Chapter 1. General introduction

1.1 Introduction to genetic screening and transposons

1.1.1 The human genome and functional studies

The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the success of the Human Genome Project (HGP), which was completed in 2003 (1,2). Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes. These genomic sequences, including approximately 20,000–25,000 genes in the human genome that have been annotated by contemporary advanced Bioinformatics technology, must be functionally annotated to assign genes with pathways and functions. Functional studies have become a major trend in the post-genomic sequencing era and can be divided into many branches such as genetics, biochemistry, developmental biology and structural biology. Each specialist subject represents a unique aspect of biology and uses specific technologies to explore the function of a gene and the corresponding protein product. Biological processes such as protein synthesis, cell division and embryogenesis are realised

by the interaction and coordination of thousands of proteins and other small molecules. In the case of malfunction, the biological system is disrupted and may generate an abnormal phenotype or lead to disease as a whole. Therefore, understanding the function of each gene and their encoded product could not only help us better understand the mechanism of biological activities, but could also improve the prevention and treatment of diseases.

1.1.2 Genetic screening is a powerful tool for functional studies

Genetic screening is an intervention that generates large numbers of genetic changes (mutations), and thereby helps identify the genes that are responsible for certain biological activities or phenotypes. Genetic screening has proved its usefulness in biological studies in a number of classical experiments: using genetic screening, geneticists have specified the mechanisms and genes responsible for cell-cycle control in yeast (3,4) , the genes involved in embryonic development in flies (5), and the genes involved in programmed cell death in the worm (6). These Nobel prize-winning experiments, which identified core genes and pathways responsible for how cells function and how organisms develop, have paved the way for current biological research into more advanced scenarios.

The initial and fundamental step in genetic screens is the introduction of genetic changes i.e. mutations. At the organism level, mutations can be classified as hypomorphic (reduced gene function, of which a null mutation is the most extreme example), hypermorphic (increased gene function) or neomorphic (changed gene functions). The mutation type and frequency with which they occur in the genome are largely dependent on the mutagen used. For example, the classical chemical mutagen N-ethyl-N-nitrosourea (ENU) can be used to generate point mutations or small deletions of 20-50 base pairs in the germ line at a frequency of 1.5×10^{-3} per locus, per generation of offspring. The murine leukaemia virus (MuLV), which is a retrovirus, can be used to disrupt endogenous gene expression or generate gain-of-function mutations to overexpress genes.

1.1.3 Dominant and recessive genetic screens

Genetic screens can be classified as dominant or recessive screens. In a dominant screen, or gain-of-function screen, a gene is ectopically expressed or expressed in a different location or at a different time point in development, generating a gain-of-function phenotype of that gene

to study its function. These ‘hypermorphic’ or dominant mutations are normally generated by using insertional promoters or creating dominant point mutations to activate gene function.

In contrast, in a recessive or loss-of-function screen, gene expression is reduced, of which a “null” mutation is the most extreme example, which results in a loss-of-function phenotype for functional study. In a recessive screen, the ‘hypomorphic’ or recessive mutations are normally generated by insertional mutagens which disrupt gene expression, or from a point mutation which creates a stop codon in the open reading frame.

Both dominant (gain-of-function) or recessive (loss-of-function) genetic screens can be powerful tools for dissecting gene function, especially in haploid systems such as bacteria. However, screening has so far proved difficult in mammalian cell culture, due to the difficulty in generating homozygous loss-of-function mutations.

1.1.4 Forward and reverse genetic screens

When considering the screening process genetic screens can be categorised into forward or reverse genetic screens. Forward or traditional genetic screens involve the introduction of mutations at random, and then cells or organisms are screened for a particular phenotype and the genes associated with the phenotype of interest are subsequently identified. The advantage of forward genetic screens is that the generation of mutations is quick and inexpensive when using chemical or insertional mutagens. However the mapping of each mutation can be tedious and time-consuming.

The availability of complete human and model organism genome sequences has allowed us to assess the phenotype from specific gene(s), usually by generating gene knockouts using homologous recombination or gene knockdowns using RNA interference (RNAi). This approach is called ‘reverse genetics’, the advantage of which is that the objects of study are specific and so their functions are relatively easy to evaluate. In contrast to the forward genetics approach, the problem of reverse genetics is that it is much harder to generate the specific mutation in the first place. In the meantime, because the phenotypes may be cell-type or developmental-stage specific, the reverse genetic screen is normally taken place in a defined biological context, making it very difficult to identify rare mutations associated with certain phenotypes.

1.1.5 Reverse genetic screening for functional studies

In a reverse genetics study, or candidate gene approach, a specific gene is defined and the work is to identify the phenotype associated with this gene and therefore deduce the gene function. The most common approach for reverse genetics studies is to delete a gene in the genome, therefore depleting the gene coding product, and look for its loss-of-function phenotype. Homologous recombination, or 'gene targeting' is routinely used to disrupt genes in yeast gene function screens. With the development of gene targeting technology in mouse embryonic stem (ES) cells, the genes in the mouse genome can be easily deleted using homologous recombination. Two ambitious projects to systematically delete every mouse gene in the genome were launched in 2004 (7,8). These modified ES cells could then generate a whole mouse with this gene deletion to perform the *in vivo* loss-of-function study in a more advanced organism. In addition, homologous recombination can also be used to modify the gene coding sequence, eg. by creating a point mutation(s) or by adding or removing a functional domain to study gene function in its modified state.

Another example of loss-of-function studies involves RNAi, a technology developed by Andrew Fire and Craig C. Mello in 1998 to deplete the endogenous messenger RNA by using double-stranded RNA injected into host cells (9). The mechanism of RNA-mediated interference involves hybridization and degradation of the endogenous mRNA by DICER and the RNA Induced Silencing Complex (RISC), therefore depleting the cells of the gene coding product (10-12). When compared with gene targeting, RNAi does not require the generation of targeting vectors as small hairpin RNAs (shRNA) or small interfering RNAs (siRNA) can be synthesised *de novo*, therefore simplifying the process of generating loss-of-function lines. However, the efficiency of RNAi can vary between individual genes. Furthermore, a phenotype may develop as the result of 'off-target' effects, which are caused by the cross–reaction of the small interfering RNA (siRNA) to other mRNAs with sequence homology to the candidate gene. Extreme caution is therefore required in the design of siRNAs for RNAi experiments (13,14).

Gain-of-function approaches can also be used in reverse genetic studies to investigate gene function by over-expressing a particular gene of interest and observing the phenotype. This can be done by simply introducing expression vectors into cells to make transgenic cell lines or by over-expressing genes *in vivo* in transgenic animals. Nevertheless, the phenotype which is observed by overexpression methods may not represent the protein function at its

physiological level. Therefore, in more advanced studies, a large fragment of genomic sequence that includes the gene coding region as well as neighbouring regulatory sequences is normally used to investigate gene function. These large fragments are catalogued within a bacterial artificial clone (BAC) library.

1.1.6 Classical forward genetic screens

Although the reverse genetic approach is a rational strategy for gene identification, the preparation of large numbers of targeting constructs or siRNA/shRNA for RNAi is an expensive and time-consuming job. In addition, reverse genetic screens largely depend on the capacity of the screen itself (e.g. how many open reading frames to target); this largely restricts the identification of candidate genes. Apart from these issues, the complexity of the genes, pathways and networks that dictate many cellular phenotypes rarely makes it possible to employ a one-by-one candidate gene (reverse genetic) approach to identify potential mediators of a biological process. In contrast, genome-wide forward genetic screens which may be performed without making *a priori* assumptions about the candidature of individual genes in a process, and therefore represents a powerful approach for gene discovery.

Classical forward genetic screens in higher order organisms have been performed using ionizing radiation or chemical mutagens to generate point mutations or deletions to target a full spectrum of genes in the genome. While these approaches can be extremely efficient at generating mutant cell lines with a phenotype of interest the subsequent identification of causal mutations is often cumbersome. This is particularly the case for traditional chemical mutagens such as N-ethyl-N-nitrosourea (ENU) and ethyl methane sulphate (EMS), which generate genome-wide point mutations. In identifying the mutation responsible for the phenotype of interest validation must be carried out due to the significant levels background noise. Ionizing radiation is a powerful tool for mutagenesis, generating sufficiently small chromosomal rearrangements so that a candidate gene can be identified using approaches such as comparative genomic hybridisation (CGH). However, it requires high doses of radiation to be used which generates a significant number of rearrangements, the majority of which represent background. Lower doses produce rearrangements of large chromosomal regions, in some cases containing hundreds of genes, complicating follow-up analysis. The following sections will discuss each of these mutation strategies in detail, with emphasis on their applications in higher eukaryotic systems such as mice.

1.1.6.1 N-ethyl-N-nitrosourea (ENU) mutagenesis screen

ENU has been used as a mutagen in forward genetic screens for many years. In addition to the high mutagenesis rate and ability to generate point mutations and small deletions (15), this mutagen is easy to prepare and handle, has low toxicity and can be used to generate germ-line mutations if necessary. The mutagenicity of ENU is due to its capacity to transfer an ethyl group to oxygen or nitrogen radicals in the DNA nitrogenous base, which causes nucleotide mismatches and ultimately results in base pair substitutions or base pair losses sometimes. These single nucleotide mutations include A/T to T/A, A/T to G/C, G/C to A/T, G/C to C/G, A/T to C/G and G/C to T/A (16).

ENU is the most potent mutagen used in mice and mammalian cells, with a mutation rate of 1 in 1,000 gametes (17,18); approximate 5 times more efficient than X-ray irradiation. By the late 1970s, a large collection of mutant mice were established by the international research community for study needs. In an ENU genetic screen, usually male mice (G0) are injected with ENU to introduce mutations into the genomes of their gametes. Mating the ENU-treated male mice with untreated female mice produces the first generation offspring (G1), which carry mutations and are thus ready for a dominant screen. *Clock* – a gene that controls circadian rhythm in mice was identified in this way (19). The G1 offspring can be backcrossed with wild type mice to produce a mouse line with the same mutation (G2). Intercrossing of G2 offspring generates homozygous mutant mice for recessive screens, an example of which includes a screen that resulted in the identification of embryonic lethal mutations in mice (20).

A number of genome-wide, dominant and recessive screens that were performed in mice using ENU mutagenesis were reviewed by a series of publications (16,21,22). These studies have generated invaluable information about mouse physiology, pathology and genetics. However, ENU mutagenesis screen have several drawbacks and limitations. Firstly, ENU has a strong bias towards A/T base pairs (87%) (16). Moreover, due to the lack of a molecular tag, tracing the mutations introduced by ENU is a rather time-consuming and laborious process. Therefore, ENU is generally substituted by other mutagenesis methods in contemporary research.

1.1.6.2 X-ray irradiation

X-ray is a form of electromagnetic radiation with a wavelength in the range of 0.01 to 10 nanometres. In many languages, X-radiation is called Röntgen radiation, after Wilhelm Conrad Röntgen, who discovered and named X-rays to signify an unknown type of radiation. Irradiation causes both direct and indirect effects on DNA. Direct effects lead to ionization of bases after the direct absorption of the radiation energy by DNA. Indirect effects are created when DNA reacts with surrounding ionized molecules. Around 65% of DNA damage is caused by the indirect effects and 35% by direct ionization. Ionizing radiation in cells normally causes a huge variety of DNA lesions, such as DNA-protein cross-links, base damage, and single and double-strand breaks that can result in deletions (23,24).

As early as the 1920s, X-rays have been used to irradiate mice to induce mutations (25). It was first used in large-scale mouse mutagenesis experiments in the Oak Ridge National Laboratory (USA) and the Medical Research Council Radiobiological Research Unit (UK) (26). Both programmes were initiated to investigate the effects of various forms of radiation on mice. Although the X-rays have been used for decades, the relationship between deletion length and irradiation dosage has not been identified. As DNA is packed around nucleosomes and organized in chromatin, radical clusters of irradiation can produce double-strand breaks at sites that are several kilobase pairs (kb) or even 700 kb apart (27). X-rays have been shown to introduce large deletions (200-700 kb) around the *Hprt* (hypoxanthine-guanine phosphoribosyltransferase) locus on the X chromosome in mouse ES cell lines (28) and experiments cells under drug selection showed that deletions could be as large as 70 Mb (29,30).

X-ray mutagenesis is a highly efficient method for introducing genome-wide mutations. The mutation rate for X-ray irradiation (13-50 $\times 10^{-5}$ per locus) is about 20-100 times higher than the spontaneous mutation rate (5×10^{-6} per locus) in the mouse, which makes it easy to saturate the genome and generate a large range of mutations, including deletions, duplications, inversions and translocations. When combined with recent whole genome technologies such as comparative genome hybridisation (CGH) and gene expression arrays, X-ray irradiation is a powerful tool for mutagenesis studies. What is more, X-ray irradiation causes chromosome rearrangements, which leaves a molecular marker for localising the mutated genes. However, as X-rays mainly generate deletions in the genome they can affect multiple genes, therefore it is hard to dissect individual gene function using this method. In addition, the irradiation

dosage is difficult to control; germ line mutagenesis requires high doses of irradiation which causes cell death in cells that are highly sensitive to DNA-damage such as those of the bone marrow.

1.1.7 Insertional mutagenesis screens

1.1.7.1 Introduction to insertional mutagenesis

ENU and X-ray irradiation can mutagenize genomic DNA in a highly-efficient and near unbiased manner. Using these methods, many genes responsible for key pathways and biological functions have been identified. However genetic screening by these classical methods is normally considered 'dirty' since there is a high level of background mutation and a huge effort is required to trace the gene mutation(s) that cause the phenotype. When compared to classical mutagenesis using ENU or X-ray, insertional mutagenesis is a much 'cleaner' and more delicate method of genetic screening. Insertional mutagenesis involves the insertion of an exogenous DNA fragment (insertional mutagens) into the host cell genome. These insertional mutagens could either be a gene-trap vector, retrovirus DNA or transposon. While ENU or X-ray mutagenesis predominantly generates loss-of-function mutations, insertional mutagens may induce either loss-of-function or gain-of-function depending on the genetic elements carried. Insertional elements provide considerable flexibility for modification depending on the need of the experiment or screen. Another advantage of insertional mutagenesis is that it leaves a molecular marker for mapping the insertion sites, providing a quick and simple way for tracing candidate genes. Therefore, after the success of the first insertional mutagenesis study in 1976 (see below for further details) (31), insertional mutagenesis became increasingly popular for large-scale mutagenesis studies in mammalian systems.

1.1.7.2 Types of insertional mutagens

1.1.7.2.1 Retroviruses

Retroviruses are a class of enveloped virus that replicate their genome, a single-stranded RNA molecule, via a DNA intermediate. Following infection, the viral genome is reverse transcribed into double-stranded DNA for integration into the host genome. The retroviral genome normally contains at least three genes: *gag* to encode core proteins, *pol* to encode the reverse transcriptase and *env* to encode the protein envelope. At both ends of the viral

genome are long terminal repeats (LTRs) which contain promoter and enhancer elements, as well as other signal sequences for viral splicing and integration (**Figure 1-1 A**). The integration of a retrovirus may result in a loss-of-function mutation if it is integrated into the coding region of a gene, or a gain-of-function mutation if it inserts into a promoter region, which uncouples the gene from its endogenous promoter and expression is driven by the viral promoter/enhancer elements in the LTR region.

The first attempt to introduce an exogenous retroviral DNA into the mouse germ line was reported by Jaenisch in 1976 (31). Jaenisch used the murine Moloney leukaemia virus and found that the expression of a host gene could be increased by the viral enhancer element. Viral genomes have since been optimized to enable better rates of insertion and mutagenesis upon integration into the host genome (**Figure 1-1 B**). For example, the viral genes (*gal*, *pol* and *env*) may be replaced with transgenes of interest and the plasmid can be introduced into a packaging cell line that has been engineered to express all three genes that are required for viral reproduction (*gal*, *pol* and *env*) (**Figure 1-1 C**). The packaging cell lines then produce infectious retrovirus in the culture media which can be used to transduce other cell cultures. However, as the non-essential genes in the modified viral genome lack these packaging proteins, once introduced into the host cell the retrovirus is not able to produce virions and infect other cells (**Figure 1-1 C**).

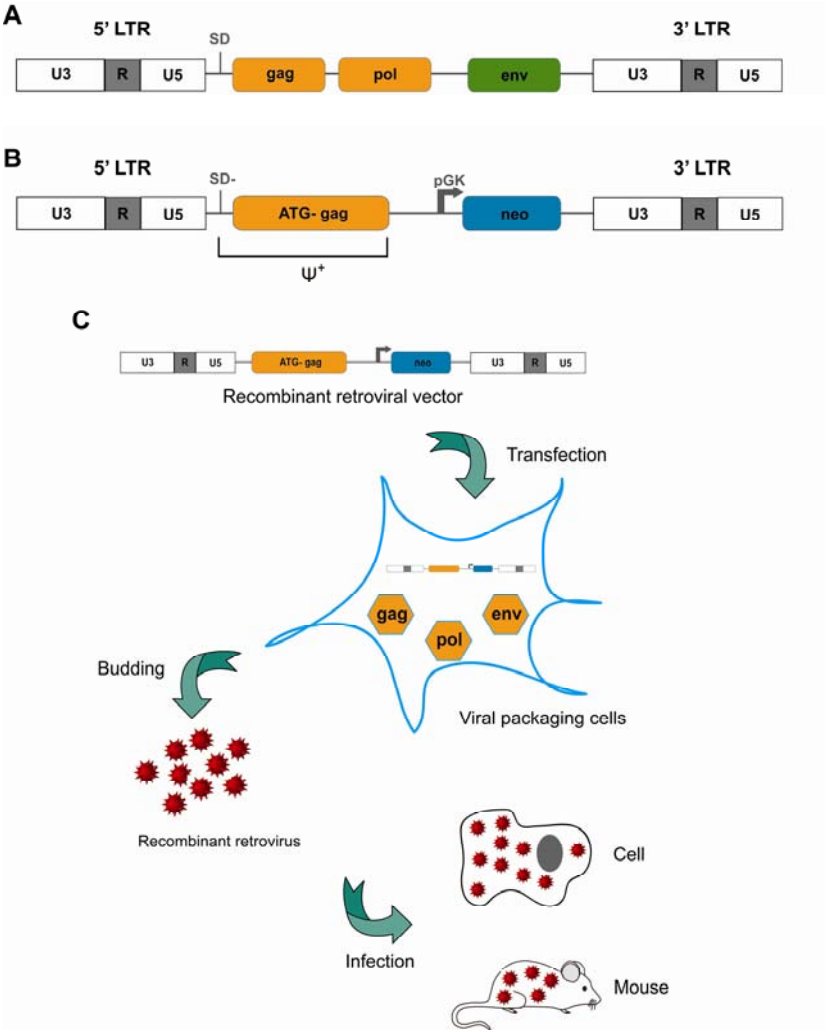

Figure 1-1. Retrovirus genome structure and recombinant retrovirus production

(A) The structure of a wild type retrovirus genome. As has been described previously, it contains the long terminal repeat (LTR) at two ends, flanking the coding sequences for viral protein core (*gag*), reverse transcriptase (*pol*) and envelope protein (*env*). SD, viral splice donor. (B) Structure of a recombinant retrovirus genome from the retroviral vector pBabe (32). Gene coding sequences including the *pol* and *env* are deleted. The splice donor and *gag* sequence is kept to facilitate viral packaging. The viral splice donor is mutated (SD-) and the start codon of the *gag* gene is deleted (ATG- *gag*). (C) Procedure for infectious retrovirus production. To produce the virus with the recombinant retrovirus genome, the recombinant retroviral vector DNA is first transfected into a viral packaging cell line, which expresses the proteins that are required for viral reproduction *in trans*. The infectious viral particles produced into cell culture are collected afterwards for infecting cell lines or mouse tissues.

1.1.7.2.2 Gene-trap mutagenesis

Since the retroviruses have a strong bias during genomic integration, the availability of mouse ES cell technology in the mid 1980s stimulated the design of new insertional mutagens for large-scale mutagenesis studies. The development of gene-trap technology enable the efficient generation of loss-of-function mutations in ES cells on a large-scale, hence this has become a popular tool for mutagenesis studies. Gene-trap vectors normally contain a promoterless selection marker or reporter gene after a strong splicing acceptor (SA). Upon integration into genome the selection marker/reporter gene is only expressed when the vectors integrates into the region downstream of the promoter/enhancer of an endogenous gene so that the gene-trap vector can utilize the endogenous transcriptional elements for expression.

The basic gene-trap vector includes an enhancer-trap, promoter-trap and polyadenylation signal (polyA) trap (**Figure 1-2 A-C**). Enhancer-trap vectors contain a minimal promoter that is not functional. The selection marker is only expressed when inserted next to an endogenous enhancer element. Because the enhancer elements may be localized far away from the coding region of a gene, enhancer-trap vectors do not normally integrate into the coding region, therefore these types of vector are not widely used for mutagenesis studies. Promoter-trap vectors contain a promoter-less reporter gene immediately after a strong splicing acceptor (SA) site. This design results in activation of reporter gene expression if the vector integrates downstream of an endogenous promoter. The vector normally contains a polyA sequence for terminating the expression of a trapped gene, thus resulting in a loss-of-function mutation. A polyA-trap vector contains a reporter gene with its own promoter but which lacks the polyA signal. The reporter gene is expressed but the transcript is not stable unless the vector inserts into an endogenous gene, upstream of a splicing acceptor and a polyA signal.

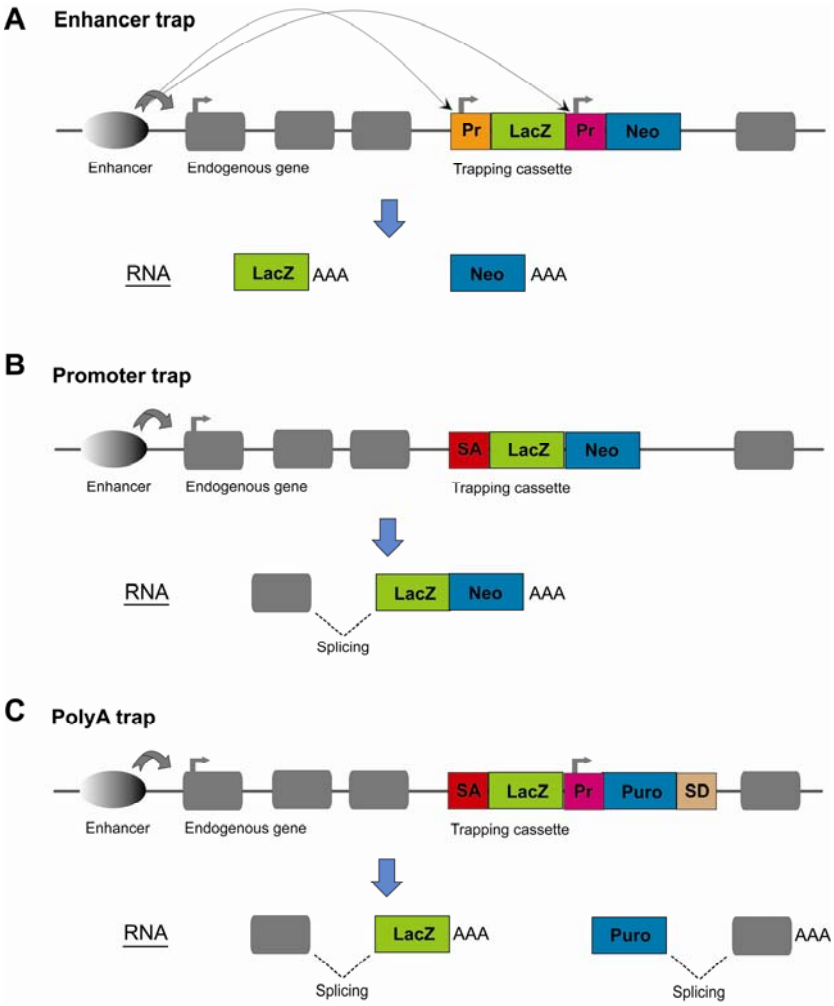

Figure 1-2. Schematic of three basic gene trap strategies

(A) Enhancer trap. The *lacZ* and *Neo* reporter genes are driven by minimal promoters (Pr) to synthesize *LacZ* and *Neo* transcripts separately. The expression level is largely enhanced by the endogenous enhancer during integration. (B) Promoter trap. The expression of *LacZ* and *Neo* fusion transcript is driven by the endogenous gene promoter while integrating into gene coding sequences. (C) PolyA trap. *Puro* is transcribed from an autonomous promoter (Pr) and spliced from the splice donor (SD) into endogenous genes while integrating into gene coding sequences. *LacZ* trap cassette may also be combined to monitor the integration into endogenous genes. SA – Splice accepter; SD – splice donor. Pr – promoter or minimal promoter.

1.1.7.2.3 Electroporation versus retroviral based gene-traps

Trapping vectors can be introduced into the genome by either electroporation or retroviral infection. The simplest way to perform gene-trap mutagenesis is to electroporate the linearized gene-trap vector directly into mammalian cells, which does not require the produce of virion particles. Gene-trap vectors that are introduced into cells by electroporation can integrate into the genome randomly, but the biggest disadvantage is that integrations are always accompanied by DNA concatermerization, which results in ectopic reporter expression and can complicate the identification of the insertion sites by 5' RACE or linker-based PCR. Theoretically there is no limitation on the size of the trapping vector, however sometimes the vector can be truncated during electroporation, for example the loss of flanking sequences can make mapping the insertion sites problematic.

The high infection rate and low cost have made retroviruses a powerful tool for delivering gene-trap vectors into host cells. The first retroviral gene-trap vector was designed by Von Melchner *et al.* in 1989 (33). In this design, the gene-trap cassette is inserted into the U3 region of the 3' long terminal repeat (LTR) and replaces the viral enhancer. After viral integration, the provirus carries a duplicated gene-trap cassette in both of the 5' and 3' LTRs. The cassette in the 5' LTR is situated just 30 bp from the host genome and is activated by transcriptional read-through rather than splicing. Two years later Friedrich *et al.* designed another version of retroviral gene-trap vector called ROSA (reverse orientation splice acceptor). In the ROSA vector, the gene-trap cassette was placed between viral LTRs in the opposite orientation relative to viral transcription. In this design, the cassette is activated only by a splicing event (34).

In contrast to electroporation which results in the formation of concatemers during integration, gene-trap mutagenesis using a retroviral vector results in the integration of a single copy of retrovirus into one genomic locus. In addition, conditions can be optimized for retroviral based gene-trapping so that most of the cells will only contain a single copy of the gene-trap vector. Retroviruses have a propensity to integrate into the 5' portion of a gene, which is more likely to generate null alleles. However, retroviral vectors also have limitations. Firstly, the packaging size of the retrovirus is highly limited and the packing efficiency drops significantly as the size increases. Secondly, the viral insertion can induce retroviral-mediated gene silencing in the genome. Thirdly, retroviral integration could result in trapping 'hot-

spots', but the same problem also exists for the electroporation-based gene-traps and can be somehow solved by using different trapping vectors.

1.1.7.2.4 Transposons

Transposons are mobile genetic elements formed during genetic evolution. They represent another class of widely used insertional mutagens and will be discussed in the following section.

1.1.8 Transposon-mediated mutagenesis

1.1.8.1 Introduction to transposons

Transposons, or transposable elements are mobile genetic elements which have been identified in many organisms including maize, insects, worms and humans. More than 40 % of the human and mouse genomes are composed of transposon-derived sequences (1,35). Transposons were first discovered in the maize genome by Barbara McClintock (36), for which she was awarded the Nobel Prize in 1983. In her studies, she identified the *Ac/Ds* transposons, two members of a family with around 100 transposons. The *Ds*, or dissociation locus, was the first mobile locus to be discovered, but it was incapable of transposition by itself. The second locus to be discovered - *Ac*, or activator locus, is an autonomous element that is capable of transposing itself and can also induce the transposition of non-autonomous elements (such as *Ds*). The idea of transposable DNA elements was not fully accepted until the insertion sequence (*IS*), a transposon-mediated resistance to antibiotics, was discovered in bacteria in 1975 (37).

Transposons can be classified into two large groups based on their mechanism of transposition. Class I transposons, or retrotransposons, transpose in the genome by a copy-and-paste mechanism: they first transcribe themselves into RNA molecules, then reverse transcribe back into DNA by reverse transcriptase at the site of integration. Class II transposons are called DNA transposons. In contrast to the retrotransposons, they transpose from one position to another in the genome by a cut-and-paste mechanism. In addition, there is a third class transposon called Miniature Inverted-repeat Transposable Elements (IMTEs) that have been recently discovered in the rice and *C.elegans* genomes. These are short recurring motifs of about 400 base pairs flanked by 15 base pairs inverted repeats. MITEs are

too small to encode any protein. The mechanism of how they copy themselves and move around in the genome is still uncertain.

1.1.8.2 Types of different transposon systems used as mutagenesis tools

1.1.8.2.1 P elements

The P elements were first cloned in 1982 from *Drosophila melanogaster* (38), as the genetic cause of hybrid dysgenesis in *Drosophila* (39,40). Around 30-50 copies of P elements were found to be well dispersed throughout the major chromosome arms in the fly genome. The full length of these autonomous elements is 2.9 kb with two 31-bp inverted terminal repeats (38,41). Due to their alternate splicing structure, the P elements transpose only in germ line cells. There are three exons and three introns in the operon of P elements. Introns 1 and 2 are spliced out in somatic cells, resulting in the expression of a transposase inhibitor, which binds to exon 3 to prevent splicing of intron 3. In contrast, all three introns are spliced out in germ line cells, leading to translation of the P element transpose. With a cut-and-past manner, P elements could function as a vehicle for insertional mutagenesis elements and are important tools in the study of *Drosophila* genetics. Like many transposons, P elements are non-functional outside their normal host range, indicating that host factors are involved in transposition (42).

1.1.8.2.2 *Tc1* transposable elements

The *Tc1* elements belong to a large transposon superfamily, the *Tc1/mariner* family (43). The first member of the family was discovered in 1983 as a repeat sequence in the genome *C. elegans* (44). Homologues of *Tc1* have been found in the genomes of *Drosophila mauritiana*, fungi, plants, fish, frogs and humans (45,46).*Tc1* elements, as well as other members of the *Tc1/mariner* family have been widely used in genetic studies in *C.elegans* and many other lower organisms.

1.1.8.2.3 *Sleeping Beauty* transposon

Although transposons have been widely used in the study of many lower organisms since their discovery, they are seldom used in mammalian system for mutagenesis studies due to the host factors required and low activity. *Sleeping Beauty* (SB) is first ‘active’ transposon system suitable for use in mammalian cells. SB is a *Tc1*-like transposon that was recovered

by comparative sequence reconstruction from teleost fish (47).The SB molecule is composed of a 1.6 kb DNA element flanked by 250 bp IR/DR terminal repeat sequences encoding a single protein, the SB transposase, which catalyses the mobilization of the SB transposon from one genomic locus to another (**Figure 1-3**). In the laboratory application, the SB transposase is normally separately expressed and the central region between the IR/DR repeats on the transposon is replaced with the gene of interest (**Figure 1-3**).

The synthetic SB was the first cut-and-paste transposon to show activity in many vertebrate genomes, including fish, mouse and human cells. It has also been shown to be active in both the somatic and germ line cells of mice (48,49). It was found that SB tends to insert into TA-rich regions and the sequence of 'ANNTANNT' is the preferred motif for SB integrations (49). Although SB can transpose to almost all locations within the genome, there is a 10 kb cargo capacity limit for SB. Also it has been found that SB has a strong propensity for *'*local hopping*'*. Over 70% of SB insertions are found to be within the same chromosome as the donor locus in mice (50). These factors have limited the application of SB as a genome-wide mutagenesis system. Nevertheless, SB has been successfully used as an insertional mutagen to drive cancer formation in mice which will be described later in detail.

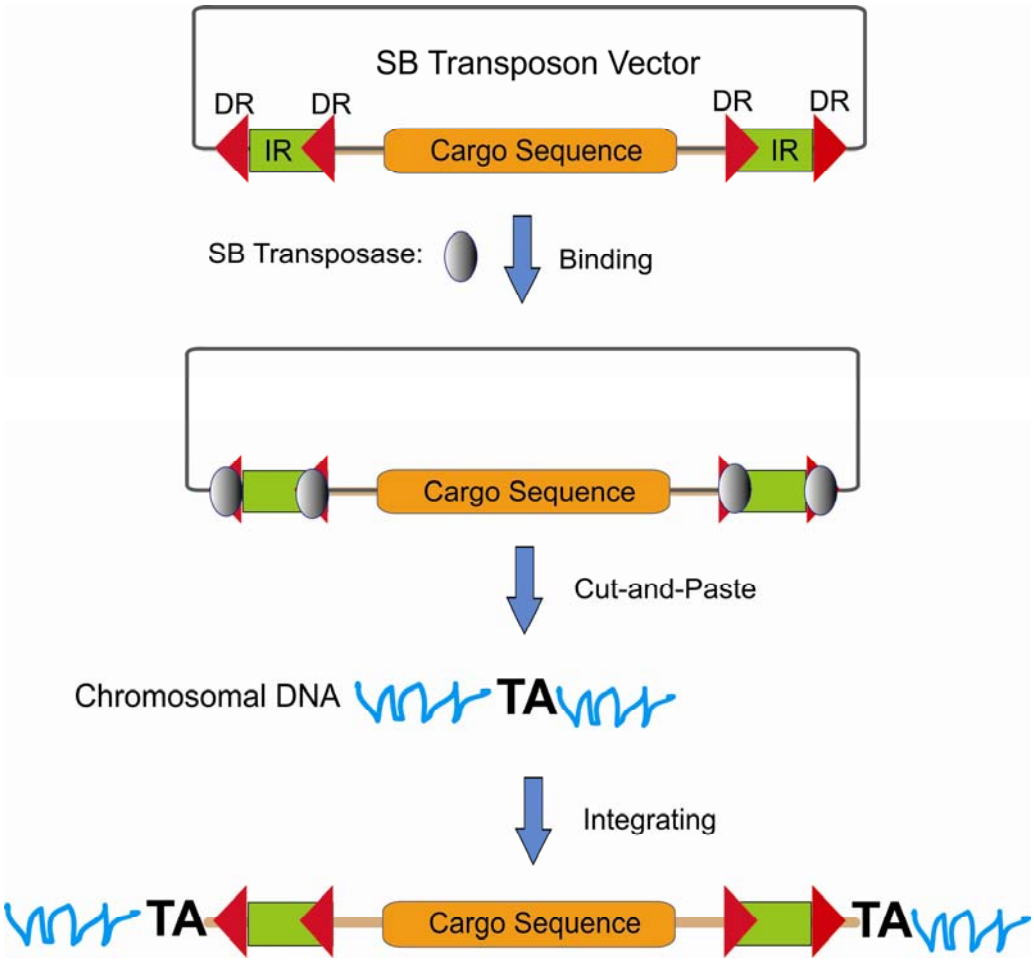

Figure 1-3. Molecular structure of the Sleeping Beauty transposon and the SB transposition system

In the laboratory applications the SB transposon and transposase are normally separated from each other. The transposon is engineered on a vector carrying the gene of interest or a cargo sequence in between the two IR/DR sequences. The transposase enzyme binds the two DR repeats (represented with red arrow heads) on each IR/DR terminal repeat sequence to carry out cut-and-paste function and integrate the SB transposon into genome with a preference towards TA-rich sites. IR – inverted repeat; DR – direct repeat. Figure is modified from Geurts *et al.* 2003 (51).

1.1.8.2.4 *piggyBac* transposon

The transposable element *piggyBac* (PB) was first discovered in the moth *Trichoplusia ni* which encodes a 594 amino acid transposase (52) flanked by two 13-bp inverted terminal sequences (ITR) (**Figure 1-4**). PB can carry transgenes up to 50 kb (unpublished results; Bradley laboratory, Wellcome Trust Sanger Institute, Cambridge, UK), which is much bigger than the maximum capacity of retroviral vectors or *Sleeping Beauty*. It has been shown that PB transposons insert into the tetranucleotide TTAA site, which is then duplicated after insertion (53). PB shows no obvious integration bias and local hopping has not been observed. In addition, unlike the *Sleeping Beauty* which leaves a TA footprint upon re-integration, PB is faithfully spliced from the donor site during mobilization. These facts suggest that PB has a unique advantage as a tool for mutagenesis in mammalian systems. More characters and applications about *piggyBac* transposon will be discussed in Chapter 3.

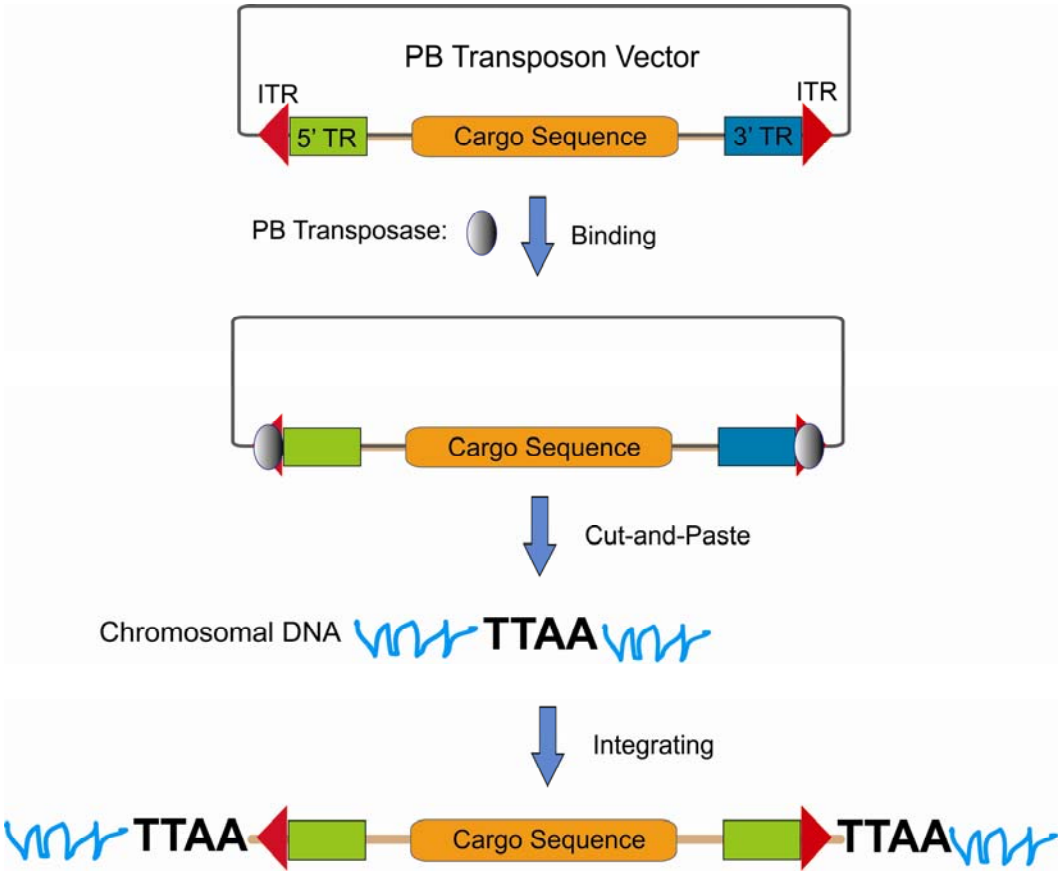

Figure 1-4. Molecular structure of the PB transposon and the PB transposition system

Similar to the SB transposition system, the PB transposon and transposase are also separated from each other in normal laboratory applications. The transposon contains two 13-bp exactly identical but inverted terminal repeats (ITR, represented with red arrow heads). The PB transposon also contains two general terminal repeats (5' TR and 3' TR) which could be used to identify the orientation during integration. Different from the SB transposon, the PB transposon leaves no footprint during excision and specifically integrates into TTAA site, which is duplicated upon integration.

1.1.8.3 Transposon mutagenesis in model organisms

1.1.8.3.1 Transposon mutagenesis in yeast

Studies with the budding yeast *Saccharomyces cerevisiae* have achieved some milestone discoveries in modern biology: the first eukaryote to be transformed by a plasmid, the first eukaryote for which gene-targeting became possible, the first eukaryotic genome to be completely sequenced (54). Although many innovative approaches have been developed to exploit the sequence data and yield information about this organism, the function of many of the > 6,000 genes in the *S.cerevisiae* genome still remains unknown, despite the sequencing of this organism being completed over 14 years ago. However, *S.cerevisiae* is still an important organism for studying gene expression and regulation, protein signalling and function of the entire genome.

Although targeted mutation of the yeast genome by homologous recombination is highly efficient in the budding yeast, insertional mutagenesis using transposons has also generated fruitful results in yeast gene functional studies. A pioneering study with yeast using transposon was performed in 1994 by Burns *et al.*(55), who constructed 2,800 yeast strains carrying translational fusions of *lacZ* to random genes using a mini-*Tn3::LEU2* transposon system and then localized the β-galactosidase fusion proteins to detect protein subcellular localization. Using immunofluorescence microscopy, distinct staining patterns were detected in 68% of the fusion proteins and 10% of the fusions were localized to discrete subcellular locations. Based on the frequency of cells expressing *lacZ* and assuming random integration, they estimated that around 74% of the ORFs in the *S.cerevisiae* genome are expressed under vegetative growth conditions. Another large-scale experiment in yeast that utilized transposon tagging was performed by Ross-Macdonald *et al.* in 1999 (56). In this study a modified *lacZ* trapping minitranposon, mTn, was used for genome-wide analysis of disruption phenotypes, gene expression and protein localization . A large collection of refined yeast mutants (over 11,000 strains) was produced, each carrying a single transposon inserted into the yeast genome. This collection has been used to determine disruption phenotypes under different growth conditions and identified over 300 previously non-annotated ORFs, constituting the largest functional analysis of the yeast genome ever undertaken.

1.1.8.3.2 Transposon mutagenesis in fruitfly

As a model organism, the *Drosophila* offers many advantages for post-genomics study, including husbandry, a relatively small genome size of which many disease genes are homologous to humans, and a range of genetic tools for manipulation of their genome. One of the key genetic tools is the P elements, a transposable element first developed as a transgenesis tool in *Drosophila* in 1982 (57). There are several key features which make the P element especially well suitable to functional studies in fly: the existence of M strains allows the creation of stocks containing only selected P elements; the transposase can easily be added or removed genetically; and P elements are highly mobile despite drastic modifications to their internal sequences.

One of the early uses for P elements for large-scale mutagenesis screening was to use naturally-occurring chromosomes containing many non-autonomous elements, however this quickly gave way to a more refined strategy using single engineered elements (58). Once injected into an embryo and incorporated into the genome, a P element construct can be easily mobilised using a separate source of transposase, creating many lines with a single element inserted randomly in the genome. Elements that transpose into genes may disrupt their function producing visible or lethal phenotypes. Mutagenesis efforts have culminated in the Gene Disruption Project – which was launched to disrupt every gene in the *Drosophila* genome with P elements (59,60). In 2004, the project has achieved single P element insertion associated with about 40% of the total genes in *Drosophila* genome (61). Whether the goal of obtaining insertions with full genomic coverage is achievable with P elements is a matter of debate, as P element has bias during integration into the genome. P elements prefer to transpose into the 5' region of the gene and have a bias toward a particular sequence motif (41). The P element preferentially inserts near existing P elements. In addition to this, there is a well documented preference for some genes, so called 'hot spots', and a distinct dislike of other genes, so called 'cold spots'. Therefore, some alternative transposon elements such as the *piggyBac* and *Minos* – a *Drosophila hydei* transposon are also used to complement the use of P elements in *Drosophila* mutagenesis studies.

1.1.8.3.3 Transposon screens in nematodes

The nematode *C. elegans* has a relatively small genome, only 20 times the size of *E. coli*. As a matter of fact, when analysing the genome of *C. elegans* it was discovered that

approximately 12 % of the *C. elegans* genome is derived from transposable elements (62,63). However, many of these sequences are fossil remnants that are no longer mobile in the genome. Among the transposons that are still active, the *Tc1* and *Tc3* are the most active and best characterized transposons in *C. elegans*. *Tc1* and *Tc3* are part of the *Tc1/mariner* superfamily of transposable elements which are named after its two best-studied members: *Tc1* and the related transposon *mariner* which were identified in *Drosophila* (45). It is probably the most widespread DNA transposon superfamily to occur in nature. Other active transposons in *C. elegans* include *Tc2, Tc4, Tc5, Tc7* and *CemaT1* elements. In addition, some transposons from other organisms have been shown to mobilize in the *C. elegans* genome, such as the Drosophila transposon *Mos1* (64).

Since the discovery of *Tc1*, the first transposon to be identified in the *C. elegans*, transposons have been used widely as genomic tools to drive *C. elegans* research while providing insight into some of the molecular mechanisms in genome evolution, surveillance and RNAi. Insertional mutagenesis with transposons generates mutant alleles that are tagged by the presence of a transposon. This molecular tag can subsequently be used to identify the mutant gene. *lin-12*, a nematode homeotic gene which controls certain binary decisions during development, was identified by means of *Tc1* transposon tagging (65). In a genetic background permissive for *Tc1* transposition, seven independent mutations were found to be associated with *Tc1* insertion events and all mutations were mapped to a single 2.9 kb restriction fragment. This DNA region contains three exons encoding 11 *lin-12* peptides homologous to a set of mammalian proteins that includes epidermal growth factor (EGF). Another similar application was the identification and molecular cloning of the muscle gene *unc-22* from *Tc-1* transposon tagging experiments in *C. elegans* (66). *Tc* elements can be used in combination with PCR to amplify the genomic sequence that flanks a mutagenic insertion and identifying the mutated gene without genetic mapping (67).

Using *Tc* elements as mutagens in *C. elegans* also has some drawbacks. First, the mobilization of *Tc* transposons is not restricted to a single class of elements in mutator strains. Second, there are several copies of each transposon in the genome which complicates the identification of the mutagenic insertion. Third, in the mutator strains that are used, transposition is removed from the mature mRNA by aberrant splicing (68). Spontaneous re-excision can generate mutagenic footprints that generate a stronger phenotype but can no longer be detected in a transposon tagging strategy. Nevertheless, these limitations can be partially circumvented by mobilizing the *Mos1* transposon in the germ line of *C. elegans* (64).

Although *Mos1* mutagenesis is 10 times less efficient than chemical mutagens, the cloning of mutated genes is easy to isolate since *Mos1* represents rare tags in the *C. elegans* genome. *Mos1* mobilization is also easy to control by conditional expression of the *Mos1* transposase. In addition to mutation of the genome by random insertion, transposons can also be used for targeted gene inactivation and screening by PCR to identify the gene of interest (69), as well as providing a means to engineer site-directed mutations into the *C. elegans* genome (70,71).

1.1.8.3.4 Transposon mutagenesis in mammalian systems

Although transposon systems have been routinely used for mutagenesis screening in lower organisms, it is only in recent years that they have been used as a genetic tool in mammalian systems; this is mainly due to the lack of activity of known transposon systems in mammalian cells. As a matter of fact, vertebrate and mammalian genomes are similar to invertebrate genomes as they also contain a large number of transposable elements, however, they are all in an inactive format due to a process called 'vertical inactivation' (72). The development of *Sleeping Beauty* (SB) from the fish genome by molecular reconstruction provided a valuable genetic tool for mutagenesis studies in mammalian systems (47). The SB transposon had relatively high activity in zebrafish, mouse and human cells, making it a powerful mutagen for generating somatic mutations. Because the SB transposon can carry a cargo sequence in between the terminal repeats, a gene trap cassette with a reporter gene can be loaded to provide loss-of-function mutagenesis to the host cell. In the meantime, the SB transposon can also carry an exogenous promoter to cause a gain-of-function.

The first screen in mammalian cell culture using the SB transposon system was carried out in a HEK293-derived cell line using a plasmid-based transposon delivery system to co-transfect two plasmids: one containing the SB transposon plasmid and the other containing the SB transposase (73). The transposon contained a CMV promoter which could drive over-expression of genes downstream of the insertion site. The screen identified a transposon that had inserted into the gene encoding the receptor-interacting protein kinase 1 (RIP1) and resulted in expression of a truncated version called 'PIP1', which lacked the N- terminal putative kinase domain and could constitutively activate NFκB in cultured cells (73).

In theory and practice, the SB transposon system is an efficient tool for gene discovery, however there are several problems associated with the application of the SB transposon system in mammalian cell culture systems. Firstly, although the SB transposon has been

found to be active in most mammalian cells, there is still some controversy as to whether SB is suitable for performing highly efficient mutagenesis screens. Secondly, the delivery of SB by plasmid co-transfection also restricts the efficiency of this system and makes downstream insertion sites analysis difficult. What is more, the transposon undergoes constitutive jumping, a phenomenon that occurs due to constitutive expression of the transposase in the cell, which further complicates downstream analysis of the candidate genes. These problems need to be solved before the SB transposon-based mutagenesis system can be routinely used for mutagenesis screens in cell cultures. There are, however, improved versions of SB such as the hyperactive version of *Sleeping Beauty* - SB100 which increases the transposition activity over a hundred-fold (74).

Recently the development of the *piggyBac* transposon as a genetic tool that is applicable to mouse and human cells provides a promising alternative for mammalian cell culture screens (75). *piggyBac* has been shown to be hundreds of times more active than the original SB, and at least 10 times more active than the hyperactive SB100 (76). A pioneering study using *piggyBac* system in a mosaic screen was carried out by Schuldiner *et al.* to identify genes responsible to regulate developmental axon pruning in γ mushroom body neurons (77). They first constructed an insertion library of over 2,000 genes using an engineered *piggyBac* mutator. Using this library they identified two cohesion subunits (SMC1 and SA) as being essential for axon pruning since mutations in these two genes disrupted axon pruning and caused neuroblast-proliferation defects.

1.2 Insertional mutagenesis screen for cancer gene identification

1.2.1 A summary of cancer

Cancer is a class of diseases that is responsible for about 13% of deaths each year according to the World Health Organization report (Retrieved 2011-01-08). Cancer affects people of all ages, although the risk for most types of cancer increases with age. Cancers are caused by genetic abnormalities in the genome, which result in a group of cells displaying uncontrolled growth, invasion and metastasis, which are three main properties of cancer cells. The genetic abnormalities found in cancer typically affect two general classes of genes: oncogenes and tumour suppressor genes. Oncogenes are a group of cancer-promoting genes typically activated or overexpressed in cancer cells, giving those cells new properties, such as

hyperactive growth and division/resistant to programmed cell death (uncontrolled growth), loss of respect for normal tissue boundaries (invasion), and the ability to become established in diverse tissue environments (metastasis). Tumour suppressor genes are a group of genes inactivated in cancer cells, resulting in the loss of normal functions in those cells, such as DNA replication or proof-reading, cell cycle control, orientation and adhesion within tissues, and interaction with protective cells of the immune system.

Cancer formation in cells is a multistep and complicated process. During cancer progression, each genetic change confers a specific cancer-related phenotype and eventually results in transformation of the cell and the formation of cancer (78). These genetic changes may include base-pair mutation, DNA fragment deletion, inversion or chromosome rearrangement. It is still unclear exactly how many genetic changes within a single cell are required for cancer and how many genes are involved in tumourigenesis in each cancer type. One recent review has estimated that 1 % of human genes have been shown to be directly involved in cancer formation (79). Therefore, identification of the oncogenes and tumour suppressors that collaborate in the formation and progression of cancer will undoubtedly help in the identification of crucial therapeutic targets.

Insertional mutagenesis based on retroviruses is a widely used approach for cancer gene discovery. Insertional mutagenesis is a mechanism of cancer initiation, as well as being an experimental tool. There are several oncogenic viruses that are implicated in human cancers including the human papilloma virus, the human T-cell lymphotropic virus (HTLV1), the hepatitis family of viruses, and the human immunodeficiency virus (HIV). In most cases, the viruse has integrated into the genome near a cancer related gene and caused ectopic gene over-expression to induce cancer. Insertional mutagenesis has also been directly proven in human patients who have received retroviral gene therapy for SCID-X1, a severe combined immunodeficiency disease. Some of the patients developed T-cell acute lymphoblastic leukaemia (T-ALL) after the retroviruses inserted upstream of *LMO2*, implicating this gene to be an oncogene in humans (80).

Research into identification of cancer genes by insertional mutagenesis usually involves three approaches: cell culture transformation assays, retrovirus-based mutagenesis and transposon-based mutagenesis. In recent years, research in this field has been markedly accelerated by the completion of human and model organism genome sequences. In particular, the development of high-throughput insertion site analysis and mapping technologies, aided by

computational tools, have made insertional mutagenesis a powerful tool for cancer gene discovery; it may be possible to profile the entire cancer genome in the near future.

1.2.2 Methods for cancer gene identification

1.2.2.1 Transformation assays for cancer gene identification

As early as in 1980s, the efforts to identify cancer genes were focused on screening for sequences isolated from human tumour cells capable of transforming NIH 3T3 fibroblasts *in vitro* (81,82). The assay involves the use of a retrovirus to deliver transforming genes into NIH 3T3 fibroblasts to yield a cell population capable of proliferation independently of both internal and external signals that normally restrain their growth. Traditionally the soft agar colony formation assay has been used to monitor cell transformation and anchorage-independent growth, with manual counting of proliferated cells after 3-4 weeks of cell growth. This method is still used as a standard protocol for the evaluation of a gene's ability to induce cell transformation. Although the transformation assay has been successful in the identification of oncogenes, the identification of tumour suppressors using *in vitro* screens is much more challenging owning to the difficulty of loss-of-function genetics. Other technologies such as RNAi have made it possible to silence the expression of a gene of interest, therefore the effects of tumour suppressor genes in transformation assay can be studied.

In the past, *in vitro* transformation assays have been used more as a method to valid the transformation ability of a gene rather than to screen for oncogenes in a cell line. This is largely because traditional retroviral mutagenesis screens have limitations in cell culture-based systems. With the development of highly active transposon systems such as *Sleeping Beauty* and *piggyBac*, it is now possible to directly screen for oncogenic insertions in an in vitro transformation assay, or develop a system to deliver the cancer related gene mutations conditionally for more precise validation using this assay.

1.2.2.2 Genes involved in transformation by retrovirus

In model organisms, transforming retroviruses have been valuable tools for cancer gene discovery. Retroviruses that function *in vivo* can be classified into acute transforming retroviruses and slow transforming retroviruses. Studies with the acute transforming

retrovirus Rous avian sarcoma virus, resulted in first discovery of a cancer gene, *v-src* about 30 years ago (83). Since then, several cancer genes have been discovered in a similar fashion, including *v-raf, v-myc, v-abl* and so on (84). Unlike the acute transforming retroviruses, which carry the genes required to transform their host cell within their genome, slow transforming retroviruses induce transformation by inserting into the host genome and are therefore amenable for genome-wide screens for cancer gene identification. For insertional mutagenesis screens performed using retroviruses, a gene harbouring insertions in multiple independent tumours is likely to be a cancer gene which is activated or disrupted by retroviral integration. The most commonly studied slow transforming retroviruses are the mouse mammary tumour viruses (MMTV) and the murine leukaemia viruses (MuLV). Many cancer genes such as *MYC, NF1, HOXA9 and EVI1* have been identified using these viruses (85). In addition, high-throughput retroviral insertional mutagenesis screens can also reveal networks of significantly collaboration between mutations and mutually exclusive interactions between cancer genes (86).

Nevertheless, the ability of retroviruses to act as cancer gene discovery tools is limited by their ability to effectively infect only a limited type of cells or tissues *in vivo*. The most efficient tissues for retrovirus mutagenesis are hematopoietic system and mammary gland. Besides these tissues, retroviruses have only limited success in cancer gene identification. In addition to this, the slow transforming retroviruses showed significant preference for inserting near the 5' end of actively transcribed genes in the host genome (87). This insertion sites bias might limit the amount of the genome accessible to retroviral mutagenesis.

1.2.2.3 DNA transposons – a new somatic mutagen for cancer gene identification

With the molecular reconstruction of *Sleeping Beauty*, a DNA transposon of *Tc1/mariner* transposon super family also active in the mouse soma, this novel genetic tool was soon tested in two experiments to drive tumour formation in mice (88,89). The SB transposons used in these two experiments are named *T2/Onc* or *T2/Onc2*, which were designed to mimic the provirual integration. The 5' part transposon contains a splicing accepter (SA) followed by a polyadenylation signal (polyA) sequence to disrupt the endogenous gene transcription for loss-of-function mutagenesis. The 3' part transposon contains a murine stem cell virus (MSCV) LTR and a splicing donor (SD) to over-express downstream gene coding sequences while integrating into open reading frame. The transposases used in these two experiments are different in activity, which might be the cause for the different results obtained in these

two studies. In the experiment of Dupuy *et al.*, a highly active transposase SB11 was knocked into the *Rosa26* locus to drive transposon mobilization. This resulted in tumour formation in the wild type background mice when crossed the transposon mice with the *T2/Onc2* transposon mice, among those the majority diseases were B- and T-cell lymphoma, by the age up to 114 days. In contrast, Collier *et al.* used a less active transposase SB10 in their experiment. This design at first did not generate tumour formation in wild type background mice. However, mobilization of the transposon did accelerate tumour formation in mice deficient for the tumour suppressor $p19^{Arf}$. The tumour spectrum in this experiment, consistent with the previously reported tumour spectrum for $p19^{Arf-/-}$, was mainly sarcomas. Therefore, these two experiments have set up the milestone for cancer gene discovery using a transposon based mutagenesis system.

In practice, the development of transposon systems could allow mice tumour studies to be designed in such a way that the transposase enzyme is specifically expressed from a tissue-specific promoter so that the mutagenesis could be studied in certain cell types. This strategy has given the transposon system having obvious advantage over the retroviruses in cancer study. One of the potential drawbacks for SB transposon system is the local hopping, as in previous studies over half of the insertion sites were mapped on the same donor chromosome. New transposon systems such as *piggyBac* and *TcBuster* were recently developed showing higher activity than SB and no detectable *'*local-hopping*'* effect. These new transposon systems are now being tested in different labs around the world under different genetic backgrounds for modelling specific diseases.

1.3 The experimental mouse as a model organism

1.3.1 The mouse and human genome

The laboratory mouse *Mus Musculus* has a genome of 3.4×10^9 base pairs (NCBI m37.1, July 2007), which is very similar to the genome size of a human (3.2×10^9 bases, NCBI 36.2, Sept 2006). Mouse and human diverged from a common ancestor about 65 million years ago and their genomes are highly conserved. 99% of human genes are represented by an identifiable mouse homologue, and 80% of mouse genes have a single human orthologue. More than 90% of the mouse and human genomes can be clustered into chromosomal segments of conserved synteny, reflecting the conservation of gene organization (35). Based on cDNA and comparative genomics study, both mouse and human have about 22,000

known genes (www.ensembl.org). All these data indicate that both mouse and human share a very similar genetic background with each other.

1.3.2 Mouse as a model organism

The laboratory mouse served as a model organism for studying human diseases and biological processes for many years. Besides its small size and relatively short generation time, mice are quite similar to humans in anatomy, physiology and genetic background. For a long time, research was limited to a few visible spontaneous mutations such as *agouti*, *reeler* and *obese* (8). Work on these spontaneous mutations has provided important insights into the molecular mechanisms of the relevant human diseases. However, spontaneous mutations in mice happen very rarely and do not provide enough mutations for functional studies. Many different methods have been developed to generate mutants in mice at a higher rate, such as chemical mutagens, X-ray irradiation and retrovirus mutagenesis.

1.3.3 Mouse embryonic stem cell as a genetic tool

The widespread use of the mouse for modern biomedical research is largely due to the isolation of mouse embryonic stem cells (ES Cell). ES cells are derived from mouse blastocysts by Evans and Kaufman in 1981(90). They are pluripotent cells that can derive into cells of three germ layers by *in vitro* culturing (ectoderm, endoderm and mesoderm). More importantly, ES cells can transmit through the mouse germ line when reintroduced into mouse blastocysts (91), this property allows genetic modified ES cells to derive into a mouse line for functional study. Another important advance in ES cell technology is the development of homologous recombination protocol in ES cells (92-95), which could allow precise engineering of loss- or gain-of-function mutations in the mouse genome through manipulation of ES cells, and then the cell line can be bred into mutant mice. This technique, together with the gene-trap mutagenesis which is able to randomly mutagenize the mouse genes in a large-scale and cost efficient manner, offers the possibility to disrupt every gene in the mouse genome for loss-of-function study. Two international consortiums was set up by this effort: The International Knockout Mouse Project, or so called KOMP (8) and the European Mouse Genome Mutagenesis Program, or so called EUCOMM (7).

1.3.4 Strategies used in ES cells for mouse genetics study

1.3.4.1 Generation of genetically modified mice by homologous recombination

That ES cells have become a key tool for mouse genetics is largely due to the development of the method to precisely modify the mouse genome by homologous recombination. Interestingly, laboratory mice were the first multi-cellular organisms in which artificial homologous recombination become possible. The targeting strategy is relatively simple and straightforward. Two homologous arms are used to flank a genetically engineered cassette in the targeting construct. The length of the homologous arms is usually between 3-5 kb, although experiments have shown that the arm can be less than 1.5 kb in length to allow successful targeting. The central cassette contains a selection marker (neomycin, puromycin or blasticidin are normally used), allowing ES cells colonies incorporating the targeting cassette to be identified by drug selection. The central cassette may also contains other genetic elements, depending on the experiment purpose. After introducing the targeting construct into ES cells followed by drug selection, correctly targeted ES cell clones are then be identified using specific techniques such as long range PCR or southern blot.

Homologous recombination is more frequently used to delete a selected gene in the genome. For this the 5' of the central cassette normally contains a SA – polyA gene-trapping cassette to disrupt gene transcription. Homologous recombination can also be useful to knock-in or express certain ectopic genes in the genome. Although the retrovirus can also be used for this application, it is sometimes required that the gene is knocked in under the endogenous promoter to be expressed at a physiological level. Gene coding regions can also be introduced into mouse genome using bacterial artificial chromosome (BAC), which has a much larger capacity and may also carry the regulatory elements for physiological expression.

Homologous recombination is an extremely important technique for studying human cancer. The most obvious advantage is that homologous recombination can generate null mutations for studying tumour suppressor genes such as *p53* and *RB*. Homologous recombination has also been used to precisely modify the genome mimicking human patient to generate mouse models for different type of cancers, for example by engineering point mutations, removing small coding fragments and gene knocking-in. As the knockout of some genes results in embryo lethality and the inability of the mice to breed, homologous recombination is often used together with the Cre-loxP system to generate a conditional knockout, which will be discussed later in this section.

1.3.4.2 Cre-loxP system for conditional mouse model

Site-specific recombination involving Cre-loxP is a type of genetic recombination in which DNA strand exchange takes place between segments possessing a certain limit of sequence homology. The most widely used site-specific recombination method in mouse is based on a P1 bacteriophase derived recombinase called Cre, which could specifically catalyze the recombination between two loxP sites. The loxP site is a 34 bp consensus sequence (ATAACTTCGTATA-GCATACAT-TATACGAAGTTAT), which includes two inverted 13 bp flanking sequences on both sides of an 8 bp core spacer sequence. The core spacer decides the orientation of the loxP site, but the flanking sequences are the actual binding site of Cre. The Cre-loxP system could act in mouse cells in three modes: inversion, deletion and translocation depending on the location and orientation of the loxP sites on the DNA.

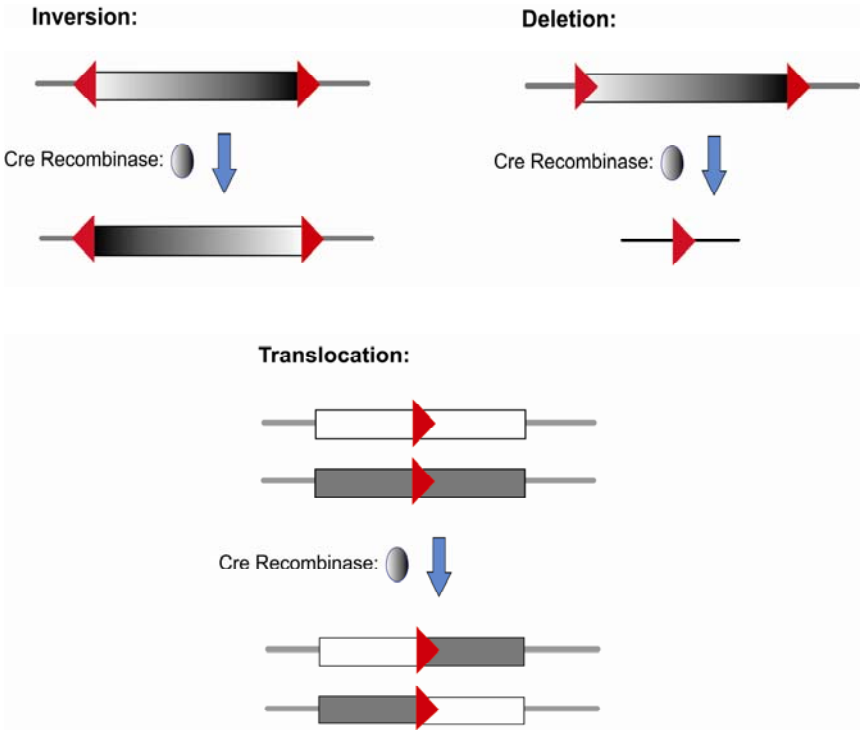

Figure 1-5. The Cre-loxP system and three applications in the eukaryote genome

Depending on the location and orientation of the loxP sites on the DNA, the Cre-loxP system could generate three applications in the eukaryotes genome: inversion (two loxP sites oriented face-to-face), deletion (two loxP sites oriented in the same direction) or translocation (two loxP sites located on different chromosomes).

Cre-mediated recombination is a very efficient system both in *in vitro* and *in vivo* studies. *In vitro*, Cre-mediated recombination is efficient enough to excise genomic regions as large as 400 kb (96). Cre recombinase is also very efficient *in vivo* and has been used to generate many mouse Cre transgenic lines. The 34 bp loxP site is short enough to be put into large introns without disrupting the transcription of the gene. It is also long enough to avoid the random occurrence of intrinsic loxP site in the mouse genome. With the completion of the sequencing of several major model organisms, searches reveal that no perfectly matched loxP site has even been found in any organisms other than the P1 bacteriophage.

One of the most common uses of the Cre/loxP is to generate conditional mouse knockout strains. This is essential for functional study of genes *in vivo*, especially for the genes that cause lethality at early stages when disrupted. The method for this usage is quite simple; two loxP sites in the same orientation are placed on both sides of the most important functional domain of the gene when designing the targeting construct. After gene targeting using homologous recombination, the ES cells and the mice carrying loxP sites in the genes of interest should be perfectly normal. When the animals are crossed to a Cre-expressing transgenic line, the progeny that carries both the Cre recombinase and the loxP sites will excise the loxP-flanked DNA fragment and result in gene knockout.

In mouse cancer studies, besides generating conditional knockouts for tumour suppress genes, the Cre/loxP conditional system could allow cancers to be modelled in a specific tissue using a tissue-specific Cre expression. The reason for this is because cancer is not only a gene-specific, but also a tissue specific disease which requires the exact genetic changes to take place in right tissues at the right time. Ubiquitous expression of an oncogene or deletion of tumour suppressor genes would result in complicated phenotype which may result in mouse death before cancer arises. To allow the right mutation take place the relevant location time, tissue specific promoters or inducible promoters are used to drive Cre expression to introduce mutations in a specific tissue at a specific time to induce cancer formation.

1.3.5 Mouse as a model for human cancer

For the reasons described above, it is not surprising that laboratory mice have been chosen as one of the primary model organisms for studying human cancer. There are many advantages for mice to be used so popular for cancer studies. The use of mouse models over comes the ethical issues involved in direct human studies on cancer; although cancer studies can be

carried out *in vitro*, it is essential to study the metabolic changes and tumour progression *in vivo* which is may not possible in human patients. In addition, mice are small with a short life cycle, which makes rapid, economical experiments become possible. Since mice genomes are very accessible to genetic manipulation, genetic modified mice could be generated to mimic human genetic changes in cancer and allow them to have a greater susceptibility to certain cancers. In the past, mouse models of cancer have produced fundamental insights into various aspects of cancer, including the identification of many oncogenes and tumour suppressors, understanding the biology of tumour-host cell interactions, the factors that influence cellular responsiveness to chemotherapy, as well as the role of stem cells in cancer development and progression.

There are many similarities between cancer characteristics in human and mice which have made cancer studies in mice possible. Both mice and humans exhibit low rates of cancer incidence rare in youth, and increased rates in old age. Many chemical and infectious agents that are carcinogenic in human are also carcinogenic in the mice. Importantly, several key genes and pathways lead to cancer in human are also functioning in mice, such as the tumour suppressor genes *p53* and retinoblastoma gene (RB). However, mice are not modelling perfect model of human cancer and there are also differences between mice and humans in cancer spectrum and progression. Mice tend to develop cancers in cells of mesenchymal tissues, resulting in lymphomas and sarcomas. In contrast, most cancers in humans tend to arise from epithelial cells and lead to carcinomas. Therefore in certain studies rats are used to substitute mice to model some cancer types. Another big difference is in the genetic pathways in mice and human. For instance, in human the telomeres decreases in size with age until the point where they can no longer function. However the telomerase in mice remains active in most cells, thereby helps cells to achieve immortality. Therefore, results obtained in mice model studies need to be treated with caution and sometimes analysis of human patents is needed for validation for oncogenes or tumour suppressors identified in mice.

1.4 Insertional sites analysis and mapping

Although insertional mutagenesis has proved its efficiency as a genetic tool for functional study and gene discovery, up-to-date analysis and mapping technologies are required to identify the retrovirus or transposon insertion sites and thereby to identify the gene of interest.

1.4.1 Isolation of the insertion sites

Several methods have been developed for isolation of insertion sites, including genomic DNA library screening, ligation-mediated PCR (LM-PCR) (97), inverse PCR (98), viral insertion site amplification (VISA), and single nucleotide polymorphism (SNP)-based mapping. Although these methods have been widely used and generated large numbers of insertion sites, there are limitations associated with each of these methods as an efficient technique for insertion site isolation.

1.4.1.1 Genomic DNA library screening

Genomic DNA library screening is the first method that has been introduced to isolate insertion sites in retrovirus induced mouse tumours. To perform the screen a DNA library of each tumour was first prepared in *E. coli* and each clone in the library was screened by colony lifting using viral long-terminal repeat (LTR) sequences as probe (97). Colonies harbouring retroviral insertions were subsequently sequenced to identify the insertion sites. Alternatively, an *E.coli* replication origin could be included in the insertional mutagen sequence and genomic DNA fragments are subject to self-ligation. Only the fragments containing insertional mutagen could be replicated in *E.coli* to form colonies. The efforts required for generating a DNA library are considerable, not to mention the subsequent screening work which is extremely time-consuming. The later application could be made more efficient since colonies harbouring insertion sites could be automatically generated, but the replication origin sequence might have negative effects on the host genome which could impair the screen efficiency. Nevertheless, both methods are depending on restriction digestion for preparing the library and the DNA fragments could vary greatly in size, which largely affects the efficiency of isolating the insertion site.

1.4.1.2 Inverse PCR

Inverse PCR is a polymerase chain reaction (PCR) based method for rapid amplification and identification of unknown sequences flanking transposable elements (98). The method has primers oriented in the reverse direction of the usual orientation. The template for the reverse primers is a restriction fragment that has been ligated with itself to form a circle. Normal PCR using these inverse primer pairs is then carried on and flanking genomic sequences read from PCR products by sequencing. Since the PCR can only be used to amplify regions of

limited size, this method is limited by the uneven distribution of the restriction sites along the genome therefore cannot provide a comprehensive amplification of all the insertion sites in the genome. Nevertheless, the inverse PCR has been used as a popular method for insertion sites identification for two decades from its discovery.

1.4.1.3 Linker-based PCR: vectorette PCR and splinkerette PCR

With the availability of reference genomes for human and other model organisms and the advent in sequencing technology, it is possible to amplify a small genomic DNA fragment flanking the mutagen insertion sites for mapping the insertion sites on the genome. Several linker-based technologies were developed for high-throughput insertional sites analysis. Among them vectorette PCR (99)and splinkerette PCR (100) are the most frequently used. Vectorette PCR can be highly sensitive, but its proneness to the amplification of contaminants by *'*end-repair priming*'*, which involves the free cohesive ends of unligated vectorettes annealing to each other to initiate priming. When this happens exponential unspecific PCR amplification may occur without the amplification of the specific PCR product.

Splinkerette PCR is a variant of linker-based PCR developed to overcome the problem of *'*end-repair priming*'* by using a splinkerette *'*hairpin loop*'*. The hairpin structure of the linker sequence is the key to the splinkerette PCR, which will prevent amplification of self-annealed linker sequence linker sequences annealed in a wrong orientation. A cohesive end is introduced to the linker for ligation with the genomic DNA. To perform splinkerette PCR the genomic DNA containing insertional mutagens was first randomly digested by restriction enzyme into DNA fragment and ligated with the linker sequence. The partial sequence tag together with flanking genomic DNA is then amplified by convention PCR using primers on the mutagen tag and the linker sequence and subjected for sequencing to identify the insertion sites. Although the splinkerette PCR still has the same problems with other PCR methods such as amplification bias and contaminations, it has become the most widely used technique for large-scale insertion sites amplification both for retroviral and transposon insertions.

1.4.2 Sequence mapping

In the linker-based PCR, after the insertion sites been isolated from the host genome and subject for sequencing the sequencing reads containing part of the genomic sequence flanking

the insertion sites needs to be mapped onto genomic sequence to identify the insertion site on the genome. Traditionally, this process can be done by using genome browsers such as Ensembl or the University of California Santa Cruz genome browser (UCSC), or a genome browser of a specific organism. In each of these browsers, users are first asked to submit a query sequence (normally one or several a time) for genome mapping. Alternatively, this process can be done by high-throughput genome query with Bioinformatics support. In either case, a mapping algorithms needs to be determined to achieve the best mapping efficiency. The original algorithm BLAST, which was developed for comparison of evolutionarily diverged sequences, is prohibitively slow in this application. As the high-throughput methods for insertional mutagenesis study often generate short sequences from the parallel sequencing platforms (Illumina-Solexa, SOLiD or 454 sequencing), several recently developed mapping algorithms for genomic sequence assembly from short sequencing reads maybe applied for genome mapping of these short insertional sequences.

1.4.2.1 MegaBLAST

MegaBLAST is similar to BLAST in that it splits a query sequence into non-overlapping fragments and searches for exact matches to the genome for regions with the highest identity. These perfect matches are then expanded to align the longest region of significant similarity. MegaBLAST uses a comprehensive algorithm that incorporates simplified gap and insertion/deletion penalties relative to BLAST and limits the number of alignments to be explored in extending the alignment beyond a perfect match seed. These alterations are justified because of the high levels of similarity expected between query and database sequences and the expectation that the alignment will not contain many mismatches or gaps. For sequences with greater than 97% identity, MegaBLAST is an order of magnitude faster than BLAST without any loss of alignment accuracy (101).

1.4.2.2 SSAHA

SSAHA (Sequence Search and Alignment by Hashing Algorithm) uses a different approach to take advantage of the high similarity expected between a query sequence and the genome. An index of all non-overlapping fragments of a set length (k) is created from the genome sequence and stored with the associated positions. The query sequence and its reverse complement are broken into all possible fragments of length k, including overlapping fragments, and compared with the genome index to identify exact matches. Matches are

sorted to find contiguous matching segments that are reported if they exceed a threshold, set by default to 2k. SSAHA is extremely fast, but due to the need to store the genome index and fragment locations, has relatively large memory requirements (102).

1.4.2.3 BLAT

BLAT (the BLAST-Like Alignment Tool) uses a multi-stage algorithm which searches for regions of similarity, aligns those regions, aggregates aligned regions in close proximity, and adjusts the boundaries of aligned regions to correspond with canonical splice sites. The initial search stage operates in a manner very similar to SSAHA. The genome database is broken into non-overlapping fragments of length k, then all k-length fragments of the query sequence and its reverse complement are associated with matching locations in the genome. The matches are sorted and grouped by proximity and those regions of the genome with a minimum of 2k contiguous matches are aligned with the query sequence. The alignment stage extends matching regions as far as possible, merges overlapping matches, links matches that fall in order on the genome into a single alignment, and fills in regions of the alignment corresponding to gaps of identical length in the query and genome sequences. Positions of gaps in the alignment, which may correspond to introns, are matched to the consensus splice site GT/AG wherever possible (103).

1.5 My PhD project overview

During my PhD studies, I worked on a series of projects to employ transposons as a high-throughput genetic tool for insertional mutagenesis study, and have applied these methods in *in vitro* and *in vivo* mutagenesis screen for gene discovery (**Figure 1-6**). To begin with, I did a rotation project in Dr. Bradley's lab and analyzed hundreds of insertional site data generated from *Sleeping Beauty* and *piggyBac* transposons mobilized from the mouse *Hprt* locus. This work resulted in a co-author publication which provides the first direct comparison of the insertional sites data from *Sleeping Beauty* and *piggyBac* transposons (76) (Appendix C). While analyzing hundreds of these sequencing reads manually and mapping them onto mouse genome to identify the genes involved, I was inspired by the idea to write a bioinformatics programme for automated analyzing insertional mutagenesis data. This resulted in the publication of *iMapper* in 2008– a freely accessible web application for

automated high-throughput analysing and mapping of the insertional mutagenesis sequencing reads (104) (Appendix D).

Figure 1-6. An outline of my PhD projects

Following my interest in transposon mediated insertional mutagenesis, I joined Dr. Adams' lab to work on an *in vitro* and an *in vivo* transposon screening project. The *in vitro* project was aimed at generating a high-efficient inducible mutagenesis system based on the *piggyBac* transposon for cell culture mutagenesis screen. The *in vivo* projects were to generate a *Tel-AML1* knockin mouse model for human acute lymphoblastic leukaemia (cALL) and a *Brd4-NUT* knockin mouse model for human midline carcinoma.

Up to the submission of my thesis, I have successfully generated an *in vitro* transposon mutagenesis system termed *'*Slingshot*'*, which could function as a stable integration in the cell genome and could be activated by tamoxifen induction for gene discovery in proof of function screens. This project in the meantime resulted in a recent publication in *Nucleic Acid Research* (105) (Appendix E). The *Tel-AML1* mouse project was a collaboration with my colleague Louise van der Weyden and Brian Huntly in MRC-CIMR Cambridge. In this project I was mainly responsible for generating the mouse model and carrying out experiments for validating this model. Since the mouse has successfully developed pro-B cell leukaemia, I have included some part of the *in vivo* analysis work from my collaborators to

prove that this mouse model has been successful for modelling cALL in human. I have carried out experiments independently to study the *Brd4-NUT* mouse model – a conditional knockin model for studying solid tumour in human. The mouse model generated promising phenotype in the ES cell that caused the cell proliferation to be completely blocked at G2-M stage with *Brd4-NUT* expression. However, although extensive attempts have been made to develop a germ line transmission for this mouse model, all have thus far ended in failure (more information will be provided in the *Brd4-NUT* chapter). Therefore I was unable to continue my work on the *Brd4-NUT* mouse for tumour study. To prove that the *Brd4-NUT* model could be a functional working model for the intended purpose, I also included some *in vitro* data for this part of experiments in my thesis.