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.

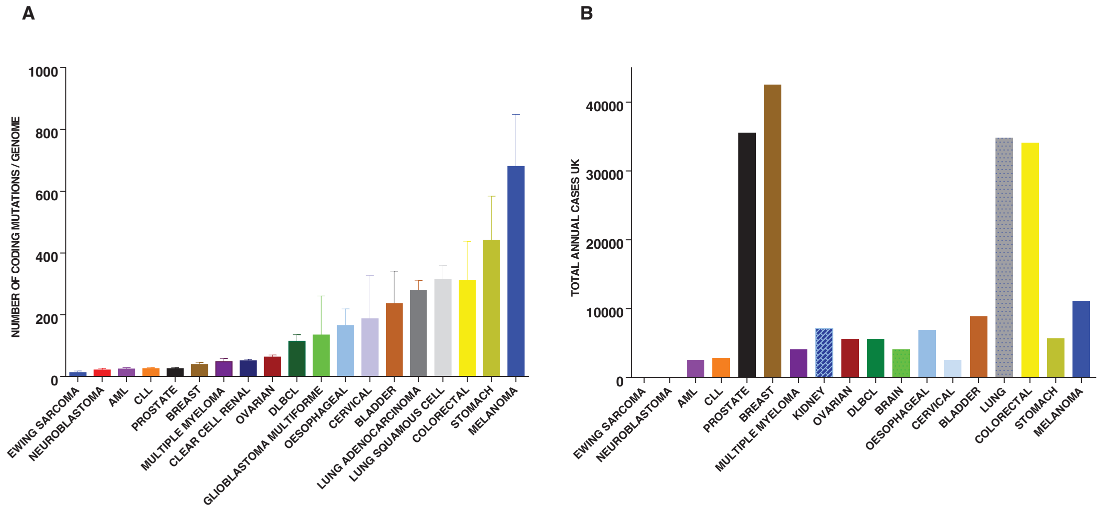

FIGURE 1.1: MUTATION BURDEN AND CANCER INCIDENCE (A) A comparison of the mean number of non-coding mutations per genome across various tumour types. The raw data is taken from Lawrence et al, 2013. Error bars show the standard error of the mean. **(B)** UK annual incident cases of various malignancies 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.

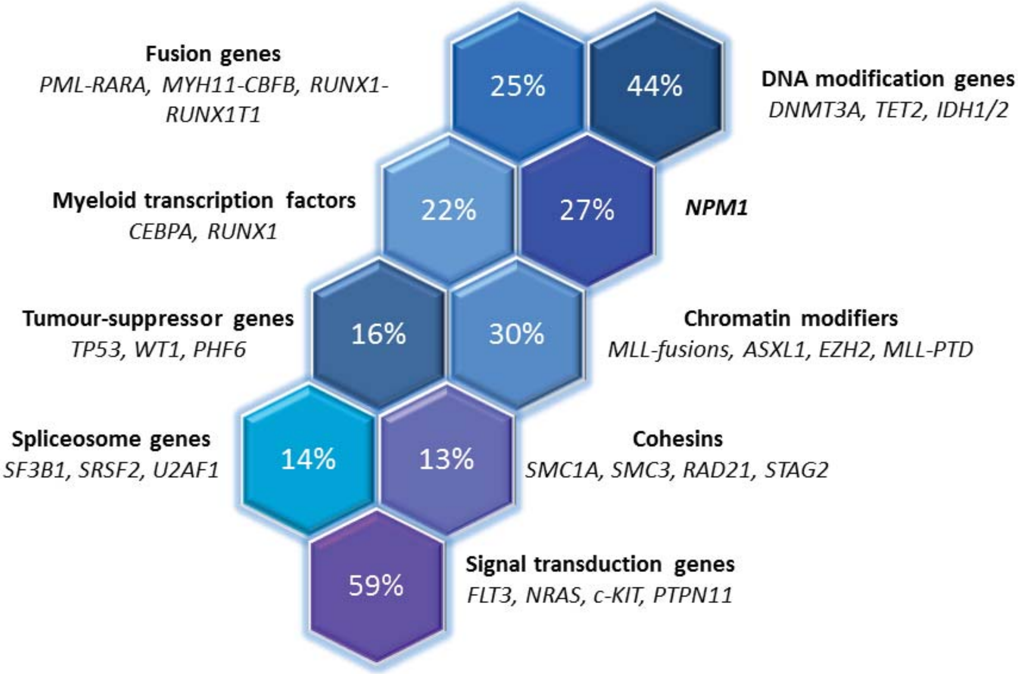

Figure 1.2: RECURRENT MUTATION GROUPS IN AML. Data on the prevalence of different mutation groups in AML (based on data from TCGA Research Network, 2013) (Grove and Vassiliou, 2014).

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.

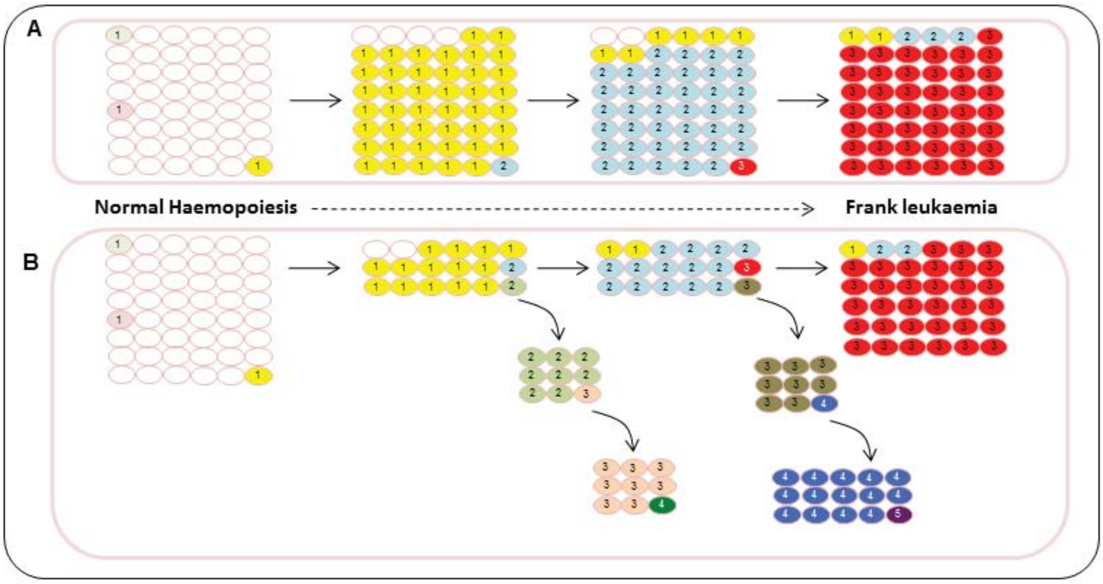

FIGURE 1.3: LINEAR AND BRANCHING CLONAL EVOLUTION
(A) Linear evolution: Sequential dominant clones (clonal sweep) result in a linear architecture with stepwise accumulation of driver mutations. The final tumour carries all mutations arising during evolutionary history and overwhelms earlier clones carrying 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 heterozygous 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).

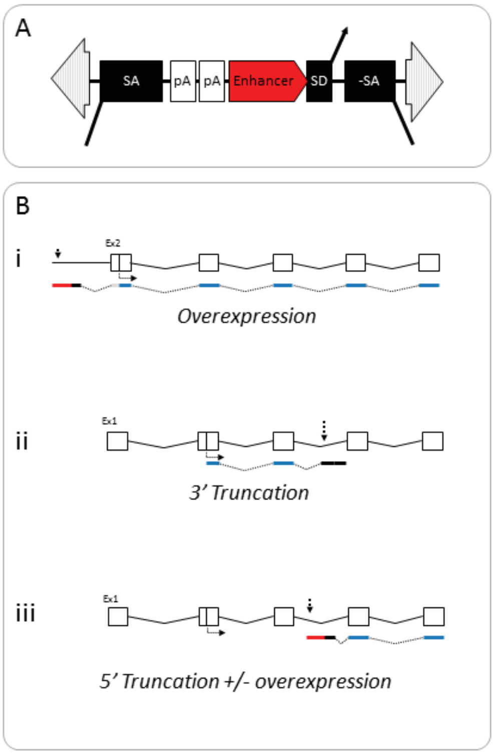

Figure 1.4 Examples of the mutagenic effects of transposons. A: A transposon with a unidirectional activating element followed by a splice donor, but dual splice acceptors and polyadenylation signals such that gene expression may be blocked regardless of the direction of transposon insertion. **B:** 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.