Transposon-mediated Insertional Mutagenesis in Gene Discovery and Cancer

This dissertation is submitted for the degree of Doctor of Philosophy

By Jun Kong

Team 113, Experimental Caner Genetics Group

The Wellcome Trust Sanger Institute

Wellcome Trust Genome Campus

Hinxton, Cambridge CB10 1SA

Darwin College

University of Cambridge

Silver Street

Cambridge CB3 9EU

DECLARATION

I hereby declare that this dissertation is the results of my own work and includes nothing which is the outcome of work done in collaboration, except the tumour watch and disease profiling studies for the *Tel-AML1* mouse project in Chapter 4. This work was a collaboration with Dr. Louise van der Weyden, my colleague in the lab and Dr. Brian Huntly in MRC-CIMR Cambridge. I have included some of their *in vivo* work to prove that this mouse model has been successful for modelling cALL (childhood Acute Lymphoblastic Leukemia) in human, and is included in Figure 4-9.

None of the material presented herein has been submitted previously for the purpose of obtaining another degree. Material included in Chapter 2 has been published as Kong et al. (2008) Bioinformatics 24:2923-2925, material from Chapter 3 in Kong et al., (2010) Nucleic Acids Research 2010;38;18;e173 and Liang et al., (2009) Genesis 47:404-408. These publications are included as appendices. This dissertation does not exceed the word limit for the respective Degree Committee.

Jun Kong

September 20, 2010

ACKNOWLEDGEMENT

I would like to take this opportunity to extend my sincere thanks to all the people who generously supported me during my PhD study.

I would like to first express my deepest thanks to my supervisor Dr. David Adams. Since the first time I met him in the Sanger institute, he has been given me so much help and support all through the years. While I was working with the mice from the beginning, he opened up my mind to explore other areas based on my interest and under his guidance most efforts afterwards achieved successful results in the end. He was also quite patient with me when I made some mistakes with the experiment. It was a great pleasure and superior experience to have an opportunity to work with David during these four years.

I would like also to give my special thanks to Louise Van Der Weyden, Theodore Whipp, Catherine Wilson, Rebecca McIntyre, Chi Wong and all other members in Team 113 who have provided me enormous help and support during the time. Louise and Cathy showed me how to handle mice, Theo taught me how to grow cell culture and helped me prepare all the routine lab reagents and equipment for my experiment, Rebecca helped discuss and edit my thesis, Chi offered lots of general analysis and suggestions. Most importantly, their friendship has made the Team 113 such a wonderful place for me to carry out my PhD.

The advice and feedback from my second supervisor and the PhD committee members was precious. I am grateful to Dr. Mark Arends from Cambridge University Department of Pathology, Dr. Derek Stemple, Dr. Allan Bradley, Dr. Jos Jonkers from Netherland Cancer Institute (NKI) for their precious time and enlightening advice during my project.

In addition, I received quite a lot of help and advice from other members in Sanger and from the University. I would like to thank Dr. Brian Huntly from MRC-CIMR Cambridge for his advice and collaboration on the *Tel-AML1* project, for Mark Bushell who directed me with the immunoprecipitation experiment design, for Barry Rosen with the construct supply and many suggestions, for Jim Stalker with the informatics support on *iMapper*, for Bee Ling with help and advice on FACS analysis. Without these generous help I could have never achieved so much during my PhD years.

Finally, I would like to thank my parents, Jiefang Kong and Yulin Chen, as well as my fiancée Yi Yao. Their endless and unconditional love and support has accompanied me all through this most exciting period of my life.

SUMMARY

The advent of DNA sequencing has significantly accelerated biological research and discovery. Complete genomic sequences, together with approximately 20,000–25,000 annotated genes in the human genome, analysed through contemporary bioinformatic technology, must be functionally annotated by up-to-date biological methods to assign genes to pathways and functions. During my PhD, I combined *in vivo* and *in vitro* studies, together with the power of bioinformatics, to dissect gene functions under different contexts.

The fundamental basis of my PhD research is utilizing a system called transposon mutagenesis. Transposons are mobile genetic elements that represent a large portion of the repetitive sequences in the human genome. In *in vitro* cell culture studies, I developed a novel system called 'Slingshot' that is based on the *piggyBac* transposon system, which is capable of randomly mutagenizing the genome of many cell types in a 'gain-of-function' or 'loss-of-function' manner. Using this system, I performed drug resistance screens in the mouse embryonic stem cells. Subsequently, several drug transporter genes were identified in these screens that provide drug resistance to puromycin and the anti-cancer drug vincristine. I have also validated the efficiency of this transposon system using human somatic cell lines. In the *in vivo* studies, in collaboration with other colleagues, a *Tel-AML1* knockin mouse was generated to model childhood acute lymphoblastic leukaemia (cALL) that is characterized by a chromosomal translocation which results in the expression of a TEL-AML1 fusion protein. When crossed with the *Sleeping Beauty* transposon mice for cooperative mutations, some of these *Tel-AML1* mice derived the appropriate type of B cell leukaemia under tumour watch analysis. Another conditional *Brd4-NUT* mouse model for human midline carcinoma was generated using a similar knockin approach. Although this model did not transmit through the germ line for *in vivo* studies, *in vitro* experiments have revealed a strong cell growth arrest phenotype associated with the *Brd4-NUT* expression.

In addition, to provide better analysis of insertion sites for the transposon studies, I have developed an online web-based tool called '*iMapper'*, which analyzes large numbers of transposon integration sites from sequence reads and maps them to the appropriate Ensembl genome. I have successfully used this bioinformatics tool to analyze the insertion sites in sequence reads generated by my own experiments. This online resource is freely accessible and could facilitate the analyses of sequence reads and mapping of insertion sites for mutagenesis studies performed world-wide.

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF CONTENTS

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**).

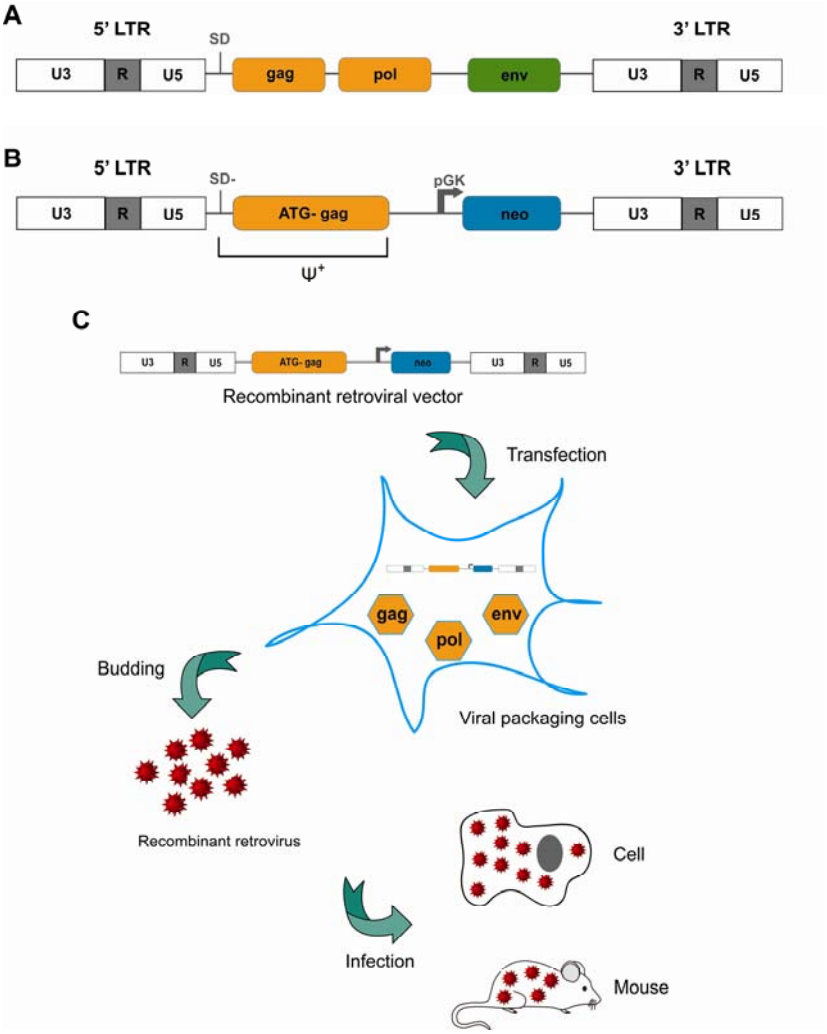


Figure 1-1. Retrovirus genome structure and recombinant retrovirus production

(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.

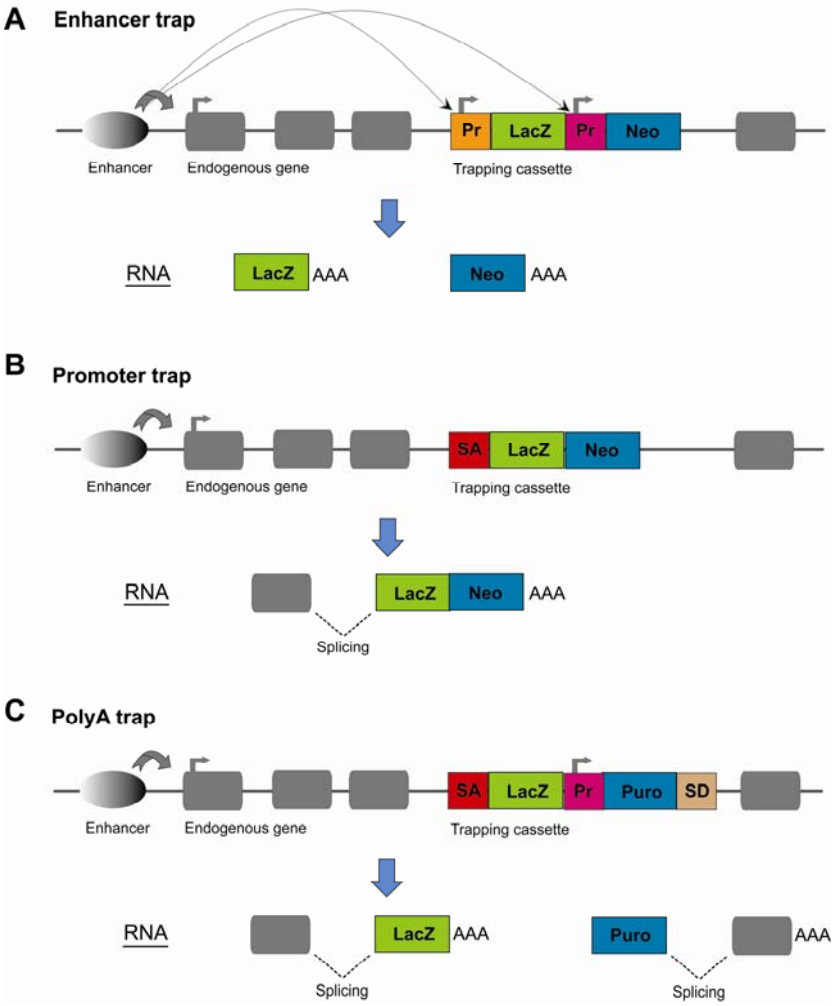


Figure 1-2. Schematic of three basic gene trap strategies

(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.

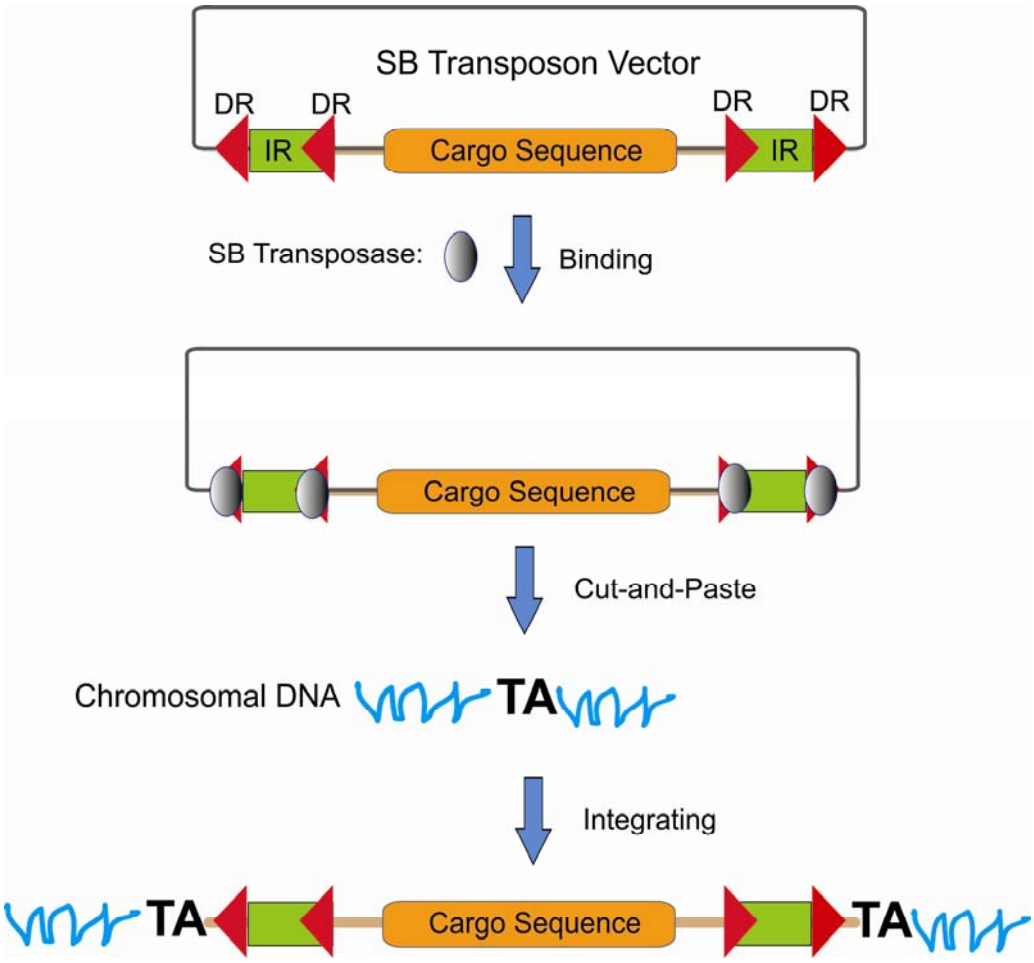


Figure 1-3. Molecular structure of the Sleeping Beauty transposon and the SB transposition system

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.

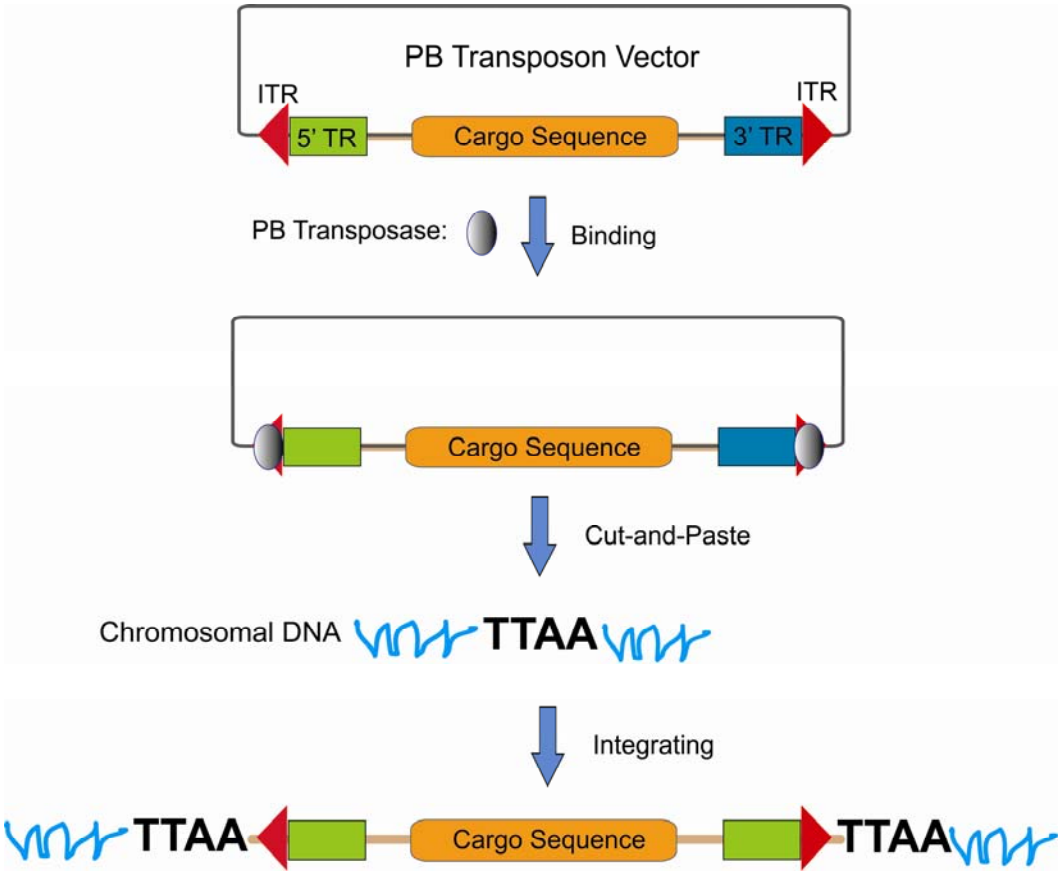


Figure 1-4. Molecular structure of the PB transposon and the PB transposition system

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.

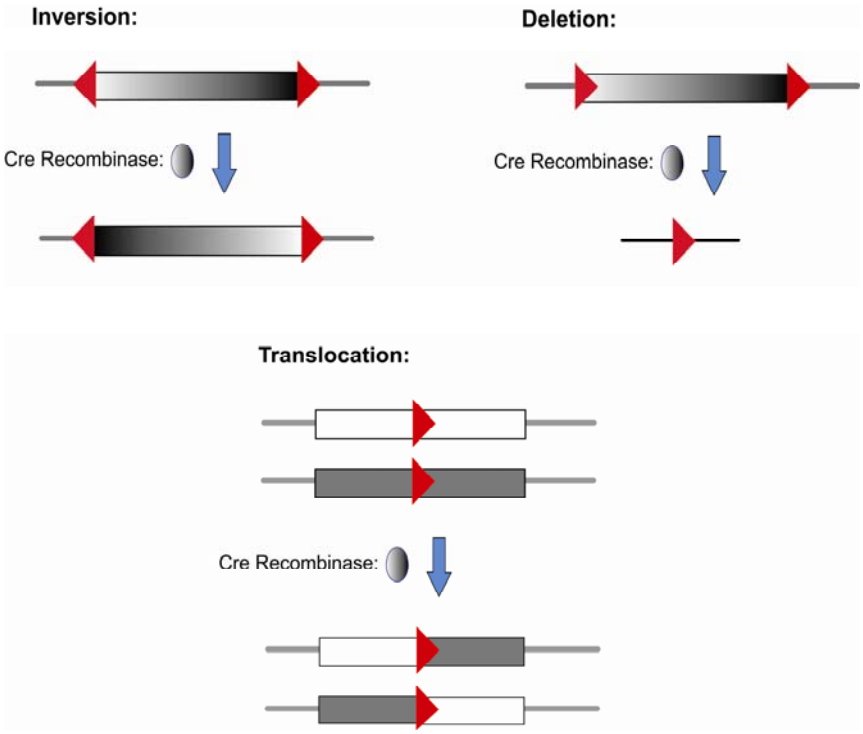


Figure 1-5. The Cre-loxP system and three applications in the eukaryote genome

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.

Chapter 2. *iMapper*: A web server for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes

2.1 Introduction

Large-scale insertional mutagenesis screening in mice or cultured cells is a rapid and efficient approach for gene discovery. Unlike other gene discovery approaches, the fact that insertional mutagens modify the genome directly, as opposed to overexpressing cDNAs from plasmids or viruses or knocking down genes by shRNA or siRNA transfection, means that they are capable of simultaneously informing us about the function of protein coding genes, regulatory regions, non-coding RNAs, miRNA or indeed any other element of the genome.

Retroviral-based insertional mutagenesis screens have been a valuable tool for the discovery of oncogenes and tumour supressors in mice (106) and also for gene discovery in cultured cells (107). More recently transposon-based approaches such as the use of the *Tc1*-family transposon *Sleeping Beauty* (88,89,108), the *Trichopulsia*-derived transposon *piggyBac* (109), and the *Tribolium castaneum*-derived *TcBuster* (110) have been developed increasing the repertoire of insertional mutagens available as gene discovery tools. To determine where in

the genome an insertional mutagen has inserted, the usual approach is to use a linker-based PCR method, such as vectorette or splinkerette (100). For any insertional mutagenesis screen to cover a significant proportion of the genome it is desirable to perform a screen using hundreds of mice and hundreds if not thousands of cell clones. Thus insertional mutagenesis screens may involve the generation and analysis of tens of thousands of DNA sequence reads from insertion sites. Although linker-based PCR methods are generally specific, non-specific PCR products, chimeric sequences and sequences derived from transposon concatemeric arrays can all represent contaminating sequences within pools of insertion site PCR products (**Figure 2-1**), thus without processing of sequence data the direct mapping of insertion site sequences to the genome may result in the identification of false-positive insertion sites. Therefore, insertion site sequences need to be processed prior to downstream mapping and analysis.

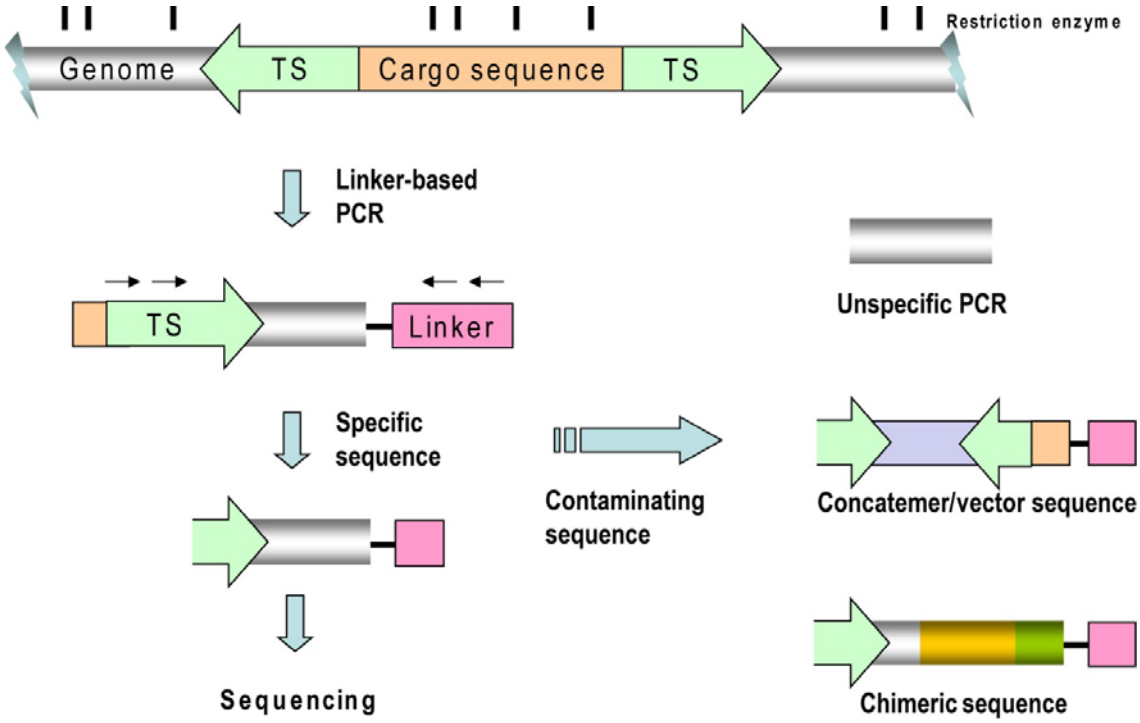


Figure 2-1. Schematic diagram of linker-based PCR procedure and contaminating sequences

To perform linker-based PCR, genomic DNA is first digested randomly with a frequent restriction cutter (average size of 300 – 500 bp). The genomic DNA fragments are first ligated with a linker sequence of various design and then subject to PCR amplification. The amplified DNA sequences containing insertional mutagens are subjected for sequencing and genome mapping. The contaminating sequences from linker-based PCR could be derived from non-specific PCR reactions, concatemer sequences from adjacent transposon arrays or chimeric sequences from inter-connection of genomic DNA during ligation step.

A major time-limiting step for insertional mutagenesis studies is the processing of tens of thousands of sequence reads in an accurate and efficient way, which includes: (1) identifying and eliminating the contaminating sequences (mainly concatemeric or chimeric sequences); (2) identifying the mutagen tag sequence in the sequencing reads to avoid any unspecific PCR products; (3) mapping the processed sequencing reads to the host genome to identify the insertion sites; and (4) obtaining a list of candidate genes overlapping with these insertion sites. Insertional mutagenesis screens are usually performed by experimental biologists who may not have the computing knowledge required to process large-scale sequencing reads, and therefore have to rely on collaborations with computational biologists who have the software and the computational skills to do large-scale genomic mapping. This has greatly limited the efficiency of insertional mutagenesis screens that have been carried out in laboratories lacking bioinformatic support. Although there are online genome browsers such as Ensembl or UCSC genome browser which enable the mapping individual genomic sequences, so far there has been no software or online web tools that could specifically facilitate the analysis and mapping of large numbers of sequencing reads generated from insertional mutagenesis screens. Therefore, developing a bioinformatics tool specifically for insertional mutagenesis sequence analysis could not only facilitate the analysis of insertional mutagenesis data for my PhD projects, but could also provide a solution for the community in better analyzing and processing their experiment results.

2.2 Aim and summary of the project

The aim of this project is to develop a powerful but simple to use bioinformatics tool for insertional mutagenesis dataset analysis. The main features of this tool should include:

1. Efficient and accurate processing of insertion site sequence data and analysis against genomes of many model organisms, including human, mouse, rat, zebrafish, *Drosophila*, and *Saccharomyces cerevisiae* genomes.

2. Output of annotated sequence reads with links to genome browsers so that insertion sites can be viewed in the context of gene structures and other genomic features.

3. Output of processed sequence data in FASTA and GFF file format to allow insertion site sequence data to be analyzed in any sequence analysis package and could be displayed as a DAS track against the Ensembl genome.

4. Output a graphical chromosome *KaryoView* showing insertion sites against an ideogram of each chromosome.

5. Since this tool should be open to public access, ideally this tool is to be developed as an online web-based tool. The interface should be friendly and the operation should be easy to use by scientists with limited computer knowledge.

Based on these expectations, a web-based server called *iMapper* (Insertional Mutagenesis Mapping and Analysis Tool) was developed for the efficient analysis of insertion site sequence reads against vertebrate and invertebrate Ensembl genomes. Taking linker-based PCR sequence reads as input, *iMapper* first scans the sequence to identify a tag sequence (TS) derived from the end of the insertional mutagen using a local sequence alignment (LSA) algorithm. *iMapper* then scans the downstream sequence for user defined contaminating sequences, then processes the sequences to identify the restriction site sequence used for linker ligation during the insertion site PCR, clips out the genomic sequence between the tag sequence and first restriction enzyme cutting site and presents this sequence to a rapid mapping algorithm called SSAHA (102). Insertion sites can then be navigated in Ensembl in the context of other genomic features such as gene structures. *iMapper* also generates FASTA and GFF files of the clipped sequence reads and provides a graphical overview of the mapped insertion sites against a *karyotype*. *iMapper* is designed for high-throughput applications and can efficiently process tens of thousands of DNA sequence reads in a short time.

2.3 Materials and Methods

2.3.1 Architecture of the program

The *iMapper* interface is web-based (**Figure 2-3**) to accept sequence information and other parameters defined by the users. The sequence analysis module within *iMapper* is developed using Perl and CGI script to accept and process information passed from the interface. The processed sequence is then passed to a SAHHA server to map the sequence to an Ensembl genome. Processed sequence information including the tag sequence, genomic sequence and mapping positions are then fed back to the user's computer to generate an output webpage with processed sequence information. The output webpage also generates links to run against the Ensembl applications to generate additional output information such as the Ensembl *ContigView* and generation of a chromosome *KaryoView* graph of the insertion sites.

2.3.2 Sequence Processing

The procedure used by the code for sequence processing is shown in **Figure 2-2**. Taking linker-based PCR sequence reads as input, *iMapper* first scans the sequence to identify a tag sequence (TS) derived from the end of the insertional mutagen using a local sequence alignment (LSA) algorithm. *iMapper* then scans the downstream sequence for user defined contaminating sequences such as concatemeric sequences or vector sequences from the end of insertional mutagens, then processes the sequences to identify the restriction site sequence used for linker ligation during the insertion site PCR, clips out the genomic sequence between the tag sequence and first restriction enzyme cutting site and presents this sequence to a rapid mapping algorithm called SSAHA (102). The annotated sequence is then displayed on the output webpage with links to other features of the insertion sequence including: (1) navigation in the Ensembl *ContigView* to view other genomic features such as gene structures; (2) FASTA and GFF files of the clipped sequence reads; (3) a graphical overview of the mapped insertion sites against a *KaryoView* picture generated by Ensembl.

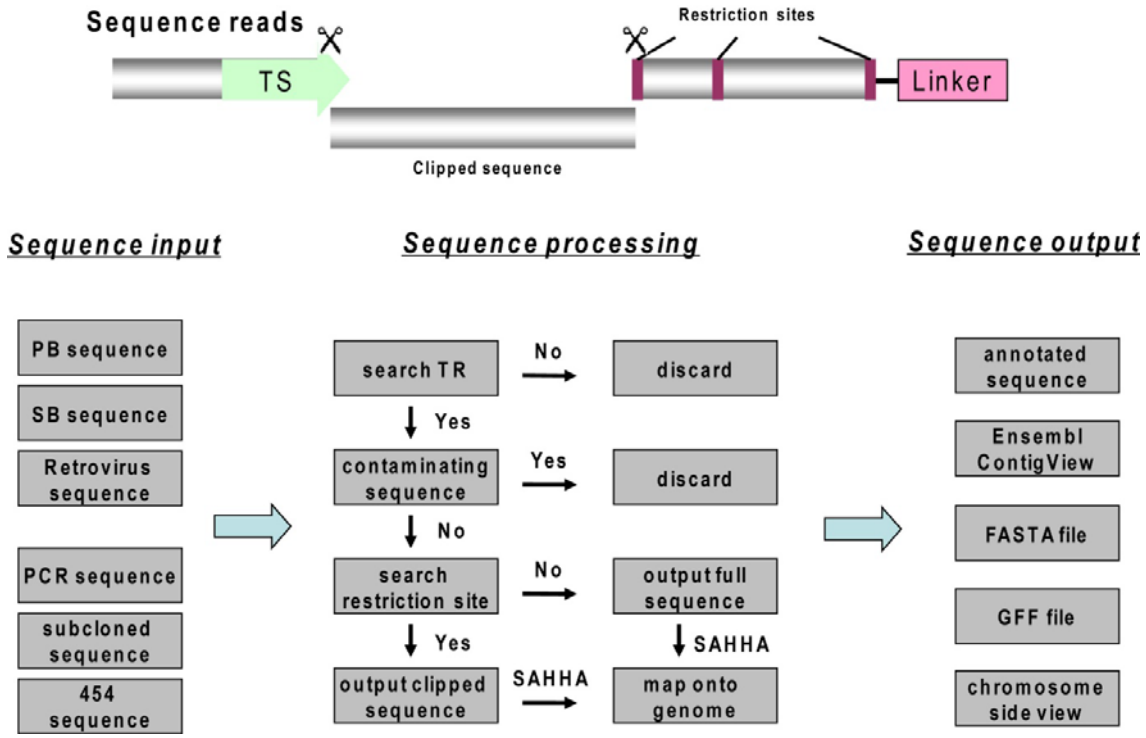


Figure 2-2. The workflow of *iMapper*

iMapper is a sequence mapping and analysis tool designed to process insertion site sequence read data generated using ligation mediated PCR methods such as Vectorette and Splinkerette. *iMapper* identifies a user-defined tag sequence and restriction enzyme site within a DNA sequence read and maps the intervening sequence to a user-defined Ensembl genome. The sequence input, processing and output formats of the software are shown.

2.3.3 Performance Test

To determine the optimal length of the mutagen tag sequence and optimal percentage threshold for sequence tag identification, a published dataset of 1920 *piggyBac* traces was chosen for analysis (109). The optimal length of the mutagen tag sequence was tested by fixing the percentage threshold to 80% and generating a graph of *'*Accuracy*'* vs. *'*Coverage*'* by increasing the tag sequence length for tag sequence identification using local sequence alignment algorithm. The optimal percentage threshold for sequence tag identification was tested by fixing the tag sequence length to 17 bp (PB tag sequence used: TATCTTTCTAGGGTTAA) and generating a graph of *'*Accuracy*'* vs. *'*Coverage*'* by increasing the percentage threshold for tag sequence identification. The value for *'*Accuracy*'* was determined by the presence of *piggyBac '*TTAA*'* signature sequence at the end of the tag sequences been identified, the value for *'*Coverage*'* was calculated by using the number of sequences containing tag sequences divided by the number of total sequences.

2.3.4 Chromosome *KaryoView* graph

To display the chromosome side view graph of mapped insertion sites a *Sleeping Beauty* dataset containing 4032 sequence reads generated from mouse tumours using shot gun cloning was used for analysis. After *iMapper* processing, users can click on the *'*Generate chromosome side view graph*'* link on the output page to enter the Ensembl *KaryoView* browser. In the next page, users click *'*continue*'* to enter a configure page: here users can define the format and display to generate the side view graph. Users may then click the *'*Finish*'* button and a *KaryoView* picture is generated. To generate a side view graph for multiple datasets, users can choose *'*Add more data*'* in the first setup page and then copy and paste the GFF file generated from another dataset to the text box before following the same procedure to display the multiple datasets on one *KaryoView* picture.

2.4 Results

2.4.1 Program interface

The front end of *iMapper* is a user-accessible web interface generated using Perl and CGI (**Figure 2-3**), allowing users to submit their sequence data to the *iMapper* server in FASTA format or plain text. There are also different sections and options on the interface page where

users can define *iMapper* working parameters. After filling out the form, users can press *'*Submit Query*'* to submit their sequence data to a back-end server and start mapping sequence reads using *iMapper*. More detailed descriptions on how to use *iMapper* can be found in the online help page.

Figure 2-3. The interface of *iMapper*

The interface of *iMapper* is web-based with advanced options (smaller window on the left) expanded to illustrate the parameters that are adjustable within the web tool. More instructions on the usage of *iMapper* interface are described below.

2.4.1.1 Basic setup window

The basic setup window allows users to readily use *iMapper* to process their sequencing reads by submitting their sequence data and define a few simple working parameters. Sequence data is imported into *iMapper* in FASTA format. Sequence data in this format can either be pasted into a text box provided or imported using the file upload option. After sequence input a user can define the species against which they would like their insertion site data analysed from Human, mouse, rat, zebrafish, *Drosophila*, and *Saccharomyces cerevisiae*. The orientation of the tag in the sequence can then be chosen, and the output option selected. Selecting *'*output good sequences*'* will exclude those sequences that do not contain the tag sequence, sequences that do not map to the genome and sequences that are identified to contain contaminating sequences. The sequence of the mutagen *'*tag*'* sequence can then be specified, or a pre-validated tag sequence can be selected from the drop down menu. We provide tag sequences for *Sleeping Beauty* and *piggyBac* transposons, and for the U3 long terminal repeat (U3LTR) of the MuLV retrovirus. The sequence of the restriction site can then be specified. Alternatively the sequence of the linker could be entered. At this point the user has defined the boundaries of the sequence which will be mapped to the genome as the sequence between the tag and the restriction site or linker. These basic setups will allow users to work with their sequences submitted to *iMapper*.

2.4.1.2 Advanced options

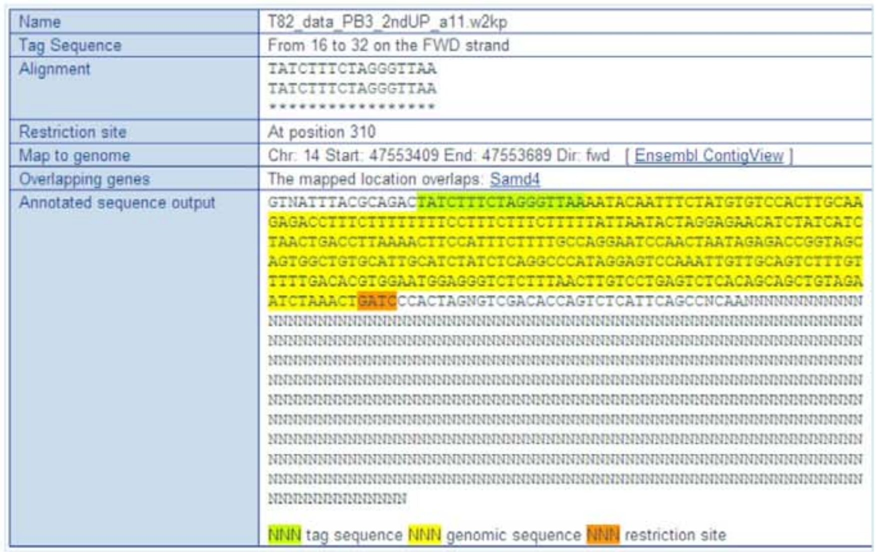
In addition, advanced options for *iMapper* can be specified in the unfolded window *'*Show Advanced Options*'*. These include the tag alignment parameters and the sequence of contaminating sequences in a tabular format. It is also possible to specify the parameters used by the SSAHA algorithm for matching the genomic sequence between the tag and the restriction site or linker, to the genome. These parameters can make a dramatic difference for SSAHA mapping especially when for analyzing short sequences such as those generated by 454/Roche sequencing. Finally, it is possible to specify the criteria for *'*gene overlaps*'*. By specifying *'*gene overlaps*'* it is possible to vary the spatial criteria for defining what constitutes a transposon insertion event in or near to a gene. For example it may be desirable to identify insertion events that mutate in or upstream of a gene, but not downstream.

2.4.2 Program Output

2.4.2.1 Output page for annotated sequence

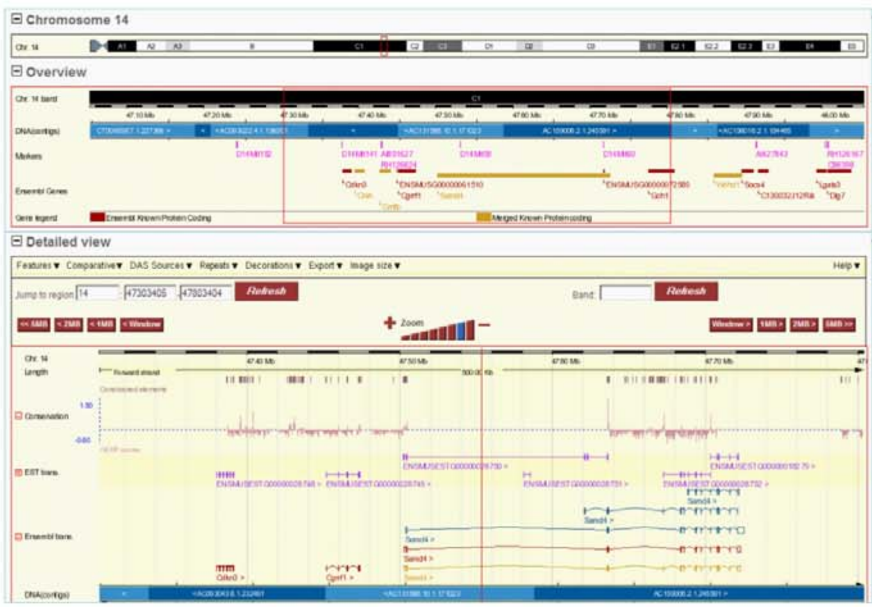
After submitting the sequence data, *iMapper* generates a html-based output of the analyzed sequence data in tabular format (**Figure 2-4 A**). Sequences such as the tag sequence, the restriction site and genomic sequence in between are highlighted in this view. If a unique genomic mapping is identified, links to the Ensembl *ContigView* are provided to view the insertion site against Ensembl genome structures (**Figure 2-4 B**). In addition, *iMapper* also identifies if the insertion site overlaps with any Ensembl known genes (within 10kb up or downstream by default), and if so a link to gene pages is also provided. *iMapper* uses a real-time display algorithm and annotated sequences are streamed to the browser as they are generated, allowing users to view their results before whole sequences are processed. After analyzed all the input sequence data, *iMapper* summarises the results to show the total number sequences that have been analyzed and then number of sequences in each category, including how many reads contain the tag sequence, contaminating sequence and how many can be mapped to genome or overlap with any Ensembl known genes. On the bottom of the output page, *iMapper* provides three useful links to display additional information. First is a link to generate clipped genomic sequences in FASTA format for all the processed reads containing tag sequences. The second link is for generating a GFF file for all the sequence reads that can be mapped onto a genome, including genomic locations, length, whether the sequence overlaps with a gene and the gene name. The gene feature list is recognised by Excel which can be used for further data processing and sorting. The third link will generate a chromosome side view graph for all the mapped sequences using Ensembl *KaryoView* browser (**Figure 2-4 C**).

A



Name	T82_data_PB3_2ndUP_a11.w2kp
Tag Sequence	From 16 to 32 on the FWD strand
Alignment	TATCTTTCTAGGGTTAA TATCTTTCTAGGGTTAA *****************
Restriction site	At position 310
Map to genome	Chr: 14 Start: 47553409 End: 47553689 Dir: fwd [Ensembl ContigView]
Overlapping genes	The mapped location overlaps: Samd4
Annotated sequence output	GTNATTTACGCAGACTATCTTTCTAGGGTTAAAATACAATTTCTATGTGTCCACTTGCAA GAGACCTTTCTTTTTTTTCCTTTCTTTCTTTTTATTAATACTAGGAGAACATCTATCATC TAACTGACCTTAAAACTTCCATTTCTTTTGCCAGGAATCCAACTAATAGAGACCGGTAGC AGTGGCTGTGCATTGCATCTATCTCAGGCCCATAGGAGTCCAAATTGTTGCAGTCTTTGT TTTTGACACGTGGAATGGAGGGTCTCTTTAACTTGTCCTGAGTCTCACAGCAGCTGTAGA ATCTAAACTGATCCCACTAGNGTCGACACCAGTCTCATTCAGCCNCAANNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNN NNN tag sequence NNN genomic sequence NNN restriction site

B



C

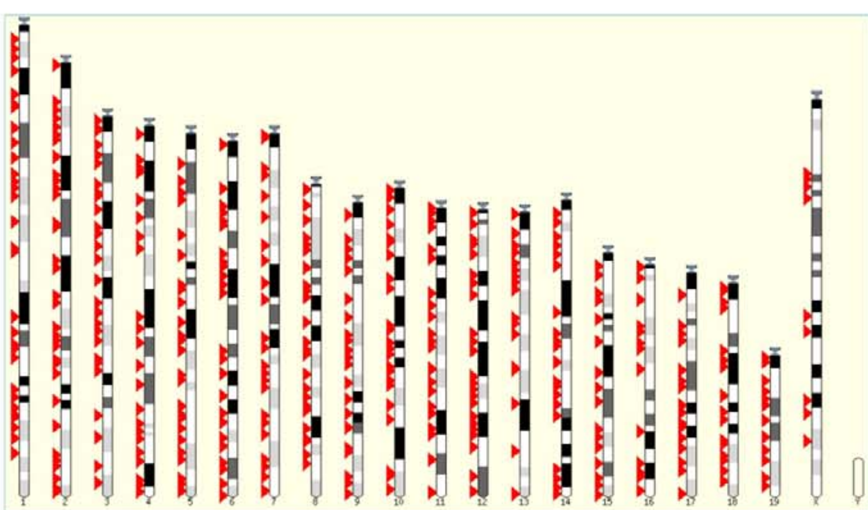


Figure 2-4. The output of *iMapper*

(A) Example of the tabular output for one sequence read processed by *iMapper*. The sequence is annotated in different colours representing tag sequence (green), clipped genomic sequence (yellow) and restriction site (brown). Other information such as the position of the tag sequence and restriction site, the location on the genome where this sequence read maps to, and the links to Ensembl *ContigView* and gene information page are all provided. (B) Detailed view of an insertion site in relation to gene structure in Ensembl *ContigView*. The red bar represents the position of the insertion site. (C) *KaryoView* picture of insertion sites generated by *iMapper* via Ensembl *KaryoView*.

2.4.2.2 FASTA and GFF formats

When the analysis run is complete, links to a FASTA file of the processed traces and a GFF file of the data are provided on the bottom of the output page. The link to the FASTA file generates clipped genomic sequences between the tag sequence and the first restriction site, or the full length sequence after the tag sequence if no restriction site can be identified. These processed genomic sequences are particularly useful for analysis by other sequence processing packages or mapping tools. The second link generates a GFF file for all the sequence reads which can be mapped onto an Ensembl genome. GFF is a file format used for describing genes and other features associated with DNA sequences such as the genomic location, length, whether the sequence overlaps with a gene and the gene names. More information on GFF files can be found at http://www.sanger.ac.uk/resources/software/gff/. This gene feature list is recognised by Microsoft Excel which can be used for further data processing and sorting. Furthermore, the GFF file format can be uploaded and displayed as a DAS track against an Ensembl genome.

2.4.2.3 KaryoView graph

To obtain a global overview of the sequence data *iMapper* has a link to an Ensembl *KaryoView* to generate a side view graph of the data against a chromosomal ideogram. This graph gives a clear and direct view of the distribution of insertion sites on different chromosomes in the genome. For instructions on how to generate a chromosome side view graph such as in **Figure 2-4 C** please refer to section 2.2 Materials and Methods.

2.4.3 Performance of *iMapper*

2.4.3.1 Determining the optimal working parameters for iMapper

iMapper uses the Smith-Waterman local sequence alignment (111) algorithm to identify the mutagen tag sequences. The specificity of tag identification depends on the length of the tag sequence entered, and the pre-defined thresholds specified for sequence tag identification including the percentage alignment threshold, gap penalty, match and mis-match score. Longer tag sequence inputs, higher alignment percentages and more stringent gap and mis-match scores will result in more accurate tag sequence matching. We have tested the optimal tag sequence length and percentage threshold using a dataset of 1920 *piggyBac* insertion site sequence reads (109). Because *piggyBac* integrations invariably occur at *'*TTAA*'* sites a

precisely identified tag sequence will always be proceeded by the sequence TTAA. As shown in **Figure 2-5 A** the minimal advisable tag sequence length is 15 bp. Next we determined the optimal percentage threshold for sequence tag identification to be used as the default, and determined this to be 80 % (**Figure 2-5 B**). Finally we optimized the SSAHA sequence parameters to be used as the default and found that for sequences from splinkerette PCR reactions that contain genomic junction fragments of, on average, 200 bp in length, the optimal SSAHA score is 35. This score should be ideal for insertion site sequences generated by capillary read sequencing but may need to be lowered to 20 for shorter reads such as those generated by 454 sequencing. It is advisable to optimize the SSAHA mapping score for each dataset and select a score that generates the highest number of uniquely mapped reads. This is important because the default mapping parameters used by *iMapper* are stringent and will return only those reads that map to unique, unambiguous genomic locations.

Figure 2-5. Performance of *iMapper: analysis of* 1920 *piggyBac* traces

(A) Determination of the optimal length of the mutagen tag sequence. (B) Determination of the optimal percentage threshold for sequence tag identification. The *'*Accuracy*'* is determined by the presence of the *piggyBac* *'*TTAA*'* signature sequence at the end of the tag sequence that has been identified, the *'*Coverage*'* is calculated by using the number of sequences containing tag sequences divided by the total number of sequences. (C) The data analyzed by *iMapper* using optimized parameters (analyzed using tag sequence TATCTTTCTAGGGTTAA (17 bp), percentage threshold = 80 %).

2.4.3.2 Timing and capacity

iMapper is a highly efficient program for the annotation of large numbers of insertion sites when compared with manual annotation. Generally, it takes the program less than half a second to process each sequence read and another second to map the sequences using the SSAHA server. Parameters that affect *iMapper's* processing time include the input tag sequence length, whether it is used to search for tag sequences in a defined orientation or whether to exclude contaminating sequences. **Table 2-1** lists the time required for *iMapper* to process two different datasets using a variety of settings. The capacity of *iMapper* to process sequence reads in one submission depends on the web browsers capacity and computer memory. We have analysed over 10,000 traces in one submission using a Mac OS system or PC. However to guarantee the stability of the operating system, we have limited the submission scale to 10,000. More sequence traces can be processed in multiple batches or using a command version *iMapper* which is available upon request.

Table 2-1. Time required for *iMapper* to analyze two datasets with different settings.

	Tag length	Define orientation?	Contaminating sequence	SAHHA mapping	Time
582 PB sequences	17	Yes	-	No	2min30s
	17	Yes	-	Yes	8min50s
	17	No	-	Yes	11min25s
4032 SB sequences	17	No	-	No	32min
	17	No	-	Yes	1h35min
	17	No	2	Yes	1h55min

2.4.3.3 Sequence analysis of real insertion site data using iMapper

We used the optimal length for a *Sleeping Beauty* (SB) tag sequence (*'*TTCCGACTTCAACTGTA*'*, 17 bp) and the optimal percentage threshold (80 %) to test a *Sleeping Beauty* dataset containing 4032 sequence reads generated from mouse tumours using shotgun cloning (**Figure 2-5 C**). After *iMapper* processing, SB tag sequences were identified in 3576 sequence reads (89 %), of which nearly half of the reads (1531, 44 %) could be mapped to the Ensembl mouse genome using SSAHA (mapping score = 35), and two-thirds (1020 reads) of which overlapped to within 10 kb of a known Ensembl gene. An additional 128 reads could be mapped to the genome when using a score of 20 for SSAHA mapping rather than the default score of 35. The majority of these reads were short sequences of 25-45 bp, indicating that a smaller SSAHA mapping score could potentially benefit short sequence analysis such as those derived from 454 or Illumina-Solexa sequencing. These results demonstrate that *iMapper* is extremely efficient for the analysis of large-scale insertion data large-scale and provides a useful overview of the insertion sites as well as detailed information about each insertion site identified.

2.5 Discussion

Large-scale insertional mutagenesis screens require thousands, if not tens of thousands of sequencing reads to be processed and mapped in an accurate and efficient way. This has always been a time-limiting step for insertional mutagenesis experiments even with the extensive involvement of computational biologists. The fact that sequencing reads that are generated by linker-based PCR need to be carefully processed to get rid of contaminating sequences, chimeric sequences and tag sequences before sending the real genomic fragments for mapping, have complicated the analysis of insertion site sequence data. Despite the great power and popularity of insertional mutagenesis as a tool for screening candidate genes in model organisms and cell culture systems, there is no tailor-made software to facilitate the analysis and processing of insertion site data. As a result, each lab has developed their own approach and methods for analysing insertion site sequence data using different software, mapping algorithms and parameters. This has complicated the comparison and sharing of results.

As a freely accessible web tool, *iMapper* provides a simple solution for the processing of insertional mutagenesis data by experimental biologists who do not have a computational

background. *iMapper* also provides the user with the possibility to optimise the analysis of their data by defining some useful parameters during sequence processing and mapping, such as the length of mutagen tag sequence, the parameters used for local sequence alignment and SSAHA mapping. The quick and easy-to-use *iMapper* software could also provide different labs from around the world with a standard solution for insertional mutagenesis data processing, making it possible to compare and share mutagenesis screen results. Furthermore, by working closely with the Ensembl genome browser, *iMapper* also provides additional information and options for downstream analysis, such as viewing the genomic structure surrounding the insertion site, generating a *KaryoView* picture of the insertion sites and obtaining information on the insertion site gene. The fact that *iMapper* is maintained by the Wellcome Trust Sanger Institute, one of the largest bioinformatics centres in Europe will provide regular updates on sequence information and technical support.

One of the most obvious advantages for *iMapper* is the speed at which sequences are analyzed. It takes the software approximately 1.5 seconds to process and analyze each sequence read, the majority of which is spent on mapping the clipped genomic sequence using SSAHA (see results section). However, different setup parameters also affect the processing speed, for example, using longer tag sequences would result in a larger matrix in the local sequence alignment which, could slow down the speed for tag sequence identification, whereas searching for the tag sequence in a single orientation could speed up data processing since otherwise the program would search both the forward and reverse sequences. Other setups such as searching for contaminating sequence could also affect the speed for sequence processing using *iMapper*.

iMapper is a convenient tool with many useful functions for insertion site data analysis. The output page of *iMapper* lists the annotated sequences in a user-friendly tabular format, where detailed information and a link to Ensembl *ContigView* page are provided for each mapped sequence. *iMapper* also summarises the analysed data for downstream analysis, for example, *iMapper* is able to generate a *KaryoView* picture of all the mapped insertion sites which could give a direct overview of the results. In addition, *iMapper* generates clipped genomic sequences in FASTA format which could be used for genomic mapping by other softwares. The GFF file for all the mapped sequencing reads generated by *iMapper* provides useful information for further analysis of the insertion site data and can be uploaded as a DAS track against the Ensembl genome.

In *iMapper*, only sequences containing the tag sequence are processed for analysis and mapping on to the genome. Therefore the parameters for tag sequence identification are critical for the performance of *iMapper*. Following a performance test using different tag sequence lengths and percentage thresholds, the optimal length for tag sequence identification was found to be ≥15 bp and the percentage threshold, ≥80 % (**Figure 2-4 A and B**). These optimal values were determined when the accuracy and coverage curves reached a plateau, indicating that tag sequence identification was saturated and that further increasing the value would not result in a dramatic improvement in accuracy or coverage. To obtain sequencing reads that are at least 15 bp in length at the tag sequence end, the PCR primers and the sequencing primers should be more than 15 bp away from the end of the tag sequence. Appendix A lists the primer sequences that are recommended for splinkerette PCR on *piggyBac* and Sleeping Beauty samples. The same principles should be followed when designing PCR primers for splinkerette PCR on retrovirus samples. In addition, the parameters for tag sequence identification and genomic sequence mapping should also be determined using real sequence data. For example, if the sequencing is of poor quality or experimental noise is introduced to the tag sequence region, which is at the end of the sequencing reads, it is recommended that a longer tag sequence is used and a lower percentage threshold for tag sequence identification. Furthermore, if the sequence data are predominantly short sequences from 454 or Illumina-Solexa sequencing, it is recommended that a smaller SSAHA mapping score is used, which favours the identification of shorter sequences between 25-45 bp.

In conclusion, *iMapper* is designed for high-throughput applications and can efficiently process tens of thousands of DNA sequence reads in a short time. The web-based design of *iMapper* allows world-wide, free access to this software and the friendly user interface and operation functions allow users without a bioinformatics background to easily use *iMapper* to process their insertion site data and to change essential parameters to optimise their experimental results. In addition, *iMapper* provides each annotated sequence in a tabular format and links the mapped insertion sites to the Ensembl genome browser for further downstream analysis.

Availability: *iMapper* is web based and can be accessed at http://www.sanger.ac.uk/cgi-bin/teams/team113/*iMapper*.cgi. The code and algorithms are also available from this website.

Chapter 3. Slingshot: A *piggyBac* based transposon system for tamoxifen-inducible 'self-inactivating' insertional mutagenesis

3.1 Introduction

The complexity of the genes, pathways and networks that dictate many cellular phenotypes rarely makes it possible to employ a candidate gene (reverse genetic) approach to identify potential mediators of biological processes. 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, represents a powerful approach for gene discovery. Classically forward genetic screens in higher order organisms have been performed using ionizing radiation, chemical mutagens, or viruses each of which are likely to target a different repertoire of genes. While these approaches can be extremely efficient at generating mutants 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 ENU and EMS, which generate genome-wide point mutations and in so doing significant background noise from which a candidate gene carrying a mutation must be identified and then validated. Similarly, while ionizing radiation is a powerful tool for mutagenesis that generates sufficiently small chromosomal rearrangements so that a candidate gene can be identified

using approaches such as comparative genomic hybridization (CGH), 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. Even viruses, which can be introduced as single copy integrants into the genome, may function many megabases distal to their integration site. Viruses also exhibit strong insertional biases preferring to insert into active promoters and open chromatin, which effectively reduces the compendium of genes that they can mutate.

Transposons are mobile genetic elements that constitute a major part of the repetitive sequence of eukaryotes genomes (1). Transposons may be classified into two groups; DNA transposons and retrotransposons. DNA transposons consist of two terminal repeats flanking a transposable element, which allows them to be mobilized and relocated to other locations in the genome by a *'*cut-and-paste*'* mechanism. Retrotransposons function by a *'*copy-and-paste*'* mechanism. In lower organisms (worms, bacteria, *Drosophila*), DNA transposons have been used extensively for genetic manipulation and for mutagenesis (59,112,113). In recent years, DNA transposons have also been used for insertional mutagenesis screens in vertebrates, for example in zebrafish, *Tol2* has gained popularity, but has limited activity in mammalian cell cultures and *in vivo* (114). Other transposons including *Minos* have also been trialled in mammalian cells but with mixed success (115). In contrast the *Tc1*-family transposon, *Sleeping Beauty* (SB), has been shown to be effective for cancer gene discovery screens *in vivo* (88,89), and is an active mutagen in the germ line (116-118). However, while *Sleeping Beauty* is active in some cell lines such as 293T and HeLa cells, it appears to be weakly active in embryonic stem cells (108). *Sleeping Beauty* also exhibits significant 'local hopping', a phenomenon whereby a mobilised transposon re-integrates close to the donor locus (108,116,118). It also has a limited cargo capacity with mobilisation being significantly reduced when elements of more than 3kb are cloned between the inverted repeats/direct repeats (IR/DR) of the transposon (119). These factors, coupled with overexpression inhibition (where overexpression of the transposase inhibits transposition and optimal transposition is only obtainable within a narrow window of transposase expression) limits the utility of *Sleeping Beauty* as a universal mutagen. Despite these factors *Sleeping Beauty*-mediated screens have been successfully performed in cell culture systems (73).

More recently considerable effort has been invested in developing the transposon *piggyBac* (PB) from the moth *Trichoplusia ni*, as a tool for insertional mutagenesis (75,76,109,120-

122). Mobilisation of *piggyBac* from donor loci results in a more random distribution of transposon insertion events around the genome than is obtainable with *Sleeping Beauty* or *Tol2*. *piggyBac* can mobilise large cargo containing transposons of up to 50 kb (unpublished results; Bradley laboratory, Wellcome Trust Sanger Institute, Cambridge, UK) and unlike *Sleeping Beauty,* expression of the *piggyBac* transposase at high levels does not appear to result in overexpression inhibition (109). Another advantage of *piggyBac* is that the *piggyBac* transposase is still active when fused with other proteins such as the estrogen receptor ligand-binding domain (ERt2), which opens up a range of possibilities for temporally controlled mutagenesis (122). In addition to these factors, *piggyBac* appears to be highly active in mammalian cells and generates multiple independent insertions in cells into which the transposon is introduced. Collectively these factors suggest that *piggyBac* is a powerful mutagen that complements tools that are currently available for genetic screens in mammals.

The most common method for introducing *piggyBac*, or indeed other transposons, into mammalian cells, is to co-transfect a plasmid expressing the transposase ('helper' plasmid) and another plasmid carrying the transposon ('donor' plasmid) (**Figure 3-1 A**) (123). Once transfected into the host cells the transposase enzyme mobilizes the transposon from the donor plasmid and integrates it into the host genome. This plasmid based transposon system in cell culture has several drawbacks. Obviously, if only the 'helper' or 'donor' plasmid presented in a cell, transposition would not take place (**Figure 3-1 B and C**). Although the number of transposon integration events can be controlled to some extent by titering the amount of plasmid and the ratio of the helper and donor plasmids, integration patterns are frequently complex, thus increasing the difficult in isolating the useful insertion sites (**Figure 3-1 D**) (121). Similarly once integrated the continued expression of the transposase can re-mobilize transposon integration events, generating a complex pattern of integrations in subsequent cell divisions. Furthermore, efficient transfection of both *Sleeping Beauty* and *piggyBac* has proven difficult in some somatic cell lines which have limited the use of transposon systems for insertional mutagenesis screens in these cell lines. Therefore, although the transposon systems including *Sleeping Beauty* and *piggyBac* represent novel and powerful tools for mutagenesis study in mammalian cells, there is a lack of efficient methods to introduce these transposon systems into cell cultures for high efficiency insertional mutagenesis screening.

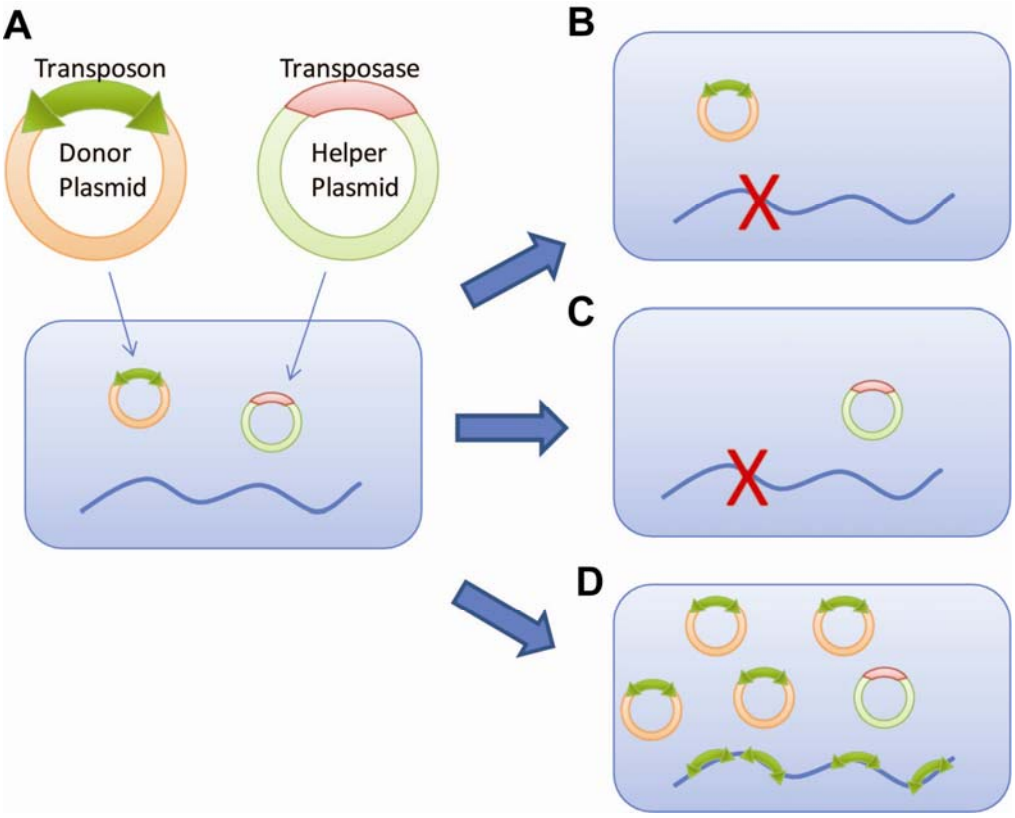


Figure 3-1. Plasmid based PB Transposon system in cell culture

(A) The Plasmid based PB Transposon system requires both the 'donor' and 'helper' plasmids to be introduced into one cell for transpostion taking place. If only 'donor' plasmid or 'helper' plasmid were present in a cell no transposition would result (B and C). Multiple copies of 'donor' plasmids in a cell would result in multiple jumping in the genome (D), therefore complicating downstream analysis.

3.2 Aim and summary of the project

To overcome the above-mentioned limitations, I set out to develop a novel, transposon-based insertional mutagenesis system with high efficiency for cell culture applications. The *piggyBac* (PB) transposon system was chosen for this purpose since it shows much higher activity than *Sleeping Beauty* and other transposon systems in different types of mammalian cells (123). In addition, the PB transposase shows favourable properties for molecular engineering. In a recent application, an estrogen receptor ligand-binding domain (ERt2) was added to the C-terminal of the PB transposase (mPB-L3-ERt2) to provide conditional activation of the transposase while retaining the efficiency of this system (122). Mouse ES cells were used for developing and testing this system because of their superior colony forming ability, and because the system has many potential applications in ES culture systems. For this project I will aim to increase the efficiency of PB transposition by combining both the transposon and inducible ERt2 transposase (mPB-L3-ERt2) in a single plasmid and use it to identify candidate genes in an insertional mutagenesis screen in ES cells.

The novel transposon system should have following advantages for cell culture applications:

1. The system should be easily introduced into cells, including ES cells and other somatic cell lines, for insertional mutagenesis screens.

2. The system should have superior transposition activity, and subsequent screening should be easy to carry out on a large scale.

3. When considering the difficulty in isolating the common insertion sites from a complicated insertion site pattern, it is preferable that the system introduces a low copy number of integrations per cell and the copy number should be controllable during the screen.

4. Ideally, the new transposon system should avoid re-mobilisation, which is caused by constitutive expression of the transposase enzyme. This should be accomplished by switching-off transposase expression after the first integration.

Based on these criteria, I have designed a self-inactivating *piggyBac* transposon system for tamoxifen inducible insertional mutagenesis. This system, which we have named *'*Slingshot*'* (referring to a single shot handheld weapon), contains a PB transposon upstream of a transposase sequence in one cassette as shown in **Figure 3-2 A**. The transposase (mPB-L3-ERt2), driven by the CAGGS promoter is tamoxifen inducible and has been described

previously (122). After stable integration of the Slingshot cassette into the genome and Blasticidin (BSD) selection (**Figure 3-2 B**), transposition can be initiated by the administration of 4-Hydroxytamoxifen (4-OHT). The Slingshot transposon contains a ubiquitous CAGGS promoter and a splicing donor (SD) (124) for gain of function mutagenesis (**Figure 3-2 D**), and elements for gene trapping to generate loss of function mutations upon integration into the genome (**Figure 3-2 C and E**). Mobilization, however, translocates the CAGGS promoter away from the Slingshot donor site, eliminating further expression of the transposase and effectively prevents remobilization (**Figure 3-2 F**).

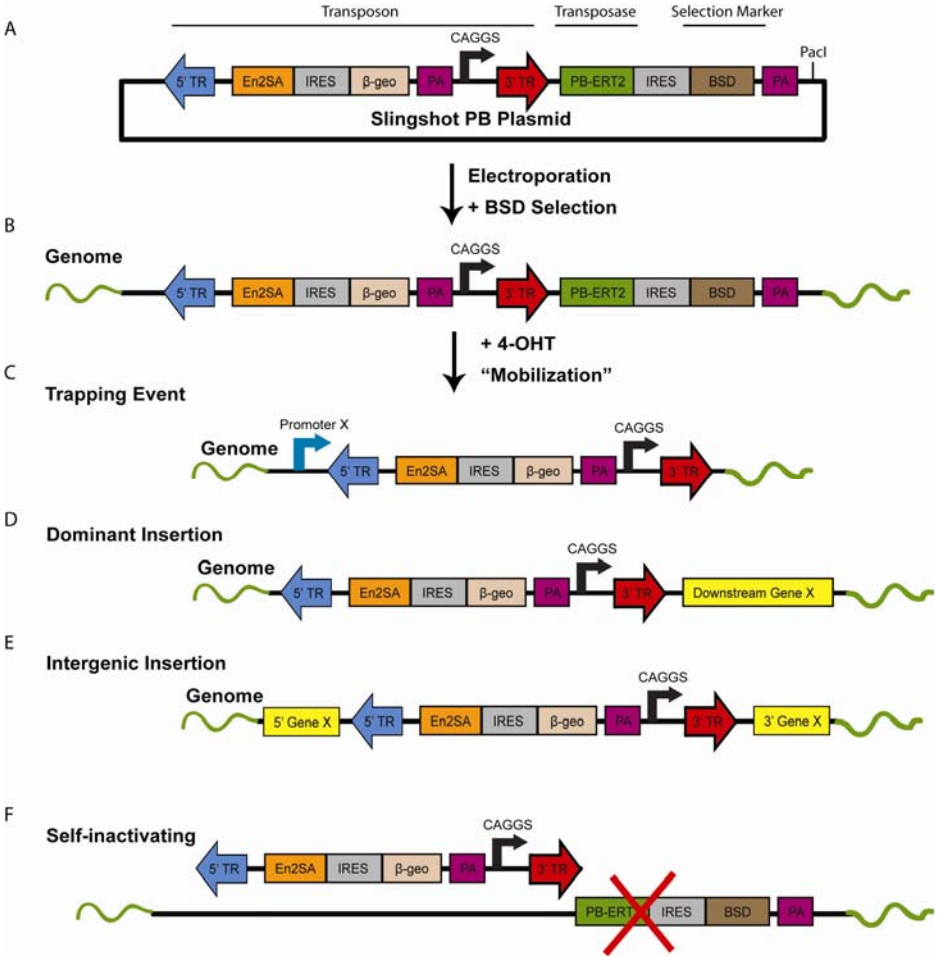


Figure 3-2. Schematic diagram of the Slingshot plasmid and the mutagenesis scenarios possible with this system

(A) The Slingshot system was constructed using a pBluescript II SK(+) vector as backbone The construct can be linearized using the restriction enzyme *Pac*I and stably integrated into a host cell genome using Blasticidin (BSD) selection. (B) The Slingshot transposon can be mobilized from the stable integrated Slingshot plasmid by treating cells with 4-OHT which translocates the PB-ERt2 fusion protein to the nucleus and mobilizes the transposon. (C) If the Slingshot transposon re-integrates downstream of an endogenous promoter 'X' it can hijack the promoter resulting in G418 resistance. (D) If the Slingshot transposon re-integrates upstream of a gene it can overexpress the gene by expression from the CAGGS promoter. (E) It is also possible for the Slingshot transposon to generate neomorphic alleles such as dominant negative alleles and loss of function events by generating intergenic insertions. (F) Mobilization of the transposon translocates the CAGGS promoter away from the PB-ERt2 preventing further re-mobilization events.

In this chapter I have develop and further characterize the Slingshot transposon system described above. My results show that the Slingshot transposon can be efficiently mobilized from a range of chromosomal loci with high inducibility and low background generating insertions that are randomly dispersed throughout the genome. Transposition and trapping efficiency is extremely high such that as many as 30,000 clonal insertion events can be generated from a single 10 cm plate of ES cells, making it theoretically feasible to screen all permissive genes and regions of the genome in just a few plates of cells in culture. To illustrate the efficacy of Slingshot as a screening tool experiments were set out to identify mediators of resistance to puromycin and the chemotherapeutic drug vincristine by performing a gain-of-function screen in mouse ES cells. From these genome-wide screens multiple independent insertions were identified in the multidrug resistance transporter genes *Abcb1a/b* and *Abcg2* and these insertions were shown to up-regulate the expression of these genes conferring resistance to drug treatment. Importantly the Slingshot transposon system was also been shown to be functional in other human somatic cell lines, suggesting that it may be used in a range of cell culture systems for genetic screens. From above results, Slingshot represents a flexible and potent system for genome-wide transposon-mediated mutagenesis with many potential applications.

3.3 Materials and Methods

3.3.1 Plasmids construction

The PB transposon element was constructed with 5' and 3' PB terminal repeats flanking a promoterless *β-geo* and a CAGGS promoter using elements derived from the plasmids 5'-PTK-3', pGTo1xr and pcDNA3.1. The *piggyBac* transposase estrogen receptor fusion (mPB-L3-ERt2) cDNA was obtained from Cadinanos *et al*. (122). The final plasmid was constructed on a pBluescript II SK(+) vector backbone using standard molecular cloning approaches and was sequenced in full. Genbank Accession: GU937109.

3.3.2 Cell culture media

Mouse embryonic stem (ES) cells were cultured in KNOCKOUT^{TM} DMEM (Invitrogen Cat. No.: 10829018) containing 2 0% Foetal Bovine Serum (Invitrogen Cat. No.: 16000044), 1× GPS, 1× BME and Leukemia inhibitory factor (LIF, Activity Varies). The human embryonic kidney cell line HEK293, and the ovarian carcinoma cell lines OVCAR-3 and PE01 were

cultured in KNOCKOUT^{TM} DMEM medium containing 8 % Foetal Bovine Serum (Invitrogen Cat. No.: 16000044) and 1× GPS.

3.3.3 Generation of an cell lines with stable integration of the Slingshot plasmid

To generate stable integrants carrying the Slingshot plasmid, 10^7 E14TG2a ES cells were electroporated at 230 V, 500 mF (BIO-RAD Gene Pulser II Electroporator) with 40 μg of Slingshot plasmid DNA following linearisationed with *Pac* I. Cells were cultured in 10 cm culture dishes for two days after transfection and subsequently selected with 15 μg/ml Blasticidin (BSD) for 10 days. Individual clones were picked into 96-well plates for downstream analysis. For transfection of HEK293, OVCAR-3 and PE01, 40 μg of *Pac* I linearised Slingshot plasmid was introduced into 10^7 cells by electroporation (300V, 800 μF) and stable integrants were selected with Blasticidin, (8 μg/ml for HEK293, 4 μg/ml for OVCAR-3 and 3 μg/ml for PE01). Round monolayer colonies formed after 2-4 weeks of selection. Cells were treated for two days with 4-OHT and G418 selection was carried out for 2-3 weeks at 500 μg/ml (HEK293), 400 μg/ml (OVCAR-3) and 200 μg/ml (PE01) to generate data on the efficiency test in these cells lines.

3.3.4 Trapping efficiency test and mobilisation assay

Mobilisation activity was tested by plating cell lines in duplicate into 6-well plates and treating with 1 μM 4-OHT (Sigma: H7904) or vehicle control (95 % EtOH) for two days. Cells were then cultured under G418 selection (175 μg/ml) for 8 days to select for gene trap events. The number of colonies was counted and the ratio of these events calculated to determine the transposon mobilisation/trapping activity. To further assess the trapping efficiency for each cell line, cells were treated with 1 μM 4-OHT for two days before a defined numbers of cells were plated on 10 cm culture dishes and treated with or without 175 μg/ml G418. Colony numbers were counted and the trapping efficiency was calculated by dividing the colony number from the plate treated with G418 (G418 resistant colonies)/by the colony number from the plate without G418 treatment (total colonies).

3.3.5 Excision PCR on DNA for 4-OHT treated cells

Excision PCR was performed to determine whether the transposon had been excised from the donor site. PCR was performed using Thermo start *Taq* Polymerase (Abgene) with primers

(forward: 5'-AAGTGTAGCGGTCACGCTGC-3' and reverse: 5'-CTCGATCACGTTCTGCTCGT-3') and the following reaction conditions: 95 °C 15 min/ 95 °C 0.5 min, 62 °C 0.5 min, 72 °C 1.5 min: 40 cycles/ 72 °C 5 min. The forward and reverse PCR primers flanked the transposon sequence in the Slingshot cassette, generating a PCR band of 526 bp if the transposon had been excised from the donor site. If the transposon remained intact, a much larger band (> 6 kb) or no band was generated.

3.3.6 Drug resistance screen using puromycin

To screen for genes responsible for puromycin resistance, 10^6 cells from clone 'PB/PB-1', which showed stable integration of Slingshot PB, were plated on 10 cm dishes. Cells were first treated with 1 μM 4-OHT or vehicle control (95 % EtOH) for two days and then cultured in normal M-15 media for another two days. Cells were then selected with 1 μg/ml puromycin for 10 days. Colonies from 4-OHT treated plates were then picked and genomic DNA was extracted to perform splinkerette PCRs to identify the insertion sites.

3.3.7 Drug resistance screen using vincristine

To screen for genes responsible for vincristine resistance, a titration test was performed to determine the lowest vincristine (Sigma Cat. no. V8388) concentration required to kill all ES cells. 10^6 ES cells from clone 'PB/PB-1' were plated on 10 cm dishes and treated with 1 μM 4-OHT or vehicle control (95 % EtOH) for two days before culturing in normal M-15 media for another two days. Cells were then selected with 10 pg/ml vincristine for 10 days. Colonies from 4-OHT treated plates were then picked and genomic DNA was extracted to perform splinkerette PCRs to identify the insertion sites.

3.3.8 Splinkerette PCR and insertion sites analysis

For splinkerette PCR 4 μg genomic DNA was digested overnight with 20 units of *Sau*3AI (NEB) in a 50 μl reaction volume. After heat inactivation at 65 °C for 20 min, 2.5 μl of the digestion mix was ligated with 1.5 μl of annealed linker oligonucleotides and 1 μl DNA ligase (NEB: M0202L) overnight. To prepare the linker oligos, 1.5 μM forward and 1.5 μM reverse linker oligonucleotides (in water) were boiled at 100 °C for 10 minutes and cooled to room temperature. One-microlitre of the ligation mixture was used for the first PCR using the 1st run splinkerette PCR primers. One-microlitre of the first PCR reaction was used as

template for the second PCR using 2nd run splinkerette primers (for all primers and oligo sequences used for splinkerette PCR see **Appendix A).** All PCRs were performed using Thermo-Start *Taq* DNA Polymerase (ABgene: AB-0908/B) with standard formula on the user's manuscript. PCR conditions were: 95 °C 15 min/ 95 °C 0.5 min, 62 °C 0.5 min, 72 °C 1.5 min: 40 cycles/ 72 °C 5 min. The PCR products were separated by agarose gel electrophoresis. Bands containing the junction site genomic DNA sequences were cut from the gel and sequenced using linker and PB sequencing primers. Sequencing reads were analyzed using the *iMapper* online web tool for identification of insertion sites (104).

3.3.9 Western blotting

Cells were lysed in NP-40 lysis buffer with 1 mM phenylmethanesulfonyl fluoride (PMSF). The protein concentration of each lysate was determined using the BCA (bicinchoninic acid) Protein Assay (Pierce). Approximately 50 μg of each lysate was fractionated on an SDS-PAGE gel. The protein was transferred to an Immobilon-P PVDF membrane (Millipore). The membrane was blocked with 5 % milk in TBST for 1 hour at room temperature and then incubated with primary antibodies [anti-Estrogen receptor (ER) (Santa Cruz Technology) and β-Actin (Cell signaling)] and then HRP-conjugated secondary antibody. The protein bands were visualized using the ECL Western blotting system (GE Healthcare).

3.4 Results

3.4.1 Generation of the Slingshot PB system and Slingshot ES cell lines

The Slingshot PB system was constructed as shown in **Figure 3-2**. To test whether the Slingshot PB cassette could introduce tamoxifen inducible transposition activity, 40 μg of Slingshot PB plasmid was transfected into ES cells. After transfection, half of the cells were plated on 10 cm dishes with 4-OHT treatment and half were plated without treatment. Two days later medium was changed to G418 selection at 175 μg/ml for another 10 days. If transposons re-integrated into an endogenous promoter (**Figure 3-2 C**), the β-geo gene in the transposon will be expressed to provide G418 resistance to the cell. After selection around 1000 G418 resistant colonies were generated on the 4-OHT treated plate while the non-treated plate had very few colonies formed (**Figure 3-3**). A similar result was obtained by

electroporating the PB transposon (neomycin + CAGGS version) and mPB-L3-ERt2 inducible transposase plasmids separately into cells (20 μg each), indicating that the Slingshot PB plasmid alone could generate comparable inducible transposition activity by combining the PB transposon and transposase on one cassette.

Since the PB transposon and transposase are located on one cassette in the Slingshot plasmid, in theory this design could increase the jumping efficiency when compared to a method which delivers the transposon and transposase via separate plasmids. However, the efficiency of the slingshot system could also be compromised by the large size of the plasmid itself (over 15 kb in length). To further improve the efficiency of the Slingshot system, I investigated the performance of Slingshot in cells with a stable integration (**Figure 3-2 B**). The Slingshot plasmids were electroporated into ES cells and selected with BSD (15 μg/ml) to generate stable integrated cell lines (see Materials and Methods). 53±5 (mean ± SD) BSD resistant colonies were formed from three independent electroporations. These colonies should all contain at least one stable Slingshot PB integration to enable the CAGGS promoter in the transposon region to drive the expression of the BSD selection marker in the cassette and provide BSD resistance. In total 18 colonies were picked and derived into cell lines for further characterisation.

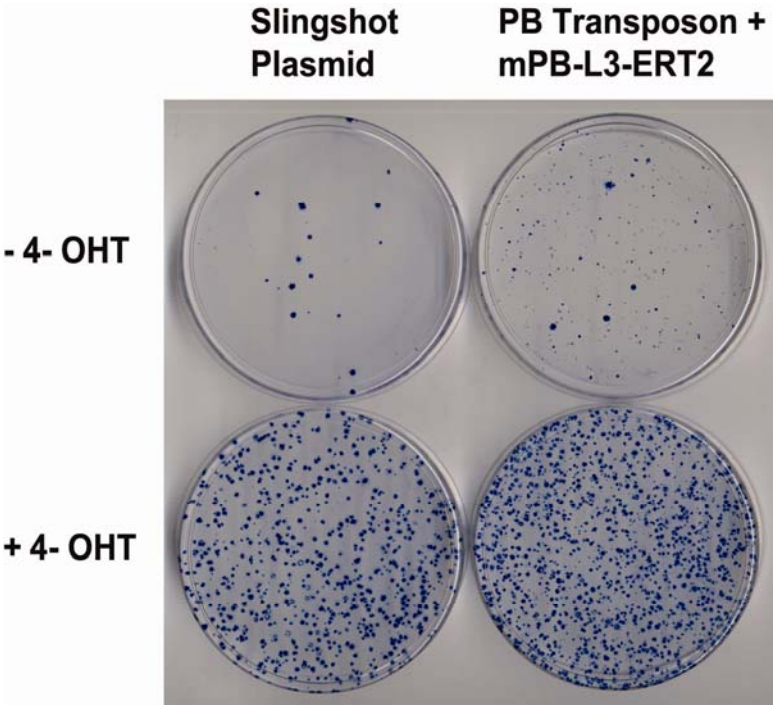


Figure 3-3. Functional test for the Slingshot PB plasmid

To test the activity of the Slingshot transposon system, the Slingshot plasmid (40 μg) was transfected into 10^7 mouse ES cells followed by 4-OHT treatment to activate transposition. After selection with G418 colonies were formed in the plate due to mobilization and re-integration of the PB transposon. As a positive control for this experiment, ES cells were transfected with two separate plasmids containing the PB transposon (neomycin + CAGGS version) and mPB-L3-ERt2 transposase (20 μg each).

3.4.2 Evaluating the jumping efficiency of the Slingshot transposon from a stable donor

To test the transposition activity from the 18 Slingshot ES cell lines showing stable integration of the Slingshot transposon system, each cell line was amplified and plated on a six-well plate and treated with 4-OHT to activate transposition events or vehicle control (95 % EtOH). After G418 selection, colonies were counted and colony numbers from 4-OHT treated and non-treated wells were compared. The cell lines with large numbers of colonies following 4-OHT treated but few colonies in non-treated wells represent cell lines with a high jumping activity that is tightly controlled for tamoxifen inducibility. To simplify the quantification, we defined a ratio of 2 (colony number in 4-OHT well/ colony number in non-treated well) as the threshold for a cell line having transposition activity.

As shown in **Figure 3-4 A**, 10 cell lines out of 18 showed jumping activity after 4-OHT treatment (colony number in 4-OHT well/ colony number in non-treated well > 2). Some cell lines were BSD resistant but showed no obvious jumping activity, which might suggest that the Slingshot PB array has been damaged during integration or that the plasmid has integrated into a site in the genome where expression from the CAGGS promoter is not permissive. Several cell lines formed colonies in 4-OHT treated wells (clones 3, 12, 18), but also formed colonies in vehicle treated wells suggesting that Slingshot has integrated into a site where β-geo is promiscuously expressed therefore resulting in G418 resistance. We selected clone 1 (PB/PB-1) for further analysis because it showed low background and high inducibility which was maintained over 12 passages (**Figure 3-4 B**). The transposition activity in PB/PB-1 clones after 4-OHT treatment and G418 selection was demonstrated by excision PCR; DNA were extracted from eight randomly picked colonies and PCR was performed to prove that transposon has been excised from the original donor site (**Figure 3-4 C**). The Slingshot plasmid integration sites for this clone were characterised by Southern Blotting and Splinkerette PCR; only one donor site was identified which was subsequently mapped to chromosome 18 (See **Figure 3-4 D**, red arrow). To assess the transposon re-integration we picked 48 G418 resistant PB/PB-1 clones and amplified their transposon insertion sites using splinkerette PCR. *iMapper* (104) was used to generate a *KaryoView* picture of the insertion sites and showed that there does not appear to be any local hopping from the original donor site on chromosome 18 for this clone, and that insertions appeared to be widely distributed throughout the genome (**Figure 3-4 D**).

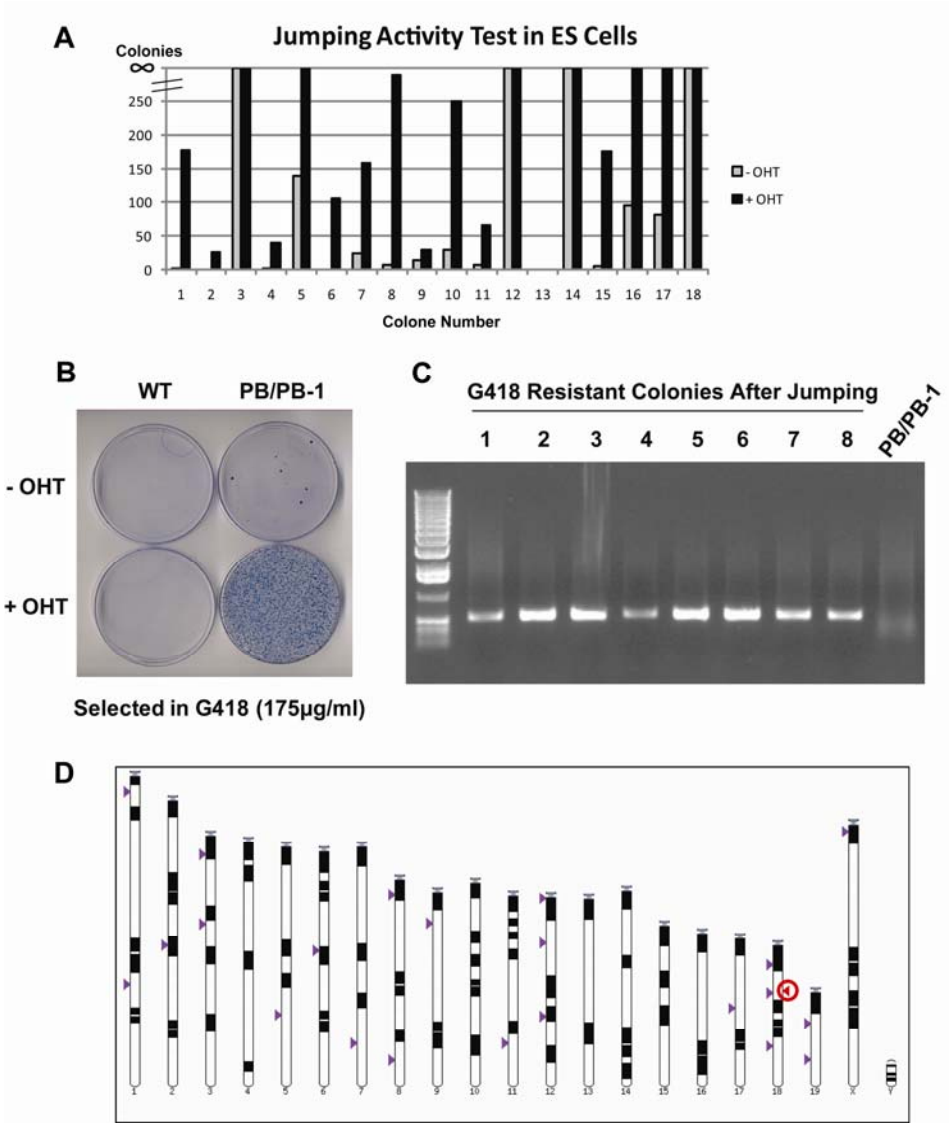


Figure 3-4. Testing of transposition activity and re-integration sites in Slingshot PB integrated ES cell lines

(A) The transposition activity in 18 ES cells lines showing stable integration of Slingshot PB was tested. The x-axis shows the individual ES colony number and the y-axis shows the colony counts for 4-OHT treated and non-treated cells cultured in six-well plates after G418 selection. (B) Colony one (PB/PB-1) ES cells treated with 4-OHT had high transposition activity and formed G418 resistant colonies in 10 cm culture dishes. (C) Excision PCR of genomic DNA for 8 PB/PB-1 G418 resistant colonies after 4-OHT treatment resulted in a band of 526 bp. The last lane is the PB/PB-1 control DNA without 4-OHT treatment. (D) *KaryoView* picture of the insertion sites isolated from 48 PB/PB-1 G418 resistant colonies generated by *iMapper*. The red arrow indicates the original Slingshot transposon donor site.

To further characterise the transposition efficiency from the Slingshot donor we performed a time series of mobilisation tests to quantify jumping efficiency with 4-OHT treatment (**Table 3-1**). From these tests we obtained mobilisation efficiencies ranging from 0.2-0.4 %, with longer 4-OHT treatment resulting in slightly higher jumping efficiency. This means that a confluent 10 cm dish with 10^7 ES cells would theoretically be sufficient for whole genome coverage (about 3×10^4 insertions). Although the Slingshot donor uses the mPB-L3-ERt2 transposase which has lower activity compared with the wild type version, Slingshot transposition is considerably higher when compared with other studies using plasmid based delivery methods (76,109,122). It is worth noting in these experiments that we are equating trapping efficiency with mobilisation efficiency. It is likely that the mobilisation efficiency of Slingshot is orders of magnitude higher than the systems' trapping efficiency.

Table 3-1. Evaluation of PB/PB-1 transposition efficiency using a time Series of 4-OHT treatment

4-OHT treatment	Cell number plated	Colony number - G418 plate *	Colony number + G418 plate **	Jumping Efficiency (Trapping) ***
1-Day	1 X 10^4	~ 3000	5 ± 2	> 0.17 %
2-Day	1 X 10^4	~ 3000	7 ± 2	> 0.23 %
4-Day	1 X 10^4	~ 3000	11 ± 2	> 0.37 %

* The colony numbers in non-G418 treated plates were estimated

** The colony numbers in G418 treated plates were calculated by mean ± SD in three parallel plates

*** The jumping efficiency is represented as trapping efficiency therefore the actual transposon jumping efficiency should be much higher

3.4.3 Drug resistance screens using the Slingshot system

The transposon used in the Slingshot system contains a CAGGS promoter and splicing acceptor- polyA 'trapping' element to generate both gain and loss of function transposition events. The trapping experiments described above essentially validate the loss-of-function elements of the transposon. Here the gain of function capabilities of Slingshot were tested by performing drug resistance screens *in vitro* using two agents, the aminonucleoside antibiotic puromycin and vincristine, an antimicrotubule spindle poison.

3.4.3.1 Puromycin resistance screen

For the puromycin screen, 18 and 26 colonies from two independent experiments (using 8 and 20 plates each) were derived after puromycin selection (**Figure 3-5 A and B**). Insertion sites were isolated from these clones by splinkerette PCR (**Figure 3-5 C**) and mapped to the genome with *iMapper*. From a total of 44 colonies, the majority of the insertion sites mapped independently to two genomic loci . Twenty-one of the insertions mapped to chromosome 5 between 8.5 - 8.8 MB (**Figure 3-5 D**), where genes encoding the ABC Transporters *Abcb1a* and *Abcb1b* are localised. Sixteen of these insertion sites were located upstream of the gene *Abcb1a,* ranging from 0.2 Mb away from the first exon to the first intron. Five were located upstream of *Abcb1b* before the first exon. The transposons all inserted in a sense orientation at this locus and the CAGGS promoters were all facing the gene orientation. Since the genes *Abcb1a* and *Abcb1b* are paralogous, having evolved as a result of a gene duplication, their drug transporter activity is essentially identical. The other locus that had multiple insertion sites was located on chromosome 6 between 58.54 - 58.59 MB (**Figure 3-5 E**), where another five transposon insertion sites were located. These transposition events are predicted to drive overexpression of *Abcg2*. *Abcb1a/b* and *Abcg2* are ABC drug transporters and these results indicate that that puromycin is a substrate for both of these drug efflux pumps (125). Six clones derived from control plates were also analyzed by splinkerette PCR and no insertion sites could be identified.

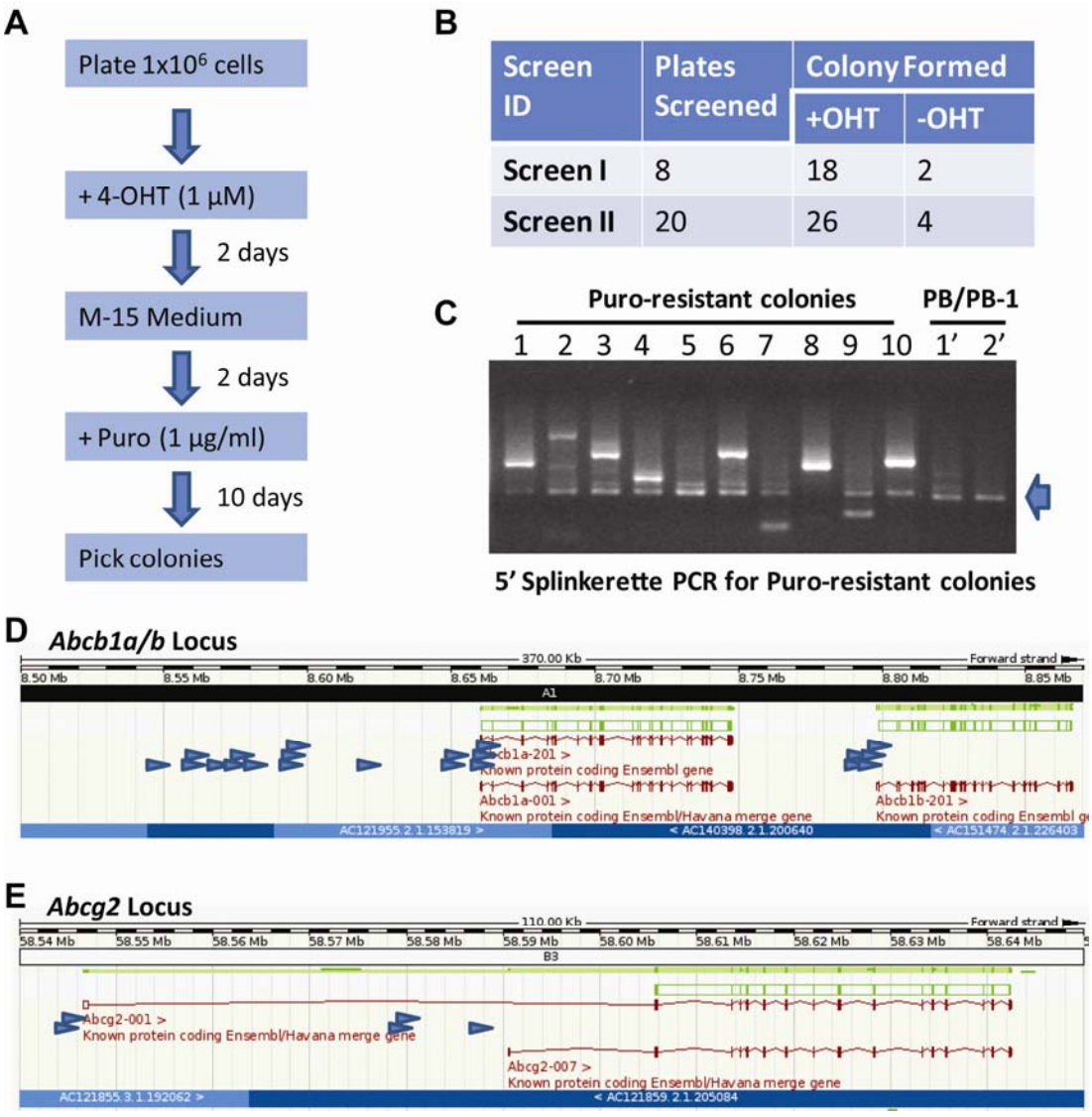
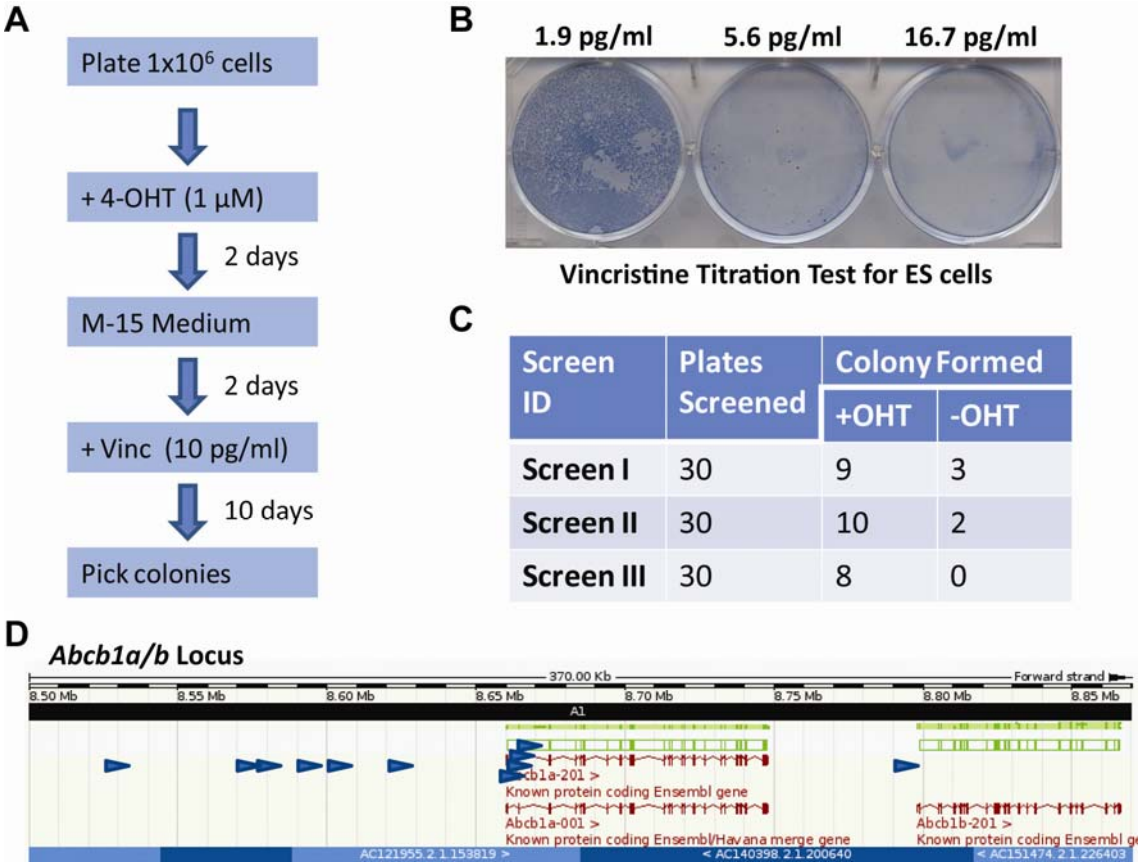


Figure 3-5. Drug resistance screen in Slingshot PB cell line PB/PB-1 using puromycin

(A) Flow chat showing overview of the drug resistance screen in PB/PB-1 using puromycin. (B) Information and results of two independent experiments. (C) Splinkerette PCR gel picture using 5 prime PCR primers as shown in Appendix A. Colonies 1-8 are puromycin resistant colonies isolated from the screen. Lane 1' and 2' are control DNA from PB/PB-1. The blue arrow indicates a background PCR band cloned from the Slingshot plasmid sequence. (D) Twenty-one independent insertion sites (indicated in blue arrowhead) were mapped to *Abcb1a/b* Locus on chromosome 5 between 8.5-8.8 MB. (E) Five independent insertion sites (indicated by blue arrowhead) were mapped to the *Abcg2* Locus on chromosome 6 between 58.54 - 58.59 MB. The orientation of the arrows indicates the CAGGS promoter orientation after transposon integration.

3.4.3.2 Vincristine resistance screen

For the vincristine resistance screen using a titration test was first performed to establish the optimal concentration of drug required to kill all cells (**Figure 3-6 B**). From this test we determined that a drug concentration of 10 pg/ml is sufficient to kill all the ES cells seeded at a density of 10^7 cells in 10 cm plates within 4 days and this concentration was used in subsequent experiments. In total three independent experiments were carried out using 30 plates each time and in total 27 colonies were obtained (**Figure 3-6 C**). The insertion sites were cloned by splinkerette PCR and 11 independent insertions were mapped to the *Abcb1a/b* locus on chromosome 5 between 8.5 - 8.8 MB (**Figure 3-6 D**). Therefore this screen has independently identified *Abcb1a/b* as a gene for anti-vincristine resistance. In both the puromycin screen and also in the vincristine screen insertions from several clones could not be mapped. This was generally due to the short sequence length of the splinkerette product or repetitive sequences that could not be mapped.



Screen ID	Plates Screened	Colony Formed	
		+OHT	-OHT
Screen I	30	9	3
Screen II	30	10	2
Screen III	30	8	0

Figure 3-6. Screen for vincristine resistance in the Slingshot PB cell line PB/PB-1

(A) Overview of the screen. (B) Titration test to determine the optimal vincristine concentration in ES cells. (C) Results of three independent screening experiments. (D) 11 independent insertion sites (indicated by blue arrowhead) were mapped to the *Abcb1a/b* Locus on chromosome 5 between 8.5-8.8 MB. The orientation of the arrows indicates the CAGGS promoter orientation after transposon integration.

3.4.4 Self-inactivation of the transposon after transposition

Transposon remobilisation caused by constitutive expression of a transposase is a major problem in insertional mutagenesis studies. In mouse tumour studies the *Sleeping Beauty* transposon may 'jump' multiple times before landing in the identified insertion sites. At each site of integration, *Sleeping Beauty* leaves a TA footprint which could cause a frame shift in the coding region. *piggyBac* has an advantage over *Sleeping Beauty* in that it is faithfully excised and leaves no footprint in the donor site. However remobilisation of *piggyBac* during transposition could still generate a complex insertion site profile and cause difficulty in isolating the common insertion sites.

During transposition the Slingshot transposon and its CAGGS promoter sequence are excised from the original donor site and re-integrate elsewhere in the genome. In theory this translocates the CAGGS promoter away from the transposase shutting down further expression to prevent re-mobilization of the transposon. To prove this theory I used puromycin resistant colonies generated using PB/PB-1 that had integrated upstream of *Abcb1a* but were G418 sensitive (puromycin resistant clones 1, 8 and 14). The presence of the transposon in these three clones was reconfirmed by splinkerette PCR and further verified by genomic PCR (data not shown). These cells were treated with 4-OHT and then selected in G418 to identify re-mobilization events, however none were detected. These data illustrate that transposase activity is completely shut down following mobilisation of the Slingshot transposon (**Figure 3-7 A**). To prove that re-mobilisation from the *Abcb1a* locus is possible, puromycin resistant clones 1, 8 and14 were transfected with the mPB-L3-ERt2 plasmid by electroporation, cells were treated with 4-OHT for 2 days, and then selected in G418. All three cell lines treated in this way generated hundreds of colonies per plate (**Figure 3-7 A**), indicating that the lack of transposition in these puromycin resistant clones was due to lack of PB transposase expression in these cells. Eight colonies were picked from each of the three puromycin resistant cultures and mobilisation of the Slingshot transposon from the *Abcb1a* locus was confirmed by excision PCR (data not shown). Finally, while PB-ERt2 expression was easily detected in the PB/PB-1 clone using an anti-ER antibody (see Materials and Methods), PB-ERt2 transposase expression was absent in whole cell lysates from puromycin resistant clones 1, 8 and 14 further confirming that transposition from the Slingshot donor represses further transposase expression (**Figure 3-7 B**).

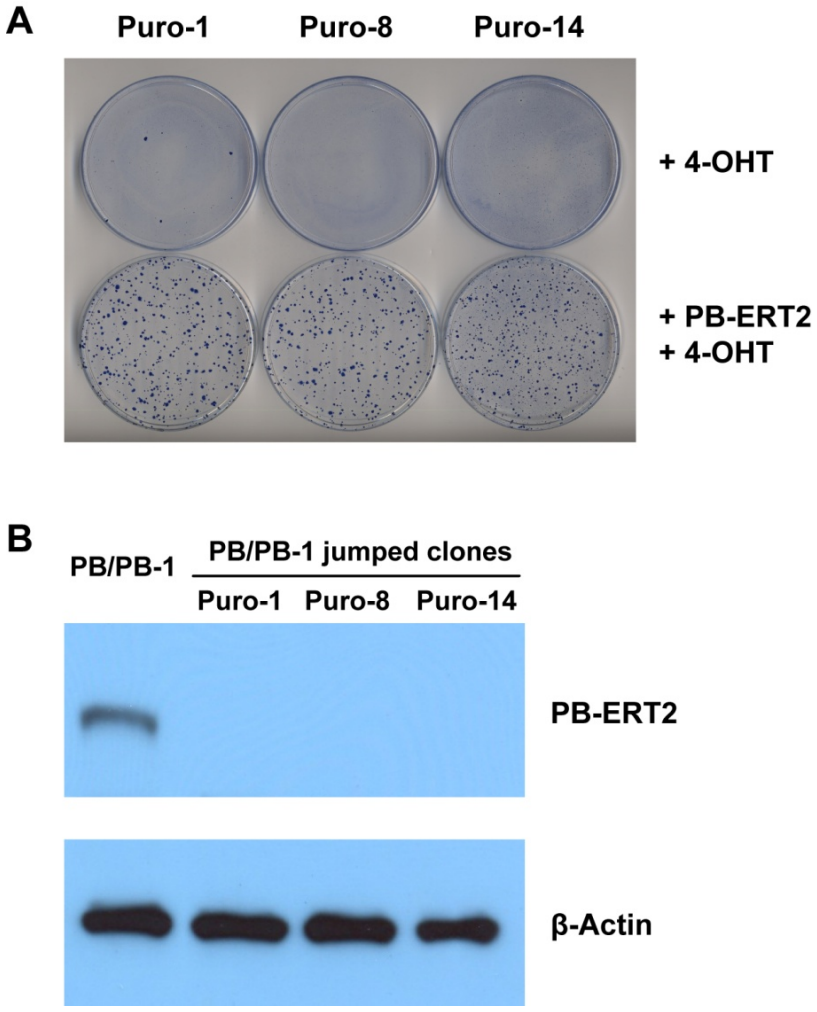


Figure 3-7. Slingshot is a self-inactivating transposon system

(A) Colony forming assay to test remobilization in three puromycin resistant clones, Puro-1, -8, -14. These three clones are derived from PB/PB-1 and all contain a PB transposon copy inserted upstream of the ABC transporter gene *Abcb1a* but are G418 sensitive. While there was no transposition after 4-OHT treatment of Puro-1, -8, -14 (top plates) transfection of a PB-ERt2 plasmid into these cell lines followed by 4-OHT treatment reveals that the PB transposon can be remobilised from the *Abcb1a* locus (bottom plates) (B) Western Blot for PB-ERt2 transposase expression using an anti-ER antibody. Expression of PB-ERt2 is readily detectable in the PB/PB-1 control but completely absent from Puro-1, -8, -14.

3.4.5 The Slingshot transposon system is active in somatic cell lines

To expand the application of the Slingshot system to somatic cell lines, in which many cell culture screening systems have been established, the Slingshot donor plasmid was also introduced into three commonly used human experimental somatic cell lines to test their transposition activities: the human embryonic kidney cell line HEK293, and the ovarian carcinoma cell lines OVCAR-3 and PE01. All three cell lines were transfected with 40 μg of linearized Slingshot plasmid which was introduced into 10^7 cells by electroporation (300V, 800 μF) and stable integrants were selected with Blasticidin, treated with 4-OHT for two days and selected with G418 for 2-3 weeks.

All three human somatic cell lines showed considerable trapping activity after 4-OHT treatment (**Figure 3-8**): 6 colonies out of 19 for HEK293, 2 colonies out of 16 for OVCAR-3 and 3 colonies out of 11 for PE01 cells had obvious transposition activity (colony number in 4-OHT well/ colony number in non-treated well > 2). The excision of the transposon from the Slingshot cassette in these three cell lines was confirmed by excision PCR (**Figure 3-8 C, F, I**). The constitutive expression of the PB-ERt2 transposase in OVCAR-3 cell lines was detectable by Western blot even after 12 passages (**Figure 3-8 J**). These results indicate that the Slingshot transposon system could be an efficient tool for mutagenesis studies in these somatic cell lines.

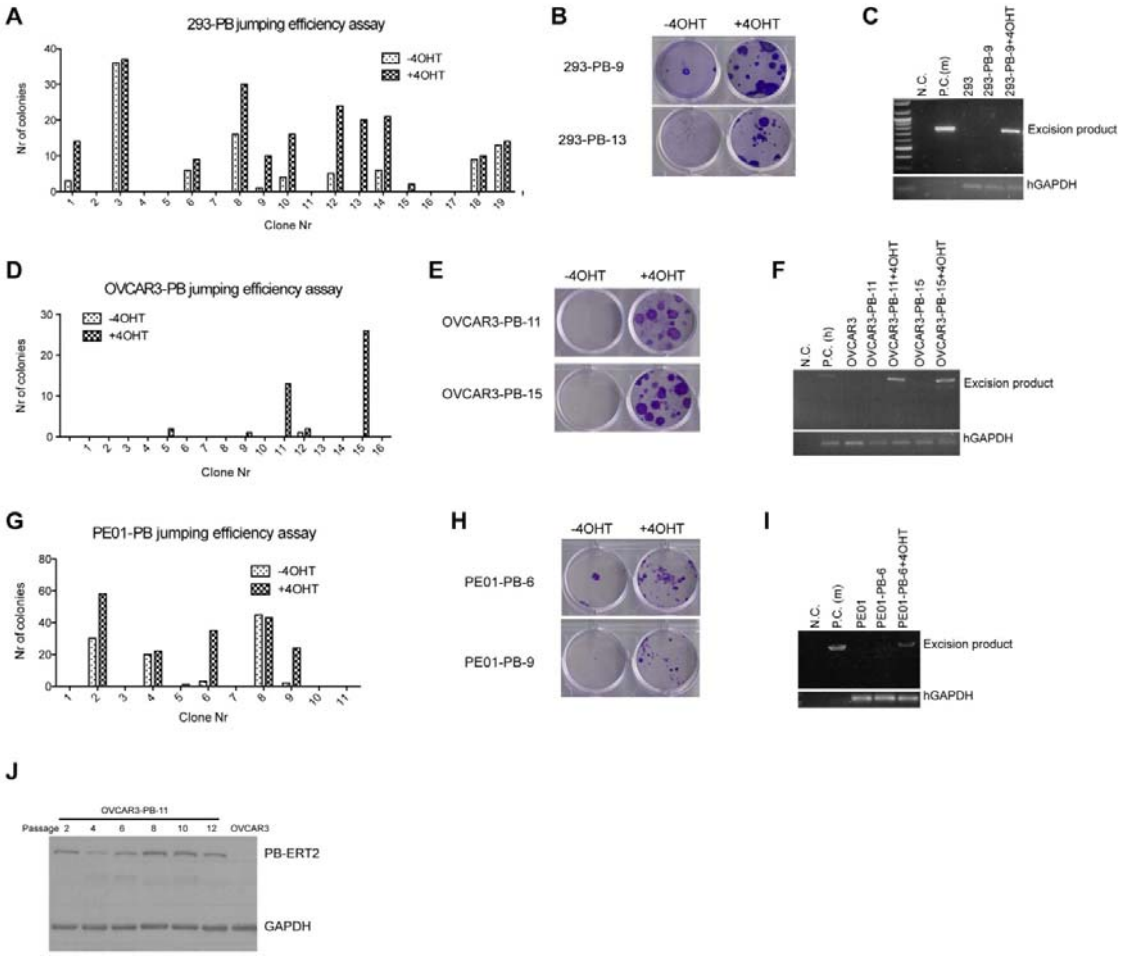


Figure 3-8. Slingshot is functional in three human somatic cell lines

HEK293, OVCAR3 and PE01 cells were stably transfected with the Slingshot plasmid and trapping efficiency assays and excision PCRs were performed. The Slingshot transposon was mobilised in HEK293 (**A, B, C**), OVCAR3 (**D, E, F**) and also in PE01 (**G, H, I**) following analysis of 19, 18 and 11 clones from these cell lines, respectively. Excision PCR was performed on DNA extracted from clones from each cell line pre- and post- 4-OHT treatments. N.C., negative control; P.C., positive control. (DNA from the Puro-1 cell line) Human GAPDH was used as a gDNA loading control. **(J)** Western blot analysis shows expression of the fusion protein PB-ERt2 in a highly active PB-carrying OVCAR3 clone following serial passage. The fusion protein was detected using an antibody recognising the ERt2 domain. Parental OVCAR3 cells were used as a negative control. GAPDH was used as a loading control.

3.4.6 Increasing the integration efficiency using chicken insulator sequence cHS4

In most transgenic experiments, the expression of integrated transgenic elements is subject to the influence of surrounding chromatin structure, a phenotype called chromosomal position effects (126). Chicken hypersensitive site 4 (cHS4), a well characterised insulator sequence was used to flank and protect the Slingshot PB cassette from chromosomal position effects (**Figure 3-9 A**). The cHS4 sequence is identified from the 5' element of the chicken β-globin domain and has been shown to improve the expression of integrating gene transfer vectors by reducing the position effects (127-129). To test the efficiency of cHS4 sequence, identical amounts of cHS4 flanked Slingshot or non-flanked Slingshot were electroporated into ES cells (40 μg DNA per electroporation) and selected with BSD (15 μg/ml). In three independent electroporations, 885 ± 48 (Means ± SD) colonies formed on plates seeded with ES cells transfected with cHS4 flanked Slingshot PB plasmid, a 17-fold increase when compared to the number of colonies formed on plates transfected with the non-flanked Slingshot (53 ± 5) (**Figure 3-9 B and C)**. To compare the transposition activity for the cHS4-flanked and non-flanked Slingshot clones, 24 clones were picked from each for mobilization activity test. 16 of non-flanked and 18 of cHS4 flanked Slingshot colonies showed obvious jumping activity after 4-OHT induction followed by G418 selection and there was no obvious increase in colony numbers for the cHS4 flanked Slingshot colonies. From the above results the chicken insulator sequence cHS4 significantly increased the colony number after electroporation, indicating that it could protect the Slingshot cassette from position effects while integrating into cells. Among cell lines with stable Slingshot integration, however, there seems little difference in transposition efficiency between colonies transfected with the cHS4-flanked and non-flanked Slingshot cassette. Of course, it is possible that the cHS4-flanked Slingshot could improve mobilisation efficiency in somatic cell lines, in which the transgenic elements are more likely to be silenced by position effects.

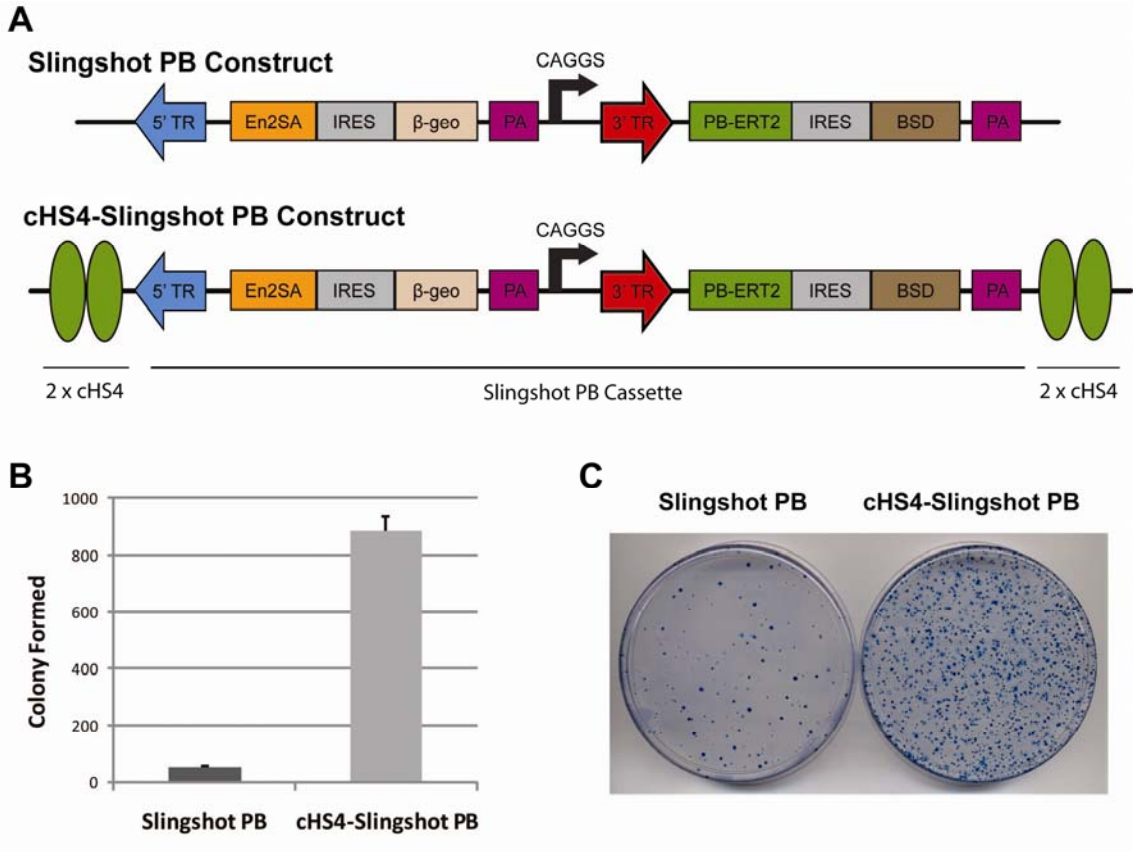


Figure 3-9. Improvement of the Slingshot PB donor colony formation efficiency using chicken insulator sequence cHS4.

(A) Slingshot PB and cHS4-Slingshot PB constructs. The cHS4-Slingshot PB is flanked with two cHS4 sequences. (B) 10^7 ES cells were electroporated at 230 V, 500 mF with Slingshot PB and cHS4-Slingshot PB constructs and the number of colonies counted following BSD selection (mean ± SD from three experiments) (C) Representative example of colony plates.

3.5 Discussion

The recent discovery of *piggyBac* (PB), a transposon derived from the cabbage looper moth *Trichoplusia ni*, which is active in mammalian cells has opened up new opportunities for insertional mutagenesis in mammalian (75). When compared to other insertional mutagens *piggyBac* has been shown to exhibit much higher rates of transposition, less local hopping and can carry large cargo sequences. In addition the *piggyBac* transposase can be fused to other sequences which has made it possible to control the transposition temporally (75,76,109,123). These advantages have made the PB system an ideal insertional mutagen for genetic screens.

I have developed a system called Slingshot, which can be used for self-inactivating insertional mutagenesis in mouse embryonic stem cells and also in a range of human somatic cell lines. The Slingshot PB cassette showed stable integration into host cell genomes and the transposition activity of each Slingshot donor cell line was easily evaluated using a colony formation assay. This 4-OHT inducible Slingshot PB system has many advantages over the PB system carried by donor and helper plasmids for cell culture applications. By stable integration into the host cell genome, the Slingshot PB cassette provides transposition activity in almost every cell, which significantly increased the transposition efficiency compared with plasmid-based transposition experiments. Because the transposon is mobilized from the integrated Slingshot PB cassette, the transposon copy number is determined by the actual copy of integrated Slingshot PB cassette per cell and this could be determined by splinkerette or Southern blot. In addition, since transposition is controlled by 4-OHT, transposition could be terminated by withdrawing 4-OHT to avoid multiple re-integration events. I have also shown that once the transposon has jumped out of the Slingshot cassette, transposase activity is shut down to prevent re-mobilisation.

To further characterise the transposition ability of the Slingshot PB system, a cell line with stable Slingshot PB integration (clone ‘PB/PB-1’) was used to characterise the jumping efficiency by colony formation assay. A single copy of the Slingshot PB cassette was identified by splinkerette PCR and mapped to chromosome 18. We identified the integration sites in 48 clones and found that the insertion sites were randomly distributed throughout the genome with no obvious local hopping. This is a particularly important feature of *piggyBac* that has previously been described and makes it ideal for PB to transpose from a defined genomic locus. The Slingshot PB cell line showed a trapping efficiency of around 0.3%.

Since the assay only detects trapping rather than mobilisation events, the actual jumping efficiency of the Slingshot system could be much higher. Nevertheless, this efficiency is much higher than that reported for studies using the plasmid transformation method and a PB trapping cassette (121,122), though it is slightly lower than the efficiency reported for mobilisation of PB from the HPRT locus, which measures the jumping efficiency rather than trapping efficiency (76). In addition, the Slingshot PB system showed promising activity in human somatic cell lines such as the human embryonic kidney HEK293 and human ovarian carcinoma cell lines OVCAR-3 and PE01. Although the actual jumping efficiency is difficult to quantify in these somatic cell lines as they can only form mono-layer colonies which are unfavourable for the colony formation assay, these experiments demonstrated that the stable integration of the Slingshot PB system is an ideal tool for high-efficient mutagenesis studies in ES cells as well as other somatic cell lines.

Here I have shown that Slingshot can be used to identify candidate genes in genome-wide mutagenesis studies by performing two screens to identify mediators of resistance to the compounds puromycin and vincristine. In the screen for puromycin resistance three ABC drug transporter genes *Abcb1a*, *Abcb1b* and *Abcg2* were hit multiple times by independent transposon insertions. *Abcb1a*, *Abcb1b* encode the same drug transporter protein ABCB1 and are located next to each other in the genome. *Abcg2* encodes the drug transporter protein ABCG2. The orientation of the CAGGS promoter in all of the insertions faced the gene orientation, indicating that the transposon activates these drug transporter genes via its CAG promoter sequence. In the screen for vincristine resistance, the drug transporter genes *Abcb1a*, *Abcb1b* were also hit multiple times suggesting that activation of these genes also results in vincristine resistance. We did not identify all of the genes that have previously been shown to result in vincristine resistance, such as the Multi-Drug Resistance Protein coding genes *MRP-1* and *MRP-2*. However, this may be because the pilot screen was not large enough or the *MPR* gene loci are not accessible for PB integration. Nevertheless, these screens have shown that the Slingshot PB system is an efficient tool for identifying activating mutations. In theory the Slingshot PB system could also be used to identify recessive mutations via the disruption of gene transcription. Since the Slingshot PB cell line we used for these drug resistance screens contained only a single copy of Slingshot, this cell line is more suited to the identification of dominant mutations.

We have found that the insertion sites generated by low copy numbers of transposons are easier to isolate and characterise by standard splinkerette protocols, while the insertions

generated by high copy numbers of transposons require more complicated, deep sequencing technology. Due to its large size, the Slingshot PB system is more likely to introduce a low copy number of integrations per cell. In fact, of all the cell lines we analysed, we did not identify any with more than one copy of Slingshot. Therefore the Slingshot PB system would be an ideal tool for the identification of phenotypes caused by single-cell events under defined cell culture conditions, such as the identification of drug resistant genes or cooperating gene mutations in a defined cellular background. More complicated cellular processes that are caused by cooperative genetic events, however, will only be identified using mutagenesis tools such as retroviruses or multi-copy transposon systems.

Although the identification of a candidate gene from insertional mutagenesis screens largely depends on the accessibility of the genomic locus and DNA modifications, methylation, acetylation status etc., the efficiency of the insertional system plays a very important role in mutagenesis screens. Therefore it is worthwhile investigating strategies that would further improve the transposition efficiency of the Slingshot system. The most obvious approach would be to reduce the size of transposon itself by, for example, removing the LacZ sequence in the transposon region, since mobilisation of the transposon is exponentially dependent on the size of the transposition molecule. Furthermore, it would also be possible to add a GFP tag to indicate any mobilisation events within a cell. This would enable the isolation and enrichment of cells by flow-assisted cell sorting. Last but not least, most of the Slingshot donor cell lines identified during these screens had only a single copy Slingshot integration. It is worth considering an alternative method of transfection such as Lipofectamine to increase copy numbers which could potentially improve the transposition efficiency several fold.

In summary, Slingshot is a stable, self-inactivating mutagenesis system that has several potential applications. Firstly, as I have shown here, Slingshot represents a useful tool for genome-wide screens. Secondly, since the transposon is mobilised from a stable chromosomal donor by adding tamoxifen to the culture medium, Slingshot can be used in heterogeneous populations of cells, in three dimensional culture systems, or where it is impractical to transiently transfect donor and helper plasmids into cells. Thirdly, since the transposon is mobilised only once following the administration of tamoxifen the Slingshot system can be used to *'barcode'* populations of cells making it possible to track the dynamics of growth of the population over time. When combined with high-throughput sequencing which could read the barcodes or insertion sites from a population of cells, it would be possible to use Slingshot in synthetic genetic screens such as those performed in yeast.

Chapter 4. Modelling *Tel-AML1* oncogenic translocation using knockin mice and transposon-mediated insertional mutagenesis

4.1 Introduction

Chromosome translocation is an event which results in exchange of genetic material between the arms of two non-homologous chromosomes. Occasionally this is accompanied by expression of a fusion chimeric protein produced at the translocation join point. Chromosome translocations can be classified into two types. Balanced chromosome translocation is the exchange of genetic materials between two chromosomes without the gain or loss of genetic information (**Figure 4-1 A**). In contrast, unbalanced chromosome translocation results in the gain or loss of genetic information through unequal chromosome exchange. Chromosome translocation has been suggested to be the cause of many types of genetic disorders including cancer, infertility and Downs syndrome (130-132). In particular, balanced chromosomal translocations have been identified as predisposing events in many types of haematological malignancies (133). For example, the ETS domain encoding genes are involved in several chromosomal arrangements (134-136).

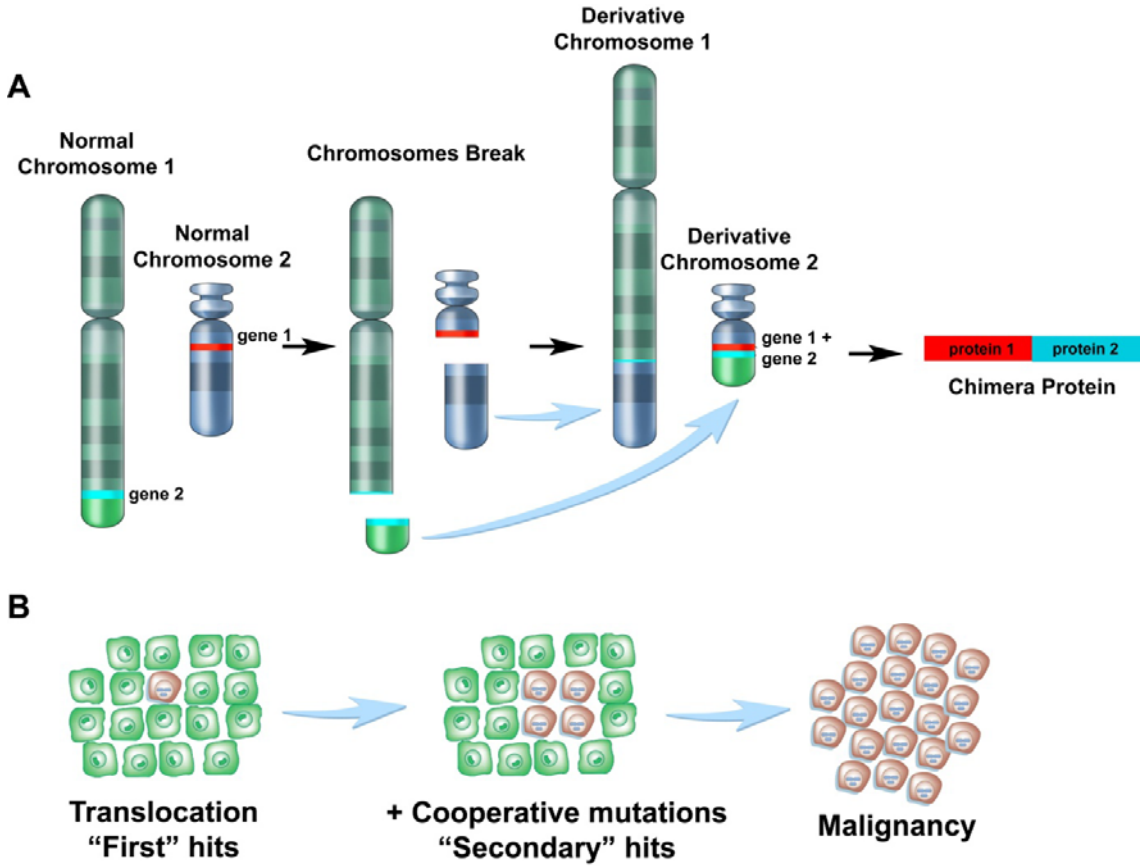


Figure 4-1. Balanced chromosome translocation and cancer cell malignancy initiated by chromosome translocation.

(A) Schematic cartoon of balanced chromosome translocation. During chromosome translocation, the chromosome breakpoints recombine to form chimeric chromosomes and express a chimeric fusion protein from fusion junction site. (B) The progression of cancer cell from harbouring a chromosome translocation to full malignancy.

The t(12;21)(p13;q22) translocation is the most common chromosomal translocation in paediatric cancers, occurring in approximately 25 % of cases of childhood pro-B cell acute lymphoblastic leukaemia (cALL) (137). The rearrangement results in the in-frame fusion of the 5' terminal of the ETS transcription factor TEL (also known as ETV6), to almost the entirety of the AML1 gene (also known as RUNX1). *AML1* encodes one of the DNA binding subunits of the core binding factor (CBF) and is related to the Drosophila gene *RUNT*. AML1 has been shown to play a role in regulating lymphoid and myeloid development (138). Clinical studies have found that the TEL-AML1 translocation occurs *in utero*, followed by a protracted time delay for leukaemia to develop (139). The disease has been recently identified to originate from a $CD34^{+}CD38^{-}$/low$CD19^{+}$ rare blood cell population (140).

The human *TEL* gene encodes a 452 amino acid ETS family transcription factor which was isolated as a fusion partner with the β chain of PDGF Receptor in t(5;12) chronic myelomonocytic leukaemia (135). In addition to its C terminus 85 amino acid ETS DNA binding domain, the N terminus of TEL contains a conserved interaction domain, which is responsible for oligomerization, and also for maintaining transcriptional activity (141). Apart from TEL-AML fusion-mediated malignancies, the TEL gene is involved in several 12p13 chromosomes translocations associated with a range of human malignancies, both as a N-terminal and a C-terminal fusion partner (142). AML1, like TEL, has been found to be rearranged with a number of different genes in leukaemogenic translocations including ETO, ETO-related MTG16 and EVI1 (143-145).

Although the formation of balanced chromosomal translocations is a frequent event in the pathogenesis of human malignancy, it is commonly believed the eventual formation of cancer normally requires additional mutations, or *'*secondary hits*'* (146) (**Figure 4-1 B**). In t(12;21) cALL, the *TEL-AML1* fusion was shown to occur in haematopoietic cells of a strikingly high proportion of live births (147). However, only a small fraction of these neonates go on to develop leukaemia (approximately 1/100), often with a long latent period, suggesting that although the *TEL-AML1* fusion may be acting as an initiating mutation, it is not sufficient to cause the disease. In addition, animal studies where the *TEL-AML1* fusion was expressed under the control of the IGH enhancer failed to cause leukaemia in mouse (148), indicating that additional mutations might be required to cooperate with the *TEL-AML1* fusion for cALL to develop. Research into the nature of mutations that can cooperate with *TEL-AML1* to cause cALL has yielded interesting results. In particular, it appears that the most common such mutation is a complete or partial deletion of the second allele of *TEL* which occurs in

approximately 70 % of cases (149). In addition, it has also been proposed that deletion or mutation of *Paired-Box-Containing Gene 5* (*PAX5)* participates with *TEL-AML1* in the induction of cALL (150) but again this has not been experimentally verified. The co-operating mutations that cause the most aggressive treatment refractory forms of cALL are yet to be identified.

Identification of mutations that cooperate with the *TEL-AML1* in the development of cALL remains far from complete, but will ultimately help to understand the mechanism of this disease and to find new treatments. One technique that can help to identify cooperative mutations is *in vivo* transposon-mediated insertional mutagenesis. In lower organisms and cell cultures, transposon systems have been routinely used for genetic manipulation and mutagenesis screens (55) (59,60). Recent publications have indicated transposons can also be engineered as mutagens in higher eukaryotes (47). The well characterized *Tc1*-like transposon *Sleeping Beauty* (SB) system consists of two components: the transposase, the enzyme responsible for mobilisation ('jumping'), and the transposon, the actual mobilised piece of DNA. It was recently shown that the SB transposon is an effective somatic insertional mutagen for studying oncogenesis and for identifying novel candidate cancer genes in mice (88,89). This was achieved through generation of two independent transgenic mouse lines, one harbouring chromosomal concatamers of the transposon DNA, and the other expressing SB transposase under the control of a ubiquitous promoter. The crossing of transposase and transposon mice yielded experimental mice in which the transposon is mobilised in the soma. The tumours were found to harbour transposon integration sites in both novel genes, as well as known human cancer genes (88,89).

4.2 Aims and summary of the project

In this chapter I plan to generate a *Tel-AML1* knockin mouse to model childhood pro-B cell acute lymphoblastic leukaemia (cALL), one of the most common paediatric cancers in human patients. The first stage of this project is to generate the model in mouse ES cells and validate the ES cell line *in vitro* before testing them *in vivo*. The second stage of the project is to model the cancer initiation and progression in the *Tel-AML1* mouse models, and combing this model with the *Sleeping Beauty* transposon system to identify *'*secondary hits*'*. I then aim to identify candidate genes and pathways underlining the cancer progression and formation process. Specifically the aims of this project are:

1. Validate the *TEL-AML1* knockin system i*n vitro* in ES cell for fusion transcripts and protein expression.

2. Cross *Tel-AML1* knockin mice with *T2/Onc* transposon mice (contains *Sleeping Beauty* transposon array) to initiate transposon mutagenesis and monitor the tumour formation by tumour watch.

3. Validate the tumour types using histology and FACS analysis

4. Combine the *TEL-AML1* system with the *Sleeping Beauty* transposon system to identify genes that when either inactivated or over-expressed represent *'*secondary hits*'* and cooperate with expression of the *TEL-AML1* to form ALL in mice.

A mouse model of *Tel-AML1* was generated by knocking-in the human *AML1* cDNA into the locus of *Tel* by DNA engineering (**Figure 4-2**), allowing expression of the fusion protein from the endogenous *Tel* promoter constitutively. A cDNA sequence encoding the *Sleeping Beauty* transposase was also knocked into the *Tel* locus after an internal ribosomal entry site (IRES), allowing deployment of transposon-mediated mutagenesis for screening the 'secondary hits'. The *loxP* sequences flanking the knockin cassette were originally designed to make a 'conditional' targeting construct. As the constitutive expressing *Tel-AML1* ES cells are viable and could be transmitted through the germ line, the conditional version was not used in this study. A backup model for *TEL-AML1* mouse model was also generated, where *TEL-AML1* was knocked into the Rosa 26 locus under a ubiquitous promoter. This mouse model will be further discussed in the results section.

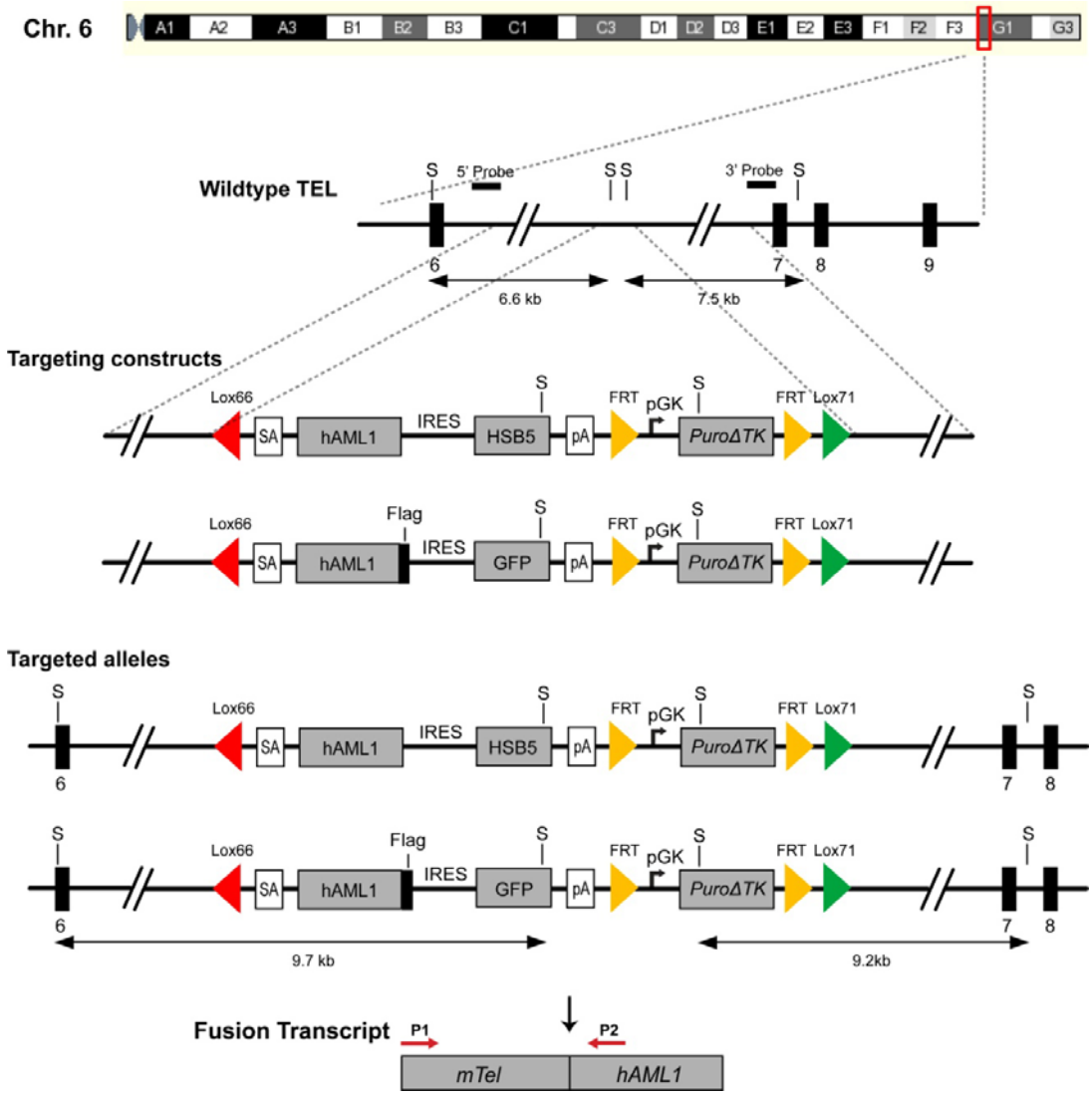


Figure 4-2. Schematic diagrams of *Tel-AML1* targeting constructs and targeted alleles

Both the SB transposon and GFP versions targeting constructs and targeted alleles are shown. The GFP version targeting construct is modified from the original SB version construct by adding a Flag tag sequence to the C-terminal of human AML sequence and replacing the HSB5 with a GFP sequence by homologous recombination. The GFP version construct was mostly used for *in vitro* validation experiments and the SB transposon version was used to generate knockin mice and for tumour watch experiment. S – *Stu*I restriction sites. The double head arrows below the schematic graph indicate the DNA fragment length in southern blot after *Stu*I digestion.

Characterization work (see results) showed that the *AML1* cDNA transcript is precisely fused with *Tel*, resulting in expression of *Tel-AML1* transcript from the *Tel* locus *in vivo* at levels approximating endogenous expression. By immunoprecipitation using an anti-Flag tag antibody, the fusion protein can be readily detected by western blotting. As part of the strategy to identify the 'secondary hits', the *Sleeping Beauty* transposase is also expressed from *Tel* locus and expression of the transposase results *'*jumping*' in vivo* in the experimental mouse after crossing of the *Tel-AML1* mouse with the *T2/Onc* transposon mouse.

The goal for *in vivo* study with this mouse model is to derive of disease mimicking human ALL and to identify cooperative mutations associated with *Tel-AML1*. The *Tel-AML1*$^{+/-}$ knockin mice were crossed with the *T2/Onc* transposon mice to generate experimental mice for a tumour watch study (**Figure 4-3**). After tumour generation in these mice the tumour types were analyzed by histological methods and FACS for disease identification. In the end DNA could be extracted from these tumours to identify cooperative mutations by splinkerette PCR (**Figure 4-3**).

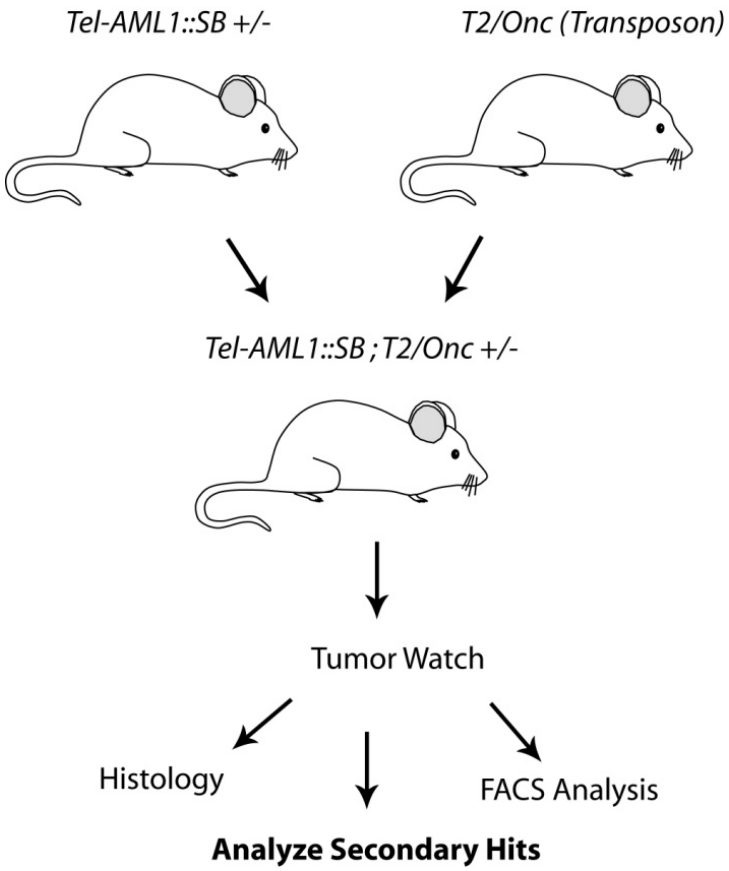


Figure 4-3. Crossing strategy for tumour watch and subsequent characterization studies in the *Tel-AML1* mouse model.

The tumour watch was initiated by crossing *Tel-AML1* knockin mouse with the *T2/Onc* transposon mouse to activate transposition and derive leukaemia formation in the experimental mice. The tumours generated were subjected to histology and FACS analysis to identify the tumour types. These tumours will subsequently been analyzed to identify secondary mutations using splinkerette PCR.

4.3 Materials and Methods

4.3.1 Targeting construct generation

To generate the *Tel-AML1* targeting construct (**Figure 4-2**), a genomic fragment of *Tel* was cloned into pBlueScript SK+ vector by gap repair from a 129S7 BAC clone (bMQ-66F22) using two homologous arms (PCR primers for 5' arm: FWD: 5'-GACAAAGTAGATGG CACCAGTGCAGTG-3', REV: 5'-GATGAGTGGTCAGGGGGGCAAAGAAGGAAAAA AAACCTTACAGAAA-3' ; 3' arm: FWD: 5'-GGTTGAAGGGCAGAGCTCTAGTGTCAA TTTG-3', REV: 5'-GGCTGGAGGGCAAACCAGGTACCATTACAGCAC TAGAAACCA GAGA-3') of 497 bp each. The AML1-SB-Puro cassette and the Lox66 and Lox71 sites were synthesized by GENEART. This cassette was inserted as a *Hpa*I fragment into a *Stu*I site within the *Tel* genomic fragment. This targeting construct was subsequently modified by molecular cloning to add a Flag tag to the end of *AML1* sequence: First a Flag tag sequence was introduced at the C-terminus of AML1 with primers: FWD: 5'-GCTCGCCGCCGCGCATCCT-3'; REV: 5'-GGCCTTAATTAATCACTTGTCGTCATCGTCCT-3'. The PCR product was then digested overnight with restriction enzyme *Pac*I. A subsequent double digestion of the targeting construct was performed overnight with *Afe*I/*Pac*I, and products were ligated at 4°C for 16 hours. The resultant plasmid was transfected into *E. Coli*, and colonies picked for PCR screening. Positive clones were verified by sequencing.

To generate a GFP version construct of the *Tel-AML1* I a *ccdB* negative selection marker was cloned by PCR and used to exchange the transposase sequence by recombineering. Two *Sbf*I restriction sites were introduced on the two ends of the *ccdB* cassette. A GFP sequence was cloned by PCR with *Sbf*I restriction ends and was used to exchange the ccdB fragment by *Sbf*I digestion and ligation. All constructs were sequenced in full to ensure that PCR had not introduced any mutations. PCR primers:

CCDB: 5'-CGGGGACGTGGTTTTCCTTTGAAAAACACGATGATAATATGGCCACAA CCCCTGCAGGGCATTAGGCACCCCAGGCT TTACAC-3' (FWD) and 5'-TAGATGCA TGCTCGA GCGGCCGCCAGTGTGATGGATATCTGCAGAGAATTCCTGCAGGTGCA GACTGGC TGTGTATAAGGGAG-3' (REV)

GFP: 5'-GTTTTCCTTTGAAAAACACGATGATAATATGGCCACAACCCCTGCAGG ATGGTGAGCAAGGGCGAGGAGCTGT-3' (FWD) and 5'-CTCGAGCGGCCGCCAGTG

TGAT GGATATCTGCAGAGAATTCCTGCAGGTTACTTGTACAGCTCGTCCATGC CG-3' (REV).

4.3.2 ES cell transfection and selection

For transfection of the targeting construct into embryonic stem cells, wild type E14 cells (Strain Name: 129P2 Ola) were fed in M-15 medium till 70-80 % confluent. To harvest cells, one culture dish of cells (~3×10^7 cells) was washed twice with phosphate buffered saline (PBS) and then treated with 3 ml Trypsin (1×) at 37°C for 7 minutes. The trypsin reaction was stopped by adding 10 ml M-15 medium and dissociated cells were spun down at 400×g for 4 min. The supernatant was removed and cells were washed two times with PBS. After the second wash, cell pellet was dissolved in PBS to give a final concentration of 1.4×10^7 cells/ml. Plasmid DNA was prepared using a JETstar Plasmid Purification Maxi Kit (Cat. No. 220020) and linearized with *Pvu*I restriction enzyme prior to electroporation. 0.9 ml of the cell suspension was added into a 0.4 cm BIO-RAD Gene Pulser Cuvette. 40 µg plasmid DNA was mixed well with cells in the cuvette. Electroporation was performed at 800 V, 25 mF with a BIO-RAD Gene Pulser II Electroporator. Cells were rested for 5 minutes after electroporation and were plated on 10 cm culture dish with confluent feeder cells. Cells were cultured in M-15 medium for two days, and then 3 µg/ml puromycin added to the medium to allow selection over a further 10 days of growth. Colonies were picked on day 12 and correctly targeted clones were screened for by southern blot analysis.

4.3.3 Generation of pMSCV expression constructs

The pMSCV-GFP expression vector was constructed by inserting an IRES2-EGFP sequence into the *Bgl*II restriction site of pMSCVneo (Clontech) downstream of the 5' LTR. The human *TEL-AML1* sequence was amplified from a GFP-*TEL-AML1* construct described previously (151) using the following primers (5' primer: 5'-GGCCGAATTCATGTCTGA GACTCCTGCTCA-3'; 3' primer: 5'- AAGATCTTCACTTGTCGTCATCGTCCTTGTAGT CCCGCGGGTAGGGCCTCCACACGGCCT-3'), thus incorporating a Flag tag at the C-Terminus of *TEL-AML1*. These constructs were sequenced in full to ensure that PCR had not introduced any mutations. The two mouse alternative initiation transcripts of *Tel-AML1* sequences (M1 and M43) were amplified from cDNA of targeted ES cells and were cloned into the pMSCV-human Tel-AML1-GFP construct using *EcoR*I and *Bcl*I restriction sites to insert the mouse *Tel-AML1* sequence into the vector (**Figure 4-5 A**).

4.3.4 Immunocytochemistry

The immunostaining using anti-Flag M2 antibody (Sigma, F1804) was performed following manufacturer's instructions (Sigma). Briefly, human 293T embryonic kidney (H293T) cells were seeded one day before transfection on cover slips in 10 cm dishes. The next day 10 µg plasmid DNA for each expression construct were transfected into H293T cells by calcium phosphate transfection using a ViraPack transfection reagent (Stratagene). Two days after transfection, cover slips were fixed in 6 well plates with 4 % paraformaldehyde in PBS (Sigma) with 4 % sucrose (BDH). Cells were then permeabilized with 0.25 % Triton X-100 (Sigma). Blocking was performed using 10 % bovine serum albumin (BSA) in PBS for 30 min at 37 ˚C. Coverslips were then incubated with ANTI-FLAG M2 antibody (1:1000 dilution) in 3 % BSA/PBS for 2 hours at 37 ˚C in a humidified chamber. After three washes in PBS, cells were incubated with Alexa Fluor 568 anti-mouse secondary antibody (Molecular Probes) for 45 min at 37 ˚C and mounted in VECTASHIELD (Vector) with DAPI. Slides were dried at room temperature and analyzed by fluorescent microscopy.

4.3.5 Immunoprecipitation of Flag Tagged Proteins

The Dynabeads protein G (Invitrogen, 0.5 ml beads) were first washed three times (buffer: 24.5mM Citric Acid, 51.7 mM Dibasic Sodium phosphate (Na_2HPO4) dehydrate, pH = 5.5). One microgram of anti-Flag M2 antibody (Sigma, F1804) was incubated with the beads in 20 µl of bead wash buffer for 40 min at room temperature. After incubation beads were washed three times with beads wash buffer with 0.1 % Tween-20 (Sigma). To prepare a cell lysate, cells were treated with protein lysis buffer (50 mM Tris pH 8.0, 450 mM NaCl, 0.2 % Nonidet P-40 (Igepal), 1 mM DTT, 1 mM EDTA, 1X Protease inhibitor (Roche) for 15 min on ice, then collected by centrfiguation at maximum speed using a desktop centrifuge for 15 min at 4 ˚C. Cell lysate was collected and incubated with the antibody conjugated beads for 1 hour at 4 ˚C with gentle shaking. The beads were collected after incubation and washed three times with protein wash buffer (same formula as protein lysis buffer except using 150 mM NaCl and 0.1 % NP-40 concentration). For Western blotting, 30 µl loading buffer were added to the beads after pull down. The beads were then boiled for 10 min at 95 ˚C and supernatants were loaded directly on SDS page gels (5 %, Bio-rad). Western blotting was using anti-Flag M2 antibody (Sigma, F1804) and performed following manufacturer's instructions from Sigma.

4.3.6 RNA isolation and cDNA preparation

The total RNA from cell culture or mouse tissue was isolated using TRIZOL reagent (Invitrogen, Cat. No. 15596-018) and manufacturer's standard protocol. Briefly, tissue samples collected from mice were crushed through a 70 □m nylon cell strainer (BD Falcon, Cat. No. 352350) in PBS to create a single-cell suspension. The cells were then pelleted (1,500 rpm for 5 min at 4 ˚C) and resuspended in 1 ml TRIZOL reagent. Cell pellet was lysed in TRIZOL for 5 minutes at room temperature and 0.2 ml chloroform added. After vigorously shaking the tube was stood at room temperature for approximately 5 minutes and then spun down by centrifuge at 12,000 g for 15 minutes at 4 ˚C. After centrifugation, the upper aqueous phase was carefully transferred into a new tube. The RNA was precipitated by adding 0.6 ml isopropyl alcohol and centrifuged at 12,000 g for 10 minutes at 4 ˚C. The RNA pellet was washed by 75 % ethanol and allowed to dry at room temperature. RNA was dissolved in 100 µl H_2O and stored at -20 ˚C for quantitative PCR. The first strand total cDNA was reverse transcribed using a SuperScript First-Strand RT-PCR kit from Invitrogen (Cat. No. 11904-018) following the Random Hexamers first-stand synthesis protocol from the manufacturer.

4.3.7 Quantitative PCR

The quantitative PCR was performed using an ABsolute™ QPCR ROX Mix kit (Cat. No. AB-4138/B) according to previously described procedure with slight modifications (152). The probe was ordered from MWG Operon which was labelled with FAM at 5' end and TAMRA at 3' end. The probe was first diluted in ddH_2O to 4 pmol/µl and then master mix was prepared using following formula per reaction: QPCR Mix kit 12.5 µl; Forward/Reverse Primer (100 nM) 0.25 µl each; Probe (4 pmol/µl) 0.5 µl; ddH_2O 6.5 µl. The reactions were set up in a standard 96 well plate. 20 µl master mix was added into each well with 5 µl reverse transcribed total cDNA. Two parallel reactions were setup for each cDNA sample, and 5 µl ddH_2O was mixed with 20 µl master mix as negative control. The plate was sealed using a MicroAmp Optical Adhesive Film (ABgene, 4311971) and PCR reaction was performed on the ABI PRISM 7900HT sequence detection system (Applied Biosystems). Thermal cycling was initiated with an incubation step at 50 °C for 2 min, followed by a first denaturation step at 95 °C for 15 minutes, and continued with 40 cycles of 95 °C for 15 seconds, 60 °C for 1 minute. The fluorescent signal representing the transcripts expression level was normalized with the β-Actin control. Probe and primer sequences were as follows:

Tel-AML1 Probe: 5'-AGCACGCCATGCCCATTGGG-3';
FWD Primer: 5'-CTTGAACCACATCATGGTCTCTATG-3';
REV Primer: 5'-TCGTGCTGGCATCTGCTATT-3'.
Tel Probe: 5'- CACGCCATGCCCATTGGGAGAA-3';
FWD Primer: 5'-TCTCTATGTCCCCACCGGAAG-3';
REV Primer: 5'-CATAATCCCAAAGCAGTCTACAGTCT-3'.
β-Actin Probe: 5'-TTTGAGACCTTCAACACCCCAGCCA-3';
FWD Primer: 5'-CGTGAAAAGATGACCCAGATCA-3';
REV Primer: 5'-CACAGCCTGGATGGCTACGT-3'.

4.4 Results

4.4.1 Generating the *Tel-AML1* knockin mouse model and characterizing the targeted ES cells

A *Tel-AML1* knockin targeting construct was first generated by conventional DNA cloning and recombineering in mouse ES cells. In the targeting construct, the human *AML1* cDNA was inserted into intron 6 of mouse *Tel* allele (Ensembl transcript ID: ENSMUST00000111963) so that a *Tel-AML1* fusion would be generated by splicing of the *Tel* transcript to *AML1* from the endogenous mouse *Tel* locus (**Figure 4-2**). The hyperactive *Sleeping Beauty* transposase variant HSB5 (provided by Steven Yant, Stanford) was cloned behind the *AML1* cDNA after an internal ribosomal entry site (IRES). This design results in the production of a bicistronic messenger composed of the *Tel-AML1* fusion and the *Sleeping Beauty* transposase (**Figure 4-2**). After introducing this construct into ES cells, targeting was analysed by southern blot analysis, with successfully targeted constructs having a 9.7 kb band (**Figure 4-4 A**). Using RT-PCR (Figure 4-2, primers P1 + P2) the expression of fusion transcript could be detected as two splicing variants (**Figure 4-4, B**), representing the alternative splicing forms M1 and M43 from *Tel* exon 2 (**Figure 4-4, D**). Sequencing results subsequently confirmed a precise in-frame fusion between *Tel* and *AML1* (**Figure 4-4, C**). The targeted ES cells were injected and the allele had been successfully transmitted through the germ line.

4.4.2 Detection of the Tel-AML1 fusion protein by immunoprecipitation

Although the *Tel-AML1* fusion transcript was easily detected from the knockin ES cells, at the protein level the fusion protein could not be detected by western blotting in cell lysate using either the anti-TEL or anti-AML1 antibody (data not shown). This may be due to very low level expression of the protein from the endogenous *Tel* locus, or alternatively that the fusion protein is not stable in ES cells.

The fact that the *Tel-AML1* fusion transcript but not the protein could be easily detected, a more sensitive detection method was required. Since only the presence of the oncogenic fusion protein itself would be the direct validation for the mouse model the original *Tel-AML1* targeting construct was modified by recombination to add a Flag tag sequence after the *AML1* cDNA (**Figure 4-2 A**) so that a highly-specific anti-Flag antibody could be used to detect the fusion protein. After targeting of this construct into ES cells and concentration of the fusion protein from cell lysate by immunoprecipitation (see Materials and Methods), the Flag tagged Tel-AML1 was successfully detected by western blotting. Two splicing isoforms were identified at ~100 and ~110 KDa (**Figure 4-2 E**) in two targeted cell lines (BM2 and BM3), representing the translation started from M1 and M43 alternative start codons, respectively (153).

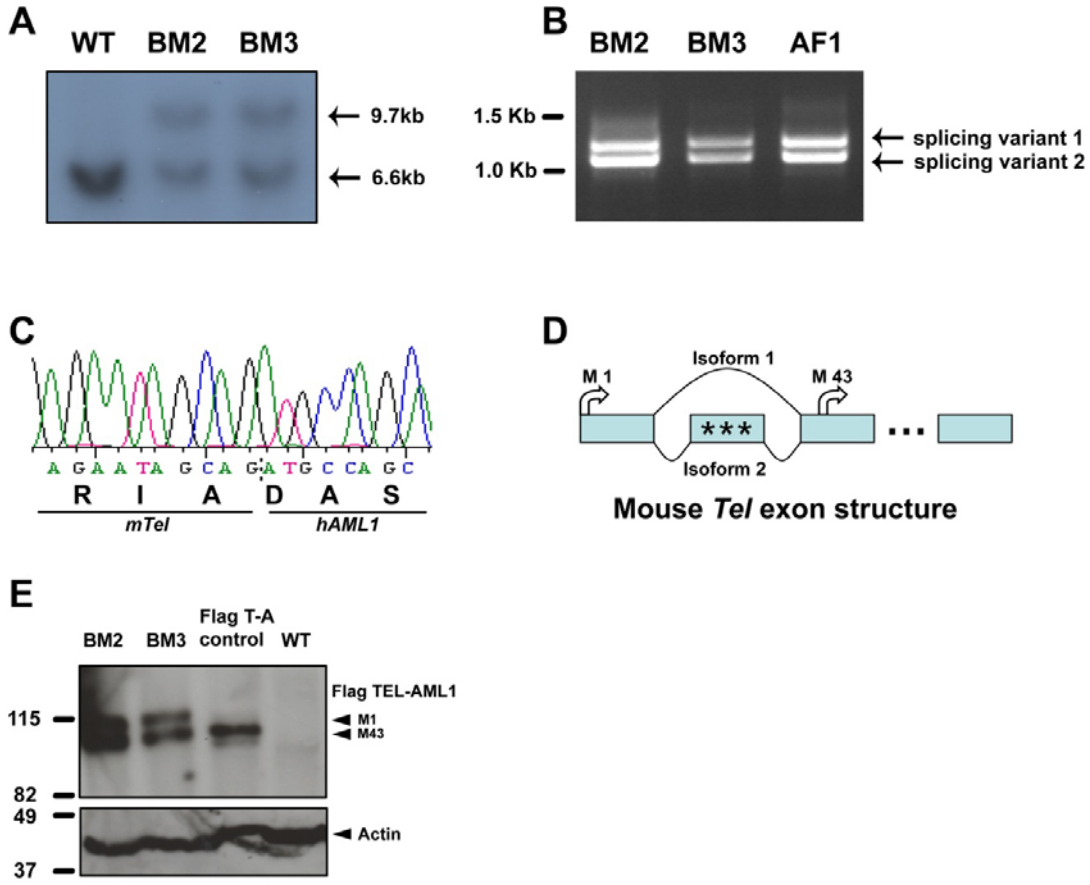


Figure 4-4. Characterization of *Tel-AML1* targeted ES cells

(A) Southern blotting detected a targeted band at 9.7 kb for two targeted ES cell clones (BM2 and BM3), in addition to the wild type band at 6.6 kb. (B) RT-PCR detected two splicing variants in three targeted ES cell clones. The clone AF1 is the actual injected clone used to derive the F1 mice. (C) Sequencing trace for the RT-PCR product showing in-frame fusion between *Tel* and *AML1* transcripts. (D) Schematic diagram of mouse Tel exon structure. Two alternative start codons (M1 and M43) are indicated. Three stop codons on exon 2 are represented with *'*'*. (E) Western blot showing two isoforms of FLAG tagged TEL-AML1 fusion proteins (M1 and M43) detected in two targeted ES cell clones after immunoprecipitation. The Flag T-A positive control is the *in vitro* expressed TEL-AML1-FLAG protein, and WT is the E14 wild type ES cells which served as a negative control.

4.4.3 *In vitro* characterization of the mouse TEL – human AML1 fusion protein

Previous studies have suggested that expression of the human TEL-AML1 is oncogenic in both mouse and zebrafish (151,154,155), however no study has ever characterized the mouse TEL-human AML1 fusion protein which is expressed in our mouse model. To prove that the mouse-human fusion is biologically identical to human TEL-AML1, localisation studies of the mouse and human TEL-AML1 fusions were performed, since localization study could be a good indication to protein's biological function. The full length mouse version transcripts, including both the M1 and M43 splicing variants, were cloned from targeted ES cells and inserted into the pMSCV-GFP expression vector (**Figure 4-5, A**). The human version of TEL-AML1 was generated as a positive control and an empty GFP vector was used as a negative control. All the fusion proteins contained a Flag tag at the C terminus and could be detected using an anti-FLAG antibody on a Western blot after transfection into human embryonic kidney 293T cells (**Figure 4-5, B**). The localization of these fusion proteins was examined using immunocytochemistry. Both the human TEL-AML1 and two isoforms of mouse-human TEL-AML1 were localized in the cytoplasm in 293T cells (**Figure 4-5, C**). This localization for human TEL-AML1 has been suggested from previous reports (156). In addition, no obvious localization difference between the M1 and M43 isoforms can be seen (**Figure 4-5, C**). Although this localization study is not a directly proof of the function, it demonstrates that the fusion transcript expressed in the knockin mouse can give rise to a stable protein which shares the same localization pattern with the human fusion protein in 293T cells.

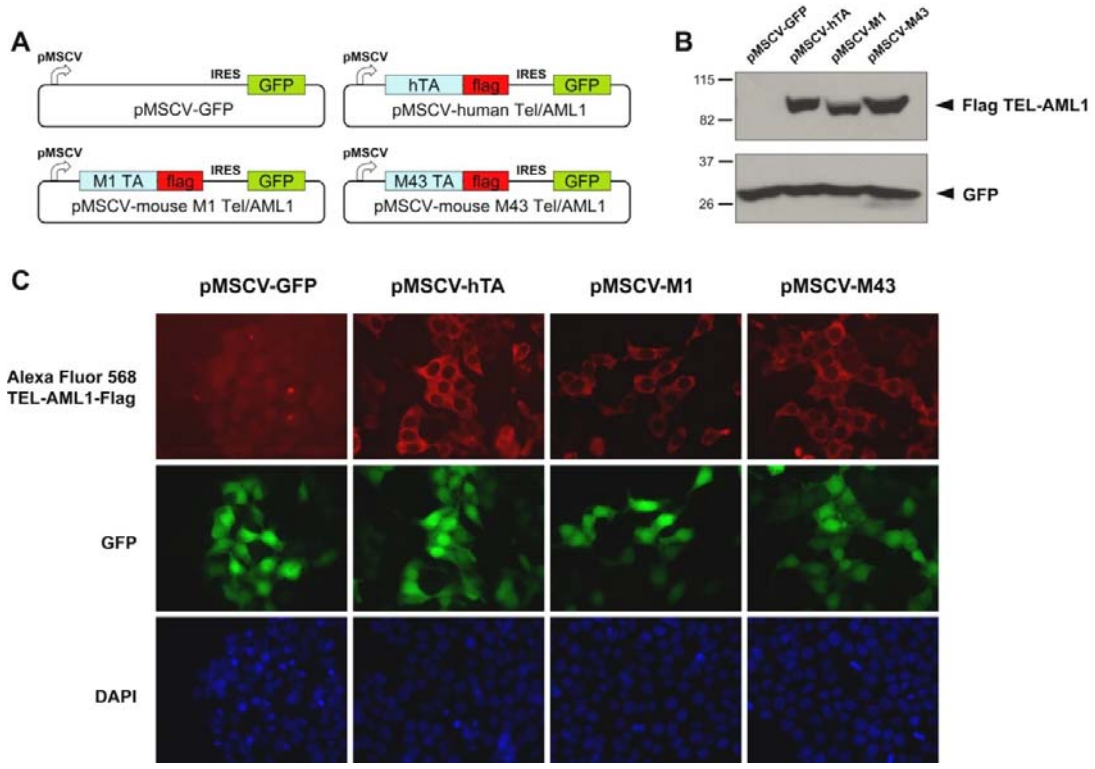


Figure 4-5. Characterization of the mouse-human TEL-AML1 protein by *in vitro* studies

(A) Generation of the pMSCV overexpression constructs for expressing TEL-AML1 fusion proteins. The mouse Tel-AML1 was expressed as two isoforms (M1 and M43). Human TEL-AML1 was used as a positive control and the original pMSCV-GFP construct was used as a negative control. (B) Western blot to verify expression of the fusion proteins in 293T cells using anti-FLAG antibody. (C) Immunostaining in 293T cells transfected with empty vector, human TEL-AML1 or two isoforms of mouse Tel-AML1 (M1 and M43).

4.4.4 Analysis of the *Sleeping Beauty* Transposon system in the knockin mouse

In our *Tel-AML1* mouse model, a transposon mediated insertional mutagenesis system was designed to incorporate with TEL-AML1 expression as a combined strategy to identify the cooperative mutations associated with formation of the *TEL-AML1*. To achieve this, a hyperactive variant of the *Sleeping Beauty* transposase (HSB5) was introduced into the *Tel-AML1* knockin mouse (**Figure 4-2**). *In vitro* tests were first performed to characterize the transposon system before *in vivo* application. By western blotting, the SB transposase protein expression could be detected in targeted ES cells (**Figure 4-6, A**). To determine if the transposon could be mobilized *in vivo*, a splicing PCR was performed on mouse tail DNA using primers flanking the SB transposon. A 225 bp 'jumping' band was detected on the electrophoresis gel with the *Tel-AML1::SB* $^{+/-}$ mouse crossed with the *T2/Onc* transposon mouse (88) but not the *T2/Onc* transposon mouse alone (**Figure 4-6, B**). These results indicated that the transposon system is active in our *Tel-AML1* knockin mouse.

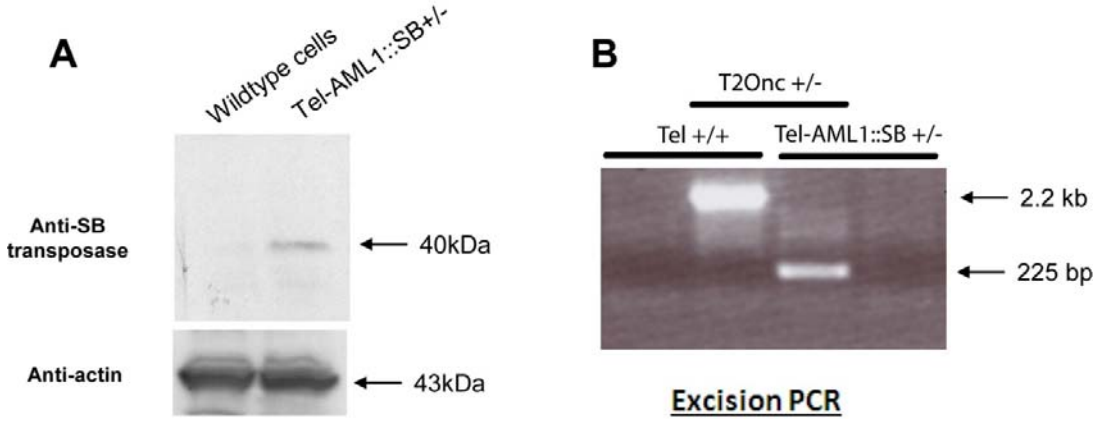


Figure 4-6. Characterization of the transposon system in *Tel-AML1* mouse model

(A) Western blotting on *Sleeping Beauty* transposase in the targeted *Tel-AML1* ES cells showing a band at 40 kDa representing the transposase protein. (B) Excision PCR for detecting transposon 'jumping' in the knockin mice using primer pairs surrounding the transposon sequence that has been described (88). A 225 bp band could be amplified in the knockin mouse crossed with *T2/Onc* transposon mouse indicating the transposon has been mobilized *in vivo* (lane 3). The *T2/Onc* transposon mouse along only generated a 2.2 kb 'unjumped' band (lane 2). The knockin mouse DNA and wild type mouse DNA was served as negative control and generated no band (lane 1 and 4).

4.4.5 Validation of the *Tel-AML1* expression level by real-time qPCR

The *in vitro* validation experiment in targeted ES cells detected the expression of the correct fusion transcripts from endogenous *Tel* locus, but was not a quantitative analysis. A real-time qPCR experiment was carried out to quantify and compare the expression level of the *Tel-AML1* transcript with the wild type *Tel* expressed from the alternate allele (**Figure 4-7 A**). The experiment was performed on hematopoetic tissues (bone marrow, spleen, thymus) as they are mostly relevant to the disease in question. To detect the fusion transcript expression two-week old mice were sacrificed and tissues were taken to prepare RNA and cDNA. The cDNA sample were first validated by Reverse Transcriptase PCR (RT-PCR) for the presence of fusion transcript in three hematopoetic tissues in the knockin mice (**Figure 4-7 B**). The real-time qPCR was performed using an ABsolute™ QPCR ROX Mix kit following the protocol described (see Materials and Methods). The expression level of *Tel-AML1* and *Tel* cDNA were normalized with β-Actin expression level and compared. From the results the *Tel-AML1* is expressed at a lower but within a comparable range (less than one fold difference in expression level) with the *Tel* transcript expressed from the other allele in the knockin mice (**Figure 4-7 C**). This result indicates that the knockin allele strategy was working *in vivo*, resulting in expression of comparable amount of *Tel-AML1*.

The expression level of *Tel* transcript in five knockin mice and five wild type mice was compared. Although *Tel* was expressed from only one allele in the knockin mice rather than from both alleles in the wild type mice, the expression level was similar indicating that there might be a compensation mechanism for *Tel* expression in the knockin mice (**Figure 4-7 D**).

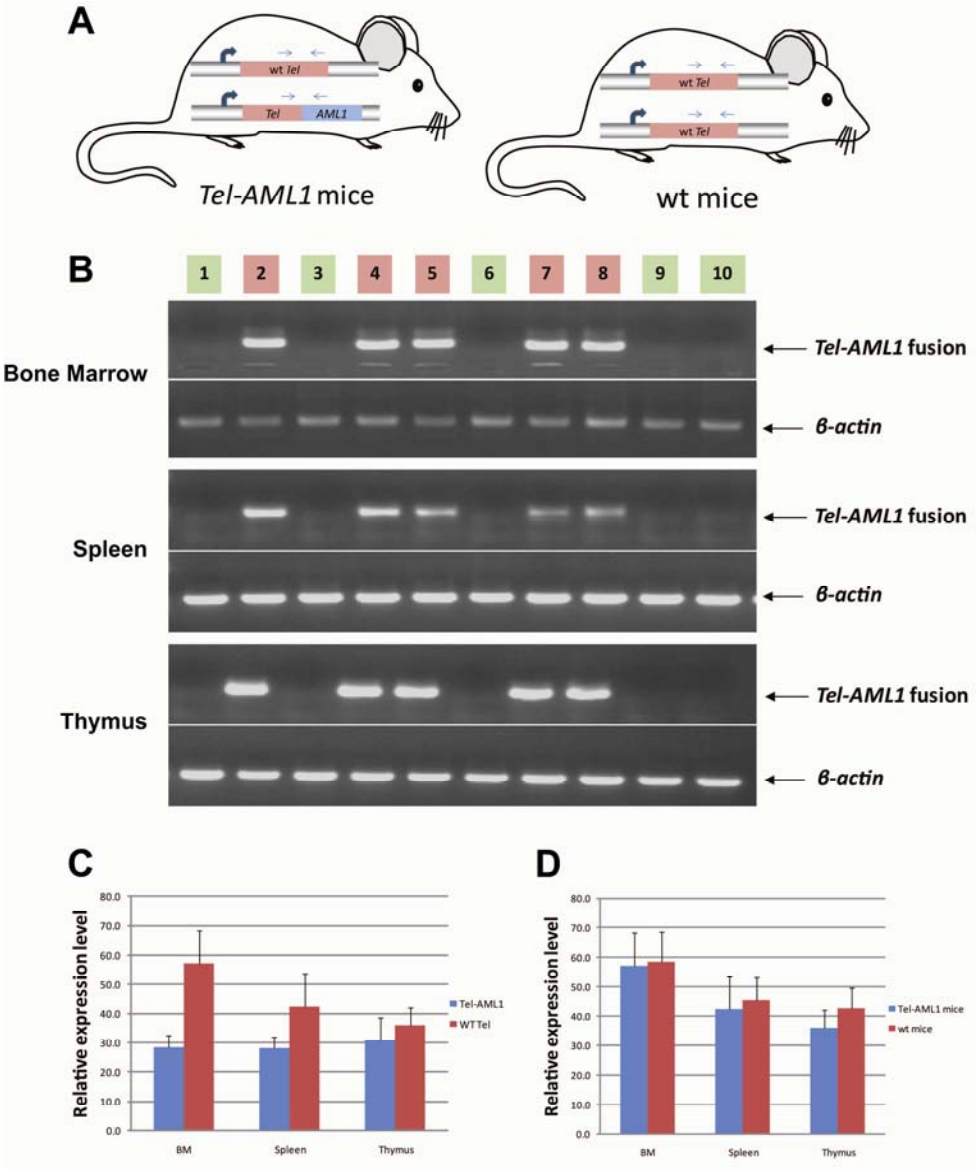


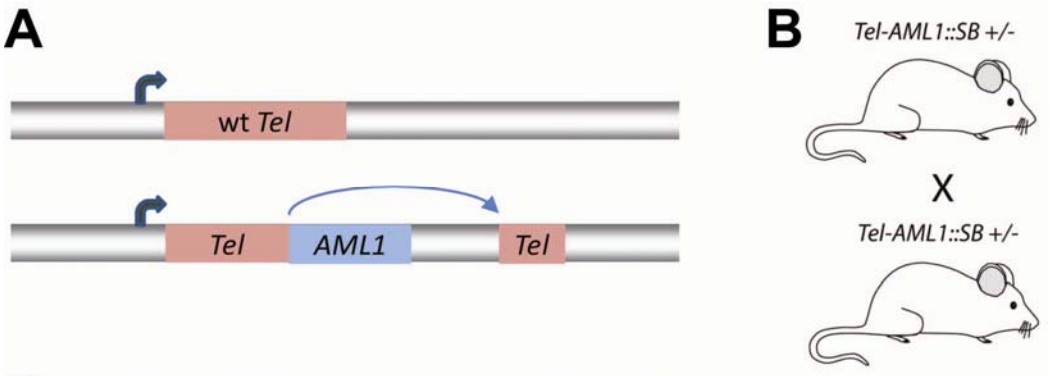
Figure 4-7. *In vivo* validation of the *Tel-AML1* expression strategy and real-time qPCR

(A) Genotypes of the *Tel-AML1* mice and wild type mice. The arrows represent the primer pairs used for RT-PCR analysis; (B) Detection of expression of *Tel-AML1* fusion transcripts by RT-PCR in *Tel-AML1* knockin mice (n = 5) and wild type mice (n = 5). The numbers in red box represent the *Tel-AML1* mouse number and in green box represent the wt mouse number; (C) Comparison of the relative expression level of the *Tel-AML1* transcript and wild type *Tel* transcript in *Tel-AML1* knockin mice (n = 5); (D) Compare the relative expression level of wild type *Tel* transcript in *Tel-AML1* mice (n = 5) and wild type mice (n = 5). The relative expression level was normalized to β-Actin.

4.4.6 Analysis of cryptic splicing in *Tel-AML1* knockin mice

One of the potential problems in designing a knockin experiment is cryptic splicing over the knockin allele i.e. part or full length of the knockin transcript could not be properly spliced. In *the Tel-AML1* knockin, the human *AML1* sequence was inserted into the *Tel* intron sequence between exon 6 and 7, leaving the 3' *Tel* sequence intact. Theoretically the splicing could jump over the *AML1* sequence and splice into the endogenous 3' *Tel* splicing site to generate wild type *Tel* transcript from the knockin allele (**Figure 4-8 A**). To assess cryptic splicing in the knockin mice, two heterozygous mice were crossed to generate homozygous embryos (**Figure 4-8 B**). According to Mendelian principles of inheritance this cross would generate progenies with three genotypes for *Tel-AML1* (+/+, +/-,-/-) at a rate of 1:2:1. However, genotyping for 95 live born mice resulted in no live born homozygous *Tel-AML1* mice (**Figure 4-8 C**), indicating an embryo lethal phenotype associated with the *Tel-AML1*$^{-/-}$ embryo.

To confirm the embryo lethality of *Tel-AML1*$^{-/-}$ mice, embryos were examined 10.5 days into pregnancy. The double knockin embryo, if no cryptic splicing taking place, should be identical to homozygous *Tel* knockout embryos which are embryonic lethal at E10.5 day (157). 27 embryos were harvested at E10.5 day and a ratio of 8:13:6 was obtained from three genotypes, close to the Mendelian ratio of 1:2:1 (**Figure 4-8 C**). From the embryo imaging all the wild type and heterozygous embryos have quite typical E10.5 day embryo morphology but the homozygous *Tel-AML1* embryos were smaller than the wildtypes and heterozygotes and some were under degradation indicating the embryo lethality (**Figure 4-8 D**). These results confirmed that the *Tel-AML1* knockin allele should be non-permissive for cryptic splicing.



C ***Tel-AML1* Intercross Results**

	$Tel^{+/+}$	$Tel^{+/AML1}$	$Tel^{AML1/AML1}$
Genotype at E10.5d	8	13	6
Genotype at live born	45	50	0

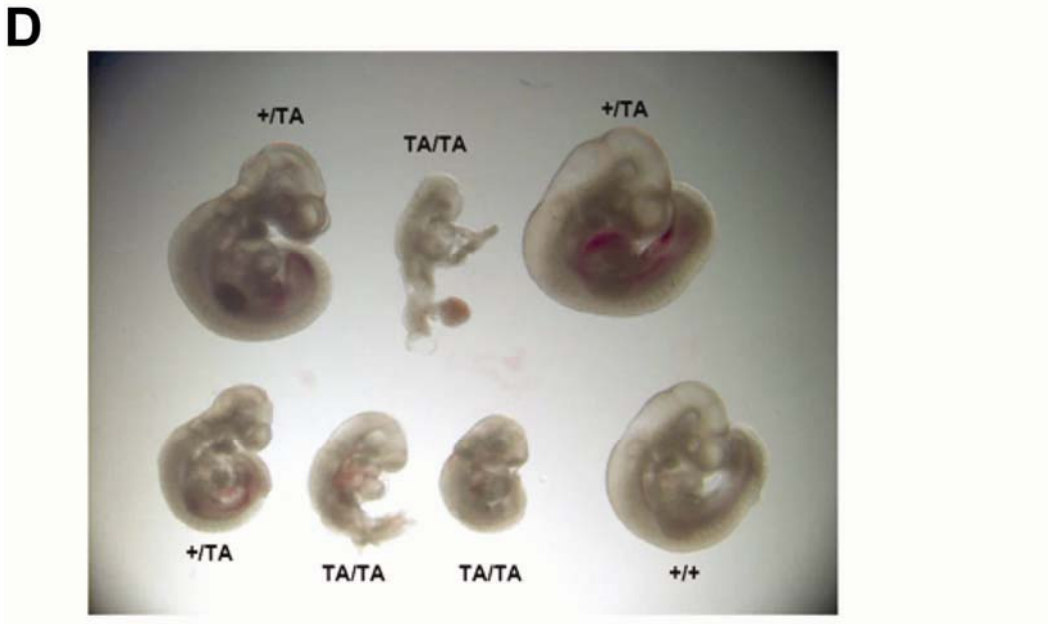


Figure 4-8. Investigation of cryptic splicing in *Tel-AML1* knockin mice by intercrossing

(A) Cartoon of the wild type and knockin alleles in the *Tel-AML1* knockin Embryo and heterozygous mice. Potentially, the splicing could jump over the AML1 knockin sequence and splice into the endogenous 3' Tel sequence. (B) The strategy of the intercrossing experiment. (C) Results of the intercrossing experiment showing the number of embryo (mouse) obtained at two different stages (Embryo day 10.5 and live born). (D) Morphology of the embryo at E10.5 day with genotypes indicated.

4.4.7 Tumour watch study in *Tel-AML1* knockin mice

To investigate the oncogenic property of *Tel-AML1* for acute lymphoblastic leukaemia (ALL) in the knockin mouse model, the *Tel-AML1*$^{+/-}$ mice were crossed with T2Onc$^{+/-}$ transposon mice to generate transposon 'jumping' mice (*Tel-AML1*$^{+/-}$; T2Onc$^{+/-}$ or *TAOnc*, *n* = 90) (**Figure 4-3**) or non-jumping mice (*Tel-AML1*$^{+/-}$; T2Onc$^{+/+}$ or *TA*, *n* = 54). The progenies of other two cohorts without *Tel-AML1* knockin (*Tel-AML1*$^{+/+}$; T2Onc$^{+/-}$ or *Onc* (n = 50), *Tel-AML1*$^{+/+}$; T2Onc$^{+/+}$ or *WT*, *n* = 22)) from the same cross were served as control groups. The experiment and control cohort mice were raised in an animal facility receiving standard care and routine check on a daily basis. Once the phenotype of sickness or cancer was identified, mice were sacrificed for histological and haematological studies for disease identification. In the 'jumping ' cohort mice (*TA*$^{+/-}$,*T2Onc*$^{+/-}$) tumour progression started as early as 25 weeks and the population also declined the quickest among the four cohorts as depicted in a Kaplan-Meier curve (**Figure 4-9 A**). All the 'non-jumping' cohorts have very similar curve patterns with each other except the *Tel-AML1* knockin cohort had slightly higher accelerated death rate than the other two control cohorts. This result indicates that the *Tel-AML1* knockin and *Sleeping Beauty* transposon system co-ordinately accelerated the tumour progressing.

The tumours isolated from these mice were subsequently identified by Dr. Brian Huntly - a haematologist from Cambridge Addenbrooke's Hospital following standard Bethesda criteria for lymphoid and non-lymphoid murine malignancies. Control mice in the *Onc* and *WT* cohorts occasionally presented with acute leukemias (10/72 mice, 14 %), however the incidence was significantly increased in the cohorts expressing *Tel-AML1* (73/90 mice, 83 % for the *TAOnc* cohort and 13/54, 24 % for *TA*) (**Figure 4-9 B**). Among the *Tel-AML1* knockin cohorts, a significant proportion of *TAOnc* mice (15/73, 21 %) developed B cell progenitor (BCP)-ALL, as did a slightly smaller proportion of the TA cohort (2/16 cases, 13 %). Importantly, BCP-ALL disease was only seen in mice carrying the *TEL-AML1* allele. FACS plots from the bone marrow of a representative mouse demonstrate only background Gr-1/Mac1 myeloid cells, with the majority of cells having a B220^{+}/CD19^{-}/sIg^{-} phenotype, in keeping with BCP-ALL (**Figure 4-9 C**). The histology study showed the presence of lymphoblasts in the peripheral blood, as is the infiltration of the spleen, bone marrow and liver, with effacement of the normal cellular architecture and replacement by nucleolated blasts (**Figure 4-9 D**). These results showed the *Tel-AML1* knockin mouse model could be able to derive BCP-ALL, i.e. the modelling strategy was successful. For the other types of

leukaemia identified from *TAOnc* mice, acute myeloid leukemia (AML) is predominant (34/75 cases, 44 %), with T-cell ALL also seen (14/75, 19 %) (**Figure 4-9 B**).

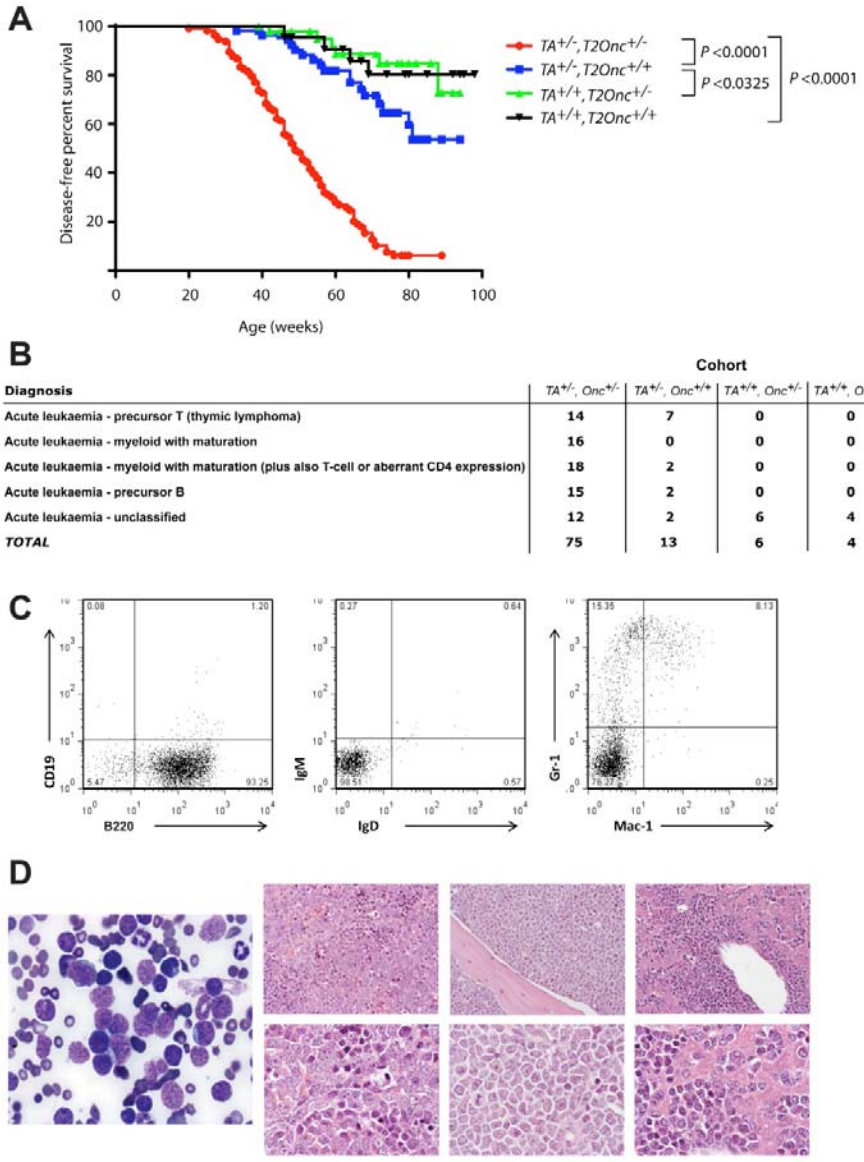


Figure 4-9. Tumour progression and hiostological analysis in *Tel-AML1* knockin mice

(A) Kaplan-Meier curves showing the tumour latency of 'jumping' Tel+/AML1; T2Onc+/- mice and 'non-jumping' control mice. (B) Classification of the malignancies developed by mice according to the Bethesda criteria for lymphoid and non-lymphoid murine malignancies (158,159). (C) FACS analysis from the bone marrow of a representative *Tel-AML1*$^{+/-}$ mice where transposon mutagenesis resulted in development of B cell precursor acute lymphoblastic leukemia; only background Gr-1/Mac1 myeloid cells and a majority of B220+/CD19-/sIg- cells are detected, in keeping with B-ALL. (D) Peripheral blood, spleen, bone marrow and liver from a representative mouse with B-ALL, showing phenotypes recapitulating features of the human disease.

4.4.8 An alternative mouse model of TEL-AML1

As an alternative to the knockin mouse model, I also generated a *'*conditional*'* *TEL-AML1* mouse model where a GFP tagged TEL-AML1 is expressed from the Rosa26 locus and also bicistronically the HSB5 *Sleeping Beauty* transposase (**Figure 4-10**). This mouse model contains a human version *TEL-AML1* fusion flanked by two inverted *LoxP* sites, which can flip the *GFP-TEL-AML1* fusion into the 'active' orientation following expression of Cre recombinase. In this construct the GFP TEL-AML1 fusion protein is expressed from the CAGGS promoter and as such is an overexpression model. Because TEL-AML1 is tagged with GFP this model may prove useful in downstream validation experiments. ES cells carrying this construct have also been injected and the chimeras are being bred for germ line transmission. However, as the knockin mouse successfully reproduced the characteristics of human ALL, I did not pursue my study on this backup mouse model after it was made.

Figure 4-10. The targeting construct of Rosa 26 GFP-TEL-AML1 mouse model.

The human version *TEL-AML1* with a N-terminus GFP tag is inserted as a reverse orientation into Rosa 26 targeting vector downstream of CAGGS promoter, together with a hyperactive transposase coding sequence (HSB5) following an IRES site. The TEL-AML1-HSB5 cassette is flanked by two inverted LoxP sites to allow *in vivo* switching on using Cre recombinase. This targeting construct is targeted into Rosa 26 locus by two homologous arms of 4.0 and 4.2 kb, respectively.

4.5 Discussion

4.5.1 Advantages of knockin mouse model for characterizing human *TEL-AML1* oncogenic translocation

As a result of this research, a *Tel-AML1* mouse model was generated to study the *in vivo* consequence(s) of this oncogenic translocation. The *TEL-AML1* is one of the most frequent translocations identified in paediatric cancer and occurring in 25 % of cases of childhood ALL. Although many studies indicated *TEL-AML1* might cooperate with secondary mutations to cause cancer, the detailed molecular mechanism required to induce leukaemogenesis is still remains unclear. In addition, previous transgenic *TEL-AML1* mouse models have all been unsuccessful, creating an opportunity to generate a knockin mouse model that better recapitulates this disease. Therefore, our mouse models bring the opportunity to help understanding the molecular mechanism of a common human cancer.

Compare to mouse models been generated by other strategies, this mouse model was generated through direct knockin of a downstream fusion partner into the original translocation locus have several advantages: 1) The fusion transcript is expressed at the endogenous level. The oncoprotein expression level has already been implicated to be critical for cancer development in KRAS and other mouse models (160), therefore appropriate expression levels will hopefully result the development of leukaemia in the mouse model; 2) The knockin allele also contains a transposase coding sequence. By one cross with the transposon mouse, secondary mutation(s) can be generated to cooperate in the cancer development; linker-based PCR technology could allow quickly identify these *'*secondary hits*'* in mouse tumours; 3) The expression of fusion protein and transposase are temporally and spatially regulated in the same way as in endogenous disease induction, therefore the cancer should develop at the right lineage and right stage.

4.5.2 Expression level of oncogenic fusion protein

Difficulty was experienced in identifying expression of the Tel-AML1 fusion protein in the mouse model and therefore analysis of a large amount of cell lysate by pull down assay was required to identify the fusion protein. This is probably due to low endogenous expression level, and ES cells may not be a good host for expression of the oncogenic fusion protein. Another possibility for low expression level is due to cryptic splicing over the knockin allele, however the *Tel-AML* transcript was able to be detected by real time qPCR, indicating

comparable expression to the endogenous *Tel*. In addition, the homozygous *Tel-AML1* were embryonic lethal, suggesting the cryptic splicing might not count for the low expression of the fusion protein. Therefore a functional correlation may exist between the oncogenic property of the fusion protein and low expression level. This is also indicated by previous publications that oncogenic fusion proteins expressed at a much lower level compared with their wild type fusion partners (153,161,162). It has also been shown that the low expression level(s) might be due to a low rate of synthesis but not a decreased half-life (161). One possible mechanism is that the fusion protein is expressed at low level in the predisposition stage of cancer formation but could be up-regulated by cooperative mutations during malignancy. In this mouse model the fusion protein incorporates a Flag tag, allowing the future testing of this hypothesis in mouse tumours.

4.5.3 The choice of mouse or human AML1

The mouse model presented here expressed a *Tel-AML1* fusion between mouse *Tel* and human *AML1*. It was decided to knockin the human *AML1* cDNA because other mouse models using human fusion protein developed a pre-leukaemic phenotype with an expanded pro-B cell population, consistent with the phenotype of *TEL-AML1* associated cALL in human. Furthermore, the zebrafish model of *TEL-AML1* developed leukaemia, indicating the human *TEL-AML1* could induce leukaemia formation in other model organisms. Both TEL and AML1 protein sequences between mouse and human are highly similar, sharing 88 % and 95 % sequence identity in the TEL and AML1 sequence respectively. Therefore it would be reasonable to assume both the mouse and human version TEL-AML1 fusion protein would have very similar oncogenic properties. However, as part of the fusion protein that has already been characterized, human *AML1* was used sequence in the mouse model.

4.5.4 The *Tel-AML1* knockin mice as a model for human ALL

The *Tel-AML1* knockin mice have derived certain numbers of B cell leukaemia under transposon-mediated mutagenesis, suggesting this knockin strategy could be successfully used for modelling human cALL by *TEL-AML1* translocation, a disease which could not be modelled by previously studies in mice. The previous *TEL-AML* mouse models, however, have based on retrovirus-mediated transplantation or transgenic method for ectopically delivering and expressing *TEL-AML1*. One hypothesis to explain the lack of success of these *TEL-AML* mouse models is that the oncogenic fusion protein needs to be expressed at a

physiological level to induce the right type of disease. In our mouse model, the *Tel-AML1* fusion is expressed from the *Tel* locus at physiological level, thus recapitulating the *TEL-AML1* oncogenic fusion expressed from the human patient. In addition, from the tumour watch studies it seems that SB transposon-mediated mutagenesis could greatly accelerate the disease in the knockin mouse, suggesting the requirement of secondary mutations to derive B cell leukaemia. The next step of this study would be to isolate insertion sites from the appropriate tumour type and identify these cooperative mutations. In turn candidate genes as 'secondary hits' could then be validated using this *Tel-AML1* mice for understanding the pathogenesis of this disease and investigate treatment strategies for use in human patients.

In contrast to human *TEL-AML1* patients where the disease is predominantly derived from pre-B cells, the acute leukaemia derived from the *Tel-AML* mice can be originated from other cellular types such as myeloid and precursor T cells. The reason for this different disease spectrum could be well caused by the differences between mouse and human in hematopoietic system. For instance, the common cell surface markers for undifferentiated hematopoietic stem cells in mouse are $CD34^{low/-}CD38^{+}$ lin^{-} but for human are $CD34^{+}CD38^{low/-}$ lin^{-}. In addition, the cell signalling pathways are also quite different between mouse and human in hematopoietic system.

Chapter 5. Modelling the consequences of *Brd4-NUT* oncogenic translocation in mouse ES cells using a conditional knockin strategy

5.1 Introduction

Chromosome translocations and structural aberrations giving rise to fusion oncogenes are some of the most common mechanisms in oncogenesis. However, although these types of chromosome rearrangements have long been identified in hematological and soft-tissue malignancies, they have only rarely been described in the corresponding carcinomas, which are responsible for around 80 % of the human cancer cases. These recurrent oncogenic fusions include the *RET* and *NTRK1* genes involved in papillary thyroid carcinomas (163), *PAX8-PPARG* fusion involved in follicular thyroid carcinoma (164), *MECT1-MAML2* fusion involved in mucoepidermoid carcinoma (165,166), *ETV6-NTRK3* fusion involved in secretory breast carcinomas (167,168) . Similar to hematological and soft-tissue malignancies, the most common types of genes involved in fusion oncogenes in carcinomas are transcription factors and tyrosine kinases. However, the mechanisms behind most of these chromosome fusions in carcinogenesis have far from been clarified, mostly because these

types of somatic fusions have only started to be identified very recently and have not been well studied in human and mice.

One particularly interesting chromosomal translocation in solid tumours is the t(15;19)(q13;p13) which results in the expression of a fusion transcript between the bromodomain transcription factor *BRD4* and a novel gene *NUT*. Expression of the *BRD4-NUT* fusion transcript is diagnostic of a highly lethal carcinoma (average survival after diagnosis in human patients = 28 weeks) called midline carcinoma arise from the epithelia midline structures (169-171). The human *BRD4* gene encodes a 1400 amino acid BET family protein containing two bromodomains in the N terminus and an ET domain in the C terminus. The bromodomain has been shown to interact with chromatin via acetylated histone H3 and H4 (172). BRD4 plays a role at several stages of the cell cycle progression, including during the G_1/S and G_2/M transition (173,174). In contrast, the cellular function of the *BRD4* fusion partner, *NUT*, is largely unknown, but recent studies have identified *NUT* rearranged with other genes in a subgroup of *NUT* midline carcinomas (NMC) (175).

The mechanism for *BRD4-NUT* fusion in the generation of carcinomas is largely unknown. Cytogenetic analysis suggests that the BRD4-NUT fusion protein always binds to chromatin through a NUT mediated interaction, indicating the oncogenic property of BRD4-NUT could be due to this chromatin binding activity associated with NUT (176). Interestingly, rather than other carcinomas which are normally associated with genomic instability, the karyotype(s) associated with all *BRD4-NUT* translocations are remarkably simple, often having the t(15;19) as the sole aberration (177-180) (**Figure 5-1**), suggesting that few additional mutations are required for this fusion to initiate cancer formation. In addition, *NUT* has also been identified to fuse with *BRD3*, a BRD family protein with very similar structure to *BRD4* (176), suggesting the fusion between *BRD4* and *NUT* is not due to a structural bias on the chromosome. Therefore, a mouse model of *BRD4-NUT* fusion is required for modelling this disease and better understanding the molecular mechanism(s) associated with the *BRD4-NUT* translocation.

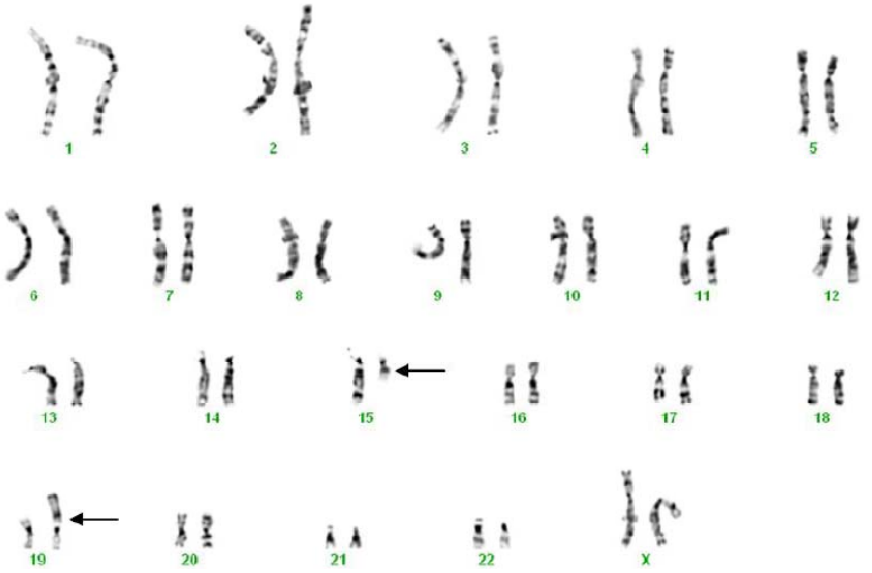


Figure 5-1. Karyotype of t(15;19) BRD4-NUT translocation in a 30-year-old midline carcinoma patient. Cited from Engleson, J, et. al, 2006. (171)

5.2 Aims and summary of the project

This primary aim of this project is generation of a *BRD4-NUT* knockin mouse model for studying the associated phenotype *in vitro* and the tumourigenesis of BRD4-NUT *in vivo*. Specifically the aims for this project are:

1. Evaluation of the phenotype of *BRD4-NUT* expression in a cell culture based assay, specifically with respect to cell cycle and cell growth.
2. Generate conditional knockin mice and tumour watch studies by crossing the *BRD4-NUT* +/- mice with the *CreERt2 +/- mice*.
3. Cross the *BRD4-NUT* +/- mice with the *T2/Onc3* mice - a recently modified *Sleeping Beauty* transposon mouse line developed by *Dupuy et al.* (181) for screening secondary mutations.
4. Establishment of a collaboration with pathologists at Addenbrooke's hospital to characterize human *BRD4-NUT* carcinomas and therefore further understand clinical aspects of midline carcinoma.

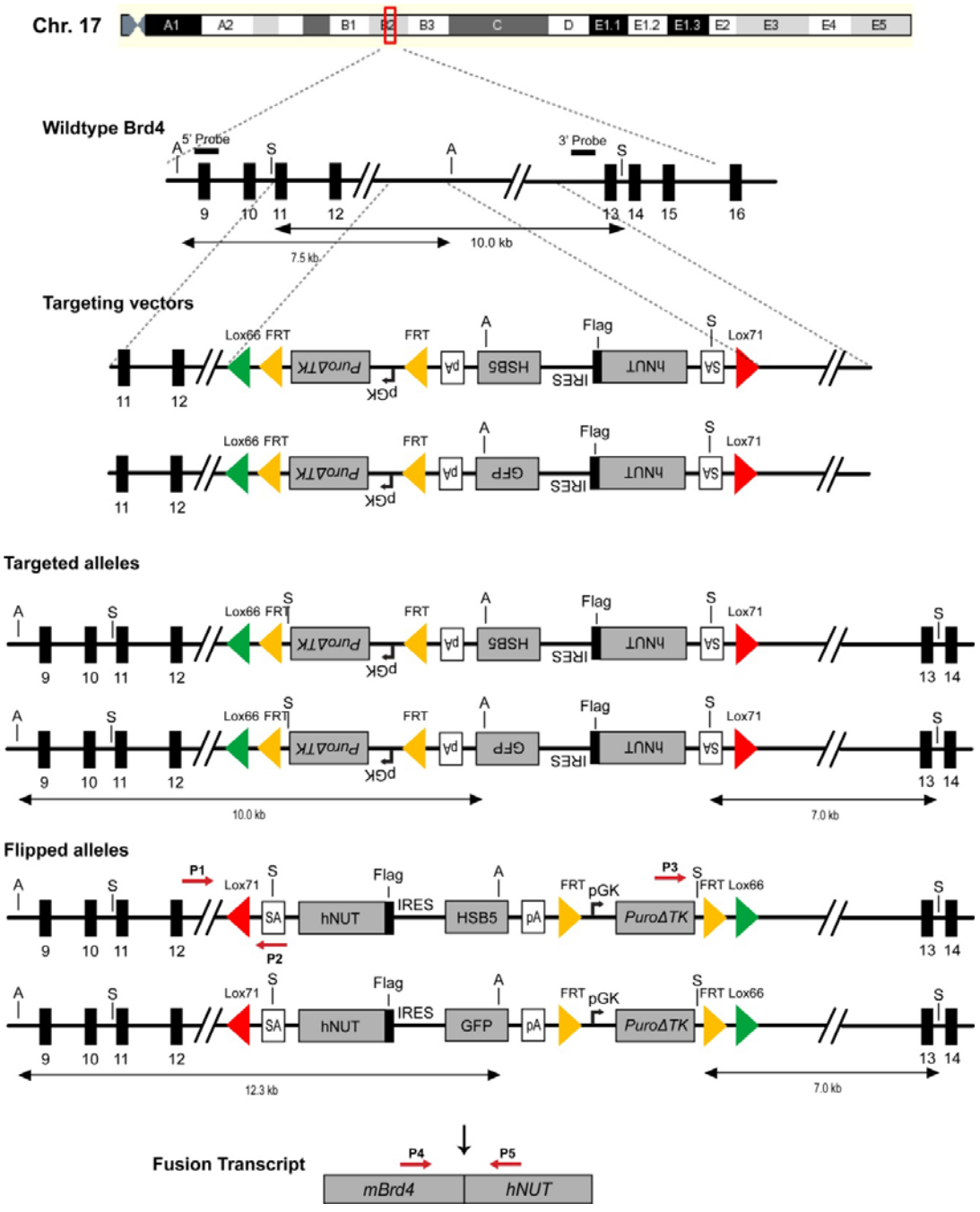


Figure 5-2. Schematic diagrams of Brd4-NUT targeting constructs and targeted alleles.

Both the SB and GFP versions targeting constructs and targeted alleles are shown. The double head arrows under each allele represent southern blot DNA fragments length after enzymatic digestion. S – *Sph*I, A – *Afl*II restriction sites.

Similar to the *Tel-AML1* mouse model described in Chapter 4, to generate the *BRD4-NUT* fusion the human *NUT* cDNA was knocked into mouse *Brd4* locus followed by an IRES and a transposase coding sequence (**Figure 5-2**). As the expression of *BRD4-NUT* has been shown to affect cell cycle progression (161), the knockin cassette is flanked by two *loxP* sites to allow activation of the *BRD4-NUT* in an conditional manner. After the knockin allele is transmitted through the germ line, this line will be crossed with a CreERt2 mouse model (182) to induce BRD4-NUT and transposase expression following treatment with 4-hydroxytamoxifen (4-OHT) (**Figure 5-3**).

After targeting and screening, the targeted ES cell clones for *Brd4-NUT* were sent for blastocyst injection to derive germ line transmissions for the *BRD4-NUT* mouse line. However, although 14 clones were injected and over 30 chimera mice were generated with chimeric rate between 20-90 percent, knockin alleles could not be transmitted through the germ line (more information will be presented in the result section for this part of work). Therefore *in vivo* characterization of this mouse model could not be performed.

In *in vitro* studies, following targeting of this allele into a lab made ES cell line carrying the CreERt2 allele, I have shown that after 4-OHT application, fusion transcript is expressed at a comparable level with the endogenous *Brd4* expressed from the untargeted locus. However, only weak expression of the fusion protein could be identified in the targeted cells using immunoprecipitation, indicating the fusion protein is either not stable in ES cells or expressed at too low level to be identified. Cell culture characterization studies showed the cell growth is severely impaired and the colony formation is completely blocked upon expression of *Brd4-NUT*. Cell cycle analysis using Propidium iodide (PI) staining subsequently indicated the arrest is taking place at the G_2/M phase.

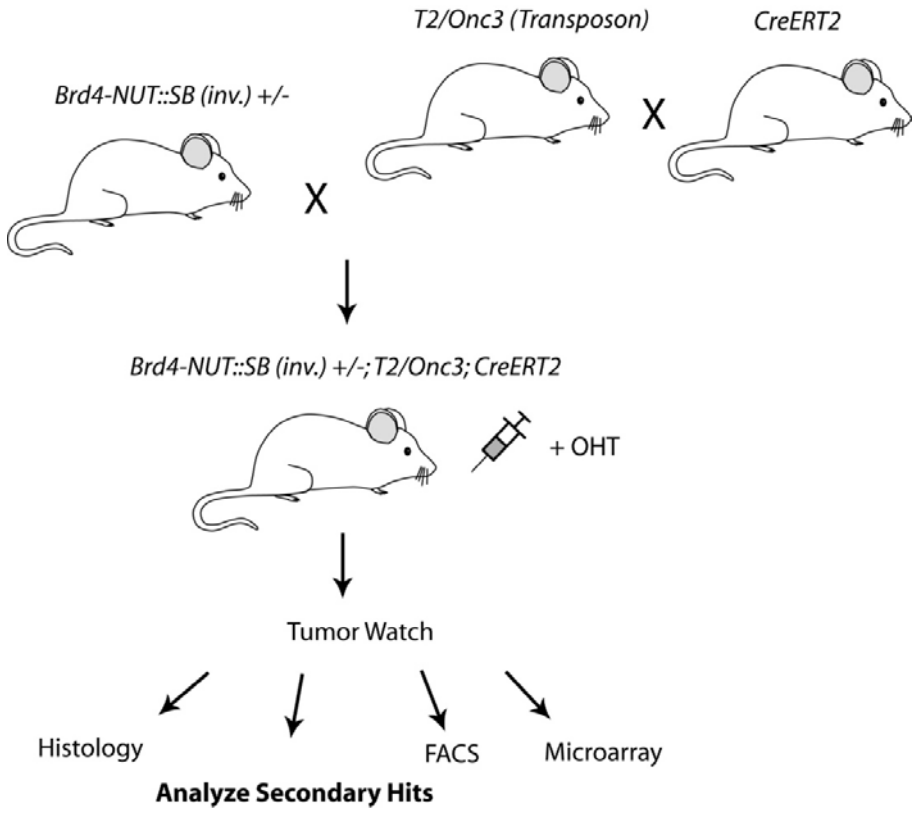


Figure 5-3. Crossing strategy for tumour watch and subsequent characterization studies in *Brd4-NUT* mouse model.

After the F1 mice are derived, the *Brd4-NUT-SB* mouse will be crossed with the *T2/Onc3;;CreERt2* mouse. Tumour watch studies in experimental mice treated with 4-OHT administration at different stage to turn-on BRD4-NUT expression. After tumours are derived, tumours will be identified and characterized. Secondary hits by transposon mutagenesis will also be identified using splinkerette PCR and 454 sequencing.

5.3 Materials and Methods

5.3.1 Targeting vector construction

For generating the *Brd4-NUT* targeting construct (**Figure 5-2**), the human *NUT* cDNA was cloned from the GFP-BRD4-NUT expression vector (161) and was used to exchanged with the *AML1* sequence in the vector described above (see Chapter 4). Two genomic fragments from the *Brd4* locus were amplified by High Fidelity PCR and inserted into pBlueScript SK+ construct as targeting arms (5' arm: FWD 5'-AAAGCGGCCGCGGCCAAAAAGGCCTTGGCT TCAGTCACCAGTCTGGGTTGGTGCCCTATCATACGCA-3', REV 5'-AAAAAAA GGCCGGCCACGCGTAAGCTTGACTGGCAATAAAAGTG AAAAGTCAGTG-3'; 3' arm: FWD 5'- AAAAAATTAATTAAACGCGTAAGCTTAAGAATCCAAGTGTT CAAATGACAATCCCAGAGACTGACCCT-3', REV 5'- AAAAAAAAAAGGCGCG CCTGGCTCCCCAACAGGATCCCAGCTGGTATACTAAGGCTT-3'). The NUT-SB-Puro cassette flanked by the Lox66 and Lox71 sites was inserted into engineered *Mlu*I restriction sites in a reverse orientation relative to *Brd4* transcription. To generate a Flag tag-GFP version construct of the *Brd4-NUT* construct a similar recombineering strategy was used to that described in Chapter 4. The constructs generated were sequenced in full to ensure that PCR had not introduced any mutations.

5.3.2 Immunoprecipitation of FLAG Tagged Proteins

The Dynabeads protein G (Invitrogen, 0.5 ml beads) were first washed three times (buffer: 24.5mM Citric Acid, 51.7 mM Dibasic Sodium phosphate (Na_2HPO4) dehydrate, pH = 5.5). One microgram of anti-Flag M2 antibody (Sigma, F1804) was incubated with the beads in 20 µl of bead wash buffer for 40 min at room temperature. After incubation beads were washed three times with beads wash buffer with 0.1 % Tween-20 (Sigma). To prepare a cell lysate, cells were treated with protein lysis buffer (50 mM Tris pH 8.0, 450 mM NaCl, 0.2 % Nonidet P-40 (Igepal), 1 mM DTT, 1 mM EDTA, 1X Protease inhibitor (Roche) for 15 min on ice, then collected by centrfiguation at maximum speed using a desktop centrifuge for 15 min at 4 ˚C. Cell lysate was collected and incubated with the antibody conjugated beads for 1 hour at 4 ˚C with gentle shaking. The beads were collected after incubation and washed three times with protein wash buffer (same formula as protein lysis buffer except using 150 mM NaCl and 0.1 % NP-40 concentration). For Western blotting, 30 µl loading buffer were added

to the beads after pull down. The beads were then boiled for 10 min at 95 ˚C and supernatants were loaded directly on SDS page gels (5 %, Bio-rad). Western blotting was using anti-Flag M2 antibody (Sigma, F1804) and performed following manufacturer's instructions from Sigma.

5.3.3 Cell proliferation and colony formation assay

Cells were seeded into T25 flasks at a density of $4x10^6$. After two days these cultures had reached 80 % confluence. Cells were then treated with trypsin and counted using a Beckman Coulter cell counter. For cell proliferation experiments, $2x10^4$ cells were plated on each well of a 6 well plate in ES cell culture medium supplemented with 1 μM 4-OHT (Sigma) or the same volume of vehicle solution (95 % EtOH). Triplicate samples were prepared for each cell line. Media was changed every two days and cell number was calculated at day 6. For colony formation assays, 1000 cells were seeded into 10 cm culture dishes at single cell density. Triplicate plates were prepared for each cell line. Cultures were supplemented with 1 μM 4-OHT or 95 % EtOH for 10 days. Media was changed every three days and colonies were stained with methane blue at day 10 and colony numbers counted.

5.3.4 Cell Cycle Analysis by Flow Cytometry

Wild type E14 ES cells and *Brd4-NUT-GFP* targeted ES cells (CE-37) were treated two days with 4-OHT or 95 % EtOH (vehicle) and then subjected to flow cytometry analysis for GFP which is bicistronically expressed from the *Brd4-NUT* fusion transcript. GFP positive and negative cells were sorted by FACS into cold 70 % EtOH and fixed overnight at 4˚C. Unsorted wild type and CE-37 cells were fixed overnight with 70 % EtOH. Cells were washed two times with PBS and stained with 500 μl PI solution (0.1 % Triton X-100 (v/v), 50 μg/ml Rnase A (Sigma), 25 μg/ml Propidium Iodide (Aldrich,)) overnight at 4˚C then analysed by flow cytometry.

5.4 Results

5.4.1 Generating "conditional" *Brd4-NUT* knockin mouse model and characterizing targeted ES cells

The construct design for the *Brd4-NUT* knockin mouse model used a similar approach to the *Tel-AML1* mouse, except the human *NUT* and SB transposase cDNA sequence were

introduced in an inverted orientation into mouse *Brd4* locus flanked by two mutant *LoxP* sites, 66 and 71 (**Figure** 5-2), so that Cre-mediated recombination could be applied to flip the cassette 'on' and thus express *Brd4-NUT* conditionally on application of 4-OHT. To validate this conditional allele, this construct has also been targeted into an ES cell line expressing an inducible CreERt2 so that 4-hydroxytamoxifen (4-OHT) could be used to flip the cassette and study the cellular phenotype of the fusion protein *in vitro*. Verification that the ES cells been successfully targeted and that the flipping of the cassette occurred after induction of Cre via 4-OHT administration was determined through Southern blotting (**Figure 5-4, A)**. Using primer pairs flanking the flipped junction sites (Primers P1 + P2), a 0.5 Kb PCR fragment was detected in targeted ES cells after 4-OHT treatment but not in the absence of 4-OHT (**Figure 5-4, B**), indicating the CreERt2 cell line was tightly controlled and spontaneously flipping events were low. Using primers flanking the transcript junction site (**Figure 5-2**, primer P4 + P5), the fusion transcript could also be detected by RT-PCR and the fusion between *Brd4* and *NUT* transcripts was verified by subsequent sequencing (**Figure 5-4, C and D**).

By immunoprecipitation using an anti-BRD4 antibody described previously (173), a faint band at around 200 kDa could be identified representing the BRD4-NUT fusion protein, which was not identified in either the wild type or untreated *Brd4-NUT* ES cells (**Figure 5-4 E**). Western blotting using the anti-Flag M2 antibody could not detect any fusion protein expression, despite the addition of a Flag tag to the BRD4-NUT fusion protein (**Figure 5-4 E**), the reasons for which are unclear.

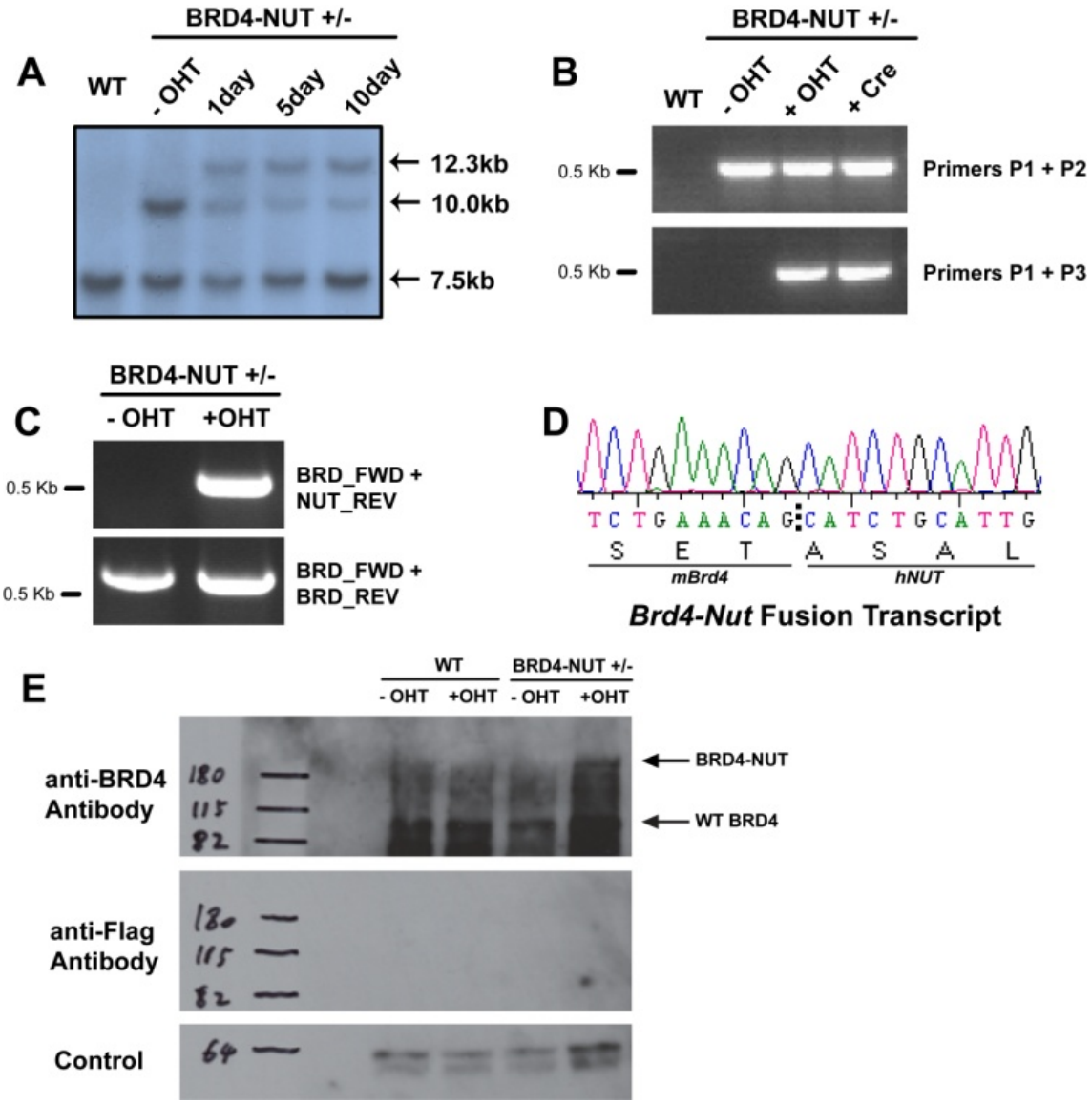


Figure 5-4. Characterization of the *Brd4-NUT* targeted ES cells.

(A) Southern blotting confirmed correct targeting and flipping of the targeted ES cell clone showing a targeted band at 10.0 kb in the targeted cells and 12.3 kb flipped band after treatment with 4-OHT for 1 day, 5 days and 10 days. (B) Genomic PCR using primer pairs flanking the flipping junction site (see Figure 5-2) in genomic DNA for the targeted cells before and after flipping using 4-OHT treatment or Cre-expressing plasmids. (C) RT-PCR showing a PCR band could be amplified using primer pairs flanking the fusion transcript junction site (see Figure 5-2) in the targeted cells treated with 4-OHT. (D) Sequencing trace confirmed the in-frame fusion between Brd4 and NUT transcript. (E) Immunoprecipitation and western blot for the BRD4-NUT fusion protein using an anti-BRD4 or anti-Flag antibody.

5.4.2 Analysis of Brd4-NUT expression using Quantitative PCR

As the BRD4-NUT fusion protein was only detectable at a very low level, a real-time qPCR experiment was setup to validate the expression level of *Brd4-NUT* at the transcriptional level. The *Brd4* gene is expressed in cells as a short and long isoform (161). To compare the relative expression of the two Brd4 isoforms and the fusion transcript, qPCR primers were designed to flank the junction points on three transcripts as indicated in **Figure 5-5**. The Brd4 short isoform was expressed at a much lower level than the long isoform in knockin cells both with or without 4-OHT treatment. The expression level of both isoforms was decreased when cells were treated with 4-OHT, which would be expected as a consequence of the inversion of the knockin allele in response to 4-OHT treatment, thus disrupting one of the Brd4 alleles. In the meantime, the Brd4-NUT transcript was expressed upon 4-OHT treatment, at a lower but comparable level to the Brd4 long isoforms. These results indicate that the conditional *Brd4-NUT* knockin construct is functional.

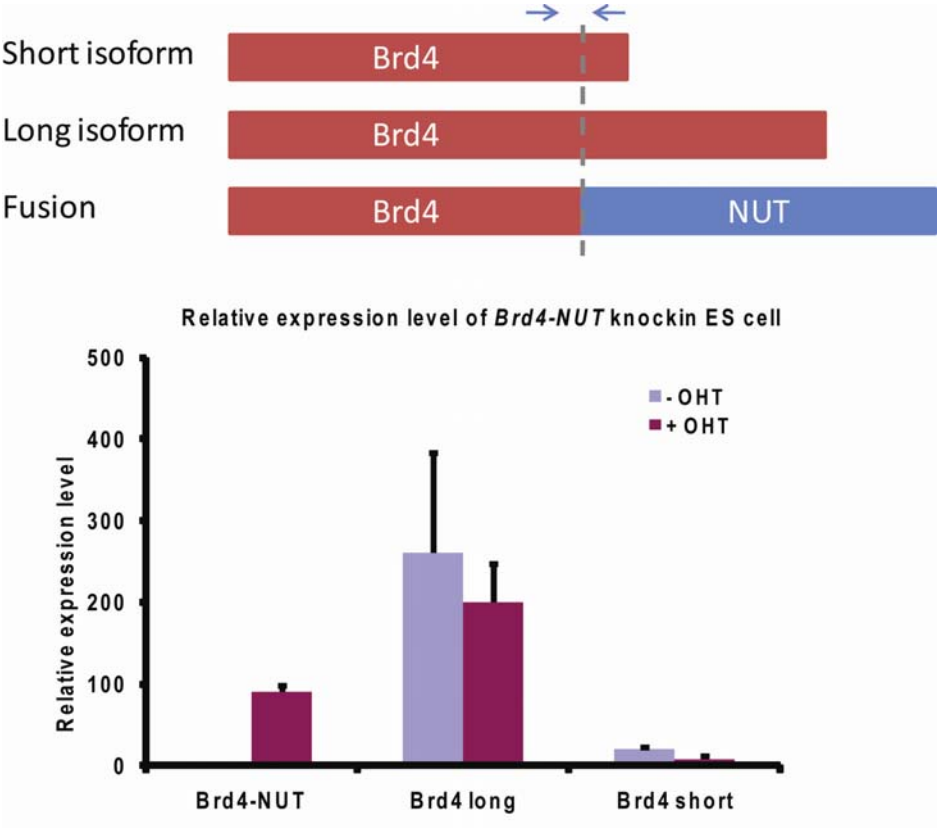


Figure 5-5. Evaluation of the *Brd4-NUT* expression level by qPCR

ES cells were treated with 4-OHT and 95 % EtOH (control) for two days before harvesting and cDNA preparation . For each type of transcript (*Brd4* short, *Brd4* long and *Brd4-NUT*), the expression level of control ES cells was shown in blue and the 4-OHT treated ES cells in purple. The blue arrows represent the primer pairs used for qPCR analysis on each transcript. The experiments were performed in triplicate. Error bars indicate mean expression level ± SD.

5.4.3 Deriving germ line transmission with the Brd4-NUT knockin ES cell lines

14 ES cell clones were injected to derive *Brd4-NUT* knockin lines, however none were successful in obtaining germ line transmission. These ES cell clones were generated from three different genetic backgrounds in over 5 different targeting batches: cell strains E14J – the ES cell strain from the Netherland Cancer Institute (NKI), B6 Blue –the commonly used ES cell strain and JM8.N4 from the mouse knockout project (COMP). Some of these ES cell clones have generated chimeras with quite high chimeric rate (80 - 90 %, as examined by animal facility staff), but none of these chimeras could produce F1 mice with *Brd4-NUT* transmitted through the germ line. Below is a list of all the targeted clones injected for attempted generation of F1 mice.

Table 5-1. Brd4-NUT clone microinjection information

Date	Team 113 ID	Cell Strain	Passage	RSF ID
13/07/2007	AX0015	E14J	15	BRD4-NUT_1
26/07/2007	AX0016	E14J	17	BRD4-NUT_2
14/09/2007	BI0067	E14J	11	BRD4-NUT_3
26/10/2007	BN0005	E14J	9	Cta_BN5
29/10/2007	BK0056	E14J	11	Nut-Flag-Ires-SB-BK56
30/10/2007	BK0004	E14J	11	Nut-Flag-Ires-SB-BK4
21/12/2007	BK0018	E14J	11	Nut-Flag-Ires-SB-BK18
03/03/2008	CB0015	B6 Blu	n/a	Nut-flag-IRES-T
31/03/2008	CB0061	B6 Blu	n/a	Nut-flag-IRES-T
10/06/2008	CB0061	B6 Blu	17	Brd4-NUT-Ires-Transposase
08/01/2009	CL0166	JM8.N4 (Cre-ERt2)	n/a	Brd4_Nut_CL166
13/01/2009	CL0180	JM8.N4 (Cre-ERt2)	n/a	Brd4_Nut_CL180
16/01/2009	CL0180	JM8.N4 (Cre-ERt2)	n/a	Brd4_Nut_CL180
27/01/2009	CL0121	JM8.N4 (Cre-ERt2)	n/a	Brd4_Nut_CL121

5.4.4 Expression of *Brd4-NUT* fusion impaired cell growth in ES cells

Although the efforts to derive a *Brd4-NUT* knockin mouse was unsuccessful, it was observed that during *in vitro* culturing of these *Brd4-NUT* knockin ES cells there was a strong cell growth arrest phenotype associated with 4-OHT induced *Brd4-NUT*. *In vitro* experiments were therefore carried out to evaluate the effects of *Brd4-NUT* expression in ES cells. It has been reported previously that ectopic expression of BRD4-NUT fusion protein could interfere with cell growth and inhibit S phase *in vitro* (161). To test the effects of endogenous *Brd4-NUT* expression in the knockin ES cells, the targeting construct was modified to exchange the SB sequence with a GFP or Neomycin coding sequence (to avoid the possible side-effects of SB transposase expression in cells), and these constructs were subsequently targeted into a CreERt2 ES cell line which has *CreERt2* cDNA targeted into *Rosa26* locus under a CMV promoter. In these cells the *Cre*ERt2 could therefore be activated by 4-OHT treatment to induce *Brd4-NUT* expression.

Four *Brd4-NUT* knockin cell lines derived from targeted ES cells were used for a cell plating assay (see Materials and Methods) to evaluate the effects of *Brd4-NUT* expression on cell growth (**Figure 5-6 A**). At 6 days after plating the original CreERt2 line and the control cell lines (Random 1-3,) which contained a random integrated targeting construct, the cell numbers in the 4-OHT treated samples have dropped to about 60 % of the control cell numbers (grown in vesicle medium with 95 % EtOH). This drop in cell number in 4-OHT treated plates could either be due to a Cre toxicity (183) or 4-OHT toxicity effect, or both. In contrast, all four Brd4-NUT lines (BN-GFP and BN-Neo1-3) treated with 4-OHT to induce expression of the fusion protein had a clear cell proliferation arrest phenotype; producing only 20 % of the number of cells compared with control cells treated in a vesicle medium. (**Figure 5-6 B**). This indicates that the *Brd4-NUT* expression severely impaired cell growth ability of ES cells. In addition, all four targeted ES cells grew slower than control cell lines in the absence of 4-OHT treatment (data not shown); this could be due to disruption of the knockin allele in one *Brd4* allele causing a negative effect on cell growth independent of *Brd4-NUT* expression.

During activation of the *Brd4-NUT* expression by 4-OHT, the flip of the knockin cassette could also disrupt *Brd4* gene transcription from the targeted allele. This has previously been demonstrated to cause haploinsufficiency for the wild type *Brd4*, therefore could be a possible mechanism causing the cell proliferation phenotype observed (184). To confirm this

cell growth arrest phenotype is associated with *Brd4-NUT* expression, the *Brd4* gene was knocked out using a targeting construct obtained from the EUCOMM project which targets exon 5 at the open reading frame of *Brd4*. In this targeting construct the *Brd4* exon 5 is flanked by two *loxP* sites which could be removed by Cre expression. After targeting of this construct into the CreERt2 cell line and validation by Southern blotting (data not show), two targeted ES cells were selected as a control in the cell proliferation assay. The two Brd4 +/- knockout cell lines had very similar cell growth pattern to the Cre-ERT2 and random insertion control cells after 4-OHT treatment (**Figure 5-6 B**), indicating that the disruption of one *Brd4* allele is sufficient reason to cause cell proliferation arrest. As further proof that the phenotype in cell culture is caused by *Brd4-NUT* expression, the knockin targeting construct was modified to truncate ~2700 base pairs from the C-terminal *NUT* cDNA using a *BstE*II restriction site. This construct expressed a truncated BRD4-NUT protein (Brd-ΔNUT), with a 200 amino acid NUT sequence forming a C-terminal fusion to BRD4, in contrast with the over 1000 amino acid NUT sequence of the original construct. Surprisingly, when the knockin ES cells with the truncated NUT cDNA were treated with 4-OHT, the cell growth rate was only slightly slower than the wild type controls but much higher than the growth rate of all four *Brd4-NUT* knockin cell lines (**Figure 5-6 B**). This indicates that full length *Brd4-NUT* is essential for causing the cell proliferation arrest phenotype.

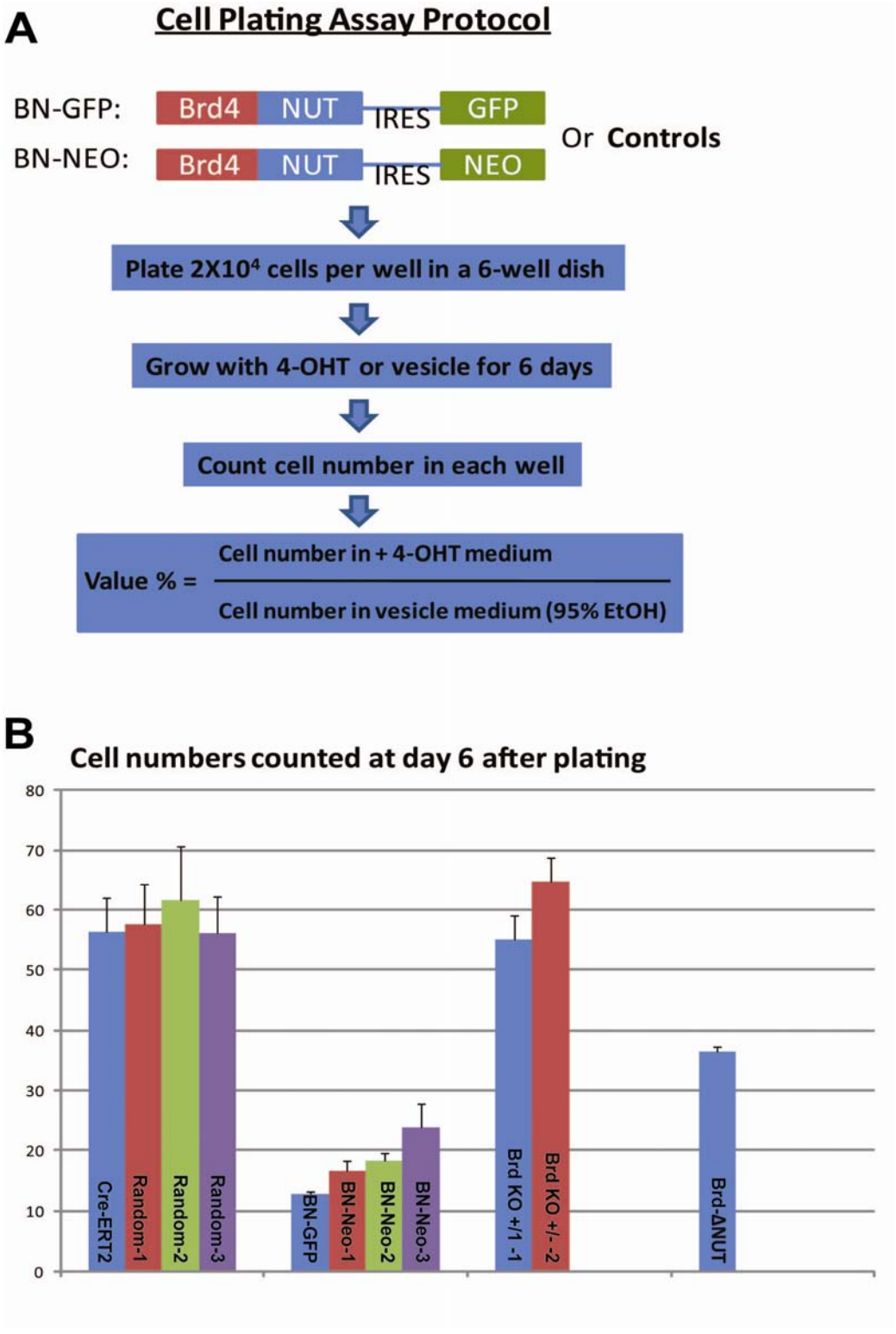

A
Cell Plating Assay Protocol
BN-GFP:
Brd4
NUT
IRES
GFP
BN-NEO:
Brd4
NUT
IRES
NEO
Or Controls
Plate 2X10⁴ cells per well in a 6-well dish
Grow with 4-OHT or vesicle for 6 days
Count cell number in each well
Value % = Cell number in + 4-OHT medium / Cell number in vesicle medium (95% EtOH)
B
Cell numbers counted at day 6 after plating
80
70
60
50
40
30
20
10
0
Cre-ERT2
Random-1
Random-2
Random-3
BN-GFP
BN-Neo-1
BN-Neo-2
BN-Neo-3
Brd KO +/1 -1
Brd KO +/- -2
Brd-ΔNUT

Figure 5-6. Cell plating assay in *Brd4-NUT* ES cells.

(A) Experiment strategy of the cell plating assay. (B) Cell number counting results at day 6 after plating for *Brd4-NUT* knockin ES cells and control groups. The Y-axial indicates the *'*value %*'* for each cell line (*n* = 3) calculated using the cell numbers in medium treated with 4-OHT divided by the cell numbers in medium treated with vesicle solutions (95 % EtOH) at day 6. Means ± SD are shown for three parallel experiments. Cre-ERT2: original ES cell line used for the targeting experiments; Random 1-3: cell line with random integration of the targeting constructs; BN-GFP: *Brd4-NUT* targeted ES cell expressing a GFP protein after IRES site; BN-Neo 1-3: *Brd4-NUT* targeted ES cell expressing a Neomycin drug resistance protein after IRES site; Brd KO: Brd4 conditional knockout cell lines; Brd-ΔNUT: *Brd4-NUT* targeted ES cell expressing a truncated NUT protein fused to the C-terminal BRD4.

5.4.5 Expression of *Brd4-NUT* fusion blocks colony formation ability in ES cells

To further validate the growth arrest phenotype associated with *Brd4-NUT* expression, a colony formation assay (see Material and Methods) was carried out to evaluate the colony formation ability of ES cells expressing *Brd4-NUT*. For this cell lines with or without 4-OHT treatment were plated at single cell density to result in individual colony formation, and colony numbers were counted after 10 days (**Figure 5-7 A**). Treatment with 4-OHT completely blocked colony formation ability in targeted ES cells (BN-GFP and BN-Neo-1, -2, -3) but only caused a moderate drop (~40 %) in control cell lines (Random-1, -2, -3 and the original Cre-ERt2 cell line) (**Figure 5-7 B and C**). Similar to the results from the cell plating assay, the number of colonies formed for the two *Brd4* KO +/- cell lines were similar to the control groups (**Figure 5-7 B and C**), indicating disruption of the *Brd4* allele during *Brd4-NUT* activation is not a cause of cell growth arrest in ES cells. Interestingly, albeit a moderate delay in cell growth was observed in the NUT truncated knockin ES cells (Brd4-ΔNUT), this cell line was able to form some colonies when treated with 4-OHT to induce truncated Brd4-NUT expression, compared with the full length Brd4-NUT lines where cell proliferation was almost completely arrested after treatment with 4-OHT. This demonstrates that the NUT sequence in the BRD4-NUT is essential to cause the growth arrest phenotype in ES cells (**Figure 5-7 B and C**).

In another experiment to derive a flipped *Brd4-NUT* clone (**Figure 5-7 D**), the targeted ES cells (BN-GFP) were treated for two days with 4-OHT so that the cell population contains a mix of flipped and unflipped cells. These cells were seeded to form single colonies and 100 colonies were picked at day 10 for genotyping and southern blot. In alignment with the colony formation results, none of these colonies contains the flipped allele indicating that expression of *Brd4-NUT* transcript has completely blocked colony formation ability.

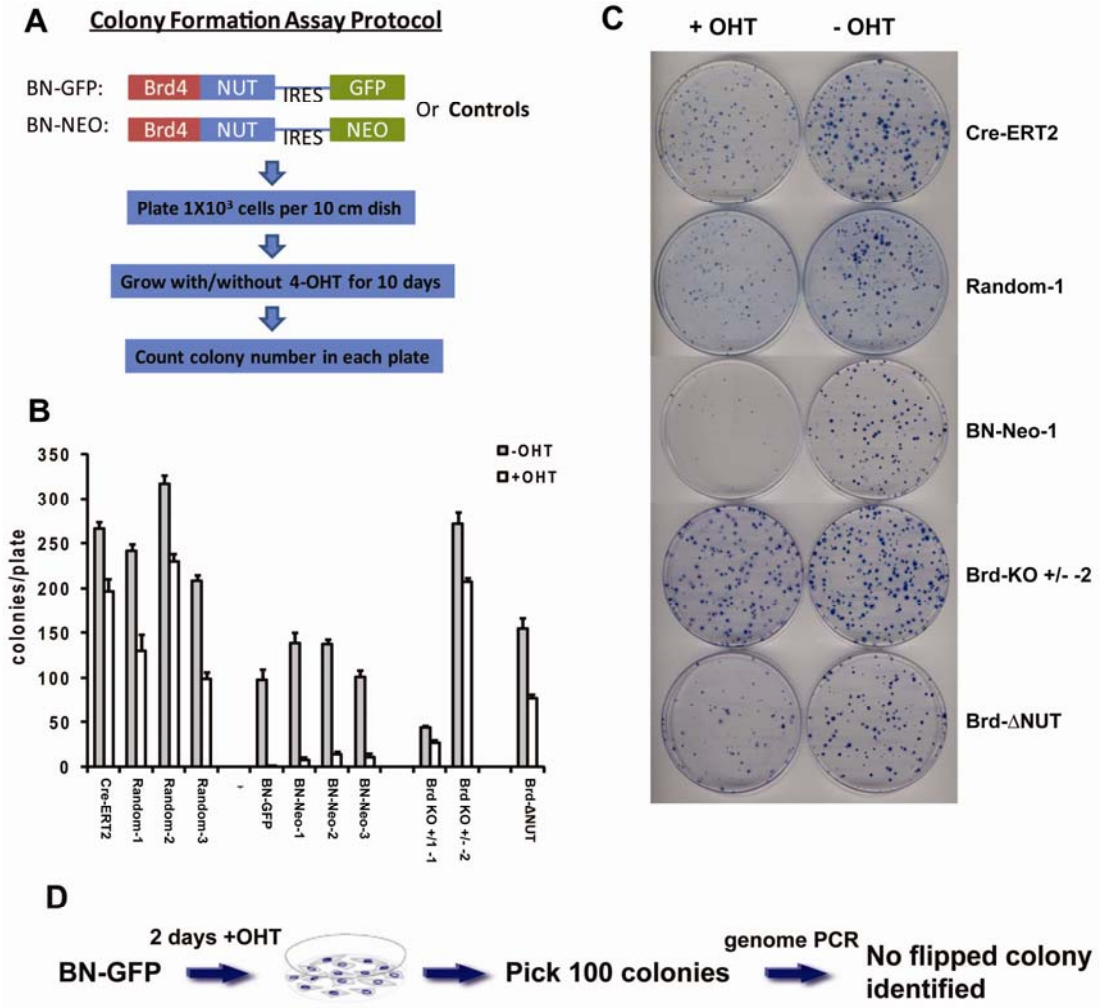


Figure 5-7. Colony formation assay with *Brd4-NUT* ES cells

(A) Experiment strategy for the colony formation assay. (B) Quantification of colony numbers of experiment and control group cell lines without or with 4-OHT treatment. Means ± SD are shown for three parallel experiments. (C) Representing plates from colony formation assay. (D) Procedure for colony plating assay to derive flipped *Brd4-NUT* clone. Cre-ERT2: original ES cell line used for the targeting experiments; Random 1-3: cell line with random integration of the targeting constructs; BN-GFP: *Brd4-NUT* targeted ES cell expressing a GFP protein after IRES site; BN-Neo 1-3: *Brd4-NUT* targeted ES cell expressing a Neomycin drug resistance protein after IRES site; Brd KO: Brd4 conditional knockout cell lines; Brd-ΔNUT: *Brd4-NUT* targeted ES cell expressing a truncated NUT protein fused to the C-terminal BRD4.

5.4.6 Expression of *Brd4-NUT* arrested cell cycle at G2/M phase

Since the targeted BN-GFP cell line also expressed a GFP protein from the targeted allele after treatment of 4-OHT, GFP signal could be used to indicate the *'*switched on*'* or *'*switched off*'* state of the *Brd4-NUT* expression. To further characterize the *Brd4-NUT* induced cell growth arrest, flow cytometry was used to separate green and non-green populations after 4-OHT treatment. Propidium Iodide (PI) staining was performed alongside to analyze progression of the cell cycle (**Figure 5-8 A**). Although the GFP signal is weak due to low expression from the endogenous *Brd4* locus, a broadened GFP spectrum could be seen after 4-OHT treatment in the BN-GFP cell line compare to the CreERt2 control ES cells treated with 4-OHT (**Figure 5-8 B**). After sorting, the green and non-green populations in 4-OHT treated BN-GFP cells were PI stained and analysed to determine cell cycle progression. An obvious change in cell cycle stage was observed in the green cell population with a much stronger peak at G_2/M phase (**Figure 5-8 D**), indicating a strong cell cycle arrest at this stage. However the cell cycle appeared normal in the non-green cells sorted from the same cell population. This G_2/M arrest was only present in the targeted BN-GFP ES cells after 4-OHT treatment but not in the control CreERt2 cell. The G2/M cell population was increased in the 4-OHT treated BN-GFP cells after FACS sorting, which was enriched in the GFP::Brd4-NUT cell population (**Figure 5-8 C**).

This experiment suggested that *Brd4-NUT* expression in ES cells results in a significant cell cycle arrest at G_2/M phase, which could be the cause of cell growth arrest phenotypes observed in the previous experiments. Due to the weak expression of the GFP marker, the construct has been re-engineered by exchanging GFP with the more sensitive *Venus* YFP cell marker (185). However, due to YFP filters not being available, separation experiments based on Venus YFP could not be performed at the present time.

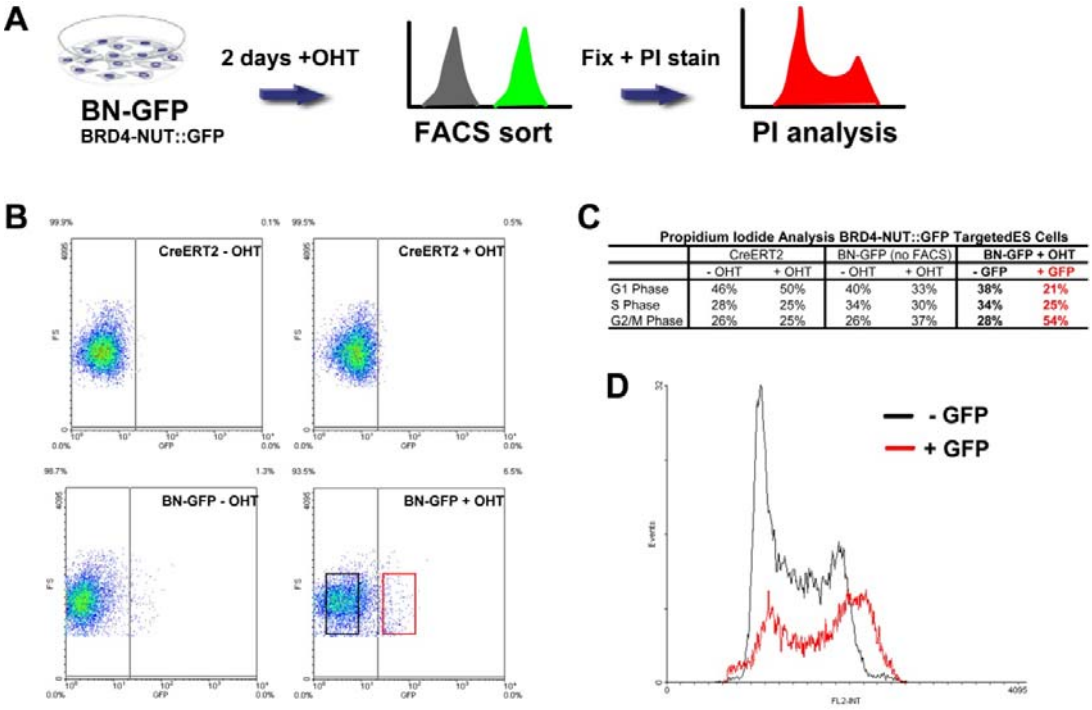


Figure 5-8. Cell cycle analysis of *Brd4-NUT* ES cells

(A) Experimental strategy for cell cycle analysis of *Brd4-NUT* ES cells. (B) FACS sorting for GFP signal of *Brd4-NUT* ES cells without or with 4-OHT treatment. (C) Quantification of cell cycle population in cell samples without or with 4-OHT treatment (Column 1: CreERt2 ES cells, Column 2: BN-GFP ES cells, Column 3: 4-OHT treated BN-GFP ES cells after FACS sorting). (D) PI analysis for FACS sorted green (red) and non-green (black) populations in targeted *Brd4-NUT* ES cells treated with 4-OHT.

5.5 Discussion

5.5.1 Models of choices for characterizing human carcinogenic translocations

This study aimed to generate a *Brd4-NUT* mouse model to study the oncogenic translocation of BRD4-NUT *in vivo*. In contrast to *TEL-AML1* which has been well studied, *BRD4-NUT* is a relatively rare translocation. Only a few clinical cases have been documented although it is likely to be under-reported because karyotype analysis is rarely performed on solid tumours. However, as a recurrent translocation identified in midline carcinoma, *BRD4-NUT* might represent a novel mechanism in the pathogenesis of solid tumours that has been underestimated. The mouse model described in this chapter brings the potential to identify a novel pathway in *BRD4-NUT* induced midline carcinoma, a rare human disease.

The *BRD4* gene is present as a long and a short isoform in both the mouse and human genome. In human patients with *BRD4-NUT* translocation, the break point has been identified in intron 12 of the *BRD4* gene, which disrupted both the short and long isoforms (161). Although the mouse *Brd4* short isoform only has 12 exons in Ensembl database (Transcript ID: ENST00000360016), in-house sequence comparison revealed a 13 exon after the NUT insertion site. The NUT insertion site in our knockin cell line is in mouse *Brd4* intron 12, at a similar position with the genomic break point found in human patients. Therefore, activation the NUT cassette would also disrupt both short and long *Brd4* isoforms in our mouse model, which exactly mimics the events taking place in human *BRD4-NUT* translocation.

5.5.2 Germ line transmission of the *Brd4-NUT* knockin ES cells

In this study I attempted to derive germ line transmission for the *Brd4-NUT* knockin mouse. Although 14 cell clones under three genetic backgrounds were used for microinjection the knockin allele could still not transmit through the germ line. Given that the *NUT* cDNA is knocked-in at a reverse orientation, is it highly unlikely that the problem is caused by *Brd4-NUT* expression in the targeted ES cells without induction. Nevertheless, knocking-in of *NUT* cDNA into the wild type mouse *Brd4* locus could still disrupt the original locus to some extent and cause haploinsufficiency of *Brd4*. However, it has been also shown that the heterozygous *Brd4* knockout could also be generated and be transmitted through the germ line (184). It is worth to notice that in the cell plating assay, the growth rate of all four *Brd4-NUT* knockin ES cell lines were approximately 50 % slower than the growth rate of control cell lines in the medium without 4-OHT. Therefore the difficulties in obtaining *Brd4-NUT*

knockin ES cells for germ line transmission might reasons associated with disruption of the mouse *Brd4* locus which results in a reduction in cell growth rate.

5.5.3 *Brd4-NUT* induced cell growth arrest

The *Brd4-NUT* mouse model had a strong cell growth arrest phenotype in ES cells, which was observed in both cell growth assays and cell cycle analysis. In the *Brd4-NUT* knockin allele, the *NUT* cDNA was inserted into intron 12 on *Brd4* locus (**Figure** 5-2). Activation of the NUT cassette results in expression of the fusion protein but also causes *Brd4* heterozygosity. We believed that the cell growth arrest phenotype is caused by expression of the fusion protein rather than *Brd4* haploinsufficiency, because although the *Brd4* haploinsufficiency has been indicated to slow down cell growth rate (184), it did not cause a strong cell growth arrest in our experiments. In addition, the *Brd4* null allele could be derived by gene trapping (184), which further indicates that the *Brd4* haploinsufficiency could not block colony formation ability. To better discriminate the effects between fusion protein expression and *Brd4* heterozygosity, we generated *Brd4* +/- cell lines in CreERt2 background and demonstrated that the cell growth rate of *Brd4* +/- cell lines was similar to the rate of control cell lines in cell plating and colony formation assays.

The cell cycle arrest was identified to be at the G2/M phase, which is different from the results obtained by Haruki et al. who claimed that ectopic expression of BRD4-NUT in H293T cells caused cell cycle arrest at S phase (161). However, another publication showed that inhibition of wild type BRD4 protein using an anti-BRD4 antibody could inhibit HeLa cells entering mitosis (173). Considering the *BRD4* itself is a multi-stage cell-cycle controlling gene and is affected by haploinsufficiency (184), it is possible that the *Brd4-NUT* expression in different cells and at different levels could result in different phenotypes. Therefore the actual function of *Brd4-NUT* in cancer inducing cells requires the *Brd4-NUT* expression system to be tested in further *in vitro* and *in vivo* studies.

Chapter 6. Summary and future directions

6.1 Generation of a high-efficient insertional mutagenesis pipeline

In the research presented in this thesis, I have explored different aspects of insertional mutagenesis and attempted to generate a pipeline to facilitate insertional mutagenesis research. I have developed *'iMapper'*, a freely accessible web tool for large-scale automated analysis and mapping of insertion sites. This software could uptake thousands of splinkerette PCR sequencing reads all in once and generate processed output in various formats, therefore greatly increasing the ease of analysis of sequencing data for insertional mutagenesis using retrovirus or transposons. I have also generated a self-inactivating transposon system called *'*Slingshot*'*, which can be conditionally activated by 4-OHT treatment for high-efficient mutagenesis screen in a variety of cell lines. I have performed proof-of-function screens and successfully identified genes responsible for resistance to puromycin and the anticancer drug vincristine. Therefore in cell culture systems, combining the Slingshot PB transposon for high-efficient mutagenesis and *iMapper* for automated insertion sites data analysis, it is

possible to perform mutagenesis screen for gene discovery in a time and cost-efficient manner. One obvious advantage of this pipeline in cell culture based mutagenesis screen is that the screen is easy to carry out with very little technology or equipment restriction. In principle, any research lab could carry out their own screen experiment by introducing the Slingshot PB system into their cellular background, performing the screening, and using *iMapper* for insertion site(s) analysis. Furthermore, the *iMapper* web tool is designed in a way that users are able to change the parameters for analysis to adapt their specific needs, such that for analyzing short sequences from 454/Roche or Illumina-Solexa sequencing.

6.2 Slingshot PB system in cell culture mutagenesis screen

In the *in vitro* cell culture screens, I have shown that Slingshot PB system is extremely efficient for gene discovery in ES cells. The efficacy of this system in other somatic cell lines, however, has not tested. One of the drawbacks for using this system in somatic cell culture screen is that somatic cells do not preferentially form colonies. One solution could be to harvest somatic cells in a pool, then identify and quantify the insertion sites by deep-sequencing technology such as 454/Roche. This solution is based on two assumptions: 1) The functional insertion sites can be enriched by selection in a pool; 2) All the insertion sites can be amplified by splinkerette PCR without bias. From unpublished results in our lab and other labs, splinkerette PCR is biased to some insertion sites based on size and genomic structure of the insertion site sequence, therefore the real functional insertion sites could be masked during amplification. Breaking genomic DNA into uniform sized fragment by glass bead sonication could be a solution to this problem, at least could reduce the PCR amplification bias towards short DNA fragments. If this pool based protocol for insertion site analysis can be established, Slingshot will become a powerful tool for gene identification in drug-resistance screens, cell differentiation studies and many other applications.

It should be noted that the Slingshot PB system has an obvious bias to gain-of-function mutagenesis. The majority of insertions during my proof-of-function screen experiments result in a gain-of-function. Due to the diploid nature of the mammalian genome, loss-of-functions insertions might not result in a phenotype, which would require disruption of both alleles in a cell. Considering the mutagenesis rate of our Slingshot system, the actual efficiency for causing a homozygous mutation could be as low as 9×10^{-9} (considering two copies of Slingshot PB are presented in one cell). It is possible to increase the recessive

mutagenesis rate for Slingshot system by delivering more copies of Slingshot cassette per cell using lipofectamine vesicle transformation, or alternatively performing the screen in Bloom ES cells which has a much higher efficiency for mitotic recombination to cause homozygous mutations (186).

6.3 Transposon-mediated insertional mutagenesis for cancer mouse model

During my research I have also tried to incorporate insertional mutagenesis tools for *in vivo* studies to generate mouse models of various cancer types. These experimental strategies are novel in that mutagenesis is taking place under a specific genetic background to trigger disease. These experiments are designed to keep a *'right'* balance between the oncogenic fusion protein and mutagenesis rate, so that mutagenesis rate is not too high to override the oncogenic effect of the fusion protein, and is also not too low to lose the efficiency of transformation. Although the *Brd4-NUT* mouse model did not transmit through the germ line after many attempts, B-cell leukaemia was derived from the *Tel-AML1* model allowing isolation of insertion sites from these tumours to identify the secondary mutations. Although this mouse model requires more characterization and validation work in the future, this is the first established *Tel-AML1* mouse model to induce disease similar to human cALL. Therefore, these experimental strategies proved that insertional mutagenesis system based on the *Sleeping Beauty* transposon can be efficiently used to model specific human cancers as well as cancer gene discovery in the mouse.

A trend in modern mouse cancer research is to mimic specific type of diseases by performing mutagenesis under a specific genetic background (such as the *Tel-AML1* model and *Brd4-NUT* models), or in a specific tissue type (such as liver or intestine) so that a specific cancer type could be modelled. For the latter application, transposon systems based on *Sleeping Beauty* or *piggyBac* could be extremely useful. By using various tissue-specific Cre mouse lines that are already available, it is possible to restrict expression of the transposase to specific tissue types to model specific cancers. A recent study demonstrated that using a hepatocyte-specific *Albumin*-Cre could drive SB-mediated mutagenesis restricted to liver cells, and could induce the hepatocellular carcinoma (HCC) (187). Similarly, another group showed that activation of the transposon in the gastrointestinal tract epithelium could give rise to neoplasia, adenomas and adenocarcinomas (188). These recent applications based on

transposon mutagenesis highlighted the potential of this strategy to identify tissue-specific cancer genes that are relevant for human cancer.

6.4 iMapper: improvements and future directions

With the advancement of parallel deep sequencing technology based on 454/Roche or Illumina-Solexa platform, it is possible to obtain more sequencing reads from insertional mutagenesis experiments. While the 454/Roche sequencing technology is able to generate relatively long sequencing reads (100-300 bp), the other cost efficient parallel sequencing platforms, Illumina-Solexa or SOLiD, can only generate short sequencing reads of around 35-60 bp. Therefore, developing appropriate mapping strategies for large numbers of short sequencing reads is a new challenge in insertional mutagenesis. The *iMapper* tool allows the adjustment of SSAHA mapping parameters to improve mapping results for short sequencing reads, however alternative mapping algorithms to map short sequences should be investigated to improve efficiency and accuracy. Maq and Short OligoNucleotide Alignment Program (SOAP) are recently developed mapping tools specifically for mapping short reads generated by parallel sequencing (189,190). Both Maq and SOAP take the same basic algorithm to build a hash table of short oligomers to align reads quickly with high sensitivity. The need to align short reads more efficiently has also arisen from recent human genome resequencing studies and the requirement for more efficient whole-genome comparison. 'Bowtie', an ultrafast, memory-efficient alignment software program for aligning short DNA sequence reads to large genomes, has been developed for this purpose. Unlike Maq or SOAP, Bowtie employs a Burrows-Wheeler index-based index that is 35 times faster than Maq and 300 times faster than SOAP for the alignment of short sequences around 35 bp, under the same conditions (191). In addition, studies that have compared different mapping tools for analysing short sequence reads also recommend a commercial tool called 'CLC NGS Cell' (www.clcbio.com) when considering both the mapping efficiency and accuracy (192). In the future it is possible that *iMapper* could incorporate different mapping algorithms to achieve the optimal mapping results for analysing sequencing reads.

In mouse cancer studies, high-throughput protocols have been developed to analyse tens and hundreds of mouse tumour insertion sites using splinkerette PCR and the 454/Roche parallel sequencing platform (193). In these studies a barcode sequence is incorporated into the splinkerette primers to enable the splinkerette products from multiples tumours to be pooled

together for deep sequencing and to use the barcode as a tag to determine which sequence belongs to which tumour within the pool. The functionality of *iMapper* would be enhanced if a barcode sequence recognition step was implemented during sequence processing. If so, when a mutagen tag sequence is identified *iMapper* could search the upstream sequence to identify the barcode and assign this sequence to a specific tumour. In the GFF file output, *iMapper* could generate a column of barcode sequences associated with each identified insertion site, and users categorise these insertion sites with specific tumour samples using commercial software such as Excel.

Insertional mutagenesis studies have shown that in multiple independent samples some regions of the genome are hit by viral or transposon insertions significantly more than expected by chance. These regions are termed Common Insertion Sites (CIS) and they are highly likely to contain candidate genes identified from the screen. These CISs arise from the random nature of the insertion process, as well as the bias - stemming from preferential insertion sites present in the genome by transposons or retroviruses. There are several statistical models to determine CISs from both the background noise. One of the most common models uses the kernel convolution (KC) framework to find CISs in a noisy and biased environment using a predefined significance level while controlling the probability of detecting false CISs (194). With the increased capacity for *iMapper* to process large numbers of sequencing reads, it would also be worthwhile to incorporate a kernel convolution model into *iMapper* for automatic detection of CISs.

REFERENCES

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
2. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.
3. Hartwell, L.H., Culotti, J. and Reid, B. (1970) Genetic control of the cell-division cycle in yeast. I. Detection of mutants. *Proceedings of the National Academy of Sciences of the United States of America*, **66**, 352-359.
4. Nurse, P., Thuriaux, P. and Nasmyth, K. (1976) Genetic control of the cell division cycle in the fission yeast Schizosaccharomyces pombe. *Mol Gen Genet*, **146**, 167-178.
5. Nusslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in Drosophila. *Nature*, **287**, 795-801.
6. Ellis, H.M. and Horvitz, H.R. (1986) Genetic control of programmed cell death in the nematode C. elegans. *Cell*, **44**, 817-829.
7. Auwerx, J., Avner, P., Baldock, R., Ballabio, A., Balling, R., Barbacid, M., Berns, A., Bradley, A., Brown, S., Carmeliet, P. *et al.* (2004) The European dimension for the mouse genome mutagenesis program. *Nature genetics*, **36**, 925-927.
8. Austin, C.P., Battey, J.F., Bradley, A., Bucan, M., Capecchi, M., Collins, F.S., Dove, W.F., Duyk, G., Dymecki, S., Eppig, J.T. *et al.* (2004) The knockout mouse project. *Nature genetics*, **36**, 921-924.
9. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806-811.
10. Bernstein, E., Caudy, A.A., Hammond, S.M. and Hannon, G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363-366.
11. Liu, Q., Rand, T.A., Kalidas, S., Du, F., Kim, H.E., Smith, D.P. and Wang, X. (2003) R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science*, **301**, 1921-1925.
12. Zamore, P.D., Tuschl, T., Sharp, P.A. and Bartel, D.P. (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, **101**, 25-33.
13. Naito, Y., Yamada, T., Matsumiya, T., Ui-Tei, K., Saigo, K. and Morishita, S. (2005) dsCheck: highly sensitive off-target search software for double-stranded RNA-mediated RNA interference. *Nucleic acids research*, **33**, W589-591.
14. Henschel, A., Buchholz, F. and Habermann, B. (2004) DEQOR: a web-based tool for the design and quality control of siRNAs. *Nucleic acids research*, **32**, W113-120.
15. Stanford, W.L., Cohn, J.B. and Cordes, S.P. (2001) Gene-trap mutagenesis: past, present and beyond. *Nature reviews*, **2**, 756-768.
16. Justice, M.J., Noveroske, J.K., Weber, J.S., Zheng, B. and Bradley, A. (1999) Mouse ENU mutagenesis. *Human molecular genetics*, **8**, 1955-1963.
17. Chen, Y., Yee, D., Dains, K., Chatterjee, A., Cavalcoli, J., Schneider, E., Om, J., Woychik, R.P. and Magnuson, T. (2000) Genotype-based screen for ENU-induced mutations in mouse embryonic stem cells. *Nature genetics*, **24**, 314-317.
18. Guenet, J.L. (2004) Chemical mutagenesis of the mouse genome: an overview. *Genetica*, **122**, 9-24.
19. Vitaterna, M.H., King, D.P., Chang, A.M., Kornhauser, J.M., Lowrey, P.L., McDonald, J.D., Dove, W.F., Pinto, L.H., Turek, F.W. and Takahashi, J.S. (1994) Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. *Science (New York, N.Y*, **264**, 719-725.

20. Kasarskis, A., Manova, K. and Anderson, K.V. (1998) A phenotype-based screen for embryonic lethal mutations in the mouse. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 7485-7490.
21. Cordes, S.P. (2005) N-ethyl-N-nitrosourea mutagenesis: boarding the mouse mutant express. *Microbiol Mol Biol Rev*, **69**, 426-439.
22. van der Weyden, L., Adams, D.J. and Bradley, A. (2002) Tools for targeted manipulation of the mouse genome. *Physiological genomics*, **11**, 133-164.
23. Frankenberg-Schwager, M. (1990) Induction, repair and biological relevance of radiation-induced DNA lesions in eukaryotic cells. *Radiation and environmental biophysics*, **29**, 273-292.
24. Goodhead, D.T. (1989) The initial physical damage produced by ionizing radiations. *International journal of radiation biology*, **56**, 623-634.
25. Muller, H.J. (1927) Artificial Transmutation of the Gene. *Science (New York, N.Y*, **66**, 84-87.
26. Green, E.L., Roderick, T. H. . (1966) *Radiation Genetics, Biology of the Laboratory Mouse*. New York, McGraw-Hill.
27. Lobrich, M., Cooper, P.K. and Rydberg, B. (1996) Non-random distribution of DNA double-strand breaks induced by particle irradiation. *International journal of radiation biology*, **70**, 493-503.
28. Kushi, A., Edamura, K., Noguchi, M., Akiyama, K., Nishi, Y. and Sasai, H. (1998) Generation of mutant mice with large chromosomal deletion by use of irradiated ES cells--analysis of large deletion around hprt locus of ES cell. *Mamm Genome*, **9**, 269-273.
29. Chick, W.S., Mentzer, S.E., Carpenter, D.A., Rinchik, E.M., Johnson, D. and You, Y. (2005) X-ray-induced deletion complexes in embryonic stem cells on mouse chromosome 15. *Mamm Genome*, **16**, 661-671.
30. Schimenti, J.C., Libby, B.J., Bergstrom, R.A., Wilson, L.A., Naf, D., Tarantino, L.M., Alavizadeh, A., Lengeling, A. and Bucan, M. (2000) Interdigitated deletion complexes on mouse chromosome 5 induced by irradiation of embryonic stem cells. *Genome research*, **10**, 1043-1050.
31. Jaenisch, R. (1976) Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. *Proceedings of the National Academy of Sciences of the United States of America*, **73**, 1260-1264.
32. Morgenstern, J.P. and Land, H. (1990) Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic acids research*, **18**, 3587-3596.
33. von Melchner, H. and Ruley, H.E. (1989) Identification of cellular promoters by using a retrovirus promoter trap. *Journal of virology*, **63**, 3227-3233.
34. Friedrich, G. and Soriano, P. (1991) Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes & development*, **5**, 1513-1523.
35. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.
36. Mc, C.B. (1950) The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, **36**, 344-355.
37. Hu, S., Otsubo, E., Davidson, N. and Saedler, H. (1975) Electron microscope heteroduplex studies of sequence relations among bacterial plasmids: identification and mapping of the insertion sequences IS1 and IS2 in F and R plasmids. *Journal of bacteriology*, **122**, 764-775.
38. Spradling, A.C. and Rubin, G.M. (1982) Transposition of cloned P elements into Drosophila germ line chromosomes. *Science (New York, N.Y*, **218**, 341-347.
39. Bingham, P.M., Kidwell, M.G. and Rubin, G.M. (1982) The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell*, **29**, 995-1004.

40. Rubin, G.M., Kidwell, M.G. and Bingham, P.M. (1982) The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*, **29**, 987-994.
41. O'Hare, K. and Rubin, G.M. (1983) Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome. *Cell*, **34**, 25-35.
42. Handler, A.M., Gomez, S.P. and O'Brochta, D.A. (1993) A functional analysis of the P-element gene-transfer vector in insects. *Archives of insect biochemistry and physiology*, **22**, 373-384.
43. Langin, T., Capy, P. and Daboussi, M.J. (1995) The transposable element impala, a fungal member of the Tc1-mariner superfamily. *Mol Gen Genet*, **246**, 19-28.
44. Emmons, S.W., Yesner, L., Ruan, K.S. and Katzenberg, D. (1983) Evidence for a transposon in Caenorhabditis elegans. *Cell*, **32**, 55-65.
45. Jacobson, J.W., Medhora, M.M. and Hartl, D.L. (1986) Molecular structure of a somatically unstable transposable element in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 8684-8688.
46. Plasterk, R.H., Izsvak, Z. and Ivics, Z. (1999) Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet*, **15**, 326-332.
47. Ivics, Z., Hackett, P.B., Plasterk, R.H. and Izsvak, Z. (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**, 501-510.
48. Yant, S.R., Meuse, L., Chiu, W., Ivics, Z., Izsvak, Z. and Kay, M.A. (2000) Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nature genetics*, **25**, 35-41.
49. Carlson, C.M., Dupuy, A.J., Fritz, S., Roberg-Perez, K.J., Fletcher, C.F. and Largaespada, D.A. (2003) Transposon mutagenesis of the mouse germline. *Genetics*, **165**, 243-256.
50. Horie, K., Yusa, K., Yae, K., Odajima, J., Fischer, S.E., Keng, V.W., Hayakawa, T., Mizuno, S., Kondoh, G., Ijiri, T. *et al.* (2003) Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Molecular and cellular biology*, **23**, 9189-9207.
51. Geurts, A.M., Yang, Y., Clark, K.J., Liu, G., Cui, Z., Dupuy, A.J., Bell, J.B., Largaespada, D.A. and Hackett, P.B. (2003) Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther*, **8**, 108-117.
52. Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E. and Fraser, M.J. (1989) Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*, **172**, 156-169.
53. Fraser, M.J., Cary, L., Boonvisudhi, K. and Wang, H.G. (1995) Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. *Virology*, **211**, 397-407.
54. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. *et al.* (1996) Life with 6000 genes. *Science (New York, N.Y*, **274**, 546, 563-547.
55. Burns, N., Grimwade, B., Ross-Macdonald, P.B., Choi, E.Y., Finberg, K., Roeder, G.S. and Snyder, M. (1994) Large-scale analysis of gene expression, protein localization, and gene disruption in Saccharomyces cerevisiae. *Genes & development*, **8**, 1087-1105.
56. Ross-Macdonald, P., Coelho, P.S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413-418.
57. Rubin, G.M. and Spradling, A.C. (1982) Genetic transformation of Drosophila with transposable element vectors. *Science (New York, N.Y*, **218**, 348-353.
58. Cooley, L., Kelley, R. and Spradling, A. (1988) Insertional mutagenesis of the Drosophila genome with single P elements. *Science (New York, N.Y*, **239**, 1121-1128.
59. Spradling, A.C., Stern, D.M., Kiss, I., Roote, J., Laverty, T. and Rubin, G.M. (1995) Gene disruptions using P transposable elements: an integral component of the Drosophila genome

project. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 10824-10830.
60. Spradling, A.C., Stern, D., Beaton, A., Rhem, E.J., Laverty, T., Mozden, N., Misra, S. and Rubin, G.M. (1999) The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics*, **153**, 135-177.
61. Bellen, H.J., Levis, R.W., Liao, G., He, Y., Carlson, J.W., Tsang, G., Evans-Holm, M., Hiesinger, P.R., Schulze, K.L., Rubin, G.M. *et al.* (2004) The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. *Genetics*, **167**, 761-781.
62. Sijen, T. and Plasterk, R.H. (2003) Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature*, **426**, 310-314.
63. Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A. *et al.* (2003) The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS biology*, **1**, E45.
64. Bessereau, J.L., Wright, A., Williams, D.C., Schuske, K., Davis, M.W. and Jorgensen, E.M. (2001) Mobilization of a Drosophila transposon in the Caenorhabditis elegans germ line. *Nature*, **413**, 70-74.
65. Greenwald, I. (1985) lin-12, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. *Cell*, **43**, 583-590.
66. Moerman, D.G., Benian, G.M. and Waterston, R.H. (1986) Molecular cloning of the muscle gene unc-22 in Caenorhabditis elegans by Tc1 transposon tagging. *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 2579-2583.
67. Wicks, S.R., de Vries, C.J., van Luenen, H.G. and Plasterk, R.H. (2000) CHE-3, a cytosolic dynein heavy chain, is required for sensory cilia structure and function in Caenorhabditis elegans. *Developmental biology*, **221**, 295-307.
68. Rushforth, A.M. and Anderson, P. (1996) Splicing removes the Caenorhabditis elegans transposon Tc1 from most mutant pre-mRNAs. *Molecular and cellular biology*, **16**, 422-429.
69. Rushforth, A.M., Saari, B. and Anderson, P. (1993) Site-selected insertion of the transposon Tc1 into a Caenorhabditis elegans myosin light chain gene. *Molecular and cellular biology*, **13**, 902-910.
70. Plasterk, R.H. and Groenen, J.T. (1992) Targeted alterations of the Caenorhabditis elegans genome by transgene instructed DNA double strand break repair following Tc1 excision. *The EMBO journal*, **11**, 287-290.
71. Barrett, P.L., Fleming, J.T. and Gobel, V. (2004) Targeted gene alteration in Caenorhabditis elegans by gene conversion. *Nature genetics*, **36**, 1231-1237.
72. Lohe, A.R., Moriyama, E.N., Lidholm, D.A. and Hartl, D.L. (1995) Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Molecular biology and evolution*, **12**, 62-72.
73. Dasgupta, M., Agarwal, M.K., Varley, P., Lu, T., Stark, G.R. and Kandel, E.S. (2008) Transposon-based mutagenesis identifies short RIP1 as an activator of NFkappaB. *Cell cycle (Georgetown, Tex*, **7**, 2249-2256.
74. Izsvak, Z., Chuah, M.K., Vandendriessche, T. and Ivics, Z. (2009) Efficient stable gene transfer into human cells by the Sleeping Beauty transposon vectors. *Methods (San Diego, Calif*, **49**, 287-297.
75. Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y. and Xu, T. (2005) Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell*, **122**, 473-483.
76. Liang, Q., Kong, J., Stalker, J. and Bradley, A. (2009) Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *Genesis*, **47**, 404-408.
77. Schuldiner, O., Berdnik, D., Levy, J.M., Wu, J.S., Luginbuhl, D., Gontang, A.C. and Luo, L. (2008) piggyBac-based mosaic screen identifies a postmitotic function for cohesin in regulating developmental axon pruning. *Developmental cell*, **14**, 227-238.

78. Kinzler, K.W. and Vogelstein, B. (1996) Lessons from hereditary colorectal cancer. *Cell*, **87**, 159-170.
79. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nature reviews*, **4**, 177-183.
80. Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K. *et al.* (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *The Journal of clinical investigation*, **118**, 3132-3142.
81. Santos, E., Tronick, S.R., Aaronson, S.A., Pulciani, S. and Barbacid, M. (1982) T24 human bladder carcinoma oncogene is an activated form of the normal human homologue of BALB- and Harvey-MSV transforming genes. *Nature*, **298**, 343-347.
82. Tabin, C.J., Bradley, S.M., Bargmann, C.I., Weinberg, R.A., Papageorge, A.G., Scolnick, E.M., Dhar, R., Lowy, D.R. and Chang, E.H. (1982) Mechanism of activation of a human oncogene. *Nature*, **300**, 143-149.
83. Stehelin, D., Varmus, H.E., Bishop, J.M. and Vogt, P.K. (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, **260**, 170-173.
84. Muriaux, D. and Rein, A. (2003) Encapsidation and transduction of cellular genes by retroviruses. *Front Biosci*, **8**, d135-142.
85. Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A. and Copeland, N.G. (2004) RTCGD: retroviral tagged cancer gene database. *Nucleic acids research*, **32**, D523-527.
86. Uren, A.G., Kool, J., Matentzoglu, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T. *et al.* (2008) Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell*, **133**, 727-741.
87. Wu, X., Li, Y., Crise, B. and Burgess, S.M. (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science (New York, N.Y*, **300**, 1749-1751.
88. Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. and Largaespada, D.A. (2005) Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature*, **436**, 272-276.
89. Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. and Jenkins, N.A. (2005) Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature*, **436**, 221-226.
90. Evans, M.J. and Kaufman, M.H. (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature*, **292**, 154-156.
91. Bradley, A., Evans, M., Kaufman, M.H. and Robertson, E. (1984) Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*, **309**, 255-256.
92. Zijlstra, M., Li, E., Sajjadi, F., Subramani, S. and Jaenisch, R. (1989) Germ-line transmission of a disrupted beta 2-microglobulin gene produced by homologous recombination in embryonic stem cells. *Nature*, **342**, 435-438.
93. Koller, B.H., Marrack, P., Kappler, J.W. and Smithies, O. (1990) Normal development of mice deficient in beta 2M, MHC class I proteins, and CD8+ T cells. *Science (New York, N.Y*, **248**, 1227-1230.
94. McMahon, A.P. and Bradley, A. (1990) The Wnt-1 (int-1) proto-oncogene is required for development of a large region of the mouse brain. *Cell*, **62**, 1073-1085.
95. Schwartzberg, P.L., Robertson, E.J. and Goff, S.P. (1990) Targeted gene disruption of the endogenous c-abl locus by homologous recombination with DNA encoding a selectable fusion protein. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 3210-3214.
96. Nagy, A. (2000) Cre recombinase: the universal reagent for genome tailoring. *Genesis*, **26**, 99-109.

97. Cuypers, H.T., Selten, G., Quint, W., Zijlstra, M., Maandag, E.R., Boelens, W., van Wezenbeek, P., Melief, C. and Berns, A. (1984) Murine leukemia virus-induced T-cell lymphomagenesis: integration of proviruses in a distinct chromosomal region. *Cell*, **37**, 141-150.
98. Ochman, H., Gerber, A.S. and Hartl, D.L. (1988) Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120**, 621-623.
99. McAleer, M.A., Coffey, A.J. and Dunham, I. (2003) DNA rescue by the vectorette method. *Methods in molecular biology (Clifton, N.J*, **226**, 393-400.
100. Devon, R.S., Porteous, D.J. and Brookes, A.J. (1995) Splinkerettes--improved vectorettes for greater efficiency in PCR walking. *Nucleic acids research*, **23**, 1644-1645.
101. Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol*, **7**, 203-214.
102. Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725-1729.
103. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome research*, **12**, 656-664.
104. Kong, J., Zhu, F., Stalker, J. and Adams, D.J. (2008) iMapper: a web application for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes. *Bioinformatics*, **24**, 2923-2925.
105. Kong, J., Wang, F., Brenton, J.D. and Adams, D.J. Slingshot: a PiggyBac based transposon system for tamoxifen-inducible 'self-inactivating' insertional mutagenesis. *Nucleic acids research*.
106. Mikkers, H. and Berns, A. (2003) Retroviral insertional mutagenesis: tagging cancer pathways. *Adv Cancer Res*, **88**, 53-99.
107. Du, Y., Jenkins, N.A. and Copeland, N.G. (2005) Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. *Blood*, **106**, 3932-3939.
108. Luo, G., Ivics, Z., Izsvak, Z. and Bradley, A. (1998) Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 10769-10773.
109. Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A. and Liu, P. (2008) Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc Natl Acad Sci U S A*, **105**, 9290-9295.
110. Craig, N.L., Atkinson, P. (2008), *Sixth Annual Internation Conference on Transposition and Animal Biotechnology*, Berlin, pp. 17.
111. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.
112. Plasterk, R.H. (1996) The Tc1/mariner transposon family. *Current topics in microbiology and immunology*, **204**, 125-143.
113. Osborne, B.I. and Baker, B. (1995) Movers and shakers: maize transposons as tools for analyzing other plant genomes. *Current opinion in cell biology*, **7**, 406-413.
114. Kawakami, K. (2007) Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol*, **8 Suppl 1**, S7.
115. Pavlopoulos, A., Oehler, S., Kapetanaki, M.G. and Savakis, C. (2007) The DNA transposon Minos as a tool for transgenesis and functional genomic analysis in vertebrates and invertebrates. *Genome Biol*, **8 Suppl 1**, S2.
116. Fischer, S.E., Wienholds, E. and Plasterk, R.H. (2001) Regulated transposition of a fish transposon in the mouse germ line. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 6759-6764.
117. Dupuy, A.J., Fritz, S. and Largaespada, D.A. (2001) Transposition and gene disruption in the male germline of the mouse. *Genesis*, **30**, 82-88.
118. Horie, K., Kuroiwa, A., Ikawa, M., Okabe, M., Kondoh, G., Matsuda, Y. and Takeda, J. (2001) Efficient chromosomal transposition of a Tc1/mariner- like transposon Sleeping Beauty in

mice. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 9191-9196.
119. Izsvak, Z., Ivics, Z. and Plasterk, R.H. (2000) Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *Journal of molecular biology*, **302**, 93-102.
120. Yusa, K., Rad, R., Takeda, J. and Bradley, A. (2009) Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat Methods*, **6**, 363-369.
121. Wang, W., Bradley, A. and Huang, Y. (2009) A piggyBac transposon-based genome-wide library of insertionally mutated Blm-deficient murine ES cells. *Genome Res*, **19**, 667-673.
122. Cadinanos, J. and Bradley, A. (2007) Generation of an inducible and optimized piggyBac transposon system. *Nucleic acids research*, **35**, e87.
123. Wu, S.C., Meir, Y.J., Coates, C.J., Handler, A.M., Pelczar, P., Moisyadi, S. and Kaminski, J.M. (2006) piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 15008-15013.
124. Okabe, M., Ikawa, M., Kominami, K., Nakanishi, T. and Nishimune, Y. (1997) 'Green mice' as a source of ubiquitous green cells. *FEBS letters*, **407**, 313-319.
125. Theile, D., Staffen, B. and Weiss, J. ATP-binding cassette transporters as pitfalls in selection of transgenic cells. *Analytical biochemistry*, **399**, 246-250.
126. Pannell, D. and Ellis, J. (2001) Silencing of gene expression: implications for design of retrovirus vectors. *Reviews in medical virology*, **11**, 205-217.
127. Chung, J.H., Whiteley, M. and Felsenfeld, G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, **74**, 505-514.
128. Li, C.L. and Emery, D.W. (2008) The cHS4 chromatin insulator reduces gammaretroviral vector silencing by epigenetic modifications of integrated provirus. *Gene therapy*, **15**, 49-53.
129. Emery, D.W., Yannaki, E., Tubb, J. and Stamatoyannopoulos, G. (2000) A chromatin insulator protects retrovirus vectors from chromosomal position effects. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 9150-9155.
130. Taj, A.S., Ross, F.M., Vickers, M., Choudhury, D.N., Harvey, J.F., Barber, J.C., Barton, C. and Smith, A.G. (1995) t(8;21) myelodysplasia, an early presentation of M2 AML. *British journal of haematology*, **89**, 890-892.
131. Kurzrock, R., Kantarjian, H.M., Druker, B.J. and Talpaz, M. (2003) Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics. *Annals of internal medicine*, **138**, 819-830.
132. De Braekeleer, M. and Dao, T.N. (1991) Cytogenetic studies in male infertility: a review. *Human reproduction (Oxford, England)*, **6**, 245-250.
133. Rowley, J.D. (1998) The critical role of chromosome translocations in human leukemias. *Annu Rev Genet*, **32**, 495-519.
134. Papadopoulos, P., Ridge, S.A., Boucher, C.A., Stocking, C. and Wiedemann, L.M. (1995) The novel activation of ABL by fusion to an ets-related gene, TEL. *Cancer Res*, **55**, 34-38.
135. Golub, T.R., Barker, G.F., Lovett, M. and Gilliland, D.G. (1994) Fusion of PDGF receptor beta to a novel ets-like gene, tel, in chronic myelomonocytic leukemia with t(5;12) chromosomal translocation. *Cell*, **77**, 307-316.
136. Buijs, A., Sherr, S., van Baal, S., van Bezouw, S., van der Plas, D., Geurts van Kessel, A., Riegman, P., Lekanne Deprez, R., Zwarthoff, E., Hagemeijer, A. *et al.* (1995) Translocation (12;22) (p13;q11) in myeloproliferative disorders results in fusion of the ETS-like TEL gene on 12p13 to the MN1 gene on 22q11. *Oncogene*, **10**, 1511-1519.
137. Romana, S.P., Poirel, H., Leconiat, M., Flexor, M.A., Mauchauffe, M., Jonveaux, P., Macintyre, E.A., Berger, R. and Bernard, O.A. (1995) High frequency of t(12;21) in childhood B-lineage acute lymphoblastic leukemia. *Blood*, **86**, 4263-4269.

138. Licht, J.D. (2001) AML1 and the AML1-ETO fusion protein in the pathogenesis of t(8;21) AML. *Oncogene*, **20**, 5660-5679.
139. Greaves, M.F., Maia, A.T., Wiemels, J.L. and Ford, A.M. (2003) Leukemia in twins: lessons in natural history. *Blood*, **102**, 2321-2333.
140. Hong, D., Gupta, R., Ancliff, P., Atzberger, A., Brown, J., Soneji, S., Green, J., Colman, S., Piacibello, W., Buckle, V. *et al.* (2008) Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science*, **319**, 336-339.
141. Jousset, C., Carron, C., Boureux, A., Quang, C.T., Oury, C., Dusanter-Fourt, I., Charon, M., Levin, J., Bernard, O. and Ghysdael, J. (1997) A domain of TEL conserved in a subset of ETS proteins defines a specific oligomerization interface essential to the mitogenic properties of the TEL-PDGFR beta oncoprotein. *Embo J*, **16**, 69-82.
142. Bernard, O.A., Romana, S.P., Poirel, H. and Berger, R. (1996) Molecular cytogenetics of t(12;21) (p13;q22). *Leuk Lymphoma*, **23**, 459-465.
143. Miyoshi, H., Shimizu, K., Kozu, T., Maseki, N., Kaneko, Y. and Ohki, M. (1991) t(8;21) breakpoints on chromosome 21 in acute myeloid leukemia are clustered within a limited region of a single gene, AML1. *Proc Natl Acad Sci U S A*, **88**, 10431-10434.
144. Mitani, K., Ogawa, S., Tanaka, T., Miyoshi, H., Kurokawa, M., Mano, H., Yazaki, Y., Ohki, M. and Hirai, H. (1994) Generation of the AML1-EVI-1 fusion gene in the t(3;21)(q26;q22) causes blastic crisis in chronic myelocytic leukemia. *Embo J*, **13**, 504-510.
145. Gamou, T., Kitamura, E., Hosoda, F., Shimizu, K., Shinohara, K., Hayashi, Y., Nagase, T., Yokoyama, Y. and Ohki, M. (1998) The partner gene of AML1 in t(16;21) myeloid malignancies is a novel member of the MTG8(ETO) family. *Blood*, **91**, 4028-4037.
146. Speck, N.A. and Gilliland, D.G. (2002) Core-binding factors in haematopoiesis and leukaemia. *Nat Rev Cancer*, **2**, 502-513.
147. Mori, H., Colman, S.M., Xiao, Z., Ford, A.M., Healy, L.E., Donaldson, C., Hows, J.M., Navarrete, C. and Greaves, M. (2002) Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci U S A*, **99**, 8242-8247.
148. Andreasson, P., Schwaller, J., Anastasiadou, E., Aster, J. and Gilliland, D.G. (2001) The expression of ETV6/CBFA2 (TEL/AML1) is not sufficient for the transformation of hematopoietic cell lines in vitro or the induction of hematologic disease in vivo. *Cancer Genet Cytogenet*, **130**, 93-104.
149. Raynaud, S., Cave, H., Baens, M., Bastard, C., Cacheux, V., Grosgeorge, J., Guidal-Giroux, C., Guo, C., Vilmer, E., Marynen, P. *et al.* (1996) The 12;21 translocation involving TEL and deletion of the other TEL allele: two frequently associated alterations found in childhood acute lymphoblastic leukemia. *Blood*, **87**, 2891-2899.
150. Tsuzuki, S., Karnan, S., Horibe, K., Matsumoto, K., Kato, K., Inukai, T., Goi, K., Sugita, K., Nakazawa, S., Kasugai, Y. *et al.* (2007) Genetic abnormalities involved in t(12;21) TEL-AML1 acute lymphoblastic leukemia: analysis by means of array-based comparative genomic hybridization. *Cancer Sci*, **98**, 698-706.
151. Sabaawy, H.E., Azuma, M., Embree, L.J., Tsai, H.J., Starost, M.F. and Hickstein, D.D. (2006) TEL-AML1 transgenic zebrafish model of precursor B cell acute lymphoblastic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 15166-15171.
152. Rad, R., Gerhard, M., Lang, R., Schoniger, M., Rosch, T., Schepp, W., Becker, I., Wagner, H. and Prinz, C. (2002) The Helicobacter pylori blood group antigen-binding adhesin facilitates bacterial colonization and augments a nonspecific immune response. *J Immunol*, **168**, 3033-3041.
153. Poirel, H., Lacronique, V., Mauchauffe, M., Le Coniat, M., Raffoux, E., Daniel, M.T., Erickson, P., Drabkin, H., MacLeod, R.A., Drexler, H.G. *et al.* (1998) Analysis of TEL proteins in human leukemias. *Oncogene*, **16**, 2895-2903.

154. Tsuzuki, S., Seto, M., Greaves, M. and Enver, T. (2004) Modeling first-hit functions of the t(12;21) TEL-AML1 translocation in mice. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 8443-8448.
155. Fischer, M., Schwieger, M., Horn, S., Niebuhr, B., Ford, A., Roscher, S., Bergholz, U., Greaves, M., Lohler, J. and Stocking, C. (2005) Defining the oncogenic function of the TEL/AML1 (ETV6/RUNX1) fusion protein in a mouse model. *Oncogene*, **24**, 7579-7591.
156. Rho, J.K., Kim, J.H., Yu, J. and Choe, S.Y. (2002) Correlation between cellular localization of TEL/AML1 fusion protein and repression of AML1-mediated transactivation of CR1 gene. *Biochem Biophys Res Commun*, **297**, 91-95.
157. Wang, L.C., Kuo, F., Fujiwara, Y., Gilliland, D.G., Golub, T.R. and Orkin, S.H. (1997) Yolk sac angiogenic defect and intra-embryonic apoptosis in mice lacking the Ets-related factor TEL. *Embo J*, **16**, 4374-4383.
158. Morse, S.F. and Marston, P.L. (2002) Backscattering of transients by tilted truncated cylindrical shells: time-frequency identification of ray contributions from measurements. *The Journal of the Acoustical Society of America*, **111**, 1289-1294.
159. Kogan, S.C., Ward, J.M., Anver, M.R., Berman, J.J., Brayton, C., Cardiff, R.D., Carter, J.S., de Coronado, S., Downing, J.R., Fredrickson, T.N. *et al.* (2002) Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice. *Blood*, **100**, 238-245.
160. Hingorani, S.R., Petricoin, E.F., Maitra, A., Rajapakse, V., King, C., Jacobetz, M.A., Ross, S., Conrads, T.P., Veenstra, T.D., Hitt, B.A. *et al.* (2003) Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell*, **4**, 437-450.
161. Haruki, N., Kawaguchi, K.S., Eichenberger, S., Massion, P.P., Gonzalez, A., Gazdar, A.F., Minna, J.D., Carbone, D.P. and Dang, T.P. (2005) Cloned fusion product from a rare t(15;19)(q13.2;p13.1) inhibit S phase in vitro. *J Med Genet*, **42**, 558-564.
162. Li, Z., Tognon, C.E., Godinho, F.J., Yasaitis, L., Hock, H., Herschkowitz, J.I., Lannon, C.L., Cho, E., Kim, S.J., Bronson, R.T. *et al.* (2007) ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex. *Cancer Cell*, **12**, 542-558.
163. Pierotti, M.A. (2001) Chromosomal rearrangements in thyroid carcinomas: a recombination or death dilemma. *Cancer letters*, **166**, 1-7.
164. Kroll, T.G., Sarraf, P., Pecciarini, L., Chen, C.J., Mueller, E., Spiegelman, B.M. and Fletcher, J.A. (2000) PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma [corrected]. *Science (New York, N.Y.*, **289**, 1357-1360.
165. Martins, C., Cavaco, B., Tonon, G., Kaye, F.J., Soares, J. and Fonseca, I. (2004) A study of MECT1-MAML2 in mucoepidermoid carcinoma and Warthin's tumor of salivary glands. *J Mol Diagn*, **6**, 205-210.
166. Behboudi, A., Enlund, F., Winnes, M., Andren, Y., Nordkvist, A., Leivo, I., Flaberg, E., Szekely, L., Makitie, A., Grenman, R. *et al.* (2006) Molecular classification of mucoepidermoid carcinomas-prognostic significance of the MECT1-MAML2 fusion oncogene. *Genes, chromosomes & cancer*, **45**, 470-481.
167. Tognon, C., Knezevich, S.R., Huntsman, D., Roskelley, C.D., Melnyk, N., Mathers, J.A., Becker, L., Carneiro, F., MacPherson, N., Horsman, D. *et al.* (2002) Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer cell*, **2**, 367-376.
168. Makretsov, N., He, M., Hayes, M., Chia, S., Horsman, D.E., Sorensen, P.H. and Huntsman, D.G. (2004) A fluorescence in situ hybridization study of ETV6-NTRK3 fusion gene in secretory breast carcinoma. *Genes, chromosomes & cancer*, **40**, 152-157.
169. French, C.A., Kutok, J.L., Faquin, W.C., Toretsky, J.A., Antonescu, C.R., Griffin, C.A., Nose, V., Vargas, S.O., Moschovi, M., Tzortzatou-Stathopoulou, F. *et al.* (2004) Midline carcinoma of children and young adults with NUT rearrangement. *J Clin Oncol*, **22**, 4135-4139.

170. French, C.A., Miyoshi, I., Aster, J.C., Kubonishi, I., Kroll, T.G., Dal Cin, P., Vargas, S.O., Perez-Atayde, A.R. and Fletcher, J.A. (2001) BRD4 bromodomain gene rearrangement in aggressive carcinoma with translocation t(15;19). *The American journal of pathology*, **159**, 1987-1992.
171. Engleson, J., Soller, M., Panagopoulos, I., Dahlen, A., Dictor, M. and Jerkeman, M. (2006) Midline carcinoma with t(15;19) and BRD4-NUT fusion oncogene in a 30-year-old female with response to docetaxel and radiotherapy. *BMC cancer*, **6**, 69.
172. Winston, F. and Allis, C.D. (1999) The bromodomain: a chromatin-targeting module? *Nature structural biology*, **6**, 601-604.
173. Dey, A., Ellenberg, J., Farina, A., Coleman, A.E., Maruyama, T., Sciortino, S., Lippincott-Schwartz, J. and Ozato, K. (2000) A bromodomain protein, MCAP, associates with mitotic chromosomes and affects G(2)-to-M transition. *Molecular and cellular biology*, **20**, 6537-6549.
174. Maruyama, T., Farina, A., Dey, A., Cheong, J., Bermudez, V.P., Tamura, T., Sciortino, S., Shuman, J., Hurwitz, J. and Ozato, K. (2002) A Mammalian bromodomain protein, brd4, interacts with replication factor C and inhibits progression to S phase. *Molecular and cellular biology*, **22**, 6509-6520.
175. French, C.A., Ramirez, C.L., Kolmakova, J., Hickman, T.T., Cameron, M.J., Thyne, M.E., Kutok, J.L., Toretsky, J.A., Tadavarthy, A.K., Kees, U.R. *et al.* (2007) BRD-NUT oncoproteins: a family of closely related nuclear proteins that block epithelial differentiation and maintain the growth of carcinoma cells. *Oncogene*.
176. French, C.A., Ramirez, C.L., Kolmakova, J., Hickman, T.T., Cameron, M.J., Thyne, M.E., Kutok, J.L., Toretsky, J.A., Tadavarthy, A.K., Kees, U.R. *et al.* (2008) BRD-NUT oncoproteins: a family of closely related nuclear proteins that block epithelial differentiation and maintain the growth of carcinoma cells. *Oncogene*, **27**, 2237-2242.
177. Kees, U.R., Mulcahy, M.T. and Willoughby, M.L. (1991) Intrathoracic carcinoma in an 11-year-old girl showing a translocation t(15;19). *The American journal of pediatric hematology/oncology*, **13**, 459-464.
178. Lee, A.C., Kwong, Y.I., Fu, K.H., Chan, G.C., Ma, L. and Lau, Y.L. (1993) Disseminated mediastinal carcinoma with chromosomal translocation (15;19). A distinctive clinicopathologic syndrome. *Cancer*, **72**, 2273-2276.
179. Kubonishi, I., Takehara, N., Iwata, J., Sonobe, H., Ohtsuki, Y., Abe, T. and Miyoshi, I. (1991) Novel t(15;19)(q15;p13) chromosome abnormality in a thymic carcinoma. *Cancer research*, **51**, 3327-3328.
180. Vargas, S.O., French, C.A., Faul, P.N., Fletcher, J.A., Davis, I.J., Dal Cin, P. and Perez-Atayde, A.R. (2001) Upper respiratory tract carcinoma with chromosomal translocation 15;19: evidence for a distinct disease entity of young patients with a rapidly fatal course. *Cancer*, **92**, 1195-1203.
181. Dupuy, A.J., Rogers, L.M., Kim, J., Nannapaneni, K., Starr, T.K., Liu, P., Largaespada, D.A., Scheetz, T.E., Jenkins, N.A. and Copeland, N.G. (2009) A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. *Cancer research*, **69**, 8150-8156.
182. Hameyer, D., Loonstra, A., Eshkind, L., Schmitt, S., Antunes, C., Groen, A., Bindels, E., Jonkers, J., Krimpenfort, P., Meuwissen, R. *et al.* (2007) Toxicity of ligand-dependent Cre recombinases and generation of a conditional Cre deleter mouse allowing mosaic recombination in peripheral tissues. *Physiological genomics*, **31**, 32-41.
183. Schmidt-Supprian, M. and Rajewsky, K. (2007) Vagaries of conditional gene targeting. *Nature immunology*, **8**, 665-668.
184. Houzelstein, D., Bullock, S.L., Lynch, D.E., Grigorieva, E.F., Wilson, V.A. and Beddington, R.S. (2002) Growth and early postimplantation defects in mice deficient for the bromodomain-containing protein Brd4. *Molecular and cellular biology*, **22**, 3794-3802.

185. Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K. and Miyawaki, A. (2002) A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature biotechnology*, **20**, 87-90.
186. Guo, G., Wang, W. and Bradley, A. (2004) Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells. *Nature*, **429**, 891-895.
187. Keng, V.W., Villanueva, A., Chiang, D.Y., Dupuy, A.J., Ryan, B.J., Matise, I., Silverstein, K.A., Sarver, A., Starr, T.K., Akagi, K. *et al.* (2009) A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nature biotechnology*, **27**, 264-274.
188. Starr, T.K., Allaei, R., Silverstein, K.A., Staggs, R.A., Sarver, A.L., Bergemann, T.L., Gupta, M., O'Sullivan, M.G., Matise, I., Dupuy, A.J. *et al.* (2009) A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science (New York, N.Y.*, **323**, 1747-1750.
189. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, **18**, 1851-1858.
190. Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, **24**, 713-714.
191. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
192. Palmieri, N. and Schlotterer, C. (2009) Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PloS one*, **4**, e6323.
193. Uren, A.G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A.H., Wilson, C.H., Rance, R., Jonkers, J., van Lohuizen, M., Berns, A. *et al.* (2009) A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nature protocols*, **4**, 789-798.
194. de Ridder, J., Uren, A., Kool, J., Reinders, M. and Wessels, L. (2006) Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS computational biology*, **2**, e166.

Appendix A. Primers and linker sequences used for Splinkerette PCR

			*Sleeping Beauty**	*piggyBac*
Linker Sequences	FWD		CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGTGTCGACACTAGTGG	
	REV		GATCCCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCAAAAAAA	
Primers on the linker	1st PCR		CGAAGAGTAACCGTTGCTAGGAGAGACC	
	2nd PCR		GTGGCTGAATGAGACTGGTGTCGAC	
	Sequencing**		ATGAGACTGGTGTCGACACTAGTG	
Primers on the transposon	1st PCR	5' TR	GTGTCATGCACAAAGTAGATG	TAAATAAACCTCGATATACAGACCGATAAA
		3' TR		CAAAATCAGTGACACTTACCGCATTGACAA
	2nd PCR	5' TR	GATGTCCTAACTGACTTGCC	ATATACAGACCGATAAAACACATGCGTCAA
		3' TR		CTTACCGCATTGACAAGCACGCCTCACGGG
	Sequencing	5' TR	GATGTCCTAACTGACTTGCC	TTTTACGCATGATTATCTTTAACGTACGTC
		3' TR		TTAGAAAGAGAGAGCAATATTTCAAGAATG

* For *Sleeping Beauty* uses the same primers for 5' and 3' PCR since the sequences for 5' and 3' terminal repeats are identical for *Sleeping Beauty*.

** Sequencing primers are primers used for sequencing the splinkerette PCR products. They could be either identical or downstream of the 2nd PCR primers.

Appendix B. ABBREVIATIONS

4-OHT – 4-Hydroxytamoxifen

AML1 – Acute Myeloid Leukemia 1

BAC – bacterial artificial clone

BLAT – BLAST-Like Alignment Tool

BSD – Blasticidin

cALL – childhood Acute Lymphoblastic Leukemia

CGH – comparative genomic hybridisation

cHS4 – chicken hypersensitive site 4 (an insulator sequence of the chicken β-like globin gene cluster)

CIS – Common Insertion Sites

EMS – ethyl methane sulphate

ENU – N-ethyl-N-nitrosourea

ERt2 – estrogen receptor ligand-binding domain

EtOH – ethanol

EUCOMM – European Conditional Mouse Mutagenesis Program

FACS – fluorescence-activated cell sorting

GFP – green fluorescent protein

HGP – Human Genome Project

HIV – human immunodeficiency virus

HTLV – human T-cell lymphotropic virus

iMapper – Insertional Mutagenesis Mapping and Analysis Tool

IRES – internal ribosomal entry site

KOMP – Knockout Mouse Project

LM-PCR – linker mediated PCR

LSA – local sequence alignment

LTR – long terminal repeat

MITEs – Miniature Inverted-repeat Transposable Elements

MMTV – mouse mammary tumour viruses

MSCV – murine stem cell virus

MuLV – murine leukaemia virus

ORF – open reading frame

PAX5 – Paired-Box-Containing Gene 5

PB – *piggyBac*

PBS – phosphate buffered saline

PCR – polymerase chain reaction

PI – propidium iodide

polyA – polyadenylation signal

RB – retinoblastoma

RISC – RNA Induced Silencing Complex

RIP1 – receptor-interacting protein kinase 1

RNAi – RNA interference

RT-PCR – Reverse Transcriptase Polymerase Chain Reaction

SA – splicing acceptor

SB – *Sleeping Beauty*

SD – splicing donor

shRNA – small hairpin RNA

siRNA – small interfering RNA

SNP – single nucleotide polymorphism

SOAP – Short OligoNucleotide Alignment Prgram

SSAHA – Sequence Search and Alignment by Hashing Algorithm

TR – terminal repeat

UCSC – University of California Santa Cruz genome browser

Appendix C. Publication 1 - Chromosomal mobilization and reintegration of *Sleeping Beauty* and *piggyBac* transposons

Appendix D. Publication 2 - *iMapper*: a web application for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes

Appendix E. Publication 3 - Slingshot: a *piggyBac* based transposon system for tamoxifen-inducible 'self-inactivating' insertional mutagenesis