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.