers of their type have been studied or because their genomic location or specific sequence context leaves them relatively resistant to common mutational processes or prevents their capture/identification by current sequencing methods. It is also probable that the number and type of mutations that can confer a fitness advantage is highly variable between genes.

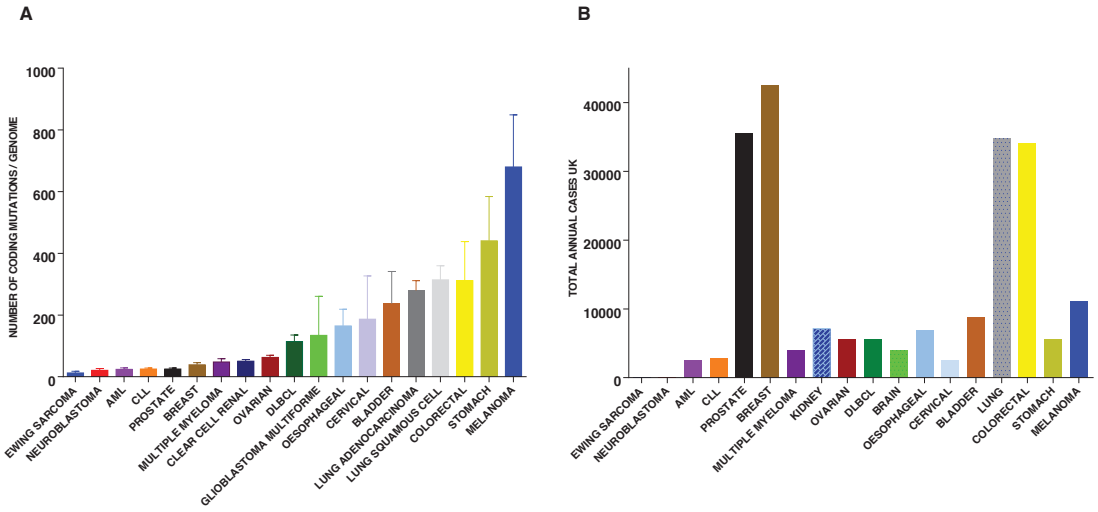


FIGURE 1.1: MUTATION BURDEN AND CANCER INCIDENCE (A) A comparison of the mean number of non-coding mutations per genome across various tumour types. The raw data is taken from Lawrence et al, 2013. Error bars show the standard error of the mean. **(B)** UK annual incident cases of various malignancies taken from the Cancer Registry Statistics (2011) (http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-302299) and Cancer Research UK (http://www.cancerresearchuk.org/cancer-info/cancerstats/). Patterned bars depict the incidence for the entire tissue rather than the specific cancer sub-type shown in A (eg lung cancer rather than lung adenocarcinoma)(Grove and Vassiliou, 2014).

The importance of timing and genetic context in identifying driver mutation status is exemplified by transient myeloproliferative disease of the newborn (TMD) and Down syndrome associated acute megakaryocytic leukaemia (DS-AMKL). TMD develops in up to 10% of newborn infants with Down syndrome, with most presenting in the first week of life (Gamis et al., 2011; Malinge et al., 2009; Pine et al., 2007). Affected children develop a megakaryocytic leukaemia, which typically spontaneously regresses within three months (Malinge et al., 2009). Around 20% of children who had TMD will develop DS-AMKL by the age of four and this occurs when the dormant TMD clone accumulates additional leukaemogenic mutations, although the mean number of somatic mutations in DS-AMKL is still much lower than in most other cancers (Yoshida et al., 2013). Intratumoral heterogeneity in mutations is described in both TMD and DS-AMKL and progression to leukaemia originating from major or minor TMD sub-clones has been reported (Yoshida et al., 2013). Both TMD and DS-AMKL are associated with truncating mutations in *GATA1* that arise in utero (Malinge et al., 2009; Nikolaev et al., 2013; Pine et al., 2007). Exome sequencing studies suggest

that the combination of trisomy 21 with a truncating *GATA1* mutation is sufficient to cause TMD, although additional putative driver mutations may also be seen, without disease progression to DS-AMKL (Nikolaev et al., 2013; Yoshida et al., 2013). Presumably the changes induced by trisomy 21 render cells susceptible to additional transforming events and/or alter the phenotypic consequences of these events as germ-line *GATA1* mutations in the absence of trisomy 21 do not associate with leukaemia (Malinge et al., 2009) and this mutation is found in only around 10% of AMKL cases in the absence of Down Syndrome (Gruber et al., 2012). Trisomy 21 is also associated with genome-wide hypomethylation and additional methylation abnormalities are detected at the TMD stage, although it is uncertain if the epigenetic changes reflect the genetic lesions or contribute to disease (Malinge et al., 2013). The transcriptional and epigenetic programs of TMD and DS-AMKL are very similar(Malinge et al., 2013).

1.2 AML as an exemplar of clonal evolution

1.2.1 How many driver mutations are required for leukaemogenesis?

In AML, there is a relatively well-defined group of recurrent mutations, most of which fall into functional categories (figure 1.2) (TCGA_Research_Network, 2013). The variation in the identity of co-occurring driver mutations is in keeping with the stochastic nature of myeloid leukaemogenesis, yet the identifiable patterns of co-occurrence and mutual exclusivity between specific mutations hint respectively to molecular synergy and redundancy between them (TCGA_Research_Network, 2013).

Gilliland and Griffin proposed the two hit model of leukaemogenesis (Gilliland and Griffin, 2002). In their model two mutations each belonging to a different class, collaborate to cause AML when neither is sufficient to do so in isolation. Class I mutations such as *FLT3-ITD* or *N-RAS* mutations confer a proliferative advantage but have no effect on differentiation. Class II mutations, represented by specific fusion genes in the original model impair haematopoietic differentiation and subsequent apoptosis. The initiating lesions in these AMLs are thought to be Class II mutations, for example *PML/RARα* and *MLL* fusions, whereas Class I mutations are typically later events. This model has provided a useful framework to conceptualise the pathogenesis of AML as a disease in which differentiation is blocked and proliferation is increased. Although most of the recently identified mutations do not fit

neatly into one of the two classes, they are thought to collaboratively produce the equivalent effects leading to the AML phenotype.

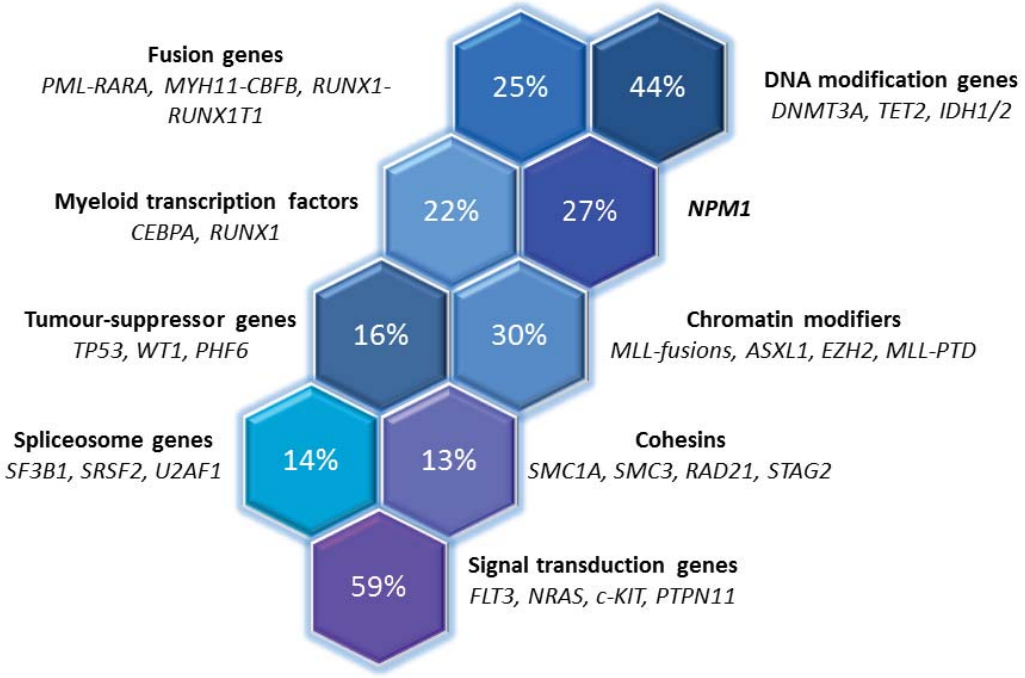


Figure 1.2: RECURRENT MUTATION GROUPS IN AML. Data on the prevalence of different mutation groups in AML (based on data from TCGA Research Network, 2013) (Grove and Vassiliou, 2014).

The number of identifiable driver mutations differs between AML cases. In a study of 200 AMLs using whole genome and whole exome sequencing the authors describe a mean of 13 (range 0-51) tier 1 (coding, splice-site and RNA gene) mutations (TCGA_Research_Network, 2013). On average five of these were in genes which are recurrently mutated in AML and thus presumed to represent driver events. The number of recurrent tier 1 mutations was lower in the presence of specific translocations while higher numbers were observed in cases with *RUNX1-RUNX1T1* fusions and those without fusion genes(TCGA_Research_Network, 2013). Co-occurrence analysis showed some common mutations such as *NPM1*, *DNMT3A*, *CEBPA*, *IDH1/2* and *RUNX1* were mutually exclusive of the transcription factor fusions and the authors proposed that these mutations may have a role in the initiation of AML (TCGA_Research_Network, 2013).

Although difficult to validate, evidence from mouse models suggests that as few as two highly complementary mutations may be sufficient to generate leukaemia (Mupo

et al., 2013; Wartman et al., 2011). In a knock-in mouse model, the combination of *Npm1c* and *Flt3-ITD* caused universal leukaemia, with all mice becoming moribund with AML in 31-68 days (Mupo et al., 2013). In another model the co-expression of *PML-RARα* and *Jak1 V657F* mutations in mice resulted in a rapid onset of acute promyelocytic leukaemia (APL) like leukaemia with a mean latency of 35 days (range 28-52 days) (Wartman et al., 2011). Compared to single mutant controls, both models demonstrated a markedly accelerated disease, increased penetrance and a change in phenotype in the double mutant mice. Although these observations suggest that specific combinations of two mutations may be sufficient to drive AML, the possibility that additional mutations are rapidly acquired even within such short latencies cannot be ruled out. In fact, in the former model most AMLs displayed acquired loss-of-heterozygosity for *Flt3-ITD*.

Similarly, human sequencing data describes many AMLs with only one or two identifiable driver mutations. Whole genome sequencing of twelve human samples of APL included one case in which *FLT3-ITD* and *PML-RARα* were the only recurrent cancer- or AML-associated tier 1 somatic mutations expressed in the tumour (Welch et al., 2012). In a mouse model, *PML-RARα* and *FLT3-ITD* induced an APL-like disease with complete penetrance and a short latency which is consistent with the hypothesis that these two mutations are sufficient for disease development (Kelly et al., 2002). Interpreting human sequencing data is compounded by the real possibility that driver mutations were missed or misclassified as passengers because of their rarity. The possibility that additional non-recurrent driver mutations contributed to the pathogenesis cannot be excluded and in a further four cases of APL with these mutations additional cancer associated tier 1 somatic mutations were identified. Additionally, in support of the premise that driver mutations may be missed, examples of AML with only one identifiable AML-recurrent mutation in the whole genome were described more recently (TCGA_Research_Network, 2013). Nevertheless, it remains possible that specific combinations of two mutations may be sufficient for leukaemogenesis, although most cases harbour three or more identifiable drivers at the time of clinical presentation (Welch et al., 2012).

1.2.2 Genotype Phenotype Correlations and Myeloid Malignancy

Many common mutations driving myeloid neoplasms are found in several phenotypically distinct diseases. For example, *TET2* mutations are found recurrently in AML, MDS, MPD and CMML as well as occurring in lymphoid tumours(Delhommeau et al., 2009; Quivoron et al., 2011). This raises two important questions; first to what extent can the disease phenotype be deduced from its complement of somatic mutations and second, how do shared initiating mutations evolve into distinct neoplasms.

Although the LSC is the cell of origin for AML, selective pressures are applied to tumour cells at all stages of differentiation in the mixed tumour population. Itzykson et al analysed candidate genes in single-cell-derived colonies from CMML patients to characterise the distribution of mutations at various stages of progenitor differentiation (Itzykson et al., 2013b). Sub-clones with a greater number of mutations were over-represented in the granulocyte-monocyte progenitors (GMP) compared to the HSC/multipotent progenitor (MPP) compartment, even though CMML is a disease of HSC origin and clonal dominance of the malignant clone is evident at the HSC/MPP stage(Itzykson et al., 2013b). Therefore, it appears these mutations present in only some of the LSCs, provide an additional clonal advantage to differentiating progeny. A comparison of *TET2* mutant CMML and MDS samples found the peripheral monocyte count correlated with the proportion of *TET2* mutated CD34+/CD38- cells suggesting that the extent of dominance of the *TET2* mutated clone in the HSC/MPP compartments influences the clinical phenotype (Itzykson et al., 2013b). However, the serial analysis of samples from individual patients also provided evidence that changes in the clonal composition of the HSC/MPP compartment are not always evident in the disease phenotype. For example, some patients showed a significant increase in the proportion of double mutant HSC/MPP clones over time even though the clinical phenotype was unchanged (Itzykson et al., 2013b).

Together such findings indicate that varied selective pressures and fitness determinants drive clonal outgrowth at different stages of the myeloid stem and progenitor cell hierarchy. This is relevant to sequencing studies as the distribution of mutations detected in the mass tumour population will not necessarily reflect their frequency in LSCs. Furthermore, when evaluating treatment it is important to

recognise that therapies which remove the proliferative advantage of a sub-clone during differentiation may have a phenotypic benefit, but will not necessarily have the same impact on LSCs.

1.2.3 Linear Versus Branching Evolution and Clonal Hierarchy

Cancer dynamics depend on the rate of acquisition of fitness conferring mutations, the relative selective advantage they give and the size of the susceptible cell population. A mutation that confers a strong selective advantage could allow a clone to expand and dominate the haematopoietic compartment in a 'selective sweep' especially if there is a long lag time before additional driver mutations occur. With sequential dominant clones leukaemia evolution would be represented by an essentially linear architecture with stepwise accumulation of driver mutations (figure 1.3A). However, deep sequencing methods have revealed that cancers, including AML, are characterised by significant mutational complexity and that the diversity and relative dominance of sub-clones varies throughout the course of disease (Anderson et al., 2011; Campbell et al., 2008; Campbell et al., 2010; Ding et al., 2012; Gerlinger et al., 2012; Nik-Zainal et al., 2012; Notta et al., 2011). The sub-clones with the highest numbers of genetic abnormalities are not necessarily numerically dominant within the tumour (Anderson et al., 2011; Jan et al., 2012; Walter et al., 2012). Cancers can be traced back to a single cell, but the continuous acquisition of mutations and associated expansions in population sizes dramatically increase genetic and clonal heterogeneity and it is likely that most cancers evolve with a complex, branching architecture (figure 1.3B).

In deep sequencing studies of mixed tumour cell populations the variant allele frequencies can be used to size sub-clones. In the whole genome sequencing of 24 primary AML samples between one and four clusters of mutations were detected based on variant allele frequency, although the number of variants specific to individual sub-clones was small (average only 40) (Welch et al., 2012). Most AML-associated mutations are generally shared by all leukaemic clones/cells, as the initiating lesion arises in a cell with a mutational history (Welch et al., 2012). Exome sequencing of the progeny of single haematopoietic stem/progenitor cells (HSPCs) from healthy individuals revealed that the number of mutations increases near-linearly with age and is very similar to that found in AML. This suggests that AML develops stochastically in a cell which fortuitously accrues a transforming

combination of mutations (Welch et al., 2012). Therefore it does not seem surprising that somatic single nucleotide variants (SNVs) in sub-clones accounted for only 14% of the total SNV per genome (Welch et al., 2012). Interestingly, in other tumours the proportion of SNVs that are specific to sub-clones is much higher (Gerlinger et al., 2012; Nik-Zainal et al., 2012). Possible explanations include a longer latency between the initiating lesion and clinically overt disease and higher rates of somatic mutation acquisition.

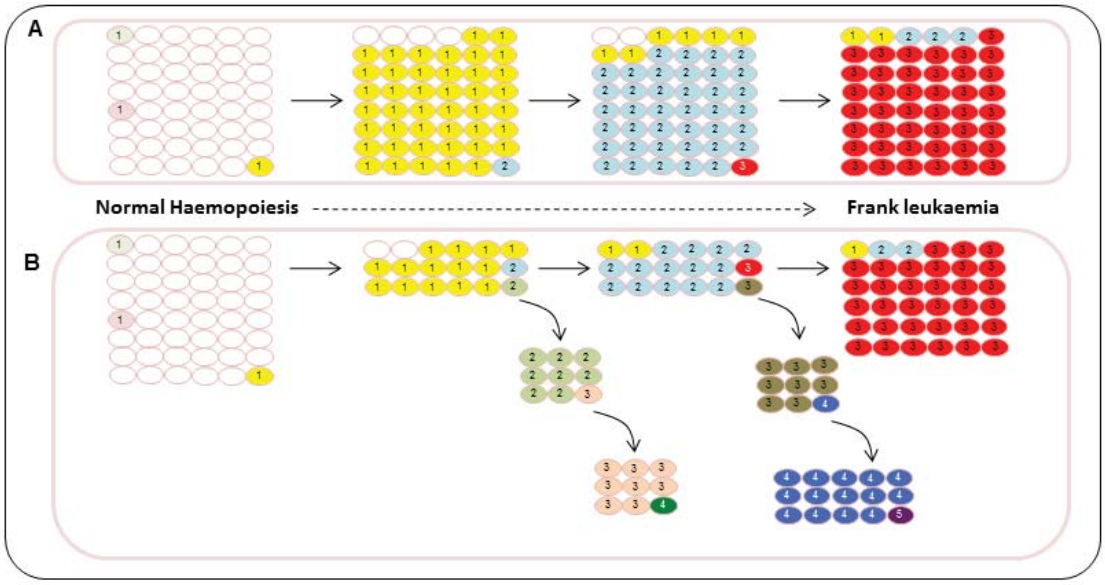


FIGURE 1.3: LINEAR AND BRANCHING CLONAL EVOLUTION
(A) Linear evolution: Sequential dominant clones (clonal sweep) result in a linear architecture with stepwise accumulation of driver mutations. The final tumour carries all mutations arising during evolutionary history and overwhelms earlier clones carrying only some of the mutations. (B) Branching evolution: The final leukaemia/cancer may be dominated by a single clone, but others which have followed divergent mutational pathways are also evident. Small sub-clones may fall below the limit of detection, in which case the complexity of the branching is underestimated. Branching evolution is favoured by faster acquisition and smaller effects of mutations. Numerals indicate the number of mutations in cells. Cells carrying identical mutations are represented in the same colour (Grove and Vassiliou, 2014).

It is likely that genetic heterogeneity is significantly under-reported in cancer genome sequencing studies because mutations in small sub-clones fall below the limit of detection. For example, the expected allelic frequency of a heterozygous mutation in a clone that represents 5% of the total tumour mass is only 2.5% and therefore at

40x coverage only one read is expected to have the mutant allele. Either ultra-deep or single cell genomic sequencing methods will be required to fully elucidate tumour architecture (Hou et al., 2012; Navin et al., 2011; Nik-Zainal et al., 2012).

The prevailing dogma is that the evolution of cancer occurs through a complex branching pattern of mutation acquisition (Greaves and Maley, 2012), although there is evidence for dominance of both linear and branching pathways in individual AMLs. A comparison of paired primary and relapsed AML samples by whole genome sequencing revealed two patterns of clonal evolution during relapse (Ding et al., 2012). In some cases only a single mutation cluster was found in the primary tumour. In these cases the single clone gained additional mutations at relapse, consistent with a linear pattern of evolution, although minor branching sub-clones may have been present below the limit of detection. In the remaining cases, multiple mutation clusters corresponding to different sub-clones were detected in the primary sample. A sub-clone survived therapy, gained additional mutations and expanded at relapse (branching evolution). In comparison to primary tumour mutations, there was an increase in transversions in the relapse-specific mutations and it is thought that these arose due to DNA damage caused by cytotoxic therapy (Ding et al., 2012).

Similarly, two studies comparing acquired copy number aberrations (CNA) and copy neutral LOH in paired diagnosis and relapse samples in *NPM1* mutant (Krönke et al., 2013) and unselected cases of AML (Parkin et al., 2013) found that re-emergence or evolution of a founder or ancestral clone is typical in relapsed AML. This is in contrast to findings in ALL where genetically distinct clones are occasionally observed (Mullighan et al., 2008). One patient from the *NPM1* mutant AML study was found to have different *NRAS* mutations at diagnosis and relapse, indicating either these represented independent clones and branching evolution, or that the mutation was lost and a new mutation acquired in the same gene in an earlier clone that was not eradicated by therapy. Similarly, the specific *FLT3-ITD* mutations differed between diagnosis and relapse in 3/24 patients (Krönke et al., 2013). In any event, both examples show that convergent evolution operates frequently in AML.

In another study, antecedent MDS bone marrow samples were genotyped for mutations identified on whole genome sequencing of a secondary AML from the same patient (Walter et al., 2012). Most MDS samples were oligoclonal, but each

clone carried all of the pre-existing driver and passenger mutations (Walter et al., 2012). The MDS founding clone was outcompeted by daughter clones in some cases, but it always persisted in the AML sample. Progression to AML was associated with the persistence of a founding clone containing 182 - 660 somatic mutations, and the outgrowth of at least one sub-clone with tens to hundreds of new mutations including at least one new tier 1 mutation (Walter et al., 2012). The proportion of secondary AML specific mutations was smaller in the subjects who progressed to secondary AML in less than 6 months(6.7%) than in those with slow progression (>20 months)(37.8%)(Walter et al., 2012).

In chronic myelomonocytic leukaemia (CMML), a condition which progresses to AML in 15-30%(Swerdlow, 2008) of patients, a predominantly linear pattern of acquisition of mutations is described, with limited branching through LOH (Itzykson et al., 2013b). In this study 18 candidate genes were analysed in single cell derived colonies from 28 patients. Only one patient showed somatic mosaicism with independent acquisition of *NRAS* and *KRAS* mutations in distinct sub-clones. Although the candidate gene approach may underestimate true clonal diversity, this work does suggest that in CMML the dominant tumour clone mostly results from sequential waves of mutation acquisition and expansion, with only minor branching sub-clones generated.

1.2.4 The Timeframe for AML Evolution

Available evidence suggests that cancer evolution is an inefficient process with a highly variable rate of progression (Stratton et al., 2009). AML is an uncommon cancer (figure 1.1), whose incidence rises with age, although it can occur at any age with 15% of cases in people under 40(2012; Bhayat et al., 2009; Dores et al., 2012; Shah et al., 2013). The rarity of the disease probably reflects the small mutational burden of AMLs and may reflect a paucity of external mutagens in the HSC niche or an unusual level of protection against them. One possible explanation for the latter is the ability of a small fraction of HSCs to sustain haematopoiesis at any time, allowing HSCs to remain quiescent for most of their lifespan and in so doing reducing their total number of divisions. This is only possible because of the very high proliferative capacity of later progenitors, whose limited lifespan and self-renewal minimises their own risk of transformation.

Pre-malignant clones arise with surprising frequency during foetal development. The in-utero acquisition of leukaemogenic mutations was first reported in concordant twins with ALL whose haematopoietic cells shared a unique somatic MLL rearrangement (Ford et al., 1993). Subsequently, clonotypic AML1-ETO fusion sequences were detected in Guthrie spots in cases of childhood AML (Wiemels et al., 2002). However, the prevalence of detectable *AML1-ETO* and *TEL-AML1* in cord bloods is 100-fold greater than the risk of the corresponding leukaemia and the frequency of positive cells (10^{-4} to 10^{-3}) indicates substantial clonal expansion of the abnormal progenitor population (Mori et al., 2002). This is because these fusion genes are not sufficient for disease development, as evident by protracted post-natal latencies, non-concordant phenotypes in monozygotic twins(Wiemels et al., 1999; Wiemels et al., 2002) and the lack of overt leukaemia in genetically-modified/transgenic mice(Rhoades et al., 2000). Therefore, secondary genetic events appear necessary for tumour development. It is unknown whether foetal acquisition of *AML1-ETO* can lead to adult-onset AML, but it is possible that long-lived HSCs progress only in later life, for example following chemotherapy in therapy related AML. In fact adults treated for *AML1-ETO* positive AML can exhibit persistence of the fusion in the blood for years in the absence of disease relapse (Kusec et al., 1994; Miyamoto et al., 1996).

The presence of detectable oncogenic mutations in blood in the absence of haematological disease is not unique to childhood. For example, somatic *TET2* inactivating mutations were identified in 10 of 182 females aged over 65 with skewed X-chromosome inactivation patterns (XCIP) and normal haematopoietic parameters (Busque et al., 2012). Mice with *Tet2* deletion exhibit increased HSC self-renewal potential, without detectable changes in standard haematological parameters, paralleling what happens in the aforementioned human cases (Moran-Crusio et al., 2011; Quivoron et al., 2011). After follow up of seven *TET2* mutant individuals for at least 5 years, one developed evidence of a haematological malignancy; a *JAK2V617F* mutant MPN (Busque et al., 2012).

The above findings show that somatic mutations, a universal feature of normal ageing, can drive the expansion of individual HSCs to the point of dominating haematopoiesis without causing disease. Nevertheless, the onward development of a haematological malignancy although not inevitable, becomes much more likely.

This observation is not unique to *TET2* mutations, but is also a feature of other somatic mutations such as large chromosomal deletions/amplifications which also increase in frequency with age (Jacobs et al., 2012; Laurie et al., 2012; Schick et al., 2013). In fact, there is a 5-10 fold increase in the risk of developing a haematological malignancy in the decade after the detection of mosaicism for such chromosomal changes in blood leukocyte DNA (Laurie et al., 2012; Schick et al., 2013).

Some studies which have analysed the clonal composition of blood from healthy women using X-inactivation markers suggest this is stable over time even in the elderly (Prchal et al., 1996; Swierczek et al., 2008). However, a study of the serial composition of copy number variants (CNV) in people without diagnosed haematopoietic disorders showed clear fluctuations in the proportion of nucleated blood cells with aberrations over time (Forsberg et al., 2012). In one person with a 20q deletion, the variant was barely detectable at 71 years of age, accounted for 50% of cells at 75 years, but only 36% at 88 years of age (Forsberg et al., 2012). In the longitudinal study of CSF3R mutations in congenital neutropenia, the independent acquisition of several different CSF3R mutations in different cells was demonstrated (Beekman et al., 2012; Campbell et al., 2010). Serial analysis of patient samples shows that one mutation/clone dominates at a time, but new mutations are able to replace previously dominant ones and mutations that fall below the limit of detection are sometimes detectable on subsequent samples (Campbell et al., 2010). It is unknown whether the clonal expansion of cells containing genetic abnormalities is always due to positive selection or reflects stochastic fluctuations in the frequency of HSC progeny or simply cycles of quiescence and active division of HSCs.

1.2.5 Initiating Mutations and Order of Acquisition

There are limited human studies which trace the presence of mutations in sequential samples from AML patients. For obvious reasons those that do compare relapsed versus primary tumours, or secondary AML versus a preceding haematological disorder, rather than profiling the pre-leukaemic evolution of primary or de novo AML (Ding et al., 2012; TCGA_Research_Network, 2013; Walter et al., 2012). The initiating lesion is only definitively known in familial AML however the dynamics of clonal evolution are likely to be different as all HSPCs carry the initiating mutation. Our understanding of initiating mutations in *de novo* AML is derived from studies of mutational allelic burden at presentation, stability of mutations through the disease

course, patterns of co-occurrence between mutations in leukaemia blasts and pre-leukaemic HSCs, and mechanistic studies of the properties of specific mutations. Generally it is thought that proliferative (type I) mutations are secondary events that co-operate with a variety of initiating lesions to produce disease. However, it is clear that at least some lesions can occur as either early or late events in the same tumour type, suggesting they are not acquired in any strict order(Anderson et al., 2011). In AML there are examples of 'early' mutations lost at relapse and 'late' mutations which are acquired first (Krönke et al., 2013).

The pattern of co-occurring mutations in residual HSCs or leukaemia cells has been used to determine the order of acquisition of mutations (Itzykson et al., 2013b; Jan et al., 2012). In one study residual HSCs were screened for patient specific mutations identified by tumour exome sequencing in six patients with de novo, *FLT3-ITD* mutant, normal karyotype AML (Jan et al., 2012). Many AML-associated mutations including *NPM1*, *TET2* and *SMC1A* were detectable in the residual HSC, but others, such as the *FLT3-ITD* and *IDH1*, were not, indicating these were probably late events. The population of residual HSCs showed varying allele frequencies for each of the detectable mutations and by comparing the patterns of mutations at the single cell level, researchers were able to reconstruct the phylogenetic tree in several cases (Jan et al., 2012).

Mutations in *NPM1* are often considered early events in AML pathogenesis largely because of their stability through the disease course and their mutually exclusivity with the most well established type of initiating mutations, chromosomal translocations (Falini et al., 2005; TCGA_Research_Network, 2013). However, recent studies have provided evidence that mutations in *NPM1* are not necessarily an initiating event and often follow *DNMT3A* mutation (Krönke et al., 2013; Shlush et al., 2014). Although *DNMT3A* and *NPM1* mutations frequently co-occur in AML blasts, stem cells purified from the blood of AML patients with both mutations showed recurrent *DNMT3A* mutations at high allele frequency without co-incident *NPM1* mutation(Shlush et al., 2014). These single mutant stem cells had a multi-lineage repopulation advantage over un-mutated HSC and persisted in post chemotherapy remission samples. Similarly, in a study comparing copy number aberrations and recurrent mutations in paired diagnosis and relapse samples of 53 *NPM1* mutant AMLs, mutations in *DNMT3A* were the most stable lesion. Persistence of *DNMT3A*

was found in five patients who lost the *NPM1* mutation at relapse suggesting that the *DNMT3A* mutations preceded those affecting *NPM1* (Krönke et al., 2013). In mice, knockout of *DNMT3A* in HSC induced increased self-renewal but did not lead to AML, suggesting co-operating mutations were required (Challen et al., 2012). Similarly, in the human study the long latency to relapse in cases which lost the *NPM1* mutation suggests the residual clone needed to acquire additional mutations before relapse occurred (Krönke et al., 2013). Notably, there was also a single case where *DNMT3A* was lost at relapse and the *NPM1* mutation was maintained, which implies that the mutation order is not strict.

So why are some mutations more often early and others more often late events in the pathogenesis of AML? It is very likely that in the great majority of AMLs the initiating mutation happens stochastically. However, this might alter the probability and type of secondary mutations en route to a malignancy. Potential mechanisms include a restriction in the cellular pathways through which secondary mutations could imbue additional fitness, but are not limited to this. For example, induced changes in the epigenetic program or the microenvironment may alter the phenotypic consequences of secondary mutations or the nature of selective pressures. Evidence of convergent evolution in multiple tumour types (Anderson et al., 2011; Gerlinger et al., 2012) suggests that either (i) those mutations are targeted by a specific mechanism of mutation, for example the off target effects of activation induced deaminase (AID), (ii) such mutations are recurrently selected due to their strong fitness advantage in a situation of high mutational diversity (parallel evolution) or (iii) the spectrum of co-operating lesions is severely limited in the context of pre-existing mutations. It is probable that there are no set rules governing the order of acquisition of mutations in AML, but that the specific effects/consequences of individual mutations make them more or less likely to facilitate subsequent evolution to leukaemia/cancer.

1.3 AML with mutated *NPM1*

Somatic mutation in *NPM1*, which encodes Nucleophosmin, is found in around 30% of AMLs, making it one of the most frequent mutations in this disease (Falini et al., 2005; TCGA_Research_Network, 2013). The prevalence of the *NPM1* mutation increases with age and it is found in approximately 50% of normal karyotype AMLs in

adults (Falini et al., 2005; Suzuki et al., 2005; Swerdlow, 2008; Verhaak et al., 2005). Mutations in *NPM1* define a distinct subgroup of AML with typical clinical, pathological and molecular characteristics and consequently *'AML with mutated NPM1'* has recently been included as a provisional entity in the WHO classification of tumours of the haematopoietic and lymphoid tissues (Swerdlow, 2008).

Although several different types of mutations in *NPM1* have been described in AML, these are localised to exon 12 and consistently result in nucleotide gain at the C-terminus (Falini et al., 2006; Falini et al., 2005). This disrupts the normal nucleolar localisation signal and generates a novel nuclear export signal, resulting in cytoplasmic dislocation of nucleophosmin (2005; Falini et al., 2006; Falini et al., 2005). The most common such mutation (Type A), is a TCTG duplication and accounts for approximately 80% of *NPM1* mutations in human AML (Verhaak et al., 2005). Although cytoplasmic dislocation of nucleophosmin is the universal consequence of *NPM1* mutations found in human AML, how this contributes to the pathogenesis of leukaemia is not yet understood. This is a subject of great interest due to the high prevalence of this mutation in human AML and because *NPM1* mutations are thought to be crucial events in leukaemic evolution.

Several groups have attempted to model the effect of mutant *NPM1* in the mouse. *Npm1* haploinsufficiency in heterozyogous knock-out mice resulted in an increase in the HSC number (Raval et al.). In a transgenic mouse model the expression of the type A *NPM1* mutation was driven by a myeloid specific human promoter *MRP8I* (Cheng et al., 2010). This resulted in the development of myeloproliferative changes in haematopoietic organs in 27% of mice. These changes were first evident from six months of age but none of the transgenic mice went on to develop leukaemia over the course of 24 months. The lack of AML development may result from differences in the expression level and pattern of the *NPM1* mutant protein in transgenic mice compared to the human disease, or could reflect a requirement for co-operating mutations. In a recent model the type A *Npm1* mutation was conditionally expressed from the *Rosa26* locus using a CAG promoter and *MxCre* (Sportoletti et al., 2013). These mice developed thrombocytopenia and an expansion of megakaryocyte precursors in haematopoietic organs, but did not develop AML after 1.5 years of follow-up.

Our group published the only *Npm1* mutant mouse model which has successfully recapitulated the major features of the human disease(Vassiliou et al., 2011). In this model, a conditional knock-in of the type A *NPM1* mutation (*Npm1*^{cA}) caused *Hox* gene overexpression, enhanced self-renewal and expanded myelopoiesis (Vassiliou et al., 2011). Approximately one third of these mice developed AML, but only after a long latency (median survival 617 days), which suggests that co-operating mutations were required. To identify these mutations we employed transposon insertional mutagenesis (IM) using the *Sleeping Beauty (SB)* transposon. In the absence of the *Npm1*^{cA} mutation *SB* caused predominantly lymphoid leukaemias, however the combination of *SB* and *Npm1*^{cA} resulted in rapid onset AML in 80% of mice. Several known and novel putative driver mutations were identified using this approach (Vassiliou et al., 2011).

1.4 Transposons as tools for gene discovery in the study of cancer

Transposons are mobile genetic elements that were first described by Barbara McLintock in the 1950's (McClintock, 1950); a discovery for which she was awarded the Nobel Prize in Medicine and Physiology in 1983. The genomes of most eukaryotic and prokaryotic species are known to contain significant numbers of transposable elements (Bire and Rouleux-Bonnin, 2012) and in humans it is estimated that almost half the genome is derived from them, although these are predominantly transpositionally inactive (2001).

Transposable elements are categorised into two classes based on their mechanism of transposition. Class I elements or retro-transposons mobilise through a 'copy and paste' mechanism and use an RNA intermediate which is reverse transcribed prior to re-insertion (Ivics et al., 2009). In contrast, class II elements or DNA transposons move by a cut and paste mechanism and are characterised by inverted terminal repeat sequences. These class II elements have been recently developed into powerful gene discovery tools and have been applied to the study of different cancers by many groups including ours..

In nature, DNA transposons consist of a single gene encoding the transposase protein, surrounded by inverted terminal repeat sequences which contain the recognition sequence for the transposase (Izsvák et al., 2002; Jacobson et al., 1986). The excision and re-integration of the transposon by a cut-and-paste mechanism is

catalysed by the transposase protein. Inverted repeats are found at each end of the mobile sequence and constitute the transposase binding sequences, which are necessary and sufficient for DNA mobilisation. Therefore, it is possible to replace the transposase gene with alternate DNA cargo as long as this is located between the repeat sequences. In these non-autonomous systems, the DNA cargo is mobilised widely throughout the genome by the transposase, which is supplied in *trans* (Ivics et al., 2009). This is the key stratagem through which transposons are applied as a tool for insertional mutagenesis, genome manipulation and transgenesis.

Transposon systems have been used for these applications in invertebrate animal models for several decades. It wasn't until the development of the synthetic *Sleeping Beauty* (*SB*) transposon that transposition efficiency in vertebrate cells was sufficient for insertional mutagenesis and transgenesis in mammalian systems(Ivics et al., 1997). *SB* was initially resurrected from multiple inactive Tc1/mariner element fossil sequences found in fish genomes. The other widely used transposon system for insertional mutagenesis in murine models is *PiggyBac* (*PB*), which was derived from the cabbage looper moth *Trichoplusia ni* (Ding et al., 2005). Subsequent genetic engineering/modification of both the transposon and transposase has resulted in significant improvement in the transposition efficiency of both the *SB* and *PB* systems. These modifications included changes to specific amino acid residues and species optimisation for codon usage (Baus et al., 2005; Cadinanos and Bradley, 2007; Geurts et al., 2003; Mates et al., 2009; Yant et al., 2007; Yant et al., 2004; Yusa et al., 2011; Zayed et al., 2004).As a result, *SB* and *PB* transposon systems can integrate efficiently into chromosomes of somatic, germ and embryonic stem cells.

For insertional mutagenesis screens in mice the *SB* or *PB* transposon is typically introduced by zygote pronuclear injection and inserts as a concatamer at a random site in the mouse genome (Mann et al., 2014). Transgenic lines containing concatemeric transposon casettes are then crossed with lines which express the transposase. Tissue targeted insertional mutagenesis can be achieved by using either a tissue specific promoter to control transposase expression or by employing an inducible allele. In *Cre* inducible systems the transposase is usually inserted into an endogenous ubiquitously expressing locus, with conditionality imparted by either an upstream stop cassette flanked by *loxP* sites (Dupuy et al., 2009; Starr et al., 2009) or by the use of an invertible transposase cDNA flanked by mutant *loxP* sites (March

et al., 2011; Vassiliou et al., 2011). A tissue specific *Cre* recombinase is used to either remove the *lox*-stop-*lox* cassette or invert the transposase resulting in permanent expression of the sense-oriented cDNA.

Both the *SB* and *PB* transposons integrate widely throughout the genome and have been used successfully for cancer gene discovery in murine models (Collier et al., 2009; Collier et al., 2005; Dupuy et al., 2005; Rad, 2010). The *SB* and *PB* systems reportedly differ in terms of insertion site sequence, mobilisation efficiency, the size of the insert that can be mobilised, the degree of local hopping and the footprint that remains after excision (Ding et al., 2005; Liang et al., 2009; Wang et al., 2008). Genomic integration is largely random but is dependent on a minimal sequence; in the case of *SB* a TA dinucleotide and for *PB* a TTAA tetranucleotide, although around 2% of *PB* insertions were found to occur at non-canonical sequences in one screen(Li et al., 2013). *PB* has a higher integration preference for actively transcribed genes compared to *SB* (Liang et al., 2009; Wang et al., 2008). *SB* preferentially inserts in TA rich regions and with the consensus ANNTANNT although only the TA is an absolute requirement (Carlson et al., 2003). As a result of local hopping approximately 30-50% of *SB* integrations have been reported to map to the donor chromosome (Collier et al., 2009; Starr et al., 2009). Although local hopping still occurs (Li et al., 2013; Wang et al., 2008) it is less of an issue for *PB* and the effected region is much smaller (Friedel et al., 2013; Rad, 2010). The local hopping interval for SB has been estimated at 3-15Mb (Carlson et al., 2003; Horie et al., 2003) compared to ~100kB for PB (Wang et al., 2008).

An important difference between *SB* and *PB*, which is of relevance to forward mutagenesis screens because of the potential for occult mutagenesis, is the DNA footprint. Transposon excision results in a double strand DNA break, which is mended by the cell's endogenous repair machinery. The canonical *SB* footprint is a five base pair insertion creating a TACAGTA or TACTGTA sequence at the TA integration site, although deletions, insertions and non-canonical footprints also occur, albeit less frequently (Liu et al., 2004; Luo et al., 1998). Although the predominant footprint differs between cell types, in zebrafish embryos, mouse embryonic stem cells and cells of the adult mouse liver 90% of the footprints add five base pairs leading to a frame shift (Liu et al., 2004). In contrast, *PB* almost always

excises itself completely leaving no footprint, with excision induced genomic alterations detected in as few as 0.8% of excisions(Yusa et al., 2011).

There are several advantages of transposon IM over chemical and retroviral mutagens for performing forward genetic screens in mice. In contrast to chemical mutagens, the mutated locus is easily identifiable as the transposon itself serves as a tag. Furthermore, transposons integrate widely allowing extensive and largely unbiased coverage of the genome and can be targeted to a variety of tissues in a spatially and temporally controlled manner. The DNA cargo can also be manipulated to achieve gain or loss of function mutations, or both (Collier et al., 2005; Dupuy et al., 2005; Rad et al., 2010; Vassiliou et al., 2011). For such bi-functional transposons the precise mutagenic effect will depend both on the orientation in which they insert and their spatial relationship to surrounding gene(s) (figure 1.4). The frequency of integration in a given orientation around a particular gene can be used to surmise if it is acting to activate or inactivate target genes.

The conventional approach for analysing transposon insertional mutagenesis screens is to use ligation-mediated splinkerette polymerase chain reaction (PCR) followed by deep sequencing (Uren et al., 2009). One or multiple frequent cutter restriction enzymes are used to digest genomic DNA and restriction products that contain both genomic and transposon sequence are then PCR amplified after ligation of a linker sequence. After massively parallel sequencing across multiple tumours, the sites in which the transposon integrated more frequently than would be expected by chance are identified using one of several statistical methods. Genes within these common integration sites (CIS) are taken as the putative tumour drivers.

Whole body transposon IM screens frequently result in haematopoietic tumours, in particular T cell malignancies (Collier et al., 2009; Dupuy et al., 2005; Dupuy et al., 2009). In addition to using tissue specific expression of the transposase, the spectrum of induced tumours can also be influenced by the structure of the *SB* or *PB* transposon; more specifically the choice of promoter (Dupuy et al., 2009; Rad et al., 2010). The aforementioned model of *Npm1* mutant AML is the first published model in which IM has been targeted to induce AML (Vassiliou et al., 2011).

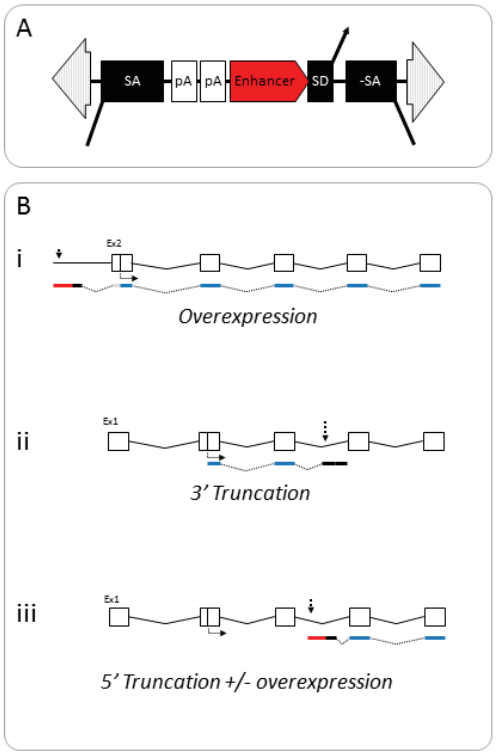


Figure 1.4 Examples of the mutagenic effects of transposons. A: A transposon with a unidirectional activating element followed by a splice donor, but dual splice acceptors and polyadenylation signals such that gene expression may be blocked regardless of the direction of transposon insertion. **B:** The effect of the transposon shown in A will depend on the direction in which it inserts and its spatial relationship to surrounding genes. Some examples are shown. i) The transposon inserts in the forward orientation 5' to the gene, causing overexpression of the gene. ii) The transposon inserts in the reverse orientation in the middle of the gene resulting in a 3' truncated gene product. iii) Insertion in the forward orientation mid-gene may result in overexpression of a 5' truncated product. Figure modified from the original provided by G Vassiliou.

Transposons used in cancer gene discovery are designed to mutate genes primarily by overexpression or truncation and therefore they cannot recapitulate the specific naturally-occurring mutations seen in human disease. This has been cited as a limitation of transposon insertional mutagenesis screens, but it is also a potential advantage. Although the precise mutations seen in the human disease were not seen, in previous transposon screens genes such as *Ras* and *Flt3* were still identified as important targets for up-regulation or activation by 5' truncation (Rad, 2010; Vassiliou et al., 2011). In addition, the transposon IM approach allows for pathway analysis, facilitating the identification of associated targets up- or down-stream of known cancer causing mutations. For example, in a *SB* driven pancreatic cancer model 20 CIS genes were identified that significantly predicted poor survival in humans, although only two of these were found to be mutated in human pancreatic cancer (Mann et al., 2012). IM screens are also useful for elucidating the functional consequences of mutations. Cancer typically involves multiple mutational events that include simultaneous activation of oncogenes and inactivation of tumour suppressor genes. Transposons are often designed so that their orientation can be used to

understand if a gene is activated or inactivated by an insertion. For example, by incorporating a unidirectional activating element (i.e. enhancer/promoter) within a transposon one can identify recurrent activating insertions by the fact that insertions always (or nearly always) face in the forward orientation with respect to their target gene (Rad, 2010; Vassiliou et al., 2011).

1.5 Using transposon insertional mutagenesis to study the molecular pathogenesis of AML

It seems likely that similar to human tumours, transposon driven cancers also evolve in stepwise manner akin to Darwinian evolution. Mobilisation of transposons is a continual process in the presence of on-going transposase expression. During tumour development cells in which a transposon integrates in a position where it gives a growth or survival advantage will be clonally selected and form premalignant clones. Amongst such clones subsequent transposon mobilisation will lead to serial clonal selection until full blown cancer becomes manifest.

The sub-clonal architecture of transposon driven tumours has not previously been assessed in detail because of the lack of a quantitative method for deriving the equivalent of mutant allelic burden. Conventional restriction-based splinkerette PCR methodology for capturing transposon insertions is not quantitative. Recent protocols have introduced shearing-based methods to fragment genomic DNA in an unbiased manner and this has significantly reduced the level of amplification bias and has, for the first time, allowed semi-quantitative analysis of transposon and retroviral integrations (Klijn et al., 2013; Koudijs et al., 2011). This, combined with significant reductions in sequencing costs will enable transposon insertions to be used as a marker of clonal size and by extension help decipher the clonal architecture of transposon-driven cancers (Friedel et al., 2013).

Another potential application of transposon-driven IM is in the study of tumour evolution. For the study of AML in particular, this could be performed in real time, as insertions can be readily and serially identified in blood samples taken prior to the onset of overt leukaemia. A model such as the *Npm1c* mutant mouse provides an ideal platform in which to study the clonal evolution of transposon-driven tumours (Vassiliou et al., 2011). As alluded to earlier, an improved understanding of the clonal evolution of AML could offer important clinical insights. For example, it has

implications for selecting appropriate markers for minimal residual disease (MRD) monitoring and for predicting the progression of pre-leukaemic clonal expansions and haematopoietic disorders.

1.6 Transposon insertional mutagenesis for cancer gene discovery in mature B cell malignancies

1.6.1 Normal B cell development

Lymphocytes, like other haematopoietic cells, are derived through lineage specific differentiation of HSCs and downstream proliferation. Normal B cell development in the bone marrow involves a process of V(D)J recombination, in which B-cell progenitors assemble the variable regions of antibody heavy and light chains from the numerous different V, D and J segments present in the germline loci. DNA sequences located between the recombined elements are deleted in the process. The endonuclease which mediates this process is encoded by the recombinase activating genes (RAG). B cells which express autoreactive receptors either undergo secondary V(D)J rearrangements or apoptosis, while those with in-frame V(D)J rearrangements and non-auto-reactive receptors leave the bone marrow to become mature, naïve B cells (Küppers et al., 1999).

When a naïve B cell recognizes an antigen it moves to the germinal centre (GC) of a secondary lymphoid organ. Within the GC, the B cell DNA is subjected to various types of DNA modification which may alter the antigen receptor specificity or the antibody type and effector function. Somatic hypermutation (SHM) introduces mutations within the variable region sequences with high frequency. This may result in increased antibody affinity for the antigen, positive selection and release of the cell into the periphery as an antibody producing plasma cell or a long-lived memory B cell. Alternatively, SHM may reduce the function of the antibody, which typically results in apoptosis of the mutated germinal centre B cell. Also, some B cells within the germinal centre will undergo class switching recombination (CSR), which is mediated by a recombination event between repeat sequences located 5' of the constant region of Ig heavy chain genes. Such recombination events leave the specificity of the antibody unaltered but switch the B cell to express other classes of immunoglobulin heavy chains and thereby change immune effector functions. Both CSR and SHM are mediated by activation induced deaminase (AID), but whereas

SHM is thought to be largely restricted to the GC, CSR can also occur elsewhere (MacLennan et al., 2003). V(D)J recombination may also take place within the germinal centre, allowing receptor editing (Han et al., 1997).

1.6.2 Correlation of lymphoma phenotypes with normal B cell development

Within the GC, normal B cells are subjected to molecular processes designed to initiate double-strand breaks and also to otherwise modify their DNA. As well as modifying the antigen receptor loci, both RAG recombinase and AID can introduce illegitimate off-target damage. Such damage, coupled with the significant proliferative expansion of B cells within the germinal centre, make this a high-risk stage of B cell development for acquisition of cancer driver mutations. B cell NHLs typically harbour translocations that juxtapose an oncogene to an Ig receptor locus. Many of these are thought to arise due to aberrant class switch recombination (CSR) or somatic hypermutation (SHM) mediated by AID, while others, such as *BCL-2* translocations in follicular lymphoma, probably arise due to errors in V(D)J recombination as indicated by the position of the breakpoint within the Ig gene. The relative rarity of T cell lymphomas may relate to the fact that normal T cells do not undergo SHM or CSR (Küppers et al., 1999). Nevertheless, it is also likely that at least some mutations within mature lymphoid neoplasms are acquired by earlier uncommitted haematopoietic progenitors (Weigert and Weinstock, 2012).

B cell lymphomas that are considered to be of GC or post-GC origin carry switched and hypermutated Ig heavy chain alleles (IgH). Somatically mutated variable region sequences are typical of many types of non-Hodgkin's B cell lymphomas (B-NHL) including follicular lymphomas, Burkitt's lymphomas, diffuse large B cell lymphomas, prolymphocytic leukaemia and lymphoplasmacytoid lymphoma, as well as chronic lymphocytic leukaemia and multiple myeloma. The differentiation between GC and post-GC origin is largely based on growth pattern, surface marker expression and the presence or absence of ongoing somatic hypermutation within the tumour clone (Küppers et al., 1999). Among the mature B cell neoplasms, unmutated variable region genes are only reported in mantle zone lymphomas and some cases of CLL (Küppers et al., 1999). In lymphomas such as Waldenstrom macroglobulinaemia or splenic marginal zone lymphoma, the malignant B cells may have undergone SHM but not CSR and IgH translocations are not typical.

1.6.3 Modelling Mature B cell Neoplasms in the Mouse

Lymphomas are also among the most common tumours in many strains of laboratory mice, with an incidence of 10-60% in aging C57BL/6 mice (Brayton et al., 2012; Szymanska et al., 2013; Ward, 2006). However, accurately recapitulating the features of human B cell neoplasms in mouse models has proven difficult. As well as the challenge of introducing recurrent somatic mutations to B cells at the appropriate stage of development, it is also evident that the constitutional genome of the mouse is important, as demonstrated by differences in disease incidence and phenotype between strains. Furthermore, the housing of experimental lines in specific pathogen free conditions may affect the spectrum and incidence of tumours, as immune activation has a role in the pathogenesis of many lymphoid tumours. There are also fundamental differences in the structure of the primary and secondary lymphoid organs between mouse and man, which must be considered. For example, in mice extramedullary haematopoiesis is normal in the spleen throughout life and continues in the thymus into adulthood. However despite these differences, the Bethesda proposals for the classification of lymphoid neoplasms in mice do highlight significant parallels between mouse and human B-lineage tumours (Morse et al., 2002).

The difficulties of modelling mature B cell neoplasms in the mouse are exemplified by plasma cell neoplasms. Multiple myeloma (MM) is a malignancy of terminally differentiated, immunoglobulin producing B cells (plasma cells). It comprises approximately 1% of all human cancers, is incurable with conventional therapy and causes nearly 2% of cancer deaths(Jemal et al.). Clinical features include osteoporosis, lytic bone lesions, renal impairment, immune paresis, hypercalcaemia and anaemia. MM is preceded by monoclonal gammopathy of uncertain significance (MGUS), an asymptomatic state characterised by the presence in serum of a monoclonal protein, which occurs in 3% of people over the age of 50 (Kyle et al., 2006). Transformation of MGUS to MM occurs at a rate of approximately 1% per year(Kyle et al., 2002), but the molecular mechanisms that drive progression are largely unknown.

Modelling MM in the mouse has proven particularly challenging, because the precise differentiation stage of the 'myeloma stem cell' remains unknown and targeting cancer genes to the mature B cell compartment is difficult. Previous mouse models

of MM have relied on the rare spontaneous development of plasma cell neoplasms in predisposed backgrounds and transplantation of transformed plasma cells(Janz, 2008). Xenograft models are useful for pre-clinical testing of novel therapies, but cannot model pre-malignant neoplastic stages and do not recapitulate normal tumour-stroma interactions. Forward genetic screens have identified various cancer genes involved in the pathogenesis of leukaemia and lymphoblastic lymphoma (Dupuy et al., 2005; Erkeland et al., 2004; Kool et al.; Li et al., 1999), but viral insertional mutagenesis of the plasma cell compartment has not been possible, probably because viruses with B-lineage tropism target less mature B-cells leading to lymphomagenesis. Transgenic mouse models in which oncogenes are targeted to the B cell compartment have frequently resulted in lymphomas with an immature or transitional cell phenotype (Adams et al., 1985; Butzler et al., 1997; Kovalchuk et al., 2000; Palomo et al., 1999). Those which cause a neoplastic plasma cell phenotype produce predominantly extraosseous tumours and do not recapitulate the typical bone marrow tumour growth of human MM (Janz, 2008).

In 2008 the Bergsagl group described a mouse model of MM, which was the first to accurately recapitulate many of the clinical features of human disease and show therapeutic fidelity (Chesi et al., 2008). This transgenic model placed the human *c-MYC* gene under the transcriptional control of the Vk promoter (*Vk*MYC*) and maintained the kappa light chain gene regulatory elements, which are required for targeting by somatic hypermutation(SHM) (Betz et al., 1994; Papavasiliou and Schatz, 2000). In the *Vk*MYC* model the pool of transgene expressing cells was restricted to B cells in a late stage of development, from which MM is believed to arise(Brennan and Matsui, 2009; Huff and Matsui, 2008). The construct contained a V-kappa exon, which spliced in frame to human *MYC* exons, however the third codon of V-kappa was mutated to a stop codon, thus preventing translation of the downstream *MYC* codons. This stop codon was engineered to overlap with a preferential target sequence for endogenous Activation Induced Deaminase (AID), the enzyme responsible for class switch recombination and SHM during B cell development (Delker et al., 2009; Maul and Gearhart; Rogozin and Diaz, 2004). The transgene was expressed in only a minority of mature B cells, yet with age all mice developed progressive monoclonal plasma cell expansion. The MM tumours did not show intraclonal heterogeneity, suggesting they were not subject to ongoing SHM. Notably

the incidence of Burkitt lymphoma was low and there were no aggressive pro-B lymphomas.

1.6.4 Targeting insertional mutagenesis to the mature B cell compartment

Mice subjected to whole body transposon insertional mutagenesis frequently develop lymphoma, but these are most commonly aggressive T cell lymphomas (Dupuy et al., 2009). In our AML insertional mutagenesis study in which an inducible *SB* transposon was targeted to the haematopoietic compartment using *Mx1-Cre*, B cell tumours were more common than T cell lymphomas (Vassiliou et al., 2011). However, even in $Npm1^{WT}$ mice only around a third developed B cell neoplasms and these were typically of high grade.

Insertional mutagenesis has previously been targeted to germinal centre B cells using a conditional transposase and an *Aid-Cre* knock in allele in which the Cre recombinase cDNA is fused to the activation-induced cytidine deaminase gene(Dupuy et al., 2009). Of the eighteen insertional mutagenesis mice for which results are published, eight (44%) developed B cell neoplasms, which included diffuse large cell, follicular and pre-B lymphomas, but no plasma cell neoplasms. Interestingly myeloid, T cell and solid tumours were also detected.

The approach used in the *Vk*MYC* mouse model provides an alternative method for targeting insertional mutagenesis to the mature B cell compartment. By modifiying the *Vk*MYC* construct to express a transposase in place of, or in addition to the *MYC* transgene, one would expect to generate a forward mutagenesis screen which is highly specific for mature B cell malignancies. This specificity is predicted because the activation of the transposase is thought to be dependent on AID induced reversion of the stop codon.

1.7 Aims

The aims of the first part of this thesis are to investigate the clonal evolution and sub-clonal architecture of AML by studying the timing and pattern of acquisition of mutations in *NPM1* mutant AML. Both human tumour samples and a mouse insertional mutagenesis model were used to investigate the order of acquisition of mutations in serial samples. Firstly, a detailed studied of an informative case of human CMML evolving to AML is described and the implications about clonal

evolution and leukaemic transformation discussed. Subsequently, using the mouse model, pre-leukaemic blood samples were studied to identify i) when integrations in putative co-operating driver genes were first evident, ii) whether such integrations resulted in any detectable changes in the blood parameters, iii) the time lag between first detection of such driver integrations and the development of overt leukaemia and iv) whether the order of acquisition of co-operating mutations followed a set pattern in different mice. I also discuss the extent to which such driver integrations were shared between different leukaemia cells within the tumour population.

The second major aim of this work is to generate a *PB* IM mouse model of MM for cancer gene discovery. Two related models were developed for this study. In the first the *PB* transposase replaced the *MYC* transgene in the *Vk*MYC* model. The second expressed *MYC* and *PB* together from the same cistron, in order to study genes co-operating with *MYC* in disease pathogenesis.

Chapter 2: Materials and Methods

2.1 Sequencing of human leukaemia samples

2.1.1Exome Sequencing and genomic alignment

The human samples were collected with the written informed consent of the patient and after ethics approval (REC 07/MRE05/44: The causes of clonal blood cell disorders). The protein coding exome of non-amplified whole bone marrow (BM) DNA was sequenced in three samples from the same patient; i) at diagnosis with CMML (day1) ii) at diagnosis of AML (day 83) and iii) in first complete remission (day 112).

Library preparation, sequencing and variant calling were done through the Sanger pipeline. Genomic libraries, enriched for protein coding exons were generated by hybridisation to RNA baits using Agilent SureSelect Human Exon 50Mb Kit (Agilent, S02972011). The libraries were analysed on the Illumina HiSeq2000 sequencing platform. Paired 75bp reads were generated, which were aligned to the human genome(NCBI build 37) using the BWA algorithm(Li and Durbin, 2010). Reads which were unmapped or outside the target region were excluded from analysis as were PCR duplicates.

The day 112 clinical remission sample was used as the reference for the identification of somatic mutations. A modification of the Pindel algorithm was used to identify insertions and deletions as previously described (Bolli et al., 2014; Ye et al., 2009). The CaVEMan (Cancer Variants through Expectation Maximisation) algorithm was used to call single nucleotide substitutions (Papaemmanuil et al., 2011; Varela et al., 2011) and copy number analysis was performed by Peter Van Loo using ASCAT (allele-specific copy number analysis of tumors) (Van Loo et al., 2010). The variant clustering on the exome data was done by David Wedge using a previously developed Bayesian Dirichlet process (Nik-Zainal et al., 2012).

2.1.2 Re-Sequencing Using Non-allele Specific PCR and MiSeq

Purified DNA from nine blood and BM samples from the same patient were obtained from Addenbrooke's Hospital. Presumed driver mutations were selected for re-sequencing along with mutations that clustered with them on the Bayesian Dirichlet

analysis of the exome sequencing data. The PCR primers were designed using Primer3web (http://primer3.ut.ee) and a 33 or 32 nucleotide sequencing adaptor was added to the forward and reverse primer sequences respectively (table 2.1). The first and second round PCR reactions and pooling of products was performed by Nicla Manes. The first round reaction used 20ng of DNA (10ng/μL), 1μL of each primer (10 μM) and 48μL of Platinum® PCR SuperMix High Fidelity (Life Technologies). PCR conditions were: 95°C 5min; then 36 cycles of (95°C 30s, 50°C 30s, 72°C 30s); with a final extension of 7 minutes at 72°C. 15uL of the PCR product was run on a 2.5% ethidium bromide gel with loading buffer (4 μL). Samples that failed were re-run with an annealing temperature of 57°C. The *CEBPA* PCR failed a second time, but was subsequently successful using Platinum® *Taq* DNA Polymerase High Fidelity with an annealing temperature of 60°C and a total reaction volume of 40μL: 20ng DNA; 0.8μL each primer; 4μL 2.5mM dNTP mixture; 4 μL buffer; 0.4μL Platinum®*Taq* (5U/μL); 3μL $MgCl_2$ (50mM) and 2.5μL 5% DMSO.

Between 7 and 30μL of each first round PCR product was pooled for each of the nine time points depending on the relative strength of the gel band for each reaction. The pooled PCR products were then purified using the Qiagen PCR purification kit and quantified by NanoDrop. Addition of the barcoded indexing primers was performed in a second PCR enrichment step, using primers designed by Mike Quail (table 2.2). This reaction was performed in a total volume of 50μL as follows: 25μL 2xKAPA HiFi HotStart ReadyMix; 200pg of pooled PCR product; 2μL of each primer (5μM). PCR conditions were: 98°C 30s; then 12 cycles of (98°C 10s, 66°C 15s, 72°C 20s); followed by final extension at 72°C for 5 minutes. Size selection was then performed using SPRI beads. A total of 31.5μL of SPRI beads was added and after 5 minutes at room temperature to allow binding, the PCR plate was placed on a magnetic plate for 3 minutes for bead capture. The liquid was then removed and two washes with 80% ethanol were performed before the samples were left to air dry for 5 minutes. The plate was then removed from the magnet and 35μL of EB buffer (Qiagen) was added. After mixing and incubation to allow release of the DNA, the plate was placed back on the magnet and 30μL of DNA solution was collected from each well. The samples were then sequenced on the MiSeq platform.

MiSeq Primer Name	Genomic Distance	Mutation Target	Full Primer
MiSeq_API5_F	292	149	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAAGTTGCCTTTTCGTCAGT
MiSeq_API5_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCGAGGGATTGAAGGTCTGT
MiSeq_UBN2_F	137	29	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGGACCAGAAAACTCCAACA
MiSeq_UBN2_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTATCTAGTGAGTCGTCGAGGC
MiSeq_Fam171a110_F	272	78	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGCTGACATAGGAGTGGTC
MiSeq_ Fam171a110_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCACGGGAAGCAAACTCACC
MiSeq_Ap4s1_F	162	43	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCACCAATGAACAGCACAGT
MiSeq_Ap4s1_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGCATTCCAGTCTAGCCCAA
MiSeq_Abca4_F	256	166	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATGGGGGCCCTCAAATCAGA
MiSeq_Abca4_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGGTGTCTCATTGCCTCAGA
MiSeq_ITPKB_F	225	124	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGATGTGCGCCTCAAACATG
MiSeq_ITPKB_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGCAGGCTGAATAGTAGCA
MiSeq_Acrc_F	194	121	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCTGCCAGAGAAACTACG
MiSeq_Acrc_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGGTGTCAGAAGGAAAGAGC
MiSeq_Speg_F	180	58	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCCCTAAGTCTGCAGAACC
MiSeq_Speg_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTCTGAGCATAGGGTGTGAGG
MiSeq_Ptchd2_F	262	112	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCCATCTCCCTGTCCATC
MiSeq_Ptchd2_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGGGATCAGCTTTGGGAAAC
MiSeq_Clcn1_F2	203	132	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCCACCCTTTCTGCTTCTT
MiSeq_Clcn1_R2			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGTTGCTAGTGTCAGGAGCA
MiSeq_Thoc2_F	198	84	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAACATGGACGCTGCCTTC
MiSeq_Thoc2_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGGGTTTTGCTAGGGGAACT
MiSeq_Acsl6_F	220	169	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCCCTGGTATCATGCTT
MiSeq_Acsl6_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCTTGGCTTGTAGGACAGTG
MiSeq_Zxdb_F	244	115	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCTTCCTGGTGCTGCTTG
MiSeq_Zxdb_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGTGGATAGTGACTGTGCCC
MiSeq_Ptpn11_F3	230	55	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGGTTGTCCTACACGATGGT
MiSeq_Ptpn11_r3			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGGGCTTTGAATTGTTGCAC
MiSeq_Smc3_F	259	123	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCATAGAAAATGTTGGCAGT
MiSeq_Smc3_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGCTTCTGCCTATTTGGACA
MiSeq_SMC3_Fs	113	61	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAACTTAATGAGCTGAGAGAGA
MiSeq_SMC3_Rs			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTCACCTTCTGATCGTGCCAT
MiSeq_Tet2_F	254	76	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCATGGGAGCCACCTCTA
MiSeq_Tet2_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGTAAGCCTCCTTGGACACA
MiSeq_TET2_Fs	118	51	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCACCCAATCTGAGCAATCC
MiSeq_TET2_Rs			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTAGGGCATGAAGAGAGCTGTT
MiSeq_FLT3_ITD_F	325	295	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAATTTAGGTATGAAAGCCAGC
MiSeq_FLT3_ITD_+325_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTTTCAGCATTTTGACGGCAACC
MiSeq_NPM1_F	248	69	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTCTATGAAGTGTTGTGGTTCC
MiSeq_NPM1_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGGACAACACATTCTTGGCA
MiSeq_CEBPA_F	191	85	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGTAGGCGCTGATGTCGAT
MiSeq_CEBPA_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGACTTCTACGAGGCGGA
MiSeq_DNMT3A_F	199	125	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGTCGCTACCTCAGTTTGC
MiSeq_DNMT3A_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTGCCCTCTCTGCCTTTTCT
MiSeq_nRAS_F	184	126	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGACAAGTGAGAGACAGGA
MiSeq_nRAS_R			TCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCAACAGGTTCTTGCTGGTG

Table 2.1: Primer sequences used for re-sequencing target genes in the human serial samples. The gene specific sequence is shown in red. The length of the PCR products is shown, along with the position of the mutation target. Two sets of *TET2* and *SMC3* primers were used.

Alignment of MiSeq reads to the reference genome was performed by Ignacio Varela (Universidad de Cantabria). A modified Bayesian Dirichlet process to allow for multiple sample analysis was performed by David Wedge. In brief, subclonal clusters of mutations were identified using a previously described Dirichlet process, implemented using a Markov Chain Monte Carlo (MCMC) method (Bolli et al., 2014; Nik-Zainal et al., 2012). The method is summarised by David Wedge as follows: 'From the MCMC assignment of mutations to clusters, the most likely configuration of clusters and node assignments was obtained using a stepwise, greedy expectation-maximization (EM) algorithm which alternately added a node and shuffled mutations

between nodes until no further improvement in the agreement with the posterior distribution from the MCMC sampling could be made. The best set of clusters was then chosen using the Bayesian information criterion(Schwarz, 1978). The Dirichlet process was run 5 times, each time for 10000 MCMC iterations. Mutations were assigned to the same clique if every mutation in the clique appeared in the same cluster as every other mutation within the clique in most (i.e. 3 or more) of the runs'.

Primer	Sequence
PE1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
iPCRtagT1	CAAGCAGAAGACGGCATACGAGATAACGTGATGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T
iPCRtagT2	CAAGCAGAAGACGGCATACGAGATAAACATCGGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T

Table 2.2: Indexing primers used for the second round PCR for MiSeq. The barcode sequences are underlined. Only the first two of the ten barcoded primers are shown.

2.2 Mice

2.2.1 Mouse Strains used in the *Sleeping Beauty* Study

The *Npm1*$^{flox\text{-}CA/+}$, *Rosa26* conditional *Sleeping Beauty* transposase (*Rosa26*$^{flox\text{-}SB}$) and *Mx1-Cre* have been previously described (Kuhn et al., 1995; Li et al., 2010; Rad et al., 2010; Vassiliou et al., 2011). The low copy transposon line was generated by George Vassiliou and differs to the published *GrOnc* high copy (GRH) model only in donor site (Chr16 vs Chr19) and transposon copy number (15 vs 80 copies) (Vassiliou et al., 2011). The mutagenesis cohort (*Npm1*$^{flox\text{-}cA/+}$, *Rosa*$^{floxSB/+}$; *GrOnc*$^{+}$; *Mx1-Cre*$^{+}$) given four to six intraperitoneal injections of polyinosinic-polycytidylic acid (pIpC)(500µg) to activate the mutant *Npm1*^{cA} and *SB* at 8-12 weeks of age. The NOD Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice, called NOD-SCID Gamma (NSG) mice were purchased from Jackson Laboratories (Bar Harbor, ME). All mice were maintained in accordance with Home Office requirements under project licenses 80/2477 and 80/2564.

2.2.2 Transplant of NSG mice

Frozen spleen cells were thawed at 37°C and re-suspended in 5mL RPMI media. An aliquot, mixed 1:1 with 0.4% trypan blue was counted using a haemocytometer. The cells were spun at 250g for 5 minutes and resuspended in 1mL of RPMI. This cell suspension was used to generate aliquots of the required number of cells in

250µL of RPMI media. Transplants were performed by tail vein injection using a 26G needle and a 1mL syringe.

2.2.3 Mice in the *PiggyBac* Study: Cloning *Vk*hPB* and *Vk*MYC-TA-hPB*

The *Vk*Myc* construct was generously provided by Leif Bergsagel (Mayo Clinic) (Chesi et al., 2008) (figure 2.1). The hyperactive *PiggyBac* (*hPB*) cDNA was generated by Kosuke Yusa by modification of *mPB* cDNA (Cadinanos and Bradley, 2007; Yusa et al., 2011). Linker sequences containing restriction enzyme sites and the start of the *hPB* sequence were synthesised to order (GENEART) (Appendix 2A). The ***Vk*hPB* only** construct was designed to replace the *hMYC* coding region (exons 2&3) with *hPB* cDNA. The ***Vk*MYC-TA-hPB*** construct was designed to introduce the *T2A-hPB* cDNA in frame, after the penultimate codon of *hMYC*, thus removing the stop codon (figure 2.1). Insect virus *Thosea asigna* 2A peptide (T2A) and similar 2A like peptides from other viruses enable translation of multiple proteins from a single mRNA (Szymczak et al., 2004). As the T2A linker peptide is hydrolysed soon after translation this construct will generate two separate proteins, *hMYC* and h*PB*.

The linker sequences were digested with *KpnI* and *PmlI* and cloned into the *PB* plasmid. PCR with a high fidelity polymerase (Phusion, Finnzymes) was performed using suitable primers to add an *FseI* restriction site (table 2.3). PCR products containing the *hPB* sequence were cloned into the PGEM-T-Easy vector system, which was then digested with either *FseI* and *ClaI* (*MYC-TA-hPB*) or *FseI* and *BbvcI* (*hPB* only) and the relevant fragment cloned into the *Vk*Myc* backbone. Sequence was verified using capillary sequencing and the constructs were digested (*MluI/PmeI* for *Vk*MYC-TA-hPB*, *MluI/EcoRV* for *Vk*hPB*) and the correct fragment purified and sent for pro-nuclear injection (PolyGene Transgenics, Switzerland). Three transgenic C57Bl/6N mice were generated for the *Vk*MYC-TA-hPB* construct and four for the *Vk*hPB* construct. These were imported and re-derived into the clean area of the WTSI animal facility. One of each of these lines was chosen to generate the insertional mutagenesis cohorts by mating with low copy *GrOnc* transposon mice. 40% of the IM mice in the *Vk*hPB* cohort and 47% in the *Vk*MYC-TA-hPB* cohort, along with a matched number of controls, received a single intra-peritoneal injection with 500µL of a 2% solution of sheep red blood cells (Sigma-Aldrich R3378) in PBS at 8-16 weeks of age as a form of antigen stimulation.

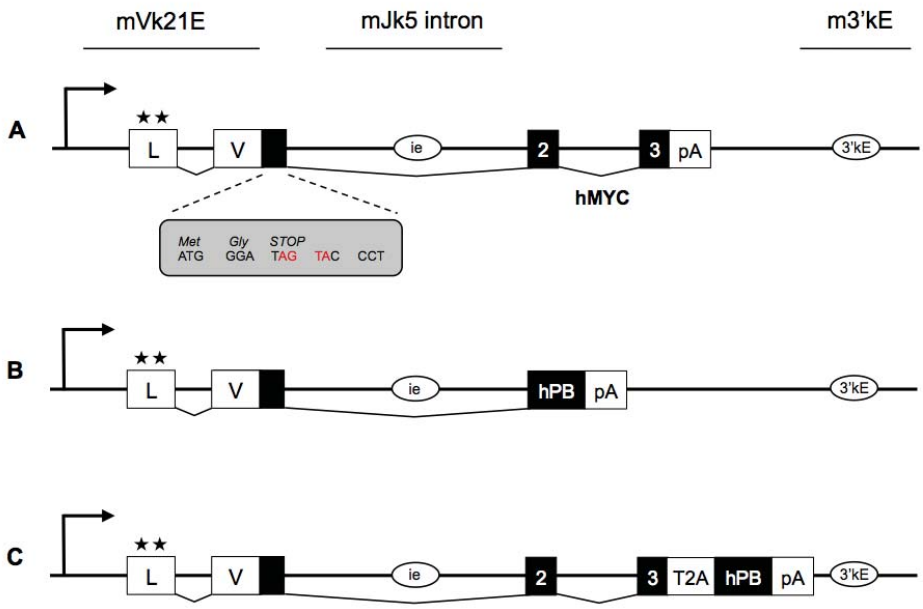


Figure 2.1: The *Vk*MYC*, *Vk*hPB* and *Vk*MYC-TA-hPB* constructs.

(A) In the *Vk*MYC* construct published by Chesi et al, the Jk5 exon in the rearranged mouse Vk21 kappa light chain gene was replaced by a short coding exon containing a Kozak ATG(Chesi et al., 2008) . Human *MYC* exons 2 and 3 replaced the C1k region. Transcription initiates at the Vk21e proximal promoter (↱), extends to the leader (L) and Vk (V) exons, splices in frame to human *MYC* (*hMYC)* and terminates at the endogenous polyA signal (PA). ATG codons (*) in L were mutated to ACG to stop initiation of translation at these positions. Intronic (ie) and 3'kappa (3'kE) enhancers are maintained. The DNA sequence immediately downstream of the Vk21 ATG is depicted. Nucleotides in red letters fit the DGYW consensus for AID targeting.

(B) In *Vk*hPB, hMYC* is replaced by the *hPB* cDNA, carrying a splice acceptor signal that leads to splicing of *hPB* mRNA in-frame with the reading frame "opened" by AID mutation of the upstream TAG stop codon.

(C) In *Vk*MYC-TA-hPB* the cDNA for the self-cleaving peptide T2A links *hPB* in-frame to *hMYC*. The chimaeric polypeptide produced from a single cistron is predicted to spontaneously dissociate into hMYC and hPB proteins.

Primer	Sequence
hMYCTAHyperPB_F	TCACTATAGGGAGACCCAAGC
hMYCTAHyperPB_FseI_R	ACTTCAGGCCGGCCCATAGAGCCCACCGCATC
VkmPBHyperPB_F	ATTCTTCCTCAGCCCCTCAA
VkmPBHyperPB_FseI_R	ACTTCAGGCCGGCCCATAGAGCCCACCGCATC

Table 2.3 Primers to add FseI restriction sites to the linker-*hPB* plasmids

2.2.4 Genotyping Transgenic Mice

Ear or tail biopsies, lysed overnight at 55˚C in ear lysis buffer with proteinase K (300ng/µL), were used as the DNA source. In most cases the genotype was later validated in tumour or spleen samples from diseased mice. Genotyping primers used in the *Vk*MYC-TA-hPB* and *Vk*hPB* IM cohorts are shown in table 2.4. Genotyping primers for the SB cohort were as previously described (Vassiliou et al., 2011). The primers to assess mobilisation of the transposon were also as published (Rad, 2010).

Standard PCR reactions contained 11µL REDTaq ReadyMix (Sigma-Aldrich), 7 µL H_20, 2µL DNA (or cDNA) and 1µL each of the forward and reverse primers. Primers were obtained from Sigma Genosys. Standard PCR conditions were 94˚C for 2 minutes, 36 cycles of 94˚C/57˚C/72˚C each for 30s, then 72˚C for 10 minutes. Amplified DNA was loaded directly onto a 2% agarose gel. The 'Jump' and 'No-Jump' PCRs were suboptimal using this standard protocol and were performed using the KAPA mouse genotyping kit with a 25µL reaction volume (12.5µL 2x KAPA2G fast genotyping mix, 1.25µL each primer (10µM), 1µL template DNA and 9µL H_20) with PCR conditions 95 ˚C for 3 min, 35 cycles of 95˚C/57 ˚C/72˚C each for 15s, then 72 ˚C for 10min.

Primer	Sequence
Genotyping Primers *Vk*MYC-TA-HPB* Construct	
Myc-ex3_to_HPB_F	AAGAGGACTTGTTGCGGAAA
Myc-ex3_to_HPB_+257R	CTCCTCGGTGTCGGACTG
Myc-ex3_to_HPB_F2	GGAAACGACGAGAACAGTTGA
Myc-ex3_to_HPB_+279R2	TGGTAGGCTGCACCTCGT
Genotyping Primers Vk*hPB Construct	
VkHPB_3345F	CATCCTCTGTGCTTCCTTCC
VkHPB_3729R	CTGGCTTCTCACGATGTTCA
VKHPB_3233F	TGGCCCATTGTTCCTTATCT
VKHPB_3578R	TTCTGCTCGTCCAGGATCTC
Primer for detection of mobilised and non-mobilised transposons†	
Jump 2F	GGGCCTCTTCGCTATTACG
Jump 2R	GGTCGAGTAAAGCGCAAATC
No-jump 1F	GGGCCTCTTCGCTATTACT
No-jump 1R	CCGATAAAACACATGCGTCA
GrOnc genotyping PCR#	
LunSD_F	CGCGAGGATCTCTCAGGTAA
LunSD_R	AACCTCTGCCCTTTCTCCTC

Table 2.4: Genotyping primers. Previously published primers are indicated † (Rad, 2010) and # (Vassiliou et al., 2011).

2.3 Sample Collection and Processing

2.3.1 Collection and processing of blood samples from live mice

The mice in the ***SB*** serial analysis study were bled fortnightly from a tail vein after pIpC injection. A small incision was made over a tail vein and approximately 75µL of blood was collected into a Microvette 200µL potassium EDTA capillary tube (Sarstedt). Blood counts were performed on a VetABC Haematology Analyser (Horiba ABX), an air dried blood smear was prepared, and the remaining sample was processed for DNA extraction as described below.

A cohort of mice in each of the ***PB*** insertional mutagenesis cohorts were bled monthly by tail vein injection and approximately 100µL of blood was collected into a Microvette 200µL BD SST clot activator gel additive tube (Sarstedt). Samples were mixed and left at room temperature for 30 minutes, before being spun down at 6.5g for 90s. The serum was transferred to an Eppendorf and stored at -20°C. Serum protein electrophoresis was subsequently performed in batches using the SAS-MX SP-10 gel kit and chamber (Helena Biosciences) according to the manufacturer's instructions.

Blood sampling in the ***PB*** mice was done by the animal technicians, whereas I performed the majority of the tail bleeds in the ***SB*** insertional mutagenesis cohort.

2.3.2 Necropsy of sick mice, sample collection and processing

The insertional mutagenesis cohorts were checked twice daily by our animal technicians for signs of illness. Timely euthanasia of sick mice was performed using rising concentrations of carbon dioxide after signs of significant illness or distress were observed. Subjective physical signs included, but were not limited to; inactivity, hunched posture, poor grooming, pallor, visible masses, abdominal distension, respiratory difficulty and hind-limb paralysis.

Blood was collected at necropsy by intra-cardiac aspiration, placed into EDTA and serum tubes and processed as described in 2.3.1. At necropsy, gross examination of the internal organs was performed with particular attention to the spleen, thymus, lymph nodes, liver and kidney. Macroscopic abnormalities and the weights of the whole mouse, one kidney (*PB* cohort only), spleen and liver were recorded. Samples of spleen, tail and any macroscopically abnormal tissue were collected into RNAlater and kept at room temperature for 24-48 hours before storage at -20°C. For later

collection of live cells, bone marrow (BM), spleen or other abnormal tissue were collected into PBS ± 2% fetal calf serum (FCS).

Sections of spleen, liver, kidney and spine as well as the heart, lung, thymus and femur were routinely placed in buffered formalin (10%) and transferred to the Addenbrooke's Hospital Tissue Bank for processing. Formalin-fixed, paraffin-embedded sections were stained with haematoxylin and eosin and haematopoietic tumours were stained for T, B and myeloid markers using rabbit anti-mouse CD3 (Abcam; UK), rat anti-mouse B220 (CD45R; R&D systems) and rabbit anti myeloperoxidase (Dako). The secondary antibodies were Biotin-conjugated donkey anti-rabbit IgG and Biotin conjugated donkey anti-rat IgG (Jackson ImmunoResearch). Anti-cMyc staining was performed using c-Myc (N-262) rabbit polyclonal antibody (Santa Cruz Biotechnology). Attempts to perform c-Myc immunohistochemistry using antibodies directed to the 9E10 epitope, specific to human Myc were unsuccessful(Evan et al., 1985) (Jac6, NB600-704, Novus Biologicals and ab10910, Abcam).

The histopathology and the majority of the immunohistochemistry were reviewed by an experienced histopathologist, Gary Hoffman, who was blinded to the mouse genotypes.

2.3.3 Processing of live cells

After removing the tip of the head of the femur (or tibia), BM was flushed on ice using a 19G needle and 10mL PBS with 2% FCS. The BM suspension was passed through a 40μm filter and spun at 250g. The cell pellet was re-suspended in 5mL 0.85% ammonium chloride for 5 min to lyse red blood cells (RBC) then processed in a similar manner to spleen and tumour cells.

Spleen and tumour samples were gently squashed in 5mL of 0.85% ammonium chloride using the end of a 5mL syringe, and the solution was then pipetted up and down to create a cell suspension. This was filtered to remove cell clumps, transferred into a 15mL falcon tube and RPMI 1640 media (+10% FCS and 1% glutamine-penicillin-streptomycin) was added and the red cell lysed samples were spun for 5 minutes at 250g. A wash step was performed and an aliquot of this solution was taken for counting. Cells were re-suspended at a concentration of 2 x 10^7 cells/mL, frozen in 10 million cell aliquots in a 50:50 mix with 2x RPMI freezing

media (60% RPMI, 20% FCS and 20% dimethylsulphoxide (DMSO)), and stored initially at -80°C before transfer to liquid nitrogen. Aliquots of 0.5mL of the cell suspension were also spun down in microtubes and the cell pellets were re-suspended in 1mL Trizol and stored at -80°C, or stored as a frozen cell pellet at -80°C.

2.3.4 Generation of single cell derived haematopoietic colonies for transplant

Frozen spleen cells from leukaemic mice were thawed and resuspended at a concentration of 150000cells/mL in Iscove's Modified Dulbecco's Media (IMDM). Aliquots of 100µL and 300µL of this suspension were each mixed into 3mL of MethoCult® GF M3434 media, plated across two wells of a 6-well plate and incubated at 37°C. After nine days of growth, ten discrete colonies for each primary tumour were picked into 1.5mL of RPMI media and incubated at 37°C for 30 minutes. The tubes were then spun down at 250g and the supernatant removed leaving 100µL of media in which the cells were re-suspended and injected into NSG mice via the tail vein.

2.3.5 Preparation of Metaphase Spreads and FISH analysis

Spleen samples were collected at necropsy from leukaemic mice and placed in PBS with 2% FCS. Red cell lysis, filtering and resuspension in RPMI media was performed as described above and the cell suspension, in 5mL RPMI media was transferred to a single well of a six well plate. Demecholchicine (D1925, 10µg/mL, 100µL) was added and after mixing the cells were incubated at 37°C for three to four hours. The cells were then spun down at 250g for 4 minutes in a 15mL Falcon tube, the supernatant removed and the tube flicked to break up the pellet. 5mL of hypotonic KCL (0.56%) was drip-added whilst mixing to avoid clumping and this suspension was incubated for 15min at 37°C to swell the cells. Freshly made methanol and acetic acid fixative (3:1) was then added (5mL) and the sample spun down at 300G for 4 minutes. This fixing process was repeated once, before the cell pellet was re-suspended in 2mL of fixative and transferred to a 2mL Eppendorf for storage at -20°C or dropped onto slides to make chromosome spreads.

The transposon specific probe was prepared by digesting the *pA6GrOnc* plasmid with *AflII* to generate a 2.5kb transposon specific probe. Probe amplification, labelling and

preparation and processing of slides was performed by Ruby Banerjee (WTSI FISH facility) as previously described (Rad, 2010).

2.3.6 DNA extraction

For spleen and tumour samples, a 2mm diameter section of tissue from the RNAlater sample or a $1x10^7$ cell pellet was lysed overnight at 55°C in 500µL Qiagen cell lysis solution with proteinase K (3µL of 20mg/mL solution). For the blood samples, red cell lysis was performed by incubating in 0.85% ammonium chloride for 3 minutes, the white cell pellet was washed in PBS and lysed overnight. The next day 3µL of RNaseA solution (Qiagen) was mixed into the cell lysate before incubation at 37°C for one hour. After cooling on ice protein was removed using 200µL protein precipitation solution (Qiagen). The samples were vortexed vigorously, spun at 15000rpm for three minutes and the supernatant was moved into a clean Eppendorf. Isopropanol (600µL) (±1µLpellet paint) was added before mixing and spinning at 15000rpm for 1 minute. The supernatant was removed and the pellet washed in 70% ethanol, before air drying. The DNA pellet was re-suspended in 50µL of water and quantified using Nanodrop or Qubit.

2.3.7 Exome Sequencing of Mouse *SB* Tumours

Library preparation and sequencing was done through the Sanger pipeline using the Illumina HiSeq 2000 sequencer to generate 75bp paired reads. The alignment and variant call analysis were done by Dr Ignacio Varela. The reads were aligned using the BWA algorithm (Li and Durbin, 2010), against a modified version of GRCm38 mouse reference genome in which an extra register with the Sleeping Beauty transposon sequence was included. PCR duplicates were marked and ignored using Picard tools (http://picard.sourceforge.net), and local realignment was performed using GATK (McKenna et al., 2010). Both tumour and normal DNA samples were sequenced in order to identify somatic mutations. Additionally, a collection of normal DNA samples from syngeneic mice was used to improve identification of germline variants. Substitutions were called using an in house written Perl script (Conte et al., 2013) and indels were called using Pindel (Ye et al., 2009).

2.3.8 Comparitive Genomic Hybridisation (CGH)

CGH was performed using the Agilent Mouse CGH 244K array (014695). The amplification, labelling, microarray hybridisation, scanning, data extraction, QC and analysis was performed by the Microarray facility.

2.3.9 RNA extraction

Trizol samples were spun at 12000g for 5 minutes at 4˚C and the supernatant transferred to a clean tube. After standing for 5 minutes at room temperature, 0.2mL of chloroform was added for each 1mL of trizol, before mixing vigorously. Samples were then stood for 10 minutes at room temperature, before centrifugation at 12000g for 15 minutes. The aqueous phase was transferred to a fresh tube, 0.5mL of isopropanol was added and mixed before standing at room temperature for 7 minutes. The sample was then spun at 12000g for 10minutes at 4˚C. The precipitated RNA pellet was washed in 75% ethanol, air dried and dissolved in 100µL RNase- free water.

2.4 Sequencing transposon integration sites: the Roche 454 Method

2.4.1 Splinkerette PCR to identify transposon integration sites

As the identity of the sequence is only known at the transposon end, a linker-based PCR method is used to amplify the transposon integration sites. The splinkerette is a double stranded linker, which also contains an unpaired region, with the unpaired extension of one strand forming a hairpin within itself(Devon, 1995). DNA was digested with *MboI*, a restriction enzyme which cuts frequently throughout the genome and leaves a GATC 5' overhang. The splinkerette adaptors were ligated to digested genomic DNA forming the template for PCR amplification. The PCR uses one primer complimentary to the transposon sequence and a second which is identical to the unpaired region of the non-hairpin strand of the adaptor (figure 2.2). The specificity of the PCR amplification for transposon integration sites is because the second primer cannot anneal and initiate priming until the complement of the unpaired region of the adaptor is generated by extension from the transposon. A nested second-round, barcoded, 454-ready PCR step further improves specificity.

The Splinkerette adaptors were prepared by combining 150pmol of each oligonucleotide (HMSpAa/HMSpBb) in 5µL of buffer 2 (NEB) and adding water to a total volume of 100µL. The solution was heated to 100°C for ten minutes and then allowed to cool slowly to room temperature before storage at -20°C. The restriction digests were performed in a 96-well plate in a total volume of 10µL using 1µL of *MboI* (NEB). *MboI* cuts at 'GATC' sites and leaves a 5' 'GATC' overhang. Restriction digests were performed overnight at 37°C before the enzyme was heat-inactivated.

For the ligation reaction, 5μL of the digested product was annealed with 3μL of the pre-annealed splinkerette oligonucleotides in a total volume of 10μL. The ligation was performed overnight (16°C), before heat inactivation of the T4 ligase. First round PCR reactions were performed using Sigma REDTaq ReadyMix, 2μL of template and 2μL each of the primers (10μM) in a reaction volume of 40μL. PCR conditions were as follows: 94°C 60s; 68°C 30s; 72°C 60s for 2 cycles then 94°C 30s; 65°C 30s; 72°C 2min for 30 cycles, followed by final extension at 72°C for 10 minutes. For the second round PCR, 3μL of 1 in 100 diluted first round product was amplified in a total volume of 31μL using 3μL of each primer (10μM). The PCR conditions were identical to the first round except for omission of the first two cycles. The Splinkerette linker and primer sequences are given in table 2.5.

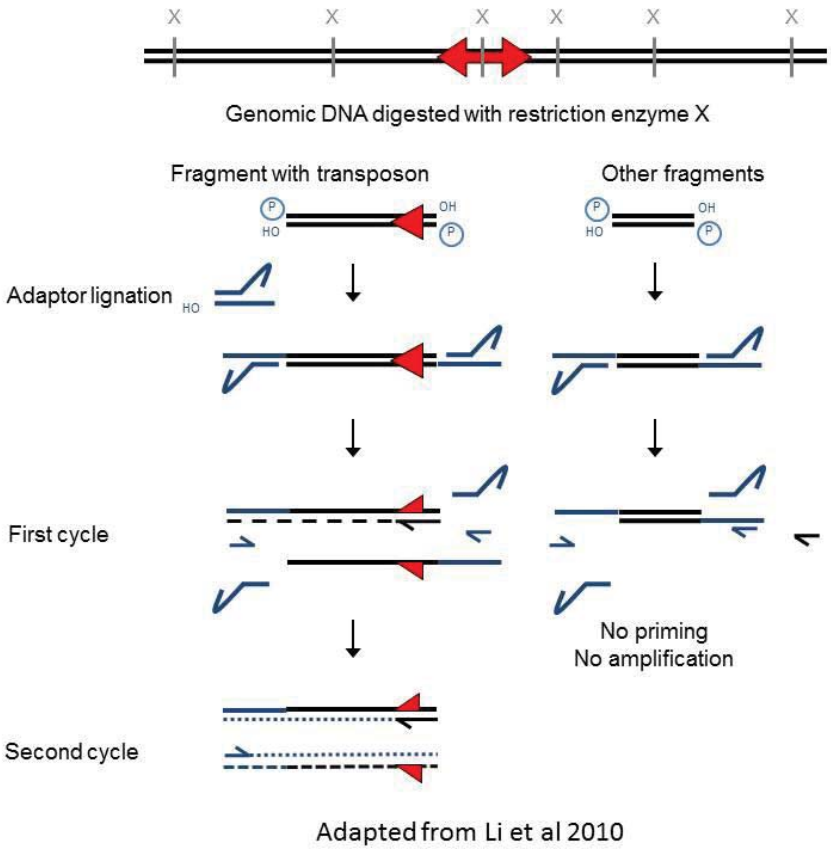


Figure 2.2: Principle of the Splinkerette PCR

An 8μL aliquot of each second round PCR product was run on an agarose gel to ensure adequate amplification and the remainder was pooled, purified through a Qiagen column and submitted for sequencing on the 454 platform (Roche). Sequencing reads were mapped to the mouse genome using the Genomic Insertion Annotation Tool ("GIANT") algorithm created by Stephen Rice (WTSI, Core Informatics Group) (Vassiliou et al., 2011).

Splinkerette Linkers	
HMSpAa	CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGTGT CGACACTAGTGG
HMSpBb-Sau3AI	GATCCCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCAAAAAAA
Splinkerette Primers	
HMSp1	CGAAGAGTAACCGTTGCTAGGAGAGACC
SB-5'-Sp1	TAGTGTATGTAAACTTCTGACCCACTGGA
SB3'_altP1	AACTGACCTTAAGACAGGGAATCTT
HMSp2_454_new2010_R	CTATGCGCCTTGCCAGCCCGCTCAGGTGGCTGAATGAGACTTGGTGTCG AC
BC454-SB1	CGTATCGCCTCCCTCGCGCCATCAGACACATACGCGTGTATGTAAACTTC CGACTTCAAC

Table 2.5: Splinkerette linkers and primers

2.4.2 Transposon mapping and common integration site (CIS) analysis of 454 data

Sequences were filtered to include only reads which contained the primer sequence, then the end of the SB repeat, followed by genomic sequence. Each raw sequence read was screened for the SB primer and the 10bp barcode by blasting against a database of the barcode-primer sequences. The best hit was identified for each read and the alignment data was used to identify those where the primer and barcode were at the beginning of the read sequence. Reads with less than 93% identity with the primer sequence were discarded. Reads were included in further analysis only if the barcode sequence was unambiguous. Reads which satisfied these filtering criteria were then trimmed at the 5' end to remove the primer sequence before further analysis.

Each read was checked for a GATC sequence (MboI restriction site). To remove multiple-ligation artefact, any sequence downstream of the first GATC was removed. The length of the remaining sequence was assessed and reads of less than 20bp were removed from further analysis. Reads which did not contain a GATC sequence were also excluded if they had less than 50bp mapping to the genome. Finally any reads that did not start with the expected TGTA sequence (end of the SB repeat and integration site) were excluded.

GrOnc insertions passing this initial filtering process were mapped to the mouse genome (NCBI37/mm9) using SSAHA2 (http://www.sanger.ac.uk/resources/software/ssaha2/). To quantitatively assess the uniqueness of the alignment a normalised score difference (NSD) was calculated as follows:

NSD = [(Score of best hit) – (Score of second-best hit)] / query length * 100

A previous analysis performed by Stephen Rice on a randomised set of 5000 mouse genomic fragments found that 96.5% of correctly-mapped reads and only 1.5% of wrongly mapped reads had an NSD ≥4. Reads with NSD <4 were removed from analysis.

The genomic co-ordinate at the start of the alignment was determined as the integration site for each read. Reads were then grouped according to barcode and listed by integration site. Redundant sequences mapping to the same location in the same tumour were 'collapsed' into a single integration. The script I used to process the data up to this point was written by Stephen Rice. All of the data were stored on a MySql database.

To identify common integration sites (CIS) non-redundant insertions were analysed by Stephen Rice, using the CIMPL R package provided by Jelle ten Hoeve. The common insertion site mapping platform (CIMPL) is based on the Gaussian Kernel Convolution (GKC) framework (de Ridder et al., 2006). Data were analysed using 10kb, 30kb, 60kb and 100kb scales (windows), with the significance threshold set at 5%. Bonferroni multiple testing correction was applied.

Due to the local hopping phenomenon a single tumour could contain multiple integrations around a site. To minimise the impact of local hopping, integrations in a

single tumour that occurred within a 10Kb window were collapsed to a single integration for the CIS analysis. This was the same method as used in the published *Npm1*^{cA} high copy (GRH) transposon model (Vassiliou et al., 2011).

The CIS identified from all windows in the analysis were merged to compile a CIS list for comparison to the published AML insertional mutagenesis cohort. The default analysis was the 10Kb 'lockout' including all reads with NSD of ≥4. These CIS were reviewed manually to remove questionable CIS sites, as was done in the published *Npm1*^{cA} GRH transposon model. Reasons for editing CIS from the list included:

1. The genomic Engrailed homeobox 2 (En2) locus was excluded because part of the gene sequence is present within the transposon cassette.
2. Occasionally multiple integrations from the same tumour were seen in the CIS window despite the 10kb 'lockout'. A second analysis was performed in which reads within a 100kb window in the same tumour were excluded. CISs where the 10kb analysis included multiple integrations from the same tumour were excluded if they were lost on the re-analysis using the 100kb 'lockout'.
3. In rare cases the hits contributing to the CIS were all from tumours analysed on the same 96-well plate in the same 454 sequencing run and mapped to the same base position. It is probable such integrations were a result of cross- contamination and such sites were excluded from the CIS list. CIS were excluded on this basis if more than half of all hits contributing to a CIS came from a single experiment and more than half of hits from that experiment were at an identical site.
4. If the identity of the alignment was ≤98% in at least 20% of hits, the CIS was excluded.

Despite the NSD threshold, several reads with an NSD score between 4 and 7 were found to blast with good alignment to multiple locations in the genome. Therefore the CIMPL analysis was re-run using an NSD cut off of ≥7.

The CIMPL program also has an inbuilt local hopping correction (LHC) method. This works on the basis that when the distance between two neighbouring insertions is less than three kernel widths, the insertion with the smallest 'contig_depth' is considered 'hopped'. Using read count as 'contig_depth' Stephen Rice re-ran the

CIMPL analysis with the LHC filter on. It is important to note that the read depth does not directly correlate with the number of DNA molecules in the tumour with that transposon insertion because the method involves numerous rounds of PCR amplification. This will introduce PCR amplification bias, for example, due to variation in the proximity of the nearest *MboI* digestion site. The CIS identified on the LHC CIMPL run were also reviewed manually and integrations where multiple tumours from the same run had the same integration site were removed as previously described, as well as integrations were the identity of the read was <98% in over 20% of hits. The CIS identified using each of these CIMPL methods (original as per GRH, NSD ≥7 and LHC) were compared.

The CIS analysis on the pre-leukaemic blood samples was performed using the same analysis method as used in the published GRH cohort (Vassiliou et al., 2011). All of the serially bled *Npm1*^{cA} mutant insertional mutagenesis mice that received any pIpC were included in this analysis. Samples from these mice were grouped by the number of days prior to sacrifice that the blood sample was taken.

2.4.3 Detecting Intra-GrOnc Jumping using PCR, Splinkerette and Sequencing

If one of the *SB* repeats is 'lost' or mutated, the transposon, or part of it will be unable to re-mobilise. This could cause a persisting integration on serial blood or transplant samples, even if the integration was not leukaemogenic (and therefore selected during tumour evolution). For example, this could happen if a transposon re-inserts into a neighbouring transposon sequence by local hopping and then on remobilising jumps using two 'non-partner' repeats (figures 2.3 and 2.4). To explore this possibility I first designed primer pairs to determine if transposons were jumping into each other (figure 2.3B). Additionally, I designed primers to detect whether after such an event transposons can jump out of each other again, using previously unpaired repeats (figure 2.3C).

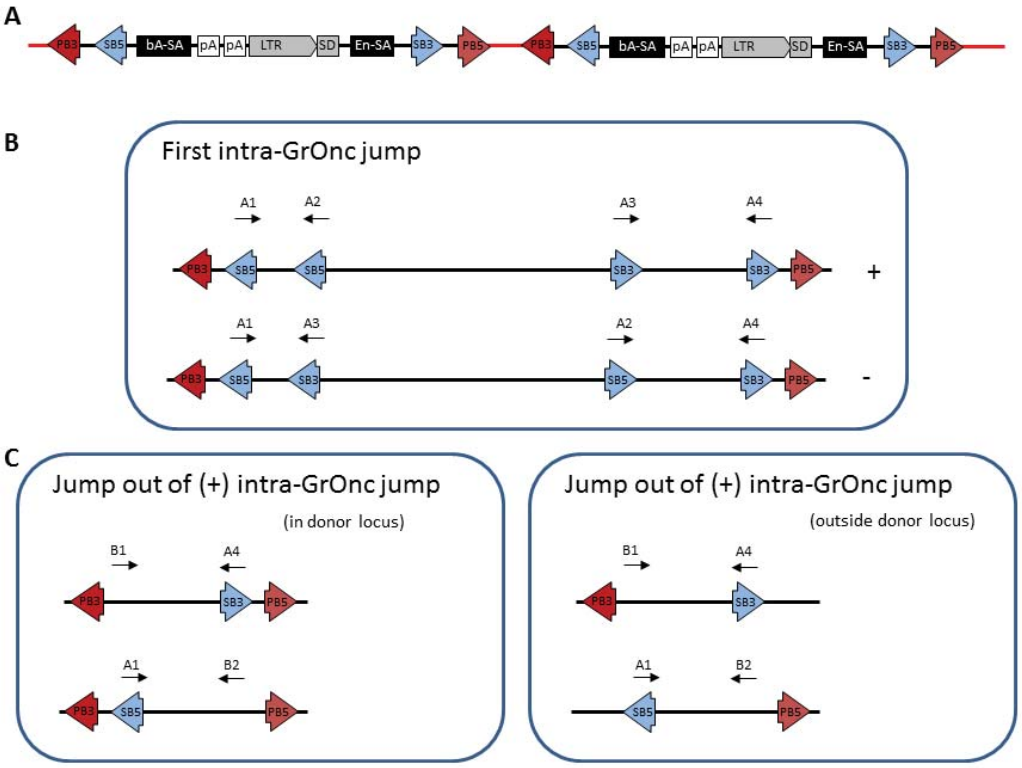


Figure 2.3: Positioning of primers to detect jumping of the SB transposons into adjacent transposons. Two adjacent *GrOnc* transposons in the donor locus (A), which have the outer PB repeats present unlike two transposons mobilising together to a new locus by SB. Example of one GrOnc transposon jumping into another at the donor locus and primers designed to 'capture' such an event (B). Example of jump out of the '+' host transposon shown in B, using SB5 and SB3 repeats that previously belonged to different transposons and primers designed to 'capture ' such an event (C, left). The same primers would capture this event happening outside the donor locus (C, right).
+: the orientation of the host and inserted transposons are the same.
-: the orientation of host and inserted transposons are opposite.

If more than one transposon mobilised together from the donor site, this would move adjacent 5' and 3' *PB* repeats together into a genomic locus where they would become 'fixed' on remobilisation of the individual transposons (figure 2.4A). If an insertion of one transposon into another happened after they had mobilised together from the donor site this would also leave the *PB* repeats fixed in the genomic locus (figure 2.4B), and if a remobilisation happened after this insertion, this could generate specific *PB-SB* repeat configurations (figure 2.4C). With such events, the 'blunt' end of the *PB* repeats would be left adjacent to genomic DNA and could therefore be mapped using splinkerette PCR. The protocol used for Splinkerette from the blunt end of the *PB* repeat was identical to that used for the standard

Splinkerette PCR, except for the modified primer sequences and an additional *AflII* digestion step. This digestion was used to cut between adjacent *PB* and *SB* repeats found in native unmobilised transposons at the donor locus (figure 2.3A), and avoid amplification of such *PB-SB* junctions. Primer sequences are shown in table 2.6.

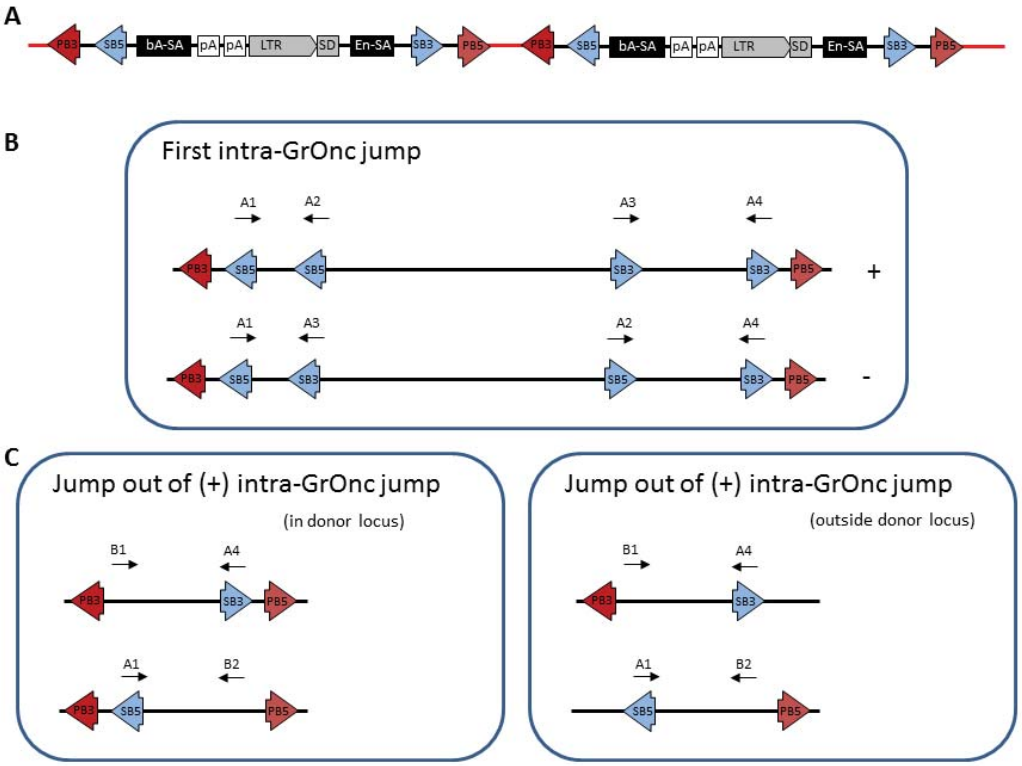


Figure 2.4: Transposon Neopartnerships. A. Two adjacent transposons that have jumped together into a genomic locus. Note the opposite facing *PB* repeats are transported as 'cargo' between them. **B**. One *SB* transposon integrates into the adjacent transposon by local hopping. **C.** Two possible re-mobilisation events using *SB*5 and *SB*3 repeats that previously belonged to different transposons (neopartnerships) and leaving a lone *SB* repeat 'sequence-fixed' at the genomic locus. Red lines represent the sequence found between adjacent transposons in the donor locus.

Primer	Sequence
SB5_A1	CTGTGCCTTTAAACAGCTTGG
SB5_A1b	CAGCTTGGAAAATTCCAGAAA
SB5_A2	TGTCCTAACTGACTTGCCAAAA
SB3_A3	GACAGGGAATCTTTACTCGGATT
SB3_A4	GAGGTCAGAGCTTTGTGATGG
PB_B1	CGCATGTGTTTTATCGGTCT
PB_B2	TGACGAGCTTGTTGGCTAGA
PB5_Bl_Sp1	TGAGCATATCCTCTCTGCTCTTC
PB5_blunt_sp2new	ATGACGAGCTTGTTGGCTAGA

Table 2.6: Primers used to screen for intra-*GrOnc* jumping and for splinkerette from the blunt end of the PB transposon.

2.5 Illumina Sequencing of Transposon Integrations

2.5.1 Library Preparation

DNA was extracted from blood, spleen or tumour mass samples as previously described. Samples were quantified by Qubit fluorometer (Life Technologies) using 1µL of sample DNA and the ds-DNA Broad Range dye-buffer mix, following the manufacturer's instructions. A quantity of 2µg of DNA was used for library preparation, diluted to 100µL in sterile MilliQ water.

A method for eukaryotic transposon direct insert sequencing (TraDIS) was developed by Iraad Bonner (Sequencing Research and Development team, WTSI). The library preparation on the *SB* and *PB* samples was performed by him using the method described below and the primer and adaptor sequences shown in appendix 2B.

Genomic DNA was sheared in a Covaris 96microTUBE plate with the following settings to shear at 250bp: duty cycle 20%; intensity 5, cycles per burst 200, time 60s, temperature 4°C to 7°C. After shearing the samples were spun briefly at 1000rpm and then purified using the QIAquick column system according to the manufacturer's instructions in an elution volume of 80µL. The shearing quality was then assessed using the Agilent 2100 Bioanalyser and DNA 7500 chip and reagent kit.

End-repair of the sheared and cleaned DNA was performed in a total volume of 100µL: 10 x T4 DNA ligase buffer, 10µL; T4 DNA polymerase, 7 µL; T4 PNK, 7µL; dNTPs, 6µL; Klenow DNA polymerase 2µL; sheared DNA 78µL. After incubation at 20°C for 30 minutes the samples were purified using the QIAGEN 96 well plate column system and eluted in 27µL of EB buffer. An A tail was added using the following reaction: Klenow fragment exo-, 4.5µL; dATP, 15µL; 10 x Klenow buffer, 5µL; incubate at 37°C, 1 hour. The samples were eluted through MinElute columns in 20µL of EB buffer and 1µL of the sample was run on a 7500 Agilent chip. The splinkerette adaptors were ligated using NEBNext(TM) DNA Sample Prep Reagent Set 1 (NEB: E6000B-SS) using the reaction parameters as follows: 18µL A-tailed DNA; 25µL 2x ligation buffer; 1µL MilliQ water; 5µL ligase; 30 to 60 minute incubation at 20°C. Clean up of the DNA was performed using a double SPRI bead (AMPure XP) purification, using 50µL beads to 50µL of adapter ligated DNA in the

first reaction and 40µL (0.8x) in the second in a similar method to that described above (SPRI bead purification of MiSeq library). The final volume of the purified, adapter ligated DNA in EB buffer was 30µL and 1µL of this sample was run on the 7500 Agilent chip to determine whether the adapter was successfully ligated. The size of the adapter ligated DNA should be approximately 100bp larger than the pre-ligated library.

Two paired rounds of PCR were then performed on the adaptor ligated DNA to generate the 3' and 5' transposon sequencing libraries. The first round PCR mix was prepared on ice to a total volume of 50µL: 7µL adaptor ligated DNA; 25µL 2x Kapa HiFi HS ReadyMix; 17µL sterile MilliQ water; 0.5µL transposon specific primer (100µM) and 0.5µL Splinkerette adapter nested primer 1 (SplAP1) (100µM). PCR conditions were as follows: 95°C, 2 min; 95°C 20s, °C 20s, 72°C for 40s repeat for 18 cycles; 72°C 5min. An AMPure XP bead purification step was performed on the first round PCR products using 0.8x beads and eluting into 25µL of EB buffer and 1µL of this product was run on an Agilent High Sensitivity chip. The second round PCR was also in a total volume of 50µL: PCR1 product 24µL; 2x Kapa HiFi HS ReadyMix 25µL; Transposon specific primer 2 (100µM) 0.3µL; Splinkerette adapter nested primer 2 (100µM) (SplAP2) 0.3µL. Reaction conditions were as follows: 95°C, 2 min; 95°C 20s, °C 20s, 72°C for 40s repeat for 12 cycles; 72°C 5min. The PCR products were again cleaned up using 40µL of AMPure XP beads and two 500µL, 80% ethanol washes before elution in 30µL of EB buffer. 1µL of the purified DNA was run on an Agilent High Sensitivity chip.

A quantitative PCR was then performed on each sample in triplicate to determine the quantity of transposon specific template within the library and guide the quantity of each sample to be used in pooling. Two reactions were performed with details as follows: KAPA SYBR Fast qPCR Mix, 10µL; sterile MiliQ water, 5.2µL; qPCR primers (10µM), 0.4µL of each. Both reactions used a generic library qPCR reverse primer (2.2), but the first used a generic forward primer (2.1) and the second a transposon specific sequencing primer. For PB libraries a third qPCR reaction was performed to quantify non-specific product in the library using the same volumes and generic reverse primer and a transposon sequencing primer from the opposite end. An aliquot of 4µL of 1:1000 diluted library DNA was used in the qPCR reactions and the reactions were performed in a MicroAmp Fast Optical 96-well reaction plate (Life

Technologies: 4346906) on a StepOnePlus Real-Time PCR System (Life Technologies).

The quantities of each sample were then standardised and pooled for sequencing. Each 96 well plate of samples was run on two 75bp paired end MiSeq runs; one for the 5' and one for the 3' library.

2.5.2 Transposon mapping and CIS analysis of Illumina data

The Illumina transposon sequencing analysis was performed by Hannes Postingl (Core Informatics Group, WTSI). As the majority of reads start with an identical transposon sequence each was sequenced in two parts to allow the Illumina software to correctly align clusters on the MiSeq. The full length read was re-assembled and those which did not start with the expected transposon specific sequence were filtered at this stage.

Illumina paired-end sequencing reads were trimmed of the transposon sequence and mapped to the mouse genome (GRCm38) using the SMALT alignment software (smalt.sourceforge.net). A hash index of 13 base pair words, sampled every fourth base pair along the reference mouse genome, was used and an expected insert range up to 800 nucleotides was specified. The software identified potentially matching segments in the reference genome from the hashed words and aligned them with the read using a banded Smith-Waterman algorithm. The quality score for the reliability of the mapping took into account the expected insert range. The analysis was performed twice, with and without the removal of putative PCR duplicates.

Read pairs which mapped in the expected orientation on the same chromosome, irrespective of the insert range, were included in further analysis. All other reads, including those where the placement was ambiguous or where mates aligned to different chromosomes, were discarded. The reads were filtered further by applying thresholds of the mapping quality score of 20 (expected mapping error rates of less than 1 in 100) and of the Smith-Waterman alignment score of 30 (using standard affine gap penalties of 1, -2, -4 and -3, respectively, for matches, mismatches, gap openings and gap extensions). In addition the presence of the integration motif, TA for *SB*, TTAA for *PB*, at the mapped location of the 1st mate (sequenced out of the

transposon) of each pair was checked. After this filtering step the 1st mates were sorted by barcode and integration site. A putative insertion site had to be covered by a least two independent reads mapping to the same location (twenty reads for non-duplicate filtered data).

CIS analysis was performed using the same CIMPL program as described for the 454 sequencing. Ten kernel widths sizes were chosen at 10,000 base pair intervals between 10,000 and 100,000. The analysis was performed using the in-built local hopping filter with default settings.

2.6 Additional methods for the *Vk*MYC-TA-hPB* and *Vk*hPB* models

2.6.1 Validation of splicing in the transgenic constructs

The human myeloma cell line U-266 ($1x10^7$ cells) was transfected with 10µg of construct DNA by electroporation in a 0.4cm cuvette at 220V and 900µF. Cells were then transferred to a 10mL flask and incubated in 90% RPMI/10% FCS media at 37˚C for 24 hours, harvested, washed in PBS and lysed in Trizol (Invitrogen). RNA was extracted, treated with DNaseI, reverse transcribed (Superscript II, Invitrogen) and subjected to RT-PCR using primers in Vκ, *hPB* and in different exons of *hMYC* (table 2.7). Controls included samples not treated with DNaseI and/or reverse transcriptase. RT-PCR products were run on a 2% agarose gel at 150V.

The same primers were subsequently used to check for reversion of the stop codon and to validate the splicing of the transgene in tumour samples from IM mice. These RT-PCR products were sent for sequencing using the Sanger or MiSeq sequencing platforms.

RT-PCR on in vitro samples: 16µL RNA was treated with 2µL DNAseI and 2µL 10x buffer for 15 minutes at room temperature, before adding 2µL stop solution and incubating at 70˚C for 10 minutes. Matched control samples were not treated with DNAseI. 10µL of random hexamers (100µM) were then added to 20µL of RNA, incubated at 65˚C for 10minutes and then placed on ice for 2minutes. The reverse transcription reaction used 15µL of DNAseI treated RNA, 8.4µL water, 8µL 5xMMLV buffer, 4µL DTT, 1.6µLdNTP, 1µL Rnasin (Promega-40U/µL) and 2µL reverse transcriptase(Gibco-GRL). A control reaction, without reverse transcriptase was also

performed. RT-PCR samples were incubated at 37˚C for 90minutes, 70˚C for 10 minutes and then stored at -20˚C until PCR.

RT-PCR of tumour samples: Reactions were prepared on ice in 0.2mL micro-tubes and containing 4µL qScript^{TM} cDNA SuperMix (5x) (Quanta Biosciences), 5µL RNA template and 11µL of RNase/DNase-free water. After mixing, samples were incubated at 25°C for 5 min, 42°C for 30 min and 85°C for 5 min. cDNA was stored at -20°C.

RT-PCR Primers	
VkMycexon1F	TGCTGACACAGTCTCCTGCT
VkMycexon2R	CAGCAGCTCGAATTTCTTCC
VkMycexon2F	CCTACCCTCTCAACGACAGC
VkMycexon3R	ACTCTGACCTTTTGCCAGGA
VkMycexon3Rb	CTCTGACCTTTTGCCAGGAG
SAHPBR (hPB)	CTCACGTGGTCGCTGATCT
SAHPBRCORTAHPBR (TA-hPB)	CTCACGTGGTCGCTCACCT

Table 2.7: Primers used to assess splicing of the Vk* constructs and reversion of the stop codon

2.6.2 In vitro verification of *hPB* activity in the *Vk*MYC* T2A linked construct: HAT resistance assay

MYC-TA-hPB cDNA was generated from U-266 cells previously transfected with *Vk*MYC-TA-hPB,* by high fidelity RT-PCR (Phusion, Finnzymes) using a modified forward primer to revert the in-frame stop codon (table 2.8). This cDNA was cloned into a pGEM-T-Easy vector and used to transform chemically competent E. Coli (One Shot Mach T, Invitrogen). X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) and IPTG (isopropyl-β-D-thiogalactopyranoside) were used for colour selection of recombinants and correct colonies were confirmed by diagnostic digest. One of these clones was digested using *Not1* and *FspI*, run on a 2% agarose gel and the 3000bp band was cloned into the eukaryotic expression vector pcDNA3. The correct clones were selected by diagnostic digest and verified by capillary sequencing.

Primer Name	Sequence
*Vk*MYCTAHPB*-cDNA_F1	CACCATGGGAAAGTACCCTTATGATGTGC
*Vk*MYCTAHPB*-cDNA_R1	GCTCATCAGAAACAGCTCTGG
*Vk*MYCTAHPB*-cDNA_R2	CCGCTCATCAGAAACAGCTC
*Vk*MYCTAHPB*-cDNA_F2	ACCATGGGAAAGTACCCTTATGATG
*Vk*MYCTAHPB*-cDNA_R3	CGCTCATCAGAAACAGCTCTGG
*Vk*MYCTAHPB*-cDNA_F1	CACCATGGGAAAGTACCCTTATGATGTGC

Table 2.8: Primers used for reversion of the stop codon in MYC-TA-hPB cDNA

A hypoxanthine-aminopterin-thymidine (HAT) resistance assay was used to assess the function of the T2A linked *hPB* in ES cells(Liang et al., 2009; Wang et al., 2008). Male AB1:HprtE3 (PB-FL-puΔtk:neo) cells harboured a *PB* transposon within the X-linked hypoxanthine-guanine phosphoribosyltransferase (*Hprt*) locus, inactivating the only copy of the gene (figure2.5). Removal of the transposon by *PB* restores *Hprt* activity and permits growth in HAT media. $1x10^7$ AB1:HprtE3 cells were electroporated (230V/500uF) with pcDNA3-MYCTAHPB (2, 10 or 20µg), pcDNA3-HPB (10µg) or water. Cells were divided into 1/10 and 9/10 aliquots after transfection and grown on 10cm feeder plates in M15 media for 48 hours post transfection, then switched to HAT media. After 7 days HAT-resistant colonies were stained with 2% methylene blue in methanol, washed, air dried and counted.

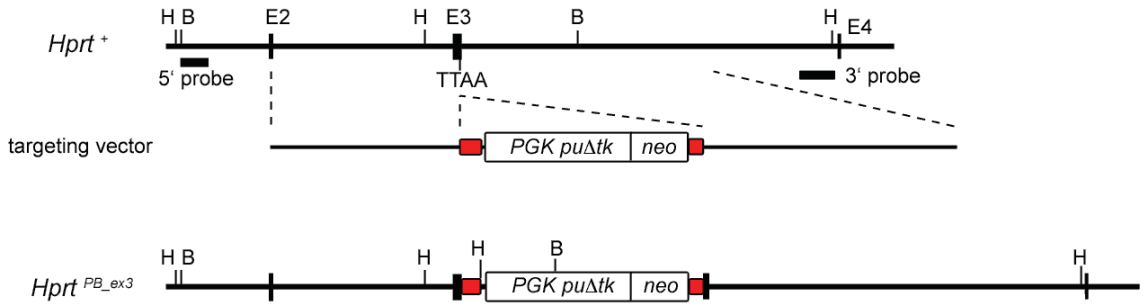


Figure 2.5: *Hprt* locus of ES cells used for the HAT resistance assay: The *PB* transposon, with inverted terminal repeats (red) positioned within exon 3 of the *Hprt* locus. Excision by *PB* transposase restores normal configuration and function of the *Hprt* locus. (Figure and cells courtesy of Kosuke Yusa, Stem Cell Genetics Team, WTSI).

2.6.3 Flow Cytometry

Flow cytometry was performed on the *Vk*MYC-TA-hPB* and *Vk*hPB* lines with the assistance of George Giotopolos and Sarah Horton (CIMR, Huntly lab). Frozen cells were thawed and re-supended in PBS with DNase I to minimise clumping, then spun at 300g for 5min. The cell pellet was re-suspended in PBS with blocking agent (1.2µL of 2.4G2 per mL) and left on ice while the antibodies (Cambridge Biosciences) were prepared. Three antibody panels were used as follows: i) B220 APC 640 670, CD19 PE 561 582, CD3 PECy7, Mac1 FITC, Gr1 PB ii) B220 APC, CD 19 PE Cy7, CD43/ AA4.1 PE, CD24 Pacific blue and iii) B220 Pacific green, CD19 PE Cy7, BP1 PE, IgM FITC, IgD APC. A 100µL aliquot of cell suspension, pooled from multiple samples was used in the control tubes. The samples were made up to a total volume of 300µL in PBS and 95µL was added to the antibody mix (1:100). After mixing the cells were incubated for 45minutes in the dark, before being washed, re-suspended in 300µL of PBS and filtered. Samples were analysed on a BD LSR Fortessa flow cytometer, gating on AAD negative cells and analysis was performed using FlowJo software.

Flow sorting of mouse bone marrow and spleen cells was performed with the assistance of David Kent (CIMR, Green Lab) or Bee Ling Ng (WTSI) using the following panels; (1) CD45 Pacific blue, CD19 PE, B220 APC Alexa 750, Kit APC, Gr1-CD11b FITC and CD3 PerCPCy5.5. or (2) CD34 FITC, CD19 PE, B220 Alexa Fluor 750, CD3 PerCPCy5.5 and CD11b Alexa Fluor 647. Cells were sorted into i)

granulocytes (CD45+, Mac1Gr1+, CD19-), ii) T cells (Mac1Gr1 and CD19 negative, CD3 positive) iii) CD19+,B220+ B cells iv) B220+, CD19- B cells and v) c-kit/CD34 positive progenitor cells, with a target of 100 000 cells in each.

2.6.4 Western Blotting

Protein samples were prepared from stored RNA-later tissues. Tissue was lysed in RIPA buffer supplemented with protease inhibitors. Approximately 30 µg of protein was added per well to a NuPAGE Bis-tris mini gel and transferred to a PVDF membrane. Primary antibodies were diluted in PBS with 5% bovine serum albumin (BSA): human cMyc (Covance) 1:200, mouse c-Myc (Abcam) 1:1000 and β-actin (Abcam) 1:5000. The secondary antibodies were anti mouse IgG HRP (human cMyc and β actin) and anti-rabbit IgG HRP (mouse c-Myc), diluted 1:5000. The detection was performed using an ECL western blotting substrate kit.

2.6.5 B cell receptor repertoire analysis

This analysis was performed by Rachael Bashford-Rogers using DNA supplied to her. Rearranged IgH genes were amplified using a multiplex PCR containing 11 forward primers, each specific to a group of functional IgHV genes, and two reverse primers specific to IgHJ genes (table 2.9). The forward primers were grouped into two pools based on similar melting temperatures and PCR amplification was performed using 70ng DNA, JH reverse primers (25µM) and each of the forward pools (25µM), using 0.5µl Phusion® High-Fidelity DNA Polymerase (Finnzymes), 1µl dNTPs (0.25mM), 1µl DTT (0.25mM), per 50µl reaction. The PCR reaction conditions were as follows: 3 min at 94^{o}C, 35 cycles of 15 sec at 94^{o}C, 30 sec at 60^{o}C and 30 min at 72 ^{o}C, 7 min at 72 ^{o}C. The two amplified sets of PCR products were pooled before sequencing.

Sequencing libraries were prepared and sequenced using standard protocols and the MiSeq platform to generate 300bp paired end reads. Filtering and repertoire analysis was performed as follows: “Raw reads were filtered for base quality (median >34) using the QUASR program (http://sourceforge.net/projects/quasr/). MiSeq paired end reads were co-joined at their overlapping region, and non-immunoglobulin sequences were removed, retaining only reads with significant similarity to mouse IgH from the IMGT reference database(Lefranc et al., 2009) using BLAST(Altschul et al., 1990) (1x10^{-10} E-value threshold). Sequences were trimmed to remove primer sequences, and sequences with a minimum length of 180bp were retained. IgH

sequence network generation was performed according to Bashford-Rogers et al.(Bashford-Rogers et al., 2013a)".

Group 1 forward primers	
VH-for11	CAGATKCAGCTTMAGGAGTC
VH-for13	CAGGTTCACCTACAACAGTC
VH-for15	GARGTGMAGCTGKTGGAGAC
VH-for2	CAGGTGCAAMTGMAGSAGTC
VH-for5	GAKGTGCAGCTTCAGSAGTC
VH-for8	GAGGTGMAGCTASTTGAGWC
Group 2 forward primers	
VH-for1	GAGGTTCDSCTGCAACAGTY
VH-for12	CAGGCTTATCTGCAGCAGTC
VH-for14	CAGGTGCAGCTTGTAGAGAC
VH-for3	GAVGTGMWGCTGGTGGAGTC
VH-for7	CAGRTCCAACTGCAGCAGYC
J Reverse primers	
JH-1_reverse	TCACCGTCTCCTCAGGTAAG
JH-2_reverse	TCACTGTCTCTGCAGGTAAG

Table 2.9: Primers used for the B cell receptor repertoire analysis

3. Whole exome sequencing reveals rapid acquisition of driver mutations and branching evolution in a case of NPM1 positive CMML transforming to AML

3.1 Introduction

Chronic myelomonocytic leukaemia (CMML) is a clonal disorder characterised by the accumulation of monocytes in the peripheral blood together with abnormal myeloid differentiation, which is either dysplastic or proliferative or both. CMML develops due to the stepwise accumulation of genetic mutations in haematopoietic stem cells (HSC). Although several recurrent mutations have been described, none are specific to CMML (Itzykson and Solary, 2013). Cancers are thought to evolve with a complex branching clonal architecture, however the evidence so far in CMML is that the majority of mutations accumulate in a linear manner, with limited branching through loss of heterozygosity (LOH) (Itzykson et al., 2013b). CMML progresses to secondary AML in 20-30% of cases. This is thought to be driven by a clone acquiring a novel fitness conferring mutation, although the expansion of a clone with high fitness in the absence of new genetic lesions has not been excluded (Itzykson and Solary, 2013).

Nucleophosmin (*NPM1*) mutations are described in around 30% of cases of AML, making it one of the commonest driver lesions in this disease (Falini et al., 2005; TCGA_Research_Network, 2013). Mutations in *NPM1* are thought to be an early although not necessarily initiating event (Shlush et al., 2014) and define a large subgroup of AML with distinct clinical, pathological and molecular characteristics (Swerdlow, 2008). These mutations result in cytoplasmic dislocation of the *NPM1* protein and are mutually exclusive to the fusion genes, which are presumed initiating lesions in several other subtypes of AML (TCGA_Research_Network, 2013).

NPM1 is frequently overexpressed in solid malignancies and is involved in chromosomal translocations in various haematological and solid tumours, however *NPM1* mutations are generally considered specific to AML (Falini et al., 2011; Grisendi

et al., 2006). Therapy related AMLs and AMLs secondary to myeloproliferative neoplasms (MPN) or myelodysplasia (MDS) are sometimes found to have cytoplasmic *NPM1*, but *NPM1* mutations are much more common in de novo disease (Falini et al., 2005; Fernandez-Mercado et al., 2012; Gale et al., 2008; Schnittger et al., 2011). Also, *NPM1* mutations have been reported in a small proportion of cases of MDS(Bains et al., 2011; Falini et al., 2011; Zhang et al., 2007). However, *NPM1* mutant AML may have dysplastic features (Falini et al., 2011) and as the distinction between MDS and AML is based on a 20% threshold of detectable blasts, the diagnosis is subject to sampling and inter-observer variation. It is therefore unclear if these cases represent true MDS or early evolving AML. Similarly there are occasional reports of *NPM1* mutant CMML (Bains et al., 2011; Caudill et al., 2006; Courville et al., 2013; Itzykson et al., 2013a), but these cases generally progress rapidly to AML (Bains et al., 2011; Caudill et al., 2006; Courville et al., 2013). Similarly, some experts question whether these cases are CMML or AML which has been detected in an early, sub-clinical phase, accompanied by marked monocytic differentiation (Falini et al., 2011).

In the acute leukaemia clinic at Addenbrooke's Hospital we were treating one such patient, who was initially diagnosed with CMML, but progressed to clinically overt AML within three months. Routine diagnostic tests performed at the hospital at the time of AML presentation detected both *NPM1* and *FLT3-ITD* mutations, but on retrospective assessment of the CMML sample only the *NPM1* mutation was identified. In order to understand the nature of the CMML to AML progression in this uncommon situation where the *NPM1* mutation was detectable prior to the onset of overt AML, I studied the paired CMML and AML diagnostic samples using deep sequencing.

3.2 Clinical Case

A 50y.o. woman presented with an eight week history of non-specific symptoms. She was anaemic (haemoglobin 7.5g/dL) and had a peripheral blood monocytosis (1.74×10^9/L) with normal neutrophil count (5.48×10^9/L) and a total white cell count (WCC) of 8.7×10^9/L. Her bone marrow was hypercellular, with myeloid to erythroid ratio of >10:1 and dysplastic changes, but <5% blasts. There were no high risk features on FISH analysis and the cytogenetic study was normal. However 83 days later she was admitted to hospital with fevers, abdominal symptoms and a WCC of

209x10^9/L (82% blasts). The diagnosis of AML was confirmed on bone marrow examination and molecular testing identified *FLT3-ITD* and *NPM1* mutations. She was treated according to the AML17 trial protocol with cytarabine, daunorubicin and etoposide (ADE) chemotherapy and initially went into complete remission with no detectable *FLT3-ITD* or *NPM1* mutation by PCR. However, following the first consolidation cycle of ADE chemotherapy she relapsed with 38% blasts on her bone marrow. She responded to salvage treatment with FLAG-IDA (fludarabine, cytarabine, idarubicin and G-CSF) followed by high dose cytarabine and was in complete remission with incomplete peripheral recovery of blood counts at the time of allogeneic transplant, 300 days after her initial presentation with CMML. Unfortunately three months after transplant she died from relapsed disease. We studied samples taken at different time points during her clinical course. The timing of the analysed samples along with her peripheral blood WCC, haemoglobin and clinical course are shown in figure 3.1.

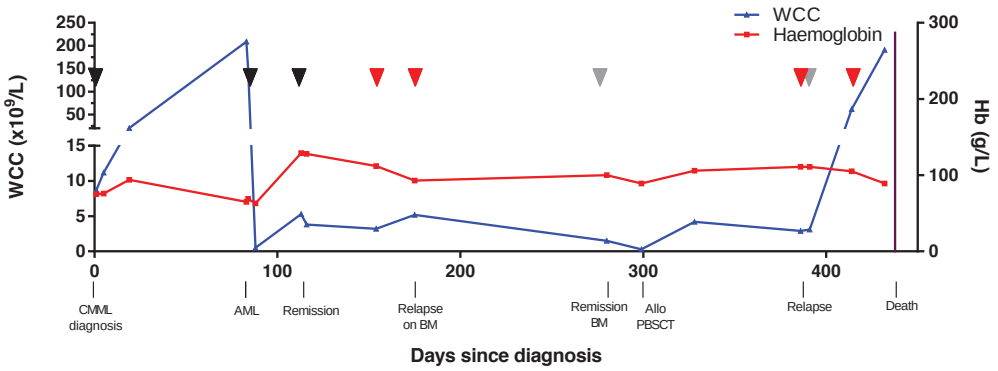


Figure 3.1: Disease timecourse. Major clinical events are noted. Black arrowheads indicate BM samples used in both exome sequencing and PCR validation. Those indicated by red and grey arrowheads were used for PCR and MiSeq only. Red = blood, grey = BM

3.3 Results

Exome sequencing was performed on whole bone marrow samples taken at diagnosis of both CMML and AML and during the first complete remission, using Agilent SureSelect Human Exon 50Mb Kit baits and the Illumina HiSeq2000 sequencing platform. This generated between 84 and 90 million, 75bp, paired end reads in all three samples. In each sample 88% of all reads were mapped to the genome covering over 98% of the targeted regions. The median coverage was 100 fold across the exome with 80% 40 fold or higher coverage in all three samples.

After comparison to the remission sample and standard filtering, ten insertions and deletions were identified on Pindel analysis in the AML samples and eight in the CMML (appendix 3A). These included a four nucleotide TCTG insertion at 5:170837547 in both samples, consistent with a type A *NPM1* mutation. The only other shared aberration called by Pindel was a complex abnormality involving 2127 bases within CDC27 on chromosome 17, of uncertain significance.

Caveman analysis for single nucleotide variants (SNV) identified 43 mutant calls in the AML sample and 62 in the CMML sample, of which 23 were common to both (table 3.1) (appendix 3B). As each read of an Illumina sequencing run derives from a single molecule of genomic DNA, the proportion of independent sequencing reads reporting a variant allele can be used to estimate the proportion of cells in a DNA sample carrying that mutation(Campbell et al., 2008). Exome sequencing of the CMML sample revealed three mutations which are likely to have a driver role in leukaemogenesis, with an allelic frequency suggesting they were present in the dominant clone; *DNMT3A*, *TET2* and *NPM1* (figure 3.2).

In addition to these three driver lesions, mutations in *PTPN11* and *SMC3*, which are both genes that are recurrently mutated in AML, were also found in the CMML sample. However, these mutations had a low variant allele frequency (VAF) in the CMML sample, suggesting they were occurring in minor sub-clones (figure 3.2). Sequencing of the AML sample demonstrated an expansion of the *SMC3* clone, but a reduction in *PTPN11* mutant reads. In fact the *PTPN11* mutation was detected in such small number in the AML sample that it did not pass filtering to be included on the output list on Caveman analysis, but mutant reads were evident on review of the raw sequencing data mpileup (Niccolo Bolli). The contrasting pattern in VAF of these two

mutations on serial sampling suggests they may not be co-occurring within a single tumour sub-clone although with this level of coverage they were still grouped together on the Dirichilet analysis.

						Allele		Depth			% Mutant in		
Gene	CHR	Position	cDNA	Protein	Type	WT	MT	Normal	CMML	AML	Normal	CMML	AML
PGLYRP4	1	153317825	c.173G>A	p.R58H	Misssense	C	T	88	100	97	0	42	45.36
DNMT3A	2	25457243	c.2644C>T	p.R882C	Misssense	G	A	90	70	73	0	42.86	47.95
ARL6	3	97503812	c.268A>G	p.I90V	Misssense	A	G	93	87	115	0	29.89	43.48
MFI2	3	196746550	c.835G>A	p.V279I	Misssense	C	T	198	210	230	0.51	32.38	40.87
TET2	4	106196940	c.5273C>G	p.S1758*	Nonsense	C	G	164	140	167	0	42.86	46.11
ACCN5	4	156775338	c.476C>T	p.A159V	Misssense	G	A	137	116	172	1.46	34.48	50
ACSL6	5	131326623	c.308G>A	p.G103D	Misssense	C	T	69	60	73	0	43.33	47.95
HLA-A	6	29911272	c.571T>G	p.W191G	Misssense	T	G	48	60	67	2.08	11.67	7.46
CLCN1	7	143029920	c.1355C>T	p.P452L	Misssense	C	T	229	188	197	0	35.11	38.58
ZMAT4	8	40532213	c.577+10G>T	p.?	Splice	C	A	101	81	107	0.99	45.68	43.93
AQP7	9	33395107	c.113T>C	p.L38P	Misssense	A	G	106	129	133	2.83	6.98	9.02
ZCCHC7	9	37356935	c.1302G>T	p.W434C	Misssense	G	T	106	98	84	1.89	26.53	50
EPB41L4B	9	112003846	Non-coding	r.2171u>c	3'UTR	A	G	35	35	53	0	45.71	45.28
SMC3	10	112360841	c.2597T>C	p.L866P	Misssense	T	C	500	500	500	0	4.2	14
ENSG00000196779	11	57876606	c.528C>G	p.F176L	Misssense	G	C	108	135	128	0	42.96	52.34
STRCP1	15	43892822	c.4903G>T	p.V1635F	Misssense	C	A	72	118	78	2.78	8.47	8.97
KRT14	17	39742894	c.193C>T	p.L65L	Silent	G	A	16	28	27	0	28.57	33.33
KRT14	17	39742898	c.189C>T	p.C63C	Silent	G	A	15	26	27	0	26.92	37.04
DIRAS1	19	2717179	Non-coding	r.784a>c	3'UTR	T	G	54	70	41	3.7	10	21.95
PSG8	19	43258444	Non-coding	r.1284g>a	3' UTR	C	T	391	378	361	0.26	36.51	40.17
ZNF41	X	47308797	c.372C>T	p.P124P	Silent	G	A	72	72	80	0	30.56	40
ZXDB	X	57618870	c.389G>A	p.G130D	Misssense	G	A	17	25	36	0	20	19.44
THOC2	X	122754807	c.4226G>A	p.R1409H	Misssense	C	T	399	374	440	0	36.9	42.5

Table 3.1: SNV shared by both the CMML and AML samples as detected by Caveman Analysis. CHR = Chromosome, WT = wildtype and MT mutant allele

There were several recurrent leukaemia associated mutations that were identified in the AML sample and were not detected at the CMML stage on standard Caveman and Pindel analysis. The dinucleotide insertion in *CEBPA* is one such example. Although this was evident in occasional reads in the CMML sample when the data was reviewed in mpileup, this mutation represented a much higher proportion of reads in the AML sample. It is possible the acquisition of a bi-allelic *CEBPA* mutation within a sub-clone contributed to this increased VAF, however the degree of change indicates there was expansion of this *CEBPA* containing clone. Analysis of the AML sample also revealed an additional *NRAS* mutation, which was not detected at the CMML stage. The allelic frequency indicates this mutation was present within a small AML sub-clone.

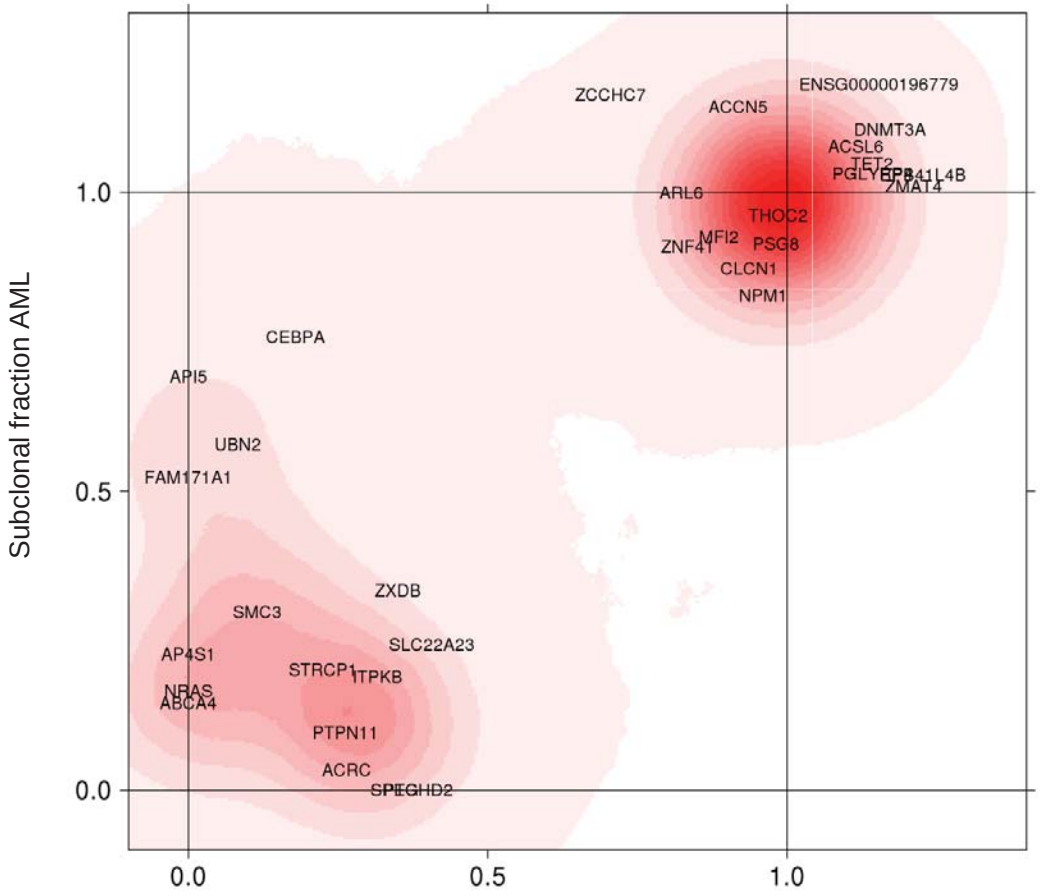


FIGURE 3.2: Two dimensional density plot showing the fraction of tumour cells carrying the mutations indicated, and their clustering. Increasing intensity of red indicates high posterior probability of a cluster. The *FLT3-ITD* mutation is not represented as it was not detected by Pindel or Caveman analysis. Manual annotation (Niccolo Bolli) was used to determine the subclonal fractions for *CEBPA* and *NPM1*.

Interestingly, the *FLT3-ITD* mutation which was picked up in the AML sample by PCR and agarose gel analysis in the diagnostic laboratory (figure 3.3) was not detected using exome sequencing and Pindel analysis. This is a recurrent problem with this specific type of mutation (Dr Eli Papaemmanuil, personal communication). Sanger sequencing was performed on DNA from the mutant band and identified the *FLT3-ITD* sequence as an 81bp duplication of 13:28608238-28608319. However, re-

analysis of the exome data performed by Dr Eli Papaemmanuil, specifically screening for this *FLT3-ITD* sequence, failed to detect mutant reads.

Copy number analysis (ASCAT) (Van Loo et al., 2010) was also performed based on the exome sequencing data. This showed sub-clonal copy-neutral loss of heterozygosity (LOH) in chromosome 1p and 13 in the AML sample, which would be consistent with acquired uniparental disomy in the *NRAS* and *FLT3* loci (figure 3.4).

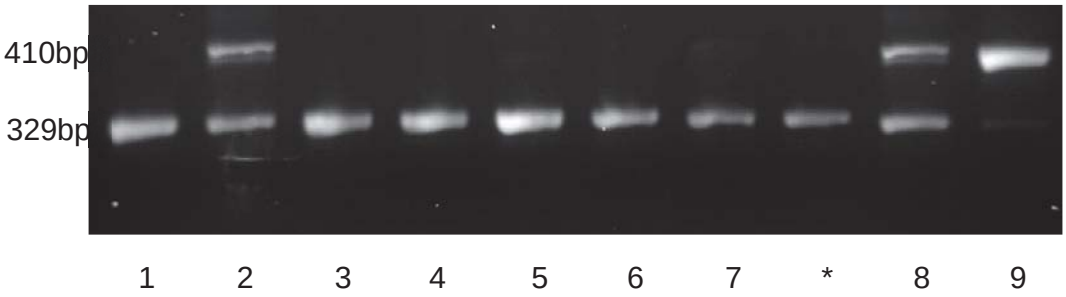


FIGURE 3.3: PCR and gel electrophoresis for the FLT3-ITD mutation performed at the diagnostic laboratory (Anthony Bench). Samples 1-9 correspond to the samples used for MiSeq analysis.

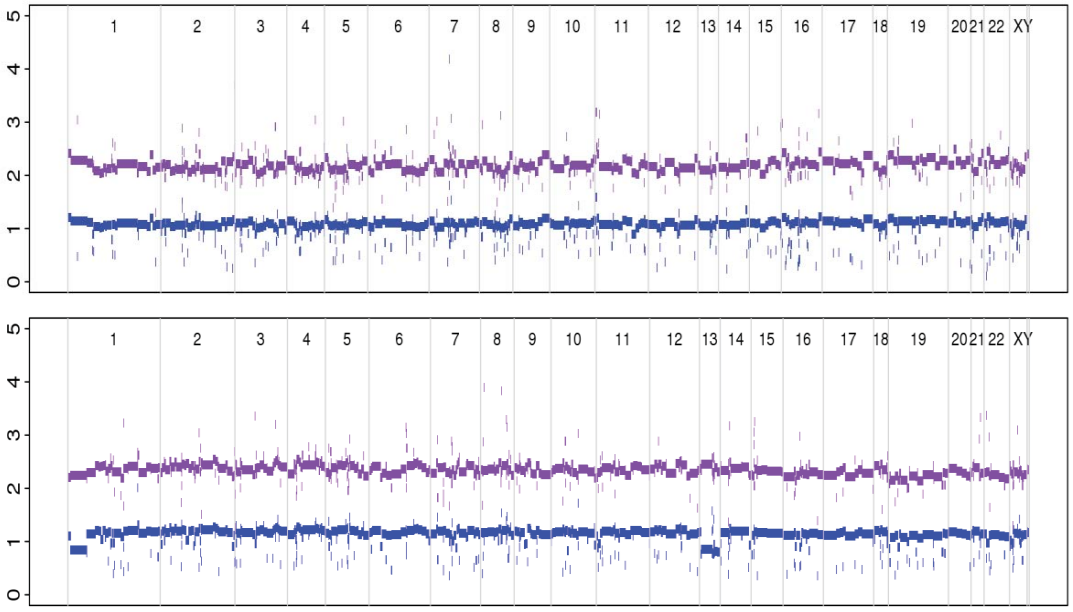


FIGURE 3.4: Copy number changes between CMML (top) and AML (bottom) The CMML had normal karyotype but there was sub-clonal copy neutral LOH in chromosome 1p and 13 in the AML sample.

The eight presumed driver mutations that have been described already (figure 3.2) along with thirteen probable passenger mutations that clustered with them were validated using read counting of Illumina sequencing (MiSeq) of non-allele specific PCR products. This analysis was performed on nine blood and bone marrow samples that were collected through the clinical course (figure 3.1) including the remission and diagnostic samples used for exome sequencing, and a control DNA from a person with no diagnosis of haematological malignancy. After excluding samples with <1000 reads for a given gene, the remaining gene and sample combinations gave an average of 65 000 fold coverage at the mutant loci (median 51023). As the *ZXDB* locus amplified poorly and failed in the majority of samples, it was excluded from further analysis (figure 3.5).

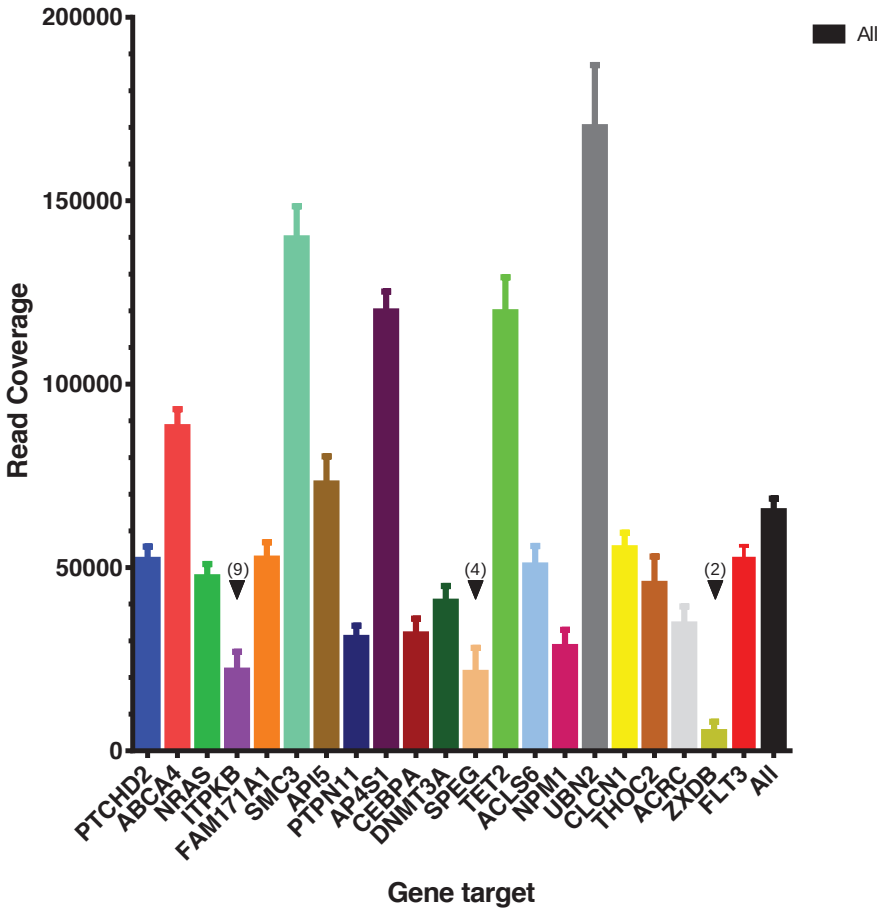


Figure 3.5: Coverage for the 10 samples across 21 target genes. The mean and standard error of the mean are shown for each site. Where a sample gave <1000 reads at a site this was excluded from the analysis. When less than10 samples were included the number of samples analysed are given in parenthesis.

Using these deep sequencing results we were able to track the allelic frequency through the serial samples. On Dirichlet analysis twelve clusters were identified of which six contained only a single mutation (figure 3.6). The findings for *CEBPA* in the CMML sample differed to the exome sequencing where only occasional *CEBPA* mutant reads were detected. Notably, the VAF of the *CEBPA* mutation rose above 50% in the last relapse sample. The VAF agreed closely with the exome data for all three samples studied and for all genes except *CEBPA* thus confirming the quantitative nature of the PCR method.

There were five mutations, *TET2*, *THOC2*, *DNMT3A*, *NPM1* and *CEBPA*, which had allelic frequencies of over 30% in both the CMML and AML samples and rose to a similar level in the last relapse sample, although these were clustered into five separate groups. The *TET2*, *THOC2* and *DNMT3A* mutations were detectable above baseline at low level in the day 175 peripheral blood sample; the time of relapse 1 (sample 5), whereas the *CEBPA* and *NPM1* mutations were not (table 3.2). *SMC3* and *UBN2*, which were grouped together on Dirichilet analysis, were also increased at relapse 1 but unlike *TET2*, *DNMT3A* and *THOC2* these were not detected in the later relapse samples. Both of these mutations were detected at low level in the CMML sample and were present in around 30% of cells in the initial AML sample (sample 2). The allelic frequencies of *TET2*, *DNMT3A, THOC2*, *SMC3* and *UBN2* in the relapse 1 sample (sample 5) were low (around 1.5% in most) and comparable to the level seen in the initial remission sample used for exome sequencing (sample 3) for *TET2* and *DNMT3A*. However, this represented a 3-12 fold increase and an absolute rise in mutant read number of several hundred compared to the mean coverage in the deep remission samples (sample 4 and 6) and control sample for each of these genes. Such changes were not seen in sample 5 for *CEBPA*, *NPM1*, *NRAS* or *FLT3*.

	TET2		THOC2		DNMT3A		SMC3		UBN2		CEBPA		NPM1		NRAS		FLT3	
Sample	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)
1	64014	40.35	12352	40.93	27901	44.37	4148	4.11	8013	4.36	8466	34.16	17210	43.11	848	1.55	12	0.07
2	53654	46.58	21338	43.98	16895	50.57	22440	14.39	28614	13.77	10313	35.50	5635	44.42	5325	12.13	5400	23.81
3	1278	1.19	329	0.31	684	2.11	290	0.27	366	0.17	2	0.01	0	0.00	73	0.20	10	0.03
4	611	0.69	105	0.33	53	0.24	298	0.24	224	0.19	2	0.01	72	0.31	113	0.20	7	0.04
5	2717	1.79	2163	4.63	987	1.79	2351	1.51	2440	1.49	16	0.03	0	0.00	69	0.16	3	0.01
6	250	0.28	127	0.51	65	0.25	265	0.20	285	0.11	3	0.02	-	0.00	51	0.16	6	0.03
7	5041	4.06	4715	8.01	1817	4.69	284	0.21	95	0.07	1819	6.31	1045	3.78	63	0.19	1008	4.40
8	32897	19.93	10229	26.06	8710	21.46	410	0.24	88	0.05	10062	24.50	3126	16.79	86	0.15	8148	24.25
9	53843	47.10	20911	47.02	17428	45.31	401	0.21	217	0.14	20764	56.57	9832	46.09	111	0.23	16832	86.75
Control	473	0.57	397	1.58	123	0.22	256	0.20	52	0.06	11	0.02	0	0.00	123	0.19	10	0.03

Table 3.2: The absolute number and proportion of reads assigned to the variant allele for nine of the mutations. The first relapse sample (sample 5) is shaded.

Although *NPM1* and *CEBPA* were clustered separately on the Dirichilet analysis, they were both detected in the CMML, AML and late relapse samples but not in the initial relapse (sample 5). Cluster 9, which contained *CLCN1* and *ACLS6* had a similar pattern of occurrence to *NPM1* and *CEBPA*, but were present at roughly half the allelic frequency. *FLT3-ITD* was also clustered separately on Dirichilet analysis. It had an allelic frequency of 24% in the AML diagnosis sample, was not detected in the initial relapse and had an allelic frequency of 87% in the final sample, which suggests the majority of cells had bi-allelic mutations at that time. *NRAS* clustered with *ABCA4* and *AP4S1* but is in a separate clone to the other driver mutations. It was present in the initial AML sample, but not in either of the relapses.

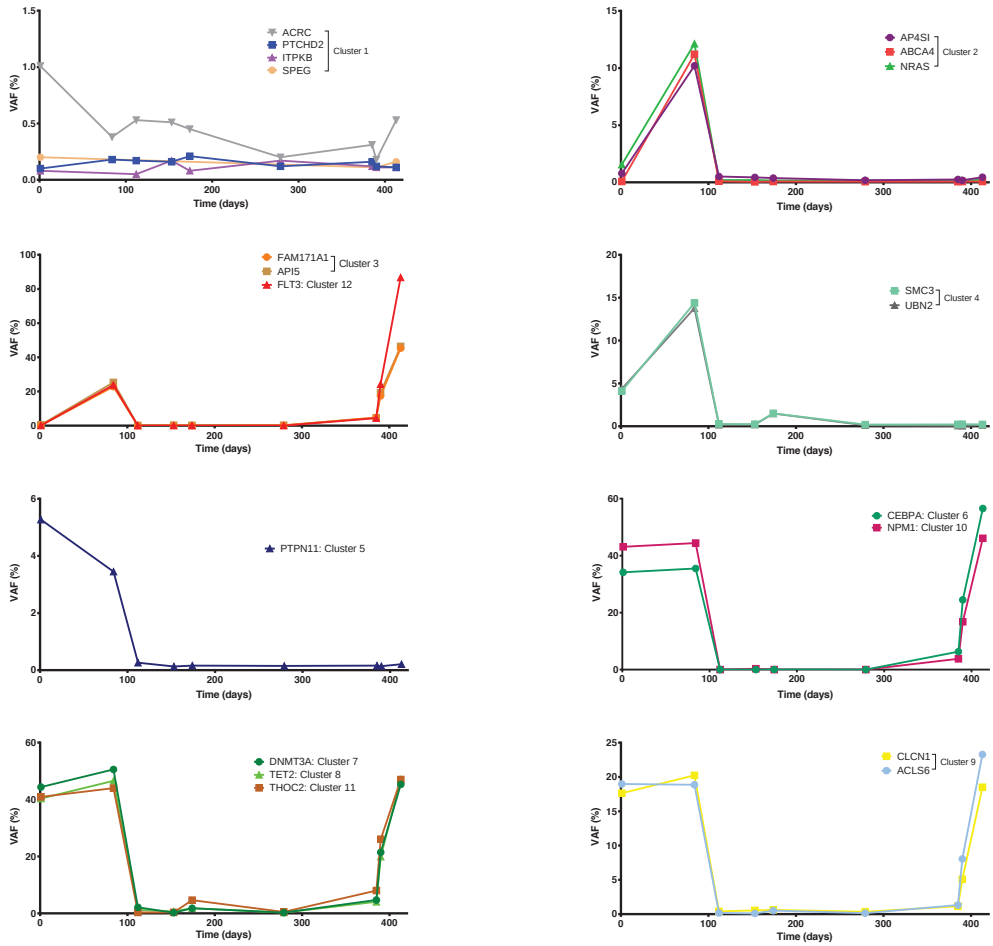


Figure 3.6: Variant allele frequencies (VAF) for the various mutations over time. The cluster to which the mutation was assigned on the Dirichilet analysis is also shown.

3.4 Discussion

This case is striking for the sheer number of AML associated mutations which were identified in the CMML genome. The detection of a combination of *TET2*, *DNMT3A* and *NPM1* mutations in the predominant clone at the time of CMML was surprising as one might have expected these three powerful mutations to cause full blown AML. It has been suggested that as few as two driver mutations may be sufficient to generate leukaemia (Welch et al., 2012). However, it also appears that rather than occurring due to the simple expansion of this CMML clone, the clinical progression to AML was driven by the acquisition of new mutations. The *FLT3-ITD* containing clone, or a sub-clone that evolved from it, also led to the fatal relapse following bone marrow transplantation. The evolution of pre-leukaemia clones to AML is thought to be a stochastic process. However, in this case with multiple pre-existing mutations the independent acquisition of *FLT3-ITD* and *NRAS* in separate clones in such a short time suggests that the evolution to AML was almost inevitable, akin to a deterministic process.

The combination of whole exome sequencing to identify disease specific mutations, with a targeted gene re-sequencing approach across multiple serial samples has proven effective in deciphering the branching clonal evolution of this individual leukaemia. In AML disease relapse typically arises from a pre-existing clone (Ding et al., 2012). In this study we used a targeted sequencing approach on the relapse samples and therefore we could not identify new mutations specific to the relapse. However, the pattern of allele frequencies for the various mutations during the disease course provided evidence for which mutations were co-occurring within a single disease clone and the order of acquisition of mutations (figure 3.7).

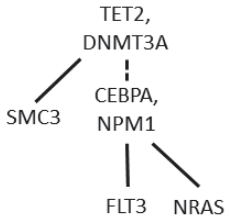


Figure 3.7: Phylogram showing the order of acquisition of the driver mutations. *TET2*, *DNMT3A*, *CEBPA* and *NPM1* are all in the major clone. The distinction between *TET2/DNMT3A* and *CEBPA/NPM1* is based on their presence in the early relapse. The differing patterns of recurrence of *SMC3*, *FLT3* and *NRAS* indicate these were likely to be in separate sub-clones.

Non-allele specific PCR and deep sequencing is a powerful approach for analysis of specific mutations at high read depth. Although whole genome and whole exome sequencing approaches have provided unique insights into the heterogeneity of mutations and sub-clonal architecture of diseases such as AML, these approaches are limited in their ability to detect mutations occurring in minor sub-clones with VAFs of less than 10%(TCGA_Research_Network, 2013). Although the proportion of mutant reads detected in the initial relapse sample from this patient were in the order of 1.5-4.6%, this represented a clear increase above baseline at these sites (table 3.2).

The pattern of mutations detected in the first relapse sample suggests that the *TET2* and *DNMT3A* mutations were acquired before the *NPM1* and *CEBPA* mutations in clonal evolution. This concurs with recent evidence that the former are early events (Busque et al., 2012; Jan et al., 2012; Shlush et al., 2014). Nevertheless, the mutant VAFs were closely concordant at CMML and AML diagnosis and in the second relapse. The first relapse sample was a peripheral blood sample and the detected mutations were at low VAF despite the fact there were 38% blasts reported on a bone marrow taken at the time. Unfortunately we were unable to obtain any DNA from the marrow sample. One potential criticism of the targeted re-sequencing approach is that PCR may introduce bias to the VAF and it is possible the *NPM1* and *CEBPA* mutations were only absent due to reduced sensitivity for detection compared to the other mutations, hence the dotted line in figure 3.7. PCR bias is unlikely to be a problem for SNVs, and the mutated base was positioned in the middle of the PCR product where possible. Although PCR bias is potentially more of an issue with insertions and deletions both the *NPM1* mutation (a tetra-nucleotide repeat) and the *CEBPA* insertion (two base pairs) are very short. It is unlikely that such small changes introduce a significant bias to the PCR reaction and with such deep coverage we would expect to find some evidence of the mutation at the initial relapse. There were over 50 000 MiSeq reads for each of these genes at this time-point, so for a heterozygous variant in 3% of cells there should be 750 mutant reads. Even if the PCR favoured the wild-type allele 10:1, 75 mutant reads would be detectable with this depth of read coverage. Furthermore, both mutations were clearly evident in the seventh sample, with VAF of 3.7% and 6.3% respectively. These results are in line with the VAF of the *TET2* (4.1%) and *DNMT3A* (4.6%)

mutations in this blood sample, when all of these mutations are present in the relapsing clone.

The data suggests that the two AML relapses occurred from different sub-clones, the first of which contained *SMC3* but not the second. The absence of *SMC3* in relapse 2 indicates this mutation was in a branching sub-clone from the one that progressed causing fatal disease, as was the *NRAS* mutation seen at diagnosis (sample 2). The final relapse sample contained *TET2*, *THOC2*, *DNMT3A*, *NPM1*, *CEBPA*, *FAM171A1*, *API5* and *FLT3-ITD* mutations at a VAF that implies they were present in close to 100% of nucleated cells in the peripheral blood.

Several of the mutations that were targeted for re-sequencing were thought to be passenger mutations. These were selected as they appeared to track with driver mutations based on their VAF in the exome sequencing. Including these mutations improved the confidence with which we could track sub-clones, nevertheless some of these, such as THOC2 were grouped separately in the MiSeq analysis. Most of the mutations with a sub-clonal fraction of close to one on the exome sequencing were clustered separately in the Dirichilet analysis of the MiSeq data. This is likely to be a consequence of the deeper read coverage and does not imply these are in separate clones.

The *FLT3-ITD* was not detected on exome sequencing, but on PCR re-sequencing had an allelic frequency of 24% in the AML diagnosis sample. The copy number analysis on the AML exome sample also suggests LOH at the *FLT3* locus. The *FLT3-ITD* sequence could not be detected in the exome sequencing output despite specifically searching for this and other *FLT3-ITD* mutations in the raw data. A possible explanation is that the 81bp duplication affected the 'pull down' of the mutant allele during hybridisation to the RNA baits, therefore selecting against the mutant allele in the exome sequencing protocol. In the MiSeq analysis it is also possible that the PCR performed ahead of re-sequencing introduced an amplification bias. However, given the larger size of the mutant PCR product any such bias is more likely to favour the wild-type allele and the mutant allele frequency of 87% in the final relapse sample suggests that this mutation is not strongly selected against in the MiSeq library preparation. Furthermore, the MiSeq data correlates well with the PCR gel electrophoresis which was performed by the diagnostic laboratory.

DNMT3A and *NPM1* mutations are among the commonest mutations in de novo AML and frequently co-occur with each other and with *FLT3* mutations within the same tumour(TCGA_Research_Network, 2013). In fact this combination of mutations was reported in the first published case of AML analysed by whole genome sequencing(Ley et al., 2010; Ley et al., 2008). The *NPM1* mutation identified in this case is the commonest, type A mutation which affects the critical 288 and 290 tryptophan residues disrupting the nucleolar localisation signal (Falini et al., 2005) as well as introducing a new nuclear export signal(2005; Falini et al., 2006). *DNMT3A* encodes one of a group of DNA methyltransferases, which catalyse the addition of a methyl group to cytosine residues of CpG dinucleotides. Increased methylation of CpG islands is typically associated with reduced expression of downstream genes(Ley et al., 2010). $DNMT3A^{R882C}$ is a missense mutation commonly found in AML and previously described in CMML-derived AML(Jankowska et al., 2011; Ley et al., 2010).

TET2 mutations also have an effect on epigenetic regulation. *TET2* catalyses the conversion of 5-methylcytosine to 5-hydroxymethylcytosine (5hmC) and therefore plays are role in DNA demethylation (Figueroa et al., 2010; Ko et al., 2010). *TET2* mutant samples display low levels of 5hmC (Ko et al., 2010). *TET2* mutations are one of the commonest lesions in CMML (Jankowska et al., 2011; Meggendorfer et al., 2012), are reported in 7-23% of de novo AML cases and seem to be more prevalent in elderly patients (Gaidzik et al., 2012; Metzeler et al., 2011; TCGA_Research_Network, 2013). The particular nonsense *TET2* mutation found in this case was in exon 11, and to our knowledge it has not been previously reported. However, exon 11 encodes the alpha ketoglutarate binding domain and is one of the most frequently mutated *TET2* exons in AML(Gaidzik et al., 2012). As opposed to *TET2* and *IDH1/2* mutations which are mutually exclusive in AML(Figueroa et al., 2010), *TET2* and DNMT3A are known to co-occur(TCGA_Research_Network, 2013). In fact the combination of *NPM1*, *DNMT3A*, *TET2* and *FLT3* mutations was reported in one case in the recent study of 200 AML patients using exome and whole genome sequencing(TCGA_Research_Network, 2013).

Only about 2-4% of *NPM1* mutated AML cases also carry a *CEBPA* mutation and over 90% of these are single *CEBPA* mutations (Dufour et al., 2010; Green et al., 2010; Taskesen et al., 2011; TCGA_Research_Network, 2013). In contrast across all non-

acute promyelocytic leukaemia (APL) cases of AML around 7% of patients are found to have *CEBPA* mutations of which over half are double mutations (Dufour et al., 2010; Green et al., 2010). It appears that the good prognostic impact of *CEBPA* mutations are limited to this double mutant group, which has distinct molecular characteristics (Green et al., 2010; Taskesen et al., 2011; Wouters et al., 2009). Surprisingly, in this patient with a *NPM1* mutation the VAF of the CEBPA mutation is 57% in the final relapse sample. This suggests either our patient has acquired a bi-allelic *CEBPA* mutation or lost her wildtype allele or that there is a bias for this variant in the MiSeq PCR library preparation.

The mutations in *NRAS* and *SMC3* found in this patient are probable driver lesions, even though they were not found in the major disease clone at AML diagnosis or in the final relapse. The $NRAS^{G12D}$ mutation is frequently described in human cancers including AML and has been shown to co-operate with other mutations in mice to induce AML(Li et al., 2011; Ward et al., 2012). Cohesin complex genes, including *SMC3* are mutated in 6-13% of cytogenetically normal AML(Kon et al., 2013; TCGA_Research_Network, 2013) and in 10% of cases of CMML. In AML they frequently co-exist with *NPM1* mutations (Ding et al., 2012; TCGA_Research_Network, 2013; Thol et al., 2013) and across all myeloid malignancies they are commonly found in association with mutations in *TET2*, *ASXL1* and *EZH2* (Kon et al., 2013). The cohesin complex mutations result in reduced amounts of chromatin bound cohesin components and are thought to have global effects on gene expression (Kon et al., 2013). The missense mutation of $SMC3^{L866P}$ found in this case is not reported in the COSMIC database. However, to date no particular mutation hotspot has been identified in *SMC3* in AML or other myeloid malignancies and the majority of the reported mutations are of the missense type as found in this case (Cosmic Database).

Although mutations in *PTPN11* which encodes the protein tyrosine phosphatase SHP-2 are widely reported in AML and other myeloid malignancies (Loh et al., 2004; Tartaglia et al., 2003) the 'driver' credentials of the particular mutation found in this case are less clear. The $PTPN11^{N308D}$ mutation was detected in the CMML sample with a VAF around 5%, had a lower allelic ratio in the AML sample and was not detectable above baseline in the remaining samples. The missense mutation $PTPN11^{N308D}$ is described as a germ-line mutation in Noonan Syndrome, but has not

been associated with AML(Tartaglia et al., 2003). Phosphatase assays using wild type and mutant SHP-2 proteins have shown the phosphatase activity of the *N308D* mutant is less than JMML associated SHP-2 mutants but greater than the wild-type protein. In *in vitro* assays, proliferation in cells transiently expressing the *N308D* mutation was also intermediate between normal cells and those with a JMML associated mutation. The prevalence of *PTPN11* mutations is lower in adult than in childhood AML (Tartaglia and Gelb, 2005) and whether this particular somatic mutation is a true driver in the context of adult AML is uncertain.

Regardless of the veracity of the *PTPN11* mutation as a driver, this case clearly demonstrates the molecular complexity of AML and shows a branching clonal evolution. Although the dominant clone at the time of CMML diagnosis already contained co-occurring *TET2*, *DNMT3A* and *NPM1* mutations, progression to frank AML was associated with the clear acquisition of new driver mutations. Three distinct sub-clones carrying unique driver mutations were detected in the AML sample. The initial relapse appears to have developed from a clone carrying *SMC3* along with the *TET2* and *DNMT3A* mutations. The ultimate relapse leading to death was from the re-emergence of a clone containing at least five driver mutations; *FLT3-ITD*, *CEBPA*, *NPM1*, *TET2* and *DNMT3A* but not *SMC3* or *NRAS*. This case highlights that clonal evolution is a dynamic process and forces us to question if the progression to AML is inevitable in the setting of a CMML clone with a high mutational burden.

4. *Sleeping Beauty* driven leukaemogenesis follows a rapid Darwinian-like evolution in a mouse model of Npm1c+ acute myeloid leukaemia

4.1 Introduction

Leukaemia, like other cancers, arises through the sequential acquisition of 'fitness' conferring mutations within a single cell. This clonal evolution framework underlies the molecular heterogeneity evident within the whole tumour DNA. Current knowledge regarding the order of acquisition of co-operating mutations during leukaemogenesis is inferred from; i) the variant allele frequencies (VAF) of mutations within the mass tumour, ii) observation of the pattern of mutations in single tumour cells or residual haematopoietic stem cells (HSC) (Jan et al., 2012) iii) knowledge of the biological effects of recurrent mutations, iv) the pattern of mutation co-occurrence across different tumours(TCGA_Research_Network, 2013) and v) studies of the mutational profile of serial samples either in relapsed disease or in secondary AML(Ding et al., 2012; Walter et al., 2012). However, this order of acquisition has not been monitored in real time in *de novo* AML as the presentation is acute and the disease is rare, unpredictable and arises without a prodrome.

Transposon insertional mutagenesis is a valuable technique with which to study tumorigenic mutations in mouse models. To date the predominant application has been for cancer gene discovery, analogous to retroviral mutagenesis. Putative tumour drivers are identified by ascertaining genes and regions in which the transposon or retrovirus integrates more frequently than is expected by chance alone; the common integration sites (CIS). This approach has proven effective for both solid and haematopoietic malignancies, but the *in vivo* kinetics of transposon integration is poorly understood. Unlike retroviruses, transposon mobilization continues throughout the life of a host cell and critical oncogenic events may potentially occur after a significant latency. Typically multiple copies of the

transposon cassette are supplied in a concatamer and the rate of new integrations (i.e. transposition) relative to the rate of cell division is largely unknown. The timing and order of acquisition of oncogenic integrations in mouse insertional mutagenesis models has not been studied to date.

Heterozygous somatic mutations in the terminal exon of *NPM1*, the gene for Nucleophosmin, are found in up to 35% of cases of human AML(Falini et al., 2005). *Sleeping Beauty* (*SB*) was used to identify genes that collaborate with *Npm1* in an insertional mutagenesis (IM) mouse model of AML developed by our lab (Vassiliou et al., 2011). In brief, the conditional $Npm1^{flox\text{-}cA}$ allele was designed to minimise interference with the native locus, but to switch to $Npm1^{cA}$ after Cre-loxP recombination (figure 4.1A). Approximately one third of $Npm1^{cA}$ mutant mice developed myeloid leukaemia but only after a protracted latency suggesting the need for co-operating mutations. A conditional Rosa26 *SB* transposase allele (figure 4.1B) was used to mobilise *GrOnc*, a bi-functional *PB*/*SB* transposon capable of both gene activation and disruption, in $NPM1^{cA}$ mice (figure 4.1c). In this model the $Npm1^{cA}$ mutation and the *SB* transposase, activated by the haemopoietic *Mx1Cre* (Kuhn et al., 1995), caused rapid onset AML in 80% of mice(Vassiliou et al., 2011). CIS were identified at known and novel cancer genes including insertions near *Csf2*, *Flt3*, *Rasgrp1*, *Kras*, *Bach2*, *Nf1* and *Nup98*(Vassiliou et al., 2011). Some of these recurrent integrations were largely mutually exclusive, suggesting their effects in leukaemia pathogenesis are redundant.

The $Npm1^{cA}$ IM model provides a useful platform in which to investigate the *in vivo* behaviour of transposons and the clonal evolution of AML. Blood can be sampled sequentially throughout the life of the mice, abnormalities in blood parameters can be monitored quantitatively and tumour cells from effected mice can be serially transplanted into recipients. Determining when the mutations first appear during tumour evolution and if they persist on transplantation, should improve the confidence for distinguishing driver and passenger mutations, define the order of mutation acquisition and enhance our understanding of how transposons operate as well as giving insights into the clonal evolution of AML.

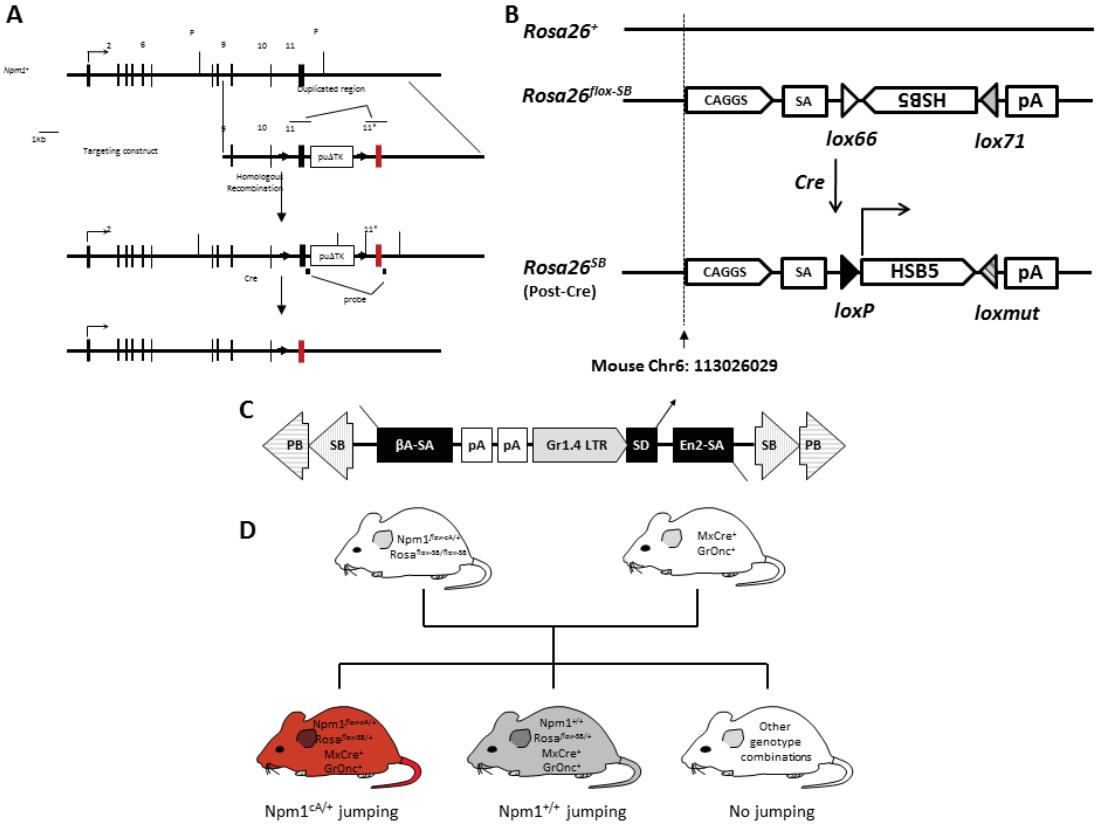


Figure 4.1: Generation of *Npm1^{cA}* insertional mutagenesis mice. A: Targeting construct for *Npm1^{cA}*. B: Conditional Rosa26 *SB* transposase. Upon Cre activation the SB cDNA flips to the sense orientation and cannot take part in any further Cre mediated recombination. C: The GrOnc transposon flanked by PB and SB repeats and gene activating and inactivating elements. Gr1.4LTR = Graffi 1.4 MuLV long terminal repeat, SD = splice donor, SA = splice acceptor. D: Mating scheme to generate insertional mutagenesis mice. Pictures courtesy of George Vassiliou (GV)(Vassiliou et al., 2011).

4.2 Results

4.2.1 *Npm1^{cA}* mutant mice with a low copy number Sleeping Beauty transposon develop myeloid leukaemias

Mice created by GV, with a humanised conditional knock-in of *NPM1^{cA}*, *SB* transposase and *GrOnc* were used in this study (figure 4.1). The model is closely related to the one used in the published work, and differs only in the transposon copy number and donor locus. The mice described here have only 15 copies of the *GrOnc* transposon resident at a donor locus within the centromere of chromosome 16 (GRL) (figure 4.2). This lower transposon copy number was selected to try to prolong the latency until tumour development, whilst the centromeric location has the potential advantage of being distant from cancer genes.

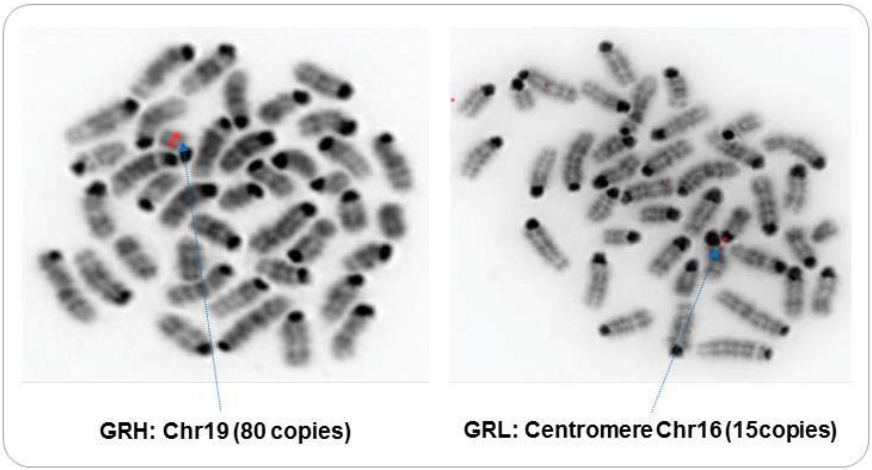


Figure 4.2: FISH analysis of the GrOnc constructs. On the left the published high copy number construct (GRH) and on the right the low copy number construct used for this work (GRL). FISH results were supplied by Ruby Banerjee and GV.

The mice were mated using the same scheme as the published model (figure 4.1D) to generate 71 quadruple transgenic mice (*Npm1*$^{floxcA/+}$, *Mx1Cre*, *Rosa*$^{floxSB/+}$, *GRL*). Of these 52 received a full course of four to six polyinosinic-polycytidylic acid (pIpC) injections at 8-12 weeks of age to activate the *Npm1*$^{flox\text{-}cA/+}$ and *Rosa*$^{floxSB/+}$ conditional alleles. These mice had a shortened lifespan compared with *Npm1*^{WT} non IM (hereafter called WT) mice (median survival 215 v 597 days $p<0.0001$) (figure 4.3). Sixteen mice with all of the mutant alleles did not receive pIpC injections and the survival of these mice was not significantly different to the WT cohort with median survival of 483 days ($p=0.07$). Two mice received an incomplete course of pIpC (only 2 injections) and these were excluded from survival and phenotyping analysis. The low copy transposon cohort developed myeloid leukaemias with similar prevalence to the published cohort, but with a longer median survival of 215 compared to 99 days (figure 4.3 and 4.4).

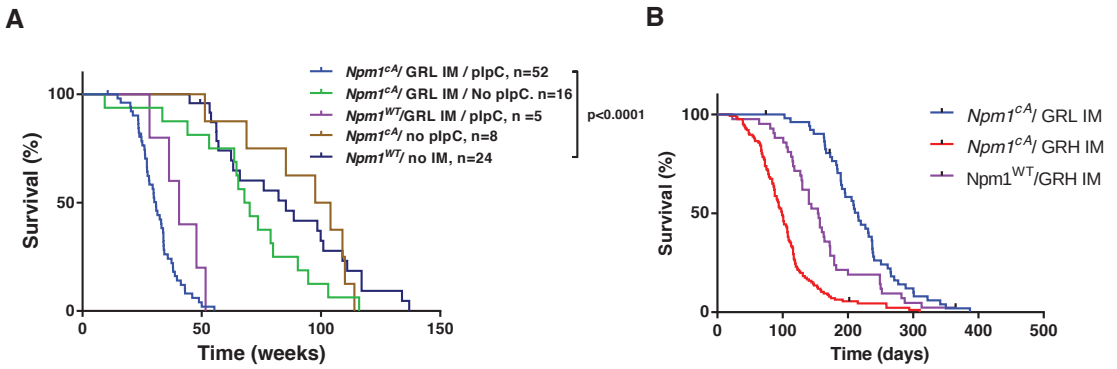


Figure 4.3: Survival Curves. A) Survival in the GRL cohorts. Mice that received an incomplete course of pIpC injections are not included in the analysis. **B)** Survival compared to the GRH IM cohorts.

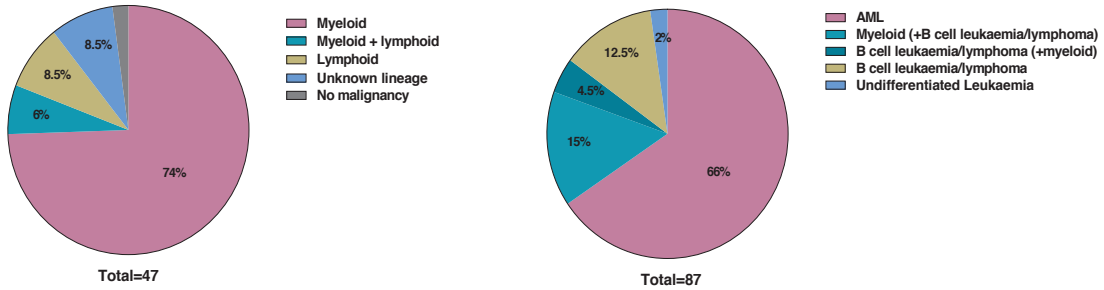


Figure 4.4: Disease phenotype in the *Npm1*^{cA} IM cohorts. On the left the GRL cohort and on the right the published GRH colony.

The *Npm1*^{cA}/IM cohort had a higher WCC and mean cell volume (MCV) at death compared to wild type (non-pIpC treated and WT) mice (mean WCC 227 ± 34 v 76 ± 23, p=0.0005, MCV 70.0 ± 2.1 v 57.8 ± 1.5, p<0.0001) and lower platelet count (561 ± 138 v 1034 ± 109 p=0.0098) but there was no significant difference in haemoglobin (p=0.9271) (figure 4.5). The mice with myeloid leukaemia had variable proportions of blasts. In some the morphology was more akin to a myeloproliferative neoplasm (MPN) or CMML (figure 4.6a and b), whereas many had acute leukaemia with a very high percentage of blasts (figure 4.6c).

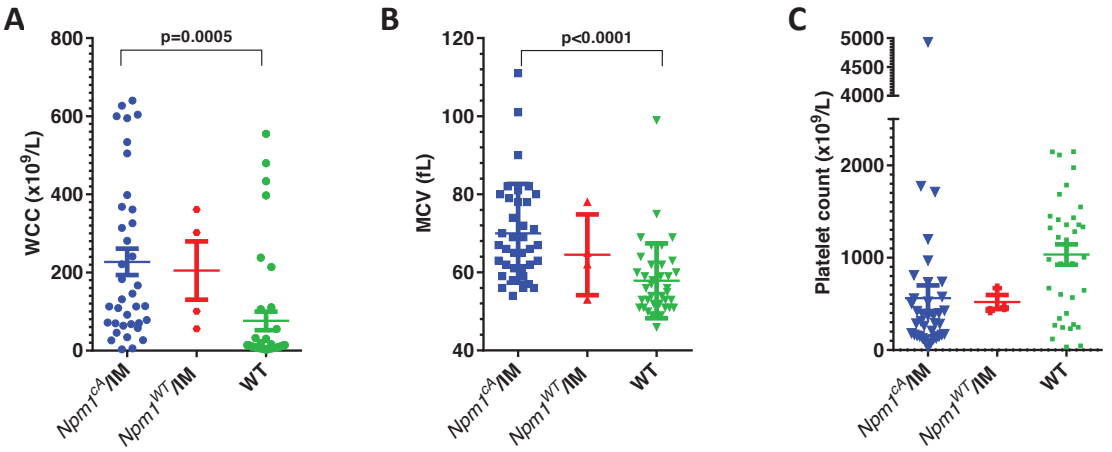


Figure 4.5: FBC parameters in the GRL IM cohort at death: **A)** White cell count **B)** Mean cell volume and **C)** Platelet count. The WT cohort includes mice that did not receive any pIpC injections, in addition to those with a WT genotype.

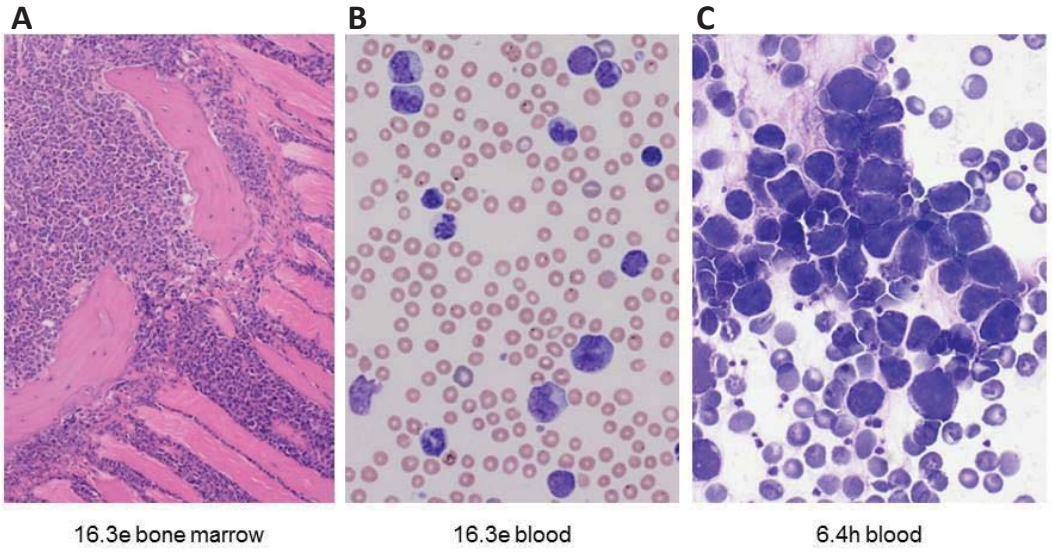


Figure 4.6: Spectrum of morphology in the *Npm1*^{cA} GRL IM cohort. Although the bone marrow is packed with myeloid cells, many of the mice have blasts and maturing cells on the peripheral blood smear as in 16.3e (**A** and **B**). Some have a high percentage of frank blasts, as shown here in the tail of the blood film from 6.4h (**C**). Picture A provided by G Hoffman.

4.2.2 GRL verifies CISs identified by GRH and identifies additional ones

Common integrations sites were identified using the CIMPL program and the parameters described (Materials and Methods 2.4.2). There were 27 CIS genes identified by all three methods (figure 4.7 and table 4.1). These CIS showed significant overlap with the published model (figure 4.8). However, several additional CIS regions were identified, indicating that the utility of the *SB* IM approach to identify genes co-operating with *Npm1*^{cA} in leukaemogenesis was not exhausted in the initial study. These additional CIS included some at the sites of well-established leukaemia associated genes such as *Mll1* and *Phf6*. The additional nine CIS identified by only one or two methods are shown in the appendix 4A.

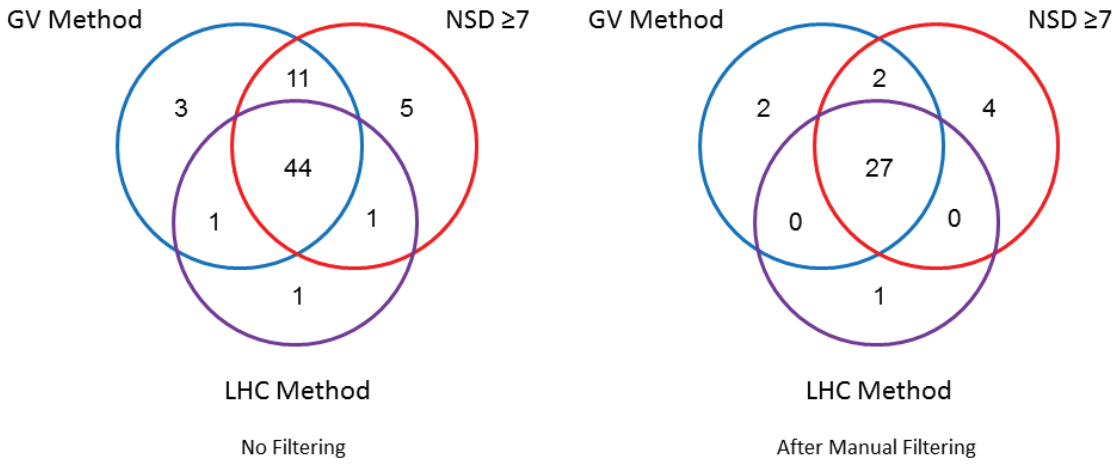


Figure 4.7: CIS identified by the three CIMPL methods. The total number of sites identified by each method before manual filtering are shown on the left, and after manual filtering on the right. GV method = published GRH model, NSD≥7 as per GRH but with higher NSD value, LHC method with built in local hopping correction.

Gene nearest to CIS peak	Kernel sizes at which CIS was identified (x1000)	Chromosome	peak location*	peak height*	start*	end*	CIS width*	Number of tumours with hits*	P value*	Genes in CIS*	Orientation +	Orientation -	Notes	Kernel Size*
Uggt1	30, 60	1	36186907	6.379633631	36151416	36204652	53237	9	3.327E-05	Hs6st1 Uggt1	5	0	Unknown target, most hits 3' of both genes in CIS	60000
A330023F24Rik	30, 60, 100	1	196863803	10.01382946	196774966	196932898	157933	17	5.51E-06	Cd34 Gm16897 A330023F24Rik Mir29b-2 Mir29c Cd46 Cr1l	7	2	Unknown target, all integrations not localised within or 5' to any annotated gene	100000
Bmi1	10, 30	2	18601379	4.313937047	18594498	18605311	10814	8	6.564E-06	Commd3 **Bmi1**	3	0	Bmi1 intron 1 forward	10000
Sec16a	60, 100	2	26295695	5.965099531	26272063	26307510	35448	8	5.414E-05	Sec16a 0610009E02Rik	1	0	unknown, ?linc RNA	60000
Ptpn1	30	2	167772933	4.676492542	167761129	167781786	20658	5	2.953E-05	**Ptpn1**	1	1	intron 1	30000
Mbnl1	10	3	60376070	4.048271321	60372155	60377049	4895	5	2.969E-05	**Mbnl1**	0	3	intron 2	10000
Bach2	30	4	32475828	5.857357482	32452359	32493430	41072	9	8.273E-06	**Bach2**	5	0	All intron 2 or forward just upstream of shorter protein coding transcript	30000
Pax5	10, 30, 60, 100	4	44676720	7.228924759	44644450	44703122	58673	9	1.11E-16	**Pax5** Gm12462	0	7	most intron 5 reverse, cis spans introns 3 to intron 6	30000
Pum1	30, 60	4	130250177	5.711543795	130232576	130264846	32271	6	1.119E-05	**Pum1**	4	1	introns 1 to 3, most hits intron 2	30000
Gnb1	30, 60, 100	4	154906562	7.808311057	154857534	154926173	68640	13	7.326E-05	**Gnb1** Gm13171	5	1	5'- intron 8	100000
Flt3	10, 30, 60, 100	5	148188504	8.513737312	148139416	148188504	49089	18	0.0001188	**Flt3**	14	0	all these hits in intron 9 although CIS spans exon 3 to 3' end	100000
Mir183	100	6	30131693	8.343864493	30062821	30180888	118068	15	5.851E-05	Nrf1 7SK Mir182 Mir96 Mir183 Ube2h	2	3	Unknown target	100000
Nup98	10, 30, 60, 100	7	109299575	17.61338072	109181604	109388054	206451	29	0	Rnf121 Trpc2 Art5 Art1 Chrna10 **Nup98** Pgap2 Rhog	6	17	Nup98, CIS extends 5' upstream and down of Nup98, but only one hit 5', rest are in introns 10-11 to 31 -32 of Nup98	100000
3930402G23Rik	10, 30, 60, 100	8	10854515	7.231662386	10819318	10904377	85060	14	1.11E-16	intergenic	6	4	no genes, true intergenic	30000
Zfp423	10, 30, 60, 100	8	90423494	5.562595401	90397096	90461624	64529	11	2.544E-06	**Zfp423**	8	0	all hits in intron 1 or 2	30000
Mll1	10, 30	9	44644326	6.948226871	44617971	44664825	46855	12	4.653E-08	**Mll1**	5	1	Reverse hit is in intron 27, the rest are forward in introns 8-10	30000
Gm12223	10, 30, 60, 100	11	54065045	32.09905927	53965301	54164790	199490	53	0	Gm12221 4933405E24Rik Gm12222 **Csf2** Gm12223 **Il3** Acsl6 Gm12224	21	1	Cluster upstream of Csf2 or Il3 in forward orientation	60000
Nf1	10, 30, 60, 100	11	79259360	16.20923841	79159615	79347370	187756	61	0	**Nf1** Gm11198 Gm11199 AU040972 Omg Evi2b Evi2a	13	6	full CIS within Nf1 gene	60000
Stat5b	10	11	100711661	4.976040066	100704833	100717514	12682	20	5.41E-06	**Stat5b**	5	0	all intron 1-2 or 5'	10000
4933426M11Rik	60, 100	12	81943475	6.695593233	81902472	81966906	64435	10	5.695E-05	**4933426M11Rik**	1	4	Intron 1-3	60000
Ube2e2	60, 100	14	19593489	7.146837391	19564855	19610669	45815	10	7.875E-05	**Ube2e2**	2	3	whole CIS in intron 3	60000
Il2rb	10, 30	15	78323417	5.051719596	78313713	78331181	17469	5	2.22E-16	**Il2rb**	5	0	Intron 1-2 or 5'	10000
Tnrc6b	10, 60, 100	15	80666091	8.087976291	80613530	80718653	105124	17	8.84E-06	**Tnrc6b**	4.5	1.5	CIS Intron 1 -10, hits mainly in intron 2 or 4	60000
Crebbp	10, 30, 60, 100	16	4189640	16.35247926	4082759	4257656	174898	26	9.503E-07	**Crebbp** Gm5766	5	8	in Crebbp or 5' of it	100000
Ubn1	60, 100	16	5066328	10.12297524	5031438	5089587	58150	20	3.723E-05	Glyr1 Ubn1 U6 Ppl	2	2	Visible hits all in Ubn1 introns although CIS extends beyond	60000
Rps6ka2	10, 30, 60, 100	17	7349648	8.553799318	7291252	7388579	97328	12	5.454E-05	Gm1604b **Rps6ka2**	0	5	all upstream and reverse to Rps6ka2	100000
Phf6	10, 30, 60, 100	X	50285172	6.47547139	50256016	50285172	29157	11	0.0001601	**Phf6**	3	3	5' - exon 5	100000

Table 4.1 Details of the CIS identified in the *Npm1^{cA} GRL* IM cohort. The genes indicated in bold are the presumed target gene

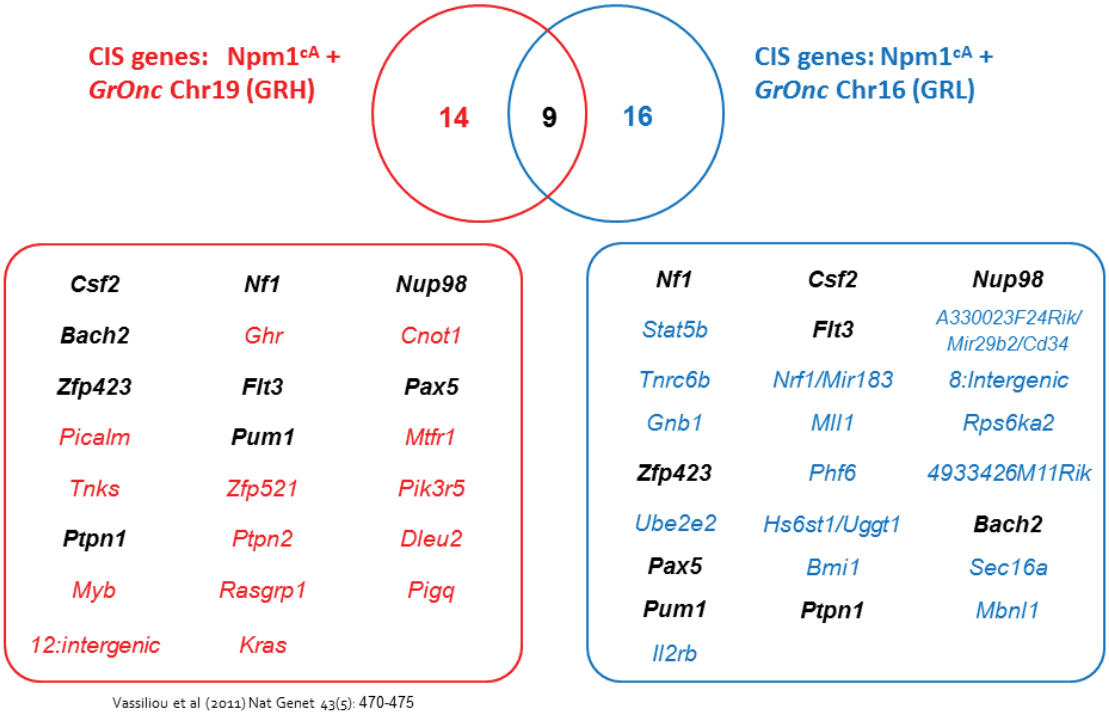


Figure 4.8: CIS genes identified on the two screens. The genes are ordered left to right according to the frequency with which they were hit. *Pten* is excluded from the GRH and *Crebbp* and *Ubn1* from the GRL list as these were near the donor site and may reflect local hopping.

The distribution of integrations across the genome in the final tumour samples for the low copy cohort is shown in figure 4.9. The *SB* transposon is known to exhibit local hopping. Less local hopping was mapped in the GRL cohort than in the published model, which probably relates to the location of the donor site within the centromere. However, the number of integrations on chromosome 16 was still double the expected number and it is difficult to be certain of the validity of the CIS involving *Crebbp* and *Ubn1* given the higher background integration rate along chromosome 16.

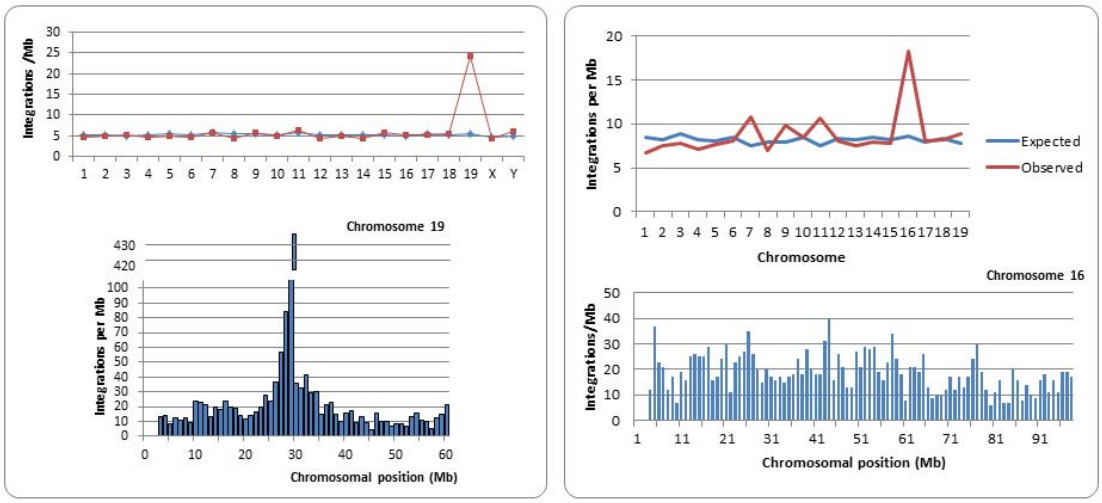


Figure 4.9: Distribution of integrations across the genome (top) and within the donor chromosome (bottom) for the published GRH cohort (left) and the GRL cohort (right).

4.2.3 *Sleeping Beauty* driven leukaemia develops suddenly without detectable antecedent abnormalities in the peripheral blood

Twenty five mice were bled fortnightly from the time of pIpC injection until the development of leukaemia or other illness. This included 17 mice with all conditional alleles, two mice that had the *SB* transposon but were *Npm1*^{WT} (hereafter called *SB* only) and six WT mouse. Of the mice that contained both *Npm1*^{cA} and the transposon, fourteen received the full course of pIpC injections (*Npm1*^{cA} IM mice). One mouse (6.4g) received only two and two mice (7.7a and 7.7b) had no pIpC injections to activate the conditional alleles. Both of the *SB* only mice received the full course of pIpC.

The fourteen *Npm1*^{cA} IM mice showed a marked variation in tumour latency. All but one of these mice had a stable white cell count (WCC) until the final fortnight, when it increased sharply (figure 4.10a). The other blood count parameters were also typically normal in the pre-leukaemic phase, with the exception of 7.5c. This mouse showed progressive polycythaemia and thrombocytosis across serial samples (figure 4.11). The two *Npm1*^{WT} IM mice showed a similar pattern to the *Npm1*^{cA} IM mice; reaching an inflection point where the WCC rapidly rose, but for the WT mice the WCC was stable until death (figure 4.10b). The clinical details of the mice which were serially bled are shown in appendix 4B.

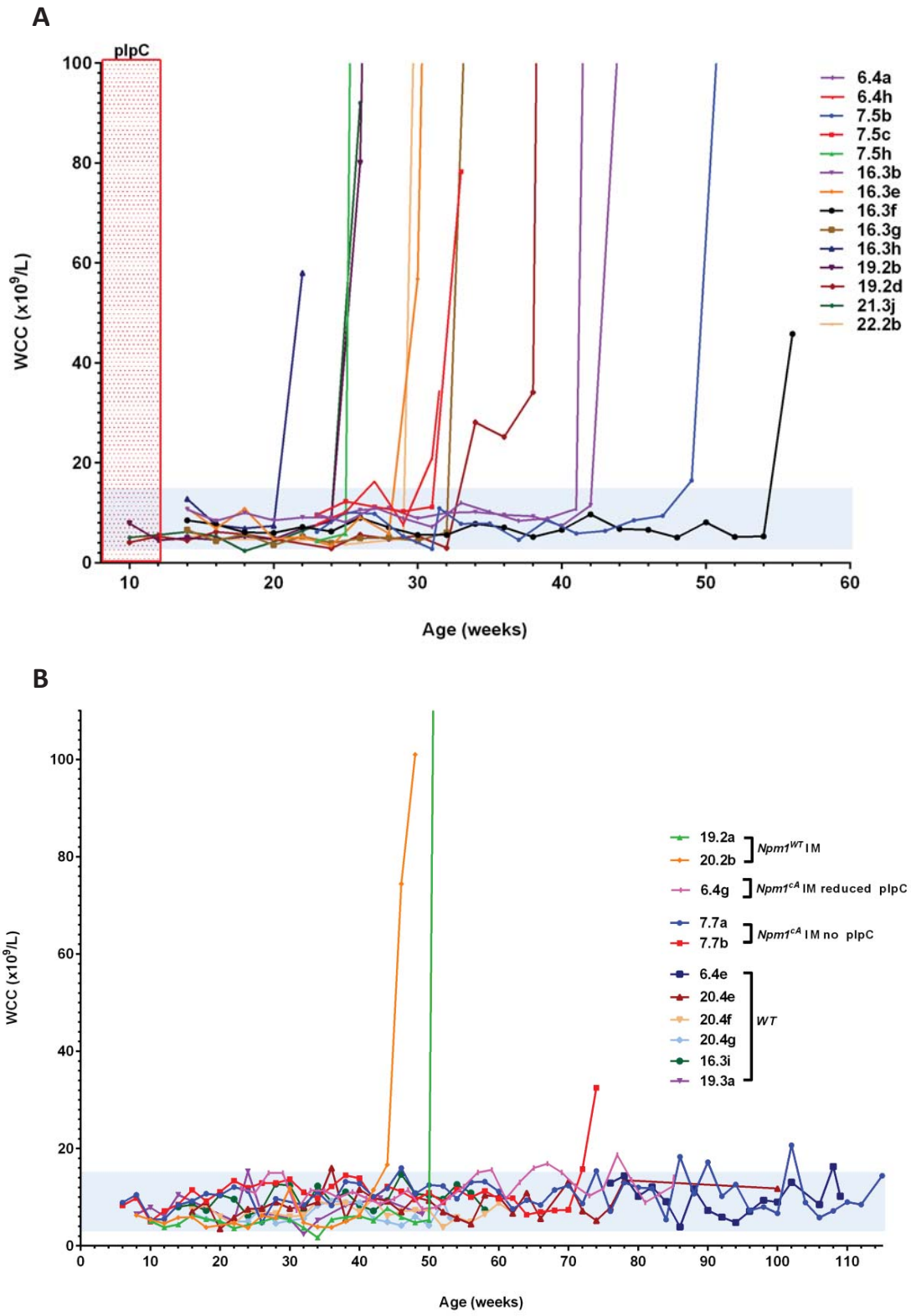


Figure 4.10: WCC in serial blood tests from the Npm1cA GRL IM cohort (A) and the mice wildtype for at least one allele or with incomplete course of pIpC (B). The normal range is indicated in blue. Timing of the pIpC injections is indicated in red. Injections in an individual mouse were given over 2 weeks.

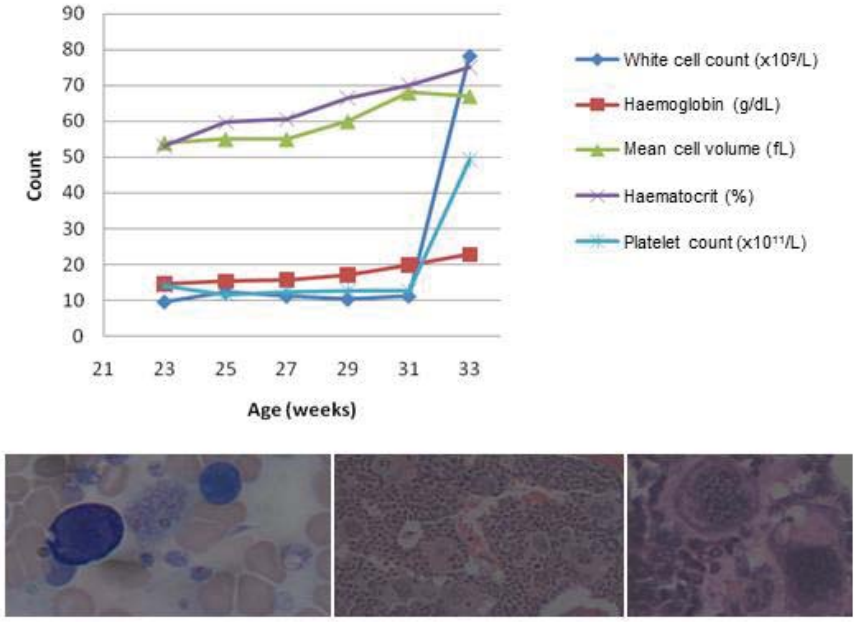


Figure 4.11: Unusual characteristics of mouse 7.5c. Top: Blood parameters. **Bottom:** Blood film and bone marrow pathology showing giant platelets and increased megakaryocyte number, some with atypical morphology. Photographs provided by G Hoffman.

4.2.4 Transposon mobilisation begins early and continues throughout the pre-leukaemic period

SB integrations presented in this chapter were detected using the non-quantitative digestion, splinkerette and 454 sequencing approach. With this method *SB* integrations were detectable throughout the genome even in the first blood samples. Local hopping was evident with a larger number of integrations within chromosome 16, particularly in the earlier blood samples (figure 4.12). In 172 pre-leukaemic blood samples from IM mice that went on to developed leukaemia, on average 504 unique integration sites were identified per sample. The mean total read number per sample was 2992, although this varied widely (standard error (SE) 152) (figure 4.13). This compares to a mean of 516 unique integrations and 3146 (SE 284) total reads in 50 insertional mutagenesis spleen samples taken at the time of death (p=0.86 and 0.63 respectively). The specific overlapping integrations for two mice are shown in figure 6.14.

Figure 4.12: Transposon distribution over time. The number of transposon integrations in 1Mb bins are shown across the genome in samples taken 2 (left) and 12 (right) weeks after completion of pIpC injections. The data are pooled from sibling mice 16.3e, 16.3g and 16.3h.

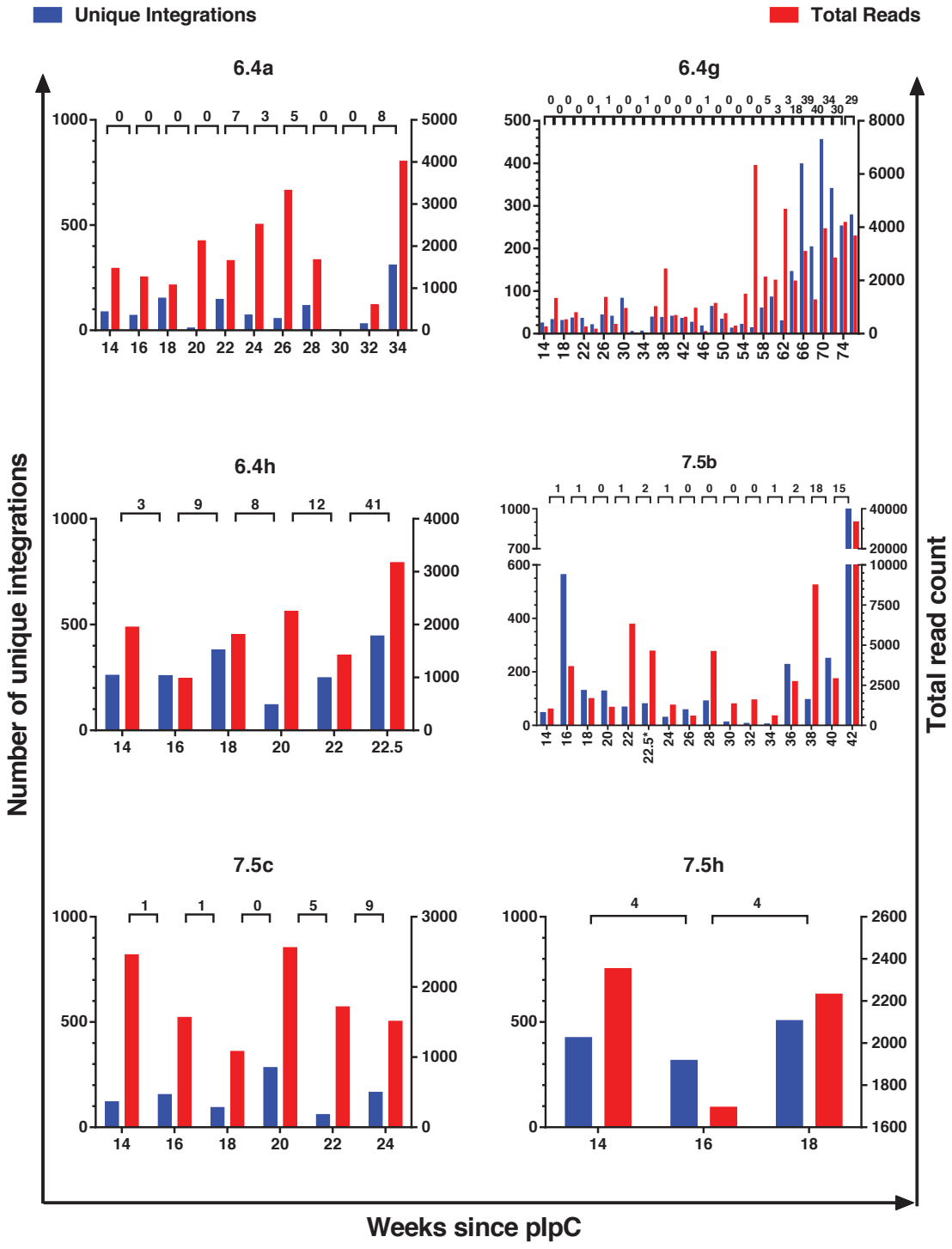
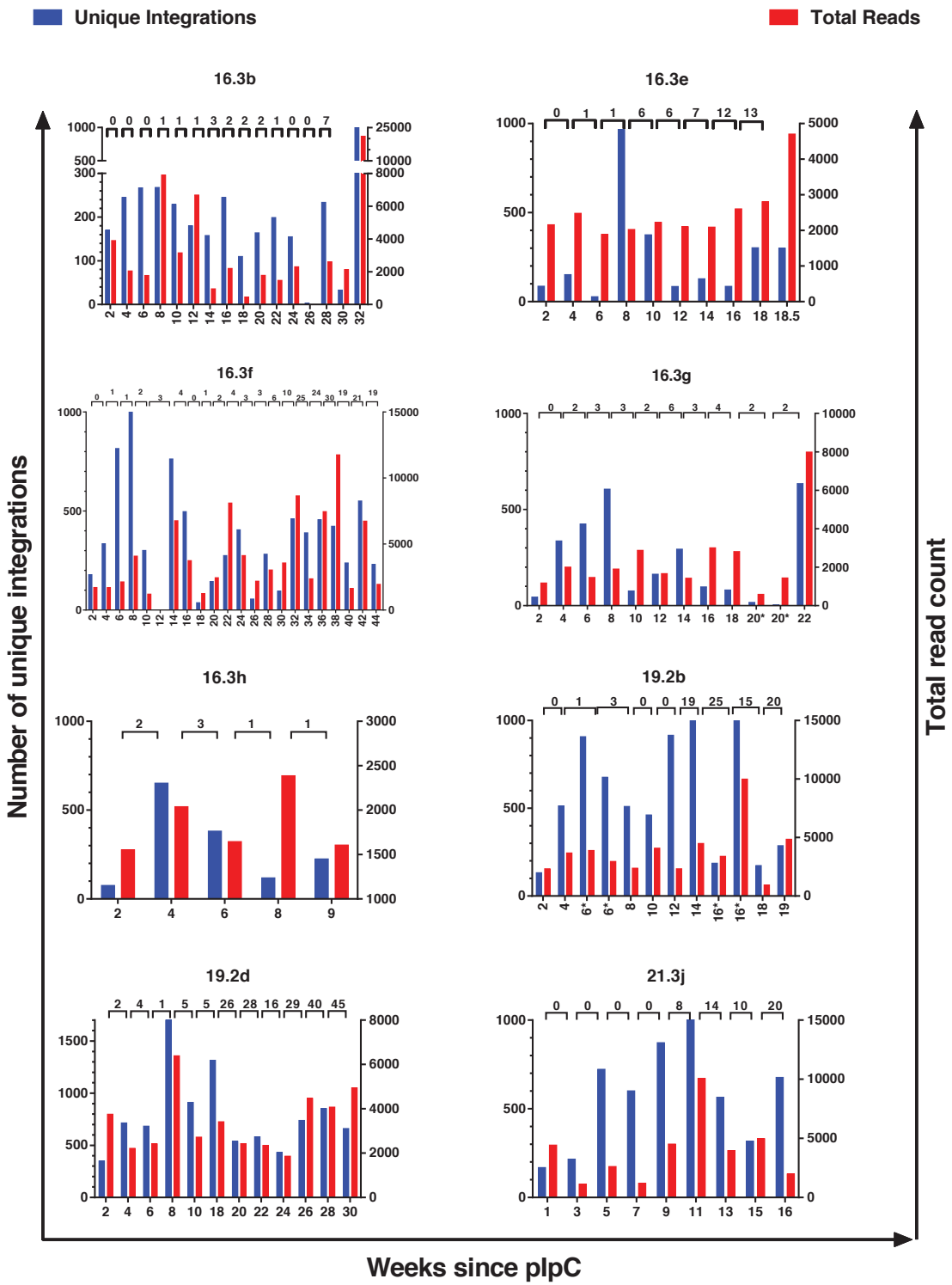


Figure 4.13: Number of reads mapped and overlapping integrations between blood samples in the serially bled mice. Overlapping integrations between consecutive samples are shown at the top. For some specimens the sample was run in duplicate or samples were repeated less than two weeks apart prior to the onset of leukaemia and these are indicated (*). The time after the midpoint of pIpC injections is shown on the X axis, except for 7.7b where the mouse age is given as this mouse did not receive pIpC.


Unique Integrations
Total Reads
16.3b
16.3e
16.3f
16.3g
16.3h
19.2b
19.2d
21.3j
Number of unique integrations
Total read count
Weeks since pIpC

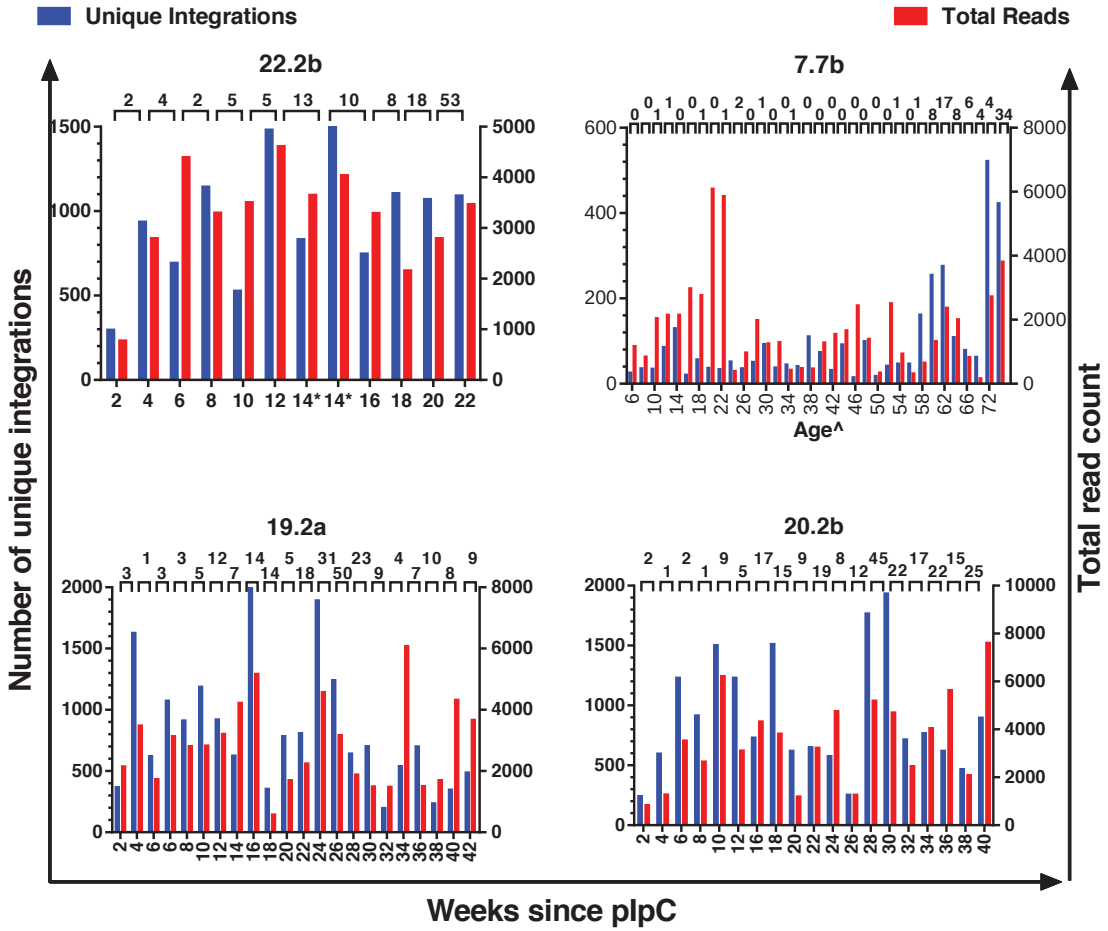
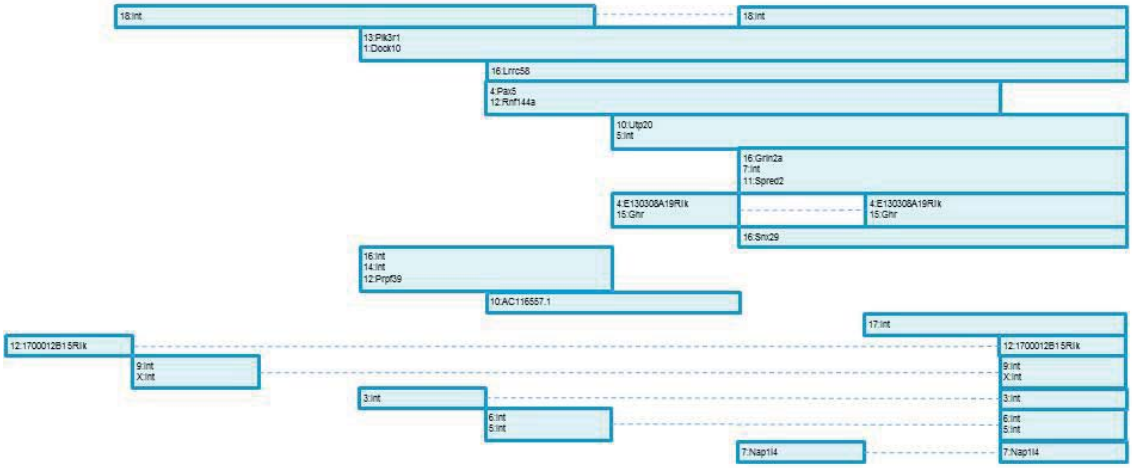


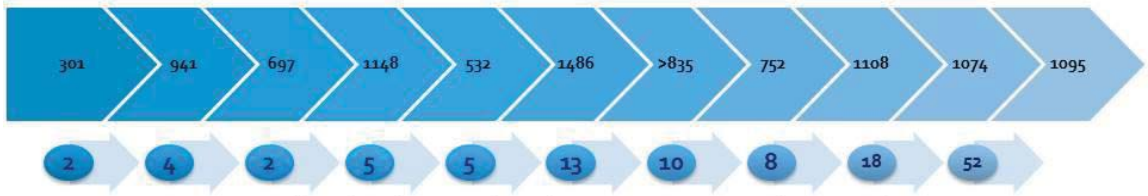
Figure 4.14 (next page): Overlapping integrations in 16.3e (top) and 22.2b (bottom). The arrows show the number of reads per blood sample and the number of overlapping integrations between consecutive samples. The chromosome and gene names are given. Int = intergenic. Each intergenic site listed in separate rows is different.

Number of unique transposon integrations per blood sample:16.3e

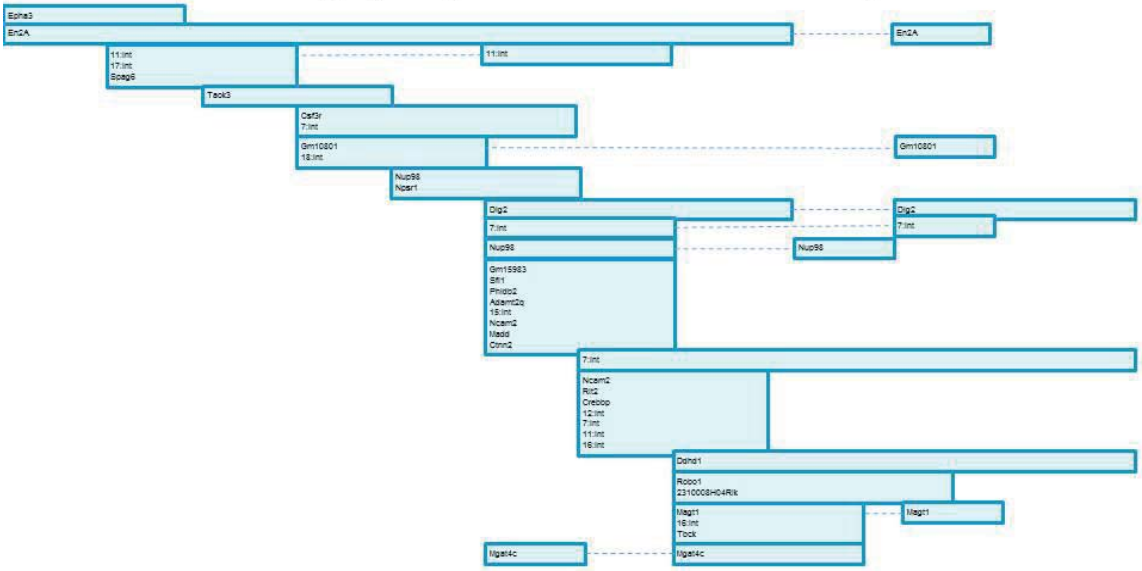
Overlapping integrations between consecutive samples



Number of unique transposon integrations per blood sample: 22.2b



Overlapping integrations between consecutive samples



4.2.5 A small number of transposon integrations occur early and persist in the pre-leukaemic samples and on serial transplantation of leukaemia cells

The vast majority of transposon integrations appeared only transiently. A minority of transposon integrations persisted on sequential samples from the same mouse, although the number of persisting integrations typically increased over time (figure 4.13). Pre-leukaemic blood samples were analysed for transposon integrations identified in leukaemic spleen and other blood samples from the same mouse. All of the transposon integrations shared between the tumour sample and the pre-leukemic blood tests are shown in figure 4.14 for two of the mice that were serially bled. There were several examples where a mutation was evident in the blood and persisted on all blood samples for two months prior to the diagnosis of leukaemia (figures 4.14 and 4.15). The persisting integrations in the serially bled mice not included in figure 4.15 are shown in appendix 4C.

Spleen cells from mice with leukaemia were serially transplanted into NSG mice by tail vein injection. The details of all of the transplants and the cell doses used are shown in appendix 4D. The transposon integrations in the recipient tumours were compared with those from the primary tumour (figures 4.15 and 4.16 and appendices 4C and E). Some but not all of the integrations that were found in serial blood samples persisted in the transplant tumours. The integrations that persisted on serial blood and/or transplant samples were enriched for CIS genes, however not all CIS integrations in the primary tumour persisted on transplant (figure 4.15 and 4.16, table 4.2 and appendices 4C and 4E).

Transplants were performed at varying cell doses from 1 million down to 100 cells. The transplanted cells generally engrafted and generated leukaemia when a dose of at least 10^4 unsorted spleen cells were used. Upon transplantation of 1000 cells, only 12 of 21 transplants generated leukaemia and with 100 cells this dropped 5 of 19 transplants.

Figure 4.15 (next pages): Shared integrations on serial blood and transplant recipient tumours for mice 6.4a, 16.3b and 16.3e (next pages). The precise position of each integration is shown across the top. Integrations in a position are indicated by the coloured squares (blue = serial blood or primary tumour spleen, yellow = recipient tumour). The integration sites that fall within CIS are indicated in red. The age of the mouse is shown in weeks for each of the blood samples. IDs of the recipient tumours are indicated. Integrations are shown by the order in which they accumulated and only integrations that persisted on multiple samples, including the tumour are shown.

6.4a
Multiple Genes
2:169984351:intergenic
1:32755148:intergenic
16:62363976:intergenic
1:91333600:intergenic
14:82887891:intergenic
6:362703365:RP23-315E20.1
16:14405801:Abcc1
18:25462704:AW554918
3:46277061:intergenic
8:46188381:intergenic
10:25971688:Samd3
11:21774655:AV249152
8:10852483:intergenic
4:154885120:Gnb1
1:63131127:Ino80d
5:37354156:intergenic
16:64253688:intergenic
11:79223829:Nf1
6:30084965:Nf1
11:79323292:Nf1
10:89173387:intergenic
6:30261690:1700095J07Rik
10:5128701 2:intergenic
2:104783648:Cdc73
7:152239044:intergenic
1:80260343:intergenic
18:50035824:Dmxl1
4:106547362:intergenic
12:32823591:intergenic
10:83971240:intergenic
8:37122561:intergenic
18:81680072:intergenic
7:132982444:D430042O09Rik
13:96629191:Igap2
14:26428815:Zmiz1
4:129409116:Kndtbs1
2:84616172:intergenic
3:94295751:Celf3
11:117160836:Sept9
7:110830355:intergenic
18:55517632:intergenic
2:51135622:Gm13497
7:100402062:intergenic
16:42681266:intergenic
7:133741858:intergenic
16:80879715:intergenic
23
25
27
29
31
33
35
37
41
43
Spleen
CIS
1.1
1.1.1
1.1.2
1.1.3
1.1.1.3
1.1.1.1
1.1.1.2
1.1.1.1.1
1.1.1.1.2
1.1.1.3
1.1.2.1
1.1.2.3
1.2
1.2.1
1.2.2
1.2.3
1.3
1.3.1
1.3.2
1.3.3

16.3b
16:10346337:intergenic
7:65973402:Atp10a
11:48968483:intergenic
13:55618124:intergenic
10:80425364:Gng7
X:124392673:intergenic
11:46288952:Havcr2
2:4163697:Lrp1b
19:28519955:Glis3
10:62661999:Mypn
6:149502566:Bicd1
4:118172338:intergenic
14:80186680:intergenic
17:62698266:intergenic
8:26500846:intergenic
7:43911103:Rdh13
5:126063091:intergenic
6:148179673:intergenic
4:35556564:intergenic
7:74182211:Adamts17
12:78606726:intergenic
16:76962433:intergenic
17:92877728:intergenic
18:35286717:Chrna1
19:34138283:intergenic
X:87226500:intergenic
9:1117172:Dhx3
9:66657688:intergenic
15:41608768:Oxr1
11:48675784:Irgm1
5:64695168:Tbc1d1
X:126262397:intergenic
5:24193714:Nub1
11:54064947:Gm12223
15:77820769:intergenic
4:95037030:intergenic
9:44650101:Mll1
13:76622013:Mcph1
1:186852488:intergenic
Multiple Genes
14
16
18
20
22
24
26
28
30
32
34
36
40
42
44
Spleen
CIS
1.1
1.1.1
1.1.2 (1000)*
1.1.3 (100)
1.2
1.2.1
1.2.2 (10 000)
1.2.3 (1000)*
*no tumour

16.3e
18:14143512:intergenic
13:102459811:Pik3r1
1:80623054:Dock10
4:44688758:Pax5
16:37872548:Lrrc58
12:27007562:Rnf144a
5:97376868:intergenic
10:88230868:Utp20
7:121331422:intergenic
15:34387756:Ghr
4:59655757:E130308A19Rik
16:99241144:Grin2a
11:198354833:Spred2
16:116631893:Snx29
7:150708587:Nap1l4
11:54064951:Csf2
9:75039016:Myo5a
10:75796148:Dip2a
11:23208833:Usp34
8:46188381:intergenic
8:10863347:intergenic
7:56129265:E2f8
6:29573899:intergenic
X:102735473:intergenic
5:23362319:intergenic
9:78726091:intergenic
6:122457471:Mfap5
13:80854844:intergenic
9:23970864:Npsr1
3:71312161:intergenic
13
15
17
19
21
23
25
27
29
29,5
29.5 Spleen
CIS
1.1
1.2
1.3
1.3.1
1.3.2
1.3.3

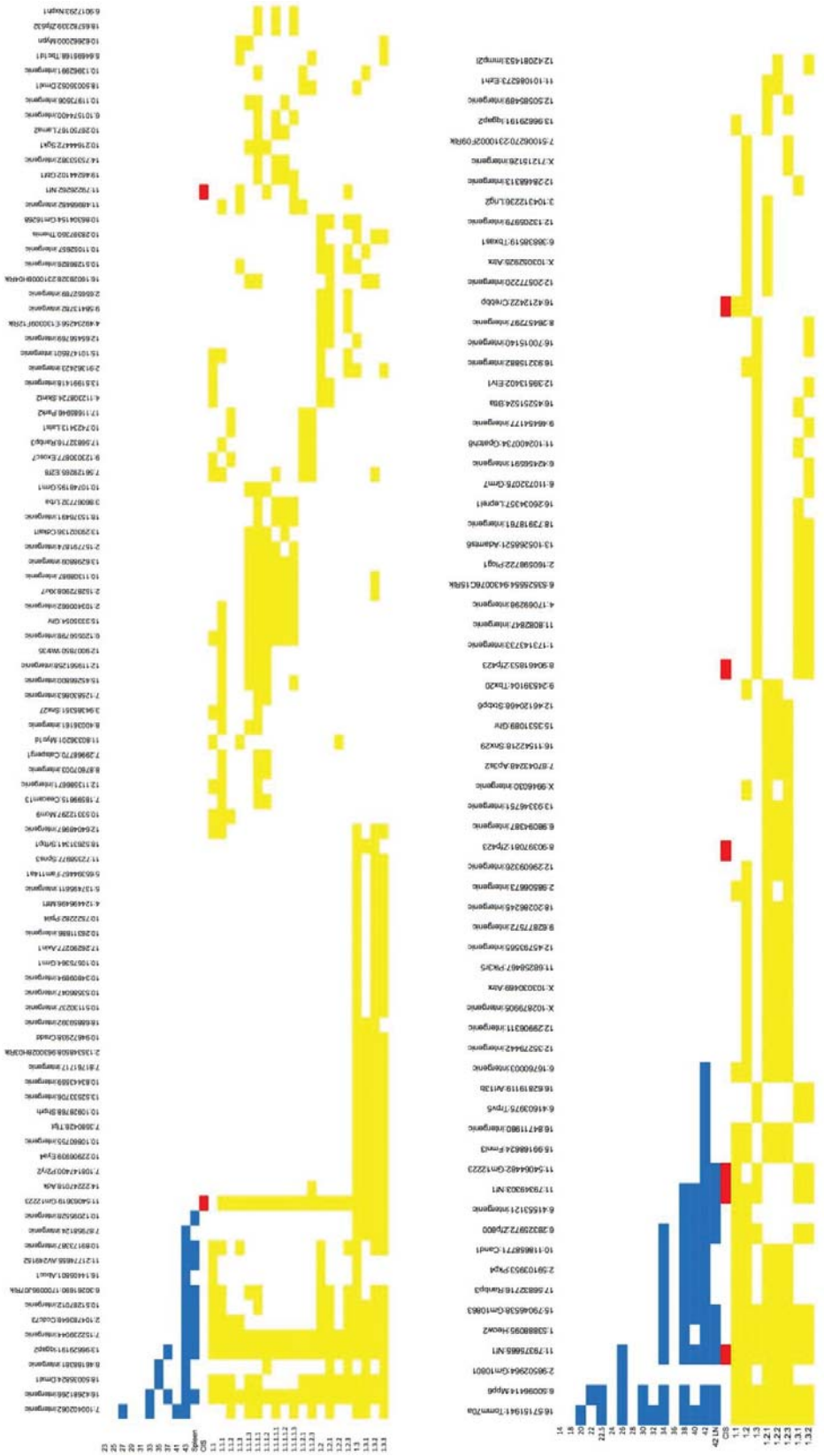
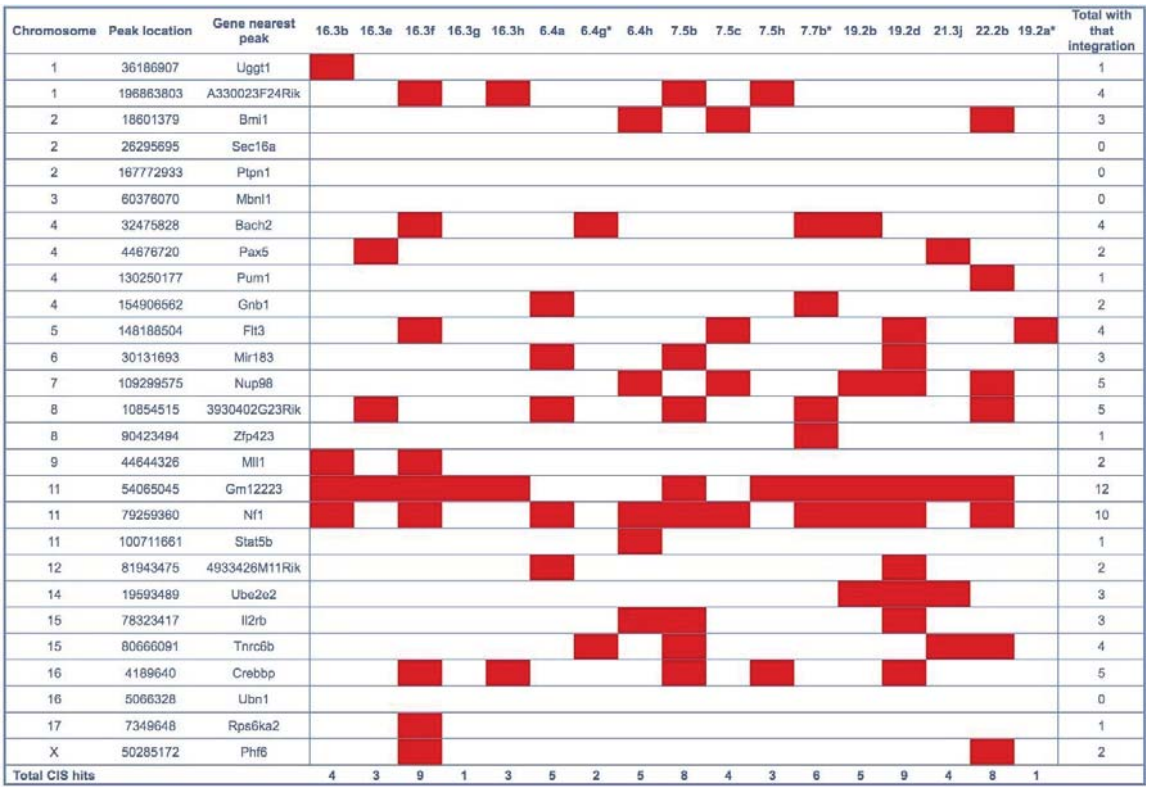


Figure 4.16: Shared integrations on serial transplants from 6.4a (top) and 7.5b (bottom). Integrations are represented if they were found in ≥2 transplant samples. The format is otherwise similar to figure 4.15. Several integrations were shared in multiple recipient tumours even though they were not detected in the primary tumour sample. Tables for several other mice are shown in appendix 4E.



Chromosome	Peak location	Gene nearest peak	16.3b	16.3e	16.3f	16.3g	16.3h	6.4a	6.4g*	6.4h	7.5b	7.5c	7.5h	7.7b*	19.2b	19.2d	21.3j	22.2b	19.2a*	Total with that integration
1	36186907	Uggt1																		1
1	196863803	A330023F24Rik																		4
2	18601379	Bmi1																		3
2	26295695	Sec16a																		0
2	167772933	Ptpn1																		0
3	60376070	Mbnl1																		0
4	32475828	Bach2																		4
4	44676720	Pax5																		2
4	130250177	Pum1																		1
4	154906562	Gnb1																		2
5	148188504	Flt3																		4
6	30131693	Mir183																		3
7	109299575	Nup98																		5
8	10854515	3930402G23Rik																		5
8	90423494	Zfp423																		1
9	44644326	Mll1																		2
11	54065045	Gm12223																		12
11	79259360	Nf1																		10
11	100711661	Stat5b																		1
12	81943475	4933426M11Rik																		2
14	19593489	Ube2e2																		3
15	78323417	Il2rb																		3
15	80666091	Tnrc6b																		4
16	4189640	Crebbp																		5
16	5066328	Ubn1																		0
17	7349648	Rps6ka2																		1
X	50285172	Phf6																		2
Total CIS hits			4	3	9	1	3	5	2	5	8	4	3	6	5	9	4	8	1	

Table 4.2: CIS identified in spleen sample from serially bled mice. The mice with * were not standard *Npm1*^{cA}/IM mouse and were not included in the CIS analysis. 6.4g was not found to have leukaemia on histopathology and received a reduced number of pIpC injections. 7.7b had no pIpC and 19.2a was *Npm1* wildtype. The spleen from mouse 20.2b was not analysed with this protocol and it is therefore excluded from the table.

4.2.6 CIS in the pre-leukaemic blood samples

A kernel analysis was performed on the blood samples from the cohort of serially bled *Npm1*^{cA} IM mice at selected time points prior to the euthanasia of sick mice. The full results are shown in appendix 4F, and the integrations that overlapped with the CIS identified in tumours from the whole cohort are summarised in table 4.3. The detection of CIS was limited due to the small number of samples, but it is notable that the top three CIS were all identified in the analysis taken on blood samples 24-33 days prior to death.

Sample	Number of samples included in analysis	CIS
Final tumour	15	Csf2, Nf1, Nup98, Mll1, Nrf1, A330023F24Rik
24-33 days pre-tumour	15	Csf2, Nf1, Nup98, Pax5, Bmi1/Commd3
51-61 days pre-tumour	14	Bach2
79-88 days pre-tumour	11	
91-113 days pre-tumour	11	

Table 4.3: CIS from the tumour analysis that were detected in the pre-leukaemic blood

4.2.7 Some transposons loose the capacity to re-mobilise

It is likely that the transposon integrations driving leukaemogenesis are among the small group that persist on serial blood samples and transplants. However, it is also probable that the persisting integrations include passengers. Such passenger integrations may have persisted because either (i) they preceded drivers and did not have time to remobilise or (ii) the transposon lost the ability to re-mobilise. Possible explanations for this include mutation of the repeat sequences or if a transposon jumps inside another and then remobilises using a non-contiguous *SB* repeat, a phenomenon I have termed '*neopartnerships*' (see Materials and Methods figures 2.3 and 2.4). It is possible this happens inside the donor locus, in which case these events cannot be mapped, however if it happened elsewhere in the genome the unpaired *SB* repeat would be 'stuck' and the integration site would be mapped on sequential samples.

To look for evidence of transposons jumping within adjacent transposons I initially performed PCRs from the repeat sequences of the transposon. The length of the *GrOnc* transposon between *SB* repeats is ~2700bp and so amplification from adjacent transposons would be expected to give a large product. In fact, amplifying from a forward and reverse primer both positioned within the *SB* 5' repeat gave several bands of varying size from 200bp to just over 1000bp. Similarly PCR from a forward primer in the *SB* 5' repeat to reverse primers in the *PB* 5' repeat or between the *SB* 3' an *PB* 5' repeats also gave multiple bands of under 1000bp, suggesting re-insertion within other transposons does occur (figure 4.17).

To further investigate if more than one transposon was mobilising as a single unit from the donor locus I designed a splinkerette to sequence from the blunt end of the

*PB*5' repeat. I identified an average of 23.4 *PB* integration sites across the chromosome in the 17 IM samples which suggests that mobilisation of more than one transposon together from the donor site is not an exceptional event. There was a detectable false positive rate with an average of 3.75 *PB* integrations detected in three *Cre* negative and one transposase negative control, however 12 of these 15 integrations were detected in a single sample and generally the number and read coverage of *PB* integrations detected in the IM samples is higher. Overall I identified four examples where *PB5'* blunt end integrations were found at the same site as a *SB* integration which was present in the serial blood, tumour or transplant samples (appendix 4G). One example was in mouse 6.4g at position 14:120987953, and this integration had very high read coverage (appendix 4G). However, in these four cases the integration site was typically detected from both ends of the *SB* transposon, rather than a single end which would be expected with the '*neopartnership'* scenario I described. In the one example (7.5h) where the integration site was only detected from one end of the *SB* transposon, this was the 5' and not the 3' *SB* repeat.

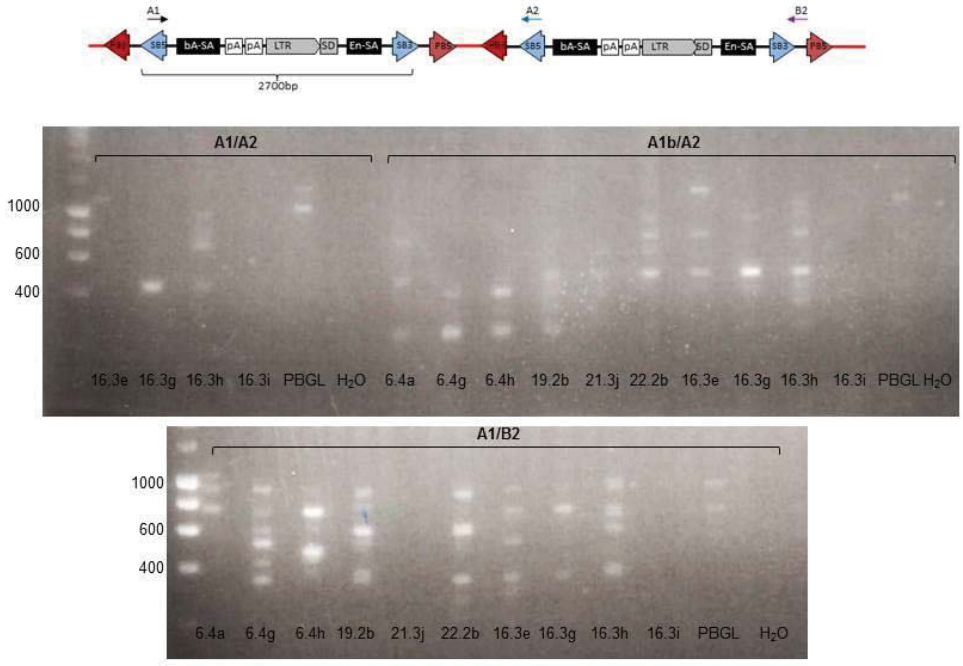


Figure 4.17: PCR to detect hopping into other transposons. The position of the primers is shown. 16.3i is a no Cre control and PBGL contains the low copy GrOnc but is mobilised with *PB*.

Together, this data suggests that transposons do integrate within other transposons and that multiple *SB* transposons can mobilise together from the donor site and re-integrate elsewhere in the genome. However, the '*neopartnership*' scenario is not a major cause for persisting integrations on serial sampling.

4.2.8 Searching for alternative drivers in transposon IM mice

It was noted, both in the GRH and GRL IM cohorts, that occasional tumours did not contain transposon integrations in any of the CIS genes. It remains possible that the non-CIS insertions in these leukaemias were drivers, but it was also considered that other mechanisms could be driving these tumours. The potential alternative mechanisms include (i) the *SB* transposon leaving a footprint after it re-mobilised that disrupted a gene but was no longer identifiable on routine analysis, (ii) acquisition of sporadic (non-transposon) coding mutations and (iii) acquisition of chromosomal aberrations or copy number changes. We therefore performed exome sequencing and CGH on selected tumours to look for evidence of such changes.

We performed exome sequencing on ten primary tumour spleen samples from the mice that were serially bled as well as four transplant recipient tumours and four non-IM mice from the cohort that did not develop AML. The canonical *SB* footprint is CTGTA, but other footprints are possible, particularly 5bp insertions and small deletions. The analysis was performed by Ignacio Varela. Although over 1000 insertions and deletions were mapped in coding regions in these samples, not a single canonical *SB* footprint was identified. The majority of the identified abnormalities were shared by multiple samples, suggesting these were polymorphisms rather than true insertions and deletions. Amongst the small number of frameshift mutations that were unique to specific samples, we did not identify any in recognisable tumour genes. Therefore, although gene knockout due to a transposon footprint in a coding gene is possible, I did not find evidence of this in this cohort and we also did not identify any recognisable sporadic mutations in known AML drivers.

To investigate whether copy number aberrations were occurring in these mice we performed comparative genomic hybridisation (CGH) on four primary tumour samples (6.4a, 6.4h, 7.5b and 19.2b) and one recipient tumour sample from mouse

21.3j. Samples 7.5b and 6.4a had evidence of monosomy 7 on CGH, but overall there was no evidence of significant copy number changes.

4.3 Discussion

These results confirm and extend many of the important findings from the published GRH insertional mutagenesis model of *Npm1*^{cA} mutant AML. Using a new donor locus and a reduced number of transposons I was able to validate many of the major CIS genes identified in the published work. On serial blood sampling I demonstrated that the blood leucocytosis occurred suddenly, without antecedent abnormalities in the full blood count (FBC) in the majority of mice. Furthermore, I was able to study the order of acquisition of mutations and by comparison of integrations in serial blood, primary and recipient leukaemia samples I identified a narrow pool of integrations amongst which the driver mutations for that primary tumour appear to reside.

The serial assessment of blood and tumour samples clearly shows that transposon mobilisation is continuous both before and after leukaemia develops. Although only a small number of integrations persist in all of the serial transplants, there are other groups of mutations which track down particular lines of recipient mice but are not evident in the primary tumour. Either these integrations were newly acquired in the recipient mouse or they were only present in rare cells in the primary tumour and fell below the limit of detection. The majority of the detected integrations were not shared between mice, even recipients transplanted with the same primary tumour, which suggests that many transposons are in 'passenger' positions within sub-clones and that transposons are continuing to re-mobilise. However, it is also true that when the same sample was run twice, although there was considerable overlap, the detected integrations were not identical. The differences in integrations from identical samples run in duplicate may result from the limited amount of input DNA used or because the sequencing depth was insufficient to detect all of the integrations present in small sub-clones. I limited the amount of DNA used in serial blood splinkerette analysis to attempt to standardise this between blood samples, realising that the DNA yield from some samples would be small. The quantity of DNA used (100ng) may have been insufficient to capture the full heterogeneity of integrations. However, generally the integrations that persisted on transplant were found in both runs.

Typically the pre-leukaemic blood and tumour samples contained hundreds of unique transposon integrations, despite the fact that each individual cell started with only 15 copies of the *GrOnc* transposon. The number of transposon integrations per cell is likely to fall over time as re-integration of *SB* is not 100% efficient (Liang et al., 2009; Luo et al., 1998). Therefore, rather than a homogenous population, the leukaemia likely contains a large number of sub-clones with varying malignant potential, which are competing for resources and 'real estate', akin to Darwinian evolution. On serial transplantation of 1 million mixed tumour cells a small set of recurrent integrations were consistently detected suggesting these include the driver mutations for both the original and the re-emergent clone/s.

In this analysis the number of reads assigned to any transposon integration cannot be used to estimate the fraction of cells with a particular integration as the method was not linear and relied on DNA digestion and 62 rounds of PCR amplification prior to sequencing. The ability to amplify transposon integration sites varies depending on the location of the nearest restriction site. Although the *MboI* enzyme is a frequent cutter, the distance between the end of the transposon and the nearest restriction site will vary widely. Also, factors such as GC content will bias the PCR reaction to amplify some integration sites more efficiently than others. This is evident from the variable read coverage detected for specific integrations on analysis from the 3' and 5' ends of the transposon. In fact, it was not uncommon for a particular integration site to be mapped only from one end of the transposon.

It is also evident from the serial blood sample data in figure 4.13, that increased read depth does not necessarily correlate with a significantly higher number of unique transposon integrations, although as expected, few unique integrations are mapped when the read depth is poor. Again this variation likely reflects the preferential amplification of particular transposon integrations. When the splinkerette PCR amplifies a small number of integrations very well, a large proportion of the 454 sequencing reads are taken up by these sites. In spite of the good overall read number, coverage of other integration sites is limited. In some samples there was also an artefact due to reads which were amplifications from the transposon primer directly into the Splinkerette linker, without intervening DNA or the transposon end sequence. This is presumed to occur due to non-specific annealing of the primer and in a few samples this accounted for 25-30% of total reads, whereas in most it

was <1%. Varying the amount of linker used in the ligation reaction did not have a consistent effect on the proportion of reads with this artefact. Although these reads were filtered out in the analysis because they did not start with 'TG' corresponding with the end of the transposon sequence, they did reduce the read coverage for true transposon integrations in some samples.

When a transposon integrates into a genomic location it can freely re-mobilise. The persistence of certain integrations on serial sampling reflects the Darwinian evolution of transposon driven tumours. Although it occurs, re-mobilisation of the transposon from a 'driver' position is selected against as cells in which this happens will lose any advantage. The persistence of integrations will also depend on the relative rate of cell division and re-mobilisation of the transposon (figure 4.18). For this reason it is likely that not all of the persisting integrations are drivers. Akin to passenger mutations, 'passenger integrations' persist because they happened to be present in the clone when it acquired a driver integration that led to clonal expansion. Although re-mobilisation of such passengers is not selected against, the integration is likely to be detected on serial sampling because it will not remobilise from all clonal cells before their next cell division. As a clone accumulates 'driver' integrations the rate of cell division is likely to increase, which would make it less likely for 'passenger' transposon integrations to be lost from every cell within the clone.

Transposon integrations would also persist if one of the *SB* repeats was mutated or 'lost'. Although there is an exponential drop in efficiency of transposition with larger elements, transposition is still reasonably efficient with a transposon length of 5kb (Izsvák et al., 2000), so mobilisation of multiple *SB* elements together is plausible. The 'reverse' Splinkerette experiment provides supportive evidence that this occurs, although the 'neopartnership' scenario was not a significant cause of the serially persisting integrations. We also questioned whether leukaemogenesis was driven through mechanisms independent of the mapped transposons, but the canonical *SB* footprint was not identified on exome sequencing, nor did we find evidence of major chromosomal aberrations on CGH analysis.

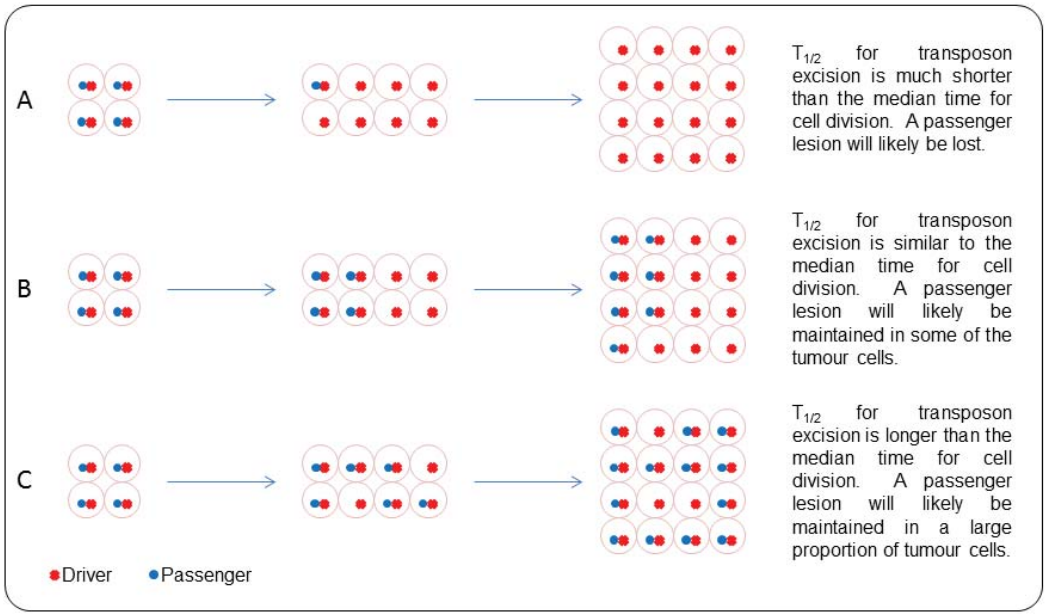


Figure 6.18: Effect of rate of transposon excision relative to cell division on the persistence of passenger integrations in the final tumour sample. In B and C an equilibrium is likely to be reached where the passenger integrations is maintained and continuously detectable in tumour DNA samples.

In several cases some of the transposon integrations found in the final tumour were present in the sequential blood samples for over two months prior to the development of overt leukaemia. This implies the continuous contribution by HSCs with these mutations to haematopoiesis during this interval. The disappearance and re-emergence of other integrations may reflect the proliferative cycles of HSCs with these mutations, or more commonly, the small percentage of cells carrying them and their variable detection by the assay. In some instances this is because of the variable DNA yield from the serial blood samples. For example, in sample 16.3g the paucity of integrations in the 32 week blood sample (20 weeks post pIpC) is likely to be due to a low DNA yield. Although this sample was processed twice for splinkerette and sequencing, a low number of reads were mapped on both occasions. However, there are also gaps in the presence of mutations in some samples in which a large number of integrations were mapped. This may occur because these integrations are carried in only a small proportion of cells and they fall below the limit of detection or because of a paucity of *Mbo1* restriction sites near to the particular integration which biases the PCR amplification against detection of

these integrations. For example, in case 19.2D the integration in *Mll3* (5:25000472) is only 19 nucleotides from the nearest restriction site at one end and there is no restriction site within a thousand bases on the other side of the transposon. If the DNA is completely digested, the transposon DNA fragments containing this integration will be either too short to map or too long to PCR efficiently in the Splinkerette protocol. Samples will vary in the number of easily amplified integration sites, which could differentially affect the rate of PCR amplification of integrations sites that are more difficult to amplify, such as those several hundred bases from the nearest *Mbo1* restriction site. In mouse 6.4g, there are a group of integrations 'missing' from the week 75, 77, 79, 81 and 83 blood samples even though all of these had comparatively good read depth and many other integrations did persist. The finding that these integrations are present in the final tumour, including two which persist on the serial transplants (Dtx2 and Cdyl2), indicates these integrations were not lost, as it would be unlikely that a new integration event occurred at exactly the same site and even more improbable that this happened at multiple positions. It is possible that these integrations occurred in a small clone that fell below the limit of detection, but re-emerged upon acquisition of a driver later on.

Despite such limitations there is important information to be gained by studying the order of acquisition of mutations in these serially bled samples. The serial CIS analysis revealed for the first time, that the *Csf2*, *Nup98* and *Nf1* CIS were all identifiable two weeks before death. *Csf2* integrations appeared as either early events several weeks before any demonstrable changes in blood counts (19.2b, 19.2d, 21.3j, 20.2b) or as late events just as the WCC became abnormal (16.3b, 16.3e, 16.3f, 7.5b, 22.2b). This is also the case for *Nup98* in which integrations were often first detected several weeks before death (19.2b, 19.2d, 6.4g, 7.5b) but were also seen as a late event (7.5c). Similarly *Nf1* integrations occurred as both early and late events and often multiple integrations were detected in the same tumour (e.g. 7.5b, 19.2b, 19.2d). In contrast, where *Mll1* integrations were detected in the serial blood samples they were always late (16.3b, 16.3f, 21.3j). Integrations in *Flt3* were universally in intron 9 in the forward orientation as in the published study suggesting these are activating integrations. Typically these were late events (7.5c, 19.2d, 19.2a) except in 16.3f where a single read was detected in the week 37 blood sample. This integration was not detected again until a blood sample a day before

death. It is difficult to be certain of the veracity of this early integration, particularly as it was a single read detected from one end of the transposon, however if real this would imply the combination of *Npm1*^{cA} and up-regulation of *Flt3* alone is insufficient for leukaemogenesis. However, the combination of *Npm1*^{cA} and *Flt3-ITD* mutations in mice was previously found to cause highly penetrant AML with an explosive onset and short latency (Mupo et al., 2013).

The transposon system provides a platform of rapid mutation acquisition, which in the setting of a predisposing *Npm1*^{cA} background makes the development of leukaemia inevitable. The longer latency in the GRL compared to the GRH cohort reflects the lower mutation rate per cell. However, given the accelerated mutagenesis in both models, it is likely that there are multiple related and possibly unrelated tumour cell populations at different stages of evolution towards leukaemia. Some CIS integrations do not persist on serial transplant, for example the *Csf2* and *Bach2* integrations in 19.2b. The loss of the *Csf2* integration on transplant seems surprising as this is one of the most frequently hit genes in both the GRH and GRL screens. Similarly the *Flt3* integration described above in 16.3f was not detected in the majority of recipient tumours. However, rather than indicating these are not driver integrations in the particular primary tumour, the loss of apparent drivers in recipient mouse tumours is more likely to represent the presence of more than one clone capable of generating leukaemia in the mixed tumour cell population. Even when high cell doses were used, not all clones were necessarily re-established in the transplant model. In sample 21.3j two integrations were identified immediately 5' to *Csf2* and both were in the forward orientation. Although both integrations are found in several blood as well as the final tumour samples, only one persists in all of the transplants. The other *Csf2* integration, along with another CIS hit in *Mll1,* is only found in one of the one million cell transplants (21.3j1.1). It is highly likely that the two *Csf2* integrations occurred in independent clones, one of which co-occurs in the same clone as the *Mll1* integration, while the other was in a clone that dominated the recipient tumours. However, single cell experiments are required to definitively prove this hypothesis.

Precisely how polyclonal these tumours are and how the transposon copy number effects this clonality is yet to be determined. It is likely the number of potentially leukaemogenic clones varies between mice and this may, in part, explain the large

variation in the number of CIS integrations identified in each mouse in this and the GRH model. It is also likely that the different integrations are associated with varying levels of fitness advantage. Some mice may develop leukaemia with only one or two 'strong' integrations whereas others may have multiple 'driver' integrations in a single clone as each one provides only a 'weak' advantage. Whether the differences in the CIS identified between the two cohorts reflects a biological difference due to the mutation rate is uncertain.

Upon transplantation of primary tumours into NSG mice some of the integrations which persisted for an extensive period in the pre-leukaemic serial blood samples were not evident in the recipient tumours. There are at least two plausible explanations for this. The first is that the tumours are oligoclonal and not all clones engraft as discussed above. Secondly, the transplant experiments provide an additional opportunity for dispersion/loss of passenger integrations, particularly when a low number of primary tumour cells are injected. If the number of leukaemia initiating cells (LIC) injected into the recipient mouse is small, then a large number of tumour cell divisions are required before the leukaemia becomes clinically apparent as is evidenced by the longer latency to tumour development. It is likely that any single passenger integration has dispersed from a significant proportion of LIC even in the primary tumour. Such passengers may be lost on transplant either because they were not represented in the LICs that successfully engraft, or because the growth kinetics of the tumour in the recipient mouse are such that the passenger is able to fully dispersed from the tumour clone. I observed that for the integrations that did not persist on transplant, the read number was typically falling in the later serial blood samples (data not shown). This may suggest they only persisted in a diminishing fraction of tumour cells or were not in the dominant expanding clone. However, as previously discussed the analysis method used here is not quantitative and this observation could also be explained by new integrations that PCR amplify easily, resulting in a relatively reduced read number without any change in the number of cells with these integrations.

It is striking that the top three CIS; *Csf2*, *Nf1* and *Nup98,* are identical between the GRL and GRH screens, although of these only *Nf1* is recurrently mutated in human myeloid leukaemia(Boudry-Labis et al., 2013; Haferlach et al., 2012; Parkin et al., 2010). Transposons do not recapitulate the type of mutations seen in human disease. This

is often seen as a limitation of transposon mutagenesis screens, however it is also an advantage as genes and pathways that are less susceptible to natural mutational processes may be identified. The recurrent integrations upstream of *Csf2*, the gene which encodes GM-CSF, are likely to be one such example. In the GRL model presented here, *Csf2* integrations were almost universally in the forward orientation suggesting these are activating mutations. Furthermore, on serial transplantation they typically persisted as opposed to the vast majority of integrations, suggesting they have a driver role in these tumours. Similarly, in the published GRH model the *Csf2* integrations were universally in the forward orientation and resulted in marked overexpression of *Csf2* mRNA and increased GM-CSF levels in leukaemia cell supernatants(Vassiliou et al., 2011).

GM-CSF is a cytokine that regulates the proliferation, differentiation, survival and functional activation of myeloid cells by binding its receptor and activating downstream signalling pathways including JAK-STAT, PI3K and RAS/MAPK(Javadi et al., 2013). *CSF2* has not been identified as a recurrent mutation target in human AML and in animal models sustained elevations in GM-CSF lead to granulocyte and macrophage hyperplasia, but not leukaemia(Johnson et al., 1989; Lang et al., 1987). However, GM-CSF stimulation is required for the in vitro proliferation of the majority of leukaemic cells from human chronic and acute myeloid leukaemia and mouse leukaemia (Metcalf, 2013; Metcalf et al., 2013). Furthermore, in some cases of AML that proliferate autonomously in vitro, GM-CSF is produced endogenously by leukaemia blasts (Bradbury et al., 1992; Young and Griffin, 1986) and in others, insertion or activation of GM-CSF or IL-3 in cell lines transforms these into leukaemic populations(Metcalf, 2013). Myelomonocytic leukaemia cells from mice generate both GM-CSF dependent and independent progeny and move between autonomous and factor dependent states (Metcalf et al., 2013). Recently secretion of growth-arrest specific gene 6 (Gas6), the ligand for Axl, a TAM family receptor tyrosine kinase, was found to be secreted by bone marrow stromal cells in response to AML mediated M-CSF(Ben-Batalla et al., 2013). Gas6 promotes tumour cell proliferation and survival in vitro and together with Axl upregulation it induces chemoresistance of AML cells. This positive feedback loop is being investigated as a therapeutic target in AML and it is possible that GM-CSF mediates a non-cell autonomous effect

through tumour-stromal cell feedback loops akin to those described for M-CSF (Ben-Batalla et al., 2013).

Together the studies detailed above suggest that although *CSF2* is not recurrently mutated in human AML it has a role in leukaemogenesis and may provide a therapeutic target. The absence of mutations in human disease may be because *CSF2* is relatively 'protected' from mutation or because naturally occurring mutations result in decreased cell viability which for example, may occur due to the disruption of important nearby genes such as *IL-3* in addition to *CSF2*. As Csf2 is a ligand, rather than coding mutations, other forms of transformation such as translocation would be needed to cause its overexpression. The activation of *Csf2* in the mouse model probably mimics the biological effect of human mutations in related pathways. For example *CBL* mutations, were recently found to enhance GM-CSF signalling (Javadi et al., 2013).

Heterozygous germline mutations in the neurofibromatosis-1 (*NF1*) gene are found in the autosomal disorder neurofibromatosis type 1 and children with this condition have an increased risk of juvenile myelomonocytic leukaemia (JMML) which may progress to AML. In these cases mutation or loss of the remaining wild-type *NF1* allele is characteristic, consistent with a tumour suppressor function of *NF1*. In human AML mono-allelic deletion of NF1 is reported in around 5% of cases, but mutation in the remaining allele is not identified in a large proportion of cases and the impact of mono-allelic loss is not fully elucidated(Haferlach et al., 2012; Parkin et al., 2010). NF-1 is a negative regulator of *RAS* signalling but mono-allelic *NF-1* loss has been found to co-occur with activating *RAS* mutations and it is thought these mutations can co-operate to up-regulate *RAS* signalling(Haferlach et al., 2010). In mice, the simultaneous expression of *K-Ras*^{G12D} and inactivation of *Nf1* in haematopoietic cells results in AML (Cutts et al., 2009), whereas either mutation alone results in a myeloproliferative disease(Braun et al., 2004; Le et al., 2004). Furthermore, bone marrow and spleen cells from mice with somatic inactivation of *Nf1*also show hypersensitivity to GM-CSF on in vitro culture(Le et al., 2004) and the absence of GM-CSF attenuates the MPD that arises in recipient mice on adoptive transfer of *Nf1*$^{-/-}$ fetal liver cells(Birnbaum et al., 2000). Hypersensitivity to GM-CSF is a feature of JMML.

In this mouse model the integrations within *Nf1* are in both orientations, suggesting these are inactivating mutations and there are frequently multiple *Nf1* hits in a single tumour. In the absence of single cell analysis it is difficult to be certain if these integrations are occurring in both *Nf1* alleles. The likelihood is that the majority represent local hopping events, which have a similar effect to the initial mutation.

Although coding mutations in the Nucleoporin 98 gene (*NUP98*) are not described in AML, *NUP98* is involved in structural chromosomal re-arrangements in a wide variety of haematopoietic malignancies including AML, MDS and T-ALL. These translocations are more common in myeloid malignancies, in which they are associated with poor prognosis. The fusion protein usually retains the amino terminus of NUP98 (Gough et al., 2011). Several fusion partner genes have been described and around half encode homeodomain proteins. The NUP98 protein is a structural component of the multi-protein nuclear pore complex that traverses the nuclear membrane, but it is also found diffusely throughout the nucleus (although not in the nucleolus). NUP98 binds CREB-binding protein (CBP) and it is thought the amino terminal portion of NUP98 has a role in active transcription as has been demonstrated in Drosophila cells(Gough et al., 2011). Leukaemogenesis is dependent on the GLFG repeats that recruit the transcriptional coactivator complex CBP/p300(Gough et al., 2011). The NUP98 fusion proteins are predominantly located in the nucleus as opposed to the wild-type protein which is mainly in the nuclear pore and these are thought to act primarily as aberrant transcriptional regulators (Gough et al., 2011).

In our mouse models the integrations in *Nup98* are bi-directional, suggesting these are inactivating integrations and they are spread through multiple introns of this gene. It is possible these integrations are affecting the role of Nup98 in transcription regulation, but this may not be the mechanism of action. As part of the nuclear pore complex NUP98 has a role in the passage of small ions and polypeptides by diffusion and the active transport of larger macroproteins across the nuclear membrane mediated by karyopherins(Gough et al., 2011). As such it is a chaperone in the transport of messenger ribonucleoprotein particles to the cytoplasm. The *Npm1cA* mutation results in cytoplasmic dislocation of the *Npm1* protein although how this leads to leukaemia is not understood. It seems plausible that the high prevalence of inactivating transposon integrations in *Nup98* in this model, could relate to its nuclear

transport function and that this may have an additive effect with *Npm1* on the mis-localisation of additional proteins or nucleic acids.

The CIS analysis is a statistical approach and as such it will not identify all driver mutations in these mice. 'Driver integrations' that occur infrequently across the cohort will not be detected by this method even if they have a powerful effect in an individual tumour. One likely example is the chromosome7:35894357 integration in 22.2b. This was one of the early integrations detected in the blood samples and persisted in all seven recipient tumours from this mouse. This integration is just 5' and in the reverse orientation to *Cebpa* (35904 312 – 35906945). *CEBPA* mutations are found in around 10% of human AML(Kihara et al., 2014; TCGA_Research_Network, 2013) and knock-in mice carrying bi-allelic *CEBPA* mutations or lacking the 42kDa CEBPA isoform develop leukaemia (Bereshchenko et al., 2009; Kihara et al., 2014). Although *Cebpa* is not identified as a CIS gene in either the GRL or GRH studies it is probable that it has a driver role in this tumour.

The reverse situation, that not all of the identified CIS represent true drivers, may also be true. Large genes or those that are actively transcribed(Liang et al., 2009) are likely to have more frequent hits irrespective of their role in leukaemogenesis. Although such hits may be found in only a small number of tumour cells, in the absence of quantitative data all integrations are treated equally in the CIS analysis. Read depth was only taken into account in the local hopping filtering. Some of the CIS identified do not contain any mapped genes or microRNA. It would be helpful to identify that these integrations are occurring in a major tumour clone before deciding to investigate them further. If the analysis approach was quantitative, integrations that were only found in a small number of cells could reasonably be excluded from the CIS analysis. This should improve the confidence that the identified genes are true driver integrations.

5. Development and Validation of a Protocol for Quantitative Analysis of Transposon Integrations

5.1 Introduction

In IM-driven cancers integrations that function as true drivers are expected to occur in a significant proportion of tumour cells. In my work, a small proportion of transposon integrations persist on serial transplantation of transposon-driven AMLs, suggesting that these contain the major drivers for leukaemogenesis. By contrast a much larger number of integrations are "lost" in leukaemias developing in AML-transplant recipients. Also, recipients of the same primary tumour can show different patterns of transposon integrations and occasionally even 'driver' integrations are "lost" in recipient tumours. These observations provide evidence that these IM-driven tumours may contain more than one clone capable of leukaemogenesis.

A major limitation of the conventional transposon-sequencing approach used in the previous chapter is that the read depth does not correlate with the number of cells in the tumour which carry a particular integration. It was previously reported that on restriction-based splinkerette analysis of tumour samples, an average of between 100 and 150 *SB* insertions were detected in each tumour, of which 50-80% are represented by a single sequence read (Dupuy et al 2009). Furthermore, the ability to amplify transposon integrations is dependent on there being a nearby restriction site and it is possible that important integrations are underrepresented or even missed simply because there is no restriction site in close proximity. A DNA shearing approach should overcome this problem and reduce the PCR amplification bias. A method for *transposon direct insert sequencing* (TraDIS) had previously been developed for bacterial genomes by the Sequencing Research and Development Team at the Wellcome Trust Sanger Institute (Langridge et al., 2009). I worked closely with them to adapt this method for insertional mutagenesis of mammalian cells. The team used AML samples from my *Npm1*^{cA} insertional mutagenesis study to adapt the protocol for mapping *Sleeping Beauty* integrations in mouse tumours. I was involved in troubleshooting of experiments and analysis of results.

5.2 Results

5.2.1 The TraDIS Illumina Sequencing Protocol Generates High Coverage and Quantitative Data

The TraDIS protocol gives high sequencing coverage when 96 samples are pooled and sequenced on a single MiSeq run for each end of the transposon. After filtering as described in Methods, including removal of PCR duplicates, there was an average of approximately 27000 reads per barcoded sample obtained from the first 96-well plate analysed. As with the 454 sequencing protocol, integrations were mapped from both ends of the *SB* transposon in two independent experiments. The reproducibility of the data from these two experiments was used to decipher how quantitative the TraDIS protocol is. The identity of the 'top' hits ranked by read number correlated well between the two experiments, as did the 5' and 3' read proportions for the majority of these hits (figure 5.1). Only 414 of the 475 integrations were used for this analysis as the others were only captured from one end of the transposon.

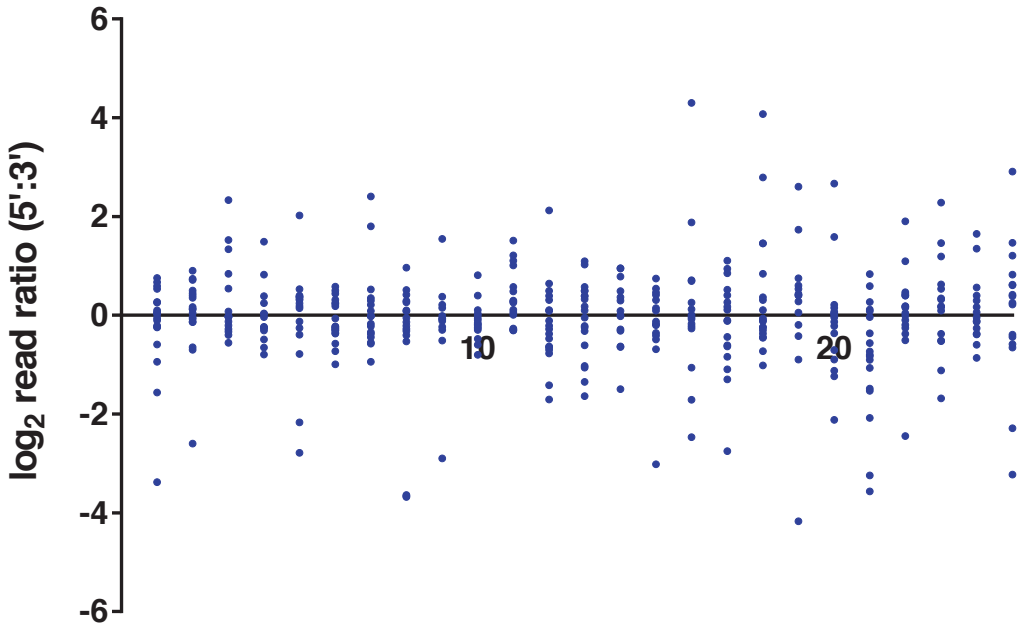


Figure 5.1: Correlation of 5' and 3' reads. The 5' to 3' ratio for the 25 integrations with highest coverage in each sample after removal of duplicates are shown for the leukaemias from 19 IM mice in the serial bleed study (chapter 4). The $\log_2$ of the ratio of the 5' to 3' reads is shown. Each blue dot represents the read ratio for the correspondingly ranked hit from one leukaemia.

Typically at least 1000 reads were obtained for the integration with the highest coverage. The number of reads per integration fell away sharply after the first few integrations in most cases. Often this occurred in a 'step-wise' manner, where several integrations had similar coverage and then there was a fall from a top tier to the next tier of integrations (figure 5.2).

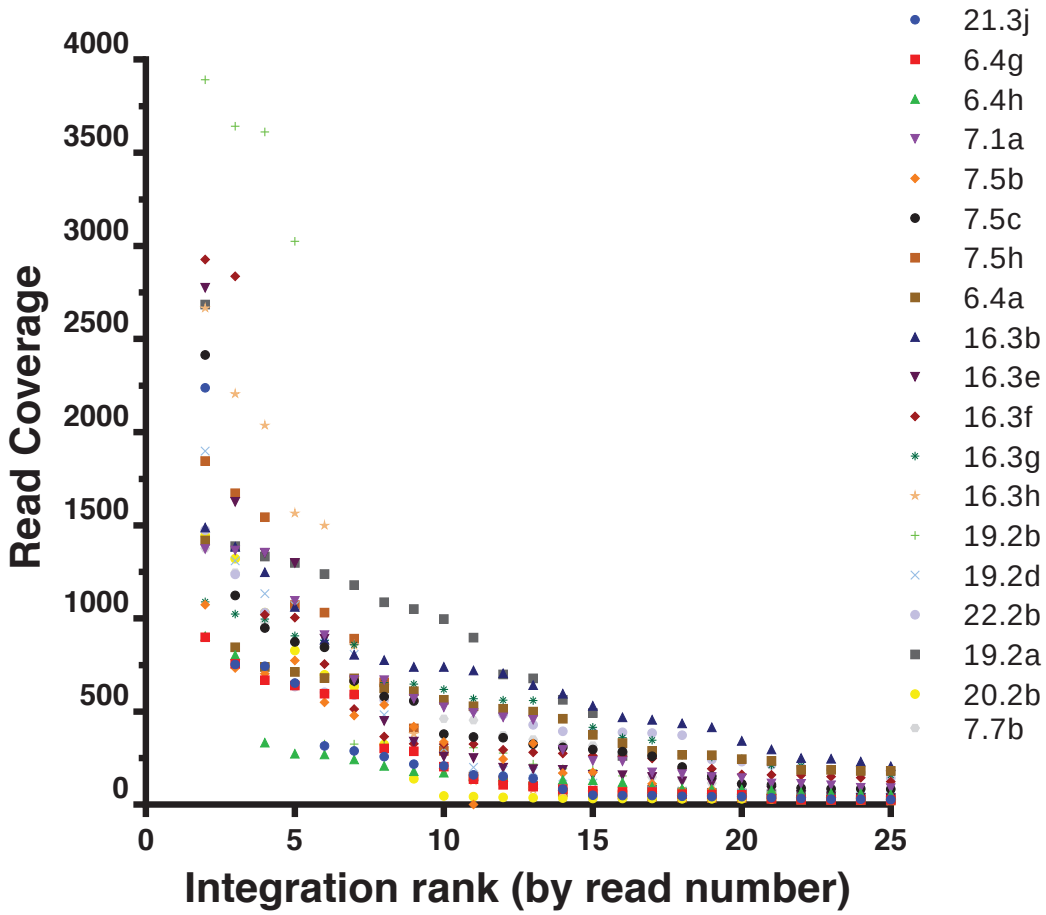


Figure 5.2: Number of reads per integration. Data is shown for the top 25 integrations by read number in the leukaemias from 19 serially bled mice (Chapter 4) after removal of PCR duplicates in the analysis.

5.2.2 TraDIS Identifies Additional CIS Compared to Restriction-Based Mapping

The set of 46 *Npm1*^{cA} *GRL* IM tumours presented in the previous chapter were analysed using the TraDIS approach and CIMPL analysis was performed using the in-built local hopping filter. After duplicate removal, all integrations with two or more reads were included in the initial 'all reads' analysis. This analysis required a massive amount of computing power and the CIMPL analysis repeatedly failed for small kernel widths, probably as a consequence of the quantity of data. As a result,

data sets for kernel windows of 40000bp or less in size were incomplete. Even so, over 100 CIS were identified for this cohort (appendix 5A). It is probable that not all of these CIS represent true driver integrations as a large number of integrations occurred at low read number in each of these tumours.

The CIS analysis on the TraDIS/Illumina data was therefore repeated using various thresholds of the number of integrations to be included from each sample. The integrations were ranked by read number and the top 10, top 25 and top 100 integrations were used for analyses. The number of CIS identified increased as the number of included integrations increased, but generally the most frequently hit sites were detected by all three analyses (table 5.1, 5.2, 5.3, 5.4 and figure 5.3). All of the CIS identified in multiple kernel scales using the top 10 hits were also detected using 25 or 100 integrations, and some of the integrations excluded from the final 'top 10' CIS list because they were only observed at one kernel scale were also identified with lower thresholds. Of note, the integrations upstream of *Csf2* (*Gm12223*) and within *Nf1* are the most frequent, regardless of the threshold. The CIS which were excluded from the final list in the analysis using the top 10 integrations are shown in table 5.1. The excluded CIS and the reasons for their exclusion are shown in appendix 5b for the 25 and 100 integration analyses.

The TraDIS/Illumina analysis identified several additional CIS that were not detected on analysis of the Splinkerette/454 data (figure 5.4). These included some genes, such as *Ets1, Pik3r5* and *Rasgrp1* that were identified on all Illumina analyses thresholds. Overall, *Ets1* integrations were detected in 11 spleen samples using the TraDIS protocol. All were in intron 1 and nine were in the forward and three in the reverse orientation (one sample had integrations mapping in both orientations). In three tumours *Ets1* was in the top 10 hits and it accounted for between 1 and 15% of reads in these mice. Review of the 454 data revealed that an *Ets1* integration was detected in only one of these three cases. In the other two, *Mbo1* restriction sites were present within 201 bases of one end of the transposon and it is therefore surprising that these integrations were not detected on 454 sequencing. *Pik3r5* integrations were detected by 454 sequencing in 7.2i, 16.3f and 19.1i however in 19.1i these sequences failed quality filtering. Both of the tumours with top 10 hits in *Rasgrp1* by TraDIS analysis were also found to have this integration on 454 analysis, but this did not reach significance as a CIS.

Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	Start	End	Number of Insertions	Number of tumours with hits within CIS	Smallest p Value	Kernel scales	Genes - largest CIS	Genes - smallest CIS	Gene Nearest Peak
11	54253024	54258966	22.09-25.97	5.4E+07	5.4E+07	29	25	0	10-100	[illegible] 4933405E24Rik Gm12222 Csf2 Gm12223 Il3 Acsl6 Gm12224 4930404A10Rik Gm12226 Gm12225 [illegible]	Csf2 Gm12223 Il3	Gm12223
11	79443113	79559960	4.09-7.33	7.9E+07	8E+07	10	8	0	10-100	Nf1 Gm11198 Gm11199 AU040972 Omg Gm21975 Evi2b Evi2a [illegible]	Nf1 Gm11199 AU040972 Omg Gm21975 Evi2b [illegible]	Nf1
4	44651711	44661723	2.7-5.69	4.5E+07	4.5E+07	7	6	0	10-100	Pax5 Mir5120 Gm12462	Pax5 Gm12462	Pax5
5	147360003	147366108	6.68-6.74	1.5E+08	1.5E+08	6	5	0	10-100	2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 [illegible]	Flt3	Flt3
9	32691960	32696580	2.9-3.23	3.3E+07	3.3E+07	5	4	0	10-100	Ets1	Ets1	Ets1
7	102154092	102159377	3.36-4.16	1E+08	1E+08	4	4	0	20-100	21 Trpc2 Art5 Art1 Chrna10 Nup98 Pg	Nup98	Nup98
8	87901290	87948876	2.92-3.88	8.8E+07	8.8E+07	4	3	0	10-100	Zfp423	Zfp423	Zfp423
7	89162170	89182469	2.13-2.47	8.9E+07	8.9E+07	3	3	1.84728E-07	50-80	Tmem135	Tmem135	Tmem135
11	68415938	68421185	2.83-3.23	6.8E+07	6.8E+07	3	2	0	10-100	Ntn1 Pik3r5	Intergenic	Pik3r5
16	52113030	52140452	2.14-2.63	5.2E+07	5.2E+07	3	2	1.22378E-06	50, 90, 100	Cblb Gm23513	Cblb Gm23513	Cblb
1	53815786	53819583	1.9-1.96	5.4E+07	5.4E+07	2	2	1.44462E-09	40, 60, 80, 90	Stk17b Hecw2	Hecw2	Hecw2
2	18677534	18681013	1.99-2	1.9E+07	1.9E+07	2	2	0	20, 30, 50, 80, 100	Gm13352 Commd3 Bmi1 Gm13334 E	Commd3 Bmi1	Bmi1
2	117340093	117343198	1.98-2	1.2E+08	1.2E+08	2	2	0	10-30, 50-80, 100	Rasgrp1	Rasgrp1	Rasgrp1
6	30131012	30135930	2.21-2.33	3E+07	3E+07	2	2	0	10, 30, 70-100	Nrf1 Gm25580	Nrf1	Nrf1
7	49334156	49336083	2.1-2.11	4.9E+07	4.9E+07	2	2	0	10-40, 90	Nav2	Nav2	Nav2
9	44839321	44844715	2.16	4.5E+07	4.5E+07	2	2	0	10, 30-80	Ifi46 Tmem25 Tlc36 Mll1	Mll1	Mll1
13	95858547	95862570	2	9.6E+07	9.6E+07	2	2	0	10, 20, 50-100	Iqgap2 Gm24032	Iqgap2	Iqgap2
Excluded as single kernel scale only												
10	28468708	28550243	1.05-1.87	2.8E+07	2.9E+07	2		7.2198E-06	30	Ptprk		Ptprk
11	100841456		2.114388983	1E+08	1E+08	2		0.001177284	30	Stat5b		Stat5b
1	131261820		0.998941663	1.3E+08	1.3E+08	1		0.002589907	50	Ikbke		Ikbke
1	156686105		0.999341849	1.6E+08	1.6E+08	1		0.000501378	50	Acbd6		Acbd6
2	93624365		0.998228839	9.4E+07	9.4E+07	1		1.46105E-13	10	intergenic		Alx4
3	101726186		1.048855483	1E+08	1E+08	1		3.04714E-07	20	intergenic		Mab21l3
10	80962233		0.998051646	8.1E+07	8.1E+07	1		5.51262E-11	10	Gng7		Gng7
10	82636647		0.998401113	8.3E+07	8.3E+07	1		0.00072215	30	Tdg		Tdg
12	26325493		0.999272518	2.6E+07	2.6E+07	1		0.003789851	10	Rnf144a		Rnf144a
12	83393123		0.999481953	8.3E+07	8.3E+07	1		0.003728404	10	Dpf3		Dpf3
13	55514966		0.998067754	5.6E+07	5.6E+07	1		8.23785E-13	40	Pdlim7		Pdlim7
14	14731810		1.11	1.5E+07	1.5E+07	1		0.000907633	10	Slc4a7		Slc4a7
17	34199660		1	3.4E+07	3.4E+07	1		7.09741E-11	10	Psmb8 Gm20496		Psmb8
X	94110024		0.99890371	9.4E+07	9.4E+07	1		1.62295E-07	80	Zfx		Zfx
CIS excluded as reported as recurrent CIS across multiple screens												
5	28156079	28159326	2.06-2.22	2.8E+07	2.8E+07	2	2	3.46375E-09	30, 60-80	En2	intergenic	En2
11	3143280	3145927	2.12-2.15	3096983	3148142	2	2	0	10, 20, 40, 60, 70, 100	Pisd-ps1 Sfi1 Gm11399	Sfi1 Gm11399	Sfi1
CIS excluded as shared by samples 9.1b and 9.1d only												
4	95019018	95021985	2.2-2.31	9.5E+07	9.5E+07	2		0	10-30, 50-80	Gm12694	Gm12694	Gm12694
8	13775791	13778487	1.99920558	1.4E+07	1.4E+07	2		0	20, 30, 50-70	Cdc16 Upf3a	Cdc16	Cdc16
9	41123677	41125414	2.16	4.1E+07	4.1E+07	2		0	10-30, 50-70	Ubash3b	Ubash3b	Ubash3b
9	114908243	114911668	2.16	1.1E+08	1.1E+08	2		0	10-30, 50-70	Gpd1l	Gpd1l	Gpd1l
2	180206734	180207115	2	1.8E+08	1.8E+08	2		0	10-100	drm1 Lama5 Rps21 Mir3091 Cables2	Lama5	Lama5
18	35554457	35558064	2	3.6E+07	3.6E+07	2		0	10, 20, 60, 70, 100	239 AC121821.1 Mir1949 Snora74a M	AC121821.1 Snora74a	AC121821.1
CIS excluded as low hits on review												
18	61102640	61105241	1	6.1E+07	6.1E+07	1		0.000155719	20, 50, 80	intergenic		Csf1r

Table 5.1: CIS identified using only the top 10 integrations from each tumour. The CIS are listed in descending order according to the number of tumours with hits within the CIS boundaries. The kernel scales for which each CIS was significant are shown. The maximum and minimum peak location show the extremes in position of the kernel peak across the various scales. The variation in peak height across the scales and the maximum boundaries of the CIS across all scales are also represented. The genes included in the largest single CIS

Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	CIS Start	CIS End	Number of hits	Number of tumours	Smallest p Value	Kernel scales	Genes - largest CIS	Genes - smallest CIS	Gene Nearest Peak
11	54252824	54257097	28.8-33.4	54031991	54481302	35	29	0	10-100	Pdlim4 P4ha2 Gm12221 4933405E24Rik Gm12222 Csf2 Gm12223 Il3 Acsl6 Gm12224 4930404A10Rik Gm12226 Gm12225 Fnip1 Gm24198	Csf2 Gm12223 Il3	Gm12223
11	79417426	79513928	3.87-10.4	79261773	79642711	18	11	0	10-100	Gm9964 Nf1 Gm11198 Gm11199 AU040972 Omg Gm21975 Evi2b Evi2a Rab11fip4 Gm23293 Gm11202 Gm25867 Gm24887	Nf1 Gm11198 Gm11199 AU040972 Omg Gm21975 Evi2b Evi2a	Nf1
4	44652521	44659165	3.83-7.01	44502146	44777575	8	6	0	10-100	Pax5 Mir5120 Gm12462 Gm12463 Zcchc7	Pax5 Gm12462	Pax5
5	147360792	147366837	6.35-6.41	147221433	147465514	6	6	0	10-100	2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 Gm6054 Pan3	Flt3	Flt3
7	102152936	102163816	3.5-6.22	102029488	102272934	6	6	0	10-100	Rnf121 Trpc2 Art5 Art1 Chrna10 Nup98 Pgap2 Rhog Stim1	Nup98	Nup98
9	44835625	44841770	5.58-6.97	44704657	44963893	7	6	0	10-100	Phldb1 Gm24166 Arcn1 Ift46 Tmem25 Ttc36 Mll1 Gm26249 Atp5l Ube4a	Mll1	Mll1
11	100828696	100845869	4.3-6.64	100691949	100947329	10	6	0	10-100	Zfp385c Gm11547 Dhx58 Kat2a Hspb9 Rab5c Kcnh4 Hcrt Ghdc Gm24358 Stat5b Stat5a Stat3	Stat5b	Stat5b
8	87901290	87955161	3.59-4.93	87807807	88043130	5	5	0	10-100	Zfp423	Zfp423	Zfp423
4	32365242	32390811	2.56-3.93	32229851	32461951	4	4	0	10-100	Bach2 D130062J21Rik Gm11932	Bach2	Bach2
8	10852960	10865434	2.28-3.83	10760518	10954694	4	4	0	80	Rps16-ps3 3930402G23Rik	intergenic	Rps16-ps3
9	32696057	32703004	3.65-4.65	32587788	32784678	8	4	0	10-100	Ets1	Ets1	Ets1
11	68417614	68421441	3.31-4.7	68322030	68497937	4	4	0	10-100	Ntn1 Gm25251 Pik3r5	intergenic	Pik3r5
1	195020274	195027633	2.08-2.98	194945019	195084049	3	3	0	10-90	Cd34 AC162692.1 mmu-mir-29b-2 Mir29b-2 Mir29c Cd46	mmu-mir-29b-2	mmu-mir-29b-2
6	103646278	103652356	3.39-3.57	103559246	103714384	4	3	0	10-70, 90, 100	Chl1 Gm24784	Chl1	Chl1
8	95761257	95768202	2.65-2.92	95688021	95824989	3	3	0	50-90	Ndrg4 Setd6 Cnot1 Gm26493 Gm26265 4930513N10Rik	Cnot1 Gm26493 Gm26265	Cnot1
10	28542121	28547286	2.13-3.18	28489877	28595859	3	3	0	10, 20, 40, 60, 80	Ptprk	Ptprk	Ptprk
12	26323898	26341410	2.05-2.8	26284154	26387214	3	3	0	10-60	Rnf144a	Rnf144a	Rnf144a
2	18678913	18681688	2.06	18670031	18681688	2	2	0	10-40	Commd3 Bmi1	Bmi1	Bmi1
2	117340555	117341622	2.05-2.06	117337642	117341622	2	2	0	20, 30	Rasgrp1	Rasgrp1	Rasgrp1
3	60568733	60571084	2.13	60558371	60572463	2	2	0	10, 20, 50-70	Mbnl1	Mbnl1	Mbnl1
3	103055968	103060038	2.13	103038253	103064522	3	2	0	10, 20, 50, 60	Csde1 Nras	Csde1	Nras/Csde1
4	101198692	101200474	2.12-2.36	101172591	101207392	2	2	2.2721E-05	20, 30, 70, 90	Jak1 Gm24468	Jak1	Jak1
6	30130727	30135829	2.24-2.38	30123433	30135829	2	2	5.35127E-14	10, 60, 80	Nrf1	Nrf1	Nrf1
7	49334864	49336103	2.06-2.08	49328274	49336103	2	2	2.22045E-16	10-40	Nav2	Nav2	Nav2
7	89149517	89160578	3-3.32	89087898	89214490	2	2	0.000101861	80-100	Tmem135	Tmem135	Tmem135
7	145052698	145053068	2.06-2.08	145044927	145054911	2	2	0	10-40	intergenic	intergenic	Ccnd1
10	70102414	70107759	1.73-2.14	70070696	70134748	3	2	2.25919E-10	20, 30, 50-80, 100	Ccdc6	Ccdc6	Ccdc6
10	81488765	81495513	1.82-2.1	81396388	81581917	2	2	0	40-100	Nfic Gm16104 Celf5 Gm16105 Ncln S1pr4 Gna15 Gm16106 Gm25595 Gna11 Aes	Celf5 Ncln S1pr4 Gna15 Gm16106 Gm25595 Gna11	Ncln
12	3106831	3108524	1.99-2.08	3085438	3122595	2	2	0	10-30, 60, 80	Gm9075	intergenic	Gm9075
13	95858402	95862570	2-2.07	95844373	95866638	2	2	0	10, 20, 40, 50, 70-90	Iqgap2	Iqgap2	Iqgap2
15	3497624	3516488	2.01-2.79	3450326	3549570	4	2	1.61039E-09	30, 40, 60, 70, 90	Ghr	Ghr	Ghr
15	78495079	78497022	1.96-2.04	78459322	78532882	2	2	0	10, 50, 70, 90, 100	Tmprss6 Il2rb C1qtnf6	Il2rb	Il2rb
16	93672689	93698069	2.12-2.94	93640758	93745482	2	2	1.07982E-06	30, 50, 70-100	Gm22344 Cbr3 Dopey2	Gm22344 Cbr3 Dopey2	Cbr3
X	75239657	75249369	2.12-2.36	75212709	75276865	3	2	0	10, 20, 40, 50, 70, 80, 100	F8 Gm6039 Gm8522	F8 Gm6039 Gm8522	F8

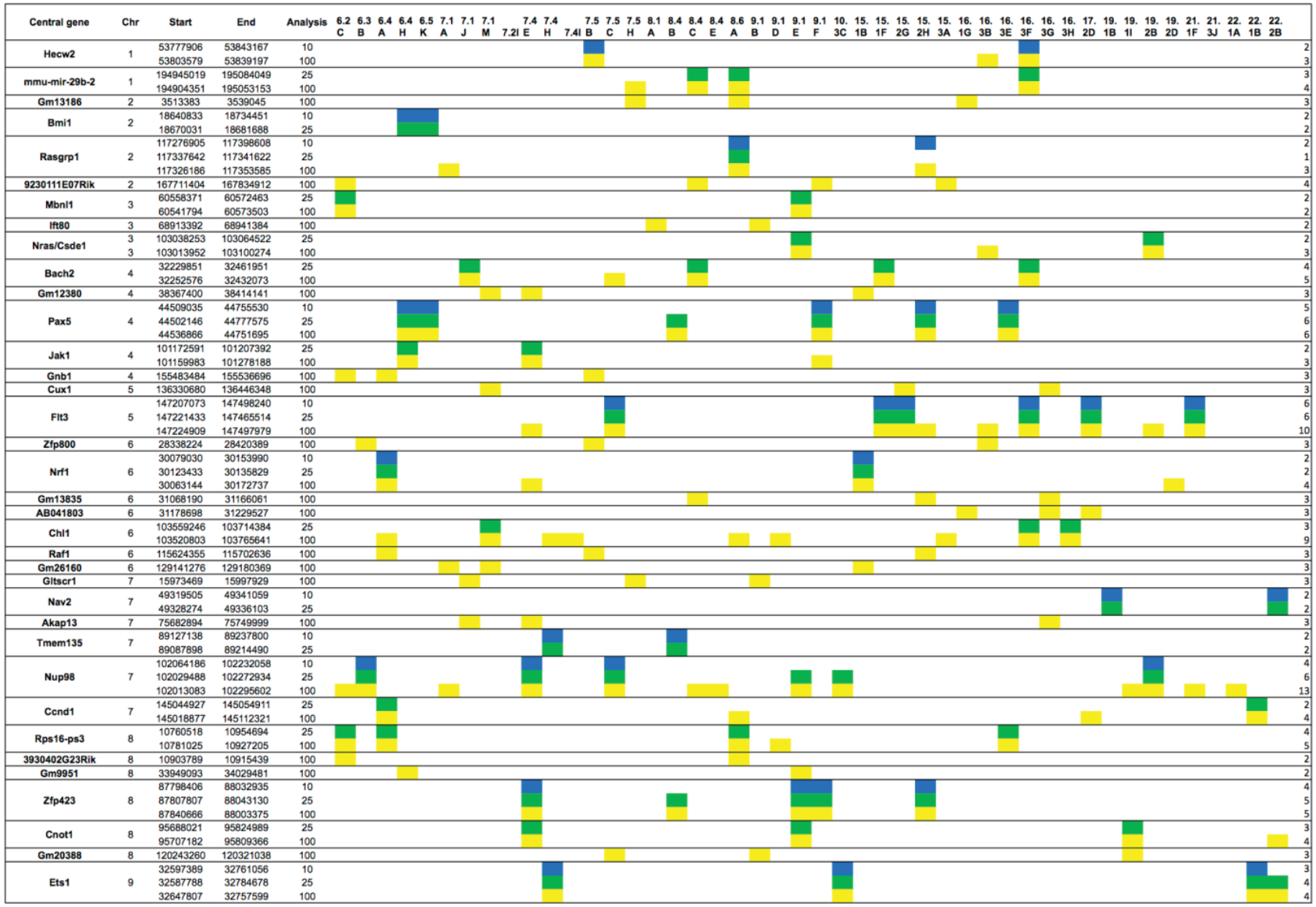
Table 5.2: CIS integrations identified with the top 25 integrations per sample. CIS that were excluded and the reason for their exclusion are shown in appendix 5b. Otherwise the features of the table are similar to table 5.1

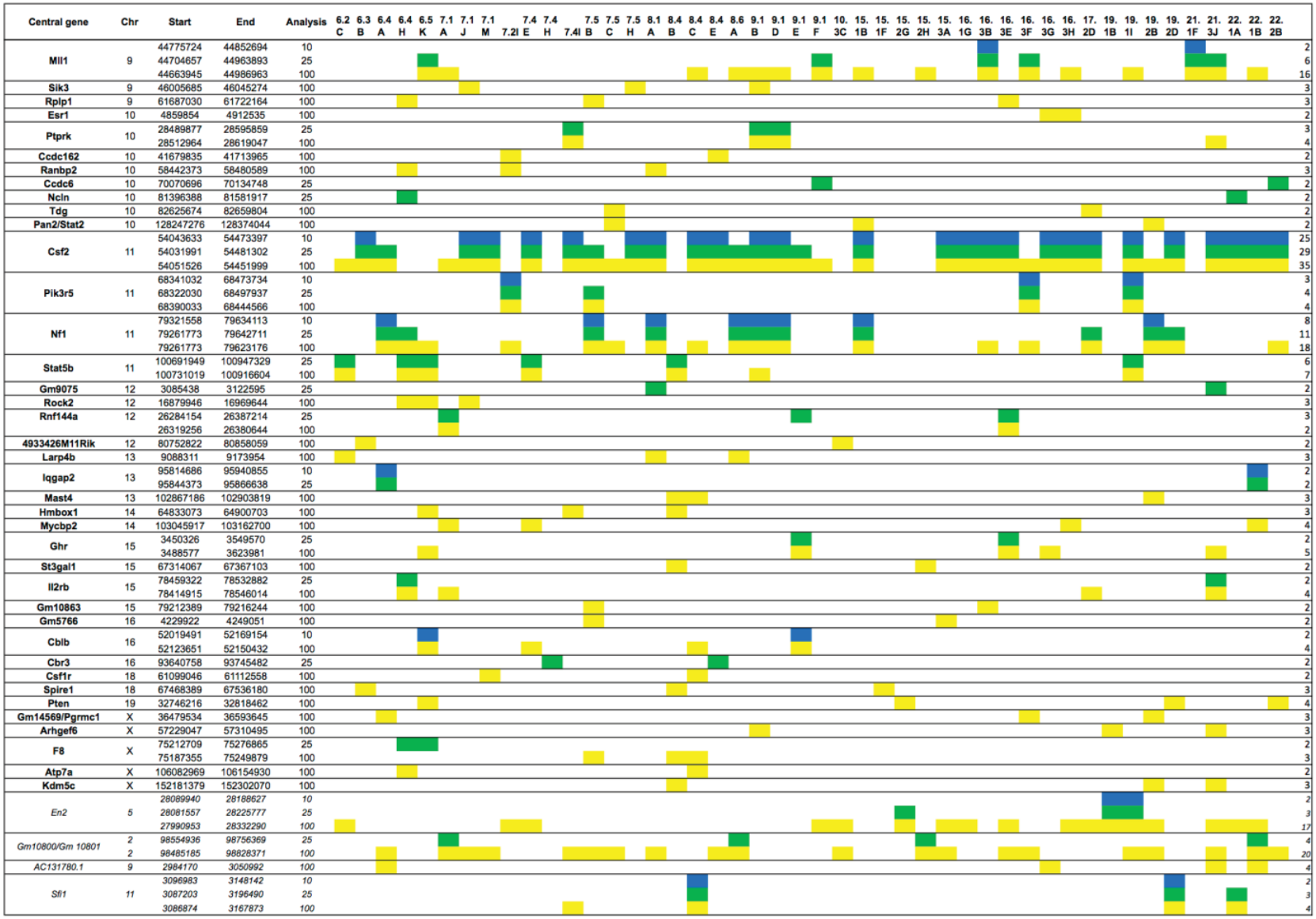
Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	CIS Start	CIS End	Number of hits	Number of tumours	Smallest p Value	Kernel scales	Genes - largest CIS	Genes - smallest CIS	Gene Nearest Peak
11	54252824	54257097	36.4-44.1	54051526	54451999	61	35	0	10-100	Pdlim4 P4ha2 Gm12221 4933405E24Rik Gm12222 Csf2 Gm12223 Il3 Acsl6 Gm12224 4930404A10Rik Gm12226 Gm12225 Fnip1	Csf2 Gm12223 Il3	Gm12223
11	79356547	79557818	12.6-18.9	79261773	79623176	55	18	0	10-100	Gm9964 Nf1 Gm11198 Gm11199 AU040972 Omg Gm21975 Evi2b Evi2a Rab11fip4 Gm23293 Gm11202	Nf1 Gm11198 Gm11199 AU040972 Omg Gm21975 Evi2b Evi2a	Nf1
9	44835970	44841958	13.16-17.48	44663945	44986963	19	16	0	10-100	Treh Phldb1 Gm24166 Arcn1 Ift46 Tmem25 Ttc36 Mll1 Gm26249 Atp5l Ube4a Cd3g Cd3d		Mll1
7	102154321	102178583	8.91-13.17	102013083	102295602	17	13	0	10-100	Numa1 Il18bp Rnf121 Trpc2 Art5 Art1 Chrna10 Nup98 Pgap2 Rhog Stim1	Art1 Chrna10 Nup98 Pgap2	Nup98
5	147361777	147371197	9.42-10.50	147224909	147497979	10	10	0	10-100	2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 Gm6054 Pan3	Flt3	Flt3
6	103647351	103650428	9.6-10.11	103520803	103765641	11	9	0	10-100	Chl1 Gm24784	Chl1	Chl1
11	100833102	100845302	5.71-8.67	100731019	100916804	19	7	0	10-100	Rab5c Kcnh4 Hcrt Ghdc Gm24358 Stat5b Stat5a Stat3	Stat5b Stat5a Stat3	Stat5b
4	44651919	44658982	3.72-6.77	44536866	44751695	9	6	0	10-100	Pax5 Mir5120 Gm12462	Pax5 Gm12462	Pax5
4	32359990	32390507	2.49-5.07	32252576	32432073	6	5	0	10, 20, 40-100	Bach2 Gm11932	Bach2	Bach2
8	10850344	10862768	2.67-5.2	10781025	10927205	8	5	0	10-100	Rps16-ps3 3930402G23Rik	intergenic	Rps16-ps3
8	87928338	87955027	2.3-5.24	87840666	88003375	5	5	0	10-100	Zfp423	Zfp423	Zfp423
15	3562591	3580782	3.14-5.59	3488577	3623981	13	5	0	10, 30-100	Ghr Gm22031	Ghr	Ghr
1	194981637	195023896	2.71-4.57	194904351	195053153	5	4	0	20-100	Cd34 AC162692.1 mmu-mir-29b-2 Mir29b-2 Mir29c Cd46	mmu-mir-29b-2 Mir29b-2 Mir29c	mmu-mir-29b-2
2	167760276	167790802	2.71-3.82	167711404	167834912	6	4	0	20-100	A530013C23Rik Gm14321 9230111E07Rik 1200007C13Rik	Gm14321 9230111E07Rik	9230111E07Rik
6	30125308	30136838	2.6-4.32	30063144	30172737	5	4	0	10, 30, 50-100	Nrf1 Gm25580 Mir182 Mir96 Mir183	Nrf1 Gm25580	Nrf1
7	145050552	145073386	2.37-5.21	145018877	145112321	6	4	5.32907E-15	10-100	intergenic	intergenic	Ccnd1
8	95758410	95766160	3.32-4.22	95707182	95809366	4	4	0	30-100	Ndrg4 Setd6 Cnot1 Gm26493 Gm26265 4930513N10Rik	Cnot1 Gm26493 Gm26265	Cnot1
9	32697818	32713572	3.54-5.13	32647807	32757599	11	4	0	10-100	Ets1	Ets1	Ets1
10	28542126	28576872	2.16-4.66	28512964	28619047	8	4	5.97081E-06	10, 30-100	Ptprk	Ptprk	Ptprk
11	68419331	68424188	4.45-5.1	68390033	68444566	5	4	0	20-60	Ntn1 Pik3r5	intergenic	Pik3r5
14	103106718	103112323	3.47-4.34	103045917	103162700	4	4	0	10-100	Irg1 Cln5 Fbxl3 Mycbp2	Fbxl3 Mycbp2	Mycbp2
15	78482617	78495392	4.11-4.78	78414915	78546014	6	4	0	10-100	Kctd17 Tmprss6 Il2rb C1qtnf6	Tmprss6 Il2rb C1qtnf6	Il2rb
19	32785873	32790504	2.79-4.4	32746216	32818462	4	4	2.08644E-05	20-80	Pten	Pten	Pten
1	53820361	53830505	2.12-3.11	53803579	53839197	3	3	7.55063E-13	10, 40-80	Hecw2	Hecw2	Hecw2
2	3531225	3533472	2.83-3.21	3513383	3539045	3	3	3.71833E-06	10, 40, 50	Cdnf Gm13186	Cdnf Gm13186	Gm13186
2	117341091	117345757	3.04-3.24	117326186	117353585	6	3	6.28628E-05	10, 30-60	Rasgrp1	Rasgrp1	Rasgrp1
3	103045366	103084560	2.09-3.49	103013952	103100274	6	3	1.52272E-05	10, 30-100	Csde1 Nras Ampd1 Gm23820	Csde1 Nras Ampd1 Gm23820	Csde1
4	38392383	38393692	3.2-3.34	38367400	38414141	3	3	1.55702E-08	30, 40, 60	intergenic	intergenic	Gm12380
4	101210052	101267740	2.32-4.33	101159983	101278188	7	3	0	10, 20, 40-100	Jak1 Gm24468 Gm12785	Jak1 Gm24468 Gm12785	Jak1
4	155510484	155517815	3.04-3.39	155483484	155536696	4	3	3.85086E-05	10, 20, 40-90	Gnb1 Gm13171	Gnb1	Gnb1
5	136398203	136429836	3-4.78	136330680	136446348	6	3	2.20049E-05	40-100	Cux1 Gm16599 A430110C17Rik	Cux1 Gm16599	Cux1
6	28371524	28390252	3.2-4.14	28338224	28420389	4	3	0	10-30, 50-90	Zfp800 Gm5303 Gcc1	Zfp800	Zfp800
6	31104629	31138673	2.17-4.26	31068190	31166061	11	3	4.09672E-13	10-40, 70-100	Gm13834 Gm13833 Gm13835	Gm13833 Gm13835 AB041803	Gm13835
6	31202144	31217858	2.32-3.06	31178698	31229527	3	3	4.99745E-06	20-40	AB041803 2210408F21Rik	AB041803	AB041803
6	115654757	115674919	2.56-3.85	115624355	115702636	5	3	8.88178E-16	10-90	Raf1 Gm14335 D830050J10Rik	Raf1 Gm14335	Raf1
6	129164416	129166894	2.69-3.29	129141276	129180369	4	3	6.57729E-10	10, 40-60	Gm26160	intergenic	Gm26160
7	15990366	15991560	3.12-3.19	15973469	15997929	3	3	0	10-40	Gltscr1	Gltscr1	Gltscr1
7	75690185	75732515	3.04-5.25	75682894	75749999	8	3	7.50217E-06	20-70, 90, 100	Akap13	Akap13	Akap13
8	120281773	122213687	2.09-3.05	120243260	120321038	3	3	2.77556E-15	20-70	Gm20388 Gse1	Gm20388	Gm20388
9	46027171	46029108	3.2-3.26	46005685	46045274	3	3	0	10-50	Sik3	Sik3	Sik3
9	61707050	61710453	2.87-3.25	61687030	61722164	3	3	0	10-50	intergenic	intergenic	Rplp1
10	58465874	58465976	3.04-3.16	58442373	58480589	3	3	2.69562E-13	30-60	Ranbp2	Ranbp2	Ranbp2
10	128272785	128316427	2.05-3.89	128247276	128374044	5	3	1.38041E-05	10, 50-100	Timeless Apon Apof Stat2 Il23a Gm23241 Pan2 Cnpy2 Gm24320 Cs Gm23182 Coq10a	Il23a Gm23241 Pan2 Cnpy2	Pan2
12	16911109	16963775	3.056952488	16879946	16969644	5	3	0	10-30, 80	Rock2	Rock2	Rock2
13	9119384	9137169	2.58-3.73	9088311	9173954	4	3	5.09354E-07	20-100	Larp4b Gm23553	Larp4b	Larp4b
13	102888265	102892192	2.43-3.03	102867186	102903819	3	3	0	10, 50-70	Mast4	Mast4	Mast4
14	64868124	64883109	3.04-3.45	64833073	64900703	6	3	0	10-40, 60-90	Hmbox1	Hmbox1	Hmbox1
18	67494037	67512056	2.06-3.69	67468389	67536180	6	3	0	10, 60-90	Slmo1 Spire1	Spire1	Spire1
X	36538800	36558268	2.08-3.11	36479534	36593645	4	3	0	10-30, 70, 90, 100	Gm14569	intergenic	Gm14569/Pgrmc1
X	57274240	57276958	3.06-3.21	57229047	57310495	3	3	0	20-100	Arhgef6	Arhgef6	Arhgef6
X	75223823	75242142	2.08-3.17	75187355	75249879	5	3	1.72497E-08	10, 20, 40, 60, 80-100	F8 Gm6039	F8 Gm6039	F8
X	152241516	152248503	2.14-3.2	152181379	152302070	3	3	0	10-100	Iqsec2 Kdm5c	Kdm5c	Kdm5c
3	60556491	60570684	2.09-2.93	60541794	60573503	3	2	5.06625E-05	10, 30, 50	Mbnl1	Mbnl1	Mbnl1
3	68915348	68929622	2.09-3.01	68913392	68941384	4	2	1.03275E-05	10, 60	Ift80	Ift80	Ift80
8	10911556	10913543	2.07-2.16	10903789	10915439	4	2	0	10, 20	intergenic	intergenic	3930402G23Rik
8	33981036	33997363	2.06-3.1	33949093	34029481	5	2	0	10-70	intergenic	intergenic	Gm9951

10	4889121	4898675	2.95-3.53	4859854	4912535	4	2	0.000115329	40-70	Esr1	Esr1	Esr1
10	41694462	41695701	2.99-3.05	41679835	41713965	3	2	0.000193599	50, 60	Ccdc162	Ccdc162	Ccdc162
10	82637274	82650052	2.08-3.09	82625674	82659804	3	2	1.46826E-06	10, 50, 60	Tdg Glt8d2	Tdg	Tdg
12	26343368	26368177	2.82-3.4	26319256	26380644	4	2	0.000184807	60-100	Rnf144a	Rnf144a	Rnf144a
12	80808583	80818725	2.95-3.76	80752822	80858059	4	2	4.46476E-05	50-100	Ccdc177 AC134537.1 4933426M11Rik	4933426M11Rik	4933426M11Rik
15	67342996	67348380	3.14-3.68	67314067	67367103	4	2	8.73651E-05	40-60	intergenic	intergenic	St3gal1
15	79215992	79216244	2.12-2.15	79212389	79216244	2	2	2.42334E-05	10, 40	Gm10863	Gm10863	Gm10863
16	4239552	4243313	3.29-3.8	4229922	4249051	9	2	4.53303E-05	20, 30	intergenic	intergenic	Gm5766
16	52137565	52139386	3.71-4.26	52123651	52150432	4	2	0	10-30	Cblb	Cblb	Cblb
18	61106767	61107511	2.01-2.06	61099046	61112558	2	2	0	10, 20	Csf1r		Csf1r
X	106126153	106136594	2.09-3.11	106082969	106154930	4	2	3.63461E-07	40-60, 90, 100	Atp7a Tlr13	Atp7a	Atp7a

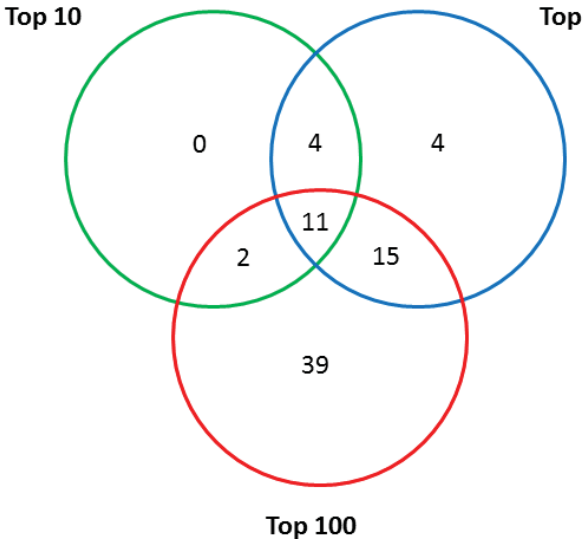
Table 5.3: CIS integrations identified with the top 100 integrations per sample. CIS that were excluded and the reason for their exclusion are shown in appendix 5b. Otherwise the features of the table are similar to table 5.1.

Table 5.4 (next page): Common integrations sites identified using the various thresholds for analysis. (Next two pages) The central gene in the CIS, maximum CIS boundaries and analysis in which the CIS were identified are shown. The tumours which had integrations within the designated CIS boundary and the number of tumours with hits within the CIS are indicated, however integrations from outside these limits also contribute to the CIS. After the analysis was completed it was noted that samples 9.1B and 9.1D gave very similar data. CIS that were based on these integrations were excluded when these were the only hits. Those with additional hits contributing to the CIS are included, but the validity of some of these CIS needs to be confirmed. Sites identified as 'false' CIS by ourselves and others are shown in italics at the bottom. CIS that were identified on one kernel scale only were excluded.


Central gene
Chr
Start
End
Analysis
6.2 C
6.3 B
6.4 A
6.4 H
6.5 K
7.1 A
7.1 J
7.1 M
7.2I
7.4 E
7.4 H
7.4I
7.5 B
7.5 C
7.5 H
8.1 A
8.4 B
8.4 C
8.4 E
8.6 A
9.1 B
9.1 D
9.1 E
9.1 F
10. 3C
15. 1B
15. 1F
15. 2G
15. 2H
15. 3A
16. 1G
16. 3B
16. 3E
16. 3F
16. 3G
16. 3H
17. 2D
19. 1B
19. 1I
19. 2B
19. 2D
21. 1F
21. 3J
22. 1A
22. 1B
22. 2B
Hecw2 1 53777906 53843167 10 2
53803579 53839197 100 3
mmu-mir-29b-2 1 194945019 195084049 25 3
194904351 195053153 100 4
Gm13186 2 3513383 3539045 100 3
Bmi1 2 18640833 18734451 10 2
18670031 18681688 25 2
Rasgrp1 2 117276905 117398608 10 2
117337642 117341622 25 1
117326186 117353585 100 3
9230111E07Rik 2 167711404 167834912 100 4
Mbnl1 3 60558371 60572463 25 2
60541794 60573503 100 2
Ift80 3 68913392 68941384 100 2
Nras/Csde1 3 103038253 103064522 25 2
3 103013952 103100274 100 3
Bach2 4 32229851 32461951 25 4
32252576 32432073 100 5
Gm12380 4 38367400 38414141 100 3
Pax5 4 44509035 44755530 10 5
44502146 44777575 25 6
44536866 44751695 100 6
Jak1 4 101172591 101207392 25 2
101159983 101278188 100 3
Gnb1 4 155483484 155536696 100 3
Cux1 5 136330680 136446348 100 3
Flt3 5 147207073 147498240 10 6
147221433 147465514 25 6
147224909 147497979 100 10
Zfp800 6 28338224 28420389 100 3
Nrf1 6 30079030 30153990 10 2
30123433 30135829 25 2
30063144 30172737 100 4
Gm13835 6 31068190 31166061 100 3
AB041803 6 31178698 31229527 100 3
Chl1 6 103559246 103714384 25 3
103520803 103765641 100 9
Raf1 6 115624355 115702636 100 3
Gm26160 6 129141276 129180369 100 3
Gltscr1 7 15973469 15997929 100 3
Nav2 7 49319505 49341059 10 2
49328274 49336103 25 2
Akap13 7 75682894 75749999 100 3
Tmem135 7 89127138 89237800 10 2
89087898 89214490 25 2
Nup98 7 102064186 102232058 10 4
102029488 102272934 25 6
102013083 102295602 100 13
Ccnd1 7 145044927 145054911 25 2
145018877 145112321 100 4
Rps16-ps3 8 10760518 10954694 25 4
10781025 10927205 100 5
3930402G23Rik 8 10903789 10915439 100 2
Gm9951 8 33949093 34029481 100 2
Zfp423 8 87798406 88032935 10 4
87807807 88043130 25 5
87840666 88003375 100 5
Cnot1 8 95688021 95824989 25 3
95707182 95809366 100 4
Gm20388 8 120243260 120321038 100 3
Ets1 9 32597389 32761056 10 3
32587788 32784678 25 4
32647807 32757599 100 4

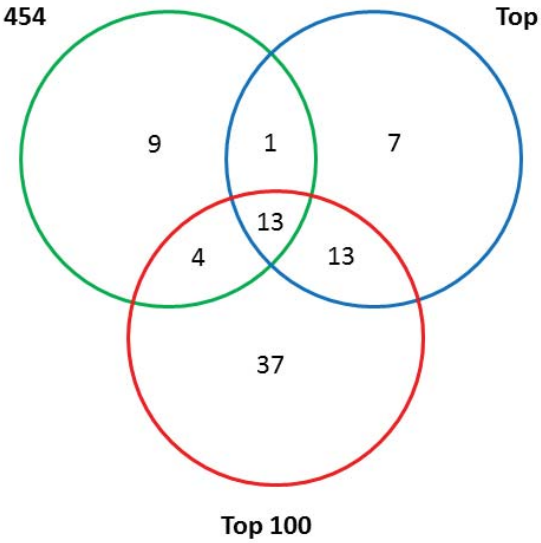


Central gene	Chr	Start	End	Analysis	
Mll1	9	44775724	44852694	10	2
		44704657	44963893	25	6
		44663945	44986963	100	16
Sik3	9	46005685	46045274	100	3
Rplp1	9	61687030	61722164	100	3
Esr1	10	4859854	4912535	100	2
Ptprk	10	28489877	28595859	25	3
		28512964	28619047	100	4
Ccdc162	10	41679835	41713965	100	2
Ranbp2	10	58442373	58480589	100	3
Ccdc6	10	70070696	70134748	25	2
Ncln	10	81396388	81581917	25	2
Tdg	10	82625674	82659804	100	2
Pan2/Stat2	10	128247276	128374044	100	2
Csf2	11	54043633	54473397	10	25
		54031991	54481302	25	29
		54051526	54451999	100	35
Pik3r5	11	68341032	68473734	10	3
		68322030	68497937	25	4
		68390033	68444566	100	4
Nf1	11	79321558	79634113	10	8
		79261773	79642711	25	11
		79261773	79623176	100	18
Stat5b	11	100691949	100947329	25	6
		100731019	100916604	100	7
Gm9075	12	3085438	3122595	25	2
Rock2	12	16879946	16969644	100	3
Rnf144a	12	26284154	26387214	25	3
		26319256	26380644	100	2
4933426M11Rik	12	80752822	80858059	100	2
Larp4b	13	9088311	9173954	100	3
Iqgap2	13	95814686	95940855	10	2
		95844373	95866638	25	2
Mast4	13	102867186	102903819	100	3
Hmbox1	14	64833073	64900703	100	3
Mycbp2	14	103045917	103162700	100	4
Ghr	15	3450326	3549570	25	2
		3488577	3623981	100	5
St3gal1	15	67314067	67367103	100	2
Il2rb	15	78459322	78532882	25	2
		78414915	78546014	100	4
Gm10863	15	79212389	79216244	100	2
Gm5766	16	4229922	4249051	100	2
Cblb	16	52019491	52169154	10	2
		52123651	52150432	100	4
Cbr3	16	93640758	93745482	25	2
Csf1r	18	61099046	61112558	100	2
Spire1	18	67468389	67536180	100	3
Pten	19	32746216	32818462	100	4
Gm14569/Pgrmc1	X	36479534	36593645	100	3
Arhgef6	X	57229047	57310495	100	3
F8	X	75212709	75276865	25	2
		75187355	75249879	100	3
Atp7a	X	106082969	106154930	100	2
Kdm5c	X	152181379	152302070	100	3
En2	*5*	*28089940*	*28188627*	*10*	*2*
		28081557	*28225777*	*25*	*3*
		27990953	*28332290*	*100*	*17*
Gm10800/Gm 10801	*2*	*98554936*	*98756369*	*25*	*4*
	2	*98485185*	*98828371*	*100*	*20*
AC131780.1	*9*	*2984170*	*3050992*	*100*	*4*
Sfi1	*11*	*3096983*	*3148142*	*10*	*2*
		3087203	*3196490*	*25*	*3*
		3086874	*3167873*	*100*	*4*



ALL	Top 10/25	Top10/100	Top 25/100
Csf2	Bmi1	Cblb	Bach2
Ets1	Iqgap2	Hecw2	Ccnd1
Flt3	nav2		Cnot1
Mll1	Tmem135		F8
Nf1			Ghr
Nrf1			Il2rb
Nup98			Jak1
Pax5			Mbnl1
Pik3r5			Nras/Csde1
Rasgrp1			Ptprk
Zfp423			Rnf144a
			Rps16-ps3
			Stat5b
			mmu-mir-29b-2

Figure 5.3: Overlapping CIS at different thresholds of the number of integrations included in the analysis.



All	454 only	454 and top 25 or top 100	Top 10 not on 454
Bach2	Phf6	Bmi1	Iqgap2
Csf2	Ptpn1	2930402G23Rik	Nav2
Flt3	Pum1	4933426M11Rik	Tmem135
Il2rb	Rps6ka2	Gm5766	Cblb
Mbnl1	Sec16a	Gnb1	Hecw2
Mll1	Tnrc6b		Ets1
Nf1	Ube2e2		Pik3r5
Nrf1	Ubn1		Rasgrp1
Nup98	Uggt1		
Pax5			
Stat5b			
Zfp423			
mmu-mir-29b-2			

Figure 5.4: Overlapping CIS integrations between the 454 and Illumina sequencing data. The integrations which were identified as CIS in the published GRH (high copy) IM cohort are indicated in red.

Some of the CIS detected on TraDIS sequencing analysis were initially detected on the 454 analysis but were removed on manual filtering. This was for various reasons including multiple hits in the same tumour (*Ghr*) and most hits mapping to the same site and occurring in the same sequencing run (*Tmem135* and *Ptprk*). The Illumina

data allows further analysis of these sites. For example, although multiple integrations in *Ghr* were mapped in sample 9.1e, there were several other samples in which reads could be mapped to *Ghr* in low number (table 5.5). However, there was only one tumour (6.5k) in which over 5% of reads mapped to *Ghr*. Two tumours had *Tmem135* integrations at different sites in their top 10 hits, which suggests that this integration may have a driver role, although not all of the top hits are necessarily drivers (some are likely to be passengers acquired in a cell prior to acquisition of the first or subsequent driver).

Tumour ID	Chromosome	Integration Site	Read coverage 3'	Read coverage 5'	Read Coverage	Proportion of total reads (%)
21.3j	15	3494201	0	14	14	0.007
7.4i	15	3373368	35	0	35	0.013
	15	3529909	22	0	22	0.008
8.4e	15	3415087	4	0	4	0.006
7.4e	15	3411065	0	3	3	0.006
	15	3494216	0	3	3	0.006
	15	3501723	0	3	3	0.006
15.2h	15	3465435	2	0	2	0.001
	15	3576758	4	0	4	0.002
	15	3577277	2	0	2	0.001
	15	3416879	4	0	4	0.007
6.3b	15	3330447	0	6	6	0.004
7.5h	15	3461781	0	3	3	0.003
	15	3494198	0	5	5	0.005
7.4h	15	3458490	0	57	57	0.065
6.2c	15	3434477	23	7	30	0.021
	15	3486515	0	13	13	0.008
9.1d	15	3489821	2	0	2	0.001
	15	3498749	3	0	3	0.002
	15	3581169	3	0	3	0.002
16.3e	15	3488756	76	84	160	0.831
8.6a	15	3354121	3	0	3	0.002
	15	3475237	0	4	4	0.003
6.5k	15	3577266	7991	3878	11869	7.337
16.3g	15	3573269	14	37	51	0.200
22.1b	15	3385054	8	7	15	0.040
7.2l	15	3473658	0	2	2	0.002
6.4a	15	3533456	0	2	2	0.007
9.1e	15	3462886	4	0	4	0.002
	15	3463392	12	30	42	0.021
	15	3463839	125	22	147	0.072
	15	3464889	8	0	8	0.004
	15	3466431	0	68	68	0.035
	15	3467843	0	34	34	0.018
	15	3468753	6	0	6	0.003
	15	3473164	583	361	944	0.468
	15	3484431	6	0	6	0.003
	15	3494215	474	1009	1483	0.753
	15	3501724	799	1001	1800	0.905
	15	3510821	0	18	18	0.009
	15	3531525	0	12	12	0.006
	15	3581145	927	112	1039	0.504

Table 5.5. Integrations in the *Ghr* locus. All of the primary tumour samples in which 2 or more reads (after PCR duplicate removal) were mapped to this locus are shown. The samples in which this was a top 100 hit are shaded. Also note the correlation between 5' and 3' reads is poor at low read number.

The observation of local hopping within a CIS was not unique to the *Ghr* locus. In fact, it was typical to see some evidence of local hopping around major integrations. As an example, the hits immediately upstream of *Csf2* in spleen samples for twelve of the mice which were serially bled are shown in table 5.6.

Mouse	Integration site	Orientation relative to *Csf2*	3' reads	5' reads	Total reads	Proportion of total reads (%)
21.3j	54250980	Forward	9	10	19	0.091
	54252890	Forward	1323	915	2238	10.605
6.4g	54254757	Forward	0	3	3	0.029
	54269566	Forward	2	0	2	0.016
19.2d	54250978	Forward	13	17	30	0.127
	54251445	Forward	2	0	2	0.009
	54253305	Forward	84	102	186	0.786
	54254757	Forward	2	5	7	0.029
	54268794	Forward	4	2	6	0.026
16.3h	54250118	Forward	0	3	3	0.012
	54250980	Forward	3	5	8	0.032
	54252781	Forward	1182	1023	2205	8.877
6.4a	54250117	Forward	1114	1032	2146	8.720
	54269567	Forward	2	3	5	0.020
16.3b	54250979	Forward	3	0	3	0.007
	54251445	Forward	414	326	740	1.647
	54254597	Forward	12	11	23	0.051
16.3f	54250979	Forward	58	45	103	0.371
16.3g	54250979	Forward	6	4	10	0.041
	54252778	Forward	437	444	881	3.553
	54269566	Forward	21	18	39	0.158
	54272909	Forward	16	6	22	0.091
19.2b	54252119	Forward	7	2	9	0.032
	54254757	Forward	0	2	2	0.007
	54269563	Forward	0	2	2	0.007
22.2b	54251894	Forward	263	325	588	1.580
	54252890	Forward	3	0	3	0.008
6.4a	54250118	Forward	252	263	515	1.766
	54252891	Forward	0	2	2	0.007
7.5b	54250591	Forward	3	0	3	0.010
	54250979	Forward	82	64	146	0.507
	54254598	Forward	244	302	546	1.895
	54254757	Forward	2	7	9	0.031

Table 5.6: Integrations upstream of Csf2 in 12 of the serially bled mice. Multiple integrations at this locus were detected in some, but not all of these tumours. Read counts and proportions are shown for duplicate filtered data.

5.2.3 PCR duplicate removal decreases the proportion of reads attributed to the top hits but does not significantly alter ranking of integration sites

The number of unique positions at which shearing of genomic DNA could result in successful capture of an integration by subsequent PCR is limited to a few hundred bases either side of the transposon. If the major integrations are common to the majority of cells in a tumour sample, then the number of unique reads could be limited by the number of possible shear sites. In other words, shearing will lead to cutting of the genome at exactly the same position in independent DNA fragments and this can appear as a PCR duplicate. In this instance, the true clonal representation of the major integrations may be underestimated by analysis of duplicate-filtered data. To investigate this, some of the Illumina sequencing was also analysed without removal of duplicate reads.

In the plate of samples presented above in 5.2.1 there was a mean of 138781 reads per barcode, with 70244 reads from the 3' and 68537 reads from the 5' end before removal of the PCR duplicates. Therefore, the removal of PCR duplicates resulted in a five-fold reduction in read number at both ends of the transposon. Typically over 5000 reads were obtained for the integration with the highest coverage in the non-duplicate filtered data (figure 5.5). There were only minor changes in the rank order of the top integrations (table 5.7). In most (e.g. 16.3f, 19.2b), but not all samples (e.g.16.3e), the proportion of reads taken by the top few integrations was higher when duplicate reads were included in the analysis (table 5.7).

In the unfiltered data there was still good correlation between the ratio of reads from the 5' and 3' ends of the transposon for the top integrations where both ends were mapped, particularly for the top ten hits (figure 5.6). There was an issue with the read correlation in both duplicate and non-duplicate filtered data sets in that around 1 in every 10 of the top integrations were only mapped to one end of the transposon. As these integrations did not return a read ratio they were not evident in figures 5.1 and 5.6. Although in some instances there was only data from one end of the transposon, in others the hit was mapped at both ends, but failed final pooling into pairs on the analysis. This seems to have occurred because amongst the thousands of aligned reads for that site, there were a handful of reads that were very long and looked aberrant. The integration site was excluded in the processing because of

these suspicious overlapping reads, even though the vast majority of reads at the same site looked real.

16.3e with duplicates						16.3e no duplicates					
Chr	Integration Site (base position)	Read Coverage	3' read coverage	5' read coverage	Proportion of total reads (%)	Chr	Integration Site (base position)	Read Coverage	3' read coverage	5' read coverage	Proportion of total reads (%)
5	96947849	12268	6162	6106	13.67	5	96947849	2775	2191	584	14.13
7	114187908	6234	2577	3657	6.95	7	106954971	1625	1625	0	8.28
10	88768122	5683	3022	2661	6.33	7	114187908	1352	414	938	6.89
11	54251450	3698	2438	1260	4.12	10	88768122	1298	404	894	6.61
7	106954971	3186	3186	0	3.55	11	23308832	892	212	680	4.54
11	23308832	2931	1544	1387	3.27	11	54251450	662	396	266	3.37
11	19935480	2307	1214	1093	2.57	8	10863348	450	87	363	2.29
4	44675886	1464	807	657	1.63	11	19935480	338	207	131	1.72
8	10863348	1460	607	853	1.63	19	11989275	258	78	180	1.31
15	19543899	1458	589	869	1.62	4	44675886	251	123	128	1.28
18	13985002	1431	844	587	1.59	18	13985002	198	107	91	1.01
1	80626479	1371	718	653	1.53	15	19543899	190	90	100	0.97
15	3488755	1262	726	536	1.41	4	59642885	188	89	99	0.96
4	59642885	1096	505	591	1.22	15	3488755	160	76	84	0.81
19	11989275	920	486	434	1.03	1	80626479	159	85	74	0.81
13	101689856	914	10	904	1.02	13	101689856	149	9	140	0.76
16	9924050	736	393	343	0.82	7	143522682	126	92	34	0.64
7	143522682	671	373	298	0.75	16	9924050	117	58	59	0.60
9	75191210	536	277	259	0.60	9	75191210	90	44	46	0.46
16	8647666	237	76	161	0.26	16	8647666	74	22	52	0.38
17	13001835	193	180	13	0.22	17	13001835	52	48	4	0.26
12	**26322697**	191	146	45	0.21	**9**	**61702075**	46	22	24	0.23
16	37872462	188	99	89	0.21	**10**	**14189688**	38	19	19	0.19
7	**83819908**	187	47	140	0.21	**15**	**4210806**	36	0	36	0.18
X	**169396799**	185	85	100	0.21	16	37872462	33	15	18	0.17

16.3f with duplicates						16.3f no duplicates					
Chr	Integration Site (base position)	Read Coverage	3' read coverage	5' read coverage	Proportion of total reads (%)	Chr	Integration Site (base position)	Read Coverage	3' read coverage	5' read coverage	Proportion of total reads (%)
1	195006589	22321	11335	10986	17.47	11	68423465	2927	1326	1601	10.38
11	68423465	21776	9531	12245	17.05	1	195006589	2837	1629	1208	10.06
14	***21998733***	4255	4255	0	3.33	16	33497860	1020	259	761	3.62
16	**52750011**	2901	136	2765	2.27	14	21998898	1004	0	1004	3.56
16	33497860	2804	1195	1609	2.20	3	30190155	755	325	430	2.68
1	53806440	2435	1336	1099	1.91	1	53806440	514	302	212	1.82
3	30190155	2140	829	1311	1.68	5	147365882	366	204	162	1.30
14	21998898	2089	0	2089	1.64	*6*	*103649266*	328	300	28	1.16
6	*103649149*	2030	2030	0	1.59	17	69679119	326	0	326	1.16
5	147365882	1451	607	844	1.14	4	3730090	325	177	148	1.15
4	32392357	1415	733	682	1.11	4	14790887	294	68	226	1.04
4	3730090	1333	579	754	1.04	4	32392357	282	144	138	1.00
4	8591429	1331	676	655	1.04	3	132797213	276	138	138	0.98
17	69679119	1168	0	1168	0.91	4	8591429	264	123	141	0.94
13	46673640	990	376	614	0.78	14	103701736	260	100	160	0.92
19	21418798	920	343	577	0.72	13	46673640	248	86	162	0.88
14	103701736	899	530	369	0.70	4	14861952	195	94	101	0.69
4	14861952	896	443	453	0.70	9	44841823	192	91	101	0.68
9	44841823	813	361	452	0.64	16	29806260	163	0	163	0.58
16	24923843	798	432	366	0.62	19	21418798	160	72	88	0.57
16	29806260	776	0	776	0.61	16	24923843	158	93	65	0.56
3	132797213	697	222	475	0.55	1	77218988	151	83	68	0.54
1	77218988	635	326	309	0.50	17	49029188	144	76	68	0.51
17	49029188	628	313	315	0.49	**16**	**4256175**	124	0	124	0.44
4	14790887	520	116	404	0.41	**11**	**54250979**	103	58	45	0.37

19.2b with duplicates						19.2b no duplicates					
Chr	Integration Site (base position)	Read Coverage	3' read coverage	5' read coverage	Proportion of total reads (%)	Chr	Integration Site (base position)	Read Coverage	3' read coverage	5' read coverage	Proportion of total reads (%)
7	102152650	36791	16789	20002	20.06	11	79558613	3891	1555	2336	13.71
5	62721650	29587	14950	14637	16.13	5	62721650	3642	1566	2076	12.83
10	122441998	27960	14805	13155	15.25	7	102152650	3612	1949	1663	12.73
11	79558613	25321	11418	13903	13.81	10	122441998	3024	1740	1284	10.66
9	89969596	2063	808	1255	1.12	9	89969596	325	139	186	1.15
14	14732190	1891	914	977	1.03	X	94113041	325	175	150	1.15
X	94113041	1827	933	894	1.00	8	70790441	324	191	133	1.14
4	6219875	1714	806	908	0.93	7	27240784	323	144	179	1.14
8	70790441	1658	799	859	0.90	14	14732190	320	158	162	1.13
7	27240784	1629	647	982	0.89	4	6219875	306	163	143	1.08
4	97975213	1506	604	902	0.82	4	97975213	277	109	168	0.98
1	86683437	1235	441	794	0.67	1	86683437	217	83	134	0.76
X	70339543	1079	419	660	0.59	X	70339543	193	84	109	0.68
14	16024808	858	388	470	0.47	3	103057430	187	112	75	0.66
3	103057430	697	237	460	0.38	14	16024808	154	79	75	0.54
X	152259929	662	313	349	0.36	X	152259929	113	57	56	0.40
19	4666291	471	220	251	0.26	19	16925277	90	51	39	0.32
19	16925277	464	257	207	0.25	19	4666291	84	43	41	0.30
4	145341339	421	0	421	0.23	4	145341339	73	0	73	0.26
5	41669778	331	331	0	0.18	5	41669778	70	70	0	0.25
10	74372435	319	173	146	0.17	**4**	**145341264**	68	68	0	0.24
X	36558250	319	138	181	0.17	X	36558250	63	32	31	0.22
11	**79418213**	284	111	173	0.15	10	74372435	57	33	24	0.20
4	**145341417**	227	227	0	0.12	14	81786706	46	0	46	0.16
14	81786706	205	0	205	0.11	**3**	**103057616**	44	5	39	0.16

Table 5.7. Comparison of duplicate filtered and non-filtered data sets from three primary tumours. The top 25 integrations are shown for each. Integrations a coloured by rank in the 'with duplicates' data for easier visualisation of the corresponding integrations in the 'no duplicates' data; red=top 5, blue = 6-10, green = 11-15, purple = 16-20, black= 21-25. Integrations sites that are not in the top 25 hits in both data sets are shown in bold.

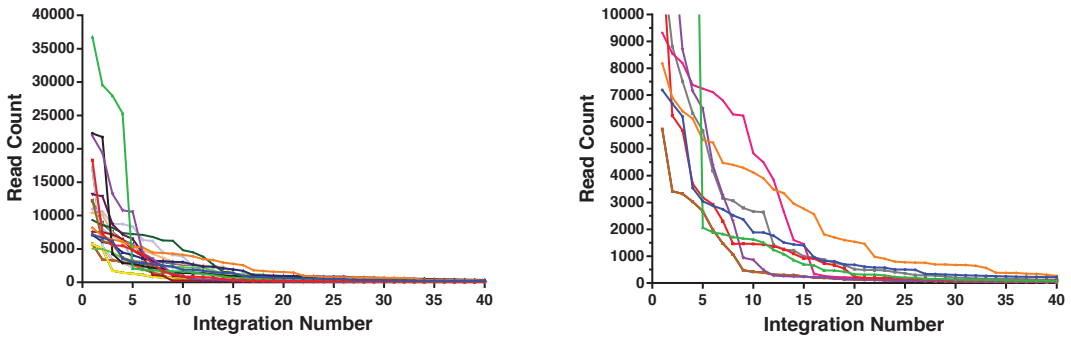


Figure 5.5: Read coverage for the major integrations without removal of duplicates. Left: Total 5' plus 3' read coverage for the top 40 integrations in the spleen samples from the 19 mice in the serial bleed study (chapter 4). **Right:** Closer view of the fall in read count in 8 selected samples from this group. In most samples there was a sharp fall in read count after the top few integrations, but in some this drop off was more gradual. In all cases the read coverage fell below 400 reads by the 40th integration and in most it was under 200.

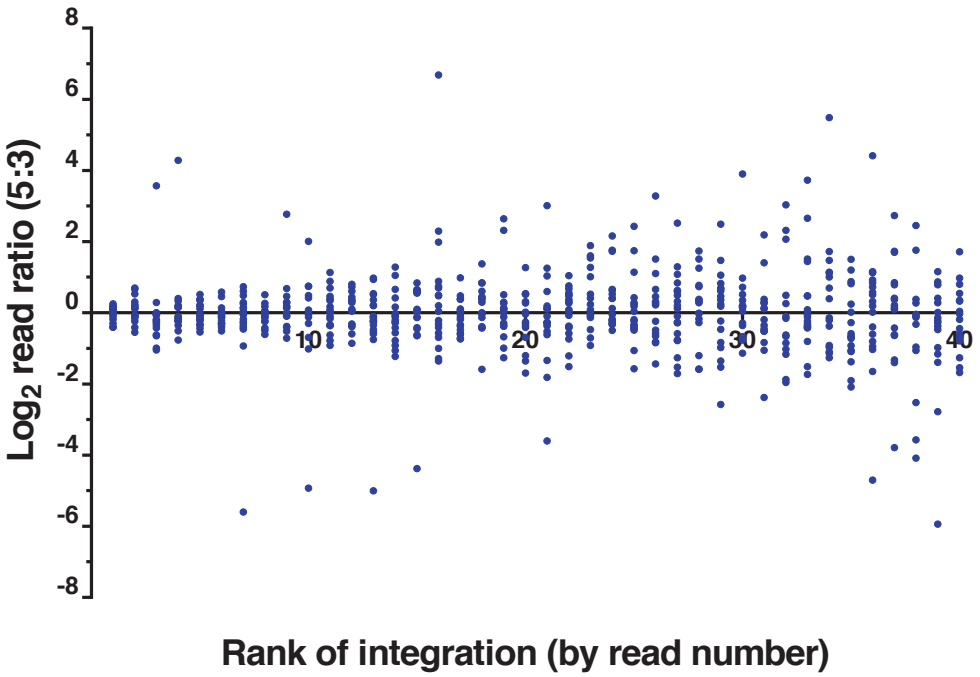


Figure 5.6: Correlation of 5' and 3' reads in the non-duplicate filtered analysis. The 5' to 3' ratio for the 40 integrations with highest coverage in each sample are shown for the 19 mice in the serial bleed study (chapter 4). The $\log_2$ of 5' and 3' read ratio is shown. 95 of the 760 integrations were excluded from analysis of 5' to 3' ratios as they only mapped to one end of the transposon.

5.2.4 Integrations that persisted on serial sampling generally had high read coverage using TraDIS

In general, the integrations which persisted on serial blood samples and recipient tumours gave high read number using the TraDIS method. Selected examples from mice which had serial sampling are described below.

5.2.4.1 Npm1^{cA}/GRL 19.2B

Mouse 19.2b is an interesting example because four integrations each account for over 10% of the total sequencing reads from this primary tumour, while all other integrations had read coverage of less than 1.5%. In all mice transplanted with tumour 19.2b, the recipient tumour contained these same four integrations which accounted for the majority of sequencing reads (figure 5.7). In the two 1000-cell transplants (1.5 and 1.6), there was not a single other integration that had over ten reads after duplicate removal and only 28 other integrations were mapped in total between these two samples. The four top integrations were located in i) intron 17 of

Nup98 (reverse orientation), ii) intron 49 of *Nf1 (*forward orientation), iii) intron 6 of *Arap2* (reverse orientation) and iv) an intergenic location on chromosome 10 just upstream of *Avpr1a*. It is likely that the driver integrations for this tumour are among these four sites and both *Nup98* and *Nf1* were located in CIS for this cohort of mice.

In the serial blood samples from this mouse which were analysed by Illumina sequencing, the *Nup98* integration was already the major integration on the week 20 blood sample taken seven weeks before the mouse died and the *Nf1* integration was the ninth integration at that time. By the week 22 sample these were the top two integrations by read number and the integrations in *Arap2* and chromosome 10 were detected for the first time in much lower read numbers. None of these integrations were detected in the week 18 sample, although an alternative integration in *Nup98* was detected in low numbers. This correlates reasonably well with the 454 sequencing data in which only the *Nup98* integration was apparent in the week 20 blood sample. Using the 454 sequencing method *Nf1* and the intergenic integration on chromosome 10 were first detected at week 22 and the *Arap2* integration at week 24.

Together these results reveal that it took several weeks after acquiring all four mutations for the mouse to develop frank leukaemia. The *Arap2* and chromosome 10 intergenic lesions are not obvious candidate drivers. In the absence of this serial data it would be easy to assume they were passengers present at the time the *Nf1* and *Nup98* integrations were acquired. However, although the Arap2 and chromosome 10 integrations are in similar proportion to the *Nf1* and *Nup98* integrations in the final tumour, the TraDIS data shows these integrations expanded in read number over a different time course and in that sense behaved like at least one of them was a driver. Alternatively, a non-transposon driver mutation may have occurred in a cell carrying the two lesions as passengers.

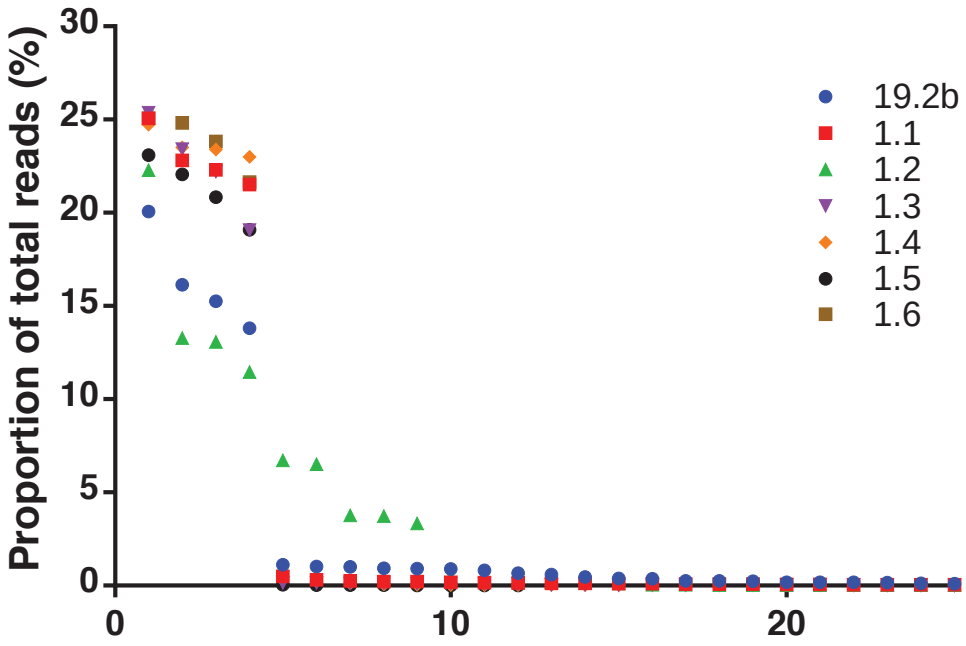


Figure 5.7: Proportion of total reads taken by the top 25 integrations in tumour 19.2b and associated recipient tumours. In each tumour the top four hits were identical and it was only in the primary tumour and 19.2b.1.2 that other transposon integrations were found in any number.

5.2.4.2 Npm1^{cA}/GRL 21.3j

As highlighted in chapter 4, mouse 21.3j had two separate transposon integrations upstream of *Csf2* (table 5.6), although only one persisted in the majority of transplants. Five recipient tumours from 21.3j were analysed using TraDIS; namely two 10^6 cell transplants and one transplant each of 10^4, 10^3 and 10^2 cells (figure 5.8). The persisting *Csf2* integration (11:54252890) was the top integration by read number in the primary tumour and was the only integration which was shared by all of the recipient tumours (figure 5.8). The second *Csf2* integration (11:54250980) was the 40th integration in the primary tumour and seemed to track with *Mll1* which was the 24th ranked integration. Of the recipient leukaemias, only 1.1 and 1.2 had the *Mll1* or *Csf2* 11:54250980 integrations and both were present in similar read numbers in each case. However, these two tumours also had the *Csf2* 11:54352890 integration as their top hit.

To determine if these *Csf2* integrations were co-occurring in the same clone I generated single cell derived colonies from frozen spleen cells of the primary tumour.

After eight days of growth in semisolid media (M3434), ten single-cell derived colonies were picked and re-suspended in RPMI media for tail vein injection into NSG mice. Of the ten recipient mice, four developed leukaemia after a latency of 36-42 days (appendix 4D). Three of these tumours were sequenced using the TraDIS protocol and in all three cases the 11:54250980 and *Mll1* integrations were among the top three hits, but the 11:54252890 integration was not detected (figure 5.9). The third top three hit varied between the colony-derived recipient tumours. Also, although several of the transposon integrations in colony-derived leukaemias were shared with the primary, most were not; which indicates that transposons were still active during colony generation and/or within the recipient mice.

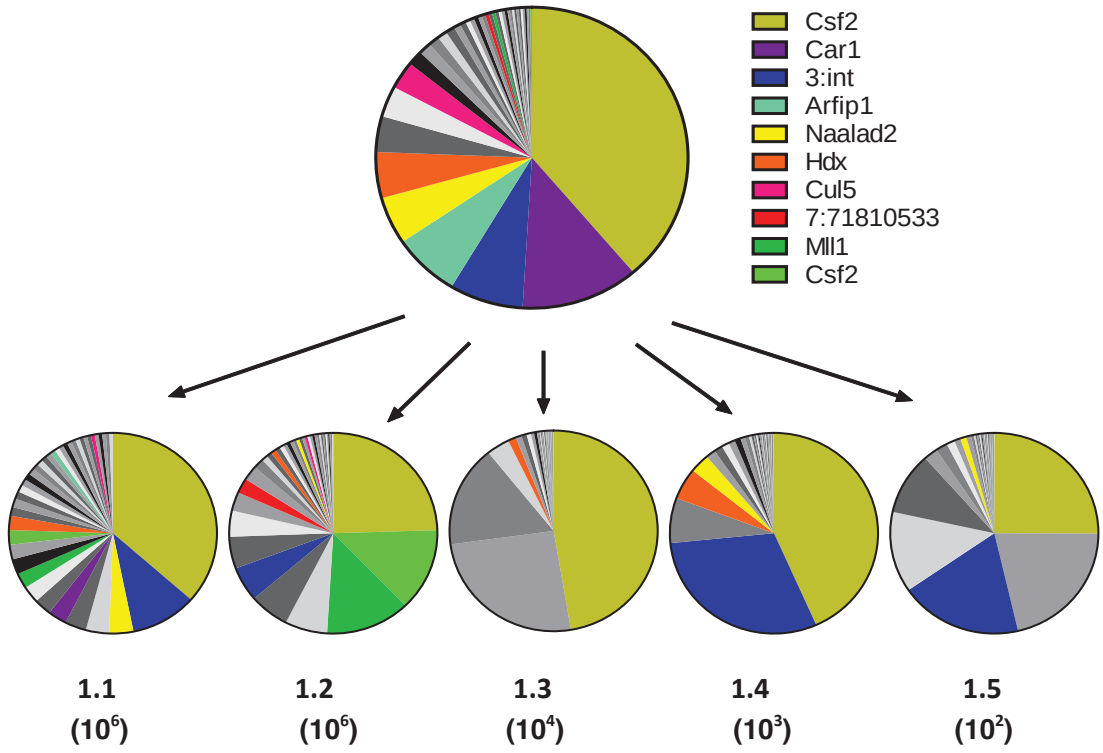


Figure 5.8: Shared integrations in primary tumour 21.3j and five recipient tumours. The top 40 integrations by read number are represented. Those shown in colour are shared between different tumours, but those in greyscale are not. The integrations are represented as a proportion of the total reads taken by the top 40 integrations. The number of spleen cells transplanted into each recipient mouse is shown.

Two serial blood samples from 21.3j were also analysed using the TraDIS protocol; the week 20 and 24 samples. In the week 20 blood sample the *Mll1* integration was ranked 8th according to read count and the *Csf2* integration at 11:54250980 was 18th,

while the *Csf2* integration that dominated the final tumour sample was only detectable at low count. Of note, a third *Csf2* integration at 11:54250118 was the 15th transposon integration at that time. By the week 24 blood sample, one week pre-death, the 11:54252890 integration had expanded to become the top read, while *Mll1* was 15th and the second *Csf2* integration was 38th. The third integration that was the most prominent of the *Csf2* integrations (15th) in the week 20 sample was no longer detected.

Together these results indicate that there were multiple transposon integrations in *Csf2* in mouse 21.3j during the pre-leukaemic period. In the final tumour the two detectable *Csf2* integrations occurred in separate clones. The clone containing the 11:54252890 integration dominated the final tumour sample mixed cell transplants. However, in colony transplants a different leukaemic clone, containing the *Mll1* and 11:54250980 integrations dominated. Also, in the 10^6 cell transplants the latter clone seemed to be growing faster than the former, although during leukaemic evolution the opposite appeared to be happening.

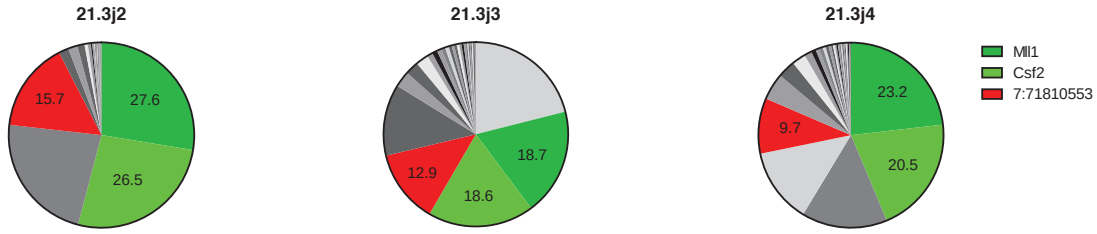


Figure 5.9: Transposon integrations in leukaemias generated after transplantation of one of three single cell-derived colonies from primary 21.3j. Identical integrations are depicted in the same colour (also used in figure 5.9) in three different recipient leukaemias. Numerals represent percentages of all reads from the top 30 integrations. Integrations not shared between the leukaemias are depicted in grey.

5.2.4.3 Npm1cA/GRL 16.3f

Mouse 16.3f had atypical results on 454 analysis because it had detectable transposon integrations in multiple CIS genes several months prior to the onset of leukaemia, however most of these did not persist in serial transplants. The TraDIS sequencing data shows that many of the main integrations in the tumour sample were those that had persisted in serial blood samples. However, it seems that the major primary tumour clone(s) was outcompeted in the transplant experiments. The

integrations that were shared by all transplant recipient tumours each accounted for less than 0.5% of the total reads in the primary tumour (table 5.8). This also shows that some of the CIS hits that went missing were in a major clone in the primary tumour (eg *Flt3*, *mmu-mir-29b-2*), whereas others such as the *Nf1* integration 11: 79447002 (11:79260504 on Gm37 version) were not.

Insertion site	Gene	49	51	53	55 (spl)	1.2	1.2.1	1.2.2	1.3	1.4	1.4.1	1.4.3	1.5
11_684234	Intergenic	0.02	0.16	0.08	9.22								
1_1950065	mmu-mir-2	5.04	4.24	11.22	8.93								
16_334978	Zfp148	0.98	0.18	0.97	3.21					0.09			
14_219988	Intergenic				3.16								
3_3019015	Mecom	1.43	2.71	0.78	2.38					0.04			
1_5380644	Intergenic	2.33	2.32	4.20	1.62					0.10			
5_1473658	Flt3				1.15					0.11			
6_1036492	Chl1	0.32	1.75		1.03	1.47	0.41	0.25		1.29	0.46	0.64	1.25
17_696791	Intergenic	2.17	1.75	0.28	1.03					0.03			
4_3730091	Lyn			1.03	1.02								
4_1479088	Lrrc69			0.16	0.93								
4_3239235	Bach2	0.52	0.60	0.25	0.89								

Insertion site	Gene	49	51	53	55 (spl)	1.2	1.2.1	1.2.2	1.3	1.4	1.4.1	1.4.3	1.5
16_249238	Lpp				0.50	9.79	12.06	10.46	7.05	10.20	11.00	10.26	8.74
19_557646	Tcf7l2				0.30	8.80	7.69	8.18	6.57	9.00	8.29	7.19	8.76
9_4484182	Mll1				0.60	8.32	8.93	8.55	6.13	8.27	8.78	9.93	8.75
11_542509	Csf2				0.32	7.71	6.81	7.83	5.89	8.12	8.74	8.52	8.84
16_425617	Intergenic				0.39	3.93	1.29	1.53	2.90	2.78	1.92	1.99	3.63
16_160282	2310008H04Rik				0.27	1.75	3.49	2.88	3.17	5.48	3.58	2.57	5.21

Table 5.8: Major integration sites in the primary and recipient tumours from 16.3f. The top 12 hits from the primary tumour and their coverage in six transplant leukaemias are shown at the top. In the bottom table the top six hits in the transplant leukaemias and their coverage in mouse blood at weeks 49, 51, 53 and from its spleen at the time of death are shown. The numbers refer to the proportion of total reads in a sample assigned to that integration. The results for the week 49, 51 and 53 blood samples and spleen samples from the primary and recipient tumours are included. The clone containing *Mll1* and *Csf2* that was detected in all the recipient tumour samples, was different to the one containing the mmu-mir-29b-2 integration which was prominent in the late serial blood and primary tumour samples.

It is important to highlight that case 16.3f is an exception rather than the rule. In most cases the integrations which persisted on serial transplant were high ranking integrations in the primary tumour. Often the pattern of the major transposon integrations was very similar in the primary and recipient tumours.

5.2.4.4 Npm1^{cA}/GRL 6.4a

Case 6.4a is a much more typical example, where the major integrations in the primary also predominated in the recipient tumours. The TraDIS sequencing results from nine of the 15 recipient tumours are represented in figure 5.10. Although the proportion of reads for the *Dmxl1* integration fell in the third generation transplants, and the intergenic integration in chromosome 10 was more prominent in tumour

1.2.1, overall the major integrations were shared in similar proportions in all tumours. Of note, in the 454 sequencing analysis the *Csf2* integration was not detected in the primary tumour sample, although it was detected in the majority of transplants. It is surprising this was mapped in any of the samples given that the nearest *Mbo1* restriction site is 764 bases from the *Csf2* integration. The 7:93253552 (7:100402062 on Gm37) and 10:11589188 (10:11308987 on Gm37) (see figure 4.15) were only detected in some transplants on the 454 analysis even though there was an *Mbo1* restriction site within 300 bases of both of these integrations.

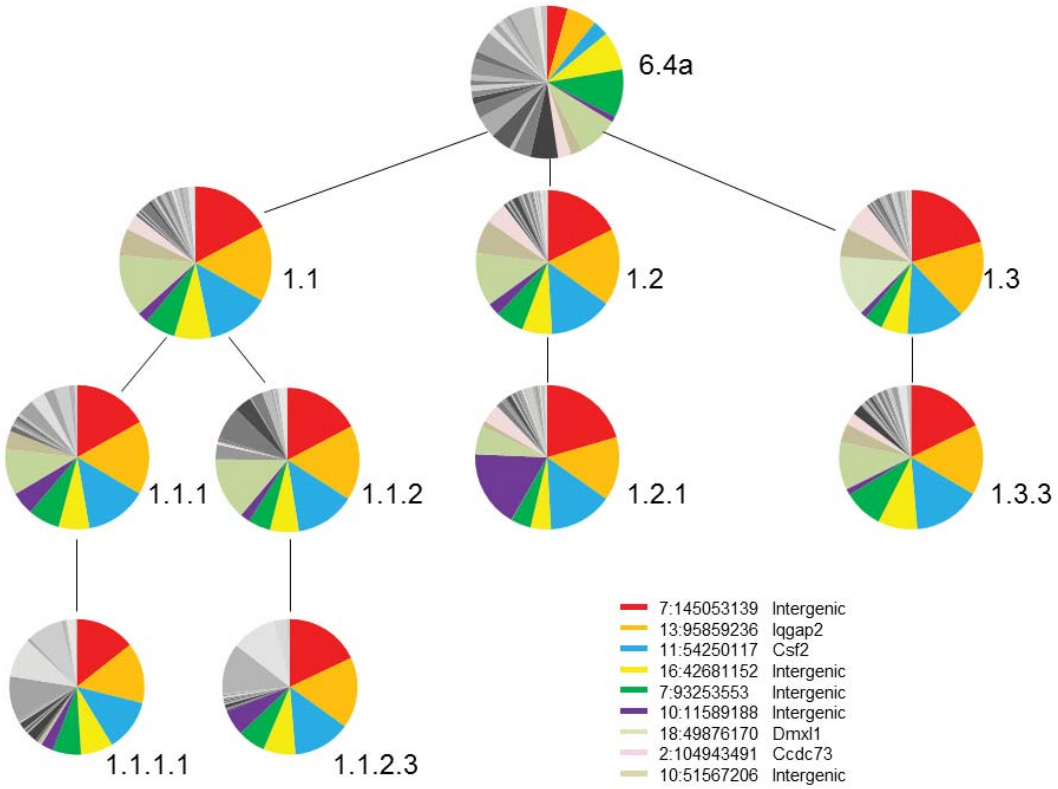


Figure 5.10: Shared transposon integrations in primary tumour 6.4a and 9 of its recipient tumours. The shared integrations are plotted in colour and the identity of these integrations is indicated. Integrations shown in grey-scale differ between the tumours.

5.2.4.5 Npm1^{cA}/GRL19.2d

On the 454 sequencing analysis of mouse leukaemia 19.2d several CIS genes were identified in the serial blood and final tumour samples including *Nup98*, *Nrf1* and multiple integrations near *Csf2* (*Gm12223*) and within *Nf1*. However, none of these persisted on multiple transplants. The TraDIS data reveals that all of these

integrations, with the exception of one that was downstream of *Csf2*, were represented by very small numbers of reads in the final tumour.

All six of the recipient tumours from this mouse, as well as seven pre-leukaemic blood samples, were analysed by TraDIS sequencing. Once again, the major transposon integrations in the primary tumour were those that were shared by all of the recipient tumours (figure 5.11). The proportion of reads taken by each of these integrations in the serial blood samples are shown in table 5.9.

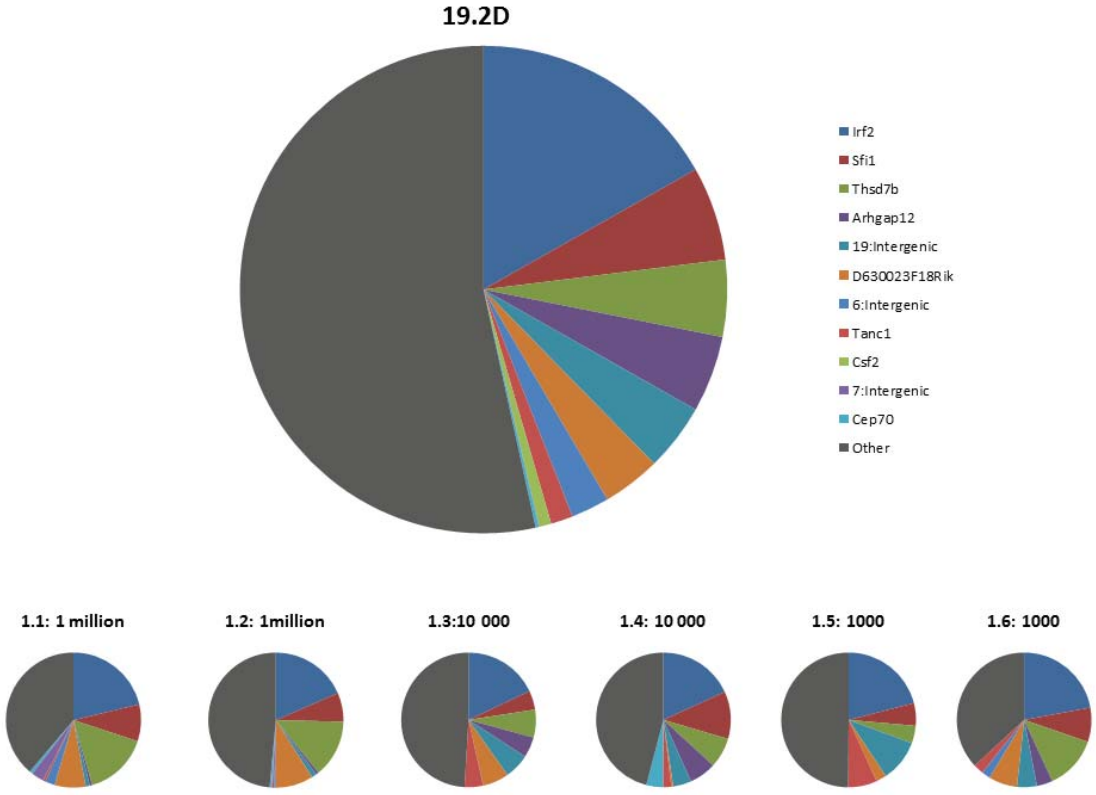


Figure 5.11: Major transposon integrations in 19.2d and its recipient tumours. The cell doses for each of the transplants are shown.

A

		11_3143139 Sfi1	8_46809981 Irf2	18_6047212 Arhgap12	6_78788203 intergenic	19_26987577 intergenic	1_65119090 D630023F18Rik	1_129345638 Thsd7b	2_59802310 Tanc1	11_54253304 Gm12223 (Csf2)
Blood	wk16	1.75	7.78							
	wk18	0.06	0.02							
	wk26	2.21	9.00							
	wk32	1.73	7.89	2.39	1.14	0.02				
	wk34	7.97	10.16	8.06	3.98	0.29	1.07	0.82		
	wk36	4.34	16.13	7.00	3.65	0.09	1.14	0.50		
	wk38	4.02	15.17	6.04	2.75	0.79	2.18	0.72	0.35	0.05
	19.2d	4.08	6.84	3.12	1.74	3.85	2.91	4.72	1.09	0.67
Transplants	1.1	7.02	14.54	0.42	1.63	0.67	5.17	10.78	0.39	0.01
	1.2	4.78	12.94	0.47	0.29	0.81	6.30	9.65	0.28	
	1.3	3.23	12.74	3.50	0.02	4.29	4.43	4.72	3.16	
	1.4	6.73	12.80	4.40	0.00	3.00	0.30	4.91	1.39	
	1.5	3.93	15.13	0.00	0.00	7.26	1.82	3.04	4.97	
	1.6	6.10	15.54	2.68	1.50	3.36	4.85	9.06	1.71	

B

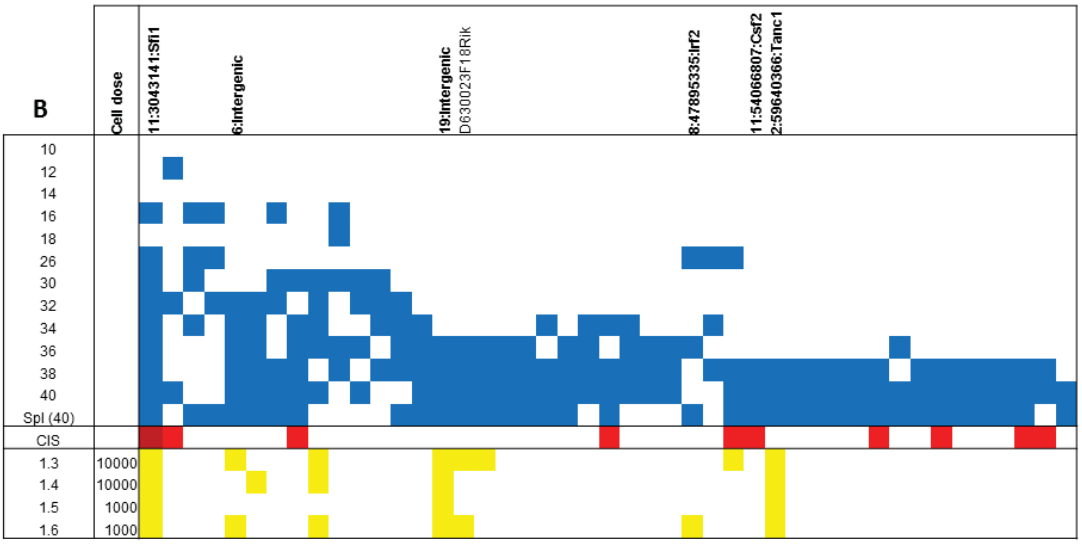


Table *5.9:* Timing of major tumour integrations in the serial blood samples (A) The proportion of reads taken by the transposon integrations that persisted in multiple recipient tumours are shown for each of the serial blood and tumour samples. **(B)** The presence of these integrations in the same samples analysed with the 454 protocol. The integration positions correlate, but the precise coordinates differ as the 454 and Illumina analyses were analysed using different versions of the mouse genome (GRCm37 v GRCm38).

5.2.5 TraDIS analysis of *Npm1^{cA}/GRL* primary tumours that did not transplant

Mouse 7.5c was one of the two serially bled cases in which transplant of primary spleen cells into NSG mice failed to initiate leukaemia in the majority of recipients. This was the mouse with MPD-like changes in the pre-leukaemic blood samples (figure 4.11). The TraDIS analysis of the primary tumour identified the major integrations as i) *Flt3*, ii) 2:72469204 intergenic (missed by 454 analysis), iii) 16:54136662 intergenic (=16:54136774), iv) *Nup98*, v) 16:52008898 intergenic (=16:52009011) and vi) 11:112705632 BC006965 (missed by 454). Each of these

integrations accounted for over 2% of non-duplicate Illumina sequencing reads. The viability of the spleen cells was noted to be poor on thawing (<10%). The recipient mice that became sick did so after a prolonged latency and typically did not have signs of leukaemia at necropsy, although some showed myeloproliferative changes on histopathology. Two of these mice were analysed by the TraDIS protocol but their integrations showed little overlap with the primary tumour.

The other sample that failed to generate myeloid leukaemia in the majority of recipients was from **mouse 16.3h**. Two of the recipient spleen samples were analysed by TraDIS even though they were not found to have leukaemia on histopathology and blood film examination (appendix 4D). One of these samples (1.4) showed no major overlap in transposon integrations with the primary tumour, however the other (1.1) shared the top four integrations including one upstream of *Csf2*, and these were in similar proportion to the primary tumour (table 5.10).

Integration site	Gene	16.3h (Spleen)	16.3h (liver)	1.1
14_103113828	Mycbp2	8.18	11.13	6.58
11_54252781	Csf2	6.77	8.30	8.53
16_76591594	Intergenic	6.25	8.69	2.65
16_37185445	Stxbp5l	4.80	6.70	2.60
3_102196149	Vangl1	4.61	4.39	0.00

Table 5.10: Shared integrations between 16.3h and one recipient. This recipient failed to develop overt leukaemia despite sharing several major integrations with the primary tumour.

Mouse 7.5h also had several transplants that failed to generate leukaemia. Mouse 1.2, which was transplanted with 10^6 cells, eventually developed a poorly differentiated myeloid leukaemia but only after a latency of 99 days, which was much delayed compared to the timing of recipient tumour development in most other cases. This tumour was successfully transplanted on to three further mice which developed leukaemia after a latency of only 25-36 days. I was able to map a typical number of transposon integration sites in the primary tumour, but we were unable to identify transposon integrations in the recipient tumours, despite generating good quality DNA and repeating the analysis (both 454 and Illumina) on multiple occasions. Transposon integration sites were not amplified in the TraDIS library preparation and following the qPCR results the samples were excluded from pooling

for sequencing. Therefore, it appeared that these recipient tumours were not transposon driven.

To further investigate the mechanism of leukaemogenesis in the transplants from mouse 7.5h we performed karyotyping and FISH analysis on three recipient tumours. All showed complex chromosomal abnormalities including Robertsonian translocations involving the donor and other chromosomes (figure 5.12). Stored metaphases on the primary tumour were therefore examined and although Robertsonian translocations were not identified, this was found to have a transposition of the centromere of chromosome 16 into the long arm of chromosome 16 in eight of the ten metaphases analysed. An additional del(3), der(3)t(3:16) was found in one metaphase (figure 5.13).

A

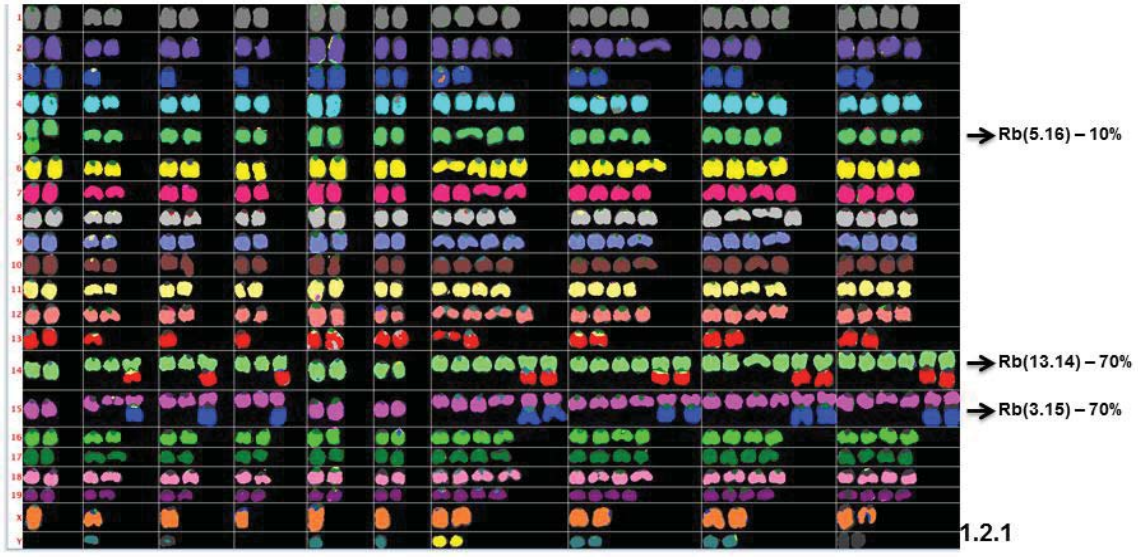
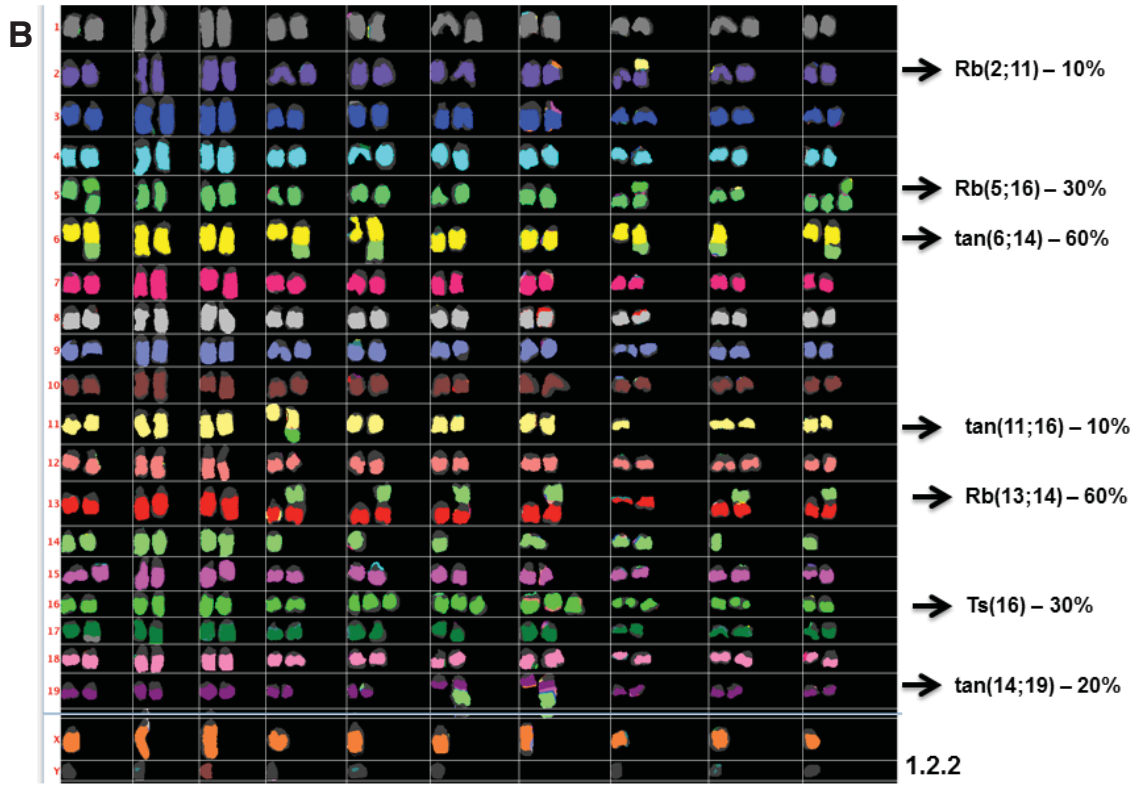
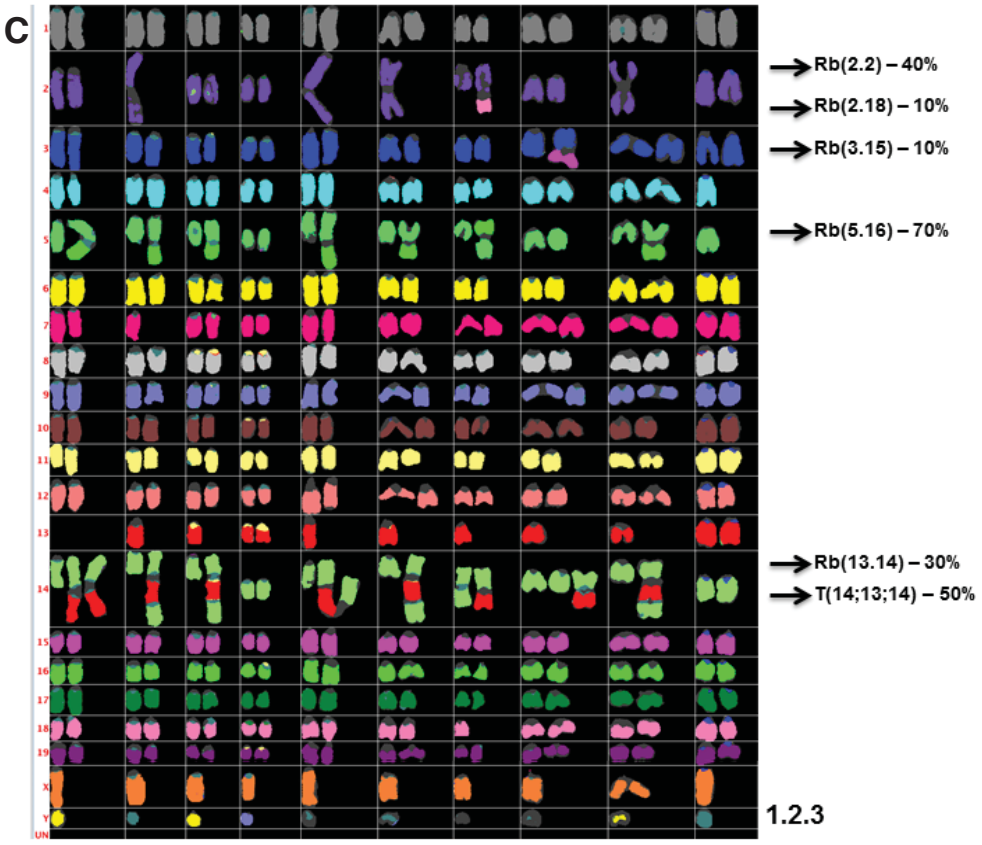


Figure 5.12: Metaphase paint images of transplants 1.2.1 (A, above), 1.2.2 (B, next page) and 1.2.3 (C, next page), showing Robertsonian translocations in all cases. In 1.2.1 there is tetraploidy in 4 metaphases in addition to the indicated Robertsonian translocations involving chromosomes 3, 5, 13, 14, 15 and 16. In 1.2.2 the abnormalities in addition to the indicated Robertsonian translocations include trisomy of chromosome 16 and tandem translocations between chromosomes 6 and 14, 11 and 16 and 14 and 19. In 1.2.3 there are several Robertsonian translocations, including one between chromosomes 13 and 14, that also has telomeric association between chromosomes 13 and 14 (T 14; 13; 14). The FISH was performed by Ruby Banerjee who supplied these images.


B
Rb(2;11) – 10%
Rb(5;16) – 30%
tan(6;14) – 60%
tan(11;16) – 10%
Rb(13;14) – 60%
Ts(16) – 30%
tan(14;19) – 20%
1.2.2


C
Rb(2.2) – 40%
Rb(2.18) – 10%
Rb(3.15) – 10%
Rb(5.16) – 70%
Rb(13.14) – 30%
T(14;13;14) – 50%
1.2.3

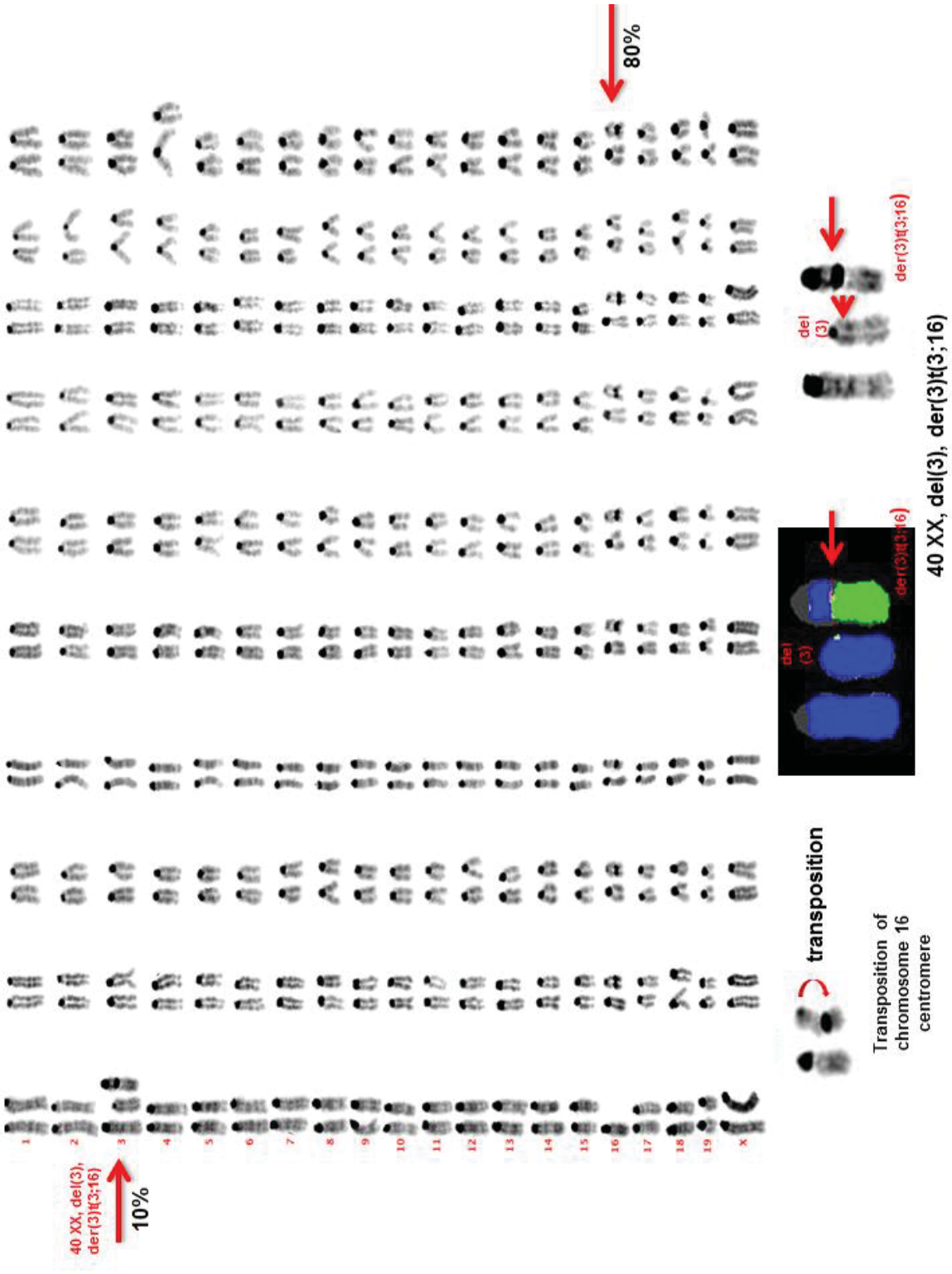
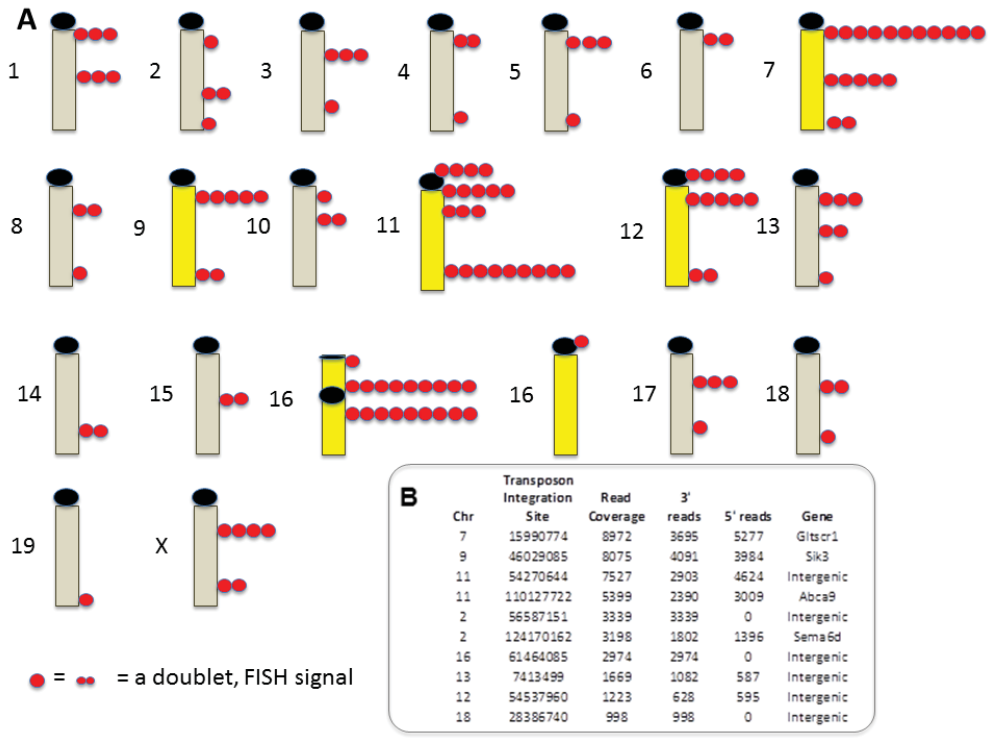


Figure 5.13: Metaphase karyotyping of primary tumour 7.5h. This showed del(3), der(3)t(3;16) in one metaphase and transposition of chromosome 16 centromere within the long arm of chromosome 16 in eight. Close up images of the abnormalities are shown at the bottom, including a metaphase paint image of the translocation. Images provided by R. Banerjee

We generated a fluorescently labelled probe directed at the *GrOnc* transposon and used this to investigate if these structural abnormalities were occurring at transposon integration sites. FISH analysis was performed by Ruby Banerjee. In the analysis of 10 metaphases from the primary tumour, transposon FISH signals were detected at the transposed chromosome 16 centromere in all nine metaphases with this abnormality. She also reported transposon integrations in chromosomes 7, 9, 11 and 12 in a large proportion of metaphases. The top three integrations by read number on the TraDIS sequencing data were on these chromosomes (figure 5.14). Furthermore, analysis of the transplant recipient metaphases with the same probe showed that transposons were localised within the centromeres of multiple chromosomes, but were not found with confidence at other sites (figure 5.15). This suggests the transposon may have a role in generating the Robertsonian translocations and that these tumours may have been transposon driven, even though transposon integrations were not mapped on TraDIS or 454 sequencing.



B

Chr	Transposon Integration Site	Read Coverage	3' reads	5' reads	Gene
7	15990774	8972	3695	5277	Gltscr1
9	46029085	8075	4091	3984	Sik3
11	54270644	7527	2903	4624	Intergenic
11	110127722	5399	2390	3009	Abca9
2	56587151	3339	3339	0	Intergenic
2	124170162	3198	1802	1396	Sema6d
16	61464085	2974	2974	0	Intergenic
13	7413499	1669	1082	587	Intergenic
12	54537960	1223	628	595	Intergenic
18	28386740	998	998	0	Intergenic

Figure 5.14: Transposon FISH analysis of 7.5h. (A) Diagramatic representation of the positions at which transposons were recorded on FISH analysis by Ruby Banerjee. Each red dot indicates a transposon integration. The chromosomes with the largest number of integrations are shown in yellow. There were integrations in the transposed centromere of chromosome 16 in nine of the ten metaphases. **(B)** The top hits by read count on TraDIS sequencing were in chromosomes 7, 9 and 11.

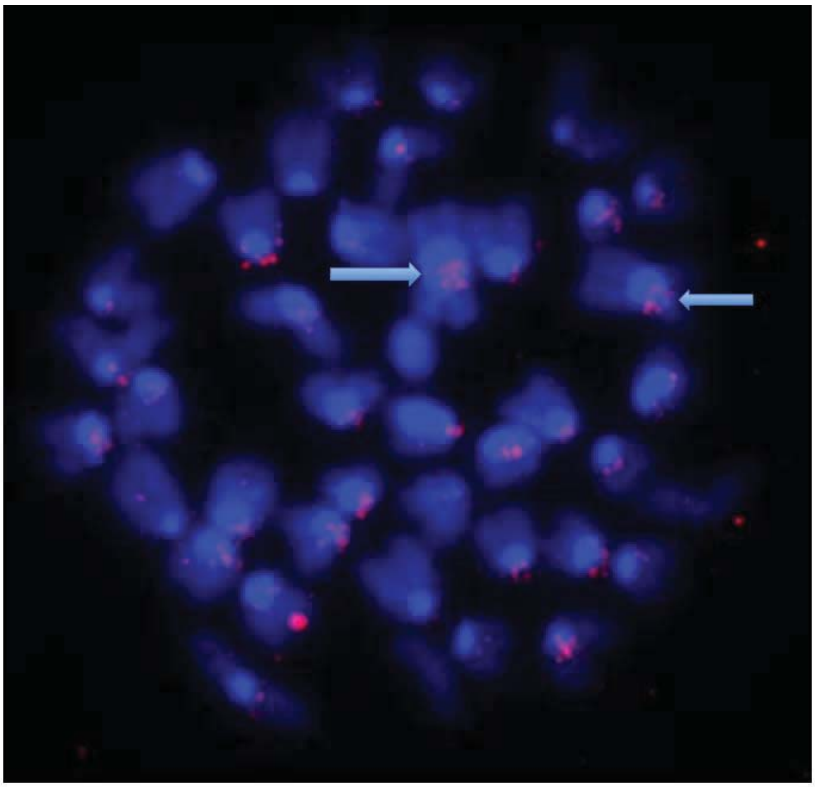


Figure 5.15: FISH of a metaphase from 7.5h recipient tumour 1.2.3. The transposon integrations are indicated by red double dots and the centromeres fluoresce bright blue. The arrows indicate some of the clear transposon integrations within centromeres.

5.3 Discussion

In this chapter I have presented the results of a re-analysis of the *Npm1*^{cA} GRL IM cohort using TraDIS, a method employing DNA shearing followed by Illumina sequencing. The CIS analysis identified 18 of the 27 CIS found in the 454 analysis and added several additional CIS of interest. The advantage of this sequencing approach was that read depths of major integrations correlated with the size of the leukaemic clone/sub-clone harbouring them. This was the result of the fact that the TraDIS protocol uses shearing to perform fragmentation of genomic DNA, which generates a smooth distribution of ligation sites around transposon integrations, and also requires significantly fewer rounds of PCR amplification (30 vs 62). Together these factors significantly reduce the problem of PCR amplification bias seen with the restriction/454 protocol. I have shown that this method is at least semi-quantitative, by demonstrating a good correlation between the proportions of reads from the major integrations mapped from the 5' vs 3' end of the transposon. The

major integrations by read number were also reproducible on re-sequencing DNA from a given tumour and on sequencing primary and recipient tumours.

This dataset was analysed both with and without removal of PCR duplicates. In reality the presence or absence of duplicates made little difference to the order of the top hits. The reason for analysing without removal of the PCR duplicates was because of concern that the clonal representation of the major integrations would be underestimated, due to the finite number of unique ligation points around any individual transposon. It seemed likely that all possible shearing positions could be utilised around integrations present in the majority of tumour cells. Although this did occur, the small overall effect it introduced was to reduce the read proportion taken by the top few hits, without changing their order significantly.

Regardless of whether or not PCR duplicates were included in the analysis, the typical pattern was one of a few 'step-wise' drops in the proportion of reads assigned to each of the top 10-20 integrations in a tumour. The much larger number of integration after these top10-20 were detected by small numbers of reads. The number of integrations in each 'step' or 'tier' did vary from case to case, but generally there were around three significant 'drop-offs' in read coverage amongst the top 20 integrations.

These quantitative read results were used to infer which integrations were present in the major clones and which were found in only a small number of cells. It was evident that the quantitative nature of the data did not hold well for minor integrations. It was not possible to draw conclusions about the possible co-occurrence of particular integrations in the same clone when they were represented by lower, but similar levels of coverage, as the presence of more than one sub-clone of similar size would lead to similar results. Groups of integrations that co-occurred together in transplant recipient tumours could be traced back to the primary tumour and were often found to have a similar read coverage in that tumour, for example integrations in 16.3f. However, it is not possible to pre-emptively pick these out as a single clone in the absence of the transplant data. Even with the evidence from the transplants that these mutations tracked together, it is still theoretically possible that they were occurring in multiple sub-clones, each of which expanding at a similar rate in the recipient mice.

It was not surprising that some of the top hits account for more than 7% of the total reads even though each cell starts with 15 copies of the transposon. Some transposons may remain un-mobilised in the donor locus and the re-integration efficiency for *SB* transposons is not 100%, so over time the number of transposons per cell is expected to fall. Therefore, it is not possible to determine a read proportion that equates to an integration being shared by all cells within a tumour. Also, the number of integrations in the major clone will affect the read coverage assigned to each of them.

The core aim of IM analysis is to distinguish true driver CIS integrations from ones that arise due to random clustering of insertions. Increasing the read coverage can in principle increase the problem of false-positive CIS, unless appropriate filtering is applied to exclude spurious and/or low level reads. This could be achieved by giving more weight to the integrations which account for a high proportion of reads and are therefore more widely represented in the tumour cell population. As I have shown, the integrations that have high read coverage are typically the ones that persist on serial transplant experiments and therefore are the group of integrations amongst which the major drivers for an individual tumour are likely to reside.

There are various published methods for performing CIS analysis on transposon and retroviral IM screens. However, there is no consensus strategy and with the current shift to Illumina based sequencing approaches the problem of false positive CIS is only likely to grow. In the literature there are few references to applying cut-offs to sequencing data to eliminate insertions that are only read a few times and therefore likely represent non-clonal insertions. TAPDANCE is a publicly available software that aims to fully automate the analysis of CIS and rank their importance (Sarver et al., 2012). In the analysis of Illumina sequencing data TAPDANCE uses a cut-off based on the percentage of total mapable reads. The recommendation is that this cut-off be set at 1/10 000, so only insertions with at least 10 reads will be included in the CIS analysis if there are 100 000 sequencing reads for the region. Another study used the number of unique adaptor ligation points on Roche 454 sequencing of sheared DNA to estimate the clonality of individual insertions (Koudijs et al., 2011). On analysis of *PB* insertions in a clonal embryonic stem cell line they found that the number of unique ligation points correlated with the expected number from permutation analysis in more samples than the raw read count. On mixing studies of

two clonal cell lines with mouse mammary tumour virus (MMTV) insertions they showed a strong correlation between the DNA mixing ratios and the number of unique ligation points at five of six MMTV insertion sites and had a sensitivity of approximately 10% for detecting bi-clonal tumours. On comparative analysis of sheared and digested splinkerette data from *SB* induced lymphomas they showed that this protocol could be used to enrich for biologically relevant insertions by excluding random insertions represented by single ligation points and likely occurring at low frequency within the tumour mass.

It is debatable as to how best to apply the 'cut-off' for reads to include in the CIS analysis. I chose to include the top 10, top 25 and top 100 insertions per sample to allow for variation in read coverage. If the cut-off was set based on read number, the number of integrations per tumour would be expected to vary, not only as a function of clonality, but also due to variation in sequencing depth. The cut off applied here of the top 10, 25 or 100 hits was chosen as it was easy to apply and used the same number of integrations per tumour regardless of sequencing depth. A reasonable, but more difficult alternative would be to apply a cut-off based on read proportion, for example, including all integrations that account for over 0.5% of the total reads within a tumour.

Going forward it is difficult to know what threshold of reads to recommend for CIS analysis. Certainly, there seems to be no need to include all of the integrations found in each tumour sample. The TraDIS protocol allowed very deep sequencing coverage and including all of the hits added unnecessary burdens to computer processing requirements and significantly extended the list of CIS hits, but probably at the cost of including a number of false positive CIS. As the number of included integrations per tumour was increased, the number of identified CIS also increased. Limiting the analysis to the top ten hits allowed identification of a small set of CIS that are likely to be important. However, it is also probable that some drivers will be missed with this approach. As I have shown in tumour 21.3j and 16.3f, integrations which account for <1% of reads in the primary tumour, may not be in the dominant tumour clone, but may be present in a smaller clone which was still capable of initiating leukaemia in recipient mice. It is therefore helpful to have the analysis performed at multiple cut off levels.

There were notable differences in the number of tumour hits and the CIS identified using the various analysis cut-offs that I applied. Although the CIS at *Mll1* is common to all lists and was found in 16 tumours overall, insertions in *Mll1* were amongst the top 10 hits in only two tumours, and in the top 25 in six. This suggests that although integrations around this well-known leukaemia associated gene are common, the integration is not in the dominant primary tumour clone in the majority of cases. Similarly, the integrations in *Nup98* and *Nf1* did not appear to be in the major clone in most tumours with these integrations, although they were in some cases.

In contrast, integrations in other CIS genes were typically amongst the top 10 hits by read number when they were detected in the top 100, which suggests that when present, they are usually in the major clone. For example, *Pax5* was in the top 10 hits in five of the six tumours it was found in, *Zfp423* in four of five and *Flt3* in six of ten. Integrations upstream of *Csf2* were found in the top ten hits in 25 tumours and were only found in the top 100 in ten further cases. Therefore, *Csf2* was among the integrations in a major tumour clone in over 50% of cases and it was amongst the top 100 integrations by read number in around 76%. *Bmi1, Iqgap2, Nav2* and *Tmem135* were only detected among the top 100 hits in two cases each, but in both cases they were in the top 10 hits. The significance of these hits as a CIS was therefore lost when 100 integrations were included in the analysis. Of these integrations, only *Bmi1* was identified as a CIS on the 454 analysis.

Overall there were nine CIS identified using only the top 10 integrations that were not detected in the 454 analysis. Amongst these was *Ets1*, a member of the ETS protein family of helix-loop-helix domain transcription factors. This has previously been identified as a CIS gene in a *SB* transposon IM screen of erythro-megakaryocytic leukaemia (Tang et al., 2013). In cases of AML with 11q23 amplification, the *ETS1* gene is in the amplified region(Poppe et al., 2004; Rovigatti et al., 1986) and over expression of *ETS1* has been demonstrated in CD34+ haematopoietic progenitor cells from patients with AML, while decreased expression was shown to be associated with differentiation of leukaemia cells(Lulli et al., 2010). Furthermore Ets-1 is among the transcription factors known to be important in regulation of the *GM-CSF* promoter (Thomas et al., 1995) and the autocrine production of GM-CSF in the leukaemic progenitor cell line KG1a was recently shown to be mediated by *ETS1*(Bade-Döding et al., 2014). In this context, it is noteworthy that two of the three

tumours with *Ets1* integrations as a top 10 hit did not have *Csf2* integrations, even though *Csf2* was the most frequently hit CIS in this screen and was amongst the top 100 integrations in three quarters of the tumours. *Ets1* is therefore an interesting CIS for further study, which was not apparent on the 454 analysis.

The other CIS that came up on the top 10 Illumina analysis, but were not identified as CIS in the 454 data, include *Pik3r5*, *Rasgrp1*, *Cblb* and *Hecw2*. *Pik3r5*, which encodes a regulatory subunit of the PI3K gamma complex and *Rasgrp1*, a nucleotide exchange factor involved in activating *Ras* and the Erk/MAPK pathway, were both described as CIS in the published *Npm1*^{cA} GRH IM model. *RASGRP1* has previously been identified as a gene-expression marker that can be used to predict response to the farnesyl transferase inhibitor, tipifarnib in AML(Raponi et al., 2008) and has been identified as a resistance gene for therapy with MEK inhibitors in a mouse model of AML(Lauchle et al., 2009). *Cblb* is an E3 ubiquitin protein ligase, which transfers ubiquitin to targets, including activated tyrosine kinases. Both *c-CBL* and *CBL-b* mutations have been described in human AML(Caligiuri et al., 2007). *Hecw2* is also believed to have ubiquitin ligase function and although it is not known to have a role in leukaemogenesis, it was recently found to be mutated in a single case of germline *GATA-2* mutation which evolved to MDS/AML(Fujiwara et al., 2014).

Although there is no consensus in the literature on how it should be performed, CIS analysis is the accepted method for analysing insertional mutagenesis screens. However, I have shown that the detailed analysis of tumours with serial sampling and transplant experiments can be a useful complementary approach to defining the driver mutations in an individual tumour. For example in tumour 19.2d, although multiple integrations in CIS genes were identified in the final tumour, only one of these, the integration in *Csf2* was among the top ten hits on Illumina analysis. Additionally, the integrations which persisted on transplantation included one at *Irf2*, which is a plausible driver of this individual tumour. *IRF2* codes for a transcriptional suppressor of type 1 interferon signalling and normally suppresses IFN signalling in HSCs, which is essential for maintaining HSCs in a quiescent state (Sato et al., 2009). IFN-α has been shown to stimulate the proliferation of dormant HSCs *in vivo* and mice deficient for *Irf2* show a reduction in HSC number and an increase in immature progenitor cells (Sato et al., 2009). Furthermore, in the leukaemia cell line TF-1, *IRF2* knock-down was associated with growth inhibition and induction of differentiation

(Choo et al., 2008). Therefore, although *Irf2* was not detected as a CIS gene, it was the integration with the highest read coverage in the primary and all of the recipient tumours in this line and is a likely leukaemia driver in this individual leukaemia.

In conclusion, in this chapter I have shown that the TraDIS sequencing approach is a quantitative method, which allows clonally expanded integrations to be distinguished from the numerous background transposon insertions present in tumour DNA. The integrations that have high read coverage are enriched for the driver integrations, although not all clonally expanded integrations are necessarily drivers. The performance of CIS analysis using only the top 10 or 25 integrations from each tumour allowed identification of a small set of CIS genes which were likely to be significant, while minimising the rate of false positive CIS that could arise if the large number of background mutations were included in the analysis. The quantitative analysis of serial samples allowed identification of additional integrations (e.g. *Irf2*), that were likely to have a driver role, but occurred infrequently across the whole cohort and therefore were not identified on CIS analysis.

Chapter 6: *PiggyBac* Insertional Mutagenesis of the Mature B Cell Compartment

6.1 Introduction

Multiple myeloma (MM) is a plasma cell malignancy that is incurable with conventional therapy and causes nearly 2% of cancer deaths (Jemal et al.). It is preceded by the asymptomatic presence of a monoclonal protein in serum/plasma; termed the monoclonal gammopathy of uncertain significance (MGUS). MGUS occurs in 3% of people over the age of 50 (Kyle et al., 2006) and transforms to MM at a rate of approximately 1% per year (Kyle et al., 2002), but the molecular mechanisms that drive progression are largely unknown.

MM is a heterogenous disease. A hyperdiploid karyotype occurs in approximately 50% of cases, but the driver for chromosome accumulation is not known (Chng et al., 2007). Recurrent chromosomal translocations involving the immunoglobulin loci are found in approximately 70% of non-hyperdiploid tumours (Avet-Loiseau, 2007). These translocations are thought to represent primary oncogenic events that occur in normal B cells during germinal center development. Breakpoints are usually within or near the immunoglobulin switch regions or VDJ sequences (Chng et al., 2007). Recurrent translocation partners include Cyclin D, MAF and MMSET/FGFR3 (Chng et al., 2007). These genetic sub-groups of MM can be used to predict clinical response and guide treatment decisions (Avet-Loiseau et al., 2007; Palumbo and Rajkumar, 2009).

Massive parallel sequencing studies have recently highlighted the remarkable molecular heterogeneity of multiple myeloma and described several additional molecular abnormalities. Deep sequencing of 38 tumour-normal pairs revealed frequent mutations in genes involved in protein translation, histone methylation and blood coagulation(Chapman et al., 2011). Along with previously reported recurrent mutations in MM such as *KRAS*, *NRAS* and *TP53*, this paper described several additional point mutations which may act as driver lesions including *CCND1*, *DIS3*, *FAM46C*, *BRAF* and *IRF4* (Chapman et al., 2011). Another group described further

candidate driver genes including truncating mutations of *SP140*, *LTB*, *ROBO1* and missense mutations in *EGR1*(Bolli et al., 2014). The striking features in all of the sequencing studies performed to date is the heterogeneity of mutational spectra between cases and the large burden of variants within each tumour.

In 2008 Chesi and colleagues described a novel mouse model, which was the first to accurately recapitulate many of the clinical features of human MM (Chesi et al., 2008). In this model the human *c-MYC* transgene was placed under the transcriptional control of the Vk promoter (*Vk*MYC*) and expressed in late B-cells. The third codon of the lead V-kappa exon was mutated to a stop codon (figure 6.1), so although it spliced in frame to the human *MYC* exons, translation of the MYC protein did not occur in tested tissues. The stop codon was engineered to overlap with a preferential target sequence for endogenous Activation Induced Deaminase (AID) and the native kappa light chain gene regulatory elements were maintained to invite targeting by AID. Therefore, it was anticipated that the stop codon would be mutated in a small proportion of B cells during germinal centre development allowing expression of the *MYC* transgene in these cells (Betz et al., 1994; Papavasiliou and Schatz, 2000). With age all mice developed progressive monoclonal plasma cell expansion and as the tumours did not show intra-clonal heterogeneity of B cell receptor sequences, the authors concluded they were not subject to ongoing somatic hypermutation (SHM).

MYC is a global transcriptional regulator that controls cell proliferation, differentiation, growth and survival (Larsson and Henriksson; Meyer and Penn, 2008). Translocations involving *c-MYC* that lead to its inappropriate expression are an initiating event in human Burkitt lymphoma, an aggressive, mature B cell non-Hodgkin lymphoma, but are not unique to this disease (Dalla-Favera et al., 1982) (Au et al., 2004; Kanungo et al., 2005). Notably the incidence of Burkitt lymphoma was low in the *Vk*MYC* mice and no cases of aggressive pro-B lymphoma were reported (Chesi et al., 2008). In mice, *Myc*-Immunoglobulin translocations are an initiating event in plasmacytoma (Ohno et al., 1979; Shen-Ong et al., 1982), whilst constitutive *MYC* expression in early B cells often leads to aggressive pro-B or diffuse high-grade blastic B-cell lymphoma (Adams et al., 1985; Butzler et al., 1997; Chesi et al., 2008; Harris et al., 1988; Kovalchuk et al., 2000; Palomo et al., 1999; Park et al., 2005; Refaeli et al., 2008; Zingone et al.). However, transgenic expression of *MYC* alone does not appear to be sufficient for lymphoma

development as *Eu-Myc* transgenic mice initially demonstrate a benign expansion in pre-B cells (Harris et al., 1988).

Translocations involving *MYC* do occur in human MM, but they are thought to be late progression events(Chng et al., 2007) and are rare in MGUS. In contrast, they occur in 15% of MM tumours and nearly 90% of human myeloma cell lines (Chng et al., 2007). However, mutations or translocations involving the *MYC* locus are not required for *MYC* activation (Meyer and Penn, 2008) and *c-MYC* over-expression due to stimulation by IL-6 and other mechanisms, occurs early in MM(Chesi et al., 2008). Compared to normal bone marrow plasma cells *MYC* expression is significantly higher in cells from MGUS and even higher in MM (Chesi et al., 2008; Zingone et al.). Gene expression profiling data suggests that patients with MM that expresses *N-MYC* or very high levels of *c-MYC* have worse survival (Chng et al., 2007; Janz, 2008).

Transposon insertional mutagenesis (IM) provides a powerful approach for the identification and validation of cancer drivers that compliments human sequencing efforts. In order to further investigate genes involved in the pathogenesis of multiple myeloma I adapted the *Vk*MYC* model to target (*hyper*)*piggyBac* (*hPB*) IM to the mature B cell compartment (figure 6.1).

The *PB* transposon system was chosen because of its efficiency, extensive access to the genome, lack of excision footprint and high rate of intragenic insertions (Cadinanos and Bradley, 2007; Liang et al., 2009; Rad, 2010; Wang et al., 2008). The hyperactive *PB* (*hPB*) cDNA was used, which is an enhanced version of the mouse codon optimised *PB* (*mPB*) but with a ten-fold higher transposition efficiency versus *mPB*, which is itself much more efficient than the native *Trichoplusia Ni* version (Cadinanos and Bradley, 2007; Liang et al., 2009; Yusa et al., 2011).

Two related constructs were generated. In the first, the coding exons of *MYC* were replaced by the *hPB* transposase (figure 6.1B). In the second, *MYC* and *hPB* were expressed together from the same cistron using a *Thosea asigna* 2A (T2A) linker peptide that is hydrolysed very quickly after translation to generate equimolar amounts of the MYC and hPB proteins (Szymczak et al., 2004). In the *Vk*MYC-TA-hPB* mice there was minimal disruption of the original *Vk*MYC* construct as the latter generated highly penetrant plasma cell tumours (Chesi et al., 2008). *T2A-hPB* cDNA

was introduced in frame, after the penultimate codon of *MYC*(thus removing the stop codon), but the remainder of the *Vk*MYC* construct was left intact.

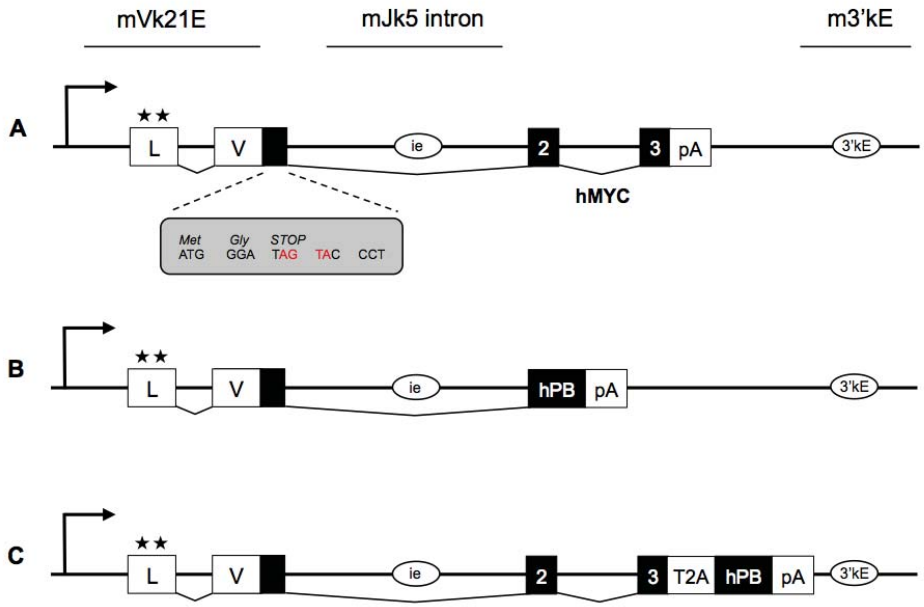


Figure 6.1: The *Vk*MYC*, *Vk*hPB* and *Vk*MYC-TA-hPB* constructs. This figure is repeated from Chapter 2 (figure 2.1) for ease of reference

(A) In the *Vk*MYC* construct published by Chesi et al, the Jk5 exon in the rearranged mouse Vk21 kappa light chain gene was replaced by a short coding exon containing a Kozak ATG(Chesi et al., 2008). Human *MYC* exons 2 and 3 replaced the C1k region. Transcription initiates at the Vk21e proximal promoter (↱), extends to the leader (L) and Vk (V) exons, splices in frame to human *MYC* (*hMYC)* and terminates at the endogenous polyA signal (PA). ATG codons (*) in L were mutated to ACG to stop initiation of translation at these positions. Intronic (ie) and 3'kappa (3'kE) enhancers are maintained. The DNA sequence immediately downstream of the Vk21 ATG is depicted. Nucleotides in red letters fit the DGYW consensus for AID targeting.

(B) In *Vk*hPB, hMYC* is replaced by the *hPB* cDNA, carrying a splice acceptor signal that leads to splicing of *hPB* mRNA in-frame with the reading frame "opened" by AID mutation of the upstream TAG stop codon.

(C) In *Vk*MYC-TA-hPB* the cDNA for the self-cleaving peptide T2A links *hPB* in-frame to *hMYC*. The chimaeric polypeptide produced from a single cistron is predicted to spontaneously dissociate into hMYC and hPB proteins.

The decision to generate two models was taken to enable comparisons between tumours derived in the presence and absence of MYC and because it was uncertain whether the *MYC* transgene would be critical for the development of MM. As the Vk regulatory elements in combination with the early stop codon requiring mutation by SHM should have, in principle, ensured that transgene activation was restricted to late B-cells, it was considered probable that *hPB* could drive the development of MM as well as other mature B cell tumours such as follicular or Burkitt lymphoma (figure 6.2 (Weigert and Weinstock, 2012)). Information on the role of *MYC* in initiating and driving these tumours would therefore be derived by comparing the tumour phenotype, latency and transposon integrations between the *Vk*MYC-TA-hPB* and *Vk*hPB* cohorts.

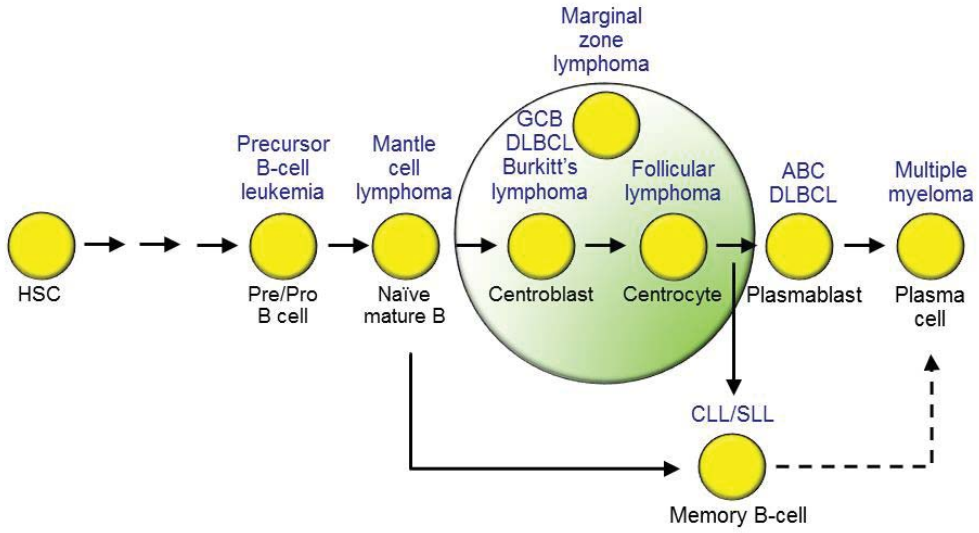


Figure 6.2: B cell maturation and lymphoma phenotypes. The stages of normal maturation are indicated in black and the corresponding neoplasm in blue text. The germinal centre is represented by the area shaded in green. GCB = germinal centre B cell like; ABC = activated B cell like; DLBCL = diffuse large B cell lymphoma; CLL/SLL = chronic lymphocytic leukaemia/small lymphocytic lymphoma. From Weigert and Weinstock, 2012.

6.2 Results

6.2.1 Cloning Vk*HPB and Vk*MYC-TA-HPB

The *Vk*MYC-TA-HPB* and *Vk*HPB* constructs were generated as described in Materials and Methods, sequenced by Sanger sequencing, linearised and sent for pronuclear injection at PolyGene AG (Switzerland). The *hPB* sequence in *Vk*MYC-TA-hPB* was identical to the expected sequence. The *Vk*hPB* construct had an S to G substitution at position 1520 in *hPB*. This corresponds to an intermediate stage of *hPB* development between *mPB* and the final *hPB,* and is expected to have a very slightly reduced transposition efficiency compared to the final *hPB* version, but still higher than mPB (Kosuke Yusa personal communication).

Three of the 46 tail samples received after pro-nuclear injection were positive for the *Vk*MYC-TA-hPB* transgene and four of 40 samples were positive for the *Vk*hPB* construct by PCR analysis. The seven founder (F0) mice were imported to our quarantine facility and colonies generated for re-derivation inside the WTSI animal facility. One line from each construct was selected for expansion and mating with low copy *GrOnc* transposon lines. This decision was primarily based on the fecundity of initial matings.

6.2.2 Validation of splicing in the transgenic constructs

Correct splicing of the constructs was initially confirmed by RT-PCR of RNA from U266 cells transiently transfected with each of the constructs. PCR products using primers annealing to exon 2 of *MYC* and to *hPB cDNA* are shown in figure 6.3. For the *Vk*HPB* construct, the V exon forward primer (VkMycexon1F) and the reverse primer in *hPB* (SAHPBR) also generated the expected RT-PCR product (190bp) (image not shown). The splicing was subsequently confirmed on mouse tumour RNA samples using the same method and further verified using capillary sequencing of PCR products.

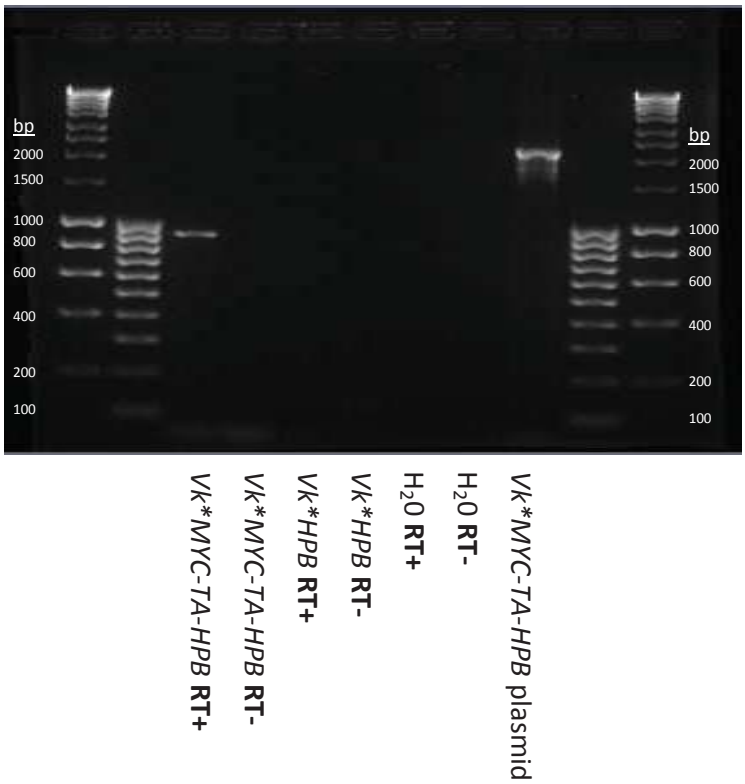


Figure 6.3: RT-PCR using primers VkMycexon2F and SAHPBRCORTAHPBR. RNA from the *Vk*MYC-TA-hPB* transfected U-266 cells generated the expected 900bp RT-PCR product. By contrast when *Vk*MYC-TA-hPB* plasmid DNA was used as a template a 2200bp PCR product was generated. As expected, no RT-PCR product was amplified from non-reverse transcribed RNA (RT-) or from RNA from Vk*HPB transfected cells.

6.2.3 The *Vk*MYC-TA-HPB* construct generates an active *PB* transposase: HAT resistance assay

After hydrolysis, the T2A linker leaves a single proline at the 5' end of *hPB(Szymczak et al., 2004)*. To test if this affected *PB* transposase activity, I performed a HAT resistance assay. In this assay, when an active transposase removes a *PB* transposon from within the X-linked *Hprt* gene locus of male ES cells, *Hprt* activity is restored permitting growth in HAT media. Transfection with the Vk**MYC-TA-hPB* cDNA resulted in growth of HAT resistant colonies, in numbers proportional to the amount of transfected DNA. Colony numbers were comparable to those seen after transfection with *hPB* cDNA (positive control), confirming that an active transposase is generated from the bi-cistronic *MYC-TA-hPB* construct (Figure 6.4).

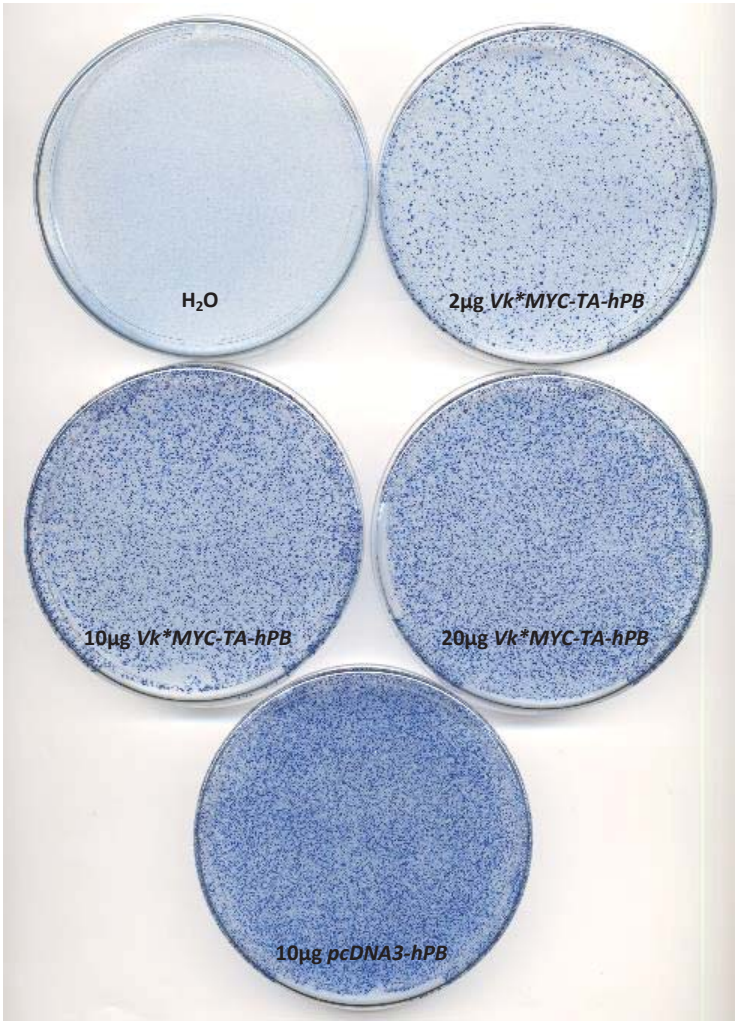


Figure 6.4: HAT resistance assay. Male ES cells harbouring a *PB* transposon in the *Hprt* locus were electroporated with H_20 or constructs Vk**MYC-TA-hPB* or *pcDNA3-hPB* (positive control) using the indicated amounts of plasmid. Photographs are of colony growth from 1/10 platings ($1/10^{th}$ of $1x10^7$ electroporated cells plated).

6.2.4 The *hPB* transposase is active *in vivo*, although transposon mobilisation is not limited to the mature B cell compartment

The *Vk*Myc-TA-hPB* and *Vk*hPB* constructs were designed to specifically express *hPB* protein in the mature B cell compartment. This specificity would be imparted by the *Vk* regulatory elements and the presence of the early stop codon, designed to prevent expression of the *hPB* transposase in the absence of mutation by SHM. The stop codon was positioned such that it created a preferential target sequence for AID and it was anticipated this would be reverted in a small percentage of B cells during germinal centre development as previously described (Chesi et al., 2008). However, using "jump" PCR, we found that mobilisation of the transposon was also occurring in non-haematopoietic tissues (figure 6.5) and some insertional mutagenesis mice were observed to develop non-haematopoietic tumours. Furthermore, when spleen and bone marrow cells from IM mice were flow sorted and DNA was extracted separately from CD34+ (progenitor), CD3+ (T), Gr1+/Mac1+

(granulocytes), B220+/CD19+ (pro/mature B) and B220+/CD19- (pre-pro B) cells, the 'jump' PCR was positive in all lineages tested in the *MYC-TA-hPB* mouse and in all cells except granulocytes in the *hPB* mouse tested.

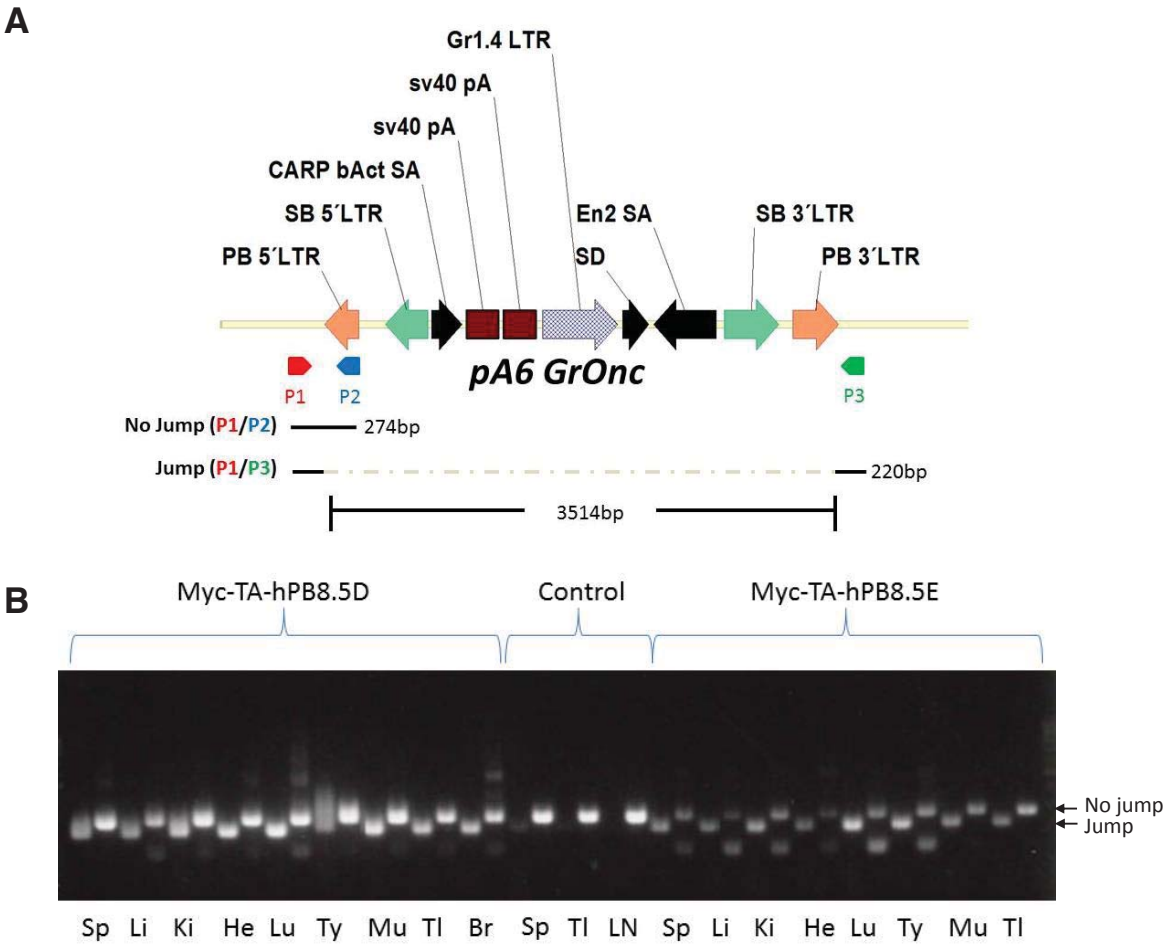


Figure 6.5: Tissue specificity of the *hPB* transposon. (A) Design of 'jump' and 'no jump' PCRs. The yellow line represents the plasmid backbone. When one or more transposons mobilise, the region flanked by the PB repeats is removed and a 220bp product is generated by the P1 and P3 primers (Jump PCR). In the absence of jumping these primers are separated by the full length of the *GrOnc* transposon (3.5kb) and no PCR product is generated. By contrast, the P1 and P2 primers will produce a 274bp product only when one or more transposons do not mobilise (No jump PCR). (B) Results from alternating 'jump' and 'no jump' PCRs on DNA extracted from various tissues from two *Vk*Myc-TA-hPB* IM mice (8.5D and 8.5E) and a control mouse which carried the *GrOnc* transposon, but no transposase. Sp-spleen, Li-liver, Ki-kidney, He-heart, Lu-lung, Ty-thymus, Mu-muscle, Tl-tail, Br-brain, LN-lymph node.

6.2.5 Insertional mutagenesis mice have increased lymphoma-associated mortality

The *Vk*MYC-TA-hPB* and *Vk*hPB* IM mice were born at expected Mendelian ratios. They died at a similar rate (p=0.57), which was significantly accelerated compared to non-IM mice (median survival 75.4, 71.9 and 91.1 weeks respectively) (figure 6.6). Of note, the survival of the *Vk*MYC-TA-hPB* only mice was also no different to the GRL (GrOnc) only mice which lacked the *MYC* transgene (median survival 94.3 v 93.2 weeks, p = 0.61 Mantel-Cox test). Intraperitoneal injection of sheep red blood cells did not alter the survival or tumour spectrum and mice were considered together in the results, regardless of whether or not they received antigen stimulation.

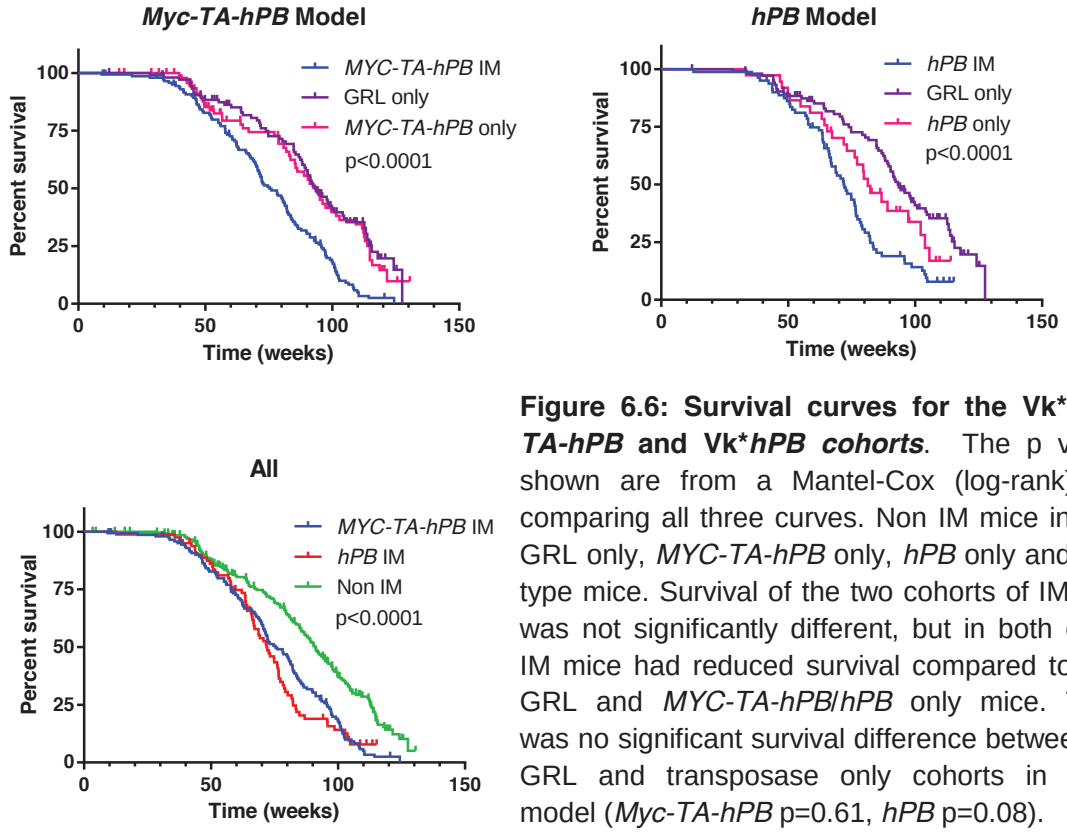


Figure 6.6: Survival curves for the Vk**MYC-TA-hPB* **and Vk****hPB cohorts***. The p values shown are from a Mantel-Cox (log-rank) test comparing all three curves. Non IM mice include GRL only, *MYC-TA-hPB* only, *hPB* only and wild-type mice. Survival of the two cohorts of IM mice was not significantly different, but in both cases IM mice had reduced survival compared to both GRL and *MYC-TA-hPB*/*hPB* only mice. There was no significant survival difference between the GRL and transposase only cohorts in either model (*Myc-TA-hPB* p=0.61, *hPB* p=0.08).

Histopathology assessment was performed on 170 of the first 184 mice in the *Vk*MYC-TA-hPB* cohort culled due to illness or found dead. Of the fourteen cases that were not reviewed, eight mice were found dead and considered too decomposed for useful histological analysis. Histopathology samples were collected in three cases but blocks and slides could not be located. No histopathology or

necropsy records were identified for two mice. One mouse was culled due to an ulcerated eye but had no other abnormality. Of the 170 cases reviewed 100 were insertional mutagenesis mice, 40 had *Vk*Myc-TA-hPB* only and 30 had *GRL* only. The spectrum of tumours is shown in figure 6.7.

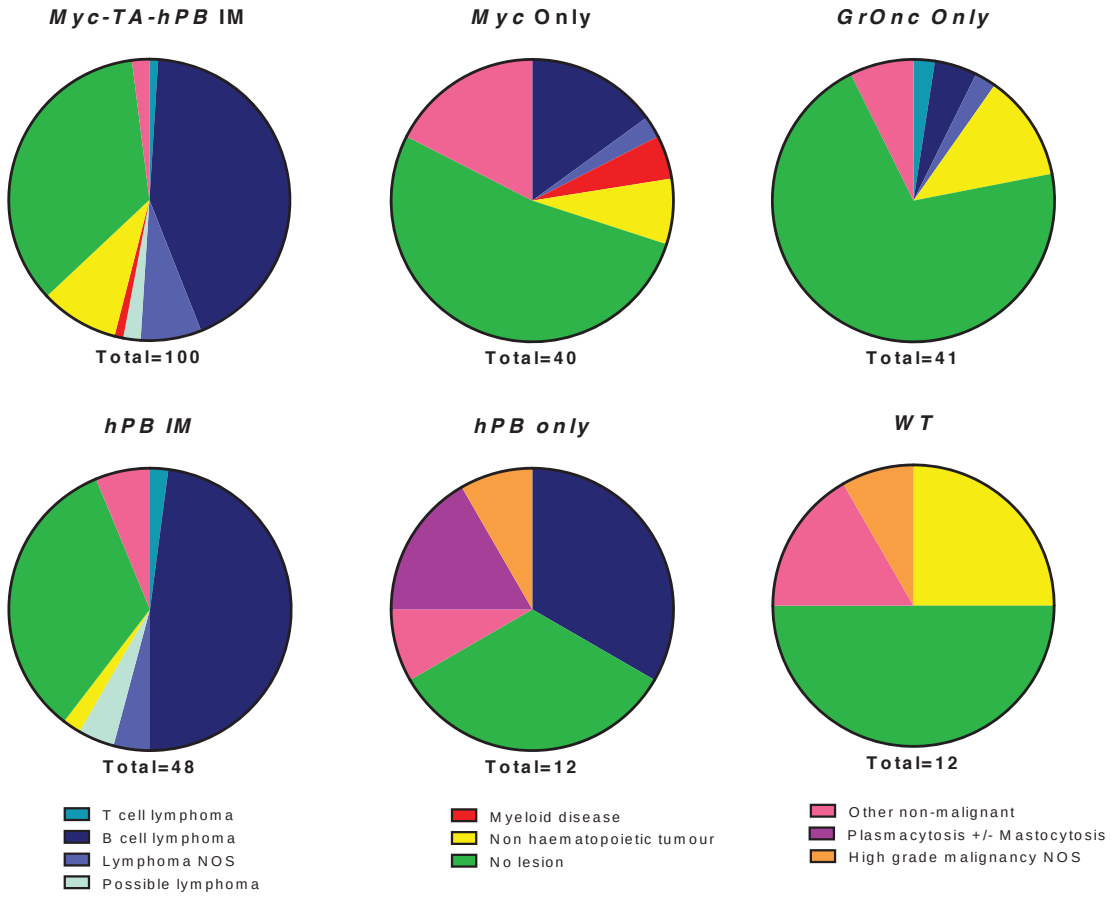


Figure 6.7: Diagnoses in the *Vk*MYC-TA-hPB* and *Vk*hPB* mice. The disease classification is based on independent review by a histopathologist (who was blinded to genotype) and immunophenotyping where this was performed. The *GRL* (*GrOnc*) only mice were pooled between the two cohorts. The diagnoses in twelve wild-type littermates which were aged along with the study animals are also shown. The *hPB* group includes one mouse that had a benign lesion only that was culled at 97.7 weeks for experimental reasons. Similarly, one of the *hPB* IM mice was culled for experimental reasons at 94.1 weeks of age but was found to have lymphoma.

In the *Vk*-hPB* cohort histopathology examination was performed on 69 of the first 96 mice to be culled sick or found dead and a further two mice which were culled for experimental reasons. Of the mice that did not have histopathology examination, two

were not necropsied (died aged 12.7 and 48.4 weeks). A further thirteen were found dead and considered too decomposed for useful histological analysis. The histopathology blocks and slides could not be found in four cases. In the remaining eight mice, histopathology samples were not recorded as received by our tissue bank.

Vk-Myc-TA-hPB mice*

Of the 100 *Myc-TA-hPB* IM mice on which histopathology was performed, 51 were reported to have lymphoma. Of these, six had additional non-haematopoietic tumours (carcinomas (4), sarcoma (1), probable neural or smooth muscle tumour(1)), two had focal increases in plasma cells in the bone marrow and one had a lymph node plasmacytosis of uncertain significance. Overall, there was no evidence of multiple myeloma in any of the *Vk*Myc-TA-hPB* IM mice.

Immunohistochemistry (IHC) staining for B220 and CD3 was performed on 46 of the mice with lymphoma and 43 were determined to have B cell lymphoma based on this staining. The other three lymphomas were of uncertain lineage. Flow cytometry was performed on spleen cells from two of these and one was found to be a T cell tumour, while in the other the spleen was not definitely involved with lymphoma and flow cytometry was unhelpful. The third tumour was reported to have a sclerotic pattern with giant cells on histopathology, more similar in morphology to human Hodgkin lymphoma. IHC was also performed on a further two mice that were called 'possible lymphoma' on initial histopathology review, but a definitive diagnosis could still not be made. Both of these mice were culled sick; one had a swollen abdomen with pale liver and splenomegaly (0.6g), the other was culled due to respiratory difficulties and had marginal thymomegaly with a spleen size of 0.34g.

Of the remaining 47 *Myc-TA-hPB* IM mice, only nine were diagnosed with other malignancies. These included myeloid leukaemia (1), squamous cell carcinoma (2), skin appendage carcinoma and hydronephrosis with probable carcinoma in the urinary tract (1), sarcoma (2) and papillary adenocarcinoma (3) of which one was also reported to have evidence of a myeloproliferative disease in the bone marrow. The remaining 38 mice were reported as having no (35) or benign lesions only (3). Of these, nine mice were found dead at ages of between 33 and 63.1 weeks. In one of these cases the spleen was enlarged (0.66g) and a large abdominal mass was

found at necropsy which is suggestive that this mouse also had lymphoma, however the tissue was too autolysed to allow a histopathological diagnosis. In the remaining eight mice that were found dead, the spleen was 0.1g or less and no masses or lymphadenopathy were identified at necropsy. The other 27 mice in which no abnormality was diagnosed on histopathology were culled due to illness. The commonest reason was a swollen abdomen (13 mice) but at necropsy the only abdominal finding was enlarged seminal vesicles in the majority of these cases. Some had a very full bladder and one had a distended bowel without an overt mass. The other mice were culled sick due to tachypnoea, severe scratch marks, piloerection, being hunched and inactive and one mouse each with limping and anal prolapse. Only one was reported as moribund by the animal technicians and this animal was found to have a left inguinal mass and splenomegaly (0.63g) at necropsy but no diagnosis could be reached on histopathology assessment. The spleen was less than 0.22g in weight in all of the remaining mice that were culled sick and reported as having 'no lesion' on histopathology. Figure 6.8 shows the spleen weight at death for the various cohorts. A bimodal distribution of spleen sizes across the *Vk*MYC-TA-hPB* cohort is evident, which largely represents cases with and without lymphoma.

Vk-hPB mice*

Histopathology assessment was performed on 48 IM mice from the *Vk*hPB* cohort, including one which was culled for experimental purposes and had features of lymphoma at death. Of these, 26 were found to have lymphoma on review of the H&E slides. Immunohistochemistry, flow cytometry or both was performed on 25 of these samples and the lineage was confirmed as B cell in 23 cases. In the most striking case of B cell lymphoma the spleen was 11.8g and there were also liver lesions, thymomegaly, and a mesenteric mass (figure 6.8 and 6.9). One mouse with B cell lymphoma was also noted to have a plasmacytosis in the thoracic and mesenteric lymph nodes and the liver. One of the remaining cases showed double staining for CD3 and B220 on IHC but the flow cytometry of the spleen suggested this was a T cell tumour (66% CD4+, 32% CD19+/B220+). Of the two lymphoma cases not assigned a lineage one was thought to be a B cell lymphoma based on morphology and distribution, however it has not been assessed by immunophenotyping. The flow cytometry of the other case is most suggestive of a T

cell tumour, but around 40% of spleen cells were negative for all markers used (CD3, CD19, B220, Mac1 and Gr1). There were two further mice which were reported as 'possible lymphoma', but this diagnosis was not confirmed even after flow cytometry and IHC. These mice both had enlarged spleens (2.07g and 1g) and one had widespread lymphadenopathy. Of the remaining IM cases, no lesion was identified in sixteen and none of these had macroscopic features of lymphoma. There were two cases with papillary adenomas of the lung and one with extensive invasive adenocarcinoma of the lung. The final case had unusual morphology with reactive and apoptotic changes and it was concluded this was most likely reactive.

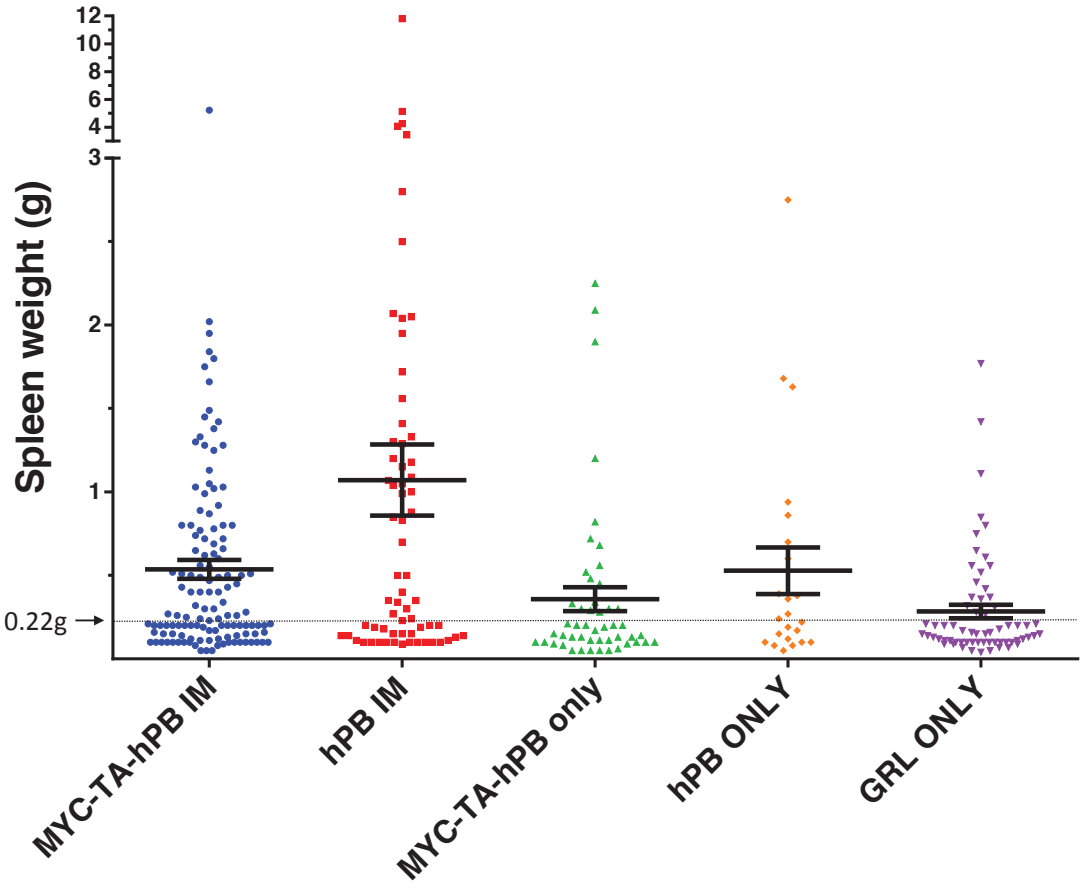


Figure 6.8: Spleen weight for the various colonies. Each mouse is represented by a point. The mean and the standard error of the mean for each colony are shown in black. The difference between *hPB* IM and *MYC-TA-hPB* IM is statistically significant (p=0.0022). For the *hPB* IM and *MYC-TA-hPB* IM mice the difference versus *GRL* only mice was statistically significant (p=0.0004 and p=0.003 respectively), but not quite so versus the *hPB* only (p=0.15) or *MYC-TA-hPB* only (p=0.0785) mice.

In total there were 41 GRL only (*GrOnc*) mice that had histopathology assessment between these two cohorts. Of these, no lesion was identified on histopathology in 29 cases, including one which was culled due to an eye defect and was not considered to have reached the survival endpoint. Of the cases in which no lesion was identified, nine were found dead. One of these had an enlarged spleen (0.52g) with no other findings at necropsy, but the tissues showed autolysis making it unsuitable for histopathology. Four of the remaining 12 *GrOnc* mice had lymphoma; two B cell, one T cell and one of uncertain lineage which was negative for B220 and CD3 on IHC but involved the thymus, heart, lung, kidney, spleen, liver, lymph nodes and blood. One of the mice with B cell lymphoma also had a focal increase in plasma cells in the BM and protein deposits in the glomeruli. Five of the GRL mice had non-haematopoietic tumours, one a lung adenoma and one fatty liver only. The final mouse had thoracic lymphadenopathy and splenomegaly, but these changes were thought to be reactive based on histopathology and IHC.

Of the 40 mice with the *Vk*MYC-TA-hPB* transgene but no *GrOnc*, only seven were found to have lymphoma. IHC was performed in six of these and confirmed B cell lymphoma in all cases. Two of these had focal areas of plasmacytosis in the bone marrow in addition to the lymphoma. No lesion was identified on histopathology in a further 21 mice and all had spleens that weighed less than 0.35g at death. Diagnoses in the remaining twelve mice were; i)probable myeloproliferative disorder (MPD) ii) chronic inflammation and immune complex glomerulopathy iii) groin abscess iv) periarteritis in lung and muscle v) lung adenoma only (n=3) vi) lung adenocarcinoma (n=3) vii) transitional cell papilloma of kidney and viii) possible myeloid leukaemia.

To date, histopathology examination has been performed on twelve mice with the *Vk*hPB* transposase alone. Of these four had B cell lymphoma, four had no lesion, and one each had plasmacytosis with mastocytosis, a skin lesion of uncertain aetiology with a plasmacytosis in a nearby lymph node, a high grade malignant mass in the abdomen of uncertain lineage and a lung adenoma. One of the mice with lymphoma was also noted to have plasmacytosis in the thoracic lymph nodes.

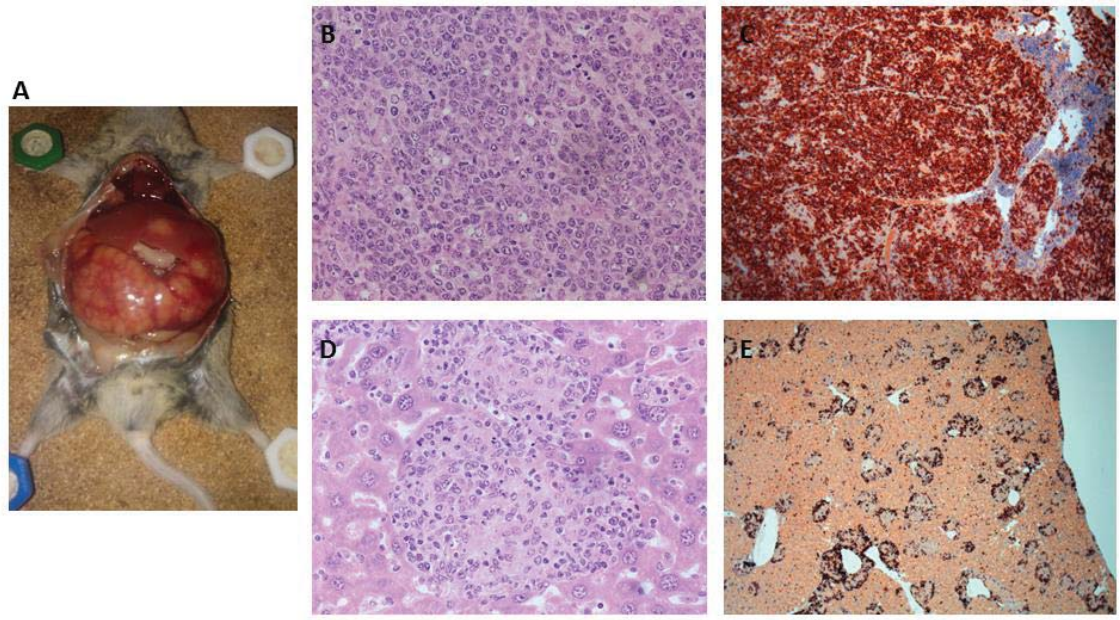


Figure 6.9: Mouse *Vk*hPB* IM 4.1c. (A) Necropsy findings included an 11.8g spleen and 4.75g liver. (B – E) Histopathology shows a diffuse infiltrate of B cells in the spleen [x200 H&E (B) and x 100 B220 IHC (C)] and peri-portal infiltrates in the liver [x400 H&E (D) and x 50 B220 IHC(E)].

Twelve wild-type littermates were aged along with these mice and tissues were taken for histopathology at death. Of these only one had a possible haematopoietic malignancy; an undifferentiated malignant infiltrate involving the liver, lymph node, lung, spleen, blood and kidney, which was negative for B220, CD3 and MPO on IHC.

6.2.6 Immunophenotyping to determine developmental stage of the B cell tumours

The histopathology and immunophenotyping confirmed an increase in B cell malignancies in the two insertional mutagenesis cohorts. However, the morphology of these lymphomas was variable and included follicular, diffuse and sclerotic tumours and some that varied in pattern between different regions of the same tumour. The cell size also varied from small to large cell both between and within tumours. Some lymphomas also had leukaemic changes, including some with blastic morphology.

With the assistance of Dr George Giotopolos (Huntly laboratory, University of Cambridge) I performed flow cytometry on CD19 and B220 positive tumours using CD24, CD43, AA4.1 (CD93), BP-1, IgM and IgD to determine what stage of B cell development these tumours correspond to (figure 6.10). Nine *Vk*MYC-TA-hPB* and 12 *Vk*hPB* IM tumours were selected for further analysis along with one *Vk*MYC-TA-hPB* only, one *Vk*hPB* only and two *GrOnc* only tumours. These included B cell lymphomas which were described on morphology as mature and immature, small and large cell and follicular, nodular and diffuse in pattern.

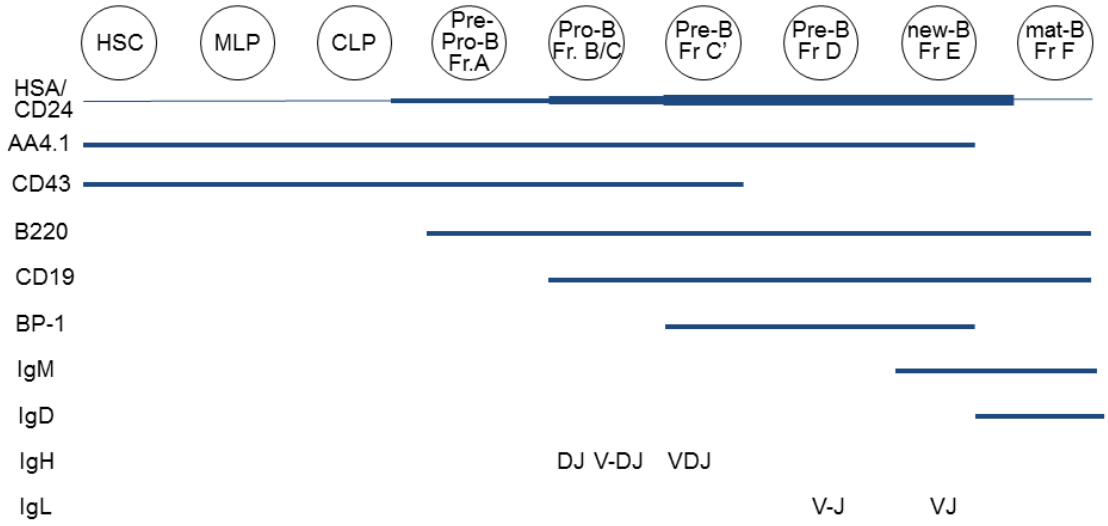


Figure 6.10: Stages of B cell development in mouse bone marrow. B cell development stages are shown according to the Hardy classification system (Hardy and Hayakawa, 2001). Corresponding surface antigens are indicated along with the timing of immunoglobulin gene re-arrangement for the heavy (IgH) and light (IgL) chain genes. Adapted from Hardy and Hayakawa, 2011.

Three samples described as 'blastic' or 'poorly differentiated' were selected for flow cytometry; two from the *Vk*MYC-TA-hPB* and one from the *Vk*hPB* IM cohorts. Of these, two were clearly positive for B220 and CD19, while one of the *Vk*MYC-TA-hPB* samples was positive for B220 but not for CD19 (figure 6.11A). Unfortunately, further flow cytometry could not be performed on the *Vk*MYC-TA-hPB* 12.4a and *Vk*MYC-TA-hPB* 10.1h samples due to a lack of viable cells. The other sample with blastic morphology *Vk*hPB* 5.1e was strongly positive for CD24, predominantly negative for BP-1 and showed variable staining for AA4.1, CD43, IgM and IgD. This

suggests that a significant proportion of the B cells in this tumour had a mature B cell phenotype (figure 6.11B).

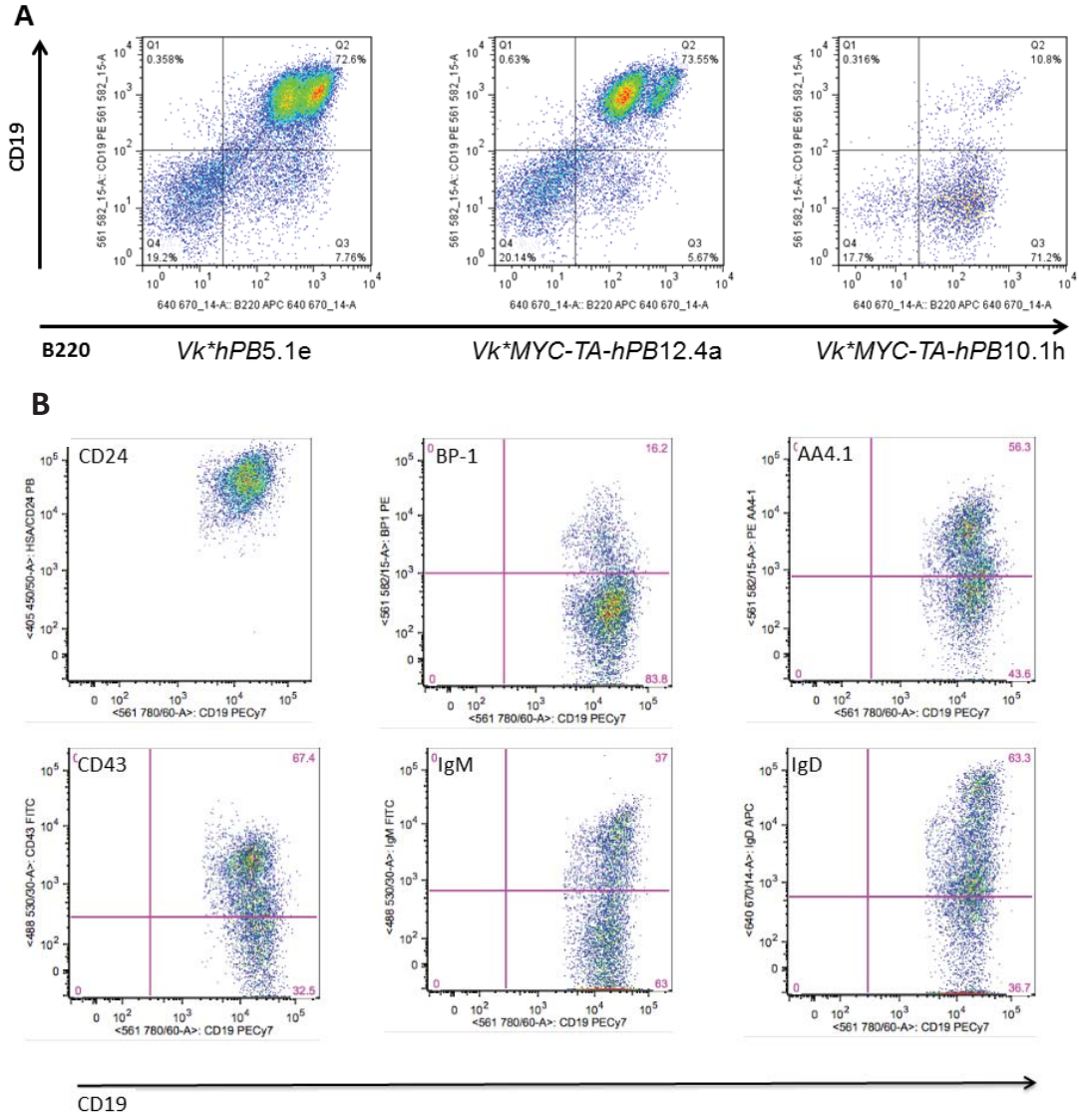


Figure 6.11: Flow cytometry of spleen samples from three mice with blastic morphology (A) CD19 (vertical axis) and B220 (horizontal) staining on the three mice **(B)** Further analysis of *Vk*hPB*5.1e. Clockwise from top left CD24, BP1, AA4.1, IgD, IgM and CD43 on the vertical axis; all shown against CD19 (horizontal).

The nine *Vk*MYC-TA-hPB* IM and one *Vk*MYC-TA-hPB* only tumours investigated using the extended flow panel were uniformly negative for AA4.1 and BP-1 and positive for CD24. Representative flow cytometry results are shown in figure 6.12. IgM was positive in most cases and the IgD staining varied from weak to strong but was positive in all cases, with the exception of *Vk*MYC-TA-hPB* 10.4f (figure 6.13),

which was reported as a large cell high grade lymphoma on morphology. The CD43 was negative in most cases, including tumours classified as both small and large cell, and low and high grade on morphological appearance. However, in samples *Vk*MYC-TA-hPB* 14.1G, the *Vk*MYC-TA-hPB* only mouse, and *Vk*MYC-TA-hPB* 12.1C (figure 6.12), there was some weak CD43 staining. A third sample also had a small number of B cells with weak CD43, although the majority were negative. All of these tumours were described as follicular lymphomas on histopathology, although 12.1C was also noted to have sclerotic changes. Despite these differences, in general, the flow patterns were similar across the various morphological subtypes. Therefore, flow cytometry analysis showed that the majority of *Vk*MYC-TA-hPB* B cell malignancies had a mature B cell phenotype.

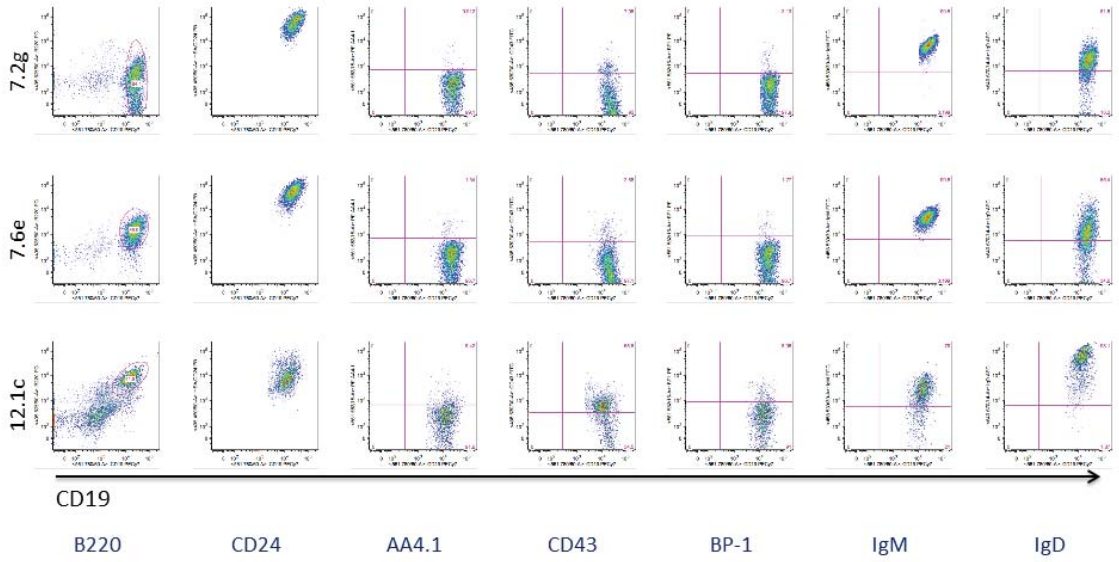


Figure 6.12: Representative flow cytometry from *MYC-TA-hPB* IM cases. The samples are *Vk*MYC-TA-hPB* 7.2g (top), *Vk*MYC-TA-hPB* 7.6e (middle) and *Vk*MYC-TA-hPB* 12.1c (bottom). Each tumour is represented in a separate row. From left to right the columns show; i)gating on B cells by B220 and CD19, ii) CD24, iii) AA4.1, iv) CD43, v) BP-1, vi) IgM and vii) IgD. CD19 is shown on the horizontal axis in all.

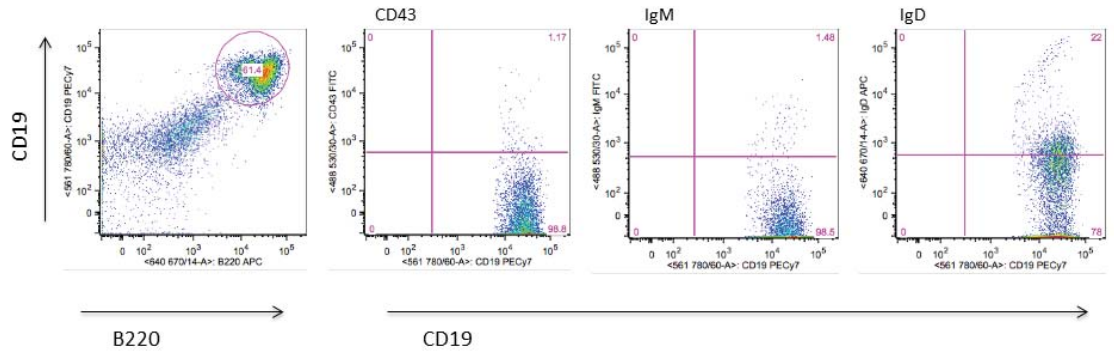


Figure 6.13: *MYC-TA-hPB*14.3f. Like most *Vk*MYC-TA-hPB* samples 14.3f was negative for CD43, but In contrast to the other samples it was negative for IgM and had negative/weak IgD.

The flow cytometry of the twelve *Vk*hPB* IM samples was more variable. Although these samples were uniformly positive for CD24, there was a spectrum of CD43 staining, with some samples clearly positive (e.g. *Vk*hPB* 4.2b, *Vk*hPB* 6.2l), others negative (e.g. *Vk*hPB* 5.1d, *hPB* 7.7d) and most showing a mixture of positive and negative cells (figure 6.14). *Vk*hPB* 4.2b and 6.2l were both uniformly positive for AA4.1 and 6.5c and 5.1e were positive in a significant proportion of cells, however all the other samples were negative. BP-1 was negative in the majority of B cells in all cases, but two had BP-1 positivity in up to 30% of cells (4.2b and 6.5c). Regardless of their CD43 expression, the *Vk*hPB* IM tumours were mostly IgD positive. Even in samples 4.2b and 6.2l, a number of cells were IgD positive (figure 6.15). The fact that these tumours are positive for IgD suggests they have a mature B cell phenotype, but CD43 is usually negative from the late pro-B stage of development. Therefore these tumours are not easy to classify in the spectrum of normal B cell maturation (figure 6.10). The finding of CD43 with sIgM and IgD may reflect an aberrant phenotype or is consistent with these being B1 cells(Wells et al., 1994).

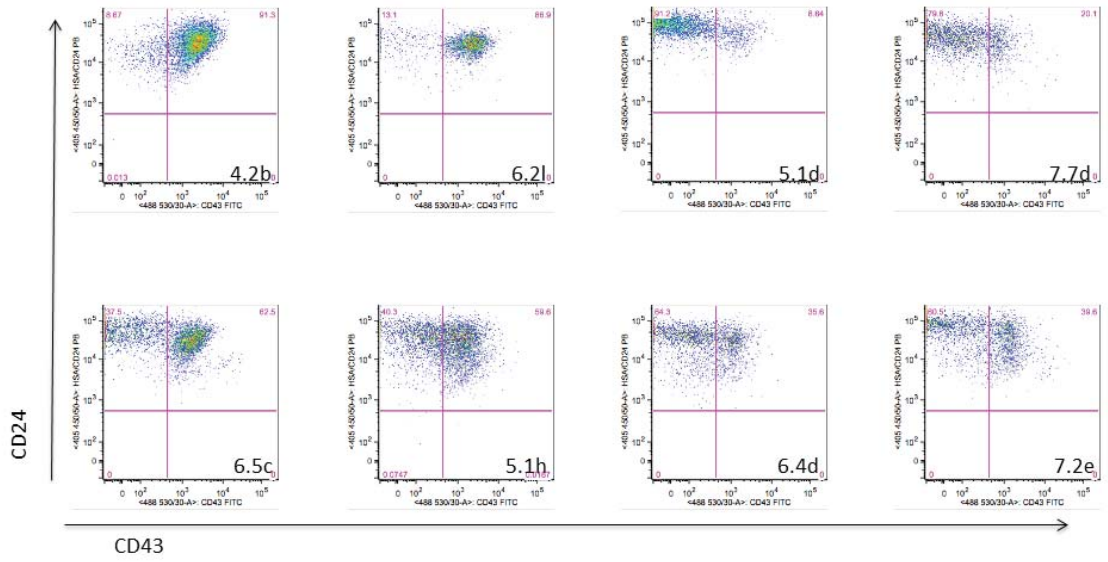


Figure 6.14: CD24 and CD43 flow cytometry on *hPB* IM samples. All analyses were performed on CD19 and B220 double positive cells. The mouse ID is indicated in each. Although all samples were CD24 positive there was significant variation in CD43.

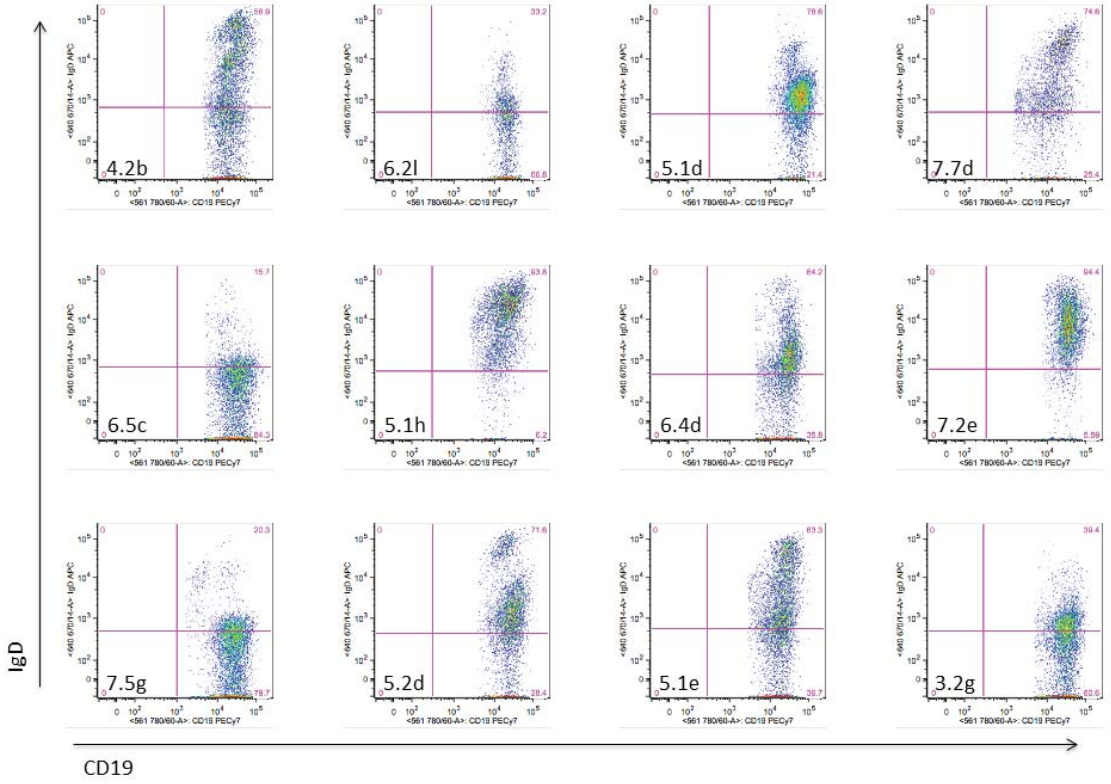


Figure 6.15: IgD expression in the *Vk*hPB* tumours. The data and mouse ID are shown for each of the twelve *Vk*hPB* IM mice analysed with the extended flow panel.

6.2.7 The MYC-TA-hPB tumours are not universally MYC dependent

In order to establish if the tumours in the *Vk*MYC-TA-hPB* cohort were *MYC*-dependent we performed Western blotting using various anti-Myc antibodies. This work was performed with Nicla Manes. Although both the human and mouse cMyc were detected by the appropriate antibodies in positive control lysates, neither was detected in the mouse samples despite abundant protein (figure 6.16). Western blotting for the *PB* transposase was also attempted using an in-house antibody, but this was unsuccessful.

Immunohistochemistry was performed using the same antibody as described in the Chesi paper, which detects both mouse and human Myc(Chesi et al., 2008). *MYC* was found to be expressed in many, but not in all of the *Vk*MYC-TA-hPB* IM tumours (figures 6.17 – 6.19). The *Vk*MYC-TA-hPB* IM samples which were described as having immature morphology typically had a high proportion of Myc positive cells (e.g. 10.1h, 12.4a and 3.4c) (figures 6.17 and 6.18). Many of the *GrOnc* only control mice or IM mouse from the *Vk*hPB* cohort did not stain strongly for Myc, but some of these tumours were positive (figures 6.20). Eight samples from *Vk*MYC-TA-hPB* mice that lacked the *GrOnc* transposon were also stained. Four of these had no lesion detected on histopathology and none of these had increased Myc staining. Of the four samples that were reported as lymphoma, only one had a noticeable increase in Myc staining (figure 6.21). Attempts to perform IHC using an antibody specific to human Myc were unsuccessful.

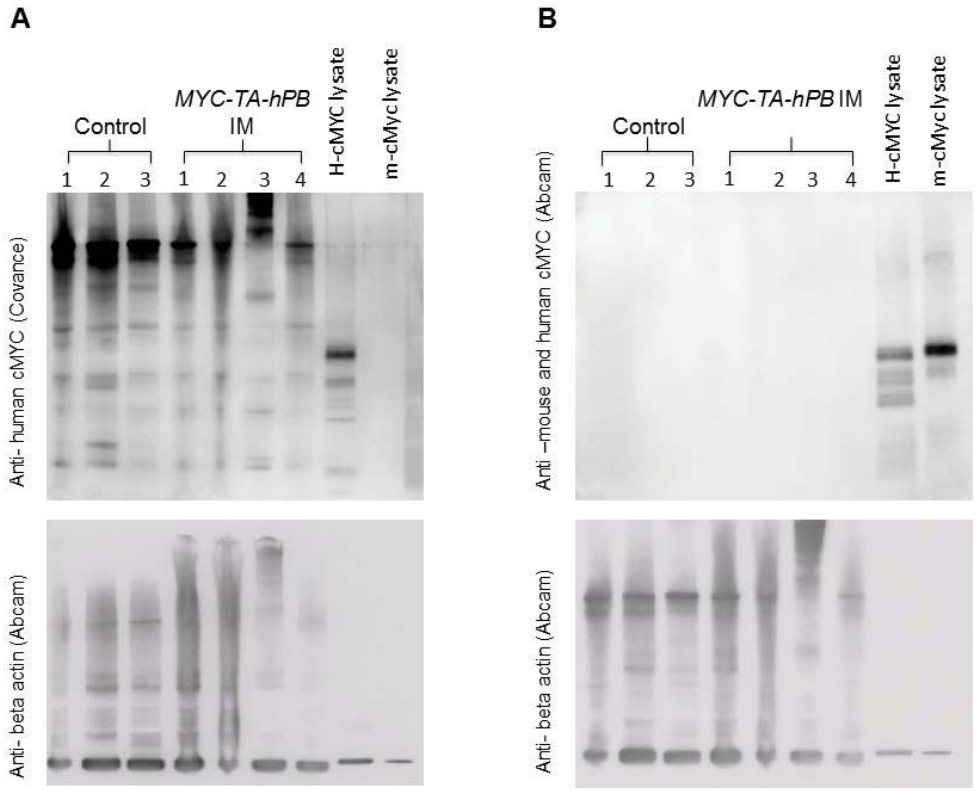


Figure 6.16: Western Blot for human and mouse Myc protein. **(A)** The Covance antibody is specific for human Myc. Despite a strong signal in the control lysate, no human MYC protein was detected in the transgenic IM mice. However, neither human or mouse MYC protein was detected with the non-specific Myc antibody in any of the tumour samples, although it was positive in control lysates **(B)**. The bottom panels show the beta-actin staining, indicating adequate protein was present. The control samples were a lymphoma from a *Ras*^{G12D} GRL insertional mutagenesis mouse (1), and tumours from GrOnc only mice from this study (2 and 3). The *Vk*MYC-TA-hPB* IM lysates were all from spleen samples from mice with lymphoma, including three with follicular morphology (1-3) and one with blastic morphology (4). These blots were generated by Nicla Manes.

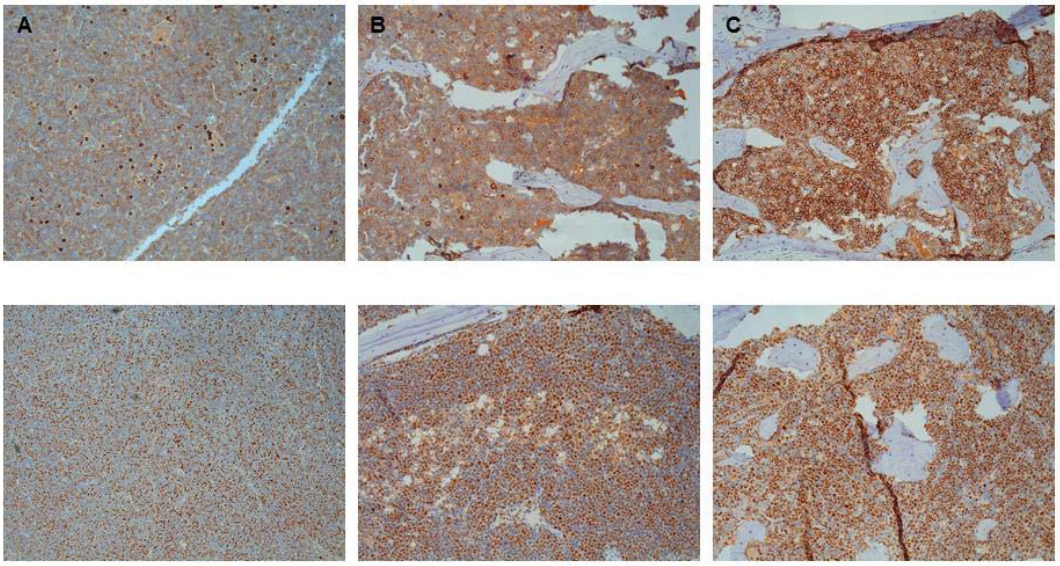


Figure 6.17: IHC staining of three *Vk*MYC-TA-hPB IM* tumours. The top row shows B220 staining and the bottom c-Myc. (A) Axillary lymph node from *Vk*MYC-TA-hPB* 10.1e, which has described as a large cell lymphoma by our histopathologist. The majority of cells are positive for B220 and many are also positive for c-Myc. Bone marrow samples from *Vk*MYC-TA-hPB*10.1h (B) and *Vk*MYC-TA-hPB*3.4c (C) showing almost complete replacement of the marrow with B220 and c-Myc positive cells. Both of these tumours were described as blastic on morphological appearance.

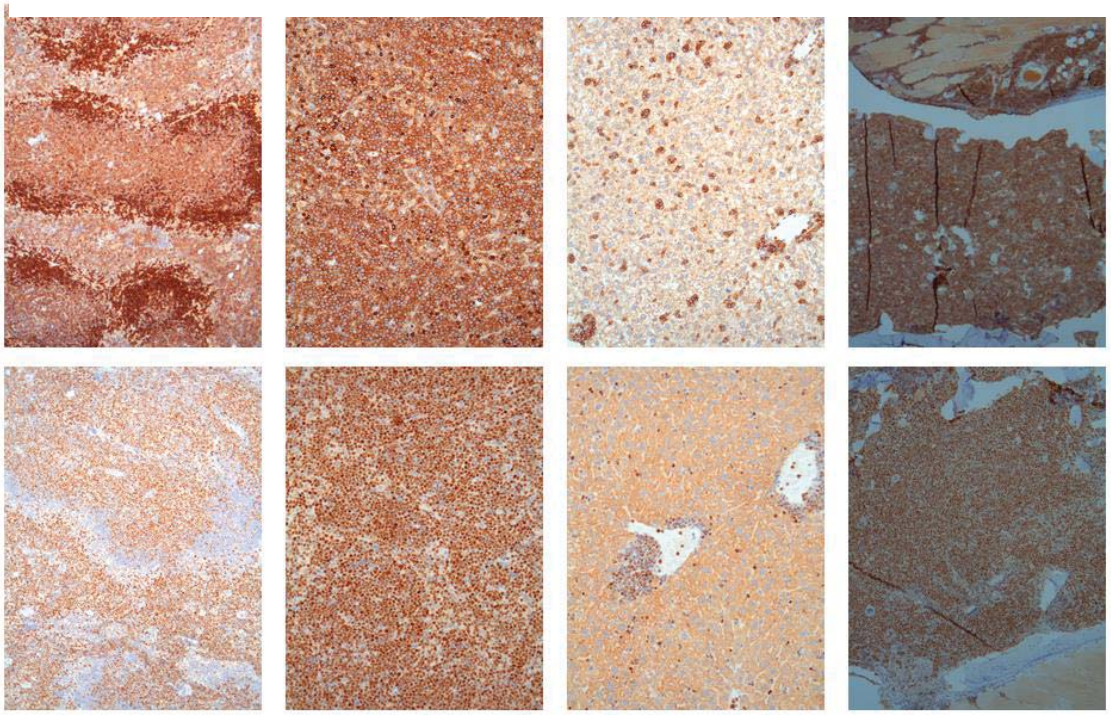


Figure 6.18: IHC of *Vk*MYC-TA-hPB* 12.4a, an IM tumour with blastic morphology. From left to right the tissues are spleen, lymph node, liver and bone marrow. Top panel: B220. Bottom panel: c-Myc.

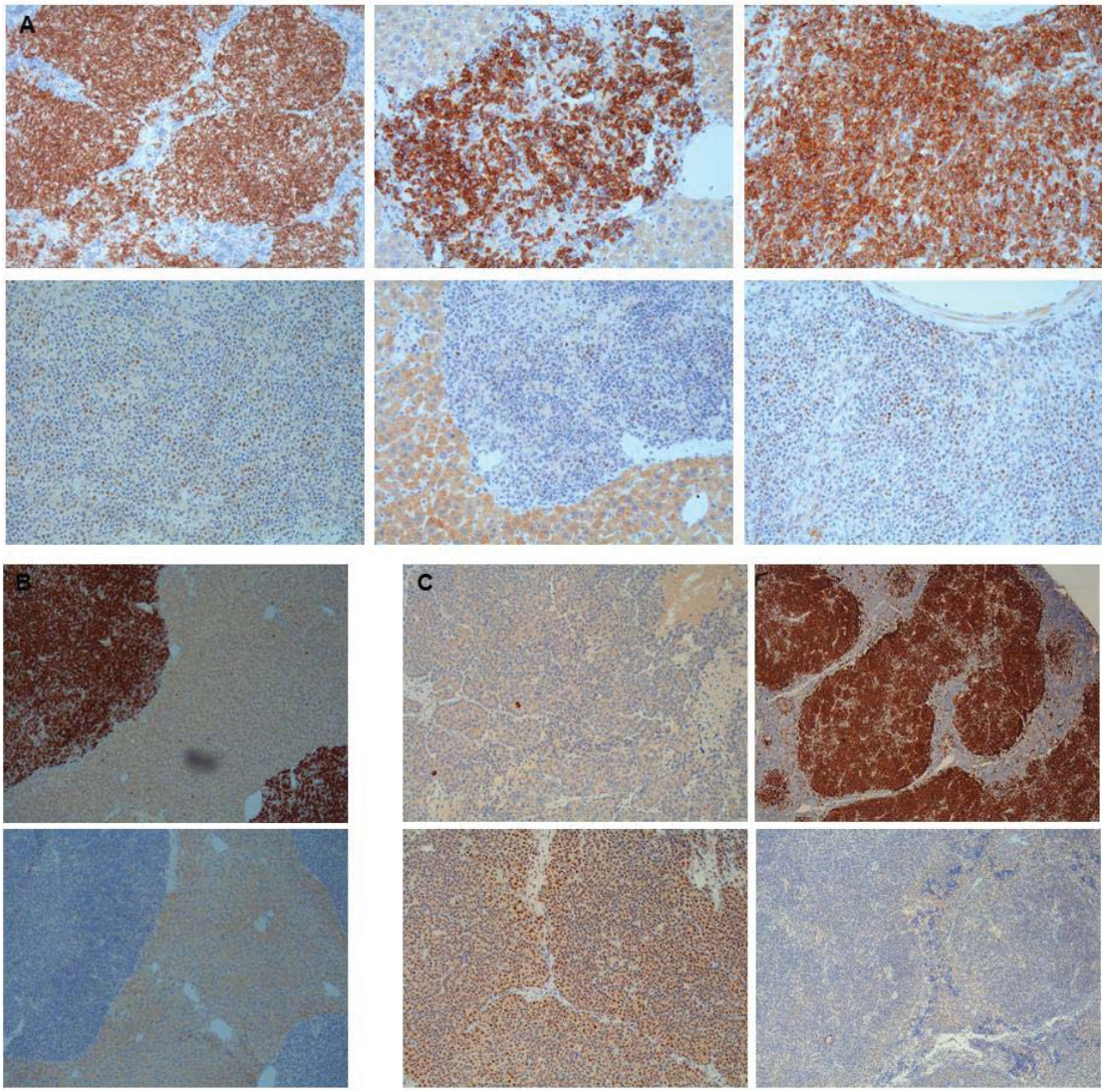


Figure 6.19: IHC of tumours from *Vk*MYC-TA-hPB* IM mice diagnosed with lymphoma, which were negative for Myc. B220 staining (top row) and c-Myc (bottom row) in each. **(A)** The malignant B cells in *Vk*MYC-TA-hPB*14.1h are negative for Myc in the spleen (left), liver (center) and mesenteric lymph node (right). **(B)** The findings were similar in *Vk*MYC-TA-hPB*1.1b. Liver immunohistochemistry is shown. **(C)** *Vk*MYC-TA-hPB*12.3g had an axillary tumour (left) (thought to be a carcinoma of a skin appendage or breast) in addition to a B cell lymphoma which involved the spleen (right). The axillary tumour is positive for Myc, but the B220 positive spleen cells are all negative.

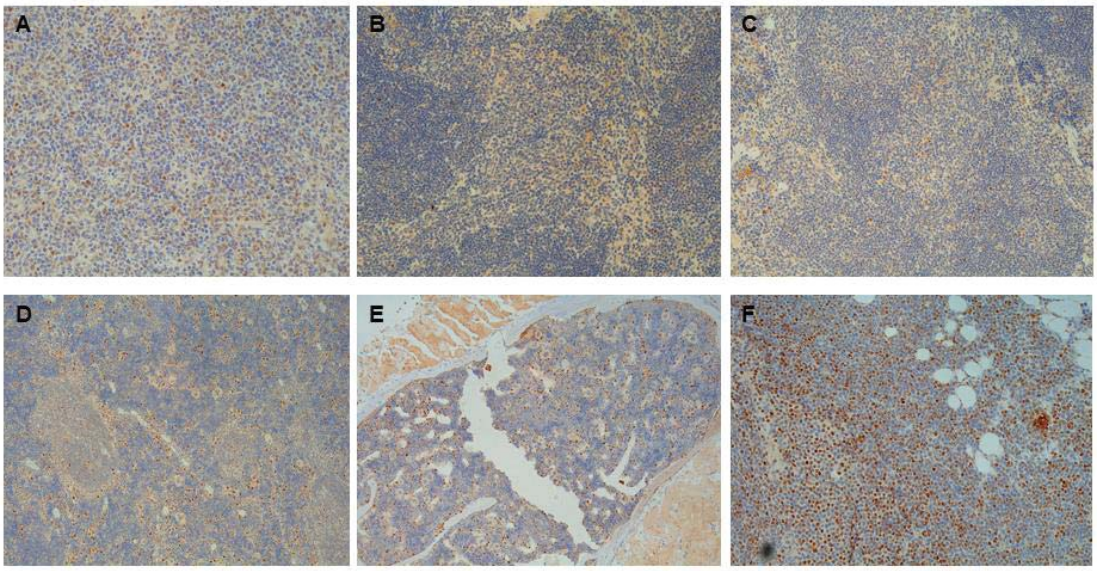


Figure 6.20: Myc IHC shows variable numbers of MYC positive cells in control mice. (A) Spleen sample from *Vk*hPB*5.1a, an IM mouse without the *MYC* transgene that developed lymphoma (x200). (B/C) Spleen samples from *GrOnc* only mice 4.1f (x200) and 2.5d (x100) showing only occasional Myc positive cells. (D) Spleen (x100) and (E) bone marrow (x100) from a wild type mouse showing scattered Myc positive cells. (F) A large number of Myc positive cells are seen in the bone marrow from *Vk*hPB*5.1e (x200), an IM mouse without the *MYC* transgene which developed lymphoma that was described as blastic in morphology.

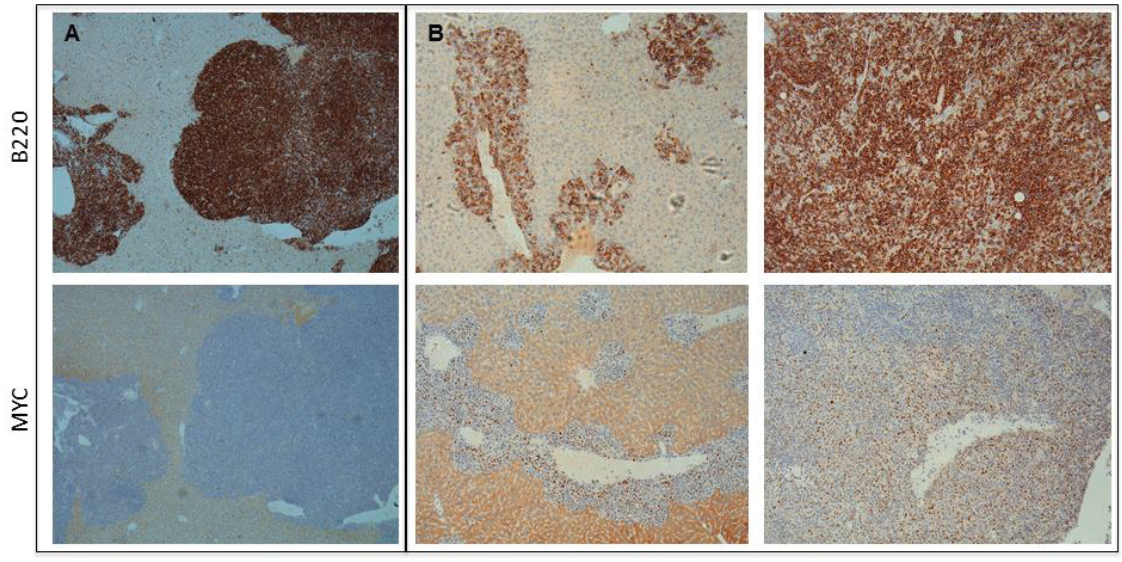


Figure 6.21: Myc IHC in *Vk*MYC-TA-hPB* mice that lacked the *GrOnc* transposon. (A) Of the four tumours tested three had only occasional Myc positive cells as seen here in *Vk*MYC-TA-hPB*14.1g. (B) *Vk*MYC-TA-hPB*10.1g was the only one to have some increase in Myc staining. Shown is the IHC performed on the liver (left) and spleen (right). This was reported as a high grade large cell lymphoma. Top panel B220; bottom Myc.

6.2.8 Stop codon reversion was not seen in *Vk*hPB* and *Vk*MYC-TA-hPB* tumours

In order to investigate if the stop codon was reverted in the *hPB* and *MYC-TA-hPB* mice I performed RT-PCR followed by capillary sequencing on RNA from tumour samples in three *Vk*hPB* and four *Vk*MYC-TA-hPB* mice using the same primers as in 6.2.2, which were used to validate splicing from the Vk exon into *MYC* or *hPB*. The results showed clear sequence with maintenance of the stop codon in all tumour samples. Reversion of the stop codon in a minor sub-clone could not be excluded using this capillary sequencing approach. However, as the construct was designed such that activation of the transposase should have been dependent on reversion of the stop codon, we would expect this to be evident in the major tumour clone if it had occurred.

To further investigate reversion of the stop codon in tumour samples I performed the same RT-PCR in eight *Vk*hPB* and ten *Vk*MYC-TA-hPB* samples followed by sequencing of the PCR products using the MiSeq platform. The RNA was derived from tumour samples in 16 of 18 cases. There was one spleen sample from each IM cohort where the histopathology showed 'no lesion'. At least 350 000 reads were mapped to the target sequence in each sample, which represented between 67 and 86% of the total reads. Although there was evidence of mutation around the stop codon, this was always in a minority of reads (table 6.1 and 6.2). In every case at least 88% of the reads had the wildtype sequence which resulted in a protein sequence of ATMG-YPYDV around the stop codon. Typically this wild type sequence accounted for over 95% of reads (table 6.1).

*Vk*MYC-TA-hPB*	Annotated reads	Top hit (% of reads)	*Vk*hPB*	Annotated reads	Top hit (% of reads)
1.1b spleen	569878	95.37	3.1d spleen	670795	88.36
1.6e spleen	452051	97.10	3.2c spleen	845021	97.31
7.6e mass	490535	96.03	3.2g spleen	833492	97.45
8.1d spleen	361669	97.53	5.2d spleen	633469	97.50
10.1g	435312	94.84	6.2b spleen	865972	97.37
10.4f spleen	371716	96.79	6.2l spleen	531630	96.76
12.3g	399241	96.87	7.2e spleen	754682	97.36
13.2f spleen	350713	96.54	7.7d spleen	899094	97.18
14.3f spleen	542318	96.85			
14.3f spleen	475863	96.77			
14.5f spleen	408151	96.89			
14.5f spleen	381979	96.79			

Table 6.1: Total reads around the stop codon in each sample from sequencing the RT-PCR products on the MiSeq platform. In each case the top hit was the unmutated sequence and this accounted for at least 94% of reads in all but one case.

6.2.9 The *hPB* and *MYC-TA-hPB* IM tumours are clonal and have undergone somatic hypermutation

In order to determine if these tumours were clonally re-arranged and had undergone SHM, BCR repertoire analysis was performed by Rachael Bashford-Rogers using the method she has established (Bashford-Rogers et al., 2013b). This analysis was performed on DNA from spleen, lymph node or abdominal mass in samples from five *Vk*hPB* and five *Vk*MYC-TA-hPB* IM mice. All samples except one returned more than 21000 reads that passed filtering, which is sufficient to derive robust B cell repertoires. The sample that failed was *Vk*MYC-TA-hPB1.1b*, which returned only 635 reads after filtering (16% of the all reads with its unique barcode). The DNA from this sample was taken from an abdominal mass, which turned out to be a non-B cell tumour, with morphology consistent with a neural or smooth muscle tumour. The mouse also developed lymphoma in the spleen and liver, however the BCR repertoire analysis was not performed on these samples. The remaining samples gave a mean of 130732 reads post filtering (range 21123-239144).

1.1b spleen		1.6e spleen		7.6e mass		8.1d spleen		10.1g		10.4f spleen		12.3g		13.2f spleen		14.3f spleen		14.5f spleen	
B cell lymphoma		Large cell lymphoma with low mitotic rate		Small cell lymphoma		No lesion		Large cell lymphoma		Large cell high grade lymphoma		Lymphoma		Follicular lymphoma		Follicular large cell lymphoma		Blastic	
ATMG-YPYDV	543519	ATMG-YPYDV	438951	ATMG-YPYDV	471083	ATMG-YPYDV	352739	ATMG-YPYDV	412840	ATMG-YPYDV	359781	ATMG-YPYDV	386745	ATMG-YPYDV	338589	ATMG-YPYDV	525218	ATMG-YPYDV	395458
ATMG-YPYGV	1847	AAMG-YPYDV	1259	ATMG-YPYGV	1058	ATMGWYPYDV	1240	ATMG-YPYGV	1004	ATMG-YPL-C	2113	ATMG-YPYGV	826	ATMG-CPYDV	559	AAMG-YPYDV	1257	AAMG-YPYDV	704
ATMG-YPYDA	1217	ATMG-YPYDA	1136	AAMG-YPYDV	979	ATMG-YPYGV	817	ATMG-YPYDA	853	ATMGWYPYDV	744	ATMG-YPYDA	677	ATMG-YPYDA	409	ATMG-YPYGV	1055	ATMG-YPYDA	701
AAMG-YPYDV	1083	ATMGWYPYDV	905	ATMG-YPYDA	956	AAMG-YPYDV	560	AAMG-YPYDV	775	AAMG-YPYDV	677	AAMG-YPYDV	635	TTMG-YPYDV	359	ATMG-YPYDA	989	ATMG-YPYGV	636
ATVG-YPYDV	1039	ATMG-YPYGV	865	ATVG-YPYDV	726	ATMG-CPYDV	528	ATVG-YPYDV	618	ATTG-YPYDV	562	ATMGWYPYDV	576	ATMG-YPYGV	320	ATMGWYPYDV	958	ATMGWYPYDV	590
ATTG-YPYDV	884	ATVG-YPYDV	829	ATTG-YPYDV	716	ATMG-YPYDA	506	ATMGWYPYDV	563	ATMG-YPYGV	554	ATTG-YPYDV	534	AAMG-YPYDV	309	ATVG-YPYDV	840	ATTG-YPYDV	497
ATMG-HPYDV	773	SAMG-YPYDV	596	ATMG-CPYDV	666	ATTG-YPYDV	396	ATMG-CPYDV	539	ATMG-YPYDA	532	TTMG-YPYDV	512	VTMG-YPYDV	300	ATMG-CPYDV	810	ATVG-YPYDV	454
ATMGWYPYDV	728	ATTG-YPYDV	565	ATMGWYPYDV	646	TTMG-YPYDV	366	ATTG-YPYDV	525	ATMG-HPYDV	506	ATVG-YPYDV	480	ATVG-YPYDV	246	ATTG-YPYDV	662	ATMG-YPHDV	454
ATMG-CPYDV	685	TTMG-YPYDV	547	ATMG-HPYDV	599	ATVG-YPYDV	357	ATMG-HPYDV	473	ATVG-YPYDV	477	ATMG-HPYDV	448	ATTG-YPYDV	241	ATMG-HPYDV	570	ATMG-CPYDV	452
ATMG-YPHDV	661	ATMG-HPYDV	540	ATMG-YPHDV	556	ATMG-YPHDV	320	TTMG-YPYDV	372	ATMG-CPYDV	384	ATMG-CPYDV	360	ATMGWYPYDV	229	TTMG-YPYDV	515	TTMG-YPYDV	373

3.1d spleen		3.2c spleen		3.2g spleen		5.2d spleen		6.2b spleen		6.2l spleen		7.2e spleen		7.7d spleen	
Lymphoma?		Pleiomorphic B cell lymphoma		Follicular lymphoma		B cell lymphoma		No lesion		Burkitt like lymphoma		Large cell high grade		Follicular large cell high grade	
ATMG-YPYDV	592695	ATMG-YPYDV	822303	ATMG-YPYDV	812238	ATMG-YPYDV	617610	ATMG-YPYDV	843166	ATMG-YPYDV	514425	ATMG-YPYDV	734725	ATMG-YPYDV	873706
ATMG-YPSDV	53938	ATMG-YPYGV	1808	ATMG-YPYGV	1890	ATMG-YPYGV	1348	ATMG-YPYGV	1883	ATMGEYPYDV	2209	ATMG-YPYGV	1759	ATMG-YPYGV	2021
ATMGFYPYDV	4154	ATMG-YPYDA	1777	ATMG-YPYDA	1758	AAMG-YPYDV	1210	ATMG-YPYDA	1727	ATMG-YPYGV	1274	ATMG-YPYDA	1576	ATMG-YPYDA	1739
ATVG-YPYDV	2434	AAMG-YPYDV	1769	AAMG-YPYDV	1469	ATMG-YPYDA	1198	AAMG-YPYDV	1431	AAMG-YPYDV	1202	AAMG-YPYDV	1567	AAMG-YPYDV	1723
ATMG-YPYGV	1520	ATMG-HPYDV	1375	ATVG-YPYDV	1239	ATMGWYPYDV	1008	ATMGWYPYDV	1316	ATMG-YPYDA	1088	ATTG-YPYDV	1330	ATTG-YPYDV	1347
ATMG-YPYDA	1255	ATMGWYPYDV	1354	ATMG-CPYDV	1233	ATVG-YPYDV	996	ATVG-YPYDV	1272	ATVG-YPYDV	918	ATMGWYPYDV	1239	ATVG-YPYDV	1304
AAMG-YPYDV	1224	ATTG-YPYDV	1244	ATMGWYPYDV	1205	ATMG-CPYDV	980	ATTG-YPYDV	1250	ATMGWYPYDV	820	ATVG-YPYDV	1056	ATMGWYPYDV	1288
ATTG-YPYDV	961	ATVG-YPYDV	1194	ATTG-YPYDV	1161	ATTG-YPYDV	953	ATMG-HPYDV	1178	ATMG-YHYDV	772	ATMG-CPYDV	994	TTMG-YPYDV	1224
ATMG-HPYDV	864	ATMG-CPYDV	1009	ATMG-HPYDV	1121	ATMG-HPYDV	696	ATMG-CPYDV	1170	ATTG-YPYDV	754	ATMG-HPYDV	969	ATMG-HPYDV	1081
VTMG-YPYDV	856	TTMG-YPYDV	823	ATMG-YPHDV	969	TTMG-YPYDV	648	ATMG-YPHDV	893	ATMG-HPYDV	747	TTMG-YPYDV	923	ATMG-YPHDV	1054

Table 6.2: Results from RT-PCR and sequencing around the stop codon in 10 *Vk*MYC-TA-hPB* and 8 *Vk*hPB* samples. The diagnosis is shown in each case along with the protein translation of the major sequences identified using MiSeq sequencing. Although some sequences were identified in which the stop codon is reverted, these were always in a minority of reads. The bioinformatic analysis was performed by Rachael Bashford-Rogers.

The *Vk*hPB* samples included two controls that had no clinical features of lymphoma at death and no identifiable abnormality on histopathology. The spleen DNA from these samples showed diverse B cell repertoires without evidence of clonal expansion (figure 6.22). The largest cluster in the samples without an identified tumour was 2.1%. By contrast, the most monoclonal of the samples was *Vk*MYC-TA-hPB*10.1h, the sample which was positive for B220 but negative for CD19 on flow cytometry (figure 6.11a) and was positive for Myc on IHC (figure 6.17). In this sample 95% of the B cell repertoire was represented by a single cluster. This cluster shared the V gene IGHV14-2*02 and the J gene IGHJ4*01 and there were also an average of 7.94 mutations in the V gene, indicating the B cells had undergone SHM.

Of the six other tumour samples that were analysed four had a dominant cluster that accounted for at least 50% of the repertoire (table 6.1). In sample *Vk*hPB*3.3a there were two clusters that accounted for 16% and 10% of the repertoire and this was reported as a diffuse large cell lymphoma based on H&E stained tissue sections. However, IHC showed double staining for CD3 and B220 and on flow cytometry only 33% of viable cells were CD19 and B220 positive, while 66% were CD4 positive. On the basis of the flow cytometry results this was called a T cell tumour, but these results indicate there was also a clonal expansion of B cells. In the remaining sample analysed for B cell repertoires (*Vk*MYC-TA-hPB*13.2f) the largest two clusters accounted for 4% of reads each. This sample was reported as follicular lymphoma based on the H&E stained specimens, but IHC was not performed. Flow cytometry on the spleen showed 42% B cells and 35% T cells and no definite lineage was assigned to this tumour. The B cell repertoire results on the spleen are consistent with a small clonal expansion of B cells. Repeat analysis was subsequently performed using BM RNA from the same mouse and the largest cluster identified was 18%, which provides further evidence that this mouse did have a clonal B cell disorder.

It is important to note that as our main analysis was performed on DNA it is possible that two clusters relate to a single B cell clone, with one representing an abortive non-functional rearrangement. This is less likely when the clusters are clearly different in size such as for *Vk*hPB* 6.4d, than in cases such as *Vk*hPB* 6.5c. However, on flow cytometry of *Vk*hPB*6.5c two distinct B cell populations were

evident (figure 6.23) and it is also possible that these clusters came from separate B cell clones.

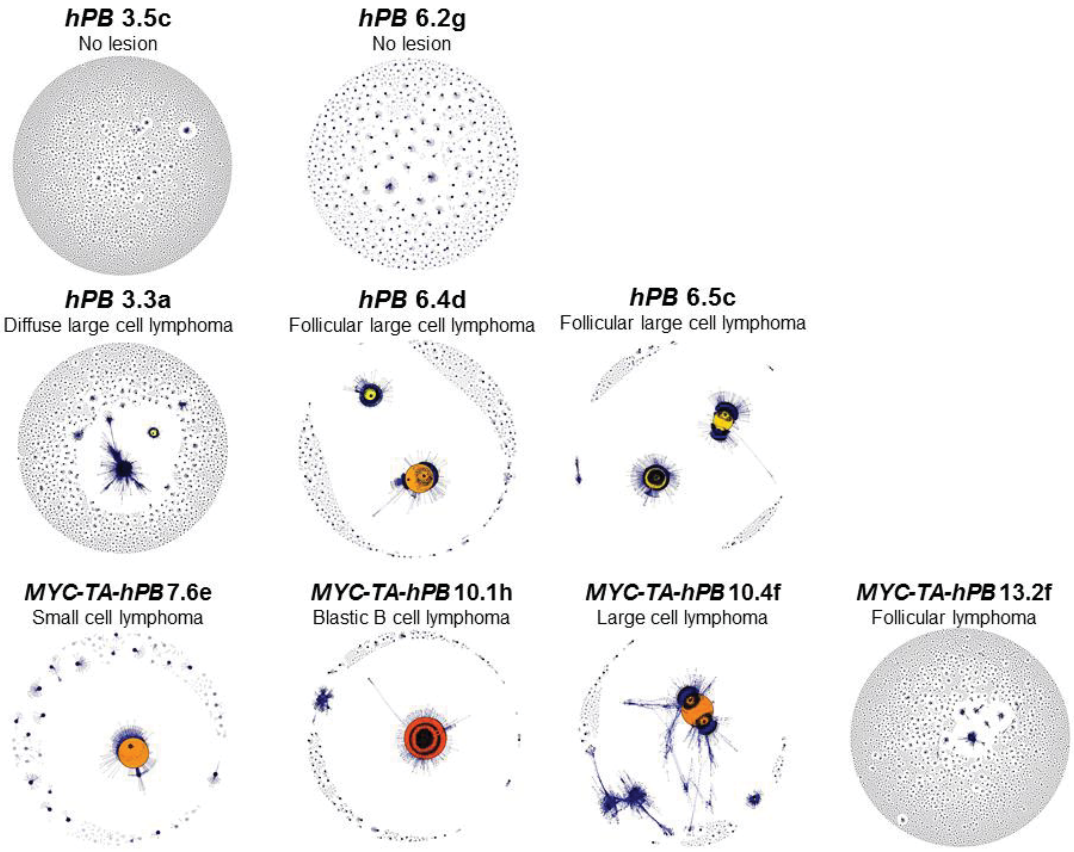


Figure 6.22: B cell receptor repertoires from different samples. Each vertex represents a unique sequence and the relative size of the vertex is proportional to the number of reads that shared this identical sequence. Vertices that differ by one base are connected by edges and groups of connected vertices are grouped into clusters. The colour of the major vertex indicates the proportion of reads that share that sequence (yellow <40%, orange 40-90%, red>90%)(Bashford-Rogers et al., 2013b). The sample ID and histopathology diagnosis are indicated for each sample. The total sequence reads and proportion assigned to the major clusters are shown in table 6.3. Plots courtesy of R. Bashford-Rogers.

Sample ID	Cluster ID	% of Repertoire	Number of Reads	Number of Vertices	V gene	J gene	Mean mutations in V
*Vk*hPB* 3.5c spleen	1	**1.05**	1753	235	IGHV3-6*01	IGHJ4*01	12.00
*Vk*hPB* 6.2g spleen	1	1.93	706	155	IGHV5-15*01	IGHJ1*03	1.37
	2	**2.12**	773	152	IGHV5-9-1*02	IGHJ4*01	1.18
	3	1.20	439	77	IGHV3-1*02	IGHJ4*01	5.06
	4	1.06	386	72	IGHV3-1*02	IGHJ4*01	6.11
	5	1.20	438	68	IGHV15-2*02	IGHJ4*01	1.03
	9	1.16	424	58	IGHV3-1*02	IGHJ4*01	5.21
*Vk*MYC-TA-hPB* 13.2f spleen	1	3.83	5901	745	IGHV11-2*01	IGHJ1*03	1.44
	2	**3.99**	6148	421	IGHV3-1*02	IGHJ2*01	5.92
	3	1.03	1582	202	IGHV11-2*01	IGHJ4*01	4.89
*Vk*hPB* 3.3a spleen	1	10.04	12740	3463	IGHV1-15*01	32	31.00
	2	**15.57**	19746	1126	IGHV2-4-1*01	IGHJ3*01	6.33
*Vk*hPB* 6.5c spleen	1	**51.44**	123026	6061	IGHV5-4*03	IGHJ3*01	1.93
	2	46.15	110354	5791	IGHV1-55*01	IGHJ3*01	5.70
*Vk*hPB* 6.4d spleen	1	**63.76**	88237	4352	IGHV15-2*01	IGHJ1*03	1.63
	2	27.92	38637	2191	IGHV14-2*02	IGHJ3*01	7.60
*Vk*MYC-TA-hPB* 7.6e mesenteric mass	1	**68.74**	14519	1220	IGHV1-15*01	IGHJ2*01	3.38
	2	2.82	596	90	IGHV3-1*02	IGHJ4*01	7.13
	3	2.54	537	79	IGHV3-3*01	IGHJ4*01	1.23
	4	1.98	419	75	IGHV12-3*01	IGHJ1*03	1.19
	5	2.25	476	66	IGHV3-6*01	IGHJ1*03	1.12
	6	1.28	270	49	IGHV3-1*02	IGHJ3*01	11.94
	7	2.27	479	47	IGHV11-2*01	IGHJ1*03	0.91
	8	1.35	285	47	IGHV11-2*01	IGHJ4*01	4.00
	9	1.12	236	45	IGHV5-6-2*01	IGHJ4*01	41.36
	10	1.30	274	37	IGHV12-3*01	IGHJ3*01	2.92
	12	1.08	228	30	IGHV5-2*02	IGHJ4*01	1.17
*Vk*MYC-TA-hPB* 10.4f spleen	1	**82.34**	72794	5986	IGHV1-53*01	IGHJ2*01	26.14
	2	7.85	6940	1757	IGHV1-53*01	IGHJ2*01	27.32
	3	4.54	4013	470	IGHV12-3*01	IGHJ1*03	1.35
*Vk*MYC-TA-hPB* 10.1h LN	1	**94.95**	193947	7615	IGHV14-2*02	IGHJ4*01	7.94

Table 6.3: Cluster assignment and B cell repertoires for samples depicted in Figure 6.22. All clusters which accounted for at least 1% of the reads for a sample are shown. The V and J genes assigned to the cluster and the mean number of mutations in the variable region of sequences in that cluster are shown.

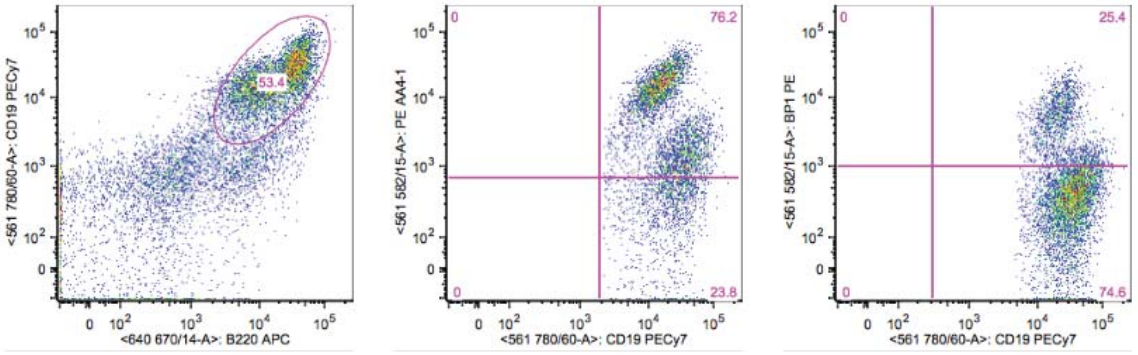


Figure 6.23: Flow cytometry on *hPB* 6.5c. Two distinct B cell populations are evident.

6.2.10 Serum protein electrophoresis of *Vk*MYC-TA-hPB* and *Vk*hPB* mice

Serum samples were collected at monthly intervals from a cohort of mice to look for the development of a paraprotein. In addition, where possible, a serum sample was taken from all mice at the time of death. The rate of detection of monoclonal proteins was lower than that reported by Chesi et al who found that over 50% of *Vk*MYC* mice, but less than 10% of controls had a detectable paraprotein at 30weeks of age (Chesi et al., 2008). In that study monoclonal proteins were first detectable from 20 weeks of age and increased in intensity over time. By 50 weeks 80% of *Vk*MYC* mice had a monoclonal band (Chesi et al., 2008). Although I detected occasional monoclonal bands these did not always increase in intensity and were sometimes lost on subsequent sampling (figure 6.24). For example *Vk*MYC-TA-hPB* 10.1b had a clear monoclonal band on the week 37 serum sample, but it was negative when the mouse died of lymphoma at 84.5 weeks of age.

Protein electrophoresis was performed on blood samples taken at death from 49 *Vk*MYC-TA-hPB* mice (all genotypes) of which only eleven had a definite monoclonal band. Of these, no diagnosis of malignancy was made based on histopathology in four, while the other seven were found to have lymphoma. Only one was found to have a bone marrow plasmacytosis on histopathology examination and this mouse also had lymphoma. Of the 17 *Vk*MYC-TA-hPB* only mice (i.e. no IM), which had protein electrophoresis performed on the final blood sample, four had a clearly detectable paraprotein of which one had a bone marrow plasmacytosis on histopathology. Of 16 *GRL* only or WT mice, four had a clear paraprotein at death.

Protein electrophoresis was performed on serum samples taken at death from seventeen *Vk*hPB* mice (including two non-IM) and six had a paraprotein detected. The one IM mouse reported to have a plasmacytosis in the lymph nodes and liver on histopathology had an increase in gamma globulins but no discrete monoclonal band.

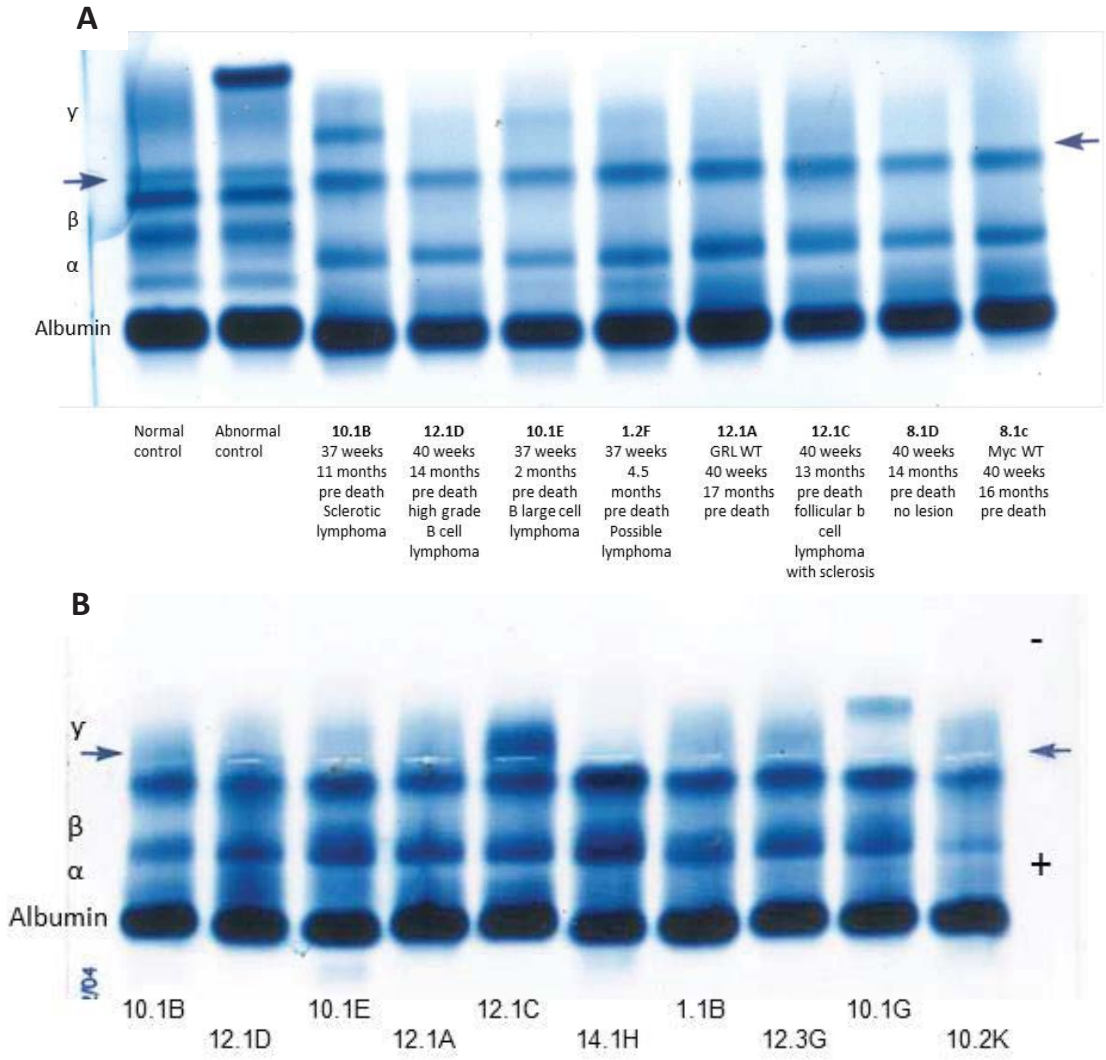


Figure 6.24: Serum protein electrophoresis on *Vk*MYC-TA-hPB* mice. A: Week 37-40 results from a selection of live mice which were serially bled. **B:** Results from sera obtained at death from ten mice including five mice from A (left). In these terminal blood samples only 10.1G has a clear monoclonal band. This mouse was a *Vk*MYC-TA-hPB* only mouse (No IM) that died with large cell high grade lymphoma. Mouse 12.1C had a lymph node plasmacytosis on histopathology and shows increased gamma globulins without a discrete band. 10.1B, 14.1H, 1.1b, 12.3G and 10.2K were all IM mice that died with lymphoma. Note that a monoclonal band was evident in mouse 10.1B at week 37 (as well as in several other serial samples). This was lost at death when the mouse had lymphoma. The arrows indicate the point of sample application and the position of α, β and γ globulins are indicated. The controls are Kemtrol Normal and abnormal serum controls.

6.2.11 Common integration site analysis identifies known and novel lymphoma genes

CIS were analysed using the top 10, 25 and 100 integrations from each tumour sample. Samples included in the analysis were from IM mice which had been diagnosed with lymphoma based on histopathology analysis or, where histopathology results were not available, from mice with necropsy findings which were suggestive of lymphoma (splenomegaly (≥0.4g) and/or significant lymphadenopathy). Only a single tissue sample from each mouse was used in the analysis and non-lymphoid tumours were excluded. The CIS identified using the top 25 hits in each cohort are shown in tables 6.3 and 6.4. Similar lists for the top 10 and 100 hit analysis are shown in appendix 6A.

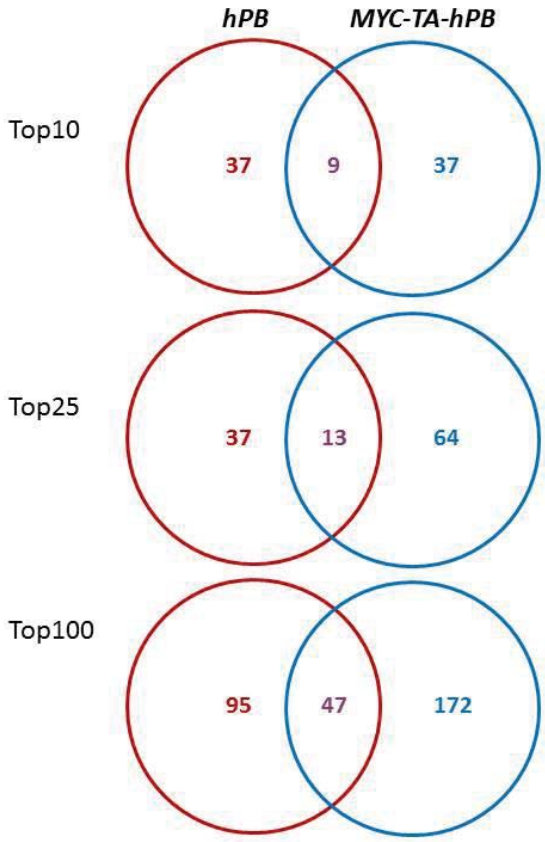
The CIS identified by CIMPL analysis showed significant overlap between the *Vk*hPB* and *Vk*MYC-TA-hPB* IM cohorts (figure 6.25). The overlapping integrations between the two cohorts using the top 10 hits from each tumour included five of the top hits in the *Vk*Myc-TA-hPB* cohort and four of the top six in the *Vk*hPB* cohort. The number of CIS sites identified increased with the number of integrations from each sample used in the analysis (figure 6.26). The most frequently hit CIS were detected regardless of the cut-off used and there were 26 CIS integrations found in all three analyses on *Vk*MYC-TA-hPB* cohort and 22 for the *Vk*hPB* tumours (figure 6.26). The samples that harboured integrations contributing to these CIS are shown in tables 6.5 and 6.6.

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
13	37608797	38048760	Rreb1 Ssr1 Cage1 Riok1	Rreb1	35	25	0	10-100
16	23824262	24334780	Sst Rtp2 Bcl6	Bcl6	19	15	0	10-100
4	44552079	44884132	Pex5 Mir5120 Gm12462 Gm12463 Zcchc7 Gm22639	Gm12463	8	7	0	10-30, 50-100
4	32211207	32533941	Bach2 D130062J21Rik Gm11932 Gm24371	Bach2	9	6	0	10-40, 60-100
11	44649490	44834554	Ebf1 Gm12158	Ebf1	6	6	4.16E-11	40, 60, 80-100
1	58595725	58840495	Ndufb3 Gm10068 Als2cr12 Cflar Casp8	Cflar	5	5	0	10-100
7	24973525	25245544	Atp1a3 Grik5 Zfp574 Pou2f2 D930028M14Rik Dedd2 Zfp526 Gsk3a 9130221H12Rik Erf	Pou2f2	5	5	0	10-100
11	62337694	62503277	Ncor1 Pigl Gm12278	Ncor1	6	4	0	10-60, 80, 100
2	18620014	18835286	Gm13355 Gm13352 Commd3 Bmi1 Gm13334 BC061194 Gm20539 Gm13333 RP23-396N6.8	BC061194	4	4	0	50, 70, 100
6	98999663	99220090	Foxp1	Foxp1	4	4	0	50-100
18	65351483	65529809	Alpk2 Gm22567 Malt1 Gm26114	Malt1	4	3	0	10-100
18	80544846	80723172	Nfatc1	Nfatc1	4	3	0	10-100
X	52821588	53000206	Gm14607 Gm6539 Rps2-ps13 Phf6 Hprt	Phf6	4	3	0	10-100
1	9647221	9732895	Mybl1 Vcpip1	Mybl1	3	3	0	10, 100
2	170043831	170229747	Tshz2 AL731822.1 Zfp217 AL844576.1	Zfp217	3	3	0	10-100
6	129115904	129259023	Klrb1-ps1 Gm26160 Clec2d AC142191.1	Clec2d	3	3	3.106E-12	10, 20, 40, 60-80, 100
10	18920536	19091825	Tnfaip3	Tnfaip3	3	3	0	10-100
11	20037469	20142449	Actr2	Actr2	3	3	0	10, 30-70
17	47684419	47814375	Frs3 Gm14873 Pgc Tfeb	Tfeb	3	3	0	40, 50, 70-100
18	60673185	60937918	Ndst1 Rps14 Gm8731 Cd74 Mir5107 Tcof1 Arsi Camk2a	Cd74	3	3	0	10-100
5	147333730	147365844	Flt3	Flt3	3	2	0	10-30, 50, 60, 90
14	115020837	115042641	intergenic	Mir17hg	3	2	0	20, 30, 70, 90, 100
2	98663779	98667685	Gm10801 Gm10800	Gm10800	2	2	0	10
2	104072250	104098503	Cd59b	Cd59b	2	2	0	10, 20, 40, 60, 80, 90
2	163223462	163288022	Tox2	Tox2	2	2	0	10, 30, 40, 60, 80, 90
2	163598659	163692531	Ttpal Serinc3 0610039K10Rik Pkig Gm16316	0610039K10Rik	2	2	1.792E-09	80
2	165953289	166047161	Gm11462 Gm11463 Gm11464 Ncoa3	Ncoa3	2	2	0	20, 40-60, 80, 90
3	101227944	101368445	Gm12490 Cd2 Gm10355	Cd2	2	2	0	10-100
4	138035572	138067992	Eif4g3	Eif4g3	2	2	0	10-40, 60, 70, 90
5	24779372	24802663	intergenic	Rheb	2	2	0.0002003	80
7	80139008	80250546	Gm24012 Mir1965 Sema4b Cib1 Gdpgp1 Gm15504 Ttll13	Sema4b	2	2	0	20, 30, 50-100
7	126136157	126229106	Xpo6 Gm17137 Gm17136	Xpo6	2	2	0.0011571	80
11	44993577	45002073	Ebf1	Ebf1	2	2	6.277E-05	40, 60
14	121914942	121993177	Ubac2 Gpr18 Gpr183	Ubac2	2	2	0	20, 40-100
15	97756840	97814214	Rapgef3 Gm16257 Gm16256 Slc48a1 Hdac7	Slc48a1	2	2	5.632E-05	50
17	45513141	45606050	Aars2 Tcte1 Tmem151b Nfkbie Slc35b2 Hsp90ab1 Slc29a1 Gm17080 Gm7325	Nfkbie	2	2	0	10-100
17	80424772	80439617	Sos1	Sos1	2	2	6.359E-05	80
18	50018926	50029554	Tnfaip8	Tnfaip8	2	2	2.987E-05	10, 40, 90
18	54955971	54955971	Zfp608	Zfp608	2	2	0.0021753	90
X	38527423	38569712	Cul4b	Cul4b	2	2	1.863E-13	30, 60-90
3	94956980	94964305	Rfx5 B230398E01Rik	B230398E01Rik	1	1	5.055E-05	40
3	100479915	100483577	Fam46c	Fam46c	1	1	9.02E-10	40
8	84907402	84908305	intergenic	Dnase2a	1	1	6.443E-06	10
9	82971653	82972578	Phip	Phip	1	1	2.891E-06	10
10	80108736	80112531	intergenic	Stk11	1	1	3.692E-06	40
14	76672486	76720129	intergenic	Serp2	1	1	9.473E-08	50
15	97731212	97738858	Endou	Endou	1	1	0.0018575	40
17	56598174	56605596	Safb	Safb	1	1	0.0016065	80
X	7811171	7813045	Gripap1	Gripap1	1	1	3.688E-08	20
X	12133903	12138595	Bcor 2908C10Rik	Bcor	1	1	0	50
X	48414879	48419570	Elf4	Elf4	1	1	1.11E-16	50

Table 6.3: Complete list of CIS generated by CIMPL analysis using the top 25 hits in each of the *Vk*hPB* IM lymphoma samples. The start and end boundaries encompass all analysis windows in which each locus was identified as a CIS. The gene shown as nearest to peak was the central gene in the majority of kernel windows (scales) detecting the CIS, but is not necessarily the target gene for the CIS. Due to local hopping the total number of insertions occasionally included multiple integrations from the same tumour, so both the total number and the number after correction for local hopping are shown. The smallest p value identified at any scale is shown along with the analysis scales at which the CIS was detected (x1000).

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
13	37605904	38076948	Rreb1 Ssr1 Cage1 Riok1	Rreb1	45	36	0	10-100
16	23859730	24306844	Sst Rtp2 Bcl6	Bcl6	19	16	0	10-100
11	3018303	3350841	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1 Patz1 Gm12592 Gm11944 Pik3ip1 Limk2	Sfi1	19	15	0	10-100
1	58523329	58882113	Fam126b Ndufb3 Gm10068 Als2cr12 Cflar Casp8	Cflar	14	14	0	10-100
4	32215887	32568511	Bach2 D130062J21Rik Gm11932 Gm24371 BC024582	Bach2	13	10	0	10-100
5	147199841	147483174	Gm24556 2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 Gm8054 Pan3	Flt3	10	7	0	10-100
3	135522156	135794340	Manba Oaz2-ps Nfkb1 Gm9799	Nfkb1	7	7	0	10-100
9	32456146	32777632	Fli1 Ets1	Ets1	7	7	0	30-100
3	102923643	103141390	Sycp1 Nr1h5 Gm22826 Sike1 Csde1 Nras Ampd1 Gm23820 Dennd2c	Csde1	6	6	0	10-100
11	109273985	109479376	Rgs9 Gm11696 Gna13 9930022D16Rik Amz2 Gm15642 Slc16a6 Gm25540 Arsg	Gna13	6	5	0	10-100
16	36530447	36730895	Casr Cd86 Ildr1	Cd86	6	5	0	10-100
16	55728468	55937755	Gm26369 Nfkbiz Nxpe3 Cep97	Nfkbiz	5	5	0	10-100
15	61857394	62060960	Myc Pvt1	Myc	6	4	0	10-100
6	98915128	99120608	Foxp1	Foxp1	5	4	0	20-100
14	114940488	115120931	Gm24073 Mir17hg Mir17 Mir18 Mir19a Mir20a Mir19b-1 Mir92-1 Gpc5	Mir17hg	5	4	0	10-100
3	51341797	51447464	Mgarp Ndufc1	Naa15	4	4	1.145E-11	40-70, 90, 100
6	99277164	99463074	Foxp1 Gm20696 Gm20705	Foxp1	4	4	0	30-100
10	68079912	68275650	Arid5b	Arid5b	4	4	0	10-100
11	62305055	62437030	Ttc19 Gm12275 Ncor1 Gm12276	Ncor1	4	4	0	10-30, 50-100
16	4486960	4590661	Srl Gm15885 Tfap4	Tfap4	4	4	0.0004811	60-100
17	46715492	46941709	CT030702.1 Pex6 Gnmt Cnpy3 Ptcra 2310039H08Rik Rpl7l1 Gltscr1l A330017A19Rik Tbcc Gm23797 Prph2 Ubr2	Gltscr1l	4	4	0	10-100
X	18090577	18280335	Dusp21 Kdm6a	Kdm6a	4	4	0	10, 30-100
12	58923991	59079189	AC163296.1 Sec23a Gemin2 Gm22973 Trappc6b Pnn	Sec23a	5	3	0	10-90
11	89052813	89086993	Dgke Gm24974 A930013B10Rik Gm525	Dgke	4	3	0.000567	70
1	175819610	175931858	Exo1 Gm23805	Exo1	3	3	0	20-80, 100
2	45048065	45075385	Zeb2	Zeb2	3	3	0.0001208	70, 80
5	64572308	64721730	Gm25306 RP24-448C16.1 RP24-448C16.2	Gm25306	3	3	0	30-100
6	100333917	100373056	intergenic	Rybp	3	3	0.000447	100
6	148905409	148980057	Fam60a 3010003L21Rik Gm23395 Gm25539	Fam60a	3	3	5.598E-11	10, 20, 40, 80-100
7	3186234	3265222	Mir290 Mir291a Mir292 Mir291b Mir293 Mir294 Mir295 Nlrp12	Nlrp12	3	3	4.186E-08	60, 90
7	25054513	25140292	Grik5 Zfp574 Pou2f2	Pou2f2	3	3	0	20, 50-100
10	128758739	128885969	Wibg Rpsa-ps2 Mmp19 Tmem198b Dnajc14 Ormdl2 Sarnp Gdf11	Mmp19	3	3	9.983E-06	10, 40, 100
11	61072860	61168564	Tnfrsf13b Gm12269 Usp22	Tnfrsf13b	3	3	0	10-100
13	48131563	48256184	AC140293.1	AC140293.1	3	3	0.0002004	80-100
14	121864043	121981443	Ubac2 Gpr18 Gpr183	Ubac2	3	3	0	70, 80
16	10433615	10587629	Tvp23a Cilta Dexi Clec16a	Cilta	3	3	0	10-100
16	38485797	38538966	Cd80 Timmdc1 Gm15953 Poglut1	Timmdc1	3	3	0.0007035	70
17	5336726	5478112	Arid1b Tmem242 Gm22475	Tmem242	3	3	0.0005588	90, 100
17	24338911	24395466	Abca17 Abca3 Gm25618	Abca3	3	3	0.0005634	100
17	75407283	75510966	Rasgrp3	Rasgrp3	3	3	0	30-100
18	65400269	65488624	Gm22567 Malt1 Gm26114	Malt1	3	3	0	10-80, 100
19	4297360	4471645	Adrbk1 Kdm2a Rhod A930001C03Rik Syt12	Kdm2a	3	3	0.0001512	10-100
19	34171155	34283560	Stambpl1 Acta2	Acta2	3	3	0.0010686	90, 100
19	44300363	44474648	Sod2 Mir5114 Sod4 Sod1	Sod1	3	3	0	10-100
19	6324570	6471336	Cdc42bpg Men1 Map4k2 Gm14966 Gm22278 Sf1 Pygm Rasgrp2 Gm14965 Nrxn2 Gm26470	Rasgrp2	4	2	0	10-60, 80-100
7	110230822	110242509	Swap70 Gm22185	Swap70	3	2	0.0005304	60
X	11961414	12151172	Gm14512 Bcor 2900008C10Rik	Bcor	3	2	0	10-50, 70-100
1	11345739	11352165	intergenic	A830018L16Rik	2	2	4.477E-08	10-30
2	32635236	32664468	Ak1 Eng Mir1954	Eng	2	2	0	10, 30, 70, 90
2	168568329	168609254	Nfatc2	Nfatc2	2	2	7.648E-09	20, 30
2	179522970	179542465	Cdh4	Cdh4	2	2	5.577E-09	10, 30, 40
3	9610064	9651853	intergenic	Zfp704	2	2	7.179E-07	30-50, 70, 80
4	154885055	154990673	Mmel1 Fam213b Tnfrsf14 Gm20421 Hes5 Pank4 Plch2	Tnfrsf14	2	2	0	10, 20, 40-70, 90, 100
6	86905261	86907211	Aak1	Aak1	2	2	1.694E-11	20
6	87825230	87855075	Isy1 Cnbp	Cnbp	2	2	4.707E-06	60, 70, 90
7	43339452	43356983	Siglec5	Siglec5	2	2	0	10, 20, 60
8	72308249	72342268	Klf2 Eps15l1	Klf2	2	2	0.0009592	70
8	105164301	105174567	Cbfb Gm22063	Cbfb	2	2	1.038E-07	50, 60, 90
9	88439513	88464196	Gm20537 4932427H20Rik Syncrip	Gm20537	2	2	8.416E-06	10, 20
10	81312878	81424445	Cactin Tbxa2r Gipc3 Hmg20b Mfsd12 4930404N11Rik Fzr1 Dohh 2210404O07Rik Nfic Gm16104	Dohh	2	2	0	10-100
11	103231757	103243649	Map3k14	Map3k14	2	2	1.998E-15	20-40
11	107932992	107936888	Prkca	Prkca	2	2	3.25E-09	20
12	54975064	54997395	Baz1a Gm20403	Baz1a	2	2	4.465E-05	20, 40, 60, 90
12	72673685	72696616	intergenic	Dhrs7	2	2	2.154E-05	40, 90
12	92880073	92883895	intergenic	Gm23249	2	2	0.0003176	40
12	98973696	98994165	Ttc8	Ttc8	2	2	7.88E-05	30, 90
17	23584741	23603145	Zfp13	Zfp13	2	2	0	20, 50, 60
18	2966613	3021996	intergenic	Vmn1r-ps151	2	2	0	20, 30, 50, 70, 100
18	34933618	34947414	Hspa9 Gm22200 Gm26109	Hspa9	2	2	6.965E-11	10, 50
18	46917651	46930883	Arl14epl	Arl14epl	2	2	0	10, 20, 50
18	60705221	60896148	Ndst1 Rps14 Gm8731 Cd74 Mir5107 Tcof1	Cd74	2	2	0	10-100
19	6046035	6117281	Syvn1 Mrpl49 Fau Znhit2 Tm7sf2 Vps51 Zfp11 BC048609 Cdca5 Naaladl1 Sac3d1	Naaladl1	2	2	0.0007123	60, 90, 100
19	26532985	26543910	intergenic	Smarca2	2	2	1.439E-07	10, 30, 50
X	136092361	136111302	5730412P04Rik	5730412P04Rik	2	2	3.593E-07	50, 90
3	88480531	88505902	Lmna	Lmna	1	1	1.102E-07	60, 70
7	101407282	101408254	Arap1	Arap1	1	1	4.123E-06	10
10	127578664	127579639	Lrp1	Lrp1	1	1	4.693E-08	10
14	31153984	31156222	Stab1	Stab1	1	1	1.11E-16	10, 20

Table 6.4: Complete list of CIS generated by CIMPL analysis using the top 25 hits in each of the *Vk*MYC-TA-hPB* IM lymphoma samples. Columns are as described in table 6.3.



Bach2	Bcl6	Cd2
Cflar	Ciita	Flt3
Foxp1	Malt1	Rreb1

Bach2	Bcl6	Bcor
Cd74	Cflar	Flt3
Foxp1	Malt1	Mir17hg
Ncor1	Pou2f2	Rreb1
	Ubac2	

Akt2, Arid1b, Arid5b, Bach2, Bcl6, Bcor, Cd74, Cdkn2a, Cflar, Cul4b, Ebf1/ 4930597A21Rik, Elf4, Elmsan1, Ets1, Fbxo11, Flt3, Foxp1, Gltscr1l, Gm10800, Gm20521/H2-0b, Grb2, Hivep2, Kdm6a, Lyn, Malt1, Map3k14, Map3k8, March3/Lmnb1, Mir17hg, Ncoa3, Ncor1, Nfatc1, Nfkb1, Nsd1, Pisd-ps2, Pou2f2, Ptma, Rreb1, Sfi1, Sos1, Stk4/Tomm34, Tnfaip3, Tnfrsf14, Ubac2, Zcchc7/Gm12678, Zeb2, Zfp217

Figure 6.25: Overlapping integrations between the two cohorts on analysis using the top 10, 25 or 100 integrations in each. The identity of the central gene in the CIS is shown on the right, for each of the shared integrations.

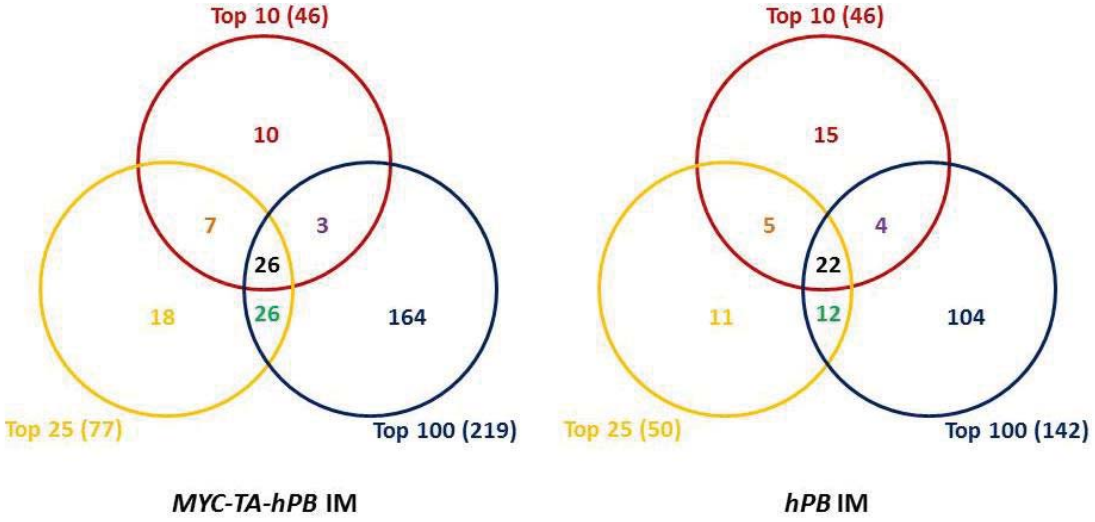


Figure 6.26: Number of integrations shared between analysis using different cut-offs for the number of integrations per tumour included in the analysis. The total number of CIS detected in each analysis is shown in brackets. This includes all CIS including those detected only at a single kernel window.

Table 6.5: CIS identified in the *Vk*hPB* cohort using the top 10, 25 or 100 integrations for CIMPL analysis. Only CIS that were detected in at least two analyses are shown. The start and end CIS boundaries (columns 3& 4) encompass all windows in an analysis in which each locus was identified as a CIS and differed between analyses. Tumours that had integrations within the CIS boundaries are highlighted (red = top 10 integration, blue = top 25, green = top 100). The gene nearest to the centre of the CIS is listed. Where this differed between analyses, more than one gene name is shown.

Table 6.6: CIS identified in the *Vk*MYC-TA-hPB* cohort using the top 10, 25 or 100 integrations for CIMPL analysis. Only CIS that were detected in at least two analyses are shown. The CIS boundaries differed between analyses as shown. Tumours that had integrations within the defined CIS boundaries are highlighted (red = top 10 integration, blue = top 25, green = top 100). The gene nearest to the centre of the CIS is listed. Where this differed between analyses, more than one gene name is shown.

The CIS with the highest number of integrations in both cohorts was at the position of the gene for *Rreb1* (Table 6.3-6.6). *Rreb1* integrations were in the forward orientation within intron 1 in the vast majority of cases suggesting this was an activating integration. Another frequently hit gene in both cohorts was the known B cell lymphoma oncogene *Bcl6*. Although the larger CIS windows report a single CIS spanning the *Bcl6* gene and its flanking regions, the smaller windows report multiple CIS around *Bcl6*. The integrations are localised within intron one and the intergenic region 5' of *Bcl6*. For example, in the *Vk*hPB* CIS data, using the top 25 integrations the 10kb kernel scale reports three discrete CIS; i)centred at 23988419, with 10 hits, ii) centred at ~24144166 with four hits and iii) at ~24188921 with four hits. The first location involves the integrations within intron 1 of *Bcl6*, but the latter two sites are 5' of the gene. It is noteworthy that although the integrations in intron one are almost universally in the forward orientation, which suggests this is an activating integration, those 5' to the gene occur in both orientations.

6.2.12 Read depth and correlation with sample clonality

In addition to sequencing the *Vk*hPB* and *Vk*MYC-TA-hPB* IM mice that developed tumours I also included many spleen samples from mice which were not found to have an abnormality on histopathology analysis. Although the read coverage was generally lower in these samples compared to mice with lymphoma (mean 55491 vs 113460 reads, p=0.0048) this still provided deep coverage of transposon integrations. A mean of 92 unique transposon integration sites were mapped with 2 or more reads after removal of duplicates in these samples compared to 466 in the *Vk*hPB* and *Vk*MYC-TA-hPB* IM mice (p<0.0001). It is notable that the proportion of reads assigned to the top integrations in these samples is significantly lower than in the mice with lymphoma (4.6% vs 18% p=0.0008 in the non-duplicate filtered data) (figure 6.28) although there were a few outlier samples. The sample with the lowest read coverage (8933 reads including duplicates) was unremarkable on histopathology, but over 23% of reads mapped to a single integration (6:71585574, intergenic). It is possible this was an artefact related to the low read coverage.

The drop off in the read count in data without removal of the PCR-duplicates was reviewed for the top 50 integrations in samples which also underwent BCR repertoire analysis (figure 6.29). In the two samples in which the largest clone was less than 2.5% on B cell repertoire analysis, the top transposon integration accounted for

under 1% of total reads. For *Vk*hPB* 3.3a, which had two clusters of 10 and 15%, the top integration accounted for 19% of reads, but the fall off in read counts for the next most common integrations was marked. It is evident that the pattern of fall in read count is different between samples and this did not appear to directly correlate with the number and size of clusters on B cell repertoire analysis. Therefore, from the transposon integration read counts it is not possible to tell if the major integrations are shared in a single or multiple clones although it is possible to infer which samples are unlikely to contain a clonal expansion.

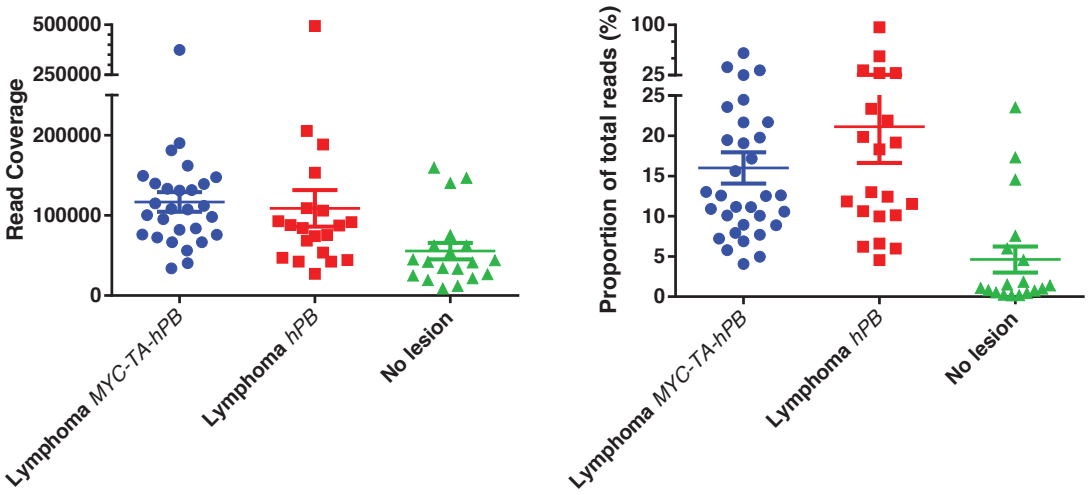


Figure 6.28: Sequence coverage and proportion of reads assigned to the top insertion
Read number (left) and proportion of reads assigned to the top hit (right) in IM mice with either no malignant lesion or a diagnosis of lymphoma. Only mice sequenced on the first two (of three) TraDIS sequencing runs are represented. Each dot represents a single mouse. The mean and standard error of the mean are indicated PCR duplicates were not removed for this analysis.

Multiple samples from a single mouse were analysed by TraDIS in 22 *Vk*MYC-TA-hPB* and nine *Vk*hPB* cases. In those where lymphoma was the only tumour identified on histopathology the major transposon integrations were similar between the different tissue samples. In samples where a non-haematopoeitic tumour was also detected, transposon integrations were often detected in these tumours, but the integrations typically differed to those in the lymphoma samples.

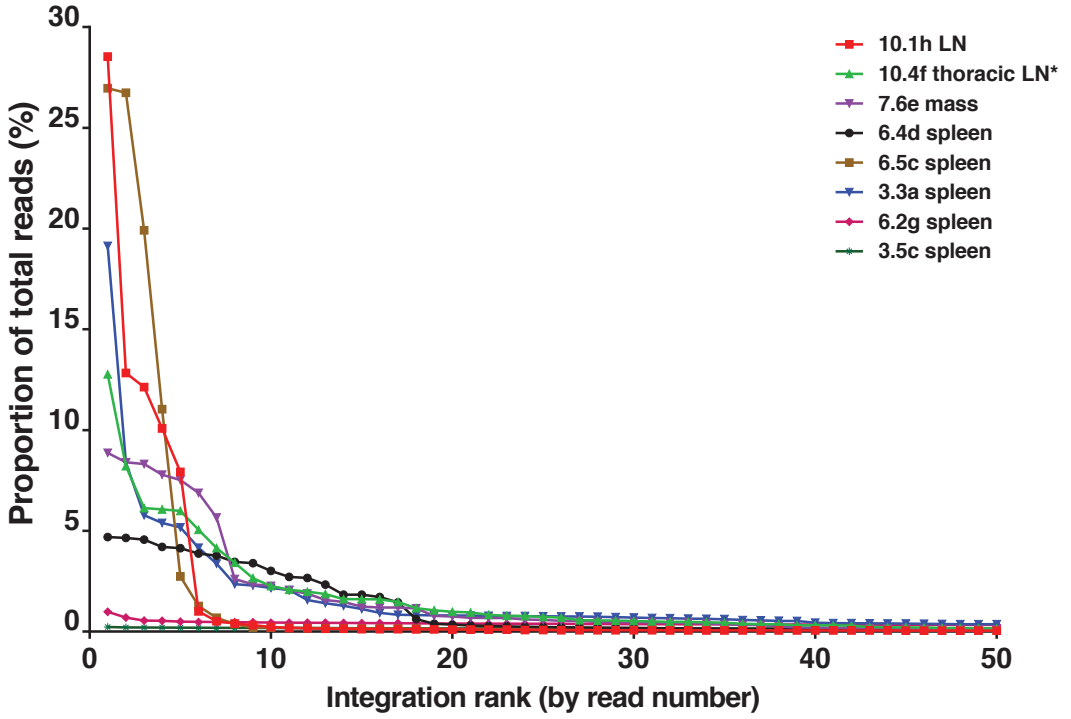


Figure 6.29: The proportion of reads assigned to each of the top 50 transposon integrations in mice that had B cell repertoire analysis. The mice are listed in order of size of the dominant clone as determined by BCR analysis. The tissue that was sequenced is shown and was the same tissue used for B cell repertoire analysis except for 10.4f (*) in which thoracic LN was used for integration site analysis and spleen for the BCR analysis. Mice 6.2g and 3.5c had no abnormality identified on histopathology.

6.3 Discussion

Two *PB* IM mouse cohorts based on *Vk*MYC* mice, which were reported to develop highly penetrant plasma cell malignancies (Chesi et al., 2008), had significantly reduced survival largely due to the development of mature B cell lymphomas, but did not develop multiple myeloma. Although the *PB* transposase was clearly active, the mechanism of activation was not through reversion of the stop codon by SHM, as had been anticipated given the design of the construct. Furthermore, there was evidence of jumping in multiple tissues analysed by PCR, which suggests that activation of *PB* was not specific to the mature B cell compartment. Although it is possible that positive jump PCRs reflect the presence of mature B cells in non-haematopoietic tissues, the finding of large numbers of transposon integrations in

malignant samples which were not B cell lymphomas, and the positive jump qPCR in different haematopoietic lineages is evidence that the transposase was active in cells other than B cells. Nevertheless, the significant incidence of B cell lymphomas suggests that transposition was most active in B lymphocytes.

In the original *Vk*MYC* model as published in *Cancer Cell* (Chesi et al., 2008), it was predicted that translation of the transgene would be stopped by the engineered stop codon in the third codon of the V-kappa exon. However, this stop codon was also engineered as part of a DGYW motif, a preferential target sequence for AID-mediated SHM, in order for it to be sporadically reverted to allow *MYC* translation in a small proportion of germinal and post-germinal centre B cells. In the original paper transfection experiments in 293T cells were used to confirm that translation of *MYC* was absent, and that it did not initiate from a downstream AUG in the setting of the engineered stop codon (Chesi et al., 2008). Furthermore, *hMYC* mRNA expression was only detectable in spleen and BM and was up-regulated by LPS stimulation to induce plasma cell differentiation. The authors reported that 100% of *Vk*MYC* mice developed a monoclonal expansion of bone marrow plasma cells with age, with approximately 80% of necropsied mice found to have more than 5% (range 2%-62%). In contrast, transgenic mice with *MYC* but without the engineered stop codon developed aggressive pro-B lymphomas (Chesi et al., 2008).

To evaluate reversion of the stop codon, Chesi et al performed single colony sequencing on the heavy chain locus VDJ fragment and the transgenic Vκ region from CD138 selected tumour plasma cells. They found evidence of SHM at both loci, with a median of 2.6% mutations in the VH gene and 2.4% in the transgenic Vκ region, but no evidence of SHM in the *MYC* exons. Of the ten mice on which data was presented, reversion of the stop codon was detected in at least one colony in six, including both mice with Burkitt-like lymphoma. Although not every colony showed reversion of the stop codon, they thought this was because multiple copies of the transgene were expressed (either 20 or 8 copies) and concluded that reversion did occur in every plasma cell because they expressed Myc by IHC. However, although it was raised against amino acids 1-262 of the human protein, the anti-Myc antibody used (sc-764_N-262) is not specific for human Myc (http://datasheets.scbt.com/sc-764.pd).

In the *PB* IM mouse cohorts I did not identify reversion of the stop codon in any of the tumours analysed using PCR and capillary sequencing of DNA from the bulk tumour. RT-PCR and deep sequencing on MiSeq, showed that over 94% of reads in each tumour had an identical sequence, with the stop codon intact. Although a small number of reads with reversion of the stop codon were detected in most tumours, these accounted for less than 1% of the total reads in each case, and therefore could not be responsible for activation of the transposase in the main tumour clone.

The precise mechanism by which the transposase is activated in these transgenic mice is yet to be determined. Alternate start codons in the leader sequence were mutated to prevent premature initiation and there are no in-frame initiation codons downstream of the engineered stop codon. The presumption is that there is a cryptic initiation codon and alternate splicing which allows bypass of the stop codon. This may occur from a distant methionine, but translation initiation at non-AUG triplets has also been described in mammalian cells (Peabody, 1989; Starck et al., 2008). These issues were discussed with Leif Bergsagl, senior author on the *Vk*MYC* paper. He acknowledged they are also now suspicious that reversion of the stop codon is not the main mechanism for transgene expression and this may only be true in the infrequent mice which develop Burkitt lymphoma (2/122 mice in the original paper) (Leif Bergsagl, personal communication). Their evidence for this is that the Myc protein is frequently truncated in the plasma cell tumours. Furthermore, in a modified *Vk*MYC* model in which *Cre* was expressed instead of *MYC*, *Cre* activation occurred in 15% of plasma cells, which seems too high for SHM (Leif Bergsagl, personal communication).

The other striking feature of our *PB* IM mice cohorts is the similar median survival and tumour rate between the *Vk*MYC-TA-hPB* and *Vk*hPB* mice, suggesting that the *MYC* transgene did not have a strong additional tumourigenic effect. In keeping with this, the *Vk*MYC-TA-hPB* IM mice without the *GRL* transposon did not have shortened survival compared to the *GRL* only mice. Less than a quarter of these mice died with lymphoma and there were no cases of plasma cell malignancy at death. Therefore, even in the absence of transposition, the *MYC* transgene did not have a powerful tumorigenic effect in the B cell compartment. One possible explanation is that the T2A linker adversely effected Myc function, as after cleavage it is expected to leave a 17 amino acid tail on the C terminal end of Myc(Szymczak et al., 2004).

Interestingly, we identified a CIS centred at *Myc* in the *Vk*MYC-TA-hPB* cohort, but not in the *Vk*hPB* IM mice. Although there were some integrations around this site detected in *Vk*hPB* mice, these were always in low number (<0.5% of reads). In three *Vk*MYC-TA-hPB* mice (7.3h, 8.1b and 16.2a) this was amongst the top 10 hits by read number. The transposon integration site was at position 15:61983792, 2kB upstream of the *Myc* gene in all three cases. If the *MYC* transgene was having a strong lymphomagenic effect, it would be surprising to have further selection for integrations at the *Myc* locus during disease development.

Although the *Vk*MYC-TA-hPB* and *Vk*hPB* mice did not develop the spectrum of tumours we anticipated, insertional mutagenesis was still active in the B cell compartment and resulted in tumour formation. Furthermore, these tumours were predominantly mature B cell lymphomas. CIS analysis performed on integrations from these tumours identified several known lymphoma associated genes (e.g. *Bcl6* and *Malt1*) and novel genes of interest. In this regard, the study was successful in identifying putative cancer genes of relevance to mature B cell malignancies, which are candidates for further study.

The CIS genes included several genes with known roles in lymphomagenesis. The CIS with the second highest number of hits in both cohorts was at the *Bcl6* locus. *Bcl6* is a transcription factor and a proto-oncogene that is expressed in normal germinal centre (GC) B cells and follicular helper T cells(Wagner et al., 2011). *Bcl6* deficient mice display normal B cell development except they are unable to form GC (Dent et al., 1997; Ye et al., 1997). Within the B lymphocyte lineage, expression of *Bcl6* is restricted to GC B cells and its down-regulation is thought to be required for differentiation into plasma cells or memory B cells. Normally, expression of *Bcl6* is regulated by T cell induced CD40 signalling, which down-regulates *Bcl6* via *NF-κB* and *IRF4* activation (Saito et al., 2007), and B cell receptor signalling which leads to *Bcl6* phosphorylation by MAPK and targets it for degradation by the ubiquitin proteasome pathway(Niu et al., 1998). *Bcl6* is expressed in cases of diffuse large B cell lymphoma (DLBCL), follicular, Burkitt's, primary mediastinal B cell, and some Hodgkin lymphomas. Overall about 20-35% of DLBCL have chromosomal translocations involving *BCL6* (Iqbal et al., 2007; Lo Coco et al., 1994; Offit et al., 1994). These translocations are more common in the activated B cell type, than germinal centre cases of DLBCL, although gene expression studies have associated

overexpression of *Bcl6* with a GC signature(Wagner et al., 2011). Mouse studies have demonstrated an auto-regulatory site in the first exon of *Bcl6* and around 10-15% of DLBCL have mutations in this auto-regulatory sequence which is another mechanism of *Bcl6* overexpression (Iqbal et al., 2007; Wang et al., 2002). The 5' non coding region of *BCL6* is also a target for SHM and mutations are reported in around 60% of cases of DLBCL, although the functional significance of many of these mutations is not known (Iqbal et al., 2007).

Bcl6 is a transcriptional repressor. Its activity is dependent on binding to specific DNA sequences via its C-terminal zinc finger domain and recruiting co-repressor molecules including SMRT, NCoR and BCor (Wagner et al., 2011). This co-repressor complex then recruits histone deacetylases, leading to transcriptional repression of target genes, including *ATR*, *TP53* and *CDKN1A*(Wagner et al., 2011). Interestingly both *Ncor1* and *Bcor* were also identified as CIS genes in the IM cohorts.

The pathogenic role of deregulated *Bcl6* expression in lymphomagenesis has been demonstrated in a transgenic mouse model in which the full length of the murine *Bcl6* coding sequence was expressed under the control of the immunoglobulin heavy chain (Iμ) promoter(Cattoretti et al., 2005). *Bcl6* transgenic mice had an increased number of GCs and at six months of age, showed abnormal polyclonal B cell expansions, with partial effacement of the follicular architecture of lymphoid organs consistent with a benign lymphoproliferative disorder. From 13 months onwards, *Bcl6* transgenic mice showed increased mortality due to clonal B cell lymphomas, which had a mature B cell phenotype (IgM+IgD+CD43-) with histology similar to human DLBCL in most cases (Cattoretti et al., 2005).

Another CIS common to both screens and with a well described role in lymphomagenesis is the mucosa associated lymphoid tissue lymphoma translocation gene 1 (*Malt1*). The detected integrations were in the forward orientation and predominantly in intron 6 of *Malt1* and are likely to be activating integrations. *Malt1* was initially described due to its recurrent translocation, t(11;18)(q21;q21), in MALT lymphoma(Dierlamm et al., 1999). The translocation created a fusion protein and resulted in constitutive activation of the canonical NF-κB pathway. It was subsequently recognised that *MALT1* binds to and synergises with *BCL10* to promote canonical NF-κB activation as part of the CARD11-BCL10-MALT1

(CBM) complex (McAllister-Lucas et al., 2011). This activates the inhibitor of kappa B kinase (IKK) complex, leading to degradation of inhibitor of kappa-Bα (IκBα) and release of active NF-κB dimers into the nucleus. Therefore, as part of the CBM complex *MALT1* activates canonical NF-κB activation downstream of both T and B cell receptor stimulation, resulting in cytokine production and cellular proliferation in response to antigen stimulation(McAllister-Lucas et al., 2011). In addition to its role in MALT lymphoma, it has recently been recognised that *MALT1* is important in the pathogenesis of DLBCL, even in the absence of described mutations or translocations in this disease. Activated B cell-like DLBCL (ABC) is characterised by constitutive NF-κB activity and in an shRNA screen, *CARD11*, *BCL10* and *MALT1* were found to be essential to the survival of ABC-DLBCL cells (Ngo et al., 2006). Furthermore, inhibition of *MALT1* protease activity was toxic for ABC-DLBCL(Ferch et al., 2009). In this context it is also noteworthy that other genes involved in the regulation of NF-κB transcription factor complexes were also identified as CIS genes in this IM screen including *NFκBIZ* and *NFκB1*.

Mir17hg (mir17-92 cluster of microRNAs) is another CIS which has previously been implicated in lymphomagenesis. In humans, *Mir17hg* is in a region at 13q31-q32 which is frequently amplified in DLBCL, follicular, mantle and other lymphoma subtypes (He et al., 2005). Overexpression of *Mir17hg* in transgenic mice resulted in increased lymphocyte proliferation and decreased cell death, and lymphoproliferative and autoimmune disease (Xiao et al., 2008). Overexpresssion of *Mir17hg* with *Myc* in *EuMyc* mice resulted in accelerated development of aggressive lymphoid malignancies (He et al., 2005).

Other CIS identified in our models included well-known transcription factors involved in B cell development including *Foxp1*, *Ets1*, *Pax5* and *Bach2*. *Foxp1* is involved in chromosomal translocations in DLBCL and MALT lymphoma, but it may also be overexpressed by other mechanisms(Goatly et al., 2008). It is reported to be expressed in 40-60% of DLBCL and its overexpression is associated with poor prognosis (Banham et al., 2005; Barrans et al., 2004). Recurrent gains at 11q24.3 were recently described in 23% of DLBCL cases and were associated with significantly higher expression of *ETS1*(Bonetti et al., 2013). *PAX5* is an important regulator of B cell differentiation, which, among other roles, activates *BACH2* and initiates the GC reaction. Translocations involving *PAX5* are described in human B cell lymphomas

(Cobaleda et al., 2007) and loss of heterozgosity of *BACH2* has been described in around 20% of B cell lymphoma(Ichikawa et al., 2014). The overexpression of *BACH2* has been associated with poor prognosis in DLBCL in one study (Ichikawa et al., 2014) although the reverse finding was reported in another (Sakane-Ishikawa et al., 2005).

Along with CIS genes with a well described role in lymphomagenesis, there were a number other CIS that warrant further investigation. The most frequently hit CIS in both cohorts was centred on the gene *Ras* responsive element binding protein 1 (*Rreb1*). *Rreb1* is a paralog of *ZNF821* and encodes a zinc finger transcription factor that binds specifically to the *RAS* responsive elements of gene promoters. It has previously been identified as a putative oncogene in a retroviral and a transposon insertional mutagenesis screen (Starr et al., 2009; Uren et al., 2008) and has been implicated in the pathogenesis of several tumours including childhood ALL (Xiao et al., 2014), thyroid (Thiagalingam et al., 1996), pancreatic (Costello et al., 2012) and colorectal (Kent et al., 2013) malignancies. A three way translocation involving MLL-ENL and *RREB1* was also recently described in paediatric AML (Tuborgh et al., 2013).

RREB1 Is thought to be activated by *RAS* signalling downstream of the MAPK pathway (Kent et al., 2013; Thiagalingam et al., 1996; Zhang et al., 2003). In a recent study of childhood ALL, *RREB1* was shown to bind to the *PTPRG* promoter and the methylation status of the *PTPRG* locus was found to be a complementary event in oncogenesis and was associated with *RAS* mutation status (Xiao et al., 2014). The differential susceptibility of Balb/c mice to pristane induced plasma cell tumours has been attributed to polymorphisms in the $p16^{INK4a}$ gene promoter that effect the *Rreb1* binding site and increase *Rreb1* mediated repression of $p16^{INK}$ (Zhang et al., 2003). *RREB1* silencing of ZIP3 is thought to be an early event in the evolution of pancreatic adenocarcinoma, which results in reduced zinc levels in the ductal and acinar epithelium (Costello and Franklin, 2013). A similar role in the down-regulation of the zinc transporter *hZIP1* has been described for *RREB-1* in prostate cancer (Milon et al., 2010). *RREB1* also represses miR-143/miR-145 promoter activity(Kent et al., 2010) and the loss of miR-145 is seen in *K-RAS* mutated pancreatic cancers(Sureban et al., 2013) and colorectal carcinoma(Kent et al., 2013). *K-RAS* medicated activation of *RREB1* is thought to directly inhibit transcription of the miR-143/145 cluster. This appears to be a feed forward mechanism to potentiate *RAS* signalling in these

tumours as *K-RAS* and *RREB1* are targets of miR-145 and therapeutic restoration of miR-145 abrogates tumorigenesis.

Despite the numerous recent accounts of the oncogenic role for *RREB1*, I could not identify any reports of mutations or dysregulation of *RREB1* in human B cell lymphomas. Nevertheless, in our IM cohorts 23/33 *Vk*hPB* IM and 25/65 *Vk*MYC-TA-hPB* IM tumours had integrations within the *Rreb1* CIS in the top ten hits and these were almost universally in the forward orientation suggesting they are activating mutations. Although *Rreb1* has been described as a CIS in some previous IM screens (Starr et al., 2009; Uren et al., 2008) it was only seen in small numbers of cancers and its specificity to any particular cancer was not determined. By contrast, in my data *Rreb1* was the most commonly hit CIS and this may suggest an important role of this gene as an effector of activated Ras and MAPK signalling. Mutant genes signalling through Ras such as *EGFR*, *FGFR2*, *KRAS* and *BRAF* have been found to be recurrently mutated in mature B cell malignancies (Chapman et al., 2011; Vaqué et al., 2014) and at least in some instances their key oncogenic effects may operate through *RREB1* overexpression. The ability of transposons to activate gene expression in a very different manner to naturally occurring human cancer mutations may be the reason this gene was identified by our studies, but has not been found to be recurrently mutated in human lymphomas.

Notably there were other CIS involving *RAS* pathway genes in the *Vk*MYC-TA-hPB* cohort including *Nras* itself. Although *Csde1* is listed as the central gene in this CIS, the majority of integrations are either in the terminal exon (exon 20) of *Csde1* or intron 1 of *Nras* and these are almost universally in the forward orientation for *Nras* suggesting activating integrations. These *Nras* integrations were amongst the top hits in samples that did not have *Rreb1* integrations as a top 25 hit, suggesting that mutations in the two genes were mutually exclusive, in turn alluding to them having similar and therefore redundant effects. Similarly, two further tumours had top ten integrations in the *Ras* activator *Rasgrp2*, and neither of these tumours had *Rreb1* integrations in significant number. Another CIS in the *MYC-TA-hPB* cohort involved *Rasgrp3*, a member of the RAS family of GTPases, and this did co-occur with *Rreb1* as one of the top hits in some, but not all of the tumours in which it was present. Notably, none of these sites were detected as CIS in the *hPB* cohort, which had a much higher rate of *Rreb1* integrations.

Evidence from a conditional *Kras* mutant mouse model is that *Ras* activation is insufficient to transform primary germinal centre B cells. These mice had a conditional $KRas^{G12D}$ mutation that was induced by two different *Cre* recombinases thought to be specific to germinal centre B cells (Cγ1-Cre and AID-Cre). The mice developed T cell lymphomas, lung adenomas and sarcomas but not B cell lymphomas or plasma cell tumours despite the presence of activated *Kras* in the B lineage cells (Mullins et al., 2013). Even in a tumour-prone Arf-null genetic background or following sub-lethal irradiation, the $KRas^{G12D}$ mutation failed to induce a mature B cell malignant phenotype, which suggests that *Ras* activation is insufficient to transform primary germinal centre B cells (Mullins et al., 2013). *Rreb1* integrations were also detected in several of the IM mice that died with normal spleen size and no detectable lesion on histopathology at death, which suggests that this is an early and commonly occurring integration in these mice, but is insufficient itself to generate lymphoma.

Another CIS gene of interest is *Cflar* (CASP8 and FADD-like apoptosis regulator), also known as *cFLIP*, which was hit by insertions which were almost universally in the forward orientation in intron 2. This gene was originally identified as a competitive inhibitor of death signalling as it blocks recruitment of caspase-8 to the death inducing signalling complex (DISC) (Budd et al., 2006). Subsequently it has also been found to form a heterodimer with caspase-8 and to activate caspase-8. This heterodimer also links T cell receptor signalling to activation of NF-κB through the CMP complex (Budd et al., 2006). Notably, lymphoma cell lines with constitutively activated NF-κB are resistant to induction of apoptosis and *Cflar* is one of the down-stream targets whose overexpression is associated with this resistance (Bernal-Mizrachi et al., 2006). Retroviral expression of *Cflar* in B cells was reported to reduce CD95 mediated B cell death and cause retention of activated B cells in the germinal centres, while transduction of B-lymphoma cells with viral *FLIP* resulted in highly aggressive tumours which were resistant to CD95 induced cell death (Budd et al., 2006; Djerbi et al., 1999). The expression of *Cflar* has been associated with poor outcome in Burkitt lymphoma and in DLBCL of both GC and non-GCB sub-types (Harris et al., 2012; Valnet-Rabier et al., 2005).

In conclusion, although the *MYC-TA-hPB* and *hPB* IM mice did not develop plasma cell tumours, expression of the transposase under the control of the Vκ gene

regulatory elements led to the development of mature B cell lymphomas as the predominant tumours in our mice. The identified CIS genes were similar between the two cohorts, suggesting that the *MYC* transgene was not operative. The identified driver genes included transcription factors, microRNAs and apoptosis regulators, many of which have been found to be de-regulated in human lymphoma. These CIS genes also include some, such as *Rreb1*, which could be the focus of future studies, to better understand the mechanisms by which they contribute to lymphomagenesis and to develop targeted therapies.

7. Discussion

There were two central themes to this thesis. Firstly, the use of the *Npm1*^{cA}*/ GRL*, *Vk*MYC-TA-hPB* and *Vk*hPB* IM models as tools for discovery and validation of tumour associated genes. Secondly, the use of IM as a tool for studying the clonal evolution and architecture of cancer and its relation to human malignancies.

7.1 Transposon IM as a tool for cancer gene discovery

The transposon IM models presented in this thesis were analysed to identify CIS genes as putative drivers for these tumours. For the purposes of this discussion the *Vk*MYC-TA-hPB* and *Vk*hPB* CIS will be considered together unless specifically stated otherwise. This is because there was significant overlap in the CIS identified in these screens and there was no convincing evidence that the *MYC* transgene had a strong collaborative effect in driving the *Vk*MYC-TA-hPB* tumours.

It is notable that in both the AML and lymphoma IM screens over 75% of tumours had integrations in the single most frequently hit CIS (*Csf2* and *Rreb1* respectively) as one of the top 100 integrations on TraDIS analysis. Furthermore, in approximately 50% of these tumours, integrations in these genes were amongst the top ten hits. This indicates that these loci were frequently hit, and that these integrations were strongly selected for in the respective tumours. Neither *Csf2* nor *Rreb1* have been reported to be mutated in the corresponding human tumours and therefore some may dismiss these genes as irrelevant to the human diseases. However, the validity of these novel driver integrations is supported by the fact that several other well-known human disease-associated genes were also identified in the CIS list for each of the cohorts. Immediately identifiable examples include recurrent integrations in *Flt3*, *Mll1* and *Nf1* in the myeloid leukaemia mice, and *Bcl6*, *Mir17hg* and *Malt1* in the lymphoma cohorts. Several other CIS integrations identified in the lymphoma screen have also recently been identified as significantly mutated genes in human and cell line sequencing studies of diffuse large B cell lymphoma (DLBCL). These include *GNA13*, *TNFRSF14*, *CIITA*, *POU2F2*, *EBF1*, *ETS1* and *TNFAIP3* (Lohr et al., 2012; Morin et al., 2013; Pasqualucci et al., 2011; Zhang et al., 2013)

A possible explanation for the fact that neither of the highly prevalent top hits have been identified in the respective human cancers, may be found in differences between transposon-induced mutagenesis and sporadic somatic mutations found in human tumours. Transposons can cause gene knockout, or overexpression of a full length or truncated gene product, but they cannot introduce point mutations, which are a common mechanism of somatic mutation in human disease. Although this is often put forward as a weakness of transposon IM screens, it may also be a major strength. Transposons are likely to identify the targets genes or pathways of point mutations seen in human cancer and therefore can help to inform understanding of the biological mechanisms involved in tumourgenesis. Although there are rare examples to the contrary, such as *Bcl6* in DLBCL *(Wang et al., 2002)*, it is unusual for point mutations to directly up-regulate human genes. The effect of the transposon integrations around both *Csf2* and *Rreb1* appears to be gene up-regulation, and such an effect cannot be recapitulated by point mutations in these genes. Therefore, it is not surprising that such mutations have not been identified in the human diseases. Translocations are a recurrent type of mutation associated with gene overexpression in human haemopoietic cancers, but gene targets for translocations are relatively limited and regulatory elements or the location of these genes may protect them from this mechanism of mutation. For example, the very close genomic proximity (10kb) and co-regulation of the *CSF2 and IL3* genes in human and mouse, could be preventing any translocation from upregulating one gene without disrupting the other. This could therefore "protect" the locus from such an event. By contrast, the small size of transposons enables them to overexpress *Csf2* without a significant effect on *Il3* expression.

The absence of detectable mutations in *Csf2* or *Rreb1* in the human diseases may reflect the difficulty of achieving up-regulation of these genes by the mechanisms of mutation that regularly occur in the human genome, however this does not make them irrelevant as potential therapeutic targets. For example, the bromodomain and extraterminal (BET) protein, *BRD4* is a general transcriptional regulator that is rarely mutated in human cancers and recurrent mutations in this gene have not been described in haemopoietic malignancies (Shi and Vakoc, 2014). However, pharmacological inhibition of BET proteins shows therapeutic activity in a variety of human cancers, including diverse genetic subtypes of haematological malignancies

and the protein product of the wild-type *BRD4* gene is believed to be the therapeutic target(Dawson et al., 2011; Shi and Vakoc, 2014). It is possible that the transposon integrations in *Rreb1* and *Csf2* are highlighting important common pathways in the pathogenesis of human haematopoietic malignancies which could be targeted therapeutically.

Csf2 is the gene which encodes GM-CSF, a cytokine that regulates myeloid cells by binding its receptor and activating downstream signalling pathways. *Csf2* was the most frequently hit CIS in both the GRL and the published GRH model(Vassiliou et al., 2011) and the transposon was in the forward orientation relative to the gene, suggesting these are activating integrations. In the GRH model these integrations were demonstrated to result in marked overexpression of *Csf2* mRNA and increased GM-CSF levels in leukaemia cell supernatants(Vassiliou et al., 2011). Although the role of GM-CSF has not been extensively evaluated in human myeloid leukaemia, there is some evidence that up-regulation of GM-CSF signalling has a pathogenic role. High expression levels of the common beta chain subunit of the GM-CSF receptor are frequently found in *FLT3-ITD* mutant AML(Riccioni et al., 2009) and hypersensitivity to GM-CSF is a feature of juvenile myelomonocytic leukaemia (JMML)(Bunda et al., 2013). Also, mutations affecting genes involved in GM-CSF receptor signalling including *PTPN11*, *NRAS, KRAS, NF1* and *CBL* are seen in both JMML and AML(Ward et al., 2012). GM-CSF is required for the in vitro proliferation of most leukaemia cell lines from human and mouse myeloid leukaemias(Metcalf, 2013; Metcalf et al., 2013). It is also noteworthy that *Ets1*, which was also identified as a CIS on the Illumina analysis, is known to have a role in regulating the GM-CSF promoter(Thomas et al., 1995) and has recently been reported to mediate autocrine GM-CSF production in the KG1a leukaemia cell line(Bade-Döding et al., 2014).

The frequency with which the *Csf2* integrations occurred in our models indicates this is an important event in the pathogenesis of these mouse leukaemias and the role of *Csf2* signalling is therefore a focus of ongoing research in our laboratory. The finding that *Csf2* integrations, when present, are typically among the top ten hits, suggests that this integration is selected for and that over-expression of *Csf2* in a minor sub-clone of cells is insufficient to drive leukaemia proliferation in the bulk tumour. It remains to be determined if the up-regulation of *Csf2* is having a cell-autonomous effect, with the leukaemic cells secreting GM-CSF which then binds the GM-CSF

receptor on their surface for its action. An alternative possibility is raised by recent work highlighting a non-cell autonomous role of AML mediated M-CSF, acting on stromal cells and causing them to secrete cytokines that can stimulate leukaemic cell growth(Ben-Batalla et al., 2013).

To further investigate whether the effect of *Csf2* is dependent on a leukaemia-stromal cell interaction we have recently imported B6.129S1-Csf2rb1^{tm1Cgb}/Csf2rb^{tm1Clsc}/J mice. These mice have a knockout of the βc and β-IL3 loci which are required for formation of high affinity receptors for GM-CSF, IL-3 and IL-5 (Nicola et al., 1996; Robb et al., 1995; Scott et al., 2000). IM tumour cells with *Csf2* integrations have recently been transplanted into these mice. If the *Csf2* integrations are acting in a cell autonomous manner, we anticipate these tumours will engraft, however if the GM-CSF effect is dependent on a tumour-stroma interaction, they would not engraft or would do so much more slowly. If the findings suggest a non-cell autonomous effect, confirmation could come from experiments to suppress expression or knock-out the gene for GM-CSF receptor in AML cells and demonstrate that this does not affect tumour growth in a normal host.

The other CIS which was identified in the myeloid leukaemia cohort and is the focus of ongoing work in our laboratory is *Nup98*. Translocations, but not point mutations have been described in *NUP98* in human haematopoietic malignancies. *Nup98* was a frequently hit CIS gene in the 454 analysis and in serially bled mice integrations in *Nup98* were often evident for several weeks prior to the development of leukaemia. On TraDIS analysis *Nup98* was one of the top ten hits in four of the leukaemia samples and it persisted on transplantation in all recipient mice from 19.2b as one of only four hits with high read coverage. The transposon integrations in *Nup98* were bi-directional and spread through multiple introns, suggesting they are inactivating integrations. Although *NUP98* fusion proteins are thought to act as aberrant transcriptional regulators(Gough et al., 2011), *NUP98* is part of the nuclear pore complex and it is possible that disruption of nuclear-cytoplasmic transport may be having an oncogenic effect in these tumours. As *Npm1^{cA}* mutations are known to cause cytoplasmic dislocation of Nucleophosmin our hypothesis is that these mutations either exacerbate its mislocalisation or alter the localisation of its protein partners. We have therefore generated a *Nup98* conditional knockout mouse, which

has recently been crossed with the *Npm1*^{cA} mutant mice to study their interactions in haematopoiesis and leukaemogenesis.

In the *Vk*MYC-TA-hPB* and *Vk*hPB* mice, the most frequently hit CIS gene was *Rreb1. Rreb1* encodes a zinc finger transcription factor that binds to the *RAS* responsive elements of gene promoters at the consensus sequence CCCCAAACCACCCC (Thiagalingam et al., 1996). *RAS* genes are the most frequently mutated oncogenes in human cancers, and yet there is still much to learn regarding the downstream oncogenic effects of their mutations(Stephen et al., 2014). RAS-GTPs activate multiple downstream effectors, including the RalGDS, Raf and PI3 kinase pathways (Stephen et al., 2014). So far, *RAS*-driven tumours have proven relatively resistant to therapy, and feedback systems have thwarted tumour responses to farnesyltransferase, Raf, MEK and PI3K inhibitors(Stephen et al., 2014). It is plausible that the transposon-mediated activation of *Rreb1* is affecting a subset of downstream *RAS* pathways and that this indicates a potential therapeutic target for modulating *RAS* signalling. A role of *Rreb1* has already been demonstrated in several solid tumours (Costello and Franklin, 2013; Kent et al., 2013; Sureban et al., 2013).

RAS mutations are reported to occur rarely in human mature B cell non Hodgkin lymphomas (Lohr et al., 2012; Nedergaard et al., 1997), although they are common in multiple myeloma, in which they have a prevalence of around 30% (Chng et al., 2008; Liu et al., 1996). The relative absence of these mutations in mature B cell lymphomas may reflect the extensive intracellular effects of *RAS*. Perhaps direct mutation of the *RAS* genes disrupts critical intracellular pathways in germinal centre B cells resulting in growth disadvantage or even apoptosis, rather than activating *RAS* pathways involved in lymphomagenesis. In keeping with such a scenario, in hairy cell leukaemia heterozygous mutations in *BRAF*, which cause constitutive kinase activation and increased MAPK signalling, are almost universal, yet there is no evidence for mutations in *RAS* itself (Tiacci et al., 2011). Amongst the lymphomas *BRAF* mutations are highly specific for hairy cell leukaemia, although they have also been reported at low frequency in MM(Chapman et al., 2011). This is one example of a pathogenic mutation affecting a specific pathway downstream of *RAS* that occurs with high prevalence in a sub-type of a mature B lymphoid disease. It is plausible that deregulation of specific pathways downstream of *RAS* are found in other B cell lymphomas in the absence of mutations in *RAS* itself. Overexpression of *Rreb1* by

transposon integrations may modulate a subset of the downstream pathways from the many that can be disrupted by direct *RAS* mutations. In this context, the human equivalent of *Rreb1* overexpression could be mutations of specific RAS pathway genes or target genes of RAS responsive element.

The downstream transcription targets of *Rreb1* in these IM induced lymphomas are not clear, but the prevalence of this integration across so many tumours indicates that it is worthy of further investigation. Unfortunately, due to the long latency for tumour development in these mice, there was insufficient time to further investigate the mechanisms through which *Rreb1* may be contributing to lymphoma formation during my PhD studies. Future work would include confirming overexpression of *Rreb1* mRNA in the IM mice and studying the gene expression profiles (GEP) of these mice to investigate potential targets. However, the selection of an appropriate control group for such an analysis is challenging. One option would be to use samples from mice that did not have overt lymphoma at death, but TraDIS analysis of spleen DNA from such mice also revealed frequent *Rreb1* integrations. It is presumed that this integration arises early in the pathogenesis of the transposon-driven lymphomas, but is not sufficient in itself for lymphoma formation. An alternative approach would be to compare GEP in lymphomas with and without integrations in *Rreb1*. However, many of the mice that did not have *Rreb1* integrations had hits in other *Ras* pathway genes, including *Nras*, *Rasgrp2* and *Rsgrp3*, and it is likely that these represent alternative mechanisms for activating overlapping pathways. It would also be important to investigate *Rreb1* gene expression levels in human mature B cell lymphomas, which could be done using publicly available datasets. If these investigations gave further supportive evidence of a potential pathogenic role for *Rreb1*, the next step could be to try to generate cell lines from these tumours and show that their growth is *Rreb1* dependent, or to knock down *Rreb1* and demonstrate that this inhibits lymphoma growth *in vivo* in a transplant setting.

7.2 Transposon IM as a tool for studying clonal evolution

The results of the studies described here also give new insights into the biology of transposon IM, which are of relevance for the analysis of future transposon screens performed for cancer gene discovery. Rather than a homogenous population,

transposon-driven tumours are dynamic, heterogenous collections of cells, which are constantly evolving and acquiring new integrations.

One important finding from this study is that in the *Npm1*^{cA} mutant mice AML typically develops without major antecedent abnormalities in the blood parameters, akin to *de novo* human AML. The sudden change in the white cell count (WCC) occurs despite clear evidence of tumour associated integrations for weeks, and sometimes months, prior to the onset of leukaemia. This sudden shift from a normal to an abnormal blood count was a surprising finding. Although the majority of human cases of AML arise *de novo*, a significant proportion occur in patients with pre-existing haematological disorders such as myelodysplastic or myeloproliferative neoplasms, in which there are detectable somatic mutations in haematopoietic cells. Compared to most adult tumours AML has a low burden of somatic mutations, which may reflect the paucity of external mutagens in the HSC compartment or an unusually high level of protection against them. Although the haematopoietic compartment in these mice was a target for mutagenesis, only one mouse (7.5c) developed FBC abnormalities suggestive of a myeloproliferative disorder in the pre-leukaemic phase. The rarity of myeloproliferative changes in the mouse peripheral blood samples concords with the fact that human *NPM1c*-mutant AML does not usually have an antecedent pre-leukaemic phase, although this can be seen rarely when mutations in a small set of genes co-occur with *NPM1c* as in the case of CMML transformation described in Chapter 3.

The analysis of the serial blood and tumour samples clearly demonstrates that transposon mobilisation begins early and is a continuous process, so what is the trigger for the rapid change in the peripheral blood parameters? It is possible that the full complement of leukaemia inducing integrations is acquired early and that the bone marrow is abnormal for a period of time without significant spill of malignant cells into the peripheral blood. However, the evidence from the serially bled cases is that the final hit, which provided the leukaemia clone with its full complement of driver mutations, occurred just before the rapid increase in WCC. For example, in tumour 6.4a the top hits by read number in the final tumour were in intergenic regions of chromosomes 7 (7:932553553) and 16 (16:42681152) and in the genes *Dmxl1* and *Iqgap2*. These were first detected in the week 27, 33, 35 and 37 blood samples respectively, however the top hit in all of the transplants was another

intergenic integration on chromosome 7 (7:145053139). This integration was not detected until the final blood sample at week 43, although it was one of the top ten hits by read number in the leukaemia . The 7:145053139 integration is proximal to *Ccnd1*, a known oncogene, which is overexpressed in AML and is therefore a plausible driver integration(Wang et al., 2009). Both the 7:145053139 and a *Csf2* integration were among the top three hits in all of the transplant recipient tumours. The timing of the *Csf2* integration in this tumour is uncertain as the pre-leukaemic samples were not sequenced by TraDIS and it was not detected on 454 sequencing in the serial blood or final tumour samples, most likely because the nearest *Mbo1* restriction site was over 700 bases away. A second example is tumour 6.4g. The major integrations in the primary tumour also persisted on the serial transplants and many of these were detected in several blood samples prior to tumour development on the 454 analysis. These included integrations at 9:21989714 (week 67), *Bach2* (week73), 14:120558731(week73), 5:3343787 (week 75) and *Ankrd17* (week75), but not the *Pou2f2* integration, which was first detected in the final blood sample (week 85), but was a major hit in all of the recipient tumours. *Pou2f2*, otherwise known as *Oct2*, is a homeobox containing transcription factor, which is overexpressed in a subset of AML patients and has been associated with poor prognosis(Advani et al., 2010). Tumour 6.4h is a third example, in which two integrations that were dominant in the final tumour and were shared by most of the transplants, were first detected at week 25 (8:45103026) and week 27 (Bmi1), whereas two further apparent driver integration, involving *Pax5* and *Ikzf1*, were first detected at 31 weeks, when the WCC was starting to rise. Therefore, in all three examples, the rapid rise in white cell count is associated with the first detection of additional integrations in plausible driver positions, which also persist as part of the major cell population in the recipient tumours.

It is difficult to draw major conclusions about the order of acquisition of driver mutations, given the small number of tumours and high level of variation in apparent drivers between them. Some integrations, such as those in *Flt3* and *Mll1* were typically late, while the serial CIS analysis revealed that the *Csf2*, *Nup98* and *Nf1* CIS were all identifiable at least two weeks before the onset of leukaemia. However, integrations at these sites still occurred as both early and late events, and the timing

of the integrations did not seem to influence whether or not these were top ten hits in the primary tumour on TraDIS sequencing.

The low copy *SB* IM screen was characterised by a longer latency to leukaemia development than the high copy cohort, consistent with a lower rate of mutation acquisition. Although the latency to tumour development varied widely, in most cases overt leukaemia developed within a few months of the mouse starting to accumulate integrations which persisted on the serial blood samples (presumed to reflect the development of a persistant pre-leukaemic clone). The variation in leukaemia latency seemed to largely reflect the lag to the first hit that persisted on subsequent samples, although there was also variation in the time it took took to accumulate additional persisting integrations. Mice 7.7b and 6.4g which had no, or reduced doses of pIpC, had long latencies to leukaemia. In these mice there were very few integrations which were shared by successive blood samples in the first six months of sampling, but they still accumulated several persisting integrations at later time points. These observations suggest that for the given mutagenesis rate, once the initiating mutation has been established in a clone, leukaemogenesis follows a deterministic models with regards to the leukaemia latency.

The step wise accumulation of persisting transposon integrations over time in some of the serially bled mice is a significant finding. The continuous detection of specific transposon integrations on fortnightly blood tests indicates both that the transposon integration persists at that site at least in a proportion of cells, and that that clone is continuously contributing to the production of circulating blood cells. It is unlikely that all of the persisting integrations are tumour drivers. However, it is probable that when a number of mutations are acquired in the same "step", such steps correlate with the acquisition of a 'driver' integration, with the majority of integrations representing passengers which were present in the cell at the time of acquisition of the driver. This is difficult to prove, as the allocation of each individual integration into categories of driver and passenger lesions cannot be fully substantiated. However, typically only a small proportion of the persisting integrations from each step were also found in the transplant recipient tumours.

The reasons that non-driver integrations would persist on serial sampling have been discussed in chapter 4. Integrated transposons are free to re-mobilise, but the

excision of transposons from 'driver' positions is selected against, as cells in which this happens will lose any growth or survival advantage that was due to the transposon. Although the remobilisation of passenger lesions is not selected against passenger lesions are unlikely to remobilise from all clonal cells before their next cell division if the cells are rapidly dividing (see figure 4.18). Unfortunately, there was insufficient DNA remaining from most of the pre-leukaemic blood samples to allow for re-sequencing with our quantitative approach. However, this was possible for some samples and in these cases there were examples of persisting integrations with increasing, stable or falling read proportions in the serial blood samples. Many of the integrations that persisted as top hits in the transplants tended to be stable or increase over time. However, as demonstrated in figure 4.18, this does not imply that all the integrations with stable read proportions are necessarily drivers, or that those with a falling read percentage are necessarily passengers. Some may be drivers in clones that were overtaken by other clones over time.

The results from the serial transplant experiments have helped to clarify which integrations are likely to be acting as driver mutations in individual tumours. Typically these integrations persist in multiple transplants and are found in high read number in the recipient tumours. Most of these integrations were also in high read number in the primary tumour, but this is not universally the case as demonstrated by mouse 16.3f. In this example the transplant experiments seemed to select out a clone which was only a small sub-clone in the tumour of the primary mouse. All of the integrations that dominated the transplant tumours were first detected in the final blood sample from the original mouse and represented less than 1% of the total reads in the primary tumour. It is unlikely that these integrations arose in the same clone as the intergenic chromosome 11 and mmu-mir-29b-2 integrations which were the top hits in the primary tumour, each corresponding to about 9% of the total reads, as these were not found in the recipient tumours. The presence of more than one clone which was able to drive leukaemia formation, in the mass tumour population, was clearly demonstrated in mouse 21.3j in which transplant recipients of single-cell derived colonies had a different *Csf2* integration to the one which predominated the bulk transplants. Therefore, although the persistence of a transposon integration in a high percentage of reads in multiple transplants implies that it is either a driver, or co-occuring in the same clone as a driver; the loss of integrations in recipient

tumours does not exclude these from being a driver. It may have been occurring in a different clonal population, some of which are clearly also capable of generating leukaemia.

Although it is tempting to try to draw conclusions about the collaboration of integrations based on their co-occurrence in transposon driven tumours, care must be taken to ensure such lesions are actually present within the same cell, rather than in independently arising clones within the tumour. Previously, CIS data generated using a restriction enzyme based sequencing approach had been used to try to identify genes which collaborate or are mutually exclusive in tumorigenesis (Vassiliou et al., 2011), but little attention could be paid to the clonality of these tumours. The serial quantitative data is useful in helping to determine which integrations are likely to be co-occurring in tumour sub-clones. For example, in tumour 16.3e, which was atypical because there were so many integrations which persisted in the recipient tumours, on the serial TraDIS data two groups of mutations could be distinguished by the pattern in read proportion. One group of integrations, which included the *Pax5*, *Dock10*, *Pik3r1*, *E103008A19Rik* and intergenic integrations on chromosome 18 seemed to be falling in read proportion in the late serial bloods and were in lower proportion in the final tumour (10th to 16th ranked integrations), while the two intergenic integrations in chromosome 7 and one in chromosome 5 were rising in prominence in the late serial blood samples and were the top three hits in the primary, and among the top hits in most of the recipient tumours. Such serial quantitative data can help tease out which integrations are co-occurring and which may be in separate sub-clones.

The problems of identifying which mutations are acting as drivers and defining which mutations are co-occurring within a clone are not unique to transposon driven tumours. Our use of the serial quantitative data to make inferences about the sub-clonal architecture of transposon driven tumours is akin to the use of allele burden in human genome/exome sequencing. Although the number of mutations required for tumour formation is thought to be lower in AML compared to many adult tumours, the human case presented in chapter 3 highlights that in some people at least, multiple AML associated mutations can be identified several weeks before the development of clinical features of this disease, and that a large number of AML associated drivers can co-occur in human leukaemia samples. Furthermore, the different

patterns of mutational burden that were identified in the relapse samples reinforces that driver mutations found in human sequencing are not necessarily co-occurring at a single cell level. The findings with regard to sub-clonal architecture and clonal evolution in the IM mouse model are not dissimilar to many of the observations in the human case presented here.

The transplant experiments also highlight the low frequency of tumour initiating cells within the spleen cell population. Not all of the cells in the mixed spleen cell population used in the transplants will act as leukaemia initiating cells (LIC) and the inconsistent tumour engraftment in the 100 and even the 1000 cell transplants suggests that the proportion of LIC is quite small. On serial transplantation of a million mixed tumour cells, only a small set of recurrent integrations were consistently detected, suggesting that these include the driver mutations for both the original and the re-emergent clones. Although in some cases it is likely that more than one leukaemia clone engrafted, in others this may not have been the case and a similar pattern of persisting integrations was often seen at reducing cell does down to 100 cells. The inconsistent engraftment of 100 cell transplants implies that the number of LIC is very small at this cell dose, which in turn suggests the major integrations in these recipient tumours are more likely to be co-occurring at a single cell level.

The most valid method for studying the sub-clonal composition of transposon driven tumours would be to study these integrations at the single cell level. The approach used here, was to generate single cell derived haematopoietic colonies and to transplant these into recipient NSG mice. However, the yield from this was low, with few mice developing tumours. A more cost and time effective approach would be to directly sequence a number of single cell derived colonies from each primary tumour, to directly validate which major integrations are co-occurring at a single cell level. I attempted this using a 454 sequencing approach, but this was unsuccessful as most of the colonies shared a panel of integrations, which appeared to be artefactual, with few tumour specific integrations being mapped probably because of the limited amount of DNA. We are yet to try sequencing single cell derived colonies using the TraDIS protocol. To date all the samples have been prepared starting with 2µg of DNA, but there is no reason, in theory that this could not be attempted with less DNA.

In many of the mice it took about two months to develop leukaemia following the first detection of an apparent driver transposon integration that persisted in subsequent samples. This probably reflects a requirement for several co-operating mutations for leukaemogenesis. In tumour 21.3j the only integration that was shared by all the recipient tumours was the integration in *Csf2*, and yet it took seven weeks for the primary tumour to become apparent after this integration. It may be that in this case there were various secondary driver lesions that dominated in the different recipient tumours. It is also possible that other mechanisms, such as chromosomal translocations or footprint mutations could have provided additional driver hits later in the time-course. Chromosomal translocations may occur more commonly in the setting of the frequent double strand breaks induced by transposons and on FISH analysis of case 7.5h we did find significant chromosomal abnormalities. However, in the cases examined using CGH, which included one 21.3j recipient tumour, there was little evidence of copy number change and exome sequencing of tumour samples did not find evidence of the canonical *SB* footprint in any coding regions.

Going forward, it is not practical to extensively transplant every tumour in an IM screen to validate which are the driver integrations in individual tumours. However, this approach may be helpful to try to characterise the driver integrations in specific tumours in which there are no integrations in recognised tumour-associated or CIS genes. It is also a useful approach to help validate 'novel' drivers, such as *Rreb1*, which do not have correlates in human sequencing. Furthermore, transposon IM screens could be used as a platform to explore cancer therapies and mechanisms of drug resistance and for this application it may be more useful to characterise changes in the mutation spectrum in treated vs untreated mice which have been transplanted from the same primary tumour with well characterised transposon integrations. In addition to minimising the number of mice needed for such studies, this approach would allow investigation of therapies in different sub-groups of leukaemias. For example, tumours with a known *Flt3* integration in addition to the *Npm1*^{cA} mutation could be studied separately from those with *Mll1* integrations, allowing differences in drug response or resistance mechanisms in these sub-groups to be explored.

An important question facing the IM field is how to pick out the important drivers amongst the many background integrations detected using deep sequencing and

which are only present in rare cells. Is it reasonable to identify candidate tumour drivers at the level of individual tumours just based on the read frequency of the integrations, using shearing based sequencing approaches? Although the top hits are likely to be present in a clonal cell population, as discussed above, it cannot be assumed that all of these are driver integrations. Furthermore, I have shown in the leukaemia mice that sub-clones present within the bulk tumour may also have tumourigenic potential, and it is therefore difficult to set a threshold level below which integrations are unlikely to have a driver role.

My data from the lymphoma cohorts suggests that the spread of reads for the top integrations can be used to differentiate clonal from non-clonal tissue samples. However, in the few samples in which B cell repertoire analysis was performed, the pattern of fall in read count for the top integrations did not directly correlate with the size of the mutant clone detected. It would be interesting to further investigate the relationship between read number and clonality, by performing the B cell repertoire analysis in a larger number of samples or in transplanted samples where cells are likely to become more clonal.

In their analysis of solid tumours generated by a ubiquitously expressed *PB* transposon system, Friedel et al identified the candidate cancer genes as those which had enriched sequence read frequencies, compared to that from tail DNA controls(Friedel et al., 2013). In their study, between 9 and 25 insertions had enriched sequence read frequencies above their threshold of 0.37%, which was set by calculating the average read frequencies for the top ten hits in each tail sample. As (i) there was distinct enrichment of reads in tumour samples, (ii) the clonally expanded insertions included many well defined cancer genes and (iii) the analysis of related tumours showed strong correlation of read frequencies of clonally expanded insertions, they concluded that the identification of clonally expanded insertions is a valid method for identifying candidate tumour genes.

Friedel et al also analysed integrations in tissues from various organs without overt tumours and found some did carry more expanded insertions than tail tissue, with a range of between 3 and 23 expanded insertions and an average of 11 per sample(Friedel et al., 2013). Therefore, they estimate that around two thirds of expanded insertions in tumours may reflect pre-cancer insertions and conclude that

other methods are still required to validate tumour genes. Although I agree with their conclusion that all clonally expanded integrations are not necessarily drivers, I do not think the finding of clonally expanded integrations in non-malignant tissue alone is justification for this statement. To me, the finding of clonally expanded integrations in non-tumour tissues in mice with a ubiquitously active transposon is not surprising. Furthermore, these insertions are still causing clonal expansion, and are therefore potentially of relevance in tumourigenesis. The difference is that in samples in which overt cancer has not been recognised, the full complement of integrations required for transformation is yet to be reached in individual cells. The evidence from the serially bled mice is that integrations in CIS genes were not infrequent in the pre-leukaemic blood samples, although not all of these went on to become part of the major tumour clone. Whether such clones were outcompeted during tumour evolution because they failed to acquire additional hits, or whether the order of integration acquisition is important is uncertain.

Transposon insertions that are not driver integrations may still be clonally expanded in the tumour population. One mechanism for this would be if passenger insertions co-occur with the driver integrations and do not have time to disperse, due to the rate of tumour cell division exceeding the rate of transposon remobilisation (figure 4.18). Another reason this may happen would be if transposition activity ceased, meaning that a transposon could not remobilise, but other transposon integrations caused clonal expansion of the cell in which that occurred. In the analysis of the *SB* tumours I looked for evidence of fixed integrations with the 'neopartnership' assay and found little evidence to support this as a common mechanism of fixing integrations. However, the possibility that some of the clonal integrations were fixed due to mutation of the repeat sequence, cannot be excluded. It is also important to recognise that shearing based transposon sequencing methods such as TraDIS, do still have PCR steps to enrich for transposon integrations as part of the library preparation. Therefore, there will still be some biases in read quantification as a result of PCR amplification bias (e.g. due to GC content) and due to difficulties in mapping certain integrations (e.g. when the transposon integrates in a repetitive region).

The data from Friedel et al supports my thresholds of using the top 10 or 25 integrations only, to perform the CIS analysis. It is debatable whether this cut-off for included hits should

be based on a proportion of reads per integration, rather than a ranking by absolute read number. The ideal threshold based on read proportion may vary between samples depending on the clonality of the tumour, the degree of non-tumour contamination in the sample and the number of driver integrations "sharing" the reads. Therefore, the read proportion may not be any more relevant than rank when setting the threshold as to which hits to include in CIS analysis. Although the most appropriate means for doing this may be debated, such an approach can be used to give more weight to the top hits, rather than treating all integrations equally in CIS analysis.

7.3 Concluding remarks

Haematopoeitic malignancies evolve through the serial selection of cells with a growth advantage, in a multi-step process akin to natural selection. Mutations in leukaemia associated genes have been documented in the blood of healthy adults, without causing haematological disease. Although the development of leukaemia is not inevitable, such individuals are at higher risk of haematological malignancy and it may be that in the setting of particular combinations of mutations progression to AML becomes unavoidable. In the human sequencing case presented here, multiple mutations in leukaemia associated genes were found in a woman with CMML. Given the high mutational load and the rapid acquisition of additional *FLT3* and *RAS* mutations, it seems probable that the progression of her disease was almost inevitable.

The biology of the mutagenic processes in transposon IM screens differs to those seen in human tumours. In spite, or perhaps because of this, IM provides a powerful approach for the identification and validation of cancer genes and pathways that compliments human sequencing efforts. In this work I have shown that transposon mobilisation is a continuous process during leukaemia evolution. Integrations in CIS genes are not infrequent in the pre-leukaemic samples, but only some of these persist as dominant integrations in the primary and recipient tumours. Following acquisition of the final driver there is a sudden change in blood parameters. The driver status and co-occurrence of individual integrations can be delineated using serial transplant experiments and quantitative sequencing approaches. My data suggests that only a minority of transposon integrations behave as 'drivers'. However, in the case of the *Npm1*^{cA} IM mice the development of leukaemia is almost universal as the rate of mutagenesis is sufficient for the rapid accumulation of

multiple driver integrations within a single clone. In some cases at least, the acquisition of a full complement of leukaemogenic mutations occurs in multiple independent clones within a single mouse.

The power of the IM approach for cancer gene discovery is strengthened by the recent development of quantitative methods to analyse transposon integrations, which now allows differentiation of clonally expanded integrations from background integrations for the first time. The challenge going forward is to use this quantitative data to inform the CIS analysis. Using threshold cut-offs of 10 and 25 integrations from each tumour I was able to identify CIS in many known disease associated genes. With this approach in each model I identified highly recurrent integrations in genes not known to be mutated in the human diseases, but with plausible roles in disease pathogenesis including activating integrations affecting the putative novel lymphoma oncogene *Rreb1* in 75% of B-cell tumours. Such integration sites warrant further investigation which may provide new therapeutic targets for patients and their study is currently under way.

References

(2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

(2005). Nucleophosmin in Acute Myelogenous Leukemia. *New England Journal of Medicine* **352**, 1819-1820.

(2012). Cancer survival and prevalence in Australia: period estimates from 1982 to 2010. In *Cancer Series*, (ed. A. I. o. H. a. Welfare). Canberra.

Adams, J. M., Harris, A. W., Pinkert, C. A., Corcoran, L. M., Alexander, W. S., Cory, S., Palmiter, R. D. and Brinster, R. L. (1985). The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature* **318**, 533-8.

Advani, A. S., Lim, K., Gibson, S., Shadman, M., Jin, T., Copelan, E., Kalaycio, M., Sekeres, M. A., Sobecks, R. and Hsi, E. (2010). OCT-2 expression and OCT-2/BOB.1 co-expression predict prognosis in patients with newly diagnosed acute myeloid leukemia. *Leuk Lymphoma* **51**, 606-12.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Borresen-Dale, A.-L. et al. (2013). Signatures of mutational processes in human cancer. *Nature* **500**, 415-421.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology* **215**, 403-10.

Anderson, K., Lutz, C., van Delft, F. W., Bateman, C. M., Guo, Y., Colman, S. M., Kempski, H., Moorman, A. V., Titley, I., Swansbury, J. et al. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356-361.

Arai, F., Hirao, A., Ohmura, M., Sato, H., Matsuoka, S., Takubo, K., Ito, K., Koh, G. Y. and Suda, T. (2004). Tie2/Angiopoietin-1 Signaling Regulates Hematopoietic Stem Cell Quiescence in the Bone Marrow Niche. *Cell* **118**, 149-161.

Au, W. Y., Horsman, D. E., Gascoyne, R. D., Viswanatha, D. S., Klasa, R. J. and Connors, J. M. (2004). The Spectrum of Lymphoma with 8q24 Aberrations: A Clinical, Pathological and Cytogenetic Study of 87 Consecutive Cases. *Leuk Lymphoma* **45**, 519-528.

Avet-Loiseau, H. (2007). Role of genetics in prognostication in myeloma. *Best Pract Res Clin Haematol* **20**, 625-35.

Avet-Loiseau, H., Attal, M., Moreau, P., Charbonnel, C., Garban, F., Hulin, C., Leyvraz, S., Michallet, M., Yakoub-Agha, I., Garderet, L. et al. (2007). Genetic abnormalities and survival in multiple myeloma: the experience of the Intergroupe Francophone du Myelome. *Blood* **109**, 3489-95.

Bade-Döding, C., Göttmann, W., Baigger, A., Farren, M., Lee, K. P., Blasczyk, R. and Huyton, T. (2014). Autocrine GM-CSF transcription in the leukemic progenitor cell line KG1a is mediated by the transcription factor ETS1 and is negatively regulated through SECTM1 mediated ligation of CD7. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1840**, 1004-1013.

Bains, A., Luthra, R., Medeiros, L. J. and Zuo, Z. (2011). FLT3 and NPM1 Mutations in Myelodysplastic Syndromes: Frequency and Potential Value for

Predicting Progression to Acute Myeloid Leukemia. *American Journal of Clinical Pathology* **135**, 62-69.

Banham, A. H., Connors, J. M., Brown, P. J., Cordell, J. L., Ott, G., Sreenivasan, G., Farinha, P., Horsman, D. E. and Gascoyne, R. D. (2005). Expression of the FOXP1 transcription factor is strongly associated with inferior survival in patients with diffuse large B-cell lymphoma. *Clin Cancer Res* **11**, 1065-72.

Barrans, S. L., Fenton, J. A. L., Banham, A., Owen, R. G. and Jack, A. S. (2004). Strong expression of FOXP1 identifies a distinct subset of diffuse large B-cell lymphoma (DLBCL) patients with poor outcome.

Bashford-Rogers, R. J., Palser, A. L., Huntly, B. J., Rance, R., Vassiliou, G. S., Follows, G. A. and Kellam, P. (2013a). Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome research*.

Bashford-Rogers, R. J., Palser, A. L., Huntly, B. J., Rance, R., Vassiliou, G. S., Follows, G. A. and Kellam, P. (2013b). Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* **23**, 1874-84.

Baus, J., Liu, L., Heggestad, A. D., Sanz, S. and Fletcher, B. S. (2005). Hyperactive transposase mutants of the Sleeping Beauty transposon. *Mol Ther* **12**, 1148-56.

Beekman, R., Valkhof, M. G., Sanders, M. A., van Strien, P. M. H., Haanstra, J. R., Broeders, L., Geertsma-Kleinekoort, W. M., Veerman, A. J. P., Valk, P. J. M., Verhaak, R. G. et al. (2012). Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia. *Blood* **119**, 5071-5077.

Beerman, I., Maloney, W. J., Weissmann, I. L. and Rossi, D. J. (2010). Stem cells and the aging hematopoietic system. *Current Opinion in Immunology* **22**, 500-506.

Ben-Batalla, I., Schultze, A., Wroblewski, M., Erdmann, R., Heuser, M., Waizenegger, J. S., Riecken, K., Binder, M., Schewe, D., Sawall, S. et al. (2013). Axl, a prognostic and therapeutic target in acute myeloid leukemia mediates paracrine crosstalk of leukemia cells with bone marrow stroma. *Blood* **122**, 2443-2452.

Bereshchenko, O., Mancini, E., Moore, S., Bilbao, D., Månsson, R., Luc, S., Grover, A., Jacobsen, S. E. W., Bryder, D. and Nerlov, C. (2009). Hematopoietic Stem Cell Expansion Precedes the Generation of Committed Myeloid Leukemia-Initiating Cells in C/EBPα Mutant AML. *Cancer Cell* **16**, 390-400.

Bernal-Mizrachi, L., Lovly, C. M. and Ratner, L. (2006). The role of NF-κB-1 and NF-κB-2-mediated resistance to apoptosis in lymphomas. *Proceedings of the National Academy of Sciences* **103**, 9220-9225.

Betz, A. G., Milstein, C., Gonzalez-Fernandez, A., Pannell, R., Larson, T. and Neuberger, M. S. (1994). Elements regulating somatic hypermutation of an immunoglobulin kappa gene: critical role for the intron enhancer/matrix attachment region. *Cell* **77**, 239-48.

Bhayat, F., Das-Gupta, E., Smith, C., McKeever, T. and Hubbard, R. (2009). The incidence of and mortality from leukaemias in the UK: a general population-based study. *BMC Cancer* **9**, 252.

Bire, S. and Rouleux-Bonnin, F. (2012). Transposable Elements as Tools for Reshaping the Genome: It Is a Huge World After All! In *Mobile Genetic Elements*, vol. 859 (ed. Y. Bigot), pp. 1-28: Humana Press.

Birnbaum, R. A., O'Marcaigh, A., Wardak, Z., Zhang, Y.-Y., Dranoff, G., Jacks, T., Clapp, D. W. and Shannon, K. M. (2000). Nf1 and Gmcsf Interact in Myeloid Leukemogenesis. *Mol Cell* **5**, 189-195.

Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R. et al. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**.

Bonetti, P., Testoni, M., Scandurra, M., Ponzoni, M., Piva, R., Mensah, A. A., Rinaldi, A., Kwee, I., Tibiletti, M. G., Iqbal, J. et al. (2013). Deregulation of ETS1 and FLI1 contributes to the pathogenesis of diffuse large B-cell lymphoma. *Blood* **122**, 2233-41.

Boudry-Labis, E., Roche-Lestienne, C., Nibourel, O., Boissel, N., Terre, C., Perot, C., Eclache, V., Gachard, N., Tigaud, I., Plessis, G. et al. (2013). Neurofibromatosis-1 gene deletions and mutations in de novo adult acute myeloid leukemia. *American Journal of Hematology* **88**, 306-311.

Bradbury, D., Rogers, S., Reilly, I. A., Kozlowski, R. and Russell, N. H. (1992). Role of autocrine and paracrine production of granulocyte-macrophage colony-stimulating factor and interleukin-1 beta in the autonomous growth of acute myeloblastic leukaemia cells--studies using purified CD34-positive cells. *Leukemia* **6**, 562-6.

Braun, B. S., Tuveson, D. A., Kong, N., Le, D. T., Kogan, S. C., Rozmus, J., Le Beau, M. M., Jacks, T. E. and Shannon, K. M. (2004). Somatic activation of oncogenic Kras in hematopoietic cells initiates a rapidly fatal myeloproliferative disorder. *Proc Natl Acad Sci U S A* **101**, 597-602.

Brayton, C. F., Treuting, P. M. and Ward, J. M. (2012). Pathobiology of Aging Mice and GEM: Background Strains and Experimental Design. *Veterinary Pathology Online* **49**, 85-105.

Brennan, S. K. and Matsui, W. (2009). Cancer stem cells: controversies in multiple myeloma. *Journal of Molecular Medicine-Jmm* **87**, 1079-1085.

Budd, R. C., Yeh, W.-C. and Tschopp, J. (2006). cFLIP regulation of lymphocyte activation and development. *Nat Rev Immunol* **6**, 196-204.

Bunda, S., Kang, M. W., Sybingco, S. S., Weng, J., Favre, H., Shin, D. H., Irwin, M. S., Loh, M. L. and Ohh, M. (2013). Inhibition of SRC corrects GM-CSF hypersensitivity that underlies juvenile myelomonocytic leukemia. *Cancer Res* **73**, 2540-50.

Busque, L., Patel, J. P., Figueroa, M. E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A. et al. (2012). Recurrent

somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet* **44**, 1179-1181.

Butzler, C., Zou, X., Popov, A. V. and Bruggemann, M. (1997). Rapid induction of B-cell lymphomas in mice carrying a human IgH/c-mycYAC. *Oncogene* **14**, 1383-8.

Cadinanos, J. and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res* **35**, e87.

Caligiuri, M. A., Briesewitz, R., Yu, J., Wang, L., Wei, M., Arnoczky, K. J., Marburger, T. B., Wen, J., Perrotti, D., Bloomfield, C. D. et al. (2007). Novel c-CBL and CBL-b ubiquitin ligase mutations in human acute myeloid leukemia. *Blood* **110**, 1022-4.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A. and Stratton, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences* **105**, 13081-13086.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M.-L. et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109-1113.

Carlson, C. M., Dupuy, A. J., Fritz, S., Roberg-Perez, K. J., Fletcher, C. F. and Largaespada, D. A. (2003). Transposon mutagenesis of the mouse germline. *Genetics* **165**, 243-56.

Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. and Abkowitz, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460-4466.

Cattoretti, G., Pasqualucci, L., Ballon, G., Tam, W., Nandula, S. V., Shen, Q., Mo, T., Murty, V. V. and Dalla-Favera, R. (2005). Deregulated BCL6 expression recapitulates the pathogenesis of human diffuse large B cell lymphomas in mice. *Cancer Cell* **7**, 445-455.

Caudill, J. S. C., Sternberg, A. J., Li, C.-Y., Tefferi, A., Lasho, T. L. and Steensma, D. P. (2006). C-terminal nucleophosmin mutations are uncommon in chronic myeloid disorders. *British Journal of Haematology* **133**, 638-641.

Challen, G. A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J. S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y. et al. (2012). Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* **44**, 23-31.

Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J. P., Ahmann, G. J., Adli, M. et al. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-72.

Cheng, K., Sportoletti, P., Ito, K., Clohessy, J. G., Teruya-Feldstein, J., Kutok, J. L. and Pandolfi, P. P. (2010). The cytoplasmic NPM mutant induces myeloproliferation in a transgenic mouse model. *Blood* **115**, 3341-3345.

Chesi, M., Robbiani, D. F., Sebag, M., Chng, W. J., Affer, M., Tiedemann, R., Valdez, R., Palmer, S. E., Haas, S. S., Stewart, A. K. et al. (2008). AID-dependent activation of a MYC transgene induces multiple myeloma in a conditional mouse model of post-germinal center malignancies. *Cancer Cell* **13**, 167-80.

Chng, W. J., Glebov, O., Bergsagel, P. L. and Kuehl, W. M. (2007). Genetic events in the pathogenesis of multiple myeloma. *Best Pract Res Clin Haematol* **20**, 571-96.

Chng, W. J., Gonzalez-Paz, N., Price-Troska, T., Jacobus, S., Rajkumar, S. V., Oken, M. M., Kyle, R. A., Henderson, K. J., Van Wier, S., Greipp, P. et al. (2008). Clinical and biological significance of RAS mutations in multiple myeloma. *Leukemia* **22**, 2280-2284.

Choo, A., Palladinetti, P., Holmes, T., Basu, S., Shen, S., Lock, R. B., O'Brien, T. A., Symonds, G. and Dolnikov, A. (2008). siRNA targeting the IRF2 transcription factor inhibits leukaemic cell growth. *Int J Oncol* **33**, 175-83.

Cobaleda, C., Schebesta, A., Delogu, A. and Busslinger, M. (2007). Pax5: the guardian of B cell identity and function. *Nat Immunol* **8**, 463-470.

Collier, L. S., Adams, D. J., Hackett, C. S., Bendzick, L. E., Akagi, K., Davies, M. N., Diers, M. D., Rodriguez, F. J., Bender, A. M., Tieu, C. et al. (2009). Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality. *Cancer Res* **69**, 8429-37.

Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J. and Largaespada, D. A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6.

Costello, L. C. and Franklin, R. B. (2013). A Review of the Current Status and Concept of the Emerging Implications of Zinc and Zinc Transporters in the Development of Pancreatic Cancer. *Pancreat Disord Ther* **4**, 002.

Costello, L. C., Zou, J., Desouki, M. M. and Franklin, R. B. (2012). Evidence for changes in RREB-1, ZIP3, and Zinc in the early development of pancreatic adenocarcinoma. *J Gastrointest Cancer* **43**, 570-8.

Courville, E. L., Wu, Y., Kourda, J., Roth, C. G., Brockmann, J., Muzikansky, A., Fathi, A. T., de Leval, L., Orazi, A. and Hasserjian, R. P. (2013). Clinicopathologic analysis of acute myeloid leukemia arising from chronic myelomonocytic leukemia. *Mod Pathol* **26**, 751-761.

Cozzio, A., Passegué, E., Ayton, P. M., Karsunky, H., Cleary, M. L. and Weissman, I. L. (2003). Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. *Genes Dev* **17**, 3029-3035.

Cutts, B. A., Sjogren, A.-K. M., Andersson, K. M. E., Wahlstrom, A. M., Karlsson, C., Swolin, B. and Bergo, M. O. (2009). Nf1 deficiency cooperates with oncogenic K-RAS to induce acute myeloid leukemia in mice. *Blood* **114**, 3629-3632.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C. and Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proc Natl Acad Sci U S A* **79**, 7824-7827.

Dawson, M. A., Prinjha, R. K., Dittmann, A., Giotopoulos, G., Bantscheff, M., Chan, W.-I., Robson, S. C., Chung, C.-w., Hopf, C., Savitski, M. M. et al. (2011). Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* **478**, 529-533.

de Ridder, J., Uren, A., Kool, J., Reinders, M. and Wessels, L. (2006). Detecting Statistically Significant Common Insertion Sites in Retroviral Insertional Mutagenesis Screens. *PLoS Comput Biol* **2**, e166.

Delhommeau, F., Dupont, S., Valle, V. D., James, C., Trannoy, S., Massé, A., Kosmider, O., Le Couedic, J.-P., Robert, F., Alberdi, A. et al. (2009). Mutation in TET2 in Myeloid Cancers. *New England Journal of Medicine* **360**, 2289-2301.

Delker, R. K., Fugmann, S. D. and Papavasiliou, F. N. (2009). A coming-of-age story: activation-induced cytidine deaminase turns 10. *Nat Immunol* **10**, 1147-53.

Dent, A. L., Shaffer, A. L., Yu, X., Allman, D. and Staudt, L. M. (1997). Control of Inflammation, Cytokine Expression, and Germinal Center Formation by BCL-6. *Science* **276**, 589-592.

Devon, R. S., Porteous, D.J, Brookes, A.J. (1995). Splinkerettes-improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res* **23**, 1644-1645.

Dierlamm, J., Baens, M., Wlodarska, I., Stefanova-Ouzounova, M., Hernandez, J. M., Hossfeld, D. K., De Wolf-Peeters, C., Hagemeijer, A., Van den Berghe, H. and Marynen, P. (1999). The Apoptosis Inhibitor Gene API2 and a Novel 18q Gene,MLT, Are Recurrently Rearranged in the t(11;18)(q21;q21) Associated With Mucosa-Associated Lymphoid Tissue Lymphomas.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D. et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y. and Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-83.

Djerbi, M., Screpanti, V., Catrina, A. I., Bogen, B., Biberfeld, P. and Grandien, A. (1999). The Inhibitor of Death Receptor Signaling, Flice-Inhibitory Protein Defines a New Class of Tumor Progression Factors. *J Exp Med* **190**, 1025-1032.

Dores, G. M., Devesa, S. S., Curtis, R. E., Linet, M. S. and Morton, L. M. (2012). Acute leukemia incidence and patient survival among children and adults in the United States, 2001-2007. *Blood* **119**, 34-43.

Dufour, A., Schneider, F., Metzeler, K. H., Hoster, E., Schneider, S., Zellmeier, E., Benthaus, T., Sauerland, M.-C., Berdel, W. E., Büchner, T. et al. (2010). Acute Myeloid Leukemia With Biallelic CEBPA Gene Mutations and Normal Karyotype Represents a Distinct Genetic Entity Associated With a Favorable Clinical Outcome. *Journal of Clinical Oncology* **28**, 570-577.

Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G. and Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6.

Dupuy, A. J., Rogers, L. M., Kim, J., Nannapaneni, K., Starr, T. K., Liu, P., Largaespada, D. A., Scheetz, T. E., Jenkins, N. A. and Copeland, N. G. (2009). A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. *Cancer Res* **69**, 8150-6.

Erkeland, S. J., Valkhof, M., Heijmans-Antonissen, C., van Hoven-Beijen, A., Delwel, R., Hermans, M. H. and Touw, I. P. (2004). Large-scale identification of disease genes involved in acute myeloid leukemia. *J Virol* **78**, 1971-80.

Evan, G. I., Lewis, G. K., Ramsay, G. and Bishop, J. M. (1985). Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. *Mol Cell Biol* **5**, 3610-6.

Falini, B., Bolli, N., Shan, J., Martelli, M. P., Liso, A., Pucciarini, A., Bigerna, B., Pasqualucci, L., Mannucci, R., Rosati, R. et al. (2006). Both carboxy-terminus NES motif and mutated tryptophan(s) are crucial for aberrant nuclear export of nucleophosmin leukemic mutants in NPMc+ AML. *Blood* **107**, 4514-4523.

Falini, B., Martelli, M. P., Bolli, N., Sportoletti, P., Liso, A., Tiacci, E. and Haferlach, T. (2011). Acute myeloid leukemia with mutated nucleophosmin (NPM1): is it a distinct entity? *Blood* **117**, 1109-1120.

Falini, B., Mecucci, C., Tiacci, E., Alcalay, M., Rosati, R., Pasqualucci, L., La Starza, R., Diverio, D., Colombo, E., Santucci, A. et al. (2005). Cytoplasmic Nucleophosmin in Acute Myelogenous Leukemia with a Normal Karyotype. *New England Journal of Medicine* **352**, 254-266.

Ferch, U., Kloo, B., Gewies, A., Pfänder, V., Düwel, M., Peschel, C., Krappmann, D. and Ruland, J. (2009). Inhibition of MALT1 protease activity is selectively toxic for activated B cell–like diffuse large B cell lymphoma cells. *J Exp Med* **206**, 2313-2320.

Fernandez-Mercado, M., Yip, B. H., Pellagatti, A., Davies, C., Larrayoz, M. J., Kondo, T., Pérez, C., Killick, S., McDonald, E.-J., Odero, M. D. et al. (2012). Mutation Patterns of 16 Genes in Primary and Secondary Acute Myeloid Leukemia (AML) with Normal Cytogenetics. *Plos One* **7**, e42334.

Figueroa, M. E., Abdel-Wahab, O., Lu, C., Ward, P. S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H. F. et al. (2010). Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell* **18**, 553-567.

Ford, A. M., Ridge, S. A., Cabrera, M. E., Mahmoud, H., Steel, C. M., Chan, L. C. and Greaves, M. (1993). In utero rearrangements in the trithorax-related oncogene in infant leukaemias. *Nature* **363**, 358-360.

Forsberg, Lars A., Rasi, C., Razzaghian, Hamid R., Pakalapati, G., Waite, L., Thilbeault, Krista S., Ronowicz, A., Wineinger, Nathan E., Tiwari, Hemant K., Boomsma, D. et al. (2012). Age-Related Somatic Structural Changes in the Nuclear Genome of Human Blood Cells. *The American Journal of Human Genetics* **90**, 217-228.

Friedel, R. H., Friedel, C. C., Bonfert, T., Shi, R., Rad, R. and Soriano, P. (2013). Clonal Expansion Analysis of Transposon Insertions by High-Throughput Sequencing Identifies Candidate Cancer Genes in a PiggyBac Mutagenesis Screen. *Plos One* **8**, e72338.

Fujiwara, T., Fukuhara, N., Funayama, R., Nariai, N., Kamata, M., Nagashima, T., Kojima, K., Onishi, Y., Sasahara, Y., Ishizawa, K. et al. (2014). Identification of

acquired mutations by whole-genome sequencing in GATA-2 deficiency evolving into myelodysplasia and acute leukemia. *Ann Hematol*, 1-8.

Gaidzik, V. I., Paschka , P., Späth, D., Habdank, M., Köhne, C.-H., Germing, U., von Lilienfeld-Toal, M., Held, G., Horst, H.-A., Haase, D. et al. (2012). TET2 Mutations in Acute Myeloid Leukemia (AML): Results From a Comprehensive Genetic and Clinical Analysis of the AML Study Group. *Journal of Clinical Oncology* **30**, 1350-1357.

Gale, R. E., Green, C., Allen, C., Mead, A. J., Burnett, A. K., Hills, R. K. and Linch, D. C. (2008). The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood* **111**, 2776-2784.

Gamis, A. S., Alonzo, T. A., Gerbing, R. B., Hilden, J. M., Sorrell, A. D., Sharma, M., Loew, T. W., Arceci, R. J., Barnard, D., Doyle, J. et al. (2011). Natural history of transient myeloproliferative disorder clinically diagnosed in Down syndrome neonates: a report from the Children's Oncology Group Study A2971. *Blood* **118**, 6752-6759.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P. et al. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892.

Geurts, A. M., Yang, Y., Clark, K. J., Liu, G., Cui, Z., Dupuy, A. J., Bell, J. B., Largaespada, D. A. and Hackett, P. B. (2003). Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther* **8**, 108-17.

Gilliland, D. G. and Griffin, J. D. (2002). The roles of FLT3 in hematopoiesis and leukemia. *Blood* **100**, 1532-1542.

Goatly, A., Bacon, C. M., Nakamura, S., Ye, H., Kim, I., Brown, P. J., Ruskone-Fourmestraux, A., Cervera, P., Streubel, B., Banham, A. H. et al. (2008). FOXP1 abnormalities in lymphoma: translocation breakpoint mapping reveals insights into deregulated transcriptional control. *Mod Pathol* **21**, 902-911.

Gough, S. M., Slape, C. I. and Aplan, P. D. (2011). NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. *Blood* **118**, 6247-6257.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature* **481**, 306-313.

Green, C. L., Koo, K. K., Hills, R. K., Burnett, A. K., Linch, D. C. and Gale, R. E. (2010). Prognostic Significance of CEBPA Mutations in a Large Cohort of Younger Adult Patients With Acute Myeloid Leukemia: Impact of Double CEBPA Mutations and the Interaction With FLT3 and NPM1 Mutations. *Journal of Clinical Oncology* **28**, 2739-2747.

Grisendi, S., Mecucci, C., Falini, B. and Pandolfi, P. P. (2006). Nucleophosmin and cancer. *Nat Rev Cancer* **6**, 493-505.

Grove, C. S. and Vassiliou, G. S. (2014). Acute myeloid leukaemia: a paradigm for the clonal evolution of cancer? *Disease Models & Mechanisms* **7**, 941-951.

Gruber, Tanja A., Larson Gedman, A., Zhang, J., Koss, Cary S., Marada, S., Ta, Huy Q., Chen, S.-C., Su, X., Ogden, Stacey K., Dang, J. et al. (2012). An Inv(16)(p13.3q24.3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute Megakaryoblastic Leukemia. *Cancer Cell* **22**, 683-697.

Haferlach, C., Dicker, F., Kohlmann, A., Schindela, S., Weiss, T., Kern, W., Schnittger, S. and Haferlach, T. (2010). AML with CBFB-MYH11 rearrangement demonstrate RAS pathway alterations in 92% of all cases including a high frequency of NF1 deletions. *Leukemia* **24**, 1065-1069.

Haferlach, C., Grossmann, V., Kohlmann, A., Schindela, S., Kern, W., Schnittger, S. and Haferlach, T. (2012). Deletion of the tumor-suppressor gene NF1 occurs in 5% of myeloid malignancies and is accompanied by a mutation in the remaining allele in half of the cases. *Leukemia* **26**, 834-839.

Han, S., Dillon, S. R., Zheng, B., Shimoda, M., Schlissel, M. S. and Kelsoe, G. (1997). V(D)J Recombinase Activity in a Subset of Germinal Center B Lymphocytes. *Science* **278**, 301-305.

Hanahan, D. and Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell* **100**, 57-70.

Hardy, R. R. and Hayakawa, K. (2001). B cell development pathways. *Annu Rev Immunol* **19**, 595-621.

Harris, A. W., Pinkert, C. A., Crawford, M., Langdon, W. Y., Brinster, R. L. and Adams, J. M. (1988). The E mu-myc transgenic mouse. A model for high-incidence spontaneous lymphoma and leukemia of early B cells. *J Exp Med* **167**, 353-371.

Harris, J., Ibrahim, H., Amen, F., Karadimitris, A., Naresh, K. N. and Macdonald, D. H. (2012). Cellular (FLICE) like inhibitory protein (cFLIP) expression in diffuse large B-cell lymphoma identifies a poor prognostic subset, but fails to predict the molecular subtype. *Hematological Oncology* **30**, 8-12.

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. et al. (2005). A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828-833.

Horie, K., Yusa, K., Yae, K., Odajima, J., Fischer, S. E., Keng, V. W., Hayakawa, T., Mizuno, S., Kondoh, G., Ijiri, T. et al. (2003). Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Mol Cell Biol* **23**, 9189-207.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D. et al. (2012). Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. *Cell* **148**, 873-885.

Huff, C. A. and Matsui, W. (2008). Multiple myeloma cancer stem cells. *Journal of Clinical Oncology* **26**, 2895-2900.

Huntly, B. J. P., Shigematsu, H., Deguchi, K., Lee, B. H., Mizuno, S., Duclos, N., Rowan, R., Amaral, S., Curley, D., Williams, I. R. et al. (2004). MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. *Cancer Cell* **6**, 587-596.

Ichikawa, S., Fukuhara, N., Katsushima, H., Takahashi, T., Yamamoto, J., Yokoyama, H., Sasaki, O., Fukuhara, O., Nomura, J., Ishizawa, K. et al. (2014). Association between BACH2 expression and clinical prognosis in diffuse large B-cell lymphoma. *Cancer Science* **105**, 437-444.

Iqbal, J., Greiner, T. C., Patel, K., Dave, B. J., Smith, L., Ji, J., Wright, G., Sanger, W. G., Pickering, D. L., Jain, S. et al. (2007). Distinctive patterns of BCL6 molecular alterations and their functional consequences in different subgroups of diffuse large B-cell lymphoma. *Leukemia* **21**, 2332-2343.

Itzykson, R., Kosmider, O., Renneville, A., Gelsi-Boyer, V., Meggendorfer, M., Morabito, M., Berthon, C., Adès, L., Fenaux, P., Beyne-Rauzy, O. et al. (2013a). Prognostic Score Including Gene Mutations in Chronic Myelomonocytic Leukemia. *Journal of Clinical Oncology* **31**, 2428-2436.

Itzykson, R., Kosmider, O., Renneville, A., Morabito, M., Preudhomme, C., Berthon, C., Adès, L., Fenaux, P., Platzbecker, U., Gagey, O. et al. (2013b). Clonal architecture of chronic myelomonocytic leukemias. *Blood* **121**, 2186-2198.

Itzykson, R. and Solary, E. (2013). An evolutionary perspective on chronic myelomonocytic leukemia. *Leukemia*.

Ivics, Z., Hackett, P. B., Plasterk, R. H. and Izsvák, Z. (1997). Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* **91**, 501-510.

Ivics, Z., Li, M. A., Mates, L., Boeke, J. D., Nagy, A., Bradley, A. and Izsvak, Z. (2009). Transposon-mediated genome manipulation in vertebrates. *Nat Meth* **6**, 415-422.

Izsvák, Z., Ivics, Z. and Plasterk, R. H. (2000). Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *Journal of Molecular Biology* **302**, 93-102.

Izsvák, Z., Khare, D., Behlke, J., Heinemann, U., Plasterk, R. H. and Ivics, Z. (2002). Involvement of a Bifunctional, Paired-like DNA-binding Domain and a Transpositional Enhancer in Sleeping BeautyTransposition. *Journal of Biological Chemistry* **277**, 34581-34588.

Jacobs, K. B. Yeager, M. Zhou, W. Wacholder, S. Wang, Z. Rodriguez-Santiago, B. Hutchinson, A. Deng, X. Liu, C. Horner, M.-J. et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet* **44**, 651-658.

Jacobson, J. W., Medhora, M. M. and Hartl, D. L. (1986). Molecular structure of a somatically unstable transposable element in Drosophila. *Proceedings of the National Academy of Sciences* **83**, 8684-8688.

Jan, M., Snyder, T. M., Corces-Zimmerman, M. R., Vyas, P., Weissman, I. L., Quake, S. R. and Majeti, R. (2012). Clonal Evolution of Preleukemic Hematopoietic Stem Cells Precedes Human Acute Myeloid Leukemia. *Science Translational Medicine* **4**, 149ra118.

Jankowska, A. M., Makishima, H., Tiu, R. V., Szpurka, H., Huang, Y., Traina, F., Visconte, V., Sugimoto, Y., Prince, C., O'Keefe, C. et al. (2011). Mutational spectrum analysis of chronic myelomonocytic leukemia includes genes associated with epigenetic regulation: UTX, EZH2, and DNMT3A. *Blood* **118**, 3932-3941.

Janz, S. M. I., H.C. Teitell, M.A. (2008). Mouse Models of Human Mature B Cell and Plasma Cell Neoplasms. *Mouse Models of Human Blood Cancers*.

Javadi, M., Richmond, T. D., Huang, K. and Barber, D. L. (2013). CBL Linker Region and RING Finger Mutations Lead to Enhanced Granulocyte-Macrophage Colony-stimulating Factor (GM-CSF) Signaling via Elevated Levels of JAK2 and LYN. *Journal of Biological Chemistry* **288**, 19459-19470.

Jemal, A., Siegel, R., Xu, J. and Ward, E. Cancer statistics, 2010. *CA Cancer J Clin* **60**, 277-300.

Johnson, G. R., Gonda, T. J., Metcalf, D., Kariharan, I. K. and Cory, S. (1989). A lethal myeloproliferative syndrome in mice transplanted with bone marrow cells infected with a retrovirus expressing granulocyte-macrophage colony stimulating factor. *EMBO Journal* **8**, 441-448.

Kanungo, A., Medeiros, L. J., Abruzzo, L. V. and Lin, P. (2005). Lymphoid neoplasms associated with concurrent t(14;18) and 8q24//c-MYC translocation generally have a poor prognosis. *Mod Pathol* **19**, 25-33.

Kelly, L. M., Kutok, J. L., Williams, I. R., Boulton, C. L., Amaral, S. M., Curley, D. P., Ley, T. J. and Gilliland, D. G. (2002). PML/RARα and FLT3-ITD induce an APL-like disease in a mouse model. *Proceedings of the National Academy of Sciences* **99**, 8283-8288.

Kent, O. A., Chivukula, R. R., Mullendore, M., Wentzel, E. A., Feldmann, G., Lee, K. H., Liu, S., Leach, S. D., Maitra, A. and Mendell, J. T. (2010). Repression of the miR-143/145 cluster by oncogenic Ras initiates a tumor-promoting feed-forward pathway. *Genes Dev* **24**, 2754-9.

Kent, O. A., Fox-Talbot, K. and Halushka, M. K. (2013). RREB1 repressed miR-143/145 modulates KRAS signaling through downregulation of multiple targets. *Oncogene* **32**, 2576-85.

Kihara, R., Nagata, Y., Kiyoi, H., Kato, T., Yamamoto, E., Suzuki, K., Chen, F., Asou, N., Ohtake, S., Miyawaki, S. et al. (2014). Comprehensive analysis of genetic alterations and their prognostic impacts in adult acute myeloid leukemia patients. *Leukemia*.

Klijn, C., Koudijs, M. J., Kool, J., ten Hoeve, J., Boer, M., de Moes, J., Akhtar, W., van Miltenburg, M., Vendel-Zwaagstra, A., Reinders, M. J. et al. (2013). Analysis of tumor heterogeneity and cancer gene networks using deep sequencing of MMTV-induced mouse mammary tumors. *Plos One* **8**.

Ko, M., Huang, Y., Jankowska, A. M., Pape, U. J., Tahiliani, M., Bandukwala, H. S., An, J., Lamperti, E. D., Koh, K. P., Ganetzky, R. et al. (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-843.

Kon, A., Shih, L.-Y., Minamino, M., Sanada, M., Shiraishi, Y., Nagata, Y., Yoshida, K., Okuno, Y., Bando, M., Nakato, R. et al. (2013). Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat Genet* **45**, 1232-1237.

Kool, J., Uren, A. G., Martins, C. P., Sie, D., de Ridder, J., Turner, G., van Uitert, M., Matentzoglu, K., Lagcher, W., Krimpenfort, P. et al. Insertional mutagenesis in

mice deficient for p15Ink4b, p16Ink4a, p21Cip1, and p27Kip1 reveals cancer gene interactions and correlations with tumor phenotypes. *Cancer Res* **70**, 520-31.

Koudijs, M. J., Klijn, C., van der Weyden, L., Kool, J., ten Hoeve, J., Sie, D., Prasetyanti, P. R., Schut, E., Kas, S., Whipp, T. et al. (2011). High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. *Genome Research* **21**, 2181-2189.

Kovalchuk, A. L., Qi, C. F., Torrey, T. A., Taddesse-Heath, L., Feigenbaum, L., Park, S. S., Gerbitz, A., Klobeck, G., Hoertnagel, K., Polack, A. et al. (2000). Burkitt lymphoma in the mouse. *J Exp Med* **192**, 1183-90.

Krönke, J., Bullinger, L., Teleanu, V., Tschürtz, F., Gaidzik, V. I., Kühn, M. W. M., Rücker, F. G., Holzmann, K., Paschka, P., Kapp-Schwörer, S. et al. (2013). Clonal evolution in relapsed NPM1 mutated acute myeloid leukemia. *Blood*.

Kuhn, R., Schwenk, F., Aguet, M. and Rajewsky, K. (1995). Inducible gene targeting in mice. *Science* **269**, 1427-9.

Küppers, R., Klein, U., Hansmann, M.-L. and Rajewsky, K. (1999). Cellular Origin of Human B-Cell Lymphomas. *New England Journal of Medicine* **341**, 1520-1529.

Kusec, R., Laczika, K., Knöbl, P., Friedl, J., Greinix, H., Kahls, P., Linkesch, W., Schwarzinger, I., Mitterbauer, G. and Purtscher, B. (1994). AML1/ETO fusion mRNA can be detected in remission blood samples of all patients with t(8;21) acute myeloid leukemia after chemotherapy or autologous bone marrow transplantation. *Leukemia* **8**, 735-739.

Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Larson, D. R., Plevak, M. F., Offord, J. R., Dispenzieri, A., Katzmann, J. A. and Melton, L. J., 3rd. (2006). Prevalence of monoclonal gammopathy of undetermined significance. *N Engl J Med* **354**, 1362-9.

Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Offord, J. R., Larson, D. R., Plevak, M. F. and Melton, L. J., 3rd. (2002). A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *N Engl J Med* **346**, 564-9.

Lang, R. A., Metcalf, D., Cuthbertson, R. A., Lyons, I., Stanley, E., Kelso, A., Kannourakis, G., Williamson, D. J., Klintworth, G. K., Gonda, T. J. et al. (1987). Transgenic mice expressing a hemopoietic growth factor gene (GM-CSF) develop accumulations of macrophages, blindness, and a fatal syndrome of tissue damage. *Cell* **51**, 675-686.

Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G. et al. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research* **19**, 2308-2316.

Lapidot, T., Sirard, C., Vormoor, J., Murdoch, B., Hoang, T., Caceres-Cortes, J., Minden, M., Paterson, B., Caligiuri, M. A. and Dick, J. E. (1994). A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645-648.

Larsson, L. G. and Henriksson, M. A. The Yin and Yang functions of the Myc oncoprotein in cancer development and as targets for therapy. *Exp Cell Res* **316**, 1429-37.

Lauchle, J. O., Kim, D., Le, D. T., Akagi, K., Crone, M., Krisman, K., Warner, K., Bonifas, J. M., Li, Q., Coakley, K. M. et al. (2009). Response and resistance to MEK inhibition in leukaemias initiated by hyperactive Ras. *Nature* **461**, 411-4.

Laurie, C. C., Laurie, C. A., Rice, K., Doheny, K. F., Zelnick, L. R., McHugh, C. P., Ling, H., Hetrick, K. N., Pugh, E. W., Amos, C. et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* **44**, 642-650.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A. et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218.

Le, D. T., Kong, N., Zhu, Y., Lauchle, J. O., Aiyigari, A., Braun, B. S., Wang, E., Kogan, S. C., Le Beau, M. M., Parada, L. et al. (2004). Somatic inactivation of Nf1 in hematopoietic cells results in a progressive myeloproliferative disorder. *Blood* **103**, 4243-4250.

Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J. et al. (2009). IMGT, the international ImMunoGeneTics information system. *Nucleic acids research* **37**, D1006-12.

Ley, T. J., Ding, L., Walter, M. J., McLellan, M. D., Lamprecht, T., Larson, D. E., Kandoth, C., Payton, J. E., Baty, J., Welch, J. et al. (2010). DNMT3A Mutations in Acute Myeloid Leukemia. *New England Journal of Medicine* **363**, 2424-2433.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M. et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95.

Li, J., Shen, H., Himmel, K. L., Dupuy, A. J., Largaespada, D. A., Nakamura, T., Shaughnessy, J. D., Jr., Jenkins, N. A. and Copeland, N. G. (1999). Leukaemia disease genes: large-scale cloning and pathway predictions. *Nat Genet* **23**, 348-53.

Li, J., Spensberger, D., Ahn, J. S., Anand, S., Beer, P. A., Ghevaert, C., Chen, E., Forrai, A., Scott, L. M., Ferreira, R. et al. (2010). JAK2 V617F impairs hematopoietic stem cell function in a conditional knock-in mouse model of JAK2 V617F–positive essential thrombocythemia. *Blood* **116**, 1528-1538.

Li, L. and Clevers, H. (2010). Coexistence of Quiescent and Active Adult Stem Cells in Mammals. *Science* **327**, 542-545.

Li, M. A., Pettitt, S. J., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J., Yusa, K., Conte, N. and Bradley, A. (2013). The piggyBac Transposon Displays Local and Distant Reintegration Preferences and Can Cause Mutations at Noncanonical Integration Sites. *Mol Cell Biol* **33**, 1317-1330.

Li, Q., Haigis, K. M., McDaniel, A., Harding-Theobald, E., Kogan, S. C., Akagi, K., Wong, J. C. Y., Braun, B. S., Wolff, L., Jacks, T. et al. (2011). Hematopoiesis

and leukemogenesis in mice expressing oncogenic NrasG12D from the endogenous locus. *Blood* **117**, 2022-2032.
Liang, Q., Kong, J., Stalker, J. and Bradley, A. (2009). Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *Genesis* **47**, 404-8.
Liu, G., Aronovich, E. L., Cui, Z., Whitley, C. B. and Hackett, P. B. (2004). Excision of Sleeping Beauty transposons: parameters and applications to gene therapy. *The Journal of Gene Medicine* **6**, 574-583.
Liu, P., Leong, T., Quam, L., Billadeau, D., Kay, N. E., Greipp, P., Kyle, R. A., Oken, M. M. and Van Ness, B. (1996). Activating mutations of N- and K-ras in multiple myeloma show different clinical associations: analysis of the Eastern Cooperative Oncology Group Phase III Trial. *Blood* **88**, 2699-706.
Lo Coco, F., Ye, B. H., Lista, F., Corradini, P., Offit, K., Knowles, D. M., Chaganti, R. S. and Dalla-Favera, R. (1994). Rearrangements of the BCL6 gene in diffuse large cell non-Hodgkin's lymphoma. *Blood* **83**, 1757-9.
Loh, M. L., Vattikuti, S., Schubbert, S., Reynolds, M. G., Carlson, E., Lieuw, K. H., Cheng, J. W., Lee, C. M., Stokoe, D., Bonifas, J. M. et al. (2004). Mutations in PTPN11 implicate the SHP-2 phosphatase in leukemogenesis. *Blood* **103**, 2325-2331.
Lohr, J. G., Stojanov, P., Lawrence, M. S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y. W., Slager, S. L. et al. (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences* **109**, 3879-3884.
Lulli, V., Romania, P., Riccioni, R., Boe, A., Lo-Coco, F., Testa, U. and Marziali, G. (2010). Transcriptional silencing of the ETS1 oncogene contributes to human granulocytic differentiation. *Haematologica* **95**, 1633-1641.
Luo, G., Ivics, Z., Izsvák, Z. and Bradley, A. (1998). Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences* **95**, 10769-10773.
MacLennan, I. C., Toellner, K. M., Cunningham, A. F., Serre, K., Sze, D. M., Zuniga, E., Cook, M. C. and Vinuesa, C. G. (2003). Extrafollicular antibody responses. *Immunol Rev* **194**, 8-18.
Malinge, S., Chlon, T., Doré, L. C., Ketterling, R. P., Tallman, M. S., Paietta, E., Gamis, A. S., Taub, J. W., Chou, S. T., Weiss, M. J. et al. (2013). Development of acute megakaryoblastic leukemia in Down syndrome is associated with sequential epigenetic changes. *Blood* **122**, e33-e43.
Malinge, S., Izraeli, S. and Crispino, J. D. (2009). Insights into the manifestations, outcomes, and mechanisms of leukemogenesis in Down syndrome. *Blood* **113**, 2619-2628.
Mann, K. M., Jenkins, N. A., Copeland, N. G. and Mann, M. B. (2014). Transposon Insertional Mutagenesis Models of Cancer. *Cold Spring Harbor Protocols* **2014**, pdb.top069849.
Mann, K. M., Ward, J. M., Yew, C. C. K., Kovochich, A., Dawson, D. W., Black, M. A., Brett, B. T., Sheetz, T. E., Dupuy, A. J., Initiative, A. P. C. G. et al. (2012).

Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. *Proceedings of the National Academy of Sciences* **109**, 5934-5941.

March, H. N., Rust, A. G., Wright, N. A., ten Hoeve, J., de Ridder, J., Eldridge, M., van der Weyden, L., Berns, A., Gadiot, J., Uren, A. et al. (2011). Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nat Genet* **43**, 1202-1209.

Mates, L., Chuah, M. K. L., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D. P., Schmitt, A., Becker, K., Matrai, J. et al. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet* **41**, 753-761.

Maul, R. W. and Gearhart, P. J. AID and somatic hypermutation. *Adv Immunol* **105**, 159-91.

McAllister-Lucas, L. M., Baens, M. and Lucas, P. C. (2011). MALT1 Protease: A New Therapeutic Target in B Lymphoma and Beyond? *Clinical Cancer Research* **17**, 6623-6631.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36**, 344-355.

McKenzie, J. L., Gan, O. I., Doedens, M., Wang, J. C. Y. and Dick, J. E. (2006). Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. *Nat Immunol* **7**, 1225-1233.

Meggendorfer, M., Roller, A., Haferlach, T., Eder, C., Dicker, F., Grossmann, V., Kohlmann, A., Alpermann, T., Yoshida, K., Ogawa, S. et al. (2012). SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). *Blood* **120**, 3080-3088.

Metcalf, D. (2013). The Colony-Stimulating Factors and Cancer. *Cancer Immunology Research* **1**, 351-356.

Metcalf, D., Glaser, S. P., Xu, Z., Di Rago, L. and Mifsud, S. (2013). Reversible growth factor dependency and autonomy during murine myelomonocytic leukemia induced by oncogenes. *Proceedings of the National Academy of Sciences* **110**, 17029-17034.

Metzeler, K. H., Maharry, K., Radmacher, M. D., Mrózek, K., Margeson, D., Becker, H., Curfman, J., Holland, K. B., Schwind, S., Whitman, S. P. et al. (2011). TET2 Mutations Improve the New European LeukemiaNet Risk Classification of Acute Myeloid Leukemia: A Cancer and Leukemia Group B Study. *Journal of Clinical Oncology* **29**, 1373-1381.

Meyer, N. and Penn, L. Z. (2008). Reflecting on 25 years with MYC. *Nat Rev Cancer* **8**, 976-90.

Milon, B. C., Agyapong, A., Bautista, R., Costello, L. C. and Franklin, R. B. (2010). Ras responsive element binding protein-1 (RREB-1) down-regulates hZIP1 expression in prostate cancer cells. *Prostate* **70**, 288-96.

Miyamoto, T., Nagafuji, K., Akashi, K., Harada, M., Kyo, T., Akashi, T., Takenaka, K., Mizuno, S., Gondo, H., Okamura, T. et al. (1996). Persistence of

multipotent progenitors expressing AML1/ETO transcripts in long-term remission patients with t(8;21) acute myelogenous leukemia. *Blood* **87**, 4789-4796.
Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, Maria E., Vasanthakumar, A., Patel, J., Zhao, X. et al. (2011). Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. *Cancer Cell* **20**, 11-24.
Mori, H., Colman, S. M., Xiao, Z., Ford, A. M., Healy, L. E., Donaldson, C., Hows, J. M., Navarrete, C. and Greaves, M. (2002). Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proceedings of the National Academy of Sciences* **99**, 8242-8247.
Morin, R. D., Mungall, K., Pleasance, E., Mungall, A. J., Goya, R., Huff, R. D., Scott, D. W., Ding, J., Roth, A., Chiu, R. et al. (2013). Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing.
Morse, H. C., Anver, M. R., Fredrickson, T. N., Haines, D. C., Harris, A. W., Harris, N. L., Jaffe, E. S., Kogan, S. C., MacLennan, I. C. M., Pattengale, P. K. et al. (2002). Bethesda proposals for classification of lymphoid neoplasms in mice. *Blood* **100**, 246-258.
Mullighan, C. G., Phillips, L. A., Su, X., Ma, J., Miller, C. B., Shurtleff, S. A. and Downing, J. R. (2008). Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. *Science* **322**, 1377-1380.
Mullins, C. D., Su, M. Y., Hucthagowder, V., Chu, L., Lu, L., Kulkarni, S., Novack, D., Vij, R. and Tomasson, M. H. (2013). Germinal Center B-Cells Resist Transformation by <italic>Kras</italic> Independently of Tumor Suppressor <italic>Arf</italic>. *Plos One* **8**, e67941.
Mupo, A., Celani, L., Dovey, O., Cooper, J. L., Grove, C., Rad, R., Sportoletti, P., Falini, B., Bradley, A. and Vassiliou, G. S. (2013). A powerful molecular synergy between mutant Nucleophosmin and Flt3-ITD drives acute myeloid leukemia in mice. *Leukemia* **27**, 1917-1920.
Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D. et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94.
Nedergaard, T., Guldberg, P., Ralfkiaer, E. and Zeuthen, J. (1997). A one-step DGGE scanning method for detection of mutations in the K-, N-, and H-ras oncogenes: mutations at codons 12, 13 and 61 are rare in B-cell non-Hodgkin's lymphoma. *Int J Cancer* **71**, 364-9.
Ngo, V. N., Davis, R. E., Lamy, L., Yu, X., Zhao, H., Lenz, G., Lam, L. T., Dave, S., Yang, L., Powell, J. et al. (2006). A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441**, 106-110.
Nik-Zainal, S., Van Loo, P., Wedge, David C., Alexandrov, Ludmil B., Greenman, Christopher D., Lau, King W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M. et al. (2012). The Life History of 21 Breast Cancers. *Cell* **149**, 994-1007.
Nikolaev, S. I., Santoni, F., Vannier, A., Falconnet, E., Giarin, E., Basso, G., Hoischen, A., Veltman, J. A., Groet, J., Nizetic, D. et al. (2013). Exome

sequencing identifies putative drivers of progression of transient myeloproliferative disorder to AMKL in infants with Down syndrome. *Blood* **122**, 554-561.
Niu, H., Ye, B. H. and Dalla-Favera, R. (1998). Antigen receptor signaling induces MAP kinase-mediated phosphorylation and degradation of the BCL-6 transcription factor. *Genes Dev* **12**, 1953-1961.
Notta, F., Mullighan, C. G., Wang, J. C. Y., Poeppl, A., Doulatov, S., Phillips, L. A., Ma, J., Minden, M. D., Downing, J. R. and Dick, J. E. (2011). Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. *Nature* **469**, 362-367.
Nowell, P. (1976). The clonal evolution of tumor cell populations. *Science* **194**, 23-28.
Offit, K., Lo Coco, F., Louie, D. C., Parsa, N. Z., Leung, D., Portlock, C., Ye, B. H., Lista, F., Filippa, D. A., Rosenbaum, A. et al. (1994). Rearrangement of the bcl-6 gene as a prognostic marker in diffuse large-cell lymphoma. *N Engl J Med* **331**, 74-80.
Ohno, S., Babonits, M., Wiener, F., Spira, J., Klein, G. and Potter, M. (1979). Nonrandom chromosome changes involving the Ig gene-carrying chromosomes 12 and 6 in pristane-induced mouse plasmacytomas. *Cell* **18**, 1001-1007.
Palomo, C., Zou, X., Nicholson, I. C., Butzler, C. and Bruggemann, M. (1999). B-cell tumorigenesis in mice carrying a yeast artificial chromosome-based immunoglobulin heavy/c-myc translocus is independent of the heavy chain intron enhancer (Emu). *Cancer Res* **59**, 5625-8.
Palumbo, A. and Rajkumar, S. V. (2009). Treatment of newly diagnosed myeloma. *Leukemia* **23**, 449-56.
Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J. S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C. et al. (2011). Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts. *New England Journal of Medicine* **365**, 1384-1395.
Papaemmanuil, E., Rapado, I., Li, Y., Potter, N. E., Wedge, D. C., Tubio, J., Alexandrov, L. B., Van Loo, P., Cooke, S. L., Marshall, J. et al. (2014). RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet* **46**, 116-125.
Papavasiliou, F. N. and Schatz, D. G. (2000). Cell-cycle-regulated DNA double-stranded breaks in somatic hypermutation of immunoglobulin genes. *Nature* **408**, 216-21.
Park, S. S., Kim, J. S., Tessarollo, L., Owens, J. D., Peng, L., Han, S. S., Tae Chung, S., Torrey, T. A., Cheung, W. C., Polakiewicz, R. D. et al. (2005). Insertion of c-Myc into Igh induces B-cell and plasma-cell neoplasms in mice. *Cancer Res* **65**, 1306-15.
Parkin, B., Ouillette, P., Li, Y., Keller, J., Lam, C., Roulston, D., Li, C., Shedden, K. and Malek, S. N. (2013). Clonal evolution and devolution after chemotherapy in adult acute myelogenous leukemia. *Blood* **121**, 369-377.
Parkin, B., Ouillette, P., Wang, Y., Liu, Y., Wright, W., Roulston, D., Purkayastha, A., Dressel, A., Karp, J., Bockenstedt, P. et al. (2010). NF1

Inactivation in Adult Acute Myelogenous Leukemia. *Clinical Cancer Research* **16**, 4135-4147.

Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V. A., Grunn, A., Messina, M., Elliot, O. et al. (2011). Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat Genet* **43**, 830-837.

Peabody, D. S. (1989). Translation initiation at non-AUG triplets in mammalian cells. *J Biol Chem* **264**, 5031-5.

Pine, S. R., Guo, Q., Yin, C., Jayabose, S., Druschel, C. M. and Sandoval, C. (2007). Incidence and clinical implications of GATA1 mutations in newborns with Down syndrome. *Blood* **110**, 2128-2131.

Poppe, B., Vandesompele, J., Schoch, C., Lindvall, C., Mrózek, K., Bloomfield, C. D., Beverloo, H. B., Michaux, L., Dastugue, N., Herens, C. et al. (2004). Expression analyses identify MLL as a prominent target of 11q23 amplification and support an etiologic role for MLL gain of function in myeloid malignancies. *Blood* **103**, 229-235.

Prchal, J. T., Prchal, J. F., Belickova, M., Chen, S., Guan, Y., Gartland, G. L. and Cooper, M. D. (1996). Clonal stability of blood cell lineages indicated by X-chromosomal transcriptional polymorphism. *J Exp Med* **183**, 561-567.

Quivoron, C., Couronné, L., Della Valle, V., Lopez, Cécile K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.-H. et al. (2011). TET2 Inactivation Results in Pleiotropic Hematopoietic Abnormalities in Mouse and Is a Recurrent Event during Human Lymphomagenesis. *Cancer Cell* **20**, 25-38.

Rad, R., Rad, L., Wang, W., Cadinanos, J., Vassiliou, G., Rice, S., Campos, L. S., Yusa, K., Banerjee, R., Li, M. A. et al. (2010). PiggyBac Transposon Mutagenesis: A Tool for Cancer Gene Discovery in Mice. *Science* **330**, 1104-1107.

Rad, R., Rad, L, Wang, W, Cadinanos, J et al. (2010). *PiggyBac* Transposon Mutagenesis: A Tool for Cancer Gene Discovery in Mice. *Science* **In press**.

Raponi, M., Lancet, J. E., Fan, H., Dossey, L., Lee, G., Gojo, I., Feldman, E. J., Gotlib, J., Morris, L. E., Greenberg, P. L. et al. (2008). A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia.

Raval, A., Kusler, B., Pang, W. W., Weissman, I. L., Mitchell, B. S. and Park, C. Y. Effect of nucleophosmin1 haploinsufficiency on hematopoietic stem cells: Leukemia. 2012 Apr;26(4):853-5. doi: 10.1038/leu.2011.270. Epub 2011 Oct 7.

Refaeli, Y., Young, R. M., Turner, B. C., Duda, J., Field, K. A. and Bishop, J. M. (2008). The B cell antigen receptor and overexpression of MYC can cooperate in the genesis of B cell lymphomas. *PLoS Biol* **6**, e152.

Rhoades, K. L., Hetherington, C. J., Harakawa, N., Yergeau, D. A., Zhou, L., Liu, L.-Q., Little, M.-T., Tenen, D. G. and Zhang, D.-E. (2000). Analysis of the role of AML1-ETO in leukemogenesis, using an inducible transgenic mouse model. *Blood* **96**, 2108-2115.

Riccioni, R., Diverio, D., Riti, V., Buffolino, S., Mariani, G., Boe, A., Cedrone, M., Ottone, T., Foà, R. and Testa, U. (2009). Interleukin (IL)-3/granulocyte

macrophage-colony stimulating factor/IL-5 receptor alpha and beta chains are preferentially expressed in acute myeloid leukaemias with mutated FMS-related tyrosine kinase 3 receptor. *British Journal of Haematology* **144**, 376-387.

Rogozin, I. B. and Diaz, M. (2004). Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol* **172**, 3382-4.

Rovigatti, U., Watson, D. K. and Yunis, J. J. (1986). Amplification and rearrangement of Hu-ets-1 in leukemia and lymphoma with involvement of 11q23. *Science* **232**, 398-400.

Saito, M., Gao, J., Basso, K., Kitagawa, Y., Smith, P. M., Bhagat, G., Pernis, A., Pasqualucci, L. and Dalla-Favera, R. (2007). A Signaling Pathway Mediating Downregulation of BCL6 in Germinal Center B Cells Is Blocked by BCL6 Gene Alterations in B Cell Lymphoma. *Cancer Cell* **12**, 280-292.

Sakane-Ishikawa, E., Nakatsuka, S., Tomita, Y., Fujita, S., Nakamichi, I., Takakuwa, T., Sugiyama, H., Fukuhara, S., Hino, M., Kanamaru, A. et al. (2005). Prognostic significance of BACH2 expression in diffuse large B-cell lymphoma: a study of the Osaka Lymphoma Study Group. *J Clin Oncol* **23**, 8012-7.

Sarver, A., Erdman, J., Starr, T., Largaespada, D. and Silverstein, K. A. (2012). TAPDANCE: An automated tool to identify and annotate transposon insertion CISs and associations between CISs from next generation sequence data. *BMC Bioinformatics* **13**, 154.

Sato, T., Onai, N., Yoshihara, H., Arai, F., Suda, T. and Ohteki, T. (2009). Interferon regulatory factor-2 protects quiescent hematopoietic stem cells from type I interferon-dependent exhaustion. *Nat Med* **15**, 696-700.

Schick, U. M., McDavid, A., Crane, P. K., Weston, N., Ehrlich, K., Newton, K. M., Wallace, R., Bookman, E., Harrison, T., Aragaki, A. et al. (2013). Confirmation of the Reported Association of Clonal Chromosomal Mosaicism with an Increased Risk of Incident Hematologic Cancer. *Plos One* **8**, e59823.

Schnittger, S., Bacher, U., Haferlach, C., Alpermann, T., Dicker, F., Sundermann, J., Kern, W. and Haferlach, T. (2011). Characterization of NPM1-mutated AML with a history of myelodysplastic syndromes or myeloproliferative neoplasms. *Leukemia* **25**, 615-621.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* **6**, 461-464.

Shah, A., Andersson, T. M. L., Rachet, B., Björkholm, M. and Lambert, P. C. (2013). Survival and cure of acute myeloid leukaemia in England, 1971-2006: a population–based study. *British Journal of Haematology* **162**, 509-516.

Shen-Ong, G. L. C., Keath, E. J., Piccoli, S. P. and Cole, M. D. (1982). Novel myc oncogene RNA from abortive immunoglobulin-gene recombination in mouse plasmacytomas. *Cell* **31**, 443-452.

Shi, J. and Vakoc, Christopher R. (2014). The Mechanisms behind the Therapeutic Activity of BET Bromodomain Inhibition. *Mol Cell* **54**, 728-736.

Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., Kennedy, J. A., Schimmer, A. D., Schuh, A. C., Yee, K. W. et al. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **advance online publication**.

Sportoletti, P., Varasano, E., Rossi, R., Bereshchenko, O., Cecchini, D., Gionfriddo, I., Bolli, N., Tiacci, E., Intermesoli, T., Zanghì, P. et al. (2013). The human NPM1 mutation A perturbs megakaryopoiesis in a conditional mouse model. *Blood* **121**, 3447-3458.

Starck, S. R., Ow, Y., Jiang, V., Tokuyama, M., Rivera, M., Qi, X., Roberts, R. W. and Shastri, N. (2008). A Distinct Translation Initiation Mechanism Generates Cryptic Peptides for Immune Surveillance. *Plos One* **3**, e3460.

Starr, T., Allaei, R., Silverstein, K., Staggs, R., Sarver, A., Bergemann, T., Gupta, M., O'Sullivan, M., Matise, I., Dupuy, A. et al. (2009). A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323**, 1747 - 1750.

Stephen, Andrew G., Esposito, D., Bagni, Rachel K. and McCormick, F. (2014). Dragging Ras Back in the Ring. *Cancer Cell* **25**, 272-281.

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

Sureban, S. M., May, R., Qu, D., Weygant, N., Chandrakesan, P., Ali, N., Lightfoot, S. A., Pantazis, P., Rao, C. V., Postier, R. G. et al. (2013). DCLK1 regulates pluripotency and angiogenic factors via microRNA-dependent mechanisms in pancreatic cancer. *Plos One* **8**.

Suzuki, T., Kiyoi, H., Ozeki, K., Tomita, A., Yamaji, S., Suzuki, R., Kodera, Y., Miyawaki, S., Asou, N., Kuriyama, K. et al. (2005). Clinical characteristics and prognostic implications of NPM1 mutations in acute myeloid leukemia. *Blood* **106**, 2854-61.

Swerdlow, S. H., Campo, E., Harris, N.L., Jaffe, E.S., Pileri, S. A., Stein, H., Thiele, J., Vardiman, J.W. (2008). WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues: WHO.

Swierczek, S. I., Agarwal, N., Nussenzveig, R. H., Rothstein, G., Wilson, A., Artz, A. and Prchal, J. T. (2008). Hematopoiesis is not clonal in healthy elderly women. *Blood* **112**, 3186-3193.

Szymanska, H., Lechowska-Piskorowska, J., Krysiak, E., Strzalkowska, A., Unrug-Bielawska, K., Grygalewicz, B., Skurzak, H., Pienkowska-Grela, B. and Gajewska, M. (2013). Neoplastic and Nonneoplastic Lesions in Aging Mice of Unique and Common Inbred Strains Contribution to Modeling of Human Neoplastic Diseases. *Vet Pathol* **16**, 16.

Szymczak, A. L., Workman, C. J., Wang, Y., Vignali, K. M., Dilioglou, S., Vanin, E. F. and Vignali, D. A. (2004). Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat Biotechnol* **22**, 589-94.

Tang, J. Z., Carmichael, C. L., Shi, W., Metcalf, D., Ng, A. P., Hyland, C. D., Jenkins, N. A., Copeland, N. G., Howell, V. M., Zhao, Z. J. et al. (2013). Transposon mutagenesis reveals cooperation of ETS family transcription factors with

signaling pathways in erythro-megakaryocytic leukemia. *Proceedings of the National Academy of Sciences* **110**, 6091-6096.

Tartaglia, M. and Gelb, B. D. (2005). Germ-line and somatic PTPN11 mutations in human disease. *European Journal of Medical Genetics* **48**, 81-96.

Tartaglia, M., Niemeyer, C. M., Fragale, A., Song, X., Buechner, J., Jung, A., Hahlen, K., Hasle, H., Licht, J. D. and Gelb, B. D. (2003). Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nat Genet* **34**, 148-150.

Taskesen, E., Bullinger, L., Corbacioglu, A., Sanders, M. A., Erpelinck, C. A. J., Wouters, B. J., van der Poel-van de Luytgaarde, S. C., Damm, F., Krauter, J., Ganser, A. et al. (2011). Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* **117**, 2469-2475.

TCGA_Research_Network. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine* **368**, 2059-2074.

Thiagalingam, A., De Bustros, A., Borges, M., Jasti, R., Compton, D., Diamond, L., Mabry, M., Ball, D. W., Baylin, S. B. and Nelkin, B. D. (1996). RREB-1, a novel zinc finger protein, is involved in the differentiation response to Ras in human medullary thyroid carcinomas. *Mol Cell Biol* **16**, 5335-45.

Thol, F., Bollin, R., Gehlhaar, M., Walter, C., Dugas, M., Suchanek, K. J., Kirchner, A., Huang, L., Chaturvedi, A., Wichmann, M. et al. (2013). Mutations in the cohesin complex in acute myeloid leukemia: clinical and prognostic implications. *Blood*.

Thomas, R. S., Tymms, M. J., Seth, A., Shannon, M. F. and Kola, I. (1995). ETS1 transactevates the human GM-CSF promoter in Jurkat T cells stimulated with PMA and ionomycin. *Oncogene* **11**, 2135-2143.

Tiacci, E., Trifonov, V., Schiavoni, G., Holmes, A., Kern, W., Martelli, M. P., Pucciarini, A., Bigerna, B., Pacini, R., Wells, V. A. et al. (2011). BRAF Mutations in Hairy-Cell Leukemia. *New England Journal of Medicine* **364**, 2305-2315.

Tuborgh, A., Meyer, C., Marschalek, R., Preiss, B., Hasle, H. and Kjeldsen, E. (2013). Complex three-way translocation involving MLL, ELL, RREB1, and CMAHP genes in an infant with acute myeloid leukemia and t(6;19;11)(p22.2;p13.1;q23.3). *Cytogenet Genome Res* **141**, 7-15.

Uren, A. G., Kool, J., Matentzoglu, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T. et al. (2008). Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133**, 727-41.

Uren, A. G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A. H., Wilson, C. H., Rance, R., Jonkers, J., van Lohuizen, M., Berns, A. et al. (2009). A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protocols* **4**, 789-798.

Valnet-Rabier, M.-B., Challier, B., Thiebault, S., Angonin, R., Margueritte, G., Mougin, C., Kantelip, B., Deconinck, E., Cahn, J.-Y. and Fest, T. (2005). c-Flip protein expression in Burkitt's lymphomas is associated with a poor clinical outcome. *British Journal of Haematology* **128**, 767-773.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B. et al. (2010). Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5.

Vaqué, J. P., Martínez, N., Batlle-López, A., Pérez, C., Montes-Moreno, S., Sánchez-Beato, M. and Piris, M. A. (2014). B-cell lymphoma mutations: improving diagnostics and enabling targeted therapies. *Haematologica* **99**, 222-231.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M.-L., Teague, J. et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539-542.

Vassiliou, G. S., Cooper, J. L., Rad, R., Li, J., Rice, S., Uren, A., Rad, L., Ellis, P., Andrews, R., Banerjee, R. et al. (2011). Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice. *Nat Genet* **43**, 470-475.

Verhaak, R. G. W., Goudswaard, C. S., van Putten, W., Bijl, M. A., Sanders, M. A., Hugens, W., Uitterlinden, A. G., Erpelinck, C. A. J., Delwel, R., Löwenberg, B. et al. (2005). Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* **106**, 3747-3754.

Wagner, S. D., Ahearne, M. and Ferrigno, P. K. (2011). The role of BCL6 in lymphomas and routes to therapy. *British Journal of Haematology* **152**, 3-12.

Walter, M. J., Shen, D., Ding, L., Shao, J., Koboldt, D. C., Chen, K., Larson, D. E., McLellan, M. D., Dooling, D., Abbott, R. et al. (2012). Clonal Architecture of Secondary Acute Myeloid Leukemia. *New England Journal of Medicine* **366**, 1090-1098.

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A. and Liu, P. (2008). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* **105**, 9290-5.

Wang, X., Li, Z., Naganuma, A. and Ye, B. H. (2002). Negative autoregulation of BCL-6 is bypassed by genetic alterations in diffuse large B cell lymphomas. *Proceedings of the National Academy of Sciences* **99**, 15018-15023.

Wang, Y. X., Zhang, J. H. and Gu, Z. W. (2009). [Beta-catenin and cyclin D1 mRNA levels in newly diagnosed patients with acute myeloid leukemia and their significance]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi* **17**, 304-8.

Ward, A. F., Braun, B. S. and Shannon, K. M. (2012). Targeting oncogenic Ras signaling in hematologic malignancies. *Blood* **120**, 3397-3406.

Ward, J. M. (2006). Lymphomas and leukemias in mice. *Experimental and Toxicologic Pathology* **57**, 377-381.

Wartman, L. D., Larson, D. E., Xiang, Z., Ding, L., Chen, K., Lin, L., Cahan, P., Klco, J. M., Welch, J. S., Li, C. et al. (2011). Sequencing a mouse acute

promyelocytic leukemia genome reveals genetic events relevant for disease progression. *The Journal of Clinical Investigation* **121**, 1445-1455.
Weigert, O. and Weinstock, D. M. (2012). The evolving contribution of hematopoietic progenitor cells to lymphomagenesis. *Blood* **120**, 2553-2561.
Welch, John S., Ley, Timothy J., Link, Daniel C., Miller, Christopher A., Larson, David E., Koboldt, Daniel C., Wartman, Lukas D., Lamprecht, Tamara L., Liu, F., Xia, J. et al. (2012). The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264-278.
Wells, S. M., Kantor, A. B. and Stall, A. M. (1994). CD43 (S7) expression identifies peripheral B cell subsets. *The Journal of Immunology* **153**, 5503-15.
Wiemels, J. L., Ford, A. M., Van Wering, E. R., Postma, A. and Greaves, M. (1999). Protracted and Variable Latency of Acute Lymphoblastic Leukemia AfterTEL-AML1 Gene Fusion In Utero: Presented at the American Society of Hematology Meeting, held in Miami Beach, FL, December 4-8, 1998, and published as an abstract in Blood 92:68a, 1998 (suppl 1). *Blood* **94**, 1057-1062.
Wiemels, J. L., Xiao, Z., Buffler, P. A., Maia, A. T., Ma, X., Dicks, B. M., Smith, M. T., Zhang, L., Feusner, J., Wiencke, J. et al. (2002). In utero origin of t(8;21) AML1-ETO translocations in childhood acute myeloid leukemia. *Blood* **99**, 3801-3805.
Wouters, B. J., Löwenberg, B., Erpelinck-Verschueren, C. A. J., van Putten, W. L. J., Valk, P. J. M. and Delwel, R. (2009). Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088-3091.
Xiao, C., Srinivasan, L., Calado, D. P., Patterson, H. C., Zhang, B., Wang, J., Henderson, J. M., Kutok, J. L. and Rajewsky, K. (2008). Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes. *Nat Immunol* **9**, 405-14.
Xiao, J., Lee, S. T., Xiao, Y., Ma, X., Houseman, E. A., Hsu, L. I., Roy, R., Wrensch, M., de Smith, A. J., Chokkalingam, A. et al. (2014). PTPRG inhibition by DNA methylation and cooperation with RAS gene activation in childhood acute lymphoblastic leukemia. *Int J Cancer* **4**, 28759.
Yant, S. R., Huang, Y., Akache, B. and Kay, M. A. (2007). Site-directed transposon integration in human cells. *Nucleic Acids Res* **35**, e50.
Yant, S. R., Park, J., Huang, Y., Mikkelsen, J. G. and Kay, M. A. (2004). Mutational Analysis of the N-Terminal DNA-Binding Domain of Sleeping Beauty Transposase: Critical Residues for DNA Binding and Hyperactivity in Mammalian Cells. *Mol Cell Biol* **24**, 9239-9247.
Ye, B. H., Cattoretti, G., Shen, Q., Zhang, J., Hawe, N., Waard, R. d., Leung, C., Nouri-Shirazi, M., Orazi, A., Chaganti, R. S. K. et al. (1997). The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nat Genet* **16**, 161-170.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871.

Yoshida, K., Toki, T., Okuno, Y., Kanezaki, R., Shiraishi, Y., Sato-Otsubo, A., Sanada, M., Park, M.-j., Terui, K., Suzuki, H. et al. (2013). The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nat Genet* **45**, 1293-1299.

Young, D. and Griffin, J. (1986). Autocrine secretion of GM-CSF in acute myeloblastic leukemia. *Blood* **68**, 1178-1181.

Yusa, K., Zhou, L., Li, M. A., Bradley, A. and Craig, N. L. (2011). A hyperactive piggyBac transposase for mammalian applications. *Proc Natl Acad Sci U S A* **108**, 1531-6.

Zayed, H., Izsvak, Z., Walisko, O. and Ivics, Z. (2004). Development of Hyperactive Sleeping Beauty Transposon Vectors by Mutational Analysis. *Mol Ther* **9**, 292-304.

Zhang, J., Grubor, V., Love, C. L., Banerjee, A., Richards, K. L., Mieczkowski, P. A., Dunphy, C., Choi, W., Au, W. Y., Srivastava, G. et al. (2013). Genetic heterogeneity of diffuse large B-cell lymphoma. *Proceedings of the National Academy of Sciences* **110**, 1398-1403.

Zhang, S., Qian, X., Redman, C., Bliskovski, V., Ramsay, E. S., Lowy, D. R. and Mock, B. A. (2003). p16INK4a gene promoter variation and differential binding of a repressor, the ras-responsive zinc-finger transcription factor, RREB. *Oncogene* **22**, 2285-2295.

Zhang, Y., Zhang, M., Yang, L. and Xiao, Z. (2007). NPM1 mutations in myelodysplastic syndromes and acute myeloid leukemia with normal karyotype. *Leukemia Research* **31**, 109-111.

Zingone, A., Cultraro, C. M., Shin, D. M., Bean, C. M., Morse, H. C., 3rd, Janz, S. and Kuehl, W. M. Ectopic expression of wild-type FGFR3 cooperates with MYC to accelerate development of B-cell lineage neoplasms. *Leukemia* **24**, 1171-8.

Appendices

Appendix 2A: Linker sequences from GENEART for the *Vk*hPB* and *Vk*MYC-TA-hPB* constructs

Sequence 1 – Vk*hPB linker

GGATCC*A****GAATTC***TT**CCTCAGC**CCCTCAACggcagcagcctggacgacgagcacatcctgagcgcc ctgctgcagagcgacgacgagctggtcggcgaggacagcgacagcgagatcagcgac**cacgtg**

Sequence 2 – Vk*hMYC-TA-hPB linker

GAATTCaGGTACC*aagaa**atcgat**gttgtttctgtggaaaagaggcaggctcctggcaaaaggtcagagtctggatcac cttctgctggaggccacagcaaacctcctcacagcccactggtcctcaagaggtgccacgtctccacacatcagcacaactacg cagcgcctccctccactcggaaggactatcctgctgccaagagggtcaagttggacagtgtcagagtcctgagacagatcagc aacaaccgaaaatgcaccagccccaggtcctcggacaccgaggagaatgtcaagaggcgaacacacaacgtcttggagcg ccagaggaggaacgagctaaaacggagctttttgccctgcgtgaccagatcccggagttggaaaacaatgaaaaggccccc aaggtagttatccttaaaaaagccacagcatacatcctgtccgtccaagcagaggagcaaaagctcatttctgaagaggacttgt tgcggaaacgacgagaacagttgaaacacaaacttgaacagctacggaactcttgtgcg**gagggcagaggaagtcttcta acatgcggtgacgtggaggagaatcccggccct**ggcagcagcctggacgacgagcacatcctgagcgccctgctg cagagcgacgacgagctggtcggcgaggacagcgacagcgagatcagcgac**cacgtg**agc*

Open reading frame
Restriction sites KpnI = GGTACC, PmlI = CACGTG, EcoRI = GAATTC, ClaI = ATCGAT, BbvcI = CCTCAGC
From just before start of hMYC
hMYC exon 3 (last) excluding stop codon
T2A linker peptide
mPB from 2nd codon to PmlI site

Appendix 2B: Primers used for TraDIS sequencing

These primer sequences were provided by Iraad Bronner

PiggyBac

Name	Order sequence	temp
PB5pr_1	g*atatacagaccgataaaacacatgcgtc*a	63
PB5pr_2	a*atgatacggcgaccaccgagatctacaccacgcatgattatctttaacgtacgtca*c	65
PB5pr_seq_2	c*accgagatctacaccacgcatgattatctttaacgtacgtcacaatatgattatctttt*c	-
PB3pr_1	g*acggattcgcgctatttagaaagaga*g	63
PB3pr_2	a*atgatacggcgaccaccgagatctacacatgcgtcaattttacgcagactat*c	65
PB3pr_seq_3	c*accgagatctacacatgcgtcaattttacgcagactatcttt*c	-

Sleeping Beauty

Name	Order sequence	temp
SB5pr_1	t*ttgttaacaagaaatttgtggagtagtt*g	63
SB5pr_2	a*atgatacggcgaccaccgagatctacacaaaaacgagttttaatgactccaa*c	65
SB5pr_seq_3	a*aaaacgagttttaatgactccaacttaagtgtatgtaaacttcc*g	-
SB3pr_1	a*ctgaccttaagacagggaatctttact*c	63
SB3pr_2	a*atgatacggcgaccaccgagatctacacggaatctttactcggattaaatgtca*g	65
SB3pr_seq_4b	g*tgagtttaaatgtatttggctaaggtgtatgtaaacttcc*g	-

qPCR primers

Name	sequence (for ordering)	temp
qPCR2.1	a*atgatacggcgaccaccgagat*c	60
qPCR2.2	c*aagcagaagacggcatacgaga*t	60
PB5prseqR1	t*gattatctttaacgtacgtcacaatatgattatctttt*c	60
PB3prseqR1	a*tgcgtcaattttacgcagactatcttt*c	60
SB5prseqR1	t*gactccaacttaagtgtatgtaaacttcc*g	60
SB3prseqR1	t*ttggctaaggtgtatgtaaacttcc*g	60

General Splinkerette primers and adapter primers

Name	Sequence (for ordering)	temp
SplAP1	g*ttcccatggtactactcat*a	63
Spl_rev_seq	t*aatacgactcactataggtgacagcgagcgc*t	-
Spl_tag_seq	a*gcgctcgctgtcacctatagtgagtcgtatt*a	-
Splinkerette V1.2 top strand	g*ttcccatggtactactcatataatacgactcactataggtgacagcgagcgc*t	ND
Splinkerette V1.2 bottom strand	/5Phos/g*cgctcgctgtcacctatagtgagtcgtattataattttttttcaaaaaa*a	ND

Splinkerette V1.2 index primer sequences

Only the first ten are shown

name	Sequence (for ordering)	Obtained tag sequence
P7_SplAP2_V1.1	c*aagcagaagacggcatacgagatcggtACAAGCTAtaatacgactcactatag*g	tagcttgt
P7_SplAP2_V1.2	c*aagcagaagacggcatacgagatcggtAAACATCGtaatacgactcactatag*g	cgatgttt
P7_SplAP2_V1.3	c*aagcagaagacggcatacgagatcggtACATTGGCtaatacgactcactatag*g	gccaatgt
P7_SplAP2_V1.4	c*aagcagaagacggcatacgagatcggtACCACTGTtaatacgactcactatag*g	acagtggt
P7_SplAP2_V1.5	c*aagcagaagacggcatacgagatcggtAACGTGATtaatacgactcactatag*g	atcacgtt
P7_SplAP2_V1.6	c*aagcagaagacggcatacgagatcggtCGCTGATCtaatacgactcactatag*g	gatcagcg
P7_SplAP2_V1.7	c*aagcagaagacggcatacgagatcggtCAGATCTGtaatacgactcactatag*g	cagatctg
P7_SplAP2_V1.8	c*aagcagaagacggcatacgagatcggtATGCCTAAtaatacgactcactatag*g	ttaggcat
P7_SplAP2_V1.9	c*aagcagaagacggcatacgagatcggtCTGTAGCCtaatacgactcactatag*g	ggctacag
P7_SplAP2_V1.10	c*aagcagaagacggcatacgagatcggtAGTACAAGtaatacgactcactatag*g	cttgtact

Appendix 3A: Pindel variants detected in the CMML (blue) and AML (grey) samples

													Number of unique reads called as variant by Pindel and BWA				Pindel variant reads				Total unique read depth			
													Disease		Normal		Disease		Normal		Disease		Normal	
	Chr	Call start position	Call end position	Variant	Length	Sequence altered	Sum of mapping score	Simple score	Statistical Score	Annotation	Repeat number	Start of Repeat	End of Repeat	+	-	+	-	+	-	+	-	+	-	+	-
CMML	1	26608812	26608889	D	78	ccagg<68>cggga	116	9	160.45	UBXN11|CCDS41288.1|r.1929_2006del78|c.1464_1541del78|p.G490_P515delGPGPSPGPGPGPSPGPGPGPSPCPGP	1	26608811	26608892	2	2	0	0	2	2	0	0	22	3	19	3
	5	170837543	170837544	I	4	TCTG	250	8	263.64	NPM1|CCDS4376.1|r.1104_1105insucug|c.859_860insTCTG|p.W288fs*12	1	1.71E+08	170837548	8	30	0	0	7	0	0	0	69	111	104	127
	5	170837547	170837548	I	4	TCTG	1374	27	982.50	NPM1|CCDS4376.1|r.1108_1109insucug|c.863_864insTCTG|p.W288fs*12	0	1.71E+08	170837548	0	27	0	0	0	26	0	0	65	111	103	127
	6	84061842	84061864	DI	23	AAGTAAAAAAAAAACCTCTGCA	87	4	0.00	ME1 Coding	0	84061842	84061864	0	3	0	0	0	3	0	0	12	22	9	22
	11	76751543	76751585	DI	43	GGAGCGCGCCGGCCTGGCGCCCAGCGGCCACGAGGGCATCC	87	4	0.00	B3GNT6 Coding	1	76751543	76751585	3	0	0	0	3	0	0	0	13	0	7	1
	17	45232152	45234278	DI	2127	G	83	4	0.00	CDC27 Coding	1	45232152	45234278	3	0	0	0	3	0	0	0	20	101	21	105
	17	39240803	39240804	I	15	ATCTCCAGCTGCTGT	97	4	93.74	KRTAP4-7|CCDS45673.1|r.402_403insaucuccagcugcugu|c.345_346insATCTCCAGCTGCTGT|p.C115_R116insISSCC	0	39240803	39240804	0	3	0	0	0	3	0	0	23	11	20	6
	19	36002421	36002423	D	3	acc	87	4	113.62	DMKN|CCDS12463.1|r.985_987delGGU|c.808_810delGGT|p.G270delG	1	36002418	36002426	7	7	0	3	0	3	0	0	13	12	4	9

Chr = Chromosome. Variant D = deletion, I= insertion, DI = complex

													Number of unique reads called as variant by Pindel and BWA				Pindel variant reads				Total unique read depth			
													Disease		Normal		Disease		Normal		Disease		Normal	
	Chr	Call start position	Call end position	Variant	Length	Sequence altered	Sum of map score	Simple score	Statistical Score	Annotation	Repeats	Start of Repeat	End of Repeat	+	-	+	-	+	-	+	-	+	-	+	-
AML	1	109792735	109792736	I	3	CGC	180	4	114.92	CELSR2|CCDS796.1|r.95_96inscgc|c.34_35insCGC|p.P16_L17insP	5	109792735	109792751	10	3	3	3	3	0	0	0	31	6	27	5
	5	65892767	65892768	I	3	GCC	180	4	115.27	MAST4|CCDS47224.1|r.560_561insgcc|c.284_285insGCC|p.P98_L99insP	3	65892767	65892779	3	3	0	0	3	0	0	0	4	6	0	6
	5	170837547	170837548	I	4	TCTG	1343	27	983.33	NPM1|CCDS4376.1|r.1108_1109insucug|c.863_864insTCTG|p.W288fs*12	1	170837543	170837548	18	33	0	0	0	26	0	0	97	122	104	127
	15	102292941	102292967	DI	27	GAGGCAGACCAAGGAGTTCAT	55	5	0.00	AtC107977.1 Coding	0	102292941	102292967	0	4	0	0	0	4	0	0	22	40	24	23
	17	45232152	45234278	DI	2127	G	69	4	0.00	CDC27 Coding	1	45232152	45234278	3	0	0	0	3	0	0	0	24	109	21	105
	19	36002419	36002421	D	3	cca	87	4	120.90	DMKN|CCDS12463.1|r.987_989delUGG|c.810_812delTGG|p.G271delG	2	36002418	36002426	8	4	0	3	3	0	0	0	15	15	4	9
	19	33793204	33793205	I	2	CG	240	5	156.38	CEBPA|ENST00000498907|r.266_267inscg|c.116_117insCG|p.Q41fs*120	2	33793199	33793205	0	5	0	0	0	4	0	0	5	16	5	10
	22	20779973	20779974	I	1	G	209	5	159.86	SCARF2|CCDS13779.1|r.2409_2410insc|c.2304_2305insC|p.E769fs*9	2	20779973	20779976	4	2	0	1	4	0	0	0	4	5	2	1
	X	50350686	50350713	DI	28	CTCCTCTTCCTCTTCCTCT	180	4	0.00	SHROOM4 Coding	0	50350686	50350713	3	0	0	0	3	0	0	0	13	14	14	24
	X	104464237	104464282	DI	46	AT	116	8	0.00	TEX13A Coding	1	104464237	104464282	1	3	0	0	1	3	0	0	35	21	47	10

Appendix 3B: SNV which are unique to either the CMML or AML samples on Caveman call

							Allele		Dept		% mutant in	
	GENE	CHR	Position	cDNA	Protein	Type	WT	MT	Normal	Tumour	Normal	Tumour
CMML	PTCHD2	1	11595708	c.3816+7A>C	p.?	splice	A	C	33	44	0	13.64
	PRG4	1	186276486	c.1635A>C	p.A545A	silent	A	C	132	140	3.79	10
	PRG4	1	186276589	c.1738A>C	p.T580P	misssense	A	C	171	145	4.09	8.28
	ITPKB	1	226924822	c.338T>G	p.V113G	misssense	A	C	57	53	1.75	11.32
	ARL6IP6	2	153575160	c.22T>G	p.W8G	misssense	T	G	97	89	1.03	8.99
	TLK1	2	172017001	Non-coding	r.343u>g	5' UTR	A	C	38	52	10.53	19.23
	SPEG	2	220348806	c.6621A>C	p.A2207A	silent	A	C	30	39	0	15.38
	AGAP1	2	236877171	c.1549G>C	p.D517H	misssense	G	C	83	66	3.61	15.15
	CAND2	3	12857461	c.1116T>G	p.G372G	silent	T	G	67	79	2.99	10.13
	TEX264	3	51733561	c.620A>G	p.E207G	misssense	A	G	134	156	2.24	5.77
	PLXNA1	3	126708354	c.849T>G	p.G283G	silent	T	G	195	280	2.56	7.14
	SOX2	3	181430812	c.664A>C	p.T222P	misssense	A	C	77	88	2.6	7.95
	TBC1D1	4	38016337	c.625T>G	p.S209A	misssense	T	G	75	68	5.33	14.71
	SPATA18	4	52938111	c.547G>C	p.A183P	misssense	G	C	60	67	5	10.45
	SLC22A23	6	3324119	c.1031T>G	p.V344G	misssense	A	C	21	34	0	14.71
	HLA-A	6	29911240	c.539T>A	p.L180*	nonsense	T	A	47	53	4.26	7.55
	HLA-A	6	29911271	c.570G>C	p.E190D	misssense	G	C	48	58	2.08	12.07
	WASF1	6	110423242	c.1071A>C	p.P357P	silent	T	G	45	41	0	12.2
	FAM160B2	8	21953856	c.133A>C	p.T45P	misssense	A	C	21	28	9.52	17.86
	KIAA1529	9	100071811	c.734T>G	p.V245G	misssense	T	G	65	86	6.15	8.14
	RXRA	9	137300857	c.502A>C	p.T168P	misssense	A	C	138	171	5.8	7.6
	SYT15	10	46970440	c.7+2T>G	p.?	essential splice	A	C	13	26	0	26.92
	SH2D4B	10	82363515	c.824A>C	p.D275A	misssense	A	C	26	24	3.85	16.67
	KRTAP5-2	11	1619378	c.103C>T	p.R35C	misssense	G	A	97	100	1.03	7
	KRT83	12	52713016	c.517T>G	p.C173G	misssense	A	C	149	161	4.7	6.83
	PTPN11	12	112915523	c.922A>G	p.N308D	misssense	A	G	94	92	0	9.78
	ADAMTS7	15	79059160	c.3093A>C	p.S1031S	silent	T	G	45	42	0	11.9
	ENSG00000179038	16	21817596	Non-coding	r.2542u>c	5' UTR	A	G	37	42	2.7	9.52
	ZNF646	16	31089685	c.2040T>G	p.G680G	silent	T	G	48	87	0	8.05
	NT5C3L	17	39981891	c.763C>G	p.R255G	misssense	G	C	80	85	5	10.59
	HCN2	19	590406	c.461A>G	p.E154G	misssense	A	G	14	22	0	22.73
	CYP2A7	19	41387647	c.190T>C	p.C64R	misssense	A	G	36	38	2.78	13.16
	CYP2A7	19	41387656	c.181T>A	p.F61I	misssense	A	T	30	34	0	8.82
	AP2A1	19	50285864	c.356A>C	p.D119A	misssense	A	C	115	128	4.35	8.59
	CCDC106	19	56164006	c.737A>C	p.Y246S	misssense	A	C	73	98	9.59	12.24
	NCOA6	20	33331075	c.2985A>C	p.A995A	silent	T	G	34	57	2.94	21.05
	KRTAP10-2	21	45970771	c.571G>A	p.V191I	misssense	C	T	128	156	2.34	5.77
	TRIOBP	22	38121786	c.3223T>C	p.S1075P	misssense	T	C	65	73	7.69	12.33
	ACRC	X	70830650	c.1731G>T	p.L577F	misssense	G	T	68	72	0	8.33

							Allele		Depth		% mutant in	
	GENE	CHR	Position	cDNA	Protein	Type	WT	MT	Normal	Tumour	Normal	Tumour
AML	AL355149.1	1	16863213	Non-coding	r.1449g>c	mRNA	C	G	52	65	3.85	7.69
	AL355149.1	1	16863233	Non-coding	r.1429u>g	mRNA	A	C	70	76	2.86	6.58
	ABCA4	1	94490534	c.4610C>T	p.T1537M	misssense	G	A	121	125	0	6.4
	NRAS	1	115258748	c.34G>T	p.G12C	misssense	C	A	201	204	0	7.35
	HELT	4	185940170	c.88A>C	p.T30P	misssense	A	C	29	40	0	15
	RREB1	6	7231841	c.3509T>G	p.V1170G	misssense	T	G	11	18	0	38.89
	UBN2	7	138967815	c.1915G>A	p.A639T	misssense	G	A	78	56	0	26.79
	SOX7	8	10583751	c.664T>C	p.S222P	misssense	A	G	39	44	2.56	18.18
	PDLIM2	8	22451396	c.1032A>C	p.A344A	silent	A	C	10	22	0	18.18
	FAM171A1	10	15255870	c.1717G>A	p.V573I	misssense	C	T	162	165	0	23.03
	FRG2B	10	135438806	c.634C>A	p.R212R	silent	G	T	83	98	3.61	6.12
	APIS	11	43344985	c.549A>G	p.L183L	silent	A	G	71	69	0	30.43
	KRT81	12	52681054	c.1079C>A	p.A360D	misssense	G	T	77	93	1.3	9.68
	AP4S1	14	31554147	Non-coding	r.747c>a	3' UTR	C	A	53	50	0	10
	C17orf97	17	263613	c.979C>A	p.P327T	misssense	C	A	83	82	2.41	8.54
	TMC4	19	54676732	c.79+2T>G	p.?	essential splice	A	C	116	111	5.17	8.11
	RBBP9	20	18477732	c.80T>G	p.V27G	misssense	A	C	30	52	3.33	19.23
	FAM182B	20	25755519	c.437A>C	p.H146P	misssense	T	G	32	30	0	13.33
	ARSH	X	2936713	c.901+2T>G	p.?	essential splice	T	G	61	64	1.64	7.81
	GPR50	X	150349621	c.1566G>C	p.K522N	misssense	G	C	145	118	4.83	7.63

Appendix 4A: CIS integrations that were identified on only 1 or 2 of the CIS analysis methods used for the 454 analysis.

The gene nearest to the CIS peak, the kernel sizes at which the CIS was identified, the location and height of the peak and the boundaries of the CIS are shown, along with the number of hits, the genes in the CIS and the analysis methods by which the CIS was identified

Gene nearest to CIS peak	Kernel size (x1000)	Chromosome	peak location*	peak height*	start	end	CIS width	Number of hits	P value	Genes in CIS*	Method
Rabgap1	10	2	37312541	4.029030834	37308610	37314507	5898	5	2.514E-05	Rabgap1	GV/NSD7
Nfia	100	4	97485158	8.262911755	97426325	97524380	98056	11	4.388E-05	E130114P18Rik Nfia	NSD7
Gm17091	10	5	10663101	3.910404439	10661145	10663101	1957	6	5.428E-05	intergenic	NSD7
Slco3a1	100	7	81586072	9.96738084	81546748	81586072	39325	17	0.0001721	Slco3a1 Gm7580	GV/NSD7
Akap13	30	7	82869581	6.328003393	82869581	82872522	2942	10	7.384E-05	Akap13	LHC
Cbl	30	9	43994061	5.406177651	43985303	43996980	11678	7	6.278E-05	Cbl	NSD7
Gm12068	60, 100	11	24339451	10.39816394	24280654	24359050	78397	16	9.529E-05	Gm12068	GV
Nsd1	100	13	55400553	7.402812078	55351678	55439654	87977	13	0.0001126	Nsd1 Rab24 Prelid1 Mxd3	GV
Fhl1	30	X	53993652	3.952833137	53987836	53993652	5817	6	8.257E-05	Fhl1	NSD7

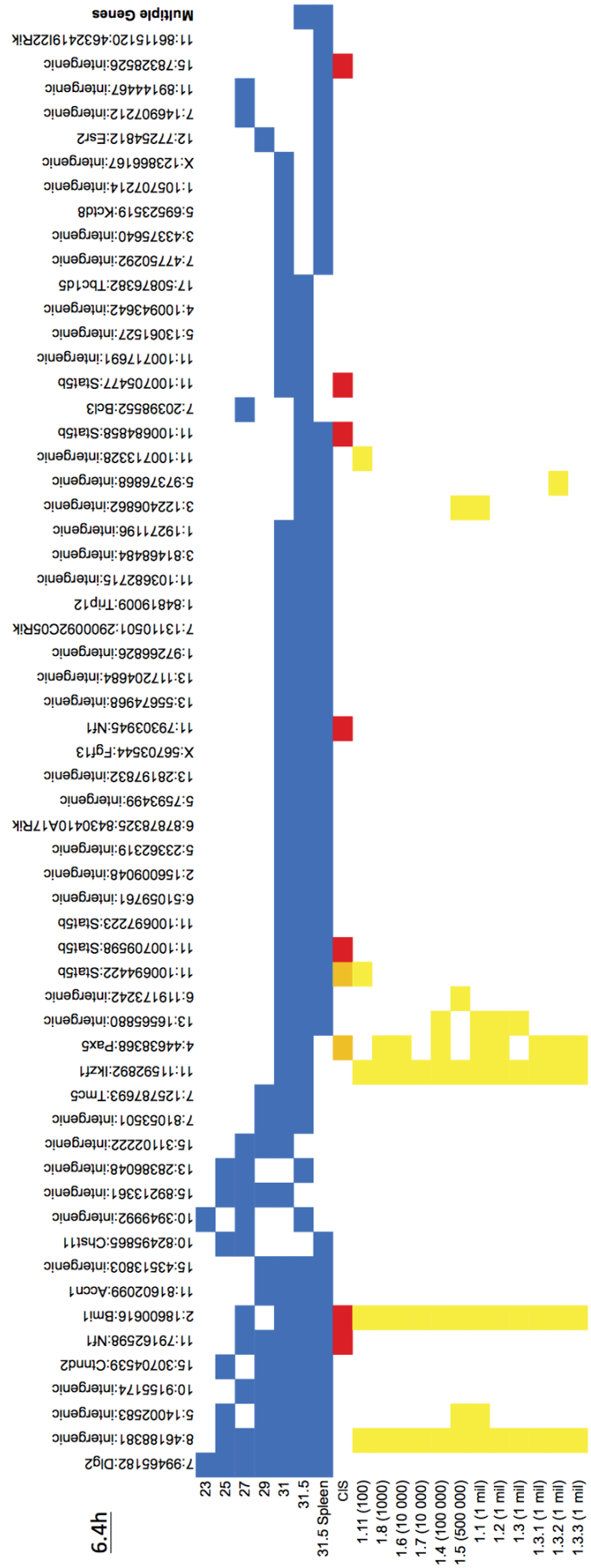
		Genotype														
Name	Sex	Npm1	Mutagenesis	pIpC	Age at death (weeks)	Mouse (g)	Spleen (g)	Liver (g)	Lymph-adenopathy	WBC	Hb	HCT	Plt	MCV	Blood film Changes	Tissue Involvement	Diagnosis
6.4a	Male	WT/c	yes	x6	42.9	44	0.7	3.1	yes	398	9.8	36.7	313	61	CMML like	Spleen, liver, kidney,heart, thymus, BM, LN, muscle	Myeloid leukaemia with maturation
6.4e	Male	WT (Cre neg)	No (Cre neg)	x2	109.3	50.5	0.3G	2.5G	no	10.2	11.1	39.8	1812	58	Normal	Mass in pancreatic node	?Lymphoma in pancreatic node. Spleen follicular hyperplasia
6.4g	Male	WT/c	yes	x2	85.1	45.4	0.5	2.5	no, tumour on leg	14.1	8.7	32.9	679	48	Normal	Spleen	Angiosarcoma leg, follicular hyperplasia
6.4h	Female	WT/c	yes	x6	31.4	45	0.5	2.3	no	34.5	13	52.6	809	58	Undifferentiated blasts	Spleen, LN, BM, ±liver	Undifferentiated leukaemia, MPO, B220 and CD3 neg
7.5b	Male	WT/c	yes	x6	50	45.9	1.2	2.4	yes	114	12.3	46	503	59	Blasts present, v high WCC with maturation	Spleen, liver, LN, BM, muscle	Myeloid leukaemia with maturation
7.5c	Male	WT/c	yes	x6	32.4	39.7	3.71	3.9	no	78.3	22.8	75	4930	67	eukerythroblastic, thrombocytosis, giant platelets	Spleen, bone marrow, liver	Myeloid leukaemia, ?MPD with progression
7.5h	Female	WT/c	yes	x6	26	29.2	1.5	4	no	627	18.3	>70	733	62	CMML like	Liver, BM, spleen, LN	Myeloid leukaemia with maturation
7.7a	Male	WT/c	yes	No	115.9	42.8	0.2	1.9	no	14.4	14.4	55.2	>2200	59	Normal	Spleen, lung	Follicular lymphoma, lung adenocarcinoma
7.7b	Male	WT/c	yes	No	73.6	45.2	1.2	2.4	no	32.5	10.1	35.1	145	67	Leukaemia with poorly differentiated blasts	BM, spleen, liver, muscle, papillary adenoma lung	Myeloid leukaemia, very poorly differentiated, MPO positive
16.3b	Male	WT/c	yes	x6	43	54.4	0.6	2.7	No	109	15.8	59	367	56	Neutrophilia, toxic granulation, left shift, ring forms ++, low blast percentage	Spleen, liver, thymus, LN, BM, muscle	Myeloid leukaemia with maturation
16.3e	Female	WT/c	yes	x6	29.4	34	0.7	2.7	yes	131.2	8.8	32.6	147	71	CMML like	Spleen, liver, LN, thymus, lung, BM, muscle	Myeloid leukaemia with maturation
16.3f	Female	WT/c	yes	x6	55.3	45.2	1	2.2	yes	45.8	11.9	47.2	173	59	Left shift, blasts, v high PMN	Spleen, liver, kidney, LN, thymus, BM. Also B cell infiltrate in lung and B and T infiltrate in liver and kidney	Myeloid leukaemia with lymphoma as probable second diagnosis
16.3g	Female	WT/c	yes	x6	33	42.2	0.8	2.9	no	167	14.5	52.8	129	62	CMML like	Spleen, liver, LN, thymus, BM	Myeloid leukaemia with maturation
16.3h	Female	WT/c	yes	x6	20.3	28.3	0.53	1.65	no	58	14.1	54.3	739	56	Undifferentiated blasts	Spleen, liver, BM	Myeloid leukaemia, AMML like
16.3i	Female	WT (Cre neg)	No (Cre neg)	x6	57	37.4	0.2	1.7	no	7.4	10.3	35.7	1219	51	Sick blood, few WBC	Culled due to tail inflammation	Borderline, ?normal
19.2a	Male	WT	yes	x4	51.6	37.1	0.7	3.1	no, thymomegaly	361	10.9	42.5	427	65	Normal	BM, spleen, LN, liver, kidney	B cell lymphoma, adenoma lung
19.2b	Male	WT/c	yes	x4	27	38.1	1.2	4	yes	221	10.7	41.4	262	79	AMML like	BM, muscle, perineural, liver, kidney, spleen, LN	Myeloid leukaemia, AMML like
19.2d	Female	WT/c	yes	x4	39.6	28.4	1.5	3	yes	595	5.9	23.8	112	101	AMML like	spleen, LN, liver, lung, BM, muscle	Myeloid leukaemia, AMML like
19.3a	Male	WT/c	No (Cre neg)	x4	49.3	24	0.1	1	no	6.5	13.4	54.2	1357	57	Normal	Culled due to eye lesion	Normal
20.2b	Male	WT	yes	x4	47.8	35.3	1.1	2.4	no, thymomegaly	101	10.9	45.7	671	62	Myeloid with blasts and maturation	Spleen, BM, liver, fat, muscle	Myeloid leukaemia with maturation and numerous blasts
20.4e	Female	WT	No (Cre neg)	No	100.9	24.19	0.63	1.45	no, tumours on kidneys	11.8	20.9	66.1	672	54	Unremarkable	Kidney, liver	B cell lymphoma, adenoma lung
20.4f	Female	WT	No (Cre neg)	No	63.2	42.6	0.2	1.7	no						Normal	Culled due to swollen abdomen, pallor	Normal
20.4g	Female	WT	No (Cre neg)	No	53.3	37.7	<0.1	1.7	no	5.8	14	54.9	696	59	Normal	Culled as pale, slow movement	Normal
21.3j	Male	WT/c	yes	x4	25.1	33.9	0.7	1.7	no	92	14.1	61.7	208	65	Myeloid disease with blasts and some maturation	BM, spleen, muscle, liver, kidney	Myeloid leukaemia with blasts and some maturation
22.2b	Male	WT/c	yes	x4	30.8	39	1.1	2.4	no	281	9.6	36.8	165	61	Very high WCC with many blasts	BM, spleen, liver, muscle	Myeloid leukaemia, with many blasts

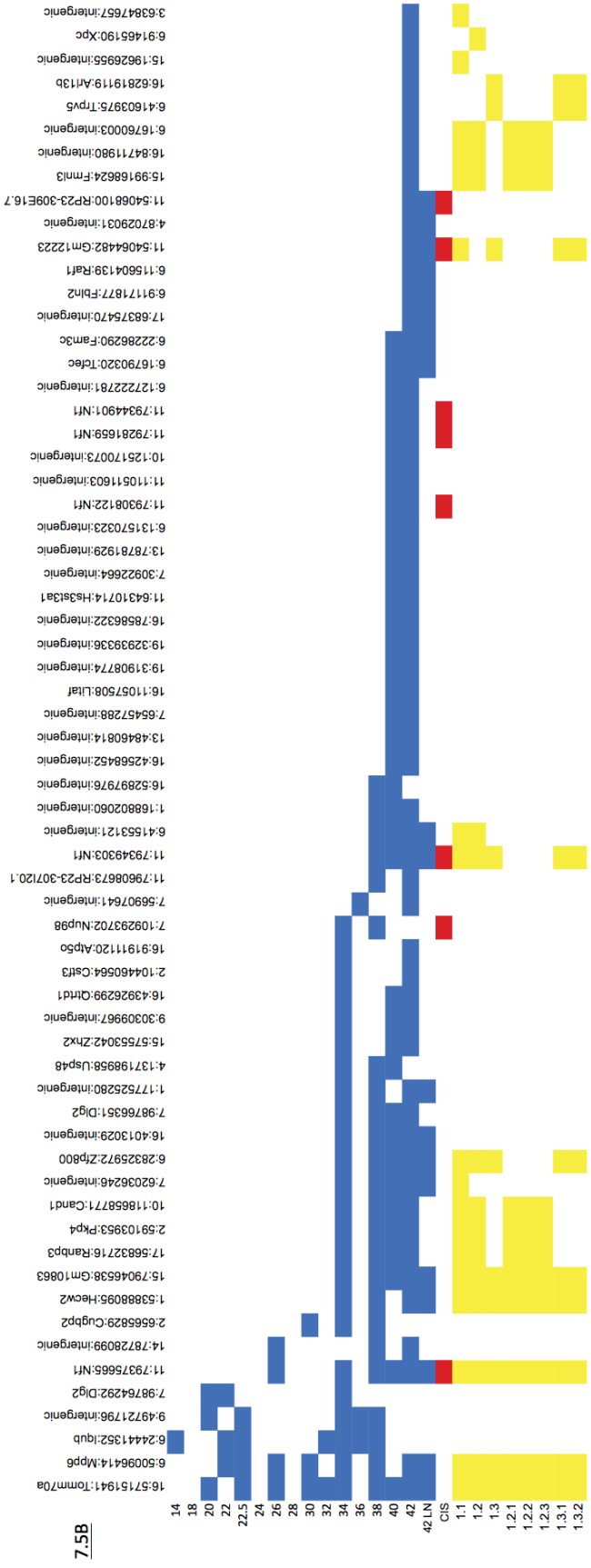
Appendix 4B: Details of the serially bled mice.

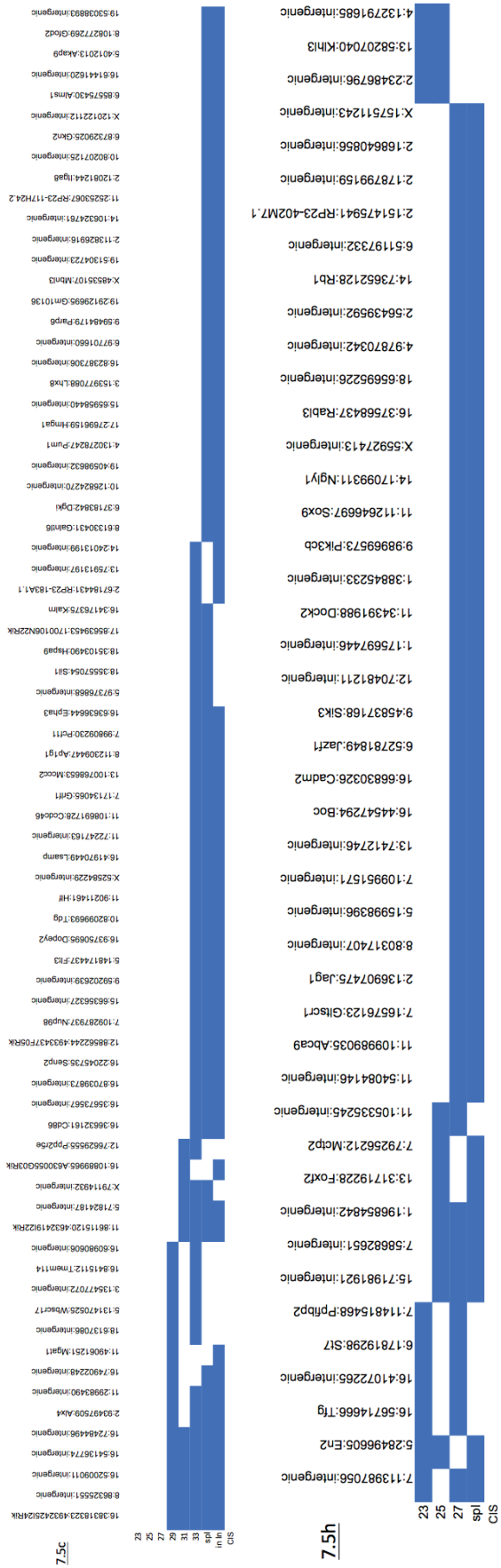
The sex, genotype, number of pIpC injections, age at death, necropsy finding and pathology findings are shown for each mouse.

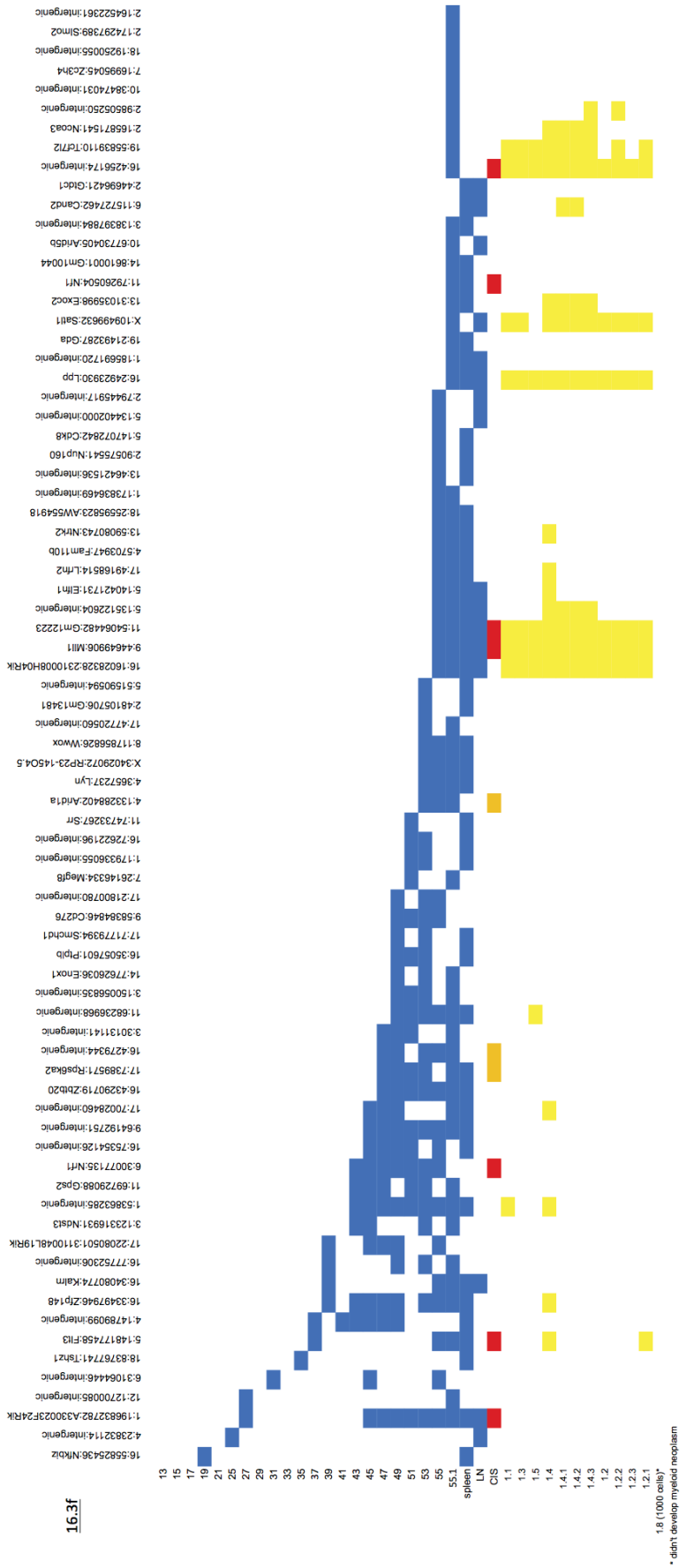
Appendix 4C:

These tables show the shared integrations on blood, primary and recipient tumours for each of the mice that were serially bled. The identity of the mouse is shown at the top left. The precise position of each integration is shown across the top. Integrations in a position are indicated by the coloured squares (blue = serial blood or primary tumour spleen or lymph node, yellow = recipient tumour). The integrations that fall within CISs are indicated in red. The age of the mouse is shown in weeks for the blood samples. IDs of the recipient tumours are indicated. Integrations are shown by the order in which they accumulated and only integrations that persisted on multiple samples are shown. Not all integrations in a given tumour could be represented in these tables.

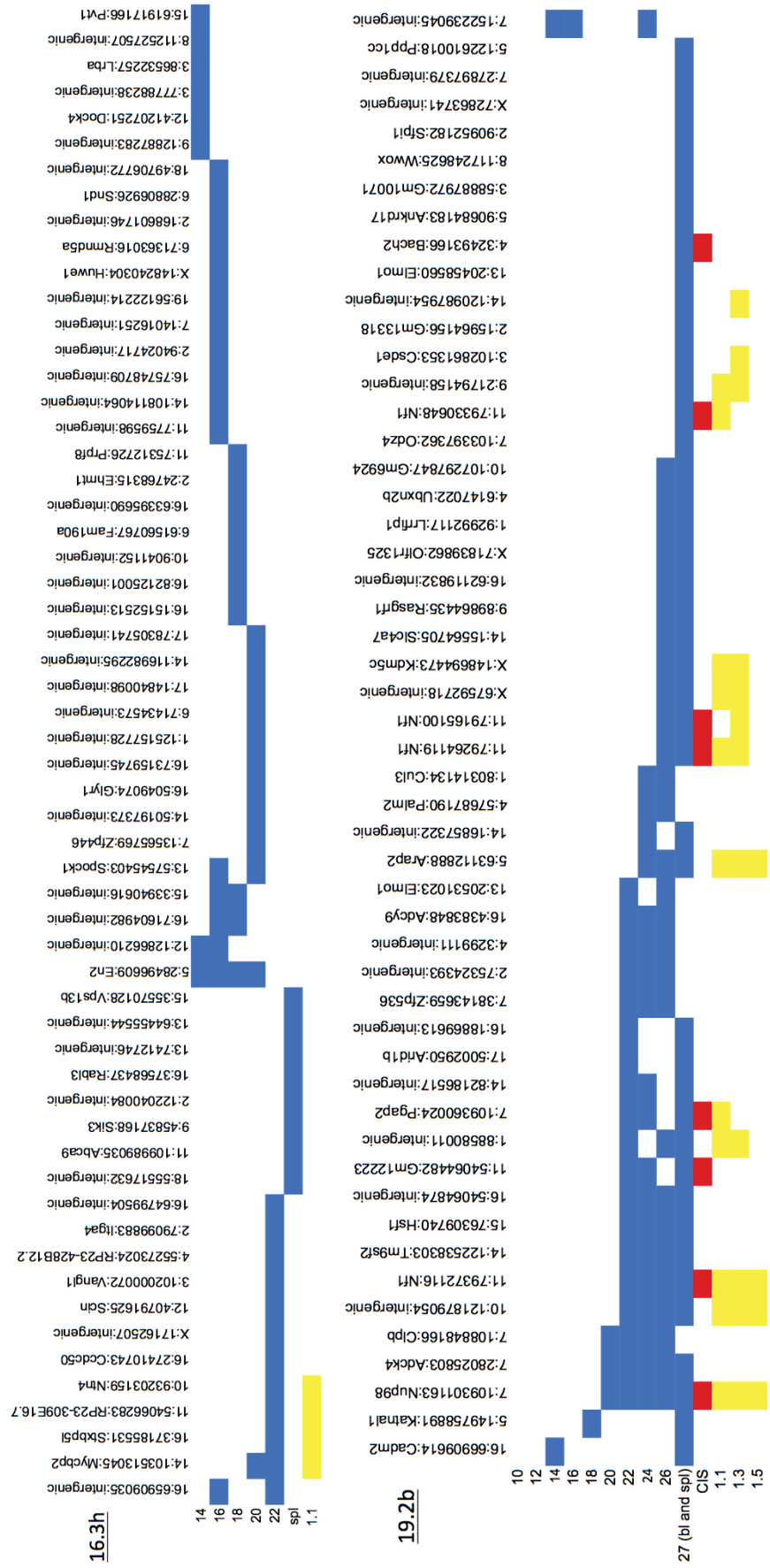

6.4h
23
25
27
29
31
31.5
31.5 Spleen
CIS
1.11 (100)
1.8 (1000)
1.6 (10 000)
1.7 (10 000)
1.4 (100 000)
1.5 (500 000)
1.1 (1 mil)
1.2 (1 mil)
1.3 (1 mil)
1.3.1 (1 mil)
1.3.2 (1 mil)
1.3.3 (1 mil)
Multiple Genes
11:86115120:46324912 2Rik
15:78328526:intergenic
11:89144467:intergenic
7:14690721 2:intergenic
12:77254812:Esr2
X:123866167:intergenic
1:105707214:intergenic
5:69523519:Kctd8
3:43375640:intergenic
7:47750292:intergenic
17:50876382:Tbc1d5
4:100943642:intergenic
5:13061527:intergenic
11:10071769 1:intergenic
11:10070547 7:Stat5b
7:20398552:Bcl3
11:10068485 8:Stat5b
11:10071332 8:intergenic
5:97376868:intergenic
3:12240686 2:intergenic
1:19277196:intergenic
3:81468484:intergenic
11:10368271 5:intergenic
1:84819009:Trip12
7:131105 01:29000920C05Rik
1:97266826:intergenic
13:11720468 4:intergenic
13:55674968:intergenic
11:79303945:Nf1
X:56703544:Fgf13
13:28197832:intergenic
5:75934 99:intergenic
6:87878325:843041 0A17Rik
5:23362319:intergenic
2:156009048:intergenic
6:51059761:intergenic
11:10069722 3:Stat5b
11:10070959 8:Stat5b
11:10069442 2:Stat5b
6:119173242:intergenic
13:16565880:intergenic
4:44638368:Pax5
11:11592892:Ikzf1
7:12578769 3:Tmc5
7:81053501:intergenic
15:31102222:intergenic
13:28386048:intergenic
15:89213361:intergenic
10:39499 92:intergenic
10:82495865:Chst11
15:43513803:intergenic
11:81602099:Accn1
2:186006 16:Bmi1
11:79162598:Nf1
15:30704539:Ctnnd2
10:91551 74:intergenic
5:14002583:intergenic
8:46188381:intergenic
7:99465182:Dlg2

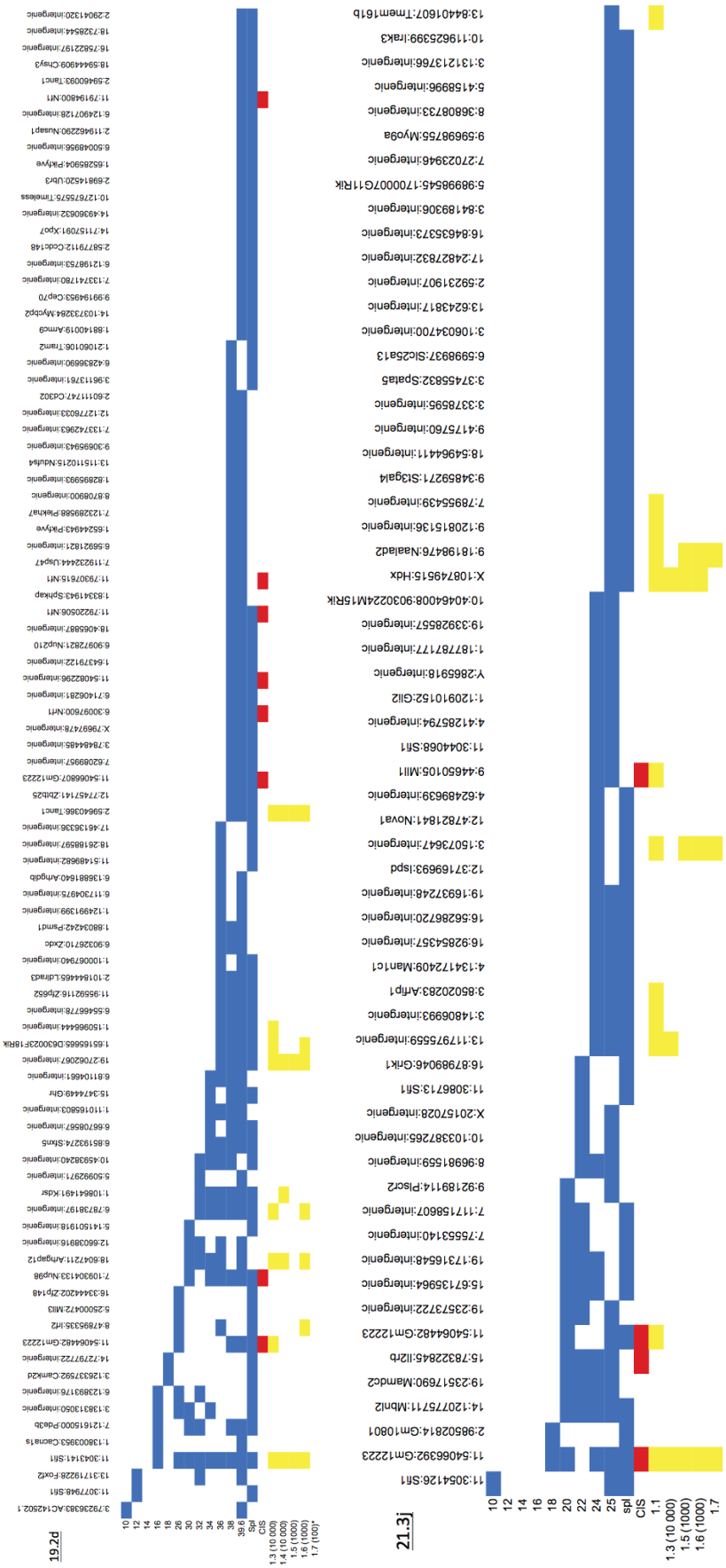

7.5B
14
18
20
22
22.5
24
26
28
30
32
34
36
38
40
42
42 LN
CIS
1.1
1.2
1.3
1.2.1
1.2.2
1.2.3
1.3.1
1.3.2
16:57151941:Tomm70a
6:50096414:Mppb
6:24441352:Iqub
9:49721796:intergenic
7:98764292:Dlg2
11:79375665:Nf1
14:78728099:intergenic
2:65658829:Cugbp2
1:53888095:Hecw2
15:79046538:Gm10863
17:56832716:Ranbp3
2:59103953:Pkp4
10:118658771:Cand1
7:62036246:intergenic
6:28325972:Zfp800
16:40130229:intergenic
7:98766351:Dlg2
1:177525280:intergenic
4:137198958:Usp48
15:57553042:Zhx2
9:30309967:intergenic
16:43926299:Gtdc1
2:104460564:Cstf3
16:9191120:Atp5o
7:109293702:Nup98
7:56907641:intergenic
11:79608673:RP23-307I20.1
11:79349303:Nf1
6:41553121:intergenic
1:168802060:intergenic
16:52897976:intergenic
16:42568452:intergenic
13:48460814:intergenic
7:65457288:intergenic
16:11057508:Ltbr
19:31908774:intergenic
19:32939336:intergenic
16:78586322:intergenic
11:64310714:Hs3st3a1
7:30922664:intergenic
13:78781929:intergenic
6:13157023:intergenic
11:79308122:Nf1
11:110511603:intergenic
10:125170073:intergenic
11:7928165 9:Nf1
11:79344901:Nf1
6:127222781:intergenic
6:16790320:Tcfec
6:22286290:Fam3c
17:68375470:intergenic
6:91171877:Fblnl2
6:11560413 9:Raf1
11:54064482:Gm12223
4:87029031:intergenic
11:54068100:RP23-309E16.7
15:99168682 4:Fmn13
16:84711980:intergenic
6:167600 03:intergenic
6:41603975:Trpv5
16:62819119:Arl13b
15:19626955:intergenic
6:91465190:Xpc
3:63847657:intergenic

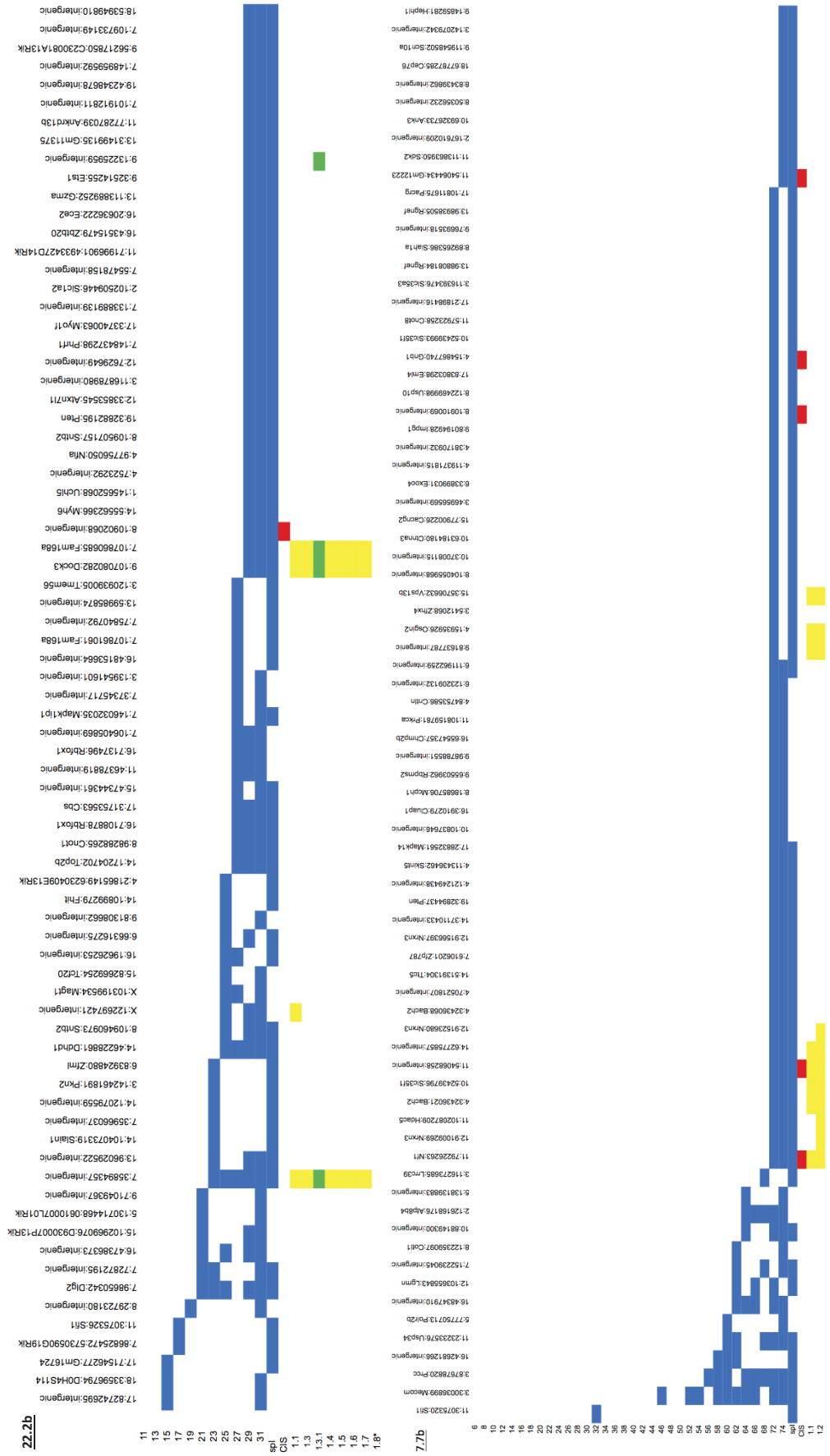

7.5h
23
25
27
spl
CIS
7:113987056:intergenic
5:28496605:En2
16:5671466:Tfg
16:41072265:intergenic
6:178192298:St7
7:114815468:Pcflbp2
15:71981921:intergenic
7:58682651:intergenic
1:196854842:intergenic
13:31719228:Foxf2
7:79256212:Mctp2
11:105335245:intergenic
11:54084146:intergenic
11:109989035:Abca9
7:16576123:Gltscr1
2:136907475:Jag1
8:80317407:intergenic
5:159983396:intergenic
7:109951571:intergenic
13:74127746:intergenic
16:44547294:Boc
16:668303326:Cadm2
6:52781849:Jazf1
9:458371685:Slk3
12:70481211:intergenic
1:175697446:intergenic
11:34391988:Dock2
1:388452533:intergenic
9:98969573:Pik3cb
11:112646697:Sox9
14:170993111:Ngly1
X:55927413:intergenic
16:37568437:Rab13
18:65695226:intergenic
4:97870342:intergenic
2:56439592:intergenic
14:73652128:Rb1
6:51197332:intergenic
2:151475941:RP23-402M7.1
2:178799159:intergenic
2:168640856:intergenic
X:157511243:intergenic
2:234867796:intergenic
13:58207040:Klhl3
4:132791685:intergenic
7.5c
23
25
27
29
31
33
spl
In In
CIS
16:38318323:49324251248Rik
8:86325551:intergenic
16:52009011:intergenic
16:5413677:intergenic
16:7248496:intergenic
2:93497509:Alx4
11:29983490:intergenic
16:7490224:intergenic
11:4906125:Mgat1
18:8137086:intergenic
5:131470525:Wbscr17
3:135477072:intergenic
16:8415112:Tmem114
16:8098060:intergenic
11:86115120:46324912248k
5:71824187:intergenic
X:79114932:intergenic
16:10889965:A630055G03Rik
12:76629555:Ppp2r5e
16:3663216:Cd86
16:3567356:intergenic
16:8703987:intergenic
16:2204735:Senp2
12:88562244:49334370F05Rik
7:109287937:Nup98
15:6635627:intergenic
9:59202563:intergenic
5:14817443:Fil3
16:9375069:Dopey2
10:8209969:Tdg
11:9021146:Hlf
X:52584229:intergenic
16:4197044:Lsamp
11:7224716:intergenic
11:10869172:Cdc42
7:1713406:Gfi1
13:10076865:Mccc2
8:11230944:Ap1g1
7:9980923:Pdf11
16:5363644:Epha3
5:9737686:intergenic
18:3557064:Sil1
18:3510349:Hspa9
17:8563945:1700106N22Rik
16:3417637:Kalrn
2:67184431:RP23-183A1.1
13:7591197:intergenic
14:2401319:intergenic
8:6133043:Galnt16
6:3718384:Dgkl
10:1268427:intergenic
19:4059863:intergenic
4:1302782:Pum1
17:2769615:Hmga1
15:8599644:intergenic
3:1539770:Lhx8
16:8238730:intergenic
6:9770166:intergenic
9:5948417:Parp8
19:2912969:Gm10136
X:48535107:Mbnl3
19:5130472:intergenic
2:113826916:intergenic
14:1063247:intergenic
11:26253067:RP23-117H24.2
2:1208124:Ilgap
10:8020712:intergenic
6:8732902:Gnb2
X:1201221:intergenic
6:8557543:Atm1
16:8144162:intergenic
5:4012013:Akap9
8:1082772:Gfod2
19:5303889:intergenic

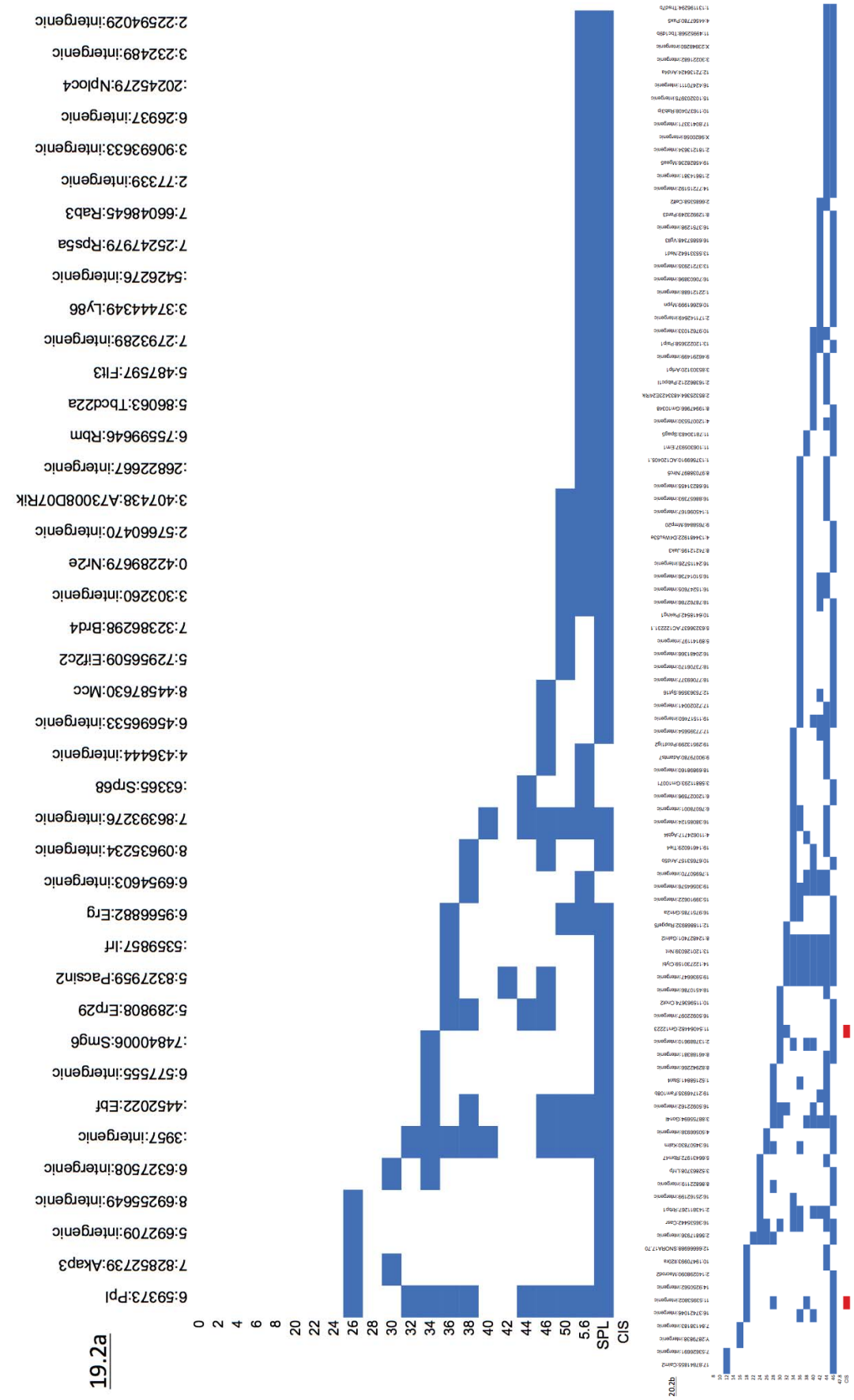

16.3f
16:55825436:NfkbIz
4:23832114:intergenic
1:196832782:A33002JF24RIK
12:12700085:intergenic
6:31064446:intergenic
18:83767741:Tshz1
5:148177458:Fli3
4:147899099:intergenic
16:33497946:Zfp148
16:34080774:Kalrn
16:77752306:intergenic
17:22080501:31100481.19RIK
3:123316931:Ndst3
1:53863285:intergenic
11:69729088:Gps2
6:30077135:Nf1
16:75354126:intergenic
9:64192751:intergenic
17:70028460:intergenic
16:43290719:Zbtb20
17:7389571:Rps6ka2
16:42793441:intergenic
3:301311411:intergenic
11:68236968:intergenic
3:150056835:intergenic
14:77626036:Enox1
16:35057601:Plxlb
17:71779394:Smchd1
9:58384846:Cdc27b
17:21800780:intergenic
7:26146334:Mogt8
1:179336055:intergenic
16:7262219b:intergenic
11:74733267:Srr
4:133288402:Arid1a
X:34029072:RP23-145O4.5
4:36572371:Lyn
8:11785568 26:Wwox
17:47720560:intergenic
2:48105706:Gm13481
5:51590594:intergenic
16:16028328:2310008H04RIK
9:44649906:Mll1
11:54064482:Gm12223
5:135122604:intergenic
5:14042171:Elln1
17:49168514:Lrrc2
4:57039347:Fam110b
13:59080743:Ntrk2
1:178364691:intergenic
18:25595823:AW554918
13:46421536:intergenic
2:90575541:Nup160
5:14707284:2:Cdk8
5:134402000:intergenic
2:79445917:intergenic
16:24923930:Lpp
1:186691720:intergenic
19:21493287:Gda
X:109499632:Satl1
13:31035998:Exoc2
11:79260504:Nf1
14:86100011:Gm10044
10:67730405:Arid5b
3:138397884:intergenic
6:11572746:2:Cand2
2:44696421:Gidc1
16:42661174:intergenic
19:55839110:Tcf7l2
2:16587154:1:Ncoa3
2:98505250:intergenic
10:38474031:intergenic
7:16995045:Zc3h4
18:19250055:intergenic
2:174297389:Slmo2
2:164522361:intergenic
13
15
17
19
21
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
55.1
spleen
LN
CIS
1.1
1.3
1.5
1.4
1.4.1
1.4.2
1.4.3
1.2
1.2.2
1.2.3
1.2.1
1.8 (1000 cells)*
* didn't develop myeloid neoplasm

16.3g
14
16
18
20
22
24
26
28
30
32
34
Spl
CIS
1.1
1.2
1.3
1.2.2
1.2.1
1.2.3
1.3.1
1.3.2
1.3.3
18:14143512:intergenic
6:119412959:Wnt5b
X:137843148:Col4a6
9:40374821:intergenic
7:121331421:intergenic
6:31042036:intergenic
19:26827046:Smarca2
6:31172687:RP23-402K12.2
3:137563155:Dnajb14
2:52633004:intergenic
5:145313552:Nptx2
17:10508393:Qk
7:56129265:E2f8
8:46188381:intergenic
6:29573899:intergenic
2:22775935:Pdss1
11:40530175:Hmmr
9:11398361:intergenic
3:145634059:intergenic
3:87213047:intergenic
9:78726091:intergenic
2:134876246:Plcb1
10:24707426:intergenic
3:51271580:Naa15
16:50468868:intergenic
6:63998490:Grid2
11:101637398:Etv4
3:60270055:intergenic
6:105396059:intergenic
1:122376667:Marco
16:25258807:intergenic
X:57926270:RP23-6C3.4
2:119012780:Dnajc17
14:26428815:Zmiz1
X:150171912:RP23-106P7.2
11:54066279:RP23-309E16.7
Multiple Genes


16.3h
16:65909035:intergenic
14:10351304 5:Mycbp2
16:37185531:Sxbp5l
11:54066283:RP23-309E16.7
10:93203159:Nth4
16:27410743:Ccdc50
X:17162507:intergenic
12:40791625:Scn
3:102000072:Vangl1
4:55273024:RP23-428B12.2
2:79098883:Itga4
16:64799504:intergenic
18:55517632:intergenic
11:109989035:Abca9
9:45837168:Slk3
2:122040084:intergenic
16:37568437:Rab13
13:37412746:intergenic
13:64455544:intergenic
15:35570128:Vps13b
5:28496609:En2
12:12866210:intergenic
16:71604982:intergenic
15:33940616:intergenic
13:57545403:Spock1
7:13565769:Zfp446
14:50193773:intergenic
16:50490774:Glyr1
16:73159745:intergenic
1:12515772 8:intergenic
6:71434573:intergenic
17:14840098:intergenic
14:11698229 5:intergenic
17:78305741:intergenic
16:15152513:intergenic
16:82125001:intergenic
10:90411452:intergenic
6:61560767:Fam190a
16:63395690:intergenic
2:24768315:Ehmt1
11:7753127 26:Prpf8
11:77595 98:intergenic
14:10811406 4:intergenic
16:75748709:intergenic
2:94024717:intergenic
7:14016251:intergenic
19:56122214:intergenic
X:14824030 4:Huwe1
6:71363016:Rmnd5a
2:16860174 6:intergenic
6:28806926:Snd1
18:49706772:intergenic
9:12887283:intergenic
12:41207251:Dock4
3:77788238:intergenic
3:86532257:Lrba
8:11252750 7:intergenic
15:6191716 6:Pvt1
spl
1.1
14
16
18
20
22
19.2b
16:66909614:Cadm2
5:14975889 1:Katnal1
7:10930116 3:Nup98
7:28025803:Adck4
7:10884816 6:Clpb
10:12187905 4:intergenic
11:79372116:Nf1
14:12253830 3:Tm9sf2
15:76309740:Hsf1
16:54064874:intergenic
11:54064482:Gm12223
1:88580011:intergenic
7:10936002 4:Pgap2
14:82186517:intergenic
17:50029 50:Arid1b
16:18869613:intergenic
7:38143659:Zfp536
2:75324393:intergenic
4:32991 11:intergenic
16:43838348:Adcy9
13:20531023:Elmo1
5:63112888:Arap2
14:16857322:intergenic
4:57687190:Palm2
1:80314134:Cul3
11:79264119:Nf1
11:79165100:Nf1
X:67592718:intergenic
X:14869447 3:Kdm5c
14:15564705:Slc4a7
9:89864435:Rasgrf1
16:62119832:intergenic
X:71839862:Olfr1325
1:92992117:Lrrfip1
4:61470 22:Ubxn2b
10:10729784 7:Gm6924
7:10339736 2:Odz4
11:79330648:Nf1
9:21794158:intergenic
3:10286135 3:Csde1
2:15964156:Gm13318
14:12098795 4:intergenic
13:20458560:Elmo1
4:32493166:Bach2
5:90684183:Ankrd17
3:58887972:Gm10071
8:11724862 5:Wwox
2:90952182:Sfpi1
X:72863741:intergenic
7:27897379:intergenic
5:12261001 8:Ppp1cc
7:15223904 5:intergenic
27 (bl and spl)
CIS
1.1
1.3
1.5
10
12
14
16
18
20
22
24
26


19.2d
21.3j


22.2b
7.7b


19.2a
0
2
4
6
8
20
22
24
26
28
30
32
34
36
38
40
42
44
46
50
5.6
SPL
CIS
6:59373:Ppl
7:82852739:Akap3
5:69270:Intergenic
8:69255649:Intergenic
6:63275508:Intergenic
:3957:Intergenic
:44520 22:Ebf
6:57755:Intergenic
:74840006:Smg6
5:28980 8:Erp29
5:83279 59:Pacsin2
:53598 57:Irf
6:95668 82:Erg
6:69546 03:Intergenic
8:09635234:Intergenic
7:86393276:Intergenic
:63365:Srp68
4:4364444:Intergenic
6:45696533:Intergenic
8:44587630:Mcc
5:72956509:Eif2c2
7:32386298:Brd4
3:30326 0:Intergenic
0:42289679:Nr2e
2:57660470:Intergenic
3:40743 8:A730 8D07Rik
:26822667:Intergenic
6:75599646:Rbm
5:86063:Tbc d22a
5:48759 7:Fli3
7:27932 89:Intergenic
3:37444349:Ly86
:54262 76:Intergenic
7:25247979:Rps5a
7:66048645:Rab3
2:77339:Intergenic
3:90693633:Intergenic
6:26937:Intergenic
:20245279:Nploc4
3:23248 9:Intergenic
2:22594029:Intergenic

Appendix 4D: Details of the transplant mice

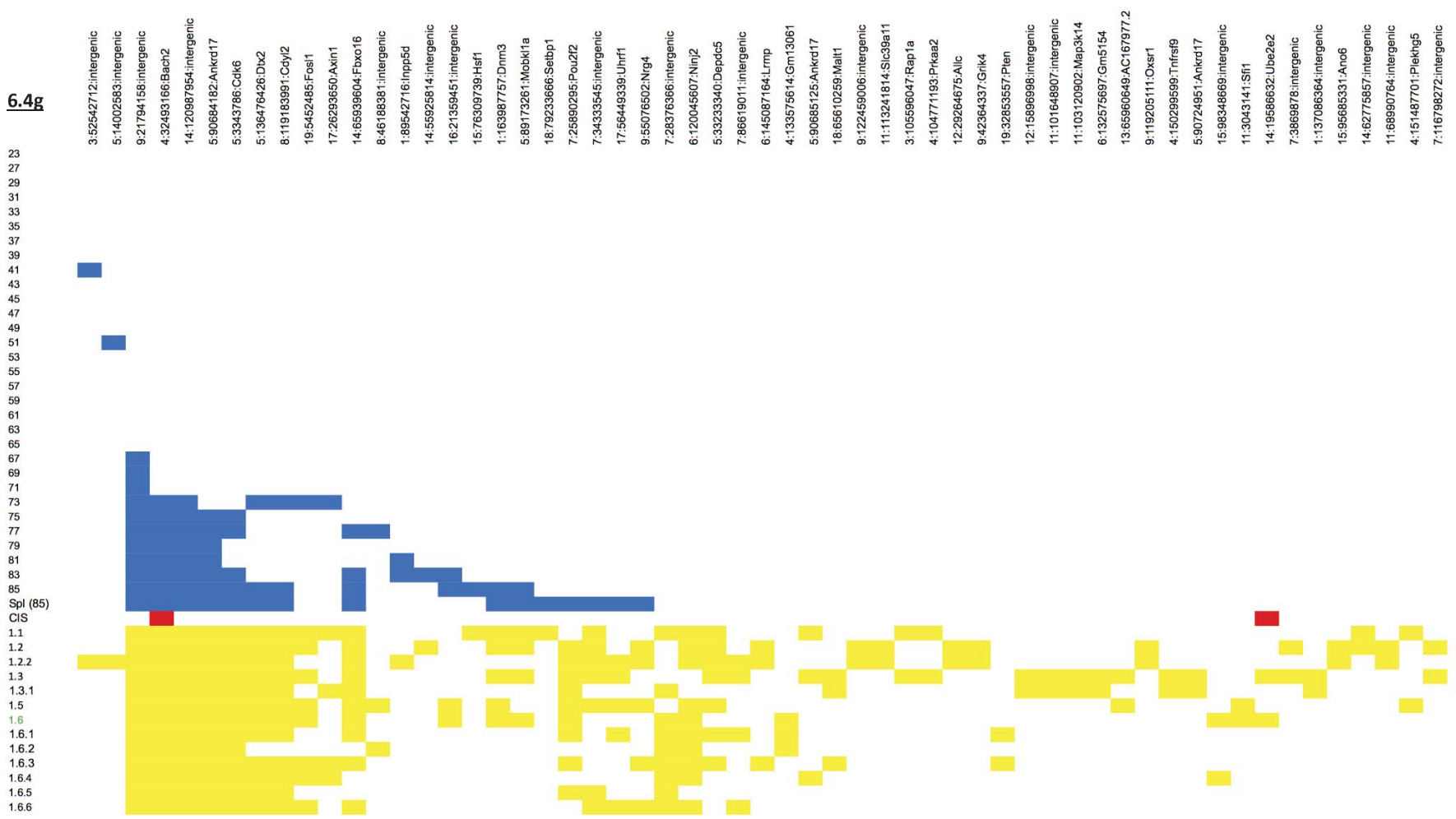
Mouse	Cell Dose	Lifetime (days) post transplant	Necropsy findings	Blood film	Histopath	WCC	Hb	Hct	Plt	MCV	Lymph/blasts	Gran
6.4a		**300**	**spleen 0.7g, liver 3.1g**	**CMML like**	**myeloid leukaemia**	**398**	**9.8**	**36.7**	**313**	**61**	**172**	**173**
1.1	1 million	26	spleen 0.5g, liver 1.6g	AMML	myeloid leukaemia in blood, bone marrow, spleen, liver, LN, bone and periosteum	488	14.1	55.6	575	55	137	288
1.2	1 million	28	spleen 0.5g, liver 2g, 3x pale areas liver, small mesenteric and Rt inguinal LN	AMML	myeloid leukaemia in blood, bone marrow, spleen, liver, kidney, lung, LN, muscle, periosteum	694	13.3	52.3	557	56	155	462
1.3	1 million	28	spleen 0.6g, liver 2.7g, ?small LN	AMML	myeloid leukaemia in blood, spleen, liver, kidney, lung, lymph node, bone marrow, muscle, meninges and periosteum	187	12	46.7	485	59	53	112
1.1.1	1 million	24	spleen 0.6g, liver 1.8g	AMML	myeloid leukaemia in blood, liver, spleen, bone marrow, muscle, minor LN only, and meninges.	248	12.4	45.4	523	60	34	192
1.1.2	1 million	37	parraspinal mass left lumbar region 1.3x0.8cm, arising off pelvis, spleen 0.6g, liver 1.7g	AMML	myeloid leukaemia in blood, spleen, liver, lymph nodes, bone marrow, muscles and meninges	266	10	42.6	467	74	230	27
1.1.3	1 million	119	macroscopic tail lesion ~1cm, spleen 0.1g, liver 1.3g	leukopaenia and thrombocytosis	spleen EMH, tail lesion ant injection site	2.7	13.7	46.7	1132	56	0.8	1.6
1.2.1	1 million	34	splenomegaly 0.4g, liver 1.5g, LN right axilla	AMML	myeloid leukaemia in blood, spleen, liver, renal LN, stomach, bone marrow and muscle	347	10.4	36.4	232	59	104	197
1.2.2	1 million*	92	rear leg paralysis, spleen 0.1g, liver 1.5g	Leukopaenia, thrombocytosis,	spleen extramedulary haematopoiesis	2	16.7	60	1848	57	0.8	0.9
1.2.3	1 million	27	dilated appendix and caecum, no mass lesion, spleen 0.4g, liver 1.7g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow and muscle	273	12.5	46.5	482	56	95	140
1.3.3	1 million	25	spleen 0.4g, liver 1.9g	AMML	myeloid leukamela in blood, spleen, liver, bone marrow and muscle	187	14.2	53.5	629	58	62	98
1.3.2	1 million	26	spleen 0.4g, liver 1.8g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow, muscle and periosteal	268	12.5	48	441	56	102	128
1.3.1	1 million	31	spleen 0.6g, liver 2.1g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow and muscle	150	15.4	57.5	577	59	37	96
1.1.1.1	1 million	19	hunched, thin, weak rear legs, spleen 0.4g, liver 1.6g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow and muscle and periosteum	455	14.4	56.5	654	62	150	245
1.1.1.2	1 million	21	dragging hind limbs, piloerection, reduced mobility spleen 0.5g liver 1.9g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow and massive tumour in muscle and meninges	265	10.8	38.8	548	60	120	110
1.1.1.3	1 million	25	spleen 0.5g, liver 1.3g, no lymphadenopathy	AMML	myeloid leukaemia in blood, spleen, liver, kidney, lung,, bone marrow and muscle and meningeal	135	13.5	51.5	658	59	57	62
1.1.1.1.1	1 million	22	spleen 0.6g, liver 1.9g	AMML	myeloid tumour in blood, spleeen, liver, bone marrow, muscle and meninges	386	12.5	45.4	426	60	114	222
1.1.1.1.2	1 million	22	spleen 0.7g, liver 2.6g	AMML	myeloid leukaemia in blood, spleen , liver, bone marrow, muscle	384	14.9	55.5	625	62	146	187
1.1.1.1.3	1 million	25	spleen 0.5g, liver 1.9g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow and muscles and meninges	446	14.2	52.6	533	58	136	252
1.1.2.1	1 million	48	spleen 0.6g, ?small ing LN, liver 1.7g	AML with maturation	myeloid leukaemia in blood, spleen, BM, liver, muscle, kidney	143	15.1	56.5	622	61	37	87
1.1.2.2.	1 million	40	intestine full of gas, pale extremities, hunched, piloerect, spleen 0.2g, liver 1.3g	AML with high % blasts	myeloid leukaemia in blood, BM, spleen, muscle, liver	208	15.4	65.5	1074	63	191	13
1.1.2.3	1 million	39	spleen 0.6g, liver 2.1g	AML with maturation	Myeloid leukaemia in blood, bone marrow, spleen, liver, muscle	152	13.8	50.8	639	58	70	59
6.4g		**596**	**lump on leg, spleen 0.5g, liver 2.5g**		**Angiosarcoma leg, follicular hyperplasia spleen**	**14.1**	**8.7**	**32.9**	**679**	**48**	**8.1**	**4.9**
1.1	1 million	99	spleen 0.7g, liver 2.1g, kidney 0.8g, large pale kidneys 2x1x1cm, mesenteric LN	Undifferentited blasts with high WCC	Leukaemia without maturation (B220 pos) BM, spleen, liver, LN, kidney, lung, bowel, ?stomach	294	15	56	552	59	196	82
1.2	1 million	91	moribund, rear leg paralysis, liver 4g, spleen 1.3g	AML with some maturation	Leukaemia with minimal myeloid differentiation (B220 pos) BM, spleen, Liver, kidney, lung, LN, muscle	247	16.4	60.2	235	56	137	90
1.3	1 million	95	moribund, spleen 0.9g, liver 2.1g, kidney enlarged, large pale kidneys 2x1x1cm	No sample	Leukaemia without maturation (B220 pos) BM, spleen, kidney, liver, muscle	no sample, culled by animal technician as moribund						
1.4	1 million	134	spleen 1.8g, liver 4.6g, enlarged pale kidneys 0.7g	AML	Leukaemia with minimal myeloid differentiation (B220 pos) BM, spleen, kidney, liver, lung	68	18.7	>70	1470	61	5.4	1.2
1.5	1 million	91	pale, liver 3.3g, spleen 2g, big lungs	AML	Leukaemia with minimal myeloid differentiation (B220 pos) bm, spleen, kidney, lung, ln, liver, muscle	451	15.4	54.1	515	59	221	194
1.6	1 million	83	spleen 0.9g, liver 2.5g, enlarged kidney 1.2g	Leukaemia	Leukaemia without maturation (B220 pos) spleen, LN, liver, kidney, lung , bone marrow , periosteal +	66.4	17.3	66.9	803	55	38.5	21.8
1.2.1	1 million	29	spleen 1g, liver 2.4g	AML	Leukaemia with minimal myeloid differentiation BM, spleen, LN, liver, lung, kidney	52.7	16.2	63	244	56	11.6	37.7
1.2.2	1 million	33	spleen 0.9g, liver 3.2g, big lungs	AML	Leukaemia with minimal myeloid differentiation BM, spleen, LN, liver, lung, kidney, adrenal	75.1	11.7	51.4	381	62	47.3	23.5
1.3.1	1 million	38	spleen 1g, liver 2.7g	AML	Leukaemia with minimal myeloid differentiation BM, spleen, liver, lung, kidney, muscle	250	11.8	53.6	429	67	204	38
1.3.1	1 million	38	spleen 0.7g, liver 2.6g, found dead 19/11/11 and necropsied 25/11/11	No sample	degenerate, likely tumour present BM, spleen, liver, lung, kidney but degenerate	no sample, found dead and delayed necropsy						
1.6.1	1 million	33	spleen 0.9g, liver 2.4g, thymomegaly, big lungs	Leukaemia	Leukaemia without maturation (B220 pos) BM, spleen, liver, lung, kidney, muscle	451	12.1	58.7	399	73	397	39
1.6.2	10 000	39	spleen 0.7g, liver 3.1g, large pale lungs	Leukaemia	Leukaemia without maturation(B220 pos) bm, spleen, liver, kidney, lung, muscle, periosteal	72	13.1	65.2	504	71	63	7
1.6.3	10 000	48	spleen 0.8g, liver 2.3g, kidney 0.5g and pale with abnormal texture	Leukaemia	Leukaemia without maturation (B220 pos) bm, spleen, liver, kidney, lung	37.1	8.8	33.7	183	51	19.5	13.8
1.6.4	1000	48	spleen 1.1g, liver 2g, kidneys 0.3g and look normal	Leukaemia	Leukaemia without maturation (B220 pos) BM, spleen, liver, kidney, lung, muscle	47.6	14.1	51.8	1027	69	34.6	10.3
1.6.5	1000	49	moribund, agonal breathing, congested vessels, spleen 0.6g, liver 1.8g	AML	Leukaemia with minimal myeloid differentiation BM, spleen, liver, lung, kidney, pancreas	100	12.5	39.5	312	73	61	31
1.6.6	100	61	spleen 1.1g, liver 3g	AML	Myeloid leukaemia BM, spleen, liver, lung, kidney	77	13	50.3	264	54	37	33

Mouse	Cell Dose	Lifetime (days) post transplant	Necropsy findings	Blood film	Histopath	WCC	Hb	Hct	Plt	MCV	Lymph/blasts	Gran
6.4H		**220**	**spleen 0.5, liver 2.3**	**Undifferentiated leukaemia**	**Undifferentiated leukaemia**	**34.5**	**13**	**52.6**	**809**	**58**	**25**	**7.2**
1.1	1 million	49	spleen 0.7, liver 4.3g	Undifferentiated leukaemia	Undifferentiated leukaemia spleen, liver,	57.9	9.2	42.5	218	70	41.4	14
1.2	1 million	49	spleen 0.6g, liver 3.4g	Undifferentiated leukaemia	Undifferentiated leukaemia spleen, liver	44.3	9.2	40.6	191	67	30.3	12
1.3	1 million	47	bloody pleural effusion, spleen 0.6g liver 1.7g	Undifferentiated leukaemia	Undifferentiated leukaemia BM, muscle, periosteum, liver, LN, spleen	13	10	38.6	451	63	7.6	4.7
1.4	1 million	39	spleen 0.2g, liver 1.6g, hydrocephalus	not sent	not sent	16.6	12.6	48.8	860	56	12.6	2.8
1.5	500 000	45	spleen 0.4g, liver 2.6g	Undifferentiated leukaemia	Undifferentiated leukaemia BM, muscle, spleen, liver, lung, destructive of bone	68.1	11.4	42.2	528	57	43.4	18.2
1.6	100 000	41	spleen 0.5g, liver 2.4g	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, liver, muscle, kidney	68.1	12.9	48.6	547	59	12.9	16.4
1.7	10 000	77	spleen 0.3g, liver 2g, left leg mass	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, LN, muscle, liver, destructive of bone, invasive ++ (LN->thymus?)	20.9	10.2	38.7	474	58	14.2	5.3
1.8	10 000	80	spleen 0.5g, liver 2.7g	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, liver, muscle, kidney and perirenal, BM patchy replacement but periosteal involvement and into muscle	66.7	12	46.4	233	61	39.7	22
1.9	1000	80	spleen pale, 0.6g, liver 5.2g, bloody ascites	Undifferentiated leukaemia	Undifferentiated leukaemia liver, spleen, BM, kidney, BM surprisingly little invasion	69.4	10.7	41.1	214	71	34.6	27.6
1.10	1000	86										
1.11	100	107	spleen 0.4g, liver 3.4g, mouse 28.9g	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, liver, muscle, kidney +/-	253	9.5	38	234	71	171	63
1.12	100	24	spleen 0.2g, liver 2.3g		no lesion	14.3	11.6	41.5	356	53	4.4	7.8
1.3.1	1 million	36	spleen 0.5g, liver 3.1g, hindlimb paralysis	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, muscle, liver	52.1	11.5	44	522	56	31.6	15.8
1.3.2	1 million	25	spleen 0.6g, liver 2.4g, dragging right hind leg	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, muscle, liver, ovary	89	8.5	31.6	349	61	54	28
1.3.3	1 million	25	spleen 0.8g, liver 2.6g	Undifferentiated leukaemia	Undifferentiated leukaemia BM, spleen, liver, muscle, periosteum,	40.8	6.9	24.3	342	62	22.4	15.4
7.5b		**350**	**spleen 1.2g, liver 2.4g, inguinal LN, splenunculus**	**high WCC with few blasts**	**Myeloid leukaemia with maturation; BM, spleen, LN, liver, muscle**	**114**	**12.3**	**46**	**503**	**59**	**22**	**83**
1.1	1 million	63	spleen 0.5, liver 2.2	Myeloid leukaemia with many blasts and maturation	Acute myeloid with some maturation: BM, spl, ln, liver, kidney, periosteum	446	5.8	23.5	228	82	251	13.8
1.2	1 million	75	spleen 1.1g, liver 3.2g	Myeloid leukaemia with many blasts and maturation	Acute myeloid with some maturation; BM, spleen, liver, kidney, lung, muscle, periosteum	299	5.4	19.7	211	71	95	164
1.2.1	1 million	46	spleen 0.6g, liver 2.1g	Myeloid leukaemia with >90% blasts	Acute myeloid with minimal differentiation; BM, spleen, LN, liver, kidney,	201	6.1	24.4	297	71	171	25
1.2.2	1 million	46	spleen 0.7g, liver 2.g	Myeloid leukaemia, mainly blasts with some maturation	Acute myeloid with minimal differentiation; perianal subcutaneous leukemia, BM, spl, liver , kidney, muscle	129	5.1	21.5	511	70	105	18
1.2.3	1 million	46	spleen 0.7g, liver 2.7g	Myeloid leukaemia with many blasts and maturation	Acute myeloid, virtually undifferentiated; BM, spleen, liver, kLN, kidney, fat, ovary, lung, muscle	438	7.1	26.2	222	74	360	62
1.3	1 million	75	spleen 0.9g, liver 2.2g	Myeloid leukaemia with many maturing cells	Myeloid leukaemia with maturation; BM, spleen, liver, muscle, periosteum, kidney	692	11.9	44.1	375	60	94	526
1.3.1	1 million	28	spleen 0.6g, liver 1.3g	Myeloid leukaemia with many maturing cells	Myeloid leukaemia with maturation; BM, spleen, liver, muscle,	429	12.8	45.5	764	56	60	324
1.3.2	1 million	28	spleen 0.5g, liver 1.4g	Myeloid leukaemia with many blasts and maturation	Acute myeloid with maturation;BM, spleen, liver, muscle, kidney	284	11.6	42.2	791	57	45	207
1.3.3	1 million	32	spleen 0.6g, liver 1.6g	Myeloid leukaemia with many blasts and maturation	Acute myeloid with maturation; BM, spl, liver, periosteum	398	10.6	39.3	661	61	51	306
7.5c		**257**	**spleen pale, 3.7g, liver 3.9g**	**Myeloproliferative with progression, giant platelets, thrombocytosis, leukoerythroblastic**	**Myeloid leukaemia with maturation, megakaryocytes +++ in spleen. Involves spleen, liver, bone marrow but not heart, lung or kidney**	**78.3**	**22.8**	**75**	**4930**	**67**	**40.6**	**34.4**
1.1	1 million	273	culled									
1.2	1 million	273	culled									
1.3	1 million	273	culled									
1.4	1 million	210	distended intestine without tumour, spleen 0.1g, liver 1.4g	thrombocytosis, otherwise normal	megakaryocytes increased in spleen ?MPD, mild bowel inflammation	8	12.1	44.5	>2200	53	2.8	4
1.5	1 million	210	stomach distended 4cm, rest normal, spleen <0.1g, liver 1.7g	MPD with eosinophilia	Myeloid leukaemia with maturation, BM, spleeen, liver, kidney, not muscle, liver, spleen sclerosis and eosinophilia, megakaryocytes increased and abnormal in BM and spleen	5.8	11	40.2	1230	55	2.7	2.6
1.6	1 million	399	pus in bladder/ureter, spleen 0.1g, liver 1.3g	thrombocytosis only	?MPD, Megakaryocytes increased BM and spleen	4.3	14.1	54	1980	56	1.9	1.9
1.7	1 million	175	spleen 0.2g, liver 5.9g	None	myeloid disease							
1.8	1 million	148	spleen <0.1g, liver 1.3g	Low WCC otherwise normal	MPD with eosinophilia, megakaryocytes and eosinophils increase BM, spleen, sclerosis and eosinophilia in lungs and liver	3.8	14.9	54.7	1094	54	1.2	2.3
1.9	1 million	238	ascites and pleural effusion, pale kidneys, spleen <0.1g, liver 1g	not available	sclerotic kidneys with ?immune deposits in glomerull	1.2	5.7	15.7	>2200	42	1	0.2
7.5h		**182**	**spleen 1.5g, liver 4g**	**Myeloid leukaemia with maturation**	**Myeloid leukaemia with maturation; liver, BM, spleen, LN**	**627**	**18.3**	**>70**	**733**	**62**	**146**	**411**
1.1	1 million	163	Culled									
1.2	1 million	99	spleen 0.9g, liver 4.2g	Very poorly differentiated, ?myeloid	Myeloid leukaemia with little differentiation: BM, spl, liver, kidney, lung, thymus	233	14.3	50.8	362	57	71	138
1.3	1 million	176	Culled									
1.4	1 million	330	Culled									
1.5	1 million	304	Culled			96	14.4	44.4	>2200	65	61	22
1.6	1 million	190	Pus filled mass ~1.2cm diameter right inguinal region. Spleen 0.2, liver 2.2		Benign: PMNs in liver, abscess in skin	17.5	11	42.3	1890	51	6.4	8.5
1.2.1	1 million	36	petechial haemorrhages, kidneys pale and abnormal, spleen 0.6g, liver 3.8g	BF+	Acute myeloid leukaemia with little differentiation: adrenal, fat,salivary gland, periosteum, BM, spl, ln, liver, lung, kidney, muscle	111	12	44.5	387	60	37	62
1.2.2	1 million	36	spleen 1.4g, liver 6g, mouse 32.7g, ?LN thorax	Acute leukaemia with very high blast count, very little differentiation	Acute myeloid leukaemia with little differentiation;BM, Spl, LN, liver, kidney, lung, muscle, periosteum	357	14.5	56.3	316	71	115	203
1.2.3	1 million	25	spleen 0.8g, liver2.4g, mouse 30.8g	Acute leukaemia with very high blast count, very little differentiation	Acute myeloid leukaemia with little differentiation; BF, BM, Spleen, Liver, Kidney, Muscle, fat	70.4	14.7	54.6	634	55	21.6	40.8
16.3b		**301**	**Bspleen 0.6, liver 2.7**	**Myeloid leukaemia with few blasts**	**Myeloid leukaemia with maturation; BM, spleen, LN, liver, muscle**	**109**	**15.8**	**59**	**367**	**56**	**54**	**43**
1.1	1 million	52	Spleen 0.5g, liver 1.8g, small right inguinal LN	**Myeloid leukaemia with maturation**	Myeloid leukaemia with maturation; BM, spleen, LN, liver, lungs, periosteum, kidney, muscle	307	14	53.6	409	60	102	156
1.2	1 million	62	spleen 0.5g, liver1.6	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with some maturation; BM, spleen, LN, liver, muscle, kidney	523	15.5	59.7	525	61	128	325
1.1.1	1 million	24	Hindlimb paralysis, spleen 0.3g, liver 1.5g	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with some maturation; BM, spleen, liver, kidney, & probable benign teratoma	40.9	16.5	59.8	571	55	12.4	22.5
1.1.2	10 000	165	found dead, no masses, pale kidneys, spleen 0.1g, liver 2.7g		Normal	there is a bm cytospin						
1.1.3	1000	102	bowel prolapse, no mass, spleen <0.1g, liver 1.7g	Reactive? Left shift of myeloid series without elevated count	Normal	6.2	9	33.2	1466	54	2.2	2.8
1.1.4	100	357	Pale, spleen 0.1g, liver 1.6g		Normal	7.7	13.2	52	1979	55	3.8	2.9
1.2.1	1 million	33	spleen 0.9g, liver 3.9g, enlarged lungs spotted with blood	Myeloid blasts with maturation	Undifferentiated leukaemia, BM, spleen, liver, LN, lung +++, muscle ++, some eos in BM							
1.2.2	10 000	54	Spleen 0.8g, liver 2g	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with some maturation: spleen, BM, muscle, liver, kidney, lung	279	10	36.9	590	57	66	183
1.2.3	1000	132	Shaking, inactive, spleen 0.1g, no lymphadenopathy	?	Some increase in myeloid cells in spleen, not diagnostic	8.6	12	41.8	532	55	1.2	69
1.2.4	100	363	Culled									

Mouse	Cell Dose	Lifetime (days) post transplant	Necropsy findings	Blood film	Histopath	WCC	Hb	Hct	Plt	MCV	Lymph/blasts	Gran
16.3e		**206**	**spleen 0.7g, liver 2.7g**	**Myeloid leukaemia with blasts and maturation**	**Myeloid leukaemia with maturation; BM, spleen, LN, liver, muscle, lung**	**131.2**	**8.8**	**32.6**	**147**	**71**	**35.4**	**78.8**
1.1	1 million	12	Ascites, spl 0.1g, liver1.5g	Normal	no histopath	3.6	13.9	54.7	1196	58	1.4	1.6
1.2	1 million	41	Spleen 0.4g, Liver 1.5g	Myeloid leukaemia with limited maturation	Acute myeloid leukaemia with limited maturation; BM, spl, LN, liver, muscle	193	8.4	32.4	215	59	74	88
1.3	1 million	62	Spleen 0.5g, liver 1.6g	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with maturation; BM, spleen, liver, muscle	47.8	12.7	39	530	63	13.8	29.3
1.3.1	1 million	39	Hindlimb paralysis, spleen 0.3g, liver 1.3g	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with maturation; BM, spleen, muscle, liver	43.5	10.8	42.3	607	57	23.9	14.7
1.3.2	1 million	39	Hindlimb paralysis, spleen 0.3g, liver 1.4g	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with maturation; BM, spleen, muscle, liver	103	9.7	35.2	640	61	51	40
1.3.3	1 million	39	spleen 0.5g, liver 1.4g	Myeloid leukaemia with numerous blasts and maturation	Acute myeloid leukaemia with maturation; BM, spleen, liver, muscle, liver	172	10.1	40	559	60	83	67
16.3f		**387**	**spleen 1g, liver 2.2g, inguinal and axillary LN**	**Myeloid leukaemia with blasts and differentiated cells**	**myeloid leukaemia with probable lymphoma as secondary diagnosis; Spleen, liver, kidney, LN, thymus, BM. Also B cell infiltrate in lung and B and T infiltrate in liver and kidney**	**45.8**	**11.9**	**47.2**	**173**	**59**	**15.1**	**24.5**
1.1	1 million	33	hindlimb paralysis, spleen 0.2g, liver 1.2g	AML or MPD	myeloid leukaemia in spleen, BM and blood	37.5	16.1	57.9	79	52	14.8	16.3
1.2	1 million	46	hydrocephalus, piloerection, spleen 0.3g, liver 1.4g	AML	myeloid leukaemia in blood, spleen, liver, kidney, bone marrow and muscle	566	8.8	35.6	687	69	233	253
1.3	1 million	69	tail mass 2cm x5mm, purpuric and wraps around tail, spleen 0.5g, liver 1.9g	AML	myeloid leukaemia in blood, spleen, liver, LN, bone marrow, muscle, meninges and fat. Probable AML in tail lesion	514	12	45.7	557	61	253	186
1.4	1 million	31	spleen 0.5g, liver 1.7g, excoriated right eye	AMML	myeloid leukaemia in blood, spleen, liver, kidney, BM and muscles and meninges	>80	9.6	37.8	740	62		
1.5	1 million	32	diarrhoea, dilated LI, SI and stomach, no overt mass, spleen 0.2g, liver 1.5g	AML or MPD	myeloid leukaemia in blood, spleen, liver, kidney, LN, stomach, periosteum, meninges, bone marrow and muscle	362	17.9	68	1025	56	121	200
1.6	1 million	369	Culled									
1.7	10 000	292	hindlimb paralysis, piloerect, liver 1.4g, spleen 0.1g	NAD	No significant lesion	11	11.8	41.3	431	52	2.1	7.5
1.8	1000	29	moribund, hindlimb paralysis, blood stained urine, two pale areas in right kidney, spleen<0.1g, liver 1.1g	Reactive changes	Repairing myocardial infartion, extramedullary haematopoiesis, no malignancy	7	12.8	42.9	234	52	1.7	4.5
1.9	1000	369	Culled									
1.10	100	369	Culled									
1.11	100	369	Culled									
1.2.1	1 million	29	spleen 0.4g, liver 1.7g	AML	myeloid leukaemia in blood, spleen, liver, kideny, bone marrow, muscle and periosteal	303	11.2	43.3	849	59	97	162
1.2.2	1 million	26	spleen 0.4g, liver 1.7g	AMML	myeloid leukaemia in blood, spleen, liver, LN, BM and muscles and may be thymus	309	11.7	45.7	787	57	125	142
1.2.3	1 million	22	spleen 0.4g, liver 2g	AMML	myeloid leukaemia in blood, spleen, liver, kidney, LN, bone marrow and muscle and periosteum	415	12.3	46.6	818	61	285	94
1.4.1	1 million	21	partial hindlimb paralysis, spleen 0.5g, liver 1.5g	AMML	myeloid leukaemia in blood, spleen, liver, kidney, skin, lung, bone marrow and muscles and meninges	340	10.5	39.4	642	60	130	164
1.4.2	1 million	21	spleen 0.4g, liver 1.3g	AMML	myeloid leukaemia in blood, spleen, liver and kidney and marrow and mucscle and meninges, great picture	420	13.5	50.9	716	60	209	161
1.4.3	1 million	21	spleen 0.4g, liver 1.8g	AMML	myeloid leukaemia in blood, spleen, liver, kidney, marrow, muscle	377	15.8	61.6	572	60	153	173
16.3G		**231**	**Spleen 0.8g, liver 2.9g, no LN or masses**	**Myeloid leukaemia CMML like**	**Myeloid leukaemia with differentiation; Spleen, liver, LN, thymus, BM**	**167**	**14.5**	**52.8**	**129**	**62**	**48**	**97**
1.1	1 million	52	dilated bowel, no mass, spleen 0.4g, liver 1.5g	CMML like	Myeloid leukaemia in blood, spleen, lung, BM, muscle and bone. CMML in and around bone	504	13.9	52.3	983	59	185	247
1.3	1 million	31	spleen 0.6g, liver 2.4g	AMML	myeloid leukaemia in blood, spleen, liver, BM, muscle and periosteum	545	18.9	>70	750	62	272	201
1.2	1 million	37	spleen 0.5g, liver 1.9g, ltt inguinal LN	AMML	myeloid leukaemia in blood, liver, kidney, spleen, lung, muscle, BM, menignes and periosteum	500	13.3	59.3	953	66	448	39
1.3.1	1 million	18	spleen 0.5g, liver 1.5g, moribund, hindlimb paralysis	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow, muscle	117	15.5	59.5	1199	59	36	64
1.3.2	1 million	19	spleen pale, 0.3g, liver 1.5g, weak rear legs	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow	72.8	15.8	56.7	1008	56	21.7	41.7
1.3.3	1 million	19	spleen 0.3g, liver 1.3g, hindlimb paralysis	AMML	myeloid leukaemia in blood, spleen,	259	16	58.4	1031	56	60	16.4
1.2.1	1 million	21	hindlimb paralysis, spleen 0.5g, liver 1.5g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow and periosteum	162	16.7	61.2	862	55	66	73
1.2.2	1 million	21	spleen 0.5g, liver 1.6g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow, muscle, bone and periosteum	193	15	55.6	747	58	65	100
1.2.3	1 million	21	hindlimb paralysis, pale spleen 0.5g, liver 1.5g	AMML	myeloid leukaemia in blood, spleen, liver, bone marrow	165	14.8	54.6	456	56	68	72
16.3h		**142**	**Spleen 0.53g, liver 1.65g**	**AMML**	**Myeloid leukaemia in blood, spleen, liver, BM**	**58**	**14.1**	**54.3**	**739**	**56**	**12.8**	**41.1**
1.1	1 million	76	Spleen <0.1g and congested ?infarcted, liver 2.3g	NAD	No tumour	3.9	15.3	56.9	1705	54	1.4	2.1
1.2	1 million*	176	culled									
1.3	1 million	176	culled									
1.4	1 million	170	piloerect, spleen <0.1g, liver 1.2g	NAD	No tumour	7.3	22.6	>70	1198	55	3.6	2.6
1.5	1 million	270	piloerect, hunched, NAD		No tumour							
1.6	1 million	256	spl <0.1g, liver 1.5g									
22.2b		**216**	**spleen 1.1g, liver 2.4g**	**Myeloid leukaemia with many blasts and maturation**	**Myeloid leukaemia with blasts;bm, spleen, liver, muscle**	**281**	**9.6**	**36.8**	**165**	**61**	**89**	**149**
1.1	1 million	25	spleen 0.4g, liver 1.9g	Myeloid leukaemia with blasts and maturation, more blasts than 22.2b	Myeloid leukaemia with some maturation; BM, spleen+++, muscle++, liver +, kidney/liver-	668	10.8	43.6	714	63	307	261
1.2	1 million	25	spleen 0.5g, liver 2.6g	Myeloid leukaemia with blasts and maturation, more blasts than 22.2b	Myeloid leukaemia with some maturation; BM, spleen +++, liver, mucle, periosteal ++, kidney, lung-	722	10.1	40	622	64	254	365
1.3	1 million	22	hindlimb paralysis, spleen 0.7g, liver 2.2g	Myeloid leukaemia with maturation	Myeloid leukaemia with maturation; BM, spleen +++, liver ++, kidney +, lung and muscle-, periosteal ++	324	14.8	57.4	531	59	178	101
1.4	1 million	33	spleen 0.6g, liver 2.3g, mesenteric LN	Myeloid leukaemia with maturation	Myeloid leukaemia with maturationBM, spleen, in, liver +++, kidney +, muscle ++, lung -	544	10.1	37.8	462	60	26.6	19.9
1.5	10 000	53	spleen 0.7g	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia with some maturation; spleen+++, liver ++, lung +, kidney?-, no bone or ln or muscle sample, perirenal +++	317	16.3	>70	679	66	282	26
1.6	10 000	33	spleen 0.5g, liver 2.1g	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia with some maturation, BM +++, spleen +++, liver, muscle++, kidney, lung -	623	11.3	43.9	677	63	38.3	15.8
1.7	1000	61	spleen 0.5g, liver 1.8g	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia with some maturation, BM +++, spleen, liver ++, muscle +, kidney, lung -, periosteal ++	171	13.9	63.5	738	67	148	17
1.8	1000	103	antalgic gait favouring right hindlimb, spleen <0.1g, liver2.3g	Benign monocytosis	Inflammation of knee	19.7	11.5	45.9	802	51	7.2	9.8
1.9	100	180	reduced activity, piloerect, lungs speckled ?patchy haemorrhages, otherwise NAD, spleen <0.1g, liver 2.1g	Unremarkable	Florid arthritis in knees	13	17.4	65.6	1109	54	3.1	8.2
1.3.1	1 million	28	partial hindlimb paralysis, spleen 0.3g, liver 1.9g	maturation	BM, spleen +++, liver, muscle ++, kidney	167	11.9	45.6	801	57	53	89
1.3.2	10000	333	Piloerect,swollen abdomen, spleen ,0.1g, liver 2.6g	Unremarkable	No tumour seen	1.3	11.4	42.5	37	55	0.8	0.4
1.3.3	1000	333	piloerect, swollen abdomen, spleen 0.3g, liver 2.4g	Unremarkable	No tumour seen	11.6	12.7	47.1	1593	54	4.7	5.3
1.3.4	100	391	piloerect, pale, immobile, spleen <0.1g, liver 0.5g									

Mouse	Cell Dose	Lifetime (days) post transplant	Necropsy findings	Blood film	Histopath	WCC	Hb	Hct	Plt	MCV	Lymph/blasts	Gran
7.7b		**515**	**spleen blotchy appearance with areas of pallor, 1.2g, liver pale throughout, 2.4g, fat 45.2g mouse**	**Numerous blasts, virtually undifferentiated, no PMNs**	**Myeloid disease, MPO positive; BM, spleen, liver, muscle, papillary adenoma lung**	**32.5**	**10.1**	**35.1**	**145**	**67**	**12.6**	**17.3**
1.3	1 million	312	Spleen 0.4g, liver 1.6g	undifferentiated blasts	Myeloid leukaemia; BM, spleen	54.9	4.5	16.9	93	64	27.4	19.9
1.2	1 million	19	spleen 0.3g, liver 1.3g, rear leg paralysis	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM, spleen, muscle, liver, infarcted bone	28.3	16	57.3	721	56	7.8	16.4
1.1	1 million	30	hindlimb paralysis, spleen 0.3g, liver 1.5g	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM, spleen, liver, muscle, kidney	32.9	17	59.3	979	54	11.1	16.9
1.4	1 million	22	hindlimb paralysis, spleen 0.3g, liver 1.9g	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM and spleen, liver, kidney. BM and spleen blasts only	44.8	16.2	60	657	54	14.7	23.4
1.5	1 million	22	spleen 0.2g, liver 1.3g	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM, spleen (blasts only, no maturing granulocytes), liver, muscle	52.7	14.4	54.3	478	55	21.1	24.3
1.6	1 million	22	spleen 0.4g, liver 1.3g, hindlimb paralysis	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia: spleen (poorly differentiated only). BM, liver, lung	44.3	16.6	62.9	1032	56	12	26
1.2.1	1 million	27	spleen 0.3g, liver 1.5g	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM & spleen blasts only, liver, muscle	52.2	17.2	66	1070	57	15.6	28.8
1.2.2	1 million	29	spleen 0.2g, liver 1.5g	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; spleen, BM, liver. BM and spleen only blasts	35.7	19	>70	1121	57	13.1	17.2
1.2.3	1 million	29	spleen 0.3g, liver1.5g, hydrocephalus	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM & spleen blasts only, liver, kidney, muscle	60	16	59.3	1169	54	22.7	29.2
1.1.1	1 million	23	spleen 0.3g, liver1.3g, full bladder, hindlimb paralysis	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM, spleen	58.7	16.5	62.3	1260	55	22	28.4
1.1.2	1 million	23	spleen 0.4g, liver 1.5g, urinary retention and hindlimb paralysis	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM & spleen blasts only, liver, muscle	44	18.7	>70	1063	55	11.6	26.1
1.1.3	1 million	23	spleen 0.3g, liver 1.3g	Myeloid leukaemia with blasts and differentiation	Myeloid leukaemia; BM & spleen blasts only, liver, kidney, muscle	38.3	16.4	65.3	1392	53	14.6	18
19.2d		**277**	**lymphadenopathy, spleen 1.5g, liver 3g, mouse 28.4g, hydrocephallic, swaying, shallow breathing**	**AMML**	**myeloid leukaemia; blood, spleen, LN, liver, lung, bm, muscle**	**595**	**5.9**	**23.8**	**112**	**101**	**445**	**117**
1.1	1 million	58	found dead, spleen 0.3g, liver 1.6g, thymomegaly	no blood film	Probable leukaemia, autolysed, cannot exclude lymphoma; spleen, liver, kidney, lung, LN, bone marrow and muscle							
1.2	1 million	53	spleen 0.6g, liver 2g	no blood film	presumed myeloid leukaemia; spleen, liver, LN, lung, bone marrow and muscle							
1.3	10 000	61	spleen 0.6g, liver 1.6g	AML	myeloid leukaemia; blood, spleen, kidney, liver, fat, lung, lymphh node, bone marrow and muscle	39.5	4.5	20.8	211	99	31.2	7.1
1.4	10 000	61	spleen 0.4g, liver 1.9g	AML, dysplastic background	myeloid leukaemia; blood, spleen, liver, kidney, BM and muscle	86	6.6	30.8	853	68	81	4
1.5	1000	62	spleen 0.5g, liver 2.1g	AML	myeloid leukaemia; blood, spleen , liver, kidney, LN, bone marrow and muscle. Adenoma of lung	47	7.3	30.5	899	63	27.7	15.8
1.6	1000	61	spleen 0.4g, liver 1.6g	Myeloid leukaemia with clumped platelets.	myeloid leukaemia: spleen, liver, kidney, adrenal gland, lung, marrow and muscle	70.8	5.6	24.9	563	65	56.4	12
1.7	100	7	antaligic gait, favouring left hind limb, no masses,	Normal	Normal	3.8	16.3	58.4	1086	54	1.9	1.6
1.8	100	343	Culled									
21.3j		**176**	**spleen 0.7g, liver 1.7g**	**Myeloid leukaemia with blasts and maturation**	**Myeloid leukaemia; BM, spleen, muscle, liver, kidney**	**92**	**14.1**	**61.7**	**208**	**65**	**67**	**21**
1.1	1 million	30	paraspinal mass left lumbar region 1cm diameter, partial hindlimb paralysis, spleen 0.5g, liver 2.1g	Myeloid leukaemia with maturation	Myeloid leukaemia; BM, spleen, LN, Liver, muscle, lung, kidney	179	17.5	65.2	248	54	38	124
1.2	1 million	26	hindlimb paralysis, spleenn0.2g, liver1.7g, pulmonary haemorrhages	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia; BM, spleen, muscle, liver, kidney	16.1	17.2	62.6	331	52	3.4	11.1
1.3	10 000	33	urinary retention, 7widening spine, spleen 0.4g, liver 1.5g	Myeloid leukaemia with maturation	Myeloid leukaemia; BM, spleen, muscle, liver, kidney	113	15.4	57.2	883	55	37	61
1.4	10 000	343	Culled									
1.5	1000	30	spleen pale, 0.4g, liver 1.5g, abnormal posture hindlimbs	Myeloid leukaemia with maturation	Myeloid leukaemia; BM, spleen, LN, liver, muscle	53.8	14.4	49.1	673	56	16.5	29.8
1.6	1000	26	spleen 0.3g, liver 1.3g, stomach distended	Myeloid leukaemia with maturation	Myeloid leukaemia; BM, splwe +++, muscle, periosteum ++, liver +, kidney, liver -	70	16.1	59.6	745	56	18.1	42.4
1.7	100	25	lesion at base of tail, hunched, spleen pale, 0.3g, liver 1.3g	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia; BM, spleen, liver	29.8	16.8	58.5	405	54	5	21.5
1.8	100	343	Culled									
21.3j1	single colony	42	Small inguinal LN, spleen 0.5g, liver 1.8g	Myeloid leukaemia, predominantly blasts	Acute myeloid leukaemia, spinal cord compression; BM, spleen, muscle, liver, kidney, lung	290	12.1	46.1	766	59	118	135
21.3j2	single colony	39	Moribund, spleen 0.4g, liver 1.3g	Myeloid leukaemia with blasts and maturation	Leukaemia poorly differentiated, favour myeloid; BM, spleen, liver, LN, liver, muscle	500	13.9	52.6	804	62	179	252
21.3j3	single colony	36	Hind limb paralysis, spleen 0.5g, inguinal LN, liver 1.6g	Myeloid leukaemia with blasts and maturation	Leukaemia poorly differentiated, favour myeloid; CSF, spleen, BM, LN, liver, muscle	488	17	68.1	486	58	160	266
21.3j4	single colony	36	Partial hind limb paralysis, spleen 0.2g, liver 1.3g	Myeloid leukaemia with blasts and maturation	AML, spinal cord compression; BM, spleen, muscle, liver, blood,	74.3	15.6	57.3	456	55	26.2	36.9
21.3j5	single colony	87	hunched and thin, spleen 0.4, liver1.4	None	Probable AML; BM, spleen, muscle, liver							
21.3j6	single colony	311	Culled									
21.3j7	single colony	314	Culled									
21.3j8	single colony	209	Culled									
21.3j9	single colony	314	Culled									
21.3j10	single colony	248	Culled									
19.2b		**189**	**Enlarged LN, spleen 1.2g, liver 4g**	**AMML**	**Myeloid Leukaemia in blood, BM, infiltrating muscle along nerve, liver, kidney, spleen and LN.**	**221**	**10.7**	**41.4**	**262**	**79**	**159**	**50**
1.1	1 million	33	spleen 0.4g, liver 2.2g, pale liver and kidneys, brown urine, faecal loading.	AMML (less differentiated than 19.2B)	Myeloid leukaemia in blood, kidney, liver, spleen, BM and Muscle, meninges of cord and peri-osteum. Spleen total replacement.	327	10.8	36.8	240	57	14.9	13.9
1.2	1 million	35	venous system dilated, spleen 0.5g, liver 1.9g	AMML	myeloid leukaemia in blood, spleen, pleura, bone marrow, kidney and liver and periosteum.	210	11.2	38.7	312	50	88	97
1.3	10 000	49	spleen 0.5g, liver 2.1g	AMML less differentiated than 19.2b	myeloid leukaemia in blood, spleen, kidney, liver, BM	133	13.6	62.8	613	66	98	24
1.4	10 000	57	spleen 0.4g, liver 2g	AMML less differentiated than 19.2b	Myeloid leukaemia in spleen, kidney, liver, fat, BM, bladder and pleura	121	12.5	47.8	483	56	30	76
1.5	1000	48	moribund, spleen 0.3g and pale, solid white mass near bladder, large kidney.	AMML less differentiated than 19.2b	Myeloid leukaemia in blood, kidney, spleen, liver, heart, BM. Necrosis in kidney	68	16.6	62.5	887	55	52	1.3
1.6	1000	131	spleen 0.6g, liver 2.4g	Undifferentiated blasts and occasional PMN	Undifferentiated leukaemia; BM, spleen, liver, kidney, lung, muscle	69.4	7.3	28.5	301	65	26.3	34.3
1.7	100	295	Piloerect, spleen 0.1g, liver 2.2g	Normal	No tumour seen	14.9	11.6	41.6	1862	54	4.7	8.1
1.8	100	253	Piloerect, pale and shaky. Pus in fat pad of suprapubic	Normal	Liver adenoma, kidney pyelonephritis	13.9	12.2	48.2	736	58	4.3	8
19.2b1	single colony	202		culled								
19.2b2	single colony	314		culled								
19.2b3	single colony	310		found dead								
19.2b4	single colony	311		culled								
19.2b5	single colony	176		found dead								
19.2b6	single colony	169		found dead								
19.2b7	single colony	178		found dead								
19.2b8	single colony	314		culled								
19.2b9	single colony	314		culled								
19.2b10	single colony	314		culled								

Mouse	Cell Dose	Lifetime (days) post transplant	Necropsy findings	Blood film	Histopath	WCC	Hb	Hct	Plt	MCV	Lymph/blasts	Gran
7.1a		**322**	**Spleen 1.3g liver 3g**	**Megakaryoblastic, low count**	**Megakaryoblastic leukaemia with some myeloid features**							
1.1	1 million	103	swollen abdomen, spl 0.4, liver 1.6, mouse 33.3	Myeloid leukaemia with blasts and some maturation	Myeloid leukaemia without much maturation; spleen, BM, liver, kidney	28.5	3.7	11.8	434	65	12.3	13.6
1.2	1 million	103	swollen abdo, spleen 0.8g, liver 2g, mouse 24.2g	Myeloid leukaemia with blasts and some maturation	Myeloid leukaemia without much maturation; BM, spleen, liver	33	2.1	6.4	376	70	8.5	23.4
1.3	100 000	101	spleen 1.4g, liver 5.3g, gelatinous texture to both	Myeloid leukaemia with maturation	Myeloid leukaemia with maturation; BM, spleen, LN, liver, kidney, lung,renal capsule, small LN, megakaryocytes increased ++	588	7.6	32.4	>2200	69	114	390
1.4	100000	111	spleen pale, gelatinous, 1.4g, liver pale,3.7g	remarkable film, no histo	Myeloid leukaemia with maturation: BM, spleen, liver, LN, lung, megakaryocytes increased +++	131	6.5	27.7	>2200	60	43	77
1.5	10 000	111	spleen 0.4g, liver 2.5g	Myeloid leukaemia with blasts and maturation	Could be megakaryoblastic, megakaryocytes increased ++ but not PMNs, immature blasts; BM, spleen, liver, lung	6.3	10.4	38.4	163	56	2.7	2.9
1.6	10 000	323	Culled									
1.7	1000	323	Culled									
1.8	1000	311	Unwell, moribund, piloerect and immobile. Splenomegaly. Mouse - 31.7, spleen 0.8, liver 2.3 and kidney 0.4	Myeloid leukaemia with some blasts and lots of maturation	AML; BM, spleen, liver, kidney, muscle	685	7.7	31.2	563	95	271	325
1.9	100	169	spleen 0.3g, liver 2.8g, mouse 37.1g, tumour mass in the urogenital system	Left shift only	Myeloid leukaemia with blasts and differentiation; spleen looks malignant but nowhere else, spleen unusual geographic pattern, megakaryocytes increased ++	10.2	8.1	27.4	375	50	1	8.7
1.10	100	132	haemothorax, retroperitoneal haematoma, mesenteric mass, thickening pleural and inner sternum, spleen 0.1g, liver 1.2g	Unremarkable, but haemothorax fluid high WCC with blasts, ?myeloid	Undifferentiated blasts, pericardium, paraspinous, BM, apoptosis +++	4	11.7	43.9	1626	54	2	1.4
1.11	10	214	found dead	PMN+++	autolysed, septic or CML; BM, liver and kidney PMNs ++							
1.12	10	323	Culled									
7.1m		**165**		**Leukocytosis with blasts**	**MPD with blasts and eosinophilia**	**600.00**	**15.00**		**423**	**66**	**400.00**	**150.00**
1.1	1 million	323	piloerect, heavy breathing, spleen 0.1g, liver 1.3g		no malignancy	17.1	12.8	49.7	>2200	52	4.3	10.2
1.2	1 million	323	culled									
1.3	10000	323	culled									
1.4	10000	323	culled									
1.5	1000	323	culled									
1.6	1000	268	Found dead, small bowel obstruction cut off at caecum but no obvious masses. Liver necrotic, spleen <0.1g	No sample	no lesion seen, no gut section							
1.7	100	323	culled									
1.8	100	323	culled									
19.2a		**361**	**Thymomegaly, spleen 0.7g, liver 3.1g**	**Normal**	**B cell lymphoma; BM, spleen, LN, liver, kidney, adenoma of lung**	**361**	**10.9**	**42.5**	**427**	**65**	**249**	**98**
1.1	1 million	13	spleen 0.3g, liver 1.8g, mouse 27.8g	Undifferentiated blasts with no other white cells	Undifferentiated leukaemia BM, spleen, liver, lung, kidney, muscle	119	13.9	62.8	883	65	93	22
1.2	1 million	12	hindlimb paralysis, spleen 0.3g, liver2g	Undifferentiated leukaemia	Undifferentiated blasts BM, spleen, liver, muscle, kidney	117	14.4	56.2	952	57	88	23
1.3	1 million	13	spleen 0.3g, liver 1.5g	Undifferentiated leukaemia	Undifferentiated blasts BM, spleen, LN, Liver, kidney, lung, muscle	79.4	13.3	57.6	866	64	57	18.9
20.2b		**335**	**Spleen 1.1g, Liver 2.4g, Thymomegaly**	**Myeloid leukaemia with blasts and maturation**	**Myeloid leukaemia; spleen, BM, liver, fat**	**101**	**10.9**	**45.7**	**671**	**62**		**26**
1.1	1 million	59	spleen 0.8g, liver 2.2g, cervical, thoracic, axillary, mesenteric and inguinal LN	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia, poorly differentiated, little maturation; BM,spleen, liver, kidney, LN, muscle, salivary gland, lung	351	13.1	48	514	62	85	224
1.2	1 million	59	spleen 1g, liver 2.5g, cervical, thoracic LN	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia, poorly differentiated; BM, spleen, LN, liver, kidney, lung, muscle, saliary gland	190	9.8	34.7	1034	66	56	115
1.3	1 million	50	spleen 0.7g, liver 2.2g	Myeloid leukaemia with blasts and maturation	Myeloid leukaemia with blasts and little maturation; BM, spleen, LN, kidney,muscle, liver, lung, fat, most infiltrate undifferentiated	318	12.3	44.5	1394	63	118	177
1.4	1 million	22	head tilt, spleen 0.2g, liver 1.8g	Myeloid increased	Moderate myeloid infiltrate with maturation; BM, spleen	13.7	12.9	48.9	91	53	4.4	7.7

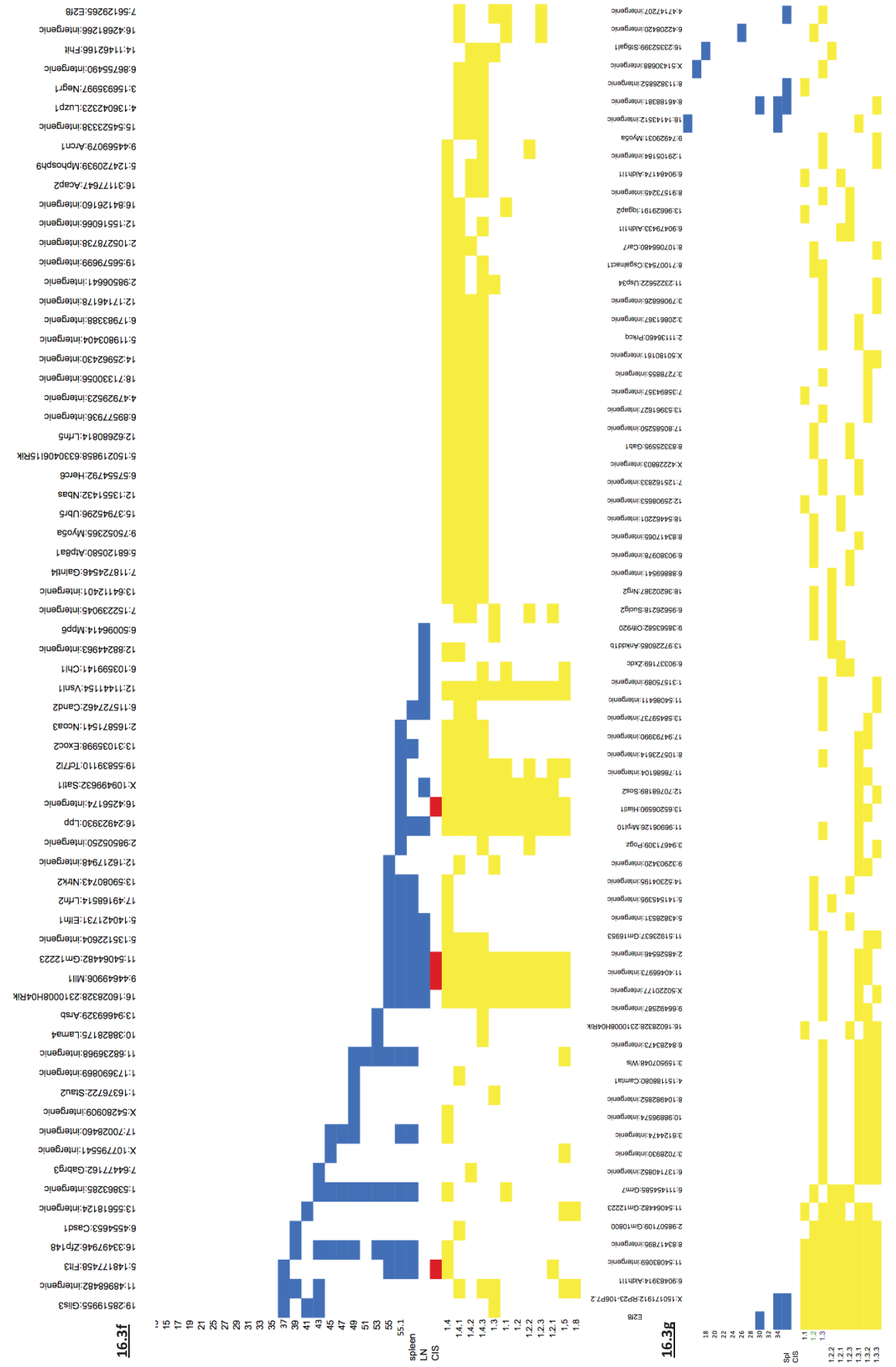


Appendix 4E: Transposon integrations ordered by their presence in multiple transplant recipients

The identity of the mouse is shown at the top left.

6.4h
23
25
27
29
31
31.5
Spleen
CIS
1.1
1.2
1.3
1.4
1.3.1
1.3.2
1.3.3
1.5
1.6
1.7
1.8
1.9
1.11
11:79308122:Nf1
8:46188381:intergenic
5:140025 83:intergenic
2:186006 16:Bmi1
13:16565880:intergenic
11:11592892:Ikzf1
4:44638368:Pax5
5:97376868:intergenic
6:119173242:intergenic
11:100694422:Stat5b
11:100713328:intergenic
7:121331422:intergenic
3:122406862:intergenic
15:78328526:intergenic
11:86115120:46324191 22Rik
2:28986270:Setx
2:22775935:Pdss1
9:64030205:Snapc5
9:32367199:intergenic
7:121507001:49334061 8Rik
6:132144187:intergenic
7:121503694:49334061 8Rik
16:160745 11:23100008H04Rik
6:91068450:intergenic
13:96629191:Iqgap2
7:152239045:intergenic
X:74467 36:Otud5
X:74387 97:Otud5
16:87228 95:Usp7
11:58516877:intergenic
2:98506673:intergenic
6:50096414:Mpp6
9:11960173 4:Scn10a
17:46148074:intergenic

16.3e
18:14143512:intergenic
3:110425087:intergenic
11:24816916:RP23-340O10.1
13:102459811:Pik3r1
1:80623054:Dock10
4:44688758:Pax5
15:34387756:Ghr
10:88230868:Utp20
4:59655757:E130308A19Rik
5:97376868:intergenic
7:121331422:intergenic
11:89772317:intergenic
11:54064951:Gm12223
11:23208833:Usp34
8:46188381:intergenic
7:56129265:E2f8
10:88249266:Utp20
8:10863347:intergenic
10:87926865:Chpt1
7:82383877:Sv2b
2:72791041:Sp3
7:130135392:Rbbp6
12:70713297:Sos2
15:67362743:intergenic
6:117923073:intergenic
10:82486836:Chst11
14:79028918:Dgkh
11:51580022:Sar1b
13:40854976:intergenic
5:144341940:intergenic
1:71903405:intergenic
11:23211927:Usp34
4:59007691:intergenic
15:19577195:intergenic
12:29906310:intergenic
7:35894357:intergenic
13
15
17
19
21
23
25
27
29
29.5
Spleen
CIS
1.1
1.2
1.3
1.3.1
1.3.2
1.3.3


16.3f
16.3g
spleen
LN
CIS
Sp
CIS

19.2b
17:47909063:intergenic
12:95369042:intergenic
7:109301163:Nup98
10:121879054:intergenic
11:79372116:Nf1
7:109360024:Pgap2
1:88580011:intergenic
5:63112888:Arap2
11:79264119:Nf1
X:67592718:intergenic
X:148694473:Kdm5c
9:21794158:intergenic
11:79165100:Nf1
11:79330648:Nf1
3:102861353:Csde1
14:120987954:intergenic
11:79254909:Nf1
8:109597411:intergenic
11:53978006:intergenic
2:145234228:intergenic
2:181375665:Prpf6
10:111193180:intergenic
1:172440847:Nos1ap
19:17422464:Gcnt1
7:96730138:intergenic
5:138541049:Zksca n1
7:88193819:intergenic
3:48469101:intergenic
19:20764557:intergenic
15:84164008:Parvg
10
12
14
16
18
20
22
24
26
27 (bl and spl)
CIS
1.3
1.1
1.5

22.2b
21.3j
10
12
14
16
18
20
22
24
25
spl
CIS
1.1
1.3
1.5
1.6
1.7
11:54066392:Gm12223
11:54064482:Gm12223
7:35894357:Intergenic
13:117975559:Intergenic
3:148069 93:Intergenic
3:85020283:Arfip1
9:44650105:Mll1
3:150736 47:Intergenic
9:120815136:Intergenic
7:78955439:Intergenic
9:18198476:Naaladl2
X:108749515:Hdx
9:21794157:Intergenic
5:90684183:Ankrd17
4:32493166:Bach2
X:108749297:Hdx
5:64695168:Tbc1d1
9:18193930:Naaladl2
9:44533618:Phldb1
9:93758 62:Intergenic
3:151511 19:Intergenic
X:108749071:Hdx

Appendix 4F: CIS results for the analysis on serial blood sample

CIS which were found in the tumour analysis of the whole cohort are shown in red. CIS which were excluded in that analysis are shown in blue. The number of samples, and the time at which the blood samples were taken relative to the onset of leukaemia are shown for each table.

Final Tumour Number samples in analysis = 15					
Gene nearest peak	**Genes in CIS**	**Chr**	**Start**	**End**	**Kernel size (kb)**
A330023F24Rik	A330023F24Rik	1	196821188	196835909	30
Dnajc17	Gchfr Dnajc17 Gm14137 Gm14138	2	119002990	119018636	10,30
Ndufc1	Elf2 4930583H14Rik Ndufc1 Naa15	3	51127253	51254938	60, 100
Cpsf3l	Aurkaip1 Gm10562 Mxra8 Dvl1 Tas1r3 Gltpd1 Cpsf3l	4	155205032	155254034	100
Kdm3a	Rnf103 Vps24 Kdm3a	6	71429134	71586550	10, 30, 60, 100
Gm13834	Gm13834 Gm13833 Gm13835 AB041803	6	31032284	31120830	10, 30, 60, 100
Nrf1	Nrf1	6	30079561	30083481	10, 30
Tet3	Tet3	6	83356893	83359834	10
9330179D12Rik	9330179D12Rik	6	127134776	127176029	60
Nup98	Art1 Chrna10 Nup98 Pgap2	7	109251626	109386388	10, 30, 60, 100
Copb1	intergenic	7	121321414	121327261	10
Il27	intergenic	7	133739132	133742056	10
Mtnr1a	intergenic	8	46184160	46190938	10
Mll1	Mll1	9	44646167	44651018	10
Sik3	Sik3	9	45808700	45832010	10, 30, 60
Gm12223	4933405E24Rik Gm12222 Csf2 Gm12223 Il3 Acsl6	11	53987119	54145289	10, 30, 60, 100
Nf1	Gm9964 Nf1 Gm11198 Gm11199 AU040972 Omg Evi2b Evi2a Rab11fip4 U6 Gm11202 U6 U6	11	79092975	79454961	10, 30, 60, 100
Sfi1	Pisd-ps1 Sfi1 Gm11399 Gm11400	11	3026257	3078900	10, 30, 60, 100
1700012B15Rik	1700012B15Rik	12	3102877	3239602	100
Nedd9	Nedd9	13	41556525	41564300	10
Foxf2	Foxf2	13	31712954	31722673	10
Abcc1	Abcc1 U6	16	14372427	14395263	10, 30, 60
Gtf2h5	Synj2 Serac1 Gtf2h5 Tulp4 Gm5812	17	6005254	6129776	100
Gm3395	intergenic	Y	2825028	2911516	10, 30, 60, 100
Zfy2	intergenic	Y	1307754	1317818	10, 30

24-33 days pre-tumour Number samples in analysis = 15					
Gene nearest peak	**Genes in CIS**	**Chr**	**Start**	**End**	**Kernel size (kb)**
Hecw2	Hecw2	1	53868275	53880081	30
Dock10	Dock10	1	80659185	80700544	60, 100
Commd3	intergenic	2	18576689	18582557	30
Pax5	Pax5 Gm12462	4	44681807	44692561	10
En2	En2	5	28486080	28504626	10, 30, 60
U2	intergenic	5	97371820	97379629	10
Mpp6	Mpp6	6	50087883	50102556	10, 30
Zfp800	Zfp800	6	28322835	28326747	10
Tjp1	Tjp1	7	72421068	72548749	30, 60, 100
Nup98	Nup98	7	109281340	109337165	30, 60, 100
Sntb2	Sntb2	8	109454987	109465693	10, 30
Mtnr1a	Mtnr1a	8	46164323	46193550	10, 30, 60
Cand1	Cand1	10	118654797	118661578	10
Chst11	Chst11	10	82420512	82488615	100
Sfi1	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1	11	2986289	3144475	10, 30, 60, 100
Gm12223	Csf2 Gm12223	11	54059597	54067389	10
Nf1	Nf1 Gm11198 Gm11199 AU040972 Omg Evi2b Evi2a	11	79157294	79382335	60, 100
Foxf2	Foxf2	13	31714003	31722694	10

51-61 days pre-tumour Number samples in analysis = 14					
Gene nearest peak	**Genes in CIS**	**Chr**	**Start**	**End**	**Kernel size (kb)**
Lrba	Lrba Gm3788	3	86446015	86514337	100
Bach2	Bach2	4	32458140	32481575	30, 60
En2	En2 Cnpy1	5	28414611	28572578	10, 30, 60, 100
9330179D12Rik	9330179D12Rik	6	127120888	127170674	30, 60, 100
Sfi1	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1	11	2986176	3161941	10, 30, 60, 100
Gphn	Gphn	12	79709137	79749966	60, 100
Foxf2	Foxf2	13	31712297	31723901	10, 30
6720401G13Rik	6720401G13Rik	X	47919483	47924376	10, 30
Gm6026	intergenic	Y	1623173	1687462	30, 60

79-88 days pre-tumour Number samples in analysis = 11					
Gene nearest peak	**Genes in CIS**	**Chr**	**Start**	**End**	**Kernel size (kb)**
En2	En2	5	28482111	28508451	10, 30, 60, 100
U3	intergenic	7	71118825	71157003	30, 60, 100
Nrbf2	intergenic	10	66783554	66905150	100
Sfi1	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1	11	3011148	3088662	10, 30, 60, 100
Gm3395	intergenic	Y	2873260	2881485	10, 30, 60, 100

98-113 days pre-tumour Number samples in analysis = 11					
Gene nearest peak	**Genes in CIS**	**Chr**	**Start**	**End**	**Kernel size (kb)**
En2	En2	5	28478970	28512057	10, 30, 60, 100
Sfi1	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1	11	3020557	3123527	10, 30, 60, 100
Gm3395	intergenic	Y	2882416	2903328	10, 30, 60, 100

Appendix 4G: Results of the splinkerette analysis out of the blunt end of PB.

Integrations which shared an insertion site in the SB analysis are highlighted in red

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
10.1D (transposase neg)	HS0580M02C05T0	8	71584302	36 - 57 (22)	100.00%	50	Slc18a1	71584281-71584302	+	-
16.3B	HS0580M02EQ9X4	1	26092751	36 - 69 (34)	100.00%	42	Intergenic	26092751-26092784	-	+
16.3B	HS0580M02D8GX6	5	28496561	36 - 289 (254)	99.61%	1	En2	28496561-28496813	-	+
16.3B	HS0580M02D8809	6	118402713	36 - 51 (16)	100.00%	12	Zfp248	118402698-118402713	+	-
16.3B	HS0580M02DIBGH	9	59783750	36 - 61 (26)	100.00%	222	Intergenic	59783725-59783750	+	-
16.3B	HS0580M02EXZ4D	13	117356737	36 - 51 (16)	100.00%	2	Intergenic	117356722-117356737	+	-
16.3B	HS0580M02DD7BB	15	47496863	37 - 66 (30)	96.67%	37	Csmd3	47496834-47496863	+	-
16.3B	HS0580M02DRY10	16	7330576	39 - 128 (90)	98.89%	25	Rbfox1	7330488-7330576	+	-
16.3B	HS0580M02EU7SS	16	16754611	36 - 79 (44)	100.00%	114	Spag6	16754568-16754611	+	-
16.3B	HS0580M02D1J9T	16	21101741	37 - 100 (64)	100.00%	42	Intergenic	21101741-21101804	-	+
16.3B	HS0580M02DGKN7	16	78764476	39 - 126 (88)	100.00%	4	Intergenic	78764476-78764563	-	+
16.3B	HS0580M02EXM45	17	4671285	39 - 263 (225)	99.56%	1	Intergenic	4671285-4671509	-	+
16.3B	HS0580M02DQSIK	X	134134562	39 - 58 (20)	100.00%	8	Il1rapl2	134134562-134134581	-	+
16.3D (cre neg)	HS0580M02ED2OY	1	75813529	39 - 62 (24)	100.00%	39	Intergenic	75813506-75813529	+	-
16.3D (cre neg)	HS0580M02DTSES	1	156588715	39 - 57 (19)	100.00%	1	Intergenic	156588697-156588715	+	-
16.3D (cre neg)	HS0580M02DGJDQ	5	50570301	39 - 64 (26)	100.00%	11	Intergenic	50570301-50570326	-	+
16.3D (cre neg)	HS0580M02CZV8D	7	129061296	36 - 56 (21)	100.00%	36	Scnn1b	129061276-129061296	+	-
16.3D (cre neg)	HS0580M02DT80F	9	52999816	36 - 58 (23)	100.00%	26	Ddx10	52999816-52999838	-	+
16.3D (cre neg)	HS0580M02C2HCN	10	108585724	36 - 66 (31)	100.00%	35	Intergenic	108585724-108585754	-	+
16.3D (cre neg)	HS0580M02C82CD	12	95013247	36 - 69 (34)	100.00%	36	Intergenic	95013247-95013280	-	+
16.3D (cre neg)	HS0580M02EI5WI	13	42714334	38 - 108 (71)	98.59%	2	Intergenic	42714334-42714403	-	+
16.3D (cre neg)	HS0580M02EEYJU	13	113569285	34 - 57 (24)	100.00%	8	Intergenic	113569285-113569308	-	+
16.3D (cre neg)	HS0580M02C241F	14	83946367	40 - 56 (17)	100.00%	6	Intergenic	83946351-83946367	+	-
16.3D (cre neg)	HS0580M02DTJL7	15	6203029	40 - 129 (90)	100.00%	1	Intergenic	6202940-6203029	+	-
16.3D (cre neg)	HS0580M02D45JX	19	17873784	36 - 54 (19)	100.00%	71	Pcsk5	17873784-17873802	-	+
16.3E	HS0580M02DUOWA	1	134034978	38 - 69 (32)	100.00%	2	Cdk18	134034947-134034978	+	-
16.3E	HS0580M02EJDJU	1	140134035	37 - 137 (101)	100.00%	15	Intergenic	140133935-140134035	+	-
16.3E	HS0580M02EUU5M	2	30082528	40 - 140 (101)	100.00%	3	Intergenic	30082428-30082528	+	-
16.3E	HS0580M02ELTGU	2	163468418	36 - 75 (40)	92.50%	32	Serinc3	163468418-163468457	-	+
16.3E	HS0580M02D507U	3	31655677	39 - 122 (84)	97.62%	2	Intergenic	31655593-31655677	+	-
16.3E	HS0580M02DO88K	5	85253057	38 - 184 (147)	99.32%	4	Intergenic	85252912-85253057	+	-
16.3E	HS0580M02DCHJG	6	34699904	34 - 63 (30)	100.00%	1	Cald1	34699904-34699933	-	+
16.3E	HS0580M02C4AJN	7	112858521	40 - 59 (20)	95.00%	8	Intergenic	112858521-112858540	-	+
16.3E	HS0580M02EOAW0	7	122827036	36 - 71 (36)	100.00%	29	Sox6	122827036-122827071	-	+
16.3E	HS0580M02DUBTB	8	30054315	39 - 128 (90)	100.00%	1	Unc5d	30054226-30054315	+	-
16.3E	HS0580M02EDTMJ	11	27303463	38 - 165 (128)	100.00%	68	Intergenic	27303336-27303463	+	-
16.3E	HS0580M02DPPIM	13	105321414	39 - 201 (163)	98.77%	407	Intergenic	105321414-105321577	-	+
16.3E	HS0580M02DRCUB	14	111547518	39 - 77 (39)	100.00%	8	Intergenic	111547518-111547556	-	+
16.3E	HS0580M02DCT36	15	7303833	38 - 125 (88)	100.00%	40	Egflam	7303833-7303920	-	+
16.3E	HS0580M02DCOKS	16	8984646	38 - 69 (32)	100.00%	203	Intergenic	8984615-8984646	+	-
16.3E	HS0580M02DTLGK	16	19094622	39 - 182 (144)	98.61%	12	Intergenic	19094480-19094622	+	-
16.3E	HS0580M02DH791	16	19110639	38 - 107 (70)	98.57%	47	Intergenic	19110639-19110708	-	+
16.3E	HS0580M02D051H	16	23827477	38 - 231 (194)	99.48%	1	Intergenic	23827477-23827671	-	+
16.3E	HS0580M02C3NRW	16	41664747	39 - 90 (52)	100.00%	2	Lsamp	41664696-41664747	+	-
16.3E	HS0580M02DDL97	16	43352874	33 - 59 (27)	100.00%	313	Zbtb20	43352874-43352900	-	+
16.3E	HS0580M02EBYC4	16	82519045	38 - 80 (43)	97.67%	114	Intergenic	82519003-82519045	+	-
16.3E	HS0580M02ESKOK	16	89105941	37 - 201 (165)	99.39%	6	Intergenic	89105941-89106106	-	+
16.3E	HS0580M02DDM2J	18	14665047	35 - 138 (104)	99.04%	16	Intergenic	14664943-14665047	+	-
16.3E	HS0580M02C63KR	19	52617444	39 - 177 (139)	100.00%	5	Intergenic	52617306-52617444	+	-
16.3F	HS0580M02ESNUL	17	22090796	37 - 249 (213)	99.53%	17	Zfp942	22090583-22090796	+	-

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
16.3G	HS0580M02D10NT	1	8739687	39 - 97 (59)	100.00%	27	Sntg1	8739687-8739745	-	+
16.3G	HS0580M02EHHTT	1	69154365	37 - 56 (20)	100.00%	148	Erbb4	69154346-69154365	+	-
16.3G	HS0580M02D47D2	2	5955067	37 - 75 (39)	100.00%	3	Upf2	5955029-5955067	+	-
16.3G	HS0580M02DX5W7	2	19708818	39 - 151 (113)	100.00%	2	Intergenic	19708818-19708930	-	+
16.3G	HS0580M02EVCR5	2	21788192	36 - 223 (188)	97.87%	502	Intergenic	21788006-21788192	+	-
16.3G	HS0580M02E0O36	2	95416344	132 - 223 (92)	100.00%	1	Intergenic	95416344-95416435	-	+
16.3G	HS0580M02DVMYV	4	5950694	39 - 209 (171)	100.00%	3	Fam110b	5950524-5950694	+	-
16.3G	HS0580M02D93LG	5	79264814	36 - 63 (28)	100.00%	2	Intergenic	79264787-79264814	+	-
16.3G	HS0580M02EMUOA	5	152179031	32 - 53 (22)	100.00%	92	1700028E10Rik	152179010-152179031	+	-
16.3G	HS0580M02DWDTN	7	13854052	38 - 199 (162)	99.38%	3	6330408A02Rik	13853891-13854052	+	-
16.3G	HS0580M02EW60U	7	55294174	39 - 164 (126)	100.00%	12	Intergenic	55294049-55294174	+	-
16.3G	HS0580M02DQIL3	7	79092022	39 - 152 (114)	98.25%	8	Intergenic	79092022-79092137	-	+
16.3G	HS0580M02DK5P6	8	109813511	40 - 75 (36)	100.00%	3	Intergenic	109813511-109813546	-	+
16.3G	HS0580M02ENWVV	12	88796809	39 - 204 (166)	99.40%	2	Snw1	88796809-88796974	-	+
16.3G	HS0580M02DTY8V	14	120714936	39 - 100 (62)	98.39%	18	Mbnl2	120714936-120714997	-	+
16.3G	HS0580M02DK5E0	16	19627197	39 - 242 (204)	100.00%	1	Intergenic	19626994-19627197	+	-
16.3G	HS0580M02DW96A	16	25110430	37 - 112 (76)	98.68%	13	Intergenic	25110354-25110430	+	-
16.3G	HS0580M02EKIU0	16	25390065	37 - 88 (52)	98.08%	68	Tprg	25390014-25390065	+	-
16.3G	HS0580M02EVDQH	16	40199206	40 - 279 (240)	99.58%	1	Intergenic	40198966-40199206	+	-
16.3G	HS0580M02EGNET	16	47432829	39 - 230 (192)	98.96%	2	Intergenic	47432636-47432829	+	-
16.3G	HS0580M02DR96J	16	54576055	39 - 235 (197)	97.97%	1	Intergenic	54576055-54576248	-	+
16.3G	HS0580M02EQO0Q	16	56686612	36 - 195 (160)	100.00%	3	Abi3bp	56686453-56686612	+	-
16.3G	HS0580M02EHEIR	16	60850001	39 - 104 (66)	100.00%	36	Intergenic	60849936-60850001	+	-
16.3G	HS0580M02ETFYL	16	65372390	38 - 229 (192)	98.96%	4	Intergenic	65372390-65372582	-	+
16.3G	HS0580M02C29KB	16	75281053	39 - 154 (116)	100.00%	17	Intergenic	75281053-75281168	-	+
16.3G	HS0580M02C7WH3	16	83410954	35 - 69 (35)	97.14%	103	Intergenic	83410921-83410954	+	-
16.3G	HS0580M02DUNZ8	17	50860488	36 - 75 (40)	97.50%	79	Intergenic	50860449-50860488	+	-
16.3G	HS0580M02DIYUA	17	52031231	38 - 63 (26)	100.00%	27	Intergenic	52031206-52031231	+	-
16.3G	HS0580M02EZ4LJ	19	30553273	39 - 125 (87)	100.00%	3	Intergenic	30553187-30553273	+	-
16.3G	HS0580M02C7AFO	X	5587469	38 - 60 (23)	100.00%	97	Intergenic	5587469-5587491	-	+
16.3H	HS0580M02DNUIL	1	32570577	38 - 88 (51)	100.00%	1	Khdrbs2	32570527-32570577	+	-
16.3H	HS0580M02DQPUA	1	173570213	39 - 54 (16)	100.00%	5	Slamf7	173570198-173570213	+	-
16.3H	HS0580M02DVN0K	2	3968183	38 - 99 (62)	98.39%	161	Frmd4a	3968183-3968244	-	+
16.3H	HS0580M02C84RT	2	16720779	36 - 54 (19)	100.00%	5	Intergenic	16720761-16720779	+	-
16.3H	HS0580M02EO9KS	2	94591245	39 - 69 (31)	100.00%	3	Intergenic	94591245-94591275	-	+
16.3H	HS0580M02EX7E6	2	163468418	39 - 78 (40)	92.50%	112	Serinc3	163468418-163468457	-	+
16.3H	HS0580M02DUSOP	4	51709280	36 - 173 (138)	99.28%	8	Intergenic	51709280-51709417	-	+
16.3H	HS0580M02D8G9W	5	130707211	39 - 198 (160)	100.00%	7	0610007L01Rik	130707052-130707211	+	-
16.3H	HS0580M02EUCT2	6	36069436	37 - 60 (24)	100.00%	313	Intergenic	36069413-36069436	+	-
16.3H	HS0580M02ESYT1	7	32329601	39 - 158 (120)	98.33%	1	Intergenic	32329601-32329718	-	+
16.3H	HS0580M02EM1LG	7	109858915	38 - 116 (79)	98.73%	7	Olfr558	109858837-109858915	+	-
16.3H	HS0580M02EOMSH	8	19718602	38 - 88 (51)	100.00%	70	Intergenic	19718602-19718652	-	+
16.3H	HS0580M02C0TJP	8	28601941	38 - 88 (51)	100.00%	1	Poteg	28601941-28601991	-	+
16.3H	HS0580M02EO0XQ	9	73847250	39 - 205 (167)	100.00%	1	Intergenic	73847084-73847250	+	-
16.3H	HS0580M02DP0Q8	11	3067893	38 - 173 (136)	98.53%	13	Sfi1	3067757-3067893	+	-
16.3H	HS0580M02EKWZX	12	116074732	39 - 155 (117)	95.73%	14	Intergenic	116074732-116074848	-	+
16.3H	HS0580M02C3LSL	13	62232047	39 - 220 (182)	99.45%	12	Zfp808	62232047-62232227	-	+
16.3H	HS0580M02EJ47H	15	70370066	39 - 200 (162)	98.77%	2	Intergenic	70370066-70370228	-	+
16.3H	HS0580M02C8UUN	16	3586910	38 - 178 (141)	99.29%	17	Intergenic	3586910-3587050	-	+
16.3H	HS0580M02C5KLD	16	39765263	37 - 57 (21)	100.00%	8	Intergenic	39765263-39765283	-	+
16.3H	HS0580M02DWOKY	16	61860006	40 - 170 (131)	100.00%	1493	Intergenic	61859876-61860006	+	-
16.3H	HS0580M02ET2IT	16	65028054	39 - 234 (196)	98.98%	8	Htr1f	65027859-65028054	+	-
16.3H	HS0580M02D66M7	16	82225648	39 - 64 (26)	96.15%	798	Intergenic	82225623-82225648	+	-
16.3H	HS0580M02EUMFP	17	11767955	35 - 56 (22)	100.00%	1	Park2	11767955-11767976	-	+
16.3H	HS0580M02DTDJR	X	36767695	37 - 214 (178)	100.00%	38	Intergenic	36767518-36767695	+	-
16.3H	HS0580M02ESAX1	X	162471395	37 - 88 (52)	100.00%	2	Intergenic	162471395-162471446	-	+

16.3I (cre neg)

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
19.2A	HS0580M02EI3BV	1	147328949	36 - 178 (143)	98.60%	2	Intergenic	147328806-147328949	+	-
19.2A	HS0580M02EJILG	2	163468410	39 - 86 (48)	89.58%	1	Serinc3	163468410-163468457	-	+
19.2A	HS0580M02EK0CU	2	163468418	39 - 78 (40)	92.50%	19	Serinc3	163468418-163468457	-	+
19.2A	HS0580M02DRYUX	2	173532393	36 - 83 (48)	100.00%	3	Rab22a	173532346-173532393	+	-
19.2A	HS0580M02DI951	3	56227096	38 - 101 (64)	95.31%	1	Intergenic	56227030-56227096	+	-
19.2A	HS0580M02DL9DB	3	128515804	38 - 101 (64)	100.00%	2	Intergenic	128515804-128515867	-	+
19.2A	HS0580M02DC7ZT	3	151561995	37 - 89 (53)	100.00%	2	Intergenic	151561995-151562047	-	+
19.2A	HS0580M02D75D8	4	109623985	39 - 65 (27)	96.30%	43	Faf1	109623985-109624011	-	+
19.2A	HS0580M02DHYRJ	5	50955421	37 - 58 (22)	100.00%	5	Intergenic	50955421-50955442	-	+
19.2A	HS0580M02DOC03	8	17238476	39 - 163 (125)	100.00%	4	Csmd1	17238352-17238476	+	-
19.2A	HS0580M02DOZDJ	8	45534346	38 - 144 (107)	100.00%	8	Intergenic	45534240-45534346	+	-
19.2A	HS0580M02D3Z9D	10	14450043	37 - 81 (45)	100.00%	5	Intergenic	14449999-14450043	+	-
19.2A	HS0580M02C0WWC	10	27108953	37 - 110 (74)	97.30%	14	Lama2	27108880-27108953	+	-
19.2A	HS0580M02EU8LX	10	71642681	37 - 96 (60)	100.00%	11	Intergenic	71642681-71642740	-	+
19.2A	HS0580M02DKFM2	11	15356413	39 - 135 (97)	100.00%	2	Intergenic	15356317-15356413	+	-
19.2A	HS0580M02EEHZG	11	49339650	39 - 104 (66)	100.00%	4	Intergenic	49339585-49339650	+	-
19.2A	HS0580M02E0GQE	13	42714334	38 - 107 (70)	100.00%	3615	Intergenic	42714334-42714403	-	+
19.2A	HS0580M02EOE5B	14	10329333	39 - 114 (76)	100.00%	15	Intergenic	10329333-10329408	-	+
19.2A	HS0580M02EQ5N8	14	40662076	38 - 143 (106)	100.00%	2	Intergenic	40661971-40662076	+	-
19.2A	HS0580M02DLP0V	14	67436099	38 - 161 (124)	100.00%	3	Dpysl2	67436099-67436222	-	+
19.2A	HS0580M02DAGD0	15	33840224	36 - 62 (27)	100.00%	1	Intergenic	33840224-33840250	-	+
19.2A	HS0580M02D2CLI	16	6526810	38 - 258 (221)	98.64%	1	Intergenic	6526590-6526810	+	-
19.2A	HS0580M02EHHYV	16	6549591	39 - 100 (62)	100.00%	2	Intergenic	6549591-6549652	-	+
19.2A	HS0580M02C8NOY	16	9590199	39 - 58 (20)	100.00%	25	Grin2a	9590180-9590199	+	-
19.2A	HS0580M02D0DIC	16	19665188	39 - 100 (62)	98.39%	64	Lamp3	19665127-19665188	+	-
19.2A	HS0580M02DA6FC	16	26554577	40 - 84 (45)	100.00%	26	Intergenic	26554577-26554621	-	+
19.2A	HS0580M02DD0G7	16	28072402	37 - 81 (45)	100.00%	19	Intergenic	28072358-28072402	+	-
19.2A	HS0580M02CZI99	16	38645261	38 - 58 (21)	100.00%	3	Arhgap31	38645261-38645281	-	+
19.2A	HS0580M02DZPA2	16	39182854	39 - 67 (29)	100.00%	1	Intergenic	39182854-39182882	-	+
19.2A	HS0580M02C73CR	16	43333520	39 - 114 (76)	100.00%	1	Zbtb20	43333520-43333595	-	+
19.2A	HS0580M02DKD09	16	49683571	36 - 59 (24)	100.00%	53	Intergenic	49683571-49683594	-	+
19.2A	HS0580M02EURPO	16	54014611	39 - 88 (50)	100.00%	4	Intergenic	54014611-54014660	-	+
19.2A	HS0580M02D3MW8	16	54797206	39 - 117 (79)	100.00%	2	Intergenic	54797128-54797206	+	-
19.2A	HS0580M02C61EA	16	59572139	40 - 105 (66)	95.45%	1	AC154473.2	59572139-59572205	-	+
19.2A	HS0580M02DDK9I	16	66515349	39 - 90 (52)	100.00%	11	Intergenic	66515298-66515349	+	-
19.2A	HS0580M02DUYRM	16	75476168	38 - 56 (19)	100.00%	5	Intergenic	75476168-75476186	-	+
19.2A	HS0580M02DVEA0	16	75598126	35 - 52 (18)	100.00%	2	Rbm11	75598109-75598126	+	-
19.2A	HS0580M02DH612	17	10392822	37 - 89 (53)	98.11%	31	Intergenic	10392771-10392822	+	-
19.2A	HS0580M02C1NBG	17	17058324	38 - 53 (16)	100.00%	1	Intergenic	17058324-17058339	-	+
19.2A	HS0580M02C1J0M	17	59927151	36 - 192 (157)	98.73%	1	Intergenic	59926995-59927151	+	-
19.2A	HS0580M02D0DB4	18	46451797	38 - 221 (184)	99.46%	1	Ccdc112	46451615-46451797	+	-
19.2A	HS0580M02DIXEH	19	30377479	39 - 98 (60)	98.33%	36	Intergenic	30377420-30377479	+	-
19.2B	HS0580M02ESWT7	1	42759589	40 - 56 (17)	100.00%	1	Intergenic	42759573-42759589	+	-
19.2B	HS0580M02DWLVC	2	82815045	38 - 54 (17)	100.00%	1	Fsip2	82815045-82815061	-	+
19.2B	HS0580M02C2OQM	2	98502693	37 - 223 (187)	98.40%	20	Gm10801	98502693-98502879	-	+
19.2B	HS0580M02D3DW2	2	179808696	38 - 76 (39)	97.44%	73	Gtpbp5	179808696-179808734	-	+
19.2B	HS0580M02C6VKP	4	37214216	38 - 93 (56)	100.00%	32	Gm12374	37214161-37214216	+	-
19.2B	HS0580M02DMZQ3	5	17319892	38 - 136 (99)	100.00%	15	Cd36	17319892-17319990	-	+
19.2B	HS0580M02DA3DX	5	106008974	38 - 52 (15)	100.00%	1	Lrrc8c	106008974-106008988	-	+
19.2B	HS0580M02D6HC8	6	131259374	38 - 272 (235)	99.15%	1	Styk1	131259138-131259374	+	-
19.2B	HS0580M02DBJGB	6	149004591	38 - 55 (18)	100.00%	970	Dennd5b	149004574-149004591	+	-
19.2B	HS0580M02C71LG	8	40743439	131 - 234 (104)	99.04%	10	Intergenic	40743439-40743542	-	+
19.2B	HS0580M02EN7B4	9	76605868	39 - 57 (19)	100.00%	12	Intergenic	76605850-76605868	+	-
19.2B	HS0580M02ELW1C	10	44369476	39 - 86 (48)	100.00%	16	Intergenic	44369476-44369523	-	+
19.2B	HS0580M02E01ZL	10	85827031	41 - 55 (15)	100.00%	1	Syn3	85827017-85827031	+	-
19.2B	HS0580M02D305S	10	101394292	36 - 55 (20)	100.00%	1	Mgat4c	101394292-101394311	-	+
19.2B	HS0580M02DTJ5O	16	72239704	38 - 168 (131)	97.71%	12	Intergenic	72239704-72239834	-	+
19.2B	HS0580M02EGDIH	16	79792058	39 - 183 (145)	99.31%	843	Intergenic	79791915-79792058	+	-
19.2B	HS0580M02C2LR8	16	88455299	38 - 55 (18)	100.00%	2	Intergenic	88455299-88455316	-	+
19.2B	HS0580M02EDXDJ	19	23517690	39 - 156 (118)	100.00%	2	Mamdc2	23517690-23517807	-	+
19.2B	HS0580M02ED34D	X	7251668	40 - 55 (16)	100.00%	1	Magix	7251668-7251683	-	+
19.2D	HS0580M02CZSRB	4	126288565	34 - 104 (71)	100.00%	2	5730409E04Rik	126288565-126288635	-	+
19.2D	HS0580M02C4I9Y	5	18843593	39 - 238 (200)	100.00%	154	Magi2	18843394-18843593	+	-
19.2D	HS0580M02DF3WV	5	18863009	39 - 231 (193)	99.48%	1	Magi2	18863009-18863201	-	+
19.2D	HS0580M02DC6RL	8	4115969	39 - 141 (103)	100.00%	36	Intergenic	4115867-4115969	+	-
19.2D	HS0580M02DD7NJ	9	6395060	38 - 171 (134)	100.00%	40	Intergenic	6395060-6395193	-	+
19.2D	HS0580M02ECD1Z	11	5764450	38 - 211 (174)	99.43%	6	Aebp1	5764450-5764623	-	+
19.2D	HS0580M02EOV82	12	13723420	35 - 215 (181)	100.00%	1	Intergenic	13723240-13723420	+	-
19.2D	HS0580M02EIN52	12	19158376	37 - 285 (249)	99.60%	53	Intergenic	19158129-19158376	+	-
19.2D	HS0580M02C909F	14	25548996	37 - 82 (46)	100.00%	1	Intergenic	25548996-25549041	-	+
19.2D	HS0580M02EB1UL	16	3833368	38 - 162 (125)	97.60%	48	Intergenic	3833245-3833368	+	-
19.2D	HS0580M02C8JQT	16	16419796	39 - 216 (178)	100.00%	124	Fgd4	16419796-16419973	-	+
19.2D	HS0580M02DSRGP	16	19540845	35 - 179 (145)	100.00%	16	Intergenic	19540701-19540845	+	-
19.2D	HS0580M02EV5HM	16	25589941	39 - 207 (169)	99.41%	13	Intergenic	25589941-25590110	-	+
19.2D	HS0580M02EWV9P	16	27253525	39 - 138 (100)	99.00%	76	Intergenic	27253525-27253625	-	+
19.2D	HS0580M02DQ0F5	16	51800123	40 - 110 (71)	100.00%	1	Intergenic	51800123-51800193	-	+
19.2D	HS0580M02DW8AV	16	53151578	35 - 186 (152)	98.68%	47	Intergenic	53151578-53151729	-	+
19.2D	HS0580M02DUPFD	16	62191056	39 - 171 (133)	100.00%	15	Intergenic	62190924-62191056	+	-
19.2D	HS0580M02C9U28	16	70643263	40 - 211 (172)	97.67%	1	Intergenic	70643095-70643263	+	-
19.2D	HS0580M02DAGUM	19	49242998	37 - 245 (209)	98.09%	4	Intergenic	49242790-49242998	+	-
19.3A (cre neg)	HS0580M02DZ4G2	2	133338268	157 - 193 (37)	97.30%	4	Intergenic	133338233-133338268	+	-
19.3A (cre neg)	HS0580M02D8IMU	7	116692993	39 - 67 (29)	100.00%	293	St5	116692993-116693021	-	+

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
20.2B	HS0580M02DW9FD	1	106253307	38 - 63 (26)	100.00%	1	Intergenic	106253307-106253332	-	+
20.2B	HS0580M02EVWPI	1	117283578	39 - 215 (177)	100.00%	30	Intergenic	117283578-117283754	-	+
20.2B	HS0580M02C36TQ	1	165393355	38 - 95 (58)	98.28%	283	Intergenic	165393355-165393412	-	+
20.2B	HS0580M02DBX0E	1	173533432	39 - 102 (64)	98.44%	4	Ly9	173533369-173533432	+	-
20.2B	HS0580M02EBOOI	2	60200295	37 - 211 (175)	100.00%	3	Ly75	60200295-60200469	-	+
20.2B	HS0580M02DHGJZ	2	107802949	37 - 58 (22)	100.00%	2	Intergenic	107802949-107802970	-	+
20.2B	HS0580M02DG98V	3	57233833	36 - 61 (26)	96.15%	2	Tm4sf4	57233808-57233833	+	-
20.2B	HS0580M02DAIH0	3	119165312	37 - 126 (90)	100.00%	41	Intergenic	119165223-119165312	+	-
20.2B	HS0580M02C9IWX	3	133691878	38 - 51 (14)	100.00%	1	AC123608.1	133691878-133691891	-	+
20.2B	HS0580M02ECZS2	4	44718860	36 - 54 (19)	100.00%	4	Pax5	44718842-44718860	+	-
20.2B	HS0580M02C6HDN	4	97960998	39 - 116 (78)	100.00%	7	Intergenic	97960998-97961075	-	+
20.2B	HS0580M02DCBUT	4	121772016	38 - 57 (20)	100.00%	23	Intergenic	121772016-121772035	-	+
20.2B	HS0580M02EIZL5	5	45181043	33 - 201 (169)	99.41%	7	Ldb2	45180875-45181043	+	-
20.2B	HS0580M02CZ7Q4	5	110101254	36 - 84 (49)	97.96%	2	Intergenic	110101254-110101302	-	+
20.2B	HS0580M02D2YL5	5	127140010	37 - 85 (49)	100.00%	1	Intergenic	127139962-127140010	+	-
20.2B	HS0580M02DDB2K	6	108031992	37 - 200 (164)	95.73%	1	Intergenic	108031992-108032156	-	+
20.2B	HS0580M02D7A5T	7	50205478	39 - 64 (26)	100.00%	1	Intergenic	50205453-50205478	+	-
20.2B	HS0580M02DKFPJ	7	83792656	38 - 153 (116)	100.00%	7	Agbl1	83792541-83792656	+	-
20.2B	HS0580M02EM8QW	7	89600876	39 - 168 (130)	98.46%	8	Intergenic	89600876-89601004	-	+
20.2B	HS0580M02DKITP	8	116207218	39 - 56 (18)	100.00%	4	Intergenic	116207218-116207235	-	+
20.2B	HS0580M02EF52C	10	21728722	38 - 61 (24)	100.00%	2	Intergenic	21728722-21728745	-	+
20.2B	HS0580M02EMO3A	10	100687183	39 - 66 (28)	100.00%	2	Intergenic	100687183-100687210	-	+
20.2B	HS0580M02EE84I	10	102594934	38 - 117 (80)	100.00%	2	Lrriq1	102594855-102594934	+	-
20.2B	HS0580M02DAEWV	10	105349232	38 - 183 (146)	100.00%	1	Intergenic	105349232-105349377	-	+
20.2B	HS0580M02EZKWW	10	107454791	37 - 237 (201)	97.51%	2	Intergenic	107454791-107454993	-	+
20.2B	HS0580M02C4U8N	11	22982250	36 - 290 (255)	99.22%	2	Intergenic	22982250-22982504	-	+
20.2B	HS0580M02D7AKV	11	63544256	38 - 64 (27)	100.00%	489	Intergenic	63544256-63544282	-	+
20.2B	HS0580M02DURI6	11	80093184	37 - 85 (49)	100.00%	5	Intergenic	80093184-80093232	-	+
20.2B	HS0580M02EVKD8	12	10032983	38 - 161 (124)	100.00%	43	Intergenic	10032860-10032983	+	-
20.2B	HS0580M02ECYKG	13	75983773	40 - 110 (71)	100.00%	48	Glrx	75983703-75983773	+	-
20.2B	HS0580M02E0L6F	13	117285171	36 - 72 (37)	100.00%	405	Intergenic	117285171-117285207	-	+
20.2B	HS0580M02EL92Z	14	17557982	36 - 57 (22)	100.00%	10	Rarb	17557982-17558003	-	+
20.2B	HS0580M02D50Z9	14	29682040	38 - 249 (212)	100.00%	2	Intergenic	29682040-29682251	-	+
20.2B	HS0580M02DDEJE	14	48390123	39 - 144 (106)	99.06%	28	Intergenic	48390018-48390123	+	-
20.2B	HS0580M02EU4TL	15	3543609	39 - 99 (61)	100.00%	12	Intergenic	3543549-3543609	+	-
20.2B	HS0580M02D2OWG	15	29245583	38 - 203 (166)	100.00%	29	Intergenic	29245418-29245583	+	-
20.2B	HS0580M02DU38Q	15	33149006	39 - 66 (28)	100.00%	5	Pgcp	33148979-33149006	+	-
20.2B	HS0580M02DCG0G	15	81845285	37 - 187 (151)	100.00%	4	Xrcc6	81845285-81845435	-	+
20.2B	HS0580M02D2JWK	16	9168411	39 - 120 (82)	100.00%	83	Intergenic	9168411-9168492	-	+
20.2B	HS0580M02CZMOO	16	16096611	39 - 153 (115)	99.13%	16	2310008H04Rik	16096611-16096726	-	+
20.2B	HS0580M02C8GNM	16	19570419	38 - 147 (110)	99.09%	7	Intergenic	19570419-19570529	-	+
20.2B	HS0580M02EB1GI	16	26640284	39 - 156 (118)	98.31%	3	Il1rap	26640284-26640400	-	+
20.2B	HS0580M02DGYKN	16	27660993	39 - 83 (45)	97.78%	266	Intergenic	27660949-27660993	+	-
20.2B	HS0580M02DOA22	16	36447389	37 - 257 (221)	99.55%	3	Intergenic	36447389-36447609	-	+
20.2B	HS0580M02DO80C	16	42791093	39 - 73 (35)	100.00%	261	Gm10809	42791093-42791127	-	+
20.2B	HS0580M02DG9X7	16	68678851	37 - 155 (119)	100.00%	12	Intergenic	68678733-68678851	+	-
20.2B	HS0580M02D85IT	16	74924402	38 - 112 (75)	98.67%	27	Intergenic	74924402-74924477	-	+
20.2B	HS0580M02CZJKS	16	78123502	37 - 115 (79)	100.00%	7	Intergenic	78123424-78123502	+	-
20.2B	HS0580M02ENPLJ	16	79161176	40 - 82 (43)	97.67%	55	Intergenic	79161134-79161176	+	-
20.2B	HS0580M02D4KV7	16	82969846	39 - 207 (169)	97.63%	2	Intergenic	82969846-82970014	-	+
20.2B	HS0580M02CZMII	16	86065632	38 - 99 (62)	98.39%	1	Intergenic	86065571-86065632	+	-
20.2B	HS0580M02EA7C6	17	55023225	37 - 99 (63)	100.00%	18	Intergenic	55023225-55023287	-	+
20.2B	HS0580M02DWS93	18	8678137	37 - 67 (31)	100.00%	70	Intergenic	8678107-8678137	+	-
20.2B	HS0580M02EJZN6	X	50304573	38 - 207 (170)	100.00%	9	Phf6	50304404-50304573	+	-
20.2B	HS0580M02DP02E	X	76474443	37 - 58 (22)	100.00%	1	Intergenic	76474422-76474443	+	-
20.2B	HS0580M02DJXUF	X	135664800	39 - 66 (28)	100.00%	1	Intergenic	135664800-135664827	-	+
20.2B	HS0580M02DZS49	X	149754908	36 - 61 (26)	96.15%	1	Klf8	149754908-149754933	-	+

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
21.3J	HS0580M02ECE1M	1	51123327	39 - 321 (283)	94.35%	5	Tmeff2	51123327-51123597	-	+
21.3J	HS0580M02D93SP	2	121805444	37 - 54 (18)	100.00%	1	Ctdspl2	121805444-121805461	-	+
21.3J	HS0580M02DESAK	3	30811966	38 - 54 (17)	100.00%	3	Phc3	30811950-30811966	+	-
21.3J	HS0580M02D0RLC	3	135974132	38 - 58 (21)	100.00%	1313	Bank1	135974112-135974132	+	-
21.3J	HS0580M02EPEY1	3	158138176	39 - 186 (148)	100.00%	6	Lrrc7	158138029-158138176	+	-
21.3J	HS0580M02D3Z6C	4	96082556	39 - 229 (191)	99.48%	12	Intergenic	96082365-96082556	+	-
21.3J	HS0580M02D38VB	5	43293053	37 - 116 (80)	100.00%	10	Intergenic	43293053-43293132	-	+
21.3J	HS0580M02EMJNW	5	98725575	38 - 92 (55)	98.18%	689	Intergenic	98725575-98725629	-	+
21.3J	HS0580M02C0RPF	5	119162831	39 - 136 (98)	98.98%	55	Med13l	119162831-119162929	-	+
21.3J	HS0580M02EXQ07	6	112258528	37 - 53 (17)	100.00%	3	Lmcd1	112258512-112258528	+	-
21.3J	HS0580M02C9B0R	8	130458544	39 - 247 (209)	99.52%	1	Intergenic	130458544-130458752	-	+
21.3J	HS0580M02C0GPZ	9	81111188	38 - 79 (42)	97.62%	9	Intergenic	81111147-81111188	+	-
21.3J	HS0580M02EOBJQ	10	8341153	39 - 63 (25)	92.00%	1	Intergenic	8341129-8341153	+	-
21.3J	HS0580M02EJSHL	10	53576357	38 - 148 (111)	100.00%	44	Intergenic	53576247-53576357	+	-
21.3J	HS0580M02DXAVH	10	90439024	39 - 133 (95)	100.00%	62	Intergenic	90438930-90439024	+	-
21.3J	HS0580M02C4EIX	11	26860241	36 - 55 (20)	95.00%	1	Intergenic	26860222-26860241	+	-
21.3J	HS0580M02EBP3Y	12	68349381	39 - 198 (160)	99.38%	1	Intergenic	68349381-68349540	-	+
21.3J	HS0580M02D4LNM	14	16024868	38 - 106 (69)	100.00%	74	Intergenic	16024800-16024868	+	-
21.3J	HS0580M02ESW9K	14	41907467	38 - 183 (146)	100.00%	16	Intergenic	41907467-41907612	-	+
21.3J	HS0580M02DRU6S	14	104585092	37 - 196 (160)	96.88%	1	Intergenic	104585092-104585250	-	+
21.3J	HS0580M02EYFKH	15	24848603	37 - 206 (170)	99.41%	11	Intergenic	24848434-24848603	+	-
21.3J	HS0580M02E0VR2	15	55046643	36 - 52 (17)	100.00%	1	Deptor	55046643-55046659	-	+
21.3J	HS0580M02EXJ2Q	16	3463251	38 - 62 (25)	100.00%	3	Intergenic	3463251-3463275	-	+
21.3J	HS0580M02D63CP	16	4004245	37 - 273 (237)	98.73%	9	Intergenic	4004245-4004480	-	+
21.3J	HS0580M02EZZ4P	16	6042397	37 - 156 (120)	100.00%	4	Intergenic	6042278-6042397	+	-
21.3J	HS0580M02DCNGI	16	9403649	36 - 51 (16)	100.00%	2	Intergenic	9403649-9403664	-	+
21.3J	HS0580M02C9DKA	16	12369247	38 - 154 (117)	100.00%	3	Intergenic	12369131-12369247	+	-
21.3J	HS0580M02DZ05Z	16	13608066	36 - 65 (30)	100.00%	49	Parn	13608037-13608066	+	-
21.3J	HS0580M02D77AC	16	21661481	39 - 158 (120)	99.17%	3	2510009E07Rik	21661481-21661600	-	+
21.3J	HS0580M02C73KM	16	25442835	36 - 93 (58)	98.28%	62	Intergenic	25442778-25442835	+	-
21.3J	HS0580M02EQVFJ	16	40069096	39 - 122 (84)	100.00%	3	Intergenic	40069013-40069096	+	-
21.3J	HS0580M02EQADJ	16	40608595	38 - 81 (44)	100.00%	13	Intergenic	40608552-40608595	+	-
21.3J	HS0580M02ELNG4	16	43317069	40 - 57 (18)	100.00%	2	Zbtb20	43317052-43317069	+	-
21.3J	HS0580M02DN9FQ	16	49833054	39 - 105 (67)	100.00%	26	Intergenic	49833054-49833120	-	+
21.3J	HS0580M02D9HUE	16	51842026	37 - 128 (92)	98.91%	62	Intergenic	51841935-51842026	+	-
21.3J	HS0580M02DPH5K	16	52817899	39 - 188 (150)	98.00%	7	Intergenic	52817899-52818047	-	+
21.3J	HS0580M02ECT02	16	54701737	39 - 58 (20)	100.00%	11	Intergenic	54701737-54701756	-	+
21.3J	HS0580M02DIWHJ	16	55238751	39 - 199 (161)	100.00%	1	Zpld1	55238591-55238751	+	-
21.3J	HS0580M02DDWEG	16	65312290	39 - 129 (91)	98.90%	122	Intergenic	65312200-65312290	+	-
21.3J	HS0580M02DK04B	16	65714484	38 - 226 (189)	100.00%	1	Intergenic	65714484-65714672	-	+
21.3J	HS0580M02EPAZW	16	70631184	35 - 182 (148)	100.00%	59	Intergenic	70631037-70631184	+	-
21.3J	HS0580M02EEO1P	17	9661791	39 - 116 (78)	97.44%	109	Intergenic	9661791-9661868	-	+
21.3J	HS0580M02C5BNY	17	15696636	39 - 134 (96)	100.00%	84	Prdm9	15696541-15696636	+	-
21.3J	HS0580M02DIO95	17	93192409	36 - 53 (18)	100.00%	2	Intergenic	93192409-93192426	-	+
21.3J	HS0580M02EYF5V	19	23517692	36 - 151 (116)	100.00%	610	Mamdc2	23517692-23517807	-	+

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
22.2B	HS0580M02D5TCC	1	38105937	36 - 54 (19)	100.00%	13	Eif5b	38105937-38105955	-	+
22.2B	HS0580M02DZLOP	2	98506401	39 - 92 (54)	100.00%	31	Intergenic	98506401-98506454	-	+
22.2B	HS0580M02DY3M4	5	57870912	35 - 74 (40)	97.50%	39	Intergenic	57870912-57870951	-	+
22.2B	HS0580M02EHYW9	5	122111705	34 - 49 (16)	100.00%	1	Brap	122111690-122111705	+	-
22.2B	HS0580M02DLA33	6	85074275	40 - 58 (19)	100.00%	2	Gm5878	85074257-85074275	+	-
22.2B	HS0580M02DKT78	7	102501575	40 - 121 (82)	100.00%	1	Intergenic	102501575-102501656	-	+
22.2B	HS0580M02C2EOU	7	110424128	39 - 172 (134)	99.25%	3	Intergenic	110424128-110424261	-	+
22.2B	HS0580M02EUTQC	7	116692993	39 - 67 (29)	100.00%	9717	St5	116692993-116693021	-	+
22.2B	HS0580M02E0M01	10	9114194	35 - 50 (16)	100.00%	1	Intergenic	9114179-9114194	+	-
22.2B	HS0580M02C7EJT	10	104323595	39 - 175 (137)	100.00%	1	Intergenic	104323595-104323731	-	+
22.2B	HS0580M02EPU30	11	3062113	40 - 159 (120)	99.17%	5	Sfi1	3062113-3062231	-	+
22.2B	HS0580M02EOHQQ	12	79432277	37 - 165 (129)	99.22%	1	Gphn	79432277-79432404	-	+
22.2B	HS0580M02DXU6F	14	51238090	38 - 64 (27)	100.00%	48	Olfr744	51238064-51238090	+	-
22.2B	HS0580M02ETLZP	16	26249551	37 - 132 (96)	98.96%	1	Intergenic	26249456-26249551	+	-
22.2B	HS0580M02DAMWM	16	51403077	36 - 160 (125)	98.40%	1	Intergenic	51402953-51403077	+	-
22.2B	HS0580M02D8AVF	17	13966972	38 - 91 (54)	98.15%	11	Mllt4	13966919-13966972	+	-

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
6.4A	HS0580M02D4DWH	1	34524182	38 - 55 (18)	100.00%	1	Ptpn18	34524165-34524182	+	-
6.4A	HS0580M02DIJ38	1	48163143	39 - 86 (48)	100.00%	4	Intergenic	48163096-48163143	+	-
6.4A	HS0580M02DDX3J	2	73757777	36 - 55 (20)	100.00%	2	Intergenic	73757777-73757796	-	+
6.4A	HS0580M02EBCPM	4	61647717	34 - 54 (21)	100.00%	4	Intergenic	61647697-61647717	+	-
6.4A	HS0580M02DGGNO	6	14082895	34 - 70 (37)	100.00%	1	Intergenic	14082859-14082895	+	-
6.4A	HS0580M02D7PDR	7	18937967	37 - 84 (48)	100.00%	17	Psg18	18937967-18938014	-	+
6.4A	HS0580M02DGUFB	7	110190119	37 - 200 (164)	98.78%	234	Olfr582	110189956-110190119	+	-
6.4A	HS0580M02EW7OH	7	110881235	36 - 233 (198)	97.98%	2	Olfr628	110881038-110881235	+	-
6.4A	HS0580M02EEDTM	8	34333155	34 - 57 (24)	100.00%	1	Intergenic	34333155-34333178	-	+
6.4A	HS0580M02DDS2H	11	89853085	35 - 59 (25)	96.00%	3	Pctp	89853085-89853109	-	+
6.4A	HS0580M02EZBLY	13	58220730	39 - 55 (17)	100.00%	1	Intergenic	58220714-58220730	+	-
6.4A	HS0580M02DSFJN	13	93073855	38 - 55 (18)	100.00%	1	Msh3	93073838-93073855	+	-
6.4A	HS0580M02EZ7I5	16	14535626	37 - 132 (96)	100.00%	20	Intergenic	14535626-14535721	-	+
6.4A	HS0580M02DPR8F	16	87225379	38 - 113 (76)	98.68%	231	Intergenic	87225304-87225379	+	-
6.4G	HS0580M02C3MIM	1	155361934	36 - 58 (23)	91.30%	4	Npl	155361912-155361934	+	-
6.4G	HS0580M02DTJ4P	6	134958884	37 - 58 (22)	95.45%	3	Intergenic	134958884-134958905	-	+
6.4G	HS0580M02ERLFV	7	12740209	39 - 105 (67)	100.00%	10	Intergenic	12740209-12740275	-	+
6.4G	HS0580M02DHI9E	12	31687642	39 - 56 (18)	100.00%	1	Intergenic	31687625-31687642	+	-
6.4G	HS0580M02D4F1U	12	100359583	131 - 247 (117)	98.29%	1	Intergenic	100359583-100359699	-	+
6.4G	HS0580M02EZ015	14	120987953	39 - 77 (39)	97.44%	5309	Intergenic	120987953-120987992	-	+
6.4G	HS0580M02DBFZZ	16	68806219	36 - 176 (141)	100.00%	5	Intergenic	68806219-68806359	-	+
6.4G	HS0580M02DNMY3	16	95487752	39 - 108 (70)	100.00%	3	Kcnj15	95487683-95487752	+	-
6.4H	HS0580M02DSXBC	1	181044763	36 - 57 (22)	95.45%	6	Smyd3	181044742-181044763	+	-
6.4H	HS0580M02C2M42	1	196740788	36 - 60 (25)	100.00%	14	Intergenic	196740788-196740812	-	+
6.4H	HS0580M02DTVIB	3	36833534	39 - 162 (124)	100.00%	6	4932438A13Rik	36833411-36833534	+	-
6.4H	HS0580M02DXMIM	3	61281675	39 - 112 (74)	100.00%	117	Intergenic	61281675-61281748	-	+
6.4H	HS0580M02DEOLS	6	43966176	36 - 88 (53)	98.11%	28	Intergenic	43966176-43966228	-	+
6.4H	HS0580M02DELND	6	137809512	36 - 63 (28)	100.00%	74	Intergenic	137809512-137809539	-	+
6.4H	HS0580M02C87ZL	7	147104540	35 - 66 (32)	100.00%	1	Kndc1	147104509-147104540	+	-
6.4H	HS0580M02DFBXT	10	9534007	39 - 174 (136)	99.26%	5	Stxbp5	9534007-9534141	-	+
6.4H	HS0580M02EGMCA	10	18096044	39 - 262 (224)	99.11%	2	Intergenic	18096044-18096265	-	+
6.4H	HS0580M02D0P85	10	27496908	39 - 148 (110)	98.18%	26	Intergenic	27496908-27497017	-	+
6.4H	HS0580M02DH5OZ	11	3028725	37 - 212 (176)	99.43%	13	Pisd-ps1	3028725-3028900	-	+
6.4H	HS0580M02D72Y2	11	33951526	36 - 59 (24)	100.00%	65	4930469K13Rik	33951526-33951549	-	+
6.4H	HS0580M02C4KTP	11	61726553	37 - 67 (31)	100.00%	2	Akap10	61726523-61726553	+	-
6.4H	HS0580M02C8BG0	12	37673080	37 - 98 (62)	100.00%	11	Intergenic	37673019-37673080	+	-
6.4H	HS0580M02DH98N	12	37891329	38 - 151 (114)	100.00%	22	Meox2	37891216-37891329	+	-
6.4H	HS0580M02EA9DN	12	46755152	37 - 88 (52)	100.00%	198	Intergenic	46755152-46755203	-	+
6.4H	HS0580M02DCKHK	13	27445337	40 - 234 (195)	100.00%	591	Prl8a2	27445337-27445531	-	+
6.4H	HS0580M02DMXRF	14	56205311	39 - 68 (30)	100.00%	20	Intergenic	56205311-56205340	-	+
6.4H	HS0580M02EQ5MV	16	9940260	39 - 73 (35)	100.00%	94	Grin2a	9940260-9940294	-	+
6.4H	HS0580M02C0OTG	16	47861517	38 - 150 (113)	100.00%	8	Intergenic	47861405-47861517	+	-
6.4H	HS0580M02DH26N	16	62189451	38 - 145 (108)	100.00%	8	Intergenic	62189451-62189558	-	+
6.4H	HS0580M02DL4N1	16	85434890	40 - 188 (149)	100.00%	1	Intergenic	85434890-85435038	-	+
6.4H	HS0580M02D7VXX	18	7177104	37 - 64 (28)	100.00%	1	Armc4	7177077-7177104	+	-
6.4H	HS0580M02EVHQX	X	20069287	35 - 113 (79)	100.00%	33	Phf16	20069209-20069287	+	-

Sample	Read Name	Chr	Transposon Integration Site	Query Start-Stop (length)	Alignment % Identity	Reads per Cluster	Hit	Hit Start-Stop	Transposon Ori	Hit Strand
7.5B	HS0580M02DLWWS	1	181165067	40 - 56 (17)	100.00%	1	Smyd3	181165067-181165083	-	+
7.5B	HS0580M02EUU60	2	163468418	36 - 75 (40)	92.50%	61	Serinc3	163468418-163468457	-	+
7.5B	HS0580M02EG9XC	5	28402713	36 - 69 (34)	100.00%	44	Insig1	28402713-28402746	-	+
7.5B	HS0580M02DQEVE	9	29834781	39 - 280 (242)	99.59%	1	Intergenic	29834541-29834781	+	-
7.5B	HS0580M02D48U1	9	91934272	36 - 161 (126)	99.21%	2	Intergenic	91934148-91934272	+	-
7.5B	HS0580M02EQ3EM	14	28172683	40 - 54 (15)	100.00%	160	Arhgef3	28172683-28172697	-	+
7.5B	HS0580M02EJNXU	16	67517810	39 - 97 (59)	100.00%	24	Cadm2	67517752-67517810	+	-
7.5C	HS0580M02C0EB0	1	16377410	36 - 208 (173)	98.27%	19	Stau2	16377410-16377582	-	+
7.5C	HS0580M02D3PV1	2	13426662	38 - 61 (24)	95.83%	1	Intergenic	13426662-13426685	-	+
7.5C	HS0580M02DBJ35	2	14335073	39 - 167 (129)	100.00%	29	Slc39a12	14335073-14335201	-	+
7.5C	HS0580M02DNW7Z	3	120706403	38 - 76 (39)	100.00%	97	Intergenic	120706403-120706441	-	+
7.5C	HS0580M02DDST1	5	9483845	37 - 196 (160)	100.00%	2	Intergenic	9483845-9484004	-	+
7.5C	HS0580M02DGGJR	7	46526510	38 - 97 (60)	98.33%	1	Intergenic	46526451-46526510	+	-
7.5C	HS0580M02DV08W	9	23351310	39 - 258 (220)	98.64%	1	Intergenic	23351310-23351529	-	+
7.5C	HS0580M02ET5E9	9	65507968	38 - 61 (24)	100.00%	1	Rbpms2	65507968-65507991	-	+
7.5C	HS0580M02DWFP5	10	9425236	39 - 176 (138)	100.00%	68	Intergenic	9425099-9425236	+	-
7.5C	HS0580M02DUTQG	12	84845364	37 - 56 (20)	100.00%	1	Intergenic	84845345-84845364	+	-
7.5C	HS0580M02C30Z4	12	101838265	37 - 58 (22)	100.00%	1	Rps6ka5	101838265-101838286	-	+
7.5C	HS0580M02D8V89	13	117356737	36 - 51 (16)	100.00%	3	Intergenic	117356722-117356737	+	-
7.5C	HS0580M02DY8LB	16	28724887	37 - 213 (177)	100.00%	3	Intergenic	28724887-28725063	-	+
7.5C	HS0580M02DNDC1	16	48508091	39 - 182 (144)	100.00%	103	Morc1	48508091-48508234	-	+
7.5C	HS0580M02EB100	16	54099065	39 - 125 (87)	98.85%	25	Intergenic	54098980-54099065	+	-
7.5C	HS0580M02DNHN1	16	56169105	39 - 146 (108)	100.00%	2	Senp7	56169105-56169212	-	+
7.5C	HS0580M02DB1HV	16	69126300	37 - 71 (35)	100.00%	32	Intergenic	69126300-69126334	-	+
7.5C	HS0580M02C8BEK	16	71424798	39 - 302 (264)	98.86%	1	Intergenic	71424798-71425062	-	+
7.5C	HS0580M02EU0FE	16	80527692	36 - 140 (105)	89.52%	18	Intergenic	80527595-80527692	+	-
7.5C	HS0580M02ENZ6S	17	17058324	38 - 53 (16)	100.00%	4	Intergenic	17058324-17058339	-	+
7.5H	HS0580M02DBXZX	1	26794463	37 - 151 (115)	100.00%	27	Intergenic	26794349-26794463	+	-
7.5H	HS0580M02DQOCK	1	141255311	39 - 128 (90)	98.89%	1	Crb1	141255223-141255311	+	-
7.5H	HS0580M02EWHQX	1	149681088	35 - 108 (74)	98.65%	10	Intergenic	149681015-149681088	+	-
7.5H	HS0580M02EQ84U	1	152145196	37 - 110 (74)	98.65%	124	Intergenic	152145123-152145196	+	-
7.5H	HS0580M02DUFBS	1	152448908	39 - 79 (41)	100.00%	60	Hmcn1	152448868-152448908	+	-
7.5H	HS0580M02D6LIL	2	60907768	37 - 54 (18)	100.00%	59	Intergenic	60907751-60907768	+	-
7.5H	HS0580M02DQ4P8	2	101705017	36 - 54 (19)	100.00%	1	Prr5l	101704999-101705017	+	-
7.5H	HS0580M02C3N5T	2	163468418	39 - 78 (40)	92.50%	95	Serinc3	163468418-163468457	-	+
7.5H	HS0580M02DZTV6	3	75350388	38 - 105 (68)	100.00%	68	Pdcd10	75350388-75350455	-	+
7.5H	HS0580M02EBYEI	3	75905868	36 - 53 (18)	100.00%	1	Fstl5	75905868-75905885	-	+
7.5H	HS0580M02EA6CH	5	28496774	38 - 77 (40)	100.00%	4	En2	28496774-28496813	-	+
7.5H	HS0580M02EE3TM	8	75325174	39 - 188 (150)	99.33%	18	Intergenic	75325025-75325174	+	-
7.5H	HS0580M02D0GYO	10	53252437	39 - 162 (124)	98.39%	14	Intergenic	53252437-53252560	-	+
7.5H	HS0580M02DFE5I	11	3067893	40 - 175 (136)	98.53%	10	Sfi1	3067757-3067893	+	-
7.5H	HS0580M02DUUYS	11	112634790	38 - 102 (65)	100.00%	227	BC006965	112634726-112634790	+	-
7.5H	HS0580M02EKHWG	11	116757245	38 - 136 (99)	100.00%	2	Intergenic	116757147-116757245	+	-
7.5H	HS0580M02ETWPD	12	45484485	37 - 55 (19)	100.00%	2	Nrcam	45484467-45484485	+	-
7.5H	HS0580M02EKSDJ	14	25548996	37 - 82 (46)	100.00%	48	Intergenic	25548996-25549041	-	+
7.5H	HS0580M02C8ZW2	15	56095373	40 - 138 (99)	100.00%	37	Intergenic	56095275-56095373	+	-
7.5H	HS0580M02C6A09	15	81218977	38 - 193 (156)	100.00%	2	St13	81218977-81219132	-	+
7.5H	HS0580M02DSVAJ	16	3988504	38 - 118 (81)	98.77%	60	Slx4	3988424-3988504	+	-
7.5H	HS0580M02D2APG	16	7392723	38 - 75 (38)	97.37%	95	Rbfox1	7392686-7392723	+	-
7.5H	HS0580M02D1CLK	16	12385827	37 - 176 (140)	99.29%	1	Intergenic	12385827-12385967	-	+
7.5H	HS0580M02EEMXI	16	16193405	38 - 202 (165)	100.00%	1	Intergenic	16193405-16193569	-	+
7.5H	HS0580M02DJZN3	16	27873374	39 - 59 (21)	100.00%	102	Intergenic	27873374-27873394	-	+
7.5H	HS0580M02DU9Z5	16	41293954	37 - 99 (63)	100.00%	193	Intergenic	41293954-41294016	-	+
7.5H	HS0580M02DGL9A	16	46720946	39 - 170 (132)	99.24%	6	Intergenic	46720815-46720946	+	-
7.5H	HS0580M02EAOSF	16	61860006	40 - 170 (131)	99.24%	701	Intergenic	61859876-61860006	+	-
7.5H	HS0580M02C0HSE	16	70016908	39 - 101 (63)	100.00%	108	Intergenic	70016908-70016970	-	+
7.5H	HS0580M02EPL5B	16	74917820	36 - 222 (187)	100.00%	3	Intergenic	74917634-74917820	+	-
7.5H	HS0580M02DGBB6	16	79815100	38 - 103 (66)	100.00%	5	Intergenic	79815035-79815100	+	-
7.5H	HS0580M02DO80U	18	25410073	36 - 185 (150)	99.33%	6	AW554918	25409924-25410073	+	-
7.5H	HS0580M02D5WEK	18	51216844	36 - 90 (55)	100.00%	1	Intergenic	51216844-51216898	-	+
7.5H	HS0580M02D5KP9	X	99382099	36 - 52 (17)	100.00%	1	Rps4x	99382083-99382099	+	-

Appendix5A: Results of the CIS analysis performed using all integrations with 2 or more reads on TraDIS

Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	Start	End	Number of Insertions	Smallest p Value	CIS Analysis Scales	Genes in CIS	Gene Nearest Peak
11	54252853	54257455	41.33-48	54030770	54470321	49	0	10-100	Pdlim4 Paha2 Gm12221 4933405E24Rik Gm12222 Csf2 Gm12223 Il3 Acsl6 Gm12224 4930404A10Rik Gm12226 Gm12229 Fnip1 Gm24198	Gm12223
11	79361616	79557057	16.76-36.01	79221934	79690789	43	0	10-100	Wsb1 Gm9964 Nf1 Gm11198 Gm11199 AU040972 Omg Gm21975 Evi2b Evi2a Rab11fip4 Gm23293 Gm11202 Gm25867 Gm24887 4930542H20Rik	Nf1
2	98661681	98670460	29.76-29.97	98465575	98857787	31	0	10-100	Gm13806 Gm10801 Gm10800	Gm10801
5	28168140	28174030	20.56-22.93	27977161	28370051	25	0	10-100	AC156021.1 Insig1 En2 Cnpy1 Rbm33	En2
7	102155068	102179487	10.93-16.86	101989801	102342775	18	0	10-100	Numa1 Il18bp Rnf121 Trpc2 Arl5 Arl1 Chrna10 Nup98 Pgap2 Rhog Stim1	Nup98
9	44835970	44841958	12.32-15.85	44663945	44996751	16	0	10-100	Treh Phldb1 Gm24166 Arcn1 Ifi46 Tmem25 Ttc36 Mll1 Gm26249 Atp5l Ube4a Cd3g Cd3d	Mll1
11	100832721	100841602	9.83-14.57	100672030	100974832	15	0	10-100	Zfp385c Gm11547 Dhx58 Katl2a Hepb8 Rabl6c Kcnh4 Hert Ghdc Gm24358 Stat5b Stat5a Stat3 Prf	Stat5b
6	30854745	31124979	2.16-6.21	30813113	31227202	11	0	10-100	Copg2 4930412F05Rik Tspa13 Klf14 Mir29a Mir29b-1 RP23-459L15.5 Gm13834 Gm13833 Gm13835 AB041803	Gm13833
9	32607818	32704786	8.82-10.61	32555878	32820752	11	0	10-100	Elsi1	Elsi1
15	3486552	3581150	7.02-9.75	3376860	3857659	11	0	10-100	Ghr Gm22031	Ghr
16	4227802	4227802	3.28-6.84	4132652	4277844	10	2.220456-16	20-100	Crebbp Gm5766 Gm24107	Crebbp
4	44647140	44656027	5.27-8.85	44500518	44774213	9	0	10-100	Pax5 Mir5120 Gm12462 Gm12463 Zcchc7	Pax5
5	147363448	147373807	7.93-8.59	147236918	147491140	9	0	10-100	2210019I11Rik Pde1 RP24-515G5.4 Cdk2 Phtoxnb Fit3 AC134441.1 Gm6054 Pan3	Fit3
8	10856322	10915557	5.31-8.51	10732768	11005860	9	0	10-100	Rpa16-ps3 3930402G23Rik Ire2	3930402G23Rik
11	3141746	3150284	5.73-7.02	3091965	3236194	8	0	10-100	Ptpn1 Sin1 Gm17399 Gm17400 Dnjc1 Gm12735 Fau-ps2 Eif4enif1	Sin1
16	15709310	15710214	5.85-6.24	15660913	15757708	8	0.000134914	90, 100	Prkdc	Prkdc
4	32379714	32394543	2.66-4.29	32291740	32445444	7	0	10-100	Bach2	Bach2
4	101233700	101260741	2.26-6.68	101118402	101341224	7	0	10-100	Ravin2 Jak1 Gm24468 Gm12785 Gm25124 Gm12801 Gm12796	Jak1
6	103647351	103650428	6.99-7.00	103540390	103746054	7	0	10-100	Chl1 Gm24784	Chl1
7	75712863	75732567	3.4-4.87	75693253	75755748	7	5.53475E-05	30-100	Akap13	Akap13
15	78485400	78496467	5.66-6.34	78359870	78592256	7	0	10-100	Gm26329 Tex33 Tsi Mpst Kctd17 Tmprss6 Il2rb C1qtnf6 Satr3 Gm6723 Rac2	Il2rb
15	98819465	98829368	2.03-3.75	98782929	98852523	7	2.87356E-07	20, 40-80, 100	Wnt1 Ddn AC161165.1 Prkag1 Mll2	Prkag1
16	52138115	52143988	4.33-5.31	52115955	52159821	7	1.9984E-15	10-70	Cbl b	Cbl b
1	194097427	195025139	3.59-5.24	194870487	195095073	6	0	20-100	Cd34 AC162692.1 mmu-mir-29b-2 Mir29c Cd46	mmu-mir-29b-2
2	117348438	117359813	3.48-5.03	117272135	117409410	6	0	10-100	Rasgrp1	Rasgrp1
3	103049284	103074742	3.13-4.55	102975203	103182730	6	0	10-100	Sike1 Cpde1 Nras Ampd1 Gm23820 Dennd2c Bcas2 RP24-409D4.2	Cpde1
8	83769992	83781126	2.91-4.58	83672218	83836833	6	0	20-100	Pigpc1 Pkn1 Ddx39 Cd97 Gm22024	Gm22024
4	97907280	97910254	3.04-3.77	97852585	97946348	5	0.000116716	60-100	Nfia	Nfia
5	136357299	136357721	3.06-4.32	136281820	136418050	5	1.64817E-06	50-100	Cux1 Gm16599 AA390100C17Rik	Cux1
6	30125308	30138281	2.43-4.03	30063144	30196385	5	0	10-100	Nrf1 Gm25580 Mir182 Mir96 Mir183	Nrf1
6	41588237	41602388	3.04-4.95	41508104	41658720	5	0	10-100	Prss2 Trbd1 Trbj1-1 Trbj1-2 Trbj1-3 Trbj1-4 Trbj1-5 Trbj1-6 Trbj1-7 Trbc1 Trbd2 Trbj2-1 Trbj2-2 Trbj2-3 Trbj2-4 Trbj2-5 Trbj2-6 Trbj2-7 Trbc2 Trbv31 Epha6 Trpv6	Epha6
7	145050270	145062409	3.26-4.86	145007114	145096226	5	0	10-100	intergenic	Cond1
8	87908734	87954889	2.18-4.82	87836599	88002357	5	0	10-100	Zfp423	Zfp423
10	28537714	28548136	2.07-4.11	28482399	28593658	5	9.93865E-09	10, 30, 50-100	Pprk	Pprk
11	85619982	85630975	2.04-5	85435600	85651447	5	0	10-100	Nhs1	Nhs1
12	26324563	26398276	3.02-3.96	26291117	26485951	5	9.57499E-07	20, 60-100	Rnf144a Rasd2 Cmpk2	Rnf144a
14	103108202	103116923	3.3-4.12	103028789	103184948	5	0	10, 20, 40-100	Irg1 Chi5 Fbxl3 Mycbp2	Mycbp2
17	62957784	62991033	2-4.1	62913289	63013452	5	4.76038E-09	10, 30-100	intergenic	Fbxl17
18	67483561	67496164	2.00-3.68	67420291	67537864	5	5.74394E-11	10-30, 50-100	Afg3l2 Gm24900 Stmn1 Spire1	Spire1
4	38399154	38399241	3.13-3.98	38345957	38433790	4	0	10-100	intergenic	Gm12380
4	116029583	116034692	2.96-3.14	115994417	116047180	4	0.000328338	80, 90	Faah Naun4 Naun4	Naun4
6	123183647	123197085	3.79-3.98	123039159	123239159	4	0	10-100	Kcnh1 Gm25160 Clec2d	Gm25160
6	141553650	141562742	2.73-3.36	141457324	141457324	4	0	20-100	Lrrk2 Gm22793 Rps6ka5 Rc3h1 Gm22793	Gm22793
8	95757207	95760219	3.12-3.89	95709802	95807192	4	0	30-100	Nanog Sesn1 Cnot1 Gm26449 Gm26065	Cnot1
8	125747008	125756930	2.07-3	125714665	125773070	4	0	10-70	Nipsr Panx2	Nipsr
9	3002194	3024029	2.07-4.03	2984170	3057834	4	0	10-100	AC131780.1 Gm10722 Gm11168 Gm10721 Gm10720 Gm10719 Gm10718 Gm10717 Gm17535	AC131780.1
10	88755802	88771056	2.61-3.37	88722427	88802305	4	1.77946E-05	20, 30, 50-90	Gm4925 Art1 Utp20	Utp20
12	16906869	16912184	2.74-3.3	16831935	16956646	4	8.672E-06	70-100	Rock2 Gm22303	Rock2
13	10286167	10318888	3.43-3.94	10236162	10352072	4	0	10, 30, 40, 60-100	intergenic	AC122624.1
15	87347759	87344743	2.91-3.4	87306968	87387926	4	6.22511E-05	40-60	intergenic	Sf3a3a1
16	9004016	9006014	3.29-3.65	4989961	5014644	4	2.28158E-05	20, 30	Sept12 Gm5480 Rogdi	Gm5480
16	42802722	42802751	3.4	42802641	42802752	4	0.000163565	20, 30	Zdhhc20	Zdhhc20
16	9263194	9263194	3.4	9262372	9263944	4	0.000163277	30	Runx1	Runx1
19	27298287	27308515	3.15-4.05	27249085	27343980	4	0	10, 20, 40-100	Vldlr Kcnv2	Kcnv2
X	75226579	75242769	2-2.96	75178401	75255486	4	8.39327E-10	10, 40, 100	F8 Gm6039	F8
X	107101017	107124183	2.6-3.41	107024321	107159221	4	2.88417E-07	40-60, 80-100	P2ry10 A630033H20Rik	P2ry10
1	53820611	53830365	2.05-2.97	53797541	53844694	3	5.01383E-11	10, 20, 50-90	Hecw2	Hecw2
2	3831073	3833472	2.67-2.98	3818278	3838045	3	1.11022E-16	10, 30-80	Cdh1 Gm13186	Gm13186
3	135444722	135449053	2.07-2.97	135427646	135472217	3	9.31143E-08	10-80	Ube2d3	Ube2d3
4	90857554	90857554	2.56-2.84	90846856	90869140	3	0.000163242	40-60	Ube2e3 Ube2e3b	Ube2e3b
4	155508331	155510215	2.82-2.97	155474608	155527308	3	8.67457E-05	20, 50, 60	Gm12632	Gm12632
5	52764605	52768347	2.09-3	52745089	52779242	3	0	10-80	Gnb1 Gm13171	Gnb1
6	115564757	115569177	2.4-2.99	115627538	115674468	3	1.12155E-13	10, 30, 50-80	Pi4k2b Zcchc4	Pi4k2b
7	15967752	15988927	2.85-2.99	15974141	15996764	3	1.81672E-09	10, 30, 50-80	Ralf1 Gm14335	Ralf1
							0	Oct-40	Gfaic1	Gfaic1

Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	Start	End	Number of Insertions	Smallest p Value	CIS Analysis Scales	Genes in CIS	Gene Nearest Peak
7	49334569	49335622	3	49322838	49341114	3	2.22045E-16	Oct-30	Nav2	Nav2
8	80321591	80332915	2-2.91	80299668	80338567	3	1.30381E-08	10, 40-60	intergenic	Gypa
8	122211602	122216151	2.99-3.01	122163585	122245982	3	0	10-90	Gm20388	Gm20388
10	37267358	37270482	2.63-2.97	37242980	37286881	3	3.77762E-10	20, 50-70	intergenic	AC153356.1
10	96700698	96704873	2.35-2.98	96676176	96718259	3	1.11022E-16	10, 20, 40-70	intergenic	Btg1
10	128338669		2.35	128291842	128367937	3	0.000272449	60	Stat2 Il23a Gm23241 Pan2 Cnpy2 Gm24320 Cs Gm23182 Coq10a	Cs
12	3104317	3107078	2.97-3	3027959	3136369	3	0	10-100	Gm9075	Gm9075
12	72752333	72763999	2.01-2.97	72717290	72788403	3	5.10854E-08	10,20, 40, 60-100	Ppm1a	Ppm1a
14	8217741	8226082	2.98-3	8189385	8252371	3	0	10-40, 60-100	Kdd8 Acox2	Kdd8
14	34710955	34715978	2.89-2.96	34673490	34738190	3	1.71759E-06	30, 40, 60-90	Wapal	Wapal
14	50292326	50304888	2.01-2.37	50285463	50320428	3	1.74416E-06	20, 40	Olfr733 Olfr734	Olfr733
14	64889089	64893894	2.67-2.91	64857947	64914320	3	0.000112464	40, 70, 80	Hmbox1	Hmbox1
16	49683599		3	49679749	49686486	3	1.29903E-07	10000	intergenic	Ift57
17	45435422	45438636	2.26-2.69	45425894	45444951	3	0.000430267	30, 50	Cdc5l AC163677.1	AC163677.1
17	71133687		2.93	71109866	71143215	3	0.000156086	50	AC126942.1 Myom1	AC126942.1
17	78884521	78885567	2.92-2.98	78866035	78899384	3	2.1763E-05	30-60	Eif2ak2 AC154274.1 Sult6b1	AC154274.1
18	*35557442*	*35569457*	*2.01-2.93*	*35537414*	*35591358*	*3*	*1.8191E-07*	*10, 30-70*	*AC121821.1 Mir1949 Snora74a Matr3*	*Slmo1*
19	9879118	9881087	2.99-3.01	9856113	9902914	3	0	10-60	Incenp	Incenp
19	32795573	32796772	2.58-2.75	32785563	32813584	3	6.90738E-05	30, 40	Pten	Pten
X	57273371	57279664	2.73-3	57227130	57305672	3	0	10, 30-50, 80-100	Arhgef6	Arhgef6
X	106128203	106130078	2.8-2.83	106099296	106147474	3	0.000354245	90, 100	Atp7a Tlr13	Atp7a
X	152255733	152273471	2.07-2.77	152225293	152319742	3	1.23124E-13	40-70, 100	Kdm5c	Kdm5c
1	39900931		1.99	39892162	39903854	2	2.74476E-05	30	Map4k4	Map4k4
1	131262493	131263176	2	131253724	131269019	2	1.11022E-16	10-30	Ikbke	Ikbke
1	*155687246*		*2*	*155686272*	*155687246*	*2*	*1.90536E-12*	*10*	*Acbd6*	*Acbd6*
2	12424960		2	12423005	12424960	2	4.47474E-05	10	intergenic	Gm13321
2	69233232		2	69231276	69233232	2	4.3996E-05	10	intergenic	Abcb11
2	124170475		2	124168520	124170475	2	1.40837E-05	10	Sema6d	Sema6d
2	163717706	163718174	2	163713795	163719662	2	1.54598E-06	10, 20	Pkig	Pkig
2	**167760337**	**167771993**	**1.06-1.91**	**167756336**	**167775907**	**2**	**0.000119615**	**10, 40**	**Intergenic**	**Gm14321**
2	180206483		2	180199663	180215281	2	0	10	Lama5	Lama5
3	24665151		2	24663194	24665151	2	2.84158E-05	10	intergenic	Gm24062
3	30190433		2	30187499	30190433	2	6.56142E-13	10	Mecom	Mecom
3	55588343	55588476	1.99-2	55584561	55588476	2	1.11022E-16	10, 20	intergenic	Nbea
3	**60570684**		**2**	**60567749**	**60570684**	**2**	**1.2449E-07**	**10**	**Mbnl1**	**Mbnl1**
3	102195869	102196177	2	102191954	102196177	2	0	10,20	Vangl1	Vangl1
4	53620575		2	53619601	53620575	2	0	10	Slc44a1	Slc44a1
4	56785049		2	56784075	56785049	2	3.76289E-10	10	Ikbkap	Ikbkap
4	95019762		2	95018788	95019762	2	6.47301E-05	10	Gm12694	Gm12694
5	30363808	30327417	2	30340418	30368367	2	0	10-30	Drc1	Drc1
5	36677116	36677217	2	36675167	36678090	2	3.98365E-06	10, 20	D5Ertd579e	D5Ertd579e
5	66232309		2	66230361	66232309	2	3.96722E-05	10	intergenic	Nsun7
6	24491388		2	24489437	24491388	2	1.92208E-06	10	Iqub	Iqub
6	28375684		2	28373732	28375684	2	4.09108E-05	10	Zfp800	Zfp800
6	54881780		2	54879828	54881780	2	1.14314E-08	10	Znrf2	Znrf2
6	117861212		2	117855355	117861212	2	1.53175E-07	20	intergenic	Zfp239
7	28007021		2	28007021	28007021	2	0.000229071	10	RP23-73F23.2 Gm4636	RP23-73F23.2
7	59278094		2	59272234	59282001	2	0	10	Ube3a	Ube3a
7	100142936		2	100140982	100142936	2	5.81581E-05	10	intergenic	Lipt2
7	101367797		2	101364866	101367797	2	3.11973E-13	10	Arap1	Arap1
7	126597590		2	126595636	126597590	2	8.99606E-08	10	intergenic	Il27
8	58821642		1.99	58820671	58821642	2	1.28616E-05	10	Galntl6	Galntl6
8	105600308	105600606	2	105596720	105601279	2	1.56342E-05	10, 20	intergenic	Fam65a
9	114911091		2	114908167	114914014	2	4.80899E-11	10	Gpd1l	Gpd1l
9	115061159		2	115057261	115062133	2	2.18714E-10	10	Osbpl10	Osbpl10
10	11586789		2	11587816	11586789	2	2.39003E-07	10	intergenic	Gm9797
10	13391280		2	13390308	13391280	2	2.4114E-06	10	Phactr2	Phactr2
12	31043009		2	31040100	31043009	2	1.90797E-05	10	intergenic	Gm25408
12	54914584		2	54912645	54914584	2	1.08962E-09	10	Baz1a	Baz1a
12	70516433	70516885	2	70511065	70516885	2	4.26816E-08	10, 20	intergenic	Tmx1
13	93908606	93909559	2	93906690	93909559	2	1.07163E-10	10, 20	Arsb	Arsb
14	39481146		1.99	39480177	39481146	2	0.000118964	10	intergenic	Gm20642
14	59737317		2	59734405	59737317	2	6.22907E-05	30	AC162936.1	AC162936.1
15	**79215343**	**79216580**	**2**	**79211488**	**79217270**	**2**	**3.34415E-10**	**10, 20**	**Gm10863**	Gm10863
17	26871097		2	26869196	26871097	2	8.03801E-14	10	Gm17382 Gm20458	Gm17382
17	29484624		2	29482722	29484624	2	0.000103657	20	AC163629.1	AC163629.1
17	50118147	50118275	2	50116245	50118275	2	1.11022E-16	10, 20	Rftn1	Rftn1
17	56698150		1.81	56697199	56705752	2	0.000248063	10	Ranbp3	Ranbp3
17	57821404		2	57820453	57821404	2	0.00033417	10	Cntnap5c	Cntnap5c
17	60763531		2	60762581	60763531	2	6.38378E-14	10	intergenic	Gm23447
18	11121284	11121728	2	11117457	11123198	2	1.27253E-07	10-20	intergenic	Gata6
18	34549606		2	34547694	34549606	2	6.44834E-10	10	Wnt8a	Wnt8a
18	68825584		2	68822715	68826540	2	0.000183073	10	intergenic	4930548C10Rik
18	80627016	80627771	2	80611735	80646908	2	0	10-40	Nfatc1	Nfatc1
19	47154357	47156074	2	47150617	47157939	2	5.74687E-09	10, 40	Calhm3	Calhm3
19	50945930		2	50941269	50949658	2	0	10	intergenic	Gm25659
X	36556293		2	36552528	36560214	2	0	10	intergenic	Pqrmc1
X	106772892	106774885	2	106770009	106774885	2	6.66134E-16	10, 30	Brwd3	Brwd3
X	143483352		2	143483352	143483352	2	0.003029448	20	intergenic	Pak3
Y	1172264	1183294	1.05-1.99	1106013	1300813	2	0	10, 30-100	Uty Ddx3y	Uty

Appendix5A: CIS analysis using all integrations with 2 or more reads from the duplicate filtered analysis of the TraDIS data.

Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	CIS Start	CIS End	Number of hits	Number of tumours	Smallest p Value	Kernel scales	Genes - largest CIS	Genes - smallest CIS	Gene Nearest Peak
CIS excluded as single kernel scale only												
7	28006812		2.064574125	28002930	28006812	2	2	1.3715E-06	20	RP23-73F23.2 Gm4636		RP23-73F23.2
7	132961846		2.080201475	132961846	132965731	2	2	0.001057554	40	Fgfr2 Zranb1		Fgfr2
2	91886251		2.048959164	91879444	91886251	2	1	9.97732E-07	50	Ambra1		Ambra1
CIS excluded as reported as recurrent CIS across multiple screens												
2	98662180	98668789	4.06-4.12	98554936	98756369	4	4	0	10-100	Gm13806 Gm10801 Gm10800	intergenic	Gm10800
5	28162029	28165870	2.73-3.18	28081557	28225777	3	3	0	20, 30, 50-100	En2 Cnpy1	En2	En2
11	3140606	3143290	3.31-3.54	3087203	3196490	3	3	0	10-100	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2	Sfi1 Gm11399	Sfi1
CIS excluded as shared by samples 9.1b and 9.1d only												
1	33365572	33368935	2	33353487	33376659	2	2	3.37952E-13	40, 80	intergenic		Gm23453
2	180204839	180206513	2.06	180171548	180259830	2	2	0	10-80, 100	Lama5 Rps21 Mir3091 Cables2	Lama5	Lama5
4	95017313	95020595	2.37	95012411	95020595	2	2	7.84138E-05	40, 60, 80	Gm12694	Gm12694	Gm12694
8	13773819	13777535	2.04	13759401	13782034	2	2	0	10, 50, 80	Cdc16	Cdc16	Cdc16
9	41123336	41124635	2.21-2.33	41097792	41138651	2	2	0	20, 40, 70	Ubash3b	Ubash3b	Ubash3b
9	114910254	114911668	2.21	114891109	114914083	2	2	0	20, 40	Gpd1l	Gpd1l	Gpd1l
9	115059581	115060843	2.21	115040437	115074897	2	2	0	20, 40	Osbpl10	Osbpl10	Osbpl10
10	20487361	20490989	2.05-2.16	20467002	20497836	2	2	0	10-30, 50-80	Pde7b	Pde7b	Pde7b
14	37011641	37014879	2.13	36993625	37014879	2	2	0	10-40, 80	intergenic	intergenic	Rgr
18	35556754	35557381	2.05-2.06	35541192	35562636	2	2	1.66533E-15	10, 20, 50, 60	AC121821.1 Mir1949 Snora74a Matr3	AC121821.1 Snora74a	AC121821.1
19	9880120	9881128	2.3-2.4	9871906	9885738	3	2	0	10, 20, 50, 70	Incenp	Incenp	Incenp
CIS excluded as low hits on review*												
11	16791417		2.234595157	16790445	16791417	2	0	0.000112297	10	Egfr		Egfr
11	67563868		2.253461297	67561923	67563868	2	0	0.000263611	20	Gas7		Gas7
X	143066829		1.061163139	143065001	143066829	1	0	9.52658E-07	20	intergenic		Rgag1
13	55487979	55514966	1.03-1.14	55376537	55525126	1	1	0	40, 80-100	Rgs14 Slc34a1 Pfn3 F12 Grk6 Prr7 Dbn1 Pdlim7 Dok3	Pdlim7	Pdlim7
14	70597955	70632810	1.07-1.26	70574426	70632810	1	1	1.47458E-05	40-60	Nudt18 Fam160b2 Epb4.9	Fam160b2	Epb4.9
7	100142866	100142962	2.06	100138984	100142962	2	0	3.1881E-06	10, 20	intergenic	intergenic	Lipt2
10	11588263	11589033	2.05-2.16	11577306	11589033	2	0	1.95399E-14	10, 40, 60	intergenic	intergenic	Gm9797
2	93639136	93641370	1.48-2.02	93604888	93654700	2	1	2.13964E-05	20, 60, 80	intergenic	Alx4	Alx4

Appendix 5Bi: CIS integrations excluded from the 'top 25 analysis ' of the TraDIS data.

Chromosome	Minimum Peak Location	Maximum Peak Location	Peak Height (range)	CIS Start	CIS End	Number of hits	Number of tumours	Smallest p Value	Kernel scales	Genes - largest CIS	Genes - smallest CIS	Gene Nearest Peak
Excluded as single kernel scale only												
1	138127240		2.060599912	138125314	138127240	2	1	5.60106E-05	10	Ptprc		Ptprc
1	155687294		2.068229491	155686331	155688257	2	2	5.38318E-06	10	Acbd6		Acbd6
1	157506974		2.067319122	157505047	157506974	2	0	0.000112911	10	Gm15486 Sec16b		Gm15486
2	26465924		2.135804824	26463968	26465924	2	0	5.13778E-05	20	Notch1		Notch1
2	52593402		2.111468975	52592425	52593402	2	0	8.64194E-05	10	Cacnb4		Cacnb4
3	55588343		2.074297935	55585408	55589321	2	1	0	10	intergenic		Nbea
6	16801593		2.944214149	16801593	16804523	4	0	0.000312162	30	intergenic		Tfec
6	50146667		2.134992326	50144715	50146667	2	2	2.80003E-05	10	Mpp6		Mpp6
10	126990164		2.068281061	126982374	126992112	2	1	1.11022E-15	20	Ctdsp2		Ctdsp2
11	11696442		2.892341426	11691580	11699359	3	2	5.27104E-06	10	Ikzf1 Gm12000		Ikzf1
15	75086350		2.113341384	75084427	75086350	2	0	1.19033E-05	10	intergenic		Ly6c2
16	4001456		2.469487774	3995721	4003368	3	1	4.83297E-05	10	Slx4		Slx4
16	42964600		3.618228578	42958861	42966513	5	2	0.000116358	20	Zbtb20		Zbtb20
17	29484841		2.118708164	29482910	29485806	2	2	1.11022E-16	10	AC163629.1		AC163629.1
17	56697917		2.071612787	56692121	56701781	2	1	2.67686E-12	20	Ranbp3		Ranbp3
18	67999404		2.053698266	67996510	68000368	2	2	1.9923E-08	10	Ldlrad4		Ldlrad4
X	56737593		2.251475302	56710744	56744306	3	1	0.000183504	70	Gm26312 Fhl1		Fhl1
X	143483102		2.091404193	143479273	143483102	2	2	1.46892E-09	20	intergenic		Pak3
Excluded as recurrent CIS on multiple screens												
2	98661681	98670460	20.96-21.95	98485185	98828371	29	20	0	10-100	Gm13806 Gm10801 Gm10800	Gm10801 Gm10800	Gm10801
5	28166020	28170958	16.41-18.64	27990953	28332290	24	17	0	10-100	AC156021.1 Insig1 En2 Cnpy1 Rbm33	AC156021.1 Insig1 En2 Cnpy1	En2
9	3002194	3012785	3.2-4.37	2984170	3050992	5	4	0	10-100	AC131780.1 Gm10722 Gm11168 Gm10721 Gm10720 Gm10719 Gm10718 Gm10717 Gm17535 Gm10715	AC131780.1 Gm10722 Gm11168 Gm10721 Gm10720 Gm10719 Gm10718 Gm10717	AC131780.1
11	3141162	3152417	4.57-5.38	3086874	3167873	5	4	0	10-70, 100	Pisd-ps1 Sfi1 Gm11399 Gm11400	Pisd-ps1 Sfi1 Gm11399 Gm11400	Sfi1
Excluded as based on 9.1b and 9.1d												
10	118512112	118520696	2.35-3.99	118489614	118546557	4	2	8.96386E-05	20, 50-90	intergenic	intergenic	Ifng
16	5012733		2.437303221	5006042	5016556	4	2	2.69012E-05	10	Gm5480 Rogdi Glyr1		Rogdi
18	35565423	35568643	2.83-3.14	35544102	35585705	3	2	0.000197264	30, 50-70	AC121821.1 Mir1949 Snora74a Matr3	Matr3	Matr3
19	9880044		2.139481557	9876283	9880044	3	0	0.000368983	20	Incenp		Incenp
2	180206483		2.113283703	180204528	180207461	2	2	0	10	Lama5		Lama5
3	65474764	65474833	2.09-2.15	65468961	65474833	2	2	1.11022E-16	10, 20	intergenic	intergenic	4931440P22Rik
4	95019375		2.19050821	95015482	95021322	2	2	4.10707E-05	20	Gm12694		Gm12694
Excluded as low number integrations on review												
5	124384575	124417649	3.15-4.24	124355318	124417649	5	1	0.000272068	60, 80, 100	Sbno1	Sbno1	Sbno1
8	80322470	80332874	2.07-3.16	80314222	80332874	4	1	0.000135844	10, 50-60	intergenic	intergenic	Gypa
6	41593147	41602388	2.16-3.23	41577588	41602920	5	0	4.75014E-06	10, 20, 50, 60	intergenic	intergenic	Ephb6
15	50848947		3.06-3.34	50844125	50853768	5	0	0.000339393	40, 50	Trps1	Trps1	Trps1

Appendix5Bii: Excluded CIS integrations from the 'top 100 analysis ' of the TraDIS data.

Appendix 6A: CIS analysis in the *Vk*hPB* and *Vk*MYC-TA-hPB* cohorts

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
13	37616615	38065245	Rreb1 Ssr1 Cage1 Rick1	Rreb1	30	23	0	10-100
16	23824481	24294739	Sst Rtp2 Bcl6	Bcl6	11	10	0	10-100
4	44508363	44855905	Pax5 Mir5120 Gm12462 Gm12463 Zcchc7	Gm12463/Zcchc7	7	6	0	10-100
2	18612595	18856281	Gm13355 Gm13352 Commd3 Bmi1 Gm13334 BC061194 Gm20539 Gm13333 RP23-396N6.8 Pip4k2a	BC061194	4	4	0	30, 60-80, 100
6	98995326	99242513	Foxp1	Foxp1	4	4	0	30, 50-100
4	32301533	32489946	Bach2	Bach2	3	3	0	60-100
7	24979620	25265769	Atp1a3 Grik5 Zfp574 Pou2f2 D930028M14Rik Dedd2 Zfp526 Gsk3a 9130221H12Rik Erf	Pou2f2	3	3	0	10-100
11	44638767	44754313	Ebf1 Gm12158	Ebf1	3	3	0.0001634	70
16	24137274	24252092	intergenic (Sst Rtp2 Bcl6)	AC116484.1	3	3	0	10-60
18	60661919	60934217	Ndst1 Rps14 Gm8731 Cd74 Mir5107 Tcof1 Arsi Camk2a	Cd74	3	3	0	10-100
18	80498405	80737697	Nfatc1 Atp9b	Nfatc1	3	3	0	10-100
5	147329096	147392896	Flt3	Flt3	3	2	0	30-60, 80, 90
11	17161778	17185094	Ppp3r1 Wdr92	Ppp3r1	3	2	0.000129	20, 80
11	62439680	62439680	Ncor1	Ncor1	3	2	0.0036251	80
1	39896022	40131226	Map4k4 AC161534.1 Il1r2	Il1r2	2	2	1.119E-06	10, 80, 90
1	58669859	58767732	Als2cr12 Cflar	Cflar	2	2	0	20, 40, 50, 70, 80, 100
2	163159736	163357047	Tox2 Jph2	Tox2	2	2	0	10-100
4	138025236	138076416	Eif4g3	Eif4g3	2	2	0	10, 40, 60, 70, 90, 100
5	23441479	23521061	Mll5 Gm25219 Srpk2	Mll5	2	2	0.0014163	70
5	72795165	72867478	Tec	Tec	2	2	1.13E-08	30, 60, 80, 90
6	113054963	113131765	Thumpd3 Gm22591 Gt(ROSA)26Sor Setd5	Setd5	2	2	7.965E-05	30, 70
9	44343501	44526252	Hmbs Vps11 Gm22141 Gm10080 Hyou1 Gm26306 Slc37a4 Trappc4 Rps25 Ccdc84 Foxr1 Upk2 Gm9830 Gm22540 C030014I23Rik Bcl9l Cxcr5	Gm9830	2	2	1.295E-06	40, 60-80, 100
17	47686315	47825200	Usp49 Tomm6 Gm21981 Gm14872 Prickle4 Frs3 Gm14873 Pgc Tfeb Mdfi	Tfeb	2	2	1.11E-16	30, 40, 60, 70, 90, 100
17	80405710	80451144	Sos1	Sos1	2	2	2.22E-16	10, 30, 50, 60, 80
18	49990796	50042738	Tnfaip8 C030005K06Rik	Tnfaip8	2	2	1.11E-16	30-50, 90
18	65406503	65484149	Gm22567 Malt1	Malt1	2	2	0	10, 40, 50, 80-100
X	52886989	52931935	Rps2-ps13 Phf6	Phf6	2	2	0	20-40, 60-80, 100
2	163620641	163623882	Serinc3	Serinc3	1	1	0.0016047	40
3	94960845	94962472	Rfx5 B230398E01Rik	B230398E01Rik	1	1	9.259E-14	20
3	100478262	100483152	Fam46c	Fam46c	1	1	0.0017189	60
3	101273850	101275477	intergenic	Cd2	1	1	1.945E-11	20
6	129180775	129181687	Clec2d	Clec2d	1	1	0.0011947	10
7	35636231	35639797	Ankrd27	Ankrd27	1	1	0.0030711	50
7	81522887	81523599	mmu-mir-1839	mmu-mir-1839	1	1	0.0012525	10
8	82317320	82318723	intergenic	Il15	1	1	0.0008985	10
8	87859028	87866085	Zfp423	Zfp423	1	1	0.0003767	100
10	17875614	17876319	intergenic	Heca	1	1	0.0002276	10
10	78379851	78386225	intergenic	Gm10146	1	1	1.239E-08	90
12	107858133	107865517	intergenic	Bcl11b	1	1	0	60
13	93170173	93170893	Papd4	Papd4	1	1	0.0015046	10
14	15046049	15048455	intergenic	Nek10	1	1	0.0051263	30
14	75188183	75190589	Lcp1	Lcp1	1	1	6.09E-05	30
15	97720057	97728761	Endou	Endou	1	1	1.11E-16	10, 70
16	8642452	8643767	Pmm2	Pmm2	1	1	0.0019735	20
16	10485822	10487137	intergenic	Cilta	1	1	0.0009845	20
17	45553024	45555498	intergenic	Nfkbie	1	1	0.0006535	30
19	32766562	32767221	Pten	Pten	1	1	4.816E-05	10

CIS analysis using the top 10 hits in the *hPB* cohort. The start and end boundaries encompass all analysis windows in which each locus was identified as a CIS. The gene shown as nearest to peak was the central gene in the majority of kernel windows (scales) detecting the CIS, but is not necessarily the target gene for the CIS. Due to local hopping the total number of insertions occasionally included multiple integrations from the same tumour, so both the total number and the number after correction for local hopping are shown. The smallest p value identified at any scale is shown along with the analysis scales at which the CIS was detected (x1000).

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
13	37624377	38058475	Rreb1 Ssr1 Cage1 Riok1	Rreb1	28	25	0	10-100
16	23863733	24281043	Sat Rtp2 Bcl6	Bcl6	9	9	0	10-100
4	32237229	32561406	Bach2 D130062J21Rik Gm11932 Gm24371 BC024582	Bach2	9	7	0	10-100
11	3055838	3302391	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1 Patz1	Sfi1	8	7	0	10-100
1	58578786	58869374	Fam126b Ndufb3 Gm10068 Als2cr12 Cflar Casp8	Cflar	7	7	0	10-100
5	147214584	147483494	2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 Gm6054 Pan3	Flt3	5	5	0	10-100
9	32500946	32758851	Fli1 Ets1	Ets1	5	5	0	10, 30-100
3	102947840	103164316	Nr1h5 Gm22826 Sike1 Csde1 Nras Ampd1 Gm23820 Dennd2c	Csde1	4	4	0	10-100
3	135562444	135765833	Manba Nfkb1 Gm9799	Nfkb1	4	4	0	10-100
16	36537254	36700550	Casr Cd86 Ildr1	Cd86	4	3	0	20-80, 100
6	98939472	99085985	Foxp1	Foxp1	3	3	0	10, 30, 50-70, 90, 100
6	99249196	99440799	Foxp1 Gm20696 Gm20705	Foxp1	3	3	0	30-100
11	109337893	109443336	Gm11696 Gna13 9930022D16Rik Amz2 Gm15642	Gna13	3	3	5.19E-08	80
15	61885470	62063475	Myc Pvt1	Myc	3	3	0	10-100
17	46746591	46933407	CT030702.1 Cnpy3 Ptcra 2310039H08Rik Rpl7l1 Gltscr1l A330017A19Rik Tbcc Gm23797 Prph2 Ubr2	Gltscr1l	3	3	0	10-100
17	75376145	75544280	Ltbp1 Rasgrp3 Fam98a	Rasgrp3	3	3	0	10, 20, 40, 50, 70-100
18	65353747	65532750	Alpk2 Gm22567 Malt1 Gm26114	Malt1	3	3	0	10-100
19	34157427	34309421	Ankrd22 Stambpl1 Acta2 Fas	Acta2	3	3	3.79E-08	60-100
2	33384683	33402619	intergenic	Zbtb34	3	2	3.86E-06	100
3	88452532	88605074	Sema4a Lmna Mex3a Mir1905 Rab25 Lamtor2 Ubqln4 Gm10704 Ssr2	Ssr2/Lmna	3	2	1.17E-08	100
7	110216552	110256898	Swap70 Gm22185	Swap70	3	2	0	10, 40, 60, 90
14	115008961	115058395	Mir17hg Mir17 Mir18 Mir19a Mir20a Mir19b-1 Mir92-1	Mir17hg	3	2	0	30-60, 80, 90
19	6313428	6474630	Cdc42bpg Men1 Map4k2 Gm14966 Gm22278 Sf1 Pygm Rasgrp2 Gm14965 Nrxn2 Gm26470	Rasgrp2	3	2	0	10-100
X	11932951	12160660	Gm14512 Bcor 2908C10Rik	Bcor	3	2	0	10-100
1	11328348	11358325	intergenic	A830018L16Rik	2	2	1.935E-11	10, 30, 50, 80
2	167431368	167440336	Slc9a8	Slc9a8	2	2	0.0012557	100
3	95480560	95511013	Arnt Ctsk	Ctsk	2	2	0.0004163	60
8	72215060	72407694	Fam32a Gm25027 Ap1m1 Gm10282 Klf2 Eps15l1	Klf2	2	2	0	30, 40, 60, 70, 90
8	105150849	105189744	Cbfb Gm22063	Cbfb	2	2	0	20-40, 60-100
9	88439642	88460300	Gm20537 4932427H20Rik Syncrip	Gm20537	2	2	4.766E-08	10, 30
14	121822642	121961940	Ubac2 Gpr18 Gpr183	Ubac2	2	2	0.0002112	40, 80, 100
16	10462546	10571037	Ciita Dexi Clec16a	Ciita	2	2	0	10, 30-50, 70-100
16	55787944	55825832	Nfkbiz	Nfkbiz	2	2	0	10-80, 100
17	23564304	23637450	Zfp13 Zscan10 Mmp25	Zfp13	2	2	0	60
18	34906925	34963881	Etf1 Hspa9 Gm22200 Gm26109	Hspa9	2	2	0	10, 20, 50-80, 100
19	4375007	4410643	Kdm2a	Kdm2a	2	2	4.001E-05	20, 50, 70, 90
2	27325880	27331253	Vav2 AA645442	Vav2	1	1	0.0009574	60
2	167356401	167399416	intergenic	B4galt5	1	1	0.0011415	80
2	167449600	167542799	Slc9a8 Spata2 Rnf114 Gm11474 Snai1	Rnf114	1	1	1.141E-06	80
3	95517352	95525796	intergenic	Ctss	1	1	0.0014022	20
3	101275162	101276006	Cd2	Cd2	1	1	2.068E-06	10
3	131225819	131227508	intergenic	Lef1	1	1	0.0001973	20
7	101398641	101405368	Arap1	Arap1	1	1	1.798E-06	70
7	125595408	125597326	intergenic	Il21r	1	1	0.0013434	20
12	76932603	76934474	intergenic	Max	1	1	7.882E-07	20
12	111168866	111174492	4930595D18Rik Traf3	4930595D18Rik	1	1	0.0008736	60
17	31073581	31074510	Abcg1	Abcg1	1	1	0.0003779	10

CIS analysis using the top 10 hits in the *Vk*MYC-TA-hPB* cohort. The start and end boundaries encompass all analysis windows in which each locus was identified as a CIS. The gene shown as nearest to peak was the central gene in the majority of kernel windows (scales) detecting the CIS, but is not necessarily the target gene for the CIS. Due to local hopping the total number of insertions occasionally included multiple integrations from the same tumour, so both the total number and the number after correction for local hopping are shown. The smallest p value identified at any scale is shown along with the analysis scales at which the CIS was detected (x1000).

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
13	37631295	38046935	Rreb1 Ssr1 Cage1 Riok1	Rreb1	43	26	0	10-100
16	23859243	24291463	Sst Rtp2 Bcl6	Bcl6	30	17	0	10-100
16	17001356	17231874	Ypel1 Gm9974 Ppil2 Gm15585 2610318N02Rik Mir130b Mir301b Sdf2l1 Ccdc116 Gm15646 Ydjc Ube2l3	Ube2l3	14	12	0	10-100
4	32162214	32543477	Gm11929 Bach2 D130062J21Rik Gm11932 Gm24371	Bach2	18	11	0	10-100
4	44528816	44968734	Pax5 Mir5120 Gm12462 Gm12463 Zcchc7 Gm22639 Gm12678 Gm12493	Zcchc7	14	11	0	10-100
6	98989498	99390113	Foxp1 Gm20696 Gm20705	Foxp1	12	10	9.533E-11	20, 30, 60-100
11	3083642	3295850	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1 Patz1	Sfi1	14	9	0	10-100
11	44623794	44731292	Ebf1 Gm12158	Ebf1	11	9	0.0002058	90, 100
10	13971899	14219779	Hivep2 AC158608.1	Hivep2	9	8	0	40-100
1	58564955	58830706	Fam126b Ndufb3 Gm10068 Als2cr12 Cflar Casp8	Cflar	8	8	0	10-100
11	86626041	86753084	Vmp1 Gm11478 Ptrh2 Cltc	Cltc	8	8	0.0002	40, 70, 90, 100
18	49873916	50106787	Dmxl1 Tnfaip8 C030005K06Rik Tnfaip8	Tnfaip8	8	8	0	10-100
2	170003847	170239141	Tshz2 AL731822.1 Zfp217 AL844576.1	Zfp217	13	7	0	10-100
18	65351617	65534873	Alpk2 Gm22567 Malt1 Gm26114	Malt1	9	7	0	10-100
X	52807880	53036994	Gm14607 Gm6539 Rps2-ps13 Phf6 Hprt	Phf6	9	7	0	10-100
1	138043905	138289970	Ptprc Atp6v1g3	Ptprc	8	7	0	30-100
11	62351127	62497715	Ncor1 Pigl Gm12278	Ncor1	11	6	0	10-100
4	6822900	7008859	Tox	Tox	9	6	0	10, 20, 40-100
1	54785399	54949545	Ankrd44	Ankrd44	7	6	0	20-100
3	135540681	135779228	Manba Nfkb1 Gm9799	Nfkb1	6	6	0	10-100
14	76575405	76760351	intergenic	Serp2	6	6	0	10-40, 60-100
18	60873300	60906171	Ndst1 Rps14 Gm8731 Cd74 Mir5107 Tcof1	Cd74	6	6	0	10-100
13	44678280	44858001	Jarid2 Gm22213	Jarid2	10	5	0	10, 30-100
5	147260198	147465893	2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 Gm6054 Pan3	Flt3	9	5	0	10-100
18	80554647	80736453	Nfatc1	Nfatc1	9	5	0	10-100
2	61553867	61658017	Tank	Tank	6	5	2.978E-10	30-100
3	60426844	60627224	Mbnl1	Mbnl1	6	5	0	20, 30, 60-100
3	138908972	139118893	Rap1gds1	Rap1gds1	6	5	0	20-60, 80-100
6	129092979	129263511	Clec2e Klrb1-ps1 Gm26160 Clec2d AC142191.1	AC142191.1	6	5	0	10-30, 50-100
2	163987014	164124206	Ywhab Pabpc1l Tomm34 Stk4	Tomm34	5	5	0	10, 30-100
4	134673032	134878328	Man1c1 Ldlrap1 Gm25751 Tmem57 Rhd	Ldlrap1	5	5	0	10-100
7	24983081	25236202	Atp1a3 Grik5 Zfp574 Pou2f2 D930028M14Rik Dedd2 Zfp526 Gsk3a 9130221H12Rik	Pou2f2	5	5	0	10-100
7	80075460	80326626	Zfp710 Idh2 Gm24012 Mir1965 Sema4b Cib1 Gdpgp1 Gm15504 Ttll13 Ngm Vps33b Prc1 AC109232.1 Rcod1 Uno45a	Sema4b	5	5	0	10-100
9	44441602	44657005	Upk2 Gm9830 Gm22540 C030014I23Rik Bcl9l Cxcr5 Ddx6	Cxcr5	5	5	5.709E-11	60-100
10	18900904	19101115	Tnfaip3	Tnfaip3	5	5	0	10-100
10	68086087	68276764	Arid5b	Arid5b	5	5	1.012E-11	20, 50-100
11	20026387	20128287	Actr2	Actr2	5	5	0	10-100
15	63961054	64090574	Fam49b Gm25628 Asap1	Fam49b	5	5	0	10, 20, 40-100
15	96315030	96463187	Arid2 Gm25397 Scaf11	Arid2	5	5	5.926E-05	50-100
16	20082219	20110676	Klhl24	Klhl24	5	5	1.677E-08	10, 30, 40, 90
17	17503487	17647345	Lnpep	Lnpep	5	5	0	20, 30, 50-100
17	34134185	34294435	H2-DMa H2-DMb2 H2-DMb1 Psmb9 Tap1 Psmb8 Gm20496 Tap2 Gm15821 H2-Ob Gm20506 H2-Ab1 H2-Aa Gm20513	H2-Ob	5	5	3.095E-05	60-100
X	38484044	38657041	Cul4b Mcts1 C1galt1c1	Cul4b	5	5	0	30, 50-100
2	98607167	98709956	Gm10801 Gm10800	Gm10800	7	4	0	10, 20, 40-100
3	101161215	101361595	Gm12486 Tpt1-ps1 Gm12490 Cd2 Gm10355	Cd2	6	4	0	10-100
11	17140025	17202381	Ppp3r1 Wdr92	Ppp3r1	5	4	0	10-70
1	80285208	80382419	Cul3	Cul3	4	4	0.0001432	20, 40-100
2	18856573	19006456	Pip4k2a 4930426L09Rik	Pip4k2a	4	4	1.067E-07	20, 30, 50-90
2	45004516	45130230	Zeb2 Gm13476	Zeb2	4	4	0	40-80, 100
2	57218345	57304521	A930012O16Rik Gpd2 Gm13535	Gpd2	4	4	4.677E-05	20, 30, 50-100
2	163565640	163712631	Hnf4a Ttpal Serinc3 0610039K10Rik Pkig Gm16316	0610039K10Rik	4	4	0	20-100
2	165936596	166077663	Gm11462 Gm11463 Gm11464 Ncoa3 Sulf2	Ncoa3	4	4	0	10-100
5	23425205	23549291	5031425E22Rik Mll5 Gm25219 Srpk2	Mll5	4	4	2.115E-08	20, 50, 70-100
6	98951602	98999201	Foxp1	Foxp1	4	4	1E-06	20-60
7	27460972	27652163	Blvrb Pgam1-ps2 Sertad3 Sertad1 Prx Gm15541 Hipk4 Pld3 2310022A10Rik Akt2	Akt2	4	4	0	10-100
7	128366459	128521590	Rgs10 Gm15503 Tial1 Gm24365	Tial1	4	4	1.447E-05	40, 50, 70-100
10	24096858	24230332	Taar8c Taar9 Gm15137 Stx7 Gm23051 Moxd1	Stx7	4	4	0	10-100
11	115641082	115687895	Grb2 Gm11702	Grb2	4	4	0.0001701	50, 60
12	76317716	76446880	Mthfd1 Akap5 Gm23809 Zbtb25 Zbtb1 AC124453.1 Hspa2 Ppp1r36 Gm25563 Gm10451	Zbtb1	4	4	0	10-100
13	13562664	13661221	Lyst	Lyst	4	4	0.0002391	60-100
14	31337607	31453720	Capn7 Sh3bp5	Sh3bp5	4	4	0	10-100

Top 100 CIS analysis for the *Vk* hPB* cohort

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
14	52047277	52150505	Gm22354 Hnrnpc Rpgrip1	Rpgrip1	4	4	0.0001649	30, 40, 60-100
14	121885297	122001108	Ubac2 Gpr18 Gpr183	Ubac2	4	4	5.035E-08	10, 30, 70-100
16	44183346	44237757	Gm608	Gm608	4	4	0	10-60
X	11966280	12186907	Gm14512 Bcor 2900008C10Rik Gm14521	Bcor	4	4	0	10-40, 60-100
8	84805604	84864152	Dand5 Gadd45gip1 Rad23a Calr 1700122E12Rik Farsa	Calr	4	3	0.0004937	100
12	3762122	3790779	Dtnb Gm20448	Dtnb	4	3	0.0004999	60
14	114996533	115062680	Mir17hg Mir17 Mir18 Mir19a Mir20a Mir19b-1 Mir92-1	Mir17hg	4	3	0	10-40, 60-100
17	35948636	36002251	Gm8801 Abcf1 Mir877 Prr3 Gnl1 Gm20508 A930015D03Rik Gm17414	Gnl1/A930015D03Rik	4	3	1.62E-05	10, 60
17	88025946	88082349	Fbxo11	Fbxo11	4	3	2.503E-09	10, 30, 50-70
X	77573612	77685726	Tbl1x Gm23121 Gm6927	Tbl1x	4	3	0	20-100
1	86441861	86521838	intergenic	Ptma	3	3	0	10-100
1	178297668	178344854	B230369F24Rik Cox20 Gm16586 Hnrnpu	Hnrnpu	3	3	3.443E-05	40, 60-80
2	44316687	44350842	Arhgap15	Arhgap15	3	3	2.755E-06	10-30, 50-70
2	163206247	163294322	Tox2	Tox2	3	3	7.55E-15	10-90
3	22125657	22201992	Tbl1xr1	Tbl1xr1	3	3	1.135E-05	20, 30, 50, 70-100
4	154850558	155000070	Ttc34 Gm13112 Mmel1 Fam213b Tnfrsf14 Gm20421 Hes5 Pank4 Plch2	Tnfrsf14	3	3	0	10-100
5	124462994	124551113	Rilpl2 Gm15621 Snmp35 Rilpl1 Tmed2	Rilpl1	3	3	0	10, 20, 40, 50, 80-100
5	139319937	139408057	Adap1 Cox19 Cyp2w1 3110082I17Rik Mir339 Gpr146 D830046C22Rik C130050O18Rik	3110082I17Rik	3	3	0.0003036	80-100
6	31047063	31118227	Mir29a Mir29b-1 RP23-459L15.5 Gm13834 Gm13833	Gm13834	3	3	0	10-50, 70-100
6	88441116	88524981	Eefsec Ruvbl1 Sec61a1	Ruvbl1	3	3	6.439E-15	10-100
6	108622048	108713537	0610040F04Rik Gm17055 Bhlhe40	Bhlhe40	3	3	2.859E-13	10-100
6	113074509	113127878	Gt(ROSA)26Sor Setd5	Setd5	3	3	0.0002616	70-90
6	127161199	127272713	AC163747.1 Gm4968	AC163747.1	3	3	0	10-50, 70-100
7	55870242	55919848	Cyfip1 Gm17907	Cyfip1	3	3	9.933E-05	40, 90
7	88266311	88310521	Ctsc AC124322.1	Ctsc	3	3	0	10-40, 60, 80, 90
7	101408504	101530865	Arap1 Pde2a Mir139	Pde2a	3	3	0	10-60, 80-100
7	121961555	122022388	Cog7 Gga2	Gga2	3	3	0.0005918	90000
9	32678724	32756584	Ets1	Ets1	3	3	0	10-30, 50-100
10	121385462	121478759	Gns Rassf3	Rassf3	3	3	5.016E-05	40, 70-90
11	22822155	22849435	Gm12057 Commd1 B3gnt2 Gm20456 Gm23772	Gm12057	3	3	4.406E-13	20-50
11	54759778	54783161	Cdc42se2	Cdc42se2	3	3	0.0001665	40
11	69062925	69089813	Tmem107 Snord118 Gm12306 Gm25371 Vamp2	Tmem107	3	3	0.0003239	40, 50
11	103217687	103267541	Spata32 Map3k14	Map3k14	3	3	0	10-50
12	71015257	71055377	3110056K07Rik Arid4a	Arid4a	3	3	4.547E-06	30, 40, 60-80
12	84167783	84225068	Elmsan1	Elmsan1	3	3	0	10-90
13	24600626	24660044	Fam65b Gm11346 AL513014.1	Fam65b	3	3	3.824E-05	30, 40, 60, 70
13	55225968	55282527	Nsd1	Nsd1	3	3	0.0002778	60-90
13	114114766	114182636	Arl15	Arl15	3	3	6.406E-06	10, 20, 40-90
15	12141435	12152927	Zfr	Zfr	3	3	0.0004746	60
15	80525845	80600344	Enthd1	Grap2	3	3	0	10-90
15	96723871	96769541	intergenic	Gm8888	3	3	5.586E-05	40, 60
15	103216952	103285905	Cbx5 Hnrnpa1 Nfe2 Copz1 Mir148b	Nfe2	3	3	4.462E-12	10, 20, 40-90
17	3031580	3104904	Gm10232 Gm16569 Gm5479 Gm25909 Pisd-ps2	Pisd-ps2	3	3	0	10, 30-80
17	5045583	5098593	Arid1b	Arid1b	3	3	2.709E-08	30-70
17	80422760	80462242	Sos1	Sos1	3	3	1.042E-08	10-70
17	83383941	83417783	Eml4	Eml4	3	3	1.188E-08	10-70
18	4303849	4350385	Map3k8	Map3k8	3	3	5.229E-06	30-70
18	34497513	34505233	Fam13b	Fam13b	3	3	8.306E-10	20
18	56692100	56743580	Lmnb1	Lmnb1	3	3	0	10-80
19	5658200	5756891	Sipa1 Pcnxl3 Map3k11 Kcnk7 Ehbp1l1 Gm16538 Fam89b Sssca1 Ltbp3	Ehbp1l1	3	3	0	10, 30-60, 80-100
19	47394832	47480281	Sh3pxd2a	Sh3pxd2a	3	3	0	10-80
1	138297320	138299284	intergenic	Atp6v1g3	2	2	4.725E-06	10
2	29633002	29640825	Rapgef1	Rapgef1	2	2	4.848E-05	40
2	128139757	128141764	Bcl2l11	Bcl2l11	2	2	1.478E-07	10, 20
4	3679034	3682930	Lyn AL772401.1	Lyn	2	2	1.289E-12	10
4	89278557	89282652	Cdkn2a	Cdkn2a	2	2	4.21E-08	10-30
4	138050803	138054699	Eif4g3	Eif4g3	2	2	2.22E-16	10, 20
5	148326668	148330577	Slc7a1	Slc7a1	2	2	2.029E-07	40
6	72510858	72529870	Sh2d6	Sh2d6	2	2	0	10, 20, 50
7	19809875	19825853	Bcl3 Gm16175 Gm16174	Bcl3	2	2	0	10, 40
7	28408589	28417957	Samd4b	Samd4b	2	2	8.882E-14	10, 30, 40
8	13094016	13108003	Pcid2 Cul4a	Pcid2	2	2	3.21E-11	10-50
8	25585755	25593539	Letm2	Letm2	2	2	0.0001626	40
8	45988083	45991971	Ufsp2	Ufsp2	2	2	5.249E-05	10, 20
8	105998484	106008218	Dus2l	Dus2l	2	2	1.668E-07	30-50
9	66421892	66422824	Herc1	Herc1	2	2	2.247E-12	10
9	72017621	72022328	Tcf12	Tcf12	2	2	0.0002163	10, 20
14	101462295	101469824	Tbc1d4	Tbc1d4	2	2	3.746E-08	10, 30
15	38077393	38078348	Ubr5	Ubr5	2	2	0	10
17	19433362	19435236	intergenic	Vmn2r-ps121	2	2	9.739E-05	10
17	45548216	45565129	Nfkbie Slc35b2	Nfkbie	2	2	0	10, 20
17	46856114	46858924	intergenic	Gltscr1l	2	2	2.442E-15	10
18	67240763	67242692	Mppe1	Mppe1	2	2	6.867E-06	10
X	18164384	18258541	Kdm6a	Kdm6a	2	2	0	20-80, 100
X	48390733	48444672	Bcorl1 Elf4	Elf4	2	2	0	10-30, 70, 80, 100
X	134572730	134577437	Btk	Btk	2	2	0.0003822	50
1	75476608	75477590	Chpf	Chpf	1	1	6.157E-09	10
8	94999314	95000286	Gpr56	Gpr56	1	1	1.518E-06	10
14	25530577	25533869	Zmiz1	Zmiz1	1	1	1.11E-16	20, 30
14	31155243	31156159	Stab1	Stab1	1	1	1.354E-14	10

CIS identified using the top 100 integrations in the *Vk*hPB* cohort

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
13	37406200	38068534	Ly86 Rreb1 Ssr1 Cage1 Riok1	Rreb1	94	56	0	10-100
11	3016121	3354935	Pisd-ps1 Sfi1 Gm11399 Gm11400 Drg1 Gm12735 Fau-ps2 Eif4enif1 Patz1 Gm12592 Gm11944 Pik3ip1 Limk2	Sfi1	53	33	0	10-100
16	23850461	24316468	Sst Rtp2 Bcl6	Bcl6	47	25	0	10-100
4	32157117	32578306	Gm11929 Bach2 D130062J21Rik Gm11932 Gm24371 BC024582	Bach2	40	25	0	10-100
6	98908797	99565200	Foxp1 Gm20696 Gm20705 Gm22328	Foxp1	37	24	0	10-100
5	147180300	147522254	Gsx1 Gm24556 2210019I11Rik Pdx1 RP24-510G5.4 Cdx2 Prhoxnb Flt3 AC134441.1 Gm6054 Pan3	Flt3	34	19	0	10-100
1	58521691	58888508	Gm15834 Fam126b Ndufb3 Gm10068 Als2cr12 Cflar Casp8	Cflar	32	25	0	10-100
15	61862678	62244510	Myc Pvt1 H2afy3	Myc	21	14	0	10-100
11	61013919	61238882	Kcnj12 Tnfrsf13b Gm12269 Usp22 Rps13-ps5 Aldh3a1 Aldh3a2	Tnfrsf13b	21	10	0	10-100
2	44886061	45190365	Gtdc1 Zeb2 Mir5129 Gm13476	Zeb2	20	16	0	10-100
9	32472788	32781639	Fli1 Ets1	Ets1	19	13	0	10-100
2	170004430	170269470	Tshz2 AL731822.1 Zfp217 AL844576.1	Zfp217	18	12	0	10-100
3	135525355	135795415	Manba Oaz2-ps Nfkb1 Gm9799	Nfkb1	18	11	0	10-100
7	110068190	110323094	Zfp143 Wee1 Swap70 Gm22185 Sbf2	Swap70	14	12	0	10-100
11	62265887	62490850	Adora2b Zswim7 Ttc19 Gm12275 Ncor1 Gm12276 Pigl Gm12278	Ncor1	13	11	0	10-100
X	11996100	12154043	Bcor 2900008C10Rik	Bcor	13	4	9.992E-16	10-100
3	102913279	103164714	Sycp1 Nr1h5 Gm22826 Sike1 Csde1 Nras Ampd1 Gm23820 Dennd2c	Csde1	12	11	0	10-100
2	98551691	98757833	Gm13806 Gm10801 Gm10800	Gm10800	12	10	0	10-100
11	100754115	100959516	Kcnh4 Hcrt Ghdc Gm24358 Stat5b Stat5a Stat3 Ptrf	Stat5a	12	9	0	10-100
16	24472671	24490144	Lpp	Lpp	12	5	0.000274	60
10	68021190	68285437	Rtkn2 Gm16212 Arid5b	Arid5b	11	10	0	10-100
14	114929360	115151053	Gm24073 Mir17hg Mir17 Mir18 Mir19a Mir20a Mir19b-1 Mir92-1 Gpc5	Mir17hg	11	9	0	10-100
16	36549167	36714211	Casr Cd86 Ildr1	Cd86	11	9	4.754E-09	10-100
16	43529572	43638277	Zbtb20 Zbtb20	Zbtb20	11	8	0.0001117	30, 40, 80-100
19	37366712	37564885	Kif11 Hhex Gm23026 Exoc6	Gm23026	11	8	0	10-100
6	148913009	149148138	Fam60a 3010003L21Rik Gm23395 Gm25539 Dennd5b AC140327.1 Gm15779 Gm15781 Gm22578 Gm15780 Gm23462 Gm10203 Mettl20 Gm10011	Dennd5b	10	10	1.11E-16	10-100
10	18890914	19125800	Tnfaip3	Tnfaip3	10	10	0	10-100
11	44493813	44699214	Rnf145 4930597A21Rik Ebf1 Ebf1 Gm12158	930597A21Ri	10	9	1.11E-16	10, 20, 50-100
17	5185325	5424715	Arid1b Tmem242 Gm22475	Arid1b	10	9	9.601E-06	50-100
11	88928888	89056041	4930405P13Rik Scpep1 Gm26471 Gm15698 Coil 2210409E12Rik Gm11496 Trim25 Dgke Gm24974	Trim25	10	8	5.525E-05	50-100
11	109292926	109468984	Rgs9 Gm11696 Gna13 9930022D16Rik Amz2 Gm15642 Slc16a6 Gm25540	Gna13	10	8	0	10-100
X	20125306	20398134	Slc9a7 Gm9083 Gm14529 Rp2h Gm9085 Gm14537	Slc9a7	10	8	0	10-100
4	44846231	45006757	Zcchc7 Gm22639 Gm12678 Gm12493 Gm12679 Grhpr Zbtb5	Gm12678	9	9	0	10, 30-100
10	108333463	108392184	Pawr Gm23105	Pawr	9	9	0.0002443	90, 100
16	49830383	50005136	Cd47 Cd47	Cd47	9	9	3.31E-06	10-100
11	88459400	88664801	Msi2	Msi2	9	8	0	20-100
17	75374410	75527620	Ltbp1 Rasgrp3	Rasgrp3	9	7	0	10-100
2	18656976	18698107	Commd3 Bmi1 Gm13334	Bmi1	9	4	0	10-50
2	163977231	164134292	Ywhab Pabpc1l Tomm34 Stk4	Stk4	8	8	0	20-100
4	3614197	3800304	Tgs1 Gm22541 2210414B05Rik Lyn AL772401.1 Gm11805 Gm22781	Lyn	8	8	0	10-100
10	13948526	14124690	Hivep2 AC158608.1	Hivep2	8	8	0	20-100
15	98837538	99032605	Mll2 Rhebl1 Dhh Lmbr1l Tuba1b Tuba1a AC157610.1 Gm8973 Tuba1c	Tuba1a	8	8	0	10, 30-100
17	29382845	29574357	AC163629.1 Pim1 Gm17657 Tmem217 Tbc1d22b Ftsjd2 Gm25932	AC163629.1	8	8	3.423E-08	30-100
18	65347674	65590538	Alpk2 Gm22567 Malt1 Gm26114 Zfp532	Malt1	8	8	0	10-100
19	44208392	44434875	Scd3 Gm25573 Scd2 Mir5114 Scd4 Scd1	Scd4	8	8	0	10-100
4	46506937	46702839	Trim14 Gm16731 Coro2a Tbc1d2 Gabbr2	Tbc1d2	8	7	0	10-100
4	135789267	135955784	Myom3 Gm13000 Srsf10 Pnrc2 Gm13006 Cnr2 Fuca1 Hmgcl	Gm13006	8	7	0	10-100
11	115601670	115689699	Gga3 Gm25364 Mrps7 Mif4gd Slc25a19 AL645470.1 Grb2 Gm11702	Grb2	8	7	0	10-100
14	70138277	70326278	2610301G19Rik 9930012K11Rik Pdlim2 Sorbs3 AC151836.1 Ppp3cc Slc39a14	Ppp3cc	8	7	2.524E-05	20, 40-100
15	97316019	97442812	Pced1b	Pced1b	8	7	0	10-100
17	34189792	34362152	Tap1 Psmb8 Gm20496 Tap2 Gm15821 H2-Ob Gm20506 H2-Ab1 H2-Aa Gm20513 H2-Eb1 H2-Eb2 H2-Ea-ps Btnl2	Gm20513	8	7	0	10-100
19	4243544	4407802	Ssh3 Ankrd13d Adrbk1 Kdm2a	Kdm2a	8	7	0	10-40, 70-100
8	72222010	72408911	Gm25027 Ap1m1 Gm10282 Klf2 Eps15l1	Gm10282	8	6	0	10, 30-100
1	175806574	175914455	Wdr64 Exo1	Exo1	7	7	5.696E-05	10-100

Top 100 CIS analysis for the *Vk* MYC-TA-hPB* cohort

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
2	30976385	31182527	BC005624 Usp20 Fnbp1 D330023K18Rik Gpr107	Fnbp1	7	7	0	20-100
3	89959710	90155271	Atp8b2 Gm24046 Hax1 Ubap2l Gm24608 4933434E20Rik Gm16540 1700094D03Rik Mir190b Tpm3 Nup210l Gm23723	Tpm3	7	7	0	30-100
4	101144001	101287620	Raver2 Jak1 Gm24468 Gm12785 Gm25124 Gm12801	Jak1	7	7	0	10-100
12	84095746	84281330	Acot6 Dnalc1 Pnma1 Elmsan1 Gm5436 Gm23399	Elmsan1	7	7	0	10-100
16	4501438	4656774	Srl Gm15885 Tfap4 Glis2 Pam16 Coro7 Vasn Dnaja3	Glis2	7	7	0	10, 30-100
16	55790823	55878971	Nfkbiz Nxpe3	Nxpe3	7	7	2.204E-09	20-100
18	53602766	53700860	Cep120	Cep120	7	7	0	10-100
18	56701712	56857145	Lmnb1 March3 Gm15345	March3	7	7	1.086E-06	20-100
19	32706826	32805685	Atad1 Pten	Pten	7	7	7.519E-11	10-90
16	38430619	38523703	Pla1a Adprh AC209577.1 Cd80 Timmdc1 Gm15953	Cd80	7	6	6.68E-05	60-90
13	32996598	33035439	Serpinb9 Serpinb9b	Serpinb9	7	5	9.119E-05	30-50
2	33384539	33431562	Zbtb34	Zbtb34	7	4	2.253E-08	10-50
2	131953952	132074305	Rassf2 Slc23a2	Rassf2	6	6	0	10-100
4	154821145	155023699	Gm13113 Ttc34 Gm13112 Mmel1 Fam213b Tnfrsf14 Gm20421 Hes5 Pank4 Plch2	Tnfrsf14	6	6	0	10-100
5	64574032	64671733	Gm25306	Gm25306	6	6	1.365E-05	40-100
7	24998900	25194979	Atp1a3 Grik5 Zfp574 Pou2f2 D930028M14Rik	Pou2f2	6	6	0	10-100
7	73469866	73616926	Chd2	Chd2	6	6	0	10, 20, 40-100
7	101156356	101283808	Fchsd2 Gm15673	Fchsd2	6	6	0	10-100
8	126758661	126914832	intergenic	Tomm20	6	6	0	10, 40-100
10	67047049	67117244	Reep3 Jmjd1c	Reep3	6	6	4.538E-05	30-50
10	116485957	116613186	Cnot2 Gm25190 5330438D12Rik	Cnot2	6	6	0	10-100
13	112757086	112834352	Ppap2a	Ppap2a	6	6	4.074E-06	20-100
16	10471364	10608079	Ciita Dexi Clec16a	Ciita	6	6	7.764E-12	10-30, 60-100
19	41490594	41509467	Lcor	Lcor	6	6	0.0007105	100
X	48282962	48509619	AL672246.1 Gm22528 Gm7212 Gm22612 Bcorl1 Gm23868 Elf4 Aifm1	Elf4	6	6	0	10-100
3	15092332	15147922	intergenic	RP23-3D20.1	6	5	9.803E-14	10-70
17	46762543	46867707	CT030702.1 Ptcra 2310039H08Rik Rpl7l1 Gltscr1l	Gltscr1l	6	5	0	10, 30-100
11	103232095	103270106	Map3k14 1700028N14Rik	Map3k14	6	4	2.958E-06	20-50
19	23122809	23149458	2410080I02Rik Klf9	2410080I02Rik	6	3	1.862E-07	10-30
1	130766427	130864401	Fcamr Gm15848 Pigr	Pigr	5	5	0	10-100
1	131939530	131947248	intergenic	Nucks1	5	5	0.0001735	80
2	165965542	166071525	Gm11463 Gm11464 Ncoa3	Ncoa3	5	5	2.794E-05	30-100
4	24481982	24547094	Mms22l	Mms22l	5	5	0	10, 30-70
9	44188518	44256803	Cbl Ccdc153 Pdzd3 Nlrx1	Cbl	5	5	0.0001421	10, 30, 40, 90, 100
9	82937551	82977892	Phip	Phip	5	5	0	20-50
11	98430738	98497056	Erbb2 Mien1 Gm12352 Grb7 Ikzf3 Gm25106	Ikzf3	5	5	0	10-100
14	74857270	74997034	Lrch1	Lrch1	5	5	2.461E-08	20, 40-100
15	85680301	85826601	Mirlet7c-2 Mirlet7b Ppara Cdpf1 Pkdrej	Ppara	5	5	0	10-100
16	6610665	6643072	intergenic	Rbfox1	5	5	0.0001292	30, 40, 60
17	87981043	88042082	Msh6 Fbxo11	Fbxo11	5	5	3.515E-05	30-100
18	2965325	3061432	Vmn1r-ps151	Vmn1r-ps151	5	5	0	10-100
18	4258355	4351005	Map3k8	Map3k8	5	5	2.712E-05	10, 30-100
X	18117256	18272468	Dusp21 Kdm6a	Kdm6a	5	5	0	10, 30-100
8	84877039	84987478	Syce2 Gcdh Klf1 Dnase2a Mast1 Gm24197 Rtbdn Rnaseh2a Prdx2 Junb	Rtbdn	5	4	1.11E-16	10, 20, 40-100
14	121889438	121941085	Ubac2 Gpr18	Ubac2	5	4	0.0001779	40-70
15	10041942	10076929	Gm26350	Gm26350	5	4	0.0002286	10, 40
17	24269479	24395191	Abca17 Gm24427 Abca3 Gm25618	Abca3	5	4	0	10-100
6	145237661	145253240	Kras	Kras	5	3	0	10-30
9	89807593	89823131	intergenic	Mir184	5	3	3.527E-06	10
11	86586697	86592542	Vmp1	Vmp1	5	3	0.0001278	20
12	58996936	59020305	Sec23a Gemin2	Sec23a	5	3	0	10, 20, 40, 50
12	111195708	111213225	Traf3	Traf3	5	3	0.0002902	30
17	35189651	35232539	Ltb Tnf Lta Nfkbil1 Gm16181	Nfkbil1	5	3	6.264E-05	30, 40
X	75068539	75111147	Gab3 Dkc1 Gm25520	Dkc1	5	3	0	10-80
1	152856551	152895091	Smg7	Smg7	4	4	8.055E-05	30-50
1	156614107	156637231	Abl2	Abl2	4	4	0.0001158	40
1	180339684	180407148	Itpkb	Itpkb	4	4	0	10-90
2	75638412	75667388	Gm24574 Hnrnpa3	Gm24574	4	4	4.82E-06	10-40
3	51391351	51437818	Mgarp Ndufc1 Naa15	Naa15	4	4	9.383E-05	10, 40-60
3	95469842	95534896	Arnt Ctsk Ctss Hmgb1-ps5	Ctsk	4	4	9.465E-09	10-90
5	115191991	115253307	Cabp1 Gm13828 Pop5 Rnf10	Pop5	4	4	0	10-70
5	123379752	123397309	Gm15747 Mlxip	Gm15747	4	4	0.000215	60
6	83898801	83932150	Zfml	Zfml	4	4	1.226E-07	10, 30-50
6	115630869	115704194	Raf1 Gm14335 D830050J10Rik	Raf1	4	4	0	10-90
7	45029512	45119443	Prr12 Prrg2 Nosip Rcn3 Fcgrt	Nosip	4	4	1.417E-06	10, 40-100
7	80002302	80048301	Zfp710	Zfp710	4	4	0	10-60
7	90110130	90168806	AC130210.1 Picalm	Picalm	4	4	1.432E-14	10-60
7	125566031	125656085	Il4ra Il21r Gtf3c1	Il21r	4	4	0	10-100

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
8	81828232	81857377	Inpp4b Gm17072	Inpp4b	4	4	0.0002971	60
8	105632825	105734280	RP24-242N1.1 Ctcf Gm5915 Gm24324 Rltpr Acd Pard6a Enkd1 4933405L10Rik Gfod2	Rltpr	4	4	0	10-100
8	111787035	111811148	Cfdp1	Cfdp1	4	4	0.000321	30, 50, 70
9	36750164	36777397	Stt3a AC155921.1	Stt3a	4	4	0.0001403	40
9	51179551	51241797	Pou2af1	Pou2af1	4	4	0	10-70
10	59551903	59590940	Mcu	Mcu	4	4	0.0001301	30-50
11	97124004	97156168	Tbkbp1	Tbkbp1	4	4	3.906E-05	10-40
12	54949179	55022245	Baz1a RP23-454K24.2 Gm20403 Gm24296	Baz1a	4	4	0	20, 40-80
12	98631007	98677794	Spata7 Ptpn21	Spata7	4	4	0.0001923	50, 60
13	30704965	30741259	Dusp22 Gm11370	Gm11370	4	4	2.315E-05	20-50
14	54611840	54703924	Psmb5 Mir686 Psmb11 Cdh24 Gm20726 Gm17606 Acin1 4930579G18Rik 1700123O20Rik	Acin1	4	4	4.694E-05	40
14	101675585	101686647	Uchl3	Uchl3	4	4	0.0003306	60
16	3331981	3361784	Gm22862	Gm22862	4	4	3.136E-12	10-30
16	75890059	75913298	Samsn1	Samsn1	4	4	3.744E-05	10-40
17	23554762	23592911	AC154766.1 Zfp213 Zfp13	Zfp13	4	4	0.0001529	30-50
17	24125845	24221601	Pdpk1 Amdhd2 Atp6v0c Tbc1d24 Ntn3 BC028777 1602H07Rik	Pdpk1	4	4	0.0002129	40, 50, 100
17	80422683	80465661	Sos1	Sos1	4	4	8.291E-10	10-50
19	44349092	44416617	Scd1	Scd1	4	4	1.11E-16	10-40
X	38446906	38556678	Lamp2 Gm7598 Cul4b	Cul4b	4	4	2.914E-13	10, 40, 60-100
18	80623833	80691754	Nfatc1	Nfatc1	4	3	3.973E-09	40
19	6389763	6411265	Pygm Rasgrp2 Gm14965	Rasgrp2	4	2	0	10-30
1	46850357	46859983	Slc39a10	Slc39a10	3	3	1.11E-16	10
1	85929740	85967303	4933407L21Rik Gpr55	Gpr55	3	3	1.891E-06	30, 40
1	170859704	170874151	Atf6 Gm9929 Dusp12	Gm9929	3	3	3.815E-05	30, 40
1	172142625	172159962	Gm10171 Dcaf8	Dcaf8	3	3	4.282E-05	30, 40
2	168570953	168609146	Nfatc2	Nfatc2	3	3	2.335E-09	20, 30
2	180692870	180716374	Dido1 Gm22502 Gid8	Dido1	3	3	0	10-40
4	40839296	40852965	B4galt1 Mir5123 Gm24112 Gm25931	B4galt1	3	3	7.815E-05	20, 30
4	89270241	89288424	Gm12606 Cdkn2a	Cdkn2a	3	3	2.998E-15	10, 30, 40
4	131869305	131889818	Srsf4	Srsf4	3	3	5.38E-05	30, 40
4	133152527	133176240	Wasf2 Gm24636	Wasf2	3	3	6.894E-07	10-40
5	128987750	129030633	Stx2 Ran	Ran	3	3	0	10, 20, 40-60
6	37726166	37740819	intergenic	Gm15487	3	3	3.042E-06	10-40
6	70715778	70735561	Igkj1 Igkj2 Igkj3 Igkj4 Igkj5 Igkc	Igkc	3	3	4.441E-16	10-40
6	128981486	129002969	Clec2g BC064078	BC064078	3	3	9.626E-14	10-40
7	19569930	19618847	Gemin7 Zfp296 Clasrp AC149052.1 Relb	Clasrp	3	3	0.0001263	10, 30, 50
7	27558164	27607168	2310022A10Rik Akt2	Akt2	3	3	0	10-70
7	45540241	45591093	Plekha4 Gm16022 Hsd17b14 0610005C13Rik Bcat2	)610005C13Ri	3	3	2.22E-16	10, 30-60
7	84672736	84687489	Zfand6	Zfand6	3	3	4.203E-09	10, 20, 40
8	106936921	106961196	Sntb2	Sntb2	3	3	0.0002554	40, 50
9	2983343	3024328	AC131780.1 Gm10722 Gm11168 Gm10721 Gm10720 Gm10719 Gm10718	AC131780.1	3	3	0	10-60, 80, 100
10	81374496	81414705	Fzr1 Dohh 2210404O07Rik Nfic Gm16104	Dohh	3	3	0	20, 30, 50
11	34031196	34044834	4930469K13Rik	:930469K13Ri	3	3	1.016E-09	10, 20
12	92873788	92890209	intergenic	Gm23249	3	3	4.856E-13	10, 20, 40
13	20131395	20147778	Elmo1	Elmo1	3	3	1.388E-08	10, 30
13	43781816	43793456	Cd83	Cd83	3	3	5.217E-05	20
13	52624650	52657631	Syk	Syk	3	3	7.327E-15	10-30
14	7888640	7947574	Flnb	Flnb	3	3	5.329E-15	20-50
14	27279241	27310905	Arhgef3	Arhgef3	3	3	0.000158	40, 50
14	72652782	72658298	Fndc3a	Fndc3a	3	3	3.491E-06	10
14	75183309	75209092	Lcp1 Gm15629	Lcp1	3	3	0.000152	40
14	79397052	79404419	Mtrf1	Mtrf1	3	3	0.0002793	40
15	80724280	80743766	Tnrc6b	Tnrc6b	3	3	0.000233	30, 40
16	8563262	8569071	Abat	Abat	3	3	0.0001961	20
17	3069051	3086207	Pisd-ps2	Pisd-ps2	3	3	8.654E-11	10, 30
17	46755239	46758099	CT030702.1 Ptcra	CT030702.1	3	3	0.0003076	30
17	49994748	50017622	Rftn1	Rftn1	3	3	0	10-40
19	60131251	60153735	E330013P04Rik	:330013P04Ri	3	3	0	10-30
X	7817728	7934754	Gripap1 Kcnd1 Otud5 Pim2 Slc35a2 Pqbp1 Timm17b Gm10491 Gm10490 Pcsk1n	Otud5	3	3	0	10-100
Y	90690526	90840981	Gm21860 Gm21857 Erdr1 Gm21748	Erdr1	3	3	0	20-100
1	37079612	37080574	intergenic	Vwa3b	2	2	8.187E-07	10
1	85598375	85601263	Sp110 Gm16094 Sp140	Sp140	2	2	1.078E-06	10
1	86501238	86504125	intergenic	Ptma	2	2	0	10
2	6209267	6212203	Echdc3 A230108P19Rik	Echdc3	2	2	4.935E-12	10
2	49516360	49519296	Epc2	Epc2	2	2	3.433E-05	10
2	152828225	152831161	Bcl2l1	Bcl2l1	2	2	4.418E-08	10
2	173270207	173271185	Pmepa1	Pmepa1	2	2	9.517E-05	10
3	27454968	27457752	Fndc3b	Fndc3b	2	2	3.086E-05	10
3	90110557	90121695	Nup210l	Nup210l	2	2	8.203E-05	20
4	130984054	130988934	Gm12973	Gm12973	2	2	2.065E-11	10
5	27340	27341947	Dpp6	Dpp6	2	2	3.618E-05	10
5	29368625	29371545	Lmbr1	Lmbr1	2	2	1.529E-05	10
5	116954811	116957731	intergenic	Suds3	2	2	2.126E-10	10

Chr	Start	End	Genes in CIS	Gene nearest to peak	Number insertions	Number insertions (no hop)	p value	Scale
5	124010717	124016560	Vps37b	Vps37b	2	2	2.029E-11	10, 20
6	85433259	85435212	Smyd5	Smyd5	2	2	8.111E-05	20
7	43351349	43354279	Siglec5	Siglec5	2	2	3.969E-07	10
7	142550850	142583108	Nctc1 H19 Mir675	Nctc1	2	2	0	10-40
9	35303218	35305161	intergenic	Gm5614	2	2	4.79E-05	10
9	75434947	75436889	intergenic	4130057D12Ri	2	2	5.613E-05	10
10	12919478	12921427	AL691505.1	AL691505.1	2	2	1.138E-05	10
10	21080184	21082133	Ahi1	Ahi1	2	2	1.126E-05	10
10	60160497	60164397	intergenic	Chst3	2	2	2.419E-05	20
10	117904515	117906464	intergenic	Rap1b	2	2	4.946E-05	10
12	79464884	79466829	Rad51b	Rad51b	2	2	7.865E-06	10
12	107864502	107865474	intergenic	Bcl11b	2	2	5.481E-08	10
13	51730278	51736098	Sema4d	Sema4d	2	2	9.598E-07	10-20
13	55197182	55199122	intergenic	Nsd1	2	2	1.356E-05	20
13	59655460	59657400	Golm1	Golm1	2	2	0.0001651	20
14	63501275	63503114	Tdh	Tdh	2	2	0.0001257	10
15	27930407	27932348	Trio	Trio	2	2	3.537E-05	10
15	59292264	59295175	intergenic	Sqle	2	2	9.555E-06	10
15	83189091	83191031	intergenic	Cyb5r3	2	2	0.0001275	10
18	36148244	36154037	Nrg2	Nrg2	2	2	3.242E-14	10, 20
18	60802736	60820157	Cd74 Mir5107	Cd74	2	2	0	10-30
18	65581279	65587075	Zfp532	Zfp532	2	2	5.142E-08	20
19	5841548	5849035	Neat1 Gm9783	Neat1	2	2	5.675E-05	20
19	40990396	40991330	Blnk	Blnk	2	2	7.674E-05	10
19	57332933	57335738	AC131756.1	AC131756.1	2	2	1.923E-07	10
2	27748312	27749291	Rxra	Rxra	1	1	3.517E-13	10