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.

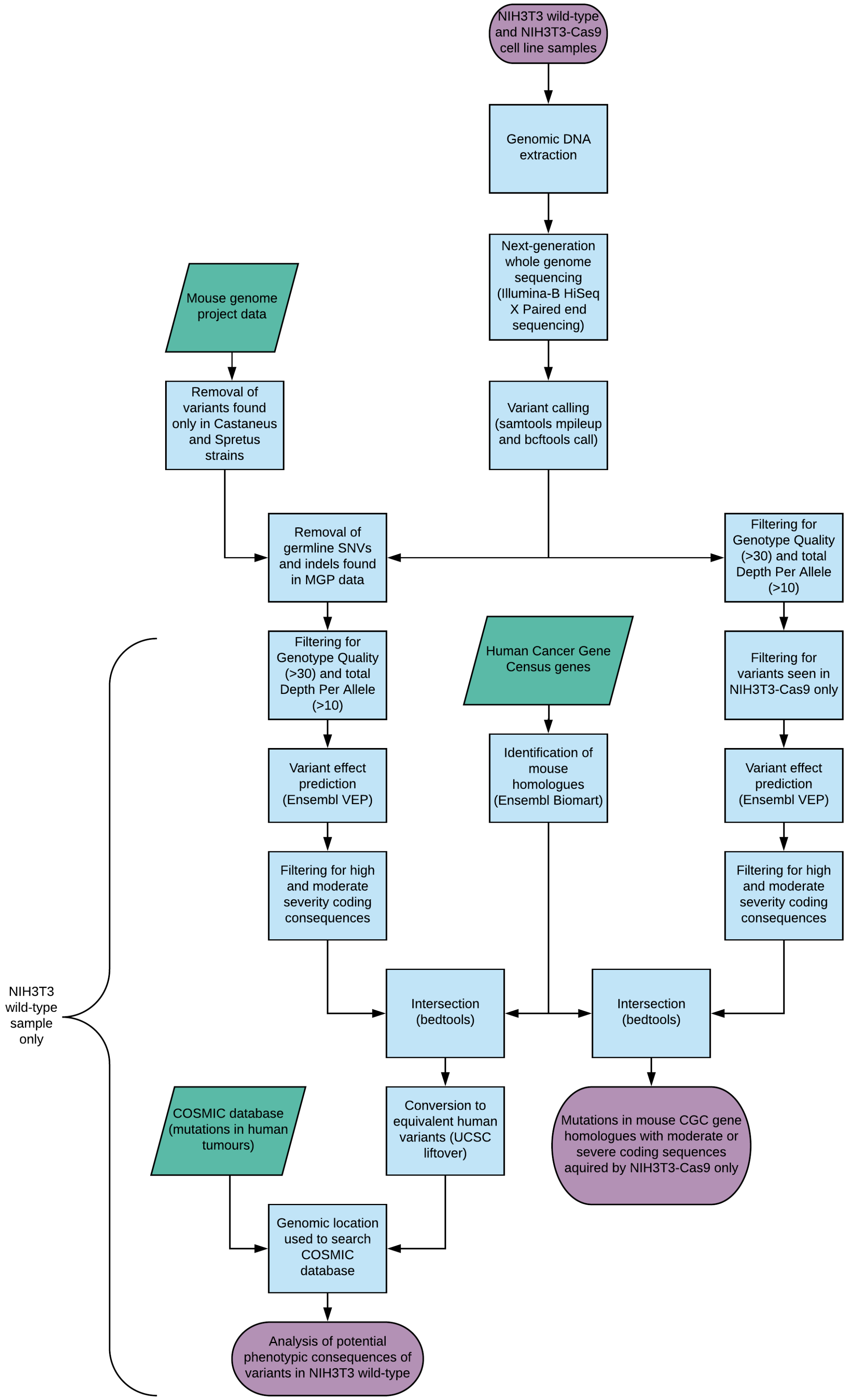
NIH3T3 wild-type and NIH3T3-Cas9 cell line samples
Genomic DNA extraction
Next-generation whole genome sequencing (Illumina-B HiSeq X Paired end sequencing)
Mouse genome project data
Removal of variants found only in Castaneus and Spretus strains
Variant calling (samtools mpileup and bcftools call)
Removal of germline SNVs and indels found in MGP data
Filtering for Genotype Quality (>30) and total Depth Per Allele (>10)
Filtering for Genotype Quality (>30) and total Depth Per Allele (>10)
Human Cancer Gene Census genes
Filtering for variants seen in NIH3T3-Cas9 only
Variant effect prediction (Ensembl VEP)
Identification of mouse homologues (Ensembl Biomart)
Variant effect prediction (Ensembl VEP)
Filtering for high and moderate severity coding consequences
Filtering for high and moderate severity coding consequences
NIH3T3 wild-type sample only
Intersection (bedtools)
Intersection (bedtools)
COSMIC database (mutations in human tumours)
Conversion to equivalent human variants (UCSC liftover)
Mutations in mouse CGC gene homologues with moderate or severe coding sequences aquired by NIH3T3-Cas9 only
Genomic location used to search COSMIC database
Analysis of potential phenotypic consequences of variants in NIH3T3 wild-type

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).

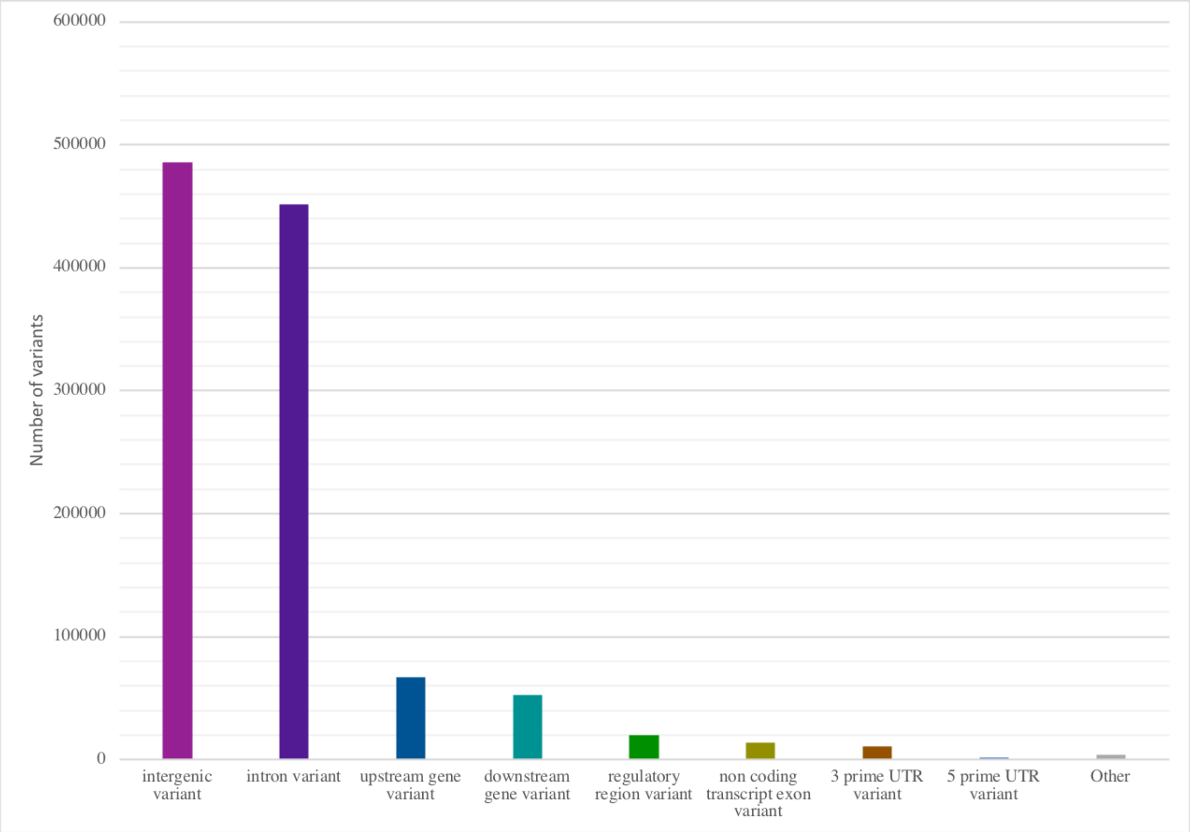

Figure 2.2: **Putative non-germline variants in NIH3T3 wild-type by consequence type** Non-germline variants in NIH3T3 wild-type were identified by filtering to exclude variants at positions that were polymorphic in the Mouse Genome Project (Adams et al., 2015) data on 36 inbred laboratory mouse strains. 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_varian 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

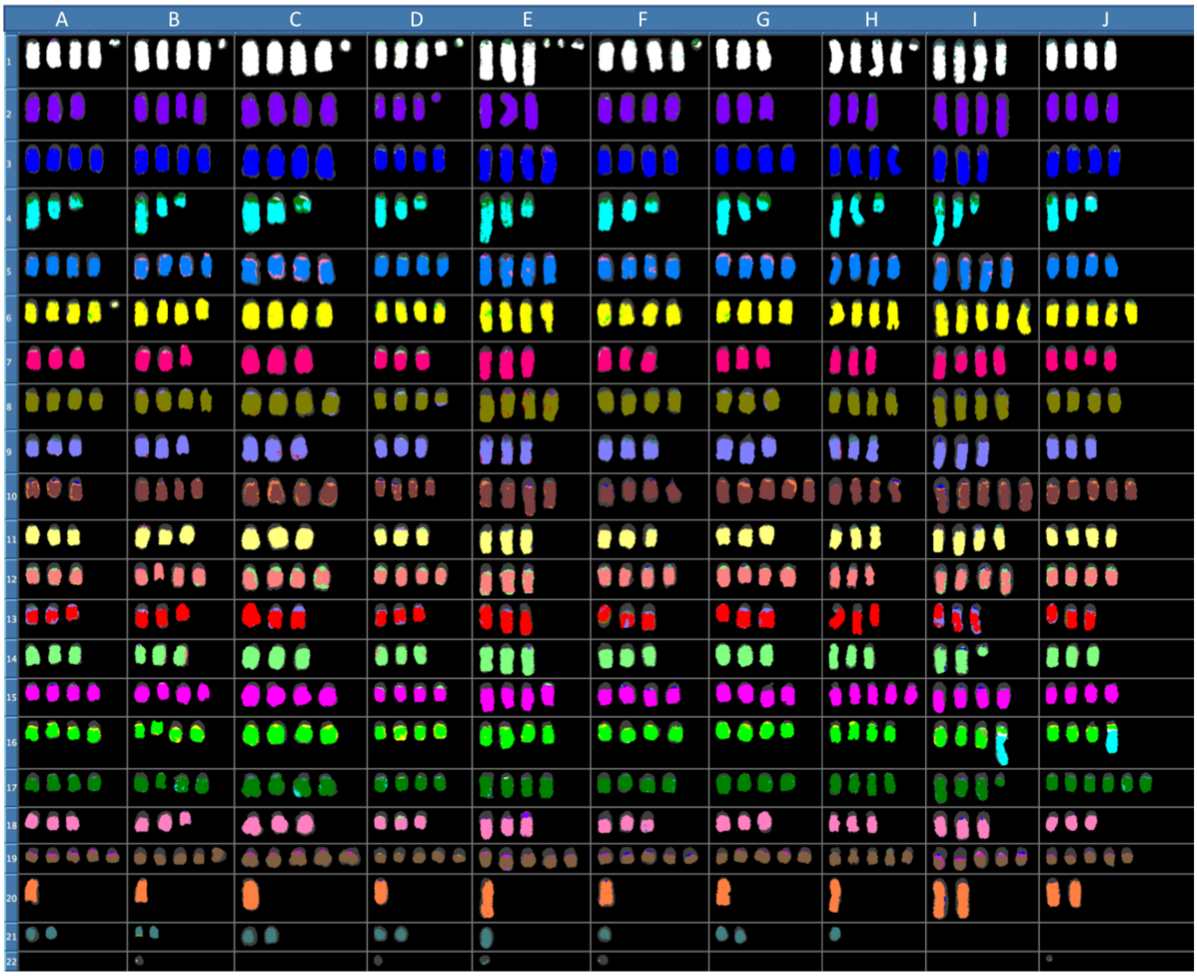

Figure 2.3: **Multiplex Fluorescence *In Situ* Hybridisation of 10 randomly selected NIH3T3-Cas9 cells at metaphase**
10 metaphase chromosome spreads (A-J) were randomly selected from a pool of NIH3T3-Cas9 cells and karyotyped using Multiplex Fluorescence *In Situ* Hybridisation with 21-colour mouse 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

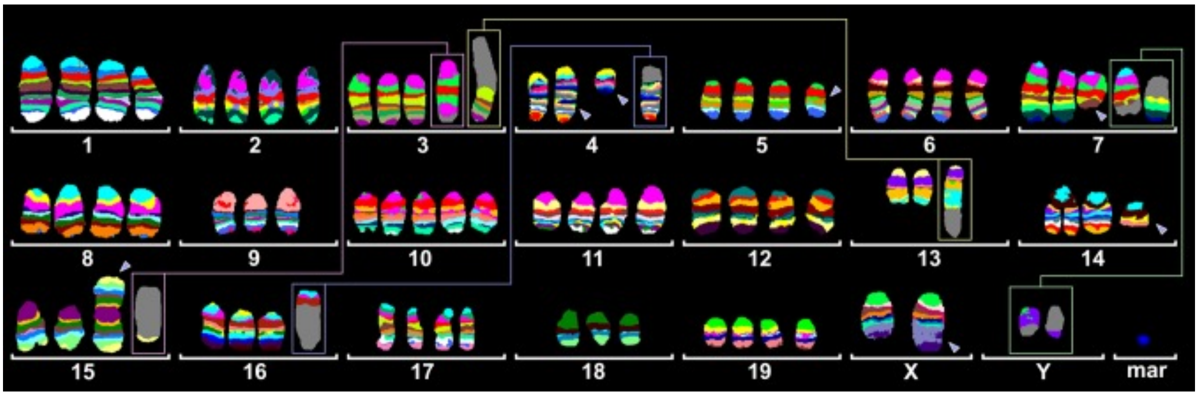

Figure 2.4: **Results of NIH3T3 karyotyping by murine multicolour banding from Leibiger et al., 2013**
Murine multicolour banding was applied to 25 individual metaphase NIH3T3 chromosome spreads. This figure shows the typical pseudocolour banding pattern for each chromosome (karyotype shown is present 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 sequence 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.