Isolation of homozygous mutant mouse embryonic stem cells by selection for copy number increase

Stephen J. Pettitt

Wellcome Trust Sanger Institute

and

Girton College, University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy.

June 2, 2011

Acknowledgements

It has been a great privilege to be part of the Wellcome Trust Sanger Institute, and I am very grateful to the Wellcome Trust for funding my time here. I have met many talented and inspirational scientists here, without whom nothing in this thesis would have been possible.

Foremost among these is my supervisor, Allan Bradley, for sharing his wealth of knowledge and experience. He has given me great ideas and opportunities for research projects, as well as a lot of freedom to pursue my own interests. I am very grateful for his trust and support, and consider myself very lucky to have been his student.

I am also very grateful to my other supervisors at the Sanger Institute. Gavin Wright, Christian Söllner, Jon Teague, Adam Butler, Pentao Liu and Mariaestela Ortiz. Haydn Prosser deserves special mention for putting up with my mistakes and questions, while teaching me everything about gene targeting. I was also fortunate to collaborate with Bill Skarnes, who has continued to be a great source of advice. The enthusiasm of my external supervisor, Steve Jackson, has been infectious, and he gave me a lot of useful suggestions for DNA repair experiments. I also thank Junji Takeda (Osaka University) for an interesting discussion of ideas during his visit last year.

Being surrounded by so many successful and bright people in the lab is great motivation. I would particularly like to thank Kosuke, Yue, Amy and Wei, who were always ready to give advice and shared materials and results with me. Frances, James, Alastair and Hiroko all helped my cell culture go smoothly, and Nathalie and Charles read my thesis. Thanks also to EPT for looking after my cells and (sometimes) cooking me curry, Floris for making the espresso machine happen, Mekayla for making me go outside and run around, Roland, who can advise on absolutely anything, and George for some truly terrible jokes.

I thank the fellows, resident and honorary, of the Cavendish Institute for their numerous collaborations in the lower impact, but essential, research fields such as European geography, anthropology, oenology, pyrotechnics, beardomics and bioinformatics. Long may they continue. I am also lucky to have one long-term collaborator in particular, who truly understands the demands of ES cell culture.

Finally, I would like to thank my parents, who have always encouraged and supported me in anything I wanted to do. Their interest in my work means a lot to me; I hope they enjoy reading!

SJP
September 2010

Abstract and declaration

Forward genetic screens are a powerful method to determine which genes are responsible for a particular phenotype in many model organisms. However, a simple method to conduct genetic screens in a mammalian system has been difficult to develop, due to the problem of making random homozygous mutations in the diploid mammalian genome. Mouse embryonic stem (ES) cells provide a convenient model for mammalian cell biology. Previous studies showed that heterozygous mutants in ES cells with mutations in the *Blm* gene segregate homozygous mutants at a low rate, due to an increase in mitotic crossovers.

Using the piggyBac DNA transposon (PB) for initial single copy heterozygous mutagenesis, I describe a method to isolate the rare homozygous cells based on selection for transposon copy number, which increases to two in homozygotes. I successfully isolated homozygous mutants using this system, but my experiments revealed aneuploidy as an alternative copy number gain pathway in ES cells. By extensive engineering of the ES cell line and PB transposase, I developed a method to allow many different homozygous mutants to be generated in a pooled format. This minimises the problem of background from aneuploidy and allows isolation of clonally pure mutants suitable for genetic screens.

I also investigated the properties of the PB transposon. By sequencing and mapping thousands of insertion sites I have investigated the insertion site preferences of PB. This method can also be used to fully define coverage of mutant libraries. I showed that precise excision of PB from the genome depends on the nonhomologous end joining pathway, and present data indicating that transposition can occur throughout the cell cycle.

The methods and tools presented will be useful for study of gene function in mammalian cells, and are also applicable for the study of DNA double strand break repair and copy number stability.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The human and mouse genomes

Modern molecular biology is defined by the analysis of the human genome sequence, published in draft form in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). The availability of a reference genome sequence has changed the way research is conducted. However, the initial analysis of the genome was also humbling in some ways, revealing how little was known, and how much is still to be discovered. For example, the number of genes in the genome had to be revised sharply downwards from pre-genome estimates of over 100,000 to the current consensus of just under 22,000 (protein coding genes, Flicek *et al.* (2010)). In contrast, the known extent of transcript diversity—revealed by mapping transcribed sequences back to the reference genome—has increased, as has the number of known genes such as microRNAs that do not code for proteins (Gardner *et al.*, 2009). Even if the full complement of genes can be identified, there is still very little information about what they all do. The next step is to address this, by annotating the genome with functional information.

1.1.1 New genetic approaches

The availability of a reference genome sequence has transformed the study of the genetic basis of disease. One approach that has been enabled is the genome-wide association study (GWAS). By genotyping variants in large cohorts of patients and controls, loci can be identified that associate with disease. Many such studies have been published, identifying variants associated with a wide range of diseases and traits (Wellcome Trust Case-Control Consortium, 2007). The approach is essentially an observational one on a large scale. Still greater resolution is required however, as these studies usually only identify a small region, and cannot formally distinguish between a genotyped variant and a closely-linked causal variant. New technology is allowing a wider range of variants to be genotyped (Wellcome Trust Case-Control Consortium *et al.*, 2010). Occasionally, a variant may be in a gene and make sense, for example the identification of *BCL11A* variants that cause elevated foetal haemoglobin levels in adults (Menzel *et al.*, 2007), or the implication of *IL23R* variants in inflammatory bowel disease (Duerr *et al.*, 2006). However, further mechanistic studies are required to confirm the causal variants.

Sequencing technology and capacity continues to advance, bringing more resequencing approaches for discovery of variants associated with disease within reach in terms of time and cost. For rare diseases inherited in a Mendelian fashion, the causal variant can often be found by sequencing all exons of just a handful of affected individuals. This can now be done for well under $10,000 (Ng *et al.*, 2010; Lupski *et al.*, 2010). Another application is in the study of cancer, where large scale sequencing of tumours can be used to completely catalogue the somatically-acquired mutations present (Sjöblom *et al.*, 2006; Wood *et al.*, 2007; Ley *et al.*, 2008; Dalgliesh *et al.*, 2010; Pleasance *et al.*, 2010b,a). It is now possible to sequence sufficient numbers of samples at high enough coverage to distinguish recurrent 'driver' mutations from background 'passenger' mutations by statistical methods (Greenman *et al.*, 2007). However, in order to conclusively prove oncogenic function and further investigate the mechanism, experimental approaches are still required (Su *et al.*, 2008).

To test any hypothesis about the function of a gene, it is usually necessary to do an experiment. This may not be possible in humans, therefore another important source of genome annotation is by homology, extending experimental findings about the function of a gene in model organisms to the homologous gene in humans. For this reason, the mouse genome sequence, published shortly after the human sequence, was eagerly awaited (Mouse Genome Sequencing Consortium, 2002).

1.1.2 Importance of the mouse genome

The biology and history of the laboratory mouse make it the ideal mammalian model organism. Being a mammal, many aspects of physiology are similar to humans, meaning that higher-level functions can be studied compared to more distantly related model organisms. Crucially this also means that

mice are susceptible to many of the same diseases and pathogens as humans, and can be used to model these.

Analysis of the mouse genome confirmed many similarities with the human sequence. Syntenic regions, in which the order of genes is preserved, can be identified for 90% of the human and mouse genomes. One or more human homologues can be identified for 99% of mouse genes, and in 80% of cases the human counterpart is unique and syntenic. Homologues are much harder to identify in other model organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans*, reflecting their much earlier common ancestor with humans—About 700 and 1,000 million years ago respectively, compared to 65 million years ago for mouse (Rubin *et al.*, 2000; Silver, 1995).

Practically speaking, mice are small and easy to house, and have a short generation time for a mammal (around 10 weeks). This relatively short breeding time means that genetic experiments are possible, and there are excellent genetic resources and technologies available to pursue these, described below. Many experimental techniques in mice that were once laborious are now routine, thanks to the reference genome sequence. I have outlined some of these techniques, and how they can be used to assign function to genes, below. Several of these approaches were originally developed in other model organisms, and have been extended to the mouse. The experiments described in this thesis form part of this ongoing effort to transfer the range of genetic tricks available in yeast, *Drosophila* and *C. elegans* to mammalian systems.

1.1.3 Experimental approaches to analyse gene function

When a experimental geneticist plans an investigation into a biological system or process, the first question that comes to mind may well be "how can this go wrong?". The rationale is that by discovering and studying the basis of defects in the process, the crucial elements will be revealed. The geneticist therefore seeks to obtain mutant organisms to study. The terms *forward genetics* and *reverse genetics* are used to describe the two fundamental ways of obtaining artificial mutants for study. In the forward genetic approach a population of random mutants is generated and individuals from the population, which carry different mutations, are examined until individuals showing the phenotype of interest are found. This process is known as genetic screening. The principles were first described by Muller (1927), and perhaps the best known example is the Nobel prize-winning screen for mutations affecting patterning of the *Drosophila* embryo (Nüsslein-Volhard and Wieschaus, 1980). For some phenotypes, the process may be simplified by an appropriate selection step which kills all mutants which do not show the phenotype of interest. For example, mutants of the bacterium *Escherichia coli* (*E. coli*) that are resistant to bacteriophage λ can be selected for simply by infecting a population with the phage. Surviving bacteria have mutations in the receptor for the phage (Randall-Hazelbauer and Schwartz, 1973). Once mutants have been identified, the molecular basis can be established—this normally involves finding the molecular lesion in the DNA and predicting the gene and protein that is affected. Thus, the starting point for forward genetics is a mutant phenotype, which leads to identification of a mutant genotype.

The reverse genetic approach begins with introducing a known mutation in the DNA. Reverse genetics is often more hypothesis driven than the forward approach, as for many organisms it is not possible or efficient to generate targeted mutations on a sufficiently large scale. In most cases therefore the gene has already been implicated in some way in the process of interest and is being mutated in order to study it in more detail. Once the mutant has been generated, unexpected phenotypes may be observed. Reverse genetics therefore leads from genotype to phenotype.

The two approaches should be properly thought of as complementary. The choice between them will often come down to how much is known about the process and which model organism is being used to investigate it. The great advantage of forward genetics screens is that unknown or unexpected components of a pathway can be identified. The ideal forward genetic screen, at complete saturation, would allow identification of all genes that are essential for the phenotype in question.

These broad approaches to the study of gene function were first developed in simple model organisms, such as phage, bacteria and yeast. In the following section I discuss how these can be applied to the mammalian model organism of choice, the mouse.

1.2 Reverse genetics in mice

Disrupting (commonly referred to as 'knocking out') a specific gene in a mammal requires extraordinary precision. The mouse genome is 2.5 Gbp (gigabase

pairs) in size, yet it is now possible to specifically change a single one of these base pairs as a result of developments in gene targeting technology. To do this in every cell of a full-grown animal would be an even more daunting task, so it is necessary to access the germ cells from which development begins. Isolation and culture of cells from the early embryo was the first step in making genetically modified mice. The development of these technologies, which is discussed below, was recognised by the Nobel prize for Medicine in 2007.

1.2.1 Embryonic stem cells

Mouse embryonic stem cells (ES cells) were first isolated from the inner cell mass of 3.5 dpc (days post coitum) blastocysts (Evans and Kaufman, 1981). ES cells can be cultured indefinitely, and like their counterparts of the inner cell mass they are pluripotent, with the ability to differentiate into cells from any of the three germ layers, ectoderm, mesoderm and endoderm. This can be demonstrated by injection of ES cells into syngenic mice, where they form teratomas—tumours consisting of different cell types (Evans and Kaufman, 1981). Another assay for pluripotency is injection into blastocysts and reintroduction to a foster mother, which results in chimaeric pups in which tissues are made up of a mixture of cells derived from the injected cells and the host blastocyst (Gardner, 1968). Cells derived from ES cells can be seen in the coat and eyes as pigmented regions if an albino blastocyst is used as the host, and use of genetic markers shows that this extends to internal organs. Examination of the injected embryos at later stages showed that ES cells can also contribute to extra-embryonic lineages (Beddington and Robertson, 1989). Crucially, ES cells retain the ability to contribute to the germ cell lineage and therefore these chimaeric mice can produce ES cell-derived sperm and oocytes, making it possible to transmit a haploid segregant of the ES cell genome to the F1 generation (Bradley *et al.*, 1984).

These technological advances opened up the possibility of genetic engineering in mice, as growing ES cells in culture provides an opportunity to make modifications. Shortly after the establishment of germline chimaeras, it was shown that these could also be derived from ES cells that had been modified by insertion of a retrovirus into the genome (Robertson *et al.*, 1986). The location of the insertion is random, although some experiments selected specifically for insertions at the X-linked *Hprt* locus by selection of ES cells in 6-thioguanine (6-TG). *Hprt*-null cells are resistant to 6-TG (see Chapter 2). Insertion at this specific locus is a rare event, but single cells can be isolated by 6-TG selection and expanded clonally prior to blastocyst injection. This selection does not compromise the ability of chimaeras to contribute to the germline (Kuehn *et al.*, 1987). The ability of ES cells to be continuously subcloned in this way makes the use of comparatively inefficient techniques for genome modification feasible, given a suitable selection scheme.

1.2.2 Gene targeting

The ability to reintroduce modified ES cells to the mouse germ line led to increased interest in methods to make specific modifications to the genome of mammalian cells. In yeast, introduction of plasmids with homology to chromosomal sequence had been shown to direct plasmid integrations to that sequence, particularly if a break was present in the plasmid homology (Orr-Weaver *et al.*, 1981). Early attempts to extend the technology to mammalian cells were inefficient. DNA also readily integrates into the genome of mammalian cells at random, and the early constructs used did not efficiently compete with this process, meaning that large numbers of random integrations were observed for every genuine gene targeting event. A targeted insertion at the β-globin locus in human cells used a plasmid containing an 11.1 kbp (kilobase pairs) homology fragment and a neomycin resistance gene (*neo*). The approach worked, but only 0.1% of G418-resistant (neo^+) cells had the targeted insertion (Smithies *et al.*, 1985). Using an artificially introduced chromosomal substrate in mouse cells to specifically select correct recombinants, an absolute efficiency of 0.1% of transfected cells (in this case by individual microinjection) was obtained. Considering the frequency of random integration, this is equivalent to 1% targeted integrations (Thomas *et al.*, 1986)

These approaches were extended to ES cells, again making use of the *Hprt* locus to easily select targeted integrations either by disruption of the *Hprt* gene, or rescue of a previously isolated spontaneous mutation (Thomas and Capecchi, 1987; Doetschman *et al.*, 1987). These experiments used insertion type vectors transfected by electroporation, obtaining targeting efficiencies (ratio of targeted to total transformed cells) ranging from less than 0.1% to 14% (Thomas and Capecchi, 1987; Doetschman *et al.*, 1987). It was also shown that the gene targeting procedure could be performed without compromising the potential of ES cells to contribute to the germ line of chimaeras (Thompson *et al.*, 1989; Koller

et al., 1989). These experiments paved the way for the study of mice with defined genetic modifications.

Although the *Hprt* locus was used for convenience in these early experiments, direct selection for the mutant phenotype was not essential, and targeting of many other loci was soon reported (Koller and Smithies, 1989; Johnson *et al.*, 1989; Joyner *et al.*, 1989; Schwartzberg *et al.*, 1989; McMahon and Bradley, 1990). Technical improvements to the method resulted in increased efficiencies. It was shown that insertion vectors (as used in many of the experiments described above) are generally more efficient than replacement vectors (Hasty *et al.*, 1991c). However, as insertion vectors conserve all sequence at the locus, and do not delete or modify DNA, the range of mutations that can be obtained with replacement vectors is greater. The differences are the position of the selectable marker gene (plasmid backbone for insertion, inside replaced region for replacement) and the restriction site used to break the targeting vector prior to transfection (inside the homology at the point of insertion for insertion vectors, outside the homology for replacement, Figure 1.1).

Targeting vectors

Several investigators carried out experiments to define the features of an efficient targeting vector. Close to 100% sequence identity, rather than simply homology, was found to be important (te Riele *et al.*, 1992). This can be accomplished by preparing targeting vector plasmids from genomic DNA libraries made from the same mouse strain as the ES cells to be used. Several such libraries exist, including some made from commonly used ES cell lines which should be as close to isogenic as possible (Adams *et al.*, 2005). More recent protocols to construct targeting vectors wholly within bacterial cells by the process of recombineering may also reduce the risk of mutations occurring during *in vitro* manipulation steps (Liu *et al.*, 2003).

Other important considerations in targeting vector design include the total length of homologous sequence. Experiments with different sized vectors targeting the *Hprt* locus demonstrated a linear increase in targeting efficiency with homology length above a minimum length of 1.9 kbp (Hasty *et al.*, 1991b). Generally at least 6 kbp of homology will result in a good targeting frequency while being easy to manipulate and maintain in *E. coli* by standard molecular biology methods. The homology can be distributed unevenly in replacement vectors as a long and short arm to aid genotyping. The short arm can be just 472 bp, although it is usually at least 1 kbp in practice (Hasty *et al.*, 1991b).

With the use of more advanced targeting vectors, gene targeting can be very precise, and is not limited simply to knockouts. Subtle mutations can be made using a two step insertion and reversion method named 'hit and run' (Hasty *et al.*, 1991a). Although selectable marker genes are still necessary even with the higher efficiencies obtained with better vector design, these can be removed using site-specific recombinases to leave a minimal impact on the locus. The most widely used recombinases are Cre and Flp (Sauer and Henderson, 1988; Schaft *et al.*, 2001). The expression of these recombinases can be restricted temporally or based on cell type. By positioning the recombinase target sites to flank critical regions of the targeted gene a conditional allele can be constructed, which is phenotypically wild type until expression of the appropriate recombinase (Adams and van der Weyden, 2008).

Study of knockout phenotypes

Long-term culture of ES cells runs the risk of abnormal variants arising in the culture that are not capable of contribution to the germline (Liu *et al.*, 1997; Liang *et al.*, 2008). Therefore to obtain a homozygous knockout, chimaeras are typically made from heterozygous ES cells. Once germline transmission has been confirmed, F1 offspring can be intercrossed to obtain homozygous F2 mice. Formation of chimaeras with high percentage contribution from ES cells depends on the injected ES cells successfully out-competing host cells in the blastocyst (Schwartzberg *et al.*, 1989). ES cells with a homozygous mutation may be at a fitness disadvantage and not form good chimaeras. Mice can be made directly from homozygous ES cells by the alternative technique of tetraploid complementation, although this method appears to only work effectively with hybrid ES cell lines (i.e. derived from an F1 outcross). This technique depends on the ES cells rescuing development of a tetraploid embryo formed by fusion, which is otherwise only competent to form extra-embryonic cell lineages (Nagy *et al.*, 1990).

Gene targeting requires knowledge of the sequence of the gene in question. It is in this area that the genome sequence has contributed. Instead of laboriously cloning a gene, with enough flanking genomic sequence from which to make a targeting vector, the sequence required can now be looked up directly. Moreover, large bacterial artificial chromosome (BAC) libraries, consisting of *E. coli* vectors with 100–200 kbp mouse genomic inserts, were used during the sequencing projects. These repre-

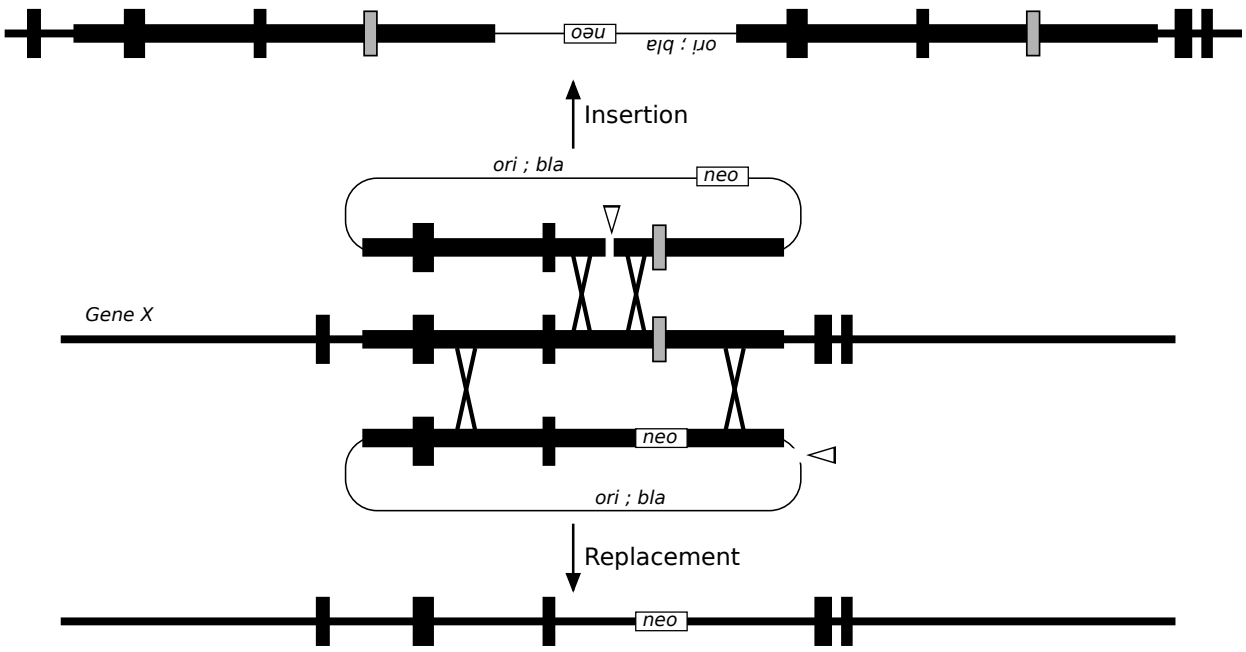


Figure 1.1: Insertion and replacement targeting vectors. The structures of insertion (top) and replacement (bottom) vectors targeting a hypothetical gene are shown. An open arrowhead indicates the site for linearisation by restriction digest. Thick line indicates homology between the genome and the targeting vector. *ori*, bacterial replication origin in plasmid; *bla*, bacterial ampicillin resistance gene.

sent ideal physical sources of DNA for vector construction and are indexed by genome position. The shotgun subcloning approaches have even been developed to make indexed libraries of insertional targeting vectors for mutagenesis and chromosome engineering (Zheng *et al.*, 1999; Adams *et al.*, 2004). Designing and synthesising a targeting vector for every known gene in the mouse is now feasible, and is being undertaken by an international consortium (International Mouse Knockout Consortium *et al.*, 2007). Thus the genome sequence has been a boon for the already fruitful area of reverse genetics in mice.

1.3 Forward genetics in mice

Gene targeting has been the flagship experimental method in mouse genetics. However forward genetic screens are also possible in mice, and may be due a renaissance in the light of the genome sequence.

1.3.1 Inbred strains

Mice have been used as a model organism for mammalian genetics for over a century, since Mendel's laws were first shown to apply to mouse coat colour mutations at the turn of the 19th century (Cuénot, 1902, 1903). Like most sexually reproducing organisms, mouse chromosomes recombine and reassort at meiosis during gamete formation, to produce genetic diversity. The pioneers of mouse genetics quickly realised that pure-bred lines of mice, homozygous at all loci across the genome, would be essential to provide a defined, invariant genetic background on which to conduct experiments. These inbred strains are obtained by many generations of brother-sister matings. The first experiments of this type, resulting in the DBA strain, were carried out by C.C. Little, founder of the Jackson Laboratory, in 1909. After 20 generations of such matings, 98.7% of the genome will be fixed (homozygous) (Silver, 1995). Stocks of inbred strains from commercial mouse breeders have been maintained for over 200 filial generations. Mutations isolated in diverse genetic backgrounds can be crossed back to an inbred strain to form a congenic strain, which contains only the mutant region on an otherwise known genetic background. This allows comparisons to be made between mutations without confounding effects from differing genetic backgrounds. One early success of mouse genetics, which relied entirely on the availability of inbred strains, was the charac-

terisation of the genetics of the major histocompatibility complexes by transplanting tumours between different inbred and hybrid strains (Snell and Stimpfling, 1966).

The process of inbreeding can isolate naturally-occurring mutations. As all alleles eventually become homozygous, the effects of recessive alleles will be observable. Some alleles are isolated by design of the process, e.g. the coat colours used to identify mice (DBA above stands for *dilute, brown, non-agouti*), and alleles with effects on reproductive fitness. However a large number of other, unknown mutations were also fixed during the production of these strains, which included susceptibilities to cancer and various other diseases (Murphy, 1966; Russell and Meier, 1966). These mutants provided valuable models of human disease for study of pathology. In fact, the susceptibility of the 129 mouse strain to testicular teratomas, which occur in about 1% of males (Stevens and Little, 1954), was the start of research leading to the derivation of the first ES cell lines from this strain. The particular ease of deriving ES cells from 129 mice may be linked to this mutation (or mutations), but its molecular basis is still unclear.

Determination of the genetic basis of the mutations had to wait for the development of more advanced molecular biology techniques associated with recombinant DNA technology. Discovery of restriction fragment and simple satellite length polymorphisms allowed linkage maps of the mouse to be drawn up (Dietrich *et al.*, 1992). This allows the mutations present in inbred strains to be mapped more precisely, and eventually cloned and the exact lesion determined. Many single gene traits were cloned using this process, although this was not always trivial even for well known mutations such as coat colour alleles (Jenkins *et al.*, 1981; Bultman *et al.*, 1992). The nature of the naturally occurring mutations in these strains (deletions, base substitutions, insertions etc.) is unknown and can vary. A project begun recently aims to fully sequence a number of inbred strains in full, which should identify more of these mutations[1] (Turner *et al.*, 2009; Sudbery *et al.*, 2009). However, with the development of experimental mouse genetics, it is unlikely that new inbred strains carrying naturally occurring mutations will be isolated for the direct analysis of phenotype in future. An exception is the collaborative cross, which aims to isolate over 1,000 new inbred strains derived from a mixed population of eight classic inbred strains to study more complex traits in these strains (Churchill *et al.*, 2004).

The limitations of using naturally-occurring 'mutant' alleles led to the development of experimental mutagenesis protocols. When making experimental mutants for study, a mutagen which causes well-defined and easily mappable lesions needs to be used. Using a mutagen also increases the number of mutations that can be generated, as the natural mutation rate is very low, of the order of 10^{-8} mutations/nucleotide/generation (Haldane, 1935; Xue *et al.*, 2009). Some mutagens that can be used are discussed below.

1.3.2 ENU mutagenesis

Alkylating agents such as *N*-ethyl-*N*-nitrosurea (ENU) are chemicals that directly alkylate bases in DNA. Most mutations caused by ENU are transition point mutations (A to G, C to T or vice versa). A major advantage of ENU mutagenesis is that it can introduce subtle mutations that can be either loss of gain of function. It is therefore possible to recover a variety of alleles for the same locus, which can be valuable for later analysis. However single base mutations such as these are notoriously difficult to map, a process that requires extensive outcrossing and subsequent genotyping of polymorphic markers. Although this has become easier with denser polymorphic markers and the availability of genome sequence, mapping can still take years.

A number of screens have successfully used ENU mutagenesis. The usual method is to generate mutations in spermatogonial stem cells by ENU injection. These mice then act as founder stock, and can be bred to a wild type female to give heterozygous G1 mutants. Dominant mutations will be picked up in these mice. Further breeding allows homozygous mutants to be recovered, in which the effect of recessive mutations can be seen.

Some examples of successful ENU mutagenesis screens include the identification of the *Min* allele of the *Apc* tumour suppressor gene (Moser *et al.*, 1990; Su *et al.*, 1992) and the cloning of the circadian rhythm regulator *Clock* (Vitaterna *et al.*, 1994). Several centres have generated large series of mutants with various phenotypes (Rastan *et al.*, 2004; Hrabé de Angelis *et al.*, 2000), although the effort to map these mutations is still ongoing.

A number of new technologies are improving ENU mutagenesis. One is the development of mouse balancer chromosomes that allow recessive lethal mutations to be isolated in a specific region. Balancer chromosomes were originally developed in *Drosophila* screens. They are engineered chromosomes with

[1] http://www.sanger.ac.uk/resources/mouse/genomes/

two main features: First, a large inversion typically spanning ten million or more base pairs of gene rich sequence. This is the "balanced" region in which recessive lethal mutants can be easily isolated. The inversion suppresses meiotic crossover in this region, such that a mutation in the homologous region on the normal chromosome will never transfer to the balancer chromosome by crossing over. If crossing over does occur, a lethal dicentric chromosome will result. The second element is a linked recessive lethal mutation that prevents recovery of animals homozygous for the balancer chromosome. Other linked markers, such as coat colour, may be included so that animals carrying one copy of the balancer are easily identified. When an animal carrying a recessive lethal mutation in the balanced region is crossed to the balancer stock, this can be identified if all progeny carry the balancer coat colour—i.e. no progeny with two non-balancer chromosomes are identified.

Balancer chromosomes, by their nature, do not help for a genome wide screen but are useful for studying particular areas of interest. The most complete balancer screen conducted so far has resulted in hundreds of developmentally lethal mutants in an interval on mouse chromosome 11 (Kile *et al.*, 2003).

Another technology that may lead to a renaissance in ENU mutagenesis screens is the continuing improvement and cost-efficiency of sequencing. Cheaper sequencing of whole genomes, or of candidate regions by microarray capture of DNA corresponding to the region (Albert *et al.*, 2007), may simplify mapping of ENU-induced mutations. Any improvement in mapping, especially without involving breeding, will greatly strengthen the case for ENU mutagenesis. The range of mutations obtainable with ENU is the greatest strength of the method compared to the others below, which generally produce (or at least aim to produce) straight knockouts. Currently, the requirement for breeding to map mutations by linkage analysis means that ENU is not ideal for mutagenesis in cell lines.

1.3.3 Irradiation

Gamma radiation is a potent mutagen that causes a number of DNA lesions, including double strand breaks (DSBs). Inaccurate repair of DSBs can result in chromosomal imbalances—deletions, duplications, or translocations where part of one chromosome is joined to another. Deletions are the most useful in terms of creating mutants. Deletions can be large or small, and can affect many genes at once. A full gene deletion is the most robust knockout mutation, as there is absolutely no possibility of residual activity of the affected gene(s). However, as with ENU, the problem lies in mapping the mutation. The possibility of affecting multiple genes could be viewed as an advantage, but in most cases a deletion spanning multiple genes complicates analysis, making additional experiments necessary to establish which deleted gene causes the phenotype.

Mapping of deletions has improved with the development of increasingly high resolution comparative genomic hybridisation (CGH) arrays (Pinkel *et al.*, 1998). CGH compares copy number across the genome between two DNA samples by competitive hybridisation of probes labelled with two different fluorescent dyes. The first generation of CGH arrays used spotted bacterial artificial chromosomes (BACs) to make microarrays for the hybridisation and thus had a resolution of only around 100 kb (Cai *et al.*, 2002), however current arrays use oligonucleotide probes synthesised in parallel directly on the slide (Barrett *et al.*, 2004). As well as allowing only specific regions to be investigated, this improves resolution to the order of ten bases. New sequencing technologies can also be used to investigate copy number variation and rearrangements (Korbel *et al.*, 2007).

Even with improvements in mapping, the problem of formally establishing causality still remains for irradiation mutants. Technologies such as recombinase mediated cassette exchange (RMCE, Seibler *et al.* (1998); Prosser *et al.* (2008)), which allows reintroduction of BACs into an engineered locus to test for phenotype rescue, may help. However as deletions induced by irradiation can be very large, many BACs may need to be tested, and for experiments in cell lines a suitable acceptor locus must be engineered before mutagenesis (Xiong, 2008).

1.3.4 Insertional Mutagenesis

A variety of DNA elements are available that can insert into genomic DNA. This is a great advantage for a mutagen, as the inserted DNA is of known sequence and therefore tags the mutated locus. Various simple linker-based PCR-based methods can be used to amplify neighbouring genomic DNA which can then be sequenced to map the mutation (see Methods). The nature of the mutation is determined by the "cargo" of the insertional element. If insertion occurs in an exon, although this is comparatively unlikely given the low proportion of exons in the genome, the element will disrupt genes. Natural or engineered promoters or enhancers in the cargo can increase gene expression or ectopically express

genes if the insertion is in an appropriate position. Loss of function mutations are also possible if the cargo contains a strong splice acceptor and the insertion occurs in the correct orientation in an intron.

Such splice acceptor constructs can be linked to a reporter gene to allow selection of insertions that express the reporter—these are known as gene trap constructs (Figure 1.2A,B; von Melchner and Ruley (1989); Gossler *et al.* (1989)). Gene traps are useful both for gene discovery and for mutagenesis. The general procedure is to transfect cells with a suitable vector, then select for the reporter gene. This selection step ensures that only insertions of the gene trap construct in genes in the appropriate orientation are isolated. Various international gene trap resources in ES cells have isolated mutations in more than 10,000 genes. Constructs with different cargoes can be used to expand the range of genes that can be trapped. For example, using the scheme above only genes expressed at the time of selection will be trapped, as expression of the reporter gene depends on trapping an active cellular promoter. Using a construct with its own promoter, but no polyadenylation (polyA) signal can trap genes that are not expressed at the time of mutagenesis (Figure 1.2C). Mutants isolated using these constructs tend to have insertions at the 3′ end of genes, so may not disrupt expression as reliably as promoter traps, which tend to be at the 5′ end. This can either be because of unstable reporter gene transcripts due to nonsense-mediated mRNA decay (Shigeoka *et al.*, 2005), or because a sufficient portion of the wild-type RNA is transcribed to form a functional protein.

Unlike deletion or substitution mutations, there is no loss of genetic information when making an insertion mutation. Mutations can therefore be designed to be revertible, by removing some or all of the inserted DNA. In a forward genetic screen, led by phenotype, both mutations caused by the insertion and naturally occurring background mutations will be picked up. By showing that the removal of the insertion rescues the phenotype, causality can be formally established. In some cases, the vector itself supports reversion (e.g. transposons, see below). In other cases, loxP sites can be incorporated to remove the cargo by Cre-mediated recombination. Although this leaves some sequence behind as a single copy of the target site, this is rarely sufficient to disrupt splicing as most insertions are in introns.

The fact that no information is lost in an insertion mutant can also be a disadvantage. It means that there there is potential leakiness, for example if the mutagen can be spliced out during transcription, restoring the wild type transcript (Voss *et al.*, 1998). Therefore, insertion mutagens need to have efficient splice acceptors to reduce the risk of this.

The choice of vector is another important factor in insertional mutagenesis. Retroviruses have been used with considerable success, and have the advantage of being easily introduced into a variety of cell types (Soriano *et al.*, 1991). Retroviruses enter the cell by binding to a surface receptor, and once inside the cell their genome is integrated into the host chromosomal DNA through the action of encoded enzymes. Retroviruses do exhibit strong site preferences for insertion however, with both hot and cold spots. From results of the gene trapping project, a limit is seen on recovery of new, non-redundant, insertions after around 100,000 clones have been screened (Skarnes *et al.*, 2004; Hansen *et al.*, 2008). In the resource described by Hansen *et al.*, a total of 10,433 genes are represented by over 350,000 clones. However, 2,793 of these are only represented by one gene trap clone, meaning that approximately 75% of the trapped genes are represented by larger numbers of redundant clones. Therefore, the coverage of the genome by retroviruses is uneven, with some genes being mutated at a relatively high frequency and others only rarely.

Results of screens carried out with libraries of mutants made using these retroviruses suggest that they do not completely cover the genome (Guo, 2004). As a result, various vectors have been used for gene trapping in an effort to expand coverage of the genome. ES cells are efficiently transfected, for example by electroporation, and a proportion of the transfected DNA will randomly integrate into the genome. Therefore it is possible to simply use plasmid DNA as a vector in cases where gene traps can be selected for. However, over the last decade efficient transposons for mammalian systems have been discovered and engineered, and these are quickly establishing themselves as the insertional mutagen of choice in mice and ES cells.

1.3.5 Transposons active in mammalian cells

DNA transposons of the cut-and-paste type are valuable reagents for insertional mutagenesis, particularly in bacteria and *Drosophila*. In their natural form, these transposons exist as two short repetitive DNA sequences that flank a gene encoding a transposase enzyme. When expressed, this enzyme recognises the transposon sequences, cuts the intervening sequence out of the chromosome and catalyses its reintegration elsewhere in the genome. The

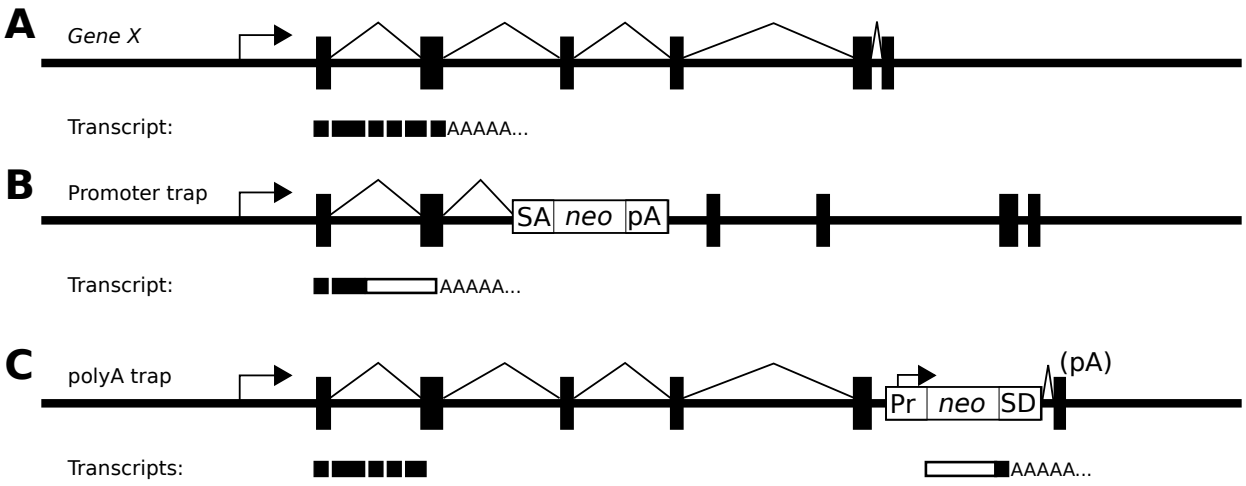


Figure 1.2: Types of gene trap vector. A—A hypothetical gene showing splicing pattern. Exons represented as black boxes. B—Promoter trap vector, consisting of splice acceptor (SA), reporter gene (*neo* in this case), and a polyadenylation signal (pA). C—polyA trap vector, with its own promoter (Pr) and splice donor (SD) splicing into the endogenous polyA site of gene X. A partial transcript from Gene X is produced, but is unlikely to be polyadenylated unless a cryptic site exists.

transposase gene is dispensable for transposition if the transposase enzyme is provided from another source. This allows transposons to be engineered for use as vectors in a similar way to retroviruses.

Although a large fraction of mammalian genomes is derived from transposable elements, none of these are known to still be active, with the possible exception of some L1 retrotransposons and the 'domesticated' RAG recombinase (Coufal *et al.*, 2009; Agrawal *et al.*, 1998). However, in the past ten years several transposons from other organisms have been shown to transpose effectively in mammalian cells.

Tol2

The only known active transposon that is naturally present in a vertebrate is *Tol2*. *Tol2* was isolated in an albino mutant of the medaka fish (*Oryzias latipes*) and is a member of the hAT family of transposons (Koga *et al.*, 1996). It has been extensively used for transgenesis in fish, and also shown to be active in mammalian cells, including mouse ES and germ cells (Kawakami and Noda, 2004; Keng *et al.*, 2009). Although efficiency in mammalian cells is reasonable, and the cargo capacity relatively high (at least 10 kbp; Balciunas *et al.* (2006)), the development of *Tol2* as a mammalian technology has not proceeded at the pace of the other transposons described below.

Sleeping Beauty

The genomes of salmonid fish contain a large number of inactivated transposable elements of the Tc1-Mariner family. By aligning these sequences, Ivics *et al.* deduced and synthesised the sequence of the ancestral transposon, which proved to be active not only in fish but also in mammalian cells. The 1.6 kbp element, which consists of two 250 bp terminal DNA elements containing inverted repeats (IRs) flanking an open reading frame encoding a transposase enzyme, was named Sleeping Beauty (SB).

SB duplicates its target site, a TA dinucleotide, upon insertion into the genome. Excision produces incompatible 3 nt overhangs, and therefore SB leaves a 'footprint' mutation for each round of transposition (Luo *et al.*, 1998). SB is active in mice and ES cells (Luo *et al.*, 1998; Dupuy *et al.*, 2001; Fischer *et al.*, 2001; Horie *et al.*, 2001). Constant improvements to the transposase enzyme are being made to compensate for the differences in codon usage and body temperature between fish and mammals (Mátés *et al.*, 2009). When the SB transposon is mobilised from an extrachromosomal plasmid in ES cells, it integrates at a wide range of genomic locations. However, when mobilised from a site on the chromosome, reintegration events occur preferentially at sites nearby. In one experiment using the *Hprt* locus, 25% of the recovered insertions were within 4 Mb of *Hprt* (Liang *et al.*, 2009). This effect has been called local hopping. Although a disad-

vantage in some situations, this property has been exploited for localised mutagenesis screens, in which SB is used to insert loxP sites near the transposon donor locus. These can then be used to make a series of nested deletions to study the requirements for sequences around the donor locus (Kokubu *et al.*, 2009).

Other interesting properties of SB include an increase in transposition efficiency when the donor DNA is methylated (Yusa *et al.*, 2004). SB appears to transpose in a variety of adult tissues and has been used as a mutagen in mice for cancer gene identification (Dupuy *et al.*, 2005). Some studies have looked for insertion preferences of SB beyond the TA target site. SB insertions do not appear to associate with genes (Liang *et al.*, 2009), but an association with a parameter predicting physical 'deformability' of DNA by proteins has been noted (Geurts *et al.*, 2006).

piggyBac

The piggyBac transposon (PB) is an active transposon isolated from the cabbage looper moth, *Trichoplusia ni* (Fraser *et al.*, 1996). PB was active without any further modifications in human and mouse cells (Ding *et al.*, 2005). Chromosomal excision of PB is more efficient than SB in the same setting (Wang *et al.*, 2008), although further improvements to both transposases are being developed (Mátés *et al.* (2009) and K. Yusa, unpublished). Methylation of the transposon reduces excision frequency (Wang *et al.*, 2008). Wang *et al.* also found that 95% of chromosomal PB excision sites were repaired accurately in ES cells. Thus, PB transposition will not generally leave footprint mutations. This has led to the use of PB as a tool for reversible introduction of transgenes, specifically the reprogramming (Yamanaka) factors required to produce induced pluripotent stem cells (Woltjen *et al.*, 2009; Yusa *et al.*, 2009; Takahashi and Yamanaka, 2006). Using PB to introduce the required transgenes means that stem cell lines with a 'clean' genome can be obtained after reprogramming.

PB inserts into a TTAA tetranucleotide. A weak preference for T 5′ of the TTAA and A on the 3′ side has also been described (Ding *et al.*, 2005). Around half of PB integrations occur in known genes, and there is a further enrichment of integrations in expressed genes (Ding *et al.*, 2005; Wang *et al.*, 2008; Liang *et al.*, 2009). The problem of local hopping, where a transposon mobilised from a chromosomal position reintegrates nearby, does not appear to be so severe for PB. No local hopping was observed in mobilisations from the *Hprt* locus in mouse ES cells, although 9% of the insertions were within 100 kbp for mobilisations from a reporter construct integrated at the *Rosa26* locus (Wang *et al.*, 2008). This difference is probably due to the relative sizes of the reporter loci, which must be fully reconstituted in order for transposition events to be recovered. The endogenous *Hprt* coding sequence spans 33.5 kbp, whereas the PGK-*puro* reporter gene used at *Rosa26* is smaller than 3 kbp. It is not known how many rounds of transposition may take place in these assays but it is likely that transposons proceed away from the donor locus by multiple rounds of excision and reintegration. If this is the case, the differences in local hopping between PB and SB could be explained by differences in the activity of the transposases.

PB has a cargo capacity of at least 9.1 kbp (Ding *et al.*, 2005), and therefore can be used to introduce large constructs carrying multiple transgenes. The transposase itself has been fused to other proteins for specialised applications. Adding a modified oestrogen receptor domain (ERT2) resulted in a transposase that can be induced by treatment with 4-hydroxytamoxifen (Cadiñanos and Bradley, 2007). A fusion with a *GAL4* DNA binding domain can be used to direct integrations to a chromosomally integrated UAS sequence (Maragathavally *et al.*, 2006).

1.3.6 Comparison of transposons

The properties of PB make it the ideal mutagen for ES cells (Table 1.1). Specifically, when compared to retroviral mutagens, PB has been shown to insert into genes that have not previously been mutated by retroviral gene traps (Wang *et al.*, 2009). The large cargo capacity means that design of mutagenesis constructs is not constrained by size requirements. Although PB is very efficient, this is a secondary consideration for ES cells, as generating large numbers of cells is not a problem. An especially valuable property of PB is its precise excision from the genome. This means that repeated transposition is unlikely to leave point mutations at loci that the transposon may 'visit' before it integrates at the site eventually observed. Such mutations could potentially cause background mutations in screens, where a mutant cell is identified but the mutation causing the phenotype is not due to the transposon. This leads to the another advantage of PB for screens—whether or not the transposon is causing the mutation can be easily tested by simply remobilising the transposon. This should rescue the phenotype if the transposon insertion causes it.

Mutations that do not revert are likely to be due to background mutations of an unknown nature, which are generally more difficult to map. These properties of PB make it ideal as a mutagen to use in genetic screens (Li *et al.*, 2010).

1.4 Genetic screens in embryonic stem cells

1.4.1 Practicality of genome-wide screens in mice

Despite improvements in mutagenesis, and the availability of the reference genome sequence to facilitate mapping, genetic screens in mice have remained something of a "cottage industry" (Kile and Hilton, 2005). The reason for this is simply the resources required to house and analyse sufficient mice to obtain enough mutants to screen a good portion of the genome. A notable recent exception is cancer gene discovery using insertion mutagens. This has the advantage that many loci can be sampled in a single mouse, with the resulting tumour acting as a simple device to clonally expand cells with the relevant mutation (Mattison *et al.*, 2009; Dupuy *et al.*, 2005; Collier *et al.*, 2005; Vassiliou *et al.*, 2010).

One solution to the problem could be to do genetic screens in ES cells. ES cells can easily be grown in quantities greater than the number of genes in the genome. Many aspects of mammalian cell biology can be accessed in ES cells, therefore such screens can still give useful functional information about mammalian genes. Given the goal of knocking out all genes in mice and making the mutants available as a public resource, the priority is to obtain information about gene function as a way to prioritise study of these mutants. I discuss below how ES cells can be used for genetic screens.

1.4.2 Suitability of embryonic stem cells as a model

Experiments using any cultured cell line are subject to caveats, as the cells are growing in an alien environment. It is well known that prolonged periods of culture can select for variants in the cell population that have a growth advantage. One characteristic of ES cells is that they maintain a relatively stable karyotype, although there is certainly potential for chromosome instability to arise (Liu *et al.*, 1997; Liang *et al.*, 2008). Many other cell lines used for experiments have severe aneuploidy and chromosomal instability, particularly those derived from tumours.

Unlike most cells, ES cells can be expanded infinitely in culture without large scale cell death or senescence. Most somatic cells will only replicate a limited number of times in culture, unless 'transformed' or 'immortalised', for example by an oncogenic virus (e.g. simian virus 40, SV40). Cell lines can often be established from primary tumours, but these are likely to have undergone a transformation-like change *in vivo*, and also to have other cancer hallmarks such as chromosome instability or mutator phenotypes. It is common to observe so-called 'crisis' events soon after the establishment of cell lines, where a large proportion of the culture dies or enters senescence, leaving only a few cells that recover (Sherr and DePinho, 2000). These are likely to be abnormal variants. This is not observed in the establishment of ES cell lines from blastocysts; thus ES cells are naturally immortal. Furthermore, the fact that ES cells can be reintroduced to blastocysts and contribute to normal development shows that ES cells are not irreversibly transformed, and that controlled growth can be re-established as part of normal development.

Multiple rounds of cell division in any cell causes problems, particularly at telomeres, the structures that cap chromosome ends (Blackburn, 1991). Every round of replication shortens the chromosome, as DNA synthesis does not proceed right to the end. This eventually results in chromosome instability and fusions between chromosomes once the protective telomere is eroded. Eventually a chromosome end is exposed, which can lead to chromosomal fusions, and cell death or senescence due to the DNA damage response (Counter *et al.*, 1992). Telomerase is a reverse transcriptase enzyme that can resynthesise telomeres, and is thus one way to solve this problem (Greider and Blackburn, 1987). Telomerase is active in human ES and iPS cells. In humans, telomerase is down-regulated during differentiation, and its reactivation is a hallmark of transformation or cancer (Hanahan and Weinberg, 2000). Telomerase is also active in mouse ES cells, although mice and other rodents appear to retain telomerase expression throughout adulthood, and thus generally have longer telomeres than humans (Forsyth *et al.*, 2002).

Another fact to bear in mind is that most ES cell lines used for making knockout mice are derived from male blastocysts. This is useful for obtaining germline transmission due to the greater breeding potential of male chimaeras made using the ES cells. It also means that most ES cell lines are XY, and thus only have a single gene dose of X chromosome genes along with genes unique to the Y chromo-

Mutagen	Coverage	Easy to map	Revertible	Cargo capacity	Footprints
Chemical	good	no	no	NA	NA
Irradiation	good	no	no	NA	NA
Retrovirus	uneven	yes	yes (Cre-loxP)	low	NA
SB	local hopping	yes	yes	low	yes
PB	gene bias	yes	yes	high	no

Table 1.1: Comparison of mutagens described in text. NA—not applicable.

some. Female ES cell lines do exist, and are pre-X inactivation—in fact they represent an excellent model for this phenomenon (Rastan and Robertson, 1985), but are not in general use for other applications.

It is well known that ES cells have an unusual cell cycle (Burdon *et al.*, 2002) . ES cells do not stop growing when confluent (contact inhibition) as fibroblasts and many other adherent cell lines do. ES cells have very low levels of D type (G1-specific) cyclins and Cdk4 is inactive (Savatier *et al.*, 1994). The G1 to S transition is controlled by the retinoblastoma protein (Rb), a Cdk4 phosphorylation target. ES cell proliferation is unaffected by knockout of all three Rb family members (Dannenberg *et al.*, 2000; Sage *et al.*, 2000). Thus ES cells lack the normal G1/S checkpoint.

Bearing in mind these differences, many pathways for normal cellular function are retained in ES cells. Some evidence for this is discussed in the context of genetic screens, below. ES cells express about 10,000 genes (Mikkelsen *et al.*, 2007). Furthermore, ES cells can be specifically differentiated into other cell types *in vitro* to access other aspects of biology. Particularly good protocols exist for differentiation into neural lineages, mesoderm and endothelium in bulk culture (Pollard *et al.*, 2006; Nishikawa *et al.*, 1998). Many other lineages are accessible through the formation of embryoid bodies—cystic aggregates formed by suspension culture of ES cells, which resemble the early embryo. Thus, any phenotype observed in ES cells can be easily investigated in differentiated cell types.

It could be argued that all cell lines are abnormal, as they do not grow under physiological conditions of matrix attachment, blood supply and so on. Alternatively, it could be said that ES cells are abnormal as they represent a unique and very specialised cell type that is not typical of most cells in the body. ES cells at least have the advantage of being very well studied, so some of their unusual features are well-documented.

It should be noted that the discussion above concerns mouse ES cells. Human ES cells, and more recently iPS cells, have been derived and in principle represent a better model for human biology. The reason that mouse ES cells remain an attractive model system is the availability of a well-developed genetic toolkit, and the constant genetic background guaranteed by the use of inbred strains. Gene targeting by homologous recombination in particular is not well developed in human cells, due to the requirement for isogenicity discussed above. Zinc finger nucleases, which can be designed to induce breaks at defined loci, are being developed as an alternative technology (Kim *et al.*, 1996; Porteus and Baltimore, 2003). The experiments described in this thesis could be extended to human cells in principle, but depended heavily on gene targeting and thus were carried out in mouse ES cells. In the following section I discuss the wide range of mouse genetic 'tricks' available that make ES cells useful for genetic screens.

1.4.3 Dominant and recessive screens

Mutations, and the screens in which they are generated and analysed, can be broadly classified as dominant or recessive.

Dominant screens

The definition of a dominant mutation is a mutation that affects phenotype even in the presence of a wild-type allele. This could include ectopic or increased expression of the wild-type gene. Alleles of this type can be generated by mutations in promoter regions, introduction of strong promoters or enhancers into endogenous loci, or by simply expressing cDNAs from a strong promoter. Dominant alleles involving coding sequence changes could be point mutations that increase enzyme activity, deletions of negative regulatory regions or disruption of homodimerisation domains of the protein.

Dominant screens are the most technically straightforward. By definition, only one round of mutagenesis is required and the resulting mutants can be im-

mediately assayed for phenotype. A common example of a dominant screen is cDNA cloning, in which a large pool of cDNAs is transfected into cells. Usually this is used where a cDNA would be expected to confer a phenotype that can be selected for, such as resistance to radiation or a drug. An example of this approach in ES cells is the identification of *Nanog* as a regulator of pluripotency (Chambers *et al.*, 2003). Introducing ectopic promoters by insertional mutagenesis is another example, as in the oncogene discovery screens mentioned above. This has also been applied in ES cells (Kong *et al.*, 2010; Bouwman *et al.*, 2010).

Recessive screens

A recessive mutant is a mutation that can be compensated for by the wild type allele. Such mutations usually disrupt or abolish normal expression of the gene. Recessive screens are more challenging because most model organisms are diploid, therefore in a random mutagenesis experiment most mutants will still have an intact wild type allele of the mutated gene. These are unlikely to show a strong loss-of-function phenotype, except in rare cases where the other allele is epigenetically inactivated. In many model organisms this can be circumvented by intercrossing mutants to obtain homozygotes, however this is a major undertaking in a mammal for a genome-wide screen. ES cells cannot be bred to homozygosity as such, but there are other ways of obtaining homozygous mutants. I have outlined these below, with reference to their scaleability to a genome-wide screen. However, I will first discuss several other systems that can be used for studying loss-of-function phenotypes in mammalian cells.

Chinese hamster ovary cells

An ovarian cell line from the Chinese hamster *Cricetulus griseus* has been extensively used, particularly for protein production for biochemistry, but also in early cytogenetics where it was attractive due to its low chromosome number ($2n = 11$, Tjio and Puck (1958)). However, it also proved easy to isolate recessive mutations for certain autosomal loci, such as *Tk* and *Aprt*, at frequencies similar to those expected for single copy genes (Siminovitch, 1976). Chinese Hamster ovary (CHO) cells are functionally hemizygous for large regions of the genome, either due to large deletions or epigenetic silencing of one copy of some genes (Holliday and Ho, 2002). Although some domains of hemizygosity have been mapped, particularly those surrounding isolated mutants (for example on Chinese hamster chromosome 9), the extent of hemizygosity is unknown. Thus the exact proportion of the genome available for recessive screens in these cells is unknown. Screens in CHO cells, mainly using EMS mutagenesis, have been particularly well applied in the field of DNA repair. Several lines sensitive to UV or ionising radiation were isolated in the early 1980s, assigned to complementation groups by somatic cell hybridisation and the genes responsible eventually identified by cDNA cloning (Thompson *et al.*, 1980; Busch *et al.*, 1980; Jeggo and Kemp, 1983; Thompson, 1998). These screens identified a number of key players in the DNA damage response: the excision repair cross-complementing (*Ercc*) series of genes and the X-ray sensitivity cross complementing (*Xrcc*) series.

Although CHO screens have been productive, the difficulty of cloning mutations and the lack of a complete genome sequence or reverse genetic technology makes them less attractive for new screens. The cells themselves are also unusual, and the lack of definition in the hemizygous region means that screens are not truly genome-wide.

RNA interference

The first indication that RNA could regulate gene expression came from studies of silencing of genes after viral infection in plants, which was shown to be associated with production of small RNAs (Hamilton and Baulcombe, 1999). These small RNAs had complementarity to the silenced genes. The first demonstration in animals, where the effect was named RNA interference (RNAi), was in *C. elegans*. Introduction of double-stranded RNA into cells in catalytic amounts silenced translation of the corresponding gene (Fire *et al.*, 1998). Studies on *C. elegans* mutants also helped to define the mechanism, in which the double-stranded RNA is cleaved into smaller 21-nt effector molecules, which are then used to confer specificity to the RNA-induced silencing complex (RISC). This binds and cleaves or prevents translation of the target mRNA (Novina and Sharp, 2004).

C. elegans possess connections between cells, meaning that RNAi actually has a systemic effect (Winston *et al.*, 2002). This means that RNAi is an excellent tool for screens in *C. elegans*, particularly as the effect can be produced simply by feeding animals on bacteria engineered to express the double stranded RNA. Thus, even though conventional forward genetics in *C. elegans* is well developed, RNAi

screens have been widely used due to the relative technical ease (Fraser *et al.*, 2000; Kamath *et al.*, 2003).

Extending the technique to mammalian cells was more problematic, as introduction of double stranded RNA induces an innate immune response. This can be overcome by pre-synthesising the short 21 nt effector molecules, and transfecting them directly (Elbashir *et al.*, 2001). These are termed short interfering RNAs (siRNAs). While specificity is generally good in *C. elegans*, where a long dsRNA can be processed into multiple effector molecules, this advantage is not available when using a single siRNA. More recent approaches transfect pools of siRNAs, typically four, targeting the same gene. However, suppression of translation is often incomplete, and in the cases of pooled siRNAs it is typical that only one or two are effective. While this may be still be sufficient to see a knockout phenotype, there is a further problem of specificity. siRNAs have been shown to have significant 'off-target effects', due to homology with other transcripts other than the intended target (Jackson *et al.*, 2003). In some screens, even very strong hits have been shown to be due to off-target effects. In fact, it may be possible to rationalise these based on analysis of the 'seed' region (nucleotides 2–8 of the siRNA) of the siRNA sequences that give hits, as these often have complementarity to the real target (Lin *et al.*, 2007; Sudbery *et al.*, 2010).

Screens in mammalian cells using siRNA offer huge promise if the problems above can be overcome. Synthesis of siRNAs was expensive initially, but DNA constructs can now be used that express a short hairpin RNA (shRNA), which is processed into a single stranded siRNA by the cell. As a technique for study of single genes, or small sets of genes, where knockdown can be optimised and the potential for false positives is low, siRNA has been a very useful approach, allowing analysis of loss of a gene of interest in a very short time, and in human cells. siRNA screens have also been applied on a genome wide scale. In this case, it is typical to find hundreds or thousands of siRNAs showing a phenotype ('hits'). These typically include siRNAs targeting several genes expected to show a knockdown phenotype, but identifying new genes involves extensive secondary screens and statistical analysis. This is likely to be a combined effect of highly variable knockdown and transfection efficiency and off-target effects. The fact that knockdown is often incomplete (and not measurable in a general way, as antibodies to each protein would be required) precludes setting of overly stringent statistical thresholds, leading to a large number of false positives from off-target effects.

Several high profile siRNAs and shRNA screens have recently been published, and studying the results of these shows the strengths and weaknesses of the method. Identification of host cell factors required for infection by pathogens is an area of great interest, and several groups have conducted screens for viral infection. Three groups published genome-wide screens for siRNAs conferring resistance to HIV, for example (König *et al.*, 2008; Brass *et al.*, 2008; Zhou *et al.*, 2008). Each identified hundreds of siRNAs affecting infection, but in each case, most of these were not shared between the other screens—the Brass screen had only 13 and 15 hits in common with the König and Zhou screens respectively (Goff, 2008). Differences in cell type, endpoint and other experimental conditions can account for some of these, but many hits could turn out to be false positives due to off-target effects. Furthermore, in each case a series of filters was applied to reduce the initial number of hits, which numbered around 2,000 in each case. This used prior information to determine likely hits, for example siRNAs targeting pathways already associated with the virus, or expression of the targets in T cells (the *in vivo* target of HIV). By taking this approach, the ability of these screens to identify completely novel factors is compromised, unless the knockdown is very good and the effect very large.

The true value of genome wide siRNA screens will be apparent once the hits have been investigated more thoroughly. As the link between siRNA sequence and gene is only a prediction, and there may be unanticipated other targets, it is important to carry out functional rescue experiments, such as rescue of the knockdown phenotype by expression of a cDNA with a 3′ UTR that does not have a binding site for the siRNA. In fact this was only carried out in one of the above papers, and only for a subset of nine attractive drug targets, only four of which confirmed this important gene-phenotype link (Zhou *et al.*, 2008). The results of genome-wide siRNA screens represent a useful starting point for further analysis, but require proper confirmation before reaching firm conclusions (Bushman *et al.*, 2009). False negatives (where expected genes are not found) are another problem that certainly exists, for example, a known HIV cofactor (LEDGF) was not picked up in any of the screens above.

It should be noted that whole genome siRNA screens have had successes in cases where individual hits have been followed up and confirmed, for example from two screens for modulators of the DNA

damage response (Smogorzewska *et al.*, 2010; Kolas *et al.*, 2007). The effort required to conduct genome wide screens is considerable using current methods, and is unlikely to be widely available to individual investigators interested in specific questions of basic biology, as yeast genetic screens currently are. siRNA screens represent the best available method for large scale gene function analysis, despite their drawbacks.

In principle RNAi represents almost the ideal mutagenesis strategy, in which it is possible to knock a gene out using only a short, easily synthesisable, length of DNA to confer specificity. The shortcomings are the off-target effects, and the weak link between genotype and phenotype. RNAi is also not a genuine forward genetic approach, and is more properly thought of as reverse genetics on a large scale (see section 1.4.4).

Haploid cell lines

Recently two studies have described haploid cell lines from normally diploid organisms, which may also be of use for recessive genetic screens. One is a medaka ES cell line (Yi *et al.*, 2009). The other is a human leukaemia cell line, haploid for all chromosomes except chromosome eight (Carette *et al.*, 2009). These were successfully used to identify mutants resistant to influenza infection and bacterial toxins. Although full details of screens have not yet been published, these cells represent an attractive system for studying loss-of-function mutants, despite the fact that the cells are clearly abnormal. This study underlines the limitation imposed on screens by the diploid mammalian genome, and shows the possibilities for annotation of gene function if this can be circumvented.

1.4.4 Making homozygous mutations in ES cells

Serial gene targeting

The International Knockout Mouse Consortium aims to produce a publically-available collection of mouse knockouts in every gene (International Mouse Knockout Consortium *et al.*, 2007). At the time of writing (September 2010), 17,753 targeting vectors had been generated and 10,230 heterozygous knockout ES cell lines produced[2]. Therefore, obtaining gene targeted ES cells is more straightforward than in the past. Moreover, the targeting vector resource is adaptable to the use of different selectable markers, or recycling of the original one, for a second round of gene targeting. Thus, it should be possible to produce libraries of ES cells with null mutations in known genes using this resource. These vectors result in conditional deletion mutants, in which a critical exon is deleted after expression of a site-specific recombinase. Therefore, they are likely to cause robust null mutations. In the future, all genes may be knocked out homozygously in the resource. Until then, sub-genomic libraries can be generated by investigators performing second round targeting for a subset of genes of interest. This still requires considerable effort, but the availability of validated targeting vectors should greatly ease the process.

This approach is not a genuine forward genetic approach, as all mutations are known to begin with. In this respect, serial gene targeting has the same drawbacks as siRNA screens, although the mutagenesis is much more robust for targeted alleles. The ability to do large scale reverse genetics blurs the boundaries of the traditional genetic approaches. However, it also means that by definition only known genes, and only the designed mutations in those genes, can be accessed by targeted libraries. A strength of forward genetics is that completely unexpected genetic elements can be identified—the discovery of animal microRNAs via the *lin-4* mutant in *C. elegans* is one famous example (Lee *et al.*, 1993).

Loss of heterozygosity

Another way to generate homozygous mutants for recessive screens would be to make random heterozygous mutations, and somehow convert these to homozygosity. A number of events can lead to loss of heterozygosity (LOH) in cells. LOH is used to describe the situation where one allele of a heterozygous locus or region is lost. LOH can affect single loci, large chromosome regions or entire chromosomes. A number of events can lead to LOH.

LOH at a single locus could occur by gene conversion (Figure 1.3A). This can happen as an outcome of the homologous recombination (HR) pathway, which is involved in the repair of DNA double strand breaks that occur in S and G2 phases of the cell cycle. Usually the recently-replicated sister chromatid would be used as a template to copy sequence information from—this would result in accurate, conservative repair. However, in rare cases the homologous chromosome could be used, and any sequence variants specific to that chromosome would be copied to the repaired molecule (Moynahan and Jasin, 1997). Thus, the original variants on the re-

[2] http://www.knockoutmouse.org

paired chromosome will be lost. The cell is now a homozygous mutant for any mutations encompassed by the synthesis occurring during repair. This type of event is very rare in ES cells—even when a double strand break is artificially induced at a specific locus, the frequency of LOH is just one per 10^6 cells (Moynahan and Jasin, 1997). In this case the selection scheme required the modification of both alleles; thus this method is not generally applicable to random mutagenesis, where only one allele can be modified to begin with.

Other events during cell division can lead to LOH across larger regions, or entire chromosomes. Several studies have measured LOH in various cell types using selectable autosomal loci. Thymidine kinase (*Tk1*) and Adenine phosphoribosyltransferase (*Aprt*) are commonly used for this purpose, as homozygous loss-of-function mutants are selectable in each case, using toxic thymine or adenine analogues respectively. Other loci can be investigated by insertion of a mutant *neo* gene and selection in very high concentrations of G418 (high [G418], Mortensen *et al.* (1992)). By isolating homozygous mutants from heterozygous starting populations, the mechanism of LOH can be examined by looking at polymorphisms linked to the selectable locus (in F_1 hybrid ES cells). Three categories of LOH event are generally detected in such experiments: No change in the flanking markers, homozygosity of all linked markers, or homozygosity of a subset of markers from some point between the centromere and the selectable locus, often all the way to the telomere (Lefebvre *et al.*, 2001; Cervantes *et al.*, 2002).

Clones with no change in flanking markers have usually acquired a 'second-hit' spontaneous mutation in the wild-type copy of the gene. This category can only be observed using loss-of-function systems, and therefore not using the high [G418] method. The cases in which all markers on the chromosome in question are homozygous can be interpreted as loss of the entire chromosome bearing the wild type allele, with a duplication of the chromosome with the mutant allele. It is likely that this proceeds through a trisomic intermediate cell, as monosomic cells are very rarely observed. This outcome is referred to as uniparental disomy (UPD), as both copies of the chromosome are now derived from a single parent and are identical to each other. Finally, the cases where only distal markers become homozygous can be explained by a mitotic recombination event followed by crossover.

All of these events are rare in ES cells; in particular the rate of spontaneous mutation is very low ($< 10^{-9}$ events/cell/generation at *Hprt*, although mutations are more readily detected at *Aprt*). A study of extensive LOH events at *Aprt* in ES cells found a rate of the order of 10^{-7} events/cell/generation (Cervantes *et al.*, 2002). The proportion of mitotic recombination events was 41%, compared to 57% UPD. These events represent a way to generate homozygous mutants from a starting population of heterozygotes. However, the rate is very low. Several approaches have been taken to increase the frequency, particularly focusing on mitotic recombination events.

Induced mitotic recombination

In mitosis, homologous recombination (HR) is induced as a response to DNA damage. Unlike HR in meiosis, the homologous chromosome is rarely used as the template for repair. Mitotic HR occurs mainly in S and G2 phases of the cell cycle, therefore a sister chromatid is available and is the preferred template for repair (Johnson and Jasin, 2000). HR in mitosis and meiosis also differs in the regulation of crossing over, the process by which homologous sequences on either side of the repair site are exchanged between maternal and paternal chromosomes. There is at least one obligate crossover per chromosome during meiosis, which helps to generate genetic diversity among gametes. In contrast, crossing over is suppressed during mitotic recombination (see below).

There are several known recombinase enzymes that are sufficient to recombine two specific sequences, with crossover. The most widely used of these in mouse is the Cre recombinase of bacteriophage P1 (Sternberg and Hamilton, 1981). Cre catalyses recombination between 34 bp loxP elements, and always induces crossing over of the flanking sequences. Strategic positioning of loxP sites in the genome can be used to generate large rearrangements not possible by gene targeting alone. LoxP sites have an orientation, defined by an 8 bp spacer element at the center of the site. Positioning two loxP sites on a chromosome in the same orientation will delete the intervening sequence when Cre recombinase is expressed in G1 phase, leaving a single loxP site. The intervening sequence is excised as a closed circle containing a single loxP site. Alternatively, loxP sites in opposite orientations can be used to reversibly invert the sequence that they flank. The two sites can also be placed on different chromosomes. If oriented in the same direction relative to their respective centromeres, the action of Cre will produce a balanced translocation. Cre recombination is very efficient over distances of up to a few kbp, and can

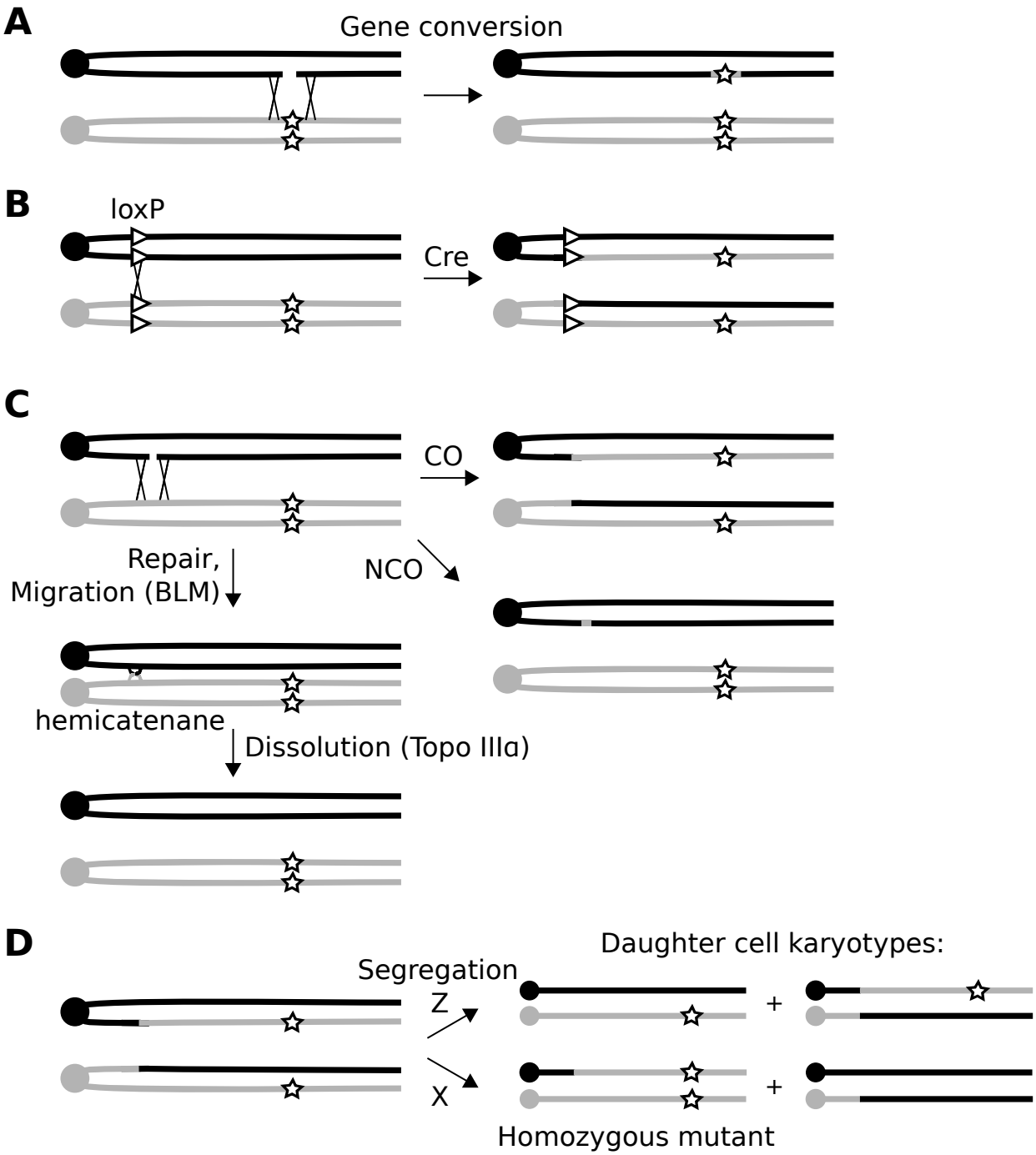


Figure 1.3: Mitotic recombination leading to LOH in heterozygous cells. A—Gene conversion, B—Induced mitotic recombination, C—Mitotic recombination in *Blm*-deficient cells. Homologous chromosomes are indicated in black and grey, and are shown after replication, so consist of a pair of sister chromatids. CO—crossover outcome, NCO—noncrossover outcome. D—segregation of recombinant chromatids to different daughter cells (X segregation) can produce a homozygous mutant.

still occur at a frequency of around 10% up to 1 Mbp, but selection for the recombination product is necessary for long distances or between chromosomes (Ramírez-Solis *et al.*, 1995).

Site-specific mitotic recombination has been used in *Drosophila* for generation of mosaics to study cell fate. Mitotic recombination in G2 phase in *Drosophila* cells affects segregation of the recombinant chromatids. After induction of recombination by the FLP recombinase, the recombinant chromatids segregate to different daughter cells (this is termed X segregation). This is the outcome necessary to generate a wild type and homozygous mutant in the daughter cells, instead of two heterozygotes. This effect is likely to be a result of spatial constraints imposed by the tight pairing of sister chromatids and the recombination event (Beumer *et al.*, 1998). If a heterozygous pigmentation mutant is used, for example, one of the cells segregated after LOH will become homozygous for the mutation and lack pigmentation. This can be used for fate mapping, as this cell will give rise to a clone of unpigmented cells (Xu and Rubin, 1993).

This technique has been extended to mouse ES cells using the Cre/loxP system, with loxP sites targeted to allelic positions on homologous chromosomes. Strong selection is necessary to isolate recombinant cells. Both high [G418] and a scheme that reconstitutes an active *HPRT* gene on recombination have been used for this purpose (Koike *et al.*, 2002; Liu *et al.*, 2002). It appears that at least at some loci, a bias towards X segregation after recombination also applies in mice (Liu *et al.*, 2002). In the best case from these experiments, a frequency of 1/20 cells was obtained, although this varied by locus and the number of loxP sites (or variants thereof) introduced. This method could be used to convert heterozygous mutations on a specified chromosome to homozygosity. Targeting of loxP sites to centromeric regions of both homologous chromosomes would result in an easy system to isolate LOH events at any distal locus on that chromosome (Figure 1.3B).

The drawback of using this method to generate genome-wide collections of mutants is that a centromeric locus with high recombination efficiency needs to be identified, and an appropriate cell line constructed, for each chromosome (except X and Y). Also, a suitable selection scheme would need to be used, as selection for the recombination event using the separated *HPRT* gene used by Liu *et al.* does not guarantee selection for the homozygous mutant daughter cell (as opposed to the homozygous wild type). Koike *et al.* did select directly for the homozygous cell using high [G418], but this selection is rarely complete, as it depends on the base level of *neo* expression, which varies at different loci. Thus high [G418] selection is useful on a small scale where conditions can be titrated for a specific locus, but is not a suitable selection strategy in a genome-wide context.

To extend the use of LOH via mitotic recombination to the whole genome, a mechanism to increase the frequency of recombination and crossover across the whole genome is required. This is known to be a property of cells from patients with a rare cancer-prone condition, Bloom's syndrome. In the following section I describe the biology of Bloom's syndrome and its associated gene *BLM*, and discuss the use of *Blm*-deficient mouse ES cells for generating homozygous mutants.

1.4.5 Biology of cells with mutations in the *BLM* gene

Bloom's syndrome

Bloom's syndrome is a rare condition, mainly prevalent among the small population of Ashkenazi Jews. The symptoms include small stature and growth defects, telangiectasia (dilation of surface blood vessels), light-sensitivity and a susceptibility to different forms of cancer (Bloom, 1966). Bloom's syndrome also has a distinctive cytogenetic phenotype—an increased frequency of sister chromatid exchanges (SCEs, Chaganti *et al.* (1974)). SCEs are points of crossover between sister chromatids generated during S or G2 phase. SCEs are measured by a cell culture assay, in which cells are grown for two generations in the presence of radiolabelled deoxythymidine, or an analogue such as bromodeoxyuridine (BrdU, Pinkel *et al.* (1985)). After the first round of DNA synthesis, each sister chromatid has one strand labelled in approximately equal amounts. After division and a second round of synthesis, one chromatid will have both strands labelled while the other, which was synthesised from the unlabelled template, will have only one labelled strand. Thus, the sister chromatids can be distinguished, and any exchanges of DNA between them can be seen by a switch from light to dark staining at a distinct point on the chromatid.

SCEs clearly represent the outcome of crossing over, but an increase in SCEs does not necessarily mean an increase in the liklihood of crossing over occurring. SCEs are increased by treatment with a variety of mutagens, particularly those that cause single stranded breaks. A single strand break en-

countered during replication is converted to a double strand break, which can be repaired by HR using the sister chromatid (Wilson and Thompson, 2007). Thus, a general increase in damage repaired by HR can also lead to increased SCEs. It is the proportion of repair events that result in crossover that is of interest in the context of LOH. Furthermore these must be interchromosomal events, rather than sister chromatid exchanges. Therefore an increase in SCEs does not necessarily indicate increased LOH unless the mechanism is also applicable to crossovers after interchromosomal recombination. For example, cells with a homozygous mutation in the *Recql5* gene have an increase in SCE but not LOH (Hu *et al.*, 2005, 2007).

In lymphocytes from Bloom's syndrome patients, where the increase in SCEs was first observed, there were also indications that the Bloom's syndrome defect did lead to increased crossing over, and that this could apply to interchromosomal events. In some patients, a small subpopulation of lymphocytes showed normal SCE levels. These patients turned out to be compound heterozygotes for the mutant *BLM* gene, having inherited a different *BLM* allele from each parent. Recombination between the *BLM* genes on the homologous chromosomes had reconstituted a functional *BLM* gene in this subpopulation (Ellis *et al.*, 1995b). This remarkable event actually assisted in mapping the *BLM* gene to chromosome 15q and cloning its cDNA (Ellis *et al.*, 1995a). The resulting sequence indicated that BLM was homologous to the RecQ helicase of *E. coli*.

Molecular biology of Bloom's syndrome

It is now apparent that *BLM* is a member of a group of RecQ paralogues in eukaryotes (Hickson, 2003). The *E. coli recQ* mutant was initially identified as a component of the recF recombination pathway, and was shown to be an ATP-dependent 3′ to 5′ DNA helicase *in vitro* (Nakayama *et al.*, 1984; Umezu *et al.*, 1990). The budding yeast (*S. cerevisiae*) homologue, *sgs1* (slow growth suppressor), was identified independently of studies of Bloom's syndrome as a suppressor of the growth defects in strains with mutations in *top3a*, which encodes DNA topoisomerase IIIα (Gangloff *et al.*, 1994). Indeed, Sgs1p interacts with topoisomerase IIIα, and the mammalian homologues also form a complex, along with two other proteins, RMI1 and RMI2 (Wu *et al.*, 2000; Singh *et al.*, 2008; Xu *et al.*, 2008).

It is this complex that carries out the best understood function of BLM, which is likely to be responsible for the increase in SCEs in *BLM* mutants. Using purified proteins, it was shown *in vitro* that BLM could cause unwinding of several DNA structures (Sun *et al.*, 1998; Karow *et al.*, 2000). BLM showed a preference for binding a synthetic version of a DNA recombination intermediate called a Holliday junction (Karow *et al.*, 2000).

Holliday junctions are four-stranded DNA structures formed at the point of strand transfer between two homologous duplexes. A Holliday junction is formed during repair by HR, when a single strand from the broken molecule invades the homologous template with the assistance of the Rad51 protein, which forms a filament on the single stranded DNA. As the sequences adjacent to the junction are homologous, the junction point can migrate by unwinding two of the duplexes and rehybridising the opposing duplexes. This migration is catalysed by BLM (Karow *et al.*, 2000). Single Holliday junctions are formed from single-ended breaks, such as those that occur when a replication fork hits a single strand nick. Resolution of these junctions to restart replication can result in template switching, which produces the observed SCEs (see Wilson and Thompson (2007) and Mankouri and Hickson (2007) for a discussion of this mechanism). It has been proposed that BLM could act to migrate the junction in the reverse direction, to allow the nick to be repaired and replication to continue without formation of a double strand break (Karow *et al.*, 2000).

Repair of a double strand break with two free ends, both of which invade the homologous duplex, will form two separate Holliday junctions, which are referred to as a double Holliday junction (dHJ) once repair synthesis and ligation has taken place (Figure 1.4A). BLM also catalyses migration of HJs in this situation. When the two HJs collide, a special DNA structure called a hemicatenane is formed. This consists of two almost complete duplexes, with a minimal strand exchange region where the exchanged strands simply loop over each other. In *in vitro* experiments this structure, formed from a synthetic dHJ, is a substrate for Topo IIIα which separates the two duplexes, a process stimulated by BLM (Wu and Hickson, 2003). This is termed HJ dissolution.

Importantly, dissolution of dHJs in this way can only produce noncrossover products (Figure 1.4B). Several other pathways exist to resolve (distinct from dissolve) HJs by endonucleolytic cleavage. The first to be discovered in mammalian cells was the MUS81-EME1 complex (Blais *et al.*, 2004), which is responsible for generating crossovers in meiosis but also acts in mitosis. More recently, the GEN1 protein was identified as being responsible for a previously

characterised resolvase activity in mammalian cell extracts (Constantinou *et al.*, 2002; Ip *et al.*, 2008). Finally, several groups identified a SLX4-containing complex possessing HJ resolution activity (Fekairi *et al.*, 2009; Muñoz *et al.*, 2009; Andersen *et al.*, 2009; Svendsen *et al.*, 2009). All of these nucleases cleave two strands in HJ, which are then religated to resolve the two duplexes. A dHJ is resolved by two independent cleavages. Depending on the relative orientation of the two cleavages, this can result in a crossover product (Figure 1.4B). Thus, in the absence of BLM, one of these nucleolytic pathways must resolve dHJs, which has the potential to result in crossing over (Figure 1.3C).

BLM has several other roles in regulation of recombination. It has been shown that BLM can disrupt Rad51-ssDNA filaments *in vitro*, which may function to divert double strand breaks to pathways other than HR that do not result in crossover (for example nonhomologous end joining or single strand annealing, see Chapter 7 and Wu *et al.* (2001); Bugreev *et al.* (2007); Krejci *et al.* (2003)). Thus BLM deficiency may also result in more breaks being repaired by HR in the first place, as well as a higher rate of crossover later in the process. BLM also forms a complex with DNA exonuclease I (ExoI), which mediates the early resection of DNA ends that is the beginning of the HR pathway (Gravel *et al.*, 2008; Nimonkar *et al.*, 2008). Thus BLM is involved in the regulation of HR at several stages, both positively and negatively. The BLM complex interacts, via RMI1 and FANCM, with the Fanconi anaemia complex which mediates repair of interstrand crosslinks, a complex lesion requiring several steps to repair (Deans and West, 2009). BLM may also have a role during anaphase. It has been shown that ultra-fine bridges of DNA that connect separated chromatids at anaphase are coated in BLM protein. These bridges link fragile sites and centromeres in particular. It is possible that BLM is required to decatenate tangled chromatids to allow complete separation at anaphase, which could explain the chromosomal instability observed in BLM-deficient cells (Chan *et al.*, 2007, 2009).

Mouse models of Bloom's syndrome

Although many human alleles of *BLM* are predicted to be null, homozygous knockout of the mouse homologue, *Blm*, resulted in embryonic lethality (Chester *et al.*, 1998). Homozygous embryos could be recovered, and were smaller than heterozygotes, possibly mirroring the Bloom's syndrome growth defects. Fibroblasts from homozygotes did show the expected high frequency of SCE.

Another mouse model used an allele derived from a complex insertion of the targeting vector, which resulted in a duplication of exon three, after the selection cassette and vector backbone were removed by Cre/loxP recombination. Mice homozygous for this allele were susceptible to multiple cancer types (Luo *et al.*, 2000). This mutation also accelerated the onset of colon cancer in the $Apc^{Min/+}$ mouse model, in which LOH at the *Apc* locus is commonly observed (Moser *et al.*, 1990). Cross breeding the two Bloom's syndrome mouse models suggests that the exon three duplication allele is actually a hypomorph (McDaniel *et al.*, 2003); however these mice appear to represent a good model for Bloom's syndrome. Another specifically modelled the mutant allele found in the Ashkenazi population by deleting exons 10, 11 and 12, replacing them with an *HPRT* minigene. Homozygosity for this allele also caused embryonic lethality, but the heterozygotes showed accelerated T cell lymphoma formation and, on an $Apc^{Min/+}$ background, increased numbers of intestinal tumours (Goss *et al.*, 2002).

Homozygous ES cells were also constructed, with one allele having a genuine deletion of *Blm* exon two, and one having the duplication described above. These ES cells showed an increased rate of SCE and LOH. As described above, LOH can lead to the generation of a homozygous mutant from a heterozygous starting cell. Therefore, there was interest in applying these cells to convert random heterozygous mutations to homozygosity for use in genetic screens.

Genetic screens using *Blm*-deficient ES cells

The first genetic screens using *Blm*-deficient ES cells were published in 2004. Using the cell line described above, recessive mutations in the DNA mismatch repair pathway were isolated by selecting for resistance to 6-TG in *Hprt*-positive cells (Guo *et al.*, 2004). A retroviral gene trap vector was used as a mutagen, and mutants were recovered with insertions in the known mismatch repair genes *Msh2* and *Msh6*. *Dnmt1*, a *de novo* DNA methyltransferase was also recovered and identified as a mismatch repair gene.

Another group generated a new *Blm* allele, making use of the tet-off system to temporarily suppress Blm expression (Hayakawa *et al.*, 2006; Yusa *et al.*, 2004). This has the advantage that Blm expression can be reactivated after homozygous mutants have been generated. This reduces the risk of genome instability associated with mutations in *Blm*, and also

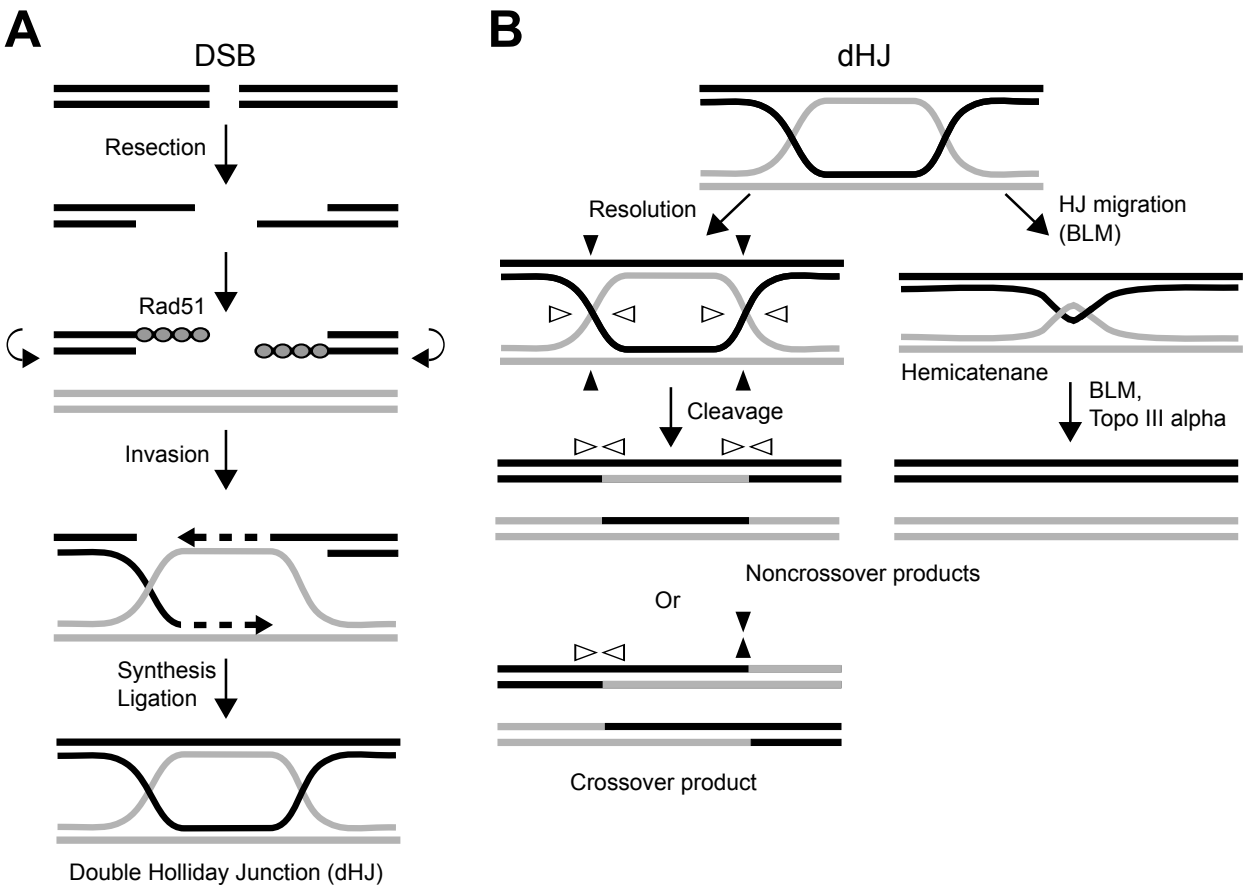


Figure 1.4: Formation and resolution of double Holliday junctions. A—Formation of a dHJ during double strand break (DSB) repair by homologous recombination. B—Pathways for resolution by structure specific endonucleases or dissolution by BLM/Topo IIIα. Products of cleavage in the orientations indicated by arrowheads are shown. Symmetrical cleavages are shown, but MUS81-EME1 cleaves asymmetrically.

ensures that any phenotype identified is relevant on a wild type background and does not interact with *Blm*. The published screen looked for mutations in the glycosylphosphatidylinositol (GPI) anchor synthesis pathway. Cells lacking GPI anchored proteins can be selected for using aerolysin. The study identified 12 out of 23 of the known genes involved in GPI anchor synthesis. Mutagenesis in this case used ENU—therefore mutations were mapped by cDNA complementation. The cell line used is a F1 hybrid (129 × C57BL/6), so polymorphisms between these strains could also be used to map mutations by crossover position.

Three other screens using *Blm*-deficient cells have since been published. A library of mutants (generated with a retroviral mutagen) was infected with a retrovirus to identify components of the infection pathway. This identified the receptor for the virus (Wang and Bradley, 2007). Another mismatch repair screen was also published, this time using piggyBac as the mutagen—this screen identified all the previously known components of the pathway (Wang *et al.*, 2008). Finally, a reporter gene approach was used to identify components of the RNA interference pathway (Trombly *et al.*, 2009). This used a cell line that contained a synthetic short hairpin RNA that suppresses expression of a reporter gene (*Hprt*). Selection for mutations that restored expression of *Hprt* isolated several mutations in the *Ago2* gene, which encodes a component of the RNAi processing pathway.

All the published screens so far have investigated a phenotype that is selectable, either directly or via a reporter construct (Figure 1.5). Thus they are not screens in the strict sense of the word, which would involve examining each mutant individually, and should be properly referred to as selections (Grimm, 2004). The reason for this is that the frequency of 'useful' cells, i.e. homozygous mutants, in cultures of *Blm*-deficient cells is still extremely low. Each homozygous mutant is likely to be outnumbered by thousands of its heterozygous progenitors, and the vast majority of the mutants in the culture will be irrelevant to the phenotype being selected for. Therefore an 'interesting' mutant cell could be literally one in a million, and very strong selection for the mutant phenotype is required to isolate such mutants. This requirement for a selectable phenotype limits the scope of these screens.

Most loss-of-function phenotypes are not directly selectable. It is perhaps more likely that loss-of-function mutants display a hypersensitivity phenotype, for example in conditions that cause dependence on a particular pathway in which the mutant gene acts. However, since the assay in such a situation would kill the cells of interest, this is of no use when cells are only present at a low level in a large and complex pool. To conduct such a screen, homozygous mutants would have to be individually isolated, replica plated and treated with (say) a drug to identify sensitive mutants. These could then be recovered from the replicate. This would be a classic genetic screen, but in order to apply it the recovery of homozygous mutants needs to be uncoupled from the screen for phenotype. This was the motivation to develop a technique to isolate homozygous mutants independent of their phenotype. These can then be screened in a separate step.

1.5 Isolation of homozygous mutants by selection for copy number increase

In this thesis, I present a method to isolate homozygous cells from pools of heterozygous mutants in a *Blm*-deficient genetic background. In Chapter 3 I describe a selection scheme to recover homozygous mutants based on their copy number, similar to the high [G418] strategy described above but much more stringent and applicable to a wide range of loci. The vector is based on the PB transposon and contains a novel mutagen designed to increase the number of mutable locations in the genome. I present data on coverage of the vector with regard to PB insertion site preferences (Chapter 3) and distance from the centromere (Chapter 4). Chapters 5 and 6 show the use of this vector to isolate homozygous cells. Finally, in Chapter 7 I present the results of studies to determine the basis of precise excision of the PB transposon.

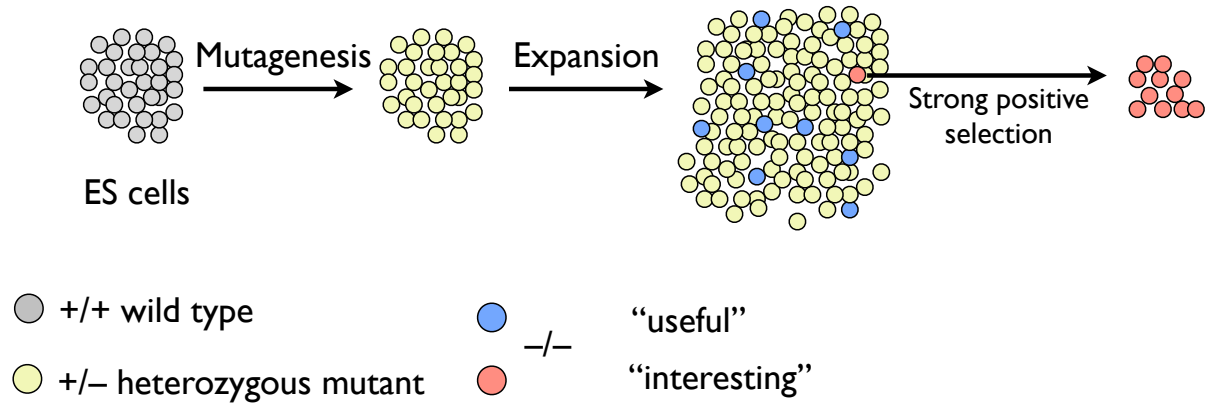


Figure 1.5: Screens for selectable phenotypes in *Blm*-deficient cells. Expansion of a population of random heterozygous mutants results in rare homozygous cells segregating. These can be isolated if the mutant phenotype is strongly selectable.

Chapter 2

Materials and Methods

2.1 Embryonic stem cell lines

2.1.1 Wild-type cell lines

AB2.2 is derived from a 129S6 blastocyst (McMahon and Bradley, 1990). The cell line carries an inactivating mutation in the *Hprt* gene on the X chromosome.
AB1 is derived from the same mouse strain as AB2.2, but has an active *Hprt* gene.
JM8 is derived from a C57BL/6N blastocyst. Feeder-independent (JM8.N4) and dependent (JM8.F6) subclones are available. I derived JM8A3 from JM8.F6 by fixing the naturally occurring *nonagouti* (*a*) coat colour mutation (Pettitt *et al.*, 2009); this cell line is used as wild type in some experiments.

2.1.2 *Blm*-deficient cell lines

NN5 was derived from AB2.2 cells by gene targeting (Luo *et al.*, 2000; Guo *et al.*, 2004). These cells are compound heterozygotes at the *Blm* locus, genotype $Blm^{tm3Brd/tm4Brd}$. The m4 allele is a deletion of exon two. The m3 allele results from an insertion event followed by Cre recombinase treatment, the net result being a duplication of exon 3.
NRB2 and RECE8 are NN5 cells containing a Cre-ERT2 gene integrated by gene targeting at the *Rosa26* locus (Figure 2.1). I used the same targeting method described in Vooijs *et al.* (2001), although I used a *bsd*-expressing version of the targeting vector obtained from David Adams. The two lines were derived from the same targeting.

Figure 2.1: Targeting NN5 cells with a *Rosa26*:Cre-ERT2 construct. Digest, probe as in Vooijs *et al.* (2001). Left, NRB2; right NN5.

$\boldsymbol{Blm}^{e/e}$ was derived from JM8.F6 by Amy Meng Li (Li, 2010). The *Blm* locus is homozygously targeted in these cells, and incorporates a blasticidin S deaminase (*bsd*) selectable marker gene and an enhanced green fluorescent protein gene (*EGFP*), both of which are constitutively expressed. These cells display the increase in sister chromatid exchanges (SCEs) characteristic of Bloom syndrome and do not express detectable Blm protein.

2.1.3 Other mutant cell lines

$\boldsymbol{Xrcc4}^{-/-}$ **and** $\boldsymbol{Xlf}^{\Delta/\Delta}$ are derived from the TC1 wild type cell line (129S7 strain) (Zha *et al.*, 2007). These cells were a kind gift from Fred Alt and Shan Zha (Children's Hospital, Harvard Medical School).

2.2 Cell culture

2.2.1 Culture conditions

ES cells were maintained in DMEM supplemented with 15% serum, 2 mM L-glutamine and 100 μM β-mercaptoethanol (M15 medium) on a layer of irradiated SNL76/7 feeder fibroblasts as previously described (Ramírez-Solis *et al.*, 1993). Medium was changed daily. For JM8 and its derivatives, recombinant mouse leukaemia inhibitory factor (LIF) was added to growth medium at 100 U/ml. For routine passaging cells were treated with 0.1% trypsin-EDTA in phosphate buffered saline (PBS) for 15 minutes (10 minutes for JM8 derivatives), quenched with an equal volume of M15 medium, clumps disrupted by pipetting and cells then transferred to a fresh plate pre-fed with M15.

2.2.2 Selective media

Drugs used for selection and their concentrations are listed in Table 2.1. For convenience, several abbreviations for drug-containing M15 media are used as follows: **DBL**, G418 (200 μg/ml) and Puromycin (3 μg/ml). **HGFL**; HAT, G418 and FIAU [L—LIF]. **HTGL**; HT (hypoxanthine and thymidine, i.e. HAT without aminopterin) and G418.

Drug	Concentration	Purpose
G418 (Geneticin)	180–200 μg/ml	Selects for *neo* expression
Puromycin	3 μg/ml	Selects for *puro* expression
Blasticidin S	10 μg/ml	Selects for *bsd* expression
HAT	0.1 mM/0.4 μM/16 μM	Selects for *Hprt* expression
6-Thioguanine	10 μM	Selects against *Hprt* expression
FIAU	200 nM	Selects against hsvTK (ΔTK)
Bleomycin	0.1–1 μg/ml	Causes double strand breaks

Table 2.1: Drugs used in selective media and concentrations. HAT—Hypoxanthine/Aminopterin/Thymidine mixture; FIAU—1-(2-deoxy-2-fluoro-1-D-arabinofuranosyl)-5-iodouracil

Mechanism of resistance

Most of the resistance genes encode an enzyme with activity that metabolises the associated drug, rendering it non-toxic. The exception is selection involving *Hprt*. HAT medium, a mixture of hypoxanthine, aminopterin and thymidine is used to select for *Hprt* (positive selection). There are two cellular pathways for guanine and adenine (the purine bases in DNA) synthesis, the *de novo* pathway which synthesises purines from simple metabolites, and the salvage pathway, which recovers purine ring compounds from other pathways. *Hprt* encodes an enzyme in the salvage pathway, hypoxanthine/guanine phosphoribosyltransferase, which adds a ribose sugar and phosphate to recovered bases to form a nucleotide that can be incorporated into RNA or reduced to form a deoxyribonucleotide for DNA synthesis. The salvage pathway can support cell growth and division on its own if the *de novo* pathway is blocked, provided there are enough purines around for salvage, but in this situation *Hprt* becomes an essential gene. This is the basis of HAT selection: aminopterin is a small molecule inhibitor of dihydrofolate reductase (DHFR), a key enzyme in the *de novo* pathway. Therefore, when cells are grown in aminopterin they are dependent on the salvage pathway, and thus on a functional copy of *Hprt*. A high concentration of hypoxanthine is included as a substrate for Hprt. After HAT selection, the medium is supplemented with hypoxanthine for two further days (I use 1× HT supplement, Invitrogen), to allow DHFR activity to recover. Although HAT and HT media contains thymidine, it is not relevant in this case; it is included to enable a similar selective strategy to be used with the pyrimidine synthesis pathway and the thymidine kinase gene.

Selection against *Hprt* function uses 6-thioguanine (6-TG). This is metabolised to 6-thioguanosine by Hprt, which can be incorporated into DNA. This is recognised by the mismatch repair machinery, leading to a persistent DNA damage response that eventually results in cell death. FIAU works on a similar basis as a toxic uracil mimic (dUMP can be metabolised to dTMP and incorporated into DNA).

2.2.3 Transfection of ES cells

Electroporation was carried out as described previously (Ramírez-Solis *et al.*, 1993). Typically, a suspension of 1×10^7 cells in 0.9 ml PBS, pre-mixed with DNA, was electroporated at 230 V, 500 μF in a BioRad GenePulser. After incubation at room temperature for five minutes, cells were transferred to a plate with feeder cells and M15 medium.

Lipofection using Lipofectamine 2000 (Invitrogen) was either done using 90% confluent adherent ES cells using the manufacturer's protocol or, with generally better results, using trypsinised cells in suspension. Cells were fed two hours prior to transfection. One hour later, Lipofectamine-DNA complexes were prepared as recommended by the manufacturer and left to incubate at room temperature (100 μl total volume for a 24-well plate). Cells were trypsinised and resuspended in OptiMEM (Invitrogen, 500 μl for a 24-well plate, around 500,000 cells), added to a fresh plate and mixed with the Lipofectamine-DNA solution. After incubation at 37° for three hours, 1 ml M15 medium was added. Cells were usually passaged the next day to a larger plate. Amaxa transfections or lipofections using the Transmessenger reagent (Qiagen) were per manufacturers' protocols.

2.2.4 Cellular analysis

Flow cytometry

For flow cytometry of live cells, cells were harvested by trypsinisation and resuspended in PBS with 1% FCS. The suspension was filtered through a 30 μm mesh (Partec CellTrics) immediately prior to flow

cytometry. A Beckman-Coulter FC-500 was used for flow cytometry and data analysed using Flo-Jo software.

For DNA content analysis, cells were fixed by pipetting a small volume of cell suspension in PBS directly into 5 ml 70% ethanol at –20°C. After fixing overnight at –20°C, the resulting nuclei were resuspended in PBS containing 2 μg/ml propidium iodide and 0.5 mg/ml RNAse A, and incubated at room temperature to digest RNA.

Growth analysis

To stain colonies, I rinsed plates once in PBS and added a small amount of 1% (w/v) methylene blue in 70% ethanol. After 15 minutes I rinsed plates by submerging several times in tap water, and left to destain in water overnight. For measurements of cell viability I used the MTT test. MTT ((3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) was dissolved in PBS, with sonication, to make a 5 mg/ml stock. Cells were fed with M15 medium and one tenth of the volume of MTT solution added two hours later. After two hours, a purple precipitate forms in actively respiring cells. The precipitate was dissolved in 1:1 DMSO:Ethanol by shaking the plate for 2 hours at room temperature. Absorbance at 540 nm was measured in a plate reader, and a background reading at 620 nm subtracted.

Preparation of metaphase spreads

Actively growing cells in a 6-well plate (fed two hours previously) were treated with demecolcine (1 μg/ml) for at least one hour. Cells were harvested by trypsinisation and washed with PBS. The suspension was centrifuged and cells resuspended in the residual PBS in 14 ml round-bottom Falcon tubes. Five millilitres of 0.56% KCl was added and cells incubated at room temperature for seven minutes to swell cells. The cells were spun at 400×g for five minutes, the supernatant decanted and resuspended in the residual KCl solution by tapping the tube. To fix, 5 ml of 40% methanol:10% acetic acid fixative was added dropwise, with constant agitation using a vortex mixer set to a low setting. The preparation was centrifuged as above and this fixing process repeated one. After a final spin, the nuclei were resuspended in 200 μl of fixative and dropped onto slides to make chromosome spreads. Fixed nuclei were stored in fixative at -20°C.

2.2.5 Isolation of nucleic acids and proteins

Preparation of DNA for enzyme digestion

From 96-well plates, I followed the protocol described in Ramírez-Solis *et al.* (1993). For larger cultures, I harvested cells by trypsinisation, washed with PBS and lysed overnight in ES cell lysis buffer at 55°C. The next day, an equal volume of isopropanol was added to precipitate DNA. The aggregate was retrieved using a sealed glass capillary, rinsed in 70% and 100% ethanol and dried for five minutes at room temperature. DNA was redissolved in 5 mM Tris-HCl pH 8.0 or 10 mM Tris-HCl, 0.1 M EDTA pH 8.0 and stored at 4°C.

Preparation of cell lysates for PCR

For preparation of lysates directly from colonies, I picked colonies into 50 μl trypsin as usual, quenched the trypsin with an equal volume of M15 medium and pipetted to form a single cell suspension. Eighty microlitres of this was transferred to a 96-well plate for expansion. To the remainder, I added 180 μl PBS and span the plate at 800 ×g for five minutes. The supernatant was removed and cells resuspended in a tiny drop of PBS. Fifty microlitres of PCR lysis buffer (1× PCR buffer with 0.45% NP-40, 0.45% Tween-20 and proteinase K, McMahon and Bradley (1990)) were added. The plate was incubated in a humid atmosphere overnight at 55°C, and heated to 95°C for 20 minutes the next day to denature the proteinase K. Up to $\frac{1}{5}$ of the PCR volume was used as template.

Preparation of RNA

RNA was prepared using Trizol (Invitrogen) using the manufacturer's protocol.

Preparation of lysates for Western blotting

After washing in PBS, cells were lysed in ELB buffer (150 mM NaCl, 50 mM HEPES pH 7.5, 5 mM EDTA, 0.1% NP-40 including Complete protease inhibitor cocktail (Roche); 10^6 cells per ml) on ice for 30 minutes. Tubes were spun briefly to pellet debris, and the supernatant removed and stored at –20°C. Protein was quantified using D_C reagent (BioRad) with BSA as a standard.

2.3 ES cell genotyping

2.3.1 PCR and long range PCR

Conventional PCR used ThermoStart polymerase (Thermo Scientific). For long range PCR to confirm gene targeting, I used Extensor PCR (Thermo Scientific) with a protocol as follows: 92°C 2 minutes; 10 cycles of: 92°C 30 s, 55°C 30 s, 68°C 4 minutes; 20 cycles of: 92°C 30 s, 55°C 30 s, 68°C 4 minutes plus 10 seconds per cycle; 5 minutes 68°C. For genotyping the *Rosa26*:ERT2-iCre-ERT2 targeting I used LA Taq (Takara) as per manufacturer's instructions.

2.3.2 Mapping transposon integration sites by splinkerette PCR

Splinkerette PCR is a linker based PCR method to amplify a product where the sequence is only known at one end, i.e. the transposon (Devon *et al.*, 1995). The splinkerette is a double-stranded adaptor oligonucleotide that contains an unpaired region, in which one strand forms a hairpin with itself. Genomic DNA is digested with a restriction enzyme, usually a frequent cutter with a four base pair recognition site, and the splinkerette adaptors ligated. The ligation products are then used as template for PCR using one primer extending outwards from the transposon sequence, and one of identical sequence to the unpaired region of the non-hairpin strand of the splinkerette. This second primer is of no use until its complement has been synthesised by extension of the transposon primer (Figure 2.2, (Li *et al.*, 2010)). This ensures that only fragments that contain the transposon sequence are amplified. A further nested PCR step also improves specificity.

To prepare Splinkerette adaptors, I combined 150 pmol of each oligonucleotide in 100 μl of water and heated to 95°C for five minutes. The solution was allowed to cool slowly to room temperature, then stored at –20°C. I carried out restriction digests in 96-well plates overnight as for Southern blots, or in tubes using 5 μg of genomic DNA, using a total volume of 50 μl. I usually used *Sau*3AI or *Bfu*CI restriction enzymes, both of which leave a 5′ GATC overhang. After digestion, the enzyme was heat inactivated and 1.5 μl used in a ligation reaction with 2.5 μl adaptor solution in a total volume of 10 μl. Ligation was at 16°C overnight, and the reaction was heat inactivated the next day. One microlitre was used as template for PCR were carried out using ThermoStart polymerase, in a volume of 25 μl with 2 mM $MgCl_2$. Cycling conditions were as follows: 94°C, 30 s; 62°C 30 s; 72°C 90 s, 30 cycles followed by five minutes final extension at 72°C. One microlitre was used as template for the secondary PCR, using the same conditions. Splinkerette and primer sequences are given in Appendix B.

2.3.3 Southern blot

Probes were designed to be at least 300 bp, preferably 800-1000 bp long. For probes hybridising to genomic sequence, RepeatMasker[1] was used to exclude repetitive regions from the probe. PCR products were amplified from BACs where available, or by two rounds of PCR from genomic DNA. For internal transposon probles, restriction fragments of plasmids were used. All probes were gel purified. **Labelling** used the random primer method using α^{32}P-dCTP (PrimeIt II kit, Agilent). Twenty five nanograms of probe were used in a 50 μl labelling reaction; one labelling reaction was used for up to three hybridisations carried out in parallel. Five hundred picograms of 1 kb λ ladder DNA (Invitrogen) was included in the labelling reaction to show the molecular weight markers. **DNA** was prepared as above. Either an entire 96-well plate (following the procedure described in Ramírez-Solis *et al.* (1993)), or 5–10 μg, was digested overnight with 30 units of restriction enzyme in the appropriate buffer (50 μl volume). **Electrophoresis and transfer** used 0.6–0.8% agarose gels run for at least five hours or overnight at low voltage in 1X TAE buffer. Five nanograms of 1 kb λ ladder was run as a marker; a larger amount was typically run in a lane as far as possible from the samples for visualisation with ethidium bromide staining. Gels were soaked in denaturing solution (1.5 M NaCl, 0.5 M NaOH) for 1–2 hours. After saturation of the gel with denaturing solution, it was placed upside down on a sheet of cling film and the following placed on top: one sheet Hybond XL charged nylon membrane (GE healthcare), two sheets filter paper (Whatman), paper towels to a height of 10–15 cm. The entire transfer apparatus was covered with cling film, and the gel tray placed on top as a weight. This was left overnight to transfer. The following day, the membrane was washed in 2× SSC for five minutes and baked at 80°C for at least 30 minutes to dry out. **Hybridisation.** The buffer used for hybridisations was: 1.5× SSPE, 1% SDS, 1% (w/v) skimmed milk powder (final concentrations). Sheared, freshly boiled, salmon sperm DNA was added just before prehybridisation to a final concentration of 200 μg/ml.

[1] http://www.repeatmasker.org

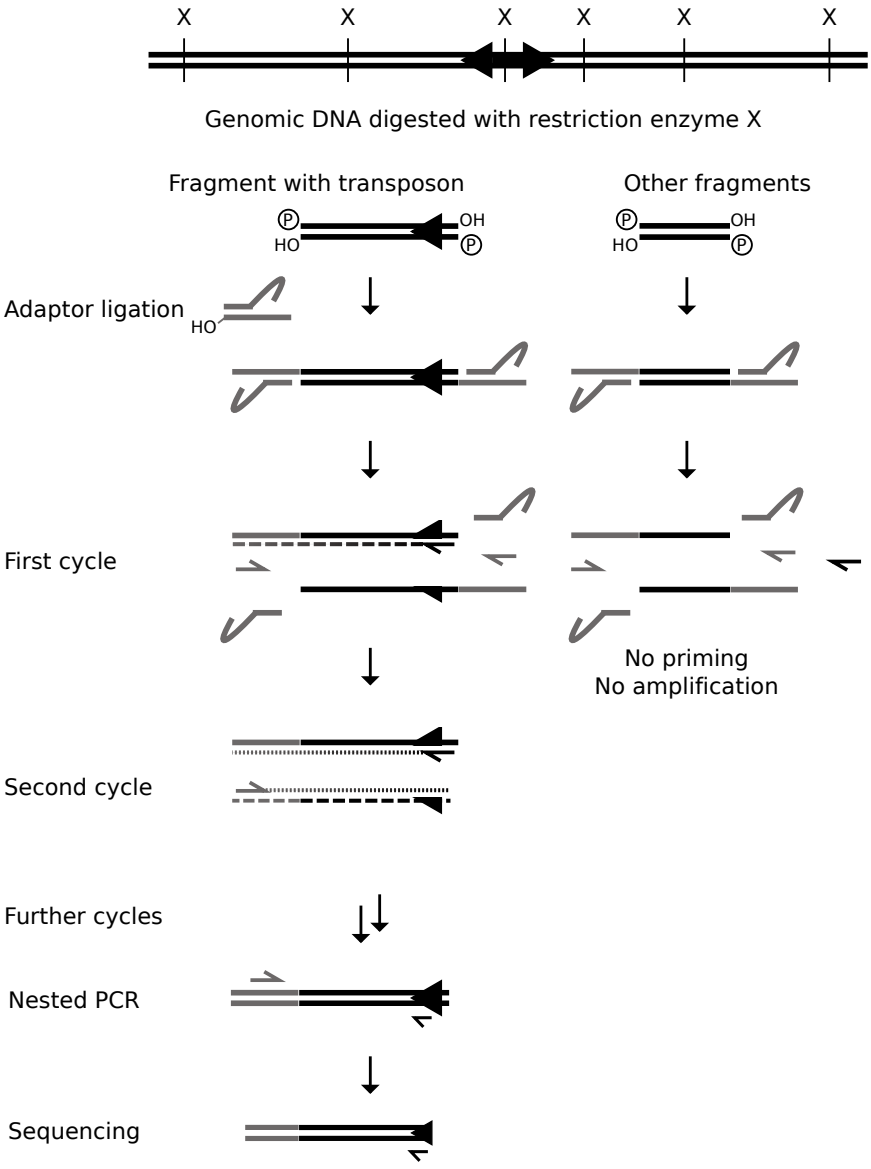


Figure 2.2: Splinkerette PCR method

:

Eight millilitres was typically used for a membrane of 30 × 10 cm. Prehybridisation was carried out for one hour at 68°C in a rotisserie oven, after which the (boiled) probe was added directly to the prehybridisation buffer. Hybridisation was at 68°C overnight.
Washing. The membrane was briefly rinsed twice at room temperature in 2× SSC, 1% SDS. This was followed by two washes at 65°C for 30 minutes each. Membranes were rinsed briefly in 2× SSC, sealed in bags and exposed to film for 1–5 days at –80°C.

2.3.4 RT-PCR

Three micrograms of RNA was used for reverse transcription (SuperScript II, Invitrogen) using oligo-dT primers, following the manufacturer's protocol. The reaction was diluted 1:5, and 1 μl used as template for PCR using standard conditions.

2.3.5 Western blot

Proteins were separated on pre-cast 4–12% Bis-Tris PAGE gels (NuPAGE, Invitrogen) using MOPS buffer. The proteins were transferred to a PVDF membrane. The membrane was blocked in PBST buffer (0.1% Tween-20) with 5% (w/v) skimmed milk for one hour at room temperature. The primary antiserum, diluted 1/200 in blocking buffer, was applied in a total volume of 2 ml, held on the protein surface of the blot by surface tension, and incubated overnight at 4°C. The membrane was washed three times in PBST prior to incubation with horseradish peroxidase (HRP) conjugated secondary antibody (1/1000 dilution) for one hour at room temperature. ECL+ chemiluminescence reagents were used for visualisation.

2.4 Molecular biology

2.4.1 Recombineering

Principle

Recombineering refers to manipulation of DNA in bacteria using recombination. Most commonly used lab strains of *E. coli* have a *recA* mutation. The *recA* gene product is homologous to eukaryotic Rad51, and forms a single stranded protein-DNA filament that begins the process of homologous recombination. This pathway needs to be knocked out to allow high copy number plasmids to be stably maintained without recombining with each other. Therefore, *recA* mutant bacteria form a stable environment to maintain and propagate plasmids. DNA manipulations (restriction digests, ligation etc.) are usually carried out *in vitro* and the products used to transform bacteria.

The recombineering method takes a different approach, and is essentially analogous to gene targeting in bacteria. The method works by transiently rescuing the *recA* mutation. At this point, homologous sequences that are present in the bacterium recombine with high frequency. Thus by designing suitable targeting constructs, BACs and plasmids can be manipulated. Importantly, as few as 30 nt of homology is sufficient for recombination, so these constructs can be easily synthesised as oligonucleotides, or tailed PCR primers.

I used the **EL350**, or its derivative **SW106**, strain for recombineering (Lee *et al.*, 2001; Warming *et al.*, 2005). EL350 contains an integration of a defective λ prophage, encoding the phage genes *exo*, *bet* and *gam*. The *exo* and *bet* genes encode a 5′ to 3′ exonuclease that resects DNA ends, exposing single stranded DNA to which *bet*, which substitutes for *recA*, can bind. The *gam* gene product is an inhibitor of the bacterial RecBCD nuclease complex, and when expressed prevents degradation of linear DNA (i.e. the introduced targeting construct) by RecBCD. The phage recombination genes are under the control of a mutant cI promoter that is repressed at 32°C and de-repressed at 42°C. Therefore, bacteria for recombineering are always grown at 32°C (there is some leaky expression of the recombination operon at 37°C), and heat shocked at 42°C immediately prior to transfection of the targeting construct.

Protocol

I typically grew 25 ml bacterial cultures, inoculated from an overnight starter culture, in baffled conical flasks at 32°C until an OD_{585} of 0.4–0.6 was reached. Then the culture was split in two, and one half grown in a 42°C shaking waterbath for 15 minutes, with the other (control) half remaining at 32°C. Both flasks were then transferred to an ice bath and swirled for five minutes to cool. To make electrocompetent cells, I then washed twice with ice cold distilled water (or 10% glycerol) in 14 ml round bottom Falcon tubes. Using round bottom tubes allows the bacteria to be resuspended very gently by swirling the tube in an ice-water slush. Electrocompetent cells were electroporated at 1.8 kV in a BioRad GenePulser, using 50 μl cell suspension in a 0.1 cm cuvette. Typically 1–10 ng of plasmid or targeting construct was used for transformation, and 50–100 ng for BAC. After electroporation 900 μl SOC medium was added, the culture transferred to a 14 ml Falcon tube and recovered in a 32°C shaking

incubator for at least one hour prior to plating.

2.4.2 Conventional cloning

Plasmid manipulation was carried out using standard procedures, using restriction endonucleases, antarctic or calf intestinal phosphatases, T4 polynucleotide kinase and T4 ligase purchased from NEB (Maniatis *et al.*, 1982). For gel purifications I used a kit from ZymoClean. Plasmids were usually maintained in DH5α *E. coli* purchased as chemically competent cells from Invitrogen and following their protocol for transformation. Ampicillin selection (*bla* gene) used 100 μg/ml ampicillin in LB or 2×TY medium, blasticidin selection (for *EM7-bsd*) using 50 μg/ml in low salt LB (Invivogen).

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

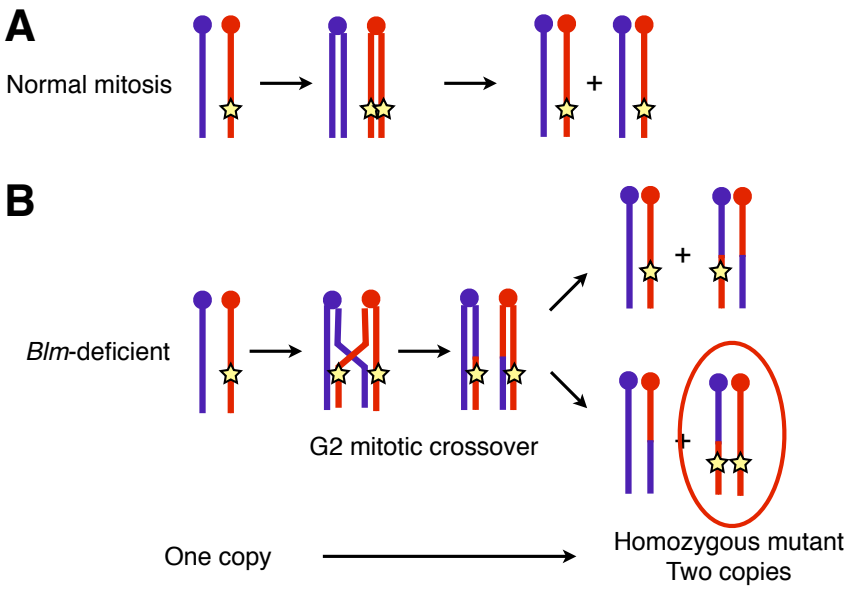


Figure 3.1: Copy number gain during loss of heterozygosity. Possible daughter cells arising from a single copy heterozygous mutant during: A—Normal mitosis, B—Mitosis with recombination and crossover in G2 phase.

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-

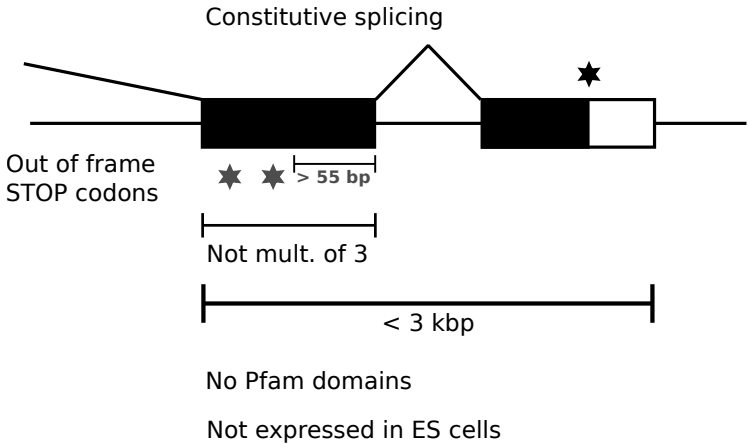


Figure 3.2: Features considered in design of the mutagen. Dark boxes—natively translated exons, empty boxes—untranslated exons. Asterisks represent stop codons.

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

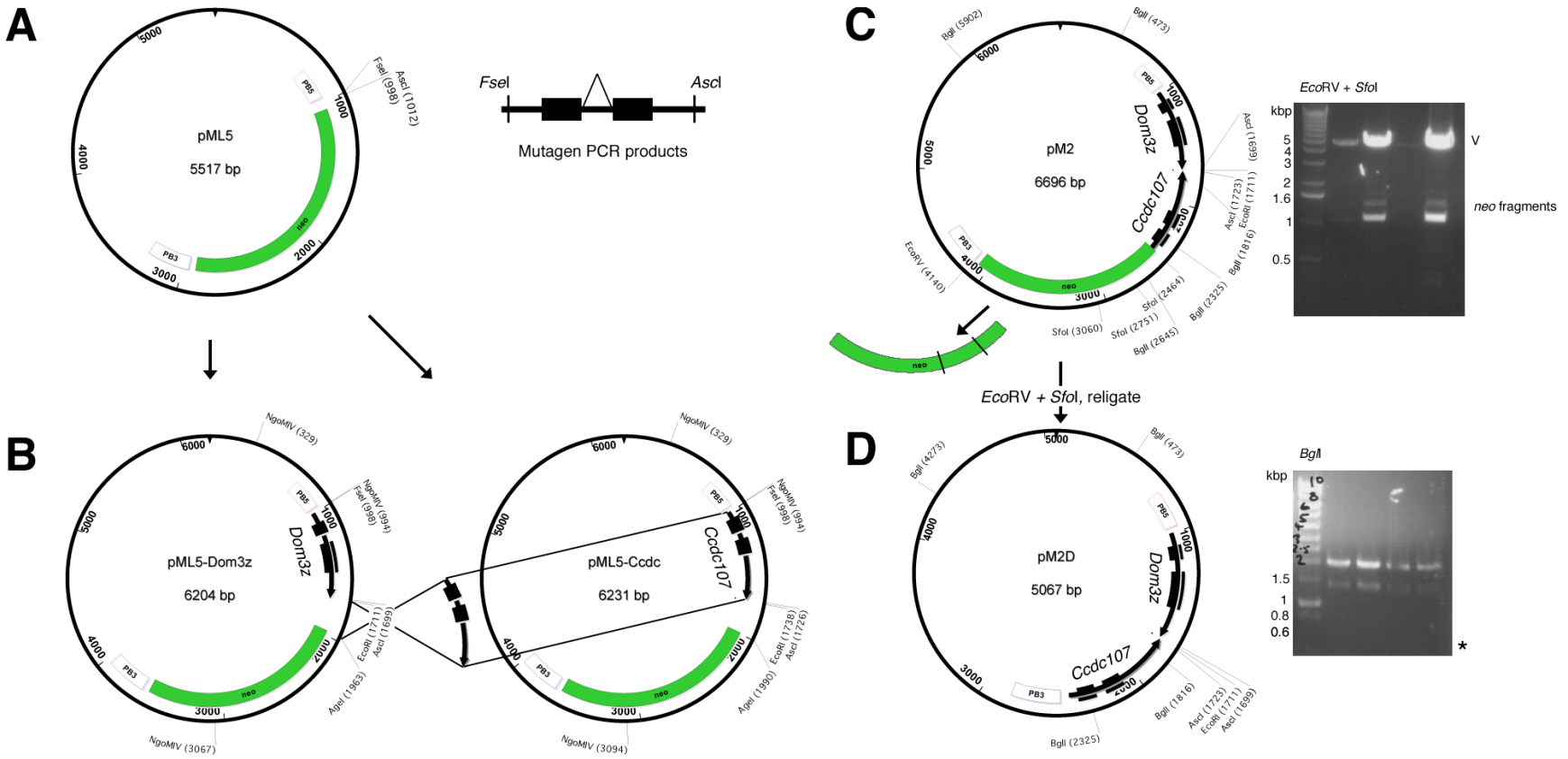


Figure 3.3: Cloning scheme for PiggyBac mutagenesis vector. A—Exon pairs amplified by PCR are cloned into pML5. B—*Ccdc107* is excised and transferred to the *Dom3z* vector in the opposite orientation. C—The resulting plasmid is digested with the enzymes indicated, the V (vector) band purified and religated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

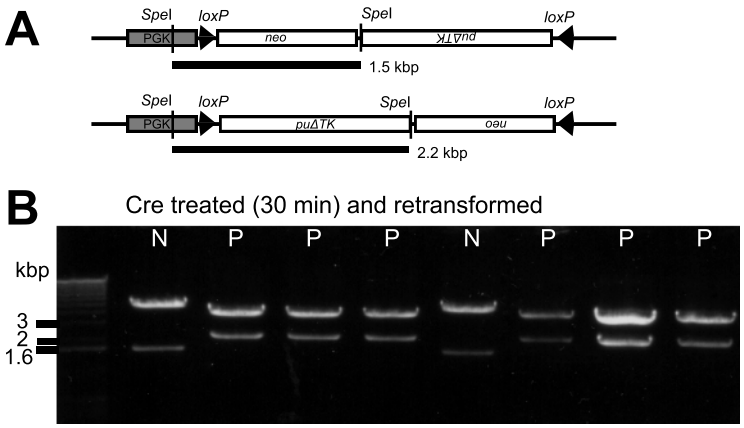


Figure 3.4: loxP sites are functional in the inverter construct. A—Map showing digest used. B—The inverter construct (Figure 3.5D) was treated with recombinant Cre and transformed into bacteria. Plasmids were digested with *Spe*I. A mixture of both possible orientations is seen, consistent with reversible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

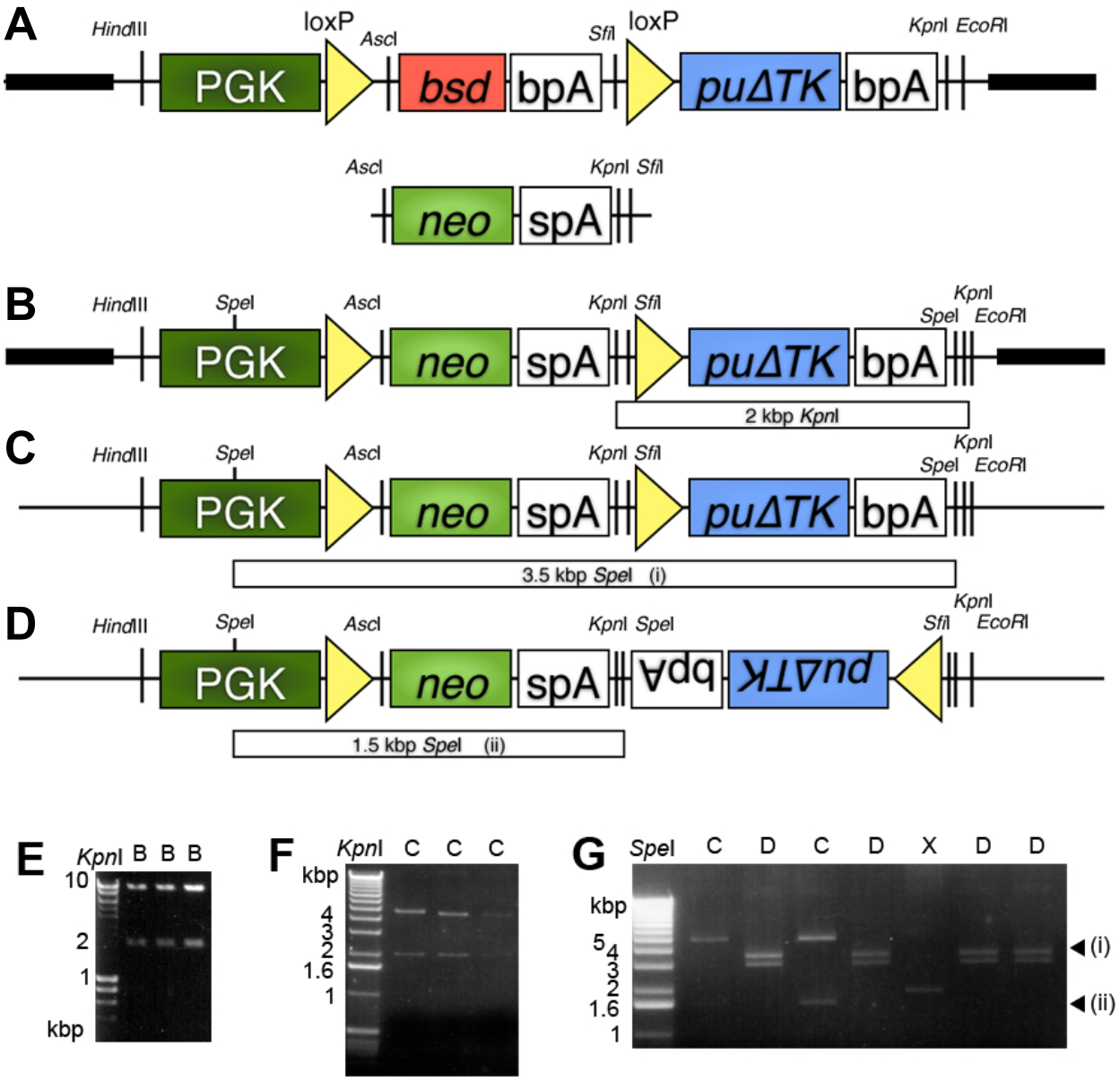


Figure 3.5: Cloning the inverter construct. A—pYTC85 and the PCR-amplified *neo*-SV40pA (spA). B—Result of replacement of *bsd*. C—Selection cassette moved into pUC19 backbone. D—Result of *Kpn*I digest and religation to give inverter construct. Lower panel, restriction digests using indicated enzyme 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.

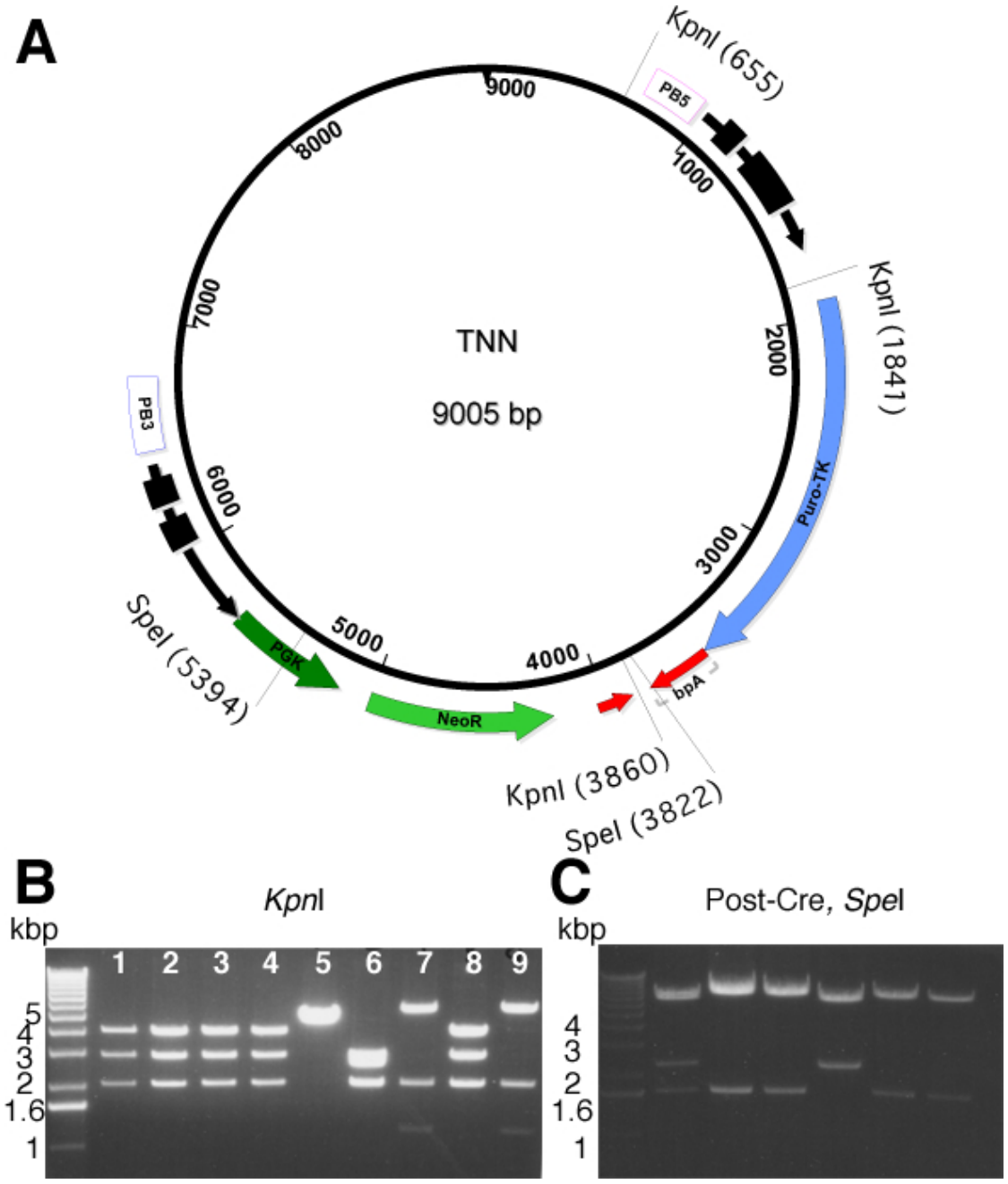


Figure 3.6: Cloning of the TNN plasmid. A—map of the plasmid. B—Results of ligation of the inverter construct into *Asc*I-digested and blunted pSDM-Pci. Lanes 7 and 9 are the desired orientation (shown in A), with the PGK promoter aligned with the promoter activity end of the PB transposon. Lanes 1–4 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.

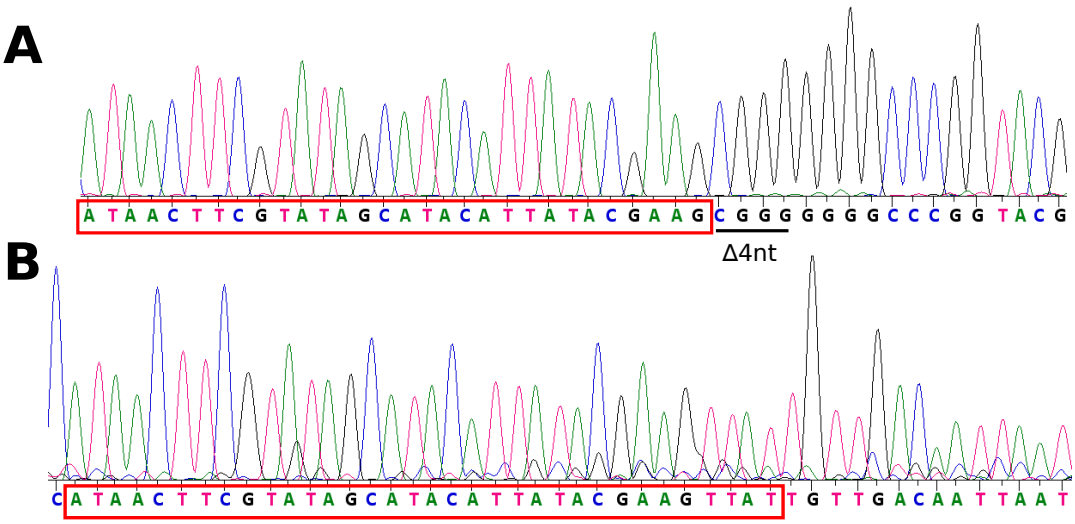


Figure 3.7: Fixing a mutation in a loxP site. A—A four base pair deletion present in the original plasmid. B—Sequence after replacement with the targeting construct. The downstream sequence is different as the *bsd* gene has not been removed in this plasmid.

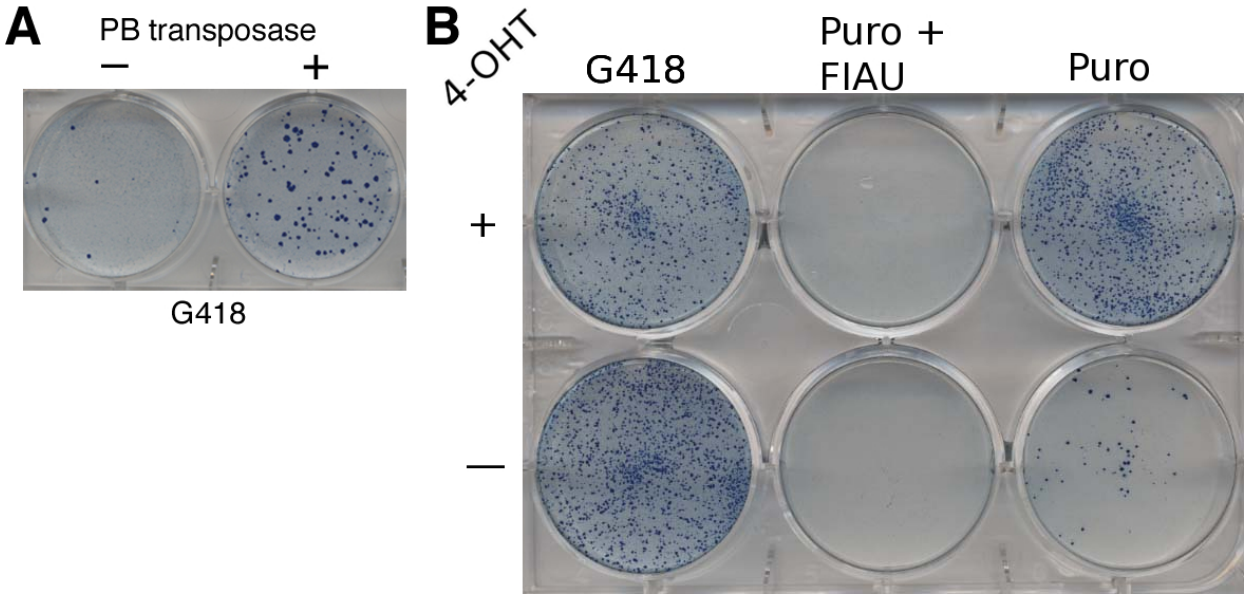


Figure 3.8: Function of the transposon construct in ES cells. A—ES cells were transfected with the TNN transposon construct with (+) or without (–) a transposase expression plasmid and selected in G418. B—NRB2 cells transfected with both plasmids were plated in the indicated drugs, and treated with 4-OHT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

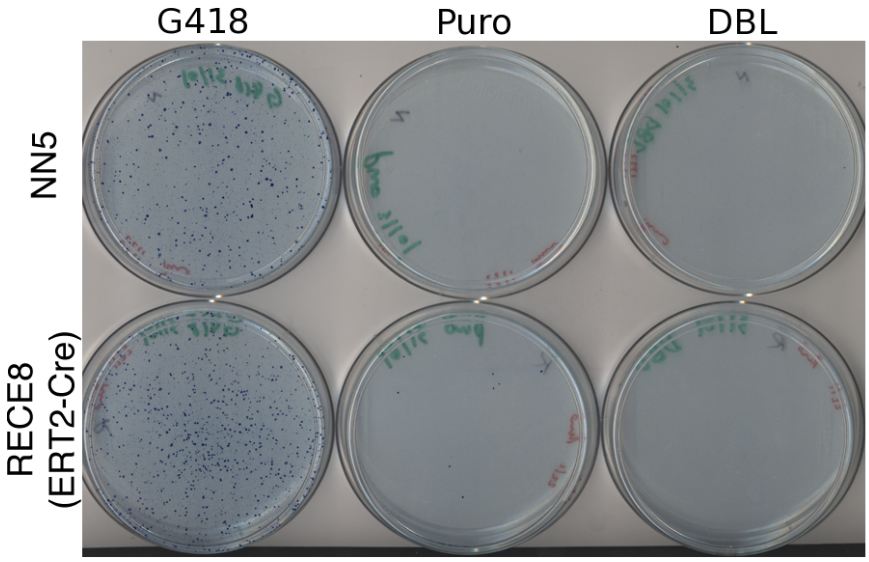


Figure 3.9: Background resistance from leaky ERT2-Cre activity. Top row: NN5 cells (no ERT2-Cre); Bottom row: NN5 cells targeted with *Rosa26::ERT2-Cre* (RECE8 cells). Cells were transiently transfected with TNN and transposase and selected in the indicated drug(s). NB—neither was treated with 4-OHT.

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

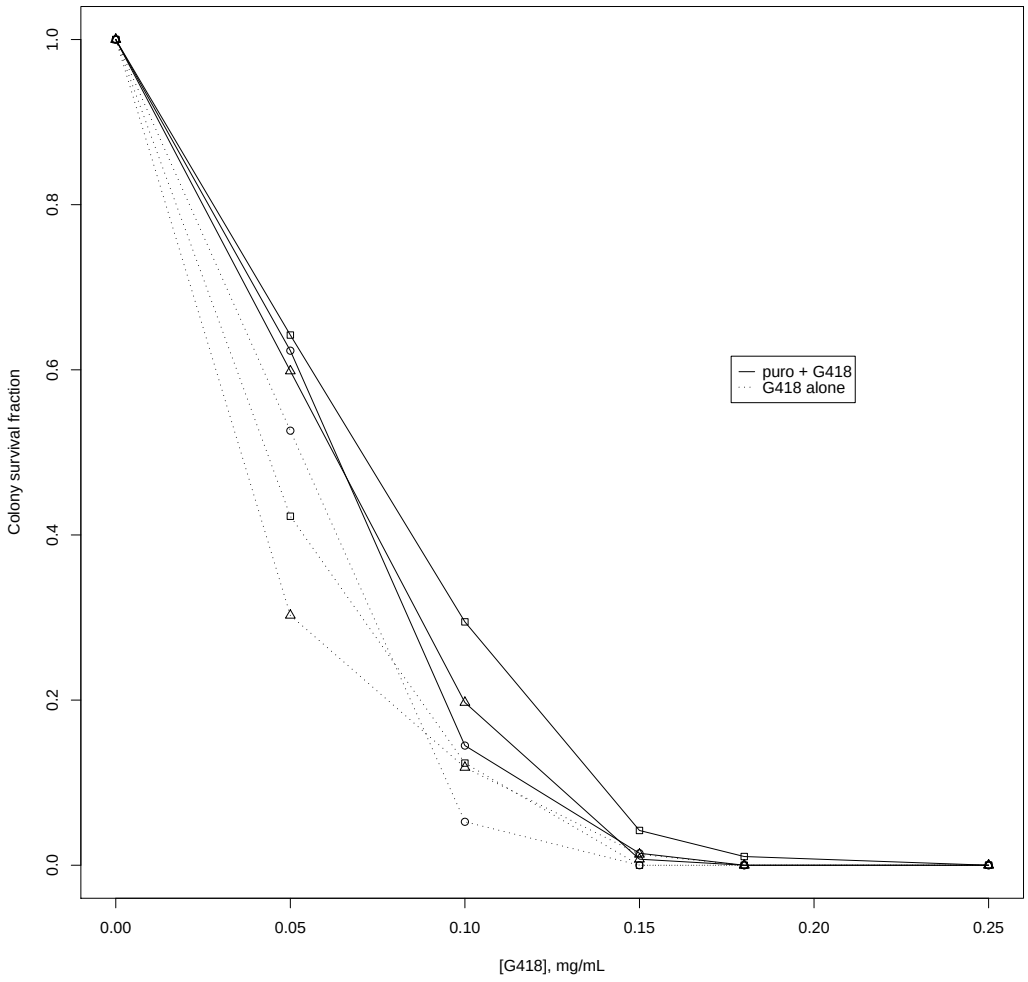


Figure 3.10: G418 kill curve for *puro*-expressing cells. Colony survival, as a fraction of unselected cells or with puromycin only as appropriate. Dashed lines: G418 in medium at indicated concentration; Solid lines: With 3 μg/ml puromycin. Three replicates are shown for each condition.

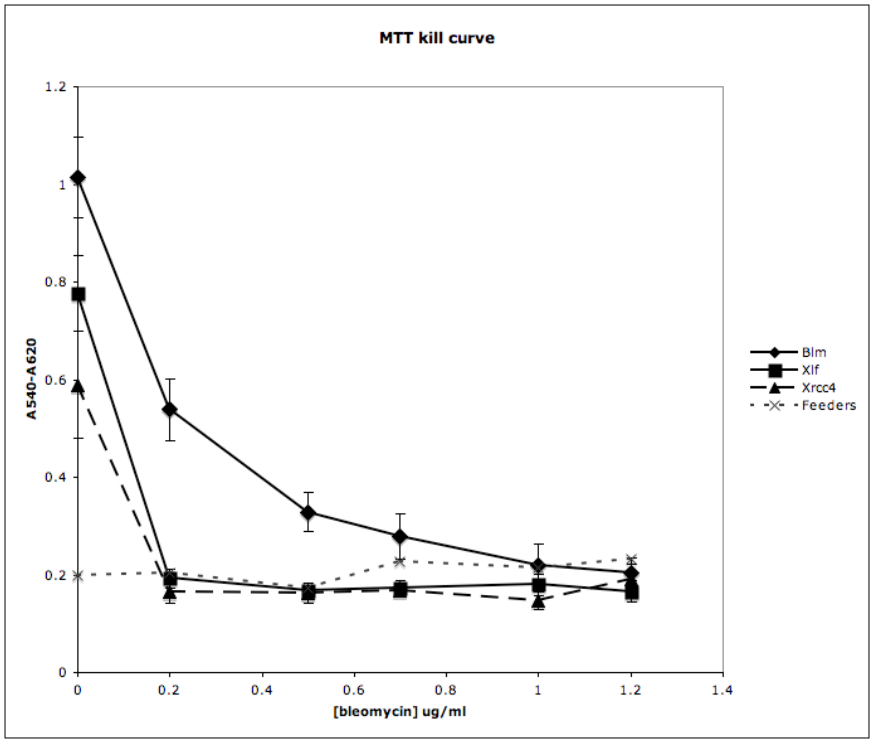


Figure 3.11: Sensitivity of *Xrcc4* and *Xlf* mutant cells to bleomycin. Results of the MTT test, reflecting electron transport chain activity and thus live cells, are plotted on the Y-axis. The indicated cell lines were treated with bleomycin for three days and allowed to recover for three further days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

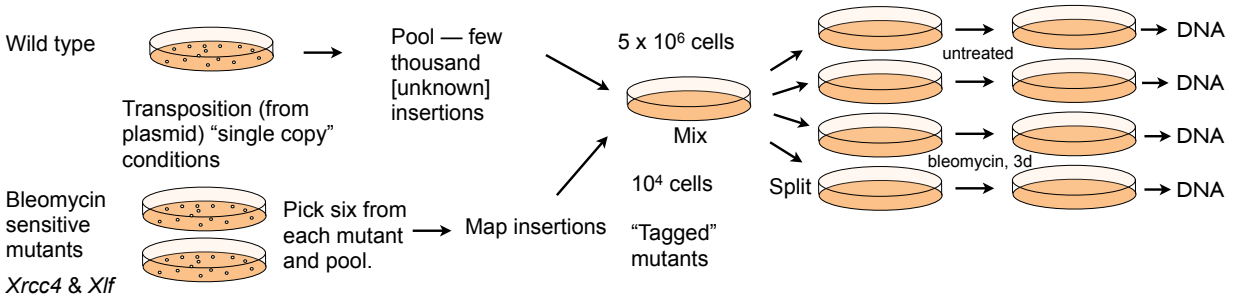


Figure 3.12: Setup of pilot experiment for Illumina sequencing and dropout screens

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

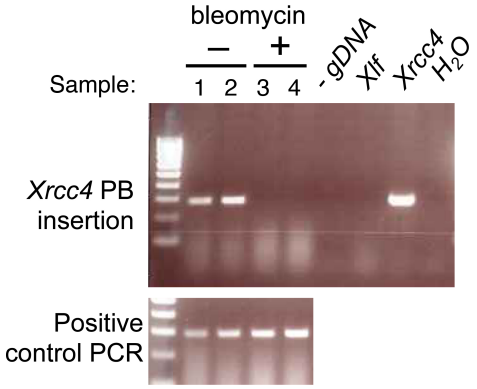


Figure 3.13: Detection of loss of a tagged *Xrcc4* mutant by PCR. Lanes 1–4: DNA prepared from the four pools of treated (+) or untreated (–) mutants; 5: Negative control DNA; 6: *Xlf* mutant pool; 7: *Xrcc4* mutant pool; 8: No template.

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.

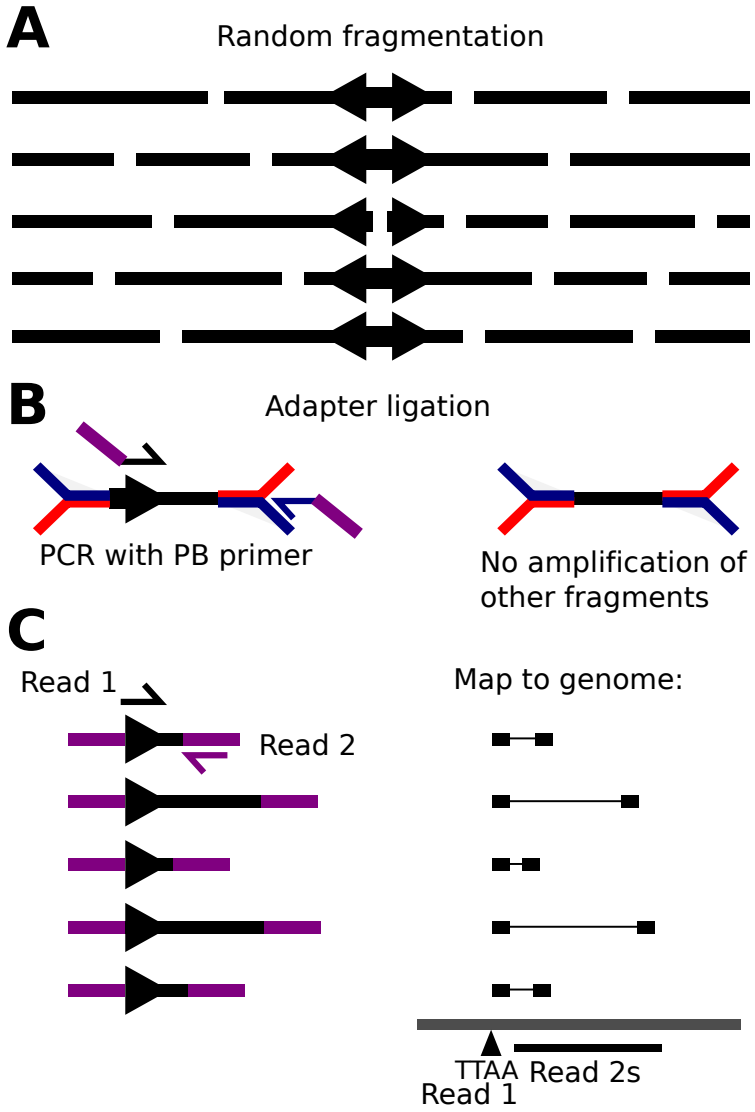


Figure 3.14: Distribution of paired ends in Illumina sequencing of PB insertions. Five molecules of DNA containing identical PB insertion sites are shown. A—Random fragmentation produces different breakpoints in each original molecule. B—Adapter ligation and PCR selects only fragment with the transposon. 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

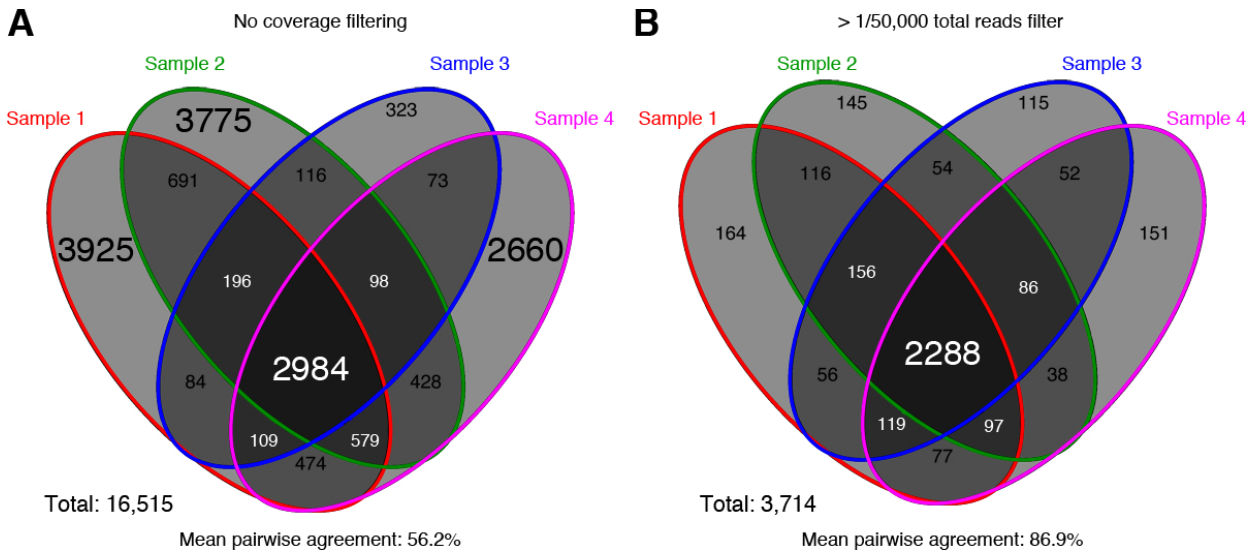


Figure 3.15: Venn diagram showing effect of applying coverage filter. Filtering by minimum coverage (B) mainly removes the insertion sites that are private to one sample. Sets with over 1,000 insertion sites are shown with larger text.

	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.

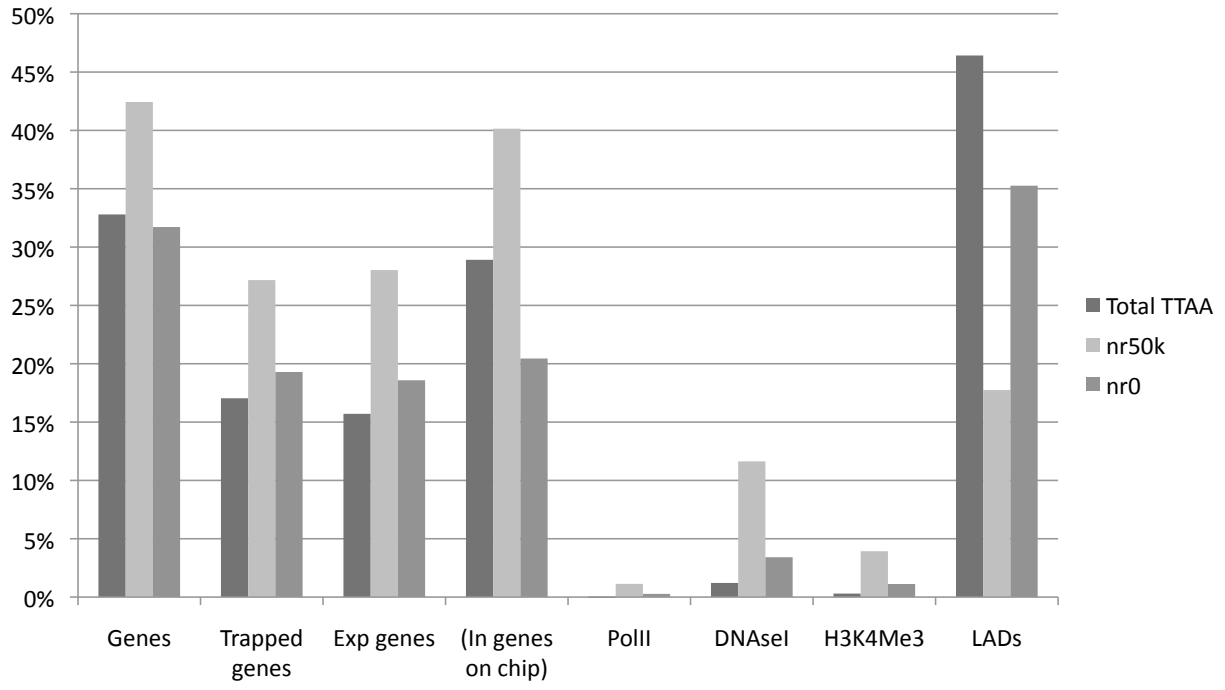


Figure 3.16: Graph of associations of PB insertions with genes and chromatin features.

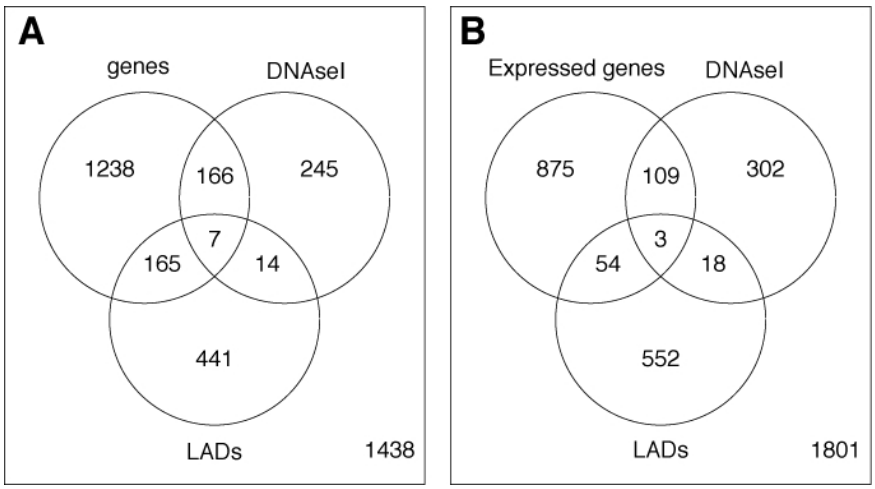


Figure 3.17: Venn diagrams illustrating insertion sites that overlap multiple features. A—Some DNAseI sites are associated with genes, but not generally with LADs. B—Although some sites in LADs are also in genes (see A), only 33% of these genes (57/172) are expressed in ES cells. This compares to 65%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

Chapter 4

The rate of loss of heterozygosity in *Blm*-deficient ES cells

4.1 Introduction

The rate of loss of heterozygosity (LOH) in *Blm*-deficient cells has been calculated previously as 4.2×10^{-4} and 2.3×10^{-4} events/locus/cell/generation respectively in the two *Blm*-deficient ES cell lines generated (Guo *et al.*, 2004; Yusa *et al.*, 2004). These measurements are based on a single locus in each case, *Gdf9* and *Fasl* respectively. The model of LOH by crossover after mitotic recombination predicts that LOH rate should vary by position on the chromosome. As LOH occurs at all loci distal to the point of crossover, loci located closer to the telomere should have an increased chance of a crossover occurring at a proximal position and thus an increased rate of LOH. For loci very close to the centromere most mitotic recombination events, if randomly distributed, will occur distally and not affect the centromeric locus.

If the rate of LOH does vary significantly across the genome, the effective coverage of the libraries will be affected. The chance of recovering homozygous mutations in genes close to centromeres may be reduced, and genes close to telomeres increased. I decided to investigate this by determining the rate of LOH at several different chromosomal positions. Working on the assumption that mitotic chiasmata and crossovers are distributed randomly, I chose three loci along the length of chromosome 11 to investigate, including the previously measured *Gdf9* locus.

LOH rates in this context are typically measured by inserting a selectable marker at the locus to be tested. For the *Gdf9*, a *HPRT* minigene was used and LOH assessed by its loss, which produces a 6-thioguanine-resistant cell (Luo *et al.*, 2000). For *Fasl*, a mutant *neo** gene was used, and high G418 selection used to select homozygous *neo/neo* cells (Yusa *et al.*, 2004). The rates measured were similar. For the *neo** selection there was a high background of surviving *neo/+* cells that had to be corrected for by genotyping resistant cells. Negative selection may also have background, for example if the spontaneous mutation rate is high, and thus this method works on the assumption that mitotic recombination and crossover is the primary mechanism in *Blm* cells.

4.1.1 Using fluctuation analysis to measure the rate of rare events in cell culture

LOH is a rare event, so to measure it a large number of cells need to be analysed. This presents a problem as LOH can occur during the expansion of the cells to a sufficient number. This results in a large variance in the number of cells that have undergone LOH observed in the culture. It is impossible to say from a single culture whether the number of resistant cells resulted from a single early LOH event, giving rise to a cell that expanded clonally over the remaining generations, or from multiple later events. If multiple cultures are set up, each beginning from a single cell, the number of resistant cells after a set time will fluctuate between cultures, due to the the disproportionate effect of early events on the final number of resistant cells. In a seminal paper, Luria and Delbrück developed a formula to explain this fluctuation in the situation of spontaneous mutation to phage resistance in bacteria (Luria and Delbrück, 1943). LOH is analogous to a spontaneous mutation, and the same formulae and method can be used to calculate the rate.

Luria and Delbrück derived two equations that can be used to calculate the mutation rate. Both result from methods to deal with the large number of resistant cells obtained when a mutation occurs very early in the culture. The first, known as the p_0 method, simply ignores all cultures in which mutations occur and instead considers the number of cultures without mutants. The total mutations in the experiment distribute across all cultures according to a Poisson distribution, therefore the probability that no mutants occur in a culture can be calculated for a given mutation rate. Conversely, using the observed fraction of cultures that show no mutations, a mutation rate can be calculated. The p_0 method does not make efficient use of the information gathered, but is at least straightforward. Its main drawback is that a very large number of cul-

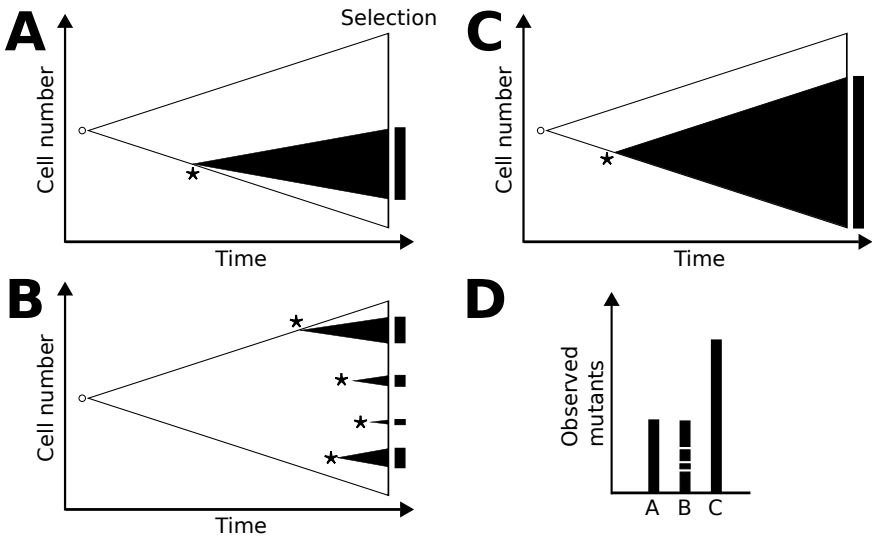


Figure 4.1: Measuring rare events in cell culture. Cultures are depicted as expansions of a single cell. Mutations (or LOH events) that arise, indicated by asterisks (*) continue to expand clonally. The number of mutant cells at the end of the culture period could result from one early event (A) or multiple later events (B). An extreme example of the effect of a very early mutation is shown in C. The figure is schematic, note that the Y axis should be a log scale if cells are growing exponentially—thus early mutations have a large effect on the final number of observed mutants (D).

tures is required to calculate a rate with any accuracy, and it relies on plating the entire culture to ensure all cells are interrogated. Both of these conditions are difficult to achieve using mammalian cell cultures.

The second method originally presented is the method of means. This uses the mean of the resistant cells to calculate a mutation rate. The problem of the long tail of the distribution is dealt with by assuming that none of the experimental cultures, which represent a small sample of the distribution of all possible mutant frequencies, are in the extreme tail. This is done in the derivation of the formula by only considering cultures that were mutation-free after a certain critical time. This allows a mutation rate to be calculated, but it is likely to be an overestimate. However, this can be used in situations where all cultures show mutations, and also in cases where only a proportion of the culture is plated.

Although the methods described in the original fluctuation analysis paper continue to be used today, several adaptations have been published for mammalian cells. As noted above, some assumptions that are acceptable for bacteria are not for mammalian cells. This applies especially to the assumption that the entire culture is plated—ES cell cultures typically have a plating efficiency of only 30–50%.

Jones *et al.* extended the principle of the p_0 method to provide an estimator of the mutation rate using the median number of resistant cells per culture. This also allows for plating of only a portion of the culture, thus plating efficiency can be incorporated. Moreover, they show that optimising the dilution such that roughly half the cultures have no mutants allows the rate to be calculated accurately with relatively few cultures (Jones *et al.*, 1994).

4.2 Results

4.2.1 Choice of loci

Two of the cell lines I used were generated as part of the TNP100 library (see Chapter 5). These are named by their well positions, D8 and F8, and both have TNP (i.e. $pu\Delta TK$-expressing) transposon integrations on chromosome 11. The co-ordinates of the insertion sites are 11:20,780,891 and 11:95,552,974 respectively (NCBI m37). I also used a cell line with $pu\Delta TK$ integrated by gene targeting at the *Gdf9* locus, which was generated by Amy Li (Li, 2010). As this locus was originally used to measure the rate of LOH in *Blm*-deficient cells (Luo *et al.*, 2000), using this cell line will allow my results to be compared directly with this rate. *Gdf9* also has the advantage of mapping almost exactly in the middle of D8 and

F8 (54 Mb from the centromere), providing a good test of whether or not LOH rate varies with distance from the centromere.

4.2.2 Calculation of mutation rate

I trypsinised cultures of these cells for at least 15 minutes and dispersed them to a single cell suspension by pipetting. I then plated 1,000 cells per 90 mm plate to obtain colonies. Each colony is a culture started from a single cell. I picked 24 colonies from each cell line after 10 days, and expanded them to a 24-well plate (via one passage on a 96-well plate). The average cell count at this stage was 751,571. Cultures with large differences from this value were discarded at this point, as the mutation rate calculation assumes that all cultures were equally expanded. One tenth of each culture was plated directly in FIAU selective medium, and the remainder diluted for counting and plating at low density (150 cells per plate) in non-selective medium to calculate the plating efficiency.

I calculated the average number of mutations per culture as follows, using the $\hat{m}_h$ median estimator derived by Jones *et al.*. For each series of cultures, I calculated the median number of FIAU resistant colonies r_m, and the mean cloning efficiency. The cloning efficiency was multiplied by the plated fraction (0.1) to obtain the effective plating p_e. The average number of mutations per culture is then given by equation (5) in Jones *et al.* (1994):

$$\hat{m}_h = \frac{r_m/p_e - ln(2)}{ln(r_m/p_e) - ln(ln(2))} \qquad (4.1)$$

The calculated mutation rates are shown in Table 4.1. The rate does appear to increase with distance from the centromere. However, the rates calculated are generally lower than those previously determined, as can be shown by comparing my rate for *Gdf9*, 2.5×10^{-5} events/cell/generation with that calculated by Luo *et al.*, 4.2×10^{-4}. Possible reasons for this are discussed below.

4.3 Discussion

4.3.1 Comparison with previously calculated rates

The rates of LOH that I calculated here are much lower than those previously determined. There are several possible reasons. First is that the different method employed here may be underestimating the number of mutations per culture. However the median estimator method gives similar results to the original formulae on other datasets, so should be applicable (Jones *et al.*, 1994). In any case, the magnitude of the difference is probably too large to be explained by features specific to one estimator. As a precaution, I did directly genotype the cell lines used in the experiment to ensure they were *Blm* mutants (not shown).

As I included the originally-measured *Gdf9* locus in my experiments, the discrepancy cannot be due to a locus-specific effect. A more likely reason is the difference in selection used between my experiments here and the previous rate calculations. I made use of the ΔTK gene for negative selection, whereas the previously reported calculations used *HPRT* or *neo** as described above. A possible mechanism by which this could affect the number of mutants recovered per culture is if the *puro*ΔTK mRNA or protein is more stable than *HPRT*, and therefore persists for longer after LOH occurs and removes the DNA. It is likely that resistance genes are expressed at high levels, as this has been artificially selected for in the choice of promoters and polyadenylation sites used in cloning vectors. If the cells are still functionally TK^+ for one or two generations after LOH, this will affect the numbers of FIAU-resistant colonies that can be obtained. Thus, using FIAU selection could result in a systematic underestimation of the mutation rate. Measuring LOH in wild-type cells using *puro*ΔTK would show whether this is the case. The off-rate of *HPRT* and *puro*ΔTK could be tested experimentally to investigate this further using, for example, Cre mediated deletion and measurement of the time taken to recover the maximum number of deleted clones. However, this is essentially the same experiment as the *Gdf9* comparison carried out here. Experiments presented in Chapter 5, also suggest that the actual rate of LOH (or at least copy number increase) is higher than that calculated here, further arguing for an effect of FIAU negative selection.

4.3.2 Implications for library coverage

The rate calculated for the most proximal locus in this analysis (D8) is about one third of the *Gdf9* rate. How relevant is this difference, over a distance of 33 Mb, with respect to library coverage? One way this can be interpreted is by considering how representative these loci are of all the genes in the genome. Plotting the positions of all genes in the genome reveals that *Gdf9* represents approximately the 30th percentile and the D8 locus approximately the 8th (Figure 4.3). Therefore, the current proto-

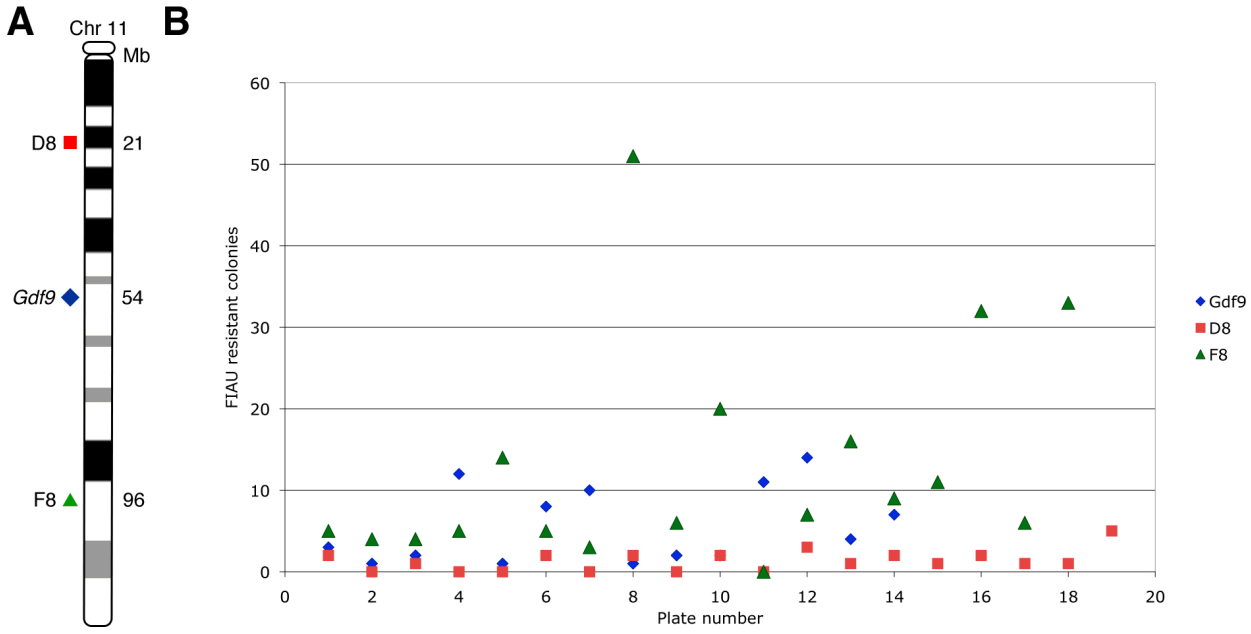


Figure 4.2: Number of LOH events observed for three loci on chromosome 11. A—Loci studied. B—Number of FIAU-resistant colonies obtained in replicate cultures for each locus. One-tenth of cultures expanded to a confluent 24-well plate was selected in each case. The ordering on the x axis is random.

Locus	Mb	Cultures	Median $FIAU^R$	Mean cloning	Mean cells/culture	$\hat{m}_h$: LOH events/culture	LOH rate
D8	21	19	1	0.35	752,914	7.53	9.9×10^{-6}
Gdf9	54	14	3.5	0.38	762,000	18.56	2.5×10^{-5}
F8	96	18	6.5	0.27	739,800	40.95	5.5×10^{-5}

Table 4.1: Calculation of LOH rate. LOH rate is events/cell/generation.

cols for library construction that are based on data for *Gdf9* should be sufficient for 70% of genes. However, the rate for the D8 locus should apply to 92% of all genes. As the rate is not drastically lower in practical terms, it should be possible to isolate LOH events at such loci with only slightly longer expansion times. These data provide a better guide for library construction, and support the hypothesis that the number of opportunities for initiation of proximal homologous recombination determines the probability of LOH at a locus.

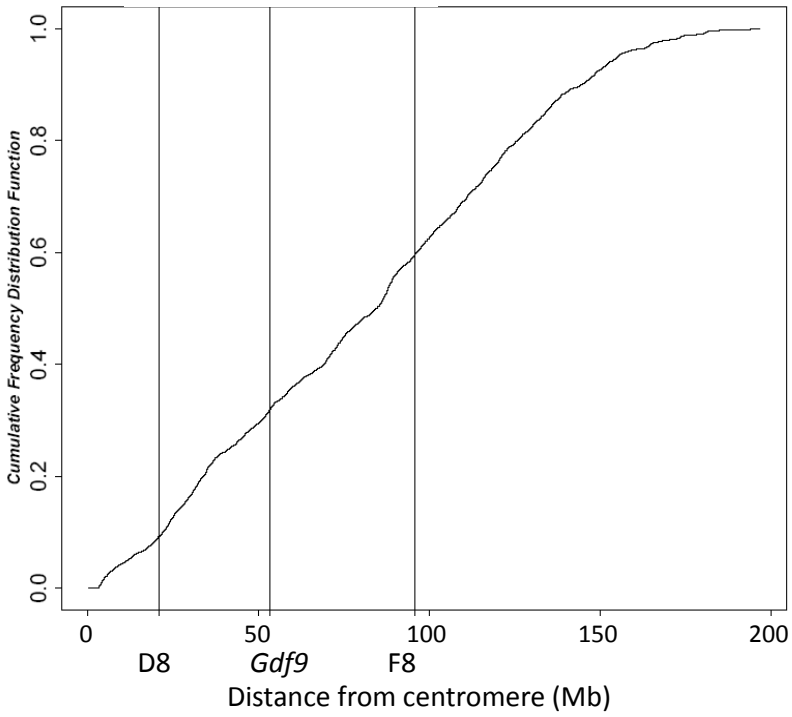


Figure 4.3: Plot of distance from centromere for all Vega curated mouse genes (Wilming *et al.*, 2008). The cumulative frequency of genes with their start (5′ end) at or before the value on the x axis is plotted. The positions of the three loci for which LOH rate was calculated are shown by vertical lines.

Chapter 5

Isolation of homozygous mutants in *Blm*-deficient ES cells based on copy number

5.1 Introduction

In this chapter I will describe preliminary experiments that I conducted to test the general method, and my construct in particular, for isolation of homozygous mutant ES cells. I created a library of single copy heterozygous mutations in *Blm* ES cells, and mapped the mutations by sequencing transposon-genome junction fragments. These clones were then used to test whether homozygous mutants could be recovered after expansion by selection based on the copy number of the transposon, which will be two in homozygous mutants but one in the heterozygous starting population. By conducting experiments on a small scale clone-by-clone basis I aimed to verify the mutagenicity and utility of my transposon construct, gain an understanding of how the loss of heterozygosity (LOH) process occurs, and potentially isolate some interesting mutants.

5.1.1 Copy number based selection

The method for isolation of homozygous mutations depends critically on a single copy insertion in the starting population of cells. As *Blm*-deficient cells with a heterozygous transposon mutation are expanded, they will segregate homozygous mutants at a low frequency as described earlier. These homozygotes will contain two copies of the transposon construct. The purpose of my transposon construct is to allow selective discrimination between cells with one and two copies. Cells with one copy will form the majority of the culture after expansion, with a minority of cells being homozygotes with two allelic copies that are "useful" for genetic screens. The culture will also contain cells that have lost the insertion and reverted to wild type as a consequence of the reciprocal LOH event that generates the homozygous mutants. As described in Chapter 3, the transposon construct contains a selection cassette encoding two mutually exclusively expressed resistant genes. Only homozygotes, which have two copies, are able to express both genes simultaneously after Cre recombinase treatment; these cells can therefore be selected in a combination of G418 and puromycin.

5.2 Results

5.2.1 Generation of single copy insertions

The experimental design is shown in Figure 5.1. For copy number selection to work it is important to limit the transposon to a single copy to begin with. This could be accomplished by mobilising the transposon from the single copy *Hprt* locus on the X chromosome. However, for simplicity in these experiments, I decided to generate the initial single copy clones by mobilisation of the transposon from a limiting quantity of plasmid coelectroporated with the transposase expression plasmid. This has been shown to result in mostly single copy insertions (Wang *et al.*, 2008). I used 100 ng of TNP transposon plasmid (i.e. *puro*-expressing construct, described in Chapter 3) with 10 μg pCMV-mPBase (Cadiñanos and Bradley, 2007) to transfect ten million NN5 ES cells in a volume of 0.9 ml. Cells were selected with puromycin for eight days and colonies picked. Analysis of the clones by Southern blot using a probe and restriction digest that allows discrimination of different insertion sites showed that the resulting colonies contained more than one insertion. However, the bands were clearly of different intensities, suggesting that the copy numbers of the corresponding insertions within the colony were different (Figure 5.2A, B). This could occur if the colony is in fact an unequal mixture of cells carrying different single copy insertions. To test this possibility, I repeated the process but replated the cells four days after transfection. Most subclones picked from this experiment bore single copy insertions (Figure 5.2C). The multiple bands seen in the first experiment likely arose from secondary genome to genome transposition events before the transposase activity was lost, resulting in mosaic clones. Another possibility is that two or more plasmid to genome transposition events occurred early in the growth of the colony, but after the founding cell had divided.

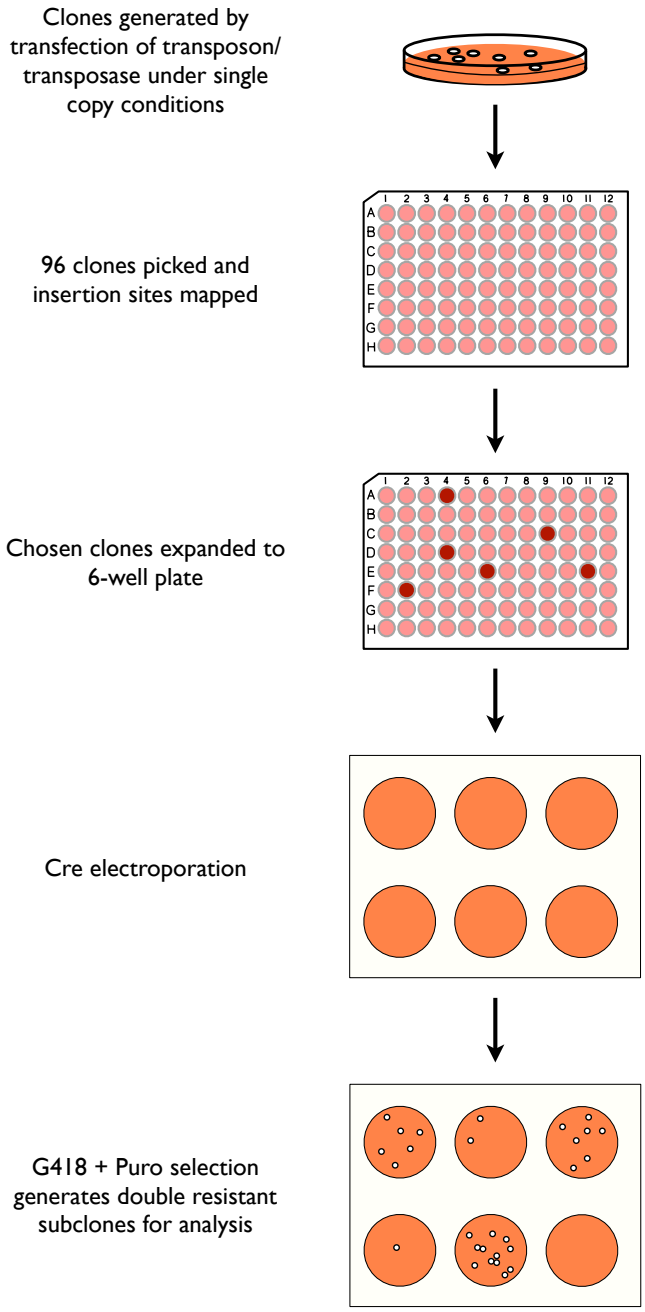


Figure 5.1: Experimental scheme for clone-by-clone isolation of homozygous mutants.

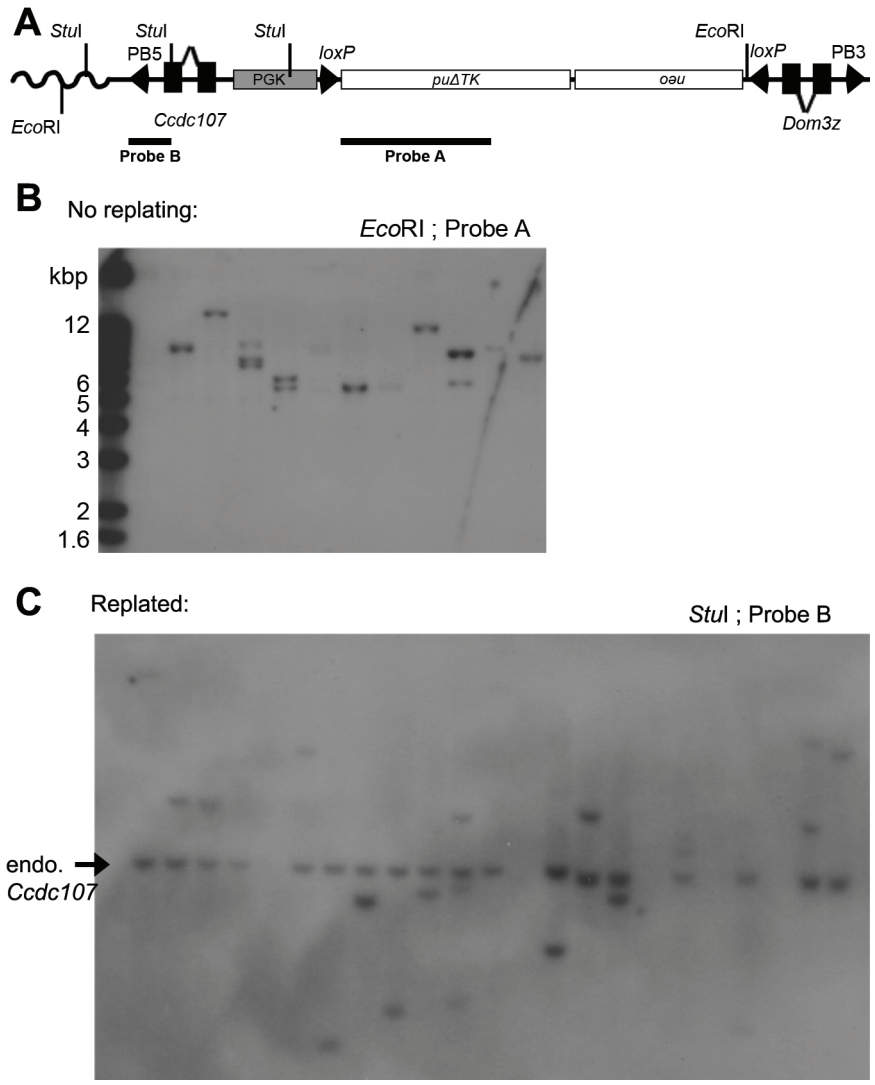


Figure 5.2: A—Map of transposon construct showing probes and restriction sites used. B—Clones picked without replating contain more than one insertion, but in different proportions. C—Replating after transfection resolves the multiple bands and reveals most clones to have a single insertion. Using probe B also detects a band corresponding to the endogenous copy of the *Ccdc107* gene.

Replating the cells after allowing time for the transposase activity to subside ensures that the colony picked is truly clonal (i.e. derived from a single cell with a stably-integrated transposon). I picked 96 clones from this second experiment to form the TNP100 arrayed library of heterozygous clones.

5.2.2 Mapping of insertion sites

I prepared DNA from a replica plate of this library and used splinkerette PCR (*Sau*3AI digest; see Methods) to amplify transposon-genome fragments. Sixty three PCR reactions gave a unique product (Figure 5.3A). I sequenced these fragments, and processed the sequences by clipping transposon sequence before the TTAA site and genomic sequence after any observed *Sau*3AI site. This removes chimaeric fragments that arise when two genomic fragments manage to ligate to each other before ligation to the splinkerette adaptors. I mapped the resulting fragments to the genome using SSAHA (Ning *et al.*, 2001). As both 5′ and 3′ fragments were amplified for each insertion, mapping confidence is highest when the fragments either side of the transposon map to the same locus, on opposite strands. In cases where only one side amplified a product, this can still be mapped. To ensure accurate mappings, I looked for a clearly visible transposon end and transition into genomic sequence before mapping these cases, and also required that the full length mapping was unique in the genome. Unambiguous mappings were obtained for 57 clones (Figure 5.3B). The insertion sites were spread across 17 chromosomes.

5.2.3 Generation of double resistant clones

I picked clones with successfully mapped insertions and expanded them to allow loss of heterozygosity to occur. I allowed the clones to expand to around five million cells on a 30 mm diameter (6-well) tissue culture plate, transfected a PGK-Cre expression plasmid by lipofection and transferred the cells to a 90 mm plate. This expansion is likely to be more than sufficient for LOH in most clones—based on the rate of LOH previously calculated (Luo *et al.*, 2000) an expansion to around 5,000 cells should be sufficient to observe one or more LOH events (Figure 5.4). However, as the transfection and locus-specific efficiency of Cre in this system will vary, I opted for a longer expansion period in these test experiments to increase the chance of observing and capturing homozygous mutants.

The day after plating I changed the medium to DBL medium (200 μg/ml G418 and 3 μg/ml puromycin). Some clones produced large numbers of double resistant cells, comparable to the number of cells plated. These clones are likely to have two copies of the construct, and were not analysed further. Some clones did not yield double resistant cells at all; in these cases the mutation could be homozygous lethal, or no LOH event occurred in the culture. However, some clones produced varying numbers of double resistant colonies, ranging from just a few to a few hundred (Figure 5.5).

The best way to characterise a population of clonogenic cells, such as the double resistant populations isolated here, is to pick and analyse subclones. I picked several colonies for each clone and genotyped them to investigate whether these cells represented real homozygous mutants.

5.2.4 Genotyping double-resistant clones

Southern blot to detect allelic transposon insertions

I designed a Southern blot probe to allow me to identify clones with two allelic copies of the transposon, and the relative amounts of TNP (*puro* oriented transposon) and TNN (*neo* orientation) contained in the cells (Figure 5.6A). The probe is a 1 kb *Sac*II–*Xmn*I restriction fragment of the transposon vector spanning the PB repeat and the *Ccdc107* exons. An *Nco*I site is present in this region that the probe will hybridise to, and also at the 5′ end of the *puro*Δ*TK* gene. Therefore, a different size *Nco*I fragment will be detected depending on the orientation of the resistance cassette: 1.7 kbp for TNN and 1.3 kbp for TNP. The other fragment detected by the probe is formed by the cut within the probe region and the closest *Nco*I site in the genome. The size of this fragment depends on the position of the insertion, and therefore allows discrimination between sites. Additionally, the probe detects two fragments of constant size from the endogenous *Ccdc107* gene—these can be used as a loading control.

I digested genomic DNA from double-resistant clones with *Nco*I, and probed the separated fragments with the probe described above. Two example clones are shown in Figure 5.6B. All subclones shown here contain four constant bands. Two of these are the predicted size for the endogenous *Ccdc107* bands, and the other two represent the TNN and TNP specific bands, as shown by hybridisation to digested plasmid. Homozygous mutants should have two copies of the transposon, one in each orientation, at a single locus. With the two endogenous bands, this should give five bands in total.

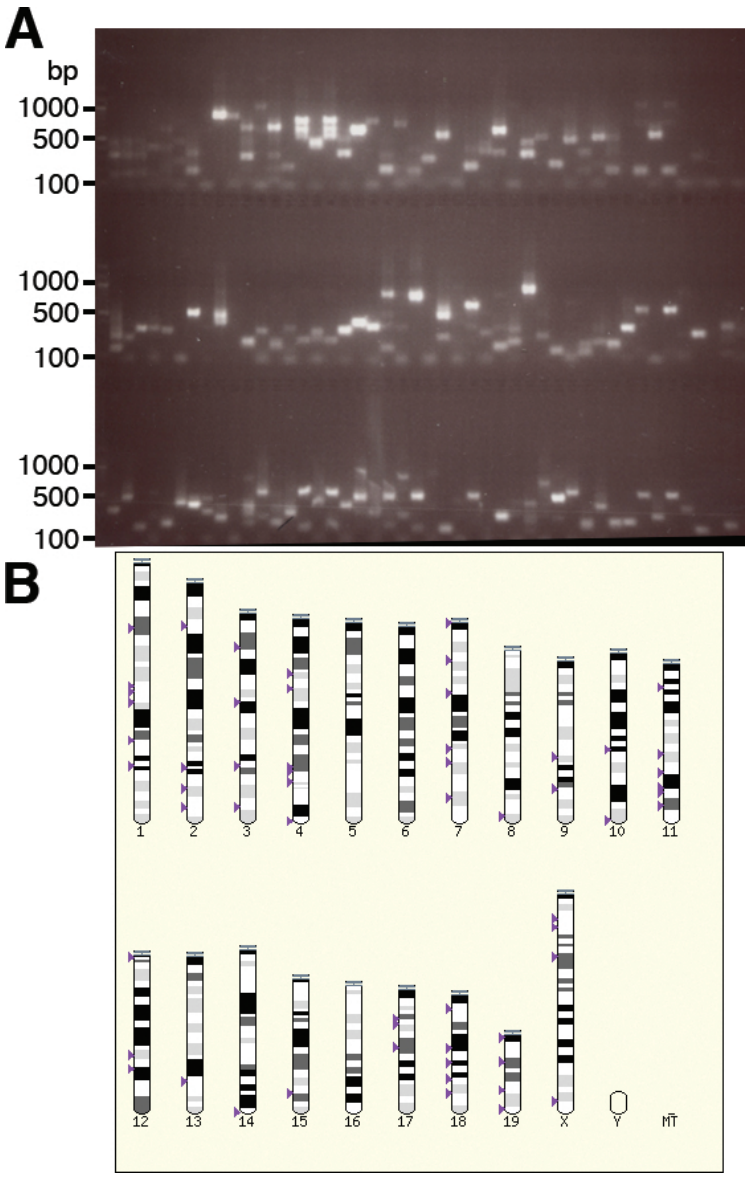


Figure 5.3: A—Second round splinkerette PCR products for the TNP100 set of clones. 5′ and 3′ products for each clone are loaded next to each other. B—Locations of successfully mapped PCR products

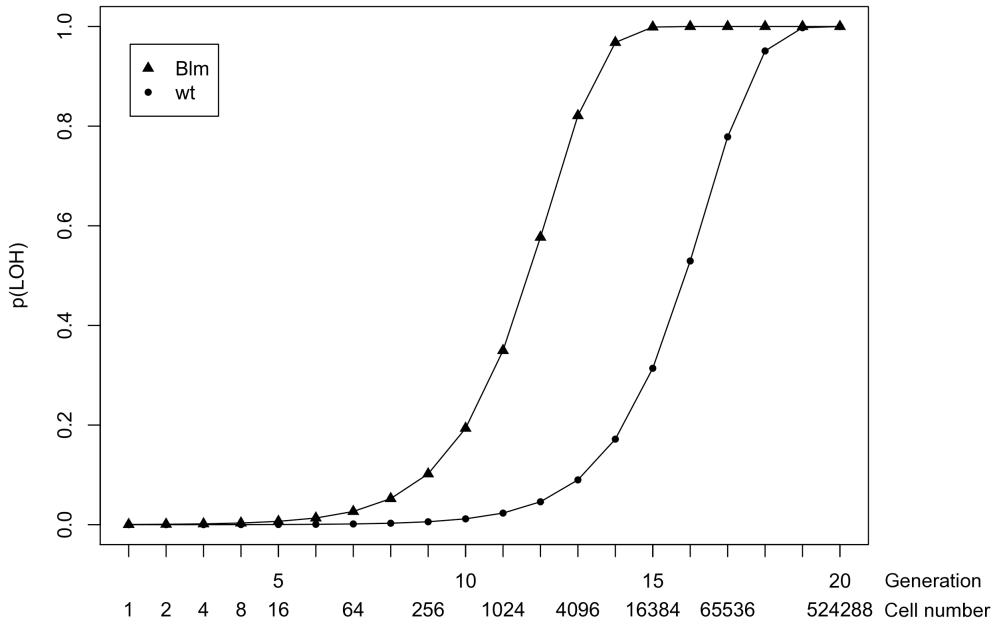


Figure 5.4: Predicted expansion time required to observe LOH events. The probability of at least one LOH event occurring at the specified generation is plotted: $1-(1-l)^n$ where n is the cell number at that generation and l is the LOH rate (rates for the *Gdf9* locus from Luo *et al.* 2000).

Clones with two copies of the transposon to begin with, at different loci, will have two locus-specific bands and therefore six in total. Both categories can be seen on the blot (Figure 5.6B). The clone with two non-allelic copies (F4) also contained two copies in the starting population as shown by Southern blot of clones from G418 selection only (Figure 5.6B lanes 1 and 2).

Selection background in initial experiments

For other clones from this experiment I observed a different result on the Southern blot. All subclones from one clone, and two out of six from another did not have both the *neo* and *puro* bands, despite surviving double selection (Figure 5.7A). These clones showed only the TNP band, indicating that they did not express *neo* from the PGK promoter, despite surviving G418 selection. However, during this experiment cells grew very slowly while under double selection. From other observations it emerged that this was due to the use of degraded L-glutamine in the lab culture media, rather than the double selection itself.

As G418 only kills actively dividing cells effectively, I considered whether slow growth when starved of L-glutamine, an essential amino acid, could explain the selection background, as I had not observed any background G418 resistance in previous experiments. By thawing replica plates of the double resistant subclones and reselecting in media containing fresh L-glutamine, I found that these cells were sensitive to G418 (and DBL) when grown in optimal culture conditions (Figure 5.7B). This highlights the importance of culture conditions in these selection experiments.

Thus, only double resistant subclones from one clone (G5) showed the expected band pattern in these experiments. I went on to analyse these in more detail.

5.2.5 Two classes of mutants are present in the double resistant population

Clones with two allelic copies of the transposon are potentially homozygous. To verify this, I checked to see whether the wild type locus was also present in these clones. I used a PCR assay with three primers in total—two locus-specific primers flanking the insertion site and one that hybridises to the PB transposon and extends into the genomic sequence. Homozygous mutants should only amplify the PB-

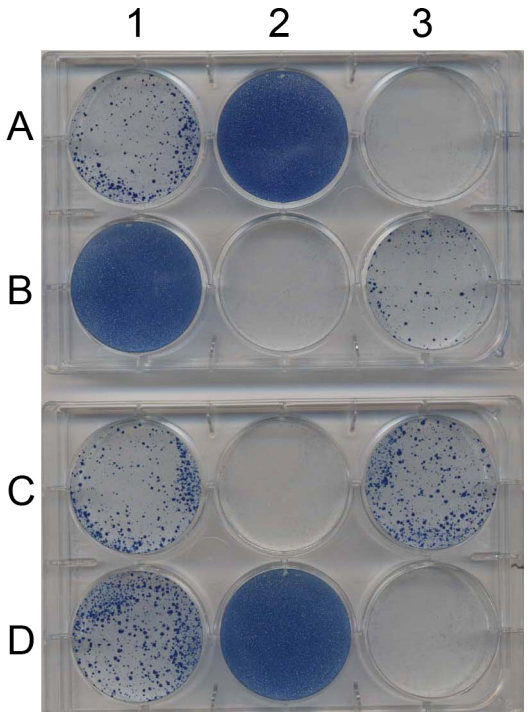


Figure 5.5: Typical results of double drug selection. Three classes of clone are visible: Those for which most cells plated are double resistant (e.g. row A, well 2), Clones that yield no double resistant cells (e.g. row A, well 3) and clones with varying numbers of double resistant colonies (e.g. row A, well 1; row B, well 3).

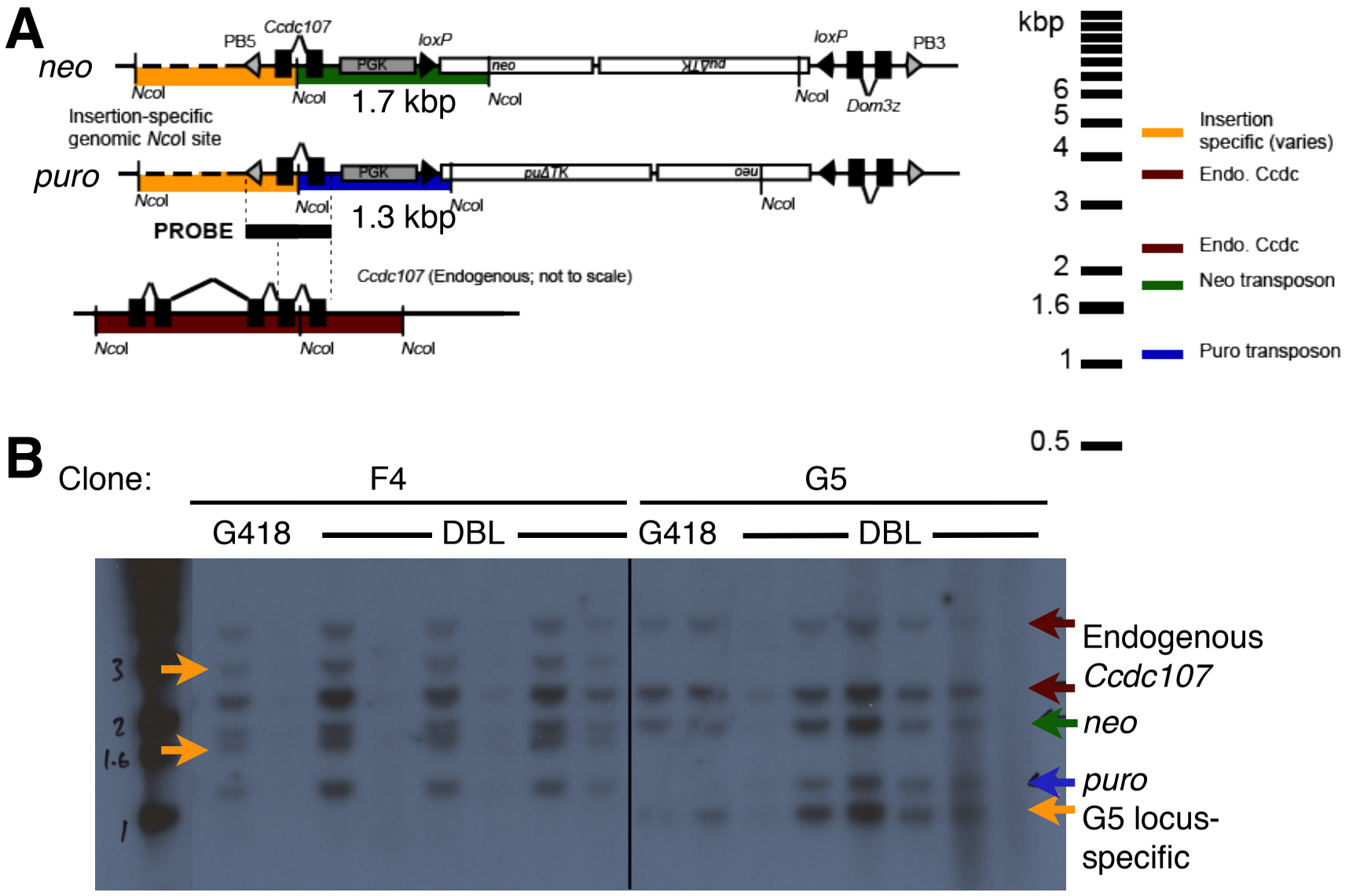


Figure 5.6: A—Probe used and predicted bands (see text) B—Examples of a clone containing two insertions at different loci (F4, left) and a potential homozygote with both *neo* and *puro* transposons at a single locus (G5, right).

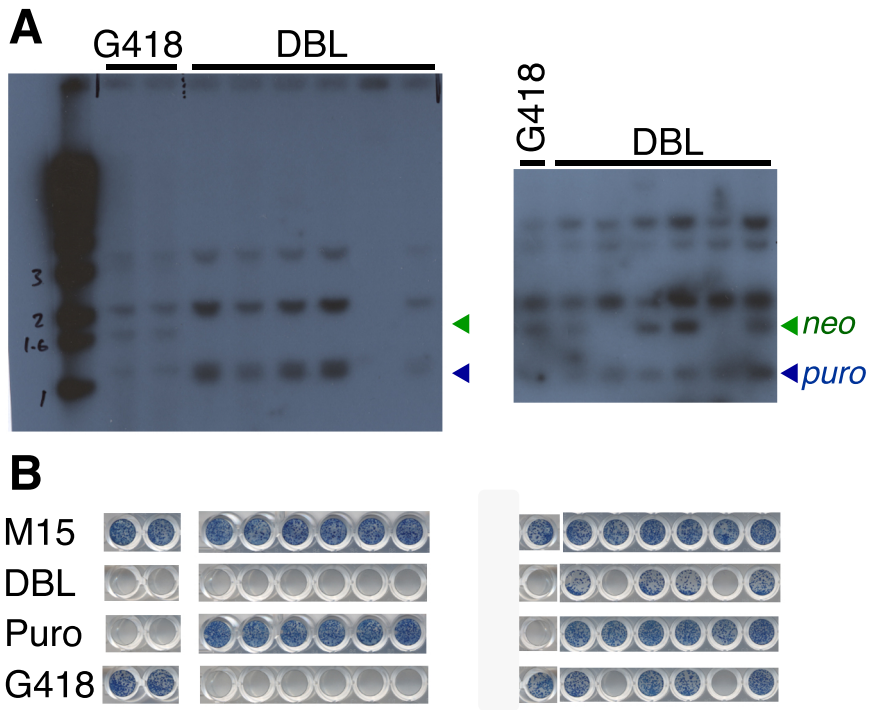


Figure 5.7: A—G418 and double-resistant (DBL) subclones from two separate clones, showing lack of TNN (*neo*) band. B—These clones show sensitivity to G418 and DBL when reselected.

genome junction product. A typical result is shown in Figure 5.8 for clone G5 with an insertion in the *Dymeclin* (*Dym*) gene. Three subclones do not amplify a wild type band, and are therefore homozygous. However, three subclones with an identical Southern blot pattern indicating two transposons at the *Dym* locus also amplified a clear wild type band. This suggests that more than two *Dym* alleles may be present in these cells—two mutant and at least one wild type.

I repeated the expansion and double selection procedure to obtain more double resistant cells for study and to ensure these results were not due to incomplete selection in the experiment above. Conditions for expansion and selection were the same, although I used electroporation to transfect the Cre plasmid. This time all double resistant clones had both *neo* and *puro* bands when analysed by Southern blot as above (Figure 5.9). This indicates that the selection worked effectively this time, when the cells grew at a normal rate. To simplify the process of isolating a larger set of double resistant clones for analysis, I also used a *Blm*-deficient cell line (NRB2) expressing a 4-hydroxytamoxifen (4-OHT) inducible Cre protein (see Chapter 2). This allows shorter expansion times to be used, as Cre induction is very efficient even in small cultures.

5.2.6 Summary of isolated double resistant clones

Altogether I isolated double resistant cells from 16 clones (Table 5.1). However, the results of PCR genotyping showed that some double resistant subclones still retained the wild type locus (Figure 5.10). The double resistant subclones generally comprised a mixture of genuine homozygous cells and cells that retain a wild type band in the PCR assay. Differences in expansion time, locus or Cre provision method did not appear to affect the general pattern, although these results do not allow this to be analysed systematically. The average clonal proportion of homozygotes obtained in all of these experiments was 34%, although as can be seen from the table, this can vary from 0–100%.

5.2.7 Double resistant clones retaining a wild type locus

It is possible that the wild type band in these PCR assays arises from a small proportion of wild type cells in the culture, either leftover feeder cells or cross-contamination from another mutant. To ad-

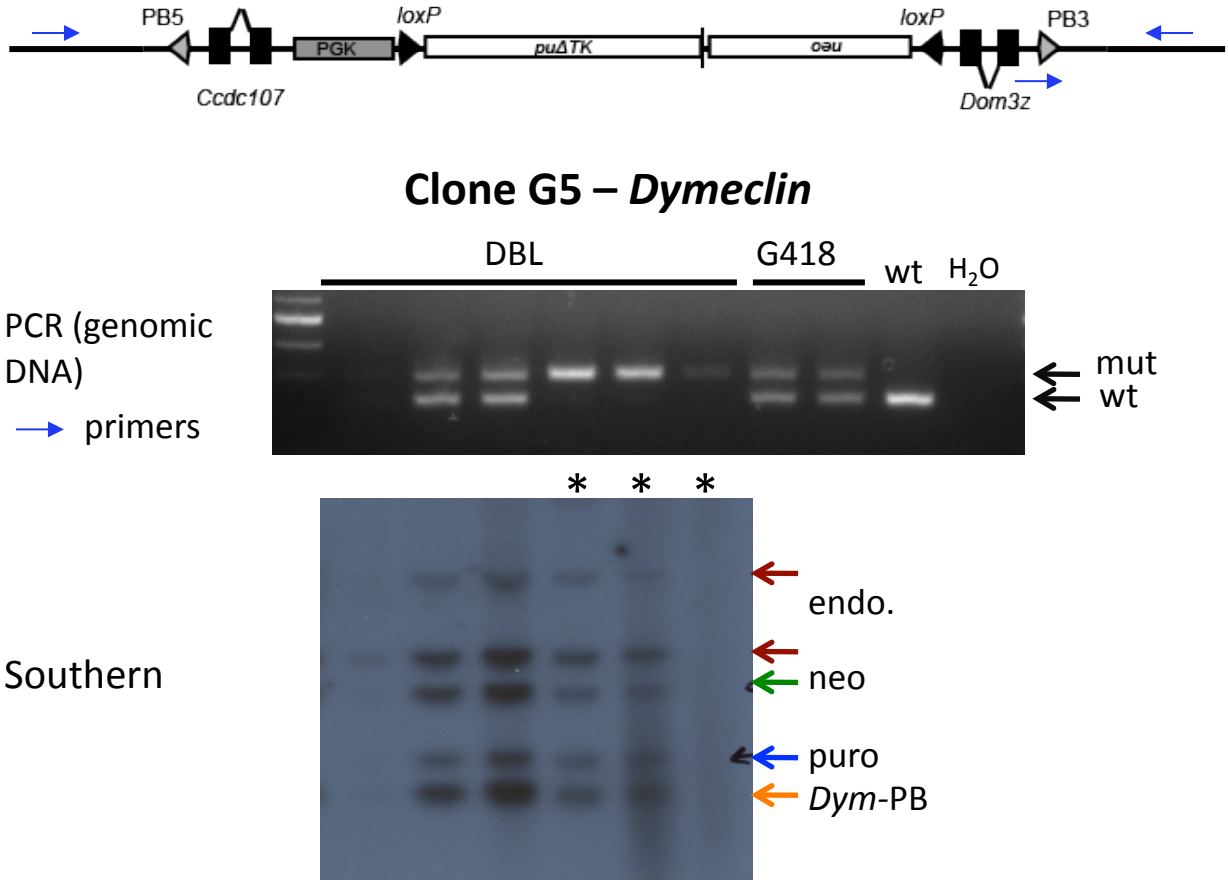


Figure 5.8: A—PCR using primers flanking insertion site and a transposon primer, indicated by arrows. In this case the mutant band is larger than the wild type band. Double resistant (DBL) clones 4–6(∗) only amplify the mutant band. B—Southern blot of the same subclones (reproduced from Figure 5.6).

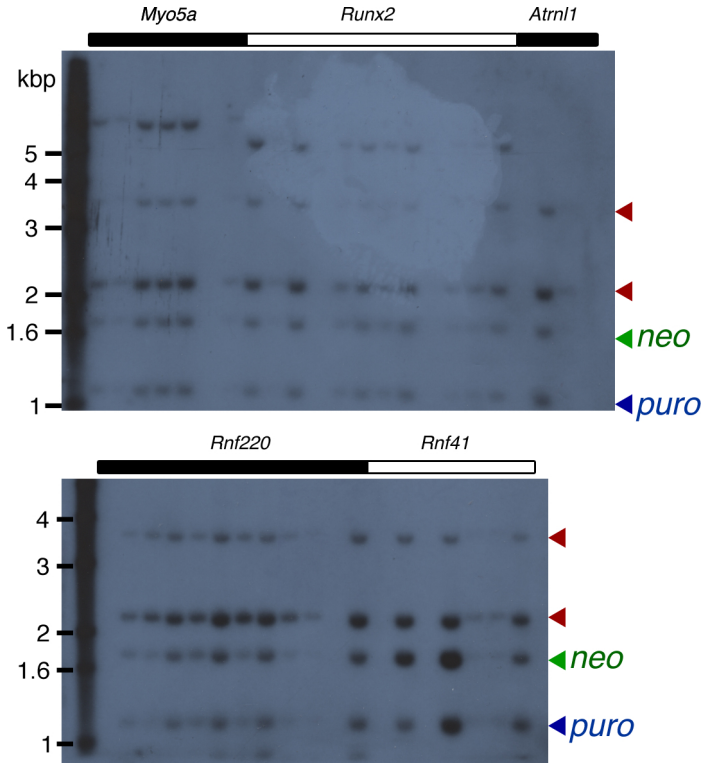


Figure 5.9: No selection background under normal selection conditions. Results for double resistant subclones from five clones with the indicated locus of insertion are shown, using the same Southern blot scheme as in Figure 5.6

Locus	Chr	Co-ordinates	Cre	Expansion	−/−	−/−/+$_n$	Fraction
Runx2	17	44,913,586	Electroporation	6w	10	0	100%
No gene (FL2.A11)	8	127,801,265	ERT2	96w	8	0	100%
Myo5a	9	74,990,083	Electroporation	6w	6	0	100%
Acpp	9	104,240,468	ERT2	96w	6	1	86%
No gene (FL2.C6)	9	15,258,570	ERT2	24w	4	3	57%
Dym	18	75,432,480	Lipofection	6w	3	3	50%
Arrb2	11	70,249,198	Electroporation	6w	2	4	33%
Rnf220	4	117,117,646	Electroporation	6w	2	6	25%
Sall1	8	91,577,321	ERT2	96w	2	5	29%
No gene (FL2.C1)	8	127,801,265	ERT2	24w	2	6	25%
Ddt	10	75,236,416	Lipofection	6w	1	3	25%
Itgb6	2	60,436,094	ERT2	24w	1	6	14%
Picalm	7	97,279,369	Electroporation	6w	0	1	0%
Atrn11	19	57,986,911	Electroporation	6w	0	1	0%
Macrod2/Flrt3	2	140,500,155	Electroporation	6w	0	0	N/A
Rnf41	10	127,863,548	Electroporation	6w	0	0	N/A
Srgap3	6	112,750,941	ERT2	96w	0	6	0%
Myh7b	2	155,429,901	ERT2	96w	0	6	0%
Tera	6	148,887,008	ERT2	96w	0	6	0%
Mrp133	5	31,916,879	ERT2	96w	0	6	0%
Tera	6	148,887,008	ERT2	96w	0	4	0%

Table 5.1: Results of genotyping for all double resistant clones

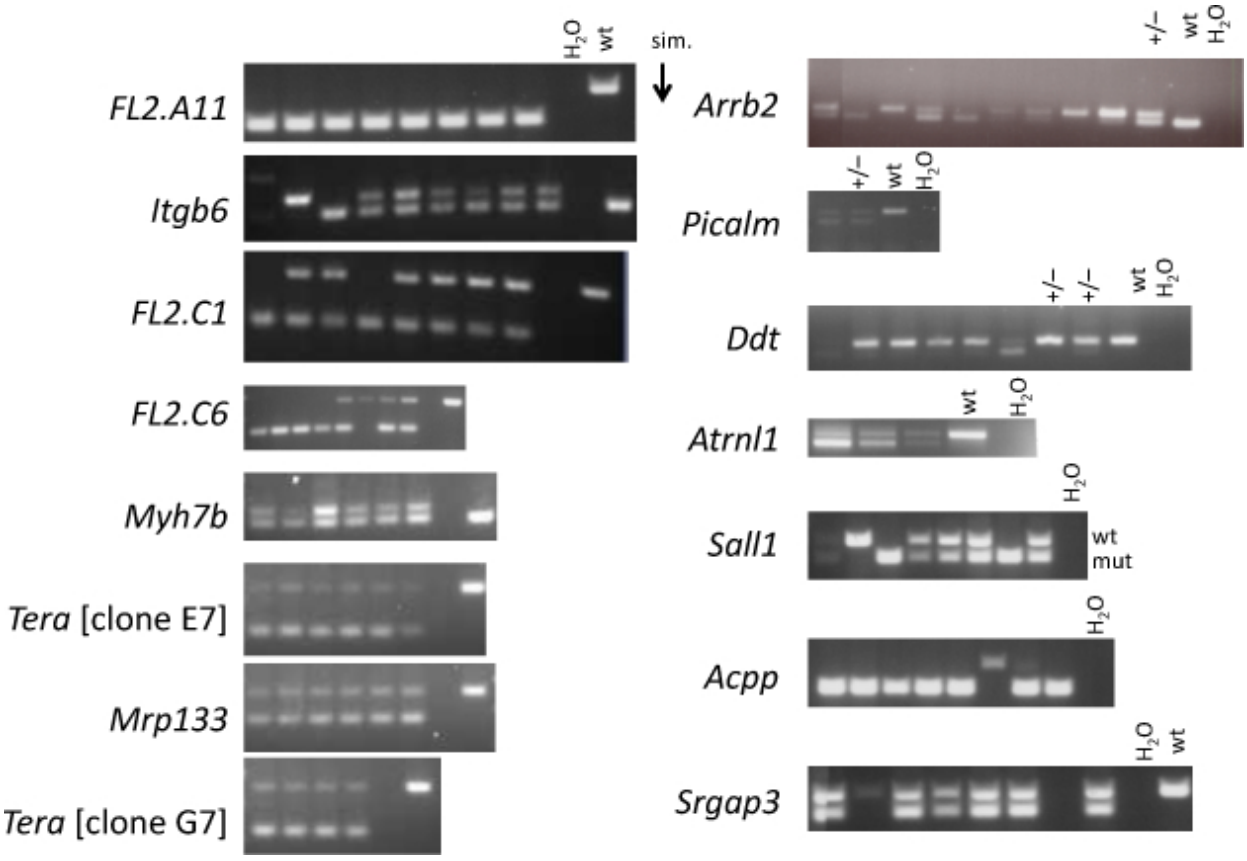


Figure 5.10: PCR genotyping of double resistant subclones from clones with an insertion at the indicated locus. Primers are designed to flank the insertion site, with an additional transposon primer as in Figure 5.8. H_2O, PCR without template; wt, PCR using wild type template DNA; +/–, PCR using DNA from cells heterozygous for the specific insertion.

dress this, I used Southern blotting with a probe specific to the individual insertion site, rather than the general transposon probe above. As the signal from a Southern blot is directly proportional to the amount of DNA this gives a more accurate representation of the relative amounts of mutant/wild type chromosomes in the culture. I analysed several clones in this way by stripping the original blot and reprobing with a probe designed to detect a different sized band for the wild type locus, the mutant with the *neo* transposon and the mutant with the *puro* transposon (Figure 5.11). For the *Myo5a* mutants, all subclones were homozygous as expected from the PCR result. In the case of the *Runx2* mutants, no wild type band was detected on the Southern blot, despite a clear band in the PCR assay. Therefore these are likely to be true homozygous mutants, and the wild type band is likely to result from contaminating cells below the level detectable by Southern blot.

Most interesting were the subclones from the *Rnf220* mutants for which three different classes can be seen on the blot (Figure 5.11B, right). Three bands were seen in wild type retaining clones (Figure 5.11B, (i)), corresponding to the *neo*, *puro* and wild type loci. The wild type band was approximately twice as intense as the others, indicating a ratio of 2:1:1 wild type:*neo*:*puro* chromosomes, and therefore possible tetraploidy. Two subclones homozygous by PCR assay were confirmed as such (Figure 5.11B, (iii)). These results show that two separate outcomes are possible after double selection, copy number increase with loss of wild type locus, presumably by *Blm*-related LOH, and copy number increase with retention of the wild type locus, which may be by acquisition of an abnormal karyotype.

These locus specific blots also highlighted the shortcomings of using PCR to assess homozygosity. The *Runx2* clones that gave a wild type band in the three-primer PCR were in fact homozygous when assessed by Southern blot. PCR is a much more sensitive technique than Southern blotting, so a small amount of contamination by wild type cells (which could be ES cells or cells from the feeder layer) may result in a wild type PCR product. Such low level contamination would not be detected on a Southern blot, where signal is directly proportional to the amount of DNA present. Therefore PCR genotyping alone may underestimate the real number of homozygous mutants, as in the case of the *Runx2* mutants in Figure 5.11 (middle).

Finally, some clones showed only a wild type band in PCR genotyping: for example *Myo5a* clone 6, *Runx2* clones 4 and 9 and *Rnf220* clones 4 and 5. In some cases (*Myo5a* and *Runx2*) no or very little DNA was isolated from these wells when I prepared DNA for Southern blots, so it is likely that these clones did not survive. When picking colonies I made a conscious effort to pick all kinds of morphologies, as to only pick "healthy looking" or large colonies may inadvertently select against genuine mutants. The wild type band in these cases where no ES cells grew may result from leftover feeder cells. However in the case of the *Rnf220* mutants, these "wild type only" subclones do show signal on the Southern blot, but do not in fact have an insertion at the *Rnf220* locus (see locus specific blot, Figure 5.11B(ii) and A. In part A the *Rnf220*-specific band is just visible at the bottom of the blot and appears to be absent in lanes 4 and 5). Therefore these may have arisen from mosaicism in the clone, despite the replating step.

As the result from the locus specific Southern blot indicated that the wild type retaining subclones may be tetraploid, I prepared metaphase spreads to check the karyotype of these subclones.

5.2.8 Karyotype of wild type retaining clones

Metaphase spreads prepared from wild type retaining subclones showed a clear near-tetraploid karyotype, whereas the genuine homozygotes isolated from the same clones (*Rnf220* and *Sall1*) had a normal diploid karyotype (Figure 5.12). Therefore in this case a change in ploidy had resulted in the transposon copy number increase that was then selected for. As both the diploid homozygotes and these tetraploid "wild type retainers" originated from a single cell with the PB insertion, this starting cell must have been euploid, and both LOH and ploidy changes must have occurred during the expansion phase.

5.2.9 DNA content analysis of wild type retaining subclones

As the wild type retaining double-resistant subclones examined above were tetraploid, I decided to explore whether these could be discriminated by DNA content analysis, as if this were possible then fluorescence activated cell sorting (FACS) could potentially be used to isolate the double resistant cells with a normal DNA content—i.e. homozygotes. Staining fixed nuclei with the DNA binding dye propidium iodide effectively discriminated the known near-tetraploid subclones from normal diploid cells (Figure 5.13A,B). However, running this analysis on a

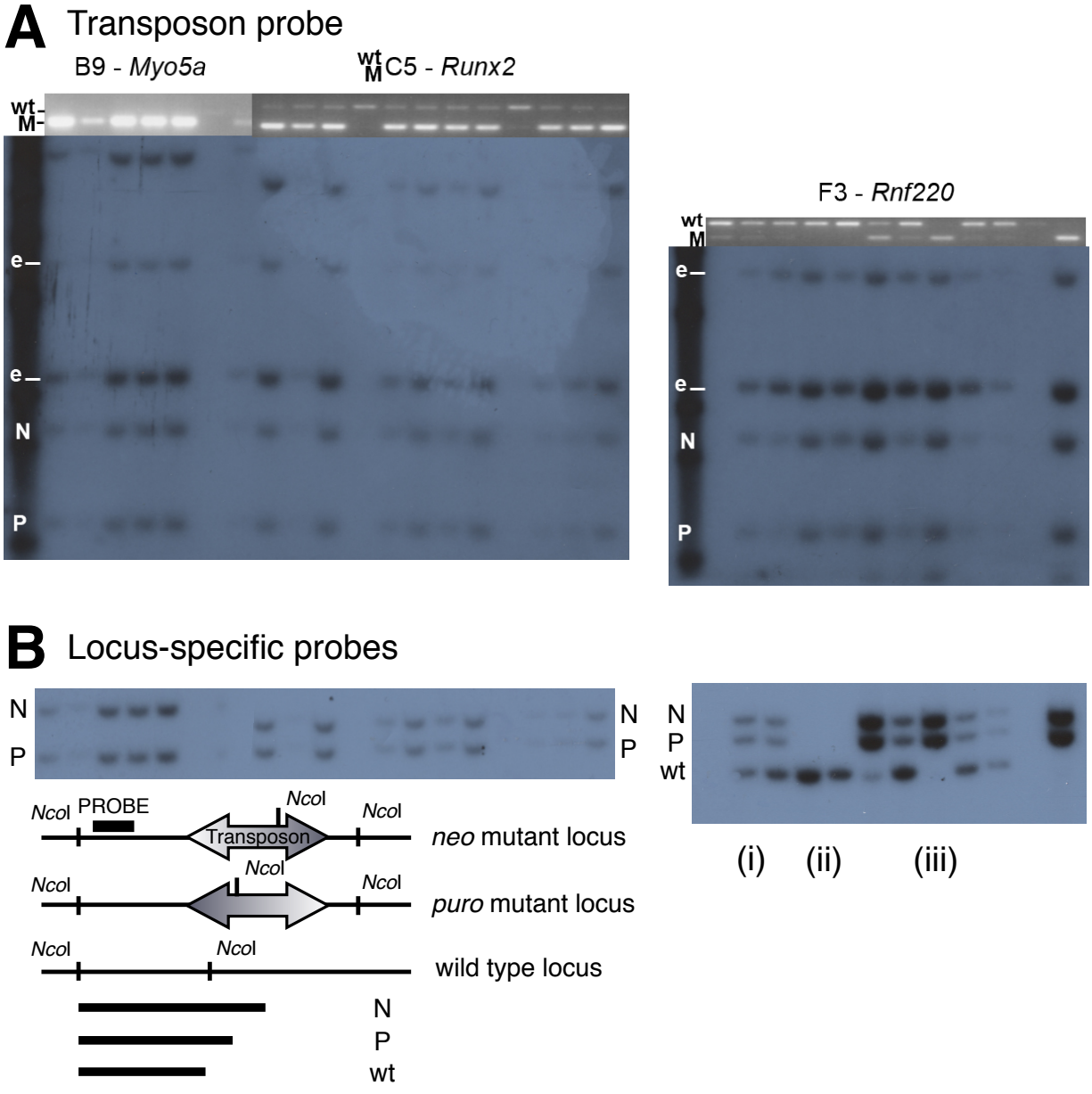


Figure 5.11: Double resistant cells analysed by Southern blot. A—Genomic PCR (original genotyping) and blot with transposon probe as in Figure 5.6. B—Reprobing of blot with a locus specific probe designed as shown. Two bands are seen for homozygotes, three for clones that genuinely retain the wild type locus. (i)–(iii): three genotype classes for *Rnf220* mutants; see text

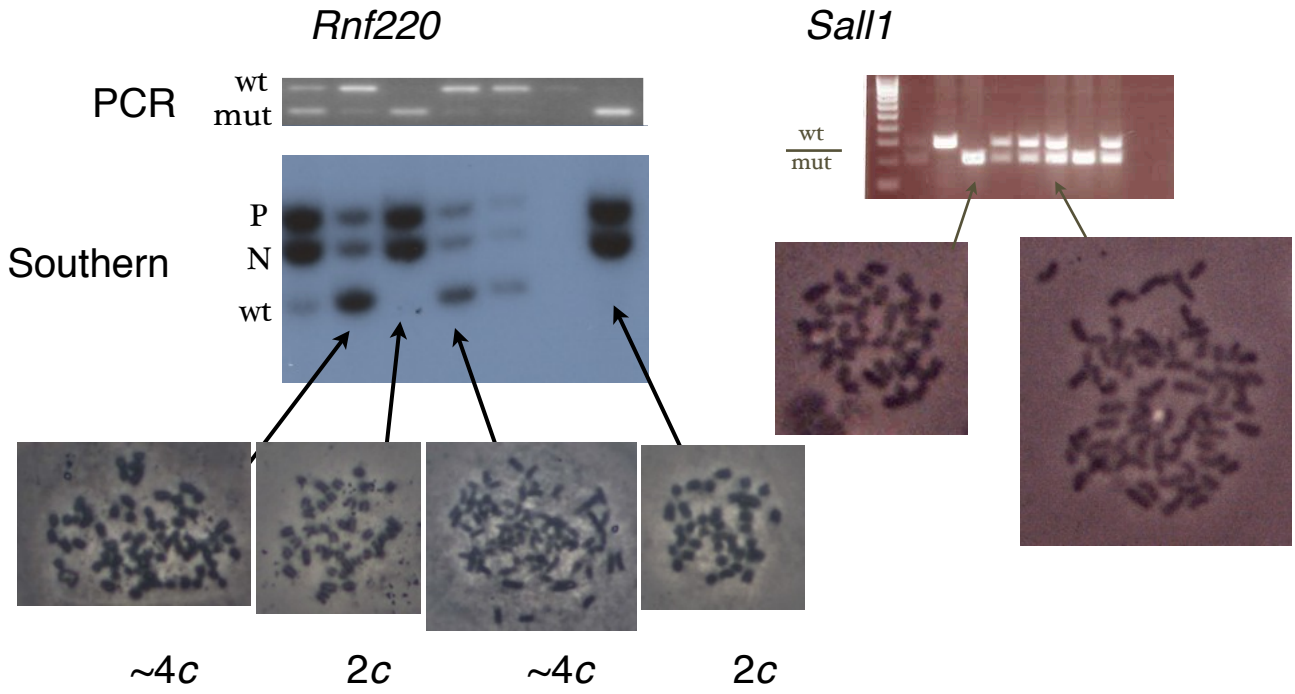


Figure 5.12: Representative chromosome spreads shown for the indicated clones. Spreads from clones that retain a wild type locus in the genotyping assays are tetraploid, whereas homozygous sister clones are euploid.

larger set of clones that had been determined to retain the wild type locus by PCR showed that most of these actually had a staining profile that resembled that of the known diploid subclones (Figure 5.13C).

It is possible that these clones had a less severe chromosome abnormality, such as a trisomy of the chromosome with the insertion or a segmental duplication. Indeed, a colleague's (Y. Huang, personal communication) double selection experiments isolated one such trisomic clone. Alternatively, the PCR assay used may be giving false negative results due to low level contamination and these clones may in fact be genuine homozygotes. The only way to be certain is to do the type of locus-specific Southern blot experiments above, which is labour intensive even on this small scale, and completely impossible on a genome wide scale.

5.2.10 The transposon disrupts transcription of genes when inserted into introns

The homozygous mutants isolated above gave me the opportunity to see if my transposon vector was mutagenic. I prepared RNA from double resistant subclones from three separate mutants with insertions in an intron—*Dym*, *Arrb2* and *Myo5a*. Using oligo-dT primers, I prepared cDNA by reverse transcription and used primers to exons flanking the intron with the insertion to see if a transcript was detectable. All clones that had been determined to be genuine homozygotes failed to amplify a PCR product (Figure 5.14). Therefore the transposon construct is mutagenic at the mRNA level. As this was the case in all three randomly picked insertion sites, the construct is likely to be mutagenic in most cases in which the insertion is in an intron.

5.3 Discussion

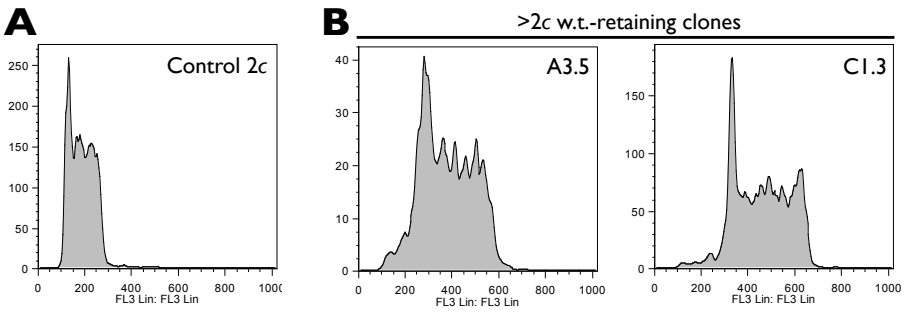
5.3.1 Paths to increase transposon copy number in *Blm* cells

Following the scheme above, I successfully isolated double resistant cells for many clones. Selection for cells with both *neo* and *puro* versions of the transposon was faithful, as on Southern blots I observed no background clones with only one version of the transposon. As expected, some clones initially contained two copies of the transposon due to the plasmid mobilisation system used. This would explain the clones that gave very large numbers of double resistant cells, although some clones which gave few enough colonies to pick also proved to have two insertions (Figure 5.6). This may reflect poor Cre efficiency at that particular locus, or poor Cre transfection efficiency for those clones.

Both genuine homozygous mutants and wild type retaining subclones were generally isolated from the double resistant population. The wild type retaining clones have increased the transposon copy by a non-LOH pathway, seemingly numerical chromosome instability (CIN). As both euploid and aneuploid cells were isolated from the same clone, which began as a single transfected cell, the original cell is likely to have been euploid. Therefore LOH and numerical CIN are competing pathways for transposon copy number increase in *Blm*-deficient ES cells. As both classes of double resistant subclones occur with similar frequencies, it could be inferred that the two processes have similar rates. However many of these experiments used relatively long expansions, and it is possible that tetraploid or trisomic cells may grow faster, as has been reported for some trisomies (Liu *et al.*, 1997). This would lead to increased representation in the selected population. Equally, some mutants may be at a fitness advantage or disadvantage, so the proportions of mutant and wild type retaining cells in the final population may not directly reflect the rate at which they arose.

5.3.2 Clones for which double resistant cells were not isolated

For 17 out of the 42 of the clones tested, no double resistant clones were isolated. This is unexpected, as even if the LOH rate at these loci is very low (e.g. if they are very close to the centromere), my results show that tetraploidy and trisomy are possible methods to acquire double resistance. Tetraploidy affects every chromosome. Therefore no location should be immune to copy number gain by this mechanism. Although it would have to be quite serious, a Cre position effect is a possibility. More likely is that LOH/other copy number gain is sufficiently rare for it not to occur in some cases, even though the expansion is quite prolonged in these cases. It is also possible that the gene is homozygous lethal when mutated, but even in this case it should still be possible to isolate aneuploid cells. For both cases where I had mapped the insertion (for the ERT-Cre experiments I only mapped the insertions after the double selection) but failed to isolated any double resistant cells, the insertions were on chromosomes for which I had previously isolated wild type retaining cells (*Macrod2* and *Rnf41*, Table 5.1). Thus, there does not appear to be a barrier to isolating



C

Clone	# wt-retaining subclones	2c	>2c	ND
A3	5	0	3	2
A4	1	1	0	0
A7	6	5	0	1
B4	6	5	0	1
C1	6	5	1	0

Figure 5.13: A—Wild type control DNA content profile. B—Examples of clones with near-tetraploid DNA content. C—Most wild type retaining clones have a near-2*c* DNA content.

cells with abnormal copy number of these chromosomes.

One further explanation could be a position effect with respect to expression of the resistance genes. Some loci may express sufficient levels of puro protein but not neo, due to their chromatin context or the influence of nearby regulatory elements. Although the two resistance genes are under the control of the same promoter and polyadenylation signals, the stability of mRNA and protein, and the amount required to confer resistance, is likely to differ.

5.3.3 Implications for creation of homozygous mutant libraries

In these experiments, where the expansion and double selection steps were done on a clone-by-clone basis, a mixed double resistant population was obtained in most cases. In most cases, the mixed population would not contain a sufficiently high proportion of homozygotes for genetic screens. Whether a proportion of 34% (the per clone average) would be sufficient to see a loss of function phenotype will depend on the assay used. For optimum performance, the double resistant population would have to be subcloned in order to create an arrayed library of pure mutant cells for genetic screens.

There is no way to select against cells with a wild type allele on a general basis. Therefore, to make a clonally pure library using the methods described here, double resistant subclones would need to be genotyped in order to identify the homozygous mutants. This also means that each insertion site would have to be mapped and a separate genotyping protocol designed. From a practical point of view, the effort required would be similar to serially targeting all known genes using the targeting vector and heterozygous ES cell resources that are quickly becoming available (International Mouse Knockout Consortium *et al.*, 2007). Ideally the library generation step would generate clones that could be picked and screened directly, and the insertion site only mapped once mutants of interest had been isolated.

An alternative strategy might be to reduce the expansion time to a critical level, such that only one LOH/CIN event is expected to occur (Figure 5.15). This would represent an expansion to a few thousand cells, roughly corresponding to a colony just visible to the naked eye. In this situation, a heterozygous clone would only rarely give rise to the mixed double resistant population and instead produce a double resistant population composed of either all homozygotes or all wild type retaining cells

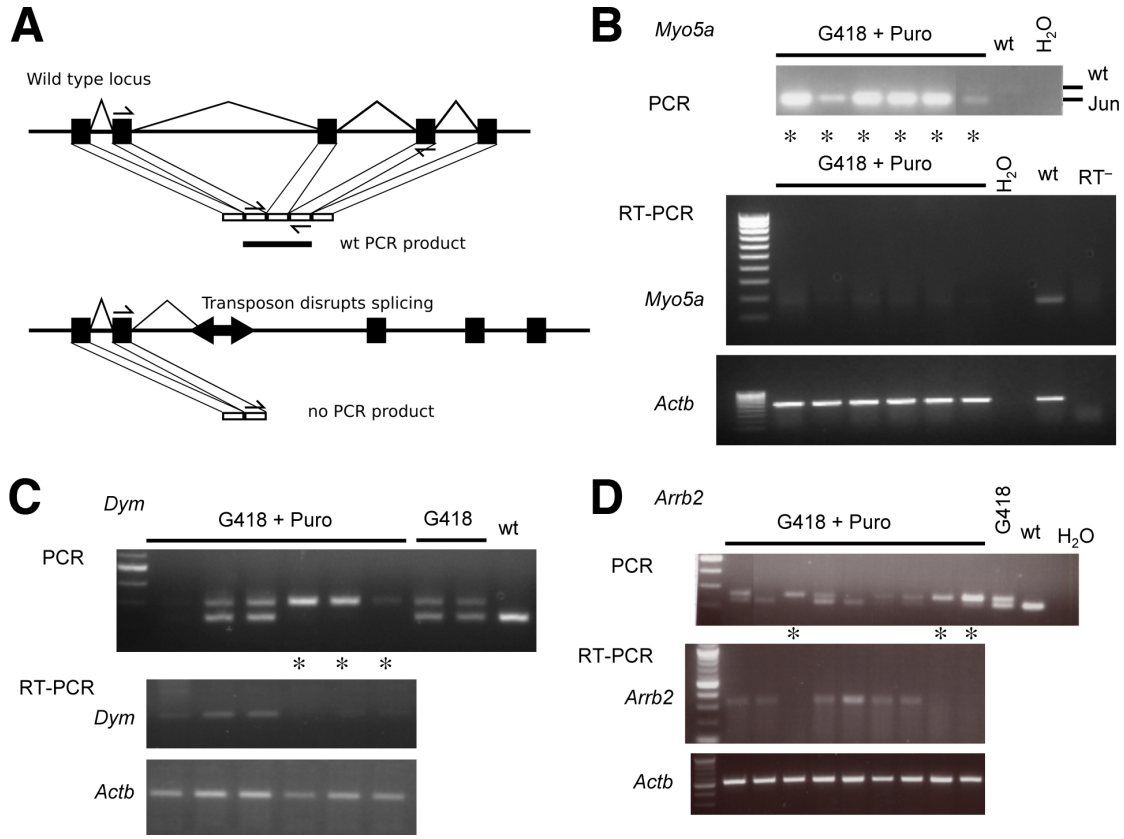


Figure 5.14: RT-PCR analysis of cDNA prepared from double resistant subclones. The corresponding genotyping PCR on genomic DNA is shown at the top of each sub-figure. RT-PCR reaction shown below, with *Actb* positive control at the bottom. A—*Dym*. B—*Myo5a*. C—*Arrb2*.

(Figure 5.15B,C). In this situation, the double resistant cells would be a pure population that could be screened directly. Retaining cells would become "passengers" in the library, so it is essential that the frequency of homozygotes obtained at this limit is high enough to give a complex and useful library.

However, using this limited expansion, the probability of isolating any double resistant cells at all from a given clone would also drop. For this reason it may be better to carry out the expansion and double selection in parallel in a pooled format, to avoid expanding many individual clones that do not yield double resistant cells. This would require the copy number of the transposon to be strictly limited to one at the start, so that there would be no clones with two copies from the beginning of the expansion, which would dominate over the low number of homozygous cells. Targeting the transposon to the X chromosome and mobilising from there would be one way to do this.

5.3.4 Conclusions

In this chapter I have demonstrated that the TNN-TNP inverter construct can be used to isolate homozygous mutants from expanded populations of *Blm*-deficient ES cells. Additionally, the construct disrupts transcription when inserted into introns and is thus likely to be an effective mutagen. An alternative pathway to increase transposon copy number exists via numerical chromosomal instability, thus the double resistant population is not purely homozygous mutants. The average clonal proportion of homozygotes was 34%, representing a significant enrichment for homozygous mutants. The clone-by-clone method for homozygote enrichment described in this chapter requires two subcloning steps to obtain pure homozygous populations, which is not practical on a genome wide scale. The next steps, described in Chapter 6 were to make a suitable transposon donor locus on the X chromosome to limit the initial copy number for library generation on a large scale.

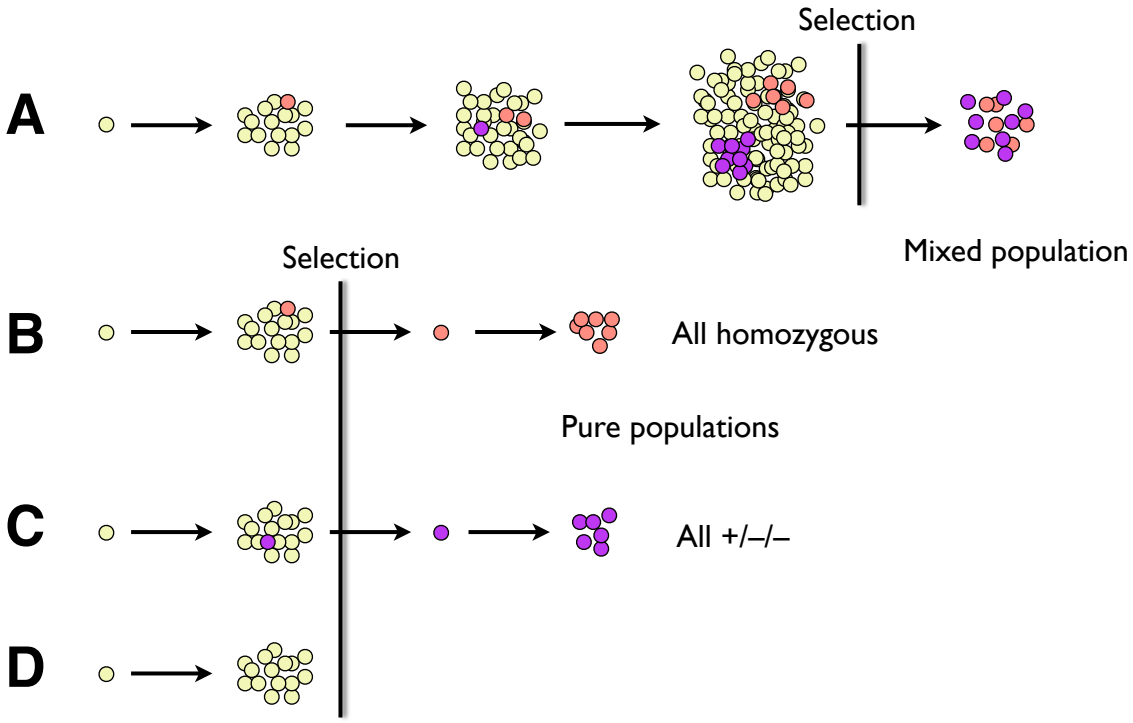


Figure 5.15: Possible consequences of over-expanding clones prior to double selection. A—Long expansion times allow two or more events to occur, resulting in a mixed double-resistant population. B, C—Ideal situation where only one event occurs, resulting in pure clonal double resistant populations. D—However, in many cases no LOH or copy number gain will occur.

Chapter 6

Isolating large numbers of homozygous mutants in parallel

6.1 Introduction

6.1.1 Aims

In the previous chapter I described the isolation of homozygous mutants by selection for copy number gain in *Blm* deficient cells. Analysis of the selected population revealed that not all the surviving cells are homozygotes, although all have two copies of the selection construct. Thus, using the scheme described, two subcloning steps are necessary to obtain pure clonal homozygous mutants for screening: one initial step to isolate a colony for expansion, and one after double selection to obtain homozygous mutants. This is not practical on a genome wide scale. In this chapter I describe several technical advances to solve this problem by allowing the initial subcloning step to be avoided. Mutants are expanded in a pool, in parallel, and Cre treatment and double selection carried out on this pool. Mutant clones are picked directly from the double selection, resulting in a complex pool with low background. I discuss uses of the libraries generated in this way.

6.1.2 Clonal expansion

When working with a complex pool of cells that will be expanded and selected over a period of time, the problem of clonal expansion arises. This is where cells that will survive the late selection step are present early in the culture, and thus have time to expand to a clone that dominates the selected population (Figure 6.1). The selection steps need to be absolutely stringent, and furthermore the full range of events occurring in the culture needs to be known. The reason for this is best illustrated by considering some numbers. In the homozygote isolation process, there are two selection steps: One to select for a PB integration (G418 for 10 days), and one double G418+Puro step (A further 10 days). A cell that had two integrations will be potentially double resistant from the start, and thus at the time of double selection this clone will have $2^{10} = 1024$ cells. At the opposite end of the scale, a clone undergoing LOH only in the generation immediately prior to selection will have just a single potentially double resistant cell. Thus, if these two clones had been pooled at the start of the experiment, the (useless) cells in the double resistant population with two insertions will vastly outnumber the real homozygote.

6.2 Results

6.2.1 C57BL/6 targeting vector to insert the transposon at the *Hprt* locus.

Choice of cell line

In order to limit the initial copy number of the transposon immediately after mutagenesis, I decided to target it to the X chromosome. Most ES cell lines used for making genetically modified mice are XY, and thus should maintain all loci on the X chromosome outside the pseudoautosomal region as single copy. The obvious choice of locus is *Hprt*, as targeting vectors are available, and the transposon could be inserted such that it disrupts *Hprt* function. Transposition would then be coupled to restoration of *Hprt* function, allowing it to be selected for with HAT.

However, the original *Blm*-deficient cells were from the AB2.2 cell line. This contains a complex mutation at *Hprt* that is not revertible. Therefore the endogenous *Hprt* gene cannot be used for HAT selection in these cells; the usual procedure is to use a human *HPRT* minigene. Because of this mutation, AB2.2-derived cells are not ideally suited for HAT selection in my proposed context. I had originally cloned my transposon into the intron of the *HPRT* minigene, and was planning to insert this at the endogenous (non-functional) *Hprt* locus (see Chapter 3).

Another problem with the original *Blm*-deficient ES cells is that they are a compound heterozygote with respect to *Blm* (Luo *et al.*, 2000). Furthermore, there is evidence that one of these alleles (*m3*) is a hypomorph (McDaniel *et al.*, 2003). This is not necessarily a problem, as the cells do show a Bloom syndrome phenotype, but it is an aspect that could

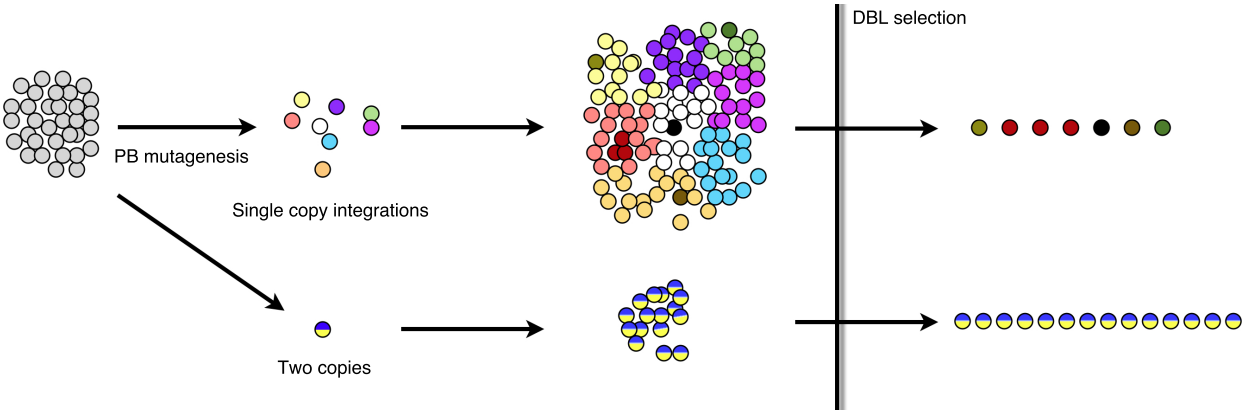


Figure 6.1: Problems caused by clonal expansion in heterogeneous pools of cells. The effect of having two transposon copies present in a minority proportion of cells prior to expansion is illustrated. Different colours represent clones of cells with different transposon insertion sites. All cells in the clone with two copies (bottom) can potentially survive double selection, indicated by a vertical line, while in the single copy clones only the rare homozygotes (represented by dark colours) survive. This leads to the clone with two copies dominating the double resistant population.

be improved on. One potential consequence could be that subclones arise in the culture that have undergone LOH at *Blm*, and therefore display slightly different phenotypes with respect to further LOH or genome instability. Recently, a new *Blm* mutant cell line, $Blm^{e/e}$ was generated in our laboratory by Amy Meng Li (Li, 2010). This cell line has two advantages: (a) it is derived from the JM8.F6 ES cell line (Pettitt *et al.*, 2009), and therefore has a functional *Hprt* gene and (b) the *Blm* locus is homozygous for a genuine null allele. It should be noted that the cells also express GFP and *bsd* constitutively from transgenes at the *Blm* locus; however this is not a problem for my method.

As this cell line is in the C57BL/6 genetic background, I constructed an isogenic targeting vector to insert my transposon (TNN) in such a way as to disrupt *Hprt*.

Retrieval of a *Hprt* fragment from a C57BL/6 BAC

I obtained the tiling path BAC RP23-252K15, which contains the complete *Hprt* coding sequence, and used it to transform EL350 bacteria (see Methods). Using long oligonucleotide primers, I amplified a pBS backbone with 70 bp homology arms. This linear PCR product can be thought of as a circular plasmid, containing two colinear fragments homologous to *Hprt* separated by several kbp with a break between them (Figure 6.2A). When electroporated into recombination competent bacteria containing the *Hprt* BAC, the 'gap' is repaired using the BAC as a homologous template. This resulted in retrieval of the fragment defined by the homology arms into the plasmid backbone (Figure 6.2B,C).

There is a naturally-occurring *Hind*III restriction site in intron two of the mouse *Hprt* gene (NCBI m37 X:50,357,636). I used an adaptor oligonucleotide sequence to introduce a *Nsi*I site into this locus in order to clone TNN as a *Nsi*I fragment from the P2-HPRT-Tn plasmid (Chapter 3). This fragment contains 125 bp of sequence from outside the vector, derived from the human *HPRT* minigene intron. As *Nsi*I cleaves 41 bp inside the PB3 end of the transposon, I had this fragment synthesised with the adaptor oligonucleotide, so the correct PB3 sequence was reconstituted upon ligation (Figure 6.3A). I verified the correct orientation and PB3 reconstitution by restriction digest and sequencing (Figure 6.3B). This targeting vector is named TV28. I also verified the function of the transposon by a transposition assay in ES cells (Figure 6.3C).

Targeting ES cells

I electroporated 10^7 $Blm^{e/e}$ ES cells with 15 μg *Srf*I-linearised TV28 (see Methods). G418-resistant clones were picked after ten days and screened for correct targeting events by PCR directly from colony lysates (Figure 6.4). Ten out of 36 tested were correct, nine of which I tested for targeting at the other junction by Southern blot (Figure 6.5A). All of these confirmed correct targeting, giving a targeting fre-

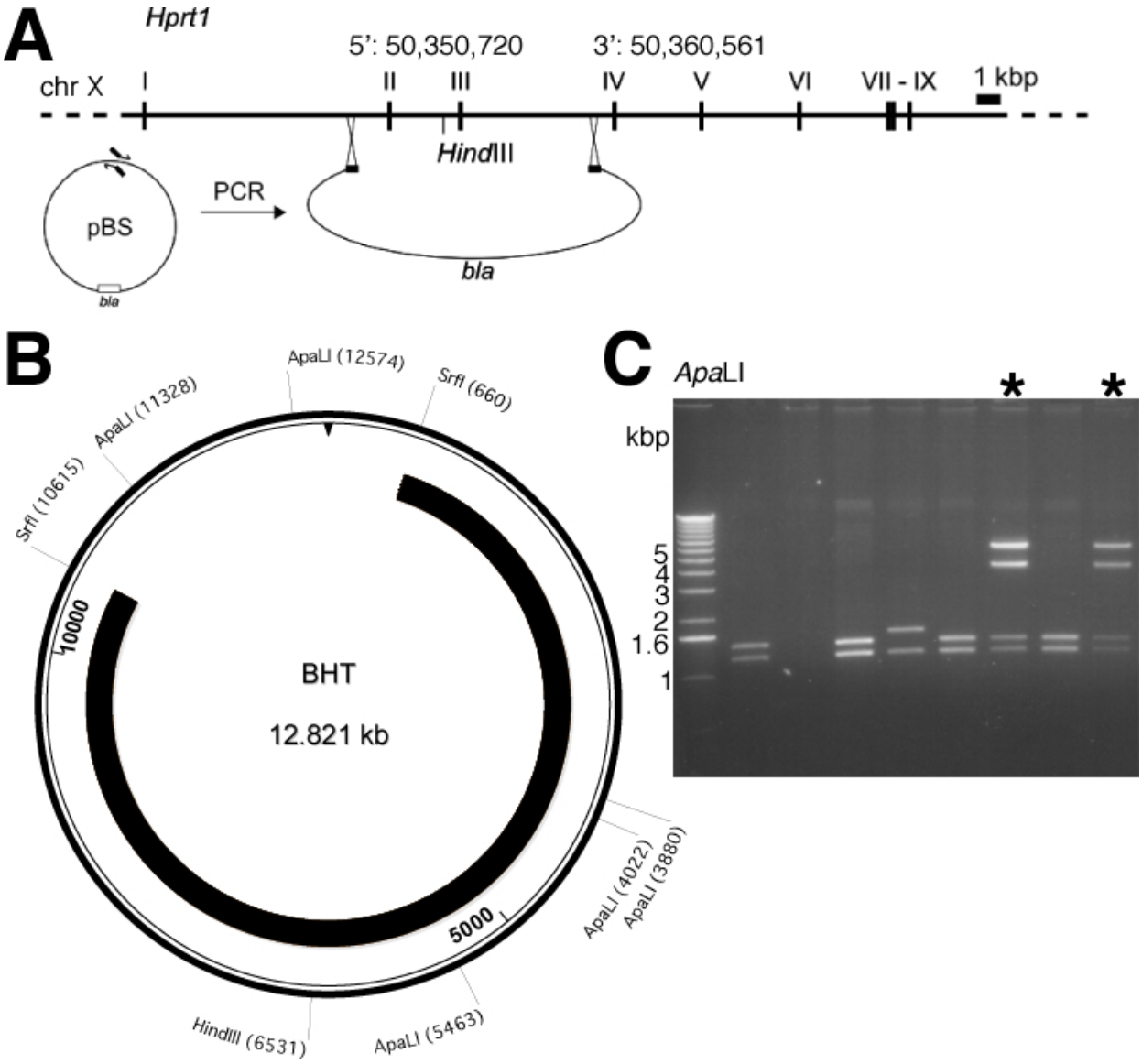


Figure 6.2: Retrieval of a fragment of the C57BL/6 *Hprt* locus. A—Design of the capture vector, produced by amplification of a pBS plasmid backbone with primers tailed with appropriately oriented homologous sequence. B—Map of the plasmid with retrieved fragment. *Srf*I sites were present on the primers used and can be used to linearise the construct for targeting. C—*Apa*LI digest showing correct structure of retrieved fragment. *—correct clone; other clones have only the *Apa*LI bands from the plasmid backbone and may arise from contaminating circular plasmid or recombination-induced recircularisation of the capture vector.

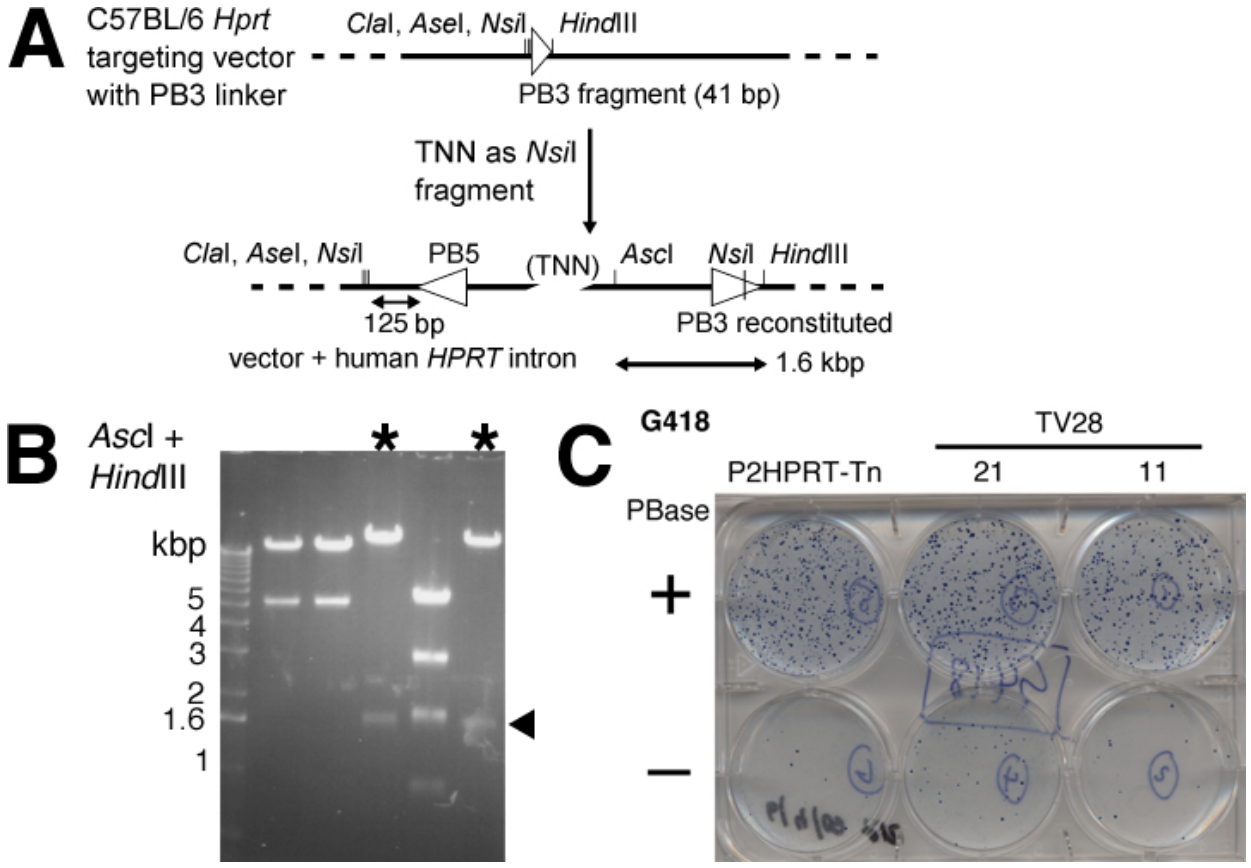


Figure 6.3: Cloning the TV28 targeting vector. A—Cloning scheme. A linker was inserted into the *Hind*III site in the *Hprt* targeting vector, containing part of PB3 distal to the *Nsi*I site. The TNN transposon was cloned as an *Nsi*I fragment from P2-HPRT-Tn. B—Screen of clones from the ligation. Asterisk (*) marks correct clones. Arrowhead indicates the 1.6 kbp band showing that the insert is in the correct orientation to reconstitute PB3. C—The transposon is reconstituted and functional in ES cells. Transient transposition assay with (+) or without (–) pCMV-hyPBase and using the indicated TV28 donor plasmids (two subclones, 11 and 21) or P2-HPRT-Tn.

quency of at least 25%. All clones also showed the expected HAT-sensitive/6TG resistant phenotype, although some contained a small proportion of HAT-resistant cells, presumably wild type (Figure 6.5B). Therefore, to remove the possibility of background HAT resistance from contaminating untargeted cells, I further subcloned the line prior to transposition for library generation. Transfection of subcloned cells with PBase resulted in HAT-resistant subclones as predicted, with no background observed in untransfected cells (Figure 6.5C). This cell line is named B6BTV.

Introduction of an inducible Cre gene into C57BL/6 *Blm* ES cells

In order to use short expansion times in homozygous mutant generation, it is important that the few homozygotes that do segregate in the limited expansion can be efficiently isolated. Transfection and expression of Cre is a limiting factor for the double selection procedure. To improve on this, I introduced an ERT2-Cre gene as for the 129 strain ES cells in the previous chapter. This time I used a vector obtained from Junji Takeda (Osaka University), which introduces an ERT2-iCre-ERT2 gene into the *Rosa26* locus. This gene consists of a mammalian codon-optimised (improved, iCre) Cre coding sequence fused to ERT2 at both termini (Shimshek *et al.*, 2002). Using two ERT2 moieties appears to reduce leakiness (J. Takeda, K. Yusa; personal communication). The targeting involved two steps: One to insert the transgene into the locus, and one to remove the F3-flanked *neo* selection marker using Flp recombinase (FlpO). To make a more generally useful cell line, I targeted this construct to the original $Blm^{e/e}$ cells and then retargeted the resulting cell line (named BRic, **B**L/6 *Blm* **R**osa26-**iC**re) with TV28 as above.

Results of the Cre targeting are summarised in Figure 6.6. I genotyped targeted clones using a PCR assay on colony lysates, which specifically detects the junction on the short arm side of the targeting vector (Figure 6.6A). Rapid PCR genotyping enabled me to directly expand correctly targeted clones, transfect them with a FlpO expression plasmid (PGK-FlpO, Raymond and Soriano (2007)) and screen unselected subclones by PCR for loss of *neo* (Figure 6.6B). As deletion of the *neo* gene was not selected for, it is probable that not all cells in the clone have undergone the deletion, as deletion could have occurred after the first division of the founding cell of the colony. This was evident from G418 sensitivity tests, therefore in order to make a stable line I subcloned some PCR-positive cells. The subclones were a mixture of deleted and undeleted cells as predicted (Figure 6.6C). In parallel to this subcloning, I took one clone that showed the fewest remaining G418-resistant cells for targeting with TV28 as above.

Library generation cell lines

To target the transposon to the BRic cell line, I used a different strategy to simultaneously recover a pure clone that had deleted the *neo* gene from the Cre targeting. I electroporated unsubcloned BRic cells with linearised TV28 as above, and plated the cells in 4-OHT containing medium for 24h. After three days, I replated the cells in puromycin. As the transposon in TV28 was originally in the *neo* orientation, using this scheme provides an internal functional test for inducible Cre activity in the cells and function of the loxP sites and resistance genes in the transposon. No puromycin resistant cells were obtained in a parallel experiment without 4-OHT treatment (Figure 6.7).

I verified correct targeting by PCR screening and functional (HAT sensitivity) test as before (Figure 6.8). Multiple targeted subclones were frozen. These cells were named BRic.TVP or, later, LGP (**L**ibrary **G**eneration **P**uro). In summary, these cells have all the elements desirable for making homozygous mutants easily, i.e.:

- Homozygous null mutation at the *Blm* locus.
- 4-OHT-inducible codon-optimised Cre expressed from the endogenous *Rosa26* promoter (heterozygous).
- TNP transposon integrated at *Hprt* intron two, excision selectable with HAT.
- C57BL/6 genetic background—the same as the reference genome—for easy mapping and comparability.

6.2.2 Generating libraries with the LGP cell line

The scheme for library generation is shown in Figure 6.9. I expanded two subclones, A2 and B4, of the LGP cell line independently to 6-well plates. I transfected adherent cells with 1 μg capped *in vitro* transcribed hyPBase mRNA using the TransMessenger lipofection reagent (Qiagen). The reason for using mRNA was to avoid the possibility of the transposase expression plasmid integrating into the genome of some cells (see Chapter 3). I replated

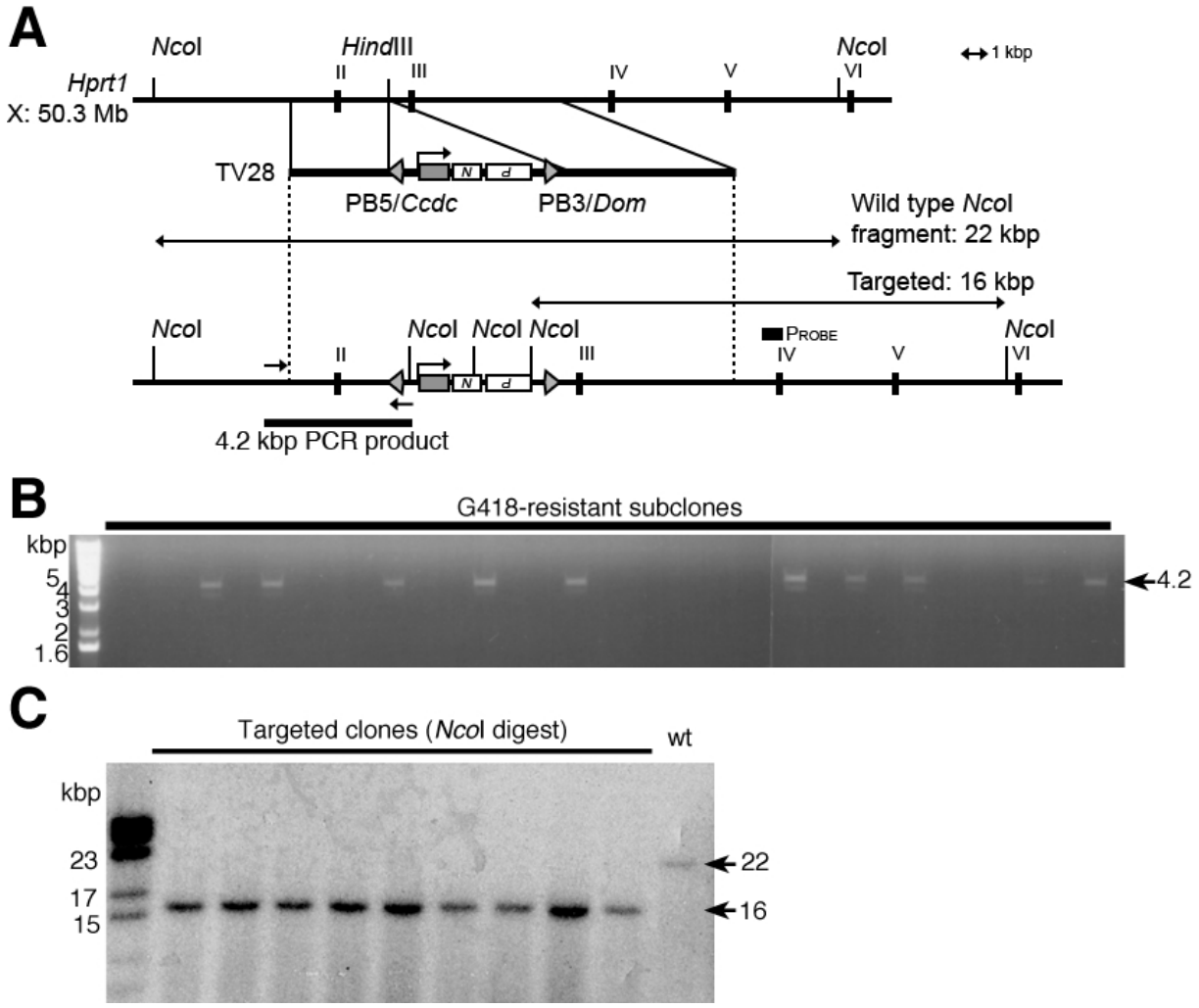


Figure 6.4: Targeting the transposon to *Hprt* in $Blm^{e/e}$ cells using the TV28 vector. A—Targeting scheme. B—PCR screen of lysates from G418-resistant colonies. A 4.2 kbp product is amplified from correctly targeted clones; another product of about 3.5 kbp, still specific to the targeted clones, is also amplified. This may arise from priming elsewhere in the PB repeats. C—Southern blot confirming correct targeting at the 3′ end in nine targeted clones. The probe used is shown in Figure 6.4.

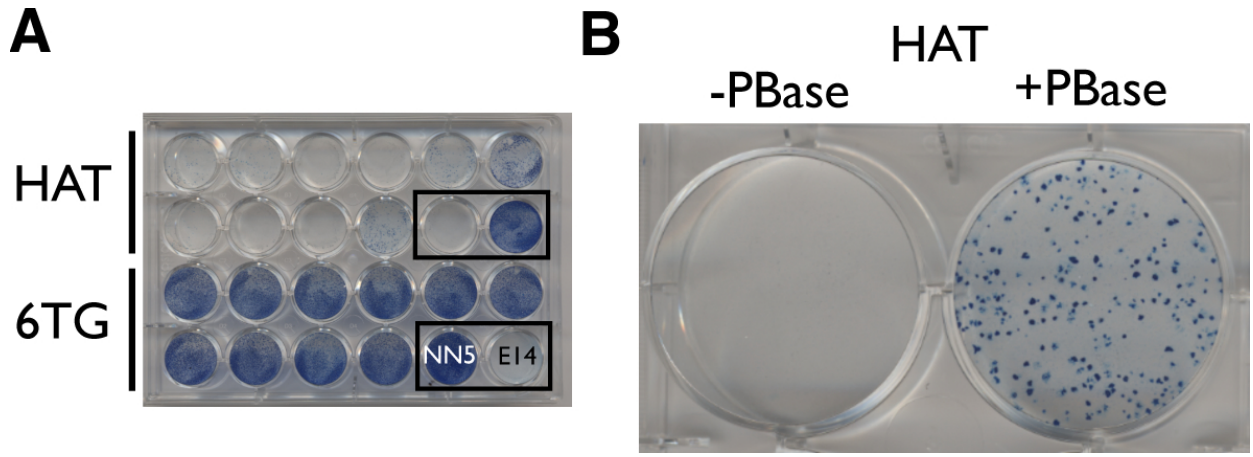


Figure 6.5: Confirmation of targeting and testing transposition from the *Hprt*^{PB} locus. A—Targeted clones are HAT sensitive and 6TG resistant. These have not been subcloned and therefore some HAT-resistant background can be seen in some wells. NN5 and E14 are *Hprt*-negative and positive controls respectively. B—Transfection with PBase results in HAT-resistance from clones that have excised the transposon. The cells used are a subclone from an originally identified targeted clone; no HAT-resistant background is observed without PBase transfection.

the cells 24 hours post transfection and selected in HAT+puromycin to isolate clones in which the transposon had excised and reintegrated elsewhere in the genome. These cells were expanded under HAT and puromycin selection for eight days and in M15 thereafter (with two days in HT medium to allow recovery from HAT selection). The cells were replated during this expansion/selection phase, at which point some were transferred to another plate at low density for counting and analysis of the mobilisation by Southern blot (Figure 6.10).

After expansion for 14 days in total, I changed the medium to 1 μM 4-OHT overnight. The next day I changed the medium to M15 and allowed one day for Cre activity to subside. At this stage I picked a few colonies from a low density plate to assess Cre induction by Southern blot. For the higher density plate I harvested the cells, counted and replated in DBL medium, again at both high and low densities (half and one tenth of the total respectively). I picked colonies from a lower density plating and analysed by Southern blot to determine transposon copy number, orientation and insertion site using the same digest and probe as in Chapter 5. I included clones from the initial mobilisation, background plates and unselected 4-OHT treated plates in the analysis to get a complete picture of each stage of the process.

6.2.3 Sources of background in double-resistant population

Copy number increase on transposition

From Southern blot analysis of double resistant clones generated in this experiment, problems were immediately apparent. First, several clones with two non-allelic insertion sites could be seen even in the HAT+puro resistant clones from immediately after mobilisation (Figure 6.11; lanes 1, 3, 6, 8). Therefore mobilisation from the X chromosome failed to effectively limit the transposon copy number to one. Analysis of genomic DNA from unmobilised LGP cells confirmed that there is only one copy of the transposon (at *Hprt*) in the starting cells (Figure 6.11, far right lanes).

Selection background

Predictably, having clones with two copies present at the start resulted in a background of clones with two non-allelic insertions in the double-resistant population. Moreover, there were some clones that survived double selection but did not have both *neo* and *puro* forms of the transposon as assessed by Southern blot (Figure 6.11, lanes with asterisk). This was surprising, as I had never observed this in the clone-by-clone experiments described in the previous chapter. Therefore it is likely that this background is specific to a comparatively rare set of loci that may have high transcription activity on

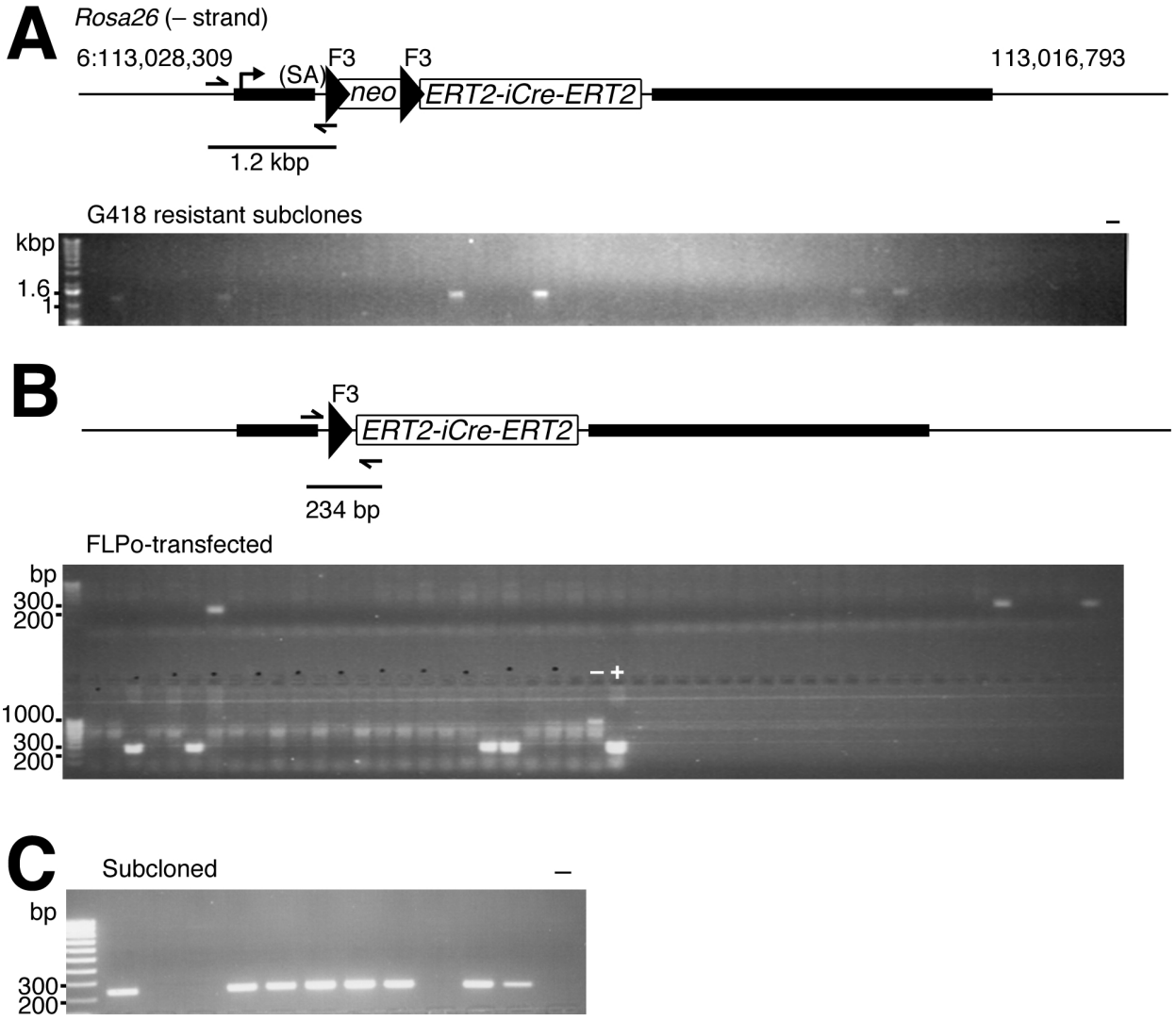


Figure 6.6: Targeting an inducible Cre gene in *Blm*$^{e/e}$ cells. A—Structure of targeted locus. *Rosa26* is shown 5′ to 3′, left to right, although it is on the reverse strand. Homology arms in the targeting vector are shown with a thick line. PCR screen for targeted clones is shown. B—Structure of targeted locus after Flp-mediated removal of *neo*. PCR screen shown. +: DNA from correctly 'popped out' cells (K. Yusa); −: DNA from untransfected cells. A PCR product of around 1 kbp is visible in the negative control and faintly in the clones that have not undergone Flp-mediated deletion. C—Repeated PCR screen on subclones of a correctly 'popped' clone from (B). The *neo* product was visible at long exposure in the negative subclones.

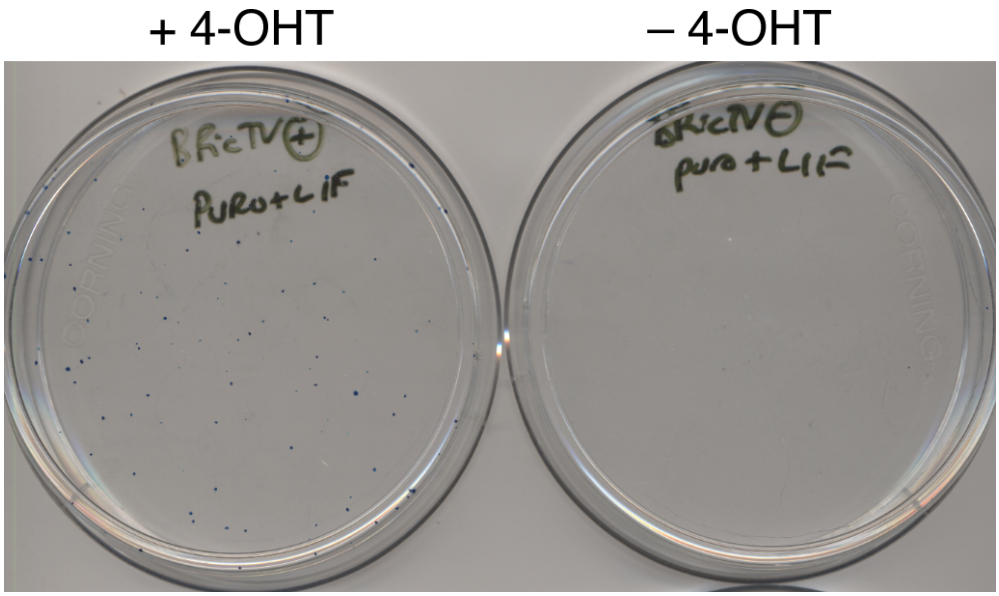


Figure 6.7: Gene targeting combined with 4-OHT treatment to insert the transposon in BRic cells. Puromycin resistant clones were only obtained when cells transfected with the TV28 targeting vector were also treated with 4-OHT.

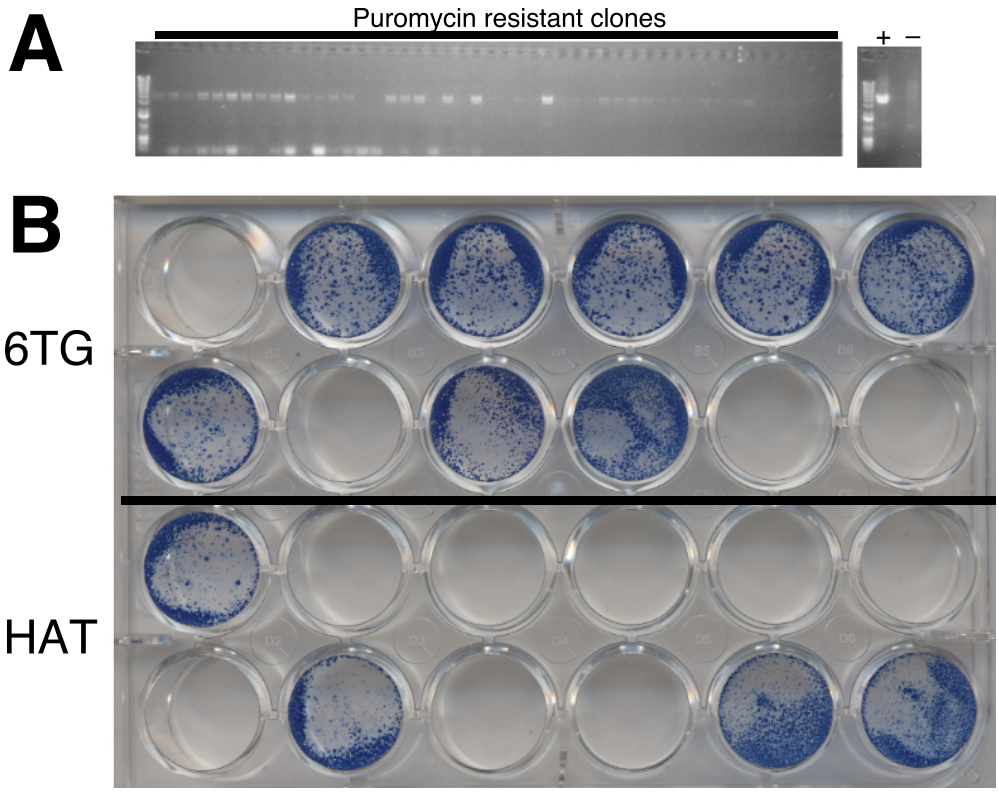


Figure 6.8: Confirmation of targeting in LGP cells picked from plates in Figure 6.7. A—PCR screen at 5′ end. This showed a large number of positives, therefore a functional test was also performed (B). Two HAT-sensitive/6-TG resistant clones were subcloned and frozen in small aliquots.

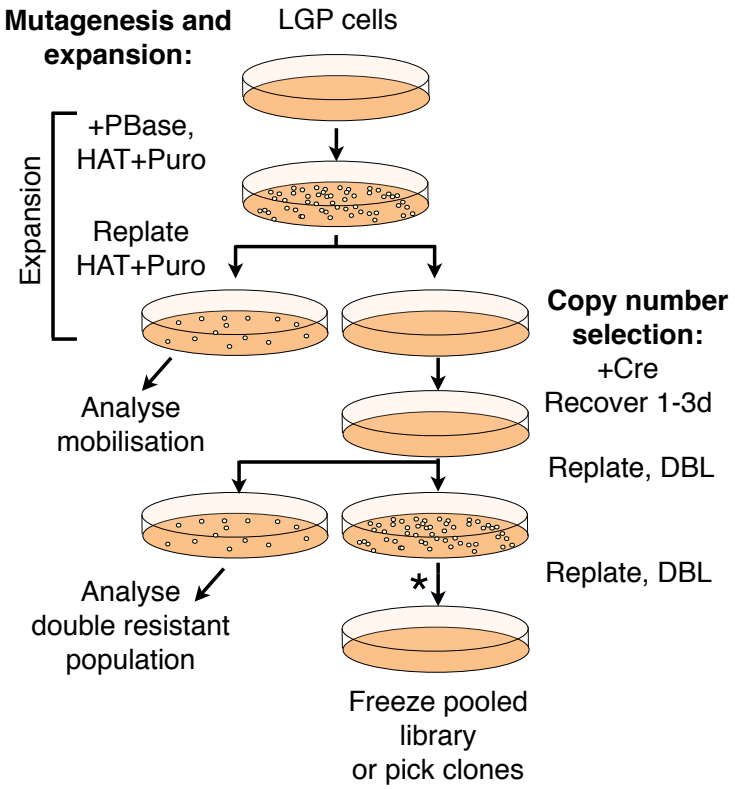


Figure 6.9: General scheme for library generation using the LGP cell line. For LGN cells, HAT+G418+FIAU selection would be used instead of HAT+Puro. The plating scheme shown is for the LGP libraries analysed (Figures 6.11 and 6.13). For the LGNL library analysed in Figure 6.19, clones were replated under DBL selection, at the stage indicated by an asterisk (*).

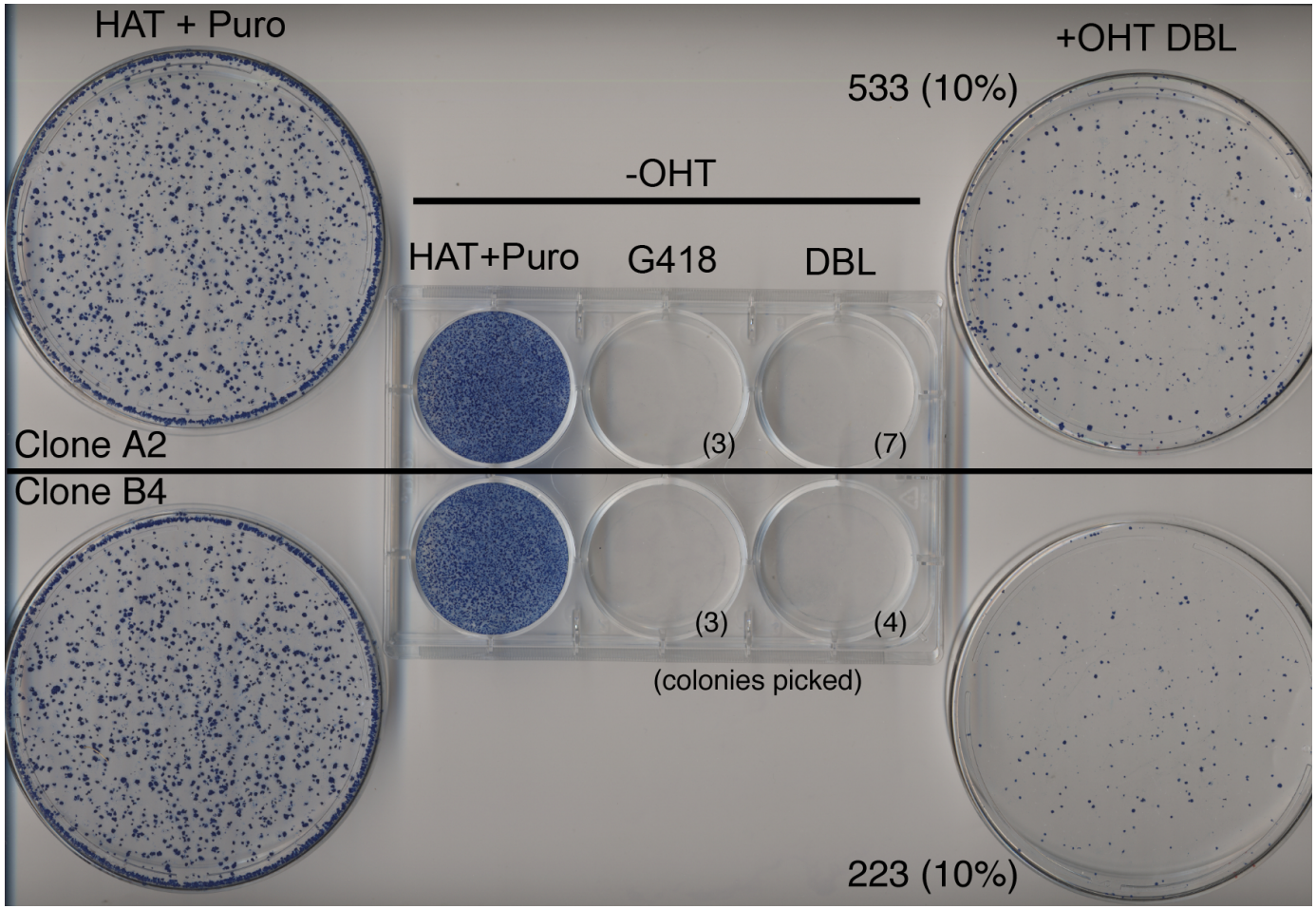


Figure 6.10: Colonies stained at various stages of the library generation process in LGP cells. Two parallel library generations from different starting subclones (A2 and B4) are shown. The plates shown are small proportions of the culture split to a separate plate at each passaging stage and stained 8–10 days later. From left, HAT+puro resistant cells that have mobilised the transposon. These cells are not generally G418- or DBL-resistant without 4-OHT treatment (six-well plate). A few colonies (numbers in brackets) did grow on these plates but were picked for genotyping (Figure 6.11). Finally, a plate containing 10% of the DBL-selected culture is shown with colony numbers.

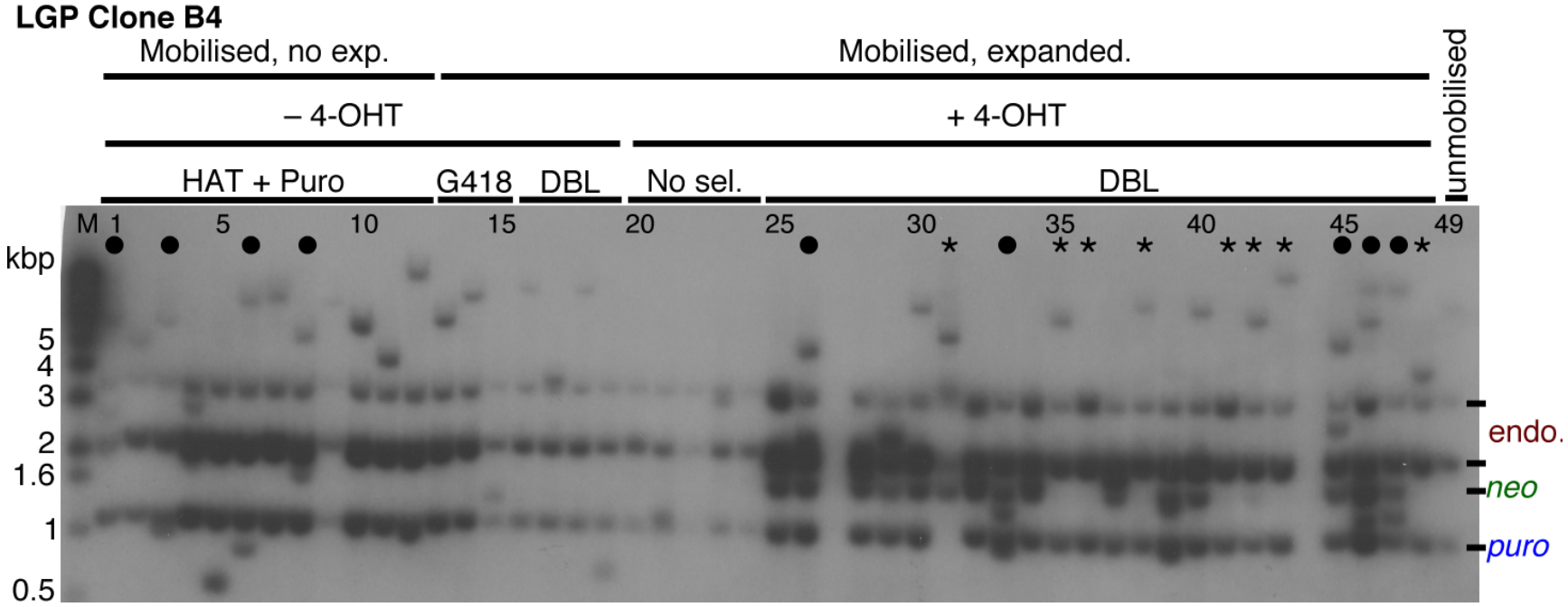


Figure 6.11: Background in double-resistant clones from pooled libraries. Clones from each stage in the generation process from LGP B4 cells were analysed by Southern blot using *Nco*I digested genomic DNA and probe as in Figure 5.6. Several mobilised clones (HAT+Puro) have more than one insertion site (lanes labelled with a dot). Background colonies that are G418- or DBL-resistant prior to 4-OHT treatment have not inverted the resistant construct, as shown by the lack of *puro* band in these clones. Unselected clones treated with 4-OHT show little or no recombination to the *puro* version, suggesting that this was not very efficient. DBL resistant clones that do not show both *neo* and *puro* bands are shown with an asterisk. The far right lane is LGP DNA (unmobilised).

the reverse strand sufficient to express the 'inactive' selectable marker (Figure 6.12A). In these experiments I am interrogating thousands of genomic loci, so it is possible that loci with these properties would be picked up in this experiment but missed in a clone-by-clone context.

If this explanation is correct, there is a simple work-around. If the initial transposition was carried out with a TNN transposon (i.e. in the *neo* expressing orientation rather than *puro*), FIAU could be used to select against loci that can express the $pu\Delta TK$ from a genomic promoter. To convert the TNP transposon in the LGP cell line to TNN, I simply treated cells with 4-OHT and selected in G418. I picked a number of resistant subclones, forming the LGN cell line with the transposon in the *neo* orientation to allow the proposed FIAU counterselection.

Comparison of libraries generated from different clones

I generated libraries starting from two distinct LGP subclones—A2 and B4. The A2 library yielded more double resistant cells, but also had much higher background. This could be due to a problem with the A2 line prior to mobilisation, e.g. aneuploidy, or a stochastic event leading to background that only occurred in the A2 clone. Particularly, 11/24 double resistant clones in this library appeared to have a single insertion site, and arose from selection background (Figure 6.13; lanes 21, 22, 23, 26 etc.).

I expected that clones picked from the 'background' selection plates (i.e. G418 or DBL selection without 4-OHT treatment) would dominate the double resistant population. However, this was not generally the case. In the A2 library, two such clones can be identified (Figure 6.13; lanes 14, 24, 39 and lanes 11 and 36). The B4 library appears relatively complex, despite the background. This indicates that genuine expansion-dependent copy number gain events occur frequently enough to dilute the effect of the background clones to some extent.

Treatment with 4-OHT appeared more effective in the A2 clone compared to the B4 clone (Compare Figure 6.13 lanes 19 and 20 with Figure 6.11 lanes 20–24). In the B4 clone, the intensities of the *neo* and *puro* bands were approximately equal, whereas many subclones of the A2 line displayed bands of unequal intensity. It is possible that some Cre activity remained after plating of the culture in DBL medium, and therefore that some cells continued to switch. An extra replating step or a longer recovery period from 4-OHT induction should be added to ensure that the clones analysed are pure.

6.2.4 A G1-specific transposase to conserve copy number during transposition

PB is known to transpose by a cut-and-paste mechanism, which should be non-replicative. Nonetheless, my data clearly indicate that copy number can increase upon transposition. Considered together with my data from sequencing of transposon excision sites (see Figure 7.9 and discussion in Chapter 7), a possible explanation is that transposition occurs after DNA synthesis, in S or G2 phase. The transposon does increase in copy number after DNA replication, as does every other locus in the genome. While the two copies would normally segregate to different daughter cells, transposition at this stage with reintegration on another chromosome (or the sister chromatid) could result in a daughter cell with two copies (Figure 6.12B). It is not known whether PB transposition is regulated based on cell cycle stage.

If this hypothesis is correct, the copy number increase could be avoided by limiting transposition to G1 phase of the cell cycle. An interesting study describing a fluorescent based cell cycle indicator suggested a possible way to achieve this. Sakaue-Sawano *et al.* made two complementary fluorescent protein fusions fused to degradation signal sequences from two cell cycle regulated proteins. Cells expressing the GFP derivative-Geminin fusion that they describe are green, except in G1 and early S phase, where Geminin is ubiquitylated and degraded. Similarly, cells that express a RFP derivative fused to a CDT1 fragment fluoresce red, but only in G1 and early S phase, after which CDT1, a replication origin licensing factor, is degraded. The fusion fragments were from human genes, but worked effectively in mice too. Therefore, fusing the PB transposase to the CDT1 fragment in a similar way might also limit its expression to G1 and early S.

I prepared cDNA from human iPS cell RNA and PCR amplified the human *CDT1* fragment encoding amino acids 30–120 as described in Sakaue-Sawano *et al.* (2008), using primers tailed with 50 bp arms with homology to the hyPBase expression plasmid. My strategy was to fuse the CDT1 to the PBase C-terminus, and introduce a linker of three amino acids between the two (Figure 6.14A). At the DNA level, this linker included an *Asc*I site to allow cloning of longer linkers if required, as previous attempts to make a PBase-ERT2 fusion had shown that PBase activity was affected by the fusion—in this case only a C terminal fusion with a positively-

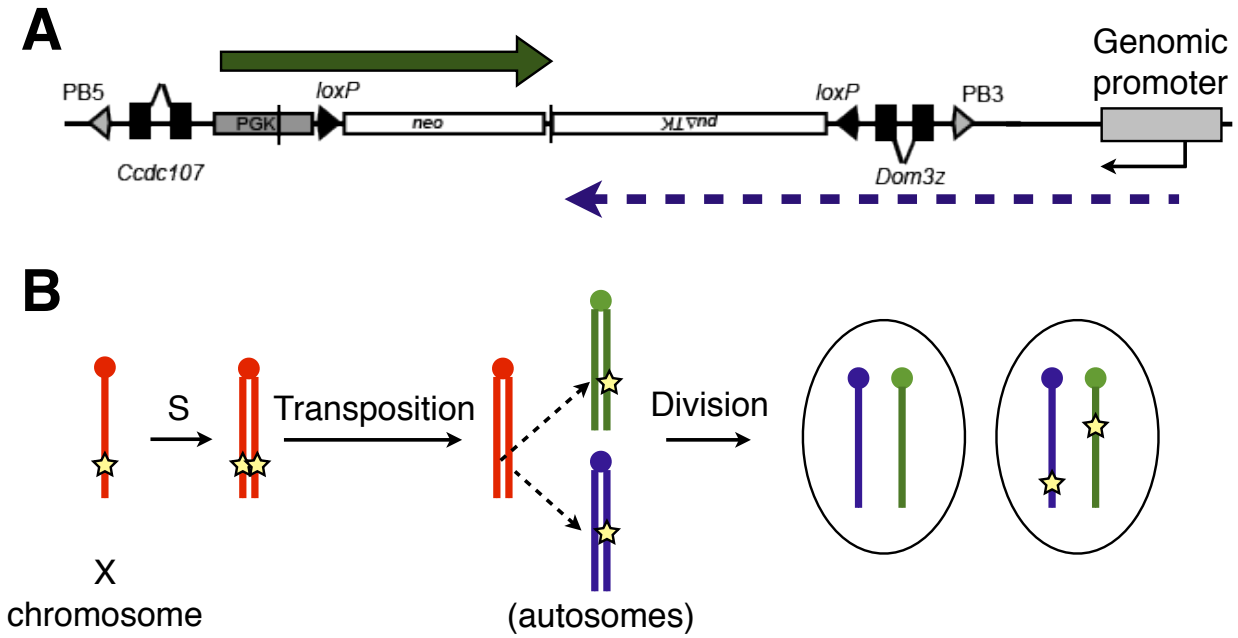


Figure 6.12: Possible sources of background in pooled libraries. A—Transcription of the reverse strand at certain genomic loci leading to expression of the 'off' resistance gene. B—Possible mechanism for the isolation of cells with two non-allelic copies, based on transposition occurring after DNA synthesis (S phase).

charged linker showed activity (Cadiñanos and Bradley, 2007). However, this could have been associated with the ERT2 domain rather than the PBase.

I linearised pCMV-hyPBase with *Not*I and co-electroporated heat-induced SW106 bacteria (see Methods) with the gel-purified fragment with the CDT1 PCR product with homology arms. Recombination in the bacteria reconstitutes a circular plasmid with the correct fusion gene, selected for by growth on ampicillin. Colonies were screened by PCR and checked by restriction digest after retransformation.

The fusion protein was active as a transposase in ES cells, displaying slightly lower activity than hyPBase without the fusion partner (Figure 6.14B). There are no commercial antibodies available that recognise the PB transposase. I tried using a commercially available antiserum raised against amino acids 7–106 of human CDT1 (Abcam ab52731). However this reacted with many different proteins on a Western blot of ES cell extracts, and did not detect any extra proteins in cells that had been transfected with a PB-CDT1 expression plasmid (Figure 6.15). Furthermore, I could not detect a band of the predicted size for endogenous CDT1 in protein extracts from a human cell line. Therefore I was unable to show cell cycle regulation of the fusion protein by Western blot. In an attempt to detect the fusion protein I made a plasmid designed to express a N-terminal FLAG epitope-tagged version of PB-CDT1. However, this transposase was not active in a transposition assay in ES cells (not shown). Further work needs to be done to verify the cell cycle regulation at the protein level.

6.2.5 Mobilisation using G1-specific transposase and FIAU counterselection

I went ahead with a mobilisation experiment using LGN cells and PB-CDT1 *in vitro* transcribed mRNA. This time I used medium containing HAT, G418 and FIAU (HGF) to select clones with excised and reintegrated transposons at loci that will not result in selection background. In this case, mobilisation efficiency was very low—only 142 clones were obtained. However, all clones had single copy integrations when analysed by Southern blot (Figure 6.16). Whereas seven of 20 clones analysed from mobilisations using hyPBase had two non-allelic insertions, all 18 analysed using PB-CDT1 had single copy integrations. This shows that the PB-CDT1 fusion protein does effectively limit the copy number of the transposon during transposition. I pooled

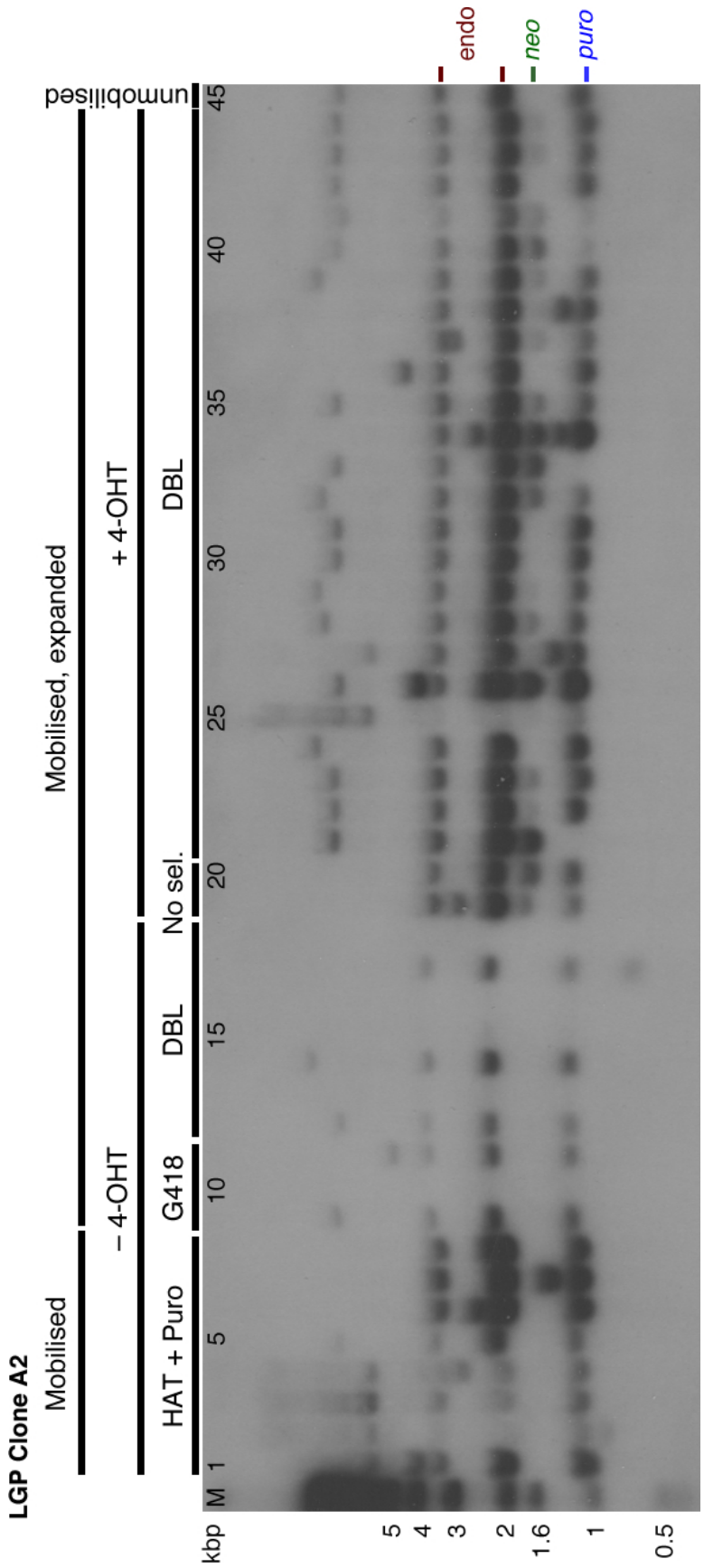


Figure 6.13: Analysis of clones from the LGP A2 library. Clones from various stages of library generation were analysed by Southern blot as in Figure 6.11. This library showed much higher background in the double resistant population compared to the LGP B4 library. Cre induction, however, appeared more efficient (lanes 19 and 20).

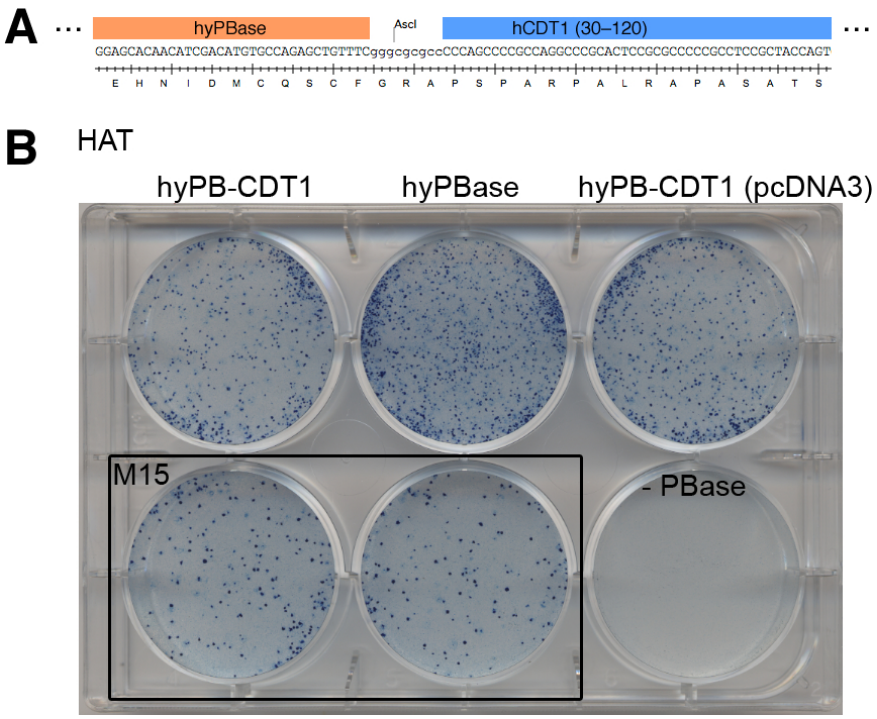


Figure 6.14: Structure and function of PBase-CDT1 fusion protein. A—DNA and protein sequence at the junction between the hyPBase coding sequence and the fused human CDT1 fragment. B—The fusion does not impair function of the transposase. Results of mobilisation in LGP cells are shown. ES cells were electroporated with 10 μg of the plasmid indicated. Expression of the transposase is driven by the CMV promoter in all cases. Top row—HAT selection, indicating transposition. Bottom row—cloning efficiency plates for hyPB-CDT1 and hyPBase transfections (unselected), and negative control (HAT selected). The top right well differs only in that PB-CDT1 is expressed from pcDNA3, which also uses the CMV promoter but contains an independently expressed *neo* gene.

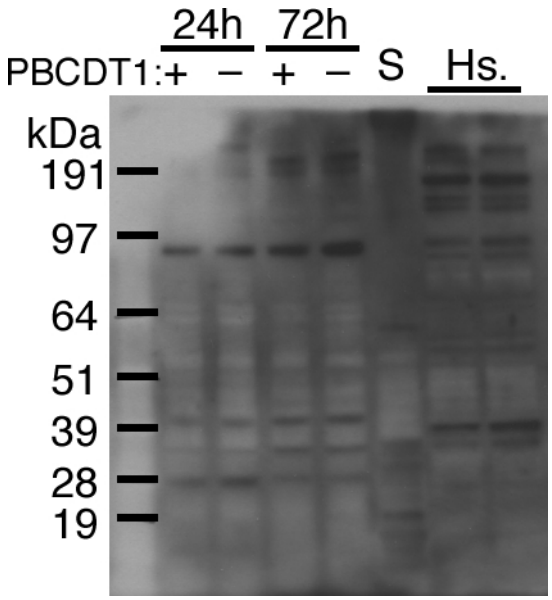


Figure 6.15: Western blot using anti-CDT1 antibody. Protein extracts tested (10 μg loading each) are JM8A3 cells at 24 or 72h post transfection with a PB-CDT1 (+) or GFP (–) expression plasmid. S—a stable G418-resistant cell line following pcDNA3-PB-CDT1 transfection [protein appears degraded]. Hs—control protein extracts from human cells. Predicted sizes—Human/mouse endogenous CDT1/Cdt1: 63 kDa; PB-CDT1 fusion: 87 kDa. The secondary antibody was HRP-conjugated rabbit anti-mouse IgG (Abcam ab6728).

all clones from this mobilisation experiment and expanded them for double selection.

The low efficiency of mobilisation in this case is likely to have been a technical problem with this particular experiment, as reintegration was readily observed in an experiment where LGP cells were transfected with PB-CDT1 (Figure 6.17A). Moreover, including FIAU counterselection did not adversely affect the number of recovered reintegration events in a pilot experiment conducted in B6BTV cells (Figure 6.17B).

6.2.6 Results of double selection with PB-CDT1 and FIAU counterselection

I expanded the cells for 20 days, as the initial cell number was low, then replated the cells in medium containing 1 μM 4-OHT. As the efficiency of Cre induction was low in the LGP B4 clone previously, where adherent cells were induced, I treated cells in suspension this time. After incubation overnight, during which cells attached to the plate, I replaced the medium with M15. After two further days I replated cells in DBL. Counting colonies from a replica plating at low density indicated that the culture contained 7,200 DBL-resistant colony-forming units (cfu) at this point. I checked the efficiency of Cre induction, and also verified that no double resistant colonies grew without 4-OHT treatment (Figure 6.18). In this case the Cre induction was very efficient, as opposed to the previous experiment with LGP cells where OHT treatment was carried out on adherent cells (see Figure 6.11, lanes 20–24).

To ensure that I picked pure clones, after four days I replated the cells under DBL selection. I picked double-resistant colonies and analysed them by Southern blot as before. This time, all clones analysed displayed fragments from both *neo* and *puro* forms of the transposon. This shows that FIAU counterselection effectively removed loci with expression of the resistance gene that is not oriented with the PGK promoter (Figure 6.19).

Despite starting from an apparently all single copy population, 19 of 45 double resistant subclones analysed from this library still have two non-allelic copies of the transposon (Figures 6.19 and 6.20). However, these are distinguished from the clones with two copies in the first library by the fact that all but one (lane 16) share one band, the size of which is consistent with the $Hprt^{PB}$ donor locus. Such clones were not detected in the initial mobilisation, suggesting they were only present at a low

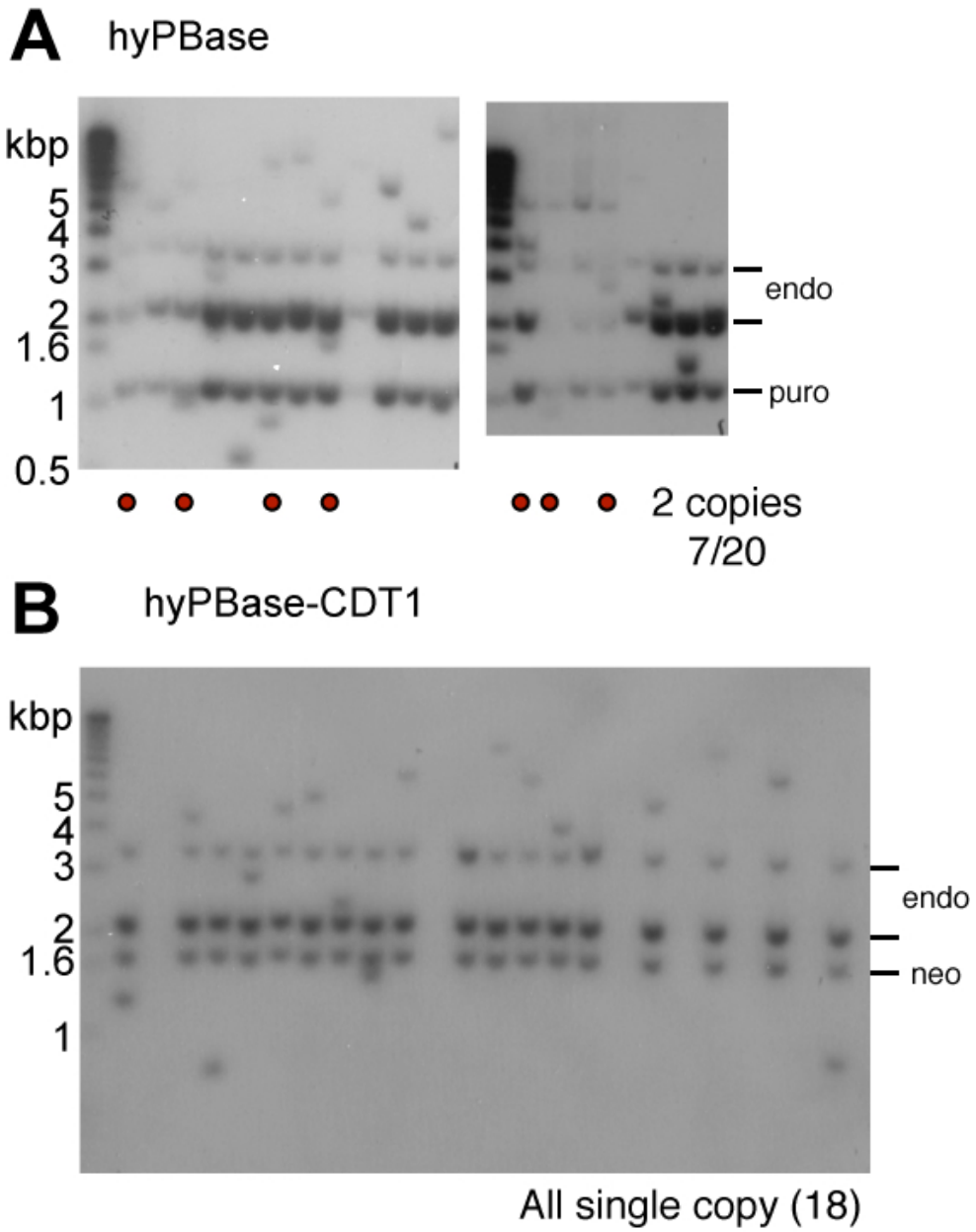


Figure 6.16: Mobilisation using the PB-CDT1 fusion protein preserves copy number. *Nco*I digests probed with the PB-CCdc probe as in Figure 6.11. A—HAT+Puro resistant clones from two mobilisation experiments in LGP cells using hyPBase mRNA. Several clones with two insertion sites are visible (lanes marked with a red dot). B—HAT+G418+FIAU resistant clones from mobilisation in LGN cells using PB-CDT1 mRNA. All clones have a single insertion site.

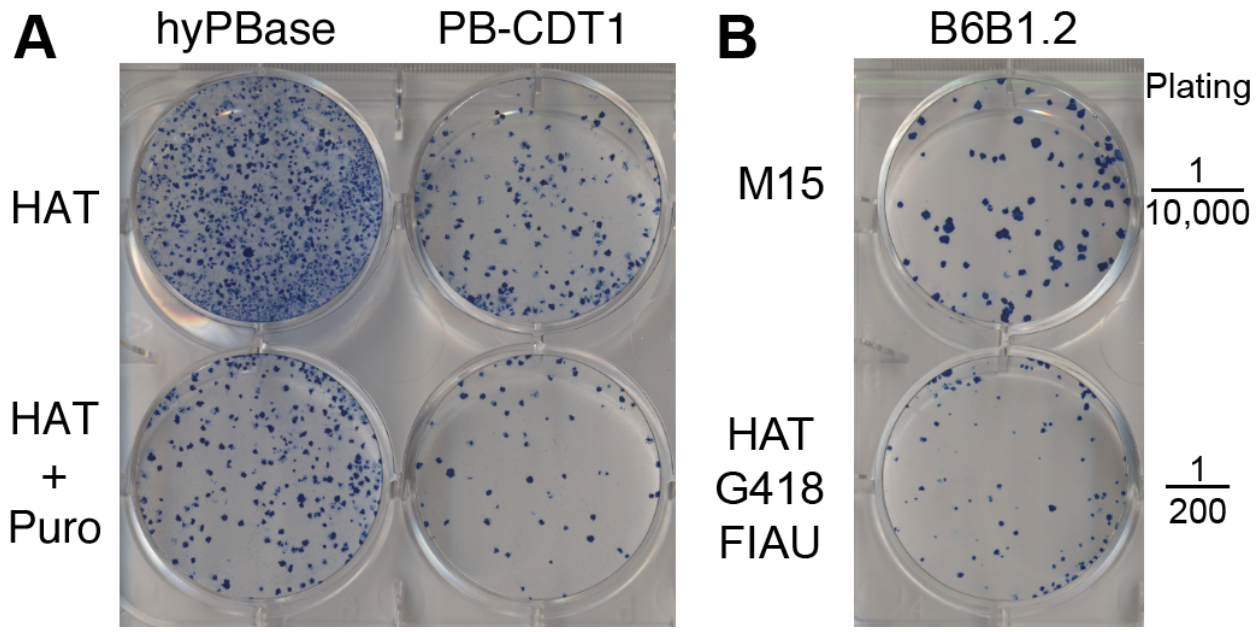


Figure 6.17: A—The fusion with the CDT fragment does not compromise transposon reintegration. LGP cells were transfected with hyPBase or PB-CDT1 expression plasmid as indicated and selected in HAT (excision) or HAT+Puro (excision and reintegration). B—An experiment using B6BTV cells, identical to LGN except for the lack of an inducible Cre gene, showing that reintegration events can be efficiently recovered using FIAU counterselection. The proportion of the mobilised culture plated is shown on the right

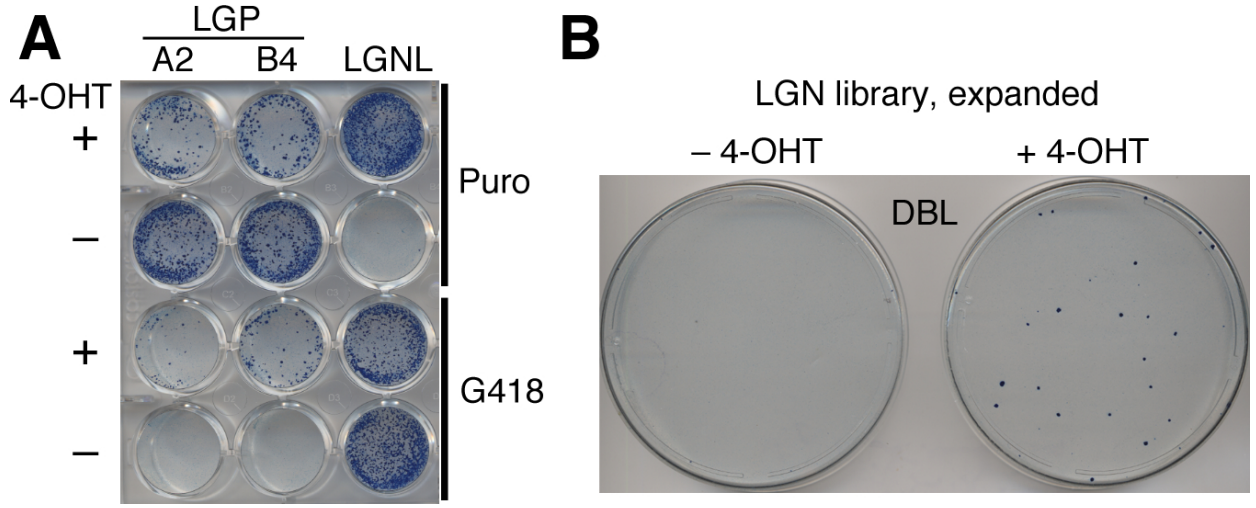


Figure 6.18: Verification of 4-OHT sensitivity and specificity in the LGNL1 library. A—4-OHT treatment efficiently induces switching to G418 resistance. Identically treated LGP cells are plated for comparison; in this case the switch is from puro to G418 resistance. B—4-OHT treatment is required for the isolation of double resistant cells.

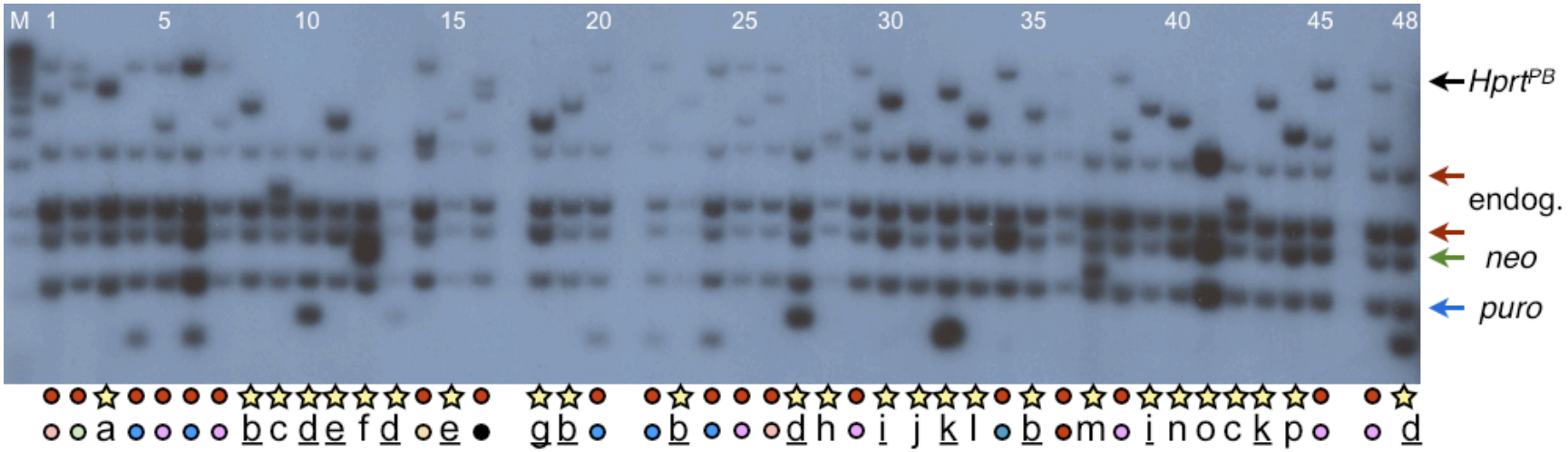


Figure 6.19: Analysis of double-resistant clones generated with G1-specific mobilisation and FIAU counterselection. Southern blot analysis with *Nco*I and PB-Ccdc probe as before. All clones show *neo* and *puro*-expressing forms of the transposon; therefore there is no selection background in this library. Most clones that show two non-allelic insertions (lanes marked below with dots) still have a band consistent with the $Hprt^{PB}$ donor locus (10 kbp). Starred lanes have allelic insertions and are potential homozygous mutants. Clonal relationships are indicated by letters (allelic insertions) or colours (non-allelic). Letters are underlined where the clonal relationship is supported by mapping data rather than band size. Further analysis in Figure 6.20.

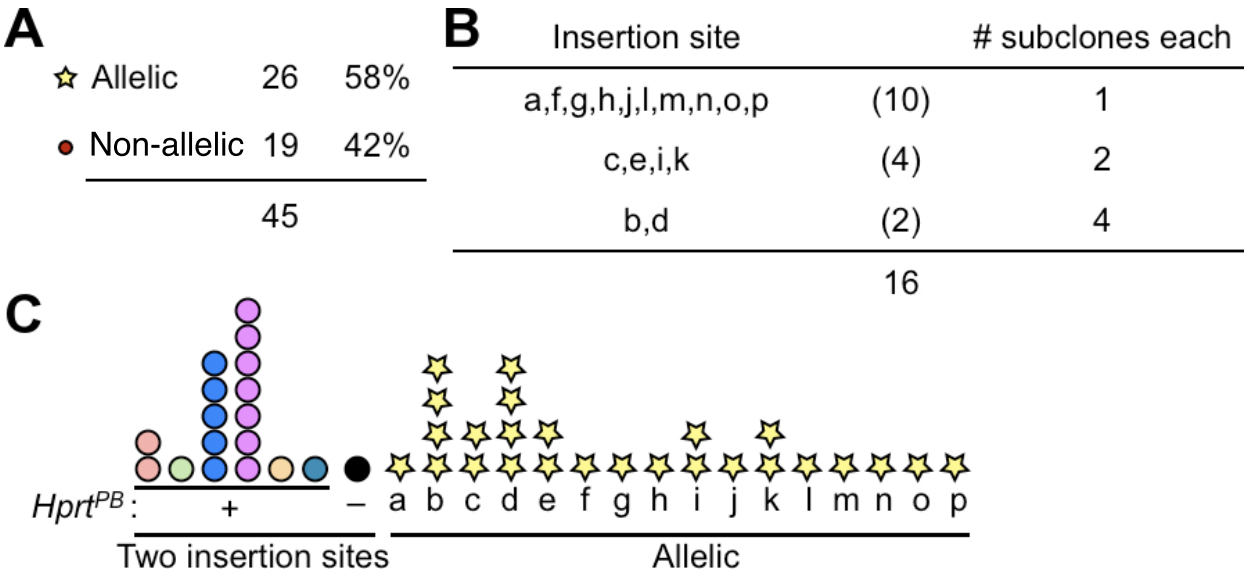


Figure 6.20: Analysis of clonal relationships between LGNL1 double-resistant clones. A—Table showing proportion of clones allelic vs. non-allelic insertions. B—Breakdown of allelic clones by insertion site. The number of subclones representing each insertion site is given in the right column. C—graphical representation of clonality data for both non-allelic (dots; colours correspond to Figure 6.19) and allelic insertions (stars). Each symbol represents a single subclone, each column a different insertion site (or site combination for non-allelic clones).

level prior to double selection.

The remaining 26 clones (58%) displayed the expected band pattern for cells with two allelic transposon insertions. There were at least 16 different insertion sites among the clones analysed on the blot, although this is a lower limit as there is some ambiguity for clones with similar size insertion-specific bands (Figure 6.20). Analysis of the clonal relationships was supported in some cases by mapping the insertion sites for clones with allelic insertions (Table 6.1). This shows that the method can generate complex libraries that are not dominated by clonal expansion from early LOH events. Some insertion sites mapped to the X chromosome. These clones presumably arise from aneuploidy in the culture. The double resistant allelic subclones need to be genotyped individually in order to determine how many are genuine homozygotes and how many arise from aneuploidy and retain the wild type locus.

6.2.7 Some allelic mutants retain the wild type locus

I genotyped the allelic mutants that I was able to map to see if they were genuine homozygous mutants. By PCR I was able to show that five of the double resistant subclones retained the wild type locus, and five were genuine homozygous mutants (Figure 6.21A). Of the five homozygous subclones identified, three were from one parental clone (clone b in Figure 6.19, Figure 6.21B).

6.3 Discussion

6.3.1 Sources of background in mutants isolated from complex pools

The experiments described in Chapter 5, in which mutants were expanded in isolation rather than as a pool, had not revealed any selection background. However, when I applied the method on a large scale, I isolated some clones that had not undergone Cre-mediated inversion of the selection construct. As the selection conditions are similar (with respect to cell density, drug concentration etc.), I interpret this as a consequence of sampling hundreds or thousands of loci simultaneously. The effect was not specific to either of the selectable markers or drugs, further suggesting a problem with the regulatory elements of the construct rather than the G418+puro selection itself.

A potential explanation is that there is a small fraction of potential transposon insertion sites that can express the resistance gene in the opposite di-

Lane	Gene	Chr	Position	Ori	Mapped ends	Clone
3	u/s of *Fubp3*	2	31,426,112	+	5+3	a
8	No gene	18	56,838,595	-	5+3	b
9	*Herc1*	9	66,347,379	+	3 only	c
10	No gene	17	34,010,713	+	3 only	d
11	*Trim47*	11	115,977,569	-	5 only	e
12	No gene	X	34,351,776	-	3 only	f
13	No gene	17	34,010,713	+	3 only	d
15	*Trim47*	11	115,977,572	-	3 only	e
18	No gene	X	12,924,110	-	5+3	g
19	No gene	18	56,838,595	-	5+3	b
23	No gene	18	56,838,595	-	5+3	b
27	No gene	17	34,010,713	+	3 only	d
28	*Gabbr1*	17	37,200,698	-	3 only	h
30	No gene	7	38,812,480	+	3 only	i
31	Refseq transcript	14	22,289,968	+	5+3	j
32	u/s of *Jarid2*	13	44,814,764	+	5+3	k
33	No gene	9	58,117,021	+	3 only	l
35	No gene	18	56,838,595	-	5+3	b
37	*Zmym3*	X	98,615,188	-	5+3	m
39	No gene	7	38,812,480	+	5+3	i
40	No gene	18	38,539,945	+	5 only	n
41	No gene	18	35,183,268	+	5+3	o
43	u/s of *Jarid2*	13	44,814,764	+	5+3	k
44	No gene	18	56,838,595	-	5+3	p
48	No gene	17	34,010,713	+	3 only	d

Table 6.1: Mapping data for LGNL1 clones with allelic insertions. Lane and clone columns refer to Figure 6.19. Ori, orientation: + indicates PB5 centromeric, PB3 telomeric.

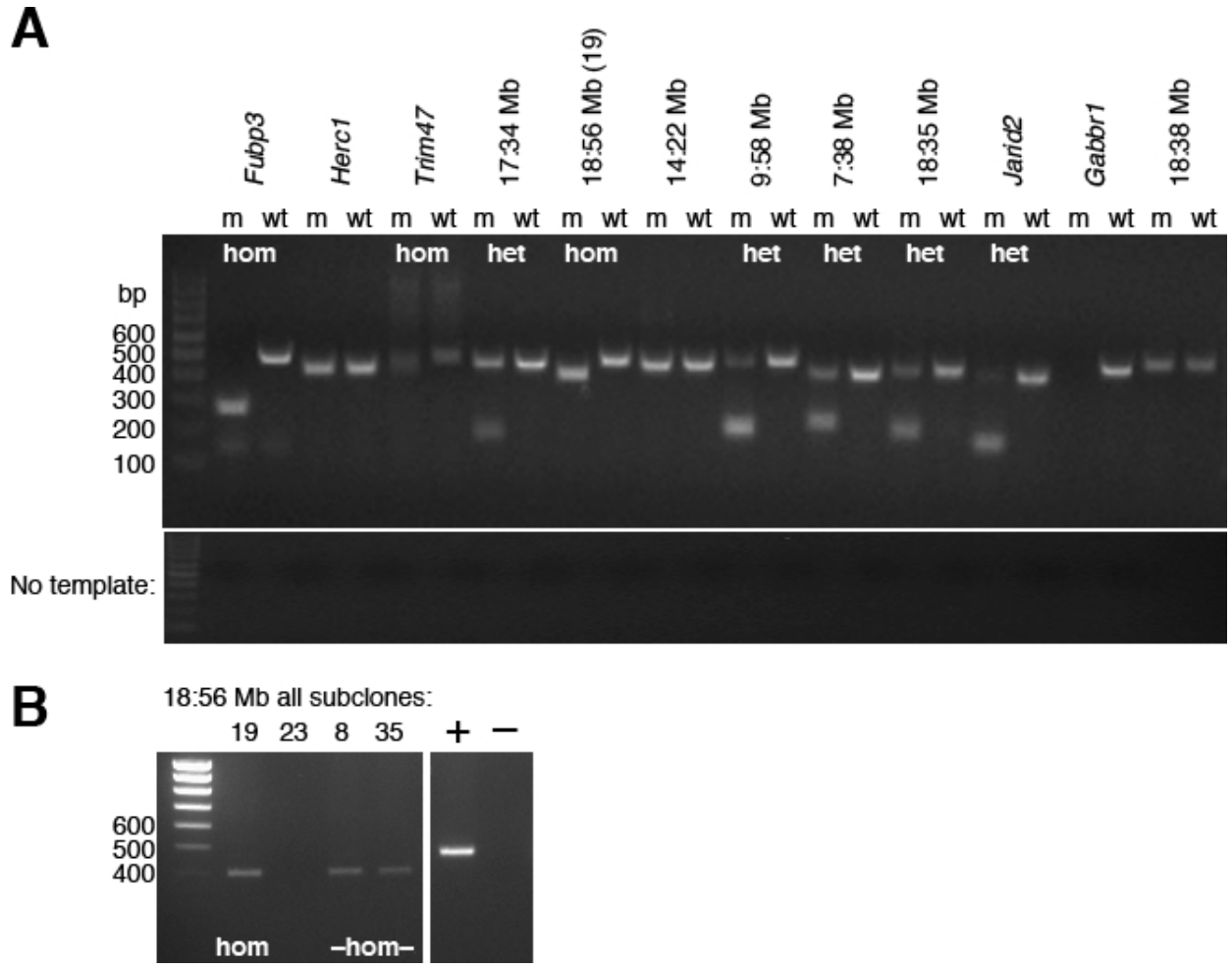


Figure 6.21: PCR genotyping of allelic LGNL1 subclones. A—PCR assays using three primers to detect mutant and wild type alleles for mapped allelic LGNL1 clones. m: DNA from isolated mutant. +/wt: wild type LGN DNA. The mutant product is smaller than the wild type in each case. Some PCRs do not amplify a product from the mutant, this could be due to incorrect mapping or primer incompatibility. Different primer sets and mutants are indicated by gene for insertions in or near a gene, and chromosome : position (Mb) for intergenic insertions. B—PCR results for additional subclones of clone b (18:56 from part A).

rection to the PGK promoter. This may be due to high levels of transcription or translation initiation on the appropriate strand, despite the presence of stop codons in the mutagen, which immediately precedes the 'unexpressed' resistance gene. There is accumulating evidence that a large proportion of the genome is transcribed to some degree (Cheng *et al.*, 2005; Gustincich *et al.*, 2006), and the presence of polyadenylation signals in the transcripts from the resistance genes would stabilise such transcripts. I added a counter-selection step with FIAU as an interim solution to this problem. However, in the long term the construct could be designed more intelligently—one such design would be to use a single polyadenylation site as well as a single promoter for both resistance genes, which would mean that any inappropriate transcripts of the resistance gene are unlikely to be stable. Such a construct has been developed, although not yet tested in a pooled format to my knowledge (K. Horie, J. Takeda *et al.*, unpublished).

These problems illustrate the difficulties arising from strong selection for a trait (double resistance) from a complex pool. In the clone-by-clone method, such 'rare' events as two-copy mutagenesis or locus-specific background are not serious, as even if they do occur, the other mutants are being cultured separately and are protected from contamination by the resulting mass of double resistant cells (see Figures 5.5 and 6.1). In the pooled format, these events adversely affect the complexity and usefulness of the double-selected libraries, as they result in potential double resistant cells present at the start. Therefore, the method needs to be refined to precisely target the event of interest (transposon copy number increase) and avoid the possibility of background completely—i.e. making these 'rare' events 'impossible'. My approach was to use FIAU counterselection mentioned above along with a novel cell-specific transposase to limit copy number in the mutagenesis step.

6.3.2 PB transposition and the cell cycle

Little is known about the cell cycle dependence of PB transposition, if any. My hypothesis for the transposon copy number increase post mobilisation is that transposition can occur after DNA is replicated. With this in mind, I modified the PB transposase by adding a degradation signal from the G1 and early S phase specific CDT1 protein. Although I was unable to characterise the expression pattern of the fusion protein directly using the reagents available, the fusion protein appeared to have the desired effect of limiting the transposon to one copy per cell after transposition from *Hprt* (Figure 6.16). This supports the idea that transposition to different loci after replication can result in cells with two non-allelic copies. Once again, this was not apparent in the clone-by-clone experiments, or rather clones with two non-allelic copies in this case were assumed to arise from repeated plasmid-to-genome transposition.

A side-by-side comparison of the PB-CDT1 fusion protein and the hyPBase protein that it is derived from indicates that the activity of the fusion, measured in terms of excision to give HAT resistance in LGP cells, is about half that of hyPBase (Figure 6.14B). A change of this magnitude is consistent with loss of expression in mid-S to G2, but is not formally separable from another effect of the CDT1 moiety (e.g. steric hindrance of the PBase).

6.3.3 Generation and uses of pooled mutant libraries

Library generation

The various library generation experiments that I have done have resulted in constant refinement of the protocol. A step-by-step experimental protocol, representing my current methods, is provided in the appendix to this thesis. This protocol incorporates the following elements that are required to obtain useful libraries:

- Use of *in vitro* transcribed mRNA to prevent integration of the transposase plasmid.
- Use of the PB-CDT1 fusion protein to limit initial copy number.
- FIAU counterselection to remove locus-specific selection background.
- Measurement of key parameters: starting clone number, average clonal expansion, Cre efficiency, double-resistant clone number.
- Characterisation of complexity and usefulness of libraries by Southern blot using the PB-Ccdc probe.

Some parameters still need to be fully optimised, particularly the expansion time to obtain sufficient complexity and a high proportion of homozygotes. This is currently based on theoretical consideration of the LOH rate, rather than empirical evidence.

The final experiments described in this chapter demonstrate a mutant library where 58% of double-resistant subclones are allelic mutants. These are

potential homozygotes, although some will still retain a wild type locus due to aneuploidy as described in Chapter 5. The results of PCR genotyping indicate that some are genuine homozygotes and some still retain the wild type allele, as for the clone-by-clone experiments. It is encouraging that all subclones were genuine mutants for the insertion site that was identified in several double resistant subclones (Figure 6.21). The library was of reasonable clonal complexity, with 16 different insertion sites identified from 26 double resistant subclones. This indicates that many clones are undergoing LOH. An expansion of 20 days was used in this experiment; a shorter expansion could reduce the redundancy still further, as discussed in Chapter 5. Of the 42% of the subclones that had two non-allelic insertions, almost all appeared to retain the $Hprt^{PB}$ donor locus. There may be further improvements to the protocol that could circumvent this. If these clones do indeed arise from aneuploidy *prior* to mobilisation, the cells could be sorted by DNA content prior to PBase transfection. Alternatively, if an arrayed library is to be made, a simple PCR screen that detects the unjumped locus could identify these clones. It may be possible to design a quantitative PCR (or Southern blot) assay to determine the relative amount of unjumped *Hprt* in a number of pooled libraries, and thus choose the best to use in screens.

Uses of pooled mutant libraries enriched for homozygotes

Libraries generated using this method are enriched for homozygous mutants by several orders of magnitude. In the expanded population prior to double selection, it is likely that there are of the order of 1,000 heterozygous cells for each homozygote. This is the population previously used for dominant screens in *Blm*-deficient cells. After my double selection procedure, clones with two allelic copies are readily visible in clones picked and analysed by Southern blot. Fifty-eight per cent were potential homozygotes. Therefore, depending on the actual percentage of real homozygotes compared to aneuploid cells, the proportion of useful cells is likely to be between half and one quarter of the library. This represents an enrichment of 250–500×.

I have documented a number of problems above, which mean that the enrichment is not complete. Is the level of enrichment obtained sufficient to use these libraries for genetic screens?

One obvious improvement is that the number of cells required for screening is vastly reduced. In an unenriched library, the order of 10^8 cells are typically screened (10,000 clones × 10,000 cells per clone; Guo (2004)). Using enriched libraries, screening the order of ten cells per clone should be sufficient (Figure 6.22). Thus, good genome coverage could be obtained by screening around 100,000 cells, which are easily accommodated in one well of a six-well plate, requiring only a few ml of medium. This may be important for some applications—for example, where resistance to dangerous or hard to obtain substances is being studied.

A second application of these libraries that I am interested in investigating is annotation of weak resistance or sensitivity phenotypes. If the library is grown for a period of time under stress or weak selection, mutants with a fitness advantage will increase their relative representation in the pool, and *vice versa*. Such screens have been successfully carried out in yeast and bacteria, and their extension to a mammalian system would greatly assist in functional annotation of the genome. As detailed in Chapter 3, Illumina sequencing can be used to determine the composition of mutants in a pool. Thus, sequencing pools of mutants expanded with or without stress or selection should allow the appropriate measurements to be made (Figure 6.22). This is only possible with enriched libraries, as the presence of a mutant is determined by insertion site sequencing, and therefore heterozygous cells will also contribute to the signal.

This method is likely to work best on libraries with short expansion times, as it depends on a high proportion of cells with each measured insertion site being homozygous mutants. If the library contains wild type retaining cells with the same insertion site as genuine homozygous mutants (see Chapter 5), these may mask an effect and lead to a loss in sensitivity. Although the mutant would not be recovered in this case, sequencing is very high throughput and should allow large libraries to be investigated. Multiple insertions per gene showing the same change in abundance would give confidence in the results. As many genes are now knocked out or have targeting vectors available (International Mouse Knockout Consortium *et al.*, 2007), methods that provide functional information are vital in prioritising further studies.

Finally, is this method suitable for the construction of an arrayed homozygous library, in a manner similar to the yeast deletion collection? Clearly in any such library there will be some 'junk', arising either from clones with non-allelic insertions, aneuploid clones that retain a wild type locus, or severe redundancy in the library (Figure 6.22). This needs to be taken into account when picking clones for an

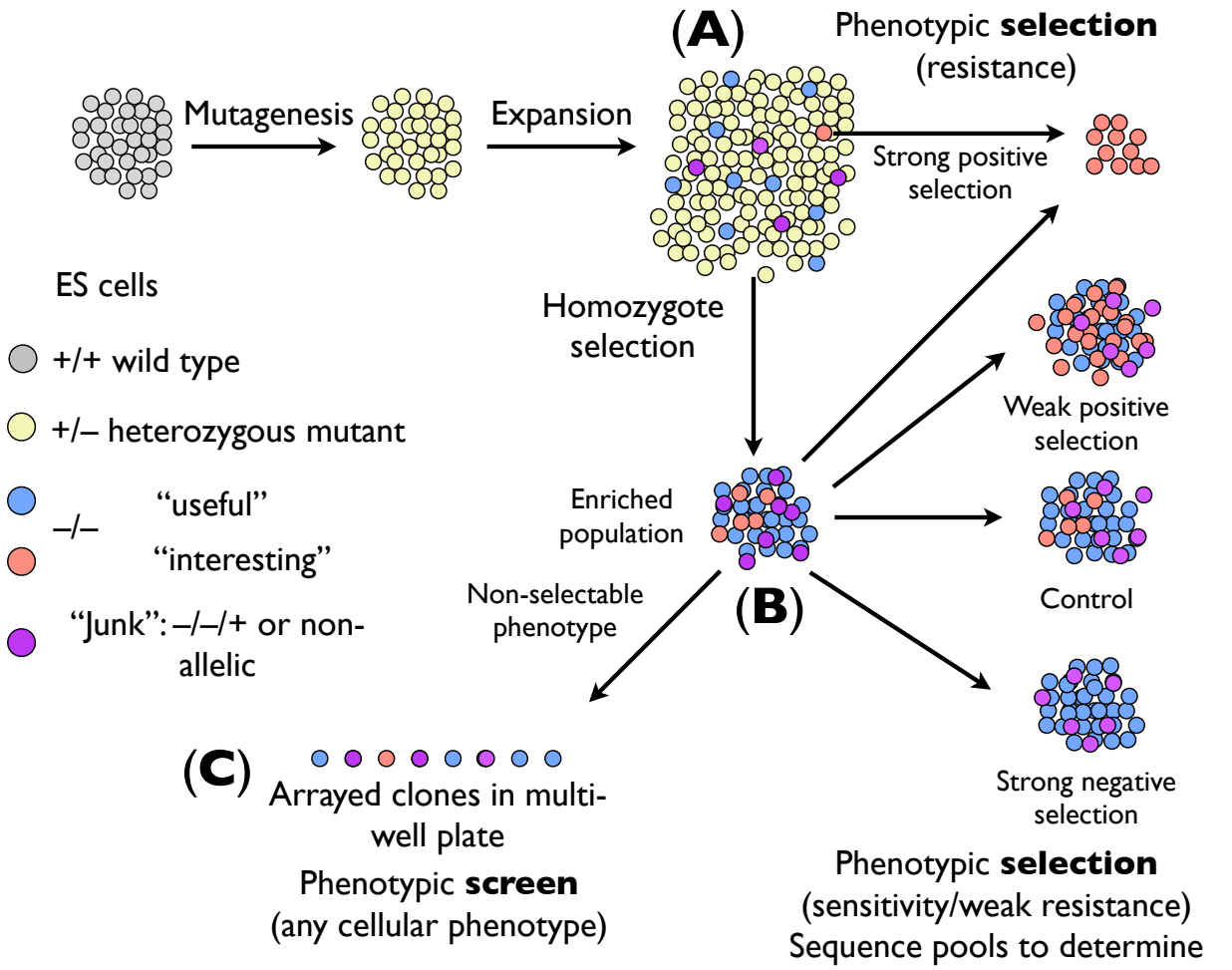


Figure 6.22: Uses of enriched libraries. A—Unenriched library, screening only by strong positive selection. B—Homozygote-enriched population isolated by copy number selection, screen by positive selection (fewer cells) or by sequencing more weakly or negatively selected pools (see text). C—Arrayed library made by subcloning homozygote-enriched population.

arrayed library, and the decision on how much junk to tolerate will be a logistical one. The method presented here to characterise the complexity and usefulness should be a guide in this regard. It appears that this may vary between experiments due to the stochastic nature of events that give rise to double resistant cells. Therefore I have taken care to isolate multiple subclones of the library generation cell lines, which can be used to generate many different libraries in parallel. These can then be characterised, and the best ones chosen to take further for screens.

6.3.4 Conclusions

There are several sources of background encountered when selecting for copy number increase on a genome-wide scale. These include copy number gain during transposition, and inappropriate expression of the second resistance gene in the construct. Copy number during transposition appears to be conserved when transposition is limited to the expression period defined by a CDT1 fragment, i.e. G1 and early S phase. This represents evidence that PB transposition can occur throughout the cell cycle. Using the $puro\Delta TK$ gene to select against inappropriate expression of the second gene eliminates this source of background. Several other adjustments to the library generation method can be used to reduce the background. The method described results in incomplete, but significant, enrichment of allelic two-copy mutants in the culture.

Chapter 7

Repair of DNA double strand breaks caused by piggyBac transposition

7.1 Introduction

One of the most striking properties of the piggyBac transposon (PB) is its precise excision from the genome (Ding *et al.*, 2005). In mouse ES cells, 95% of excision events were found to be precise (Wang *et al.*, 2008). This is unusual for a transposable element and has led to investigation of a novel use for PB—removal of transgenes from a genome. Two studies have used PB to introduce reprogramming transgenes for the creation of induced pluripotent stem cells (iPS cells, Yusa *et al.* (2009); Woltjen *et al.* (2009)). In these studies, the transposon is remobilised in the resulting iPS cells and subclones isolated where the transposon has not reintegrated. Due to the precise repair of the donor site, these cells are proposed to have a 'clean' genome, and thus are potentially suitable for therapeutic use, e.g. transplantation. In a related application, currently being pursued in our laboratory, PB can be used as an alternative to Cre-loxP or Flp-FRT recombinase systems for removal of selectable markers after gene targeting. Using site-specific recombinases for this purpose always leaves a single copy of the target site after removal. This is not optimal, as it is difficult to be sure that the remaining target site does not disrupt a functional element in some way. Furthermore, in extensively engineered cells or mice there may be many copies of the site in the genome, which could potentially recombine to cause inversions, translocations or deletions. In contrast, using PB to remove a selectable marker after targeting will leave no other mutation at the locus, provided that the PB is engineered into an endogenous TTAA site.

If such methods are to be used clinically, it is important to understand them thoroughly. Very little is known about the biochemistry of PB excision, essentially all coming from one published study (Mitra *et al.*, 2008). Repair of the donor site was not addressed in this study. In this chapter I describe the use of the *Hprt-PB* reporter locus developed for library generation to study repair of the break produced by PB excision. I found a genetic requirement for the nonhomologous end joining (NHEJ) factors *Xrcc4* and *Xlf* in accurate repair. The tools that I have developed constitute a new method to program and study the repair of double strand breaks in mammalian cells.

7.1.1 Excision of transposons

There are several known families of transposons with different mechanisms of excision. Transposases such as SB cleave both strands of DNA, usually at staggered positions, and thus cause a double strand break. The structure of the end produced depends on the exact position of cleavage and the ends are not necessarily compatible—SBase produces a three nucleotide non-complementary 3′ overhang (Luo *et al.*, 1998).

Other transposases make a single stranded nick, exposing a 3′ hydroxyl group, which is then used to break the second strand by nucleophilic attack. This produces terminal hairpins at the site of attack, which must be processed before the site is repaired. PB is an example of this category, which also includes the RAG1/RAG2 recombinase—a domesticated transposase used in V(D)J recombination in lymphocyte development.

An *in vitro* study using purified recombinant PBase and a minimal PB element has characterised the mechanism of PB excision and integration (Mitra *et al.*, 2008). PB leaves four nucleotide 5′ overhangs, and these are compatible as the PB insertion site (TTAA) is four bp in length and duplicated upon insertion. The two ends should, therefore be directly ligatable. Several host double strand break repair pathways could potentially handle this type of break, discussed below.

7.1.2 Cellular double strand break repair pathways

A single unrepaired double strand break (DSB) is a lethal lesion (Bennett *et al.*, 1993) because of signalling events that stall cell cycle progression and eventually cause apoptosis in response to DNA damage. This DNA damage response (DDR) is tailored

to the cell cycle phase and type of damage occurring (Jackson, 2002). In mammalian cells, the two major pathways of double strand break repair are nonhomologous end joining (NHEJ) and homologous recombination (HR; Figure 7.1).

Nonhomologous end joining

NHEJ, as the name implies, joins free DNA ends together without the use of sequence homology to guide pairing. In this sense, it can be considered an error-prone pathway, as two ends that do not belong together could be joined. NHEJ can also introduce mutations at the break point if the ends are processed before joining. This processing may involve removal or addition of nucleotides.

Nonhomologous end joining is used much more widely in mammalian cells compared to yeast, and many of the essential proteins were identified in mammalian systems, by a combination of biochemistry and complementation analysis of X-ray sensitive Chinese hamster ovary (CHO) cells (Jeggo and Kemp, 1983). T and B lymphocyte development relies on NHEJ for repair of developmentally programmed breaks in V(D)J recombination and class switch recombination, two processes that generate diversity at the immunoglobulin and T-cell receptor (TCR) loci (Dudley *et al.*, 2005). Therefore, lymphocyte developmental defects have also been useful for the study of NHEJ. The first factor involved is the heterodimeric Ku protein complex, which binds tightly to free DNA ends (Mimori *et al.*, 1986). These proteins form part of the DNA-dependent protein kinase (DNA-PK), which is completed by binding of the third component, the catalytic subunit DNA-PKcs (Gottlieb and Jackson, 1993).

Mutations in the *Prkdc* gene, which encodes DNA-PKcs, were found to be responsible for the phenotype of severe combined immunodeficiency (SCID) mice (Blunt *et al.*, 1995, 1996). SCID mice, as well as mice with targeted mutations in *Prkdc* are defective in T and B lymphocyte development (Gao *et al.*, 1998; Taccioli *et al.*, 1998). More specifically, such mice are deficient in processing and joining coding ends in V(D)J recombination—the process by which different segments of the immunoglobulin genes are juxtaposed to generate TCR and antibody diversity. This involves resolution of a hairpin intermediate, much like that generated at the ends of the excised PB transposon. In contrast, the blunt signal ends of the excised sequence in the V(D)J recombination process are repaired normally in SCID and $Prkdc^{-/-}$ mice. Targeted knockouts of either of the two Ku subunits also lead to immunodeficiency; however in this case both coding and signal joins are affected (Nussenzweig *et al.*, 1996; Gu *et al.*, 1997a,b; Zhu *et al.*, 1996).

DNA-PK regulates the ongoing DNA damage response by activating itself by autophosphorylation in *trans*, and also by phosphorylating other proteins required for end processing. Many physiological DNA breaks will contain complex structures that can not be ligated, so require processing by nucleases and polymerases. Several enzymes that process ends are known, including the Artemis nuclease (which forms a complex with DNA-PKcs), terminal deoxynucleotidyl transferase (TdT) and the polymerases μ and λ. *Artemis* knockout mice have a similar phenotype to *Prkdc* knockouts, supporting an essential role in repair of ends that require processing prior to ligation (Rooney *et al.*, 2003).

Another important kinase in the signalling response to DNA damage is ATM (Ataxia Telangiectasia Mutated). ATM is activated in response to very low levels of DNA damage, corresponding to just a few breaks per cell (Bakkenist and Kastan, 2003) and phosphorylates a number of cell cycle regulators and DNA repair proteins (Shiloh, 2003). Although ATM is not directly involved in repair of DSBs, cells from ataxia telangiectasia (AT) patients are radiosensitive, as are *Atm*-deficient ES cells (Xu and Baltimore, 1996). A subset of DSBs caused by IR persists in ATM-deficient cells and cells treated with an ATM inhibitor. This set of ATM-dependent DSBs may represent breaks occurring in heterochromatin or breaks with complex structures (Goodarzi *et al.*, 2008).

The ligation itself in NHEJ is carried out by DNA Ligase IV. Although DNA Ligase IV is sufficient for the ligation of certain substrates *in vitro*, in cells it forms a complex with the XRCC4 protein. In the absence of XRCC4, DNA Ligase IV protein is destabilised and its ligation activity reduced (Grawunder *et al.*, 1997; Bryans *et al.*, 1999). The two proteins are thus functionally linked. More recently, a new component of the ligation complex was identified: XRCC4-like factor (XLF, also known as Cernunnos. Ahnesorg *et al.* (2006); Buck *et al.* (2006)). Purified XLF stimulates the activity of XRCC4-DNA Ligase IV in *in vitro* assays. XLF can be considered a core NHEJ component, as XLF-deficient cells display increased radiosensitivity (Ahnesorg *et al.*, 2006).

In contrast to the early-acting NHEJ factors, knocking out *Xrcc4* or *Lig4* in mice results in embryonic lethality (Frank *et al.*, 1998; Gao *et al.*, 1998; Zha *et al.*, 2007). There appears to be a particular requirement for these factors in the de-

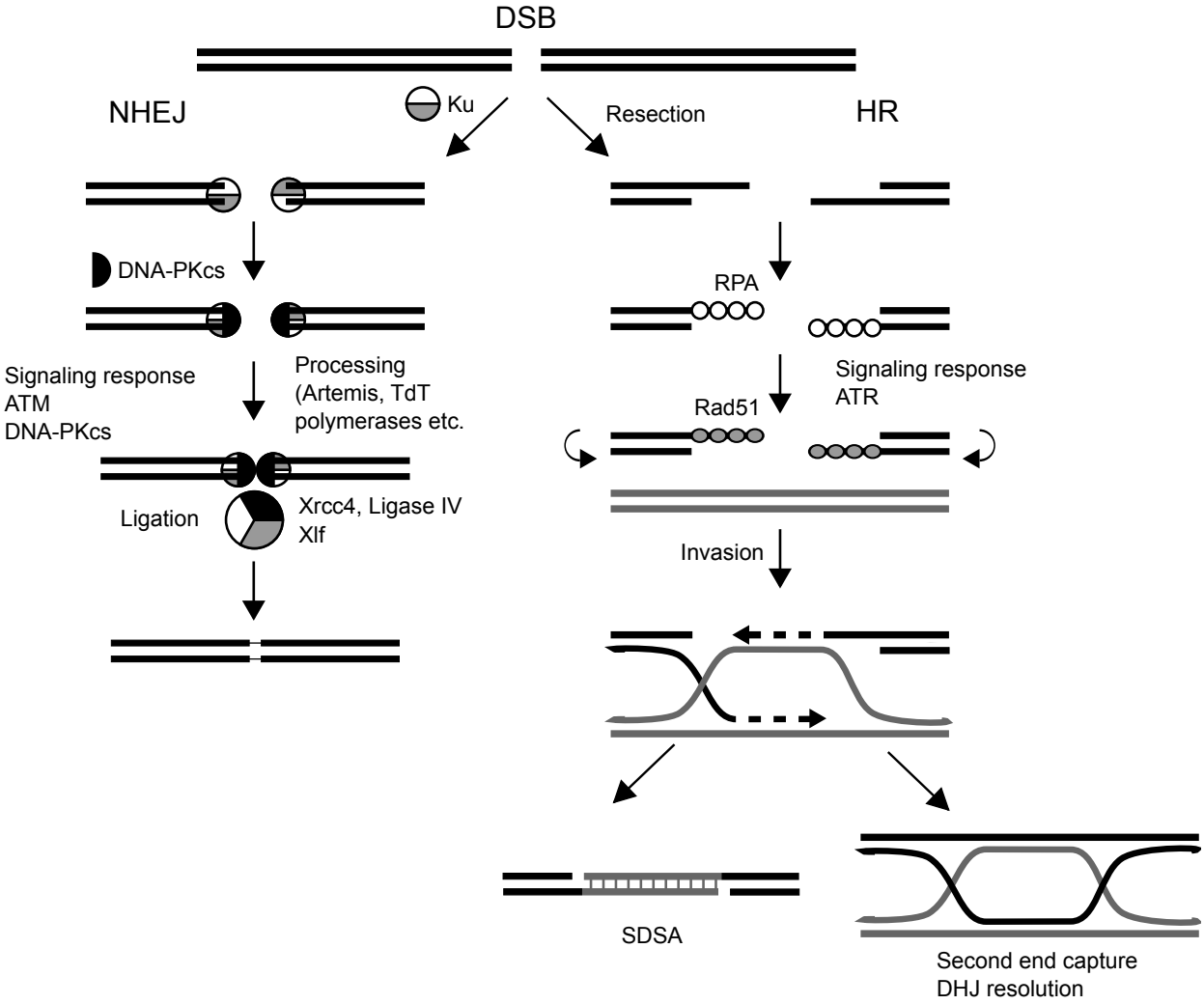


Figure 7.1: Double strand break repair pathways in mammalian cells. See text for details. Protein complexes are simplified and do not represent the exact stoichiometry or contacts.

veloping nervous system, as homozygous embryos display massive apoptosis in the developing nervous system. On a p53-deficient background, this apoptosis and the embryonic lethality is rescued and animals develop medulloblastomas and pro-B cell lymphomas (Frank *et al.*, 2000; Gao *et al.*, 2000). In the lymphoid lineage (in conditional knockouts or on a p53-deficient background), knockouts display the expected V(D)J recombination and class switch recombination defects.

Despite a strong radiosensitive phenotype in *Xlf*-deficient cells, the corresponding knockout mice do not show severe defects in V(D)J recombination, although they show some defects in class switch recombination later in development (Zha *et al.*, 2007; Li *et al.*, 2008). This distinguishes *Xlf* from the other 'core' NHEJ factors. What determines the requirement for *Xlf* in repair of DSBs, or rescues repair in its absence, remains unclear.

Some joining can still occur in the absence of NHEJ components. Cells lacking DNA-PKcs, Xrcc4 or Ligase IV can still repair a large fraction of IR-induced DSBs, albeit with slower kinetics compared to wild type cells (DiBiase *et al.*, 2000; Wang *et al.*, 2006). A low level of joining activity also occurs in V(D)J recombination on extrachromosomal substrates, and at double strand breaks induced by the I-*Sce*I nuclease (see below). Using these assays, the structure of the products can be examined by sequencing. In these mutant backgrounds, larger deletions are observed at the site of the break, often accompanied by apparently untemplated insertions of a few base pairs (Weinstock and Jasin, 2006; Guirouilh-Barbat *et al.*, 2007; Yan *et al.*, 2007). The deletions are often flanked by 'microhomology' of a few (2–6) base pairs, and are proposed to arise by annealing of these homologous sequences either side of the break. This has been proposed to provide synapsis, which may be lacking in cells deficient in core NHEJ components, and thus hold the ends together long enough for joining by another ligase, with deletion of the intervening sequence. The deletions and untemplated insertions may reflect multiple cycles of nucleolytic degradation and addition in the absence of repair. It has been suggested that this process may generate novel microhomology.

A likely candidate for the ligase in this so-called backup NHEJ (B-NHEJ) pathway is DNA Ligase III. Depletion of Ligase III from extracts from cells defective in the core NHEJ components further reduces joining activity (Wang *et al.*, 2005). Another factor implicated in the B-NHEJ pathway is poly(ADP-ribose) polymerase (PARP). There are several genes encoding PARPs in humans and mice, with PARP-1 and PARP-2 likely to represent the main activity in DNA repair (Amé *et al.*, 2004). Inhibiting PARP activity with small molecule inhibitors reduces end joining in Ku-deficient cells, but not in cells lacking Ligase IV (Wang *et al.*, 2006). This raises the possibility that PARP may act early in the pathway choice, at the same stage as Ku, and therefore inhibition has no effect in Ku-proficient cells.

NHEJ is active in all phases of the cell cycle. However, in late S phase and G2 phase, where a homologous template (the newly synthesised chromatid) is available, double strand breaks are more likely to be repaired by the process of homologous recombination (Rothkamm *et al.*, 2003).

Homologous recombination

Homologous recombination is the process of repairing DNA damage using sequence information from a homologue elsewhere in the DNA. Usually this is the allelic position on the sister chromatid; thus homologous recombination is only a major pathway of DNA repair in mammalian cells after replication has occurred, i.e. in S and G2 phases (Figure 7.1, Johnson and Jasin (2000)).

The process begins with 5′ to 3′ nucleolytic resection of the DNA flanking the double strand break. This resection produces 3′ single stranded DNA (ssDNA) overhangs. These are bound by a series of RPA protein monomers to form a protein-DNA filament. The presence of single stranded DNA activates the ATR (Ataxia telangiectasia related) kinase, which promotes downstream HR events and cell cycle arrest (Zou and Elledge, 2003). RPA is replaced by RAD51, a process dependent on BRCA2, which interacts functionally and physically with RAD51 (Scully *et al.*, 1997; Sharan *et al.*, 1997). This RAD51-ssDNA filament promotes homology searching and strand invasion on the homologous DNA. Synthesis to extend the invading strand allows use of sequence information from the homologue to fill any gaps and effect error-free repair. Both ends can be extended and ligated while still invading the homologue, producing a double Holliday junction (DHJ) which needs to be resolved, usually via BLM-TOP3α-RMI1/2. Alternatively, repair can be accomplished by extension of both ends templated by the homologue to create a compatible overlap, which can then anneal and be filled in and ligated. This is known as synthesis dependent strand annealing (SDSA). Both these pathways usually yield non-crossover products, although the DHJ pathway has the potential to produce crossovers if resolved by other enzymes. For further details, see Filippo *et al.* (2008).

7.1.3 Experimental induction of DNA double strand breaks

Random breaks

The most common technique to directly induce double strand breaks is to use ionising radiation (IR) or a radiomimetic drug such as bleomycin. IR produces double strand breaks, as well as single strand breaks and other complex damage, whereas drugs like bleomycin cause direct DNA breaks, which are often converted to double strand breaks (Steighner and Povirk, 1990). Such direct DNA damaging agents cause breaks throughout the cell cycle, and at many different loci. Studies using these methods usually look for repair *en masse*, either by using cellular survival as a proxy for successful repair or by looking at the extent of DNA breakage directly by electrophoresis techniques (Singh *et al.*, 1988). Individual breaks can be studied to an extent by looking at accumulation of DNA damage response proteins in nuclear foci, particularly phosphorylated histone H2AX (γ-H2AX), which are thought to form even in response to a single break (Rothkamm *et al.*, 2003). Irradiation of only part of the nucleus can also be accomplished using a laser, providing slightly more control over the induced damage (Kong *et al.*, 2009). Some cell cycle specificity can also be achieved by using drugs that cause single strand nicks, such as camptothecin. These are converted to double strand breaks when a replication fork passes—i.e. in S phase.

Locus-specific breaks

Experiments using the DNA damaging agents described above have provided many useful insights into the biology of DNA repair. However, in most cases the amount of damage caused is far in excess of any normal physiological setting and therefore the repair pathways may be unduly stretched. The main limitation of such assays is that as the locations of the breaks are not known, it is difficult to get information on the accuracy of the repair by sequencing repaired loci. This was the main incentive for the development of methods to experimentally induce single breaks at defined positions.

As mentioned above, B lymphocyte development involves induction of breaks at the IgH and IgL loci. These are programmed by recombination signal sequences (RSSs) in the DNA in the case of V(D)J recombination or switch (S) regions in the case of class switch recombination (CSR). T lymphocyte development also involves programmed breaks at the TCR loci. Although the breaks in these cases do occur in a defined region of the genome, the exact nucleotide position can vary. There are a number of possible RSS and S sites that can be cleaved, and in the case of S regions cleavage can occur at multiple positions within the S region. However, as the resulting joins can be cloned and sequenced, this has resulted in a number of important observations about end joining pathways—for example, that junctions often contain microhomology.

The most widely used mammalian experimental system to induce DSBs at a defined locus uses the I-*Sce*I restriction endonuclease. This has an 18 bp recognition site that is not present in the mouse or human genome. The recognition site is introduced as a transgene or on a plasmid, typically combined with suitable reporter genes. Transfecting cells with an I-*Sce*I expression plasmid results in cleavage at the recognition site (Rouet *et al.*, 1994). Several reporter constructs have been developed to allow different types of repair events to be recovered and measured. These have been used to discern the relative contributions of NHEJ and HR to repair (Liang *et al.*, 1998) and to investigate repair template choice (predominantly the sister chromatid, Johnson and Jasin (2000)), to name but two. The ability to program breaks at known loci by targeting I-*Sce*I sites into the genome has also been used for other purposes, such as the demonstration that double strand breaks at the targeted locus increase gene targeting frequency (Smih *et al.*, 1995).

The cleaved I-*Sce*I site has compatible 3′ four nucleotide overhangs. Precise ligation regenerates the cleavage site, and therefore the break may persist. This could result in a bias towards inaccurate repair in the recovered events. A recent study in which the Trex1 exonuclease was co-expressed with I-*Sce*I seems to support this theory (Bennardo *et al.*, 2009).

Another similar approach that has recently become available is the use of zinc finger nucleases. These can be designed to target specific sequences in the genome (via the zinc finger domains) and cause breaks by bringing a fusion partner, *Fok*I, into proximity of the targeted locus. The great attraction is that they do not require the introduction of an ectopic recognition site, and thus are being exploited as tools to create knockouts by simply cleaving and screening for inaccurate repair events. They have also been used to stimulate gene targeting and generate translocations as well as for the study of DSB repair (Porteus and Baltimore, 2003; Bibikova *et al.*, 2003; Brunet *et al.*, 2009). As for I-*Sce*I, there is a minor caveat about persistence of the break in this context, as accurate repair regenerates the cleavage

site.

7.1.4 Aims

In this chapter I describe experiments to determine whether or not the host DNA repair pathways are involved in repair of the PB-induced DSB. I found that the classical NHEJ pathway repairs all detectable breaks in the reporter system I describe, and therefore show that PB can be used to induce DSBs at known loci. I also argue that PB has several unique properties compared to other methods of DSB induction.

7.2 Results

7.2.1 Reporter cell lines with DNA repair deficiencies

Given that PB appears to leave compatible 5′ overhangs *in vitro* (Mitra *et al.*, 2008), and therefore does not require processing prior to ligation (although this is not ruled out), I decided to first investigate the ligation step of NHEJ as a likely host pathway to handle this lesion. I used two NHEJ-deficient ES cell lines —*Xrcc4*$^{-/-}$ and *Xlf*$^{\Delta/\Delta}$, both with homozygous mutations in components of the ligation complex (Zha *et al.*, 2007). These cell lines were a kind gift from Fred Alt and Shan Zha (Harvard). As noted above, *Xrcc4*$^{-/-}$ are effectively also Ligase IV deficient.

I already had a suitable reporter construct for excision, in the form of the TV28 targeting vector used to create the *Hprt*^{PB} reporter locus in Chapter 6 (Figure 7.2). The transposon is 667 bp from the nearest *Hprt* exon in this construct. Therefore, HAT selection can be used to isolate cells which have successfully repaired the PB-induced break. Even cells which repair the break inaccurately could be isolated, provided that transcription of *Hprt* is not disrupted. Finally, as the transposon contains the *puro*Δ*TK* gene, it is also possible to select cells that have lost the transposon and *not* regained *Hprt* function (6-TG selection). This should allow recovery of larger deletions that disrupt *Hprt* function (Table 7.1).

I used my transposon targeting vector (TV28) as before to insert the transposon into the *Hprt* locus in these cell lines, and also into JM8A3 wild type cells to use as a control. For JM8A3 and *Xrcc4*$^{-/-}$ cells, the targeting vector was in the TNN (*neo* expressing) orientation. Therefore I also transfected the targeted subclones with Cre to obtain the *puro*Δ*TK* expressing transposon required for FIAU selection. For the *Xlf* mutant targeting, I treated the targeting vector plasmid with recombinant Cre *in vitro* and transfected this linearised plasmid into cells as before. As the targeting efficiency was high in the previous experiment (at least 25%; Figure 6.4), I did not use 6TG selection to directly select for *Hprt* mutants.

Targeted clones were identified by PCR genotyping at the 5′ end relative to *Hprt* (Figure 7.3). As expected, the targeting efficiency was lower in the 129-derived cell lines that are not isogenic with the targeting vector (Figure 7.3 and Table 7.2). However, even under these suboptimal conditions, the targeting frequency was still at least 12% of G418 or Puro resistant subclones. Targeting was very efficient in the C57BL/6 cell line, in which almost 70% of G418-resistant clones were targeted—more than in the *Blm*-deficient background (see Figure 6.4). For all subsequent mobilisation experiments I used multiple targeted subclones as biological replicates. All clones used were checked to confirm correct targeting at the 3′ end and resistance to 10 μM 6-TG.

7.2.2 *Xrcc4* and *Xlf* are required for survival after transposition

I transfected 10^7 cells from each cell line with 15 μg of pCMV-hyPBase expression plasmid to mobilise the transposon. I plated a small fraction of transfected cells in M15 medium (non-selective) to determine the total colony forming units in the transfected cells. The remainder of the culture was plated at a higher density, and HAT selection begun 24 hours post transfection (Figure 7.4). As a negative control, I transfected cells with an equal amount of a GFP expression plasmid. These cells were selected as above, and in addition the transfection efficiency was determined in unselected cells by flow cytometry at 48 hours post transfection. Transfection efficiency (fraction of GFP-positive cells at 48 h) ranged from 37–52%.

There was a striking drop in the proportion of HAT resistant cells obtained for the mutant lines compared to the wild type (Figure 7.5). To compare the lines, I normalised the number of HAT resistant colonies for each subclone by plating and transfection efficiency. This analysis indicated that the mean survival after transposase selection and HAT selection in *Xrcc4* mutants is only 5% of the wild type value. For *Xlf* mutants, the surviving fraction was slightly higher at 11% of wild type. This demonstrates that reconstitution of a functional *Hprt* gene after transposon excision requires NHEJ.

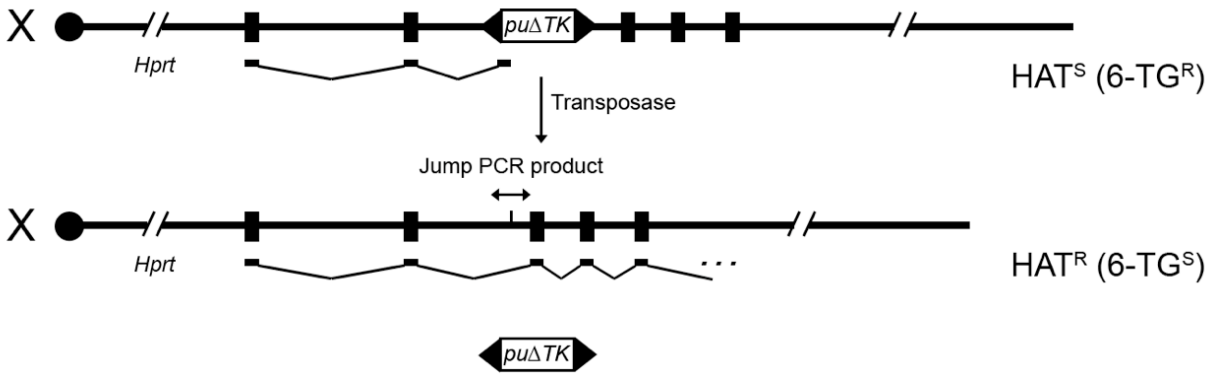


Figure 7.2: TV28 reporter locus for excision. See Chapter 6 for details

Event	Genotype	Resistance
Excision, successful repair (or small del)	*Hprt*$^+$	HAT
... with reintegration	*Hprt*$^+$, PB$^+$	HAT+Puro
Excision, no reintegration, large deletion	*Hprt*$^-$	FIAU+6-TG

Table 7.1: Transposition outcomes using the TV28 reporter locus. Selection schemes to detect transposition accompanied by successful repair (accurate or inaccurate)

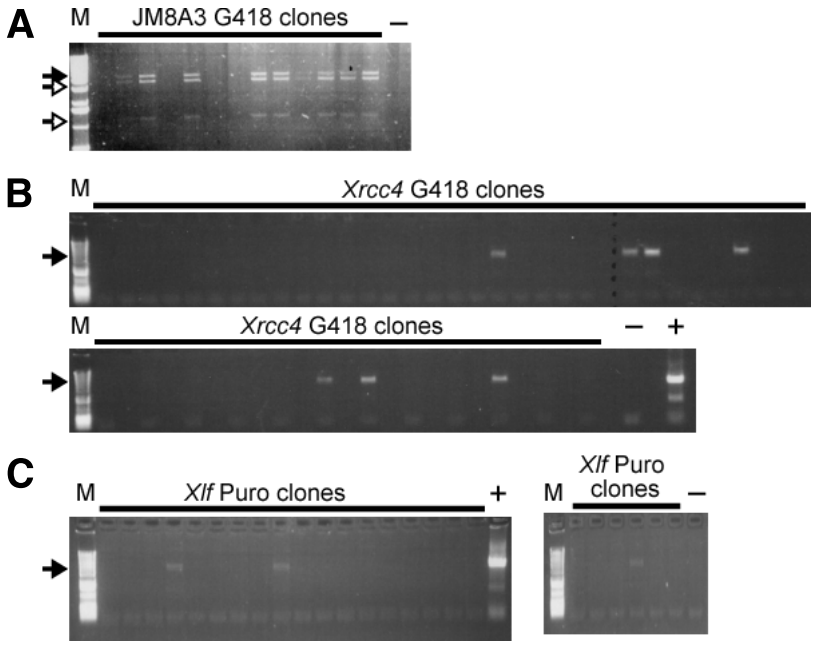


Figure 7.3: Targeting the *Hprt*^{PB} locus in NHEJ-deficient cells. PCR genotyping (as Figure 6.4) at the 5′ end of the targeting vector for A—JM8A3 (*neo* targeting vector) B—*Xrcc4*$^{-/-}$ (*neo* targeting vector) and C—*Xlf*$^{\Delta/\Delta}$ (*puro* targeting vector). The expected 4.2 kbp PCR product is shown with a filled arrow. Two smaller products (open arrows) are also amplified, although only in targeted clones, and probably arise due to one primer hybridising to a repetitive region of the PB repeat.

Cell line	Background	Genotyped	Targeted	Efficiency
JM8A3	C57BL/6N	13	9	69%
Xrcc4$^{-/-}$	129S7	60	7	12%
Xlf$^{\Delta/\Delta}$	129S7	24	3	13%

Table 7.2: Targeting efficiency in NHEJ-deficient cell lines

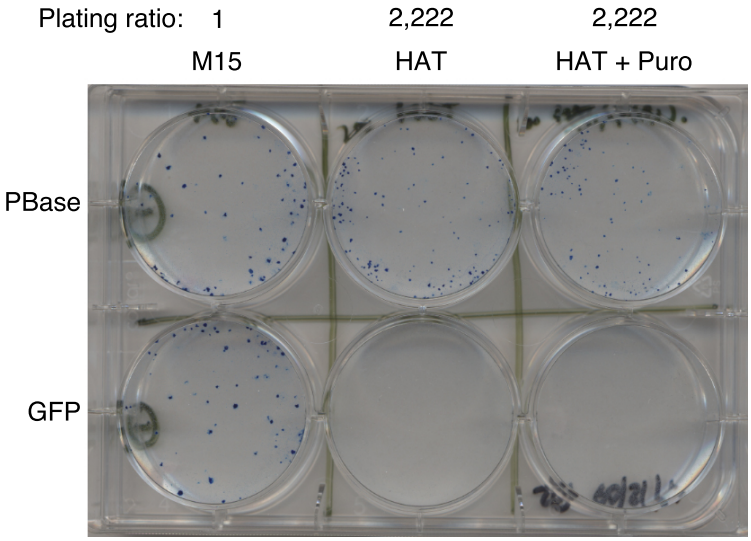


Figure 7.4: Example of transposition assay. *Xrcc4* reporter cells transfected with 15 μg of hyPBase or GFP expression plasmid are shown. M15—unselected cells to determine plating efficiency. 2,222 times as many cells are plated on the selected plates in this case. A much lower plating ratio (around 1:40) would be used for wild type cells.

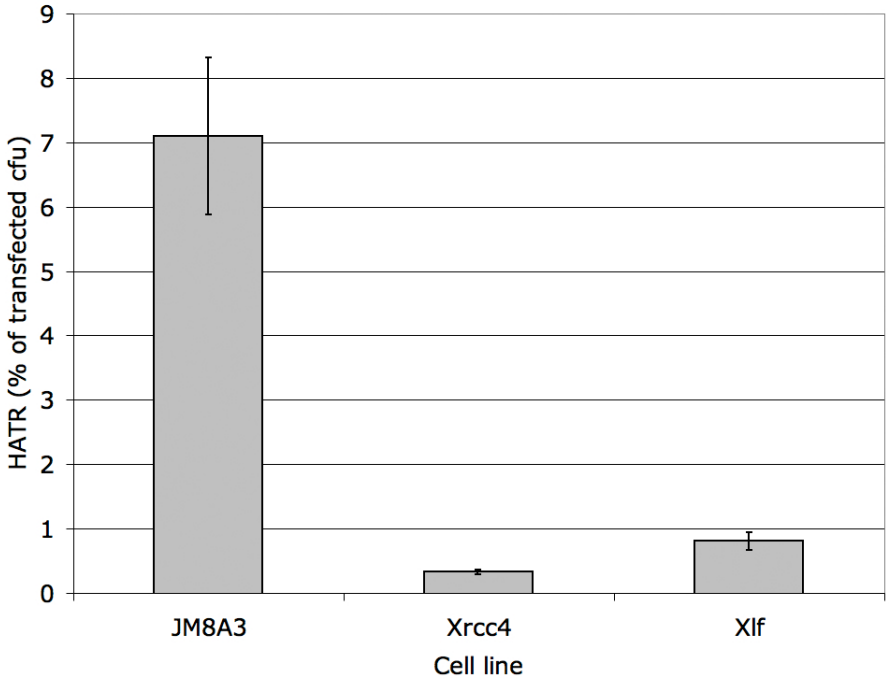


Figure 7.5: Survival in HAT medium following transfection of NHEJ reporter cell lines. The indicated wild type (JM8A3) or mutant cell lines were transfected with 15 μg pCMV-hyPBase as described in the text. The value plotted is corrected for transfection and plating efficiency. $n = 4, 6, 2$ respectively for JM8A3, *Xrcc4*$^{-/-}$, *Xlf*$^{\Delta/\Delta}$. Error bars show 95% confidence interval.

7.2.3 Mutations at the donor locus in *Xrcc4* mutants

Using primers that flank the donor site, I amplified a fragment from HAT resistant *Xrcc4*$^{-/-}$ subclones that had mobilised the transposon. PCR products from different clones, representing different transposition events, had clear size differences when separated on a 2% agarose gel (Figure 7.6A). The corresponding fragments from wild type cells were all of equal size, and sequencing revealed no mutations (Figure 7.6B and data not shown). When I sequenced the *Xrcc4*$^{-/-}$ PCR products, I found that all clones tested had deletions at the donor site, sometimes accompanied by a short insertion (Figure 7.7 and Table 7.3). The deletions without insertions were often flanked by 2–4 bp microhomologies. Some events (defined by the extent of deletion on each side) were recurrent, particularly those without insertions and with microhomology flanking the deletion. Events with insertions were more variable with respect to the extent of deletion.

As microhomologies are short by definition, and usually less than 5 bp in length, it is difficult to be sure that they do not simply occur by chance, and thus whether they are really characteristic of the repair. A formula has been developed to address this, although it assumes a random sequence of a given GC content, as it was developed to analyse non-site specific breaks (Roth *et al.*, 1985). The sequence surrounding the break in this case is always the same, and furthermore is not random, as there is some vector sequence present close to the break from the cloning procedure (see Chapter 6). Therefore a better approach would be to consider microhomology use in the context of this particular sequence. To address this, I generated a distribution of the microhomology that would be expected by chance. Taking the sequence surrounding the breakpoint, I modelled a random resection of up to 20 nt at each end to determine how often microhomologies of 1–4 nt would be encountered if resection was randomly terminated. Plotting these with the experimental data shows a clear increase of junction microhomology of two or more nucleotides in the sequenced junctions compared to that expected by chance (Figure 7.8).

In other cases the deletion was accompanied by an insertion. In most cases these were short and not obviously derived from surrounding sequence. I only isolated a single event that could be classified as accurate with respect to the TTAA site, but this had a single base pair deletion immediately downstream, so could also have resulted from a deletion and reinsertion of nucleotides. These data indicate that *Xrcc4* is required for accurate repair of all PB induced breaks.

In most cases where a deletion flanking the donor locus was accompanied by an insertion, this was short and not uniquely mappable to the genome. These could arise from untemplated nucleotide additions by polymerase enzymes during end processing. It is possible that such additions could generate new microhomologies, which can then be used to anneal the two ends to each other. However, in three cases I observed larger mutations with a clear structure. Two of these had a duplication of sequence from both sides of the break. However, the arrangement of the sequences from either side of the break was shuffled (Figure 7.9 and Figure 7.10A,B). One possible way to generate this structure might be a duplication after repair is complete, although the event shown in Figure 7.10B is not perfectly duplicated.

Another event with a large insertion turned out to have the terminal 245 bp of the PB5 end of the transposon remaining in the locus, which was then joined to downstream genomic sequence. Two nucleotides of microhomology were present at the site of joining (Figure 7.11). One possible explanation for this is that only one end of the transposon was cleaved, followed by extensive degradation and rejoining. However, as the genomic end adjacent to PB5 was not degraded, a more likely possibility may be that this event occurred in late S/G2 phase and involved some homology-directed repair from the unjumped sister chromatid, followed by microhomology-mediated joining to the other free end.

7.2.4 Low frequency of mutations at the donor locus in *Xlf* mutants

I also sequenced the donor locus in 44 subclones from the *Xlf* mutant cells (Figure 7.12 and Table 7.4). In contrast to the *Xrcc4* mutants, most (37/44) repair events were precise in these cells. This is in broad agreement with the results of an extrachromosomal V(D)J recombination assay in these cells (Zha *et al.*, 2007). Three clones had deletions with clear flanking microhomologies, and two events were also recovered with structured insertions—one with a 72 bp repetitive insertion, and one with a duplication of 16 bp of sequence from one side of the break.

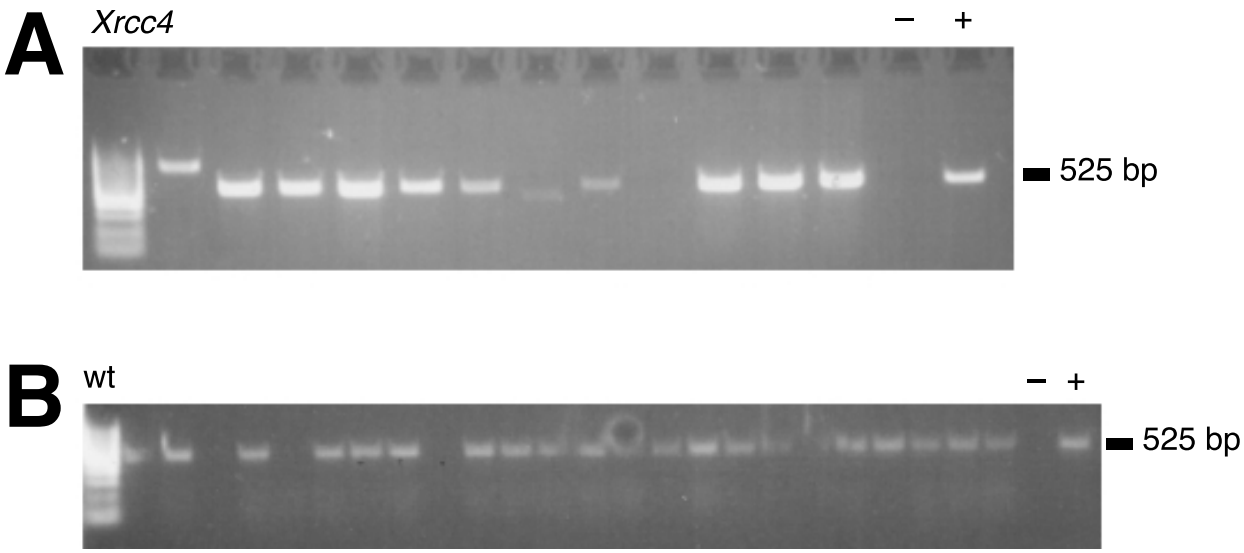


Figure 7.6: PCR amplification of donor locus after transposition. A—Products from *Xrcc4* mutants are different sizes, indicating insertions and/or deletions have occurred. B—All products from wild type cells are normal. +: Template DNA from known mobilised LGN cells, –: No template DNA added.

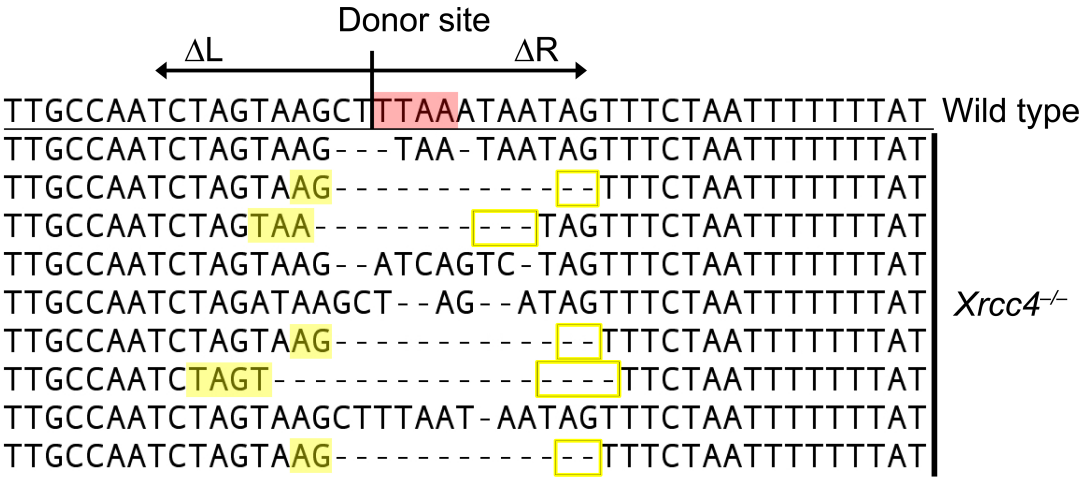


Figure 7.7: Examples of mutations at the donor site in *Xrcc4* mutants. The cleavage point on the strand shown is indicated by a vertical bar. ΔL,R show the deleted base pairs as summarised in Table 7.3. Microhomologies flanking the deleted sequence are highlighted in yellow on the left of the deletion, the position of the corresponding identical sequence on the right, which is deleted, is boxed (refer to the aligned accurate repair sequence on the top line).

# clones	ΔL	Insertion	ΔR	μ-hom
1	0	AG	4	
1	0	AG	5	
1	0		6	T
1	0	AG	7	
1	0		8	
1	0		9	T
1	0	AACA	10	
1	0		16	CT
1	0	TTAA	5**	
1	1	19 bp*	5	
1	1	AAACTAA	5	
1	1	30 bp*	7	
1	2	275 bp*	0	
2	2	T	3	
1	2		5	
1	2	T	7	
1	2	ATCAGTC	8	
7	2		11	AG
1	2	TAATAACTGATT	105	
1	3		5	
4	3		8	TAA
1	3		11	
2	5		12	TAGT
1	7	A	15	
1	10		9	AAT
36				

Table 7.3: Sequencing of the repair site in *Xrcc4* mutants. For the wild type sequence see Figure 7.7. ΔL—number of base pairs deleted on the 'left' side (5′ with respect to *Hprt*), ΔR—size of deletion on 'right' side, μhom—microhomology observed flanking deletion. Insertions marked with * are shown in more detail in Figure 7.9. ** Could also be classified as accurate, with a 1 bp deletion.

# clones	ΔL	Insertion	ΔR	μ-hom
37	0		0	
1	0		3	
1	0	TAGATTAGTTTCTAAT	8	
1	0		9	T
1	3	$(CCCTAA)_{12}$	5	TAA
1	3	*	8	TAA
1	4		5	TA
1	5		12	TAGT
44				

Table 7.4: Mutations at donor locus in *Xlf* mutants. * Duplication of sequence adjacent to the breakpoint as shown in Figure 7.10C. The insertion shown in row 3 was not uniquely mappable.

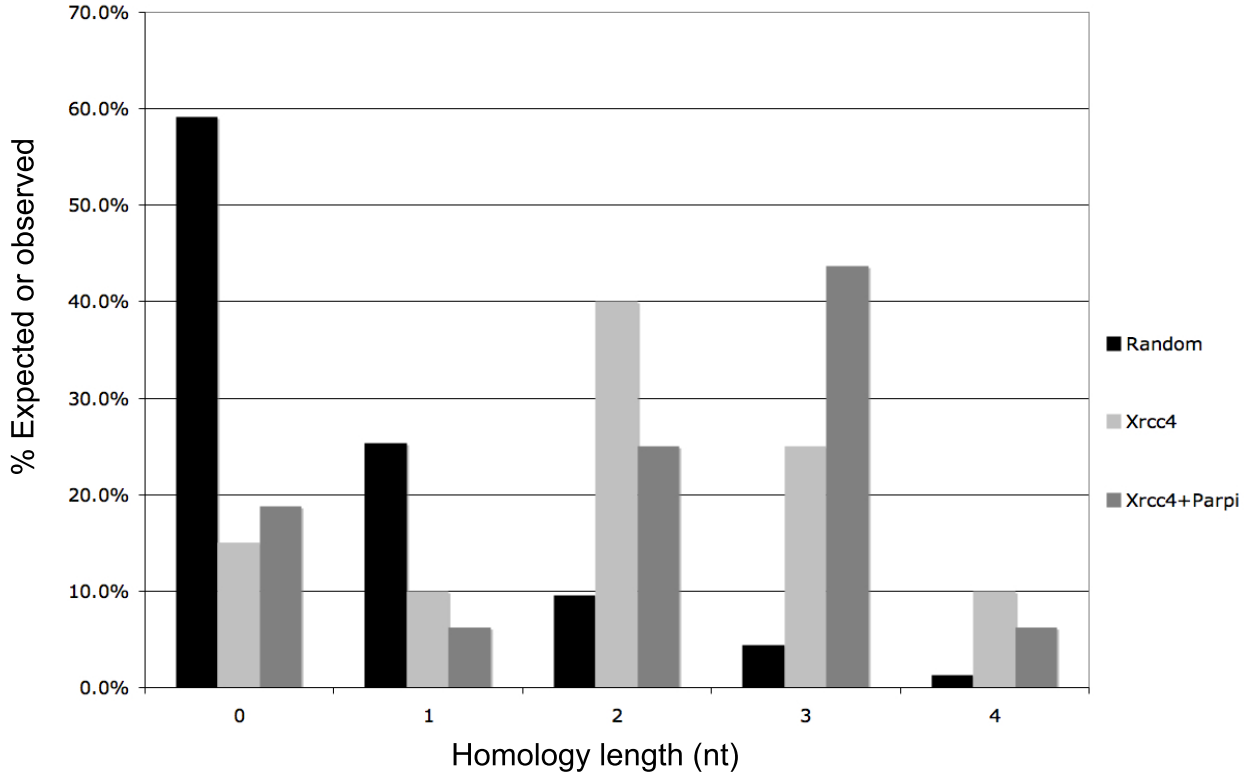


Figure 7.8: Expected and observed use of microhomology of the indicated lengths at repair sites. Expected value (black bars) is calculated based on a random resection of the break (up to 20 nt). The observed distribution in *Xrcc4* mutants is plotted in light grey. Dark grey bars—*Xrcc4* mutants treated with PARP inhibitor as described in text. Repair events that also contained an insertion are omitted; the inserted nucleotides may have generated novel microhomology, but this cannot be concluded from the final sequence.

	Xrcc4	*Xlf*
Accurate	1	37
Deletion (w/μhom)	21 (17)	4 (3)
Deletion + insertion	11	1
Deletion + insertion with clear structure:		
Duplication of sequence from both sides	2	
Truncated transposon (275bp)	1	
One-sided duplication		1
Short repeats (CCCTAA)x12		1
	36	44

Figure 7.9: Structure of insertions at excision site in NHEJ mutants. The transposon is shown as a dark arrow. Not to scale.

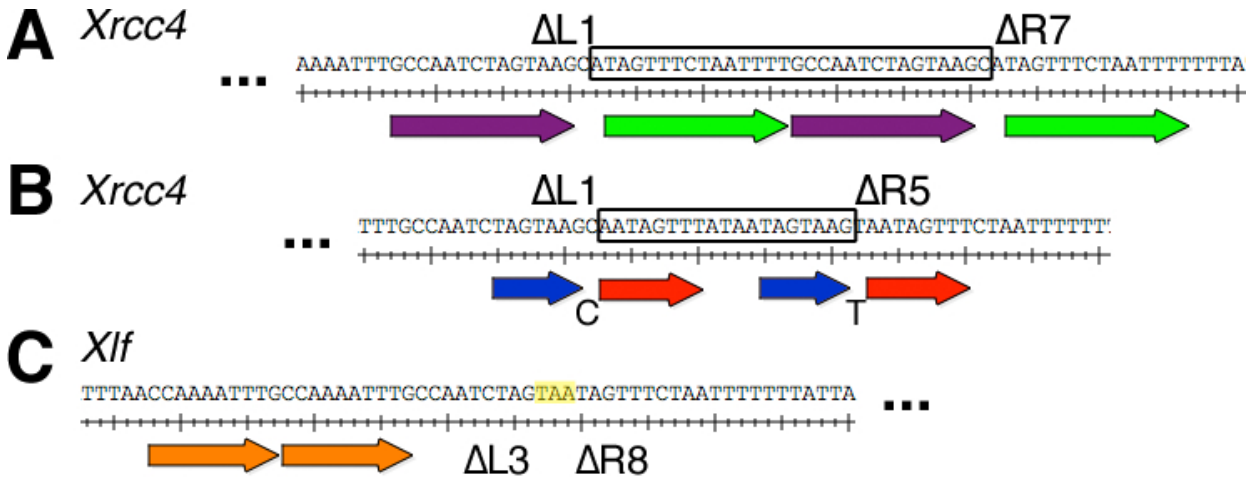


Figure 7.10: Sequence of insertions with clear structure. A, B—duplications from *Xrcc4* mutants. The duplication in A is perfect while the sequence in B has differences in the nucleotides separating the individual and pairs of repeats. The inserted sequence is shown in a box. C—Sequence from *Xlf* mutant showing a tandem duplication upstream of the excision site. There is a deletion at the excision site (with associated TAA microhomology highlighted in yellow). ΔL, ΔR give sizes of deletion observed either side of the break, as Figure 7.7.

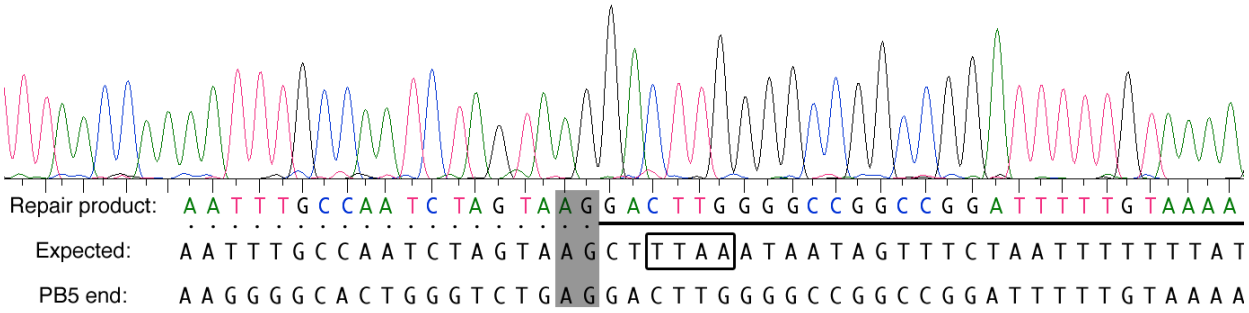


Figure 7.11: Junction sequence of a repair event retaining part of the transposon, isolated from the *Xrcc4* mutant. Sequence and chromatogram of the PCR product sequence is shown, aligned with the expected sequence in the case of accurate repair, and the proximal PB5 end sequence. Potential microhomology (AG) at the site of joining is highlighted; the excision site is shown in a box in the expected sequence.

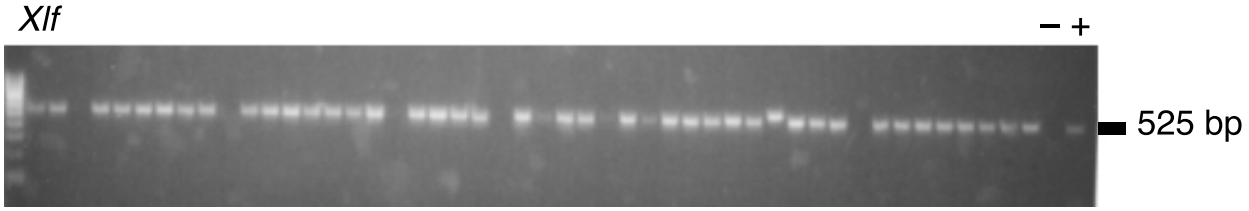


Figure 7.12: PCR amplification of donor locus from *Xlf* mutants, as Figure 7.6.

7.2.5 No evidence for larger deletions related to transposition in *Xrcc4* mutants

As many repair products from *Xrcc4* mutants had deletions, it is possible that the decrease in HAT resistant clones is due to large deletions that destroy *Hprt* function. To test this, I repeated the transfection of transposase and subsequently cultured the cells for three days without selection. I then replated the culture in selective medium containing 6-TG and FIAU at low density (1×10^5 cells per 90 mm plate). This low density is necessary to avoid cross-killing of *Hprt* negative cells by nearby *Hprt* positive cells that can metabolise 6-TG.

Colonies were obtained on the selective plates in this experiment even without transfection of the PBase plasmid (Figure 7.13). These may arise from a high mutation rate or silencing affecting the *Puro-ΔTK* gene in the *Xrcc4* mutant cells. However, there was no obvious increase in the number of colonies in cells transfected with the transposase plasmid. This suggests that the 'missing' transposition events that are not recovered under HAT selection in the *Xrcc4* mutant cells do not arise from large deletions that destroy *Hprt*.

7.2.6 PARP inhibition does not affect repair in the absence of *Xrcc4*

An increased use of microhomology during end joining has been reported for cells defective in several of the components of the core NHEJ pathway. This has been termed, variously, backup end joining (B-NHEJ) or alternative end joining (Alt- or A-NHEJ). However, as the factors responsible have not been conclusively identified, it remains to be seen whether describing alt-NHEJ as a distinct pathway is accurate. One obvious requirement is a ligase. There are only two other ligases apart from Ligase IV in mammals: Ligase I and Ligase III. Biochemical experiments indicate that Ligase III is probably responsible for joining in the absence of Ligase IV (and also in the absence of XRCC4). Ligase III is an essential gene in mammalian cells (Puebla-Osorio *et al.*, 2006), and no specific inhibitors are currently available. This precludes further analysis using my system. Another factor implicated in the alt-NHEJ process is PARP activity. Inhibition of PARP in Ku deficient cells resulted in a further reduction in end joining, although this was not seen in Ligase IV deficient cells (Wang *et al.*, 2006). As highly potent and specific PARP inhibitors are available, I decided to see if these would affect end joining in my *Xrcc4*-deficient system. I treated reporter cells for two hours prior to transfection with 10 μM KU-0058948 (A gift from S.P. Jackson and KuDOS/AstraZeneca, (Farmer *et al.*, 2005)), then carried out electroporations as above. The cells were maintained in medium with the inhibitor for 24 hours, then selected in HAT medium without inhibitor. Treatment with the inhibitor affected cell viability, but HAT resistant colonies could still be obtained (Figure 7.14A, B). Taking the lower viability into account, neither excision nor reintegration appeared to be affected in cells treated with the inhibitor (Figure 7.14C). I amplified and sequenced 21 donor sites in total. The mutation spectrum appeared similar to the untreated *Xrcc4*-deficient cells, with no significant change in deletion length, types of mutation observed or microhomology use (Table 7.5 and Figures 7.8 and 7.16). There was a slightly higher proportion of microhomology-mediated deletions relative to deletions accompanied by short insertions, but this was not statistically significant ($P = 0.13$, one-sided binomial test). Recurrent events were not observed in the presence of the inhibitor (Table 7.5). This could reflect a role of PARP in regulating end processing, leading to a higher diversity of joining events, but more events would need to be analysed to investigate this. These minor alterations in the types of repair event aside, PARP activity does not seem to be required for repair in the absence of *Xrcc4*. This is in agreement with the results obtained in Ligase IV deficient cells (Wang *et al.*, 2006). These data support the conclusion that PARP acts upstream of Xrcc4-Ligase IV in end-joining pathway choice.

7.2.7 Excision and reintegration are not affected by inhibitors of PARP, ATM or DNA-PKcs in wild type cells

The recent development of potent and specific inhibitors of DNA repair enzymes for potential therapeutic use has provided a new set of tools for the study of DNA repair (Jackson, 2009). The PB system for precise induction of DSBs complements these drugs well, as a variety of perturbations can be studied using the same reporter cell line and the same break. I used my wild type reporter cell line with some of these small molecule inhibitors to address several questions.

First, although the experiments described above demonstrate that the host NHEJ machinery is involved in repair of the transposon excision site, whether the host DNA repair machinery is involved in reintegration *in vivo* remains to be determined. The

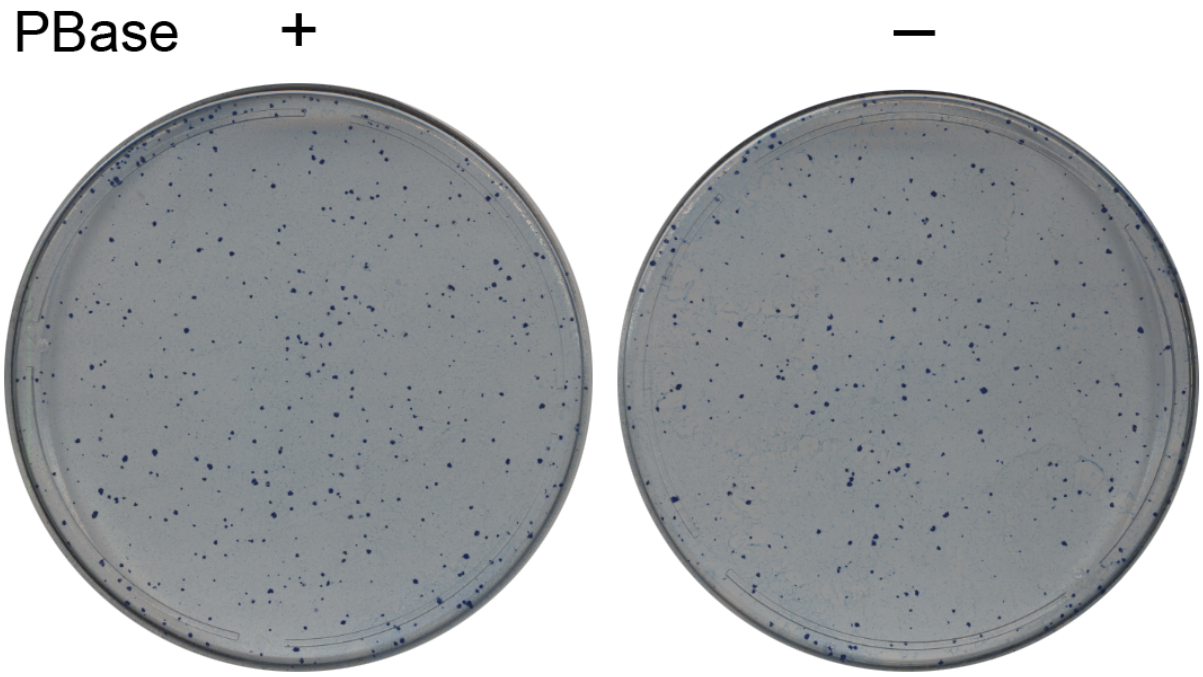


Figure 7.13: FIAU+6-TG selection to detect large deletions. *Xrcc4*$^{-/-}$ cells were transfected with hyPBase (+) or GFP (–) expression plasmids, and replated in FIAU+6-TG after three days. There is a high background, but no obvious PBase-dependent increase.

# clones	ΔL	Insertion	ΔR	μ-hom
1	0		0	
1	0	CT	4	
1	0	TATAATTA	4	
1	0		7	
1	0	TTTATTAG	13	
1	1		3	TTA
2	1		6	
4	1		10	TTA
1	1		21	A
1	4		10	TA
1	5		8	CT
1	5	TACTAATTGAATTG(AAAAATTAGA)AGCT	8	
1	6		11	AC
1	12		17	ATT
1	18		1	TAAA
1	20		10	TA
1	42		7	AGC
21				

Table 7.5: Mutations at the site of repair in *Xrcc4* cells treated with a PARP inhibitor. The bracketed portion of one insertion is mappable, to a sequence just downstream of the break on the reverse strand; otherwise the insertions are not uniquely mappable.

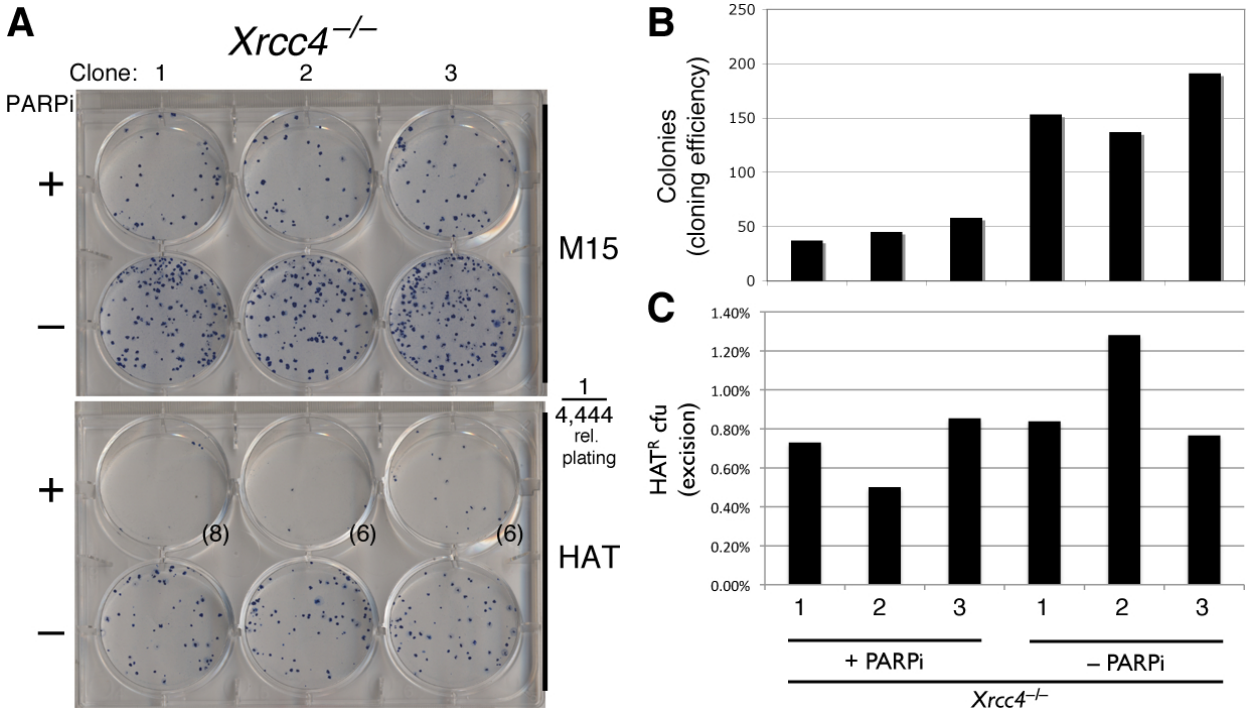


Figure 7.14: Results of transposition in *Xrcc4* mutant cells treated with PARP inhibitor. A—A low frequency of HAT resistant colonies are obtained, showing that excision is not completely abolished. PARPi treatment also reduces the number of colonies on the untransfected plate (M15). Numbers in brackets for plates with few colonies show number of colonies that were picked for analysis. B—Colony counts of unselected colonies. C—Frequency of HAT resistant cells post-transposition (corrected for cloning efficiency) is not affected by PARPi treatment. Results are from three independent subclones.

Type	*Xrcc4*$^{-/-}$	*Xrcc4*$^{-/-}$ + PARPi	*Xlf*$^{\Delta/\Delta}$	w.t.
Normal	1	1	37	17
μhom del	17	13	3	0
other del	3	3	1	0
del + ins $<$ 5nt	9	1	0	0
del + ins $>$ 5nt	6	3	3	0
Total analysed	36	21	44	17

Table 7.6: Summary of types of event observed in different mutants

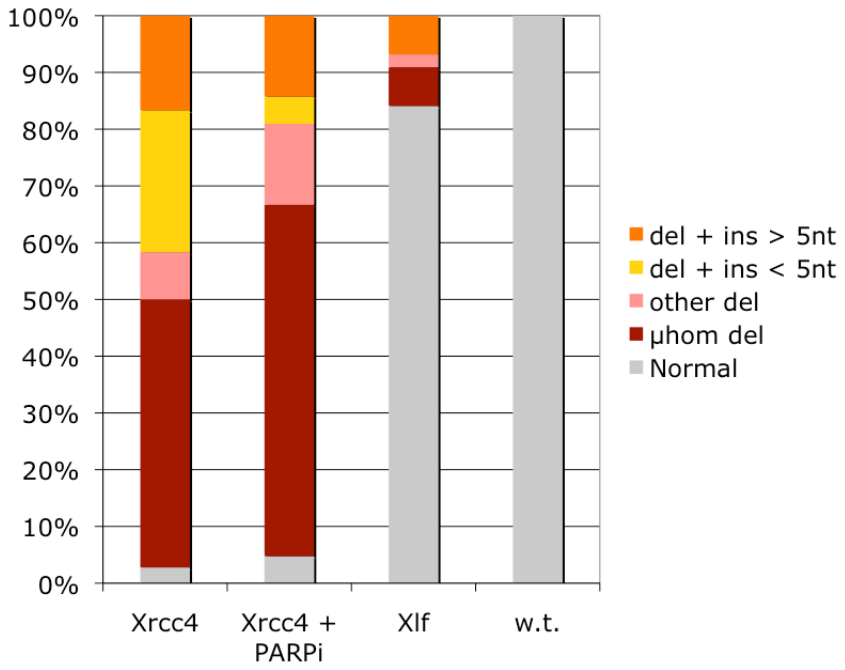


Figure 7.15: Graph showing frequency of repair event classes in different mutants

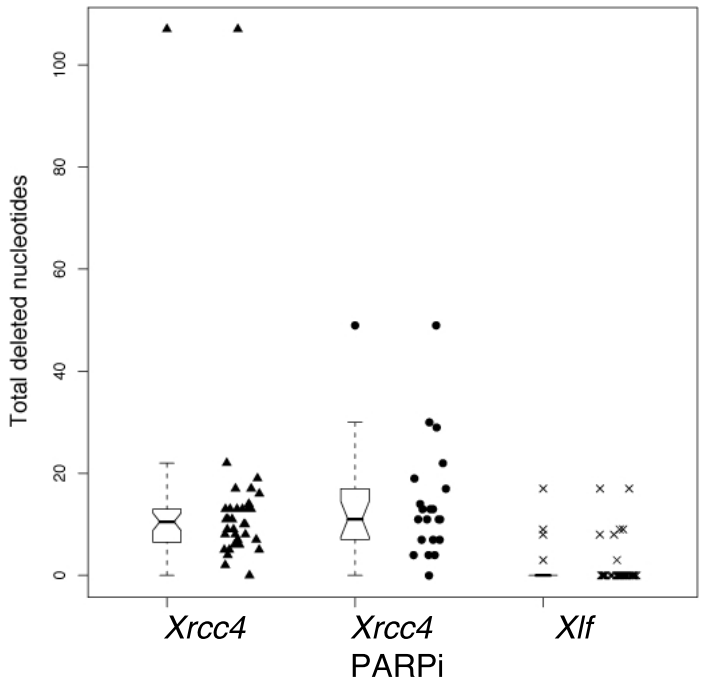


Figure 7.16: Distribution of total deletion size (ΔL + ΔR) at repair site in NHEJ mutants.

transposase is sufficient to join at least one strand *in vitro* (Mitra *et al.*, 2008). As the excised transposon is likely to be capped by hairpins, these could be dependent on DNA-PKcs for processing by analogy with coding ends in V(D)J recombination. I treated cells with the DNA-PKcs inhibitor NU-7441 (Tocris Bioscience, used at 1 μM, 2 h pre-transfection, 24 h post-transfection) and carried out the transposition assay as above. Both HAT and HAT+Puro resistant clones were obtained, at frequencies similar to untreated cells. This indicates that reintegration is not dependent on DNA-PKcs (Figure 7.17), and suggests that the observation that the transposase is sufficient to join excised transposons to the target site *in vitro* applies *in vivo*.

I also tested inhibitors of ATM and PARP using this system. ATM is required for the repair for some DSBs, but what determines whether or not it is required for a particular break is unclear. Using the ATM inhibitor KU-55933 (Tocris Bioscience, used at 10 μM with pretreatment as for NU-7441, above), I determined that ATM is not required for the repair of the PB-induced break in my reporter cells (Figure 7.17).

I also checked the effect of PARP inhibitors, and obtained similar results to the *Xrcc4* mutants above—i.e. a decrease in cell survival that was not PB-dependent, and no clear change in the excision or reintegration frequency (Figure 7.17).

7.2.8 Homologous recombination repair of PB-induced breaks

The cell lines described in this chapter do not allow repair of the break by HR to be assessed directly. The most likely template for HR is the sister chromatid. In the reporter cells, the sister chromatid will also contain a transposon—therefore repair of the break by gene conversion will restore the transposon in the *Hprt* locus and not result in HAT resistance. Even if the transposons on both sister chromatids are mobilised, there is no way to tell sister chromatids apart at the sequence level. Previous methods for detection of HR using the sister chromatid have used reporters with direct repeats, where crossing over between the distal repeat unit on one chromatid with the proximal on the other results in three copies of the repeat on one of the resulting chromatids (Johnson and Jasin, 2000). HR using a homologous chromosome is not possible in my system, as the reporter is on the X chromosome.

7.3 Discussion

7.3.1 Requirement for host repair pathways in repair of PB-induced breaks

Experiments using the reporter cell lines described in this chapter show that *Xrcc4* and *Xlf*, components of the NHEJ pathway, are required for recovery of HAT-resistant clones after transposition. There are two possible reasons for the failure to recover HAT-resistant clones, assuming the same number of excision events from the *Hprt* locus. One explanation is that NHEJ mutant cells cannot repair the break, and subsequently die or enter senescence and do not form a colony. Alternatively, the break could be repaired imprecisely and in the process destroy *Hprt* function, for example by causing a large deletion.

However, although the system is able to detect deletions of at least 100 bp (Figure 7.16), almost all deletions are distributed in the 1–20 bp interval. If this size distribution is in fact bimodal, with a second peak of undetected large deletions, this could explain the results. It is difficult to envisage a mechanism for such a distribution based on known DNA repair mutant phenotypes, and the results of FIAU+6-TG selection (Figure 7.13) suggest that there are not a large number of cells bearing large deletions affecting *Hprt*.

Another alternative to death would be restoration of a transposon at *Hprt* from the sister chromatid by HR, if excision occurs in S or G2 phase. These would also not be picked up by the HAT selection system, nor by FIAU+6-TG selection. One potential improvement to the reporter system would be to select for the excision independently of *Hprt* function. This could be accomplished by using a gene trap transposon at a known locus where a gene is not trapped. Mobilisation of this transposon could be selected for by selection for reintegration events that do trap a gene. The original locus could then be examined by PCR or Southern blotting, allowing the full range of mutations to be detected.

It should be noted that although this system allows a single repair event to be studied at the donor locus, there may be multiple breaks elsewhere in the cell if the transposon reintegrates and jumps again. This may affect the sensitivity measurements, as there may be more than one break in some cells (if the transposon jumps again before repair of the previous break), or breaks induced persistently over the expression period of the transposase. Little is known about the kinetics of PB transposition, so the effect is hard to predict.

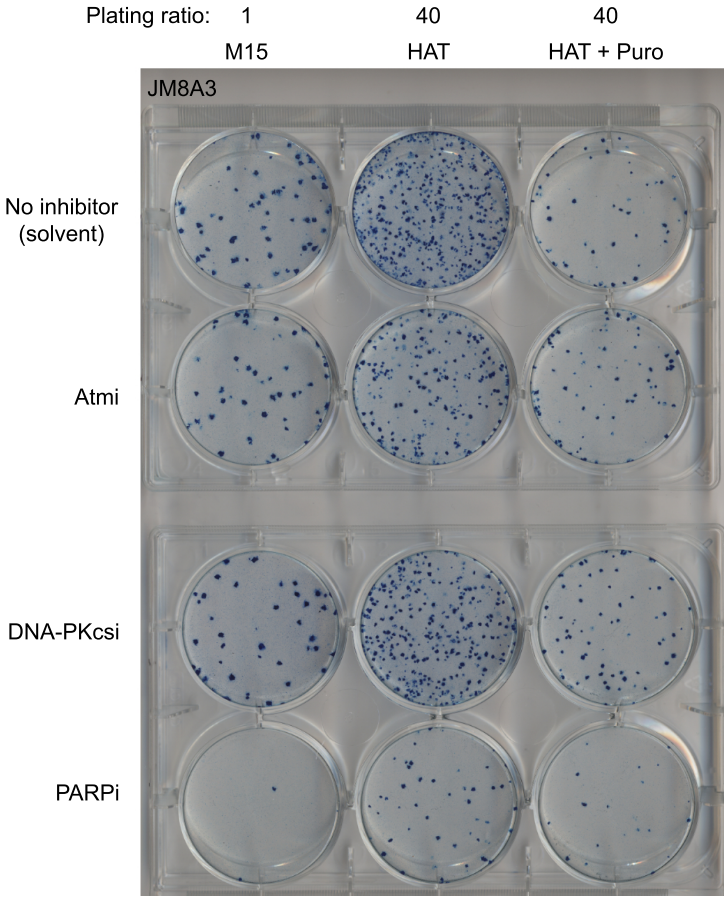


Figure 7.17: Transposition assay in wild type cells treated with ATM, DNA-PKcs and PARP inhibitors as indicated.

It would be interesting to investigate whether there is any involvement of HR in repair of the transposon. Gene targeting is dependent on components of the HR pathway (Essers *et al.*, 1997; de Wind *et al.*, 1995), and double strand breaks introduced by I-*Sce*I or zinc finger nucleases stimulate gene targeting at the locus of the break (Smih *et al.*, 1995). Therefore, if PB induced breaks are indeed processed in the same way as an endogenous break, they should also stimulate gene targeting at the donor locus. This could be investigated by attempting to stimulate targeting at the *Hprt* locus in the reporter cell line by transfection with PBase.

Potential effect of genetic background

The two mutant cell lines used are from the 129S7 genetic background, while the wild type cells used for comparison are C57BL/6N. Therefore any differences could potentially arise from different genetic backgrounds. Similar transposition assays have been carried out in 129S6 and 129S6×C57BL/6J genetic backgrounds by colleagues (Wang *et al.* (2008); Liang *et al.* (2009) and K. Yusa, unpublished) with similar excision efficiencies obtained. No differences in DNA repair have been documented between the genetic backgrounds used, and as the *Xrcc4* mutant defect is so severe and produces a known phenotype with respect to the structure of the recovered products, it is unlikely that genetic background alone could be responsible for the difference. However to formally prove this, the experiment should be repeated in TC1 wild type cells [the 129S7 cell line that the NHEJ mutants were derived from], or in complemented cells expressing *Xrcc4* or *Xlf* transgenes as appropriate.

7.3.2 Differential requirement for *Xrcc4* and *Xlf* at PB-induced breaks

The two NHEJ mutant cell lines studied *are* derived from the same parental cell line, and can be directly compared. It had been previously noted that $Xlf^{\Delta/\Delta}$ cells had a less severe radiosensitivity compared to $Xrcc4^{-/-}$ cells. Western blotting and overexpression experiments have confirmed that the *Xlf* allele is a genuine null (Li *et al.*, 2008). This implies that *Xlf* is dispensable for repair of some IR-induced lesions. IR causes different types of break, often with complex structures, at different loci as well as causing multiple lesions per cell.

My results show that the difference in IR sensitivity between these two mutant lines extends to the single break caused by PB transposition. The constant nature and location of the break in my system raises the question of what the basis is for this differential requirement for the two NHEJ factors. One possible explanation may be a difference in NHEJ at different stages of the cell cycle; this is something that could be addressed using the G1-specific PBase-CDT1 fusion protein described in the previous chapter.

7.3.3 DNA repair requirements in V(D)J recombination and PB transposition

The mechanism of PB excision is similar to V(D)J recombination, in that hairpin ends are produced that then need to be resolved before joining. In V(D)J recombination, these hairpin ends (coding ends) require DNA-PKcs and Artemis for repair. Structurally these are analogous to the PB transposon ends (Figure 7.18). I therefore asked whether PB reintegration, as opposed to excision, had similar requirements. As I did not have DNA-PKcs or Artemis deficient ES cells available, I used a recently developed DNA-PKcs inhibitor instead. The experiment did not show an effect (Figure 7.17). It should be noted that DNA-PKcs inhibitors or kinase-dead DNA-PKcs mutants do not always reproduce the phenotype of DNA-PKcs knockouts, suggesting that DNA-PKcs also performs a structural role in DSB repair that is separable from its kinase activity. However, as the transposase is sufficient for hairpin resolution in a defined *in vitro* system, I favour the explanation that this extends to the *in vivo* situation. Thus the host DNA repair pathway is only responsible for repair of the excision site and not involved in reintegration.

This is perhaps reasonable considered in the light of evolution: DNA repair pathways are efficient and highly conserved across vertebrates; therefore there is probably no need for the transposon itself to handle excision site repair, and no selective pressure for this function. The DNA repair machinery does suppress translocations and is likely to fuse free transposon ends to form a circle, as with V(D)J recombination signal ends, so the transposase needs to take control of the reintegration. Interestingly, both the RAG1/2 recombinase and SB transposase appear to interact with Ku, which is proposed to channel repair of the excision site into the NHEJ pathway. It would be interesting to see if PB has a similar association.

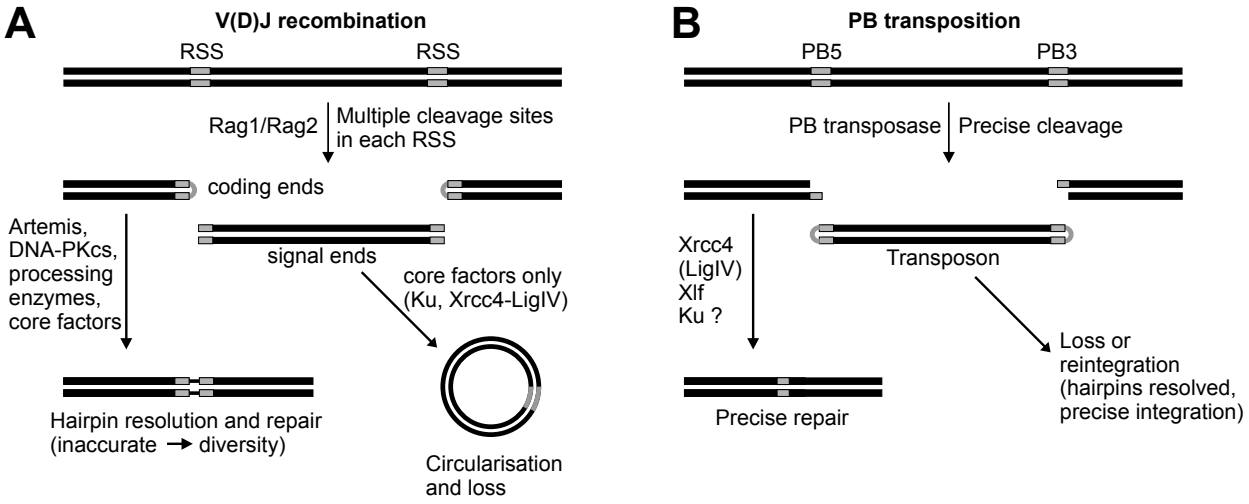


Figure 7.18: Requirements of NHEJ factors in V(D)J recombination and PB transposition. The structure of the different DNA ends formed and subsequent products are shown (not to scale). A—V(D)J recombination, B—PB transposition

7.3.4 DNA repair requirements in SB transposition

The requirements for various DNA repair factors in SB transposition have previously been investigated (Izsvák *et al.*, 2004). It should be emphasised that while SB and PB are often thought of as similar in terms of their role as mouse genetic tools, they belong to distinct transposon families with different mechanisms. However, as both cause double strand breaks, there is likely to be some similarity in the requirement for host repair pathways. Izsvák *et al.* used an integration assay in Chinese hamster ovary (CHO) cells with various DNA repair defects. Using this system, the authors concluded that SB transposition required Ku80, XRCC4, ATM (in some cases) and DNA-PKcs, although not DNA-PKcs kinase activity. Although transposition in this case was from a plasmid, junctions could be recovered by PCR and sequenced. In all the mutant lines for which junctions were recovered, there were deletions flanking the excision site, and for ATM and Ku80 mutants these appeared to be flanked by microhomology. Interestingly the authors of this study were unable to recover junctions from XRCC4 mutants, whereas I could readily amplify PCR products.

There are several differences between the assays used that could explain this. First, the breaks left by the two transposons may not be dealt with in the same way. SB leaves incompatible overhangs, and as mentioned above, the transposase may directly interact with Ku to affect repair. Second, the break in the SB assay is on a plasmid, while in my system, the break is genomic and thus presumably occurs in an appropriate chromatin context. Transiently transfected plasmids may not reflect the situation for genomic breaks. Finally, my system incorporates selection for repair and subcloning prior to PCR and sequencing, which may make it easier to isolate rare repair events, e.g. those in *Xrcc4* mutants. A further advantage is that the distribution of accurate/inaccurate repair events in the products is not affected by bias in PCR, gel cutting or subcloning.

7.3.5 Advantages of PB for programming double strand breaks

The experiments described here suggest an alternative use of PB: as a system to create locus-specific DSBs in cells and study their repair. PB has a number of features that distinguish it from other enzymes used for this purpose (Table 7.7). Using PB, cell lines with single copy insertions at known random positions can be easily generated (see e.g. Figure 5.2). With a transposon carrying suitable negative selectable markers to detect excision, DSBs at a variety of different loci could be studied. It would be interesting to use such a method to determine if breaks at some loci require Atm activity for repair—in my system, the break is at an expressed locus. IR-induced breaks that require Atm for repair are associated with heterochromatin, suggesting that knowing the exact locus of a break and its

dependence on Atm would be useful in investigating this further (Goodarzi *et al.*, 2008). This could be done in many different cell types without the need for gene targeting or extensive screening for single copy transgenics to introduce enzyme recognition sites for I-*Sce*I. One potential caveat would be that PB may have a preference for euchromatin (see Chapter 3), so the SB transposon, which has different epigenetic preferences (Wang *et al.*, 2008), may be a better option for such a study.

One unique aspect of PB is that the excision site cannot be recleaved after transposition. This is in contrast to the endonucleases where accurate repair reconstitutes the recognition site, which can then be recleaved. Thus I-*Sce*I-induced breaks, for example, are persistently recleaved until they are repaired inaccurately. This could lead to a bias towards inaccurate repair events in the observed products, which does not reflect the actual accuracy of the process. In my system, each HAT-resistant colony represents a single repair event that can be easily subcloned and analysed. These attributes of PB make it a useful tool for making careful measurements of repair accuracy under different circumstances, and provide a method for simple analysis of mutations at the sequence level. It would be particularly interesting to use cell cycle specific transposase enzymes, as described in Chapter 6, to induce and study cell cycle specific breaks.

Enzyme	Rec. site	End structure	Persistant?	Transgenic req?
I-*Sce*I	18 bp	3′ 4 nt	Yes	Yes
Zn-finger-*Fok*I	customisable	5′ 4 nt	Yes	No
PBase	4 bp + Tn	5′ 4 nt	No	No

Table 7.7: Comparison of different site specific nuclease systems for causing experimental DSBs

Chapter 8

Discussion

8.1 Enrichment for homozygous mutants in *Blm*-deficient ES cells

In Chapters 5 and 6 I described the development of a method to enrich for homozygous mutants by selection for the copy number increase that occurs during loss-of-heterozygosity (LOH). I developed a transposon carrying a double selection construct that can be used to isolate cells with two (or more) copies of the construct. I first used clonal cultures, in which all cells had the same heterozygous transposon insertion site to begin with. After expansion of the culture, I was able to isolate double resistant cells that had increased the copy number of the construct.

From these experiments it was clear that LOH leading to segregation of a homozygous daughter cell is not the only pathway for copy number gain in *Blm*-deficient ES cells. The wild type copy of the mutated locus could still be detected by PCR or Southern blot in some subclones isolated from the double resistant population. This was in addition to the two distinct forms of the selection construct, indicating that three alleles were present in some cells, including a wild-type. I refer to these subclones as 'wild-type retaining' clones to distinguish them from the genuine homozygous mutants. I interpreted these results as arising from chromosomal instability, and indeed I found that some of the wild-type retaining subclones had a near-tetraploid karyotype.

The average proportion of genuine homozygous subclones isolated from a given clone was 34%. While this level of enrichment is sufficient to easily obtain homozygous mutants by subcloning, this may not be enough to use the unsubcloned double resistant population directly in screening assays. Therefore I sought to adapt the method to produce clonally pure populations that would be suitable for screening directly. This required several technical improvements to strictly limit the initial copy number of the transposon to one by chromosomal mobilisation of the transposon specifically in G1 phase of the cell cycle (Chapter 6). Doing this removes the requirement to subclone cultures immediately after mutagenesis, and enables a mixed pool of mutants to be grown together. Analysis of 45 double resistant subclones revealed 19 mutants with two allelic insertions, representing 16 different insertion sites. Thus, this procedure can produce clonally pure mutants with a single subcloning step, without severe redundancy with respect to the number of different insertion sites.

8.1.1 Future improvements to library generation

The most obvious improvement required is the generation of large libraries with tens of thousands of mutants. The limiting factor in the experiments reported here was the low mobilisation efficiency using mRNA. As thousands of new transposon insertions can be obtained from transfection of 10^7 cells with PB-CDT1 plasmid, it should be possible to improve the efficiency of mobilisation using mRNA.

In the library analysed, the remaining 26 of the 45 subclones had two non-allelic insertions. However, in 25 of these cases, one of the insertions remained at the donor locus. Finding a way to eliminate these would help to increase the proportion of useful mutants in the enriched library. I plan to investigate these to see if these cells arise from aneuploidy present prior to mobilisation, and thus whether sorting cells by DNA content could reduce the problem.

8.2 Using enriched libraries for screens

Although not complete, the level of enrichment for homozygous mutants achieved here is high enough to consider using these libraries to investigate phenotypes that are not strongly positively selectable.

8.2.1 New technologies applicable to genetic screens

The traditional way to screen collections of homozygous mutants would be to pick and assay each individually. In cell culture, this means using multiwell plates (96- or 384-well for high throughput). However, some new technologies incorporate elegant

solutions to this requirement, particularly new sequencing technologies. With high throughput sequencing of transposon insertion sites, for example using the Illumina method described in Chapter 3, the number of cells belonging to each clone can be determined by counting the number of reads from their associated insertion site. This makes it a promising method for investigating phenotypes linked to survival or fitness in prolonged culture. Some phenotypes that could be interesting to investigate are differentiation into different lineages. Using a suitable differentiation protocol that is efficient in bulk culture, such as those for neural or mesodermal lineages, a differentiated library could be isolated. Sequencing all insertion sites in the differentiated population, and comparing to the starting population and an expanded, undifferentiated population, could identify mutants unable to progress to the differentiated stage. Assays for sensitivity to drugs should also be possible—the experiment described in Chapter 3 is a proof of principle of this type of screen.

Another class of phenotype that can be screened by this system is weak positive selection. Mutant clones with fitness advantages under a selective condition will expand and increase their representation in the pool. One potential area of application is screens for infection by viruses and other pathogens, or resistance to toxins. ES cells are not the natural hosts for pathogens, and may not be killed effectively enough to conduct a traditional resistance screen using a non-enriched library. Using an enriched library with a chronic treatment may produce better results than relying on complete acute killing.

I conducted one pilot experiment for this approach to screening, which suggested several improvements (Chapter 3). First, all transposons in the library need to be stably integrated, such that no *de novo* events occur after library generation. For this reason, I generated all subsequent libraries using mRNA to express the transposase, to remove the possibility of stable expression of the transposase in some cells that integrate the expression plasmid that I used previously. Combined with further technical and biological replicates, and the addition of a sample prior to expansion, a high confidence set of transposon sites present at the start of the experiment could be formed to compare the treated population against. This should increase confidence in the identified insertion sites. Using larger libraries, with more than one insertion site per gene would also strengthen the evidence that loss-of-function mutations in that gene are causing the phenotype.

A drawback of this approach is that the mutant cell line cannot be directly obtained. In a traditional genetic screen, this would be a problem, but as single mutant ES cell lines can be obtained easily from the public resource for rapid confirmation, it is less important now. In my view, the emphasis should be on obtaining rapid leads to gene function, which is the role played by the screening systems described here. In any case, mutants need to be reconfirmed on a wild-type (*Blm*-proficient) background using the cell line described here. Transferring the system to the $Blm^{tet/tet}$ line, for example, would be a further improvement.

Another potential improvement for screening assays could be the use of micro-patterned agar to array single cells for screening (Wood *et al.*, 2010). This technique provides a simple method to seed single cells in a grid pattern. This then allows single cells to be screened, and reliably located by a computer-controlled microscope. Screening single cells, rather than a population, has many advantages. For example, in my clone-by-clone isolation experiments (Chapter 5) a mixture of homozygotes and aneuploid cells was obtained. Screening this population, for example for sensitivity to a drug conferred by the mutation, would show an intermediate survival phenotype depending on the relative amounts of homozygotes to aneuploid cells. At the single cell level, the structure of the population can be seen more accurately—in fact in the paper above, cells can also be stained for DNA content, raising the possibility that tetraploid cells could be detected directly during the analysis.

8.2.2 Comparison to other systems for recessive genetic screens

The system described here has several advantages compared to siRNA screens. By picking colonies from homozygote-enriched libraries, a clonally arrayed library of mutants could be constructed that would be usable in similar situations to siRNA screens. The main advantage here is robust mutagenesis. The transposon construct that I used effectively abolished transcription of the wild type allele when homozygous and inserted into an intron (Figure 5.14). When using siRNA the knockdown is often incomplete, and it is also possible that not all cells are transfected, or receive different amounts of siRNA (this can be improved to some extent by shRNA approaches with selection for transformation).

As enrichment of the library is incomplete, there will be some 'junk' clones in such a library, which will manifest as false negatives. This is also a prob-

lem for siRNA screens, however, as the effectiveness of a particular knockdown cannot be guaranteed. Furthermore, transposon mutagenesis deals effectively with false positives, as the insertion can easily be removed by remobilising the transposon (Li *et al.*, 2010). This provides a simple test for causality, which is not available with siRNA.

The recent discovery of a human haploid cell line may represent a powerful alternative system for screens, although it remains to be seen how these cells behave (Carette *et al.*, 2009). Transposon mutagenesis should be readily applicable in this cell line, and the generation of loss-of-function mutants is much more straightforward compared to the *Blm*-deficient ES cell system. The limitation of screens to a single cell type, derived from a tumour, appears to be the only major limitation of this system. Certainly screens for differentiation are not possible in these cells, and they may also have other mutations acquired during tumourigenesis that could make them unsuitable for screening other phenotypes. In these situations, using the ES cell system described here will be necessary.

8.3 Other uses of the copy number selection transposon

The ability to select for copy number increase using the construct that I developed could find wider applications in the field of chromosome instability and copy number variation. Such effects could be easily investigated at different loci in different cell lines through use of the transposon to make stable, single copy integrations. Some of the ES cell lines generated as part of this work could be used as reporters for induction of copy number instability by drugs or mutagens, or new ones could easily be generated in other mutant backgrounds. As described in Chapter 7, PB also induces double strand breaks that are repaired by the host machinery, so my construct can also be used to investigate repair of locus specific DNA damage in a similar way.

8.4 Conclusions

The experimental systems and protocols that I describe in this thesis further extend the genetic toolkit available for analysis of gene function in mice. The main technology, homozygote enrichment by copy number selection, will be useful for conducting recessive genetic screens, a powerful technique from other model organisms that has still not been completely translated to mammalian systems. Technical improvements that were necessary to solve problems associated with copy number instability in ES cells during this process could prove to be more generally useful for the study of genome instability and DNA repair. Application of the technologies that I have developed will assist in the ongoing task of functionally annotating the mammalian genome.

Bibliography

Adams D, Quail M, Cox T *et al.*, 2005. A genome-wide, end-sequenced 129Sv BAC library resource for targeting vector construction. *Genomics*, **86**(6):753. doi:10.1016/j.ygeno.2005.08.003.

Adams DJ, Biggs PJ, Cox T *et al.*, 2004. Mutagenic insertion and chromosome engineering resource (MICER). *Nat Genet*, **36**(8):867–71. doi:10.1038/ng1388.

Adams DJ and van der Weyden L, 2008. Contemporary approaches for modifying the mouse genome. *Physiol Genomics*, **34**(3):225–38. doi:10.1152/physiolgenomics.90242.2008.

Agrawal A, Eastman QM and Schatz DG, 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, **394**(6695):744–51. doi:10.1038/29457.

Ahnesorg P, Smith P and Jackson SP, 2006. XLF interacts with the XRCC4-DNA ligase IV complex to promote DNA nonhomologous end-joining. *Cell*, **124**(2):301–13. doi:10.1016/j.cell.2005.12.031.

Albert TJ, Molla MN, Muzny DM *et al.*, 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Meth*, **4**(11):903–5. doi:10.1038/nmeth1111.

Amé JC, Spenlehauer C and de Murcia G, 2004. The PARP superfamily. *BioEssays*, **26**(8):882–93. doi:10.1002/bies.20085.

Andersen SL, Bergstralh DT, Kohl KP, LaRocque JR, Moore CB and Sekelsky J, 2009. Drosophila MUS312 and the vertebrate ortholog BTBD12 interact with DNA structure-specific endonucleases in DNA repair and recombination. *Molecular Cell*, **35**(1):128–35. doi:10.1016/j.molcel.2009.06.019.

Bakkenist CJ and Kastan MB, 2003. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature*, **421**(6922):499–506. doi:10.1038/nature01368.

Balciunas D, Wangensteen KJ, Wilber A *et al.*, 2006. Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS Genet*, **2**(11):e169. doi:10.1371/journal.pgen.0020169.

Barash Y, Calarco JA, Gao W *et al.*, 2010. Deciphering the splicing code. *Nature*, **465**(7294):53–9. doi:10.1038/nature09000.

Barrett MT, Scheffer A, Ben-Dor A *et al.*, 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *PNAS*, **101**(51):17,765–70. doi:10.1073/pnas.0407979101.

Beddington RS and Robertson EJ, 1989. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. *Development*, **105**(4):733–7.

Bennardo N, Gunn A, Cheng A, Hasty P and Stark JM, 2009. Limiting the persistence of a chromosome break diminishes its mutagenic potential. *PLoS Genet*, **5**(10):e1000,683. doi:10.1371/journal.pgen.1000683.

Bennett CB, Lewis AL, Baldwin KK and Resnick MA, 1993. Lethality induced by a single site-specific double-strand break in a dispensable yeast plasmid. *PNAS*, **90**(12):5613–7.

Beumer KJ, Pimpinelli S and Golic KG, 1998. Induced chromosomal exchange directs the segregation of recombinant chromatids in mitosis of Drosophila. *Genetics*, **150**(1):173–88.

Bibikova M, Beumer K, Trautman JK and Carroll D, 2003. Enhancing gene targeting with designed zinc finger nucleases. *Science*, **300**(5620):764. doi:10.1126/science.1079512.

Blackburn EH, 1991. Structure and function of telomeres. *Nature*, **350**(6319):569–73. doi:10.1038/350569a0.

Blais V, Gao H, Elwell CA *et al.*, 2004. RNA interference inhibition of Mus81 reduces mitotic recombination in human cells. *Molecular Biology of the Cell*, **15**(2):552–62. doi:10.1091/mbc.E03-08-0580.

Bloom D, 1966. The syndrome of congenital telangiectatic erythema and stunted growth. *J Pediatr*, **68**(1):103–13.

Blunt T, Finnie NJ, Taccioli GE *et al.*, 1995. Defective DNA-dependent protein kinase activity is linked to V(D)J recombination and DNA repair defects associated with the murine scid mutation. *Cell*, **80**(5):813–23.

Blunt T, Gell D, Fox M *et al.*, 1996. Identification of a nonsense mutation in the carboxyl-terminal region of DNA-dependent protein kinase catalytic subunit in the scid mouse. *PNAS*, **93**(19):10,285–90.

Borovinskaya MA, Pai RD, Zhang W *et al.*, 2007. Structural basis for aminoglycoside inhibition of bacterial ribosome recycling. *Nature Structural & Molecular Biology*, **14**(8):727–32. doi:10.1038/nsmb1271.

Bouwman P, Aly A, Escandell JM *et al.*, 2010. 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. *Nature Structural & Molecular Biology*. doi:10.1038/nsmb.1831.

Bradley A, Evans M, Kaufman MH and Robertson E, 1984. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*, **309**(5965):255–6.

Brass AL, Dykxhoorn DM, Benita Y *et al.*, 2008. Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, **319**(5865):921–6. doi:10.1126/science.1152725.

Brunet E, Simsek D, Tomishima M *et al.*, 2009. Chromosomal translocations induced at specified loci in human stem cells. *PNAS*, **106**(26):10,620–5. doi:10.1073/pnas.0902076106.

Bryans M, Valenzano MC and Stamato TD, 1999. Absence of DNA ligase IV protein in XR-1 cells: evidence for stabilization by XRCC4. *Mutat Res*, **433**(1):53–8.

Buck D, Malivert L, de Chasseval R *et al.*, 2006. Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immunodeficiency with microcephaly. *Cell*, **124**(2):287–99. doi:10.1016/j.cell.2005.12.030.

Bugreev DV, Yu X, Egelman EH and Mazin AV, 2007. Novel pro- and anti-recombination activities of the Bloom's syndrome helicase. *Genes Dev*, **21**(23):3085–94. doi:10.1101/gad.1609007.

Bultman SJ, Michaud EJ and Woychik RP, 1992. Molecular characterization of the mouse agouti locus. *Cell*, **71**(7):1195–204.

Burdon T, Smith A and Savatier P, 2002. Signalling, cell cycle and pluripotency in embryonic stem cells. *Trends in Cell Biology*, **12**(9):432–8.

Busch DB, Cleaver JE and Glaser DA, 1980. Large-scale isolation of UV-sensitive clones of CHO cells. *Somatic Cell Genet*, **6**(3):407–18.

Bushman FD, Malani N, Fernandes J *et al.*, 2009. Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathogens*, **5**(5):e1000,437. doi:10.1371/journal.ppat.1000437.

Cadiñanos J and Bradley A, 2007. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Research*, **35**(12):e87. doi:10.1093/nar/gkm446.

Cai WW, Mao JH, Chow CW, Damani S, Balmain A and Bradley A, 2002. Genome-wide detection of chromosomal imbalances in tumors using BAC microarrays. *Nat Biotechnol*, **20**(4):393–6. doi:10.1038/nbt0402-393.

Carette JE, Guimaraes CP, Varadarajan M *et al.*, 2009. Haploid genetic screens in human cells identify host factors used by pathogens. *Science*, **326**(5957):1231–5. doi:10.1126/science.1178955.

Carter MS, Li S and Wilkinson MF, 1996. A splicing-dependent regulatory mechanism that detects translation signals. *EMBO J*, **15**(21):5965–75.

Cervantes RB, Stringer JR, Shao C, Tischfield JA and Stambrook PJ, 2002. Embryonic stem cells and somatic cells differ in mutation frequency and type. *PNAS*, **99**(6):3586–90. doi:10.1073/pnas.062527199.

Chaganti RS, Schonberg S and German J, 1974. A manyfold increase in sister chromatid exchanges in Bloom's syndrome lymphocytes. *PNAS*, **71**(11):4508–12.

Chambers I, Colby D, Robertson M *et al.*, 2003. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, **113**(5):643–55.

Chan KL, North P and Hickson I, 2007. BLM is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges. *EMBO Journal*, **26**(14):3397–3409.

Chan KL, Palmai-Pallag T, Ying S and Hickson ID, 2009. Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. *Nature Cell Biology*, **11**(6):753–60. doi:10.1038/ncb1882.

Cheng J, Kapranov P, Drenkow J *et al.*, 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**(5725):1149–54. doi:10.1126/science.1108625.

Chester N, Kuo F, Kozak C, O'Hara CD and Leder P, 1998. Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. *Genes Dev*, **12**(21):3382–93.

Churchill GA, Airey DC, Allayee H *et al.*, 2004. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet*, **36**(11):1133–7. doi:10.1038/ng1104-1133.

Collier LS, Carlson CM, Ravimohan S, Dupuy AJ and Largaespada DA, 2005. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature*, **436**(7048):272–6. doi:10.1038/nature03681.

Constantinou A, Chen XB, McGowan CH and West SC, 2002. Holliday junction resolution in human cells: two junction endonucleases with distinct substrate specificities. *The EMBO Journal*, **21**(20):5577–85.

Coufal NG, Garcia-Perez JL, Peng GE *et al.*, 2009. L1 retrotransposition in human neural progenitor cells. *Nature*, **460**(7259):1127–1131. doi:10.1038/nature08248.

Counter CM, Avilion AA, LeFeuvre CE *et al.*, 1992. Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *The EMBO Journal*, **11**(5):1921–9.

Cuénot L, 1902. La loi de Mendel et l'hérédité de la pigmentation chez les souris. *Arch Zoo Exp Gen*, **4**:33–38.

Cuénot L, 1903. L'Hérédité de la pigmentation chez les souris, 2me note. *Arch Zoo Exp Gen*, **4**:33–38.

Dalgliesh GL, Furge K, Greenman C *et al.*, 2010. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, **463**(7279):360–3. doi:10.1038/nature08672.

Dannenberg JH, van Rossum A, Schuijff L and te Riele H, 2000. Ablation of the retinoblastoma gene family deregulates G(1) control causing immortalization and increased cell turnover under growth-restricting conditions. *Genes Dev*, **14**(23):3051–64.

de Wind N, Dekker M, Berns A, Radman M and Riele HT, 1995. Inactivation of the mouse Msh2 gene results in mismatch repair deficiency, methylation tolerance, hyperrecombination, and predisposition to cancer. *Cell*, **82**(2):321–30.

Deans AJ and West SC, 2009. FANCM connects the genome instability disorders Bloom's Syndrome and Fanconi Anemia. *Molecular Cell*, **36**(6):943–53. doi:10.1016/j.molcel.2009.12.006.

Devon RS, Porteous DJ and Brookes AJ, 1995. Splinkerettes–improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Research*, **23**(9):1644–5.

DiBiase SJ, Zeng ZC, Chen R, Hyslop T, Curran WJ and Iliakis G, 2000. DNA-dependent protein kinase stimulates an independently active, nonhomologous, end-joining apparatus. *Cancer Res*, **60**(5):1245–53.

Dietrich W, Katz H, Lincoln SE *et al.*, 1992. A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics*, **131**(2):423–47.

Ding S, Wu X, Li G, Han M, Zhuang Y and Xu T, 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell*, **122**(3):473–83. doi:10.1016/j.cell.2005.07.013.

Doetschman T, Gregg RG, Maeda N *et al.*, 1987. Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. *Nature*, **330**(6148):576–8. doi:10.1038/330576a0.

Down TA and Hubbard TJP, 2005. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Research*, **33**(5):1445–53. doi:10.1093/nar/gki282.

Dudley DD, Chaudhuri J, Bassing CH and Alt FW, 2005. Mechanism and control of V(D)J recombination versus class switch recombination: similarities and differences. *Adv Immunol*, **86**:43–112. doi:10.1016/S0065-2776(04)86002-4.

Duerr RH, Taylor KD, Brant SR *et al.*, 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**(5804):1461–3. doi:10.1126/science.1135245.

Dupuy AJ, Akagi K, Largaespada DA, Copeland NG and Jenkins NA, 2005. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature*, **436**(7048):221–6. doi:10.1038/nature03691.

Dupuy AJ, Fritz S and Largaespada DA, 2001. Transposition and gene disruption in the male germline of the mouse. *genesis*, **30**(2):82–8.

Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K and Tuschl T, 2001. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**(6836):494–8. doi:10.1038/35078107.

Ellis NA, Groden J, Ye TZ *et al.*, 1995a. The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell*, **83**(4):655–66.

Ellis NA, Lennon DJ, Proytcheva M, Alhadeff B, Henderson EE and German J, 1995b. Somatic intragenic recombination within the mutated locus BLM can correct the high sister-chromatid exchange phenotype of Bloom syndrome cells. *Am J Hum Genet*, **57**(5):1019–27.

Essers J, Hendriks RW, Swagemakers SM *et al.*, 1997. Disruption of mouse RAD54 reduces ionizing radiation resistance and homologous recombination. *Cell*, **89**(2):195–204.

Evans MJ and Kaufman MH, 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, **292**(5819):154–6.

Farmer H, McCabe N, Lord C *et al.*, 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**(7035):917–921. doi:10.1038/nature03445.

Fekairi S, Scaglione S, Chahwan C *et al.*, 2009. Human SLX4 Is a Holliday Junction Resolvase Subunit that Binds Multiple DNA Repair/Recombination Endonucleases. *Cell*, **138**(1):78–89. doi:10.1016/j.cell.2009.06.029.

Filippo JS, Sung P and Klein H, 2008. Mechanism of eukaryotic homologous recombination. *Annual review of biochemistry*, **77**:229–57. doi:10.1146/annurev.biochem.77.061306.125255.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE and Mello CC, 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**(6669):806–11. doi:10.1038/35888.

Fischer SE, Wienholds E and Plasterk RH, 2001. Regulated transposition of a fish transposon in the mouse germ line. *PNAS*, **98**(12):6759–64. doi:10.1073/pnas.121569298.

Flicek P, Aken BL, Ballester B *et al.*, 2010. Ensembl's 10th year. *Nucleic Acids Research*, **38**(Database issue):D557–62. doi:10.1093/nar/gkp972.

Forsyth NR, Wright WE and Shay JW, 2002. Telomerase and differentiation in multicellular organisms: turn it off, turn it on, and turn it off again. *Differentiation*, **69**(4-5):188–97. doi:10.1046/j.1432-0436.2002.690412.x.

Frank KM, Sekiguchi JM, Seidl KJ *et al.*, 1998. Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature*, **396**(6707):173–7. doi:10.1038/24172.

Frank KM, Sharpless NE, Gao Y *et al.*, 2000. DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. *Molecular Cell*, **5**(6):993–1002.

Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M and Ahringer J, 2000. Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature*, **408**(6810):325–30. doi:10.1038/35042517.

Fraser MJ, Ciszczon T, Elick T and Bauser C, 1996. Precise excision of TTAA-specific lepidopteran transposons piggyBac (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol Biol*, **5**(2):141–51.

Gangloff S, McDonald JP, Bendixen C, Arthur L and Rothstein R, 1994. The yeast type I topoisomerase Top3 interacts with Sgs1, a DNA helicase homolog: a potential eukaryotic reverse gyrase. *Molecular and Cellular Biology*, **14**(12):8391–8.

Gao Y, Chaudhuri J, Zhu C, Davidson L, Weaver DT and Alt FW, 1998. A targeted DNA-PKcs-null mutation reveals DNA-PK-independent functions for KU in V(D)J recombination. *Immunity*, **9**(3):367–76.

Gao Y, Ferguson DO, Xie W *et al.*, 2000. Interplay of p53 and DNA-repair protein XRCC4 in tumorigenesis, genomic stability and development. *Nature*, **404**(6780):897–900. doi:10.1038/35009138.

Gardner PP, Daub J, Tate JG *et al.*, 2009. Rfam: updates to the RNA families database. *Nucleic Acids Research*, **37**(Database issue):D136–40. doi:10.1093/nar/gkn766.

Gardner RL, 1968. Mouse chimeras obtained by the injection of cells into the blastocyst. *Nature*, **220**(5167):596–7.

Geurts AM, Hackett CS, Bell JB *et al.*, 2006. Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Research*, **34**(9):2803–11. doi:10.1093/nar/gkl301.

Goff S, 2008. Knockdown screens to knockout HIV-1. *Cell*, **135**:417–420.

Goodarzi AA, Noon AT, Deckbar D *et al.*, 2008. ATM signaling facilitates repair of DNA double-strand breaks associated with heterochromatin. *Molecular Cell*, **31**(2):167–77. doi:10.1016/j.molcel.2008.05.017.

Goss KH, Risinger MA, Kordich JJ *et al.*, 2002. Enhanced tumor formation in mice heterozygous for Blm mutation. *Science*, **297**(5589):2051–3. doi:10.1126/science.1074340.

Gossler A, Joyner AL, Rossant J and Skarnes WC, 1989. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science*, **244**(4903):463–5.

Gottlieb TM and Jackson SP, 1993. The DNA-dependent protein kinase: requirement for DNA ends and association with Ku antigen. *Cell*, **72**(1):131–42.

Gravel S, Chapman JR, Magill C and Jackson SP, 2008. DNA helicases Sgs1 and BLM promote DNA double-strand break resection. *Genes Dev*, **22**(20):2767–72. doi:10.1101/gad.503108.

Grawunder U, Wilm M, Wu X *et al.*, 1997. Activity of DNA ligase IV stimulated by complex formation with XRCC4 protein in mammalian cells. *Nature*, **388**(6641):492–5. doi:10.1038/41358.

Greenman C, Stephens P, Smith R *et al.*, 2007. Patterns of somatic mutation in human cancer genomes. *Nature*, **446**(7132):153–8. doi:10.1038/nature05610.

Greider CW and Blackburn EH, 1987. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell*, **51**(6):887–98.

Grimm S, 2004. The art and design of genetic screens: mammalian culture cells. *Nat Rev Genet*, **5**(3):179–89. doi:10.1038/nrg1291.

Gu Y, Jin S, Gao Y, Weaver DT and Alt FW, 1997a. Ku70-deficient embryonic stem cells have increased ionizing radiosensitivity, defective DNA end-binding activity, and inability to support V(D)J recombination. *PNAS*, **94**(15):8076–81.

Gu Y, Seidl KJ, Rathbun GA *et al.*, 1997b. Growth retardation and leaky SCID phenotype of Ku70-deficient mice. *Immunity*, **7**(5):653–65.

Guirouilh-Barbat J, Rass E, Plo I, Bertrand P and Lopez BS, 2007. Defects in XRCC4 and KU80 differentially affect the joining of distal nonhomologous ends. *PNAS*, **104**(52):20,902–7. doi:10.1073/pnas.0708541104.

Guo G, 2004. Recessive genetic screen for mismatch repair components in BLM-deficient ES cells. Ph.D. thesis, Cambridge University.
URL http://www.sanger.ac.uk/research/publications/theses.html

Guo G, Wang W and Bradley A, 2004. Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells. *Nature*, **429**(6994):891–5. doi:10.1038/nature02653.

Gustincich S, Sandelin A, Plessy C *et al.*, 2006. The complexity of the mammalian transcriptome. *J Physiol (Lond)*, **575**(Pt 2):321–32. doi:10.1113/jphysiol.2006.115568.

Haldane J, 1935. The Rate of Spontaneous Mutation of a Human Gene. *J Genet*, **31**:317–326.

Hamilton AJ and Baulcombe DC, 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, **286**(5441):950–2.

Hanahan D and Weinberg RA, 2000. The hallmarks of cancer. *Cell*, **100**(1):57–70.

Hansen G, Markesich D, Burnett M *et al.*, 2008. Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. *Genome Research*, **18**(10):1670–1679. doi:10.1101/gr.078352.108.

Hasty P, Ramírez-Solis R, Krumlauf R and Bradley A, 1991a. Introduction of a subtle mutation into the Hox-2.6 locus in embryonic stem cells. *Nature*, **350**(6315):243–6. doi:10.1038/350243a0.

Hasty P, Rivera-Pérez J and Bradley A, 1991b. The length of homology required for gene targeting in embryonic stem cells. *Molecular and Cellular Biology*, **11**(11):5586–91.

Hasty P, Rivera-Pérez J, Chang C and Bradley A, 1991c. Target frequency and integration pattern for insertion and replacement vectors in embryonic stem cells. *Molecular and Cellular Biology*, **11**(9):4509–17.

Hayakawa T, Yusa K, Kouno M, Takeda J and Horie K, 2006. Bloom's syndrome gene-deficient phenotype in mouse primary cells induced by a modified tetracycline-controlled trans-silencer. *Gene*, **369**:80–9. doi:10.1016/j.gene.2005.10.041.

Hickson ID, 2003. RecQ helicases: caretakers of the genome. *Nat Rev Cancer*, **3**(3):169–78. doi:10.1038/nrc1012.

Holliday R and Ho T, 2002. DNA methylation and epigenetic inheritance. *Methods*, **27**(2):179–83.

Horie K, Kuroiwa A, Ikawa M *et al.*, 2001. Efficient chromosomal transposition of a Tc1/mariner- like transposon Sleeping Beauty in mice. *PNAS*, **98**(16):9191–6. doi:10.1073/pnas.161071798.

Hrabé de Angelis MH, Flaswinkel H, Fuchs H *et al.*, 2000. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet*, **25**(4):444–7. doi:10.1038/78146.

Hu Y, Lu X, Barnes E, Yan M, Lou H and Luo G, 2005. Recql5 and Blm RecQ DNA helicases have nonredundant roles in suppressing crossovers. *Molecular and Cellular Biology*, **25**(9):3431–42. doi:10.1128/MCB.25.9.3431-3442.2005.

Hu Y, Raynard S, Sehorn MG *et al.*, 2007. RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments. *Genes Dev*, **21**(23):3073–84. doi:10.1101/gad.1609107.

Huang X, Guo H, Tammana S *et al.*, 2010. Gene Transfer Efficiency and Genome-Wide Integration Profiling of Sleeping Beauty, Tol2, and PiggyBac Transposons in Human Primary T Cells. *Molecular Therapy*. doi:10.1038/mt.2010.141.

International Mouse Knockout Consortium, Collins FS, Rossant J and Wurst W, 2007. A mouse for all reasons. *Cell*, **128**(1):9–13. doi:10.1016/j.cell.2006.12.018.

Ip SCY, Rass U, Blanco MG, Flynn HR, Skehel JM and West SC, 2008. Identification of Holliday junction resolvases from humans and yeast. *Nature*, **456**(7220):357–61. doi:10.1038/nature07470.

Ivics Z, Hackett PB, Plasterk RH and Izsvák Z, 1997. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**(4):501–10.

Izsvák Z, Stüwe EE, Fiedler D, Katzer A, Jeggo PA and Ivics Z, 2004. Healing the wounds inflicted by sleeping beauty transposition by double-strand break repair in mammalian somatic cells. *Molecular Cell*, **13**(2):279–90.

Jackson AL, Bartz SR, Schelter J *et al.*, 2003. Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotech*, **21**(6):635–7. doi:10.1038/nbt831.

Jackson SP, 2002. Sensing and repairing DNA double-strand breaks. *Carcinogenesis*, **23**(5):687–96.

Jackson SP, 2009. The DNA-damage response: new molecular insights and new approaches to cancer therapy. *Biochem Soc Trans*, **37**(Pt 3):483–94. doi:10.1042/BST0370483.

Jeggo PA and Kemp LM, 1983. X-ray-sensitive mutants of Chinese hamster ovary cell line. Isolation and cross-sensitivity to other DNA-damaging agents. *Mutat Res*, **112**(6):313–27.

Jenkins NA, Copeland NG, Taylor BA and Lee BK, 1981. Dilute (d) coat colour mutation of DBA/2J mice is associated with the site of integration of an ecotropic MuLV genome. *Nature*, **293**(5831):370–4.

Johnson RD and Jasin M, 2000. Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *EMBO J*, **19**(13):3398–407. doi:10.1093/emboj/19.13.3398.

Johnson RS, Sheng M, Greenberg ME, Kolodner RD, Papaioannou VE and Spiegelman BM, 1989. Targeting of nonexpressed genes in embryonic stem cells via homologous recombination. *Science*, **245**(4923):1234–6.

Jones ME, Thomas SM and Rogers A, 1994. Luria-Delbrück fluctuation experiments: design and analysis. *Genetics*, **136**(3):1209–16.

Joyner AL, Skarnes WC and Rossant J, 1989. Production of a mutation in mouse En-2 gene by homologous recombination in embryonic stem cells. *Nature*, **338**(6211):153–6. doi:10.1038/338153a0.

Kamath RS, Fraser AG, Dong Y *et al.*, 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**(6920):231–7. doi:10.1038/nature01278.

Karow JK, Constantinou A, Li JL, West SC and Hickson ID, 2000. The Bloom's syndrome gene product promotes branch migration of holliday junctions. *PNAS*, **97**(12):6504–8. doi:10.1073/pnas.100448097.

Kawakami K and Noda T, 2004. Transposition of the Tol2 element, an Ac-like element from the Japanese medaka fish Oryzias latipes, in mouse embryonic stem cells. *Genetics*, **166**(2):895–9.

Keng VW, Ryan BJ, Wangensteen KJ *et al.*, 2009. Efficient transposition of Tol2 in the mouse germline. *Genetics*, **183**(4):1565–73. doi:10.1534/genetics.109.100768.

Kile BT, Hentges KE, Clark AT *et al.*, 2003. Functional genetic analysis of mouse chromosome 11. *Nature*, **425**(6953):81–6. doi:10.1038/nature01865.

Kile BT and Hilton DJ, 2005. The art and design of genetic screens: mouse. *Nat Rev Genet*, **6**(7):557–67. doi:10.1038/nrg1636.

Kim YG, Cha J and Chandrasegaran S, 1996. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *PNAS*, **93**(3):1156–60.

Koga A, Suzuki M, Inagaki H, Bessho Y and Hori H, 1996. Transposable element in fish. *Nature*, **383**(6595):30. doi:10.1038/383030a0.

Koike H, Horie K, Fukuyama H, Kondoh G, Nagata S and Takeda J, 2002. Efficient biallelic mutagenesis with Cre/loxP-mediated inter-chromosomal recombination. *EMBO Reports*, **3**(5):433–7. doi:10.1093/embo-reports/kvf097.

Kokubu C, Horie K, Abe K *et al.*, 2009. A transposon-based chromosomal engineering method to survey a large cis-regulatory landscape in mice. *Nat Genet*, **41**(8):946–952. doi:10.1038/ng.397.

Kolas NK, Chapman JR, Nakada S *et al.*, 2007. Orchestration of the DNA-damage response by the RNF8 ubiquitin ligase. *Science*, **318**(5856):1637–40. doi:10.1126/science.1150034.

Koller BH, Hagemann LJ, Doetschman T *et al.*, 1989. Germline transmission of a planned alteration made in a hypoxanthine phosphoribosyltransferase gene by homologous recombination in embryonic stem cells. *PNAS*, **86**(22):8927–31.

Koller BH and Smithies O, 1989. Inactivating the beta 2-microglobulin locus in mouse embryonic stem cells by homologous recombination. *PNAS*, **86**(22):8932–5.

Kong J, Wang F, Brenton JD and Adams DJ, 2010. Slingshot: a PiggyBac based transposon system for tamoxifen-inducible 'self-inactivating' insertional mutagenesis. *Nucleic Acids Research*. doi:10.1093/nar/gkq658.

Kong X, Mohanty SK, Stephens J *et al.*, 2009. Comparative analysis of different laser systems to study cellular responses to DNA damage in mammalian cells. *Nucleic Acids Research*, **37**(9):e68. doi:10.1093/nar/gkp221.

König R, Zhou Y, Elleder D *et al.*, 2008. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, **135**(1):49–60. doi:10.1016/j.cell.2008.07.032.

Korbel JO, Urban AE, Affourtit JP *et al.*, 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**(5849):420–6. doi:10.1126/science.1149504.

Krejci L, Komen SV, Li Y *et al.*, 2003. DNA helicase Srs2 disrupts the Rad51 presynaptic filament. *Nature*, **423**(6937):305–9. doi:10.1038/nature01577.

Kuehn MR, Bradley A, Robertson EJ and Evans MJ, 1987. A potential animal model for Lesch-Nyhan syndrome through introduction of HPRT mutations into mice. *Nature*, **326**(6110):295–8. doi:10.1038/326295a0.

Lander ES, Linton LM, Birren B *et al.*, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921. doi:10.1038/35057062.

Langridge GC, Phan MD, Turner DJ *et al.*, 2009. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res*, **19**(12):2308–16. doi:10.1101/gr.097097.109.

Lee EC, Yu D, de Velasco JM *et al.*, 2001. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics*, **73**(1):56–65. doi:10.1006/geno.2000.6451.

Lee RC, Feinbaum RL and Ambros V, 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**(5):843–54.

Lee Y, Barnes DE, Lindahl T and McKinnon PJ, 2000. Defective neurogenesis resulting from DNA ligase IV deficiency requires Atm. *Genes Dev*, **14**(20):2576–80.

Leeds P, Peltz SW, Jacobson A and Culbertson MR, 1991. The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. *Genes Dev*, **5**(12A):2303–14.

Lefebvre L, Dionne N, Karaskova J, Squire JA and Nagy A, 2001. Selection for transgene homozygosity in embryonic stem cells results in extensive loss of heterozygosity. *Nat Genet*, **27**(3):257–8. doi:10.1038/85808.

Ley TJ, Mardis ER, Ding L *et al.*, 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**(7218):66–72. doi:10.1038/nature07485.

Li G, Alt FW, Cheng HL *et al.*, 2008. Lymphocyte-specific compensation for XLF/cernunnos end-joining functions in V(D)J recombination. *Molecular Cell*, **31**(5):631–40. doi:10.1016/j.molcel.2008.07.017.

Li MA, 2010. A recessive genetic screen to discover components in the miRNA pathways using the piggyBac-mediated insertional mutagenesis in Blm-deficient mouse embryonic stem cells. Ph.D. thesis, Cambridge University. URL http://www.sanger.ac.uk/research/publications/theses.html

Li MA, Pettitt SJ, Yusa K and Bradley A, 2010. Genome-Wide Forward Genetic Screens in Mouse ES Cells. *Methods Enzymol*, **477C**:217–242. doi:10.1016/S0076-6879(10)77012-9.

Liang F, Han M and Jasin M, 1998. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *PNAS*, **95**(9):5172–7.

Liang Q, Conte N, Skarnes W and Bradley A, 2008. Extensive genomic copy number variation in embryonic stem cells. *PNAS*, **105**(45):17,453. doi:10.1073/pnas.0805638105.

Liang Q, Kong J, Stalker J and Bradley A, 2009. Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *genesis*, **47**(6):404–8. doi:10.1002/dvg.20508.

Lin X, Morgan-Lappe S, Huang X *et al.*, 2007. 'Seed' analysis of off-target siRNAs reveals an essential role of Mcl-1 in resistance to the small-molecule Bcl-2/Bcl-XL inhibitor ABT-737. *Oncogene*, **26**(27):3972–9. doi:10.1038/sj.onc.1210166.

Liu P, Jenkins NA and Copeland NG, 2002. Efficient Cre-loxP-induced mitotic recombination in mouse embryonic stem cells. *Nat Genet*, **30**(1):66–72. doi:10.1038/ng788.

Liu P, Jenkins NA and Copeland NG, 2003. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Research*, **13**(3):476–84. doi:10.1101/gr.749203.

Liu X, Wu H, Loring J *et al.*, 1997. Trisomy eight in ES cells is a common potential problem in gene targeting and interferes with germ line transmission. *Dev Dyn*, **209**(1):85–91. doi:10.1002/(SICI)1097-0177(199705)209:1⟨85::AID-AJA8⟩3.0.CO;2-T.

Luo G, Ivics Z, Izsvák Z and Bradley A, 1998. Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *PNAS*, **95**(18):10,769–73.

Luo G, Santoro IM, McDaniel LD *et al.*, 2000. Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nat Genet*, **26**(4):424–9. doi:10.1038/82548.

Lupski JR, Reid JG, Gonzaga-Jauregui C *et al.*, 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, **362**(13):1181–91. doi:10.1056/NEJMoa0908094.

Luria SE and Delbrück M, 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, **28**(6):491–511.

Maniatis T, Fritsch EF and Sambrook J, 1982. *Molecular Cloning, A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Mankouri HW and Hickson ID, 2007. The RecQ helicase-topoisomerase III-Rmi1 complex: a DNA structure-specific 'dissolvasome'? *Trends Biochem Sci*, **32**(12):538–46. doi:10.1016/j.tibs.2007.09.009.

Maquat LE, 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol*, **5**(2):89–99. doi:10.1038/nrm1310.

Maragathavally KJ, Kaminski JM and Coates CJ, 2006. Chimeric Mos1 and piggyBac transposases result in site-directed integration. *FASEB J*, **20**(11):1880–2. doi:10.1096/fj.05-5485fje.

Mátés L, Chuah MKL, Belay E *et al.*, 2009. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet*, **41**(6):753–761. doi:10.1038/ng.343.

Mattison J, van der Weyden L, Hubbard T and Adams DJ, 2009. Cancer gene discovery in mouse and man. *Biochim Biophys Acta*, **1796**(2):140–61. doi:10.1016/j.bbcan.2009.03.001.

McDaniel LD, Chester N, Watson M, Borowsky AD, Leder P and Schultz RA, 2003. Chromosome instability and tumor predisposition inversely correlate with BLM protein levels. *DNA Repair*, **2**(12):1387–404.

McMahon A and Bradley A, 1990. The Wnt-1 (int-1) proto-oncogene is required for development of a large region of the mouse brain. *Cell*, **62**(6):1073.

Menzel S, Garner C, Gut I *et al.*, 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet*, **39**(10):1197–9. doi:10.1038/ng2108.

Mikkelsen TS, Ku M, Jaffe DB *et al.*, 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**(7153):553–60. doi:10.1038/nature06008.

Mimori T, Hardin JA and Steitz JA, 1986. Characterization of the DNA-binding protein antigen Ku recognized by autoantibodies from patients with rheumatic disorders. *J Biol Chem*, **261**(5):2274–8.

Mitra R, Fain-Thornton J and Craig NL, 2008. piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J*, **27**(7):1097–1109. doi:10.1038/emboj.2008.41.

Mortensen RM, Conner DA, Chao S, Geisterfer-Lowrance AA and Seidman JG, 1992. Production of homozygous mutant ES cells with a single targeting construct. *Molecular and Cellular Biology*, **12**(5):2391–5.

Moser AR, Pitot HC and Dove WF, 1990. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. *Science*, **247**(4940):322–4.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915):520–562. doi:10.1038/nature01262.

Moynahan ME and Jasin M, 1997. Loss of heterozygosity induced by a chromosomal double-strand break. *PNAS*, **94**(17):8988–93.

Muller HJ, 1927. Artificial transmutation of the gene. *Science*, **66**(1699):84–7. doi:10.1126/science.66.1699.84.

Muñoz IM, Hain K, Déclais AC *et al.*, 2009. Coordination of structure-specific nucleases by human SLX4/BTBD12 is required for DNA repair. *Molecular Cell*, **35**(1):116–27. doi:10.1016/j.molcel.2009.06.020.

Murphy ED, 1966. Characteristic tumors. In EL Green, editor, *Biology of the Laboratory Mouse*, chapter 27. Dover Publications, Inc. (New York).
URL http://www.informatics.jax.org/greenbook

Nagy A, Gócza E, Diaz EM *et al.*, 1990. Embryonic stem cells alone are able to support fetal development in the mouse. *Development*, **110**(3):815–21.

Nakayama H, Nakayama K, Nakayama R, Irino N, Nakayama Y and Hanawalt PC, 1984. Isolation and genetic characterization of a thymineless death-resistant mutant of Escherichia coli K12: identification of a new mutation (recQ1) that blocks the RecF recombination pathway. *Mol Gen Genet*, **195**(3):474–80.

Nathans D, 1964. Puromycin inhibition of protein synthesis: Incorporation of puromycin into peptide chains. *PNAS*, **51**:585–92.

Neilan EG and Barsh GS, 1999. Gene trap insertional mutagenesis in mice: new vectors and germ line mutations in two novel genes. *Transgenic Res*, **8**(6):451–8.

Ng SB, Buckingham KJ, Lee C *et al.*, 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, **42**(1):30–5. doi:10.1038/ng.499.

Nimonkar AV, Ozsoy AZ, Genschel J, Modrich P and Kowalczykowski SC, 2008. Human exonuclease 1 and BLM helicase interact to resect DNA and initiate DNA repair. *PNAS*, **105**(44):16,906–11. doi:10.1073/pnas.0809380105.

Ning Z, Cox AJ and Mullikin JC, 2001. SSAHA: a fast search method for large DNA databases. *Genome Res*, **11**(10):1725–9. doi:10.1101/gr.194201.

Nishikawa SI, Nishikawa S, Hirashima M, Matsuyoshi N and Kodama H, 1998. Progressive lineage analysis by cell sorting and culture identifies FLK1+VE-cadherin+ cells at a diverging point of endothelial and hemopoietic lineages. *Development*, **125**(9):1747–57.

Novina CD and Sharp PA, 2004. The RNAi revolution. *Nature*, **430**(6996):161–4. doi:10.1038/430161a.

Nussenzweig A, Chen C, da Costa Soares V *et al.*, 1996. Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. *Nature*, **382**(6591):551–5. doi:10.1038/382551a0.

Nüsslein-Volhard C and Wieschaus E, 1980. Mutations affecting segment number and polarity in Drosophila. *Nature*, **287**(5785):795–801.

Ooi SL, Shoemaker DD and Boeke JD, 2001. A DNA microarray-based genetic screen for nonhomologous end-joining mutants in Saccharomyces cerevisiae. *Science*, **294**(5551):2552–6. doi:10.1126/science.1065672.

Orr-Weaver TL, Szostak JW and Rothstein RJ, 1981. Yeast transformation: a model system for the study of recombination. *PNAS*, **78**(10):6354–8.

Peric-Hupkes D, Meuleman W, Pagie L *et al.*, 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular Cell*, **38**(4):603–13. doi:10.1016/j.molcel.2010.03.016.

Pettitt SJ, Liang Q, Rairdan XY *et al.*, 2009. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat Methods*, **6**(7):493–5. doi:10.1038/nmeth.1342.

Pinkel D, Segraves R, Sudar D *et al.*, 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20**(2):207–11. doi:10.1038/2524.

Pinkel D, Thompson LH, Gray JW and Vanderlaan M, 1985. Measurement of sister chromatid exchanges at very low bromodeoxyuridine substitution levels using a monoclonal antibody in Chinese hamster ovary cells. *Cancer Research*, **45**(11 Pt 2):5795–8.

Pleasance ED, Cheetham RK, Stephens PJ *et al.*, 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**(7278):191–6. doi:10.1038/nature08658.

Pleasance ED, Stephens PJ, O'Meara S *et al.*, 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**(7278):184–90. doi:10.1038/nature08629.

Pollard SM, Benchoua A and Lowell S, 2006. Neural stem cells, neurons, and glia. *Methods Enzymol*, **418**:151–69. doi:10.1016/S0076-6879(06)18010-6.

Porteus MH and Baltimore D, 2003. Chimeric nucleases stimulate gene targeting in human cells. *Science*, **300**(5620):763. doi:10.1126/science.1078395.

Prosser HM, Rzadzinska AK, Steel KP and Bradley A, 2008. Mosaic complementation demonstrates a regulatory role for myosin VIIa in actin dynamics of stereocilia. *Molecular and Cellular Biology*, **28**(5):1702–12. doi:10.1128/MCB.01282-07.

Puebla-Osorio N, Lacey DB, Alt FW and Zhu C, 2006. Early embryonic lethality due to targeted inactivation of DNA ligase III. *Molecular and Cellular Biology*, **26**(10):3935–41. doi:10.1128/MCB.26.10.3935-3941.2006.

Quail MA, Kozarewa I, Smith F *et al.*, 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Meth*, **5**(12):1005–10. doi:10.1038/nmeth.1270.

Ramírez-Solis R, Davis AC and Bradley A, 1993. Gene targeting in embryonic stem cells. *Meth Enzymol*, **225**:855–78.

Ramírez-Solis R, Liu P and Bradley A, 1995. Chromosome engineering in mice. *Nature*, **378**(6558):720–4. doi:10.1038/378720a0.

Randall-Hazelbauer L and Schwartz M, 1973. Isolation of the bacteriophage lambda receptor from Escherichia coli. *J Bacteriol*, **116**(3):1436–46.

Rastan S, Hough T, Kierman A *et al.*, 2004. Towards a mutant map of the mouse–new models of neurological, behavioural, deafness, bone, renal and blood disorders. *Genetica*, **122**(1):47–9.

Rastan S and Robertson EJ, 1985. X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation. *J Embryol Exp Morphol*, **90**:379–88.

Raymond CS and Soriano P, 2007. High-efficiency FLP and PhiC31 site-specific recombination in mammalian cells. *PLoS ONE*, **2**(1):e162. doi:10.1371/journal.pone.0000162.

Robertson E, Bradley A, Kuehn M and Evans M, 1986. Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. *Nature*, **323**(6087):445–8. doi:10.1038/323445a0.

Rooney S, Alt FW, Lombard D *et al.*, 2003. Defective DNA repair and increased genomic instability in Artemis-deficient murine cells. *J Exp Med*, **197**(5):553–65.

Roth DB, Porter TN and Wilson JH, 1985. Mechanisms of nonhomologous recombination in mammalian cells. *Molecular and Cellular Biology*, **5**(10):2599–607.

Rothkamm K, Krüger I, Thompson LH and Löbrich M, 2003. Pathways of DNA double-strand break repair during the mammalian cell cycle. *Molecular and Cellular Biology*, **23**(16):5706–15.

Rouet P, Smih F and Jasin M, 1994. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Molecular and Cellular Biology*, **14**(12):8096–106.

Rubin GM, Yandell MD, Wortman JR *et al.*, 2000. Comparative genomics of the eukaryotes. *Science*, **287**(5461):2204–15.

Russell ES and Meier H, 1966. Constitutional diseases. In EL Green, editor, *Biology of the Laboratory Mouse*, chapter 29. Dover Publications, Inc. (New York).
URL http://www.informatics.jax.org/greenbook

Sage J, Mulligan GJ, Attardi LD *et al.*, 2000. Targeted disruption of the three Rb-related genes leads to loss of G(1) control and immortalization. *Genes Dev*, **14**(23):3037–50.

Sakaue-Sawano A, Kurokawa H, Morimura T *et al.*, 2008. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, **132**(3):487–98. doi:10.1016/j.cell.2007.12.033.

Sauer B and Henderson N, 1988. Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. *PNAS*, **85**(14):5166–70.

Savatier P, Huang S, Szekely L, Wiman KG and Samarut J, 1994. Contrasting patterns of retinoblastoma protein expression in mouse embryonic stem cells and embryonic fibroblasts. *Oncogene*, **9**(3):809–18.

Schaft J, Ashery-Padan R, van der Hoeven F, Gruss P and Stewart AF, 2001. Efficient FLP recombination in mouse ES cells and oocytes. *genesis*, **31**(1):6–10.

Schwartzberg PL, Goff SP and Robertson EJ, 1989. Germline transmission of a c-abl mutation produced by targeted gene disruption in ES cells. *Science*, **246**(4931):799–803.

Scully R, Chen J, Plug A *et al.*, 1997. Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell*, **88**(2):265–75.

Seibler J, Schübeler D, Fiering S, Groudine M and Bode J, 1998. DNA cassette exchange in ES cells mediated by Flp recombinase: an efficient strategy for repeated modification of tagged loci by marker-free constructs. *Biochemistry*, **37**(18):6229–34. doi:10.1021/bi980288t.

Sharan SK, Morimatsu M, Albrecht U *et al.*, 1997. Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. *Nature*, **386**(6627):804–10. doi:10.1038/386804a0.

Shawlot W, Deng JM, Fohn LE and Behringer RR, 1998. Restricted beta-galactosidase expression of a hygromycin-lacZ gene targeted to the beta-actin locus and embryonic lethality of beta-actin mutant mice. *Transgenic Res*, **7**(2):95–103.

Sherr CJ and DePinho RA, 2000. Cellular senescence: mitotic clock or culture shock? *Cell*, **102**(4):407–10.

Shigeoka T, Kawaichi M and Ishida Y, 2005. Suppression of nonsense-mediated mRNA decay permits unbiased gene trapping in mouse embryonic stem cells. *Nucleic Acids Res*, **33**(2):e20. doi:10.1093/nar/gni022.

Shiloh Y, 2003. ATM and related protein kinases: safeguarding genome integrity. *Nat Rev Cancer*, **3**(3):155–68. doi:10.1038/nrc1011.

Shimshek DR, Kim J, Hübner MR *et al.*, 2002. Codon-improved Cre recombinase (iCre) expression in the mouse. *genesis*, **32**(1):19–26.

Silver LM, 1995. *Mouse Genetics—Concepts and Applications*. Oxford University Press.

Siminovitch L, 1976. On the nature of hereditable variation in cultured somatic cells. *Cell*, **7**(1):1–11.

Singh NP, McCoy MT, Tice RR and Schneider EL, 1988. A simple technique for quantitation of low levels of DNA damage in individual cells. *Experimental Cell Research*, **175**(1):184–91.

Singh TR, Ali AM, Busygina V *et al.*, 2008. BLAP18/RMI2, a novel OB-fold-containing protein, is an essential component of the Bloom helicase-double Holliday junction dissolvasome. *Genes Dev*, **22**(20):2856–68. doi:10.1101/gad.1725108.

Sjöblom T, Jones S, Wood LD *et al.*, 2006. The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**(5797):268–74. doi:10.1126/science.1133427.

Skarnes WC, Auerbach BA and Joyner AL, 1992. A gene trap approach in mouse embryonic stem cells: the lacZ reported is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. *Genes Dev*, **6**(6):903–18.

Skarnes WC, von Melchner H, Wurst W *et al.*, 2004. A public gene trap resource for mouse functional genomics. *Nat Genet*, **36**(6):543–4. doi:10.1038/ng0604-543.

Smih F, Rouet P and Jasin M, 1995. Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells. *Nucleic Acids Research*, **23**(24):5012–9.

Smithies O, Gregg RG, Boggs SS, Koralewski MA and Kucherlapati RS, 1985. Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature*, **317**(6034):230–4.

Smogorzewska A, Desetty R, Saito TT *et al.*, 2010. A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Molecular Cell*, **39**(1):36–47. doi:10.1016/j.molcel.2010.06.023.

Snell GD and Stimpfling JH, 1966. Genetics of tissue transplantation. In EL Green, editor, *Biology of the Laboratory Mouse*, chapter 24. Dover Publications, Inc. (New York). URL http://www.informatics.jax.org/greenbook

Soriano P, Friedrich G and Lawinger P, 1991. Promoter interactions in retrovirus vectors introduced into fibroblasts and embryonic stem cells. *J Virol*, **65**(5):2314–9.

Steighner RJ and Povirk LF, 1990. Bleomycin-induced DNA lesions at mutational hot spots: implications for the mechanism of double-strand cleavage. *PNAS*, **87**(21):8350–4.

Sternberg N and Hamilton D, 1981. Bacteriophage P1 site-specific recombination. I. Recombination between loxP sites. *Journal of Molecular Biology*, **150**(4):467–86.

Stevens LC and Little CC, 1954. Spontaneous Testicular Teratomas in an Inbred Strain of Mice. *PNAS*, **40**(11):1080–7.

Su LK, Kinzler KW, Vogelstein B *et al.*, 1992. Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. *Science*, **256**(5057):668–70.

Su Q, Prosser HM, Campos LS *et al.*, 2008. A DNA transposon-based approach to validate oncogenic mutations in the mouse. *PNAS*, **105**(50):19,904–9. doi:10.1073/pnas.0807785105.

Sudbery I, Enright AJ, Fraser AG and Dunham I, 2010. Systematic analysis of off-target effects in an RNAi screen reveals microRNAs affecting sensitivity to TRAIL-induced apoptosis. *BMC Genomics*, **11**:175. doi:10.1186/1471-2164-11-175.

Sudbery I, Stalker J, Simpson JT *et al.*, 2009. Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels. *Genome Biol*, **10**(10):R112. doi:10.1186/gb-2009-10-10-r112.

Sun H, Karow JK, Hickson ID and Maizels N, 1998. The Bloom's syndrome helicase unwinds G4 DNA. *J Biol Chem*, **273**(42):27,587–92.

Svendsen JM, Smogorzewska A, Sowa ME *et al.*, 2009. Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. *Cell*, **138**(1):63–77. doi:10.1016/j.cell.2009.06.030.

Taccioli GE, Amatucci AG, Beamish HJ *et al.*, 1998. Targeted disruption of the catalytic subunit of the DNA-PK gene in mice confers severe combined immunodeficiency and radiosensitivity. *Immunity*, **9**(3):355–66.

Takahashi K and Yamanaka S, 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**(4):663–76. doi:10.1016/j.cell.2006.07.024.

te Riele H, Maandag ER and Berns A, 1992. Highly efficient gene targeting in embryonic stem cells through homologous recombination with isogenic DNA constructs. *PNAS*, **89**(11):5128–32.

Thomas KR and Capecchi MR, 1987. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*, **51**(3):503–12.

Thomas KR, Folger KR and Capecchi MR, 1986. High frequency targeting of genes to specific sites in the mammalian genome. *Cell*, **44**(3):419–28.

Thompson LH, 1998. Chinese hamster cells meet DNA repair: an entirely acceptable affair. *BioEssays*, **20**(7):589–97. doi:10.1002/(SICI)1521-1878(199807)20:7⟨589::AID-BIES11⟩3.0.CO;2-W.

Thompson LH, Rubin JS, Cleaver JE, Whitmore GF and Brookman K, 1980. A screening method for isolating DNA repair-deficient mutants of CHO cells. *Somatic Cell Genet*, **6**(3):391–405.

Thompson S, Clarke AR, Pow AM, Hooper ML and Melton DW, 1989. Germ line transmission and expression of a corrected HPRT gene produced by gene targeting in embryonic stem cells. *Cell*, **56**(2):313–21.

Tjio J and Puck T, 1958. Genetics of somatic mammalian cells. II. Chromosomal constitution of cells in tissue culture. *The Journal of Experimental Medicine*, **108**(2):259–68.

Trombly MI, Su H and Wang X, 2009. A genetic screen for components of the mammalian RNA interference pathway in Bloom-deficient mouse embryonic stem cells. *Nucleic Acids Research*, **37**(4):e34. doi:10.1093/nar/gkp019.

Turner DJ, Keane TM, Sudbery I and Adams DJ, 2009. Next-generation sequencing of vertebrate experimental organisms. *Mamm Genome*, **20**(6):327–38. doi:10.1007/s00335-009-9187-4.

Umezu K, Nakayama K and Nakayama H, 1990. Escherichia coli RecQ protein is a DNA helicase. *PNAS*, **87**(14):5363–7.

Vassiliou G, Rad R and Bradley A, 2010. The Use of DNA Transposons for Cancer Gene Discovery in Mice. *Methods Enzymol*, **477C**:91–106. doi:10.1016/S0076-6879(10)77006-3.

Venter JC, Adams MD, Myers EW *et al.*, 2001. The sequence of the human genome. *Science*, **291**(5507):1304–51. doi:10.1126/science.1058040.

Vitaterna MH, King DP, Chang AM *et al.*, 1994. Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. *Science*, **264**(5159):719–25.

von Melchner H and Ruley HE, 1989. Identification of cellular promoters by using a retrovirus promoter trap. *J Virol*, **63**(8):3227–33.

Vooijs M, Jonkers J and Berns A, 2001. A highly efficient ligand-regulated Cre recombinase mouse line shows that LoxP recombination is position dependent. *EMBO Reports*, **2**(4):292–7. doi:10.1093/embo-reports/kve064.

Voss AK, Thomas T and Gruss P, 1998. Efficiency assessment of the gene trap approach. *Dev Dyn*, **212**(2):171–80. doi:10.1002/(SICI)1097-0177(199806)212:2⟨171::AID-AJA3⟩3.0.CO;2-E.

Wang H, Rosidi B, Perrault R *et al.*, 2005. DNA ligase III as a candidate component of backup pathways of nonhomologous end joining. *Cancer Res*, **65**(10):4020–30. doi:10.1158/0008-5472.CAN-04-3055.

Wang M, Wu W, Wu W *et al.*, 2006. PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. *Nucleic Acids Research*, **34**(21):6170–82. doi:10.1093/nar/gkl840.

Wang W and Bradley A, 2007. A recessive genetic screen for host factors required for retroviral infection in a library of insertionally mutated Blm-deficient embryonic stem cells. *Genome Biol*, **8**(4):R48. doi:10.1186/gb-2007-8-4-r48.

Wang W, Bradley A and Huang Y, 2009. A piggyBac transposon-based genome-wide library of insertionally mutated Blm-deficient murine ES cells. *Genome Res*, **19**(4):667–73. doi:10.1101/gr.085621.108.

Wang W, Lin C, Lu D *et al.*, 2008. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *PNAS*, **105**(27):9290–5. doi:10.1073/pnas.0801017105.

Warming S, Costantino N, Court DL, Jenkins NA and Copeland NG, 2005. Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Research*, **33**(4):e36. doi:10.1093/nar/gni035.

Weinstock DM and Jasin M, 2006. Alternative pathways for the repair of RAG-induced DNA breaks. *Molecular and Cellular Biology*, **26**(1):131–9. doi:10.1128/MCB.26.1.131-139.2006.

Wellcome Trust Case-Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145):661–78. doi:10.1038/nature05911.

Wellcome Trust Case-Control Consortium, Craddock N, Hurles ME *et al.*, 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**(7289):713–20. doi:10.1038/nature08979.

Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T and Harrow JL, 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Research*, **36**(Database issue):D753–60. doi:10.1093/nar/gkm987.

Wilson D and Thompson L, 2007. Molecular mechanisms of sister-chromatid exchange. *Mutation Research*, **616**(1-2):11–23.

Winston WM, Molodowitch C and Hunter CP, 2002. Systemic RNAi in C. elegans requires the putative transmembrane protein SID-1. *Science*, **295**(5564):2456–9. doi:10.1126/science.1068836.

Woltjen K, Michael IP, Mohseni P *et al.*, 2009. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*, **458**(7239):766–70. doi:10.1038/nature07863.

Wood DK, Weingeist DM, Bhatia SN and Engelward BP, 2010. Single cell trapping and DNA damage analysis using microwell arrays. *PNAS*, **107**(22):10,008–13. doi:10.1073/pnas.1004056107.

Wood LD, Parsons DW, Jones S *et al.*, 2007. The genomic landscapes of human breast and colorectal cancers. *Science*, **318**(5853):1108–13. doi:10.1126/science.1145720.

Wu L, Davies SL, Levitt NC and Hickson ID, 2001. Potential role for the BLM helicase in recombinational repair via a conserved interaction with RAD51. *J Biol Chem*, **276**(22):19,375–81. doi:10.1074/jbc.M009471200.

Wu L, Davies SL, North PS *et al.*, 2000. The Bloom's syndrome gene product interacts with topoisomerase III. *J Biol Chem*, **275**(13):9636–44.

Wu L and Hickson ID, 2003. The Bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature*, **426**(6968):870–4. doi:10.1038/nature02253.

Xiong Z, 2008. Genome-wide recessive screens for DNA mismatch repair genes in mouse ES cells. Ph.D. thesis, Cambridge University.
URL http://www.sanger.ac.uk/research/publications/theses.html

Xu T and Rubin GM, 1993. Analysis of genetic mosaics in developing and adult Drosophila tissues. *Development*, **117**(4):1223–37.

Xu X, D'Hoker J, Stangé G *et al.*, 2008. cells can be generated from endogenous progenitors in injured adult mouse pancreas. *Cell*, **132**(2):197–207.

Xu Y and Baltimore D, 1996. Dual roles of ATM in the cellular response to radiation and in cell growth control. *Genes Dev*, **10**(19):2401–10.

Xue Y, Wang Q, Long Q *et al.*, 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol*, **19**(17):1453–7. doi:10.1016/j.cub.2009.07.032.

Yan CT, Boboila C, Souza EK *et al.*, 2007. IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature*, **449**(7161):478–82. doi:10.1038/nature06020.

Yi M, Hong N and Hong Y, 2009. Generation of medaka fish haploid embryonic stem cells. *Science*, **326**(5951):430–3. doi:10.1126/science.1175151.

Yusa K, Rad R, Takeda J and Bradley A, 2009. Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat Meth*, **6**(5):363–9. doi:10.1038/nmeth.1323.

Yusa K, Takeda J and Horie K, 2004. Enhancement of Sleeping Beauty transposition by CpG methylation: possible role of heterochromatin formation. *Molecular and Cellular Biology*, **24**(9):4004–18.

Zha S, Alt F, Cheng HL, Brush J and Li G, 2007. Defective DNA repair and increased genomic instability in Cernunnos-XLF-deficient murine ES cells. *PNAS*, **104**(11):4518–4523. doi:10.1073/pnas.0611734104.

Zhang J, Sun X, Qian Y and Maquat LE, 1998. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA*, **4**(7):801–15.

Zheng B, Larkin DW, Albrecht U *et al.*, 1999. The mPer2 gene encodes a functional component of the mammalian circadian clock. *Nature*, **400**(6740):169–73. doi:10.1038/22118.

Zhou H, Xu M, Huang Q *et al.*, 2008. Genome-Scale RNAi Screen for Host Factors Required for HIV Replication. *Cell Host and Microbe*, **4**(5):495–504. doi:doi:10.1016/j.chom.2008.10.004.

Zhu C, Bogue MA, Lim DS, Hasty P and Roth DB, 1996. Ku86-deficient mice exhibit severe combined immunodeficiency and defective processing of V(D)J recombination intermediates. *Cell*, **86**(3):379–89.

Zou L and Elledge SJ, 2003. Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science*, **300**(5625):1542–8. doi:10.1126/science.1083430.

Appendix A

Protocol: Generating libraries using the LGN cell line

Materials required

LGN cells ready for transfection on six well cell culture plate(s).

Feeder plates resistant to both G418 and puromycin.

PBC expression plasmid Qiagen maxi prep or similar, at least 1 mg/ml. PBC = pCMV-hyPBase-hCDT1. PBC mRNA can also be used.

4-hydroxytamoxifen 1 mM solution in ethanol.

M15 medium supplemented with 100 U/ml LIF, and derivatives below (all media contain LIF):

HGF medium M15 medium supplemented with 200 μg/ml G418, 1X HAT, 200 nM FIAU.

HTG medium M15 medium with 200 μg/ml G418 and 1X HT supplement (Invitrogen).

DBL medium M15 with 200 μg/ml G418 and 3 μg/ml puromycin.

Protocol

1. Wash cells twice with PBS and add 500 μl trypsin solution. Incubate at 37C for 10 minutes.
2. Quench trypsin with 500 μl M15 medium, pipette to break up cell clumps. Wash cells with PBS and resuspend in 900 μl PBS with 15 μg PBC plasmid. Transfer to an electroporation cuvette (0.4 cm BioRad).
3. Electroporate (230 V, 500 μF). Incubate at room temperature for five minutes. Transfer cells to a 10 cm feeder plate with M15 medium. Plate 1/100 of each electroporation to a six well plate to estimate the number of new insertions obtained (select in HGF until colonies are visible, stain and count). Alternatively, use Qiagen Transmessenger reagent to transfect 1–2 μg capped *in vitro* transcribed mRNA (Use Ambion mMessage mMachine T7 kit with *Avr*II-linearised pCMV-hyPBase or pCMV-PBCDT1).
4. The next day, change medium to **HGF**. Change medium daily. Passage cells at a ratio of 1:2–1:4 if they become confluent, maintaining selection at all times.
5. After eight days of selection, change medium to **HTG** for two days.
6. (Day 10) Change medium to M15, supplemented with G418 alone (200 μg/ml).
7. Cells can be expanded further if required. I typically expand until day 12–14, as cells grow more slowly than normal in HAT-containing medium.
8. Once the required expansion has been reached, harvest the cells with trypsin as above. Ensure colonies are effectively dispersed by pipetting. Count the cells and record the number (this can be used to estimate the number of cells per mutant clone). Plate the entire culture onto 10 cm plates containing M15 supplemented with 1 μM 4-OHT, at no more than 5×10^6 cells per plate. Incubate overnight.
9. The next day, change medium to M15. After two days, trypsinise the cells and replate half of each culture to a 10 cm plate containing **DBL** medium. Again, plating some cells at low density is useful to estimate the number of double-resistant clones in the library. A small number of cells can be plated in puromycin to check the induction of Cre was efficient.
10. Change medium daily and select for at least 10 days. Passage cells under selection at least once, or more if they become confluent. When passaging, plate some at low density to pick clones for analysis[1].
11. Freeze the enriched library. As the library has been expanded by over 1000× since beginning DBL selection, small aliquots can be used without affecting representation. Colonies can be picked from the low density plate and analysed by Southern blot to determine the complexity of the library.

[1]There is no need to replate the entire culture every time—for every 2–3 days of growth, the proportion to replate can be reduced by 1/4–1/2. As long as the representation of clones is retained, it is not necessary to retain all cells from every clone.

Appendix B

Primer sequences

Name	Sequence (5′ to 3′)	Purpose
ActB-F	ATGGGTCAGAAGGACTCCTA	Beta actin RT-PCR primer
ActB-R	CAACATAGCACAGCTTCTCT	Beta actin RT-PCR primer
Ccdc107-FseI-f	GCATTTAGGCCGGCCGAGCCAAGGAGACAG- -ACTGG	Primers to amplify Ccdc107 mutagenic exons
CCdc107-AscI-r	GGAATCGGCGCGCCTTTATTTCCCCACTGG- -ATCTT	—
Dom3z-f	GCATTTAGGCCGGCCCCAAGTCCTCAGACC- -CAGTG	Primers to amplify Dom3z mutagenic exons
Dom3z-r	GGAATCGGCGCGCCGCCAGCCTCTACACCC- -AGTA	—
XhoI-H3-adaptA	TCGAGATCGATACATGTA	adaptor oligo - goes into XhoI + HindIII, adds PciI and ClaI
XhoI-H3-adaptB	ACGTTACATGTATCGATC	rev comp to above
PacI-PmeI-palilinker	GTTTAAACAT	Goes into PacI site (3' AT OH), adds a PmeI site. Palindromic oligo
g1101a-a1103t	GAATGACCGAGAAGGCTGAATTCCCTCTGT--GTGCATGAA	Site directed mutagenesis primers for mutagens
g1101a-a1103t-antisense	TTCATGCACACAGAGGGAATTCAGCCTTCT--CGGTCATTC	—
c2379a-c2381a-c2388t-a2394t	GTTTTGTGTCTAGAAGTTCCATATGGGTTT--CAACCTAAGTCGTCACCCTGTAGAAA	—
c2379a-c2381a-c2388t-a2394t-antisense	TTTCTACAGGGTGACGACTTAGGTTGAAAC--CCATATGGAACTTCTAGACACAAAAC	—
HmSpAa-SPCR	CGAAGAGTAACCGTTGCTAGGAGAGACCGT--GGCTGAATGAGACTGGTGTCGACACTAGTG- -G	Splinkerette linker
HmSpBb-SPCR-GATC	GATCCCACTAGTGTCGACACCAGTCTCTAA--TTTTTTTTTTCAAAAAAA	Splinkerette linker with GATC overhang (for Sau3AI)
HmSpBb-SPCR-TA	TACCACTAGTGTCGACACCAGTCTCTAATT--TTTTTTTTCAAAAAAA	Spinkerette linker with TA overhang
HMSp1-SPCR	CGAAGAGTAACCGTTGCTAGGAGAGACC	Primer to Splinkerette linker
HMSp2-1-SPCR	GTGGCTGAATGAGACTGGTGTCGAC	—
HMSp2-2-SPCR	ATGAGACTGGTGTCGACACTAGTG	—
PB5-1-SPCR	TAAATAAACCTCGATATACAGACCGATAAA	Primers for SPCR of PB5
PB5-2-SPCR	ATATACAGACCGATAAAACACATGCGTCAA	—
PB5-seq-SPCR	TTTTACGCATGATTATCTTTAACGTACGTC	—
PB3-1-SPCR	CAAAATCAGTGACACTTACCGCATTGACAA	Primers for SPCR of PB3
PB3-2-SPCR	CTTACCGCATTGACAAGCACGCCTCACGGG	—
PB3-seq-SPCR	TTAGAAAGAGAGAGCAATATTTCAAGAATG	—
Neo-SV40-F-AscI	ATAGGCGCGCCTTGGAGGCCTAGGCTTTTG	Amps neo-SV40pA from pcDNA3 with AscI and SfiI (specific site in YTC85) for cloning
Neo-SV40-R-RCSfiI-KpnI	TTAGGCCTGATCGGCCGGTACCTGTGGAAT--TGTGAGCGGATA	—
Dym-insertion-F	AGCATAGAGGAGGAGATAAGCACTC	gPCR primers 288bp w/PB5
Dym-insertion-R	GTTTTGGGCTCTACCATTATTTATTTT	gPCR primers - 234bp w/F
PB5-R	taaataaacctcgatatacagaccgata	gPCR primers
Ddt-insertion-F	AGGTGGCTCTGTTTTTCCCTCT	gPCR primers 169 w/PB5
Ddt-insertion-R	GTATCTTAGGGACCAGAGAGAGATG	gPCR primers 232 w/F
Picalm-F	CGCAATGGATTGTCACATTTTT	TN is -, use with PB5-R, 151bp
Picalm-R	CACCTGGACTGTGAGTGAAGAC	use with Picalm-R, 220bp
Arrb2-F	TGTTAGGGTCTTCAAGAAGTCGAG	use with PB5-R, 259bp
Arrb2-R	AAGCTTGCTTAGGAACCCAGAC	use with Arrb2, 227bp
Arrb2-e1-F	gcaccatgggagaaaaacc	RT-PCR primers
Arrb2-e5-R	cttcttcagcagtcggtcct	—
Dym-e1-F	tgacctacggaaccatggag	—
Dym-e3-R	gaaatggttgcctcttccaa	—
TNP100-F3-F	TTTAGGATGGGCTTCCCTTT	Genotyping primers for TNP100 insertions

Name	Sequence (5′ to 3′)	Purpose
TNP100-F3-R	AAGACCACCGTTTTCCCTCT	—
TNP100-G10-F	AGGGCAGCTGAGTTTAAGCA	—
TNP100-G10-R	GGCAGGAAACAGGTAGGACA	—
TNP100-H10-F	AGAAACCACCCACAAAAACG	—
TNP100-H10-R	AGGGGGTTAGCCACAAGTTT	—
TNP100-D1-F	AATCTGGTGATGGCCTTCTG	—
TNP100-D1-R	AGAGCCCTGACACTCTTCCA	—
TNP100-C5-F	CACCTGCAACCATCAAACAC	—
TNP100-C5-R	TCTGCACTGGGAGAAGGTCT	—
TNP100-B9-F	TTGCCGCATTGTCTCTATTG	—
TNP100-B9-R	CCAAACCTTTGTGAAGTCGAA	—
PB5-gPCR	taaataaacctcgatatacagaccgata	—
TNP100-C8-F	TGCAGGCAAAATTCTTTTATTG	—
TNP100-C8-R	TCTCCATATGTATTCAATTACAATTTCTC	—
TNP100-F3-probe-F	TTACGGTCTGTCCCAAGGTC	Locus-specific Southern probes for TNP100 insertions
TNP100-F3-probe-R	AATGAGGCTGCAAGAGGAAA	—
TNP100-B9-probe-F	AAAATCAGTGTGTTGTCACTACCTC	—
TNP100-B9-probe-r	CCAACAAACAAAGCCAAAAA	—
TNP100-C5-probe-F	CAGTCTTAAAAATCAAGGCTGACC	—
TNP100-C5-probe-r	CCTTTACCAGGTCTTTTCAAGC	—
TNP100-C8-probe-F	AGAAAAGGGAACCGAAAGGA	—
TNP100-C8-probe-r	AGACAGGATGGAAGCCATTG	—
TNP100-H10-probe-f	GAAGGATGGAGAGGAAGGGTA	—
TNP100-H10-probe-r	CACAGCTCCCTAACCTATAACACA	—
Myo5a-RT-F	GGCAGCCCTATGATAGAAGG	RT-PCR primers
Myo5a-RT-R	TTGTGCAGCTGTCTGAATCC	—
iCre-target-1	CCTAAAGAAGAGGCTGTGCTTTGG	Rosa26:ERT2-iCre-ERT2 primers (J. Takeda, K. Yusa)
iCre-target-2	CATCAAGGAAACCCTGGACTACTG	—
iCre-popout-1	TAAGGGATCTGTAGGGCGCAGTAGTCCAGG	—
iCre-popout-2	TAAGCTAGCTTGGGCTGCAGGTCGAGGGAC	—
AseClaNsi-linker-A	AGCTAATCGATTAATCGCATTCAATGCATG- -CGTCAATTTTACGCAGACTATCTTTCTAGG- -GTTAA	Linker to reconstruct PB3 up to NsiI site
AseClaNsi-linker-B	AGCTTTAACCCTAGAAAGATAGTCTGCGTA- -AAATTGACGCATGCATTGAATGCGATTAAT- -CGATT	—
FL2-B4-F	CCCTGTCCTTGGTTTATGGA	Genotyping primers for further clone-by-clone enrichment experiments
FL2-B4-R	TACCGCCCTTAAAGAACCAG	—
FL2-C1-F	CTCTGGGATCCCTCCTCTTC	—
FL2-C1-R	CCCAAGACTGAGTGCCATCT	—
FL2-C4-F	AACCCAGGCCTCTGAAGTTT	—
FL2-C4-R	CTCTGCCTCTGAGTGCCTTT	—
FL2-A7-F	AAGCATGGGCTACTTCTCCA	—
FL2-A7-R	ATGCAGTGTCCAGTGCTGAG	—
FL2-C6-F	AGAGACCATGGATGCCAGAC	—
FL2-C6-R	GGTATTTTGGTGGTGGTGGT	—
FL2-C9C10-F	ACTCTGCACATGGCACACAT	—
FL2-C9C10-R	GGAGGCTCCTTCCTCATTCT	—
FL2-A3-F	CGTTTGTTCTGCAGGTCTGA	—
FL2-A3-R	CAACTGAGGAGTGTGGCAGA	—
FL2-A4-F	TTTCCGGGCACATCTTTATC	—
FL2-A4-R	ATGATCCCAGATGCCTTCAG	—
FL2-A11-F	GTGGGGCTCATGTAGGAAGA	—
FL2-A11-R	GTAGCTGCCTCCCAAGACTG	—
FL2-B9C7-F	AATAGCCGCATACCTGCATC	—
FL2-B9C7-R	CGGAGCTGTTCCTTGTCATT	—
PB5-gPCR	taaataaacctcgatatacagaccgata	—
Sall1-e1-F	ACCCGGAAGAGGGAGTACAG	—
Sall1-e3-R	GGCATCCTTGCTCTTAGTGG	—
Acpp-e2-F	TTCCTACCGACCCCATTACA	—
Acpp-e4-R	ATCCCCTCTGGAGGAAACAG	—
TNP100-B6-F	GCTCTGAGCCTGGGAGATTA	—
TNP100-B8-F	ATCTTGTGGGATGGCATAGC	—
TNP100-B11-F	CCACAGCCTGGGAAACTATT	—
PB3-gPCR	acggattcgcgctatttaga	—

Name	Sequence (5′ to 3′)	Purpose
XX-Tmp-F	TGGATCAACAGAACAAGGAAA	Primer to amplify tag insertion in bleomycin-sensitive pool
TV28-3-geno-F	TAAACCTCGATATACAGACCGATAAAACAC	TV28 targeting genotyping, by long range PCR
TV28-3-geno-R	CTACCTCACACCATGCACAAAAATAAAT	—
TV28-probe-F	TGATTTAATACCAGCACATCCAAATTAT	TV28 Southern probe
TV28-probe-R	ACCTTTCCAGTTAAAGTTGAGAGATCAT	—
FL2-E7G7-F	AACCCAGGCCTCTGAAGTTT	Genotyping primers for further clone-by-clone enrichment experiments
FL2-E7G7-R	GACCAGGATCCTTGGACTCA	—
FL2-D6-F	ATGTCCCTCTCCTGTGTTGG	—
FL2-D6-R	CCTCGCTTCACCTCTGAGAC	—
FL2-F11-F	AGGGTGGGGATAGAGCAGAT	—
FL2-F11-R	CTTGCTCTTGGCAACTTGTG	—
HmSpBb-NcoI-CATG	CATGCCACTAGTGTCGACACCAGTCTCTAA--TTTTTTTTTTCAAAAAAA	Splinkerette linker with CATG overhang (for NcoI)
TV28-jump-F	CTGGTCAAGGAAATGGTGCT	Primers for amplifying Hprt donor locus after transposon jumping
TV28-jump-R	CACCAACACACCAGCTCAAC	—
CDT1-R	CTCTAGCATTTAGGTGACACTATAGAATAG--GGCCCTCTAGATGCATGCTCTCATTACAAC--TCCCCAGCATCCTGGGCACT	Amplify human CDT1 fragment plus homology arms to hyPBase plasmid
CDT1-F-AscI	AGAAGGTCATCTGCCGGGAGCACAACATCG--ACATGTGCCAGAGCTGTTTCgggcgcgccC--CCAGCCCCGCCAGGCCCGCACTCCGCGCC	—
LGNL1-A1-F	CGAACTCAGAGATCTGCTTGTCT	Genotyping primers for LGNL1 allelic clones (used with PB5-gPCR primer)
LGNL1-A1-R	GAAGGTGAGGTCACTCTGAGCTA	—
LGNL1-A2-F	CCCAAGTCCTCTGTAATTCCTCT	—
LGNL1-A2-R	TGTTTTACAGACTGGATGGCTTT	—
LGNL1-A3-F	CTGATGACATTACACCTGCGTTA	—
LGNL1-A3-R	GAGAGATGGCTCAGTGGTTAAGA	—
LGNL1-A4-F	CTCAAAAGCCTTTCTCTCCTTTC	—
LGNL1-A4-R	CTCCTTTCTCACCTCAGTAGCAA	—
LGNL1-A8-F	TGGCTTCTATCTACCCACAGCTA	—
LGNL1-A8-R	CCATCACATGTGGCCTATATTTT	—
LGNL1-B11-F	TTATGATTGCCTCAGGATCATCT	—
LGNL1-B11-R	AGCAACTCACTGCAAGAGAGAAC	—
LGNL1-B1-F	GAACCAAAGGGTAAAAGGAGAGA	—
LGNL1-B1-R	CCCAGAGCATTTTACATTTCAAG	—
LGNL1-B2-F	AAGGAAACCTGAAGAAACCAGTC	—
LGNL1-B2-R	CTAGTCAGCAGTGCCCAATATCT	—
LGNL1-B5-F	CTGGCTCTGCTGAAGATAAACAT	—
LGNL1-B5-R	CATCAGATCCCATTACAGATGGT	—
LGNL1-B6-F	TAGGGTTTCTCTGTGTAGCCTTG	—
LGNL1-B6-R	TTCTCCATGCTCAGTCACACTTA	—
LGNL1-B7-F	CCCCATCTTCTGAGACTAAAGGT	—
LGNL1-B7-R	GTGTGTTACAAGGCAAGCTCTCT	—
LGNL1-B8-F	AGTGTGTCCAAAAAGATCAAGGA	—
LGNL1-B8-R	GGTTCTAATGCCTTGGAGAAGAT	—