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

## 5.1 Introduction

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

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

## 5.2 Results

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

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

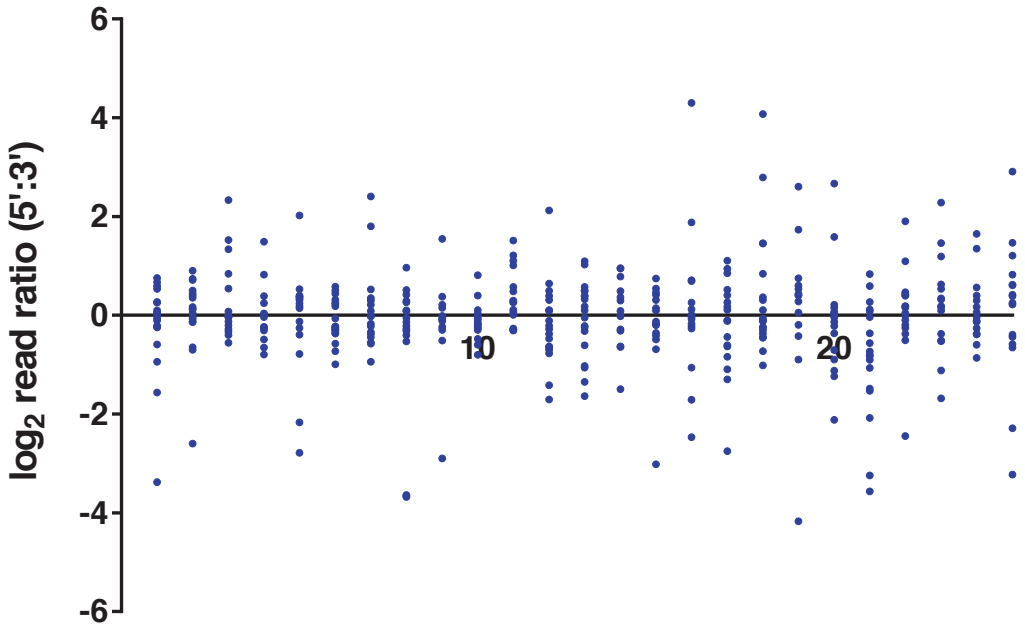



**Figure 5.1: Correlation of 5' and 3' reads.** The 5' to 3' ratio for the 25 integrations with highest coverage in each sample after removal of duplicates are shown for the leukaemias from 19 IM mice in the serial bleed study (chapter 4). The $\log_2$ of the ratio of the 5' to 3' reads is shown. Each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

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

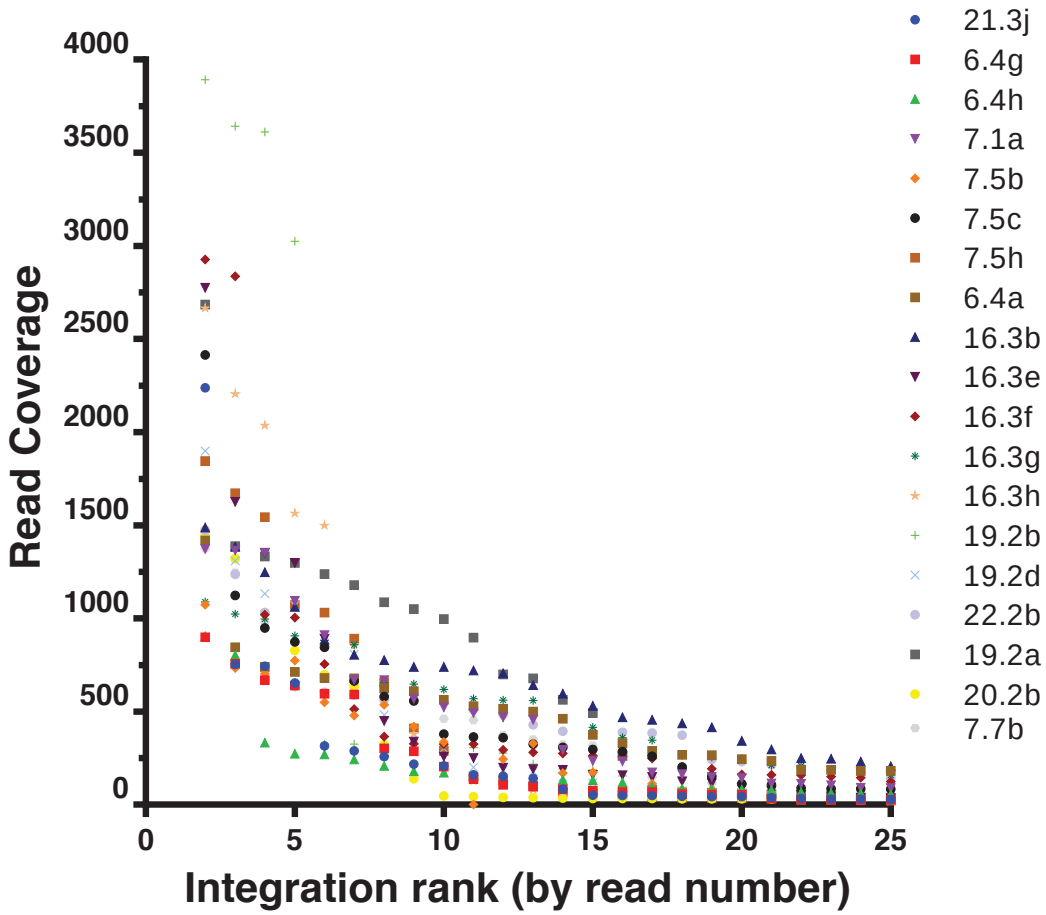



**Figure 5.2: Number of reads per integration**. Data is shown for the top 25 integrations by read number in the leukaemias from 19 serially bled mice (Chapter 4) after removal of PCR duplicates in the analysis.

### 5.2.2 TraDIS Identifies Additional CIS Compared to Restriction-Based Mapping

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

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

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

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

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

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

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

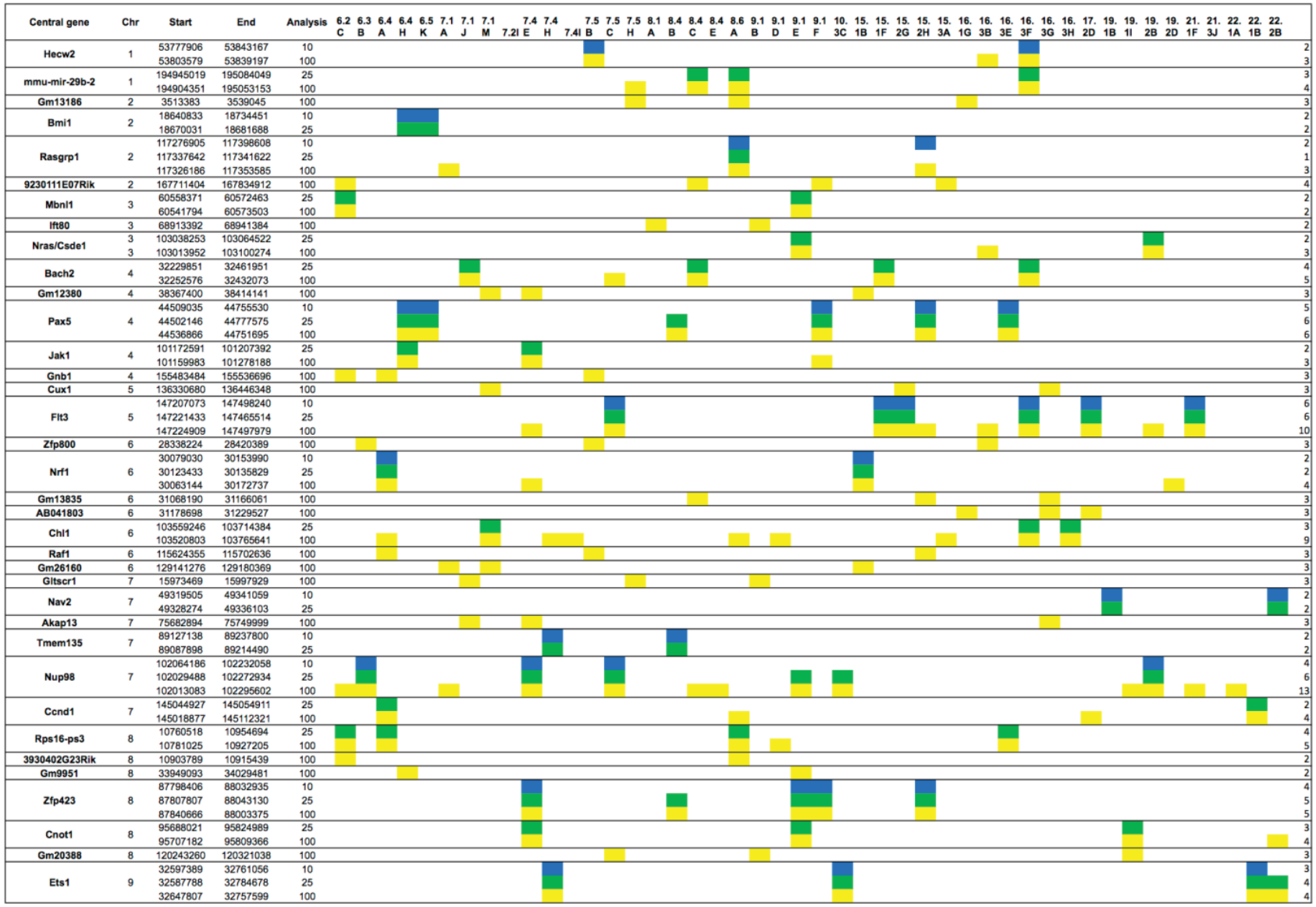
**Table 5.2: CIS integrations identified with the top 25 integrations per sample.** CIS that were excluded and the reason for their exclusion are shown in appendix 5b. Otherwise the features of the table are similar to table 5.1

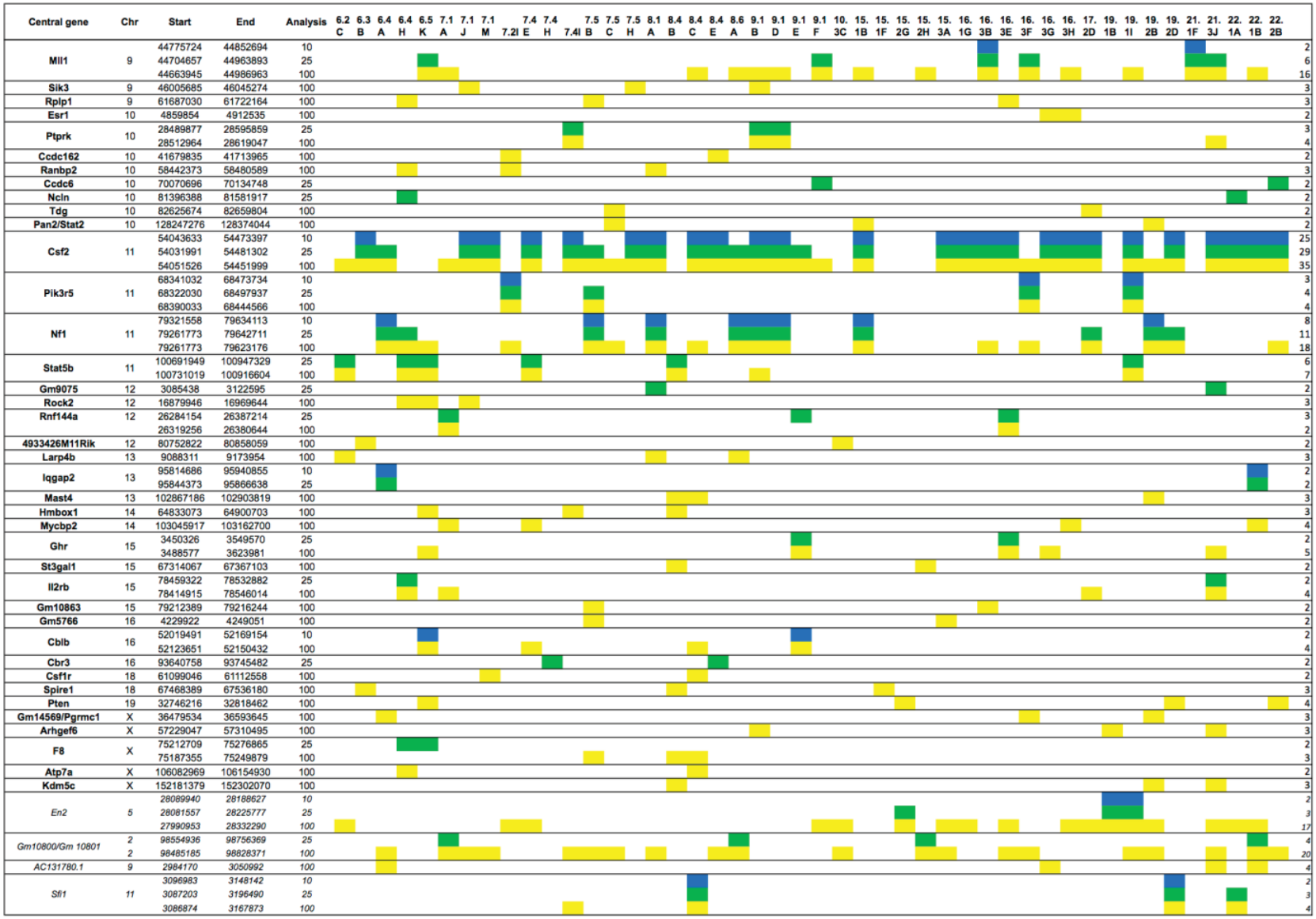
| Chromosome | Minimum Peak Location | Maximum Peak Location | Peak Height (range) | CIS Start | CIS End | Number of hits | Number of tumours | Smallest p Value | Kernel scales | Genes - largest CIS | Genes - smallest CIS | Gene Nearest Peak |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11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

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

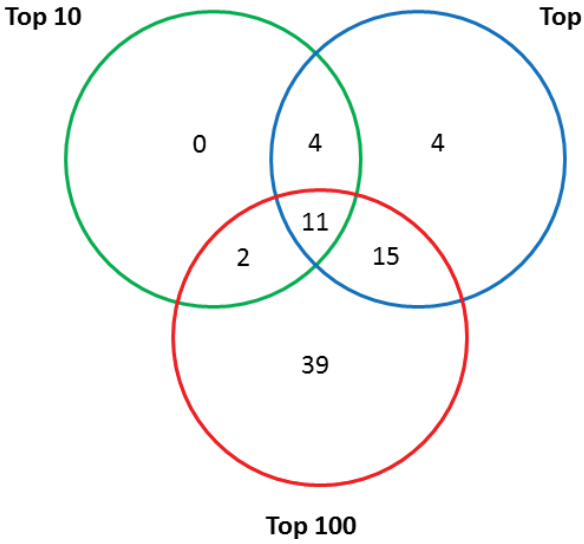
**Table 5.3: CIS integrations identified with the top 100 integrations per sample.** CIS that were excluded and the reason for their exclusion are shown in appendix 5b. Otherwise the features of the table are similar to table 5.1.

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


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

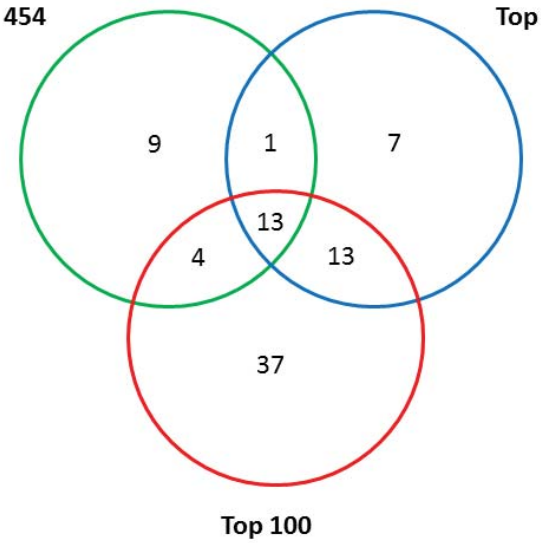


| Central gene | Chr | Start | End | Analysis | |
|---|---|---|---|---|---|
| Mll1 | 9 | 44775724 | 44852694 | 10 | 2 |
| | | 44704657 | 44963893 | 25 | 6 |
| | | 44663945 | 44986963 | 100 | 16 |
| Sik3 | 9 | 46005685 | 46045274 | 100 | 3 |
| Rplp1 | 9 | 61687030 | 61722164 | 100 | 3 |
| Esr1 | 10 |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




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

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




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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

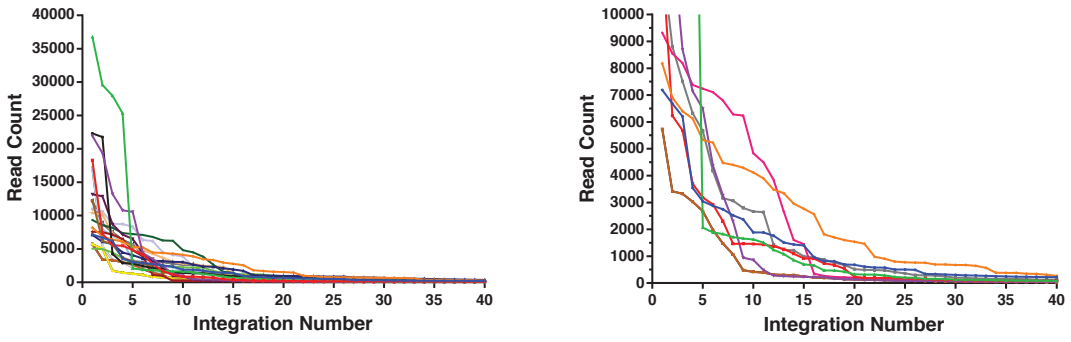



**Figure 5.5: Read coverage for the major integrations without removal of duplicates. Left:** Total 5' plus 3' read coverage for the top 40 integrations in the spleen samples from the 19 mice in the serial bleed study (chapter 4). **Right:** Closer view of the fall in read count in 8 selected samples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

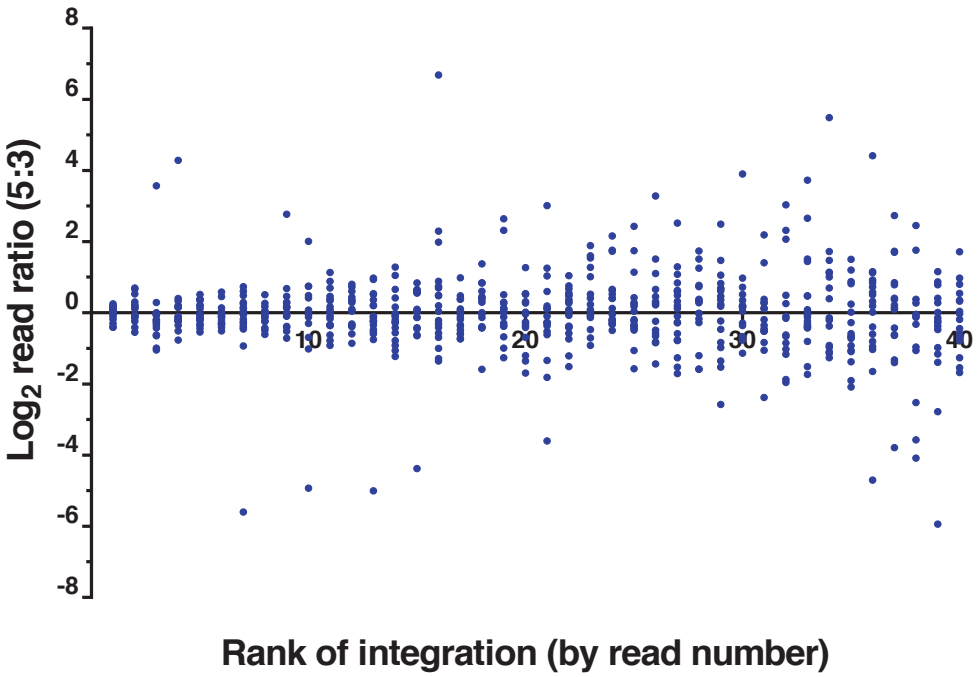



**Figure 5.6: Correlation of 5' and 3' reads in the non-duplicate filtered analysis.** The 5' to 3' ratio for the 40 integrations with highest coverage in each sample are shown for the 19 mice in the serial bleed study (chapter 4). The $\log_2$ of 5' and 3' read ratio is shown. 95 of the 760 integrations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

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

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

#### *5.2.4.1 Npm1$^{cA}$/GRL 19.2B*

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

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

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

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

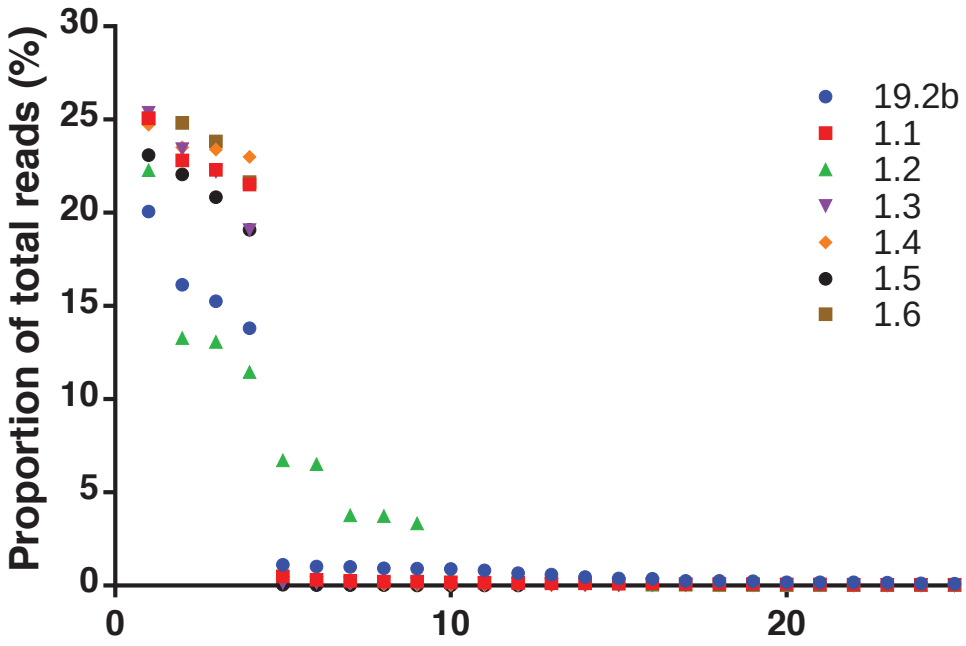



**Figure 5.7: Proportion of total reads taken by the top 25 integrations in tumour 19.2b and associated recipient tumours.** In each tumour the top four hits were identical and it was only in the primary tumour and 19.2b.1.2 that other transposon integrations were found in any number.

#### *5.2.4.2 Npm1$^{cA}$/GRL 21.3j*

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

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

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

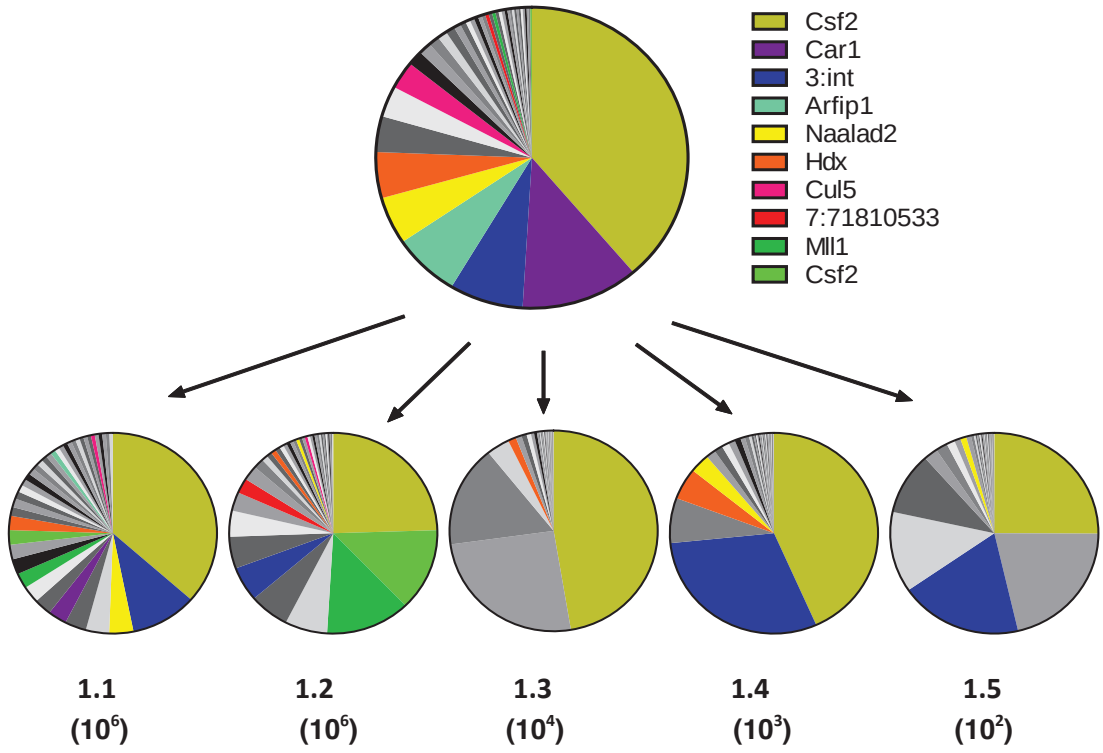



**Figure 5.8: Shared integrations in primary tumour 21.3j and five recipient tumours**. The top 40 integrations by read number are represented. Those shown in colour are shared between different tumours, but those in greyscale are not. The integrations are represented as a proportion of the total reads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

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

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

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

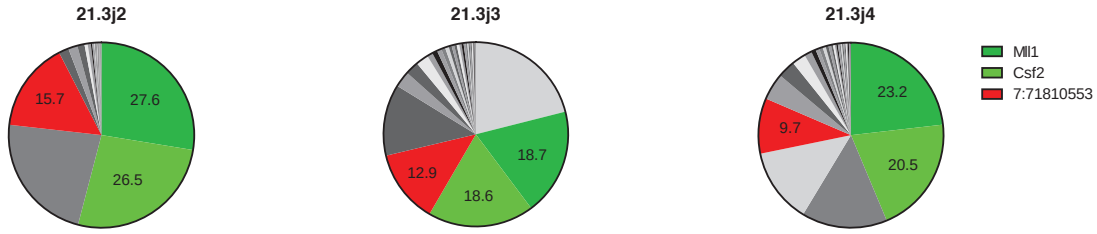



**Figure 5.9: Transposon integrations in leukaemias generated after transplantation of one of three single cell-derived colonies from primary 21.3j**. Identical integrations are depicted in the same colour (also used in figure 5.9) in three different recipient leukaemias. Numerals represent percentages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

#### *5.2.4.3 Npm1cA/GRL 16.3f*

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

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

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

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

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

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

#### *5.2.4.4 Npm1$^{cA}$/GRL 6.4a*

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

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

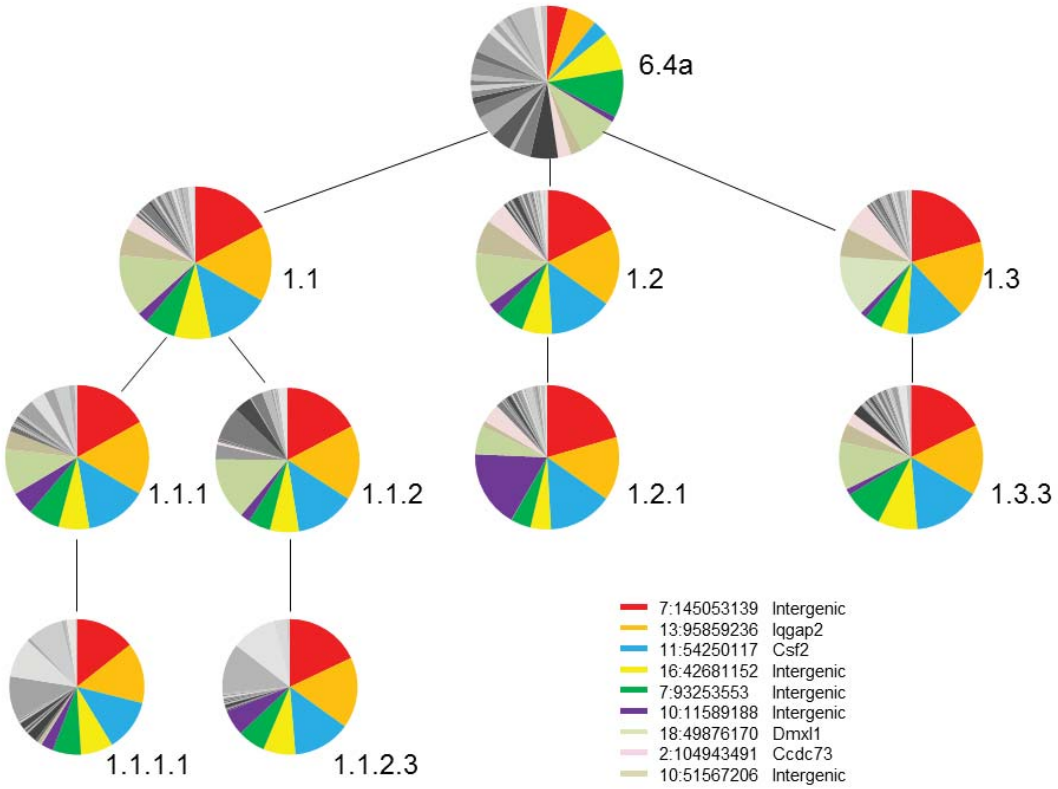



**Figure 5.10: Shared transposon integrations in primary tumour 6.4a and 9 of its recipient tumours.** The shared integrations are plotted in colour and the identity of these integrations is indicated. Integrations shown in grey-scale differ between the tumours.

### *5.2.4.5 Npm1$^{cA}$/GRL19.2d*

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

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

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

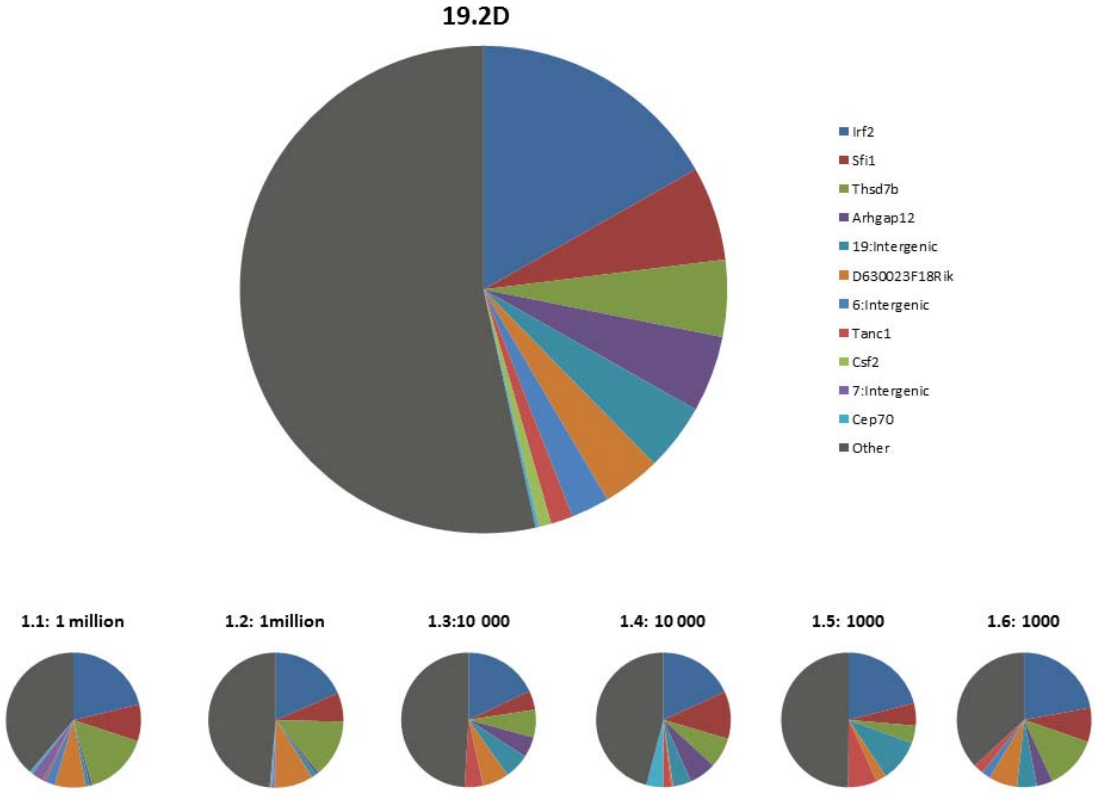



**Figure 5.11: Major transposon integrations in 19.2d and its recipient tumours.** The cell doses for each of the transplants are shown.

**A**

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

**B**

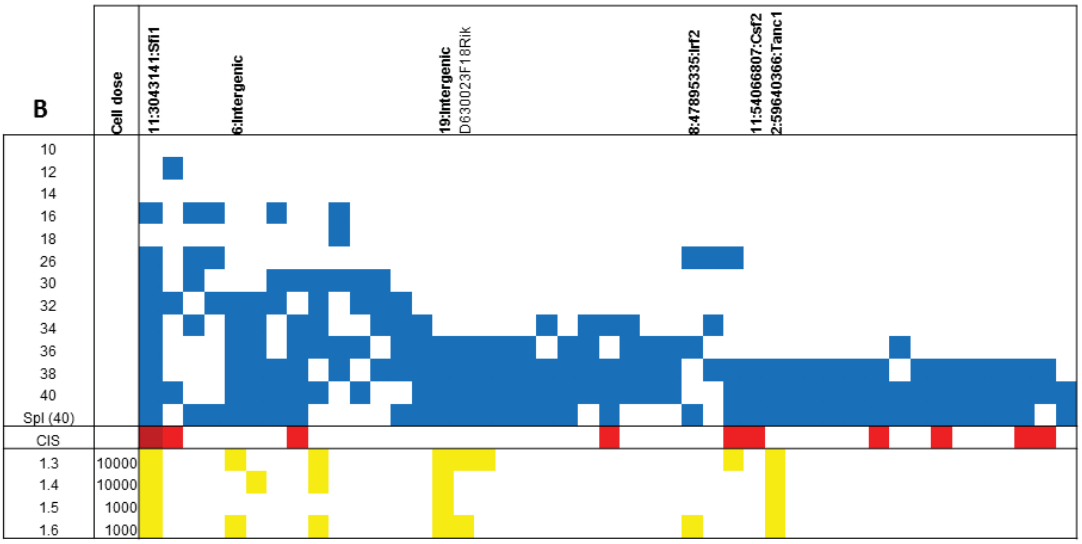



**Table *5.9:* Timing of major tumour integrations in the serial blood samples (A)** The proportion of reads taken by the transposon integrations that persisted in multiple recipient tumours are shown for each of the serial blood and tumour samples. **(B)** The presence of these integrations in the same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

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

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

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

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

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

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

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

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

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

**A**

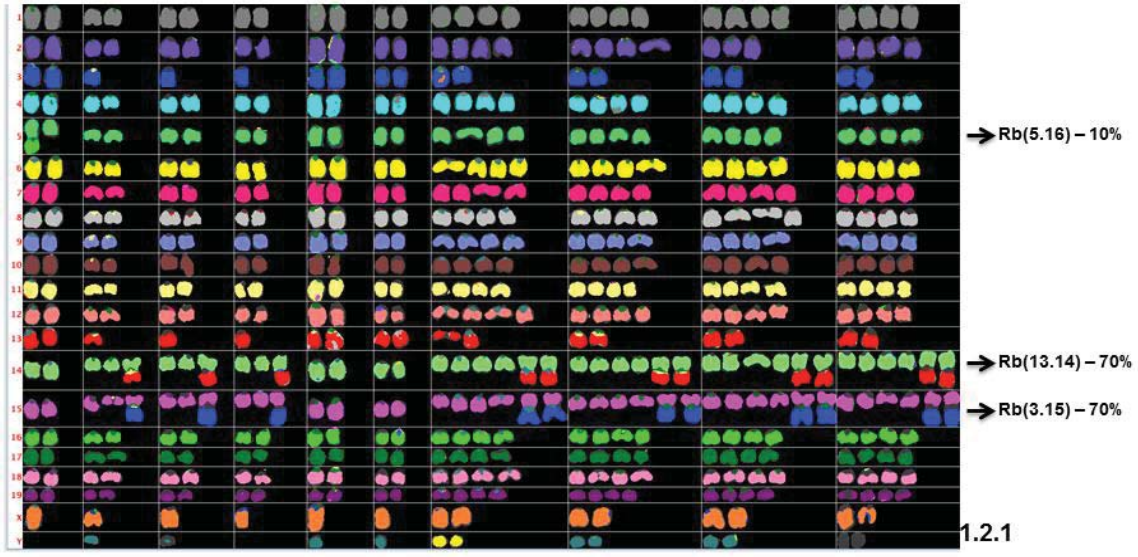
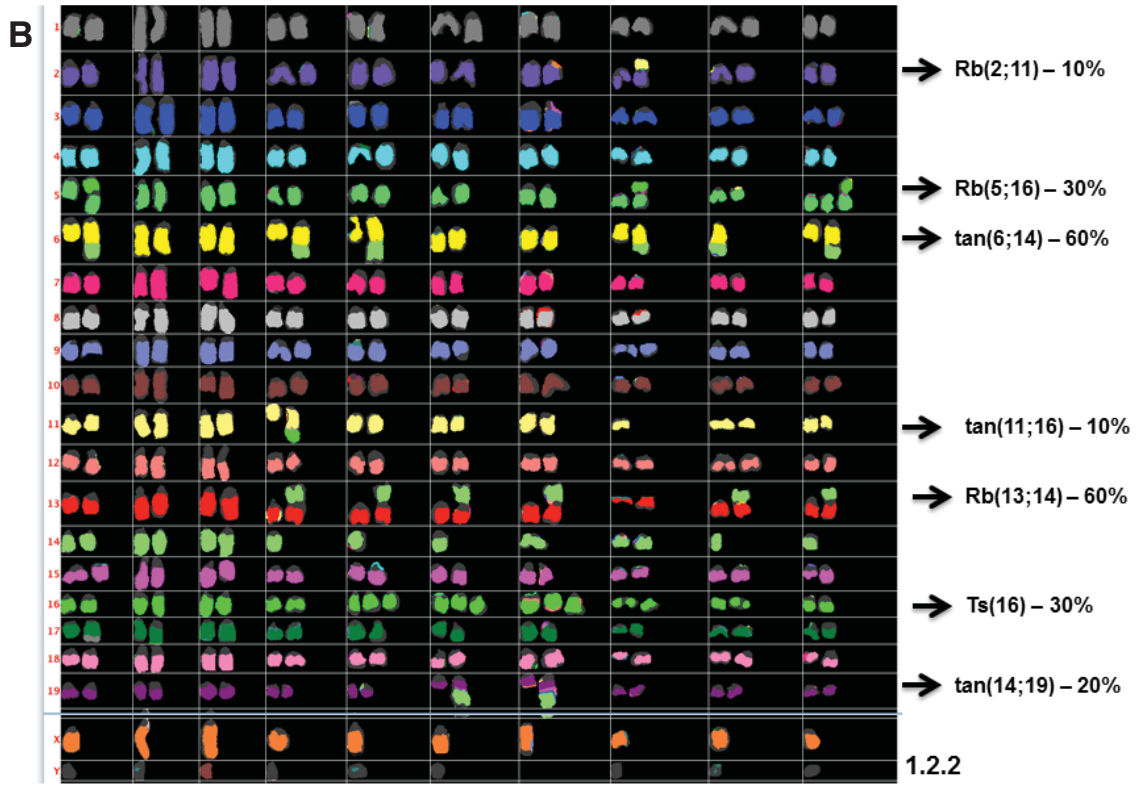
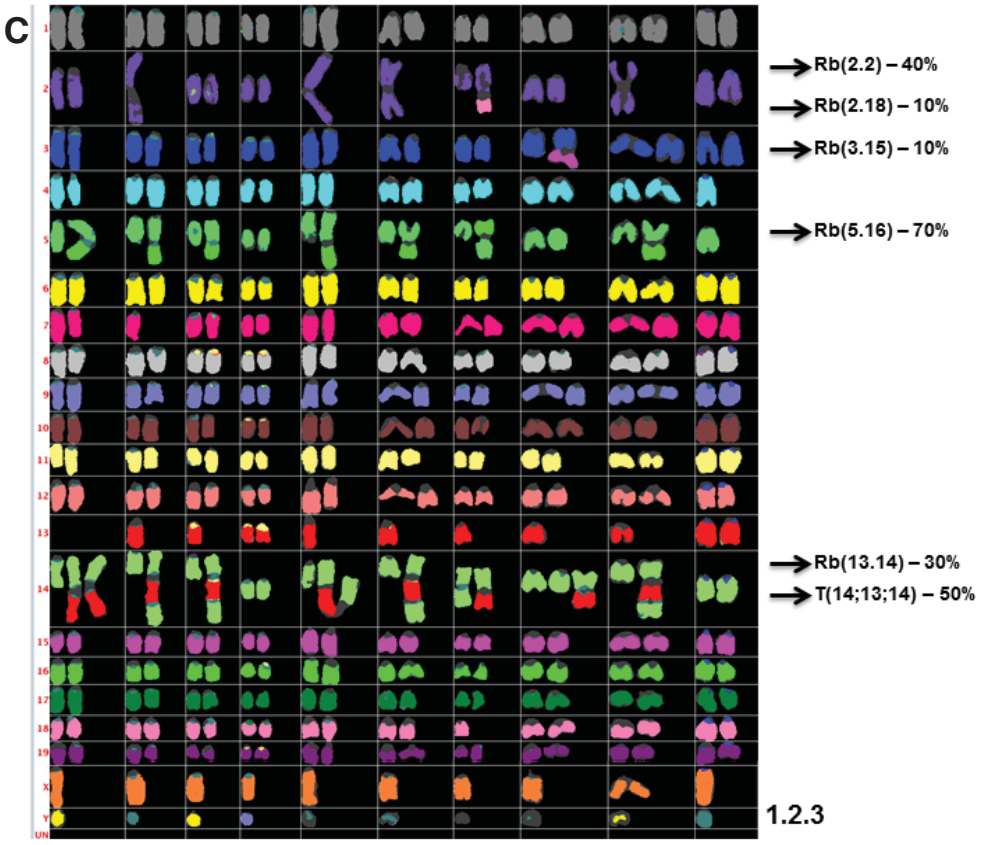



**Figure 5.12: Metaphase paint images of transplants 1.2.1 (A, above), 1.2.2 (B, next page) and 1.2.3 (C, next page), showing Robertsonian translocations in all cases.** In 1.2.1 there is tetraploidy in 4 metaphases in addition to the indicated Robertsonian translocations involving chromosomes 3, 5,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


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


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

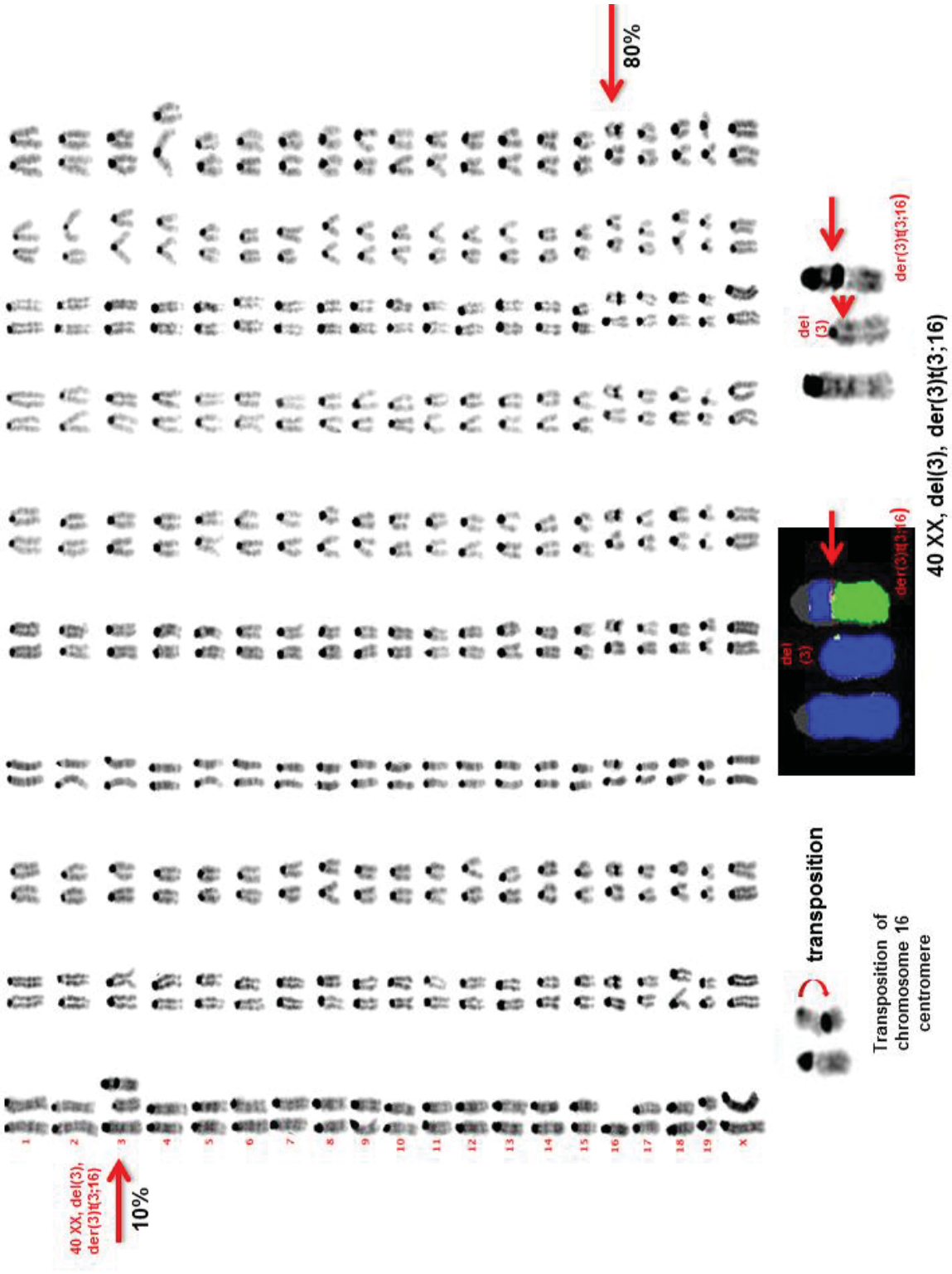
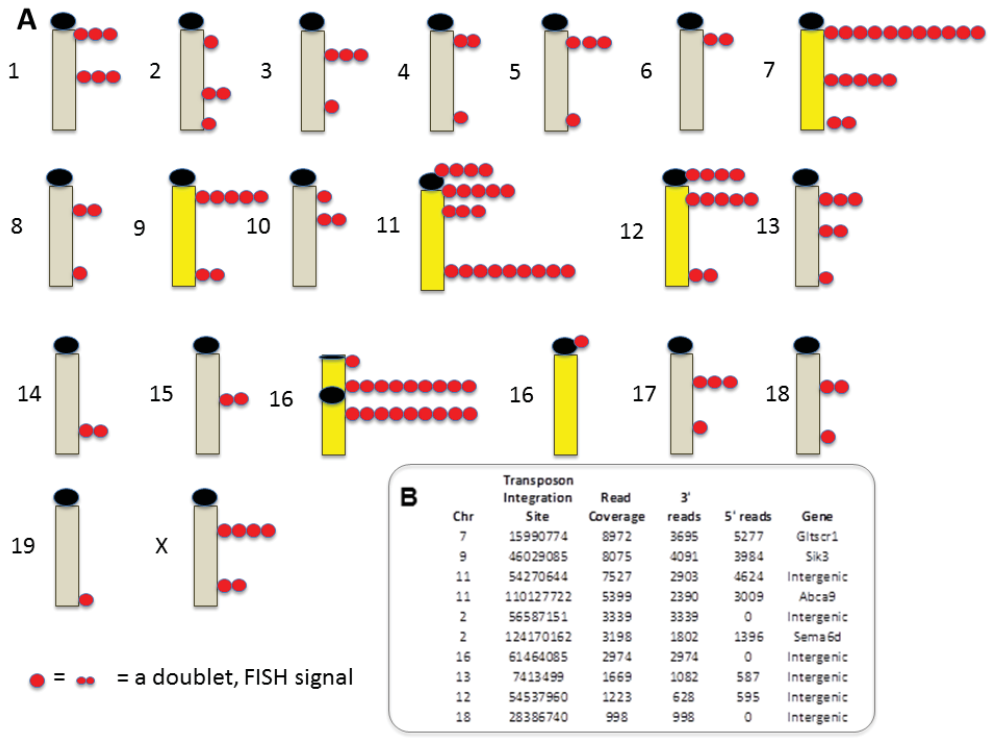



**Figure 5.13: Metaphase karyotyping of primary tumour 7.5h.** This showed del(3), der(3)t(3;16) in one metaphase and transposition of chromosome 16 centromere within the long arm of chromosome 16 in eight. Close up images of the abnormalities are shown at the bottom, including a metaphase paint image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

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




**B**

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

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

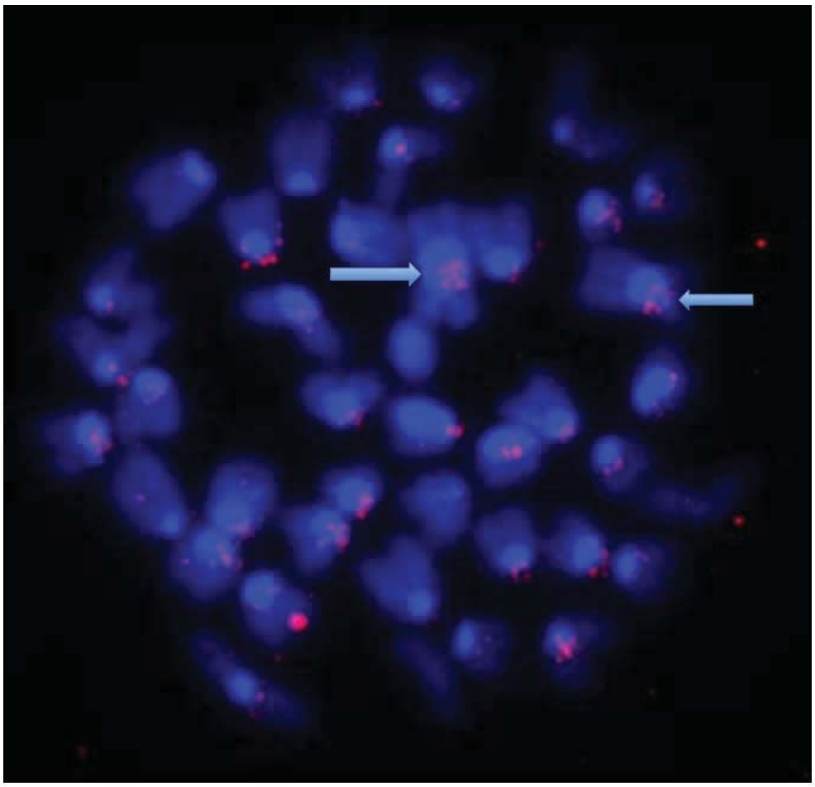


**Figure 5.15: FISH of a metaphase from 7.5h recipient tumour 1.2.3.** The transposon integrations are indicated by red double dots and the centromeres fluoresce bright blue. The arrows indicate some of the clear transposon integrations within centromeres.

## 5.3 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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