Appendix A

Software and databases

Software	Version	Citation
BCFtools	1.8	(Danecek et al., 2011)
bedtools	2.21.1	(Quinlan and Hall, 2010)
cBioportal	1.14.0	(Cerami et al., 2012)
COSMIC	85	(Forbes et al., 2017)
Ensembl Variant Effect Predictor	92	(McLaren et al., 2016)
Ensembl BioMart	92	(Zerbino et al., 2018)
MAGeCK	0.5.7	(Li et al., 2014)
pROC	1.12.1	(Robin et al., 2011)
samtools	1.8	(Li et al., 2009)

Table A.1: **Software and databases used in analyses**

Appendix B

Supplementary information

B.1 Generation of the NIH3T3-Cas9 cell line

Materials

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

Cas9 virus The Cas9 lentivirus was generated by Gemma Turner from the experimental cancer genetics group at the Wellcome Sanger Institute, using pKLV2-EF1a-Cas9Bsd-W (this plasmid was a gift from Dr. Kosuke Yusa, Addgene plasmid #68343).

Reagents

Reagent	Manufacturer
Blasticidin (10mg/mL)	InvivoGen
Dimethyl sulphoxide (DMSO)	Sigma-Aldrich
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma-Aldrich
Fetal bovine serum (FBS)	Gibco
Paraformaldehyde	Sigma-Aldrich
Penicillin, streptomycin and L-glutamine (100X)	Gibco
Phosphate-buffered saline (PBS)	Sigma-Aldrich
Polybrene (10mg/mL)	Sigma-Aldrich
TrypLE Express Enzyme	Gibco

Table B.1: **Reagents used in the generation of the NIH3T3-Cas9 cell line**

Method

Day 1 2.5 x 10^6 NIH3T3 cells were infected in suspension in 3.5mL complete DMEM (DMEM supplemented with 10% FBS and 500µg/mL penicillin, streptomycin and L-glutamine) containing polybrene at 8µg/mL. 1.5mL Cas9 virus was added and cells were seeded in a T25 tissue culture flask.

Day 2 Media was changed to complete DMEM.

Days 4-11 Cells were split every 3-4 days, 2µg/mL blasticidin was added to the media to select for Cas9 expressing cells.

Day 14 Cells were detached using TrypLE Express Enzyme and flasks were pooled before freezing in liquid nitrogen in cryopreservation medium (50% DMEM, 40% FBS, 10% DMSO).

Acknowledgement

This work was performed by Dr. Nicola Thompson from the experimental cancer genetics group at the Wellcome Sanger Institute.

B.2 Cas9 activity determination in NIH3T3-Cas9

Cas9 activity was assessed using a reporter vector expressing BFP, GFP and a gRNA targeting GFP (gGFP). Cas9 activity was determined based on the percentage of cells that are BFP positive but GFP negative, indicating successful knockout of the GFP gene by Cas9.

Plasmids

pKLV2-U6gRNA5(Empty)-PGKGFP2ABFP-W This plasmid was a gift from Dr. Kosuke Yusa (Addgene #67983).

pKLV2-U6gRNA5(gGFP)-PGKBFP2AGFP-W This plasmid was a gift from Dr. Kosuke Yusa (Addgene #67980).

Method

Day 1

500,000 NIH3T3-Cas9 cells/well were seeded in 3 wells of a 6-well plate in complete DMEM (250,000 cells/mL). 1.6μL of polybrene was added per well. For mock infection nothing further was added, for control infection 100μL of a lentivirus containing the BFP/GFP plasmid (pKLV2-U6gRNA5(Empty)-PGKGFP2ABFP-W) was added, and for the final well 100μL of a lentivirus containing the BFP/GFP/gGFP plasmid (pKLV2-U6gRNA5(gGFP)-PGKBFP2AGFP-W) was added.

Day 2

Media was changed to complete DMEM.

Day 4

Cells were harvested using TrypLE Express enzyme, fixed using 4% paraformaldehyde in PBS for 10 minutes, and centrifuged (200*xg*, 5 minutes). Cells were resuspended in 1% FBS in PBS and protein expression was assessed using flow cytometry using the following filters/detectors: BFP 450/50 (405)-A; GFP 530/30 (488)-A. Baseline values for negative/positive expression of BFP and GFP were established using the control infected and mock infected samples. The proportion of cells expressing active Cas9 was determined to be 82%, based on the percentage of BFP positive, GFP negative cells.

Acknowledgement

This work was performed by Dr. Nicola Thompson from the experimental cancer genetics group at the Wellcome Sanger Institute.

B.3 NIH3T3 wild-type variants with coding consequences overlapping mouse homologues of CGC genes

Table B.2: **NIH3T3 wild-type coding variants in mouse homologues of CGC genes**

Chromosome	Position	Reference	Alternate	Genotype	Gene	Sequence ontology term

1	138117790	G	C	0/1	*Ptprc*	missense variant
1	143701990	GAAAAAAAA	GAAAAAAA	0/1	*Cdc73*	frameshift variant
1	150443179	AACA	AACACA	1/1	*Tpr*	splice acceptor variant
1	156641457	G	C	0/1	*Abl2*	missense variant
2	112248256	G	T	0/1	*Nutm1*	missense variant
2	122151119	C	A	1/1	*B2m*	missense variant
2	126758523	T	G	0/1	*Usp8*	missense variant
3	15542385	G	C	1/1	*Sirpb1b*	missense variant
3	15832375	G	C	1/1	*Sirpb1c*	missense variant
3	103280187	AGCCCCGGCCCC GGCCCCGGCCCC GGCCCCGGCCCC	AGCCCCGGCCCCG GCCCCGGCCCCGG CCCCGGCCCCGGC CCC	1/1	*Trim33*	inframe insertion
4	75956319	G	T	0/1	*Ptprd*	missense variant
4	126178083	A	G,T	2/1	*Thrap3*	missense variant
4	133752826	CGAGGAGG	CGAGG	1/1	*Arid1a*	inframe deletion
4	141516845	TTGCTGCTG CTGCTGCTG CTGCTGCTG	TTGCTGCTGCTGC TGCTGCTGCTG	1/1	*Spen*	inframe deletion
4	143135893	GCTCCTCCT CCTCCTCCT CCTCCTC	GCTCCTCCTCCT CCTCCTCCTC	1/1	*Prdm2*	inframe deletion
4	151010416	GAC	GACGGACAC	1/1	*Per3*	protein altering variant
5	67097668	TCCC	TCCCC	0/1	*Phox2b*	frameshift variant
5	103501611	G	T	0/1	*Ptpn13*	missense variant
5	125106206	T	A	0/1	*Ncor2*	missense variant
5	147306749	A	C	0/1	*Cdx2*	missense variant
5	150541525	A	G	0/1	*Brca2*	missense variant
5	150543195	A	T	0/1	*Brca2*	missense variant
6	17533897	CTTTTTTTTTT	CTTTTTTTTTTT TT	1/1	*Met*	splice acceptor variant
6	125036455	CCAAGCTCAAGC	CCAAGC	0/1	*Zfp384*	inframe deletion
6	125036464	AGCCCAGGCCCA GGCCCAGGCCCA GGCCCAGGC	AGCCCAGGCC CAGGCCCAGG CCCAGGCCCA GGCCCAGGCC CAGGC	1/1	*Zfp384*	inframe insertion

6	125122132	CCCCCTGCCCCT GCCCCTGCCACT GCCCCTGCC	CTCCCTGCCC CTGCCCCTGC CCCTGCCACT GCCCCTGCC	1/1	*Chd4*	protein altering variant
6	143217634	GG	GGACAG	1/1	*Etnk1*	frameshift variant
7	35409642	ACTCCTCCTCCT CCTCCTCCTCCTC	ACTCCTCCTC CTCCTCCTCC TC	1/1	*Cep89*	inframe deletion
7	80098332	CCCAGGGCCAGG GCCAGGGCCAG	CCCAGGGCCA GGGCCAG	1/1	*Idh2*	inframe deletion
7	80502467	GTCATCATCATCA TCATCATCATCA	GTCATCATCAT CATCATCATCA	1/1	*Blm*	inframe deletion
7	80512904	GCCTCCTCCTCC TCCTCCTCCTCC TCCTCC	GCCTCCTCCTC CTCCTCCTCCT CCTCCTCCTCC	1/1	*Blm*	protein altering variant
7	102145442	TAGAA	TAGAAGAA	1/1	*Nup98*	inframe insertion
7	102145495	GCC	GCCTGCAGCAC TGTGCCCTCCCC TGCACTTAGTT CCC	1/1	*Nup98*	frameshift variant
7	122590166	G	T	0/1	*Prkcb*	missense variant
7	130759613	A	C	0/1	*Tacc2*	missense variant
8	70392070	G	C	0/1	*Crtc1*	missense variant
8	108956091	ACAGCAACAGC AGCA	ACAGCAACAGCA GCAACAGCAGCA	1/1	*Zfhx3*	inframe insertion
8	108956100	GCAGCAGCAAC AGCGGCAACTA CAGCA	GCAGCA	0/1	*Zfhx3*	inframe deletion
9	16376784	C	T	1/1	*Fat3*	missense variant
9	18644173	G	A	1/1	*Muc16*	missense variant
9	18654473	GTTGAAATTGAA	GTTGAA	1/1	*Muc16*	inframe deletion
9	44848133	T	A	1/1	*Kmt2a*	missense variant
9	71849844	T	C	1/1	*Tcf12*	missense variant
9	75776376	AGGAGTCGGAGT	AGGAGT	1/1	*Bmp5*	inframe deletion
9	95865570	C	G	1/1	*Atr*	missense variant
10	28493041	G	T	0/1	*Ptprk*	missense variant
10	52081998	C	A	0/1	*Ros1*	missense variant
10	93847307	T	A	1/1	*Usp44*	missense variant

10	127331182	C	A	0/1	*Gli1*	missense variant
10	127647806	C	A	0/1	*Stat6*	missense variant
11	49643527	G	C	0/1	*Flt4*	missense variant
11	75761161	CCCCCCAAA	CA	0/1	*Gm26836*	splice donor variant
11	88687463	GCCCC	GCC	1/1	*Msi2*	frameshift variant
11	119409459	CAGGAG	CAG	0/1	*Rnf213*	inframe deletion
13	67139785	CAAAAA	CAA	1/1	*Zfp759*	inframe deletion
13	112495176	CTTTTTTTTT	CTTTTTTTTTTT	1/1	*Il6st*	splice acceptor variant
14	47704466	C	G	0/1	*Ktn1*	missense variant
15	30619227	CAG	CAGGAG	1/1	*Ctnnd2*	protein altering variant
15	47847161	C	G	0/1	*Csmd3*	missense variant
15	98849587	ATGCTGCTG CTGCTGCTG CTGCTGCTG CTGCTGCTG CTGCTGCTG CTGCTG	ATGCTGCTGCTG CTGCTGCTGCTG CTGCTGCTGCTG CTGCTGCTGCTG CTGCTG	1/1	*Kmt2d*	inframe insertion
15	98851005	CCTGCTGCT GCTG	CCTGCTG	1/1	*Kmt2d*	inframe deletion
16	32753466	C	A	1/1	*Muc4*	missense variant
16	32753802	T	C	0/1	*Muc4*	missense variant
16	32753919	G	C	0/1	*Muc4*	missense variant
16	32754065	G	C	0/1	*Muc4*	missense variant
16	32754425	A	C	0/1	*Muc4*	missense variant
16	32754794	T	C	1/1	*Muc4*	missense variant
16	32756159	C	A	1/1	*Muc4*	missense variant
16	32757020	T	C	1/1	*Muc4*	missense variant
16	74035037	G	T	0/1	*Robo2*	missense variant
17	4995186	CCCACCACC ACCACCACC ACCA	CCCACCAC CACCACCA CCACCACC ACCA	1/1	*Arid1b*	inframe insertion
17	4995586	GGGCGGCGG CGGC	GGGCGGCG GCGGCGGC	1/1	*Arid1b*	inframe insertion
17	4995925	AGCAGCGGC AGCGGCAGC	AGCAGCGG CAGC	1/1	*Arid1b*	inframe deletion
17	33912659	CGATGATGAT GA	CGATGATGA	1/1	*Daxx*	inframe deletion

17	35264016	G	C	0/1	*H2-D1*	missense variant
17	35380388	G	C	0/1	*H2-Q4*	missense variant
17	35439650	C	A	0/1	*H2-Q7*	missense variant
17	35508871	CTGTG	CTG	1/1	*Pou5f1*	splice donor variant
17	36031053	C	A	1/1	*H2-T23*	stop gained
17	36032145	GCA	GCAGTCA	1/1	*H2-T23*	splice donor variant
17	36083252	C	A	0/1	*H2-Bl*	missense variant
17	36119331	T	G	1/1	*H2-T10*	missense variant
17	36120282	GTTTCCCAC TGTTTTCCC ACTGT	GTTTCCCACT GT	1/1	*H2-T10*	inframe deletion
17	36187455	G	C	1/1	*H2-T3*	missense variant
17	36189406	GAAGAACTC CA	GA	0/1	*H2-T3*	inframe deletion
17	36549178	ATTGTTGT	ATTGT	1/1	*H2-M11*	inframe deletion
17	47786091	ACAGCAGC AGCAGCAG CAGCAGCA GC	ACAGCAGCAG CAGCAGCAGC AGCAGCAGC	1/1	*Tfeb*	inframe insertion
19	39807469	GCACACACA CACACACAC ACACACACA CACACACAC ACACAC	GCACACACACA CACACACACAC ACACACACACA CACACACACACAC, GCACACACACAC ACACACACACAC ACACACACACAC ACACACAC	2,1	*Cyp2c40*	splice donor variant

This table contains variants found in NIH3T3 wild-type that overlap mouse homologues of Cancer Gene Census (Futreal et al., 2004) genes, along with their positions in the mouse genome (GRCm38). Genotypes: 0/1 = heterozygous reference and alternate allele, 1/1 = homozygous alternate allele, 2/1 = heterozygous first alternate allele and second alternate allele (comma separated). Sequence Ontology terms (Eilbeck et al., 2005) were assigned to each variant by Ensembl Variant Effect Predictor (McLaren et al., 2016), with this table listing only those with 'high' or 'moderate' coding consequences.

B.4 Verification of amplified Genome-wide Knockout CRISPR Library v2

The amplified Genome-wide Knockout CRISPR Library v2 (see section 3.2.2) was verified by sequencing (section 3.2.2), to compare its characteristics with those of the original library. Read counts were generated for each gRNA sequence, and a frequency histogram of these is plotted in figure B.1. The ratio between the 90th and 10th percentile is 4.72, indicating an acceptable level of variation in read counts between gRNAs. The number of gRNA sequences with 0 reads is 362 (0.4% of the total), showing that the gRNA sequence representation of the original library has been maintained well during amplification.

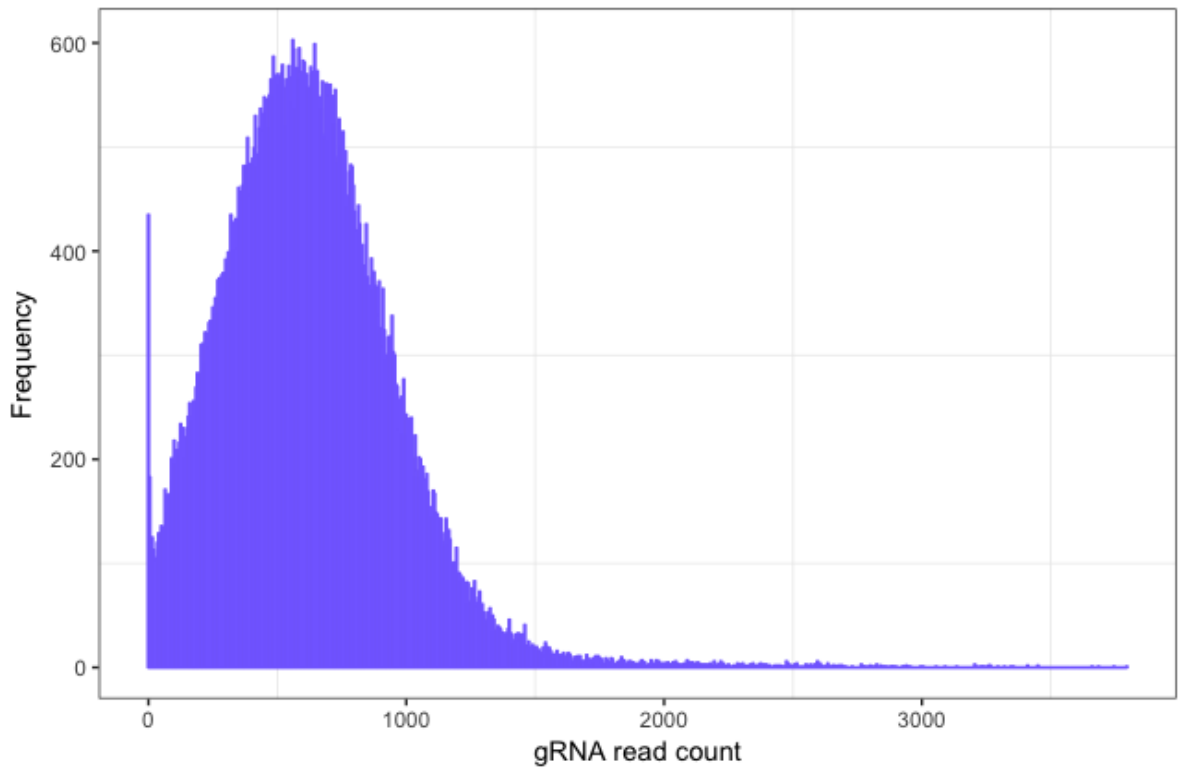

Figure B.1: **Histogram of gRNA read counts generated from sequencing of amplified Genome-wide Knockout CRISPR Library v2**
Genome-wide Knockout CRISPR Library v2 (Addgene #67988, (Koike-Yusa et al., 2013)) was amplified according to the depositor's instructions and then sequenced as detailed in section 3.2.2. In this figure, the read counts of the individual gRNA sequences present in the library are plotted against the frequency of their occurence (bin size = 5).

B.5 Primer sequences for CRISPR-Cas9 gRNA insert library preparation

B.5.1 1st round PCR - Genome-wide CRISPR-Cas9 knockout screen

Primer	Sequence
Forward primer sequence	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTGGAAAGGACGAAACA
Reverse primer sequence	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTTAAAGCGCATGCTCCAGAC

Table B.3: **Primer sequences for the 1st round PCR in the CRISPR-Cas9 gRNA insert library preparation for the genome-wide CRISPR-Cas9 knockout screen**

B.5.2 1st round PCR - Validation (pooled gRNA lentivirus)

Primer	Sequence
Forward primer sequence	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTGGAAAGGACGAAACA
Reverse primer sequence	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTACTCGGTGCCACTTTTTCAA

Table B.4: **Primer sequences for the 1st round PCR in the CRISPR-Cas9 gRNA insert library preparation for the validation using a pooled gRNA virus**

B.5.3 2nd round PCR

Primer	Seqeunce
Forward primer sequence	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATC*T
Reverse primer sequence 1	CAAGCAGAAGACGGCATACGAGATAACGTGATGAGATCGGTCTCGGCATTCC TGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 2	CAAGCAGAAGACGGCATACGAGATAAACATCGGAGATCGGTCTCGGCATTCC TGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 3	CAAGCAGAAGACGGCATACGAGATATGCCTAAGAGATCGGTCTCGGCATTCC TGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 4	CAAGCAGAAGACGGCATACGAGATAGTGGTCAGAGATCGGTCTCGGCATTC CTGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 5	CAAGCAGAAGACGGCATACGAGATACCACTGTGAGATCGGTCTCGGCATTC CTGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 6	CAAGCAGAAGACGGCATACGAGATACATTGGCGAGATCGGTCTCGGCATTC CTGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 7	CAAGCAGAAGACGGCATACGAGATCAGATCTGGAGATCGGTCTCGGCATTC CTGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 8	CAAGCAGAAGACGGCATACGAGATCATCAAGTGAGATCGGTCTCGGCATTC CTGCTGAACCGCTCTTCCGATC*T
Reverse primer sequence 9	CAAGCAGAAGACGGCATACGAGATCGCTGATCGAGATCGGTCTCGGCATTC CTGCTGAACCGCTCTTCCGATC*T

Table B.5: **Primer sequences for the 2nd round PCR in the CRISPR-Cas9 gRNA insert library preparation**

For each of the 9 samples derived from the screen a different reverse primer sequence was used, acting as a tag for sequencing. For the validation only one was needed as there was a single sample. The same forward primer was used for each sample. The C*T notation indicates a phosphorothioate bond before the terminal T residue to protect the oligonucleotide from exonuclease digestion during the library construction process.

B.6 Genes identified by the genome-wide CRISPR-Cas9 knock-out screen

Table B.6: **Putative transformation-associated genes identified by the genome-wide CRISPR-Cas9 knockout screen**

Gene	Mouse chromosome	Mouse position	Equivalent human chromosome	Equivalent human position	Genotype	Sequence ontology term
Cdc73	1	143701990	1	193122777	0/1	frameshift variant
Tpr	1	150443179	1	186322588	1/1	splice acceptor variant
Trim33	3	103280187	1	114510763	1/1	inframe insertion
Arid1a	4	133752826	1	26697179	1/1	inframe deletion
Spen	4	141516845	1	15876649	1/1	inframe deletion
Prdm2	4	143135893	1	13778600	1/1	inframe deletion
Per3	4	151010416			1/1	protein altering variant
Phox2b	5	67097668	4	41747334	0/1	frameshift variant
Met	6	17533897	7	116757424	1/1	splice acceptor variant
Zfp384	6	125036455	12	6667979	0/1	inframe deletion
Zfp384	6	125036464	12	6667952	1/1	inframe insertion
Chd4	6	125122132	12	6581155	1/1	protein altering variant
Etnk1	6	143217634			1/1	frameshift variant
Cep89	7	35409642	19	32948291	1/1	inframe deletion
Idh2	7	80098332	15	90087643	1/1	inframe deletion
Blm	7	80502467	15	90761056	1/1	inframe deletion
Blm	7	80512904	15	90749940	1/1	protein altering variant
Nup98	7	102145442	11	3712338	1/1	inframe insertion
Nup98	7	102145495	11	3712397	1/1	frameshift variant
Zfhx3	8	108956091			1/1	inframe insertion
Zfhx3	8	108956100	16	72788122	0/1	inframe deletion
Muc16	9	18654473	19	8945781	1/1	inframe deletion
Bmp5	9	75776376	6	55874571	1/1	inframe deletion
Gm26836	11	75761161			0/1	splice donor variant
Msi2	11	88687463	17	57289571	1/1	frameshift variant
Rnf213	11	119409459	17	80288688	0/1	inframe deletion
Zfp759	13	67139785	19	21972541	1/1	inframe deletion
Il6st	13	112495176	5	55954997	1/1	splice acceptor variant
Ctnnd2	15	30619227	5	11412062	1/1	protein altering variant
Kmt2d	15	98849587	12	49037507	1/1	inframe insertion

Kmt2d	15	98851005			1/1	inframe deletion
Arid1b	17	4995186	6	156778195	1/1	inframe insertion
Arid1b	17	4995586	6	156778586	1/1	inframe insertion
Arid1b	17	4995925	6	156778928	1/1	inframe deletion
Daxx	17	33912659	6	33320142	1/1	inframe deletion
Pou5f1	17	35508871			1/1	splice donor variant
Tfeb	17	47786091	6	41691081	1/1	inframe insertion
Cyp2c40	19	39807469	10	94775225	2/1	splice donor variant

This table lists genes that were significantly enriched (FDR < 0.01) in one or more of the MAGeCK ((Li et al., 2014)) comparisons between the focus formation sample from the genome-wide CRISPR-Cas9 knockout screen (see section 3.2.2), and the three other samples ("library", "14-day" and "proliferation-only"). 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.

B.7 Determination of NIH3T3 transfection efficiency

Materials

Cell lines

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

Plasmids

pmaxGFP (Lonza, catalogue #VDF-1012)

Reagents

Reagent	Manufacturer
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma-Aldrich
Fetal bovine serum (FBS)	Gibco
Penicillin, streptomycin and L-glutamine (100X, 50mg/mL)	Gibco
Trypsin-EDTA (0.05%)	Gibco
Opti-MEM™ reduced serum media	Gibco
Lipofectamine 3000 kit (Lipofectamine 3000 reagent and P3000)	Thermofisher Scientific
Phosphate-buffered saline (PBS)	Sigma-Aldrich

Table B.7: **Reagents used in the determination of NIH3T3 transfection efficiency**

Method

600,000 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, 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 following quantities of reagents (table B.8). Three wells were transfected with pmaxGFP, and three were mock transfected as a control, with the plasmid DNA replaced with an equivalent volume of Opti-MEM. After 16 hours the media was changed to complete DMEM.

After 72 hours the cells were fixed using 4% paraformaldehyde in PBS for 10 minutes, and centrifuged (200*xg*, 5 minutes). Cells were resuspended in 1% FBS in PBS and protein expression was then assessed using flow cytometry using the following filter/detector: 530/30 (488)-A. Using the mock transfected cells to establish baseline values for negative expression, the mean proportion of cells expressing GFP was determined to be 23.5%.

Reagent	Quantity per 100,000 cells
Lipofectamine 3000 Reagent	1.5μL
P3000	1μL
pmaxGFP	0.5μg

Table B.8: **Transfection reagent quantities for determination of NIH3T3 transfection efficiency**