Chapter 3

Analysis of iPSC CRISPR/Cas9 screens

3.1 Introduction

A variety of methods are available to analyse CRISPR/Cas9 knockout screens. There are no examples in the literature of genome-wide screens in iPSCs and few in hESCs, but a vast amount of data has been published for hundreds of cancer cell line screens. Using insights gained from these studies, we analysed data from our screens in the parental BOB and KO derivatives. We initially performed basic quality control measures and tested screen performance, using published datasets as a reference. To evaluate reproducibility, we compared the data from all lines including the results of biological replicates of both the parental and *TP53* KO lines. Using various filtering strategies, we identified candidate SLIs for all 15 of the TSGs studied. In addition to exploring genes that were required for cell fitness in the iPSCs, we also identified genes that, when lost, appeared to provide a proliferative advantage.

3.1.1 Aims of this chapter

- To assess the quality of our iPSC screen data and evaluate the performance of the screens in terms of recall of known fitness genes.
- To assess the reproducibility of the screens by comparing the results of all lines, with a particular focus on biological replicates of the same lines.
- To filter for dependencies that were specific to KO lines and identify candidate SLIs.
- To investigate enrichment of genes in the iPSC screens.

3.2 Screen data quality control

3.2.1 Sequencing coverage

In total, 24 screens were performed as part of this project: I performed 5 screens and a further 19 were performed by Rebecca McRae and Verity Goodwin (CGaP, WSI). As discussed in Section 2.7.2, these screens were performed using the same protocol and library but added an additional passage was added by CGaP. Each screen was carried out in technical triplicate, with the exception of the *FAT1* KO line which had only two replicates due to unexplained cell death in one replicate. All samples were processed in the same way and gRNAs were sequenced on a HiSeq 2500, with 6 samples multiplexed per run. There was slight variation in sequencing depth between different samples and runs, which was accounted for by normalisation prior to further analysis. An average of 4.07×10^7 read counts per replicate mapped to the gRNA library (Fig. 3.1). This was equivalent to ~400x coverage of the library, with no samples dropping below 200x coverage (Fig. 3.1).

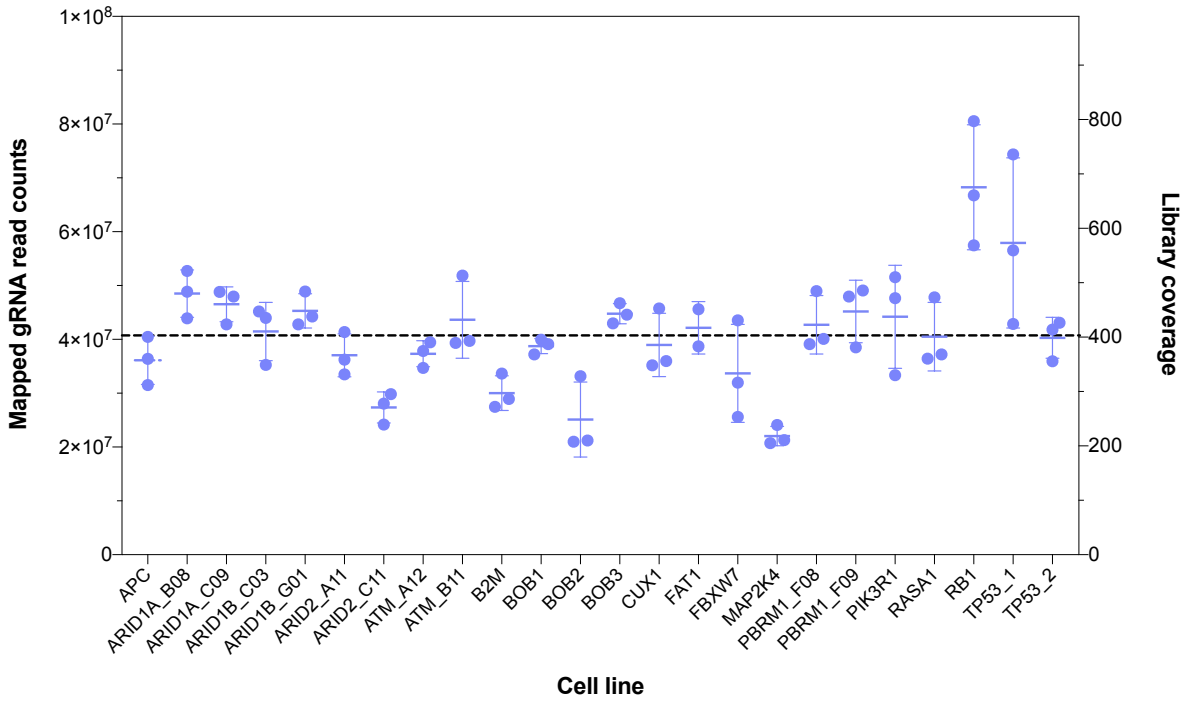

Figure 3.1. Sequencing coverage across all iPSC screens. PCR was performed to amplify gRNAs present in the genomic DNA of each screened cell line. gRNAs were then sequenced on a HiSeq 2500 and mapped to the library. The number of mapped reads is plotted for each replicate in every screened cell line (left y-axis). The corresponding library coverage that was achieved is also shown (right y-axis). The dotted line indicates the mean across all samples. BOB, BOB_2 and BOB_3 refer to replicate screens of the parental BOB line. TP53 and TP53_2 refer to replicate screens of the *TP53* KO line.

3.2.2 Enrichment of non-targeting controls

When looking at the initial gRNA read count data prior to analysis, it was noted that there was an enrichment of non-targeting control (NTC) gRNAs in the screen samples compared to the library plasmid. The $\log_2$(fold-change) in abundance between plasmid and screen for the targeting gRNAs was distributed around 0, whereas NTC gRNAs were enriched (an average of 1.2 for the screens shown in Fig. 3.2). The NTC gRNAs do not target any region in the genome, hence Cas9 would not induce DSBs in cells expressing these gRNAs. In line with our previous observation of Cas9-induced toxicity (Section 2.7.1), we hypothesised that cells expressing NTC gRNAs had a proliferative advantage due to the lack of DNA damage. Therefore, widespread depletion of cells expressing targeting gRNAs but not those expressing NTC gRNAs would cause this observed enrichment of the controls. This effect was also observed in the *TP53* KO line screen (Fig. 3.2), suggesting that depletion of TP53 was not sufficient to avoid this toxicity. Other studies have since reported similar findings (discussed in Section 5.2.2). We decided to remove the NTC gRNAs from all further analysis as they skewed the results rather than acting as controls.

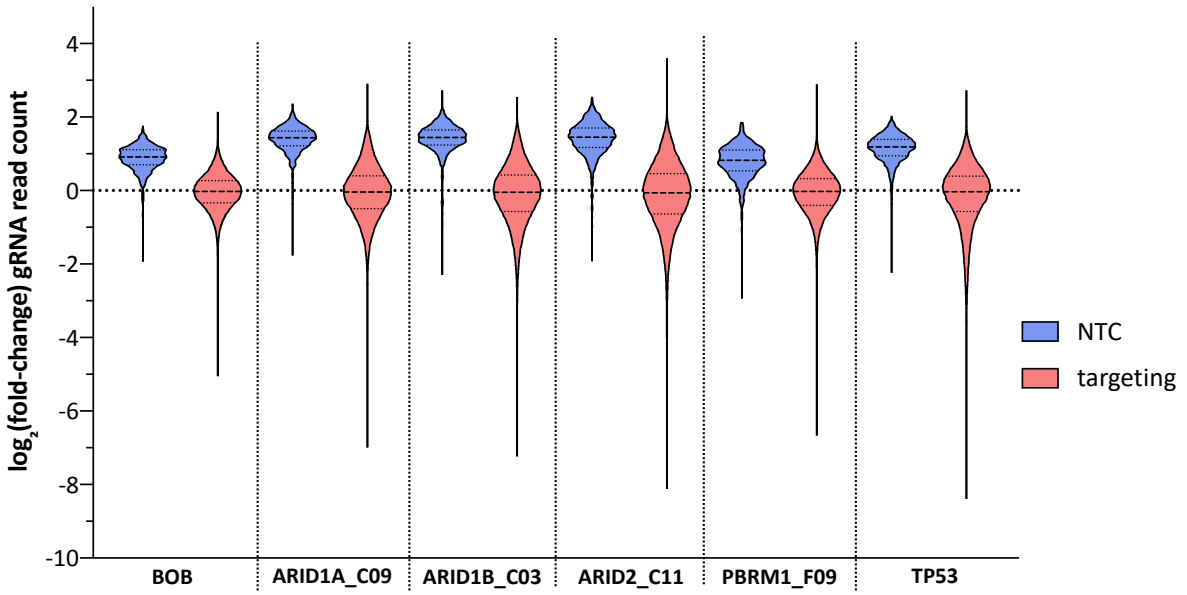

Figure 3.2. Fold-changes of non-targeting and targeting gRNAs. The $\log_2$(fold-change) in read count between the screen and the library plasmid was calculated for NTC gRNAs and all other targeting gRNAs. Results are shown for screens in the BOB, ARID1A_C09, ARID1B_C03, ARID2_C11, PBRM1_F09 and TP53 cell lines. The values shown were calculated using the average read count of the technical replicates in each screen.

3.2.3 Correlation between technical replicates

The correlation of gRNA read counts between technical replicates was measured for each cell line using Pearson's Correlation Coefficient. The average R value was calculated for each line, with a median of 0.82 across all screens (range 0.6-0.9) (Fig. 3.3a). In a recent study of 324 cancer cell lines using the Yusa v1.1 library, a median R of 0.8 was achieved.[109] However, it was noted that this correlation was not sufficient to distinguish between replicates of the same cell line and any two random cell lines. To gain a better measure of reproducibility, in that study they selected gRNAs that had an average pairwise Pearson's Correlation of > 0.6 across all screens when comparing the count fold-changes at the screen endpoint vs the plasmid library. For each replicate, they then calculated the average gene-level fold-change for only the genes targeted by these 'reproducible gRNAs'. Pearson's Correlation of these fold-changes was assessed between all replicates across all screens, and a reproducibility threshold was defined that would allow distinction between replicates of the same cell line and random lines. We repeated this analysis on our screens and identified 279 gRNAs that were reproducible (average R > 0.6). However, this measurement assumes that the cell lines being compared are independent, but the cell lines we screened were almost genetically identical. Unsurprisingly, the correlation between cell lines was still not distinct from that observed between replicates. Based on the reproducibility threshold identified in the cancer cell line screens (R=0.68), the majority of our screens passed this quality control test (Fig. 3.3b). A more thorough approach could be to repeat this analysis using a combined set of data from the iPSC screens and cancer cell line screens. Comparing the iPSCs to independent lines may provide a more reliable threshold to assess replicate reproducibility.

Notably, the initial 5 screens had the lowest correlations in both analyses, suggesting that the screen quality was improved by the addition of a passage. This may be due to a reduction in the cell death that occurred when the cells approached confluency. We also noted that replicate A of the ARID1A_C09 screen negatively impacted the average correlation. Replicates 1 and 2 had R = 0.83 for all gRNA counts and R = 0.76 for reproducible gRNA fold-changes. During this screen, the level of transduction in replicate 1 was higher and more cell death was observed, which may explain the lower correlation. We therefore decided to exclude replicate 1 of the ARID1A_C09 screen from further analysis.

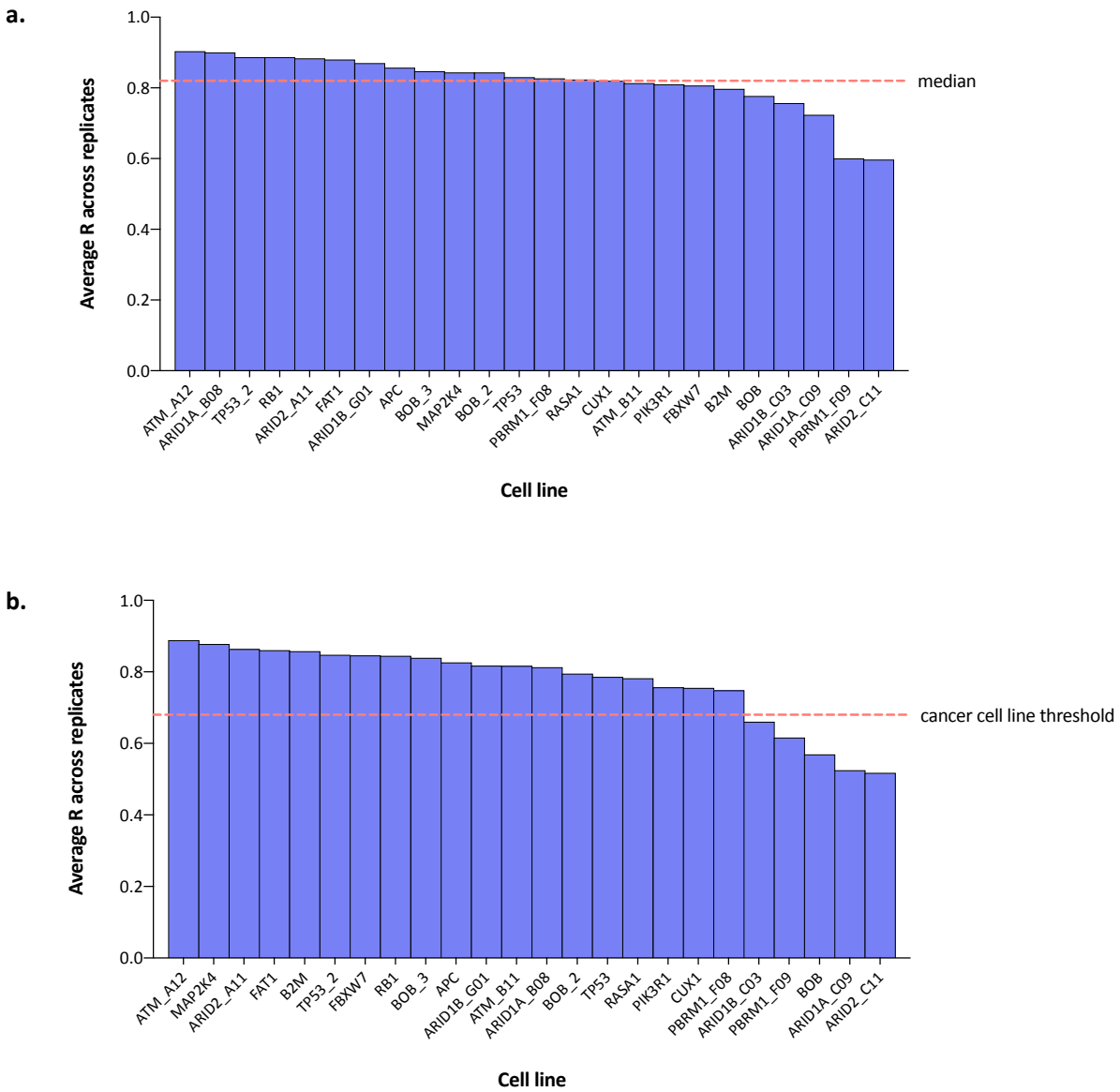

Figure 3.3. Correlation between screen technical replicates. a) Pearson's Correlation values were computed between technical replicates of each cell line using gRNA read counts. The average R value for each screen is shown, with the median represented by a dotted line, **b)** Pearson's Correlation values were computed between technical replicates of each cell line using the gene-level fold-changes only for reproducible gRNAs. The average R value for each screen is shown. The dotted line represents the reproducibility threshold that was defined by Behan *et al.* based on screens in cancer cell lines.[109]

3.3 Identifying effects of gene loss on cell fitness

Several methods are available for analysis of CRISPR/Cas9 knockout screen data. We chose two of the most well-established methods, described below, to identify genes that were significantly depleted in our screens. These analyses allowed us to compare our data to previously published cancer cell line screens that were processed in the same way after screening with the same gRNAs.[109] Our library had an average of 5 gRNAs/gene but, for a subset of ~2000 cancer-related genes, there were 10 gRNAs/gene. This may have caused a bias towards genes that were better represented and we were unclear how to account for this, so we removed the data for all additional gRNAs before further analysis.

3.3.1 Bayesian Analysis of Gene EssentiaLity

Bayesian Analysis of Gene EssentiaLity (BAGEL) is an algorithm developed to analyse genetic perturbation screens using *a priori* known training sets of 'essential' and 'nonessential' genes.[114] The 360 essential genes were defined based on their essentiality in at least 50% of a set of shRNA screens, and constitutive expression in a panel of cell lines. Using the same panel, the 972 nonessential genes were defined as those which generally lacked expression in these lines. Firstly, median-ratio normalisation is performed on all raw read counts to account for differences in sequencing coverage. A $\log_2$(fold-change) is then calculated for each gRNA, comparing the abundance at the screen endpoint to that in the library plasmid. The average $\log_2$(fold-change) for each gRNA is calculated across screen replicates. BAGEL then uses the fold-change distribution of all gRNAs targeting the essential and nonessential genes (Fig. 3.6) to calculate the likelihood that a given gRNA belongs to either set, based on the observed fold-change. The output of this probability calculation is a value termed the Bayes Factor (BF); every gRNA is assigned a BF. A recently published R implementation (BAGELR) calculates the gene-level BFs by taking the average of the gRNA values.[109] The original Python version of BAGEL calculated a sum rather than an average value.[114] A positive BF indicates that the gene is likely to be essential for cell fitness.

Bayes Factors were computed using BAGELR for all of our screens (Appendix A.5). To determine statistical significance, a threshold of 5% False Discovery Rate (FDR, 1- Precision) was defined for each screen. Genes were assigned a scaled BF (Appendix A.6), which was calculated by subtracting the BF at the 5% FDR threshold for that screen from the original BF. Any gene with a scaled BF > 0 was considered to be significantly depleted.

3.3.2 Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout

Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) is an algorithm which, unlike BAGEL, identifies significant differences in gRNA abundance using no prior knowledge.[116] Median-normalisation is performed on the read counts to account for differences in sequencing depth across replicates and conditions, and the gRNA variance is estimated. Replicates can be analysed together, with a mean read count calculated for every gRNA. Using a negative binomial model, MAGeCK then determines whether the abundance of a gRNA is significantly different between the control and treatment; in our analysis this was the library plasmid vs the screen endpoint. A robust ranking aggregation (RRA) algorithm is used to rank gRNAs by the p-value obtained from the negative binomial model (Fig. 3.4).

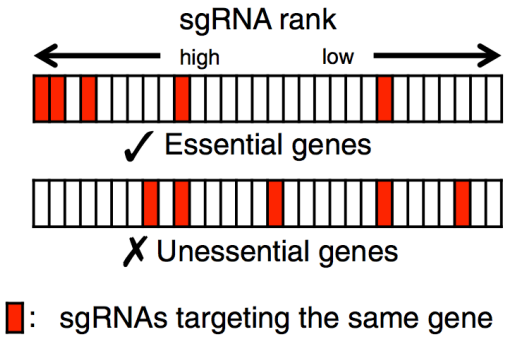

Figure 3.4. Robust rank aggregation. A negative binomial model is used to determine the significance of any change in gRNA abundance compared to control. gRNAs are then ranked based on their p-value. If a gene is essential, gRNAs targeting it should be ranked highly more often than expected. If a gene is nonessential, gRNAs targeting it should be uniformly distributed. Figure taken from [116].

If a gene has no effect on fitness, the assumption is that gRNAs targeting this gene will be evenly distributed in the rankings. If several gRNAs targeting a gene are ranked higher than expected, this gene would be considered significant. Each gene is assigned a p-value and an FDR is computed using the Benjamini-Hochberg method. MAGeCK can be used for bi-directional analysis; from one screen it can identify genes whose knockout impairs cell fitness (negative selection) and genes whose knockout induces cell proliferation (positive selection). Genes under negative selection would have a significant depletion of gRNAs compared to the control. Those under positive selection would have significant enrichment of gRNAs compared to the control. MAGeCK analysis was performed on all of our screens to calculate depletion values for every gene (Appendix A.7). A threshold of negative FDR 0.1 was applied to identify significant hits.

3.4 Assessing screen performance

3.4.1 Receiver Operating Characteristic & Precision-Recall curves

As a measure of the sensitivity and specificity of the screens, receiver operating characteristic (ROC) and precision-recall (PrRc) curves were computed (Fig. 3.5). This was done using the gene-level count fold-changes, with the average taken across technical replicates for each screen. Using the pre-defined sets of BAGEL essential and nonessential genes, these analyses can indicate how well a screen performed. ROC curves plot sensitivity (i.e. true positive rate) against specificity (1 - false positive rate). The area under the curve (AUC) is a measure of how accurately the essential and nonessential genes were identified as distinct groups. If these genes cannot be separated, the AUC would be 0.5. A screen with 100% specificity and sensitivity would have an AUC of 1. PrRc curves plot recall (i.e. true positive rate, the same as sensitivity) against precision (i.e. positive predictive value). Precision and specificity are slightly different: precision measures how many of the predicted positives are actually true positives; specificity measures how many of the expected negatives are called as negative. Similar to the ROC curve, a high AUC for the PrRc curve indicates good performance with high precision and recall. Based on these models, performance across all of the screens was fairly consistent. The median area under the ROC curve was 0.91 (Fig. 3.5a) and area under the PrRc curve was 0.87 (Fig. 3.5b). These results were similar to those obtained in the Behan *et al.* (2019) study of cancer cell lines (0.92 and 0.9, respectively).[109]

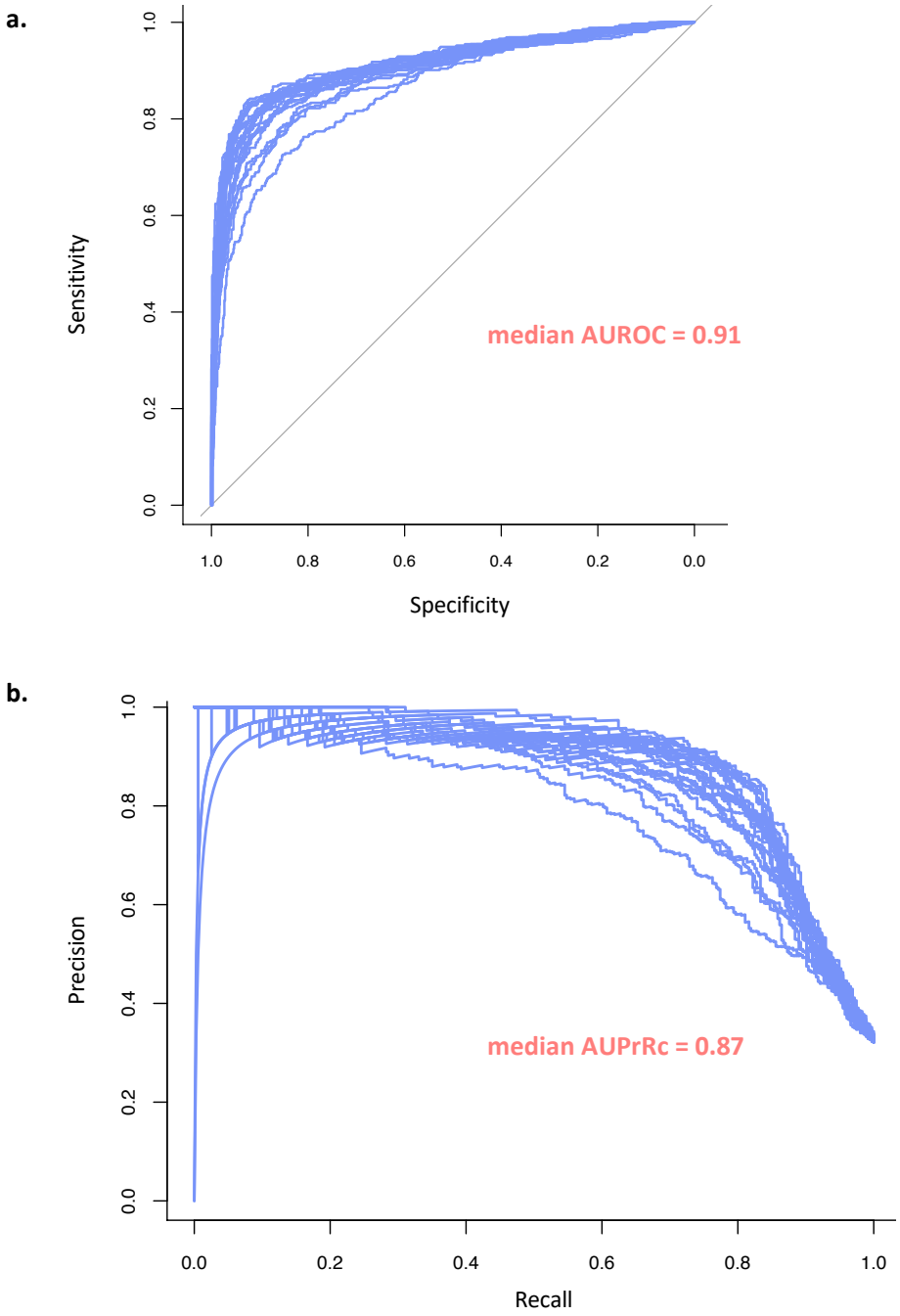

Figure 3.5. Assessment of screen performance. ROC **(a)** and PrRc **(b)** curves were plotted for all 24 screens. These were obtained by classifying the pre-defined BAGEL essential (n=354) and nonessential (n=747) genes using the gene-level fold-changes. The median AUC across all screens is shown for both. These were computed using the ROC and PrRc functions in the CRISPRcleanR package.[255]

3.4.2 Distributions of BAGEL essential and nonessential genes

Another indicator of screen performance is the separation between the results for the BAGEL essential and nonessential genes. We analysed the distribution of gene-level fold-changes (Fig. 3.6) and scaled BFs (Fig. 3.7) for each of these sets across all screens. There was a slight separation between the essential and nonessential genes but the overlap was high, although this was improved in the scaled BF distributions. It may be the case that some of these genes were not essential in this iPSC line, as these genes were identified in immortalised cell lines. We also analysed the fold-change distribution for genes encoding ribosomal proteins, which we would expect to be vital for cell function regardless of the cell type (Fig. 3.6). These were slightly more depleted and separated from the nonessential genes compared to the essentials, but in some cases (e.g. PBRM1_F09) there was still high overlap. It appeared that depletion was simply not large enough to clearly separate the sets, suggesting this was most likely an issue with screen performance. For reference, the fold-changes (Fig. 3.6) and scaled BFs (Fig. 3.7) for an ovarian cancer cell line, A2780ADR, are shown. This cell line was screened by Behan *et al.* (2019) using the same library and the data was processed in the same way.[109] This cell line passed all quality control tests and had high AUC values for the ROC (0.93) and PrRc (0.93) analyses, thus we considered it to be a good representation of a high-quality screen. For this cell line, the fold-changes for essential and ribosomal genes spread further and were more distinct from the nonessential population than in our iPSC screens. There was also a greater separation between the scaled BFs for the essential and nonessential genes in the A270ADR screen, with the majority of essential genes being correctly called as essential. Of all the lines we screened, the results for the *TP53* knockout line were most similar to this cancer cell line, indicating that this was the best performing screen.

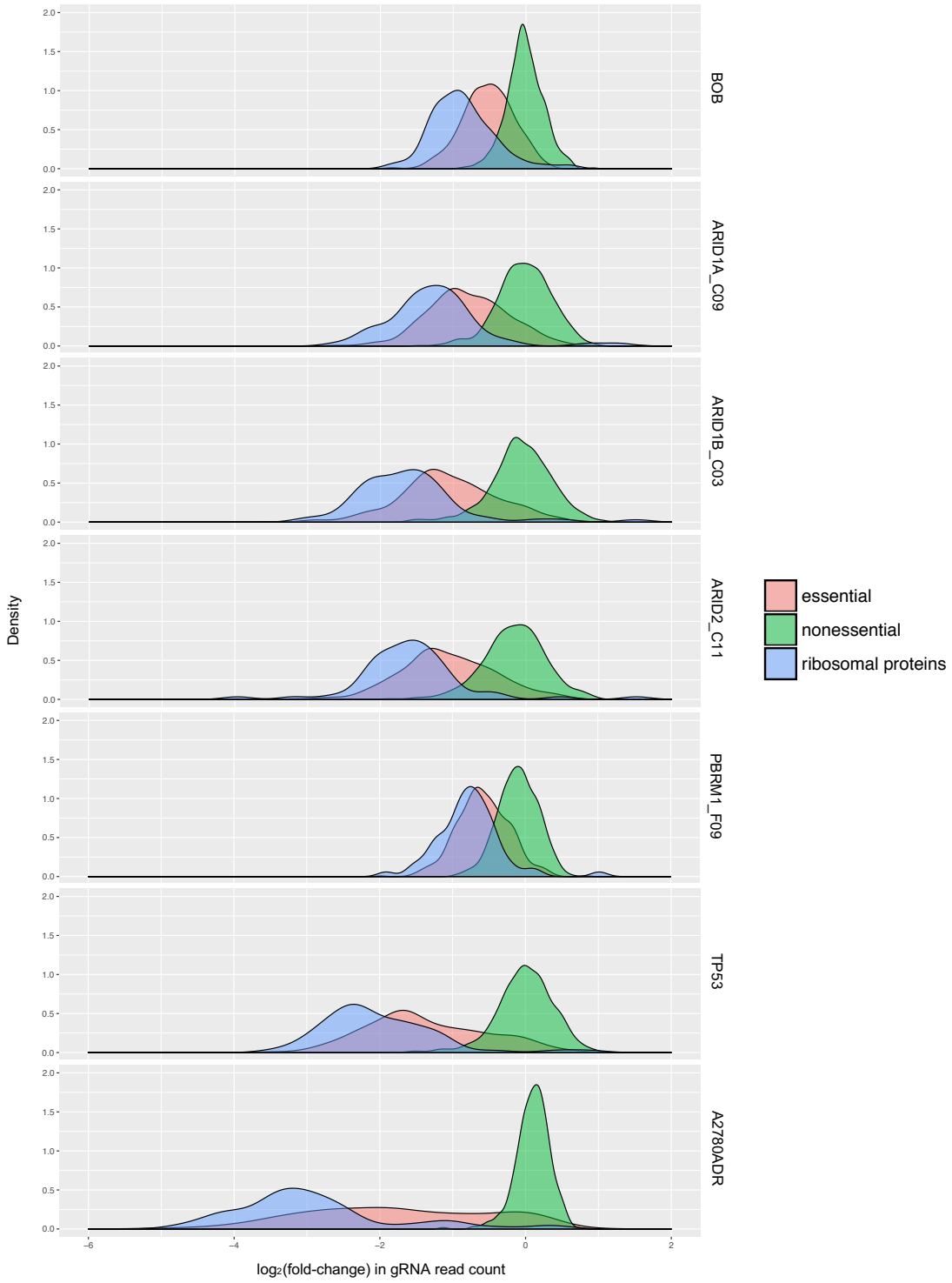

Figure 3.6. Distribution of fold-changes for BAGEL essential and nonessential genes. The fold-change of each gRNA was calculated for every screen replicate relative to the library plasmid. The average of the replicates was calculated for each cell line, and then a gene-level fold-change was calculated by taking average of the values for all of the gRNAs targeting each gene. The distribution of $\log_2$(fold-changes) of the BAGEL essential (n=354) and nonessential (n=747) genes are plotted for the parental BOB, ARID1A_C09, ARID1B_C03, ARID2_C11, PBRM1_F09 and TP53 screens. For comparison, results are also shown for the A2780ADR ovarian cancer cell line, screened by Behan *et al.* using the same library.[109] Distributions are also shown for the genes that encode ribosomal proteins.

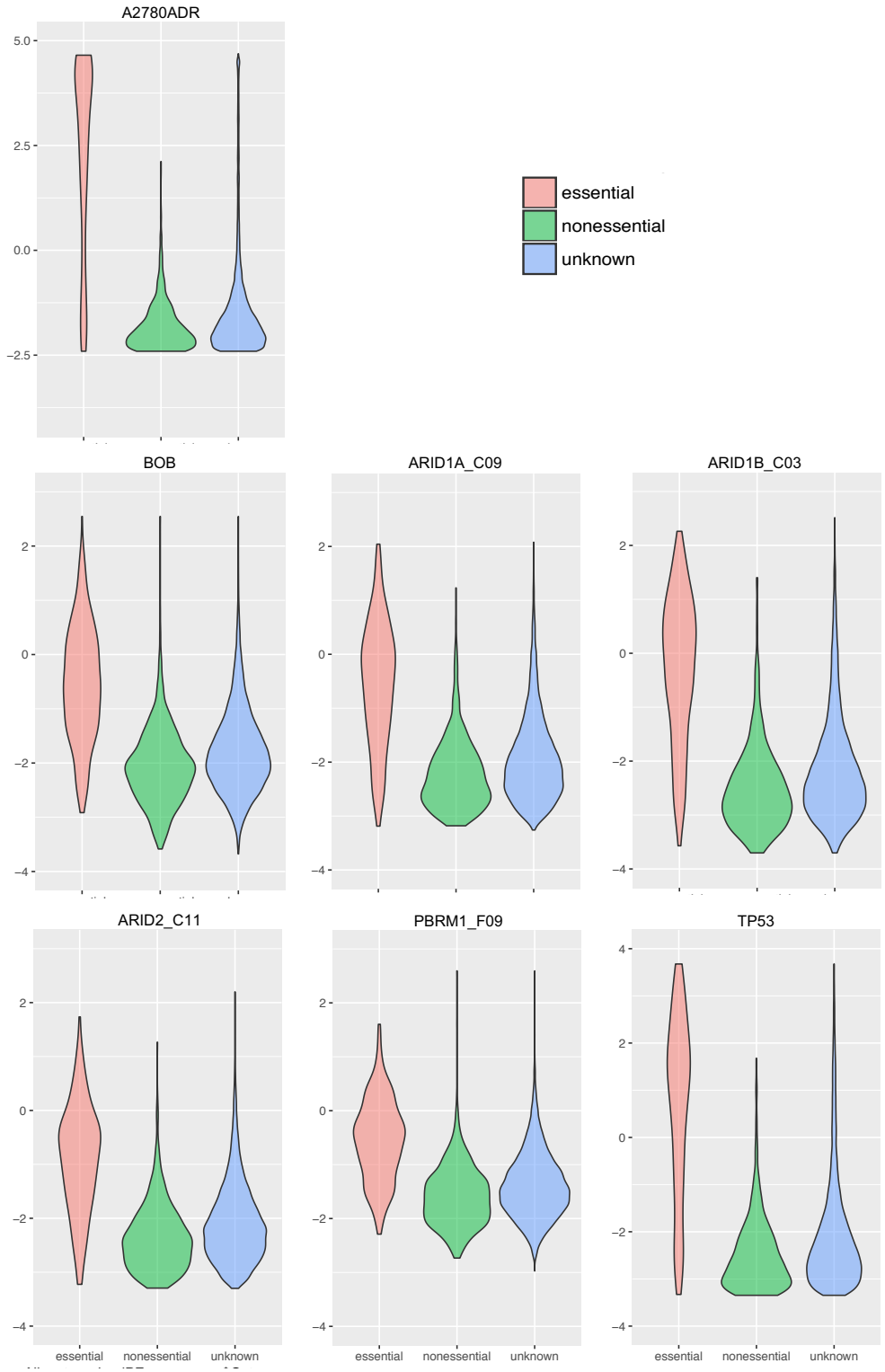

Figure 3.7. Distribution of scaled Bayes Factors. BAGEL was applied to compute BFs for all genes in every screen, calculating an average across replicates. The values were scaled using a 5% FDR threshold, with a value > 0 representing a significant hit. Plots show the distributions of the scaled BFs for the BAGEL essential (n=354) and nonessential (n=747) genes, and all other genes (unknown, n=16,906). Results are shown for the parental BOB, ARID1A_C09, ARID1B_C03, ARID2_C11, PBRM1_F09 and TP53 screens. For comparison, results are also shown for the A2780ADR ovarian cancer cell line, screened by Behan *et al.* using the same library[109].

3.4.3 Recall of known fitness genes

Whilst the ROC and PrRc curve models performed well, the distribution analyses indicated that we may have had issues with detecting known fitness genes and that performance was variable. In the Behan *et al.* (2019) pan-cancer study, a set of 552 pan-cancer core fitness genes were identified[109], providing an additional reference set for comparison. Using both the BAGEL and MAGeCK analyses outputs, we assessed exactly how many of the BAGEL essential genes and pan-cancer core fitness genes were called as hits in our screens (Table 3.1). Although these gene sets were identified in cancer cell lines, the pan essentiality across many cell types suggests that they are likely required for general cell fitness and survival, regardless of tissue type or tumourigenicity. Considering this, it was expected that many of these should also be essential in our iPSCs. As our KO lines were all derived from the same parental, we also expected that there should be a large overlap in the essential genes identified. Consistently more of the pan-cancer core fitness genes were detected than the BAGEL essential genes (Table 3.1). This may be because the BAGEL gene list was derived from shRNA screen data rather than CRISPR/Cas9 screen data. However, in general the number of essential genes identified in the parental and KO iPSC screens was highly variable. In several screens, including one of the parental screens (BOB_2), there was a high recall of core fitness genes. This suggested that many of the core genes identified in cancer cell lines were also essential in this iPSC line, but the ability to consistently detect them was impeded by variable screen performance.

Table 3.1. Recall of pre-defined essential genes in iPSC screens. MAGeCK and BAGEL were applied to identify significantly depleted genes in each screen. The % of pan-cancer core fitness genes (n=552) and BAGEL essential genes (n=354) that were called as hits are shown, based on the results of both analyses. Increasing colour intensity reflects increasing % recall.

	MAGeCK		BAGEL	
Cell line	**Core fitness**	**BAGEL essential**	**Core fitness**	**BAGEL essential**
APC	63%	48%	63%	47%
ARID1A_B08	36%	25%	15%	12%
ARID1A_C09	28%	21%	47%	35%
ARID1B_C03	44%	34%	65%	48%
ARID1B_G01	42%	31%	49%	38%
ARID2_A11	57%	42%	30%	22%
ARID2_C11	15%	10%	26%	17%
ATM_A12	67%	52%	66%	50%
ATM_B11	52%	39%	10%	9%
B2M	72%	54%	91%	66%
BOB	18%	13%	41%	32%
BOB_2	47%	36%	80%	60%
BOB_3	47%	35%	13%	12%
CUX1	60%	45%	83%	62%
FAT1	29%	21%	51%	39%
FBXW7	39%	28%	39%	28%
MAP2K4	53%	40%	58%	43%
PBRM1_F08	43%	33%	61%	46%
PBRM1_F09	6%	5%	30%	24%
PIK3R1	40%	29%	33%	21%
RASA1	61%	47%	80%	60%
RB1	68%	52%	53%	40%
TP53	70%	53%	95%	68%
TP53_2	69%	51%	91%	66%

3.5 Comparison of MAGeCK and BAGEL

The inconsistent detection of established core fitness genes indicated that there may be high variability in the overall essentiality profiles of these cell lines. As described previously, both MAGeCK and BAGEL were used to identify genes that were significantly depleted in the parental and KO lines. We calculated the total number of significant hits in each screen and compared the results from both analyses (Table 3.2). For many lines, the number of hits called by MAGeCK and BAGEL varied considerably. However, the overlap of the genes that were identified was generally high. In screens where BAGEL detected less than MAGeCK, the majority of the hits detected by BAGEL were also identified by MAGeCK, and vice versa. It is not surprising that these analyses identified different hits and it was reassuring to see that many of these were shared. However, it is not clear why there was no trend in the variability: MAGeCK detected more hits in some screens but BAGEL detected more in others. This was also reflected in the detection of *a priori* known essentials discussed earlier, with inconsistent variability between both analyses. The results of these analyses are dependent on the chosen significance threshold. Here, a threshold of FDR 0.1 was used for MAGeCK and FDR 0.05 was used for BAGEL. These can be adjusted to alter the stringency of the analysis; increasing the stringency too far will result in identification of very few hits and decreasing it may introduce noise and cause a high false positive rate. The most robust hits are likely to be those that were identified by both analyses, although this may lead to an increased false negative rate. For subsequent analysis, we considered the outputs from both MAGeCK and BAGEL rather than excluding data.

Table 3.2. Number of significantly depleted genes identified in iPSC screens. MAGeCK and BAGEL were applied to identify significantly depleted genes in each screen. The number of genes called as hits by MAGeCK using an FDR of 0.1 and by BAGEL using an FDR of 0.05 are shown for each cell line. The overlap of hits that were identified by both analyses is also shown.

Screen	MAGeCK hits	BAGEL hits	Overlap
APC	1068	863	713
ARID1A_B08	490	156	116
ARID1A_C09	413	671	343
ARID1B_C03	715	962	589
ARID1B_G01	654	637	435
ARID2_A11	1009	353	319
ARID2_C11	232	343	144
ATM_A12	1235	995	848
ATM_B11	872	95	86
B2M	1244	1506	1109
BOB	264	625	228
BOB_2	758	1379	714
BOB_3	820	151	128
CUX1	981	1384	875
FAT1	475	832	419
FBXW7	671	541	383
MAP2K4	930	885	689
PBRM1_F08	675	905	565
PBRM1_F09	86	528	75
PIK3R1	680	429	326
RASA1	1029	1244	868
RB1	1258	683	613
TP53	1178	1656	1100
TP53_2	1079	1468	991

3.6 Assessing screen reproducibility

3.6.1 Comparison of biological replicates

All screens were carried out in technical triplicate (or duplicate for the *FAT1* KO) with cells split into three populations at the passage prior to setting up the screen, and then transduced and maintained separately throughout. However, we considered that biological replicates may be more informative with regards to reproducibility. For the parental BOB line and the *TP53* KO line, screening was repeated weeks apart using cells thawed from different vials. The data was analysed as described previously, and the overlap of the results was assessed (gene lists are provided in Appendix A.8).

I carried out the initial parental BOB screen and two further biological replicates were performed by CGaP using their adapted protocol. Using MAGeCK, only 185 genes were significantly depleted in all replicates (Fig. 3.8). A further 314 hits were detected in both the second and third screens, but not in the first. Using BAGEL, fewer genes were identified in BOB_3 but almost all of them overlapped with BOB_2 (Fig. 3.8). Similarly, the majority of hits from BOB were also found in BOB_2. The detection of core fitness and BAGEL essential genes was higher in BOB_2 than in the others (Table 3.1). Thus, the incomplete overlap may be due to poorer performance in the BOB and BOB_3 screens.

The replicates of the *TP53* KO line (referred to as *TP53* and *TP53*_2) had a greater correlation, with a higher overlap between the hits identified using both BAGEL and MAGeCK (Fig. 3.9). BAGEL detected more significantly depleted genes in both replicates, but these included almost all of the genes detected by MAGeCK. When the overlap of both replicates from both analyses were compared, 847 genes were found to be significantly depleted, in comparison to only 62 in the parental overlap.

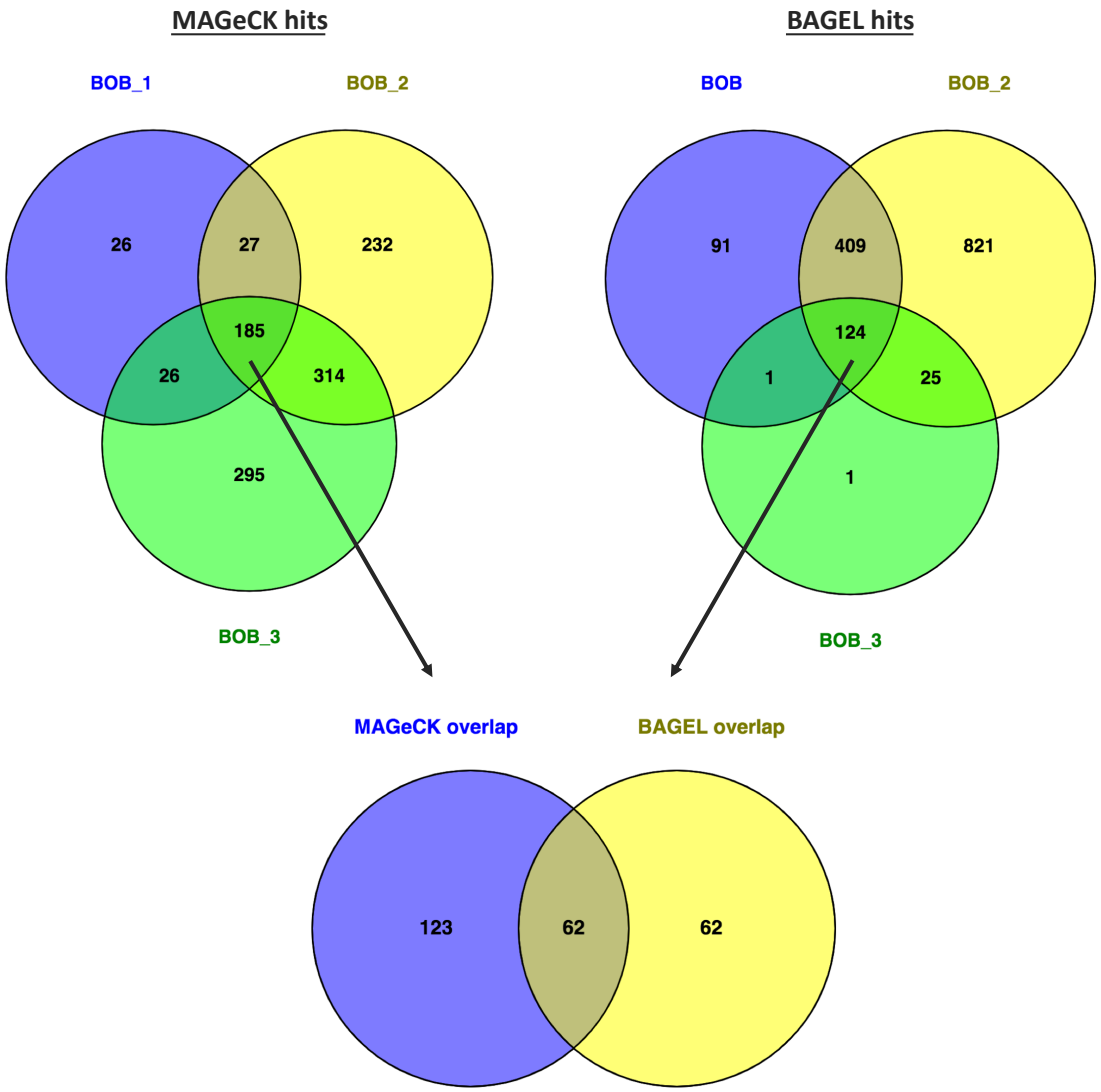

Figure 3.8. Overlapping hits in biological replicates of the parental BOB screen. The parental BOB cell line was screened three times, with technical triplicate in each case. MAGeCK and BAGEL were applied to identify genes that were significantly depleted compared to the library plasmid. The outputs for all three screens were compared to find common hits. The overlapping MAGeCK hits were compared with the overlapping BAGEL hits to assess the correlation of the two analyses. Diagram created using Venny.[256]

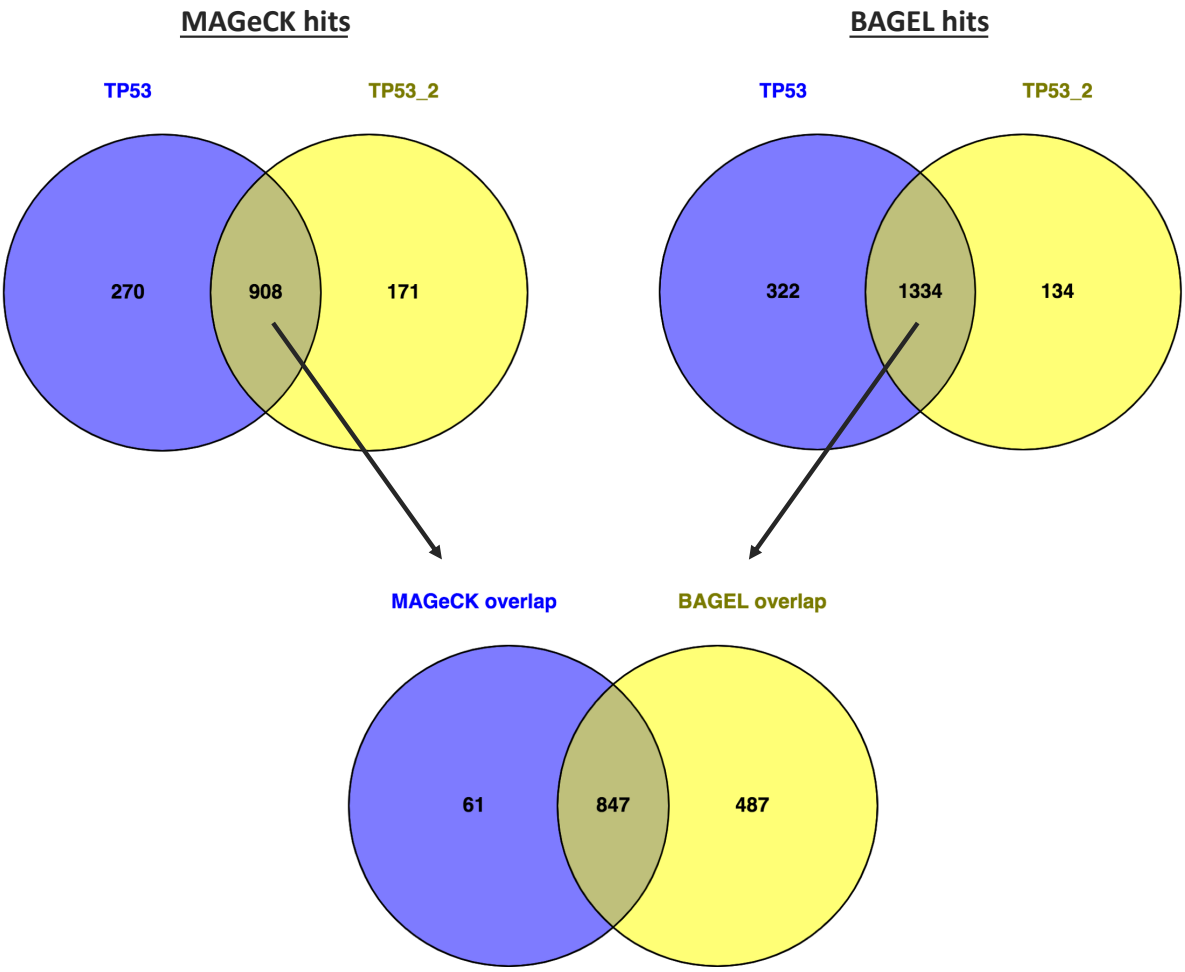

Figure 3.9. Overlapping hits in biological replicates of *TP53* KO line screen. A *TP53* KO derivative of BOB was screened twice, with technical triplicate in both experiments. MAGeCK and BAGEL were applied to identify genes that were significantly depleted compared to the library plasmid. The outputs for both screens were compared to find common hits. The overlapping MAGeCK hits were compared with the overlapping BAGEL hits to assess the correlation of the two analyses. Diagram created using Venny.[256]

3.6.2 Comparison across all screens

As all of the cell lines differed by only a single genetic change, we considered that results from screening different lines could also act in some way as biological replicates. Thus, we compared the results across all screens as another measure of reproducibility and to further define core fitness genes in the BOB iPSC line. We anticipated that the majority of the hits would be shared, however very few were identified in every screen (25 using BAGEL, 17 using MAGeCK) and 18% of genes were called only once (Fig. 3.10). We analysed all of the genes that were significant in 20-23 (out of 24) screens to determine whether specific screens consistently failed to identify common hits (Fig. 3.11). Using the BAGEL output, 5 screens in particular (ARID2_C11, PBRM1_F09, BOB_3, ARID1A_B08 and ATM_B11) consistently failed to detect hits that were identified by the majority of the other screens. The results were slightly different using the MAGeCK output, with the ARID2_C11, PBRM1_F09, BOB and ARID1A_C09 screens accounting for the majority of missed hits. Whilst BAGEL and MAGeCK differed, the results correlated well with their respective detection of known essentials/fitness genes. The screens that failed to detect the highest number of common hits also had the poorest recall (Table 3.1). In line with the previous data, this indicated that the screens were not highly reproducible. This limited our ability to accurately define core essential genes for the BOB iPSC line.

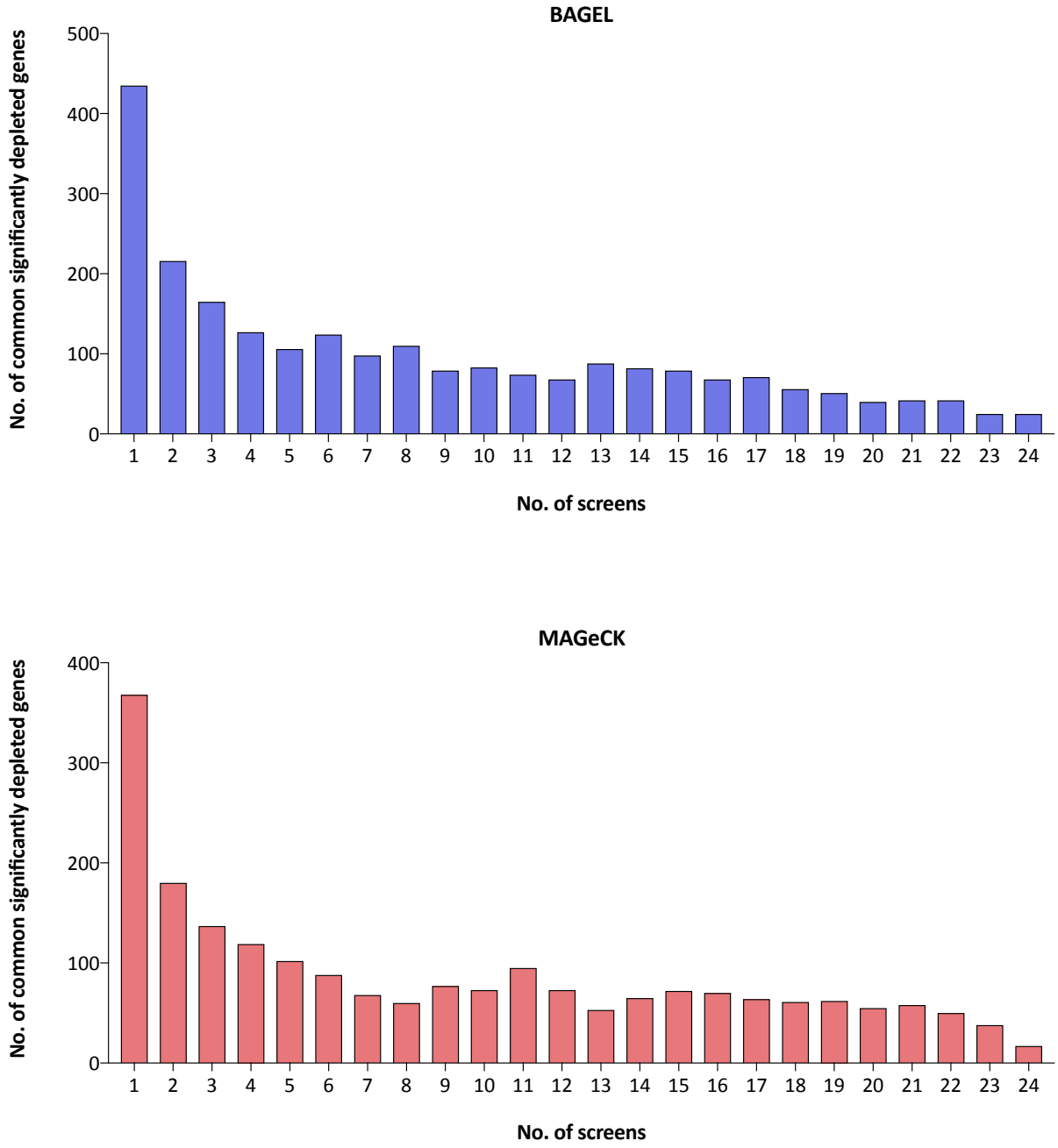

Figure 3.10. Frequency of significantly depleted genes across all iPSC screens. Across all screens, a total of 2371 genes were identified as significantly depleted by BAGEL, and 2105 by MAGeCK. Some of these hits were specific to one screen, but many were identified in multiple screens. These plots show the frequency with which genes were identified by one or more screens.

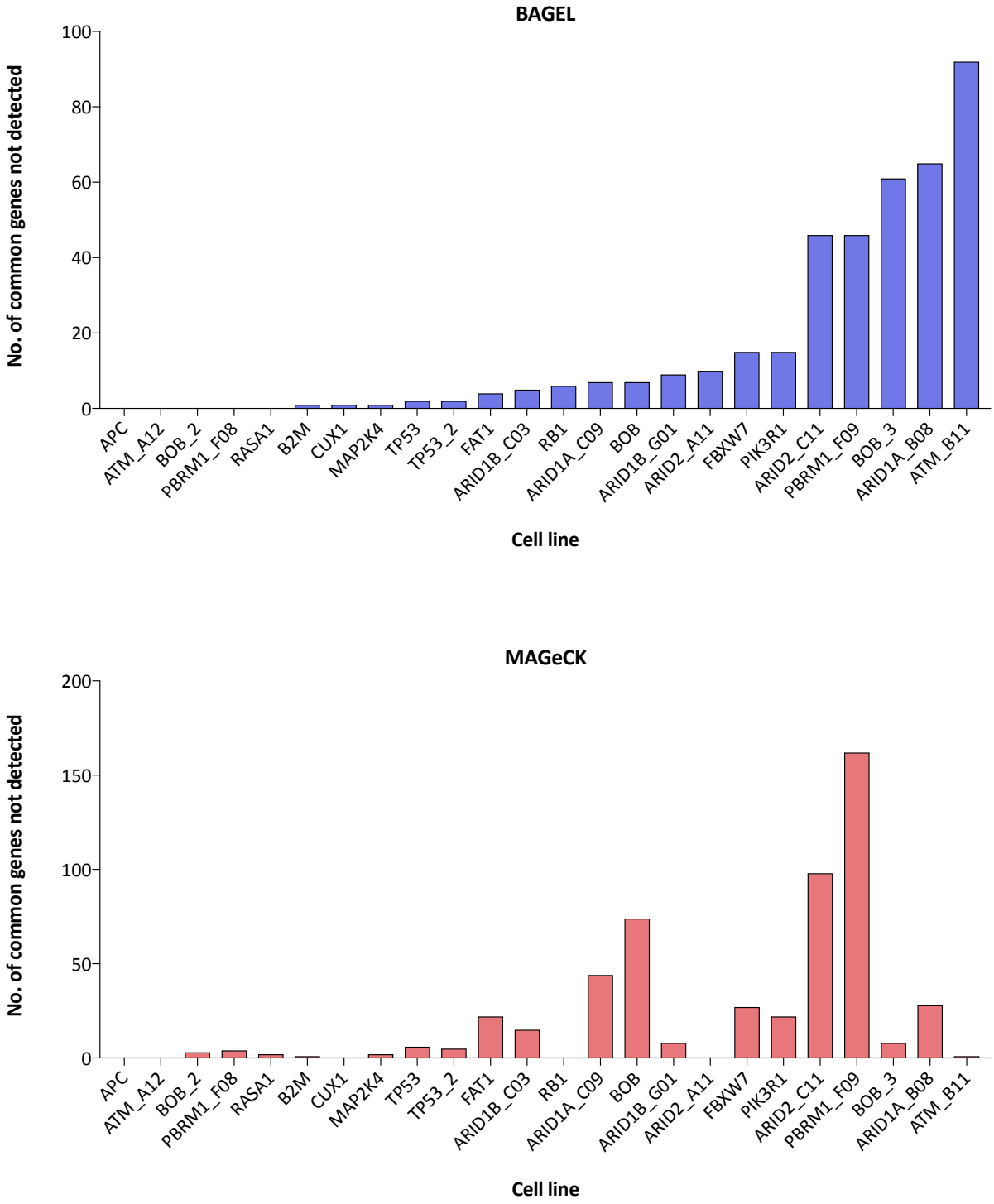

Figure 3.11. Detection of common gene hits in iPSC screens. Considering only genes that were significantly depleted in 20-23 of the iPSC screens, this plot shows the number of these genes that were not detected by each screen. Results are shown for both BAGEL and MAGeCK outputs.

3.7 Filtering for KO-specific dependencies

Despite the variability in the data, the results were not completely inconsistent and there were indications that known essentials could be detected. Therefore, we decided to continue using these data for our primary aim of identifying SLIs. To do so, we were interested in finding genes that were specifically essential in the TSG KO lines but not in the parental line. To identify KO-specific hits, we compiled lists of genes that were significantly depleted in each KO line and removed any genes that were also significant in the parental. For simplicity, I will discuss filtering using only the BAGEL outputs, but the same could be performed on the MAGeCK outputs or the overlap of both. As data was obtained for 3 biological replicates of the parental, various strategies were possible. One approach was to remove genes that were hits in every parental screen, ensuring that only high confidence hits in the parental were discarded. Another option was to exclude all genes that came up in any of the parental screens, accounting for the fact that detection of some genes may have been affected by some replicates performing poorly. A final strategy was to filter based only on the hits from BOB_2, which appeared to be the highest quality screen. Table 3.3 indicates the number of KO-specific genes identified in each screen using all of these filtering approaches. These gene lists are provided in Appendix A.9.

The screen which had the highest recall of established core fitness genes was ‘TP53’, closely followed by the biological replicate ‘TP53_2’ (Table 3.1). With this in mind, I have selected these screens to provide an example of the KO-specific gene lists. Table 3.4 shows the scaled BFs for the 20 top-ranking genes in the ‘TP53’ screen, excluding genes that were hits in any of the parental BOB screens and removing established core fitness genes. Of these, 19/20 genes were also significantly depleted in the ‘TP53_2’ replicate screen. It has been previously shown that one of these genes, *ATR,* is synthetically lethal with *TP53* (as reviewed by Qiu *et al.*, 2018).[257] This warrants further validation of the other genes, particularly those with a higher ranking, to identify novel SLIs with *TP53*.

In Chapter 4 I will discuss more advanced filtering of results from screens in the PBAF/BAF gene KO lines, and subsequent experimental validation of these genes.

Table 3.3. Number of KO-specific screen hits. The scaled BFs computed by BAGELR analysis of all screens were used to identify significantly depleted genes (scaled BF > 0). Genes that were significantly depleted in all BOB screens OR in at least one BOB screen OR in the BOB_2 screen, were removed from the list of significant hits in each KO line screen. The number of remaining genes are shown for each screen, based on each filtering strategy.

Screen	Not in every BOB screen	Not in any BOB screen	Not in BOB_2
TP53	1536	453	496
TP53_2	1346	329	367
ARID1A_C09	560	73	87
ARID1A_B08	86	1	5
ARID1B_C03	848	126	150
ARID1B_G01	528	53	72
ARID2_A11	238	5	9
ARID2_C11	262	31	42
PBRM1_F09	448	153	168
PBRM1_F08	781	90	107
FAT1	715	106	126
APC	741	60	77
FBXW7	436	66	81
ATM_A12	872	93	120
ATM_B11	41	0	1
MAP2K4	763	86	111
PIK3R1	330	27	38
RB1	564	32	43
CUX1	1260	272	308
RASA1	1121	208	239
B2M	1384	327	366

Table 3.4. Candidate synthetic lethal partners of *TP53*. The scaled BFs obtained by BAGELR analysis of the first *TP53* KO line screen were ranked from highest to lowest. The top 20 genes are shown, with scaled BFs noted for both biological replicates of this line.

Gene	TP53	TP53_2
SBNO1	2.12	0.51
HIST2H3A	2.11	2.62
SNAP23	2.06	1.57
HSD17B7	1.95	1.40
RINT1	1.90	1.53
ALDOA	1.88	1.25
HIRA	1.88	0.68
MED14	1.84	1.90
DR1	1.74	0.56
SOX2	1.73	0.54
MRPS12	1.71	1.11
ATR	1.67	0.36
ALG10	1.65	-0.47
RPP21	1.65	0.90
MRPL23	1.62	0.34
PRIM1	1.58	0.87
HNRNPA1	1.54	0.36
PRR13	1.53	1.67

3.8 Gene enrichment in iPSC screens

As described previously, MAGeCK can also be applied to identify significantly enriched genes. Considering that we observed an enrichment of NTC gRNAs in our initial data analysis (Section 3.2.2), we were interested to see whether any targeted genes were also enriched. Eleven genes were recurrently significantly enriched in at least 50% of the screens (Fig. 3.12a) (all MAGeCK enrichment values are provided in Appendix A.10). Five of these encoded for proteins that are involved in activation of apoptotic signalling in response to DNA damage, including TP53. Thus, it is logical that knockout of these genes would provide a proliferative advantage by preventing an apoptotic response to Cas9-induced DSBs.

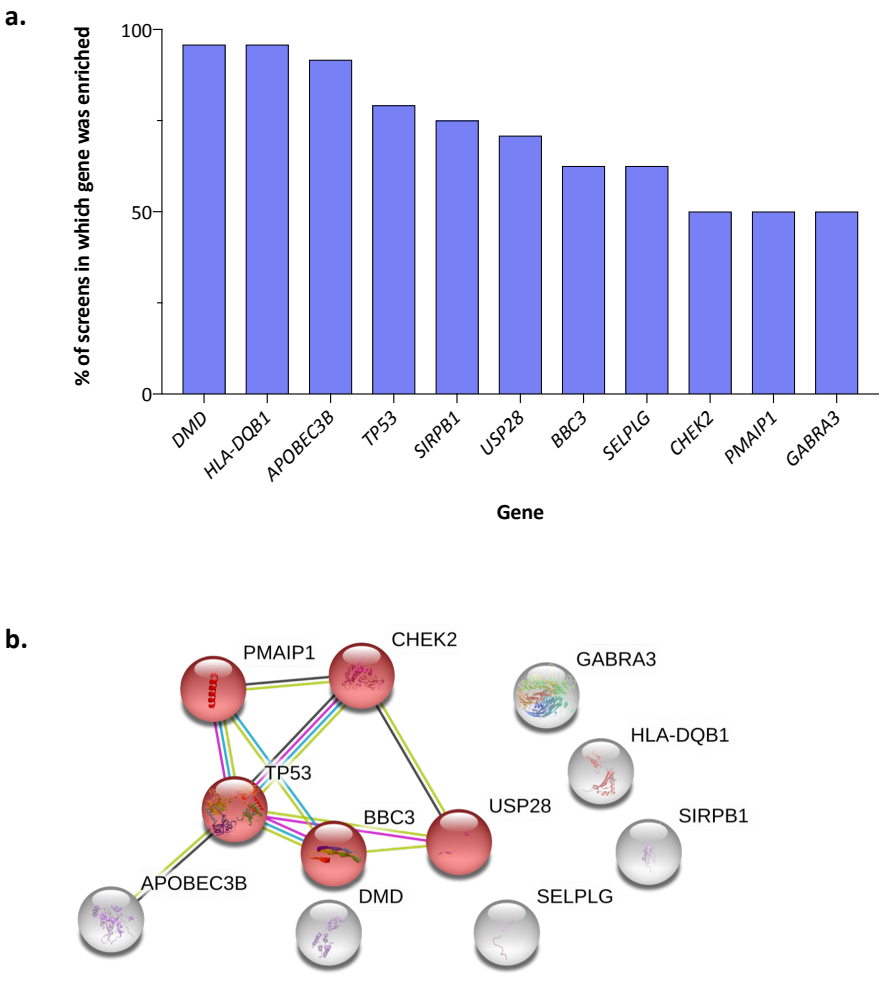

Figure 3.12. Enriched genes in iPSC screens. a) MAGeCK was applied for all iPSC screens and enrichment scores (positive FDR values) were computed for each gene. Genes with a positive FDR value < 0.1 were compared across all screens. Eleven genes were significantly enriched in at least 50% of the screens, **b)** The proteins encoded by these genes were analysed using STRING[258] to identify any interactions. Lines between the nodes indicate known/predicted interactions between proteins. Nodes highlighted in red are involved in apoptotic signalling pathways.

3.9 Summary

We performed CRISPR/Cas9 KO screens in 21 iPSC lines, with 3 biological replicates of the parental BOB line, 2 biological replicates of the *TP53* KO line and a single screen in 19 other KO lines. Due to an unexpected enrichment of non-targeting controls, which we attributed to Cas9 toxicity in the presence of targeting gRNAs, we had to remove these controls from our data. Initial quality control and screen performance tests produced results similar to published screens in cancer cell lines. However, further analysis indicated that the iPSC screens were highly variable and this made it difficult to confidently deduce which genes were essential for cell fitness. In some screens there was high recall of previously established core fitness genes, indicating that true positives could be identified but there was evidently a high risk of false negatives. Our aim was to identify genes that were specifically essential in the KO lines, and hence could be potential synthetic lethal partners. This variability made it challenging as there was a high possibility that genes identified in the KOs may be universally essential but were missed due to screen performance in the parental. Equally, low performance in the KO line screens may have led to false negatives and missed interactions. Despite this, we computed lists of KO-specific genes identified in each screen and hence have identified candidate SLIs. Results from the *TP53* KO line screen included a known SLI, which provides more confidence to the findings. However, validation is critical for any conclusions to be drawn from these datasets. As an aside, we also identified genes that were recurrently enriched in the screens. These may be informative with regards to iPSC biology and more specifically, their response to the CRISPR/Cas9 screening process.