CHAPTER 3

DEVELOPMENT OF A CRISPR-CAS9 BASED KNOCK-OUT SCREEN FOR CELLULAR RECOGNITION

3.1 Introduction

This chapter describes the design of a genome-scale knockout (KO) screening approach using the CRISPR-Cas9 system in human cell lines. The approach is subsequently used to screen a panel of monoclonal antibodies targeting structurally diverse membrane receptors to assess the potential of the approach and to determine the experimental parameters to identify cellular factors mediating cellular recognition events at the cell surface.

3.1.1 Monoclonal antibodies are ideal probes for the study of cell surface recognition

To first assess the feasibility of genome-scale KO screens for the identification of factors required for cell surface recognition events, I focused initially on the interactions mediated by monoclonal antibodies (mAbs). mAbs have already been established as powerful tools to investigate various aspects of receptor biology as exemplified by their use in the past to isolate, localise, and biochemically characterise membrane receptors, as described in Chapter 1. As mAbs are selected to bind specifically to a single epitope within a receptor with a high avidity (K_{Ds} in the nM to pM range), the binding of mAbs on a cell surface that abundantly expresses the corresponding antigen often leads to a bright staining with a high signal to noise ratio. In a CRISPR-Cas9 screen, this can be the basis for Fluorescence Activated Cell Sorting (FACS)-based phenotypic selection, allowing identification of factors, which if disrupted causes a decrease in the cell surface expression display of the mAb epitope. Strong phenotypic selections are usually desired in a pooled screening approach as they allow for the identification of candidate genes with

high confidence, thus the use of mAbs as probes will be ideal to determine the parameters of CRISPR-KO screens.

3.1.2 Genome-scale knockout screening approach has the potential to determine the specificity of mAbs.

The use of mAbs in a genome-scale KO screening approach in itself has the potential to become a novel means of determining the specificity of a given antibody. High quality antibodies should recognise their targets specifically but the batch-to-batch variability of the commercially available antibodies can lead to poor specificity, and has caused concerns in the research field [194, 195, 196, 197]. The need for target validation of mAbs is even higher now as there is a growing use of mAbs as therapeutics in human diseases such as cancer, rheumatoid arthritis and Crohn's disease [198]. Common methods to characterise antibody specificity involve immunoassays such as immunohistochemistry (IHC), immunoprecipitation (IP), Western blotting (WB), and mass spectrometry (MS) analysis [199, 200]. However, such approaches are not always applicable, especially if the mAbs are raised against the correctly folded protein epitopes such that they do not react to denatured or detergent-solubilised antigens. The genetic strategies discussed earlier in Chapter 1 to identify binding partners, such as the gain-of-function approach using a cDNA library and the loss-of-function screens using RNAi-mediated knockdown, have been used to identify targets of mAbs that recognise folded epitopes that are correctly expressed on the surface of cells [201, 202, 203]. However, the high off-target effects of RNAi systems that often lead to inconsistent results and the resource- and time-intensive nature of expression cloning approaches using cDNA libraries pose potential challenges in their applications. Genetic screening methods using the CRISPR-Cas9 KO system holds potential for identification of receptors targeted by mAbs, which could be valuable in the field of antibody characterisation. This will be explored in detail in this chapter.

3.1.3 Considerations for knockout screening approach to identify directly interacting receptors

One of the considerations for the use of genome-scale KO screens to identify directly interacting receptors is the possibility that cellular processes that contribute to general protein transport, would dominate the identified genes in all the screens and decrease the likelihood of identifying the direct receptor. Some of the cellular factors that transport plasma membrane-destined membrane

receptors from the endoplasmic reticulum (ER) lumen via the Golgi to the cell surface include: proteins of the signal recognition particle (SRP)-dependent protein targeting pathway, which mediates co-translational targeting of newly synthesised polypeptides into the ER; enzymes within the ER, which modify the polypeptides through signal peptide cleavage and initial glycosylation (also called 'core glycosylation'); and proteins that package folded polypeptides into specific vesicles for intracellular transport between organelles [204]. The disruption of these factors would affect the membrane expression of a large number of membrane proteins and therefore these factors are expected to be identified in the majority of the screens. However, many genes involved in these pathways (specifically the genes encoding the SRP-dependent protein translocation pathway proteins and the core glycosylation pathway proteins) are also known to be essential for the cells, and multiple negative selection CRISPR-KO based screens have shown that gRNAs targeting these genes usually drop-out of the mutant pool as the cells are continually grown [175, 176, 177, 205]. This would instead decrease the likelihood of these genes dominating the identified hits in the screen. In this chapter, I investigate how the experimental parameters in terms of the timing and the stringency of selection can influence the identification of genes that generally contribute to receptor expression versus genes that encode the directly interacting receptor.

3.1.4 Knockout approach used in this study

A dual vector approach is used in this study to generate genome-wide mutants. In this approach, stable Cas9-expressing cell lines are first created and then transduced with the lentiviral knockout library to generate the cell mutant library. The lentiviral library is generated from a plasmid pool obtained from Kosuke Yusa ('Yusa library'). This library consists of 90,709 gRNA targeting 18,009 genes (approximately five gRNAs/gene). The library was designed with features that have been shown to improve gRNA efficacy, such as the improved scaffold (iscaffold) on the gRNA. The conventional scaffold of a chimeric gRNA consists of a stretch of thymidines (T), which acts as a pause signal for RNA polymerase III that can potentially reduce the transcription efficiency [206]. The improved scaffold on the Yusa library is longer than the conventional scaffold by five nucleotides and avoids T stretches by mutating a single T residue within the stretch. The library has already been used for the study of essential genes in cells for the identifcation of genetic vulnerabilities in acute myeloid leukemia [176].

3.1.5 Scope of this chapter

In this chapter, I first describe the generation of a 'toolkit' required for setting up a genome-scale KO screening technology. This toolkit consists of: (a) human cell lines expressing highly efficient Cas9, and (b) a quality-controlled genome-scale library generated from the Yusa plasmid library. I next proceed to describe the optimal strategy for FACS-based phenotypic selection, before going on to describe multiple screens carried out with mAbs targeting structurally diverse membrane receptors, to identify cellular factors required for mAb binding to the cell lines. Finally, I will summarise the lessons learnt from this initial study.

3.2 Results

3.2.1 Generation of stable cell lines expressing Cas9

Determination of protein turnover time

To carry out CRISPR-Cas9 mediated KO screens, I first generated a stable Cas9-expressing HEK-293-E cell line using lentiviral transduction. I next designed a method to measure the Cas9 efficiency by making a gRNA construct expressing a single gRNA targeting *BSG*, which could be introduced to Cas9-expressing cells via lentiviral transduction (scheme for vectors used in figure 3.1A). Using this system, the decrease in cell surface expression of BSG in the transduced cells is readily established by flow cytometry using an anti-BSG mAb, which is then used as the measure of Cas9 efficiency. This flow-cytometric approach allows for rapid assessment of the efficiency of the individual cells in a pool rather than efficiency of the 'bulk' population. An important consideration in using this system is to wait long enough to allow complete protein turnover so that the previously transcribed mRNA and the translated protein is degraded. To determine the earliest timepoint where loss of the protein could be observed, polyclonal Cas9 cells infected with gRNA targeting *BSG* were checked for cell surface BSG expression on days 6, 8,10, 15 and 20, post transduction. While only a very small loss of cell surface BSG was observed at day six post infection, a clear population lacking the surface staining for BSG was observed from day eight onwards, demonstrating the expression of functional Cas9 in the cell line. However, approximately 25% cells retained surface expression of BSG on day eight and this knockout refractory population did not change even when examined at day 20 (figure 3.1B).

To assess whether eight days was a typical turnover time for other surface receptors, three more guides targeting *SDC1*, *CD55*, and *CD44*, as well as an empty construct with no targeting gRNA were cloned in the BbsI site of the reporter vector and was used to infect the Cas9-expressing HEK-293-E cells. The level of the respective surface proteins was also tested on day eight post transduction. A clear loss of the surface proteins were observed in all cases. Cells refractory to gene knockout were also observed in all cases (figure3.1C). Collectively, day eight was determined to be the earliest time-point that could be used to test the efficiency of Cas9 using the endogenous gene knockout system.

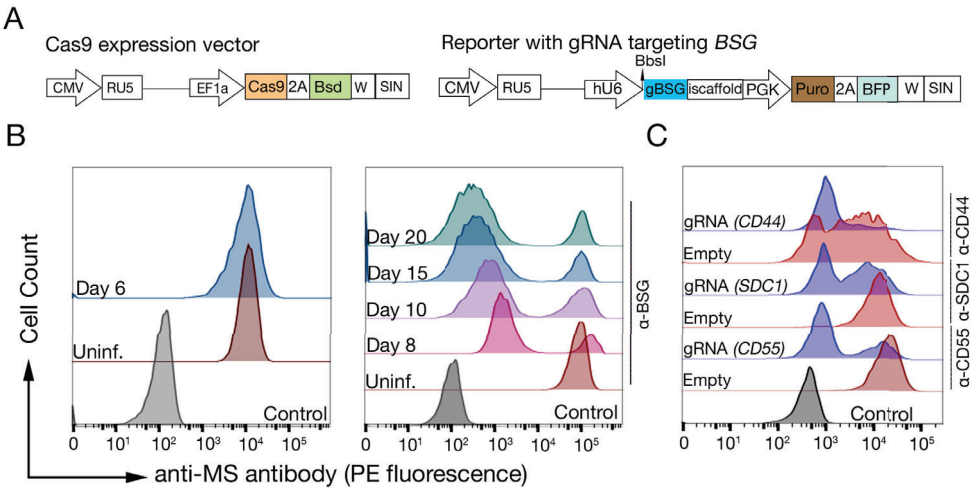

Fig. 3.1 Time- dependent decrease in cell surface expression of membrane receptors is observed in Cas9-expressing cells that are transduced with gene specific gRNAs. **A.** Schematics of lentiviral vector encoding Cas9 together with the blasticidin resistant gene at the C-terminus used to create stable Cas9 expressing lines (left panel) and Cas9 activity reporter with an improved scaffold and puro/BFP markers targeting *BSG* (right panel). **B.** Viruses with the activity reporter were used to quantify genome editing efficiency in Cas9-expressing HEK-293-E cell line. A small decrease in the surface expression of BSG was seen on transduced cells on day six post transduction. From day eight onwards, a very clear double population was observed. Approximately 25% of cells retained BSG expression on the cell surface by day eight and the frequency of this population did not change even at day 20. **C.** The surface expression of three other cell surface receptors (CD55, SDC1, and CD44) also decreased on day eight. Control: parental cell line stained with secondary anti-mouse(ms)-PE alone; Uninf: parental cell line stained with anti-BSG antibody, Empty: parental cell line stained with the indicated primary antibody. Both **B** and **C** are representative of two technical replicates. CMV: Cytomegalovirus promoter, RU5: 5' long terminal repeat lacking the U3 region, EF1a: intron-containing human elongation factor 1a promoter, Cas9: codon-optimised *Streptococcus pyogenes* Cas9, Bsd: Blasticidin resistant gene, W: Woodchuck Hepatitis Virus posttranscriptional regulatory element, SIN: ΔU3RU5 (self-inactivating 3' LTR); hU6: human U6 promoter, gBSG: gRNA targeting *BSG*, Iscaffold: Improved scaffold, 2A: self cleavage peptide, PGK: PGK1 promoter, puro: puromycin-resistant gene, BFP: blue fluorescent protein.

Cloning Cas9-expressing cell lines increases Cas9 cleavage efficiency

Having established the protocol to measure Cas9 efficiency, I proceeded to make stable Cas9 lines for a cell line originating from a different tissue source (colorectal carcinoma, NCI-SNU-1). The level of BSG expression on the surface of cells was analysed eight days post-infection by flow cytometry using an anti-BSG mAb. While the Cas9-mediated gene inactivation was evident from loss of BSG from the cell surface upon infection with gRNA targeting *BSG*, approximately 30% of cells that retained BSG expression were again observed

(figure 3.2B). One of the reasons for this incomplete knockout of *BSG* is likely to be the variability in expression of efficient Cas9-nuclease in a polyclonal cell population. In a polyclonal Cas9 line, it has been shown that some cells acquire mutations in the proviral Cas9-coding sequence with APOBEC3 mutational signature that can lead to Cas9 inactivation and decrease the overall efficiency [176]. Such inactivations within a population can potentially cause problems in genome-scale screens where Cas9 is assumed to be 100% efficient and the incorporation of a gRNA in a cell is equated to the complete loss of the targeted gene product. To reduce the heterogeneity and to select the cell with the highest Cas9 efficiency, single cells of the polyclonal line were sorted using a MoFlo-XDP sorter into 96-well tissue culture plates. Colonies appeared approximately two weeks later and as an initial test, 23 clones were tested. All except one clone had an increased population of cells that had lost surface BSG compared to the polyclonal line (five best clones depicted in figure 3.2B). The clone with the largest refractory population (clone 22) showed a broad surface expression profile with no clear population lacking cell surface BSG (figure 3.2C). In a polyclonal Cas9 line, cells with lower cleavage efficiency would decrease the overall Cas9 efficiency.

The loss of cell surface BSG expression as an indicator of Cas9 activity is useful for estimating efficiency for targeting endogenous genes but is time-consuming requiring steps involving antibody staining, and waiting for at least eight days to ensure complete protein turnover. A faster method (a GFP-BFP system) to check Cas9 efficiency has been developed using an exogenous system in which cells are transduced either with a construct expressing GFP with a gRNA targeting *GFP* or an empty gRNA as a control (vector schematics in 3.2 D) [176]. The expression of GFP can be monitored as early as three days post-transduction to determine the nuclease cleavage efficiency. To directly compare the *BSG* KO method with the *GFP* KO method, the five best clones of NCI-SNU-1 cells selected for the lowest refractory population of BSG surface expression were checked with the GFP-BFP system. All five clones showed a very high efficiency of *GFP* targeting compared to the polyclonal line (figure 3.2D). The fraction of cells from the polyclonal line that remain refractory to gene KO was lower using this system compared to the BSG KO system (6% compared to 30%). That said, the clone with the lowest refractory population under the *BSG* KO system (clone 4) also had the lowest refractory population with the GFP-BFP system.

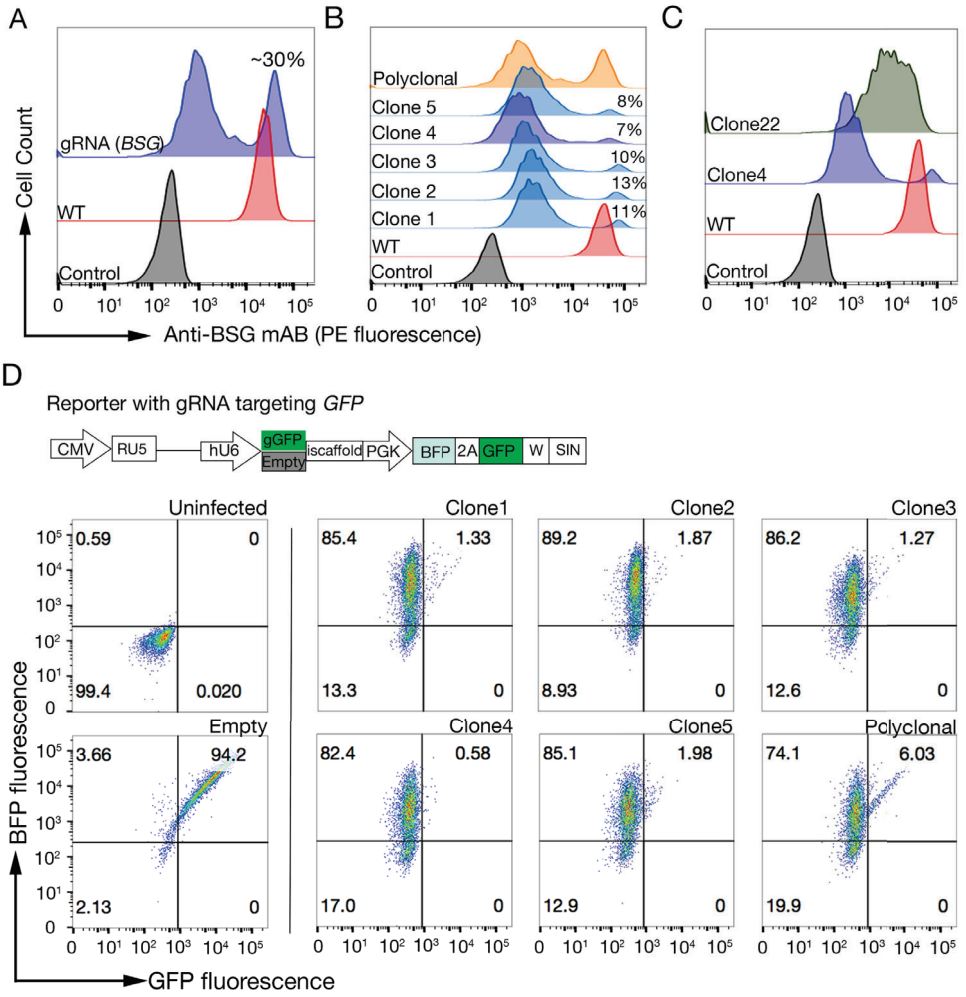

Fig. 3.2 Selecting clonal cell lines with high Cas9 activity for efficient genome-scale genetic screening. **A.** The Cas9 activity of polyclonal Cas9-expressing NCI-SNU-1 was tested by lentiviral transduction with a *BSG*-KO reporter: approximately 70% decrease in surface expression of BSG was observed. **B.** Clones of stable Cas9-NCI-SNU-1 showed variation in the fraction of the population that remained refractory to *BSG* inactivation. All five clones depicted a population that was approximately threefold lower compared to that from the polyclonal line. **C.** Comparison between the best and the worst clones from the same polyclonal line (23 clones tested). Clone 22 showed a small decrease in surface expression of BSG with no evident negative population as observed in clone 4. Control: parental NCI-SNU-1 cell line stained with secondary anti-mouse-(ms-) PE alone; WT: NCI-SNU-1 cell line stained with anti-BSG antibody. **D.** Viruses with a gRNA-targeting plasmid encoded GFP or without ('empty') (schematic depicted) were used to quantify genome editing efficiency of both polyclonal and cloned lines of Cas9-expressing NCI-SNU-1 cell line. The cells were tested for both BFP and GFP expression by flow cytometry after infection with lentivirus, or left uninfected. BFP expression marks transduced cells and the loss of GFP expression was used to quantify Cas9 activity. The profile for uninfected and 'empty' infected cells looked similar for all clones; representative profiles are depicted in the left panel. All five clones of the NCI-SNU-1 cell line show a higher loss of GFP compared to the polyclonal line (right panel), with clone 4 having the lowest refractory population.

Taken together, the GFP-BFP system used to assay Cas9 efficiecny correlated well with the *BSG* KO system, but a direct comparison between the assays suggested that the GFP-BFP system overestimates the Cas9 efficiency, whereas the *BSG* KO system provides a more realistic estimation of the efficiency of the targeting of endogenous genes. Taking this into consideration, the strategy I subsequently employed to create other stable Cas9-expressing lines was to initially use the GFP-BFP system to clone the high efficiency lines, and then to check the clone with the highest GFP cleavage efficiency with the *BSG* KO system to obtain a more accurate estimate of Cas9 efficiency. Five clones of HEK-293-E and HEL cells were each tested, and the clone with the lowest refractory population expressing GFP upon infection with gRNA targeting *GFP* was chosen to re-test with the *BSG* KO system (figure 3.3A). The chosen clones of HEL and HEK-293-E cell lines exhibited approximately 91% and 92% loss of BSG from the cell surface respectively, when transduced with a gRNA targeting *BSG* (figure 3.3B).

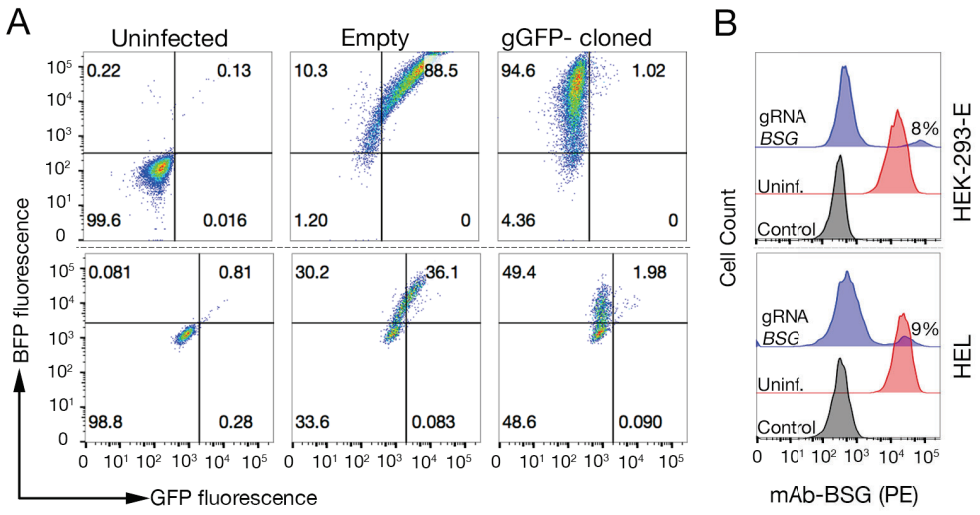

Fig. 3.3 Cas9-expressing human cell lines generated in this study demonstrate high *GFP* and *BSG* cleavage efficiency. A. Cloned Cas9-expressing HEK-293-E and HEL cell lines were tested for both BFP and GFP expression by flow cytometry after transduction with GFP-BFP reporter viruses, or alternatively left uninfected. BFP expression represents transduced cells (note the transduction rate for HEL cells was lower compared to the HEK-293-E cells). As before, the loss of GFP expression is used to quantify Cas9 activity; the line exhibiting the greatest loss of GFP expression out of at least five clones tested are shown. **B.** The cloned cell lines selected using the GFP-BFP assay were additionally tested for their ability to target an endogenous locus by infecting them with a lentivirus encoding a gRNA targeting *BSG* and quantifying expression of BSG on the cell surface with an anti-BSG mAb; both cell lines exhibited loss of BSG from >90% of the population. Control: parental cell line stained with secondary anti-ms-PE alone; Uninf: parental cell line stained with anti-BSG antibody.

3.2.2 Quality control of the genome-scale mutant cell library

Reproducible mutant cell libraries with a uniform coverage of the gRNAs can be created from plasmid gRNA library

I next set out to explore the optimal parameters to perform a pooled KO screen using the Cas9 cell lines that were generated. In the intended genome-scale screening setup, in which approximately 91,000 gRNAs are used, the library complexity plays a role in determining the success of a screen. The population of cells that are to be screened should be large enough to capture all gRNAs in the library. Usually, libraries that have high complexity (500-1000-fold representation of each guide) are used to screen for the desired phenotype. To obtain this representation, it is common to transduce 30-100 million cells at MOI of 0.3 to reduce the chance of more than one gRNA per cell [207]. The transduced cells make up the initial library. As a guiding principle, a library at least five fold larger than this library should be maintained while passaging the cells over the time that is required for the complete turnover of the proteins (at least eight days).

To ensure that gRNA representation is maintained using the cellular KO library preparation protocol, I first quantified the individual gRNA abundances in at least 5×10^7 cells on different days after transduction from two independent NCI-SNU-1-Cas9 libraries and one library of both HEK-293-E-Cas9 and HEL-Cas9 cells. The gRNA abundances in the libraries were quantified by deep sequencing and a high correlation between the biological replicates of NCI-SNU-1-Cas9 libraries and amongst the three different cell line libraries at equivalent time points was observed (figure 3.4A). This indicated that using this protocol, it was possible to reproducibly create mutant cell libraries with good representation of all gRNAs. Compared to the correlation between the different mutant cell libraries, the correlation between the original plasmid population and the mutant cell libraries was lower. This suggested that the most appropriate control for gRNA enrichment analysis was the cell line on that particular day rather than the population of gRNAs in the original plasmid population.

To further investigate the difference between the plasmid pool and the mutant cells in terms of their gRNA abundance, I next analysed the distribution of all gRNAs in each cell line and compared it to the distribution in the plasmid pool. In the plasmid library, while a small fraction of gRNAs were under- or over-represented, approximately 82% of gRNAs were uniformly distributed

with only 8-fold increase in abundance between the 10^{th} and 90^{th} percentiles. The mutant libraries generated from this plasmid library also showed a uniform coverage, but a small drop in the overall representation of the gRNA library was observed in all cell lines; this decrease was more evident for libraries on day 16 compared to those on day 9. (figure 3.4B).

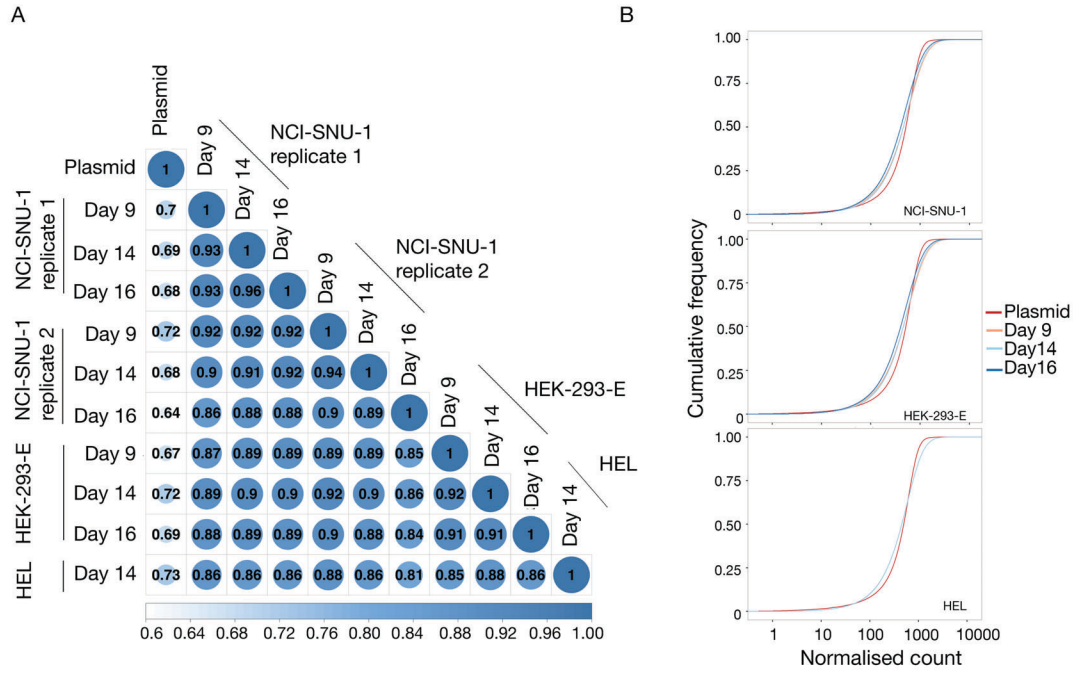

Fig. 3.4 Mutant cell libraries can be created reproducibly by transduction with the lentiviral gRNA library. A. Matrix plot depicting the correlation between normalised read counts of all gRNAs in the original plasmid library and gRNAs in the indicated cellular mutant libraries on the indicated days. Correlation between biological replicates of NCI-SNU-1 and between any cell line on any day was higher than when compared to the original plasmid pool. **B.** Cumulative distribution function plots comparing the gRNA abundance in the plasmid library to the mutant libraries of HEK-293-E and NCI-SNU-1 cells on day 9, 14 and 16 days post transduction. The differences in the curves from the mutant cell populations represent the depletion in a subset of gRNAs compared to the plasmid library.

Negative selection screens reveal essential genes

In a negative selection screen, in which the mutant cells are continually grown for an extended period of time, the cells that carry mutations in the genes that are essential for cell proliferation will deplete, which makes the corresponding gRNA to 'drop-out'. Therefore, when comparing the mutant gRNA abundance to the plasmid gRNA abundance, gRNAs that target fundamental cellular processes should, in principle, be depleted. These screens are sensitive to the library representation as losing the representation of gRNAs due to technical

reasons- such as the need to passage the library because of cell growth, which may create population restriction points, reducing gRNA library representation. Thus, investigating the depleted genes under negative selection can provide valuable insights into the quality of the mutant library.

To check whether the depleted gRNAs targeted essential genes, I next carried out a gene-level negative selection enrichment analysis to identify genes that were depleted in the mutant library compared to the plasmid library. As a quality control, I first analysed ribosomal genes (annotations from KEGG-Ribosome), which are known to be essential and are often identified robustly in similar negative selection screens [175, 176, 177]. Reassuringly, the majority of the ribosomal genes were amongst the most significantly depleted genes (False discovery rate (FDR) <10%) in all three days and in both HEK-293-E and NCI-SNU-1 cell lines (figure 3.5A). Next, I carried out pathway analysis using KEGG annotated pathways; among the most depleted pathways were essential biological processes such as the spliceosome, cell cycle, proteasome and DNA replication processes. The number of relevant genes ('hits') in these pathways was similar between the biological replicates and also between cell lines (figure 3.5B). These results provided further confidence that the cellular mutant libraries generated using the protocol retained their gRNA complexity and could be used for genome-scale screening.

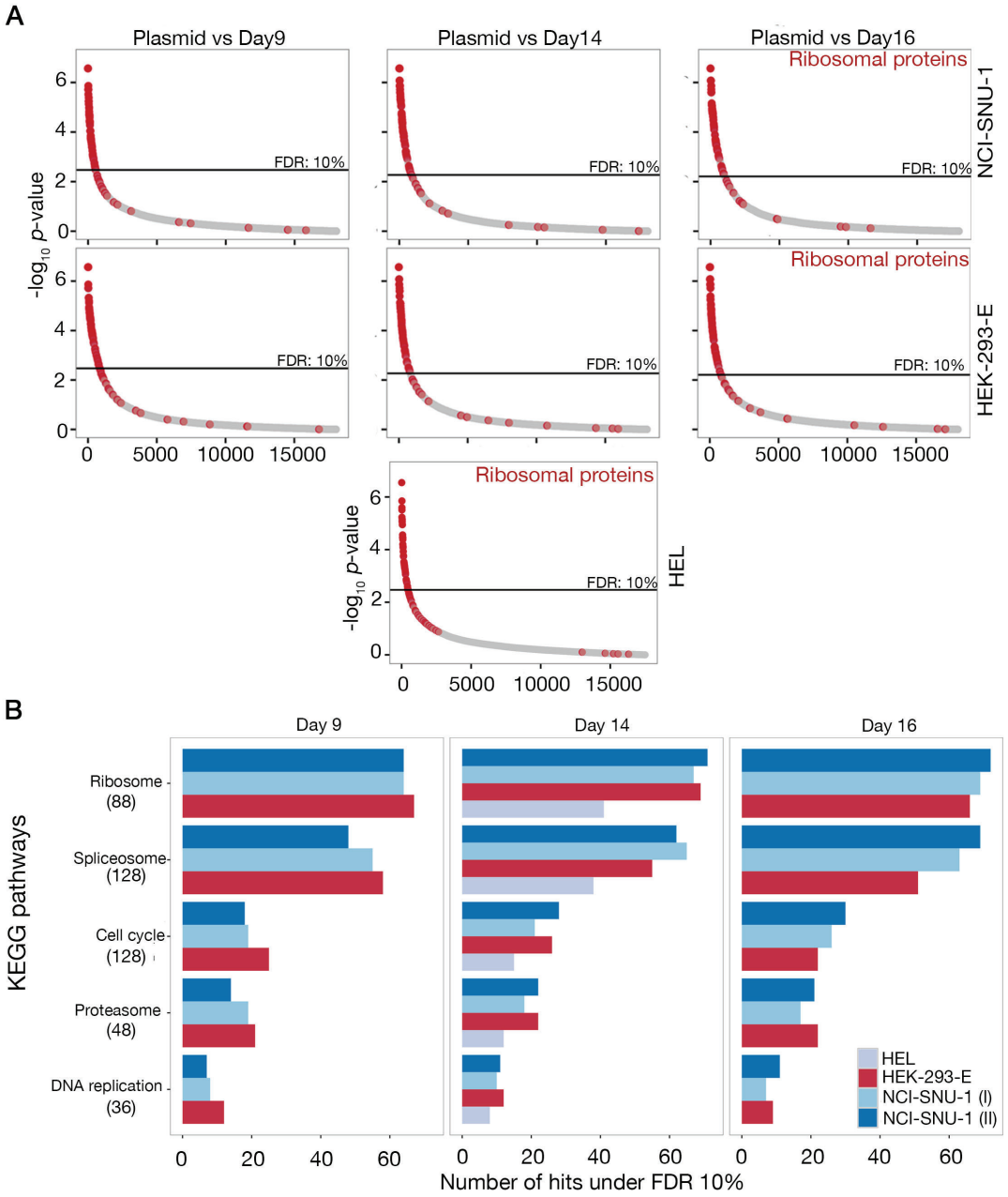

Fig. 3.5 gRNAs targeting genes involved in essential biological processes were depleted during culture of the mutant cell library. A. Gene-level enrichment analysis revealed genes involved in ribosomal biosynthesis to be among the most significantly depleted genes in NCI-SNU-1, HEL and HEK-293-E cell lines on all three time-points, days 9,14 and 16. For NCI-SNU-1, representative analysis from one of the two experiments is shown. **B.** Five KEGG-annotated pathways known to be involved in essential cell processes that were significantly depleted in all the samples are shown. The numbers under each pathway represents the total number of genes in the group. The number of genes involved in each pathway identified as being significant was similar for the replicates of NCI-SNU-1 and between the HEK-293-E, HEL and NCI-SNU-1 cell lines.

3.2.3 Genome-scale screens using monoclonal antibodies

Design of a genome-scale KO screening platform to investigate cellular recognition events at the cell surface

To identify factors required for molecular recognition events at the cell surface in the context of a plasma membrane, I initially designed a genome-scale KO screening system in human cell lines. A flow-cytometry based binding assay was initially used to identify a high-efficiency Cas9 expressing cell line that stained brightly with a mAb. I then created a genome-scale mutant cell library for the chosen cell line using a library of lentiviruses, each encoding a single gRNA from a pool of 90,709 individual gRNAs (gRNA expression vector depicted in figure 3.6A). Transduced cells that had lost the antibody epitope at the cell surface were isolated through fluorescence-activated cell sorting (FACS) and the genes responsible for this loss of binding were identified by comparing the relative abundance of the different gene-specific gRNAs present in the sorted cells compared to the total unsorted population using deep sequencing of PCR products and enrichment analysis (schematics depicted in figure 3.6).

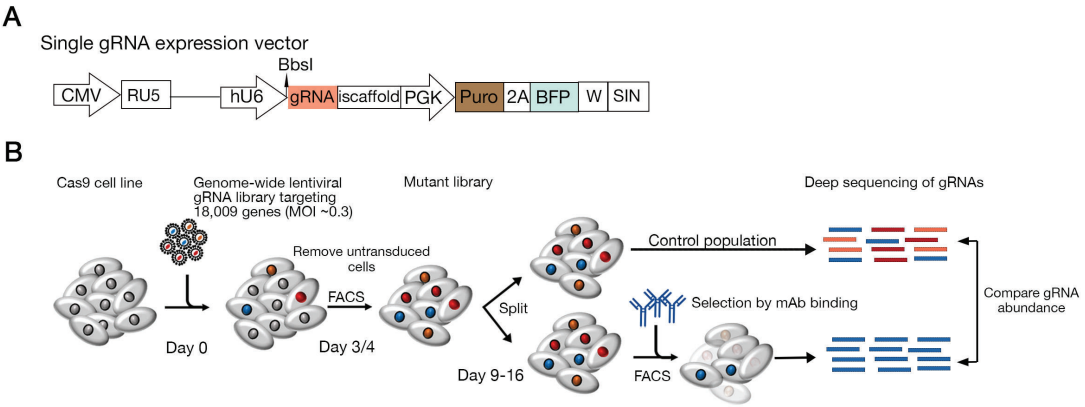

Fig. 3.6 Schematic of the genetic screening approach. A. Schematic of lentiviral single gRNA expression vector with BFP and puromycin selection cassettes with gRNAs cloned into the BbsI site. The backbone is the same as the reporter vector with gRNA targeting *BSG* described earlier (figure 3.1A). **B.** Strategy used to perform genome-scale KO screens using mAbs as selection reagents. The initial mutant library is generated by infecting cells at a low MOI of 0.3 and the infected cells expressing BFP are selected by FACS. The sorted library is kept in culture between 9 and 16 days. The total cell population on the screen day is divided into two; one half is kept as control population and the other half is selected for mAb binding. The cells from the mutant library that are refractory to binding are sorted using FACS and the abundance of gRNAs in the sorted population is compared to the control population to identify genes contributing to binding of the mAb to the cells.

A proof-of-principle screen using anti-BSG antibody identifies BSG and a chaperone required for BSG trafficking to the plasma membrane.

To test the use of the genome-scale screening approach, I carried out a proof-of-principle screen on HEK-293-E cells with an anti-BSG mAb as the selection reagent. BSG was a good model to test this system as the antibody provided a clear bright signal on the Cas9-expressing HEK-293-E cell line with a very high signal to noise ratio (figure 3.7A). The mutant library was screened 16 days post transduction and 0.22% of the population that expressed gRNA (BFP+) and lacked BSG on the surface (PE-) was sorted (figure 3.7B). Unsorted cells also from day 16 post-infection were used as a controls. gRNAs from both sorted and control populations were isolated and deep sequenced to quantify their abundance.

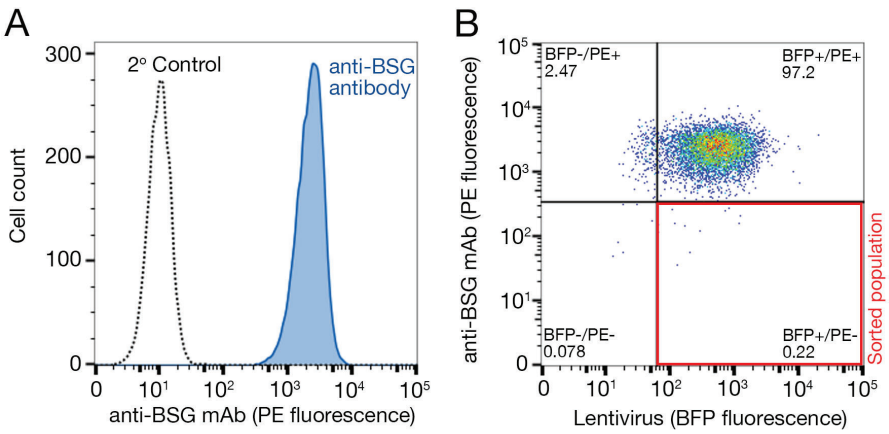

Fig. 3.7 Cell sorting strategy for a proof-of-principle genome-scale screening for recognition of BSG in HEK-293-E cells by an anti-BSG mAb A. BSG was highly expressed in HEK-293-E cells as indicated by the clear separation of cells stained with anti-BSG antibody versus secondary-only stained samples. **B.** Gating strategy for screen carried out with anti-BSG mAb in HEK-293-E cells. A very small fraction of PE-negative and BFP-positive cells were sorted. Approximately 60 million cells were sampled and 40,000 cells were collected during this screen.

A preliminary analysis using the raw abundance of gRNAs in the sorted population revealed a high enrichment of gRNAs targeting *BSG*, as expected, but also gRNAs targeting a gene encoding a monocarboxylate transporter, *SLC16A1* (or *MCT1*). Within the top ten most abundant gRNAs, four targeted *BSG* and four targeted *SLC16A1*—which is a known chaperone required for maturation and surface expression of BSG (3.8A). BSG is an unusual protein for a single transmembrane protein as it contains a charged glutamic acid residue within the transmembrane region. It has been suggested that this charged residue is important for the lateral association of BSG with the multi-

pass protein MCT1 on the plasma membrane (Schematics in figure 3.8B) [208, 209].

I next carried out a enrichment analysis using the MAGeCK software and identified four genes (*BSG, SLC16A1* and two novel factors *SPPL3* and *WDR48*) that were significantly enriched (FDR <0.05) in the sorted population compared to the control population (figure 3.8C). SPPL3 is a Golgi-resident intramembrane-cleaving protease that has been shown to regulate the activity of N-glycosylation related glycosyltransferases such as MGAT5, B3GNT5 and B4GALT1 [210]. The three N-glycosylation sites (N44Q, N152Q, and N186Q) of BSG are known to contain complex-type carbohydrate groups generated in the Golgi by the action of multiple glycosyltransferases, including MGAT5, thus it is possible that the loss of SPPL3 causes the generation of abnormal carbohydrate chains in BSG, which could lead to disregulated surface expression [211]. WDR48 is a regulator of deubiquitylating complexes, and, in this case also, there has not been a reported direct association of this protein with BSG. However, it has been suggested that BSG recruitment to the surface of the cell membrane is induced upon K63-linked ubiquitylation [212]. Therefore, it is possible that WDR48 has an effect on the pathway leading to BSG ubiquitylation thereby affecting its surface localisation. Further validation experiments are required to investigate these hypothesised roles.

This proof-of-principle study demonstrated that genome-scale KO screening approach was a suitable method to identify not only the gene encoding the antibody epitope, but also reveal factors such as chaperones that are important in expression of the receptors. However, one limitation of the method was that very few genes were identified as being significantly enriched. This was surprising because I was expecting that genes encoding proteins within the secretory pathway would be enriched in the sorted non-binding population as they would affect the surface expression of BSG. While a few members of the general secretory pathway are known to be essential, and by the day the screen was carried out, cells with mutations in those genes would have dropped-out of the mutant pool, it was surprising that none were identified. A likely explanation for the low number of significant genes could be the stringent gating threshold and the low number of cells ($\sim$40,000) that were collected, resulting in lower number of gRNA being represented in the sorted population. This could be observed by comparing the gRNA abundance in the control and sorted sample, which revealed a clear loss of many gRNAs in the sorted sample (figure 3.8D). Upon a closer look, only approximately 16,300 gRNAs

had representation (count above 10) in the sorted population compared to the 86,000 in the control population. Within the represented gRNAs, the abundance of gRNAs targeting the four significant hits were clearly higher in the sorted population compared to the control population. This suggested that using a highly stringent sorting threshold would result in the identification of few genes that have strong effects on binding loss. However, such a strategy might not be ideal for the identification of genes that have comparatively weaker size effects.

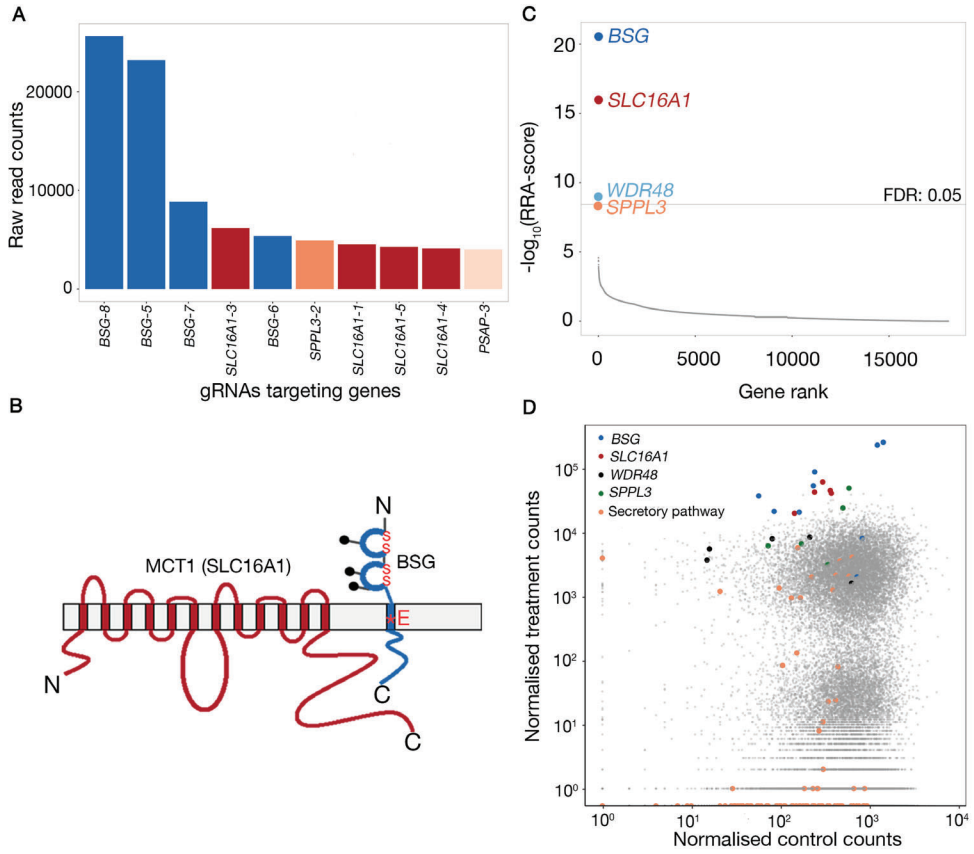

Fig. 3.8 A positive selection screen using anti-BSG mAb demonstrates the successful application of a CRISPR-based loss-of-function screen to identify factors required for epitope recognition by a mAb A. Top ten gRNAs with the highest raw read counts in the sorted non-mAb binding population. Four out of nine gRNAs targeting *BSG* and four out of five gRNAs targeting *SLC16A1* were represented in the top ten list. **B.** Schematic representation of cell surface interaction between 12-transmembrane transporter protein MCT1 and the single pass type I protein BSG. **C.** Enrichment analysis using MAGeCK. The y-axis represents the Robust Rank Aggregation (RRA) scores of genes calculated by comparing the gRNA abundance in sorted vs control cells. Genes are ordered by their RRA-score. **D.** Comparison between normalised read counts in control versus treatment (sorted) population. Individual gRNAs targeting the four significantly enriched genes but no gRNAs targeting the secretory pathway genes (KEGG-protein export) were enriched in the sorted population.

An improved screen identifies cellular pathways required for receptor expression in addition to genes directly encoding for the mAb epitope

I next explored ways to improve the screen to increase its sensitivity and identify not only the direct receptor, but also the pathways required for the expression of the receptor on the cell membrane. To test if the stringency of sorting was the reason for the low number of significant genes being identified in the screen, I aimed to decrease the stringency and to increase the mutant library size, in order to increase the number of sorted cells, thereby increasing the representation of the gRNAs. I carried out the screen using these parameters with an anti-CD59 mAb antibody. CD59 was chosen to test this system for various reasons. First, the anti-CD59 provided a bright stain on Cas9 expressing HEK-293-E cells (figure 3.9A); second, as CD59 is a GPI-anchored protein, its expression on the plasma membrane depends on the components of the cellular GPI anchor biosynthesis pathway [213]. The GPI anchor pathway is known to be non-essential for cell lines and has been robustly identified in genome-scale loss-of-function screens [173]. For this experiment, I screened 100 million cells from the mutant library and sorted 0.45% of the non-binding and BFP expressing cells (figure 3.9B). For increased representation of the library in the sorted population, I aimed to collect up to ten times more cells in the non-binding population compared to the screen with anti-BSG mAb.

When comparing the overall abundance of gRNAs in the sorted population to that from the control population, I immediately observed a strong enrichment for individual gRNAs targeting *CD59*, the GPI anchor biosynthesis pathway and the secretory pathway in the sorted population (figure 3.10A). Next, I applied gene-level enrichment analysis using the MAGeCK software and observed that *CD59* itself and 25 genes directly related or contributing to the GPI anchor biosynthesis pathway were significantly enriched (FDR <0.05) in the sorted population (figure 3.10B, GPI anchor biosynthesis schematic depicted in figure 3.10C).

Upon a closer look at the secretory pathway genes, only one member of the protein export pathway, *SLC61A1*, was identified as being significantly enriched in the sorted population. All other genes corresponded to the initial stages of the N-linked glycosylation pathway or the 'core glycosylation' pathway (identified genes in figure 3.11). The early steps in N-linked glycosylation involves the generation of a unique 14- monosaccaride ($Glc_3Man_9GlcNAc_2$) structure, which is a conserved feature of eukaryotic cells and is used as

a signalling molecule in the protein folding pathway for a wide range of N-glycosylated proteins such as CD59[1] [215]. Thus, the genes in this pathway are likely to have contributed to the proper folding of CD59 ensuring its transport to the surface of the cells.

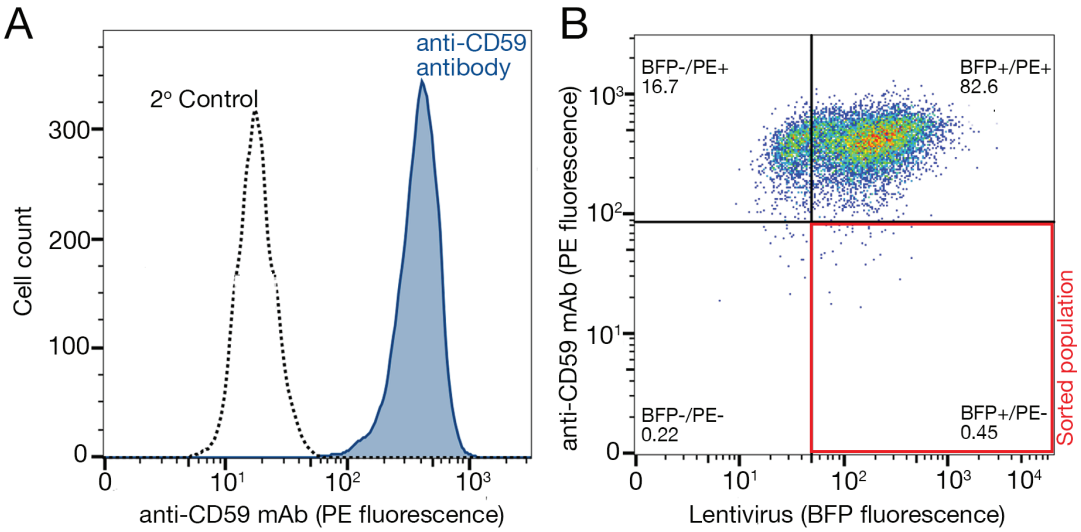

Fig. 3.9 Cell sorting profile for flow-cytometry based CRISPR-Cas9 screen in HEK-293-E cells with an anti-CD59 antibody. A. CD59 was highly expressed on the surface of HEK-293-E cells as indicated by the clear separation of cells stained with antibody compared to the secondary only stained sample. **B**. Gating strategy used in a genome-scale screen carried out using an anti-CD59 mAb. Approximately 100 million cells were sampled and 350,000 cells were collected during this screen. The mutant cell library has a clear BFP negative population as the library was made with only BFP sort and no puromycin was added to remove the contaminating non-transduced cells. The screen was carried out 15 days post infection with the gRNA lentiviral library.

The data here demonstrated that sampling enough cells from a high complexity library increases the sensitivity and allows for the identification of general and protein specific pathways required for transport of receptor to the surface in addition to the receptor itself. Increasing the sorting threshold did not influence the confidence with which the target receptor was identified as *CD59* was the still the most enriched gene in the sorted population.

[1] CD59 has two N-linked glycosylation sites and it mainly consists of a family of biantennary complex-type structures with and without lactosamine extensions and outer arm fucose residues. It has been shown that CD59 transport to the surface of the cells does not rely on this glycan being present on the protein [214]. Consistent with this, no genes in the later processing steps, which make these complex type structures were identified in the screen.

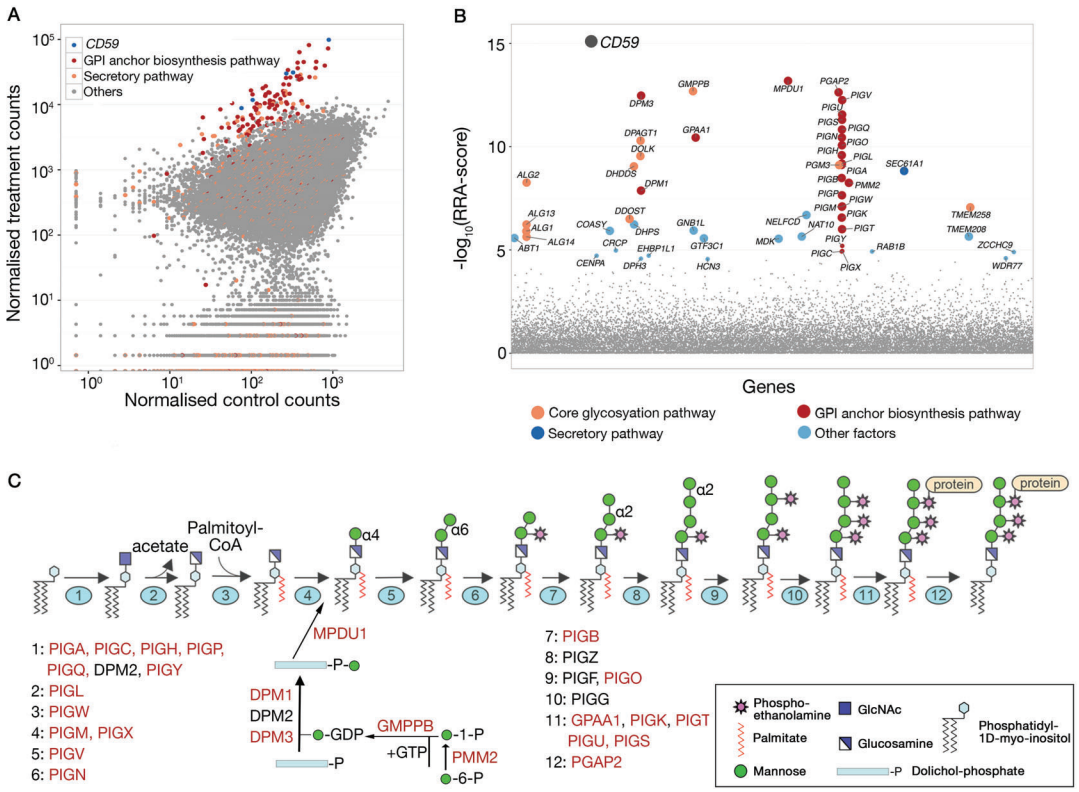

Fig. 3.10 An Improved CRISPR-mediated forward genetic screen identifies the genes required for the trafficking of the receptor in addition to the gene encoding the antibody epitope. A. Comparison of gRNAs abundance in sorted samples versus the control samples. Multiple gRNAs targeting the GPI-anchor pathway and *CD59* were clearly enriched in the sorted population. **B.** The enrichment of gRNAs targeting each gene (ordered alphabetically) is quantified as the RRA-score. Each circle represents a specific gene and the size corresponds to the FDR (large circle: FDR<1%, small circle: 1% < FDR < 5%). Only genes with FDR < 5% are labelled and the colors represent genes related by function. For better clarity, enlarged version of **B.** is also depicted in Appendix section figure A.1. **C.** Schematics of cellular GPI-anchor biosynthesis pathway (adapted from [216]). The genes that were significantly enriched are highlighted in red. Key enzymes along the pathway are labelled. PMM2 catalyses the isomerisation of mannose 6-phosphate to mannose 1-phosphate, which is a precursor to GDP-mannose necessary for the synthesis of dolichol-P-mannose by the members of dolichol-phosphate mannose (DPM) synthase complex (DPM1, DPM2 and DPM3). MPDU1 is required for the utilisation of the mannose donor dolichol-P-mannose by mannosyltransferase PIGM. Figure modified from [216].

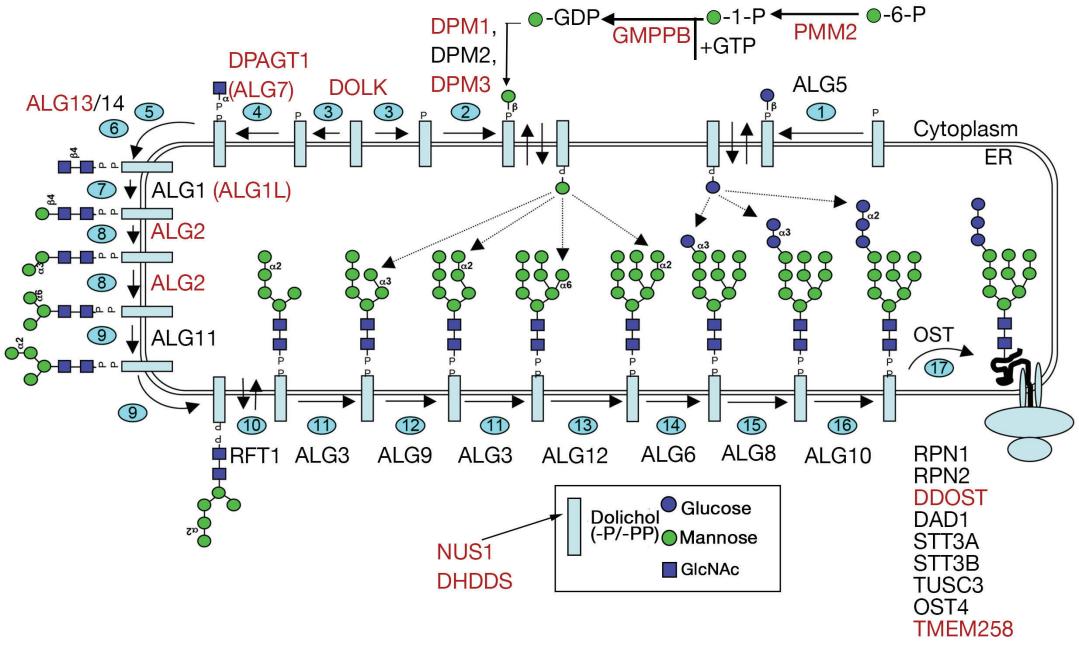

Fig. 3.11 Several genes involved in the early stages of N-glycan biosynthesis pathway were identified in the screen using an anti-CD59 mAb. Pathway depicting the core glycosylation that occurs in the ER to generate a 14-sugar-glycan precursor (dolichol-P-P-$GlcNAc_2Man_9Glc_3$), which is transferred *'en-bloc'* to the N-glycosylation consensus sequence (Asn-X-Ser/Thr) by the oligosaccharyltransferase (OST). The first committed step of the core glycan assembly is catalysed by a well-conserved essential gene DPAGT1 (or ALG7) [217]. The enzymes that catalyze each step in the biosynthesis have been identified mainly from studies carried out in mutants of the yeast *Saccharomyces cerevisiae*. The gene affected by each yeast mutation is known as an ALG gene (for altered in glycosylation). NUS1 together with DHDDS, forms the dehydrodolichyl diphosphate synthase (DDS) complex, an essential component of the dolichol monophosphate (Dol-P) biosynthetic machinery. Inhibition of protein N-glycosylation has been shown to cause cellular stresses that disrupt the ER leading to the induction of an unfolded protein response (UPR) and ER-assisted degradation (ERAD) [218]. Figure modified from [216]

Application of the screening method on a panel of antibodies reveals cellular factors required for receptor expression.

To determine whether the high sensitivity of the genome-scale method could be exploited to understand the other components of receptor biology, I next selected six mAbs targeting structurally diverse cell surface proteins expressed on either HEK-293-E or HEL cells and applied the method (summarised in table 3.1). All screens were carried out with libraries created from independent lentiviral transductions and for all screens, approximately 0.8-1% cells (250,000-400,000 cells) that had lost the ability to bind to the mAbs were collected.

Enrichment analysis using MAGeCK was carried out by comparing the gRNAs in the sorted population and the control population.

Table 3.1 Summary of mAbs used for genome-scale loss-of-function screening.

Cell line/Screen day	Antibody	Target receptor	Receptor feature
	BRIC125	CD47	
HEK-293-E cells (9 days p.i)	BRIC126	CD47	5-TM membrane protein
	B6H12	CD47	
	BRIC5	CD58	Single-Pass type I
HEK-293-E cells (10 days p.i)	P16	GP1a/IIa	(αIIβ1) Integrin heterodimer
HEL- (14 days p.i)	BRIC216	GYPA	O-glycosylated protein

Genome-scale screens carried out using mAbs identify SRP dependent ER translocation pathway. Three screens, using independently created knockout libraries, were performed with three different monoclonal antibodies targeting CD47 (BRIC125, BRIC126 and B6H12) on day nine post mutant library generation using lentiviral transduction. These screens were designed to (i) gain insights into the repeatability and robustness to biological variation in using the KO screening approach, and (ii) to assess the effect of screen day on genes that are identified in a genome-scale screen.

Enrichment analysis of the sorted population in all three screens revealed *CD47* among the most significantly enriched genes (FDR <0.01 in all cases), (figure 3.12A, B and C), demonstrating the robustness of the genetic screening method using the CRISPR-Cas9 system in identifying the gene encoding the direct receptor. In all three cases, multiple members of the general protein export pathway were also identified within FDR <0.05 (summarised in table 3.2). Excluding *CD47*, 16 genes were identified in at least two of three screens and nine of these common genes encoded the proteins involved in the general protein export pathway, specifically the components of the ER translocon protein complex (*SEC61A1, SEC61G*), the signal recognition particle (SRP) proteins (*SRP19, SRP14, SRP54, SRP68, SRPR*), and the members of the signal peptidase complex (*SPCS2* and *SPCS3*) (figure 3.12 D).

The general protein export pathway is the cellular machinery by which the majority of newly synthesised membrane and secreted proteins are folded and transported, thus in this context, their identification is to be expected. That said, it was interesting that these genes were mainly identified repeatedly in this set of screens, and not in the one with anti-CD59 antibody (where only *SLC61A1* was identified). A likely explanation for this is the day the screens

were carried out— day 9 here, versus day 15 post transduction in case of anti-CD59 mAb screen. Many identified genes along the general protein export pathway, including the ones identified here, such as *SLC61A1, SRPR, SPCS3, SRP54*, have been described as 'core fitness genes' that are essential for cellular proliferation in every cell [175]. Thus, the gRNAs targeting these genes are very likely to 'drop-out' as the cells are continually grown. Hence, the chance of identifying them could be time-dependent, as the longer the mutant library is kept, the higher the chance that the cells with mutations in essential genes will be non-viable and drop-out of the mutant pool, causing the signal to dilute.

Table 3.2 Genes involved in protein export (KEGG annotation) identified in screens carried out using anti-CD47 mAb on day nine post mutant library generation

Description	Protein	Number of genes(s) identified using mAb
Translocation channel constituents	SEC61A1	BRIC125, BRIC126, B6H12
	SEC61G	BRIC125, BRIC126
Components of SRP	SRP14	BRIC125, BRIC126, B6H12
	SRP19	BRIC125, BRIC126, B6H12
	SRP54	BRIC125, BRIC126, B6H12
	SRP68	BRIC125, BRIC126
	SRP72	B6H12
SRP-ribosome complex receptor	SRPR	BRIC126, B6H12
	SRPRB	BRIC126
Signal peptidase complex	SPCS2	BRIC125, BRIC126
	SPCS3	BRIC125, BRIC126

The other gene that was identified in all three screens was *ASCC3*, which has been recently identified to regulate gene expression after UV-induced DNA damage [219]. The role of ASCC3 in relation to CD47 expression on the cell surface is not known, but this gene has also been identified in CRISPR mediated KO screen carried out for host factors required by DENV virus [220], so it is possible that it plays a regulatory role for general protein expression rather than a CD47-specific role. Further studies would have to be done to confirm this.

The data presented here demonstrate that this method is reproducible and can tolerate biological variation especially for the identification of the direct target of the mAb. In addition, it can also repeatedly reveal both essential and non-essential cellular factors that contribute to the expression of the target receptor on the surface of the cell.

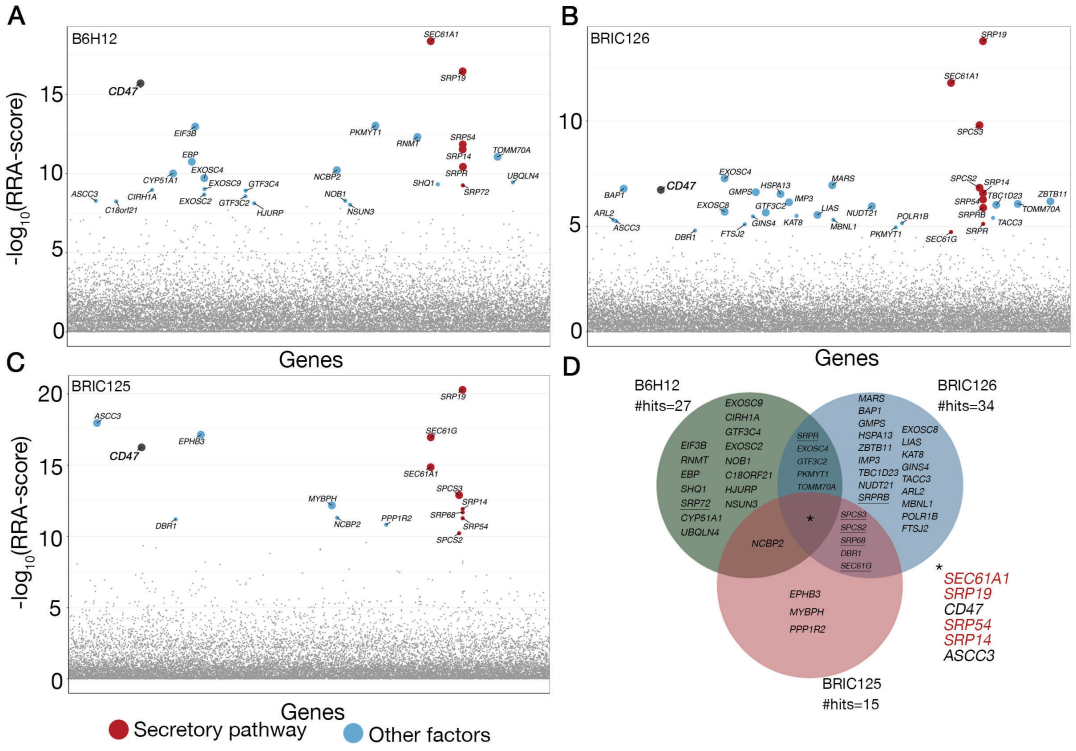

Fig. 3.12 Genome-scale CRISPR-Cas9 screen using anti-CD47 mAbs targeting cell surface protein reveal members of SRP dependent ER protein translocation pathway. Gene-level RRA-score calculated using MAGeCK in a genome-scale KO screens using anti-CD47 mAbs: B6H12 mAb (**A**), BRIC 125 (**B**) and BRIC126 (**C**). Each circle represents a specific gene and the size corresponds to the FDR (large circle: FDR<1%, small circle: 1% < FDR < 5%). Only genes with FDR < 5% are labeled and the colors represent genes related by function. In all cases, CD47 was identified amongst the most enriched genes, as expected; however, in the case of mAb clone BRIC125, an additional gene encoding for a different cell surface protein, EPHB3, was also identified with same the FDR as CD47. EPHB3 was not enriched for two other anti-CD47 mAbs, suggesting the BRIC125 epitope was present on both CD47 and EPHB3. **D.** Venn-diagram representing the genes identified with FDR < 5 % in the three screens. Six genes were common to all screens. These included genes of the SRP-dependent protein export pathway, *CD47* and *ASCC3*. All three screens were performed nine days post infection with the gRNA lentiviral library. For better clarity, an enlarged version of this figure is also depicted in Appendix section figure A.2.

Genome-scale screen using an anti-GYPA mAb reveals post-translational modifications present on GYPA required for the mAb epitope. Genome-scale screening methods provide a unique potential to study the enzymes required for cell surface post-translational modifications such as glycans. With the method described here, it should, in principle, be possible to identify genetic pathways leading to the biosynthesis and the subsequent positioning of the receptor glycans on proteins or lipids, thereby facilitating their identification.

The human erythrocyte protein, GYPA, serves as a good model to study post translational glycan modifications as it is one of the most heavily glycosylated proteins with 15 O-linked glycans, which consists predominantly of disiaylotetrasaccharide linked to a serine or threonine residue and one N-linked glycan [221]. To investigate this further, an anti-GYPA antibody (BRIC256), which recognises the extracellular epitope amino acids 41-58 on GYPA where three O-linked glycans are present (position 44, 47 and 50) [222] was used as a probe to carry out a genetic screen. The expression of GYPA is restricted to the cells of erythroid lineage, thus in this case, the screen could not be carried out in HEK-293-E cells. For this purpose, I selected the HEL cell line as these cells resemble erythroblasts (nucleated immature erythrocyte) and are known to express GYPA [223].

The gene-level enrichment analysis of the gRNAs in the cells that had lost binding the anti-GYPA antibody revealed *GYPA* as one of the most-enriched genes, as expected (figure 3.13A). In addition, genes required for the generation of the core-O-glycan structure (*C1GALT1, C1GALT1C1*) and genes involved in the addition of terminal N-acetyl neuraminic acid (Neu5Ac) modifications (*CMAS, SLC35C1*) were also identified. Schematics of the predominant O-glycan found on GYPA and the roles of the identified enzymes in generating this structure are depicted in Figure 3.13B. This suggests that the loss of GYPA itself or the loss of enzymes required for the generation of the disialotetrasaccaride present on GYPA leads to the loss of the mAb epitope. This could either happen if GYPA expression on the surface of the cells depended on the presence of O-linked glycans on the protein or if the O-linked glycans together with the protein backbone form the antibody binding epitope. As BRIC256 recognises the region of the protein where O-linked glycosylation sites are present, the latter seems more likely in this case. These results demonstrate that in situations where the interaction is mediated by glycans, the screen is able to identify the cellular pathways that are required for the biosynthesis of the glycan, thereby providing clues to the identify of the glycan.

The screen additionally identified an erythroid-specific transcription factor GATA1, which is presumably required for GYPA transcription in HEL cells. In addition, two genes, *CDIPT* and *SACM1L* were also identified. Both of these genes are required for metabolism of phosphatidylinositol (PtdIns). CDIPT protein is required for biosynthesis of phosphatidylinositol whereas SACM1 is a phosphoinositide lipid phosphatase. GYPA has been shown to be associated with PtdIns(4,5)P_2 in the cytosol [224], which suggests that the loss of phos-

phoinositides would adversely affect its membrane positioning. Additionally, the identification of *TNNT3*, a gene that encodes for the tropomyosin-binding subunit of troponin, was intriguing as tropomyosin has been previously shown to associate with the major proteins of the erythrocyte membrane skeleton (spectrin, actin, and 4.1R (human erythrocyte protein 4.1)) to form a membrane complex that includes glycophorins [225]. Further studies will have to be carried out to verify the role of TNNT3 in regulating GYPA expression.

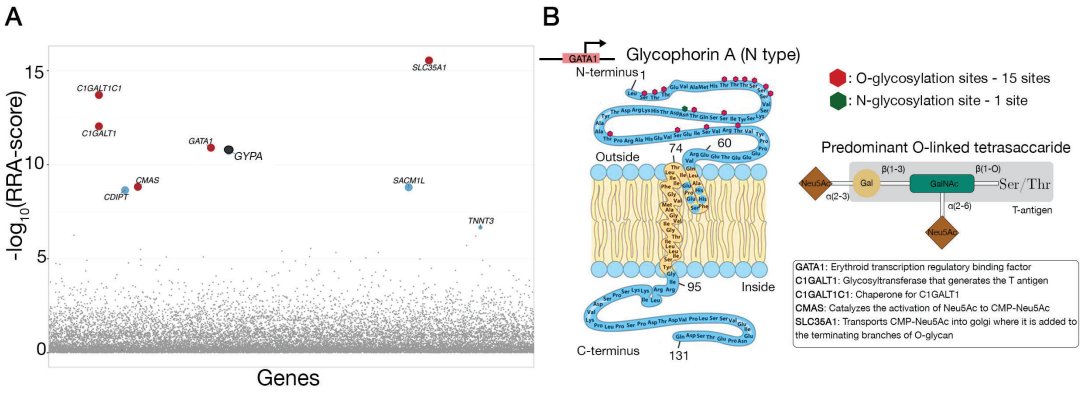

Fig. 3.13 Genome-scale KO screen using an anti-GYPA antibody reveals mAb epitope and factors required for the cell surface GYPA display. **A.** Enrichment analysis using MAGeCK revealed nine genes that were enriched in the sorted population compared to the control population (FDR< 5%). The number of cells sampled in this screen was lower compared to the previous screens carried out with HEK-293-E cells (50 million vs 100 million). The screen was performed on day 14 post infection with the gRNA lentiviral library. Genes relating to the post-translational modification present on GYPA and the gene encoding for erythroid specific transcription GATA1 are labelled in red. **B.** GYPA on the cell surface exits in either the M or N form with the M phenotype characterised by Ser-1, Gly-5 and the N-phenotype by Leu-1 and Gly-5. Schematics of the N-type GYPA is depicted in the left panel and the detailed structure of the predominant O-linked tetrasaccaride present in the protein is depicted in the upper right panel. The function of the GYPA specific genes identified in (A) are detailed in the lower right panel.

A Genome-scale screen using an anti $\alpha 2\beta 1$ mAb identifies the subunit encoding the antibody epitope and the critical requirement of actin regulation. A genome-scale screen was carried out in HEK-293-E cells using a monoclonal antibody (P16) targeting integrin $\alpha 2\beta 1$ (ITGA2/ITGB1) ten days post mutant library generation. Enrichment analysis revealed the gene encoding only the $\beta 1$ chain (*ITGB1*) of the integrin heterodimer in the sorted population demonstrating the epitope of this mAb is located within the $\beta 1$ and not $\alpha 2$ chain. In addition, genes of the general secretory pathway (SEC translocon complex and SRP components) were also significantly enriched in the sorted population and similar to the screen with anti-CD59 mAb, genes

required for early steps of N-linked glycosylation in the ER were also identified (figure 3.14A).

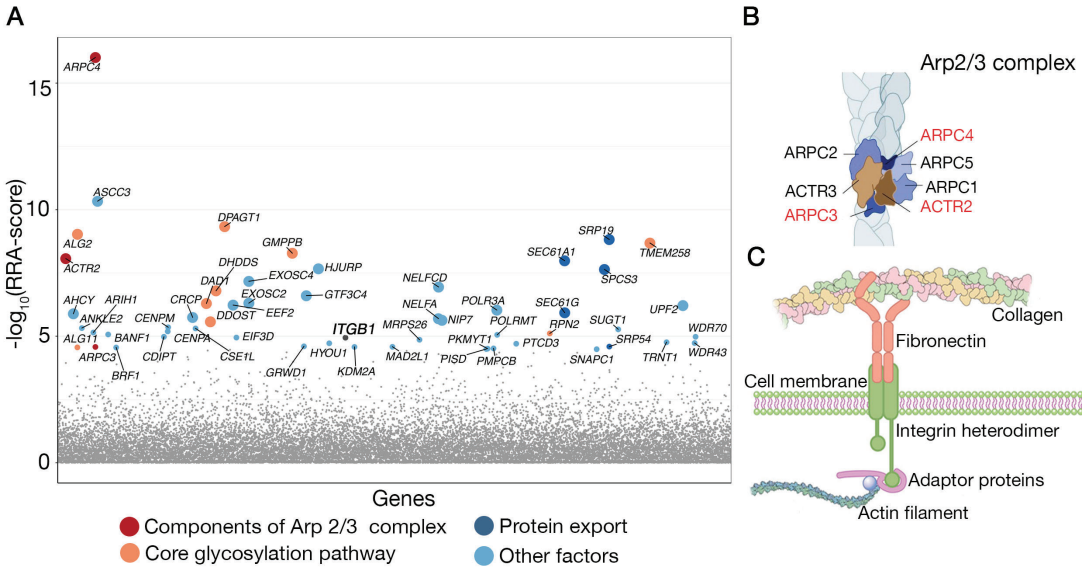

Fig. 3.14 Genome-scale loss-of-function screen using an anti-integrin αIIbβ1 mAb identifies the subunit encoding the antibody epitope and components of the cytoplasmic Arp2/3 complex. **A.** Gene-level enrichment analysis on a screen carried out using an anti-αIIbβ1 antibody. Only the β1-subunit encoding gene, *ITGB1*, was identified to be significantly enriched in the sorted population (FDR= 0.024). **B.** Schematics depicting the ARP2/3 complex, which consists of seven members, three of which were identified in the screen to be statistically significant (FDR $<$0.05) (highlighted in red, figure adapted from [226]). Depending on the combinations of α and β subunits, the extracellular domains of integrin heterodimer interact with the components of the extracellular matrix (ECM). This link between the intracellular cytoskeleton to the ECM via integrins has been shown to be important in cell motility, cell polarity, cell growth and survival (figure adapted from [227]). The screen was performed ten days post infection with the gRNA lentiviral library. For better clarity, an enlarged version of **A.** is also depicted in Appendix section figure A.3.

An interesting observation in this screen was also the identification of three genes relating to the members of Arp2/3 complex (*ARPC4, ACTR2, APRC3*). Arp2/3 protein complex consists of seven members and it is essential in the nucleation and assembly of branched actin filaments (figure3.14B) [226]. The cytoplasmic tails of β-subunits have been shown to be necessary and sufficient to link integrins to the actin cytoskeleton [228]. Actin binding proteins such as α-actinin, filamin and talin bind to actin filaments and mediate their link with cell surface integrin heterodimers (schematic depicted in figure 3.14C). Such interactions at the cytoplasmic domains of integrins have been shown to induce conformational changes in integrin extracellular domains (from a

'bent' (inactive) to 'extended' (active) conformation) that result in increased affinity for ligand in a process described as 'inside-out signalling' [229]. The screen here demonstrated that the components of the Arp2/3 complex were important for the recognition of ITGB1 at the cell surface by the mAb. This finding can be interpreted in two ways: (i) the expression of ITGB1 on the surface of cells depends on its interaction with the members of the Arp2/3 complex or (ii) the interaction with Arp2/3 complex with the cytoplasmic domain of ITGB1 changes the integrin subunit conformation that is required for the mAb epitope binding. The latter scenario is consistent with the presence of other mAbs for integrins that have been previously described to specifically recognise epitopes that are only present in the active conformation [230]. It is therefore likely that P16 falls into this class of integrin mAbs.

Summary of the genetic screens

All enrichment analysis here was done using MAGeCK software, which can use a single dataset to estimate a read count variance to determine the significantly enriched genes. I used a FDR of less than 0.05 as the cut-off point and estimated the biologically relevant genes that could be identified within that threshold. In all the screens that were carried out, the target receptor was identified with a very low FDR and well within the threshold (summarised in table 3.3).

Table 3.3 FDR of identification of the genes encoding direct receptor in a genome-scale screening approach using monoclonal antibodies targeting cell surface receptors.

Antibody	Day of screen	Target receptor	Target FDR	Genes (FDR<5%)
MEM6/6	Day 16	BSG	0.0012	4
BRIC222	Day 15	CD59	0.0002	58
BRIC125	Day 9	CD47	0.001	15
B6H12	Day 9	CD47	0.001	27
BRIC126	Day 9	CD47	0.004	35
P16	Day 10	ITGB1	0.024	52
BRIC256	Day 14	GYPA	0.001	9
BRIC5	Day 9	CD58*	0.001	14

*Full screen result for this screen is available in appendix section figure A.4

The number of significantly enriched genes identified in each screen depended on the quality of the screen performed in terms of the library size and the representation of the gRNA in the sorted cells. In the example of an

anti-BSG antibody where a stringent sorting threshold was used, very few genes were identified. This was improved when the number of sorted cells was increased in the subsequent screens.

Apart from the identification of the direct target receptor, specific cellular pathways required for the ligand recognition such as the GPI-anchor pathway (in anti-CD59 screen) and the O-glycosylation pathway (in anti-GYPA screen) were also identified. In addition, a pathway analysis of all enriched genes that were shared between antibody selections encoded proteins required for protein secretion and glycosylation, as expected, but also identified housekeeping pathways such as ribosome biosynthesis and RNA metabolism; genes identified in these pathways were grouped and labelled as 'other factors' (figure 3.15, also refer to table A.5 in the appendix section for specific 'other factors' that were enriched in at least two out the seven screens with FDR<0.05). I observed that the representation of these general pathways was often reduced when selections were performed several days later (day 15-16 rather than day 9), suggesting these genes are required for long-term cell viability in culture, and that antibody staining was reduced on moribund cells.

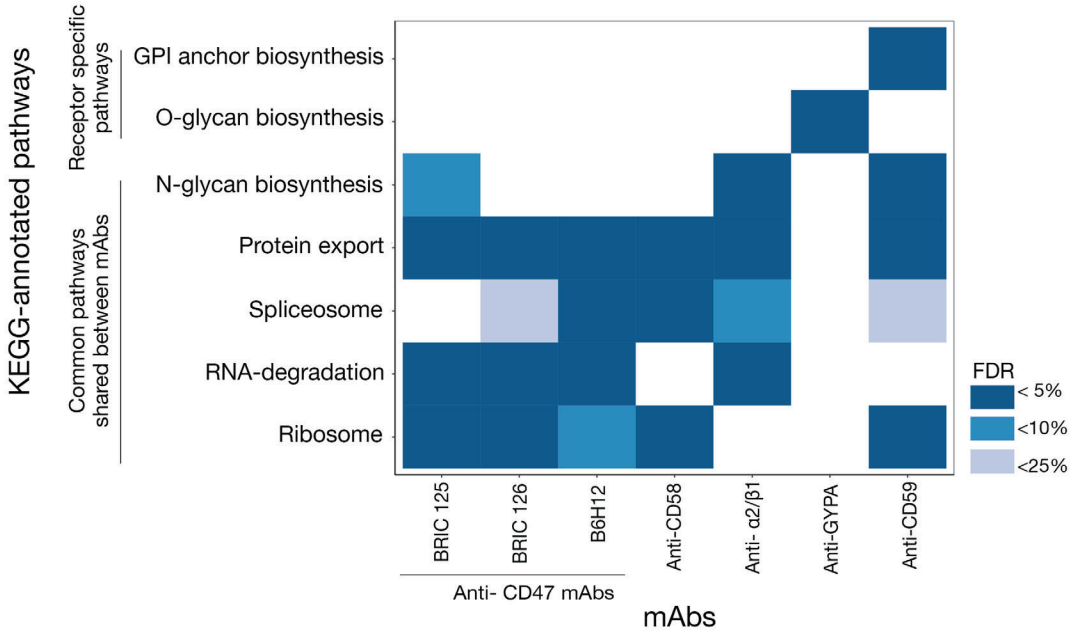

Fig. 3.15 Genes required for protein export, N-glycan biosynthesis and general housekeeping functions were enriched in, and shared between, cells selected for the loss of cell surface mAb staining. Pathway analysis on genes enriched in cells selected with the indicated mAbs identified some pathways shared between more than one antibody including protein export, N-glycan biosynthesis as well as more general housekeeping functions. No significantly enriched pathway was identified in the anti-BSG screen.

3.3 Discussion

In this chapter, I have demonstrated the use of a cell-based genome-scale CRISPR screening approach to identify genes encoding proteins required for extracellular recognition. I first generated high efficiency Cas9 cell lines by isolating stable clones grown from single cells within a polyclonal Cas9-expressing population. This approach allowed for the removal of cells which contained Cas9 with decreased efficiency resulting presumably from the acquisition of mutations that led to the inactivation of Cas9. I utilised two different methods to measure the Cas9 efficiency: (i) an endogenous gene KO method in which endogenous *BSG* was targeted, and (ii) a rapid exogenous method in which *GFP*, together with gRNA targeting *GFP*, was introduced to the cells and the efficiency of *GFP* targeting was measured. The exogenous approach provided a rapid means of determining the presence of functional Cas9 in the cells and was particularly useful during the single cell cloning steps, during which many clones had to be tested. However, targeting the single integrated GFP can be considerably more efficient than targeting both alleles of an endogenous gene (as exemplified by the lower efficiency observed for the same clonal line for *BSG* KO compared to *GFP* KO), so this approach usually represents the best-case-scenario for targeted gene KO. Additionally, endogenous targeting also depends on other factors, such as the protein-turnover time, the copy number of the target gene in a cell, and the targeting efficiency of the gRNA. To account for these factors, I opted to estimate the 'realistic' Cas9 efficiency based on the decrease in cell surface expression of BSG upon transduction with gRNA targeting *BSG*, rather than the *GFP* targeting approach.

In this work, I utilised the Yusa library to conduct genetic screens. There are several genome-wide libraries that have been described to date, which can be used to perform genome-wide KO screens [175, 176, 177, 205, 231, 232, 233]. Of these, four libraries are considered to be second generation libraries as they feature improved gRNA efficacy [234]. These are the 'Human V1 library' or the 'Yusa library' [176], the 'Whitehead library' [177], the 'Brunello library' [205] and the 'Toronto knockout libray version 3.0 (TKO3.0) library' [233]. The Whitehead library is the largest consisting of 10 gRNA/gene (182,134 gRNAs) compared to the 5 gRNAs/gene of the Yusa library (90,709 gRNAs) and 4 gRNAs/gene of the Brunello library (77,441 gRNAs) and the TKO3.0 library (71,090 gRNAs). The Yusa library differs from the other libraries as it uses the improved gRNA scaffold to avoid the T stretches as discussed in section 3.1.4. The design

of the gRNAs in the Yusa library, unlike those in the other three libraries, is not based on an on-target prediction for gRNA selection. However, this does not seem to affect the screening results as a recent comparison between the performance of these libraries in negative selection screens has shown that that these libraries identify genes with similar false discovery rates (between 14 % and 23 %) [234]. The same study has also suggested that while the higher number of gRNAs per gene, for example that of the Whitehead library, allows for better statistical confidence, this can also cause over-sensitivity of the analysis programs such as MAGeCK to call genes that have lower fold change values as statistical hits. In addition, as the number of gRNAs increases, the complexity of the library also increases, maintainance of which over an extended period of time can be practically challenging. This could instead lead to poor screen outcome. For second generation libraries it has been suggested that libraries with 6 gRNAs per gene are likely to be optimal for genome-wide CRISPR-KO screens [234]. While I only used the Yusa library in the study here, given the similarity of performance with the other libraries, it is unlikely that the other libraries would yield vastly differing results.

The genetic approach described in this chapter provides a valuable alternative to existing biochemical methods which must account for the largely insoluble nature of membrane-embedded receptors. An advantage of this method over existing methods is its ability to reveal the receptor protein at the cell surface which directly interacts with a presented ligand, but also identify other gene products that are required for the cell biology of the receptor. The screens carried out with the panel of mAbs demonstrated the ways in which the method can be used to identify the protein/glycan receptors and cellular factors such as chaperones, transcription factors, and cytoskeletal elements that are involved in the expression or correct positioning of the receptors on the surface of the cells.

The application of a genome-scale screening approach in the identification of monoclonal antibody targets can be a valuable tool for monoclonal antibody characterisation. A similar approach for this purpose has been described recently by others [235], where the cells refractory to antibody binding were enriched using two rounds of sorting and few gRNAs present in this enriched population were analysed using Sanger sequencing to identify the target receptor. Here I have extended the use of the genome-scale screening approach by using a single sorting approach that can not only capture gRNAs targeting genes encoding the antibody epitope, but also the gRNA targeting

genes contributing to the cell biology of the antibody epitope. In the example of an anti-integrin mAb, this included identifying the specific subunit of the heterodimer encoding the antibody epitope and components of the cytoplasmic Arp2/3 complex and for an anti-Glycophorin A mAb, a required role for enzymes involved in O-linked glycosylation in antibody binding.

In the genetic screens carried out here, I often identified genes that are required for the 'house-keeping' of receptors, including those involved in the secretory and glycosylation pathways. However, this did not greatly influence the confidence with which the receptor was identified. The target receptor was revealed in every case attempted within the FDR threshold of 0.05, with seven out of eight screens identifying the receptor with a very low FDR of under 0.005. The highest FDR of 0.024 was observed in the screen with the anti-ITGA2/B1 integrin antibody where admittedly, genes of the secretory and glycosylation pathway were enriched more than the target receptor. But even in such a scenario, assuming that the target of this antibody was unknown, the observation that (i) *ITGB1* was the only cell surface receptor encoding gene identified in the screen within the given threshold of 0.05; (ii) components of the Arp2/3 complex that fit with the integrin biology were identified among the most enriched genes and; (iii) the fact that genes that contribute to general receptor house-keeping can be ruled out as specific factors, suggest that it would have been possible to identity *ITGB1* as the target receptor.

One of the challenges of loss-of-function screen studies is the investigation of the effect of the genes, which are essential for cell growth and viability, on the phenotype being tested. Comprehensive studies carried out out using CRISPR–Cas screens on multiple cell lines have identified approximately 2000 'core-genes' that are designated to be 'essential' for optimal growth of the cells [175, 176, 177, 205]. Identifying the role of cell-essential genes on cellular recognition can be challenging as cells that contain mutations in essential genes become non-viable and are no longer represented in the mutant library. However, it has been suggested that the number of genes thought to be 'core-essential' could be overestimated as this includes genes that not only affect viability of the cells but also moderately affect cell growth [236, p. 357]. CRISPR-Cas9 KO screens carried out recently in the context of virus-host interactions have shown that it is possible to identify the effect of genes that have been categorised as core-essential genes in genome-scale KO screens (for example, the role of oligosaccharyl transferase complex (OST) in dengue virus infection [220]). The data here demonstrated that such screens can

indeed identify the core-essential genes involved in general protein export and glycosylation pathways, but unlike the direct receptors, genes in these categories were not identified in every case (e.g., not identified in anti-BSG and anti-GYPA screens). The timing (early time points rather than later) and the quality of the screen parameters in terms of the day post transduction the screening was carried out, the number of cells in the sorting population and the sorting stringency seemed to influence the successful identification of essential genes in such screens.

In summary, I have developed and applied a cell-based genetic method based on CRISPR-KO technology using mAbs to identify genes that mediate high affinity interactions at the cell surface. The method is able to identify the direct receptors at the cell surface robustly with high confidence and often also identifies cellular components that are related to the biology of the receptor.