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

## 4.1 Introduction

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

Transposon insertional mutagenesis is a valuable technique with which to study tumorigenic mutations in mouse models. To date the predominant application has been for cancer gene discovery, analogous to retroviral mutagenesis. Putative tumour drivers are identified by ascertaining genes and regions in which the transposon or retrovirus integrates more frequently than is expected by chance alone; the common integration sites (CIS). This approach has proven effective for both solid and haematopoietic malignancies, but the *in vivo* kinetics of transposon integration is poorly understood. Unlike retroviruses, transposon mobilization continues throughout the life of a host cell and critical oncogenic events may potentially occur after a significant latency. Typically multiple copies of the transposon cassette are supplied in a concatamer and the rate of new integrations (i.e. transposition) relative to the rate of cell division is largely unknown. The timing and order of acquisition of oncogenic integrations in mouse insertional mutagenesis models has not been studied to date.

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

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

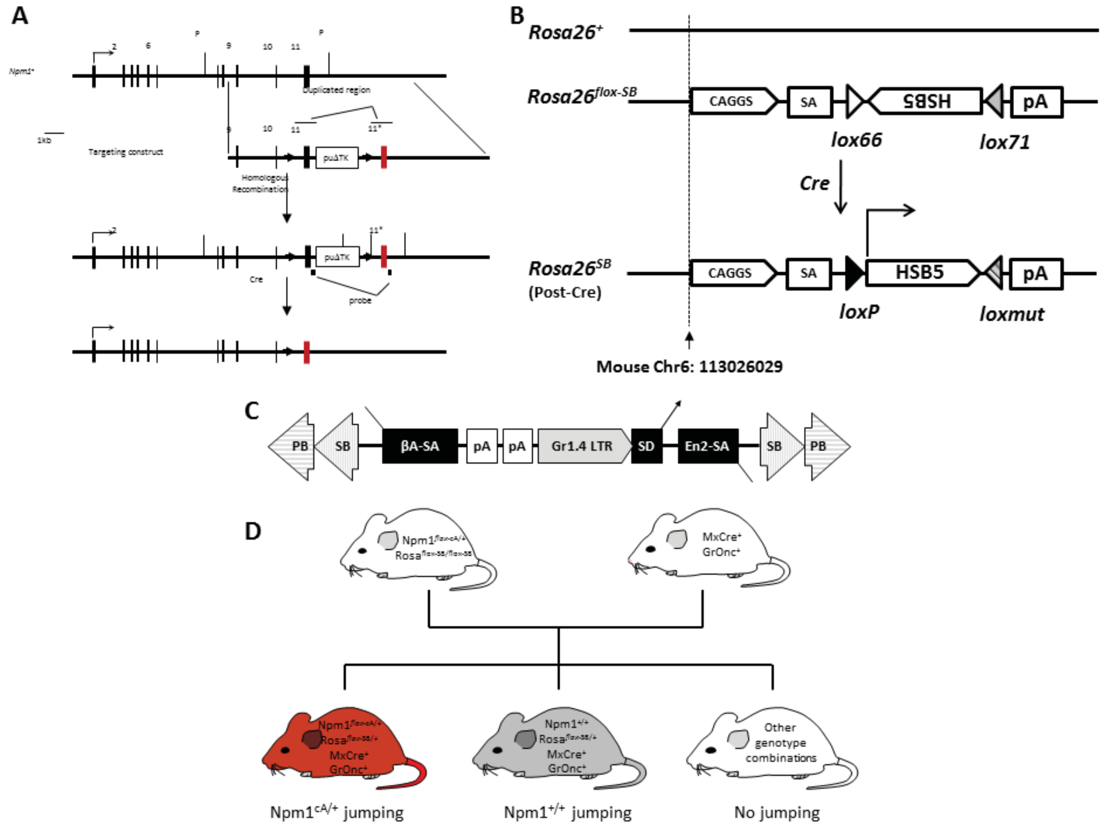


**Figure 4.1: Generation of *Npm1$^{cA}$* insertional mutagenesis mice.** A: Targeting construct for *Npm1$^{cA}$*. B: Conditional Rosa26 *SB* transposase. Upon Cre activation the SB cDNA flips to the sense orientation and cannot take part in any further Cre mediated recombination. C: The GrOnc transposon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

## 4.2 Results

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

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

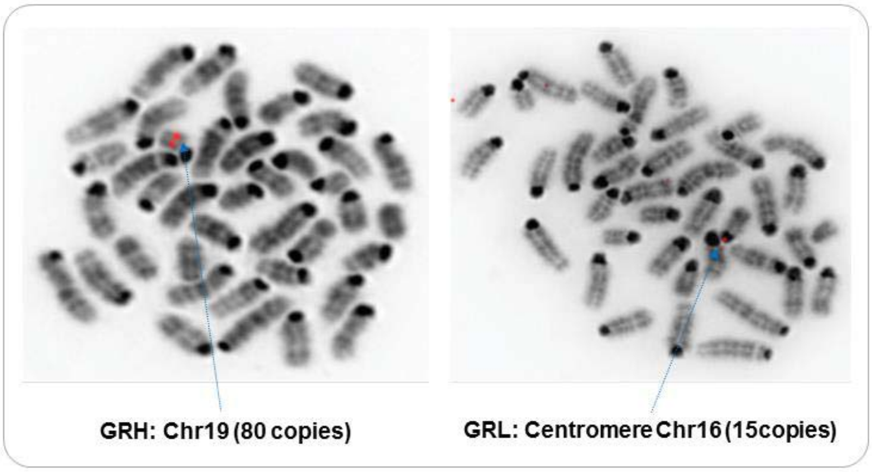


**Figure 4.2: FISH analysis of the GrOnc constructs.** On the left the published high copy number construct (GRH) and on the right the low copy number construct used for this work (GRL). FISH results were supplied by Ruby Banerjee and GV.

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

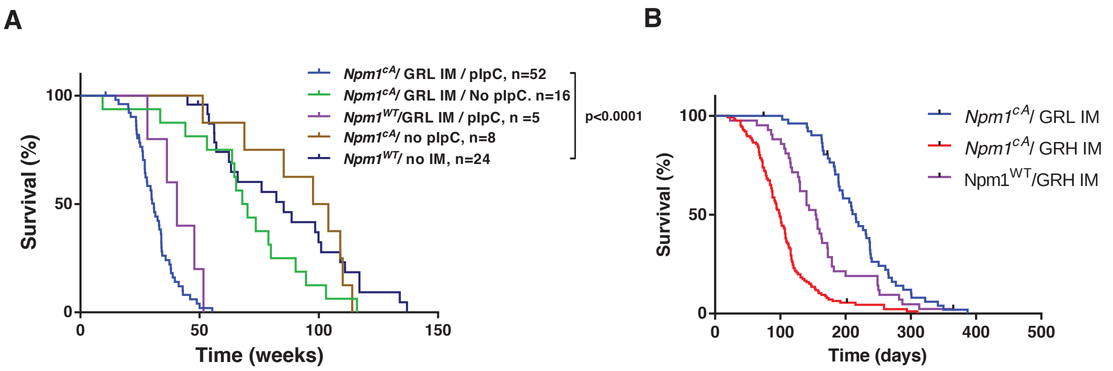


**Figure 4.3: Survival Curves. A)** Survival in the GRL cohorts. Mice that received an incomplete course of pIpC injections are not included in the analysis. **B)** Survival compared to the GRH IM cohorts.

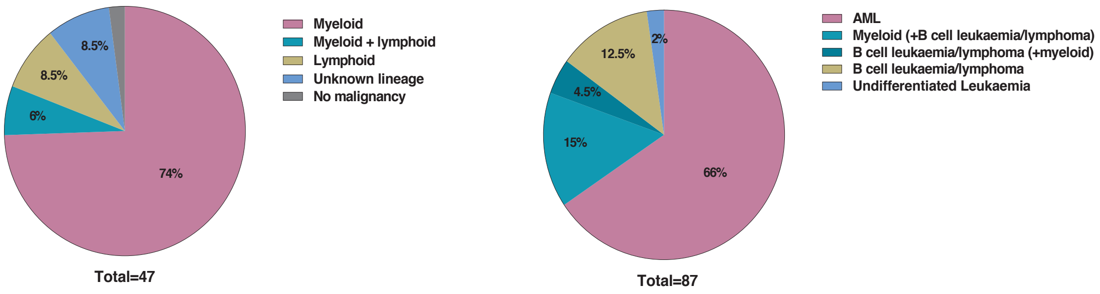


**Figure 4.4: Disease phenotype in the *Npm1$^{cA}$* IM cohorts.** On the left the GRL cohort and on the right the published GRH colony.

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

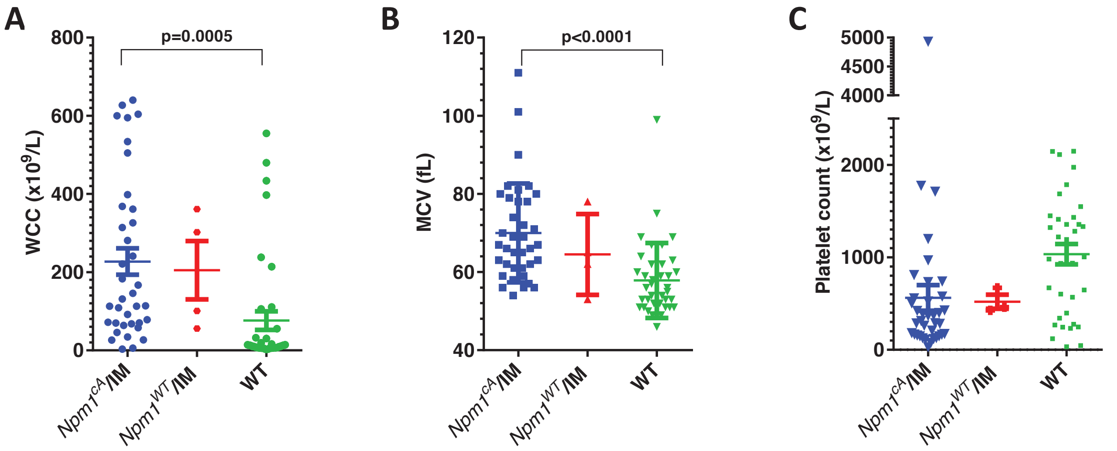


**Figure 4.5: FBC parameters in the GRL IM cohort at death**: **A)** White cell count **B)** Mean cell volume and **C)** Platelet count. The WT cohort includes mice that did not receive any pIpC injections, in addition to those with a WT genotype.

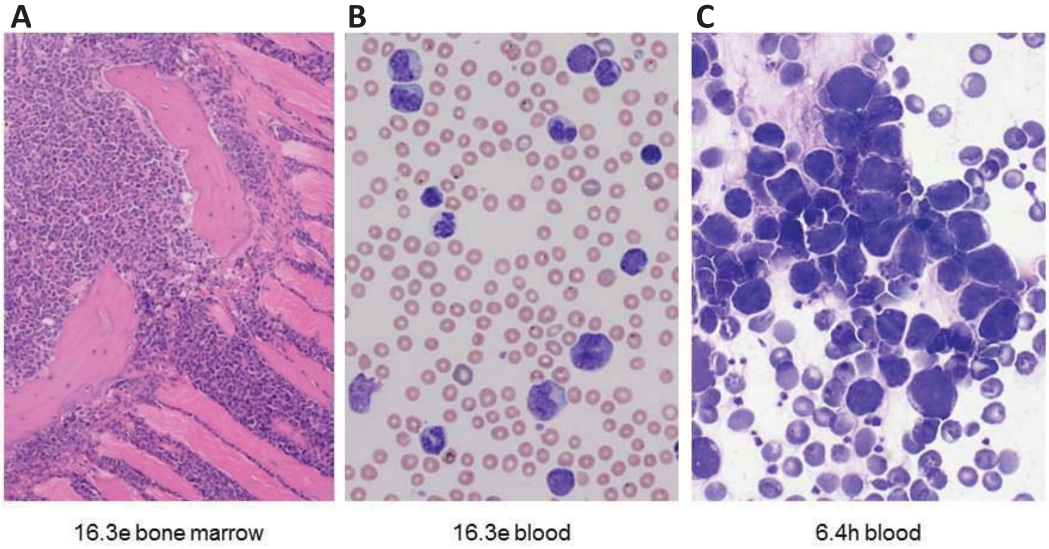


**Figure 4.6: Spectrum of morphology in the *Npm1$^{cA}$* GRL IM cohort.** Although the bone marrow is packed with myeloid cells, many of the mice have blasts and maturing cells on the peripheral blood smear as in 16.3e (**A** and **B**). Some have a high percentage of frank blasts, as shown here in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

### 4.2.2 GRL verifies CISs identified by GRH and identifies additional ones

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

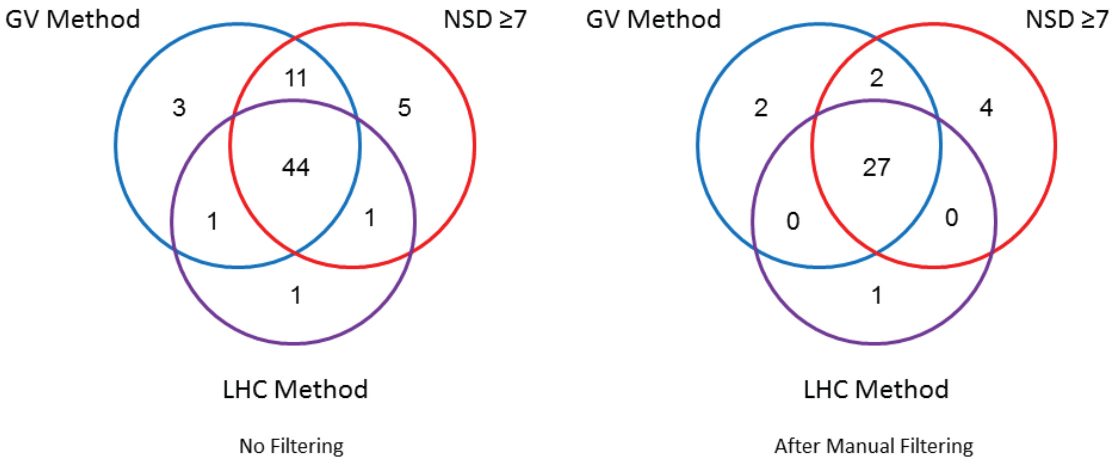
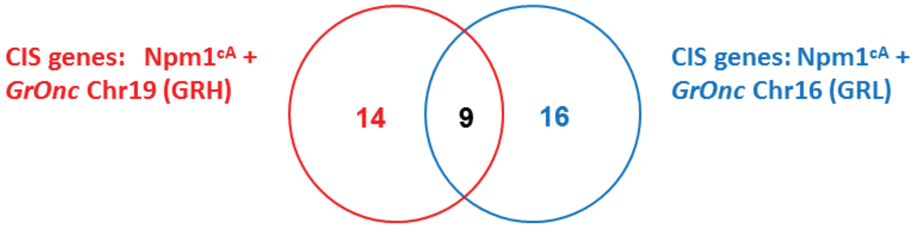


**Figure 4.7: CIS identified by the three CIMPL methods.** The total number of sites identified by each method before manual filtering are shown on the left, and after manual filtering on the right. GV method = published GRH model, NSD≥7 as per GRH but with higher NSD value, LHC method with built in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

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

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


CIS genes: $Npm1^{cA}$ + *GrOnc* Chr19 (GRH)
14
9
16
CIS genes: $Npm1^{cA}$ + *GrOnc* Chr16 (GRL)

| | | |
|---|---|---|
| ***Csf2*** | ***Nf1*** | ***Nup98*** |
| ***Bach2*** | *Ghr* | *Cnot1* |
| ***Zfp423*** | ***Flt3*** | ***Pax5*** |
| *Picalm* | ***Pum1*** | *Mtfr1* |
| *Tnks* | *Zfp521* | *Pik3r5* |
| ***Ptpn1*** | *Ptpn2* | *Dleu2* |
| *Myb* | *Rasgrp1* | *Pigq* |
| *12:intergenic* | *Kras* | |

| | | |
|---|---|---|
| ***Nf1*** | ***Csf2*** | ***Nup98*** |
| *Stat5b* | ***Flt3*** | *A330023F24Rik/ Mir29b2/Cd34* |
| *Tnrc6b* | *Nrf1/Mir183* | *8:Intergenic* |
| *Gnb1* | *Mll1* | *Rps6ka2* |
| ***Zfp423*** | *Phf6* | *4933426M11Rik* |
| *Ube2e2* | *Hs6st1/Uggt1* | ***Bach2*** |
| ***Pax5*** | *Bmi1* | *Sec16a* |
| ***Pum1*** | ***Ptpn1*** | *Mbnl1* |
| *Il2rb* | | |

Vassiliou et al (2011) Nat Genet 43(5): 470-475

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

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

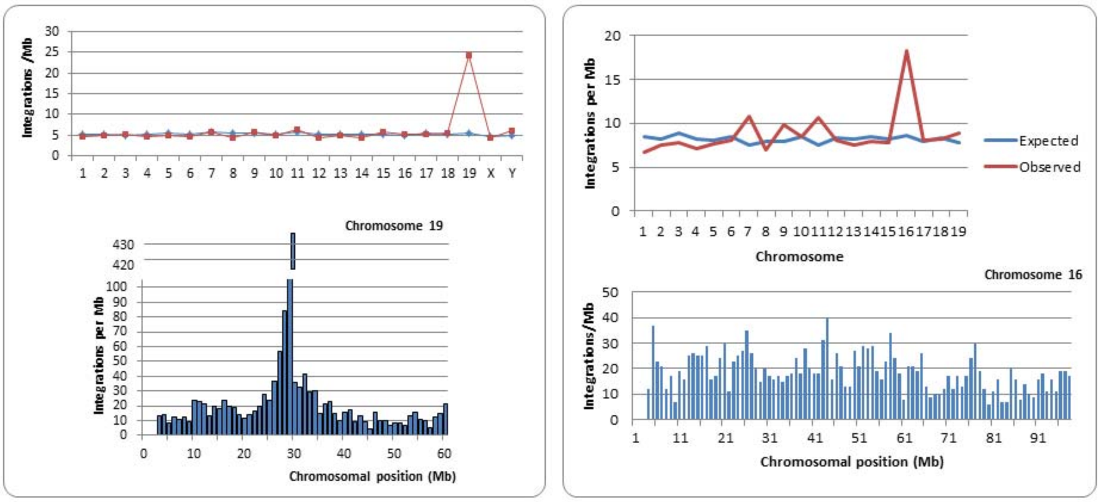


**Figure 4.9:** Distribution of integrations across the genome (top) and within the donor chromosome (bottom) for the published GRH cohort (left) and the GRL cohort (right).

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

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

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

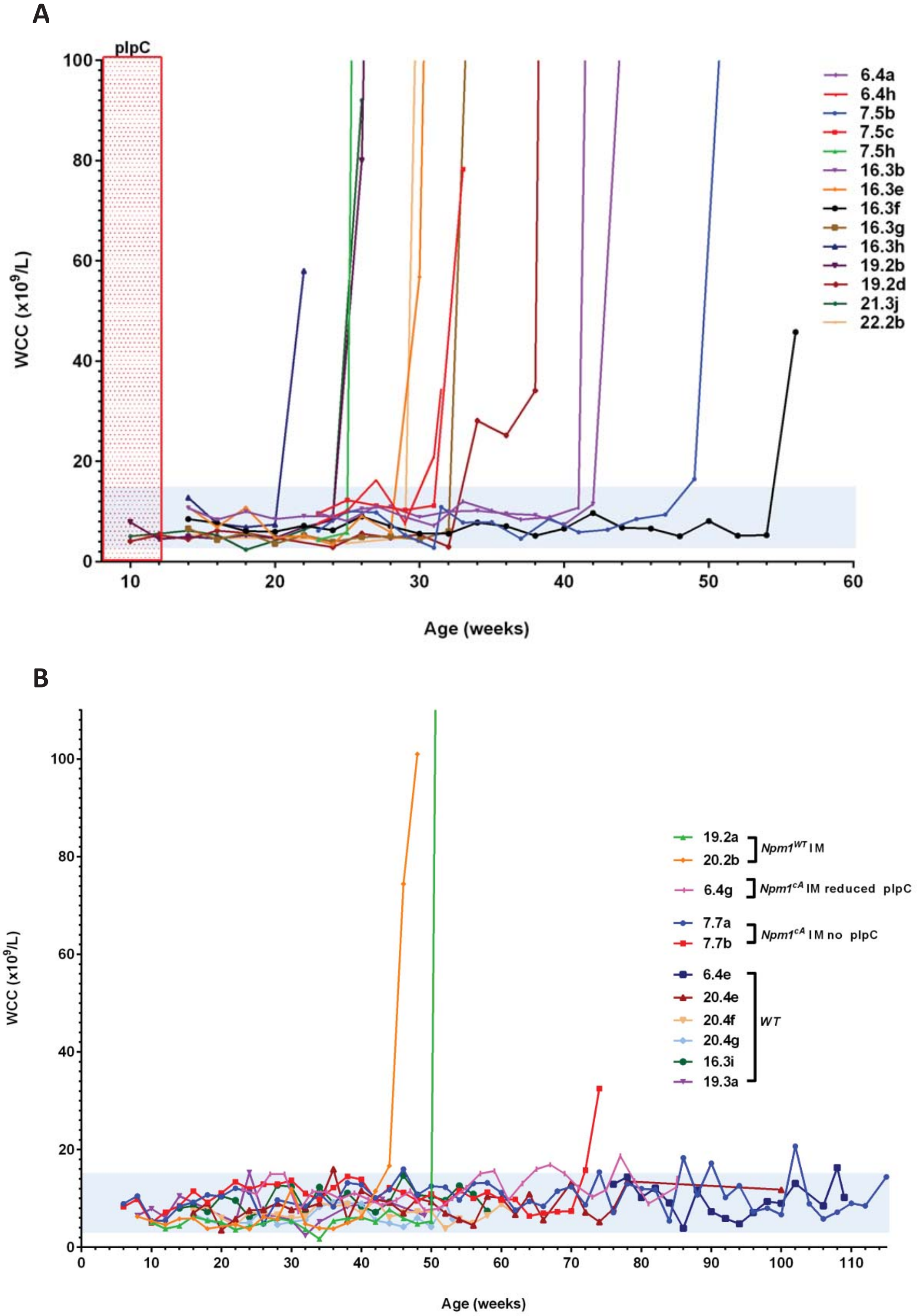


**Figure 4.10: WCC in serial blood tests from the Npm1cA GRL IM cohort (A) and the mice wildtype for at least one allele or with incomplete course of pIpC (B).** The normal range is indicated in blue. Timing of the pIpC injections is indicated in red. Injections in an individual mouse were given over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

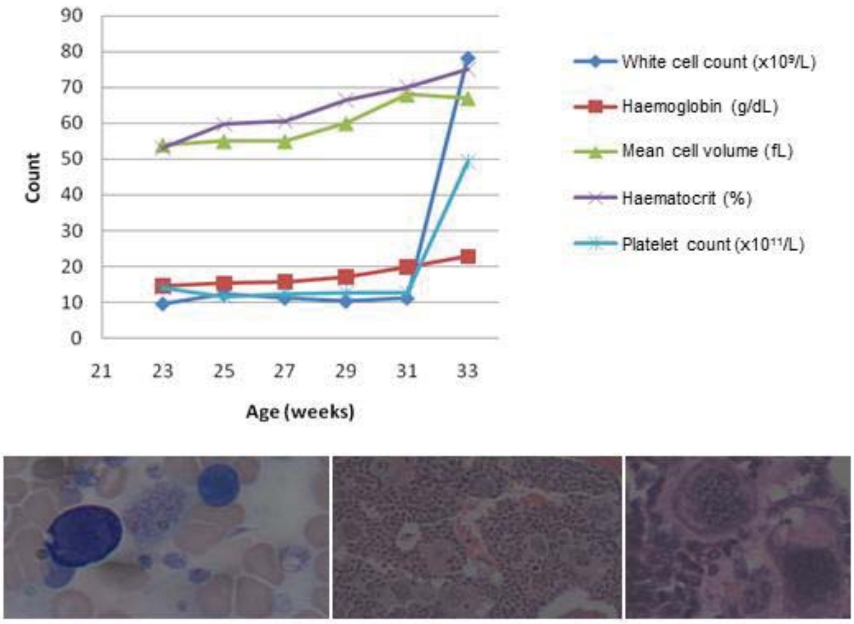


**Figure 4.11: Unusual characteristics of mouse 7.5c. Top:** Blood parameters. **Bottom:** Blood film and bone marrow pathology showing giant platelets and increased megakaryocyte number, some with atypical morphology. Photographs provided by G Hoffman.

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

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

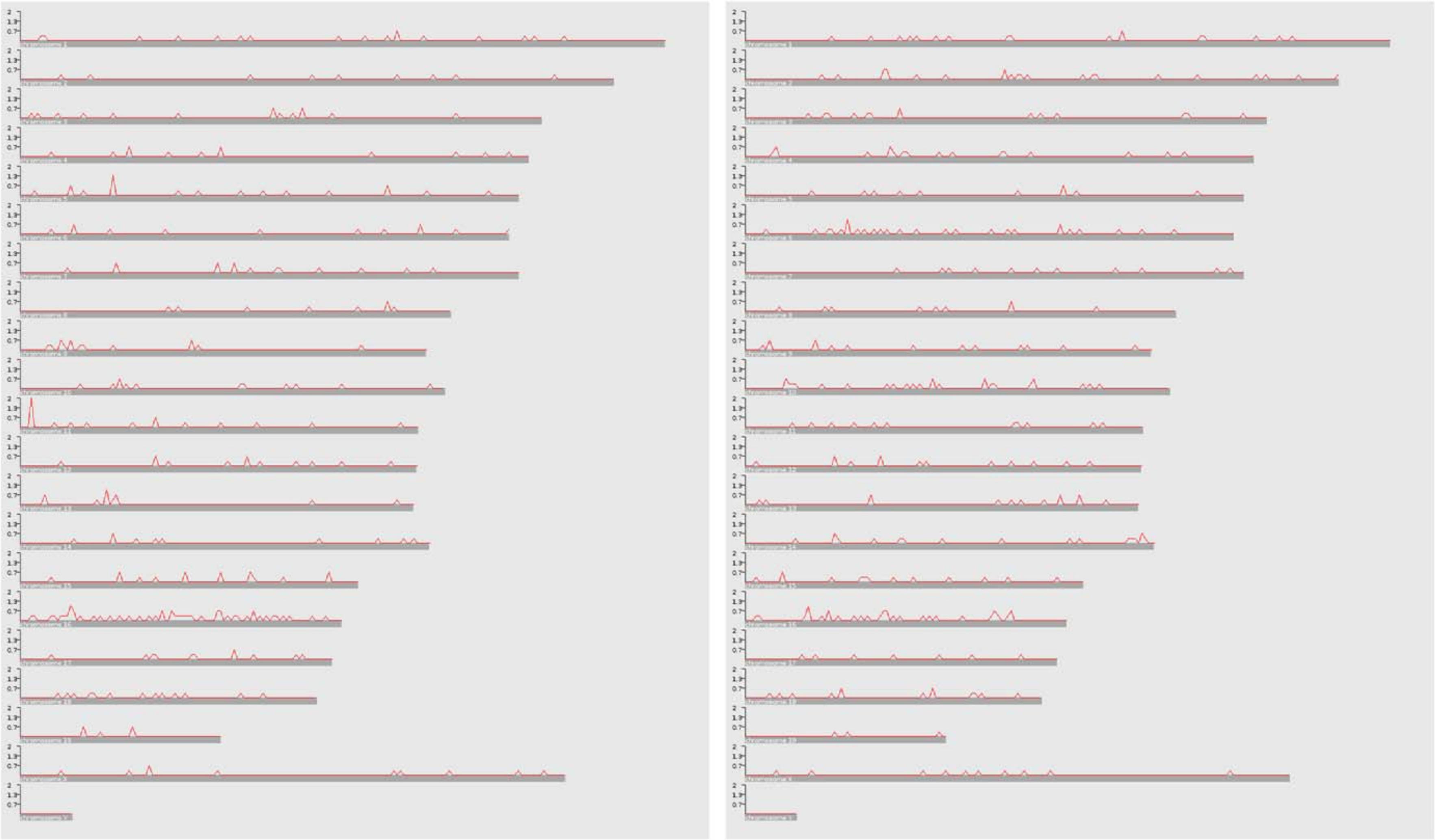


**Figure 4.12: Transposon distribution over time**. The number of transposon integrations in 1Mb bins are shown across the genome in samples taken  2 (left) and 12 (right) weeks after completion of pIpC injections.  The data are pooled from sibling mice 16.3e, 16.3g and 16.3h.

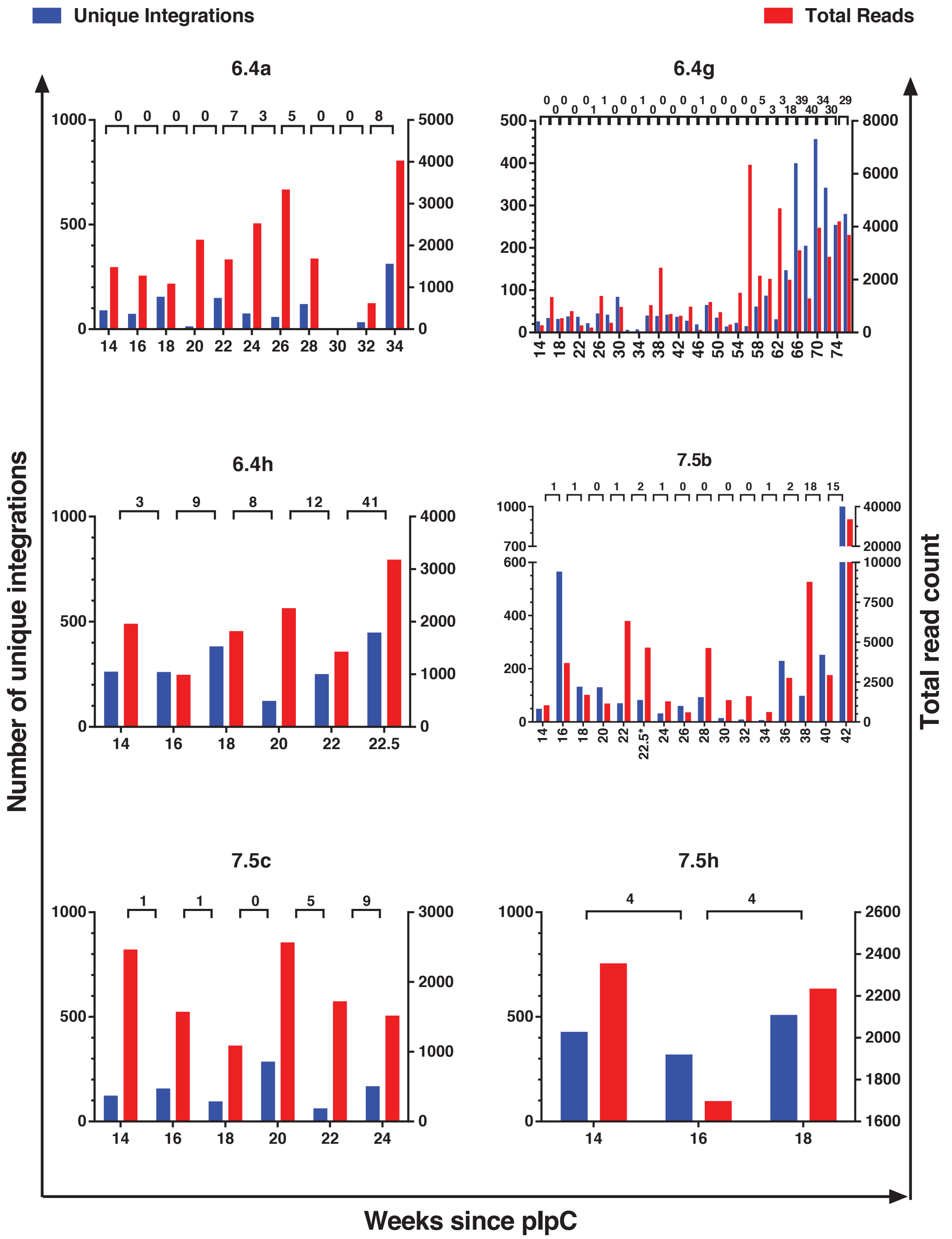
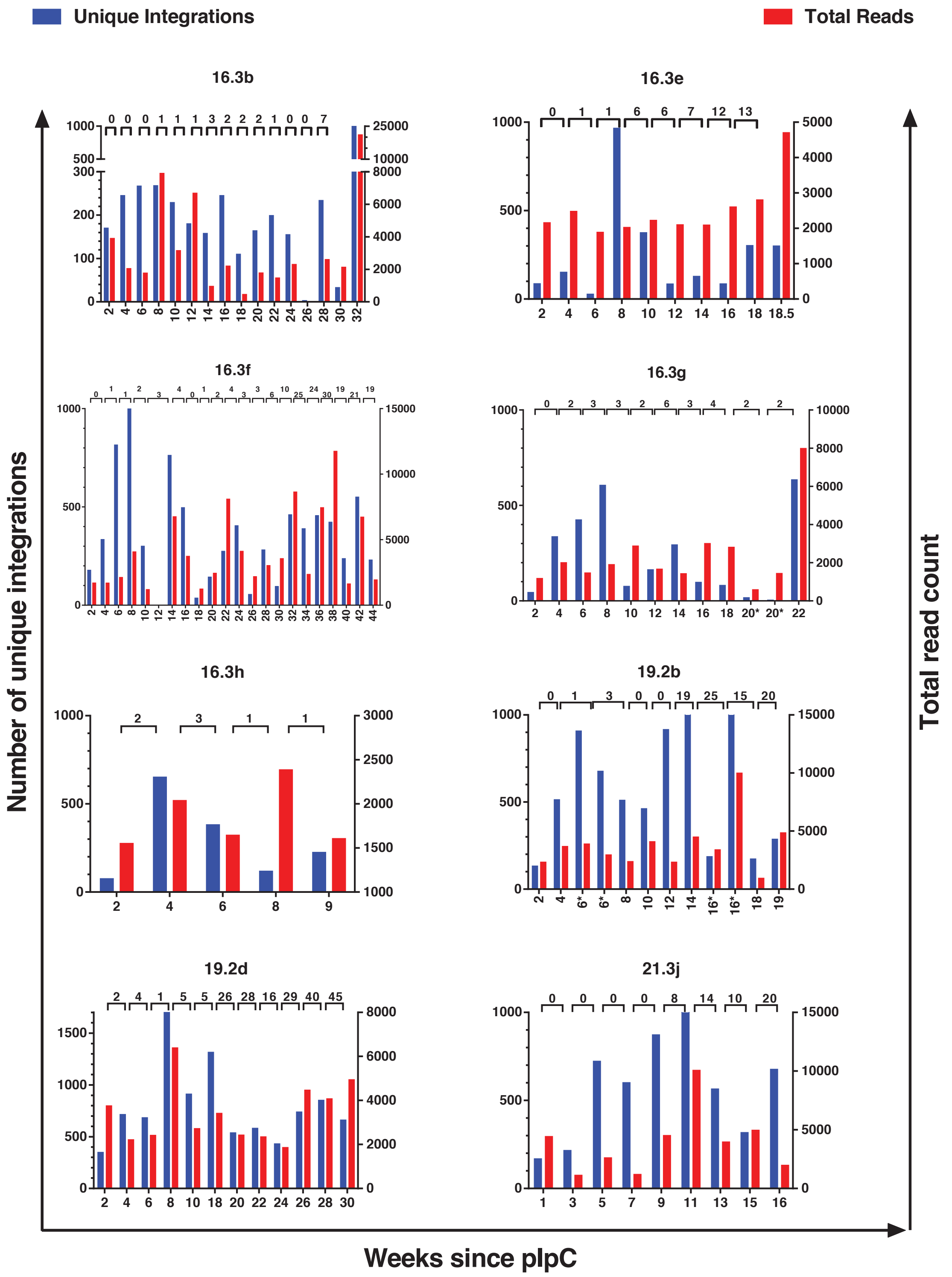


**Figure 4.13: Number of reads mapped and overlapping integrations between blood samples in the serially bled mice.** Overlapping integrations between consecutive samples are shown at the top. For some specimens the sample was run in duplicate or samples were repeated less than two weeks apart prior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



🟦 Unique Integrations
🟥 Total Reads
16.3b
0 0 0 1 1 1 3 2 2 2 1 0 0 7
1000 500 300 200 100 0
25000 10000 8000 6000 4000 2000 0
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
16.3e
0 1 1 6 6 7 12 13
1000 500 0
5000 4000 3000 2000 1000 0
2 4 6 8 10 12 14 16 18 18.5
16.3f
0 1 1 2 3 4 0 1 2 4 3 3 6 10 25 24 30 19 21 19
1000 500 0
15000 10000 5000 0
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
16.3g
0 2 3 3 2 6 3 4 2 2
1000 500 0
10000 8000 6000 4000 2000 0
2 4 6 8 10 12 14 16 18 20* 20* 22
16.3h
2 3 1 1
1000 500 0
3000 2500 2000 1500 1000
2 4 6 8 9
19.2b
0 1 3 0 0 19 25 15 20
1000 500 0
15000 10000 5000 0
2 4 6* 6* 8 10 12 14 16* 16* 18 19
19.2d
2 4 1 5 5 26 28 16 29 40 45
1500 1000 500 0
8000 6000 4000 2000 0
2 4 6 8 10 18 20 22 24 26 28 30
21.3j
0 0 0 0 8 14 10 20
1000 500 0
15000 10000 5000 0
1 3 5 7 9 11 13 15 16
Number of unique integrations
Total read count
Weeks since pIpC

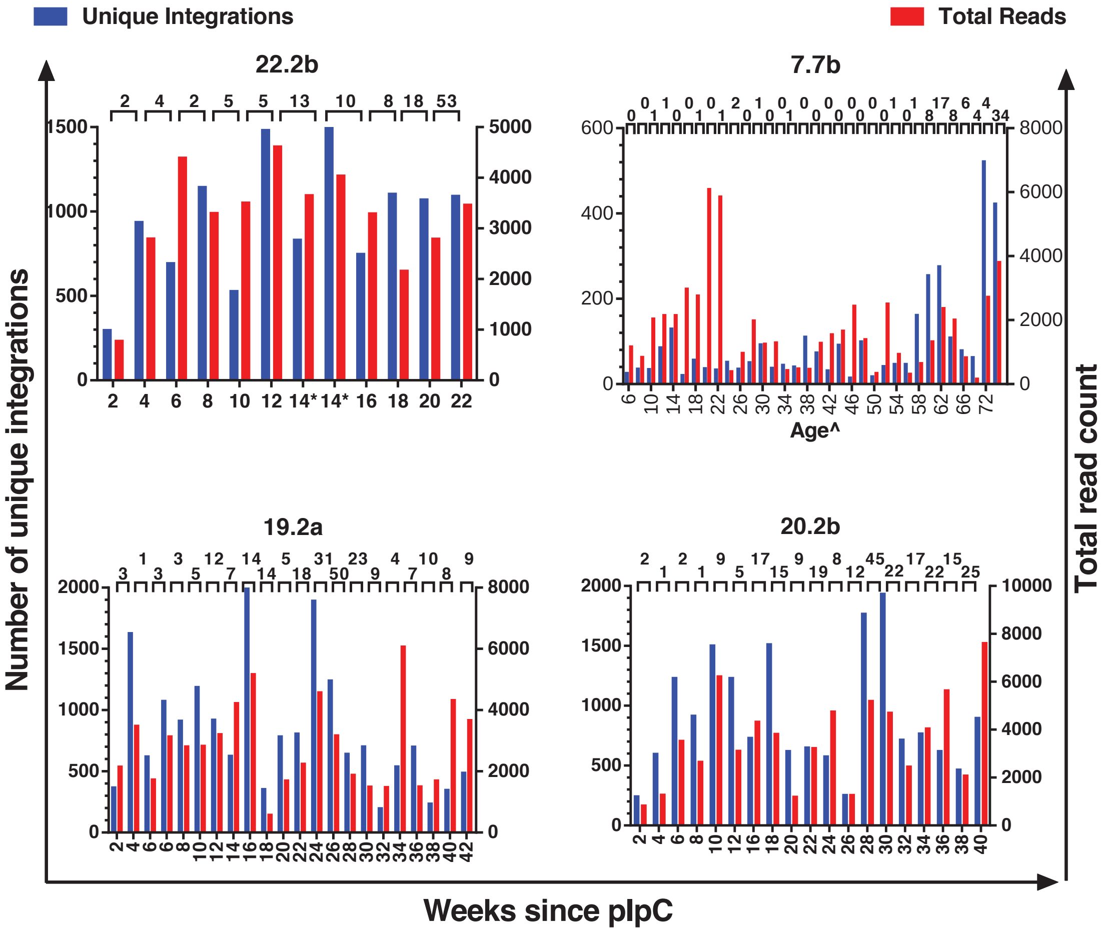
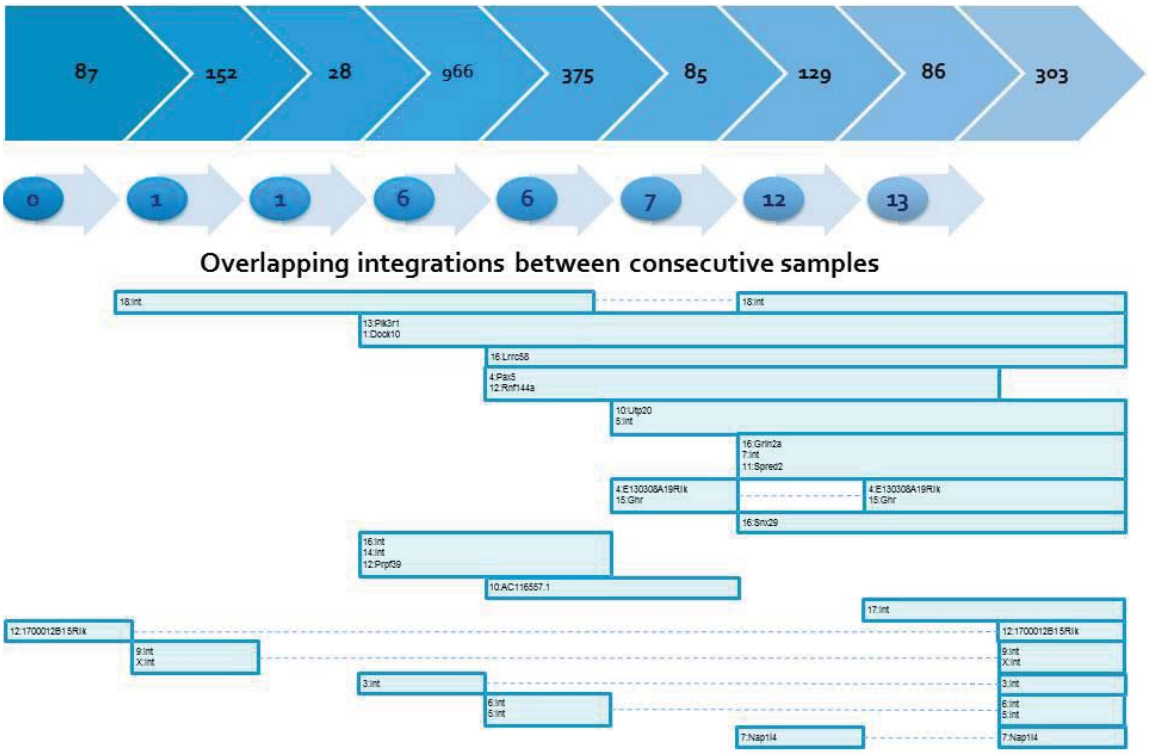
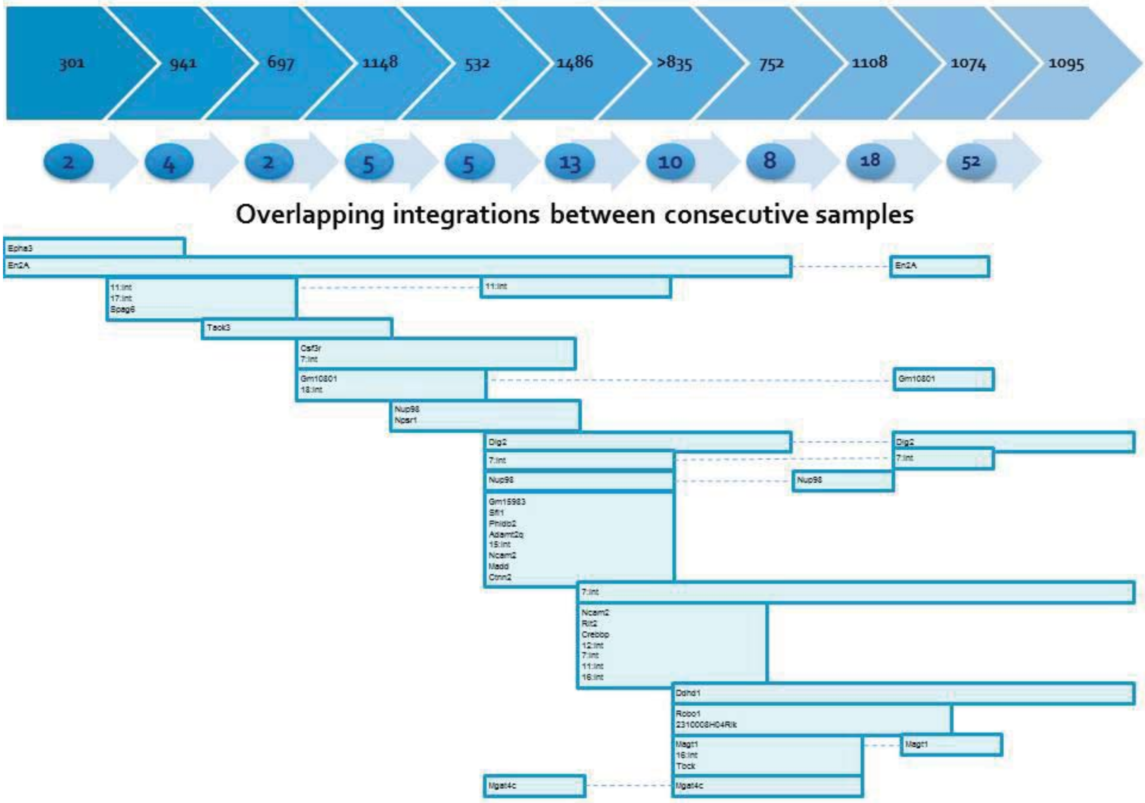


**Figure 4.14 (next page): Overlapping integrations in 16.3e (top) and 22.2b (bottom).** The arrows show the number of reads per blood sample and the number of overlapping integrations between consecutive samples. The chromosome and gene names are given. Int = intergenic. Each intergenic site listed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

Number of unique transposon integrations per blood sample:16.3e



Number of unique transposon integrations per blood sample: 22.2b



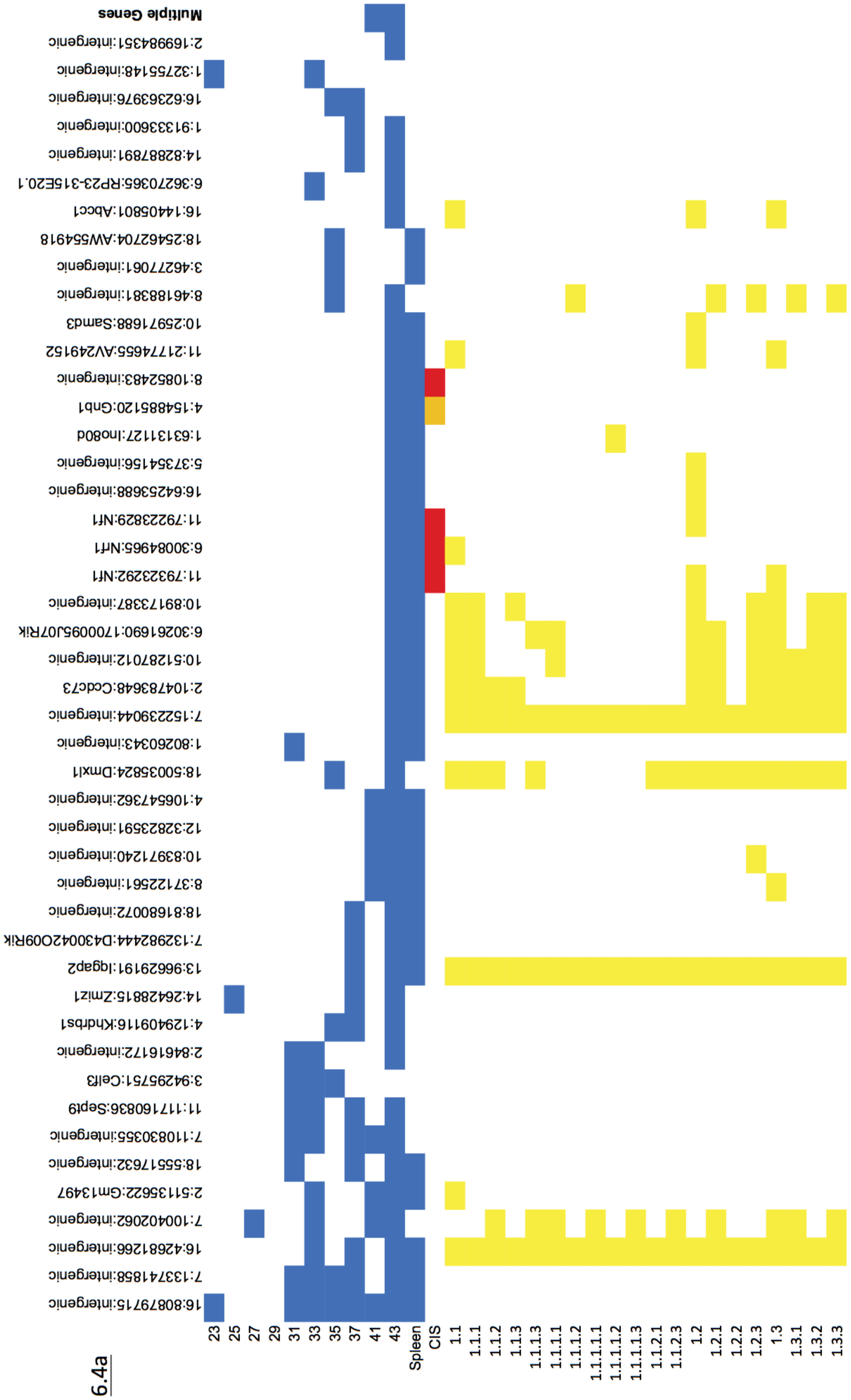
### 4.2.5 A small number of transposon integrations occur early and persist in the pre-leukaemic samples and on serial transplantation of leukaemia cells

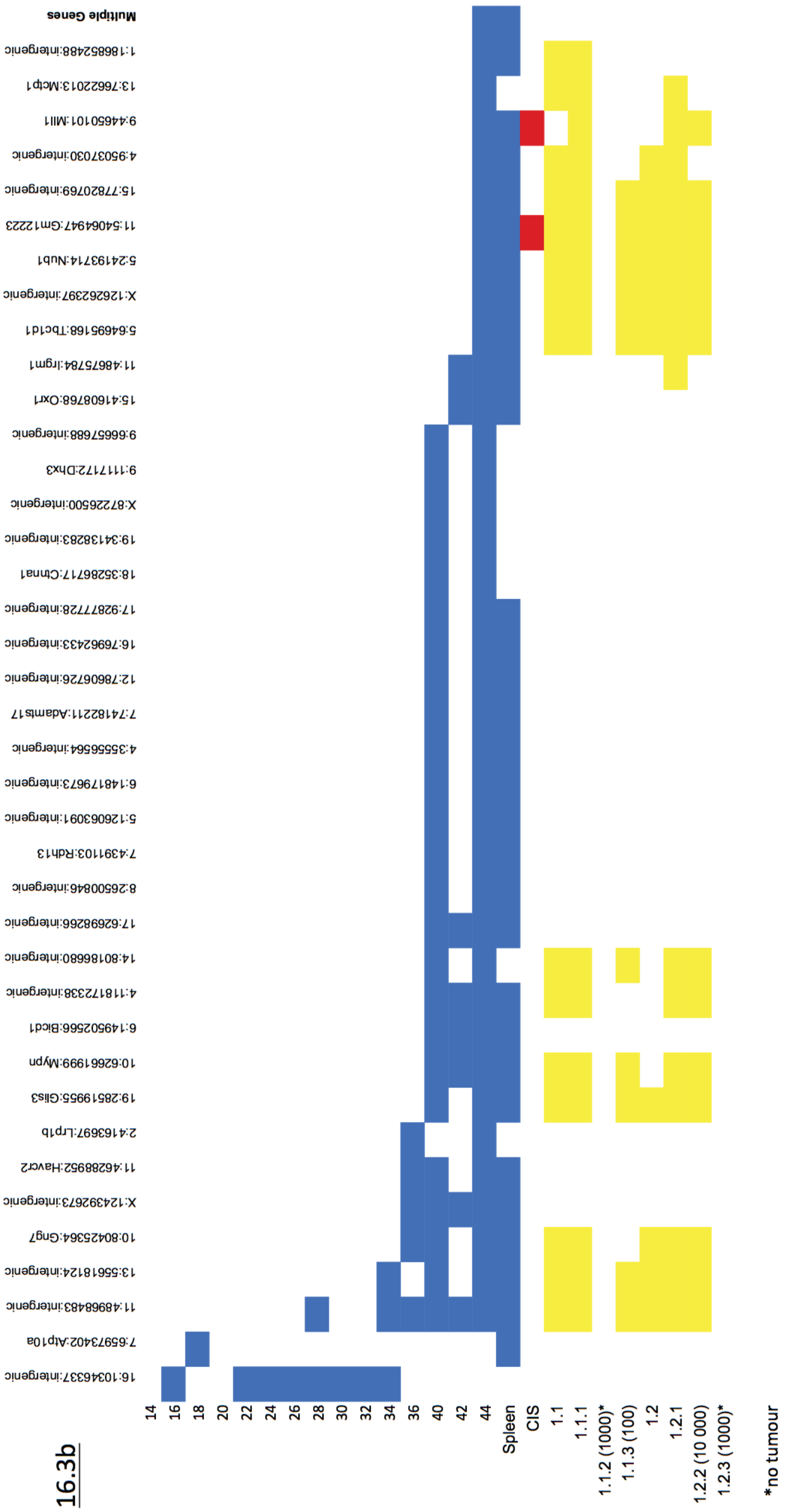
The vast majority of transposon integrations appeared only transiently. A minority of transposon integrations persisted on sequential samples from the same mouse, although the number of persisting integrations typically increased over time (figure 4.13). Pre-leukaemic blood samples were analysed for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

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

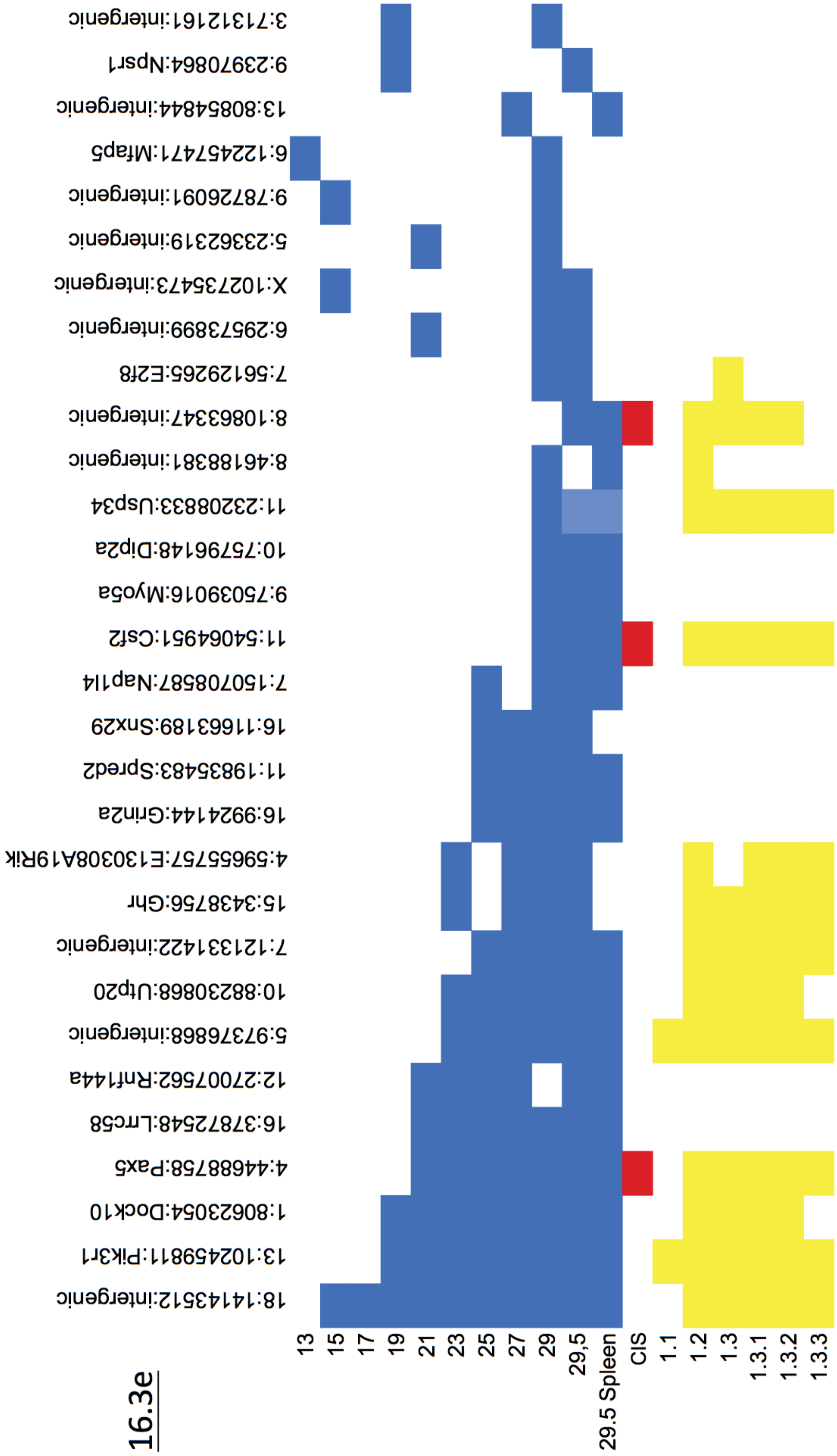
Transplants were performed at varying cell doses from 1 million down to 100 cells. The transplanted cells generally engrafted and generated leukaemia when a dose of at least $10^4$ unsorted spleen cells were used. Upon transplantation of 1000 cells, only 12 of 21 transplants generated leukaemia and with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

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


6.4a
Multiple Genes
2:169984351:intergenic
1:32755148:intergenic
16:62363976:intergenic
1:91333600:intergenic
14:82887891:intergenic
6:36270365:RP23-315E20.1
16:14405801:Abcc1
18:25462704:AW554918
3:46277061:intergenic
8:46188381:intergenic
10:25971688:Samd3
11:21774655:AV249152
8:10852483:intergenic
4:154885120:Gnb1
1:63131127:Ino80d
5:37354156:intergenic
16:64253688:intergenic
11:79223829:Nf1
6:30084965:Nf1
11:79323292:Nf1
10:89173387:intergenic
6:30261690:1700095J07Rik
10:51287012:intergenic
2:104783648:Ccdc73
7:15223904:intergenic
1:80260343:intergenic
18:50035824:Dmxl1
4:106547362:intergenic
12:32823591:intergenic
10:83971240:intergenic
8:37122561:intergenic
18:81680072:intergenic
7:132982444:D430042O09Rik
13:96629191:Iqgap2
14:26428815:Zmiz1
4:129409116:Khdrbs1
2:84616172:intergenic
3:94295751:Celf3
11:117160836:Sept9
7:110830355:intergenic
18:55517632:intergenic
2:51135622:Gm13497
7:100402062:intergenic
16:42681266:intergenic
7:133741858:intergenic
16:80879715:intergenic
23
25
27
29
31
33
35
37
41
43
Spleen
CIS
1.1
1.1.1
1.1.2
1.1.3
1.1.1.3
1.1.1.1
1.1.1.2
1.1.1.1.1
1.1.1.1.2
1.1.1.1.3
1.1.2.1
1.1.2.3
1.2
1.2.1
1.2.2
1.2.3
1.3
1.3.1
1.3.2
1.3.3


Multiple Genes
1:186852488:intergenic
13:76622013:Mctp1
9:44650101:Mll1
4:95037030:intergenic
15:77820769:intergenic
11:54064947:Gm12223
5:24193714:Nub1
X:126262397:intergenic
5:64695168:Tbcfd1
11:48675784:Irgm1
15:41608768:Oxr1
9:66657688:intergenic
9:111717172:Dhx3
X:87226500:intergenic
19:34138283:intergenic
18:35286717:Ctnna1
17:92877728:intergenic
16:76962433:intergenic
12:78606726:intergenic
7:74182211:Adamts17
4:35556564:intergenic
6:148179673:intergenic
5:126063091:intergenic
7:4391103:Rdh13
8:26500846:intergenic
17:62698266:intergenic
14:80186680:intergenic
4:118172338:intergenic
6:149502566:Bicd1
10:62661999:Mypn
19:28519955:Glis3
2:4163697:Lrp1b
11:46288952:Havcr2
X:124392673:intergenic
10:80425364:Gng7
13:55618124:intergenic
11:48968483:intergenic
7:65973402:Atp10a
16:10346337:intergenic
14
16
18
20
22
24
26
28
30
32
34
36
40
42
44
Spleen
CIS
1.1
1.1.1
1.1.2 (1000)*
1.1.3 (100)
1.2
1.2.1
1.2.2 (10 000)
1.2.3 (1000)*
*no tumour

16.3b


16.3e
3:71312161:intergenic
9:23970864:Npsr1
13:80854844:intergenic
6:122457471:Mfap5
9:78726091:intergenic
5:23362319:intergenic
X:102735473:intergenic
6:29573899:intergenic
7:56129265:E2f8
8:108633347:intergenic
8:46188381:intergenic
11:23208833:Usp34
10:75796148:Dip2a
9:75039016:Myo5a
11:54064951:Csf2
7:150708587:Nap1l4
16:11663189:Snx29
11:19835483:Spred2
16:9924144:Grin2a
4:59655757:E130308A19Rik
15:3438756:Ghr
7:121331422:intergenic
10:88230868:Utp20
5:97376868:intergenic
12:27007562:Rnf144a
16:37872548:Lrrc58
4:44688758:Pax5
1:80623054:Dock10
13:102459811:Pik3r1
18:14143512:intergenic
13
15
17
19
21
23
25
27
29
29,5
29.5 Spleen
CIS
1.1
1.2
1.3
1.3.1
1.3.2
1.3.3

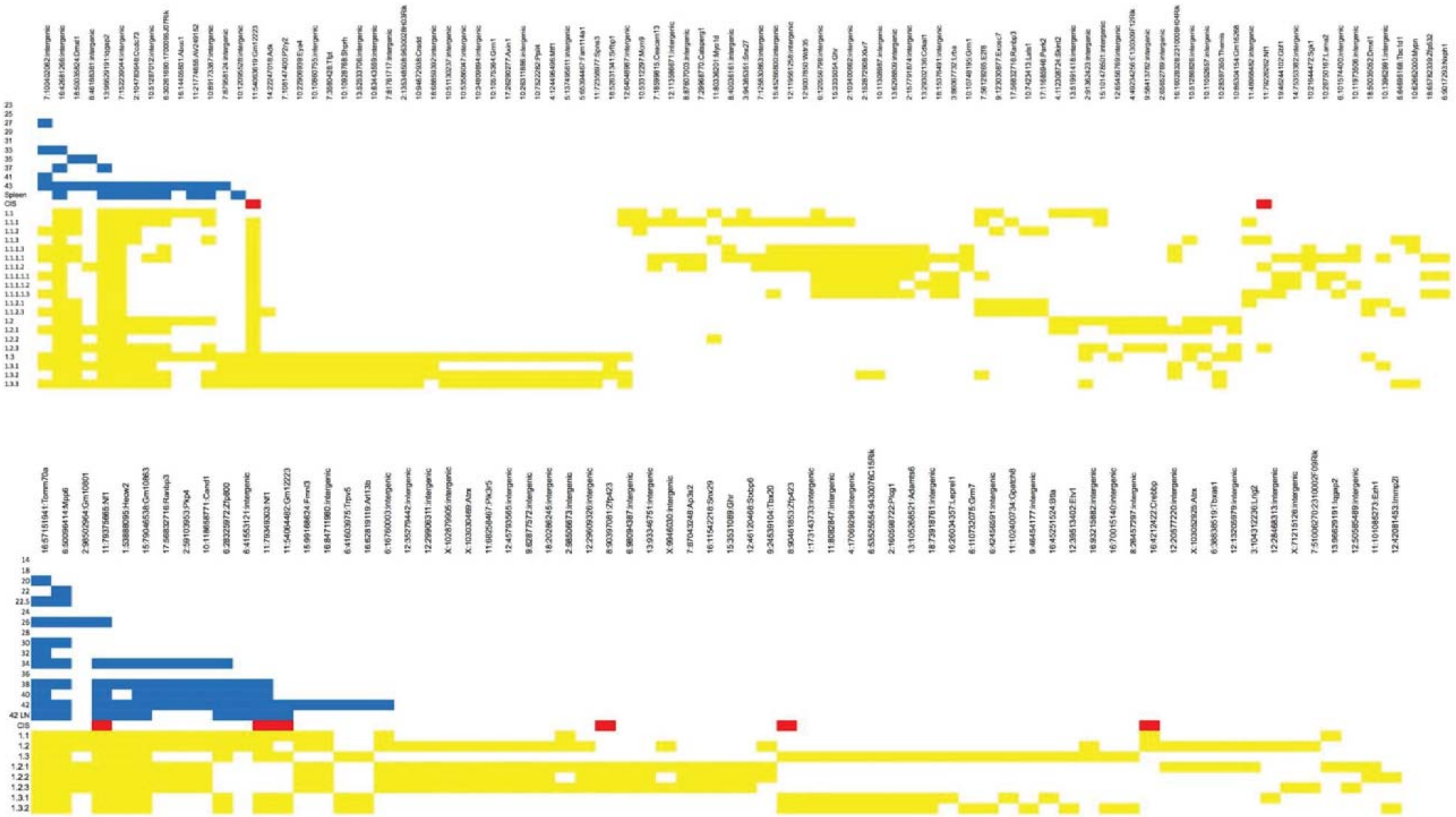
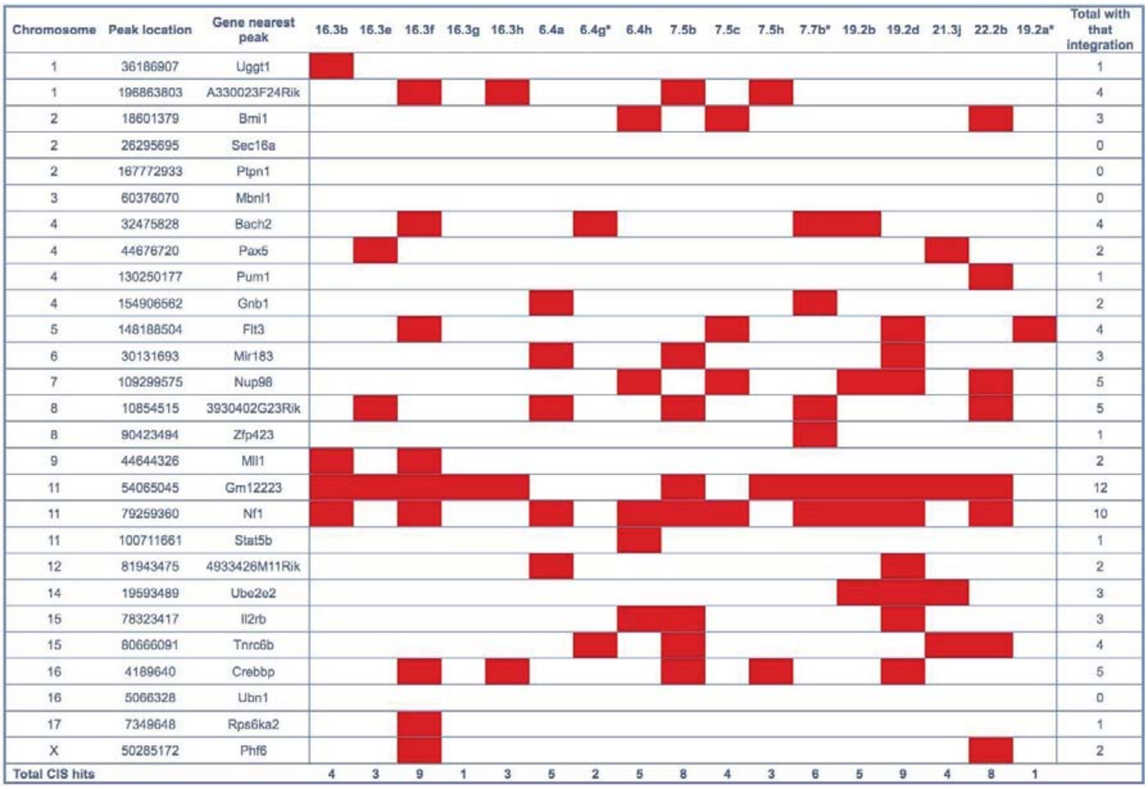


**Figure 4.16: Shared integrations on serial transplants from 6.4a (top) and 7.5b (bottom).** Integrations are represented if they were found in ≥2 transplant samples. The format is otherwise similar to figure 4.15. Several integrations were shared in multiple recipient tumours even though they were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



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

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

### 4.2.6 CIS in the pre-leukaemic blood samples

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

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

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

### 4.2.7 Some transposons loose the capacity to re-mobilise

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

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

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

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

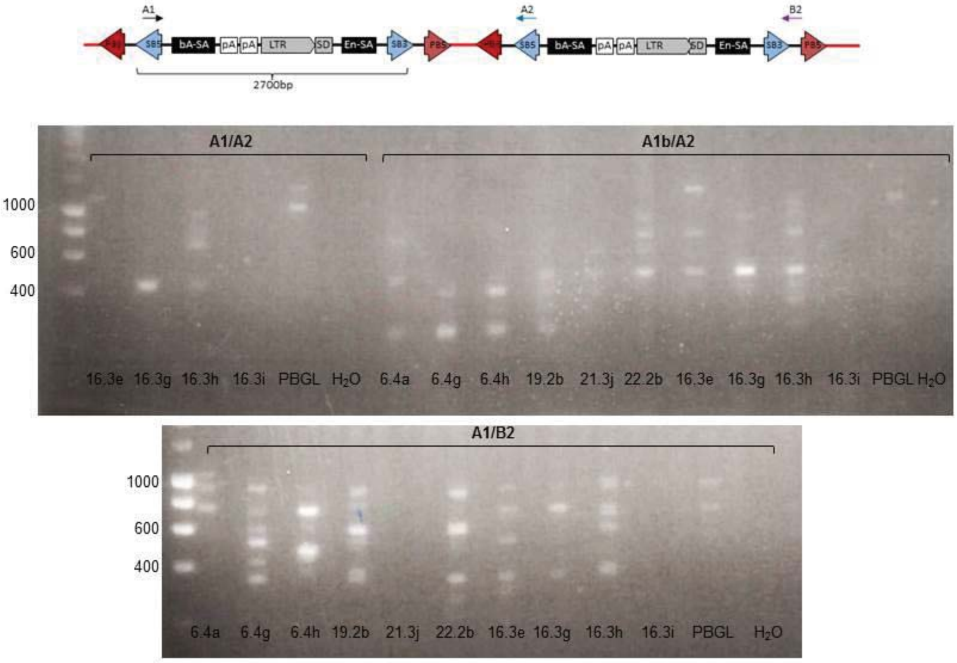

A1
A2
B2
PB3
SB5
bA-SA
pA
pA
LTR
SD
En-SA
SB3
PB5
2700bp
A1/A2
A1b/A2
1000
600
400
16.3e 16.3g 16.3h 16.3i PBGL H₂O 6.4a 6.4g 6.4h 19.2b 21.3j 22.2b 16.3e 16.3g 16.3h 16.3i PBGL H₂O
A1/B2
1000
600
400
6.4a 6.4g 6.4h 19.2b 21.3j 22.2b 16.3e 16.3g 16.3h 16.3i PBGL H₂O

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

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

### 4.2.8 Searching for alternative drivers in transposon IM mice

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

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

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

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

## 4.3 Discussion

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

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

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

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

It is also evident from the serial blood sample data in figure 4.13, that increased read depth does not necessarily correlate with a significantly higher number of unique transposon integrations, although as expected, few unique integrations are mapped when the read depth is poor. Again this variation likely reflects the preferential amplification of particular transposon integrations. When the splinkerette PCR amplifies a small number of integrations very well, a large proportion of the 454 sequencing reads are taken up by these sites. In spite of the good overall read number, coverage of other integration sites is limited. In some samples there was also an artefact due to reads which were amplifications from the transposon primer directly into the Splinkerette linker, without intervening DNA or the transposon end sequence. This is presumed to occur due to non-specific annealing of the primer and in a few samples this accounted for 25-30% of total reads, whereas in most it was <1%. Varying the amount of linker used in the ligation reaction did not have a consistent effect on the proportion of reads with this artefact. Although these reads were filtered out in the analysis because they did not start with ‘TG’ corresponding with the end of the transposon sequence, they did reduce the read coverage for true transposon integrations in some samples.

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

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

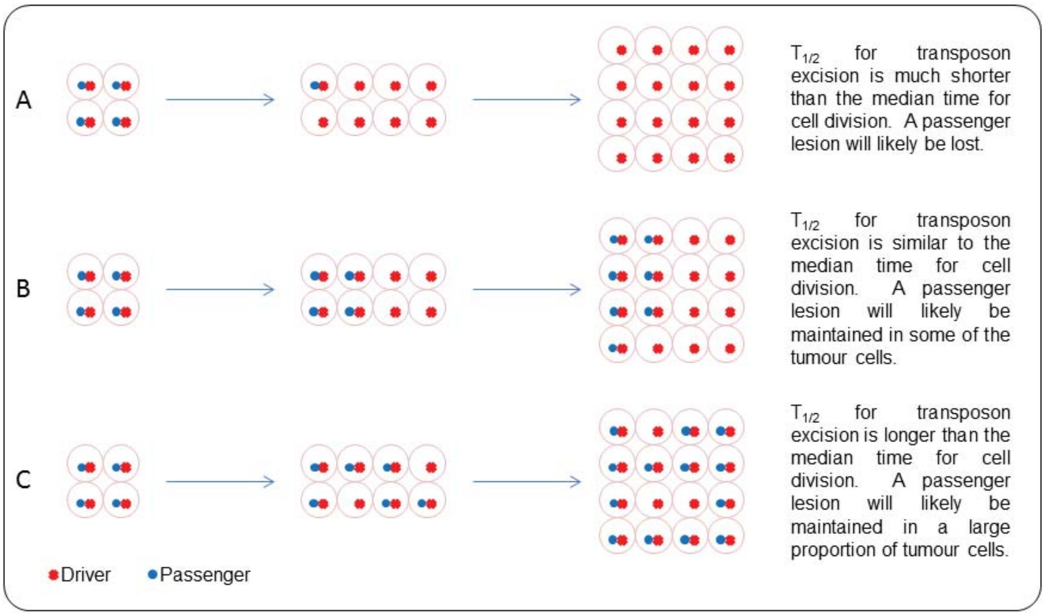

A
B
C
$T_{1/2}$ for transposon excision is much shorter than the median time for cell division. A passenger lesion will likely be lost.
$T_{1/2}$ for transposon excision is similar to the median time for cell division. A passenger lesion will likely be maintained in some of the tumour cells.
$T_{1/2}$ for transposon excision is longer than the median time for cell division. A passenger lesion will likely be maintained in a large proportion of tumour cells.
Driver
Passenger

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

In several cases some of the transposon integrations found in the final tumour were present in the sequential blood samples for over two months prior to the development of overt leukaemia. This implies the continuous contribution by HSCs with these mutations to haematopoiesis during this interval. The disappearance and re-emergence of other integrations may reflect the proliferative cycles of HSCs with these mutations, or more commonly, the small percentage of cells carrying them and their variable detection by the assay. In some instances this is because of the variable DNA yield from the serial blood samples. For example, in sample 16.3g the paucity of integrations in the 32 week blood sample (20 weeks post pIpC) is likely to be due to a low DNA yield. Although this sample was processed twice for splinkerette and sequencing, a low number of reads were mapped on both occasions. However, there are also gaps in the presence of mutations in some samples in which a large number of integrations were mapped. This may occur because these integrations are carried in only a small proportion of cells and they fall below the limit of detection or because of a paucity of *Mbo1* restriction sites near to the particular integration which biases the PCR amplification against detection of these integrations. For example, in case 19.2D the integration in *Mll3* (5:25000472) is only 19 nucleotides from the nearest restriction site at one end and there is no restriction site within a thousand bases on the other side of the transposon. If the DNA is completely digested, the transposon DNA fragments containing this integration will be either too short to map or too long to PCR efficiently in the Splinkerette protocol. Samples will vary in the number of easily amplified integration sites, which could differentially affect the rate of PCR amplification of integrations sites that are more difficult to amplify, such as those several hundred bases from the nearest *Mbo1* restriction site. In mouse 6.4g, there are a group of integrations 'missing' from the week 75, 77, 79, 81 and 83 blood samples even though all of these had comparatively good read depth and many other integrations did persist. The finding that these integrations are present in the final tumour, including two which persist on the serial transplants (Dtx2 and Cdyl2), indicates these integrations were not lost, as it would be unlikely that a new integration event occurred at exactly the same site and even more improbable that this happened at multiple positions. It is possible that these integrations occurred in a small clone that fell below the limit of detection, but re-emerged upon acquisition of a driver later on.

Despite such limitations there is important information to be gained by studying the order of acquisition of mutations in these serially bled samples. The serial CIS analysis revealed for the first time, that the *Csf2*, *Nup98* and *Nf1* CIS were all identifiable two weeks before death. *Csf2* integrations appeared as either early events several weeks before any demonstrable changes in blood counts (19.2b, 19.2d, 21.3j, 20.2b) or as late events just as the WCC became abnormal (16.3b, 16.3e, 16.3f, 7.5b, 22.2b). This is also the case for *Nup98* in which integrations were often first detected several weeks before death (19.2b, 19.2d, 6.4g, 7.5b) but were also seen as a late event (7.5c). Similarly *Nf1* integrations occurred as both early and late events and often multiple integrations were detected in the same tumour (e.g. 7.5b, 19.2b, 19.2d). In contrast, where *Mll1* integrations were detected in the serial blood samples they were always late (16.3b, 16.3f, 21.3j). Integrations in *Flt3* were universally in intron 9 in the forward orientation as in the published study suggesting these are activating integrations. Typically these were late events (7.5c, 19.2d, 19.2a) except in 16.3f where a single read was detected in the week 37 blood sample. This integration was not detected again until a blood sample a day before death. It is difficult to be certain of the veracity of this early integration, particularly as it was a single read detected from one end of the transposon, however if real this would imply the combination of *Npm1$^{cA}$* and up-regulation of *Flt3* alone is insufficient for leukaemogenesis. However, the combination of *Npm1$^{cA}$* and *Flt3-ITD* mutations in mice was previously found to cause highly penetrant AML with an explosive onset and short latency (Mupo et al., 2013).

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

Precisely how polyclonal these tumours are and how the transposon copy number effects this clonality is yet to be determined. It is likely the number of potentially leukaemogenic clones varies between mice and this may, in part, explain the large variation in the number of CIS integrations identified in each mouse in this and the GRH model. It is also likely that the different integrations are associated with varying levels of fitness advantage. Some mice may develop leukaemia with only one or two 'strong' integrations whereas others may have multiple 'driver' integrations in a single clone as each one provides only a 'weak' advantage. Whether the differences in the CIS identified between the two cohorts reflects a biological difference due to the mutation rate is uncertain.

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

It is striking that the top three CIS; *Csf2*, *Nf1* and *Nup98,* are identical between the GRL and GRH screens, although of these only *Nf1* is recurrently mutated in human myeloid leukaemia(Boudry-Labis et al., 2013; Haferlach et al., 2012; Parkin et al., 2010). Transposons do not recapitulate the type of mutations seen in human disease. This is often seen as a limitation of transposon mutagenesis screens, however it is also an advantage as genes and pathways that are less susceptible to natural mutational processes may be identified. The recurrent integrations upstream of *Csf2*, the gene which encodes GM-CSF, are likely to be one such example. In the GRL model presented here, *Csf2* integrations were almost universally in the forward orientation suggesting these are activating mutations. Furthermore, on serial transplantation they typically persisted as opposed to the vast majority of integrations, suggesting they have a driver role in these tumours. Similarly, in the published GRH model the *Csf2* integrations were universally in the forward orientation and resulted in marked overexpression of *Csf2* mRNA and increased GM-CSF levels in leukaemia cell supernatants(Vassiliou et al., 2011).

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

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

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

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

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

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

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

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