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-

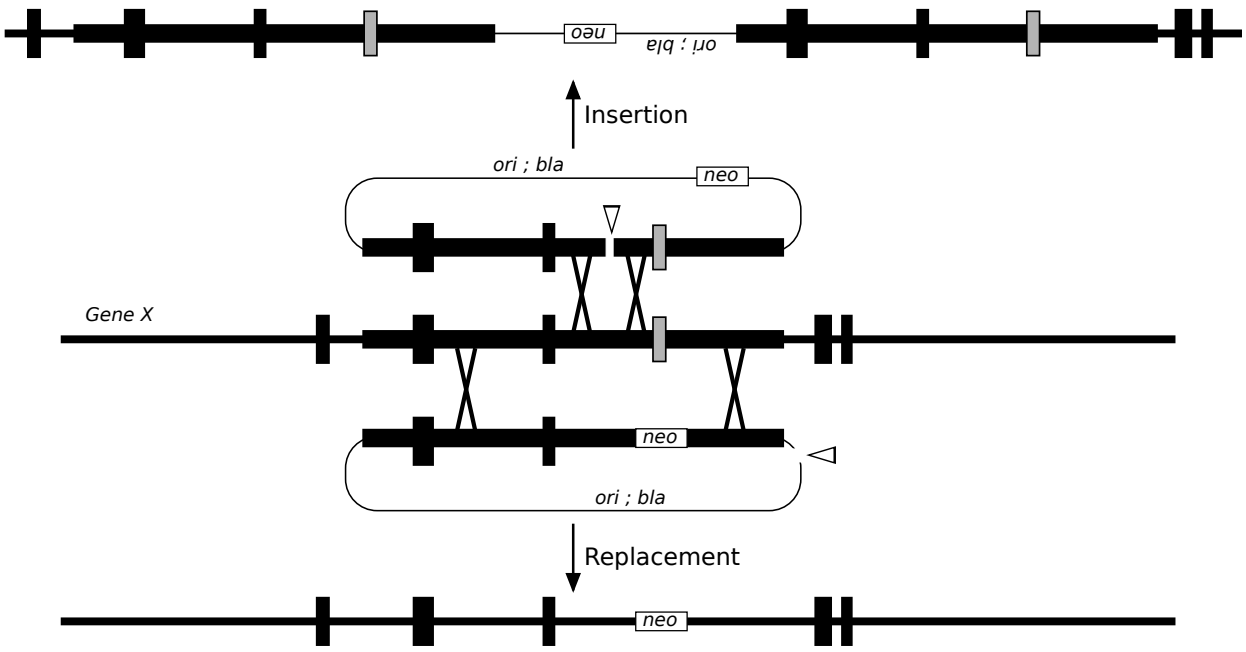

Figure 1.1: Insertion and replacement targeting vectors. The structures of insertion (top) and replacement (bottom) vectors targeting a hypothetical gene are shown. An open arrowhead indicates the site for linearisation by restriction digest. Thick line indicates homology between the genome and the targeting 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

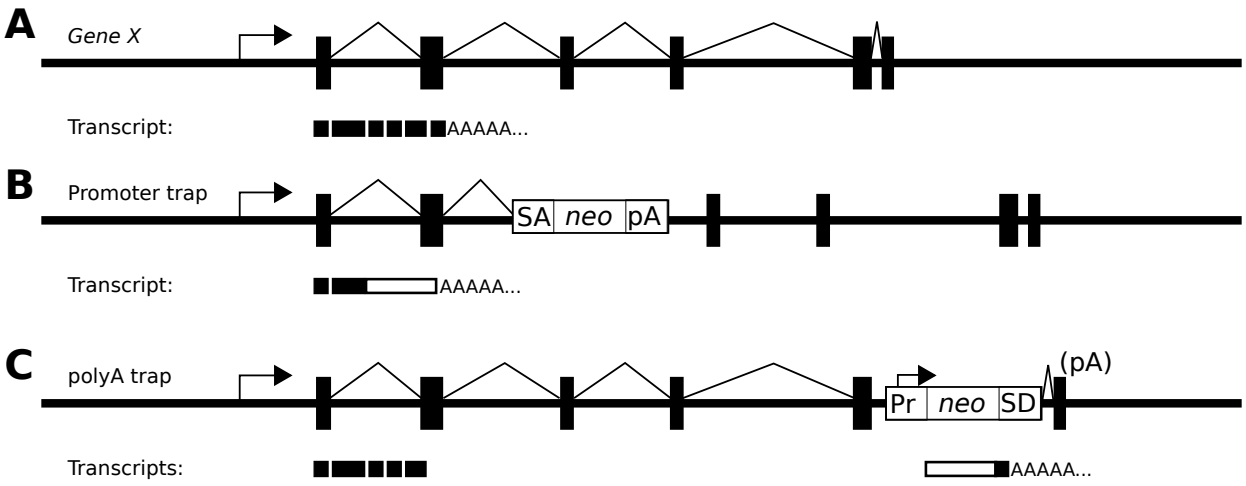

Figure 1.2: Types of gene trap vector. A—A hypothetical gene showing splicing pattern. Exons represented as black boxes. B—Promoter trap vector, consisting of splice acceptor (SA), reporter gene (*neo* in this case), and a polyadenylation signal (pA). C—polyA trap vector, with its own promoter (Pr) and 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

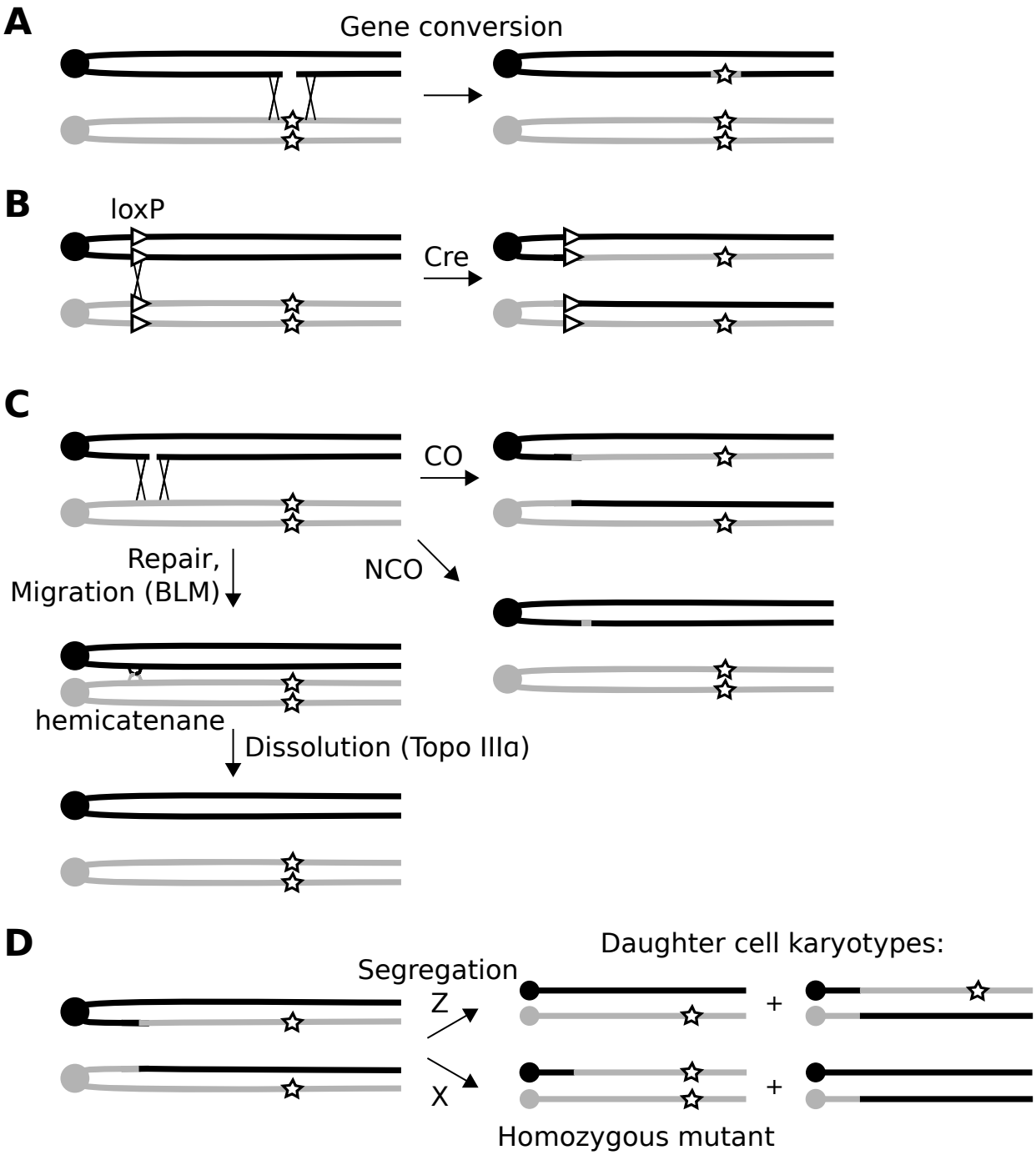

Figure 1.3: Mitotic recombination leading to LOH in heterozygous cells. A—Gene conversion, B—Induced mitotic recombination, C—Mitotic recombination in *Blm*-deficient cells. Homologous chromosomes are indicated in black and grey, and are shown after replication, so consist of a pair of sister chromatids. 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

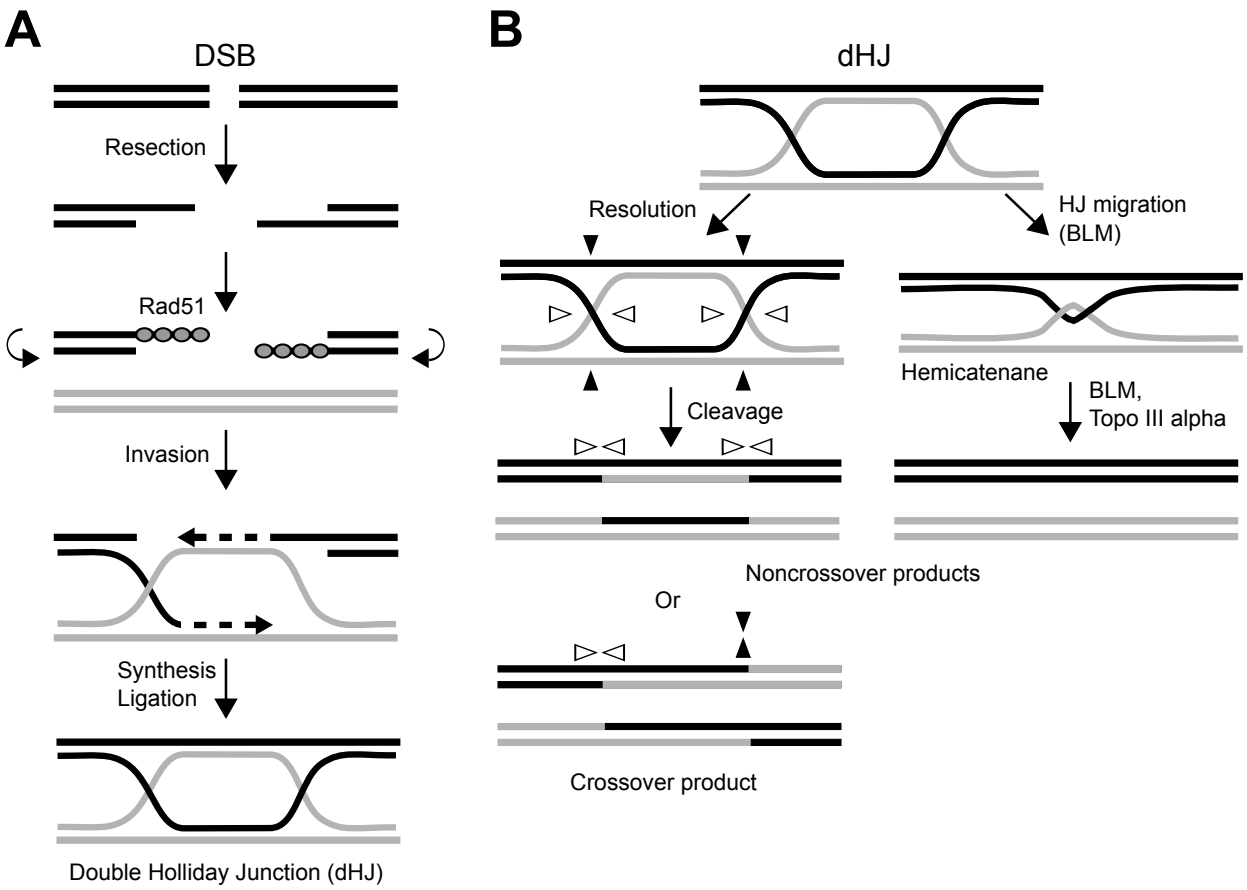

Figure 1.4: Formation and resolution of double Holliday junctions. A—Formation of a dHJ during double strand break (DSB) repair by homologous recombination. B—Pathways for resolution by structure specific endonucleases or dissolution by BLM/Topo IIIα. Products of cleavage in the orientations indicated 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.

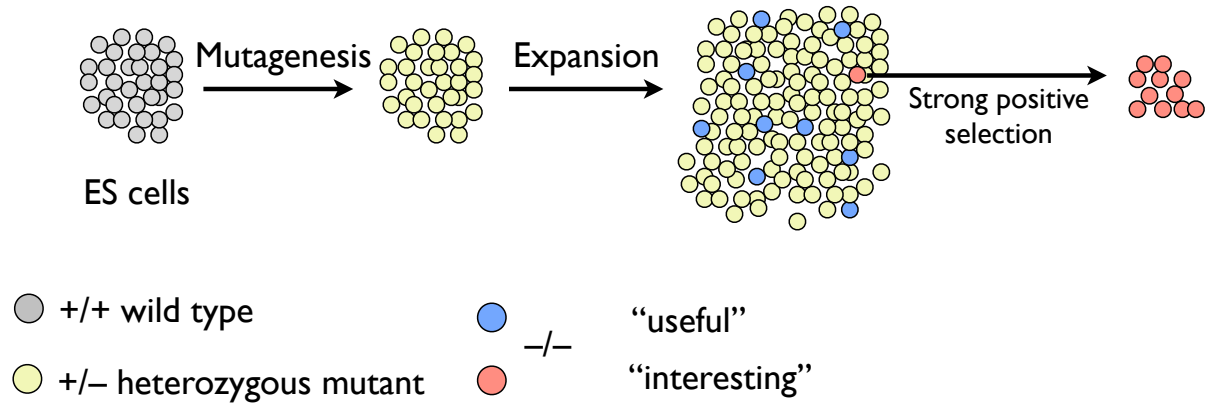

Figure 1.5: Screens for selectable phenotypes in *Blm*-deficient cells. Expansion of a population of random heterozygous mutants results in rare homozygous cells segregating. These can be isolated if the mutant phenotype is strongly selectable.