# Chapter 3

# Screening Preparation and Condition Optimisation

## 3.1 Introduction

Care must be taken with the design of a genome-wide screen to ensure the ultimate success of the endeavour. A successful screen must be robust and reproducible, with maximal sensitivity and minimal false-positives. Some of the major considerations are whether an assay needs to be used to measure the desired parameters, which marker to choose for that assay and whether the selected marker is truly indicative of that process. To monitor the very early stages of differentiation, I adopted the Rex1:GFPd2 reporter cell line kindly provided by the Smith lab. Two characteristics of the *Rex1* gene render it a desirable pluripotency marker. Firstly, as demonstrated by Masui et al., the Rex1 function is dispensable for both the development of a mouse embryo and the maintenance of ESCs[263]. Secondly, *Rex1* expression is tightly restricted to the naive pluripotency compartment and is rapidly downregulated at the onset of differentiation [417], therefore providing an accurate biological focus and faithful pluripotency readout. Loss of *Rex1* expression leads to loss of clonogenicity under 2i/LIF conditions, indicating the irreversible exit of pluripotency [30] [185]. Notably, the downregulation of other naive pluripotency markers such as *Nanog*, *Klf2* and *Tfcp2l1* upon differentiation is earlier than that of *Rex1*, but self-renewal capacity is retained as long as Rex1 is expressed [185]. The Smith group then generated the Rex1:GFPd2 reporter cell line, in which destabilised GFP with a half-life of 2 hours is expressed under the control *Rex1* promoter [450]. This reporter cell line enables almost real time monitoring of differentiation, hence providing great convenience in the fractionation of ESCs based on its naive pluripotency state by flow cytometry and further downstream analysis.

Screening parameters need to be designed to maximise the difference in gRNA abundance between treated and control samples. One of the crucial parameters is the time of examining, at which it is ideal to achieve a balance between assessing mutants with modest phenotypes and discarding a large number of outliers. In the context of a differentiation screen, conditions such as differentiation duration and FACS cut-off are crucial to success. Given the substantial complexity of the genome-wide gRNA library, it is challenging to maintain a sufficient coverage of gRNA representation during the screen. Therefore, a relatively simple differentiation protocol is favourable, especially when handling a large number of cells. A monolayer neuroectodermal differentiation method has been used in previous genetic screens to identify pluripotency regulators [469] [142] [217] [19], in which ESCs were induced to differentiate following LIF and inhibitors withdrawal. The same

can be applied in the CRISPR-Cas9 based screen and mutant candidates can be easily identified based on the persistence of Rex1GFP expression. Last but not least, it is noteworthy that there is natural variation between each cell as well as between each screen. Such variation is unavoidable and difficult to eliminate, but measures can be taken to keep it in an acceptable range. For example, it is worth checking the Cas9 activity to make sure it is uniformly active in the chosen cell line, so that the phenotype linked to a gRNA will not be masked by inactive Cas9. Including biological replicates will also help to reduce stochastic noise. Feasibility and cost needs to be taken into account when choosing the numbers of replicates.

In this chapter, I describe the preparation work for the genome-wide CRISPR-Cas9 knockout screen, including analysis of culture and differentiation conditions, generation of constitutive Cas9-expressing Rex1:GFPd2 cell line, as well as proof-of-principle studies which target well-studied pluripotency-regulating genes, namely *Tcf7l1* and *Apc*. My aim was to optimise the screening conditions to achieve the highest possible sensitivity and robustness.

## 3.2 Results

### 3.2.1 Analysis of self-renewal and differentiation conditions

Establishing the pre-screening culture condition is important to achieve a successful differentiation screen. It is preferable to maintain ESCs as a relatively homogeneous population to minimise any biases prior differentiation. I first examined the Rex1GFP expression in serum/LIF and 2i/LIF culture, with or without feeder layer (Figure 3.1). Consistent with previous findings, *Rex1* expression was homogeneous in 2i/LIF but rather heterogeneous in serum/LIF. Notably, ESCs cultured in serum/LIF without feeders were inclined to lose *Rex1* expression, indicating a less stable pluripotency status. It is thus likely that feeders play additional roles in promoting self-renewal beyond contributing LIF. As expected, the presence of feeders did not make a difference in *Rex1* expression in 2i/LIF. As 2i effectively insulates any differentiation signals, so that the pluripotency status is rather stable without any further support. The heterogenous ESC population became homogeneous after two days of culture in 2i/LIF, suggesting that the *Rex1*-negative primed cells can be either reversed to naive pluripotency or eliminated from the population, and that the two culture conditions are convertible. In observing the above, I decided to maintain ESCs in 2i/LIF before screening to achieve a feeder-free homogeneous population prior induction

of differentiation.

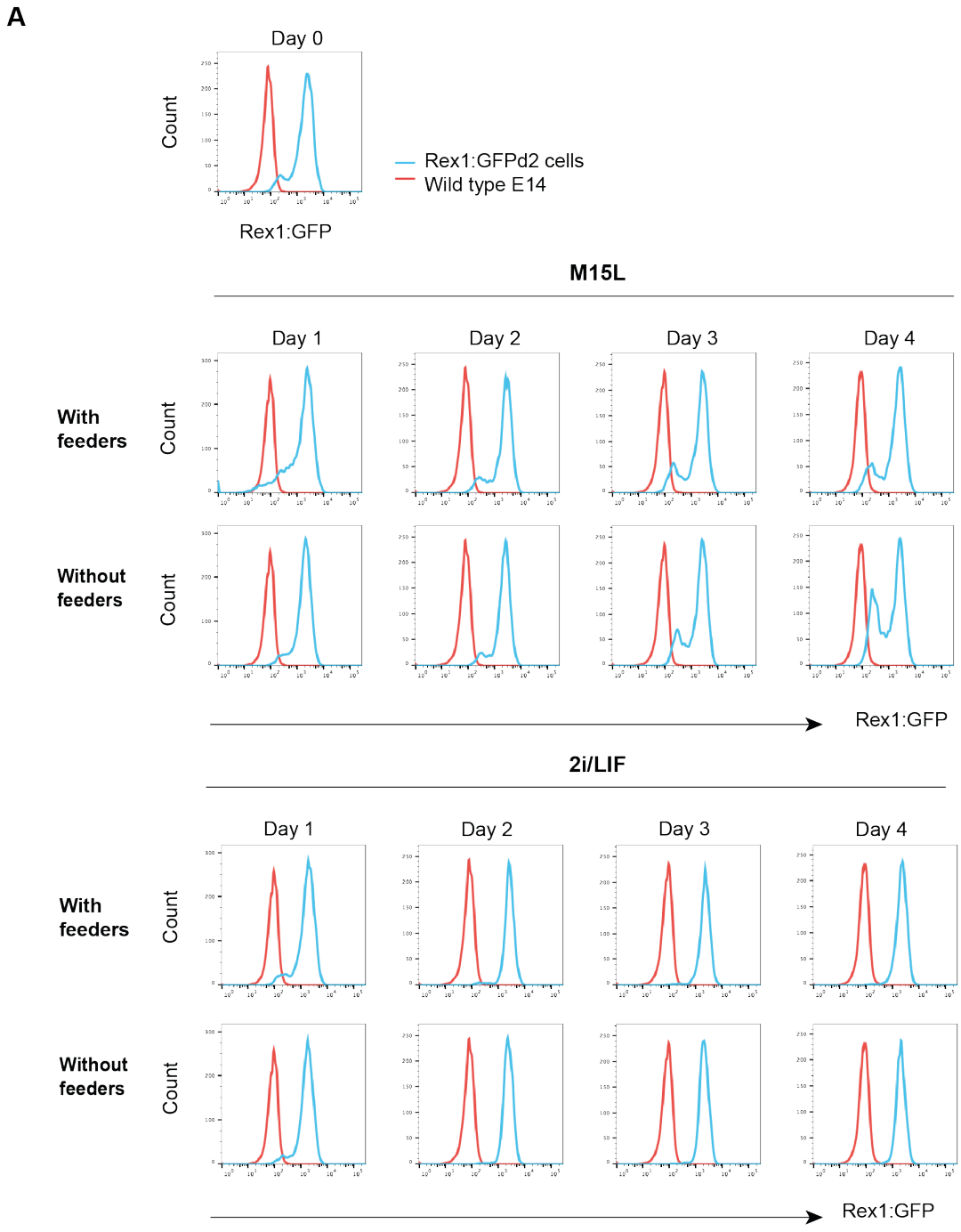


**Figure 3.1:** Analysis of Rex1GFP profile under maintenance conditions. (A) Comparison of Rex1GFP profile under different maintenance conditions. ESCs were cultured under serum/LIF condition on feeders (Day 0), and were subsequently split into four maintenance conditions: serum/LIF with or without feeders, and 2i/LIF with or without feeders. The Rex1GFP profiles of cells in each condition were measured everyday for four days. Blue - Rex1:GFPd2 cells; Red - GFP negative control, wild type ESC line E14.

Following 2i withdrawal, the pluripotency network collapsed quickly, which was reflected by the rapid down regulation of *Rex1* expression (Figure3.2 (A)). The most drastic transition occurred in the first three days, during which the majority of the cells lost *Rex1* expression. Taking the large complexity of the library into consideration, I then checked cell growth during differentiation to plan the scale of the screen and to ensure there was sufficient coverage of the gRNA representation and enough cells at the end of the screen for the extraction of genomic DNA. Cell number increased throughout the differentiation period until a plateau was reached on day five, indicating a healthy differentiation occurred in these conditions (Figure3.2 (B)).

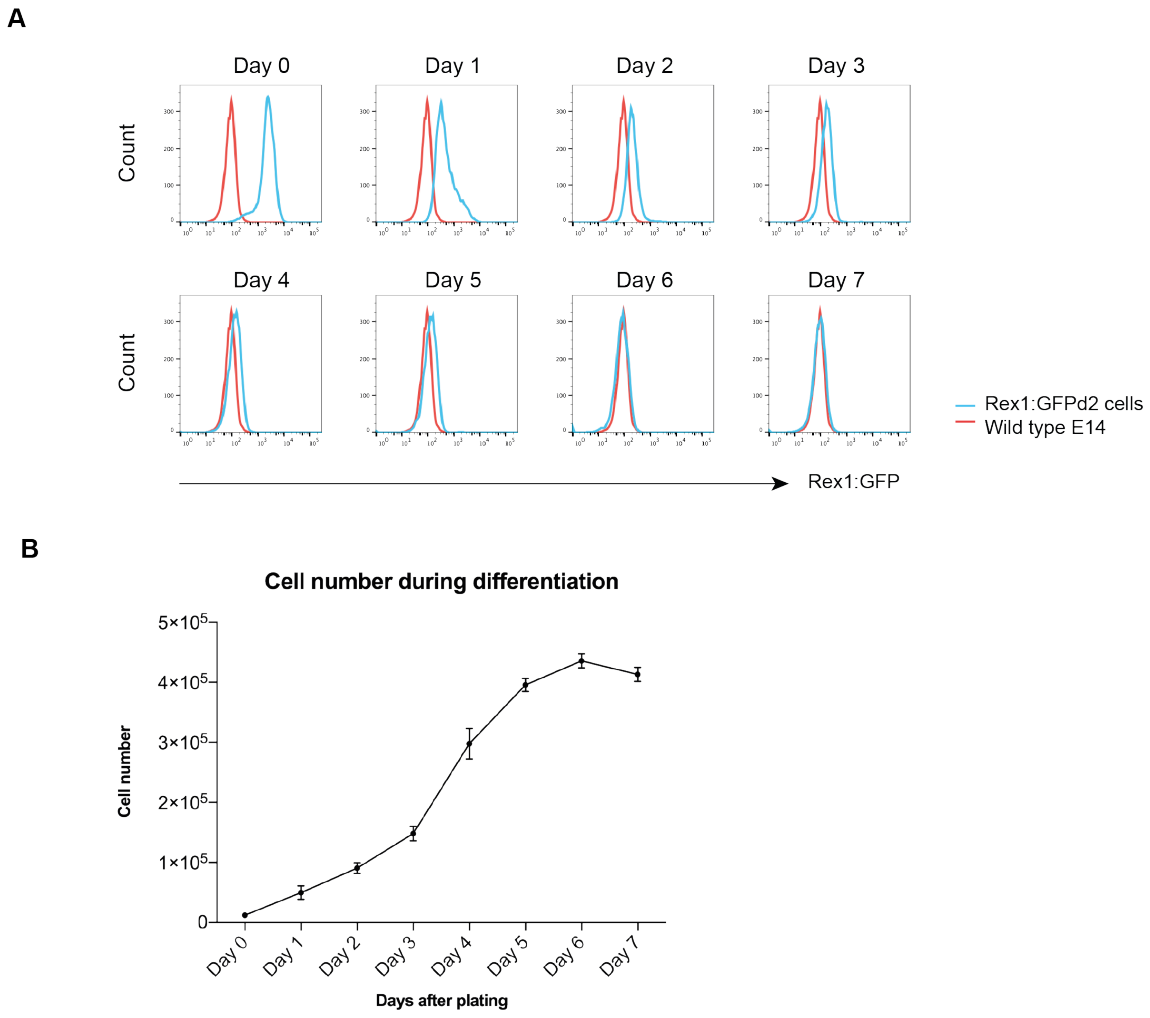


**Figure 3.2:** Analysis of Rex1GFP profile under differentiation conditions. (A) Rex1GFP expression after 2i withdrawal. The removal of 2i relieved ESCs from the shield of differentiation cues, which initiated spontaneous differentiation. ESCs were maintained in N2B27 with supplements but without 2i (described in Chapter 2). Rex1GFP expression profile was measured everyday for seven days. Blue - Rex1:GFPd2 cells; Red - GFP negative control, wild type ESC line E14. (B) ESC growth curve after 2i withdrawal. ESCs were plated on gelatin in N2B27 without LIF at a density of 10,000 cells per $cm^2$. Cells were detached everyday for enumeration. Cell number per $cm^2$ was calculated.

### 3.2.2 Establishment of Cas9 expression in Rex1:GFPd2 cell line

In order to perform the CRISPR-Cas9-based mutagenesis in the Rex1:GFPd2 cell line, I performed homologous recombination-mediated Cas9 knockin at the *Rosa26* locus using the previously published targeting vector [202] (Figure3.3 (A)). G418-resistant colonies were picked and further expanded for genotyping. Out of 24 colonies analysed, 13 showed correct PCR bands (Figure3.3 (B)).

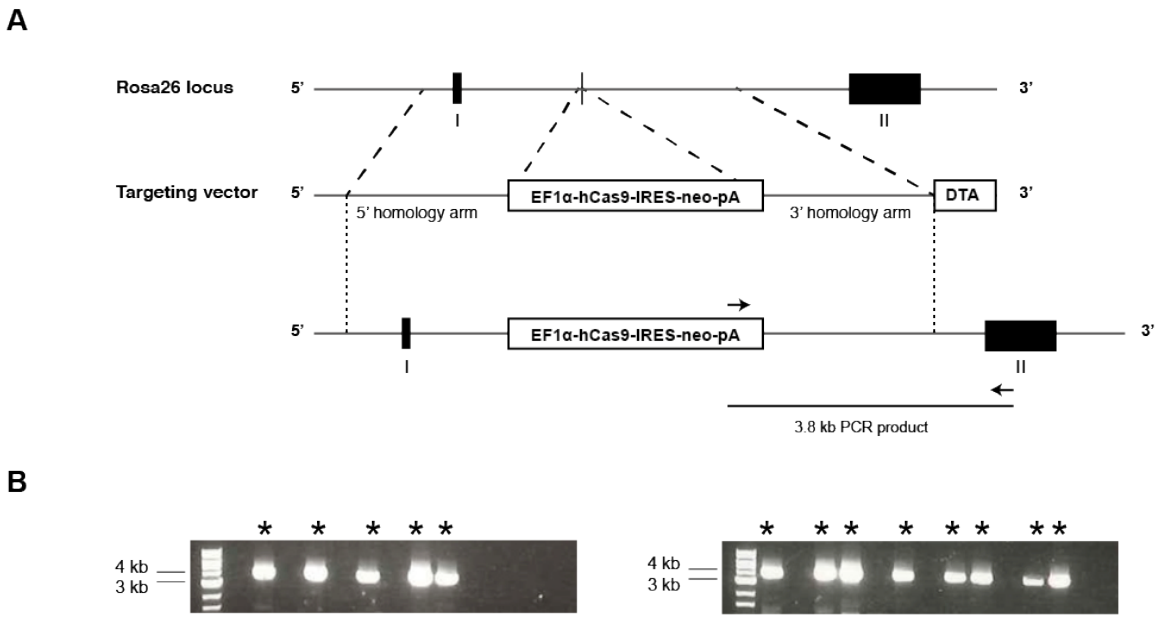


**Figure 3.3:** Cas9 knockin. (A) Schematic diagram showing the targeting scheme at the *Rosa26* locus. EF1α: elongation factor-1αpromoter; hCas9: human codon-optimised SpCas9; IRES: internal ribosomal entry site; neo: neomycin resistant gene; DTA: diptheria toxin A. Arrows indicate PCR primers. (B) PCR genotyping results. PCR bands were compared to a 1kb DNA ladder (NEB). Clones with successful knockin will produce a 3.8kb PCR product, which were labeled with asterisk.

To investigate whether Cas9 was functional in these clones, I carried out a reporter assay developed in the lab using a lentiviral vector expressing BFP, GFP, and gRNA targeting the GFP sequence (Figure3.4 (A)). An 'empty' vector, which expresses BFP and GFP but not the gRNA sequence, was included in the assay as a negative control. Transduced cells should be double positive when the 'empty' vector is used, regardless of its Cas9 function. When the vector containing gRNA targeting GFP is introduced, only BFP can be detected in Cas9-active cells, whereas both BFP and GFP will be detected in Cas9-inactive cells (Figure3.4 (B)). Two clones were analysed for Cas9 function, in which fluorescent signal was exclusively detected in the BFP quadrant, indicating positive Cas9 activity in both clones. Clone 9, which has more consistent colony morphology, was selected for further

studies (Figure3.4 (C)).

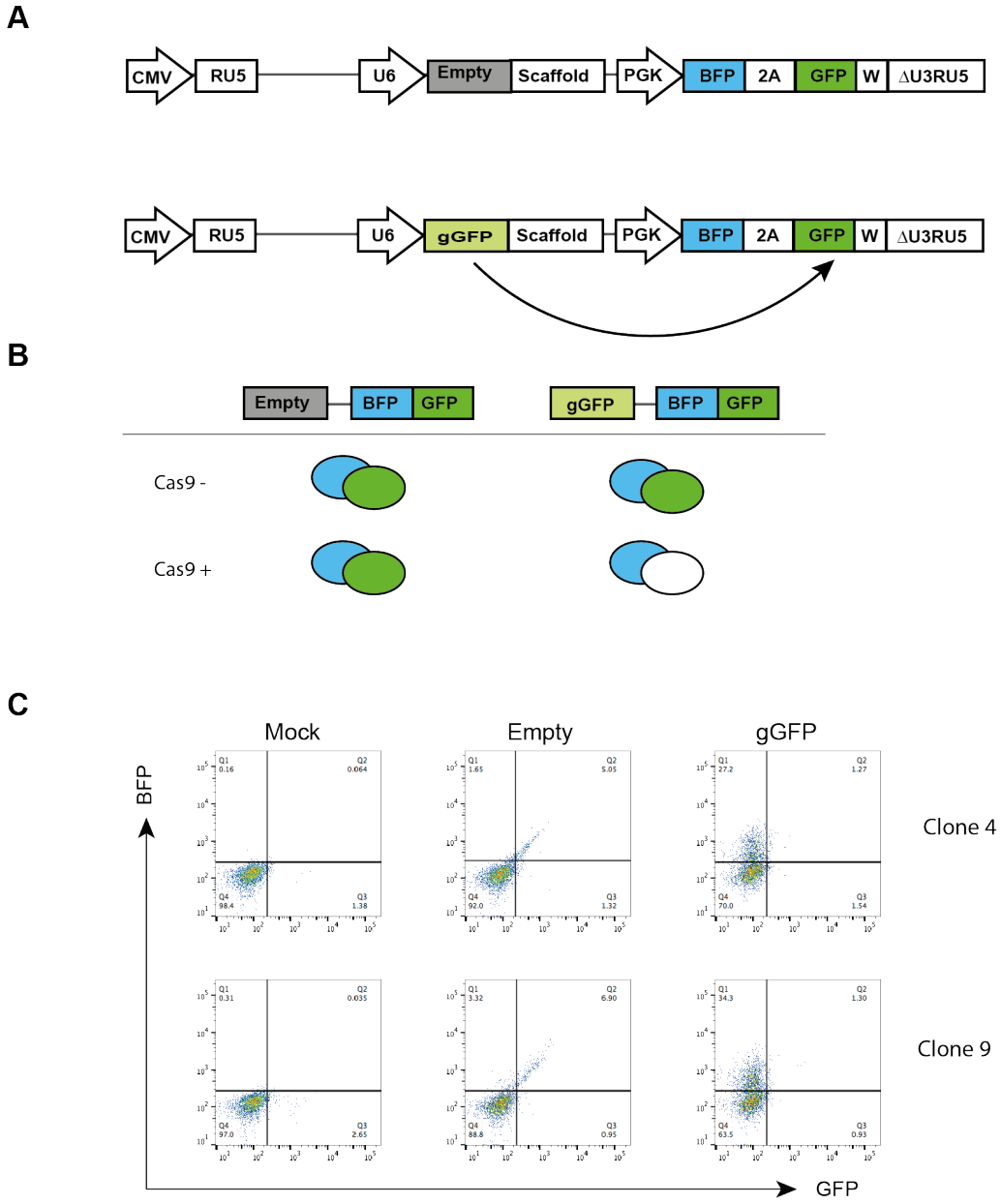


**Figure 3.4:** Cas9 function assay (A) Lentiviral vector for Cas9 function reporter assay. CMV: CMV promoter; RU5: 5′ long terminal repeat; U6: U6 promoter; gGFP: gRNA targeting GFP coding sequence; Empty: The original BbsI cloning site; scaffold: gRNA scaffold; PGK: mouse phosphoglycerate kinase 1 promoter; BFP: blue fluorescent protein; GFP: green fluorescent protein; 2A, Thosea asigna virus 2A peptides; W: Woodchuck Hepatitis Virus posttranscriptional regulatory element; ΔU3RU5: self-inactivating 3′ long terminal repeat. Empty gRNA vector doesn't express gRNA but expresses BFP and GFP. gGFP vector expresses gRNA targeting GFP, as well as BFP and GFP. (B) The expected fluorescent expression pattern. If Cas9 nuclease was inactive, cells transduced with empty and gGFP vectors would be GFP/BFP double positive. If Cas9 was functional, cells transduced with empty vector would be double positive, whilst cells transduced with gGFP vector will express BFP only. (C) The Cas9 function of two knockin clones were analysed. Culture medium was added to mock infection instead of virus supernatant. Positive Cas9 activity was observed in both clones.

### 3.2.3 Proof of principle studies

With the establishment of the constitutive Cas9-expressing Rex1:GFPd2 cell line, I moved on to design a screening strategy which includes testing differentiation and duration conditions. To achieve that, two well-studied genes were selected as positive control genes, namely *Tcf7l1* (Tcf3) and *Apc*. Tcf3 acts as a pluripotency repressor and Apc is a subunit of the β-catenin degradation complex downstream of Wnt. Knocking out *Tcf7l1* and *Apc* relieves the suppression of pluripotency genes, therefore cells were expected to exhibit enhanced self-renewal and delayed differentiation phenotype.

I first sought to generate stable knockout cell lines by deleting a 'critical exon', which is defined as a common exon expressed in all transcript variants and when deleted, creates a frame-shift mutation. Deletion was achieved by introducing DSBs on both sides of the critical exon by the CRISPR-Cas9 system. Knockout clones were identified by PCR genotyping (Figure3.5 (A) (C)). The selected clones were assessed under two conditions: serum-based and N2B27-based differentiation conditions. As expected, under both differentiation conditions, *Tcf7l1* knockout and *Apc* knockout showed impeded differentiation phenotype, reflected by their delayed downregulation of *Rex1* expression (Figure3.5 (B) (D)). It appears that N2B27-based differentiation induced a more rapid decrease in Rex1GFP expression compared to serum-based differentiation, especially in *Tcf7l1* knockout. This is probably due to the distinct differentiation mechanisms. In serum-based condition, ESCs differentiate via a mixed routes towards mesoderm, endoderm and trophectoderm, whereas in N2B27 without 2i/LIF, ESCs mainly differentiate towards neuroectoderm. N2B27-mediated differentiation has been adopted in large-scale RNAi- and transposon-mediated genetic screens, which can be used as reference for CRISPR-Cas9-mediated screen [19] [217]. It will also be interesting to cross compare different screening methods. For these reasons I decided to focus on N2B27-mediated differentiation.

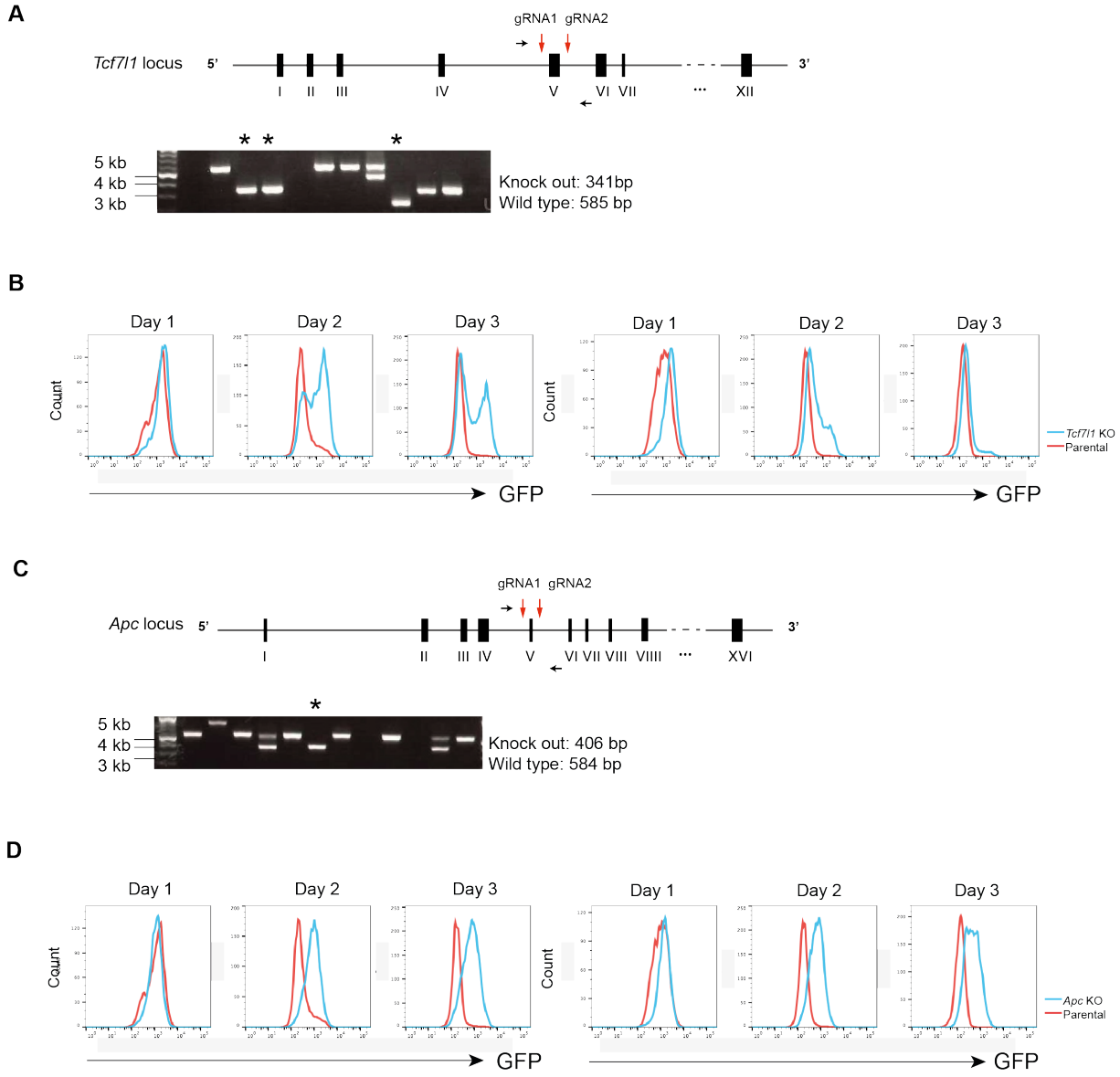


**Figure 3.5:** Positive control study with stable knockout clones. (A) Schematic diagram of critical exon deletion at the *Tcl7l1* locus and genotyping results. Black arrows indicate PCR genotyping primers. Red arrows indicate gRNA cutting sites. Asterisk indicate knockout clones. Wildtype PCR product was 585bp. Knockout PCR product was 341bp. (B) *Tcl7l1* knockout clone Rex1GFP expression profile during serum-based differentiation (left) and neuroectoderm differentiation (right). Serum-based differentiation was carried out by LIF withdrawal from serum/LIF maintenance condition. Neuroectoderm differentiation was carried out by 2i/LIF withdrawal from N2B27. Rex1GFP expression was analysed everyday for three days. Blue - *Tcl7l1*/*Apc* knockout Rex1:GFP cells; Red - wt Rex1:GFP cells. (C) Schematic diagram of critical exon deletion at *Apc* locus and genotyping result. Black arrows indicate PCR genotyping primers. Red arrows indicate gRNA cutting sites. Asterisk indicate knockout clones. Wild type PCR product was 584bp. Knockout PCR product was 406bp. (D) *Apc* knockout clone Rex1GFP expression profile during serum-based differentiation (left) and neuroectoderm differentiation (right). Serum-based differentiation was carried out by LIF withdrawal from serum/LIF maintenance condition. Neuroectoderm differentiation was carried out by 2i/LIF withdrawal from N2B27. Rex1GFP expression was analysed everyday for three days. Blue - *Tcl7l1*/*Apc* knockout Rex1:GFP cells; Red - wt Rex1:GFP cells.

In the actual screening setting, lentivirus is used as a vehicle to deliver the gRNA library and the phenotype generated by each gRNA is assessed as a mixed population with indels of various sizes. To recapitulate this, I transduced ESCs with lentivirus expressing gRNAs targeting the critical exon of *Tcf7l1* and *Apc* individually (Figure3.6 (A) (B)), followed by differentiation in N2B27 without 2i. The resulted phenotype was compared to that of the cells transduced with gRNA-negative lentivirus. Consistent with the phenotype of pure knockouts, a clear difference in Rex1GFP expression was observed between *Tcf7l1*/*Apc* targeted cells and the empty control, on differentiation day 2 and day 3 (Figure3.6 (C)). I reasoned that day 3 is the optimal time to harvest cells for gRNA representation analysis. Because vast majority of wild type cells were Rex-GFP negative on day 3, which produces a clear contrast with cells showing delayed differentiation, and a cleaner background will be obtained.

### 3.2.4 A Preliminary screen

Once the screening principle was verified by knockout of *Tcf7l1* and *Apc* via lentivirally expressed gRNAs in Rex1:GFPd2 Rosa26:Cas9 cells, I then sought to perform a preliminary screen in order to study the scale-up effect and further optimise screening conditions. Due to the reform of the manufacturing company Stem Cell Inc at the time, there was a long delay in purchasing the the basal media N2B27. Therefore, I carried out the preliminary screen based on serum differentiation. Although the mechanisms of differentiation differ, the fundamental principle and design of the screen remain the same.

The Rex1:GFPd2 Rosa26:Cas9 cells were transduced with the genome-wide lentiviral gRNA library. The transduced cells were sorted according to BFP expression two days after infection, followed by three days of expansion before plating in serum-containing medium without LIF for differentiation. Differentiation medium was replenished on day two. After three days LIF withdrawal, cells retaining Rex1GFP expression were collected by FACS and subsequently pelleted for genomic DNA extraction. The gRNA sequences were amplified by PCR from genomic DNA and sent for sequencing with the Illumina Hiseq platform. A technical replicate was carried out in parallel. Sequenced gRNAs were mapped to the library and counted, and statistical analysis was performed by the computational algorithm MAGeCK.

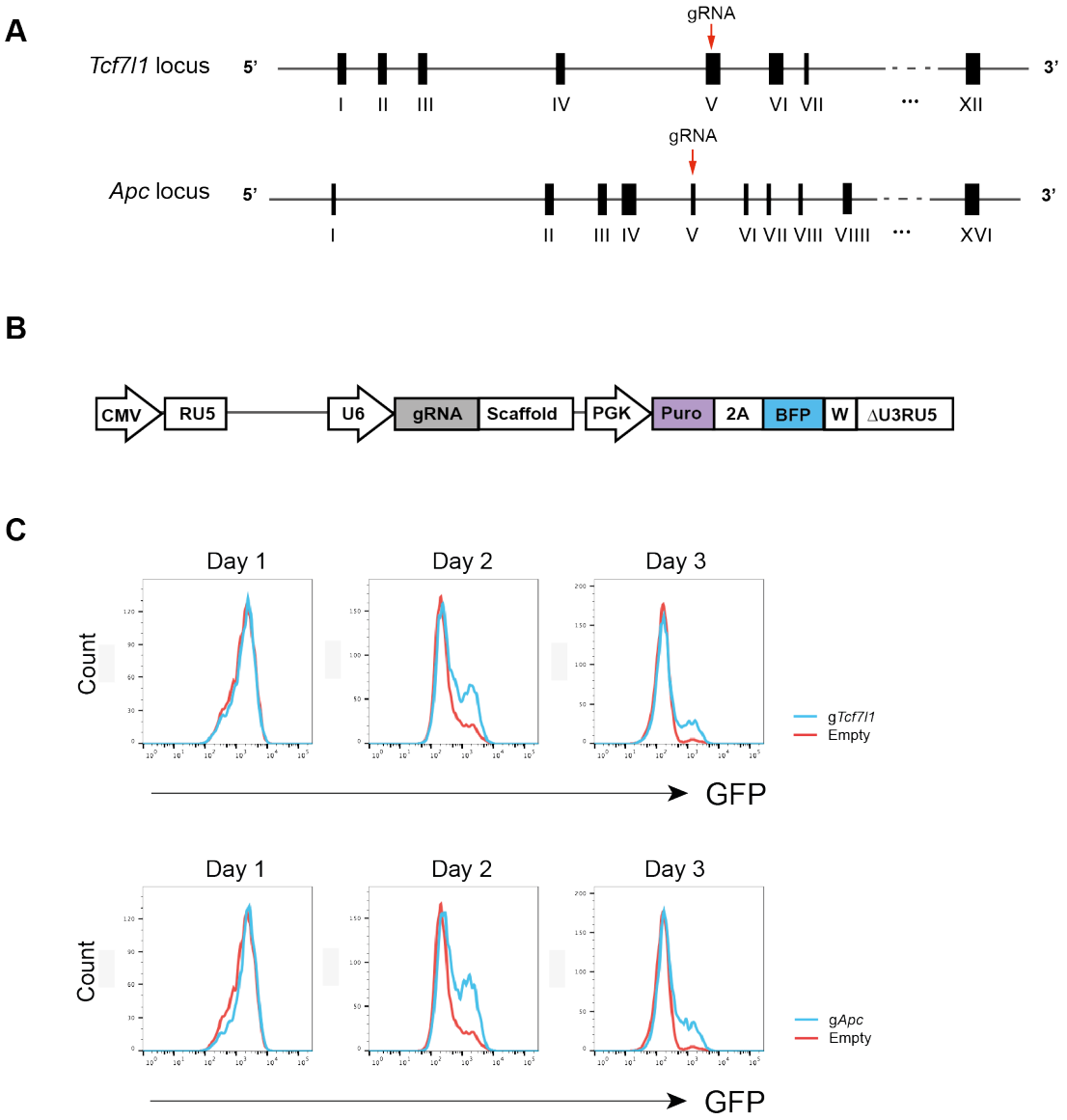


**Figure 3.6:** Positive control study with single-gRNA knockouts. (A) Schematic diagram of gRNAs targeting the critical exon of *Tcf7l1* and *Apc*. Red arrows indicate gRNA cutting sites. The gRNA sequence is included in Chapter2 section 2.1.2.4. (B) Lentiviral vector for single gRNA knockout. CMV: CMV promoter; RU5: 5′ long terminal repeat; U6: U6 promoter; gRNA: gRNA targeting *Tcl7l1* and *Apc*; PGK: mouse phosphoglycerate kinase 1 promoter; Puro: puromycin resistant gene; BFP: blue fluorescent protein; 2A, Thosea asigna virus 2A peptides; W: Woodchuck Hepatitis Virus posttranscriptional regulatory element; ΔU3RU5: self-inactivating 3′ long terminal repeat. (C) *Tcl7l1* and *Apc* knockout differentiation profiles. Differentiation was induced by 2i and LIF removal. Rex1GFP expression was analysed everyday for three days. Blue - *Tcl7l1*/*Apc* knockout Rex1:GFP cells; Red - wt Rex1:GFP cells.

Once the screening principle was verified by knockout of *Tcf7l1* and *Apc* via lentivirally expressed gRNAs in Rex1:GFPd2 Rosa26:Cas9 cells, I then sought to perform a preliminary screen in order to study the scale-up effect and further optimise screening conditions. Due to the reform of the manufacturing company Stem Cell Inc at the time, there was a long delay in purchasing the the basal media N2B27. Therefore, I carried out the preliminary screen based on serum differentiation. Although the mechanisms of differentiation differ, the fundamental principle and design of the screen remain the same.

The Rex1:GFPd2 Rosa26:Cas9 cells were transduced with the genome-wide lentiviral gRNA library (Figure3.7 (B)). The transduced cells were sorted according to BFP expression two days after infection, followed by three days of expansion before plating in serum-containing medium without LIF for differentiation. Differentiation medium was replenished on day two. After three days LIF withdrawal, cells retaining Rex1GFP expression were collected by FACS and subsequently pelleted for genomic DNA extraction. The gRNA sequences were amplified by PCR from genomic DNA and sent for sequencing with the Illumina Hiseq platform. A technical replicate was carried out in parallel. Sequenced gRNAs were mapped to the library and counted, and statistical analysis was performed by the computational algorithm MAGeCK.

A Quantile-Quantile (Q-Q) plot was generated using the gene level $P$-values calculated by MAGeCK (Figure3.7 (B)). The distribution of data points followed a curved pattern with an increasing slope that diverged from the expected $P$-value distribution, suggesting that a large number of genes showed statistical significance. Such distribution was expected based on the results of similar screens done in the past. The exit of pluripotency screen performed by Betschinger et al. has identified 70 genes that passed stringent Z score cut-off [19]. Two of other similar screens carried out by Yang et al. and Leeb et al. have revealed 272 and 113 hits with high confidence [217] [464]. Although only part of these candidate lists was validated, it suggests that there are many genes involved in the process of differentiation initiation. Therefore, a highly skewed $P$-value distribution with a large group of genes showing some degree of statistical significance would be expected. As mentioned previously, the FGF/ERK pathway and WNT/GSK3 pathway play crucial roles in the exit from pluripotency. Reassuringly, the preliminary screen identified mutants from both pathways such as Fgfr1, Mapk1 from the FGF/ERK pathway, and Tcf7l1, Crebbp and Ctbp1 from the WNT/GSK3 pathway (Figure3.7 (C)). The recovered hits also include genes that were identified and validated from previous screens, such as Flcn, Tsc2, and Hira. Screening candidates that passed the statistical cut-off were analysed

for enrichment of Gene Ontology (GO) terms. The results demonstrated a significant enrichment in terms related to embryo development and regulation of gene expression, which corroborated the hyperactive transcription status of the transition period from pluripotency to differentiation (Figure3.7 (D)).

## 3.3 Discussion and Conclusion

This chapter has described the preparation work and proof-of-principle studies for the set-up of the genome-wide CRISPR-Cas9-based exit of pluripotency screen. In verifying the knockout phenotype of *Tcf7l1* and *Apc* in the Rex1:GFPd2 Rosa26:Cas9 cell line, I was ready to perform the screen.

The proof-of-principle studies have provided valuable insights into screening design. The rapid downregulation of *Rex1* expression suggests that the dissolution of pluripotency takes place very soon after inhibitor withdrawal. To capture this fast-happening event, I reasoned that it is better to terminate the screen on day two or day three of differentiation. The Rex1GFP flow profile of *Tcf7l1* and *Apc* knockouts also suggests that the most drastic difference between knockout and wild type control occurs on day 2 or day 3. A few screens conducted by other groups have adopted the strategy of several rounds of replating and enrichment, which may not be necessary for CRISPR-Cas9-mediated screen. Given the high efficiency of gRNA and its convenience in mutant identification, the CRISPR-Cas9-mediated screen is highly sensitive to detect subtle changes in gRNA representation. Furthermore, because the gRNA counts are available which allows statistical analysis, the assessment of a specific phenotype is no longer 'black and white', but can be evaluated in a quantitative way. Another concern over several rounds of prolonged enrichment is the depletion of fitness genes or essential genes, which are required for ESC survival but may also be involved in pluripotency regulation. Therefore, I decided to perform a screen as a short-course one-round of differentiation, after which cells showing differentiation defects will be collected and sequenced.

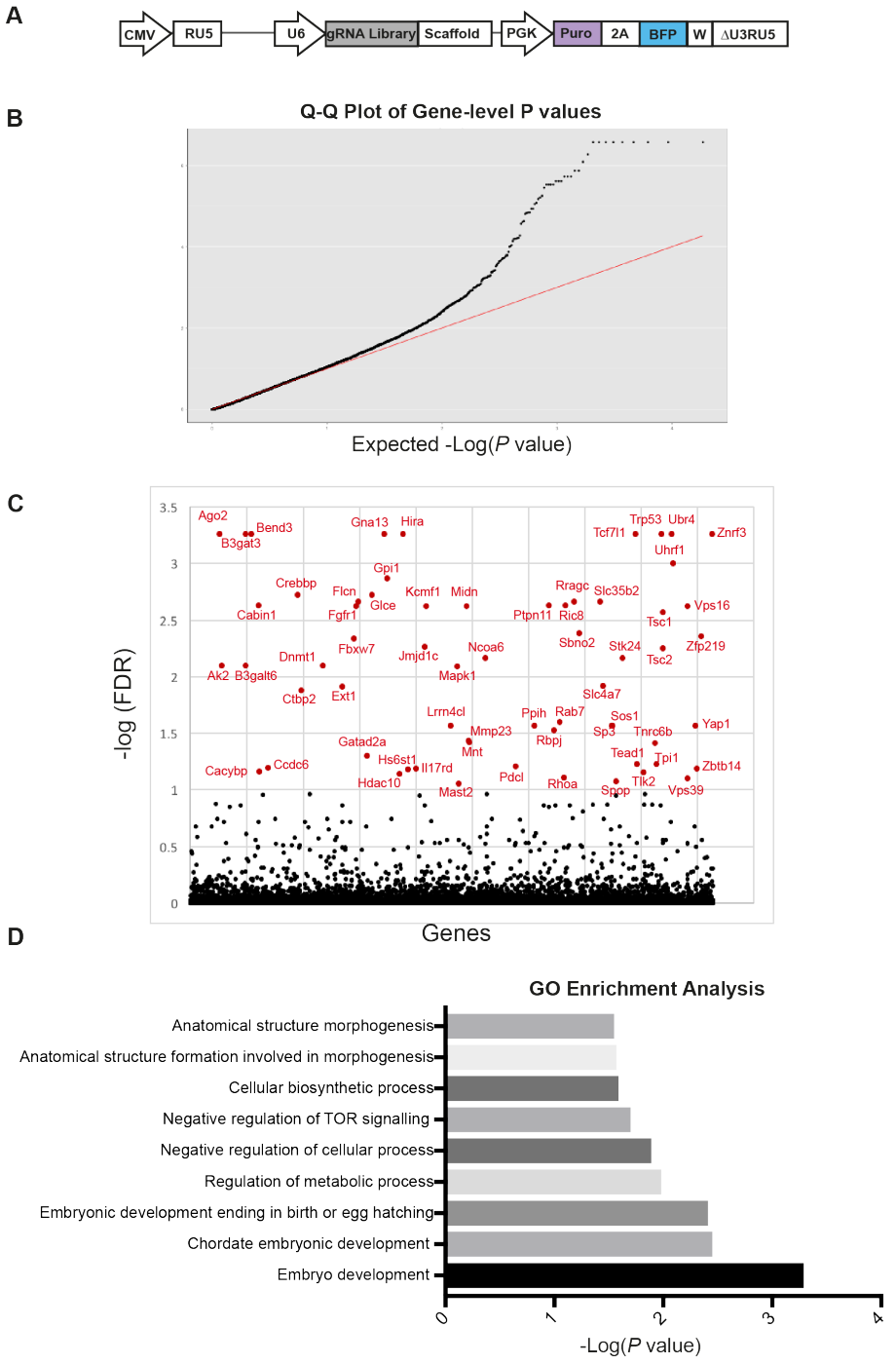


**Figure 3.7:** Preliminary screen result analysis. (A) RNA library expression vector. CMV: CMV promoter; RU5: 5′ long terminal repeat; U6: U6 promoter; gRNA library: gRNA sequence from the mouse V2 library; scaffold: gRNA scaffold; PGK: mouse phosphoglycerate kinase 1 promoter; Puro: puromycin resistant gene; BFP: blue fluorescent protein; 2A, Thosea asigna virus 2A peptides; W: Woodchuck Hepatitis Virus posttranscriptional regulatory element; ΔU3RU5: self-inactivating 3′ long terminal repeat. (B) Gene-level *P*-values Quantile-Quantile (Q-Q) plot. Gene level *P*-values were calculated by the published algorithm MAGeCK. The *P*-values were sorted in ascending order, and then plotted versus quantiles calculated from a theoretical distribution. A 95% confidence interval was used. A 45-degree reference line (red) was also plotted. Points would fall along the reference line if the observed *P*-values were randomly distributed. The Q-Q plot was generated using the gglot2 package from software R. (C) Gene-level false discovery rate (FDR). The FDR values were calculated by MAGeCK. X-axes represents genes ranked in alphabetical order. Y-axes represents -log(FDR). Genes with FDR $< 0.1$ were labelled red. (D) Gene ontology term analysis of statistically significant hits. Cutoff was defined as FDR $< 0.1$. The analysis was performed using the online GO analysis algorithm PANTHER (http://pantherdb.org/).

The preliminary screen reassured screening strategy, helped to study the scale-up effect and provided material for a data analysis run-through. An important parameter to be decided is the statistical cut-off, which is essential to reduce the number of false positives and restrict the scale of subsequent validation. Setting the cut-off at $P$-value $< 0.05$ resulted in 1095 significant candidates, which is likely to include high proportion of false positives, and were obviously impractical for further validation purposes. Also, because the screening results were generated from a genome-wide study, where thousands of hypothesis tests were conducted simultaneously, small $P$-values may occur by chance. It is thus more accurate to use the false discovery rate (FDR) to control false positives. To set a decent threshold, the enrichment fold change in the control experiment can be used as a reference, where *Tcf7l1* and *Apc* KO cells were enriched in the Rex1GFP positive population by 4.9 times and 4.3 times respectively after three days of differentiation. Assuming the selection criteria were based on FDR $< 0.1$, the shortlist of potential candidate genes were reduced to 64. *Mast2*, the candidate gene at the cut off point, has an average of 1.8 times and 3.9 times enrichment in each replicate at the gRNA level, which is slightly lower compare to the enrichment effect in the positive control. However, *Tcf7l1* and *Apc* are genes with relatively strong phenotypes. Therefore, to reduce the chance of having false negatives, the cut-off should be selected to allow the test to pick up potential candidates with weaker phenotypes. Hence, FDR $< 0.1$ would be an appropriate threshold to start with.