1 Introduction

1.1 Genetic screen in model organisms

1.1.1 Genetic screens in yeast

It is hard to imagine that so much of our knowledge of cell biology comes from the simple unicellular fungi, the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*. Budding yeast and fission yeast are very different in their biology and are used to study different issues of cell biology. *S. cerevisiae* is the ideal model for signal transduction, cell cycle control, as well as chromosome structure (Forsburg 2001). On the other hand, *S. pombe* is a favourite for studies of cell-cycle control, mitosis and meiosis, DNA repair and recombination, and the checkpoint controls important for genome integrity (Wood, Gwilliam et al. 2002). In spite of their numerous differences, these organisms share one thing in common, namely the ease of genetic manipulation. This has resulted in the widespread use of these organisms as model organism to understand the biology of more complex systems.

S. cerevisiae has some milestones in biology. It was the first eukaryote to be transformed by plasmids, the first eukaryote for which gene-targeting became possible, the first eukaryote to be completely sequenced (Goffeau, Barrell et al. 1996). But surprisingly, the function of many of the 6000 or so genes still remains unknown. *S. pombe* has a relatively shorter history and a smaller research community. It contains 4,824 genes in its 13.8 Mb genome, which is the smallest number of protein-coding genes yet recorded for an eukaryote (Wood, Gwilliam et al. 2002). It diverged from budding yeast approximately 330 million years ago, around 4000 of the 4824 genes (83%) of *S. pombe* have homologues in *S. cerevisiae*, but they share no conserved synteny. Only 681 genes (14%) seem to be unique to *S. pombe*.

Both yeast species have a life cycle that is ideally suited for classical genetic analysis. They both can grow and divide as haploids and thus the phenotype of recessive mutations can be easily discovered. On the other hand, they also have a diploid sex cycle that allows the maintenance of lethal mutations and further characterisation of these. With the sequences of both genomes

finished (Goffeau, Barrell et al. 1996; Wood, Gwilliam et al. 2002), yeast geneticists are now equipped for genomic approaches, as well as their traditional genetic tools to answer fundamental questions: what is the function of each of the several thousand genes in the genome and how they work together to make the single cell function?

Because of the high efficiency of homologous recombination in yeast, it is relatively easy to design a construct with a nutritional or drug selection marker to disrupt a specific gene. In fact, budding yeast is the only eukaryotic organism in which every open reading frame has been knocked out or trapped (Kumar and Snyder 2001). These resources have made it possible to carry out large-scale screening in an efficient way.

But a complete loss-of-function mutation of an essential gene always leads to a lethal phenotype that impedes further functional analysis. So a partial loss-of-function mutation, which is functional only under a permissive condition, is a favourite of yeast geneticists. A classical example is the temperature sensitive mutation, which can either be a thermosensitive (*ts*) or a cold-sensitive (*cs*) mutation. The defective protein only works at low temperatures or high temperatures, respectively. Temperature sensitive mutants can be easily identified by replica plating and culturing at different temperatures. This simple method has directly led to the finding of genes involved in cell-division-cycle (*cdc*) machinery (Hartwell, Culotti et al. 1970; Nurse 1975). In 2001, Leland H. Hartwell, Tim Hunt and Sir Paul Nurse were awarded the Nobel Prize "for their discoveries of key regulators of the cell cycle".

Temperature sensitive mutants are not just used to find the genes affected in certain processes, they also serve as a starting point for screening for genes in the same or parallel pathways. Even at permissive temperature, the activity of the mutated allele is often attenuated, though the cell might look perfectly normal. An additional mutation at another locus can sometimes cause lethality, even at the permissive temperature. This is called “synthetic lethality”.

Another broad approach to identify other members in the network is to use suppression analysis. The basic logic behind this approach is that if one mutation will cause the strain to die under a non-permissive condition, and another mutation can rescue the phenotype, then the second mutation is likely to be in a gene that is functionally linked to the first gene. Though the concept itself sounds straightforward, the mechanisms that result in the rescue can be very different, for example a mutation in a direct interaction partner, a mutation that activates an alternative pathway, or even a mutation of the tRNA molecule that recognizes the mutated codon and translates it to the right amino acid.

The suppressor assay is more attractive to yeast geneticists, because the rescue of the mutant phenotype can be selected, for example; survival under the original non-permissive conditions. In comparison, the synthetic lethality screen requires screening every clone under various conditions, which can be laborious and time-consuming work. Nevertheless, both methods are important to define genetic networks in yeast.

Yeast genetics and genomics studies continue to provide insights into the molecular mechanisms of eukaryotic cell biology. The small genome size and the limited homology with human genes can sometimes even be an advantage. The highly conserved genes are often the most fundamental components of the pathway. In the process of evolution, more regulators and modifiers are gradually added to pathways which allow more versatile and accurate control. If the phenotype of a mutation of a yeast gene can be rescued by its counterpart in a more complex eukaryotic organism, the two genes might have a similar if not the same function.

The arrival of the “genomics era” has rejuvenated studies in yeast. Interestingly, this organism serves as a platform to test high-throughput genomics techniques, such as genome sequencing and genome wide deletion libraries. All these techniques, once developed in yeast, have been quickly transferred to other organisms. Genetic screens in yeast have

provided and continue to provide informative insight into how to carry out similar experiments in more complex cells and organisms.

1.1.2 Genetic screens in fruitfly

Thomas Hunt Morgan's work in the fruitfly *Drosophila melanogaster* has been widely regarded as the beginning of the modern genetics. Considering the fact that he built his Nobel Prize winning chromosome theory of heredity on the spontaneous mutations isolated in a relatively short time (Rubin and Lewis 2000), the fruitfly is no doubt one of the most tractable multicellular organisms for genetic studies. Since then, generations of scientists have added numerous powerful tools to the fruitfly research, balancer chromosomes, deletion chromosomes, induced mitotic recombination, just to name a few.

Though flies and vertebrates diverged from a common ancestor about 700 million years ago, a lot of fundamental developmental processes are still essentially the same. *Drosophila melanogaster* has about 13,600 genes in its 180 Mb genome, much fewer than *Caenorhabditis elegans* (Adams, Celniker et al. 2000). But when both genomes are compared to human, the fruitfly has twice as many genes that have homologs in humans (Friedman and Hughes 2001). The fruitfly also has many homologs of human disease genes, which when mutated can provide insights of the molecular mechanisms of those human diseases. All these features have made *Drosophila melanogaster* an ideal system for genetic screens.

The Nobel Prize winning work by Christiane Nüsslein-Volhard and Eric Wieschaus (Nusslein-Volhard and Wieschaus 1980) changed the landscape of genetic studies completely. For the first time, a genome-wide mutagenesis was carried out in a multicellular organism to identify the mutations that would disrupt a given process, namely embryogenesis. Also for the first time, embryos rather than the adults were used for a genetic screen and many of the mutations found are actually the fundamental regulators of the whole development process (St Johnston 2002).

The famous Heidelberg screen has shown the tremendous power of a genetic screen. But like other successful screens in history, it also has its limitations. These limitations were either caused by the nature of the chosen organism, the design of the screen or the characteristics of the signaling pathway members. Since Christiane Nüsslein-Volhard and Eric Wieschaus had utilized this method to its logical extreme, most genetic screens in *Drosophila* after this were aimed at hunting for the mutations missed by the Heidelberg screen.

One efficient way to recover the missing parts of the genetic jigsaw is to find the mutations that are linked one way or another to known mutations. So enhancer and suppressor screens in a sensitized background, in which one component's function has already been partially disrupted, have become the favourite methods to expand knowledge of certain pathways and to fill the holes and the gaps. For most genes in the genome, the expression of one wild-type allele is enough for normal development. But if a second gene in the same pathway is also mutated, sometimes the level of expression of the first gene might no longer be enough to keep the signalling pathway running normally. In this way, dominant enhancers or suppressor of the first gene can be identified.

Compared to traditional screens, enhancer and suppressor screens have several important advantages. First of all, these can be an F1 screen rather than an F3 screen for recessive homozygous mutations, thus there is no need for a complicated breeding plan to make the mutations homozygous. Second, the function of an embryonic lethal mutation in later developmental stages can be studied. The most prominent example of a modifier screen was the identification of the components of the *Sevenless* (Sev) pathway, which controls the fate choice of the R7 photoreceptor cells in the eye (Simon 1994).

Another way to bypass the embryonic lethality limitation of a traditional screen is to perform a "mosaic screen". In this screen, induced mitotic recombination can be used to create homozygous mutations in a heterozygous background. When the recognition sites of the Flp recombinase, *FRT*s, are located at identical positions on homozygous chromosomes, FLP can mediate site-

specific recombination (Golic and Lindquist 1989) and make a mutation which lies distal to the *FRT* sites homozygous. The efficiency of this approach is surprisingly high, partly due to the fact that homologous chromosomes are paired in mitotic cells in *Drosophila.*

Interestingly, mosaic screens in *Drosophila* have identified not only those mutations that affect development, but also genes that can produce tumor outgrowth, which are almost impossible to find in a traditional screen (Xu, Wang et al. 1995). Because of the numerous tissue-specific FLP-expressing lines, *Drosophila* has become the only multicellular organism in which any cell type or developmental stages can be targeted for a genetic screen.

Even with all these powerful tools at hand, a lot of genes never show up in any of these loss-of-function screens. The challenge now is to annotate the remaining genes. Gain-of-function screens, which cause either over-expression of a gene in the right place or its mis-expression in the wrong tissue, can be used to identify the function of these genes (Rorth 1996; Rorth, Szabo et al. 1998).

The introduction of RNA Interference and the completion of the *Drosophila* genome sequence have now provided an unprecedented opportunity to carry out forward genetic screens (Carthew 2001). Nevertheless, in the foreseeable future, reverse genetic screens will still remain the favourite for the *Drosophila* community.

1.1.3 Genetic screens in nematode

Simplicity is one of the most important reasons why Sydney Brenner chose *Caenorhabditis elegans* as a new experimental organism to study the nervous system and embryonic development in 1963 (Ankeny 2001). But this small organism is not as simple as it looks. Surprisingly, this small creature with only 959 somatic nuclei (adult) has over 19,000 genes in its 97-Mb genome (1998), compared to the 23,000 genes of human and mouse. About one third of these genes have their counterparts in mammals.

This new model organism rapidly became quite a popular experimental system because of its unique hermaphroditic lifestyle, its rapid generation time, its simplicity and ease of manipulation. The nematode is the first multicellular model organism whose cell fate map has been fully described (Sulston 1976; Sulston and Horvitz 1977; Sulston, Schierenberg et al. 1983). From the fertilized egg to the 959-cell adult, every cell that appears or dies in the development process can be traced back to its origin. Thus, even a small change in cell number of a specific organ, like the vulva, can be used for genetic screens. The mutations recovered then build a bridge between one gene, its relevant development process and the aberrant anatomical structure caused by the mutation of this gene.

Though the worm was first picked with the view that this would enable studies of nervous system, many genetic screens have been carried out to elucidate various genetic pathways, such as apoptosis, RAS signalling, *Notch* signalling and sex determination. The early screens were mostly simple forward recessive screens that identified mutants with visible phenotypes. A typical example is Sydney Brenner's screen published in 1974 (Brenner 1974). Additional screens for the same phenotypes or modifier screens in the mutant lines have helped to identify more members of the same pathway.

A variety of mutagens have been used in *C. elegans* over the years. Ethyl methane sulphonate (EMS) is the commonly used mutagen. Mapping the point mutations caused by EMS largely depends on the available markers. *Mariner* elements from *Drosophila melanogaster* have also been used (Bessereau, Wright et al. 2001). The sequencing of the nematode genome (1998), which was the first multicellular organism to be completely sequenced, has greatly accelerated the identification of mutants. It has also enabled genome-wide RNAi screens (Kamath, Fraser et al. 2003).

The phenotype of the RNAi "knockdown" of a specific gene depend heavily on the timing and delivery method of dsRNA, the various characteristics of the target (protein stability and homology of the target gene to its family members) and the design of the dsRNA construct (Maine 2001). An example of the

limitations of RNAi is that the first genome-wide RNAi screen in *C. elegans* only identified mutant phenotypes of 1,722 genes out of 16,757 genes (86% of the 19,427 predicted genes) being targeted (Kamath, Fraser et al. 2003). So RNAi only serves as a complement, instead of a complete replacement to the traditional genetic "knockout" methods (Maine 2001). The lack of a reliable gene-targeting method to create null alleles for any given gene has limited the power of the worm genetics (Jorgensen and Mango 2002). A random deletion library and a standard method to recover the desired mutations have partly solved this problem (Jansen, Hazendonk et al. 1997; Liu, Spoerke et al. 1999).

Of the 19,000 to 20,000 *C. elegans* genes, it is estimated that only 6,000 will have a visible, lethal or sterile phenotype when mutated, and only a small portion of these have already been hit by various mutation methods (Jorgensen and Mango 2002). So the challenge now is to annotate the rest of the worm genome. Forward genetic screens are still the most powerful way of doing this, but new strategies need to be designed to screen for redundant genes. Combined with other genomics tools, such as cDNA microarrays and RNAi, genetic screens are now helping to discover the remaining secrets in the worm genome.

1.1.4 Genetic screen in mouse

Of all commonly used model organisms, the mouse is the closest relative to human. The conserved gene structures, sequence and the extensive comparative genetic linkage map make the mouse the best model to identify human gene function and provide models of human disease (Justice 2000).

Compared to the other model organisms, the history of large-scale genetic screens in mouse is relatively short. So it is no surprise that many genetic tools currently used in the mouse have already been applied to other model organisms a long time ago. Though the tools are old, the screens in mouse reveal the functions of many genes which are unique to mammals.

Interestingly, the mouse was the first multi-cellular organism in which gene targeting by homologous recombination became possible, owing to the development of mouse embryonic stem (ES) cell technology (Evans and Kaufman 1981; Bradley, Evans et al. 1984). Probably, it will also become the first multi-cellular organism in which all genes are systematically knocked-out, with the goal of creating a public resource that contains mutated alleles of every mouse genes (Austin, Battey et al. 2004; Auwerx, Avner et al. 2004).

In a model organism in which reverse genetics has been the major technology used for identifying gene function, it is important to notice the advantage of forward genetics and its role in future functional studies. In mouse, *N*-ethyl-*N*-nitrosourea (ENU) is the most commonly used mutagen for phenotype-driven screens. The point mutations induced by ENU can generate a wide range of alleles of a given gene, ranging from a complete loss-of-function null allele, a partial attenuated hypomorph to a gain-of-function allele. An allelic series provides researchers with a unique opportunity to dissect a gene's function.

It is no surprise that the first two large-scale genetic screens in the mouse were dominant screens for viable and visible phenotypes (Hrabe de Angelis, Flaswinkel et al. 2000; Nolan, Peters et al. 2000). Both screens identified a lot of visible mutations, including hair and skin, pigmentation, skeletal morphology and eye defects. Because the two different groups focused on different specialized phenotypes (the UK group on neurological phenotypes and the German group on haematological phenotypes), they identified a number of phenotypes directly linked to human disease. These two screens showed the efficacy of creating novel mutations of unknown genes by ENU mutagenesis. But they also revealed the bottleneck of this approach, the confirmation of the mutations and the subsequent identification of the point mutations (Justice 2000). Although the sequencing of the mouse genome and the creation of the mouse single nucleotide polymorphism (SNP) map have made mapping much easier, identification of the mutation is still a very laborious and time-consuming effort.

Most mutations in genes are recessive, which means a phenotype can only be observed when both alleles of a gene are disrupted. Dominant screens are thus limited to only a small subset of the 23,000 mouse genes. But recessive screens in mouse were quite difficult before the chromosome engineering techniques were developed (Ramirez-Solis, Liu et al. 1995; Zheng, Sage et al. 1999). A recent published recessive screen on mouse chromosome 11 (Kile, Hentges et al. 2003) has capitalized on a 24-cM inversion between the *Trp53* and *Wnt3* genes to isolate recessive mutations in this interval. The inversion served not only to suppress the recombination in this region, but also to simplify the genotyping by carrying a dominant *Agouti* coat colour marker.

This recessive screen has shown the tremendous power of a non-biased, phenotype-driven genetic screen that has already been proven in other model organisms. Though the cost and the time of mouse breeding has greatly limited the scale of the screens that can be carried out, the striking similarity between mice and humans has made every new mutation identified a human disease gene candidate. The fact that a single screen has tripled the number of the mutations in an already well characterised region confirms that ENU mutagenesis will still be a major player in the annotation of the mouse genome (Kile, Hentges et al. 2003).

1.1.5 Genetic screens in mammalian cells

About 40 years ago, extensive tests were carried out to test the suitable growth media for growing mammalian cells *in vitro* (Grimm 2004). This work has laid the foundation of modern cell biology. Cultured mammalian cells soon become a favourite tool for geneticists because of the difficulty to perform genetic screens at the organism level in mammals. Even today, when the development of mouse genetics and genomics tools has made large-scale genetic screens in mouse feasible, mammalian cells of various origins are still of fundamental importance to utilize the vast quantities of data generated by the genome projects.

As a model system for genetic studies, cultured mammalian cells are very similar to yeast. They can both be grown on defined media, which makes the

growth conditions precisely controllable. They can both grow quickly under “ideal” conditions, which make it easy to accumulate a lot of experimental material in a short time. They both have well-documented origins and a uniform make-up, which make the identification of mutations relatively easy. Last but not least, they both can be easily manipulated genetically, which facilitates all sorts of genetic screen designs.

Yeast was the first model used for studying eukaryotic gene functions. Genetic screens in yeast shed light on many different molecular mechanisms from signal transduction to cell-cycle control, chromosome structure to secretion (Forsburg 2001). But bioinformatic analysis of the yeast genome sequence has shown that many human genes do not have homologs in yeast (Goffeau, Barrell et al. 1996). For example, yeast does not have any real counterparts of certain cellular processes ranging from apoptosis, tissue specific differentiation to oncogenic transformation (Grimm 2004).

Similar limitations also exist in other model systems. Because of the phylogenetic distance between these model organisms and mammals, extra care needs to be taken when researchers try to interpret the exact mechanisms in human according to the results obtained from these systems. Even for the mouse, genetic screens cannot be easily carried out for some cellular alterations because of the complexity of the intrinsic and extrinsic environments. Thus, the cultured mammalian cells become the ideal substitute for the model organisms.

Because of efficient mutagenesis protocols, the ease with which genetic material can be introduced into mammalian cells, the uniform genetic make-up and the defined growth conditions, cultured mammalian cells are one of the most widely used biological systems (Grimm 2004). The basic logic of a genetic screen in cultured mammalian cells is essentially the same as a screen in other genetic systems, but it also has some clear differences. First, in contrast to the whole organisms, which constitute a wide range of different cell types, mammalian cell lines, no matter what their tissue origins are, comprise a phenotypically and genetically uniform population. Second,

mammalian cell lines are also stable under the proper culture conditions. For most cell lines, they will not enter a developmental pathway without induction. This is a very important characteristic which can be used for screens for the determinants of the developmental fate decision of a cell line. Third, mammalian cell lines are kept under defined conditions, which make the variation in external environments negligible. In more complicated systems, a simple genetic change can sometimes cause various phenotypes, and a lot of different cell types might be involved, which makes the phenotype description and dissection difficult. Thus, mammalian cells provide a reductionist model that eliminates external variation.

Cultured mammalian cells provide a simple model system to study gene function in a complex organism. This is a big advantage for studying some basic biological processes, such as cell cycle and apoptosis. The knowledge obtained from this simplified model system might not exactly reflect what has happened *in vivo*, however it does provide a good start point to extrapolate the possible functions *in vivo*.

1.2 Mouse as a genetic tool

1.2.1 Introduction

The laboratory mouse, *Mus Musculus*, has been used to study human disease throughout the last century. For a long time, the study was limited to a few visible spontaneous mutations such as *agouti*, *reeler* and *obese* (Austin, Battey et al. 2004). The work on these spontaneous mutations has provided important insights into the molecular mechanisms of the relevant human diseases. However spontaneous mutations in mice do not provide enough different mutants for genetic studies. Many different methods have been developed to generate mutants in mouse at a higher rate, including gene-trapping, ENU mutagenesis and gene targeting.

Since the gene targeting technology became a reality in ES cells in the late 1980s (Thomas and Capecchi 1987; Capecchi 1989), the mouse has played a prominent role in functional genetics and genomics studies. Compared to other model organisms, *Mus Musculus* has some unique advantages for

studying human biology and disease. As a mammal, its development, body plan, physiology, behaviour and even its diseases can be very similar to human. It is also one of the model organisms that have the highest homology to the human.

1.2.2 Similarity between human and mouse

With the completion of the human genome (Lander, Linton et al. 2001; Venter, Adams et al. 2001), the biggest challenge now is to annotate the 2.9 billion nucleotides and decode all the information. Mouse is undoubtedly the key player in the process. After about 75 million years of divergence, the genomes of mouse and human have been altered so much by evolution that there is nearly one substitution for every two nucleotides, as well as deletions, insertions, translocations and inversions (Waterston, Lindblad-Toh et al. 2002). In spite of the divergence rate, systematic genome comparisons can still identify the highly conserved regions between these genomes, which indicate functional importance. Comparative genomics also help to identify the key differences between these two organisms and elucidate the driving force shaping their genomes.

The mouse genome is 14% smaller than the human genome (2.5 Gb compared to 2.9 Gb). Over 90% of the mouse and human genomes can be partitioned into corresponding regions of conserved synteny. Approximately 40% of the human genome can be aligned to the mouse genome, which represents most of the orthologous sequences that remain in both lineages from a common ancestor. The mouse and human genomes each seem to contain about 23,000 protein-coding genes. Approximately 80% of mouse genes have at least one identifiable orthologue in the human genome. Less than 1% of the mouse genes do not have any homologue currently detectable in the human genome (Waterston, Lindblad-Toh et al. 2002).

1.2.3 Tools available for mouse genetics and genomic studies

The widespread use of the mouse for biomedical research is largely due to the development of many genetic and genomic tools. One of the landmarks in mouse genetics was the isolation of pluripotent mouse embryonic stem (ES)

cells from mouse blastocysts (Evans and Kaufman 1981) and the subsequent demonstration that cultured ES cells can transmit through the mouse germline when reintroduced into host blastocysts (Bradley, Evans et al. 1984). Importantly, cultured ES cells maintain their pluripotency after modification of their genome which allows these modifications to be established in mice. Initially, the targets for modification were random or limited to a couple of mouse genes whose disruption could be selected by drugs, such as the *Hprt* gene on the hemizygous X chromosome (Kuehn, Bradley et al. 1987; Thomas and Capecchi 1987). A more general technology was needed to allow the disruption of the genes that could not be selected *in vitro* (Goldstein 2001).

Then came the second important breakthrough. Several groups independently demonstrated that targeted mutations could be introduced into ES cells by homologous recombination (Zijlstra, Li et al. 1989; Koller, Marrack et al. 1990; McMahon and Bradley 1990; Schwartzberg, Robertson et al. 1990). This technique allowed the precise disruption of any of the 23,000 mouse genes.

This pioneering work has established a new era in mouse genetics. Precisely engineered loss- or gain-of-function mutations can be established in the mouse through *in vitro* manipulation of ES cells. These approaches, together with the transgenic technique of zygote injection, are all classified as reverse genetics. Interestingly, the laboratory mouse is the first multi-cellular animal model organism in which gene targeting by homologous recombination became possible. Reverse genetics has become the main approach to identify gene function in mouse. This situation is partly due to the ease of genetic manipulation of mouse ES cells. Another important reason is that the cost of mouse breeding makes forward genetic screens a lot more expensive using mice compared to other model organisms.

Many new genetic and genomic tools have since been developed to help to decipher the information encoded in the mouse genome. Some of these, for example Cre/*lox*P technology, chromosome engineering and induced mitotic recombination, will be reviewed in the following chapters.

1.3 Mouse embryonic stem cells as a genetic tool

1.3.1 A brief history

The foundation of the mouse embryonic stem cell (ES) technology can be traced back to the observations made on teratocarcinomas and the embryonal carcinoma (EC) cells derived from them in 1970s (Chambers and Smith 2004). Teratomas, which arise spontaneously from independent germ cells in mouse testis, contain different types of tissue derived from all the three germ layers. In some cases, the tumours also contain undifferentiated stem cells, and these malignant teratomas are thus named teratocarcinomas.

The “stemness” of the teratocarcinomas can be demonstrated by the ability of these cells to form secondary teratocarcinomas after transplantation. Undifferentiated teratocarcinoma cells can also be maintained *in vitro*. Moreover, if the EC cells are injected into blastocysts, they can sometimes be incorporated into the embryos and contribute to various cell types (Chambers and Smith 2004).

Because of their tumour origin, EC cells are mostly aneuploid, which greatly limits their ability to differentiate *in vitro*. But the unique characteristics of these pluripotent cells *in vitro* and *in vivo* raised an important question: Do they have a pluripotent counterpart in normal blastocysts that acts similarly? This presumed similarity eventually led to the isolation of the pluripotent mouse embryonic stem (ES) cells (Evans and Kaufman 1981). Initially, these cells were called “teratocarcinoma stem cells” because they share so much similarity with EC cells, such as their appearance, the culture conditions, their unlimited self-renewal and their ability to differentiate *in vivo* and *in vitro*. However, these embryo-derived cells were shown to be more stable and thus more controllable in genetic terms.

The pioneering work in ES cell biology has made it possible to target any mouse gene precisely and then transmit the mutation to the germline. The ease of manipulation of mouse ES cells has made them the “workhorse” of mouse genetics.

1.3.2 Mouse embryonic stem cells compared with the mouse

Compared to the other model organisms, the laboratory mouse can provide more accurate disease models for the human. But no model organism is perfect, and the mouse is no exception. Logistical and cost considerations that come with breeding large numbers of mice greatly limit the scale of mouse genetic studies. Even when the proposed International Knockout Mouse Project (Austin, Battey et al. 2004) and the European Mouse Genome Mutagenesis Program (Auwerx, Avner et al. 2004) becomes reality, the scale of the genetic screens will still be limited by the mouse lines that can be accommodated in a given institution. Different targeting strategies, ES cell origins and mouse strain background will all affect the phenotype observed even when a same gene is disrupted.

Another complication of a global knockout project is the fact that the disruption of many of the 23,000 mouse genes will cause embryonic lethality. Mouse embryos are covered by many layers of maternal tissues in the uterus. So observations can not easily be made without killing the pregnant female. The identification of the cause of embryonic lethality thus requires a lot of experimental analysis.

Mouse ES cells have become a key tool for mouse genetics. But their unique characteristics also mean that they can be used as an independent experimental system for studying early embryonic development. ES cells retain their unlimited self-renewal and differentiation capacity under appropriate culture conditions. They can also differentiate, both *in vitro* and *in vivo*, into almost all specialized cell types and their *in vitro* differentiation recapitulates the early embryogenesis (Wobus 2001).

The *in vitro* differentiation of ES cells has been widely studied to define the parallels with early embryonic development. Using a panel of markers representative of the early germ layers and late cell lineages, progressive differentiation of embryoid bodies (EB) has been correlated with early embryogenesis of mouse embryos (Leahy, Xiong et al. 1999). It is interesting to note that the temporal and spatial expression pattern of these markers is

strikingly similar between the EBs and the embryos. Thus *in vitro* differentiation of ES cells can serve as a system to study early lineage determination and organogenesis in mouse. These markers can be used to screen for mutations in ES cells that disrupt these processes.

1.3.3 Mouse embryonic stem cells compared with other cultured mammalian cells

Compared with other mammalian culture cell lines, mouse ES cells offer some unique advantages in cell-based screens.

First, many mammalian cell lines that are in use now, for example the human cervical carcinoma “Hela” cells and the mouse monkey embryonic kidney “Cos-7” cells, were either transformed *in vitro* or obtained from tumours to get immortalized lines. Most of these cell lines are aneuploid (Grimm 2004) and thus care needs to be taken when the data obtained from these cells is used to interpret what really happens in normal cells. In contrast, although ES cells exhibit unlimited growth in culture, they are stable in their genomic structure and they remain undifferentiated with a stable phenotype under the appropriate culture conditions. It is reasonable to argue that the biology of the ES cells more precisely reflects a normal biological and physiological status than that displayed by highly aneuploid transformed cell lines.

Second, homologous recombination in ES cells is much more efficient than that in the other somatic cell lines with the possible exception of DT40 cells. The ease of genetic manipulation allows the introduction of virtually any kinds of gene-targeting construct and reporter cassettes into almost any locus in ES cells.

Third, because most mammalian cell lines were derived from a variety of differentiated tissues, their differentiation capacity is highly limited (Grimm 2004). To identify the functions of a gene in different tissue contexts, cell lines from various differentiation stages need to be used. It is more efficient to disrupt the gene in ES cells and differentiate them into different cell types *in vitro* and *in vivo*.

It is estimated that more than 10,000 genes are expressed in ES cells (Sharov, Piao et al. 2003). Most of these genes are either the structural components or the essential players of basic processes common to all cells, for example metabolism, signalling, cell division and DNA repair. The others are genes that govern the special properties of embryonic stem cells. The second category of genes are especially attractive because the knowledge of these will have implications not only for academic research, but also for the clinical application of ES cells in cell replacement therapy (Ramalho-Santos, Yoon et al. 2002).

Several genetic screens have recently been published in which ES cells have been used identify genes in different pathways. Chambers *et al.* (2003) have used a gain-of-function approach to isolate self-renewal determinants in mouse ES cells (Chambers, Colby et al. 2003). Expression cloning was used to identify a homeodomain protein, *Nanog*, which when over-expressed can drive ES cell self-renewal without LIF. Another two groups have exploited the high rate of mitotic recombination in *Bloom*-deficient ES cells to screen for recessive mutations related to the DNA mismatch repair pathway (Guo, Wang et al. 2004) and glycosylphosphatidylinasitol-anchor biosynthesis pathway (Yusa, Horie et al. 2004).

1.3.4 Genetic and epigenetic instability of mouse embryonic stem cells

Although mouse embryonic stem cells show great potential in cell-based genetic screens, cautions need to be taken in designing screens in ES cells and also in interpreting the results of the screens because of the genetic and epigenetic instability of ES cells maintained in culture.

It was noticed that there is clear clonal variance in the efficiency of germ line transmission of ES cell clones derived even from the same parental cell line. The germline transmission ability of ES cells also decreases when the passage number increases (Nagy, Rossant et al. 1993). A possible explanation is that genetic alterations, especially those which will provide the mutant ES cells with growth advantages, can accumulate in ES cells

cultivated *in vitro*. When the passage number increases, these mutant cells will dominate the cell population because of their growth advantage, and thus interfere with the germline transmission.

When the growth rate, karyotype and the efficiency of germ line transmission are examined, it was found that chromosomal abnormality occurred rather frequently in ES cells (Nichols, Evans et al. 1990; Liu, Wu et al. 1997). A number of chromosomes can be randomly duplicated in culture, especially trisomy 8 and trisomy 11, which are directly associated a growth advantage *in vitro* and the failure of ES cells to contribute to the germ line. It is reasonable to predict that the mutant cells with trisomy 8 or trisomy 11 will also have abnormality in *in vitro* differentiation, which might interfere with genetic screens using ES cell differentiation.

Although the mutant cells with trisomies can have dramatic abnormalities both *in vivo* and *in vitro*, these cells can be distinguished by their accelerated growth rate or by karotype analysis. Subcloning of the parental cell line is an easy way to get a normal population of ES cells for further analysis.

Epigenetic instability in cultured ES cells can also impact on genetic screens using ES cells. Mouse embryonic stem cells were isolated from inner cell mass of the blastocysts (Evans and Kaufman 1981). Theoretically, they should carry the same epigenetic information as their *in vivo* counterparts. However, epigenetic state of the ES cell genome may not be stable under *in vitro* culture conditions. Epigenetic variance was observed in different ES cell lines and even in those cells derived from ES cells of the same subclone (Humpherys, Eggan et al. 2001). Epigenetic alterations at one imprinted locus did not necessarily predict changes at other loci, which suggests that the epigenetic instability of ES cells is more likely to be caused by random local loss of imprinting, instead of global increase or decrease of the methylation level in the ES cell genome.

The epigenetic variability was even found in the placentas of cloned mice derived from the same cell line (Humpherys, Eggan et al. 2001). However,

epigenetic instability of murine ES cells does not interfere with their germline transmission efficiency. Mammalian development may be rather tolerant to local epigenetic abnormalities, and unless a global loss of imprinting happens, the pluripotency of the ES cells will not be compromised. However, differentiation may be biased by altered imprinting (Mann, Gadi et al. 1990).

In summary, the unique intrinsic characteristics of ES cells have already made them a promising system to address a wide range of basic cell biology and developmental questions. The potential clinical application of ES cell biology will attract more and more researchers to use different methods to understand how the ES cells maintain self-renewal and how they differentiate into other cell types. However, care needs to be taken to monitor and control for the genetic and epigenetic status of cultured ES cells.

1.4 Cre/*lox*P site specific recombination

1.4.1 A brief history

Site-specific recombination in multicellular organisms was first achieved in *Drosophila* (Golic and Lindquist 1989). Flp recombinase from the yeast 2μ plasmid can efficiently mediate site-specific recombination between *FRT* (Flp recombinase target) sites in the fruitfly. The Flp/*FRT* system has been widely used for creating deletions, duplications, inversions and genetic mosaics.

The most widely used site-specific recombination method in mouse is based on another recombinase, Cre, although recent work shows that Flp/*FRT* works as efficiently in mouse as it does in the fruitfly. The recombinase, Cre, from the P1 bacteriophage belongs to the integrase family of site-specific recombinases (Hamilton and Abremski 1984). Cre can catalyze the recombination between two *lox*P sites. The *lox*P site is a 34-bp consensus sequence, which includes two inverted 13-bp flanking sequences on both sides of an 8-bp core spacer sequence. The core spacer decides the orientation of the *lox*P site, but the flanking sequences are the actual binding site of Cre.

The Cre/*lox*P site-specific recombination system was first shown to work in mammalian cells in the late 1980s (Sauer and Henderson 1988). In the early 1990s, this system was shown to work in mouse when Cre was expressed *in vivo* (Lakso, Sauer et al. 1992; Orban, Chui et al. 1992; Gu, Zou et al. 1993). The Cre/*lox*P system has been widely used in mouse genetics, combined with the gene targeting and transgenic technology, the Cre/*loxP* system has made it easy for mouse geneticists to tailor the mouse genome almost without any limitations, from one-base-pair point mutations to mega-base-level deletions, inversions, duplications and translocations.

1.4.2 Basic Characteristics of Cre/*lox*P system

The 34-bp *lox*P site is short enough to be put into large introns without disrupting the transcription of the gene. It is also long enough to avoid the random occurrence of intrinsic *loxP* site in the mouse genome. With the completion of the sequencing of several major model organisms, searches reveal that no perfectly matched *lox*P site has even been found in any organisms other than the P1 bacteriophage. It has been noted that some pseudo recognition sites exist in the mouse genome but the efficiency of recombination between wild type *lox*P sites and these pseudo sites has not been thoroughly studied.

In vitro, Cre-mediated recombination is efficient enough to excise genomic regions as large as 400 kb, and recombinants can be identified without selection (Nagy 2000). The Cre recombinase is also very efficient *in vivo*. Numerous Cre transgenic lines have been established in the last decade to facilitate efficient Cre-mediated excision in a lot of different developmental stages and different cell types. One aim is to generate more Cre transgenic lines to cover all development processes and cell types (Nagy 2000). A resource like this will greatly help the study of gene function *in vivo*, especially for the genes that cause lethality at early stages when disrupted.

1.4.3 Application of Cre/*lox*P system in mouse genetics

1.4.3.1 Conditional gene knock-out

One of the common uses of the Cre/*lox*P site-specific recombination is for conditional gene knockouts. The logic behind this powerful tool is simple: two *loxP* sites in the same orientation are placed on both sides of the most important functional domain of the gene of interest. Since the *lox*P site is only 34 bp, usually it will not affect the gene transcription if it is placed in the non-conserved region of an intron. The targeted ES cells and the animals containing such an allele are perfectly normal compared to wild type animals. But when the animals are crossed to a Cre-expressing transgenic line, the progeny that carry both the Cre transgene and the *lox*P-flanked allele will excise the *lox*P-flanked portion of the gene in the cells that express Cre (Tsien, Chen et al. 1996).

Two main issues with the conditional knockout approach are the design of the conditional targeting construct and the specificity and efficiency of the Cre line. When a conditional targeting vector is designed, the region that is selected to be flanked by *lox*P sites needs to be important enough to disrupt the gene function completely when excised. The flanked region also needs to be small enough for the two *lox*P sites to be introduced into ES cells in one targeting step. The availability of restriction enzyme sites, the size of the genomic insert that a vector can incorporate and the subsequent genotyping strategy all limit the choice of the position of *lox*P sites. The development of *E. coli* recombineering technology recently has greatly simplified the method and allowed the flexibility of design of conditional targeting vector (Copeland, Jenkins et al. 2001).

Although many Cre transgenic lines have been generated in the last decade, they are still not enough to satisfy the increasing need (an incomplete list of available Cre excision lines can be found on the webpage of Dr. Andras Nagy's lab http://www.mshri.on.ca/nagy). Even for existing Cre lines, leaky expression of Cre in the wrong cell type and/or developmental stage, or incomplete excision in target cells makes the interpretation of the phenotypes difficult, or can result in no phenotype at all, for example in a mosaic tissue.

In spite of these problems, the conditional gene knockout technology, which combines the strength of gene targeting and site-specific recombination, is still a powerful tool for the mouse geneticist, which will play an increasingly important role in functional genetic studies.

1.4.3.2 Selectable marker removal and recycling

It has been reported that in some cases, the selection cassette used for gene targeting will affect the expression of genes nearby *in vivo*. An easy way to circumvent this problem is to flank the selection marker with two *lox*P sites and "pop out" the cassette either *in vivo* or *in vitro*. This has already become a routine procedure for both traditional and conditional targeting.

Another advantage for marker removal is that the same selection marker can be reused in the subsequent manipulations of the ES cells. For studying gene function *in vitro*, especially for genetic screens, multiple gene targeting events are usually required to disrupt a number of loci or introduce reporter cassettes. If the usable markers are exhausted, this will limit downstream analysis. Positive-negative selection marker flanked by *lox*P sites can be used to generate mouse ES cells that carry multiple targeted mutations but devoid of any exogenous markers (Abuin and Bradley 1996).

1.4.3.3 Subtle change and hypomorphic alleles

Another important application of the Cre/*lox*P system is to create subtle mutations. For most of the mouse knockout lines published so far, either an important domain of a gene or even the whole gene has been deleted. This approach is more likely to create a null allele, but considering the fact that many human hereditary diseases are caused by point mutations, small deletions and small insertions, null alleles in mice might not generate an ideal model for their relevant human disease. Also, hypomorphic alleles, which partially disrupt gene function, are sometimes more useful for genetic screens.

A typical way to introduce subtle genetic changes into a gene in mouse ES cells is to make the change in one of the homology arms of the targeting vector, and include a selection marker flanked by *lox*P sites in a non-coding region. After subtle mutation has been confirmed to have been incorporated, the selection marker can be popped out either *in vitro* or *in vivo*, leaving only the small change (Nagy, Moens et al. 1998).

1.4.3.4 Chromosome rearrangement

Chromosome rearrangements happen spontaneously in almost all the eukaryotic species. They play a very important role in evolution, but in humans, they are also one of the most common causes of foetal losses, developmental disorders and cancer (Yu and Bradley 2001). Thus, engineered chromosome rearrangements in mouse generated by long range Cre/*lox*P recombination can be used to model their human counterparts and investigate the molecular mechanisms underlying a variety of different human genetic disorders.

Chromosome deletions and inversions are also useful tools for performing recessive genetic screens. Large deletions can reduce the diploid genome to areas of segmental haploidy, which allows F1 screens for recessive mutations in the deletion region (Ramirez-Solis, Liu et al. 1995). On the other hand, inversions, which serve as balancer chromosomes, can be used to maintain lethal recessive mutations in the inversion interval (Zheng, Sage et al. 1999). Creating a resource of inversions and deletions throughout the mouse genome will be important for large-scale phenotype-driven mutagenesis programs.

1.5 Chromosome engineering

1.5.1 A brief history

Chromosome engineering has its origin in *Drosophila* genetics. Spontaneous chromosome rearrangements were found and mapped by observing fruitfly salivary gland polytene chromosomes under the microscope. These rearrangements are very useful tools in genetic studies. For example, inversions could be used to maintain lethal mutations without selection.

In 1927, Muller showed that ionizing radiation could induce different kinds of genetic damage, including chromosomal rearrangements (Rubin and Lewis 2000). He was awarded the Nobel Prize in 1946 for this finding. This method has been exploited to its logical extremes by D. L. Lindsley and his colleagues in 1972 to generate an ordered set of duplications and deletions spanning the major autosomes (Lindsley, Sandler et al. 1972). This effort leads to the collection of deletion lines that provide maximal coverage of the genome in a minimum number of stocks held in the Blooming Stock Center (St Johnston 2002). This resource provides ideal starting material for region-specific mutagenesis screens. It also provides a rapid way to map recessive mutations found in genetic screens.

Though X-rays are also very efficient in inducing genetic damage in other species, a similar genome-wide chromosomal rearrangement resource is not available in any other multi-cellular organisms. This is partly due to one unique characteristic of *Drosophila*, the salivary gland polytene chromosomes, which have made the physical mapping of chromosome rearrangements significantly easier than in other species. The lack of efficient methods to determine the endpoints of rearrangements has made it hard to replicate this genome-wide resource in other species.

1.5.2 Engineering mouse chromosome with Cre/*lox*P

Spontaneous chromosome arrangements are very rare in nature, and even if they do happen, there are practical difficulties in recovering them. The same problem also exists in the arrangements induced by radiation and other chemical mutagens (Yu and Bradley 2001). Clearly, induced arrangements with pre-determined end points will be more useful for genetic studies. The Cre/*lox*P site-specific recombination system is now the most commonly used method to generate these rearrangements in mouse.

1.5.2.1 Deletions, duplications and inversions

Deletions, duplications, translocations and inversions with pre-determined end points are useful not only for creating human disease models, but also for making genetic tools for functional genomics studies.

If large genomic regions are involved, an efficient selection strategy is needed to identify the ES clones that carry the desired genomic re-arrangement. A common way to achieve this is to put the *loxP* sites into two non-functional halves of the hypoxanthine phosphoribosyl transferase (*Hprt*) mini-gene. Then the two halves are sequentially targeted to two pre-determined end points of *Hprt*-deficient ES cells. Transient Cre expression induces the site-specific recombination and restores the activity of *Hprt* mini-gene. HAT selection can directly select the clones with the desired chromosomal rearrangements (Ramirez-Solis, Liu et al. 1995; Smith, De Sousa et al. 1995).

The bottleneck for making the targeting vectors used for chromosome engineering is to isolate the end-point genomic fragments, which normally requires laborious genomic library screening. A two-library system of pre-made targeting vectors has greatly simplified the procedure (Zheng, Mills et al. 1999). Also, by incorporating the coat colour markers into the vector backbone, the mice that carry the chromosome rearrangements can be genotyped easily by eye.

To evaluate the efficiency of Cre-mediated recombination over long distances, Zheng et al. (2000) created a series of deletions, duplication and inversions on mouse chromosome 11 and compared their relative efficiency *in vitro* (Zheng, Sage et al. 2000). It has been shown that, although the site-specific recombination efficiency decreases with increasing distances, rearrangements as large as three quarters of chromosome 11 can be achieved with a proper selection strategy. The only limitation seems to be the haploinsufficiency that comes with large deletions (Liu, Zhang et al. 1998; Zheng, Sage et al. 2000). Although the recombination can still occur over these distances, HAT resistant clones are only recovered if they duplicate the wild type chromosome to compensate the loss caused by the large deletion.

1.5.2.2 Nested deletions

For human genetic disorders that are caused by spontaneous deletions, a common way to map the disease gene is to identify the end points of the deletions in a collection of patients. Since the deletions are often of different sizes and with different end points, a minimum overlapping region can be defined and used to locate the disease gene(s). The availability of the genetic material is limited, thus it is not always possible to identify the relevant gene(s). Instead, a key region and several candidate genes will often be suggested for future studies. If the minimum overlapping region is relatively small and only a few genes are involved, it is possible to knockout these genes one by one and analyse the mouse phenotype to identify the disease gene. But if the range is too big, additional steps are needed to further reduce the size of the region.

The conserved synteny between human and mouse can be used to define the region corresponding to the deletion in the human genome. Nested deletions can then be constructed to map the disease gene (Yu and Bradley 2001). First, a 5' *hprt*-*lox*P cassette was targeted to a predetermined locus to serve as an anchor point. The *lox*P-3' *hprt* cassette was then randomly integrated into the ES cell genome by retroviral infection. ES cells with nested deletions can be isolated by transient expression of Cre and subsequent HAT selection (Su, Wang et al. 2000).

In this strategy, the introduction of a second *lox*P site is random, thus only a small subset of the viral integrations will occur on the same chromosome as the fist anchor *lox*P site. The strategy to recover these rare events from the pool of random insertions in the ES genome is based on two key observations of Cre/*lox*P recombination efficiency. First, Cre efficiency does not change appreciably over the range of several megabases (Zheng, Sage et al. 2000). So this predicts that small deletions will not be generated any more efficiently than large ones. Second, Cre-mediated recombination within a few megabases on the same chromosome (*cis*) is two-three orders of magnitude more efficient than recombination between two *lox*P sites located on

homologous chromosomes (*trans*) (Ramirez-Solis, Liu et al. 1995; Liu, Zhang et al. 1998). The recombination efficiency between *lox*P sites on non-homologous chromosomes is even lower (Zheng, Sage et al. 2000).

Taking these two observations together, it is reasonable to predict that in from a pool of random insertions, the HAT-resistant clones recovered after exposure to Cre should have deletions within several megabases of the anchor point. This prediction has been proved to be true, a nested set of deletions were generated on the mouse chromosome X and chromosome 11 (Su, Wang et al. 2000). Most deletions recovered in this report were mapped within 1 cM distal or proximal to *Hprt* or E_2DH. It was also noticed that although the deletions proximal to *Hsd17b1*, the chromosome 11 anchor point, could be transmitted through the germline and maintained as heterozygotes, the deletions distal to *Hsd17b1* could not be transmitted, which might be due to the severe haploinsufficiency reported in that region (Liu, Zhang et al. 1998).

The nested deletion strategy has been successfully used to map *Tbx1*, the gene responsible for the DiGeorge Syndrome (Lindsay, Vitelli et al. 2001). The key deletion region in human encompasses around 1 Mb, and contains at least 15 genes. Since human deletion map was not helpful to determine the causal gene, several deletions were made in mouse ES cells by both traditional chromosome engineering and nested deletion methods. Some of the deletion mouse lines showed the typical clinical phenotype, while others did not. By comparing these phenotypes with the deletion regions, the critical region was reduced to 200 kb between *T10* and *Cdcrel1* genes. PAC transgenesis was then used to rescue the phenotype in these deletion lines and finally identify the disease gene as *Tbx1*.

1.5.2.3 Regional trapping

The nested deletion strategy described by Su et al. (2000) was shown to be an efficient way to create a series of deficiencies in a region of interest. But like other deletion strategies, a large portion of the deletions generated cannot be transmitted into the germline, greatly limiting the use of this strategy as a

convenient way to mutate genes (Liu, Zhang et al. 1998; Su, Wang et al. 2000; Zheng, Sage et al. 2000).

On the other hand, the success of the nested deletion strategy demonstrated that retrovirus integrations in a small region around an anchor point could be recovered using an appropriate selection procedure *in vitro*. If retroviral integrations occur in the coding region of a gene, then a variation of this strategy could be utilized to accumulate mutations around an anchor point. This has led to the idea of “regional trapping”, which use a gene-trapping virus to disrupt genes and then use site-specific recombination to select for trapping events that occur in the desired region (Wentland, unpublished data).

Cre recombination occurs efficiently between *lox*P sites on the same chromosome (*cis*) (Liu, Zhang et al. 1998; Zheng, Sage et al. 2000). For a given physical distance, the efficiency of recovery of inversions is higher than deletions because inversions do not result in the loss of any genetic material and thus ES cells with inversions do not suffer growth disadvantages compared with cell lines with deletions (Zheng, Sage et al. 2000). Theoretically, an inversion should only disrupt the genes close to the two end points. If the phenotype of the anchor point in the regional trapping is already known, any new phenotype of the inversion is most likely related to the other breakpoint, although it is possible that an inversion can sometimes disrupt long-range regulatory elements.

A strategy similar to that described for making nested deletions has been used for regional trapping. In this strategy, the anchor point was introduced into ES cells by targeting the 5’ *Hprt-lox*P cassette into the anchor point locus, *Hsd17b1*. The *loxP*-3’*Hprt* cassette was then randomly integrated into the ES cell genome by retroviral infection using a gene-trapping retroviral vector. Gene-trap insertions on chromosome 11 can be selected by transient expression of Cre which induces an inversion, cells with recombination events can be selected in HAT (Wentland, unpublished data).

In this strategy, gene-trap insertions along the whole length of the chromosome have been recovered. As predicted, about 86% of the gene-traps were concentrated on the distal part of chromosome 11, and fall within a 43 Mb region surrounding the *Hsd17b1* locus. The largest confirmed inversion was 82 Mb (Wentland, unpublished data). The range over which this strategy is effective was much larger than the range of deletions achieved by Su et al. (2000). Since the same anchor point was used for both experiments, the data shows that cells with inversions are more likely to be viable than cells with deletions of similar size.

It is interesting to note that in this study, a large portion of trapped loci were neither predicted by *in-silico* gene prediction, nor supported by any EST sequences. Some of these genes were verified to express in the embryo or adult tissues by RT-PCR, but few of them were expressed in ES cells. This data demonstrates the use of 3' trapping to mutate genes that do not normally express in undifferentiated ES cells (Wentland, unpublished data).

The "genes" which do not appear to be expressed might have been recovered with the 3' gene-trap selection because of pseudo splice acceptors and polyadenylation signals scattered in the genome. Some of the 3' RACE products matched repeats such as retrotransposons and SINEs. Though approximately 37.5% of the mouse genome is comprised of these repeat elements, this did not cause any serious background problems.

Regional trapping disrupts a trapped gene by moving the part of the gene upstream of the proviral integration site away from the downstream part. So this method should be more mutagenic than traditional trapping. But inversions have the potential complication of disrupting genes nearby whose coding region or transcriptional regulatory regions overlap or fall on either end of the trapped gene. Thus expression of the genes around the breakpoint needs to be checked to avoid misinterpretation of the phenotype (Wentland, unpublished data).

The regional trapping strategy is potentially useful for finding disease genes in a given region. Because of the conserved synteny between the human and mouse genomes, it is easy to find a mouse genomic region which corresponds to a human disease candidate region. By regional trapping, it is possible to mutate a large portion of the genes in that region in a quick and efficient way, facilitating regional screening of disease genes in the mouse genome (Wentland, unpublished data).

1.6 Induced mitotic recombination

1.6.1 A brief history

Chimaeras are individuals that are formed from cells of different origins, and genetic mosaics refer to the individuals that contains cells of different genotype but of the same origin (Rossant and Spence 1998). Mosaics are particularly important for the study of cell lineage, cell fate determination and cell-cell interactions. Some human diseases are caused by somatic loss of heterozygousity (LOH) caused by mitotic recombination or chromosome loss and re-duplication. Genetic mosaics can serve as models for these types of human diseases.

If mitotic recombination occurs during the G2 phase of the cell cycle and the recombinant chromatids segregate to different daughter cells (G2-X), the daughter cells will have either two copies of the paternal chromosome or two copies of the maternal chromosome. If the recombinant chromatids segregate to the same daughter cells, the daughter cells will still have one copy of maternal chromosome and one copy paternal chromosome (G2-Z). If the mitotic recombination happens during G1 phase of the cell cycle, this is genetically neutral as all the daughter cells will be the same as the parental cell (Golic 1991).

Historically, mosaics have been widely used to address many different developmental questions (Xu and Harrison 1994). Since the spontaneous mitotic recombination rate is too low for any practical use in genetic studies, various methods have been developed to induce mitotic recombination. X-ray irradiation is the most frequently used method to achieve this purpose. But the

shortcomings of this method greatly limit its usage as a genetic tool. First, the dosage required to induce mitotic recombination limits experimental use. High dosage of ionizing irradiation causes excessive cell death, whilst a low dosage is inefficient at inducing mitotic recombination. Second, the mitotic recombination induced by X-rays occurs randomly in the genome, so unless a genetic marker which is visible at the cellular level (such as the eye colour markers in the fruitfly and the coat colour markers in the mouse) is located near the allele of interest, it is impossible to distinguish the mosaic clones from their background.

1.6.2 Induced mitotic recombination in fruitfly

After the Flp/*FRT* system was shown to be able to mediate site-specific recombination in *Drosophila* (Golic and Lindquist 1989), this system was quickly applied to generate mitotic recombination clones by creating flies with transgenic FRT sites at the same position on homologous chromosomes (Golic 1991). Xu et al. (1993) have constructed a series of *FRT* transgenic lines to cover all the four *Drosophila* chromosomes. This resource has made it possible to create mosaic animals for 95% of *Drosophila* genes. Each of these lines carries not only an *FRT* site close to the centromere of one of the chromosomes, but also a cell-autonomous marker distal to the *FRT* site. When these *FRT* lines are crossed to a Flp recombinase transgenic line, the marker can be used to distinguish cells of different genotypes (Xu and Rubin 1993).

This method has been successfully used to screen for mutations that produce tumorous outgrowth in the imaginal discs (Xu, Wang et al. 1995). A tumour suppressor gene, *large tumor suppressor (lats)*, which encodes a protein kinase, has been discovered in this screen. This example clearly shows the usefulness of induced mitotic recombination. The *lats* gene was found to cause a wide range of defects throughout development. All the alleles of *lats* were also found to be lethal at different stages (Xu, Wang et al. 1995). So it would have been almost impossible for a traditional recessive screen to identify the tumor-suppressor function of *lats* in the adult.

Induced mitotic recombination is very efficient in *Drosophila*, largely due to its chromosome structural and mechanical characteristics (Liu, Jenkins et al. 2002). For example, in *Drosophila*, nearly two-thirds of the mitotic recombination events are G2-X, and G1 recombination accounts for the remaining one-third, whereas G2-Z events are very infrequent (Golic 1991; Xu, Wang et al. 1995). Also in *Drosophila*, the homologous chromosomes are paired in somatic cells, which makes mitotic recombination more likely to happen.

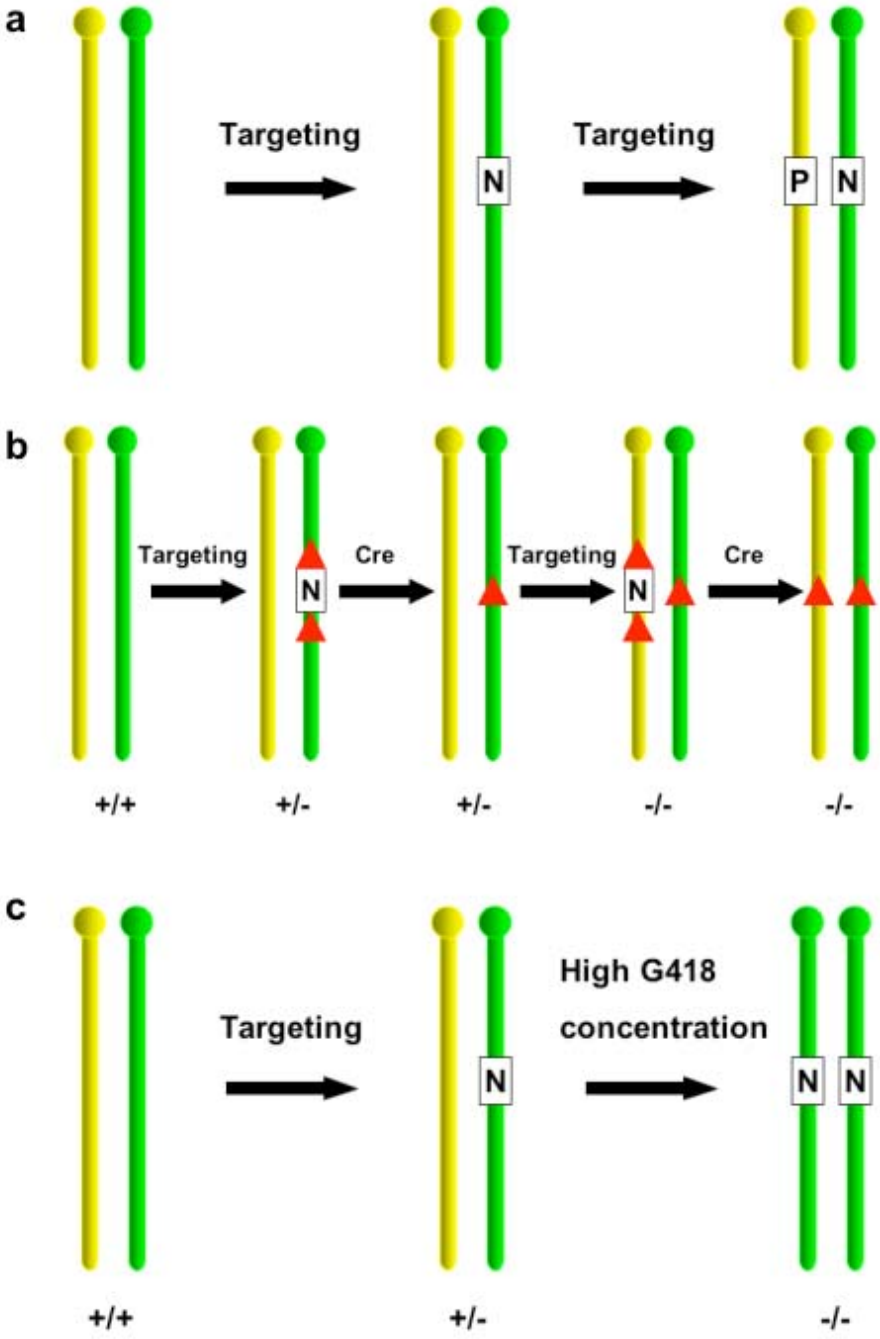
One disadvantage of the induced mitotic recombination approach is that it can only be used to screen for genes distal to the *FRT* site. Most *FRT* lines were created by random integration of a dominantly marked *FRT*-containing P element and FISH was then used to map the integration sites (Xu and Rubin 1993). Using this approach, it is hard to find integrations that are located very close to the centromeres. Also, only chromosome-specific screens can be carried out using this method. Separate screens for each of the five arms are required to cover the entire *Drosophila* genome (St Johnston 2002).

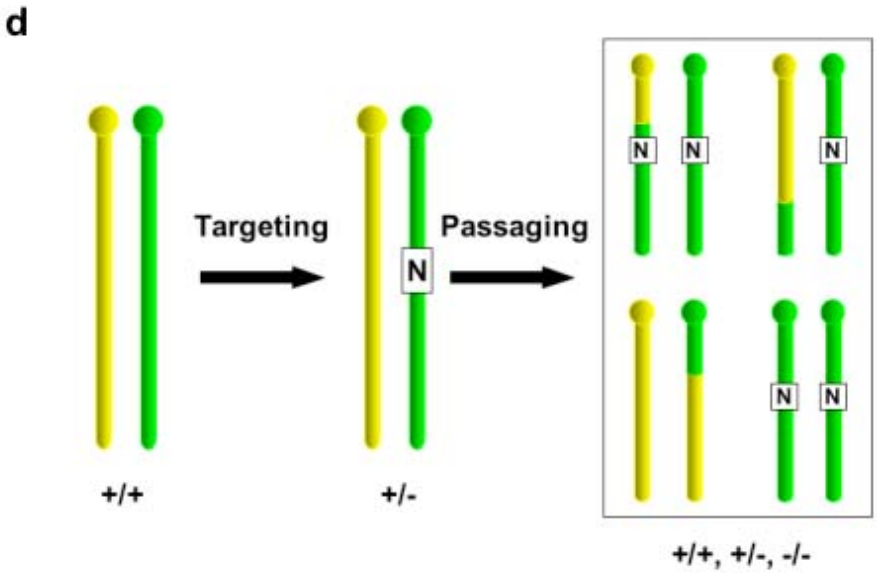
1.6.3 Traditional methods to generate homozygous mutations in mouse embryonic stem cells

Homozygously mutated ES cells are important resources to study gene function *in vitro*. To make a homozygous mutant ES cell line is a complicated process that normally requires several targeting steps and a screening strategy to identify the correctly targeted clones.

1.6.3.1 Sequential gene-targeting

The most frequently used method to create loss-of-function mutations is to sequentially target both alleles of a gene in ES cells. Two different drug selection markers are needed for the two targeting events (Fig. 1-1a) (te Riele, Maandag et al. 1990). An alternative way is to flank the drug selection cassette with two *lox*P sites and remove the selection marker after targeting by Cre-mediated recombination (Fig. 1-1b). The same targeting vector can then be reused to target the second allele (Abuin and Bradley 1996).


a
Targeting
N
Targeting
P
N
b
Targeting
N
Cre
Targeting
N
Cre
+/+
+/-
+/-
-/-
-/-
c
Targeting
N
High G418
concentration
N
N
+/+
+/-
-/-


d
Targeting
Passaging
N
+/+
+/-
+/+, +/-, -/-

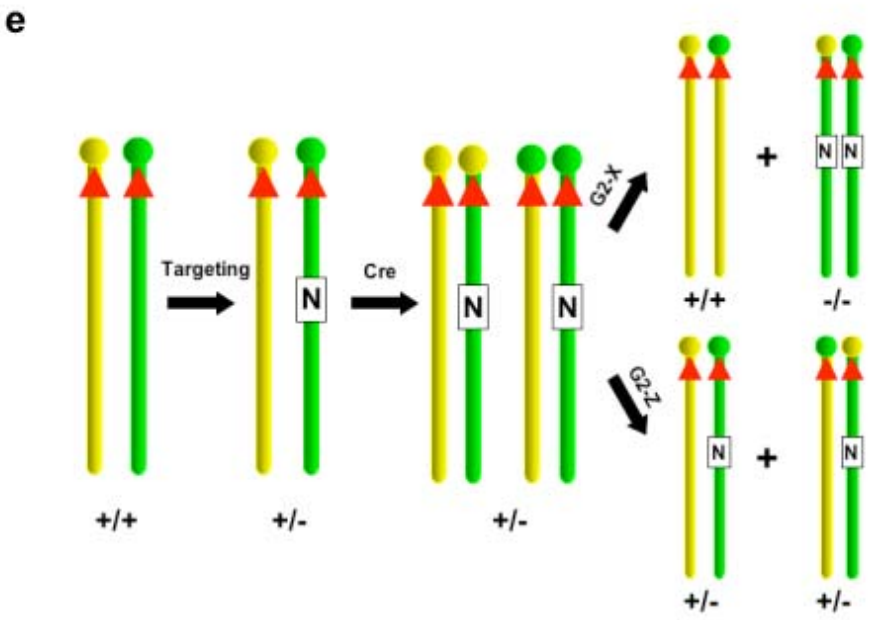

e
Targeting
Cre
N
G2-X
G2-Z
+
+/+
-/-
+/-
+/-
+/-
+/-

Fig. 1-1 Methods to generate homozygous mutations in ES cells. a. Sequential gene targeting using two different selection cassettes. Two vectors with different selection markers are used to target both alleles of a genomic locus. The two selection markers cannot be removed after targeting. **b.** Sequential gene targeting using *lox*P flanked selection cassettes. After the first allele of a genomic locus is targeted by a *lox*P flanked cassette, Cre-mediated site-specific recombination can be used to remove the cassette. The same vector can be used to target the second allele. This approach can be used to generate marker-free homozygously mutated ES cell lines. **c.** LOH induced by a high G418 concentration. After the first allele of a genomic locus is targeted by a *neo* cassette, high G418 concentration can induce the loss of the un-targeted chromosome and the subsequent duplication of the targeted one. **d.** Elevated mitotic recombination in *Blm*-deficient ES cells. The *Blm*-deficiency in mouse ES cells caused a 20-fold increase in the rate of LOH compared to wild type cells. A single allele mutation will become homozygous randomly in this genetic background. But the resulting pool of cells will be a mixture of heterozygous, homozygous and wild type clones. **e.** Induced mitotic recombination. Cre-mediated site-specific recombination can be used to induce mitotic recombination between two *lox*P sites targeted to two homologous chromosomes. Depending on the segregation pattern following the mitotic recombination, some of the daughter cells will become homozygous for the targeted locus. Yellow and green bars: homologous chromosomes; red arrow, *lox*P site; N, *neo* selection marker; P, *puro* selection marker.

The advantage for this approach is that gene targeting allows precise disruption of the gene of interest. But it also has its limitations. For targeting both alleles using different drug selection markers, the same markers cannot be reused for future steps. The selection cassette recycling approach requires an additional step of popping out the selection marker. Both methods are time-consuming and hard to scale up, so they can only be applied on a gene-by-gene basis.

1.6.3.2 High concentration G418 selection

Homozygous mutated cells also occur spontaneously from cultured mammalian somatic cells containing a heterozygous mutation, a process known as loss of heterozygosity (LOH). LOH can occur by many mechanisms including regional or whole chromosome loss, mitotic recombination and gene inactivation. But a selection strategy is needed to identify these rare events.

It has been shown that when heterozygous cells targeted with a Neomycin (*Neo*) drug resistance cassette are grown in high concentrations of G418, many of the surviving cells are homozygous for the targeted allele (Mortensen, Conner et al. 1992). This strategy provides an easy way to generate ES cells in which both alleles have been targeted. The existence of two copies of the *Neo* cassette in these cells suggests that LOH has occurred either by mitotic recombination between homologous non-sister chromatids, by chromosomal loss followed by chromosomal duplication (Fig. 1-1c) or by a local gene conversion event.

To investigate the mechanism of LOH in ES cells by high concentration G418 selection, the *Neo* cassette was targeted into six different genomic loci on four different chromosomes of a hybrid ES cell line (R1) (Lefebvre, Dionne et al. 2001). The use of a hybrid cell line allows the origin of the two homologous chromosomes to be tracked by analyzing polymorphic DNA markers. In this study, it was shown that all of the homozygous gene-targeted clones recovered by high concentration G418 selection had lost heterozygosity, not only at the targeted locus, but also at the distant linked markers. Thus LOH selected by high concentration G418 selection involves either chromosomal

loss and subsequent duplication, or mitotic recombination proximal to the locus targeted with the selection marker.

Compared to the sequential targeting method, this high G418 concentration selection approach only requires one step of gene targeting, therefore, providing a convenient way to generate homozygous mutations in ES cells. The possible mechanism underlying LOH suggests that any randomly induced mutation that lies on the same chromosome as the pre-targeted *Neo* cassette can also become homozygous under high concentrations of G418 selection. Combined with other mutagenesis methods, high G418 concentration induced LOH can be used to generate homozygous mutation in a chromosome-specific way.

However this method also has its limitations. If it is used to make targeted mutations homozygous, it will still require designing targeting vectors and probes to generate and genotype the mutations. If this method is used to make random mutations homozygous (ENU or gene-trap mutations), it will be difficult to determine the genotype of the mutated locus by just checking the genotype of the *Neo* cassette targeted locus because the range of the LOH can be different from clone to clone.

1.6.3.3 Elevated mitotic recombination in *BLM*-deficient cells

Mitotic recombination can occur spontaneously, but its efficiency is too low to be used as an efficient tool for generating homozygous mutations without strong selection. Recently, it was shown that the mitotic recombination rate can be increased in mouse ES cells that lack the function of a DNA helicase, *Blm* (Luo, Santoro et al. 2000).

Six different *Blm* knockout alleles have been published so far (Chester, Kuo et al. 1998; Luo, Santoro et al. 2000; Goss, Risinger et al. 2002; McDaniel, Chester et al. 2003). Four of these were generated by gene targeting with replacement targeting vectors(Chester, Kuo et al. 1998; Luo, Santoro et al. 2000; Goss, Risinger et al. 2002; McDaniel, Chester et al. 2003). These four alleles deleted one or more coding exons of the *Blm* gene and all of them

have been described as embryonic lethal. The other two alleles, *Blm*^{tm2Brd} and *Blm*^{tm3Brd}, are the products of an insertional gene-targeting event, which results in the duplication of exon 3 (Luo, Santoro et al. 2000). This duplication caused a frame-shift mutation. Interestingly, the *Blm*^{tm2Brd} allele is homozygous lethal but the derived *Blm*^{tm3Brd} is viable. The homozygous mice (*Blm*^{tm2Brd}/*Blm*^{tm3Brd}) exhibited genomic instability and tumor susceptibility, a phenotype mimicking the human Bloom's syndrome.

Luo et al. (2000) measured the LOH rate in *Blm*-deficient ES cells by targeting *Hprt* minigene into an autosomal genomic locus. Cells that lose the *Hprt* minigene by LOH become resistant to 6-thioguanine. By Luria-Delbruck fluctuation analysis, the rate of LOH in *Blm*-deficient ES cells was determined to be 4.2×10^{-4} (events/locus/cell/generation), compared to 2.3×10^{-5} (events/locus/cell/generation) in wild type ES cells (Luo, Santoro et al. 2000; Liu, Jenkins et al. 2002).

The *Blm* gene product is not required for cell growth or survival in culture. *Blm*-deficiency in mouse ES cells caused a 20-fold increase in the rate of LOH, which provides the basis for generating homozygous autosomal mutations from single allele mutations. By calculation, a single ES cell with a heterozygous autosomal mutation will have segregated at least one daughter cell with a homozygous mutation by the time the colony derived from this cell contains 2,000 to 5,000 cells (Guo, Wang et al. 2004). So theoretically, when an ES cell library of heterozygous mutations is expanded for more than 13 generations, the library will contain a genome-wide set of homozygous mutations (Fig. 1-1d).

Recently, two groups have used *Blm*-deficient cells to screen for recessive mutations related to the DNA mismatch repair pathway (Guo, Wang et al. 2004) and glycosylphosphatidylinasitol-anchor biosynthesis pathway (Yusa, Horie et al. 2004). The success of the two recessive genetic screens in the *Blm*-deficient ES cells has shown the utility of this system for generating genome-wide homozygous mutations that facilitate recessive genetic screens *in vitro*.

This method does not require gene targeting of the first allele for generating the homozygous alleles. So it can easily be used in combination with other large-scale mutagenesis methods, such an insertional gene-trap mutagenesis or chemical mutagenesis. However, the *Blm* system can only generate a mixture of heterozygously and homozygously mutated ES cell clones. The representation of any one particular mutation in the pool will be extremely rare (approximately 10^{-7}-10^{-8}), so the mixture can only be used when there is a strategy available to selectively isolate the clone of interest from the rest of the ES cell population. For a genetic screen without selection, mutants need to be examined one-by-one to see whether or to what extent the desired phenotype is present (Grimm 2004). So pure homozygously mutated clones are needed, *Blm*-deficient ES cells are obviously not suitable for this purpose.

1.6.4 Induced mitotic recombination in mouse

Flp/*FRT* induced mitotic recombination has provided an efficient way to generate genetic mosaics in *Drosophila*. Induced mitotic recombination makes it possible to perform F1 mosaic screens, which save the trouble of performing three generations of crosses to establish individual lines to identify potential mutants. Screening for mutations in mosaic animals also circumvents the limitation of embryonic lethality of homozygous animals, especially for the genes that might have multiple functions at different developmental stages (Theodosiou and Xu 1998).

Recently, Liu et al. (2002) have successfully developed this system for the mouse. In this study, mitotic recombination was induced in mouse ES cells via Cre-mediated recombination between targeted *lox*P sites (Fig. 1-1e). The mitotic recombination frequency varied between different genomic loci and chromosomes, ranging from 10^{-5} to 10^{-2} after transient Cre expression. However, four of five loci tested showed a relatively low frequency, ranging from 4.2 X 10^{-5} (*Snrpn*) to 5.1 X 10^{-4} (*Wnt3*) for single allelic *lox*P sites after transient expression of Cre. Even for the clones in which induced mitotic recombination did occur in G2, not all of the events were followed by X segregation. For example, only 60% of recombination events at the *D11Mit71*

locus and 23% of the events at *Snrpn* locus was a G2-X event. One explanation for the low recombination frequency and the low proportion of G2-X events compared to *Drosophila* is that homologous mouse chromosomes are not paired in the interphase.

It was also noted that the fifth locus in the study, *D7Mit178* has exceptionally high rates of induced mitotic recombination (7.0×10^{-3}). Also, for this locus, all the recombination events seemed to occur at the G2 phase and followed by X segregation (Liu, Jenkins et al. 2002). The variation of the mitotic recombination frequency and the proportion of G2-X events from one chromosome to another might be caused by the different levels of the association between homologous mouse chromosomes in interphase. It is likely that for some chromosomes, some regions may be closely associated during the S-G2 phase of the cell cycle, and this greatly promotes the recombination efficiency and the chance of G2-X segregation (Liu, Jenkins et al. 2002).

In Liu's study, it was shown that multiple allelic *lox*P sites could increase the efficiency of induced mitotic recombination. For the *D7Mit178* locus, the increase was more than seven fold (from 7.0×10^{-3} to 5.0×10^{-2}), but the proportion of G2-X segregation among the recombination events dropped from 100% to 65% (Liu, Jenkins et al. 2002).

The frequency of inducible mitotic recombination on mouse chromosome 11 is similar to the spontaneous frequency of LOH on chromosome 11 reported in *Blm*-deficient mice (Luo, Santoro et al. 2000). Considering that the Cre/*lox*P-induced mitotic recombination is achieved in a small time window by transient Cre expression, it is likely that the recombination efficiency can be significantly increased if Cre is expressed constitutively.

A possible limitation to the application of inducible mitotic recombination in mice is genome imprinting. A number of regions on several different mouse chromosomes have been identified to have imprinting effects, ranging from early embryonic lethality to various developmental defects (Cattanach and

Jones 1994). If the daughter cells generated by induced mitotic recombination carry two maternally or two paternally imprinted chromosomes, these cells will either over-express the imprinted gene(s) or not express it at all. If these imprinted chromosomes are used for genetic mosaic experiments, mitotic recombination should only be induced later at a developmental stage so that imprinted gene(s) do not cause any visible phenotypes. On the other hand, inducible mitotic recombination also provides a good way to study genomic imprinting both *in vivo* and *in vitro*. ES cells with two paternally or two maternally imprinted chromosomes can be made by mitotic recombination and injected into wild-type blastocysts or tetraploid embryos to study the contribution of these cells into different cell lineages. These ES cells can also be differentiated *in vitro* to study the effect of imprinting on the development.

It is suspected that nonspecific Cre-mediated recombination between cryptic genomic *lox*P sites could induce DNA damage and cause background problem. However, so far no dada can support this hypothesis. Transient expression of Cre recombinase should be able to minimize the effect even it does exist. Induced mitotic recombination is useful not only for generating genetic mosaic *in vivo*, but also for making homozygous mutations *in vitro*. It is compatible with a wide range of mutagenesis methods, including gene targeting and gene-trapping. Unlike the homozygous mutant clones generated in *Blm*-deficient ES cells, induced mitotic recombination can be used to generate pure homozygous clones instead of a pool of heterozygous and homozygous clones.

1.7 Gene-trap Mutagenesis

1.7.1 A brief history

Ever since the mouse fanciers began to collect mice, there have been numerous records of spontaneous mutants. When genetics became a formalized science, mouse geneticists around the world were no longer satisfied with the simple collection and documentation of spontaneous mutants (Stanford, Cohn et al. 2001). Instead, methods have been developed to generate large number of mutants in an efficient way.

X-ray mutagenesis was the first high-efficiency method that was applied to generate mutants for mouse genetic studies (Stanford, Cohn et al. 2001). The X-ray mutation rate (13~50 X 10^{-5} per locus) is about 20-100 times higher than the spontaneous mutation rate (5 X 10^{-6} per locus) in the mouse. What's more, X-rays cause chromosomal rearrangements, which leaves a molecular marker for localizing the mutated gene. But the chromosomal rearrangements often affect several genes close to the break points. Also the dosage of X-ray is limited because the high dosage required for germline mutagenesis induce massive levels of cell death to the animal.

Chemical mutagenesis with ethylnitrosourea (ENU) generates mostly point mutations, and thus affects only single genes. It is also much more efficient than the X-ray mutagenesis, with a typical mutation rate of around 150 X 10^{-5} per locus. Besides, the drug is easy to administrate. But the limitation of this approach is that it leaves no markers in the mouse genome. A complicated mating strategy is therefore needed to map the mutations.

In 1976, exogenous retroviruses were shown to be able to transmit through the mouse germline (Jaenisch 1976). The subsequent observations that the retroviral integration can disrupt endogenous genes and alter their expression have led to the widespread use of insertional mutagenesis in mouse. The integration of a retrovirus can produce a loss-of-function mutation if it integrates into the coding region of a gene. Retroviral integration can also generate gain-of-function mutations because the viral LTR contains a strong enhancer element. Wild-type retroviruses are not very efficient mutagens because the vast majority of the insertions are in the non-coding parts of the genome. Viral integrations in these locations will neither activate nor inactivate a gene, and thus these are often phenotypically "neutral" to cells.

Gene-trapping technology has successfully circumvented the limitation of insertional mutagenesis using wild-type retroviruses. Gene-trapping vectors contain a non-functional selection cassette and/or a reporter cassette.
The selection and/or reporter cassettes are designed so that they are only activated if integration occurs in the vicinity of an endogenous gene capturing

the transcriptional elements. The requirement that selection markers in the virus are only activated effectively eliminates any random integration events in non-coding regions of the genome.

Though gene targeting by homologous recombination has made it possible to precisely knockout all the genes in the mouse genome, there are still some limitations of this technology. The main drawback is that, like all the reverse genetics approaches, it is hard to predict the biological pathway from the observed phenotypes. Also in some cases, single gene knockouts might not have any obvious phenotype because of genetic redundancy. The embryonic lethality caused by ablation of some developmentally important genes also prevents their function in the adults from being characterized (Stanford, Cohn et al. 2001). So random mutagenesis is still of great interest to mouse geneticists to address many important biological questions.

1.7.2 Gene-Trap vectors

Various gene-trap vectors have been designed for different purposes. They can be divided into three main groups: enhancer-trap, promoter-trap and PolyA-trap vectors .

1.7.2.1 Enhancer-trap vectors

Enhancer-trap vectors contain a minimal promoter that is not functional. The vectors are activated when they insert next to a *cis*-acting endogenous enhancer element, which activates expression of the selection and reporter cassettes (Fig. 1-2a). Enhancer-trap vectors have not been widely used in the mouse because loss-of-function mutations generated by this type of vector are very rare. The enhancer elements are often located far away from the coding region of a gene. Analysis of the integration sites from ES cell lines that show reporter expression *in vivo* has indicated that most insertions are not in the coding regions (Gossler, Joyner et al. 1989). Thus, enhancer-trap vectors rarely disrupt the normal expression of a gene.

1.7.2.2 Promoter-trap vectors

The essential part of a promoter-trap vector is a promoter-less reporter gene that has a strong splice acceptor (SA) site immediately upstream of it. The expression of the reporter can only be driven by an endogenous promoter and the enhancer elements of a trapped gene (Fig. 1-2b). A fusion transcript is thus generated from 5' end of the endogenous gene and the reporter gene. The fusion transcript can be used to clone the 5' end of the trapped gene using the Rapid Amplification of cDNA Ends (RACE) technique. Since the splice acceptor can effectively capture the transcription from the endogenous promoter, transcription is stopped at the transcription terminator sequences in the gene-trap vector. Thus, promoter-trap vectors efficiently generate loss-of-function mutations.

The main disadvantage of promoter-trap is that insertion events which activate the selection cassettes mostly occur in introns. Alternative splicing can sometimes bypass the trapping cassette. Even very low levels of wild-type transcripts can sometimes partially rescue the phenotype and result in hypomorphic alleles (Gossler, Joyner et al. 1989; Friedrich and Soriano 1991). However, hypomorphic alleles can be very useful to characterize a gene's function, especially if the null allele causes early embryonic lethality.

The design of promoter-trap vectors restrict their use to genes which are expressed in the cell type of interest, for example undifferentiated ES cells. ES cells transcribe an abnormally high number of genes, but there are still many genes that do not express in ES cells or express at too low levels for them to activate a gene-trap cassette to a level suitable for selection. So other methods, such as PolyA-traps are needed to cover the rest of the genome.

Nevertheless, promoter-trap vectors are the most widely used trapping vectors. The International Gene-trap Consortium (IGTC; http://www-igtc.ca), a joint program of several academic groups, has successfully used a variety of plasmid and retroviral trapping vectors to achieve 32% genome coverage with 27,000 tags (Skarnes, von Melchner et al. 2004).

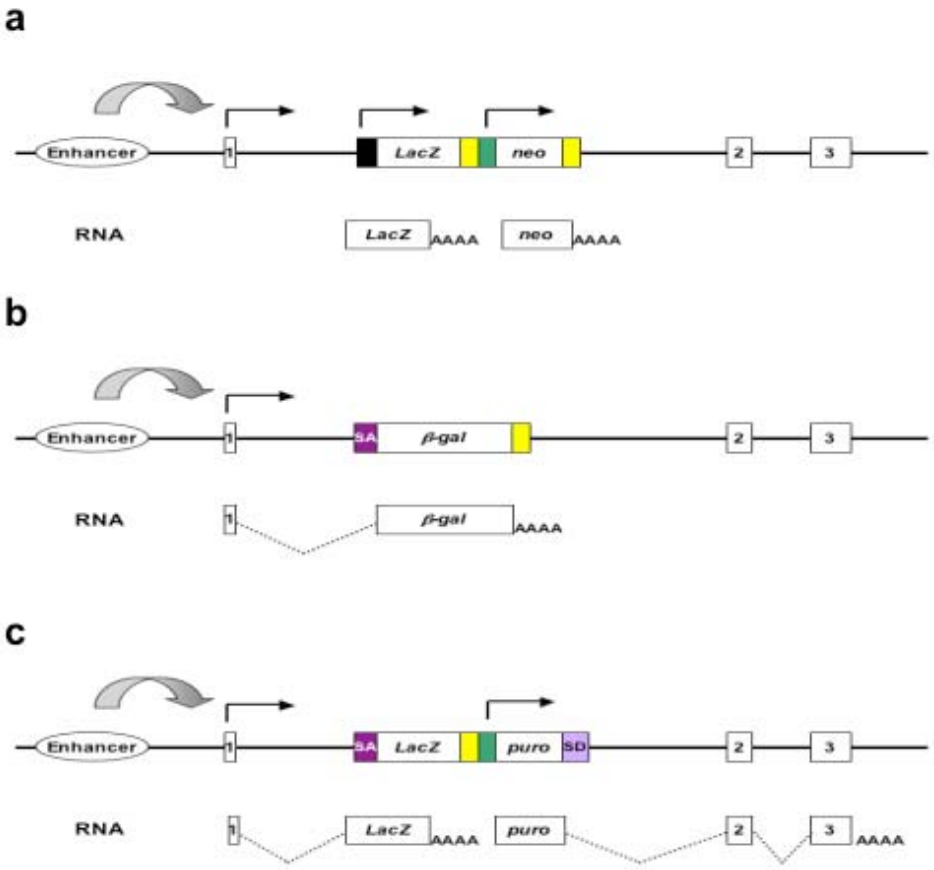


Fig. 1-2 Three main gene trap strategies. a. Enhancer trap. *LacZ* and *neo* reporter genes are driven by minimal promoters. The expression of reporter genes requires an endogenous enhancer. **b.** Promoter trap. The reporter gene doesn't have its own promoter. The expression of the reporter can only be driven by the promoter and enhancer elements of an endogenous gene. **c.** PolyA trap. The reporter gene doesn't have its own polyadenylation signal. *Puro* is transcribed from an autonomous promoter and spliced from the splice donor (SD) into an endogenous gene. Black and green boxes, different promoters; yellow box, poly A; SA, splicing acceptor; SD, Splicing donor.

1.7.2.3 PolyA-trap vectors

A polyA-trap vector contains a reporter gene lacking a polyadenylation signal, but possess a “strong” splice donor (SD) site. The reporter gene has its own promoter, but the reporter transcript is not stable unless the vector inserts into an endogenous gene upstream of a splice acceptor and a polyA signal (Fig. 1-2c). Usually, these vectors are designed so that termination codons of all the three reading frames follow the reporter gene, which prevents translation of the 3’ end of the trapped gene.

In contrast to a promoter-trap, a polyA-trap vector can mutate genes that do not normally express in undifferentiated ES cells, since the reporter is driven by an exogenous promoter, which is active in most genomic locations. The 3’ RACE technology can be used to clone the downstream exons, which is more reliable and robust than 5’ RACE.

One drawback of polyA-trap vectors is that they trap some pseudo splice acceptors and polyadenylation signals in the mouse genome. There are many fossilized gene fragments in the genome from old gene duplication events. Alternative splicing at the 3’ end of the gene is another potential problem. The mutagenicity rate of this kind of vectors is still controversial. Lexicon Genetics, a US-based biotechnology company, first used this method to create sequence-tagged mutations on a large scale (Zambrowicz, Friedrich et al. 1998). Lexicon has now achieved 60% genome coverage with 200,000 sequence tags (Zambrowicz, Abuin et al. 2003). It is interesting to note that about one-fifth of the genes trapped by IGTC are not represented in Lexicon’s tags (Skarnes, von Melchner et al. 2004). Since the number of sequence tags attained by the public domain is still relatively small, it is hard to predict how many genes trapped by Lexicon will never be represented in the promoter trapping approach pursued by IGTC. Nevertheless, the two efforts combined together have already trapped nearly two-thirds of all mouse genes (Skarnes, von Melchner et al. 2004) and provided a invaluable resource for mouse functional genomics and genetic studies.

1.7.3 Application of gene-trap mutagenesis in genetic screens

1.7.3.1 Expression screens

A genetic screen is an important method to identify a gene in any pathway or developmental event. Compared with phenotypic screens using ENU mutagenesis, gene-trapping in ES cells is a relatively inefficient method to generate germline mutations (Gossler, Joyner et al. 1989; Friedrich and Soriano 1991). But it has the advantage that the reporter, *βgeo* and human placental alkaline phosphatase (PLAP), can serve as tag for the expression of the trapped genes. The temporal and spatial expression pattern of a gene provides clues for its function. Some developmentally important genes often show highly restricted expression patterns during development.

Wurst et al. (1995) generated about 300 aggregation chimaeras using ES cell lines that contained gene-trap insertions. X-gal staining was then used to examine the expression patterns of the mutated genes in chimaeric embryos. About two-thirds of the chimaeric embryos expressed *lacZ*, which was temporally and spatially restricted for many lines. In a similar screen, Stoykova et al. (1998) has analysed 64 mouse lines generated from the gene-trap ES cell lines. About 75% of these lines showed embryonic *lacZ* expression (Stoykova, Chowdhury et al. 1998). Interestingly, for both screens, a large portion of the genes trapped in undifferentiated ES cells show *lacZ* expression in the developing nervous system.

It was noted in these studies that many gene-trap clones show widespread *lacZ* expression in *vivo*. Since many groups studying developmental questions are interested in genes that have highly restricted expression patterns in specific cell lineages, a pre-screen to enrich for genes with these characteristics would be valuable to save time and effort. For this reason, libraries of gene-trapped ES cell clones have been induced by specific growth/differentiation factors or physiological stimuli, such as nerve growth factor, retinoic acid, engrailed homeobox proteins and γ-irradiation. Those gene-trap integrations that are found to be either activated or repressed by

one of these factors are more likely to be linked to a specific signalling pathway. *LacZ* staining during embryogenesis has shown a strong enrichment of gene-trap clones that have restricted patterns *in vivo* after the induction screen procedure (Bonaldo, Chowdhury et al. 1998).

Gene-trap vectors have also been designed to trap specific classes of genes. Skarnes et al. (1995) designed a vector to trap genes that encode secreted and transmembrane proteins. In this secretory-trap vector, a transmembrane (TM) domain was placed between the splice acceptor (SA) and the *βgeo* reporter. The transmembrane domain will result in the sequestration of the *βgeo* fusion protein into the lumen of the endoplasmic reticulum (ER) of genes that encode non-secretary proteins, and thus abolish the *β-gal* activity. The *βgeo* reporter is expressed when an additional secretory signal (SS) from the trapped gene results in the *βgeo* portion of the fusion protein being positioned in the cytosol (Skarnes, Moss et al. 1995). Recently, this secretory-trap design was modified to identify and mutate receptors and ligands controlling neuronal axon guidance. Leighton et al. (2001) used a human placental alkaline phosphastase (PLAP) reporter, which is co-expressed with the *LacZ* gene-trap reporter, to label neuronal projections. By *β-gal* and PLAP staining, genes with restricted expression patterns in neuronal axons were identified. Recently, Chen et al. (2004) have used an inducible gene trapping system to screen gene trap events responding to retinoic acid (RA). 65 gene traps were identified using this method. *In vivo* analysis revealed that 85% of the retinoic acid-inducible gene traps trapped developmentally regulated genes.

1.7.3.2 Phenotype-driven screens

Phenotype-driven screens in diploid genomes require a strategy to obtain homozygous mutations. So it is difficult to perform a phenotype-driven screen in ES cells. Recently, Guo et al. (2004) has utilized the elevated mitotic recombination rate of the *Blm*-deficient ES cells to generate a genome-wide homozygous mutation library of gene-traps. This library was then used to screen for genes involved in the DNA mismatch repair pathway (Guo, Wang

et al. 2004). This strategy has provided a new way to carry out recessive genetic screens in ES cells.

1.7.3.3 Genotype-driven screens

Sequence-based screens have been made possible by the availability of large resource of ES cell clones with defined mutation. Several academic groups in the International Gene-trap Consortium (IGTC; http://www.igtc.ca) have initiated a genome-wide gene-trap program aimed at generating an international resource of embryonic stem cells with gene-trap insertions in most mouse genes. Gene-trap cell lines generated by the IGTC are freely available to the public and all the sequence tags are finely mapped on Ensemble mouse genome browser (http://www.ensembl.org/Mus_musculus) (Skarnes, von Melchner et al. 2004). ES cell lines which carry gene-trap insertions are now available for nearly 40% of mouse genes. A parallel effort is also carried out by Lexicon Genetics, but the cell lines and the mouse line derived from them are available with a fee and limitations of future work.

1.7.4 Electroporation versus retroviral infection

Trapping vectors can be introduced into the genome by either electroporation or retroviral infection. Both methods have their advantages and disadvantages. Since gene-trap screens using both methods have been reported, it is possible now to assess the two methods.

1.7.4.1 Electroporation

The simplest way to perform gene-trap mutagenesis is to electroporate the linearized gene-trap vector directly into mammalian cells. The gene-trap vectors introduced by electroporation integrate into the genome randomly, while the retroviral vectors tend to insert into the 5' portion of the gene. One advantage of the electroporation method is that it does not require the construction of a virus, which has numerous constraints discussed later. Scaling up electroporation is relatively easy because large amounts of DNA can be easily prepared. Last but not least, theoretically there is no limitation on the size of the trapping vector. Multiple reporter cassettes can be incorporated into one vector, which can be tailored for specific usages.

The biggest disadvantage of the electroporation method is that integrations are always accompanied by DNA concatermerization, though conditions can be optimized such that concatermers occur in less than 20% of the cells (Stanford, Cohn et al. 2001). Tandem insertions into the same locus happen through a recobinational process to form a concatomer followed by non-homologous end joining DNA repair (NHEJ) as the vector inserts into the genome. Multiple copies of the gene-trap vector in one locus can result in ectopic reporter expression, aberrant splicing, and can complicate the identification of the gene-trap mutations by 5' RACE. Sometimes, the gene-trap vector can be truncated during electroporation due to exonuclease digestion. The loss of different amounts of flanking sequence makes the cloning of the flanking genomic sequence by Inverse PCR problematic.

1.7.4.2 Retroviral infection

1.7.4.2.1 Retroviral life cycle

The typical retrovirus genome consists of two copies of a single-stranded RNA molecule of about 8-12 kb. The wild type murine leukaemia virus genome encodes three major proteins, Gag, Pol and Env. Gag is processed to make the core proteins. Pol has the reverse transcriptase, RNase H and integrase activities. Env encodes the viral envelope protein. A mature viral particle consists of the virus nucleoprotein core and the outer lipid-protein shell of the viral envelope (Fig. 1-3a).

Viral particles infect host cells by binding to cell surface receptors, which is mediated by the envelope proteins of the retrovirus. Infection is followed by injection of the virus nucleoprotein core into the host cell. After this, a double-stranded DNA is generated from the viral genomic RNA by the viral reverse transcriptase. Finally, the newly transcribed double strand viral DNA integrates into the host chromosome, which is catalyzed by the viral integrase. The integrated viral DNA is known as proviral DNA.

Once integrated, the virus is ready to initiate a new round of replication and infection. Full-length genomic RNA is transcribed from proviral DNA by the host cell RNA polymerase. The genomic RNA is either processed to generate mRNAs, which are translated into the viral proteins by the host protein synthesis machinery, or the full length genomic RNA remains unspliced and is packaged into new viral particles and released from the host cell by budding from the plasma membrane .

1.7.4.2.2 Recombinant retroviral, viral packaging cell lines

A retroviral vector can be used to transfer exogenous DNA into eukaryotic cells. Because the retrovirus can efficiently integrate into the host genome, exogenous genes carried by the retrovirus can be expressed. However, the wild type retrovirus is not ideal for this purpose because of the size limit of the genomic RNA that can be efficiently packaged into the virus particle.

Many recombinant retroviral vectors have been constructed. A typical recombinant retroviral vector includes the 5' long terminal repeat (5' LTR), the 3' long terminal repeat (3' LTR) and viral RNA packaging signal, known as Ψ. All the other essential components of the wild type retrovirus are deleted to make space for the exogenous DNA (Fig. 1-3b). The recombinant retrovirus vector itself is replication-deficient. To produce infectious retrovirus, the vector needs to be transfected and transcribed in a viral packaging cell line, which can express all three proteins that are required for viral reproduction, Gal/Pol and Env.

When a recombinant retrovirus construct is transiently transfected into the viral packaging cell line, the transcribed genomic RNA is recognised and assembled as an infectious particle with the viral proteins. The derived replication-deficient retrovirus particle can infect any cells that have the receptor for the virus and the vector can then integrate into the host genome. However, there is only one infection cycle because the recombinant virus lacks the Gal/Pol and Env products (Fig. 1-3c).

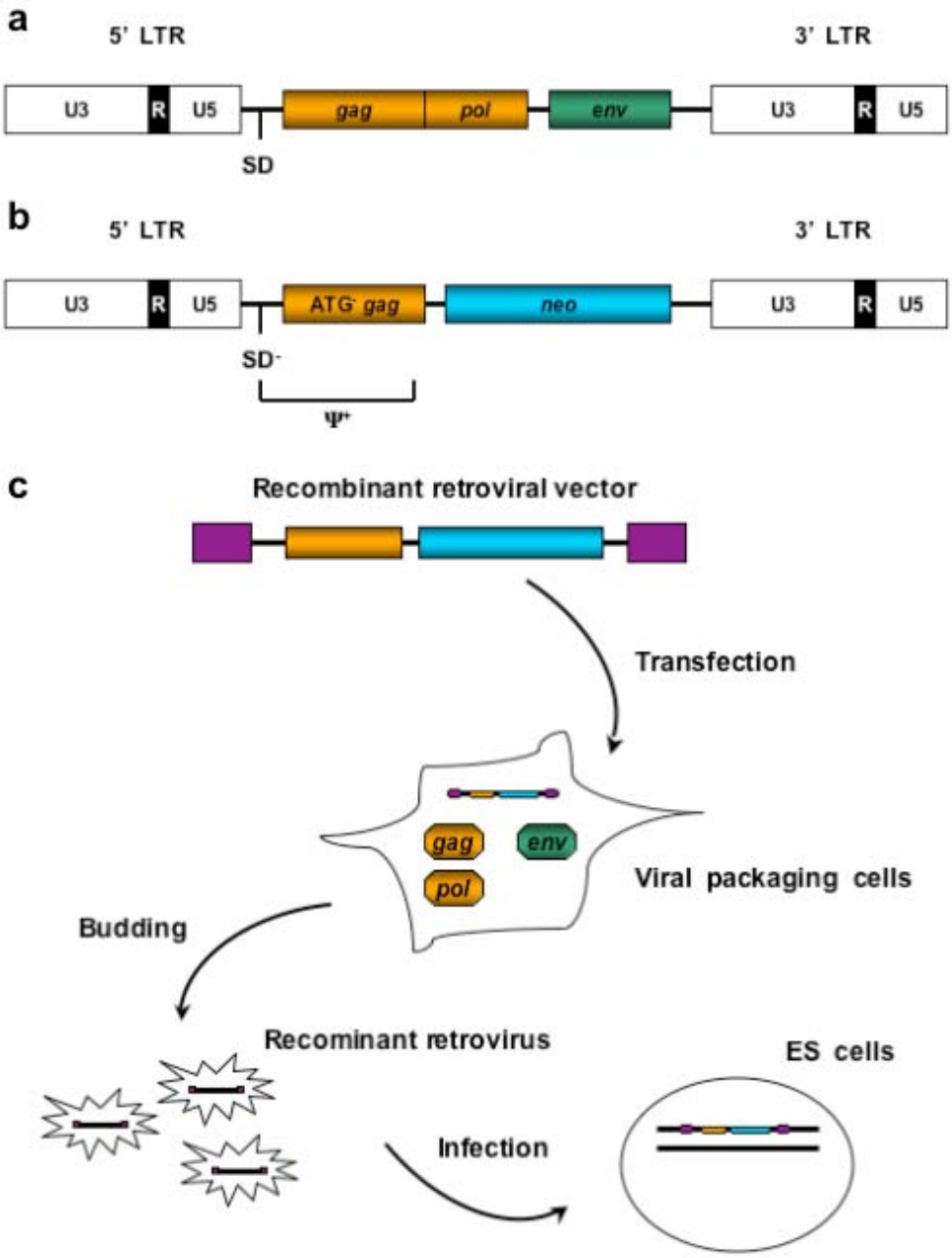

a
5' LTR
3' LTR
U3
R
U5
gag
pol
env
U3
R
U5
SD
b
5' LTR
3' LTR
U3
R
U5
ATG- gag
neo
U3
R
U5
SD-
Ψ+
c
Recombinant retroviral vector
Transfection
gag
env
pol
Viral packaging cells
Budding
Recominant retrovirus
ES cells
Infection

Fig. 1-3 Retroviral vectors and viral production. a. The structure of a wild type retrovirus. The virus genome contains genes that encode viral protein core (*gag*), reverse transcriptase (*pol*), and viral envelope protein (*env*). SD, viral splicing donor. **b.** The structure of a recombinant retroviral vector, pBabe. The viral genes *pol* and *env* are deleted. A *neo* selection marker is inserted to allow the selection of retroviral integration. The splicing donor and the *gag* gene are kept to facilitate the viral packaging. The splicing donor is mutated (SD^-) and the initiation codon of gag gene is deleted (ATG^-*gag*) to avoid the interference of internal gene expression. Ψ, viral packaging signal. **c.** Production of the recombinant retrovirus. The recombinant retrovirus is replication-deficient. To produce infectious retrovirus, the vector need to be transfected into a viral packaging cell line, which can express all the three proteins that are required for viral reproduction, Gal/Pol and Env. In the packaging cell line, the viral genomic RNA is packaged and infectious viral particles are released. The recombinant retrovirus produced can be used to infect ES cells.

1.7.4.2.3 Retroviral based gene-traps

Von Melchner et al. (1989) developed the first retroviral gene-trap vector. In this design, the gene-trap cassette is inserted in the U3 region of the 3' LTR and replaces the viral enhancer. After viral replication and integration, the provirus carries a duplicated gene-trap cassette in both of the 5' and 3' LTRs. The cassette in the 5' LTR is situated just 30 nt from the host genome and is activated by transcriptional read-through rather than splicing (von Melchner and Ruley 1989). Friedrich et al. (1991) designed another version of a retroviral gene-trap vector, called ROSA (reverse orientation splice acceptor). In the ROSA vector, the gene-trap cassette was placed between viral LTRs in the opposite orientation relative to viral transcription. This reverse orientation was essential to avoid removal of the viral packaging sequence Ψ from the full length genomic RNA by splicing from the upstream viral splice donor sequence directly to the splice acceptor in the gene-trap cassette (Friedrich and Soriano 1991). In this design, the cassette is activated only by a splicing event.

1.7.4.2.4 Advantages and disadvantages of Gene-trap via retroviral infection

Gene-trap mutagenesis using a retroviral vector has advantages compared to electroporation. In contrast to electroporation, only a single copy of retrovirus will integrate into one genomic locus. The provirus has a predictable structure which is the same in every clone, which makes the cloning of virus insertion site by PCR based methods very reliable. The trapped exons can be determined by RACE. The flanking genomic fragment of the insertion site can also be cloned by inverse PCR (Suzuki, Shen et al. 2002) or splinkerette PCR (Mikkers, Allen et al. 2002). Once the virus trapping titre is determined, it is easy to control the virus infection conditions so that most of the cells will only contain a single copy of the gene-trap vector. Retroviruses have a propensity to integrate into 5' portion of a gene. Virus insertion in the 5' untranslated region and first few introns is more likely to generate null alleles.

Retroviral vectors also have their limitations. First, the packaging size of the retroviruses is highly limiting. Even within the limit, the virus packaging efficiency drops significantly as the size of the virus increases. Second, the virus insertion can induce retroviral-mediated gene silencing. Deleting the LTR enhancer sequences can solve this problem. Third, viral splice donor sequence in the 5' LTR can cause ectopic reporter expression. This problem can be solved by putting the reporter gene in the reverse orientation. Fourth, non-random retroviral integration results in trapping "hot spots", but the same problem also exists for the plasmid-based gene-traps (Hansen, Floss et al. 2003).

1.7.4.3 Gene-trap "hot spots"

Although it was noticed a long time ago that there are preferred integration sites, or "hot spots" for gene-trap mutagenesis, the data available was not sufficient to systematically assess the problem. Recently, the German Gene-trap Consortium (GGTC) reported the generation of over 11,000 independent gene-trapped ES clones using four different gene-trap vectors, including two electroporation-based vectors and two retrovirus-based vectors. 5,142 sequence tags were obtained from gene-trap insertions, which made it possible to do a systematic analysis of gene-trap "hot spots" by both methods (Hansen, Floss et al. 2003).

It was found that there was a direct correlation between the number of gene-trap insertions on a given chromosome and the number of the genes on that chromosome, which suggests that there is no obvious bias to a single chromosome. Of all the recovered UniGene clusters, 75% of appeared only once, while the remaining 25% were hit multiple times. This data suggested that most mouse genes are randomly accessible to gene-trap mutagenesis (Hansen, Floss et al. 2003).

45% of hot spots were hit by more than one of the four vectors, suggesting that these hot spots might be caused by locus-specific factors, for example secondary chromatin structure (Hansen, Floss et al. 2003). Considering that

over half of the hot spots are vector-specific, it seems that each gene-trap vector design also has its own pool of trappable genes. Achieving saturation mutagenesis with gene-trap vectors will require the use of a combination of different gene-trap vector designs.

1.8 ES cell *in vitro* differentiation

1.8.1 Introduction

In mammals, the fertilized oocyte and the blastomeres of 2-, 4- and 8-cell stage embryos are totipotent. They can generate a complex organism of hundreds of different specialized cell types (Wobus 2001). On the other hand, the embryonic stem cells, which are derived from inner cell mass (ICM) of blastocysts, are only pluripotent. When transferred back to blastocysts, they can contribute to all the different cell types, except the placental tissues, thus they are not able to generate a complex organism by themselves.

The ability of ES cells to give rise to various cell types including germline cells has laid the basis of gene targeting technology (Bradley, Evans et al. 1984). However, ES cells not only differentiate *in vivo*, they can also form three-dimensional embryo-like aggregates which contain cells of the endodermal, ectodermal and mesodermal lineages (Wobus 2001). These three germ layers can further differentiate into a variety of specialized cell types including cardiac muscle, smooth muscle, skeletal muscle, haematopoietic, pancreatic, cartilage and neuronal cells (Czyz, Wiese et al. 2003). ES cell *in vitro* differentiation can therefore recapitulates the early mouse embryogenesis to a degree.

1.8.2 *In vitro* differentiation potential of ES cells

When ES cells are cultured on feeder layers and/or in medium supplemented with differentiation inhibitory factors such as LIF, they can remain undifferentiated indefinitely. But, once the ES cells are deprived of differentiation inhibitory factors, they will commit to a differentiation fate.

Many different protocols have been established for the *in vitro* differentiation of ES cells into different terminally differentiated cell types by the “hanging

drop" method (Wobus, Wallukat et al. 1991) (Fig. 1-4), by the "mass culture" method (Doetschman, Eistetter et al. 1985), by cultivation in methylcellulose (Wiles and Keller 1991), by stromal cell co-culture method (Nakano, Kodama et al. 1994; Kawasaki, Mizuseki et al. 2000) or by adherent monoculture method (Nishikawa, Nishikawa et al. 1998; Ying, Stavridis et al. 2003). Treatment of differentiation cultures with soluble growth factors can also help to drive differentiation into specific directions.

Compared to the "hanging drop" and "mass culture" methods, which require the generation of three-dimensional embryoid bodies (EB), the stromal cell co-culture and adherent monoculture methods are much simpler and highly efficient in inducing ES cells to differentiate into neuronal or haematopoietic lineages. The success of these methods proves that the three-dimensional structure in the embryoid body is not requisite for blood cell, endoderm and neuron cell differentiation. By adding exogenous growth factors (adherent monoculture method) or the secretion of endogenous factors (stromal cell co-culture method), ES cells can be coaxed into a specific lineage without the spatial information of the embryo (or embryoid bodies).

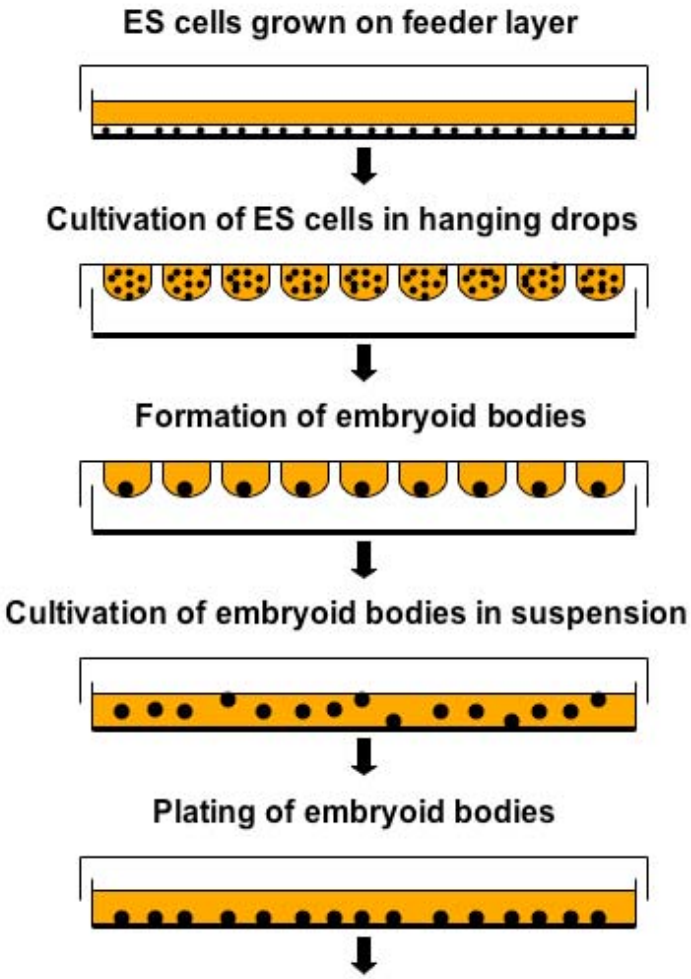
As the spatial information of the embryo is absent in the culture of ES cells, the lineage and stage of differentiation may conveniently be determined by the cell surface markers (Nishikawa, Nishikawa et al. 1998; Ying, Stavridis et al. 2003) or the GFP reporter genes tagged to an intracellular markers (Nishikawa, Nishikawa et al. 1998; Ying, Stavridis et al. 2003). Fluorescence Activated Cell Sorting (FACS) can be used to quantitatively compare the differentiation efficiency of ES cell lines in different culture conditions or with different genetic modifications.

However, these methods are used to differentiate ES cells into some specific lineages. The culture conditions, the addition of differentiation inducers and/or the stromal cell lines are different from one protocol to another. If more than one cell lineage is studied, ES cell lines need to be differentiated in many different ways which increases the complexity of the experimental design. On

the other hand, “hanging drop” method provides a simple and universal way to differentiate ES cells into multiple lineages.

Embryoid bodies (EB) can undergo specific and reproducible morphological changes. First, an outer layer of endoderm-like cells are developed over the primitive ectoderm layer. This is followed by the formation of an ectodermal rim and by the formation of the mesodermal cells. When the EBs are plated and extrinsic differentiation factors are added, differentiated specialized cells can develop in the outgrowth area of the EBs (Wobus 2001).

The morphological changes are accompanied by a dynamic change in the expression pattern of a set of lineage- and tissue-specific genes. During the first several days of EB development, the primitive ectoderm specific genes such as *Oct4* and *Fgf5* express at a high level. This is followed by up-regulation of genes characteristic for early postimplantation stages, such as *Nodal* and early endodermal genes *vHnf1*, *Hnf3*β and *Hnf4*. At the same time, genes that are characteristic of gastulation and the early mesodermal differentiation show maximum expression, such as *Brachyury*, *Goosecoid* and *Bmp4*. Tissue-specific genes that show developmental regulated expression patterns will begin to express after that. At the terminal differentiation stage, genes that are expressed only in specialized cell types are detected (Rohwedel, Guan et al. 2001).



Differentiation of endodermal, ectodermal and mesodermal cells.

RT-PCR, in situ hybridization and immunohistochemistry was performed to detect the developmentally controlled expression of mRNA and proteins.

Fig. 1-4 *In vitro* differentiation of ES cells using the "hanging drop" method. Certain number of undifferentiated ES cells are cultivated in hanging drops for 2 days to form the cell aggregates, embryoid bodies (EB). The EBs are cultivated in suspension for another 3 days before they are plated onto gelatin-coated tissue culture plates. The cellular phenotypes of differentiated cells derived from endodermal, ecotodermal and mesodermal lineages can be characterized by a variety of methods.

Leahy et al. (1999) has performed *in situ* hybridization on EBs using probes for germ layer markers (*Oct4*, *Fgf5*, *Gata4*, *Nodal* and *Brachyury*) as well as cell lineage-specific markers (*Flk-1*, *Nkx2.5*, *Eklf* and *Msx3*). Since the expression patterns of these markers has already been extensively characterized *in vivo*, the marker expression during EB formation and early embryogenesis can be correlated. By this method, different stages of EB *in vitro* differentiation can be linked to different stages of embryogenesis *in vivo* (Leahy, Xiong et al. 1999). Markers that show highly reproducible temporal and spatial distribution can thus be used in genetic screens for mutations that disrupt the normal expression pattern of these markers.

This fundamental work has made it possible to use ES cell *in vitro* differentiation as an alternative to study the functions of the genes that are important in early embryogenesis. The mutations that are caused by over-expression (gain-of-function mutations) or homozygous targeting (loss-of-function mutations) in ES cells can be analyzed in this system. The loss-of-function approach is very important for functional studies, especially for those genes that cause early embryonic lethality when both alleles are disrupted (Wobus 2001).

In vitro differentiation of ES cells can not only be used as a model system to study early embryogenesis, it also provides a promising way to generate terminally differentiated cell types for therapy. Some cell types, such as cardiomyocytes, neuronal and glial cells, and pancreatic cells, are of potential therapeutical relevance because they can be used for the treatment of cardiac diseases, neurodegenerative disorders and diabetes, respectively.

However, cystic EBs are heterogeneous, they consist of various differentiated cell types as well as undifferentiated ES cells. Numerous experiments have demonstrated that the *in vitro* differentiation of ES cells can be directed into certain lineages by controlling various parameters such as the starting ES cell numbers to form the EB aggregates, the composition of the differentiation media, the type, concentration or combinations of the growth factors,

differentiation induction factors and genetic “gain-of-function” and “loss-of-function” manipulations (Czyz, Wiese et al. 2003).

Though much progress has been made in determining the best conditions for ES cells to differentiate into certain lineages, this technology is still far from mature enough to be used for transplantation therapy. Many important developments are required, for example improving the efficiency of the differentiation protocols, the purity of cell population of the desired cell types, limiting the potential tumorigenicity of the undifferentiated ES cells, controlling the donor/recipient immune-compatibility and achieving the long-term functional engraftment of differentiated cells *in vivo* (Czyz, Wiese et al. 2003). Considering the ethical concerns surrounding the human ES cells isolated from *in vitro* fertilized human embryos and the much more advanced stages of research in the mouse system, these questions need to be answered in mouse ES cells first and then confirmed in their human counterpart.

1.8.3 Using genetic screens to study the ES cell *in vitro* differentiation

Relatively little is known about the genetic pathways that control ES cell *in vitro* differentiation and cell lineage determination. Though efficient differentiation protocols have been established for some cell types, we still do not know exactly why and how the change of concentration of certain factors can dramatically increase the percentage of certain cell types in the whole EB cell population. The lack of knowledge of the genetic pathways underlying the *in vitro* differentiation process significantly impedes further improvements in these protocols.

In vitro differentiation experiments have been performed with many mutant ES cell lines, and the phenotypes of these have been described. But in most cases, the technique was used to study the function of the genes that caused early embryonic lethality when both alleles were disrupted. Different differentiation methods (hanging drops or mass culture), different ES cell lines (R1, E14.1, and AB1) and different assay methods (RT-PCR, whole mount *in situ*, immunohistology and physiological analysis) all make the data from

different experiments not directly comparable. So a systematic approach is needed to thoroughly study the ES cell *in vitro* differentiation process.

A genetic screen is always the most powerful way to study a complex system. Screens have been successfully applied to many model organisms and mammalian cultured cell lines. Many important genetic pathways have been identified and well characterized using this approach. But very few genetic screens have been carried out on ES self-renewal and *in vitro* differentiation (Chambers, Colby et al. 2003). There has no loss-of-function screen been reported so far. One of the main reasons for this is the difficulty of producing numerous mutations in ES cells, especially homozygous mutations that are the key for recessive genetic screens.

There are many different ways to create homozygous mutations in ES cell, which have been described in chapter 1.6.3. However, selection can hardly be performed in *in vitro* differentiation studies, so single mutant ES clones need to be differentiated separately and checked one by one. *Blm*-deficient ES cells can not be used for this purpose because the resulting pool of cells will be a mixture of ES cells with heterozygous mutations with a few rare homozygous mutants. Targeting both alleles using marker recycling or different selection markers is both time-consuming and labour-intensive, limiting throughput. The high G418 concentration induction method can save the trouble of targeting the second allele, but it still requires the targeting of the first one and a genotyping strategy to distinguish the homozygous clones from the heterozygous ones. All these limitations make them unsuitable for conducting a recessive genetic screen to identify genes required for ES cell *in vitro* differentiation.

Induced mitotic recombination can be used to make homozygous mutations in a chromosome-specific way. Since all mutations distal to the mitotic recombination selection cassettes will become homozygous after Cre-induced mitotic recombination, the genotype of the mutations anywhere on the chromosome can be simply determined by the drug resistance as well as by genotyping results of the locus where the induced mitotic recombination

cassettes are targeted. But the problem remains, how does one introduce mutations on one specific chromosome which has been designed to undergo induced mitotic recombination. Gene targeting by homologous recombination is one choice, but the throughput will be limited. Insertional and ENU mutagenesis can create a lot of mutations in a random way, but most of these mutations will not be on the right chromosome, so they will not become homozygous after Cre-induced mitotic recombination. To identify clones with homozygous mutations from such a high background without selection is difficult for insertional mutagenesis and almost impossible for chemical mutagenesis. So clearly, a pre-screening strategy is needed to accumulate mutations on the desired chromosome before chromosome-specific mitotic recombination is induced by Cre expression.

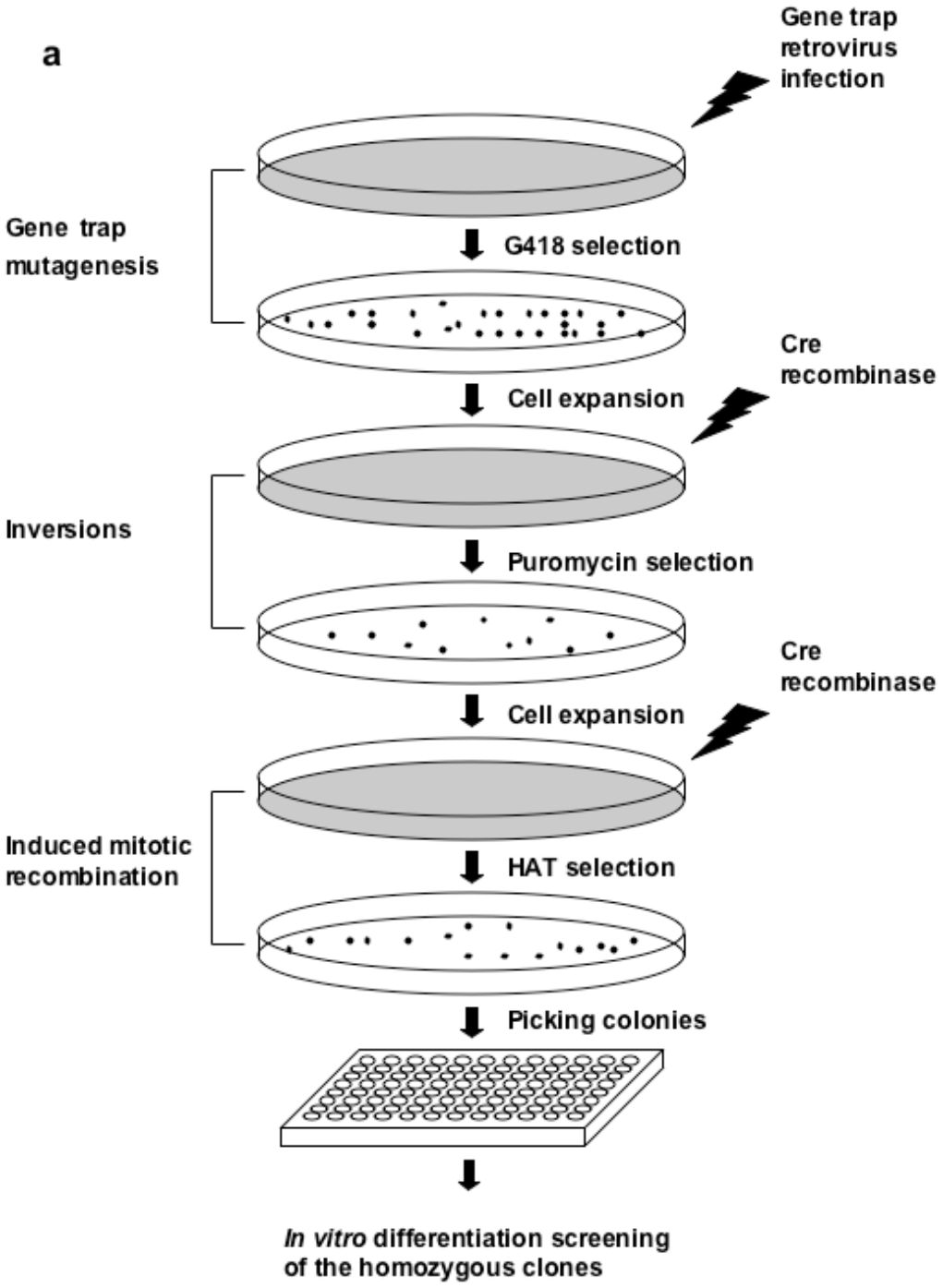
1.9 Thesis project

The primary goal of the project was to generate a large number of homozygous mutations in a genomic region of interest on chromosome 11 in mouse embryonic stem cells by exploring induced mitotic recombination and regional trapping mutagenesis methods and to investigate the application of a recessive genetic screen for genes involved in ES cell *in vitro* differentiation.

Mouse chromosome 11 was chosen for this project because of some unique characteristics. First, the distal half of the mouse chromosome 11 exhibits highly conserved linkage with human chromosome 17. Almost every gene mapped to human chromosome 17 is found on mouse chromosome 11. Second, numerous genetic tools have already been created on this chromosome (Liu, Zhang et al. 1998; Zheng, Sage et al. 1999; Su, Wang et al. 2000; Zheng, Sage et al. 2000; Liu, Jenkins et al. 2003). To date, 18 deletions and 3 inversions/balancers have been made which cover the chromosome (Yu and Bradley 2001). These resources make the downstream functional characterization of the mutations recovered from a genetic screen much easier. Third, the distal part of mouse chromosome 11 has a very high gene density which makes it an ideal target for mutagenesis studies. Forth, this chromosome is not imprinted, eliminating potential complexity associated with mono-allelic expression from maternal or paternal chromosomes.

A disadvantage of using mouse chromosome 11 for the ES cell differentiation study is that this chromosome is often found to be amplified in mouse ES cells (Nichols, Evans et al. 1990; Liu, Wu et al. 1997). ES cell clones that are trisomic for all or part of chromosome 11 exhibit accelerated cell growth and decreased efficiency of germ line transmission. Extensive engineering of this chromosome might increase the possibility of accumulating trisomic clones. These trisomic clones are likely to have abnormal differentiation potential and thus complicate the interpretation of the phenotypes. However, these clones are relatively rare in the whole culture population. If care is taken in genotyping the cell lines used to generate homozygous gene-trap mutations, it is possible to distinguish these trisomic clones by their abnormal growth rate and by Southern analysis (the ratio between the signals of targeted and un-targeted restriction fragments).

In the previous sections, the principles of induced mitotic recombination and regional trapping mutagenesis have already been discussed. The design of my project was: (i) Generate a cell line engineered to undergo induced mitotic recombination on chromosome 11 and capture a subset of gene-trap mutations that were generated on this chromosome. (ii) Generate a set of genome-wide gene-traps in this cell line and isolate those that are located on chromosome 11 using a Cre/*lox*P mediated inversion strategy. (iii) Make the inversions homozygous by Cre-induced mitotic recombination. (iv) Assess the homozygous clones for their developmental potential by an *in vitro* differentiation assay. (v) Confirm the mutation by BAC rescue or by re-generating the mutation followed by re-confirmation of the phenotype (Fig. 1-5a and b).


a
Gene trap retrovirus infection
Gene trap mutagenesis
G418 selection
Cre recombinase
Cell expansion
Inversions
Puromycin selection
Cre recombinase
Cell expansion
Induced mitotic recombination
HAT selection
Picking colonies
In vitro differentiation screening of the homozygous clones

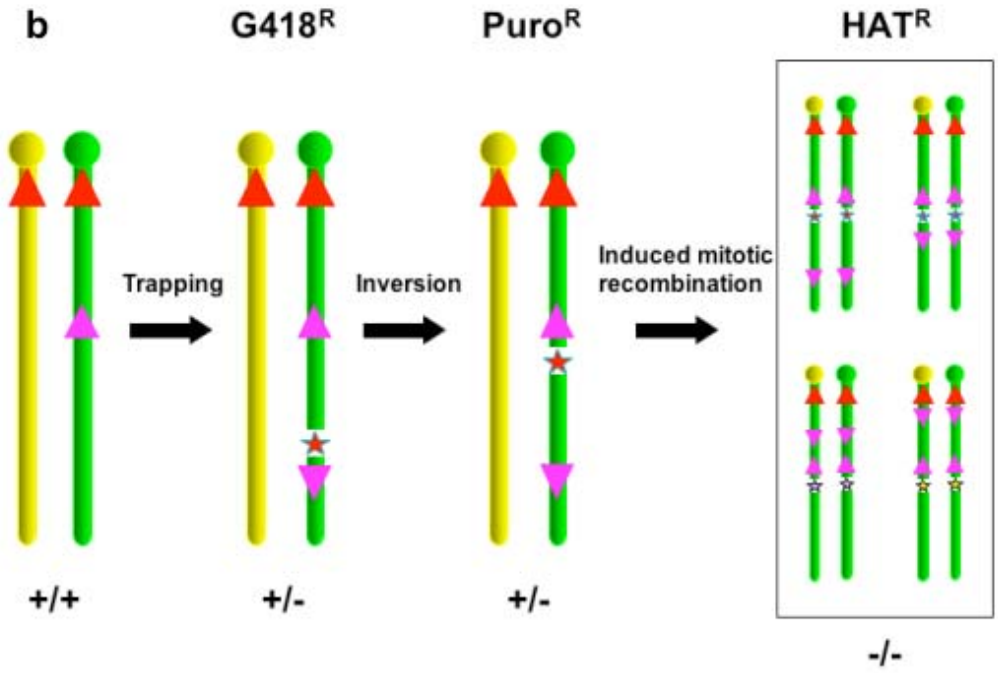


Fig. 1-5 Thesis Project. a. Schematic illustration of the screening strategy for homozygously mutated ES cells. The whole project is composed of three independent steps, gene trap mutagenesis, regional accumulation of the gene trap mutations by inversion and the induced mitotic recombination. Since each step is selected by a different drug selection marker, no genotyping is necessary for the intermediate products. **b.** Schematic illustration of the chromosomal structure of the products of each step. A cell line was engineered to undergo mitotic recombination and to capture the subset of gene trap mutations that were generated on this chromosome. A large panel of genome-wide gene traps were generated in this cell line and those that are located on chromosome 11 were specifically selected via Cre/*lox*P mediated inversions. The inversions were then made homozygous by induced mitotic recombination. Red arrows, mutant *lox* sites; pink arrows, wild type *lox*P sites; Yellow and green bars: homologous chromosomes; stars of different colours, gene-trap insertions.