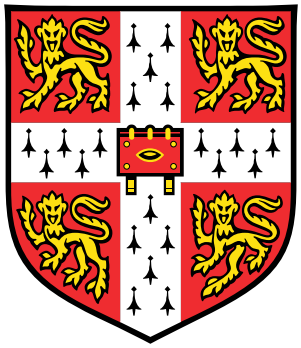
Identifying mediators of malignant transformation in human cancer using genome-wide forward genetic screening approaches



Eleanor Dunstone

Wellcome Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Master of Philosophy

Lucy Cavendish College

2018

I would like to dedicate this thesis to my parents, grandparents and my brother Jack for their ongoing love and support, and to Ben, Lily and Martha for their reassurance and encouragement during the writing process.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Declaration and specified in the text. Additionally, this thesis does not exceed the prescribed word limit for the relevant Degree Committee.

Eleanor Dunstone
2018

Acknowledgements

I would like to thank my supervisor, Dr. David Adams, for the opportunity to work in his group and for his ongoing help and guidance throughout the project. I would also like to acknowledge the other members of the experimental cancer genetics group for their support, with special thanks going to Dr. Marco Ranzani, Dr. Nicola Thompson and Gemma Turner for their advice and assistance. For their role in teaching me the bioinformatics skills required for this project, I would like to thank Aravind Sankar, Rashid Mamunur and Duncan Berger. I would also like to thank Dr. Vivek Iyer and Rashid Mamunur for their assistance in analysing the sequencing data generated from the whole genome sequencing and CRISPR-Cas9 screen. For sequencing library preparation, I would like to acknowledge the Cancer Genomes Project and Dr. Jonathan Cooper. I would also like to thank the cytogenetics team for performing the M-FISH analysis. Finally, I thank the rest of my friends and colleagues at the Wellcome Sanger Institute for making this year a thoroughly enjoyable and fulfilling experience.

Abstract

Malignant transformation is the transition of a cell from a normal state of proliferative homeostasis to a state of abnormal over-proliferation, acting as the initial step of tumourigenesis. This process is governed by the mutation of genes controlling cellular processes such as cell division, DNA replication and growth signaling. In this project, genome-wide forward genetic screening approaches were used to identify novel candidate genes involved in transformation. The model system used was the transformation-sensitive murine cell line NIH3T3, in which genes were assessed for their ability to initiate the formation of transformed foci of proliferation when mutated *in vitro*.

Firstly, the NIH3T3 genome was sequenced to characterise its genetic background and identify possible reasons for its transformation-sensitive phenotype. This was accompanied by Multiplex - Fluorescence *In Situ* Hybridisation to investigate large-scale genomic alterations. These approaches identified specific indels and single nucleotide variants in known cancer-associated genes, and large-scale genomic alterations, both of which may contribute to the transformation sensitivity of the cell line. The karyotype was discovered to be highly abnormal and heterogeneous, suggesting high chromosomal instability and continuing evolution of the line. This work has provided valuable insight into the limitations of this model and has implications for its use in this project and beyond.

The first genetic screening approach used was a pooled CRISPR-Cas9 genome-wide knockout screen, identifying candidate tumour suppressor genes by generating loss-of-function mutations. Genes causing an increase in proliferative focus formation when knocked out were identified by sequencing the guide RNA population and identifying those that were overrepresented in the cultured cells using the algorithm Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout. This identified putative genes associated with transformation, which were then compared with existing mutation data from human cancer sequencing efforts. This screen successfully identified some known cancer-associated tumour suppressor genes, along with potential novel candidates. As a complementary approach, a genome-wide transposon-based screen was also conducted, activating genes by inserting the CMV promoter at random throughout the genome using a *PiggyBac*-based transposon. Recovery of the insertion sites in the final cell population to locate sites that are overrepresented is currently in progress, aiming

to identify putative oncogenes.

While further work to validate the candidates identified is needed, this work has made some progress towards identifying novel transformation-associated genes. If these genes can be validated, they may provide useful insights into the biology of early tumourigenesis, informing further research and possible therapeutic targets.

Contents

Contents xi

List of Figures xv

List of Tables xvii

Nomenclature xviii

1 Introduction 1

1.1 Cancer genetics: a brief overview . . . 1
1.2 Malignant transformation . . . 2
1.3 Models of malignant transformation . . . 3
1.3.1 NIH3T3 cells . . . 5
1.4 Genetic screening tools . . . 5
1.4.1 CRISPR-Cas9 . . . 5
1.4.2 Transposons . . . 7
1.5 Overall aims . . . 8
1.6 Abbreviations . . . 9
1.7 Thesis Overview . . . 9

2 Genomic analysis of NIH3T3 and NIH3T3-Cas9 cells 11

2.1 Introduction . . . 11
2.1.1 Aims . . . 12
2.2 Materials and methods . . . 13
2.2.1 Materials . . . 13
2.2.2 Methods . . . 13
2.3 Results . . . 17
2.3.1 Summary of variants in NIH3T3 wild-type . . . 17

2.3.2 Comparison of NIH3T3 wild-type variants with Cancer Gene Census genes . . . 17
2.3.3 Comparison of NIH3T3 wild-type variants with COSMIC . . . 19
2.3.4 Comparison of NIH3T3 wild-type and NIH3T3-Cas9 . . . 24
2.3.5 Multiplex Fluorescence *In Situ* Hybridisation (M-FISH) of NIH3T3-Cas9 . . . 24
2.4 Discussion . . . 28
2.4.1 Comparison of NIH3T3 wild-type genome with human cancer genome data . . . 28
2.4.2 Large scale genomic alterations in NIH3T3-Cas9 . . . 29
2.4.3 Comparison of single nucleotide variants and indels in NIH3T3 wild-type and NIH3T3-Cas9 . . . 31

3 Identifying mediators of malignant transformation in cancer using genome-wide CRISPR-Cas9 knockout screening 33
3.1 Introduction . . . 33
3.1.1 Aims . . . 35
3.2 Materials and Methods . . . 35
3.2.1 Materials . . . 35
3.2.2 Methods . . . 37
3.3 Results . . . 47
3.3.1 Focus formation assay with pBabe-puro Ras-V12 . . . 47
3.3.2 Genome-wide CRISPR-Cas9 knockout screen for mediators of malignant transformation . . . 49
3.3.3 Screen quality . . . 50
3.3.4 Candidate genes . . . 51
3.3.5 Prioritising genes for validation using existing cancer genome data . . . 51
3.4 Discussion . . . 58

4 Identifying mediators of malignant transformation in cancer using genome-wide transposon-based gene activation 63
4.1 Introduction . . . 63
4.1.1 Aims . . . 64
4.2 Materials and Methods . . . 64
4.2.1 Materials . . . 64
4.2.2 Methods . . . 65
4.3 Results . . . 67

4.4 Discussion 69

5 Conclusions and further directions **71**
5.1 Characterisation of the genetic background of NIH3T3 and NIH3T3-Cas9 . . 71
5.2 Identification of candidate genes associated with malignant transformation . 73

References **75**

A Software and databases **93**

B Supplementary information **95**
B.1 Generation of the NIH3T3-Cas9 cell line 95
B.2 Cas9 activity determination in NIH3T3-Cas9 96
B.3 NIH3T3 wild-type variants with coding consequences overlapping mouse homologues of CGC genes 97
B.4 Verification of amplified Genome-wide Knockout CRISPR Library v2 102
B.5 Primer sequences for CRISPR-Cas9 gRNA insert library preparation 103
B.5.1 1st round PCR - Genome-wide CRISPR-Cas9 knockout screen 103
B.5.2 1st round PCR - Validation (pooled gRNA lentivirus) 103
B.5.3 2nd round PCR 104
B.6 Genes identified by the genome-wide CRISPR-Cas9 knockout screen 104
B.7 Determination of NIH3T3 transfection efficiency 106

List of Figures

2.1 Schematic of NIH3T3 and NIH3T3-Cas9 whole genome sequence analysis . 15
2.2 Putative non-germline variants in NIH3T3 wild-type 18
2.3 Multiplex Fluorescence In Situ Hybridisation of NIH3T3-Cas9 25
2.4 NIH3T3 karyotyping by murine multicolour banding 27

3.1 Schematic of genome-wide CRISPR-Cas9 knockout screen using NIH3T3-Cas9 cells 40
3.2 Focus formation assay using NIH3T3 cells transfected with pBabe-puro Ras-V12 48
3.3 Receiver Operating Characteristic curve for CRISPR-Cas9 knockout screen . 50
3.4 Arrayed focus formation assays for genome-wide CRISPR-Cas9 screen validation 55
3.5 Comparisons between normalised gRNA read counts in genome-wide CRISPR-Cas9 knockout screen 59
3.6 Comparison between normalised gRNA read counts in the plasmid library and focus formation samples 60
3.7 Mean normalised read counts for gRNAs against *Kdsr, Nup160* and *Smu1* . . 60

4.1 Focus formation assay for negative and positive controls of the genome-wide transposon-based activation screen 68

B.1 Histogram of gRNA read counts from sequencing of amplified Genome-wide Knockout CRISPR Library v2 102

List of Tables

1.1 Abbreviations 9

2.1 Reagents used in the methods described in Chapter two 13
2.2 PCR programme for the library preparation for whole genome sequencing of NIH3T3 wild-type and NIH3T3-Cas9 14
2.3 NIH3T3 wild-type variants by consequence 19
2.4 NIH3T3 wild-type variants in mouse homologues of CGC genes 20
2.5 Indels and truncation mutations in mouse homologues of Cancer Gene Census genes in NIH3T3 wild-type 21
2.6 Missense SNVs in mouse equivalents of Cancer Gene Census genes in NIH3T3 wild-type 23
2.7 NIH3T3-Cas9-specific variants by consequence 26
2.8 NIH3T3-Cas9-specific variants overlapping mouse CGC gene homologues . 26

3.1 Reagents used in Chapter 3 37
3.2 Transfection reagent quantities for transfection with Ras and GFP plasmids . 38
3.3 Transfection reagent quantities for Genome-wide Knockout CRISPR Library v2 lentivirus production 39
3.4 Reagent quantities for the 1st round PCR in the CRISPR-Cas9 gRNA insert library preparation 42
3.5 PCR programme for the 1st round PCR in the CRISPR-Cas9 gRNA insert library preparation 42
3.6 Reagent quantities for the 2nd round PCR in the CRISPR-Cas9 gRNA insert library preparation 42
3.7 PCR programme for the 2nd round PCR in the CRISPR-Cas9 gRNA insert library preparation 43
3.9 Transfection reagent quantities for validation focus formation assays 44

3.8 DNA sequences encoding complementary gRNAs from the arrayed plasmid library used in validation 45
3.10 Transfection reagent quantities for validation virus production 46
3.11 Genes containing recurrent deletions in multiple tumour types 53
3.12 Genes for individual validation 53
3.13 Mean normalised read counts for gRNAs against *Kdsr, Nup160* and *Smu1* . . 58

4.1 Reagents used in chapter 4 65
4.2 Transfection reagent quantities used in genome-wide transposon-based gene activation screen 67

A.1 Software and databases used in analyses 93

B.1 Reagents used in the generation of the NIH3T3-Cas9 cell line 95
B.2 NIH3T3 wild-type coding variants in mouse homologues of CGC genes . . . 97
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 . . . 103
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 103
B.5 Primer sequences for the 2nd round PCR in the CRISPR-Cas9 gRNA insert library preparation 104
B.6 Putative transformation-associated genes identified by the genome-wide CRISPR-Cas9 knockout screen 105
B.7 Reagents used in the determination of NIH3T3 transfection efficiency 107
B.8 Transfection reagent quantities for determination of NIH3T3 transfection efficiency 107

Chapter 1

Introduction

1.1 Cancer genetics: a brief overview

Cancer is a collection of diseases involving the abnormal proliferation of cells with the ability to invade other parts of the body, causing 8.8 million deaths per year worldwide (World Health Organisation, 2018). It is now the second leading cause of mortality globally due to increasing worldwide incidence, with an ageing population responsible for much of this phenomenon (Fitzmaurice et al., 2017).

Cancer is fundamentally a genetic disease, with tumour initiation and development governed by the acquisition of mutations in somatic cells during a person's lifetime. A genetic origin of cancer has been hypothesised for over a century, beginning with the observation of inheritance of incorrect chromosomal numbers and subsequent abnormal development in sea urchins by Theodor Boveri, leading him to postulate that the acquisition of similar errors in genetic material may be responsible for the abnormal proliferation seen in tumours (Balmain, 2001; Boveri, 1902). Following the discovery of genes as the units of heredity, the identification of specific genes associated with tumourigenesis began. Genes involved in cancer are often divided into two broad categories; oncogenes that promote tumourigenesis when activated or amplified, and tumour suppressor genes that are associated with cancer when subjected to loss-of-function mutations (Lodish et al., 2000). Oncogenes were first discovered when it was observed that the avian sarcoma virus genes causing malignant transformation in infected cells showed homology to normal avian genes (Stehelin et al., 1976). This illustrated that the genes responsible for tumourigenesis were mutated versions of host genes. Tumour suppressor genes were initially identified through the existence of familial cancer syndromes, where the heterozygous loss-of-function mutation of a gene that protects against cancer in the germline leads to increased susceptibility to a specific range of cancers in affected individuals (Nagy et al., 2004). For example, hereditary retinoblastoma is caused by the mutation of the

gene *RB1* in the germline, leading to the development of multiple retinal tumours during early childhood (Friend et al., 1986). The discovery of this syndrome led to the development of the 'two hit' hypothesis by Knudson in 1971, describing a statistical model of the independent loss-of-function mutation in both alleles of a tumour suppressor gene required to cause a cancer-associated phenotype (Knudson, 1971).

Recent developments in molecular biology have driven an increased understanding of the genes and biological processes involved in tumourigenesis. For example, the use of microarray technology in the analysis of human cancer samples enabled the identification of genes associated with cancer through the study of genes exhibiting copy number changes in the genome (Albertson et al., 2000), or showing alterations in gene expression (Perou et al., 2000). The advent of next-generation sequencing has revolutionised cancer gene discovery, allowing large-scale identification through the sequencing of ever-increasing numbers of tumour samples (Martincorena et al., 2017). The primary disadvantage of these genome-wide approaches that utilise patient cancer samples is that they work on the principle that functionally important genes in cancer development will be mutated more frequently than expected, and while this establishes an association between a mutation and a phenotype, it cannot determine how or when this mutation affects tumourigenesis.

Functional screening can be used to complement next-generation screening by experimentally generating mutations and identifying those that give rise to the desired phenotype. A variety of techniques have been historically used for mutation generation, including chemical or radiation-based mutagens, and transposon-based insertional mutagenesis (Friedrich et al., 2017; Moresco et al., 2013). More recently, the development of CRISPR-Cas9 genome editing technology has enabled efficient, homozygous loss-of-function mutation at specific loci, facilitating forward genetic screening for a variety of phenotypes (Koike-Yusa et al., 2013). The advantage of these experiments is that they identify genes through functional assays, showing a causal relationship between a phenotype and a mutation. The combined power of functional screening and the increasing scale of next-generation sequencing of patient tumours is likely to fuel the discovery of new cancer-related genes in future.

1.2 Malignant transformation

Malignant transformation is the the initial step in tumourigenesis, when a normal cell acquires the characteristics of cancer. Normal tissues maintain growth homeostasis by balancing proliferation, differentiation and cell death at the tissue level (Biteau et al., 2011). Malignant cells overcome these controls, allowing overproliferation and the formation of a clonal expansion, generating a tumour. This requires the disruption of multiple cellular mechanisms, producing a

set of phenotypes known as the 'hallmarks of cancer'. These consist of sustaining proliferative signalling, evading growth suppression, activating invasion and metastasis pathways, enabling replicative immortality, inducing angiogenesis and resisting cell death, with deregulation of cellular energetics and avoidance of immune detection emerging more recently as additional important traits (Hanahan and Weinberg, 2011). Genetic instability and tumour-promoting inflammation act as enabling characteristics that facilitate acquisition of these hallmarks, by increasing mutation rate and proliferation and generating genetic diversity for clonal selection to act upon.

Transformation occurs by the mutation of genes associated directly or indirectly with proliferative control, working cooperatively to generate these cancer-associated phenotypes. Commonly dysregulated pathways include growth signalling pathways such as RAS-RAF-MEK-ERK (Li et al., 2016), pathways governing proliferation and apoptosis such as PI3K-AKT (Liu et al., 2009), and those involved in cell cycle control, for example RB-E2F (Nevins, 2001). Some genes such as *TP53* influence all of these cellular functions and more, acting as hubs for the integration of cellular proliferation signalling (Lane and Levine, 2010).

Mutation of these genes occurs by a number of mechanisms, including exposure to exogenous mutagens such as ultraviolet light and chemical carcinogens (Brash et al., 1991; Miller and Miller, 1981), chronic inflammation (Coussens and Werb, 2002), or stochastic failures of DNA replication (Tomasetti et al., 2017), the rate of which increases with age (Milholland et al., 2015).

1.3 Models of malignant transformation

In order to study the earliest stages of tumourigenesis, tractable models of malignant transformation are required. *In vitro* models using cell lines are easily manipulated, allowing the introduction of genetic material via transfection or transduction, or mutations using genome editing. Immortalised but non-tumourigenic human and mouse cell lines have provided a valuable model whereby tranformation can be observed *in vitro*. For example, MCF-10A and NIH3T3 have been shown to transform in response to oncogene overexpression (Giannakourous et al., 2015; Wasylishen et al., 2011). The practicality and relatively low cost of using such cell lines often makes them the most suitable model for genome-wide screening and other high-throughput approaches. However, the utility of these models is limited by their differences from the 'normal' cells in which tumourigenesis occurs *in vivo,* which have been observed at the genetic, epigenetic, transcriptomic and phenotypic level (Hughes et al., 2007; Pellacani et al., 2016).

Induced pluripotent stem cells (iPSCs) have been used as an alternative to established

cell lines, acting as a more representative model of the somatic cells of origin of specific tumour types. For example, iPSC-derived neural progenitor cells have been transformed into glioma-initiating cells through dysregulation of receptor tyrosine kinase and p53 signalling (Sancho-Martinez et al., 2016). However, another limitation of cell-based models is that they do not fully recapitulate tumourigenesis due to the absence of the components of the cancer microenvironment, such as the immune system, that can affect the ability of a transformed cell to actually form a tumour *in vivo.*

Mouse models have the advantage of demonstrating transformation *in situ,* taking into account non-cell autonomous factors. Mouse genomes can be easily genetically manipulated in site- and time-specific ways, making them a powerful model for the investigation of early events in tumourigenesis (Balani et al., 2017). Sophisticated models allow the induction of mutations in specific cell lineages, identifying transforming mutations and cells of origin in multiple tumour types (Blanpain, 2013). However, caution must be exercised when using mouse models to make conclusions about human disease, as there are relevant genetic and physiological differences between the two species. For example, it has been shown that mouse cells incompletely recapitulate human haematopoeitic oncogenesis, partially as they are more easily transformed than their human counterparts (Beer and Eaves, 2015). One way to ameliorate this issue is to combine mouse and human models by introducing human cells into immunodeficient mice, studying the ability of cells that have been modified *in vitro* to form tumours *in vivo.* For example, the injection of non-cancerous patient-derived prostate basal cells engineered to express activated AKT and ERG and the androgen receptor has been shown to lead to the development of prostate cancer in mice (Goldstein et al., 2010). However, a further disadvantage of *in vivo* models is their lack of scalability for screening approaches.

An alternative to the use of models is the study of mutation profiles in human cancers, deconstructing the evolution of the tumour to identify the mutation(s) responsible for the initial transformation event (Aparicio and Caldas, 2013; Shlush et al., 2014). The advantage of this is that these mutations have occurred in real cases of the disease, so results are more likely to be biologically relevant. However, this approach is inherently observational, and therefore limited to detectable mutations in the available samples. This means that, unlike in model organisms, mutations cannot be experimentally manipulated to probe their effects. Additionally, the deconvolution of the mutational history of a tumour is technically challenging, especially as most cancers are detected at late stages (Cancer Research UK, 2014), by which time they have many mutations and show high genetic heterogeneity. In addition to driver mutations that promote tumour-associated phenotypes, the high mutation rate seen in many cancers leads to the acquisition of many passenger mutations (Pon and Marra, 2015). This makes identifying the mutation(s) that mediated the initial transformation difficult, especially as they may not

still be present at high frequency in the tumour.

1.3.1 NIH3T3 cells

NIH3T3 is a mouse embryonic fibroblast cell line generated in 1969 from desegregated NIH Swiss mouse embryo fibroblasts (Jainchill et al., 1969), using the same method used to generate the cell line 3T3 (Torado and Green, 1963). These cells spontaneously immortalised in culture and became tetraploid shortly after establishment (Torado and Green, 1963). Subsequent work has descibed the cells at the cytogenetic level, using multicolour banding fluorescence *in-situ* hybridisation to characterise the NIH3T3 genome at high resolution (Leibiger et al., 2013). This study found that the genome is predominantly tetrasomic (60%), but that its ploidy varies across different sites, showing a complex karyotype with four derivative chromosomes appearing since its previous characterisation in 1989 (Kasid et al., 1989).

NIH3T3 cells are sensitive to malignant transformation, transforming readily in response to overexpression of an oncogene (Giannakourous et al., 2015; Rao et al., 2014). This phenotype makes them ideal for use as an *in vitro* model of tumourigenesis, allowing the detection of mutations that are able to induce malignant transformation. In this project, NIH3T3 cells were used in genome-wide forward genetic screening approaches to identify novel genes putatively associated with transformation. The NIH3T3 genome is poorly characterised at the individual gene level, and the genetic background of the model may affect the outcome of these screens. Therefore, the mutational landscape of the cells was characterised in chapter two using whole-genome sequencing.

1.4 Genetic screening tools

1.4.1 CRISPR-Cas9

CRISPR-Cas9 based immunity in prokaryotes

Clustered regularly interspaced palindromic repeats (CRISPR) are genomic features found in certain bacteria and archaea. These sequences are one component of a prokaryotic adaptive immune mechanism that allows recognition and destruction of viral nucleic acid sequences based on previous encounters with the same sequences. The second key component of this system is the CRISPR-associated protein 9 (Cas9) endonuclease, which is directed to the targeted viral sequences by homologous RNAs produced at the CRISPR sites, cleaving the viral sequence and rendering it unable to perform its function (Barrangou et al., 2007).

In prokaryotes, two RNA components are generated from the CRISPR locus. The crRNA

contains sequences homologous to the viral invaders that the CRISPR locus is derived from, whereas the *trans*-activating CRISPR RNA (tracrRNA) is transcribed from a locus upstream, and contains a region complementary to the repeat region of the CRISPR locus. This allows binding to the crRNA, creating a double-stranded RNA which is cleaved by RNaseIII, before associating with Cas9 to form an active ribonucleoprotein complex. In the presence of 3' Protospacer Adjacent Motifs (PAMs) in the viral DNA, Cas9 then cleaves DNA at sequences that bind to the crRNA (Deltcheva et al., 2011).

CRISPR-Cas9 based gene editing

This system has now been adapted for use in the genomic editing of a variety of cell types. An engineered single guide RNA (sgRNA/gRNA) combines the functions of the crRNA and tracrRNA, targeting Cas9 to the desired genomic sequence. Cas9 derived from *Streptococcus pyogenes* is the most commonly used, creating double-stranded DNA breaks. This induces error-prone repair by non-homologous end joining, causing insertions or deletions, leading to loss-of-function mutation. For gene knockout, gRNAs are designed to target early, constitutively-expressed exons, producing a null phenotype (Jinek et al., 2013). This system has since been modified in a variety of ways to produce a wide range of effects in a sequence-specific manner. For example, engineered CRISPR-Cas9 based systems can now be used for gene activation, individual base-pair editing and epigenome modification (Adli, 2018).

Genome-wide CRISPR-Cas9 knockout screening

Methods for genome editing before the development of CRISPR-Cas9 technology, such as transcription activator-like effector nucleases (TALENs) and zinc-finger nucleases, were limited by the need to design a new set of proteins for each target sequence. The relative ease of using CRISPR-Cas9 due to the requirement only for a complementary oligonucleotide has revolutionised the field due to the improved speed, cost and scalability (Adli, 2018). One of the most powerful applications of large-scale CRISPR-Cas9 genome editing is genome-wide knockout screening. This approach employs gRNA sequence libraries targeting genes across the whole genome, allowing non-biased screening for a range of phenotypes. Libraries can be pooled (Koike-Yusa et al., 2013) or arrayed (Metzakopian et al., 2017), and are most commonly introduced into cells using lentiviral vectors. Pooled libraries are less labour-intensive to use for whole-genome screening, however a sequencing step is required for hit identification, and they are not suitable for all functional assays. In pooled screens, genes of interest are usually identified by sequencing of the cell population and detection of gRNA sequences that are either enriched or depleted, which can be used to identify genes that suppress or are

essential for the studied phenotype. In chapter three of this project, a pooled knockout screen was used to identify gRNA sequences enriched in cells that have undergone malignant transformation, indicating that the genes they target may have tumour suppressor function in the early stages of tumourigenesis.

1.4.2 Transposons

Transposons as a biological tool

Transposons are mobile genetic elements that are ubiquitous components of metazoan genomes, with sequences derived from them making up 45% of the human genome (Lander et al., 2001). There are two classes of transposon: retrotransposons that are mobilised via an RNA intermediate using a 'copy-and-paste' mechanism, and DNA transposons that use a 'cut-and-paste' process, excising the original copy from the genome. Both classes require a transposase enzyme that cuts, ligates and rejoins the DNA during this process (Ivics et al., 2009). Transposons have been widely used as a tool in molecular biology, taking advantage of their ability to insert chosen DNA sequences into the genomes of model organisms *in vivo* (Mátés et al., 2007). They have been used for applications such as the production of transgenic model organisms, and random insertional mutagenesis in forward genetic screening.

The transposons used in molecular biology are mostly of the DNA-based class. Naturally occuring versions encode a transposase in-between inverted terminal repeats (ITRs) that contain binding sites for the transposase. Experimentally, the transposase is usually supplied in *trans*, with the sequence of choice lying between the ITRs; this sequence can now be inserted efficiently into the experimental host genome (Ivics et al., 2009). Initially the use of transposons was limited to lower organisms such as *Drosophila* that have retained active DNA transposons in their genomes, however subsequent developments have produced modified systems such as *Sleeping Beauty* and *PiggyBac* that have high activity in mammalian systems (Ding et al., 2005; Ivics et al., 1997).

Transposon-based whole-genome screening

The advantage of transposons for genome-wide screening is that, unlike lentiviral approaches, transposons do not show tissue tropism so are more widely applicable to a range of cell types and tissues, both *in vitro* and *in vivo.* Transposon insertions are easily recovered by sequencing using their specific molecular characteristics, allowing for the quantification of insertion sites (Friedrich et al., 2017). In addition to insertional mutagenesis that generates loss-of-function mutations, modified transposons can be used to create a range of genome-wide modifications. For example, transposons carrying combinations of promoter and enhancer elements have

been used in mice to identify novel cancer-associated genes (Rad et al., 2010). In chapter four of this project, a transposon-based approach was be used to insert the cytomegalovirus (CMV) promoter into the NIH3T3 genome at random, increasing expression of downstream genes. This approach aimed to identify putative oncogenes that are able to mediate malignant transformation when overexpressed.

1.5 Overall aims

- To characterise the genome of cell line NIH3T3, investigating possible genetic causes behind its tranformation-sensitive phenotype.

- To compare the genomes of NIH3T3 wild-type and the daughter cell line NIH3T3-Cas9, identifying any genetic divergence that has taken place in culture and potential phenotypic effects of this.

- To identify candidate genes involved in transformation using a genome-wide CRISPR-Cas9 knockout screen in NIH3T3-Cas9.

- To identify candidate genes involved in transformation using a genome-wide transposon-based activation screen in NIH3T3.

- To compare these candidate genes with genes identified in existing cancer genome data and prioritise candidates for validation and further investigation.

1.6 Abbreviations

CGC	Cancer Gene Census
COSMIC	Catalogue Of Somatic Mutations In Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DMEM	Dulbecco's Modified Eagle's Medium
FBS	Fetal Bovine Serum
gRNA	Guide RNA
MAGeCK	Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout
M-FISH	Multiplex - Fluorescence *In Situ* Hybridisation
PBS	Phosphate Buffered Saline
SNV	Single Nucleotide Variant
TCGA	The Cancer Genome Atlas
VCF	Variant Call Format
VEP	Variant Effect Predictor (Ensembl)

Table 1.1: **Abbreviations**

1.7 Thesis Overview

In Chapter two, I describe the acquisition and analysis of the whole-genome sequence data obtained for NIH3T3 wild-type and its modified daughter cell line NIH3T3-Cas9. In brief, this consists of the identification of single nucleotide variants and indels in NIH3T3, assessment of their possible effects in coding regions, and cross-referencing of their locations with genes listed in the Cancer Gene Census and mutations catalogued in COSMIC to investigate variants that may play a role in the transformation-sensitive phenotype of the cell line. Additionally, this chapter describes the comparison of the genome of NIH3T3 with NIH3T3-Cas9 to identify any genetic differences between them, and any possible phenotypic effects of these, assessing the suitability of NIH3T3-Cas9 as a model in the CRISPR-Cas9 screen discussed in Chapter 3. NIH3T3-Cas9 was also karyotyped using Multiplex - Fluorescence *In Situ* Hybridisation to identify large scale genomic alterations.

In Chapter three the design of the genome-wide CRISPR-Cas9 knockout screen for genes involved in transformation is detailed, along with the analysis of the generated data using MAGeCK (Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout) and prioritisation of candidate genes using existing cancer genome data. Following this, efforts to validate these candidate genes are described. Chapter four covers the genome-wide transposon-based

activation screen for genes involved in transformation and future plans for the analysis of these data.

The final chapter summarises the results of the previous chapters, discussing possible further directions and wider implications of this work.

Chapter 2

Genomic analysis of NIH3T3 and NIH3T3-Cas9 cells

2.1 Introduction

NIH3T3 is a mouse embryonic fibroblast cell line, originally generated from a single cell from a mouse from the NIH Swiss strain in 1969. The cells are spontaneously immortalised but not transformed, and are sensitive to malignant transformation in culture (Jainchill et al., 1969), making them ideal for use in forward genetic screening for this phenotype. However, a limitation of using this cell line as a model is that established cell lines are not fully representative of normal organisms, due to the presence of mutations acquired during culture. For example, in NIH3T3, mutations would have been required to overcome replicative senescence to spontaneously immortalise the cell line. Existing mutations present in the cell line have the potential to affect the results of the forward genetic screening approaches used in this project in two ways. Firstly, existing mutations may interact with those induced during the screens. Secondly, the transformation-sensitive nature of this cell line means that it may have already acquired some of the properties of cancer, potentially limiting the range of genes that can be mutated to cause transformation *in vitro*. In order to investigate these possibilities, the genetic background of NIH3T3 wild-type cells was characterised using whole genome sequencing.

The aim was to analyse the genetic variants present in this cell line in order to identify those that may be responsible for its transformation-sensitive phenotype. This was done by comparing these variants with known cancer-associated genes in the Cancer Gene Census (CGC) (Futreal et al., 2004) and mutations listed by the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Forbes et al., 2017). Additionally, the cell line NIH3T3-Cas9, which expresses Cas9 as a transgene (see appendices B.1 and B.2), was investigated by Multiplex Fluorescence

In Situ Hybridisation (M-FISH) in order to identify mutations such as translocations, and large scale amplifications and deletions. Together, this information should help to inform interpretation of the screen results, and assess the suitability of this cell line as an *in vitro* model for malignant transformation.

Previous work on characterisation of NIH3T3 has shown that the cell line is predominantly tetraploid, with widespread chromosome gains and losses and five derivative chromosomes (Leibiger et al., 2013). Cytogenetic analysis using M-FISH has been compared with these results, allowing identification of chromosomal-scale variation within the cell line, and potential karyotypic evolution over time, indicating chromosomal instability (section 2.3.5).

The NIH3T3-Cas9 cell line was also sequenced to identify differences between its genome and that of its parental cell line NIH3T3 wild-type, which have been cultured independently for an estimated 20 passages. The aim was to identify mutations that have occurred since the establishment of NIH3T3-Cas9, quantifying how much genetic drift has taken place and if this may affect the characteristics of the cell line relative to NIH3T3 wild-type.

2.1.1 Aims

Overall aim: To determine the genetic background of NIH3T3 and NIH3T3-Cas9, characterising the models used in the forward genetic screening approaches applied in this project (Chapters 3 and 4).

1. To identify small genetic variants (single nucleotide variants and indels) found in NIH3T3 compared to the mouse reference genome.

2. To compare these variants with mutations found in the CGC (Futreal et al., 2004) and the COSMIC database (Forbes et al., 2017) to investigate the reasons for the transformation-sensitive phenotype of NIH3T3.

3. To determine the karyotype of NIH3T3-Cas9 and identify genomic changes such as translocations and large-scale amplifications and deletions.

4. To compare variants in NIH3T3 wild-type and NIH3T3-Cas9 to identify any genetic divergence that has occurred between the two cell lines and its possible effects on the use of the NIH3T3-Cas9 as a model for transformation.

2.2 Materials and methods

2.2.1 Materials

Cell lines

NIH3T3 wild-type 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 Thompson from the experimental cancer genetics group at the Wellcome Sanger Institute (appendix B.1).

Reagents

Reagent	Manufacturer
Agencourt AMPure XP SPRI beads	Beckman Coulter
Blood & Cell Culture DNA Mini Kit	Qiagen
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma Aldrich
Fetal bovine serum (FBS)	Gibco
KAPA HiFi HotStart ReadyMix 2X	Kapa Biosystems
NEBNext Ultra II DNA Library Prep Kit	New England Biolabs
Penicillin, streptomycin and L-glutamine (100X, 50mg/mL)	Gibco
Trypsin-EDTA (0.05%)	Gibco

Table 2.1: **Reagents used in the methods described in Chapter two**

2.2.2 Methods

DNA extraction

Cells were cultured in complete DMEM (DMEM supplemented with 10% FBS and 500µg/mL penicillin, streptomycin and L-glutamine), then detached using 0.05% trypsin-EDTA, centrifuged (200*xg*, 5 minutes) and frozen at -80°C . Genomic DNA was extracted using Qiagen Blood & Cell Culture DNA Mini Kit according to the manufacturer's instructions.

Library preparation

Library preparation was performed with the assistance of the Cancer Genome Project at the Wellcome Sanger Institute. DNA (200ng/120µl) was sheared to 450bp using a Covaris LE220

instrument and purified using Agencourt AMPure XP SPRI beads. Libraries were constructed using the NEBNext Ultra II DNA Library Prep Kit. Polymerase chain reactions (PCRs) were set up using KAPA HiFi Hot Start Mix and IDT 96 iPCR tag barcodes. DNA was amplified using the protocol in table 2.2. Post-PCR samples were purified using Agencourt AMPure XP SPRI beads.

Cycle number	Denaturing	Annealing	Extension
1	95°C, 5 minutes		
2-7	98°C, 30 seconds	65°C, 30 seconds	72°C, 1 minute
8			72°C, 10 minutes

Table 2.2: **PCR programme for the library preparation for whole genome sequencing of NIH3T3 wild-type and NIH3T3-Cas9**

Whole genome sequencing

Sequencing was performed using Illumina-B HiSeq X paired-end sequencing. The mean coverage acheived was 37.2 for the NIH3T3 wild-type sample and 37.8 for the NIH3T3-Cas9 sample.

Analysis

The analysis of the whole genome sequence data generated from NIH3T3 wild-type and NIH3T3-Cas9 is summarised in figure 2.1.

Variant calling Variant calling was performed with the assistance of Rashid Mamunur from the experimental cancer genetics group at the Wellcome Sanger Institute. Samtools (Li et al., 2009) mpileup (parameters: -C50 -pm3 -F0.2 -d2000 -L500 -r 10:0-50000000) followed by BCFtools (Danecek et al., 2011) call were used to call single nucleotide variants (SNVs) and indels, producing a variant call format (VCF) file.

Filtering Variants were filtered to remove those found in the cell line due to their presence in the NIH Swiss mouse germline. Variants found in 36 inbred mouse strains were obtained from the Mouse Genomes Project (Adams et al., 2015). Data from the Castaneus and Spretus strains (wild mouse) were discarded, leaving only those derived from 34 laboratory strains. This was used to filter the VCF file generated above using BCFtools isec (Li et al., 2009). This left variants found in the cell lines, but not in the mouse variant files, removing variants at sites known to be polymorphic between strains of laboratory mice.

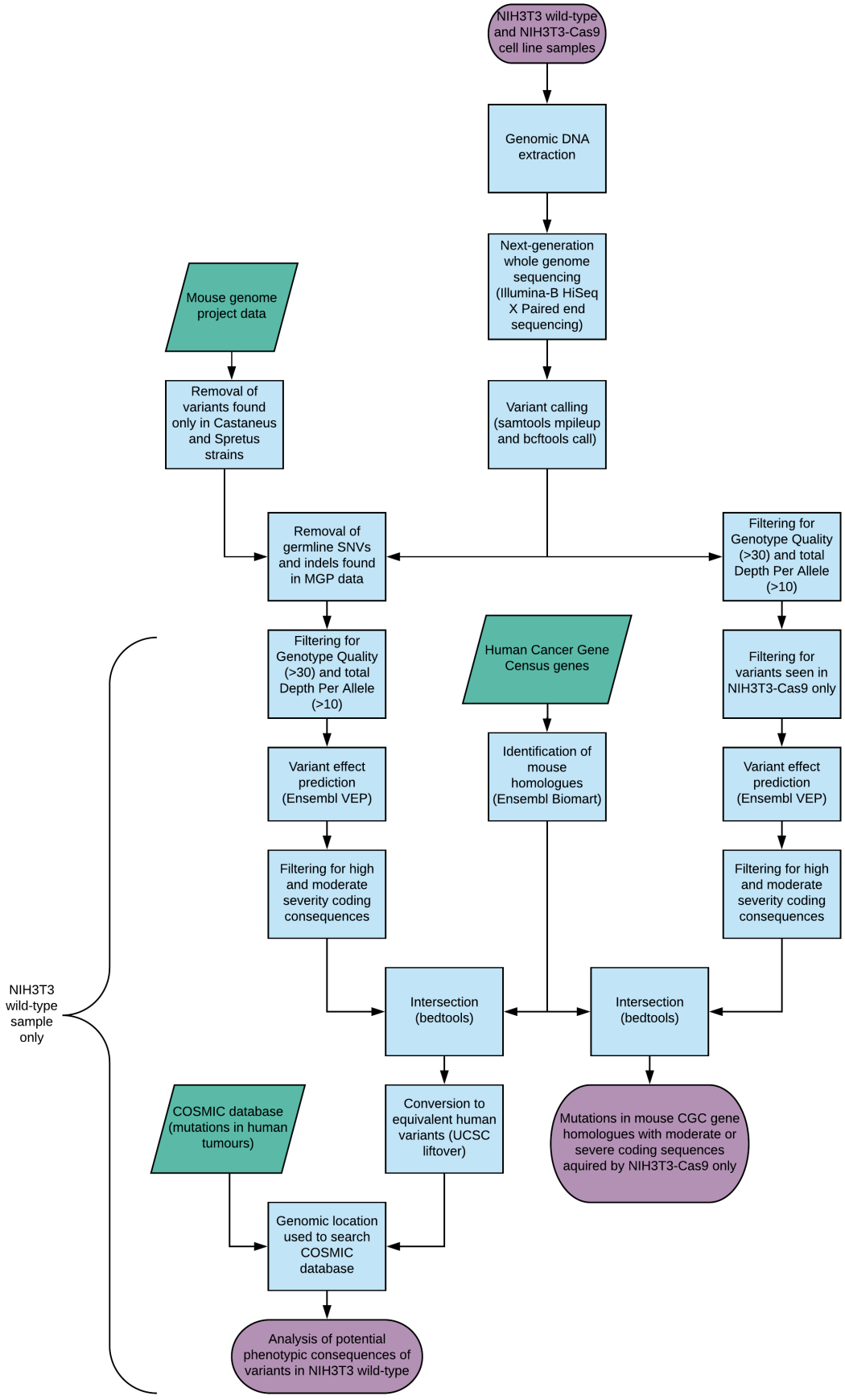


Figure 2.1: **Schematic of NIH3T3 and NIH3T3-Cas9 whole genome sequence analysis**

Variants were filtered for genotype quality (GQ > 30) and total depth per read (total DPR across all alleles > 10) using BCFtools filter (Li et al., 2009).

Prediction of variant effects For the NIH3T3 wild-type sample only, the effects of the variants were determined using Ensembl Variant Effect Predictor (VEP) standalone Perl script (McLaren et al., 2016). Variants were filtered for consequence severity using the filter_vep command, leaving only 'high' and 'moderate' severity variants. Consequence severity is based on the assignment of a Sequence Ontology (Eilbeck et al., 2005) term to a variant, describing its effect on a given transcript. The 'high' and 'moderate' severity consequence Sequence Ontology terms used were missense_variant, inframe_deletion, inframe_insertion, transcript_amplification, stop_lost, frameshift_variant, stop_gained, splice_acceptor_variant, start_lost, protein_altering_variant, splice_donor_variant and transcript_ablation.

Comparison with human cancer genome data A list of the 1439 CGC (Futreal et al., 2004) genes was obtained from COSMIC (Forbes et al., 2017). The Ensembl Gene Stable IDs of these genes were used to find mouse orthologues using Ensembl Biomart (Kinsella et al., 2011), retrieving 803 mouse genes. The genomic coordinates of these genes were obtained from Ensembl and used to create a Browser Extensible Data (BED) file containing the mouse CGC gene homologues. Bedtools intersect (Quinlan and Hall, 2010) was used to intersect the locations of these genes with the NIH3T3 wild-type variants with 'high' or 'moderate' severity coding consequences, generating a list of variants found within or overlapping the mouse CGC gene homologues.

In order to compare these variants with those listed in the COSMIC database, they were converted from the mouse GRCm38 assembly to the equivalent genomic coordinates in the human GRCh38 assembly using the Genome Browser LiftOver tool from the University of California Santa Cruz (Kent et al., 1976) (parameters: minimum ratio of bases that must remap = 0.1, minimum hit size in query = 0, minimum chain size in target = 0, minimum ratio of alignment blocks or exons that must map = 1).

The phenotypic consequence of the indels was inferred from the Sequence Ontology term previously assigned to the variant where possible. For the SNVs, the COSMIC database was searched for SNVs at this position.

Comparison between NIH3T3 wild-type and NIH3T3-Cas9 Variants from NIH3T3 wild-type and NIH3T3-Cas9 were filtered for genotype quality (GQ > 30), and total depth per read (total DPR across all alleles > 10 for both samples) using BCFtools filter (Li et al., 2009). This file was then filtered to contain only records where the NIH3T3 wild-type sample

showed the reference allele, whereas the NIH3T3-Cas9 sample showed an alternate allele. Mouse germline SNVs and indels from the 36 strains described in the Mouse Genomes Project ((Adams et al., 2015)) were removed as described above (see 2.2.2).

Ensembl VEP was used to filter these results for 'high' and 'moderate' severity coding consequences as described previously (see 2.2.2).

The resulting variants were intersected with the mouse CGC gene homologues using Bedtools (Quinlan and Hall, 2010) intersect to determine if any overlapped these genes.

Multiplex Fluorescence *In Situ* Hybridisation (M-FISH)

M-FISH analysis of NIH3T3-Cas9 cells was performed with assistance from the cytogenetics team at the Wellcome Sanger Institute using 21-colour mouse chromosome specific DNA probes (Geigl et al., 2006).

2.3 Results

2.3.1 Summary of variants in NIH3T3 wild-type

After filtering for genotype quality (GQ > 30) and total depth per read (total DPR > 10), and removing variants likely to be germline variants present in the parental mouse strain, the total number of variants called in the NIH3T3 wild-type cell line was 1,107,940. However, due to the absence of the strain-matched control (Swiss) in the mouse genome database used to filter out known germline variants, it is likely that some remain. Of these variants, 203,395 were single nucleotide variants (SNVs), and 904,347 were indels.

The vast majority of variants were in non-coding regions, as shown by figure 2.2, with coding variants making up only 0.37% of the total (contained within 'other' in figure 2.2).

2.3.2 Comparison of NIH3T3 wild-type variants with Cancer Gene Census genes

The variants from section 2.3.1 were then filtered for consequence severity using the filter_vep command (McLaren et al., 2016). Consequence severity is based on the assignment of a Sequence Ontology (Eilbeck et al., 2005) term to a variant, describing its effect on a given transcript. Filtering for only variants with a 'high' or 'moderate' consequence (see section 2.2.2), left 2018 variants. The numbers of variants are listed by consequence in table 2.3 (numbers do not total 2018 as some variants affected multiple transcripts, causing different coding consequences).

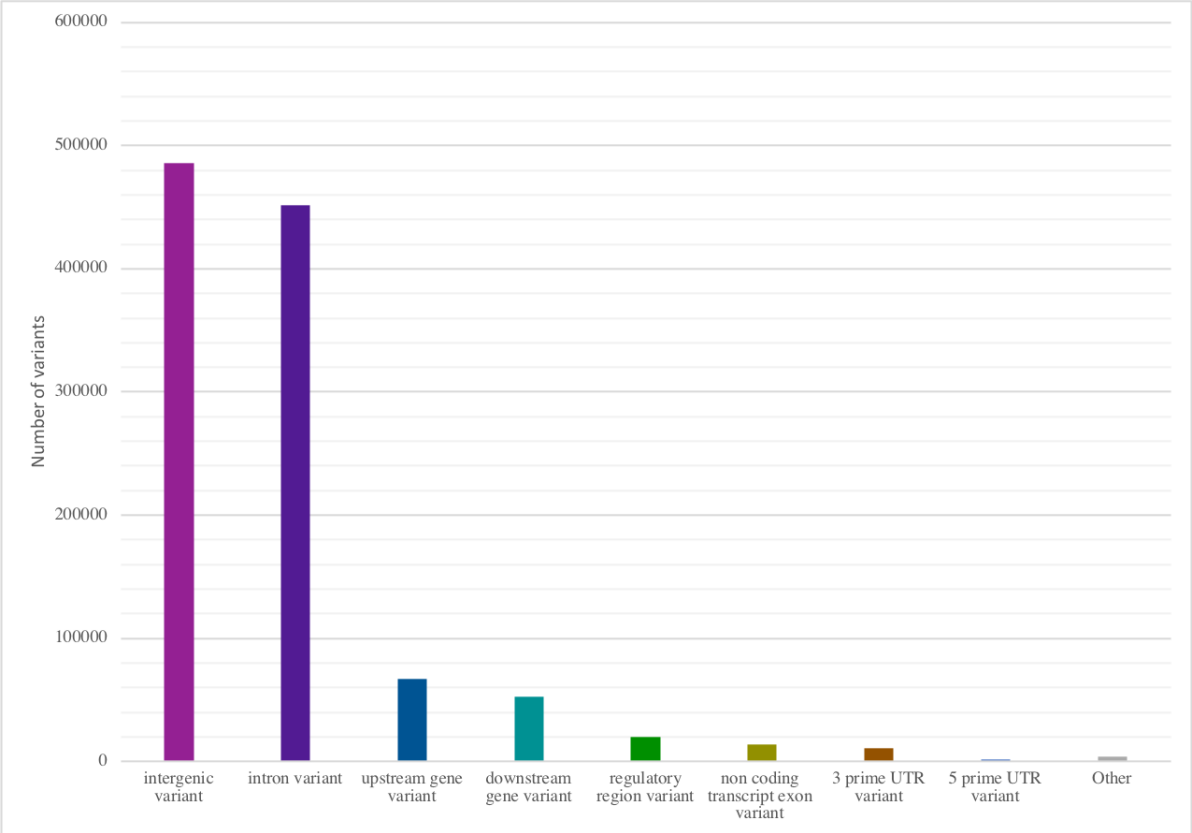


Figure 2.2: **Putative non-germline variants in NIH3T3 wild-type by consequence type** Non-germline variants in NIH3T3 wild-type were identified by filtering to exclude variants at positions that were polymorphic in the Mouse Genome Project (Adams et al., 2015) data on 36 inbred laboratory mouse strains. These variants were also filtered for quality (genotype quality >30 and total depth per read > 10). Variants are grouped by consequence, as assigned by the Ensembl Variant Effect Predictor (McLaren et al., 2016).

Consequence	Number of variants
missense variant	1095
inframe deletion	241
frameshift variant	203
inframe insertion	162
splice acceptor variant	132
splice donor variant	121
protein altering variant	44
stop gained	43
stop lost	9
start lost	2

Table 2.3: **NIH3T3 wild-type variants by consequence**
Numbers of 'high' and 'moderate' severity variants in NIH3T3 wild-type as defined by Ensembl Variant Effect Predictor (McLaren et al., 2016), categorised by consequence based on Sequece Ontology term (Eilbeck et al., 2005).

The list of variants with 'moderate' and 'high' severity coding consequences was then intersected with 803 mouse homologues of Cancer Gene Census (CGC) genes. This gave 88 variants in 69 genes (see appendix B.3 for full list). The numbers of these variants, by consequence, are listed in table 2.4. The parental mouse strain (NIH Swiss) is not included in the Mouse Genomes Project (MGP) data, therefore some germline variants may remain. For example, where many variants are present in one gene it is likely that this represents a haplotype in NIH Swiss mice that is different to that seen in the strains included in the MGP data. On this basis, variants in the *Muc4* gene were excluded from the analysis. Additionally, seven of the variants on chromosome 17 are found in members of the murine Major Histocompatibility Complex gene group, which are likely to be polymorphic between mouse strains. For this reason, variants in *H2-D1, H2-Q4, H2-Q7, H2-T23, H2-Bl, H2-T10, H2-T3* and *H2-M11* were discarded.

2.3.3 Comparison of NIH3T3 wild-type variants with COSMIC

The variants discovered in the mouse homologues of CGC genes were then investigated to try and determine their phenotypic consequences.

Sequence ontology term	Number of variants
missense variant	45
inframe deletion	29
frameshift variant	5
splice donor variant	4
protein altering variant	4
splice acceptor variant	3
stop gained	1

Table 2.4: **NIH3T3 wild-type variants in mouse homologues of CGC genes by consequence**
Numbers of 'high' and 'moderate' severity variants in NIH3T3 wild-type as defined by Ensembl Variant Effect Predictor (McLaren et al., 2016), that intersect mouse homologues of the Cancer Gene Census genes (Futreal et al., 2004). These genes are categorised by Sequece Ontology term (Eilbeck et al., 2005).

Indels and truncating mutations

For the indels (table 2.5), the Sequence Ontology (Eilbeck et al., 2005) terms assigned by Ensembl VEP (McLaren et al., 2016) were used to determine their phenotypic consequence where possible. Indels assigned the terms frameshift_variant, splice_donor_variant and splice_acceptor_variar along with a nonsense mutation (stop_gained), were considered to lead to a potential loss-of-function phenotype. The inframe_deletion, inframe_insertion and protein_altering_variant categories are harder to assign a functional consequence to based on these terms alone, and require further investigation.

The *Tpr, Met, Etnk1*, *Nup98, Msi2, Il6st, Pou5f1* and *Cyp2c40* genes have a predicted loss-of-function mutation present homozygously and are therefore the most likely candidates to have a phenotypic effect. A literature search concerning the functions of these genes in cancer suggested that *Tpr* (David-Watine, 2011), *Met (Tovar and Graveel, 2017), Etnk1* (Lasho et al., 2015), *Nup98 (Gough et al., 2011)* and *Msi2 (Li et al., 2015)* appear to act as oncogenes. Loss-of-function mutation of these genes would therefore not be typically expected to cause a cancer-associated phenotype. However, high expression of *IL6ST* in triple-negative breast cancer is associated with improved outcomes (Mathe et al., 2015), suggesting that it may act as a tumour suppressor gene, making it more likely to contribute to transformation-sensitivity in NIH3T3. Another possible tumour suppresor gene is *Cyp2c40,* which has been reported to produce anti-inflammatory metabolites in colon cancer (Albert and Bennett, 2012). However, this is less likely to have a tumour suppressing effect in the absence of immune cells *in vitro*. *Pou5f1* (also known as *Oct4)* has been reported to suppress metastatic potential in breast cancer cells (Shen et al., 2014), and promote tumourigenesis in cervical cancer cells (Wang et al.,

2013), indicating that it is not easily categorised as a tumour suppressor gene or an oncogene. This may also be the case for some of the other genes in this list, especially those that are less well characterised.

Table 2.5: **Indels and truncation mutations in mouse homologues of Cancer Gene Census genes in NIH3T3 wild-type**

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 indels, and SNVs that are predicted to cause a truncation of the protein, along with their positions in the NIH3T3 wild-type genome (GRCm38), and equivalent locations in the human genome (GRCh38) when remapped using the Genome Browser LiftOver tool from the University of California Santa Cruz (Kent et al., 1976). Where the 'human equivalent' columns are blank, this tool was unsuccessful at mapping this variant to the human genome. Genotypes: 0/1 = heterozygous reference and alternate allele, 1/1 = homozygous alternate allele, 2/1 = heterozygous first alternate allele and second alternate allele (for reference and alternate alleles for each variant, see Appendix B.3). Sequence Ontology (Eilbeck et al., 2005) terms 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.

Missense SNVs

All remaining SNVs were successfully mapped to the human genome (see table 2.6). The COSMIC (Forbes et al., 2017) database was then searched for mutations at these positions. For eight of the missense SNVs, at least one mutation was listed in the database at this position (mutations found in *Ptprc, Csmd3, Ptprd, Robo2, Gli1, Sirpb1b* and *Sirpb1c*). For details of these variants, see appendix B.3. However, none of these sites had more than three SNVs reported in COSMIC, which is insufficient to suggest selection for mutation at these sites in human cancers.

Gene	Mouse chromosome	Mouse position	Equivalent human chromosome	Equivalent human position	Genotype
Ptprc	1	138117790	1	198702494	0/1
Abl2	1	156641457	1	179108983	0/1
Nutm1	2	112248256	15	34357423	0/1
B2m	2	122151119	15	44715669	1/1
Usp8	2	126758523	15	50499005	0/1
Sirpb1b	3	15542385	20	1635512	1/1
Sirpb1c	3	15832375	20	1635512	1/1
Ptprd	4	75956319	9	8341103	0/1
Thrap3	4	126178083	1	36289543	2/1
Ptpn13	5	103501611	4	86701497	0/1
Ncor2	5	125106206	12	124466180	0/1
Cdx2	5	147306749	13	27968773	0/1
Brca2	5	150541525	13	323392245	0/1
Brca2	5	150543195	13	32340919	0/1
Prkcb	7	122590166	16	24180911	0/1
Tacc2	7	130759613	10	122249129	0/1
Crtc1	8	70392070	19	18768578	0/1
Fat3	9	16376784	11	92353569	1/1
Muc16	9	18644173	19	8939057	1/1
Kmt2a	9	44848133	11	118473583	1/1
Tcf12	9	71849844	15	57282526	1/1
Atr	9	95865570	3	142562507	1/1
Ptprk	10	28493041	6	128067665	0/1
Ros1	10	52081998	6	117324371	0/1
Usp44	10	93847307	12	95532619	1/1
Gli1	10	127331182	12	57470938	0/1
Stat6	10	127647806	12	57107622	0/1
Flt4	11	49643527	5	180612519	0/1
Ktn1	14	47704466	14	55650615	0/1
Csmd3	15	47847161	8	112503844	0/1
Robo2	16	74035037	3	77493330	0/1

Table 2.6: **Missense SNVs in mouse equivalents of Cancer Gene Census genes in NIH3T3 wild-type**

SNVs in NIH3T3 wild-type that are predicted by the Ensembl Variant Effect Predictor (McLaren et al., 2016) to cause a missense mutation in a mouse homologue of a Cancer Gene Census (Futreal et al., 2004) gene. Variants are listed along with their positions in the NIH3T3 wild-type genome (GRCm38), and equivalent locations in the human genome (GRCh38) when remapped using the Genome Browser LiftOver tool from the University of California Santa Cruz (Kent et al., 1976). Genotypes: 0/1 = heterozygous reference and alternate allele, 1/1 = homozygous alternate allele, 2/1 = heterozygous first alternate allele and second alternate allele (for reference and alternate alleles for each variant, see Appendix B.3).

2.3.4 Comparison of NIH3T3 wild-type and NIH3T3-Cas9

A total of 8,385 variants were seen in NIH3T3-Cas9 where NIH3T3 wild-type showed the reference allele, suggesting that these mutations may have occured during culture since the establishment of this cell line from the parental wild-type line. Of these, 4,356 were SNVs and 4,029 were indels. This number was higher than expected considering the relatively short time in culture (estimated <20 passages).

To assess the possible consequences of these additonal variants, they were filtered using Ensembl VEP (McLaren et al., 2016) for 'high' and 'moderate' severity coding consequences, leaving a total of 22 variants, consisting of 19 SNVs and three indels. The numbers of variants are listed by consequence in table 2.7. These variants were then intersected with the 803 mouse CGC gene homologues described above to identify coding variants in known cancer-associated genes, of which there were two, listed in table 2.8. These are heterozygous missense mutations of the CGC genes *Fat4* and *Zfhx3*.

Fat4 expression has been shown to be downregulated in gastric cancers when compared with adjacent normal tissue, with lower expression corellating with reduced survival (Cai et al., 2015), supporting a role as a tumour suppressor gene. This gene has also been reported as a putative tumour suppressor in triple negative breast cancer (Hou et al., 2016).

Zfhx3 mutation has been shown to be associated with endometrial cancer, where it predominantly undergoes loss-of-function mutation and is associated with poorer outcome (Walker et al., 2015).

These results suggest that both *Fat4* and *Zfhx3* require downregulation or homozygous loss-of-function mutation in order to generate a cancer-associated phenotype, therefore the heterozygous missense mutation seen in NIH3T3-Cas9 is unlikely to have this effect. The lack of any differences between NIH3T3 wild-type and NIH3T3-Cas9 in terms of CGC genes containing mutations similar to those seen in human cancers is reassuring, and suggests that the transformation characteristics of NIH3T3-Cas9 should be similar to those seen in NIH3T3 wild-type.

2.3.5 Multiplex Fluorescence *In Situ* Hybridisation (M-FISH) of NIH3T3-Cas9

In order to karyotype the cell line NIH3T3-Cas9, 10 randomly selected metaphase chromosome spreads were hybridised with 21-colour mouse chromosome specific DNA probes and the karyotype was determined based on M-FISH DNA probe and DAPI-banding patterns. The results are shown in figure 2.3, illustrating the abnormal karyotype of this cell line. At the whole chromosome level (mean counts across the 10 cells) 41% of the chromosomes were

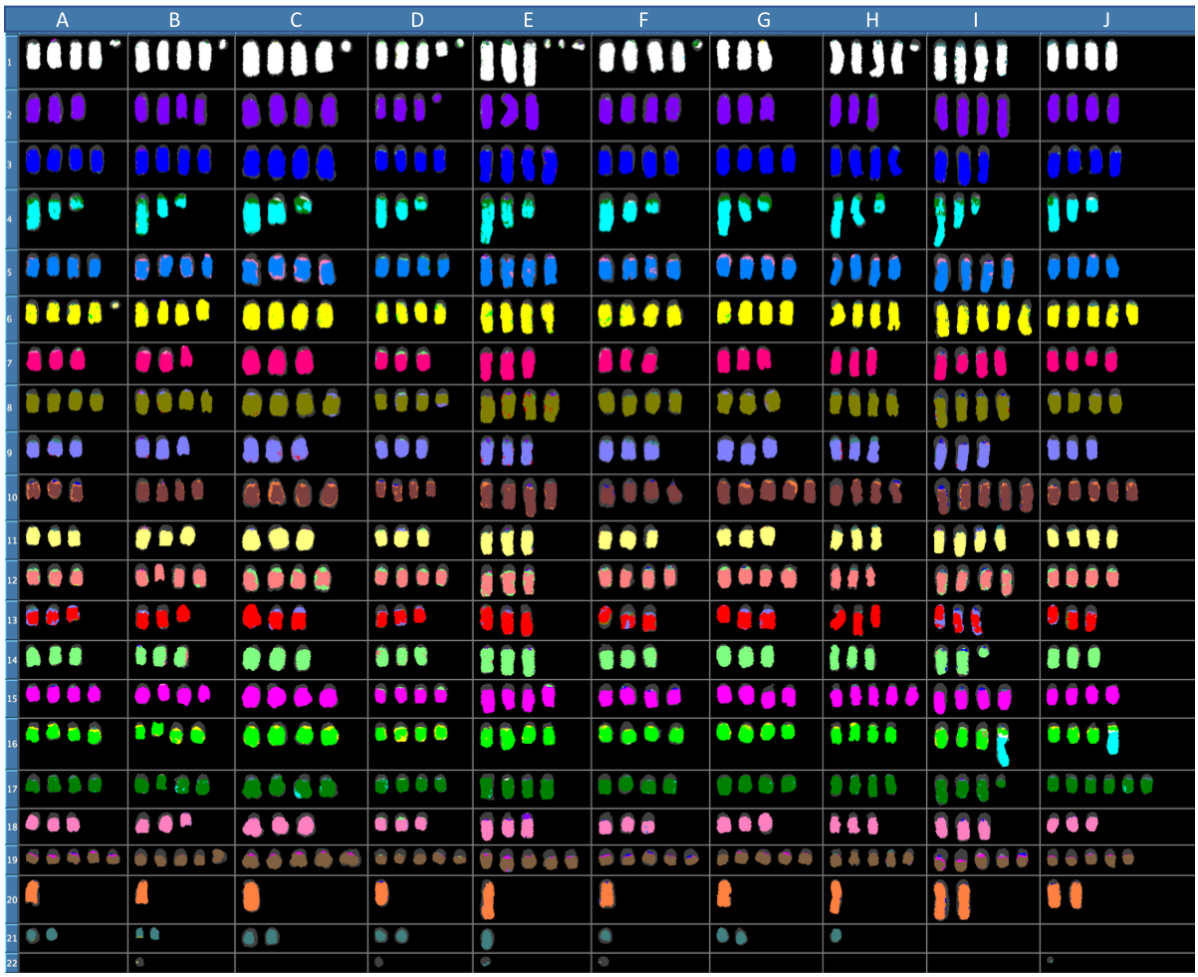


Figure 2.3: **Multiplex Fluorescence *In Situ* Hybridisation of 10 randomly selected NIH3T3-Cas9 cells at metaphase**
10 metaphase chromosome spreads (A-J) were randomly selected from a pool of NIH3T3-Cas9 cells and karyotyped using Multiplex Fluorescence *In Situ* Hybridisation with 21-colour mouse chromosome specific DNA probes (Geigl et al., 2006). This figure shows the binding patterns of the probes specific to each chromosome, each indicated by a unique colour. Where dark grey portions are seen, this indicates a chromosomal region that was bound by DAPI but did not bind any of the probes. Chromosomes 1-19 are labelled with their number, the X chromosome is numbered 20 and the Y chromosome is numbered 21. Chromosomes in row 22 are marker chromosomes that did not hybridise to any of the probe sets.

Sequence ontology term	Number of variants
missense variant	17
stop gained	2
splice acceptor variant	1
splice donor variant	1
protein altering variant	1

Table 2.7: **NIH3T3-Cas9-specific variants by consequence**
Variants present in NIH3T3-Cas9 that are not seen in the parental NIH3T3 wild-type cell line, grouped by Sequence Ontology (Eilbeck et al., 2005) term assigned by the Ensembl Variant Effect Predictor (McLaren et al., 2016).

Mouse chromosome	Position	Reference	Alternate	Genotype	Gene	Sequence ontology term
3	38980892	G	T	heterozygous	*Fat4*	missense
8	108956120	T	A	heterozygous	*Zfhx3*	missense

Table 2.8: **NIH3T3-Cas9-specific variants overlapping mouse CGC gene homologues**
Variants present in NIH3T3-Cas9 that are absent in the parental NIH3T3 wild-type cell line, that overlap mouse homologues of the Cancer Gene Census (Futreal et al., 2004) genes.

triploid, 46% were tetraploid, 12% were quintaploid, and 1% were hexaploid. This constitution suggests whole genome duplication to form a tetraploid cell line, followed by widespread single chromosome deletions to give the large number of chromosomes present in three copies, alongside duplications of others to give five or six copies.

The occurence of whole chromosome gains and losses and translocations indicates genetic instability at the chromosomal level in this cell line. Between the 10 cells, there is variation both in chromosome number (from 71-79 in total) and in features such as translocations, with a 4:16 translocation present in cells I and J only, and a marker chromosome present in five of the cells. The karyotype of each of the 10 cells is unique, indicating a high degree of karyotypic heterogeneity within this cell line. The cell line NIH3T3 was originally established from a single clone, therefore the observed variation in karyotype suggests that this line has evolved in culture, and may continue to do so.

Another way to assess chromosomal level evolution in NIH3T3 is to compare my results with those from a previous publication where the cell line was characterised using a cytogenetic approach (Leibiger et al., 2013). Here, 25 metaphases were analysed using murine multicolour banding probes, which are able to characterise the genome at a higher resolution than M-FISH using the banding patterns generated. In this study, 75% of the metaphase spreads showed the karyotype featured in figure 2.4. The two analyses have some broad similarities, for example both show a predominantly tetraploid karyotype, with some whole chromosome deletions and amplifications, and the presence of chromosomal translocations. However, the

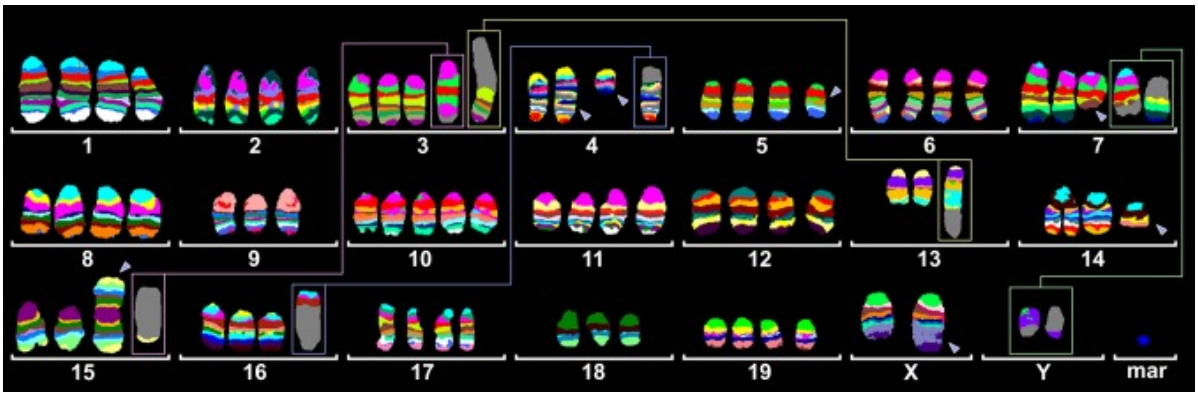


Figure 2.4: **Results of NIH3T3 karyotyping by murine multicolour banding from Leibiger et al., 2013**
Murine multicolour banding was applied to 25 individual metaphase NIH3T3 chromosome spreads. This figure shows the typical pseudocolour banding pattern for each chromosome (karyotype shown is present in 75% of cells analysed). Derivative chromosomes are shown as two segments, grouped with their chromosome of origin and surrounded by boxes, connected by a line. A marker chromosome (mar) which did not specifically stain using any of the murine multicolour binding sets is also shown.

number of translocations differs between the two analyses. In the analysis by Leibiger *et al.* the translocations present are 3:15, 3:13, 4:16 and two 7:Y fusions, whereas in my analysis none of these are present except 4:16, which is only present in two of 10 cells. These translocations could have reverted in a subpopulation of the cell line, or alternatively could have been present in a subset of cells that gave rise to the line used by Leibiger *et al.* Another difference is that there are fewer chromosomes present in three copies in the analysis by Leibiger *et al.*, and more present in four copies. These differences again indicate that the cell line is continuing to evolve at the karyotypic level. One notable difference between the two analyses is that the cells analysed by Leibiger *et al.* appeared to be more karyotypically consistent, with 75% of the metaphase spreads showing an identical karyotype, whereas all ten NIH3T3-Cas9 cells analysed were unique at the chromosomal level. However, 25% of the karyotypes analysed by Leibiger *et al.* diverged from this typical constitution, indicating that a degree of karyotypic heterogeneity was present in this NIH3T3 sample as well. This could potentially indicate increased chromosomal instability in the cell line since 2013, however since these are only two samples taken at single time-points, caution should be exercised when attempting to draw conclusions from these results in terms of the evolution of the cell line as a whole. However, these differences in karyotype both between and within the two analyses clearly show the presence of karyotypic heterogeneity in the cell line, indicating some evolution at the chromosomal level over time.

2.4 Discussion

2.4.1 Comparison of NIH3T3 wild-type genome with human cancer genome data

Single nucleotide variants and indels

The analysis of SNVs and indels present in the NIH3T3 wild-type genome revealed that there are 2018 variants predicted by Ensembl Variant Effect Predictor (McLaren et al., 2016) to have moderate or high severity coding consequences.

The genetic background of the cell line has the potential to affect the outcomes of the CRISPR-Cas9 and transposon-based screens described in chapters three and four. One factor that may influence the genes discovered in the screens is that the mutations introduced experimentally may work in combination with existing mutations to cause the observed malignant transformation. This may mean that some of the genes identified may not have the ability to cause transformation when mutated alone, but instead require other mutations that are already present in NIH3T3. Unpicking these relationships may be important in further work to determine the mechanisms of transformation underlying these novel mutations and explaining the basic cancer biology behind them. The same principle applies to the large scale mutations in NIH3T3-Cas9 that were identfied by M-FISH, which may also have unknown interactions with the hits discovered by the genetic screens.

The 88 coding mutations in NIH3T3 that affect known cancer-associated genes present in the Cancer Gene Census (Futreal et al., 2004) may be especially likely to influence the outcome of the screens. Of the genes that were subjected to homozygous loss-of-function mutation, two were suggested to be tumour suppressor genes by existing literature, *Cyp2c40* and *Il6st*. *Cyp2c40* appears unlikely to have a tumour suppressing effect in NIH3T3, as it primarily acts via the modulation of the immune microenvironment (Albert and Bennett, 2012). *Il6st* expression is associated with improved outcome in triple-negative breast cancer (Mathe et al., 2015), therefore it is possible that the homozygous loss-of-function mutation seen in NIH3T3 could contribute to cancer-like phenotypes. *Il6st* functions as a signal transduction protein (Taga and Kishimoto, 1997), acting upstream of Janus Kinase and Signal Transducer and Activation of Transcription 3 (STAT3) (Hibi et al., 1990). Both loss-of-function and gain-of-function mutation in *STAT3* have been reported as associated with cancer (Avalle et al., 2017), therefore the *Il6st* mutation seen in NIH3T3 could potentially have an effect on transformation.

Most of the loss-of-function mutations identified were in genes with evidence to suggest they are oncogenes (see section 2.3.3). However, as seen with *Pou5f1,* some genes can act as

oncogenes or tumour suppressor genes depending on context, so the effect of these mutations cannot be clearly identified.

The effects of the missense mutations identified in NIH3T3 are less clear, as they could cause loss- or gain-of-function depending on the effect of the codon change on protein structure or function, which is hard to determine from sequence data alone. None of the mutations observed occured at mutation 'hot-spots' according to human cancer data from COSMIC (Forbes et al., 2017), however this does not necessarily mean that they have no effect on cancer-related phenotypes.

Overall, few of the SNVs and indels identified by sequencing NIH3T3 appear to have clear biological relevance to the transformation-sensitive nature of the cells. Alternatively, this phenotype could be due to epigenetic changes that have accumulated in the cell line during culture, or the effects of the large scale genomic alterations discussed in section 2.4.2. The cytogenetic analysis by Leibeger *et al.* also attempted to align the amplifications and deletions seen in NIH3T3 with the homologous regions of the human genome. Similarity was identified between the alterations present in the cell line and those seen in human cancers of ectodermal origin, potentially suggesting amplifications and deletions of regions containing genes involved in tumourigenesis (Leibiger et al., 2013). It is also possible that there are variants occuring in genes with unknown or less well characterised effects on tumourigenesis, that are not currently listed as Cancer Gene Census genes.

2.4.2 Large scale genomic alterations in NIH3T3-Cas9

Genetic instability is now considered an enabling characteristic of the hallmarks of cancer, generating the genetic diversity for clonal selection to act upon, thereby facilitating the acquisition of the hallmarks (Hanahan and Weinberg, 2011). Chromosomal instability is a subtype of this phenomenon, involving an increase in the rate of gain or loss of whole chromosomes or large segments, and translocation events, a characteristic which is present in the majority of solid tumours (Bakhoum and Compton, 2012). The results of the M-FISH analysis of NIH3T3-Cas9 provides evidence of both numerical and structural chromosomal instability, showing many whole chromosome amplifications and deletions, and a translocation between chromosomes 4 and 16. These alterations vary widely between individual cells, showing that there has been a large amount of genetic divergence since the establishment of NIH3T3 from a single clone in 1969 (Jainchill et al., 1969).

It is likely that this genetic instability was also present in the NIH3T3 wild-type cell line that NIH3T3-Cas9 was derived from. NIH3T3 is a more phenotypically 'normal' cell line than cancer-derived cell lines, but the high levels of genetic instability observed could be responsible for some of the 'cancer-like' phenotypes of the line, such as its immortality and

transformation-sensitivity. In future, it would be interesting to investigate the chromosomal changes in NIH3T3 in more detail. Amplification of regions containing oncogenes, deletion of regions containing tumour suppressor genes, or the generation of fusion genes by translocation could all play a part in the characterisitcs of the cell line, and would not have been identified in my analysis of the small sequence-level mutations.

The implications of these results for the use of this cell line as a model system are wide-ranging. For the CRISPR-Cas9 and transposon-based screens in this project, the polyclonality of this cell line means that the mutations induced by CRISPR-Cas9 or transposon insertions are not working against the same genetic background in each cell in the population, which may affect the phenotypes generated due to a different combination of mutations present in any given cell. As mentioned in section 2.4.1, existing mutations may also lead to genes being picked up by the screen that require this specific genetic background to cause malignant transformation.

An alternative to the use of an established cell line for this type of screen is the use of induced pluripotent stem cells (iPSCs). These have a genetic background that is more representative of a 'normal' cell, so results obtained may be more biologically relevant, as the initial mutations of malignant transformation occur in 'normal' somatic cells. However, the reason for choosing NIH3T3 cells as a model was their established sensitivity to transformation, facilitating the identification of hits in the screens. As with many *in vitro* assays, balancing the relevance of the model to real biological systems with its feasibility of use is crucial.

Despite their clear genetic instability, these cells rarely form transformed foci of proliferation in culture spontaneously - requiring the mutation of further genes. This suggests that while the cell line's genetic instability is a potent vehicle for the acquisition of further oncogenic mutations, the instability itself is not enough to cause a malignant phenotype. This is consistent with the description of genetic instability as an enabler of the hallmarks of cancer, rather than an initiator of the transition to malignancy (Hanahan and Weinberg, 2011).

When planning the CRISPR-Cas9 screen using NIH3T3-Cas9, the initial assumption was that the NIH3T3 and NIH3T3-Cas9 cell lines would have the same susceptibility to transformation. However, it is possible that karyotypic changes that have occured between the two cell lines could have caused phenotypic changes. In order to investigate this it would be interesting to karyotype the NIH3T3 wild-type line that NIH3T3-Cas9 was generated from to compare the differences. During the CRISPR-Cas9 screen, it appeared that the NIH3T3-Cas9 cells have retained the ability to resist transformation until mutations are introduced, with no increase in background levels of transformation observed. However, a change in this susceptibility cannot be ruled out without further work to confirm this.

The abnormal karyotype of this cell line also has implications for my analysis of the se-

quence data generated from it. The variant caller used to identify mutations assumes a diploid genome when determining if a variant is present, and whether it is present heterozygously or homozygously. Due to genetic instability, in this cell line there are multiple different genotypes in different cells, and between two and six autosomes, meaning that the calling of variants and their zygosity is likely to be inaccurate in some cases.

The implications of these results go beyond the scope of this project, as they indicate that not all cell lines originally derived from a single clone retain their genetic homogeneity over years or decades in culture. This genetic evolution could lead to drastically different phenotypes, genetic interactions, and therefore results, from the same cell line. This factor could contribute to issues with non-reproducibility in cell culture-based experiments, providing a potential reason for previously described evidence on changes in morphology, gene expression and drug response in cell lines at high passage numbers (Ben-David et al., 2018; Hughes et al., 2007). To investigate the extent of this issue for NIH3T3, it would be interesting to karyotype and whole-genome sequence NIH3T3 samples from a variety of sources to get a more comprehensive picture of the genetic variability present in this supposedly clonal cell line.

2.4.3 Comparison of single nucleotide variants and indels in NIH3T3 wild-type and NIH3T3-Cas9

The comparison of these two cell lines at the level of SNVs and indels indicated that they possessed more unique mutations than expected given that they have only been cultured independently for a brief period (estimated <20 passages). Given that the cell lines appear to exhibit high levels of chromosomal instability (leading to their abnormal and variable karyotype), it is possible that the genome could also be subject to sequence instability, causing higher levels of single nucleotide and indel mutation. This would have implications for their use as a model in genetic screening, as mutations induced experimentally would be working in subtly different genetic backgrounds in each cell.

Genetic instability at the nucleotide level usually develops due to the inability of cells to detect or repair errors of replication because of mutations affecting DNA repair pathways, resulting in an increased number of SNVs and indels (Pikor et al., 2013). Mutations in genes associated with this phenotype have not been identified in this case, but it is possible that issues with DNA repair processes could have been caused by the large-scale chromosomal mutations discussed in section 2.4.2.

Further analysis of the additional variants in NIH3T3-Cas9 showed that none of these mutations are expected to cause cancer-associated phenotypes when comparing them with known

genes involved with cancer. However, this does not rule out the possibility that mutations could have occured in cancer-associated genes that are currently unknown or poorly characterised and are therefore not listed in the Cancer Gene Census.

While one interpretation of these results is that there has been genetic mutation since the establishment of the NIH3T3-Cas9 line, the number of variants that differ between the two cell lines could also simply indicate genetic heterogeneity within the NIH3T3 wild-type line. The results of the M-FISH analysis suggest that NIH3T3 is now polyclonal, despite being originally clonal, showing marked genetic variation between cells. The generation of NIH3T3-Cas9 from a subset of NIH3T3 wild-type represents a bottleneck in genetic diversity. This means that some of the 'new' mutations acquired by NIH3T3-Cas9 may have been present subclonally in the parental population, and therefore not been picked up by the variant caller. As mentioned in section 2.4.2, the deviation of this cell line from a diploid karyotype may have also caused errors in the calling of variants, and the determination of zygosity. These potential sources of error in identifying mutations that genuinely occured after the establishment of NIH3T3-Cas9 could mean that the rate of mutation in this cell line is not as high as the apparent number of new variants suggests.

To investigate the possibility of genetic change over time in these cell lines, it would be possible to sequence both lines again after a defined period of time in culture to see if mutation continues to occur at a similar rate. Alternatively, one could take two samples from the same cell line simultaneously and sequence them to determine how many variants are identified purely due to variability within a single cell line. This could help to determine whether these cell lines are actually mutating rapidly, or whether genetic heterogeneity is responsible for the inconsistent variant calls.

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.

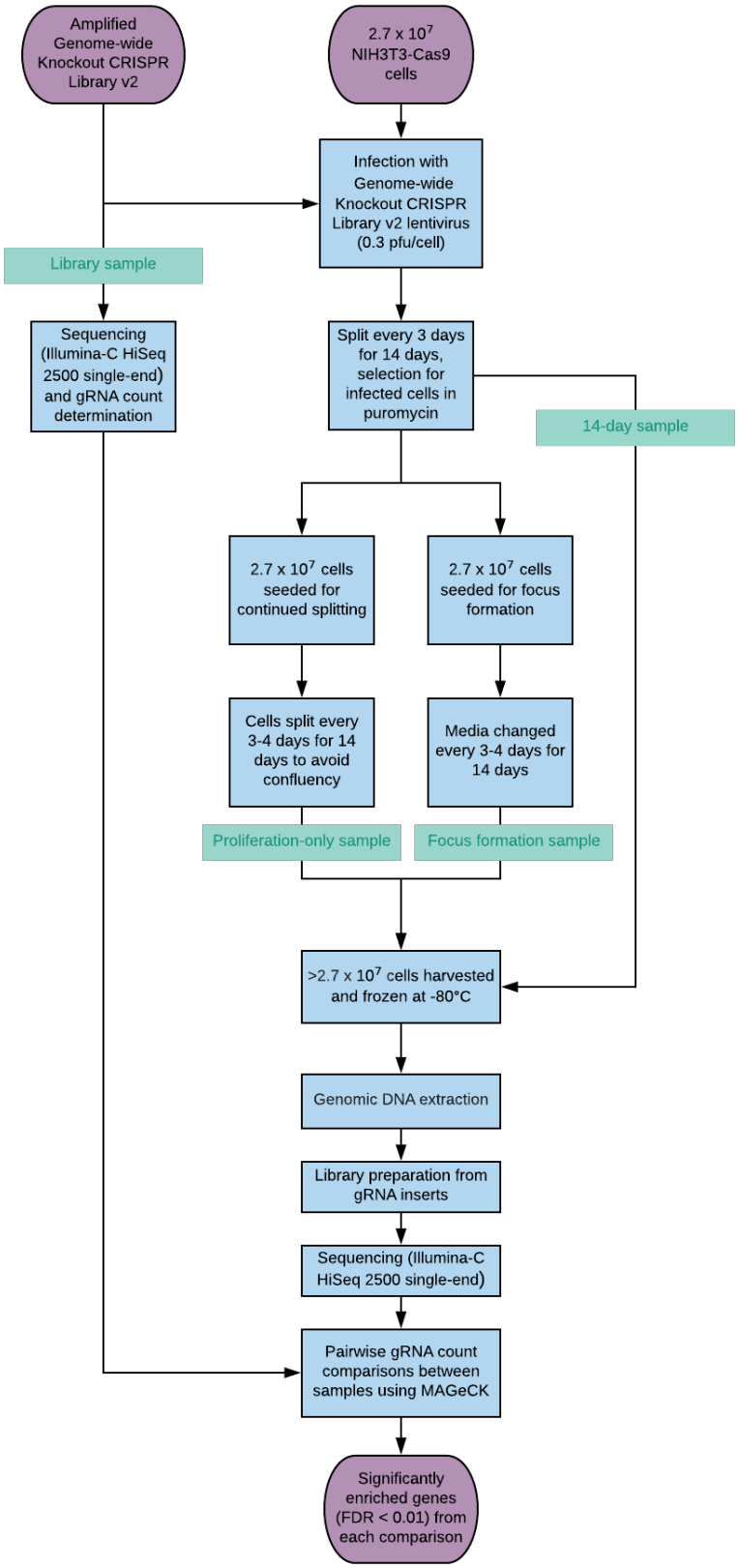


Figure 3.1: **Schematic of genome-wide CRISPR-Cas9 knockout screen using NIH3T3-Cas9 cells**

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.*

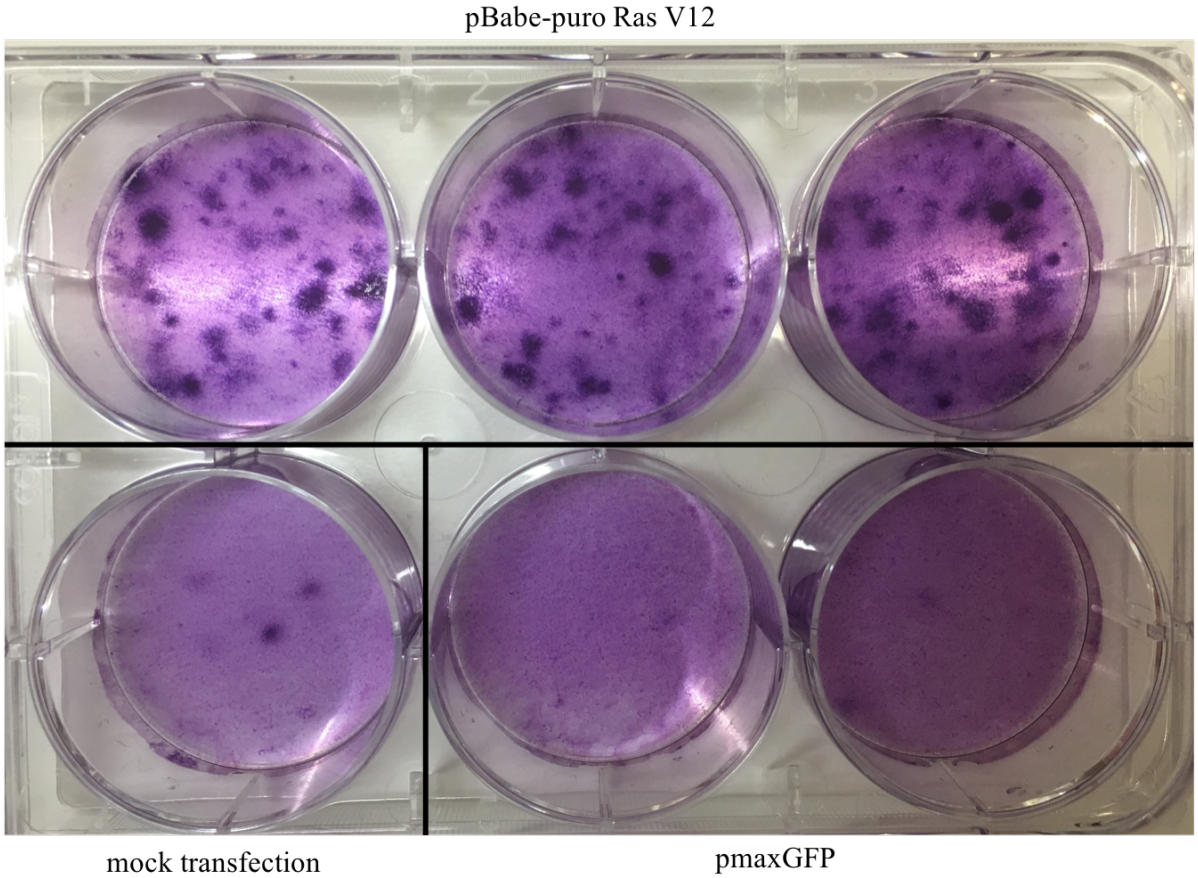


Figure 3.2: **Focus formation assay using NIH3T3 cells transfected with pBabe-puro Ras-V12**
NIH3T3 cells were transfected with pBabe-puro Ras-V12, pmaxGFP or mock transfected in order to compare the effects of *H-RAS* expression with control. After transfection, cells were cultured for 12 days before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

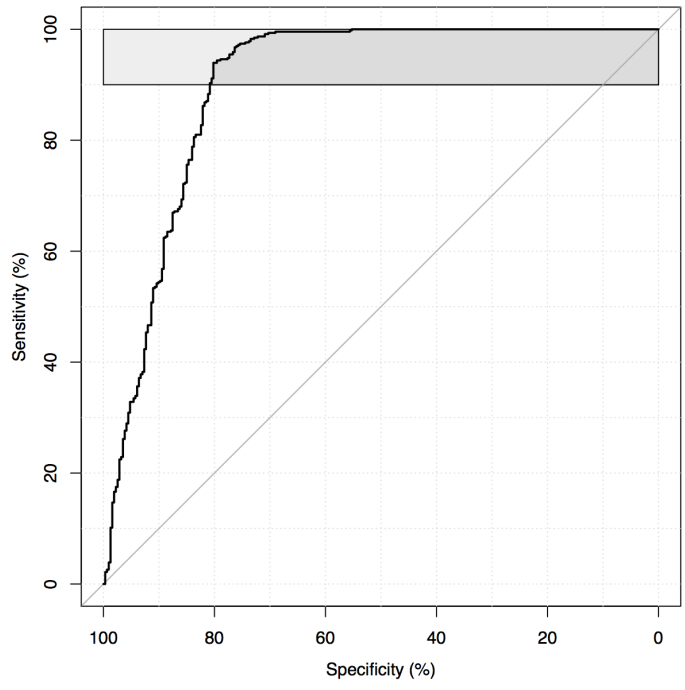


Figure 3.3: **Receiver Operating Characteristic curve based on the detection of BAGEL essential genes by the genome-wide CRISPR-Cas9 knockout screen**
This ROC curve is based on the ability of the 14-day - library MAGeCK comparison from the genome-wide CRISPR-Cas9 knockout screen (section 3.2.2) to detect 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.

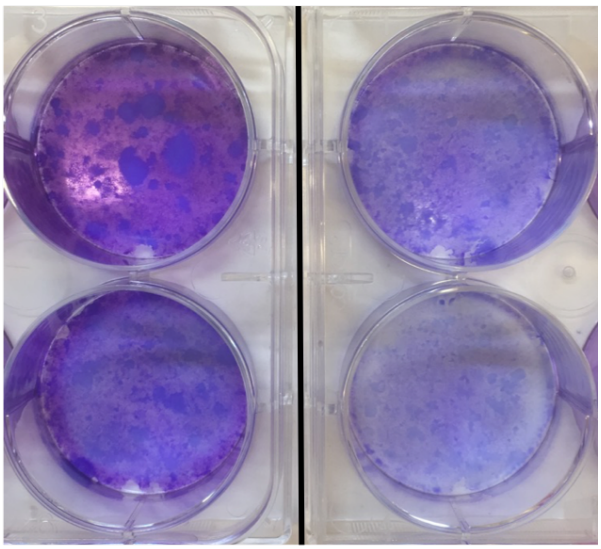
Validation - Arrayed focus formation assay

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).

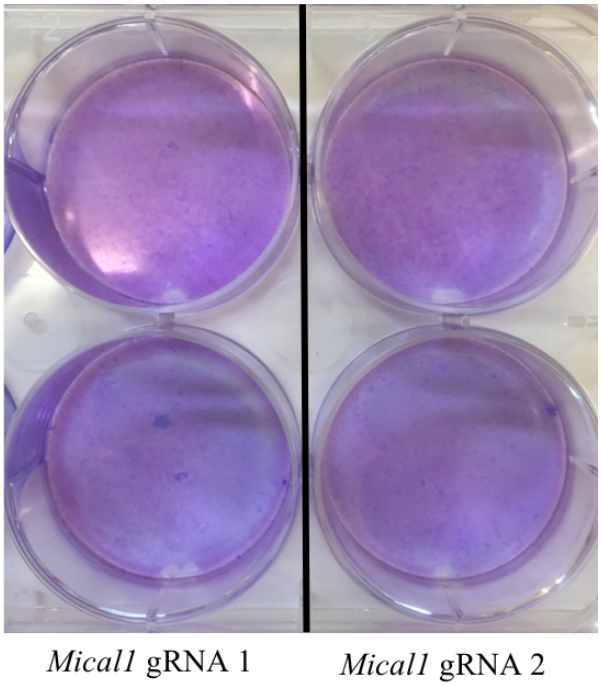
The inability of this assay to validate these hits was originally thought to be due to the difference in the way the gRNAs were introduced to the cells. In the main screen, a lentivirus carrying the Genome-wide Knockout CRISPR Library v2 was used to integrate the gRNA sequences into the genome, whereas 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).

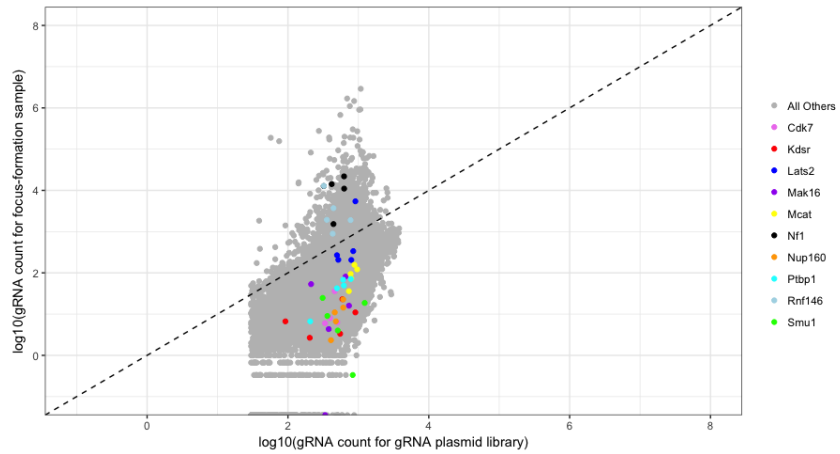
in human cancers, however it is probable that this is purely a passenger effect, potentially due to other nearby driver genes.

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.

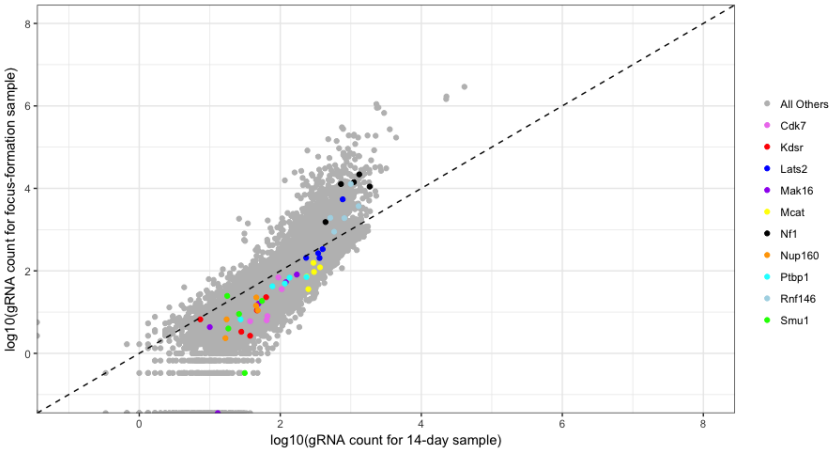
3.4 Discussion

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.

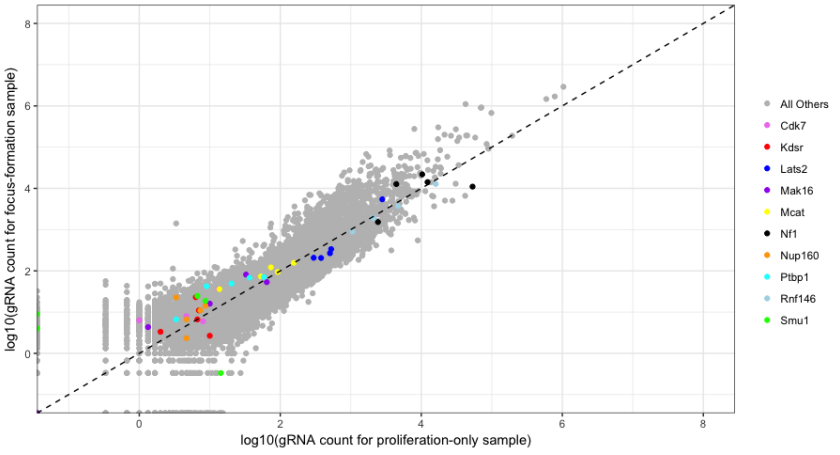
One limitation of the approach used to identify these genes is that it is only able to detect those that cause malignant transformation when mutated *alone* in the NIH3T3-Cas9 genetic background, as each gRNA causes loss-of-function mutation of a single gene. Given that cancer is a polygenic disease, 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.

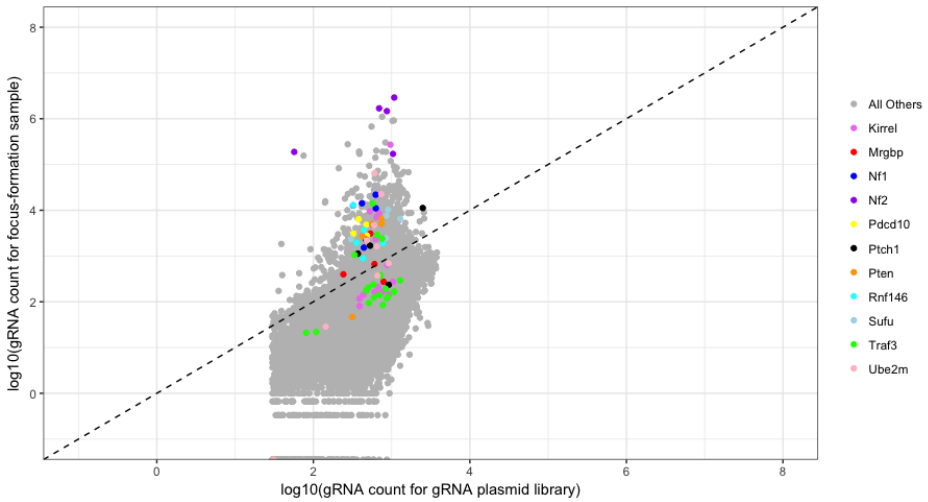


Figure 3.6: **Comparison between normalised gRNA read counts in the plasmid library and focus formation samples taken from the genome-wide CRISPR-Cas9 knockout screen for genes associated with malignant transformation**

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.

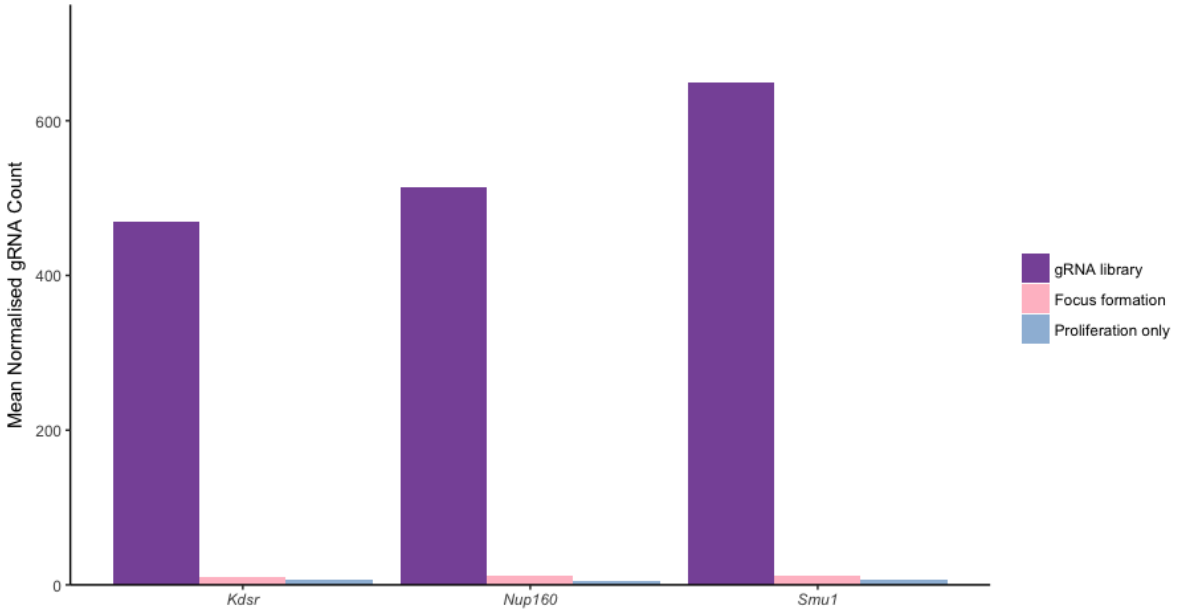


Figure 3.7: **Mean normalised read counts for gRNAs against *Kdsr, Nup160* and *Smu1* in library, proliferation-only and focus formation samples**

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.

Chapter 4

Identifying mediators of malignant transformation in cancer using genome-wide transposon-based gene activation

4.1 Introduction

The genome-wide CRISPR-Cas9 knockout screen described in chapter three has one clear limitation - it can only identify genes that cause transformation when subjected to loss-of-function mutations. Many existing mutations known to cause transformation affect oncogenes, that require overexpression, upregulation or activating point mutation to initiate tumourigenesis. These genes include those involved in pro-proliferative signalling such as *RAS* and *SRC* (Oneyama et al., 2007; Yamamoto et al., 1999), and genes where mutation allows the bypass of replicative senescence, such as *TERT*, the catalytic subunit of telomerase (Nault et al., 2014).

To detect genes where gain-of-function mutations are responsible for transformation, an approach was needed that could upregulate expression of genes genome-wide, and experimentally identify the genes of interest. Genome-wide transposon screening is a powerful approach that can be used to insert a desired sequence across the genome in a range of cell types. These insertion sites are easily recoverable due to specific sequences within the inserts, allowing the numbers of insertions at different sites in a cell population to be quantified by sequencing (Friedrich et al., 2017).

This approach has been used previously by Friedrich *et al.* (2017) for cancer gene discovery in mice, using *PiggyBac*-based transposons for genome-wide insertional mutagene-

sis. Insertions are quantified using QiSeq, which comprises DNA fragmentation by acoustic shearing, library preparation through modified splinkerette PCR (Devon et al., 1995), and custom Illumina sequencing. In this chapter, this method was used to recover insertions that transcriptionally upregulate expression of a downstream gene. The plasmid used was pPB-SB-CMV-puro-SD, which contains the human cytomegalovirus (CMV) promoter, flanked by *PiggyBac* and *Sleeping Beauty* sites (Tsutsui et al., 2015). The CMV promoter strongly upregulates downstream transcription leading to increased gene expression, aiming to model the gene amplifications seen in some tumours (Xia et al., 2006).

Similarly to the CRISPR-Cas9 screen discussed in chapter three, the principle of this screen was to induce genome-wide modifications of the NIH3T3 cells, then allow them to form transformed foci in culture. Cells containing alterations that induce transformation will therefore be overrepresented in the final population, and the numbers of mutations at different loci can be determined using sequencing to identify the genes responsible.

4.1.1 Aims

Overall aim: To identify putative oncogenes involved in malignant transformation in human cancer.

1. To identify genes that may mediate transformation *in vitro* using genome-wide transposon-based gene activation screening in NIH3T3.

2. To prioritise hits from this screen using mutation data from existing human cancer sequencing projects.

3. To functionally validate prioritised hits *in vitro.*

4.2 Materials and Methods

4.2.1 Materials

Plasmids

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

pPB-SB-CMV-puro-SD This plasmid was a gift from Professor Cyril Benes ((Tsutsui et al., 2015))

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

Cell lines

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

Reagents

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

Table 4.1: **Reagents used in the methods described in chapter 4**

4.2.2 Methods

Screen design

The aim was to acheive 100X coverage of the genome, with each gene being upregulated in a mean of 100 cells in the screen. To acheive this coverage, the rationale behind a previous screen using pPB-SB-CMV-puro-SD was used (Chen et al., 2013). Assuming that the number of insertions within a region is distributed according to the Poisson distriution, if the aim is for <5% of genes to have <100X coverage, the mean number of insertions upregulating expression of any given gene should equal 117 (see equation).

$$P(X \leq 99) = 0.05$$

$$\sum_{x=0}^{x=99} e^{-\lambda} \frac{\lambda^x}{x!} = 0.05$$

$$\lambda = 117$$

Following the assumption of Burgess et al. that the CMV promoter can upregulate transcription of a gene when <64kb upstream of the transcriptional start site, 117 insertions within this region equates to a mean 0.547kb gap between insertions. Therefore, in the 3 million kb mouse genome, 5,484,375 insertions were required. These are only functional if on the coding strand, therefore 10,968,750 insertions were needed in total.

A 1:10 ratio of *PiggyBac* transposase plasmid to *PiggyBac* transposon plasmid has been determined to generate an expected 1-10 insertions per cell (Wang et al., 2008), providing a reasonable number of insertions without causing excessive cell lethality. Making the conservative assumption that one insertion per cell is generated, 10,968,750 transfected cells would represent the same number of insertions. Given a measured transfection efficiency of NIH3T3 of 23.5% (appendix B.7), this equates to 46,675,532 cells. To account for cell loss during processing, this was rounded to 5 x 10^7.

Transfection

Day 0 5 x 10^7 NIH3T3 wild-type cells were seeded in 26 15cm-diameter culture dishes at a density of 96,100 cells/mL in complete DMEM (DMEM supplemented with 10% FBS and 500μg/mL penicillin, streptomycin and L-glutamine). For the positive and negative controls, a 6-well tissue culture plate was seeded at a density of 50,000 cells/mL, in 2mL of complete DMEM per well.

Day 1 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 reagent quantities (table 4.2). Cells in the 15cm dishes were transfected with pPB-SB-CMV-puro-SD and pCMV-hypBase, and for the positive control, 3 wells of the 6-well plate were transfected with pBabe-puro Ras V12. For the negative control, 3 wells of the 6-well plate were mock transfected, with the plasmid replaced with an equivalent volume of Opti-MEM™.

Reagent	Quantity per well (1 x 10^6 cells)	Quantity per dish (1.92 x 10^7 cells)
Lipofectamine 3000 Reagent	1.5µL	28.8µL
P3000	1µL	19.2µL
pPB-SB-CMV-puro-SD	0.45µg	8.64µg
pCMV-hypBase	45ng	0.864µg
pBabe-puro Ras V12	0.5µg	9.2µg

Table 4.2: **Transfection reagent quantities used in genome-wide transposon-based gene activation screen**

Day 2 16 hours post-transfection, media was changed to 20mL complete DMEM per dish, or 2mL per well.

Day 4 Media was changed to complete DMEM containing puromycin (2µg/ml). Media was changed every 3-4 days.

Day 30 Cells from the dishes were harvested by scraping, centrifuged (200*xg,* 5 minutes), washed with PBS and frozen at -80°C. The control cells were fixed for 1 hour using methanol, stained with 1% aqueous crystal violet for 10 seconds, washed with MilliQ water and air-dried.

DNA extraction

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

QiSeq

Transposon insertions were quantified using QiSeq (Friedrich et al., 2017). Library preparation was performed with the assistance of Dr. Jonathan Cooper from the haematological cancer genetics group at the Wellcome Sanger Institute.

4.3 Results

At 30 days there were no or few foci in the mock-transfected wells, confirming the low rate of background transformation in NIH3T3. The *RAS*-transfected positive control contained a variable number of foci per well, ranging from three to approximately 45 in total. However, the overall increased focus formation compared to the control indicates that these cells can

be transformed by known oncogenes (figure 4.1). Many foci were seen in the transposon-transfected dishes, suggesting successful transformation by transposon-mediated gene over-expression.

The results of sequencing the insertions to determine their locations have not been returned yet. Once the sequencing data is generated the aim is to analyse it using the method detailed by Friedrich *et al*. to extract the insertion sites with the most reads and correlate these with the downstream genes they are likely to induce overexpression of. Putative tranformation-associated genes will then be prioritised using existing cancer genome data and validated using an arrayed focus formation assay as decribed in sections 3.2.2 and 3.2.2 of chapter three.

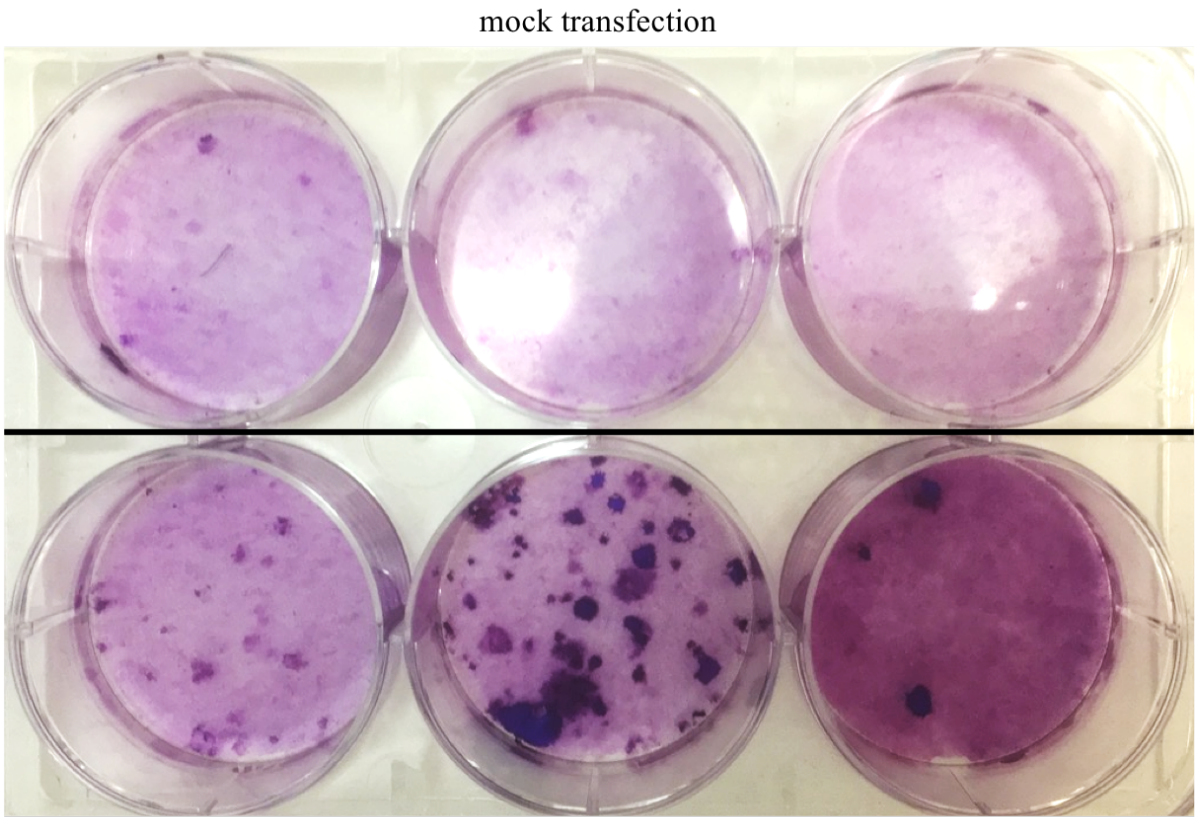


Figure 4.1: **Focus formation assay for negative and positive controls of the genome-wide transposon-based activation screen**

NIH3T3 cells were mock transfected as a negative control and transfected with a plasmid carrying known oncogene *RAS* (pBabe-puro Ras-V12) as a positive control for the genome-wide transposon-based activation screen. After culturing for 30 days, cells were stained using 1% aqueous crystal violet (see section 4.2.2).

4.4 Discussion

One limitation of this approach is that oncogenic point mutations are not accurately modelled, with the transposon insertions more closely recapitulating oncogene activation by amplification or upregulation. However, this may not preclude the identification of genes usually activated by point mutation, as many oncogenes can be be activated in multiple ways, giving similar phenotypes. For example, *RAS* is most frequently point mutated in tumours, but can also cause transformation by overexpression *in vitro* or amplification *in vivo* (Pierceall et al., 1991; Pulciani et al., 1985). There may also be issues inherent to the use of a cell-based model, as differences have been observed between the level of gene expression required for transformation *in vitro* when compared to that seen *in vivo*. For example, it has been shown that mutant human *RAS* requires over 100-fold higher expression to cause tranformation in cell lines compared to that seen in human cancers (Hua et al., 1997). However, *RAS* overexpression was used as the positive control in this screen and successfully induced transformation, which is promising.

As discussed in chapter three, the heterogeneous genetic background and complex karyotype of NIH3T3 may affect the nature of the genes identified from this screen, as the ability of a mutation to induce tranformation is likely to be dependent on other co-occurring mutations. Another limitation of the model discussed in chapter three is the absence of factors that play a part in early tumourigenesis *in vivo,* such as the immune microenvironment. As with the CRISPR-Cas9 screen, some of these issues could be resolved by future validation of any hits using mouse models. This could be done by injection of CRISPR-edited cells, or alternatively using a transgenic model that can activate the desired putative oncogene in a lineage- and time-specific manner (Blanpain, 2013).

Chapter 5

Conclusions and further directions

In this thesis I have described the genetic characterisation of the cell-based model of malignant transformation NIH3T3, and karyotyping of its daughter cell line NIH3T3-Cas9 followed by presenting two complementary forward genetic screening approaches used to identify putative oncogenes and tumour suppressor genes involved in the earliest stages of tumourigenesis.

5.1 Characterisation of the genetic background of NIH3T3 and NIH3T3-Cas9

The characterisation of the NIH3T3 genetic background consisted of whole-genome sequencing to identify SNVs and indels, comparison of these variants with known cancer-associated genes from the CGC and mutations from COSMIC, and characterisation of large-scale genomic alterations using M-FISH. 88 SNVs and indels with coding effects in CGC genes were identified; however determining the phenotypic effects of these variants and how they may influence transformation sensitivity was challenging. Some of the indels identified were predicted to have loss-of-function effects, but for many it was unclear whether this would promote tumourigenesis given the function of the gene and the mutations reported in human cancers. However the indel affecting a splice acceptor site in *Il6st* is worth further investigation due to its potential effect on known cancer-associated pathway JAK-STAT3 (Avalle et al., 2017; Hibi et al., 1990). The SNVs identified were not located at mutation hot-spots catalogued by COSMIC, however it is difficult to predict the effects of point mutations on a protein from sequence data alone. It may be possible to investigate this using computational approaches to the prediction of the effects of codon changes on protein structure and function (Tang and Thomas, 2016).

Additionally, mutations in genes that were not investigated in the analyses carried out in

this project may contribute to the transformation sensitivity of this cell line. For example, there may be murine genes without human homologues that affect transformation, or genes with unknown or poorly characterised effects on tumourigenesis. To ensure that these genes are also investigated, additional literature sources could be included, such as the positively selected driver mutations reported by Martincorena *et al.* in 2017. The transformation-sensitive phenotype of NIH3T3 could also be explained by epigenetic effects. To investigate this possibility, the epigenome of the cell line could be characterised using methods such as cytosine methylation profiling, and compared to epigenome data from patient tumour samples or cancer cell lines (Ehrich et al., 2008).

The analysis of large-scale genetic alterations in the daughter cell line NIH3T3-Cas9 showed an abnormal, predominantly tetraploid karyotype with whole chromosome amplifications, deletions and translocations. The effects of these alterations on cancer-associated genes may explain the transformation-sensitivity of the line. These results also revealed high levels of inter-cell heterogeneity, indicating genetic instability and suggesting that the cell line is continuing to evolve at the chromosomal level. This is also supported by the differences between the karyotype determined in this analysis and in previous cytogenetic characterisation by Leibiger *et al.* (2013). This finding has important implications for the use of this cell line as a model. As an enabling characteristic of the 'hallmarks of cancer' (Hanahan and Weinberg, 2011), genetic instability may influence transformation-related phenotypes and may may contribute to the transformation-sensitivity of these cells. Genetic heterogeneity may also affect the outcome of genetic screens, as experimentally-induced mutations do not act in a consistent genetic background given the polyclonality of the model.

In future, it would be informative to characterise the karyotype and genetic heterogeneity of the parental cell line NIH3T3 wild-type. These results could be compared with NIH3T3 samples from other sources to get a more complete picture of the heterogeneity present in the cell line as a whole. Large-scale alterations found consistently in the cell line could then be investigated at higher resolution to identify if any known oncogenes or tumour suppressor genes are amplified or deleted.

Overall, this work has made some progress in identifying possible reasons behind the transformation-sensitive phenotype of NIH3T3, while providing further evidence that cell lines evolve, and should not be assumed to remain genetically or phenotypically identical over time in culture.

5.2 Identification of candidate genes associated with malignant transformation

The genome-wide CRISPR-Cas9 knockout screen sought to identify novel candidate tumour suppressor genes that induce the formation of tranformed foci of proliferation when knocked out in NIH3T3. This screen successfully identified the known cancer-associated genes *Gnas, Kdsr, Sufu, Nf2, Cltc, Ptch1, Nf1* and *Pten*, along with putative novel candidates. The combined results of manual investigation of the normalised read counts and their subsequent analysis using MAGeCK showed that using the original gRNA sequence library as a control was the most successful approach for identifying gRNA seqeunces that were genuinely enriched, indicating that the corresponding genes may be involved in the formation of the transformed foci. This also highlighted an issue with using MAGeCK analysis to compare two samples that have been grown in culture, which can lead to the mis-identification of essential genes as enriched hits due to high statistical noise at low read counts. In future this could be avoided by removing gRNAs with control read counts below a certain threshold from the analysis, and by manually investigating the normalised read count data at an earlier stage to discard spurious hits. Alternatively, the screen could be redesigned to use only the gRNA sequence library as control.

The novel candidates identified from the comparison between the focus formation sample with the gRNA sequence library were *Rnf146, Kirrel, Mrgbp, Pdcd10, Traf3* and *Ube2m*. *Rnf146, Kirrel* and *Mrgbp* are described as putative oncogenes or reported as overexpressed in cancer samples (Gao et al., 2014; Yamaguchi et al., 2011; Zhang et al., 2018), whereas *Pdcd10, Traf3* and *Ube2m* are described as having tumourigenic effects when inactivated (Cukras et al., 2014; Hajek et al., 2017; Lambertz et al., 2015). This is interesting considering that all of these genes emerged from a screen employing loss-of-function mutation, suggesting that the effects of *Rnf146, Kirrel* and *Mrgbp* may be poorly characterised or context-dependent.

The planned analysis of the data generated from the transposon-based activation screen should generate candidate oncogenes involved in transformation. Once these hits and those from the CRISPR-Cas9 screen have been validated *in vitro,* successful hits could be verified *in vivo* using mouse models. A limitation of the current approach is that murine cancer-related pathways may show important differences from the human equivalents. To test genes in a human context, it would be informative to repeat any successful validation experiments using the human homologues of candidates in the human non-tumorigenic immortalised cell line MCF-10A (Qu et al., 2015). Following validation, further work to dissect the pathways containing these genes and their biological functions would be enlightening. Existing data on human cancer genomes from sources such as COSMIC could be used to investigate in which

tumour types these genes are commonly mutated, what kinds of mutations are most common, and which other mutations are significantly co-occuring or mutually exclusive. Overall, the further investigation of these candidates has the potential to inform the basic biology of malignant transformation and possible future therapeutic targets.

References

David J. Adams, Anthony G. Doran, Jingtao Lilue, and Thomas M. Keane. The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mammalian Genome*, 26(9-10):403–412, 2015. ISSN 14321777. doi: 10.1007/s00335-015-9579-6.

Mazhar Adli. The CRISPR tool kit for genome editing and beyond. *Nature communications*, 9(1):1911, may 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04252-2.

M Albert and A Bennett. The roles of CYP2C40 and CYP2C55 in preventing colon cancer. *Gut*, 61(Suppl 2):A333.2–A333, 2012. ISSN 0017-5749. doi: 10.1136/gutjnl-2012-302514d.89.

Donna G. Albertson, Bauke Ylstra, Richard Segraves, Colin Collins, Shanaz H. Dairkee, David Kowbel, Wen-Lin Kuo, Joe W. Gray, and Daniel Pinkel. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genetics*, 25(2):144–146, jun 2000. ISSN 1061-4036. doi: 10.1038/75985.

Samuel Aparicio and Carlos Caldas. The Implications of Clonal Genome Evolution for Cancer Medicine. *New England Journal of Medicine*, 368(9):842–851, feb 2013. ISSN 0028-4793. doi: 10.1056/NEJMra1204892.

Lidia Avalle, Annalisa Camporeale, Andrea Camperi, and Valeria Poli. STAT3 in cancer: A double edged sword. *Cytokine*, 98:42–50, oct 2017. ISSN 10434666. doi: 10.1016/j.cyto.2017.03.018.

Samuel F Bakhoum and Duane a Compton. Chromosomal instability and cancer : a complex relationship with therapeutic potential. *The Journal of Clinical Investigation*, 122(4):1138–1143, 2012. ISSN 1558-8238. doi: 10.1172/JCI59954.1138.

Sneha Balani, Long V. Nguyen, and Connie J. Eaves. Modeling the process of human tumorigenesis. *Nature Communications*, 8:15422, may 2017. ISSN 2041-1723. doi: 10.1038/ncomms15422.

Allan Balmain. Cancer genetics: from Boveri and Mendel to microarrays. *Nature Reviews Cancer*, 1(1):77–82, oct 2001. ISSN 1474-175X. doi: 10.1038/35094086.

Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, 315(5819): 1709–12, mar 2007. ISSN 1095-9203. doi: 10.1126/science.1138140.

Philip A. Beer and Connie J. Eaves. Modeling Normal and Disordered Human Hematopoiesis. *Trends in Cancer*, 1(3):199–210, nov 2015. ISSN 2405-8033. doi: 10.1016/J.TRECAN.2015.09.002.

Uri Ben-David, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A. Strathdee, Joshua Dempster, Nicholas J. Lyons, Robert Burns, Anwesha Nag, Guillaume Kugener, Beth Cimini, Peter Tsvetkov, Yosef E. Maruvka, Ryan O'Rourke, Anthony Garrity, Andrew A. Tubelli, Pratiti Bandopadhayay, Aviad Tsherniak, Francisca Vazquez, Bang Wong, Chet Birger, Mahmoud Ghandi, Aaron R. Thorner, Joshua A. Bittker, Matthew Meyerson, Gad Getz, Rameen Beroukhim, and Todd R. Golub. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560(7718):325–330, aug 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0409-3.

Benoit Biteau, Christine E Hochmuth, and Heinrich Jasper. Maintaining tissue homeostasis: dynamic control of somatic stem cell activity. *Cell stem cell*, 9(5):402–11, nov 2011. ISSN 1875-9777. doi: 10.1016/j.stem.2011.10.004.

Cédric Blanpain. Tracing the cellular origin of cancer. *Nature Cell Biology*, 15(2):126–134, feb 2013. ISSN 1465-7392. doi: 10.1038/ncb2657.

T Boveri. Uber mehrpolige Mitosen als Mittle zur Analyse des Zellkerns. *Verhandl Phys-med Ges (Wulzburg) NF*, 35:67–90, 1902.

D E Brash, J A Rudolph, J A Simon, A Lin, G J McKenna, H P Baden, A J Halperin, and J Pontén. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 88(22):10124–8, nov 1991. ISSN 0027-8424. doi: 10.1073/PNAS.88.22.10124.

Jian Cai, Dan Feng, Liang Hu, Haiyang Chen, Guangzhen Yang, Qingping Cai, Chunfang Gao, and Dong Wei. FAT4 functions as a tumour suppressor in gastric cancer by modulating Wnt/β-catenin signalling. *British Journal of Cancer*, 113(12):1720–1729, 2015. ISSN 15321827. doi: 10.1038/bjc.2015.367.

Cancer Research UK. Saving lives, averting costs - An analysis of the financial implications of achieving earlier diagnosis of colorectal, lung and ovarian cancer. Technical report, 2014.

Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. ISSN 21598274. doi: 10.1158/2159-8290.CD-12-0095.

Li Chen, Lynda Stuart, Toshiro K. Ohsumi, Shawn Burgess, Gaurav K. Varshney, Anahita Dastur, Mark Borowsky, Cyril Benes, Adam Lacy-Hulbert, and Emmett V. Schmidt. Transposon activation mutagenesis as a screening tool for identifying resistance to cancer therapeutics. *BMC Cancer*, 13(1):1, 2013. ISSN 14712407. doi: 10.1186/1471-2407-13-93.

Jiqiu Cheng, Jonas Demeulemeester, David C. Wedge, Hans Kristian M. Vollan, Jason J. Pitt, Hege G. Russnes, Bina P. Pandey, Gro Nilsen, Silje Nord, Graham R. Bignell, Kevin P. White, Anne Lise Børresen-Dale, Peter J. Campbell, Vessela N. Kristensen, Michael R. Stratton, Ole Christian Lingjærde, Yves Moreau, and Peter Van Loo. Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-01355-0.

Karen Cichowski and Tyler Jacks. NF1 tumor suppressor gene function: Narrowing the GAP. *Cell*, 104(4):593–604, 2001. ISSN 00928674. doi: 10.1016/S0092-8674(01)00245-8.

Lisa M Coussens and Zena Werb. Inflammation and cancer. *Nature*, 420(6917):860–7, 2002. ISSN 0028-0836. doi: 10.1038/nature01322.

Scott Cukras, Nicholas Morffy, Takbum Ohn, and Younghoon Kee. Inactivating UBE2M impacts the DNA damage response and genome integrity involving multiple cullin ligases. *PloS one*, 9(7):e101844, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0101844.

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27 (15):2156–2158, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr330.

David Dankort, David P Curley, Robert a Cartlidge, Betsy Nelson, N Anthony, William E Damsky Jr, Mingjian J You, Ronald a Depinho, and Marcus Bosenberg. BRAF V600E

cooperates with PTEN silencing to elicit metastatic melanoma. *Nature genetics*, 41(5): 544–552, 2009. ISSN 1546-1718. doi: 10.1038/ng.356.BRaf.

Brigitte David-Watine. Silencing nuclear pore protein Tpr elicits a senescent-like phenotype in cancer cells. *PloS one*, 6(7), 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0022423.

Elitza Deltcheva, Krzysztof Chylinski, Cynthia M. Sharma, Karine Gonzales, Yanjie Chao, Zaid A. Pirzada, Maria R. Eckert, Jörg Vogel, and Emmanuelle Charpentier. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, 471(7340): 602–607, mar 2011. ISSN 0028-0836. doi: 10.1038/nature09886.

Rebecca S. Devon, David J. Porteous, and Anthony J. Brookes. Splinkerettes - improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Research*, 23(9):1644–1645, 1995. ISSN 03051048. doi: 10.1093/nar/23.9.1644.

Sheng Ding, Xiaohui Wu, Gang Li, Min Han, Yuan Zhuang, and Tian Xu. Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. *Cell*, 122(3): 473–483, aug 2005. ISSN 0092-8674. doi: 10.1016/J.CELL.2005.07.013.

M. Ehrich, J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor, and D. van den Boom. Cytosine methylation profiling of cancer cell lines. *Proceedings of the National Academy of Sciences*, 105(12):4844–4849, mar 2008. ISSN 0027-8424. doi: 10.1073/pnas.0712251105.

Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5), 2005. ISSN 1465-6914. doi: 10.1186/gb-2005-6-5-r44.

Christina Fitzmaurice, Christine Allen, Ryan M. Barber, Lars Barregard, Zulfiqar A. Bhutta, Hermann Brenner, Daniel J. Dicker, Odgerel Chimed-Orchir, Rakhi Dandona, Lalit Dandona, Tom Fleming, Mohammad H. Forouzanfar, Jamie Hancock, Roderick J. Hay, Rachel Hunter-Merrill, Chantal Huynh, H. Dean Hosgood, Catherine O. Johnson, Jost B. Jonas, Jagdish Khubchandani, G. Anil Kumar, Michael Kutz, Qing Lan, Heidi J. Larson, Xiaofeng Liang, Stephen S. Lim, Alan D. Lopez, Michael F. MacIntyre, Laurie Marczak, Neal Marquez, Ali H. Mokdad, Christine Pinho, Farshad Pourmalek, Joshua A. Salomon, Juan Ramon Sanabria, Logan Sandar, Benn Sartorius, Stephen M. Schwartz, Katya A. Shackelford, Kenji Shibuya, Jeff Stanaway, Caitlyn Steiner, Jiandong Sun, Ken Takahashi, Stein Emil Vollset, Theo Vos, Joseph A. Wagner, Haidong Wang, Ronny West-

erman, Hajo Zeeb, Leo Zoeckler, Foad Abd-Allah, Muktar Beshir Ahmed, Samer Alabed, Noore K. Alam, Saleh Fahed Aldhahri, Girma Alem, Mulubirhan Assefa Alemayohu, Raghib Ali, Rajaa Al-Raddadi, Azmeraw Amare, Yaw Amoako, Al Artaman, Hamid Asayesh, Niguse Atnafu, Ashish Awasthi, Huda Ba Saleem, Aleksandra Barac, Neeraj Bedi, Isabela Bensenor, Adugnaw Berhane, Eduardo Bernabé, Balem Betsu, Agnes Binagwaho, Dube Boneya, Ismael Campos-Nonato, Carlos Castañeda-Orjuela, Ferrán Catalá-López, Peggy Chiang, Chioma Chibueze, Abdulaal Chitheer, Jee-Young Choi, Benjamin Cowie, Solomon Damtew, José das Neves, Suhojit Dey, Samath Dharmaratne, Preet Dhillon, Eric Ding, Tim Driscoll, Donatus Ekwueme, Aman Yesuf Endries, Maryam Farvid, Farshad Farzadfar, Joao Fernandes, Florian Fischer, Tsegaye Tewelde G/hiwot, Alemseged Gebru, Sameer Gopalani, Alemayehu Hailu, Masako Horino, Nobuyuki Horita, Abdullatif Husseini, Inge Huybrechts, Manami Inoue, Farhad Islami, Mihajlo Jakovljevic, Spencer James, Mehdi Javanbakht, Sun Ha Jee, Amir Kasaeian, Muktar Sano Kedir, Yousef S. Khader, Young-Ho Khang, Daniel Kim, James Leigh, Shai Linn, Raimundas Lunevicius, Hassan Magdy Abd El Razek, Reza Malekzadeh, Deborah Carvalho Malta, Wagner Marcenes, Desalegn Markos, Yohannes A. Melaku, Kidanu G Meles, Walter Mendoza, Desalegn Tadese Mengiste, Tuomo J. Meretoja, Ted R. Miller, Karzan Abdulmuhsin Mohammad, Alireza Mohammadi, Shafiu Mohammed, Maziar Moradi-Lakeh, Gabriele Nagel, Devina Nand, Quyen Le Nguyen, Sandra Nolte, Felix A. Ogbo, Kelechi E. Oladimeji, Eyal Oren, Mahesh Pa, Eun-Kee Park, David M Pereira, Dietrich Plass, Mostafa Qorbani, Amir Radfar, Anwar Rafay, Mahfuzar Rahman, Saleem M. Rana, Kjetil Søreide, Maheswar Satpathy, Monika Sawhney, Sadaf G. Sepanlou, Masood Ali Shaikh, Jun She, Ivy Shiue, Hirbo Roba Shore, Mark G. Shrime, Samuel So, Samir Soneji, Vasiliki Stathopoulou, Konstantinos Stroumpoulis, Muawiyyah Babale Sufiyan, Bryan L. Sykes, Rafael Tabarés-Seisdedos, Fentaw Tadese, Bemnet Amare Tedla, Gizachew Assefa Tessema, J. S. Thakur, Bach Xuan Tran, Kingsley Nnanna Ukwaja, Benjamin S. Chudi Uzochukwu, Vasiliy Victorovich Vlassov, Elisabete Weiderpass, Mamo Wubshet Terefe, Henock Gebremedhin Yebyo, Hassen Hamid Yimam, Naohiro Yonemoto, Mustafa Z. Younis, Chuanhua Yu, Zoubida Zaidi, Maysaa El Sayed Zaki, Zerihun Menlkalew Zenebe, Christopher J. L. Murray, and Mohsen Naghavi. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015. *JAMA Oncology*, 3(4):524, apr 2017. ISSN 2374-2437. doi: 10.1001/jamaoncol.2016.5688.

Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai YinKok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam

Thompson, Tisham De, and Peter J. Campbell. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(1):777–783, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1121.

Mathias J Friedrich, Lena Rad, Iraad F Bronner, Alexander Strong, Wei Wang, Julia Weber, Matthew Mayho, Hannes Ponstingl, Thomas Engleitner, Carolyn Grove, Anja Pfaus, Dieter Saur, Juan Cadiñanos, Michael A Quail, George S Vassiliou, Pentao Liu, Allan Bradley, and Roland Rad. Genome-wide transposon screening and quantitative insertion site sequencing for cancer gene discovery in mice. *Nature Protocols*, 12(2):289–309, jan 2017. doi: 10.1038/nprot.2016.164.

Stephen H. Friend, Rene Bernards, Snezna Rogelj, Robert A. Weinberg, Joyce M. Rapaport, Daniel M. Albert, and Thaddeus P. Dryja. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 323(6089):643–646, oct 1986. ISSN 0028-0836. doi: 10.1038/323643a0.

P.a. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. M.R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004. ISSN 1474-175X. doi: 10.1038/nrc1299.A.

Ying Gao, Chengyang Song, Linping Hui, Chun-yu Li, Junying Wang, Ye Tian, Xu Han, Yong Chen, Da-Li Tian, Xueshan Qiu, and Enhua Wang. Overexpression of RNF146 in Non-Small Cell Lung Cancer Enhances Proliferation and Invasion of Tumors through the Wnt/β-catenin Signaling Pathway. *PLoS one*, 9(1), jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0085377.

Jochen B Geigl, Sabine Uhrig, and Michael R Speicher. Multiplex-fluorescence in situ hybridization for chromosome karyotyping. *Nature Protocols*, 1(3):1172–1184, sep 2006. ISSN 1754-2189. doi: 10.1038/nprot.2006.160.

Panagiota Giannakourous, Isabelle Matte, Claudine Rancourt, and Alain Piche. Transformation of NIH3T3 mouse fibroblast cells by MUC16 mucin (CA125) is driven by its cytoplasmic tail. *International Journal of Oncology*, 46(1):91–98, jan 2015. ISSN 1019-6439. doi: 10.3892/ijo.2014.2707.

A. S. Goldstein, J. Huang, C. Guo, I. P. Garraway, and O. N. Witte. Identification of a Cell of Origin for Human Prostate Cancer. *Science*, 329(5991):568–571, jul 2010. ISSN 0036-8075. doi: 10.1126/science.1189992.

Sheryl M Gough, Christopher I Slape, and Peter D Aplan. NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. *Blood*, 118(24): 6247–57, dec 2011. ISSN 1528-0020. doi: 10.1182/blood-2011-07-328880.

David H. Gutmann, Rosalie E. Ferner, Robert H. Listernick, Bruce R. Korf, Pamela L. Wolters, and Kimberly J. Johnson. Neurofibromatosis type 1. *Nature Reviews Disease Primers*, 3: 1–18, 2017. ISSN 2056676X. doi: 10.1038/nrdp.2017.4.

Michael Hajek, Andrew Sewell, Susan Kaech, Barbara Burtness, Wendell G. Yarbrough, and Natalia Issaeva. TRAF3/CYLD mutations identify a distinct subset of human papillomavirus-associated head and neck squamous cell carcinoma. *Cancer*, 123(10):1778–1790, may 2017. doi: 10.1002/cncr.30570.

Douglas Hanahan and Robert A Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144:646–674, 2011. doi: 10.1016/j.cell.2011.02.013.

Traver Hart and Jason Moffat. BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, 17(1):1–7, 2016. ISSN 14712105. doi: 10.1186/s12859-016-1015-8.

Masahiko Hibi, Masaaki Murakami, Mikiyoshi Saito, Toshio Hirano, Tetsuya Taga, and Tadamitsu Kishimoto. Molecular cloning and expression of an IL-6 signal transducer, gp130. *Cell*, 63(6):1149–1157, dec 1990. ISSN 00928674. doi: 10.1016/0092-8674(90)90411-7.

Lingmi Hou, Maoshan Chen, Xiaobo Zhao, Jingdong Li, Shishan Deng, Jiani Hu, Hongwei Yang, and Jun Jiang. FAT4 functions as a tumor suppressor in triple-negative breast cancer. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, nov 2016. ISSN 1423-0380. doi: 10.1007/s13277-016-5421-3.

V Y Hua, W K Wang, and P H Duesberg. Dominant transformation by mutated human ras genes in vitro requires more than 100 times higher expression than is observed in cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18): 9614–9, sep 1997. ISSN 0027-8424.

Peyton Hughes, Damian Marshall, Yvonne Reid, Helen Parkes, and Cohava Gelber. The costs of using unauthenticated, over-passaged cell lines: How much more data do we need? *BioTechniques*, 43(5):575–586, 2007. ISSN 07366205. doi: 10.2144/000112598.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Julio Saez-rodriguez, Ultan Mcdermott, and Mathew J Garnett. A Landscape of Pharmacogenomic Interactions in Resource A

Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, (166):740–754, 2016. doi: 10.1016/j.cell.2016.06.017.

Z Ivics, P B Hackett, R H Plasterk, and Z Izsvák. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, 91(4):501–10, nov 1997. ISSN 0092-8674.

Zoltán Ivics, Meng Amy Li, Lajos Mátés, Jef D Boeke, Andras Nagy, Allan Bradley, and Zsuzsanna Izsvák. Transposon-mediated genome manipulation in vertebrates. *Nature Methods*, 6(6):415–422, jun 2009. ISSN 1548-7091. doi: 10.1038/nmeth.1332.

John L Jainchill, Stuart A Aaronson, and George J Todaro. Murine Sarcoma and Leukemia Viruses: Assay Using Clonal Lines of Contact-Inhibited Mouse Cells. *Journal of Virology*, 4(5):549–553, 1969. ISSN 0022-538X.

Martin Jinek, Alexandra East, Aaron Cheng, Steven Lin, Enbo Ma, and Jennifer Doudna. RNA-programmed genome editing in human cells. *eLife*, 2:e00471, jan 2013. ISSN 2050-084X. doi: 10.7554/eLife.00471.

U N Kasid, R R Weichselbaum, T Brennan, G E Mark, and A Dritschilo. Sensitivities of NIH/3T3-derived clonal cell lines to ionizing radiation: significance for gene transfer studies. *Cancer research*, 49(12):3396–400, jun 1989. ISSN 0008-5472.

W. J Kent, C. W. Sugnet, T. S. Furey, and K. M. Roskin. The Human Genome Browser at UCSC W. *Journal of medicinal chemistry*, 19(10):1228–31, 1976. ISSN 0022-2623. doi: 10.1101/gr.229102.

Rhoda J. Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*, 2011:1–9, 2011. ISSN 17580463. doi: 10.1093/database/bar030.

A G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–3, apr 1971. ISSN 0027-8424.

Hiroko Koike-Yusa, Yilong Li, E-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke Yusa. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, 32(3):267–273, 2013. ISSN 1087-0156. doi: 10.1038/nbt.2800.

Nicole Lambertz, Nicolai El Hindy, Ilonka Kreitschmann-Andermahr, Klaus Peter Stein, Philipp Dammann, Neriman Oezkan, Oliver Mueller, Ulrich Sure, and Yuan Zhu. Down-regulation of programmed cell death 10 is associated with tumor cell proliferation, hyper-angiogenesis and peritumoral edema in human glioblastoma. *BMC Cancer*, 15(1):759, dec 2015. ISSN 1471-2407. doi: 10.1186/s12885-015-1709-8.

E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szus-

takowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001. ISSN 0028-0836. doi: 10.1038/35057062.

David Lane and Arnold Levine. p53 Research: the past thirty years and the next thirty years. *Cold Spring Harbor perspectives in biology*, 2(12), dec 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a000893.

T L Lasho, C M Finke, D Zblewski, M Patnaik, R P Ketterling, D Chen, C A Hanson, A Tefferi, and A Pardanani. Novel recurrent mutations in ethanolamine kinase 1 (ETNK1) gene in systemic mastocytosis with eosinophilia and chronic myelomonocytic leukemia. *Blood cancer journal*, 5(1):e275, jan 2015. ISSN 2044-5385. doi: 10.1038/bcj.2014.94.

Christine Leibiger, Nadezda Kosyakova, Hasmik Mkrtchyan, Michael Glei, Vladimir Trifonov, and Thomas Liehr. First Molecular Cytogenetic High Resolution Characterization of the NIH 3T3 Cell Line by Murine Multicolor Banding. *Journal of Histochemistry and Cytochemistry*, 2013.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp352.

Lei Li, Guo-Dong Zhao, Zhe Shi, Li-Li Qi, Li-Yuan Zhou, and Ze-Xian Fu. The Ras/Raf/MEK/ERK signaling pathway and its role in the occurrence and development of HCC. *Oncology letters*, 12(5):3045–3050, nov 2016. ISSN 1792-1074. doi: 10.3892/ol.2016.5110.

Ning Li, Maryam Yousefi, Angela Nakauka-Ddamba, Brian D Gregory, Zhengquan Yu, and Christopher J Lengner Correspondence. The Msi Family of RNA-Binding Proteins Function Redundantly as Intestinal Oncoproteins. 2015. doi: 10.1016/j.celrep.2015.11.022.

Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S Liu, Myles Brown, and X Shirley Liu. MAGeCK enables robust identification of essential genes from genome-scale CRISPR / Cas9 knockout screens. *Genome Biology*, 15(554):1–12, 2014. doi: 10.1186/s13059-014-0554-4.

Yunfang Li, Jing Pei, Hong Xia, Hengning Ke, Hongyan Wang, and Wufan Tao. Lats2, a putative tumor suppressor, inhibits G1/S transition. *Oncogene*, 22(28):4398–4405, 2003. ISSN 09509232. doi: 10.1038/sj.onc.1206603.

Pixu Liu, Hailing Cheng, Thomas M Roberts, and Jean J Zhao. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nature reviews Drug discovery*, 8(8):627–44, aug 2009. ISSN 1474-1784. doi: 10.1038/nrd2926.

H Lodish, A Berk, and HL Zipursky. *Molecular Cell Biology*. W. H. Freeman, New York, 4th editio edition, 2000.

Marcos Malumbres and Mariano Barbacid. RAS oncogenes: The first 30 years. *Nature Reviews Cancer*, 3(6):459–465, 2003. ISSN 1474175X. doi: 10.1038/nrc1097.

Iñigo Martincorena, Helen Davies, Michael R Stratton, Peter J Campbell, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, and Peter Van Loo. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 2017. doi: 10.1016/j.cell.2017.09.042.

Lajos Mátés, Zsuzsanna Izsvák, and Zoltán Ivics. Technology transfer from worms and flies to vertebrates: transposition-based genome manipulations and their future perspectives. *Genome biology*, 8 Suppl 1(Suppl 1):S1, 2007. ISSN 1474-760X. doi: 10.1186/gb-2007-8-s1-s1.

Andrea Mathe, Michelle Wong-Brown, Brianna Morten, John F. Forbes, Stephen G. Braye, Kelly A. Avery-Kiejda, and Rodney J. Scott. Novel genes associated with lymph node metastasis in triple negative breast cancer. *Scientific Reports*, 5(1):15832, dec 2015. ISSN 2045-2322. doi: 10.1038/srep15832.

William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), jun 2016. doi: 10.1186/s13059-016-0974-4.

Emmanouil Metzakopian, Alex Strong, Vivek Iyer, Alex Hodgkins, Liliana Antunes, Mathias J Friedrich, Qiaohua Kang, Teresa Davidson, Christina Hoffman, Gregory D Davis, George S Vassiliou, C William, and Allan Bradley. Enhancing the genome editing toolbox : genome wide CRISPR arrayed libraries. *Scientific Reports*, (April):1–9, 2017. doi: 10.1038/s41598-017-01766-5.

Chrysiis Michaloglou, Liesbeth C W Vredeveld, Maria S. Soengas, Christophe Denoyelle, Thomas Kuilman, Chantal M A M Van Der Horst, Donné M. Majoor, Jerry W. Shay,

Wolter J. Mooi, and Daniel S. Peeper. BRAFE600-associated senescence-like cell cycle arrest of human naevi. *Nature*, 436(7051):720–724, 2005. ISSN 00280836. doi: 10.1038/nature03890.

Brandon Milholland, Adam Auton, Yousin Suh, and Jan Vijg. Age-related somatic mutations in the cancer genome. *Oncotarget*, 6(28):24627–35, sep 2015. ISSN 1949-2553. doi: 10.18632/oncotarget.5685.

E C Miller and J A Miller. Mechanisms of chemical carcinogenesis. *Cancer*, 47(5):1055–64, mar 1981. ISSN 0008-543X.

Eva Marie Y Moresco, Xiaohong Li, and Bruce Beutler. Going forward with genetics: recent technological advances and forward genetics in mice. *The American journal of pathology*, 182(5):1462–73, may 2013. ISSN 1525-2191. doi: 10.1016/j.ajpath.2013.02.002.

Rebecca Nagy, Kevin Sweet, and Charis Eng. Highly penetrant hereditary cancer syndromes. *Oncogene*, 23(38):6445–6470, aug 2004. ISSN 0950-9232. doi: 10.1038/sj.onc.1207714.

Jean Charles Nault, Julien Calderaro, Luca Di Tommaso, Charles Balabaud, Elie Serge Zafrani, Paulette Bioulac-Sage, Massimo Roncalli, and Jessica Zucman-Rossi. Telomerase reverse transcriptase promoter mutation is an early somatic genetic alteration in the transformation of premalignant nodules in hepatocellular carcinoma on cirrhosis. *Hepatology*, 60(6):1983–1992, dec 2014. ISSN 02709139. doi: 10.1002/hep.27372.

J. R. Nevins. The Rb/E2F pathway and cancer. *Human Molecular Genetics*, 10(7):699–703, apr 2001. ISSN 14602083. doi: 10.1093/hmg/10.7.699.

Chitose Oneyama, Tomoya Hikita, Shigeyuki Nada, and Masato Okada. Functional dissection of transformation by c-Src and v-Src. *Genes to Cells*, 13(1):1–12, dec 2007. ISSN 13569597. doi: 10.1111/j.1365-2443.2007.01145.x.

Davide Pellacani, Misha Bilenky, Nagarajan Kannan, Alireza Heravi-Moussavi, David J.H.F. Knapp, Sitanshu Gakkhar, Michelle Moksa, Annaick Carles, Richard Moore, Andrew J. Mungall, Marco A. Marra, Steven J.M. Jones, Samuel Aparicio, Martin Hirst, and Connie J. Eaves. Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Reports*, 17(8):2060–2074, nov 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.10.058.

Charles M. Perou, Therese Sørlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, Douglas T. Ross, Hilde Johnsen, Lars A. Akslen,

Øystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X. Zhu, Per E. Lønning, Anne-Lise Børresen-Dale, Patrick O. Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, aug 2000. ISSN 0028-0836. doi: 10.1038/35021093.

W E Pierceall, L H Goldberg, M A Tainsky, T Mukhopadhyay, and H N Ananthaswamy. Ras gene mutation and amplification in human nonmelanoma skin cancers. *Molecular carcinogenesis*, 4(3):196–202, 1991. ISSN 0899-1987.

Larissa Pikor, Kelsie Thu, Emily Vucic, and Wan Lam. The detection and implication of genome instability in cancer. *Cancer and Metastasis Reviews*, 32(3-4):341–352, 2013. ISSN 01677659. doi: 10.1007/s10555-013-9429-5.

Julia R. Pon and Marco A. Marra. Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10(1):25–50, jan 2015. ISSN 1553-4006. doi: 10.1146/annurev-pathol-012414-040312.

S Pulciani, E Santos, L K Long, V Sorrentino, and M Barbacid. Ras gene amplification and malignant transformation. *Molecular and cellular biology*, 5(10):2836–41, oct 1985. ISSN 0270-7306.

Ying Qu, Bingchen Han, Yi Yu, Weiwu Yao, Shikha Bose, Beth Y Karlan, Armando E Giuliano, and Xiaojiang Cui. Evaluation of MCF10A as a Reliable Model for Normal Human Mammary Epithelial Cells. *PloS one*, 10(7), 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0131285.

Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq033.

R. Rad, L. Rad, W. Wang, J. Cadinanos, G. Vassiliou, S. Rice, L. S. Campos, K. Yusa, R. Banerjee, M. A. Li, J. de la Rosa, A. Strong, D. Lu, P. Ellis, N. Conte, F. T. Yang, P. Liu, and A. Bradley. PiggyBac Transposon Mutagenesis: A Tool for Cancer Gene Discovery in Mice. *Science*, 330(6007):1104–1107, nov 2010. ISSN 0036-8075. doi: 10.1126/science.1193004.

Wei Rao, Guohua Xie, Yong Zhang, Shujun Wang, Ying Wang, Huizhen Zhang, Feifei Song, Renfeng Zhang, Qinqin Yin, Lisong Shen, and Hailiang Ge. OVA66, a Tumor Associated Protein, Induces Oncogenic Transformation of NIH3T3 Cells. *PLoS one*, 9(3):e85705, mar 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0085705.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean Charles Sanchez, and Markus Müller. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 2011. ISSN 14712105. doi: 10.1186/1471-2105-12-77.

Carlota Rubio-Perez, David Tamborero, Michael P. Schroeder, Albert A. Antolín, Jordi Deu-Pons, Christian Perez-Llamas, Jordi Mestres, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell*, 27(3):382–396, 2015. ISSN 18783686. doi: 10.1016/j.ccell.2015.02.007.

Ignacio Sancho-Martinez, Emmanuel Nivet, Yun Xia, Tomoaki Hishida, Aitor Aguirre, Alejandro Ocampo, Li Ma, Robert Morey, Marie N. Krause, Andreas Zembrzycki, Olaf Ansorge, Eric Vazquez-Ferrer, Ilir Dubova, Pradeep Reddy, David Lam, Yuriko Hishida, Min-Zu Wu, Concepcion Rodriguez Esteban, Dennis O'Leary, Geoffrey M. Wahl, Inder M. Verma, Louise C. Laurent, and Juan Carlos Izpisua Belmonte. Establishment of human iPSC-based models for the study and targeting of glioma initiating cells. *Nature Communications*, 7:10743, feb 2016. ISSN 2041-1723. doi: 10.1038/ncomms10743.

Long Shen, Kunhua Qin, Dekun Wang, Yan Zhang, Nan Bai, Shengyong Yang, Yunping Luo, Rong Xiang, and Xiaoyue Tan. Overexpression of Oct4 suppresses the metastatic potential of breast cancer cells via Rnd1 downregulation. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(11):2087–2095, nov 2014. ISSN 0925-4439. doi: 10.1016/J.BBADIS.2014.07.015.

Liran I. Shlush, Sasan Zandi, Amanda Mitchell, Weihsu Claire Chen, Joseph M. Brandwein, Vikas Gupta, James A. Kennedy, Aaron D. Schimmer, Andre C. Schuh, Karen W. Yee, Jessica L. McLeod, Monica Doedens, Jessie J. F. Medeiros, Rene Marke, Hyeoung Joon Kim, Kwon Lee, John D. McPherson, Thomas J. Hudson, The HALT Pan-Leukemia Gene Panel Consortium, Andrew M. K. Brown, Fouad Yousif, Quang M. Trinh, Lincoln D. Stein, Mark D. Minden, Jean C. Y. Wang, and John E. Dick. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*, 506(7488):328–333, feb 2014. ISSN 0028-0836. doi: 10.1038/nature13038.

D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173, mar 1976. ISSN 0028-0836. doi: 10.1038/260170a0.

Tetsuya Taga and Tadamitsu Kishimoto. Gp 130 and the Interleukin-6 Family of Cytokines.

Annual Review of Immunology, 15(1):797–819, apr 1997. ISSN 0732-0582. doi: 10.1146/annurev.immunol.15.1.797.

Haiming Tang and Paul D Thomas. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics*, 203(2):635–47, 2016. ISSN 1943-2631. doi: 10.1534/genetics.116.190033.

Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, mar 2017. ISSN 0036-8075. doi: 10.1126/science.aaf9011.

George J Torado and Howard Green. Quantative studies of the growth of mouse embryo cells in culture and their development into established lines. *The Journal of Cell Biology*, 1963.

Elizabeth A Tovar and Carrie R Graveel. MET in human cancer: germline and somatic mutations. *Annals of translational medicine*, 5(10):205, may 2017. ISSN 2305-5839. doi: 10.21037/atm.2017.03.64.

Mai Tsutsui, Hirofumi Kawakubo, Testsu Hayashida, and Kazumasa Fukuda. Comprehensive screening of genes resistant to an anticancer drug in esophageal squamous cell carcinoma. *International Journal of Oncology*, (September):867–874, 2015. doi: 10.3892/ijo.2015.3085.

Christopher J. Walker, Mario A. Miranda, Matthew J. O'Hern, Joseph P. McElroy, Kevin R. Coombes, Ralf Bundschuh, David E. Cohn, David G. Mutch, and Paul J. Goodfellow. Patterns of CTCF and ZFHX3 mutation and associated outcomes in endometrial cancer. *Journal of the National Cancer Institute*, 107(11):1–8, 2015. ISSN 14602105. doi: 10.1093/jnci/djv249.

Wei Wang, Chengyi Lin, Dong Lu, Zeming Ning, Tony Cox, David Melvin, Xiaozhong Wang, Allan Bradley, and Pentao Liu. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, 105(27):920–9295, 2008.

Y-D Wang, N Cai, X-L Wu, H-Z Cao, L-L Xie, and P-S Zheng. OCT4 promotes tumorigenesis and inhibits apoptosis of cervical cancer cells by miR-125b/BAK1 pathway. *Cell death & disease*, 4(8), aug 2013. ISSN 2041-4889. doi: 10.1038/cddis.2013.272.

A R Wasylishen, A Stojanova, S Oliveri, A C Rust, A D Schimmer, and L Z Penn. New model systems provide insights into Myc-induced transformation. *Oncogene*, 30(34):3727–3734, aug 2011. ISSN 0950-9232. doi: 10.1038/onc.2011.88.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna M Shaw, A Brad, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764.The.

World Health Organisation. Cancer, feb 2018.

Wei Xia, Peter Bringmann, John McClary, Patrick P. Jones, Warren Manzana, Ying Zhu, Soujuan Wang, Yi Liu, Susan Harvey, Mary Rose Madlansacay, Kirk McLean, Mary P. Rosser, Jean MacRobbie, Catherine L. Olsen, and Ronald R. Cobb. High levels of protein expression using different mammalian CMV promoters in several cell lines. *Protein Expression and Purification*, 45(1):115–124, jan 2006. ISSN 10465928. doi: 10.1016/j.pep.2005.07.008.

Kiyoshi Yamaguchi, Michihiro Sakai, JooHun Kim, Shin-ichiro Tsunesumi, Tomoaki Fujii, Tsuneo Ikenoue, Yoshinao Yamada, Yoshiyuki Akiyama, Yasuhiko Muto, Rui Yamaguchi, Satoru Miyano, Yusuke Nakamura, and Yoichi Furukawa. MRG-binding protein contributes to colorectal cancer development. *Cancer Science*, 102(8):1486–1492, aug 2011. ISSN 13479032. doi: 10.1111/j.1349-7006.2011.01971.x.

Takaharu Yamamoto, Shinichiro Taya, and Kozo Kaibuchi. Ras-induced transformation and signaling pathway. *Journal of Biochemistry*, 126(5):799–803, 1999. ISSN 0021924X. doi: 10.1093/oxfordjournals.jbchem.a022519.

K. Yusa, L. Zhou, M. A. Li, A. Bradley, and N. L. Craig. A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences*, 108(4): 1531–1536, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1008322108.

Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, and Paul

Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(1):754–761, 2018. ISSN 13624962. doi: 10.1093/nar/gkx1098.

Ming-Jun Zhang, Yan-Yan Hong, and Na Li. Overexpression of Kin of IRRE-Like Protein 1 (KIRREL) in Gastric Cancer and Its Clinical Prognostic Significance. *Medical science monitor : international medical journal of experimental and clinical research*, 24:2711–2719, may 2018. ISSN 1643-3750. doi: 10.12659/MSM.910386.

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.

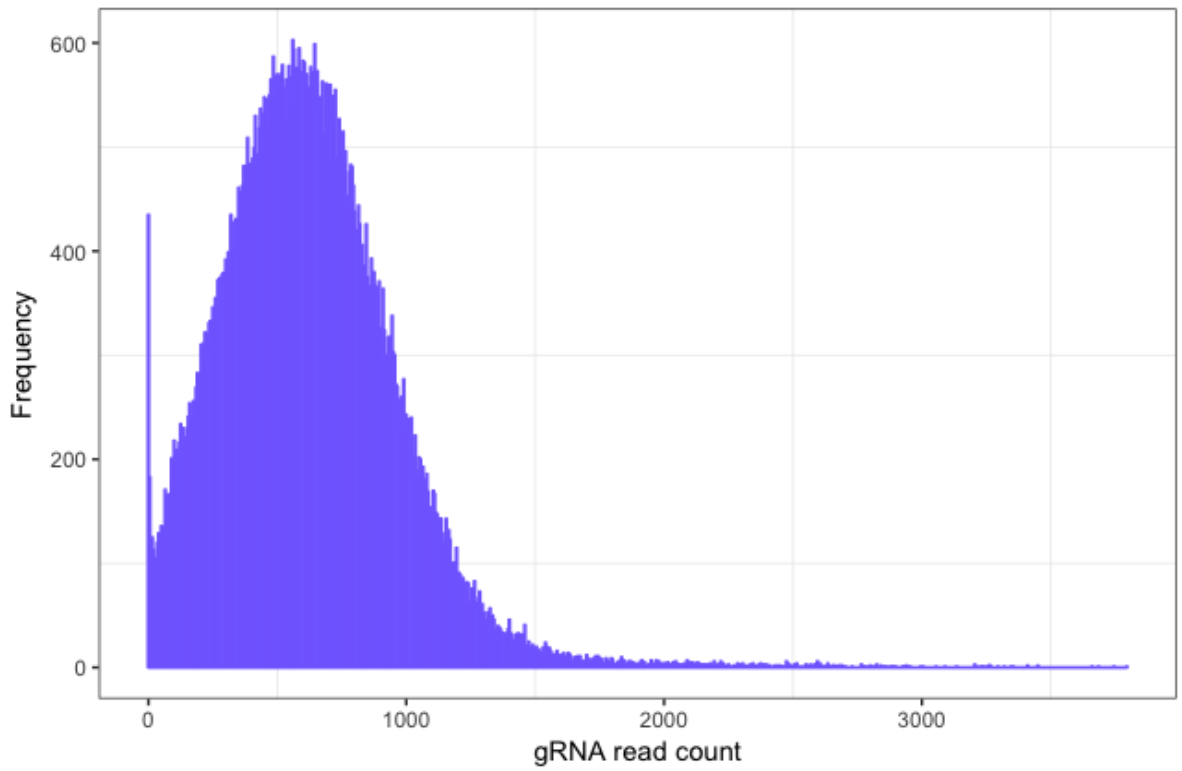


Figure B.1: **Histogram of gRNA read counts generated from sequencing of amplified Genome-wide Knockout CRISPR Library v2**
Genome-wide Knockout CRISPR Library v2 (Addgene #67988, (Koike-Yusa et al., 2013)) was amplified according to the depositor's instructions and then sequenced as detailed in section 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**