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.