Chapter 1

Introduction

1.1 Reverse Genetics

Reverse genetics refers to a gene-driven approach, which analyses the phenotypic consequences of directed mutations of a target gene. Emerged in the end of 1980s, reverse genetics set off since the development of transgenic organisms, where a genotype is designed, constructed *in vitro* and introduced to the mouse germline. The resulting transgenic animals display a phenotype dependent on the location and design of the mutation, which allows characterisation of a gene and eventual understanding of its underlying biology [213]. Since the 1990s, a large number of genes have been cloned and their knockout animals were generated. After the completion of mouse genome project, high-throughput reverse genetics became the major approach [151]. In this section, I will describe different strategies in reverse genetics, from homologous recombination (HR)-mediated modifications to RNA interference, and finally, the use of programmable nucleases, including zinc finger nucleases (ZFNs), transcription activator-like effectors nucleases (TALENs) and the revolutionary clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 technology.

1.1.1 Gene targeting

Gene targeting, defined as the introduction of site-specific modification into the genome by HR, has enabled genetic manipulations ranging from simple gene disruptions and point mutations to insertions and inversions, even conditional knockouts or knockins [32] [284]. The principle of gene targeting was developed in yeast, where DNA fragments with homology to yeast DNA sequence can integrate into its genome [306]. The first HR-based mutagenesis in mammalian cells was achieved by Smithies et al. showing successful integration by HR of a plasmid into the *beta*-globin locus of human erythroleukaemia cells [386]. In parallel with Smithies's work, Capecchi and colleagues independently achieved HR-mediated repair of a defective neomycin resistant gene in transformed mouse fibroblast cell lines [415].

Subsequent to the derivation and establishment of mouse embryonic stem cell (ESC) culture and the demonstration of its germ line transmission capacity [107] [34], the first HR in mouse ESCs was achieved at the selectable hypoxanthine phosphorybosyl transferase (Hprt) gene locus [94][414]. The targeting of non-selectable genes such as *Int-2* and *c-Abl* later also became possible [254][365]. These early experiments have provided a basis for the generation of genetically modified mice, which became invaluable tools for understanding

the functions of mammalian genes at the organism level and producing models for human diseases.

A targeting vector typically consists of three basic portions: a 5' homology arm, a positive selectable marker such as neomycin resistance gene (*neo*) and a 3' homology arm. Successfully targeted cells are positively selected by neomycin (G418) or other antibiotics such as puromycin or hygromycin depending on the selectable marker used. HR clones can be further enriched using the negative selection markers, such as the diphtheria toxin fragment A(DT-A) or thymidine kinase(TK), which are placed outside the homology arms. If the targeting vector is integrated randomly in the genome, the negative selection marker is also most likely to be retained and exerts toxicity, which will lead to cell death. Site-specific recombinases such as Cre or FLP are routinely used to remove the selection marker genes to leave minimal impact at the locus. The expression of these recombinases can be controlled in a time- and/or tissue- dependent manner using the Tamoxifen-ERT2 system and tissue specific promoter, respectively, which are particular useful for a detailed study of genes whose inactivation would be otherwise lethal.

1.1.2 RNA interference (RNAi)

RNAi is defined as the process of suppression of the expression of a target gene via specific destruction of its mRNA by exogenous or endogenous double stranded RNA (dsRNA) [100]. Its ability of gene silencing in a sequence-specific manner has made it a powerful tool for investigating the function of a gene. The RNAi pathway was first discovered by Fire and Montgomery in 1998 when they injected long dsRNA into *Caenorhabditis elegans* and observed specific cytoplasmic degradation of the mRNA molecules containing the same sequence as the injected dsRNA [114] [280]. The use of long dsRNA was soon proven to be effective in flies and plants but not in mammalian cells due to a nonspecific interferon response [379] [389]. Subsequently, it was discovered that the use of 21-28 nucleotide short interfering RNAs (siRNAs) or short hairpin shRNAs (shRNAs) can prevent the interferon response and achieve sequence-specific silencing in mammalian cells [102] [43].

To function as a silencing effector, siRNAs need to be processed from longer precursors by a member of the RNase III family called Dicer, in an ATP-dependent manner [18]. Processed siRNAs are typically 21-23 nucleotides long, which are subsequently loaded onto a group of Agonaute proteins called RNA-Inducing Silencing Complex (RISC) with some help from Dicer [149] [103]. The loaded RNA duplex then unwind itself, leaving

one strand as a guide for target recognition, whereas the other passenger strand gets discarded [478]. When the siRNA-guided RISC reaches its complementary RNA target, an endonucleolytic cleavage is induced by the PIWI domain of RISC at the 10th nucleotide counting from the 5' end [478]. Following the initial cut, the mRNA target dissociates from the siRNA, and cellular exonucleases join in to complete the degradation process [149]. It has been observed that certain imperfect matches between siRNA and mRNAs can be tolerated, which causes the off-target effect [357].

siRNAs can be chemically synthesised and introduced into mammalian cells directly by transfection. The transfected siRNAs bypass the dicing step and directly incorporate into RISC for target mRNA degradation [357]. However, synthesised siRNAs have several drawbacks such as the expensive chemical synthesis process and low transfection efficiency in certain cell types. Furthermore, the siRNA molecules are unstable and become diluted as cell divide, which cannot constitutively sustain stable gene knockdown. To circumvent these problems, a plasmid-based system was developed, where siRNAs are expressed in the form of short hairpin RNAs (shRNAs) [43]. Once the expression cassette is introduced into the cell, shRNA is constitutively transcribed by RNA polymerase III promoter and forms a stem-loop structure, which is processed by Dicer and other RNAi-related machineries to into a 20-25 nucleotides double-stranded siRNA [43]. The processed siRNAs can then be loaded to RISC and carry out target mRNA degradation as described [43]. As an alternative method to plasmid transfection, retro- and lentiviral transduction are frequently used for shRNA delivery [459] [446] [42]. The transduction efficiency can be optimised to a high level, to nearly 100% in some cell types [100]. Furthermore, it was shown that the shRNA constructs can be designed to be embedded in the context of endogenous miRNA precursor sequence, improved the knockdown efficiency up to 12-fold higher [88] [376].

1.1.3 Genome engineering with programmable nucleases

Conventional gene targeting via HR is a powerful approach to achieve gene inactivation and enable gene function interrogation. However, it is usually a tedious process given that the efficiency of HR is extremely low in higher eukaryotic cells, which lead to the need for the labour-intensive and time-consuming selection/screening procedure. A study using a rare-cutting endonuclease, I-SceI, showed that the gene targeting efficiency increased by more than 2 orders of magnitude with the expression of endonuclease [337]. This observation provided the first evidence that HR is stimulated by the introduction of DNA double-

stranded break (DSB). From there onwards, targeted genome engineering became widely adopted, which allows precise and efficient genome editing via inducing a site-specific DSB followed by generation of desired modifications during subsequent DNA repair.

1.1.3.1 The repair pathways and applications

DNA DSBs are potentially lethal to cells. Generally, they are repaired via one of the two major mechanisms: non-homologous end joining (NHEJ) or homology directed repair (HDR). NHEJ-mediated DSB repair involves direct ligation of the broken ends, which is error-prone, and often results in small insertions or deletions (indels). HDR is a template dependent pathway, which allows perfect restoration of the broken ends. Thus HDR has been exploited to achieved genetic modifications such as targeted gene insertion, correction and point mutation.

Gene disruption

Gene disruption can be achieved by the error-prone repair pathway NHEJ. Indels generated by NHEJ often give rise to frameshifts in the protein coding region, which result in premature termination followed by non-sense mediated decay, and the final consequence of gene knockout.

Gene addition or tag ligation

By co-transfecting a site-specific nuclease with a targeting vector bearing locus-specific homology arms, the transgene can be efficiently incorporated into the desired site. Alternatively, using specific nuclease that generates defined overhangs, large transgenes (up to 200kb) can be inserted into the targeted loci via NHEJ-mediated ligation.

Point mutation or gene correction

Targeting vectors or single-strand oligodeoxynucleotides (ssODNs) can be co-delivered with programmable nucleases to correct point mutations or introduce single-nucleotide variations via the HDR-mediated repair pathway. Compared to the use of targeting vector, ssODNs are much simpler in design and can be synthesised in a few days. Such advance in technology has greatly enhanced the efficiency of disease modelling, and could potentially be applied in cell and gene therapy.

Chromosomal rearrangement

Two simultaneous DSBs made on the same chromosome can result in chromosomal deletion, inversion, duplication or other rearrangements. If DSBs are introduced on two different chromosomes, chromosomal translocation can be achieved, which opens up oppor-

tunities for creating models for genetic defects caused by large chromosomal rearrangements.

1.1.3.2 Programmable nucleases before the CRISPR era

1.1.3.2.1 ZFN

Although the study performed by Rouet et al. with I-SceI demonstrated improved targeting efficiency, type II restriction enzymes are not suitable for introducing unique DSBs in eukaryotic genomes due to their short recognition sites. A novel restriction endonuclease with longer recognition site, preferably 16-18bp in length, is required for the general use of gene targeting in eukaryotes. ZFN was the first programmable nuclease that demonstrated the potential to cleave any arbitrary DNA sequences [197]. It is composed of a DNA binding domain, which is adapted from the prevalent class of eukaryotic transcription factor – zinc finger proteins (ZFPs), and a nuclease domain derived from the restriction enzyme FokI.

The versatility of ZFN arises from its DNA binding module ZFPs, which typically contains a tandem array of Cys2-His2 fingers [273]. Each zinc finger (ZF) is composed of approximately 30 amino acid residues folded to a unique ββα structure that is stabilised by a zinc ion. The crystal structure suggested that ZF binds DNA by inserting its α-helix into the major groove of the double helix and recognises a 3bp sequence via making contact of the amino acids within the α-helix and their target 3 nucleotides [312].

This modular structure has made ZFP a suitable component for the design of custom DNA binding protein. Facilitated by the discovery of a highly conserved linker sequence, researchers were able to generate ZFPs for a specific DNA sequence by identifying individual ZF modules for each triplet component and link them together. However, it soon became clear that the recognition of DNA by ZFs is not truly independent or modular, and that each ZF's activity is largely influenced by its neighbours [167][447][274]. To circumvent the constraints of simple modular assembly, strategies to generate context dependence of ZF modules, such as oligomerized pool engineering (OPEN) and context-dependent assembly (CoDA), were developed [244][349].

FokI was identified as a desirable subunit for generating programmable nuclease because its sequence recognition domain and endonuclease domain are structurally separated. This provides an opportunity for swapping the recognition domain with other DNA-binding proteins [59]. The FokI nuclease domain must form a dimer to cleave DNA [26], there-

fore two ZFNs are required to bind adjacent sites with appropriate spacing for efficient dimerisation. The dimerisation process increased DNA binding stringency, resulting in increased specificity. However, off-target cleavages can still arise from the homodimerisation of FokI monomers [275] [398]. To increase specificity further, FokI domain was engineered to cleave only as a heterodimerised pair [275] [398].

ZFN was first applied to *Drosophila melanogaster* [23][22]. Since then it has been used to modify endogenous genes in a wide range of organisms such as frog oocytes, mice, rats, plants, zebrafish and netatodes, as well as cultured cells such as human cancer cell line, mouse ES cells, human ES cells, and human iPS cells [21][230] [133] [51] [268] [283] [97] [139] [234] [493]. Furthermore, ZFN has also been applied in the development of novel therapy. The first clinical trial using ZFN to target the *CCR5* gene in T cells from HIV-infected patients has already been completed in 2014 [407].

1.1.3.2.2 TALEN

After more than a decade of research and development on ZFNs, the discovery of a simpler modular DNA recognition protein, namely transcription activator-like effectors (TALEs), provided an alternative platform as a customisable endonuclease for genome editing. TALEs are naturally occurring proteins from the plant pathogenic bacteria *Xanthomonas* [27]. The ability of this proteins to bind DNA was first discovered in 2007 [336]. The binding process is mediated by an array of highly conserved 33-35 amino acid repeats, each of which recognises a single base pair in the major groove. The nucleotide specificity of each repeat is determined by two hypervariable amino acids positioned at 12 and 13, which are named as repeat variable diresidues (RVDs) [245] [82].

The discovery of TALEs attracted great interests in the field, and its DNA recognition code was deciphered shortly afterwards [28] [145]. Four different RVD residues NN, NI, HD and NG are the most widely used for the recognition of G, A, C and T, respectively [179]. Subsequently, chimeric TALE nucleases (TALENs) were generated by combining the TALE-based DNA recognition domain and the FokI nuclease domain [74]. Like ZFNs, TALENs work as pairs with the DNA binding sites designed to locate 12-25bp apart [74].

The one-to-one correspondence of TALE-DNA binding repeats provided greater design flexibility than triplet-confined ZFNs, which renders TALENs to be designed to target almost any given DNA sequences [195]. With a comparable targeting efficiency to

ZFNs, TALENs seem to be an easier option for non-specialist researchers [348] [333] [157]. However, due to the extensive identical repeats, expression vector construction could be challenging. To overcome this problem, several strategies have been developed such as the 'Golden Gate' cloning system, high-throughput solid-phase assembly and ligation-independent cloning techniques [54] [333] [37] [362].

1.1.3.3 CRISPR-Cas systems

1.1.3.3.1 The discovery of CRISPR-Cas systems The CRISPR repeats were first identified in *Escherichia coli* in a study of *iap* enzyme in 1987. Ishino et al. observed an unusual structure of five highly homologous sequences of 29 nucleotides arranged as direct repeats with non-repetitive 32 nucleotides interspacing [168]. The biological significance of such structure remained elusive at the time. Over a decade later, Mojica et al. reported the wide spread of such short regularly spaced repeats among prokaryotic genomes [278]. Subsequently, Jansen et al. named these short repeats as clustered regularly interspaced short palindromic repeats (CRISPR) and identified the CRISPR-associated (Cas) genes [172].

In 2005, three groups independently published results showing the homology of CRISPR spacers with extrachromosomal elements such as phages and plasmids [279] [324] [29]. This observation, together with the evidence of the correlation between phage resistance and the CRISPR spacers, suggested that the acquisition of CRISPR elements may be related to foreign DNA invasion and CRISPR may function as a bacterial adaptive immune system [279] [29] [247]. This hypothesis was soon proved by Barrangou et al., who have demonstrated that the removal or addition of particular CRISPR spacers modified the phage resistance phenotype of the bacteria [12]. The natural mechanism of CRISPR system as part of the adaptive immunity is shown in Figure 1.1.

1.1.3.3.2 The diversity of CRISPR-Cas systems The CRISPR-Cas system has been classified into six types based on the configuration of their effector modules [248] [452] [373]. The six types can be further grouped into two major classes. Type I, III, and IV are considered as Class 1 system, where the targeted cleavage requires several Cas proteins and crRNAs to form an effector complex [248] [452]. Type II, V, and VI are grouped into the Class 2 system, where all functions of the effector complex are carried out by a single RNA-guided endonuclease, such as Cas9 [248] [452]. The utilisation of a single-component effector protein makes the Class 2 system a favourable choice to exploit

in genome editing applications. Among all the subtypes in the class 2 system, Cas9 from type II has be the most extensively characterised effector and widely utilised as a genome engineering tool.

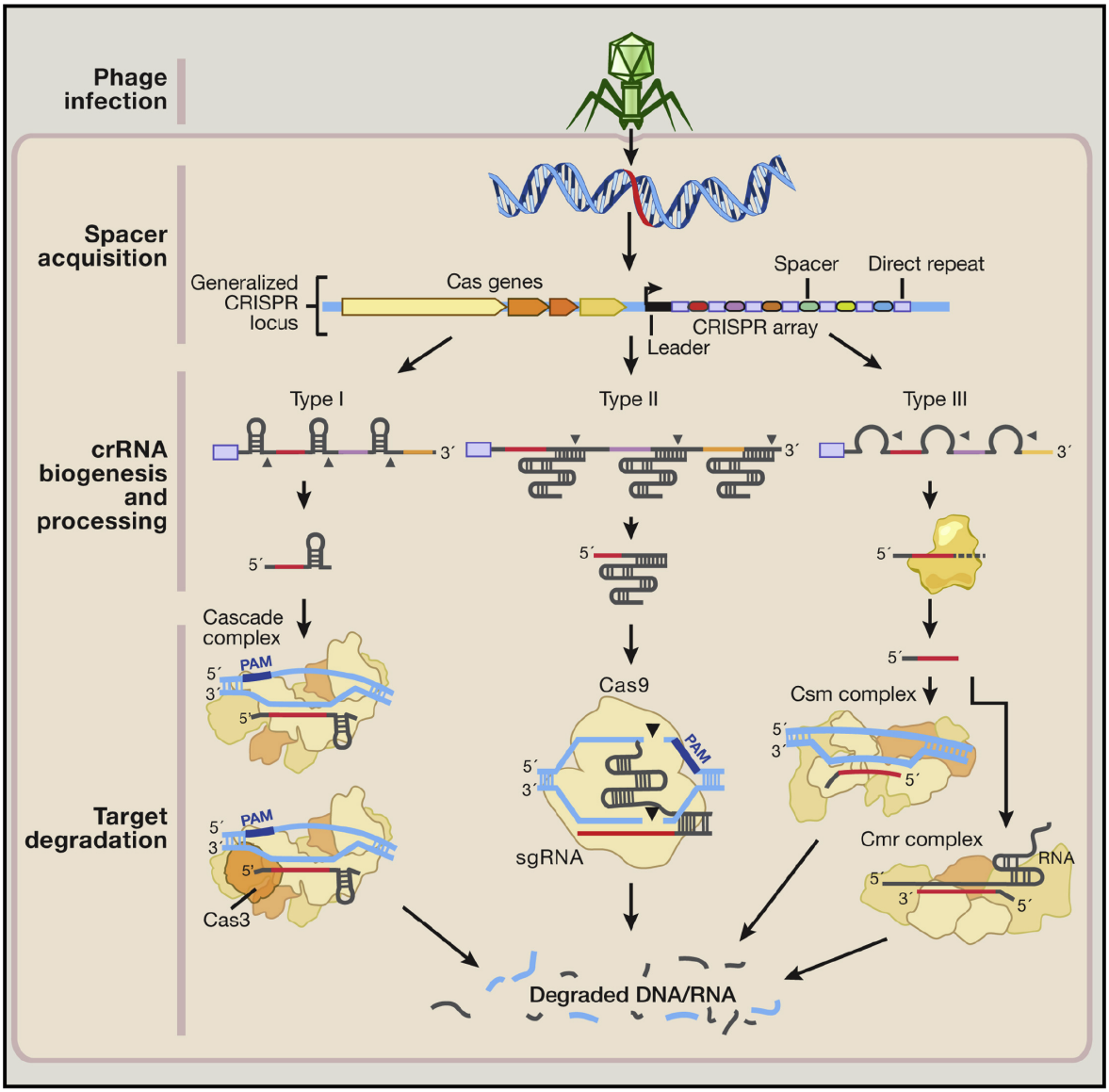

Figure 1.1: CRISPR-mediated DNA interference in microbial adaptive immunity. A typical CRISPR locus comprises a set of Cas9 genes, a unique noncoding RNA called the trans-activating CRISPR RNA (tracrRNA), and an array of repetitive sequences interspaced by a range of non-repetitive sequences referred as spacers. Following the invasion of foreign genetic elements from bacteriophages or plasmids, the Cas enzymes acquire new spacers into the exogenous protospacer sequences and install them into the CRISPR locus within the prokaryotic genome. The crRNA biogenesis and processing follow distinct pathways in each type of CRISPR system. In type I and III CRISPR systems, the pre-crRNA transcript is cleaved within the repeats and further processed to produce matured crRNA before being loaded onto effector proteins complexes for target recognition and degradation. In type II system, tracrRNA hybridises with the direct repeats which then gets cleaved and processed by RNase III and other nucleases. The processed crRNA-tracrRNA hybrid forms a complex with Cas9 to degrade DNA matching its guide RNA sequence. Image is taken and adapted from Hsu et al., 2014 [160].

1.1.3.3.3 CRISPR-Cas9 system as a genome-editing tool As the CRISPR field started to attract more interests, researchers soon unraveled more details about the molecular mechanisms of the CRISPR-Cas system. Brouns et al. showed that the spacer sequences were transcribed, cleaved by the CRISPR RNA nuclease and act as guide RNAs [39]. Marraffini et al. demonstrated for the first time that the Cas protein was able to target DNA directly [255]. Moineau and colleagues showed that the *Streptococcus thermophilus* CRISPR1/Cas system cleaves plasmid and bacteriophage double-stranded DNA at specific sites within the proto-spacer squence [129]. Subsequently, more studies were carried out and in particular, molecular mechanisms of the type II CRISPR system were extensively characterised. Charpentier and colleagues identified the trans-encoded small RNA (tracrRNA), which was required for the maturation of CRISPR RNA (crRNA) and Cas9 loading [81] (Figure 1.3 (B)). Soon afterwards, Siksnys and colleagues published detailed biochemical characterisation of Cas9-mediated cleavages. Most importantly, they demonstrated that Cas9 can be programmed to a specific target site by changing the sequence of the crRNA [130]. Like Siksnys, Doudna and Charpentier's groups also showed that the Cas9-crRNA-tracrRNA complex could cleave purified DNA *in vitro*. They also demonstrated that Cas9 could be programmed with a custom-designed crRNA to cut at a target site of their choosing. Furthermore, they showed that both crRNA and tracrRNA were required for Cas9 to function and the two RNAs could be fused as a single guide RNA (sgRNA), which has become a widely accepted concept used in genome-editing [176]. Skilsnys, Doudna and Charpentier's work unleashed the potential of the universal programmable RNA-guided DNA endonuclease, and is considered profoundly important in the field of genome editing (Figure 1.2). The final highlight of the adaptation of CRISPR-Cas9 system as a genome-editing tool is the successful demonstration of targeted genome cleavage in mammalian cells by Zhang and colleagues [78]. Similar findings were published in the same issue of Science by the Church group and in Nature Biotechnology by the Joung group and the Kim group[250] [164] [71]. Since then, the CRISPR-Cas9 system has been widely adopted by the scientific community to edit genomes of a wide range of cells and organisms.

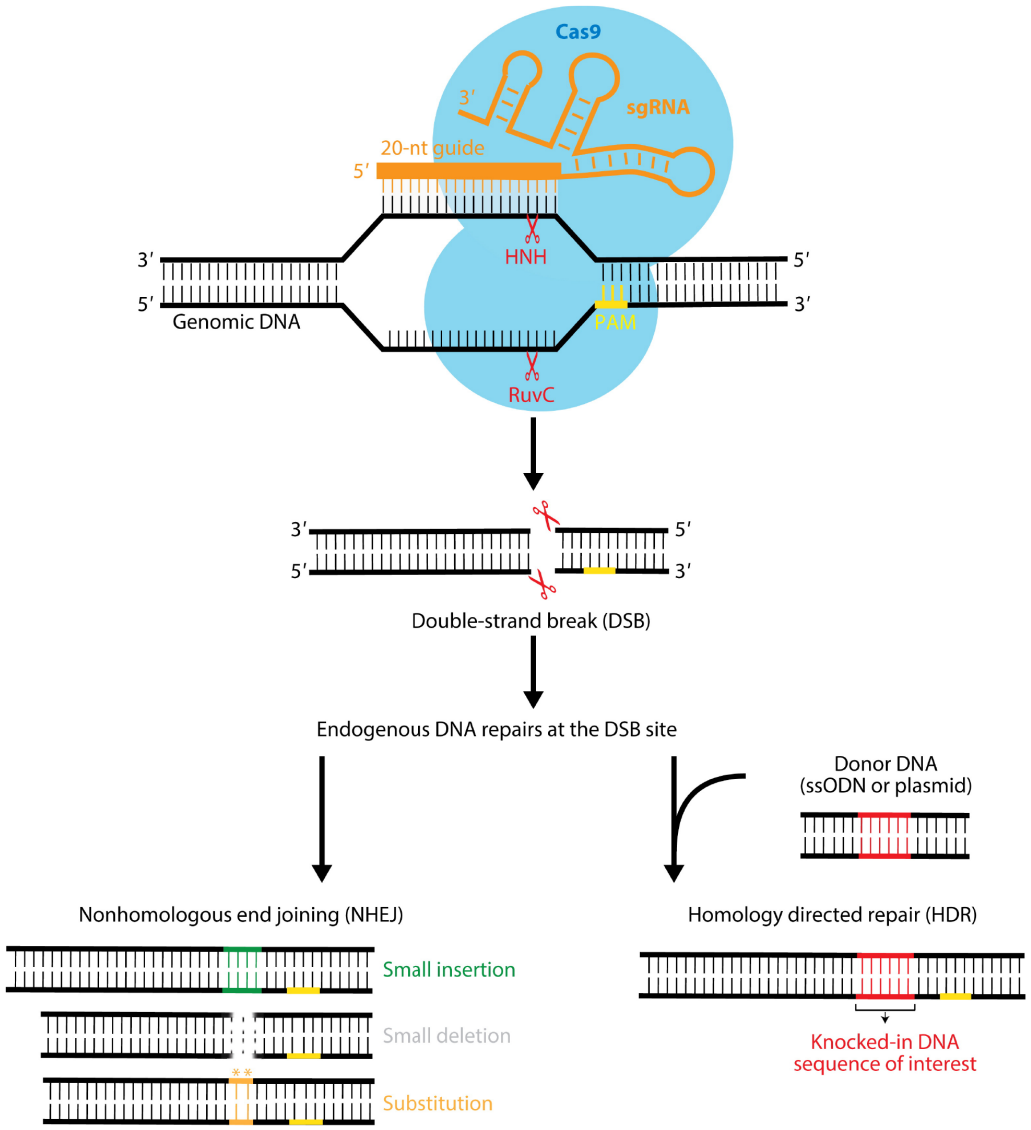

Figure 1.2: The mechanism of CRISPR-mediated genome engineering. To perform gene-editing, a guide RNA can be designed and constructed by fusing a crRNA containing the targeting sequence to a tracrRNA that facilitates DNA cleavage by Cas9. Binding of the PAM sequence and a matching target triggers Cas9 nuclease activity which allows it to produce a DSB 3bp upstream of the PAM site. DNA DSBs are typically repaired by NHEJ or HDR. In the error-prone NHEJ pathway, indels are introduced, frequently lead to the disruption of gene function. In the presence of a donor template, HDR pathway can be initiated to create desired mutations through homologous recombination, which allows precise gene modification such as gene knock-in and base correction. Image is taken and adapted from Jiang and Doudna, 2017 [174].

1.1.3.3.4 Structure and working mechanism of Cas9 The Cas9 is a bilobed enzyme consists of the α-helical recognition (REC) lobe and the nuclease (NUC) lobe [177] [293]. The REC lobe is essential for gRNA and DNA binding, whereas the NUC lobe is composed of the RuvC-like nuclease domain, HNH nuclease domain and a carboxyl-terminal domain (CTD) (Figure 1.3 (A)) [177] [293]. Jinek et al. and Gasiunas et al. have independently showed that the RuvC domain of the Cas9 enzyme cleaves the non-complementary strand while the HNH domain cleaves the complementary site [176] [130]. The carboxyl-terminal domain is responsible for the interaction with the protospacer adjacent motif (PAM) [293]. It was shown that the Cas9 enzyme adopts an auto-inhibited confirmation in the absence of gRNA. Upon RNA loading, the two lobes of Cas9 undergo a conformational rearrangement and form a central channel, which accommodates the RNA-DNA heteroduplex [177].

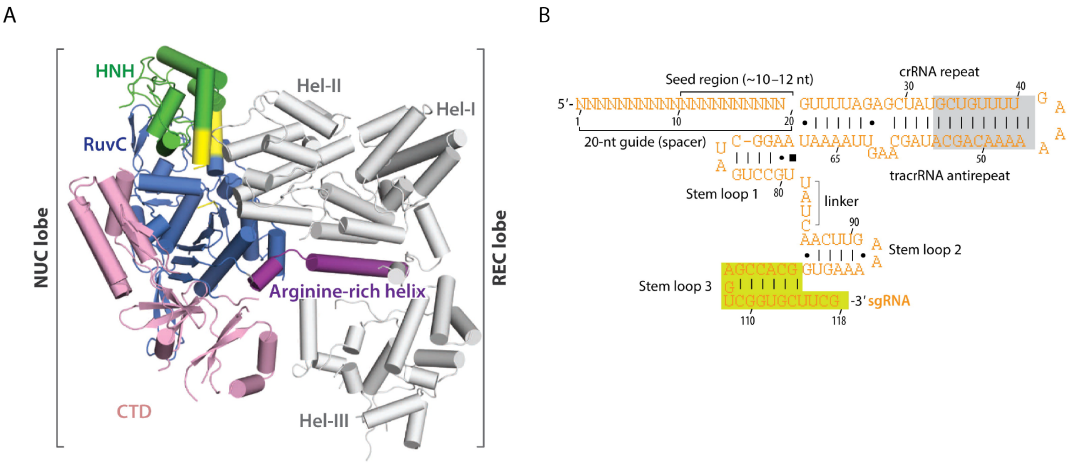

Figure 1.3: Overall structure of *Streptococcus pyogenes* Cas9 and gRNA. (A) Ribbon representation of the crystal structure of Cas9 protein. The nuclease (NUC) lobe contains the HNH and RuvC nuclease domains as well as the more variable C-terminal domain (CTD). The REC lobe consists of three alpha-helical domains (Hel-I, Hel-II, and Hel-III) and no structural similarity was found with other known proteins. (B) Schematic representation of the sgRNA secondary structure. Grey box denotes the extra repeat?antirepeat region in the full-length gRNA scaffold. Yellow box indicates the stem loop 3 of sgRNA, which is not essential for Cas9 function and was truncated in the sgRNA-bound structure. Image is taken and adapted from Jiang and Doudna, 2017 [174].

The PAM sequence plays an important role in target binding. While the exact PAM sequence is dependent on the bacteria or archea species, 5' NGG 3' is the PAM site for the most widely used *Streptococcus pyogenes* Cas9 [78] [161]. Biochemical studies showed that Cas9 uses PAM recognition to identify target sites while scanning large DNA molecules [391] . Furthermore, competition assay suggested that DNA unwinding and RNA-DNA heteroduplex formation initiated at the PAM site and proceeded towards the distal 5' end

of the target sequence [391]. The crystal structure of the REC lobe also indicated that the eight PAM-proximal nucleotides in the Cas9-bound gRNA were exposed for base-pairing with target DNA, which supports the theory that the 8-12nt PAM proximal 'seed' region is critical for target DNA recognition [293] [176] [78] [126] [161] [311] [249].

1.1.3.3.5 Off-target effect Because genome-editing results in permanent alterations within the genome, the off-target effect of Cas9 nuclease is of paticular concern, especially for clinical applications. A series of studies have investigated this issue using mismatched gRNA libraries, *in vitro* selection and reporter assays to monitor the ratio of off-target cleavage frequency [126] [161] [311] [80] [72] [249]. Beyond the previous knowledge that mismatches at the 3' 'seed' region are less tolerated, these studies further demonstrated that the overall off-target effect depends on the number, position, and distribution of mismatches within the protospacer sequence [126] [161] [311] [249]. In addition, it was shown that the 'NGG' PAM sequence was not absolutely required, as the 'NAG' PAM can also be tolerated, albeit at a lower frequency. Therefore, it is worth considering both NGG and NAG PAM sequences in off-target analysis. Several groups have designed algorithms to select gRNAs with minimal off targets based on these findings [158] [161] [9] [61]. Furthermore, the amount of Cas9 enzyme expressed in the cell will also affect the off-target effect. It was reported that high Cas9 concentration increases the chance of off-targets [126]. To improve specificity, Cas9 was converted into a nickase by mutating either the HNH or RuvC domain. The double-nicking strategy can then be adopted using a pair of gRNAs and Cas9 nickase. Such method is based on the hypothesis that two simultaneous adjacent off-target binding and cleavage is much more unlikely than a single cleavage. It was shown that Cas9 nickase improved targeting specificity by up to 1500 times compare to wild-type Cas9 [329] [369] [249]. In addition to the usage of double nickase, shorter gRNA truncated by two or three nucleotides at the distal end could also reduce off-target activity, potentially due to greater mismatch sensitivity [127].

1.1.3.3.6 Applications of CRISPR-Cas9 system The CRISPR-Cas9-mediated genome editing system has been broadly used in gene function studies, disease modelling and potentially, gene therapy. In addition to DSB-based genome editing, Cas9 nucleases was engineered into RNA-guided DNA binding protein by mutating the RuvC and HNH nuclease domains [326]. This nuclease-deficient Cas9 (dCas9) can be fused with functional effector domains such as transcriptional activators, suppressors and chromatin modifiers.

The CRISPR-dCas9-based transcriptional and epigenetic regulators allow both loss-of-function and gain-of-function perturbations precisely and rapidly without major disruption of the local genomic architecture, which supplements the wild type CRISPR-Cas9 knockout function. Collectively, wild-type Cas9 and dCas9-mediated transcriptional and epigenetic regulators form a complete toolbox for comprehensive genomic study from all directions.

Genome editing

The use of CRISPR-Cas9 platform has greatly accelerated the efficiency of generating cellular models as well as transgenic organisms. For generating cellular models, Cas9 and gRNAs can be easily introduced into the cell via transient plasmid transfection, viral transduction, or as ribonucleoproteins (RNPs). For generating transgenic animal models, it was shown that Cas9 protein and gRNA could be directed injected into fertilised zygotes, which bypassed the ESC targeting stage and shortened the generation time for mutant mice to only several weeks[462] [431]. In addition, the multiplexing capability of Cas9 provides a possibility for studying polygenic diseases, such as diabetes, schizophrenia and heart disease. Furthermore, many studies have reported using CRISPR-Cas9 system to correct disease-related mutations *in vitro* [465] [221] [364] [222] [457]. In 2016, the US National Institute of Health (NIH) approved the first clinical trial involving the use CRISPR-Cas9 to modify T cells from cancer patients. Although still a long way to reach the clinic, CRISPR-Cas9 technology holds prominent potential in treating inherit genetic disease, infectious diseases and cancer.

Transcriptional modulation

It was shown that dCas9 alone can repress gene expression by blocking transcriptional elongation, RNA polymerase binding and transcription factor binding, which is a phenomenon referred as CRISPR interference (CRISPRi) [326]. However, the knockdown effect by dCas9 itself is not effective in mammalian cells (2-fold) [326]. Gilbert et al. demonstrated that dCas9 fused to KRAB domain can specifically and stably repress endogenous genes in human cells [136]. Similarly, Konermann et al. showed that dCas9-SID4X can mediate repression of Sox2 in 293 cells [204]. Although both studies demonstrated effective dCas9-mediated gene repression, the current CRISPRi needs further improvements as the knockdown effect is still partial.

Similar to CRISPRi, dCas9 can be fused to transcription activating domains, such as VP64 and p65AD to create the effect of CRISPR-mediated gene activation (CRISPRa)

[249] [204] [136] [317]. In addition to direct fusion of dCas9 with activating domains, gRNAs can be further engineered to include protein-interacting RNA aptamers for the recruitment of RNA binding proteins fused to functional effectors such as VP64 [477]. These scaffold RNAs can be used as a defined sets to generate synthetic multi-gene transcriptional programs to rewire cell fates or engineer metabolic pathways [477]. To enhance the efficiency of CRISPRa, Joung and colleagues expressed a dCas9-VP64 fusion protein and multiple gRNAs and showed dose dependent synergistic activation [243]. Similarly, Jaenisch and colleagues created the CRISPR-on system with dCas9 fused with VP48 and showed that the a cluster of 3-4 gRNAs could achieve more efficient gene induction [70]. Alternatively, other groups have developed strategies which exploit the synergy of multiple transcriptional activators. Chavez et al. generated a tripartite activator, with dCas9 fused to VP64-p65-Rta, and used to direct neuronal differentiation from iPS cells [62]. Another example is the synergistic activation mediator (SAM) system, which consists of dCas9-VP64 and gRNA-MS2, which selectively recruit MCP fused with activating domains of p65 and HSF1 [205]. Finally, Gilbert et al. and Tanenbaum et al. developed a protein scaffold called SunTag and applied to recruit multiple copies of VP64 activator modules to a single activating site [406] [135].

Epigenetic control

Epigenetic modifications are crucial for controlling gene expression. Similar to transcriptional regulation, CRISPR-dCas9 can be used to recruit epigenetic modifiers and reshape the epigenome at a defined locus. Hilton et al. designed a programmable CRISPR-dCas9-based acetyltransferase by fusing dCas9 with the catalytic core of acetyltransferase p300. dCas9-p300 catalysed H3K27 acetylation at its target sites, resulting in specific transcriptional activation of genes from targeted enhancers [155]. Another study showed that the dCas9-LSD1 fusion can efficiently suppress the expression of genes controlled by the targeted enhancers [189]. Thakore et al. demonstrated that dCas9-KRAB is able to disrupt the activity of targeted enhancer via induction of H3K9me3 [411]. Furthermore, Liu et al. established a system where targeted DNA demethylation and methylation can be achieved by dCas9-Tet1 and dCas9-Dnmt3a, respectively. The dCas9 epigenetic effectors allow both loss-of-function and gain-of-function perturbations precisely and rapidly without major disruption of the local genomic architecture [229].

1.2 Forward genetics

In contrast to reverse genetics, forward genetics starts with isolation of mutants that show a particular phenotype and works to identify the causative genetic change. The most fundamental advantage of forward genetics is the unbiased nature of inquiry, which requires no hypothesis nor the molecular basis of the biological process to be studied, making it a powerful approach for new and unexpected discoveries [282].

Long before the post-genomic era, forward genetics was the approach in genetic research as most of the studies were based on the observations of a particular phenotype. Many genes were even named after their mutant phenotype. The phenotypes observed at the time were mostly caused by spontaneous mutations, and the subsequent identification of the causative genes usually involves positional cloning and mapping [282]. Examples include the wingless (wg) gene from Drosophila and obese (Ob) gene from mouse [368] [486]. Since spontaneous mutations arise at a rate insufficient to perform systematic genetic studies, accelerating the rate of mutagenesis became the pressing need.

Artificial mutagenesis can be achieved by exposing organisms to physical agents such as irradiation, chemical mutagens such as ENU, or biological mutagens such as the transposon system. These mutagens are able to introduce random mutations in a cell or organism at high efficiency, providing opportunities to perform loss-of-function genetic screens. A genetic screen is an experimental approach to identify individuals that exhibit the phenotype of interest in a mutant pool. Driven by genome sequence data, technologies such as RNA interference (RNAi) and most recently the CRISPR-Cas9 system allow scientists to generate multiplexed libraries and perform genome-wide screens in a targeted high-throughput manner, giving genetic screens the unprecedented power to discover novel genetic functions and biological pathways.

1.2.1 Mutagenesis using chemical and physical agents

1.2.1.1 N-ethyl-N-nitrosourea (ENU)

N-ethyl-N-nitrosourea (ENU) is alkylating agent widely utilised for generating mutations in the mouse genome[340]. Given that ENU is able to induce nonsense and missense mutations, it is possible to generate null, hypomorphic and hypermorphic alleles, which diversifies the consequent phenotypes and widens its application [181]. ENU induces mutations at a relatively high efficiency. It has been shown that the mutation rate of ENU is approximately 1 in 1000 in mouse gametes and 1 in 200 in mouse ESCs[156].

Most of the ENU-induced mutations has been identified in the coding region or intro-exon boundaries, which makes it great for studying gene functions, but not ideal for the regulatory regions [193].

1.2.1.2 Irradiation

Ionizing radiation (IR) including X-rays and γ-rays is a powerful mutagen which typically generates large deletions and complex chromosomal rearrangements such duplications, translocations and inversions. An X-ray-induced forward mutation study in Chinese hamster cells performed by Chu provided detailed analysis of the non-linear relationship between induced mutation rate and the dose of X-ray exposure [75]. It was found that X-irradiation could generate mutations at a rate of $13 - 50 \times 10^{-5}$ per locus, which is a lower yield of mutations than chemical mutagens [339] [334].

Although ENU and irradiation are a powerful tools for genome-wide genetic screens, the difficulty in identifying the causative point mutation remains the biggest hurdle. The typical strategy for causative gene identification involves genetic mapping to narrow the region containing the mutation to a manageable size, followed by DNA sequencing. To perform genetic mapping, the affected mice need to be out-crossed to a different bred strain and the resulted F1 progeny are then backcrossed or intercrossed, which can take up to 40 weeks and involves up to hundreds of animals [193] [282]. With the development and popularisation of next-generation sequencing (NGS) technology, as well as the availability of ready-made genomic fragments in vectors such as bacterial artificial chromosomes (BACs) and the complementary DNA (cDNA) library, this process could be simplified and shortened to up to two weeks, albeit still highly labour-intensive and resource-dependent [282] [416].

1.2.2 Insertional mutagenesis

Insertional mutagenesis refers to the method in which an exogenous DNA integrates into the host genome and causes disruption or alteration of genes nearby the insertion site. There are mainly two categories of insertional mutagens, namely retroviral vectors and DNA transposons. Both of the vectors are flexible with the 'cargo' they accommodate, and can be constructed to carry various molecular elements based on the experimental design. Unlike ENU and irradiation, which have the biggest bottleneck as the identification of causative mutations, insertional mutagenesis involves integration of a piece of DNA whose identity is known, which can be used as a molecular tag to identify mutated genes.

1.2.2.1 Retroviral-mediated mutagenesis

Retroviral insertional mutagenesis is an experimental approach for gene discovery, taking advantage of the retrovirus life cycle: the proviral DNA integrates into the genome and results in gene expression alteration. Integration of proviruses can introduce both loss-of-function and gain-of-function mutations. The former are resulted when the provirus inserts into exogenic regions, whereas the later are produced when the enhancer element in the long terminal repeat (LTR) region of the retrovirus drives aberrant gene expression. There are several limitations to the use of retroviral-mediated mutagenesis. Most critically, retroviruses exhibit strong preferences for integration sites. It has been shown that retroviruses have both 'hot' and 'cold' insertion spots and preferentially target the 5' end of expressed genes [276] [455]. Additionally, retroviruses carry strong enhancers in their LTRs that can deregulate genes located hundreds of kbs away, which can complicate the identification of causative gene [287].

1.2.2.2 DNA transposon-mediated mutagenesis

DNA transposons are mobilised in a non-replicative 'cut and paste' manner and has been developed as a widely used molecular tool for insertional mutagenesis [79]. Among various categories of DNA transposons, Sleeping beauty (SB) and piggyBac (PB) are more popular choices as methods of genome editing [200] [383] [171].

SB is a reconstructed DNA transposon from fossil fragments found in the salmon genome. It exclusively integrates into a TA dinucleotide, which is duplicated upon integration and flank the transposed element [171] [428]. The SB transposon can be mobilised from either an exogenous plasmid or a donor site on the chromosome, and every excision made by the SB transposase leaves a 3 bp 'footprint' [240]. PB is a moth-derived transposon system active in a wide range of organisms. Unlike SB, PB recognises a short TTAA sequence for insertion and excises the transposon without a footprint, which makes it a more precise and defining system [122]. Because PB can excise precisely from the donor site, it is especially useful when a transgene is transiently required, for example, in generating transgene-free iPSCs [449] [476]. One of the examples of PB-mediated genetic screen was a study performed by Rad and colleagues, in which PB was applied as a tool for genome-wide mutagenesis in mice, and many novel cancer genes were uncovered [327]. Both SB and PB demonstrate a strong tendency of local-hopping, meaning that the excised transposon preferentially integrates near its original location, which is unfavourable in conducting

genome-wide genetic screens [240] [99] [116].

1.2.3 The use of *Blm*-deficient and haploid cell lines in genetic screens

1.2.3.1 *Blm*-deficient ESC systems

Recessive genetic screens in mammalian systems were hampered by their diploid nature. Although homozygous mutant organisms could be generated by breeding, there was no efficient approach to induce homozygous mutations in cultured cells, until the establishment of the Bloom's syndrome protein (*Blm*)-deficient ESCs [241]. Bloom syndrome is a recessive genetic disorder associated with genomic instability and cancer-prone phenotypes [140]. It has been demonstrated that *Blm*-deficient cells have increased incidence of homologous recombination which led to the increased loss of heterozygosity [241] [140]. Performing recessive screen on *Blm*-deficient genetic background increases the chance of recovering biallelic mutations. Two groups independently exploited this phenotype and conducted genetic screens in *Blm*-deficient ESCs. In one study performed by Yusa et al., a tetracycline-induced *Blm* ESC line was combined with ENU mutagenesis to successfully identify genes that are involved in the glycosylphosphatidylinositol (GPI)-anchor synthesis [475]. In the other approach, Guo et al. applied the retrovirus gene-trap system in *Blm*-deficient ESCs to select genes in the mismatch repairs pathway [144].

1.2.3.2 Haploid cell lines

Another technical advance in overcoming the challenge of recessive screens in eukaryotic systems is the establishment of haploid cell lines. The first haploid human cell line (KBM7) was isolated from a chromic myeloid leukaemia patient in 1999, but it was not until a decade later its use as a tool for genetic screen was showcased [208] [52]. The mouse haploid ESC lines was generated in 2011 independently by two groups [104] [219]. Analysis revealed that these cells exhibited typical mouse ESC morphology and expressed pluripotency markers including Oct4, Klf4, Sox2, Nanog and Rex1 [104] [219]. Both of the studies demonstrated the utility of haploid ESCs for genetic screens. Leeb et al. conducted a pilot screen for mismatch repair genes in the presence of 6-TG using the gene trap PB transposon vector, and successfully identified *Msh2* and *Hprt* [219]. He and others then proceeded to a large-scale genome-wide screen with the haploid mutant library to study the exit of pluripotency and successfully recovered novel pluripotency regulators *Zfp706* and *Pum1* [217]. Elling et al. generated a haploid mutant library with retrovirus and challenged it with toxin ricin. As a result, they identified multiple genes involved in

ricin processing pathways, some of which had never been reported before [104]. These results illustrated the potential of haploid cells for large-scale forward genetic screens. However, haploid cells are unstable, and cells often undergo auto-diploidization, which requires regular selection such as cell sorting to maintain the haploid nature of the cell culture. Additionally, due to the limitation of the derivation process, genetic screens using haploid cells are only limited to a few cell types.

1.2.4 RNAi-mediated screens

The first large-scale genetic screen using siRNAs in mammalian cells was performed in 2003 to study the mechanism of TRAIL-induced apoptosis. Aza-Blanc et al. transfected 510 siRNAs targeting 510 genes in HeLa cells and used AlamarBlue as a cell viability read out [7]. The screen has successfully uncovered several modulator of TRAIL-induced apoptosis including DOBI and MIRSA [7]. Zerial and colleagues performed the first genome-wide transfected siRNA screen in combination with automated imaging analysis. The authors discovered a number of kinases involved in endocytosis, suggesting that signalling functions are built into the machinery of endocytosis [313]. The Bernards group developed the first large-scale virus-based shRNA library, which contained approximately 23,000 shRNAs targeting around 8000 human genes. It was used in a pooled infection screen, from which the authors could identify one known and five new modulators contributing to the p53-dependent proliferation arrest [16]. One complication about this study was that each of the individual clone contained several shRNA inserts, which required further analysis to identify the shRNA responsible for the observed phenotype.

Compared to the methods described previously, RNAi was the first technology that supports a fully-controlled targeted screen. However, with the classical genetic approaches, one can plan for gain-of-function screens or identify mutations that are not in the coding regions, which is limited in RNAi [96]. In addition, large discrepancy were often observed between the results of a similar RNAi-based screen performed by several groups [35] [206] [488]. This is probably due to the false-positive hits resulted from the off-target effect of siRNAs. Therefore, secondary screens are often necessary to identify the true hits, and the phenotype needs to be verified by a second independent siRNA targeting the same transcript. Furthermore, siRNAs almost never completely deplete the target mRNA, which often results in false negatives [96].

1.2.5 CRISPR-Cas9-mediated screens

1.2.5.1 The establishment of CRISPR-Cas9 screening technology

Soon after the successful adaptation of CRISPR-Cas9 as a genome editing tool in mammalian cells, three groups independently generated genome-wide gRNA libraries and performed functional genetic screens with Cas9 nuclease. The Yusa lab constructed a genome-wide library with 87,897 gRNAs targeting 19,150 mouse genes and applied the resulted library in a recessive screens to identify genes that modulate susceptibility to *Clostridium speticum* alpha-toxin and 6-TG [202]. As a result, all known essential components of the GPI-anchor biosynthesis pathway has been identified, together with 13 genes whose function in alpha-toxin resistance had not been reported. Analysis of the 6-TG resistance screen revealed all known factors including four MMR genes and *Hprt*. Similarly, the Zhang group designed a genome-scale CRISPR-Cas9 knockout (GeCKO) library with 64,751 gRNAs targeting 18,080 human genes and successfully identified essential genes in cancer and pluripotent stem cell lines [366]. The authors also demonstrated the use of GeCKO library for positive selections, which uncovered both known and novel genes whose loss conferred resistance to vemurafenib in melanoma cell lines [366]. The Sabatini/Lander group built a library with 73,151 gRNAs targeting 7114 genes and 100 non-targeting gRNAs as control. With this library, they screened for genes that function in the DNA MMR pathway in the presence of 6TG using the haploid cell line KBM7, and identified genes encoding four components of the MMR pathway [434]. They also screened for resistance to etoposide in diploid cell line HL60 and revealed hits including TOP2A as well as CDK6 whose role in this pathway was previously unknown [434]. Those proof-of-principle studies demonstrated the power of CRISPR-Cas9-mediated genome-wide screens and uncovered its potential to address a wide-range of biological questions.

Compared to the other mutagenesis methods described earlier in this chapter, the CRISPR-Cas9 system allows high-throughput gene knockout in a targeted manner with pre-defined cutting sites. Importantly, the Cas9-mediated mutagenesis exhibited high bi-allelic mutation efficiency, which is essential for its application in the mammalian systems. In addition, it is straightforward to identify the causative mutations as gRNA itself can serve as a molecular barcode. Direct comparison suggested that the CRISPR-Cas9 system outperformed RNAi, which is also a reprogrammable, easy-to-perform genome-editing technology [202]. This is probably due to the fact that RNAi can almost never achieve complete suppression of gene silencing and its off-target effect often complicate the analysis and results in poor

consistency. Using a set of 'gold standard' essential and non-essential genes as targeting controls, Evers et al. showed that CRISPR outperformed shRNA-based system with lower noise, better consistency and lower off-target effect [108]. Similarly, Munoz et al showed that CRISPR-based screens consistently identify more lethal genes than RNAi in cancer cell lines, indicating lower false-negative rate [285].

1.2.5.2 Screening format

1.2.5.2.1 Arrayed screening

CRISPR-Cas9-mediated screen can be carried out in two formats: arrayed or pooled format. The choice depends on the experimental aim. Arrayed screening is usually carried out in multi-well plates with each well containing one or a few gRNAs targeting a single gene. Two major advantages are associated with arrayed screens. First, the causative mutation can be easily identified as the constituents of each well are known. Second, given that each well has a single known genetic perturbation, it allows the investigation of a much wider range of phenotypes such as high-content imaging [367]. Recently, Metzakopian et al. generated the first two individually-cloned CRISPR-Cas9 genome-wide arrayed gRNA libraries with a complexity of 34,332 gRNAs for human and 40,860 gRNAs for mouse genome, covering 17,166 human and 20,430 mouse genes [272]. These libraries expanded the toolbox for comprehensive gene editing and offered an opportunity to perform screens at a single-gene level. However, arrayed screens are labour-intensive and time-consuming, as reagents have to be prepared individually. Accessibility to automated robotic equipment is often necessary for plate handling and can be its limitation on wider usage.

1.2.5.2.2 Pooled screening

In contrast to arrayed screens, pooled screening is usually less expensive and do not require highly automated equipment. Although direct phenotypic assessment is limited for each gRNA and a more careful experimental design is needed, pooled screening is a powerful and fast approach for systematically investigating plenty of biological questions. The simplicity and high efficiency of the CRISPR-Cas9 technology make it an ideal system for pooled functional genomic screen. The *in-silico* designed gRNA libraries can be chemically synthesised as a pool of oligonucleotides, which is subsequently cloned into plasmid vectors, usually lentivirus expressing vectors. A mutant cell library can be generated from transduction of lentivirus library, followed by the application of selection stress. Mutations causing the phenotype of interest can be identified by next generation sequencing based

on the representation of gRNAs.

A pooled genetic screen can be designed for either positive or negative selection of gRNAs. Positive selection screening identifies genes that are enriched after applying the selection pressure. It is most commonly used to identify perturbations that confer resistance to a toxin, inhibitor, drug, or pathogen [367]. For example, in the screen aiming to identify genes involved in the GPI-anchor biosynthesis pathway, most mutants transduced with irrelevant gRNAs were depleted from the population due to the susceptibility to alpha-toxin, whereas cells transduced with gRNAs targeting genes involved in the GPI-anchor biosynthesis pathway lacked the cellular receptor, and became resistant to alpha-toxin, thereby getting enriched after selection [202]. Positive selections usually produce clearer results as the expected hits are few. In contrast, negative selection screening is to identify genes that are depleted during the selection process. The most typical negative selection screen is the one to identify essential or fitness genes, that are are required for cell survival and/or proliferation. After a certain period of cell culture, mutants transduced with gRNAs targeting essential genes will deplete or 'drop out' from the population. Compared to positive selection screening, negative selection screening is more technically challenging and require higher screening sensitivity due to the fact that level of depletion can be limited.

1.2.5.3 Applications of CRISPR-Cas9-mediated screens

The CRISPR-based screening technology has provided a great opportunity for systematic identification of essential genes. Wang et al. constructed a genome-wide gRNA library to screen for genes required for proliferation and survival in the near-haploid KBM7 cell line [433]. The screen results were validated by an orthogonal retroviral gene-trap screen and benchmarked with functional profiling in yeast. As a result, the authors were able to uncover a group of uncharacterised essential genes in various cellular pathways [433]. Hart et al. designed and generated a second-generation CRISPR knockout library referred to as The Toronto KnockOut (TKO) library and applied it to screen for essential genes in a range of cell lines derived from different parts of the human body [153]. With this approach, they could identify five-times more fitness genes than previously described in shRNA screens, and were able to classify the 'core' fitness genes and context-specific fitness genes, which provided insights to the biological differences between cell types [153].

CRISPR-Cas9 loss-of-function screening has also been applied to the non-coding region

of the genome. Canver et al. developed a screen using tiling gRNA library for saturated mutagenesis of non-coding elements *in situ*, which provided insights into the function and organisation of the BCL11A enhancer [50]. Similar studies have been performed by others to analyse the regulatory regions of *NF1*, *NF2* and *CUL3* loci, *POU5F1* locus and TP53 and ESCR1 transcription binding sites [350] [207] [86]. In addition, the Gersbach group designed the CRISPR-Cas9-based epigenomic regulatory element screen (CERES), which utilises the dCas9-KRAB repressor and dCas9-p300 activator to target the DNase I hypersensitive sites around genes of interest. Both loss- and gain-of-function screens were conducted for the *β-blobin* and *HER2* loci, which revealed known and previously unidentified regulatory elements [198]. Similarly, another study performed by Fulco et al. exploited the CRISPRi library to screen for regulatory elements of *MYC* and *GATA1*, and identified nine distal enhancers that control gene expression and cellular proliferation. Not limited to the discovery of novel enhancer elements, the CRISPR-based screens were also utilised to study the function of lncRNA using cell growth as a readout [491] [228]. Zhu et al. constructed a paired-guide RNA library and uncovered 51 lncRNA that positively or negatively regulated cell growth [491]. Using a CRISPRi-based gRNA library, Liu et al. identified approximately 500 functional lncRNAs out of 17,000 screened [228]. These studies demonstrated the potential of CRISPR-Cas9 screens to unravel the functions of non-coding genome.

Although the most of the CRISPR-Cas9 screens were performed in *in vitro* systems, it has also been applied *ex vivo* in primary dendritic cells to study regulators of the bacterial lipopolysaccharide response [310]. Similarly, Chen et al. performed CRISPR-Cas9 screen in mice to study tumour growth and metastasis [65]. Recently, Manguso et al. performed an *in vivo* screen using a library encoding 9872 gRNAs targeting 2368 genes to identify genes that synergise with or cause resistance to PD-L1 checkpoint blockade [251].

The development of dCas9 opened up alternative options to conventional knockout screens. The Weissman lab constructed a genome-wide library targeting the transcription regulatory regions of approximately 16,000 genes and applied it with a dCas9-KRAB fusion protein to achieve CRISPR interference (CRISPRi) screens [135]. CRISPRi is not as efficient as CRISPR-Cas9 mediated gene knockout, but it could be a suitable option to screen for genes that are essential for cell viability. Also, because the CRISPRi-mediated knockdown effect is reversible if an inducible dCas9 is used, it allows extra on/off control to be incorporated in screening design [135]. The Weissman library can also be used for CRISPR activation (CRISPRa) screen when utilised with dCas9-VP64 with the SunTag system for

signal amplification. Another available version of CRISPRa library was designed by the Zhang lab, which contains 70,290 gRNAs targeting every coding isoform from the RefSeq database[205]. In this system, gene activation is mediated by dCas-VP64 combined with modified gRNA that recruits MS2-p64-HSF1, which is also referred as the Synergistic Activation Mediator (SAM) [205]. Both CRISPRa systems have exhibited ability to increase gene expression, however, the degree of increment depends of the targeted gene. Nonetheless, the CRISPRa libraries give researchers the opportunity to perform gain-of-function screens and study the biological pathways from a different angle.

1.2.5.4 Experimental design of CRISPR-Cas9-mediated screens

1.2.5.4.1 gRNA design

The outcome of a CRISPR-Cas9-based screen is directly determined by the design of the gRNA library. The key to an effective gRNA library is to maximise the overall on-target efficacy and minimise any off-target activity. Generally, the basic design process follows a set of conventional rules. First, gRNAs should designed against the the constitutive coding exons [202] [366] [434]. Second, all available gRNA candidates should be screened based on the potential off-target matches in the genome. Any gRNAs having a perfect match of its seed region elsewhere in the genome should ideally be removed from the library [202] [366] [434] [176]. Third, gRNAs with very low or high GC content, as well as homopolymer stretches should be avoided [434].

Wang et al. described some of these early rules and performed additional tests on gRNA efficacy [434]. By analysing the performance of gRNAs targeting the essential ribosomal gene sets, they found that gRNAs having pyrimidines at the last four nucleotides were disfavoured [434]. Consistently, Tzelepis et al. and Hart et al. also observed a strong bias against uridine at the last few positions of the gRNA, which is due to the premature termination of transcription by RNA polymerase III [421] [153] [456]. The Root group took an approach of a tiling library covering all possible gRNAs for a selection of cell surface markers and used flow cytometry to measure the performance of each gRNA [93] [92]. In contrast to what has been reported by Vakoc and colleagues that higher knockout efficiency can be achieved by targeting the functional domains of a protein [370], the Root lab showed that gRNAs targeting the 90% of the N-terminal protein coding sequences exhibited similar efficiency, which allowed more gRNA selection flexibility due to expanded target site window [92]. In addition to target site selection, the Yusa group demonstrated that a modification of the gRNA scaffold improved its targeting efficiency significantly,

and incorporated this change in their version 2 gRNA libraries [64][421]. Finally, the false-negative error due to lack of efficacy can be improved by increasing the number of gRNAs per gene [305]. However, a library with a larger number of gRNAs require a larger screening scale, and a balance needs to be considered by taking cost, space and number of screening conditions into consideration.

The off-target effect of gRNAs are predicted based on the position, number and nucleotide identity of potential mismatches. Although analysis in the early proof-of-principle screening studies demonstrated low off-target activity at predicted sites [202] [366] [434], the recent unbiased DBS prediction revealed unexpected potential off-target activity [125] [418]. Furthermore, a series of ChIP-seq of dCas9 coupled with various gRNAs showed a large number of off-targets binding events [211] [456]. Though such analysis is not feasible for a large-scale gRNA library, it indicates that there is a room for further improvement of gRNA specificity in library design. Potential measures include paying attention to alternative PAM sites, gRNA modification and utilisation of the double-nicking approaches. Although off-target activity is deeply concerned in clinical applications, it is unlikely to affect the performance of a genetic screen, due to the fact that any false-positive hits can be identified by comparing to the phenotype of other gRNAs targeting the same gene.

1.2.5.4.2 Cas9 and gRNA delivery

Almost all the published CRISPR-Cas9-based screens to date have unanimously used lentivirus to deliver gRNA libraries. The idea has been adapted from the delivery of shRNAs libraries in RNAi-based screens. The reason of its popularity is mainly because they can stably integrate into the host genome. The virus titre at transduction needs to be carefully titrated to achieve a reasonably low multiplicity of infection (MOI) is achieved at transduction, so that the majority of transduced cells have been infected with one virus particle.

There are mainly two strategies to generate Cas9-expressing cell line: the first is to deliver Cas9 and gRNA in a single virus as demonstrated by Shalem et al. [366], and the second is to establish a stable Cas9-expressing cell line by knockin or viral transduction. Several evidence suggested that the prior-generation of Cas9-expressing cell line is advantageous because a clonal cell line with high Cas9 activity can be selected. Tzelepis et al. showed that subcloned Cas9-expressing HT29 cell line exhibited higher efficiency of mutagenesis [421]. Similar phenomena has been observed in Huh7.5, HeLa, 293T and HT1080 cell lines [490]. Such improvement in efficiency is especially important for negative selection

screens where higher screening sensitivity is required. However, it might not be applicable in primary cells, where the one-vector approach or Cas9-transgenic mouse can be used [322].

1.2.5.4.3 Gene identification and data analysis

At the end of a pooled CRISPR-Cas9 screen, cell pellets are collected from both treated and control samples, and genomic DNA is extracted from them. The lentiviral integrant, which contains gRNA sequence, will be amplified by PCR and analysed by Illumina sequencing. The necessary sequencing depth is largely dependent on the design of the screen and resulted final cell number to be sequenced. In the case of a positive selection screen, often only a small number of cells are collected at the end of the selection process, hence only a few million reads will be enough. In contrast, negative selection screens, where the change of gRNA representation can be subtle and cell population at the end of the screen is usually large, require much deeper sequencing depth. Typically the Illumina HiSeq platform is used with 30-40 million reads for a population up to 100 million cells.

Following sequencing, statistical analysis needs to be performed to determine the significance of any changes between control and experimental samples at the gRNA-level, as well as gene-level. A range of algorithms designed for differential RNA-Seq expression analysis or shRNA screens were employed to analyse CRISPR-Cas9 screening data. Shalem et al. used the RNAi Gene Enrichment Ranking (RIGER) algorithm, which examined the positions of the gRNAs targeting the same gene and assessed whether the set was skewed towards the top of the list based on Kolmogorov-Smirnov statistic. An enrichment score was calculated based on this algorithm followed by a permutation test [239] [366]. The Wei group adopted an R package called DESeq2 to perform the statistical analysis of gRNA abundance, where gRNAs were ranked by the average fold changes [490]. Although algorithms for differential expression analysis such as DESeq2, edgeR and baySeq can be used to evaluate the statistical significance of hits in the CRISPR/Cas9-based screens, they can only perform the analysis at the gRNA level. Algorithms for shRNA screens such as RIGER and Redundant siRNA Activity (RSA) are also not ideal [225]. Followed the need of a computational algorithm suitable for CRISPR/Cas9-mediated screens, the Liu group developed an algorithm called Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) [225].

In MAGeCK, read counts from all the conditions were normalised by the median ratio method, followed by the use of a negative binomial (NB) model to estimate the statistical

significance between the treatment and control samples [225]. The gRNAs are ranked based on P-value calculated from the NB model and a modified robust ranking aggregation (α-RRA) algorithm is used to evaluate the statistical significance at the gene level [225]. Specifically, the α-RRA ranks a gene by comparing the skew of its gRNA sets to the uniform null model and calculates the false discovery rate (FDR) from an empirical permutation test using the Benjamini-Hochberg procedure [225]. The authors demonstrated more robust and sensitive CRISPR-Cas9 screening data analysis using MAGeCK compared to the other existing algorithms [225]. The reason is probably because MAGeCK has considered the fact that not all gRNAs targeting the same gene are working equally well and the α-RRA algorithm can remove the gRNAs with low targeting efficiency, which is more likely to produce more accurate gene level analysis.

1.3 Embryonic stem cells (ESCs)

1.3.1 Early development of mouse embryo

The mouse embryonic development begins in a fertilised egg packed within the protective glycoprotein layer called zona pellucida. The fertilised egg, also known as zygote, is capable of generating all embryonic and extraembryonic lineages. Such ability is defined as totipotency. The first five cell cycles of the embryonic development, which is referred to as the cleavage divisions, occurs without increase in total cell mass. Cells generated from cleavage divisions are called blastomeres. Blastomeres retain both embryonic and extraembryonic potential until the late eight-cell stage, when cell polarity is established and compaction takes place to form morula. During compaction, the spacial location of each individual cell is instructive for their subsequent lineage differentiation. Cells located on the outside develop into the first extraembyonic lineage called trophectoderm, which is essential for implantation and will subsequently differentiate into placenta. In contrast, the inner cells of a morula are biased toward forming the inner cell mass (ICM).

Specification of the trophoblast lineage appears to be mediated primarily by the Hippo signalling pathway, which conveys positional information into lineage-specific gene expression. In an early embryo, the Hippo pathway is active in the inner cells, where Yap1 is phosphorylated by Lats1/2 and degraded. As a consequence, Yap1 is unavailable to act as a co-activator for the key trophectoderm transcription factor Tead4, resulting in failure to activate the trophectodermal programme. On the contrary, in the outer cells, where Lats1/2 is inactive,Yap1 pairs with Tead4 and upregulates Cdx2, Gata3 and eomesodermin, which collectively drive commitment to the trophoblast lineage. In line with this model, Lorthongpanich et al. showed that knockdown of LATS kinases by injecting siRNA into mouse zygotes caused lineage misspecification and resulted in the generation of a TE-like lineage in the morula [236]. Once upregulated, the expression of Cdx2 and Eomes is maintained by Elf5 through a positive feedback loop to reinforce the commitment to the trophoblast lineage [288]. Despite both being important for trophectoderm (TE) specification, Cdx2 and Eomes seem to play different roles. It was shown that Cdx2-deficient blastocysts failed to repress Oct4 and Nanog in the outer cells, which led to the failure of the segregation of ICM and TE, whereas Eomes mutant blastocysts could implant and showed normal Cdx2 and Oct4 expression [394] [430]. These observations suggested that Cdx2 is the earlier TE inducer in morula and Eomes is required for further TE differentiation at the blastocyst stage.

Coincident with the specification of the TE, the establishment of ICM is under the influence of the upregulated Oct4. In the absence of Oct4, inner cells fail to develop into mature ICM but rather divert into TE [289]. Oct4 acts cooperatively with Sox2 to regulate the expression of several pluripotent genes, including *Fgf4* and *Nanog* [13] [33]. It was shown that *Sox2*-null embryo was able to develop normal ICM, but fail to maintain an epiblast or further differentiation [6].

After the segregation of trophoblast and ICM, the trophoblast pumps fluid into the blastocyst to form a cavity known as the blastocoel. At this stage, the ICM start being partitioned into the epiblast and primitive endoderm as a consequence of differential gene expression. This specification is first observed when Nanog and Gata6 begin to express in a mutually exclusive manner. Cells expressing primitive endoderm markers such as *Gata6*, *Gata4*, *Sox17* and *Pdgfr* αgradually move away from Nanog-positive cells, and eventually form a morphologically distinguishable epithelium layer adjacent to the cavity. This process is regulated by the FGF signalling pathway as embryos deficient of *Grb2*, *Fgf4* or *Fgfr2* fail to form the primitive endoderm [63] [69][291]. At the same time, *Nanog*-expressing cells remain restricted to the inner space between the trophectoderm and primitive endoderm and develop into the pluripotent epiblast. These pluripotent cells are thought to be in the 'ground state', which is characterised by their unrestricted differentiation capacity and flexibility to the formation of all embryonic lineages [258].

1.3.2 Derivation of mouse ESCs

ESCs are characterised by its ability to sustain self-renewal and remain as undifferentiated for an extended period of time in culture. When injected into adult mice, ESCs give rise to multi-differentiated terotocarcinoma. Their full differentiation potential was revealed by blastocysts injection, which yields chimeric mice with high contribution from the injected ESCs to all tissues, including functional colonisation of the germ line [34]. Competence of germline transmission suggests that ESCs can be exploited as a vehicle for introducing genetic modifications into mice [335]. The fact that ESCs are permissive to multiple rounds of sophisticated genetic manipulation and their ability of clonal expansion enables isolation of mutants with desired genetic modification. These groundbreaking discoveries led to the creation of transgenic mice, which became an immensely powerful technology for basic research and the development of new therapies.

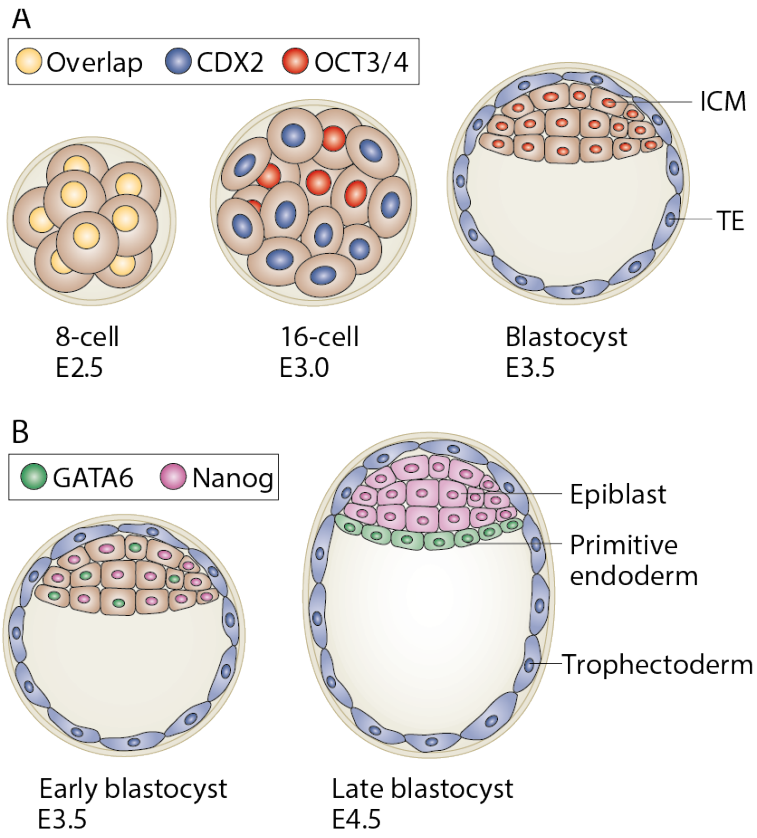

Figure 1.4: Lineage segregation in mouse blastocyst. (A) At E2.5, the eight blastomeres retain both embryonic and extraembryonic potential, which is reflected by the overlapping expression of Cdx2 and Oct4. At the blastocyst stage, Oct4 is expressed exclusively in the inner blastomeres, which leads to the formation of inner cell mass (ICM). As the early trophectoderm (TE) inducer, Cdx2 is exclusively expressed in the outer blastomeres. (B) At E3.5, the ICM shows mosaic 'salt and pepper' expression of Nanog and GATA6. GATA6-positive cells are subsequently sorted to the distal surface of the ICM, where they give rise to the primitive endoderm. Nanog-positive cells exclusively give rise to the pluripotent epiblast. Image taken and adapted from Arnold and Robertson, 2009 [4].

It was a challenge to derive ESCs directly from the embryo until Martin and Evans independently succeeded in isolation and maintenance of pluripotent cell lines [107] [260]. The original derivation of mouse ESCs involved explanting blastocysts or isolated ICMs on a layer of mitotically inactivated fibroblasts called 'feeders', in medium containing fetal calf serum. This method was developed empirically from the early research on embryonal carcinoma(EC) cells. Smith et al. later demonstrated the ability of leukaemia inhibitory factor (LIF) in replacing feeder cells both in derivation and long-term culture. The derivation process in serum-containing culture is inefficient as the emergence of ESCs only happens after dissociation and replating of the primary outgrowth. With the advent of 2 inhibitors (2i), namely GSK3 inhibitor CHIR99021 and and MEK inhibitorPD032590, the derivation process became more efficient [290].

It has been well-established that mouse ESCs are originated from the mid-blastocyst-stage at embryonic day E4.5. However, it has been shown that ESCs can also be derived from early-blastocyst-stage at E3.5 or even from eight-cell-stage blastomeres, suggesting that ESCs may represent a very early developmental stage. Attempts to derive pluripotent cell lines from implanted mouse embryos had not been successful for a long time until a different pluripotent cell type was established from the postimplantation-epiblast using a different culture condition [38] [410]. These cells are referred to as epiblast stem cells (EpiSCs). Unlike ESCs, EpiSCs do not rely on LIF or 2i to maintain pluripotency, instead require FGF and activin. EpiSCs exhibit some pluripotency features such as expression of Oct4, the ability to differentiate *in vitro* and form teratocarcinomas, but they cannot contribute effectively to blastocyst chimeras. Gene expression analysis revealed that EpiSCs exhibit relative low expression of ICM-specific genes such as *Rex1*, *Sox2* and *Nanog*, but upregulation of late epiblast markers such as *Fgf5*, *Brachyury* and *Sox17* [410] [38]. These evidence indicated that EpiSCs represent a more advanced state of pluripotency, which is called primed pluripotency

In addition to ESCs and EpiSCs, which have been derived from the epiblasts of the blastocyst, other stem cell lines have been established from other lineages of the early embryo. Examples include embryonic germ (EG) cells which can be derived from primordial germ cells (PGCs) in embryos between E8.5 and E11.5, permanent trophoblast stem cell lines from early post-implantation trophoblasts and extra-embryonic endoderm (XEN) cell lines from the primitive endoderm lineage [265] [331] [405] [209]. The establishment of these cell lines has provided powerful models for the dissection of the molecular mechanism underlying lineage specification in early embryonic development.

1.3.3 Regulation of the pluripotency state

Pluripotency is defined as a capacity of a cell to give rise to all the specialised cell types of an adult organism. The derivation of ESCs made it possible to capture pluripotency indefinitely *in vitro*, and provided an extraordinary tool to investigate the molecular mechanisms that govern pluripotency. Accumulating evidence suggested that the ESC identity is sustained through integrated actions of multiple extrinsic signalling pathways with intrinsic transcription regulatory network, reinforced by epigenetic modifiers.

1.3.3.1 Extrinsic signalling pathways

The ability of ESCs to retain pluripotency is stabilised by the continuous input of extrinsic cues. Such requirement is owing to the auto-inductive stimulus, in particular FGF, which promotes the exit from pluripotency. Multiple extrinsic factors need to be fed into the system to counterbalance the self-inductive differentiation signals and reinforce the pluripotent network (Figure 5.1).

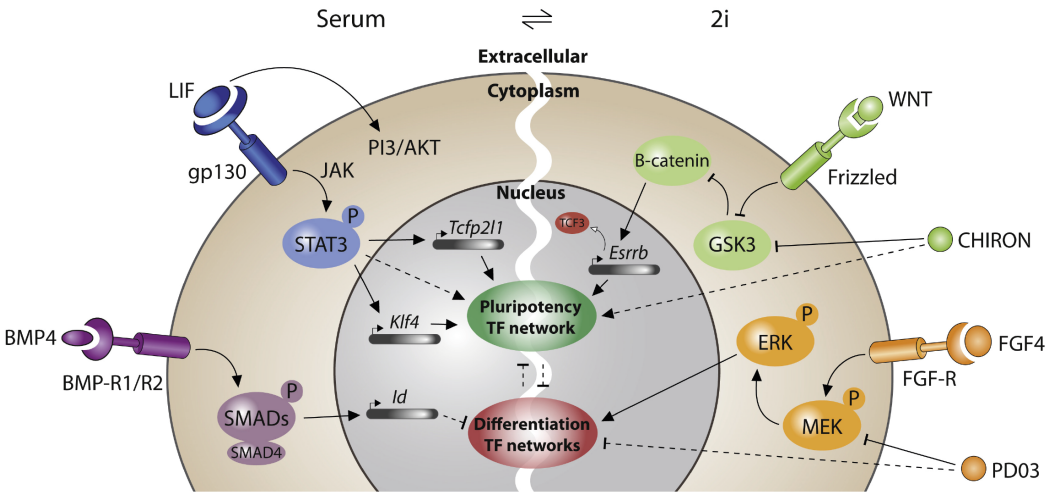

Figure 1.5: Extrinsic signalling pathways that regulate pluripotency. Filled arrows represent activation of target activity and bars indicate inhibition. Solid lines indicate a direct or known downstream target, whereas a dashed line shows an indirect or inferred effect. BMP4 signalling functions via phosphorylating Smads to activate Id genes. LIF signalling affects many pathways but its positive effect on pluripotency is primarily via JAK-mediated phosphorylation of STAT3, which activates *Tcfp2l1* and *Klf4*. Canonical WNT signalling inhibits GSK3 activity leading to stabilisation of ?-catenin, which in turn abrogates TCF3-mediated repression of pluripotency genes including Esrrb. CHIRON closely mimics WNT signaling by inhibiting GSK3. FGF signalling activates the MAPK pathway which promotes the transition to ?a primed? state. Two inhibitors (2i), CHIRON and PD03, stabilise naive pluripotency through inhibiting GSK3 and MAPK pathway respectively. Image taken and adapted from Hackett and Surani, 2014 [147]

1.3.3.1.1 LIF-mediated signalling pathway

The discovery of LIF was driven by the urge to demystify the function of feeder cells. Through screening of a range of feeder cell types, medium collected from Buffalo rat liver cells was found to effectively suppress differentiation and sustain propagation of ESCs in the absence of feeders [382]. Further analysis of this medium led to the identification of the effective factor, LIF [381] [445]. LIF is able to replace feeders in both derivation and long-term culture of mouse ESCs. It is likely that feeder cells provide extra support other than LIF as ESCs cultured in serum and LIF in the absence of feeders demonstrate slightly larger heterogeneity and more differentiated cells. However, LIF is the major

functional component of feeders given that *LIF* knockout fibroblasts were unable to sustain undifferentiated ESCs growth [392].

LIF signals through binding to a heterodimeric receptor consists of gp130 and Lifr [131]. Both gp130 and Lifr are constitutively associated with tyrosine kinases from the JAK family [444]. These kinases become activated upon binding of LIF. There are four members of JAK family, among them, Jak1 was demonstrated to be the primary downstream effector of of Lifr [388] [105]. Activated Jak1 initiates a cascade of tyrosine phosphorylation that stimulates three distinct signalling pathways: the JAK/STAT, PI(3)-kinase, MAPK pathways[388] [413] [302]. These pathways contribute to self-renewal, survival as well as differentiation [292].

The Stat proteins are a family of transcription factors, among which Stat3 is the crucial functional effector following stimulation by LIF. Stat3 is activated via phosphorylation by Jak1, which allows it to form a signalling competent dimer and translocates into the nuclease. It was shown that Stat3 is required for ESC maintenance. Over-expression of a dominant-negative Stat3 construct in ESCs led to abrogated self-renewal and differentiation, whereas over-expression of Stat3 is sufficient to sustain LIF-independent self-renewal [295] [264]. Molecular studies revealed that Stat3 promotes pluripotency via upregulating pluripotency genes such as *Klf4*, *c-Myc*, *Gbx2* and *Tfcp2l1* [148] [53] [297] [399] [466] [257]. Additionally, ChIP-Seq analysis revealed that STAT3 shares many common target site with Nanog, Oct4 and Sox2, indicating that JAK/STAT3 pathway directly feeds into the core pluripotency transcription network [68].

Activation of PI(3)-kinase pathway following LIF stimulation is driven by the association between Jak1 and the p85 subunit of PI(3)-kinase. It has been demonstrated that inhibition of PI(3)-kinase pathway in ESCs resulted in spontaneous differentiation, even in the presence of LIF. Furthermore, ESCs expressing an active form of Akt could be maintained in an undifferentiated state without LIF, indicating that constitutively active PI(3)-kinase signalling is sufficient for ESC maintenance [439].

Interestingly, in parallel to JAK/STAT and PI(3)-kinase pathway, LIF activates MAPK pathways via recruiting Shp2, which induces stimulation towards differentiation [47] [361]. It was demonstrated that *Socs3* null ESCs, which showed hyperactive MAPK pathway as a result of non-competition of binding to the receptors, were inclined to differentiate into primitive endoderm in the presence of LIF [401]. And this phenotype can be reversed by inhibition of Mek [118]. These observations suggests LIF induces competing pathways

and the self-renewal signal downstream of LIF is under a fine balance between positive and negative pluripotency regulatory pathways.

1.3.3.1.2 TGF-β-mediated pathway

The TGF-β family comprises a broad range of proteins including TGF-β, nodal, activins, Growth Differentiation Factors (GDFs), and Bone Morphogenetic Proteins (BMPs) [261]. Signalling mediated by TGF-β ligands is transduced through two types of transmembrane kinases, the type I and type II receptors. In the canonical TGF-β pathway, upon ligand binding, type II receptors phosphorylate type I receptors, which in turn phosphorylate and activate the downstream effector Smads. Smad1, Smad5 and Smad8 are activated by BMP receptors, whereas Smad2 and Smad3 are activated by TGF-β/nodal/activins receptors. These receptor-activated Smads form trimers with the common Smad, Smad4, and translocate into the nucleus, where they interact with other transcription factors, co-activators or co-repressors to regulate the expression of target genes [343].

Bmp4 signalling plays an important role in mouse ESC maintenance through upregulation of the Id proteins, which are transcription factors inhibit neuronal differentiation [469]. It was shown that Bmp4 can substitute serum in ESC maintenance in the presence of LIF, indicating that Bmp4 might be the functional component of serum [469]. However, Bmp4 is known to promote mesoderm, endoderm and trophoblast differentiation [286] [315] [342] [424] [460], and LIF can block mesoderm and endoderm differentiation but not neural differentiation [470]. These evidence led to the proposal of a mechanism that LIF and Bmp4 act cooperatively in supporting pluripotency status by each suppressing differentiation towards specific fates. Because of this counterbalancing effect, ESCs in serum/LIF condition exist in an unstable environment with competing signals, which is reflected by the pluripotency heterogeneity.

The TGF-β/activin signalling comprises another branch of canonical TGF-β pathway, which was shown to be essential for pluripotency maintenance in human ESCs and mouse EpiSCs via Smad2/3-mediated Nanog activation [423]. However, activation of TGF-β/activin signalling induces differentiation of mouse ESCs in the absence of LIF.

1.3.3.1.3 FGF/MAPK pathway

Fgf signalling is activated via binding of Fgf to the receptor tyrosine kinase Fgfr, which leads to the formation of a complex between Fgfr, Frs2α, Shp2, and Grb2. The complex formation facilitates the activation of the phosphorylation cascade through Ras-Raf-Mek-

Erk. Originally considered as a autocrine self-renewal signal, it was later confirmed that Fgf4 acts to stimulate ESCs towards lineage specification. Evidence include that *Fgf*-deficient ESCs were severely compromised in neural and mesodermal differentiation [443]. Although formation of the Fgfr-Frs2α-Grb2 can also activates PI(3)-kinase pathway via Gab1, Mek1/2 was identified as the downstream effector of the Fgf4 signal based on the fact that the phenotype of *Fgf4*-null ESCs can be reproduced using Mek1/2 inhibitor [214] [47] [390]. Furthermore, Kunath et al. demonstrated that *Erk2*-null ESCs fail to commit differentiation and retain expression of pluripotency markers Oct4, Nanog and Rex1 [210]. Consistent with the observations *in vitro*, Mek inhibitor treated embryos failed to form blastocysts and generated enlarged epiblast [291]. Similar phenotype can be observed in *Grb2*-deficient embryos [63].

These discoveries led to the hypothesis that blocking Fgf/Mapk pathway facilitates maintenance of pluripotency. As predicted, it was shown that either of the Fgfr inhibitor SU5402 or MEK1/2 inhibitor PD184352 could replace the requirement for serum/BMP and support long-term ESC maintenance [471]. However, inhibition of FGF/ERK pathway is not sufficient to maintain ESC self-renewal without LIF.

Despite the widely use of Mek inhibitor in ESC maintenance, its molecular mechanism remained unclear, until a study performed by Yeo et al. suggested that ERK2 drives differentiation through phosphorylation and destabilisation of Klf2. It was demonstrated that over-expression of Klf2 can replace Mek inhibition which allows stable culture under Gsk3 inhibition alone [467]. In addition, Tee et al. showed that Erk2 directly modulate chromatin features required for developmental gene expression via regulating PRC2 and RNAPII [409]. Notably, ESCs express Fgf4 under the regulation of Oct4 and Sox2. This indicates that the transcription factors essential for the establishment and maintenance of pluripotency also function as differentiation promoters [473].

1.3.3.1.4 Wnt signalling pathway

In the absence of Wnt, Apc, Axin and GSK3 form a complex that phosphorylates β-catenin in coordination with Ck1α, which marks it for ubiquitination and proteolysis [76]. In the presence of Wnt, Frizzled receptor forms a complex with Lrp5/6, which triggers the displacement of GSK3 from the destruction complex, allowing β-catenin to accumulate and translocate into the nucleus where it interacts with co-activators to drive transcription of target genes [451] [454].

A positive effect of Wnt signalling in promoting self-renewal was demonstrated by two studies focusing on the knockout phenotype of *Apc* and *Gsk3* in ESCs. Kielman et al. showed that constitutive activation of Wnt signalling via *Apc* mutation affected the differentiation potential of ESCs both *in vitro* and in teratomas [192]. Doble et al. generated the *Gsk3* DKO cell line, in which both Gsk3α and Gsk3β were inactivated. The DKO cell line, which has elevated β-catenin, demonstrated severe defects in differentiation [91]. Notably, in both studies, the severity of the phenotype exhibited a dose-dependent manner which correlated to the *Apc* mutation or *Gsk3* functional alleles. Furthermore, Sato et al. showed that the addition of a GSK3 inhibitor BIO could facilitate maintenance of ESC and resulted in sustained expression of Oct4, Rex1 and Nanog [352]. Ogawa et al. demonstrated that supplementation of Wnt3a helped to maintain ESC self-renewal in the presence of LIF [301]. These evidence converged to the deduction that elevated β-catenin promotes ESC self-renewal and results in differentiation defects, which was confirmed by Wray et al. by showing that the absence of β-catenin eliminated the self-renewal response to Gsk3 inhibition [450]. They also showed that the responsiveness could be restored by truncated β-catenin lacking a transactivation domain [450]. This indicates that the transcriptional activation function is not required for β-catenin to confer differentiation resistance. Instead, it was found that the role of β-catenin in pluripotency arises through direct interaction with the transcription repressor Tcf3 (gene name *Tcf7l1*) [450]. ChipSeq data revealed that Tcf3 shares binding site with Oct4 and Sox2 [109], and acts a repressor to antagonise their function [269]. Other key pluripotency factors repressed by Tcf3 include *Esrrb*, *Klf2* and *Nanog* [77] [316] [142] [450]. Interaction between β-catenin and Tcf3 abrogate its repressive effect on pluripotent genes and stabilises pluripotency programme [374] [453]. It has been shown that *Tcf7l1*-null ESCs exhibit enhanced self-renewal and differentiation defects[468] [142]. *Tcf7l1*-null embryos develop normally until profound defects was observed in axial patterning during implantation, which highlighted the prominent role of Tcf3 as a regulator for differentiation [270].

It was shown that the stimulation Wnt signalling facilitates ESC maintenance; however, activation of Wnt signalling alone is insufficient to maintain long-term ESC self-renewal [471] [301]. Remarkably, the combination of GSK3 inhibitor (CHIR9902) with the Fgfr inhibitor (SU5402) and Mek1/2 inhibitor (PD184352) could effectively maintain ESCs for an extended period of time even in the absence of LIF [471]. This system was referred to as '3i', which evolved to '2i' with the substitution of the SU5402 and PD184352 to a more potent and specific Mek inhibitor PD0325901.

1.3.3.1.5 Serum/LIF culture and 2i culture

Conventional condition (serum/LIF) is chemically undefined and often activate multiple conflicting pathways. As a result, it promotes a considerable degree of morphological, transcriptional and functional heterogeneity among cells [147]. This heterogeneity is reflected on the expression of a range of pluripotency-associated transcription factors, such as Nanog, Rex1, Esrrb, Stella, Klf4, Tbx3 and Hex [57] [417] [154] [426] [49] [297]. Functional distinction was observed between cells with different expression levels of some of these factors. For example, cells with low Nanog expression exhibit moderate expression of primitive endoderm markers such as Gata4 and Hex1, and epiblast marker Fgf5 [49] [377] [186]. Similarly, Rex1-low cells were shown to have poor ability to form chimeras following blastocyst injection [417]. These observations indicate the existence of two distinct sub-populations in ESCs cultured in serum/LIF: naive pluripotent cells, and primed cells. The later is associated with expression of lineage markers and poor performance in pluripotency assays. Notably, purification of the primed cells by cell sorting and replating assay showed that these two subpopulations are interchangeable, and the transcriptional and functional differences exist in in a dynamic equilibrium [451] [57]. Overall it suggested that ESCs in serum/LIF condition is maintained in a metastable naive pluripotent state, with a small proportion of cells cycling in and out of the 'primed', pluripotent state [147].

The development of 2i condition allowed maintenance of ESCs in stabilised naive state, characterised by its relatively spherical colony morphology with defined borders and lack of differentiating cells. 2i-cultured ESCs exhibit a homogeneous transcriptional and epigenetic state with uniform expression of *Nanog* and *Rex1* [471]. Transcriptome and epigenomic analysis showed that the 2i-cultured ES cells exhibit a profile comparable to that of E4.5 epiblast, which probably explains its higher chimera contribution [30][203]. This more robust naive pluripotency status in 2i condition is probably owing to the complete insulation of differentiation signals. It was thus proposed that ESCs in 2i represent the *in vitro* 'ground state', meaning a homogenous population with the potential to form all embryonic lineages unbiasedly [471]. The *in vitro* ground state is the most optimised state of naive pluripotency to date and the closest model to the pre-implantation epiblast [147]. However, a recent study reported that prolonged MEK1/2 suppression resulted irreversible epigenetic changes that compromise the developmental potential of ESCs [73].

1.3.3.2 Transcription factor network

1.3.3.2.1 Core pluripotency factors

Oct4

Oct4, also known as Oct3, is a member of the POU transcription factor family encoded by the *Pou5f1* gene. Oct4 regulates gene expression by binding to the octamer motif ATGCAAAT within the promoter or enhancer region and was the first factor identified as a master transcription factor in pluripotency and lineage specification regulation [363] [220] [303].

Oct4 is absolutely essential for embryogenesis as *Oct4*-deficient embryos failed to develop ICM and die at the time of implantation [289]. The detection of Oct4 was made at as early as the zygote stage, which is believed to be inherited from the oocyte [479]. Zygotic Oct4 expression can be detected at 4- or 8- cell stage in blastomeres until blastocyst formation. After the first lineage specification takes place, cells in the ICM retain Oct4 expression whereas cells in the trophoblast have little or no Oct4 expression [307] [318]. Upon implantation, transient up-regulation of Oct4 induces the formation of primitive endoderm, while in the epiblast, Oct4 expression remains uniformly and continuously high [307] [318]. During gastrulation, Oct4 expression is down-regulated and eventually confined to primordial germ cells. In cell culture systems, Oct4 is highly expressed in ESCs, ECs, and embryonic germ cells. Its expression is down-regulated upon induction of differentiation [307] [190] [318].

The critical role of Oct4 in pluripotency maintenance was uncovered by Niwa et al. For this, an inducible Oct4-expression system was established wherein Oct4 level can be modulated by the addition of tetracycline. Using this system, Niwa et al. demonstrated that a two-fold increase in Oct4 expression led to primitive endoderm differentiation, whereas repression of Oct4 caused dedifferentiation to trophectoderm [296]. Therefore, Oct4 expression needs to be tightly regulated in ESCs. It was shown that the Oct4 positive regulators include Esrrb and Sall4, whereas Tcf3, Gcnf and Cdk2 mediate its negative regulation [485] [483] [353] [84] [404] [281]. In addition to its crucial function in the maintenance of pluripotency, Oct4 plays a role in regulating early cell fate. As mentioned previously in this Chapter that expression of the autocrine differentiation signal Fgf4 was under the regulation of Oct4. It was shown that Oct4 formed a complex with Cdx2, which resulted in a reciprocal inhibition mechanism with mutually exclusive expression and facilitated the segregation of pluripotent stem cells and trophectoderm [298]. Similarly, others have

shown that sustained Oct4 expression induced specific lineage commitment in dependent on the condition. For instance, Shimozaki et al. reported that Oct4 upregulation in ESCs accelerated neurogenesis under serum-free culture condition [372]. Additionally, transient increase in Oct4 expression led to cardiac commitment [480]. Finally, Oct4 was found to play an important role in the reprogramming of somatic cells into induced pluripotent stem cells (iPSCs). As one of the groundbreaking works in the stem cell field, Takahashi and Yamanaka screened 24 factors and found that four transcription factors Oct4, Sox2, Klf4 and c-Myc were sufficient to reprogram fibroblast cells to pluripotent cells [400].

Among all pluripotency regulators, Oct4 was found to be central to the machinery. Most importantly, Oct4 acts as a fundamental coordinator that recruits factors with various functions to establish gene regulation programmes. Several mass spectrometry studies were performed and identified a large number of Oct4 interaction partners from families such as transcription factors, epigenetic modifiers, transcriptional coactivators and components of signalling pathways [425] [106] [89] [309]. In particular several chromatin remodelling complexes such as NuRD, SWI/SNF and LSD1 were found to interact with Oct4. The correlation of LSD1 and Oct4 was confirmed by Whyte et al., showing that LSD1 mediated pluripotency-related gene silencing during differentiation and this function was through recruitment by Oct4 [442]. Although many of the other proposed correlations found in these mass spectrometry analysis need to be validated, they demonstrated the prominent role of Oct4 in pluripotency regulatory network.

Sox2

The most well-known partner of Oct4 is Sox2, which is also considered as one of the core pluripotency factors. Sox2 belongs to the Sry high mobility group (HMG) box (Sox) superfamily, which interact with DNA via the HMG domain with a consensus sequence. Like Oct4, Sox2 is also required for early embryogenesis. Homozygous *Sox2* mutant embryos die shortly after implantation, due to failure of the epiblast formation [6]. Sox2 is highly expressed in mouse ESCs. *Sox2*-null ES cells differentiated primarily to trophoectoderm-like cells, similar to Oct4-null ES cells [262]. Notably, forced expression of Oct4 in *Sox2*-deficient ESCs could rescue their the phenotypes, suggesting that the role of Sox2 in pluripotency maintenance is to sustain Oct4 expression [262]. Masui et al. also showed that Sox2 positively regulated the expression of Oct4 by promoting the expression of Oct4 positive regulators such as Nr5a2 and repressing Oct4 negative regulators such as Nr2f2 [262]. Oct4 and Sox2 bind DNA cooperatively and act synergistically on many pluripotency-related genes [1] [2]. Compared to Oct4, Sox2 is more widely expressed in

the developing embryo, from epiblast to trophectoderm as well as later in the neuroectoderm [6] [191]. Sox2 has been reported to be in charge of neural differentiation by repressing other lineage regulators such as brachyury [436] [487].

Nanog

Another key regulator that contributes to the core pluripotency circuit is Nanog. Nanog is a homeodomain-containing transcription factor whose role in pluripotency was first described by two groups independently in 2003. Chambers et al. discovered Nanog from a functional screen using an ESC cDNA library and found that forced expression of Nanog from transgene is sufficient to maintain ESC pluripotency with elevated Oct4 level independent of LIF [56]. In the same issue, Mitsui et al. reported the identification of Nanog by digital differential display comparing the expressed sequence tag libraries from ESCs and somatic tissues [277].

Deletion of Nanog results in preimplantation lethality, indicating its indispensability in early mouse embryo development. *Nanog*-null embryos failed to develop epiblast, instead cells either committed to trophoblast differentiation or progress to apoptosis [375] [277] [289].The expression of Nanog in ICM follows a 'salt-and-pepper' pattern mutually exclusive with Gata6-expressing cells, which was shown to be essential for the formation of primitive endoderm and epiblast via potentiating Gata6 expression and providing support from a functional epiblast [375]. By the late blastocyst stage, Nanog expression become restricted to epiblast compartment, where it is uniformly expressed [375]. It was observed that Nanog expression in conventional ESC culture is heterogeneous. To investigate this phenomena, Chambers et al. generated heterozygous and null Nanog cell lines [57]. A reduction of self-renewal ability was observed in relation to the dosage of Nanog [57]. Surprisingly, it was observed that *Nanog*-null cells maintained the ability for self-renewal, albeit prone to differentiation [57]. Subsequent studies showed that Nanog-null cells were able to contribute to three germ laters. These findings suggested that Nanog mainly function in stabilising pluripotency by counteracting alternative gene expression states [57].

Like Oct4 and Sox2, Nanog has been shown to interact with a large number of protein partners ranging from transcription factors, chromatin modifiers and signalling pathway components, indicating its critical role as a core pluripotency regulator. Unlike Sox2 which is closely associated with Oct4, global mapping of Oct4 and Nanog binding sites showed only partial co-occupancy of Nanog and Oct4. In addition, transcriptome analysis after

shRNA knockdown of either Oct4 and Nanog showed distinct gene expression profile [232] [294]. It was thus proposed that Nanog has a complementary and partially overlapping gene regulation activities to Oct4/Sox2 [258].

The Oct4/Sox2/Nanog triumvrate

Oct4, Sox2 and Nanog crossly regulate each other, forming an interconnected auto-regulatory and feedforward circuitry known as the Oct4/Sox2/Nanog (OSN) triumvrate [33]. They function cooperatively to activate the expression of genes required to maintain pluripotency, and at the same time repress genes involved in lineage specification [472] [33]. The ability of OSN to positively or negatively regulate gene expression is based on the interaction with other transcription factors and epigenetic machineries. Chen et al. demonstrated that the OSN collaboratively activate gene expression via binding to the enhancer site [68]. Most of these binding sites are occupied with coactivator p300 and mediator [68] [24] [182]. Furthermore, Yuan et al. showed that Oct4 recruits Setdb1, which catalyses the repressive histone modification H3K9me3 at genes associated with trophectoderm differentiation, such as Cdx2 and Tcfap2a [474]. Similar findings have been reported by the Young lab [24]. Additionally, Liang et al. showed that Nanog and Oct4 associate with specific repressor protein from the NuRD complex, namely Hdac1/2 and Mta1/2, to form a complex and co-occupy Nanog target genes for developmental gene repression [226]. Importantly, pluripotency signalling pathways are wired into the OSN circuitry to deliver exogenous information to the genome in the form of activate transcription factors [472]. The OSN binding sites are correlated with the binding of Stat3, Tcf3 and Smad1, which are the effectors of LIF/STAT3, Wnt and BPM4 signalling pathways. Loss of Oct4 leads to a loss of co-binding of these transcription factors, indicating that these pathways regulate pluripotency by directly deliver signals to the core regulatory circuitry [68] [232] [77] [472].

1.3.3.2.2 Ancillary pluripotency regulators

Further to the core pluripotency circuitry established by the OSN triumvrate, ESCs also express a repertoire of 'ancillary' pluripotency regulators that are individually dispensable but collectively reinforce naive pluripotency. A large-scale RNAi screen performed by Ivanova et al. identified several ancillary pluripotency factors including Esrrb, Tbx3, Tcl1 and Dppa4 [170]. Among them Esrrb appears to play an especially crucial effect, probably because it interacts with Oct4 and is directly up-regulated by Nanog [111] [485] . Esrrb over-expression showed an enhanced self-renewal phenotype and ESC pluripotency

can be sustained without LIF [111]. During development, Esrrb is required in placenta formation but not the embryo [242], indicating that its role in pluripotency can be substituted by alternative pathways. Consistently, it was shown that *Esrrb* knockout ESCs can be isolated and propagated in serum/LIF with sustained Oct4 expression [111]. It was also reported that Esrrb is the principal target of Tcf3 and forced expression of Esrrb render ESCs propagation without GSK3 inhibitor [259]. Furthermore, Esrrb is dispensable in the presence of LIF, confirming the functional compensation by LIF/STAT3 [259]. Tbx3 is shown to be regulated by the PI3K/Akt pathway downstream of LIF [297], while Klf4 is found to be a direct target of Stat3 [148]. Forced expression of Tbx3 and Klf4 is sufficient to maintain self-renewal of ESCs in LIF-free condition [297]. A study performed by Martello et al. identified another ancillary factor Tfcp2l1, which is another non-compensable downstream target of STAT3 [257]. Other ancillary factors include but not limited to Tcf3, Klf2, Sall4, Prdm14, Pum1 and Zfp706. These factors are expressed uniformly in 2i but heregenously in serum.

1.3.3.3 Epigenetic regulation

1.3.3.3.1 DNA methylation

Methylation at CpG dinucleotide is a repressive epigenetic modification at the level of DNA [385] . Once established, DNA methylation (5mC) is stably maintained by DNA methyltransferase 1 (DNMT1) and propagated through cell division. It was found that during early development, 5mC is dynamically erased, resulting in a globally hypomethylated state in the ICM [384][380]. It was proposed that the global hypomethylation is to remove epigenetic barriers and facilitate pluripotency acquisition [147].

There are mainly two possible mechanisms of DNA demethylation: the replication-independent active DNA demethylation and the replication-dependent passive DNA demethylation [438]. As an effector involved in the active DNA demethylation, the activation-induced cytidine deaminase (AID) has been shown to demethylate the promoters of Oct4 and Nanog during human fibroblast reprogramming [20] [323]. However, controversial findings have been reported in certain mouse ESC lines [120]. The passive DNA demethylation pathway involves oxidation of 5mC to 5-hydroxy-methylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by the ten-eleven translocation (TET) family of enzymes [223]. *Tet1* and *Tet2* are highly expressed in mouse ESCs and are down-regulated upon differentiation [169] [201]. It was demonstrated that the rate and extent of DNA demethylation was compromised in Tet1 and Tet2 deficient ESCs [112]. Furthermore, si-

lencing of Tet1 resulted in downregulation of pluripotency genes such as *Nanog*, *Esrrb* and *Klf4* [169] [123] [201].

As the mouse embryonic development proceeds, DNA methylation is reestablished by Dnmt3a and Dnmt3b. Embryos with mutant Dnmt3b was normal in early developmental stages but was defective in later stages [304]. Furthermore, it was shown that Dnmt-deficient ESCs exhibited normal self-renewal ability but lost differentiation potential [67] [419] [266]. Collectively, it suggests that DNA methylation is important for lineage specification but not ESC maintenance.

Recent studies demonstrated that ESCs cultured in 2i/LIF exhibited global hypomethylation which is comparable with the ICM, whereas ESCs in serum/LIF accumulated a much higher level (approximately 3-fold) of DNA methylation, which resembles the hypermethylated state of the postimplation epiblast [113] [146]. Interestingly, the Nanog/Rex1-positive cells in serum/LIF also retain high global 5mC, indicating they may not be at the ground state, which is consistent with the heterogeneity and primed feature of serum/LIF cultured cells [113]. The difference in DNA methylation between serum/LIF and 2i/LIF indicates profound effect of exogeneic signalling pathways on the epigenetic landscape in ESC maintenance.

1.3.3.3.2 Bivalent domains

The developmental promoters in pluripotent stem cells are featured by the co-presence of activating modification H3K4me3 and the repressive modification H3K27me3, which is a phenomenon known as bivalency [458] [8] [308]. These conflicting marks are commonly observed in pluripotent stem cells but very rarely in somatic cells [17] [267]. The bivalent signature is thought to keep lineage-specific genes silenced yet maintaining a poised state so that they can be rapidly reactivated in response to differentiation cues [438]. The H3K27 methylation is catalysed by Polycomb repressive complex 2 (PRC2), which is composed of Ezh2, Eed and Suz12. It has been shown that the polycomb complexes are dispensable for ESC self renewal, but simultaneous knockout of PRC1 and PRC2 ESCs failed to differentiate into three germ layers, suggesting that the repressive epigenetic modifications are primarily function in the initiation of differentiation rather than pluripotency maintenance [55] [218]. Notably, cells cultured in 2i/LIF exhibited decreased H3K27 modifications on bivalent domains compared to cells in serum/LIF. It was thought that this could be a an effect of Erk inhibition since Erk is required for the activity of Eed [409]. The methylation of H3K4 is mediated by the Trxthorax group (TrxG) complex

such as Wdr5. It was shown that Wdr5 physically interact with Oct4 and genome-wide protein localisation analysis revealed overlapping gene regulatory functions between Oct4 and Wdr5 [3]. Additionally, depletion of Wdr5 resulted down regulation of Oct4 target genes and resulted in loss of self renewal [3]. It has been widely accepted that the bivalant histone modifications are the unique feature of pluripotent stem cells that keeps genes in an inducible state and increases robustness at the same time. However, some evidence suggested that bivalency may be functionally dispensable[83] [429] [147]. Thus more work needs to be done to directly probe the function of bivalent domains in development.

1.3.3.3.3 Heterochromatin organisation

Higher order of chromatin remodelling has been shown as an important machinery that facilitates coordinated action on gene expression. H3K9 methylation marks constitutive heterochromatin in pericentric and telomeric regions. Immunostaining of hetrochromatin protein1 (HP1) and H3K9me3 revealed a hyperdynamic and less compartmentalised structure in ESCs, indicating of chromatin reorganisation during differentiation [271]. This relatively diffused hetorochromatin structure is a functionally important hallmark of pluripotent stem cells, which helps to maintain plasticity and establish higher-order chromatin structure upon differentiation. It was shown that Oct4 regulates H3K9 methylation via upregulating *Jmjd1a* and *Jmjd2c*, which encode H3K9me2/3 demethylases [233]. Depletion of *Jmjd1a* and *Jmjd2c* resulted in down regulation of pluripotency genes and differentiation of ESCs [233].

The maintenance of pluripotency has been extensively studied over the past decades. However, up until now, there is still a lack of knowledge on the exact mechanism of the initial transition towards differentiation. With the advent of the CRISPR-Cas9 technology, I sought to perform a genome-wide knockout screen to study the exit of pluripotency in a comprehensive in-depth manner. In order to do that, specific investigation steps were undertaken. First, careful preparation and optimisation was conducted to establish the screening conditions. Second, the genome-wide CRISPR-Cas9-mediated screen was performed and result was analysed. Finally, the role of mTORC1-related pathways in the regulation of pluripotency and differentiation was investigated.