Chapter 3

A vector to make homozygous mutations with high genome coverage

3.1 Introduction

A number of screens have been previously conducted in *Blm*-deficient cells—for mutants resistant to 6-thioguanine, aerolysin and retroviral infection. These had several limitations. In all cases, the phenotype screened was selectable. Most loss-of-function phenotypes are not directly selectable, and may in fact be more likely to manifest as hypersensitivity. Thus a method to access these phenotypes would greatly increase the scope of these screens. In order to do this, the isolation of homozygous mutants needs to be uncoupled from the screen itself. A collection of homozygous mutants could be subcloned and arrayed in multiwell plates, and screened clone-by-clone for any phenotype. This would include sensitivity (lethal) phenotypes and also more subtle phenotypes, such as changes in morphology or gene expression.

The second limitation was that in the screens using an insertional mutagen, only a subset of the expected mutants was found. In the mismatch repair screen, only two of the known genes were recovered, although a novel component was also discovered (Guo *et al.*, 2004). In the case of the retroviral resistance screen, only the receptor for the virus was recovered, while other components of the infection pathway might be expected (Wang and Bradley, 2007). Notably, multiple independent mutants were obtained for genes that were identified while other expected genes were not identified at all. This suggests that the retrovirus used for mutagenesis in these cases does not efficiently mutate all loci in the genome. Therefore improvements to the mutagen are necessary to increase coverage. For the aerolysin resistance screen, which recovers mutants in the GPI anchor synthesis pathway, ENU mutagenesis was used and 12/23 known genes involved in GPI anchor synthesis were recovered (Yusa *et al.*, 2004). While this is better than the insertional mutagens, it has the disadvantage that ENU mutants are not easily mappable.

3.1.1 Estimating library coverage

Coverage of previously created libraries has been evaluated by the number of expected mutants recovered in a test screen, e.g. mismatch repair genes. This approach only examines a small number of loci (five known autosomal genes that confer 6-TG resistance when mutated: *Msh2, Msh6, Pms2, Mlh1* and *Dnmt1*), and while other parameters such as the number of independent mutations in these genes can be used to estimate complexity or saturation there is no information about other loci in the genome. It would be useful to know all insertion sites in a library prior to screening to know if any genes known to be involved in the screened phenotype are mutated.

3.1.2 Illumina sequencing technology

The Illumina Genome Analyser method (previously known as Solexa), and related technologies that combine molecular cloning and sequencing without involving a bacterial cloning step have greatly increased sequencing throughput. The Illumina method begins with a random fragmentation of DNA by nebulisation or sonication. Processing these fragments with a mixture of enzymes creates ends with a single $3'$ adenylate overhang. Illumina adaptors bearing a compatible overhang are then ligated to the fragments. A minimal PCR amplification using primers to these adaptors is usually incorporated to increase the amount of DNA available.

The adapted fragments are then denatured and the single strands hybridised to a slide coated in complementary adaptor oligonucleotides. By carefully titrating the amount of adapted fragments that are loaded, a spread of well separated single-stranded DNA molecules can be obtained on the slide. These single molecules are expanded to a cluster by an isothermal PCR reaction, using nearby adaptor oligonucleotides on the slide as primers. Thus all the PCR products are covalently linked to the slide and remain close to each other, forming a spot of identical single-stranded DNA molecules and their reverse complements. This step is analogous to the

bacterial cloning stage when making conventional sequencing libraries, but on a huge scale—a single slide contains eight lanes which can have 10^7 clusters or more each.

Sequencing of the fragments is done in parallel, by monitoring synthesis of the complementary strand. Nucleotide triphosphates are provided with reversible terminators, so only one is added at a time. Each also has a fluorescent dye, so if it is incorporated into the molecules in the cluster, the spot will fluoresce. After each step the slide is photographed to identify the clusters that have incorporated the nucleotide. The dye is then removed prior to the next addition. By analysing all the images, the sequence of each cluster can be built up. Two paired-end reads of over 100 bases each can be obtained at the time of writing, and the read length is constantly being improved.

The Illumina adaptors contain an unpaired region similar to splinkerette adaptors. This can be exploited in the same way as in splinkerette PCR to selectively amplify fragments that contain a known sequence (i.e. a PB transposon repeat). A method to do just this, resulting in PB-genome fragments flanked by Illumina adaptors ready for loading onto an Illumina flow cell, was recently developed (Langridge *et al.* (2009) and D.J. Turner, unpublished). I decided to use this method to sequence a large set of insertion sites for the TNP vector to accurately determine the potential coverage of mutant libraries.

Furthermore, this method could also be used to study changes in mutant populations, as it allows identification of all the insertion sites present in a population of cells[1]. As each insertion site tags a corresponding mutant, and the mutated gene, the number of cells present that belong to a particular mutant clone can be estimated based on the number of reads for each insertion site. An example of how this might be used is to split a library into two duplicates, and treat one with a drug while expanding the other without selection. Comparing the insertion sites in each population could allow identification of sensitive mutants (not present in treated sample) or mutants with increased resistance (relative increase in treated sample). Similar methods have been successfully used with mutant collections in yeast and bacteria (Langridge *et al.*, 2009; Ooi *et al.*, 2001). A secondary aim of these experiments was to see if this approach could work for performing screens in mammalian cells.

[1]This could perhaps be termed the transposome!

3.1.3 Mutagens

Retroviruses have clear insertion 'hot' and 'cold' spots, with higher or lower frequencies of mutation compared to the average across the genome. This is clear from the ES cell gene trap libraries (Hansen *et al.*, 2008). In these libraries, which contain hundreds of thousands of clones, some mutations are represented by thousands of independent insertion events while other genes are not hit at all. Some genes are simply not expressed in ES cells, or not expressed at high enough levels or consistently enough to be trapped, but others may missed due to some property of the chromatin that is unfavourable to retroviral insertion, or expression of the resistance genes contained within the retrovirus. This could be the cause of expected hits being missed in these screens.

The PiggyBac (PB) and Sleeping Beauty (SB) transposons seem to display no such site preference, beyond a four (TTAA) or two (TA) nucleotide acceptor site respectively. The two do differ in their preference for methylated DNA, SB apparently favouring it, but no data on insertion sites so far suggest serious hot spots. In particular, these transposons can access sites that have not been mutable by retroviral gene traps (Wang *et al.*, 2008, 2009). Therefore these transposons were ideal candidates to expand coverage of the libraries while retaining the mapping advantages of using an insertion mutagen.

One advantage of PB in particular is a slight preference for active genes. Almost half of PB insertions in ES cells are in genes expressed in ES cells (Liang *et al.*, 2009). This figure is high enough to not select for mutagenesis using a promoter trap construct. Not selecting for mutagenesis will expand coverage to genes not accessible to trapping. However, it is important that the construct used is designed to be mutagenic in as many genomic locations as possible. In this chapter I will describe the design and synthesis of such a construct.

3.1.4 Isolation of homozygous mutants

Given these recent advances in ES cell mutagenesis, the next step is to develop a method to convert these mutations to homozygosity. As mentioned above, homozygous mutants segregate spontaneously in cultures of *Blm*-deficient cells carrying heterozygous mutations. If each cell begins with a single heterozygous transposon insertion, homozygous mutants can be distinguished by copy number, as these will contain two allelic copies of the transposon (Figure 3.1). I will describe below the design

of a construct that is selectable based on copy number and would therefore be suitable for isolation of the rare homozygous cells.

As I anticipated such a construct being larger than the 3 kbp cargo capacity of SB, I designed the construct with PB in mind as a vector. PB has been shown to still transpose effectively with cargoes of up to 9 kbp (Ding *et al.*, 2005).

3.2 Results

3.2.1 An insertional mutagen for non-selectable mutagenesis

Around half of PB insertions will be in genes, the vast majority of these in introns. The other half may be in important sequence, if PB has a preference for "open" chromatin or transcribed regions. However without further information about the nature of these insertions it is difficult to design a mutagen to specifically disrupt them, beyond simply introducing ectopic sequence. I therefore focused my design on maximising the chances of disrupting transcription for insertions in introns. I designed and constructed the mutagen in collaboration with Amy Meng Li, another graduate student in the lab.

Firstly, the TTAA insertion site of PB is palindromic and the transposon can insert in either orientation. Therefore, the mutagen must be bidirectional in order to disrupt genes in either orientation relative to the insertion. To accomplish this, I chose to use two mutagenic units, one at each end of the transposon. For the mutagenic units themselves, there are several conditions to take into account. The primary consideration is that they should be of small size—i.e. less than one kilobase in length—as although PB has a relatively large cargo capacity, there must also be space for the homozygosity selection cassette (see below).

Splicing can occur over long distances, therefore simply introducing ectopic sequence may not affect splicing unless splice acceptor sequences are present. There is no single consensus splice acceptor sequence, although the two nucleotides immediately 5′ of the spliced exon are always AG. Further upstream of the splice site there is often a polypyrimidine tract, but the length and separation from the splice site vary and a clear polypyrimidine tract is not always present. Computational methods to predict splice acceptor activity have been developed (Barash *et al.*, 2010), but designing a splice acceptor from scratch would be difficult without full knowledge of the factors determining activity, which may also vary by organism and cell type. Therefore the safest method to obtain a working splice acceptor is to simply use sequences from endogenous genes.

Various splice acceptors have been used for gene traps, popular ones include those from SV-40, adenovirus, and the mouse *En-2* gene (Gossler *et al.*, 1989). As these are generally linked to a selectable marker for gene trap mutagenesis, the real efficiency with which they disrupt splicing of the wild type pre-mRNA is not known. The reason for their use is convenience, and in many cases based on what had been cloned at the time. Although they clearly work in many genomic locations (Skarnes *et al.*, 1992; Neilan and Barsh, 1999), some exceptions have been reported (Voss *et al.*, 1998; Shawlot *et al.*, 1998). As my aim was to make a construct that is mutagenic in as many genomic contexts as possible, I took this opportunity to logically select endogenous mouse splice acceptors with desirable properties for mutagenesis. I decided on the following set of parameters to search the reference genome sequence for potential mutagens.

Most previously used gene trap mutagens consist of a single exon reporter gene with associated splice acceptor. Although an insertion bias for the 5′ end of genes has been reported, many insertions do occur further along the gene. These may not be mutagenic if critical domains or sites in the protein are upstream of the truncation caused by the gene trap. It is even possible that a dominant mutation could be caused if the encoded protein has a C-terminal regulatory domain that is deleted by the truncation. An ideal mutagen would cause null mutations when inserted at any point in the coding sequence.

By exploiting the nonsense-mediated transcript decay (NMD) pathway, this may be possible. NMD is a surveillance pathway that guards against production of aberrant transcripts, and may also have a regulatory role. The pathway is activated by transcripts with an in-frame STOP codon at any position more than around 50 nt 5′ of the final intron-exon junction. Introduction of a premature termination codon (PTC) 5′ of this boundary is sufficient to direct a transcript for NMD, as is introduction of an extra intron downstream of the real termination codon (Zhang *et al.*, 1998; Carter *et al.*, 1996). Transcripts with PTCs are detected by a translation-dependent process involving the exon junction protein complex and mammalian homologues of the yeast up-frameshift proteins (UPFs, Leeds *et al.* (1991); Maquat (2004)). Due to the requirement for an exon junction complex downstream of the PTC, a mutagen designed to make use of NMD requires two exons, with the penultimate exon being at least

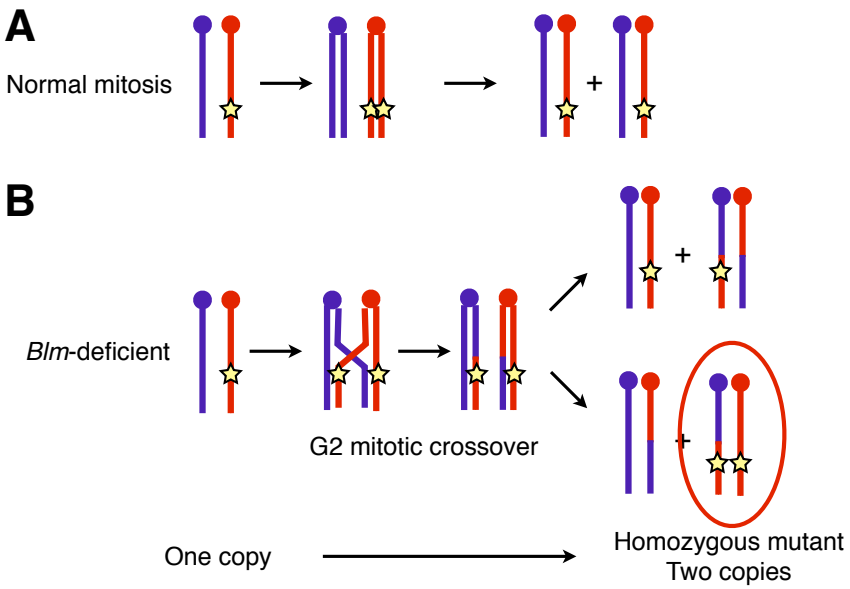

Figure 3.1: Copy number gain during loss of heterozygosity. Possible daughter cells arising from a single copy heterozygous mutant during: A—Normal mitosis, B—Mitosis with recombination and crossover in G2 phase.

50 bp in length. Therefore, I began by searching for pairs of terminal exons (i.e. the two most 3′ exons of a given gene), with a total size of less than 3 kbp.

To ensure splicing was not regulated, I also stipulated that genes from which the exons were selected had only a single annotated transcript, implying constitutive splicing. To ensure mutagenicity in all reading frames, whether by truncation or NMD, I specified that the exon pairs should have out of frame STOP codons in both non-native reading frames, and ranked the pairs by the number they contained. As an extra precaution, I considered the possibility that splicing may occur preferentially at one splice acceptor, or that splicing might continue downstream after splicing one or both exons. To guard against this, I only considered exons that begin and end in different phases, and would therefore be likely to cause a frameshift if incorporated into a longer transcript rather than at the end.

Finally, as production of a fusion protein with the endogenous gene product of these exons could have a dominant effect, I checked for the presence of annotated Pfam domains encoded by the exons and picked only exons that lack such domains. Additionally, I checked that the gene from which the exons are derived is not expressed in ES cells, as judged by lack of a gene trap clone (although see discussion of gene traps, above). This may decrease the chance that expression of part of the gene could affect normal ES cell physiology.

I incorporated these criteria into a script to search the Ensembl database (Flicek *et al.* (2010), version 43 based on the NCBI m36 mouse assembly) for candidate exon pairs (Figure 3.2). These candidates were ranked by size and number of premature STOP codons and exon pairs from *Ccdc107* and *Dom3z* chosen as the best candidates. I amplified these exon pairs from BAC templates by PCR using the proof-reading enzyme KOD and ligated them to pML5, a plasmid containing PB repeats flanking a PGK-*neo* gene (Figure 3.3A,B). I then transferred the *Ccdc107* exons (*Ngo*MIV–*Eco*RI fragment) to the *Dom3z* plasmid (*Age*I–*Eco*RI digest; *Age*I and *Ngo*MIV leave compatible ends) in the opposite orientation (Figure 3.3B,C). I then deleted the *neo* gene by excising it as an *Eco*RV–*Sfo*I fragment and religating the plasmid (Figure 3.3C,D).

To further increase the mutagenic potential of this construct, I used site-directed mutagenesis (Stratagene QuikChange) to introduce additional premature stop codons in the native reading frame of the penultimate exon. The primers incorporated additional nucleotide changes to introduce restriction sites to screen for plasmids with the changes. I carried out mutagenesis at both sites in parallel and identified several plasmids with both changes. I inserted a short oligonucleotide linker into the multiple cloning site flanking one of the PB repeats, in-

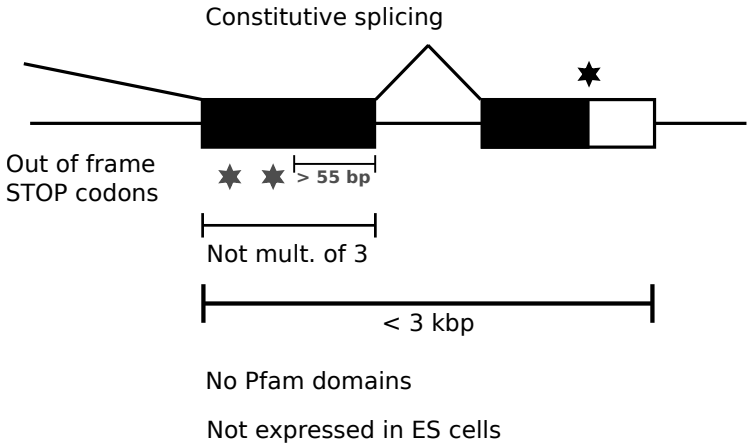

Figure 3.2: Features considered in design of the mutagen. Dark boxes—natively translated exons, empty boxes—untranslated exons. Asterisks represent stop codons.

troducing an extra *Pci*I site required for subsequent subcloning (see below), forming pSDM-Pci.

The function of this construct was tested by Amy Li (Li, 2010). Briefly, splicing occurred at both mutagens. However, in the case of the *Dom3z* end of the transposon, some splicing occurred at a cryptic splice acceptor site within the PB repeat. Therefore the construct functions to disrupt splicing *in vivo*. Further evidence for the function of this construct as a mutagen is provided in Chapter 5.

3.2.2 Dual selection cassette for copy number based selection

Strategies for selection based on copy number

A simple way to select for cells with different copy numbers of a gene would be simply based on the amount of gene product present. However, to discriminate one copy from two copies this is unlikely to be sensitive enough. The amount of protein product may be buffered to some extent by mRNA stability and translation efficiency, and is also likely to vary from cell to cell such that the distributions of protein amounts in cells with one and two copies overlap. Also, as transgenes are typically expressed at very high level, the activity of the transgene may be close to maximal even with one copy, unless careful thought is given to the promoter used, message stability etc.

Nevertheless, such dosage-sensitive selection has been used in ES cells in the past. The key requirement is a hypomorphic mutant *neo* gene, often referred to as *neo**, and in fact contained in many common vectors. The wild type gene is too active to discriminate selectively enough based on dosage, as are most other common selectable markers. Using a *very* high concentration of G418 (in the mg/ml range, corresponding to the order of 1 mM), rare cells with two alleles can be isolated (Mortensen *et al.*, 1992). However, there is typically a high background in the selection and a screening step on a scale similar to gene targeting (10–100 clones) is required. Although this is feasible for a single locus, the technique is not generally applicable on a genome-wide scale for this reason.

The alternative strategy that I decided to investigate is to select based on expression of two simultaneous selectable markers. This would be an improvement over the high-G418 scheme above, as it does not involve selecting for discrete variants (i.e. cells with one or two copies) from a continuous distribution of protein levels. However, as only a single construct can be used to make the initial mutation, this needs to be carefully designed. My selection scheme is based on a construct that can only express one of two encoded selectable markers at a time. If the construct can switch between expressing one or the other, only a cell with two (or more) copies of the construct will be able to express both simultaneously, and therefore grow in the presence of both corresponding drugs.

The design for the construct is shown in Figure 3.4A. The coding sequences of *neo* and *puro*Δ*TK* are placed in opposite orientations downstream of a PGK promoter. The two genes are flanked by inverted loxP sites, such that inversion of the intervening sequence by Cre recombinase will change which selectable marker is under the control of the promoter. This is reversible, so a cell with two copies

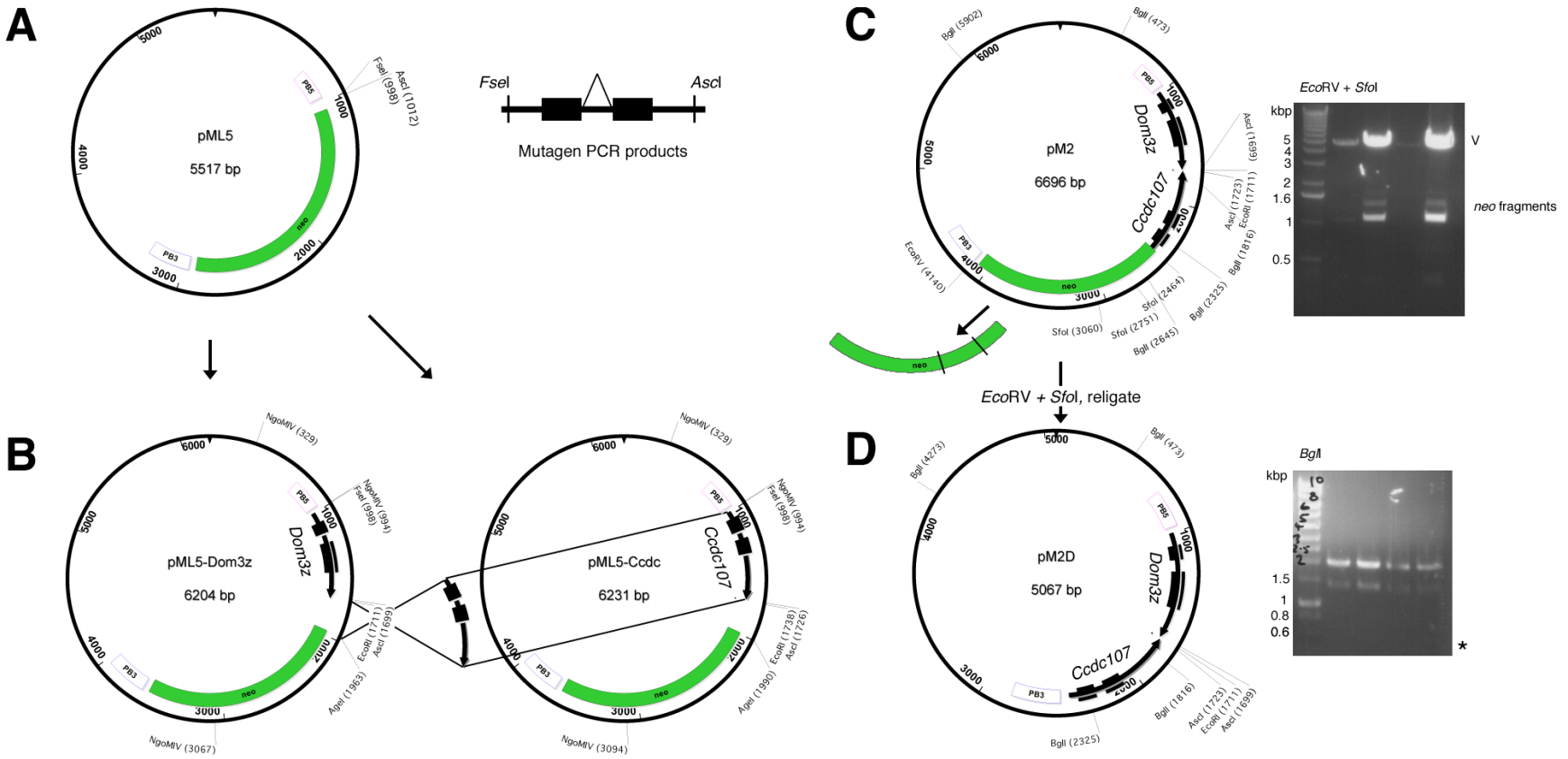

Figure 3.3: Cloning scheme for PiggyBac mutagenesis vector. A—Exon pairs amplified by PCR are cloned into pML5. B—*Ccdc107* is excised and transferred to the *Dom3z* vector in the opposite orientation. C—The resulting plasmid is digested with the enzymes indicated, the V (vector) band purified and religated to give D—Double mutagen construct lacking *neo*. Diagnostic *BglI* digest shown.

of the construct will have a 50% chance of becoming double resistant if the Cre reaction is efficient and unbiased.

Cloning of the inverter construct

The inverter construct was derived from pYTC85, a plasmid containing the *bsd* and *puro* genes in tandem. In this construct, both selectable markers have the same polyadenylation (pA) signal, derived from the bovine growth hormone gene (bpA). I switched one of these to a different pA sequence to avoid secondary structure when these pA signals become juxtaposed as inverted repeats in the inverter plasmid. This will also prevent unwanted recombination between bpA sequences during recombineering reactions to construct targeting vectors (see below). I also replaced the *bsd* with a *neo* gene, as feeder fibroblasts that are resistant to both *bsd* and *puro* were not available at the time. To do this, I replaced the entire *neo*-bpA with a *neo*-SV40pA amplified from pcDNA3 (Invitrogen). The PCR primers used contained a compatible *Sfi*I and *Asc*I restriction sites. This cloning step also introduced a *Kpn*I site after the SV40pA sequence (Figure 3.5A, B, E).

As pYTC85 is a targeting vector and therefore a large plasmid, I cloned the selection cassette in pUC19 as an *Hind*III–*Eco*RI fragment to ease handling (Figure 3.5C,F). The extra *Kpn*I site previously introduced was then used to flip the loxP-*neo*-SV40pA segment by *Kpn*I digestion and religation, forming the inverter construct (Figure 3.5D,G). I confirmed function of the loxP sites by treatment with recombinant Cre *in vitro* and preparing plasmids from bacteria transformed with the products of the reaction (Figure 3.4B).

The inverter selection cassette was excised as an *Hind*III–*Eco*RI fragment and cloned into the *Asc*I site of pSDM-Pci in a blunt ended ligation (using the Klenow fragment of DNA polymerase I to form blunt ends). This formed the TNN plasmid (as used in experiments in Chapters 4 and 5). I took care to choose the orientation in which the PGK promoter was adjacent to the PB end for which promoter activity has been reported (Cadiñanos and Bradley, 2007). This ensures that the *puro* resistance gene cannot be expressed without inversion to bring it under the control of PGK.

As I planned to select for mobilisation of the transposon, I cloned it as a *Pci*I fragment into the *Xba*I site (both blunted with Klenow) of a human *HPRT* minigene driven by the long RNA polymerase II promoter. Sequencing the construct revealed a four base pair deletion in one of the loxP sites (Figure 3.7A). Surprisingly this did not seem to abolish recombination *in vitro* or *in vivo*, and in fact the mutation was also present in the lab stock of the original pYTC85 plasmid. I decided to fix the mutation, as if there is a decrease in recombination efficiency *in vivo* the efficiency of copy number selection will also be reduced. I designed PCR primers to amplify an EM7-*bsd* (blasticidin-S deaminase) gene, flanked by wild type loxP sites and 50 bp homology arms targeting sequence either side of the mutant loxP. Co-transformation of recombination-competent EL350 bacteria with this construct and the P2-HPRT-Tn plasmid resulted in the mutated loxP site being replaced with the *bsd* gene flanked by wild-type loxP sites (Figure 3.7B). Correct recombinants were selected on low salt LB-blasticidin agar plates. The *bsd* gene was then removed by inducing Cre expression using arabinose induction in EL350 cells (Lee *et al.*, 2000), leaving a functional loxP site.

Function of the inverter construct in ES cells

I tested the function of the transposon, resistance genes and loxP sites in ES cells. I used the NRB2 ES cell line, which is *Blm*-deficient (derived from the NN5 cell line) and carries a 4-hydroxytamoxifen (4-OHT) inducible Cre gene (targeted by me using the vector and procedure in Vooijs *et al.* (2001)). I expanded duplicate cultures and treated one with 4-OHT 24 hours prior to electroporation. Electroporation with TNN plasmid, with and without a PB transposase (PBase) expression plasmid confirmed that most resulting G418-resistant colonies were PBase-dependent, indicating the transposon is functional, and most puro-resistant colonies were 4-OHT dependent, indicating the loxP sites are functional (Figure 3.8). PBase independent G418-resistant colonies are likely to result from random integration of the plasmid into the genome. All puro-resistant colonies are sensitive to FIAU, indicating that the ΔTK is also functional.

Background puro-resistant colonies are likely to be due to leaky activation of the ERT2-Cre fusion, possibly by steroid hormones in the foetal calf serum used in the culture medium. Testing the construct in cells without ERT2-Cre confirmed that the background puro resistance is due to the presence of the ERT2-Cre (Figure 3.9). I also selected the transfected cells in G418 and puromycin (without 4-OHT treatment), which confirmed that the puromycin resistant cells in this case were not resistant to G418. Therefore, even a low level of leakiness in the Cre transgene will not result in a background of double-resistant cells that only contain one copy of the

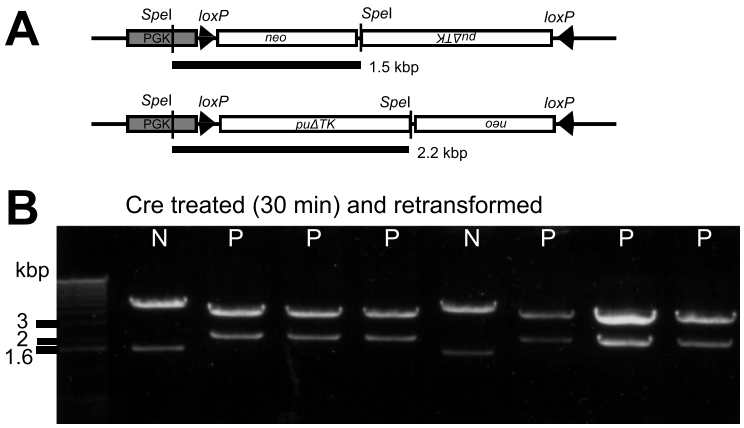

Figure 3.4: loxP sites are functional in the inverter construct. A—Map showing digest used. B—The inverter construct (Figure 3.5D) was treated with recombinant Cre and transformed into bacteria. Plasmids were digested with *Spe*I. A mixture of both possible orientations is seen, consistent with reversible recombination between the loxP sites.

transposon.

Selection conditions for G418 and puromycin selection

Puromycin and G418 both act by inhibiting protein synthesis. The mechanism of action for the aminonuceloside puromycin is well-defined: it is incorporated into the nascent peptide and acts as a chain terminator (Nathans, 1964). It contains a nucleoside moiety that can mimic an aminoacyl tRNA and cause formation of a peptide bond with the nascent chain, a property that has been instrumental in studies of the ribosome. Puromycin kills eukaryotic cells quickly, within a few days (Adams and van der Weyden, 2008).

G418 is structurally distinct from puromycin, being an aminoglycoside similar to the antibiotic neomycin. Although it also binds to the ribosome it does not bind either of the active sites as a direct mimic of an aminoacyl tRNA, but instead binds to ribosomal RNA, interfering with the decoding site and affects ribosome recycling (Borovinskaya *et al.*, 2007). G418 kills cells slowly, and cells can continue to grow and divide before widespread death begins.

Resistance to puromycin and G418 is not mediated by ribosomal variants, but by expression of enzymes derived from fungi or bacteria that inactivate the drugs. *puro* encodes puromycin N-acetyltransferase, which N-acetylates the amino group that would otherwise form a bond with the nascent peptide chain. *neo* encodes neomycin phosphotransferase II, which is also active against G418 and inactivates it by phosphorylation of the 3′ glycosidic hydroxyl groups.

As the pathways for resistance are independent and the drugs structurally distinct, it is unlikely that cross-resistance will occur. There are several reports to this effect in the literature of double targeting using G418 and puromycin, but generally the genotyping used does not distinguish between random integrations of the targeting vector and possible background resistance. To ensure that standard selection conditions could independently select for the two resistance genes I used different concentrations of G418 to kill *puro*-expressing cells, with and without puromycin in the growth medium. I also carried out the reciprocal experiments killing *neo*-expressing cells with puromycin.

Interestingly, in the case of *puro*-expressing cells killed with G418, the addition of puromycin to the medium did appear to shift the kill curve to the right, indicating decreased sensitivity to G418 (Figure 3.10). Killing was still complete in all but one replicate, which had a single surviving colony, at the standard 180 μg/ml concentration. The difference is small and cannot be evaluated as significant with the numbers used. A possible explanation could be a slowing in the growth rate in the presence of puromycin with a corresponding increase in G418 resistance, as G418 is most effective against actively dividing cells.

3.2.3 Genome coverage and insertion preferences of the TNP vector

In this section, I describe an attempt to assess coverage, i.e. the number and distribution of loca-

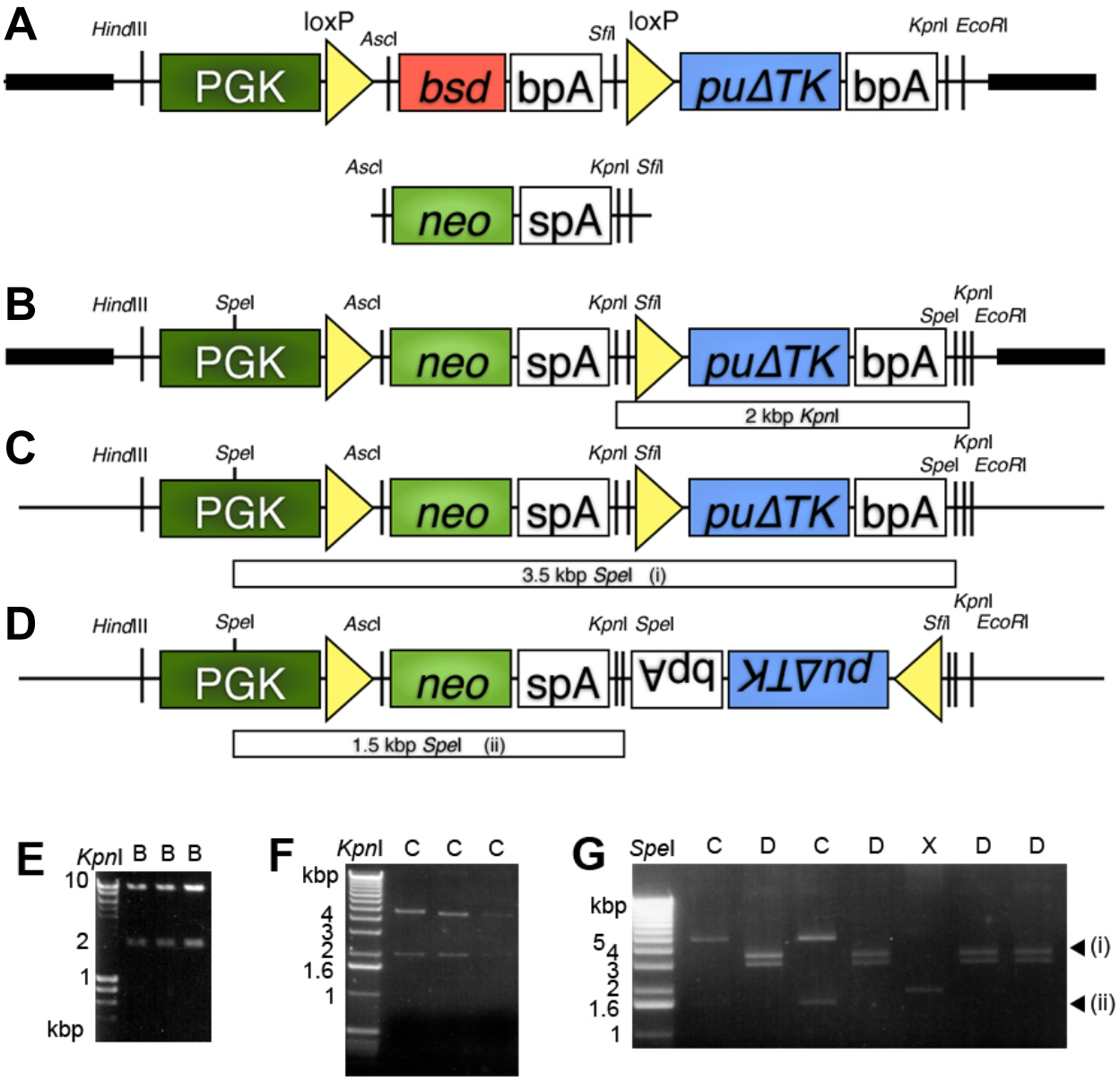

Figure 3.5: Cloning the inverter construct. A—pYTC85 and the PCR-amplified *neo*-SV40pA (spA). B—Result of replacement of *bsd*. C—Selection cassette moved into pUC19 backbone. D—Result of *Kpn*I digest and religation to give inverter construct. Lower panel, restriction digests using indicated enzyme of: E—several clones from *bsd* replacement(B); F—several clones in pUC19 backbone (C); G—*Kpn*I digested and religated (C), giving a mixture of C and D when subcloned. D is the inverter construct.

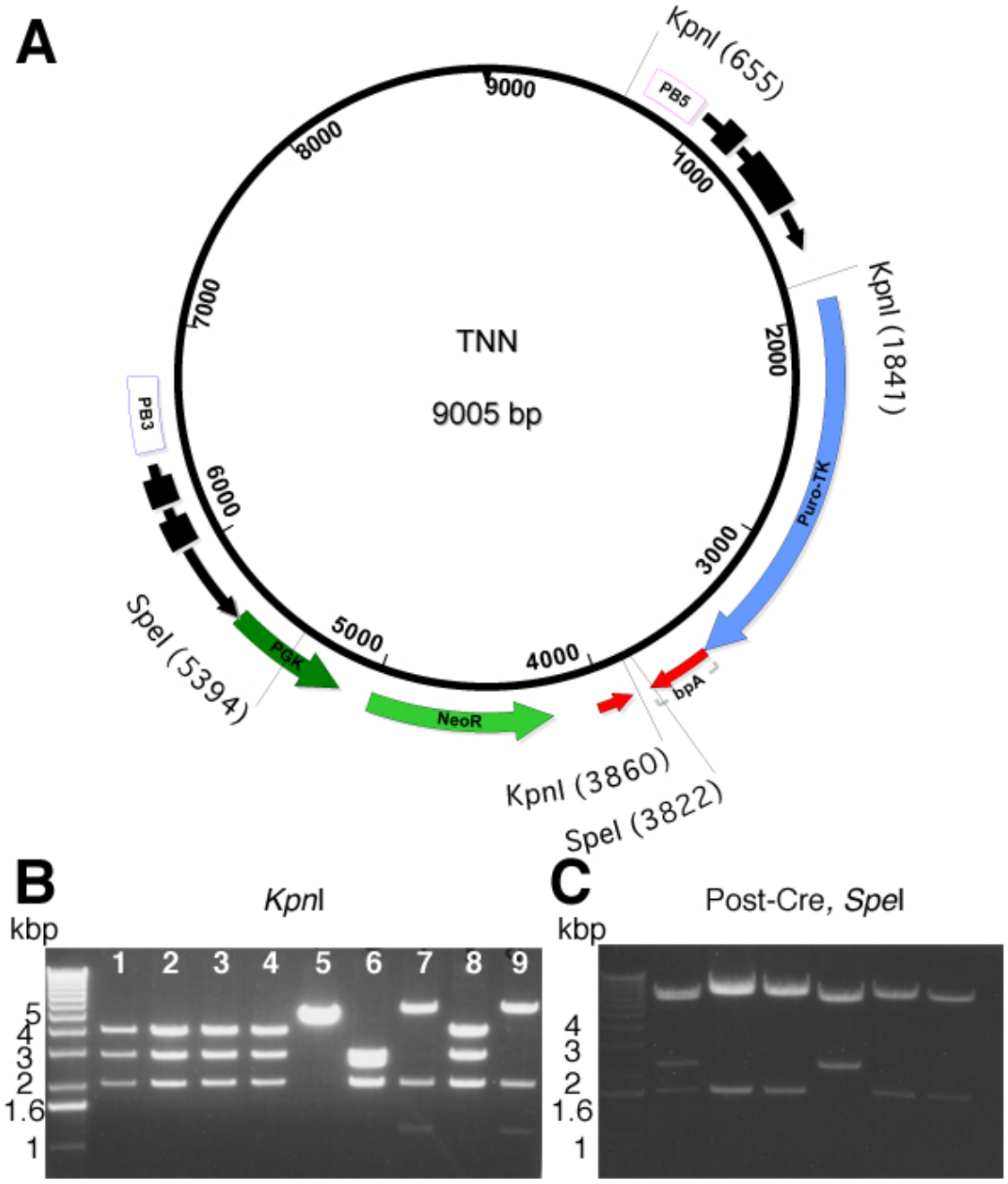

Figure 3.6: Cloning of the TNN plasmid. A—map of the plasmid. B—Results of ligation of the inverter construct into *Asc*I-digested and blunted pSDM-Pci. Lanes 7 and 9 are the desired orientation (shown in A), with the PGK promoter aligned with the promoter activity end of the PB transposon. Lanes 1–4 and 8 are the other orientation, 5 is religated pSDM-PCi. C—Plasmid from lane 7 (the TNN clone used for all other experiments) was treated with Cre and analysed as for Figure 3.4. Lane 1 contains a mixture of both products. The 2.2 kbp band arises from the *puro*-expressing version.

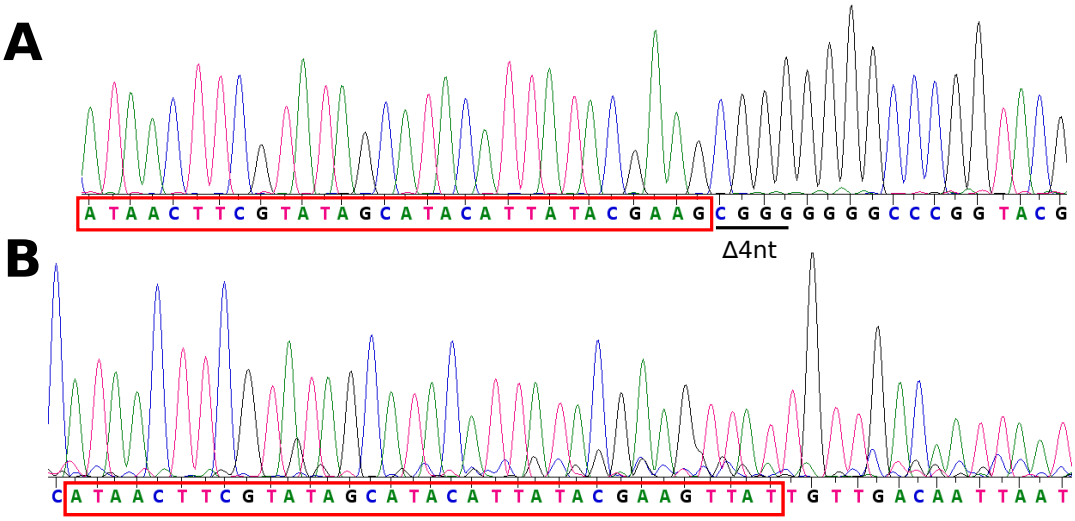

Figure 3.7: Fixing a mutation in a loxP site. A—A four base pair deletion present in the original plasmid. B—Sequence after replacement with the targeting construct. The downstream sequence is different as the *bsd* gene has not been removed in this plasmid.

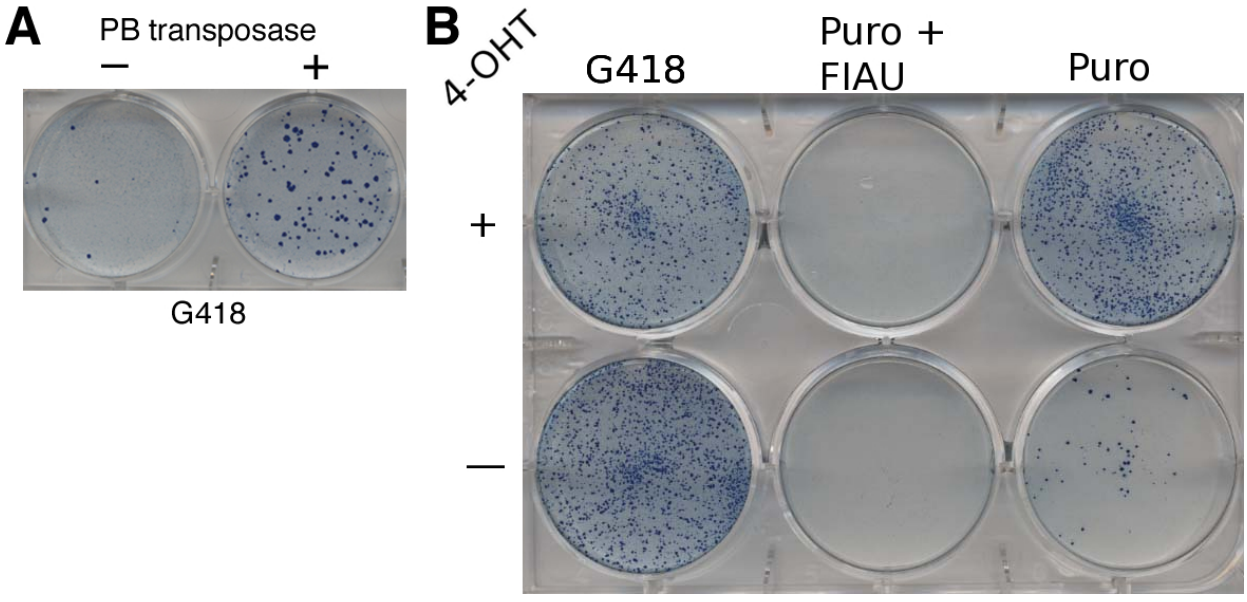

Figure 3.8: Function of the transposon construct in ES cells. A—ES cells were transfected with the TNN transposon construct with (+) or without (–) a transposase expression plasmid and selected in G418. B—NRB2 cells transfected with both plasmids were plated in the indicated drugs, and treated with 4-OHT as indicated.

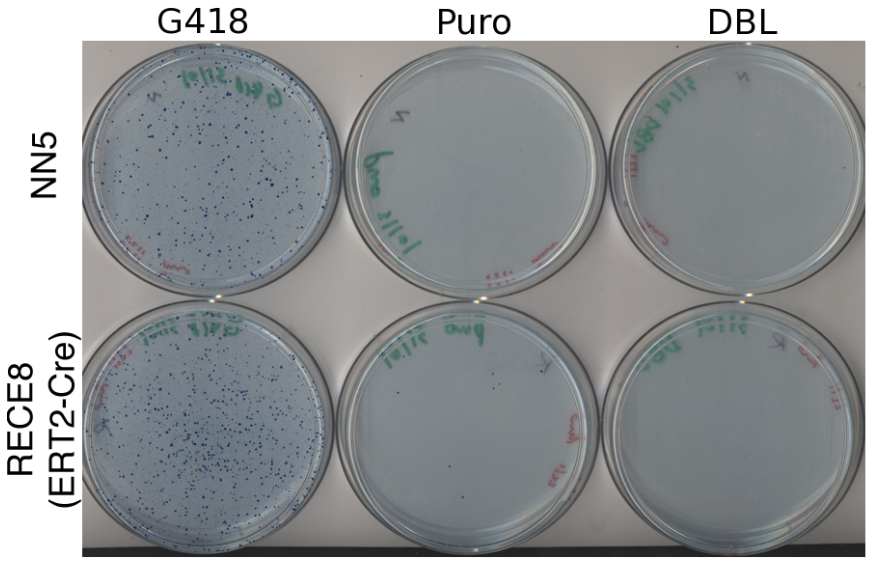

Figure 3.9: Background resistance from leaky ERT2-Cre activity. Top row: NN5 cells (no ERT2-Cre); Bottom row: NN5 cells targeted with *Rosa26::ERT2-Cre* (RECE8 cells). Cells were transiently transfected with TNN and transposase and selected in the indicated drug(s). NB—neither was treated with 4-OHT.

tions with transposon insertions, in libraries of cells mutagenised with TNN/TNP (TNN refers to the *neo*-expressing orientation, and TNP to the *puro*-expressing orientation). The best way to determine coverage is to map all of the transposon insertion sites in the library. I investigated the use of Illumina sequencing for this purpose. Coverage will also depend on how many of the mutated sites can be successfully converted to homozygosity; this is addressed in the next chapter.

ES cell transposon libraries for Illumina sequencing

I first generated a large library of heterozygous transposon insertions in *Blm*-deficient cells, by electroporation of 100 ng TNP plasmid and 15 μg mPBase transposase plasmid. These conditions result in a modal copy number of one per cell (Wang *et al.*, 2009). Therefore, most insertion sites in the pool will have been directly selected for, which mirrors the intended use of the TNP vector. Cells with insertions were selected in puromycin, and then pooled and passaged together. This formed a library of thousands of insertions to assess transposon coverage.

I used two DNA repair deficient cell lines to investigate whether changes in the abundance of certain mutants could be detected against a background of a large number of other mutants. The *Xrcc4*$^{-/-}$ and *Xlf*$^{\Delta/\Delta}$ cell lines are hypersensitive to agents that cause DNA double strand breaks (DSBs), such as ionising radiation (Zha *et al.*, 2007) and bleomycin (Figure 3.11). I transfected these cells as above and picked six independent puromycin-resistant subclones for each, which I then mixed and expanded as two pools, one for each mutant cell line. These should contain around six insertions each. I mapped these insertion sites by conventional splinkerette PCR. These known insertion sites act as a tag to follow the mutant clones, although in this experiment the insertion does not cause the mutation itself.

To create a test library I mixed the wild type library with the pooled *Xlf* and *Xrcc4* mutants in a ratio of 500:1. I expanded these together for two passages before splitting the mixed library into four duplicate plates (Figure 3.12). Two of these were treated with a chronic dose of bleomycin (400 ng/ml for three days), while the other two were expanded in normal ES cell medium. The cells were then lysed in 5 ml ES cell lysis buffer and DNA prepared by isopropanol precipitation.

To ensure that the selection worked, I designed PCR genotyping primers for one insertion site in the *Xrcc4* mutants. These primers amplified a product

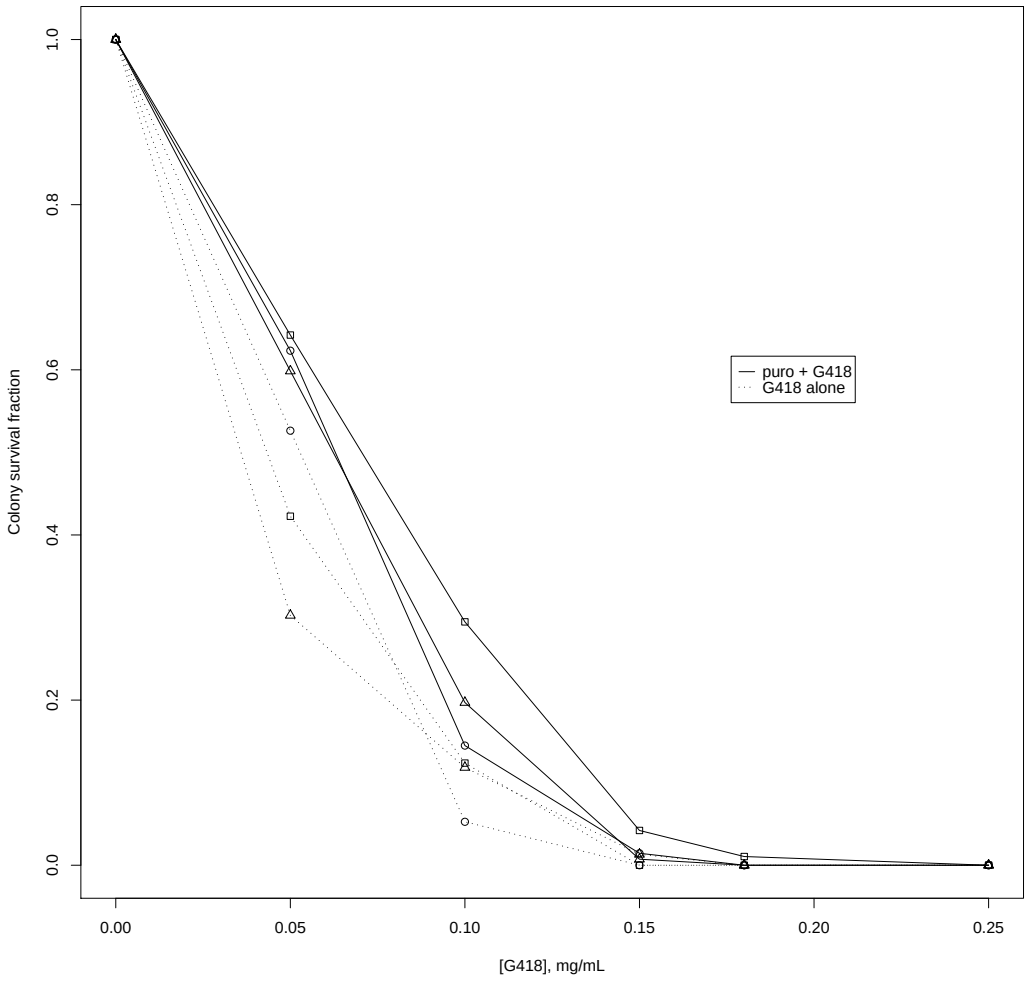

Figure 3.10: G418 kill curve for *puro*-expressing cells. Colony survival, as a fraction of unselected cells or with puromycin only as appropriate. Dashed lines: G418 in medium at indicated concentration; Solid lines: With 3 μg/ml puromycin. Three replicates are shown for each condition.

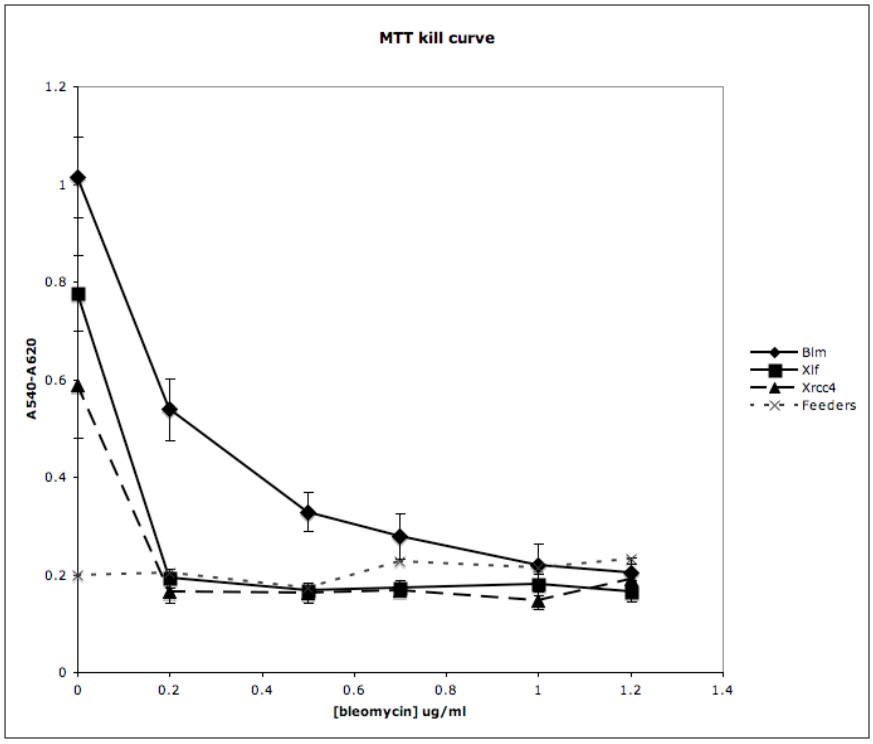

Figure 3.11: Sensitivity of *Xrcc4* and *Xlf* mutant cells to bleomycin. Results of the MTT test, reflecting electron transport chain activity and thus live cells, are plotted on the Y-axis. The indicated cell lines were treated with bleomycin for three days and allowed to recover for three further days before measurement. Error bars: standard deviation; $n = 5$ in each case

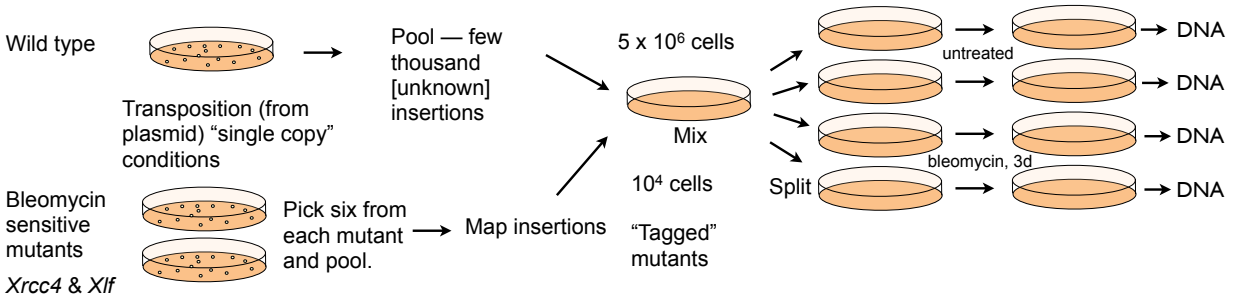

Figure 3.12: Setup of pilot experiment for Illumina sequencing and dropout screens

from the unselected libraries but not the bleomycin-treated libraries, showing that these mutant cells were no longer present (Figure 3.13). This gave me confidence that the same result would be seen in the results of the sequencing experiment.

Preparation and sequencing of transposon-genome fragments

I used the Covaris sonication system to randomly fragment 10 μg of DNA from each library. The fragmented DNA was purified using a QiaQuick column (Qiagen) and analysed on an Agilent Bioanalyzer electrophoresis chip to examine the distribution of fragment sizes. The fragments were distributed with a peak at around 190 bp. I used the Illumina library generation kit and protocol to repair the ends of these fragments add 3′-dA overhangs and ligate standard Illumina adaptors. From these adapted fragment libraries I used a nested PCR protocol to enrich for fragments containing the PB5 end of the transposon. This protocol is similar to splinkerette PCR and uses primers in the second PCR that have the Illumina adaptor sequence at the 5′ end (designed by D.J. Turner, unpublished data), such that the resulting fragments are ready to load onto the Genome Analyser flow cell.

After the second PCR step I separated fragments on a 2% agarose gel and isolated fragments in the 250–350 bp range. The DNA was recovered using a Qiagen gel purification kit, but without heating the sample above room temperature to maintain representation of AT-rich fragments as previously described (Quail *et al.*, 2008).

Prior to loading the flow cell, I used quantitative PCR (qPCR) to determine the concentration of adapted fragments relative to known standards. This is important to obtain the correct density of clusters on the flow cell (Quail *et al.*, 2008). The four samples were loaded in four separate lanes and clusters generated using the Illumina protocol. The flow cell was then sequenced using a PB-specific sequencing primer (read 1) and the standard paired-end adaptor primer (read 2). Seventy-two bases were read at each end. Use of the PB-specific sequencing primer provides further specificity for transposon-genome junction fragments, as clusters that do not contain the PB repeat despite the PCR enrichment step will not yield any sequence data.

Mapping insertion sites from Illumina sequencing data

The two untreated and one of the bleomycin treated libraries produced around four million reads each (Table 3.1). The remaining treated library did not yield enough material for sequencing after the Illumina preparation protocol as assessed by qPCR, but was sequenced with the flow cell below capacity and yielded 438,148 reads. More clusters were present on the flow cell, which has a capacity of around 14 million per lane, but only reads that could be sequenced with the PB primer, and their associated adaptor ends, were included in the results.

The first step of the analysis is to remove PCR duplicates. The PCR steps in the library preparation can result in amplification of certain fragments, so it is important to distinguish whether different fragments that map to the same locus arise from multiple cells with an insertion at that locus, or amplification during the PCR stages. As the initial fragmentation is random, each molecule of DNA is likely to have a different breakpoint. This means that although read 1 (from the transposon end) will map to the same position, read 2 should be different for each molecule of DNA present initially (Figure 3.14). For a given read 1, clusters which have the same read 2 as another cluster can therefore be assumed to have arisen from PCR amplification. From a clone of (say) 100 cells with the same heterozygous insertion, the expected result would be 100 hits at the insertion site for read 1, and 100 hits in the reverse direction distributed 200-300 bp away for read 2 (Figure 3.14). Removing read pairs with identical read 2s showed that 50–60% of sequenced fragments were PCR duplicates (Table 3.1).

Although a nested PCR step was used with transposon specific primers, it is still possible to end up with fragments that do not contain the transposon, as with splinkerette PCR. The sequencing primer terminates six nucleotides from the end of the transposon. As a further control, I examined the start of read 1 for the terminal 'GGTTAA' corresponding to the distal six nucleotides of PB, allowing some mismatch. Eighty-five percent of reads contained this sequence and were retained for mapping (Table 3.1).

I mapped the processed reads using SSAHA2 (Ning *et al.*, 2001) using options appropriate for solexa reads with paired ends within 500 bp of each other (`-rtype solexa -score 20 -kmer 13 -skip 2 -pair 2,500`). To collect reads that map to the same site, I wrote a simple program to read the mapping file and group mappings by start site. This

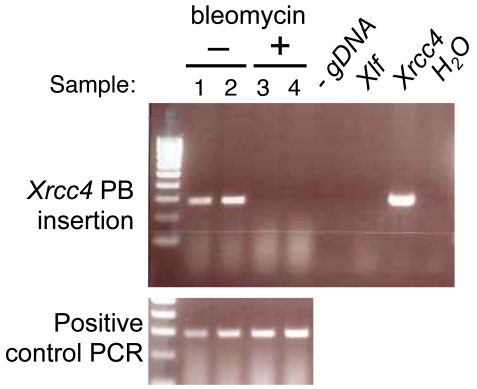

Figure 3.13: Detection of loss of a tagged *Xrcc4* mutant by PCR. Lanes 1–4: DNA prepared from the four pools of treated (+) or untreated (–) mutants; 5: Negative control DNA; 6: *Xlf* mutant pool; 7: *Xrcc4* mutant pool; 8: No template.

Sample	Read pairs	No PCR dups[a]	With GGTTAA	Mapped
1	4,276,924	2,133,392	1,865,005	1,472,646
2	4,879,456	2,124,150	1,888,844	1,514,453
3	438,148	186,183	159,371	132,239
4	3,921,779	1,714,696	1,507,346	1,198,752
	13,516,307	6,158,421	3,531,723	4,318,090

Table 3.1: Number of read pairs remaining after each stage of filtering. Samples 1 and 2 were untreated, 3 and 4 treated with bleomycin as described in the text. [a]Number of reads remaining after removal of PCR duplicates as described in text.

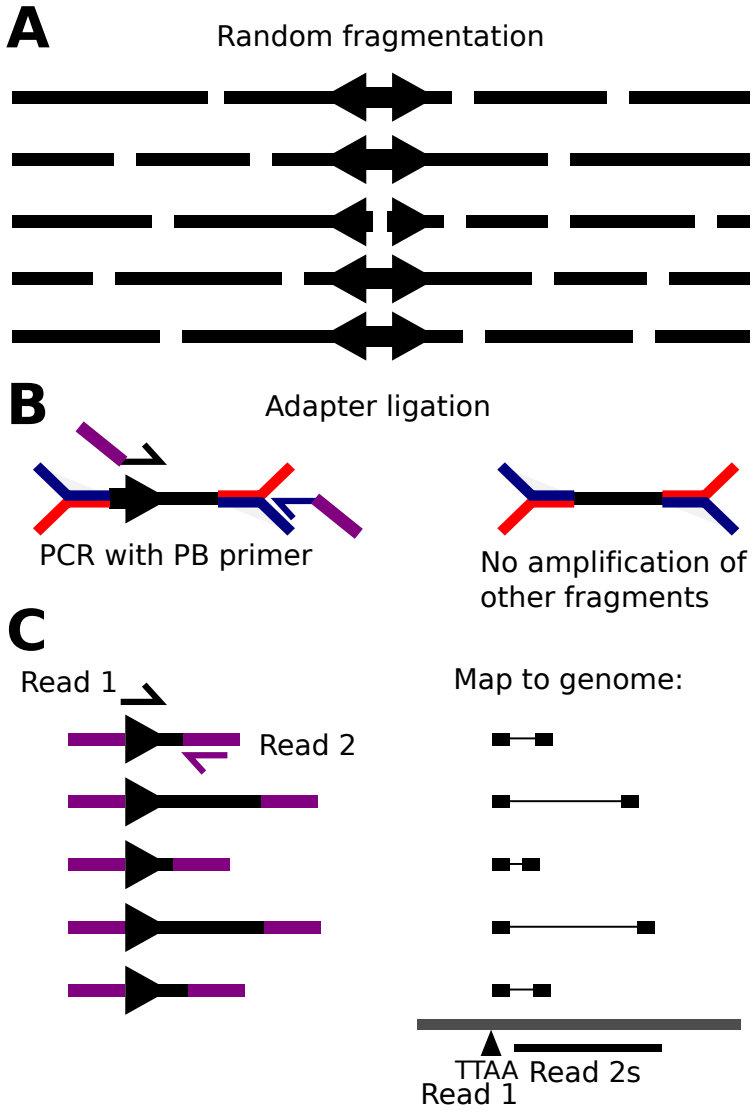

Figure 3.14: Distribution of paired ends in Illumina sequencing of PB insertions. Five molecules of DNA containing identical PB insertion sites are shown. A—Random fragmentation produces different breakpoints in each original molecule. B—Adapter ligation and PCR selects only fragment with the transposon. C—Sequencing and mapping produces a constant result for read 1, but the mapped position of read 2 is unique for each molecule present in the initial fragmentation. PCR duplicates would have identical read 2 mappings.

is much more efficient in terms of computing time and memory than the equivalent program considering the whole length of reads (`ssaha-pileup`), and is appropriate in this specific case because the only information required is the chromosome and position of the insertion site.

Reproducibility of the method

I identified a total of 16,515 insertion sites across all four samples. However, many of these sites were only present in one sample (Table 3.2, Figure 3.15). Moreover, within a sample, many insertion sites were represented by a single, or very few reads. In most cases, the sites that were seen only once also had low read coverage. Judging from the successful extension of the sequencing primer and presence of the transposon sequence in the read, these do represent genuine transposon insertion sites and are not an artefact of the library preparation process.

It is possible that the library is not adequately sampled, although the 10 μg of DNA used for library preparations is equivalent to around 3 $\times 10^6$ cells. The total number of insertion sites is higher than expected, at 16,515. Although I did not explicitly count the number of clones obtained after mobilisation, this is generally of the order of a few thousand under these conditions. Another possible explanation is that the transposase plasmid has stably integrated in a small fraction of the cells. This is almost certain to have occurred given the number of cells transfected and the amount of plasmid used. These cells will express the transposase enzyme constitutively, and therefore could continue to mobilise the transposon during expansion. As the library was split into four pools, new transposition events that occur after the split (when cells were not under puromycin selection), will appear in only one pool. The resulting clones would be of much smaller size compared to the initial set of transposon insertions, which already had thousands of cells per clone when the library was split, and therefore these *de novo* events would likely be poorly sampled, presumably corresponding to a lower coverage in terms of reads.

Many sites that appear in only one sample also have low coverage

As would be expected given this situation, pairwise agreement between libraries was very poor, only 56% on average (Table 3.3). Interestingly, although lane 7 (bleomycin treated) did not produce many reads, I could still identify 3,983 insertion sites in the library. Furthermore, most of these (82–85%) were present in the other libraries. Making the assumption that the set of insertion sites with very low coverage is an artefact of transposase integration, I applied a minimum coverage filter to the data to see if agreement between samples improved. As lane 7 had fewer reads than the others, I defined the cut-off as a fraction of the total number of reads for a lane.

Even a relatively generous requirement for inclusion of 1/100,000 of the total reads (effectively at least two reads for sample 3, and 11–15 for the others) resulted in an increase in pairwise agreement between samples increasing to 85% and above (Table 3.3). However, this is still too low to see 'drop-outs' in the treated libraries, as far too many will occur simply by chance. Looking at the mapped insertion sites for the bleomycin sensitive mutants, it can be seen that they are not present in the treated library, but the agreement between libraries is still not sufficient to determine which insertions are from the bleomycin sensitive cells without prior information (Table 3.4). Additionally, some of the known insertions in the bleomycin-sensitive mutants were only seen in one of the untreated samples. In two cases, one of the known insertions was detected in the (bleomycin-treated) sample 3, with very few reads. This may be real, due to incomplete killing in this case, or due to some low level contamination between libraries. Further experiments need to be done to determine if screens using this method would be viable (see Discussion).

Insertion site preferences

Previous investigations into the insertion preferences of piggyBac have used splinkerette PCR or similar methods to map insertions on a clone-by-clone basis (Ding *et al.*, 2005; Wang *et al.*, 2008; Liang *et al.*, 2009). A preference for insertion into active genes has been noted. These results are based on the order of 100 insertion sites. As my dataset contains thousands of insertion site sequences, I investigated whether anything further could be learned about insertion site preference. As the DNA-repair deficient cells are present as a tiny fraction of the pool, I considered all four lanes of sequencing data to be equivalent for these purposes.

I assembled a non-redundant set of insertion sites using a coverage cut-off of 1/50,000 of total reads for that lane. This set (nr50k; non-redundant, 1/50,000 cut-off) contains 3,714 insertion sites. In order to detect any bias in integration sites I also prepared a set of all the TTAA sites in the sequenced genome by

Number of samples containing insertion	All	> 1/50k coverage
1	10,687	575
2	1,866	393
3	982	458
4	2,984	2,288
Total	16,515	3,714

Table 3.2: Many insertion sites are unique to one sample and have low sequence coverage. Number of insertion sites present in 1, 2, 3 or 4 samples is shown for all identified sites and for sites filtered by coverage (more than 1/50,000 of the total reads for that lane).

Sample	Percentage pairwise agreement				Total overlapping insertion sites				
Unfiltered	1	2	3	4	1	2	3	4	Total
1	100.0	49.2	37.3	45.9	9,042	4,450	3,373	4,146	9,042
2	50.2	100.0	38.3	46.1	4,450	8,867	3,394	4,089	8,867
3	84.7	85.2	100.0	81.9	3,373	3,394	3,983	3,264	3,983
4	56.0	55.2	44.1	100.0	4,146	4,089	3,264	7,405	7,405
Average				56.2					
> 1/500k	1	2	3	4	1	2	3	4	Total
1	100.0	79.6	75.6	75.1	4,256	3,386	3,217	3,196	4,256
2	81.1	100.0	77.3	75.3	3,386	4,175	3,228	3,145	4,175
3	80.8	81.0	100.0	78.0	3,217	3,228	3,983	3,105	3,983
4	83.3	81.9	80.9	100.0	3,196	3,145	3,105	3,838	3,838
Average				79.2					
> 1/100k	1	2	3	4	1	2	3	4	Total
1	100.00	86.86	85.17	83.72	3,311	2,876	2,820	2,772	3,311
2	88.47	100.00	86.77	84.10	2,876	3,251	2,821	2,734	3,251
3	88.35	88.38	100.00	86.00	2,820	2,821	3,192	2,745	3,192
4	88.70	87.49	87.84	100.00	2,772	2,734	2,745	3,125	3,125
Average				86.8					
> 1/50k	1	2	3	4	1	2	3	4	Total
1	100.0	86.5	85.2	84.0	3,073	2,657	2,619	2,581	3,073
2	89.2	100.0	86.7	84.2	2,657	2,980	2,584	2,509	2,980
3	89.5	88.3	100.0	87.0	2,619	2,584	2,926	2,545	2,926
4	88.8	86.3	87.5	100.0	2,581	2,509	2,545	2,908	2,908
Average				86.9					

Table 3.3: Effect of minimum coverage filtering on agreement between samples

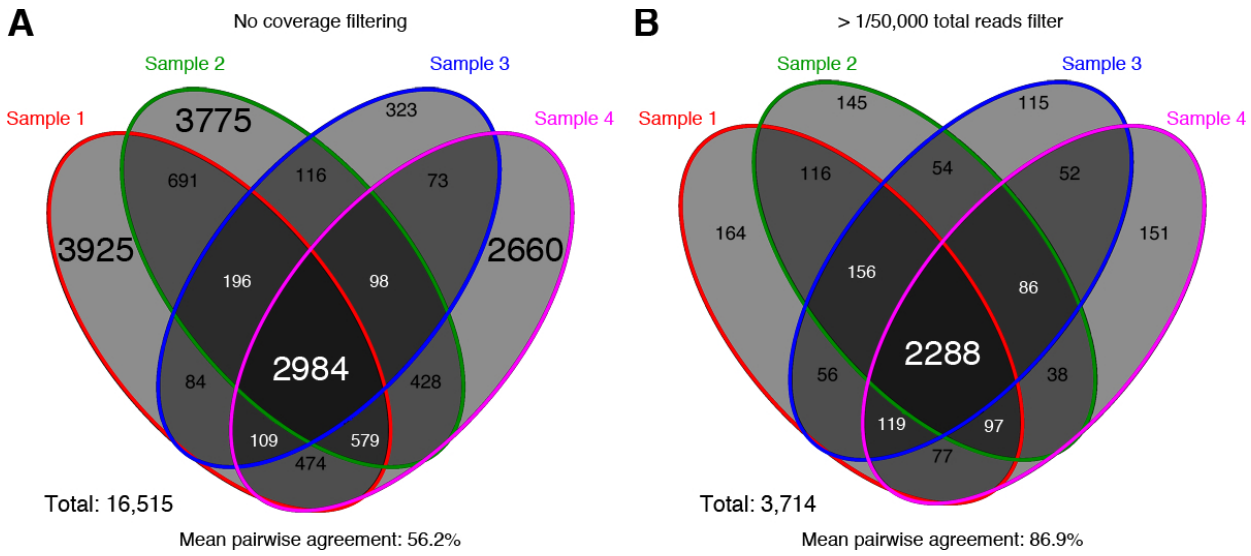

Figure 3.15: Venn diagram showing effect of applying coverage filter. Filtering by minimum coverage (B) mainly removes the insertion sites that are private to one sample. Sets with over 1,000 insertion sites are shown with larger text.

	1/50,000 coverage filter				Unfiltered			
	Untreated		Bleomycin		Untreated		Bleomycin	
Insertion site	1	2	3	4	1	2	3	4
13:73,482,861 (+)	66	52	–	–	66	52	–	–
2:18,718,350 (+)	268	–	–	–	268	–	–	–
5:86,663,315 (+)	–	–	–	–	**3**	**11**	–	–
18:37,789,859 (+)	276	216	–	–	276	216	**1**	–
17:35,417,330 (–)	45	61	–	–	45	61	–	–
16:16,036,599 (–)	34	–	–	–	34	–	–	–
17:87,847,692 (+)	–	233	–	–	**14**	233	–	–
5:36,927,022 (–)	–	–	–	–	–	–	–	–
18:40,467,674 (–)	–	–	–	–	–	–	–	–
13:38,461,580 (–)	–	–	–	–	–	–	–	–

Table 3.4: Search for mapped insertion sites in *Xrcc4* and *Xlf* mutants. The number of reads (no PCR duplicates) representing each insertion is shown for all samples in which that insertion was found. Samples 1 and 2 were untreated, 3 and 4 treated with bleomycin. – indicates that the insertion site was not detected in that sample. Entries in bold are those only detected without coverage filtering. + or – indicate the orientation, + being with PB5 nearest to the centromere.

using the nested MICA set of programs[2] to search the genome for occurrences of the TTAA motif (`nmscan` using a TTAA position weight matrix with values of 1; Down and Hubbard (2005)). I used another program in the nested MICA suite (`nmbrandfeat`) to analyse overlaps between sets as described below.

TNP insertions occur preferentially in genes

First, I checked whether the previously observed preference for piggyBac to insert into active genes was also the case for the TNP transposon. I mapped all the insertion sites in the nr50k set and found that 42.4% were in annotated coding regions of the genome (from Ensembl release 55). Only 36.3% of TTAA motifs are in genes, showing that the transposon has a preference for genes that is not explained by an uneven distribution of TTAA sites (Table 3.5). Furthermore, by filtering the genes with transposon insertions based on their expression in ES cells, as judged by presence in gene trap libraries using promoterless splice acceptor vectors, I also confirmed that piggyBac inserts into active genes more often. Seventeen percent of TTAA motifs were in the trapped gene set, compared to 27% of the experimentally determined integration sites.

It is possible that the discrepancy could result from genic sequence being intrinsically more complex than intergenic sequence, and therefore there would be a greater probability of obtaining a unique mapping for the transposon sequencing reads. To address this, I took a random sample of 50,000 TTAA sites across the genome, and retrieved 74 bp of adjacent sequence, plus 76 bp from the opposite strand at a distance of 200 bp 3′ to the TTAA. This models a paired end sequencing read (GG was added 5′ to the 74 bp end to mimic the GGTTAA transposon tag). These were processed exactly as above, to model the 'mappability' of sequence surrounding known TTAA sites. I found that 97% of these sequences were mappable using my procedure; thus differences in 'mappability' cannot explain the differences in PB insertions compared to random TTAA sites that I observed.

As gene trapping requires DNA to be introduced into the genome via some kind of vector (mostly retroviruses), it is possible that this analysis is instead detecting some common requirement between the various gene trap vectors and PB. To address the question more directly, I used gene expression data from a published microarray experiment to obtain an independent set of expressed genes (GEO accession: GSM198062, Mikkelsen *et al.* (2007)). I combined probes present in all three replicate experiments and obtained the corresponding Ensembl gene IDs for comparison with my list. This analysis gave essentially the same results as using the gene trap data, with 28% of the nr50k insertions in expressed genes compared to 15% of all TTAA sites.

[2] http://www.sanger.ac.uk/Software/analysis/nmica/

PB insertions are associated with features of 'open' chromatin

The observed preference for active genes may be linked to chromatin state rather than transcription *per se*. Chromatin state can be probed directly, by analysis of sensitivity to DNA endonuclease I (DNAseI), or indirectly by analysis of histone modifications associated with gene expression ('open' chromatin) or repression ('closed' chromatin). This information has been collected for ES cells (Mikkelsen *et al.*, 2007) and is contained in the 'regulatory features' data track in Ensembl. I filtered the data (downloaded from Ensembl release 55) to obtain lists of features that contain annotated ES cell DNAseI hypersensitive sites. I repeated this for histone 3 lysine 4 trimethylation ($H3K4Me_3$) and RNA polymerase II occupancy as determined by chromatin immunoprecipitation. The nr50k set of transposon insertions associated significantly with all of these features (One-sided binomial test, $P < 10^{-16}$ in all cases).

I also checked to see if the features examined correlated with each other, or if different features explain different subsets of PB integrations. Most insertions in DNAseI sites were intergenic with respect to the definition of gene used here—i.e. a transcribed region (Figure 3.17). DNAseI sites often occur in the promoter region of genes. However, all annotated $H3K4Me_3$ features were also associated with DNAseI hypersensitivity; thus examining this association does not give any extra information about the transposon preference.

PB insertions are under-represented in lamin-associated domains

To further test the hypothesis that chromatin state can influence PB transposition, I investigated whether PB insertions were excluded from lamin associated domains (LADs). These are regions that are spatially associated with the nuclear lamina, and are enriched for heterochromatin and unexpressed genes. Using ES cell LAD mapping data from Peric-Hupkes *et al.* (2010), I found that PB insertions are significantly underrepresented in LADs ($P < 10^{-16}$, bi-

Feature	Total TTAAs	nr50k	Total TTAA (%)	nr50k (%)
Genes	4,905,936	1,576	32.80%	42.4%
Trapped genes	2,549,063	1,009	17.04%	27.2%
Expressed genes	2,349,387	1,041	15.7%	28.0%
(Genes on chip[a])	4,324,224	1,491	28.9%	40.1%
PolII	7,004	42	0.05%	1.1%
DNAseI	181,171	432	1.21%	11.6%
$H3K4Me_3$	44,755	146	0.3%	3.9%
LADs	6,944,026	659	46.4%	17.7%
All	14,959,110	3,714		

Table 3.5: Association of PB integrations with genes and chromatin features. [a]Sites in genes for which probes were present on the microarray used for expression analysis.

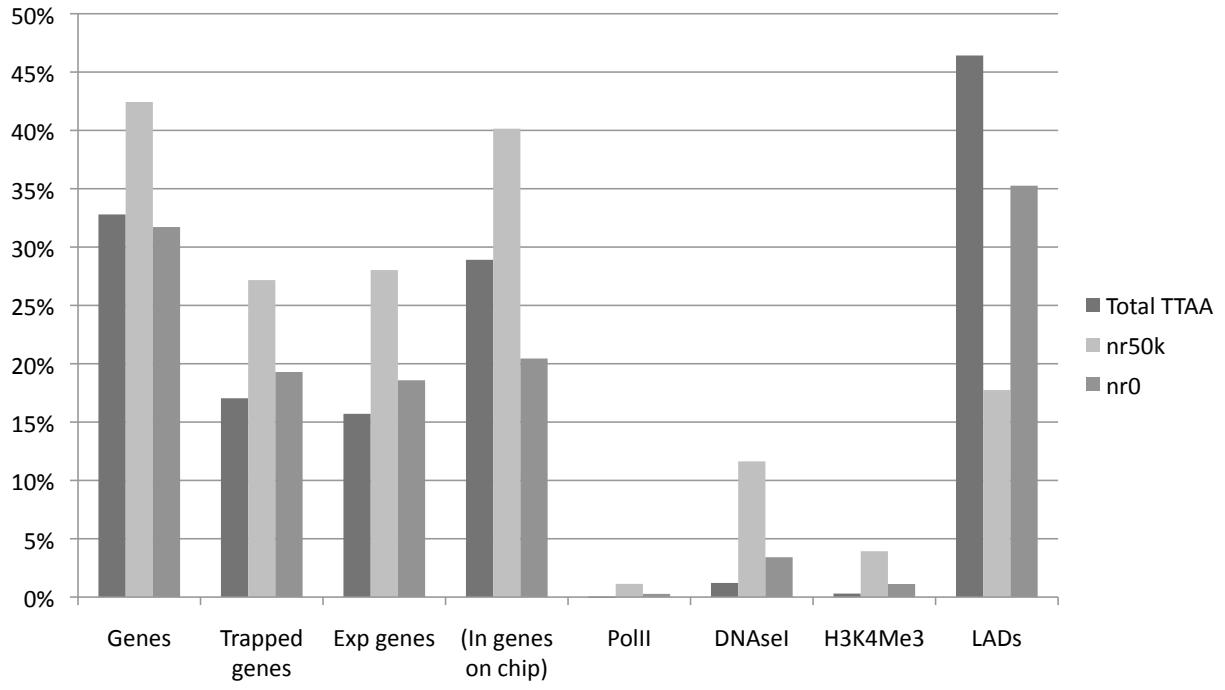

Figure 3.16: Graph of associations of PB insertions with genes and chromatin features.

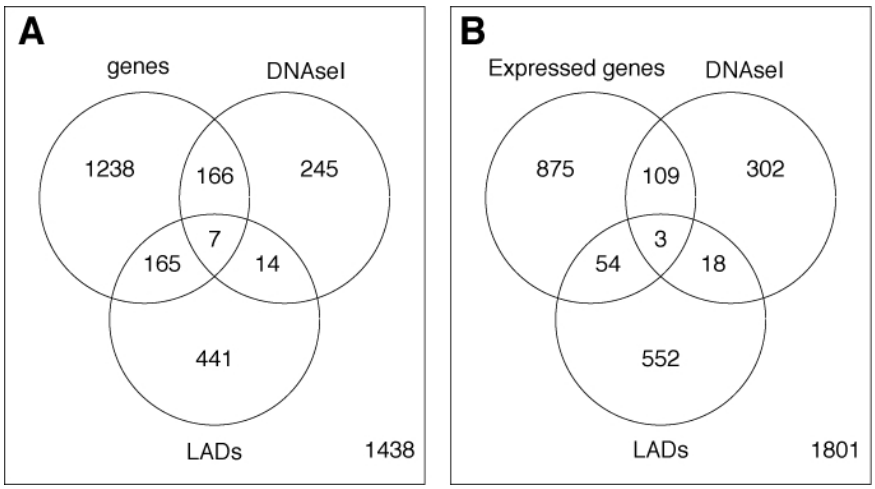

Figure 3.17: Venn diagrams illustrating insertion sites that overlap multiple features. A—Some DNAseI sites are associated with genes, but not generally with LADs. B—Although some sites in LADs are also in genes (see A), only 33% of these genes (57/172) are expressed in ES cells. This compares to 65% (112/173) of DNAseI sites in genes.

nomial test). Although 46.4% of TTAA sites are lamin-associated in ES cells, only 17.7% of nr50k insertions overlapped with a LAD. Of the PB insertions that were in LADs, there was little overlap with DNAseI sites (3.1%) or ES cell expressed genes (8.6%), although 26% overlapped when all genes were considered.

Effect of coverage filtering on observed insertion preferences

I repeated these analyses on a non-redundant set of insertions assembled without filtering by coverage (set named nr0, containing 16,515 insertions in total). All the associations became much weaker, and the distribution closer to that expected for random choice of TTAA sites (Figure 3.16). Some possible explanations for this are discussed below.

3.3 Discussion

3.3.1 The TNN/TNP transposon vector—mutagenesis

I describe above the construction of a PB transposon vector for causing loss of function mutations without the selection requirement of conventional gene trap mutagens. This should expand coverage of the libraries created to genes that are not expressed at the time of mutagenesis. Indeed, when I sequenced a large number of integration sites in the second part of this chapter, many were in genes that are not expressed in ES cells.

There are several splice acceptor elements in common use as components of mutagenesis vectors. These have generally been in use for many years and although they are clearly functional, there is little data concerning the efficiency with which they can compete for splicing with endogenous splice acceptors. To obtain this information, it is necessary to make homozygous mutations and see if the wild-type transcript can be detected. It is likely that many of these splice acceptor sequences came into use as a matter of convenience, prior to the availability of the genome sequence, depending on what had been cloned and was available. I took a logical approach to choose novel sequences to use, by scanning the mouse genome for sequences that fit a set of criteria for what could be considered an effective mutagen. The outcome was to use pairs of terminal exons, and their preceding splice acceptors, from two mouse genes—*Ccdc107* and *Dom3z*. These appear to be effective mutagens (Chapter 5 and Li (2010))

3.3.2 The TNN/TNP transposon vector—copy number selection

The second component of the vector is a dual selection cassette. The intention is to allow any increase in copy number of the transposon to be selected for. I designed the construct to switch reversibly between expressing the two resistance genes *neo* and

*pu*Δ*TK*. I refer the the transposon as TNN when it is in the *neo*-expressing orientation, and TNP when *pu*Δ*TK* is oriented with the promoter. I have shown that these two resistance genes can be independently selected for. In some cases, there was some background of cells resistant to the 'wrong' drug given the orientation of the transposon, but I showed that this arises from leaky activation of the inducible ERT2-Cre gene in the cells used. These cells were not resistant to both drugs simultaneously, indicating that the two genes can only be expressed mutually exclusively, as designed.

An alternative approach would have been to make an irreversible switching construct, by orienting the loxP sites to delete one of the resistance genes, or by using variant loxP sites to make the inversion irreversible. I decided against this as using such a method relies on inefficient Cre activity, such that Cre-mediated recombination only occurs on one copy. Cre can be very efficient in ES cells, so I decided on a reversible inverter-type construct to take advantage of this and allow me to use the most active Cre transgenes available. Cre activity does vary with locus (Vooijs *et al.*, 2001), so the best strategy is to strive for the most active Cre conditions possible in order that recombination should be maximally efficient at as many loci as possible. If recombination is efficient and goes to completion, there should be a 50% chance of a cell with two copies ending up with one TNN and one TNP copy.

3.3.3 Coverage and insertion site preferences of PB

PB associates with genes and 'open' chromatin

I carried out an experiment to sequence a large set of PB insertion sites. Previous attempts to define coverage of mutant libraries have been unsatisfactory, as they only examine a small number of loci. Sequencing all insertions in the pool using the method presented here should allow the coverage of the library to be defined completely, which would be a great improvement on previous methods.

The results of this experiment, which represents the largest set of PB insertions published so far, also allows the insertion site preference of PB to be investigated in more detail. Fortunately there are many useful ES cell datasets with genome-wide information on chromatin states and gene expression that can be used to analyse association of PB insertions with various features. I found significant associations of PB with genes and expressed genes, as had been previously reported, but also with markers of chromatin state, particularly DNAseI hypersensitivity. One recent report did note an association with DNAseI sites in T cells (Huang *et al.*, 2010). In my dataset, PB insertions also appeared to be excluded from the highly chromatinised lamin-associated domains. These results suggest that PB could be a useful tool for monitoring chromatin accessibility, although the exact parameters that govern insertion remain to be determined. Insertions in genes and annotated DNAseI hypersensitive sites made up half of the mapped insertion sites analysed here, but none of the features I investigated explained the other half. Closer examination of these may give more insights into the biology of PB transposition.

Potential effect of subpopulations in cell cultures

Although insertion sites were enriched in expressed genes and depleted in LADs, some were still found in regions that are not expressed or are lamin-associated in ES cells. This could represent genuine differences in the chromatin state at these loci, but could also arise if there is a subpopulation of differentiated cells in the culture, with differences in expression profile or chromatin changes. However, such a population would have to be expandable, as cells were subcultured several times between transposition and sequencing. As most cells under the microscope were ES cells as judged by morphology, this is unlikely to have had a large effect on the results. Another possibility is that transposition could occur after breakdown of the nuclear lamina. Further experiments could include using cell cycle specific transposases to address this (see Chapter 6) and mapping insertion sites in mutant ES cell lines lacking chromatin modifying enzymes.

A potential subpopulation that could markedly affect the results is if some cells continue to express the PB transposase, as a result of integration of the expression plasmid into the chromosome. Over the total time of the experiment, many rounds of transposition could take place in these cells, resulting in a large number of unique insertions, each represented by relatively few cells and unique to one pool. This is likely to be the reason for the low agreement between pools of insertion sites with low coverage. As the PCR amplification is minimal, the number of reads should approximate the relative numbers of cells with each insertion. I consider this to be the most likely reason for divergence between the pools. This suggests that an improvement to the protocol would be to use PBase mRNA, which would re-

move the possibility of integration. Additionally, technical replicates should be sequenced to assess how completely the pool is sampled, which could be another source of disagreements between pools. Another useful exercise would be to sequence the library prior to splitting into separate pools—this would answer the question of whether transposition is continuing after the split.

Interestingly, including these low-coverage and poorly-reproduced insertion sites in the feature enrichment analysis resulted in a distribution much closer to that expected for a random choice of TTAA (Figure 3.16). This suggests that it may be only the most highly represented and reproducible insertions that have a preference beyond the TTAA requirement. One possibility is an extra requirement imposed by the puromycin selection, rather than by the transposon itself, as expression of the resistance gene at sufficient levels may require an open chromatin context. If the low-coverage insertions do indeed arise from transposition later in the culture, these would not be subject to such selection and therefore may show a wider distribution of insertion sites. However, another study in which PB insertions were not directly selected for also showed a preference to genes similar to my puromycin-selected insertions (Liang *et al.*, 2009). This question could be further addressed by sequencing libraries without selection for insertion. However, in the situation I envisage for screens the insertions will be selected for, so this needs to be accounted for in experiments to investigate coverage.

The poor agreement between the four sequenced pools in this experiment hampered attempts to identify loss of the tagged bleomycin sensitive mutants (Table 3.4). As mentioned above, avoiding transposase plasmid integration and including technical as well as biological replicates would be necessary first steps in improving the method. An additional consideration would be to make libraries complex enough to have multiple insertion sites per gene, as this would give confidence that any change in abundance was not due to a background mutation.

3.3.4 Conclusions

I have described the construction of a vector combining the PB transposon with a mutagen designed to be effective at a wide range of loci and a double resistance cassette that should be selectable based on copy number. High throughput sequencing of insertion sites allows the coverage of libraries created with this transposon to be determined more thoroughly than previous methods. With some refinements, this method may also be applicable to screen the resulting libraries.