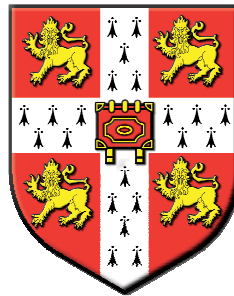


**A recessive genetic screen to discover components in miRNA
pathways using *piggyBac*-mediated insertional mutagenesis in
Blm-deficient mouse embryonic stem cells**

Meng Li

**A dissertation submitted in partial fulfilment of the requirements for the award of
the degree of**

**Doctor of Philosophy
of the
University of Cambridge**



The Wellcome Trust Sanger Institute
Sidney Sussex College, Cambridge

Declarations

This thesis is submitted on partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of Cambridge. It describes the work carried out at the Wellcome Trust Sanger Institute from October 2006 to August 2010. Unless otherwise indicated, the research is my own and not the product of collaboration.

This thesis does not exceed the word limit of 60,000 as set by the Degree Committee for the Faculty of Biology.

Meng Li

25th August 2010

Acknowledgements

My foremost gratitude goes to Prof. Allan Bradley for being an inspirational mentor with his broad vision to science. For the past four years, he has provided me with guidance, support and patience, helping me through the difficulties and frustrations. I am aware of the fact that without his open-minded approach to science and his encouragement, I would not be able to have the freedom to pursue my own ideas and build up confidence in conducting independent research.

Many thanks to the Bradley group, current members and alumni, for being great people to work with, and always being there for help and advice. Special acknowledgement goes to Kosuke Yusa, for making his resources available to me and from whom I have learnt a great deal and had so many instructive discussions. Also thanks to Qi Liang, for introducing me to many basic techniques and providing me with much of the starting materials to start my projects. Thanks to Stephen Pettitt, Yue Huang and Wei Wang (Mr.), who are also working on *piggyBac* and *Blm*-deficient ES cell systems, for all the instructive discussions and sharing of resources. I am very grateful to Frances Law, Alastair Beasley and James Cooper for tissue culture support; Haydn Prosser and Hiroko Yusa for providing me with useful plasmids; Roland Rad and George Vassiliou for useful discussions and having laughs with. I would like also to thank many great people from other labs in the Sanger Institute; Daniel Turner and Sabine Eckert for their expertise in conducting Illumina sequencing of my samples, Zeming Ning for helping me with bioinformatics analysis and Tomas Fitzgerald for array CGH analysis.

Thanks to my thesis committee members Julie Ahringer, Bill Skarnes, and David Adams for their comments and advice throughout my PhD. Thanks too to Seth Grant, Jyoti Choudhary and Jürg Bähler for hosting me in their labs during my first PhD year rotations and to Mark Colin and Marcelo Coba for guiding me through the rotations.

Robert, you are always here for me and this work would not be as it is without your constant support. I look forward to our life together as husband and wife. Thank you for proof-reading of this thesis too.

Mum and Dad, thank you so much for all your love and support emotionally and financially in whatever I chose to do and always being there for me even you are thousands of miles away in China when I am faced with difficult decisions and feeling sad.

Thanks to my future in-laws, Andy and Julie Phipps for being so welcoming, generous and treating me as a true family member for the past six years.

Abstract

Genome-wide recessive genetic screens complement reverse genetic approaches to ascribe gene function in biological pathways of interest without prior knowledge. The ability to conduct such screens in cultured embryonic stem cells requires a tractable mutagenic system combined with a *Blm*-deficient background to efficiently generate and convert heterozygous mutations to homozygosity in parallel on a genome-wide scale.

The first part of this thesis describes the establishment of a novel mutagenic strategy based on remobilisation of a single copy *piggyBac* transposon targeted within the genome. This strategy has the significant advantage over conventional co-transfection methods for transposon delivery in their ease of mutant library construction while maintaining a single copy of the mutagen per cell for the subsequent establishment of the genotype-phenotype causality.

The second part of this thesis concerns the development of reporter systems to conduct genetic screens using the established mutant pools, for the identification of novel factors in the miRNA biogenesis pathways and dissecting possible differential regulators in the two branches of the miRNA downstream effector pathways, i.e. miRNA mediated mRNA decay and translational repression. To our knowledge, this is the first attempt to use non-hypothesis driven genetic approach to identify novel components in this pathway in mammals. Preliminary screening with one of the reporter system has revealed a homozygous mutant in a known effector in the miRNA-mediated repression, *Ago2*.

A final part of this thesis presents a separate part of the research during the PhD to advance the *piggyBac* transposon technology in large genomic DNA delivery. This work has demonstrated a giant cargo capacity of up to 100 kb for *piggyBac* transposons. The integrations of giant *piggyBac* transposons are intact, they can be expressed stably and can be remobilised from the genome. Giant *piggyBac* transposons open new doors to many applications in basic mammalian genetics and gene therapy, which are not possible with existing methods.

Contents

Abbreviations	xi
List of figures	xiii
List of tables	xvi

Chapter 1 – Introduction

Chapter overview	1
1. Reverse genetics.....	2
1.1. Homologous recombination based gene targeting.....	2
1.2. Reverse genetics using RNA interference.....	3
1.1.3. Zinc finger nuclease mediated genetic alternations	5
2. Forward genetics and different screen designs	6
2.1. Germline and somatic mutagenesis	7
2.2. Dominant, recessive and genetic interaction screens.....	8
2.2.1. Dominant genetic screens.....	8
2.2.2. Recessive genetic screens	9
2.2.2.1. Germline recessive mutagenesis	9
2.2.2.2. Mitotic recombination mediated recessive mutagenesis in the soma	11
2.2.3. Genetic-interaction screens	14
3. Means of mutagenesis for forward genetics screens	17
3.1. Chemical agents.....	17
3.2. Physical agents.....	19
3.3. Biological insertional agents.....	19
3.3.1. Retroviral vectors	19
3.3.2. DNA transposons.....	20
3.3.2.1. <i>Sleeping Beauty</i> and <i>piggyBac</i> possess different characteristics	22
3.3.2.2. Transposon-mediated germline mutagenesis in mice	24
3.3.2.3. Transposon-mediated somatic mutagenesis for cancer gene discovery	26
3.4. Insertional mutagen designs	29

3.4.1. Loss-of-function designs.....	29
3.4.1.1. Promoter-trap designs.....	29
3.4.1.2. PolyA-trap designs.....	31
3.4.2. Gain-of-function designs.....	32
4. Mammalian cells as genetic models.....	35
4.1. The mouse and rat as mammalian model experimental organisms.....	35
4.2. Mammalian cells as experimental models.....	36
4.3. Mouse ES cells as an attractive mammalian cell-based model.....	37
5. Strategies for recessive genetic screens in mammalian cell culture.....	39
5.1. Loss-of-heterozygosity-based strategies.....	39
5.1.1. Induced mitotic recombination using Cre/ <i>loxP</i> system.....	40
5.1.2. High G418 selection.....	42
5.1.3. <i>Blm</i> -deficient ES cell system and further developments.....	43
5.2. Haploid mammalian cell lines for recessive genetic screens.....	49
6. miRNAs and their biogenesis pathways.....	50
6.1. The discovery of miRNAs.....	50
6.2. miRNAs and siRNAs.....	52
6.3. miRNA biogenesis.....	53
6.3.1. The canonical biogenesis pathway.....	53
6.3.2. Different roles of Dicer homologues.....	55
6.3.3. Strand selection of the miRNA: miRNA* duplex.....	56
6.3.4. Choice of Argonaute association.....	56
6.3.5. Non-Canonical biogenesis pathways.....	57
6.4. Regulation of miRNA biogenesis.....	58
6.5. Wider implications of miRNA biogenesis.....	60
6.5.1. miRNA biogenesis and cancer.....	60
6.5.2. Hijacking miRNA biogenesis by viruses.....	61
6.5.3. miRNA biogenesis pathway mutants as tools to studying miRNA functions.....	62
7. Thesis project design.....	64

Chapter 2 – Materials and methods

1. Vectors.....	67
1.1. PB transposon and transposase vectors	67
1.2. PB transposon targeting vectors	69
1.3. Vectors miRNA reporters and their targeting vectors	72
1.4. BAC vectors for giant PB transposons	77
2. ES cell lines	79
2.1. Genotype of the precursor ES cell lines	79
2.2. ES cell lines generated.....	79
3. Media and chemicals used for ES cell culture.....	81
4. AB1, AB2.2 and B6 derived ES cell culture and manipulations.....	82
4.1. Passaging, freezing down and thawing ES cells.....	82
4.2. ES cell transfection by electroporation	83
4.3. ES cell transfection by lipofection and pulse puromycin selection.....	83
4.4. Picking ES cell colonies	84
4.5. Cre or Flp mediated recombination to pop-out selection cassettes.....	84
5. <i>piggyBac</i> mobilisation in ES cells.....	84
5.1. Plasmid-to-genome mobilisation	84
5.2. Intra-genomic <i>piggyBac</i> mobilisation.....	85
5.3. <i>piggyBac</i> Transposon excision for phenotypic reversion.....	86
6. Homozygote mutant generation using <i>Blm</i> -deficient ES cells.....	86
7. FACs analysis of eGFP expressing ES cells	87
8. Cytogenetic analysis.....	87
8.1. Preparation of metaphase spread.....	87
8.2. Sister chromatid exchange analysis.....	88
9. DNA methods	89
9.1. Recombineering technology.....	89
9.1.1. <i>E. coli</i> strains	89
9.1.2. Transformation of <i>E. coli</i> by electroporation.....	90

9.1.3. DNA fragment preparation for recombineering.....	90
9.1.4. Recombineering procedure	91
9.1.5. Cre or Flp mediated cassette pop-out in <i>E. coli</i>	91
9.2. Southern blotting and hybridization	92
9.3. Southern blotting probes used in this thesis.....	93
9.4. Splinkerette PCR.....	94
9.4.1. Adaptor preparation	94
9.4.2. Genomic DNA digestion and adaptor ligation	96
9.4.3. Nested-PCR amplification	96
9.5. Genomic DNA triple-primer competition PCR.....	97
10. RNA method: RT-PCR	98
11. Protein method: western blotting	101
11.1. Whole-cell protein extraction	101
11.2. Protein blotting and antibody hybridisation	101
11.3. Antibodies used in this thesis.....	102
12. Regional high density Genomic comparative hybridisation (CGH) array.....	102
13. Illumina sequencing for PB integration sites analysis.....	103
14. Bioinformatic analysis of Illumina sequencing data.....	105

Chapter 3 – Generation of an efficient mutagenic strategy using *piggyBac*

1. Introduction.....	107
1.1. Design principles of the mutagen.....	108
1.2. DNA mismatch repair pathway as a screening model for proof-of-principle	110
2. Results	113
2.1. Mutagenic exon-pair selection.....	113
2.2. Validations of mutagen units using an <i>Hprt</i> trapping assay.....	115
2.2.1. <i>Hprt</i> trapping assay with 6-TG and G418 dual selection	117
2.2.2. 4-OHT induction optimisation	118
2.2.3. <i>Hprt</i> trapping assay with sequential 6-TG and G418 selection	122

2.3. Validations of mutagen units using a random trapping assay	125
2.3.1. Generation of a cell line with an autosomal single-copy PB transposon	126
2.3.2. The random trapping assay	127
2.4. Local hopping of PB around the <i>Gdf9</i> locus	131
2.5. Proof-of-principle of my mutagenic strategy using DNA mismatch repair screen	133
2.5.1. PB reintegration efficiency estimation in NN5-Gdf9 ^{hprtminiPB/+} cells.....	133
2.5.2. Library construction and screening	134
2.5.3. Mutant analysis and validation.....	138
3. Discussion	141
3.1. Molecular design of the gene-inactivating PB transposon	141
3.2. <i>piggyBac</i> possesses a fast transposition kinetics	142
3.3. Local hopping characteristics of <i>piggyBac</i>	144
3.4. A recessive genetic screen using the established mutagenesis strategy.....	146
3.5. Complete genotype reversion using PBase with FIAU selection.....	147

Chapter 4 – Generation of a *Blm*-deficient ES cell line with single-copy PB in *Hprt* locus

1. Introduction.....	149
2. Results	149
2.1. Generation of a new <i>Blm</i> -deficient ES cell line	149
2.2. Phenotypic assessments of the <i>Blm</i> ^{e/e} cell line	153
2.3. Gene targeting of the mutagenic PB to the <i>Hprt</i> locus in <i>Blm</i> ^{e/e} cells	154
2.4. PB-remobilisation assessment of the newly generated cell lines.....	156
3. Discussion	159

Chapter 5 – Development of reporter systems for probing the miRNA pathway

1. Introduction.....	162
1.1. Why studying the miRNA biogenesis and effector pathways	162
1.2. Challenges in using <i>Blm</i> -deficient ES cell system to study the miRNA pathway.....	164
2. Results	164

2.1. Development of the artificial miR-eGFP system	166
2.1.1. Generation of an eGFP knock-in reporter cell line	166
2.1.2. miR-30-based miR-eGFP generation and analysis	167
2.1.3. Construct optimisation for the artificial miR-eGFP construct	171
2.1.4. Generation of an artificial miR-eGFP knockin in the <i>Blm^{e/e}Hprt^{PBin2}</i> cell line	175
2.1.5. Validation of the miR-eGFP knockin <i>Blm^{e/e}Hprt^{PBin2}</i> cell line.....	177
2.2. Development of an endogenous miRNA target reporter.....	179
2.2.1. Artificial ESCC-miRNA target generation	180
2.2.2. Generation of an ESCC-miRNA target reporter knockin <i>Blm^{e/e}Hprt^{PBin2}</i> cell line	186
3. Discussion.....	189

Chapter 6 – Preliminary screen and condition optimisation for the miR-eGFP system

1. Introduction.....	193
2. Results	194
2.1. Generation of the mutant library	194
2.2. Screening strategies and optimisation	196
2.3. Mutant validation	202
3. Discussion.....	204
3.1. Future improvements on the miR-eGFP screening system.....	204
3.2. <i>Ago2</i> , Argonaute proteins and the small RNA mediated pathways.....	206

Chapter 7 – Mobilisation of giant *PiggyBac* transposons in the mouse genome

1. Introduction.....	210
2. Results	211
2.1. The generation of giant PB transposons	211
2.2. Transposition detection of giant PB transposons in ES cells.....	213
2.3. Individual clone analysis of giant PB genomic integrations	218
2.4. Analysis of chromosomal excision capability of giant PB.....	221
3. Discussion.....	223

Chapter 8 – Conclusions and future work

8.1. Conclusions..... 225

8.2. Future work..... 226

References 229

Abbreviations

°C	degrees centigrade
4-OHT	4-hydroxyltamoxifen
6-TG	6-thioguanine
Ago	Argonaute
BAC	bacterial artificial chromosome
<i>Blm</i>	Bloom syndrome gene
bp	base pair
<i>bghpA</i>	bovine growth hormone poly-adenylation signal
BrdU	5-bromo-2-deoxyuridine
<i>Bsd</i>	blasticidin resistant gene
CAG	the hybrid cytomegalovirus enhancer/chicken β actin promoter
CMV	cytomegalovirus promoter
CGH	comparative genomic hybridisation
DMSO	dimethylsulfoxide
ESCC	ES cell-specific cell cycle-regulating
FIAU	1-(2-deoxy-2-fluoro-1- D-arabinofuranosyl)-5-iodouracil
<i>Huc</i>	human ubiquitin C promoter
<i>IRES</i>	Internal ribosome entry site
LB	Luria-Bertani broth
LOH	loss of heterozygosity
miRNA	microRNA
MMR	mismatch repair
MOPS	3-morpholinopropane-1-sulfonic acid, C ₇ H ₁₅ NO ₄ S
<i>Neo</i>	neomycin resistant gene
nt	nucleotides
o/n	over night
PB	<i>piggyBac</i> transposon
PB5 or PB3	<i>piggyBac</i> 5' (or 3') inverted terminal repeat

PBase	<i>piggyBac</i> transposase
PB ITRs	<i>piggyBac</i> transposon inverted terminal repeats
PBS	Phosphate buffered saline
PCR	polymerase chain reaction
<i>PGK</i>	phosphoglycerate kinase
Pri-miRNA	Primary microRNA transcript
Pre-miRNA	Precursor microRNA transcript
<i>Puro</i>	puromycin resistant gene
<i>PuroΔtk</i>	fusion gene of puromycin <i>N</i> -acetyltransferase gene (<i>puro</i>) and truncated herpes simplex viral type 1 thymidine kinase gene (Δtk)
<i>rbgpA</i>	rabbit β -globin poly-adenylation signal
RISC	RNA-induced silencing complex
RNAi	RNA interference
r.p.m	revolutions per minute
RT-PCR	reverse transcripton- PCR
SCE	sister chromatid exchange
SDS	Sodium lauryl sulfate, $C_{12}H_{25}SO_4Na$
siRNA	small interfering RNA
TBS	Tris buffered saline

List of figures

Chapter 1

Figure 1-1: Germline and somatic mutagenesis	11
Figure 1-2: Somatic mosaic recessive screens in <i>Drosophila</i> eye	13
Figure 1-3: Synthetic suppression screens.....	15
Figure 1-4: Synthetic enhancement screens.....	17
Figure 1-5: Transposon-mediated mutagenesis	25
Figure 1-6: Molecular designs for loss- and gain-of-function insertional mutagens.....	34
Figure 1-7: Mitotic recombination following G2-X and G2-Z segregation.....	41
Figure 1-8: Generation of homozygous mutant from heterozygous counterparts.....	46
Figure 1-9: Four steps involved in conducting recessive genetic screens using <i>Blm</i> -deficient ES Cells	47
Figure 1-10: Canonical miRNA and endo-siRNA biogenesis pathway.....	55
Figure 1-11: Mirtron biogenesis pathway, bypassing the <i>Drosha</i> processing step	58
Figure 1-12: Schematic representation of the mutagenic strategy employed for this project...	65
Figure 1-13: A schematic representation of a reporter strategy to screen for miRNA biogenesis mutants	66

Chapter 2

Figure2-1: Schematic representation of the Splinkerette-PCR method	95
--	----

Chapter 3

Figure 3-1: Design of an inactivating PB transposon targeted into the <i>Hprt</i> locus of <i>Blm</i> -deficient ES cells.	107
Figure 3-2: The mechanistic basis of 6-TG mediated cytotoxicity through MMR system.....	112
Figure 3-3: Mutagen <i>Dom3z</i> and <i>Ccdc107</i> exon-pair structures	115
Figure 3-4: Schematic representations of the <i>Hprt</i> trapping assay	116
Figure 3-5: Reversion analysis of the 6-TG resistant colonies	118
Figure 3-6: 4-OHT induction protocol optimisation.....	121

Figure 5-9: An endogenous miRNA-based reporter system	180
Figure 5-10: ESCC-miRNA eGFP reporter analysis.....	181
Figure 5-11: <i>Dgcr8</i> ^{gt1/tm1} ES cell validation	183
Figure 5-12: Causality establishment of the ES cell miRNAs and the ESCC-miRNA reporter	184
Figure 5-13: ESCC-miRNA Neo reporter analysis	186
Figure 5-14: The <i>mir384</i> -locus targeting with the ESCC-miRNA reporter construct	188

Chapter 6

Figure 6-1: A hypothetical estimation of generating homozygous mutant in miRNA pathway..	194
Figure 6-2: Schematic representation of the mutant library construction.....	195
Figure 6-3: Screening strategies for the miR-eGFPsystem.....	196
Figure 6-4: Southern detection for the PB transposon reintegration patterns.....	201
Figure 6-5: <i>Ago2</i> mutant molecular analysis	203
Figure 6-6: Multiple roles of <i>Ago2</i> in the non-coding small RNA biogenesis, regulation and effector pathways	209

Chapter 7

Figure 7-1: Giant PB construction and selection scheme incorporated to detect their transposition in mouse ES cells.....	212
Figure 7-2: Transfection efficiency of the AB2.2 ES cells determined by the eGFP expression after pulse-puromycin selection	213
Figure 7-3: Outline of the experimental scheme	214
Figure 7-4: Schematic representation of the experimental platform and bioinformatics analysis to identify transposition events using the Illumina sequencing	217
Figure 7-5: Giant PB transposition with single copy integration per cell	219
Figure 7-6: Regional high density CGH array analysis to determine the PB cargo integrity	220

List of tables

Table 2-1: primers used to construct giant PB transposons	79
Table 2-2: A summary of ES cell lines generated for this thesis work	80
Table 2-3: Southern blotting probes used in this thesis work.	93
Table 2-4: Locus-specific primers used for the random trapping assay	100
Table 2-5: Antibodies used in this thesis work	102
Table 2-6: Primers used for the multiplex Illumina sequencing to identify PB integration sites	105
Table 3-1: Estimate of number of cells required to obtain <i>Hprt</i> trapping events	119
Table 3-2: Ten PB reintegration clones selected from the random trapping	128
Table 3-3: Summary of seven independent integration sites identified from the MMR screen.	137
Table 3-4: Local-hopping comparison among different genomic loci	144
Table 5-1: Summary of different miR-eGFP constructs generated.....	173
Table 6-1: Efficiency comparison of screening protocols to identify eGFP-positive clones	200
Table 6-2: Summary of the identities of the eGFP-positive clones.....	200
Table 7-1: Transposition efficiency of different sized PB transposons and versions of the PBase	216

Chapter one – Introduction

Chapter overview

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

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

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

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

Reverse and forward genetics in experimental model organisms

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

1. Reverse genetics

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

1.1. Homologous recombination-based gene targeting

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

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

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

1.2. Reverse genetics using RNAi

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

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

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

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

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

1.3. Zinc finger nuclease-mediated genetic alternations

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

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

2. Forward genetics and different screen designs

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

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

2.1. Germline and somatic mutagenesis

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

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

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

2.2. Dominant, recessive and genetic interaction screens

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

2.2.1. Dominant genetic screens

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

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

2.2.2. Recessive genetic screens

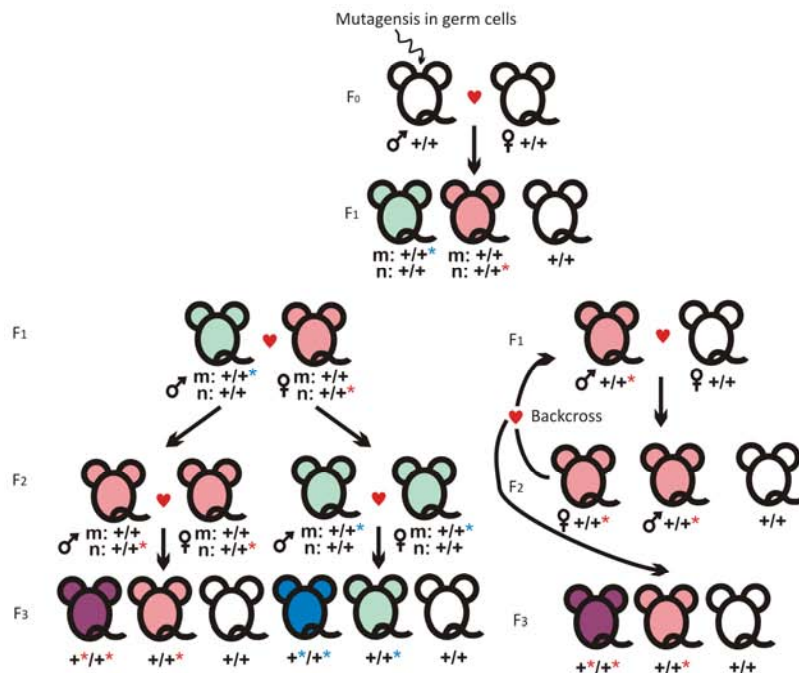
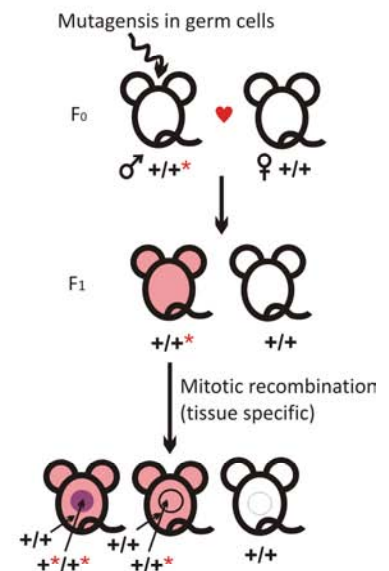
Recessive genetic screens are designed to isolate genes which show a phenotype when inactivated and this often requires the inactivation of all copies of the gene in a given genome to evoke a phenotypic change. In yeast, recessive genetic screen is more feasible to conduct than other organisms with the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

2.2.2.1. Germline recessive mutagenesis

Early screens used chemical agents such as *N*-ethyl-*N*-nitrosourea (ENU) and Ethyl methane sulphonate (EMS) to efficiently mutagenise the genome of an organism, mainly generating loss-of-function mutations. When mutations occur in the germline of the organism, it can be passed on to the offspring. By crossing the mutagenised animal with a wild-type animal, the F₁ offspring will be heterozygous mutants of different loci depending on the mutation spectrum present in each germ cell. Further mating can be conducted in two ways. The first method is to inter-cross F₁ animals in order to introduce more mutations into the pedigree,

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

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

Figure 1-1 : Germline and somatic mutagenesis.**a Germline mutagenesis****b Somatic mutagenesis**

Both mutagenesis strategies are illustrated with mice, but they are universal to all diploid multi-cellular models. The colour scheme reflects the genetic status with white represents wild-type, light pink and dark pink represent heterozygous and homozygous. The red star represents a mutation. For both strategies, only one germline mutation is illustrated here and one can imagine that each sperm from the mutagenised F₀ male can carry different mutations; therefore all heterozygous F₁ animals will have different mutations. a, germline mutagenesis to obtain mutants having identical genotype throughout the bodies. It involves mutagenising the sperms of F₀ males and heterozygous whole-body mutants can be derived in F₁ generations. There are two ways to obtain homozygous mutants. The first way (bottom left panel) is to intercross F₁ animals and in this way, more mutations can be introduced into the pedigree. The second way (bottom right panel) is to cross F₁ males with wild-type females to produce heterozygous F₂ females and these females are backcrossed to the F₁ heterozygous males to produce homozygous animals in F₃ generation. In this crossing, most of the mutations in F₁ males can be converted to homozygosity in F₃. b, somatic mutagenesis with somatic mosaic mutants. After obtaining F₁ heterozygous mutants, mitotic recombination can be induced somatically (or/and spatially), clones of homozygous somatic cells can arise in otherwise heterozygous background.

2.2.2.2. Mitotic recombination-mediated recessive mutagenesis in the soma

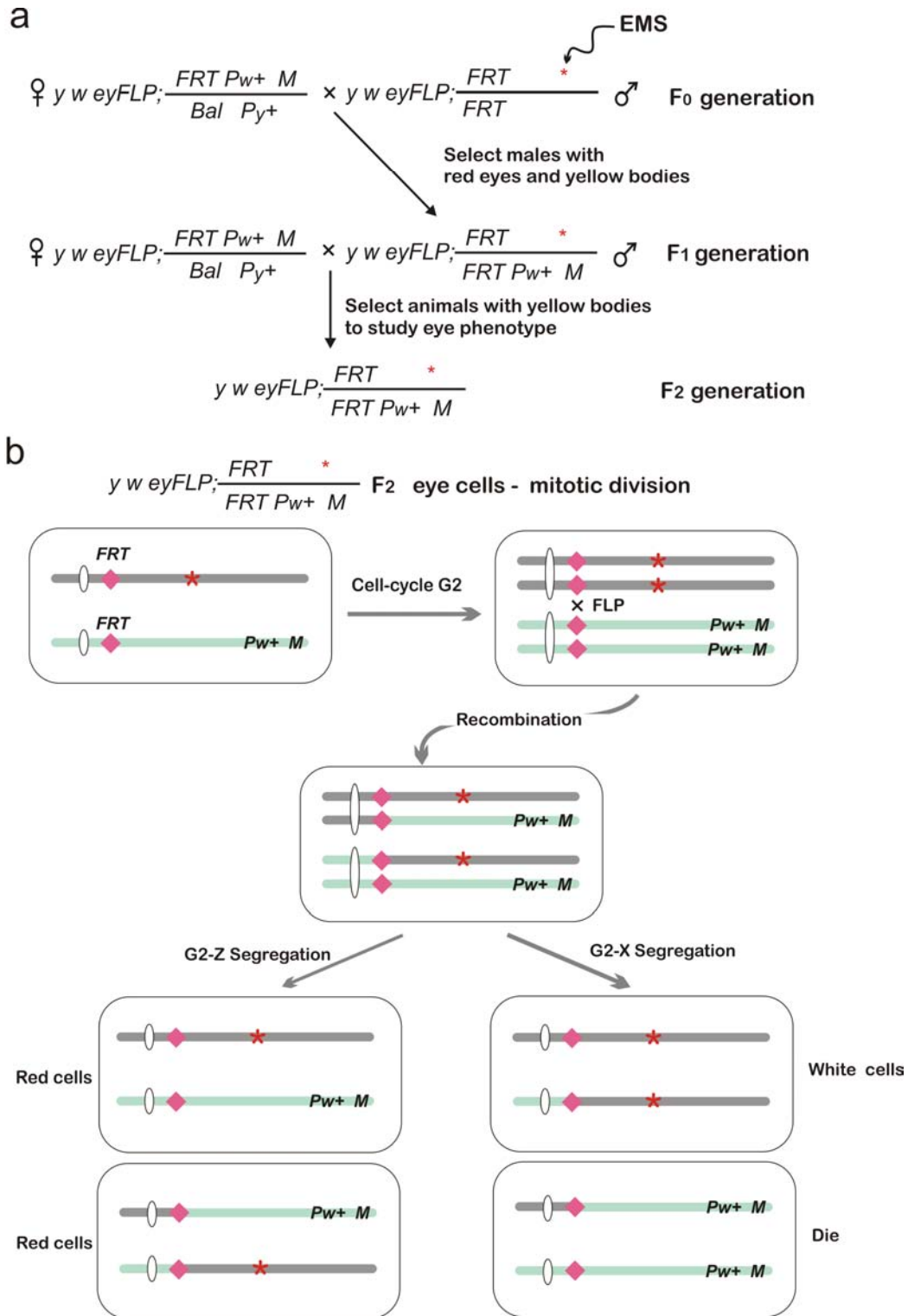
Another *in vivo* approach to conduct recessive genetic screens in diploid organisms is through the generation of somatic mosaics with homozygous daughter cells produced from a heterozygous genotype by mitotic recombination, Figure 1-1b. Mitotic recombination is a

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

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

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

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

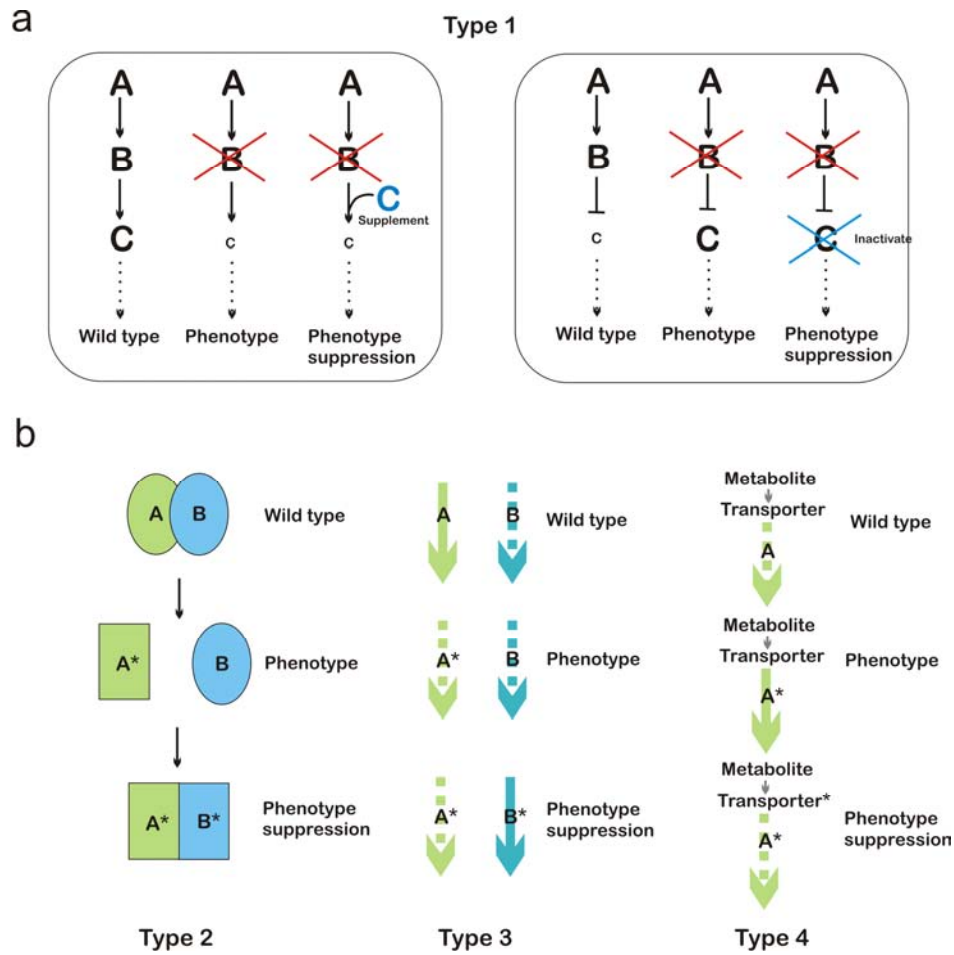


This system can also speed up the recessive genetic screens compared to conventional germline recessive mutagenesis screens, as the screen can be conducted in the somatic tissues of an F1 generation rather than in the F3 generation in a conventional scheme. The Flp can be expressed conditionally and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

2.2.3. Genetic-interaction screens

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

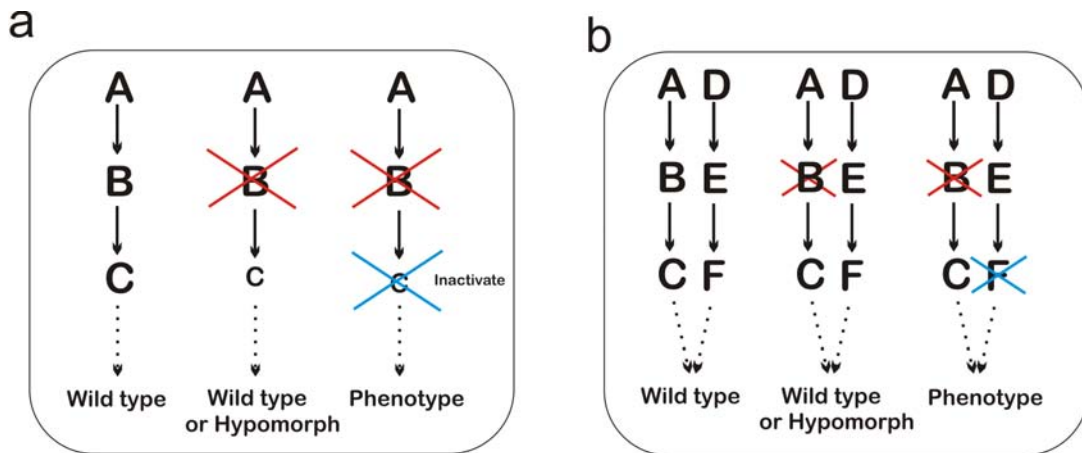
Figure 1-3: Synthetic suppression screens.



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

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

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

Figure 1-4: Synthetic enhancement screens.

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

3. Means of mutagenesis for forward genetics

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

3.1. Chemical agents

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

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

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

3.2. Physical agents

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

3.3. Biological insertional agents

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

3.3.1. Retroviral vectors

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

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

3.3.2. DNA transposons

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

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

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

The lack of active DNA transposons in mammals hindered their application of insertional mutagenesis in experimental organisms such as the mouse and the rat using existing strategies developed in other organisms. In 1997, the first mammalian-active DNA transposon, *Sleeping Beauty* (SB) a member of the Tc1/Mariner family, was re-activated based on “ancient” sequences found in fish (Ivics et al., 1997). Since then, the mammalian DNA transposon toolkit has been expanded by the discovery and development of several members

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

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

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

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

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

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

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

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

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

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

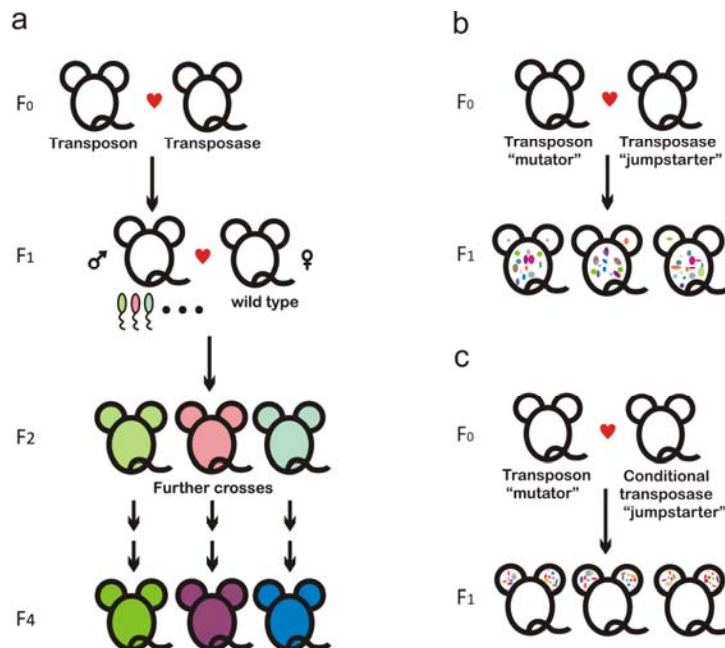
3.3.2.2. Transposon-mediated germline mutagenesis in mice

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

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

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

Figure 1-5: Transposon-mediated mutagenesis.



a, Transposon-mediated germline mutagenesis. The colour scheme reflects the genetic status with white represents wild-type, light colours and corresponding dark colours represent heterozygous and homozygous. Each sperm derived from the F₁ males carries different transposon integration sites. b,c, Transposon-mediated somatic mutagenesis, with b represents a whole-body somatic mutagenesis and c illustrate a tissue-specific mutagenesis. The coloured circles represent cells with the same clonal origins with independent transposon integration sites.

3.3.2.3. Transposon-mediated somatic mutagenesis for cancer gene discovery

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

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

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

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

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

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

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

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

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

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

3.4. Insertional mutagen designs

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

3.4.1. Loss-of-function designs

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

3.4.1.1. Promoter-trap designs

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

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

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

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

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

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

3.4.1.2. PolyA-trap designs

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

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

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

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

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

3.4.2. Gain-of-function designs

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

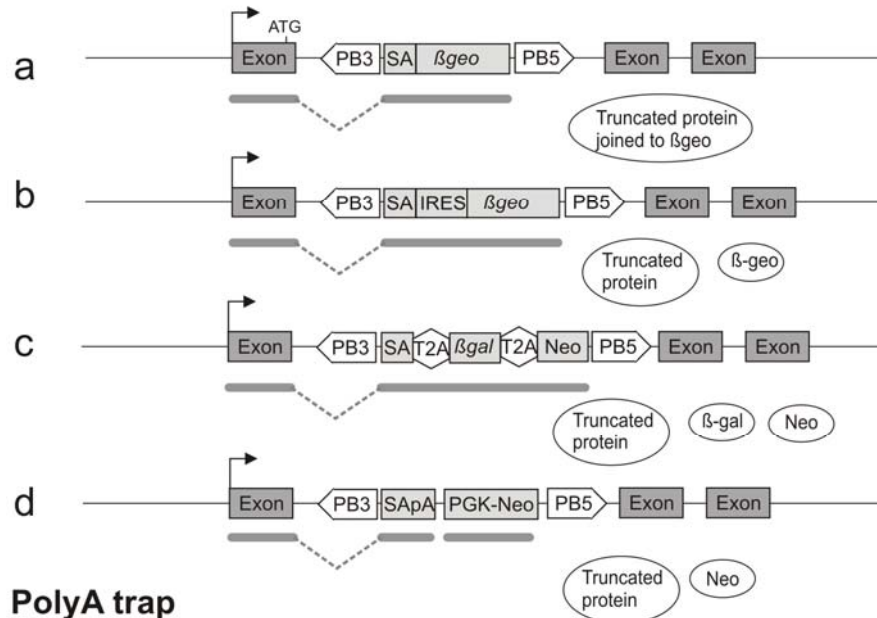
alternative method is to over-express genes from a cDNA expression cassette. A single cDNA or a group of cDNAs connected by 2A peptides can be engineered into one vector, Figure 1-6i. 2A peptides provide a level of expression which is equivalent for all of the individual cDNAs (Donnelly et al., 2001).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

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

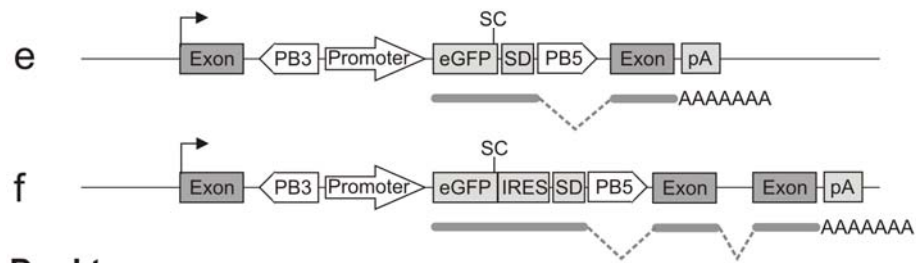
All the designs are illustrated in the context of a PB transposon based vector. SA, splice acceptor; pA, polyadenylation signal; β -geo, β -galactosidase and neomycin resistant gene fusion gene; β -gal, β -galactosidase; PGK, mouse phosphoglycerate kinase promoter; SC, stop codon; IRES, internal ribosome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

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

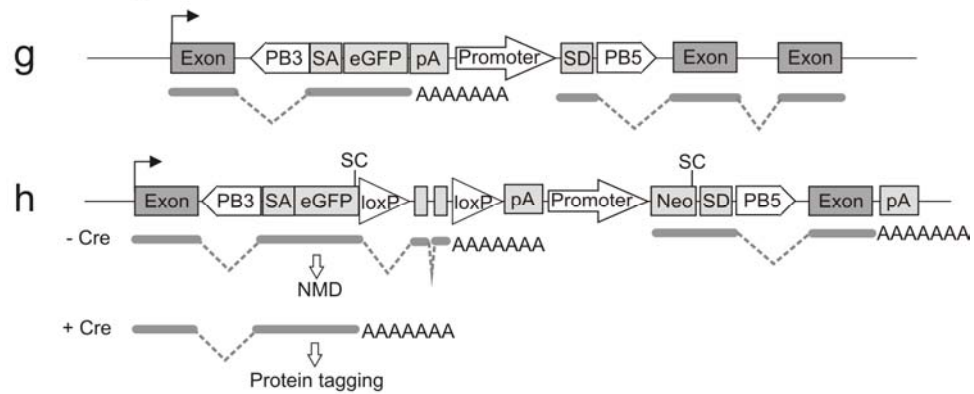
Promoter trap



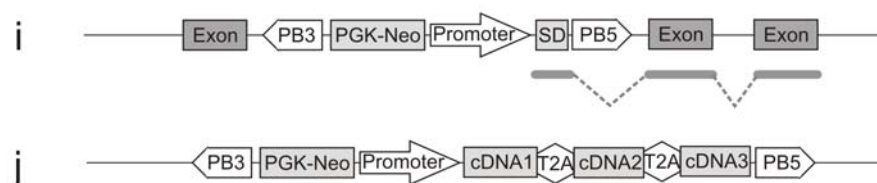
PolyA trap



Dual trap



Gain-of-Function design



4. Mammalian cells as genetic models

4.1. The mouse and rat as mammalian model experimental organisms

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

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

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

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

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

4.2. Mammalian cells as experimental models

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

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

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

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

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

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

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

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

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

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

5. Strategies for recessive genetic screens in mouse ES cells

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

5.1. Loss of heterozygosity based strategies

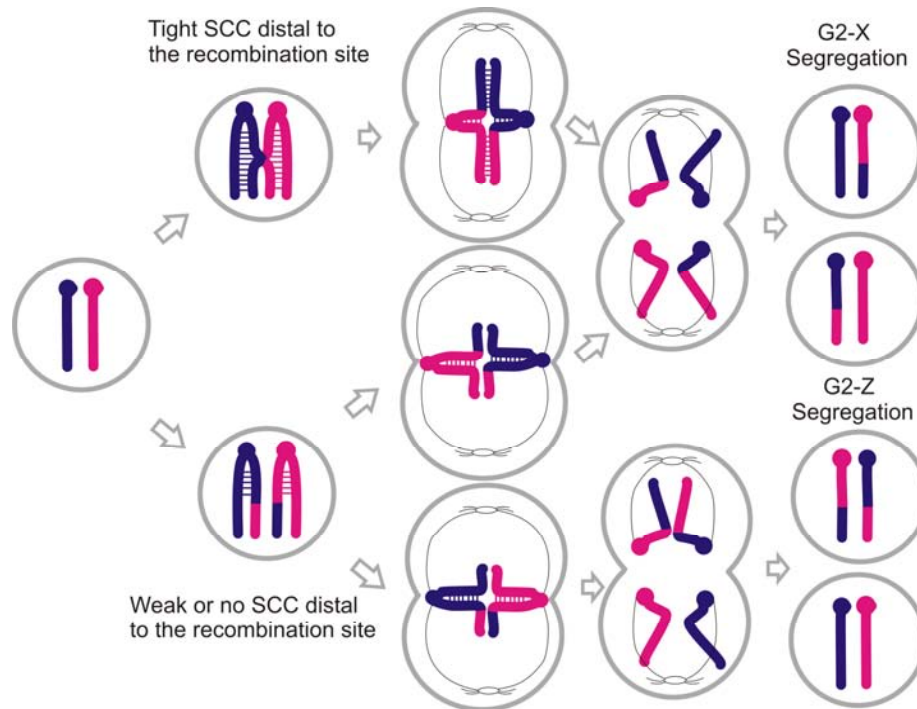
Several methods have been developed which exploit the naturally occurring phenomenon of loss of heterozygosity (LOH), to recover rare homozygous mutant cells from heterozygous cells (Mortensen *et al.*, 1992; Lefebvre *et al.*, 2001; Guo, 2004; Yusa *et al.*, 2004a). In addition, the LOH events can be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

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

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

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

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

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

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

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

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

5.1.2. High G418 selection

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

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

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

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

5.1.3. *Blm*-deficient ES cell system

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

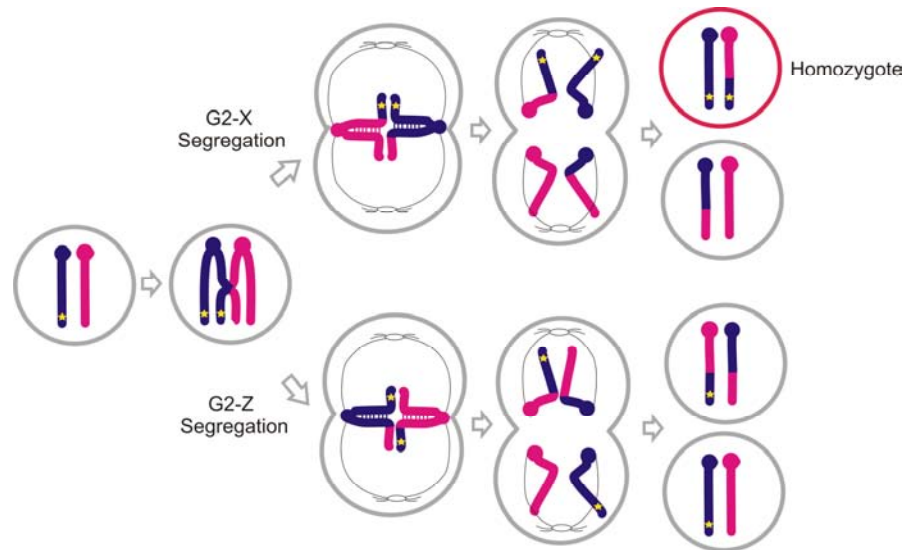
Blm-deficient mouse models recapitulate the phenotypes observed in human patients (Luo et al., 2000) and six different *Blm* mouse knockout alleles have been generated, namely, *Blm*^{tm1Brd}, *Blm*^{tm2Brd}, *Blm*^{tm3Brd}, *Blm*^{tmChes1}, *Blm*^{tmChes3}, *Blm*^{tm1Grd} (Chester et al., 1998; Luo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

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

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

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

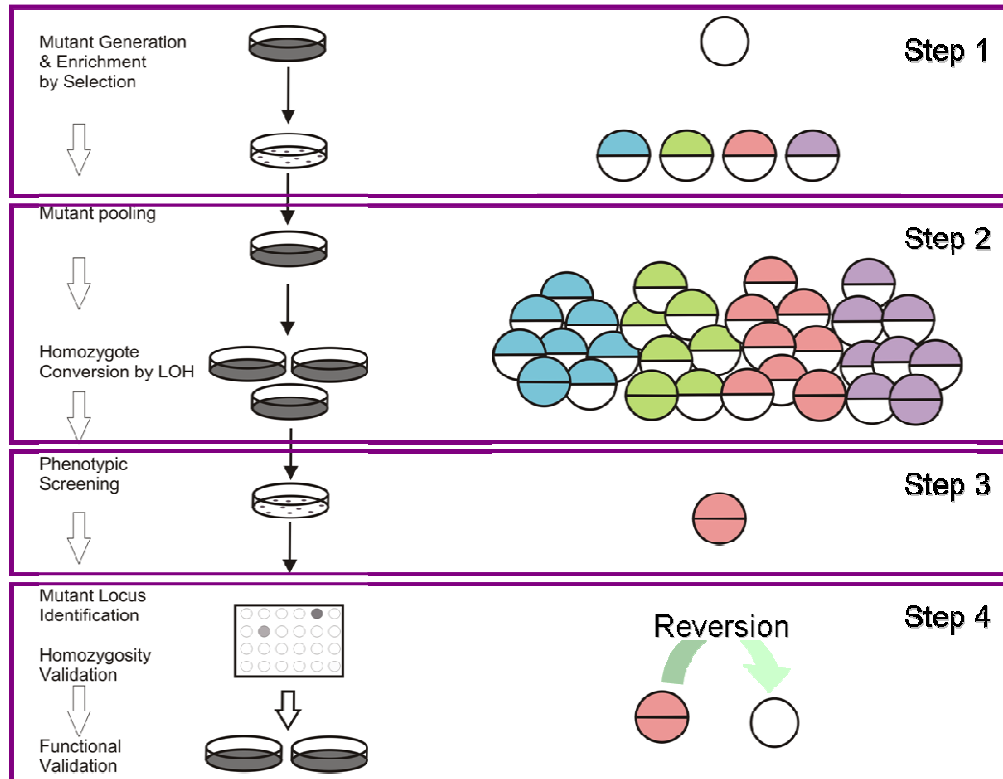
Using a *Blm*-deficient genetic background together with insertional mutagenesis, genome-wide recessive genetic screens have been successfully conducted in ES cells. These screens cover a variety of biological pathways, including DNA mismatch repair, retroviral resistance, RNAi processing and toxin resistance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

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

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

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

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



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

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

A major consideration is the phenotypic read-out in the screen design, as not all screens are suitable in the *Blm*-deficient system in a pooled format. Based on the previous calculation using a *Blm*-deficient background, the ratio of homozygote to heterozygote cells is 1 to 1×10^4 after expansion. This means that the screening method must be sensitive and specific enough to be able to isolate only a few relevant homozygote mutants from a large pool of irrelevant cells. In addition, recessive mutations causing phenotypes involving cell death or which affect growth rates of the relevant clones cannot be conducted using such pools.

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

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

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

5.3. Haploid mammalian cell lines for recessive genetic screens

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

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

6. microRNAs and their biogenesis pathways

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

6.1. The discovery of miRNAs

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

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

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

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

6.2. miRNAs and siRNAs

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

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

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

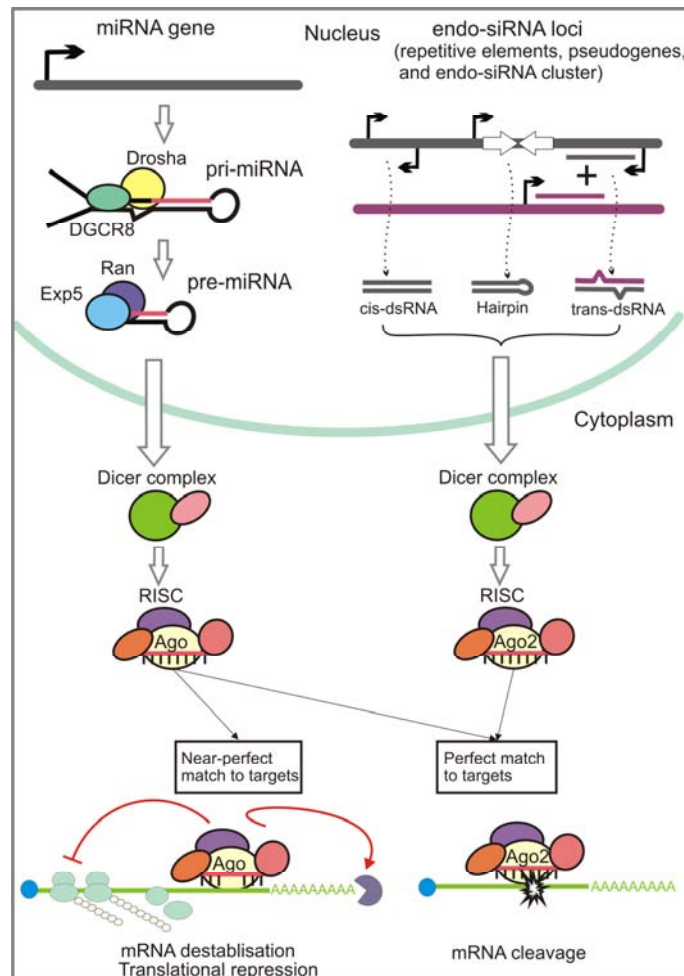
6.3. miRNA biogenesis

6.3.1. The canonical biogenesis pathway

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

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

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

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

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

6.3.2. Differential roles of Dicer homologues

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

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

6.3.3. Strand selection of the miRNA: miRNA* duplex

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

6.3.4. Choice of Argonaute (Ago) association

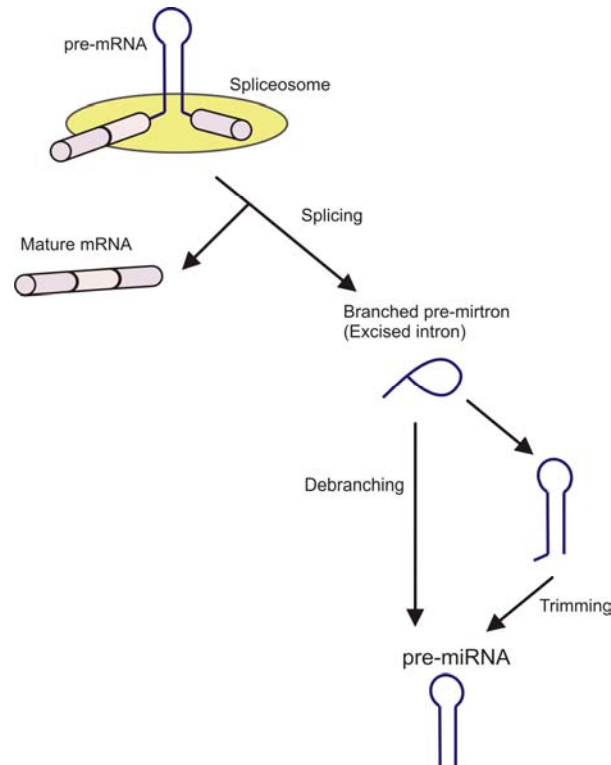
There are several homologues of Ago proteins; Ago1 and Ago2 in *Drosophila* and Ago1, Ago2, Ago3 and Ago4 in mammals. In *Drosophila*, the major factor that determines the sorting of RNA duplexes to two different Ago proteins is the degree of complementarity of the duplex (Forstemann et al., 2007; Steiner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

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

6.3.5. Non-canonical miRNA biogenesis pathways

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

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

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

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

6.4. Regulation of miRNA biogenesis

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

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

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

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

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

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

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

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

6.5. Wider implications of miRNA biogenesis

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

6.5.1. miRNA biogenesis and cancer

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

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

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

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

6.5.2. Hijacking miRNA biogenesis by viruses

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

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

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

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

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

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

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

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

7. Thesis project design

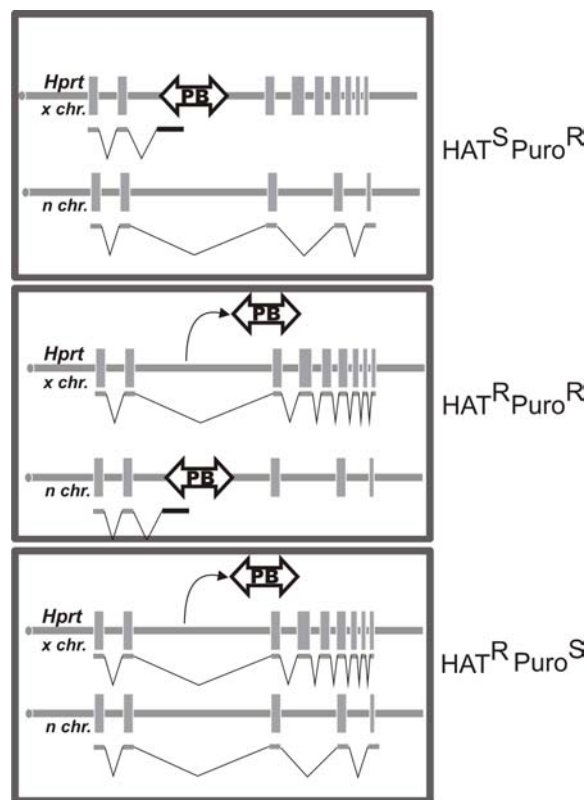
In my thesis project, I have designed a novel mutagen based on the PB transposon in *Blm*-deficient ES cells, with the aim of building a genome-wide mutant library for conducting recessive genetic screens *in vitro*. I applied this strategy using a selectable phenotypic screening to isolate components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

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

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

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

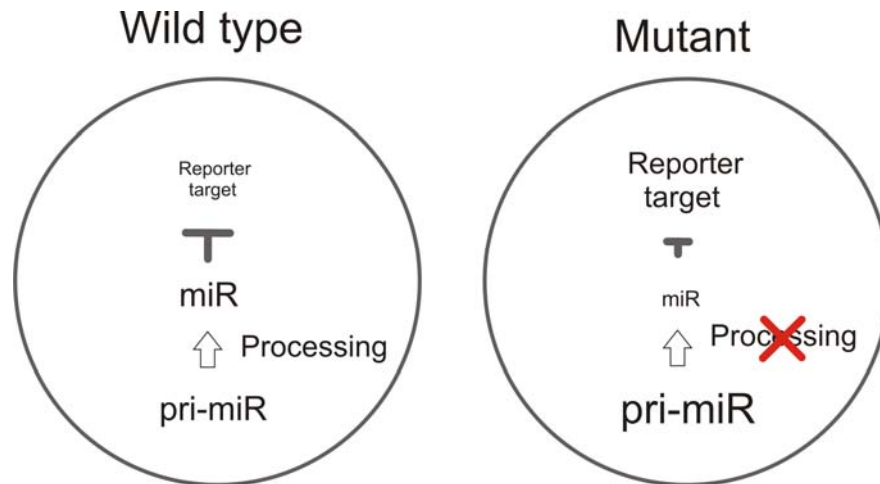
Figure 1-12: Schematic representation of the mutagenic strategy employed for this project.



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

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

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



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

Chapter Two – Materials and methods

1. Vectors

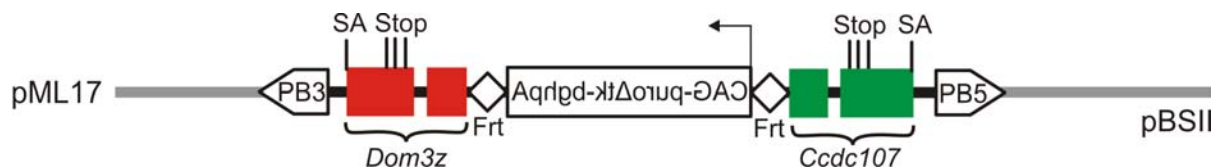
1.1. PB transposon and transposase vectors

piggyBac Transposase plasmids

A CMV promoter-driven mammalian codon-optimised *piggyBac* transposase (mPBase) expressing plasmid was obtained from Juan Cadinanos (CMV-mPBase). The plasmid was modified with the Neomycin resistant cassette removed (pmPBase Δ Neo). A CMV promoter driven hyperactive *piggyBac* transposase (HyPBase) expressing plasmid was obtained from Kosuke Yusa (CMV-hyPBase).

To construct PGK-Puro and PBase co-expressing plasmids, a PGK-Puromycin resistant cassette, the *EcoRI/NotI* fragment from plasmid pPGKPuro, was blunt ligated to the CMV promoter driven mPBase or HyPBase plasmid, *NaeI/SapI* fragment, to give CMV-mPBase-PGK-Puro and CMV-hyPBase-PGK-Puro. The control plasmid CAG-eGFP-bGHpA contains the *PGK-Puro Δ tk* cassette.

Construction of the mutagenic PB (pML17)



The PB inverted terminal repeats (PBITRs) were obtained from Qi Liang. pML6 Plasmid was generated by inserting PCR amplified PB5 ITR, using Bsd-PB5 ITR as template, into *Sall/ECORV* digested vector plasmid pQL2+PB3 (obtained from Qi Liang). The PGK promoter and bgh-PolyA signal were removed from the pML6 by PCR amplification of the EM7-Neo cassette and ligated into *NheI/BamHI*-digested pML6 to generate pML9 plasmid with a floxed EM7-Neo cassette. Two recombineering arms were PCR amplified from the human *HPRT* gene intron 2, 690 bp away from the 3' end of the *hprt* exon 2. The 5' and 3' recombineering arms were

sequentially inserted 5' to the PB3'ITR (*KpnI/XhoI*) and 3' to the PB5'ITR (*BamHI/NotI*) of pML9 to give rise to pML12. A *NotI* site was also introduced at the 5' of the 5' recombineering arm during PCR.

Dom3z and *Ccdc107* last two exons were PCR-amplified from BAC bMQ-365E12 and bMQ-242N12 respectively, and cloned into pML5 using primers Dom3zF and Dom3zR for *Dom3z*-exon cloning and *Ccdc107*F and *Ccdc107*R for *Ccdc107*-exon cloning.

Primers

*Ccdc107*F: GCATTTAGGCCGGCCGAGCCAAGGAGACAGACTGG

*Ccdc107*R: GGAATCGGCGCGCCTTTATTTCCCACTGGATCTT

Dom3zF: GCATTTAGGCCGGCCCAAGTCCTCAGACCCAGTG

Dom3zR: GGAATCGGCGCGCCGAGCCTCTACACCCAGTA

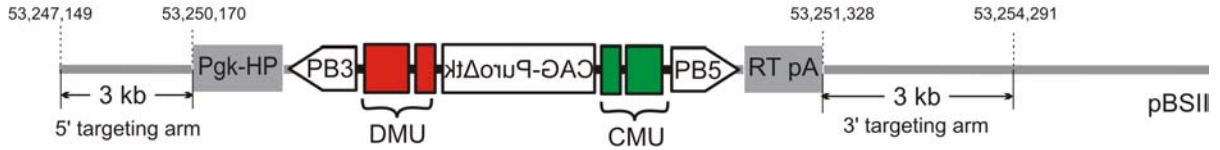
Dom3Z mutagenic unit was extracted from *AgeI/ECOR1* digested pML5-Dom3Z and it was ligated into pML5-*Ccdc107* (*NgoMIV/ECOR1*) to generate a PB transposon with a bidirectional pair of mutagen units pML5-Dom3Z-*Ccdc107*. *EcoRV/SfoI* digestion of pML5-Dom3Z-*Ccdc107* followed with blunt ligation was done to remove the residual Neo cassette in the vector. Additional stop codons were introduced to two other reading frames for both Dom3Z and *Ccdc107* penultimate exons by site directed mutagenesis.

An *FseI/SacII* fragment containing the pair of mutagen units from pML5-Dom3Z-*Ccdc107* was blunt ligated into *FseI/ECOR1* digested vector pML12 to give rise to pML13. *NheI/Agel* digested pML10+FRT plasmid containing the FRT flanked CAGG-Puro Δ TK cassette was blunt ligated into *AscI*-linearised vector pML13 to complete all the modules in the PB transposon and to give rise to pML17.

1.2. Targeting vectors

Construction of targeting vector Gdf9TVPB (pGDF9T-ML4)

pGDF9T-ML4

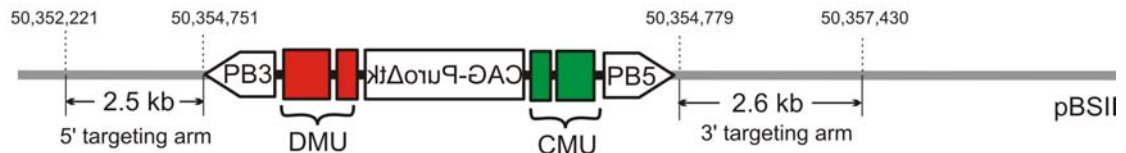


The coordinates are genomic fragments used for the targeting arms and are based on NCBI Build 37.

The *NotI* fragment of pML17 containing the mutagenic PB was used as the linear fragment to insert the PB onto plasmid GDF9T-ML2 using recombineering in the EL350 bacterial strain. The resultant plasmid GDF9T-ML3 was subject to *in vitro* Cre recombinase-mediated excision of floxed kanamycin resistant cassette, giving rise to final targeting vector pGDF9T-ML4. The targeting vector was linearised with *PmeI* before electroporation.

Construction of targeting vector HprtTVIn2PB

HprtTVIn2PB



The coordinates are genomic fragments used for the targeting arms and are based on NCBI Build 37.

The *AscI* site in pML17 was destroyed by digesting the plasmid with *AscI* followed by klenow treatment and self-ligation to give pML18. *XhoI/HindIII* and *HindIII/PstI* fragments from the pML18 were ligated with *XhoI/PstI* fragment of the ARM1 (Haydn Prosser) to give pML20. The hprt targeting arms were extracted from RMCE cassette plasmid (pCEI-3, Haydn Prosser), 2.5 kb *FseI/XhoI* fragment from pCEI-3 containing the hprt 5' homology arm was cloned into *FseI/XhoI* digested pML20 to give phprtTV-left. The 3' homology arm was PCR amplified from pCEI with the KOD system using primers ML45f and ML45r. The PCR fragment was then digested with *PacI* and *AscI* and cloned into phprtTV-left to give the final targeting vector

HprtTVIn2PB. This targeting vector was designed to insert PB transposon into intron 2 of the *Hprt* locus. The homology arms were AB1 origin. The vector was linearised with *Ascl* for gene targeting.

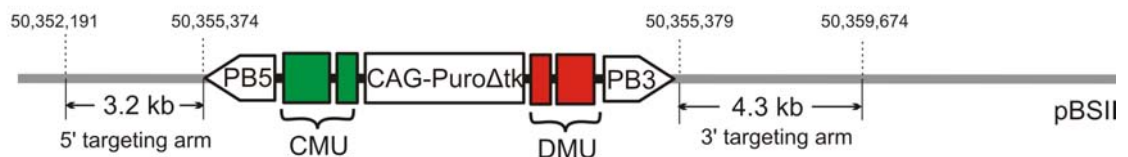
Primers

ML45f: ACCTTAATTAAGATAGATGTTATAGTGTACTCTCCTCTCC

ML45r: AAAGGCGCGCCAGGCACTCAAGATGATCCATATACT

Construction of targeting vector HprtTVE3PB

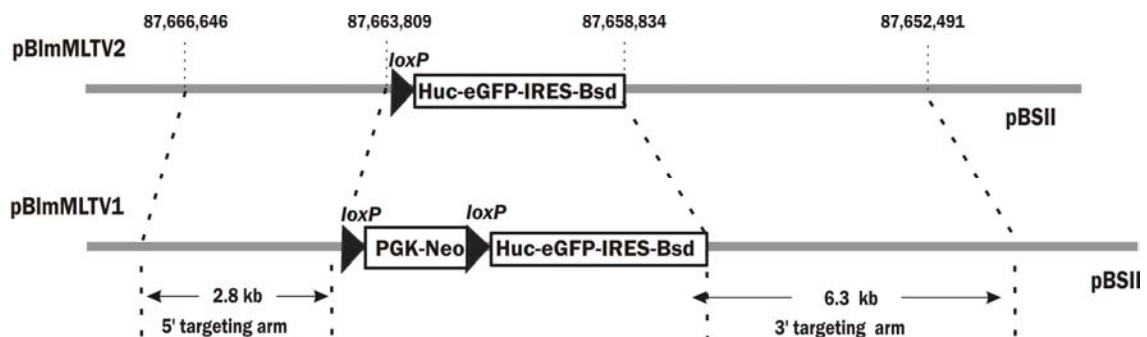
HprtTVE3PB



The coordinates are genomic fragments used for the targeting arms and are based on NCBI Build 37.

The *BspEI/BsiWI* fragment and *BspEI/NsiI* fragment of the pML20 was ligated into *NsiI/BsiWI* fragment of the targeting vector pDTAhpertE3RCAG (Kosuke Yusa). This vector contains the homology arms retrieved from B6 strain BAC, for targeted insertion of PB into a TTAA site within the exon 3 of the *hprt* locus. The vector was linearised with *PmeI* for gene targeting.

Construction for targeting vector BlmTV (BlmTV1 and BlmTV2)



The coordinates are genomic fragments used for the targeting arms and are based on NCBI Build 37.

Human ubiquitin C (Huc) promoter was PCR-amplified from human BAC RP11 214K23 using primers ML67f and ML67r with Phusion PCR system and cloned into pML27 to give pML28. pML38 (pCAG-IRES-Bsd) was made by three-way ligation of the *EcoRI* fragment from pCAG-IRES-Neo, the *EcoRI/NcoI* fragment from pCAG-IRES-Neo and the *EcoRI/NcoI* fragment from PCR amplified Bsd resistant gene from pML313. IRES-Bsd was PCR amplified with ML79f and ML78r and cloned into *BsrGI* site in pPB-Huc-eGFP to give pML40.

B6 strain BAC RP24-180J10 was used to construct the *Blm* targeting vector. Following the *Blm* targeting strategy conducted by Luo and co-workers (Luo et al., 2000), the targeting vector was designed to replace the ATG-containing exon 2 of the *Blm* gene with the *Huc-eGFP-IRES-Bsd* reporter to give rise to null *Blm* alleles. *Blm* mini homology arms for driving recombineering of the reporter construct to the BAC were sequentially PCR amplified and cloned into pML40, flanking the Huc-eGFP-IRES-Bsd together with a floxed PGK-EM7-Neo cassette (pML65). A retrieval vector (pML54) was constructed by PCR amplification of the retroviral homology arms from the *Blm*-containing BAC and subsequent cloning into pBSKSII+. The *AscI/PacI* fragment of pML65 was purified and used for engineering the *Blm*-containing BAC, for replacing the *Blm* exon 2 with the reporter. The correctly targeted BAC-containing bacteria were further used for retrieval of the BAC region containing the 5' and 3' homology arms together with the Huc-eGFP-IRES-Bsd and floxed PGK-EM7-Neo cassette. The retrieval fragment was prepared by linearising pML54 with *SacI*. The correctly retrieved plasmid gave rise to BlmTVML1. The floxed PGK-EM7-Neo cassette was popped out in EL350 and the resulting plasmid became BlmTVML2. Both BlmTVML1 and BlmTVML2 were linearised with *PmeI* before electroporations.

Primers:

ML67f: AATGGCGCGCCATCGATGGCCTCCGCGCCGGGTTTTGGCGCC

ML67r: AATGCTAGCATCGATGTCTAACAAAAAGCCAAAAACGGC

ML79f: GACGAGCTGTACAAGTAACCGCCCCCCCCCTCTCCCTCCCCCCCCCTAACGTTAC

ML78r: AACTGAATTCTGTACATTAGCCCTCCCACACATAACCAGAG

Construction of targeting vector GDF9TV-GFP-IRES-Bsd (pML46)



The coordinates are genomic fragments used for the targeting arms and are based on NCBI Build 37.

The *XhoI/NotI* fragment of pGDF9TML2, which contained the 3' *Gdf9* targeting arm, was ligated into the *Sall/NotI* fragment of the pML40, to give pML43. The 5' *Gdf9* targeting arm was PCR-amplified from the pGDF9ML2 using primers ML83f and ML83r with Phusion PCR system. The PCR product was digested with *Ascl* and *XhoI* and cloned into pML43 to give the final targeting vector pML46. The targeting vector was linearised with *NotI* before electroporation.

Primers:

ML83f: TAAAGGCGCGCCCTGACTCCAGAGCACTCTACTACAT

ML83r: TAAACTCGAGCATGGCAGTCACCCGGTCCAGGTTA

1.3. Vectors for miRNA reporters and their targeting vectors

Construction of miR-eGFP

The mouse miR-155 backbone based mirshRNAs

The mouse miR-155 precursor stem loop surrounding sequences were constructed by annealing oligos ML123f and ML123r. The annealed oligos were treated with Klenow fragment 3'-5' exo^- (NEB) to form the full mirshRNA unit with the the original miR-155 BIC 134–283 region replaced with a synthetic polylinker containing two inverted *Bbs1* sites (Chung et al., 2006). The treated oligos were then digested with *MluI* and *EagI* and ligated into pBS-LR digested with the same enzymes, to give pML64.

The target specific sequences recognising eGFP were constructed by annealing of two 64 nt oligos. The eGFP targeting sequences were either from published work or designed using

BLOCK-iT™ Pol II miR RNAi Designer (Invitrogen, online software). Five different eGFP sequences were constructed in total using five sets of oligos, ML124f and ML124r, ML125f and ML125r, ML126f and ML126r, ML127f and ML127r, and ML128f and ML128r. The miR-155 based the mirshRNA has the effective sequence residing in the 5' strand. The annealed oligos were treated with T4 PNK (NEB) and ligated into Bbs1 digested pML64, to give pML64-124, pML64-125, pML64-126, pML64-127, and pML64-128 (respectively to the oligo sequences). The constructs were verified by DNA sequencing.

The single unit of the target-specific mirshRNA was PCR amplified with four sets of primers (ML131f and ML131r, ML132f and ML132r, ML133f and ML133r, ML134f and ML134r) containing restriction sites that allowed five-way ligation of four copies of the target-specific mirshRNA into *AvrII/NdeI* digested pML29, giving rise to pML79-124, pML79-125, pML79-126, pML79-127, and pML79-128 respectively. The mirshRNAs with the same or different eGFP target sequences were placed within the intron of the human ubiquitin promoter, within the Huc-EM7-Neo-bghpA cassette flanked by PBITRs. The final constructs were verified by DNA sequencing. pML79-128 was further modified by cloning four copies more of the identical mirshRNA to give pML88, which contains a Huc-EM7-Neo cassette with 8 copies of the mirshRNA recognising eGFP polycistronically expressed from the intron 1 of the Huc promoter.

Oligos for constructing the miR-155 backbone:

ML123f:

AATTACGCGTTGGAGGCTTGCTGAAGGCTGTATGCTGTTGTCTTCAAGATCTGGAAGACACAGGACAC
AAGGCCT

ML123r:

AATTCGGCCGTTGTCATCCTCCCACGGTGGCCATTTGTTCCATGTGAGTGCTAGTAACAGGCCTTGTGT
CCTGTG

Target sequence specific oligo pairs:

ML124f: GCTGtagttgtactccagcttgtgcGTTTTGGCCACTGACTGACgcacaagctagtacaactac

ML124r: TCCTgtagttgtactagcttgtgcGTCAGTCAGTGGCCAAAACgcacaagctggagtacaacta

ML125f: GCTGcatgatatagacgttgtggctGTTTTGGCCACTGACTGACagccacaacctatatcatgc

ML125r: TCCTGcatgatataggttgtggctGTCAGTCAGTGGCCAAAACagccacaacctatatcatg

ML126f: GCTGtgaagaagtcgtgctgcttcaGTTTTGGCCACTGACTGACtgaagcagcgacttcttcac

ML126r: TCCTgtgaagaagtcgctgcttcaGTCAGTCAGTGGCCAAAACtgaagcagcacgacttcttca

ML127f: GCTGttagttgtactccagcttgtGTTTTGGCCACTGACTGACacaagctggtacaactacac

ML127r: TCCTgttagttgtaccagcttgtGTCAGTCAGTGGCCAAAACacaagctggagtacaactaca

ML128f: GCTGttgaagttcaccttgatgccgGTTTTGGCCACTGACTGACcggcatcagtgacttcaac

ML128r: TCCTgttgaagttcactgatgccgGTCAGTCAGTGGCCAAAACcggcatcaagtgacttcaa

ML124F~ML128r: the sequences shown in upper case are the backbone sequence and the loop sequence; the sequences shown in lower case are the target-specific sequences to different parts of the eGFP mRNA.

5-way ligation primers for constructing the multimers of miR-155:

ML131f: AATT CCTAGG (*AvrII*) TGGAGGCTTGCTGAAGGCTGTATGC

ML131r: AATT TGTACA (*BsrGI*) TTGTCATCCTCCCACGGTGGCCATT

ML132f: AATT TGTACA (*BsrGI*) TGGAGGCTTGCTGAAGGCTGTATGC

ML132r: AATT GAATTC (*EcoRI*) TTGTCATCCTCCCACGGTGGCCATT

ML133f: AATT GAATTC (*EcoRI*) TGGAGGCTTGCTGAAGGCTGTATGC

ML133r: AATT GGATCC (*BamHI*) TTGTCATCCTCCCACGGTGGCCATT

ML134f: AATT GGATCC (*BamHI*) TGGAGGCTTGCTGAAGGCTGTATGC

ML134r: AATT CATATG (*NdeI*) TTGTCATCCTCCCACGGTGGCCATT

The human miR-30 backbone based mirshRNAs

For the human miR-30 backbone-based construction, miR-30 surrounding context sequences were PCR amplified from the plasmid LMP (Dickins et al., 2005) using primers ML47f and ML47r and inserted into pML21 to give pML23. Two sets of oligos ML44f and ML44r and ML91f and ML91r were used to anneal to form the eGFP target-specific miR-30 stem-loop sequences. For miR-30, the 3' strand is the effector strand of the miRNA. The annealed oligos were treated with Klenow fragment 3'-5' exo⁻ (NEB) to form the full mirshRNA and were subsequently cut with *XhoI/EcoRI* to be ligated with pML23 digested with the same enzymes, to give pML23-eGFP (ML44f and ML44r) and pML23-eGFP2 (ML91f and ML91r). The constructs were verified by DNA sequencing. The target-sequence containing miR-30 were PCR amplified with four sets primers containing restriction sites that allowed four-way ligation of three copies of the target-specific mirshRNA into *AvrII/NdeI* digested pML29 as before.

Primers

ML47f: AATTCTAGAAGATCTCTCgacTAGGGATAACAGG

ML47r: ATATCTAGATTGAAAAAagtgatttaatttataccattt

Target sequence specific oligo pairs:

ML44f: ACTCTCGAGTGCTGTTGACAGTGAGCGAgcacaagctggagtacaactaTAGTGAAGCCACAG

ML44r: ATTGAATTCCGAGGCAGTAGGCAGgcacaagctggagtacaactaTACATCTGTGGCTTC

ML91f:

AAATCTCGAGAAGGTATATGCTGTTGACAGTGAGCGaagccacaacgtctatatcatgTAGTGAAGCCACAG

ML91r: ATTGAATTCCGAGGCAGTAGGCACagccacaacgtctatatcatGTACATCTGTGGCTTC

ML44f~ML91r: the sequences shown in upper case are the backbone sequence and the loop sequence; the sequences shown in lower case are the target-specific sequences to different parts of the eGFP mRNA.

Construction of mir290C reporters

pML81 Was constructed by inserting three copies of the oligo-annealed ESCC miRNA target sequence and its surrounding sequence (up to 48 bp in total, using 3 sets of oligos, ML151f and ML151r, ML152f and ML152r, ML153f and ML153r) within the 3'UTR of Cdk1na (Wang et al., 2008c) to pML80, which contains the wild type Cdk1na 3'UTR, amplified from BAC RP23-73N16. The target recognition sites were confirmed by sequencing. The 3x recognition sites were further PCR-amplified using two sets of primers (ML182f and ML182r, ML183f and ML183r) with different restriction site combinations and ligated into the *XbaI/PacI* fragment of the pML82C, between the Neo resistant gene and the *rbgpA* site to give rise to pML82M, a PBTR flanked Huc-EM7-Neo cassette with 6 copies of the ESCC miRNA recognition sites.

Oligo pairs for ESCC miRNA recognition site construction:

EcoRI to *HindIII*

ML151f: AATTCGTGTGATCCTCAGACCTGAATAGCACTTTGGAAAAATGAGTAGGACTTTGA

ML151r: AGCTTCAAAGTCCTACTCATTTCCTCAAAGTGCTATTCAGGTCTGAGGATCACACG

HindIII to *SpeI*

ML152f: AGCTTGTGATCCTCAGACCTGAATAGCACTTTGGAAAAATGAGTAGGACTTTGA

ML152r: CTAGTCAAAGTCCTACTCATTTCCTCAAAGTGCTATTCAGGTCTGAGGATCACA

SpeI to *PstI*

ML153f: CTAGTGTGATCCTCAGACCTGAATAGCACTTTGGAAAAATGAGTAGGACTTTGCTGCA

ML153r: GCAAAGTCCTACTCATTTCCTCAAAGTGCTATTCAGGTCTGAGGATCACA

Primers for multimerisation of the ESCC miRNA recognition sites:

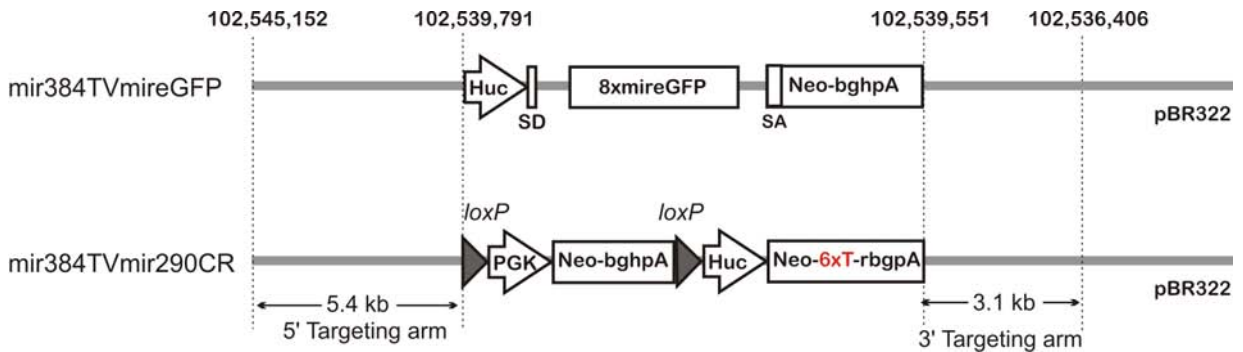
ML182f: AATT TCTAGA (*XbaI*) GTCCCCACTAGGACTTTGGAATTCG

ML182r: AATT CCATGG (*NcoI*) TAGATATGCCTCCTGCCGACCTGCA

ML183f: AATT CCATG (*NcoI*) GTCCCCACTAGGACTTTGGAATTCG

ML183r: AATT TTAATTAA (*PacI*) TAGATATGCCTCCTGCCGACCTGCA

Construction of targeting vector for miR-384 locus with miR-eGFP (miR-384TVmiR-eGFP) or mir290CR reporter (miR-384TVmir290CR)



The coordinates are genomic fragments used for the targeting arms and are based on NCBI Build 37.

pML88 Was modified by PCR cloning of mini homology arms from miR-384TVPTK (Haydn Prosser) outside of the reporter construct to give pML91. The linear *KpnI/AgeI* fragment of pML91 was used to drive homologous recombination of the reporter construct onto the miR-384TV to replace the PGK-EM7-Puro Δ tk in the recombineering strain EL350. The correct plasmid DNA was extracted and re-transformed into Stbl2 strain (Invitrogen) to isolate the correct the plasmid from the mixture. The resultant vector is miR-384TVmiR-eGFP. The targeting vector was linearised with *SbfI* for gene targeting.

The mir290CR reporter, the ESCC miRNA-regulated Neo resistant reporter was modified similarly to miR-eGFP by PCR cloning of mini homology arms into pML82M to give pML93. The *KpnI/AgeI* fragment of pML93 was used to conduct recombineering experiments to generate the miR-384TVmir290CR targeting vector. This targeting vector was also linearised with *SbfI* for gene targeting.

1.4. BAC vectors for giant PB transposons (PB-HPRT-28, PB-HPT-70 and PB-HPRT-100)

The human BAC (RP11-674A04) was obtained from the Wellcome Trust Sanger Institute BAC clone Archives. The *loxP* site on the BAC backbone was replaced by the *EM7-Zeomycin* resistant cassette (gift from Junji Takeda at Osaka University) by recombineering(Lee et al.,

2001). The plasmid containing both PB ITRs was a gift from Xiaozhong Wang at Northwestern University. PL451, PL452 and PL313 are gifts from Pentao Liu at Wellcome Trust Sanger Institute. The PB5'ITR was PCR-amplified using primers Neo-PB5-F and Neo-PB5-R to insert the PB5'ITR 5' upstream of a *loxP*-flanked *PGK-EM7-Neo* cassette in PL452. The PB3'ITR was PCR-amplified with primers Bsd-PB3-F and Bsd-PB3-R to insert PB3'ITR 3' downstream of an EM7 driven Blasticydin (Bsd) cassette in PL313. The PB5'ITR together with the *loxP* flanked *PGK-EM7-Neo* cassette was PCR-amplified with 110 bp chimeric PCR primers (PB3-forward and PB3-reverse) and inserted 10 kb downstream of the *HPRT* stop codon on the BAC by recombineering. The *loxP*-flanked Neomycin cassette was excised by *L*-arabinose induced Cre expression in the bacterial strain EL350(Lee et al., 2001). *PGK-Puro Δ tk* was cloned from YTC37(You-Tzung and Allan, 2000) into PL451. A second round of recombineering was carried out to introduce a *PGK-Puro Δ TK/Frt-PGK-EM7-Neo-Frt* using chimeric primers ML157f and ML157r to insert it immediately downstream of the PB5'ITR in the BAC. The neomycin resistant cassette was removed by *L*-arabinose induction in the EL250 strain(Lee et al., 2001). The *EM7-Bsd-PB3' ITR* fragment was inserted either 10 or 40 kb upstream of the *HPRT* start codon to generate PB transposons with a cargo capacity of 70 kb (using primers 70kb-F and 70kb-R) and 100 kb (using chimeric primers ML160f and ML160r), respectively. To construct the 28 kb PB vector, a plasmid was made by PCR with primers (ML162fx and ML162r) to introduce a *loxP* flanked *PGK-EM7-Neo* cassette 3' downstream of the human HPRT mini-gene(Ramirez-Solis et al., 1995). The DNA fragment excised from this plasmid containing the exon 3-9 portion of the *hHPRT* mini-gene together with the *loxP* flanked *PGK-EM7-Neo* cassette was used to replace the exon 3 to 9 by recombineering on the 70 kb PB-containing BAC using primers ML163f and 60kb-R. The sequences of all primers are shown below.

Name	Sequence
Neo-PB5-F	5'- CCTCGAGGGGATCCATCGTCTAAAGAACTACCCATTTTA -3'
Neo-PB5-R	5'- GAGCTCCACCGCGGAAAGTTTAGGTCGAGTAAAGCGCAA -3'
Bsd-PB3-F	5'- GTTCTAGAGCGGCCGCCTGAATAATAAAAAAATTAGAAACT -3'
Bsd-PB3-R	5'- GAGCTCCACCGCGGATTTTTGTAAGAGAGAATGTTTA -3'
PB3-forward	5'- CAGAATATTTACACAGTCTCAAATATCATTCTCGTATTAATTACAAAGGGAAAAATAGTA TGTTTATACTGGAGATCTTCGACCTGCATCCATCTAGATCCAC -3'
PB3-reverse	5'- ATATTAGGAGGCCGAGATGTCAATTTGTTTCATTAATGATGATGTTAATCTGGATTACTTTG TCAAGACAGCCTCTGCCATTTTTGTAAGAGAGAATGTTTA -3'
ML157f	5'- GATTTGCGCTTACTCGACCTAACTTTAGTAGCTATGTTTGGACCTCATTTGGATCCCAA TTCAAACAACTTTAAAATAATTCTACCGGGTAGGGGAGGCGCT -3'
ML157r	5'- AATAATTCCTTAATATCAATGGATATTCAAGTCAGTGTTCAAATCTCCAATTGTCTCATAA GTGTCATAAATGGGTGTGTTTGGATCCATGTACCTGACTGATGA -3'
70kb-F	5'- CTGGTATCTTCAACAAATAAATTGCAGGGAATAAAGACATGATGGAGGAACCTAGACTA AAAGAGACTTAAAGATGAACCCAATTCCGATCATATTCAATAACCC -3'
70kb-R	5'- GTCTCATAAGTGTCTATAAATGGGTGTGTATTTTAAAGTTTGTGTTGAATTGGGATCCAAATG AGGTCCAAACATAGCTACTAAAGTTTAGGTCGAGTAAAGCGCAA -3'
ML160f	5'- CCAGTAATAACTAGTTAGAAAATGTAAAGGAAAACCGCCTGCCATACCCCGCCAAACACA CATACACTAAAGATGAATTTTCGACCTGCATCCATCTAGATCCAC -3'
ML160r	5'- GCCTGGGTGACAGAGTGAGACCTGTCTCCCCAACACACCCCAAAACAAAGCTTTGTGTG TTTTTTTTTAAATGGAAACACCGCGGATTTTTGTAAGAGAG -3'
ML162fx	5'- AATTGGATCCGATCATATTCAATAACCCCTAAT -3'
ML162r	5'- AATTCCGCGGACGATGGATCCCCTCGAGGGACCT -3'
ML163f	5'- TGATAGATCCATTCTATGACTGTAGA -3'
60kb-R	5'- GGTCATCTTTAAGTCTCTTTTAGTCTAGTTTCTCCATCATGTCTTTATCCCTGCAATTTAT TTGTTGAAGATACCAGCTAGTCGAGCCCCAGCTGGTTCTTT -3'

Table 2-1: primers used to construct giant PB transposons.

2. ES cell lines

2.1. Genotype of the precursor ES cell lines

AB1: 129S7/SvEv^{Brd-Hprt+}, XY (Evans and Kaufman, 1981; Kuehn et al., 1987; Simpson et al., 1997)

AB2.2: 129S7/SvEv^{Brd-Hprt^b-m²}, XY (Kuehn et al., 1987; Simpson et al., 1997)

NN5: AB2.2, *Blm*^{tm3/tm4} (Luo et al., 2000; Guo, 2004)

Jm8.F6: C57BL/6N, XY (Pettitt et al., 2009)

2.2. ES cell lines generated

Table 2-2 summarises all the cell lines generated for this thesis work.

I.D.	Cell line	Locus	Vector	Long range PCR for genotyping	Comments
1	AB1- R26 ^{mPBBaseERT2/+}	<i>Rosa26</i>	N/A	N/A	Juan Cadinanos generated the ES cell line AB1-R26 ^{LSLmPBBaseERT2/+} . Lox-Stop-lox cassette was popped out from this cell line.
2	AB1- Hprt ^{PBint2} R26 ^{mPBBaseERT2/+}	<i>Hprt</i>	hprtTVPB	5' arm: ML51f and PB5-1: 3.6 kb ML51f: CCCATATCTCACTATAATTCAGTCA 3' arm: PB3-1 and ML50r: 4.2 kb ML50r: ACTAACAACCCTTTCTCTCAAGGTCTAGTT	Gene targeting of cell line #1. Long range PCR was achieved with LA taq.
3	AB1-R26 ^{mPBBaseERT2/+} -Hprt ^M	<i>Hprt</i>	N/A	N/A	6-TG resitant spontanous hprt mutant of cell line #1 selected with 6-TG.
4	AB1- R26 ^{mPBBaseERT2/+} - Hprt ^M - Gdf9 ^{hprtmimiPB/+}	<i>Gdf9</i>	Gdf9TVML4	5' arm, ML30f-ML24r: 3.9 kb ML30f: CCAAGTCATGGCACTCCCAGCAACTTCT ML24r: GGCGGAGCAGAGGAGGAGCGGAGG	Gene targeting of cell line #3.
5	NN5- Gdf9 ^{hprtmimiPB/+}	<i>Gdf9</i>	Gdf9TVML4	Same as cell line #4.	Gene targeting of cell line NN5.
6	NN5- Gdf9 ^{eGFP/+}	<i>Gdf9</i>	pML46	3' arm: IresF- ML31r, 4.0kb IresF: ACCCCATTGTATGGGATCTG	Gene targeting of cell line NN5.
7	Jm8.F6-Blm ^{e/e}	<i>Blm</i>	BlmTVML1 BlmTVML2	5' arm, ML108f-ML102rx: 3.6 kb (BlmTV2) ML108f-ML102rx: 5.5 kb (BlmTV1) ML108f: AGGGACTACTTTTCTCGCTGTTC ML102rx: AGACCATCACATACACTCCCATTTTT 3' arm, ML103f-IRES-F: 7.8 kb ML103f: AGTCACATTAGGCTGTACTGGAAAG IRES-F: ACCCCATTGTATGGGATCTG	Gene targeting of cell line Jm8.F6 sequentially with BlmTVML2 and BlmTVML1. Subsequently, the floxed PGK-Neo-pA cassette was popped out with Cre transient expression.
8	Jm8.F6-Blm ^{e/e} - Hprt ^{PBint2}	<i>Hprt</i>	hprtTVPB	ML48f-ML30r: 3.4 kb ML48f: GCAAATTTGGTAACTAGGTTCAATGAGAG ML30r: AACCTCGATATACAGACCGATAAAACACAT ML31f-ML48r: 3.3 kb ML48r: GAAAGTTACAGAGCACTAGGGACAGAAAAA ML31f: TTGCGCTTTACTCGACCTAAACTTT	Gene targeting of cell line #7.
9	Jm8.F6-Blm ^{e/e} - Hprt ^{PBex3}	<i>Hprt</i>	B6PBhprtE3TV	Kosuke Yusa supplied primers.	Gene targeting of cell line #7.
10	Jm8.F6-Blm ^{e/e} - miR-384 ^{miR-eGFP}	<i>miR-384</i>	miR-384TV- miR-eGFP	ML173f-ML173r: 3.5 kb ML173f: TAGCCAGATTTTCTCCTCTCTCTG ML173r: TTTCCAAGGTCACAGAACTGAAAA ML172f-Neo-rev2: 8.3 kb ML172f: AATGTCTCATGTGGCTAATGGCACT Neo-rev2: AGGTCGGTCTTGACAAAAAGAAC	Gene targeting of cell line #7.
11	Jm8.F6-Blm ^{e/e} - Hprt ^{PBex3} miR-384 ^{miR-eGFP}	<i>Hprt</i>	B6PBhprtE3TV	Same as cell line #9.	Gene targeting of cell line #9.
12	Jm8.F6-Blm ^{e/e} - miR-384 ^{miR-eGFP} Hprt ^{PBint2}	<i>Hprt</i>	hprtTVPB	Same as cell line #2.	Gene targeting of cell line #8.
13	Jm8.F6-Blm ^{e/e} - miR-384 ^{mir290CR}	<i>miR-384</i>	miR-384TV- mir290CRNeo	Same as cell line#10.	Gene targeting of cell line #7.

Table 2-2: A summary of ES cell lines generated for this thesis work.

3. Media and chemicals used for ES cell culture

Phosphate Buffered Saline (PBS): PBS (1x) was prepared in 10 L quantities and stored at room temperature. NaCl (80.0 g), KCl (2.0 g), Na₂HPO₄·7H₂O (10.72 g) and KH₂PO₄ (2.0 g) were dissolved in 10 litres of Milli-Q water. The PH was adjusted to 7.2 with saturated solution of Na₂HPO₄·7H₂O. Add phenol red until a peach colour is achieved.

Trypsin: trypsin solution was prepared in 5 L quantities and filter-sterilised and stored at -20°C. NaCl (35.0 g), D-glucose (5.0 g), Na₂HPO₄·7H₂O (0.9 g), KCl (1.85 g), KH₂PO₄ (1.2 g), EDTA (2.0 g), trypsin (12.5 g, Invitrogen, 840-725IL) and Tris Base (15.0 g) were mixed in 5 L MilliQ water and the PH was adjusted to 7.6 with HCl.

β-mercaptoethanol (BME): 10⁻² M stock solution (100X) was prepared by adding 72 µl of 14 M β-mercaptoethanol to 100 ml PBS and filter-sterilised.

Gelatin: 0.1 % (W/V) gelatin solution was prepared in MilliQ water and sterilized by autoclaving.

Cell staining solution: 2 % (W/V) methylene blue (Sigma) in methanol.

ES cell lysis buffer: 50 mM Tris-HCl [pH 8.0], 5 mM EDTA [pH 8.0], 200 mM NaCl, 1 % [w/v] SDS. Upon cell lysis, a final concentration of 0.1 mg/mL proteinase K (10x stock) was added.

Proteinase K (Roche, 03 115 879 001) stock: 100 mg lyophilised proteinase K was dissolved in 10 ml ddH₂O, to give 10x stock. The stock was aliquoted and stored in -20°C.

Blasticidin S HCl (Invitrogen, R210-01): 5 mg/ml stock (1000x) was made in PBS. After the powder was dissolved completely, the solution was sterilized by filtering through a 0.2 µm syringe filter and stored at -20°C.

FIAU (1-(2'-deoxy-2'-fluoro- β -D-arabinofuranosyl)-5-iodouracil): 200 μ M stock (1000x) was made in PBS and 5 M NaOH was added drop-wise until all the powder dissolved. After mixing, the solution was sterilised by filtering through a 0.2 μ m syringe filter and stored at -20°C.

G418 (or Geneticin, Invitrogen, 10131): 50 mg/ml solution The working concentration is 180 μ g/ml.

Puromycin (C₂₂H₂₉N₇O₅.2HCl, Sigma): 3 mg/ml stock (1000X) was made in milliQ water, then sterilised and stored at -20°C.

HAT (Invitrogen, 21060-17): 50X liquid stock, containing 5 mM Hypoxanthine, 20 μ M Aminopterin and 0.8 mM Thymidine. The stock was stored at -20°C.

HT (Invitrogen, 11067030): 50X liquid stock, containing 5 mM Hypoxanthine and 0.8 mM Thymidine.

4. AB1, AB2.2 and B6 derived ES cell culture conditions

The culture conditions of AB1, AB2.2 and B6 derivatives have been described in details previously by Ramirez and co-workers (Ramirez-Solis et al., 1993). These cells were maintained on γ -irradiated (60 Gray) or mitomycin C (Sigma) inactivated monolayer SNL 76/7 STO feeder cells. The ES were cultured at 37°C with 5 % CO₂ in M15 medium, which consists of Knockout™ D-MEM (Invitrogen 10829-018) and 15 % foetal bovine serum (FCS, Hyclone), 2 mM L-glutamine (Sigma), 50 units/ml penicillin (Sigma), 40 mg/ml streptomycin (Sigma) and 0.1 mM β -mercaptoethanol (Sigma).

4.1. Passaging, freezing down and thawing ES cells

For passaging, the medium was changed to M15 two hours prior to passage. The plate was washed with PBS twice before the addition of trypsin. The ES cells were trypsinised for 5 min at 37°C before the addition of M15 medium to inactivate the trypsin. The cells were suspended by pipetting the cell-containing medium up and down vigorously. The single cell

suspension was plated again on fresh feeder-containing plate. For freezing down, the ES cells were spun down after resuspension in M15 medium post trypsinisation, and resuspended in freezing medium, which consists of Knockout D-MEM, FCS, and DMSO (7:2:1, v/v). The vials were left in polystyrene boxes or a freezing pot at -80°C for at least 24 hours and transferred to liquid nitrogen subsequently. For thawing ES cells, the cells were taken out of liquid nitrogen and immediately thawed at 37°C. Cells were then spun down and the freezing medium was aspirated. The cells were re-suspended in fresh culture medium and transferred to a feeder-containing plate.

4.2. ES cell transfection by electroporation

70 % - 90 % confluent ES cells were fed 2~4 hours pre-electroporation. The cells were trypsinised as described before and the cell suspension was spun down at 1200 rpm for 3 min at room temperature. The cells were then resuspended in PBS and washed twice and finally counted to make a final concentration of 1.4×10^7 cell/ml. 0.7 ml of cells were mixed with DNA and transferred into 0.4 cm gap cuvette (Bio-Rad). The gene pulser (Bio-Rad) was configured at 230 V and 500 μ F to electroporate the cells. Cells were recovered in M15 medium and plated into 90 mm feeder-containing plates. For gene targeting, 20 μ g of linearised targeting vector was used. The next day, drug selection was applied and the drug-containing medium was changed daily to allow ES cell colonies to form. For BAC transfection, 5 μ g of BAC DNA purified using Qiagen large construct purification kit (Optional protocol without endonuclease treatment) was used. For PB mobilisation, the conditions were described in a separate section.

4.3. ES cell transfection by lipofection and pulse puromycin selection

Transient transfection of PGK-puro coexpressing PBase and hyPBase plasmids was achieved with Lipofectamine 2000 (Invitrogen). For large scale experiment, 3×10^6 cells in suspension were transfected with 24 μ g of PBase or control plasmids (CMV-hyPBase-PGK-Puro, CMV-mPBase-PGK-Puro) mixed with 60 μ l Lipofectamine 2000 reagent. The transfected cells were plated onto 90 mm feeder-containing tissue culture plate in 3 ml OptiMEM medium. One hour post plating, 7 ml of M15 was added. The next day, fresh M15 was replaced. The

transfection experiment can be scaled down according to Invitrogen's standard protocol. For pulse puromycin enrichment strategy, M15 containing 1 or 1.5 $\mu\text{g}/\text{ml}$ puromycin was supplemented 16 hours post lipofection and was sustained for 48 hours.

4.4. Picking ES cell colonies

50 μl of trypsin was added to each well of the 96-well round bottom plate. After washing the 90 mm tissue culture plate with PBS, ES cell colonies were picked using a P20 pipette manually and transferred to the trypsin-containing wells. After completing the full plate, the cells were incubated at 37°C for 15 min. 150 μl of M15 was added per well to inactivate the trypsin. The colonies were broken up into single-cell suspension by continuous pipetting the cell suspension up and down vigorously. The cells were then transferred to a feeder-containing 96-well plate.

4.5. Cre or Flp mediated recombination to pop-out selection cassettes

20 μg of Cre expressing plasmid (pCAGGS-Cre) or Flp expressing plasmid (PGK-FlpO) was electroporated into 1×10^7 cells. The electroporated cells were serially diluted in M15 and 1,000 cells were plated onto a 90 mm feeder-containing plate in duplicates. The next day, drug-containing medium was added to one of the plates while the other was grown under M15. Upon eight days selection, the differential colony number reflected how efficient the pop-out reaction had proceeded. According to the ratio, corresponding number of colonies were picked. Usually a third of the colonies contained the pop-out event.

5 *piggyBac* mobilisation in ES cells

5.1. Plasmid-to-genome mobilisation

The condition described was established for obtaining predominantly single-copy *piggyBac* integration per cell using a promoter-driven selection cassette to select for genomic integration events. Both *piggyBac* transposon-containing plasmid and transposase-expressing plasmid were prepared using commercial column based maxi-purification and dissolved in sterile Tris-EDTA buffer (TE). The transposon-containing plasmid was diluted to a final concentration 10 ng/ μl , whereas the transposase-expressing plasmid was adjusted to a final

concentration of 1 µg/µl. 100 ng of transposon donor was mixed with 10 µg mPBase expressing plasmid. The DNA was mixed with 1×10^7 ES cells and electroporated under the condition as described before. One tenth of the electroporated ES cells were plated on a 90 mm feeder-containing plate and the rest on the other. The next day, the drug-containing M15 medium was supplied to the cells. After eight days selection, the plate with a tenth of the electroporated cells was stained using the methylene-based cell staining solution to estimate the mutant complexity. The ES cell colonies could also be picked to extract DNA for *piggyBac* copy number per cell analysis using the PB5'ITR sequence as the probe.

5.2. Intra-genomic *piggyBac* mobilisation

When the transposon is pre-integrated in the genome, transfection of the transposase-containing plasmid mediates the mobilisation of the transposon from the original location to elsewhere in the genome. A selection scheme is necessary to eliminate the cells with the transposon still residing in the donor locus. This can be achieved by engineering the transposon into the X-linked *Hprt* locus or a bipartite *hprt* mini gene that is engineered in any selected genomic location. When the transposon is present in *Hprt*, it disrupts its expression; therefore such cells are sensitive to HAT. Upon excision, the function of *Hprt* is restored; therefore cells become resistant to HAT. Additionally, a positive selection marker, which is present within the *piggyBac* transposon, can be used to select for transposon re-integration events. The excision efficiency using *Hprt* as the donor locus is measured to be around 0.1% ~ 1 % of total electroporated cells using mPBase (Wang *et al.*, 2008b; Liang *et al.*, 2009).

For mobilising a genomic-residing *piggyBac* transposon, 1×10^7 ES cells were electroporated with 25 µg transposase (either mPBase or HyPBase), using the ES cell transfection by electroporation protocol as described before. One tenth of the electroporated ES cells were plated on a 90 mm feeder-containing plate and the rest on another. The next day, the drug containing M15 medium was supplied to the cells. After eight days selection, the plate with a tenth of the electroporated cells was stained using the methylene-based cell staining solution to estimate the mutant complexity.

5.3. PB Transposon excision for phenotypic reversion

The *piggyBac* transposon can be removed from the genome in the phenotypic reversion experiments in order to establish the causal link between the genotype and phenotype. The *piggyBac* transposon used in this thesis contains a *Puro Δ TK* selection cassette, which enables the negative selection of ES cells with the transposons removed from the genome. 3×10^6 ES cells were electroporated with 25 μ g transposase (either mPBase or HyPBase), using the ES cell transfection by electroporation protocol described before. As a control, 3×10^6 ES cells were electroporated with 25 μ g pBSKSII+ DNA. The cells were maintained in M15 medium for three days after the electroporation, allowing the decay of TK mRNA and protein. At day four post electroporation, the ES cells were trypsinised and 1×10^5 cells were plated on a fresh feeder-containing 90mm plate. The next day, FIAU selection was commenced and the medium was changed daily for the first six days. At day eight to ten, the colony number between the control and the PBase transfected cells were compared in order to estimate the reversion efficiency. The colonies from the PBase transfected cells were picked for further molecular analysis of the genotype and phenotype assessments.

6. Homozygote mutant generation using *Blm*-deficient ES cells

The heterozygote mutant libraries were generated by the *piggyBac* intra-genomic mobilisation method for obtaining around 1,000 colonies per 90 mm plate using both the HAT and positive selection (Puromycin, 3 μ g/ml) to obtain genome-wide *piggyBac* reintegration events as described previously. Twenty independent pools were conducted to give a total complexity of 20,000 mutants. The ES cell colonies from each pool were trypsinised for 15 min at 37°C to obtain a single cell suspension with vigorous pipetting, and plated onto a fresh 90 mm feeder-containing plate. The mutant pools were expanded until the 90 mm plate was confluent, before they were subjected to the phenotypic selection.

The total number of cells required to obtain a few homozygote mutants for each initial heterozygote mutant was calculated based on the LOH rate in *Blm*-deficient ES cells, as shown in the following equation: $N = C \times 10^4 \times m \times 4.2^{-1}$; where N is the total cell number, C is the complexity of the heterozygote mutant library, and m is the number of homozygote mutants

expected. When the mixed mutant pools were confluent on 90 mm plate, it would reach 3×10^7 cells. Therefore, the total cell numbers for all twenty mutant pools would be 6×10^8 cells, thus the number of homozygote mutants for each initial heterozygote mutant would be around 12. In practice, the heterozygote-to-homozygote conversion is a stochastic process and the LOH rate differs across the lengths of the chromosomes with the LOH rate lower than the measured LOH rate towards centromeric regions. Therefore, the calculation is only a guide to the expansion process.

7. Fluorescent Activated Cell sorting (FACs) of eGFP expressing ES cells

ES cells either transfected with eGFP or having eGFP expressing cassette knocked-in the ES cell genome were analysed live for eGFP expression. ES cells were harvested from wells of the 24-well plate by trypsinisation and were washed twice with PBS. The cells were resuspended in a 1 % (V/V) FCS-containing PBS solution to a final density of 1×10^5 cells/ml. The cells were passed through a cell-strainer-capped round bottom FACs tube (BD Falcon, 352235) to remove of the cell clumps. The tubes were left on ice until FACs analysis. For FACs analysis, either Beckman Coulter FC-500 or BD LSRII FACs machines were used.

A two-parameter dot-plot of Forward Light Scatter (FLS) vs. Side Scatter (SS) was plotted and adjusted to give the best separation of the cells. Another two-parameter dot-plot of FL1 and FLS was plotted, and the voltage of the FL1 channel was adjusted to the best signal to noise ratio for detecting green fluorescent cells. Unless specified, 10,000 ungated events were acquired for each sample. Analysis of the FACs data was conducted using FlowJo software.

8. Cytogenetic analysis

8.1. Preparation of metaphase spread

ES cells were grown in a well of a 24 well plate until 80% - 90 % confluent. Two hours prior to trypsinisation, the medium was changed to M15 with 1mg/ml cocemid (1:100 from the stock solution, Sigma). The cells were trypsinised after cocemid treatment and washed with PBS twice. The cell pellet was then resuspended in the residual PBS and 1 ml of 1 % (w/v) trisodium citrate was added to further resuspend the cells and the solution was incubated at

37°C for 20 min. The solution was cooled down on ice and the ice-chilled fixing agent was added dropwise for the first 5 min. The fixing agent contains methanol and glacial acetic acid (3:1, v/v). Then 7 ml of fixing agent was gradually added and the solution was slowly mixed with a glass pipette. The solution was washed with fixing agent twice and was spun down at 1,000 rpm for 5 min. A final residue 100 µl of fixing agent was left to resuspend the cells and one drop was applied onto a microscope slide (positively charged superfrost or normal glass slide with 100 % ethanol pretreated) which was placed in a pre-wetted paper towel to allow spreading. The slide can be stained with Dapi to visualise the chromosomes using a fluorescent microscope or directly visualised under a light microscope.

8.2. Sister chromatid exchange analysis

Sister chromatid exchanges (SCEs) can be visualised in a proliferating cells with the sister chromatids differentially labelled and stained. When proliferating cells are incubated with the thymidine analogue BrdU (5-bromo-2'-deoxyuridine), BrdU is incorporated into the newly synthesised DNA strand. Upon cell division after the first round of DNA replication, the sister chromatids are equally labeled. In the second round of DNA synthesis and at the metaphase, the chromatid containing the original template strand is darker than the chromatid with both strands of the DNA containing BrdU, arising from two rounds of cell cycle. Therefore, any SCEs can be visualised at this stage. Further cell cycles will accumulate sister chromatids with both strands labelled with BrdU, therefore the ratio of differentially labeled to undifferentially labelled chromatids will significantly reduce. Therefore for SCE analysis, it is important to know the doubling time of the cell line under investigation to in order to optimise the BrdU labelling time.

Freshly thawed BrdU stock solution (2 mM) was diluted in ES cell medium to give a final 10 µM BrdU. The medium was added to ES cells grown on a well of the 24 well feeder-containing plate. For JM8.F6 ES cells, after 34 hours BrdU incubation, the cells were prepared for metaphase spread as described before. BrdU is light sensitive; therefore the slides were kept in dark to dry. 0.01 % (w/v) Acridine orange staining solution was prepared in Sorensen's buffer (0.06 M, pH 6.5). Sorensen's buffer was made with 2.51 g of KH_2PO_4 and 5.87 g of

Na₂HPO₄ in 1 L of distilled water and the pH was adjusted. The metaphase slide was stained in Acridine orange solution for 5 min in the dark and was rinsed briefly with Sorensen's buffer. The slide was mounted with the same buffer and the excess fluid was blotted. The slide was examined using fluorescence microscope with the green channel. Acridine orange produces a green fluorescence when bound to double-stranded DNA and red fluorescence with single stranded DNA and RNA. Regions of BrdU-labelled DNA produce red fluorescence due to photo-degradation of the BrdU.

9. DNA methods

9.1. Recombineering technology

9.1.1. *E. coli* strains (Lee et al., 2001)

DH10B: F⁻ endA1 recA1 galE15 galK16 nupG rpsL ΔlacX74 Φ80lacZΔM15 araD139

Δ(ara,leu)7697 mcrA Δ(mrr-hsdRMS-mcrBC) λ⁻

EL250: DH10B [λcI857 (*cro*-*bioA*) <> *araC*-P_{BAD}*flpe*]

EL350: DH10B [λcI857 (*cro*-*bioA*) <> *araC*-P_{BAD}*cre*]

The DH10B strain lacks *recA*, which provides a good host environment for BAC or plasmid containing repetitive sequences. EL250 and EL350 were modified based on DY380 (DH10B [λcI857 (*cro*-*bioA*), *tet*]), which contains the defective λ prophage. The λ *cI857* repressor provides a temperature regulatable expression. At 32°C, the *PL* operon encoding *exo* and the red recombination genes, *bet* and *gam* are repressed by the *cI857* repressor. At 42°C, the *cI857* repressor is inactivated and the *PL* operon can be expressed. Gam inhibits the *E. coli* RecBCD nuclease from attacking the electroporated linear DNA, whereas Exo and Beta mediate the recombination activity. Therefore, incubation of DY380 at 42°C permits the homologous recombination between a linear exogenous DNA and the *E. coli* genomic DNA. EL250 and EL350 were generated based on DY380 by homologous recombination based replacement of the *tet* gene with *araC* and the arabinose-inducible *flpe* and *cre* genes, respectively.

The procedures described below can be used for BAC targeting, retrieving DNA fragment from the BAC, and plasmid fragment replacement.

9.1.2. Transformation of *E. coli* by electroporation

Either DH10B or the recombining strains were inoculated o/n from glycerol stock at 32°C in 5 ml LB broth. The next day, 0.5 ml of culture was transferred into 9.5 ml of the low salt LB (per 1 L, 10 g Bacto-tryptone, 5 g Bacto-yeast extract, 5 g NaCl, 1 L ddH₂O, and the pH is adjusted to 7.0 with NaOH, sterilised by autoclaving) and was inoculated until the OD₆₀₀ reached 0.6. The bacteria were cooled down on ice with gentle shaking and harvested by spinning down at 5,000 rpm for 5 min at 4°C. The pellet was then resuspended in 1 ml of ice cold double distilled water and washed twice with ice-cold water by spinning at 9,600 rpm for 30 seconds each time twice in the cold room. The pellet was finally resuspended in 180 µl of ice-cold water and mixed with the 10 ng plasmid or 100 ng BAC from a mini DNA preparation. The mixture was transferred to a pre-chilled 0.1 cm cuvette (BioRad Gene Pulser) and electroporated using a Gene Pulser (Bio-Rad) at 1750 V, 25 µF and 200 Ω. After electroporation, the bacteria were immediately transferred into SOC medium (2 % bacto-tryptone, 0.5 % Bacto-yeast extract, 10mM NaCl, 2.5 mM KCl, 10mM MgCl₂, 10 mM MgSO₄ and 20 mM Glucose in ddH₂O, pH =7.0, sterilised by autoclaving with Glucose added after autoclaving) and recovered at 32°C for one hour. The recovered culture was then plated on a 10 cm agar plate with the appropriate antibiotics and incubated at 32°C o/n.

9.1.3. DNA fragment preparation for recombineering

DNA fragments for recombineering could be either cut out of a plasmid or generated by PCR using chimeric PCR primers, as homologous sequences over 50 bp are sufficient for recombineering. Longer homologous sequences can improve the efficiency further. For the PCR-based method, each primer in the set was designed to contain 80 bp of sequence homologous to the target BAC or plasmid, followed by a 25 bp sequence identical to the target DNA fragment to be amplified. The KOD polymerase based PCR system (Novagen) was found to be the most reliable system to amplify target DNA fragments with long chimeric primers without the formation of primer dimers. The reaction set-up was as described by the

manufacture's protocol with addition of 5 % (V/V) final concentration of glycerol in difficult-to-amplify cases. Usually 20 ng plasmid DNA or 100 ng BAC DNA was used as the template for the PCR reaction. The PCR condition was 98°C 2 min, and followed by 30 cycles of 98°C 20 sec, 68°C 1 min (depending on the target length), and 72°C 5 min. The PCR product was column purified to remove salt and the concentration was estimated by running 2 µl on an agarose gel. For the release of fragment from a plasmid, the restriction digested fragments were separated on agarose gel and the correct fragment was extracted from the gel and column purified to remove the agarose.

9.1.4. Recombineering procedure

Strain EL250 or EL350 was inoculated o/n in 5 ml LB at 32°C. The next day, 0.5 ml of the culture was transferred to 9.5 ml fresh low salt LB and cultivated until the OD₆₀₀ reached 0.6 (approx. 2 hours). The bacterial culture was incubated at 42°C for 15 min to induce the *gam*, *bet* and *exo* expression and was spun down at 4 °C at 5,000 rpm for 5 min. The pellet was re-suspended in 1 ml ice-cold ddH₂O. The bacterial culture was further washed with ice-cold ddH₂O in the cold room (4 °C) by spinning it down at 9,800 rpm for 30 sec. The pellet was re-suspended in 200 µl ice-cold ddH₂O. 100 ng of the DNA fragment was mixed with the bacteria and the mixture was electroporated using the Gene-pulser as described in the previous section. After one hour recovery in SOC medium at 32°C, one tenth of the culture was plated on a 90 mm agar plate with the appropriate antibiotics, and the rest onto another plate. The plates were incubated at 32°C o/n. The colonies were picked and plasmid DNA was extracted which was subjected to diagnostic analysis either by restriction digestion or junction PCR from the target BAC or plasmid to the recombineering DNA fragment.

9.1.5. Cre or Flp mediated cassette pop-out in *E. coli*

EL250 or EL350 containing a BAC or plasmid with a *FRT* or *loxP* flanked cassette respectively were cultured in 5 ml LB at 32°C o/n with vigorous shaking. The next day, 0.5 ml of culture was transferred into 9.5 ml fresh low salt LB to shake for a further 2 hours or until the OD₆₀₀ reached 0.6 and 100 µl of 10 % (v/v) *L*-arabinose solution was added to the culture to give a final concentration of 0.1 % (v/v) *L*-arabinose. The culture was further shaken for 1 hour and

the bacterial culture was serially diluted to 10^{-5} of the original culture in LB media, and 100 μ l was plated on one 90 mm agar plate with the appropriate antibiotics and an equivalent amount on another plate without the antibiotics. The presence of the *loxP* or *FRT* flanked cassette confers the resistance to the antibiotics used; therefore, the antibiotic containing plate could be used to define the background level of bacteria without “pop-out” events. The plates were incubated at 32°C o/n and the next day, colonies could be picked from the plate without the antibiotics for further diagnostic analysis by either PCR or restriction digest of the BAC or plasmid DNA.

9.2. Southern blotting and hybridisation

4 μ g of genomic DNA was digested with 20 units of appropriate restriction enzyme overnight. The digested DNA was resolved on a 0.8 % agarose gel in TAE buffer either at 45 V for 10 hours or 15V o/n. The gel was rinsed with MilliQ water first and then was placed in denaturing buffer (1 M NaCl and 0.4 N NaOH in MilliQ water) with gentle agitation for 20 min at room temperature. An alkaline based transfer system was set up for o/n transfer as described in the original Southern method with the denaturing buffer as the transfer buffer. A gel tray was placed in a tank containing the transfer buffer. Three layers of the 3MM paper were soaked in the transfer buffer and were layered on top of the gel tray with their ends submerged under the transfer buffer in the tank. These Whatman 3MM paper acted as the liquid supply for the capillary transfer. The agarose gel was placed on top of the Whatman 3MM paper, followed by the Hybond XL (Amersham) membrane, which was presoaked in distilled water before placed on the agarose gel. Three layers of Whatman 3MM paper were placed on the membrane and finally a deck of tissue towels were layered on the assembly to drive the buffer upwards for the DNA transfer to the membrane. Bubbles were eliminated using a plastic pipette to roll over each layer during the assembly. At last a weight was placed on top of the tissue deck. The next day, the membrane was rinsed in the neutralising buffer (0.5 M Tris HCl, 1M NaCl in MilliQ water) for 1 min and was baked at 80°C for at least one hour for cross-linking.

The dried membrane was pre-hybridised in a hybridisation tube with 15 ml hybridisation buffer (Sigma, H7033) supplemented with 150 μ l of the denatured Salmon sperm SSDNA (Ambion) at 65°C for 1 hour while the probe was labelled. 50 ng probe DNA was labelled with [γ -³²P]dCTP using Prime-It®II Random Primer labelling kit (Stratagene product, Agilent, 300385) for one hour, following the manufacturer's protocol. The labelled probe was filtered to eliminate the unlabeled isotopes using ProbeQuant G-50 column (GE). The probe was denatured by heating at 100°C and snap cooled on ice for 10 min before it was added to the pre-hybridised membrane. Hybridisation was conducted in a rotating oven o/n at 65°C.

The next day, the membrane was rinsed with low stringency wash buffer (2xSSC, 0.1 % (v/v) SDS; 20xSSC buffer stock, 3M NaCl and 300 mM Na₃citrate.2H₂O with pH adjusted to 7.0 with 14N HCl) and washed with this buffer for 10 min at 65°C. According to the signal strength of the membrane, it was further washed with higher stringency wash buffer (1xSSC, 0.1 % SDS or 0.5 x SSC, 0.1 % SDS) until the signal was roughly around 10-20 counts per second on the mini monitor Geiger counter. The membrane was sealed and placed in a cassette with an intensifying screen and an X-ray film was placed in the cassette to be exposed o/n at -80°C.

9.3. Southern blotting probes

Table 2-3: southern blotting probes used in this thesis work.

Probe	Size (bp)	Enzyme(s) to digest gDNA	Expected pattern	Probe construction
Neo	773	N/A	N/A	<i>EagI/XbaI</i> digest from plasmid pML4
Blm	437	<i>EcoRI/NheI</i>	WT: 9kb +Neo: 6.2 kb Δ Neo: 7.4 kb	PCR from wild type genomic DNA with: ML147f: TGGACTCAACACAGTGGAGGCCCTT ML147r: GTGAAGGTCAGAGGACAACCTGAAG
Msh6	385	<i>SpeI</i>	WT: 4.7 kb Mutant: 6.8 kb	PCR from wild type genomic DNA with: ML72f: GGTGGATGCGATCCTTGTAGGCGCT ML72r: CACGGAGCAGCCTCCCCCTCCTCC
Gdf9	700	<i>NcoI</i>	WT: 6 kb Targeted: 6.5 kb	<i>SalI/BamHI</i> digest of pGDF9-P1(Dong et al., 1996).
PB5'ITR	235	N/A	N/A	<i>EcoRV/NsiI</i> digestion of pML5.

9.4. Splinkerette PCR

9.4.1. Adaptor preparation

The methodology of the Splinkerette is depicted in Figure2-1. The Splinkerette adaptor was prepared by mixing 150 pmol of HMSpAa and 150 pmol of HMSpBb with 5µl of NEB buffer 2 (10x stock contains 500 mM NaCl, 100 mM Tris-HCl, 100 mM MgCl₂, 10 mM Dithiothreitol with pH =7.9 at 25°C) in double-distilled water with a total volume of 100 µl. The mixture was heated at 95°C for 10 min and gradually cooled down to room temperature, allowing the oligos to anneal. The adaptors can then be stored at -20°C. The HMSpBb oligo can be designed with different restriction enzyme recognition sequences to suit the need. The 4 bp cutter *Sau3A*I was used in this project. There primers were not phosphorylated at the 5'.

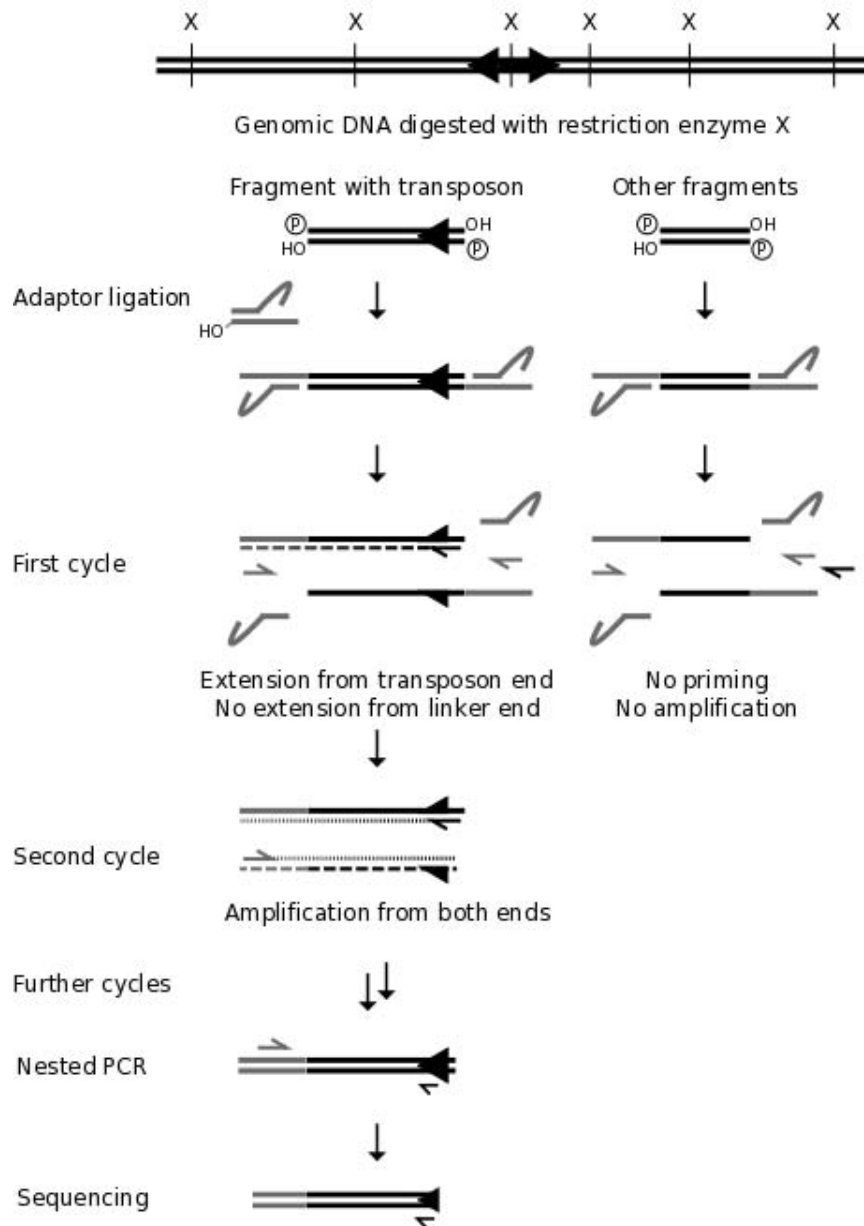
HMSpAa:

5'CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGTGTGCGACACTAGTGG

HMSpBb-*Sau3A*I:

5'gatcC**CA**CTAGTGTGCGACACCAGTCTCTAATTTTTTTTTTCAAAAAA

Figure 2-1: Schematic representation of the Splinkerette -PCR method.



A fragment of genomic DNA containing a *piggyBac* inverted terminal repeat (black double arrow) is shown. After digestion (cleavage sites marked with X), splinkerette adaptors (shown in grey) are ligated to all fragments. If the fragment contains the transposon end (left), the transposon-specific primer (black arrow) can hybridize and extend in the first round. Extension into the long-strand of the adaptor provides the template for the adaptor primer to anneal and extend. In the subsequent cycles of PCR, the transposon-specific and adaptor primers can amplify the transposon-genomic junction. A nested PCR is used to improve specificity. This scheme is adapted from review (Li et al., 2010).

9.4.2. Genomic DNA digestion and adaptor ligation

1-2 µg of genomic DNA was digested with 4 units of *Sau3AI* (NEB) in a 30 µl reaction at 37°C for 3 hours and the enzyme was heat-inactivated at 65°C for 20 min. 5µl of digested genomic DNA was mixed with 3 µl adaptor and 4 units of T4 ligase (NEB) in a 20 µl reaction and the ligation was carried out at 16°C overnight. The next day, the ligase was heat-inactivated at 65°C for 15 min. The ligation mixture was diluted ten-fold for the PCR amplification.

9.4.3. Nested-PCR amplification

A *Taq* polymerase-based Platmium®PCR supermix (Invitrogen) was used for both rounds of the PCR. The first round PCR was set with 0.5 µl of the diluted ligation mixture, 0.5 µl of the 10 µM adaptor specific primer SP1 and 0.5 µl of the 10 µM *piggyBac* transposon specific primers PB5-1 or PB3-1 (corresponding to the PB5'ITR and PB3'ITR) and 18 µl of the supermix. The PCR condition was 94°C for 2 min, followed by 2 cycles of 94°C 20 sec, 68°C 30 sec, 72°C 1min; then followed by 30 cycles of 94°C 20 sec, 60°C 30 sec, 72°C 1 min; and a final extension of 72°C 5 min.

The second round of PCR was set up similar to the first round with 1 µl of the first round PCR reaction as the template, and 0.5 µl of the 10 µM nested-adaptor primer SP2 and 0.5 µl of the 10 µM nested *piggyBac* specific primer PB5-2 or PB3-2 (corresponding to the PB3'ITR and PB5'ITR) were used. The PCR condition was 94°C for 2 min, followed by 30 cycles of 94°C 20 sec, 60°C 30 sec, 72°C 1 min; and a final extension of 72°C 5 min. 2 µl of the reaction mixed was used to run on a 1 % agarose gel to assess the PCR products. When a single product was generated, the PCR reaction mix was diluted in water and was subjected directly to capillary sequencing. When multiple products appeared, the different products were resolved by gel electrophoresis and the PCR products were extracted individually by gel extraction before capillary sequencing. The sequencing primers were further nested *piggyBac* transposon ITR specific primers PB5-seq and PB3-seq.

Primer sequences for nested PCR and sequencing:

Adaptor-specific primers:

HMSp1: 5'CGAAGAGTAACCGTTGCTAGGAGAGACC

HMSp2: 5'GTGGCTGAATGAGACTGGTGTCGAC

PB5'ITR-sepcific primers:

PB5-1: 5'CAAATCAGTGACACTTACCGCATTGACAA

PB5-2: 5'CTTACCGCATTGACAAGCACGCCTCACGGG

PB5-seq: 5'TTAGAAAGAGAGAGCAATATTTCAAGAATG

PB3'ITR-sepcific primers:

PB3-1: 5'TAAATAAACCTCGATATACAGACCGATAAA

PB3-2: 5'ATATACAGACCGATAAAACACATGCGTCAA

PB3-seq: 5'TTTTACGCATGATTATCTTTAACGTACGTC

9.5. Genomic DNA triple-primer competition PCR

To confirm the homozygosity of the mutant alleles, triple-primer PCR reactions were conducted on a locus-specific basis, based on determination of the mutagen integration site by the Splinkerette PCR. Two of the primers are locus specific and can amplify a wild-type product that spans the transposon insertion site. The third primer is mutagen-specific (primers in the PBITRs are used) and amplifies the mutagen-host junction together with one of the locus-specific primers. These three primers are used in equimolar amounts in a conventional PCR reaction set-up.

Msh6 locus:

ML71f: 5'TTTTGCCTCCGTGTATGTATGTGTG

ML71r-2: 5'CAGCAAGTTTTAGTACCTGGGTAAA

Wild type allele detection: ML71f + ML71r-2: 860bp

Mutant allele detection: ML71r-2 + PB5-1: 585bp

Ago2 locus:

Ago2F: 5'GCAGCCCCAGCCTCTTTACTTGT

Ago2R: 5'GAAAGCTCTGCCAACAGCTGGAAC

Wild type allele detection: Ago2F + Ago2R: 224bp

Mutant allele detection: PB5seq + Ago2R: 163bp

Primers for locus specific PCRs in Chapter 7

Actin locus (control):

Actin-F: 5'GTTTGGACAAAGACCCAGAGG

Actin-R: 5'CTCTCTCGTGGCTAGTACCTCAC

Giant PB integrated clones:

Bf6f: 5'GGGCCTCTTAAGATACTTTTAGCTG (for clone Bf6)

Bf6r: 5'AACCCAATTTCAAACAATTTATCA (for clone Bf6)

Cd8f: 5'TTATGGCCAAGGACATATACAAGTT (for clone Cd8)

Cd8r*: 5'AACTATGTAAACGATAACACGAAACG (for clone Cd8)

Ch2f: 5'GTACACAAGAGCTGATTCCATCC (for clone Ch2)

Ch2r*: 5'CTCCTTTGGCTGTTAGATGAACTTA (for clone Ch2)

Dc7f: 5'GACTTCTACCTGTTTGGTTGTGTTT (for clone Dc7)

Dc7r*: 5'AAAAGCTAGACCCGGATTTAAAAG (for clone Dc7)

Dc11f: 5'TGAGCTGACTAACAGAACAGAGTCA (for clone Dc11)

Dc11r*: 5'CCTAAACACACAAAACACGTAACAG (for clone Dc11)

* primers were used with PB3-seq to PCR amplify the PB5 ITR to genomic junction for each individual ES cell lines.

10. RNA method: RT-PCR

Total RNA was extracted from ES cells grown in a well of a 24-well gelatinised plate, either using a RNeasy Mini extraction kit (Qiagen) or using Trizol reagent with chloroform extraction (Invitrogen), following the manufacturer's protocol. The concentration of the RNA was determined with a NanoDrop ND-1000 machine by measuring the absorbance at 260 nm. 1 µg of the RNA was mixed with 1 µl of either random hexamer primers (10 µM stock) or an oligo-dT primer (10 µM stock) in 10 µl DEPC water (Ambion) to give a final volume of 12 µl. The mixture was heated at 100°C for 2 min and snap-cooled on ice for 2 min. The reverse transcription reaction was set up with a SuperscriptTMII cDNA synthesis kit (Invitrogen), by

adding 1 μ l 10 mM dNTP, 4 μ l of 5x synthesis buffer (250 mM Tris-HCl, 375 mM KCl, 15 mM MgCl₂, pH 8.3), 2 μ l 0.1 M DTT and 1 μ l (200 U) of SuperscriptTMII reverse transcriptase to the RNA and oligonucleotide mixture. The reverse transcription was conducted at 42°C for 1 hour and was followed by heat-inactivation at 70°C for 15 min. The cDNA reaction mixture was diluted ten-fold for use as the PCR template.

The PCR reaction was set up with 1 μ l of the tenfold diluted cDNA mixture, 1 μ l of gene specific primer mix (10 μ M stock) and 18 μ l of PCR supermix (Platmium[®] with anti-taq antibody in the mixture to permit ambient temperature PCR-setup, Invitrogen). The PCR condition was 94°C for 2 min, followed by 30 cycles of 94°C 20 sec, 60°C 30 sec, 72°C 1 min; and a final extension of 72°C 5 min.

Primer sequences

Oligo-dT primer: 5'-GGC CAC GCG TCG ACT AGT AC (T)₁₇-3'

Random hexamer primer: Invitrogen, cat no. 48190011

Gene specific primers:

Msh6

Msh6Ex1F: 5'CTTGTACAGCTTCTCCCAAGTCC

Msh6Ex3R: 5'TTACTTAAGGCCTCATCTGCACGTT

Ago2

Ago2Ex1F: 5'CTCGGCAACGCCACCATGTACT

Ago2Ex4R: 5'CGACACCCACTTGATGGATACCTTA

Ago2Ex13F: 5'CCACAGACCCTATCCAATCTCTGCT

Ago2Ex16R: TGATAGTCCTTCTCCAGCTTGATGC

Dgcr8

Dgcr8ex2F: 5'GTGGATGAAGAGGCCTTGAATTTCT

Dgcr8ex5R: 5'TTGTTCTCGTTCCTCATTGTCCTTC

Dgcr8ex9F: 5'CCTGGAAATTCTCATCCCTGACTTT

Dgcr8ex12R: 5'ACATGCGTAGTAAGGAACCCAGTT

Hprt

Hprt-Ex1F: 5'TCAGACCGCTTTTTGCCGCG

Hprt-Ex3R 5'TTATAGCCCCCTTGAGCACACAG

Hprt-Ex7F 5'GCTGGTGAAAAGGACCTCT

Hprt-Ex9R 5'CACAGGACTAGAACACCTGC

*Ccdc7**

Ccdc107-rev-1: 5'GGGACTCTTTTTACTAAACCTTCAGTGGTT

Ccdc107-rev-2: 5'TTTAGTTTTGTGTCTAGAAGTTCCTCTGG

*Dom3z**

Dom3z-rev-1: 5'GAGACCCACTTAGAAACACGCCTTTATTAT

Dom3z-rev-2: 5'CTAAGGAAGGCAGCACAGAAGTTCAT

*: Mutagenic exons used for gene inactivation. Rev-1 is from the last exon and rev-2 is from the penultimate exon of the mutagens. The reverse primers were paired with locus-specific primers for detecting trapping events, and triple-primer genomic PCR was used to assess homozygosity.

Table 2-4: Locus-specific primers used for the random trapping assay.

Gene	Chr	Mutagen	Location	Locus-specific primer
<i>Tmem131</i>	1	<i>Ccdc107</i>	Intron 4	d7-Tmem131-ex2: 5'TTCAGTCGGAGAGCATAATAGAAGT
<i>Dut</i>	2	<i>Ccdc107</i>	Intron 4	g10-Dut-ex1: 5'TGCCGGCTACGACCTATTACG
<i>Fstl4</i>	11	<i>Dom3z</i>	2 nd last intron	e8-Fstl4-ex12: 5'AGGTCCAGAGACGCTTGAAACCTAC
<i>A24Rik</i>	15	<i>Dom3z</i>	Intron 17-18	c4-A24Rik-ex15: 5'GGTGGCTACTTTCACCATCAACATC
<i>Sfrs3</i>	17	<i>Ccdc107</i>	Intron 2	e4-Sfrs3-ex1: 5'ACCGAGAATCTGTAGGAGCAGAACC
<i>Sema6a</i>	18	<i>Dom3z</i>	Intron 1	f8-Sema6a-ex1: 5'TTTCTTGAGCATTTACCTGGTCTCT
<i>Undcd3</i>	11	<i>Ccdc107</i>	Intron 1	h8-Undcd3-ex1: 5'CTTTATGACCAGGCCCTGTTGG
<i>Stk22s1</i>	7	<i>Dom3z</i>	Intron 6-7	a8-stk22s1-ex6: 5'GAAGCTAAGATACCTCCAGCAGCAA
<i>Ran</i>	5	<i>Dom3z</i>	Intron 4	f2-Ran-ex1: 5'GACAGGCGCGGAGACTCTCT
<i>Tmem50a</i>	4	<i>Ccdc107</i>	Intron 1	b3-Tmem50a-ex1: 5'GGCTGTTTTGTTTTCTTGCAAGACT

These primers were used in conjunction with *CCdc7* and *Dom3z* reverse primers to confirm trapping events.

11. Protein methods: Western blotting

11.1. Whole-cell protein extraction

ES cells were trypsinised from a gelatinised plate and washed with PBS twice. The cell number was determined and 1×10^5 cells were used per lane. Usually 1×10^6 cells were resuspended in 50 μ l of PBS. A 2x sample buffer was prepared with 25 μ l 4x NuPAGE[®] LDS sample buffer (Invitrogen, the stock contains 40 % glycerol, 4 % lithium docecyl sulfate, 0.8 M triethanolamine-Cl pH 7.6, 4 % Ficoll-400, 0.025 % phenol red, 0.025 % brilliant blue G250 and 2 mM EDTA-disodium) supplemented with 5 μ l of 14 M β -mercaptoethanol and 20 μ l PBS. 50 μ l of 2x sample buffer was mixed with the cell suspension and the mixture was heated at 95°C for 5 min and vigorously vortexed for 5 min for whole cell lysis.

11.2. Protein blotting and antibody hybridisation

10 μ l of the sample was loaded in each lane of the pre-cast 4 – 12 % Bis-Tris gel (Invitrogen) and a tenfold diluted sample was loaded for detecting β -actin as the loading control. The samples were resolved by running using 1x SDS MOPS running buffer (Invitrogen, 20x stock contains, 0.6M MOPS, 1.2 M Tris, 2 % SDS, and 50 mM sodium bisulfite) at 140 V at room temperature. The protein was transferred to Hybond ECL membrane (Amersham) using the Bio-rad transfer system with chilled transfer buffer (100 ml methanol, 50 ml 20x transfer buffer (Invitrogen), the rest MilliQ water to make up 1 L of the 1x transfer buffer) at 4°C with 90 V for a least one hour.

After transfer, the membrane was soaked with 2.5 % (w/v) milk powder in 1x TBST buffer (blocking buffer) for 30 min at room temperature with gentle shaking. TBST was made by adding 100 ml of 10x TBS (500 mM Tris.HCl, pH 7.4 and 1500 mM NaCl), 0.5 ml Tween 20 to up to 1 L of MilliQ water. The primary antibody was diluted in 1 ml of 0.5x blocking buffer (Table 2-5 shows the dilutions of each antibody used) and used to cover the protein containing site of the membrane, and covered with parafilm. The membrane was incubated with the primary antibody at 4°C o/n. The next day, the membrane was washed with TBST for 30 min with gentle shaking twice. The horseradish peroxidase (HRP)-conjugated secondary antibody (1 μ l) was diluted in 1 ml of 0.5x blocking buffer and incubated with the membrane

for 1 hour at room temperature. The membrane was then washed as before. The chemilluminescence (ECL) detection mix (Amersham) was applied to the membrane and left at room temperature for 1 min before the ECL film (Amersham) was exposed to the membrane. The exposure time varied depending on the strength of the signal, typically 10 sec, 30 sec, 1 min and 5 min exposures were used to give the best signal.

11.3. Antibodies used in this thesis

Table 2-5: Antibodies used in this thesis work.

Name	Type	Species	Clonal	Dilution	Source	Cat. No.
Anti-eGFP	Primary	Mouse	Mono	1: 1000	Roche	11 814 460 001
Anti-Dgcr8	Primary	Rabbit	Poly	1: 300	ProteinTech Group	10996-1-AP
Anti-Hprt	Primary	Rabbit	Poly	1: 500	Abcam	Ab10479-200
Anti-Blm*	Primary	Rabbit	Poly	1: 500	Abcam	Ab476-100
Anti- β actin	Primary	Mouse	Mono	1: 1000	Sigma	A5441
HRP conjugated anti-mouse IgG	Secondary	Horse	Mono	1: 1000	Cell Signalling	# 7076
HRP conjugated anti-rabbit IgG	Secondary	Horse	Mono	1: 1000	Cell Signalling	# 7074

* Blm detection only worked with homemade polyacrylamide gel, not the pre-casted gels from Invitrogen.

12. Regional high density Genomic comparative hybridisation (CGH) array

Genomic comparative hybridisation experiment was conducted in Chapter Seven to detect the copy number gain after PB-mediated cargo integration. A customer regional CGH array was designed to serve this purpose. The 230 kb human genomic region ChrX: 133,358,379-133,591,045 (hg18), covering the whole BAC (RP11-674A04) was used to design the hybridization probes for an Agilent regional CGH array (8x15K), with the criteria that the probes must pass a similarity score filter to exclude probes with secondary genomic alignments and exclusion of repetitive genomic regions. Additional criteria were adopted to avoid mouse-human cross species hybridization. The rules were: a, reject probes that have more than 90 % identity to the mouse genome; b, reject probes which have 20 bp or more of uninterrupted sequence matching to the mouse genome. In total, 1773 probes were selected from this region to provide an average detection resolution of 130 bp, and they were printed

in triplicate on the array. The remaining 9,600 probes were a random selection of probes from Agilent catalogue mouse CGH HD probes to provide the baseline normalization.

The DNA was extracted using Puregene kit (Qiagen). The DNA with the large-cargo PB integrated was compared to the DNA extracted from AB2.2. Both samples and the control were mixed in equal amount with pooled genomic DNA from human male primary cell lines. In this array, within the high density human probe region, the copy number increase of one on the Log_2 scale represents the gain of an extra copy. The raw array data was normalized using a robust cubic spline interpolation method contained inside the R[®] package aCGH.Spline (<http://cran.r-project.org/web/packages/aCGH.Spline/index.html>) to adjust for dye biases. A custom wavelet transform was applied to remove the presence of genomic waves and the true baseline was estimated using the median value reported by the 9,600 randomly selected probes.

The detailed information on the array design and the experimental procedures as well as the raw data was deposited on ArrayExpress (www.ebi.ac.uk/aerep/login). The ArrayExpress accession for the array design is A-MEXP-1849 and the accession for the experimental data is E-MEXP-2788. The login details are as follows: username, A-MEXP-1849; password, 1276737775749.

13. Illumina sequencing for PB integration sites analysis

For giant PB transposon integration analysis described in Chapter Seven, Illumina sequencing was conducted. The sequencing work was conducted by Daniel J. Turner and Sabine Eckert in the Next-generation sequencing R&D department at the Sanger Institute. Genomic DNA was extracted from pooled ES cells as described before. The DNA was sheared acoustically, and paired-end Illumina shotgun libraries were prepared, up to the point of adapter ligation. Ligated libraries were quantified by qPCR, relative to a concentration standard, using primers Ad_T_qPCR1 and Ad_B_qPCR2, which annealed to the Illumina adapters, allowing equimolar amounts of each library to be used in a semi-specific enrichment PCR. qPCR reactions were set up using a SybrGreen system following the manufacturer's recommendation (Applied

Biosystems). Approximately 750 ng of ligated library was used in the subsequent enrichment PCR. Enrichment PCR was optimized by performing a gradient PCR to find an appropriate annealing temperature, and by qPCR to establish the minimum number of cycles of amplification. In the first round of PCR, primer PB5pr_1 that was specific for the PB5'ITR and an adaptor specific primer PCR_V4, which was tailed with an Illumina-compatible sequence(Lander et al., 2001), were used. Phusion enzyme was used for the PCR and the cycling condition is as follows: 1 cycle of 94°C 2 minutes; 18 cycles of 94°C 20 seconds, 62°C 20 seconds, 72°C 40 seconds; and 1 cycle of 72°C 10 minutes. In the nested PCR round, 5 µl of first round PCR was used as the template. A PB5'ITR specific primer PB5prP5_2 was used along with primer PCR_V4.

Index sequences were designed to be error-correcting, by the use of Hamming codes(Hamming, 1950; Mamanova et al., 2010), and they were included in primer PCR_V4. The cycling condition was the same as before except that an annealing temperature of 60°C was used and 12 cycles were performed. Following PCR, reactions were cleaned up using a spin column, and were run in a 2 % agarose gel. Gel slices corresponding to a fragment size of 250-350 bp were cut, and the DNA was extracted without heating (Waterston et al., 2002). Following qPCR quantification, libraries were amplified onto an Illumina paired-end flowcell, following the manufacturer's recommended conditions, and flowcells were sequenced, using customized PB5-ITR sequencing primer PB5prSeqR1 for read 1, TranSeqR2 for read 2 and an indexing sequencing primer RInv4 to decode the barcode of each library. The sequences of all primers in this section are shown in Table 2-6.

Table 2-6: Primers used for the multiplex Illumina sequencing to identify PB integration sites.

Name	Sequence
Ad_T_qPCR1 (Sigma, desalted)	5'- CTTCCCTACACGACGCTCTTC -3'
Ad_B_qPCR2 (Sigma, desalted)	5'- ATTCCTGCTGAACCGCTCTTC -3'
PB5pr_1	5'- GACGGATTCGCGCTATTTAGAAAGAGAG -3'
Primer PE_PCR_V4	5'- CAAGCAGAAGACGGCATAACGAGATCGGT [INDEX]
(Index has below 6 versions)	ACACTCTTCCCTACACGACGCTCTCCGATCT -3'
Index V4.1	5'- ACCTAG -3'
Index V4.2	5'- ACTGCT -3'
Index V4.3	5'- ATAGTG -3'
Index V4.4	5'- ATCAGA -3'
Index V4.5	5'- CAAGGT -3'
Index V4.6	5'- CCAACA -3'
PB5prP5_2	5'- AATGATACGGCGACCACCGAGATCTACACATGCGTCAATTTT ACGCAGACTATC -3'
PB5prSeqR1	5'- ATGCGTCAATTTTACGCAGACTATCTTTC -3'
TranSeqR2	5'- ACACTCTTCCCTACACGACGCTCTCCGATCT -3'
RInv4	5'- AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -3'

14. Bioinformatic analysis of Illumina sequencing data

The Illumina sequencing analysis was conducted by Zeming Ning from the Sanger Institute. One lane of Illumina pair-end reads of 2 x 70 bp was used to sequence all six sample libraries in a multiplexing manner. Raw sequences were grouped into six bins based on the barcode sequences incorporated in primer PE_PCR_V4. Each bin was analyzed separately. Read 1 sequences that contained the end of the PB5' ITR joining directly to the BAC vector sequence, representing the BAC random integration events, were excluded. Duplicated sequences are generated in Illumina sequencing by the PCR reactions. Duplicate reads was defined as sequences that possess identical start and end mapping locations on the reference mouse genome. The code `ssaha_duply` (<ftp://ftp.sanger.ac.uk/pub/zn1/transposon/codes/>) was developed to detect and remove duplicated segments from the raw read files. It combines two paired-end read files into one file, while two paired-ends are merged into one sequence.

This program then sorts the merged sequences to check the occurrence number and only one copy is maintained while other copies are removed.

SSAHA2 (<http://www.sanger.ac.uk/resources/software/ssaha2/>) was used to map paired-end Illumina reads against the mouse reference sequence NCBI_M37.fa. Before assembly, possible chimeric reads caused by ligation of two or more genomic sequences was filtered. For a true integration event, a genomic sequence joins immediately downstream of the PB5'ITR, and the genomic sequence should start with TTAA. The condition was set so that the query start matching point must be base position 6 or 5 of the read 1 (in genomic sequence, it is possible to have GTTAA just by chance). A full read match contains read lengths bigger or equal to 65 bp. The Cleaned and uniquely mapped reads were assembled using the pile-up utility of ssaha_pileup (ftp://ftp.sanger.ac.uk/pub/zn1/ssaha_pileup) to extract information, such as read coverage.

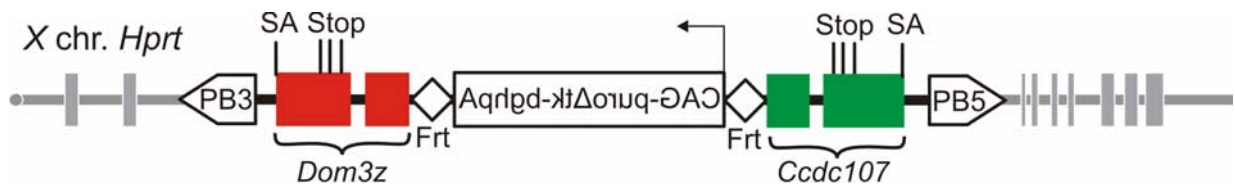
For a clean and contamination-free library, the results from the pileup file are reliable and accurate. However, cross-contamination can occur during the sample preparation steps prior to the PCR incorporation or during the barcode-primer synthesis and purification at source. Such contaminations can be directly reflected as cross-contamination in the sequences. Therefore, cross-library check for identical reads after PCR duplication filtering is an important step for the quality control.

Chapter Three - An efficient mutagenic strategy using the PB transposon

1. Introduction

The aim of this project is to establish a new mutagenic platform, coupled with the *Blm*-deficient ES cell system, to achieve genome-wide mutagenesis using the PB transposon in order to conduct recessive genetic screens. As described in Chapter One, there are several mutagenic agents and modular components available for generating efficient insertional mutagens. Four requirements have been taken into account in my mutagenic strategy: firstly, broad genome-coverage of the insertional mutagen; secondly, the requirement of highly effective gene inactivation; thirdly the ease with which the mutagen can be mapped and finally the establishment of the causal relationship of the genotype and phenotype. The molecular design of my transposon is depicted in Figure 3-1.

Figure 3-1: Design of an inactivating PB transposon targeted into the *Hprt* locus of *Blm*-deficient ES cells.



Dom3z, *ccdc107*, the terminal exon pairs and a portion of the 3' downstream sequence containing the endogenous polyadenylation signal from of the two computationally selected genes *Dom3z* and *ccdc107* respectively; Stop, three stop codons in three different reading frames; SA, endogenous splice acceptor of the penultimate exons of *Dom3z* and *Ccdc107*. The PB transposon was targeted into intron 2 of *Hprt* locus. The *Dom3z* exon pair traps the expression of *Hprt*, thus the resulting ES cells are sensitive to HAT.

1.1. Design principles of the mutagen

Broad genome-coverage of the insertional mutagen

Genome-wide mutagenesis requires the insertional mutagen to be delivered randomly with high efficiency, and two major factors determine this genome coverage. The first and major determinant is the integration characteristic of the insertional mutagen. As described in Chapter 1, the retroviral integration pattern in the mouse genome is significantly non-random, with severe integration “hot spots” (Hansen et al., 2008). The PB transposon integration pattern is significantly more random than retrovirus (Wang et al., 2008b). In addition, PB has a preference to insert into active transcribing units. This bias towards genes is advantageous for insertional mutagenesis. Together with its high transposition efficiency, PB transposon is a good candidate as a mutagen carrier for genome-wide mutagenesis.

There are two ways to deliver the PB transposon, “plasmid-to-genome” integration and intra-genomic mobilisation of a pre-engineered PB transposon within the genome. The “Plasmid-to-genome” delivery method has been successfully used to allow the recovery of homozygous mutants in *Blm*-deficient ES cells (Wang et al., 2008a). However, a careful titration is crucial to restrict the copy number of transposons to one per cell. The second method is intra-genomic mobilisation of the PB transposon. This is achieved by targeted insertion of a PB transposon within the genome, and the transposon is excised from the donor locus and re-integrates in a genome-wide fashion (Wang et al., 2008b; Liang et al., 2009). Re-mobilisation events from the donor locus are relatively inefficient (1 % of total cells electroporated), however they can be enriched using a selection marker that is activated when PB is excised. The outline of this strategy has been introduced in Chapter 1, Figure 1-12. The re-integration rate was approximately 40 % of the total excision events (Wang et al., 2008b; Liang et al., 2009). When the re-integration events were examined, local hopping was not detected when PB was mobilised from the *Hprt*, although a 10 % local hopping within 2.4 Mb window around the transposon donor site and 18 % residing on the donor chromosome have been observed at *Rosa26* locus (Wang et al., 2008b; Liang et al., 2009). Therefore, PB re-mobilisation from the *Hprt* locus for generating random mutations is an attractive strategy to conduct genome-wide mutagenesis with a tightly controlled copy number to mostly one.

The second determinant is the selection method used for cells with the insertional mutagen integrated. Various methods were discussed to select for loss-of-function mutations in Chapter 1. For instance, the use of the conventional gene trap system relies on the production of a functional reporter protein. This selection strategy eliminates integrations of the mutagen into “wrong” reading frames or “wrong” orientations with respect to the endogenous gene within which the mutagen is inserted. In addition, situations where fusion between the reporter and the endogenous gene have reduced or inactivated the reporter are also eliminated. In addition, integrations into non-expressing loci, genes with weak promoters to drive sufficient expression of the reporter, and non-coding genes will all be eliminated. Therefore, a mutagen in which selection for integration is independent from the requirement for gene inactivation will enhance the coverage of the mutagenesis. In this design (Figure 3-1), the selection marker (puro) for integration is independent from the trapping (mutagenic exon pairs).

Highly effective gene inactivation for loss-of-function screens

Because of the complexity of the mouse genome, which contains 3 billion base pairs and around 30,000 genes, it is essential to achieve efficient gene inactivation to ensure a high portion of integrations in genes result in null mutations. The high mutagenicity is ensured by the gene trapping unit containing a “strong” splice acceptor to “out-compete” the endogenous splice acceptor. The incorporation of stop codons and polyadenylation signals can lead to fusion transcript termination.

In this design (Figure 3-1), the mutagen is bi-directional, with two terminal-exon pairs as gene traps to inactivate gene expression in either orientation. The exon pairs can be selected for containing strong endogenous splice acceptors, polyA signals and stop codons in all reading frames to ensure efficient gene trapping and termination. Furthermore, the incorporation of stop codons before the final mutagen intron-exon splice junction may induce the endogenous non-sense mediated decay (NMD) for fusion transcript degradation. This may further ensure the generation of null mutation even when the mutagen is inserted towards the 3’ end of the transcript.

Mutagen mapping and establishment of causal genotype-phenotype relationships

Upon isolation of phenotypic mutants, it is essential to be able to identify the causal gene(s) disrupted by the mutagen. PB transposons provide a unique molecular tag to identify integration sites. However, if there are multiple integrations per cell, it is difficult to assign the causal mutation. Therefore, a single integration of the mutagen per cell is ideal. With the “plasmid-to-genome” integration method, single copy of PB transposon per cell can be achieved in the majority of the cells with careful titration of the amount of the transposon and transposase. However, the balance between efficient genomic integration and the single copy transposon per cell is difficult to strike. Intra-genomic mobilisation of PB transposon can maintain a single copy of mutagen per cell stably and is not restrained with the efficiency of generating sufficient number of mutations.

Precise transposon excision with *Puro Δ tk*-mediated selection

PB transposon excision does not generate any footprint. This unique property of PB can be utilised in complete mutant reversion experiments. In mutants with both alleles inactivated (homozygous), two copies of the PB transposon from both alleles of the same locus can be excised. However, the efficiency of obtaining such type of complete reversion events is very low. The incorporation of the *Puro Δ tk* cassette (Figure 3-1) allows the efficient isolation of genotype revertants using the drug FIAU to select for PB transposon removal from the genome. The reversion of the disrupted gene can directly address the genotype and phenotype causal relationship, as mutagen removal should lead to the phenotype reversal to wild-type if the mutagen is the cause of the mutant phenotype.

1.2. The DNA mismatch repair pathway as a screening model for proof-of-principle

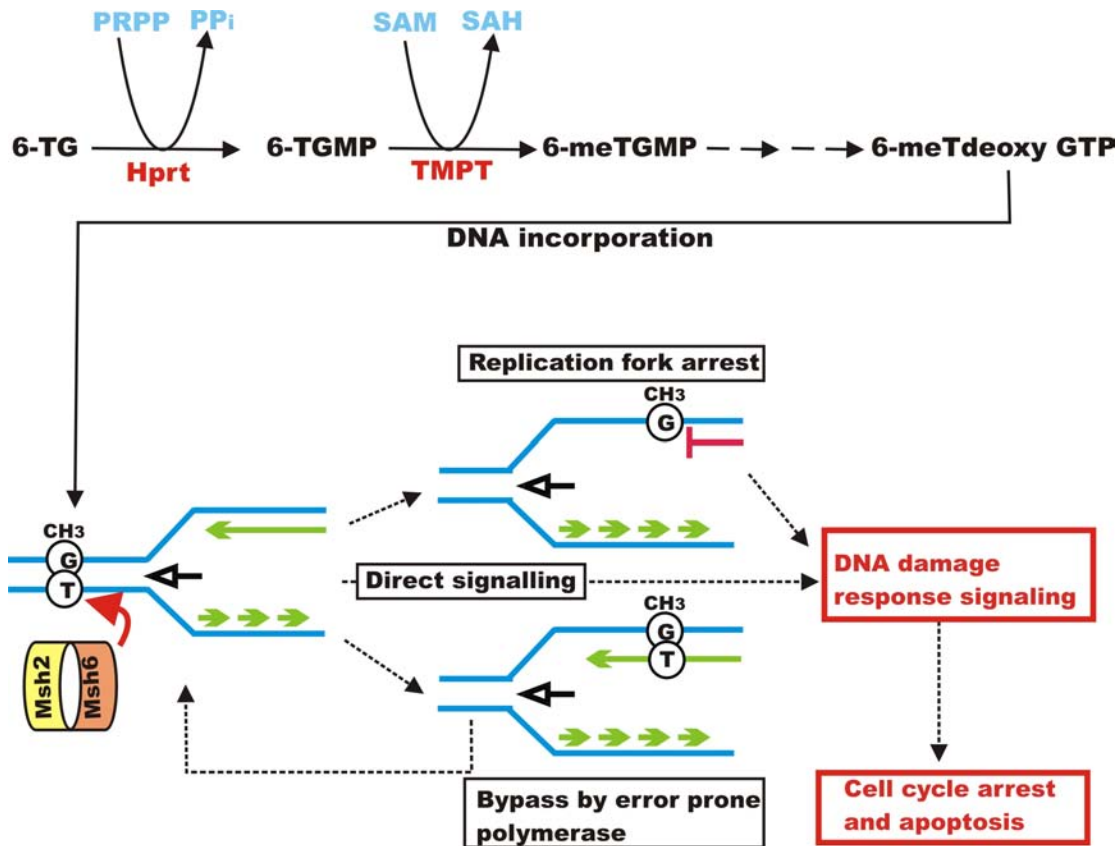
DNA mismatch repair (MMR) is an evolutionarily conserved pathway from bacteria to mammals, which exists to maintain genome integrity during DNA replication and to sense and repair certain types of DNA damage. It has been observed that MMR-deficient cells are more resistant than matched MMR-proficient cells to 6-thioguanine induced cell death (Griffin et al., 1994; Jiricny, 2006). The 6-TG resistant phenotype of MMR-deficient cells has been used

as a phenotypic selection strategy to isolate homozygous mutants in the MMR pathway (Guo, 2004; Wang et al., 2008a).

The mechanism of 6-TG mediated genotoxicity is described below and shown in Figure 3-2. The pro-drug 6-TG can be actively up-taken by cells and enter the purine salvage pathway. Hypoxanthine-guanine phosphoribosyltransferase (Hprt) catalyses the conversion of 6-TG to 6-thio guanosine monophosphate (6-TGMP). 6-TGMP can be methylated by thiopurine S-methyltransferase (TPMT) through methyl group transfer from the co-enzyme S-adenosylmethionine (SAM) to form 6-methyl thioguanine monophosphate (6-meTGMP). 6-meTGMP can then be subsequently processed to form 6-meTG deoxynucleotide, which can be incorporated into DNA during DNA replication. 6-meTG pairs with T during replication, causing a single base-pair mismatch. In wild type cells, such a mismatch can be recognised and the MMR system is activated for repair. The MMR system may also be directly involved in the cell-cycle checkpoint to trigger cell-cycle arrests and apoptosis. If the modified base is in the template strand, the mismatch will be repeatedly generated by the polymerase and subsequently cause replication fork arrests (Jiricny, 2006). ATR/CHK1-dependent checkpoint signalling cascades are responsible for sensing replication blocks and mediating cell-cycle arrest and apoptosis in unsuccessful mismatch repair attempts (Jiricny, 2006). In addition, the damage could be bypassed by an error-prone DNA polymerase and the DNA damage signalling could be triggered by the repair in the subsequent rounds of replication.

In MMR-deficient cells, the mismatch is not recognised and it does not induce DNA damage signalling; therefore these cells survive in the presence of a genotoxic agent such as 6-TG but the cells bear a large number of mutations in their genome.

Figure 3-2: The mechanistic basis of 6-TG mediated cytotoxicity through MMR system.



See main text for the detailed description. PRPP: 5-phospho-ribosyl-1 α -pyrophosphate; PPi: pyrophosphate ion; SAM: S-adenosyl methionine; SAH: S-adenosylhomocysteine; TMPT: thiopurine S-methyltransferase; 6-TGMP: 6-thio guanosine monophosphate; 6-meTGMP: 6-methylthioguanine monophosphate; 6-methyl-thio deoxy GTP: 6-methylthio deoxyl guanine triphosphate.

In this chapter, I describe how the mutagenic units used in the transposon were identified and how these cells were assessed functionally. The MMR pathway was used as a proof-of-principle screening system to validate this newly established mutagenic strategy.

2. Results

2.1. Mutagenic exon-pair selection

Selection of mutagenic exon-pair was achieved by scanning the annotated mouse genome computationally for terminal exon-intron structures which met defined criteria for efficient mutagenesis. This part of the work was done in conjunction with Steve Pettitt, another graduate student, in the Bradley group.

The computational selection criteria for the terminal exon pairs are as follows:

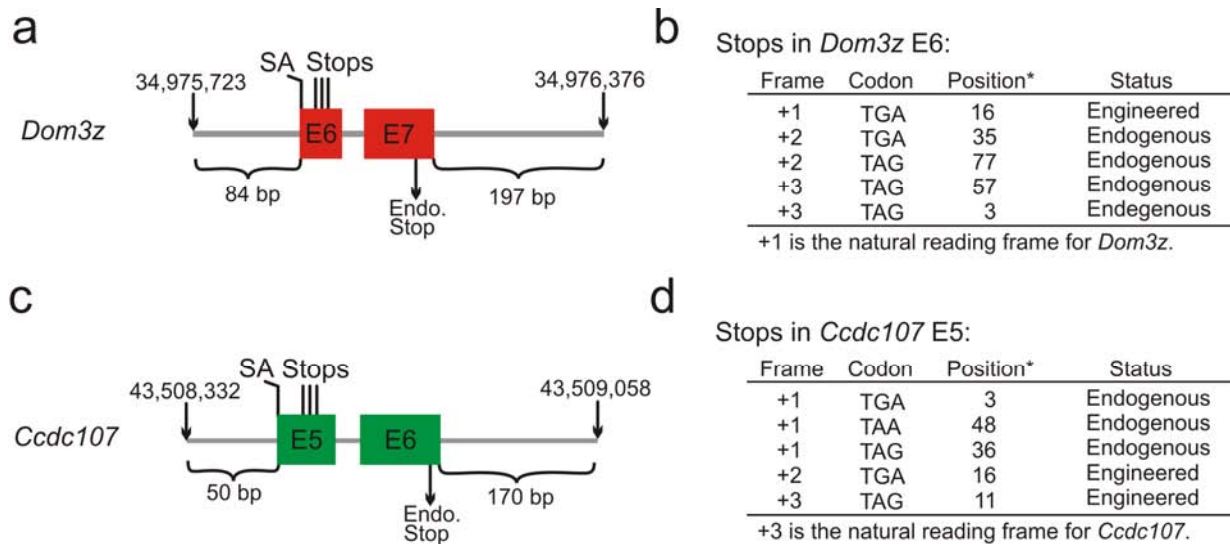
- Terminal exon pairs together with the intron in between the exons are defined as a mutagen unit.
- The mutagen unit must be relatively small (less than 2000 bp) to ensure highly efficient PB transposition.
- The mutagen unit must be present in all transcript variants of the endogenous gene, i.e. the mutagen exons are constitutively spliced, to ensure that the splice acceptors are relatively strong.
- Stop codons are present in the penultimate exon of the other two non-natural reading frames around 50 to 55 bp from the exon-intron junction.
- The mutagen unit must not contain any domain structure or structure/signal for cellular organelle localisation.

A short list of genes which contained mutagen units fitting the top four criteria were then manually curated at the protein sequence level and through literature searching to confirm the candidates that satisfy the final criterion. This was to avoid any possible fusion protein product having dominant-negative functionality which would complicate the mutant phenotype. Out of eight finalists, the two smallest mutagen units from the genes *Dom3z* and *Ccdc107* were selected.

Dom3z is located in the gene dense MHC class III regions on Chromosome 17, coding for a protein which is homologue to *C. elegans* DOM-3 and human *DOM3Z*. It binds to a nuclear

exoribonuclease, involved in the processing of 5.8S rRNA (Xue et al., 2000). *Ccdc107*, short for Coiled-coil domain-containing protein 107 precursors, resides on Chromosome 4 and this protein's function is currently unknown.

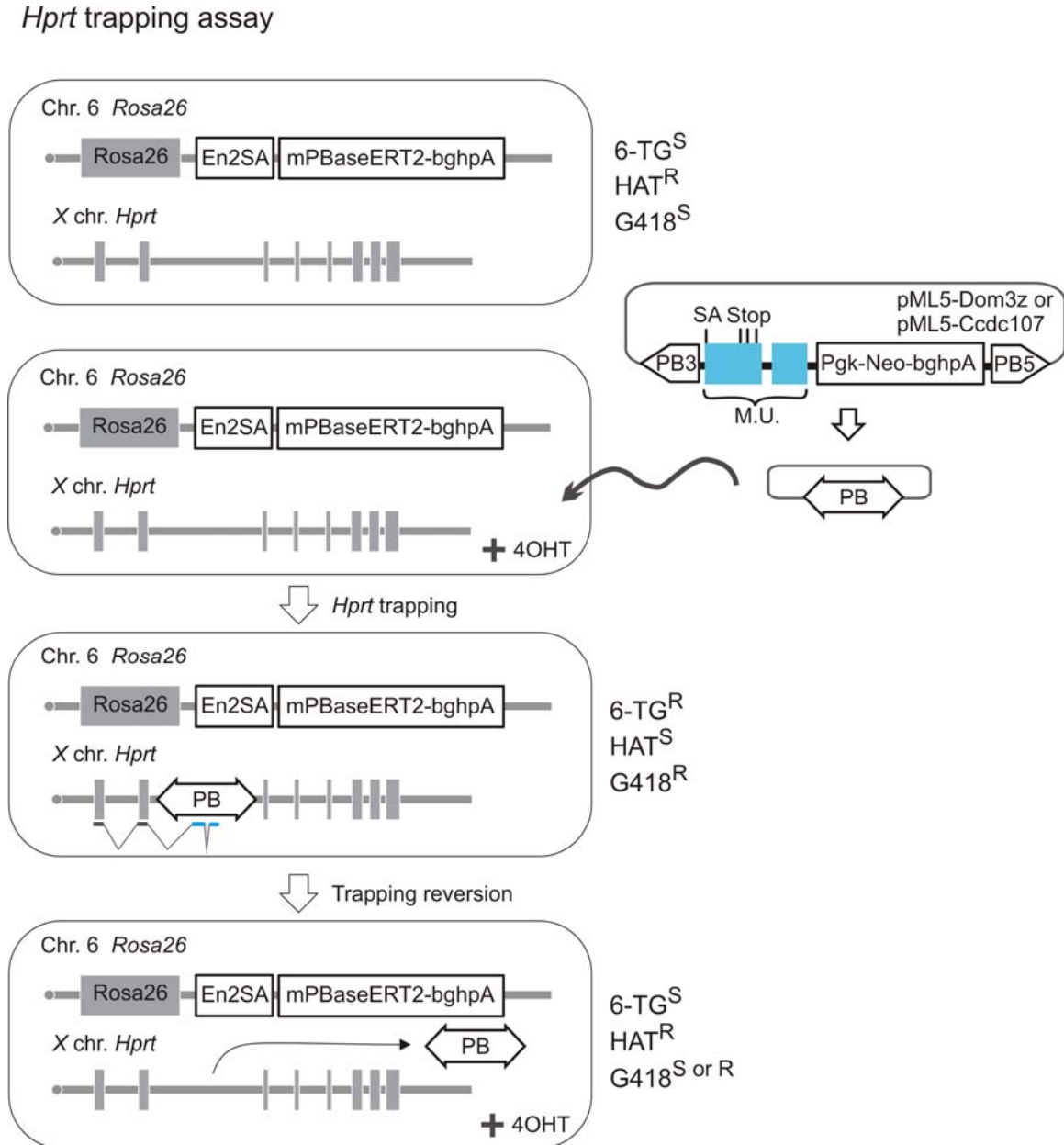
Sequence modifications were also conducted to introduce additional stop codons into the natural coding frames of the two mutagen units by site-directed mutagenesis. This was done by Steve Pettitt. Thus both mutagen units contain stop codons in all reading frames before the final exon-intron junction. This is an observed structure, which can induce a natural surveillance mechanism, nonsense mediated decay (NMD) pathway, to degrade mRNAs with premature stop codons (PTCs) (Chang et al., 2007). Thus these stop codons in the penultimate exon may be recognised as PTCs to induce the NMD pathway and degrade the trapped fusion transcripts. This feature may be particularly useful when the mutagen is inserted in the 3' end of a gene, and the fusion product may have sufficient endogenous protein structures to function normally. Figure 3-3 shows the structures and the locations of the stop codons of the mutagen pairs.

Figure 3-3: Mutagen *Dom3z* and *Ccdc107* exon-pair structures.

a and c, Schematic diagrams showing the mutagen structures. The coordinates indicate the genomic positions of the fragments based on NCBI37. SA, splice acceptor; Stops, stop codons; Endo. stop, endogenous stop codon. b,d, Endogenous and engineered stop codons within the penultimate exon of the mutagens to cover all reading frames. Position* indicates the nucleotide position of the “T” in the stop codons. Position 1 defines as the first nucleotide of the penultimate exon.

2.2. Validations of mutagen units using an *Hprt* trapping assay

Since the mutagen units were computationally selected, it was critical to assess their gene-inactivation efficiencies experimentally. The first assay, an *Hprt* trapping assay was designed to test the strength of the splice acceptor in competition with the endogenous splice acceptor of the trapped gene for generating null mutations, Figure 3-4.

Figure 3-4: Schematic representations of the *Hprt* trapping assay.

A saturating amount of PB transposon plasmids were introduced into an *Hprt*-proficient ES cell line (AB1) previously targeted with an inducible mPBBase (mPBBaseERT2) in the *Rosa26* locus, 3' to the *Rosa26* endogenous promoter. mPBBase can be induced with 4-OHT which results in its nuclear localisation. *Hprt* trapping events can be selected with drugs G418 and 6-TG. If the 6-TG resistant phenotype is caused by PB transposon inserting within *Hprt*, phenotypic rescue to HAT resistance is possible by reactivating mPBBase by 4-OHT to remobilise the PB transposon out of the *Hprt* gene.

An ES cell line expressing an inducible form of the mPBase (AB1-R26^{mPBaseERT2/+}) was used for this assay, which provides a temporal control of transposase activation. In this cell line, mPBase was fused with a modified version of the human estrogen receptor ligand binding domain to its C-terminus (Cadinanos and Bradley, 2007). The mPBaseERT2 is driven by the constitutively active endogenous *Rosa26* promoter (Cadinanos, unpublished).

2.2.1. *Hprt* trapping assay with 6-TG and G418 dual selection

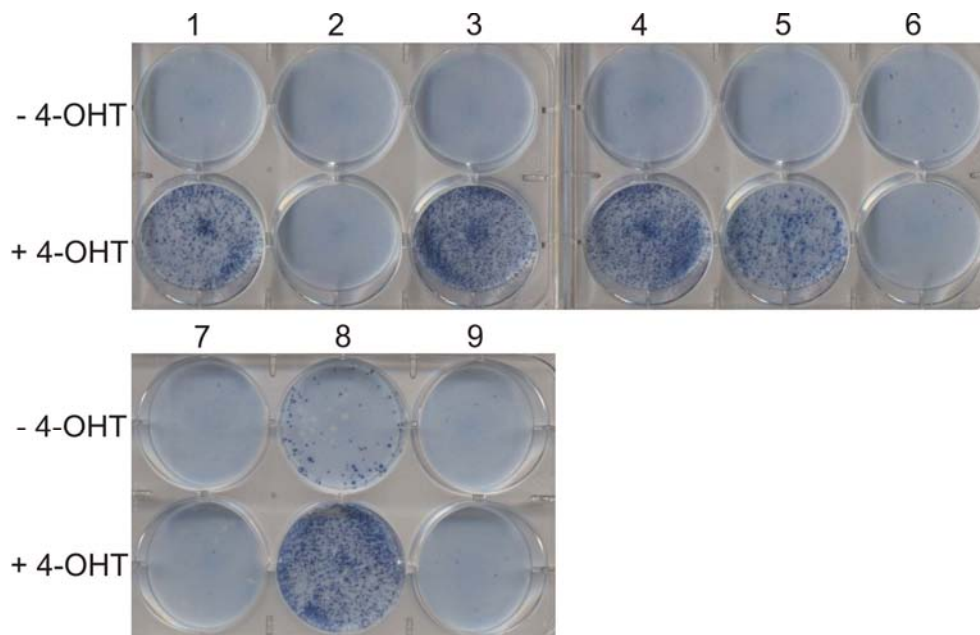
Two constructs (pML5-Dom3z and pML5-Ccdc107) were made with PB transposon harbouring the individual mutagen exon-pairs together with a NGK-Neo expression cassette. Integration of these mutagen-containing PB transposons into the *Hprt* locus with the correct orientation can inactivate the *Hprt* expression, thus these cells can be selected with 6-TG to access the trapping ability.

In order to maximise the probability of transposon integration into the *Hprt* locus, a saturating amount (50 µg) of pML5-Dom3z and pML5-Ccdc107 plasmids were independently electroporated into 1×10^7 AB1-ROSA26^{mPBaseERT2/+} cells. Electroporated cells were plated in six individual 90 mm plates. 1 µM 4-OHT was added 16 hours post-electroporation and sustained for two days. G418 selection was initiated 24 hours post-electroporation and was applied for five days to select cells with PB transposons integrated. G418 and 6-TG double selection was applied from day six onwards until colonies formed. Five clones were obtained from pML5-Dom3z transfected cells and four clones from pML5-Ccdc107 transfected cells.

The double resistant clones were further tested for the reversion of *Hprt* activity to establish the causality of PB transposon integration and the *Hprt* inactivation. The double resistant cells were plated in duplicate and 1 µM 4-OHT was applied o/n to one plate of the pair. HAT selection was added to all cells and was initiated 24 hours after the first application of 4-OHT. Five out of nine clones could be reverted to HAT resistance, suggesting the rest had inactivated *hprt* as a result of negative background mutations, Figure 3-6.

Splinkerette PCR was used to try to identify PB integration sites in the *Hprt* locus for the five *Hprt*-deficient clones which yielded revertants. In these experiments, a large amount of PB plasmid was used, thus there will be many PB integrations per cell. The Splinkerette PCR reaction does not amplify all insertion sites evenly, thus making it difficult to identify specific integrations in *Hprt*.

Figure 3-5: Reversion analysis of the 6-TG resistant colonies.



1-4 were clones isolated from pML5-Dom3z electroporation; 5-9 were clones isolated from pML5-Ccdc107 electroporation. The colonies were plated in duplicate and incubated overnight with or without 4-OHT. The next day, the cells were selected using HAT.

2.2.2. 4-OHT induction optimisation

It has been previously estimated that 47 % of the PB integrations reside in genes and 80 % of the insertions in genes land in actively transcribed genes in ES cells (Liang et al., 2009). Based on this information, the number of cells with PB insertions required to obtain cells with the *Hprt* locus inactivated can be estimated, Table 3-1. The calculation assumes that the average copy number of the independent PB integrations is five when using the PB transposon saturation transfection method, the trapping efficiency of the mutagen is 100 %, and only one

Hprt trapping event per cell can occur. Using this calculation, the total number of cells with PB inserted required to generate five *Hprt*-trapped clones is 1.5×10^5 cells. With the protocol used with the saturating amount of PB transposon, the number of cells harbouring integrated PB transposon within their genome can be over 10 % of the survival cells after electroporation (Wang et al., 2008b), i.e. at least 5×10^5 cells should contain PB integrations with the 50 % survival rate post-electroporation. Therefore, per mutagen tested, at least 15 *Hprt* trapped cells should be generated per 1×10^7 cells electroporated with the protocol used. This estimation indicates that the actual efficiency of obtaining *Hprt* trapped clones was low in this experiment.

Table 3-1: Estimate of number of cells required to obtain *Hprt* trapping events.

<i>Hprt</i> trapping efficiency estimate	
Total No. of cells with PB inserted	N
No. of cells with PB integrated, assuming 5 insertions/cell	5N
No. of insertions in genes (47 %)	2.5N
No. of insertions in actively transcribed genes in ES cells (80 %)	2N
<i>Hprt</i> size proportional to the actively transcribed genome size (34 kb/ 1×10^6 kb*)	$6.8 \times 10^{-5}N$
Correct orientation to trap <i>Hprt</i> , assuming 100% trapping efficiency (50%)	$3.4 \times 10^{-5}N$
Value of N, if <i>Hprt</i> trapping in one cell is obtained	3×10^4 cells

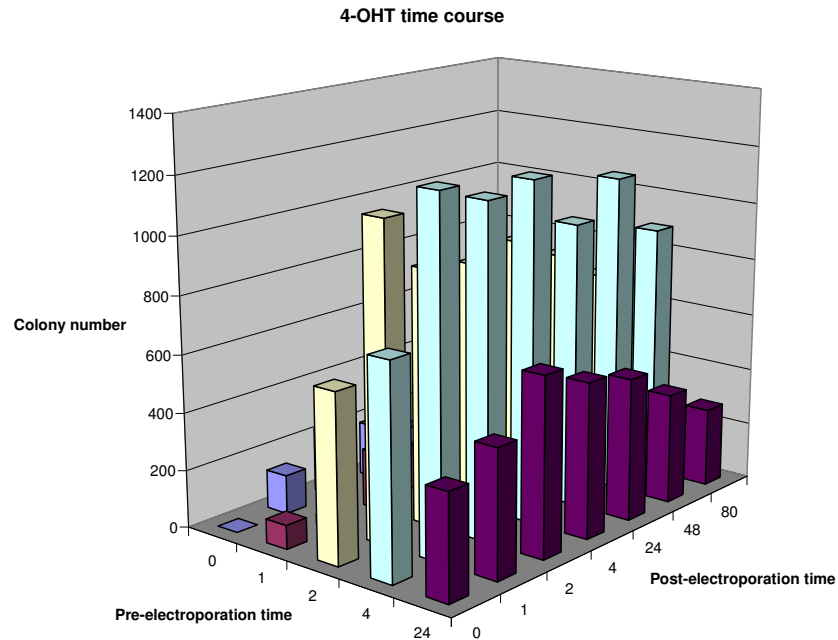
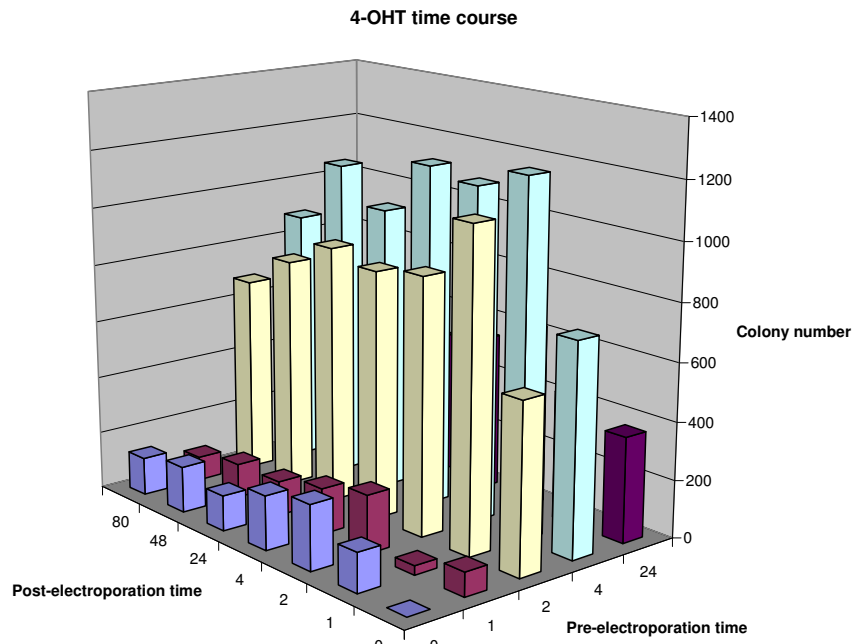
This table represents the *Hprt* trapping efficiency calculation. * 1×10^6 bp is the total size of the genes that are actively transcribed in ES cells, and the size of each gene is defined as the absolute value of the differences between start and stop codon positions. The actively transcribed gene list and their start and stop codon coordinates were obtained from the gene trap consortium (<http://www.sanger.ac.uk/Postgenomics/genetrap/>).

The low efficiency with which the 6-TG resistant colonies were obtained could be due to a suboptimal 4-OHT induction protocol. Additionally, the high cell number plated may also give false-negative results due to the cross-killing Hprt-deficient cell by adjacent Hprt-proficient cells.

In order to optimise the induction condition, a pre- and post- electroporation time-course matrix of 4-OHT was set up to obtain the optimal 4-OHT induction protocol in order to achieve high transposition efficiency. For each time point, 1×10^7 AB1-ROSA26^{mPB_{Base}ERT2/+} cells were incubated with 4-OHT for the length of time required 0, 1, 4, and 24 hours before electroporation. Subsequently 1 μ g of a plasmid with PB transposon harbouring PGK-Neo expression cassette (pML5) was electroporated. One twentieth of the electroporated cells were plated in six-well plates with 4OHT present for post-electroporation time points 0, 1, 2, 24, 48, and 80 hours. G418 selection was initiated 24 hours post electroporation. The results are summarised in Figure 3-6.

Short term pre-electroporation incubation with 4-OHT between two to four hours showed a significant elevation of transposition events overall. The dramatic effect of pre-electroporation incubation with 4-OHT provides an indication of the fast kinetics of PB transposition. The effect of post-electroporation incubation with 4-OHT was less overt, but prolonged incubation with 4-OHT showed a trend of reducing number of G418 resistant colonies. This may be due to continuous PB transposition, with each excision event having 40 - 50% probability of transposon lost. Overall, the optimal 4-OHT induction condition was a four-hour pre-electroporation incubation coupled with a 24 hour 4-OHT post-electroporation incubation.

Figure 3-6: 4-OHT induction protocol optimisation.



The top panel is viewed from the front while the bottom panel is viewed from the back.

2.2.3. *Hprt* trapping assay with sequential G418 and 6-TG selections

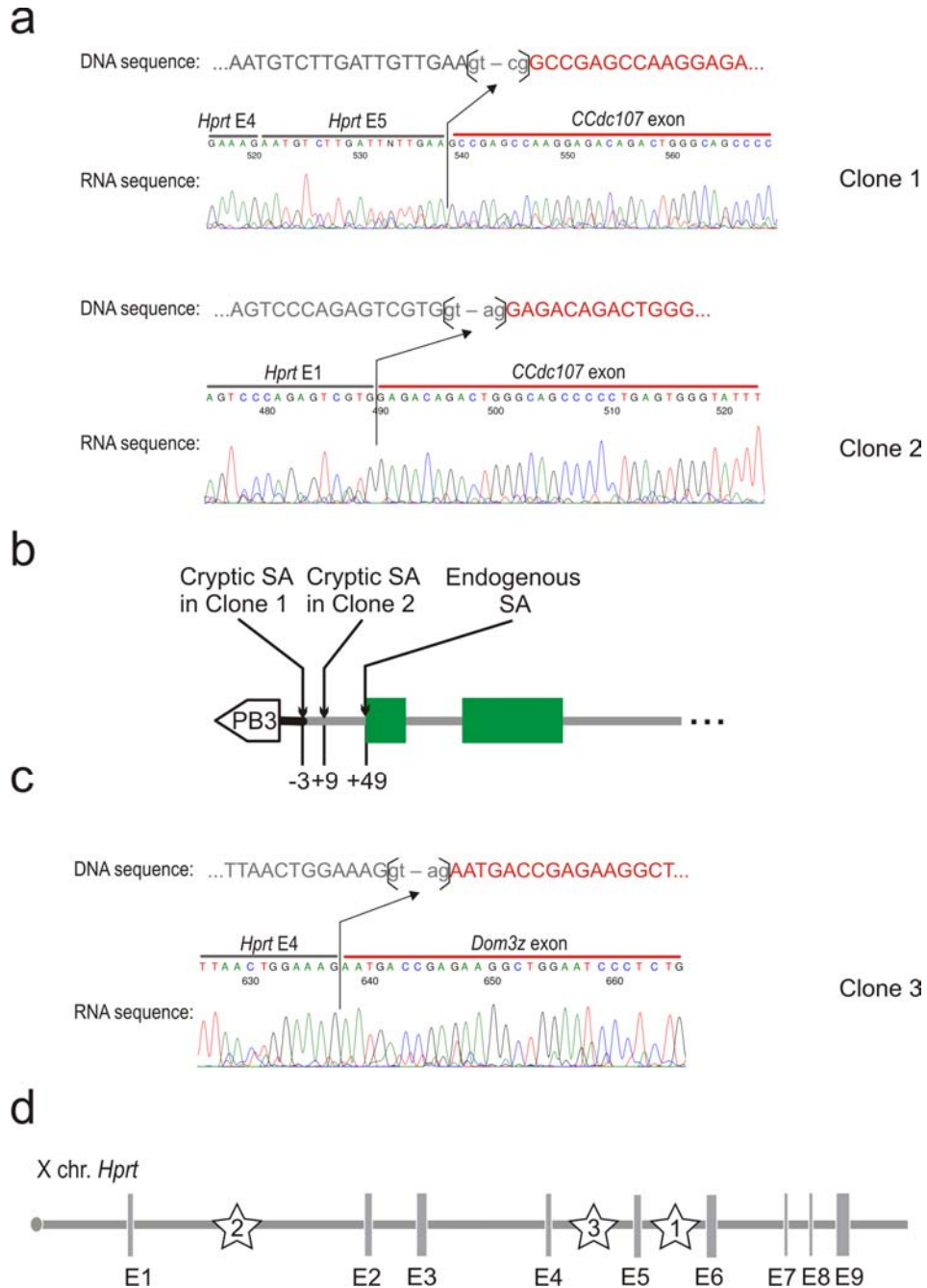
Sequential selection was used with the optimised 4-OHT induction in order to maximise the efficiency of obtaining the *Hprt* trapped clones. Selection using G418 a few days prior to G418 and 6-TG dual selection has several advantages. Firstly, initial G418 selection can eliminate a large proportion of non-transfected cells, thus enriching the cells with PB transposon integrated. Secondly, the replating of G418-selected cells at low density can reduce the cross killing of *Hprt*-deficient cells by the neighbouring *Hprt*-proficient cells. In addition, delayed selection of 6-TG allows the decay of *Hprt* mRNA and protein to avoid false negative results. Finally, PB transposon may have a delayed integration into the *Hprt* locus to give rise to a sub-population of the sibling cells within a clone containing *Hprt*-proficient cells. Without replating, these *Hprt*-trapped daughter cells can be crossed killed under 6-TG selection by sharing metabolites with their *Hprt*-proficient sibling cells.

AB1-ROSA26^{mPB_{Base}ERT2/+} cells were treated with 4-OHT following the optimised protocol to induce PB_{ase} subsequently. 50 µg of pML5-Dom3z and pML5-Ccdc107 plasmids were electroporated into 1x10⁷ 4-OHT induced cells, independently. The electroporated cells were plated initially on a 90 mm plate and G418 was initiated 24 hours post-electroporation. G418 selection was sustained for four days, and colonies started to emerge. Cells were then trypsinised and these cells were plated in eight 90 mm plates at 2.5x10⁵ cells per plate. 6-TG and G418 selection was applied directly after replating, and the selection was sustained for 12 days. In total, 20 double resistant colonies were obtained from the pML5-Dom3z transfected cells and 18 from pML5-Ccdc107 cells. The double resistant colonies were subjected to the reversion analysis as described before. Following 4-OHT induction, five clones yielded *Hprt* revertants in pML5-Dom3z transfected cells while six clones were obtained from pML5-Dom3z transfected cells. The non-revertants could due to the spontaneous null mutations of the *Hprt* locus.

Splinkerette PCR was performed to identify the PB integration sites within the *Hprt* locus. Three clones (one from pML5-Dom3z and two from pM5-Ccdc107) out of the 11 yielded the PB-*Hprt* genomic junction fragments. RT-PCR was performed using a mutagen exon-specific

primer (either Dom3z or Ccdc107 depending on the orientation of the integration in relation to the *Hprt* transcription) and an *Hprt*-exon-specific primer on these three clones to confirm the trapping of *Hprt* and this provided an opportunity to access the splicing structures of the fusion transcripts, Figure 3-7. Sequencing of the RT-PCR products of the fusion transcripts confirmed that the splice junction between the *Hprt* and Dom3z mutagenic exons were as predicted with the endogenous splice acceptor mediating the trapping (Clone 3, Figure 3-7b). However, in neither case of the Ccdc107 mutagen-mediated trapping, was the endogenous splice acceptor from the penultimate Ccdc107 exon used. Instead, two independent cryptic splice acceptors were identified (Clone 2 and Clone 3, Figure 3-7a). The cryptic splice acceptor identified from Clone 1 was present at the cloning junction between the endogenous DNA fragment containing *Ccdc107* and the vector sequence; whereas the cryptic splice acceptor observed in Clone 2 was within the endogenous intronic sequence of *Ccdc107* (Figure 3-7a,b). Both these splice acceptors were located 5' of the endogenous splice acceptor of the exon 5 of the *Ccdc107* gene.

Figure 3-7: PB transposon-mediated *Hprt* trapping.

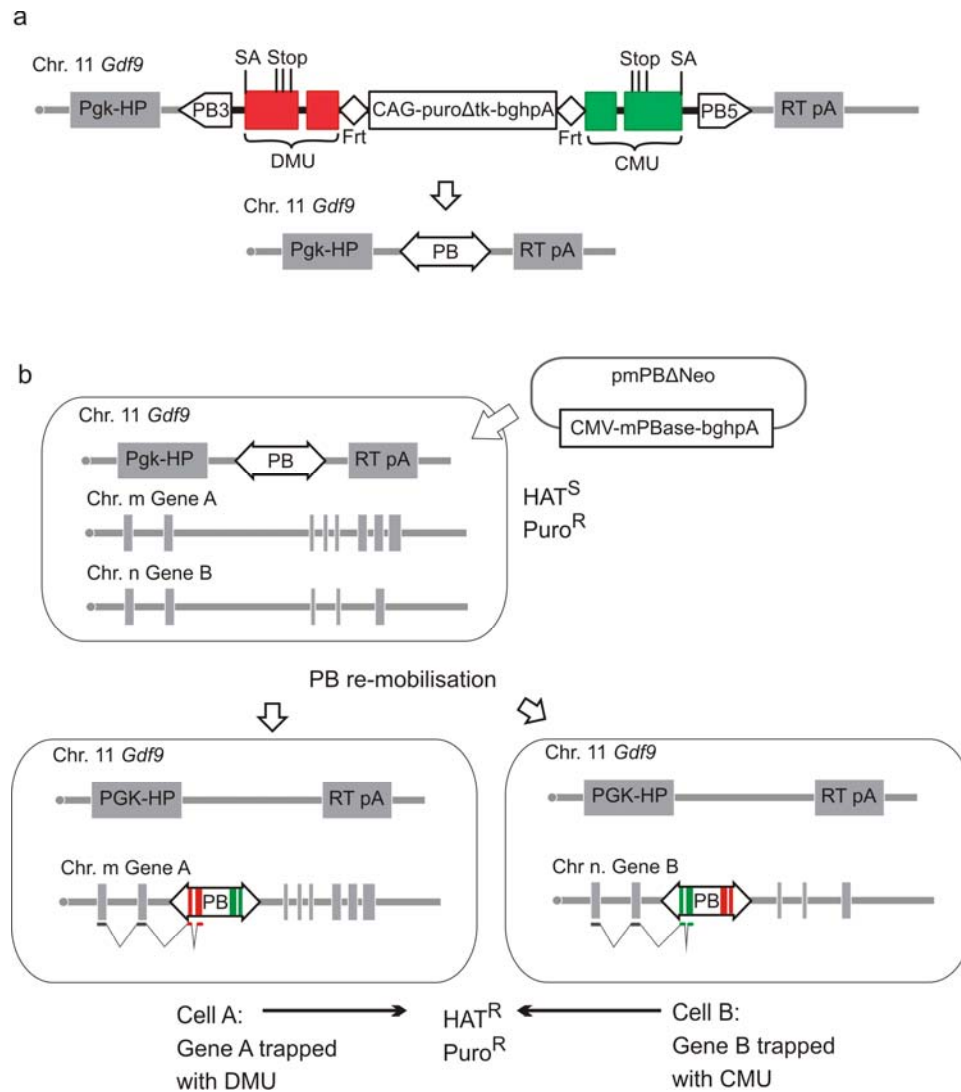


a,c, RT-PCR sequence traces of the three revertible 6-TG resistant clones. b, schematic summary of the two cryptic splice acceptors identified with the *Ccdc107* mutagen. The number indicates the location of the splice acceptors with the first base from the cloned endogenous *Ccdc107* sequence to be “0”. The base position is referred to the first “G” in the exonic sequence after the splice junction. d. Summary of the integration sites of the three revertible 6-TG resistant clones in the *Hprt* locus, with the stars representing the integration sites and the number within the star denoting the clone I.D.

2.3. Validation of mutagen units using a random trapping assay

The second validation assay, a random trapping assay was designed to test whether the endogenous splice acceptors of the penultimate exons from *Dom3z* and *Ccdc107* are capable of trapping endogenous genes in different genomic contexts. The schematic representation of the assay is shown in Figure 3-8.

Figure 3-8: Schematic representation of the random trapping assay.

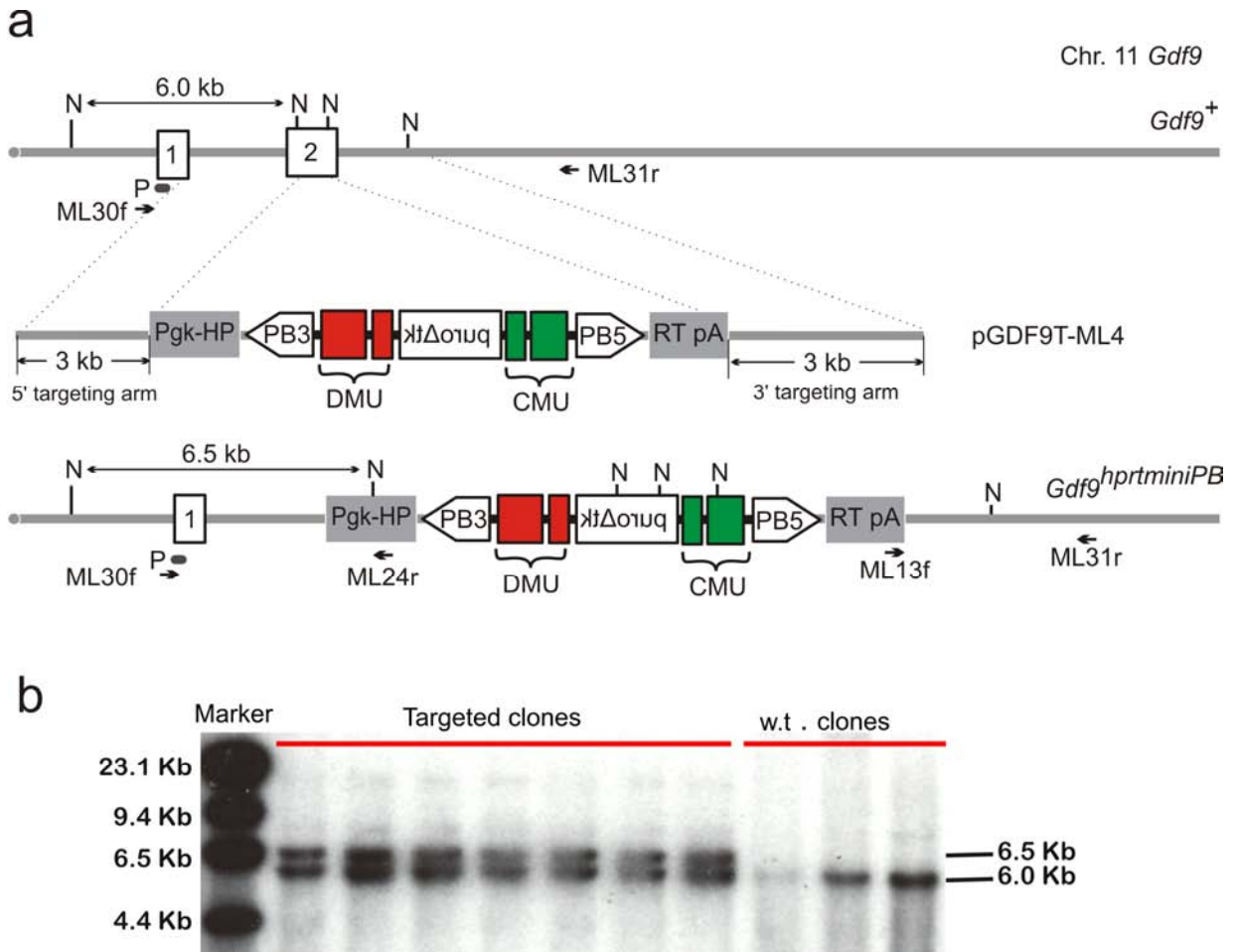


a, The mutagenic PB transposon was targeted into the *Gdf9* locus, flanked by the 5' and 3' part of the *HPRT* minigene cassette. HP, 5' part of the human *HPRT* cDNA; RT-pA, 3' part of the human *HPRT* cDNA with polyA signal. b, Post mPBbase introduction, PB re-integration events in introns can be used to characterise the trapping capability of the mutagen units (DMU, Dom3z mutagen unit and CMU, *Ccdc107* mutagen unit) in different genomic contexts. Both Dom3z and *Ccdc107* mutagen units can be assessed independently.

2.3.1. Generation of a cell line with an autosomal single-copy PB transposon

A cell line was made in order to conduct the random trapping assay. A single copy of the mutagenic PB transposon was targeted into the autosomal locus *Gdf9*, flanked by the two portions of the *HPRT* minigene, Figure 3-9a. A targeting vector pGDF9T-ML4 was constructed by inserting the mutagenic PB transposon into an existing *Gdf9* targeting vector by recombineering (Chapter Two). The targeting vector was linearised by *PmeI*, purified by ethanol precipitation and electroporated into 1×10^7 NN5 ES cells. The NN5 cell line was derived from AB2.2, in which the endogenous *Hprt* locus was inactive (Kuehn et al., 1987). Puromycin selection was applied 24 hours post-electroporation and 48 colonies were analysed by long-range PCR for both the 5' and 3' junctions of the targeting arms to the contiguous genomic regions. Seven clones with positive results for long-range PCR for both targeting arms were further confirmed by Southern blotting to be positive, thus the targeting efficiency is 15 % at this locus using this construct, Figure 3-9. The resulting cell line was termed NN5-*Gdf9*^{hprtminiPB/+}.

Figure 3-9: Gene targeting of the *Gdf9* locus to generate the NN5-*Gdf9*^{hprtminiPB/+} cell line.



a, Cartoon representation of the DNA structures of the wild type *Gdf9* allele (*Gdf9*⁺) and the targeted allele (*Gdf9*^{hprtminiPB}). N, *NcoI* recognition site; P, southern probe; DMU, Dom3z mutagenic unit; CMU, Ccdc107 mutagenic unit; ML30f and ML24r are primers used for 5' long-range PCR; ML13f and ML31r are primers used for 3' long-range PCR. b, Southern blot confirmation of the targeted clones. Marker, Lambda *HindIII* ladder.

2.3.2. The random trapping assay

1x10⁷ NN5-*Gdf9*^{hprtminiPB/+} cells were electroporated with 25 µg of mPBase containing plasmid (pmPBΔNeo), and the electroporated cells were plated in a 90 mm plate and selected with HAT and Puro the next day for the clones with PB transposon has been excised from the *Gdf9*-HPRT locus and re-integrated in the genome. In total, 1,300 colonies were obtained from the electroporation. Ninety six Double-resistant colonies were picked and were subjected to

Splinkerette PCR to identify the locations of re-integrated PB transposons. All clones with uniquely identifiable genomic locations (84 clones) were further analysed for insertions within introns in genes which were transcriptionally active in ES cells based on the genome-wide ES cell gene trap resource (with a total of 165,778 trapped events are present in the database). Clones with PB transposons present in introns were examined and five selected in which the *Dom3z* mutagenic unit is in the same orientation as the gene, while another five clones were selected on the basis of the orientation of the *Ccdc107* mutagenic unit, Table 3-2.

The ten clones were subjected to RT-PCR analysis to determine the efficiency of gene trapping in these randomly-selected genes using a gene-specific upstream primer with a mutagen specific terminal-exon primer. The chromosomal locations and the position of the PB transposons with the selected genes were described in Table 3-2, and Figure 3-10a shows the RT-PCR results.

Table 3-2: Ten PB reintegration clones selected from the random trapping.

Gene	Chr	Location	Start	End	Mutagen	Clone
<i>Tmem131</i>	1	Intron 4	36919002	36919531	<i>Ccdc107</i>	D7
<i>Dut</i>	2	Intron 4	125082204	125082534	<i>Ccdc107</i>	G10
<i>Fstl4</i>	11	2 nd last intron	52993736	52994119	<i>Dom3z</i>	E8
<i>A24Rik</i>	15	Intron 17-18	72649435	72649901	<i>Dom3z</i>	C4
<i>Sfrs3</i>	17	Intron 2	29173721	29174157	<i>Ccdc107</i>	E4
<i>Sema6a</i>	18	Intron 1	47487960	47488080	<i>Dom3z</i>	F8
<i>Undcd3</i>	11	Intron 1	6097207	6097470	<i>Ccdc107</i>	H8
<i>stk22s1</i>	7	Intron 6-7	52208693	52208790	<i>Dom3z</i>	A8
<i>Ran</i>	5	Intron 4	129527403	129527558	<i>Dom3z</i>	F2
<i>Tmem50a</i>	4	Intron 1	134469873	134470347	<i>Ccdc107</i>	B3

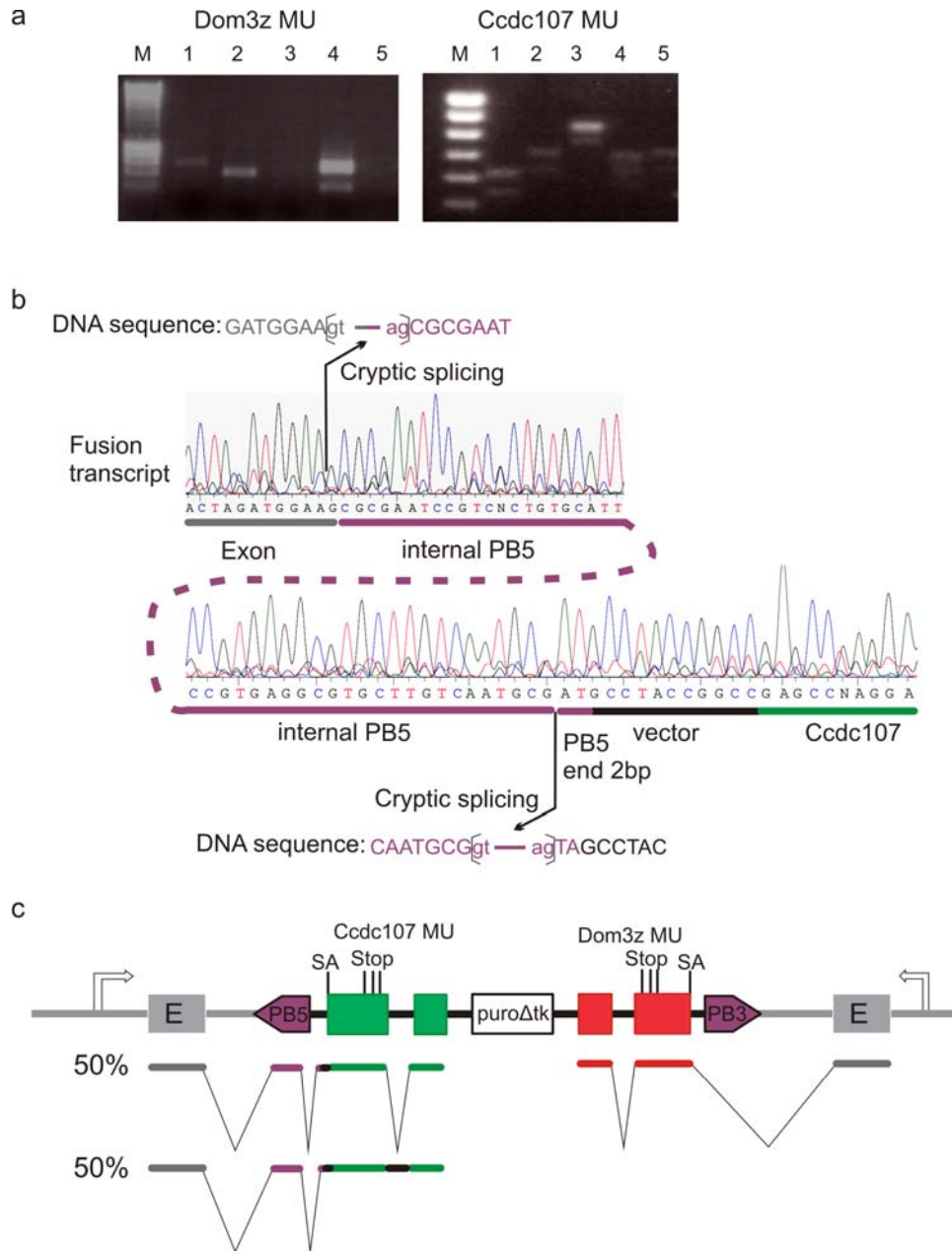
The locus-specific primers for these clones can also be found in Table 2-4 Chapter Two. NCBIm36 was used for the mapping of the PB integration sites.

In clones where Dom3z was the trap, three out of the five clones analysed showed PCR products representing the fusion products. Sequencing of the RT-PCR fusion products confirmed that the endogenous splice acceptor from the Dom3z penultimate exon was mediating the trapping as predicted. Two clones did not yield any PT-PCR products. The minor product in Clone 4 was non-specific PCR amplification. There are two possibilities to explain this, firstly, the failure for Dom3z to mediate trapping and secondly highly efficient NMD effect. Further experiments will be required to distinguish the two possibilities.

In clones where the Ccdc107 mutagen unit mediated trapping, all clones generated fusion products (Figure 3-10a). However, unexpectedly, every clone produced doublet PCR bands with similar intensity. Sequence analysis of these PCR products revealed further complexity of these fusion transcripts. The two PCR products were produced due to the alternative splicing of the intron between the two Ccdc107 terminal exons. More interestingly, a cryptic splice acceptor and donor were present within the PB5 sequence, which trapped transcripts coming from the upstream exon in all five cases, Figure 3-10b,c.

Taken together, the trapping mediated from the Ccdc107 orientation did not occur as predicted due to the presence of cryptic splice acceptors within the PB5 sequence and the alternative splicing of the intron between the mutagenic terminal exons. However, trapping is observed in all random loci tested. The trapping mediated from the Dom3z orientation occurs as predicted, however, two clones did not show fusion product amplification. Figure 3-10c summarises the trapping structures for both mutagenic units. In conclusion from this analysis, the mutagenic PB transposon should yield the gene-trap mutations 80 % of the total integration events where PB transposon land in genes.

Figure 3-10: Characterisation of the trapping events mediated by the mutagenic units.



a, RT-PCR results from ten clones studied. Genes trapped with the Dom3z mutagenic unit (left): 1, *FstI4*, 2, *A24Rik*, 3, *Sema6a*, 4, *stk22s1*, 5, *Ran*; Genes trapped for Ccc107 mutagenic unit (right): 1, *Tmem131*, 2, *Dut*, 3, *Sfrs3*, 4, *Undcd3*, 5, *Tmem50a*. b, Sequence trace from no.3 of Ccdc107 mediated trapping, when PB transposon inserted within intron 2 of *Sfrs3* on chromosome 17. One cryptic splice donor and two cryptic splice acceptors were observed within the PB5'ITR. c, Schematic summary of the trapping characteristics of the mutagenic PB transposon. The colour schemes in b and c are coordinated so that they correspond to the origins of the DNA/RNA sequences.

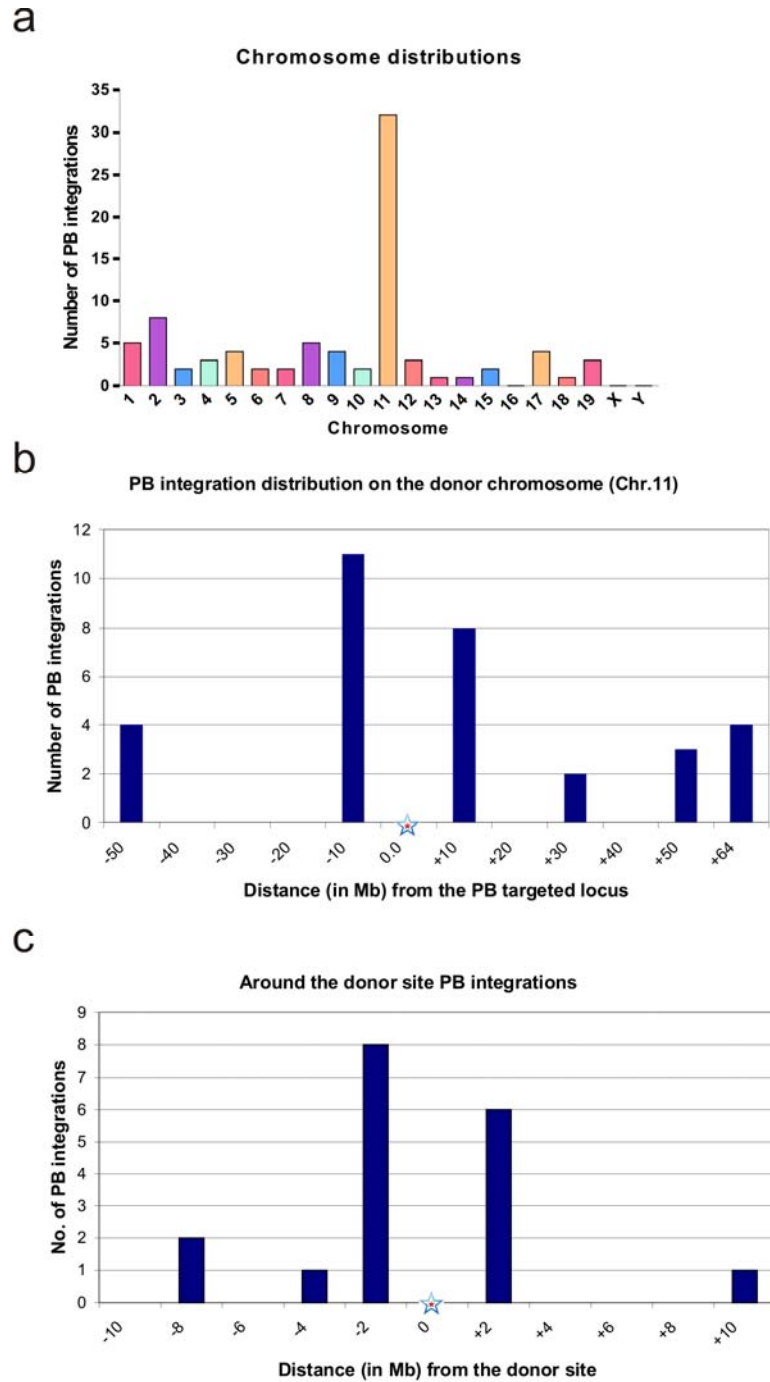
2.4. Local hopping of PB around the *Gdf9* locus

From the 96 puromycin and HAT double resistant clones generated in the random trapping assay, 86 clones were mapped to unique locations in the mouse genome with high alignment quality. Genome integration sites were analysed in order to gain insight into the PB transposon re-mobilisation characteristics following excision from the *Gdf9* locus.

Surprisingly, 32 out of 84 insertions were found on chromosome 11 (38 %), Figure 3-11a. In the vicinity of the donor site, 14 out of 32 reintegrations on chromosome 11 (44%) were within a 2.4 Mb of the donor site, Figure 3-11b,c. Although the dataset analysed here is small, the local hopping bias towards the donor chromosome and the donor site is clear. The number of centromeric and telomeric integrations also seems to be higher than other regions on chromosome 11 apart from the donor site.

Since the re-mobilisation selection strategy in this experiment is independent from the gene trapping status, the gene preference of PB integration can be analysed in an unbiased fashion. Of the 84 integrations, 42 landed in genes (50 %), 15 of which are on chromosome 11, and six were within the 2.4 Mb regions surrounding the donor site. It is known that PB-mediated integration has a bias towards genes compared with a simulated prediction for random integrations (Liang et al., 2009). For my small set of 36 insertions in genes which are not at the donor region, 30 insertions were within genes which have been trapped previously in mouse ES cells using a selection-based trapping strategy (a total of 165,778 trapping events in the database), which indicates that these genes are actively transcribed in mouse ES cells (<http://www.sanger.ac.uk/PostGenomics/genetrapp/>). Only two out of six integrations within the 2.4 Mb region have not been previously trapped. The PB transposon donor site is within a gene-dense region that is actively transcribed. The characteristics of this donor region may bias for local PB integrations within this region as the PB transposon integration is favoured to chromosomal regions that are actively transcribed, possibility due to the easy access to open chromatin structures.

Figure 3-11: Local hopping observed with PB transposon mobilised from the *Gdf9* locus.



a, Genome-wide distribution of PB re-integration, with a total of 84 events analysed. b,c, Local clustering of re-integrations surrounding the donor site. The star represents the donor locus where PB was inserted by gene targeting. “0” is the targeted locus of PB transposon. “-” is 5’ away from the donor site and “+” is 3’ downstream of the donor site.

2.5. Proof-of-principle of the mutagenic strategy in a DNA mismatch repair screen

In order to experimentally validate that the mutagenic strategy using my PB transposon could be coupled with the *Blm*-deficient ES cell system for recessive genetic screens, a proof-of-principle screen was conducted using the NN5-Gdf9^{hprtminiPB/+} cell line using the *Gdf9* locus as the donor site, to identifying components of the DNA mismatch repair pathway with 6-TG selection as the phenotypic readout.

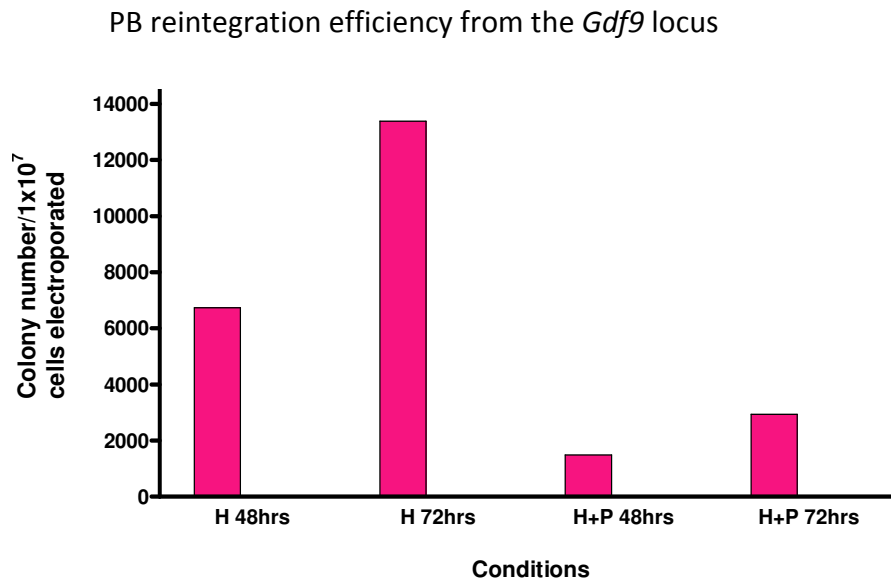
2.5.1. PB re-integration efficiency estimation in NN5- Gdf9^{hprtminiPB/+} cells

In order to provide an estimate of the efficiency of mutant generation by PB re-integration using NN5- Gdf9^{hprtminiPB/+} cells, the number of clones with PB transposon excised from the donor *Gdf9* locus and reintegrated in the genome could be measured. The mutagenic transposon was mobilised by electroporating 1×10^7 NN5- Gdf9^{hprtminiPB/+} cells with 25 μ g of mPBase expression plasmid (mPB Δ Neo). The electroporated cells were plated in ten 90 mm plates equally and the cells were selected with either HAT containing medium alone or HAT with puromycin. The selection was initiated either 48 or 72 hours post-electroporation, allowing sufficient time for PB re-integration to occur, Figure 3-12. A negative control experiment was carried out in parallel, without the electroporation of PBase expression plasmid.

In the control electroporation without the PBase, no colony was formed under any selection condition. The numbers of HAT or HAT and Puro double resistant colonies were two fold higher when selection was initiated 72 hours rather than 48 hours post-electroporation, but Hprt-deficient ES cells (i.e. cells without PB excision from the donor site) can be cross-rescued by Hprt-proficient cells (cells with PB excised from the donor site), giving rise to mixed colonies. The number of HAT and Puro double-resistant colonies was a quarter of HAT resistant colonies. The reintegration efficiency estimated here was lower than previously estimated without the positive selection for re-integration (Liang et al., 2009). The discrepancy between the two measurements is due to the timing of the puromycin selection in this experiment relative to the PB reintegration kinetics. The previous measurement was based on HAT-based excision selection (Liang et al., 2009); therefore, delayed PB

reintegration events may represent a small proportion of cells within a colony and could be detected by the Splinkerette PCR method. However, such kind of colonies can not survive under direct puromycin selection. A 48-hour post-electroporation selection scheme provided 1,500 HAT and puro double-resistant colonies per 1×10^7 cells electroporated. This condition provides good mutant complexity per pool for the mutant library construction.

Figure 3-12: PB re-integration efficiency excising from the *Gdf9* locus.



H, HAT; P, puromycin.

2.5.2. Library construction and screening

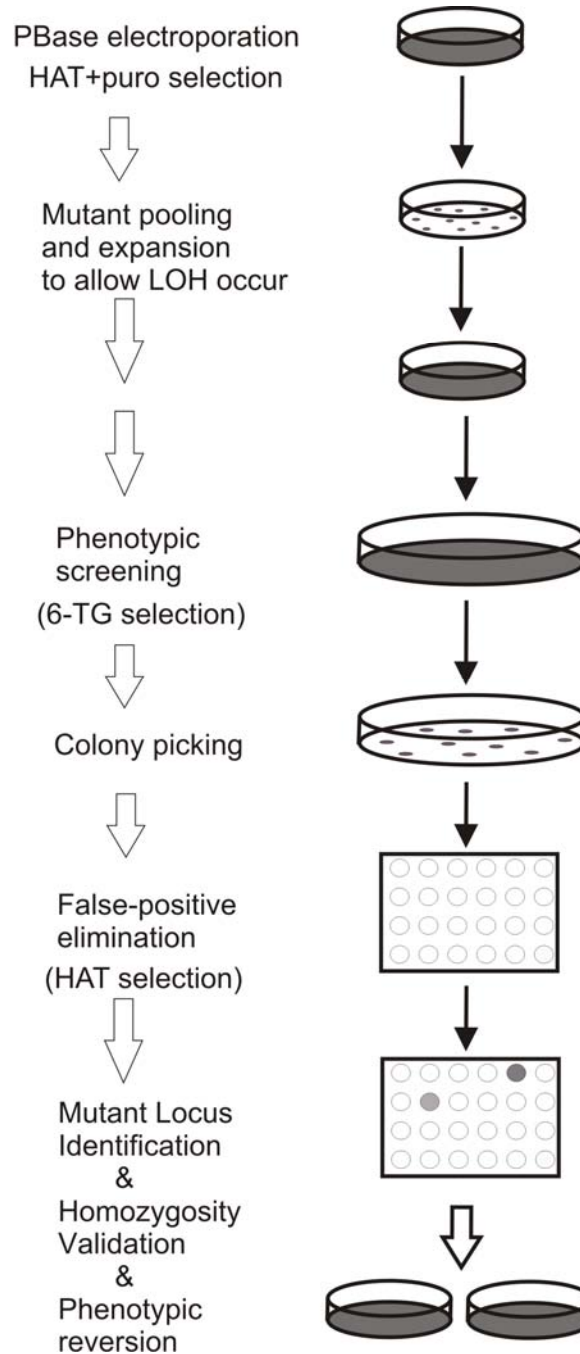
Ten electroporations were conducted with the electroporation condition described above, and the cells resulting from each electroporation was plated in a 90 mm plate were selected in HAT and puromycin containing medium until colonies were formed, which yielded approximately 1,300 heterozygous mutants per 90 mm plate. The colonies were pooled in each 90 mm plate and passaged into a fresh 90 mm plate until the plate was confluent (3×10^7 cells/90 mm plate). The cells were then passaged 1:1 from a 90mm plate to a 150 mm plate, and 2 μ M 6-TG selection was commenced 24 hours after plating to select for DNA MMR mutants. The experimental scheme is shown in Figure 3-13. MMR mutant isolation using 6-

TG selection can be conducted at high cell density, as 6-TG-induced cytotoxicity is only effective in cells with proficient MMR mechanism. MMR mutant cells are insensitive to the accumulation of mismatch repairs in the genome, therefore they survive even surrounding cells are continuously supplying them with genome-toxic metabolites derived from 6-TG.

After eight days of 6-TG selection, around 100 - 200 colonies were formed on each 150 mm plate. The number of 6-TG resistant colonies generated was too high to be real, as it is unlikely that 1,000 to 2,000 genes are involved in the MMR system. In addition, previous screens using libraries with a similar heterozygous mutant complexity to this experiment yielded less than twenty clones were produced (Guo, 2004; Wang et al., 2008a). There may be two possibilities for the generation of such large amount of 6-TG resistant clones. Firstly, LOH events within MMR relevant genes occurred very early on during the expansion, thus the homozygous daughter cells expanded several generations before 6-TG selection was initiated. Secondly, false positive clones may present due to loss of the *HPRT* minigene, as *Hprt*-deficient ES cells are resistant to 6-TG selection. The method to distinguish the two possibilities is to select these clones individually with HAT. If the clones were generated due to loss of the *HPRT* minigene, they should be sensitive to HAT selection. As the *HPRT* minigene is on one of the homologous chromosomes of an autosomal locus, the rate of losing the *HPRT* minigene in *Blm*-deficient ES cells is high, approximately one per 2,000 cells post PB transposon excision. Cells deficient in *Hprt* are 6-TG resistant irrespective to the MMR status, as 6-TG can not be converted and incorporated into DNA synthesis, thus contributing to the majority of the 6-TG resistant colonies formed.

600 Colonies were picked in total from all ten 150 mm plates. The cells in 96-well plates were selected under HAT for four days. Seventeen HAT resistant colonies were obtained from these 600 colonies. Therefore, the large proportion of false positive clones was highly likely due to loss of the *HPRT* minigene on the *Gdf9* locus by LOH.

Figure 3-13: Schematic representation of the experimental procedures for isolating homozygote MMR mutants.



The seventeen 6-TG and HAT double resistant clones were expanded and the PB integration sites were identified using the Splinkerette PCR. The results are summarised in Table 3-3. Seven independent insertions were found within the seventeen clones, and four were mapped within 600 kb region in four different genes surrounding the PB donor locus. These are likely to reflect the local hopping effect observed previously. In these clones, the PB integration sites are unlikely to relate the 6-TG resistant phenotype. One insertion was mapped to an intergenic region. One PB integration site was mapped to a gene *Rrp9* (ribosomal RNA processing 9), which is a component of a nucleolar small nuclear ribonucleoprotein particle, snoRNP, thought to participate in the processing and modification of the pre-ribosomal RNA (UniProtKB/Swiss-Prot). *Rrp9* therefore is unlikely to be a candidate of the MMR system. It is likely that spontaneous mutations in other MMR components have occurred in these clones, as *Blm* deficiency promotes mutagenesis and the conversion to homozygosity. The MMR screen using 6-TG promotes the accumulation of mutations in the genome; therefore background spontaneous mutations in this screen can be elevated compared to other screens. A final PB integration sites for three daughter clones was in intron 1 of *Msh6*, a known gene essential for the DNA mismatch recognition. The insertion sites are summarised in Table 3-3.

Table 3-3: Details of seven independent integration sites identified from the MMR screen.

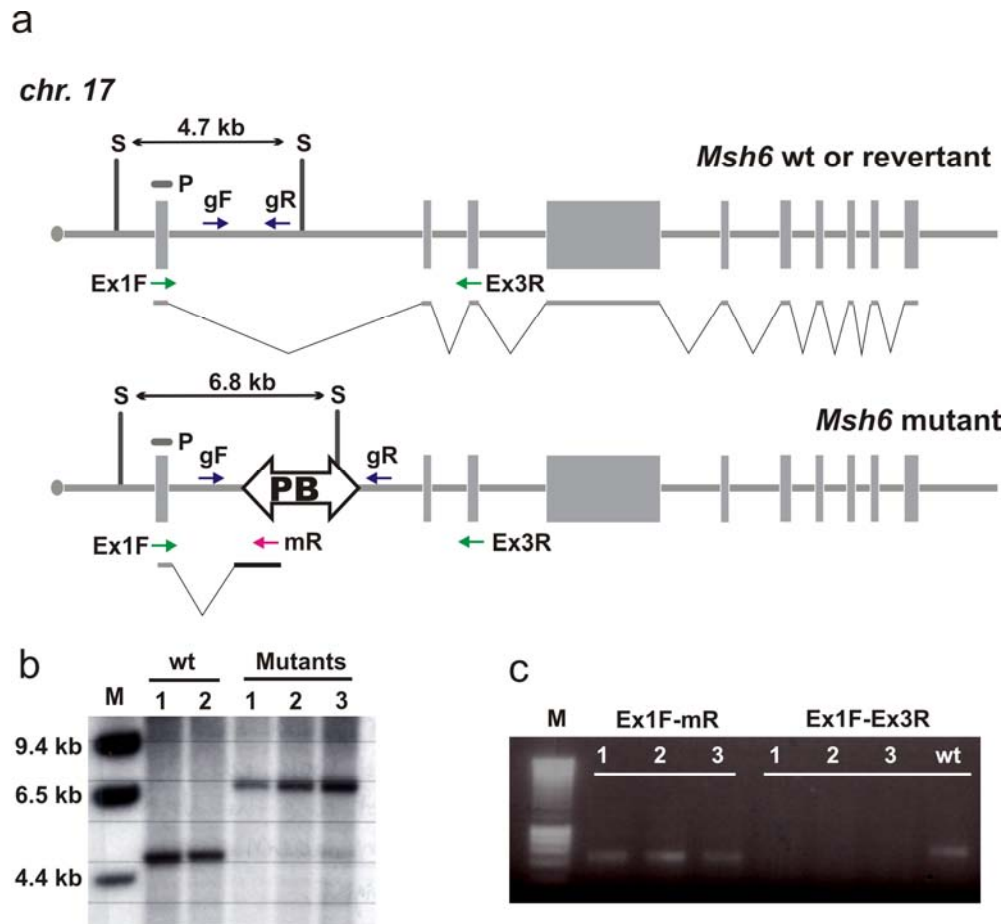
Pool	Chr.	In gene	Samples	Start	End	PBITR	Strand	Length
1	11	Aff4	1	53166717	53166863	5	C	147
3	11	Fst14	3	52847204	52847285	5	F	82
5	11	Il13	4	53447234	53447397	5	F	164
7	17	Msh6	5,6,7	88376055	88376181	5	F	127
6	9	Rrp9	8,10,14,16,17	106380210	106380327	3	C	118
6	11	Sept8	9,11,12,13	53355879	53356054	5	F	176
6	5	No	15	127578422	127578557	3	C	136

Note that the integrations on chromosome 11 are within 600 kb from the donor site. NCBI m36 was used for the integration-site mapping.

2.5.3. Mutant analysis and validation

This *Msh6* mutant was analysed for its homozygosity status. An external DNA probe was used so that both wild type and mutant alleles could be detected by Southern analysis. This clone was confirmed to be homozygous by Southern blotting, Figure 3-14a,b. Further analysis of the transcript by RT-PCR confirmed that the mutant was null for *Msh6*, Figure 3-14c.

Figure 3-14: *Msh6* mutant validation.



a, Schematic representation of the wild type (revertant) and mutant allele, with Southern blotting detection strategy (b) and the primers used for genomic PCR and RT-PCR (c). S, *SpeI* recognition site; P, Southern blotting probe. b, Southern blotting to confirm the homozygosity status of the mutant. c, RT-PCR analysis. 1, 2, 3, are three sister clones from the mutant. The primers used for the PCR reaction are indicated above the samples. wt, wild type control; M, marker. mR, reverse primer from the terminal exon of *Dom3z*.

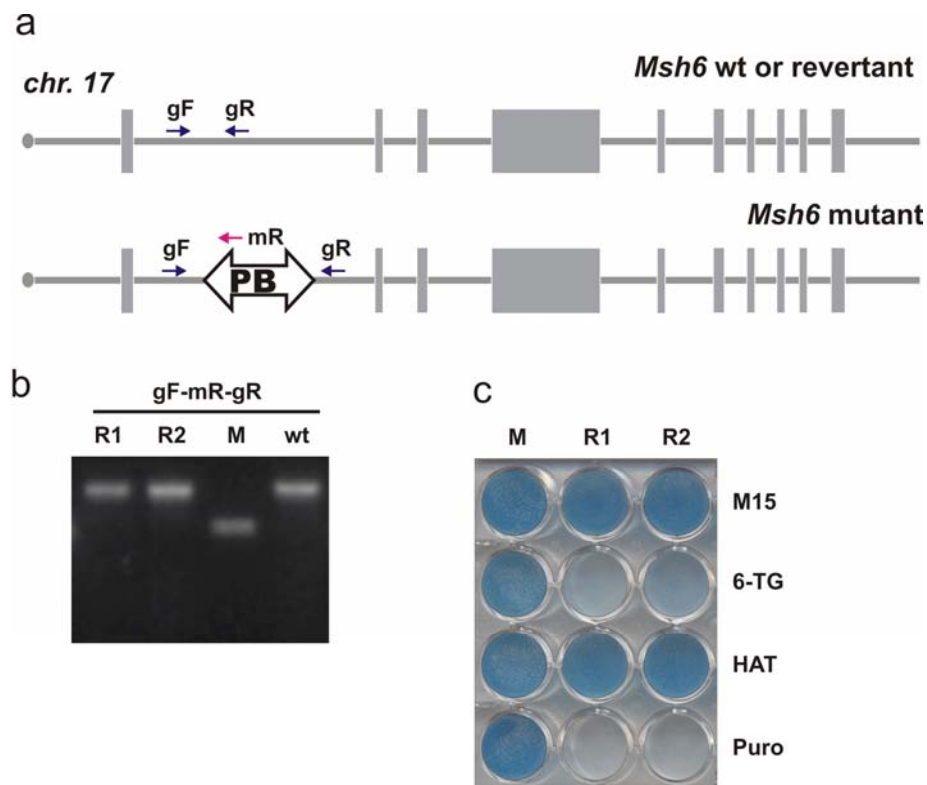
Msh6 is a known gene for the MMR pathway and homozygous mutant of this locus can be used to functionally validate the use of PBase to revert the genotype of this mutant, further confirming the genotype-phenotype causality. Removal of the mutagenic transposon from the intron of the *Msh6* gene allows the transcription of the wild type mRNA, thus the revertant cells become sensitive to 6-TG selection. A single copy of PB transposon removal allows the wild-type transcription to resume at one allele, however, in certain cases, heterozygous mutants may show a haploid insufficient phenotype, making the causality difficult to be confirmed. Therefore, complete geno-type reversion is ideal. The efficiency of the dual-transposon removal could be very low. Therefore, by coupling the PB transposon removal with a negative selection scheme, the revertant clones can be easily isolated by direct selection of cells without any PB transposons (containing *puroΔtk* expression cassette) residing in the genome.

The 3×10^6 *Msh6* mutant ES cells were electroporated with 20 μ g of mPBase expression plasmid (mPBase Δ Neo) and the cells were plated into a 90 mm plate and grown without any selection for three days to allow the transcripts for *puroΔtk* to decay prior to conducting the negative selection. To select for complete genotype revertants, 1×10^5 cells were plated in a 90 mm plate and FIAU selection was initiated the following day. As a background control, pBlueScript plasmid was electroporated instead of PBase expression plasmid. Colonies from the mPBase-electroporated cells were picked for further genotype validation and phenotypic profiling.

In total, 148 colonies were formed in the cells electroporated with PBase whereas 30 colonies were present in the control plate. This slight background observed in the control was likely due to the presence of a small fraction of ES cells without the PB transposon mixed within the mutant clone or spontaneous mutations generated within the *tk* coding region of the *puroΔk* cassette. Triple-primer PCR was conducted using locus-specific genomic primers up- and downstream of the integration site, together with transposon-specific primer from Dom3z sequence, Figure 3-15a,b. FIAU-resistant revertant cells had lost both copies of the transposons (from both homologous chromosomes) as predicted. A panel of drug selections

was also conducted to confirm the phenotype of the revertants, Figure 3-15c. Genotype-reverted cells also showed a wild-type phenotype, i.e. sensitive to 6-TG selection. The loss of both copies of the PB transposon was also reflected by the loss of puromycin resistance. Both mutant and revertants were HAT resistant, suggesting that the *HPRT* minigene was not lost in all cases.

Figure 3-15: *Msh6* mutant rescue analysis.



a, Schematic representation of the wild type/revertant and the mutant allele. b, triple-primer competition PCR to genotype the revertants. c, Phenotypic profile of the mutant and the revertant cells. R1 and R2, are two independent revertant clones; M, mutant, wt, wild type.

3. Discussion

This chapter has described the establishment and experimental validations of a new mutagen and a strategy to deploy intra-genomic re-mobilisation of a *piggyBac* transposon to generate genome-wide heterozygous mutants. This strategy incorporates the aims of unbiased genome-wide coverage, efficient mutagenesis, easy identification of the mutation and reversion to establish genotype-phenotype causality.

In this strategy, I have generated a novel mutagenic PB transposon and inserted it into the ES cell genome by gene targeting, providing a stable single copy per cell transposon. Upon the supply of PB transposase, cells with PB excised from the donor locus and re-integrated elsewhere can be enriched. This enrichment for the re-integration events is dependent on a dual positive-selection strategy for the reactivation of *Hprt* transcription upon PB excision from the donor site and a positive selection marker within the PB transposon. Because of the high transposition efficiency of PB, in theory, intra-genomic mobilisation is sufficient in this design to provide enough heterozygous mutants to cover the whole ES cell-expressed genome in merely twenty 90 mm culture plates.

3.1. Molecular design of the gene-inactivating PB transposon

The molecular design within the PB transposon is aimed at maximal gene inactivation, with a non-selective trapping in both orientations mediated by terminal-exon pairs selected from the mouse genome. It should be mutagenic in either orientation in intronic and exonic positions in all genes, irrespective of their protein-coding potential, gene expression levels or reading frames. The use of endogenous exons as efficient gene traps has genetic basis, and it was explored and experimentally validated in this chapter. Most of the widely used gene-trap cassettes contain a splice acceptor, which has been directly cloned from naturally occurring viral or mammalian sequences. For example, *βgeo* relies on the adenoviral splice acceptor from the viral major late transcript to disrupt gene expression (Friedrich and Soriano, 1991), and the mammalian *Engrailed-2* splice acceptor is also frequently used (Collier et al., 2005). The popularity of these splice acceptors is historical and other endogenous exons within the mouse genome are likely to contain suitable splice acceptors for the purpose of efficient gene

trapping in many genomic contexts. Therefore, we computationally scanned the mouse genome to identify good mutagenic terminal-exon structures with criteria that allow the selection of “strong” mutagenic units other than the conventional splice acceptors.

The final two candidates were experimentally validated for their mutagenicity in two complementary assays, an *Hprt* trapping assay and a random trapping assay. The *Hprt* trapping assay is a stringent measurement for the strength of the trapping, as only cells with a near or complete null mutation of *Hprt* can give rise to 6-TG resistance. In these experiments, several colonies were isolated with a 6-TG resistant phenotype and the phenotype was revertible upon PBase re-introduction. In many cases, the exact insertion sites could not be mapped due to the technical challenges of mapping PB integration sites when there are many in each cell. However, several independent clones with PB transposons were mapped in independent integration sites within the *Hprt* were eventually detected, and the *Hprt* inactivation in all cases was mediated by the mutagenic units. Considered together, the data suggest that the mutagenic units efficiently cause gene inactivation.

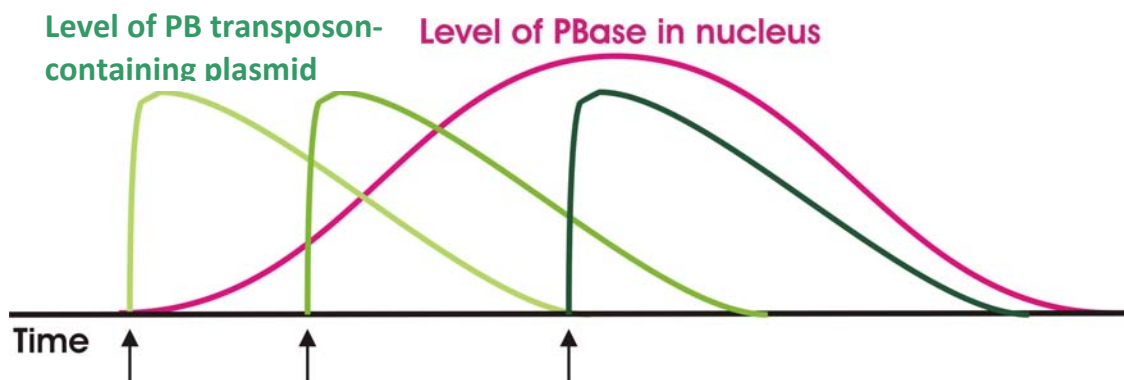
One limitation of this assay is that it does not test mutagen trapping capability in different genome contexts. Therefore, I conducted a random trapping assay to address this issue, using a cell line in which the mutagenic PB transposon was inserted into an HPRT minigene which was itself targeted into the *Gdf9* locus. PB was excised from the donor locus *Gdf9* following PBase expression, and reintegrated in a genome-wide fashion. A set of ES cell clones with PB re-integrated within genes expressed in ES cells were selected and assessed for the efficiency of the two mutagenic units within the PB transposon. Overall based on the data from two independent trapping assays, the mutagenic transposon are predicted to be able to mediate trapping 80 % of all intragenic integrations.

3.2. *piggyBac* possesses a fast transposition kinetics

A PBase inducible cell line, AB1-ROSA26^{mPBaseERT2/+}, has been established with the advantage of a tight temporal control over the PBase activity in the nucleus by 4-OHT addition. A pre- and post- PB transposon introduction time course was analysed with the aim to obtain high

efficiency of “plasmid-to-genome” transposition using the inducible PBase system. A short period of PBase induction (2-4 hours) prior to the electroporation of the PB transposon-containing plasmid dramatically elevated the transposition efficiency by five fold compared to without the pre-electroporation treatment. This suggests that PBase-catalysed transposition reaction occurs very fast when PBase and PB transposon first encounter and the longer the overlapping period between PBase and the PB transposon, the higher efficiency of obtaining transposon integrations, Figure 3-16. The post-electroporation incubation with 4-OHT was less overt. However, sustained 4-OHT incubation post-electroporation consistently decreased the overall number of G418 resistant colonies. This may reflect the continuous intra-chromosomal transpositions occurring in each cell after the initial PB integration. The rate of transposon loss during intra-chromosomal transposition is approximately 60 %, measured from the single PB transposon remobilisation from a genomic donor locus (Liang et al., 2009). This means that with the sustained supply of PBase, the copy number of the integrated transposons are gradually lost overtime *via* repeated cycles of intra-chromosomal mobilisation.

Figure 3-16: A model to account for the PB transposition kinetics.



The green curves represent the level of PB transposon-containing plasmid in the nucleus with an initial maximal level at the point of electroporation and decay overtime due to cell division and DNA degradation. The pink curve represents the level of PBase present in the nucleus induced by 4-OHT, with the initial accumulation upon 4-OHT addition and decay overtime due to protein turnover after 4-OHT withdrawal. The arrows below the timeline represent the different time points when the PB transposon containing plasmids were introduced by electroporation.

3.3. Local hopping characteristics of *piggyBac*

The PB intra-genomic mobilisation was also investigated using a single copy PB mobilised from the *Gdf9* locus on chromosome 11. Clear local hopping was observed on the donor chromosome and predominantly surrounding the donor site. The possible local hopping effect of PB intra-chromosomal transposition has been noted previously for intra-genomic mobilisation from the *Rosa26* locus on Chromosome 6 (Wang et al., 2008b) but not observed on the X-linked *Hprt* locus (Wang et al., 2008b; Liang et al., 2009), with assay set-ups similar to the one described here, i.e. without any trapping-based integration selection. The results from these different loci are directly compared and summarised in Table 3-4.

Table 3-4: Local-hopping comparison among different genomic loci.

	<i>Gdf9</i> locus		<i>Rosa26</i> locus		<i>Hprt</i> locus ^a		<i>Hprt</i> locus ^b	
	Insertions	% Total	Insertions	% Total	Insertions	% Total	Insertions	% Total
Total	84	100 %	264	84 %	93	100 %	79	100 %
PB Donor Chr.	32	38 %	47	32 %	8	9 %	5	6 %
PB Donor site*	14	17 %	25	14 %	3	3 %	0	0 %

*, 2.4 Mb region surrounding the donor site was analysed; for *Hprt* locus, 2.4 Mb window was set outside the 33.6 kb *Hprt* gene, as the integration site was not mentioned in the paper. Datasets for *Rosa26* locus and *Hprt* locus were obtained from supplementary materials of Wang and co-workers (Wang et al., 2008b). a, data from (Wang et al., 2008b) b, data from (Liang et al., 2009), and the integration events in this data were combined with two versions of the PBase, i.e. the wild-type insect PBase and the mPBase.

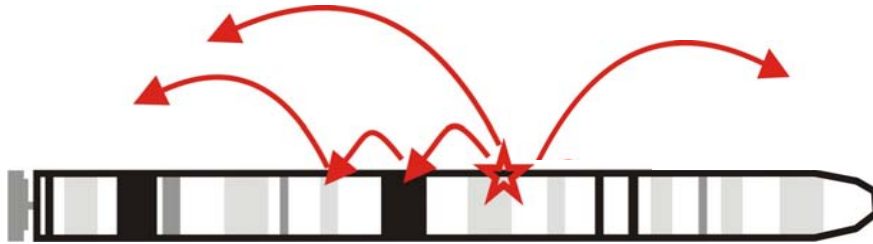
Out of the three PB transposon donor loci, the X-linked *Hprt* locus does not show a donor chromosome local hopping pattern, although there may be a slight tendency towards local re-integrations around the donor locus observed in one of the *Hprt* excision data sets (Wang et al., 2008b). Both the *Gdf9* locus and *Rosa26* locus show a local hopping effect, however at the *Gdf9* locus, this appears to be more extensive.

There are several possibilities contributing to the local hopping of PB transposon. Firstly, the experimental condition is slightly different with different versions and amounts of PBase introduced. The amount of nuclear PBase protein level can affect the pattern of transposition. The more PBase-containing plasmid transfected or the more stable PBase protein is (i.e.

mammalian codon-optimised PBase for stable expression v.s. wild-type PBase with insect-derived sequence), the longer the PBase protein can exist to catalyse the transposition events continuously. Secondly, PB mobilisation has been described to be sensitive to methylation. Methylation reduces the transposition efficiency of PB (Wang et al., 2008b). Therefore, if there is extensive methylation in an area around the donor locus, the PB transposon may “get stuck” within this region due to inefficient transposition in methylated genomic sequences. Additionally, if the transposon itself is methylated after targeted insertion into the genome, its re-mobilization efficiency can be significantly reduced. Thirdly, the chromatin structure may affect re-mobilisation pattern and efficiency. PB has a strong preference for integrating in actively transcribed genes (Liang et al., 2009), suggesting that PB has a preference of integrating into open chromatin regions. If the donor site is surrounded by “inaccessible” closed chromatin, PB transposition may also be retarded to mobilise out of the donor region. Chromosome 11 has a very high gene density and the region surrounding the *Gdf9* locus is also rich in actively transcribed genes in ES cells. The donor site/chromosome integration preference may be partially due to the open chromatin bias. A final possibility is that *Blm*-deficiency may influence PB transposition. The ES cell line used for the PB transposon re-mobilisation from the *Gdf9* locus was derived from a *Blm*-deficient ES cell line (Guo, 2004), whereas other remobilisation work was based on a *Blm*-proficient ES cell background. The influence of *Blm* deficiency and PB transposition is not yet known. However, experimental evidence shown in Chapter 4 seems to suggest that *Blm*-deficiency negatively affect the PB-mediated transposition, as identical PB transposon targeted into the same position within the *Hprt* locus in *Blm*-proficient or deficient ES cells showed differential re-mobilisation efficiency (Chapter 4, Figure 4-4). It is possible that *Blm*-deficient ES cells may not be able to repair the DNA double strand breaks (DSBs) generated upon PB excision. Continuous cycles of PB excisions and re-integrations within the same cell cause cell death if these genomic damages can not be fixed efficiently and precisely. Genome-wide remobilisation may be achieved through continuous transpositions, thus cells with less cycles of continuous transposition may be selected for in this scenario. The precise mechanism is yet to be discovered.

For the purpose of genome-wide mutagenesis, the use of the *Hprt* locus is a better choice to maximise genome-wide coverage. However, the observation of local hopping is an interesting phenomenon to provide us with an insight into the kinetics of intra-genomic transposition of *piggyBac*, Figure 3-17. Using a temporal controllable PBase expression (PBaseERT2), this kinetic aspect of the *piggyBac* transposition can be dissected. Such investigations not only can provide insight into the fundamental characteristics of PB transposition, but may also highlight potential risks in using PB in genetic studies and clinical medicine for its possible “non-tagged” mutagenesis of the genome.

Figure 3-17: Possible intra-genomic mobilisation kinetics of the PB transposon.



Genome-wide reintegration of the PB transposon can be achieved in two ways. One possibility is that PB can directly re-integrate into a locus randomly after being excised from the donor site. Another possible mechanism is through several rounds of local hopping to “escape” the donor site. This local-hopping escape model can be extended to the reintegration into a proximal chromosome, which is spatially local to the donor site.

3.4. A recessive genetic screen using the established mutagenesis strategy

A DNA mismatch repair screen was conducted using a library generated with the established mutagenesis strategy and a known gene was identified from this screen. The inability to uncover other MMR genes is mainly due to the bias re-integrations of the PB transposon surrounding the PB transposon donor site. The local hopping of PB transposon surrounding the 2.4-Mb region of the donor site was measured to be 17 % of the total number of re-integrations. This means that in this mutant library with 10,000 mutants, only 8,300 clones contain integrations away from the donor site. Thus, the number of heterozygous mutations was limited in this library. However, a known MMR gene, *Msh6*, was isolated from the screen

suggesting that the established mutagenesis strategy was sufficient to couple with the *Blm*-deficient background for recessive genetic screens. It has been previously observed that intra-chromosomal mobilisation from the *Hprt* locus did not give rise to local hopping, Table 3-4. Therefore, using the *Hprt* locus as the donor site will be better choice to mediate genome-wide mutagenesis.

In this proof-of-concept MMR screen, other HAT and 6-TG double-resistant clones were identified, which were not obvious candidates for MMR pathway and most of these clones contain PB integration sites mapped to regions locally surrounding the *Gdf9* locus. This suggests that these clones may possess background mutations in the MMR pathway genes, and the PB integrations are not the casual mutations. These PB integrations are unlikely to be causal to the 6-TG resistant phenotype and background mutations may have occurred in these clones. *Blm* deficiency promotes conversion of heterozygous to homozygosity, thus spontaneous mutations generation during ES-cell culturing may be converted to homozygosity with 20-fold enhancement in *Blm*-deficient background compared to wild-type cells. In addition, the MMR screen using 6-TG selection may enhance the random background mutation rate, as 6-TG is genotoxic pro-drug and its metabolites can be incorporated into the genomic DNA of cells to generate point mutations. Therefore, it is important to conduct genetic rescue experiment to validate the causal link between the PB integration sites and the phenotype. If the integration is irrelevant to the phenotype, excision of the PB transposon will not revert the mutant phenotype to wild type.

3.5. Complete genotype reversion using PBase with FIAU selection

For the purpose of genetic screens, it is important to be able to establish a causal relationship between the gene mutated and the phenotype. The best way to establish this connection is through genetic rescue experiments. Upon mutagen removal or complementation of a wild-type copy of the mutated gene, the mutant phenotype should also revert to wild type.

Mutagenesis mediated with DNA transposons has the significant advantage of their simple removal. PB transposition has the unique property of excision without footprint. Thus,

excision of the mutagenic PB transposon from a gene can completely revert the mutant phenotype to wild type. Even insertions in exons can be fully reverted. Because the reversion efficiency per transposon is approximately 1 %, identifying complete reversion events of homozygous mutants is not efficient enough without any selection strategy, especially in “difficult-to-excise” genomic contexts. In my transposon design, a *puroΔtk* cassette was introduced to facilitate the selection for revertants using FIAU. This strategy was demonstrated in this chapter using a homozygous *Msh6* mutant obtained from a DNA MMR genetic screen. Many genotype-reverted colonies could be readily obtained and their drug-selection profile proved that these were phenotypic reversions.

The selective excision of PB transposon is also widely useful for other purposes requiring PB transposon removal without reintegration elsewhere in the genome. For example, integration-free induced pluripotent stem (iPS) cells can be generated with PB transposon carrying the Yamanaka factors to reprogram somatic cell types. The transgenes can be subsequently removed from the genome by PBase supply and FIAU selection (Yusa et al., 2009).

Chapter Four - Generation of a *Blm*-deficient mouse ES cell line with a single copy of the mutagenic *piggyBac* transposon

1. Introduction

A *Blm*-deficient ES cell line with an intact *Hprt* locus is a very useful reagent. It can be used to conduct genome-wide recessive screens when coupled with *piggyBac*-mediated intra-genomic mobilisation using *Hprt* as a donor locus to facilitate the enrichment for PB remobilisation (Chapter 3). However, the existing *Blm*-knockout ES cell line was generated in the *Hprt*-deficient AB2.2 cell line (Chapter 2). In addition, a reporter system is required for my screens in order to select for mutants in the miRNA biogenesis and effector pathways (Chapter 1 and 5). Therefore, a new *Blm*-deficient ES cell line has been designed and generated to incorporate both requirements.

2. Results

2.1. Generation of a new *Blm*-deficient mouse ES cell line

The gene-targeting strategy used to create the *Blm*-null mutation was based on the design of the *Blm*^{tm1Brd} allele of the *Blm*-knockout ES cells, using a positive drug selection cassette to replace the start codon-containing exon by the replacement gene-targeting method (Luo et al., 2000). An eGFP and Bsd-resistant gene co-expression cassette driven by the human ubiquitin C promoter (Huc-eGFP-IRES-Bsd) was introduced into the *Blm* locus as the selection marker to replace exon 2 of the *Blm* gene Figure 4-1b. The eGFP-IRES-Bsd is constitutively expressed under the Huc promoter and the coding regions of eGFP and Bsd are connected by IRES, which allows the two coding regions to be transcribed from a single mRNA. Both alleles of *Blm* were targeted sequentially to give rise to the *Blm*-null mouse ES cell line.

The pBlmMLTV2 targeting vector (Figure 4-1a) was constructed based on a C57BL/6 BAC covering the *Blm* gene, and the detailed vector construction is shown in Chapter 2. The targeting vector was linearised with *PmeI* and electroporated into 1x10⁷ JM8.F6 cells. The cells were selected under Blasticidin for seven days and 48 colonies were picked and screened

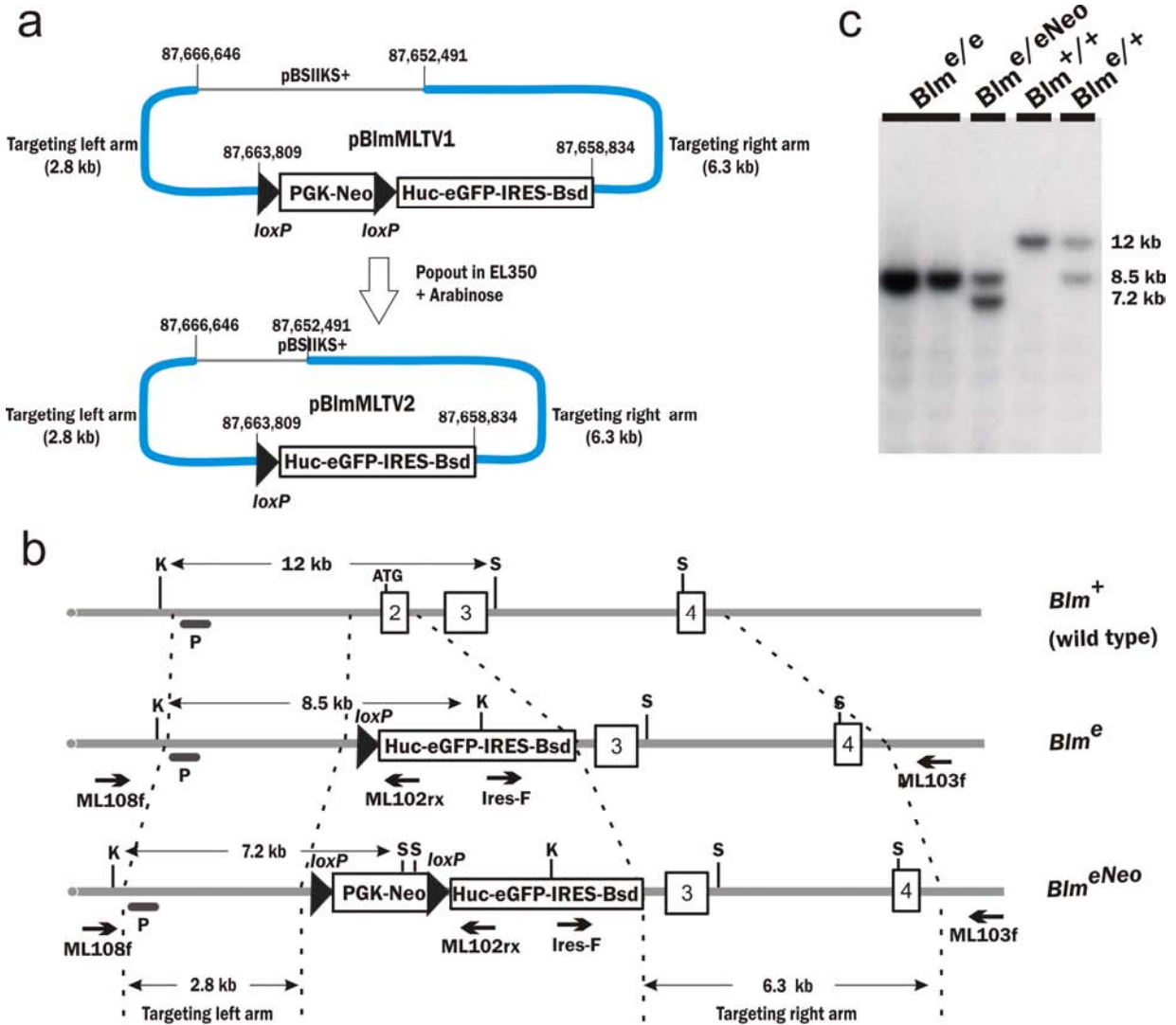
with long-range PCR (with primers ML108f and ML102rx) to detect homologous recombination on the short targeting arm end for correctly targeted clones. The targeting efficiency was 18 %. Long-range PCR (with primers Ires-F and ML103f) was also conducted to confirm the homologous recombination on the long targeting arm end. The correctly targeted cells are named *Blm^{e/+}*.

The six of the correctly targeted clones were subjected to karyotyping analysis to ensure the ES cell clones were not aneuploid during the targeting and subcloning processes. The cells were incubated with Colcemid (100µg/ml final concentration) for two hours to accumulate cells with cell-cycle arrested at metaphase before the metaphase preparation (Chapter 2 for detailed protocols). For each ES cell clone, ten metaphases were examined for the total chromosome numbers. Wild-type cells with a diploid genome should contain 40 chromosomes at metaphase of the cell cycle, and the most common aneuploidy includes the loss of Y chromosome (the cells used are male) to give rise to 39 chromosomes, and trisomy for chromosome 8 or 11, to give rise to 41 chromosomes. All six *Blm^{e/+}* clones examined had 40 chromosomes per metaphase.

One of the clones with the correct karyotype was selected for the second round targeting with the *PmeI*-linearised targeting vector pBlmMLTV1. The electroporated cells were selected using G418 for eight days and 96 colonies were picked. The efficiency of the second allele targeting is expected to be slightly lower, as the targeting of the second allele is only 50 % of all the correctly targeted events. Therefore, a larger number of colonies were screened for the correctly targeted clones than the first allele targeting. Long-range PCR (with primers ML108f and ML102rx) was conducted to screen for colonies with both alleles targeted, as both alleles can be amplified with the primer set. Two PCR products with different sizes can be observed. The second allele targeting efficiency was 17 %. The double-allele targeted ES cell clones (*Blm^{e/eNeo}*) were expanded and were subjected to karyotyping analysis. Out of the six clones analysed, no aneuploidy was observed.

One correctly targeted and karyotypically normal clone was expanded and the PGK-Neo cassette was popped out by transient transfection of a Cre expression plasmid (pCAG-Cre). 3×10^6 cells were electroporated with 25 μ g of pCAG-Cre plasmid, and the cells were plated onto a 90 mm plate. Three days post-electroporation in non-selective medium, the cells were trypsinised and replated in duplicates at the density of 1,000 cells per 90 mm plate. The next day, G418 selection was initiated on one plate and no selection was added to the other plate. The number of G418 resistant colonies provided a background level for the clones without the cassette popped out. Twelve colonies were picked from the 90 mm plate without any selection and were subjected to PCR analysis to detect the presence of the genomic junction with the cassette removed. A third of the colonies picked were positive for the cassette popout. After removal of the PGK-Neo cassette, both alleles of the *Blm* locus are identical, giving rise to the final *Blm*-deficient ES cells, *Blm*^{e/e} (Figure 4-1b). Southern blotting was also conducted using an external locus-specific probe to further confirm the locus structure after each step of the manipulation, shown in Figure 4-1c.

Figure 4-1: Double-allele targeting of the *Blm* locus of the JM8.F6 ES cell line.



a, Two targeting vectors designs for the sequential *Blm* targeting. pBlmMLTV2 was derived from pBlmMLTV1 with the PGK-Neo cassette deleted in EL350 *E coli* strain in which Cre expression is induced by *L*-arabinose. The genomic coordinates (NCBI Build 37) shown on the targeting vector are the coordinate of the start and base pair taken for the targeting arm. b, Allele structures of the *Blm* locus. The wild-type allele was targeted with pBlmMLTV2 to give *Blm*^{e/+} cells. *Blm*^{e/+} cells were further targeted with pBlmMLTV1 to give *Blm*^{e/eNeo} cells. Finally, the PGK-Neo cassette was popped out to give the final *Blm*^{e/e} cells. The locations of the long-range PCR primers used for the initial genotype screening are shown. P, DNA probe for Southern blotting; K, *Kpn*I site; S, *Spe*I site. c, Southern blot confirming the genotypes of all the intermediate and final *Blm*-deficient (*Blm*^{e/e}) ES cells.

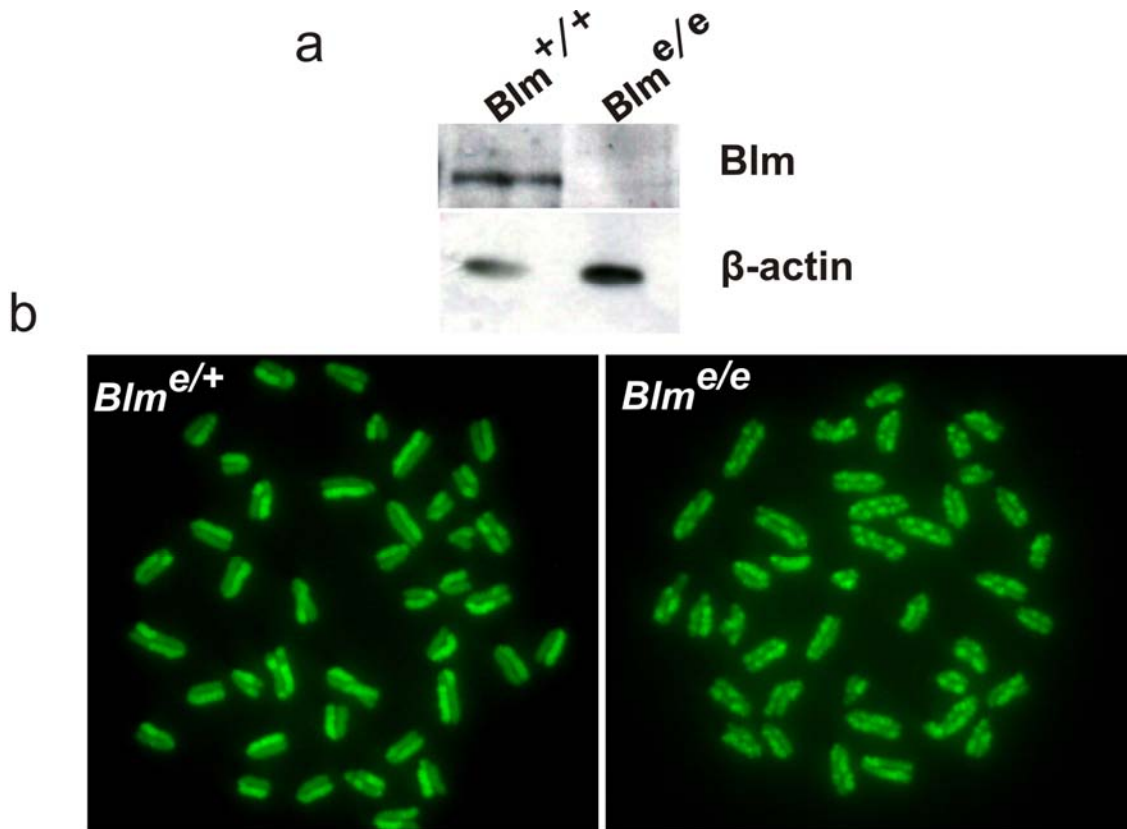
2.2. Phenotypic assessments of the *Blm*^{e/e} cell line

Although the targeted replacement of exon 2 should give rise to a null mutant of *Blm*, it is important to assess the absence of Blm protein and to confirm *Blm*-deficiency phenotypically. The Blm protein level of the *Blm*^{e/e} cell line was assessed by Western blotting, Figure 4-2a. No protein product was detected in the *Blm*^{e/e} cells, suggesting that the targeting completely abolished *Blm* expression.

One of the main characteristics of *Blm*-deficient mammalian cells is hyper-recombination between homologous chromosomes and between the sister chromosomes. Sister chromatid exchange, SCE, can be visualised on metaphase preparation of cells with the sister chromatids differentially labelled with BrdU. Both of the heterozygous and homozygous *Blm* mutants (*Blm*^{e/+} and *Blm*^{e/e}) were treated with BrdU in two consecutive cell cycles (approx. 34 hours). Metaphase spreads were prepared which were stained with Acridine orange for SCE visualisation. BrdU is a thymidine analogue, which can be incorporated into the newly synthesised DNA strand during S phase. After the first round of DNA replication in the presence of BrdU, the sister chromatids are equally labelled, with only the newly synthesized strand of the DNA containing BrdU. In the second round of DNA synthesis, the two sister chromatids are differentially labelled. One sister chromatid contains both strands of the DNA labelled with BrdU, while the other sister chromatid which inherited the original template strand (without BrdU labelling) has only one strand of the DNA labelled with BrdU. Acridine orange binds to dsDNA and emits green fluorescence upon binding. BrdU can quench the Acridine orange fluorescence, thus the sister chromatid with both DNA strands labelled with BrdU appears dimmer than the one with only single DNA strand labelled.

In *Blm*^{e/+} cells, most of the chromosomes do not have SCE, occasionally one exchange per chromosome can be observed. However, in *Blm*^{e/e} cells, SCE occurs in almost all chromosomes, and four or even more exchanges per chromosome are commonly observed. Figure 4-2b shows the typical images of metaphase spreads of *Blm*^{e/+} and *Blm*^{e/e} cells. Taken together, the newly generated ES cells with both *Blm* alleles targeted show the cellular characteristics of Blm deficiency.

Figure 4-2: Functional validation of the new *Blm*-deficient ES cell line *Blm^{e/e}*.



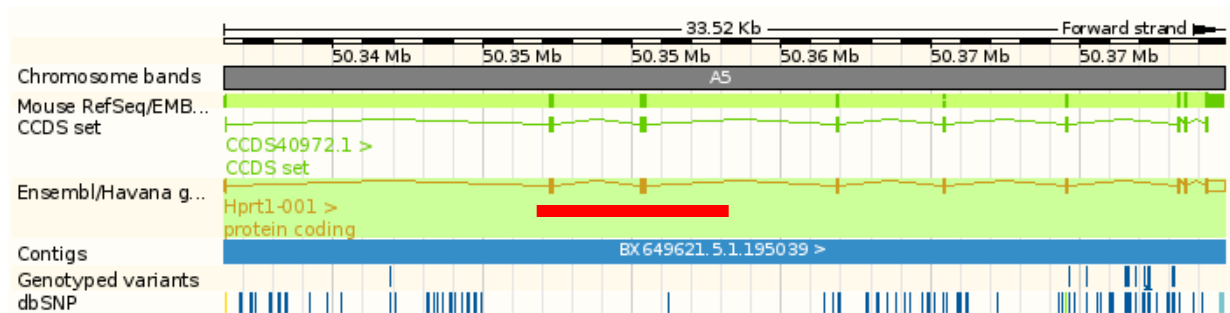
a, Western blot of wild-type (*Blm^{+/+}*) and *Blm*-deficient (*Blm^{e/e}*) ES cells. b, Sister chromatid exchange analysis of heterozygous (*Blm^{e/+}*) and homozygous (*Blm^{e/e}*) *Blm* mutant cells. The sister chromatids are differentially labelled with BrdU and visualised by staining the slides with acridine orange. A SCE is an abrupt point of colour exchange on both sister chromatids.

2.3. Gene targeting of the mutagenic PB to the *Hprt* locus in *Blm^{e/e}* cells

With the *Hprt*-proficient *Blm*-deficient ES cell line established, the mutagenic *piggyBac* transposon was introduced into the *Hprt* locus to generate a *Blm*-deficient ES cell line with a single-copy PB transposon. Two cell lines were generated with the PB transposon targeted in two different locations within the *Hprt* locus, as local surrounding sequence may affect the PB excision efficiency. The designs of both of the targeted alleles are shown in Figure 4-3a,e.

The first cell line was made by introducing my mutagenic PB within the intron 2 of the *Hprt* locus (BlmHprtPBin2 cell line). The targeting vector was constructed with homologous arms derived from a 129 strain-based *Hprt* intron 2 targeting vector obtained from Haydn Prosser (Prosser *et al.*, 2008). Strain-specific single-nucleotide polymorphisms (SNPs) introduced by the targeting arms to the ES cell genome can be present in exons and introns. The genomic region where the targeting arms reside does not show any SNPs in the SNP database (dbSNP), suggesting that this region is highly conserved (Figure 4-3). Therefore introduction of 129 strain-based DNA fragment into the C57BL/6 ES cell genome is unlikely to introduce any variation which may affect Hprt function. The PB recognition site “TTAA” was introduced immediately adjacent to the two PBITRs within the targeting arms to ensure the efficient excision of the PB from the donor site. The mutagenic PB was placed in such an orientation that the *Dom3z* mutagen unit should mediate endogenous Hprt trapping in correctly targeted clones, thus the cells should be HAT sensitive.

Figure 4-3: Genetic variations with the *Hprt* targeting region.



Ensembl screen shot of the entire *Hprt* gene. The red line highlights the region covering both of the targeting arms. The coordinates are based on NCBI Build37.

The linearised HprtTVPB targeting vector was electroporated into the 1×10^7 *Blm^{e/e}* cells and puromycin selection was initiated 24 hours post-electroporation and colonies were picked after seven days. 6-TG selection was not used for the isolation of correctly targeted clones, as homologous recombination can occur in a portion of the daughter cells of the electroporated parental cells. Under 6-TG selection, the correctly targeted cells can be cross killed by their neighbouring untargeted sister cells due to the toxic metabolite sharing. Long range PCR was

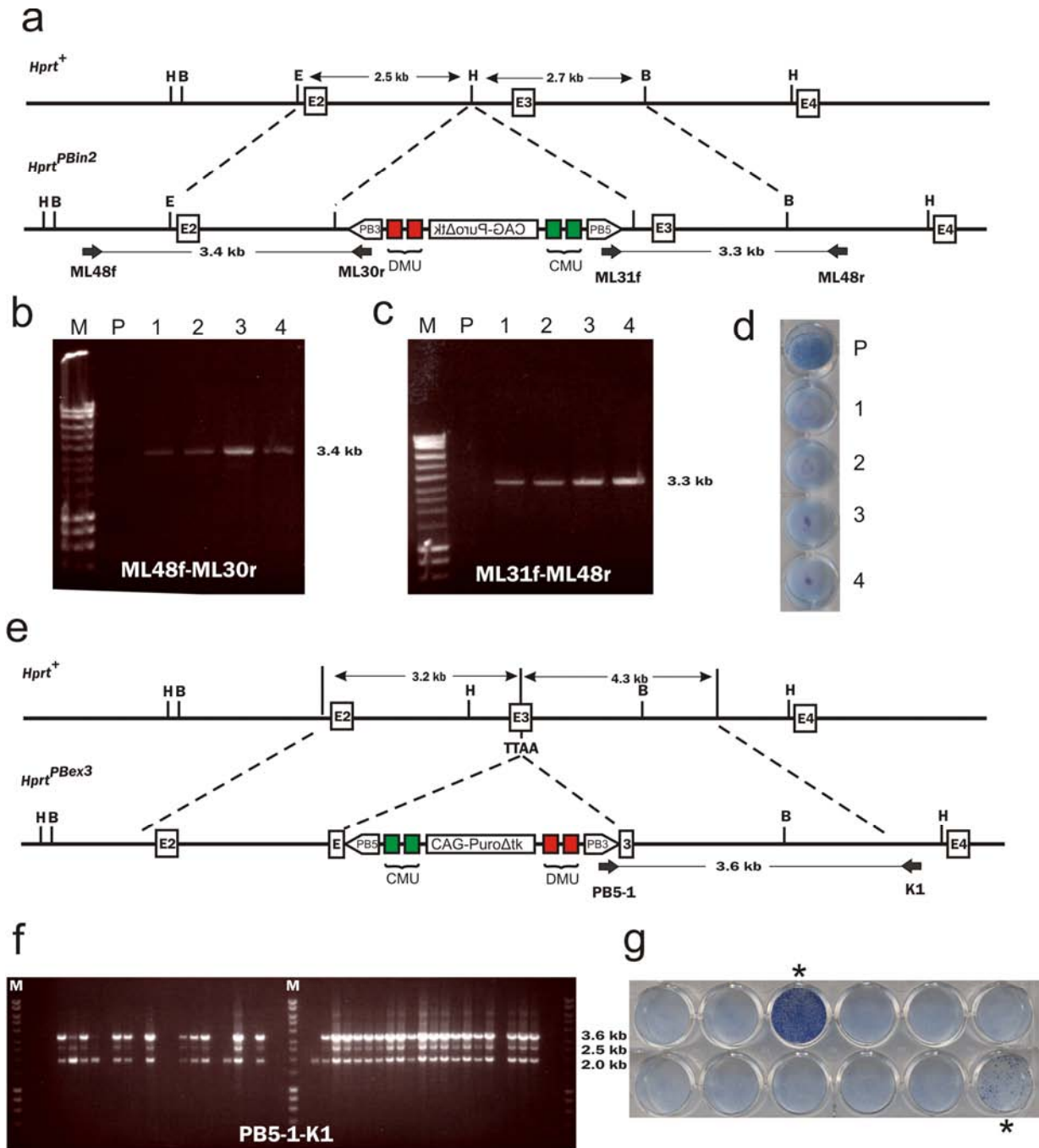
conducted to detect the homologous recombination of the short targeting arm to screen for targeting-positive clones (primers ML48f and ML30r), and the homologous recombination at the other end of the targeting arm was also confirmed by long-range PCR (primers ML48r and ML31f), Figure 4-4b,c. In correctly targeted clones, the *Dom3z* mutagenic unit should trap the *Hprt* expression, thus cells should be sensitive to HAT, whereas random integration of the targeting vectors will not cause *Hprt* inactivation. Therefore, HAT sib-selection was conducted on the PCR positive clones to further validate the gene targeting functionally, Figure 4-4d.

The second cell line was made by introducing the PB into a TTAA site within exon 3 of the *Hprt* gene (BlmHprtPBEx3 cell line), Figure 4-4e. C57BL/6N BAC DNA was used to construct the targeting vector, HprtTVE3PB. The gene targeting procedure was conducted as described above, 29 out of 48 clones were positive for the targeting events screened by the Long-range PCR (with primers PB5-1 and K1) detecting the homologous recombination of the short targeting arm, Figure 4-3f. The HAT selection was also conducted to confirm *Hprt* inactivation by targeted insertion of the PB transposon, Figure 4-4g. Most of the clones show complete sensitivity to HAT, suggesting that the clones are unlikely to contain mixed untargeted cells.

2.4. PB-remobilisation assessment of the newly generated cell lines

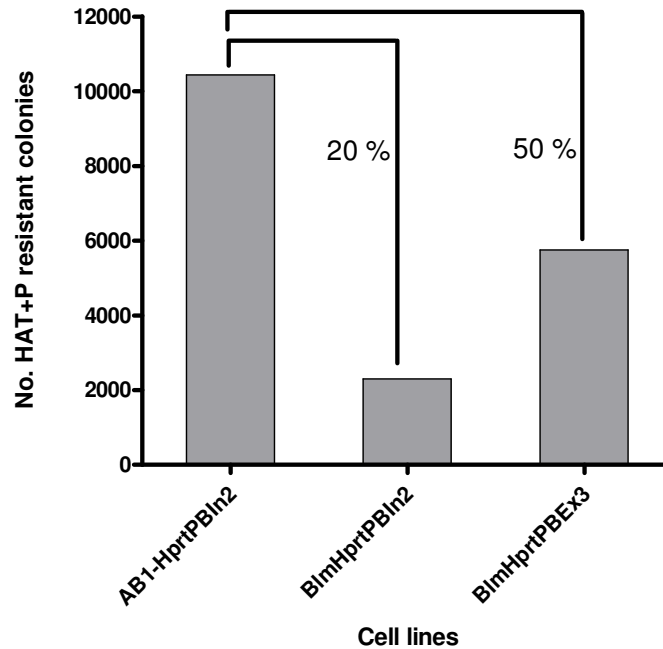
BlmHprtPBIn2 and BlmHprtPBEx3 cell lines were assessed for the efficiency of PB remobilisation. Twenty five µg of CMV-HyPBase plasmid was electroporated into 1×10^7 cells for both cell lines as well as the positive control, AB1- $Hprt^{PBIn2}$, which contains an identical targeting allele of the mutagenic PB transposon in intron 2 of the *Hprt* locus, shown in Figure 4-4a. A tenth of each electroporation was plated on a 90 mm plate and HAT and puromycin double selection was commenced 24 hours post-electroporation to select for PB excision and reintegration, respectively. The colony numbers were counted after ten days selection and are shown in Figure 4-5.

Figure 4-4: *Hprt* targeting of the mutagenic PB transposon.



a, The *Hprt* intron 2 targeted allele with the PB transposon. b,c, long-range PCR confirmations of the targeted clones 1-4. P, parental untargeted cells. d, HAT selection on the targeted clones with the parental untargeted cells as the control. e, The *Hprt* exon 3 targeting with the PB transposon. f, Long-range PCR screening of the homologous recombination on the shorter targeting arm end with PB5-1 and K1 primers. The larger molecular weight PCR product is the correct size whereas others are non-specific PCR products. g, HAT selection on 12 randomly selected PCR-positive clones. Two clones (HAT resistant, marked with *) contain non-targeted cell mixtures. M, Bio-Rad hyperladder I marker.

Figure 4-5: PB re-mobilisation efficiency assessments.



The PB re-mobilisation efficiency was compared among three cell lines with AB1-HprtPBIn2, BlmHprtPBIn2 and BlmHprtPBEx3. AB1-HprtPBIn2 was Blm proficient with PB targeted into intron 2 of the *Hprt* locus. BlmHprtPBIn2 and BlmHprtPBEx3 were derived from Blm-deficient ES cells. PB was targeted to the intron 2 of the *Hprt* locus in BlmHprtPBIn2, while PB was targeted to the intron 3 of the *Hprt* in BlmHprtPBEx3. The number of HAT and puromycin resistant colonies plotted were the total number obtained per 1×10^7 cells. The percentage shown in the graph represents the percentage of HAT and puromycin double resistant colonies in the *Blm*-deficient cell lines vs the positive control. P, puromycin.

Compared to the positive control, the PB remobilisation efficiency in the two newly generated *Blm*-deficient cell lines was lower. Similar experiment was done with HAT selection alone to measure the excision efficiency and the results were similar to that obtained with HAT and puromycin double selection. PB transposon targeting in the BlmHprtPBIn2 cell line was identical to the AB2.2-HprtPBIn2 cell line, however, PB remobilisation efficiency in the BlmHprtPBIn2 and BlmHprtPBEx3 cell lines were only 20 % and 50 % of AB2.2-HprtPBIn2 cell line, respectively.

3. Discussion

In this chapter, I have described the generation of a *Blm*-deficient but otherwise wild-type C57BL/6N ES cell line. The proficiency of wild-type *Hprt* expression permits the use of this locus as a donor site for PB mediated genome-wide re-mobilisation, previously described in Chapter 3. The *Blm*-null ES cell line was generated by sequential gene targeting of a selection cassette to replace the start codon-containing exon 2 of the *Blm* gene, a strategy that has been used previously to generate the *Blm*-knockout ES cells (Luo *et al.*, 2000; Guo, 2004). The selection cassette used here was an essential part of the reporter strategy for the miRNA biogenesis and effector pathway screening design. The detailed features of this reporter strategy are described in Chapter 5.

The new *Blm*-deficient ES cells (*Blm^{e/e}*) were confirmed to be null since *Blm* protein expression could not be detected. In addition, *Blm* deficiency was further assessed functionally by accessing the frequency of the SCE. A highly elevated level of SCE was observed in these *Blm*-deficient ES cells compared to their heterozygous counterpart. Therefore, my *Blm*-deficient ES cell line is a true *Blm* loss-of-function mutant.

The mutagenic PB transposon was introduced into the *Hprt* locus of the *Blm^{e/e}* cells in two independent loci, and both mutagenic units can mediated efficient *Hprt* trapping and gave rise to HAT sensitive targeted clones. This data confirmed further of the mutagenicity of my PB transposon. PB re-mobilisation efficiency was compared in these newly generated cell lines.

PB re-mobilisation in the AB2.2-*HprtPBIn2* cell line is five times more efficient than in the *BlmHprtPBIn2* cell line. The major difference between these two cell lines is their genetic backgrounds. The AB2.2-*HprtPBIn2* cell line was derived from the 129 SvEvBrd strain, whereas the *BlmHprtPBIn2* cell line was derived from C57BL/6 strains. Moreover, the AB2.2-*HprtPBIn2* cell line is *Blm*-proficient whereas the *BlmHprtPBIn2* cells are *Blm*-null. The PB transposition is host independent, as highly efficient transposition can occur in a wide range of species (Chapter One). Thus the difference in genetic background may not be the cause for the

reduction in PB transposition efficiency. Thus the observed difference may be due to *Blm* deficiency, although the connection between the DNA transposon excision and *Blm* has not yet been investigated.

Blm is known to be involved in DNA double-stranded break (DSB) repair using the local sequence microhomology, and *Blm* may function in unwinding the DNA at the breakpoint in order to allow the local microhomology search to repair the DSB (Langland et al., 2002). Although DSB can be repaired via other system such as non-homologous end joining, the loss of *Blm* may partially compromise the DSB repair efficiency upon PB transposon excision. Cell-cycle arrest or apoptosis may be induced in cells without efficient DSB repair, thus contributing to the reduction in colony number. If PB undergoes fast transposition kinetics with several cycles of excision and reintegration occurring per cell, cells harbouring rounds of DSB generated by PB remobilisation are more prone to cell death. Such selection pressure in the *Blm*-deficient background may also enrich cells with less PB intra-genomic mobilisation, and possibly enhance the local hopping events. In yeast, the absence of both *Sgs1* (*Blm* homologue in yeast) and *Exo1* causes a pronounced hypersensitivity to DSB inducing agents and the cells have compromised DSB resection, DNA damage-mediated signalling and strongly impaired homologous recombination-mediated repair (Gravel et al., 2008). *BlmHprtPBIn2* and *AB1-HprtPBIn2* cell lines generated here are very useful reagents for the future investigation into the PB transposon mediated DSB and *Blm*-mediated DSB repair. Research into the relation between the two could potentially reveal biological insights into the mechanisms of the mammalian DNA repair systems which are used to eliminate DNA transposition-mediated genomic insults.

PB re-mobilisation from the *BlmHprtPBEx3* cell line performed three times more efficient than in *BlmHprtPBIn2* cell line. This difference may be due to differences in the excision efficiency, as the PB transposon was introduced into slightly different genomic sequence contexts. Therefore, it is possible that AT-rich local DNA sequence context may favour the PB excision. Furthermore, local methylation status and chromatin structure may also influence the PB excision efficiency.

With respect to the genome-wide insertional mutagenesis, the BlmHprtPBEx3 cell line is a better choice of the two due to its higher PB remobilisation efficiency, Figure 4-5. Despite the reduction in PB remobilisation efficiency observed in *Blm*-deficient ES cells, per electroporation of 1×10^7 of BlmHprtPBEx3 cells, approximately 6,000 HAT and puro double resistant colonies can be generated representing the PB transposon excised from the *Hprt* locus and reintegrated elsewhere in the genome, respectively, Figure 4-5. Based on the previous estimate, 47 % PB insertions are in genes and 80 % of which land in actively transcribed genes in ES cells, assuming no local PB hopping phenomenon is present (Liang et al., 2009). Thus 40,000 heterozygous mutants can be generated in merely seven electroporations in BlmHprtPBEx3 cells, to cover approximately 15,000 PB insertions in actively transcribed genes in ES cells. Therefore in terms of generating a genome-wide mutant library, such a cell line is sufficient enough to cover the whole ES cell expressed genome, assuming half of the genome is transcribed in ES cells. Therefore, the *Blm*-deficient cell line with a mutagenic PB transposon residing in the *Hprt* locus constitutes a very simple, useful and efficient means to generate genome-wide insertional mutagenesis with a tightly controlled mutagen copy number. Libraries of mutants generated using such a cell line can support recessive genetic screens for phenotypically selectable pathways, such as phenotypes involving viral, toxin and drug resistance (Yusa *et al.*, 2004a; Wang and Bradley, 2007; Wang *et al.*, 2008a). This cell line is also highly versatile, as any genetic modification can be engineered into the cell line prior to library generation to suit screen-specific requirements for probing phenotypes which are not directly selectable. This aspect is demonstrated in the latter part of this thesis by introducing reporter systems to probe the miRNA biogenesis and effector pathway genes.

Chapter Five – Development of reporter systems for probing the miRNA pathway

1. Introduction

1.1. Why study miRNA biogenesis and effector pathways?

Understanding miRNA biogenesis and downstream effector pathways is important to reveal crucial regulation points which intersect with other molecular pathways to control phenotypic outcomes in response to external stimuli in different physiological and pathological environments.

Recent advances have brought significant understanding in the field of miRNA biogenesis and downstream effector functions. Many of the components in the miRNA biogenesis pathway have been identified from siRNA-mediated gene silencing and biochemical studies such as pulling down protein complexes associated with known components of the system (Bernstein *et al.*, 2001; Hutvagner *et al.*, 2001; Lee *et al.*, 2003; Han *et al.*, 2004). Although much has been learnt, there are likely to be novel enzymes and regulators yet to be discovered. There are many interesting but outstanding questions regarding mammalian miRNA processing and effector pathways. Questions such as, what are the determinants in specifying the two downstream effector pathways, miRNA-mediated mRNA cleavage or post-translational inhibition? How are RISC complexes containing mature miRNAs degraded or recycled? Is there an “amplifying system” in mammals, equivalent to the system observed in worms and plants, to generate secondary siRNAs to efficiently mediate gene repression?

Forward genetic screens have been conducted in the invertebrate model organisms *C. elegans* and *Drosophila* to discover genes in siRNA-mediated silencing pathway (Kim *et al.*, 2005; Dorner *et al.*, 2006). Dorner and co-workers used dsRNA-mediated RNAi to knockdown genes on a genome-wide scale in cultured *Drosophila* S2R+ cells and followed by transfection of dsRNAs targeting the luciferase reporter system as readout for detecting RNAi pathway abrogation. This dsRNA-based RNAi screening system has identified seven genes, including a known gene *Ago2* and two members of the heat shock protein family members *Hsc70-4* and

Hsp70-3 (Dorner et al., 2006). The specificity and efficacy of the dsRNA-mediated gene knockdown may be limited in this system and the precise function in the RNAi pathway of the novel genes found is yet to be discovered. In the *in vivo C. elegans* study, Kim *et al* engineered a GFP reporter and a dsRNA hairpin targeting GFP in the epithelial seam cells in *C. elegans* to give rise to the “RNAi sensor” strain (Kim et al., 2005). An RNAi screen was conducted by feeding the RNAi sensor strain with specific dsRNAs to knockdown 19,000 genes in the worm genome. Mutant worms with the GFP expression restored in the seam cells may implicate the function of targeted genes in the RNAi pathway. In this screen, 90 genes were identified to have RNAi mutant phenotype and 11 of which are known RNAi components. The newly identified genes include members of the nonsense-mediated decay (NMD) pathway, members of the pre-mRNA cleavage and polyadenylation complex, and factors involved in nuclear transport and chromatin factors. Several genes found in this screen are relevant to the miRNA pathway as miRNA and siRNA share similarities and are artificially interchangeable. However, Kim et al found that several factors are RNAi specific and had little molecular overlap to the miRNA pathway by examining the *let-7*-associated supernumerary seam cell phenotype (Kim et al., 2005). This work suggests that miRNA pathway possess unique factors that are not shared with the RNAi pathway. In addition, mammalian systems might use and regulate these pathways differently to invertebrates. Genetic screens performed in mammalian systems may be able to discover mammalian specific factors. Despite the advantages of non-hypothesis driven approaches in mammalian systems to identify components of these pathways, so far they have not been conducted.

Finally, there is a strong clinical motivation to investigate miRNA biogenesis and effector pathways. It is widely accepted that there is a tight connection between the miRNA pathway and cancer (Lu et al., 2005), which were described in detail in Chapter One. Mutations in miRNA processing genes have been found in a wide range of human cancers and some well-known tumour suppressors, such as p53, have been reported to be directly involved in the biogenesis of several miRNAs in DNA damage response (Suzuki et al., 2009) (Chapter One). Therefore, novel components in miRNA biogenesis and effector pathways may provide insights into the function of genes mutated in cancer whose functions are yet to be defined.

1.2. Challenges in using *Blm*-deficient ES cell system to study the miRNA pathway

One of the main constraints in employing current mixed population-based recessive genetic screens in using the *Blm*-deficient ES cell system is the extremely low ratio between the relevant homozygous mutants to irrelevant cells within the pool. Therefore, a potent but sensitive selection/isolation method is required in order to identify the few genuine homozygous mutants out of approximately 30 million irrelevant cells. This poses a major challenge when one is using the *Blm*-deficient ES cell system to investigate biological pathways that are not directly selectable. Therefore, screen designs, which are able to “translate” the characteristics of a mutant phenotype in the biological pathway of interest into a strong reporter system, is required to conduct the screen.

The miRNA processing and effector pathway falls into a category of “subtle” biological pathways in which mutant phenotypes can not be directly assessed in mouse ES cells. Therefore, a sensitive reporter system is pivotal for the success of the screen. This chapter describes the development of two reporter systems to probe the miRNA pathway, with each designed to specifically explore a branch of the pathway. Both reporter systems share common elements in miRNA biogenesis, thus mutants which affect biogenesis can be identified using both systems.

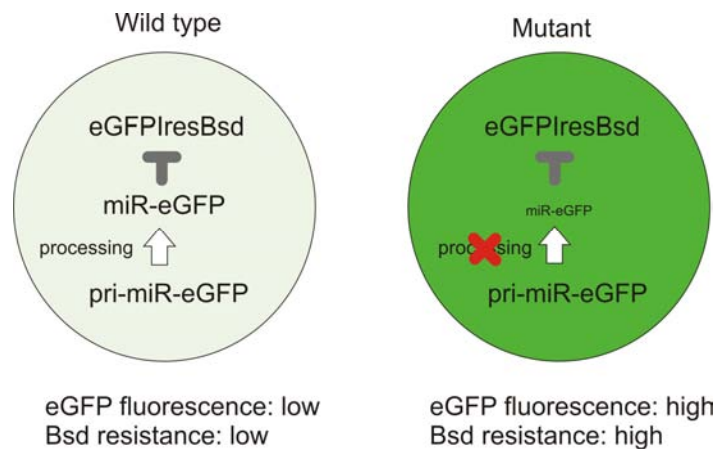
2. Results

2.1. Development of the artificial miR-eGFP system

The first reporter system I developed is based on an artificial miRNA with a seed sequence that perfectly matches the *eGFP* coding region to achieve miRNA-mediated *eGFP* mRNA cleavage. In miRNA pathway proficient cells, *eGFP* mRNA is expressed but cleaved and these cells do not express eGFP protein. A change in fluorescence intensity from non-green to green is a direct phenotypic readout of mutants which affect miRNA processing and effector pathways. Due to the extremely low ratio of mutant to background cells, a drug selection system was also incorporated into the reporter target. To achieve this, the *eGFP* coding region is linked to the Blasticidin S deaminase gene (*Bsd*) via a viral *IRES* sequence, so that *eGFP* and *Bsd* are transcribed as a single mRNA. The repression of *eGFP* caused by artificial miR-eGFP-

mediated mRNA cleavage can also prevent translation of Bsd, hence a miRNA biogenesis mutant will be more resistant to blasticidin selection than corresponding miRNA-pathway proficient cells. Thus blasticidin resistance can be used to directly select homozygous mutants from a mixed pool. Background mutants which affect Bsd resistance by other mechanisms can be excluded by further examination of the eGFP expression level. Figure 5-1 illustrates this reporter system.

Figure 5-1: Schematic representation of the artificial miR-eGFP system.



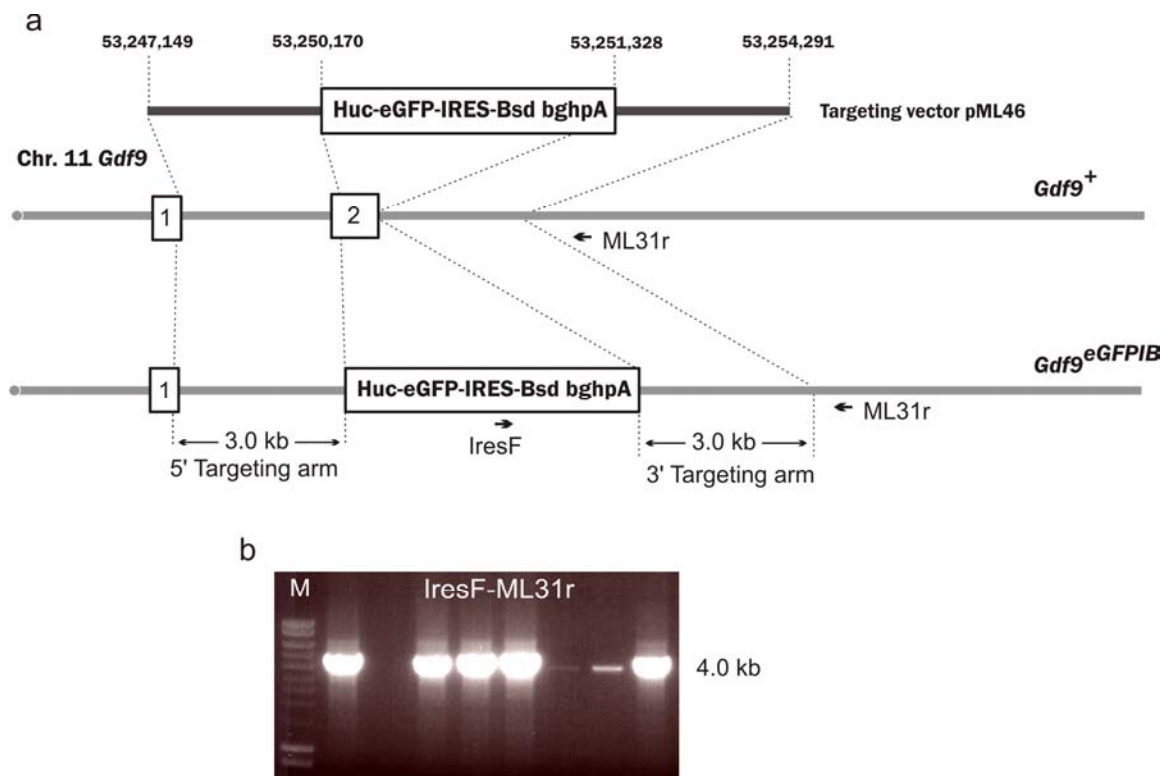
In this reporter design, the essential components are the eGFP-IRES-Bsd reporter and a miR-eGFP which efficiently represses its expression. This system needs to be stable yet sensitive enough to distinguish processing mutants from non-mutant cells. The eGFP-IRES-Bsd construct has been described in Chapter 4. It was targeted into both alleles of the *Blm* locus while making *Blm*-deficient ES cells (Chapter 4). The homozygosity of this reporter is an important design feature of this screen since in a heterozygous context, a major background in the screen would have been the loss of the heterozygosity of this reporter.

An artificial miR-eGFP system was generated based on human miR-30 and mouse miR-155 backbones to generate miRNAs with target recognition sequences to the eGFP coding region. The target recognition sequences were designed to be perfectly complementary to the eGFP sequence, thus the effector pathway downstream is expected to be miRNA-mediated mRNA cleavage.

2.1.1.1. Generation of an eGFP-knockin cell line

While the *Blm*-deficient ES cell line was under construction, an eGFP-IRES-Bsd knockin cell line was generated in order to validate the artificial miR-eGFP designs using the *Blm*-deficient NN5 ES cells. A targeting vector (pML46) was constructed to introduce the eGFP-IRES-Bsd reporter into the *Gdf9* locus, giving rise to a new cell line NN5- *Gdf9*^{eGFP/+}. The targeting vector pML46 was linearised by *NotI* and electroporated into 1×10^7 NN5 cells and plated into a 90 mm culture plate with a mono-layer of G418 and blasticidin double-resistant feeders. The following day, blasticidin selection (10 μ g/ml) was initiated. The cells were selected for seven days and 48 colonies were picked and genotyped using long-range PCR (with primers IresF and ML31r) to identify homologous recombination events at the 3' targeting arm, Figure 5-2.

Figure 5-2: Targeting *Gdf9* locus with an eGFP-IRES-Bsd reporter.

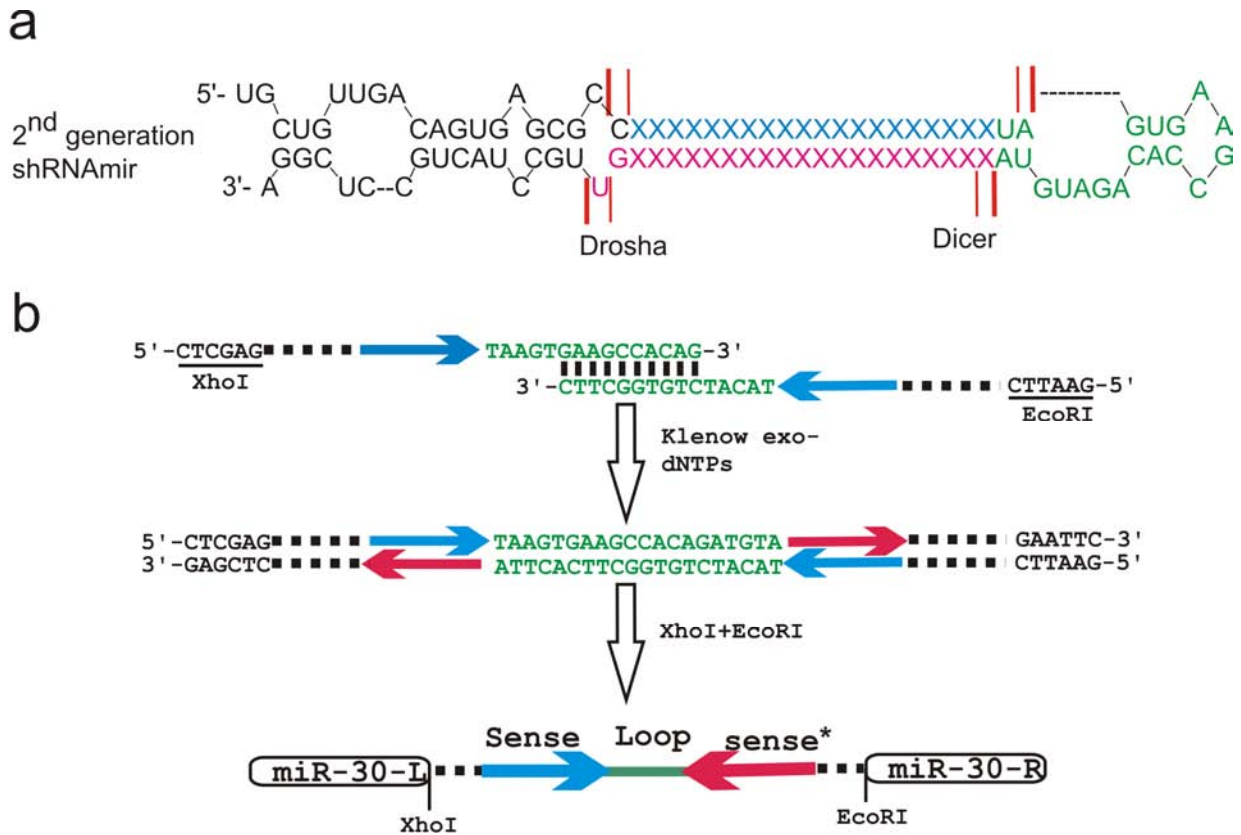


a, Targeting strategy of the *Gdf9* locus with the reporter construct **b**, Long-range PCR confirmation of positively targeted ES clones using primers Ires-F and ML31r. The genomic fragments used as targeting arms are indicated in the targeting vector with coordinates based on NCBI build37.

2.1.2. miR-30-based miR-eGFP generation and analysis

The miR-30-based shRNA (Figure 5-3a) has been shown to mediate efficient gene knockdown previously (Silva et al., 2005). A schematic representation of the procedures used to generate customised miR-eGFP using the human miR-30 backbone is shown in Figure 5-3b. A 20 nt eGFP recognition sequence and its complementary sequence were incorporated into two synthetic oligo-nucleotides juxtaposed to the miR-30 loop region. The two oligos were annealed and subsequently extended by Klenow exo⁻ fragment to give rise to a double strand DNA fragment coding for the miR-eGFP. The miR-eGFP fragment was then digested with restriction enzymes and cloned into a plasmid. The miR-eGFP backbone and its target-sequence region were confirmed by sequencing. Two different miR-eGFP target recognition sequences used in this experiment were designed based on experimentally-validated eGFP knockdown from published literature. These two target sequences recognise different parts of the eGFP coding region and were used to construct miR-30-based miR-eGFP, and they were named no.44 and no.91.

Figure 5-3: The construction of the miR-30 backbone based single-unit miR-eGFP.

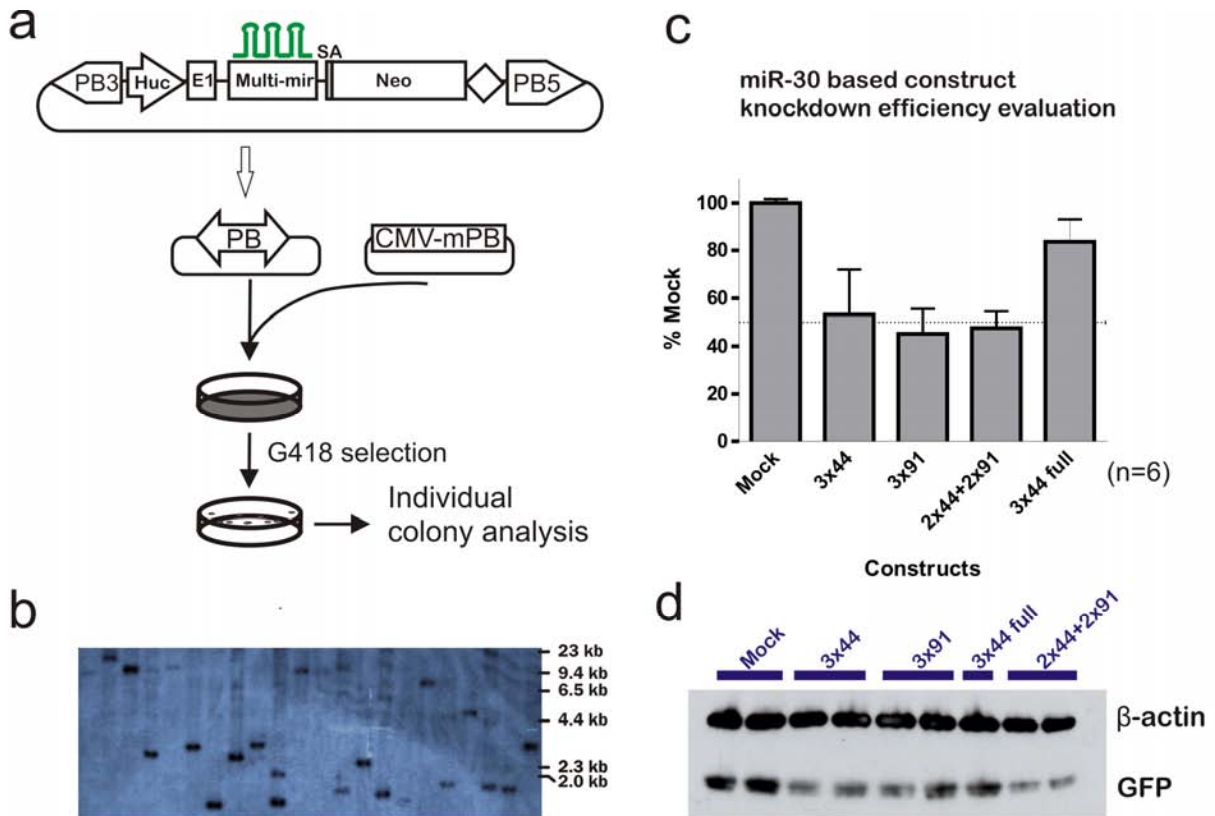


a, The human miR-30 backbone used for engineering of second-generation shRNA-miR described by Silva and co-workers, and the diagram is adapted from (Silva et al., 2005). The 3' strand (pink) is the mature miRNA post processing. The red lines represent the Drosha and Dicer processing sites while the heavy and thin red lines represent the major and minor processing sites respectively. b, The cloning method used to generate the miR-30 based miR-eGFP. Two long synthetic oligos including the 20 nt eGFP targeting sequencing (the blue and red arrows which represent the sense and anti-sense strand sequences respectively) were annealed with ten nucleotides overlapping sequence at the loop region. The partially annealed oligos were treated with Klenow fragment (exo⁺) to form a dsDNA fragment. The fragment was then cut with restriction enzyme at the sites indicated and cloned into a plasmid with the corresponding restriction enzyme sites. The colour scheme in a and b is correlated.

In order to increase the repression efficiency, multiple copies of the miR-eGFP was constructed in a head-to-tail fashion to give rise to miR-eGFP multimers “3x44” (three copies of miR-eGFP with no.44 as the eGFP recognition sequence), “3x91” (three copies of miR-eGFP with no.91 as the eGFP recognition sequence) and “2x44+2x91” (two copies of miR-eGFP with no.44 and no.91 as the eGFP recognition sequence each). Finally, a miR-eGFP multimer “3x44 full” was also generated with three copies of the miR-eGFP with no. 44 as the recognition sequence. However in this case, each miR-eGFP includes 200 bp of the DNA sequences flanking the miR-30 hairpin, as it has been proposed that this “context” can enhance the miRNA initial processing by the Drosha complex (Silva et al., 2005). The miR-eGFP multimers were inserted into intron 1 of the human ubiquitin C (huc) fragment containing the promoter region to generate a miR-eGFP and neomycin resistant gene co-expressing construct, Figure 5-4a. This architecture ensures the constant expression of miR-eGFP in G418 resistant ES cells when the construct is stably integrated. The construct was further introduced within the PB transposon, allowing the stable and simple delivery of the constructs to reporter ES cells, Figure 5-4a. Using the PB transposon and PBase co-electroporation system, stable integration of the PB transposon can be achieved with one copy per cell by electroporating 1×10^7 NN5-Gdf9^{eGFP/+} cells with 100 ng of the PB transposon plasmid and 5 μ g of the mPBase Δ Neo plasmid. Southern blotting, shown in Figure 5-4b, confirmed that the majority of the G418 resistant colonies generated under these conditions contain a single copy of the PB transposon.

Plasmids with the PB transposons carrying different designs and combinations of the miR-eGFP multimers were co-electroporated with mPBase into the NN5-Gdf9^{eGFP/+} reporter cell line. After eight days of G418 selection, six colonies were picked from each transfection corresponding to the different miR-eGFP designs and these clones were analysed for eGFP expression, Figure 5-4c. As a control, NN5-Gdf9^{eGFP/+} cells were transfected with the PB transposon containing the Huc-Neo cassette without the miR-eGFP (mock).

Figure 5-4: miR-30-based artificial miR-eGFP evaluation.



a, Three copies of the miR-30-based miR-eGFP were incorporated within the intron of a neomycin resistant cassette (Neo) cassette driven by the human ubiquitin C (Huc) promoter. The exon 1 and the intron were derived from the endogenous human ubiquitin C gene and the splice acceptor from the endogenous exon 2 of the human ubiquitin C gene was placed upstream of the Neo coding region. The entire expression cassette was cloned into the PB transposon. The miR-eGFP expression construct was delivered into the NN5- Gdf9^{eGFP/+} using the PB transposon system under single copy per cell delivery conditions. b, Southern blot using PB5'ITR as the probe to determine the copy number of the PB-miR-eGFP-Neo integrations. The genomic DNA was digested with *Pst*I. c, flow cytometry analysis of eGFP intensity in G418 resistant colonies. Six colonies were analysed for each type of construct. d, western blot confirmation of the eGFP protein level in G418 resistant colonies.

All miR-eGFP designs achieved a degree of eGFP knockdown. Both eGFP target recognition sequences ("3x44" and "3x91") performed similarly, giving approximately 50 % of the eGFP repression. The combination of the two different eGFP target recognition sequences

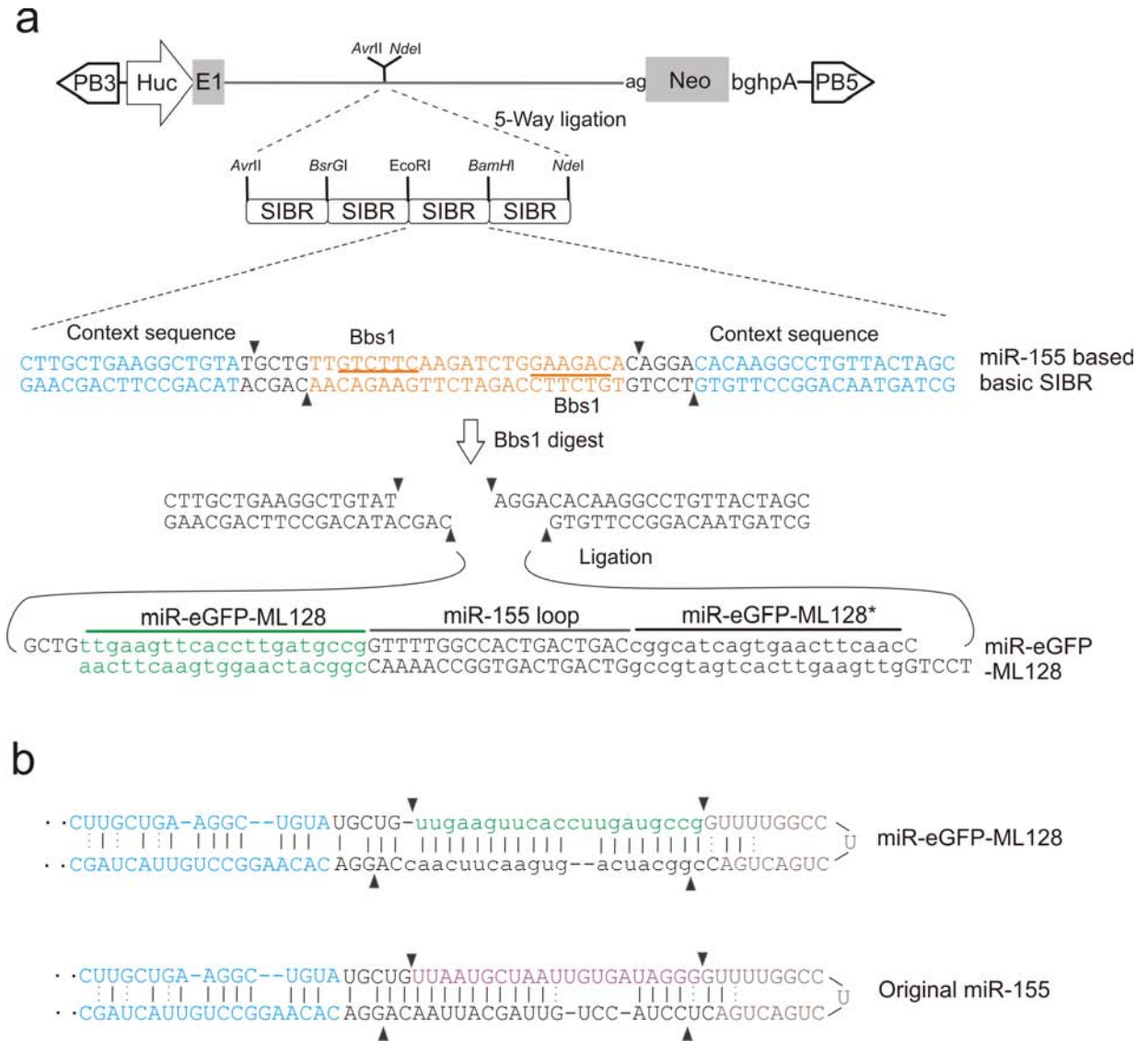
("2x44+2x91") did not result in synergistic repression, the level of repression was comparable to individual eGFP recognition sequences of similar copy numbers. However, the miR-eGFP multimer, which included the miR-30 flanking sequences ("3x44 full"), performed worse than the smaller version ("3x44"). This may reflect synergistic processing of miR-eGFP multimers, as in the "3x44 full" construct, the individual miR-eGFP hairpins were separated by 400 bp of sequence. When the miR-eGFP hairpins are directly next to each other, they might facilitate the recognition of Drosha on adjacent processing sites or they might enhance pri-miRNA processing itself. eGFP protein levels of representative clones were further analysed from all types of the miR-eGFP constructs by Western blotting. The eGFP protein levels showed a similar level of repression to the results obtained from fluorescent analysis, Figure 5-4d.

2.1.3. Construct optimisation for the artificial miR-eGFP construct

The observation of repression in eGFP suggests that the artificial miR-eGFP generated can be processed to mediate an eGFP-specific "knockdown". However, the level of repression achieved by all constructs was not sufficient for screening purposes. The inefficiency of the knockdown could be due to inefficient processing of miR-30, the target recognition sequence itself not being sufficient in mediating repression or the promoter used is not being strong enough to drive adequate levels of the miR-eGFP.

To optimise the knockdown efficiency, a set of constructs were generated using different promoters, target sequences and miRNA backbones. A panel of total 12 constructs (Table 5-1) were tested, including three different promoters (Huc, CAG and human EF1 α promoters), five different target sequences (44, 91, 126, 127, and 128, the actual sequences are described in the methods in Chapter 2), two miRNA backbones (mouse miR-155 and human miR-30) and various copy numbers of the multimer. The mouse miR-155 backbone was based on the design of synthetic inhibitory BIC-derived RNA (SIBR) cassette, constructed by Chung and co-workers (Chung et al., 2006). The precise sequence of the SIBR cassette and structure of the miR-155 backbone based miR-eGFP multimer is shown in Figure 5-5 with miR-eGFP 128 as an example. The constructs were evaluated for the eGFP knockdown efficiency as described for the miR-30 backbone (Figure 5-4a).

Figure 5-5: miR-155 backbone based miR-eGFP multimer generation.



a, The design of the miR-eGFP and Neo co-expression cassette in PB transposon. Each synthetic inhibitory BIC-derived RNA (SIBR) cassette contains a single miR-eGFP unit with the miR-155 as the backbone. The eGFP recognition target and its antisense sequence containing oligos were annealed and the double-strand fragment was ligated into *BbsI* digested SIBR cassette. b, An example of a miR-eGFP (with target sequence 128) compared to the original miR-155 hairpin structure. The wobble base pairing is shown with dotted lines, whereas the full base pairing is illustrated with straight lines. The black triangles indicate the Drosha and Dicer main processing sites.

Table 5-1: Summary of different miR-eGFP constructs generated.

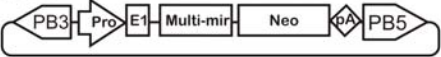
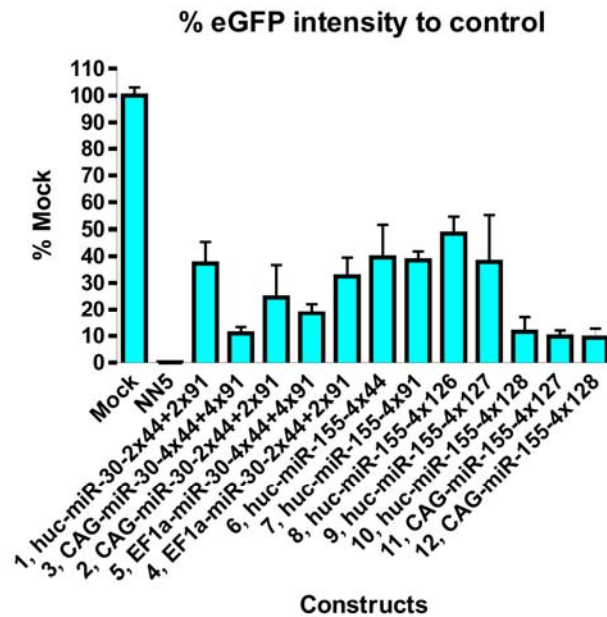
Basic structure				
				
Construct	Promoter	Backbone	Target seq.	Copy No.
1	Huc	miR-30	44 and 91	2 each
2	CAG	miR-30	44 and 91	2 each
3	CAG	miR-30	44 and 91	4 each
4	EF1a	miR-30	44 and 91	2 each
5	EF1a	miR-30	44 and 91	4 each
6	Huc	miR-155	44	4
7	Huc	miR-155	91	4
8	Huc	miR-155	126	4
9	Huc	miR-155	127	4
10	Huc	miR-155	128	4
11	CAG	miR-155	127	4
12	CAG	miR-155	128	4

Figure 5-6: Optimisation of an efficient miR-eGFP to knockdown eGFP.



Analysis of eGFP intensity for each construct. NN5, the parental cell line without the eGFP reporter knock-in. Mock, transfection of a plasmid containing the PB transposon with the PGK-Neo cassette. The values shown are an average of four clones assessed for each construct.

The performance of each miR-eGFP construct was assessed by electroporation of 1×10^7 NN5-Gdf9^{eGFP/+} cells with 100 ng of each construct together with 5 μ g of the CMV-mPB Δ Neo plasmid. The electroporated cells were plated on a 90 mm plate and selected with G418. After eight days selection, four colonies were picked from each plate and the colonies were expanded and were subjected to fluorescence analysis. The results of eGFP intensity measurements of all constructs are shown in Figure 5-6.

Comparing the eGFP repression among constructs 6-10, which have identical promoter and miR-eGFP copies but have different target sequences, target sequence 128 showed 90 % repression of the control eGFP level, whereas other sequences only resulted in 50 % - 60 % repression. Thus, the nature of the sequence is the major determinant of the knockdown efficiency.

The effect of different promoters can be compared with constructs 1, 2 and 4, which contain identical target sequences. The CAG-driven miR-eGFP multimer gave the best knockdown efficiency as expected, since CAG is the strongest promoter of the three and Huc is the weakest. High levels of expression improve the degree of knockdown by providing the miRNA processing reactions with more substrate. However, the effect is reduced when the knockdown is already relatively efficient as shown by comparing constructs 10 and 12. This may reflect the fact that in cases where the knockdown is efficient, the miRNA processing and effector silencing machineries are close to saturation.

The effect of copy number of the miR-eGFP within the multimer can also be compared with constructs 2 and 3 as well as constructs 4 and 5. Increasing numbers of hairpins provided only a moderate increase in the knockdown efficiency. The synergistic effect observed previously (Figure 5-4c, “3x44 full” vs “3x44”) is most dramatic when the copy number increases from one to three (data not shown). When additional copies were incorporated, the synergistic effect diminished. Thus the slight increase in efficiency is due to the increase in precursor miR-eGFP substrate only.

Finally, the miRNA backbone can be compared in construct 1 vs 6 and 7. It was found from the previous experiment (Figure 5-4c) that target sequences 44 and 99 perform similarly in eGFP knockdown. Thus construct 1 with the combination of two different target sequences can be compared to constructs 6 and 7 with target sequence 44 and 99 singly, respectively. miR-30 and miR-155 backbone did not show any difference in the repression efficiency.

From these comparisons, miR-eGFP with sequence 128 in a multimer format provided the best repression. Although the CAG promoter was optimal in this analysis, it may be prone to silencing in mouse ES cells due to the presence of cytomegalovirus enhancer sequence within the promoter. Since the promoter strength is less significant when an optimal target sequence was used, Huc driven miR-eGFP multimer with sequence 128 was selected for my reporter miR-eGFP construct as this will provide a comparable repression to the CAG-driven construct.

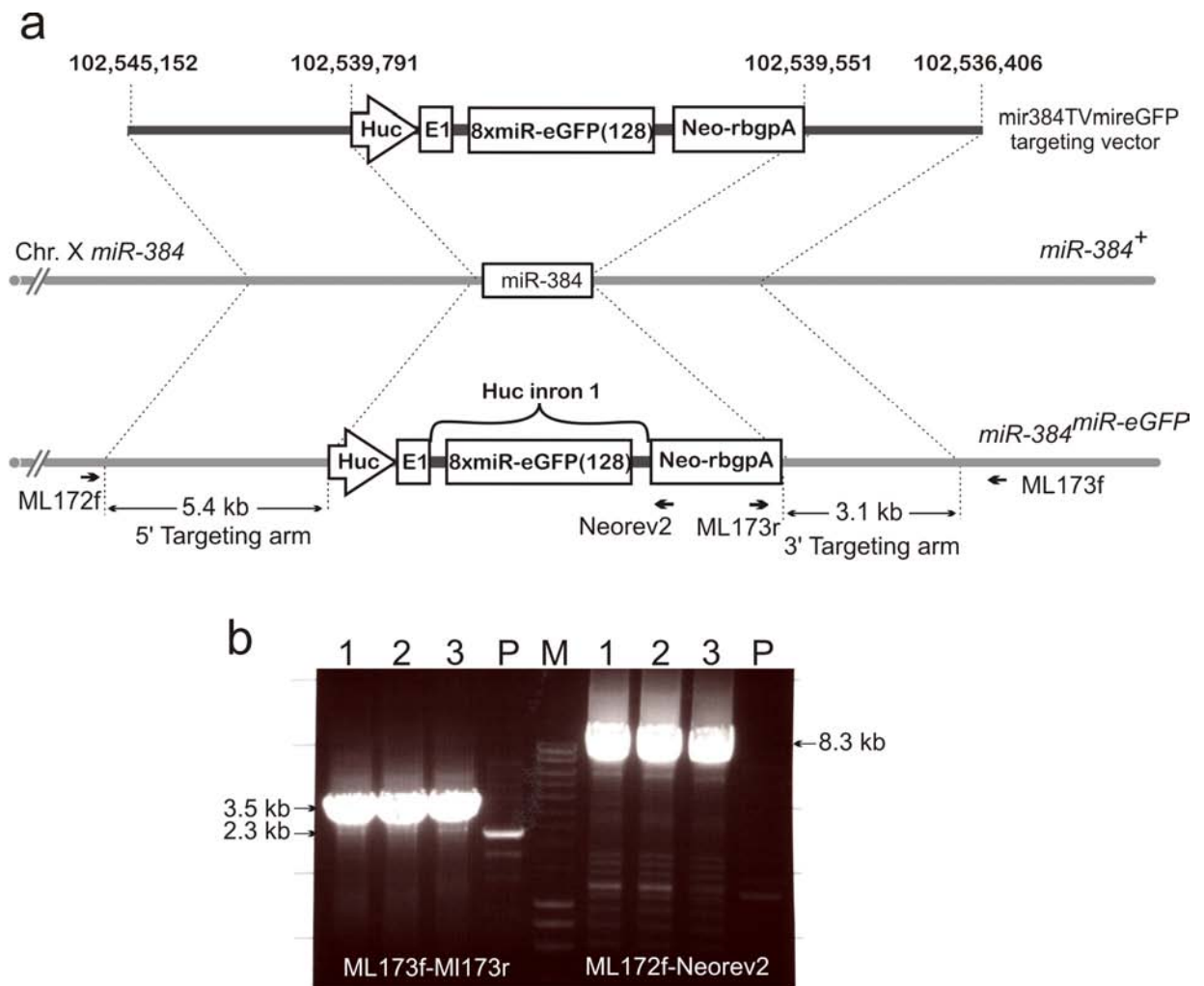
2.1.4. Generation of an artificial miR-eGFP knockin in the *Blm^{e/e}Hprt^{PBin2}* cell lines

In order to provide a uniform and constant level of miR-eGFP expression, a polycistronic expression cassette containing a Huc-driven miR-eGFP octamer with target sequence 128, residing within the intron of the Neomycin resistance gene was targeted to the X-linked *mir-384* locus. An X-linked locus was selected because this chromosome can not be lost in XY ES cells. The *mir-384* locus was selected because it is not expressed in mouse ES cells, although it is expressed in neurons (Marson et al., 2008). Therefore, disruption of *mir-384* is unlikely to result in an ES-cell phenotype.

A targeting vector miR-384TVmiR-eGFP was constructed by replacing the PGK-Puro cassette on the *mir-384* targeting vector with the Huc-driven miR-eGFP multimer and Neo resistance gene co-expressing cassette using recombineering (Details of the construction were described in methods in Chapter 2). The miR-384TVmiR-eGFP targeting vector was linearised with *SbfI* and electroporated into 1×10^7 BlmHprtPBin2 cells. The electroporated cells were selected with G418 for seven days and 48 colonies were picked and screened for correctly targeted events using long-range PCR (with primers ML173f and ML173r) across the short targeting

arm. This yielded three PCR-positive clones, giving rise to a targeting efficiency of 6 %. These three clones were further confirmed for the homologous recombination at the long targeting arm by long-range PCR (with primers ML172f and Neorev2). Figure 5-7 shows the gene targeting strategy and the long-range PCR confirmations of the correctly targeted clones.

Figure 5-7: Gene targeting of miR-eGFP multimer construct to the *miR-384* locus.



a, Gene targeting strategy with the targeting vector, wild-type *miR-384* allele and the targeted allele shown. The genomic fragments used as targeting arms are indicated in the targeting vector with coordinates based on NCBI build37. b, Long-range PCR confirmation of correctly targeted clones for both targeting arms. P, the parental untargeted cell line. 1,2,3, three independent clones. M, Bio-Rad hyper ladder I marker.

2.1.1.5. Validation of the miR-eGFP knockin *Blm^{e/e}; Hprt^{PBin2}* cell line

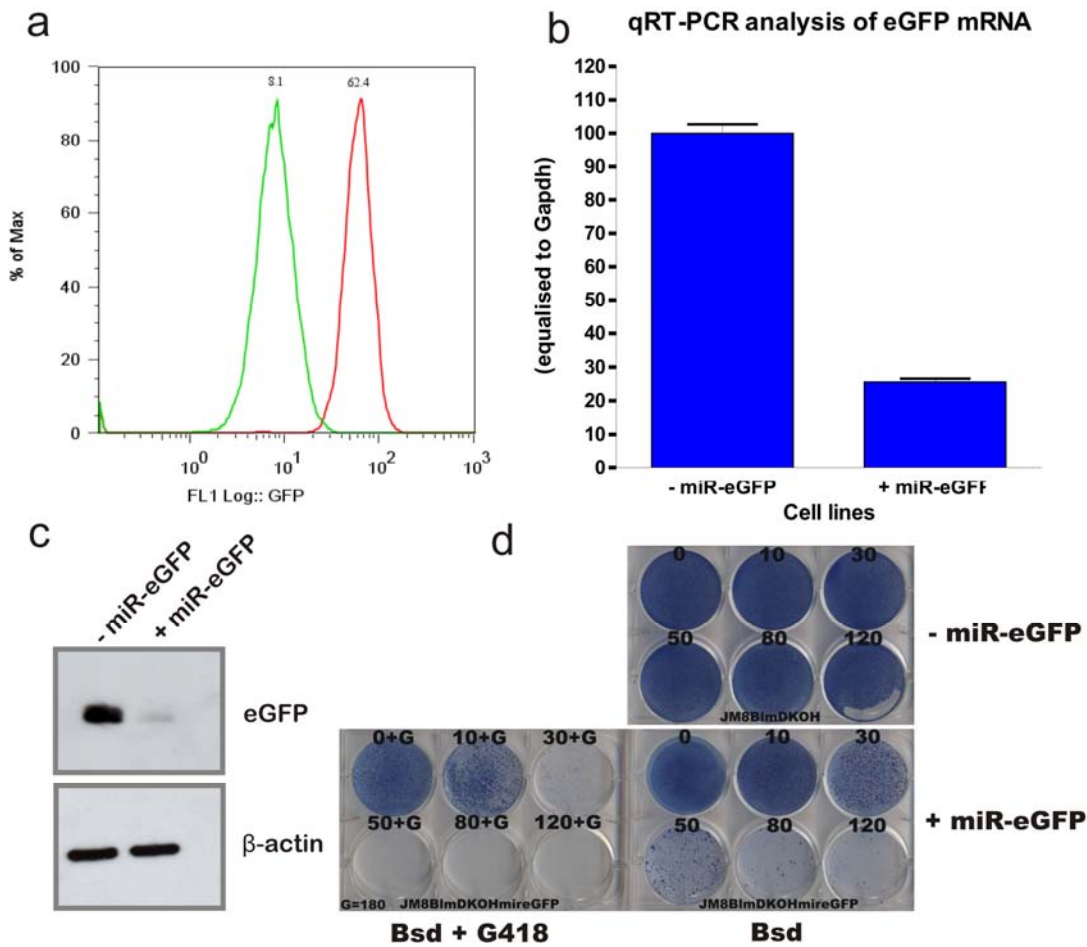
The three correctly targeted clones were subjected to karyotype analysis. Eight metaphases of each clone were examined. Two clones showed a normal karyotype with 40 chromosomes present in all metaphase spreads analysed and one clone had 40 % of the metaphases showing 39 chromosomes, possibly due to loss of the Y chromosome. One of the two clones with normal karyotype, B11, with the *Blm^{e/e}; Hprt^{PBin2}* background was chosen for subsequent experiments. The eGFP fluorescence of individual cells of this miR-eGFP targeted cell line was compared to the parental cell line without the miR-eGFP octamer. This analysis confirmed that the introduction of miR-eGFP resulted in approximately 90 % reduction in eGFP expression, Figure 5-8a. The eGFP mRNA and protein level was also confirmed to be significantly reduced, Figure 5-8b,c.

The repression of eGFP by a miR-eGFP with perfect complementarity to eGFP should mediate mRNA cleavage and subsequent degradation. Since Bsd is translated from the same mRNA as eGFP, it was hypothesised that resistance to blasticidin would be reduced in miR-eGFP knockin cells. If this is the case, the screen can be conducted using blasticidin selection to eliminate the majority of the irrelevant cells in the mutant pools. A Blasticidin titration was conducted with concentrations ranging from 10 µg/ml to 120 µg/ml. The miR-eGFP knockin cells are more sensitive to high concentrations of blasticidin, with blasticidin concentration of 80 µg/ml and above clearing the majority of the cells. Although a few colonies arose from miR-eGFP knockin cells in high blasticidin concentrations, double selection with 180 µg/ml G418 and 80 µg/ml blasticidin completely cleared the wells. This observation suggests that these clones were derived from a very small proportion of cells in which the artificial miR-eGFP has been silenced. Therefore, dual blasticidin and G418 selection eliminate this false positive background.

Taken together, a reporter cell line has been successfully generated for probing the miRNA biogenesis pathway with a mRNA-mediated cleavage effector pathway. An artificial miR-eGFP was generated and the knockin of a single copy of this miR-eGFP octamer repressed up to 90 % of the eGFP expression in *Blm^{e/e}* cells with two copies of an eGFP expression cassette. In

addition, miR-eGFP knockin cells are much more sensitive to high concentrations of the blasticidin than cells without miR-eGFP due to the miRNA mediated cleavage of the *eGFP**RES**Bsd* mRNA. This selection scheme provides a simple method to enrich homozygous mutants within mixed mutant pools.

Figure 5-8: miR-eGFP-knockin cell line validation.

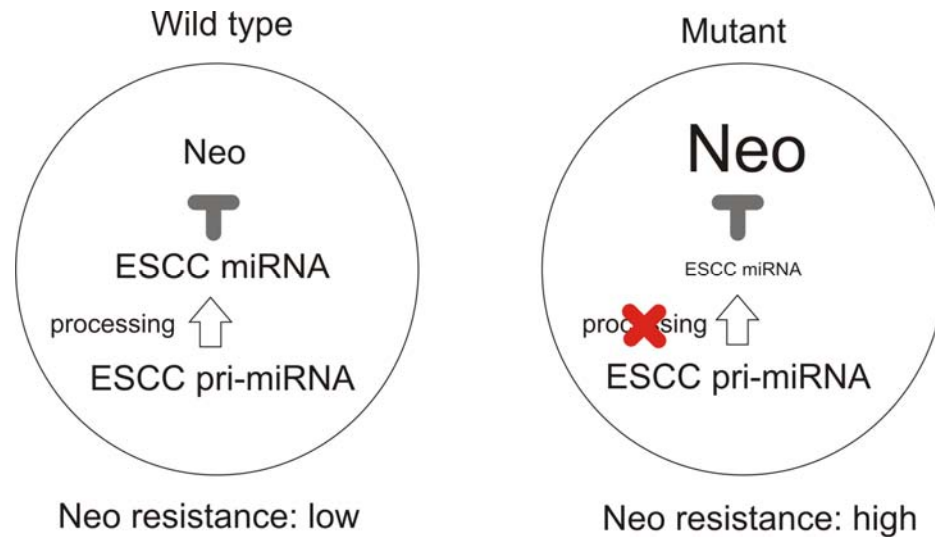


a, Fluorescence analysis of the eGFP intensity in miR-eGFP knockin cell line (green) compared to parental cell line (red). 10,000 events were gathered for the generation of the histogram. b, qRT-PCR (n=3) and c, protein level analysis of the miR-eGFP knockin cell line. d, Blasticidin resistant titration of the miR-eGFP knockin cell line (+miR-eGFP) compared to the parental cells without the miR-eGFP (-miR-eGFP). The blasticidin concentration used for each well was shown above with $\mu\text{g/ml}$ as the unit. Blasticidin and G418 double-selection was also conducted with a constant 180 $\mu\text{g/ml}$ G418.

2.2. Development of the endogenous miRNA target reporter

As previously described, a miR-eGFP system has been generated which probes one branch of the miRNA effector pathway. In order to assess the alternative effector pathway, translational repression, another reporter system is required. In lieu of an artificial miRNA reporter system, a highly expressed endogenous miRNA can be used and a reporter cassette can be generated to provide a selection scheme for mutant identification.

The mir-290 cluster is highly expressed in mouse ES cells comprising up to 70 % of all the miRNA expressed in ES cells (Marson et al., 2008). Four out of six mature miRNAs encoded within the mir-290 cluster, mir-291a-3p, mir-291b-3p, mir-294 and mir-295, share the same seed sequence AAGUGCU (Wang et al., 2008c). In addition, three miRNAs derived from the mir-320 cluster, mir-320b, mir-320c and mir320d, also share the same sequence, although the expression level of mir-320 is relatively low in mouse ES cells (Marson et al., 2008). These miRNAs are termed ES cell specific Cell Cycle regulating miRNAs (ESCC) miRNAs, and one of their functions is to promote the rapid G1-S cell-cycle transition in ES cells (Wang et al., 2008c). The abundance of the ESCC miRNAs with the identical seed sequence makes them an attractive candidate for use as a miRNA reporter, as the target suppression should be highly efficient due to this miRNA redundancy. Thus, a reporter cassette with the ESCC miRNA seed sequence recognition sites in the 3'UTR of the reporter can provide a direct repression of this reporter by the ESCC miRNA. In miRNA biogenesis mutants, the lack of miRNA elevates the reporter repression, providing a selection scheme for the screen. Figure 5-9 shows the reporter strategy.

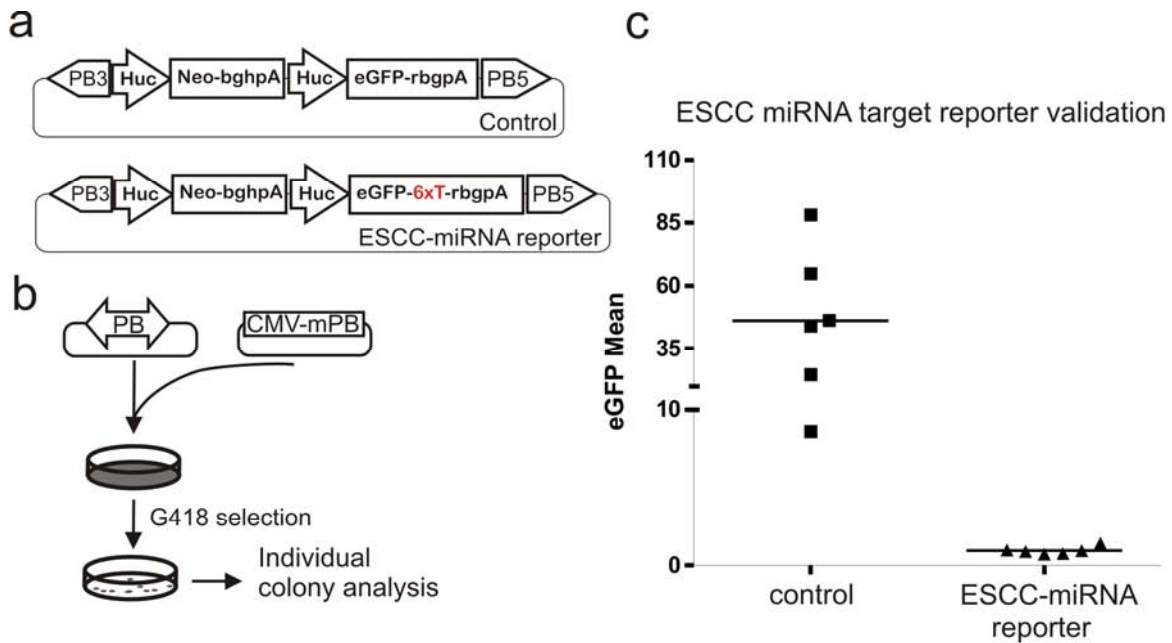
Figure 5-9: An endogenous miRNA-based reporter system.

2.2.1. Artificial ESCC-miRNA target generation and validation

Several of the ESCC miRNA targets and target sites have been validated *in vitro*, including the cell cycle regulator, the inhibitor of the cyclinE/Cdk2 complex, *Cdkn1a* (p21) (Wang et al., 2008c). One of the ESCC-miRNA target sites within the 3'UTR of *Cdkn1a* and its 50 bp surrounding sequences were PCR-amplified and six tandem copies of the target sites were engineered in between the eGFP coding sequence and the polyadenylation signal to detect reporter knockdown by the endogenous ESCC-miRNAs, Figure 5-10a. The negative control in this experiment is an eGFP expression construct without ESCC target sites. The reporter expression cassette and the control were stably delivered into the ES cell genome by the PB transposon. Co-electroporation protocol used in this experiment was designed to deliver one copy of the transposon per genome and eGFP expression was examined in G418 resistant colonies, Figure 5-10b. AB2.2 cells (1×10^7) were co-electroporated with 100 ng PB transposon-containing plasmid and the mPBase-expression plasmid. The cells were selected with G418 for eight days and the colonies were picked, expanded and analysed using fluorescent activated cell sorting (FACs) for their eGFP intensity. Colonies derived from cells transfected with control PB plasmid gave rise to eGFP with varying intensities. This may be due to the locus-specific influence on eGFP expression. The colonies derived from cells transfected with the

ESCC-miRNA reporter PB plasmid expressed little eGFP. Therefore, the ESCC-miRNA reporter seems to be efficiently repressed by the endogenous ESCC miRNAs.

Figure 5-10: ESCC-miRNA eGFP reporter analysis.

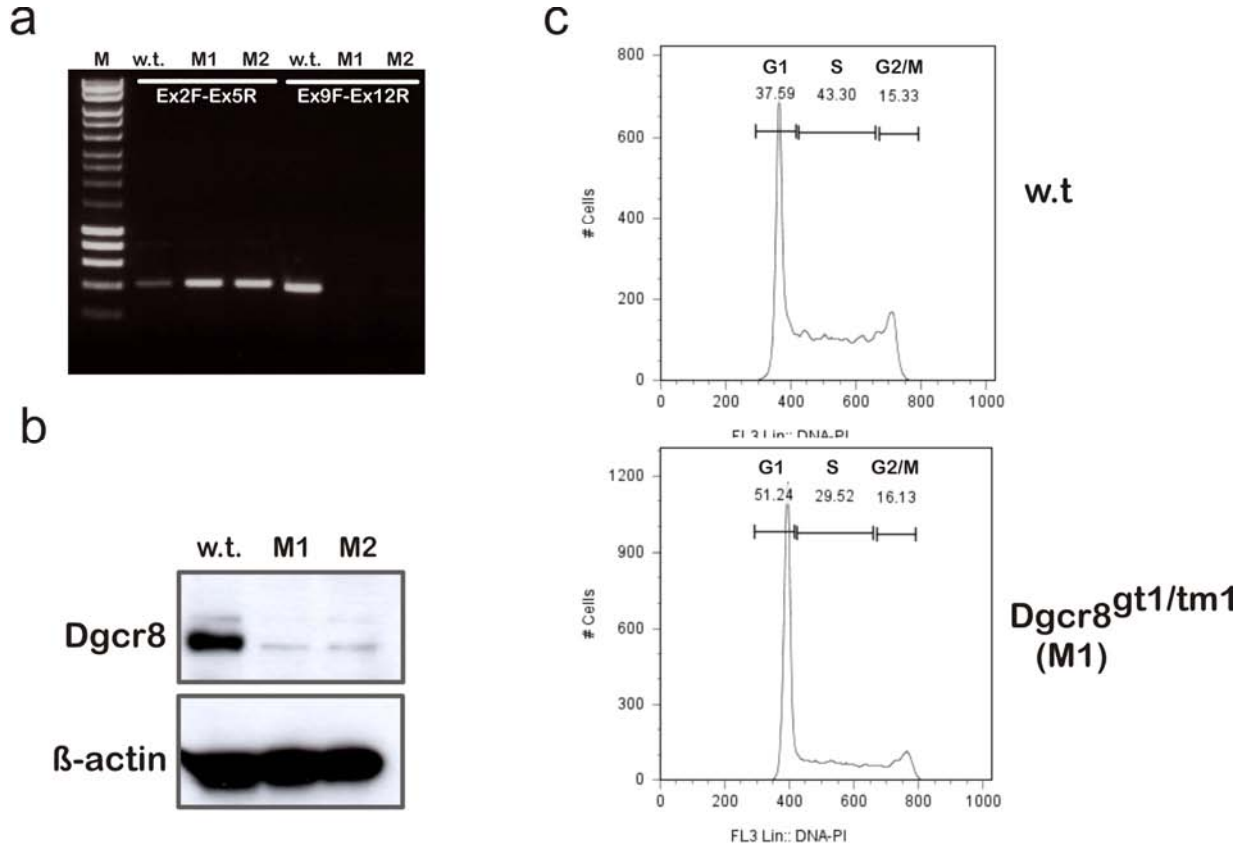


a, the ESCC-miRNA target reporter and the control in a PB transposon. b, PB transposon and transposase co-electroporation scheme to deliver the reporter and control constructs using single-copy PB delivery conditions. c, Analysis for eGFP intensity of six individual clones picked from each electroporation. The lines in the graph are the median values of the six clones.

In order to confirm that suppression of the ESCC-miRNA reporter is mediated by endogenous miRNA, the ESCC-miRNA reporter can be introduced into wild-type or *Dgcr8*-deficient ES cells. *Dgcr8*-deficient ES cells can not produce mature canonical miRNAs as *Dgcr8* is required for the Drosha processing to convert pri-miRNAs to pre-miRNAs (Wang et al., 2007). If the ESCC-miRNA reporter repression is mediated by the endogenous miRNA, the reporter repression should be abolished in *Dgcr8*-deficient cells, thus the reporter should be expressed, giving rise to eGFP fluorescence.

A *Dgcr8*-deficient ES cell line has been previously generated by Matthew Davis (Davis, 2009). One allele of the *Dgcr8* in a clone (XH157) isolated from BayGenomics resource with the *Dgcr8* allele disrupted by gene trapping with the *SA-βgeo-pA* trap cassette inserted in intron 8 of the *Dgcr8* (*Dgcr8^{gt1}* allele). The second allele was disrupted by insertional targeting of a SA-Hygromycin-pA cassette into intron 4 of the *Dgcr8* locus and the resulting allele had a duplication for the genomic region from exon 2 to exon 6 (*Dgcr8^{tm1}* allele).

The *Dgcr8* expression in this cell line was validated by RT-PCR and Western blotting to detect the production of the endogenous mRNA and protein, respectively. RT-PCR products amplified using primers Ex2F and Ex5R were detected in both wild-type and the double-allele trapped *Dgcr8* (*Dgcr8^{gt1/tm1}*) clones (M1 and M2) as both endogenous and chimeric trapped alleles produce mRNA including transcripts derived from these upstream exons. However, RT-PCR product amplified using primers Ex9F and Ex12R was not detected in mutant clones, as this portion of the transcript can only be detected in wild-type *Dgcr8* mRNA. Therefore, endogenous mRNA was not present in the *Dgcr8^{gt1/tm1}* clones, Figure 5-11b. The endogenous protein level was assayed by Western Blotting and in *Dgcr8^{gt1/tm1}* cells a faint *Dgcr8* protein signal was detected, but the level of expression was significantly weaker than the wild-type cells, Figure 5-11c. Further analysis of the cell cycle was conducted to confirm the functional loss of the *Dgcr8*, as it was shown previously that an increased of G1 phase accumulation and a reduced proportion of S phase in *Dgcr8^{gt1/tm1}* ES cells compared to wild-type cells. The *Dgcr8^{gt1/tm1}* cell line showed a significant increase in the proportion of G1 phase population and reduced S phase population compared to the wild-type cells, Figure 5-11d. Illumina sequencing was conducted by Matthew Davis to examine all mature miRNAs in *Dgcr8^{gt1/tm1}* cells compared to wild-type cells and *Dgcr8^{gt1/tm1}* cells possessed a global reduction in canonical mature miRNAs (unpublished). Taken together, the *Dgcr8^{gt1/tm1}* mutation is hypomorphic; however, this cell line behaves as a *Dgcr8* loss-of-function mutant.

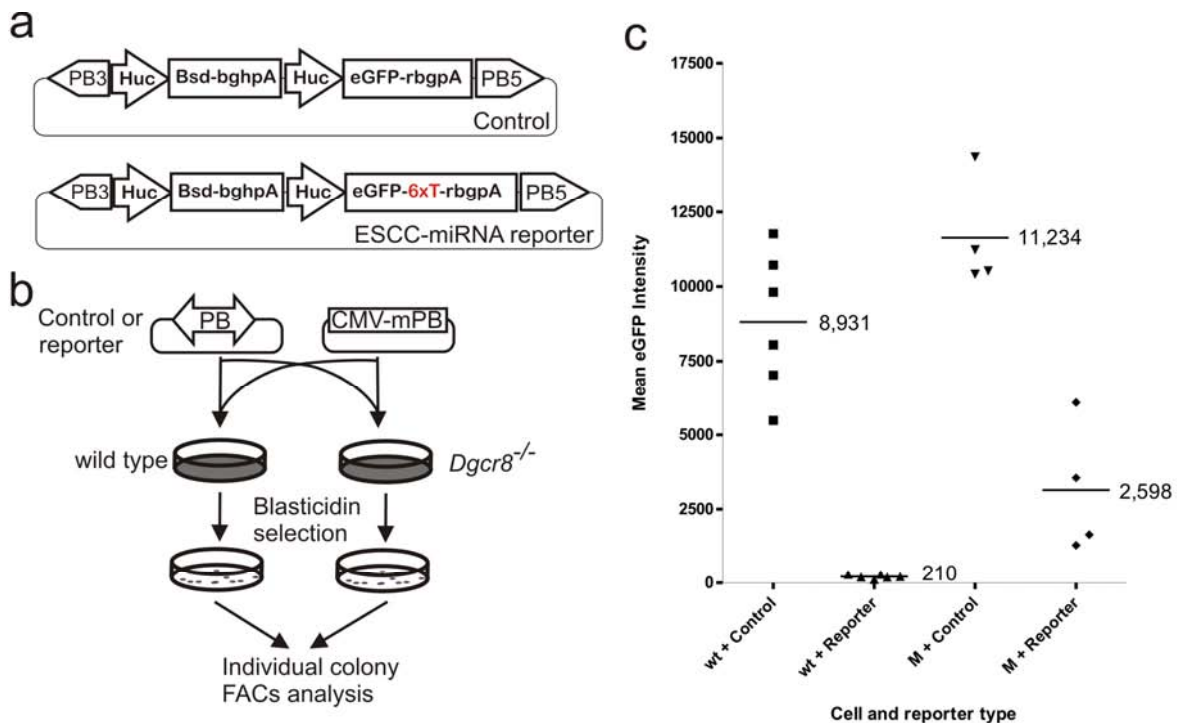
Figure 5-11: *Dgcr8*^{gt1/tm1} ES cell validation.

b, RT-PCR and c, western blotting to confirm the absence of endogenous *Dgcr8* mRNA and protein, respectively of the *Dgcr8*-deficient ES cell line (two clones M1 and M2). PCR with primers Ex2F and Ex5R can amplify products from both trapped transcript and the endogenous transcripts, whereas primers Ex9F and Ex12R can amplify only the wild-type transcript. c, cell-cycle analysis of the *Dgcr8*-null ES cells (bottom) compared to the wild-type cells (top). The percentage of cells in each phase of the cell cycle is shown in the histogram. The x-axis is the intensity of the propidium iodide (PI), which binds to the DNA and used to quantitatively measure the DNA content. The y-axis is the number of cells analysed and the total number of cells analysed is 10,000.

To confirm the causality of the endogenous miRNA mediated ESCC-miRNA reporter knockdown, the wild-type or *Dgcr8*^{gt1/tm1} ES cells were co-electroporation of with either the control or ESCC-miRNA reporter-containing PB transposons. The eGFP intensity was compared in Bsd-resistant clones, Figure 5-12. As observed previously, in colonies derived from wild-type cells (n=6), the ESCC-miRNA reporter express little eGFP, whereas in colonies derived from *Dgcr8*^{gt1/tm1} cells (n=4), an average over ten-fold eGFP expression was detected,

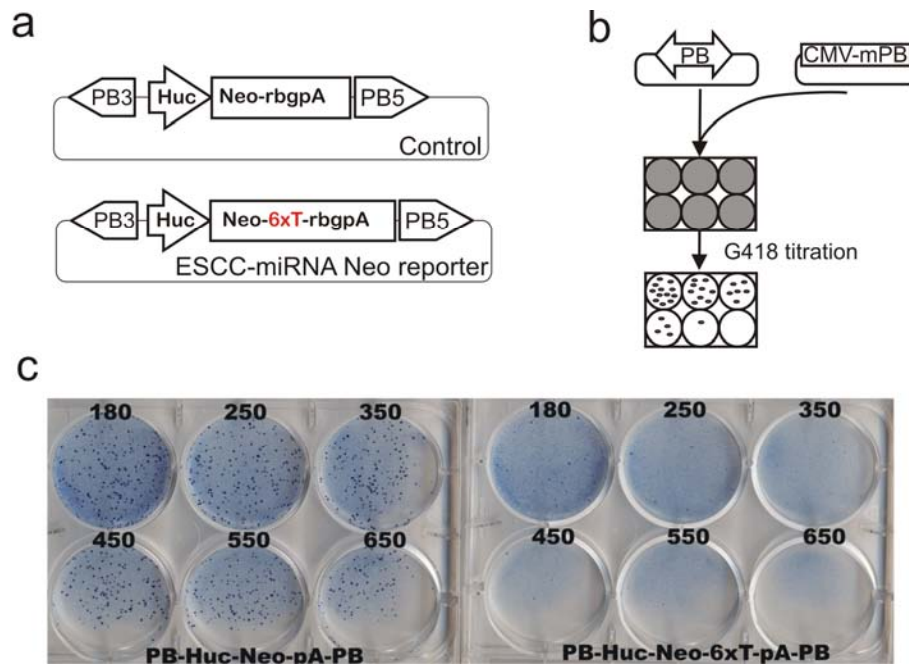
Figure 5-12c. The diverse levels of eGFP expression may be due to the locus specific influence on the eGFP reporter as the PB transposons integrate in different loci in all clones. However, the eGFP expression level of *Dgcr8*^{gt1/tm1} cells transfected with the ESCC-miRNA reporter was four fold less compared to the control reporter without the ESCC-miRNA target sequence multimer in the 3'UTR of the eGFP coding region. Taken together, these results suggest that the ESCC-miRNA reporter was knocked down by the endogenous ESCC miRNAs in ES cells. However, the ESCC-miRNA reporter expression is reduced due to the presence of the repetitive ESCC-miRNA target sequences present in its 3'UTR. Despite this reduction in expression level of the reporter, the ESCC-miRNA reporter may still provide sufficient expression for screening purposes.

Figure 5-12: Causality establishment of the ES cell miRNAs and the ESCC-miRNA reporter.



a, the ESCC-miRNA Neo reporter and control constructs. “6xT” highlighted in red is the six copies of the ESCC-miRNA recognition sites. b, PB transposon and transposase co-transfection scheme to deliver the ESCC reporter and control constructs to both wild type or *Dgcr8*^{gt1/tm1} cells. c, eGFP intensity measured in Bsd-resistant clones derived from the transfection. The median of each condition is shown on the diagram with a line presenting the median.

In order to test the feasibility of the ESCC-miRNA reporter can be differentially selected using a drug resistant cassette, a Neo reporter assay was also conducted to validate whether the ESCC-miRNA mediated reporter repression is selectable using G418. The assay was very similar to the eGFP-based system (Figure 5-10a) except that a Neo resistance gene was used in place of eGFP and the PGK-Neo cassette was deleted, Figure 5-13a. AB2.2 cells (1×10^7) were co-electroporated with 100 ng PB transposon-containing plasmids and the mPBase-expressing plasmid for the control and the reporter, independently. The electroporated cells were divided equally between six wells in a six-well culture plate. A G418 titration, ranging from 180 $\mu\text{g/ml}$ to 650 $\mu\text{g/ml}$, was conducted on these cells, with each G418 concentration tested in one well of the 6-wells, Figure 5-13b,c. The cells transfected with control plasmid were resistant to G418 at all concentrations tested. However, the cells transfected with ESCC-miRNA Neo reporter were more sensitive to G418 than control even at the lowest concentration tested, Figure 5-13c. Thus, an ESCC-miRNA Neo-resistant reporter can be distinguished using G418 selection and this provide a screening scheme using the ESCC-miRNA reporter strategy. The previous assay using the *Dgcr8*^{gt1/tm1} cells uncovered the reduced expression of the ESCC-miRNA reporter compared to the control reporter without the ESCC-miRNA target sequence in the 3'UTR, Figure 5-12. Therefore, further titration of the G418 concentration should be conducted in *Dgcr8*^{gt1/tm1} cells to identify the level of G418 resistance of the reporter containing the ESCC-miRNA target sites. However, an unrecyclable Neo cassette is already present in the *Dgcr8*^{gt1/tm1} cells, thus they are not suited for such an experiment. A *Dgcr8*-deficiency can be generated in the cell line containing the ESCC-miRNA reporter to further aid the titration of the G418 for the screening purposes.

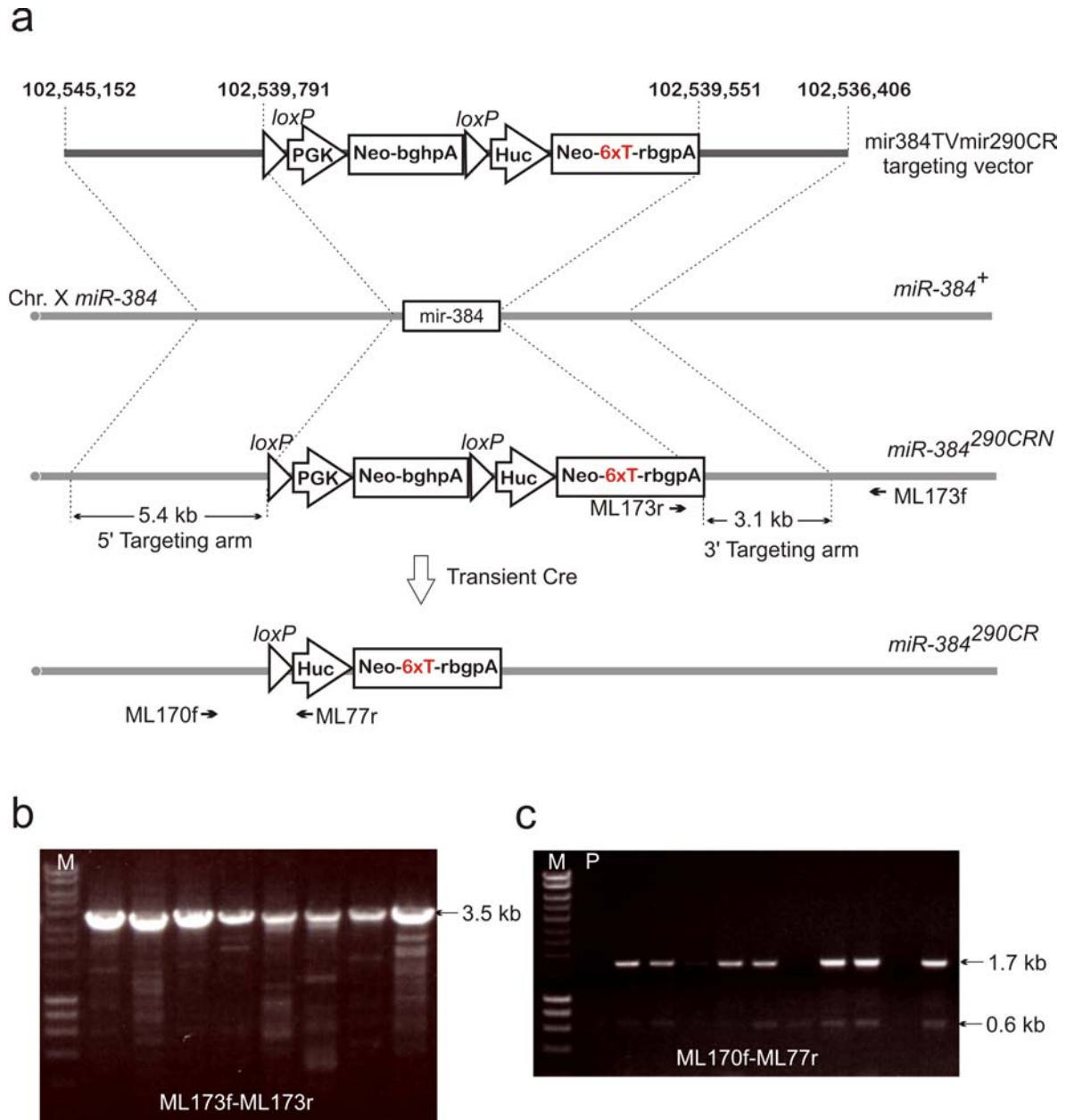
Figure 5-13: ESCC-miRNA Neo reporter analysis.

a, the ESCC-miRNA Neo reporter and control constructs. b, PB transposon and transposase co-electroporation scheme to deliver the ESCC reporter and control constructs using single-copy PB delivery conditions. Each electroporation was divided and plated equally in six wells of a 6-well plate and a G418 titration was conducted. c, The G418 titration. Left plate: control construct. Right: ESCC-miRNA Neo reporter constructs. The G418 concentration is indicated on top of each well ($\mu\text{g/ml}$).

2.2.2. Generation of an ESCC-miRNA target reporter knockin *Blm^{e/e}Hprt^{PBin2}* cell line

A stable reporter cell line was generated by targeting the ESCC-miRNA Neo reporter into the X-linked *mir-384* locus. The targeting vector, miR-384TVmir290CR, was constructed by introducing a *loxP*-flanked *PGK-Neo* cassette together with the ESCC-miRNA Neo reporter to the *mir-384* targeting vector by recombineering (Chapter 2 methods). The *PGK-Neo* cassette provides a selection scheme for the gene targeting event and it can be removed subsequently by transient Cre expression, Figure 5-14a. The targeting vector was linearised with *SbfI* and electroporated into 1×10^7 *BlmHprtPBin2* cells. The electroporated cells were selected with G418 and after seven days of selection, 48 colonies were picked and screened for correctly targeted events using long-range PCR (with primers ML173f and ML173r), Figure 5-14b. The targeting efficiency was 17 % in this experiment.

Eight metaphases of each correctly targeted clone were examined and four out of six clones had normal karyotype, with 40 chromosomes and the rest two clones had 39 chromosomes, probability due to the loss of Y chromosome. One karyotypically normal clone was chosen and 3×10^6 cells were electroporated with pCAG-Cre plasmid to pop-out *PGK-Neo* cassette. Three days post-electroporation, the transfected cells were replated at 1,000 cells per 90 mm plate in duplicates. One of the 90 mm plate was selected under 180 $\mu\text{g/ml}$ G418 and the other plate was cultured without drug selection. After eight days, colonies were picked from the plate without G418 selection. The pop-out positive clones (1.7 kb) were screened by genomic junction PCR using primers ML170f and ML77r. The cells without the *PGK-Neo* cassette popped out could not be amplified in this PCR condition, Figure 5-14c. The resulting reporter cell line will be tested to establish the minimal G418 concentration and can then be subsequently used for the screening.

Figure 5-14: The *miR-384* locus targeting with the ESCC-miRNA reporter construct.

a, Targeting vector, wild-type *miR-384* allele, targeted allele (*miR-384290^{CRN}*) and the pop-out allele (*miR-384290^{CR}*) structures. The genomic fragments used as targeting arms are indicated in the targeting vector with coordinates based on NCBI build37. b, long-range PCR confirmation of correctly targeted clones using primers ML173f and ML173r. c, PCR screening of PGK-Neo cassette popout, and the 1.7 kb product represents the popout events. The 0.6 kb product is a non-specific PCR product. P, parental cell line, *miR-384290^{CRN}*, without *Cre* transfection. M, Bio-rad Hyper ladder I marker.

3. Discussion

This chapter describes the establishment and validation of two reporter systems which enable genetic screens for the miRNA biogenesis and its downstream effector pathways. These two reporter systems complement each other in the types of the mutants that can potentially be identified. Both systems cover the common upstream miRNA processing pathway, i.e. components that play a role in pri-miRNA to pre-miRNA processing and subsequent transport from the nucleus to the cytoplasm. On the other hand, they can also probe novel components that are specific for each of the two miRNA effector pathways, i.e. the mRNA cleavage- or translational repression-mediated gene repression.

It is thought that the degree of the complementarity between the mature miRNA sequence and its target mRNA distinguishes between these two branches of the effector pathways and different Ago proteins are central to carry out the effector functions. As Ago2 is the only mammalian Ago protein out of the four homologues with the endonucleolytic (“slicing”) function, perfectly complementing miRNA and its mRNA target can induce mRNA cleavage by the Ago2-containing RISC complex. mRNAs recognised by imperfectly matched miRNAs, which are associated with Ago1, 3, and 4, can only mediate mRNA destabilisation and translational repression. However, Ago2 or Ago1, 3, and 4 are not exclusively associated with perfect and imperfect complementarity, respectively, between the miRNA and its target and it is not clear whether there is any novel component involved in guiding the choice. Furthermore, it is not yet known whether the two branches of the miRNA pathways recruit different components to conduct or regulate effector functions. In addition, downstream of RISC, little is known about how cells “treat” the miRNAs and whether the two types of effector pathway are differentially “treated”. Finally, the biogenesis and regulation of the newly discovered endo-siRNAs is only just starting to be revealed in mammalian systems (Tam et al., 2008; Watanabe et al., 2008). The resemblance of post pri-miRNA processing of the artificial miR-eGFP system to endo-siRNAs may provide some insights into the biogenesis and downstream effector pathway of this new class of endogenous small RNAs.

The establishment of sensitive reporter systems is very helpful to “translate” non-visible phenotypes to selectable schemes, thus these systems expand the possibilities of investigating many biological pathways in cell culture. When used in conjunction with the *Blm*-deficient ES cells to conduct recessive genetic screens, such reporter systems are also useful in overcoming a major limitation in the mixed pooling strategy to isolate a few pathway-related homozygous mutants from a vast number of irrelevant cells mixed within the same pool. This situation is almost like finding needles (relevant homozygous mutants) in a haystack (irrelevant cells in the pool), but with a magnet (a sensitive reporter system), the needles can be easily found.

Targeting rather than random integration of the reporter was adopted to introduce reporters into the ES cell genome. There are two main reasons for doing so. Firstly, reporters randomly integrated in different genomic loci may show different levels of expression due to the influence of locus-specific chromatin structures and methylation status. Moreover, random integration of linearised transgenes is prone to silencing. Thus, a single clone with a good expression level has to be established in order to avoid variation in the expression level; as such variations can give rise to a significant background level during the screening procedures. Secondly, random integrations almost always land in autosomal loci, as 38 out of the 40 chromosomes are autosomes. In *Blm*-deficient ES cells, the LOH rate is much higher than wild type cells, with approximately one cell in 2,000 at a specific locus losing its heterozygosity every generation. Therefore, randomly integrated reporters can be lost during culture expansion in the *Blm*-deficient background, leading to false positive and negative clones. Gene targeting provides a control over the expression uniformity. In addition, targeting a reporter into both alleles of an autosomal locus, or by targeting the reporter to X-linked loci, avoids LOH-mediated reporter loss. Recombineering technology allows the rapid construction of targeting vectors from BACs or modification of existing targeting vectors. Thus, introduction of defined genetic modifications in mouse ES cells by gene targeting is very simple experimentally and provides significant advantages for reporter expression.

An artificial miR-eGFP has been successfully generated and the miR-eGFP multimers were placed within the intron of a neomycin expression cassette. The main advantage of this architecture is to maintain miR-eGFP expressing cells can be maintained using G418 selection. Silencing of the miR-eGFP could introduce false positive hits in the screen, and thus can be eliminated in this scheme. In addition, by using G418 in conjunction with the blasticidin selection, mutants that are defective in the miR-eGFP transcription can also be excluded.

The miRNA-reporter polycistronic expression strategy described here is very useful for other applications *in vitro* and *in vivo*. The artificial or even endogenous miRNAs can be inserted into any transcription unit with tissue specific or temporally controlled promoters. *In vivo*, the expression of such miRNAs can also be monitored by incorporating reporters which provide fluorescence or luminescence readouts. Since the roles of many miRNAs are unknown, this expression strategy can provide a system to investigate miRNA function by over-expression or ectopic-expression.

During the generation of the miR-eGFP system, several factors have been tested, including promoters with different strengths, miRNA backbones, and copy numbers of the miRNA structures, in order to produce the maximum degree of repression. It was found that the target sequence or its secondary structure is the major factor influencing the repression efficiency. In addition, different combinations of conditions tested gave rise to a wide range of repression but none were able to provide complete repression of eGFP. Such a range in target repression may have biological significance. miRNAs may repress their targets with a range of efficiencies depending on the target sequences *in vivo* and subsequently produce a range of protein levels of the targets. This provides an additional dimension in regulating the biological systems, as different concentrations of the protein products may give rise to different phenotypic outcomes. This kind of biological phenomenon has been most extensively demonstrated in pattern formation with a morphogen gradient during *Drosophila* embryogenesis for example. This strategy may be used by miRNA mediated gene expression control for tuning phenotypes. Finally, this incomplete knockdown effect further highlights

the fact that shRNA-mediated knockdown is equivalent to a “hypomorphic” mutant, thus the phenotypic interpretation of these experiments has to be cautious.

The difficulty of achieving efficient knockdown with an artificial miRNA stimulated me to try an alternative approach in which endogenous miRNAs with the identical seed sequence was utilised to repress a reporter with the target sites engineered into the 3'UTR. It was noted that the ESCC-miRNA reporter with multiple ESCC miRNA target sequence in the 3'UTR showed a reduced expression compare to the otherwise identical reporter without these elements. This could due to the decreased stability of the messenger RNA and consequently a reduced protein production. However, the ESCC-miRNA reporter expression level is ten fold higher in cells lacking of the ESCC miRNAs compared to wild-type cells. This different may be sufficient enough to distinguish homozygous miRNA-pathway mutants from the irrelevant cells mixed within the pool. A cell line has been constructed which contains the neomycin resistant gene with the ESCC miRNA target sites in the 3'UTR targeted to an X-linked locus in the BlmHprtPBln2 cell line. Further G418 titration is required to identify the appropriate drug selection level for conducting the screen. The generation of a miRNA biogenesis mutant in this cell line can facilitate to achieve this goal.

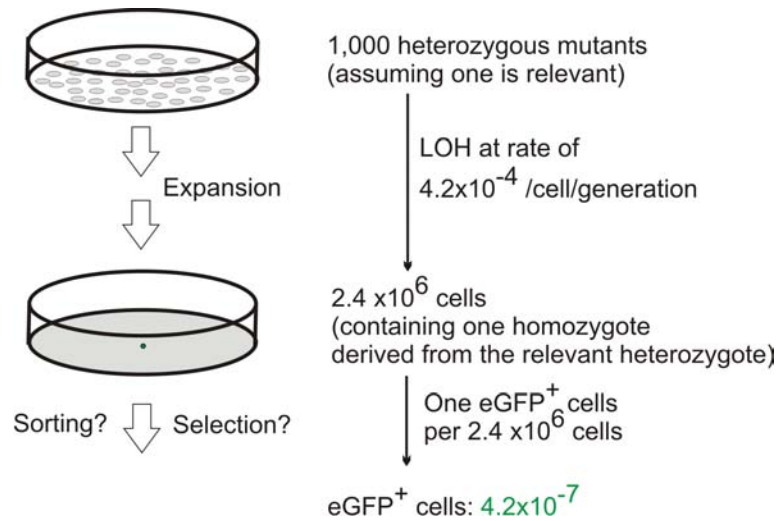
A reporter with a specific miRNA target site present in the 3'UTR has been used previously as a miRNA decoy to dampen the effect of miRNAs on their endogenous targets, a technology termed “miRNA sponges”, although such methods can only achieve a partial elevation of the expression of endogenous miRNA targets (Ebert et al., 2007). A recent finding suggests that these types of miRNA “sponges” exist naturally (Poliseno et al., 2010). Transcribed pseudogenes can act as natural “sponges”. This finding provides evidence that pseudogenes play active roles in regulating gene expression, challenging the conventional thoughts that pseudogenes are genomic junk. Furthermore, this also highlights the fact that there may be many novel avenues towards regulating miRNA-mediated target repression.

Chapter Six – Preliminary screen and condition optimisation for the miR-eGFP system

1. Introduction

As has been described in the previous chapter, an artificial miR-eGFP knockin cell line has been established, which contains a single copy of the mutagenic PB transposon knocked into the *Hprt* locus in *Blm*-deficient ES cells. ES cells with the miR-eGFP knockin express only 10 % of the eGFP level of cells without the miR-eGFP. ES cells with the miR-eGFP are sensitive to high concentrations (80 $\mu\text{g/ml}$ or above) blasticidin (Chapter 5, Section 5.2.1.5.). Thus the phenotypic screening can be conducted using either FACs sorting based on eGFP expression levels or by direct select from the mutant pools using high concentrations of blasticidin.

FACs sorting would be an ideal screening method, as not only is it a direct phenotypic readout of the mutants, but it is also able to distinguish different classes of mutants based on the levels eGFP de-repression. Sorted cells can be plated in low density and the single cell-derived colonies can be picked and further analysed individually. However, this system may not work very efficiently when coupled with a recessive genetic screen based on the *Blm*-deficient ES cell system, as the frequency of pathway-relevant homozygous mutants is too low for cell sorting, Figure 6-1. Assuming that within 1,000 heterozygous mutants generated by PB transposon-mediated mutagenesis, one mutant is relevant to the miRNA biogenesis pathway. The rate of LOH in *Blm*-deficient ES cells is estimated to be 4.2×10^{-4} per cell per generation at a specific locus. Based on conservative estimates, the mutant pool needs to be expanded to 2.4×10^6 cells to obtain a single homozygous mutant cell from one of the original heterozygous mutants. This means that the percentage of relevant mutants within the pool is 4.2×10^{-7} . This proportion of eGFP positive mutants is too low to be sorted successfully. However, if the mutants are enriched in a preliminary round of selection, a subsequent round of sorting may aid mutant isolation from the background. Therefore, using blasticidin selection as the primary screening scheme is a good choice.

Figure 6-1: Estimate of the proportion of homozygous mutants in the miRNA pathway.

A detailed explanation is described in the main text.

In this chapter, different screening protocols have been tested for the isolation of homozygous mutants that are deficient in processing the miR-eGFP reporter. A known gene connected to the miRNA pathway, *Ago2*, has been identified during the preliminary screen and the validation of this mutant was also described.

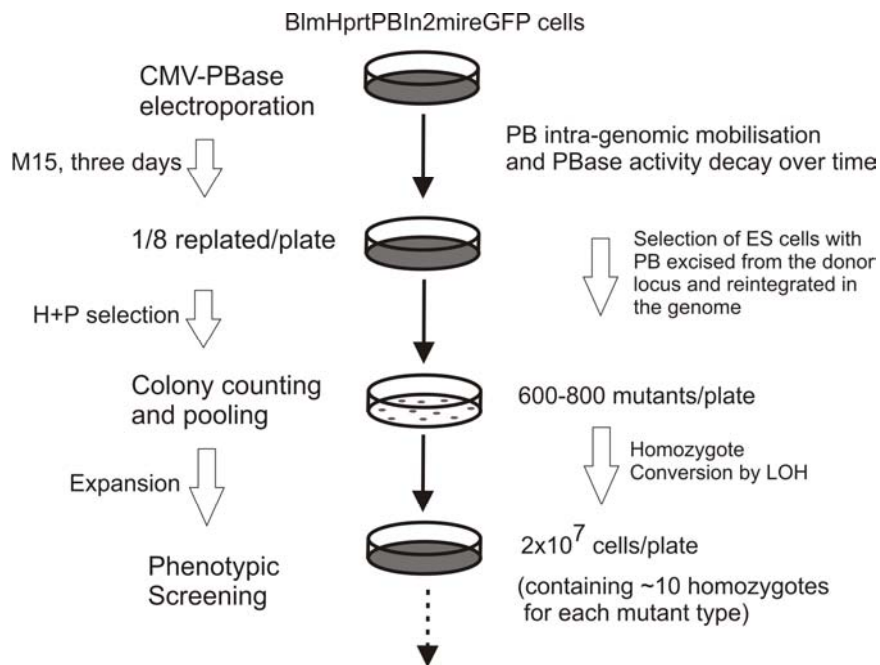
2. Results

2.1. Generation of the mutant library

Figure 6-2 shows a schematic representation of mutant library generation. BlmHprtPBIn2miR-eGFP cells (1×10^7) were electroporated with 25 μ g of a CMV-HyPBase expression plasmid and were divided and plated equally on three 90 mm plates. After three days of growth in M15, the cells from each 90 mm plate were trypsinised and an eighth of the cells were plated onto a fresh 90 mm plate. The next day, HAT and puromycin containing M15 was added to the plates and the selection continued for eight days until visible colonies formed. With this condition, around 600 to 800 colonies can be formed per 90 mm plate. The replating step conducted after electroporation with PBase was designed to allow transposition to proceed to completion before initiating selection. This disperses cells with different genotypes within a colony, so that the complexity of the mutant pool can be correctly estimated.

In total, ten electroporations were conducted and 30x 90 mm plates were selected to give rise to a total of 20,000 HAT and Puro double-resistant colonies. After HAT and puromycin selection, the colonies were recovered in M15 with HT supplemented for two days before the colonies were counted, trypanised to form single cell suspension and placed onto new 90 mm plates (one plate to one passage). When the 90 mm plates were confluent (containing about 2×10^7 cells/90 mm plate), the culture should contain at least a few homozygous mutants per locus. In this procedure, the mutant library was sectored into 30 sub-libraries which were propagated independently, this mutants arose from different sub-libraries should be derived from independent heterozygous mutants. This should avoid the daughter cells of a dominant clone to populate the entire library, making the identification of other mutants difficult. Therefore in a typical screen, a few sub-libraries should contain colonies which are derived from the same homozygous mutant, while most of the sub-libraries do not contain any mutants.

Figure 6-2: Schematic representation of the mutant library construction.

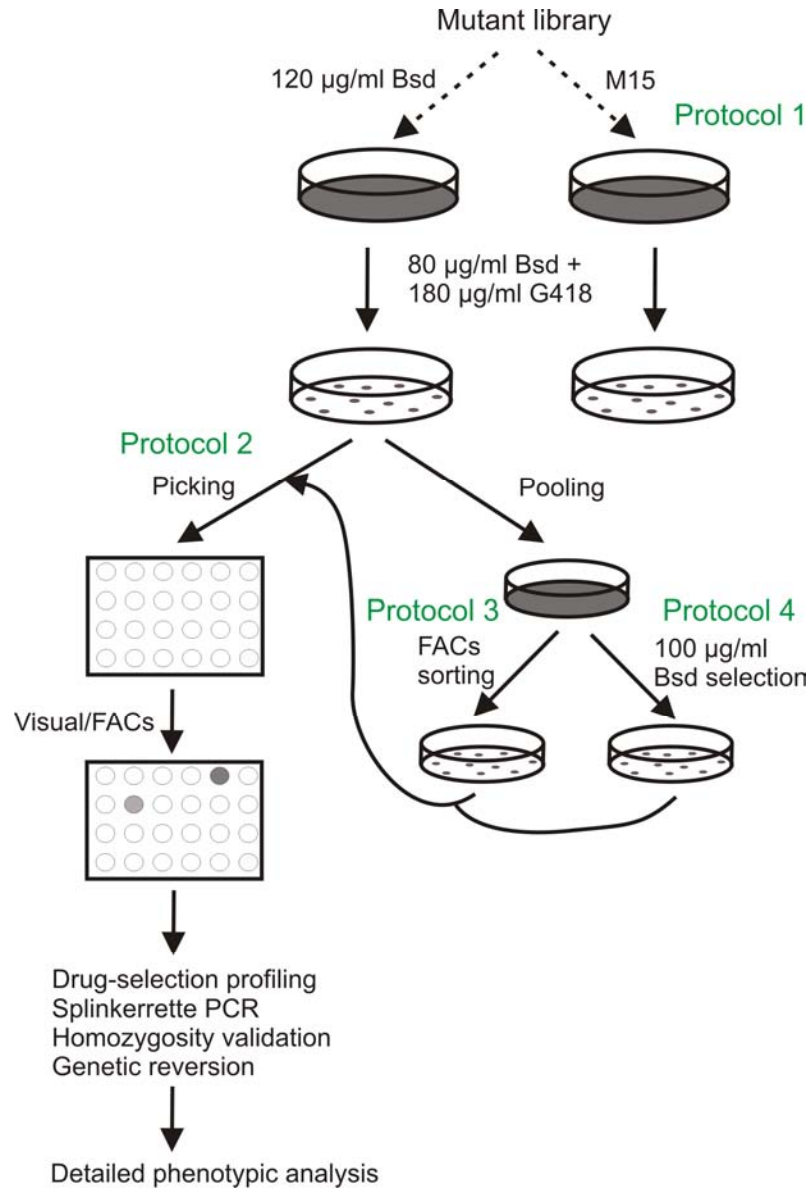


The flow chart to the left of the diagram describes the experimental procedures used for library generation; the flow chart to the right of the diagram describes the events occurring in the cell culture for each step during the library construction.

2.2. Screening strategies and optimisation

Four protocols were investigated for the phenotypic screening strategy, Figure 6-3.

Figure 6-3: Screening strategies for the miR-eGFP system.



In total four protocols were tested and a detailed description is given in the main text below.

The first protocol (Protocol 1) was conducted by directly passaging the mutant cells from the library to 150 mm culture plates with a monolayer of blasticidin and G418 resistant feeders. The next day, 80 $\mu\text{g/ml}$ blasticidin and 180 $\mu\text{g/ml}$ G418 containing M15 was applied to the cells and the selection was maintained for eight days. Many cells survived the selection and each plate was covered with approximately a thousand colonies at the end of the selection and there was no difference observed among the plates. This might have been caused by the high cell density and the inefficiency of Bsd-mediated killing of cells with residual 10% eGFP-IRES-Bsd expression. Therefore, protocol 1 was not sufficient for the relevant mutant selection.

In order to enhance the selection sensitivity, the cells from the established mutant library were passaged onto 150 mm plates with M15 containing 120 $\mu\text{g/ml}$ blasticidin. This concentration of blasticidin was tested during the cell line validation analysis (Chapter 5 Section 5.2.1.5.). Under this selection, the cells without the miR-eGFP can survive albeit a slight reduced growth rate for up to six days. The next day, 80 $\mu\text{g/ml}$ blasticidin and 180 $\mu\text{g/ml}$ G418 dual-selection was initiated and maintained for eight days. Under these conditions, there were approximately 50 ~ 300 colonies formed on each plate with some plates containing significantly more colonies than others, indicating that there may be mutants present. The majority of the plates with low number of colonies may reflect the background of this selection scheme.

Three protocols were used after this initial blasticidin selection. In Protocol 2, twelve individual colonies were directly picked from each mutant pool and in total 360 colonies were picked from the whole library and examined for eGFP intensity either manually under the fluorescent microscope or by high-throughput FACs analysis using 96 well plates. In Protocol 3 and 4, the colonies generated from the initial blasticidin selection were first pooled with three other pools to form ten super-pools. These ten pools were passaged twice and were either FACs sorted (Protocol 3), or plated onto 90 mm culture plates for a second round of blasticidin selection with slightly elevated blasticidin concentration, 100 $\mu\text{g/ml}$. The colonies

which resulted from these protocols were picked and subjected to eGFP intensity analysis as described previously.

In total, 13 eGFP-positive colonies were isolated from all three protocols, they were subjected to drug-selection profiling using HAT, puromycin and G418 to confirm their phenotype. The Splinkerette PCR was then conducted on the clones with the correct drug-selection profiles to identify the PB transposon insertion sites. Upon locus identification, the homozygosity status of the locus and further detailed phenotypic analysis could proceed.

The results of Protocol 2, 3 and 4 are shown in Table 6-1. Using protocol 2, only one out of the 360 colonies picked showed eGFP expression. By FACs analysis, the eGFP expression level in this mutant was equal to the eGFP level in the cells without miR-eGFP reporter. Splinkerette PCR indentified the PB transposon was inserted into *Ago2*, which is known for its involvement in miRNA/siRNA-mediated gene silencing. The detailed validation of this mutant is described in the next section of this chapter.

When protocols 3 and 4 were used, 12 eGFP positive colonies were identified and confirmed by FACs analysis. These colonies were derived from three main pools, thus they may be daughter clones originating from single mutant genotypes. The *Ago2* mutant was identified from pool No. 5 using Protocol 2, the eGFP positive clones obtained from pool No.13/14/15 were not *Ago2* mutants. In total from both protocols, three clones were derived from pool No.1/2/3, four clones from pool No. 13/14/15 and five clones from pool No.7/8/9 and one clone from pool No. 25/26/27.

These 12 clones were examined for their resistant or sensitivity to puromycin, G418, and HAT to further confirm their genotypes. Puromycin resistance confirms the presence of the PB transposon within the ES cell genome and all the clones analysed were puromycin resistant. The sensitivity to G418 suggests the loss or silencing of the miR-eGFP, which can give rise to elevated eGFP expression, thus, clones with this phenotype will not be true mutants. Four

clones were identified as being G418 sensitive (two clones from pool No.1/2/3 and one clone from pool No. 25/26/27), Table 6-2.

HAT resistance confirms the excision of PB transposon excision from the donor site. However, spontaneous mutation in the *Hprt* locus after PB excision from the locus can also give rise to a HAT sensitive phenotype. The nine clones from two pools (four from pool No. 13/14/15, and five from No. 7/8/9) were HAT sensitive. The Splinkerette PCR was performed on 36 clones including all the eGFP-positive clones and randomly selected eGFP-negative background clones. For 33 clones including all eGFP-positive clones, the PCR products were mapped at the donor *Hprt* locus. The four clones consisted of two independent PB integration sites with two eGFP-negative sister clones (Clone 22 and 23) and two eGFP-positive sister clones (Clone 2 and 3). The PB integration in Clone 2 and 3 was mapped to X chromosome 3' of the *Hprt* coding region with coordinate 50,380,002 337 (NCBI Build37), Table 6-2, while the PB integration in Clone 22 and 23 was mapped to a intergenic region on X chromosome with coordinate 116,735,337 (NCBI Build37).

Southern blotting was conducted on all these clones to detect the pattern of PB integrations. A PB5'TIR probe was used to detect PB intra-genomic mobilisation, Figure 6-4. The southern blotting results confirmed the Splinkerette PCR result, with most of the clones showing a band pattern identical to the control sample, the cell line before PB mobilisation. The only four clones with different band patterns matched the four eGFP-negative clones with PB away from the *Hprt* donor site, identified from the Splinkerette PCR. This data suggests that there were a significant proportion of cells present in the mutant pools without the mobilisation. Their survival under HAT selection during the library construction was most likely due to the known strong cross-rescue effect among cells close by. Therefore, the clones that were HAT sensitive but their eGFP-positive expression may be due to spontaneous mutations within the miRNA processing pathways.

Table 6-1: Efficiency comparison between screening protocols.

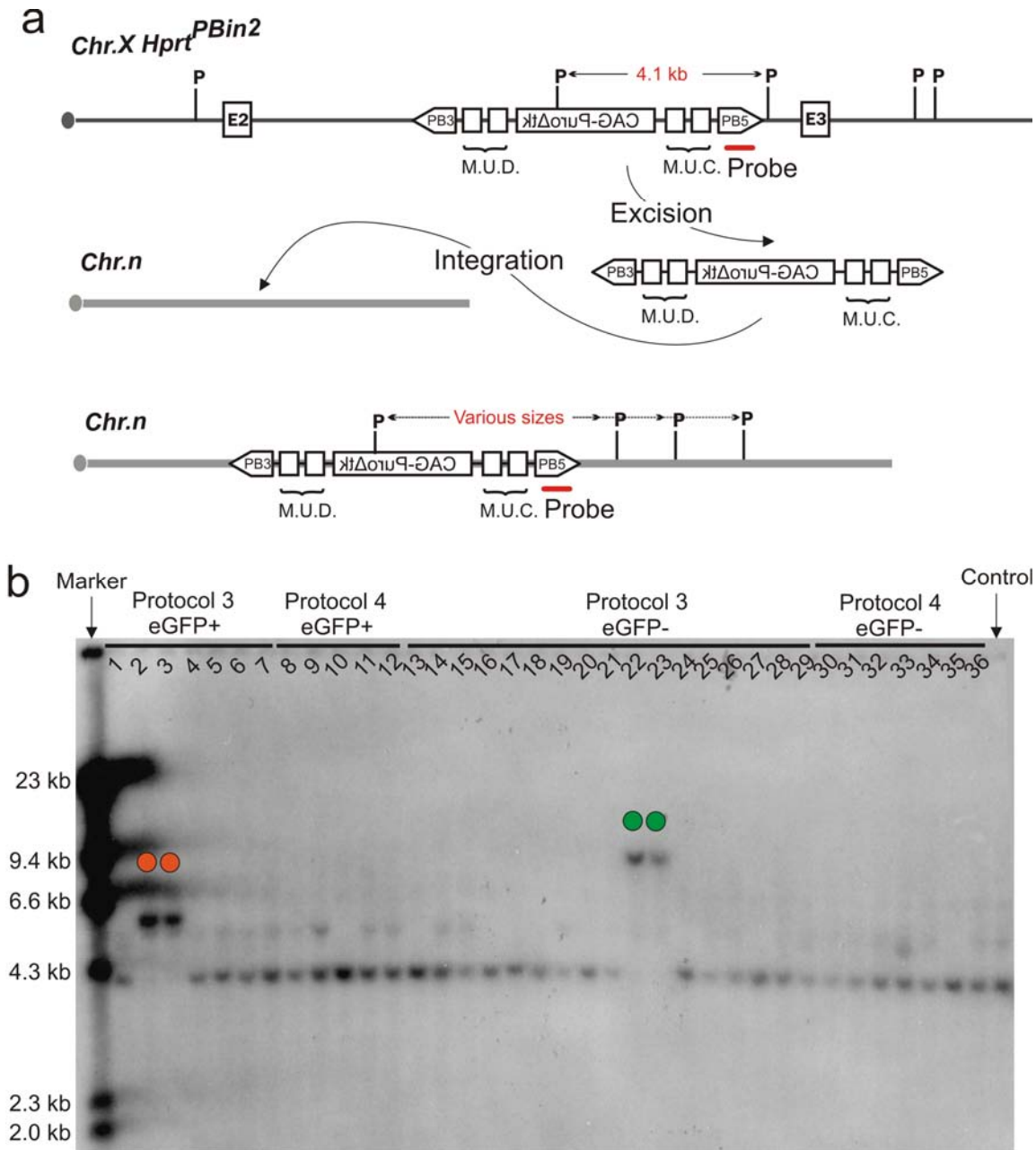
	Protocol 2	Protocol 3	Protocol 4
No. colonies	~ 6,000	~ 400	~ 60
No. picked	360	170	50
No. eGFP+	1	7	5

Table 6-2: Summary of the identities of eGFP-positive clones.

	Protocol 2	Protocol 3	Protocol 4
Total no. eGFP+	1	7	5
Clonal relation*	Pool No.13	A, 2 from pool No.1/2/3 B, 4 from pool No.13/14/15 C, 1 from pool No.25/26/27	B, 5 from pool No.7/8/9
Drug-profile	HAT ^R , Puro ^R , G418 ^R	Type A: HAT ^R , Puro ^R , G418 ^S Clone 2 and 3 Type B: HAT ^S , Puro ^R , G418 ^R Clone 4,5,6 and 7 Type C: HAT ^R , Puro ^R , G418 ^S Clone 1	Type A, HAT ^R , Puro ^R , G418 ^S Clone 8 Type B, HAT ^S , Puro ^R , G418 ^R Clone 9, 10, 11, and 12
Splinkerette PCR (Candidate locus)	<i>Ago2</i>	Type A: X chr.: 50,380,002 Others were mapped to the donor site [#] .	All were mapped to the donor site.
Comments	Intron 1 of <i>Ago2</i>	Type A, X chr.: 50,380,002 3' downstream <i>Hprt</i> [#] .	N/A

*: clonal relation between different clones isolated from the same pool was classified based on the G418, puromycin (Puro) and HAT resistant phenotype A, B, and C. The clone I.D. was followed by the drug-resistant phenotype and is consistent with the clones shown in the Southern blot in Figure 6-4b. #: the two sister eGFP-positive clones (Clone 2 and 3) were mapped to a genomic location 3' to the *Hprt* coding region and the genomic coordinate is based on NCBI Build37.

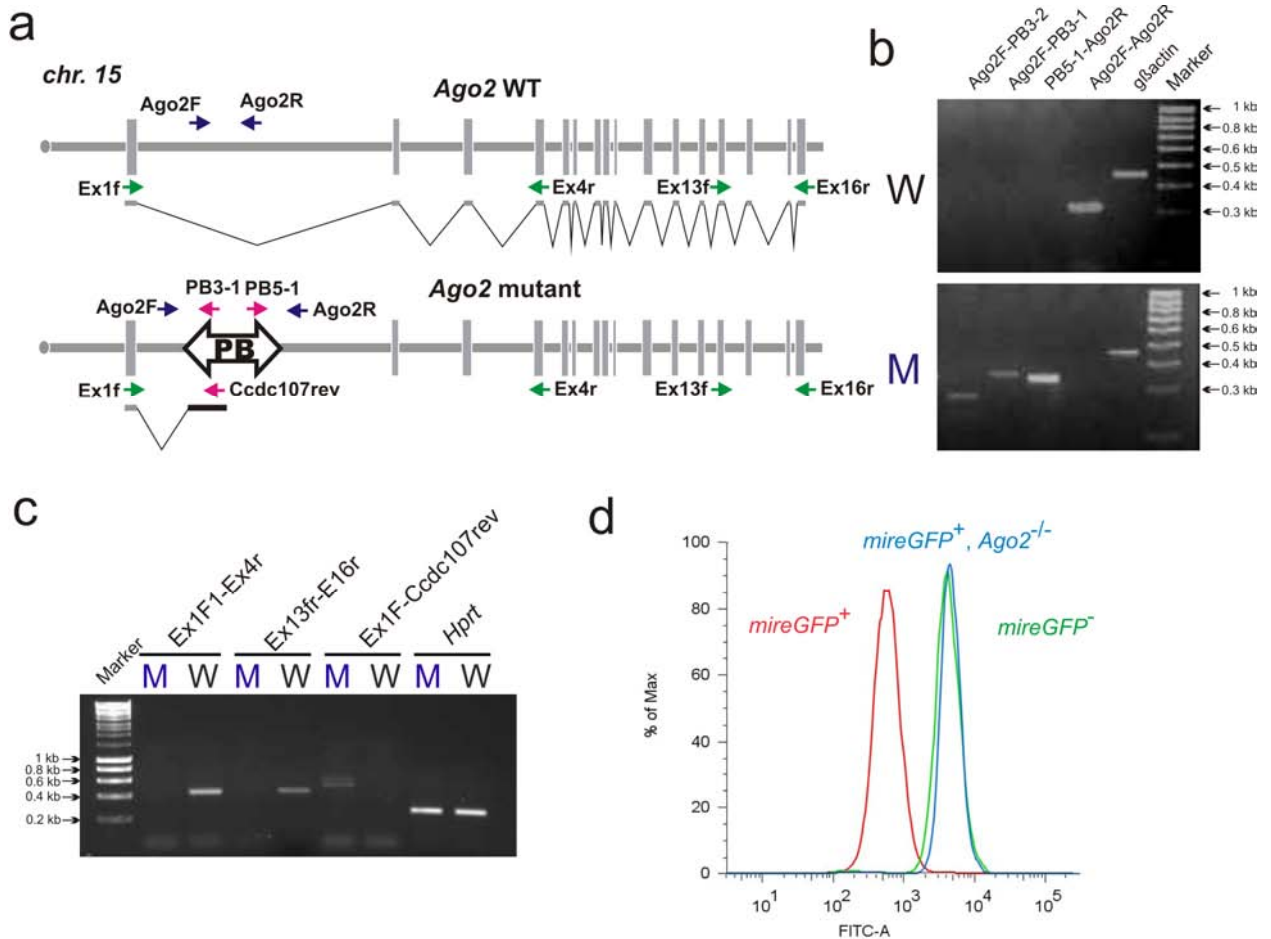
Figure 6-4: Southern detection for the PB transposon reintegration patterns.



a, Southern detection scheme for the PB integration sites. When the PB transposon was located in the *Hprt* locus, the Southern blot gave rise to a band with 4.1 kb in size when the genomic DNA was digested with *Pst*I. When the PB transposon re-integrated elsewhere in the genome, the band pattern will alter depending on the sequences of the specific integration sites. P, *Pst*I site. b, Southern blotting results of all eGFP positive clones identified from Protocol 3 and 4 and some random clones picked from the pool. Control, DNA extracted from ES cells without PB mobilisation. Only the clones with red (two sister clones from pool No.1/2/3) or green (two sister clones from pool No.19/20/21, randomly selected) dots indicated showed PB reintegration events, and the two red marked clones were derived from the same pool, and the green clones from another pool.

2.3. Mutant validation

The *Ago2* mutant identified from Protocol 2 was further analysed, Figure 6-5. The mutagenic PB transposon was inserted in intron 1 of the *Ago2* gene. The *Ago2* inactivation was mediated by the *Ccdc107* mutagen unit, based on the orientation of the PB integration and confirmed by the fusion RT-PCR product detected in the mutant clone, Figure 6-5a,c. Genomic PCRs were conducted with primers specific to the locus and PB-specific primers to validate the homozygosity status of this mutant, Figure 6-5b. In the mutant, the wild-type PCR product generated by two locus specific primers spanning the PB integration site was completely absent; whereas PCR products were produced when the PB transposon-specific primers and the locus-specific primers near the integration site were used. The opposite was observed in wild-type cells. Therefore, the mutant must be homozygous with both alleles inactivated by the mutagenic PB transposons. RT-PCR analysis on the mutant confirmed a null mutant phenotype, as no wild-type transcript was detected, when both 5' and 3' end of the transcripts were analysed. The fusion transcript detected in the mutant showed the correct fragment size as predicted. The doublet band-pattern indicates the alternative splicing in the intron of the *Ccdc107* mutagen between the two exons as observed previously (Chapter 3). Finally, the eGFP expression profile showed that the mutant had an elevated eGFP level equivalent to the reporter cell line without the miR-eGFP, Figure 6-5d. Taken together, the *Ago2* mutant clone found is truly a homozygous null mutant with the defective miR-eGFP-mediated eGFP silencing.

Figure 6-5: *Ago2* mutant molecular analysis.

a, PB integration site within the *Ago2* locus and the predicted trap-based gene inactivation. b, genomic-PCR validates the homozygosity status of this clone. W, wild type control, M, *Ago2* mutant. The absence of the wild-type PCR product suggests that *Ago2* mutant is homozygous. c, RT-PCR analysis of the *Ago2* transcript in mutant and wild-type cells. d, FACS profile on the eGFP expression of the *Ago2* mutant. The positive control was ES cells without miR-eGFP targeted and the negative control is the cells with miR-eGFP targeted but without PB mobilised.

3. Discussion

3.1. Future improvement on the miR-eGFP screening system

This chapter describes a preliminary screen using the miR-eGFP system and various experimental strategies to optimise the screening procedure. A homozygous mutant *Ago2*, a key component in the miRNA/siRNA-mediated silencing was identified from this initial attempt. The ability to identify a known gene in the pathway confirmed the success of the design and performance of the recessive-screening strategy. Further experiments can be conducted using the established protocols to identify novel components in the pathway. However, during this preliminary study, two existing problems have also been identified. These outstanding problems could be a reason why only one hit was found. These problems will be addressed before subsequent rounds of screening are conducted to increase the capacity of novel mutant identification.

Firstly, the mutant library contained a limited representation of heterozygous mutants which was not correlated with the estimated complexity. The main cause was likely to be the post-electroporation replating and selection scheme which favoured the survival of cells without PB transposon excised from the donor locus. Although such cells should be sensitive to HAT, they can survive by sharing the metabolites generated by surrounding *Hprt* proficient cells in the mixed population that had PB excised from the *Hprt* locus. One way to avoid this would be to initiate HAT selection directly after the PBase-electroporation without the replating to avoid such an issue.

Secondly, PB excision from the donor site did not seem to be very efficient, judging by the total colony number obtained after selection compared to previous data obtained using the identical locus and transposon. In addition, the two integration sites mapped in clones isolated after blasticidin selection were both having PB transposons integrated on X chromosome near the donor site. Thus, in this preliminary screen, genome-wide mutagenesis was not achieved to a satisfactory level. The reason for inefficient PB intra-chromosomal mobilisation of PB was not clear, although *Blm* deficiency may negatively affect the repair of the double-strand breaks of the genomic DNA after PB excision. Local-hopping of the PB

transposon means that more HAT and puromycin resistant mutants need to be generated in order to achieve genome-wide coverage compared to mutagenic strategies with no local-hopping effect. Therefore with the current system, the degree of local hopping should be estimated in order to generate a large enough library to provide a good coverage of the genome. Alternatively, the plasmid-to-genome mobilisation method may be reconsidered for generating the library.

Apart from the limitation posed by the mutagenesis coverage, the mixed-mutant pooling strategy used for the *Blm*-deficient ES cell system has an intrinsic limitation in isolating mutants that have growth defects. The LOH rate estimation was based on the assumption that homozygous mutants of interest possess the same growth rate as other cells present in the pool. However, mutants with reduced growth rates will expand slower than other cells in the pool, thus the number of homozygous mutant cells from such clones within the pool would be less than expected. Overtime, the proportion of this type of homozygous mutant will drop significantly as they are out-competed by others. Both miRNA-biogenesis mutants *Dgcr8* and *Dicer* show retarded growth and due to the loss of ESCC miRNAs, described in Chapter One (Kanellopoulou et al., 2005; Wang et al., 2007). Therefore, in this screen, mutants like *Dgcr8* and *Dicer* may not be easily identified, but this should still be possible, since these cells are viable and can proliferate. If LOH of these mutants occurred early during expansion, they could still contribute a decent proportion within the pool. In further screens, enrichment strategies should be conducted as early as possible to reduce the risk of such mutants being out-competed. Cell-cycle rate compensation of the mutants may also be possible by introducing transgenic cell cycle effectors that are regulated by ESCC miRNAs or parallel to ESCC miRNA regulation. Thus, upon loss of miRNAs in the mutants, the cell-cycle defect might be rescued by the exogenous expression of the introduced gene to compensate the effect downstream of ESCC miRNA regulation. For example, genes such as *C-myc* may partly compensate the loss of ESCC miRNAs since *C-myc* has been shown to be able to transcriptionally repress the cell cycle inhibitor *Cdk1na*, which is a major target of ESCC miRNAs which reduces the G1-S cell-cycle transition (Seoane et al., 2002).

Screening protocols 2, 3 and 4 were able to identify eGFP-positive cells. Despite the low efficiency of obtaining eGFP-positive clones in Protocol 2, this is the only protocol which produced a genuine hit. The failure to identify the same hits in Protocols 3 and 4 may be due to the fact the proportion of the mutant in the sub-pool may be very limited, and the further pooling of sub-pools would have reduced the proportion of this mutant and it may have been lost during passages before the pools were subjected to secondary enrichment. The ability to isolate eGFP-positive clones using protocols 3 and 4 suggest that these methods are capable of isolating potential mutants. The inability to detect sister clones from the same clonal origin may reflect the fact that there were no genuine mutants present in these pools and the false positive mutants identified may have arisen after the culture was split. The use of FACs sorting to enrich mutants in Protocol 3 may be potentially less “invasive” to cell growth than use of high concentration of blasticidin. Therefore, in the future screening, the combined use of protocols 2 and 3 could be sufficient for mutant isolation and enrichment, with minimal passages before the FACs sorting in protocol 3.

3.2. *Ago2*, Argonaute proteins and the small RNA mediated pathways

The homozygous mutant identified in this preliminary screen was *Ago2*, also known as *Eif2c2*. *Ago2* is one of the mammalian Argonaute proteins playing key roles in small RNA-mediated regulatory pathways that modulate gene expression, chromosome structure and function, and provide a defense mechanism to silence invading viruses and transposons. A *Blm*-deficient ES cell based recessive screen for siRNA processing pathway also identified *Ago2* as a key player in siRNA-mediated gene silencing (Trombly et al., 2009). In that screen, multiple but independent hits within intron 1 of *Ago2* were identified, highlighting the high degree of bias for retrovirus integrations within certain regions of the mammalian genome.

The Ago family can be sub-divided into three clades, the Piwi clade, which are closely related to *D. melanogaster* PIWI (P-element induced wimpy testis), the Wago clade, which are specific to *C. elegans* and the Ago clade, which are similar to *Arabidopsis thaliana* AGO1 (Hutvagner and Simard, 2008). The piwi clade is not present in plant, and plays a role as part of the innate immune system to silence mobile genetic elements in the nucleus. The Ago-

clade members are found in both plant and animal species and are effectors in small RNA-mediated gene regulation.

The protein structures of Argonautes are well conserved, consisting of four distinct domains, the N-terminal, PAZ, Mid, and PIWI domains. The PAZ and Mid domains facilitate the anchoring of the small RNA guide, with PAZ binding the 3' end using a series of conserved aromatic residues and the Mid domain providing a binding pocket for the 5' end. In addition, the Mid domain of metazoan Argonautes that function in the miRNA pathway contain a motif known as the MC domain, which has homology to the cap structure binding motif of the translation initiation factor eIF4E, and it may function to interfere with translation (Kiriakidou et al., 2007). The PIWI domain contains an RNase-H-like fold, which evolved to use ssRNA as a template to target RNA with highly complementary sequence. Ago brings the scissile phosphate, opposite to the 10th and 11th nucleotides of the small RNA guide, into the enzyme active site of the PIWI domain to conduct the RNA cleavage, leaving 5' P and 3' OH termini. In mammals, Ago2 is the only Ago members out of four members Ago1-4, to maintain the endonuclease (slicer) activity. Other Ago-clade members have lost their cleavage activity during evolution.

A few miRNAs have been found to use cleavage-mediated gene repression in mammals (Yekta et al., 2004), thus Ago2 is important in mediating miRNA function. However, in mammals, the majority of the miRNA-mediated gene repression uses translational repression to regulate target gene expression. Although this mode of gene regulation does not require the slicer activity of the Ago protein, several mature miRNAs which use the translational repression-mediated effector pathway showed a significant decrease in their steady-state expression level in the absence of Ago2 (Diederichs and Haber, 2007; Kaneda et al., 2009). This suggests that Ago2 may also play an active role in miRNA biogenesis or stabilisation of mature miRNAs.

A novel miRNA intermediate (ac-pre-miRNA) structure of several miRNAs was discovered which was shown to be an Ago2 cleaved pre-miRNA hairpin, and it can be subsequently processed by Dicer complex (Diederichs and Haber, 2007). Whether this intermediate is a by-

product generated during Ago2-Dicer complex association or has biological significance is unclear. This novel Ago2 processing function may be independent from the RISC-mediated effector action, aiding Dicer cleavage to enhance miRNA biogenesis. This may explain the drop in the steady-state level of some miRNAs observed when Ago2 is absent. Processing with Ago2 may also be biologically significant. So far one miRNA, mir-451, has been found to rely purely on the Ago2-dependent processing to generate a structure similar to the ac-pre-miRNA and can bypass the Dicer step in its maturation (Cheloufi et al., 2010).

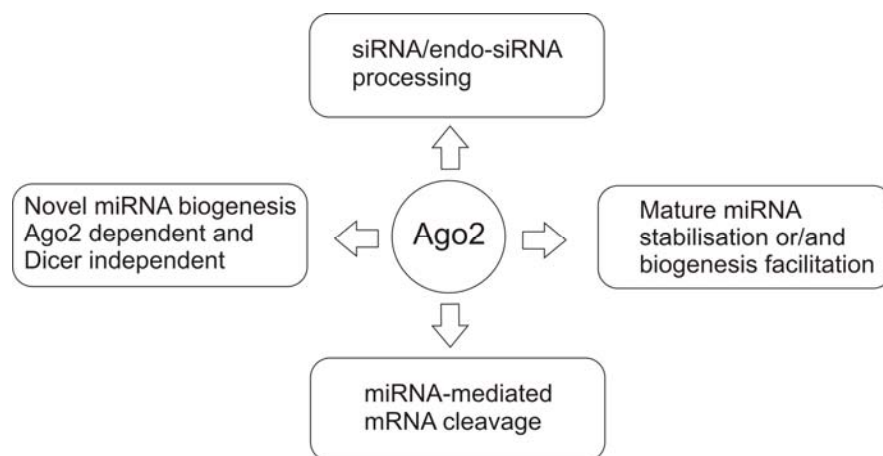
In plants and *C. elegans*, some Argonaute proteins that are involved in both exogenous and endogenous RNAi pathways can mediate secondary siRNA production, which is important in maintaining and propagating the RNAi response (Yigit *et al.*, 2006; Hutvagner and Simard, 2008). Once a *de novo* dsRNA duplex is generated, the siRNA can serve as the primer to generate new siRNAs (the secondary siRNAs) by RNA-dependent RNA polymerases (RdRPs). These products are then either cleaved by Dicer or other RNase-III-like enzymes and are subsequently loaded onto catalytic or non-catalytic Ago proteins to induce another level of gene silencing. In mammals, it is thought that such an endogenous RNAi response is not permissive due to the long-dsRNA induced interferon response, which leads to global translation termination, RNaseL induction and apoptosis. In addition, the mammalian homologue RdRP has not been found at this time.

Recently there have been two breakthroughs in proving the presence of endogenous RNAi in mammals. Firstly, mammalian endo-siRNAs have been identified in oocytes and ES cells (Babiarz et al., 2008; Tam et al., 2008; Watanabe et al., 2008), suggesting that the endogenous RNAi pathway is indeed present in mammals. The presence of endo-siRNAs in oocytes and ES cells is attributed to the lack of interferon responses in these cells. Oocytes with a knockout (both maternal and zygotic) of either *Dicer* or *Ago2* show severe defects in oocyte meiosis I and II and the global mRNA expression varied significantly from the wild-type situation (Murchison *et al.*, 2007; Kaneda *et al.*, 2009). Furthermore, knocking down Ago2 but not Ago3 or 4 in the two-cell stage mouse embryos can lead to developmental arrest (Lykke-Andersen et al., 2008). Finally, oocytes with a *Dgcr8* knockout are phenotypically normal

without any global mRNA expression change (Suh et al., 2010). These data highlight the essential roles endo-siRNAs play in mammalian oogenesis and early zygotic development. Another seminal discovery is the identification of mammalian RdRP equivalent, formed by the telomerase reverse transcriptase catalytic subunit (TERT) and the RNA component of mitochondrial RNA processing endoribonuclease (RMRP) (Maida et al., 2009). Therefore, mammals do possess the endogenous siRNA pathways and the understanding of endo-siRNA biogenesis and its effector pathways as well as their functions in mammalian early development is a very exciting area of research.

In summary, being a unique Ago-clade member with endonuclease activity, Ago2 is involved in many areas of small RNA biogenesis and regulation. Figure 6-6 summarises the different roles of Ago2 in mammalian systems. Considered together, the use of small RNA biogenesis/effector pathway mutants has provided us with significant knowledge in understanding the roles these small non-coding RNAs play in many areas in biology. However, components in the biogenesis and effector pathways are still not fully uncovered. Novel mutants will further aid our understanding of the fundamental roles non-coding small RNAs play in biology.

Figure 6-6: Multiple roles of Ago2 in the non-coding small RNA biogenesis, regulation and effector pathways.



Chapter Seven – Mobilisation of giant *piggyBac* transposons in the mouse genome

1. Introduction

The development of new technologies that allows the stable delivery of large genomic DNA fragments in mammalian systems is important for genetic research as well as applications in gene therapy. Genomic sequences contain not only protein-coding regions, but also important regulatory elements, which are critical to ensure the appropriate level as well as regulated spatial-temporal gene expression in an organism. Although heterologous-promoter driven cDNA sequences can be readily introduced as transgenic elements, these rarely provide the full repertoire of alternative isoforms, physiological-relevant expression patterns and are prone to silencing. Therefore, the delivery of large contiguous genomic sequences is essential to achieve regulated gene expression.

Episomal vectors based on Epstein-Barr virus (Wade-Martins et al., 2000) and Herpes Simplex Virus type 1 (Hibbitt and Wade-Martins, 2006), have been used to introduce large genomic sequences into mammalian cells. As episomes can be lost without selection pressure, they do not guarantee indefinite expression of the delivered cargo. Long-term expression of a transgene is most reliably achieved by stable integration. Retroviral and lentiviral vectors have been used for this purpose, but their cargo capacity is limited to 10 kb and they are not suited for the delivery of intron-containing cargos. Additionally, these viral systems have immunogenic and tumorigenic potential.

Transfection of naked DNA has been used for large-cargo delivery. Pronuclear injection of bacterial artificial chromosomes (BACs) has been successful for transgenesis of up to 300 kb. However, the integrity, integration site and copy number can not be controlled. BAC vectors have also been used for targeting large cargos to defined genomic positions in ES cells via homologous recombination (Valenzuela et al., 2003), but the efficiency is locus-dependent and can be very low. Recombinases such as Cre have also been used to deliver BACs to a pre-

defined genomic location by recombination-mediated cassette exchange (Wallace *et al.*, 2007; Prosser *et al.*, 2008); however, pre-engineering of target sites in the genome is necessary. While these methods are useful for certain applications, all have limitations and most of them are not able to revert the insertion of large DNA fragments.

DNA transposons have emerged as flexible and efficient molecular vehicles to mediate stable cargo transfer. However, the ability to carry DNA fragments greater than 10 kb is limited in most DNA transposons. The development of a DNA transposon system with the capacity of accommodating large genomic fragment will be an important technology that complements the current methods to facilitate a wide range of genetic and genomic applications. PB has previously been shown to be capable of delivering DNA fragment of up to 10 kb in mice without the significant loss of its transposition efficiency (Ding *et al.*, 2005b). In this chapter, the cargo capacity of PB transposon has been investigated with the aim to develop PB into a giant genomic cargo carrier.

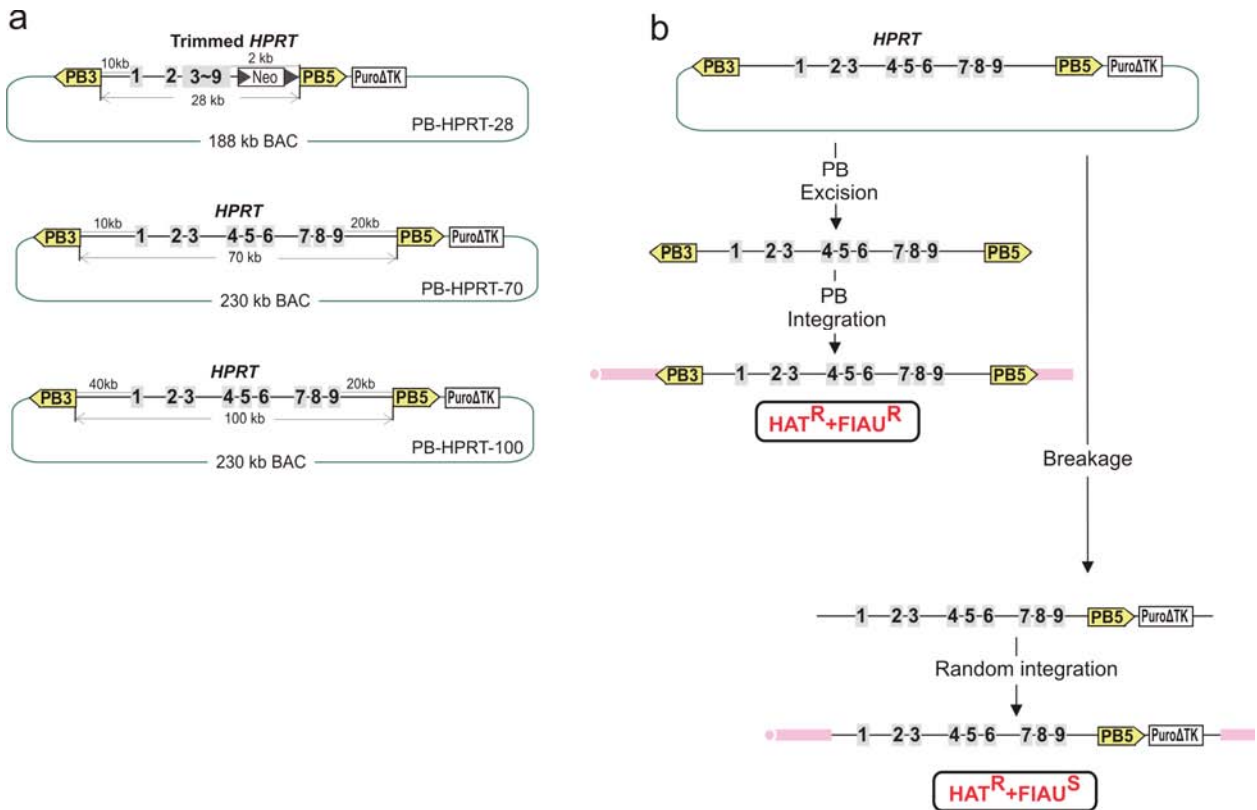
2. Results

2.1. The generation of giant PB transposons

In order to assess the cargo capacity of PB, we have constructed a series of PB transposons with sizes of 28, 70 and 100 kb on a BAC backbone for testing their transposition capability in mouse ES cells (Figure 7-1a). A BAC containing the human *HPRT* gene was used to construct these transposons with the human *HPRT* gene as a positive selection marker. When introduced into the *Hprt*-deficient ES cells, these transposons can complement the *Hprt* deficiency, so that clones in which transposition has occurred could be directly selected in HAT. This BAC was modified by insertion of both PB ITRs by recombineering technology to generate PB transposons with 70 and 100 kb cargos (PB-HPRT-70 and PB-HPRT-100). The 28 kb PB (PB-HPRT-28) was constructed by substituting the genomic regions from exon 3 to 9 with the corresponding part of the *HPRT* cDNA using the PB-HPRT-70 vector and recombineering technology, (Figure 7-1a). The BACs were further modified by insertion of a *Puro Δ tk* cassette (Chen and Bradley, 2000) immediately downstream of the PB5'ITR, so that random insertions could be counter-selected. ES cell clones in which the *HPRT* gene has been

inserted by transposition should exclude the *PuroΔtk* cassette and will be resistant to FIAU (Figure 7-1b).

Figure 7-1: Giant PB construction and selection scheme incorporated to detect their transposition in mouse ES cells.



a, Giant PB-HPRT constructs modified from the same BAC. **b**, A scheme of alternative genomic integration pathways. PBase can mediate precise excision of the giant PB from the BAC and insert this into the ES cell genome, generating cells that are resistant both to HAT and FIAU. If physical breakage of the BAC occurs the PB ITRs are separated, random integration of the BAC can occur including the *PuroΔtk* cassette. If the *HPRT* gene on the BAC is intact, the cells will be HAT resistant. If the *puroΔtk* cassette is intact, the cells will be sensitive to FIAU.

2.2. Transposition detection of giant PB transposons in ES cells

The *Hprt*-deficient AB2.2 mouse ES cell line was transiently transfected using the lipofection method with one of the two types of the *piggyBac* transposase (PBase); the mammalian codon optimized version, mPBase (Cadinanos and Bradley, 2007) or a hyperactive form HyPBase (Yusa, *et al* Submitted). These PBase-expressing plasmids also contain a puromycin selection cassette so that ES cells expressing PBase could be enriched by a pulse puromycin selection (Taniguchi et al., 1998). As a negative control, a plasmid co-expressing enhanced green fluorescent protein (eGFP) and the puromycin resistant cassette was used. With this enrichment method, greater than 50 % of the ES cells were expected to be expressing PBase given that they were eGFP positive (Figure 7-2). Three days after PBase transfection, the BACs harbouring different-sized PB transposons were introduced by electroporation and the cells were replated in HAT and FIAU containing medium to select for ES cells with a stable integration of PB. The HAT and FIAU resistant colonies were either picked and analysed individually or pooled together for high throughput PB-transposition identification and mapping using the Illumina sequencing platform. Figure 7-3 shows the entire experimental scheme.

Figure 7-2: Transfection efficiency of the AB2.2 ES cells determined by the eGFP expression after a pulse of puromycin selection.

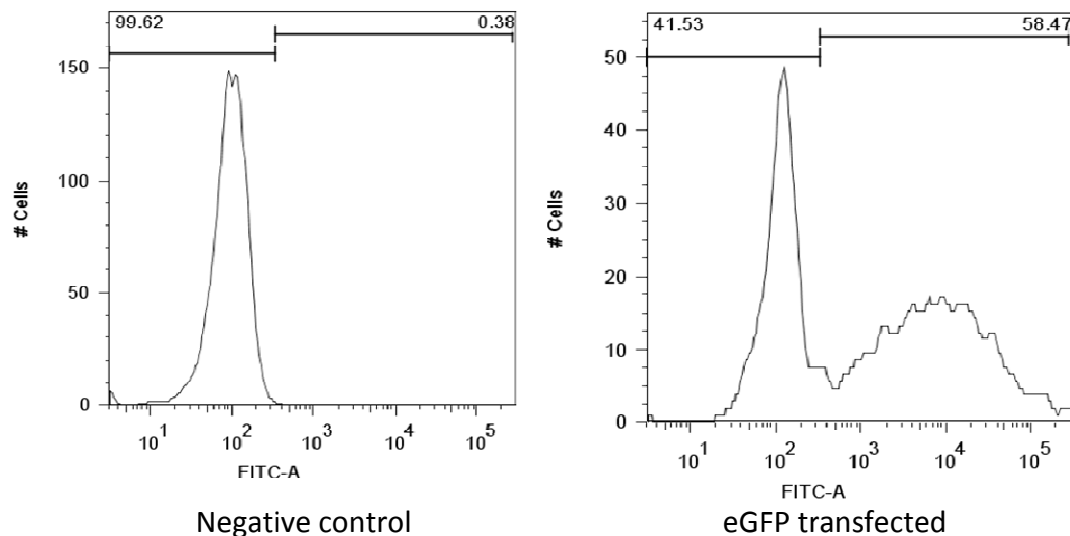
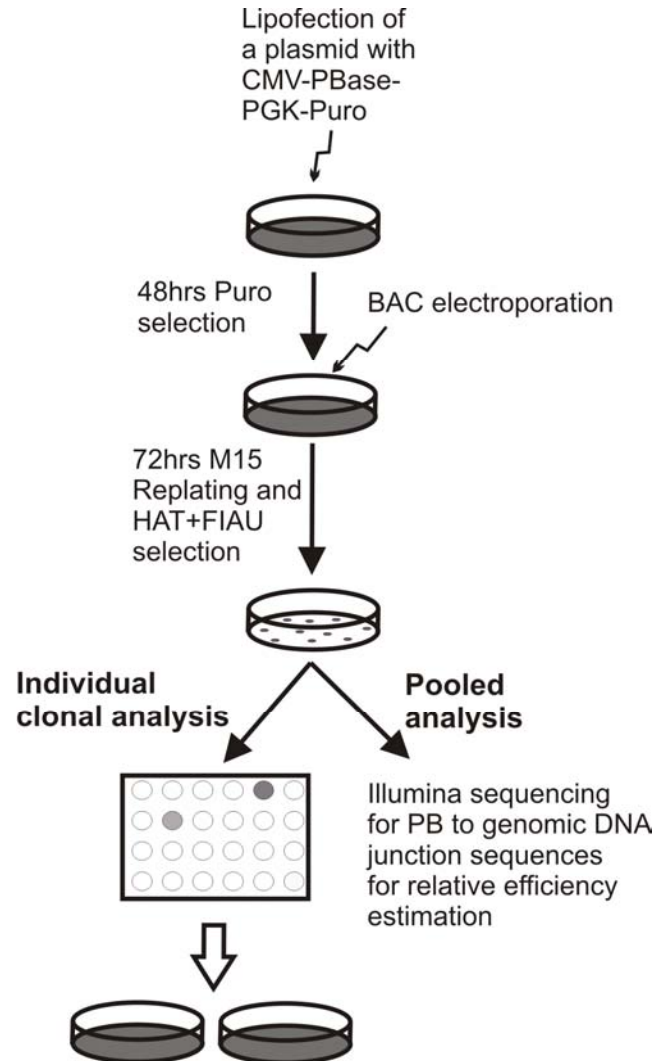


Figure 7-3: Outline of the experimental scheme.



All three PB-HPRT transposons gave rise to HAT and FIAU double-resistant colonies and the colony numbers exceeded those in the non-PBase control, Table 7-1. Unexpectedly, the number of double-resistant colonies did not vary greatly with the size of the PB transposon. However, the number of double-resistant colonies increased significantly when HyPBase was supplied compared to mPBase suggesting that transpositions occurred. HAT and FIAU resistant colonies can be generated by two competing mechanisms: transposition or random integration of the BAC with the loss of the *Puro Δ tk* cassette. The proportion of random

integration events was higher with the 70 and 100 kb PB transposons judged by the number of HAT and FIAU double-resistant colonies in the non-PBase control. The larger transposons are expected to have a higher background of HAT and FIAU double-resistant colonies because the *Puro Δ tk* cassette is located 20 kb from the 3' end of the *HPRT* gene whereas the *Puro Δ tk* cassette in the 28 kb PB is only separated from the *HPRT* stop codon by 2 kb (Figure 7-1a).

Genuine transposition can be distinguished from the random integration by analyzing sequences adjacent to the PB inverted repeats. If PBase-mediated integration occurred, both ends of the PB ITRs should be flanked by mouse genomic sequences together with PB's signature recognition site TTAA (Ding *et al.*, 2005a; Wang *et al.*, 2008b; Liang *et al.*, 2009). If random integration occurred, the original BAC vector sequences adjacent to PB ITRs will be present. Illumina sequencing technology was used to identify a large number of genuine PB transposition events from the random integrations. HAT and FIAU resistant colonies were pooled from each experimental condition; genomic DNA was extracted and subjected to paralleled paired-end sequencing to identify the PB5' ITR – genomic junctions, Figure 7-4.

Transposition events were identified for all three transposons when either mPBase or HyPBase was used, Table 7-1. The absolute number of transposition events dropped significantly as the cargo size increased from 28 kb to 70 kb, however, the 70 kb and 100 kb PB transposons showed a similar number of integration events. The proportion of transposition events among the HAT and FIAU double-resistant colonies was lower with the 70 and 100 kb PB transposons than the 28 kb PB transposon, reflecting the higher rate of random integrations of large transposons as seen in the HAT and FIAU resistant colony number in the non-PBase control. HyPBase-mediated transposition was approximately four times that of mPBase for the larger transposons and seven times for the 28 kb transposon. Albeit at lower efficiency, wild type PBase can mediate transposition with large cargos, suggesting that the large-cargo capacity is an intrinsic property to the PB system, not acquired as a result of modifications to PBase.

Table 7-1: Transposition efficiency of different sized PB transposons and versions of the PBase.

Transposons	mPBase		HyPBase		eGFP
	HAT+ FIAU	Transposon Events (%)*	HAT+ FIAU	Transposon Events (%)*	HAT+ FIAU
PB-HPRT-28	39	18 (46 %)	183	131 (72 %)	9
PB-HPRT-70	56	9 (16 %)	104	26 (25 %)	37
PB-HPRT-100	77	5 (7 %)	103	30 (29 %)	47

The number of transposition events was determined using massive-parallel sequencing from pooled HAT and FIAU resistant clones. *: Percentage of transposition events as a fraction of HAT and FIAU double-resistant colonies. The transposition events are assumed to be one per cell.

Figure 7-4: Schematic representation of the experimental platform (a) and bioinformatics (b) analysis to identify transposition events using the Illumina sequencing.

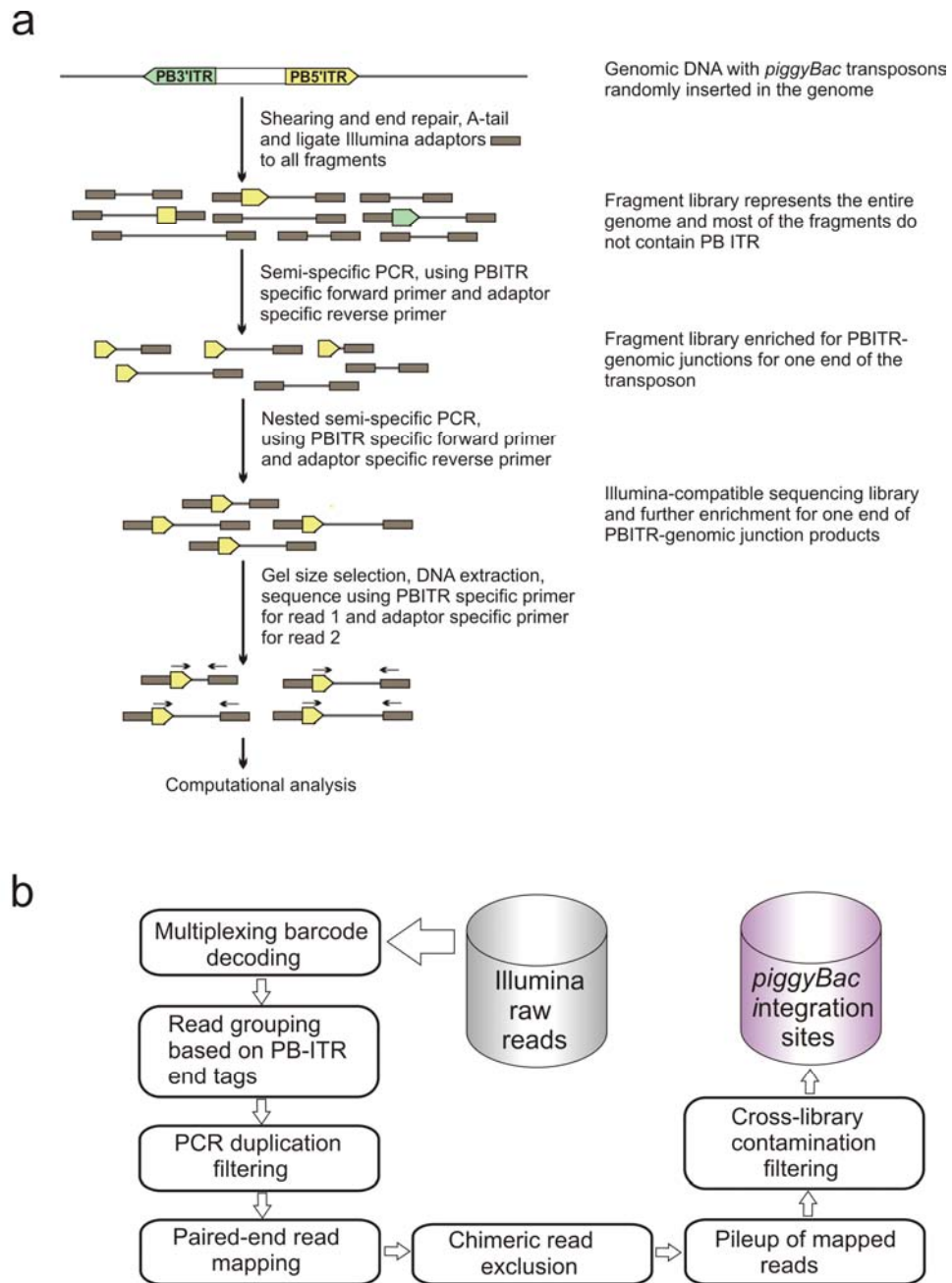


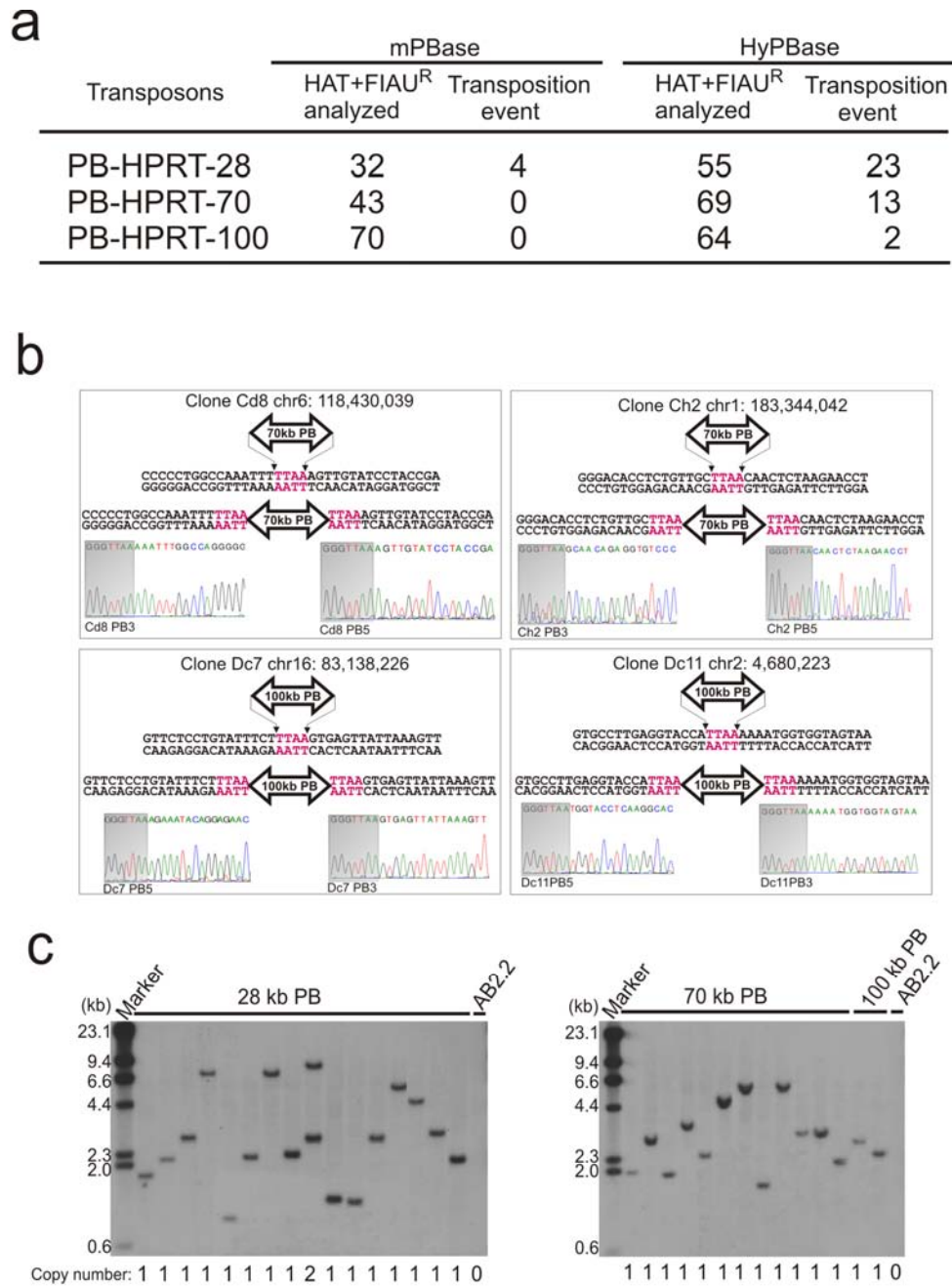
Figure **a** was adapted from Daniel J. Turner’s original figure and **b** was adapted from Zemin Ning’s original figure.

2.3. Individual clone analysis of giant PB genomic integrations

In a separate experiment, double-resistant colonies were generated and analyzed individually to identify integration sites using Splinkerette PCR, Figure 7-5a. For each transposon insertion analyzed, both the PB5' and PB3' ITR – host genome junction sequences were contiguous in the mouse genome (Figure 7-5b). Analysis of the transposon copy number in these clones by Southern blotting also revealed that almost all the PB-mediated integrations were single-copy (Figure 7-5c).

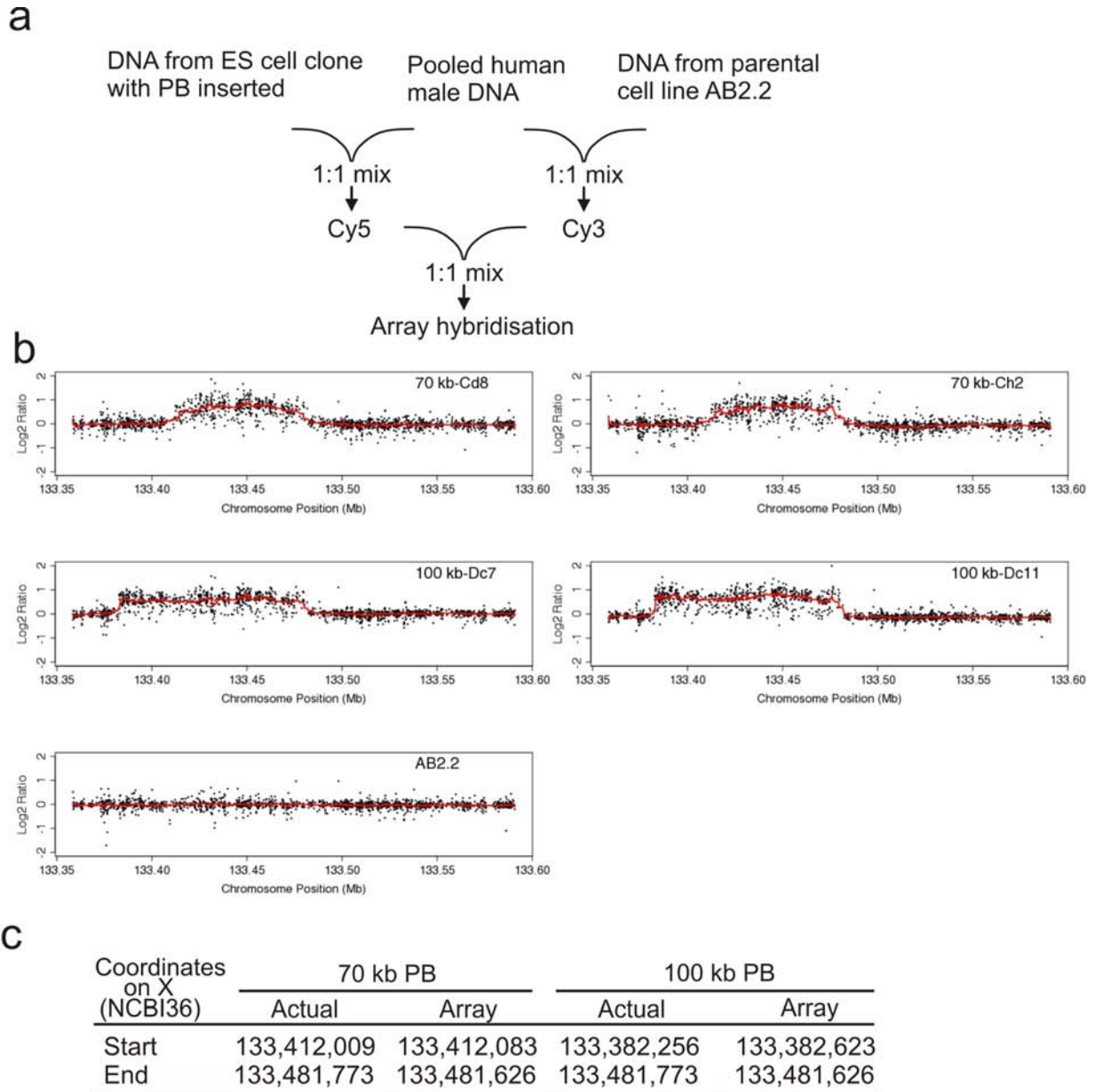
One of the major advantages of using transposition to deliver large genomic DNA fragments is that cargo integrity is expected to be maintained. To examine this, we used a custom high-resolution (average probe spacing of 130 bp) comparative genomic hybridization (CGH) array, covering the entire human *HPRT*-containing BAC. The samples and control were mixed with equal amount of human male DNA to provide a baseline for array normalization, Figure 7-6a. Four independent clones with 70 kb (Cd8 and Ch2) and 100 kb (Dc7 and Dc11) PB insertions were assessed. The regions of copy number gain in all the clones precisely matched the regions flanked by the PB ITRs (Figure 7-6b,c). Within these regions, the human DNA sequences were continuous and did not contain any detectable change (Figure 7-6b). Thus, the large cargos mobilized as PB transposons remained intact in all cases.

Figure 7-5: Giant PB transposition with single copy integration per cell.



a, Transposition events of different sized PB transposons and versions of the PBase from single-colony analysis. **b**, Precise integration of giant PB transposons at the expected TTAA site. The chromosomal coordinates of the first T corresponds to the PB recognition site TTAA are shown (NCBI m37). **c**, Southern blot illustrating the copy number of the PB mediated large-cargo integrations using the PB5'ITR as the detection probe. The genomic DNA was digested with *SpeI* and *XbaI*.

Figure 7-6: Regional high density CGH array analysis to determine the PB cargo integrity.

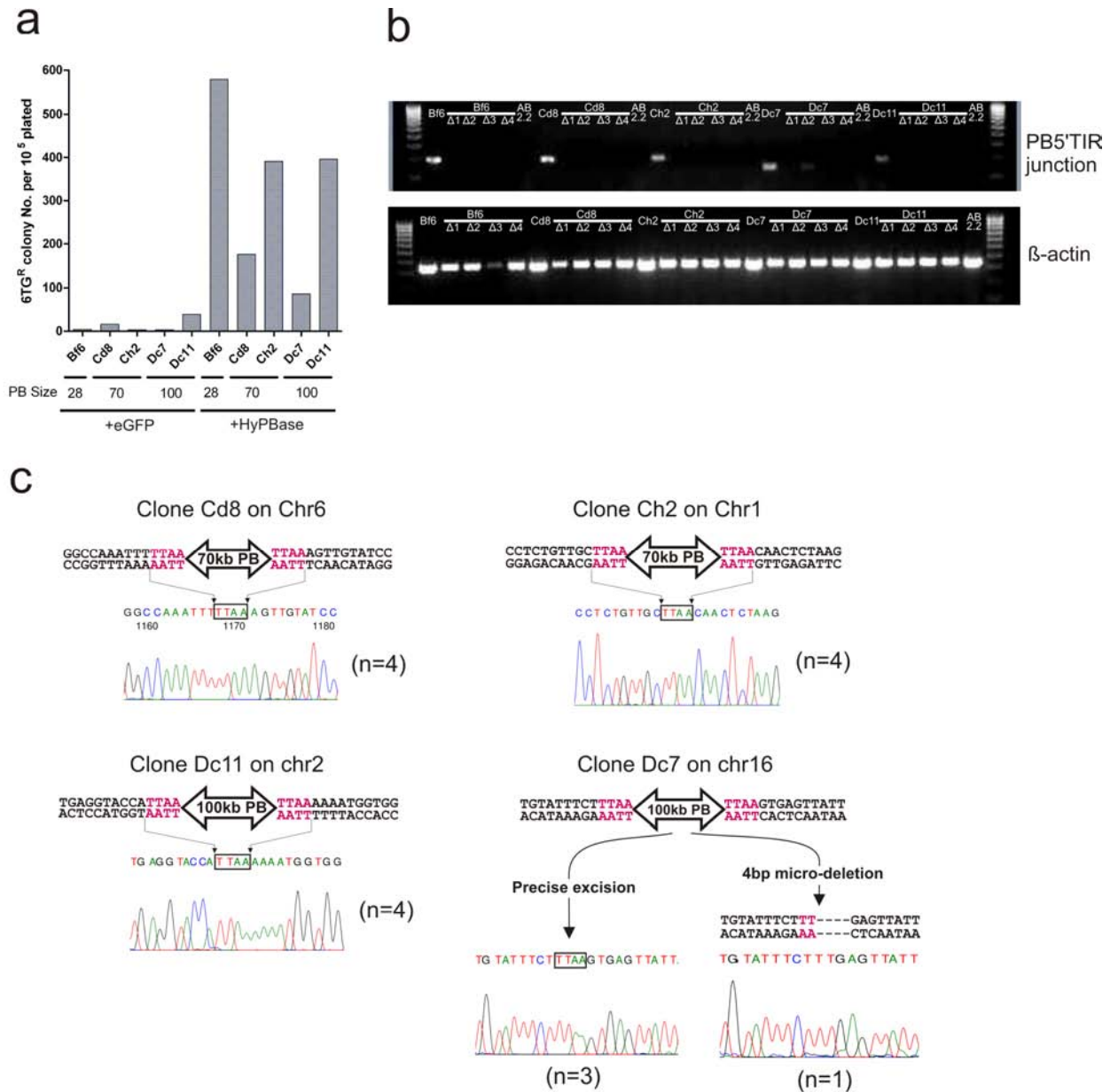


a, The pooled human male DNA was used as a reference to detect the gain of copy number for the X-linked genomic DNA in the ES cell clones with PB integrated. b, Regional CGH analysis showing the gain of an extra copy of the human *HPRT* gene delivered by PB transposition. The red line was calculated as the running median of the Log_2 value of each CGH probe to aid the visualization. c, Comparison of CGH array read-outs for the positions of copy number gain and the actual positions.

2.4. Analysis of chromosomal excision capability of giant PB

ES cell lines with 28 kb (Bf6), 70 kb (Cd8 and Ch2) or 100 kb (Dc7 and Dc11) PB transposons were transiently transfected with HyPBase and enriched for expression with a pulse puromycin selection. Following a period of culture to allow for the decay of *HPRT* mRNA and protein products, the cells were plated at low density and selected for the loss of *HPRT* in 6-TG. PB excision was observed for all clones tested with efficiencies ranging from 0.1 % to 0.6 % of the total number of cells plated (Figure 7-7a). 6-TG resistant colonies derived from each of the four parental PB-containing clones were examined by genomic PCR using transposon-host specific primer sets for each integration site, and none exhibited the PB ITR-host genomic junctions (Figure 7-7b). PB does not normally leave a footprint upon excision. We therefore sequenced the transposon excision sites in all of the 6-TG resistant clones to check for their intactness. Precise excision was observed in all clones derived from the three donor sites (Cd8, Ch2, and Dc11). However, one out of four 6-TG resistant clones derived from the Dc7 clone showed a micro-deletion (Figure 7-7c). Micro-deletions upon PB excision have been reported previously (Wang et al., 2008b), suggesting that this low frequency of imprecise excision is not due to the size of the cargo.

Figure 7-7: PBase mediated excision of giant PB transposons from the ES cell genome.



a, Genomic excision efficiency of five ES cell clones containing giant PB transposons with different cargo sizes following transfections with either HyPBase or eGFP (control) **b**, Molecular analysis of individual 6-TG resistant colonies to evaluate fidelity of excision events. Excision of PB eliminates the PB-host junction fragment amplified in the parental lines. **c**, Analysis of PB excision identified one clone with a micro-deletion (clone Dc7) following the excision of a 100 kb PB transposon. “n” represents the number of 6-TG resistant colonies with the shown sequence traces of the excision site.

3. Discussion

In this chapter, the mobilisation of giant *piggyBac* transposons of up to 100 kb from exogenous BAC vectors and endogenous genomic loci in mouse ES cells was demonstrated. Mobilisation of giant PB transposons achieves stable but precisely revertible genomic insertions. Importantly, large DNA cargos remain intact during transposition and the copy number of the delivery is predominantly one.

In the vector-to-chromosome transposition assay, the efficiency of transposition dropped as the cargo size increased, Table 7-1. This could be caused by the lack of integrity of the BAC vector during preparation and electroporation. Breakage between PB ITRs prevents transposition and stimulates random integration. The chance of a break occurring between the PB ITRs increases with the distance between them. In the chromosomal transposition assay, the frequency of excision appeared to be less dependent on the size of the transposon than the integration site. This supports the view that one of the major factors influencing vector-to-chromosome transposition is the continuity of the BAC DNA between the PB ITRs, rather than inherent limits in the transposition reaction per se.

In the system described here, transposition events can be enriched by negative selection using FIAU, because transposition uncouples the PB transposon from a negatively selectable *puro Δ tk* cassette on the BAC backbone. It follows that the tighter the linkage between the positive selection marker in the transposon and the negative selection cassette, the greater the degree of enrichment for transposition events. The tight linkage in the 28 kb transposon allowed us to achieve 70 % transposition efficiency after positive-negative selection, Table 7-1a. This selection scheme is very useful to enrich for transposition by selecting against random integrations of the BACs.

We have demonstrated that giant PB transposons effectively deliver intact large genomic DNA fragments with a controllable copy number. This is useful in many genetic applications such as BAC transgenesis and genetic complementation. Another DNA transposon, *Tol2*, has

also recently been shown to deliver a 70 kb genomic DNA for transgenesis (Suster et al., 2009). The additional ability of *piggyBac* to cleanly excise large genomic DNA fragments provides a valuable genome engineering technology for creating *in vitro* and *in vivo* gains and losses of large genomic regions.

PB-mediated integration of large genomic fragments can provide “permanent complementation” with prolonged and physiologically-regulated gene expression. It also avoids the complications of viral vectors, which can induce host-immune responses and tumorigenesis. Viral vectors contain the polIII promoter activity within the 5’ viral long terminal repeat (LTR), therefore insertions of the viral vectors may ectopically drive the expression of surrounding genes or part of the gene if insertions is present within a gene. If the surrounding gene is a proto-oncogene, such as *C-myc* or *Ras*, over-expression of such genes can initiate tumorigenesis. On the contrary, PB ITRs do not contain promoter signals, thus they have a low risk of inducing tumour formation. Although PB integration is random, specific integration sites of PB can be screened to identify permissive locations that are not likely to affect normal function for clinical applications. The development of giant PB transposons will be valuable for therapeutic gene delivery of large genomic sequences in patient-specific iPS cell lines to combat a range of human genetic diseases.

Giant PB transposons are comparatively simple to construct. In principle, a genome-wide resource of PB-BACs could be generated using recombineering technology (Chang et al., 2007). Such a resource can be used in genetic screens and in complementation studies. Transient expression of PBase to mediate giant PB transposition does not require prior genome modification, thus giant PB libraries can be used in most cell types and organisms.

Taken together, the work presented here provides a framework for using *piggyBac* to mobilise large genomic DNA fragments. This will open the door to a wide range of future applications in genetics and genomic research as well as clinical medicine, which have been difficult to conduct previously with other tools.

Chapter Eight – Conclusions, future work and projections into the future

1. Conclusions

This thesis encompasses two distinct bodies of work in advancing transposon technology, focused on the DNA transposon *PiggyBac* in mammalian genetics and genomics. The first part of my research work focused on the establishment of a genome-wide insertional mutagenesis strategy utilising efficient intra-genomic mobilisation of the PB transposon to conduct recessive genetic screens in *Blm*-deficient mouse ES cells. This technological development with the PB transposon is a further step forward towards the goal of unbiased genome-wide mutagenesis in mammalian systems. Combined with the methodology established using *Blm*-deficient ES cells for parallel conversion of heterozygous mutants to homozygosity on a genome-wide scale, the mutant library should cover a broader range of genes to facilitate genome-function assignment in many biological pathways of interests.

The second part concerns the development of novel reporter systems which enable phenotypic screens for the discovery of components of the miRNA biogenesis and effector pathways together with the established mutagenic strategy. This type of non-hypothesis driven genetic screen has so far not been conducted for the study of the mammalian miRNA pathways, although the siRNA pathway has been investigated using such an approach in *C. elegans*, *Drosophila* and mouse ES cells (Kim *et al.*, 2005; Dorner *et al.*, 2006; Trombly *et al.*, 2009). The two reporter systems established complement each other in the potential of their scope to probe two different branches of the miRNA downstream effector pathways. Upon completion of screens using both systems, it is hoped that some currently unanswered questions regarding the determinants and differential regulation of these two branches of the effector pathway can be revealed through the identification of novel factors. Although only preliminary screening results with one of the systems are described in this thesis, experiments are ongoing towards the completion of the large-scale screening work.

The final section of this thesis describes an independent body of research focusing on discovering new aspects of the *piggyBac* transposon for basic research as well as clinical

applications. The cargo capacity of the PB transposon was addressed for its use in large genomic DNA delivery. Up to 100 kb of genomic DNA cargo can be integrated and excised from the mouse genome using the PB transposon system. To our knowledge, this is the only transposon to possess such a large cargo capacity. This work could have exciting applications in correcting human genetic disorders in conjunction with the current developments in patient-specific induced pluripotent stem cells or adult stem cell derivation and transplantation.

2. Future work

The work described in this thesis has raised several interesting questions and opened up several lines of exciting research work which will be followed up in the near future. Firstly, large-scale phenotypic screening using the two developed reporter systems is still ongoing to identify the novel components within the miRNA biogenesis and effector pathways. The identification of potential novel factors can open up biological investigations into the function of these factors within the pathway. This work may reveal new aspects in the miRNA biogenesis and miRNA-mediated gene silencing.

Secondly, there are still a big scope for improvement of the current homozygous mutant conversion system using *Blm*-deficient ES cells. Although *Blm*-deficiency elevates the heterozygous to homozygous conversion rate, there is still a significant proportion of cells within the mutant pools that are irrelevant heterozygous cells. A method has been developed by Yue Huang in our laboratory to use a double-selection system to enrich the homozygous mutants by eliminating the heterozygous cells through the drug selection (Huang, *et al* and Horie *et al*, submitted). This method can successfully isolate homozygous mutants from their counterparts in a clone-by-clone fashion. However, the mixed mutant pooling method for assessing thousands of mutants at a time can not yet be coupled to this double-selection system due to the presence of aneuploid ES cells or/and cells with two-copy mutagen integrations. Through such a strong double-selection scheme, aneuploid cells or cells possessing two copies of the transposons can dominate the mutant pools. Thus methods to

allow the homozygous enrichment within a mixed mutant pool will further improve the efficiency of isolating relevant homozygous mutants.

Thirdly, the future investigation into the kinetics of *piggyBac* intra-genomic mobilisation may reveal the fundamental characteristics of DNA transposon mobilisation. PB transposition is highly efficient within the mammalian genome, and local hopping is not observed in cell culture or *in vivo* when a sufficient amount of transposase is supplied. In contrast to *piggyBac*, another DNA transposon from a separate family, *Sleeping Beauty*, is much less efficient than *piggyBac* during intra-genomic mobilisation and also shows a severe local hopping effect (Keng et al., 2005). However, *piggyBac* intra-genomic local hopping has recently been shown in Chapter Three of this work and by Wang and co-workers (Wang et al., 2008b). This observation poses the question as to whether the general mechanism of transposon intra-genomic mobilisation is dependent on continuous cycles of integration and excision gradually moving away from the donor site, until the transposition reaction is terminated by either a lack of sufficient transposase or integrations into “difficult-to-excise” genomic contexts. Understanding the transposition kinetics not only helps us to gain further understanding of the fundamental biology in DNA transposons, it may also highlight safety issues in using DNA transposons in clinical applications as every excision may bring the possibility of a mutagenic transposon foot-print.

Finally, much of the *piggyBac* transposon technology developed in mouse ES cells can be directly transferred to human cell-based research as *piggyBac* transposition is host independent. Although mice bear a high resemblance to human genetically, physiologically and anatomically, there are still major differences at the molecular, cellular and physiological levels. Furthermore, many disease models using mice can not capture the disease characteristics seen in human patients. Recent rapid developments in human ES cell culture and *in vitro* differentiation conditions allow the investigation of cellular pathways in these non-transformed human cell lines. Forward genetic screens can be conducted in human ES cells or iPS cells to uncover biological pathways, for example, the highly human-specific pathways that human immune deficiency virus (HIV) and influenza utilise for infection. The

derivation of induced human iPS cells and adult stem cells from patients opens up the possibilities of investigating these diseases in “authentic” pathological scenarios at molecular level and curing the disease using such matched cells after correction. Therefore, useful genetic tools and methodologies established in mice and cell cultures will play a vital role in helping us to understand the physio-/pathological mechanisms in the human system and providing therapeutic avenues for human patients.

References

- Abbott, A. (2004). Laboratory animals: The Renaissance rat. *Nature* **428**, 464-466.
- Abrahante, J. E., Daul, A. L., Li, M., Volk, M. L., Tennessen, J. M., Miller, E. A., and Rougvie, A. E. (2003). The *Caenorhabditis elegans* hunchback-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. *Dev Cell* **4**, 625-37.
- Alwin, S., Gere, M. B., Guhl, E., Effertz, K., Barbas, C. F., 3rd, Segal, D. J., Weitzman, M. D., and Cathomen, T. (2005). Custom zinc-finger nucleases for use in human cells. *Mol Ther* **12**, 610-7.
- Ambros, V. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*. *Cell* **57**, 49-57.
- Arbones, M. L., Austin, H. A., Capon, D. J., and Greenburg, G. (1994). Gene targeting in normal somatic cells: inactivation of the interferon-gamma receptor in myoblasts. *Nat Genet* **6**, 90-7.
- Austin, S., Ziese, M., and Sternberg, N. (1981). A novel role for site-specific recombination in maintenance of bacterial replicons. *Cell* **25**, 729-36.
- Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev* **22**, 2773-85.
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-97.
- Bender, A. M., Collier, L. S., Rodriguez, F. J., Tieu, C., Larson, J. D., Halder, C., Mahlum, E., Kollmeyer, T. M., Akagi, K., Sarkar, G., Largaespada, D. A., and Jenkins, R. B. (2010). Sleeping beauty-mediated somatic mutagenesis implicates CSF1 in the formation of high-grade astrocytomas. *Cancer Res* **70**, 3557-65.
- Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., Kerkhoven, R. M., Madiredjo, M., Nijkamp, W., Weigelt, B., Agami, R., Ge, W., Cavet, G., Linsley, P. S., Beijersbergen, R. L., and Bernards, R. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-7.
- Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363-6.
- Beumer, K. J., Pimpinelli, S., and Golic, K. G. (1998). Induced chromosomal exchange directs the segregation of recombinant chromatids in mitosis of *Drosophila*. *Genetics* **150**, 173-88.
- Bibikova, M., Golic, M., Golic, K. G., and Carroll, D. (2002). Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. *Genetics* **161**, 1169-75.
- Booher, R., and Beach, D. (1987). Interaction between *cdc13+* and *cdc2+* in the control of mitosis in fission yeast; dissociation of the G1 and G2 roles of the *cdc2+* protein kinase. *Embo J* **6**, 3441-7.
- Bradley, A., Evans, M., Kaufman, M. H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature* **309**, 255-6.
- Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71-94.

- Buehr, M., Meek, S., Blair, K., Yang, J., Ure, J., Silva, J., McLay, R., Hall, J., Ying, Q. L., and Smith, A. (2008). Capture of authentic embryonic stem cells from rat blastocysts. *Cell* **135**, 1287-98.
- Buerstedde, J. M., and Takeda, S. (1991). Increased ratio of targeted to random integration after transfection of chicken B cell lines. *Cell* **67**, 179-88.
- Burdon, T., Smith, A., and Savatier, P. (2002). Signalling, cell cycle and pluripotency in embryonic stem cells. *Trends Cell Biol* **12**, 432-8.
- Cadinanos, J., and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. *Nucl. Acids Res.* **35**, e87-.
- Carette, J. E., Guimaraes, C. P., Varadarajan, M., Park, A. S., Wuethrich, I., Godarova, A., Kotecki, M., Cochran, B. H., Spooner, E., Ploegh, H. L., and Brummelkamp, T. R. (2009). Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326**, 1231-5.
- Carroll, D. (2004). Using nucleases to stimulate homologous recombination. *Methods Mol Biol* **262**, 195-207.
- Cary, L. C., Goebel, M., Corsaro, B. G., Wang, H. G., Rosen, E., and Fraser, M. J. (1989). Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* **172**, 156-169.
- Cervantes, R. B., Stringer, J. R., Shao, C., Tischfield, J. A., and Stambrook, P. J. (2002). Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proc Natl Acad Sci U S A* **99**, 3586-90.
- Chalfie, M., Horvitz, H. R., and Sulston, J. E. (1981). Mutations that lead to reiterations in the cell lineages of *C. elegans*. *Cell* **24**, 59-69.
- Chan, W., Costantino, N., Li, R., Lee, S. C., Su, Q., Melvin, D., Court, D. L., and Liu, P. (2007). A recombineering based approach for high-throughput conditional knockout targeting vector construction. *Nucleic Acids Res* **35**, e64.
- Chang, T. C., Yu, D., Lee, Y. S., Wentzel, E. A., Arking, D. E., West, K. M., Dang, C. V., Thomas-Tikhonenko, A., and Mendell, J. T. (2008). Widespread microRNA repression by Myc contributes to tumorigenesis. *Nat Genet* **40**, 43-50.
- Chang, Y. F., Imam, J. S., and Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**, 51-74.
- Cheloufi, S., Dos Santos, C. O., Chong, M. M., and Hannon, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*.
- Chen, Y., Yee, D., Dains, K., Chatterjee, A., Cavalcoli, J., Schneider, E., Om, J., Woychik, R. P., and Magnuson, T. (2000). Genotype-based screen for ENU-induced mutations in mouse embryonic stem cells. *Nat Genet* **24**, 314-317.
- Chen, Y. T., and Bradley, A. (2000). A new positive/negative selectable marker, puDeltatk, for use in embryonic stem cells. *Genesis* **28**, 31-5.
- Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436**, 740-744.
- Chester, N., Babbe, H., Pinkas, J., Manning, C., and Leder, P. (2006). Mutation of the murine Bloom's syndrome gene produces global genome destabilization. *Mol Cell Biol* **26**, 6713-26.

-
- Chester, N., Kuo, F., Kozak, C., O'Hara, C. D., and Leder, P. (1998). Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. *Genes Dev* **12**, 3382-93.
- Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., Johnston, W. K., Russ, C., Luo, S., Babiarz, J. E., Blelloch, R., Schroth, G. P., Nusbaum, C., and Bartel, D. P. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* **24**, 992-1009.
- Chu, E. H. (1971). Mammalian cell genetics. 3. Characterization of x-ray-induced forward mutations in Chinese hamster cell cultures. *Mutat Res* **11**, 23-34.
- Chung, K. H., Hart, C. C., Al-Bassam, S., Avery, A., Taylor, J., Patel, P. D., Vojtek, A. B., and Turner, D. L. (2006). Polycistronic RNA polymerase II expression vectors for RNA interference based on BIC/miR-155. *Nucleic Acids Res* **34**, e53.
- Collier, L. S., Adams, D. J., Hackett, C. S., Bendzick, L. E., Akagi, K., Davies, M. N., Diers, M. D., Rodriguez, F. J., Bender, A. M., Tieu, C., Matise, I., Dupuy, A. J., Copeland, N. G., Jenkins, N. A., Hodgson, J. G., Weiss, W. A., Jenkins, R. B., and Largaespada, D. A. (2009). Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality. *Cancer Res* **69**, 8429-37.
- Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J., and Largaespada, D. A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-6.
- Collins, F. S. (1992). Positional cloning: let's not call it reverse anymore. *Nat Genet* **1**, 3-6.
- Cooley, L., Kelley, R., and Spradling, A. (1988). Insertional mutagenesis of the Drosophila genome with single P elements. *Science* **239**, 1121-8.
- Cullen, B. R. (2009). Viral and cellular messenger RNA targets of viral microRNAs. *Nature* **457**, 421-425.
- Daniels, S. B., McCarron, M., Love, C., and Chovnick, A. (1985). Dysgenesis-induced instability of rosy locus transformation in Drosophila melanogaster: analysis of excision events and the selective recovery of control element deletions. *Genetics* **109**, 95-117.
- Davis, A. C., Wims, M., Spotts, G. D., Hann, S. R., and Bradley, A. (1993). A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. *Genes Dev* **7**, 671-82.
- Davis, B. N., Hilyard, A. C., Lagna, G., and Hata, A. (2008). SMAD proteins control DROSHA-mediated microRNA maturation. *Nature* **454**, 56-61.
- Davis, M. P. (2009). Generation of a murine ES cell system deficient in microRNA processing for the identification of miRNA targets. In "Wellcome Trust Sanger Institute", Vol. Doctor of Philosophy. University of Cambridge.
- Debec, A. (1984). Evolution of karyotype in haploid cell lines of Drosophila melanogaster. *Exp Cell Res* **151**, 236-46.
- DeChiara, T. M., Efstratiadis, A., and Robertson, E. J. (1990). A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. *Nature* **345**, 78-80.
-

- Dickins, R. A., Hemann, M. T., Zilfou, J. T., Simpson, D. R., Ibarra, I., Hannon, G. J., and Lowe, S. W. (2005). Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. *Nat Genet* **37**, 1289-95.
- Diederichs, S., and Haber, D. A. (2007). Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression. *Cell* **131**, 1097-108.
- Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005a). Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. *Cell* **122**, 473-483.
- Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005b). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473-83.
- Doench, J. G., Petersen, C. P., and Sharp, P. A. (2003). siRNAs can function as miRNAs. *Genes Dev* **17**, 438-42.
- Dong, J., Albertini, D. F., Nishimori, K., Kumar, T. R., Lu, N., and Matzuk, M. M. (1996). Growth differentiation factor-9 is required during early ovarian folliculogenesis. *Nature* **383**, 531-5.
- Donnelly, M. L., Luke, G., Mehrotra, A., Li, X., Hughes, L. E., Gani, D., and Ryan, M. D. (2001). Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip'. *J Gen Virol* **82**, 1013-25.
- Dorner, S., Lum, L., Kim, M., Paro, R., Beachy, P. A., and Green, R. (2006). A genomewide screen for components of the RNAi pathway in *Drosophila* cultured cells. *Proc Natl Acad Sci U S A* **103**, 11880-5.
- Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G., and Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221-6.
- Dupuy, A. J., Rogers, L. M., Kim, J., Nannapaneni, K., Starr, T. K., Liu, P., Largaespada, D. A., Scheetz, T. E., Jenkins, N. A., and Copeland, N. G. (2009). A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. *Cancer Res* **69**, 8150-6.
- Ebert, M. S., Neilson, J. R., and Sharp, P. A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat Methods* **4**, 721-6.
- Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494-8.
- Evans, M. J., and Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-6.
- Fazio, T. G., Huff, J. T., and Panning, B. (2008). An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. *Cell* **134**, 162-74.
- Ferguson, E. L., and Horvitz, H. R. (1989). The multivulva phenotype of certain *Caenorhabditis elegans* mutants results from defects in two functionally redundant pathways. *Genetics* **123**, 109-21.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806-11.

- Fischer, S. E., Wienholds, E., and Plasterk, R. H. (2001). Regulated transposition of a fish transposon in the mouse germ line. *Proc Natl Acad Sci U S A* **98**, 6759-64.
- Forstemann, K., Horwich, M. D., Wee, L., Tomari, Y., and Zamore, P. D. (2007). Drosophila microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. *Cell* **130**, 287-97.
- Freed, J. J., and Mezger-Freed, L. (1970). Stable haploid cultured cell lines from frog embryos. *Proc Natl Acad Sci U S A* **65**, 337-44.
- Friedrich, G., and Soriano, P. (1991). Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev* **5**, 1513-23.
- Friedrich, G., and Soriano, P. (1993). Insertional mutagenesis by retroviruses and promoter traps in embryonic stem cells. *Methods Enzymol* **225**, 681-701.
- Gangloff, S., McDonald, J. P., Bendixen, C., Arthur, L., and Rothstein, R. (1994). The yeast type I topoisomerase Top3 interacts with Sgs1, a DNA helicase homolog: a potential eukaryotic reverse gyrase. *Mol Cell Biol* **14**, 8391-8.
- Gaspar-Maia, A., Alajem, A., Polesso, F., Sridharan, R., Mason, M. J., Heidersbach, A., Ramalho-Santos, J., McManus, M. T., Plath, K., Meshorer, E., and Ramalho-Santos, M. (2009). Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863-8.
- German, J. (1964). Cytological Evidence for Crossing-over in Vitro in Human Lymphoid Cells. *Science* **144**, 298-301.
- German, J. (1993). Bloom syndrome: a mendelian prototype of somatic mutational disease. *Medicine (Baltimore)* **72**, 393-406.
- Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H. M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fectel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K. C., Cooney, A. J., D'Souza, L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521.
- Golic, K. G. (1991). Site-specific recombination between homologous chromosomes in Drosophila. *Science* **252**, 958-61.
- Goss, K. H., Risinger, M. A., Kordich, J. J., Sanz, M. M., Straughen, J. E., Slovek, L. E., Capobianco, A. J., German, J., Boivin, G. P., and Groden, J. (2002). Enhanced tumor formation in mice heterozygous for Blm mutation. *Science* **297**, 2051-3.

- Gottwein, E., Mukherjee, N., Sachse, C., Frenzel, C., Majoros, W. H., Chi, J. T., Braich, R., Manoharan, M., Soutschek, J., Ohler, U., and Cullen, B. R. (2007). A viral microRNA functions as an orthologue of cellular miR-155. *Nature* **450**, 1096-9.
- Gravel, S., Chapman, J. R., Magill, C., and Jackson, S. P. (2008). DNA helicases Sgs1 and BLM promote DNA double-strand break resection. *Genes Dev* **22**, 2767-72.
- Griffin, S., Branch, P., Xu, Y. Z., and Karran, P. (1994). DNA mismatch binding and incision at modified guanine bases by extracts of mammalian cells: implications for tolerance to DNA methylation damage. *Biochemistry* **33**, 4787-93.
- Grunwald, D. J., and Streisinger, G. (1992). Induction of recessive lethal and specific locus mutations in the zebrafish with ethyl nitrosourea. *Genet Res* **59**, 103-16.
- Guo, G. W., W; Bradley, A. (2004). Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells. *Nature* **429**, 891-895.
- Hamming, R. (1950). Error Detecting and Error Correcting Codes. *The Bell System Technical Journal* **29**, 147-161.
- Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* **18**, 3016-27.
- Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., Sohn, S. Y., Cho, Y., Zhang, B. T., and Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**, 887-901.
- Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y. K., Yeom, K. H., Yang, W. Y., Haussler, D., Belloch, R., and Kim, V. N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. *Cell* **136**, 75-84.
- Hansen, G. M., Markesich, D. C., Burnett, M. B., Zhu, Q., Dionne, K. M., Richter, L. J., Finnell, R. H., Sands, A. T., Zambrowicz, B. P., and Abuin, A. (2008). Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. *Genome Res.* **18**, 1670-1679.
- Hartwell, L. H., Culotti, J., Pringle, J. R., and Reid, B. J. (1974). Genetic control of the cell division cycle in yeast. *Science* **183**, 46-51.
- Hayakawa, T., Yusa, K., Kouno, M., Takeda, J., and Horie, K. (2006). Bloom's syndrome gene-deficient phenotype in mouse primary cells induced by a modified tetracycline-controlled trans-silencer. *Gene* **369**, 80-9.
- He, L., and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5**, 522-31.
- Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V. N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. *Mol Cell* **32**, 276-84.
- Hibbitt, O., and Wade-Martins, R. (2006). Delivery of large genomic DNA inserts >100 kb using HSV-1 amplicons. *Curr. Gene Ther.* **6**, 325-336.
- Hill, D. A., Ivanovich, J., Priest, J. R., Gurnett, C. A., Dehner, L. P., Desruisseau, D., Jarzembowski, J. A., Wikenheiser-Brokamp, K. A., Suarez, B. K., Whelan, A. J., Williams, G., Bracamontes, D., Messinger, Y., and Goodfellow, P. J. (2009). DICER1 Mutations in Familial Pleuropulmonary Blastoma. *Science* **325**, 965-.
- Hrabe de Angelis, M. H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M., Reis, A., Richter, T., Alessandrini, F., Jakob, T., Fuchs, E., Kolb, H., Kremmer, E., Schaeuble, K., Rollinski, B., Roscher, A., Peters, C., Meitinger, T., Strom, T., Steckler, T., Holsboer, F., Klopstock, T., Gekeler, F.,

- Schindewolf, C., Jung, T., Avraham, K., Behrendt, H., Ring, J., Zimmer, A., Schughart, K., Pfeffer, K., Wolf, E., and Balling, R. (2000). Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet* **25**, 444-7.
- Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T., and Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-8.
- Hutvagner, G., and Simard, M. J. (2008). Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol* **9**, 22-32.
- Hutvagner, G., and Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**, 2056-60.
- Ikeda, R., Kokubu, C., Yusa, K., Keng, V. W., Horie, K., and Takeda, J. (2007). Sleeping beauty transposase has an affinity for heterochromatin conformation. *Mol Cell Biol* **27**, 1665-76.
- Ivics, Z., Hackett, P. B., Plasterk, R. H., and Izsvák, Z. (1997). Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* **91**, 501-510.
- Ivics, Z., Li, M. A., Mates, L., Boeke, J. D., Nagy, A., Bradley, A., and Izsvak, Z. (2009). Transposon-mediated genome manipulation in vertebrates. *Nat. Meth.* **6**, 415-422.
- Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P. S. (2003). Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* **21**, 635-7.
- Jaenisch, R. (1976). Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. *Proc Natl Acad Sci U S A* **73**, 1260-4.
- Jang, S. K., Krausslich, H. G., Nicklin, M. J., Duke, G. M., Palmenberg, A. C., and Wimmer, E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* **62**, 2636-43.
- Jasin, M. (1996). Genetic manipulation of genomes with rare-cutting endonucleases. *Trends Genet* **12**, 224-8.
- Jiang, F., Ye, X., Liu, X., Fincher, L., McKearin, D., and Liu, Q. (2005). Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. *Genes Dev* **19**, 1674-9.
- Jiricny, J. (2006). The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* **7**, 335-46.
- Kaneda, M., Tang, F., O'Carroll, D., Lao, K., and Surani, M. A. (2009). Essential role for Argonaute2 protein in mouse oogenesis. *Epigenetics Chromatin* **2**, 9.
- Kanellopoulou, C., Muljo, S. A., Kung, A. L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D. M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev* **19**, 489-501.
- Karlas, A., Machuy, N., Shin, Y., Pleissner, K. P., Artarini, A., Heuer, D., Becker, D., Khalil, H., Ogilvie, L. A., Hess, S., Maurer, A. P., Muller, E., Wolff, T., Rudel, T., and Meyer, T. F. (2010). Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature* **463**, 818-22.
- Karsi, A., Moav, B., Hackett, P., and Liu, Z. (2001). Effects of insert size on transposition efficiency of the sleeping beauty transposon in mouse cells. *Mar. Biotechnol. (NY)* **3**, 241-5.

- Keller, G. M. (1995). In vitro differentiation of embryonic stem cells. *Curr Opin Cell Biol* **7**, 862-9.
- Keng, V. W., Villanueva, A., Chiang, D. Y., Dupuy, A. J., Ryan, B. J., Matise, I., Silverstein, K. A., Sarver, A., Starr, T. K., Akagi, K., Tessarollo, L., Collier, L. S., Powers, S., Lowe, S. W., Jenkins, N. A., Copeland, N. G., Llovet, J. M., and Largaespada, D. A. (2009). A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-74.
- Keng, V. W., Yae, K., Hayakawa, T., Mizuno, S., Uno, Y., Yusa, K., Kokubu, C., Kinoshita, T., Akagi, K., Jenkins, N. A., Copeland, N. G., Horie, K., and Takeda, J. (2005). Region-specific saturation germline mutagenesis in mice using the Sleeping Beauty transposon system. *Nat Methods* **2**, 763-9.
- Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209-16.
- Kile, B. T., and Hilton, D. J. (2005). The art and design of genetic screens: mouse. *Nat Rev Genet* **6**, 557-67.
- Kim, J. K., Gabel, H. W., Kamath, R. S., Tewari, M., Pasquinelli, A., Rual, J. F., Kennedy, S., Dybbs, M., Bertin, N., Kaplan, J. M., Vidal, M., and Ruvkun, G. (2005). Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308**, 1164-7.
- Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10**, 126-39.
- Kiriakidou, M., Tan, G. S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P. T., and Mourelatos, Z. (2007). An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* **129**, 1141-51.
- Kitamura, Y., Lee, Y. M., and Coffin, J. M. (1992). Nonrandom integration of retroviral DNA in vitro: effect of CpG methylation. *Proc Natl Acad Sci U S A* **89**, 5532-6.
- Kloosterman, W. P., Wienholds, E., de Bruijn, E., Kauppinen, S., and Plasterk, R. H. (2006). In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nat Methods* **3**, 27-9.
- Koike, H., Horie, K., Fukuyama, H., Kondoh, G., Nagata, S., and Takeda, J. (2002). Efficient biallelic mutagenesis with Cre/loxP-mediated inter-chromosomal recombination. *EMBO Rep* **3**, 433-7.
- Kokubu, C., Horie, K., Abe, K., Ikeda, R., Mizuno, S., Uno, Y., Ogiwara, S., Ohtsuka, M., Isotani, A., Okabe, M., Imai, K., and Takeda, J. (2009). A transposon-based chromosomal engineering method to survey a large cis-regulatory landscape in mice. *Nat Genet* **41**, 946-52.
- Kool, J., and Berns, A. (2009). High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. *Nat Rev Cancer* **9**, 389-99.
- Kotecki, M., Reddy, P. S., and Cochran, B. H. (1999). Isolation and characterization of a near-haploid human cell line. *Exp Cell Res* **252**, 273-80.
- Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15**, 8125-48.
- Kuehn, M. R., Bradley, A., Robertson, E. J., and Evans, M. J. (1987). A potential animal model for Lesch-Nyhan syndrome through introduction of HPRT mutations into mice. *Nature* **326**, 295-8.

- Kumar, M. S., Lu, J., Mercer, K. L., Golub, T. R., and Jacks, T. (2007). Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet* **39**, 673-677.
- Kumar, M. S., Pester, R. E., Chen, C. Y., Lane, K., Chin, C., Lu, J., Kirsch, D. G., Golub, T. R., and Jacks, T. (2009). Dicer1 functions as a haploinsufficient tumor suppressor. *Genes Dev* **23**, 2700-4.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
- Langland, G., Elliott, J., Li, Y., Creaney, J., Dixon, K., and Groden, J. (2002). The BLM helicase is necessary for normal DNA double-strand break repair. *Cancer Res* **62**, 2766-70.
- Langridge, G. C., Phan, M. D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., and Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* **19**, 2308-16.
- Lee, E. C., Yu, D., Martinez de Velasco, J., Tessarollo, L., Swing, D. A., Court, D. L., Jenkins, N. A., and Copeland, N. G. (2001). A Highly Efficient Escherichia coli-Based Chromosome Engineering System Adapted for Recombinogenic Targeting and Subcloning of BAC DNA. *Genomics* **73**, 56-65.
- Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-54.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415-9.
- Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J., and Carthew, R. W. (2004). Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117**, 69-81.
- Lefebvre, L., Dionne, N., Karaskova, J., Squire, J. A., and Nagy, A. (2001). Selection for transgene homozygosity in embryonic stem cells results in extensive loss of heterozygosity. *Nat Genet* **27**, 257-258.

- Li, M. A., Pettitt, S. J., Yusa, K., and Bradley, A. (2010). Genome-Wide Forward Genetic Screens in Mouse ES Cells. *Methods Enzymol* **477C**, 217-242.
- Li, P., Tong, C., Mehrian-Shai, R., Jia, L., Wu, N., Yan, Y., Maxson, R. E., Schulze, E. N., Song, H., Hsieh, C. L., Pera, M. F., and Ying, Q. L. (2008). Germline competent embryonic stem cells derived from rat blastocysts. *Cell* **135**, 1299-310.
- Liang, Q., Kong, J., Stalker, J., and Bradley, A. (2009). Chromosomal mobilization and reintegration of Sleeping Beauty and piggyBac transposons. *Genesis* **47**, 404-408.
- Liao, X., Buchberg, A. M., Jenkins, N. A., and Copeland, N. G. (1995). Evi-5, a common site of retroviral integration in AKXD T-cell lymphomas, maps near Gfi-1 on mouse chromosome 5. *J Virol* **69**, 7132-7.
- Lin, C. H., Jackson, A. L., Guo, J., Linsley, P. S., and Eisenman, R. N. (2009). Myc-regulated microRNAs attenuate embryonic stem cell differentiation. *Embo J* **28**, 3157-70.
- Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J. J., Hammond, S. M., Joshua-Tor, L., and Hannon, G. J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437-41.
- Liu, P., Jenkins, N. A., and Copeland, N. G. (2002). Efficient Cre-loxP-induced mitotic recombination in mouse embryonic stem cells. *Nat Genet* **30**, 66-72.
- Liu, Q., Rand, T. A., Kalidas, S., Du, F., Kim, H. E., Smith, D. P., and Wang, X. (2003). R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* **301**, 1921-5.
- Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., and Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-8.
- Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear export of microRNA precursors. *Science* **303**, 95-8.
- Luo, G., Ivics, Z., Izsvak, Z., and Bradley, A. (1998). Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* **95**, 10769-73.
- Luo, G., Santoro, I. M., McDaniel, L. D., Nishijima, I., Mills, M., Youssoufian, H., Vogel, H., Schultz, R. A., and Bradley, A. (2000). Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nat Genet* **26**, 424-9.
- Luria, S. E., and Delbruck, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491-511.
- Lykke-Andersen, K., Gilchrist, M. J., Grabarek, J. B., Das, P., Miska, E., and Zernicka-Goetz, M. (2008). Maternal Argonaute 2 is essential for early mouse development at the maternal-zygotic transition. *Mol Biol Cell* **19**, 4383-92.
- Maeder, M. L., Thibodeau-Beganny, S., Osiak, A., Wright, D. A., Anthony, R. M., Eichinger, M., Jiang, T., Foley, J. E., Winfrey, R. J., Townsend, J. A., Unger-Wallace, E., Sander, J. D., Muller-Lerch, F., Fu, F., Pearlberg, J., Gobel, C., Dassie, J. P., Pruett-Miller, S. M., Porteus, M. H., Sgroi, D. C., Iafrate, A. J., Dobbs, D., McCray, P. B., Jr., Cathomen, T., Voytas, D. F., and Joung, J. K. (2008). Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* **31**, 294-301.

- Maida, Y., Yasukawa, M., Furuuchi, M., Lassmann, T., Possemato, R., Okamoto, N., Kasim, V., Hayashizaki, Y., Hahn, W. C., and Masutomi, K. (2009). An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461**, 230-5.
- Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Meth.* **7**, 111-118.
- Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., Calabrese, J. M., Dennis, L. M., Volkert, T. L., Gupta, S., Love, J., Hannett, N., Sharp, P. A., Bartel, D. P., Jaenisch, R., and Young, R. A. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-33.
- Mates, L., Chuah, M. K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D. P., Schmitt, A., Becker, K., Matrai, J., Ma, L., Samara-Kuko, E., Gysemans, C., Pryputniewicz, D., Miskey, C., Fletcher, B., Vandendriessche, T., Ivics, Z., and Izsvak, Z. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet* **41**, 753-61.
- McDaniel, L. D., Chester, N., Watson, M., Borowsky, A. D., Leder, P., and Schultz, R. A. (2003). Chromosome instability and tumor predisposition inversely correlate with BLM protein levels. *DNA Repair (Amst)* **2**, 1387-404.
- McDonald, J. D., Bode, V. C., Dove, W. F., and Shedlovsky, A. (1990). Pahhph-5: a mouse mutant deficient in phenylalanine hydroxylase. *Proc Natl Acad Sci U S A* **87**, 1965-7.
- McLeod, M., Craft, S., and Broach, J. R. (1986). Identification of the crossover site during FLP-mediated recombination in the *Saccharomyces cerevisiae* plasmid 2 microns circle. *Mol Cell Biol* **6**, 3357-67.
- McMahon, A. P., and Bradley, A. (1990). The Wnt-1 (int-1) proto-oncogene is required for development of a large region of the mouse brain. *Cell* **62**, 1073-85.
- Melo, S. A., Roper, S., Moutinho, C., Aaltonen, L. A., Yamamoto, H., Calin, G. A., Rossi, S., Fernandez, A. F., Carneiro, F., Oliveira, C., Ferreira, B., Liu, C. G., Villanueva, A., Capella, G., Schwartz, S., Jr., Shiekhata, R., and Esteller, M. (2009). A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function. *Nat Genet* **41**, 365-70.
- Melton, C., Judson, R. L., and Blelloch, R. (2010). Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature* **463**, 621-6.
- Metzger, D., Clifford, J., Chiba, H., and Chambon, P. (1995). Conditional site-specific recombination in mammalian cells using a ligand-dependent chimeric Cre recombinase. *Proc Natl Acad Sci U S A* **92**, 6991-5.
- Moreno, M. A., Chen, J., Greenblatt, I., and Dellaporta, S. L. (1992). Reconstitutive mutagenesis of the maize P gene by short-range Ac transpositions. *Genetics* **131**, 939-56.
- Mortensen, R. M., Conner, D. A., Chao, S., Geisterfer-Lowrance, A. A., and Seidman, J. G. (1992). Production of homozygous mutant ES cells with a single targeting construct. *Mol Cell Biol* **12**, 2391-5.
- Moss, E. G., Lee, R. C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* **88**, 637-46.

- Munroe, R. J., Bergstrom, R. A., Y Zheng, Q., Libby, B., Smith, R., John, S. W. M., Schimenti, K. J., Browning, V. L., and Schimenti, J. C. (2000). Mouse mutants from chemically mutagenized embryonic stem cells. *Nat Genet* **24**, 318-321.
- Murchison, E. P., Stein, P., Xuan, Z., Pan, H., Zhang, M. Q., Schultz, R. M., and Hannon, G. J. (2007). Critical roles for Dicer in the female germline. *Genes Dev* **21**, 682-93.
- Murphy, E., Vanicek, J., Robins, H., Shenk, T., and Levine, A. J. (2008). Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: implications for latency. *Proc Natl Acad Sci U S A* **105**, 5453-8.
- Newsome, T. P., Asling, B., and Dickson, B. J. (2000). Analysis of Drosophila photoreceptor axon guidance in eye-specific mosaics. *Development* **127**, 851-60.
- Nolan, P. M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., Gray, I. C., Vizor, L., Brooker, D., Whitehill, E., Washbourne, R., Hough, T., Greenaway, S., Hewitt, M., Liu, X., McCormack, S., Pickford, K., Selley, R., Wells, C., Tymowska-Lalanne, Z., Roby, P., Glenister, P., Thornton, C., Thaug, C., Stevenson, J. A., Arkell, R., Mburu, P., Hardisty, R., Kiernan, A., Erven, A., Steel, K. P., Voegeling, S., Guenet, J. L., Nickols, C., Sadri, R., Nasse, M., Isaacs, A., Davies, K., Browne, M., Fisher, E. M., Martin, J., Rastan, S., Brown, S. D., and Hunter, J. (2000). A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat Genet* **25**, 440-3.
- Nurse, P. (1975). Genetic control of cell size at cell division in yeast. *Nature* **256**, 547-51.
- Nüsslein-Volhard, C., and Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795-801.
- Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., and Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**, 89-100.
- Pettitt, S. J., Liang, Q., Rairdan, X. Y., Moran, J. L., Prosser, H. M., Beier, D. R., Lloyd, K. C., Bradley, A., and Skarnes, W. C. (2009). Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat Methods* **6**, 493-5.
- Pham, J. W., Pellino, J. L., Lee, Y. S., Carthew, R. W., and Sontheimer, E. J. (2004). A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi in Drosophila. *Cell* **117**, 83-94.
- Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8.
- Prosser, H. M., Rzadzinska, A. K., Steel, K. P., and Bradley, A. (2008). Mosaic complementation demonstrates a regulatory role for myosin VIIa in actin dynamics of stereocilia. *Mol Cell Biol* **28**, 1702-12.
- Ramirez-Solis, R., Davis, A. C., and Bradley, A. (1993). Gene targeting in embryonic stem cells. *Methods Enzymol* **225**, 855-78.
- Ramirez-Solis, R., Liu, P., and Bradley, A. (1995). Chromosome engineering in mice. *Nature* **378**, 720-4.
- Rao, P. K., Toyama, Y., Chiang, H. R., Gupta, S., Bauer, M., Medvid, R., Reinhardt, F., Liao, R., Krieger, M., Jaenisch, R., Lodish, H. F., and Blelloch, R. (2009). Loss of cardiac microRNA-mediated regulation leads to dilated cardiomyopathy and heart failure. *Circ Res* **105**, 585-94.

- Reddy, S., DeGregori, J. V., von Melchner, H., and Ruley, H. E. (1991). Retrovirus promoter-trap vector to induce lacZ gene fusions in mammalian cells. *J Virol* **65**, 1507-15.
- Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901-6.
- Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. *Nature* **323**, 445-8.
- Rorth, P., Szabo, K., Bailey, A., Laverty, T., Rehm, J., Rubin, G. M., Weigmann, K., Milan, M., Benes, V., Ansorge, W., and Cohen, S. M. (1998). Systematic gain-of-function genetics in *Drosophila*. *Development* **125**, 1049-57.
- Rothstein, R., and Gangloff, S. (1995). Hyper-recombination and Bloom's syndrome: microbes again provide clues about cancer. *Genome Res* **5**, 421-6.
- Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-6.
- Rybak, A., Fuchs, H., Smirnova, L., Brandt, C., Pohl, E. E., Nitsch, R., and Wulczyn, F. G. (2008). A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. *Nat Cell Biol* **10**, 987-93.
- Schwartzberg, P. L., Goff, S. P., and Robertson, E. J. (1989). Germ-line transmission of a c-abl mutation produced by targeted gene disruption in ES cells. *Science* **246**, 799-803.
- Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P. D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199-208.
- Seoane, J., Le, H. V., and Massague, J. (2002). Myc suppression of the p21(Cip1) Cdk inhibitor influences the outcome of the p53 response to DNA damage. *Nature* **419**, 729-34.
- Shen, H., Suzuki, T., Munroe, D. J., Stewart, C., Rasmussen, L., Gilbert, D. J., Jenkins, N. A., and Copeland, N. G. (2003). Common sites of retroviral integration in mouse hematopoietic tumors identified by high-throughput, single nucleotide polymorphism-based mapping and bacterial artificial chromosome hybridization. *J Virol* **77**, 1584-8.
- Shigeoka, T., Kawaichi, M., and Ishida, Y. (2005). Suppression of nonsense-mediated mRNA decay permits unbiased gene trapping in mouse embryonic stem cells. *Nucleic Acids Res* **33**, e20.
- Silva, J. M., Li, M. Z., Chang, K., Ge, W., Golding, M. C., Rickles, R. J., Siolas, D., Hu, G., Paddison, P. J., Schlabach, M. R., Sheth, N., Bradshaw, J., Burchard, J., Kulkarni, A., Cavet, G., Sachidanandam, R., McCombie, W. R., Cleary, M. A., Elledge, S. J., and Hannon, G. J. (2005). Second-generation shRNA libraries covering the mouse and human genomes. *Nat Genet* **37**, 1281-8.
- Simon, M. A. (1994). Signal Transduction during the Development of the *Drosophila* R7 Photoreceptor. *Developmental Biology* **166**, 431-442.
- Simon, M. A., Bowtell, D. D., Dodson, G. S., Laverty, T. R., and Rubin, G. M. (1991). Ras1 and a putative guanine nucleotide exchange factor perform crucial steps in signaling by the sevenless protein tyrosine kinase. *Cell* **67**, 701-16.
- Simon, M. A., Dodson, G. S., and Rubin, G. M. (1993). An SH3-SH2-SH3 protein is required for p21Ras1 activation and binds to sevenless and Sos proteins in vitro. *Cell* **73**, 169-77.

- Simpson, E. M., Linder, C. C., Sargent, E. E., Davisson, M. T., Mobraaten, L. E., and Sharp, J. J. (1997). Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nat Genet* **16**, 19-27.
- Skarnes, W. C., Melchner, H., Wurst, W., Hicks, G., Nord, A., Cox, T., Young, S., Ruiz, P., Soriano, P., Tessier-Lavigne, M., Conklin, B. R., Stanford, W. L., and Rossant, J. (2004). A public gene trap resource for mouse functional genomics. *Nat Genet* **36**, 543-544.
- Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A., and Kucherlapati, R. S. (1985). Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**, 230-4.
- Snouwaert, J. N., Brigman, K. K., Latour, A. M., Malouf, N. N., Boucher, R. C., Smithies, O., and Koller, B. H. (1992). An animal model for cystic fibrosis made by gene targeting. *Science* **257**, 1083-8.
- Song, G., Sharma, A. D., Roll, G. R., Ng, R., Lee, A. Y., Blelloch, R. H., Frandsen, N. M., and Willenbring, H. (2010). MicroRNAs control hepatocyte proliferation during liver regeneration. *Hepatology* **51**, 1735-43.
- Stark, G. R., Kerr, I. M., Williams, B. R., Silverman, R. H., and Schreiber, R. D. (1998). How cells respond to interferons. *Annu Rev Biochem* **67**, 227-64.
- Starr, T. K., Allaei, R., Silverstein, K. A., Staggs, R. A., Sarver, A. L., Bergemann, T. L., Gupta, M., O'Sullivan, M. G., Matise, I., Dupuy, A. J., Collier, L. S., Powers, S., Oberg, A. L., Asmann, Y. W., Thibodeau, S. N., Tessarollo, L., Copeland, N. G., Jenkins, N. A., Cormier, R. T., and Largaespada, D. A. (2009). A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323**, 1747-50.
- Steiner, F. A., Hoogstrate, S. W., Okihara, K. L., Thijssen, K. L., Ketting, R. F., Plasterk, R. H., and Sijen, T. (2007). Structural features of small RNA precursors determine Argonaute loading in *Caenorhabditis elegans*. *Nat Struct Mol Biol* **14**, 927-33.
- Stocking, C., Kollek, R., Bergholz, U., and Ostertag, W. (1985). Long terminal repeat sequences impart hematopoietic transformation properties to the myeloproliferative sarcoma virus. *Proc Natl Acad Sci U S A* **82**, 5746-50.
- Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* **458**, 719-24.
- Su, H., Trombly, M. I., Chen, J., and Wang, X. (2009). Essential and overlapping functions for mammalian Argonautes in microRNA silencing. *Genes Dev* **23**, 304-17.
- Su, Q., Prosser, H. M., Campos, L. S., Ortiz, M., Nakamura, T., Warren, M., Dupuy, A. J., Jenkins, N. A., Copeland, N. G., Bradley, A., and Liu, P. (2008). A DNA transposon-based approach to validate oncogenic mutations in the mouse. *Proc Natl Acad Sci U S A* **105**, 19904-9.
- Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., and Blelloch, R. (2010). MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr Biol* **20**, 271-7.
- Sulston, J. E., and Horvitz, H. R. (1981). Abnormal cell lineages in mutants of the nematode *Caenorhabditis elegans*. *Dev Biol* **82**, 41-55.
- Sun, L. V., Jin, K., Liu, Y., Yang, W., Xie, X., Ye, L., Wang, L., Zhu, L., Ding, S., Su, Y., Zhou, J., Han, M., Zhuang, Y., Xu, T., Wu, X., Gu, N., and Zhong, Y. (2008). PBmice: an integrated

- database system of piggyBac (PB) insertional mutations and their characterizations in mice. *Nucleic Acids Res* **36**, D729-34.
- Sung, P., and Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nat Rev Mol Cell Biol* **7**, 739-750.
- Suster, M. L., Sumiyama, K., and Kawakami, K. (2009). Transposon-mediated BAC transgenesis in zebrafish and mice. *BMC Genomics* **10**, 477.
- Suzuki, H. I., Yamagata, K., Sugimoto, K., Iwamoto, T., Kato, S., and Miyazono, K. (2009). Modulation of microRNA processing by p53. *Nature* **460**, 529-33.
- Szymczak, A. L., Workman, C. J., Wang, Y., Vignali, K. M., Dilioglou, S., Vanin, E. F., and Vignali, D. A. A. (2004). Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat Biotech* **22**, 589-594.
- Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., and Hannon, G. J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-8.
- Taniguchi, M., Sanbo, M., Watanabe, S., Naruse, I., Mishina, M., and Yagi, T. (1998). Efficient production of Cre-mediated site-directed recombinants through the utilization of the puromycin resistance gene, pac: a transient gene- integration marker for ES cells. *Nucl. Acids Res.* **26**, 679-680.
- Thompson, L. H., Brookman, K. W., Jones, N. J., Allen, S. A., and Carrano, A. V. (1990). Molecular cloning of the human XRCC1 gene, which corrects defective DNA strand break repair and sister chromatid exchange. *Mol Cell Biol* **10**, 6160-71.
- Thomson, J. M., Newman, M., Parker, J. S., Morin-Kensicki, E. M., Wright, T., and Hammond, S. M. (2006). Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev* **20**, 2202-7.
- Tong, C., Li, P., Wu, N. L., Yan, Y., and Ying, Q. L. (2010). Production of p53 gene knockout rats by homologous recombination in embryonic stem cells. *Nature*.
- Tower, J., Karpen, G. H., Craig, N., and Spradling, A. C. (1993). Preferential transposition of Drosophila P elements to nearby chromosomal sites. *Genetics* **133**, 347-59.
- Troelstra, C., Odijk, H., de Wit, J., Westerveld, A., Thompson, L. H., Bootsma, D., and Hoeijmakers, J. H. (1990). Molecular cloning of the human DNA excision repair gene ERCC-6. *Mol Cell Biol* **10**, 5806-13.
- Trombly, M. I., Su, H., and Wang, X. (2009). A genetic screen for components of the mammalian RNA interference pathway in Bloom-deficient mouse embryonic stem cells. *Nucleic Acids Res* **37**, e34.
- Uren, A. G., Kool, J., Matentzoglou, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T., Reinders, M., Hubbard, T. J., Rogers, J., Jonkers, J., Wessels, L., Adams, D. J., van Lohuizen, M., and Berns, A. (2008). Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133**, 727-41.
- Uren, A. G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A. H., Wilson, C. H., Rance, R., Jonkers, J., van Lohuizen, M., Berns, A., and Adams, D. J. (2009). A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat Protoc* **4**, 789-98.

- Urlaub, G., Mitchell, P. J., Kas, E., Chasin, L. A., Funanage, V. L., Myoda, T. T., and Hamlin, J. (1986). Effect of gamma rays at the dihydrofolate reductase locus: deletions and inversions. *Somat Cell Mol Genet* **12**, 555-66.
- Valenzuela, D. M., Murphy, A. J., Frendewey, D., Gale, N. W., Economides, A. N., Auerbach, W., Poueymirou, W. T., Adams, N. C., Rojas, J., Yasenchak, J., Chernomorsky, R., Boucher, M., Elsasser, A. L., Esau, L., Zheng, J., Griffiths, J. A., Wang, X., Su, H., Xue, Y., Dominguez, M. G., Noguera, I., Torres, R., Macdonald, L. E., Stewart, A. F., DeChiara, T. M., and Yancopoulos, G. D. (2003). High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat. Biotech.* **21**, 652-659.
- Viswanathan, S. R., Daley, G. Q., and Gregory, R. I. (2008). Selective blockade of microRNA processing by Lin28. *Science* **320**, 97-100.
- Vitaterna, M. H., King, D. P., Chang, A. M., Kornhauser, J. M., Lowrey, P. L., McDonald, J. D., Dove, W. F., Pinto, L. H., Turek, F. W., and Takahashi, J. S. (1994). Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. *Science* **264**, 719-25.
- von Melchner, H., and Ruley, H. E. (1989). Identification of cellular promoters by using a retrovirus promoter trap. *J Virol* **63**, 3227-33.
- Vooijs, M., Jonkers, J., and Berns, A. (2001). A highly efficient ligand-regulated Cre recombinase mouse line shows that LoxP recombination is position dependent. *EMBO Rep* **2**, 292-7.
- Wade-Martins, R., White, R. E., Kimura, H., Cook, P. R., and James, M. R. (2000). Stable correction of a genetic deficiency in human cells by an episome carrying a 115 kb genomic transgene. *Nat. Biotech.* **18**, 1311-1314.
- Wallace, H. A. C., Marques-Kranc, F., Richardson, M., Luna-Crespo, F., Sharpe, J. A., Hughes, J., Wood, W. G., Higgs, D. R., and Smith, A. J. H. (2007). Manipulating the Mouse Genome to Engineer Precise Functional Syntenic Replacements with Human Sequence. *Cell* **128**, 197-209.
- Wang, W., and Bradley, A. (2007). A recessive genetic screen for host factors required for retroviral infection in a library of insertionally mutated Blm-deficient embryonic stem cells. *Genome Biology* **8**, R48.
- Wang, W., Bradley, A., and Huang, Y. (2008a). A piggyBac transposon-based genome wide library of insertionally mutated Blm deficient murine ES cells. *Genome Research* **19**, 667-673.
- Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A., and Liu, P. (2008b). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences* **105**, 9290-9295.
- Wang, Y., Baskerville, S., Shenoy, A., Babiarz, J. E., Baehner, L., and Blelloch, R. (2008c). Embryonic stem cell-specific microRNAs regulate the G1-S transition and promote rapid proliferation. *Nat Genet* **40**, 1478-83.
- Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet* **39**, 380-5.
- Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M. A., Sakaki, Y., and Sasaki, H. (2008).

- Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539-43.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyas, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62.
- Weinberg, R. A. (2006). "The biology of cancer." Garland Science,
- Withers-Ward, E. S., Kitamura, Y., Barnes, J. P., and Coffin, J. M. (1994). Distribution of targets for avian retrovirus DNA integration in vivo. *Genes Dev* **8**, 1473-87.
- Woltjen, K., Michael, I. P., Mohseni, P., Desai, R., Mileikovsky, M., Hamalainen, R., Cowling, R., Wang, W., Liu, P., Gertsenstein, M., Kaji, K., Sung, H.-K., and Nagy, A. (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458**, 766-770.
- Wright, D. A., Townsend, J. A., Winfrey, R. J., Jr., Irwin, P. A., Rajagopal, J., Lonosky, P. M., Hall, B. D., Jondle, M. D., and Voytas, D. F. (2005). High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J* **44**, 693-705.
- Wu, L., and Hickson, I. D. (2003). The Bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature* **426**, 870-4.
- Wu, S., Ying, G., Wu, Q., and Capecchi, M. R. (2007). Toward simpler and faster genome-wide mutagenesis in mice. *Nat Genet* **39**, 922-30.
- Wu, S. C., Meir, Y. J., Coates, C. J., Handler, A. M., Pelczar, P., Moisyadi, S., and Kaminski, J. M. (2006). piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proc Natl Acad Sci U S A* **103**, 15008-13.
- Xu, T., and Rubin, G. M. (1993). Analysis of genetic mosaics in developing and adult *Drosophila* tissues. *Development* **117**, 1223-37.
- Xue, Y., Bai, X., Lee, I., Kallstrom, G., Ho, J., Brown, J., Stevens, A., and Johnson, A. W. (2000). *Saccharomyces cerevisiae* RAI1 (YGL246c) is homologous to human DOM3Z and encodes a protein that binds the nuclear exoribonuclease Rat1p. *Mol Cell Biol* **20**, 4006-15.
- Yamagata, K., Kato, J., Shimamoto, A., Goto, M., Furuichi, Y., and Ikeda, H. (1998). Bloom's and Werner's syndrome genes suppress hyperrecombination in yeast *sgs1* mutant:

- implication for genomic instability in human diseases. *Proc Natl Acad Sci U S A* **95**, 8733-8.
- Yant, S. R., Wu, X., Huang, Y., Garrison, B., Burgess, S. M., and Kay, M. A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell Biol* **25**, 2085-94.
- Yekta, S., Shih, I. H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304**, 594-6.
- Yi, M., Hong, N., and Hong, Y. (2009a). Generation of medaka fish haploid embryonic stem cells. *Science* **326**, 430-3.
- Yi, R., Pasolli, H. A., Landthaler, M., Hafner, M., Ojo, T., Sheridan, R., Sander, C., O'Carroll, D., Stoffel, M., Tuschl, T., and Fuchs, E. (2009b). DGCR8-dependent microRNA biogenesis is essential for skin development. *Proc Natl Acad Sci U S A* **106**, 498-502.
- Yigit, E., Batista, P. J., Bei, Y., Pang, K. M., Chen, C. C., Tolia, N. H., Joshua-Tor, L., Mitani, S., Simard, M. J., and Mello, C. C. (2006). Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127**, 747-57.
- You-Tzung, C., and Allan, B. (2000). A new positive/negative selectable marker, puDeltatk, for use in embryonic stem cells. *genesis* **28**, 31-35.
- You, Y., Bergstrom, R., Klemm, M., Lederman, B., Nelson, H., Ticknor, C., Jaenisch, R., and Schimenti, J. (1997). Chromosomal deletion complexes in mice by radiation of embryonic stem cells. *Nat Genet* **15**, 285-8.
- Yusa, K., Horie, K., Kondoh, G., Kouno, M., Maeda, Y., Kinoshita, T., and Takeda, J. (2004a). Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene. *Nature* **429**, 896-899.
- Yusa, K., Rad, R., Takeda, J., and Bradley, A. (2009). Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat. Meth.* **6**, 363-369.
- Yusa, K., Takeda, J., and Horie, K. (2004b). Enhancement of Sleeping Beauty transposition by CpG methylation: possible role of heterochromatin formation. *Mol Cell Biol* **24**, 4004-18.
- Zakharov, A. F., and Egolina, N. A. (1972). Differential spiralization along mammalian mitotic chromosomes. I. BUdR-revealed differentiation in Chinese hamster chromosomes. *Chromosoma* **38**, 341-65.
- Zhang, E. E., Liu, A. C., Hirota, T., Miraglia, L. J., Welch, G., Pongsawakul, P. Y., Liu, X., Atwood, A., Huss, J. W., 3rd, Janes, J., Su, A. I., Hogenesch, J. B., and Kay, S. A. (2009). A genome-wide RNAi screen for modifiers of the circadian clock in human cells. *Cell* **139**, 199-210.
- Zijlstra, M., Li, E., Sajjadi, F., Subramani, S., and Jaenisch, R. (1989). Germ-line transmission of a disrupted beta 2-microglobulin gene produced by homologous recombination in embryonic stem cells. *Nature* **342**, 435-8.