Chapter 3

Identifying mediators of malignant transformation in cancer using genome-wide CRISPR-Cas9 knockout screening

3.1 Introduction

Malignant transformation is the transition of a cell from a normal proliferative phenotype to an abnormal malignant state, where the cell has the potential to form a tumour *in vivo*. This involves overcoming normal growth controls and the dysregulation of the proliferative homeostasis that usually maintains a constant cell number and constrains cells to their normal location within a tissue.

This transition may occur through the mutation of genes involved in these growth control processes. These genes may be directly involved in regulation of cell division, or influence its control indirectly via other cellular processes (Hanahan and Weinberg, 2011). Known genes that are able to induce malignant transformation include both oncogenes and tumour suppressor genes. For example, the *RAS* group of genes are mutated by amplification or activating point mutation in a range of human cancers, and are also able to transform cells *in vitro* through the over-activation of signal transduction processes that result in changes in proliferation and differentiation (Yamamoto et al., 1999). Mutation of tumour suppressor genes is also able to induce transformation, through the removal of repression of proliferation. For example, *NF1* functions as a GTPase activating protein, inhibiting the activity of the Ras protein. Homozygous loss-of-function mutations of this gene are therefore able to induce

transformation through the same downstream mechanisms as *RAS* activation (Cichowski and Jacks, 2001).

Historically, these genes have been identified on an individual basis using cases where malignant transformation occurs due to naturally occuring genetic alterations. Some of the first oncogenes discovered were identified due to the presence of their homologues in the genomes of viruses that cause malignant transformation. For example, *RAS* genes were first described in 1982, resulting from research based on Harvey sarcoma virus and Kirsten sarcoma virus, which can cause sarcomas in rodents due to retroviral integration of a *RAS* homologue into the host genome (Malumbres and Barbacid, 2003). Early tumour suppressor genes were often identified through the study of familial cancer syndromes, where heterozygous germline mutations of these genes predispose individuals to the development of certain cancers. An example of this is Neurofibromatosis Type 1, a condition causing a range of nervous system tumours due to heterozygous loss of function mutations of the tumour suppressor gene *NF1* in the germline (Gutmann et al., 2017).

The advent of next-generation sequencing has led to a dramatic increase in the amount of information generated from human tumour genomes in recent years. It is now possible to identify genes that may be involved in malignant transformation by sequencing large numbers of tumours from cancer patients and analysing the somatic mutations present. The identification of genes that are frequently mutated in human tumours can indicate their involvement in tumour biology, but this alone is not able to show what role they play in cancer development. In order to isolate genes that are involved in the earliest stage of oncogenesis, a functional assay for malignant transformation is needed.

In the past, the generation of defined genetic alterations at a genome-wide scale for functional screening has been technically challenging. The development of CRISPR-Cas9 genome editing techniques have made genome-wide screening for a range of phenotypes possible, by generating mutations at the desired locations using libraries of guide RNAs (gRNAs) complementary to the regions to be altered. This chapter describes the use of a genome-wide CRISPR-Cas9 knockout screen to identify genes that can induce malignant transformation when subjected to loss-of-function mutations.

The model of transformation used in this screen was the cell line NIH3T3-Cas9, the genetic background of which is discussed in chapter two. NIH3T3 cells are an immortalised but untransformed mouse embryonic fibroblast cell line that is sensitive to maligant transformation *in vitro* (Jainchill et al., 1969). The transformation-sensitive nature of this cell line facilitates its use in a functional assay for this phenotype, as the background rate of transformation in cells that are not genetically altered is low. The assay used in this screen is the focus formation assay, where transformation is measured through the formation of clonal foci of proliferation

in cultured NIH3T3 cells. When transforming mutations are introduced, the number of these foci increases (see section 3.3.1), providing a phenotypic readout for the screen.

The principle of the screen was to compare the gRNAs that are enriched in cells that have been allowed to form these transformed foci, compared with cells that have been split regularly and therefore proliferated without focus formation, and the original gRNA library. An overview of the screen can be found in figure 3.1. The overproliferation of cells in which transforming tumour suppressor genes have been knocked out leads to overrepresentation of gRNAs against these genes in the final gRNA population. These genes are then identified by targeted sequencing of the gRNA sequences present in the cells, and analysis of the read counts using the algorithm Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout (MAGeCK) (Li et al., 2014). This approach was used to identify putative genes that can cause malignant transformation *in vitro* alone when knocked out, followed by attempting to validate these candidates individually.

3.1.1 Aims

Overall aim: To identify putative tumour suppressor genes that are involved in malignant transformation in human cancer.

1. To identify genes that may mediate transformation *in vitro* using genome-wide CRISPR-Cas9 knockout screening in NIH3T3-Cas9.

2. To prioritise hits from the CRISPR-Cas9 screen using mutation data from existing human cancer sequencing projects.

3. To functionally validate prioritised hits for transforming potential *in vitro.*

3.2 Materials and Methods

3.2.1 Materials

Cell lines

HEK293T HEK293T cells were obtained from Dr. Eugenio Montini at the San Raffaele Telethon Institute for Gene Therapy.

NIH3T3 NIH3T3 wild-type cells were obtained from the American Tissue Culture Collection (ATCC® CRL-1658™).

NIH3T3-Cas9 NIH3T3-Cas9 cells were generated by Dr. Nicola Thomspon from the experimental cancer genetics group at the Wellcome Sanger Institute (see Appendix B.1).

Plasmids

pmaxGFP (Lonza, catalogue #VDF-1012)

psPAX2 This plasmid was a gift from Dr. Didier Trono (Addgene, plasmid #12260).

pMD2.G This plasmid was a gift from Dr. Didier Trono (Addgene, plasmid #12259).

pAdVAntage™ Vector (Promega, catalogue #E1711).

pBabe-puro Ras-V12 This plasmid was a gift from Professor Bob Weinberg (Addgene, plasmid #1768).

pCMV-hyPBase This plasmid was obtained from Dr. Kosuke Yusa at the Wellcome Sanger Institute (Yusa et al., 2011).

Genome-wide Knockout CRISPR Library v2 This library was a gift from Dr. Kosuke Yusa (Addgene #67988, (Koike-Yusa et al., 2013)).

Genome-wide mouse sgRNA lentiviral-PiggyBac library This library was a gift from Dr. Emmanouil Metzakopian (Metzakopian et al., 2017).

Reagents

Reagent	Manufacturer
Agencourt AMPure XP SPRI beads	Beckman Coulter
Blasticidin (10mg/mL)	InvivoGen
Blood and Cell Culture DNA Maxi Kit	Qiagen
Crystal violet solution (1%, aqueous)	Sigma-Aldrich
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma-Aldrich
Ethanol absolute ($\geq$99.8%, AnalaR NORMAPUR)	VWR International
Fetal bovine serum (FBS)	Gibco
Gelatin solution (2%, aqueous)	Sigma-Aldrich
Lipofectamine 3000 kit (Lipofectamine 3000 reagent and P3000)	Thermofisher Scientific
KAPA HiFi HotStart ReadyMix 2X	Kapa Biosystems
Methanol ($\geq$99.8%, AnalaR NORMAPUR)	VWR International
Nuclease-free water	Sigma-Aldrich
Opti-MEM™ reduced serum media	Gibco
Penicillin, streptomycin and L-glutamine (100X, 50mg/mL)	Gibco
Phosphate-buffered saline (PBS)	Sigma-Aldrich
Polybrene ($\geq$95%)	Sigma-Aldrich
Puromycin (10mg/mL)	InvivoGen
Q5 Hot Start High-Fidelity 2X Master Mix	New England Biolabs
QIAquick PCR Purification Kit	Qiagen
Trypsin-EDTA (0.05%)	Gibco

Table 3.1: **Reagents used in the methods described in Chapter 3**

3.2.2 Methods

Focus formation assay

Transfection NIH3T3 wild-type cells were seeded at a density of 100,000 cells/well in a 6-well plate (50,000 cells/mL) in complete DMEM (DMEM supplemented with 10% FBS and 500µg/mL penicillin, streptomycin and L-glutamine), and incubated at 37°C for 24 hours. The media was changed to Opti-MEM™ reduced serum media before transfection. Cells were transfected using Lipofectamine 3000 according to the manufacturer's instructions, using the quantities of reagents listed in table 3.2. For mock transfection, the plasmid DNA was replaced with an equivalent volume of Opti-MEM™. After 16 hours the media was changed

to complete DMEM. Cells were cultured for 12 days without splitting to allow formation of foci of proliferation, changing the media every 3-4 days.

Reagent	Amount per well (100,000 cells)
Lipofectamine 3000 Reagent	1.5μL
P3000	1μL
pBabe-puro Ras V12 or pmaxGFP	0.5μg

Table 3.2: **Transfection reagent quantities for transfection with Ras and GFP plasmids**

Fixation and staining Wells were washed with 4˚C PBS and fixed for 1 hour using methanol. Cells were stained with 1% aqueous crystal violet for 10 seconds, washed with MilliQ water and air-dried.

Genome-wide Knockout CRISPR Library v2 amplification

The Genome-wide Knockout CRISPR Library v2 was amplified according to the depositor's instructions available on the product page (Addgene #67988). See Appendix B.4 for verification of the amplified library.

Genome-wide Knockout CRISPR Library v2 sequencing

The amplified Genome-wide Knockout CRISPR Library v2 was sequenced using Illumina-C HiSeq 2500 single-end sequencing at the Wellcome Sanger Institute. Sequencing libraries were prepared from the plasmid using the protocol detailed in section 3.2.2: Library preparation.

Genome-wide Knockout CRISPR Library v2 lentivirus production

Ten T150 cell culture flasks were coated with 0.1% gelatin in PBS. 2.1 x 10^8 HEK293T cells were seeded at a density of 1 x 10^6 cells/mL and incubated for 24 hours in complete DMEM. Before transfection the media was changed to Opti-MEM™ reduced serum media. Cells were transfected using Lipofectamine 3000 according to the manufacturer's instructions using the following quantities of reagents per flask (table 3.3). At 16 hours post-transfection, the media was changed to heat-inactivated complete DMEM. 72 hours post-transfection, the supernatant was filtered through a 45μm low-protein binding filter, and frozen at -80˚C.

Reagent	Amount per flask (2.1 x 10^7 cells)
Lipofectamine 3000 Reagent	120µL
P3000	105µL
pMD2.G	11.2µg
psPAX2	16.8µg
pAdVAntage™ Vector	16.8µg
Genome-wide Knockout CRISPR Library v2	29.4µg

Table 3.3: **Transfection reagent quantities for Genome-wide Knockout CRISPR Library v2 lentivirus production**

Whole genome CRISPR-Cas9 knockout screen

Infection: NIH3T3-Cas9 cells were cultured for 7 days in complete DMEM containing 5µg/mL blasticidin to select for expression of the Cas9 transgene (see appendix B.1 for details of selection marker). Cells were suspended in 536mL of heat-inactivated complete DMEM, containing 536µL of polybrene. Cells were infected with the lentivirus described in section 3.2.2 at a multiplicity of infection of 0.3 plaque forming units/cell. This mixture was split between 16 T150 flasks per replicate. For each of three replicates, 2.7 x 10^7 cells were infected, giving a mean 300X coverage per gene.

Days 1-14 With day 0 as the day of infection, the following protocol was followed.

Day 1: The media in the flasks was changed to 30mL of fresh heat-inactivated complete DMEM per flask.

Day 3: The media in the flasks was changed to 30mL of complete DMEM, containing 2µg/mL puromycin to select for infected cells.

Day 5: Repeat of day 3 protocol.

Day 7: Cells were split by detaching with 0.05% trypsin-EDTA. 2.7 x 10^7 million cells per replicate were seeded in four five-layer Falcon Cell Culture Multi-Flasks in 150mL complete DMEM containing 2µg/mL puromycin per flask.

Day 11: Repeat of day 7 protocol.

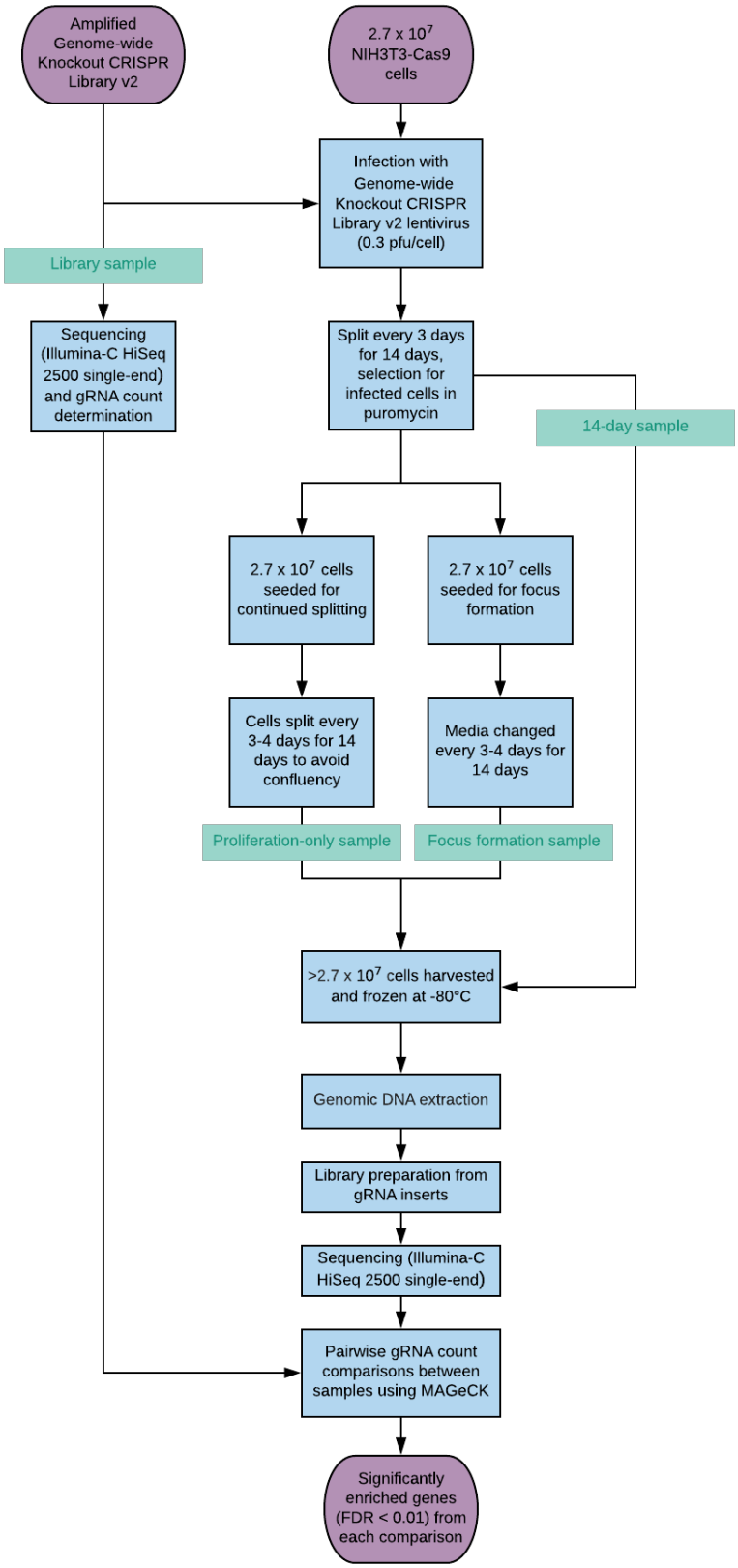


Figure 3.1: **Schematic of genome-wide CRISPR-Cas9 knockout screen using NIH3T3-Cas9 cells**

Day 14: Cells were divided into two arms of the screen (see days 15-27). The remaining cells ($\geq$ 2.7 x 10^7 per replicate) were harvested using 0.05% trypsin-EDTA, washed with PBS and centrifuged (200xg, 5 minutes). Pellets were frozen at -80°C.

Days 15-27

Arm A 2.7 x 10^7 cells per replicate were seeded in three five-layer Falcon Cell Culture Multi-Flasks in 150mL complete DMEM per flask. For the following 14 days these cells were split every 3-4 days according to the Day 7 protocol.

Arm B 2.7 x 10^7 cells per replicate were seeded in ten T150 flasks in 30mL complete DMEM. For the next 14 days these cells were not split, allowing the cells to form foci of proliferation. The media in these flasks was changed every 3-4 days.

Control One T150 flask was seeded at the same density with uninfected NIH3T3 Cas9 cells.

Day 28: Cells from arm A and arm B ($\geq$ 2.7 x 10^7 per replicate) were harvested using 0.05% trypsin-EDTA, washed with PBS and centrifuged (200xg, 5 minutes). Pellets were frozen at -80°C.

DNA extraction

Genomic DNA was extracted using the Qiagen Blood and Cell Culture DNA Maxi Kit according to the manufacturer's instructions.

Library preparation

Library preparation was carried out with the assistance of the Cancer Genome Project at the Wellcome Sanger Institute.

First round polymerase chain reaction (PCR) 36 technical replicates per sample were set up using the reagent quantities listed in table 3.4. The gRNA sequences inserted into the genomic DNA were amplified using the programme detailed in table 3.5.

Reagent	Quantity per reaction
Genomic DNA (section 3.2.2)	2µg
Nuclease-free water	24µL - (DNA volume)
Q5 Hot Start High-Fidelity 2X Master Mix	25µL
Primer mix (10µM each)	1µL

Table 3.4: **Reagent quantities for the 1st round PCR in the CRISPR-Cas9 gRNA insert library preparation**

Primer sequences can be found in appendix B.5.

Cycle number	Denaturing	Annealing	Extension
1	98°C, 30 seconds		
2-29	98°C, 10 seconds	61°C, 15 seconds	72°C, 20 seconds
30			72°C, 2 minutes

Table 3.5: **PCR programme for the 1st round PCR in the CRISPR-Cas9 gRNA insert library preparation**

PCR purification For each sample, 5µL of PCR product was taken from each of the 36 replicates and pooled. The products were then purified using QIAquick PCR Purification Kit according to the manufacturer's instructions.

Second round PCR The PCR product was diluted to 40pg/µL in nuclease-free water. For each sample, reactions were prepared as in table 3.6. The DNA was amplified using the programme detailed in table 3.7, adding sequencing adaptors.

Reagent	Quantity per reaction
1st round PCR product (40pg/µL dilution)	5µL (200pg)
Primer mix (5µM each)	2µL
Nuclease-free water	18µL
KAPA HiFi HotStart ReadyMix 2X	25µL

Table 3.6: **Reagent quantities for the 2nd round PCR in the CRISPR-Cas9 gRNA insert library preparation**

Primer sequences can be found in appendix B.5.

Cycle number	Denaturing	Annealing	Extension
1	98°C, 30 seconds		
2-9	98°C, 10 seconds	66°C, 15 seconds	72°C, 20 seconds
10			72°C, 5 minutes

Table 3.7: **PCR programme for the 2nd round PCR in the CRISPR-Cas9 gRNA insert library preparation**

Library purification Each 50µL PCR product was purified using 40µL of Agencourt AMPure XP SPRI beads according to the manufacturer's instructions.

Sequencing

Illumina-C HiSeq 2500 single-end sequencing was performed at the Wellcome Sanger Institute. The mean number of reads per replicate was 25,822,355 and read length was 20bp.

Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout (MAGeCK)

The gRNA read counts generated were analysed using the algorithm MAGeCK (Li et al., 2014) with the assistance of Dr Vivek Iyer from the experimental cancer genetics team at the Wellcome Sanger Institute. Pairwise comparisons between the different samples were conducted. Initially, the 14-day sample (test) was compared to the plasmid library (control), for use in the generation of a receiver-operating characteristic curve to assess the screen quality. The focus formation sample (test) was then compared with each of the other three samples - "library", "14-day" and "proliferation-only" (control). This was done using the test command from the MAGeCK package.

Receiver-operating characteristic curve generation

A receiver-operating characteristic (ROC) curve was generated using the data collected from the 14-day - library MAGeCK comparison. Genes that were significantly depleted (FDR < 0.01) in this comparison were compared to the list of essential genes used by the Bayesian Analysis of Gene Essentiality (BAGEL) algorithm (Hart and Moffat, 2016) to determine the relationship between the sensitivity and specificity of the screen in identifying known essential genes. The ROC curve was generated using the roc command from the pROC package (Robin et al., 2011) (partial.auc=c(100,90), partial.auc.correct=TRUE, partial.auc.focus="sens", boot.n=100).

Gene prioritisation

Genes significantly enriched in the focus formation sample when compared to any other sample in the screen (FDR < 0.01) were considered for validation. Genes were prioritised by comparison with existing cancer genome data taken from the Cancer Gene Census (Futreal et al., 2004), Intogen Cancer Drivers Database (Rubio-Perez et al., 2015), positively selected driver mutations (Martincorena et al., 2017), recurrently deleted intervals (Iorio et al., 2016), homozygously deleted regions (Cheng et al., 2017), and The Cancer Genome Atlas (Weinstein et al., 2013). The rationale for inclusion of the chosen genes in the validation is detailed in the results (section 3.3.5).

Validation (arrayed focus formation assay)

Plasmids carrying gRNA sequences against the genes in table 3.12 (with the exception of *Lats2* and *Rnf146*), along with those against 10 randomly selected genes as a negative control, were obtained from an arrayed mouse gRNA library (Metzakopian et al., 2017). Two gRNAs sequences were used per gene to help ensure successful knockout (table 3.8).

The validation was performed using a small scale focus formation assay. For each gRNA and for the mock transfection, 100,000 cells/well were seeded in 2 wells of a 6-well tissue culture plate at a density of 50,000 cells/mL and incubated for 24 hours in complete DMEM. Before transfection. the media was changed to Opti-MEM™ reduced serum media. The cells were transfected using Lipofectamine 3000 according to the manufacturer's instructions using the following quantities of reagents (table 3.9). For the mock transfection, the plasmid DNA was replaced with an equivalent volume of Opti-MEM™. After 16 hours the media was changed to complete DMEM. Cells were cultured for 12 days without splitting to allow formation of foci of proliferation, changing the media every 3-4 days.

Reagent	Amount per 100,000 cells
Lipofectamine 3000	1.5μL
P3000	1μL
gRNA plasmid DNA	500ng
pCMV-hyPBase	50ng

Table 3.9: **Transfection reagent quantities for validation focus formation assays**

Gene	gRNA 1 sequence (complementary)	gRNA 2 sequence (complementary)
Mical1	CTCAGCAGGCACTGCTTCTTGG	GCTGTCATCACAAAGTAGTGGG
Cyp2j11	ACGTAATTAGCGTGAATTTTGG	TGTTGCCTTGCAGCTAAACTGG
Lgalsl	AGTCGTGGGAAGTACACGTCGG	GTTCAATTGCCACATCGGCAGG
Slain1	CAGCACAGAGTTCACAGCGTGG	GCGGCATGCCTTTATCCAATGG
Fbrs	AGCTGGTGGGAGACCCGGAGGG	TCAGCACTGGCCCCAGTCGTGG
Zfp418	CAGCATCACATCAAGATACAGG	GTGGCAGTTTACTTCTCCCAGG
Sparc	GGTGCAGAGGAAACGGTCGAGG	GAGGAAACGGTCGAGGAGGTGG
Cyp2a22	GCTTTGGAGGACAACGCTGAGG	GCCGGGTTGTGGTGCTATATGG
Tmem160	CTTCTGTCATCCGGCATTGGGG	GACTTCTGTCATCCGGCATTGG
Top3b	TCAAGATGACGTCTGTCTGCGG	AAGTACAACAAGTGGGATAAGG
Nup160	GCAAGTGCCGCGTTTGGAACGG	TGAAGTACAGTGAGAGCGCTGG
Smu1	GACAGCATTGAAAGTTTCGTGG	CAAGTACTGCATGATTAGTCGG
Nf1	CTCTCTCAGTTGATCATATTGG	TTGATCATATTGGATACACTGG
Ptbp1	CACGTGGAGAAGAGCTCGTCGG	CTGTAAACTCCGTCCAGTCTGG
Kdsr	CTATTGAGTGCTACAAACAAGG	TCTCAAGACTATAACCAAGTGG
Mcat	GGAGAAGTTGGACTGACGCTGG	ATCCCACTGGGAACGGCTTCGG
Cdk7	AATAAATAGAACAGCCTTAAGG	GCTCCCAAATGATTTGGCCAGG
Mak16	AATCGGTCGTCCTGTCCTCTGG	TCTGACTGGTCTGTGCAATCGG

Table 3.8: **DNA sequences encoding complementary gRNAs from the arrayed plasmid library used in validation**
DNA sequences encoding complementary gRNAs used in the validation of the genome-wide CRISPR-Cas9 knockout screen discussed in section 3.2.2. Plasmids carrying these sequences were obtained from an arrayed mouse gRNA library (Metzakopian et al., 2017).

Validation - pooled gRNA lentivirus

An alternative method of validation was carried out using a lentivirus pool carrying the 36 gRNA sequences listed in table 3.8.

Validation virus production A T25 culture flask was coated with 0.1% gelatin in PBS. 3.5 x 10^6 HEK293T cells were seeded at a density of 700,000 cells/mL and incubated for 24 hours in complete DMEM. Before transfection, the media was changed to Opti-MEM™ reduced serum media. The cells were transfected using Lipofectamine 3000 according to the manufacturer's instructions using the following quantities of reagents per flask (table 3.10). At 16 hours post-transfection, the media was changed to heat-inactivated complete DMEM. 72 hours post-transfection, the supernatant was filtered through a 45μm low-protein binding filter and frozen at -80˚C.

Reagent	Amount per flask (3.5 x 10^6 cells)
Lipofectamine 3000 reagent	2μL
P3000	1.75μL
pMD2.G	187ng
psPAX2	280ng
pAdVAntage™ Vector	280ng
gRNA plasmid DNA (for each gRNA listed in table 3.8)	86.1ng

Table 3.10: **Transfection reagent quantities for validation virus production**

Infection with pooled gRNA lentivirus and focus formation assay:

Day 1: 830,000 NIH3T3-Cas9 cells were suspended in 20mL of heat-inactivated complete DMEM, containing 20μL of polybrene. The virus described in section 3.2.2 was added at a multiplicity of infection of 0.3 plaque forming units/cell. This mixture was seeded in a T75 culture flask. 830,000 cells infected with a virus pool carrying 36 different gRNAs at a multiplicity of infection of 0.3 gives a mean coverage per targeted gene of 6917X.

Day 2: After 16 hours, the media was changed to complete DMEM without polybrene.

Day 4: Media was changed to complete DMEM containing 2μg/mL puromycin to select for infected cells.

Day 8: Cells were split using 0.05% trypsin-EDTA and 830,000 were re-seeded in a T75 culture flask.

Days 9-19: Cells were cultured without splitting to allow for the formation of foci of proliferation. Media was changed every 3-4 days.

Day 20: Cells were harvested using 0.05% trypsin-EDTA, washed with PBS and centrifuged (200*xg*, 5 minutes). Pellets were frozen at -80°C.

Genomic DNA extraction: Genomic DNA was extracted using Qiagen Gentra Puregene kit according to the manufacturer's instructions.

Library preparation Library preparation and purification methods were the same as those for the genome-wide CRISPR-Cas9 knockout screen (see section 3.2.2), with the exception of the use of different primers for the first round PCR (see appendix B.5).

Sequencing Illumina-C HiSeq 2500 single-end sequencing of the above library is currently in progress.

3.3 Results

3.3.1 Focus formation assay with pBabe-puro Ras-V12

The focus formation assay used in chapters three and four uses transformation-sensitive NIH3T3 cells to detect genetic changes that induce malignant transformation *in vitro.* In order to validate this assay, cells were transfected with a plasmid expressing the known transforming oncogene *H-RAS* (Addgene #1768). When compared with the mock transfected and control pmaxGFP transfected wells, the wells transfected with pBabe-puro Ras V12 developed many more foci of proliferation during the 12 day culture period, indicating the malignant transformation of individual cells due to the expression of *H-RAS,* and the formation of clonal foci that have overcome normal growth controls (see figure 3.2). For the mock and control GFP cells, few or no foci of proliferation were visible, indicating a low level of background transformation. These results suggest that this assay is a suitable means of detecting genetic changes that induce malignant transformation *in vitro.*

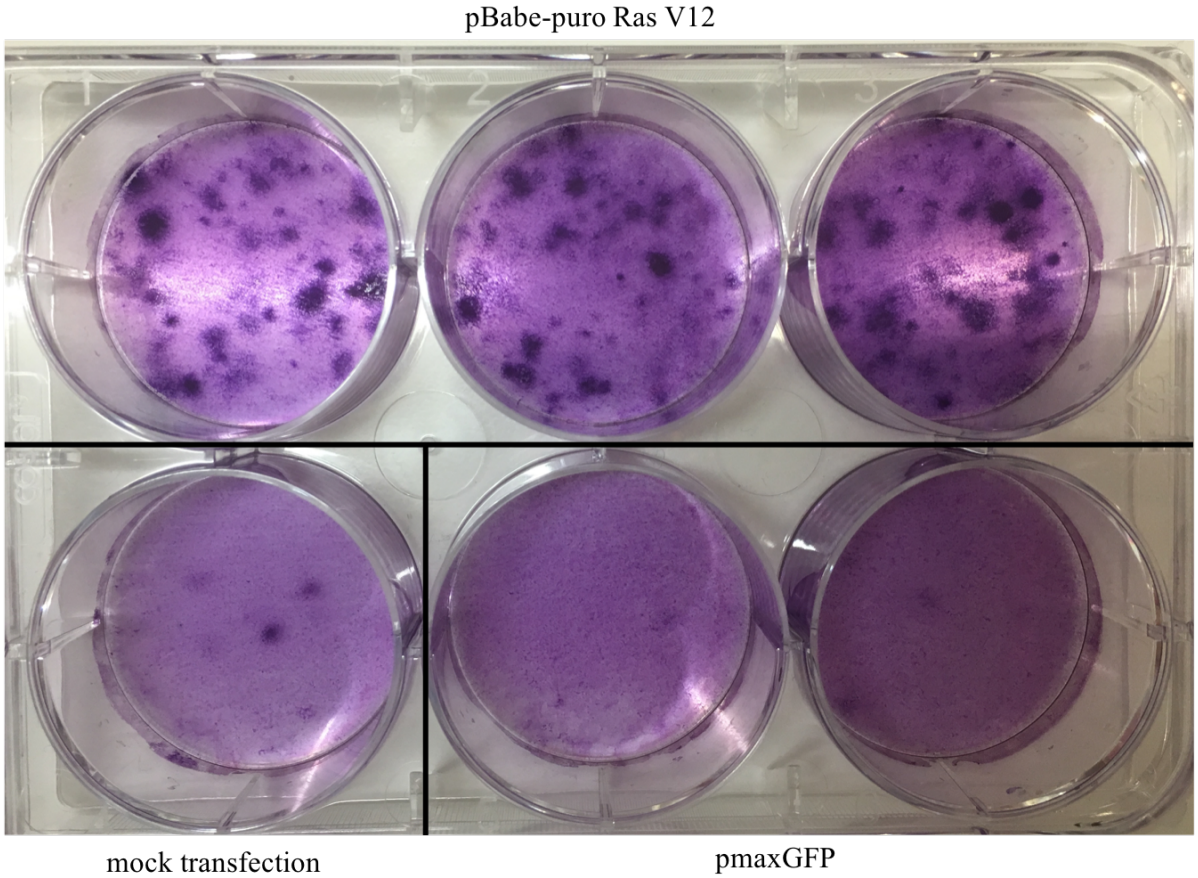


Figure 3.2: **Focus formation assay using NIH3T3 cells transfected with pBabe-puro Ras-V12**
NIH3T3 cells were transfected with pBabe-puro Ras-V12, pmaxGFP or mock transfected in order to compare the effects of *H-RAS* expression with control. After transfection, cells were cultured for 12 days before staining with crystal violet, as described in section 3.2.2.

3.3.2 Genome-wide CRISPR-Cas9 knockout screen for mediators of malignant transformation

The aim of this screen was to identify genes that can induce malignant transformation *in vitro* when subjected to loss-of-function mutation. Transformation-sensitive NIH3T3-Cas9 cells were infected with a lentivirus carrying Genome-wide Knockout CRISPR Library v2 (Koike-Yusa et al., 2013) and allowed to proliferate for 14-days, followed by splitting the cells into two arms of the screen. In the proliferation-only arm, cells were split every 3-4 days, preventing them from reaching confluency. In the focus formation arm, cells were not split, allowing them to form transformed foci of proliferation. In the latter arm, the foci of proliferation were visible macroscopically on day 28, whereas a control flask seeded with uninfected NIH3T3-Cas9 cells and cultured in parallel showed very few transformed foci. This indicates a low background rate of transformation during the screen, with the focus formation occuring due to the CRISPR-Cas9 mediated knockout of specific genes.

Comparison of the gRNA counts in cells at different stages of the screen (see figure 3.1 for an overview of the samples taken) was used to identify putative genes involved in the formation of the transformed foci of proliferation seen in the focus formation sample.

Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout (MAGeCK)

The gRNA read counts were analysed using the algorithm MAGeCK (Li et al., 2014). MAGeCK identifies genes where gRNAs against that gene were significantly enriched or depleted in one sample with respect to another sample. Initially, the 14-day sample was compared to the plasmid library to generate the data for a receiver-operating characteristic curve to assess the screen quality. The read counts from the focus formation sample were then compared to those from each of the other three samples (library, 14-day and proliferation-only). The comparisons between the focus formation sample and the library, and between the focus formation sample and the 14-day sample, are likely to identify any genes involved in either malignant transformation or the control of proliferation. However, the comparison between the focus formation sample and the proliferation-only sample was made in an attempt to identify genes specifically involved in the ability to form the clonal foci seen during the screen, which may indicate involvement in malignant transformation.

3.3.3 Screen quality

Receiver-operating characteristic (ROC) curve - ability to detect essential genes

The quality of the screen was assessed by comparing the genes determined by MAGeCK analysis to be significantly depleted (FDR < 0.01) in the 14-day sample compared to the gRNA library, with the list of essential genes used in the Bayesian Analysis of Gene Essentiality (BAGEL) algorithm (Hart and Moffat, 2016). A ROC curve was generated to measure the sensitivity and specificity of the detection of these genes. The partial area under the curve (coloured dark grey in figure 3.3) equalled 87.6%, indicating a good level of overall sensitivity and specificity. This suggests that the gRNA library has acheived knockout of genes across the genome, producing the expected phenotypes.

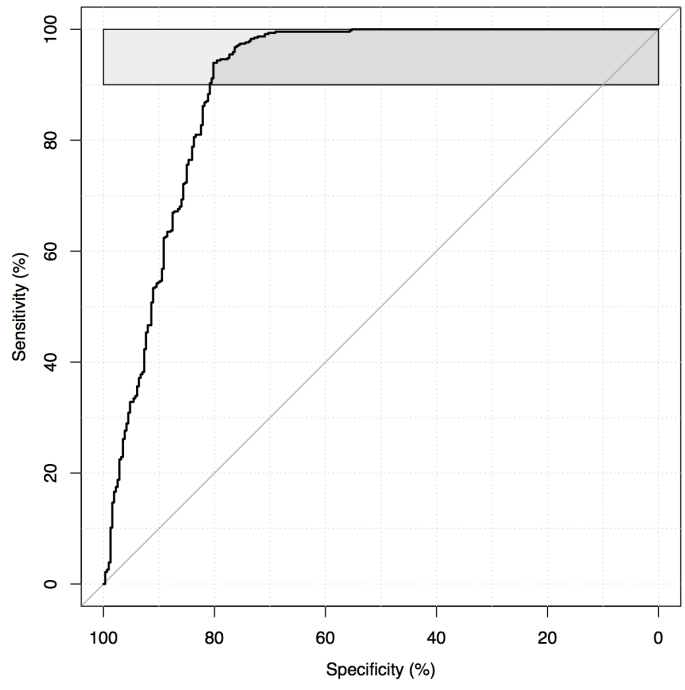


Figure 3.3: **Receiver Operating Characteristic curve based on the detection of BAGEL essential genes by the genome-wide CRISPR-Cas9 knockout screen**
This ROC curve is based on the ability of the 14-day - library MAGeCK comparison from the genome-wide CRISPR-Cas9 knockout screen (section 3.2.2) to detect the dropout of gRNAs against the essential genes used in the Bayesian Analysis of Gene Essentiality (BAGEL) algorithm (Hart and Moffat, 2016). The dark grey area indicates the partial area under the curve between 90-100% specificity.

3.3.4 Candidate genes

3.3.5 Prioritising genes for validation using existing cancer genome data

A shortlist of genes that are candidates for involvement in transformation was produced by taking genes significantly enriched (FDR < 0.01) in the focus formation sample, compared with the plasmid library, the 14-day sample, or the proliferation-only sample. This identified 50 potential hits, which are listed in appendix B.6. In order to further prioritise genes for individual validation, cancer genome data from a variety of studies was consulted to identify the strongest candidates based on existing mutation data for each gene.

Cancer Gene Census

The Cancer Gene Census (CGC) (Futreal et al., 2004) genes are a list of genes curated by COSMIC (Forbes et al., 2017), that have been shown to contain mutations that are causally implicated in human cancer. From the 77 genes identified by the screen, eight (*Gnas, Kdsr, Sufu, Nf2, Cltc, Ptch1, Nf1* and *Pten*) were found to be existing CGC genes. Some of these genes were ranked very highly in the MAGeCK analyses, for example *Gnas* was the most overrepresented gene in both the 14-day - focus formation and the proliferation-only - focus formation comparisons (see appendix B.6 for complete list of gene rankings). The appearance of these genes in the results is encouraging, as it suggests that the screen successfully identified genes that are causally involved in cancer. As these are well characterised cancer-linked genes, they were not investigated further. *Nf1* was taken forward to the validation as a positive control.

IntOGen

IntOGen is a publicly available database that identifies driver mutations in cancer from the analysis of point mutations from 4,623 cancer exomes at 13 tumour sites. Six genes (*Gnas, Nf2, Cltc, Ptch1, Nf1* and *Pten*) from the candidate list were found in IntOGen's Cancer Drivers Database (Rubio-Perez et al., 2015). However, these are all genes that had previously been identified in the Cancer Gene Census, so they were not taken forward into the validation stage.

Positively selected substitution mutations

In 2017 Martincorena *et al.* identified driver mutations in somatic tissues and cancer, looking at coding substitutions under positive selection. They found that >50% of the identified mutations were outside known cancer genes, making this a potentially promising source of data for

confirming novel hits in the CRISPR-Cas9 screen. The genes from the candidate gene list that were found to be positively selected in somatic tissues were *Lats2, Nf2, Nf1* and *Pten* (Martincorena et al., 2017). *Lats2* has been previously described as a potential tumour suppressor gene involved in inhibition of the G1/S phase transition (Li et al., 2003), but its role is not as well characterised as those listed in the Cancer Gene Census, so the gene was chosen for validation.

Deletion mutations

In addition to looking at substitution mutations it was important to consider deletions found in human cancers. If a gene on the list of hits from the CRISPR-Cas9 screen is in a recurrently deleted interval, this may indicate that it could be the driver responsible for the interval's recurrent deletion. In 2016, Iorio *et al.* published statistically significant copy number changes present across a range of cancers, including deleted intervals. This data was taken from 1869 tumours from 12 different tissue types. Six hits from the CRISPR-Cas9 knockout screen were found to be contained within one of these deleted intervals (*Nf1, Pten, Ptbp1, Mcat, Lats2* and *Nup160*) and four of these (*Ptbp1, Mcat, Lats2* and *Nup160*) were genes not listed in the CGC, notably including *Lats2* which was also listed as a gene carrying positively selected mutations in somatic tissues by Martincorena *et al.* These four genes were taken forward to the validation stage.

Homozygous deletions are rare events, which may indicate a potential tumour suppressor gene when seen in tumour genomes. In 2017 Cheng et al. published homozygous deletion intervals found in 2218 primary tumours across 12 human cancer types. When compared with the list of hits from the CRISPR-Cas9 screen, the genes *Smu1, Nf1, Pten* and *Sufu* were found to intersect with homozygously deleted intervals. As *Smu1* is not a listed CGC gene, this gene was chosen for validation.

The Cancer Genome Atlas - cBioportal

Finally, I looked at deletion data across a range of human cancers for all of the remaining hits from the CRISPR-Cas9 knockout screen using cBioportal (Cerami et al., 2012) to view data from The Cancer Genome Atlas (Weinstein et al., 2013). Here I identified a further four genes (*Cdk7, Rnf146, Mak16* and *Kdsr*) from the list of hits that contained frequent deletions in multiple tumour types, listed in table 3.11. These four genes were taken forward for validation.

Gene	Tumour types (deletion frequency)
Cdk7	Adenoid cystic carcinoma (13%), prostate (12%), pancreas (7%), malignant peripheral nerve sheath tumour (7%), ovarian (5%)
Rnf146	Diffuse large B cell lymphoma (13%), pancreas (7%), adenoid cystic carcinoma (7%)
Mak16	Prostate (14%), uterine (5%), bladder (5%), breast (7%), lung adenocarcinoma (6%), liver (6%)
Kdsr	Pancreas (19%), prostate (8%), stomach and oesophageal (7%)

Table 3.11: **Genes containing recurrent deletions in multiple tumour types**
This table lists genes from the list of hits generated by the genome-wide CRISPR-Cas9 screen (see section 3.2.2) that contain recurrent deletions in multiple tumour types. This data was taken from The Cancer Genome Atlas (Weinstein et al., 2013). Tumour types are listed where the deletion frequency >5%.

Candidate gene list for validation

Gene	Mutation Data	MAGeCK comparison(s)	Rank
Cdk7	Recurrent deletion (TGCA)	proliferation-only - focus formation	15
Rnf146	Recurrent deletion (TGCA)	library - focus formation	8
Mak16	Recurrent deletion (TGCA)	proliferation-only - focus formation	23
Kdsr	Recurrent deletion (TGCA)	proliferation-only - focus formation	26
Ptbp1	Recurrent deletion (Iorio et al., 2016)	proliferation-only - focus formation	5
Mcat	Recurrent deletion (Iorio et al., 2016)	proliferation-only - focus formation	18
Lats2	Positively selected driver mutation & recurrent deletion (Iorio et al., 2016)	14-day - focus formation	13
Nup160	Recurrent deletion (Iorio et al., 2016)	proliferation-only - focus formation	33
Smu1	Recurrent homozygous deletion (Cheng et al., 2017)	proliferation-only - focus formation	8
Nf1	Cancer Gene Census gene (positive control)	library - focus formation	3

Table 3.12: **Genes for individual validation**
Genes that were significantly enriched (FDR < 0.01) in the focus formation sample when compared using MAGeCK (Li et al., 2014) with any of the three control samples (library, 14-day or proliferation-only) were prioritised for individual validation by comparison with the sources of cancer genome data listed in section 3.3.5. This table lists the nine genes chosen for individual validation, along with the chosen positive control, *Nf1*. The MAGeCK comparison(s) the gene was enriched in are also listed, alongside its rank order in this comparison when compared with all other genes analysed in the screen.

Validation

Eight of these ten genes were included in validation experiments using individual gRNAs from an arrayed mouse gRNA library (Metzakopian et al., 2017), using two separate gRNAs per gene. gRNA sequences targeting *Lats2* and *Rnf146* are not included in this library, so further work is required to investigate these genes. The use of independent gRNA sequences to those used for the screen itself aims to reduce any off-target effects due to the specific gRNA sequences used in the screen.

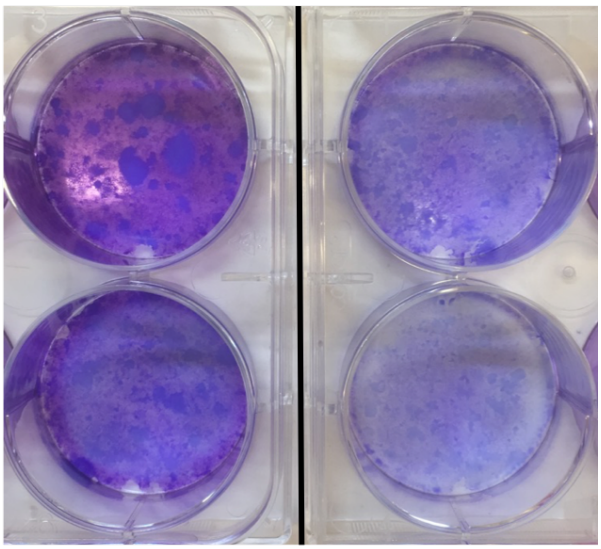
Validation - Arrayed focus formation assay

In this experiment, gRNAs against each gene were used to knock out the genes in NIH3T3-Cas9 in individual wells, followed by culturing the cells for 12 days and staining to visualise the foci of proliferation. Known transforming tumour suppressor gene *Nf1* was included as a positive control, and as a negative control 10 randomly chosen gRNA sequences from the arrayed mouse library were included. Many foci of proliferation were seen in the positive control *Nf1* (figure 3.4a), and no or few foci were seen for the randomly selected negative control genes (figure 3.4b shows *Mical1* as an example). However, no or few foci were observed in the wells transfected with gRNAs against any of the hits from the CRISPR-Cas9 knockout screen (table 3.12).

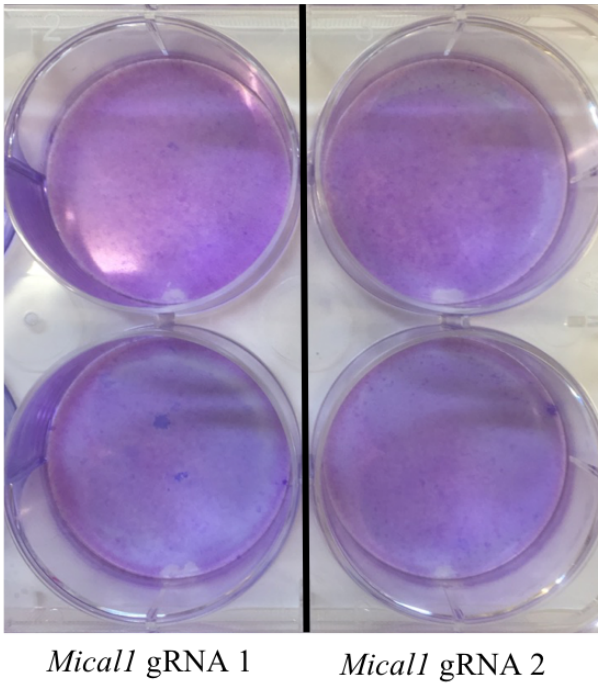
The inability of this assay to validate these hits was originally thought to be due to the difference in the way the gRNAs were introduced to the cells. In the main screen, a lentivirus carrying the Genome-wide Knockout CRISPR Library v2 was used to integrate the gRNA sequences into the genome, whereas for the validation the plasmids from the arrayed mouse library were introduced by transfection, followed by integration into the genome by pCMV-hyPBase. If the efficiency of either transfection or integration was low, then sustained expression of the gRNA may have been poor. In order to avoid this, a lentivirus pool containing gRNAs against all 18 genes was made (see section 3.3.5).

Validation - Pooled gRNA lentivirus

During the culture of the NIH3T3-Cas9 cells, visible foci of proliferation were formed. This suggests that the cells were successfully transformed by at least one of the gRNAs included in the validation virus, however this may be the positive control gRNAs against *Nf1*. The sequencing of the library prepared from the gRNA inserts is currently in progress. These data will then be analysed to determine if gRNA sequences against any of the candidate genes are overrepresented compared to the controls.



(a) **Focus formation assay using 2 gRNAs against positive control gene *Nf1***



(b) **Focus formation assay using 2 gRNAs against negative control gene *Mical1***

Figure 3.4: **Arrayed focus formation assays for genome-wide CRISPR-Cas9 screen validation**
NIH3T3-Cas9 cells were transfected with plasmids from an arrayed mouse CRISPR-Cas9 library (Metzakopian et al., 2017), carrying gRNA sequences against a positive control (known transforming tumour suppressor gene *Nf1*), 10 randomly selected genes, and eight hits from the genome-wide CRISPR-Cas9 knockout screen (see section 3.2.2). The cells were then cultured for 12 days, and stained using crystal violet to visualise foci of proliferation. a) This figure shows NIH3T3-Cas9 cells transfected with two plasmids expressing two different gRNAs against the positive control gene, *Nf1*. gRNA 1 produces visibly more and larger foci, suggesting that there may be variation in gRNA efficiency. b) This figure shows NIH3T3-Cas9 cells transfected with two plasmids expressing two different gRNAs against one of the randomly selected negative control genes, *Mical1*.

Re-analysis of gRNA read counts

After the putative transformation associated genes failed to show any difference in transforming ability from the negative control genes in the arrayed focus formation validation assay described above, I revisited the data generated by the genome-wide CRISPR-Cas9 screen, and its analysis using MAGeCK. In order to examine the original read count data that was used as an input for the MAGeCK analysis, the normalised read counts for each gRNA in the different samples were visualised (figure 3.5). These three figures compare read counts of individual gRNAs between the pairs of samples that were compared in the MAGeCK analysis described in section 3.2.2.

In these plots, gRNAs that are genuinely enriched in the focus formation sample lie above the y=x line (more frequent in the focus formation sample than in the control), and also show a high total read count in this sample. However, the figures indicate that the only genes for which this is consistently true for multiple gRNAs are *Rnf146* and the positive control *Nf1* (this is seen in the library - focus formation (figure 3.5a) and 14-day - focus formation (figure 3.5b) comparisons). Another gene where the gRNA sequences may show genuine enrichment is *Lats2,* which was determined by MAGeCK to be significantly enriched in the 14-day - focus formation comparison. In the plot of the normalised read counts for this comparison (figure 3.5b), one gRNA sequence is far more prevalent in the focus formation sample than in the 14-day sample, and the other four are present in similar amounts in both samples.

For these three genes, the interpretation of these figures is consistent with the results obtained using MAGeCK. *Nf1* and *Rnf146* were determined by MAGeCK analysis to be significantly enriched in the library - focus formation comparison, and *Lats2* was called as significantly enriched in the 14-day - focus formation comparison (see table 3.12). This indicates that MAGeCK was successful at identifying genuine hits when comparing the focus formation sample when using the library sample as a control. This was also potentially true when the 14-day sample was used as a control, although this is more uncertain as *Lats2* is not as clearly enriched as *Rnf146* or *Nf1*. The success of the MAGeCK analysis when using the library sample as a control is apparent from a similar figure highlighting all of the genes identified as significantly enriched in the focus formation sample when compared with the library (figure 3.6). Unlike the other figures, this plot shows the expected distribution of read counts for a list of genuine hits, with nearly all highlighted genes having the majority of their gRNAs above the y=x line, showing that they are more frequent in focus formation sample than in the library. Importantly, these gRNAs also show a high total read count in the focus-formation sample. This comparison identified known Cancer Gene Census (Futreal et al., 2004)genes *Sufu, Nf2, Ptch1, Nf1* and *Pten,* alongside the six other genes listed in the plot. The only gene from this list that was chosen for the validation stage was *Nf1,* which was used as a positive

control.

However, in the figure comparing read counts in the focus formation and proliferation-only samples (figure 3.5c), the genes that MAGeCK analysis determined to be significantly enriched when using the proliferation-only sample as a control (*Cdk1, Kdsr, Mak16, Mcat, Nup160, Ptbp1* and *Smu1*) are not distributed as would be expected for genes involved in transformation. While the gRNA counts lie mostly above the y=x line (more frequent in the focus formation than in the proliferation-only sample), read counts are mostly relatively low in both samples. Clearly, the expected result for a gene involved in the proliferation of the foci formed during the screen would be to have a high read count in the focus formation sample.

A potential explanation for why the MAGeCK analysis called these genes as enriched can be seen when comparing the read counts from the focus formation sample with those from the gRNA library (figure 3.5a). In this figure, the genes that were identified by MAGeCK as enriched in the focus formation sample when the proliferation-only sample was used as a control are actually *underrepresented* in the focus-formation sample when compared with the library. For three of these genes, *Kdsr, Nup160* and *Smu1*, MAGeCK actually calls them as significantly depleted (FDR < 0.01) in the focus formation sample when using the library as the control (FDR = 0.003, 0.004 and 0.005 respectively). This result implies that these genes are potentially essential or highly important genes for normal cellular survival.

This illustrates a potential issue when using MAGeCK to detect gRNA enrichment, where the read counts from the original plasmid library are not used as the control. If a gene is essential, read counts of gRNAs against it will drop dramatically between the input library, and subsequent samples, leaving a low number of reads. However, if a comparison is then made directly between two samples, both with low read counts, the relative difference can be large due to the low signal:noise ratio. This noise can be derived from biological factors such as variation in the efficiency of individual gRNAs, or random variation. For some genes, there will be a large relative difference between the two low read counts due to this noise, leading MAGeCK to call this as a significant enrichment between the two samples. Figure 3.7 and table 3.13 show the mean normalised gRNA read counts for the three genes (*Kdsr, Nup160* and *Smu1*) that were called as significantly depleted in the focus formation sample when compared with the gRNA library, but significantly enriched in the focus formation sample when compared with the proliferation-only sample. From these figures, it is clear that the genes are in fact essential genes, with much lower read counts in both cultured samples compared to the input read count from the library. However, the very low read counts in the samples have led to noise resulting in a large relative increase seen between the proliferation-only and focus formation samples, leading to the counterintuitive result of the MAGeCK analysis. These genes were included in the validation stage due to their recurrent presence in deletion intervals found

Gene	gRNA library	Proliferation-only	Focus formation
Kdsr	469.2	6.4	9.3
Nup160	513.6	5.7	11.4
Smu1	648.8	5.9	11.3

Table 3.13: **Mean normalised read counts for gRNAs against *Kdsr, Nup160* and *Smu1* in library, proliferation-only and focus formation samples**
This table shows the mean of the normalised read counts for the gRNAs against three genes (*Kdsr, Nup160* and *Smu1*) for the gRNA library, proliferation-only and focus formation samples taken from the genome-wide CRISPR-Cas9 knockout screen (see section 3.2.2).

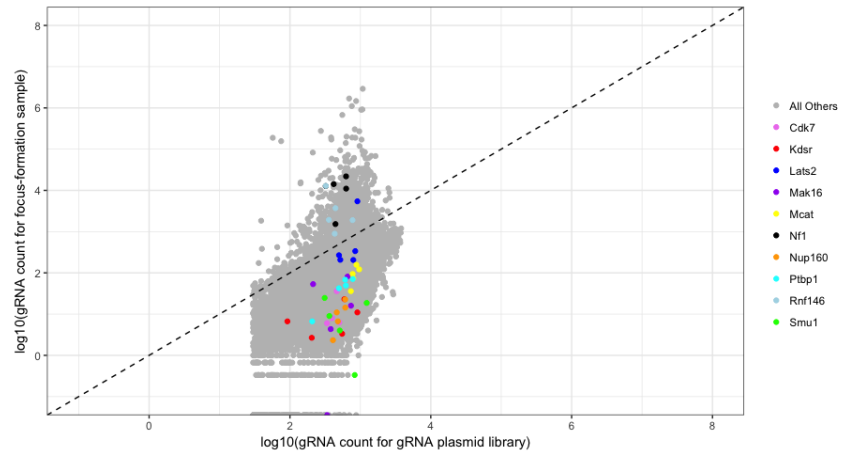
in human cancers, however it is probable that this is purely a passenger effect, potentially due to other nearby driver genes.

This phenomenon could explain why none of the hits identified from the screen were validated except for *Nf1*, which was included due to its enrichment in the focus formation sample when using the library as a control. *Rnf146* and *Lats2* are also potentially valid hits as they were also not generated from the comparison using the proliferation-only sample as a control, and showed high total read counts in the focus-formation sample. Unfortunately, as mentioned above (section 3.3.5: Validation) gRNAs against both *Rnf146* and *Lats2* were not included in the arrayed mouse gRNA library (Metzakopian et al., 2017), so these were not able to be included in the validation at this stage.

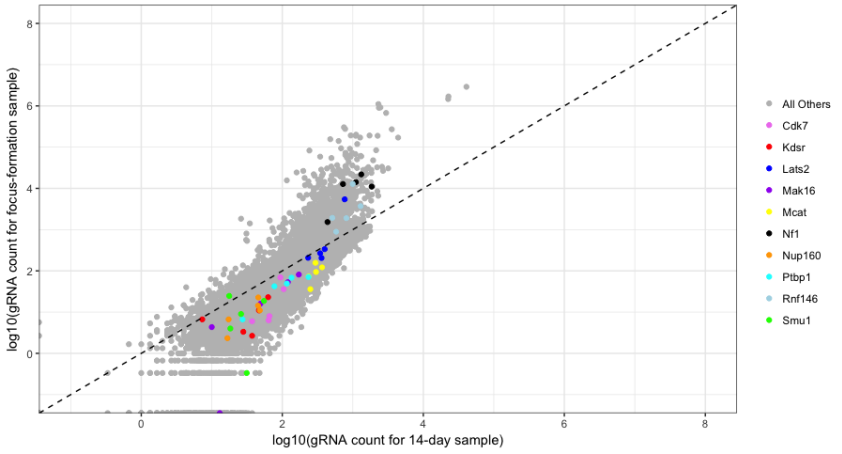
3.4 Discussion

The aim of the genome-wide CRISPR-Cas9 screen described in this chapter was to identify genes involved in the earliest stage of tumourigenesis, mediating the initial transition of a single cell to malignancy that may then clonally proliferate and form a tumour. The advantage of looking for these genes is that they will probably be present clonally in the tumour, presenting a potential therapeutic target. Further knowledge of the genes involved in transformation may also help to elucidate early mechanisms of tumourigenesis.

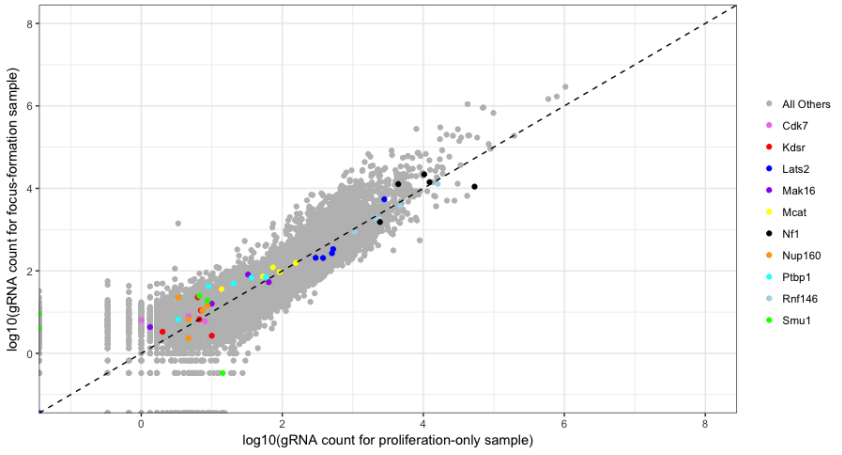
One limitation of the approach used to identify these genes is that it is only able to detect those that cause malignant transformation when mutated *alone* in the NIH3T3-Cas9 genetic background, as each gRNA causes loss-of-function mutation of a single gene. Given that cancer is a polygenic disease, it is possible that some mutations may have to work in combination to initiate transformation. For example, *BRAF* V600E is the most common initiating mutation in melanocytic neoplasms, causing the formation of naevi. Alone, this mutation does not cause malignant transformation, forming a benign lesion where proliferation is limited by cellular



(a) **gRNA plasmid library - Focus formation**



(b) **14-day - Focus formation**



(c) **Proliferation-only - Focus formation**

Figure 3.5: **Comparisons between normalised gRNA read counts in different samples taken from the genome-wide CRISPR-Cas9 knockout screen for genes associated with malignant transformation**

These figures show the $\log_{10}$(normalised gRNA count)s for each gRNA sequence in the Genome-wide Knockout CRISPR Library v2 (Koike-Yusa et al., 2013), plotting the values derived from different samples from the genome-wide CRISPR-Cas9 knockout screen against each other. Each figure compares the data from the focus formation sample with a different control sample (gRNA library, 14-day and proliferation-only), corresponding to one of the datasets analysed by MAGeCK in section 3.3.2. The coloured points represent gRNA sequences against genes that were identified by MAGeCK analysis (Li et al., 2014) as enriched in the focus formation sample with respect to one of the controls, and were then taken forward to the validation stage on the basis of comparison with existing cancer genome data (see section 3.3.5). The black dotted lines (y=x) indicate the point at which the normalised gRNA read counts are equal in the two samples, with gRNA sequences that are enriched in the focus formation sample lying above this line.

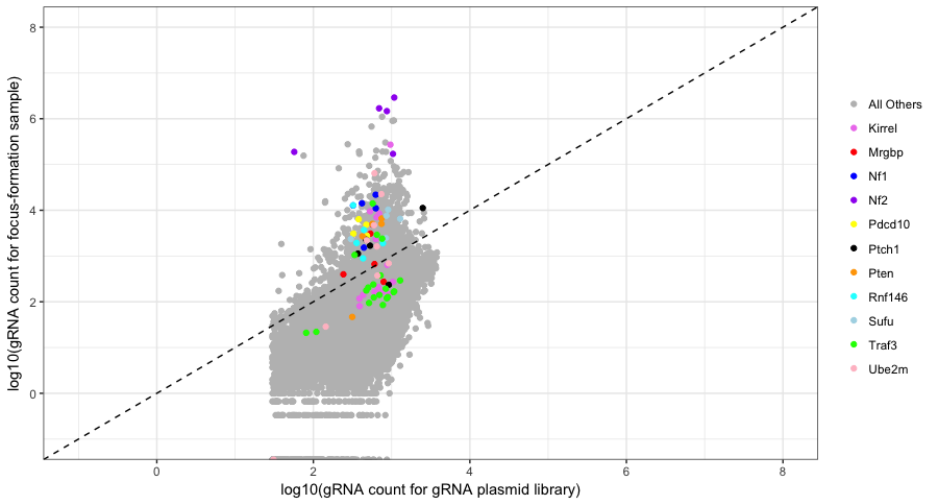


Figure 3.6: **Comparison between normalised gRNA read counts in the plasmid library and focus formation samples taken from the genome-wide CRISPR-Cas9 knockout screen for genes associated with malignant transformation**

This figure shows the $\log_{10}$(normalised gRNA count)s for each gRNA in the Genome-wide Knockout CRISPR Library v2 (Koike-Yusa et al., 2013), plotting the values derived from the plasmid library and the focus-formation sample from the genome-wide CRISPR-Cas9 knockout screen against each other. The coloured points represent gRNA sequences against genes that were identified by MAGeCK analysis (Li et al., 2014) as enriched in the focus formation sample with respect to the plasmid library. The black dotted line (y=x) indicates the point at which the normalised gRNA read counts are equal in the two samples, with gRNA sequences that are enriched in the focus formation sample lying above this line.

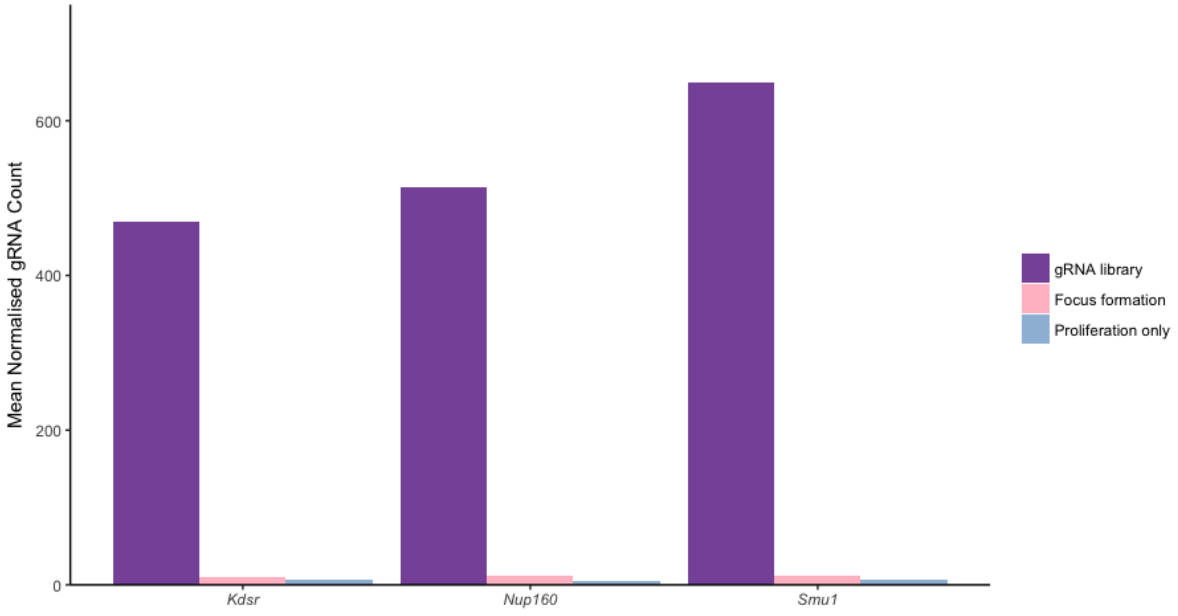


Figure 3.7: **Mean normalised read counts for gRNAs against *Kdsr, Nup160* and *Smu1* in library, proliferation-only and focus formation samples**

This figure shows the mean of the normalised read counts for the gRNAs against three genes (*Kdsr, Nup160* and *Smu1*) for the gRNA library, proliferation-only and focus formation samples taken from the genome-wide CRISPR-Cas9 knockout screen (see section 3.2.2).

senescence (Michaloglou et al., 2005). However, when accompanied by further mutations, activating *BRAF* mutation can lead to the development of malignant melanoma. For example, the combination of *BRAF* V600E and inactivating *PTEN* mutation has been shown in mice to cause metastatic melanoma (Dankort et al., 2009). This example illustrates how multiple mutations can be required to cause malignant transformation, and therefore screens using gRNAs targeting single genes may be unable to detect certain driver genes due to the requirement for accompanying genetic alterations. One way to test the effects of multiple mutations occuring simultaneously is to use a plasmid library where each plasmid carries more than one gRNA sequence. However, this approach is not suitable for exhaustive genome-wide screens as the number of gene combinations would be prohibitively high. This means that some *a priori* hypothesis about which gene combinations may be of interest is required to design a practical number of guides to make up a screening library.

Additionally, this screen was conducted in a single cell line, and therefore may fail to identify mutations that require a different genetic or epigenetic background to induce transformation. NIH3T3-Cas9 cells are transformation-sensitive and have an abnormal karyotype (see section 2.3.5), potentially presenting a lower genetic barrier to malignant transformation than genetically and phenotypically 'normal' cells. It is therefore unclear whether any identified mutations would have the same effect in 'normal' cells *in vivo*. Another potential complicating factor is the genetic heterogeneity of NIH3T3-Cas9 cells. As discussed in Chapter 2 (section 2.3.5), the cell line appears to exhibit chromosomal instability, leading to a range of large-scale alterations in the genome that differ between individual cells. Therefore, mutations in different cells within the population are acting in different genetic environments.

A further issue with screening *in vitro* is that transformation in this context may not fully recapitulate the *in vivo* phenotype. The aim of the screen is to identify genes that are involved in the transition of a cell to maligancy, forming a tumour with the ability to metastasise. It is not certain that the *in vitro* formation of proliferative foci is phenotypically equivalent to this; for example, it may not be able to differentiate between mutations that cause benign and malignant tumours. Genome-wide screening *in vivo* is not feasible, however it is possible to perform further *in vivo* validation of genes that were successfully validated *in vitro*. For example, the injection of CRISPR-Cas9 edited NIH3T3 cells into a mouse model and observation of tumour initiation over time compared to control wild-type NIH3T3 cells could determine the ability of mutations identified during the screen to initiate transformation *in vivo*. This approach would also have the advantage of accounting for factors such as the immune response that may make it more difficult for a transformed cell to establish a tumour.

The analysis of the data from this screen identified a potential issue with using the MAGeCK algorithm to detect enrichment of gRNA sequences when not using the gRNA library as a

control. The presence of gRNA sequences in the library that target essential or other highly advantageous genes meant that, for these genes, read counts dropped to very low levels in all cultured samples. This led to the identification of hits that, on closer examination, are actually likely to be essential genes, due to high levels of statistical noise (see section 3.3.5).

There are multiple approaches that could be used to attempt to avoid this issue. Examination of the distributions of the read count data using plots such as those in figure 3.5 could be used at an earlier stage in the analysis, to confirm that hits called by MAGeCK as enriched correspond to gRNA sequences present at high overall read counts in the test sample. The MAGeCK analysis could also be examined to discard any genes that are significantly depleted in the test sample when compared to the gRNA library, suggesting they may be essential. Another possibility is to use a minimum threshold for read count in the control before the data from a particular gRNA is used in the MAGeCK analysis. Alternatively, the issue could be avoided entirely by using the gRNA library as the control sample. For example, to compare the gRNA counts between the focus formation and proliferation-only samples, both samples could have been independently compared with the gRNA library using MAGeCK, followed by comparing the genes enriched in each comparison.

Two genes were both identified as enriched by the MAGeCK analysis, and also looked promising on further investigation of the original read counts - *Rnf146* and *Lats2*. Unfortunately, gRNA sequences targeting these genes were not included in the library used for validation (Metzakopian et al., 2017). In future, gRNA sequences against these genes could be cloned into the plasmid backbone used in this library, and used to validate these hits. Additionally, there are further novel genes that were enriched in the focus formation sample when compared to the library that may be of interest - *Kirrel, Mrgbp, Pdcd10, Traf3* and *Ube2m* (see figure 3.6). For these genes that are enriched compared to the library, validation using a focus formation assay is crucial to ensure that their mutation actually enables formation of transformed foci, rather than simply increasing rate of proliferation and causing enrichment of gRNA sequences targeting them in the absence of transformation.

Overall, the work described in this chapter has identified some potential genes that may be involved in malignant transformation when subjected to loss-of-function mutations. If successfully validated in future, these genes may represent useful sources of information about the early stages of tumourigenesis or even potential therapeutic targets. Additionally, this work highlighted a potential issue to be aware of when using MAGeCK to analyse CRISPR-Cas9 knockout screen data, suggesting that consideration of the original read count data alongisde the results of the algorithm is advisable in order to identify and eliminate spurious hits.