Chapter one – Introduction

Chapter overview

The rapid development of high-throughput genome-sequencing technologies in the past decade has brought a new era to biological research. Vast amounts of genomic information have been generated for diverse species, including human beings. The speed of acquisition of this type of data will only increase due to further innovations in future-generation sequencing technologies. Although this information has brought new ways to think about and approach many areas of biology, it has proved a considerable challenge for the research community to interpret the information encoded in these genomes, i.e. to understand the functions of all genes in a given genome and their coordinated activities that give rise to a complex organism.

Genetics has paved the way for generations of scientists to devise simple and effective means to manipulate genomes of model organisms, in order to understand the functions of genes through the isolation and characterisation of mutants. A small number of eukaryotic organisms, primarily yeast, the worm, the fruit fly and the mouse, have been extensively used as model organisms for genetic investigations. Many elaborate approaches have been developed over the years and these have been invaluable in contributing to our understanding of many fundamental biological processes. Such classical approaches, using model organisms, have been and will still be offering gateways to tackle the tremendous challenges in the post-genomic era.

My PhD research is exploring a forward genetic approach to discover components of the miRNA biogenesis and downstream effector pathway in cultured mammalian cells. This introductory chapter encompasses five main areas. The first part describes the concepts and principles of reverse and forward genetic approaches in experimental organisms to ascribe gene function. The second part concerns the means of mutagenesis with the particular focus on the insertional mutagenesis in mammalian systems. The third part describes the use of mammalian cells as models for forward genetic screens, focusing on the recessive genetic

screens. The fourth part of the introduction focuses on the microRNA and their biogenesis and effector pathways and followed by the design concept of this thesis project.

Reverse and forward genetics in experimental model organisms

Genetics and genomic approaches at molecular level have been revolutionised our understanding of biological processes. “Reverse genetics” describes the “gene to phenotype” approach, with which functions of a gene of interest can be investigated by disrupting the physiological expression of this gene. “Forward genetics” is a “phenotype to gene” approach without the requirement of prior knowledge. Efficient genome-wide gene disruption allows the isolation of genes which function in the phenotype of interest. Both approaches complement each other in dissecting and unveiling gene functions in biological pathways.

1. Reverse genetics

Reverse genetics is an approach with which the expression of a gene can be disturbed either by mutating the DNA sequence of the gene or knocking down the gene expression using RNA interference. The phenotypic consequences of this particular genetic perturbation can be analysed. There are three reverse approaches, namely the homologous-recombination-based gene targeting, RNA interference and the Zinc finger nucleases-based genetic alterations.

1.1. Homologous recombination-based gene targeting

The strategy of introducing defined mutations in a whole organism was piloted in the mouse twenty seven years ago, and has become the standard method to manipulate and study the mouse genome. The success of this technology in the mouse owes to two significant achievements. The first major breakthrough was the demonstration of germline transmission of cultured mouse embryonic stem cells. Martin Evans and Matthew Kaufman at the University of Cambridge established the pluripotent ES cell lines from the E3.5-E4.5 mouse blastocyst (Evans and Kaufman, 1981). Bradley *et al* subsequently showed that after prolonged culturing of these ES cells, they still maintain the ability to contribute to all cell types of an animal, including the germ cells (Bradley et al., 1984). Moreover, mutations generated in these ES cells do not affect their germline transmission property, and this work

opened up the possibilities of generating mutations in endogenous genes thereby determine their functions in the mouse (Robertson *et al.*, 1986; Kuehn *et al.*, 1987). Secondly, targeted mutagenesis via homologous recombination of an artificial targeting vector and the genomic DNA was feasible in mammalian cells, and this was first demonstrated by Smithies *et al* using the *β*-globin locus (Smithies et al., 1985). The marriage of these two advances allowed targeted manipulation of endogenous genes via homologous recombination to be carried out in ES cells to be transmitted to the whole mouse (Schwartzberg *et al.*, 1989; Zijlstra *et al.*, 1989; Snouwaert *et al.*, 1992). This began to allow the gene-function dissection and human-disease modelling in mice (DeChiara et al., 1990; McMahon and Bradley, 1990; Snouwaert et al., 1992).

This Nobel-prize winning work has since become a "gold standard" for studying gene function and provides the foundation for many subsequent developments of other mouse genetics technologies. Since the early 90's until present, the number of studies based on gene targeting has exploded. The completion and annotation of the mouse genome sequencing, the availability of the indexed bacterial artificial chromosomes (BACs), and the development of methods to use homologous recombination in *Escherichia coli* (*E. coli*) to generate targeting vectors with nucleotide precision, allow this technology to be carried out for the whole genome (Lee et al., 2001). Currently, international consortia are using gene targeting to generate ES cells and eventually mouse lines with a targeted mutation in every gene (http://www.knockoutmouse.org/about/komp). The mouse has become the only multi-cellular model organism to possess such an immense wealth of resources for reverse genetics.

1.2. Reverse genetics using RNAi

In other experimental organisms where gene targeting is not feasible, RNA interference has been widely used to investigate the function of genes of interest in a loss-of-function manner. RNAi is an evolutionarily conserved mechanism in eukaryotic cells to silence gene expression. It was initially observed in plants and then in animals, it was first described in *Caenorhabditis elegans* (*C. elegans*) by Fire and co-workers (Fire et al., 1998). Long double-stranded RNAs (dsRNA) introduced into the worm led to targeted degradation of a mRNA. Although

successful with several experimental invertebrate and vertebrate organisms, the application of dsRNA-induced RNAi was not feasible in mammals. Long dsRNAs induce global gene suppression by dsRNA-induced activation of the interferon response in mammalian cells, which leads to an overall blockage of translation and apoptosis (Stark et al., 1998). However, small interfering RNAs (siRNAs), approximately 21 base-pair double stranded RNAs, can elicit RNAi in mammalian cells without inducing the interferon response (Elbashir et al., 2001). This discovery has sparked intense development of this technique and tools have been developed to study individual genes and conduct large scale screens. With the availability of full genome sequence of many species, RNAi libraries have been constructed to target all genes in a given genome. High throughput synthesis of oligonucleotides and their cloning into vectors has been established to produce shRNAs on a large scale. Large shRNA collections (http://www.openbiosystems.com/rnai/) with different vector designs are available for both genome-wide and gene family investigations. Incorporation of barcode tags into the shRNA design also allows negative selection screens to be conducted, where knockdown of a gene causes cell death or reduced proliferation.

Using mammalian cell culture systems, numerous large-scale RNAi screens have also been conducted to study a wide range of biological pathways. The first screen reported was conducted in mammalian cells to identify genes involved in p53-mediated cell cycle arrest (Berns et al., 2004). Recent examples include investigations into many areas of research, such as human host factors crucial for influenza virus replication (Karlas et al., 2010); chromatin factors that regulate ES cell identity (Fazzio et al., 2008; Gaspar-Maia et al., 2009); and modifier screen for the circadian clock in human cells (Zhang et al., 2009).

There are two major concerns using RNAi to study gene functions. Firstly, most of the cases, RNAi-mediated silencing is incomplete, thus it is known as “knockdown” and gives rise to hypomorphic phenotypes. Therefore the phenotypic interpretation may be complicated. The second major limitation is its off-target effects. Sequence-specific off-target silencing of mRNA sequences that have partial complementarity to the siRNA can occur. As few as 11 contiguous nucleotides, which are complementary to a target sequence, have been observed

to evoke off-target silencing and both the sense and antisense strands of the siRNA can induce off-target effects (Jackson et al., 2003). Therefore, siRNA sequences must be chosen carefully by screening for homologous sequences in the genome of interest to ensure gene silencing efficiency while avoiding non-specific off-target effects. Multiple siRNAs with different sequences should be used for single candidate gene to distinguish the off-target effect. In addition to the sequence-based off-target effect, siRNA can also induce non-specific effects on gene expression profiles. Several factors can contribute to this, such as the concentration of the delivered RNAi and the delivery method.

1.3. Zinc finger nuclease-mediated genetic alternations

A final and newly developed approach is the zinc finger nuclease (ZFN)-mediated genetic alternations. ZFN consists of a synthetic zinc finger DNA-binding domain, composed of three or four fingers, fused to the nuclease domain of the *Fok*I restriction endonuclease. The ZFN functions as a homo- or hetero- dimer to recognise a particular stretch of DNA sequence and to induce double strand breaks (DSBs), thereby promoting site-specific homologous recombination (Jasin, 1996; Carroll, 2004). A repair DNA template can be supplied to direct repair of the DSBs to incorporate specific genetic alternations into the defined genomic loci (Jasin, 1996). In addition, ZFNs can be used to direct mutagenesis to specific loci without the template based on the error-prone non-homologous end joining (NHEJ) repair system to generate loss-of-function mutations. Because the specificity of the DNA binding can be achieved by engineering the finger arrays to recognise specific DNA sequence, any sequence combination is theoretically recognisable by the synthetic ZFNs. Much work has shown that ZFNs can be used to direct locus-specific genetic alternations in many organisms, including human, plant and *Drosophila* cells (Bibikova *et al.*, 2002; Alwin *et al.*, 2005; Wright *et al.*, 2005). Therefore, it offers an alternative genetic manipulation approach to conventional gene targeting and may be useful in cell types that homologous-recombination-based gene targeting is not efficient enough to obtain desired mutations. International ZFN consortia is underway to use combinatorial-based selection method for making zinc finger arrays and screening for combinations of fingers which can provide high level of activity and sequence specificity (Maeder et al., 2008).

Although this technology is potentially very powerful, one of the major concerns is the sequence specificity for cleavage. Because a functional ZFN dimer only replies on an 18 bp sequence to define the recognition specificity, complex genomes such as mouse and human may contain many sequence matches to the ZFN recognition sequence in different genomic loci, which are not the intended locations for targeted genetic alternations. Cleavage of these sites by ZFNs will likely to introduce point mutations, small insertions and deletions upon DNA repair if the endogenous error-prone non-homologous end joining (NHEJ) system is used by the cells to repair the DSBs. Thus, the generation of these "unintended" mutations in the genome can complicate phenotype interpretations.

2. Forward genetics and different screen designs

Forward genetics is a discovery process that identifies gene function in a non-hypothesis driven fashion, therefore, it can be a powerful approach for investigating a biological process without any prior knowledge on the molecular nature. This classical genetic approach has a long history and has led to many landmark discoveries in model organisms before genome sequences were available. Such an approach replies on randomly mutagenising the genome of an organism, then to isolate mutants with phenotypic changes. The presence of a particular mutant phenotype provides geneticists with an entry point to a biological process. Subsequent identification of the gene being mutated can establish the functional connections of these genes to the biological process under investigation. The mutant itself is a valuable resource for subsequent gene-function dissection. There are several means of mutagenesis including chemical, physical or biological agents, with each having their own characteristics with respect to the nature of mutations, efficiency of mutagenesis and genome coverage. The details of the mutagenesis are covered in Section 3 of this chapter. Using the unicellular yeast as the model organism, a large proportion of genes in the cell cycle were discovered by performing forward genetic screens, isolating mutants that show a cell-cycle arrest or modified cell-cycle behaviours (Hartwell et al., 1974; Nurse, 1975). *C. elegans* and *Drosophila* have also been the test beds in exploring various technologies and elaborate screen designs to uncover gene functions in multi-cellular model organisms.

The designs for forward genetic screens can be classified broadly in two ways. The first way to categorise the screen designs is based on the clonality of the mutation and the designs can be divided into germ-line and somatic mutagenesis. In germ-line mutagenesis, mutations are generated in the gametes of mature adults. After mating, the mutation originated from a gamete will be present in every cell of an offspring. In contrast to germline mutagenesis, somatic mutations are generated in tissues of an organism and an individual can also harbour different types of mutations. Therefore, the organism is genetically mosaic and the mutations can only be passed onto offspring if they occur in germ cells. The second way to classify the screen designs is based on the types of the mutations generated, and the screens can be broadly classified into recessive (loss of function) and dominant (gain of function) screens. More complex screen designs such as modifier screens and synthetic lethal screens can be built on the basic loss- or gain-of-function screens.

2.1. Germline and somatic mutagenesis

Early generation of geneticists relied on the isolation of visible mutants generated from the natural population spontaneously. Forward genetic screens are not possible based on spontaneous mutations as the frequency of such events is very low. The forward genetics only became feasible when efficient means of mutagenesis was available. A classic genetic screen involves the generation of mutations in germ lines of an organism using chemical mutagens and, by propagating progenies, mutations can be transmitted and segregated in the subsequent generations. A phenotypic screen can then be conducted on these organisms. This method is widely applicable to experimental organisms such as *C. elegans* and *Drosopholia*, as they have fast generation times and produce many progenies. Several of the early developments using this approach have led to Nobel prizes, and most of these focused on the discovery of loss-of-function mutants. Using *C. elegans* as the model, Sydney Brenner showed that random mutagenesis using the chemical mutagen Ethyl methane sulphonate (EMS) gave rise to many visible phenotypes efficiently (Brenner, 1974). John Sulston and Robert Horvitz used this method and discovered mutants with defects in vulva differentiation (Sulston and Horvitz, 1981). Using *Drosophila* as a model organism and EMS mutagenesis, Christiane Nüsslein-Volhard and Eric Wieschaus isolated mutants that affect the patterning of

the embryo using (Nüsslein-Volhard and Wieschaus, 1980). These approaches and discoveries from these forward genetic screens have transformed subsequent research in model organisms, and many of the genes and pathways discovered in these early screens are still of interest to the research community today. However, one class of screen can not pull out all the genes involved in a biological process and more elaborate screen designs have been developed to expand the genes that can be functionally assigned.

In a multi-cellular organism, a single gene can play multiple roles in different biological pathways in different cell types and tissues. Using germ-line mutagenesis, many genes can not be recovered due to their crucial roles in early development, which may be irrelevant to the roles they have in biological process of interest later on. Therefore, germ-line mutagenesis can only provide information on the first essential role of a given gene. Another type of strategy, namely somatic mutagenesis, can overcome this limitation by introducing mutations conditionally in appropriate cell types or times to bypass lethality.

2.2. Dominant, recessive and genetic interaction screens

Dominant and recessive genetic screens are distinguished by the nature of the mutation generated. Dominant (or gain-of-function) screens are designed to identify mutant phenotype when the genes are abnormally activated either through over-expression or ectopic expression. Recessive (or loss-of-function) screens are designed to isolate genes showing a phenotype of interest when inactivated. These two types of screen designs can complement each other in expanding the repertoire of genes which can be functionally assigned.

2.2.1. Dominant genetic screens

Dominant screens use exogenous factors to achieve phenotype conversion by the activity of single genes or combinations of genes from a genome-wide library or from a knowledge-based pre-selected genomic or cDNA library. In addition, dominant screens can be very useful in studying genes, as the loss-of-function mutation of these genes may be lethal or do not provide a phenotypic change due to the presence of other genes which are functionally redundant. Such kind of screens can be performed through germ-line mutagenesis or

somatically in a multi-cellular organism. For example, in *Drosophila*, several genes important for the eye and wing development were isolated from the tissue-specific misexpression (Rorth et al., 1998). In mice, large-scale forward genetic screens using ENU has also been conducted to isolate dominant mutants in mice, and the first gene identified to be involved in the circadian rhythm in mammals, *Clock*, was identified through ENU-mediated forward genetic screen coupled with mutant identification by positional cloning (Vitaterna et al., 1994). Two large centres in Europe, Helmholtz Zentrum in Germany and the UK Medical Research Council centre in Harwel, are dedicated in producing large numbers of dominant germline mutations using ENU mutagenesis approach (Hrabe de Angelis et al., 2000; Nolan et al., 2000).

2.2.2. Recessive genetic screens

Recessive genetic screens are designed to isolate genes which show a phenotype when inactivated and this often requires the inactivation of all copies of the gene in a given genome to evoke a phenotypic change. In yeast, recessive genetic screen is more feasible to conduct than other organisms with the diploid genome, as they can exist as haploid. In organisms with stable diploid genome including mammalian systems, recessive genetic screens are challenged by the inactivation of both alleles of the genes. Although loss-of-function screens can be conducted using genome-wide RNAi libraries in *C. elegans*, *Drosophila* and mammalian cells and large number of genes are typically identified in a single screen, the major problem in such types of screens are the off-target effects and the subsequent validation of all the "hits", which has been discussed earlier in this chapter.

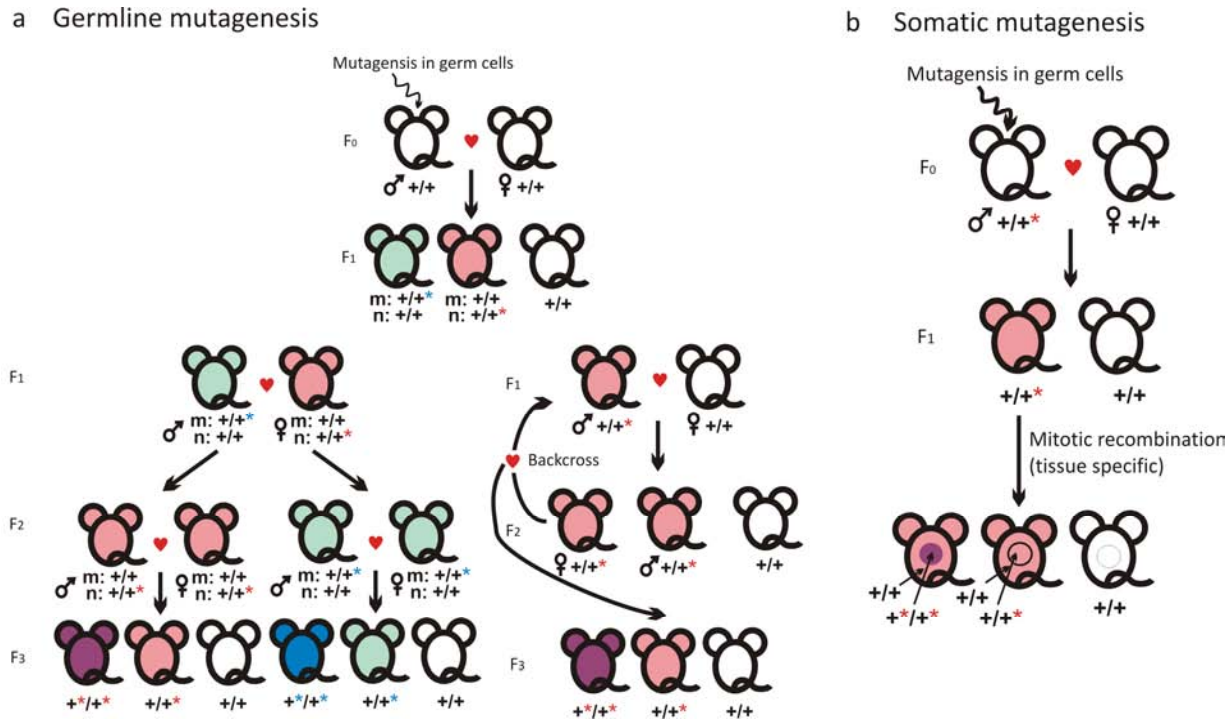
2.2.2.1. Germline recessive mutagenesis

Early screens used chemical agents such as *N*-ethyl-*N*-nitrosourea (ENU) and Ethyl methane sulphonate (EMS) to efficiently mutagenise the genome of an organism, mainly generating loss-of-function mutations. When mutations occur in the germline of the organism, it can be passed on to the offspring. By crossing the mutagenised animal with a wild-type animal, the F_1 offspring will be heterozygous mutants of different loci depending on the mutation spectrum present in each germ cell. Further mating can be conduced in two ways. The first method is to inter-cross F_1 animals in order to introduce more mutations into the pedigree,

Figure 1-1a. Further inter-crossing F_2 animals can produce recessive mutants in F_3 generation. The second method is to cross the F_1 heterozygous males with wild-type females, producing F_2 heterozygous females which can then used to mate with the heterozygous F_1 males to produce F_3 homozygous mutants, Figure 1-1a. In this method, most of the mutations present in the F1 males can be converted to homozygosity in F_3 generation.

Such mutagenesis and mating strategy can be efficiently conducted in small model organisms with short generation time and relatively small genome compared to mammalian systems, such as *C. elegans* and *Drosophila*. Many early discoveries in these organisms have been conducted in this way (Brenner, 1974; Nüsslein-Volhard and Wieschaus, 1980; Sulston and Horvitz, 1981; Grunwald and Streisinger, 1992). In mice, ENU can efficiently mutagenise the genome, at the rate of one mutation per every 1-2 Mb and one loss-of-function mutation at a given locus in one sperm per 1,000 (Kile and Hilton, 2005). Using such a method, a recessive genetic screen has been performed in mice in isolating homozygous mutant in the phenylalanine hydroxylase (*Pah*) locus and the loss-of-function of which models the human phenylketonuria (McDonald et al., 1990). Although screens in both dominant and recessive manner have been conducted in mice, such an approach is very time consuming and labour intensive due to the complexity of the mammalian genome, long generation time and the high cost for husbandry. In addition, the mutant identification procedure is difficult and this involves identifying the physical location of the mutation by linkage mapping, and subsequent sequencing of the region where the mutation is residing. Thus, the identification of mutations and demonstration of their causality from such large scale mutagenesis can take many years.

Figure 1-1 : Germline and somatic mutagenesis.



Both mutagenesis strategies are illustrated with mice, but they are universal to all diploid multi-cellular models. The colour scheme reflects the genetic status with white represents wild-type, light pink and dark pink represent heterozygous and homozygous. The red star represents a mutation. For both strategies, only one germline mutation is illustrated here and one can imagine that each sperm from the mutagenised F_0 male can carry different mutations; therefore all heterozygous F_1 animals will have different mutations. a, germline mutagenesis to obtain mutants having identical genotype throughout the bodies. It involves mutagenising the sperms of F_0 males and heterozygous whole-body mutants can be derived in F_1 generations. There are two ways to obtain homozygous mutants. The first way (bottom left panel) is to intercross F_1 animals and in this way, more mutations can be introduced into the pedigree. The second way (bottom right panel) is to cross F_1 males with wild-type females to produce heterozygous F_2 females and these females are backcrossed to the F_1 heterozygous males to produce homozygous animals in F_3 generation. In this crossing, most of the mutations in F_1 males can be converted to homozygosity in F_3. b, somatic mutagenesis with somatic mosaic mutants. After obtaining F_1 heterozygous mutants, mitotic recombination can be induced somatically (or/and spatially), clones of homozygous somatic cells can arise in otherwise heterozygous background.

2.2.2.2. Mitotic recombination-mediated recessive mutagenesis in the soma

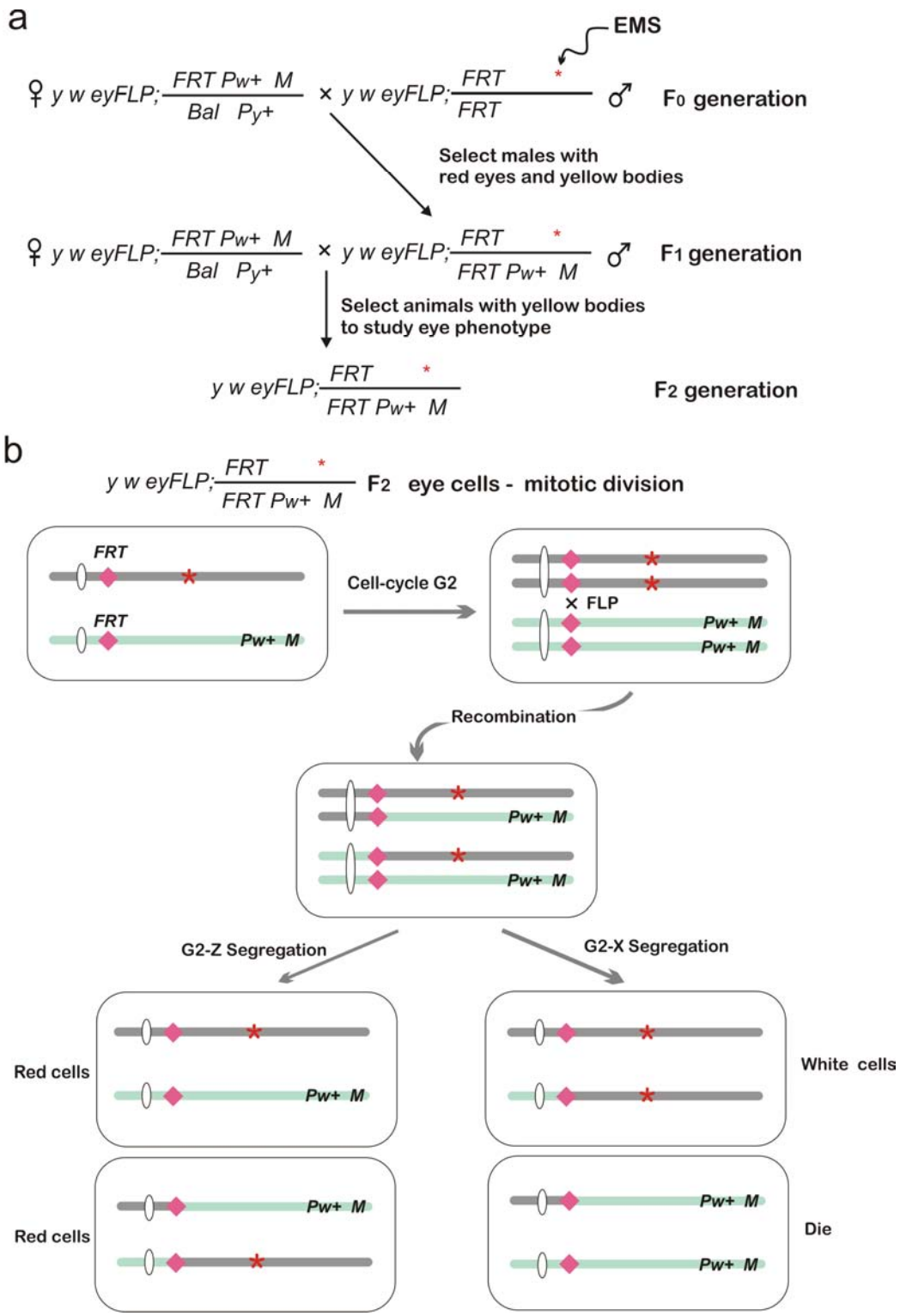
Another *in vivo* approach to conduct recessive genetic screens in diploid organisms is through the generation of somatic mosaics with homozygous daughter cells produced from a heterozygous genotype by mitotic recombination, Figure 1-1b. Mitotic recombination is a

natural occurring phenomenon which was first discovered in *Drosophila* and subsequently has been detected in a variety of species including yeast, mouse and human. Site-specific mitotic recombination can be induced by recombination systems such as FLP/*FRT* and Cre/*loxP*. The system was first demonstrated in *Drosophila*, to convert chromosomal regions distal to the *FRT* sites to homozygosity by FLP/*FRT* induced mitotic-recombination (Golic, 1991; Xu and Rubin, 1993). The *FRT* sites can be engineered in centromeric regions of all chromosomes independently to maximise the number of loci which can be converted to homozygosity. Mitotic recombination can be induced by Flp expression and in the G2 phase of the cell cycle, some daughter cells will be homozygous for loci distal to the *FRT* sites after segregation, Figure 1-2b. A selection marker or visible marker can be incorporated into this system to aid in the identification and enrichment of the homozygous mutant clones. In *Drosophila*, this system can be easily adapted to a genome-wide level, due to the high efficiency of Flp/*FRT* mediated mitotic recombination and a small number (four) of chromosomes. An elegant example to illustrate the genome-wide approach using this system to study recessive genes is a screen conducted in *Drosophila* for the identification of genes in the photoreceptor axon guidance (Newsome et al., 2000), and Figure 1-2 shows the screening strategy and the use of selection markers.

Figure 1-2 legend (figure on next page): Somatic mosaic recessive screens in *Drosophila* eye.

a, a mating and screening strategy. This part is adapted from Newsome et al, 2000. EMS, ethylmethanesulphonate. Drosophila genes and markers: *y, X*-linked yellow gene, a dominant body pigmentation marker with *y+* animals being brown and *y-* being yellow ; *w*, white gene, a dominant eye pigmentation marker, *w+* animals have red eyes and *w-* having white eyes; *M*, minute gene, a gene associated with developmental retardation, a recessive marker and cells without *M* having retarded growth and reduced viability. These markers help to select animals with correct genotypes, distinguish the homozygous cells from the heterozygous background and to eliminate undesired homozygous cells without mutations. Pw^+ and Py^+ represent that the markers were integrated *P* element mediated transpositions. *Bal*, represents the Balancer chromosome to suppress spontaneous mitotic recombination. *eyFLP*, *FLP* is driven by a eye specific promoter for spatial-specific expression. b, schematic representations of FLP/*FRT* induced mitotic recombination to obtain eye cells with homozygous mutations.

Figure 1-2: Somatic mosaic recessive screens in *Drosophila* eye.

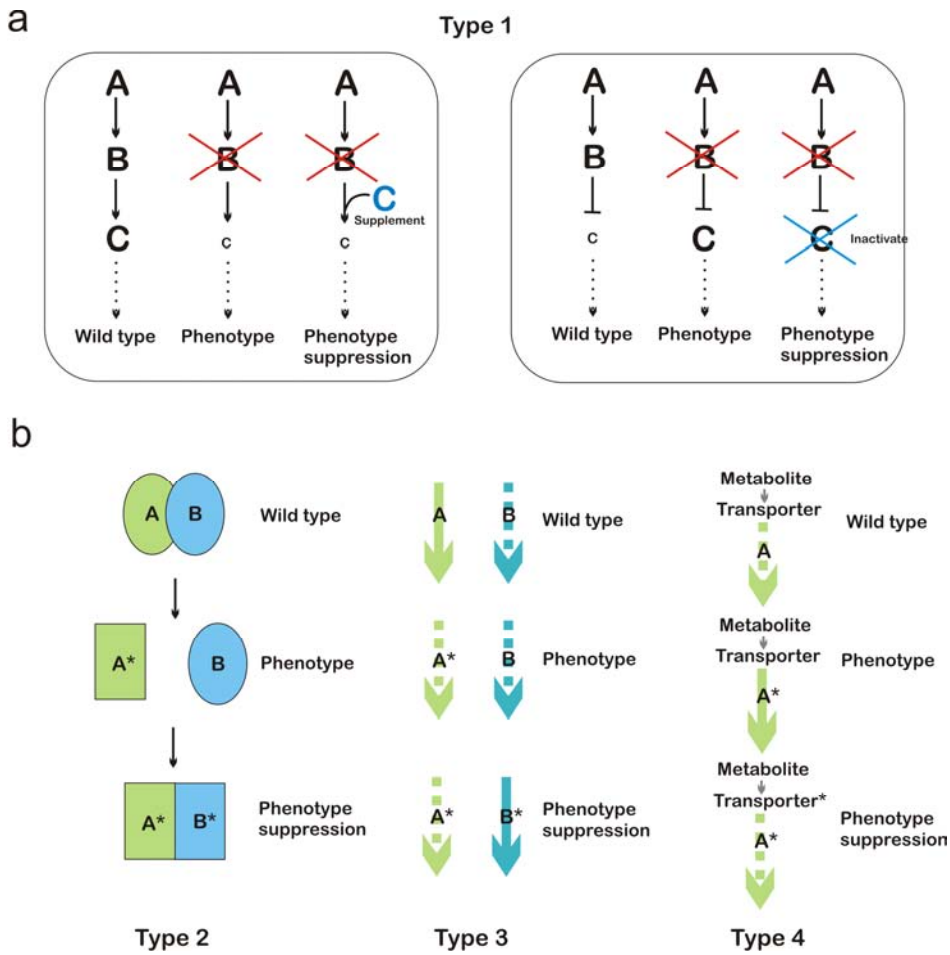


This system can also speed up the recessive genetic screens compared to conventional germ-line recessive mutagenesis screens, as the screen can be conducted in the somatic tissues of an F1 generation rather than in the F3 generation in a conventional scheme. The Flp can be expressed conditionally and therefore the mitotic recombination events can be specified in a spatiotemporally-controlled manner. In this way, multiple functions of a single gene can be dissected in different cellular contexts and developmental stages. The use of Cre/*loxP* site-specific recombination system can also generate induced mitotic recombination in mouse ES cells and in somatic cells in mice, albeit with much lower efficiency than in *Drosophila*, the details of which are described in Section 5.1.1. of this chapter.

2.2.3. Genetic-interaction screens

Using simple dominant suppressor or enhancer screens are powerful means to gain further information on the genetic interactions in a biological pathway. There are two broad approaches to delineate a genetic interaction network. The first approach is the use of synthetic-suppression screen, in which over-expression or inactivation of a gene can rescue an observed phenotype caused by another gene, thus identifying genes that act in the same biological processes, Figure 1-3a. There are four types of interactions that may give rise to the genetic suppression interaction, Figure 1-3b. The first type is that the two genes functions in the same pathway and this is the most useful interaction to delineate a biological pathway, Figure 1-3a. The second type is the direct physical interaction between the gene products, and not all the mutations in the original gene can be rescued by the mutation of the second gene. The direct interaction between Cdc2 kinase and Cdc13 cyclin were predicted using this approach in yeast (Booher and Beach, 1987). The second type of interaction is alternative pathway activation which can function in a similar manner to the first pathway that is blocked due to mutations of a gene in this pathway. The final type of interaction is non-specific rescue, and the mutant identified is not related to the biological pathway of interest.

Figure 1-3: Synthetic suppression screens.

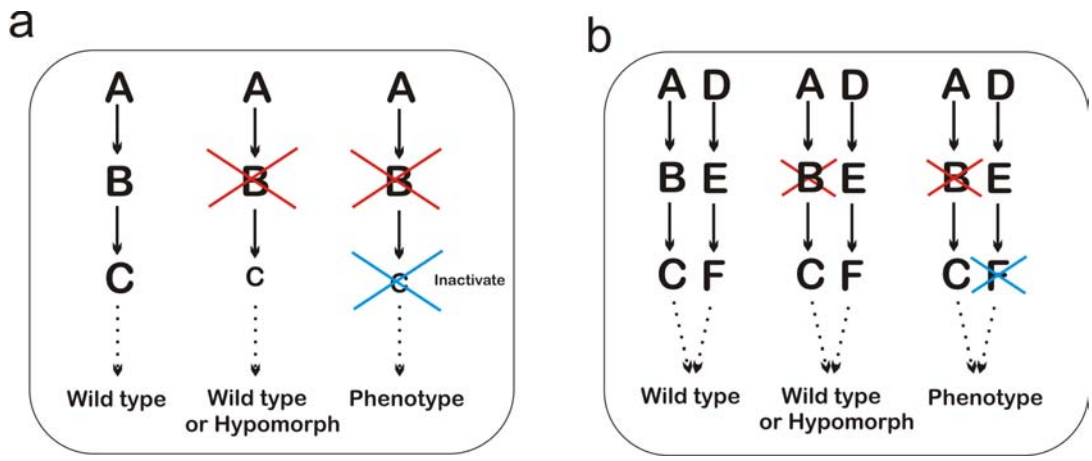


a, In synthetic suppression screen, either over-expression or inactivation of a second gene can rescue the observed phenotype caused by the first mutation. In this screen, a biological pathway can be delineated based on the genetic interactions. b, three other types of interactions which can be isolated from the synthetic suppression screen. Type 2 identifies a physical interaction partner of the first gene, but not all mutations in the interaction partner can rescue the phenotype. Type 3 identifies the alternative pathways, with the mutation in a gene in the alternative pathway rescues the inactivation of the first pathway. The final type is non-specific interaction, as the mutation in the second gene is not playing a role in the pathway of interest and an example using a transporter to demonstrate this. The pathway is inactivated in the presence of a metabolite. Mutations in a gene A in the pathway activates the pathway, however, mutation to inactivate a non-specific transporter can lead to pathway inactivation, thus rescuing the phenotype. However, the transporter of the metabolite is not involved in the pathway.

The second approach is synthetic enhancement (synthetic-enhancer screen), in which mutating the second gene can further attenuate an observed phenotype caused by the mutation in the first gene, in some cases, to the point of lethality, i.e. synthetic-lethal screen. This approach can be very useful for pathway delineation, Figure 1-4a. For example, several synthetic enhancer screens were conducted in *Drosophila* to identify the components downstream of Sevenless (Sev), which controls the cell-fate choice in the eye formation (Simon, 1994). The screen uses a temperature-sensitive Sev, a hypomorphic allele, as the sensitized genetic background to hunt for components involved in the photoreceptor R7 cell-fate determination. In this background, a heterozygous mutant of a gene in this pathway, with 50 % loss in expression is sufficient to produce a failure in the Sev-mediated signalling pathway. The mutants from the screen demonstrated that Sev and other receptor tyrosine kinases are upstream of Son of Sevenless (Sos) to activate the Ras signalling pathway (Simon *et al.*, 1991; Simon *et al.*, 1993; Simon, 1994).

In addition, synthetic-enhancer screens can be also very useful for identifying redundant genetic pathways in a biological process, as the inactivating of one pathway does not show a phenotype due to the presence of the redundant pathway to support the wild-type phenotype. However, double mutations in both pathways will be show a phenotype, Figure 1-4b. One could use a null mutation in one pathway as the genetic background to screen for components in the complementing pathway which gives a phenotype in this background. Such a type of screen performed in *C. elegans* led to the identification of two redundant classes of the Synthetic Multivulval genes (Ferguson and Horvitz, 1989). Neither first nor second class of mutants displays phenotypes on their own or in combination with mutants within the same class and the synthetic multivulval phenotype is revealed only when mutants are present in genes from both classes (Ferguson and Horvitz, 1989).

Figure 1-4: Synthetic enhancement screens.



a: pathway delineation; b, redundant-pathway identification.

3. Means of mutagenesis for forward genetics

In a forward genetic screen, the function of a gene can be assigned to a specific biological process by analysing the phenotypic consequences when the gene activity is altered by a mutagen. Three main categories of agents can be used to achieve genome-wide mutagenesis, namely chemical, physical and biological mutagens. Each has its own characteristics with respect to the nature of mutations, the efficiency of mutagenesis and genome coverage.

3.1. Chemical agents

Chemical agents such as *N*-ethyl-*N*-nitrosourea (ENU) and ethylmethanesulphonate (EMS) have been most widely used as efficient mutagens and many of the classical forward genetic screens have been conducted using chemical mutagens in most of the model organisms. The details are described in the previous section of this chapter. These DNA alkylation mutagens generate a range of alterations, including point mutations, several-nucleotide insertions and deletions (Chen et al., 2000; Munroe et al., 2000). Mutants caused by these chemicals result mainly in loss-of-function mutations, which can be complete or partial loss of function, gain-of-function mutations can also be recovered. In male mouse ES cell cultures, the mutagenesis efficiency of ENU and EMS were estimated tested on the *X*-linked *Hprt* locus (Chen et al., 2000; Munroe et al., 2000) and the mutation rate per locus was measured to be one in 200

cells and one in 1,200 cells for ENU and EMS respectively. Therefore, ENU-mediated mutagenesis possesses a high mutation rate. Coupled with the unbiased genome-wide distribution of mutations, the complete genome coverage (saturation) of mutagenesis can be achieved. However, the main drawback is the difficulty in identifying the causal mutation due to the mutation load per cell can obscure causality and the difficulty in tracing the mutations. According to the mutation rate of one gene mutated in every 200 cells measured in mouse ES cells, the number of genes mutated per cell will be around 150 assuming 30,000 genes are present in the mouse genome. Such a large number of mutations per cell can make the identification of the causal mutation very difficult.

There are two ways to narrow down to the genomic region with the causal mutation. The first method is genetic complementation. In cell culture systems or unicellular organism such as yeast, wild-type genomic DNA is transferred to the mutant cells to suppress the observed phenotype. The identity of the gene mutated can be identified by isolating the genes in the complementation groups. This can be achieved by cloning of the genomic fragment using cosmid or bacteriophage vectors. Several genes that function in the nucleotide excision repair pathway and DNA single and double strand break repair pathway were identified this way (Thompson et al., 1990; Troelstra et al., 1990). With the completion of the whole genome sequencing of many experimental model organisms, complementation assay becomes much simpler as ready-made genomic fragments in vectors such as bacterial artificial chromosomes (BACs) or the complementary DNA (cDNA) library with known sequences can be used directly. The second method is to narrow down the genomic region containing the causal mutation by mating mutants with wild-type animals and followed the linkage between genetic markers with the phenotype. Once a small genomic region is identified by linkage analysis, subsequently sequencing of the region can be conducted to isolate the causal mutation (Collins, 1992; Vitaterna et al., 1994). This process is very labour intensive and time consuming. The development in whole-genome, exome and RNA sequencing technologies will facilitate mutant identification process.

3.2. Physical agents

Physical agents such as gamma-ray irradiation have been used to efficiently generate genome-wide mutations (Chu, 1971; Urlaub *et al.*, 1986; You *et al.*, 1997; Munroe *et al.*, 2000). The mutations generated by gamma-rays are typically large deletions, duplications, amplifications, translocations and more complex rearrangements, causing both loss- and gain-of-function mutations. One advantage of large deletions is that the whole genome can be covered with a relatively small number of mutants. However, identifying a causal gene-phenotype relationship is difficult, because of the large number of genes affected in each clone. Techniques such as comparative genomic hybridisation (CGH) arrays can be used to locate the regions of alterations in the genome. Causal regions can be further narrowed down by identifying the commonly altered region in independent cell lines followed by complementation assays to re-introduce the genes within the region to rescue the phenotype. Such a method is also routinely used in human genetics in isolating disease causing genes in patient cohorts with overlapping regions of the chromosome deleted.

3.3. Biological insertional agents

Retroviral and DNA transposons are commonly used as recombinant vectors to mutate the host genome. These vectors are flexible and can accommodate different molecular designs to achieve mutagenesis. Additionally, they serve as molecular tags to identify the mutated gene, a significant advantage over chemical and physical mutagenesis.

3.3.1. Retroviral vectors

Retroviral vectors have a long history of use for the introducing exogenous DNA into mammals efficiently. Exogenous retroviruses were first used to experimentally alter the mouse germ line in the 1970's, and this started the insertional mutagenesis research in mice (Jaenisch, 1976). The observations of leukaemia in mutagenised mice led to the recognition that retroviral insertions could alter the activity of endogenous genes. Somatic mutagenesis using retroviral vectors by injection of them into newborn pups can also give rise to cancer and the cloning of common viral insertion sites subsequently led to the identification of causal genes to these cancer (Liao et al., 1995; Shen et al., 2003; Uren et al., 2008).

The integration of a retrovirus in protein-coding regions can disrupt gene expression, leading to loss-of-function mutants. Retroviral integration can also provide gain-of-function mutants due to the fact that viral LTR contains strong enhancer element, which can ectopically drive the expression of genes nearby (Stocking et al., 1985). However, wild-type retroviruses are not efficient mutagens of the mammalian genome, thus cells with such retroviral integrations are phenotypically neutral. The incorporation of elements within the retrovirus to increase the frequency of mutagenesis improved their mutagenicity compared to wild-type retroviruses (von Melchner and Ruley, 1989; Reddy *et al.*, 1991). This led to the widespread adoption of insertional mutagenesis in the mammalian genome. The classical design such as promoter trap, which will be described later, has also been adapted to DNA transposon-mediated insertional mutagenesis (Collier et al., 2005; Dupuy et al., 2005; Keng et al., 2005). However, it has become increasingly apparent that retroviral integrations have a severe non-random genome distribution, with both "hot-" and "cold-" integration spots in the host genome (Kitamura *et al.*, 1992; Withers-Ward *et al.*, 1994; Guo, 2004; Hansen *et al.*, 2008). The large resource of retroviral gene-trap clones, TIGM OmniBank II, provides a useful dataset for analysing the retroviral integration patterns in ES cells (Hansen et al., 2008). The bank possesses over 350,000 ES cell clones, with insertions in 10,433 unique genes. The trapping events do not seem to have any chromosomal bias for integration. However, only 27 % of the genes in this resource have been trapped once and the rest of the genes trapped at multiple times with several clones with insertions in the same gene a few hundred times. With such highly uneven integration patterns, mutating genes in the retroviral integration "cold-spots" is difficult and requires highly redundant coverage of the genome and many genes will still remain un-touched. In addition to their non-random genome distribution, retroviral vectors also have several other limitations, including a restricted cargo capacity less than 10 kb, restriction on delivery of intron-containing cargos, some viral LTRs are prone to silencing and RNA intermediates are not always stable.

3.3.2. DNA transposons

Transposable elements are "mobile" genetic elements that are major components of the mammalian genome. There are two classes of transposons that are distinguished based on

the mechanism of mobilisation. Class I elements, retrotransposons, transpose with a "copy-and-paste" mechanism via an RNA intermediate. Class II elements are DNA transposons using a "cut-and-paste" mechanism. Transposable element-derived sequences make up about 45 % of the human genome (Lander et al., 2001) and 37.5 % of the mouse genome (Waterston et al., 2002), of which the majority are retrotransposon-derived sequences. These transposable elements are likely to be derived from horizontal transfer from bacteria to vertebrate lineages. In the human genome, there has been a marked decline in the activities of DNA transposons which appear to become completely inactive compared to those in the mouse genome measured by the lineage-specific transposons (transposons that are present in the mouse but not in human) versus the ancestral elements (Lander et al., 2001).

DNA transposons encode a transposase protein flanked by inverted terminal repeats (ITRs). The transposase binds to the terminal inverted repeats and excises the element from the donor locus and insert it in a new location elsewhere in the genome. The transposons can also function in a bi-partite system, in which the transposase can be separated from the ITRs and supplied *in trans*, thereby creating a non-autonomous transposon vector that can harbour unrelated DNA cargo. This unique property has been harnessed extensively as a molecular vehicle for transgenesis and insertional mutagenesis in a wide range of model organisms. In bacteria, high-density insertional mutagenesis with DNA transposons *Tn5* has achieved the genome-wide saturation mutagenesis (Langridge et al., 2009). In *Drosophila*, *P* element has been extensively used for the generation of random insertions to cause gene inactivation either by insertion of the element itself or by subsequent imprecise excision of the primary insertion events (Daniels *et al.*, 1985; Cooley *et al.*, 1988).

The lack of active DNA transposons in mammals hindered their application of insertional mutagenesis in experimental organisms such as the mouse and the rat using existing strategies developed in other organisms. In 1997, the first mammalian-active DNA transposon, *Sleeping Beauty* (SB) a member of the Tcl/Mariner family, was re-activated based on "ancient" sequences found in fish (Ivics et al., 1997). Since then, the mammalian DNA transposon toolkit has been expanded by the discovery and development of several members

from different families, including native transposons such as *Tol2*, *piggyBac*, and re-constructed transposons such as *Frog Prince* and *Hsmar1* (Ivics et al., 2009). Not only can DNA transposons facilitate mammalian genetic and genomic research, but their application to gene therapy may potentially confer significant advantages over the viral-mediated gene transfer for many diseases.

3.3.2.1. *Sleeping Beauty* and *piggyBac* possess different characteristics

As well as SB, which is widely used in mammals, *piggyBac* (PB), a transposon system from the *piggyBac* transposon family, has also been utilised in the mouse and cultured mammalian cells. *piggyBac*, originally isolated from the cabbage looper moth *Trichoplusia ni* (Cary et al., 1989), exhibits a highly efficient transposition in diverse genera of insects and vertebrates. *Sleeping Beauty* and *piggyBac* have different characteristics.

With respect to integration preference, SB shows a small bias towards genes than intergenic regions, whereas PB has a stronger bias toward intragenic integrations in both "vector-to-genome" and intra-chromosomal mobilisations without selection for actively transcribed regions of the genome (Yant *et al.*, 2005; Liang *et al.*, 2009). For intra-chromosomal mobilisation, SB has a strong tendency to land into *cis*-linked sites in the vicinity of the donor locus; a phenomenon termed "local hopping". In studies conducted both *in vitro* and *in vivo*, over half of the SB transposons excised from the donor locus landed in the donor chromosome, within the 4-Mb region near the donor site having this highest density of insertions (Keng et al., 2005; Kokubu et al., 2009; Liang et al., 2009). Local hopping is also observed with PB-mediated intra-chromosomal mobilisation, although to a much lesser extent than with SB (Wang et al., 2008b). Local hopping has been demonstrated with other transposon systems such as *P* element of *Drosophila* and *Ac/Ds* elements of *Zea mays* at several different donor locations (Moreno et al., 1992; Tower et al., 1993). Therefore, local hopping is likely to be a universal phenomenon during intra-chromosome transpositions of DNA transposons.

The transposition efficiency of *piggyBac* has been shown to be the highest for both vector-to-genome mobilisation and intra-chromosomal transposition in several direct comparison studies in mammalian cells (Wu *et al.*, 2006; Liang *et al.*, 2009). Significant efforts have been made to improve the SB transposition efficiency using a random mutagenesis method to generate transposase mutants. The most recent version of the hyperactive SB transposase (SB100x) showed a 100-fold increase in intra-chromosomal transposition efficiency compared to the first generation (Mates et al., 2009). However, in mouse ES cells, a direct comparison of intra-chromosomal transposition efficiency was conducted for *piggyBac* and *Sleeping Beauty* using *Hprt* locus as the donor site for identical cargo carried by either of the transposons. *piggyBac* showed an over 100-times higher transposition efficiency than *Sleeping Beauty*, even when the hyperactive *Sleeping Beauty* transposase, SB100X , was used (Liang et al., 2009). Progress has also been made in generating a mammalian hyperactive PB transposase with a ten-fold increase in the excision efficiency (Yusa, K, unpublished).

Although the transposition activity of SB was found to be very low in mouse ES cells (Luo *et al.*, 1998; Liang *et al.*, 2009), its transposition is much higher in the mouse germ line and somatic cells (Collier et al., 2005; Dupuy et al., 2005). This may due the epigenetic status of the transposon. SB transposase can mobilise transposons that are methylated with 100-fold higher activity than the non-methylated elements (Yusa *et al.*, 2004b; Ikeda *et al.*, 2007). Transgenic mice harbouring the transposon concatemers are likely to be methylated at the donor site; therefore SB transposition may be significantly enhanced *in vivo*.

There are several other unique features of PB that are advantageous in certain applications. PB possesses a very large cargo capacity, whereas SB shows diminished transposition when its cargo size reaches 10 kb (Karsi et al., 2001). It has been shown that PB can transpose with a 14.3 kb cargo with minor loss of transposition efficiency (Ding *et al.*, 2005b). Genomic cargo size up to 100 kb can be mobilised in a “vector-to-genome” integration assay and be excised from the genome in mouse ES cells (Li, MA, unpublished, chapter 7 of this thesis). This superior cargo capacity of PB will facilitate many areas of research in transgenesis, chromosome engineering, complementation, and therapeutic gene delivery.

In contrast to most DNA transposons, PB excision does not leave any footprint; therefore, the genome is intact after transposon re-mobilisation in the host genome (Ding *et al.*, 2005b). This property has been exploited to generate transgene-free induced pluripotent stem (iPS) cells with minimal genome modification (Woltjen et al., 2009; Yusa et al., 2009).

Another important feature of the PB transposase (PBase) is that it is tolerant to molecular engineering. Fusion of domains to the C-terminal of the PBase protein is well tolerated in PB transposase in contrast to SB transposase (Wu et al., 2006). A useful inducible PB transposase has been generated by the fusion of the modified human estrogen receptor ligand-binding domain (ERT2). This inducible PB transposase can be very useful to temporally regulate transposition *in vitro* and *in vivo* with 4-hydroxytamoxifen administration (Cadinanos and Bradley, 2007).

The comparison of the integration bias between PB and retroviral vector was also compared in mouse ES cells and PB shows a much more random than comparable retroviral vectors (Wang et al., 2008a; Wang et al., 2008b). Even in small libraries (approximately 280 clones) of PB-mediated gene-trap clones, 8 % of the trapped genes were not previously identified in the retroviral based gene trap resource OmniBank II (Wang et al., 2008a). Thus PB integrations provide access to genes which have not been tagged in a more than 20-fold saturated retroviral insertion library. Comparable DNA mismatch repair screens have been conducted using gene-trap libraries constructed with either retroviral or PB vectors. Similar complexity libraries yielded all known mis-match repair genes in the PB-based library whereas just one of the known genes was identified in the retroviral library (Guo, 2004; Wang et al., 2008a).

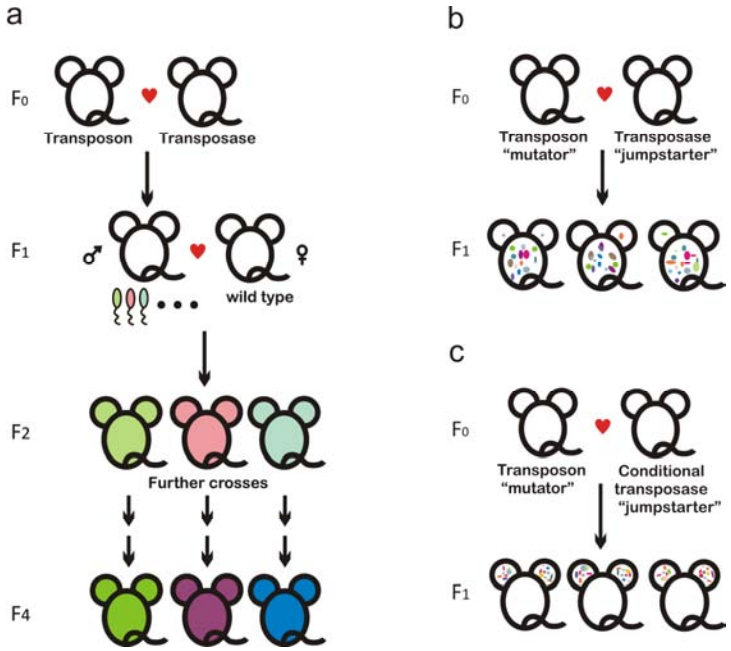
3.3.2.2. Transposon-mediated germline mutagenesis in mice

Germ-line mutagenesis in mice with a DNA transposon first established with SB, as SB was the first available mammalian-active DNA transposon. The mutagenesis is activated by crossing a transgenic mouse with a SB transposase (driven either by a constitutive promoter or germ-line specific promoter) with transposon transgenic line to generate double-transgenic mice in which the transposition is activated in the germline, Figure 1-5a. These double transgenic

mice are then further bred to wild-type mice to generate offspring with germline mutants (Fischer et al., 2001; Keng et al., 2005). Although "local hopping" phenomenon can limit the genome-wide mutagenesis *in vivo*, independent transgenic lines harbouring the transposons may be established covering different chromosomes to achieve genome-wide mutagenesis. Efficient germline mutagenesis has also been established using PB with similar approaches.

Efficient germ-line mutagenesis has also been achieved with PB. This was first demonstrated by co-injecting PB transposon and PB transposase under the control of the male germ-line specific promoter (*protamine 1*, *prm1*) to generate males with transpositions occurring in the male germ cells (Ding *et al.*, 2005b; Wu *et al.*, 2007). A large PB insertional mutagenesis project in mice is on-going in which new PB insertions are generated in large numbers and crossed to homozygosity in order to discover recessive gene function *in vivo* (Sun et al., 2008).

Figure 1-5: Transposon-mediated mutagenesis.



a, Transposon-mediated germline mutagenesis. The colour scheme reflects the genetic status with white represents wild-type, light colours and corresponding dark colours represent heterozygous and homozygous. Each sperm derived from the F_1 males carries different transposon integration sites. b,c, Transposon-mediated somatic mutagenesis, with b represents a whole-body somatic mutagenesis and c illustrate a tissue-specific mutagenesis. The coloured circles represent cells with the same clonal origins with independent transposon integration sites.

3.3.2.3. Transposon-mediated somatic mutagenesis for cancer gene discovery

Cancer is an evolutionary process in which cancer cells accumulate advantageous mutations over time and eventually they are able to proliferate autonomously, invade tissues and metastasise (Weinberg, 2006). Recent advances in DNA sequencing technologies have enhanced our ability to identify these mutations in cancer. However, whole cancer genome sequencing studies have identified two classes of mutations, the "driver" mutations and bulk of "passenger" mutations that do not contribute to the pathogenesis of the disease. The key to cancer gene discovery is to distinguish the "passenger" mutations from the real "driver" mutations (Stratton et al., 2009).

Somatic mutagenesis in mice has offered experimental test beds to identify and validate "drivers" in cancer formation and progression. Random insertional somatic mutagenesis in mice mimics the sporadic somatic mutations in human cancer, but with a much higher carcinogenesis rate due to use of efficient mutagens, and provide easily identifiable tags. Over the lifetime of the mouse, somatic mutations induced by constitutively active insertional mutagens accumulate, to a point that cells bearing "permissive" mutation collections expand clonally. Slow transforming retroviruses have provides a tool in cancer gene discovery in mammals (Kool and Berns, 2009), they have two major disadvantages, firstly their tropism, i.e. limited host tissue range for tumorigenesis and secondly significant biases in their integration sites, which limit the spectrum of genes identified with retroviruses. Transposon systems offer a flexible alternative approach that can overcome the limitations of the retroviral approach. The use of different promoters to drive the transposase expression, transposon-based *in vivo* somatic mutagenesis with either SB or PB generates a wide spectrum of tumour types (Collier et al., 2005; Dupuy et al., 2005). Although any one type of transposon may have preferences within the mouse genome, transposons from different families can be used in combination to achieve more complete genome coverage.

Somatic mutagenesis has also been coupled with pre-engineered genetic lesions or treatment with certain anti-cancer drugs to identify collaborative or mutually exclusive mutations (Uren et al., 2008) or mutations that confer drug resistance. This type of genetic interaction analyses

allow researchers to address questions about the genes and signaling networks involved in the dynamic process of tumour development.

Mutagenesis in the soma using transposons can be achieved using a classical breeding strategy of "jumpstarter" and "mutator" stocks. Transgenic "Mutator" lines carrying non-autonomous PB transposons can be crossed with a "jumpstarter" containing a transposase. The expression of the transposase can be either constitutive in the whole animal or controlled in a tissue specific manner, Figure 1-5 b and c. Constitutive expression of the transposase leads to continuous whole-body mutagenesis, resulting in cancer in many tissues could result (Collier et al., 2005; Dupuy et al., 2005). However, one major limitation of whole-body mutagenesis is early lethality in embryonic stages before cancer can develop. The extent of lethality is affected by the copy number of the transposons, and the activity of the transposase (Collier et al., 2009).

Two strategies have been to achieve spatial control of the transposition. One is to use a tissue specific promoter to drive the expression of the transposase. The other method is to use a conditional transposase allele with a floxed intervening cassette (*lox-stop-lox*) between the promoter and the transposase. Such mouse lines have been established for both SB transposase (*Rosa26-LSL-SB11*) and PB transposase (*Rosa26-LSL-mPBase*, Cadinanos et al, unpublished). When a Cre line with a tissue specific promoter is crossed to the transposon/transposase double transgenic animal, the deletion of the floxed intervening cassette can activate transposition in particular tissues. The advantage of the latter system is that a strong promoter can be used to drive the expression of the transposase to ensure the high transposition efficiency in the whole animal and many Cre lines are readily available. Several tissue-specific screens have been conducted to identify cancer genes in specific tumours, including colorectal (Starr et al., 2009), liver (Keng et al., 2009), hematopoietic (Dupuy et al., 2009) and neuronal (Bender et al., 2010) cancers using SB. The conditional activation of the transposase leads to permanent expression of the transposase, therefore continuous mutagenesis occurs in the tissue of interest throughout the development. Additional temporal control of the transposition events allows the mutagenesis to be turned

"on" and "off" at different developmental stages. *In vivo* temporal control of transposition can be achieved in principle by using the inducible forms of Cre or PB transposase (Metzger *et al.*, 1995; Vooijs *et al.*, 2001; Cadinanos and Bradley, 2007).

Using high-throughput sequencing technologies on a large number of tumour samples, it is possible to obtain a complete picture of the insertion sites in each tumour type at reasonable cost (Uren et al., 2009). With an ever increasing amount of data generated by this approach, the evidence of causality needs to be strong. In some cases, more than 100 insertion sites can be indentified from a single tumor sample, suggesting that most of the tumors may be polyclonal or many "passenger" insertion sites are present. The poly-clonality of the tumor samples can also lead to false positive interpretations of co-occurring pairs of mutations, as insertions that are found in the same tumour may not be in the same cell. Single cell analysis will be required to address this issue. Sequencing of micro-dissected tumour samples may reduce the complexity of insertional mutagenesis data and hence reduce the false positive calls for co-occurrence. The number of "passenger" integrations can be restricted by decreasing the number of transposons per cell used for mutagenesis. The identification of the common insertion sites (CIS) in most studies so far are either based on fixed windows or smooth Gaussian windows, and assume random distribution of the insertions in the genome to determine the number of insertions that define a CIS (Kool and Berns, 2009). However, all insertional mutagens have biases. Therefore, mutagen-specific integration patterns should be adopted to assign statistically significant CIS. Additionally, large candidate gene lists for various tumours have been generated; validation strategies are needed to understand how mis-regulation of these genes can transforms normal cells into tumour cells in what order and/or combinations, allow them to metastasise and resist therapy.

Another strategy has also been developed to validate putative cancer "driver" mutations identified in human cancer genome sequencing project. A promoterless cDNA array of candidate oncogenes harbouring the mutant versions of these genes has been mobilised by transposition in mice (Su et al., 2008). Mobilisation of the oncogenic array in the genome can result in the activation of the candidate genes driven by an endogenous promoter. The

correct level and spatial-temporal expression of the mutant form of the candidate gene can result in cancer, thereby confirming or refining the consequences of the observed mutations.

3.4. Insertional mutagen designs

Insertional mutagens such as retroviral and transposon vectors are often carry modular molecular designs in order to achieve high efficiency of mutagenicity. There are two basic types of designs determined by their means to inactivate (loss-of-function) or activate (gain-of-function) an endogenous gene. These two types can also be used in combination to achieve maximise the mutagenesis and such a strategy has been extensively used in somatic mutagenesis in mice for cancer discovery, which is described earlier in this chapter (Collier *et al.*, 2005; Dupuy *et al.*, 2005; Uren *et al.*, 2008; Starr *et al.*, 2009).

3.4.1. Loss-of-function designs

There are two main classes of gene-trap designs which give rise to loss-of-function mutants, promoter trap and polyA trap. Enhancer traps mainly serve the purpose of mapping enhancer elements, which seldom disrupt normal gene expression, and thus are not described here.

3.4.1.1. Promoter-trap designs

A promoter trap vector uses a reporter gene that is activated when the reporter has integrated in an intron or exon of a transcribed gene, in the correct orientation so that the reporter is transcribed under the control of the "trapped" gene. A promoter trap offers a means to select these integration events from a large number of random insertions in non-transcribed regions of the genome, and it also allows the regulation of the "trapped" gene to be investigated by assaying the activity of the reporter gene. To achieve this, a promoter trap vector contains a strong splice acceptor (SA), followed by a promoter-less reporter gene (βgeo is most commonly used) with a polyadenylation signal (pA) (Friedrich and Soriano, 1991; Friedrich and Soriano, 1993). Upon insertion in the correct orientation of a transcribed gene, the promoter of the endogenous gene drives the expression of a fusion transcript of mRNA from the endogenous exon(s) upstream of the trap, spliced onto the reporter and terminated at the pA signal 3' to the reporter. Translation of this fusion mRNA is initiated

from the initiation codon of the endogenous gene, Figure 1-6a. If insertion occurs just downstream of a 5' untranslated exon, by including the mammalian initiator codon (ATG) within a Kozak consensus sequence (Kozak, 1987) in the reporter gene, the resulting fusion transcript can also be translated.

One of the limitations of the promoter trap designs is reading frame restrictions which result in a functional reporter. Translation of a fusion mRNA can also only result in a functional reporter gene when the upstream endogenous exon is in the same reading frame and correct orientation as the reporter gene. Therefore, one in six of the trapping events can be selected for using the reporter. The other issue is the variable functionality of the reporter gene, when fused to protein products from the translation of upstream exons. Further improvements to these vectors have been achieved by incorporating viral elements, such as Internal Ribosome Entry Site (IRES) or viral self-cleaving 2A peptides, between the SA and the reporter. IRES from encephalomyocarditis virus (EMCV) is a non-coding RNA fragment noted for its ability to initiate high levels of cap-independent protein synthesis in mammalian cells (Jang et al., 1988). The incorporation of the IRES sequence allows the reporter gene to be independently translated from the upstream exons without any reading-frame restriction, although the translation of the ORF after the IRES tends to be at reduced level compared to the upstream ORF, Figure 1-6b.

The 2A peptide sequences derived from foot-and-mouth disease virus (F2A), equine rhinitis A virus (E2A), *Thosea asigna* virus (T2A) and porcine teschovirus-1 (P2A), contain a consensus motif which results in polypeptide cleavage between the 2A glycine and the 2B proline (2A, Asp-Val/Ile-Glu-X-Asn-Pro-Gly; 2B, Pro) (Szymczak et al., 2004). Through a ribosomal skip mechanism, the 2A peptide impairs the normal peptide bond formation between the 2A glycine and the 2B proline without affecting the translation (Donnelly et al., 2001). By inserting a 2A peptide sequence in the correct reading frame between the SA and the reporter, the trapped exons are fused with the reporter in a single transcript, but fusion proteins are not produced, Figure 1-6c. In this way, the reporter function is not compromised by a chimeric fusion with the translated portion from the upstream exon(s).

Another limitation of promoter trap vector is that the exogenous SA is in competition with the endogenous SA for trapping the gene. In some cases, trapping and endogenous expression can co-occur, resulting in some level of wild-type expression. Depending on the degree of leakiness, homozygous mutants do not always exhibit a loss-of-function phenotype. In rare situations, the gene trap cassette may be completely bypassed.

Promoter-trap-based mutagenesis can only mutagenise expressed genes, as reporter expression is dependent on endogenous gene expression. Thus promoter-trap based mutagenesis is only comprehensive in phenotypic screens in which used the cell type screened is the one used for mutagenesis. Screens involving differentiation and reprogramming will be limited in their coverage when promoter-trap based mutagenesis is used. However, if a phenotypic screen is conducted in the same cell type as mutagenesis, the use of promoter trap vector enriches for the expressed genome.

3.4.1.2. PolyA-trap designs

In contrast to promoter traps, polyA trap vectors are not restricted to mutagenising expressed genes. PolyA trap vectors consist of exogenous-promoter driven reporters followed by a strong splice donor (SD), but lacking a signal for transcription termination. The reporter gene produces a stable spliced transcript when the vector inserts into the correction orientation in an intron, capturing a termination and polyadenylation signal. Usually, stop codons in all three reading frames are also incorporated in the reporter, Figure 1-6d.

A strong bias for last intron insertion has been observed when using polyA trap vectors in mouse ES cells (Shigeoka et al., 2005). Thus, very few of the insertions result in null mutations, because only the small proportion of C-terminal proteins is truncated. This non-random distribution of trapping events is due to mRNA surveillance mechanisms, in nonsense mediated decay (NMD) of the reporter (Shigeoka et al., 2005). In mammalian cells, a stop codon is recognised as premature if it is located greater than 60 nucleotides 5’ to the last exon–exon junction, and a mRNA containing such a premature stop codon is degraded by

NMD. Therefore, the stop codon in the reporter gene is recognised as premature when a polyA trap is inserted in an intron other than the last one. Vectors have been developed to correct this conventional polyA trap bias and this has been achieved by adding an IRES sequence and three initiation codons in all three reading frames 3' of the reporter gene and the SD in the conventional polyA trap design, Figure 1-6e (Shigeoka et al., 2005).

Promoter and polyA traps can also be used in combination for gene-trapping and tagging the expression pattern of the trapped gene, Figure 1-6f. Using both traps, another strategy was developed which utilises NMD to degrade the trapped gene by engineering floxed internal exons containing premature stop codons downstream of a fluorescent reporter, Figure 1-6g (Skarnes et al., 2004). In this design, although trapping events enriched by the reporter are still biased to the 3' end of genes, NMD will cause destabilisation of the transcript when the trapped gene is expressed. Cre/*loxP*–based deletion of the internal exons containing the premature stop codon will stabilise the transcript which will be translated with a tag.

Gene trapping is a powerful technology that permits the generation of mutants on a large scale, allowing the investigation of gene function using either forward or reverse genetic approaches. Genome-wide mouse ES cell gene trapping resources using retroviral vectors have been established in the commercial sector as well as within the academic community. Lexicon Genetics, a mouse genetics-based biotechnology company, was the first to transform gene-trap technology into a high-throughput platform, generating more than 350,000 mouse ES cell clones, with 10,433 unique genes trapped (Hansen et al., 2008). Academic groups have also formed a consortium, the International Gene Trap Consortium (IGTC), to generate annotated gene-trap ES cells. Currently, this resource contains more than 430,000 clones, covering 12,431 genes (http://www.genetrap.org/).

3.4.2. Gain-of-function designs

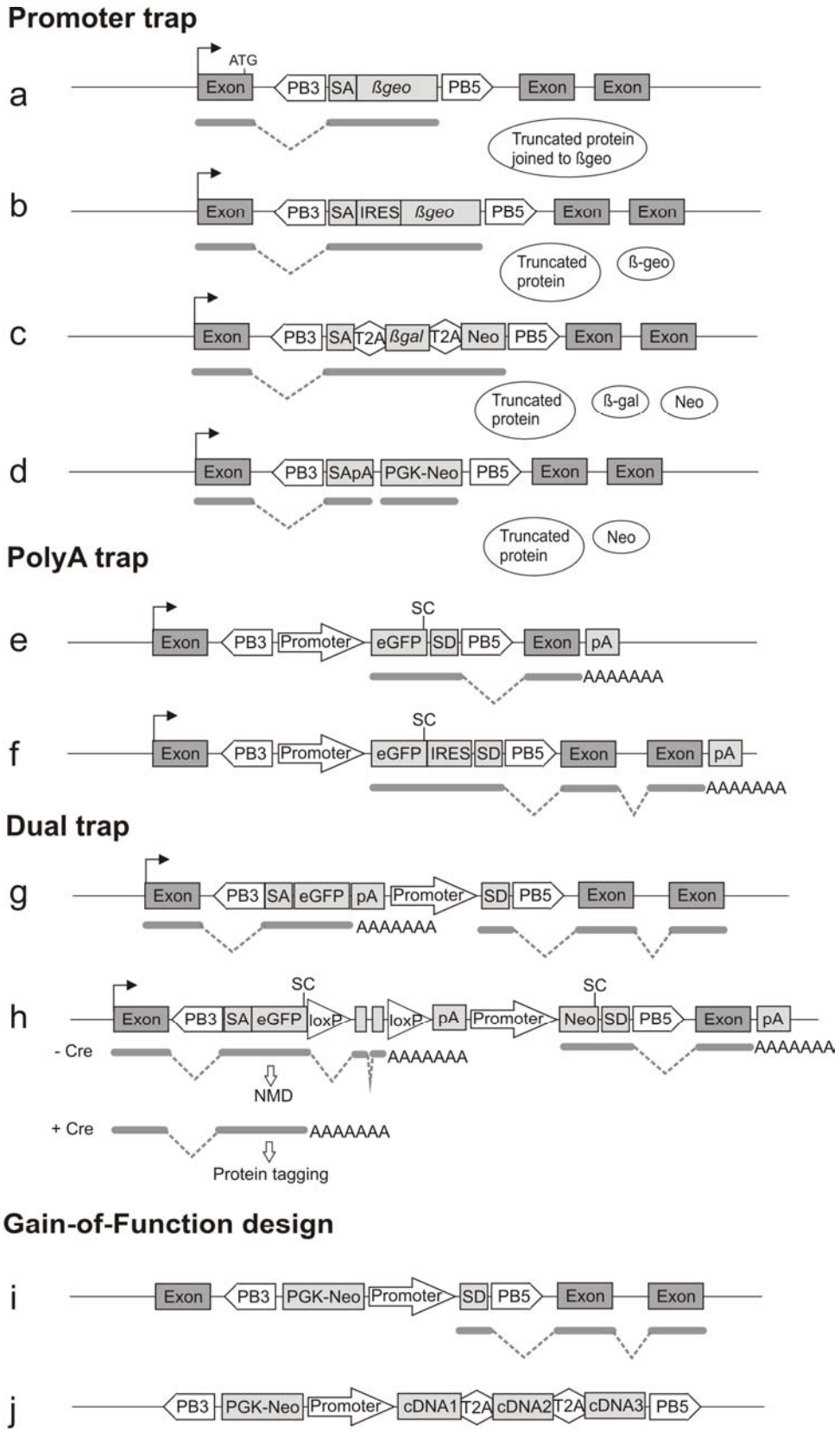
For generating gain-of-function mutants, a strong exogenous promoter can be engineered to drive the over-expression of either a full length or truncated gene product (depending on where the insertional mutagen lands with respect to the transcription unit, Figure 1-6h. An

alternative method is to over-express genes from a cDNA expression cassette. A single cDNA or a group of cDNAs connected by 2A peptides can be engineered into one vector, Figure 1-6i. 2A peptides provide a level of expression which is equivalent for all of the individual cDNAs (Donnelly et al., 2001). In this way, a cDNA library can be screened for the phenotype using either individual genes or groups of genes in combination.

Figure 1-6 legend (Figure on next page): Molecular designs for loss- and gain-of-function insertional mutagens.

All the designs are illustrated in the context of a PB transposon based vector. SA, splice acceptor; pA, polyadenylation signal; β-geo, β-galactosidase and neomycin resistant gene fusion gene; β-gal, β-galactosidase; PGK, mouse phosphoglycerate kinase promoter; SC, stop codon; IRES, internal ribosome entry site; T2A, *Thosea asigna* virus self-cleaving 2A peptide sequence; NMD, non-sense mediated mRNA decay; PB5 and PB3, *piggBac* transposon 5' and 3' ITR respectively. The grey line under each design represents the transcribed mRNA. The white circles for a~d represent the translated protein products.

Figure 1-6: Molecular designs for loss- and gain-of-function insertional mutagens.



4. Mammalian cells as genetic models

4.1. The mouse and rat as mammalian model experimental organisms

The studies in model organisms such as yeast, *C. elegans*, and *Drosophila* have provided an immense amount of knowledge on the molecular players in many evolutionarily conserved biological pathways. However, many features are unique in mammals compared to other organisms, such as the complex central nerve system, highly developed circulatory and respiratory systems, and the advanced immune responses, therefore, using mammalian model organisms is important for an understanding of these unique features. The mouse and rat are both commonly used mammalian model organisms. Despite having diverged from human approximately 75 million years ago for mouse and 12-24 million years ago for rat, they are both similar to human at the DNA sequence, anatomical and physiological levels (Waterston *et al.*, 2002; Gibbs *et al.*, 2004).

With the completion of the genome sequences for human, mouse and the rat, the detailed comparisons between the human and these laboratory mammalian models were conducted at genomic sequence level (Waterston *et al.*, 2002; Gibbs *et al.*, 2004). Over 90 % of the three genomes can be partitioned into large orthologue chromosomal segments with conserved linkage and identical gene orders, i.e. syntenic regions, ranging from hundred kilobases to multiple megabases. At the nucleotide level, approximately 40 % of the human genome can be aligned to both the mouse and the rat genomes, constituting mainly the coding regions and the regulatory regions of the genomes, despite the high rate of divergence of one base-pair substitutions as well as deletions and insertions. On a gene level, the mouse, rat and human genomes encode similar numbers of genes with highly conserved exonic and intronic structures. The proportion of mouse and rat genes with a single identifiable orthologue in the human is approximately over 80 %. These orthologues are also highly conserved at DNA sequence level (85 % median identity) and at protein level (88 % median identity), suggesting a highly likelihood of functional conservation.

In addition, the mouse and rat have a long history being used as experimental organisms dating back to the early 1800 (Simpson *et al.*, 1997; Waterston *et al.*, 2002; Gibbs *et al.*,

2004). With the rediscovery of Mendel's law of inheritance, geneticists used domesticated mice and rats to set up mating to test the theories of inheritance using the coat colour trait. These types of mating programs resulted in many inbred strains, lines selected for particular traits available for the study of human health and disease, and many modern strains are derived from those mating.

The mouse became the dominant mammalian model organism for geneticists due to the significant achievements in the isolation and culturing of the mouse embryonic stem cells and the genetic alterations introduced to these cells for germline transmission (details are described previously in this chapter). Although the rat is believed to be better models in certain human diseases such as arthritis, cardiac dysfunction, hypertension and neuroscience (Abbott, 2004), rat genetics has been hindered until very recently by the lack of embryonic stem cells for the generation of defined genetic alterations. In 2008, the derivation of authentic rat ES cells was achieved using an inhibitor cocktail (2i) within a molecularly defined culture condition that is only permissive to true pluripotent ES cells (Buehr et al., 2008; Li et al., 2008). The first knockout rat with p53 gene inactivated has recently been established using the 2i-derived rat ES cells (Tong et al., 2010). The availability of rat ES cells and possibility of conducting gene targeting will transform the genetic studies in the rat.

4.2. Mammalian cells as experimental models

Although the mouse and rat *in vivo* models have provide much knowledge in the mammalian molecular genetics, anatomy and physiology, their long generation time and requirement for large facilities in husbandry have made forward genetic approach *in vivo* very costly. Since the first establishment of medium formulations that support the continuous growth of mammalian cells *in vitro*, they have become the mostly widely used biological system. The development of techniques that allow genetic material to be easily delivered to mammalian cells further boosts the use of cell-based models for gene function characterisation. Cell-based models are simpler than whole animals due to their phenotypic and to some extent genetic uniformity and defined culture conditions within a controlled environment. Cell lines

are also highly scalable for genetic and biochemical analysis, have a shorter discovery time scale and are lower cost than whole-animal based classical genetic studies.

Studies using mammalian cell-based models differ from those in model organisms in several ways. Firstly, cell lines are derived from different tissues and developmental stages. Therefore, it is important to know the cell line origin and genotype as well as whether the chosen cell line possess the biological pathway of interest in order to produce a desired phenotype when mutagenised. Secondly, cultured mammalian cells display limited phenotypes for direct phenotypic screening, such as growth, differentiation, apoptosis, and senescence. The use of reporter genes such as green fluorescent protein (GFP) and β-galactosidase reporter enzyme, and selection markers can expand and diversify functional read-outs for phenotypes investigated *in vitro*. Finally, mammalian cell lines do not go through meiosis, thus their genomes are always predominantly diploid. This diploid nature poses a significant challenge in conducting genetic screens to isolate recessive genes, as the inactivation of both alleles of such a locus is required to evoke a phenotype. Several technologies have been developed to address this technical challenge, including utilisation of the natural occurring phenomena of loss of heterozygosity (LOH) and taking advantage of the haploidy in certain cell types. These different approaches are described in detail in Section 5 of this chapter.

4.3. Mouse ES cells as an attractive mammalian cell-based model

Pluripotent mouse ES cell lines possess several unique features that make them particularly attractive as cellular model systems for genetic screens. ES cells differ in several ways from many mammalian cell lines, which are immortalised cell lines, which are either transformed *in vitro* (e.g. Cos-7) or derived from human cancers (e.g. Hela). Firstly, ES cells are cellular entities that are physiologically relevant with *in vivo* counterparts during embryo development (Bradley et al., 1984). They offer the advantages of indefinite proliferation symmetrically, i.e. the daughter cells are identical to their parental cells, like other transformed mammalian cells. However, even with prolonged *in vitro* culturing, they maintain pluripotency and still behave like the cells in the inner cell mass of a blastocyst, generating all

cell types of a mouse (Bradley et al., 1984). For this reason, using ES cells as models is more physiological relevant than using transformed cell lines.

Secondly, transformed cell lines are often aneuploid with regional amplifications, deletions and rearrangements. Therefore, there are many pre-existing mutations in their genomes which may interfere with phenotype of interests. Unlike these cell lines, ES cells can maintain a stable diploid genome for many doublings without undergoing crisis or senescence. However, care must be taken in culturing ES cells and regular karyotyping analysis and subcloning is important to maintain a normal ES cell population. It has been observed that the rate of germ-line transmission drops when the passage number increase, due to random genetic alterations occurring during normal culturing. Trisomy for chromosome 8 and 11 are often observed in cultured ES cells and these genetic changes accelerate the growth rate, so that these abnormal clones can dominate the entire culture.

Thirdly, ES cells possess many unique features, such as their differentiation capacity to form an array of different cell types *in vitro* (Keller, 1995) and *in vivo* (Bradley et al., 1984), the ability to maintain their genome stability (Cervantes et al., 2002), their shortened G1 phase cell-cycles (Burdon et al., 2002). Therefore, ES cells not only provide a good model for investigating many biological pathways shared with other somatic cell types, but they also offer a panel of unique phenotypes that can be explored using genetic screens. The elucidation of the mechanisms which underline these ES cell specific properties will shed light on some pathological mechanisms in cancer and the aging process.

Finally, homologous recombination is two or three order of magnitude more efficient in ES cells than most other somatic cell types (Smithies *et al.*, 1985; Arbones *et al.*, 1994) , with the targeting efficiency ranging from 20 % - 90 % depending on the targeting vector designs and the locus accessibility. In other cell types, random integrations of the targeting vector are much more frequent than homologous recombination, impeding the isolation of gene targeting events. The only reported somatic cell line which shows comparable targeting efficiency to mouse ES cell is the DT40 cell line, a chicken B cell derived lymphoma cell line

(Buerstedde and Takeda, 1991). The high amenability of ES cells to multiple rounds of sophisticated genetic manipulation without compromising their pluripotency and genome stability allows the introduction of molecular designs into any locus to facilitate the requirements for a genetic screen.

5. Strategies for recessive genetic screens in mouse ES cells

In mouse ES cells, the simplest method for generating loss-of-function mutations is to sequentially target both alleles of a gene (Davis et al., 1993). The advantage of this approach is that gene targeting allows precise inactivation of the gene of interest, with the flexibility to generate conditional knockouts, hypomorphic alleles and to introduce point mutations. Another advantage is that these ES cells can be injected into blastocysts to derive homozygous mutant mice. However, this method is a lengthy and labour-intensive process, which requires two rounds of gene targeting with individual clones being isolated and genotyped at each step. With the availability of the genome-wide BAC libraries, the availability of accurate gene structures, and development of high-throughput recombineering technology (Chan et al., 2007), targeted mutagenesis can be conducted on a large scale. An international consortium has achieved single allele knock-out of thousands of genes (http://www.knockoutmouse.org/aboutkomp). Plans have been made to perform second allele targeting on a large scale. Once completed, this indexed homozygote ES cell mutant library will constitute a powerful resource for both forward and reverse genetic approaches to investigating gene function. Despite all of the recent advances, this method is still very costly and time consuming. Additionally, screen specific designs such as loss- or gain-of function mutations and the incorporation of reporter genes can not easily be incorporated into a pre-existing mutant ES cell library.

5.1. Loss of heterozygosity based strategies

Several methods have been developed which exploit the naturally occurring phenomenon of loss of heterozygosity (LOH), to recover rare homozygous mutant cells from heterozygous cells (Mortensen *et al.*, 1992; Lefebvre *et al.*, 2001; Guo, 2004; Yusa *et al.*, 2004a). In addition, the LOH events can be controlled at specific locations using site-specific recombination

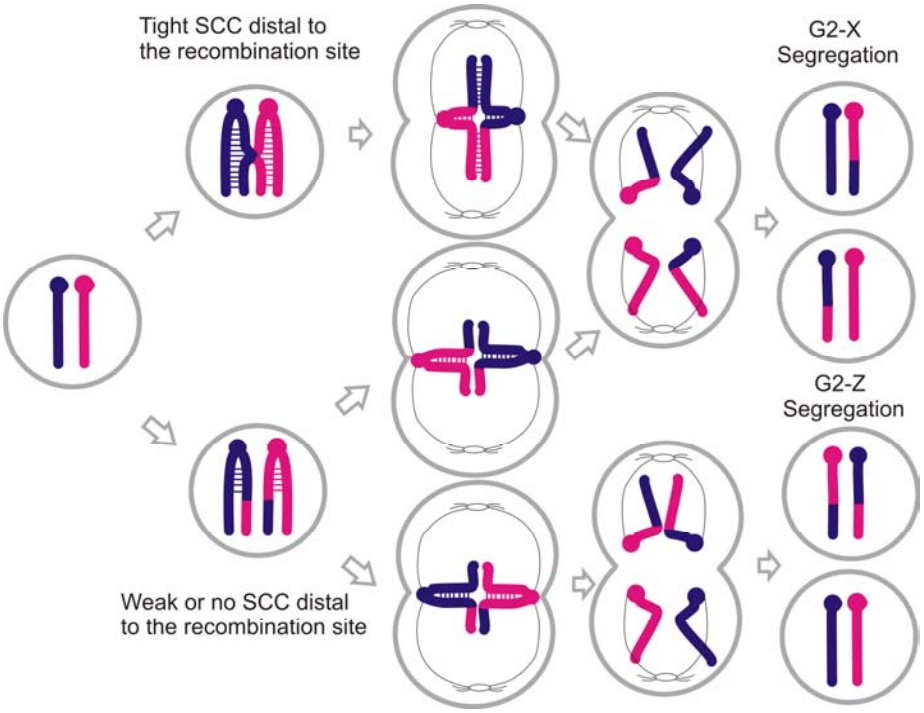
systems such as Cre/*loxP* and Flp/*FRT*. LOH arises by three possible mechanisms, mitotic recombination, gene conversion, or regional or whole chromosome loss and duplication. Because LOH is a rare event in wild type cells, a strong selection strategy is required for the isolation of LOH events.

5.1.1. Induced mitotic recombination using Cre/*loxP* system

Somatic mosaicism generated by mitotic recombination has been extensively used in *Drosophila* to conduct somatic recessive screens (Section 2.2.2.2. in this chapter). Mitotic recombination followed by G2-X segregation is a very useful genetic technique to obtain homozygote daughter cells from parental cells with a heterozygous genotype. Isolation of such homozygous cells provides an avenue to study recessive gene function.

In mitotic divisions in *Drosophila*, G2 X-segregation (recombinant chromatids segregate away from each other) occur in more than two thirds of the mitotic recombination events due to the unique characteristics of somatic chromosome pairing and the universal sister chromatid cohesion (SCC) effect, depicted in Figure 1-7 (Beumer et al., 1998). Mitotic spindles from each end of the spindle pole are physically constrained by the mitotic chiasma and attach to the kinetochores of one recombinant chromatid and the non-recombinant chromatid adjacent to it, but not two recombinant chromatids at the same time. However, if mitotic recombination occurs near the tip of the chromatids, the SCC is weak or non-existent and G2-X segregation is not favoured.

Figure 1-7: Mitotic recombination following G2-X and G2-Z segregation.



One of the forces driving G2-X segregation is sister chromatid cohesion (SCC) between the chromatid distal to the mitotic recombination site, with tighter SCC predominantly giving rise G2-X segregation. This is due to the physical constraint of mitotic spindle attachments to the kinetochores during segregation. G2-X segregation may not be favoured when crossing over occurs near the tips of the homologous chromosomes. G1 mitotic recombination can also occur (not shown), resulting in heterozygote daughter cells, which can not be distinguished from G2-Z segregations. The figure is adapted from (Liu et al., 2002).

A more controlled site-specific mitotic recombination can be achieved with the use of site-specific recombination systems such as Flp/*FRT* (McLeod et al., 1986) and Cre/*loxP* (Austin et al., 1981), which were originated from yeast and P1 bacteriophage, respectively. This system was first demonstrated in *Drosophila*, utilising pre-engineered *FRT* sites on the identical locations of the homologous chromosomes to obtain homozygous clones of cells somatically (Golic, 1991) and recessive screens were conducted using this system subsequently (Xu and Rubin, 1993). The details of this type of screen design in *Drosophila* have been described in Section 2.2.2.2. of this chapter, with an elegant example demonstrating the genome-wide approach of this system in recovering recessive mutations in a tissue-specific manner.

In mouse ES cells, mitotic recombination has been achieved with Cre/*loxP*, albeit with much lower efficiency than in Drosophila (Koike et al., 2002; Liu et al., 2002). The system designed in a way such that G2-X mitotic recombination events can be directly selected using drug selection markers (Liu et al., 2002). Although one locus investigated, *D7Mit178*, showed almost 100-fold higher rate for obtaining homozygous segregants using Cre/*loxP*-induced mitotic recombination than spontaneous LOH rate, four other loci studied showed much lower induced mitotic recombination efficiency consistently. Consequently, the application of this method to conduct recessive genetic screens on a genome-wide level is restricted. Firstly, individual *loxP* or *loxP* variants first have to be engineered in the chromosomes in the correct orientations to mediate mitotic recombination. In addition, if each chromosome is studied independently, 20 cell lines have to be made in order to cover the mouse genome. Several chromosomes can be engineered in one cell line, different *loxP* sites have to be used to avoid translocation events, which can also be enriched by selection. Finally, the rates of Cre/*loxP* mediated mitotic recombination vary significantly in different loci and many loci have comparable rates to the spontaneous LOH rate in wild type mouse ES cells, thus isolation of mitotic recombination events on a genome-wide scale is not feasible. Single-chromosome recessive genetic screens coupled with a chromosome-specific insertional mutagen such as *Sleeping Beauty* may be practical using the Cre/*loxP*-induced mitotic recombination method.

5.1.2. High G418 selection

One strategy developed to select for homozygote ES cells from cells with a single allele targeted with a Neomycin (*neo*) resistant selection cassette (Mortensen et al., 1992) relied on the level of Geneticin (G418) resistance. The mutant cells with a double dosage of *Neo* can be selected for by growing the heterozygous cell line in the presence of high concentrations of G418 (0.75 - 2 mg/ml) which selects for rare homozygous mutants arising via LOH (Mortensen et al., 1992). This method is simple to conduct and only requires the generation of heterozygous mutant allele expressing Neo.

Lefebvre and co-workers further developed the high G418 method using a hybrid ES cell line, R1, obtained from an F1 embryo from two 129 inbred substrains (129X1×129S3), to

discriminate between the homologous chromosomes using simple sequence length polymorphism (SSLP) (Lefebvre et al., 2001). Using these markers, they were able to identify homozygote mutants from six targeted *neo* insertions in four different chromosomes. It was also observed in their study that the LOH not only occurred at the targeted locus, but it extended to distant linked SSLPs 16-66 cM away. Thus possible mechanisms to generate LOH may be mitotic recombination, gene conversion, or regional or whole chromosome loss and duplication.

Although this method can be effective in selecting homozygous clones at some loci, it is difficult to select for homozygote conversion in parallel. Independent loci require different G418 concentrations to succeed in selection possibly due to the effect of the local genomic context on *neo* expression and homozygosity cannot be selected in many loci. Therefore, selection for homozygote mutants from a genome-wide randomly generated heterozygote mutant pool is not be efficient enough to eliminate the vast background of heterozygote cells and it will eliminate loci that require lower dose of G418.

5.1.3. *Blm*-deficient ES cell system

Patients with the autosomal recessive disorder Bloom's syndrome are due to mutations in the *BLM* gene (German, 1993). The cells derived from patients with Bloom's syndrome show a characteristic phenotype of hyper-recombination and genomic instability (German, 1993), and this can be visualised by cytogenetic analysis on metaphase spreads for homologous chromosome and sister chromatid exchanges (German, 1964; Zakharov and Egolina, 1972). *BLM* encodes a member of the ATP-dependent RecQ helicase family, which is highly conserved in evolution from bacteria to human and functions to unwind the DNA helix. Evidence on how Blm suppresses hyper-recombination comes from *in vitro* assays and genetic studies on its yeast homologue *SGS1*, where Blm interacts with TOPIIIα and cooperates with the strand cleavage and unwinding activities of this type I topoisomerase to resolve a double Holliday junction structure, suppressing exchanges between flanking DNA sequences (Gangloff *et al.*, 1994; Rothstein and Gangloff, 1995; Yamagata *et al.*, 1998; Wu and Hickson, 2003; Sung and Klein, 2006).

Blm-deficient mouse models recapitulate the phenotypes observed in human patients (Luo et al., 2000) and six different *Blm* mouse knockout alleles have been generated, namely, Blm^{tm1Brd}, Blm^{tm2Brd} , Blm^{tm3Brd}, $Blm^{tmChes1}$, $Blm^{tmChes3}$, Blm^{tm1Grd} (Chester *et al.*, 1998; Luo *et al.*, 1998; Goss *et al.*, 2002; McDaniel *et al.*, 2003). Four of these alleles were generated with replacement gene targeting using a drug selection cassette to substitute one or more exons of the *Blm* gene. All the replacement-based targeted alleles are homozygous lethal during embryonic development. The other alleles Blm^{tm2Brd} and Blm^{tm3Brd} were generated through an insertional targeting event, resulting in the duplication of exon 3 of the *Blm* gene and causing a frame shift for Blm translation (Luo et al., 2000). Blm^{tm3Brd} was derived from Blm^{tm2Brd} by Cre-mediated deletion of the floxed drug resistant selection marker. Although mice that are homozygous for the Blm^{tm2Brd} allele die during embryogenesis, $Blm^{tm3Brd/tm3Brd}$ mice are viable and cancer prone, mimicking one of the unique phenotypes in Bloom syndrome patients (Luo et al., 2000). The difference may resides in the PGK-driven *Neo* cassette, which is present in the Blm^{tm2Brd} allele, affects the expression of surrounding genes. It is also possible that the Blm^{tm3Brd} is a hypomorphic allele.

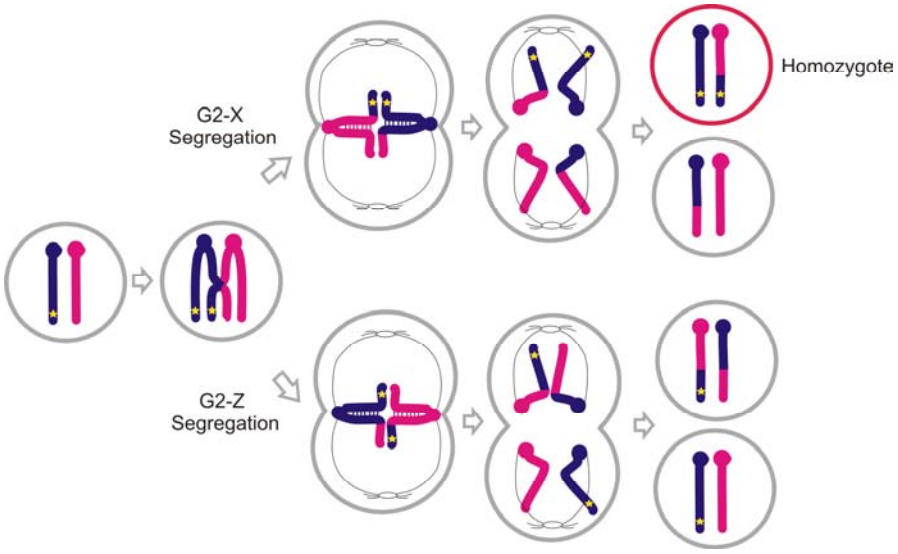
Conditional *Blm* alleles have also been generated, namely $Blm^{tmChes4}$ and Blm^{tet} (Yusa *et al.*, 2004a; Chester *et al.*, 2006). The Blm^{tet} allele has a tet-off cassette inserted upstream of the initiation codon of the Blm protein (Yusa et al., 2004a). In this allele, the expression of Blm is under the control of doxycycline which inhibits the binding of tTA to the TRE, thus the transcription of *Blm* mRNA is repressed. After withdrawal of doxycycline from the culture medium, tTA proteins bind to TRE and re-activate the *Blm* expression. These cells offer the opportunity to "switch off" *Blm* when mitotic recombination is required during the expansion for homozygote conversion, but to keep Blm "on" normally to maintain genome stability. Although a successful genome-wide recessive genetic screen has been conducted using the $Blm^{tet/tet}$ ES cells (Yusa et al., 2004a), Blm^{tet} has been shown to be leaky in mouse primary fibroblast cells (Hayakawa et al., 2006).

Blm-deficient mouse ES cells also display a genome-wide hyper-recombination phenotype and consequently an elevated LOH rate. The high frequency of crossing-over between

homologous non-sister chromosomes and the subsequent G2-X segregation in *Blm*-deficient ES cells generates homozygous mutant cells from their heterozygote counterparts, Figure 1-8. In a wild-type background, the rate of mitotic recombination has measured to be 3.5 x 10^{-5} events/cell/generation using Luria-Delbrück fluctuation analysis (Luria and Delbruck, 1943). The frequencies calculated based on $Blm^{tm1Brd/tm3Brd}$ and $Blm^{tet/tet}$ ES cells using different loci on different chromosomes are highly similar, and were measured to be 4.2 x 10^{-4} events/cell/generation (Luo et al., 2000; Yusa et al., 2004a). In other words, a single LOH event at a defined locus can be expected in every 2,400 divisions, i.e. if one heterozygous mutant cell is expanded to 1 x 10^4 cells, a few homozygous mutants will be converted in the culture. Whereas in wild-type cells, a heterozygote mutant cell has be expanded to around 2x 10^5 cells to produce one homozygous mutant, the 12-fold increase in the rate of LOH in *Blm*-deficient ES cells greatly enhances the rate of homozygous mutant production. This genetic background offers a simple means to derive homozygous mutants in parallel on a genome-wide scale enabling recessive genetic screens in mammalian cells (Guo, 2004; Yusa et al., 2004a).

Using a *Blm*-deficient genetic background together with insertional mutagenesis, genome-wide recessive genetic screens have been successfully conducted in ES cells. These screens cover a variety of biological pathways, including DNA mismatch repair, retroviral resistance, RNAi processing and toxin resistance screens (Guo, 2004; Wang and Bradley, 2007; Wang *et al.*, 2008a; Trombly *et al.*, 2009).

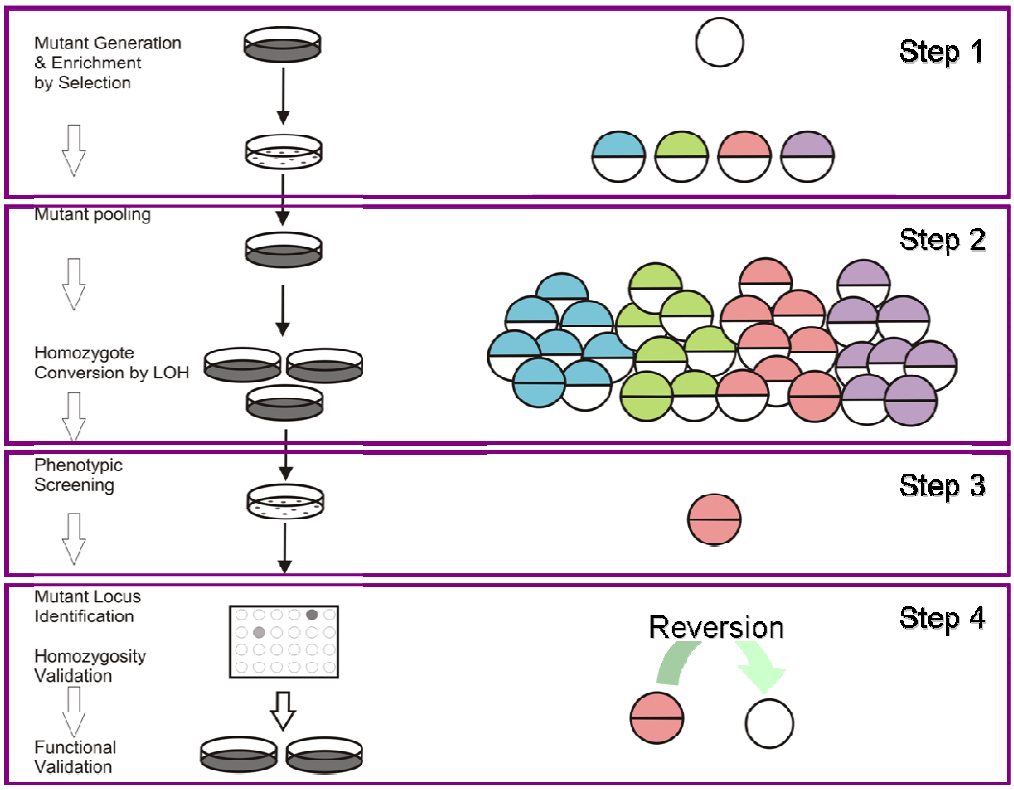
Figure 1-8: Generation of homozygous mutant from heterozygous counterparts.



In *Blm*-deficient ES cells, the rate of loss of heterozygosity induced by mitotic recombination is significantly elevated, thus increasing the probability of obtaining homozygous mutants.

Using *Blm*-deficient ES cells to conduct recessive genetic screens involve four main steps. Firstly, the parallel generation of genome-wide heterozygous mutations is conducted and mutant cells are selected. Secondly, the heterozygous mutant cells are pooled and expanded to allow LOH to occur in order to generate homozygous mutants. Thirdly, phenotypic screening is conducted on the mutant pools and finally the validation of the mutants for their genotype and functional relevance to the phenotype of interest, Figure 1-9.

Figure 1-9: Four steps involved in conducting recessive genetic screens using *Blm*-deficient ES cells.



The coloured balls on the right represent the genetic status of the mutants at particular stages during the screening processes, indicated by the purple frames. Half-coloured balls represent heterozygous mutants, whereas the fully coloured balls illustrate the homozygotes.

There are several considerations when using the *Blm*-deficient ES cell system to conduct a successful screen. Firstly, the LOH rate differs along the length of the chromosomes, with LOH rates higher towards telomeres compared with centromeric ends due to physical constrains during crossing over. The LOH rate in *Blm*-deficient cells was estimated based on three independent loci (*FasI* on Chr.1, *Nanog* on Chr.6 and *Gdf9* on Chr.11), with two (*FasI* and *Nanog*) are close to the telomere ends and Gdf9 is toward the middle of Chr. 11. Obtaining homozygous mutants for genes residing closer to the centromeres may require more cell doublings to cover the probabilities of obtaining homozygous mutants of centromeric loci. Secondly, heterozygote-to-homozygote conversion is a stochastic process. Therefore, culturing independent mutant pools is important and should limit "jack-pot" effects of a single mutant which converts to homozygosity early during expansion and dominate the pool.

A major consideration is the phenotypic read-out in the screen design, as not all screens are suitable in the *Blm*-deficient system in a pooled format. Based on the previous calculation using a *Blm*-deficient background, the ratio of homozygote to heterozygote cells is 1 to $1x10^4$ after expansion. This means that the screening method must be sensitive and specific enough to be able to isolate only a few relevant homozygote mutants from a large pool of irrelevant cells. In addition, recessive mutations causing phenotypes involving cell death or which affect growth rates of the relevant clones cannot be conducted using such pools.

A final consideration is the *Blm*-deficient background itself. Because of the hyper-recombination phenotype, spontaneous mutations generated during the screening processes can also be converted to homozygosity. If a mutation is present in the gene which is relevant to the pathway of interest, cells harbouring this homozygous mutation can be isolated in the screen, however, the mutagen present in these cells are irrelevant to the observed phenotype. Therefore, confirmations on the homozygosity status of the mutagen and the causality between the mutagen and the phenotype are important means to identify such false positive background.

A further development to the existing *Blm*-deficient ES cell system in order to further enrich for homozygous mutants in pooled libraries using a double selection strategy (Huang, et al, unpublished). In this system, screens can be conducted in pooled formats because the homozygotes are heavily enriched. At the same time, this strategy allows provide a means to build an indexed homozygous mutant library.

The selection strategy of this method is based on the incorporation of a "switchable" or "deletable" selection marker pair delivered by an insertional mutagen. After heterozygous mutant expansion using the *Blm*-deficient ES cells, the rare homozygous mutants with these selection systems can be enriched and selected from the pool because homozygous mutants can express two selection markers simultaneously while the heterozygous mutants can only express one. This method has already achieved isolation of homozygous mutants in many independent loci on different chromosomes. However, the strong selection scheme for dual copies of the insertional mutagen also favours two other background events apart from the true homozygous mutants. Firstly, it was observed that some clones isolated with this selection scheme are aneuploid, including trisomy and tetraploidy (Huang, et al, unpublished). Such cells are functionally heterozygous. Secondly, if the insertional mutagen copy number is more than one per cell, these cells can dominate the pool as homozygosity is not required for such cells to confer double-drug resistance. This constraint imposes a technical challenge during the generation of mutants to ensure that only single copy of the mutagen per cell is achieved.

5.3. Haploid mammalian cell lines for recessive genetic screens

Another approach for conducting recessive genetic screens in mammalian systems is to use haploid mammalian cells. One of the main strength of yeast as a genetic tool is the ease with which recessive mutations can be isolated at its haploid life stage. Karyotypically stable haploid cell lines have been established in amphibians and insects (Freed and Mezger-Freed, 1970; Debec, 1984), and recently haploid medaka fish ES cell lines have also been established (Yi et al., 2009a). However, mammalian cells are rarely haploid sufficient. Occasionally, some tumour cells can survive with a near-haploid genome. A human KBM-7 chronic myeloid

leukaemia (CML) cell line subcloned from a heterogeneous population was established (Kotecki et al., 1999). This cell line has a haploid karyotype except a disomy from chromosome 8 and also contains a Philadelphia translocation. Up to 12 weeks in culture, more than 50 % of the cells can maintain as near-haploid (Kotecki et al., 1999). Using this cell line, genome-wide loss-of-function screens have been conducted by mutagenising the genome with an inactivating insertional mutagen. Carette and co-workers demonstrated the feasibility of this strategy to identify host factors used by several pathogens (Carette et al., 2009). One major limitation of this cell line is the fact that these cells are karyotypically and genetically not stable. They have a tendency to increase in ploidy with time in culture, as diploidisation offers grow advantages over haploid (Kotecki et al., 1999). Tumour cells are often loaded with mutations such as insertions, deletions as well as chromosomal amplifications and deletions. Therefore, many biological pathways may have been mutated in this genetic background, limiting the success of phenotypic screens in this type of cell line. Another concern is the physiological relevance of these cells for certain biological pathways of interest, as these cells are originated from cancer, possibly several biological pathways are disregulated.

6. microRNAs and their biogenesis pathways

The discovery of non-coding RNAs changed one of the traditional views on the central dogma of molecular biology. The RNA family is much broader than just the coding mRNAs that function as a "messenger". Many non-coding RNAs produce transcripts that function directly as structural, catalytic or regulatory RNAs. Comparative genome analysis and various experimental approaches such as cDNA cloning and high throughput sequencing have unveiled an abundance of non-coding RNAs. MicroRNAs (miRNAs) are among the many classes of non-coding RNAs including endogenous small interference RNAs (siRNAs), piwi interacting RNAs (piRNAs), small nucleolar RNA (snoRNAs), ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs).

6.1. The discovery of miRNAs

The founding member of the miRNA family, *lin-4*, was first discovered in *C. elegans* based on its role in postembryonic development (Chalfie *et al.*, 1981; Ambros, 1989). The seam cells in

C. elegans go through four distinct larval-stages (L1-L4) and exhibit stage-specific characteristics for their cell divisions. *lin-4* Loss-of-function causes the seam cells to reiterate the L1 cell stage at later stages, resulting in extra larval molts and absence of adult structures (Chalfie et al., 1981). Another mutant *lin-14*, which has the opposite phenotype to *lin-4*, shows L1 stage skipping and premature entrance into the L2 stage. *lin-14* encodes a nuclear protein, which is down regulated at the end of the L1 stage in order to allow progression to the L2 stage (Lee et al., 1993). The cloning of *lin-4* revealed a 22 nt RNA that has a partially complementary sequence to the 3' UTR of *lin-14*, and the negative regulation of *lin-4* on *lin-14* protein synthesis is dependent on the intact 3' UTR of *lin-14* (Lee et al., 1993). *lin-4* was also found to negatively regulate another protein, *lin-28*, which functions to initiate the developmental transition from the L2 to L3 stage (Moss et al., 1997).

The discovery of *lin-4* mediated target-specific translational repression defined a new mechanism of gene regulation in development. Seven years after the cloning of *lin-4,* the second miRNA, *let-7* was also discovered using forward genetic screen for developmental regulators of the larval L4 stage to adult transition (Reinhart et al., 2000). *let-7* regulates the translational repression of *lin-41* and *lin-57*, by binding to their 3' UTRs (Reinhart *et al.*, 2000; Abrahante *et al.*, 2003). At this point, it was realised that miRNA mediated target-specific translational repression may be a universal mechanism, not restricted to developmental controls in C elegans (He and Hannon, 2004).

The RNA structures and the regulatory mechanism of *lin-4* and *let-7* have provided rules to enable subsequent miRNA discoveries by comparative genomic analysis, *in silico* prediction and high throughput sequencing. Hundreds of miRNAs have been since identified in many organisms, and a large proportion of the mammalian transcriptome are predicted to be under miRNA regulation, although critical experimental validation is required to formally prove the existence of all the predicted miRNAs and the effect they play on their predicted targets (Chiang et al., 2010).

6.2. miRNAs and siRNAs

MicroRNAs are 19- to 25-nucleotide-long single-stranded non-coding RNA molecules that are derived from larger precursor molecules with a stem-loop structure (Bartel, 2004). These miRNA precursors are transcribed from specific genomic locations by RNA polymerase II. The pre-miRNAs are usually a few kb long with 5' caps and polyA tails. Endogenous siRNAs differ from the miRNAs in their origin. siRNAs are processed from long double stranded RNAs (dsRNAs) that are either exogenously introduced dsRNAs or are transcribed from the bi-directionally transcribed endogenous RNAs that are annealed to form dsRNAs. The dsRNAs are then enzymatically processed by Dicer and giving rise to siRNAs.

It was originally thought that siRNAs and miRNAs act in distinct pathways and the degree of complementary of the siRNA/miRNA with their target sequence determine their mechanisms of silencing. siRNAs have near-perfect complementarity to their target sequences and cause the cleavage of their targeted mRNAs, whereas miRNAs tend to be partially complementary to their target mRNAs and evoke translational repression of the target proteins without affecting the stability of the target mRNAs. However, numerous findings suggest that there is no clear distinction between the siRNA/miRNA mediated silencing. Most plant miRNAs have near-perfect complementarity to their target mRNA and mediate mRNA cleavage to silence their targets. Although animal miRNAs tend to be partially complementary, miR-196, possesses near-perfect complementary to the *Hoxb8* mRNA (Yekta et al., 2004). The miRNA *let-7*, which normally acts through translational repression *in vivo*, can also enter the RNAi pathway *in vitro* if complementary target RNA is supplied (Hutvagner and Zamore, 2002). Conversely, siRNAs with imperfect complementarity can evoke translational inhibition in mammalian tissue culture (Doench et al., 2003). Recently, endogenous siRNAs and shRNAs have also been found in mouse oocytes and mouse ES cells (Babiarz et al., 2008; Tam et al., 2008; Watanabe et al., 2008). The shRNAs are often produced through the transcriptional read-through of the inverted SINE elements. The siRNAs found in mouse oocytes are produced from long dsRNAs formed *in trans* by pseudogene/gene pairing (*trans*-nat-siRNAs) or *in cis* by antisense transcription (*cis*-nat-siRNAs).

Therefore, siRNAs and miRNAs are fundamentally similar in terms of their molecular characteristics and mechanism of action and the distinction between the two may be arbitrary.

6.3. miRNA biogenesis

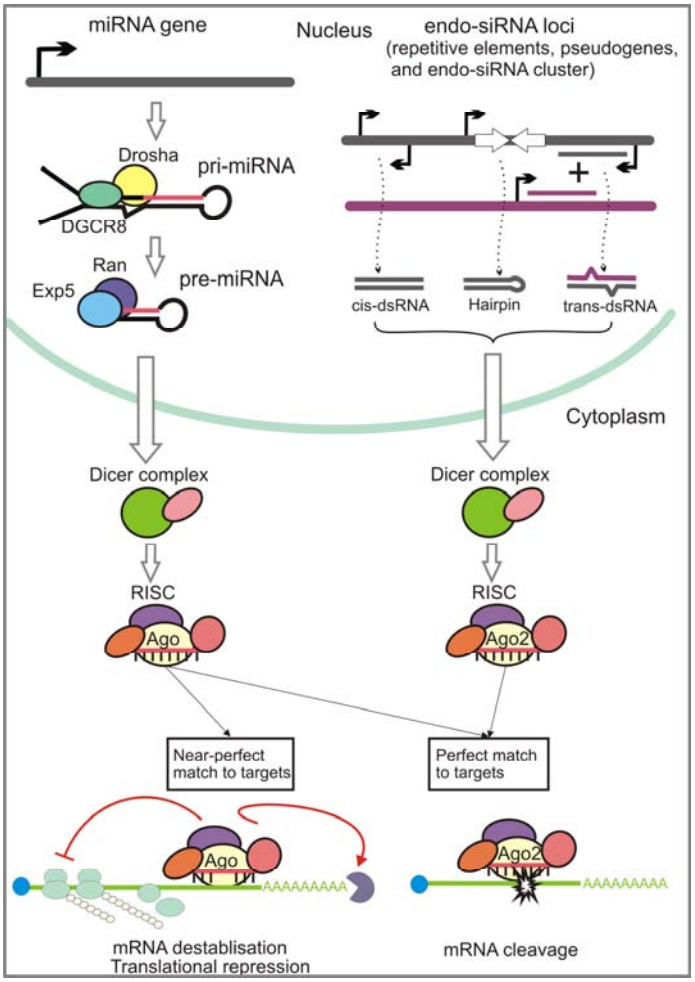
6.3.1. The canonical biogenesis pathway

miRNAs are encoded in the genome and firstly are transcribed as primary miRNAs (pri-miRNAs). They can reside in introns and exons of protein coding genes and non-coding genes or as independent loci. They can be located in a genome solely, or multiple miRNAs are located closely together to form clusters which are transcribed polycistronically. The pri-mRNAs form stem-loop structures with large unpaired segments on the opposite ends. The pri-miRNAs are processed into precursor miRNAs (pre-miRNAs), approximately 70 nt in length, regardless of their specific sequences, by a nuclear ribonuclease (RNase) III-like enzyme Drosha. The specificity of Drosha cleavage is guided by its partner Dgcr8 (known as Pasha in *Drosophila*), acting as a "molecular ruler", directing Drosha-mediated cleavage. Dgcr8 "measures" the distance (approximately 11 bp) from the flanking ssRNA segment to stem junction, and anchors Drosha to cleave the stem of the pri-miRNAs at that position specifically (Han et al., 2006).

After initial cleavage by Drosha in the nucleus, pre-miRNAs are transported to the cytoplasm by exportin-5 (Exp5), a ran-GTP dependent transporter (Lund et al., 2004). Once they have reached the cytoplasm, the pre-miRNA stem-loops are cleaved by Dicer, another RNase-III enzyme, to generate 21-25 nt dsRNAs that contain the mature miRNA and the passenger strand, named miRNA*. Dicer itself exhibits little sequence specificity for cleavage, the specificity of Dicer cleavage site on pre-miRNAs is based on the Drosha cleavage site, which is approximately 22 nt away from the 3' 2-nt overhang. Human immunodeficiency virus (HIV)-1 trans-activating response (TAR) RNA-binding protein (TRBP) recruits the Dicer complex to Ago to form the RNA-induced silencing complex (RISC) which achieves downstream effector functions (Chendrimada et al., 2005), Figure 1-10.

The use of miRNA biogenesis mutants has proven to be a very useful genetic tool to investigate the functions of miRNAs (Wang *et al.*, 2008c; Melton *et al.*, 2010) and endogenous siRNAs (endo-siRNAs) (Babiarz et al., 2008; Tam et al., 2008; Watanabe et al., 2008) (Chapter One, Section 6.5.3). Loss-of-function mutants in different biogenesis components can abolish the production of only the canonical miRNAs (Wang et al., 2007) or both miRNAs and endo-siRNAs (Kanellopoulou et al., 2005), enabling a functional dissection of these small RNAs in different biological systems (Wang *et al.*, 2008c; Rao *et al.*, 2009; Yi *et al.*, 2009b; Melton *et al.*, 2010; Song *et al.*, 2010). Endo-siRNA precursors are derived from transcripts of repetitive elements, pseudogenes or long stem-loop DNA structures. Several steps of their processing are shared with the miRNA processing pathways, such as Dicer cleavage and Ago2 association (Tam et al., 2008; Watanabe et al., 2008). Figure 1-10 shows the comparison of the canonical miRNA and endo-siRNA biogenesis pathways. However, the details of the endo-siRNA biogenesis are not completely clear and the biological functions of these molecules have not been elucidated. Identification of components that differentially regulate the miRNA or endo-siRNA production will facilitate understanding of small RNA processing and enable future research directed at understanding the roles of these small RNAs in mammalian development and physiology.

Figure 1-10: Canonical miRNA and endo-siRNA biogenesis pathway.



Mammalian Ago1-4 can associate with RISC in mediating miRNA effector pathways, although Ago2 is the only Ago protein with endonuclease activity which mediates mRNA cleavage. Ago2 has been shown to be associated with the endo-siRNA processing, which has not been shown for other Ago proteins. The endo-siRNAs are processed from transcripts derived from repetitive sequences and pseudogenes within the genome. Dicer and Ago2 have been shown to be involved in their processing and mediating gene silencing.

6.3.2. Differential roles of Dicer homologues

The multiple Dicer homologues present in some genomes can have different functions. Genetic and biochemical analysis of the two Drosophila Dicer homologues, Dicer1 and Dicer2, illustrate this point. Both Dicer1 and Dicer2 function in miRNA and siRNA production to facilitate RISC-mediated gene silencing. However, a loss-of-function mutant of Dicer1 exhibits

disrupted processing of pre-miRNAs, whereas loss of Dicer2 function affects siRNA maturation without compromising miRNA processing (Lee et al., 2004; Pham et al., 2004). Dicer1 requires co-factor R3D1 to process pri-miRNAs (Jiang et al., 2005), whereas Dicer2 forms a complex with R2D2 to enhance the target mRNA cleavage (Liu et al., 2003). In mammals, there is only one Dicer (Dcr-1) gene, therefore there may not be an equivalent Dicer functional distinction in mammalian systems.

6.3.3. Strand selection of the miRNA: miRNA* duplex

Following Dicer cleavage, the resulting 21-23 nt dsRNA duplex is loaded onto Ago protein to generate the RISC complex. One strand remains associated with Ago, whilst the other strand is degraded. The strand selection is based on the differential thermodynamic stability of the 5' end of the two arms of the miRNA : miRNA* duplex (Khvorova et al., 2003; Schwarz et al., 2003). The miRNA is mostly derived from the least stable of the 5' ends, suggesting that 5' end instability promotes the incorporation of the miRNA into the RISC. The instability of the 5' end provides an entry point for the RNA helicase to unwind the duplex, and the asymmetrical entry of the helicase determines the symmetry of the miRNA strand recruitment to the RISC. When the two strands have similar thermodynamic stability, both strands of the duplex are incorporated into the RISC at similar frequencies (Khvorova et al., 2003; Schwarz et al., 2003). siRNA strand selection is also based on this thermodynamic rule (Khvorova et al., 2003; Schwarz et al., 2003).

6.3.4. Choice of Argonaute (Ago) association

There are several homologues of Ago proteins; Ago1 and Ago2 in *Drosophila* and Ago1, Ago2, Ago3 and Ago4 in mammals. In *Drosophila*, the major factor that determines the sorting of RNA duplexes to two different Ago proteins is the degree of complementarity of the duplex (Forstemann et al., 2007; Steiner et al., 2007). In *Drosophila*, miRNA duplexes with central mismatches are preferentially sorted into Ago1, whereas perfectly matched siRNA duplexes are incorporated into Ago2 (Forstemann et al., 2007). In *Drosophila*, both Ago1 and Ago2 possess endonucleolytic enzymatic (splicer) activity, however in mammals, only Ago2 mediates endonucleolytic cleavage of mRNAs (Liu et al., 2004). In contrast to *Drosophila*, all

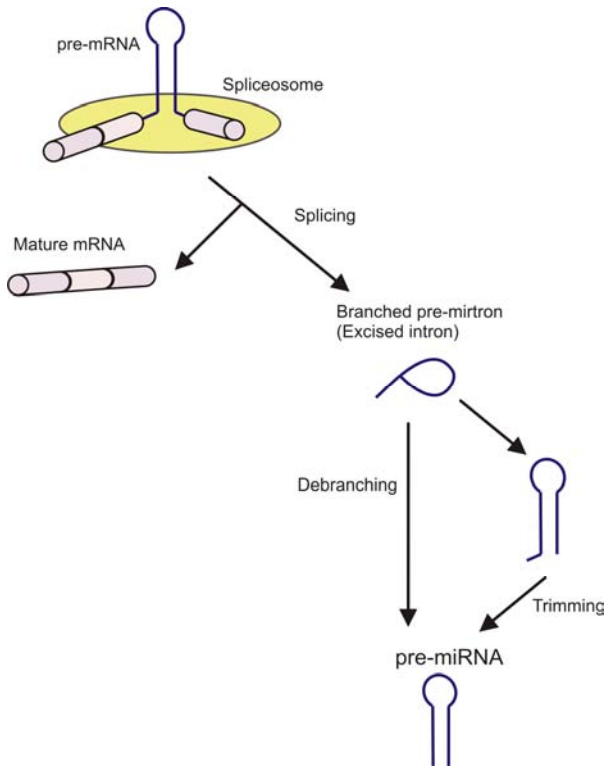
four Ago proteins in mammals seem to bind with miRNA indistinguishably and appear to have overlapping functions in miRNA-mediated translational repression, whereas Ago1 and Ago2 are preferentially involved in perfectly matched siRNA-mediated mRNA degradation (Liu et al., 2004; Su et al., 2009).

6.3.5. Non-canonical miRNA biogenesis pathways

Non-canonical pathways for miRNA biogenesis have also been observed. A class of intronic miRNAs known as mirtrons are produced by a Drosha-independent pathway (Okamura et al., 2007; Ruby et al., 2007). Mirtrons were first observed in *Drosophila* and *C. elegans*, and have also been found in mammals (Babiarz et al., 2008). Typical mirtrons are approximately 65 nts in length and resemble canonical pri-miRNAs, but they lack the lower stem of the pre-miRNA. Instead, the hairpin ends precisely match the splice sites. The "AG" splice acceptor of mirtronic introns typically adopts a 2-nt 3′ overhang to these hairpins, thereby mimicking a Drosha product. After splicing, mirtronic introns are de-branched from the lariat structure, further folded and trimmed in certain cases to give rise to pre-miRNA-like structures, which are then exported to the cytoplasm. The cytoplasmic processing of mirtrons requires Dicer and resembles canonical miRNA processing. Therefore, mirtrons are the endogenous equivalents of shRNAs, processed independent of Drosha/Dgcr8, Figure 1-11.

Another non-canonical pathway of Dicer-independent but Ago2-dependent biogenesis is observed with miR-451 in mice (Cheloufi et al., 2010). A dramatic loss of miR-451 was identified when comparing wild-type mice with mice possessing an inactive catalytic unit of Ago2. The miR-451 hairpin has a unique structure with a stem region of only 17 nt and with the mature miRNA sequence extended into the loop region (Cheloufi et al., 2010). The Pri-miRNA of miR-451 is processed normally by Drosha, but the Dicer step is skipped with pri-miRNA directly loaded into Ago2 for further trimming by Ago2 to mature (Cheloufi et al., 2010). The degree of usage of this non-canonical miRNA biogenesis pathway still needs to be determined.

Figure 1-11: Mirtron biogenesis pathway, bypassing the *Drosha* processing step.



This figure is adapted from (Kim et al., 2009).

6.4. Regulation of miRNA biogenesis

miRNA biogenesis is regulated both at transcriptional and post-transcriptional levels. Transcriptional regulation is a major regulation point to determine the spatial-temporal expression of miRNAs. This is achieved by RNA polymerase II (PolII) associated transcription factors binding to the promoter regions of the miRNA loci. For example, the miR-290 cluster has a mouse ES cell specific expression, and its promoter is associated with the pluripotent core transcription factors, Oct4, Nanog, and Tcf3 (Marson et al., 2008).

Post-transcriptional regulation also plays a crucial role in regulating miRNA function. In theory, any step of the miRNA biogenesis can be regulated. So far, relatively little is known about the mechanisms involved in the post-transcriptional regulation. Drosha processing is one of the most studied steps. One example is the regulation of miR-21 by bone morphogenetic protein (BMP)/transforming growth factor-β (TGfβ) signalling pathway in human vascular smooth muscle cells (Davis et al., 2008). In this study, SMAD proteins

activated by BMP/TGfβ were found to interact with Drosha and p68 to stimulate Drosha processing, although the detailed mechanism is still unknown (Davis et al., 2008).

Another example is p53's role in modulating miRNA processing to modify global expression against genome damage. In response to DNA damage, p53, a central tumour suppressor, has also been shown to facilitate the biogenesis of several miRNAs with growth-suppressive functions via an interaction with the Drosha processing complex (Suzuki et al., 2009).

The *let-7* cluster is also regulated post-transcriptionally during mouse development with highly elevated expression of mature *let-7* at 10.5 days of gestation (Thomson et al., 2006). The pri-*let-7g* is expressed at a constant level in both undifferentiated ES cells and during ES cell differentiation (Thomson et al., 2006). The RNA binding protein Lin28 was found to be responsible for the regulation of *let-7* maturation from protein pull-down experiments in embryonic carcinoma (EC) cell extracts using *pre-let-7g* as the bait (Viswanathan et al., 2008). Although the precise mechanism by which Lin28 selectively blocks *pir-let-7* processing is still unknown, several different actions of Lin28 have been proposed, including blockage of Drosha processing (Thomson et al., 2006; Viswanathan et al., 2008) or by inducing terminal uridylation of *pre-let-7*, which subsequently leads to blockage of Dicer processing and the decay of *pre-let-7* (Heo et al., 2008).

miRNA biogenesis can also be controlled in complex feedback loops that involve the biogenesis factors, the miRNA targets and themselves. Drosha and DGCR8 form a regulatory circuit to maintain miRNA production homeostasis. Drosha downregulates DGCR8 by cleaving *DGCR8* mRNA, whereas DGCR8 upregulates Drosha through protein stabilisation (Han et al., 2009).

A double-negative feedback loop is also used in miRNA biogenesis control to achieve efficient bi-stable switching during cell type commitment upon differentiation. One such example is the feedback regulation between *let-7* and Lin28 during neural stem cell commitment and ES cell differentiation. Lin28 selectively blocks *let-7* maturation in undifferentiated ES cells

(Thomson *et al.*, 2006; Heo *et al.*, 2008; Viswanathan *et al.*, 2008). In differentiated cells, mature *let-7* suppresses the Lin28 protein synthesis (Rybak *et al.*, 2008; Melton *et al.*, 2010).

6.5. Wider implications of miRNA biogenesis

Understanding the mechanism and regulation of the miRNA biogenesis pathway has wide implications in the understanding of pathological mechanisms of disease such as cancer and viral infections.

6.5.1. miRNA biogenesis and cancer

During normal mammalian development, only a handful of miRNAs are expressed in early embryos. During mid to late embryonic development, a large number of miRNAs are induced in a spatiotemporal manner (Kloosterman et al., 2006). In adult tissues, a large proportion of miRNAs are expressed, reflecting the differentiation status of different tissues. However, in many human cancers, miRNAs are reduced globally compared to normal tissues, reflecting de-differentiation cellular states in many cancers (Lu et al., 2005; Thomson et al., 2006). Therefore, it has been speculated that miRNA biogenesis pathways and their regulation are pivotal to maintain normal tissue homeostasis and cancer can evolve to "shut down" miRNA biogenesis which promotes unregulated cell growth.

Several lines of evidence coming from mouse models, human cancer cell lines and human genetics studies support this hypothesis. In a mouse lung cancer model, the knock-down of several key players of the miRNA processing pathway, Dicer, Dgcr8, and Drosha, caused tumorigenesis (Kumar et al., 2007). Although the precise mechanism of cancer initiation is still unknown, it has been proposed that the loss of the *let-7* family of miRNAs triggered the up-regulation of several oncogenes such as *c-myc* and *Ras*, which are both targets of the *let-7* family members. Mutations in TARBP2, a component of the Dicer1 complex, have been identified in a mismatch repair-deficient colon cancer cell line and when this tumor cell line was complemented with wild-type TARBP2, the tumor formation capacity in nude mice of these cells were reduced (Melo et al., 2009). A family linkage study has identified heterozygous germline point mutations in *DICER1* in patients with pleuropulmonary blastoma

(PPB) (Hill et al., 2009). Hemizygote DICER1 mutations are frequently found (approximately one in three human cancer cell lines) in the copy number data compiled on many tumour types from the Cancer Genome Project at the Sanger Institute (Kumar et al., 2009). Taken together, all these studies suggest that disruption of the miRNA biogenesis pathway can facilitate tumour progression, but it is not clear whether disrupted miRNA processing is the cause of the tumor initiation.

Conditional deletion of miRNA processing components has provided a useful means of examining the role of miRNAs in tumorigenesis. A conditional *Dicer1* allele has been combined with a $Kras^{LSL\text{-}G12D}$ background to use lung tumour formation as a model (Kumar et al., 2009). Heterozygous *Dicer1* mutants promote tumorigenesis, but homozygote *Dicer1* mutations are selected against in tumours. This suggests that *Dicer1* is a haplo-insufficient tumour suppressor. Partial loss of Dicer1 and possibly other effectors in the miRNA processing machinery is sufficient to cause a global reduction in miRNAs which contributes to cancer progression. Tumour burden was significantly decreased in $Dicer1^{fl/fl}$ mice after Lenti-Cre infection and the tumours arose from the $Dicer1^{fl/fl}$ mice were incomplete Dicer1 deletions. This selection against total loss of Dicer1 in tumours indicates that some miRNAs, which are expressed in normal tissues or induced upon cellular de-differentiation in cancerous cells, can act as oncogenes and contribute to tumour survival and growth.

6.5.2. Hijacking miRNA biogenesis by viruses

All herpes viruses express viral miRNAs which hijack the host miRNA processing pathways to support infection (Cullen, 2009). The viral miRNAs are not only advantageous not being recognised by the host immune systems, but also provide an efficient method to down-regulate key genes in the host immune systems and to regulate the entry to and exit from the latent stage of the viral life cycle (Gottwein et al., 2007; Murphy et al., 2008). Therefore, identifying novel components in the miRNA biogenesis pathway not only sheds light on the mechanistic insights into the miRNA processing, but also helps us to understand and identify potential targets in pathological scenarios.

6.5.3. miRNA biogenesis pathway mutants as tools to studying miRNA functions

As previously explained, *Dicer1* mutants have been a useful model to understand global miRNA repression in relation to tumorigenesis (Kumar et al., 2007; Kumar et al., 2009). Loss-of-function mutants in the miRNA biogenesis pathway also provide an avenue to access individual miRNA functions, as different miRNAs are believed to have functional redundancy, judging from their identical seed sequences. Therefore, loss-of-function mutants of a single miRNA may not show any phenotype. MiRNA biogenesis mutants can be complemented with miRNAs one at a time, to access their function and explore redundancy. Conditional *Dicer1* and *Dgcr8* mutants have been used to study individual miRNA function in mouse ES cells as well as in adult tissues. In addition, the role of non-canonical miRNAs and endogenous siRNAs can be catalogued in *Dgcr8* and *Dicer1* mutants in different tissues and developmental stages (Babiarz et al., 2008).

Dicer1 and *Dgcr8* null ES cells have been used to study miRNA function in cell cycle progression and control of the switch between self-renewal and differentiation. Dicer is required for the biogenesis of endogenous siRNAs and non-canonical miRNAs (Drosha-independent but Dicer-dependent processing) in mammals, so Dicer knockout defects can be attributed to both loss of canonical and non-canonical miRNAs as well as endogenous siRNAs. However, *Dgcr8* is exclusively involved in canonical miRNA processing. Both *Dicer1* and *Dgcr8* null mutant ES cells are viable and share several similar phenotypes such as retarded cell growth and resistant to differentiation (Kanellopoulou et al., 2005; Wang et al., 2007). The similarities in these phenotypes suggest that miRNAs are playing pivotal roles in the regulation of cell cycle progression and differentiation. Despite the similarities, *Dicer1*-null ES cells are more profound in growth arrest and differentiation phenotypes. In addition, the *Dicer1* mutant also exhibits epigenetic silencing of centromeric repeat sequences and reduced expression of homologous small dsRNAs (Kanellopoulou et al., 2005). The differences between the *Dicer1* and *Dgcr8* null ES cells are likely to be linked to endogenous siRNAs or miRNAs generated from the non-canonical pathways.

Cell cycle arrest at the G1 phase can be rescued by complementing *Dcgr8* null ES cells with members of the miR-290 cluster with a specific seed sequence (AAGUGCU), termed ES cell cycle regulating (ESCC) miRNAs. These ESCC miRNAs directly target and translationally repress the inhibitors, such as p21, of the Cdk2-cyclin E complex that control the G1 to S phase cell cycle transition. Therefore, these ESCC miRNAs promote the cell cycle transition at the G1 to S phase (Wang et al., 2008c).

Resistance to differentiation of *Dicer1* and *Dgcr8* null ES cells is observed both *in vitro* as well as *in vivo*. Dicer null ES cells fail to make chimeric mice when introduced into blastocysts, and upon subcutaneous injection into nude mice, they did not to give rise to teratomas with a heterogeneous mix of differentiated cell types (Kanellopoulou et al., 2005; Wang et al., 2007). Wild-type ES cells show a progressive loss in expression of pluripotent factors, such as Oct4, Nanog, Sox2, and Rex1 during *in vitro* differentiation in embryoid body (EB) formation assay or in response to differentiation inducing agents such as retinoic acid. However, *Dicer1* mutants showed sustained Oct4 expression even after five days of EB formation, and *Dgcr8* mutant ES cells show persistent expression of Oct4, Nanog, Sox2, and Rex1 in retinoic acid induced differentiation up to eight days from induction (Kanellopoulou et al., 2005; Wang et al., 2007). Under these differentiation inducing conditions, wild-type ES cells shut down the expression of these pluripotency factors within the first two days. This delay in switching off the pluripotent programs and initiating differentiation suggests that some miRNAs are directly and indirectly involved in one or both of these processes.

Some evidence suggests that the switch between pluripotency and differentiation can be regulated by two classes of miRNAs with opposing roles; members of the *miR-290* cluster and the *let-7* family members (Melton et al., 2010). In *Dcgr8* null ES cells, introduction of *let-7* can suppress pluripotent factors such as Oct4, Sox2, and Nanog, but this suppression does not occur in wild type ES cells. Introduction of *miR-294* together with *let-7* into Dgcr8 null ES cells blocks *let-7* mediated self-renewal suppression. Introduction of *miR-294* up-regulates Lin28 and *n-* and *c-Myc* (Melton et al., 2010). Both Lin28 and Myc have inhibitory roles in *let-7* expression (Thomson *et al.*, 2006; Chang *et al.*, 2008; Heo *et al.*, 2008; Viswanathan *et al.*,

2008; Lin *et al.*, 2009). Myc upregulation also forms a positive feedback loop in promoting the ESCC miRNA expression. In addition, Myc is reported to suppress miRNAs that are expressed in differentiated cells and upregulates others expressed in ES cells to attenuate differentiation (Chang et al., 2008; Lin et al., 2009).

7. Thesis project design

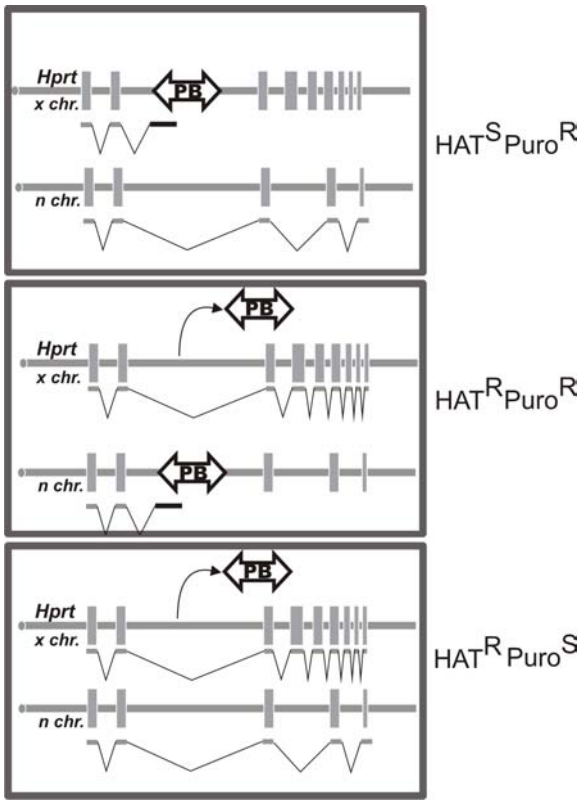
In my thesis project, I have designed a novel mutagen based on the PB transposon in *Blm*-deficient ES cells, with the aim of building a genome-wide mutant library for conducting recessive genetic screens *in vitro*. I applied this strategy using a selectable phenotypic screening to isolate components of the canonical miRNA biogenesis and effector pathways combined with my library.

The main considerations in designing a strategy for conducting genome-wide recessive screens is achieving broad genome coverage of the mutagenesis, obtaining a single insertional mutation per cell and the effectiveness of the mutagen to perturb the gene function. Good genome coverage by the mutagen provides a higher probability of mutating a relevant gene in the biological pathway of interest. Although no insertional mutagen can achieve full genome coverage in vertebrates, this class of mutation offers the great advantage of a molecular tag for mutant identification. A single copy of an insertional mutagen per cell is ideal for establishing the genotype-phenotype causal relationship. The effectiveness of the insertional mutagen is also crucial for the success of the recessive screen, as cells with insufficient gene inactivation of the mutagen may have a wild-type phenotype.

As previously described, the PB transposon is an efficient insertional mutagen with a more random genome distribution than retroviral vectors. In this project, a mutagenic PB transposon was introduced into *Blm*-deficient ES cells by gene targeting. When PB transposase (PBase) is supplied, the transposon can be excised from the donor locus and re-integrated in the host genome to evoke genome-wide mutagenesis. The initial excision event can be enriched by a positive selection scheme using *Hprt* as a selection marker. At the donor locus, the PB transposon inactivates the expression of *Hprt*, and renders the cells sensitive to

HAT. Upon PB excision, *Hprt* expression is restored and the cells become HAT resistant. The re-integration events can also be selected using a selection marker cassette carried by the transposon. In this design, the copy number of the PB transposon is maintained to be predominantly one. It has been demonstrated that PB excision of the mouse endogenous *Hprt* locus can result in genome-wide re-integration. Therefore, this mutagenic strategy should provide genome-wide mutagenesis with a single copy of the mutagen per cell. Figure 1-12 shows the strategy and the steps involved in conducting a recessive genetic screen.

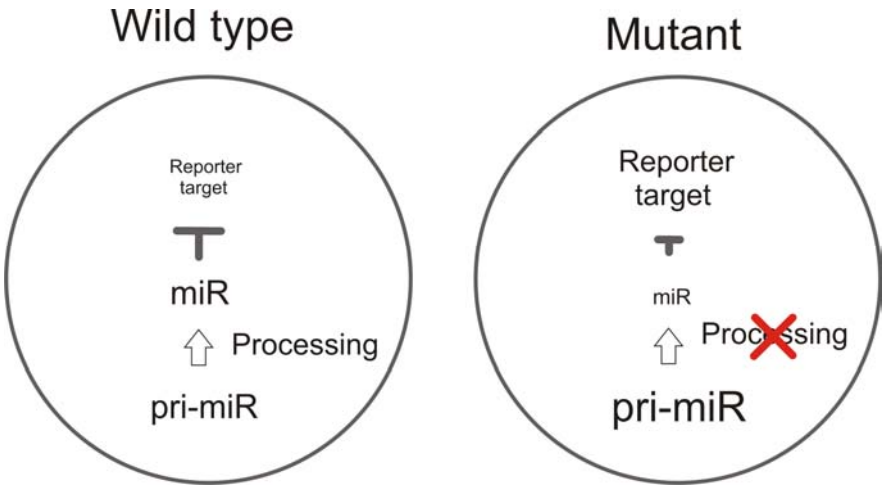
Figure 1-12: Schematic representation of the mutagenic strategy employed for this project.



Upper panel, *Blm*-deficient cells with a mutagenic PB transposon knocked into the intron 2 of the *Hprt* locus, rendering cells sensitive to HAT. Within the PB, an exogenous promoter-driven puromycin resistant cassette is also present (not depicted). Middle panel, upon transposase exposure, the PB transposon excises from the donor locus and re-integrates elsewhere in the genome. In many cases, PB lands into a gene, causing the gene to be inactivated. Lower panel, upon excision from the donor locus, re-integration may not occur and the transposon is lost. This scenario is selected against using the HAT and Puromycin double-selection.

Several successful screens have been conducted using *Blm*-deficient ES cells coupled with a chemical, viral and PB mutagenesis. In this project, I have designed a screen for isolating factors involved in the canonical miRNA biogenesis and its downstream effector pathways. The screening strategy involved the design of a miRNA reporter system so that, when the miRNA biogenesis pathway is perturbed and miRNAs can not be generated the reporter expression looses repression and becomes active, Figure 1-13. This provides a selection and phenotypic assessment strategy for isolating rare homozygote events from a pool of irrelevant cells.

Figure 1-13: A schematic representation of a reporter strategy to screen for miRNA biogenesis mutants.



In wild-type cells (left), an artificial or endogenous miRNA is processed normally. The mature miRNA can mediate the targeted reporter knockdown. In miRNA biogenesis mutant cells, the pri-miRNAs are accumulated due to the inability of miRNA processing; thereby the reporter target of the miRNA is not repressed, providing readout for the processing mutant phenotype.