Chapter 2: Materials and Methods

2.1 Sequencing of human leukaemia samples

2.1.1Exome Sequencing and genomic alignment

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

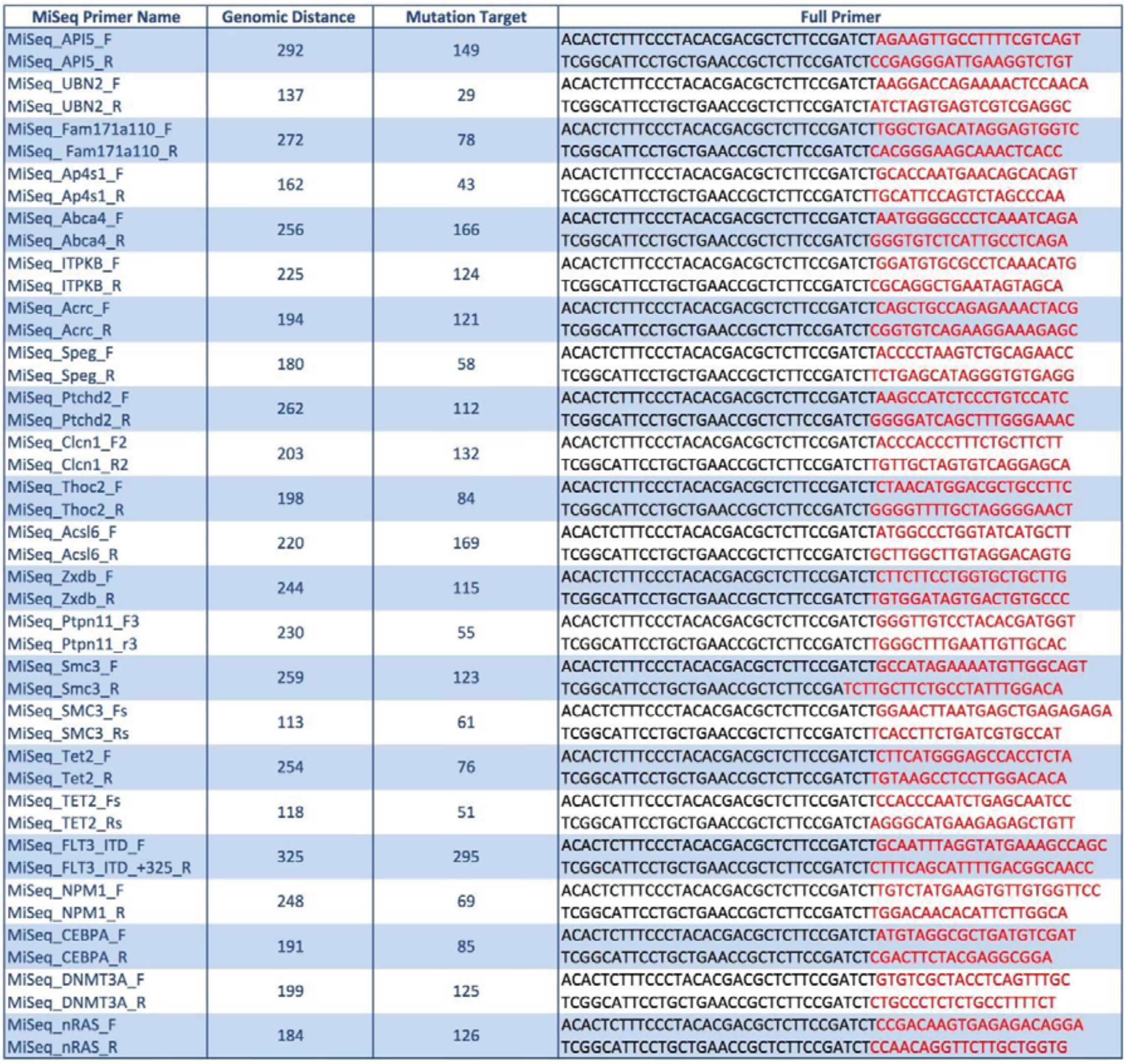
Library preparation, sequencing and variant calling were done through the Sanger pipeline. Genomic libraries, enriched for protein coding exons were generated by hybridisation to RNA baits using Agilent SureSelect Human Exon 50Mb Kit (Agilent, S02972011). The libraries were analysed on the Illumina HiSeq2000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

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

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

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

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



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

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

Alignment of MiSeq reads to the reference genome was performed by Ignacio Varela (Universidad de Cantabria). A modified Bayesian Dirichlet process to allow for multiple sample analysis was performed by David Wedge. In brief, subclonal clusters of mutations were identified using a previously described Dirichlet process, implemented using a Markov Chain Monte Carlo (MCMC) method (Bolli et al., 2014; Nik-Zainal et al., 2012). The method is summarised by David Wedge as follows: 'From the MCMC assignment of mutations to clusters, the most likely configuration of clusters and node assignments was obtained using a stepwise, greedy expectation-maximization (EM) algorithm which alternately added a node and shuffled mutations between nodes until no further improvement in the agreement with the posterior distribution from the MCMC sampling could be made. The best set of clusters was then chosen using the Bayesian information criterion(Schwarz, 1978). The Dirichlet process was run 5 times, each time for 10000 MCMC iterations. Mutations were assigned to the same clique if every mutation in the clique appeared in the same cluster as every other mutation within the clique in most (i.e. 3 or more) of the runs'.

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

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

2.2 Mice

2.2.1 Mouse Strains used in the *Sleeping Beauty* Study

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

2.2.2 Transplant of NSG mice

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

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

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

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

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

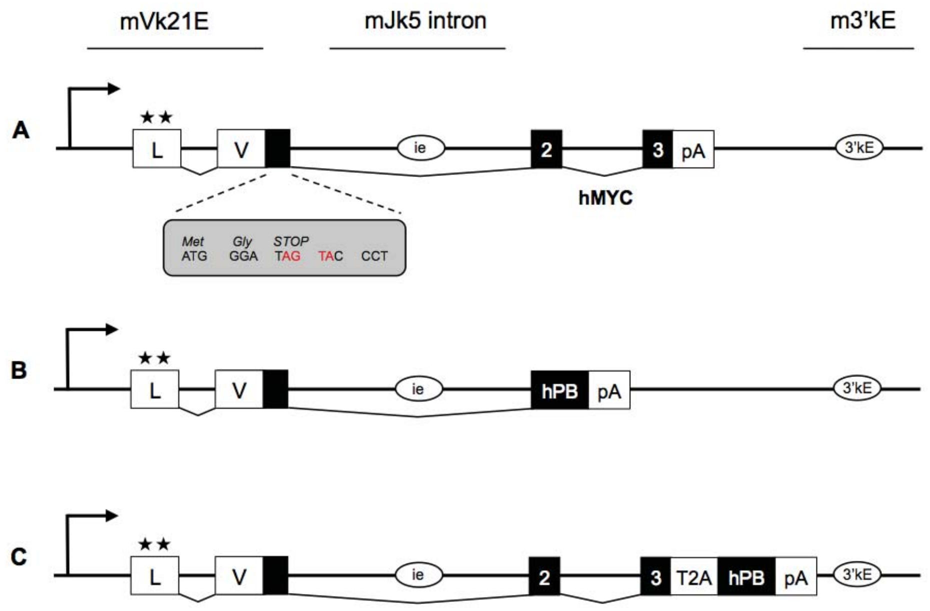


Figure 2.1: The *Vk*MYC, Vk*hPB* and *Vk*MYC-TA-hPB* constructs.

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

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

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

Primer	Sequence
hMYCTAHyperPB_F	TCACTATAGGGAGACCCAAGC
hMYCTAHyperPB_FseI_R	ACTTCAGGCCGGCCCATAGAGCCCACCGCATC
VkmPBHyperPB_F	ATTCTTCCTCAGCCCCTCAA
VkmPBHyperPB_FseI_R	ACTTCAGGCCGGCCCATAGAGCCCACCGCATC

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

2.2.4 Genotyping Transgenic Mice

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

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

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

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

2.3 Sample Collection and Processing

2.3.1 Collection and processing of blood samples from live mice

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

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

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

2.3.2 Necropsy of sick mice, sample collection and processing

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

Blood was collected at necropsy by intra-cardiac aspiration, placed into EDTA and serum tubes and processed as described in 2.3.1. At necropsy, gross examination of the internal organs was performed with particular attention to the spleen, thymus, lymph nodes, liver and kidney. Macroscopic abnormalities and the weights of the whole mouse, one kidney (*PB* cohort only), spleen and liver were recorded. Samples of spleen, tail and any macroscopically abnormal tissue were collected into RNAlater and kept at room temperature for 24-48 hours before storage at -20°C. For later collection of live cells, bone marrow (BM), spleen or other abnormal tissue were collected into PBS ± 2% fetal calf serum (FCS).

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

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

2.3.3 Processing of live cells

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

Spleen and tumour samples were gently squashed in 5mL of 0.85% ammonium chloride using the end of a 5mL syringe, and the solution was then pipetted up and down to create a cell suspension. This was filtered to remove cell clumps, transferred into a 15mL falcon tube and RPMI 1640 media (+10% FCS and 1% glutamine-penicillin-streptomycin) was added and the red cell lysed samples were spun for 5 minutes at 250g. A wash step was performed and an aliquot of this solution was taken for counting. Cells were re-suspended at a concentration of 2 x 10^7 cells/mL, frozen in 10 million cell aliquots in a 50:50 mix with 2x RPMI freezing media (60% RPMI, 20% FCS and 20% dimethylsulphoxide (DMSO)), and stored initially at -80°C before transfer to liquid nitrogen. Aliquots of 0.5mL of the cell suspension were also spun down in microtubes and the cell pellets were re-suspended in 1mL Trizol and stored at -80°C, or stored as a frozen cell pellet at -80°C.

2.3.4 Generation of single cell derived haematopoietic colonies for transplant

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

2.3.5 Preparation of Metaphase Spreads and FISH analysis

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

The transposon specific probe was prepared by digesting the *pA6GrOnc* plasmid with *AflII* to generate a 2.5kb transposon specific probe. Probe amplification, labelling and preparation and processing of slides was performed by Ruby Banerjee (WTSI FISH facility) as previously described (Rad, 2010).

2.3.6 DNA extraction

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

2.3.7 Exome Sequencing of Mouse *SB* Tumours

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

2.3.8 Comparitive Genomic Hybridisation (CGH)

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

2.3.9 RNA extraction

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

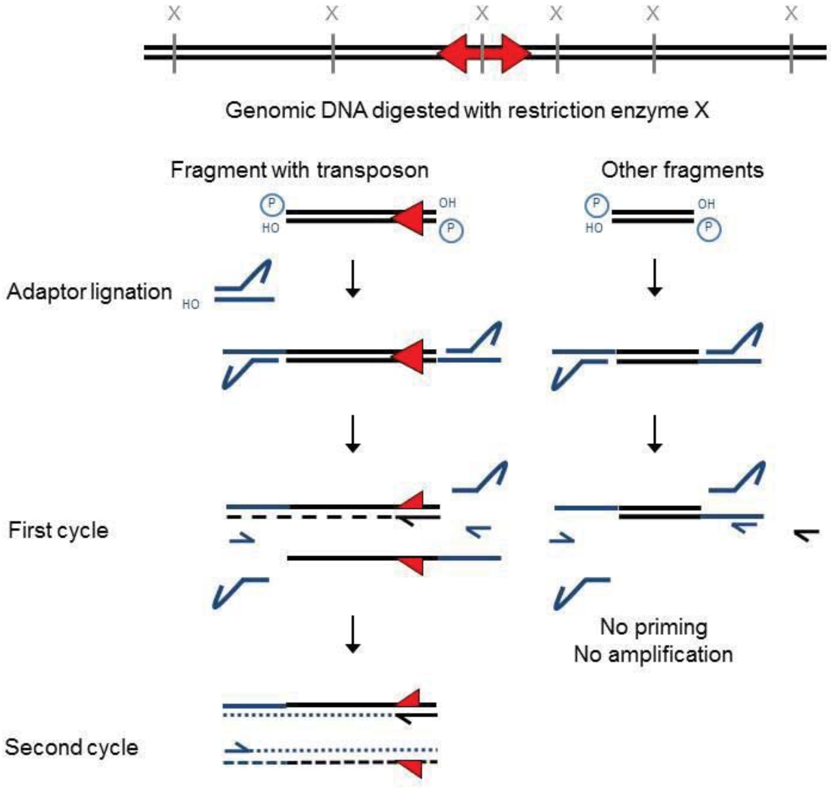
2.4 Sequencing transposon integration sites: the Roche 454 Method

2.4.1 Splinkerette PCR to identify transposon integration sites

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

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

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



Adapted from Li et al 2010

Figure 2.2: Principle of the Splinkerette PCR

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

Splinkerette Linkers	
HMSpAa	CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGTGTCGACACTAGTGG
HMSpBb-Sau3AI	GATCCCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCAAAAAAA
Splinkerette Primers	
HMSp1	CGAAGAGTAACCGTTGCTAGGAGAGACC
SB-5'-Sp1	TAGTGTATGTAAACTTCTGACCCACTGGA
SB3'_altP1	AACTGACCTTAAGACAGGGAATCTT
HMSp2_454_new2010_R	CTATGCGCCTTGCCAGCCCGCTCAGGTGGCTGAATGAGACTTGGTGTCGAC
BC454-SB1	CGTATCGCCTCCCTCGCGCCATCAGACACATACGCGTGTATGTAAACTTCCGACTTCAAC

Table 2.5: Splinkerette linkers and primers

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

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

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

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

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

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

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

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

Due to the local hopping phenomenon a single tumour could contain multiple integrations around a site. To minimise the impact of local hopping, integrations in a single tumour that occurred within a 10Kb window were collapsed to a single integration for the CIS analysis. This was the same method as used in the published *Npm1*^{cA} high copy (GRH) transposon model (Vassiliou et al., 2011).

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

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

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

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

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

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

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

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

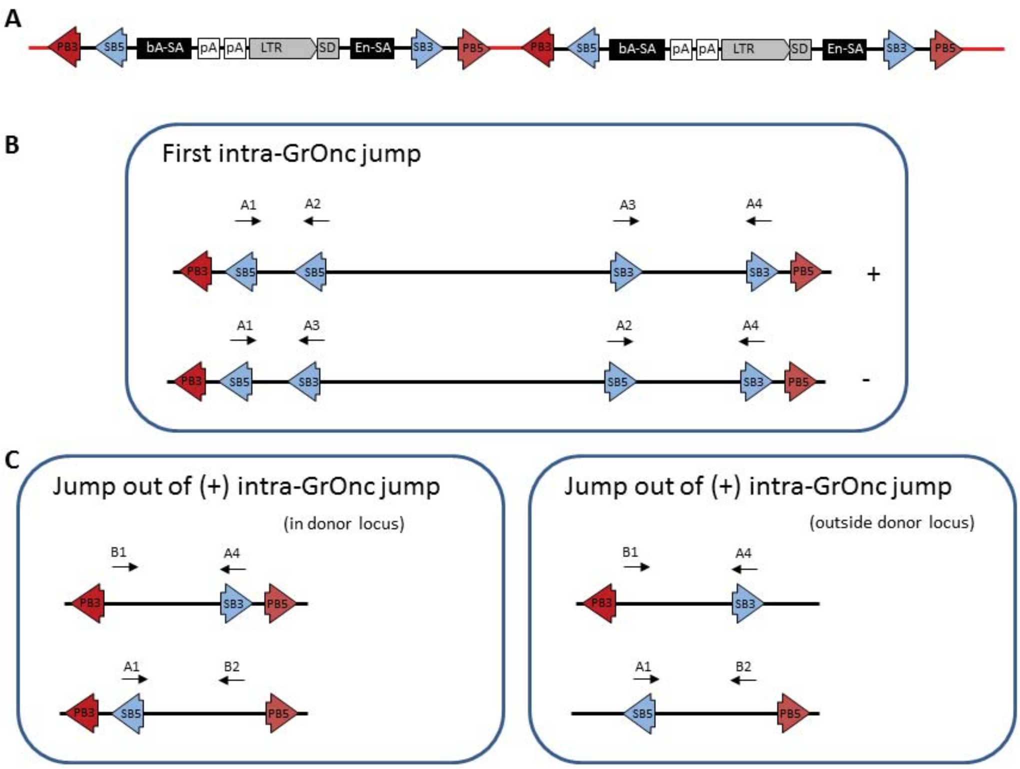


Figure 2.3: Positioning of primers to detect jumping of the SB transposons into adjacent transposons. Two adjacent *GrOnc* transposons in the donor locus (A), which have the outer PB repeats present unlike two transposons mobilising together to a new locus by SB. Example of one GrOnc transposon jumping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

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

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

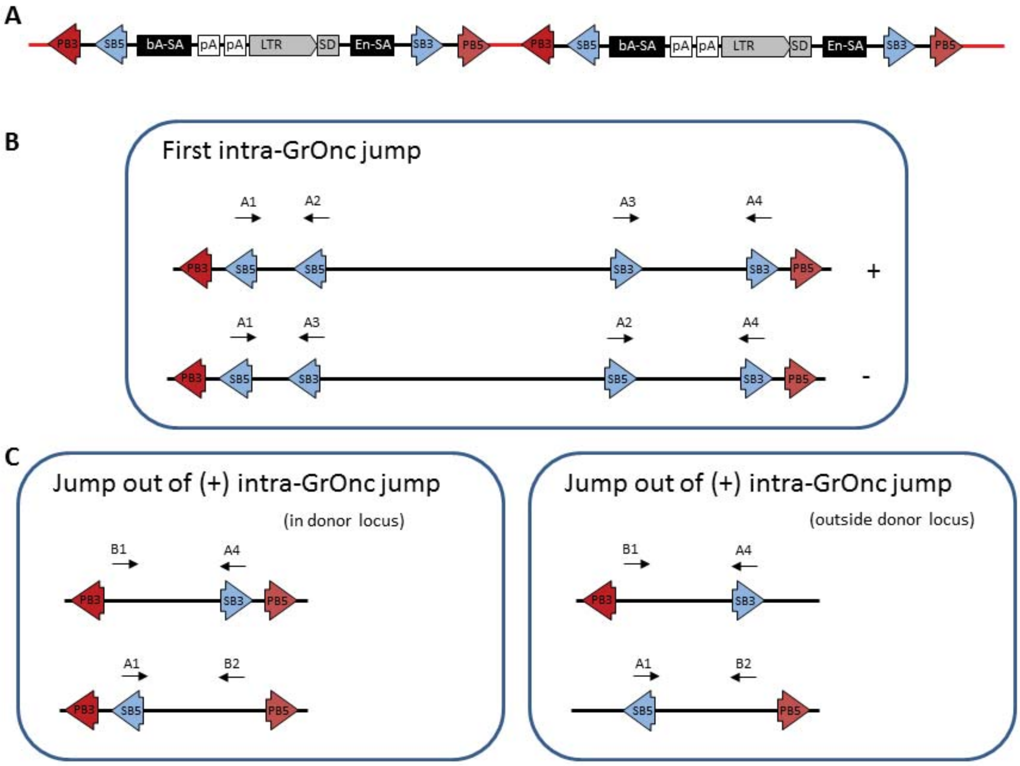


Figure 2.4: Transposon Neopartnerships. A. Two adjacent transposons that have jumped together into a genomic locus. Note the opposite facing *PB* repeats are transported as 'cargo' between them. **B**. One *SB* transposon integrates into the adjacent transposon by local hopping. **C.** Two possible re-mobilisation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

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

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

2.5 Illumina Sequencing of Transposon Integrations

2.5.1 Library Preparation

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

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

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

End-repair of the sheared and cleaned DNA was performed in a total volume of 100µL: 10 x T4 DNA ligase buffer, 10µL; T4 DNA polymerase, 7 µL; T4 PNK, 7µL; dNTPs, 6µL; Klenow DNA polymerase 2µL; sheared DNA 78µL. After incubation at 20°C for 30 minutes the samples were purified using the QIAGEN 96 well plate column system and eluted in 27µL of EB buffer. An A tail was added using the following reaction: Klenow fragment exo-, 4.5µL; dATP, 15µL; 10 x Klenow buffer, 5µL; incubate at 37°C, 1 hour. The samples were eluted through MinElute columns in 20µL of EB buffer and 1µL of the sample was run on a 7500 Agilent chip. The splinkerette adaptors were ligated using NEBNext(TM) DNA Sample Prep Reagent Set 1 (NEB: E6000B-SS) using the reaction parameters as follows: 18µL A-tailed DNA; 25µL 2x ligation buffer; 1µL MilliQ water; 5µL ligase; 30 to 60 minute incubation at 20°C. Clean up of the DNA was performed using a double SPRI bead (AMPure XP) purification, using 50µL beads to 50µL of adapter ligated DNA in the first reaction and 40µL (0.8x) in the second in a similar method to that described above (SPRI bead purification of MiSeq library). The final volume of the purified, adapter ligated DNA in EB buffer was 30µL and 1µL of this sample was run on the 7500 Agilent chip to determine whether the adapter was successfully ligated. The size of the adapter ligated DNA should be approximately 100bp larger than the pre-ligated library.

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

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

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

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

2.5.2 Transposon mapping and CIS analysis of Illumina data

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

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

Read pairs which mapped in the expected orientation on the same chromosome, irrespective of the insert range, were included in further analysis. All other reads, including those where the placement was ambiguous or where mates aligned to different chromosomes, were discarded. The reads were filtered further by applying thresholds of the mapping quality score of 20 (expected mapping error rates of less than 1 in 100) and of the Smith-Waterman alignment score of 30 (using standard affine gap penalties of 1, -2, -4 and -3, respectively, for matches, mismatches, gap openings and gap extensions). In addition the presence of the integration motif, TA for *SB*, TTAA for *PB*, at the mapped location of the 1st mate (sequenced out of the transposon) of each pair was checked. After this filtering step the 1st mates were sorted by barcode and integration site. A putative insertion site had to be covered by a least two independent reads mapping to the same location (twenty reads for non-duplicate filtered data).

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

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

2.6.1 Validation of splicing in the transgenic constructs

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

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

RT-PCR on in vitro samples: 16µL RNA was treated with 2µL DNAseI and 2µL 10x buffer for 15 minutes at room temperature, before adding 2µL stop solution and incubating at 70˚C for 10 minutes. Matched control samples were not treated with DNAseI. 10µL of random hexamers (100µM) were then added to 20µL of RNA, incubated at 65˚C for 10minutes and then placed on ice for 2minutes. The reverse transcription reaction used 15µL of DNAseI treated RNA, 8.4µL water, 8µL 5xMMLV buffer, 4µL DTT, 1.6µLdNTP, 1µL Rnasin (Promega-40U/µL) and 2µL reverse transcriptase(Gibco-GRL). A control reaction, without reverse transcriptase was also performed. RT-PCR samples were incubated at 37˚C for 90minutes, 70˚C for 10 minutes and then stored at -20˚C until PCR.

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

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

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

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

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

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

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

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

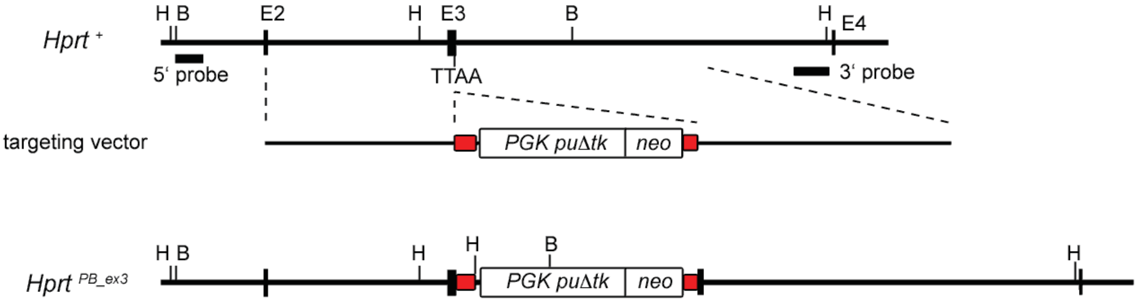


Figure 2.5: *Hprt* locus of ES cells used for the HAT resistance assay: The *PB* transposon, with inverted terminal repeats (red) positioned within exon 3 of the *Hprt* locus. Excision by *PB* transposase restores normal configuration and function of the *Hprt* locus. (Figure and cells courtesy of Kosuke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

2.6.3 Flow Cytometry

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

Flow sorting of mouse bone marrow and spleen cells was performed with the assistance of David Kent (CIMR, Green Lab) or Bee Ling Ng (WTSI) using the following panels; (1) CD45 Pacific blue, CD19 PE, B220 APC Alexa 750, Kit APC, Gr1-CD11b FITC and CD3 PerCPCy5.5. or (2) CD34 FITC, CD19 PE, B220 Alexa Fluor 750, CD3 PerCPCy5.5 and CD11b Alexa Fluor 647. Cells were sorted into i) granulocytes (CD45+, Mac1Gr1+, CD19-), ii) T cells (Mac1Gr1 and CD19 negative, CD3 positive) iii) CD19+,B220+ B cells iv) B220+, CD19- B cells and v) c-kit/CD34 positive progenitor cells, with a target of 100 000 cells in each.

2.6.4 Western Blotting

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

2.6.5 B cell receptor repertoire analysis

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

Sequencing libraries were prepared and sequenced using standard protocols and the MiSeq platform to generate 300bp paired end reads. Filtering and repertoire analysis was performed as follows: "Raw reads were filtered for base quality (median >34) using the QUASR program (http://sourceforge.net/projects/quasr/). MiSeq paired end reads were co-joined at their overlapping region, and non-immunoglobulin sequences were removed, retaining only reads with significant similarity to mouse IgH from the IMGT reference database(Lefranc et al., 2009) using BLAST(Altschul et al., 1990) ($1x10^{-10}$ E-value threshold). Sequences were trimmed to remove primer sequences, and sequences with a minimum length of 180bp were retained. IgH sequence network generation was performed according to Bashford-Rogers et al.(Bashford-Rogers et al., 2013a)".

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

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