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.