Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Bioinformatic identification of novel melanoma drug targets and investigation of potential inhibitors using high throughput screening

Alexander Joseph Trevarton

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Molecular Medicine, The University of Auckland, 2017
Abstract

A vast and growing amount of genomic, molecular and clinical information about melanoma is available to researchers. This information may be especially valuable for the identification of novel drug targets. However, due to the challenge of assembling, integrating and interpreting this dispersed information, it is difficult to leverage to use for drug target prioritisation or to aid rational experimental design. In this project, multiple sources of information about melanoma have been integrated into a comprehensive bioinformatics database. This database allows complex links to be drawn between somatic variants in individual melanomas revealed by DNA sequencing, associations between gene expression in melanoma and patient outcome, data concerning drug targets, biomarkers, protein druggability, the biomedical literature and clinical trials.

While recent years have seen the advent of many new therapies for melanoma including the immune checkpoint inhibitors, factors such as tumour heterogeneity and drug resistance mean that treatment remains challenging. Development of new, varied therapeutic options are still urgently required. Therefore, the melanoma database was used to identify a short list of putative melanoma drug targets. From this list, one candidate, the transcription factor YB-1, was selected as a target for high throughput screening.

YB-1 expression in melanoma has a strong association with patient outcome and plays distinct roles in multiple aspects of cancer biology. This makes YB-1 a valuable but potentially difficult target, requiring a multi-pronged approach to identify lead compounds. Two novel high throughput assays were developed to detect compounds that might interfere with YB-1 binding to nucleic acid. The first assay is an in vitro cell-based, luciferase reporter gene assay that detects the activation of the E2F1 promoter fragment by YB-1. The second assay is based on the AlphaScreen system, detecting YB-1 binding to a single-stranded DNA oligonucleotide. These orthogonal assays complement one another by measuring the binding of YB-1 to two discrete nucleic acid sequences with completely different read-outs.

These assays were used in high throughput screening of over 7,000 small molecule compounds to identify eight compounds for follow-up study and development.
Acknowledgements

I would like to thank my supervisors Professor Cristin Print and Dr Annette Lasham for the advice, guidance, feedback and continuous encouragement during my PhD. This thesis could not have been completed without their knowledge and valuable input.

I would like to thank Professor Wang Ming-Wei for his support and generosity that allowed the high throughput screen to be performed, and supported my stay at the Chinese National Compound Library (CNCL) in Shanghai. I would also like to thank him for his feedback and insight into the project management of a screen.

I would like to thank Yang Dehua, Zhou Yan and the team at the laboratory at CNCL for their collaboration at the lab bench, and also for making my stay in Shanghai a pleasant one.

I would like to thank my advisor Dr Maggie Kalev for her suggestions and chats that helped me step back to see the overall picture.

I would like to thank all of the members of the Print lab for their continuous friendliness and willingness to help, that make the Print lab a pleasant working environment. Thanks also go to all of my friends and fellow students in the surrounding labs here on the 2nd floor.

I would like to thank my parents for their support, encouragement and belief in me.

Finally, I would like to thank my wonderful wife, Catherine, for accompanying me on this PhD journey, giving me daily support, and who inspires me to keep going and give it my best.
# Table of Contents

ABSTRACT.................................................................................................................................II

ACKNOWLEDGEMENTS..................................................................................................................III

TABLE OF CONTENTS ........................................................................................................................ IV

LIST OF FIGURES................................................................................................................................. VIII

LIST OF TABLES........................................................................................................................................ XI

ABBREVIATIONS........................................................................................................................................... XIII

CHAPTER 1 INTRODUCTION...................................................................................................................... 1

1.1. OUTLINE ............................................................................................................................................ 1

1.2. MELANOMA .................................................................................................................................... 1

1.2.1. Melanoma overview ......................................................................................................................... 1

1.2.2. Incidence ....................................................................................................................................... 2

1.2.3. Melanoma risk factors ....................................................................................................................... 2

1.2.4. Signalling pathways altered in melanoma .......................................................................................... 3

1.3. MELANOMA DATA ............................................................................................................................ 8

1.3.1. The growth and complexity of melanoma data ................................................................................... 8

1.3.2. Integration of separate data types mutually enhances the interpretation of each ................................ 9

1.3.3. Potential clinical use of molecular pathway data from individual tumours ...................................... 10

1.4. INTRODUCTION TO THERAPY FOR MELANOMA.......................................................................... 12

1.4.1. Targeted therapy of melanoma ......................................................................................................... 13

1.4.2. Drug resistance ............................................................................................................................... 16

1.5. DRUG DEVELOPMENT: TARGET AND LEAD IDENTIFICATION .................................................... 21

1.5.1. Targets and leads as starting points ................................................................................................... 21

1.5.2. High throughput screening to identify lead compounds ................................................................... 22

1.5.3. Biomarkers ..................................................................................................................................... 24

1.6. THESIS PROJECT OBJECTIVE ....................................................................................................... 24

CHAPTER 2 MATERIALS AND METHODS.............................................................................................. 26

2.1. INTRODUCTION .................................................................................................................................. 26

2.2. MATERIALS ....................................................................................................................................... 26

2.2.1. Molecular biology .......................................................................................................................... 26

2.2.2. Cell culture .................................................................................................................................... 31

2.2.3. YB-1 binding nucleic acid assays ................................................................................................... 31

2.3. METHODS ......................................................................................................................................... 32

2.3.1. Bioinformatics ............................................................................................................................... 32

2.3.2. Molecular biology .......................................................................................................................... 36

2.3.3. Cell culture .................................................................................................................................... 40

2.3.4. YB-1 binding nucleic acid assays ................................................................................................... 46

2.3.5. Computational filtering ................................................................................................................. 53

CHAPTER 3 DATABASE DEVELOPMENT AND CREATION OF MELANOMADB ............................... 54

3.1. INTRODUCTION ............................................................................................................................... 54

3.1.1. Basis of the database ....................................................................................................................... 54

3.1.2. Existing database tools for study of cancer genomics .................................................................... 54
3.1.3. Genomic data .......................................................... 55
3.1.4. Consultation with researchers and clinicians to guide collection of data ........................................... 56
3.2. DATA COLLECTION AND PROCESSING ........................................ 57
3.2.1. Overview .................................................................. 57
3.2.2. RNA expression data ...................................................... 57
3.2.3. Drug and biomarker data ................................................. 61
3.2.4. Druggability data .......................................................... 62
3.2.5. Literature and genomic data relationship information .......................................................... 63
3.2.6. Somatic variant data ....................................................... 64
3.3. MELANOMADB .................................................................. 66
3.3.1. Amalgamation of all data into gene sets ............................................. 66
3.3.2. SQL database generation ............................................... 67
3.3.3. Assembly of information for individual tumours ......................................................... 67
3.3.4. Data visualisation ........................................................... 68
3.4. MELANOMA PROFILER ...................................................... 68
3.5. DISCUSSION ................................................................ 68
3.5.1. Integration of high level data ............................................... 68

CHAPTER 4 DATABASE APPLICATION ..................................... 72
4.1. INTRODUCTION ............................................................... 72
4.2. INSIGHTS INTO MELANOMA OBTAINED FROM AN OVERVIEW OF COLLECTED MELANOMA SAMPLES ........................................ 72
4.2.1. Genes most frequently altered in metastatic melanoma tumours .................................................. 75
4.2.2. Putative melanoma tumour suppressors and oncogenes ......................................................... 75
4.2.3. Gene alterations in sun-exposed primary tumours .......................................................... 75
4.2.4. Melanoma gene network .................................................. 78
4.2.5. Cross-validation of database use examples with TCGA melanoma data ........................................ 78
4.3. ANALYSIS OF SIGNALLING PATHWAYS RELEVANT TO MELANOMA VIA GENESETDB AND MELANOMADB 84
4.3.1. Incorporation of KEGG signalling pathways into MelanomaDB ................................................... 84
4.3.2. Analysis of melanoma signalling pathways in individual tumours .............................................. 85
4.4. PUTATIVE MELANOMA DRUG TARGETS ................................. 88
4.4.1. Inhibition of an over-active, wild-type target ................................................................................. 89
4.4.2. RNA expression associated with metastasis or patient survival ............................................... 89
4.4.3. Selection for targets predicted to be druggable ............................................................................. 90
4.4.4. Direct or implicit literature relationship to 'metastatic melanoma' ............................................. 90
4.4.5. Corroboration of 14 putative drug targets using new data ..................................................... 91
4.4.6. Literature connecting cancer to 14 putative drug targets and selection of single target for further investigation ............................................................................. 92
4.5. INTRODUCTION TO YB-1 ..................................................... 95
4.5.1. Protein Structure ............................................................. 95
4.5.2. YB-1 in cancer ................................................................ 96
4.5.3. Possible approaches to drug YB-1 ...................................................................................... 99

CHAPTER 5 DEVELOPMENT OF AN IN VITRO CELL-BASED ASSAY FOR YB-1
TRANSSCRIPTIONAL ACTIVITY ................................................. 101
5.1. INTRODUCTION TO ASSAY DEVELOPMENT ............................... 101
5.2. ASSAY OVERVIEW .................................................................. 101
5.2.1. Modulation of transcription factor activity .................................................................................. 102
5.3. OPTIMISATION OF TRANSFECTION CONDITIONS FOR THE ASSAY ........................................ 102
5.3.1. Cell number per well .......................................................... 102
5.3.2. Kill curve assay .................................................................. 103
List of Figures

Figure 1-1 A highly simplified overview of major pathways involved in melanoma ..........6
Figure 1-2 Plotted are the cumulative number of all data series (left panel) and individual samples (right panel) deposited into the GEO database between the years 2001 and 2016 inclusive...............................................................8
Figure 1-3 Project overview. Flow-chart summarising the progression of the project and defining the focus of each Chapter.................................................................25
Figure 2-1 Overview of the transformation of E.coli by DNA plasmids........................38
Figure 2-2 Overview of transfection for high-throughput screening, followed by addition of screening compounds and measurement of luciferase .........................44
Figure 2-3 Dot blot procedure. Described is the treatment of nitrocellulose membrane, following attachment of oligonucleotides by UV cross-linking..........................49
Figure 2-4 50μL AlphaScreen reaction procedure .....................................................50
Figure 2-5 Final AlphaScreen procedure ................................................................51
Figure 2-6 TruHits assay procedure ....................................................................52
Figure 2-7 Procedure for make-shift assay for false positives.................................53
Figure 4-1 An overview of the genes mutated across the 310 cell-line, primary and metastatic melanoma samples .............................................................................75
Figure 4-2 The proportion of somatic variant-containing tumours derived from sun-exposed primary tumours (y-axis) compared to the proportion of somatic variant-containing tumours derived from non sun-exposed primary tumours (x-axis). ..........77
Figure 4-3 Putative melanoma gene network ..........................................................78
Figure 4-4 Percentage of TCGA and MelanomaDB melanoma samples with protein-altering variants for each gene ........................................................................78
Figure 4-5 Analyses produced lists of genes with mRNA expression associated with patient survival for three datasets: TCGA (blue circle), Bogunovic et al (red circle), Jonsson et al (green circle) .......................................................................................80
Figure 4-6 The proportion of somatic variant-containing tumours derived from sun-exposed primary tumours (y-axis) compared to the proportion of somatic variant-containing tumours derived from non sun-exposed primary tumours (x-axis). Data shown here is from the melanoma samples contained within TCGA ........................................80
Figure 4-7 KEGG pathway ‘Melanoma’ with genes shaded red according to the percentage of melanoma samples in MelanomaDB that contain somatic variations ..........................85
Figure 4-8 Clustered heatmap for genes encoding proteins of the KEGG “Melanoma” signalling pathway ....................................................................................86
Figure 4-9 Neuroactive ligand-receptor interaction pathway for tumour sample ME029......88
Figure 4-10 Outline of the sequential filtering performed on the database to generate a short list of 14 putative melanoma drug targets ......................................................91
Figure 5-1. Kill curve for HCT116 cells exposed to a range of G418 concentrations. Cellular confluence was assessed visually at 96 hours after addition of G418 to culture media. Error bars represent the standard error of observations over three replicate wells. Live
cells were not found in any wells exposed to G418 concentrations equal to or greater than 1.25mg/mL................................................................. 103

Figure 5-2 The E2F1 promoter:luciferase reporter gene activity in the presence of variable levels of E2F1 and YB-1, across 3 experiments................................................................. 107

Figure 5-3 Results from two experiments showing that the Renilla luciferase luminescent signal, driven by the TK promoter, is increased in the presence of YB-1 over-expression, and decreased in the presence of 1xCohen competitor oligonucleotide, that inhibits YB-1 activity .................................................................................. 109

Figure 5-4 Synergistic activation of the E2F1 promoter by E2F1 and YB-1.............................. 111

Figure 5-5 Average luminescent signal for transfections using a range of ratios of YB-1 to E2F1 over-expressing plasmids.............................. 113

Figure 5-6 Luminescent signal decreases as competitor amount per well is increased....... 115

Figure 5-7 Luminescent signal decreases with increasing time between transfection and luciferase activity measurement, regardless of the time of DMSO addition................................. 119

Figure 5-8 The standard deviation of the luminescent signal decreases with increasing time between the transfection and the measurement of luciferase, and is influenced by the time at which DMSO is added to the cells. ................................................................. 120

Figure 5-9 Luminescent signal standard deviation decreases with increasing time between DMSO addition and measurement of luciferase........................................................................ 121

Figure 5-10 Within four hours of the addition of compounds, luminescent signal is decreased. No decrease in luminescent signal is seen when luminescent signal is measured immediately after compounds are added at 24 hours. ............................................ 123

Figure 6-1 Graphical overview of the AlphaScreen assay..................................................... 127

Figure 6-2 Graphical overview of the TruHits assay. Streptavidin-coated Donor beads bind biotinylated Acceptor beads to form a complex that emits light at 570nm upon laser irradiation at 680nm. Compounds that interfere with the assay system will reduce the intensity of this emitted light................................................................. 128

Figure 6-3 Overview of the AlphaScreen assay, showing points at which false positives can interfere with the assay system (red arrows). ................................................................. 130

Figure 6-4 Dot blot of YB-1 binding to 2 different YB-1 binding site oligonucleotides..... 132

Figure 6-5 Titration of 1xCohen competitor from concentrations ranging from 0 to 1000 fmol/μL................................................................. 141

Figure 6-6 Results of an AlphaScreen assay evaluating the effect of variable BSA levels on YB-1 binding to 3xCohen oligonucleotide, and non-specific binding between other reaction components................................................................. 142

Figure 6-7 AlphaScreen signal across a range of concentrations of two cellular lysates to determine the lysate concentration at which luminescent signal is highest. ................................ 143

Figure 6-8 Luminescent signal in the AlphaScreen assay is decreased by increasing concentrations of DMSO................................................................. 144

Figure 6-9 Comparison of AlphaScreen signal between Acceptor beads NZ3 and Shanghai 1 (SH1), including different negative controls................................................................. 148

Figure 6-10 Luminescent AlphaScreen signals from the comparison between Acceptor beads conjugated in New Zealand (NZ3), and those conjugated in Shanghai (SH2-8) in three batches (SH2, SH3-5 and SH6-8)................................................................. 148
Figure 7-1 Graph showing the percentage inhibition of 23 compounds........................................ 152
Figure 7-2 Venn diagram showing the compounds hits ................................................................. 153
Figure 7-3 Representation of the sequential filtering of compounds interfering with
YB1:ssDNA binding, as tested by screening through AlphaScreen assays. ......................... 154
Figure 7-4 Plot showing the distribution of percentage signal reduction resulting from
inhibition of luciferase activity by 7,360 compounds across 23 384-well plates used
during primary screening. .................................................................................................................. 159
Figure 7-5 The distribution of percentage signal reduction resulting from inhibition of
luciferase activity by 342 compounds tested in a validation screen following their
identification as hits during the primary screen. The percentage decrease in luminescent
signal in the population had a median of 62 and a standard deviation of 61.78.................. 160
Figure 7-6 Plot showing the distribution of percentage inhibition of luminescent signal by
251 compounds screened in an AlphaScreen assay using purified YB-1 binding to
3xCohen oligonucleotide. The population had a median of 32, and a mean of 33.2 with a
standard deviation of 32.77. .................................................................................................................. 162
Figure 7-7 Graphical overview of screening process. ................................................................. 164
List of Tables

Table 1-1-1 Examples of targeted drugs used in the treatment of metastatic melanoma .......... 16
Table 2-1 Composition of buffers used in in vitro protein-nucleic acid binding assays .......... 26
Table 2-2 HEPES buffers used in protein-nucleic acid binding assays .................................. 27
Table 2-3 PBS buffers .............................................................................................................. 27
Table 2-4 Plasmids ................................................................................................................... 29
Table 2-5 Oligonucleotide primer sequences .......................................................................... 30
Table 2-6 YB-1 binding sequences ......................................................................................... 30
Table 2-7 Human cell lines and culture media in which they were grown .............................. 31
Table 2-8 List of antibodies used in this study ........................................................................ 32
Table 2-9 Different transfection conditions used. For both forward and reverse transfection methods, different volumes of LF2000 and LF3000 were used ........................................ 41
Table 2-10 Restriction digest by KpnI ..................................................................................... 48
Table 3-1 Sources of RNA expression data ............................................................................. 58
Table 3-2 Sources of melanoma biomarkers data .................................................................... 62
Table 3-3 Sources of druggability data .................................................................................... 63
Table 3-4 Sources of somatic variant data ............................................................................... 64
Table 5-1 Transfection efficiencies for HCT116 and A375 cells under different transfection methods, transfection reagents and transfection reagent volumes ........................................... 104
Table 5-2 Different combinations of plasmid and oligonucleotide used in 24-well plates .. 106
Table 5-3 Transfection conditions used in 96-well plate wells to test synergistic activation of E2F1 promoter by simultaneous E2F1 and YB-1 over-expression .................................. 110
Table 5-4 Average luminescent signals from cells transfected with 80 or 100ng of pGL4.17-E2F1-728 plasmid per well, in the presence or absence of the 1xCohen competitor.... 116
Table 5-5 Average and standard deviation of luminescent signal at different times between transfections using 50 or 80ng of pGL4.17-E2F1-728 per well, with and without inhibition by 50pmol of 1xCohen competitor per well. ........................................ 117
Table 5-6 Experiment using different volumes of cell culture media in 384-plate wells shows that reaction volume does not have a strong effect on the variation of luminescent signal. ........................................................................................................ 124
Table 5-7 Coefficient of variation of luminescent signal decreases with increasing numbers of cells seeded per well. ........................................................................................................ 124
Table 6-1 A comparison of two buffers used in AlphaScreen experiments detecting YB1 binding to 3xCohen oligonucleotide ................................................................................. 134
Table 6-2 AlphaScreen assay testing a range of both 3xCohen oligonucleotide and purified YB-1 protein concentrations ......................................................................................... 136
Table 6-3 AlphaScreen assay testing a range of both 3xCohen oligonucleotide and purified YB-1 protein concentrations, performed using two batches of acceptor bead ........... 136
Table 6-4 AlphaScreen assays testing a range of both acceptor bead and donor bead concentrations, performed using two batches of acceptor bead ........................................... 138
Table 7-1 Showing the variation in Z-factor across the 23 plates used during primary screening. ........................................................................................................................ 157
Table 7-2 Percentage reduction in luminescent signal during screening assay on 8 “true” positive compounds. ........................................................................................................... 162
Abbreviations

CDK  cyclin-dependant kinase
CNCL  Chinese National Compound Library
COSMIC  Catalogue of Somatic Mutations in Cancer
DMSO  dimethyl sulfoxide
ECM  Extracellular matrix
GEO  Gene Expression Omnibus
GFP  green fluorescent protein
GPCR  G protein-coupled receptor
MAPK  mitogen-activated protein kinase
MHCII  major histocompatibility complex II
MMP  matrix metallopeptidase
NCBI  National Centre for Biotechnology Information
NGS  Next-generation sequencing
PI3K  phosphatidylinositol 3-kinase
PIK3CA  phosphoinositide-3-kinase, catalytic, alpha polypeptide
PKA  protein kinase A
PTEN  Phosphatase and tensin homolog
RTK  receptor tyrosine kinases
TCGA  The Cancer Genome Atlas
UV  ultraviolet
Chapter 1 Introduction

1.1. Outline

This Introduction Chapter begins with an overview of melanoma and a brief description of the signalling pathways that are altered in this disease, and the insights in melanoma biology provided by genomic studies. Following this is an explanation of the challenge presented by the immense amount of available melanoma data, and the potential benefits to be had through integrated utilisation of this resource. Then, melanoma therapy is discussed, with emphasis on the problem of drug resistance, and the need for more therapeutic options. Finally, the process of drug target identification is briefly introduced, before summarising the direction taken by this project.

1.2. Melanoma

1.2.1. Melanoma overview

Melanoma, also termed malignant melanoma, results from the neoplastic growth of melanocytes, neural crest derived cells predominantly found in the basal layer of the epidermis, but also in the eye, inner ear, meninges, bones and heart. In response to ultraviolet (UV) radiation, melanocytes produce melanin, which provides protection from UV radiation-induced DNA damage by absorbing UV radiation. Benign neoplasms derived from melanocytes (naevi, singular: nevus, commonly called moles,) may develop into malignant melanoma. In a seminal paper, Clark et al. described this development as a step-wise transformation. In the first step, the radial (or horizontal) growth phase, the nevus grows peripherally within the epidermis. An invasive sub-set of these cells will proceed to the vertical growth phase, growing into the dermis. Melanoma cells acquire the ability to metastasise most commonly during the vertical growth phase and less commonly during the radial growth phase. They do this by dissociating from the primary tumour and migrating through the stroma to invade lymphatic and blood vessels before forming a metastatic tumour at a distant site. Over the years, melanoma researchers have built on these basic ideas, understanding progressively more about the cascades of histological, molecular and genomic changes that occur during melanoma genesis and development. In particular, they have
gained an understanding of the highly metastatic nature of melanoma, possibly related to the origin of melanocytes as highly motile neural crest cells.

1.2.2. Incidence

Melanoma is the fourth most commonly registered cancer in New Zealand and the sixth most common cause of cancer death. While the more common and less life threatening forms of skin cancer, basal cell carcinoma and squamous cell carcinoma, do not require registration in New Zealand, data available from the United Kingdom suggest that melanoma is responsible for over 75% of skin cancer deaths despite comprising only 2.3% of skin cancers. New Zealand and Australia have the highest melanoma incidence in the world, possibly due to higher levels of ultraviolet radiation. The mortality rate for New Zealand males is 7.8 deaths per 100,000 males, more than twice that for females. In males aged 25 to 44, melanoma is the most common cancer, and the second most common cause of cancer death.

While the incidence of melanoma is rising globally, it may possibly have reached a peak, and a future decline in mortality has been predicted by some researchers, due to reduced childhood UV radiation exposure in more recent generations, improved surveillance, better diagnostic imaging and surgical techniques and new treatment options available to medical oncologists. These include first the targeted therapies such as Vemurafenib and more recently in New Zealand the limited Government funding of immune checkpoint inhibitor drugs such as Nivolumab. Nevertheless, melanoma will remain a devastating and important tumour to address in New Zealand through laboratory research and through advances in clinical care for the foreseeable future.

1.2.3. Melanoma risk factors

1.2.3.1. Environmental factors

A comprehensive meta-analysis found that intermittent sun exposure and history of sunburn, which is a marker of intermittent sun exposure, doubled melanoma risk. However, interestingly, an inverse association has been observed between melanoma and high continuous sun exposure – this remains to be fully explained.

UV light can cause C to T/CC to TT mutations. These direct C to T transitions represent the most obvious type of mutation in melanoma related to sun exposure, however they are just one of many types of mutation found in malignant melanoma. For example, one of the
most studied driver mutations in melanoma is the BRAF gene T to A transversion at nucleotide 1796\textsuperscript{11}, which may in some cases occur through indirect mechanisms such as UV-induced inflammation\textsuperscript{12}.

1.2.3.2. Patient factors

A family history increases the risk for melanoma by 1.7 to 2 times according to epidemiological meta-analyses\textsuperscript{13,14}. Several studies have suggested that higher numbers of naevi, and clinically atypical naevi, are associated with increased melanoma risk\textsuperscript{15-17}, although two recent studies found that some patterns of high naevi numbers may be associated with more favourable prognosis\textsuperscript{18,19}. Light-coloured eyes, skin and hair are associated with increased relative risk for melanoma\textsuperscript{13} - phenotypic traits also associated with sensitivity to UV radiation\textsuperscript{15}.

Patients with a family history may also have inherited highly penetrant melanoma susceptibility genetic variants. Melanoma susceptibility variants are known to exist in genes including CDKN2A\textsuperscript{6}, CDK4\textsuperscript{20}, MITF\textsuperscript{21}, POT1\textsuperscript{22,23}, TERT\textsuperscript{24}, BAP1\textsuperscript{25}, BRCA2\textsuperscript{26} and MC1R\textsuperscript{27}.

Germ-line mutations in these high penetrance genes are implicated in approximately 50% of familial melanoma cases (with 20-40% attributable to CDKN2A), with the remainder likely due to polygenic contribution of multiple low-risk alleles\textsuperscript{28}. Such alleles may be potential targets for therapeutic intervention. Identifying new melanoma therapeutic targets is a major theme of Chapter 4 of this thesis.

Immunosuppression is also a relative risk factor for melanoma. For example, for immunosuppressed transplant recipients the incidence of melanoma is increased 2.4-fold\textsuperscript{29}. This observation is important when considering the dramatic responses of melanoma patients to immune checkpoint inhibitor drugs and other immunotherapies, discussed below.

1.2.4. Signalling pathways altered in melanoma

Several intersecting cellular signalling pathways are involved in melanoma initiation, progression and maintenance\textsuperscript{3}. The mitogen-activated protein kinase (MAPK) signalling pathway, also known as the Ras/Raf/MEK/ERK pathway, is the most commonly altered signalling pathway in melanoma cells. Increased proliferation results from the abnormal activation of this pathway, most commonly due to mutually-exclusive mutations in either NRAS (mutated in approximately 15% of melanomas)\textsuperscript{30} or BRAF (mutated in approximately 42% of melanomas, most frequently with the V600E or V600K mutations\textsuperscript{31}). Similar
frequencies of *BRAF* mutations are found in benign naevi, primary melanoma tumour and metastatic melanomas. As most naevi do not proliferate indefinitely, the progression to malignancy probably requires further mutations, perhaps disrupting oncogene-induced cell senescence.

To illustrate the complex cross-talk involved in melanoma progression, a subset of the molecular pathways involved are described below. The MAPK pathway intersects with MITF signalling through ERK, and with the PI3K (phosphatidylinositol 3-kinase) pathway at multiple points, most importantly through NRAS. MITF, the master regulator of melanocytes, promotes cell survival through regulating anti-apoptotic and stress-attenuating genes. The PI3K (PI3K/PIP3/AKT/mTOR) pathway regulates cell proliferation, survival, growth and motility, as well as T-cell activation, and is inhibited by lipid phosphatase PTEN, and receptors PD-1 and CTLA-4. The tumour suppressor gene encoding PTEN is frequently mutated in non-familial melanoma, with loss of PTEN function associated with genetic instability. The PI3K pathway activates anti-apoptotic BCL2 in the mitochondria, and is involved in T-cell activation. Both the MAPK and PI3K pathways are stimulated by receptor tyrosine kinases, such as the gene *KIT*, which is frequently mutated or amplified in melanoma.

The *CDKN2A* gene, frequently mutated in familial melanoma, encodes INK4A (p16) and ARF on an alternative reading frame. MITF activates INK4A expression. INK4A suppresses aberrant cellular proliferation in cells with damaged DNA or activated oncogenes. Tumour suppressor ARF allows p53 to accumulate by sequestering the p53 regulator MDM2. If not sequestered by ARF, MDM2 causes the destruction of p53 through ubiquination.

WNT proteins are secreted proteins that regulate cell proliferation and migration during embryonic development. In the context of melanoma, WNT protein binding to WNT receptors leads to dissociation of β-catenin (encoded by *CTNNB1*) from transmembrane E-cadherin (encoded by *CDH1*), causing β-catenin to localise to the nucleus and up-regulate the genes *MITF* and *CCND1* (encodes Cyclin D1), leading to increased cellular proliferation and survival. β-catenin signalling is also stimulated by the decrease of E-cadherin expression. The shift from the radial to the vertical growth phase, outlined in section 1.1.1, is associated with decreased E-cadherin expression, and increased expression of N-cadherin (encoded by *CHD2*) expression. This permits invasion and subsequent metastasis by allowing melanoma
cells to interact with other N-cadherin expressing cells in the dermis and vascular endothelium. Similarly, transition to the invasive vertical growth phase is associated with increased αVβ3 integrin expression, which alters the cytoskeleton to increase cell motility, increases expression of anti-apoptotic BCL-2 and the collagen degrading MMP2.

Some of the same complex signalling pathways appear to operate in the stromal cells of malignant melanomas as in the tumour cells. One example is interferon-α and interferon-γ signalling stimulating anti-tumour immune responses via the JAK/STAT pathway. An anti-tumour immune response follows CDK4+ T-cell recognition of peptides bound to cell-surface major histocompatibility complex II (MHC II), the expression of which is induced by interferon gamma via the JAK/STAT pathway. Melanomas may potentially escape immune responses by disrupting this pathway.

To highlight the interconnected nature of the signalling pathways described in this section in a simplified way, these are summarised graphically in Figure 1-1 below. While these signalling pathways are irrefutably involved in melanoma progression, in order to find novel targeted therapies for melanoma it may be necessary to extend the search outside of these canonical melanoma pathways.
Figure 1-1 A highly simplified overview of major pathways involved in melanoma. Nodes represent genes (or groups of genes/their encoded proteins), with arrows pointing to the genes/proteins that they regulate. These interactions represent any form of regulation at DNA, RNA or protein levels, which may be facilitated via intermediaries. Although this Figure represents an extreme oversimplification of melanoma-associated molecular pathways, it is shown to emphasise the general interconnectivity of both protein-based and RNA expression-based pathways involved in melanoma progression.

1.2.4.1. Genomic insights into melanoma biology

The well-known melanoma pathways described above are frequently disrupted in tumours, and therefore well catalogued. Less frequently mutated pathways have been detected during comprehensive genomic studies, which have highlighted the biological heterogeneity of melanoma\textsuperscript{47,48}. Larger sequencing studies provide the depth required to detect the more uncommon melanoma genes, such as those described by Hodis et al.\textsuperscript{48} (\textit{PP6C, RAC1, SNX31, TACC1, STK19, ARID2}). Large genomic studies reveal features of tumour biology such as the immense variation in the total number of mutations between different melanoma
genomes. A large TCGA (The Cancer Genome Atlas) study of melanoma implicated genes previously unidentified as having a role in melanoma, such as DDX3X. Large genomic studies have also provided confirmation of sunlight exposure as a melanoma risk factor: mutation hotspots in the melanoma genes identified by Hodis et al. were enriched C>T and G>T transitions, indicative of UVB/UVA mutagenesis. Another large genomic study linked the level of sun exposure to tumour mutation count, and found an enrichment of UV damage signature mutations in sun exposed tumours.

A melanoma may contain multiple mutations in multiple genes belonging to complementary melanoma pathways. Mutations may be conditional upon co-occurrence of other mutations, such as mutations that co-operate with the common BRAF V600E mutation to drive melanogenesis. A genomic study by Mann et al. found that genes frequently mutated in the background of a BRAF V600E mutation are connected to one another through molecular links, and that in some cases, disruption of an overall molecular pathway may be more important in cancer development than disruption of specific members of that pathway. In a related analysis, gene association analysis by Krauthammer et al. found frequent association between mutations in RAC1 and DCC, suggesting that certain mutations may cooperate to promote malignancy. Redundant and dependant oncogenic driver roles are revealed by such patterns of mutation. Analysis of mutation patterns can suggest putative drug targets for new therapies, for example, targeting one of a pair of genes that share synthetic lethality, such as the targeting of APE1, a gene encoding a DNA base excision repair protein, in PTEN deficient melanoma cells.

The TCGA genomic study discussed above also provided a broad genomic classification of melanoma. This study identified four genomic subtypes, BRAF, RAS, NF1 and Triple Wild-Type. An interesting insight was that the most copy-number changes and complex structural arrangements were observed in the Triple Wild-Type subtype. Clinical traits were also associated with these genomic sub-types; for example patient age was lower in the BRAF subtype, and higher in the NF1 sub-type (rank sum p = 0.008).
1.3. Melanoma data

1.3.1. The growth and complexity of melanoma data

A rapidly growing amount of molecular and clinical data is available to melanoma researchers. Genomic data in particular is rapidly accumulating in publically available, international repositories; for example the Gene Expression Omnibus (GEO)\(^5\) and ArrayExpress\(^2\) for gene expression data, as well as the NCBI’s Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) and Illumina’s BaseSpace data repository (https://basespace.illumina.com/home/index) for sequencing data. A survey of all data accumulated in GEO, including melanoma data series and samples, illustrates this rapid growth (Figure 1-2). Meaningful utilisation of rapidly accumulating genomic data represents a challenge for researchers.

![Cumulative number of GEO series and samples](image)

**Figure 1-2** Plotted are the cumulative number of all data series (left panel) and individual samples (right panel) deposited into the GEO database between the years 2001 and 2016 inclusive.

Access to such large collections of data have allowed studies of genes involved in melanoma progression to advance from focused investigations of candidate genes to studies on a whole genome scale. Next-generation sequencing (NGS) alone has unleashed a wave of data, from the first published complete melanoma genome in 2010\(^5\) to more recent studies containing whole-exome sequences of more than 100 tumours\(^4\).\(^5\). Large-scale projects such as Australia’s Melanoma Genome Project (http://www.melanoma.org.au/research/melanoma-
genome-project.html) and the melanoma data contained in the TCGA project (http://cancergenome.nih.gov) and many others have further increased the amount of melanoma genomic data available to researchers.

Although there is a vast amount of melanoma data available to researchers, this data exists in a diverse range of formats and locations. Even when limited to a single gene of interest, relevant data may still include information about mutational frequency, genome-wide association, chromosomal aberrations, clinical SNP associations, associations between RNA expression and patient metastasis, treatment response and survival, literature-mined associations, as well as the encoded protein’s structure, location, function and contribution to molecular pathways. Even within the single data type of tumour DNA sequencing, data is produced in multiple different formats when different processing methods are used to associate genes with melanoma initiation and progression. Until now, researchers have been unable to utilise these diverse forms of data in an integrated manner, allowing them to triangulate in on clinically important genes. Even though the advent of next generation sequencing has allowed simultaneous identification of mutation, copy number and expression, integration of this data is still a challenge.

A further challenge is that the high mutation rate of melanoma, likely due to UV induced DNA damage and defects in DNA repair, results in gene mutation data that is especially dense. Furthermore, sequencing studies suggest that malignant melanoma is relatively heterogeneous, with a range of driver mutations.

A massive amount of data is brought together within repositories such as Oncomine, Ingenuity Pathways Analysis, the Catalogue of Somatic Mutations in Cancer (COSMIC), and the Broad Institute’s Melanoma Genomics Portal. However, these resources do not yet allow all researchers to access the full potential of an integrated analysis of diverse melanoma data types, and in some cases these repositories represent data ‘collections' rather than meaningfully integrated databases (as described below). The challenge of data integration is the focus of Chapter 3 of this thesis.

1.3.2. Integration of separate data types mutually enhances the interpretation of each

Analyses of different data types permit different insights into cancer. For example, cancers can be divided into sub-types according to their mutations, or according to differences in gene expression between groups. Yet a single data type can have its interpretation enhanced
by integration with another data type. Somatic mutations in regulatory sequences alter gene expression, so that in addition to other factors, gene expression is a functional consequence of the entire set of mutations within a tumour. For example, a novel mutation in the non-protein coding 5'-untranslated regulatory regions of BRAF could be identified as an oncogenic activating mutation if its presence also correlated with an increase in BRAF mRNA levels. Predictive models can also benefit from the mutual enhancement of integrated data types. Gomez-Rueda et al found that prognostic models of cancer patient survival that integrated mRNA expression, miRNA expression, mutation and copy number data had greater predictive power, with a higher concordance between predictions and actual outcomes in the test data set, than models created using only one data type60.

Certain biological relationships may only be detectable using integrated data. Two genes may appear unrelated when examining only mutation data, with their link only becoming apparent with inclusion of expression data, such as the presence of a mutation in one gene correlating with high expression of the other gene. Similarly, the study of Muniategui et al improved the prediction of miRNA targets by integrating mRNA and miRNA expression and sequence data61. Integrated models were able to predict a greater number of experimentally validated miRNA targets.

Furthermore, one data type can potentially be used to validate another. For example, it would be reasonable to expect copy number to positively correlate with gene expression data. Ensuring data veracity may also be more easily achieved with integrated data, with cross-checking between different data cohorts or data types flagging outlying relationships, thereby reducing the perpetuation of errors in large data sets.

1.3.3. Potential clinical use of molecular pathway data from individual tumours

Multiple protein-encoding genes and non-coding RNAs operate in molecular pathways involved in tumour development. Methods such as gene set analysis (GSA)62 can be used to infer molecular pathway activity from tumour genomic data. Gene sets, which are frequently conserved across species, may consist of co-expressed genes that share a common upstream molecular pathway or transcription factor binding site in their promoters63. They can therefore act as biomarkers for upstream molecular pathway activity or transcription factor activity64,65, with activity levels being inferred from statistical summaries of these gene sets. Several molecular pathways associated with melanoma have been identified by GSA66,67, including putative functional changes caused by the mutation, DNA gain or loss, and/or
altered gene expression within a particular patient’s tumour. For example, in their analysis of gene expression in 6 cell-lines of advanced melanomas, Hoek et al. found down-regulation of the expression of 21 out of 22 genes in the interferon pathway, which is likely to undermine patient response to interferon therapy. GATHER\textsuperscript{68}, DAVID\textsuperscript{69}, GSEA\textsuperscript{70} and GeneSetDB (software from our own laboratory)\textsuperscript{71} are popular GSA tools that are publically available.

The diverse genetic drivers that can contribute to melanoma are much larger in range than those currently targeted by therapy. Nevertheless, the development of these drugs targeting even a small subset of melanoma drivers has been a major clinical advance. The development of new drugs against a wider range of targets is crucial for patients. Molecular pathway data is important in selecting new therapeutic targets, and can also assist with the strategic use of existing therapies. For example, the drug Vemurafenib targets the V600E and V600K mutations in \textit{BRAF}\textsuperscript{72} to decrease activity of the MAPK pathway. However, the majority of patients treated with Vemurafenib show relatively short term remission with relapse almost certainly due to the re-activation of the MAPK-related pathways, frequently due to mutations in \textit{NRAS} or \textit{PDGFRB}\textsuperscript{73}. Integration of molecular pathway data, at both the population and individual tumour levels, could aid understanding of melanoma and allow molecular stratification of patients to which tailored combinatorial therapies could be rationally assigned. Popular, publically available molecular pathway databases include KEGG\textsuperscript{74} (www.kegg.jp), WikiPathways\textsuperscript{75} (www.wikipathways.org), Reactome\textsuperscript{76} (www.reactome.org), Comparative Toxicogenomics Database\textsuperscript{77} (http://ctdbase.org/) and Ingenuity Pathways Analysis\textsuperscript{53}.

\subsection{Integration of pathway data with cancer registry data}
Cancer patient registries are used to measure cancer incidence and mortality, track patterns of treatment, and construct risk models for predicting patient outcomes\textsuperscript{78}. Clinical data routinely collected by patient registries can be used to make broad epidemiological observations, such as associations between melanoma and prostate cancer incidence\textsuperscript{79}, or between mutation status and treatment outcome. If clinical data is annotated further with pathway data, then new insights may be possible – for example, treatment outcome may be associated with not only the mutational status of one gene, but with the genes belonging to a certain pathway. At the individual patient level, additional annotation could for example allow informed usage of off-label drug treatment.
1.4. Introduction to therapy for melanoma

Therapy for melanoma can be broadly divided into surgery, chemotherapy, immunotherapy and radiotherapy. Chemotherapy can be further divided into traditional, non-targeted cytotoxic chemotherapy, targeted therapies directed towards tumour cells and targeted therapies directed to enhance an anti-melanoma immune response. Targeted therapy (which includes both targeted chemotherapy and immunotherapy) is a core focus of this thesis and is described in the section below. Traditional chemotherapy interferes with the proliferation or survival of cancer cells, but usually also affects all rapidly dividing cells to some extent. Such non-targeted, cytotoxic chemotherapies are still used to treat melanoma, (e.g. the DNA alkylating agent Dacarbazine) but are increasingly being used in conjunction with, or replaced by targeted therapy.

It is standard treatment to surgically remove the primary tumour, and many can be successfully removed. However, metastatic melanoma, which develops from the primary tumour, poses the greater threat to patient survival, and is therefore the focus of therapy following surgery. If no metastasis has been previously detected, then biopsy may be taken from ‘sentinel’ lymph nodes that receive drainage directly from the primary tumour site. As metastases travel through the lymph system, their presence in a ‘sentinel’ lymph node indicates that other metastases may be present elsewhere in the patient; though the utility of conducting sentinel node biopsies in all cases has been questioned. Metastases, if operable, will be removed by surgery. The possibility of undetected metastases, makes adjuvant systemic therapy desirable in addition to surgery in many instances. Commonly used adjuvant therapy are immunotherapy regimens including Interleukin-2, Interferon α-2b (Intron A) and Dacarbazine.

Local radiotherapy may also be used on inoperable, high-risk metastases and in a palliative setting, particularly in brain and bone metastases. As radiotherapy causes inflammatory effects, its combination with newer targeted therapies, such as Ipilimumab, may potentially yield better therapeutic outcomes.

A more recently developed immunotherapy is Adoptive T-cell therapy. Adoptive T-cell therapy is the *ex vivo* expansion of tumour-infiltrating lymphocytes – autologous or allogenic T or NK cells with affinity to tumour antigens - for infusion into patients. Trials of this
approach in melanoma have had encouraging results\(^8^6\). The high mutational load of melanoma can be exploited by identifying tumour-specific, mutation-encoded antigens recognised by tumour-infiltrating lymphocytes\(^8^7\). An alternative to direct infusion with T-cells, is to attract endogenous, polyclonal T-cells to the tumour by use of MCSP/CD3 bispecific T-cell engager (BiTE) antibodies that bind to T-cells’ CD3 co-receptor and the melanoma-associated chondroitin sulphate proteoglycan (MCSP)\(^8^8\).

1.4.1. Targeted therapy of melanoma

1.4.1.1 Rationale and concepts of targeted therapy
Cancer contains a diverse range of alterations within a network of numerous interacting pathways. Melanoma in particular is known for its high mutation rate and associated high number of genomic alterations\(^8^9\). However, many of these alterations can be termed ‘passenger mutations’ as they do not confer any selective advantage to a cancer cell, implying that these mutations may have arisen incidentally and that drug targeting of these mutations may not affect tumour progression. This is in contrast to ‘driver mutations’ that confer selective advantages to tumour cells such as increased proliferation or survival, or an increased ability to metastasise\(^9^0\). Ideally, for the development of new targeted therapies, the critical, functional nodes within molecular networks commonly hijacked by melanoma cells to support their growth will be identified\(^9^1\). In addition to targeting key nodes, new drugs should ideally be targeted to cancer cells but have minimal effect on normal cells – that is, targeted drugs should constrain cancer cells by inducing apoptosis, necrosis, senescence, or differentiation, while ideally having limited effect on normal cells\(^9^1\).

Described below are four concepts distilled from the literature to aid identification of therapeutic targets within molecular pathways:

i) Oncogene addiction

ii) Tumour suppressor gene hypersensitivity

iii) Non-oncogene addiction

iv) Lineage-survival oncogenes.

Oncogene addiction is the concept that cancer cells’ growth and survival depends upon a certain over-active pathway more than normal cells, and therefore depends on the over-active oncogene(s) driving that pathways\(^3^,^9^2^,^9^3\). The rationale for the efficacy of many targeted
therapies is that cancer cells will be more sensitive than normal cells to the inhibition of these onco-genes by drug binding. For example, melanoma cells containing the activating V600E mutation in the \textit{BRAF} onco-gene are dependent on this mutated BRAF to drive proliferation via the MAPK pathway. Oncogene addiction is exploited when such melanomas are targeted with BRAF inhibitors such as Vemurafenib and Dabrafenib. These drugs are specific to the protein encoded by a \textit{BRAF} variant gene.

An inverse concept to oncogene addiction can be termed “tumour suppressor gene hypersensitivity”\textsuperscript{93}, when cancer cells are hypersensitive to re-activation of their inactivated tumour suppressor genes. However, it is typically more difficult to pharmacologically activate a protein target, than inhibit it\textsuperscript{91}. Receptors are the exception to this. Therefore, pharmacological inactivation of inhibitors of tumour suppressor genes may be a viable therapeutic approach provided that within the tumour, the under-active tumour suppressor gene does not contain an inactivating mutation and the molecular pathway in which the protein it encodes operates still remains functional. For example, the presence of a mutation in the \textit{TP53} gene is a strong negative predictor of therapeutic response to MDM2 inhibition\textsuperscript{94}.

Solimini et al. described the concept of ‘non-oncogene addiction’\textsuperscript{95}. Identifying and targeting key pathway nodes or ‘driver mutations’ is the basis of targeted therapy, however, such drivers may be rare, present in only a subset of cells, and may contribute fractionally to the cancer phenotype in concert with other driver mutations. Targeting such drivers alone may not be sufficient to halt cancer growth or may do so only in a subset of cells. According to Luo et al.\textsuperscript{91} “Large-scale sequencing of multiple cancers has so far failed to identify new, high-frequency mutation targets in addition to those previously identified. Rather, these studies found that every tumour harbours a complex combination of low-frequency mutations thought to drive the cancer phenotype”. That implies that in some tumours a combination of multiple genes and pathways may collectively contribute to the cancer phenotype. Some of these genes or pathways may be essential to a cancer cell’s survival, but not required to the same degree for the viability of normal cells\textsuperscript{91}. These combinations of genes might not even contain oncogenic mutations themselves, but may simply have abnormally increased expression. Such combinations of ‘non-oncogenes’ may interact in such a way that targeting combinations of them results in synthetic lethality within tumours, but not in normal cells. The high mutational load in melanoma may mean a high number of synthetically lethal combinations. Thus, these non-oncogenes represent a special case of potential drug targets.
A further similar concept is that of ‘lineage-survival oncogenes’, in which cancer cell survival is dependent on the deregulation of genes key to the development of the normal cells from which the cancer arises\textsuperscript{96}. The lineage-restricted, melanocyte master regulator MITF can become a melanoma oncogene – that is, MITF is only an oncogene in melanoma and in general not other cancer types because it has a central role in the regulation of melanocytes, but less so cells of other lineages. 15% of melanomas have MITF amplification, and the majority of melanomas are thought to be dependent on MITF for survival\textsuperscript{97}. A small molecule MITF inhibitor has been described in the literature\textsuperscript{97}.

1.4.1.1.2 Expansion of melanoma therapy options by targeted therapy

Before 2011, there was no systemic therapy in common clinical use that conclusively improved survival in metastatic melanoma patients. An objective response rate of as low as 15% response could be achieved by the international standard, cytotoxic chemotherapy Dacarbazine, so due to the lack of other treatment options, enrolment in a clinical trial was considered a first-line option\textsuperscript{5}. At that time, Interferon-\(\alpha\) was the only effective adjuvant therapy for patients who had previously had thick primary tumours resected\textsuperscript{98}. Licensing of new targeted therapies for metastatic melanoma began in 2011 with immunotherapies targeting CTLA4 and PD-1, and continued in 2012 with chemotherapies targeting the MAP kinase pathway\textsuperscript{5}. As described in section 1.2.4 above, the MAPK pathway promotes proliferation, so therapies inhibiting the activity of this pathway can restrict tumour growth. CTLA4 and PD-1 facilitate negative feedback on the T-cell mediated immune response. Blocking this negative feedback can enable a sustained immune attack on tumour cells. Combinations with other immunotherapeutic strategies have been suggested to enhance\textsuperscript{99} or complement\textsuperscript{100} PD-1 inhibition.

Table 1-1 shows examples of targeted drugs approved by the FDA for use in the treatment of melanoma. Note that Dacarbazine is listed as an example of a drug toxic to dividing cells, in contrast to those that target pathways aberrant in melanoma. These targeted therapies can be divided into sub-groups of targeted chemotherapy (Dabrafenib, Trametinib, Vemurafenib) and immunotherapy (Aldesleukin, Interferon \(\alpha\)-2b, Pembrolizumab, Nivolumab, Ipilimumab).
In addition to drugs in Table 1-1, phase II trials in melanoma have been successfully completed for three re-purposed leukaemia drugs, Imatinib, Dasatinib and Nilotinib, which target the receptor c-Kit \(^ {101}\). Similar re-purposing attempts have been made in phase II melanoma trials of the mTOR inhibitor Temsirolimus\(^ {102}\), and RAS kinase inhibitor Tipifarnib\(^ {103}\). Trials are currently being conducted with the breast cancer drug Palbociclib, a CDK4/CDK6 inhibitor, for use in treating melanoma (ClinicalTrials.gov Identifiers: NCT02202200 and NCT02065063) as well as other drugs.

While drugs may be approved by the FDA, they may not be licensed in New Zealand. Even when licenced for use, the cost of a drug may be prohibitive for many New Zealand patients, introducing potential health service inequities. Development of new therapies may reduce the cost burden to patients.

A challenge for any targeted therapy is that an expanding subset of the cancer can adapt and develop resistance to treatment, which is the topic of the following section.

### 1.4.2. Drug resistance

#### 1.4.2.1. Theories of drug resistance

**1.4.2.1.1 Clonal evolution**

Many tumours contain a heterogeneous population of cells, each containing different sets of genetic alterations which in some cancers appear to allow evolution by natural selection. Cell populations within tumours evolve by natural selection. For example, a cell containing a mutation that increases proliferation may divide into more daughter cells per unit of time than
a similar cell without that mutation. Similarly, a cell with an anti-apoptotic mutation may have a selective advantage that would also result in a greater number of progeny cells also containing that mutation. Therefore, the clonal evolution theory predicts that cancer can arise as a result of the iterative accumulation of such advantageous genetic alterations\textsuperscript{104}, while disadvantageous mutations are theoretically ‘selected out’, and neutral mutations remain static unless linked to an advantageous or disadvantageous mutation or lost by a chance event.

Cancer therapy of any kind exerts a selective pressure upon cancer cells. The clonal evolution theory suggests that genetically diverse tumours may include sub-populations of cells not responsive to the therapy that will expand. For example, BRAF inhibitors such as Vemurafenib may limit the growth of tumour cells that harbour the V600E BRAF mutation, but not those with other alterations in the PI3K pathway, such as loss of function of the PTEN\textsuperscript{105} protein. The tumours of approximately 15% of melanoma patients treated with BRAF inhibitors do not go into regression in response to the inhibition, due to this pre-existing ‘intrinsic’ or ‘primary’ resistance. That is, the tumour cell population includes Vemurafenib-resistant sub-populations, such as those with inactive PTEN\textsuperscript{105}. Another group of patients may initially respond to therapy but then develop ‘acquired’ or ‘secondary’ resistance. For BRAF inhibitors, acquired resistance is most frequently caused by reactivation of the MAPK pathway – not just by alterations in pathway components downstream of BRAF (such as activating MEK1/2 mutations), but also by alterations in upstream components (e.g. NRAS activating mutations) or within BRAF itself\textsuperscript{105}. Additionally, there is evidence from tumour cell line studies that Vemurafenib, while inhibiting the MAPK pathway in V600E BRAF mutant cells, may actually induce the pathway in cells with wild type BRAF\textsuperscript{106,107}.

1.4.2.1.2 Cancer stem cells
Drug resistance may also arise from the presence of cancer stem cells, which may be either an earlier, quiescent stage of pre-malignant cells giving rise to metastatic malignant melanoma, or a sub-set of the cells in a malignant melanoma tumour\textsuperscript{108}. That is, the cancer stem cells may be a less advanced, non-malignant cell population that repetitively gives rise to malignancies, or a sub-set within the malignancy that can replenish space vacated by dying cancer cells with new cells. In either scenario, therapies that target rapidly dividing cells or advanced tumour signalling pathways, may not kill quiescent cancer stem cells that have less proliferative phenotypes.
1.4.2.1.3 Reduced spontaneous apoptosis
Melanoma cells have lower levels of spontaneous apoptosis relative to other cancer cells, which may contribute to their high resistance to chemotherapies, the majority of which work directly or indirectly by inducing apoptosis. That is, the entire population of melanoma cells may be inherently more resistant to chemotherapies than other cancers, even prior to chemotherapy exerting selection pressure in favour of more resistant sub-populations.

1.4.2.2 Strategies for overcoming drug resistance

1.4.2.2.1 Combinatorial therapy
The identification of dominant oncogenes within biopsied tumours, and subsequent treatment by single targeted therapies may only treat a sub-population of cells within the heterogeneous cancer cell population. In order to treat multiple sub-populations, a combination of different therapies that target different pathways can be used. It has been recently suggested that simultaneous targeting of MAPK pathway and Notch signalling may be effective in reducing melanoma growth. Similarly, simultaneously targeting multiple nodes within one pathway may ensure that this pathway’s activity is still modulated even when cells develop resistance against therapy targeting one node. For example, the treatment of metastatic melanoma with a combination of Dabrafenib and Trametinib, which target different nodes within the MAPK pathway, has been shown to be more effective than therapy with just one of these drugs alone. However, while the sequencing of the drugs seems to be important in many cases, combinatorial therapy may still not be completely successful, as resistance can still develop.

1.4.2.2.2 Immune-mediated bystander effects
Even when multiple, orthogonal targeted therapies are concurrently applied, some cells within the heterogeneous cancer cell population will not have any of the targeted alterations and therefore will not be targeted. However, in some tumours apoptosis of the targeted cells can trigger immune-mediated bystander effects that affect neighbouring cells not targeted by the therapy. This immune activation may be triggered by damage-associated molecules released from damaged or dying tumour cells. This is in addition to any immune response triggered by tumour antigens. In melanoma patients for example, BRAF inhibitor treatment may lead to these immune stimulatory bystander effects. Combining BRAF inhibitor treatment with MEK inhibition may further enhance this immune activation, possibly by inhibiting the expression of PD-L1, a ligand of PD-1. Additionally, the immune system can be directly targeted. Given the evidence that certain oncogenes modulate expression of
immune regulatory genes, and thus interfere with the local immune environment, it has been proposed that oncogene targeting drugs be combined with immune therapy\textsuperscript{114}. Clinical trials in melanoma are underway combining Dabrafenib and Trametinib with Iplimunumab (ClinicalTrials.gov Identifiers NCT01940809, NCT01767454) or Pembrolizumab (NCT02130466) or Nivolumab (NCT02357732) or both Iplimunumab and Nivolumab (NCT02224781). Two clinical trials (NCT01400451, NCT02095652) simultaneously administering Vemurafenib and Iplimunumab were terminated due to dose-limiting toxicities, although a Phase II trial of Vemurafenib followed by Iplimunumab is on-going as of May 2015 (NCT01673854). Monitoring response to therapy through biomarker\textsuperscript{115} or biopsy feedback may in theory be particularly useful during this type of sequential therapy, in order to determine the best time to switch between drugs.

1.4.2.2.3 Discontinuous dosing
Another drug treatment strategy is discontinuous dosing. Drug resistant cells may have a ‘fitness deficit’ in the absence of the drug\textsuperscript{116}. For example, to produce proteins conferring drug resistance, a cell may incur a metabolic cost in resources otherwise used for proliferation. Therefore, non-resistant cells that do not incur that metabolic cost have a proliferative advantage over a period when the drug is withdrawn. A discontinuous dosing strategy can delay the onset of drug-resistant disease by allowing periods without drugs during which drug-resistant cells have their growth restricted by their being out-competed by their non-resistant neighbouring cells. These non-resistant cells are then killed when drug dosing resumes. The anti-tumour effect of withdrawing drug therapy may be due to the cells being not just resistant to the drug, but drug-dependant. That is, the growth of some resistant cells may be dependent upon continued exposure to a drug, this has been seen in both melanoma tumours and cell lines in response to Vemurafenib\textsuperscript{117}. This melanoma drug addiction was observed to be more robust in cell lines selected for resistance to both BRAF and MEK1 inhibition, suggesting that acute withdrawal of multiple drugs may be more effective than withdrawal of a single targeted therapy drug\textsuperscript{110}. As a further example of discontinuous dosing, outside of melanoma, withdrawal of anti-androgen drugs is an established clinical practice in the treatment of prostate cancer\textsuperscript{118}.

1.4.2.2.4 Targeting non-tumour cells
An alternative to directly targeting cancer cells is targeting non-tumour cells in the tumour microenvironment. Genetically and epigenetically normal non-tumour cells adjacent to or within a tumour, such as stromal fibroblasts or vascular endothelial cells, are required to
support tumour growth\textsuperscript{119}. Targeting these non-tumour cells can limit tumour growth. As these non-tumour cells do not share the genetic instability of tumour cells, they are less likely to develop drug resistance. Furthermore, targeting non-tumour cells can limit tumour growth regardless of cancer cell genotype, so such a strategy avoids the complication of tumour heterogeneity and its associated tendency for the development of drug resistance. Potential therapeutic strategies of this type for treating melanoma include inhibition of PAR-1\textsuperscript{119}, and disruption of cancer assisting fibroblasts\textsuperscript{120}.

1.4.2.2.5 Drugging wild-type proteins
Not all drivers of cancer result from disruptive mutations in gene coding regions. Cancer can be driven not only by structurally altered cancer-specific proteins, but also by disruption of the normal regulation of wild-type proteins. These proteins are found in non-tumour cells, but their level of activity or abundance has been altered in cancer. This may be due to mutations in upstream regulators, such as the promoter or upstream transcription factors, or disruption of downstream degradation systems. Drugs targeting altered proteins are specific to the alteration they target, however drugs targeting wild-type proteins affect that protein regardless. The drugging of wild-type proteins is addressed again later in the thesis (Chapter 4, Section 4.4.1).

1.4.2.3. Combination of targeted therapy and immune checkpoint inhibition
It is an attractive approach to combine targeted immunotherapy with some form of non-targeted, anti-tumour therapy, provided the non-targeted therapy does not indiscriminately kill cells of the immune system. Previously, anti-tumour cell targeted therapies, such as BRAF inhibition, and immunotherapies, such as inhibition of CTLA-4 or PD-1, were thought to have distinct mechanisms of action. However, responses to BRAF inhibition are now thought to be partly facilitated by the host immune system, with indications that resistance to BRAF inhibition is associated with up-regulation of PD-1\textsuperscript{121}. Furthermore, BRAF inhibition results in a range of immune responses, such as an increase in T-cells within the tumour\textsuperscript{122}. It seems likely that the near future of melanoma therapy lies in a strategic combination of immunotherapy with targeted therapy.

1.4.2.3.1 Therapy sequence
Immunotherapy can induce a durable, long-term but sometimes slow response. In contrast, targeted therapy can induce a faster, short-term response that is halted by the development of
resistance. Tumours that develop resistance to BRAF inhibition can rapidly progress in a
time-frame that may be too short for any subsequent immunotherapy to induce a response\textsuperscript{123},
or may have entered an immune-evasive state that would undermine immunotherapy\textsuperscript{124}. Therefore, it has been argued that immunotherapy should be first in the sequence of
treatment, with targeted therapy like BRAF inhibition being applied to patients who fail immunotherapy\textsuperscript{124}. The optimal sequence of therapies is currently under on-going study and further prospective, randomised studies are required\textsuperscript{125}.

1.5. Drug development: target and lead identification

1.5.1. Targets and leads as starting points

Lead compounds are chemicals with activity that may be refined to form a therapeutically
useful drug. This activity is affected through the interaction of the compound with its target.
Drug discovery can start from the identification of a lead compound (for which a target, and
thus mechanism of action, is sought) or identification of a target (to which a lead compound
is sought). There is a wide scope for definition of a target. Though targets are most
commonly proteins, they can also be, for example, a molecular structure within a biological
system, an interaction between components, or a physicochemical mechanism\textsuperscript{126}. The
selection of a target from this wide range of possibilities represents one of the most important
and consequential decisions in drug development\textsuperscript{127}. This point has been a core focus of my
thesis work.

1.5.1.1. Academic target identification

Academic research can yield insights that uncover potential new drug targets. An academic
investigator engaged in exploratory, basic research can investigate areas of scientific interest
not practicable for pharmaceutical companies concerned with profitability\textsuperscript{128}. For an
academic group, a lead compound may provide a valuable reagent to study a target gene even
if not of clinical relevance. In contrast, a pharmaceutical company would not pursue a target
or a lead compound that was unlikely to result in a clinically valuable drug. Traditionally,
lead compounds developed in academia are passed onto pharmaceutical companies that have
the resources required for the long process of developing these compounds into drugs\textsuperscript{128}.
Sometimes this relationship is explicitly organised– for example, the Manchester Institute
Drug Discovery Unit has defined their role as “to ‘derisk’ novel, emerging biology through
the discovery and development of pharmacologically active, drug-like small molecules that
would serve both as tools to deliver robust target validation, perhaps confirming data obtained from genetic methods, and to deliver potential new therapeutic agents for preclinical and ultimately clinical evaluation”\textsuperscript{129}.

1.5.1.2. Methods for identifying targets

The identification of potential new drug targets in academia can be incidental. However, if an academic lab were to actively set out with the intention to find novel drug targets, there are many possible approaches. Observational phenotypic screens can find compounds that perturb a disease model, and these active compounds can have their target(s) deconvoluted\textsuperscript{130}. Genomics-based approaches can include broad screens of gene knock-down (e.g. by RNAi\textsuperscript{130} or by CRISPR-Cas9\textsuperscript{131}) in disease models to identify genes where knock-down results in a desired phenotype (e.g. reduced proliferation in a cancer cell line), or compare gene expression between disease and control groups. There are also data and literature-based approaches that set out to identify (i) clinically relevant, but not yet drugged (“undrugged”) members of a commonly drugged protein family, such as the G-protein coupled receptors\textsuperscript{132}, (ii) undrugged nodes within relevant molecular pathways, or (iii) critical bridging nodes in molecular networks\textsuperscript{133}. Since there is no well established, best standard approach, academic groups generally use methods most familiar or accessible to them. For this reason, the starting point for this PhD project was a data-based approach.

Data-based approaches usually benefit from integration of multiple types of data generated by a variety of laboratory techniques\textsuperscript{134}. Of course bioinformatic inferences and hypotheses ultimately need to be confirmed in the wet lab, or by validation of the target in a disease model. For a protein target implicated in cancer, an example would be inhibition of xenograph tumour growth by RNAi-mediated down-regulation of the protein’s gene’s expression. Once a target is validated, then an assay can be developed to facilitate screening of compounds that bind to the target.

1.5.2. High throughput screening to identify lead compounds

High-throughput screening (HTS) is a broad term referring to the combined use of automation, assay miniaturisation and data processing to screen large collections of compounds for activity against biological targets. While specific methods vary, the concept of HTS itself is well established, having been used for lead discovery by pharmaceutical companies for over 20 years, and more recently in academic research\textsuperscript{135}. 
In the early 1990s, the widespread use of combinatorial chemistry and the commercial availability of compound collections increased the size and diversity of compounds available for use in screening\textsuperscript{136}. Utilising the potential of these larger compound libraries, while avoiding a corresponding increase in screening cost, required the miniaturisation of screening assays. Automation, through integration of time-saving robotic processing, has moved HTS formats from the 96-well microtiter plates of the 1990s to 384-well and 1536-well plates, with a reduction of reaction volumes from 100-200µL to 25-100µL and 2.5-10µL respectively\textsuperscript{135}. Costs can also be reduced by multiplexing, that is, pooling multiple compounds in single wells, without having significant impact upon the success of a screen\textsuperscript{137}. This is especially appropriate if one or more assay reagents are in restricted supply.

A high-throughput screen can be run against a full compound library or a smaller sub-set. A sub-set may be selected by ligand- or target-based computational modelling of binding, or by a trait of the target, for example, screening a kinase target with novel kinase inhibitors. Alternatively, an initial sub-set with the greatest chemical diversity may be selected to represent the variety of compounds within the library\textsuperscript{138}. Conversely, instead of selecting for a sub-set, filtering may be applied to a compound library to remove highly similar compounds or undesirable compounds such as ‘frequent hitter’ promiscuous compounds, also known as pan-assay interference compounds (aptly abbreviated PAINS)\textsuperscript{139}. Sub-sets may especially be used when there is a restricted supply of one assay reagent (e.g. protein), or during ‘high-content screens’ that generate high quality data (relevant, for example, to the desired mechanism of compound action) but require costly down-stream analysis\textsuperscript{140}.

The results or ‘hits’ from a screen of one sub-set can be used to build statistical models to select the next sub-set for a later screen. Thus, screens can be run iteratively, adding data from each screen to build cumulative models that can select compounds for subsequent screens\textsuperscript{138}. Screening cycles then cease upon obtaining a sufficient number or quality of hits. However, for some sufficiently-automated, large industrial HTS facilities, it is now easier and more desirable to screen a full library than iteratively run sub-sets\textsuperscript{141}.

Trends in HTS during the early 2000s saw a shift in focus from improving time and cost to improving quality\textsuperscript{135}. An HTS assay must not only consider metrics of quality such as the rate of false positive and negatives, signal to background ratio, and Z-factor (a commonly used metric of signal dynamic range and measurement variability)\textsuperscript{142}, but also the design and content of the assay and its subsequent follow-up. Carefully considered assay design is
required to ensure the assay results are biologically relevant. The ‘hit’ compounds that an assay identifies as affecting its output signal, must be evaluated by an independent, secondary, orthogonal screen or screens. Secondary screens may include counter-screens and selectivity screens. Counter-screens aim to eliminate artefacts, such as false positives, due to the technology used in the assay. Selectivity screens aim to differentiate between compounds that interact only with the target of interest, and compounds that interact with the target of interest and other, possibly related, targets. Once a lead is identified, it can be optimised either for development into a research tool or a pre-clinical drug candidate.

1.5.3. Biomarkers

Biomarkers are biological indicators that provide diagnostic or prognostic information to guide therapeutic decisions. For example, positive PD-L1 immunohistochemistry staining has been used with varying success as a biomarker for treatment of tumours with PD-1 inhibitors. Biomarkers may predict melanoma progression through stages of increased angiogenesis (e.g. VEGF-A, FGF-2, MMP) or invasion and metastasis (e.g. S100B, LDH, CRP, ICAM-1, MCAM).

Ideally, clinical trial patients may be selected for inclusion and stratified on the basis of companion biomarkers specific to the drug target. Furthermore, trials of a pre-clinical drug candidate can identify biomarkers that can be used later to gauge response to the drug. Qualification and validation of biomarkers at the early stages of drug development can reduce cost during later clinical trials.

1.6. Thesis project objective

Recent years have seen an expansion of targeted melanoma therapies. However, the challenges of tumour heterogeneity and drug resistance need to be met with a diverse range of drug combinations. Despite the promise of immune checkpoint inhibitors, there is still an urgent need to find new melanoma cell drug targets and to extend the drugged cancer genome.

The amount of available molecular, genomic and clinical information about melanoma is rapidly growing. It would be helpful to bring this dispersed and disparate information into an integrated, manageable resource, so that new insights can be gained to inform rationale experimental design. One use of such a resource could be to identify putative new drug
targets. This was the aim of the thesis, and the flow-chart below in Figure 1-3 summarises the progression of this project, which is described in subsequent chapters.

Figure 1-3 Project overview. Flow-chart summarising the progression of the project and defining the focus of each Chapter.

Chapter 3 describes the construction of a database following the collection of diverse data types and their subsequent integration. Chapter 4 describes example database applications, including the identification of a short list of putative drug targets, from which a single drug target was selected for further investigation. Two assays were developed to measure the activity of that target: a cell-based luciferase reporter gene assay described in Chapter 5, and an AlphaScreen assay described in Chapter 6. Finally, Chapter 7 describes the application of these assays for high-throughput screening.
Chapter 2 Materials and Methods

2.1. Introduction

Described here are the standard methods drawn upon for general use during this project. These general methods were frequently modified for specific uses. Descriptions of these modifications, and the specific purposes for which they were developed, are given in subsequent Chapters. The materials used in this project are given first, followed by the general methods. The methods are divided into sections describing those used during bioinformatics, molecular biology, cell culture and YB-1 binding nucleic acid assays.

2.2. Materials

2.2.1. Molecular biology

2.2.1.1. Buffers

Table 2-1 Composition of buffers used in *in vitro* protein-nucleic acid binding assays

<table>
<thead>
<tr>
<th>Buffer:</th>
<th>Tris-HCl pH 7.9 (mM)</th>
<th>HEPES KOH pH 7.9 (mM)</th>
<th>KCl (mM)</th>
<th>NaCl (mM)</th>
<th>DTT (mM)</th>
<th>EDTA (mM)</th>
<th>Glyc-erol (v/v)</th>
<th>MgCl2 (mM)</th>
<th>BSA (w/v)</th>
<th>Dextran (w/v)</th>
<th>Gelatin (w/v)</th>
<th>Tween-20 (v/v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>South-Western (SW)</td>
<td>4</td>
<td>12</td>
<td>100</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2</td>
<td>4</td>
<td>12</td>
<td>100</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2-KCl</td>
<td>4</td>
<td>12</td>
<td>-</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2-KCl-gly</td>
<td>4</td>
<td>12</td>
<td>-</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2-KCl-gly-EDTA</td>
<td>4</td>
<td>12</td>
<td>-</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>0.5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2-KCl-gly + 0.5% BSA</td>
<td>4</td>
<td>12</td>
<td>-</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>0.5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2-KCl-gly + 0.5% dextran</td>
<td>4</td>
<td>12</td>
<td>-</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>0.5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SW+MgCl2-KCl-gly + 0.5% gelatin</td>
<td>4</td>
<td>12</td>
<td>-</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>12%</td>
<td>2</td>
<td>-</td>
<td>0.5%</td>
<td>-</td>
<td>0.5%</td>
</tr>
<tr>
<td>Thermo RNA-Protein binding buffer</td>
<td>20</td>
<td>-</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>12%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.1%</td>
</tr>
<tr>
<td>TBS + 0.1% Tween</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>150</td>
<td>-</td>
<td>-</td>
<td>12%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.1%</td>
</tr>
</tbody>
</table>
### Table 2-2 HEPES buffers used in protein-nucleic acid binding assays

<table>
<thead>
<tr>
<th>Buffer:</th>
<th>HEPES pH 7.4</th>
<th>NaCl (v/v)</th>
<th>Casein (v/v)</th>
<th>Dextran-500 (v/v)</th>
<th>Triton X-100 (v/v)</th>
<th>Proclin-300 (v/v)</th>
<th>Gelatin (v/v)</th>
<th>BSA (v/v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImmunoAssay</td>
<td>25mM</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.05%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NaCl HEPES</td>
<td>25mM</td>
<td>0.5M</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.05%</td>
<td>0.5%</td>
<td></td>
</tr>
<tr>
<td>HiBlock</td>
<td>25mM</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.05%</td>
<td>0.5%</td>
<td>0.5%</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2-3 PBS buffers

<table>
<thead>
<tr>
<th>Buffer:</th>
<th>KCl (mM)</th>
<th>NaCl (mM)</th>
<th>Na₂HPO₄ (mM)</th>
<th>KH₂PO₄ (mM)</th>
<th>MgCl₂ (mM)</th>
<th>BSA (v/v)</th>
<th>Dextran (v/v)</th>
<th>Gelatin (v/v)</th>
<th>Tween -20 (v/v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS</td>
<td>2.7</td>
<td>137</td>
<td>10</td>
<td>1.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PBS + 0.2% BSA</td>
<td>2.7</td>
<td>137</td>
<td>10</td>
<td>1.8</td>
<td>2</td>
<td>0.2%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PBS + 0.1% BSA</td>
<td>2.7</td>
<td>137</td>
<td>10</td>
<td>1.8</td>
<td>2</td>
<td>0.1%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PBS + 0.1% BSA + Tween</td>
<td>2.7</td>
<td>137</td>
<td>10</td>
<td>1.8</td>
<td>2</td>
<td>0.1%</td>
<td>-</td>
<td>-</td>
<td>0.1%</td>
</tr>
<tr>
<td>PBS + 0.5% dextran</td>
<td>2.7</td>
<td>137</td>
<td>10</td>
<td>1.8</td>
<td>2</td>
<td>-</td>
<td>0.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PBS + 0.5% gelatin</td>
<td>2.7</td>
<td>137</td>
<td>10</td>
<td>1.8</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.5%</td>
<td>-</td>
</tr>
</tbody>
</table>

The sources of the reagents given within Tables 2-1 to 2-3 are stated in brackets here. Sources are given once during this Chapter, at the first instance of each reagent: Tris (Invitrogen), HEPES (N-2-hydroxyethylpiperazine-N-2-ethanesulfonic acid; Sigma-Aldrich), KCl (British Drug Houses, BDH, now a part of Merck group), NaCl (CalbioChem), DTT (Invitrogen), EDTA (Sigma-Aldrich), Glycerol (BDH), MgCl₂ (Sigma-Aldrich), BSA (Sigma-Aldrich), Dextran (Sigma-Aldrich), Gelatin (Sigma-Aldrich), Tween-20 (Sigma-Aldrich), Casein (Sigma-Aldrich), Triton X-100 (BDH), Proclin-300 (Sigma-Aldrich), Na₂HPO₄ (Sigma-Aldrich), KH₂PO₄ (Sigma-Aldrich).
Buffers for Qiagen preps, all sourced from Qiagen:
Miniprep solution A  25 mM Tris HCl pH 8.0, 50 mM glucose, 10 mM EDTA
Miniprep solution B  0.2 M NaOH, 1% (w/v) SDS
Miniprep solution C  3 M KOAc, 11.5% (v/v) glacial acetic acid

Buffers used in Antibody purification:
20mM sodium phosphate (Sigma-Aldrich), pH 7.0
0.1M glycine (Merck), pH 2.7
1M Tris, pH 9.0

Buffers used in cellular fractionation:
Fractionation Buffer A  10 mM HEPES-KOH pH 7.9, 10 mM KCl, 1.5 mM MgCl₂, 0.5 mM DTT, 0.1% (v/v) NP-40 (Sigma-Aldrich)
Fractionation Buffer C  20 mM HEPES-KOH pH 7.9, 25% glycerol, 420 mM NaCl, 1.5 mM MgCl₂, 0.5 mM DTT, 0.2 mM EDTA, 0.1% (v/v) NP-40

2.2.1.2. Electrophoresis reagents
6x DNA loading dye  30% (v/v) glycerol, 0.25% (w/v) bromophenol blue (Sigma-Aldrich), 0.25% (w/v) xylene cyanol (Merck)
GelRed Nucleic Acid stain  10,000x solution in water, Biotium Inc, Cat#41003
TAE buffer  40 mM Tris pH 8.0, 2 mM EDTA, 0.1% (v/v) glacial acetic acid (Merck)

Reagents used in Coomassie Blue SDS-PAGE protein separation:
Reducing buffer  125 mM Tris HCl, pH 6.8, 20% (v/v) glycerol, 15% (v/v) β-mercaptoethanol (Sigma-Aldrich), 4.1% (w/v) SDS, 1.8% (w/v) bromophenol blue
SDS running buffer  25 mM Tris base, 25 mM glycine, 0.1% (w/v) SDS
Acrylamide/Bis  30% (w/v) acrylamide (Merck), 0.8% (w/v) bisacrylamide (Merck)
10mL 10% SDS-PAGE gel 4.2mL water, 3.3mL Acrylamide/Bis, 2.5mL 0.4% (w/v) SDS in 1.5 M Tris HCl pH 8.8, 60μL Ammonium Persulfate (BioRad), 8μL tetramethylethlenediamine (Sigma-Aldrich)

3.4mL Stacking gel 2mL water, 0.83mL 0.4% (w/v) SDS in 0.5 M Tris HCl pH 6.8, 0.5mL Acrylamide/Bis, 40μL 10% Ammonium Persulfate, 3.3μL tetramethylethlenediamine

Coomassie stain 50% (v/v) ethanol, 7.5% (v/v) acetic acid, 0.06% (w/v) brilliant blue R-250 (Sigma-Aldrich)

Coomassie destain 25% (v/v) ethanol, 8% (w/v) acetic acid

2.2.1.3. Enzymes
DNA Polymerase I, Large (Klenow) Fragment (5 units/μL; New England BioLabs)

High Fidelity KAPA Taq DNA polymerase (1 unit/μL; Kapa Biosystems)

HindIII restriction endonuclease (20 units/μL; New England BioLabs)

KpnI restriction endonuclease (10 units/μL; New England BioLabs)

SacI restriction endonuclease (20 units/μL; New England BioLabs)

2.2.1.4. Plasmids
Table 2-4 Plasmids

<table>
<thead>
<tr>
<th>Plasmid name</th>
<th>Use of plasmid</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pcDNA3:YB1</td>
<td>Generation of human <em>YBX1</em> RNA by <em>in vitro</em> transcription using T3 RNA polymerase promoter</td>
<td>145</td>
</tr>
<tr>
<td>pGL4.17</td>
<td>Base plasmid into which human <em>E2F1</em> promoter fragment was cloned, control plasmid</td>
<td>Promega</td>
</tr>
<tr>
<td>728 pE2F1-Luc</td>
<td>Source of 728bp human <em>E2F1</em> promoter fragment</td>
<td>146</td>
</tr>
<tr>
<td>pcDNA3.1</td>
<td>Plasmid DNA used to control the amount of DNA transfected between transfections</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>pGL4.17-E2F1-728</td>
<td>Luciferase reporter for human <em>E2F1</em> promoter activity</td>
<td>Constructed during this project</td>
</tr>
<tr>
<td>pIRES2 eGFP</td>
<td>GFP reporter to determine transfection efficiency</td>
<td>Clontech</td>
</tr>
</tbody>
</table>
2.2.1.5. **Oligonucleotides**

Oligonucleotides were synthesised by Integrated DNA Technologies. Primers used for PCR are listed in Table 2-5. The restriction site in each sequence is underlined.

**Table 2-5 Oligonucleotide primer sequences**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence 5’-3’</th>
<th>Restriction site</th>
</tr>
</thead>
<tbody>
<tr>
<td>4N-Sacl-E2F1promoter-728Forwrd</td>
<td>aataagctctcagggccacgaattga</td>
<td>SacI</td>
</tr>
<tr>
<td>4N-HindIII-E2F1promoter-4revrs</td>
<td>atctagcttacgcccacatcttt</td>
<td>HindIII</td>
</tr>
</tbody>
</table>

**Table 2-6 YB-1 binding sequences**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence 5’-3’</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1xCohen competitor</td>
<td>cctcccacccctccccaccctccccaccctcccc</td>
<td>149</td>
</tr>
<tr>
<td>1xCohen competitor (with phosphorothioate-modified bases)</td>
<td>cctcccacccctccccaccctccccaccctcccc</td>
<td>149</td>
</tr>
<tr>
<td>3xCohen (biotinylated)</td>
<td>cctcccacccctccccaccctccccaccctccccaccctcccc</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>cctcccacccctccccaccctccccaccctccccaccctcccc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cctcccacccctccccaccctccccaccctccccaccctccccaccctcccc</td>
<td>149</td>
</tr>
<tr>
<td>3xFasR (biotinylated)</td>
<td>gtctggaactgcatccaaattcaggttctctggactgcatc</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>ccaatcaggttcgaactgcatccaaattcaggttcgct</td>
<td></td>
</tr>
</tbody>
</table>

2.2.1.6. **Bacterial strains and isolates**

*Escherichia coli* DH5α is an ATCC (American Type Culture Collection) strain used for cloning. The DH5α strain contains the following mutations: dlacZ Delta M15 Delta(lacZYA-argF) U169 recA1 endA1 hsdR17(rK-mK+) supE44 thi-1 gyrA96 relA1

2.2.1.7. **Bacterial culture media**

Luria-Bertani (LB) broth (Invitrogen) 1% (w/v) bacto-tryptone, 0.5% (w/v) bacto-yeast extract, 1% (w/v) NaCl
2.2.1.8. **Selective antibiotics**

Ampicillin (Amp; Gibco) Final concentration of 100 µg/ml for killing wild-type *E. coli*, and selecting for cells transformed with pcDNA3.1, pGL4.17, pGL4.17-E2F1-728 or pCMV E2F1 plasmids.

Kanamycin (Kan; Gibco) Final concentration of 100 µg/ml for killing wild-type *E. coli*, and selecting for cells transformed with pIRES2 eGFP or pIRES2 eGFP YB1 plasmids.

2.2.2. **Cell culture**

G418 Antibiotic (Gibco) Final concentration of 1250µg/mL for killing wild-type HCT116 and A375 cells and selecting for cells transformed with plasmids containing genes conferring resistance

2.2.2.1. **Cell lines**

**Table 2-7 Human cell lines and culture media in which they were grown**

<table>
<thead>
<tr>
<th>Name</th>
<th>Cell type</th>
<th>Culture medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>293T</td>
<td>Embryonic kidney</td>
<td>Complete Dulbecco’s Modified Eagle Medium supplemented with 5% Fetal Bovine Serum (v/v) and 1% Penicillin-Streptomycin-Glutamine (v/v)</td>
</tr>
<tr>
<td>HCT116</td>
<td>Colon cancer</td>
<td>RPMI 1640 supplemented with 5% Fetal Bovine Serum (v/v) and 1% Penicillin-Streptomycin-Glutamine (v/v)</td>
</tr>
<tr>
<td>A375</td>
<td>Melanoma</td>
<td>Complete Dulbecco’s Modified Eagle Medium supplemented with 5% Fetal Bovine Serum (v/v) and 1% Penicillin-Streptomycin-Glutamine (v/v)</td>
</tr>
</tbody>
</table>

2.2.3. **YB-1 binding nucleic acid assays**

Purified human YB-1 protein, with a C-terminal GST tag, kindly provided by Prof. Braithwaite, University of Otago

NuPAGE MES SDS Running buffer (20x)  Life Technologies, Cat# NP0002
6x Laemmli loading dye 375mM Tris-HCl pH 6.8, 9% SDS (w/v), 50% (v/v) glycerol, 0.03% (w/v) bromophenol blue

2.2.3.1. Antibodies

Table 2-8 List of antibodies used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Species</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-YB-1 N-terminal</td>
<td>Rabbit</td>
<td>Provided by Prof. Braithwaite, University of Otago¹⁴⁹</td>
</tr>
<tr>
<td>Anti-YB-1 N-terminal</td>
<td>Sheep</td>
<td>Serum provided by Prof. Braithwaite, University of Otago, raised to the same epitope as the rabbit antibody above</td>
</tr>
</tbody>
</table>

2.3. Methods

2.3.1. Bioinformatics

2.3.1.1. Software and programming languages

Microsoft Excel 2010 was used for storing, sorting, filtering and plotting data, however where possible its use was minimised by using alternative scripted methods, which are more reproducible and where results of operations can be more easily checked. For example, Perl was used for text processing and mining data from websites and the R programming language (version 2.13.2) was used for statistical analysis¹⁵⁰, using the R Revolution Enterprise 5.0 (Revolution Analytics) graphical user interface. Data visualisation in R was performed using the R packages:

- ‘gplots’ (http://cran.r-project.org/web/packages/gplots/index.html),

MySQL, Java and HTML were used in the construction of MelanomaDB.

Scripts used for data analysis have been included in the Appendix.

2.3.1.2. Source databases

Data from source databases was most frequently downloaded from FTP sites, or web-accessible download interfaces, with exceptions addressed in sub-sections below. Most frequently, data was downloaded in table format, such as comma-separated value (CSV) or
tab-delineated text files. These were imported directly into Excel if possible, or pre-processed as text using Perl. Source database managers were contacted for permission to use their publically accessible data.

2.3.1.2.1 Data-mining EBI druggability data
Perl code was used to download a series of webpages from the EMBI-EBI website as JSON (JavaScript Object Notation) files, and subsequently to extract druggability data. The full script used for this is provided in the Appendix. Pseudocode is given below as an outline of the process. Note that more recently EMBI-EBI has made their druggability data able to be directly downloaded, eliminating the need to interrogate each JSON file as described here.

```
Download every domain accession from EBI's File Transfer Protocol site at ftp://ftp.ebi.ac.uk/pub/databases/chembl/DrugEBIlity/
For each $accession get
    JSON from http://www.ebi.ac.uk/chembl/drugability/domain/$accession.json
    For each JSON get
        protein database code -> print to output file
        domain accessions -> print to output file
        average drugebility score section
            For each average drugebility score section get
                tractable score -> print to output file
                druggable score -> print to output file
                ensemble score -> print to output file
```

2.3.1.2.2 Clinical trials identifiers
All records from ClinicalTrials.gov that featured the word 'melanoma' were retrieved. Perl scripts were used to extract drug interventions from records. DrugBank was used to identify the targets of the drugs listed within trials. In the amalgamated database, each gene was annotated with the identifiers of clinical trials that used drugs targeting that gene. Due to variation in drug naming, this process was curated manually. The full script used to extract drug interventions is provided in the Appendix. Pseudocode is given below as an outline of the process.

```
Download all trials containing 'melanoma' from https://clinicaltrials.gov/ct2/resources/download
Split download into separate trials
For each trial
    get National Clinical Trial number -> print to output file
    get all text between term "Drug:" and section break -> print to output file
```
2.3.1.3. Gene identifier conversion

Entrez Gene ID was used as the unique gene identifier. Data containing other gene identifiers (e.g. Ensembl Gene ID, HUGO Gene symbol) or other object identifiers linked to genes (e.g. Affy ID, UniProt ID) were converted or linked to Entrez Gene ID using the online bioDBnet db2db151 and DAVID69 identifier converters. In order to prevent errors during conversion, gene identifiers were converted using both of these conversion tools. For each identifier, the converter identifier would be accepted if yielded from both tools. Mismatches, including identifiers converted by just one of the tools, were manually checked. DAVID and bioDBnet were selected for use due to their ease of use, and the ease of processing their output files.

2.3.1.4. Survival analysis

RNA expression data from Jonsson et al.152 and Bogunovic et al.153 were used in survival analysis that is detailed in Chapter 3. R scripts used during this analysis are included in the Appendix. Pseudocode is given below as an outline of the process.

Load data and process data format to fit input for analysis
From data, create an ordered list of survival time of each patient
From data, create an ordered list of survival status of each patient
From data, create a matrix with score for each probe set for every patient
For each probe set
   -> Print probe set name to output file
   Rank patients by probe set score
   For each decile of probe set score
      Divide patients into two groups at current decile
      Create survival curves for both groups
      Perform log-rank test for difference between two survival curves
      -> print log-rank p-value to output file
   Fit a Cox proportional hazards regression model for censored survival time on the predictor of probe set score
   -> print p-value of fitted model to output file

2.3.1.5. Literature and genomic data relationship strength

The IRIDESCENT154 and GAMMA155 software packages were used to derive data on literature relationship strength. These were used cautiously in this thesis as discovery tools.

2.3.1.5.1 Use of IRIDESCENT software

IRIDESCENT searches every abstract within MEDLINE for associations between objects, which include, for example, genes, diseases, phenotypes, ontology categories, chemical compounds and drugs. A network is created linking these objects via tentative relationships. The frequency in which two objects appear together in the same sentence or abstract
determines the relative strength of association between them. In this study, that network was interrogated for the strength of association between the name of each gene and the terms ‘melanoma’ or ‘metastatic melanoma’. The association network was created by IRIDESCENT’s creator Jonathan Wren (Oklahoma Medical Research Foundation and University of Oklahoma Health Sciences Center) and interrogated at my request for the purpose of this project.

IRIDESCENT returned to each gene a score of their direct literature relationship with the object 'melanoma' or ‘metastatic melanoma’ (Lit Str) and two scores of their implicit relationship (Obs/Exp, Min MIM). Lit Str (Literature strength) is a sum of the incidences where 'melanoma' or ‘metastatic melanoma’ is mentioned with the gene name or one of its synonyms; a gene scores 0.5 for every abstract and 0.8 for every sentence that co-mentions 'melanoma' or ‘metastatic melanoma’ and the gene term. For genes without a direct relationship, Lit Str is given as "Published Relationship below threshold". Obs/Exp is the number of observed relationships between two objects (the gene and melanoma/metastatic melanoma) in the network divided by the number expected by chance. A ratio less than 1 indicates that objects were observed together fewer times than were expected by chance. While the direct Lit Str measure is biased towards abundantly studied genes, the Obs/Exp score minimises historical bias of selection – thus it indicates how statistically exceptional those connections are, and reduces the score of genes long implicated in a variety of cancers.

Min. MIM (Minimum mutual information measure) is a measure of the cumulative minimum informativeness of the relationships shared by 'melanoma' and the gene, in slight contrast to literature strength indicated by Lit Str and Obs/Exp. For every implied connection between two objects there exists a chain of explicit connections between objects that link the two implicitly connected objects. For example, in this case, ‘melanoma’ and a gene may both be explicitly connected to a third object, such as a drug, a chemical or another gene. This common connection implicitly connects ‘melanoma’ and the gene. Connections may be through multiple intermediate links. Scores of mutual information are calculated between each connected link in the chain between ‘melanoma’ and the gene. The lowest of these scores within the chain is given as the minimal mutual information, which correlates well with the chain’s mutual information and correlates best with the observed strength and frequency of known associations\textsuperscript{156}.  

35
2.3.1.5.2 Use of GAMMA software

GAMMA conducts a meta-analysis of gene expression behaviour across 16,000 microarray experiments to infer function. GAMMA identifies co-expression connections between genes that are consistent across all different experiment types. If a gene has co-expression connections to other genes with published associations with melanoma, then it has an inferred connection to melanoma. The strength of published associations with 'melanoma' is calculated by IRIDESCENT.

2.3.2. Molecular biology

2.3.2.1. Polymerase chain reaction (PCR)

PCR reactions were conducted according to High Fidelity KAPA manufacturer’s instructions:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x KAPA HiFi buffer</td>
<td>5µL</td>
</tr>
<tr>
<td>Forward primer</td>
<td>300nM</td>
</tr>
<tr>
<td>Reverse Primer</td>
<td>300nM</td>
</tr>
<tr>
<td>dNTPs</td>
<td>300µM</td>
</tr>
<tr>
<td>Plasmid template</td>
<td>10ng</td>
</tr>
<tr>
<td>High Fidelity KAPA Taq DNA polymerase</td>
<td>0.5 Units</td>
</tr>
<tr>
<td>Water</td>
<td>To 25µL total reaction volume</td>
</tr>
</tbody>
</table>

After denaturing for 30 seconds at 94°C, the following steps were cycled through 25 times: 30 seconds of denaturation at 94°C, 60 seconds of annealing at 57°C, and 60 seconds of extension at 72°C. After a final extension period of 300 seconds at 72°C, the reactions were rapidly cooled, and then stored at 4°C.

2.3.2.2. DNA agarose gel electrophoresis

1% (w/v) agarose gels were prepared by dissolving agarose powder in TAE buffer in a beaker by heating to boil in a microwave. The boiled, liquid gel was cooled slightly by briefly running cold water outside the beaker. A small amount (<0.5µL) of GelRed Nucleic acid stain was added to approximately 100mL of still molten gel, mixed well and then poured into a mould. The mould was a flat, horizontal surface with a well comb inserted into the liquid.
gel. The gel was allowed to set at room temperature. The gel was placed into the gel tank, submerged in TAE buffer, and then the comb was removed. Then 7μL of each PCR reaction was mixed with 1.5μL of 6x DNA loading dye and loaded by pipette into the wells of the gel. A 1kb DNA ladder or mass ladder was added to another well for size or mass reference. DNA fragments were separated by electrophoresis at 100mV using TAE buffer for 30 to 90 minutes, depending on the size of DNA fragments to be viewed.

2.3.2.3. Isolation of DNA fragments by gel extraction
Following agarose gel electrophoresis of the PCR reaction, DNA bands were visualised on a UV light box and the desired band excised with a razor blade. The excised, DNA-containing agarose was put through a QiaQuick column according to the manufacturer’s instructions. The DNA fragments were eluted into 20 to 40μL of water, and then quantified by NanoDrop spectrophotometer.

2.3.2.4. Restriction endonuclease digestion
In a 25μL reaction volume, 1μg of plasmid or PCR product was incubated with 1 Unit of restriction enzyme in manufacturer-supplied ligation buffer (New England Biolabs) for an hour at the temperature recommended by the manufacturer. After the restriction digestion was complete, the reaction was incubated at 80°C for 15 minutes to inactivate the restriction enzyme.

2.3.2.5. Ligation
In 20μL reactions in manufacturer-supplied ligase buffer (New England Biolabs), 20-50ng of linear DNA was incubated with 1 Unit of T4 DNA ligase for 10 minutes at room temperature for ligation, and then 10 minutes at 65°C to inactivate the enzyme.

2.3.2.6. Transformation of chemically competent TOP10 E.coli cells
Chemically competent TOP10 E.coli cells were created and provided by Sandra Fitzgerald. 100μL of cells were incubated on ice with 10-100ng of plasmid for 30 minutes. Cells were subjected to heat shock for 2 minutes at 42°C, followed by 5 minutes incubation on ice. To this, 900μL of SOC media was added and the cells were incubated at 37°C for an hour rotating at 300RPM, before 100μL was removed and spread onto LB agar plates containing the appropriate antibiotic, either 100μg/ml ampicillin or kanamycin. The selective antibiotic used was that to which the plasmid confers resistance. The remaining cells were pelleted by brief centrifugation for 10 seconds at 1000 RPM, re-suspended in 100μL SOC media and
spread on a second LB agar plate. Plates were incubated at 37°C overnight. From these plates, colonies were picked and transferred into 5mL of LB broth containing selective antibiotic and grown for over 8 hours at 37°C with shaking at 200 RPM. Then, a 1.5mL aliquot of the culture from each colony was applied to Qiagen MiniPreps to isolate plasmid DNA. Plasmid DNA was digested by restriction enzymes and then DNA fragments were separated by gel electrophoresis to confirm transformation of the desired plasmid. The remaining cultures of the positive colonies were each transferred into 25mL of LB broth containing selective antibiotic, and then grown overnight at 37°C with shaking at 200 RPM. This is represented graphically in Figure 2-1.

Figure 2-1 Overview of the transformation of *E.coli* by DNA plasmids.

### 2.3.2.7. Isolation of plasmid DNA

As described in section 2.3.2.6 above, plasmid-transformed *E.coli* were grown in 25mL of LB broth containing selective antibiotic at 37°C overnight with shaking at 200 RPM. Plasmid DNA was extracted from bacterial cells by putting the 25mL cultures through a Qiagen
MidiPrep or MaxiPrep column, according to the manufacturer’s instructions. The concentration of the eluted plasmid DNA was quantified by using either a NanoDrop spectrophotometer or by comparison to the intensity of DNA bands from an Invitrogen High DNA Mass Ladder or Invitrogen Low DNA Mass Ladder following agarose gel electrophoresis.

2.3.2.8. Creation of new E2F1 promoter: luciferase reporter plasmid

Plasmids pGL4.17 and pIRES2 eGFP were transformed into TOP10 E.coli in order to prepare large quantities of these plasmids for downstream work. The 728 base-pair E2F1 promoter fragment was amplified by PCR from the plasmid ‘728 pE2F1-Luc’ (originally constructed by Johnson, Ohtani and Nevins\textsuperscript{146}), using the primers ‘4N-SacI-E2F1promoter-728Forwrd’ and ‘4N-HindIII-E2F1promoter-4revrs’. These primers include SacI and HindIII restriction sites, to introduce these sites onto the ends of the PCR product. The 728 base-pair E2F1 promoter fragment was thus flanked by SacI and HindIII restriction sites, which enabled directional cloning of the E2F1 promoter PCR product into the pGL4.17 vector. The E2F1 promoter PCR product and pGL4.17 vector were both digested with restriction enzymes SacI and HindIII as described above. The digested PCR product and pGL4.17 plasmid were separated by gel electrophoresis, and for each the appropriate bands were cut from the gel and the DNA extracted as described above. The PCR product was ligated into pGL4.17 to create the plasmid pGL4.17-E2F1-728. The new plasmid was amplified by transformation of E.coli cells that were spread onto LB agar plates containing ampicillin antibiotic, as described above. Colonies were picked, grown in culture before plasmids were extracted. In order to screen for plasmids into which the E2F1 promoter fragment had been inserted, the plasmids extracted from these colonies were digested with restriction enzymes and separated by gel electrophoresis. Plasmids digestions showing DNA bands of the appropriate size were considered to probably contain the E2F1 promoter fragment. In order to confirm this, the colonies containing these plasmids were further grown at 37°C overnight, before having larger quantities of plasmid extracted. Then Sanger sequencing was performed on 1000mg of isolated plasmid, by the DNA Sequencing Centre, Faculty of Science, University of Auckland. The DNA sequences were analysed using the Chromas Lite software (version 2.1.1).
2.3.3. Cell culture

2.3.3.1. Cell culture of human cell lines
The culture media for each cell line are listed in Table 2-6. The handling and sub-culturing procedure for all human cell lines were performed as instructed by ATCC (www.atcc.org). During maintenance of cells, media was supplemented with antibiotics penicillin (50 International Units/mL) and streptomycin (50μg/mL). Cells were incubated in a humidified incubator at 37°C with 5% (v/v) CO₂.

2.3.3.2. Transfection during assay development

2.3.3.2.1 Cell seeding
A375 or HCT116 cells were grown in T75 flasks until between 25% and 50% confluent. The culture media was discarded and the cells were rinsed with PBS. 2mL of trypsin was added to the cells, which were then incubated at 37°C for 5 minutes. After confirming that the cells had detached from the flask surface, 3mL of media was added to the cells to neutralise the trypsin, the cell suspension was then transferred to a tube, then the cells were pelleted by centrifugation at 1,000 RPM for 5 minutes. The cell pellet was gently re-suspended in 5mL of media and two 10μL samples were taken and dispensed onto a haemocytometer. The haemocytometer was used to obtain a cell count to determine the concentration of cells. The media was then diluted to give concentrations of 44,000 cells/mL for HCT116 cells, and 15,000 cells/mL for A375 cells.

2.3.3.2.2 Cell number per well
In 24-well plate wells, A375 and HCT116 cells were seeded at different numbers of cells per well to find the number of seeded cells that reach confluence at 96 hours after seeding. Wells were seeded with between 5,000 and 60,000 cells each. Optimal cell seeding numbers were determined so that experiments that concluded at 96 hours had the maximum number of cells per well. This was done to maximise influence of those cells upon experimental metrics, without complications caused by over-confluence – that is, exhaustion of growth media by an excess of non-dividing, unhealthy and dying cells.

2.3.3.2.3 Kill curve assay
A kill curve assay was used to determine the minimum concentration of the antibiotic G418 that would kill all wild-type A375 or HCT116 cells within 96 hours, so that this concentration could be later used to select for successfully transfected cells. Two thousand A375 or HCT116 cells were seeded into each well of a 96-well plate. G418 at concentrations ranging
from 0 to 2mg/mL (separated by increments of 250μg/mL) were added to the wells. The volume of the G418 solution accounted for less than 5% of the total volume of the culture media. The extent of cell death was observed over 96 hours.

2.3.3.2.4 Optimisation of transfection conditions

A number of transfection variables were tested. Initially, a comparison of transfection reagents and procedures was made on both A375 and HCT116 cells. Once those variables were selected, reagent quantities were optimised.

2.3.3.2.4.1 Selection of cell type, transfection method and lipofectamine reagent

In the wells of 24-well plates, HCT116 and A375 cells were seeded at 22,000 and 7,500 cells per well respectively. Cells were transfected either concurrently with seeding (termed ‘reverse transfection’) or 16-20 hours after seeding (‘forward transfection’), using two different volumes of either Lipofectamine 2000 (LF2000) or Lipofectamine 3000 (LF3000). In accordance with the manufacturer’s recommendations, 3μL or 4μL of LF2000 was incubated with 2μg of pIRES2 eGFP, and 0.75μL or 1.5μL of LF3000 was incubated with 1μg of pIRES2 eGFP. The variable transfection conditions, used for both A375 and HCT116 cells, are shown in Table 2-9 below:

Table 2-9 Different transfection conditions used. For both forward and reverse transfection methods, different volumes of LF2000 and LF3000 were used.

<table>
<thead>
<tr>
<th>Transfection method</th>
<th>Lipofectamine</th>
<th>Lipofectamine volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverse transfection</td>
<td>LF2000</td>
<td>4μL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3μL</td>
</tr>
<tr>
<td></td>
<td>LF3000</td>
<td>1.5μL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75μL</td>
</tr>
<tr>
<td>Forward transfection</td>
<td>LF2000</td>
<td>4μL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3μL</td>
</tr>
<tr>
<td></td>
<td>LF3000</td>
<td>1.5μL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75μL</td>
</tr>
</tbody>
</table>

The media was replaced at 24 hours after transfection. At 48 hours after transfection, the approximate and relative transfection efficiencies of the cells were estimated by microscopic visualisation of the number of green fluorescent cells from pIRES2 eGFP transfection.
2.3.3.2.4 Optimised transfection method for HCT116 cells

After this experiment, the use of A375 was discontinued in favour of using the more efficiently transfecting HCT116 cells. HCT116 cells were transfected using 0.75µL or 0.375µL of LF3000 with a range of pIRES2 eGFP plasmid amounts (1µg, 0.75µg, 0.5µg and 0.25µg). Microscope observations to determine transfection efficiency were performed 48 hours after transfection.

2.3.3.3 Development of clonal, stably transfected HCT116 cell lines

Stable transfected cell lines were desired in order to reduce the plasmids required in each transfection, and to reduce variation between experiments. HCT116 cells were transfected with pGL4.17 or pGL4.17-E2F1-728 in 6-well plates. 24 hours after transfection, the culture media of RPMI with 5% FBS was replaced with RPMI with 5% FBS and 1.25mg/mL of G418 antibiotic. Cells transfected with the pGL4.17 or pGL4.17-E2F1-728 plasmid and that had incorporated the plasmid into their genome would be resistant to the antibiotic. Media, containing G418 at the same concentration, was replaced approximately every three days. After 17 days, cells were lifted and serially diluted into 96-well plate wells, with the aim of obtaining some wells containing a single cell per well. Wells were inspected by microscope to identify those containing a single cell. Cells were grown and when confluent split into multiple wells. One well of each stable line was lysed and had luminescent signal measured. Stable lines that exhibited luciferase luminescent signal above background were grown further, before aliquots were frozen for later use.

2.3.3.4 Transfection for high-throughput screening assays

2.3.3.4.1 Transfection for pilot experiments in 96-well format

For pilot experiments prior to high throughput screening, 100µL of HCT116 cells at a density of 160,000 cells/mL were dispensed into each well of a 96-well plate (16,000 cells per well), in media without antibiotics. A pooled master mix was created, containing 150ng of plasmid DNA, 0.2µL of reagent P3000 and 5µL of OptiMem per well. This solution was added to a second master mix containing 0.15µL of Lipofectamine 3000 in 5µL of OptiMem per well. The solution was mixed, and incubated at room temperature for 5 minutes. To the cells in each 96-well plate well, 10µL of this solution was added dropwise. Cells were then incubated. Luciferase was measured by adding Dual-Glo Luciferase Substrate (Promega) 48 hours after transfection, as described in Section 2.3.3.5 below.
2.3.3.4.2 Transfection in 100mm cell culture dishes
For high throughput screening, transfections were then scaled up to 100mm cell culture dishes and transfection conditions re-optimised. HCT116 cells were lifted and seeded into 100mm cell culture dishes 12 to 18 hours prior to being transfected at a confluence of approximately 70%. In order to obtain a plate with confluence closest to 70%, cells were seeded into dishes at different cell densities. 12 to 18 hours later, the dish with confluence closest to 70% was selected for transfection.

For each transfection, three sets of wells were set up as described below. Those sets were:

1) Experimental cells transfected with plasmid, that later received compound
2) Control cells transfected with plasmid, that later did not receive compound
3) Control cells transfected with both plasmid and 1xCohen oligonucleotide “competitor”

8μg of plasmid DNA was mixed with 20μL of reagent P3000 in a volume of 500μL of OptiMem. This solution was added to 500μL of OptiMem containing 15μL of Lipofectamine 3000, mixed, and incubated at room temperature for 5 minutes. The solution was then added to the 100mm dish containing cells in 10mL of media, at approximately 70% confluence. These cells would become the sets 1 and 2, described above.

The control set 3 was simultaneously transfected with 1xCohen oligonucleotide “competitor” in addition to the DNA plasmid, using the same amount of LF3000. In this set, each well received 50pmol of 1xCohen competitor in addition to the plasmids. The 1xCohen competitor oligonucleotides used during screening had phosphorothioate bonds between the three bases at both the 3’ and 5’ ends.

The cells were incubated for 5 hours at 37°C with 5% CO₂, and then the media was replaced fresh RPMI with 5% FBS. The cells were incubated for an additional hour. Cells were then lifted with trypsin and re-suspended at a concentration of 266,666 cells/mL, and 30μL (8,000 cells) were dispensed into each well of a 384-well plate. Two hours later (8 hours after transfection) a robot was used to add screening compounds to the plate. Compounds were in DMSO, so 2μL of each compound was initially diluted in 18μL of media, and then 1.5μL of each diluted compound solution was added to a well of the 384-well plate containing the cells. This results in the culture media containing a final concentration of 0.5% (v/v) DMSO.
As a control, equivalent amounts of DMSO without any compound were added to the control wells of sets 2 and 3. The level of luciferase produced by the cells was measured at 36 hours after transfection, which is 28 hours after the compounds were added, by adding SteadyGlo Luciferase Substrate (Promega), as described in Section 2.3.3.5 below. This transfection process is shown graphically in Figure 2-2.

Figure 2-2 Overview of transfection for high-throughput screening, followed by addition of screening compounds and measurement of luciferase.

Picture of EnSpire plate reader taken from www.perkinelmer.com

2.3.3.5. Measurement of luciferase

48 hours after transfection, SteadyGlo Luciferase Substrate was added to each well at a volume equal to the volume of media in the well (e.g. 30µL in the 384-well plates used during high throughput screening), and the plate was incubated for 15 to 20 minutes at room temperature.
temperature before luminescence was measured using a Perkin Elmer EnSpire plate reader. Within this thesis, luciferase refers to firefly luciferase (as expressed by the pGL4.17 plasmids), unless explicitly referred to as Renilla luciferase (as expressed by the pRL-TK plasmid). When experiments used Renilla luciferase, the luminescent signal produced by the substrate for Renilla luciferase was measured after the luminescent signal produced by the substrate for firefly luciferase was measured, in accordance with the protocol for Promega’s Dual-Glo Luciferase Assay System.

2.3.3.6. Assessment of high throughput screen assay quality

The quality of a high throughput screen assay can be assessed by calculating the Z-factor. The Z-factor is a statistical measure that compares the means and variation of the positive and negative controls. The formula is:

$$Z\text{-factor} = 1 - \frac{3(\sigma_p + \sigma_n)}{|\mu_p - \mu_n|}$$

That is, 1 minus 3 times the sum of the standard deviation of the positive and negative controls, divided by the difference between the mean values of the positive and negative controls. An assay with a Z-factor below 0 is considered inappropriate for HTS, while a Z-factor between 0 and 0.5 is marginal, and between 0.5 and 1 is excellent.

2.3.3.7. Preparation of cytoplasmic and nucleic subcellular protein lysates

All processes were performed on ice, with buffers pre-chilled to 4°C. Fractionation Buffers A and C were supplemented with Roche Complete Proteinase and PhosphoSTOP Phosphatase Inhibitor Cocktail tabs. PBS was used to twice wash the cells. The cells were then scraped from the plate surface and transferred to a 15mL Falcon tube on ice. This tube was centrifuged for 3 minutes at 900RPM, and the supernatant was removed. The cell pellet was resuspended in 350μL of Fractionation Buffer A. Cells were incubated on ice for 10 minutes, then vortexed for 10 seconds and centrifuged at 20,000RPM for 10 seconds at 4°C. The supernatant, containing the cytoplasmic protein fraction, was transferred to a new tube on ice. The remaining nuclear pellet was re-suspended in 150μL of Fractionation Buffer C, and incubated on ice for 20 minutes, before it was centrifuged at 20,000RPM for 2 minutes at 4°C. The supernatant, containing the nuclear protein fraction, was transferred to a new tube on ice. Both fractions were split into smaller aliquots and stored at -80°C. Quantification of
protein lysates was performed using the Pierce BCA Protein Assay kit, according to manufacturer’s instructions. Briefly, Reagents A and B are mixed and added to samples of protein. Protein catalyses the reduction of copper to an ion, that can be detected by light absorbance. The rate of reduction is proportional to the amount of protein present. Samples are compared to a standard curve derived from known amounts of albumin.

2.3.4. YB-1 binding nucleic acid assays

2.3.4.1. Antibody purification from sheep serum

Serum from sheep inoculated with an N-terminal peptide from the human YB-1 protein was supplied by collaborator Professor Antony Braithwaite, Department of Pathology, University of Otago. Polyclonal antibodies were isolated by passing the serum through a Hi-Trap Protein G column within an AKTA fast protein liquid chromatography system as follows: 5mL of serum was diluted with 5mL of 20mM sodium phosphate buffer (pH 7.0) and then filtered through a 0.22μm filter. The protein G column was pre-equilibrated with the sodium phosphate buffer and then the serum sample was applied. Column-bound IgG antibodies were eluted using 0.1M glycine (pH 2.7). The eluted solution was eluted into a volume of 1M Tris (pH 9.0), that constituted approximately 10% of the volume of the eluted solution. As described below in Section 2.2.4.1.1, a sample of the eluted solution was separated by electrophoresis on a Coomassie blue gel, to confirm the removal of proteins other than IgG. The eluted solution was then dialysed, primarily to remove amines, as follows. The eluted solution was loaded into dialysis tubing and the tubing was placed into a large beaker containing 3 litres of PBS. This reaction was left at 4°C overnight, to affect buffer exchange.

2.3.4.1.1 Coomassie blue gel

To confirm that the only protein remaining after purification was IgG, a sample of the column eluted solution from antibody purification was separated on a Coomassie blue gel.

A 10% SDS page gel was cast in a Hoefer Dual Gel Caster, and when set a stacking gel was layered overtop and a comb inserted. 100μL of the column elution was mixed with 100μL of reducing buffer and incubated at 95°C for 5 minutes. This solution was used to fill wells on the stacking gel. Proteins were separated by electrophoresis in SDS running buffer, at 200V (constant voltage). After electrophoresis, the gel was stained in Coomassie Blue stain for 30 minutes at room temperature, with gentle shaking (approximately 50 RPM). The gel was then
rinsed in water and de-stained in Coomassie de-stain solution. The bands were visualised using a Gel Doc EZ (Biorad).

The antibody concentration in the dialysed elution solution was quantified by NanoDrop spectrophotometer, and validated by western blot using cellular lysates, as described below in Section 2.2.4.2.

2.3.4.2. Western blotting

3µL of 6x Laemmli loading buffer and 2µL of 1M DTT were added to each 15µL cell lysate sample containing 10-20µg of protein. These solutions were then heated at 90°C for ten minutes. A pre-cast NuPAGE Novex 4-12% Bis-Tris Protein Gel and the XCell SureLock Mini-Cell Electrophoresis System running tank were assembled and then NuPAGE MES SDS Running buffer was poured into the buffer reservoirs. The heated samples were then loaded into the wells and the gel run at 50V (constant voltage) initially for 15 minutes to allow the samples to enter the gel. Once samples were in the gel, 100V (constant voltage) was applied to the gel for 2 hours. Following this, the electrophoresis set up was dissembled and the gel was removed from the frame and transferred to an Invitrogen iBlot Gel Transfer Device. Using the Invitrogen iBlot Gel Transfer Device, the protein bands were transferred to nitrocellulose, according to manufacturer’s instructions. The nitrocellulose was incubated for one hour in a blocking solution of TBS-T with 5% (w/v) non-fat milk powder. The nitrocellulose was then incubated overnight at 4°C with gentle rocking (~50 RPM) in TBS-T containing the primary antibody at a dilution of 1:1000. The nitrocellulose was washed four times in TBS-T for 15 minutes per wash, at room temperature with gentle rocking. The nitrocellulose was then incubated for an hour at room temperature with gentle rocking in TBS-T with a 1:5000 dilution of secondary antibody. The nitrocellulose was washed again four times in TBS-T for 15 minutes per wash, at room temperature with gentle rocking. The surface of the nitrocellulose was then covered with WesternBright ECL HRP substrate (Advansta) and allowed to sit for 2 minutes. The excess substrate solution was drained, before the nitrocellulose was covered in plastic wrap. Luminescence visualised using an ImageQuant LAS 4000.
2.3.4.3. Generation of \textit{YBX1} RNA by \textit{in vitro} transcription

The plasmid “pcDNA3:YB1”, containing the human \textit{YBX1} gene adjacent to a T3 promoter was linearised by setting up the following restriction digest and incubating at 37°C for one hour:

\textbf{Table 2-10 Restriction digest by KpnI}

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmid</td>
<td>4(\mu)L (2(\mu)g)</td>
</tr>
<tr>
<td>BSA (20mg/mL)</td>
<td>0.2(\mu)L</td>
</tr>
<tr>
<td>10x New England Biolabs Buffer 2</td>
<td>2(\mu)L</td>
</tr>
<tr>
<td>KpnI restriction enzyme</td>
<td>0.2(\mu)L (2 units)</td>
</tr>
<tr>
<td>Water</td>
<td>To 20(\mu)L</td>
</tr>
</tbody>
</table>

In order to fill in the overhanging ends on the digested plasmid, the plasmid was incubated with 1 Unit of DNA Polymerase I, Large (Klenow) Fragment for a further 15 minutes at 37°C. The linearised plasmid was then purified using the Roche PCR purification kit according to the manufacturer’s instructions. 1\(\mu\)g of this linearised, blunt-end plasmid was used to synthesise RNA, using Promega’s Riboprobe In Vitro Transcription system (T3 RNA polymerase), according to the manufacturer’s instructions.

2.3.4.4. \textit{YB-1} Protein:nucleic acid Dot Blot Assays

A number of buffers (SW, SW+MgCl2, SW+MgCl2-KCl, SW+MgCl2-KCl-gly or Thermo RNA-protein binding buffers, see Section 2.2.1.1, Table 2-1) were trialled in this protocol to determine which buffer was optimal for YB-1 protein binding to nucleic acid. Buffers were always used cold (4°C), following the method described below.

A grid was drawn onto a nitrocellulose membrane. A narrow-mouthed pipette tip was used to dispense 2\(\mu\)L to 4\(\mu\)L of either DNA oligonucleotide (2pmol to 200pmol) or \textit{YBX1} RNA (0.7pmol to 1.4pmol) into each cell of the grid, and allowed to dry for 5 minutes. DNA oligonucleotides were cross-linked to the nitrocellulose membrane by exposure to UV light at 120,000 microjoules per cm\(^3\), and \textit{YBX1} RNA was cross-linked at 70,000 microjoules per cm\(^3\). The nitrocellulose membrane was treated as shown in Figure 2-3:
Figure 2-3 Dot blot procedure. Described is the treatment of nitrocellulose membrane, following attachment of oligonucleotides by UV cross-linking.

The enhanced chemiluminescence, horse-radish peroxidase substrate used was WesternBright ECL HRP substrate (Advansta).

2.3.4.5. ELISA-style plate-based Assays

Biotinylated 3xCohen oligonucleotides were diluted in MilliQ water to a concentration of 13nM and then 100μL (13pmol) were added to each well of a streptavidin-coated 96-well plate (Pierce, binding capacity 10pmol biotin/well). The plate was incubated at 50°C for 30 minutes, before being rinsed six times with cold SW buffer. 100μL of SW buffer containing 5nM purified GST-tagged YB-1 protein was added to each well, and the plate was incubated for 30 minutes at room temperature. Wells were rinsed six times with cold SW buffer. An anti-GST antibody conjugated to fluorescent marker (Fluorescein) diluted 1:2000 in SW buffer was added to the wells, and the plate was incubated for 30 minutes, before wells were rinsed six times with cold SW buffer. Plates were read on a Perkin Elmer EnSpire plate reader with fluorescence excitation wavelength set to 495nm, and emission wavelength set at 519nm.
2.3.4.6. **AlphaScreen assays**

2.3.4.6.1 **Preparation of acceptor beads conjugated to YB-1 antibody**

To generate AlphaScreen acceptor beads able to bind to YB-1, 100μg of the sheep YB-1 N-terminal antibody was conjugated to 1mg of AlphaScreen acceptor beads, according to the manufacturer’s instructions. This conjugation reaction was performed several times, when further acceptor beads were required to replenish stocks. Small batches were prepared in this way, rather than as one large batch, to maximise shelf-life of the beads.

2.3.4.6.2 **Initial AlphaScreen procedure**

During development of the protocol, many alternative protocols were trialled. Where alternative protocols were used, their descriptions accompany their results given in the Results section. The initial procedure is described here. This protocol was used with a variety of buffers, most commonly SW+MgCl2-KCl-gly buffer, SW+MgCl2-KCl-gly-EDTA buffer or PBS+BSA buffer. The majority of reactions were performed in Perkin Elmer OptiPlate-384, white opaque 384-well microplates, however some early experiments trialled Perkin Elmer OptiPlate-96, white opaque 96-well microplates. All incubations were performed in darkness, and at room temperature. The procedure for each 50μL reaction is described in Figure 2-4.

![Figure 2-4 50μL AlphaScreen reaction procedure](image)
2.3.4.6.3 Final optimised AlphaScreen procedure

Reactions were performed in Perkin Elmer OptiPlate-384, white opaque 384-well microplates. All incubations were performed at room temperature. The procedure for each 50μL reaction is described in Figure 2-5.

10μL of buffer containing purified YB-1 protein (final reaction concentration: 40fmol/μL) or whole cellular lysate (final reaction concentration: 12.5ng/μL) dispensed into each well of a 384-well plate (Perkin Elmer OptiPlate)

10μL of buffer containing 1x Cohen competitor oligonucleotide (final reaction concentration: 1pmol/μL) was added to control wells. 10μL of buffer was added to all other wells

Incubation at room temperature for 30 minutes

10μL of buffer containing AlphaScreen acceptor beads (final reaction concentration: 20ng/μL) and biotinylated 3x Cohen (final reaction concentration: 2.5fmol/μL) was added to each well

Incubation at room temperature for 60 minutes in darkness

20μL of buffer containing donor beads (final reaction concentration: 20ng/μL) was added to each well

Incubation at room temperature for 60 minutes in darkness

Well signals read on Perkin Elmer Enspire plate reader, using an excitation wavelength of 680nm, and detecting an emission wavelength of 570nm

Figure 2-5 Final AlphaScreen procedure
### 2.3.4.7. TruHits assay

The AlphaScreen TruHits kit was used to detect false positive compounds that interfere with the AlphaScreen assay components. The procedure for each 25μL reaction is described in Figure 2-6.

**Figure 2-6 TruHits assay procedure**

1. 10μL of PBS + 0.2% (w/v) BSA containing 25μg/mL Streptavidin-donor beads was dispensed into 384-well plate wells (Perkin Elmer OptiPlate-384).

2. Compound stocks (in DMSO) were diluted in PBS + 0.2% (w/v) BSA so that 5μL of compound was added to each well, resulting in a final reaction compound concentration of 20μmol/L, and a final reaction DMSO concentration of 1%. 5μL of PBS + 0.2% (w/v) BSA with DMSO (without compounds) was added to control wells.

3. Incubation at room temperature for 30 minutes.

4. 10μL of PBS + 0.2% (w/v) BSA and containing Biotin-acceptor beads was added to each well, for a final reaction concentration of 12.5μg/mL.

5. Incubation at room temperature for 30 minutes.

2.3.4.7.1 Make-shift assay for false positives

Prior to the laboratory’s acquisition of the TruHits kit, an alternative assay for false positives was devised. The procedure for each 25μL reaction is described in Figure 2-7.

![Figure 2-7 Procedure for make-shift assay for false positives](image)

2.3.5. Computational filtering

Computational filtering can be performed on compound structures to remove compounds possessing specified traits or containing certain substructures. Filters were applied using the SYBYL-X 2.11 software, with compound structure inputted in Structure Data File (SDF) format.

The first filter applied was to eliminate compounds containing substructures identified as pan-assay interference compounds (PAINS) \(^\text{157}\) that would indicate they are likely false positives. Five increasingly stringent filters were applied to eliminate groups unfavourable for drug development, such as groups with toxicity, poor pharmacokinetic behaviour, or that are highly electrophilic. Those filters, from least stringent to most stringent, were: WEHL_93K, Baell 2013 Filters 1, 2 and 3, and the CTX filter \(^\text{158}\).
Chapter 3 Database development and creation of MelanomaDB

3.1. Introduction

As described in the introduction to this thesis, there is a need for both integration of diverse melanoma data, and development of drugs to previously un-drugged melanoma drivers. To this end, the aim of the work presented in this chapter was development of a database to aid drug target selection. Integration of diverse melanoma data into a coherent database could also be a useful aid for hypothesis generation across a varied range of melanoma research questions. This chapter introduces the motivation and practical considerations for this data integration exercise, explains what was done, then discusses its potential utility, potential criticisms of the database and aspects of the database that require caution when it is used.

3.1.1. Basis of the database

The initial objective of integrating diverse melanoma data was to create a tool that could aid discovery of novel melanoma drug targets. As melanoma involves disruption of the normal functioning of melanocyte cellular pathways, a sensible drug target would be a protein component of these pathways; therefore, the basic unit of the database should be either protein or gene identifiers.

3.1.2. Existing database tools for study of cancer genomics

Web-based genomic database resources have been established for a range of purposes\textsuperscript{159}. Resources such as COSMIC\textsuperscript{57,58}, TCGA (http://cancergenome.nih.gov/) or CBioPortal\textsuperscript{160} act as a repository of cancer genomic information, while others such as UCSC cancer browser\textsuperscript{161} act as tools for data visualisation. Tools for data analysis and integration are offered by databases such as CanEvolve\textsuperscript{162} and the Broad Institute’s Firehose (https://confluence.broadinstitute.org/display/GDAC/Home), which integrates genomic data with a limited amount of clinical data. However, these databases do not specifically set out to integrate data in the manner most useful for drug development, and therefore lack annotations of proteins with targeting drugs, or druggability estimations. Current databases that focus on applying genomics to drug discovery include Genomics of Drug Sensitivity in Cancer (GDSC)\textsuperscript{163}, and CanSar\textsuperscript{164}. At the time this project began, GDSC was not yet available, and CanSar was under development.
At the beginning of this project, a survey found only one web-based melanoma genomic database, The Melanoma Molecular Map project\textsuperscript{165}. This database provides an excellent interactive review of melanoma, with links to molecular pathways and targeted therapies for melanoma as well as a patient database. However, this database does not integrate a diverse range of genomic data with other data types, such as the drug data that it provides separately. Two other melanoma genomics databases have arisen during this project: MelGene\textsuperscript{166} and the Melanoma Gene Database\textsuperscript{167}. However, these are almost exclusively focused on genomic data, without integration into other data types.

3.1.3. Genomic data

Genomic data was a major part of the publically available data that this project sought to integrate. This included sequencing, gene variant and expression data from melanoma tumours. Therefore, it was decided that the database should be gene-based. Entrez gene ID was selected as the unique gene identifier. Entrez Gene ID numbers are unique, and their unambiguous numerical identifiers are less prone to error than systems using gene symbols (such as those of the HUGO Gene Nomenclature Committee (HGNC)). That is, genes with similar names can be confused when using gene symbols, but less so when they are labelled with unrelated numerical identifiers.

An annotated list of all Entrez Gene IDs for homo sapiens were downloaded from (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/; Accessed: 30/7/2012) to form the starting point for database construction.

3.1.3.1. Conversion of gene identifiers

As described above, it was decided to integrate genomic data from studies that use different gene identifiers, by converting/cross-referencing all identifiers into Entrez Gene IDs. However, it was found that conversion of gene identifiers is not trivial, and downstream analyses can be undermined by distorted input due to incorrect identifier conversion, or errors introduced during conversion\textsuperscript{168}.

A number of online tools are available for conversion of gene identifiers, including DAVID\textsuperscript{69}, bioDBnet’s db2db\textsuperscript{151} and BioMart\textsuperscript{169}. In order to prevent errors during conversion, gene identifiers were converted using two conversion tools. For each identifier, the converter identifier would be accepted if yielded from both tools. Mismatches, including identifiers
converted by just one of the tools, were manually checked. DAVID and bioDBnet were selected for use due to their ease of use, and the ease of processing their output files.

### 3.1.4. Consultation with researchers and clinicians to guide collection of data

One of the main objectives of the database was the discovery of new drug targets for melanoma. Therefore, to avoid generating a superfluous tool that would not be utilised, additional relevant information was sought, including data on drugs, drug targets, protein druggability, and input from translational cancer researchers and clinicians. Researchers and clinicians were consulted in depth to determine what information they would find useful, so that data collection could be directed to include relevant sources.

Researchers were relatively easy to consult. The concentration of research labs within the University of Auckland environment is conducive to casual, spontaneous discussion. Suggestions were also drawn from published literature. Clinicians were more difficult to consult, and in this regard the project is grateful to Dr Ben Lawrence (Medical Oncologist) and Dr Gavin Harris (Anatomical Pathologist) for their time. Dr Lawrence was interviewed on two occasions, initially identifying the type of data and level of detail that is of use to an oncologist, and secondly giving feedback on the accessibility of the data’s presentation. Dr Harris was formally consulted once and informally multiple times, providing insight into how pathology may incorporate bioinformatic approaches. These interviews revealed several unexpected and valuable insights, such as that these clinicians wished to be able to visualise gene pathway diagrams annotated with data (such as mutation status or RNA expression) from the individual patient including clinically relevant data, notably identification of those proteins that can be therapeutically targeted. The interviews found that also valuable would be an overview of the genomic data from a collection of tumours in the form of an annotated, clustered heat map, providing context into which would fit genomic data from an individual patient’s tumour. Both of these points, along with other requirements that we have already envisaged, were incorporated into the thesis work described in this Chapter.
3.2. Data collection and processing

3.2.1. Overview

Since the output of this chapter was the generation of a database, unlike a laboratory-based thesis chapter, this chapter does not have a separate methods and Results section. Instead it combines those two sections into this Section, 3.2, which describes the data that was collected, the sources from which that data was taken, and how it was processed and integrated into the combined database. The types of data collected were:

i) RNA expression data
ii) Drug and biomarker data
iii) Druggability data
iv) Literature and genomic data relationship information
v) Somatic variant data

3.2.2. RNA expression data

A survey was undertaken of literature and data depositories (e.g. Gene Expression Omnibus (GEO)\textsuperscript{170} and ArrayExpress\textsuperscript{52}) for publicly available melanoma RNA expression microarray data. A commonly used microarray platform is the Affymetrix Human Genome U133 Plus 2.0 Array. This is the platform for which the highest number of samples have been submitted to GEO (http://www.ncbi.nlm.nih.gov/geo/browse/?view=platforms&display=20&zsort=samples; Accessed: 5/4/2016). Annotations for the U133 Plus 2.0 array probe sets were retrieved from Affymetrix (http://www.affymetrix.com/support/technical/annotationfilesmain.affx; Accessed: 25/7/2012), and used to annotate the gene entries of the database.

Data sources containing RNA-Seq data, such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) were not integrated into the database. At the time of this survey, this data had a relative lack of clinical annotation. In particular, the lack of patient survival information was considered a critical limitation.
3.2.2.1. Sources of RNA expression data

Survey of literature and two microarray data depositories yielded the sets of melanoma microarray data from gene expression studies, listed in Table 3-1, for which patient survival data was collected:

Table 3-1 Sources of RNA expression data

<table>
<thead>
<tr>
<th>Publication first author name and year of publication</th>
<th>Data depository and accession number</th>
<th>Microarray Platform</th>
<th>Number of patients</th>
<th>Melanoma sample stage</th>
<th>Survival data availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Augustine 2010(^{171})</td>
<td>GEO: GSE19293</td>
<td>Affymetrix U133 Plus 2.0</td>
<td>29</td>
<td>IV</td>
<td>Not available</td>
</tr>
<tr>
<td>Bogunovic 2009(^{153})</td>
<td>GEO: GSE19234</td>
<td>Affymetrix U133 Plus 2.0</td>
<td>38</td>
<td>III, IV</td>
<td>Available</td>
</tr>
<tr>
<td>John 2008(^{172})</td>
<td>ArrayExpress: E-TABM-403</td>
<td>Custom array: MWG Human 30K</td>
<td>30</td>
<td>IIIB, IIC</td>
<td>Available</td>
</tr>
<tr>
<td>Jonsson 2010(^{152})</td>
<td>GEO: GSE22153</td>
<td>Illumina human-6 v2.0</td>
<td>57</td>
<td>IV</td>
<td>Available</td>
</tr>
<tr>
<td>Journe, 2011(^{173})</td>
<td>Not available</td>
<td>Affymetrix U133 Plus 2.0</td>
<td>32</td>
<td>III</td>
<td>Available</td>
</tr>
<tr>
<td>Mandruzzato 2006(^{174})</td>
<td>Not available</td>
<td>Custom array: 17,500 probes</td>
<td>38</td>
<td>III, IV</td>
<td>Available</td>
</tr>
<tr>
<td>Mansfield 2010</td>
<td>GEO: GSE23376</td>
<td>Affymetrix U133 Plus 2.0</td>
<td>22</td>
<td>IV</td>
<td>Not available</td>
</tr>
<tr>
<td>Martins 2011(^{175})</td>
<td>GEO: GSE17275</td>
<td>Custom array: 4,608 probes</td>
<td>60</td>
<td>I, II, III, IV</td>
<td>Not available</td>
</tr>
<tr>
<td>Riker 2008(^{176})</td>
<td>GEO: GSE7553</td>
<td>Affymetrix U133 Plus 2.0</td>
<td>40</td>
<td>IV</td>
<td>Not available</td>
</tr>
<tr>
<td>Xu 2008(^{177})</td>
<td>GEO: GSE8401</td>
<td>Affymetrix U133A</td>
<td>30</td>
<td>unspecified</td>
<td>Not available</td>
</tr>
</tbody>
</table>

Data from the five studies for which melanoma-specific patient survival data was available were incorporated into the database. From these studies, the strength of statistical associations between RNA abundance and melanoma-specific survival were determined. Associations were either taken directly from the original conclusions of these studies, or from our additional statistical analysis. In addition to these patient survival studies, the database also incorporated data from a review by Timar et al.\(^{178}\) of studies associating RNA abundance with metastasis in melanoma.

3.2.2.1.1 Missing value estimation

Raw microarray data can contain missing values, and the analyses performed in this project contained functions that cannot process missing values. Therefore, the missing values must be estimated. For example, the Affymetrix microarray expression data from Bogunovic et al.
included one tumour for which the reading of one gene expression probe set was given as an
abnormally large value. This value was replaced with an average value for the probe set
across all tumours in the data from that paper. Another probe set contained identical,
extremely low values for all tumours. This was presumed to be an artefact of processing, and
excluded from later analysis. If the values were real, then they represent a probe set without
any variation, from which meaningful information could not be derived.

Since missing data or severely outlying data was sparse, and since we wished to alter the
genomic data sets as little as possible, more sophisticated methods for replacing missing data
such as K-Nearest Neighbours, Singular Value Decomposition\textsuperscript{179} were not required.

\textbf{3.2.2.1.2 Normalisation of microarray data}

Normalisation is required to correct for variation between raw microarray data due to
systematic biases that exist between experiments; such as those caused by different plate
readers or different batches of dye-labelled probes. Multiple normalisation procedures are
available for microarray data, including RMA, GCRMA and MAS5\textsuperscript{180}. The ‘best’
normalisation method depends upon the aims of an experiment (which are known) and on the
properties of the data set (which are largely unknown). The general view of bioinformaticians
is that if a strong signal is present in the data, then post-normalisation statistical analysis is
still likely to detect it, regardless of which method is used. If a result of analysis is not robust
to different normalisation methods, then it may not represent a real trend in the data. For this
project all Affymetrix microarray expression data were normalised using the RMA method
with background correction (see Section 3.2.2.2.1 below).

\textbf{3.2.2.1.3 Quality control on data sets}

Technical laboratory procedures can be performed unsatisfactorily, resulting in inaccurate
data containing technical artefacts and high levels of noise that disguise true signals. Quality
control algorithms exist that can detect data produced by unsatisfactorily performed
laboratory procedures, such as the R package named ‘AffyQCReport’\textsuperscript{181}
(https://www.bioconductor.org/packages/release/bioc/html/affyQCReport.html). This was
used to generate a quality control report for the data from Bogunovic et al.\textsuperscript{153}, and indicated
that the data was of sufficient quality to proceed to statistical analysis.

The authors of Jonsson et al\textsuperscript{152} processed the data of their study using Illumina’s Beadstudio
v3 software before making it publically available. Quality control was applied during this
processing - loci were only retained if expression measurements could be made across at least 80% of samples.

3.2.2.2 Integration of published gene expression associations and analyses of RNA expression data

Associations between gene expression in melanoma and patient survival were taken directly from the results reported by John et al.\textsuperscript{172}, Mandruzato et al.\textsuperscript{174} and Journe et al.\textsuperscript{173}. Associations between gene expression and metastasis were taken directly from Timar et al.\textsuperscript{178}. The database was annotated with these associations linked to the relevant genes. The raw microarray and patient survival data of Bogunovic et al.\textsuperscript{153} and Jonsson et al.\textsuperscript{152} were used in survival analyses.

3.2.2.2.1 Analysis of Bogunovic et al. data

The raw Affymetrix microarray data of the Bogunovic et al.\textsuperscript{153} study was normalised using RMA normalisation performed using the ‘affy’ package in the R statistical software\textsuperscript{182}. Using R, the patients were split into two groups, and survival objects were created for each group and compared using a Log Rank test. This splitting was performed 9 times for each probe set, once at each RNA abundance decile across the patient population. R was also used to fit a Cox proportional hazards regression model to each probe set. Each gene was annotated in the database with the lowest Cox proportional hazards p-value from any of its probe sets (using the ‘survival package in R with Benjamini–Hochberg multiple testing correction applied; https://cran.r-project.org/web/packages/survival/survival.pdf).

After examining the distribution of p-values, when generating the database a probe set was considered associated with survival if its Cox proportional hazards p-value was less than $1 \times 10^{-4}$, or if any Log Rank p-value was less than $1 \times 10^{-6}$. Furthermore, a probe was also considered to be associated with survival if it had both a log-rank p-value lower than $1 \times 10^{-4}$ and a Cox proportional hazards p-value lower than $1 \times 10^{-3}$. Subsequently, 317 genes were annotated as having probe sets associated with survival. For these genes associated with survival, database annotations included whether poor patient survival was associated with either high or low gene expression. These statistical cut-offs were more stringent that may strictly be required, chosen to minimise false positive signals in the resulting database. The arbitrary nature of this decision must be acknowledged, however a decision was nevertheless
required and was based on careful visual examination of survival curves and discussion with a statistician (Dr Mik Black, University of Otago).

3.2.2.2 Analysis of Jonsson data
The Illumina microarray data from the Jonsson et al. study was obtained in a normalised format, however, three patients without survival data were removed, and, in order to eliminate negative values, all microarray values were adjusted by adding a minimum value equal to the lowest negative value. This Jonsson data was analysed in the same manner as the Bogunovic data, as described in Section 3.2.2.2.1 above. For each probe set, p-values were obtained for Log Rank tests performed at 9 decile splits, and for Cox proportional hazards regression models.

After examining the distribution of p-values (as described above), a probe was considered associated with survival if the Cox proportional hazards p-value was less than $5 \times 10^{-4}$. Additionally, the tumours were divided into two groups for log rank tests, once at each decile (10-quantile). Subsequently, 9 log rank tests were performed for each probe using the ‘survdiff’ function in R. A probe was considered associated with survival if any log-rank p-value was less than $1 \times 10^{-6}$. Furthermore, a probe was also considered to be associated with survival if it had both a log-rank p-value lower than $1 \times 10^{-4}$ and a Cox proportional hazards p-value lower than $5 \times 10^{-3}$.

3.2.3. Drug and biomarker data

3.2.3.1 DrugBank drug targets
A list of all drugs and their targets was retrieved from DrugBank (DrugBank.ca; Accessed: 15/02/2012), and used to annotate each gene with the drugs that directly interact with its encoded protein. Drugs have intended targets and also off-target interactions. Genes were annotated with DrugBank identifiers for drugs that bind their encoded proteins, regardless of whether this interaction was the intended target of the drugs. For example, the drug Cetuximab binds its intended target (human epidermal growth factor receptor) but also binds complement components and Fc receptors, as is expected of an antibody drug (http://www.drugbank.ca/drugs/DB00002).

3.2.3.2 Protein targets of melanoma drugs
Genes were annotated if their encoded proteins or RNA products are targets of approved melanoma drugs, or targets of any drugs that have been clinically trialled for use in
melanoma. Data on approved melanoma drugs was taken from DrugBank and KEGG DRUG\textsuperscript{74}. Data on drugs used during clinical trials on melanoma was taken from Therapeutic Targets Database\textsuperscript{184} and Clinicaltrials.govt.

3.2.3.3. Melanoma biomarkers
Genes were annotated if their encoded proteins or RNA products have been identified as melanoma biomarkers according to any of the following review studies listed in Table 3-2. If a review indicated that the biomarker was for a specific aspect of melanoma, then the annotation included this data.

Table 3-2 Sources of melanoma biomarkers data

<table>
<thead>
<tr>
<th>Source</th>
<th>Types of Biomarker</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gould Rothberg \textit{et al.} 2009</td>
<td>Categorised according to Hanahan-Weinberg functional capabilities</td>
<td>\textsuperscript{185}</td>
</tr>
<tr>
<td>Schramm and Mann 2011</td>
<td>Prognostic</td>
<td>\textsuperscript{186}</td>
</tr>
<tr>
<td>Utikal \textit{et al.} 2007</td>
<td>Serologic, Immunohistochemical</td>
<td>\textsuperscript{187}</td>
</tr>
<tr>
<td>Mehta \textit{et al.} 2010</td>
<td>Prognostic</td>
<td>\textsuperscript{188}</td>
</tr>
<tr>
<td>KEGG BRITE</td>
<td>Basal cell carcinoma, Squamous cell carcinoma</td>
<td>\textsuperscript{74}</td>
</tr>
</tbody>
</table>

Additionally, genes were annotated as melanomagenesis oncogenes or tumour suppressors if they were identified as such in the review by Flaherty, Hodi and Fisher 2012\textsuperscript{189}.

3.2.4. Druggability data
A survey was taken of publically available sources that estimate the tractability of proteins to modulation by drugs. Sources of druggability data that were included in the database are listed in Table 3-3.
Table 3-3 Sources of druggability data

<table>
<thead>
<tr>
<th>Source</th>
<th>Data type</th>
<th>Literature reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sophic Integrated Druggable Genome Database</td>
<td>Database</td>
<td>190</td>
</tr>
<tr>
<td>EBI’s DrugEBIlity Database</td>
<td>Database</td>
<td>191</td>
</tr>
<tr>
<td>Uniprot</td>
<td>Database</td>
<td>192</td>
</tr>
<tr>
<td>Secreted Protein Database</td>
<td>Database</td>
<td>193</td>
</tr>
<tr>
<td>KinBase</td>
<td>Database</td>
<td>194</td>
</tr>
<tr>
<td>Li and Lai 2007</td>
<td>Paper</td>
<td>195</td>
</tr>
<tr>
<td>Tiedemann et al. 2012</td>
<td>Paper</td>
<td>196</td>
</tr>
</tbody>
</table>

Genes were also annotated if their encoded proteins are kinases or are secreted proteins, as these characteristics are relevant to their druggability. Kinase information was downloaded in FASTA format from Kinbase\(^{194}\). Data on secreted proteins was downloaded from UniProt\(^{192}\) and the Secreted Protein Database\(^{193}\).

3.2.5. Literature and genomic data relationship information

3.2.5.1. Use of IRIDESCENT software

IRIDESCENT identified in published literature 370 genes having a direct or implicit relationship to the object ‘melanoma’ above a set threshold. Each of these 370 genes was annotated with a score of its direct literature relationship with the object 'melanoma' (Lit Str) and two scores of the implicit relationship (Obs/Exp, Min MIM). In this instance, sensitivity (rather than selectivity) to genes associated with ‘melanoma’ was desired, with the provision that the database user could set their own filtering threshold. Thus, a Obs/Exp score over the threshold of 0.74 (i.e. less than 1) was required for a gene’s annotation.

This process and the resulting annotations were also performed for the object ‘metastatic melanoma’. The term ‘metastatic melanoma’ was central to this study, therefore it was desired that the database user could set the threshold for association. Additionally, the term ‘metastatic melanoma’ is more specific than the broader term ‘melanoma’, so that weaker association could still be of interest to a database user. A minimal Obs/Exp threshold of 0.01 was thus used, so that 12703 genes having any literature relationship to ‘metastatic melanoma’ were subsequently annotated.
3.2.5.2. Use of GAMMA software

GAMMA inferred a relationship to melanoma for 2,428 genes. Each of the 2,428 genes was annotated with the Score for the strength of their connection to melanoma, Obs/Exp, and the Lit Str of that gene's published connection to melanoma, as given by IRIDESCENT.

3.2.6. Somatic variant data

Multiple sources reporting melanoma variants were collated for inclusion in this database. 11 exome sequencing studies were identified from the literature, and variants from 3 databases were included. Sources of variant data are listed in Table 3-4.

Table 3-4 Sources of somatic variant data

<table>
<thead>
<tr>
<th>First author of study and year of publication</th>
<th>Melanoma samples studied</th>
<th>Tables used (T = Table, ST = Supplementary Table)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berger (2010)</td>
<td>8 patient-derived short-term cultures and 2 cell lines</td>
<td>T1 (Novel Gene fusions), T2 (readthrough transcripts), ST2 (All Gene Fusions), ST5 (nonsynonymous coding variants)</td>
<td>197</td>
</tr>
<tr>
<td>Berger (2012)</td>
<td>25 germline-matched metastases</td>
<td>ST4 (Somatic base pair mutations in protein coding regions), ST6 (Small indels in protein coding regions), ST7 (Somatic structural rearrangements in 25 melanoma genomes)</td>
<td>89</td>
</tr>
<tr>
<td>Nikolaev (2011)</td>
<td>6 germline-matched cell lines</td>
<td>ST1 (List of all somatic mutations in the studied melanomas), ST1-Somatic indels, ST4 (Copy number of genes from aCGH and SNP arrays and transcript expression)</td>
<td>198</td>
</tr>
<tr>
<td>Palavalli (2009)</td>
<td>79 germline-matched metastases</td>
<td>T1</td>
<td>199</td>
</tr>
<tr>
<td>Prickett (2009)</td>
<td>79 germline-matched metastases</td>
<td>T1</td>
<td>200</td>
</tr>
<tr>
<td>Prickett (2011)</td>
<td>11 germline-matched</td>
<td>T1 (Somatic mutations identified in recurrently mutated GPCRs), ST1 (Somatic mutations identified in GPCR</td>
<td>201</td>
</tr>
<tr>
<td>Study</td>
<td>Cell lines matched</td>
<td>Cell lines retrieved</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Stark (2011)</td>
<td>8 germline-matched cell lines</td>
<td>ST3 (Somatic mutations identified via exome sequencing)</td>
<td></td>
</tr>
<tr>
<td>Turajlic (2012)</td>
<td>1 primary and metastatic tumour pair</td>
<td>ST2a (Non-synonymous germline variants annotated for dbSNP and Cancer Gene Census entry), ST3 (Validated structural variations)</td>
<td></td>
</tr>
<tr>
<td>Wei (2011)</td>
<td>14 germline-matched metastases</td>
<td>T1 (Recurrent mutations identified in melanoma whole exome sequencing and prevalence screen), ST2-WGS alterations (Somatic mutations identified in the Discovery Screen - WGS alterations &gt;05)</td>
<td></td>
</tr>
<tr>
<td>Hodis (2012)</td>
<td>15 germline-matched primary tumours, 30 metastases, 76 metastases-derived short-term cultures</td>
<td>ST4A (All called somatic substitutions and small indels in 121 melanomas, Related to Figure 1), ST4B (Manually reviewed coding small indels in 121 melanomas, Related to Figure 1), ST9 (Recurrent somatic mutations, Related to Figure 3)</td>
<td></td>
</tr>
<tr>
<td>Krauthammer (2012)</td>
<td>48 primary and 99 metastatic tumours and cell lines</td>
<td>ST1a (SNVs), ST1b (InDels), ST1c (Splice Site variants), ST1d (Additional Mutations)</td>
<td></td>
</tr>
<tr>
<td>CCLE</td>
<td>6 cell lines</td>
<td>COLO-829, CP50-MEL-B, CP66-MEL, LB2518-MEL, LB373-MEL-D, MZ7-mel</td>
<td></td>
</tr>
<tr>
<td>COSMIC</td>
<td>52 cell lines</td>
<td>A101D_SKIN, A2058_SKIN, A375_SKIN, C32_SKIN, CHL1_SKIN, CJM_SKIN, COLO679_SKIN, COLO741_SKIN, COLO783_SKIN, COLO792_SKIN, COLO818_SKIN, COLO829_SKIN, G361_SKIN, HS600T_SKIN, HS688AT_SKIN, HS695T_SKIN, HS834T_SKIN, HS839T_SKIN, HS852T_SKIN</td>
<td></td>
</tr>
</tbody>
</table>
In total, data was collected on 58 established melanoma cell lines, 119 primary ‘short-passage’ cell lines, 38 primary tumours and 96 metastatic melanoma tumours. The majority of samples have been paired with matched normal samples to ensure that the variants reported are somatic; the exceptions to this are some samples from COSMIC, and the 10 samples from the Berger et al. 2010 study. Non-synonymous coding mutations, insertions and deletions (indels; both in-frame and frame-shift), splice-site mutations and structural rearrangements (including gene fusions and read-through transcripts) were integrated into the database. Synonymous coding mutations were not included. In total, somatic variant data included more than 35,000 non-synonymous coding mutations and more than 3,500 structural rearrangements, insertions and deletions, across 16,488 genes.

### 3.3. MelanomaDB

#### 3.3.1. Amalgamation of all data into gene sets

All of the data collected was combined into a single gene-based matrix using Entrez Gene ID as the unique identifier for each gene. Each gene was represented by one row, with each column containing data from a single source.

A number of gene sets were derived from this data matrix. Most frequently, the columns of the matrix were converted directly into gene sets that included every gene containing an annotation in that column. For example, the DrugBank Drug ID gene set contained every gene that had been annotated as the binding partner of any DrugBank Drug. For some
columns, such as statistical associations between RNA expression and patient survival, a cut-off was used to define gene set membership.

3.3.2. SQL database generation

Working in collaboration with bioinformatics programmer Christoph Knapp (University of Auckland), a web-accessible SQL database named MelanomaDB was created using my generated gene sets (http://genesetdb.auckland.ac.nz/melanomadb/about.html). MelanomaDB facilitates access, combination and filtering of these different types of data, and allows interpretation of the data in terms of molecular pathways and functional categories. Gene sets were derived from the data matrix as sets of gene identifiers, which were then fit into a suitable format for input into MelanomaDB. The web interface to this database was created solely by Christoph Knapp and was implemented using Apache, PHP, Javascript and HTML. MelanomaDB can directly access the meta-gene set database GeneSetDB\textsuperscript{206}, previously generated by our research group, in order to identify the intersection between melanoma-specific gene sets and gene sets related to molecular pathways and biological functions.

With MelanomaDB it is easy to find the union or intersection between any number of melanoma gene sets, and any user-submitted gene lists. These lists, unions or intersections can be piped into GeneSetDB to interrogate molecular pathways for which they are enriched. Multiple iterations are possible; for example, a union of some melanoma-associated gene sets may be compared against the union of some other melanoma-associated gene sets to find their intersection, which can then be piped into GeneSetDB to identify enriched molecular pathways.

3.3.3. Assembly of information for individual tumours

A tumour-based matrix was constructed from the assembled exome and whole-genome sequencing information. Each row was a gene, and each column a tumour, with cells describing somatic variants present in that gene for that tumour. This somatic variant data included 310 samples, including 173 and 64 samples which had somatic alterations in the \textit{BRAF} and \textit{NRAS} genes respectively. The union of somatic variants found across multiple samples was used when multiple sequenced tumours or cell lines from the same patient were available.
3.3.4. Data visualisation

A clustered heat map of tumour variants for genes included in the KEGG ‘Melanoma’ signalling pathway was constructed using the ‘gplots’ package (http://cran.r-project.org/web/packages/gplots/index.html) of the statistical software R. The ‘binary’ method was used for distance calculation, and the ‘single’ method was used for clustering. R was also used to draw gene network diagrams. The ‘graphviz’ package (http://www.bioconductor.org/packages/2.11/bioc/html/Rgraphviz.html) was used to plot KEGG pathways included in the graphite R package (http://www.bioconductor.org/packages/release/bioc/html/graphite.html).

3.4. Melanoma Profiler

A selected sub-set of MelanomaDB data was used to create a freely available Illumina BaseSpace (REF) application called Melanoma Profiler. Melanoma Profiler was created, in close collaboration with me, by Auckland biological data visualisation company BioMatters, in collaboration with my supervisor Prof. Cris Print. Melanoma Profiler identifies gene variants from a normal-matched tumour provided by the BaseSpace user or taken from the BaseSpace archive. This list of variant genes is entered into GeneSetDB to retrieve molecular pathways and gene sets with statistically significant pathway enrichment for that list of genes. Pathways can be visualised relative to targeting drugs and other clinically-related information, and to show the tumour’s variant genes in the context of these molecular pathways. Heatmaps can be generated to compare the sample tumour to the 310 melanoma samples catalogued in MelanomaDB. Melanoma profiler is available at https://apps.biomatters.com/melanoma-profiler/welcome.

3.5. Discussion

3.5.1. Integration of high level data

There is a question of which ‘level’ data should be integrated. The lowest level of gene expression data from a set of tumours is the raw microarray intensities counts that come directly off the scanner at the time of experiment. This data then undergoes sequential processing, and after each step the data can be considered of a higher level. That sequential processing includes these steps:
Data cleaning → Log transformation → Missing value estimation → Normalisation → Analysis → Results and visualisation → Conclusions

At each step of microarray data processing there are a variety of methods that can be applied. The processing methods used result in different outputs, that subsequently effect downstream processing and analyses\textsuperscript{180,207}. One possible approach is to integrate low level raw data from different studies, and process all raw data using the same methods. However, it is difficult to select uniform processing methods suitable for diverse data sets. In addition, technical variation due to batch effects will still have a significant influence on the data even after normalisation, evident during hierarchical clustering that groups microarrays according to their laboratory and batch of origin\textsuperscript{208}. Several batch correction methods exist to reduce this batch effect\textsuperscript{209}.

Alternatively, data can be integrated at a higher level. For example, analyses can be performed separately on individual sets of data, with output generated in a format easily integrated. An advantage of this approach is that data is processed by the methods selected by the original researchers responsible for its generation, who may select methods based on knowledge of the data that is not publically available. That is, without complete knowledge of how the raw data was originally generated, it is difficult for the integrator of many data sources to select appropriate processing methods. This difficulty is avoided if higher level data is integrated. For this project, in general higher level integration of individual studies’ data was preferred where possible.

3.5.1.1. Extremely high level integration of data was not performed

It would have been possible to perform integration at an even higher level, such as integration of raw data from separate studies prior to analysis. At first, this may seem justifiable – for example, studies that look at only a tightly defined range of tumour types can lack the variation in dependant variables required to make inferences. Each individual study may show no relationship between gene and survival, but when the studies are combined a trend can be inferred.

However, different study parameters (e.g. patient groups, anatomical tumour sites, histological tumour types) make it difficult to draw conclusions that are universal across multiple studies\textsuperscript{67,178}. For example, Wang et al.\textsuperscript{67} found little overlap in genes used in prognostic gene signatures between studies (and in different metastatic and invasive
signatures). Combination of multiple strong, but disparate, signals may result in those signals being lost amongst each other.

Furthermore, combining raw microarray data from different studies can lead to artefacts; results may depend not so much upon the data and more upon how the data was put together. For example, gene-matched data combined from different microarray platforms (e.g. U133, U95), even if normalised and standardised, tend to cluster with data from the same platform.

Therefore, this project applied the same analysis separately on data taken from different studies, and compared the results of these analyses. The reported results of different studies are also integrated and compared. That is, biological indicators can be taken directly from the reported results of a publication, and also from analyses performed on the data from a publication.

3.5.1.2. Limitations of an integrated database approach

The inherent danger of any collation of data from multiple sources is that the collated data inherits any flaws in its component parts. That is, it is difficult to assess the reliability of source data, yet conclusions are reliant upon that data. This risk is difficult to control. However, all the sources of data used in this study have associated peer-reviewed papers, and have thus been through the peer review process. Additionally, this study selected for use constituent databases that are extensively used, and on which it is possible to perform spot checks. Where possible, intersecting multiple independent data sources of similar information may reduce the chance of propagating random errors, which are (theoretically) unlikely to co-occur in independently generated sources.

One of the ‘open review’ reviewers who had accepted the paper describing the MelanomaDB database\textsuperscript{210} for publication in the journal ‘Frontiers in Genetics’ (Dr William Reinhold) subsequently wrote a commentary about the database, which appeared in the 13 August 2013 issue of the same journal\textsuperscript{211}. This commentary, in addition to describing the usefulness of our integrative database approach, was also critical of the potential for websites that integrate data from other websites to perpetuate inaccuracies inherent in the original data and amplify their influence in the field. Dr Reinhold points out that mRNA expression data may be especially difficult to assess, since data problems are difficult to recognise. This criticism and caution has been taken on board. As Dr Reinhold says, this problem affects any integrative database initiative of this type. It seems that the best protection may be care in selection of original data, and careful quality control where possible. Dr Reinhold may not have been
aware that an advantage of Affymetrix array data is the ability to go back to raw data files (cel files) as I did in this case and only included files that passed the quality thresholds found in the ‘AffyQCReport’ R package\textsuperscript{181}. Specifically, each of the arrays signal intensity density plot distributions were unimodal and either normal or right-skewed, with none having an extremely low average intensity. The glyceraldehyde 3-phosphate dehydrogenase (GAPDH) 3’ to 5’ ratio for every array was below 1.25, and the centre of intensity co-ordinates were all within a magnitude of 0.5 of centre. Nevertheless his caution seems well-founded and should be taken into account when using any database like MelanomaDB.

3.5.1.3. The necessity of periodic updates to an integrated database

The creation of the database described in this Chapter was followed by its utilisation for various applications, described in Chapter 4. A sub-set of these applications lead to the work described in the later Chapters of this thesis. During the time that this work was pursued, more data has become publically available. In particular, The Cancer Genome Atlas has made available a large amount of melanoma genomic data not available at the start of this project. Furthermore, some of the component databases that were integrated into this database have themselves been updated. Therefore an update of MelanomaDB will be undertaken in the near future. The applications described in the following Chapter were made using the database in the form it existed at the time of its publication.
Chapter 4 Database Application

4.1. Introduction

A database was created in order to build a tool that can aid identification of putative drug targets of treatment of melanoma. Data from various sources were assembled as described in Chapter 3. The assembled data takes the simple form of a data matrix, with annotations added to each unique gene identifier. Each row is a gene, and each column contains data from one source. In addition to access via a web interface and via the SQL query language, the database can be opened in spreadsheet applications, such as Microsoft Excel, and easily sorted, filtered and utilised. This facilitates use of the data in exploratory analyses for hypothesis generation. This Chapter describes examples of hypothesis generation performed during this project that illustrate the utility and versatility of the database, including examples utilising the functions of MelanomaDB. The Chapter concludes with the analysis that primarily motivated the construction of the database: that is, utilisation of the integrated data to propose some putative melanoma drug targets.

4.2. Insights into melanoma obtained from an overview of collected melanoma samples

With 310 melanoma samples from numerous studies collected into one database, it is possible to plot the assembled data to gain an overview of the ‘biological range’ inherent in these samples. In total, data was collected on 58 established melanoma cell lines, 119 primary ‘short-passage’ cell lines, 38 primary tumours and 96 metastatic melanoma tumours. Non-synonymous coding mutations, indels, and/or structural rearrangements were reported in 16,488 genes.

The majority of genes show somatic variations in only a small number of tumours. This can be seen by plotting the distribution of the number of tumours containing a somatic variant in each gene (Figure 4-1, A below). Two genes, TTN (Titan) and BRAF are altered in the majority of the 310 samples (60% and 56%, respectively), but 97.6% of genes are altered in 5 or fewer samples. Further insight is gained by plotting total exon length against somatic variation frequency (Figure 4-1, B). A statistically significant, but weak correlation (Pearson’s correlation coefficient = 0.47, p≤0.001) exists between total exon length of genes.
and the somatic variation frequency. Variations implicated as melanoma drivers in large genes such as *TTN* may frequently occur in melanomas due to large gene size increasing the likelihood of passenger mutations. The *BRAF* gene, which is of medium exon length, stands out as frequently mutated in melanomas.

There is large variation in the number of genes that can be altered in a tumour (Figure 4-1, C), with some tumours having alterations in nearly 3000 genes. All tumours with alterations in more than 100 genes were associated with primary tumours in sun-exposed areas of skin.
Figure 4-1 An overview of the genes mutated across the 310 cell-line, primary and metastatic melanoma samples
(A) The distribution of the number of tumours with somatic alterations in each individual gene. (B) Each gene’s total exon length in base pairs (y-axis) versus the number of the 310 tumours with a mutation in that gene (x-axis). (C) The distribution of the number of genes with somatic alterations in each individual tumour.

4.2.1. Genes most frequently altered in metastatic melanoma tumours

The database allows calculation of the proportion of melanomas that carry somatic variations in each gene/locus on a genome-wide scale. The database was used to identify those genes in which over 10% of the 96 sequenced metastatic (not primary) melanomas carry non-synonymous somatic variations. This identified a list of 245 genes, including genes that have been the focus of recent publications describing mutations in melanoma, such as PREX2\textsuperscript{55}, GRM3\textsuperscript{201}, and ERBB4\textsuperscript{200} (other melanoma-associated genes such as MAP3K5/9\textsuperscript{202}, MAP2K1/2\textsuperscript{198} and RAC1\textsuperscript{48,54,55} are included as mutated genes in human tumours in MelanomaDB but fall outside this list of 245 genes).

4.2.2. Putative melanoma tumour suppressors and oncogenes

The database was then used to generate lists of putative melanoma tumour suppressor genes and melanoma oncogenes. Several approaches could be used. One approach for example, generates a list of genes mutated in more than 10% of melanoma metastases in which expression is associated with shorter melanoma-free patient survival. Known tumour suppressors and oncogenes (NRAS, KIT, and WNT family members) were removed from this list to yield a list of genes that may warrant further investigation in the laboratory. This approach suggested two putative tumour suppressors, COL14A1 and THEMIS, as suggested by their low expression associating with shorter melanoma-free patient survival; and four putative oncogenes, SALL1, PAPPA, SACS and NETO1, as suggested by their high expression associating with shorter melanoma-free patient survival.

4.2.3. Gene alterations in sun-exposed primary tumours

In cases where data was available, melanomas can be sub-divided based on the location of the primary tumour into sun-exposed and non sun-exposed cohorts. Recognising the limitations of this approach, primary tumours located on the legs, arms (including shoulders), neck and
head were considered sun-exposed, while samples with primary tumours located on the trunk, groin and feet were considered non sun-exposed.

The tumours could be further subdivided based on this assessment of the location of the primary tumour (sun-exposed versus non sun-exposed skin). For each gene, the proportion of somatic variant-containing tumours that originate from sun-exposed primary tumours was plotted against the proportion of somatic variant-containing tumours derived from non sun-exposed primary tumours (Figure 4-2). The diagonal line in the Figure represents equal proportions of tumours derived from sun-exposed and non sun-exposed primary tumours. As expected due to ultraviolet light induced DNA damage, most genes are plotted as dots above that line, indicating a greater proportion of mutations in tumours derived from sun exposed primary tumours.
Figure 4-2 The proportion of somatic variant-containing tumours derived from sun-exposed primary tumours (y-axis) compared to the proportion of somatic variant-containing tumours derived from non sun-exposed primary tumours (x-axis).

Genes are represented as open circles (in some cases the genes overlie one another). Genes lying along the blue diagonal line represent the situation when somatic mutations are present in a gene at the same proportion of tumours derived from sun-exposed and non sun-exposed primary tumours. Labelled in blue are genes that are mutated in the highest proportion of sun-exposed tumours (TTN, DNAH5, BRAF, LRP1B), non sun-exposed tumours (BRAF, TTN, MUC16, LRP1B) and genes with the greatest difference of mutated proportion between sun-exposed and non sun-exposed groups (DNAH5, TTN, PREX2, ERBB4).
4.2.4. Melanoma gene network

A putative gene network involved in melanoma was derived from the literature\textsuperscript{3,30,39}. The MelanomaDB database was used to annotate the network, indicating the percentage of melanoma tumours included in the database that are mutated in each gene, and the genes with RNA expression associated with patient survival (Figure 4-3).

![Figure 4-3 Putative melanoma gene network.](image)

Genes are shaded according to the percentage of tumours in which each gene is mutated, with darker shades of red indicating a higher percentage of mutated tumours. Square nodes indicate that the expression of these genes is associated with patient survival.

4.2.5. Cross-validation of database use examples with TCGA melanoma data

The database described in Chapter 3 was constructed before melanoma data from The Cancer Genome Atlas (TCGA) was made easily accessible. It is useful in retrospect to validate some of the findings of this Chapter with analyses performed on the now available melanoma data from TCGA.
Of the 245 genes mutated in more than 10% of the 96 metastatic melanoma samples within MelanomaDB (described in Section 4.2.1), 196 are also mutated in more than 10% of the 470 melanoma samples within TCGA (Skin Cutaneous Melanoma, Mutation Annotation Format, version 2.4). This includes the melanoma-associated genes emphasised as the subjects of recent studies \textit{PREX2}, \textit{GRM3}, and \textit{ERBB4} and similarly excludes other melanoma-associated genes \textit{MAP3K5/9, MAP2K1/2} and \textit{RAC1}. A high correlation exists between MelanomaDB and TCGA for the percentage of melanoma samples with protein-altering variants in each gene (Pearsons correlation coefficient = 0.757). The percentage variant-containing melanoma samples from each study is plotted in Figure 4-4.

\textbf{Percentage of melanomas with protein-altering variants for each gene}

![Graph showing the correlation between MelanomaDB and TCGA samples containing variant(s)]
Figure 4-4 Percentage of TCGA and MelanomaDB melanoma samples with protein-altering variants for each gene.
Each gene is given as a data-point showing the percentage of TCGA samples in which that gene contains at least one variant (y-axis), plotted against the percentage of MelanomaDB samples in which that same gene contains at least one variant (x-axis). The line-of-best-fit is shown in red.

Section 3.2.2.2 described analyses performed on mRNA expression data from the studies by Bogunovic et al\textsuperscript{153} and Jonsson et al\textsuperscript{152} that associated mRNA abundance in melanoma tumours with patient survival. The same analysis was also performed on the equivalent data from the TCGA melanoma cohort to associate mRNA expression with patient survival. These analyses produced lists of genes with mRNA expression associated with patient survival for each of the three datasets. The size of these gene lists and the intersection between them is shown in Figure 4-5. The hypergeometric distribution\textsuperscript{212} was used to determine the probability of the size of the intersections between these gene lists being due to chance alone. None of these intersections are larger than what can be expected to occur by chance. Furthermore, there is no statistically significant correlation between the Cox proportional hazards regression model or Log Rank test p-values of analyses performed on any two of the three datasets.
Figure 4-5 Analyses produced lists of genes with mRNA expression associated with patient survival for three datasets: TCGA (blue circle), Bogunovic et al (red circle), Jonsson et al (green circle).

The numbers given in the circles indicate the number of genes with mRNA associated with patient survival in just a single dataset. The genes named at the intersections of the coloured circles representing the three datasets are those that had mRNA expression associated with patient survival in two datasets. No genes had mRNA associated with patient survival in all three analyses of all three datasets. Hypergeometric tests indicated that the size of the intersections between these gene lists was not larger than that expected by chance alone (p-values for each test given adjacent to the intersections for which they were calculated).

It has been previously noted by Wang et al\textsuperscript{67} and Timar et al\textsuperscript{178} that there is little concordance between studies associating the expression of individual mRNA in melanoma tumours with patient survival. This lack of concordance may in part be due to the different patient groups, histopathological tumour types and tumour sites used during different studies. The study of
Bogunovic et al examined lymph node metastases (39 from patients with stage III disease, 5 from patients with stage IV disease), while the study of Jonsson et al examined both lymph node and subcutaneous metastases, the majority of which (54 out of 57) were from patients with stage IV disease. The 470 TCGA melanoma samples includes 103 primary tumours, and metastases taken from a wider range of sites (including intestine, bone, brain and liver) beyond lymph nodes or subcutaneous sites. The TCGA cohort also includes patients who received neo-adjuvant therapy administered prior to sample resection. While differences between these datasets may preclude confirmation of their separate conclusions, a broader set of associations are enabled by the amalgamation of analyses from orthogonal datasets. That is, some potentially clinically useful associations may be detectable only in a restricted patient subset or in tumours from a particular site. The union of results from different datasets is appropriate when identifying as many associations as possible.

The data visualisation tool OncoLnc (http://www.oncolnc.org) was used to investigate whether the melanoma data within TCGA could verify the two putative tumour suppressors and four putative oncogenes suggested in Section 2.4.2. Of these, only the putative tumour suppressor gene THEMIS was validated by the RNA expression and clinical annotation data from TCGA melanoma project. Within the TCGA melanoma samples, higher than median expression of THEMIS was associated with longer melanoma-free patient survival (p-value = 1.36*10^{-7}). The intersection of this result between different datasets suggests that the association is robust.

It is possible to recreate the visual data summary shown in Figure 4-2 by plotting the same parameters using TCGA melanoma data (Figure 4-6). The plot of this data also suggests that most genes are altered in a higher proportion of sun-exposed than non-sun-exposed tumours. Of the 13,830 genes which contain a protein-altering somatic variant in at least one TCGA melanoma sample, 8,120 genes are mutated in a higher proportion of sun-exposed samples than non sun-exposed samples, a higher number than what is expected by chance alone (p-value = 2.485*10^{-93}).
Figure 4-6 The proportion of somatic variant-containing tumours derived from sun-exposed primary tumours (y-axis) compared to the proportion of somatic variant-containing tumours derived from non sun-exposed primary tumours (x-axis). Data shown here is from the melanoma samples contained within TCGA.

Genes are represented as open circles (in some cases the genes overlie one another). Genes lying along the blue diagonal line represent the situation when somatic mutations are present in a gene at the same proportion of tumours derived from sun-exposed and non sun-exposed primary tumours. Labelled in blue are genes that are mutated in the highest proportion of both sun-exposed and non sun-exposed tumours (*TTN, MUC16, BRAF, DNAH5, PCLO, LRP1B*) and genes with the greatest difference of mutated proportion between sun-exposed and non sun-exposed groups (*BRAF, THSD7B, SPTA1*).
4.3. Analysis of signalling pathways relevant to melanoma via GeneSetDB and MelanomaDB

Further insights are possible when use of the database is combined with the functions provided by gene set analysis tool GeneSetDB\textsuperscript{71}. For example, Section 4.2.1 described a composite list of 245 genes that carrying non-synonymous somatic variations in over 10% of the 96 sequenced metastatic melanomas present in the database. This list of genes is available from MelanomaDB. By submitting that list of 245 frequently mutated genes to GeneSetDB for gene set enrichment, it is possible to identify biological functions that may be commonly disrupted in melanoma. When false discovery rate was set to 0.01, this list of 245 genes was found to be significantly enriched for 56 Pathway and Gene Ontology gene sets. Amongst these 56 gene sets, the highest ranked gene set was homophilic cell adhesion (GO:0007156), with 11 other gene sets associated with cell adhesion. The second highest ranked gene set was extracellular matrix structural constituent (GO:0005201), with 4 other gene sets associated with the extracellular matrix. Additionally, 7 gene sets were associated with axon guidance or neurite out-growth, 6 gene sets with microtubule-based movement, and 11 sets with actin or myosin filaments. It is possible that these groups represent pathways utilised by melanoma cells for migration during metastasis.

4.3.1. Incorporation of KEGG signalling pathways into MelanomaDB

The KEGG signalling pathways\textsuperscript{74} contained within the R ‘graphviz’ package were annotated with data collected in MelanomaDB. This annotation function was added to MelanomaDB, and an example is shown below in Figure 4-7. The nodes of the KEGG pathway ‘Melanoma’ have been coloured with shades of red according to the frequency of non-synonymous somatic variants for each gene in melanoma. 13 nodes were plotted as boxes rather than circles to indicate that the expression of the gene was significantly associated with patient survival in the analysis performed on data taken from Bogunovic \textit{et al.}\textsuperscript{153} (Cox proportional hazards model, p≤0.05, no multiple correction applied). In accordance with the known importance of the signalling events represented in this pathway upon melanoma formation and progression, significantly more of the genes in this pathway carried somatic variants than expected due to chance alone (Fisher’s exact test with right-tailed hypergeometric distribution, p≤0.002).
4.3.2. Analysis of melanoma signalling pathways in individual tumours

Sequencing information from an individual patient’s tumour can be viewed in the context of all of the tumours collated in MelanomaDB by using this pathway-specific information. Other information from MelanomaDB can be added as annotations. For example, a clustered heatmap featuring the genes from the KEGG ‘Melanoma’ signalling pathway was plotted and annotated with druggability indices, current drug targets, COSMIC census genes, gene-survival associations, known melanoma driver mutations and somatic variant frequency in melanoma. This plot, shown below in Figure 4-8, was created using the Rgraphviz package of R, and is provided as a function in MelanomaDB. The clustering in this heatmap was driven by somatic gene variants, potentially stratifying patients into groups with common biological changes, which may be susceptible to particular pathway-targeted therapies. For example, there is a cluster of tumours with BRAF as the only somatic variant in the pathway. These tumours may be suitable for Vemurafenib therapy, given the lack of other somatic
variants in this pathway that could potentially contribute to Vemurafenib resistance. A new tumour's mutation profile could be added to an existing clustering analysis to identify which previously studied tumours have similar mutation profiles. This may aid prognostication and treatment stratification.

Figure 4-8 Clustered heatmap for genes encoding proteins of the KEGG “Melanoma” signalling pathway.

Each column represents a gene within the KEGG Melanoma signalling pathway and each row represents an individual tumour. Blue bars indicate the presence of a protein-altering somatic variant in a particular gene within a tumour. Based on this variant information, the clustering of genes and tumours was performed using single linkage clustering with binary distance. The clustered figure was then annotated with additional information above the heatmap. The first row shows those genes encoding known drug targets according to version 3 of the DrugBank database (shown as red blocks). In the second row yellow blocks mark genes encoding
potentially druggable proteins, as indicated by the MelanomaDB gene set “Druggability: Sophic ENSEMBL list”\textsuperscript{214}. In the third row, orange and red blocks indicate genes mutated in \( \geq 1\% \) or \( \geq 5\% \) of the 310 melanomas in our database, respectively. In the fourth row, the blue blocks indicate those genes with expression significantly associated with patient survival (\( p \leq 0.05 \) no multiple testing correction applied, Cox proportion hazards model, Bogunovic et al. data\textsuperscript{153}). In the fifth row, the brown blocks indicate genes that are members of the Wellcome Trust Cosmic “Cancer Gene Census” gene set, as on 1st March 2013 (http://cancer.sanger.ac.uk/cancergenome/projects/census/). In the sixth row, the purple blocks mark genes thought to be melanoma drivers when mutated (MelanomaDB gene set “Melanomagenesis Drivers”\textsuperscript{189}). The figure is available in higher resolution from MelanomaDB (http://genesetdb.auckland.ac.nz/tumour-pathway-analysis/select.php).

An additional function of MelanomaDB is the plotting of a specific tumour’s somatic alterations in a specific molecular pathway. Visualisation of a protein-altering variant in the context of the encoded protein’s position in a molecular pathway may be relevant to specific targeted therapies. For example, a known activating mutation in \( KRAS \), downstream of a drug target may indicate potential for resistance to the drug (e.g. Cetuximab targeting EGFR).

The list of somatic variants in a single tumour can be used to perform gene set enrichment analysis, to suggest the pathway useful for visualisation. This was performed for one sequenced metastatic melanoma: sample ME029 from the Berger et al. 2012\textsuperscript{89} cohort. The list of somatic variants within the tumour was submitted to the GATHER web tool\textsuperscript{68}, which identified the “Neuroactive ligand-receptor interaction” as significantly enriched. This pathway was then plotted, with genes annotated with mutational status within that tumour (Figure 4-9).
4.4. Putative melanoma drug targets

A major motivation for assembling this database was the need to identify new drug targets for melanoma therapy. A series of filters can be imposed upon the assembled data to sift down to a short list of genes meeting certain criteria, such as those with traits desirable for a putative drug target. There are numerous ways to prioritise desirable traits for a drug target – that is, what traits are considered important, and how much weight should be placed upon each trait. It is difficult to objectively state which traits are most important, and a number of ways to do this were considered. One approach is to identify traits common to existing drug targets and then seek novel targets possessing those traits. However, this may generate a list of putative targets that have already received sufficient attention from researchers, and exclude unusual targets that lack a common trait. Another approach is to permutate selection criteria and
generate multiple short lists, and look for the genes that occur most frequently. However, this still requires subjective selection of criteria, and also selects against an unusual, novel target. Selection of criteria proceeded with recognition that subjective choices were unavoidable; the justifications for the selection of each trait are given here.

4.4.1. Inhibition of an over-active, wild-type target

Many oncogenes come to be recognised due to their high frequency of mutations in tumours. Subsequently, many drugs are inhibitors of over-active targets encoded by variant genes containing activating mutations. A well-known example in melanoma is inhibition of the oncogenic effect of the activating V600E \textit{BRAF} mutation by targeted therapies such as Vemurafenib.

The activity of a protein is more easily inhibited than increased by drug binding, and thus the majority of existing drugs are inhibitors targeting over-active proteins. Likewise, this project aims towards developing an inhibitor drug.

Given that common melanoma mutations are more likely to reduce activity of their encoded protein and are thus unsuitable for therapeutic inhibition, it could be contended that selection of a drug target should exclude frequently mutated genes.

There is a further reason why a drug inhibitor is a more desirable goal than a drug that activates its target. A drug that inhibits an over-active wild-type protein can be, in effect, countering multiple mutations in the upstream regulators of the protein that have caused its increased activity. Furthermore, further mutations in these upstream regulators will not confer resistance to the drug. Mutations in the target itself would have to be within the drug’s binding site to confer drug resistance; mutations elsewhere in the target are likely to reduce activity, and thus additionally inhibit protein activity.

4.4.2. RNA expression associated with metastasis or patient survival

Because of the reasons given above, it was decided that candidate drug targets identified in this project should be proteins that are over-active and in wild-type form in melanoma. While it is not feasible to take direct measurements of protein activity from all proteins in the cell, RNA expression levels are an available representation of protein activity. As therapy is applied to advanced melanomas, there should be a link between the activity of putative drug targets and patient survival.
Data from studies by Jonsson et al.\textsuperscript{152} and Bogunovic et al.\textsuperscript{153}, that was analysed as described in Section 3.2.2.2.), yielded survival associations that were directional: that is, either high or low expression of certain genes was associated with poor survival. As an over-active target was desired, the assembled database was queried for genes with an association between high RNA expression and patient survival in either of these two data-sets. This query yielded a list of 465 genes that were taken as a starting point for a list of candidate drug targets.

### 4.4.3. Selection for targets predicted to be druggable

A prospective drug target needs to have some indication that it is amenable to binding by small molecule compounds. The database incorporates 10 sources of data concerning protein druggability. The list of 465 genes was further filtered to remove genes if they were not predicted to be druggable by any of these sources. This filter reduced the list to 261 genes. Following this, genes encoding proteins that are already the target of existing drugs were also removed, reducing the list to 227 genes.

The list of 227 genes was then filtered for protein characteristics relevant to druggability. Genes were retained if their encoded protein is one of the following: a transmembrane protein, a secreted protein, or a protein kinase. These characteristics are common among the targets of existing drugs. This reduced the gene list to 61 genes.

### 4.4.4. Direct or implicit literature relationship to ‘metastatic melanoma’

The remaining 61 genes were ordered according to the strength of their literature relationships to ‘metastatic melanoma’, as inferred by the IRIDESCENT method. As described in the methods section, IRIDESCENT scores each gene for its direct literature relationship to ‘metastatic melanoma’ and gives two scores of their implicit relationship to ‘metastatic melanoma’. It was desired that the genes have a stronger explicit or implicit relationship to ‘metastatic melanoma’. First, genes were removed if they had no explicit or implicit relationship to ‘metastatic melanoma’. This reduced the gene list to 52 genes. Next, the upper quartile was calculated for the Min. MIM and Obs/Exp scores of these last 52 genes. If a gene had either Min. MIM and Obs/Exp score above the upper quartile, then it was retained. This reduced the list to 14 genes. A schematic showing the process of gene selection is outlined in Figure 4-10 below.
The 14 genes remaining at the end of this process were *AURKB, BNIP3, CCNB1, CD163, FANCA, FZD8, KCNG1, MERTK, MKI67, RETN, SCARB1, SLC11A1, TRPC3* and *YBX1*.

### 4.4.5. Corroboration of 14 putative drug targets using new data

The publication of newer, larger data-sets, and of additional methods and analyses for interrogating that data, have continually occurred in the time since this database was created. Distinct data and data-based methods were selected to retrospectively corroborate the result of the sequential filtering outlined in Figure 4-10. Data from sources not integrated into the database were used to verify the druggability, implicit relationship to melanoma, and RNA expression to patient survival association, for each of the 14 putative drug targets. The data sets used in this section may be integrated into MelanomaDB during its next update.

Clinical and mRNA expression data from the TCGA melanoma dataset were used to corroborate the association between the high RNA expression of the 14 putative drug targets and poor patient survival in melanoma patients. For each of the 14 putative drug targets, Cox proportional hazards regression models for censored survival time were fitted to mRNA
expression data for each of the genes encoding the 14 putative drug targets. Of these, those with high expression significantly associated with poor patient survival in the TCGA melanoma cohort were \textit{AURKB, CCNB1, MKI67, SCARB1} and \textit{YBX1}. The statistically significant association with patient survival in the TCGA dataset of 5 out of 14 genes is unlikely to be due to chance alone (hypergeometric test p-value = 0.0257). Interestingly, low expression of \textit{CD163} was associated with poor survival in this dataset.

The database CanSar\textsuperscript{164}, which was previously not available for inclusion into MelanomaDB, was used for cross-validation based on structural druggability and ligand-based druggability. The former was calculated using data from ChEMBL’s DrugEBility database, which was also used as the basis of druggability metrics incorporated into MelanomaDB. Therefore, only CanSar’s ligand-based druggability scores were used during this cross-validation. Ligand-based druggability is based on the drug-like properties of compounds that have been tested for binding a protein and/or its homologues. If a protein (or its homologues) bind drug-like compounds, then it is scored higher than a protein that only binds compounds with properties that are not drug-like. The targets 
\begin{math}
\text{AURKB, CCNB1, KCNG1, MERTK, MKI67, SCARB1} \text{ and } \text{TRPC3}
\end{math} were also predicted to be druggable based upon this metric.

The recently published Catalogue of Cancer Genes database\textsuperscript{215} is a manually curated amalgamation of data on known and putative cancer genes from 84 published papers and databases. This database connects \textit{CD163, FANCA, MKI67} and \textit{YBX1} to melanoma.

\textbf{4.4.6. Literature connecting cancer to 14 putative drug targets and selection of single target for further investigation}

Published literature on each of the 14 putative drug targets were briefly examined with the aim of selecting one target for further investigation.

\textit{AURKB} is a known target with elevated expression in numerous cancer types, including lung cancer\textsuperscript{216}, neuroblastoma\textsuperscript{217} and leukemia\textsuperscript{218}. Clinical trials on an \textit{AURKB} inhibitor have already advanced to phase II\textsuperscript{218-220}. While this confirmed that the data-driven method employed can select useful cancer drug targets, this project aimed to pursue a target that does not already have drugs under development.

\textit{BNIP3} is also a cancer gene involved in numerous cancer types, including breast\textsuperscript{221} and pancreatic\textsuperscript{222} cancer. However, in these cancers, low expression of \textit{BNIP3} correlated to poor prognosis. In melanoma, high \textit{BNIP3} expression has been linked to melanoma cell
migration\textsuperscript{223}. While inhibition of BNIP3 may be of therapeutic use in melanoma, it is unlikely to be of use in treatment of other cancer types. Ideally, development of an inhibitor drug should be applicable to multiple cancer types.

CCNB1 is a prognostic biomarker for breast cancer\textsuperscript{224}. CCNB1 over-expression in colorectal cancer promotes cell proliferation and tumour growth\textsuperscript{225,226}.

CD163 expression is associated with shorter survival times in breast cancer\textsuperscript{227} and rectal cancer\textsuperscript{228}, and the encoded protein has been proposed as a potential cancer drug target\textsuperscript{229}.

\textit{FANCA} is a breast cancer susceptibility gene\textsuperscript{230}. \textit{FANCA} variants have been associated with survival in melanoma patients\textsuperscript{231}. Patients with homozygous null-mutations in \textit{FANCA} have an increased susceptibility to leukaemia\textsuperscript{232}, suggesting that this target may not suitable for therapeutic inhibition. However, as \textit{FANCA} expression has been linked to resistance to chemotherapy drug cisplatin, Chirnomas \textit{et al.}\textsuperscript{233} developed a cell-based screen that identified \textit{FANCA} inhibitors.

\textit{FZD8} is a known cancer gene, and the encoded protein has already been suggested as a possible therapeutic drug target in colorectal\textsuperscript{234}, lung\textsuperscript{235}, breast\textsuperscript{236} and head and neck\textsuperscript{237} cancers. A monoclonal antibody, Vantictumab, is currently in Phase I trials in breast (NCT01973309), pancreatic (NCT02005315), and lung (NCT01957007) cancers.

There is no explicit association in the literature between \textit{KCNG1} and any type of cancer, making it an unlikely candidate for cancer drug development.

\textit{MERTK} has been proposed as a therapeutic target in gastric cancer\textsuperscript{238}, and head and neck cancer\textsuperscript{239}, and a recently developed small molecule inhibitor against MERTK has shown promise in pre-clinical trials in lung cancer\textsuperscript{240}.

MKI67 is a marker of proliferation, a predictive and prognosis marker in cancer and has been suggested as a therapeutic target\textsuperscript{241}.

Increased \textit{RTEN} expression is associated with poor patient survival in breast cancer\textsuperscript{242}, and elevated expression is found in colorectal cancer\textsuperscript{243}. The protein it encodes has been considered a drug target for diabetes\textsuperscript{244} as early as 2001\textsuperscript{245}, yet no small molecule inhibitor drugs have yet been reported.
SCARB1 is a potential biomarker for nasopharyngeal carcinoma\textsuperscript{246}, and SNP variants are associated with increased renal cell carcinoma\textsuperscript{247} and thyroid cancer\textsuperscript{248} risk. Knockdown of SCARB1 in prostate cancer reduces prostate cancer cell viability\textsuperscript{249}.

SLC11A1 may be linked to cancer via the immune system\textsuperscript{250}, but otherwise little association with cancer can be found in the literature.

In ovarian cancer, TRPC3 expression levels correlate with worse prognosis, early relapse and poor disease-free survival, and it has been suggested as a cancer drug target\textsuperscript{251}. A small molecular inhibitor of TRPC3 has been discovered by a group that suggested its use in treating cardiac hypertrophy\textsuperscript{252}.

Increased YBX1 expression has been observed in many cancer types and has been suggested as a therapeutic target\textsuperscript{253,254}. In a xenograft mouse model, transplanted cells transfected with an anti-YB-1 siRNA developed into tumours that grew more slowly than siRNA control-transfected cancer cells\textsuperscript{255,256}.

In summary, drugs are already under development for AURKB, FANCA, FZD8, MERTZ and TRPC3. The targets KCNG1 and SLC11A1 have only implicit, but not explicit links to cancer in the literature, making them risky targets for drug development. BNIP3 is implicated in cancer, but its decreased expression is associated with poor prognosis, making it an unsuitable target for drug inhibition. Despite long term interest in RTEN as a drug target, no small molecule inhibitors have yet been reported. This possibly suggests that it is an intractable target. This narrows the list of targets to CCNB1, CD163, MKI67, SCARB1 and YBX1.

From these 5 targets, YBX1, which encodes the YB-1 protein, was selected. Certain favourable features of YB-1 allowed it to pass the sequential filtering process, described in Figure 4-10, that was used to eliminate proteins not suitable as putative melanoma drug targets. Firstly, high RNA expression of YBX1 is associated with poor patient disease-free survival in melanoma according to analysis performed on data from Bogunovic et al\textsuperscript{153}. Secondly, YB-1 was predicted to be druggable by the DrugEBlility tractable score, and by Sophic’s BioLT-Drugbank. The DrugEBlility tractable score is based on the suitability of binding sites on protein domain structures for small molecules under Lipinski’s Rule of Five. The cold shock domain of YB-1 was assessed as suitable for small molecule binding. BioLT is a literature mining tool that identified suspected druggable genes by mining Pubmed for
gene names that co-occurred with drug names and the term “inhibit*”. Thirdly, YB-1 is not the target of any existing, clinically approved drugs. Fourthly, YB-1 was annotated as a secreted protein by Uniprot, based on the study of Frye et al. Finally, the IRIDESCENT method implied a connection between YBX1 and ‘metastatic melanoma’, with calculated Min. MIM and Obs/Exp scores both above the upper quartile of the scores of the remaining genes. This resulted in the inclusion of YB-1 in the final list of 14 putative melanoma drug targets.

Dr Annette Lasham has had a long term interest in this molecule, and her experience and expertise was considered to be invaluable to the development of YB-1 as a drug target. This knowledge gave YB-1 a considerable advantage over any of the other putative drug targets. It was at this point in this PhD project that Dr Lasham agreed to join as co-supervisor.

4.5. Introduction to YB-1

YB-1 is a multifunctional transcription factor that preferentially binds single stranded DNA (ssDNA), and also binds RNA and also double stranded DNA (dsDNA) but with much lower affinity, to regulate gene expression at both the transcriptional and translational levels\(^\text{257}\). While most transcription factors bind dsDNA\(^\text{258}\), short lengths of ssDNA occur in most transcriptionally active gene promoters in mammalian cells, and are bound by several protein families\(^\text{259}\), including Y-box transcription factors\(^\text{260-262}\), such as YB-1\(^\text{258}\). In addition to directly modulating transcription, YB-1 may alter DNA structure, destabilising dsDNA in gene promoter regions adjacent to ssDNA regions, thereby potentially inhibiting the binding of traditional dsDNA binding proteins\(^\text{254}\). YB-1 is also thought to be involved in the splicing, packaging and stabilisation of mRNA, as well as DNA replication and repair\(^\text{263}\). Through these diverse functions, YB-1 is thought to play a role in many of the cellular processes that are disturbed during cancer\(^\text{253}\).

4.5.1. Protein Structure

YB-1 consists of a short, 51-residue N-terminal domain (NTD), a 78-residue cold shock domain (CSD), and a large, 195-residue C-terminal domain (CTD)\(^\text{264}\). The CSD is evolutionarily conserved with homologues found across mammalian species like primates, rodents, rabbits, bats and cats\(^\text{263}\). While recently a prediction of the structure of YB-1 has been made\(^\text{263}\), only the CSD has had its structure determined using NMR, unlike the CTD or NTD\(^\text{265}\). The 3D structure of the NTD and CTD are still unknown, probably because they are disordered, have high intra-molecular mobility and lack a unique tertiary structure packing.
The CSD also has a low stability: in solution at 25°C, only 70% of the CSD molecules are in the native state\textsuperscript{265}. The lack of a rigid structure may enhance YB-1’s capacity to interact specifically with a variety of ligands\textsuperscript{254}.

The CTD is thought to have a role in the specific and non-specific binding to DNA and RNA, and has high affinity to nucleic acids\textsuperscript{263}. The CSD is purportedly responsible for specific binding with nucleic acids, while CTD and NTD are thought to stabilise this reaction\textsuperscript{254}. The isolated CSD binds DNA less tightly and with lower specificity than intact YB-1, indicating that the other domains have a role in DNA binding\textsuperscript{264}. Through binding nucleic acids, YB-1 has a role in most processes dependent upon DNA or mRNA, including transcription, translation, mRNA splicing and packing, and DNA replication and repair\textsuperscript{263}.

The full length YB-1 protein has a tendency to form oligomers. It has been proposed that intermolecular interactions between the charged regions of the CTD promote YB-1 oligomerisation\textsuperscript{264}. It is thought that greater numbers of multimers form when the YB-1 to mRNA ratio is high, and these multimers, which have been shown to comprise about twenty monomeric YB-1 subunits, are known to bind to mRNA predominantly with their NTDs\textsuperscript{266}. When the YB-1 to mRNA ratio is low, the protein binding to mRNA prevails over its self-association\textsuperscript{264}.

Two sections of the CTD are thought to regulate its distribution between cytosol and nucleus, the nuclear localisation signal (NLS) and the cytoplasmic retention site (CRS). The effect of the CRS prevails over that of the NLS, so YB-1 is typically in the cytoplasm. The CTD is also believed to be responsible for YB-1 secretion\textsuperscript{254}.

4.5.2. YB-1 in cancer

YB-1 is involved in many cancer types including tumours of the breast, ovary, intestine, lung, liver, prostate, skin and blood\textsuperscript{254,267}. In breast cancer, the highest expression levels of \textit{YBX1} RNA are found in the most aggressive and rapidly proliferating tumour subtypes\textsuperscript{256}. Overall cellular YB-1 level in breast cancer, which appears to be predominantly cytoplasmic, is a significant indicator of patient prognosis\textsuperscript{268-270}. Nuclear localisation of YB-1 is associated with drug resistance in breast cancer\textsuperscript{271,272} and with tumour progression and aggressiveness\textsuperscript{253}. Phosphorylation of YB-1 at certain sites may be required for YB-1’s translocation into the nucleus\textsuperscript{273}.
During melanoma progression, YB-1 has been shown to be up-regulated and translocated to the nucleus\(^2\), and YB-1 levels correlate with tumour progression\(^2\). In an *in vitro* study of metastatic melanoma cells, down-regulating YB-1 using shRNA reduced the rate of proliferation, decreased resistance to chemotherapeutic drugs, and increased apoptosis. The same study showed decreased migration and invasion in an *in vitro* skin reconstruct model\(^2\). Another study suggested that the melanoma inhibitory activity (MIA) protein, which is involved in the progression of melanoma, may exert its effects on melanoma through activation of YB-1\(^2\).

The ‘hallmarks of cancer’ proposed by Hanahan and Weinberg\(^1\) are capabilities acquired by tumours during their development. YB-1 plays a facilitating role across many of these hallmarks\(^2\), as described below.

### 4.5.2.1. Involvement of YB-1 in cellular proliferation through the E2F, MAPK and PI3K pathways

In tumour cells YB-1 appears to preferentially transactivate genes encoding proteins involved in cellular proliferation\(^2\), including cyclin and E2F family members\(^2\). In line with this function, the human YB-1 is more highly expressed in tumours with a high mitotic index\(^2\). YB-1 activates members of the MAPK pathway\(^2\), including *EGFR*\(^2\). In a positive feedback circuit, it has been proposed that YB-1 transcriptionally activates *PIK3CA* which in turn activates Akt which phosphorylates YB-1, causing the nuclear translocation of YB-1 and further activation of *PIK3CA* transcription\(^2\). In addition to this, YB-1 also appears to control the translation or mRNA stability of mTOR, a downstream element of the PI3K pathway\(^2\).

### 4.5.2.2. YB-1 represses RB and p53 and thus promotes genomic instability and the evasion of growth suppressors and cell-cycle checkpoints

YB-1 has been shown to reduce the tumour suppressive activities of the retinoblastoma (RB) protein by controlling the expression of RB’s upstream regulators cyclin D1\(^2\), CDK1, CDK2\(^2\) and by transcriptionally regulating several E2F S-phase genes, so that RB activity is insufficient to inhibit the activity of these E2F molecules\(^2\).

The tumour suppressor protein p53 is a transcription factor that regulates expression of numerous survival and proliferation genes, and can cause cell-cycle arrest following DNA damage\(^2\). YB-1 acts as a negative regulator of p53, reducing endogenous levels of p53 expression, both by repressing transcription at the *TP53* promoter\(^2\) and by directly interacting with p53 protein\(^2\). Furthermore, YB-1 regulates some of the same genes as
p53 but in an opposing manner. Through this repression of p53 and RB, YB-1 also plays a role in promoting replicative immortality and genomic instability. YB-1 overexpression in transgenic mice is also associated with genomic instability and provokes development of breast adenocarcinomas.285

4.5.2.3. YB-1 provides resistance to apoptosis

YB-1 inhibits expression of the pro-apoptotic BAX gene. Transfection of melanoma cells with ‘decoy’ YB-1 binding sites to sequester YB-1 from endogenous promoters leads to an increase in p53 and BAX levels, and an increase in p53-dependant apoptosis.282

YB-1 also inhibits the Fas-mediated apoptosis pathway by acting as a transcriptional repressor of the FAS promoter145, and may also repress transcription of CASP7.256

In addition, YB-1 in some contexts has been shown to promote MDRI gene expression which enhances resistance to chemotherapeutic agents.286

4.5.2.4. Involvement of YB-1 in invasion and metastasis

In vitro studies have shown that reducing YB-1 levels inhibits the invasive properties of cancer cell lines.287-289 YB-1 has been shown to promote the translation of SNAI1, LEF1 and TWIST1 which all repress the expression of the gene encoding E-cadherin CDH1.290-292 E-cadherin normally sequesters β-catenin at the cell membrane thereby repressing expression of EMT-inducing transcription factors. The CDH1 gene is frequently inactivated in metastatic cancers, and the indirect reduction in E-cadherin function due to high YB-1 activity provides an alternative pathway promoting tumour cell invasiveness.293

YB-1 also binds to the promoters of a number of Wnt pathway members.294 The Wnt pathway is involved in driving epithelial-mesenchymal transition, a process that can contribute to metastasis.295

Tumour cell migration is constrained by the extra-cellular matrix (ECM) and the basement membrane (BM).296,297 Proteolysis is one mechanism that dissociates the ECM and BM. The urokinase-type plasminogen activator system proteolysis pathway appears to be activated indirectly by YB-1, while other studies suggest that YB-1 transcriptionally regulates a number of matrix metalloproteinase genes.298-300
4.5.2.5. **YB-1 contributes to evasion of the immune response**

MHC class II and Fas/CD95 have been observed at lower levels in tumours\(^{301,302}\). These intrinsic immunity genes are transcriptionally repressed by YB-1\(^{145,303}\). YB-1 is also thought to promote the translation of extrinsic immunity gene \(TGFB1\) \(^{256,304}\).

### 4.5.3. Possible approaches to drug YB-1

For the reasons described above, as well as from the shortlist of putative drug targets identified by the database, YB-1 seems like a good therapeutic target. However, possible approaches to drugging over-expressed YB-1 in tumours include disrupting:

i) Potential activation of YB-1 by phosphorylation,

ii) YB-1 homo-multimerisation,

iii) YB-1 nuclear localisation or

iv) YB-1 nucleic acid binding\(^{253}\).

Phosphorylation of YB-1 at Ser102 is thought to be required for YB-1’s translocation into the nucleus\(^{273}\), where it can perform its functions as a transcriptional regulator. Therefore, inhibiting phosphorylation of YB-1 by upstream regulators AKT and RSK\(^{305}\) is a possible approach to modulating YB-1 activity.

In xenograft mouse models using lung carcinoma or breast cancer cells, cells transfected with an anti-YB-1 siRNA developed into tumours that grew more slowly than siRNA control-transfected cancer cells\(^{255,256}\). This suggests that siRNA may provide a therapeutic strategy against YB-1. However, in addition to the challenge of delivery, many siRNAs therapies still need to overcome problems with stability and bioavailability\(^{306}\), therefore siRNAs are not an attractive option for inhibiting YB-1 function in human tumours at present.

Previous attempts to modulate the activity of YB-1 have also included using ‘decoy’ ssDNA YB-1 binding sites to sequester the YB-1 protein\(^{307}\). As with siRNAs, advances in the delivery of nucleic acids as therapeutic agents may make this a viable therapeutic option in the future\(^{306}\), however effective delivery remains a significant barrier at present. A small molecule that could similarly bind YB-1 to interfere with the binding of YB-1 to nucleic acid, but without the delivery difficulties of the DNA ‘decoy’, presently is a more attractive strategy.
While YB-1 structures have been predicted, without NMR determined 3D structures of the NTD and CTD, it was not possible to take a rational, structure-based approach to drug design. Therefore, development of functional assays measuring YB-1 activity was required. Two assays were developed: a cell-based luciferase reporter gene assay, and a nucleic acid binding assay based on the AlphaScreen system. These assays are described in the following Chapters.
Chapter 5 Development of an *in vitro* cell-based assay for YB-1 transcriptional activity

5.1. Introduction to assay development

The multifunctional transcription factor YB-1 exerts many of its functions through the binding of nucleic acids. The 3D structure of YB-1 has not been resolved, and the mechanism of YB-1 binding to nucleic acid is not fully understood. This makes YB-1 a difficult target, demanding a multi-pronged approach that uses multiple assays that are able to examine different aspects of the binding of YB-1 to nucleic acids. Orthogonal assays are also an advantage during high throughput screening, in which two separate assay systems can complement each other. False positives that arise from small molecule interference in an assay system can be detected when they are submitted to a different assay system. Therefore, two complementary assays were developed to detect YB-1 binding to nucleic acid, and the inhibition of that binding by small molecules. Importantly, these two assays detect different aspects of YB-1 binding to nucleic acids. The first assay, described in this Chapter, detects activation of a YB-1-regulated double stranded promoter via a reporter gene, in a cell-based assay. The second assay, an AlphaScreen assay described in the next Chapter, detects purified YB-1 protein binding to a single stranded oligonucleotide in a plate format.

These assays were developed and optimised in Auckland and then transferred to the Chinese National Compound Library (CNCL) in Shanghai for further optimisation and use in high throughput screening. Transfections and some AlphaScreen optimisation experiments performed in Shanghai were completed with assistance from CNCL researcher Zhou Yan. Robotic dispensing of compounds was performed with Wu Ji and Zhou Yan. Associate Professor Yang Dehua contributed to discussions of the direction of assay development.

5.2. Assay overview

This *in vitro* assay is based on the binding of endogenous YB-1 to the promoter fragment of the human *E2F1* gene to activate transcription of a reporter gene in HCT116 colorectal cancer cells. pGL4.17-E2F1-728, a plasmid with 728bp of the human *E2F1* promoter driving transcription of a firefly luciferase reporter gene, was created for use in this assay. In addition to YB-1, the transcription factor E2F1 autonomously binds and increases the activity of the *E2F1* promoter. YB-1 and E2F1 are able to bind the 728bp fragment of human *E2F1*
promoter in the pGL4.17-E2F1-728 plasmid, which activates/facilitates the transcription of the luciferase gene. Increased transcription of this luciferase gene can result in a greater amount of luciferase protein, and when the cells are mixed with the appropriate substrate, the level of luciferase can be measured as a luminescent signal. In this Chapter, the term ‘luminescent signal’ refers to the amount of measured luminescence produced as a result of bioluminescent reactions catalysed by the activity of luciferase upon its substrate.

5.2.1. Modulation of transcription factor activity

During the experiments described below, the plasmid pCMV E2F1\textsuperscript{147} was used to over-express E2F1 and the plasmid pIRES2 eGFP YB1\textsuperscript{148} was used for YB-1 over-expression. A method for inhibiting YB-1 activity was also required as a control to which small molecule inhibition of YB-1 could be compared. Two methods were considered: use of siRNA to reduce YB-1 levels, and sequestering of YB-1 by oligonucleotides featuring YB-1 binding sites. Efficient transfection with siRNA requires specialised, siRNA transfection reagents to be used in addition to the lipofectamine used for plasmid transfection, while oligonucleotides can be co-transfected with plasmids, making inhibition of YB-1 by oligonucleotides more convenient, which avoids the stress on the cell that would occur as a result of a double transfection. YB-1 has previously been shown to strongly bind (Kd ~4nM) to a promoter fragment of the human γ-globin genes either in cells or cell free systems\textsuperscript{149,308}. This same sequence was used as a competitive inhibitor of YB-1 binding to nucleic acid and is referred to below as the 1xCohen competitor oligonucleotide.

5.3. Optimisation of transfection conditions for the assay

5.3.1. Cell number per well

It was desired that wells be seeded with enough cells to achieve confluence after 96 hours of growth following transfection, without reaching over-confluence at which cell death is apparent by inspection by microscope. The suitable cell number for seeding 24-well plates was found to be 22,000 cells per well for HCT116 cells, and 7,500 cells per well for A375 cells.
5.3.2. Kill curve assay

The pGL4.17, pGL4.17-E2F1-728, pIRES2 eGFP and pIRES2 eGFP YB1 plasmids contain a gene encoding a protein that provides host cells with resistance to the G418 antibiotic. Transfected cells are resistant to G418, so the antibiotic was used to select for transfected cells by killing wild-type cells. A kill curve assay determined that 1.25mg/mL is the minimum concentration of G418 that kills all wild-type A375 and HCT116 within 96 hours. Data from HCT116 cells are plotted in Figure 5.1. Cellular confluence was assessed by visual inspection. Note that no live cells were found at 96 hours after exposure to G418 concentrations equal to or greater than 1.25mg/mL.

![Figure 5-1. Kill curve for HCT116 cells exposed to a range of G418 concentrations. Cellular confluence was assessed visually at 96 hours after addition of G418 to culture media. Error bars represent the standard error of observations over three replicate wells. Live cells were not found in any wells exposed to G418 concentrations equal to or greater than 1.25mg/mL.](image-url)
5.3.3. Selection of cell type, transfection method and lipofectamine reagent

A greater luminescent signal with greater dynamic range would be achieved if more cells were successfully transfected with the pGL4.17-E2F1-728 plasmid that encodes luciferase; therefore, high transfection efficiencies were desired. A375 and HCT116 cells were selected for experimentation due to their ease of transfection, and the experience in their use within our lab. Additionally, the CNCL lab in Shanghai had experience using the HCT116 cell line. In order to determine the conditions in which transfection efficiency was highest, experiments were conducted comparing cell type, transfection method and lipofectamine reagent.

Transfection efficiency was much higher for HCT116 cells than A375 cells. Reverse transfection efficiencies appeared higher than those in forward transfections. Transfection efficiencies were higher when using LF3000 than LF2000; in addition, cells appeared more healthy and regular using LF3000 than when using LF2000. A375 cells transfected using 0.75µL of LF3000 had higher transfection efficiencies than those transfected using 1.5µL. In HCT116 cells, transfection efficiencies were similar at the two concentrations of LF3000 (1.5µL and 0.75µL), however cells appeared healthier in wells using 0.75µL. Transfection efficiencies under these different conditions are shown in Table 5-1 below.

Table 5-1 Transfection efficiencies for HCT116 and A375 cells under different transfection methods, transfection reagents and transfection reagent volumes.

<table>
<thead>
<tr>
<th>Transfection method</th>
<th>HCT116</th>
<th>A375</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverse Transfection</td>
<td>30% to 90%</td>
<td>5% to 20%</td>
</tr>
<tr>
<td>Forward transfection</td>
<td>10% to 70%</td>
<td>5% to 10%</td>
</tr>
<tr>
<td>Transfection reagent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipofectamine 3000 (1.5µL)</td>
<td>50% to 90%</td>
<td>10% to 20%</td>
</tr>
<tr>
<td>Lipofectamine 3000 (0.75µL)</td>
<td>50% to 90%</td>
<td>5% to 10%</td>
</tr>
<tr>
<td>Lipofectamine 3000 (0.375µL)</td>
<td>10% to 40%</td>
<td>Not tested</td>
</tr>
<tr>
<td>Lipofectamine 2000 (4µL)</td>
<td>10% to 60%</td>
<td>5% to 10%</td>
</tr>
<tr>
<td>Lipofectamine 2000 (3µL)</td>
<td>10% to 50%</td>
<td>5% to 10%</td>
</tr>
</tbody>
</table>

Cells successfully transfected with the pIRES2-GFP plasmid express Green Fluorescent Protein. The cell populations were visually assessed via microscope, and estimations were made of the percentage of GFP-expressing cells.
Observations by microscope determined that cells transfected with 0.75µg of pIRES2-GFP appeared healthier than wells transfected with 1µg, with transfection efficiency only slightly reduced. Therefore, the optimised procedure used subsequently was reverse transfection of HCT116 with a total plasmid amount of 0.75µg using 0.75µL of LF3000 per 24-well plate well. While using A375 cells would have provided a melanoma context to this assay, the aim of the assay was to detect inhibition of YB-1 activity via the $E2F1$ promoter. The higher transfection efficiency of HCT116 cells was considered a greater advantage for that purpose.

5.4. Generation of HCT116 stably transfected with pGL4.17 or pGL4.17-E2F1-728

Stably transfected HCT116 lines were desired to facilitate future experiments by reducing the number of plasmids required during transient transfections. Furthermore, using the same stably transfected cell line can reduce variation between experiments. Several stable transfected clonal lines of HCT116, containing either pGL4.17 or pGL4.17-E2F1-728, were developed as described in the Methods section.

Seven wells contained a single cell transfected with pGL4.17 and 10 wells contained a single cell transfected with pGL4.17-E2F1-728. These cells were maintained in media as previously described in order to expand these cell lines. Cell lines stably transfected with the pGL4.17-E2F1-728 plasmid should produce luciferase, indicating that the luciferase gene has been incorporated into the host genome and is still being transcribed. Stable lines transfected with the promoter-less pGL4.17 were expected to not express luciferase. A sample of each cell line was lysed so that luminescent signal could be measured. Unexpectedly, two out of seven of the pGL4.17 stable lines displayed luciferase activity, with luminescent signals of 23 and 15 times the background signal. This may be due to plasmid insertion near an active host promoter. Only two out of ten of the pGL4.17-E2F1-728 stable cell lines displayed luciferase activity, with luminescent signals of 212 and 19 times the background signal. While the stable lines expressing background levels of luciferase still maintained resistance to G418 (implying retention and expression of the NeoR gene on the pGL4.17 or pGL4.17-E2F1-728 plasmids), their luciferase expression, driven by the $E2F1$ promoter in pGL4.17-E2F1-728, may have been silenced by the host cell. Only the two cell lines stably transfected with pGL4.17-E2F1-728 and expressing high levels of luciferase were retained.
Next, the response of these two pGL4.17-E2F1-728 stable lines to over-expression of YB-1 was determined by transfection with pIRES2 eGFP YB1. Over-expression of YB-1 in the stable line resulted in an increase in luminescent signal that was smaller than that resulting from transiently transfecting wild-type cells with pIRES2 eGFP YB1 in addition to pGL4.17-E2F1-728. Additionally, the stable transfected cell lines exhibited less overall luciferase than transiently transfected HCT116 cells. It may be that transiently transfected wild-type cells contain more copies of the luciferase producing pGL4.17-E2F1-728 plasmid per cell than stably transfected cells, and are thus produce a greater increase in luciferase in response to YB-1 overexpression. Subsequently, the use of the stably-transfected cell lines was abandoned in favour of using transient transfection for later experiments, which was feasible because of the high transfection efficiency of the HCT116 cells.

5.5. Overexpression of YB-1 or E2F1 activates transcription driven by the 728bp E2F1 promoter sequence

5.5.1. Optimisation of reporter gene assays

In order to observe the effects of YB-1 and E2F1 over-expression on the 728bp E2F1 promoter driving luciferase expression, three experiments were performed by transfecting HCT116 cells with the combinations of plasmids and oligonucleotide as shown in Table 5-2 (below). Cells were transfected with the E2F1 promoter:luciferase reporter plasmid pGL4.17-E2F1-728 and either a control plasmid (pIRES2 eGFP) or plasmids over-expressing either human E2F1 (pCMV E2F1) or human YB-1 (pIRES2 eGFP YB1). As a control for YB-1 inhibition, 1xCohen oligonucleotide was used to sequester YB-1 in cells with or without YB-1 over-expression.

Table 5-2 Different combinations of plasmid and oligonucleotide used in 24-well plates

<table>
<thead>
<tr>
<th>Amount of each plasmid used for transfection in 24-well plates (ng/well)</th>
<th>pGL4.17</th>
<th>pGL4.17-E2F1-728</th>
<th>pIRES2 eGFP</th>
<th>pCMV E2F1</th>
<th>pIRES2 eGFP YB1</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>80</td>
</tr>
</tbody>
</table>
The amount of luminescence produced by these conditions is plotted in Figure 5-2 below:

Figure 5-2 The E2F1 promoter: luciferase reporter gene activity in the presence of variable levels of E2F1 and YB-1, across 3 experiments.

The level of luminescent signal was increased by E2F1 or YB-1 over-expression, and reduced by 1xCohen oligonucleotide. Note that the 1xCohen oligonucleotide was used only in the 2nd and 3rd experiments.

These results suggest that endogenous transcription factors, including YB-1, act upon the E2F1 promoter fragment in pGL4.17-E2F1-728 to drive expression of the luciferase reporter gene. Over-expression of E2F1 increased luminescent signal 8.6-fold in the first experiment, 4.8-fold in the second experiment, and 4.6-fold in the third experiment over the signals measured from cells transfected with the pGL4.17-E2F1-728 plasmid alone. Over-expression of YB-1 increased luminescent signal 2.1-fold, 2.3-fold and 2.1-fold across the three experiments. In cells with YB-1 over-expression, addition of the 1xCohen competitor
oligonucleotide to cells over-expressing YB-1 reduced the luminescent signal 1.7-fold in the second experiment, and 1.1-fold in the third experiment, probably through inhibition of YB-1.

As over-expression of YB-1 or E2F1 is thought to increase cell proliferation\textsuperscript{256}, and thus could increase luminescent signal directly (through activating the 728bp \textit{E2F1} promoter in pGL4.17-E2F1-728) or indirectly (by increasing cell number), it was decided that a control for cell number was required.

5.5.1.1. \textbf{YB-1 activates expression from the thymidine kinase promoter}

A constitutively active promoter driving a reporter gene can generate a luminescent signal proportional to the number of transfected cells. A signal proportional to the number of transfected cells can be used to adjust other signals from the same experiment to control for cell number. The pRL-TK plasmid contains a thymidine kinase (TK) promoter adjacent to a discrete luciferase derived from \textit{Renilla reniformis}. This distinct luciferase acts upon a different substrate than the firefly luciferase expressed by the pGL4.17-E2F1-728 plasmid. pRL-TK constitutively expresses this Renilla luciferase so that signal is proportional to the number of transfected cells, and thus acts as an internal control for cell number. Cells in two of the experiments described above were also co-transfected with the pRL-TK plasmid. Results from those two experiments are portrayed in Figure 5-3 below.
Figure 5-3 Results from two experiments showing that the Renilla luciferase luminescent signal, driven by the TK promoter, is increased in the presence of YB-1 over-expression, and decreased in the presence of 1xCohen competitor oligonucleotide, that inhibits YB-1 activity

YB-1 over-expression was expected to further activate expression driven by the E2F1 promoter fragment, resulting in a higher firefly luciferase luminescent signal, but not affect the Renilla luciferase luminescent signal. YB-1 over-expression also increases cell number, resulting in both higher firefly luciferase and Renilla luciferase luminescent signals. Unexpectedly, it was observed that YB-1 over-expression resulted in a greater fold increase in Renilla luciferase luminescent signal (a 3-fold increase in the second experiment, and a 2.6-fold increase in the third experiment) than in firefly luciferase luminescent signal (a 2.3-fold increase in the second experiment, and a 2.1-fold increase in the third experiment). E2F1 over-expression had no effect on TK promoter activity. An equivalent Renilla and firefly luciferase luminescent signal fold increase would imply the difference in signal resulted from an influence on cell number, or an equivalent direct influence on both the firefly luciferase promoter (E2F1-728) and the Renilla luciferase promoter (TK). To test whether YB-1 over-
expression was increasing TK promoter activity, one of the experiments described above measured both firefly luciferase and Renilla luciferase luminescent signal, and also measured total protein following cell lysis. Total protein is a rough indicator of cell number that is independent to Renilla luciferase luminescent signal.

After adjusting for total protein, YB-1 over-expression resulted in a 2-fold increase in Renilla luciferase luminescent signal. Consistent with this, transfection with 1xCohen competitor decreased Renilla luciferase luminescent signal 1.5-fold after adjusting for total protein. Ladomery and Sommerville\textsuperscript{309} infer from the presence of a Y-box sequence in the human thymidine kinase promoter, that YB-1 regulates TK expression. These experimental results support that inference. This suggests that Renilla luciferase luminescent signal driven by the TK promoter is not a suitable indicator of cell number when cells have varying levels of YB-1.

5.5.1.2. Simultaneous over-expression of E2F1 and YB-1 synergistically activates the \textit{E2F1} promoter

In the experiments described above, it was observed that, based on luciferase reporter gene levels, over-expression of E2F1 resulted in a 4.6 to 8.6-fold increase in activity of the \textit{E2F1} promoter, while over-expression of YB-1 resulted in a 2.1 to 2.3-fold increase in activity of the \textit{E2F1} promoter. In order to determine whether E2F1 and YB1 act synergistically to activate the \textit{E2F1} promoter, a series of transfections were performed, as shown in Table 5-3 below. The results from this experiment are shown in Figure 5-4.

\textbf{Table 5-3 Transfection conditions used in 96-well plate wells to test synergistic activation of \textit{E2F1} promoter by simultaneous E2F1 and YB-1 over-expression}

<table>
<thead>
<tr>
<th>Amount of each plasmid used for transfection in 96-well plates (ng/well)</th>
<th>pGL4.17-E2F1-728</th>
<th>pIRES2 eGFP</th>
<th>pCMV E2F1</th>
<th>pIRES2 eGFP YB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>pIRES2 eGFP</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>pIRES2 eGFP YB1</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>
Figure 5-4 Synergistic activation of the $E2F1$ promoter by E2F1 and YB-1.

The $E2F1$ promoter: luciferase reporter construct was co-transfected in cells that over-express YB-1 or E2F1 or both.

The effect of YB-1 over-expression on the activity of the $E2F1$ promoter results in a 1.5-fold increase in luminescent signal about the luminescent signal observed in cells transfected with pGL4.17-E2F1-728 alone, whilst E2F1 over-expression increased luminescent signal 4.7-fold. The sum of these is still less than the 8.7-fold increase in signal seen when both YB-1 and E2F1 were over-expressed. This suggests that activation of this $E2F1$ promoter fragment by E2F1 and YB-1 is not simply additive, but that these two transcription factors act synergistically to drive expression from this promoter.

5.5.2. Optimisation of reporter gene assay conditions for use in HTS

During high throughput screening, two controls are required that represent signal that is not inhibited, and signal that is completely inhibited. There should be minimal variation within these controls, and a large difference in signal between them. Experiments were performed to determine the conditions that would maximise the signal from the not inhibited control,
which are the cells transfected without addition of the 1xCohen competitor oligonucleotide. The completely inhibited control would be based on inhibition of YB-1 activity by the 1xCohen competitor. In order to prepare the assay for the stringent requirements of HTS, assay conditions were optimised to:

i) Reproducibility between experiments

ii) Increase the luminescent signal ratio between not inhibited and completely inhibited controls

iii) Reduce signal variability within control groups

Experimental conditions that were tested during optimisation included: quantities of plasmid used in transfections; co-transfection with different plasmid combinations; forward versus reverse transfection methods; cell concentrations; plate types and formats; and evaluation of timing between transfection, the addition of compounds and measuring luciferase activity.

5.5.2.1. Optimisation of synergistic activation of the E2F1 promoter by E2F1 and YB1

In order to follow up on the results that suggested that YB-1 and E2F1 may act synergistically to activate the E2F1 promoter, the next experiment evaluated the effect of varying amounts of E2F1 and YB1 on the activity of the E2F1 promoter:luciferase reporter construct. HCT116 cells were transfected with both pIRES2 eGFP YB1 and pCMV E2F1 plasmids, at molar ratios ranging from 10:1 to 1:10. To create an empty vector control, pIRES2 eGFP (that is, without YB-1) was also transfected to control for the equivalent amount of DNA being transfected per condition. Transfections using a range of molar ratios of YB-1 and E2F1 over-expressing plasmids resulted in average luminescent signals plotted in Figure 5-5.
Figure 5-5 Average luminescent signal for transfections using a range of ratios of YB-1 to E2F1 over-expressing plasmids.

HCT116 cells were transfected with both pIRES2 eGFP YB1 and pCMV E2F1 plasmids, at molar ratios ranging from 10:1 to 1:10. The control on the left is an empty vector control transfected with pIRES2 eGFP that does not express YB-1 or E2F1.

The results of this experiment confirm that YB1 and E2F1 act synergistically to drive expression from the E2F1 promoter. This conclusion was based on the results which showed that the sum of the level of activation of the E2F1 promoter by YB-1 and by E2F1 was less than the level of activation of the promoter observed following simultaneous overexpression of both YB-1 and E2F1. While a YB-1:E2F1 ratio of approximately 1:1 yielded the highest luminescent signal, any ratio from 5:1 to 1:5 resulted in luciferase activity driven by the E2F1 promoter to be higher than was observed when either YB-1 or E2F1 were individually overexpressed.

5.5.2.2. Assay development in HTS laboratory

At this point, further optimisation of this assay was performed in the HTS lab in Shanghai.
One objective of assay development was to create an assay sensitive to inhibition of YB-1. At this point in development, two contradictory speculations were made. The first was that for the assay to be focused on YB-1 inhibition, YB-1 must be abundant and the dominant activator of the \( E2F1 \) promoter, and therefore the dominant driver of luciferase reporter activity. Alternatively, it was speculated that YB-1 should be at a low level, in order to be the rate limiting step of synergistic activation of the \( E2F1 \) promoter. To get further insight into these ideas, an experiment was performed in which HCT116 cells were transfected with a high and low ratio of YB-1 to E2F1 over-expressing plasmids. Performed in parallel were a set of transfections that also included 20pmol of 1xCohen competitor per well. Equivalent amounts of DNA were transfected between all conditions. Inclusion of the 1xCohen competitor decreased luminescent signal 1.3-fold in cells only over-expressing YB-1 compared to the parallel cells (over-expressing YB-1) in which the 1xCohen competitor was absent, 1.2 fold in cells with a high ratio of YB1 to E2F1 over-expression, 1.6 fold in cells with a low ratio of YB1 to E2F1 over-expression, and 1.6-fold in cells only over-expressing E2F1. In the control well, which contained no YB-1 or E2F1 expression, the 1xCohen competitor decreased luminescent signal 1.9-fold.

While the effect of the 1xCohen competitor was modest in this experiment, a pattern can be observed. Wells with low or no YB-1 over-expression show greater fold decreases in signal when 1xCohen competitor was included. This suggests that the assay is more sensitive to YB-1 inhibition when lower amounts of YB-1 are over-expressed.

5.5.2.3. Determining competitor oligonucleotide concentration

In order to determine the concentration of 1xCohen competitor oligonucleotide required to effect maximum signal inhibition, 1xCohen competitor oligonucleotide amounts ranging from 0 to 80pmol per well were applied to transfected wells of these three conditions described in the experiment above: Control, High YB-1:E2F1, Low YB-1:E2F1. The average luminescent signals from each of these three conditions are plotted in Figure 5-6.
This result shows that the effect of 1xCohen competitor is dose-dependent. In this experiment, 1xCohen competitor similarly inhibits activation of the E2F1 promoter in the presence of both high and low ratios of YB-1:E2F1 over-expression. Using a greater amount of 1xCohen oligonucleotide may be desirable to maximise the ratio of luminescent signal between wells without and with inhibition of YB-1 activity. Conversely, using less competitor oligonucleotide was desirable in order to avoid potential off-target effects. Unexpectedly, the highest luminescent signal can be observed in the control wells which do not have over-expressed YB-1 or E2F1. This may be because YB-1 and E2F1 over-expression results in increased rates of cellular proliferation, which dilutes the E2F1 promoter: luciferase reporter construct resulting in fewer plasmids per cell. Alternatively, it may be that the initial amount of pGL4.17-E2F1-728 plasmid are too high, resulting in high expression of luciferase, regardless of the levels of YB-1 and E2F1.

A later experiment tested the effect of inhibition with 50 and 70pmol of 1xCohen competitor oligonucleotide per well with 80 or 100ng of pGL4.17-E2F1-728 plasmid per well. The
motivation for this experiment was to increase luminescent signal by producing more luciferase while reducing the amount of YB-1 expressed, theoretically making the assay more sensitive to inhibition of YB-1. The conditions tested are tabulated in Table 5-4 below with their average signals and ratios of the not inhibited/inhibited cells (i.e. signals of wells without 1xCohen competitor divided by signal of wells with 1xCohen competitor).

Table 5-4 Average luminescent signals from cells transfected with 80 or 100ng of pGL4.17-E2F1-728 plasmid per well, in the presence or absence of the 1xCohen competitor.

The grid cell colour represents the strength of the signal, with higher signals coloured with darker shades of red, and lower signals coloured with shades approaching white. Given also is the ratio of luminescent signal from cells that are not inhibited divided by signal from inhibited cells.

<table>
<thead>
<tr>
<th>Plasmid amounts added per well (ng)</th>
<th>Signal (not inhibited)</th>
<th>Signal (50pmol competitor)</th>
<th>Signal ratio (not inhibited/50 pmol competitor)</th>
<th>Signal (70pmol competitor)</th>
<th>Signal ratio (not inhibited/70 pmol competitor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGL4.17-E2F1-728</td>
<td>pcDNA 3.1</td>
<td>pIRE2/eGFP</td>
<td>YB1</td>
<td>pCMV E2F1</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

In this experiment, similar inhibition of YB-1 activity is effected by 50pmol and 70pmol of 1xCohen competitor. Luminescent signals are higher from cells transfected with higher concentrations of pGL4.17-E2F1-728, but show similar ratios of signals between cells that are not inhibited and those that are inhibited. The control wells, without YB-1 or E2F1 over-expression, yield signal higher than wells containing YB-1 over-expression alone, but not higher than wells containing YB-1 and E2F1 over-expression together. This variation between experiments may be due to YB-1 overexpression having a complex and unpredictable effect on the cell. Wells containing both YB-1 and E2F1 over-expression produce an excessively high signal that is only slightly inhibited by the 1xCohen competitor. Additionally, the highest not inhibited/inhibited signal ratios (6.49:1, 5.31:1, 4.75:1, 3.59:1) were all achieved by the control wells. As these ratios are sufficiently high for high-throughput screening, it was hypothesised at this point that a simpler transfection of only
pGL4.17-E2F1-728, without overexpression of YB-1 or E2F1, may be less variable between experiments and therefore an appropriate option for HTS.

5.5.2.4. **Determining the times between transfection, addition of compounds and measuring luciferase activity**

The quality of an HTS assay is determined both by its signal to background ratio, and the variation between replicates. As signal ratios were sufficiently high in the previous experiments, ranging between 2.42:1 to 6.49:1, the focus of assay development moved to reducing the variation in luminescent signals between replicate wells. In previous experiments, cells were reverse transfected, however it was thought that forward transfection, in which the lifted cells are given time to recover and adhere to a new dish before transfection, would yield healthier cells, and thus less variation due to variable cell health between replicate wells. It was also theorised that if less time is allowed to elapse between transfection and measurement of luciferase, then fewer stochastic cellular events can occur to increase differences between replicate wells. Also, less time for proliferation reduces the effect of variation in the number of cells in each replicate. To test this theory, cells were forward transfected with either pGL4.17-E2F1-728 alone, or in combination with competitor at a concentration equivalent to 50pmol per well. Then luciferase activity was measured at three time points: 10 hours, 24 hours and 48 hours after transfection. Two conditions were tested in parallel; transfection with 50 or 80ng of pGL4.17-E2F1-728 per well. The average and standard deviation at each time point are given in Table 5-5.

Table 5-5 Average and standard deviation of luminescent signal at different times between transfections using 50 or 80ng of pGL4.17-E2F1-728 per well, with and without inhibition by 50pmol of 1xCohen competitor per well.

The grid cell colour represents size of the standard deviation with higher standard deviations coloured with darker shades of red, and lower standard deviations coloured with shades approaching white.
These results showed that the standard deviation generally increased with increased time between transfection and luciferase measurement. The luminescent signal was highest at 24 hours after transfection. Similar results were seen when this experiment was repeated. This experiment additionally tested whether reduction of the percentage of FBS in the growth media from 5% to 2% would reduce variability between replicate wells by slowing cell proliferation rates. The results of this showed, however, that the variation between replicate wells was similar regardless of FBS concentration.

The next experiment investigated how the time of addition of the compound and DMSO influenced the variation between replicate wells. Cells were forward transfected with pGL4.17-E2F1-728 in the absence or presence of 1xCohen competitor. DMSO was then added at either 8, 20 or 24 hours after transfection and luciferase activity was measured at 24, 36 or 48 hours after transfection. The results showed that the inhibition of luciferase activity by the 1xCohen competitor was similar, regardless of the time between transfection and addition of DMSO or measurement of luciferase activity. The mean and standard deviations for wells that were transfected without the 1xCohen competitor are plotted below in Figures 5-7 and 5-8:
Figure 5-7  Luminescent signal decreases with increasing time between transfection and luciferase activity measurement, regardless of the time of DMSO addition. HCT116 cells were forward transfected with the pGL4.17-E2F1-728 plasmid, with DMSO added at 8, 20 or 24 hours after transfection. Cells were lysed, luciferase substrate added, and the resulting luminescent signal measured at 24, 36 or 48 hours after transfection (x-axis). Each y-axis value is the mean luminescent signal of 14 replicate wells for each experimental group. The results from this experiment suggest that the luminescent signal decreases with time, but is not influenced by the time at which DMSO is applied.
Figure 5-8 The standard deviation of the luminescent signal decreases with increasing time between the transfection and the measurement of luciferase, and is influenced by the time at which DMSO is added to the cells.

HCT116 cells were forward transfected with the pGL4.17-E2F1-728 plasmid, with DMSO added at 8, 20 or 24 hours after transfection. Cells were lysed, luciferase substrate added, and the resulting luminescent signal measured at 24, 36 or 48 hours after transfection (x-axis). Each y-axis value is the standard deviation of the luminescent signal of 14 replicate wells for each experimental group.

This data suggests that standard deviation of luminescent signal decreases with time. In this experiment, deviation appears to reach a minimum of around 600 luminescent units. That minimum is reached earliest, at 36 hours, when DMSO is added 8 hours after transfection. Plotting standard deviation against time between addition of DMSO and luciferase measurement is more revealing, as shown in Figure 5-9:
Figure 5-9 Luminescent signal standard deviation decreases with increasing time between DMSO addition and measurement of luciferase.

HCT116 cells were forward transfected with the pGL4.17-E2F1-728 plasmid, with DMSO added at 8, 20 or 24 hours after transfection. Cells were lysed, luciferase substrate added, and the resulting luminescent signal measured at 24, 36 or 48 hours after transfection. The x-axis displays the hours between the addition of DMSO and the measurement of luminescent signal. Each y-axis value is the standard deviation of the luminescent signal of 14 replicate wells for each experimental group.

This data shows that variation decreases with greater time between DMSO addition and luciferase measurement. However, the average luminescent signal is higher the earlier luciferase is measured. Therefore, a balance must be found between the decreasing signal and the decreasing variation. Clearly, adding DMSO earlier allows more time between that and the measurement of luciferase; it is more difficult to decide the time of luciferase measurement. For each data point, the standard deviation was divided by the average luminescent signal to give the coefficient of variation. The lowest coefficient of variation was
when DMSO was added at 8 hours after transfection, and luciferase measured at 36 hours after transfection.

Additionally, 3 compounds, identified as inhibitors of the E2F1 promoter: luciferase reporter assay by pilot screening, were added to wells in this experiment to establish the time required for a potential inhibiting compound to influence luciferase activity. Compounds were added at 8, 20 and 24 hours after transfection and luciferase activity was measured at 24, 36 and 48 hours after transfection. As a control, DMSO only was tested in parallel. The compounds inhibited the luminescent signal to a level equal or lower than the inhibition displayed by 1xCohen competitor in all instances except for two:

1) There was no inhibition observed when the compound was added to the cells at 24 hours after transfection, and luciferase activity measured immediately.

2) When the compound was added to the cells 20 hours after transfection, and luciferase activity measured at 24 hours after transfection, the compound was found to have inhibited the luminescent signal by 31% of the inhibition effected by the 1xCohen competitor.

From this experiment it was concluded that the compounds should be added to wells at least 4 hours before the luciferase activity is measured. The decrease in luminescent signal following addition of compounds at different times after transfection is plotted in Figure 5-10 below.
Figure 5-10 Within four hours of the addition of compounds, luminescent signal is decreased. No decrease in luminescent signal is seen when luminescent signal is measured immediately after compounds are added at 24 hours.

HCT116 cells were forward transfected with the pGL4.17-E2F1-728 plasmid, with DMSO or compounds added at 8, 20 or 24 hours after transfection. Cells were lysed, luciferase substrate added, and the resulting luminescent signal measured at 24, 36 or 48 hours after transfection. The x-axis displays the hours between the addition of DMSO or compounds and the measurement of luminescent signal. Each y-axis value is the standard deviation of the luminescent signal of 14 replicate wells for each DMSO experimental group, and 3 replicates for each compound experimental group.

5.5.2.5. Determination of the reaction volume of each well of transfected cells

In order to determine the effects of the well reaction volume on the luminescent signal and its standard deviation, an experiment was performed using three different reaction volumes. A dish of cells was transfected with pGL4.17-E2F1-728 and 8,000 cells were dispensed into each well of a 384-well plate in media volumes of 20, 30 and 40μL. The results of this experiment are given in Table 5-6.
Table 5-6 Experiment using different volumes of cell culture media in 384-plate wells shows that reaction volume does not have a strong effect on the variation of luminescent signal. Average and standard deviations of luminescent signal are shown for 13 replicate wells for each reaction volume.

<table>
<thead>
<tr>
<th>Reaction volume</th>
<th>Average luminescent signal</th>
<th>Standard deviation</th>
<th>Coefficient of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>20μL</td>
<td>78404</td>
<td>12500.9</td>
<td>0.159442</td>
</tr>
<tr>
<td>30μL</td>
<td>55715</td>
<td>8386.3</td>
<td>0.150521</td>
</tr>
<tr>
<td>40μL</td>
<td>61929</td>
<td>13748</td>
<td>0.221996</td>
</tr>
</tbody>
</table>

The coefficients of variation are similar in this experiment, that is, the standard deviation is proportional to the average luminescent signal. This experiment suggested that reaction volume does not have a strong influence on variation. Previous experiments used 30μL reaction volumes, and this was continued for consistency.

5.5.2.6. Determination of initial cell number

Previous experiments had seeded 384-well plate wells with 8,000 cells per well. In order to investigate whether altering this initial cell number influences the variation of the luminescent signal, well volumes of 30μL were maintained but a range of seeded cell numbers from 4,000 to 12,000 cells per well were tested. The results of this experiment are given in Table 5-7.

Table 5-7 Coefficient of variation of luminescent signal decreases with increasing numbers of cells seeded per well.

The grid cell colour represents the size of the coefficient of variation, with higher coefficient of variation coloured with darker shades of red, and lower coefficient of variation coloured with shades approaching white. 14 replicate wells were measured for each starting cell number.

<table>
<thead>
<tr>
<th>Starting cell number</th>
<th>Average luminescent signal</th>
<th>Standard deviation</th>
<th>Coefficient of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4000</td>
<td>31557</td>
<td>6631.6</td>
<td>0.21015</td>
</tr>
<tr>
<td>6000</td>
<td>42065</td>
<td>8390.5</td>
<td>0.19947</td>
</tr>
<tr>
<td>8000</td>
<td>55715</td>
<td>8386.3</td>
<td>0.15052</td>
</tr>
<tr>
<td>10000</td>
<td>63949</td>
<td>9913</td>
<td>0.15501</td>
</tr>
<tr>
<td>12000</td>
<td>95041</td>
<td>9841</td>
<td>0.10355</td>
</tr>
</tbody>
</table>
This experiment suggested that increasing the cell number to 12,000 cells per well may increase assay quality. However, higher numbers of cells may be less sensitive to inhibition by the same amount of compound. Therefore, the use of 8000 cells per well was continued into screening, consistent with previous experiments.

5.6. Summary

This Chapter described development of an in vitro cell-based reporter gene assay responsive to the activity of YB-1. HCT116 cells, transiently transfected using Lipofectamine 3000, were selected for use over A375 cells due to their higher transfection efficiency. Transiently transfected HCT116 cells yielded higher luminescent signals than stably transfected HCT116 cells. The experimental conditions used for this in vitro cell-based assay during high-throughput screening are described in detail in Section 2.3.3.4.2. The conditions were optimised with the aims of increasing the luminescent signal ratio between control groups (signal to background ratio), and reducing the signal variation (i.e. standard deviation) between replicates of the control groups. Sufficiently high ratios were achieved first (Section 5.5.2.3), using 50pmol of 1xCohen competitor per replicate well of the fully inhibited controls. Later experiments focused on reducing variation within control groups. Variation was lowest during forward transfections of only pGL4.17-E2F1-728, without co-transfection of plasmid over-expressing E2F1 or YB-1. Variation increased with greater time between addition of DMSO and measurement of luciferase, and was lowest when DMSO was added at 8 hours after transfection, and luciferase measured at 36 hours after transfection. Reaction volume did not have a strong influence on variation within control groups. This series of optimising experiments increased the quality of the assay to a level sufficient for the high-throughput screening described in Chapter 7.
Chapter 6 Development of a novel AlphaScreen assay to detect YB-1 binding of nucleic acid

6.1. Introduction

This Chapter describes an AlphaScreen assay that detects YB-1 binding to a ssDNA oligonucleotide. As orthogonal assays are a requirement for high throughput screening, this assay uses a system separate from the assay described in the previous Chapter.

The Chapter begins with an introduction to the AlphaScreen system and the complementary TruHits assay, before a description of the optimisation of the assay conditions. A series of experiments are described in which the components of the system are altered and tested to develop a sensitive, high-quality assay. While representative experiments are presented in this Chapter; the Appendix contains an outline of the over 200 experiments performed during development of this assay. Finally, the Chapter ends with reflection on the process of assay development.

6.2. Outline of methods

6.2.1. Introduction to the AlphaScreen system

The AlphaScreen assay is based on a series of binding interactions between its components, as portrayed in Figure 6-1 below. When excited by light at the excitation wavelength of 680nm, the donor beads emit unstable oxygen singlet molecules. The singlet oxygen has a half-life of only 4 µsec, and can diffuse approximately 200 nm in solution. If an acceptor bead is brought into proximity of a donor bead by the series of binding interactions, then energy is transferred to the acceptor bead so that it emits a signal at 570nm, which is detected by a plate reader. The AlphaScreen assay system was adapted to screen for inhibitors of YB1 binding to ssDNA. The ssDNA was the biotinylated 3xCohen oligonucleotide which bound to the streptavidin-coated AlphaScreen donor beads. The acceptor beads were conjugated to sheep anti-YB-1 antibodies, which bind to the YB-1 protein. The beads are brought into proximity by the binding of YB-1 to the 3xCohen oligonucleotide.
AlphaScreen assay

Figure 6-1 Graphical overview of the AlphaScreen assay.

The biotinylated 3xCohen oligonucleotide binds the streptavidin-coated AlphaScreen donor beads. The acceptor beads are conjugated to sheep anti-YB-1 antibodies, which bind to the YB-1 protein. The beads are brought into close proximity by the binding of YB-1 to the 3xCohen oligonucleotide, and the resulting emitted light is detected. Modified from Alpha Technology Brochure from Perkin Elmer, available at http://www.perkinelmer.com/labsolutions/resources/docs/BRO_AphaTechnology.pdf

6.2.1.1. AlphaScreen in previous HTS campaigns
AlphaScreen was designed to be suitable for high throughput screening\textsuperscript{310}. The assay requires no wash steps, and the AlphaScreen beads do not contain large epitope tags that may interfere with the binding of the subject proteins. AlphaScreen’s relatively long excitation and emission wavelengths reduce the number of false positives caused by compound fluorescence\textsuperscript{311}, as compound libraries tend to contain a greater percentage of compounds that fluoresce at wavelengths shorter than 500nm\textsuperscript{312}. 


AlphaScreen has been used to screen compounds that interfere with protein to peptide interactions\textsuperscript{313} and protein to protein interactions\textsuperscript{314}, including transcription factors\textsuperscript{315-317}. AlphaScreen was originally developed for detecting protein to protein interactions\textsuperscript{310}, and has only more recently been used in screens detecting protein to nucleic acid binding\textsuperscript{318-321}. The assay that has been developed during this project is a unique use of this assay system, as it required developing an assay for protein binding to ssDNA.

6.2.2. TruHits assay

In order to eliminate false positives amongst the compound hits identified during HTS, the AlphaScreen TruHits kit was used. This kit consists of streptavidin-coated AlphaScreen donor beads, and biotinylated donor beads. As previously described, donor beads are excited by light to generate oxygen singlet molecules which induce proximal acceptor beads to emit a detectable signal. In the TruHits kit, the streptavidin:biotin binding causes the two types of beads to come into the proximity required for them to generate a signal, as illustrated in Figure 6-2 below.

![Excitation 680nm Emission 570nm](image)

Figure 6-2 Graphical overview of the TruHits assay. Streptavidin-coated Donor beads bind biotinylated Acceptor beads to form a complex that emits light at 570nm upon laser irradiation at 680nm. Compounds that interfere with the assay system will reduce the intensity of this emitted light.
As YB-1 and the 3xCohen oligonucleotide are not involved in this system, the only way that a compound can inhibit signal is by interfering with the assay system itself. Compounds that interfere directly with AlphaScreen assays (and are thus false positives) and can be detected by the TruHits assay, are of four classes:

1. Singlet oxygen quenchers. These are compounds that bind or react with the singlet oxygen generated by the donor bead, preventing the singlet oxygen from inducing the acceptor bead to emit a signal.

2. Colour quenchers. These are compounds that can absorb either the 680nm wavelength light that excites the donor bead, or the 570nm signal emitted by the acceptor bead.

3. Light scatterers. These are compounds that bind together to form large aggregates that diffract either the excitation or the emission light.

4. Biotin mimetics. These are compounds that compete with biotin for binding to the streptavidin coating the assay beads.

While the TruHits kit tests for the above 4 classes of interfering, false positive compounds, it does not test for fifth class termed acceptor bead competitors. These compounds interact with the anti-YB-1 antibodies which are not a part of the TruHits kit system. If a compound binds to the anti-YB1 antibodies, then it can prevent binding between the acceptor bead and the target protein YB-1.

The potential interference of these five classes of compounds on the YB-1 AlphaScreen assay is shown graphically in Figure 6-3 below.
Figure 6-3 Overview of the AlphaScreen assay, showing points at which false positives can interfere with the assay system (red arrows).

When excited by light, the donor beads emit unstable oxygen singlet molecules that can diffuse only approximately 200nm. If the donor bead is bound to oligonucleotide bound by YB-1, and that YB-1 is bound by antibodies on the acceptor bead, then the oxygen singlet is able to reach and excite the acceptor bead to emit a detectable signal. 5 classes of false positive compounds, written in red, can interfere with this assay system. Colour quenchers and light scatterers block the excitation or emission light used by the system. Singlet oxygen quenchers react with the singlet oxygen generated by the donor bead, preventing the singlet oxygen from inducing the acceptor bead to emit a signal. Biotin mimetics bind the streptavidin coating the donor beads, prevent them from being bound by the biotinylated ssDNA. Acceptor bead competitors bind the antibodies that coat the acceptor bead, preventing them from binding the YB-1 protein.

6.2.2.1. Make-shift assay for false positives

Prior to the laboratory’s acquisition of the TruHits kit, an alternative assay for false positives was devised. This assay exploited the background luminescent signal observed in YB-1-lacking negative control wells containing acceptor beads, donor beads and the biotinylated oligonucleotide. If a compound is able to inhibit this background signal, that suggests that it
is interfering with the assay components, rather than inhibiting YB-1. This assay was only used on the compound hits from the initial, pilot compound screen and was superseded by the TruHits kit, which was used in later screening.

6.3. Optimisation of the YB-1:ssDNA AlphaScreen assay conditions

6.3.1. The process of experimental condition optimisation

The objective of optimisation was to identify the conditions that yield the highest ratio between positive control wells, containing YB-1 binding to ssDNA oligonucleotide, and background control wells. The conditions that were optimised included assay buffer, all reagent concentrations, order of reagent addition to the reactions, incubation periods. Furthermore, because of the interactions between the component reagents, a change to the concentration of one may have altered the optimal concentration of another. Therefore, the optimisation of AlphaScreen experimental conditions for YB-1 binding of ssDNA was an iterative process. A single condition was determined not by a single experiment in one day, but rather was resolved and adjusted during a course of experiments over time. Subsequent experiments testing one condition would have their other conditions updated to reflect optimisation of these other conditions in the time since earlier experiments.

The conditions optimised are presented here in chronological order. However, this optimisation process was iterative, and the optimising of different conditions did over-lap and often occurred concurrently, so this chronological ordering should be considered to be approximate. A spread sheet showing the objectives of each of the optimising experiments, emphasising the over-lap between condition optimisation, is given in the Appendix.

6.3.2. 3xCohen DNA oligonucleotide selected over 3xFasR oligonucleotide

In order to maximise YB-1 to nucleic acid binding, it was necessary to select an oligonucleotide for which YB-1 has a high binding efficiency. Two YB-1 binding sequences had been previously studied by Dr Annette Lasham and Prof. Antony Brathwaite145,149 and two oligonucleotides, 3xCohen and 3xFasR, were obtained consisting of three repeats of these sequences. The use of 3x multimers was to increase YB-1 binding in two ways. Firstly, three binding sites may allow more copies of YB-1 to bind. Secondly, bringing multiple molecules of YB-1 into close proximity using these adjacent binding sites may facilitate the multimerisation of YB-1. The binding of YB-1 to the 3xCohen and 3xFasR DNA
oligonucleotides was compared by performing a dot blot with different amounts of each oligonucleotide immobilised on a nitrocellulose membrane, incubating with YB-1 protein in South-Western (SW) buffer, and using anti-YB-1 antibodies to detect YB-1 bound to the immobilised oligonucleotide. Although endogenous YB-1 has previously been shown to bind 3xFasR in SW buffer\textsuperscript{145}, this dot blot also tested the whether the purified recombinant YB-1 protein was functional and able to bind to these oligonucleotides in SW buffer. YB-1 protein directly blotted onto the nitrocellulose membrane was used as a positive control to confirm detection of YB-1 protein by the YB-1 antibody. The dot blot is shown in Figure 6-4 below.

This blot confirmed that YB-1 does bind both oligonucleotides. It was clear that more YB-1 binds 3xCohen than 3xFasR, or that YB-1 binds 3xCohen with higher affinity than it binds 3xFasR.
YB-1 is also known to bind its own mRNA. In order to assess the viability of developing RNA-binding assays for YB-1, the dot blot was repeated using YBX1 RNA synthesised by *in vitro* transcription. Blots were treated according to the described method using four different buffers: SW+MgCl₂, SW+MgCl₂-KCl, SW+MgCl₂-KCl-gly and Thermo RNA-protein binding buffer. The Thermo RNA-protein buffer produced the blot with strongest signal and lowest background. However, the signal was not as strong as the dot blots using DNA oligonucleotides. Because of this, it was decided to continue development using just DNA oligonucleotides.

These dot blot experiments were followed by AlphaScreen experiments that trialled both DNA oligonucleotides and also confirmed that the signal after YB-1 binding to 3xCohen was higher than YB-1 binding to 3xFasR (data not shown). Use of 3xFasR was discontinued, with later experiments using only 3xCohen.

Initial experiments showed that wells containing all four components of acceptor beads, donor beads, 3xCohen and purified YB-1 protein generated a signal higher than any wells containing 3 or fewer of these components. These wells were referred to as ‘positive’ control wells. Of the negative control wells containing only 3 components, wells containing acceptor beads, donor beads and 3xCohen generated the highest signal. Those negative control wells were referred to as YB-1-lacking negative controls.

### 6.3.3. Assay viability established in 96-well plates

Initial experiments that established the viability of this AlphaScreen assay were performed in 96-well plates. As the 384-well plate format is more desirable for high-throughput screening, all experiments after the initial experiments were performed in 384-well plates. Similar signal ratios were observed between experiments performed on 96-well and 384-well plates.

### 6.3.4. Determination of reagent addition order and selection of incubation times

The last reagent added during the initial format protocol, described in the Method section, was the acceptor beads. However, it was thought that of the three sets of binding occurring during the AlphaScreen reactions (acceptor beads’ antibodies to YB-1 protein, YB-1 protein to biotinylated oligonucleotide, biotinylated oligonucleotide to streptavidin-coated donor bead), the binding of antibodies to YB-1 protein may take the longest time to reach maximum binding, and may be hindered by donor beads binding the oligonucleotide first. Therefore, the
reagent addition order was altered so that donor beads were added last, and acceptor beads were added earlier, to allow this binding to occur earlier in the course of the reaction. Additionally, the donor beads are light-sensitive, so adding them last permits earlier steps of the experiment to be performed outside of a dark-room.

Using 96-well plates, the initial format protocol’s 60 minute incubation following addition of its final reagent was determined by an experiment using replicate wells that were read on the plate reader simultaneously, but set up so that final incubations were 15, 30, 45, 60 or 75 minutes. Compared to reactions that had all components apart from YB-1 protein (negative control), the reactions that also contained YB-1 protein (positive reaction or positive control) ratios over negative wells were 1.23:1, 1.62:1, 1.68:1, 1.76:1, 1.44:1 respectively. Subsequent reactions used final incubations of 60 minutes.

6.3.5. Selection of assay buffer

Buffers were tested to ascertain optimal YB1:3xCohen binding conditions in the AlphaScreen assay. Reactions containing purified YB-1 protein were considered positive controls, and reactions without purified YB-1 protein were considered negative controls. Many buffers were based on SW buffer which had previously been used successfully to get YB-1 to bind DNA in vitro. Modified buffers were prepared and tested on advice from Perkin Elmer AlphaScreen technical expert Margareta Sutija. The buffers tested included: SW buffer, SW+MgCl2 buffer, SW+MgCl2-KCl buffer, SW+MgCl2-KCl-gly buffer, SW+MgCl2-KCl-gly-EDTA buffer, SW+MgCl2-KCl-gly + 0.5% BSA buffer, SW+MgCl2-KCl-gly + 0.5% dextran buffer, SW+MgCl2-KCl-gly + 0.5% gelatin buffer, PBS, PBS + 0.2% BSA, PBS + 0.1% BSA, PBS + 0.1% BSA + Tween, PBS + 0.5% dextran, PBS + 0.5% gelatin, TBS+Tween, ImmunoAssay buffer, NaCl HEPES buffer, and HiBlock buffer. Buffers were assessed by comparing the ratio of signals from the YB-1:3xCohen binding positive controls and the YB-1-lacking negative controls. The two highest ratios generated with SW+MgCl2-KCl-gly and PBS + 0.1% BSA buffer were 1.81:1 and 2.07:1 respectively. These two buffers were then further evaluated in three subsequent experiments. The average positive and negative control signals for this series of four experiments are given below in Table 6-1.

Table 6-1 A comparison of two buffers used in AlphaScreen experiments detecting YB1 binding to 3xCohen oligonucleotide.

Tabulated below are the average luminescent signals across 4 AlphaScreen experiments comparing two buffers. The grid cell of each ratio is coloured according to its size, with high
ratios represented by darker shades of red, and low ratios represented with shades approaching white.

<table>
<thead>
<tr>
<th>Experiment:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW+MgCl2-KCl-gly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average positive control</td>
<td>59282</td>
<td>48580</td>
<td>78576</td>
<td>43869</td>
</tr>
<tr>
<td>Average negative control</td>
<td>32712</td>
<td>20756</td>
<td>18963</td>
<td>12137</td>
</tr>
<tr>
<td>Ratio positive/negative</td>
<td>1.81:1</td>
<td>2.34:1</td>
<td>4.14:1</td>
<td>3.61:1</td>
</tr>
<tr>
<td>PBS + 0.1% BSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average positive control</td>
<td>12573</td>
<td>5455</td>
<td>6082</td>
<td>9662</td>
</tr>
<tr>
<td>Average negative control</td>
<td>6052</td>
<td>2082</td>
<td>2360</td>
<td>2403</td>
</tr>
<tr>
<td>Ratio positive/negative</td>
<td>2.08:1</td>
<td>2.62:1</td>
<td>2.58:1</td>
<td>4.02:1</td>
</tr>
</tbody>
</table>

6.3.5.1.1 Background luminescence associated with DNA oligonucleotide 3xCohen

The results from the AlphaScreen assay shown above demonstrated that in the presence of SW+MgCl2-KCl-gly buffer, the luminescent signals of both positive and negative controls are much higher than those in PBS + 0.1% BSA. However, the ratios between positive and YB-1-lacking negative controls were similar for both buffers. The high signal in the YB-1-lacking negative control using SW+MgCl2-KCl-gly buffer was considered to be above background luminescence. This background signal is not present in wells that do not contain the DNA oligonucleotide 3xCohen – that is, wells that contain only purified YB-1 protein, acceptor beads and donor beads and buffer have a much lower signal than wells containing 3xCohen, acceptor beads and donor beads. Although this was also the case when the assay buffer was PBS + 0.1% BSA, the background signal was proportionately higher when SW+MgCl2-KCl-gly buffer was used. Experiments titrating the amount of DNA oligonucleotide used in the assay showed that this background signal increased in proportion to the concentration of 3xCohen. Attempts were made to lower this background luminescence by supplementing the SW+MgCl2-KCl-gly buffer with BSA, dextran or gelatin, but these additives had no effect. Without an experimentally validated theory of the origin of the background signal in the SW+MgCl2-KCl-gly buffer, PBS + 0.1% BSA, in which ratio of positive to negative control signal was the same but the background signal much lower, was chosen for use for future experiments.
6.3.6. Evaluating a range of 3xCohen oligonucleotide concentrations

In preliminary experiments, oligonucleotide and purified YB-1 concentration were set high in a local optimum that was not discovered until the two variables were tested together in a series of matrices, shown below in Table 6-2.

Table 6-2 AlphaScreen assay testing a range of both 3xCohen oligonucleotide and purified YB-1 protein concentrations.

Ratios on the right show signals divided by signal of negative control wells. On the left, the values of the luminescent signal are given, colour-coded to represent the amount of signal. High signals are represented by red shading, and low signals represented by blue. On the right, the ratios of positive control signals divided by signal from the negative control wells. The grid cell of each ratio is coloured according to its size, with high ratios represented by red, and low ratios represented by blue.

<table>
<thead>
<tr>
<th>Oligo conc. (fmol/μL)</th>
<th>Oligo conc. (fmol/μL)</th>
<th>Ratios (to 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 10 20 40</td>
<td>0 10 20 40</td>
<td></td>
</tr>
<tr>
<td>YB-1 conc. (fmol/μL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 5307 29793 36823</td>
<td>35514</td>
<td></td>
</tr>
<tr>
<td>10 2030 35766 42583</td>
<td>39841</td>
<td>0.38 1.2 1.16 1.12</td>
</tr>
<tr>
<td>20 1972 37490 43604</td>
<td>39938</td>
<td>0.37 1.26 1.18 1.12</td>
</tr>
<tr>
<td>40 2670 42875 45989</td>
<td>40113</td>
<td>0.5 1.44 1.25 1.13</td>
</tr>
</tbody>
</table>

The ratio of the luminescent signal of the positive control wells divided by that of the negative control wells was highest when the oligonucleotide concentration was 10fmol/μL, and the YB-1 concentration was 40fmol/μL. This was a result of the background signal in the negative control being lowest when oligonucleotide concentration was lowest (10fmol/μL). Therefore, the nest experiment extended on these findings and tested the concentrations of YB-1 to 80fmol/μL and also trialled further decreasing 3xCohen oligonucleotide to 1.25fmol/μL. This experiment also evaluated two different batches of YB-1 antibody-conjugated acceptor beads. Results are shown below in Table 6-3.

Table 6-3 AlphaScreen assay testing a range of both 3xCohen oligonucleotide and purified YB-1 protein concentrations, performed using two batches of acceptor bead.

The grid cell colour represents the size of the signal (or ratio), with high signals (and ratios) represented by red, and low signals (and ratios) represented by blue.
The results of the first of these experiments showed that the positive to negative signal ratios increased as the oligonucleotide concentration decreased, while increasing the concentration of YB-1 did not have a large impact on the positive to negative signal ratio. The last experiment again showed the positive to negative signal ratio increased with decreasing oligonucleotide concentration.

### 6.3.7. Optimisation of the AlphaScreen acceptor and donor bead concentrations

Perkin Elmer technical expert Margareta Sutija also advised trialling of various concentrations of the AlphaScreen acceptor and donor beads to increase the luminescent signal, including after new batches of acceptor beads were prepared, or donor beads...
purchased. Frequent optimisations of acceptor and donor bead concentrations were therefore made during the assay development process using matrices simultaneously varying both beads’ concentrations. Given below in Table 6-4 is an example of this type of experiment, performed using oligonucleotide concentration of 1.25fmol/μL and using two different batches of acceptor bead. For each batch of acceptor bead, shown below are three 4x4 matrices with donor bead concentration varying along the X-axis, and acceptor bead concentration varying along the Y-axis. The left matrices shows the signal of wells containing both oligonucleotide and YB-1; the middle matrices shows YB-1-lacking negative control wells containing oligonucleotide but no YB-1, and the right matrices show the ratio between these two. Note that there are three missing data points lost to technical error.

Table 6-4 AlphaScreen assays testing a range of both acceptor bead and donor bead concentrations, performed using two batches of acceptor bead.

Given for each batch of acceptor bead are three 4x4 matrices with donor bead concentration varying along the X-axis, and acceptor bead concentration varying along the Y-axis. The left matrices shows the signal of wells containing both 3xCohen oligonucleotide and YB-1; the middle matrices shows YB-1-lacking negative control wells containing 3xCohen oligonucleotide but no YB-1, and the right matrices show the ratio between these two. Note that there are three missing data points lost to technical error. The grid cell colour represents the size of the signal or ratio, with high signals and ratios represented by darker shades of red, and low signals and ratios represented with shades approaching white.
These results showed that although the acceptor bead concentration had more influence on raw signal, the signal ratios were more influenced by donor bead ratio. In acceptor bead batch 3, the highest ratios are observed when the donor bead concentration is 20ng/μL, while with acceptor bead batch 4 it is not clear whether higher ratios are achieved by donor bead concentrations of 20ng/μL or 40ng/μL. Based on these results, it was concluded that concentrations of 20ng/μL were appropriate for both acceptor and donor beads and so were used in subsequent experiments.

### 6.3.8. Optimisation of reaction conditions with an oligonucleotide competitor

1xCohen competitor is a DNA oligonucleotide with the sequence of a YB-1 binding site that is not conjugated to biotin, so will not become bound by donor beads. 1xCohen competitor is not complementary to 3xCohen and so will compete with 3xCohen oligonucleotide for the binding by YB-1. An excess of 1xCohen oligonucleotide, as used in the luciferase reporter gene assays (Chapter 5) should compete with the 3xCohen oligonucleotide for binding by YB-1. To test this, experiments were performed with increasing concentrations of 1xCohen competitor, from 0 to 1000fmol/μL. These experiments used 2.5fmol/μL of 3xCohen oligonucleotide with three or four concentrations of YB-1 (0, 20, 40 and 60 fmol/μL). A 1xCohen competitor oligonucleotide concentration of 1000fmol/μL represents a 400-fold molar excess over the 2.5fmol/μL of 3xCohen oligonucleotide. However, because the 3xCohen oligonucleotide contains three times the number of YB-1 binding sites, the 1000fmol/μL of 1xCohen competitor represents 133.3-fold excess of YB-1 binding sites. The results from two of these experiments, performed in different laboratories using different batches of acceptor beads, are shown in Figure 6-5 below. These experiments showed that increasing concentrations of 1xCohen competitor reduced the AlphaScreen luminescent signal.
Figure 6-5 Titration of 1xCohen competitor from concentrations ranging from 0 to 1000 fmol/μL. This experiment used 2.5fmol/μL of 3xCohen oligonucleotide with three concentrations of YB-1 (0, 20 and 40 fmol/μL).

A 1xCohen competitor concentration of 1000fmol/μL was sufficient to reduce the signal from reactions with either 60, 40 or 20fmol/μL YB-1 to the level of background signal of negative control reactions, and was thus the concentration used in later experiments.

6.3.9. Evaluating the effect of the BSA content of assay buffer on luminescent AlphaScreen signal

PBS + 0.1% BSA buffer were used in previous experiments described above. A BSA titration experiment using purified YB-1 binding to 3xCohen oligonucleotide was performed to test whether increasing the concentration of BSA might decrease the background luminescence seen in the negative control reactions. BSA might lower background luminescent signal by lowering non-specific binding that may occur between, for example, donor beads and YB-1, or 3xCohen oligonucleotide and acceptor beads. Excess amounts of BSA might also inhibit specific binding between YB-1 and the 3xCohen oligonucleotide. The results of this experiment are given in Figure 6-6 below.
Figure 6-6 Results of an AlphaScreen assay evaluating the effect of variable BSA levels on YB-1 binding to 3xCohen oligonucleotide, and non-specific binding between other reaction components. These results showed that increasing BSA reduced the background signal, but at high concentrations also inhibited signal from YB-1 binding to oligonucleotide. The signal/background ratio was highest with a minimal amount of BSA added to the buffer (0.0625%), although subsequent experiments used 0.2% BSA to ensure a greater reduction of background luminescence, and to ensure that the majority of luminescent signal could be attributed to the specific binding of YB-1 to 3xCohen oligonucleotide.

6.3.10. Different cellular lysates have different optimal concentrations

AlphaScreen assays were also developed to use cell lysates instead of purified YB-1 protein. 293T cells were transfected with pIRES2 eGFP YB1 and nuclear cellular lysates were extracted, as described in the Methods section. Experiments with lysates showed that while generally a lysate concentration of 50ng/μL yields good results under the experimental conditions described above, the true optimal concentration for each lysate preparation is
different for each lysate. Low concentrations result in less than optimal YB-1, while high concentrations result in too much non-specific binding, dampening the signal in the same manner as increasing concentrations of BSA. Shown as an example below in Figure 6-7 is a pair of lysate concentration optimising experiments performed on two lysates side-by-side.

![Graph](image)

**Figure 6-7** AlphaScreen signal across a range of concentrations of two cellular lysates to determine the lysate concentration at which luminescent signal is highest.

6.3.11. Evaluating the effect of DMSO on the luminescent AlphaScreen signal

As described previously, the compounds for screening are stored at DMSO, so the tolerance of the YB1:3xCohen AlphaScreen assay to DMSO was tested. Figure 6-8 presents two experiments using cellular lysates that show the effect of increasing amounts of DMSO on the AlphaScreen signal. The percentages of DMSO used were those that would be encountered during screening. Note that these two experiments were performed in different laboratories, using different 293T cell lysates, and different batches of acceptor bead.
Figure 6-8 Luminescent signal in the AlphaScreen assay is decreased by increasing concentrations of DMSO.
Two AlphaScreen assays, one performed in Shanghai (upper graph) and the other performed in Auckland (lower graph), incubating 293T cell lysates with 3xCohen oligonucleotide in the presence of increasing amounts of DMSO. The left-most data point is a negative control containing no lysate.

While lower levels of DMSO (0.5%) yielded higher luminescent signals than high levels of DMSO, adding less compound per well would reduce the sensitivity of the assay to compounds that maybe weaker inhibitors of YB-1. Therefore, it was decided to screen using 1% DMSO and compound. Although, in this experiment, 1% DMSO reduced the signal by nearly 50%, the signal is still clearly present and much higher than that of the negative control.

The optimised AlphaScreen assay developed here was next used in HTS to identify inhibitors of YB-1 binding to ssDNA, as described in Chapter 7.

6.4. Discussion

In total, over 200 AlphaScreen experiments were performed. Included in the Appendix is a table that displays the chronological order of the experiments, showing the main aspect tested by each experiment and where experiments failed, and the suspected reason for each experimental failure. With hindsight, it is possible to make a number of observations about the process of developing this assay. Described below are additional results and observations that were made during the development of these assays.

6.4.1. Local optima

When experiments individually test and optimise only one aspect, then it is possible to fall into a local optimum. For example, if the optimal concentrations of two components are being determined sequentially, then the determined concentration of the second component will depend upon the first. That is, if Condition A is first determined to exhibit best results when at X, and then later experiments to optimise Condition B will only give the optimal setting of Condition B given A=X. However, these individually optimised settings for A and B, may be inferior to those determined if A and B were determined simultaneously in a matrix where every possible setting of B was tested at every possible setting of A. For three,
four and increasing numbers of conditions, an exponentially greater number of possible combinations of settings are possible. However, practical constraints of experimental reagents and time make it too costly to simultaneously optimise multiple conditions.

An example of a local optimum occurred during assay development when the ssDNA oligonucleotide concentration was being optimised. Oligonucleotide concentration was titrated in experiments that were using high concentrations (2pmol/μL) of purified YB-1 protein. As a result, the oligonucleotide concentration was ‘optimised’ at a concentration of YB-1 that, in retrospect, was much too high. Then, purified YB-1 concentrations were titrated for optimisation during experiments that used this excessively high concentration of oligonucleotide, resulting in YB-1 concentration that was optimised to this excessively high oligonucleotide concentration. Thus, both oligonucleotide and YB-1 concentrations were, in hindsight, too high relative to the concentrations of acceptor and donor beads being used. As these beads were the rate-limiting reagents, this may have caused signal to be overly sensitive to bead quality and concentration, which, in cases of extreme sensitivity, may even vary within a single experiment’s bead master mix if beads settle and clump together (due to insufficiently frequent vortex mixing). This local optimum resulted in a string of failed experiments that was only resolved when concentrations of oligonucleotide and YB-1 were simultaneously varied in a matrix. Incidentally, during this string of failed experiments, a second problem arose (within the acceptor beads) and the two problems mutually convoluted the solving of the other. This illustrates how while a local optimum may still allow viable experiments, those experiments then have increased sensitivity to further problems.

6.4.2. Continual adjustment to new component batches

The sections above illustrate the direction of the assay’s optimisation, but some data from only a small fraction of the over 200 experiments performed in preparation for high-throughput screening. The types of experiments described above were performed numerous times to ensure conditions were still optimised whenever a new batch of an assay component was used for the first time. In particular, many different batches of acceptor beads and cellular lysates were prepared and consumed during development.

6.4.3. Optimal reagent concentration dependent upon reagent quality

If an aliquot of a reagent can degrade over time, then later experiments using that aliquot will have concentrations lower than calculated by the experiment’s protocol, resulting in reagents
being optimised to concentrations higher than the true concentration. In this assay, the reagents most vulnerable to degradation at room temperature or from freeze-thaw cycles were the purified YB-1 protein, and the cellular lysates. In later experiments, when this came to be appreciated, protein and lysates were stored as smaller, more numerous, single-use aliquots.

For some reagents made in batches that differ in quality, the optimal concentration in experiments will also differ. AlphaScreen acceptor beads were conjugated to anti-YB-1 antibodies. Batches of resulting acceptor beads may differ according to variation between different conjugation reactions, or between component aliquots of antibody and unconjugated beads used in those reactions. Therefore, different batches of acceptor bead may separately yield optimal results at different conjugations.

6.4.3.1. Experiments comparing different batches of Acceptor bead

New batches of acceptor beads were conjugated to anti-YB-1 antibodies as required. The ability of newly conjugated beads to bind YB-1 and thereby produce AlphaScreen signal was tested after each conjugation by comparison alongside previously conjugated beads. It was found that variation existed between acceptor beads, as illustrated by the sample experiments shown below in Figures 6-9 and 6-10.
Figure 6-9 Comparison of AlphaScreen signal between Acceptor beads NZ3 and Shanghai 1 (SH1), including different negative controls.

Signal is produced when the AlphaScreen Acceptor beads are brought into proximity of the Donor beads via the binding of purified YB-1 protein to the 3xCohen oligonucleotide sequence. The y-axis value of each bar is the mean AlphaScreen signal from 3 replicate wells. Error bars show standard error. Negative controls show signal when components are omitted: either oligonucleotide, purified YB-1 protein, or both oligonucleotide and protein.

Figure 6-10 Luminescent AlphaScreen signals from the comparison between Acceptor beads conjugated in New Zealand (NZ3), and those conjugated in Shanghai (SH2-8) in three batches (SH2, SH3-5 and SH6-8).

These results showed that in AlphaScreen assays using purified YB-1 binding to 3xCohen oligonucleotide, none of the acceptor beads conjugated with anti-YB-1 antibody in Shanghai, except SH2, gave luminescent signals as high as those conjugated in New Zealand. The conjugation protocol was not complicated, and assay specialists from Perkin Elmer reported little variation between their conjugations. Reagents used in the conjugation reactions were replaced to eliminate the possibility of their degradation. Variation may have existed between
the supplied, unconjugated beads, and the aliquots of antibody to which the beads were being conjugated.

6.4.4. Reduction of experimental volume

In order to reduce reagent consumption, experiments were performed to test the effect of reducing experimental volume from 50μL to 40, 30, 25 or 20μL. Lower signals were observed as volume was reduced. Signals from positive and negative controls decreased proportionately, so that the ratio between positive and negative reactions remained constant. However the ratios of experiments using reduced volumes were more vulnerable to variation in background signal, due to their lower signal. That is, an increase in background signal in a single experiment, due to stochastic variation, would cause the signal/background ratio to decrease more in a lower volume experiment where overall signals are lower. Larger volume experiments with higher luminescent signals were more robust to variation in background signal.

6.5. Summary

This Chapter described development of an AlphaScreen assay sensitive to inhibition of YB-1 binding of a target oligonucleotide sequence, 3xCohen. The experimental conditions used for this AlphaScreen assay during high-throughput screening are described in detail in Section 2.3.4.6.3. PBS + 0.1% BSA was chosen over other assay buffers because it gave the lowest background signal. To reduce background further, the BSA concentration was later increased to 0.2%. The tolerance of the assay for DMSO, the solute of the high-throughput screening compounds, was tested.

The concentrations of each assay component, the 1xCohen and 3xCohen oligonucleotides, purified YB-1 or cellular lysate, and the donor and acceptor beads, were optimised to maximise the signal ratio between positive and negative control groups. 2-dimensional matrices varying two components of the assay (e.g. 3xCohen oligonucleotide and purified YB-1 protein concentration) were required to find the optimal concentration of each component in the context of the concentration of the other component. While it is possible that higher dimensional matrices (e.g. with extra dimensions representing acceptor bead and
donor bead concentrations) may be more accurate, this is not feasible with limited components.

The component that most influences the quality of the assay, as assessed by the ratio between positive control wells and the negative control wells, was the quality of the batch of acceptor beads. This series of optimising experiments increased the quality of the assay to a level sufficient for the high-throughput screening described in Chapter 7.
Chapter 7 High-throughput screening using optimised assays

7.1. Introduction

During the process of high throughput screening (HTS), an initial, primary screen is used to test a large number of compounds. The ‘hit’ compounds identified as affecting the assay signal during the primary screen are then evaluated by an independent, orthogonal, secondary screen. A portion of the hit compounds identified by the primary screen will be false positives that interfere with the assay system, rather than the target of interest. The secondary screen uses a different assay system to test the hit compounds, and can detect the false positives that only interfere with the assay system of the primary screen, while identifying the compounds that affect the assay signal of both primary and secondary screens.

7.2. Pilot screening

Originally, the AlphaScreen assay was intended for use as the primary screen. Four 384-well plates of 320 compounds were used for a pilot screen to show that the assay had been optimised to a level suitable for high-throughput screening. An assay is suitable if it has a large difference in the signals of its positive and negative controls, and low variation within those control groups. The first two plates of compounds were screened using both purified YB-1 protein and cellular lysates, in parallel. For use in cellular lysate experiments, 293T cells were transfected with pIRES2 eGFP YB1 to increase the amount of YB-1 protein before nuclear lysates were extracted, as described in the Methods section. Assays using purified YB-1 should detect compounds that inhibit YB-1 binding to ssDNA, or that inhibit YB-1 multimerisation. As YB-1 is capable of multimerisation, which in this AlphaScreen assay would presumably increase the luminescent signal by bringing more YB-1-bound acceptor beads within proximity of donor beads. Assays using cellular lysate, instead of purified YB-1 protein, incubated with 3xCohen oligonucleotide, could detect compounds that inhibit YB-1 binding to ssDNA, inhibit YB-1 multimerisation, and additionally inhibit YB-1:co-factor binding (if the co-factor influenced YB-1 binding to the ssDNA).

Initially, one 384-well plate of 320 compounds was screened in AlphaScreen assays using purified YB-1 protein or cellular lysate binding to the 3xCohen oligonucleotide. From this screen, 23 compounds were identified as decreasing the assay signal. These 23 compounds
were subsequently re-tested and 19 compounds were confirmed that they inhibited the assay signal by more than 45%. Of these, 13 compounds inhibited the signal in both AlphaScreen assays with purified YB-1 protein and cellular lysate. Five compounds only inhibited luminescent signal from purified YB-1 protein binding to ssDNA, and one compound only inhibited the AlphaScreen signal in reactions with cellular lysate and ssDNA. Figure 7-1 below plots the percentage inhibition of the AlphaScreen luminescent signal by each ‘hit’ compound in the assays with purified protein and cellular lysate incubated with 3xCohen oligonucleotide. Figure 7-2 shows the intersection between the ‘hits’ from each of the two assays.

Figure 7-1 Graph showing the percentage inhibition of 23 compounds tested in AlphaScreen assays with either purified YB-1 protein or cellular lysates incubated with 3xCohen oligonucleotide. 13 compounds (black circles) inhibited signal of both AlphaScreen assays with purified YB-1 protein and cellular lysate. Five compounds (blue circles) only inhibited signal in reactions with purified YB-1 protein binding, and one compound (yellow circle) only inhibited signal in reactions with cellular lysate. Four compounds (grey circles) were found to not strongly inhibit signal in either assay.
Figure 7-2 Venn diagram showing the compounds hits in the purified protein and cell lysate assays. 13 compounds inhibited signal of both AlphaScreen assays with purified YB-1 protein (blue circle) and cellular lysate (yellow circle). Five compounds only inhibited signal in reactions with purified YB-1 protein binding, and one compound only inhibited signal in reactions with cellular lysate.

A second 384-well plate of 320 compounds was screened in two sets of AlphaScreen assays using either purified YB-1 protein or cellular lysate assays incubated with 3xCohen oligonucleotide. Hits from either set of assay were re-tested, yielding one hit compound that inhibited the purified YB-1 protein assay, and three hit compounds that inhibited signal in both the purified YB-1 protein assay and the cellular lysate assay.

From these first two plates, the 23 hit compounds were subjected to the make-shift assay for false positives, described in Section 6.2.2.1. 18 of these compounds appeared to inhibit the background signal emitted by wells containing only acceptor beads, donor beads and biotinylated oligonucleotide, suggesting that they were interfering with the assay system’s components and were false positives. 5 compounds were accepted as confirmed hits.

As most hit compounds were active in both AlphaScreen assays with purified YB1 protein and cellular lysates. The next round to screen a further 640 compounds used only the cellular lysate assay. Two 384-well plates, comprising 320 compounds each, were subjected to the AlphaScreen assay using cellular lysate incubated with 3xCohen oligonucleotide. Initial hit
compounds were re-tested for their inhibition of AlphaScreen reactions that used both purified YB-1 and cellular lysate assays. Of the initial 53 hits, for 20 compounds the inhibition of AlphaScreen assay signal was confirmed. However, 12 of these also inhibited signal in the make-shift assay for false positives, suggesting that they were false positives. This left 8 compounds as confirmed hits, to add to the 5 hits identified in the previous set of 640 compounds. The reduction of these 1280 total compounds to 13 hits is illustrated in Figure 7-3 below.

Figure 7-3 Representation of the sequential filtering of compounds interfering with YB1:ssDNA binding, as tested by screening through AlphaScreen assays.

Left triangle: the first 640 compounds were screened with AlphaScreen assays using both purified YB-1 and cell lysate binding to 3xCohen oligonucleotide. Right triangle: the second 640 compounds were screened only with AlphaScreen assays using the purified YB-1 binding to 3xCohen oligonucleotide. The initial hits of these screens were re-tested to confirm inhibition. The compounds with confirmed inhibition were tested in the make-shift assay for false positives, which eliminated all but 13 compounds.

Subsequently, these 13 compounds were repeatedly tested in AlphaScreen assays using both purified YB-1 protein and cellular lysates, as well as the make-shift assay for false positives. 3 compounds were removed as having weak signal inhibition, or as suspected false positives.

Once the TruHits kit was acquired, it was used to test these remaining 10 compounds. 6 compounds were identified by this kit as being false positives interfering with the assay components.
7.2.1. Decision to use cell-based luciferase assay as the primary screen

This pilot screen suggested that hits identified by the AlphaScreen for inhibition of YB-1 binding to ssDNA may include a high proportion of false positives. Additionally, acceptor beads conjugated to antibodies in Shanghai were generally inferior to beads conjugated in Auckland, for unknown reasons. AlphaScreen experiments performed in Shanghai using acceptor beads conjugated in Shanghai had lower ratios of experimental luminescent signal to background luminescence than those performed with acceptor beads generated in Auckland.

In Auckland, a new batch of antibodies were isolated from serum, and successfully conjugated to acceptor beads. These antibodies were couriered to Shanghai, and the reagents of the conjugation reaction were replaced with new batches. However, conjugation of acceptor beads to antibodies in Shanghai still yielded acceptor beads that gave lower luminescent signal on binding to YB-1, than those acceptor beads which were conjugated to antibodies in Auckland. It is possible the antibodies were damaged during shipping.

As a result of this, the acceptor beads conjugated in Shanghai required a greater concentration of acceptor beads and purified YB-1 protein be used in each AlphaScreen reaction, to maintain a signal to background ratio sufficiently high for HTS. This increased the cost per reaction.

Primarily because of the high cost per reaction of the AlphaScreen assay, it was decided that the cheaper cell-based luciferase reporter gene assay would be used as the primary screen. This cell-based assay has less specificity for YB-1 than the AlphaScreen assay, as compounds can inhibit luciferase luminescent signal by inhibiting any proteins that up-regulate the E2F1 promoter, or interfere with off-target components of the complex cell system. Tangentially, discovery of an inhibitor of E2F1 expression would still be a useful outcome. However, as a primary screen, the cell-based assay would yield hits that are more biologically relevant for future in vivo studies (for example, by excluding compounds that cannot pass the cell membrane).

The purified YB1:3xCohen oligonucleotide AlphaScreen assay would then be used as a secondary screen, so that off-target hits from the E2F1 promoter:luciferase reporter assay that did not inhibit YB-1 would be eliminated. The use of two separate, orthogonal screens would reduce the chances of retaining false positives; that is, only false positives that can disrupt both assay systems would be retained, for example, light-absorbing compound aggregates.
Furthermore, in each assay the DNA sequence to which YB-1 binds is different: in the AlphaScreen assay the 3xCohen oligonucleotide sequence is used and in the cell based assay the E2F1 promoter sequence is used.

7.3. Primary screen: cell-based luciferase assay

Using the optimised protocol described in chapter 5, HCT116 cells were transfected with plasmid pGL4.17-E2F1-728, incubated in the presence of compounds, and the luminescent signal measured. HCT116 cells have moderately high endogenous levels of both YBX1 and E2F1, based on expression data obtained from the Cancer Cell Line Encyclopedia\textsuperscript{205}. Small molecule compounds that inhibit E2F1, YB-1 or other transcription factor activators binding to the E2F1 promoter would decrease the luminescent signal. An excess of the 1xCohen oligonucleotide competitor was used as a model for the effect of YB-1 inhibition upon reporter gene activity. Cells transfected with the pGL4.17-E2F1-728 plasmid alone were considered to exhibit luminescent signal representing reporter gene activity that has not been inhibited, while cells co-transfected with pGL4.17-E2F1-728 plasmid and 1xCohen oligonucleotide were considered to exhibit luminescent signal representing reporter gene activity with 100% inhibition.

7.3.1. Optimisation of the cell-based assay for HTS

Cell-based assay conditions were selected for reproducibility between experiments, to increase the luminescent signal ratio between not inhibited and 100% inhibited controls, and to reduce signal variability within the control wells exhibiting luminescent signal that has not been inhibited. Experimental conditions that were tested during the optimisation described in Chapter 5 included: transfection with different plasmid amounts, plasmid co-transfection with YB-1 and E2F1 over-expressing plasmids, forward versus reverse transfection, varied cell concentrations and numbers per plate well, plate types and formats, variation of time gaps between transfection, addition of compounds and luciferase reading. Experiments testing these parameters were largely aimed at reducing the variability of the high luminescent signal from the controls that were not inhibited (thus increasing Z-factor). It was observed that variability between these control wells increased with time after transfection, suggesting a kind of stochastic drift outside of experimental parameters.
In total, 7,360 compounds were screened on 23x 384-well plates. This involved 6 independent transfections and was performed in collaboration with Zhou Yan at the Chinese National Compound Library.

7.3.2. Primary screen assay quality

The Z-factor is a measurement of HTS assay quality that is described in Section 2.3.3.6. The Z-factors for the 23 plates were within the range of 0 to 0.5, defining this as a marginal assay. The transfections for the screens were performed in sets, so that plates in the same transfection set used cells from in the same batches of transfected cells. There is strong evidence that the transfection set had a significant effect upon Z-factor (p-value = 0.000191). Z-factor, while calculated from the averages and standard deviations of both sets of controls, was most closely correlated to standard deviation of the luminescent signals from controls that were not inhibited with 1xCohen oligonucleotide (Correlation coefficient = -0.94). For a large-scale screen, this may result in too many hits proceeding to the secondary screen. However, for this moderately-sized screen, confirming hits by re-testing was feasible within the time-constraints of the project. Z-factors for each of the 23x 384-well plates used during the luciferase screen are given in Table 7-1 below.

Table 7-1 Showing the variation in Z-factor across the 23 plates used during primary screening. Transfection reactions were performed in sets that were used with multiple 384-well compound plates. Shown below for each 384-well plate, are the average luminescent signals and their standard deviations from controls groups in which inhibition from the 1xCohen competitor oligonucleotide was absent or present. Coefficient of variation is also shown. The average luminescent signal from the control group that was not inhibited has been divided by the average luminescent signal from the inhibited control group to give an ‘Average signal Ratio’ that is the fold difference between controls. Z-factor has been calculated as an indicator of the quality of each plate. Shown on the bottom row of the table are the Pearson’s correlation coefficient between the metric in each column and the Z-factor.
<table>
<thead>
<tr>
<th>Transfection set</th>
<th>384-well Plate</th>
<th>Luminescent signal from controls that were not inhibited</th>
<th>Luminescent signal from controls that were 100% inhibited by the 1xCohen competitor</th>
<th>Average</th>
<th>Standard deviation</th>
<th>Coefficient of variation</th>
<th>Average</th>
<th>Standard deviation</th>
<th>Coefficient of variation</th>
<th>Average signal Ratio (to 1)</th>
<th>Z-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>RUS0201-0204</td>
<td>7938.13</td>
<td>1589.21</td>
<td>0.2</td>
<td>207.86</td>
<td>125.24</td>
<td>0.6</td>
<td>38.19</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>RUS0205-0208</td>
<td>30619</td>
<td>4970.94</td>
<td>0.16</td>
<td>1305.31</td>
<td>423.98</td>
<td>0.32</td>
<td>23.46</td>
<td>0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>JK0391-0394</td>
<td>31014.81</td>
<td>6437.69</td>
<td>0.21</td>
<td>2386.56</td>
<td>842.81</td>
<td>0.35</td>
<td>13</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>JK0395-0398</td>
<td>30350</td>
<td>7398.57</td>
<td>0.24</td>
<td>2257.19</td>
<td>669.82</td>
<td>0.3</td>
<td>13.45</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>RUS0209-0212</td>
<td>33613</td>
<td>7044.73</td>
<td>0.21</td>
<td>2791.88</td>
<td>936.57</td>
<td>0.34</td>
<td>12.04</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>WNN0001-0004</td>
<td>25434.29</td>
<td>5205</td>
<td>0.2</td>
<td>2721.43</td>
<td>661.51</td>
<td>0.32</td>
<td>9.35</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>WNN0101-0104</td>
<td>29396.15</td>
<td>6437.69</td>
<td>0.24</td>
<td>2386.56</td>
<td>842.81</td>
<td>0.35</td>
<td>13</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>WNN0201-0204</td>
<td>31014.81</td>
<td>6437.69</td>
<td>0.21</td>
<td>2386.56</td>
<td>842.81</td>
<td>0.35</td>
<td>13</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>RUS0101-0104</td>
<td>3231</td>
<td>693.56</td>
<td>0.18</td>
<td>1756.55</td>
<td>452.57</td>
<td>0.26</td>
<td>25.5</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>RUS0105-0108</td>
<td>44788.75</td>
<td>8273.23</td>
<td>0.18</td>
<td>1756.55</td>
<td>452.57</td>
<td>0.26</td>
<td>25.5</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>RUS0109-0112</td>
<td>45369.35</td>
<td>9332.51</td>
<td>0.21</td>
<td>1861.03</td>
<td>562.77</td>
<td>0.3</td>
<td>24.38</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>RUS0113-0116</td>
<td>51765.71</td>
<td>9559.68</td>
<td>0.18</td>
<td>1984.19</td>
<td>654.07</td>
<td>0.33</td>
<td>26.09</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>RUS0117-0120</td>
<td>44380</td>
<td>8407.03</td>
<td>0.19</td>
<td>1888.97</td>
<td>547.62</td>
<td>0.29</td>
<td>23.49</td>
<td>0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0001-0004</td>
<td>44675.33</td>
<td>11524.75</td>
<td>0.26</td>
<td>2073</td>
<td>514.03</td>
<td>0.25</td>
<td>21.55</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0005-0008</td>
<td>46110</td>
<td>10746.38</td>
<td>0.23</td>
<td>2093.45</td>
<td>641.14</td>
<td>0.31</td>
<td>22.03</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0009-0012</td>
<td>43096.55</td>
<td>9930.35</td>
<td>0.23</td>
<td>2419</td>
<td>751.96</td>
<td>0.31</td>
<td>17.82</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0013-0016</td>
<td>45061</td>
<td>11423.59</td>
<td>0.25</td>
<td>2184.83</td>
<td>670.41</td>
<td>0.31</td>
<td>20.62</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0017-0020</td>
<td>41292.41</td>
<td>9479.08</td>
<td>0.23</td>
<td>2035.67</td>
<td>524.76</td>
<td>0.26</td>
<td>20.28</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0021-0024</td>
<td>42612.14</td>
<td>12725.55</td>
<td>0.3</td>
<td>1907.14</td>
<td>471.91</td>
<td>0.25</td>
<td>22.34</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0025-0028</td>
<td>46156.55</td>
<td>11255.38</td>
<td>0.24</td>
<td>2406.79</td>
<td>733.3</td>
<td>0.3</td>
<td>19.18</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0029-0032</td>
<td>45261.79</td>
<td>10736.7</td>
<td>0.24</td>
<td>1994.33</td>
<td>463.19</td>
<td>0.23</td>
<td>22.7</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>RUS0033-0036</td>
<td>65396.3</td>
<td>18677.49</td>
<td>0.29</td>
<td>3518.33</td>
<td>1340.01</td>
<td>0.38</td>
<td>18.59</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Pearson's correlation coefficient between given metric and Z-factor | -0.22 | -0.58 | -0.94 | -0.53 | -0.43 | 0.19 | 0.40 | 1 |

### 7.3.3. Results from the primary screen

In order to determine the inhibition of luciferase activity caused by each compound, the percentage decrease was calculated for the luminescent signal measured in each well. For the plot below, the calculated percentage inhibitions for every compound on each of the 23 plates were pooled, and their distribution plotted. The percentage decrease in luminescent signal in
the population had a median of 38, and a mean of 33.8 with a standard deviation of 32.77. Distribution was plotted (Figure 7-4) and the data was found to be left-skewed. The long left-tail may partially be due to fluorescent compounds.

![Density of Percentage Inhibition](image)

**Figure 7-4** Plot showing the distribution of percentage signal reduction resulting from inhibition of luciferase activity by 7,360 compounds across 23 384-well plates used during primary screening.

The percentage decrease in luminescent signal in the population had a median of 38, and a mean of 33.8 with a standard deviation of 32.77. The data was left-skewed.

The high median and mean percentage reduction in signal may reflect the sensitivity of the luminescent signal to cell proliferation. That is, the luminescent signal is dependant not only on promoter activity, but also cell number/proliferation. Both the E2F1 promoter activity and cell proliferation rates are dependent upon E2F1 and YB-1 protein activity, but many other factors also influence proliferation. Compounds that influence cell proliferation will influence the luminescent signal, as will compounds that interfere with other cellular processes, such as translation. Before the primary screen was run, it was expected that some compounds would inhibit luminescent signal independent of E2F1 or YB-1 protein inhibition, and that any hits not associated with YB-1 (including E2F1 inhibitors) would be filtered out during the YB-1-specific secondary screen.
For each plate, the distribution of the percentage signal reduction was considered in order to select plate-specific cut-offs (ranging from 60% to 100%) that designate hits. The number of hit compounds from each 384-well plate of 320 compounds ranged from 0 to 19. These 342 hit compounds were re-tested to confirm inhibition of luciferase activity indicated by a reduction in luminescent signal, and from this, 251 compounds proceeded into the secondary screen. The distribution of the percentage signal inhibition for these 342 compounds is plotted in Figure 7-5 below.

![Density of percentage inhibition of 342 compounds during validation screen](image)

**Figure 7-5** The distribution of percentage signal reduction resulting from inhibition of luciferase activity by 342 compounds tested in a validation screen following their identification as hits during the primary screen. The percentage decrease in luminescent signal in the population had a median of 62 and a standard deviation of 61.78.

### 7.4. Secondary screen: AlphaScreen using purified YB-1 protein

In order to determine if the reduction in luciferase activity was due to inhibition of YB-1 binding to DNA, the 251 hits from the luciferase reporter assay primary screen were tested in the AlphaScreen assay using purified YB-1 protein bound to ssDNA. As described previously, compounds that inhibit the binding of YB-1 to the 3xCohen oligonucleotide will reduce AlphaScreen luminescent signal. The 1xCohen oligonucleotide was used as a model for inhibition of YB-1 binding to the 3xCohen oligonucleotide. AlphaScreen reactions that included a 400-fold molar excess of 1xCohen produced lower luminescent signals than reactions containing only YB-1, 3xCohen, Acceptor and Donor beads. In the presence of the 1xCohen oligonucleotide, there would be less YB-1 available to bind to biotinylated
3xCohen, resulting in reduced AlphaScreen signal. For the “100% inhibited” control reactions, the purified YB-1 protein was incubated with 1xCohen before the addition of any other assay components. These control reactions were considered to represent total (100%) inhibition of the signal generated by YB-1 binding to the 3xCohen oligonucleotide. The “not inhibited” control reactions representing the absence of inhibition of YB-1 binding to 3xCohen oligonucleotide, had purified YB-1 incubated in assay buffer prior to addition of remaining assay components.

As a secondary screen, this AlphaScreen assay, detecting YB-1 binding to ssDNA, was considered as both a counter-screen and a selectivity screen. As a counter-screen, this AlphaScreen assay removed some false positives particular to the luciferase assay system used in the primary screen. As a selectivity screen, this AlphaScreen would differentiate between compounds that inhibit only YB-1, and compounds that interact with other proteins that influence the luminescent signal of the primary screen, such as E2F1.

7.4.1. Results from the secondary screening (AlphaScreen) assays

The 251 hit compounds from the cell-based assay screen were tested in this AlphaScreen assay in three sets. Each set contained a group of “not inhibited” control reactions that represented the level of luminescent signal when inhibition was absent, and a “100% inhibited” control group with the level of luminescent signal reduced due to inhibition of YB-1 by the 1xCohen competitor oligonucleotide. The Z-factor was calculated from the controls in each set and found to be 0.20, 0.37 and 0.56. The reduction in luminescent signal observed in each of the 251 reactions incubated with a compound was converted to a percentage. The calculated percentage of signal reduction for the three sets of compounds were pooled. The population of percentage signal reduction had a median of 32, and a mean of 33.2 with a standard deviation of 32.77. The distribution of AlphaScreen luminescent signal, representing the inhibition of YB1:ssDNA binding, effected by 251 compounds is plotted in Figure 7-6 below.
Figure 7-6 Plot showing the distribution of percentage inhibition of luminescent signal by 251 compounds screened in an AlphaScreen assay using purified YB-1 binding to 3xCohen oligonucleotide. The population had a median of 32, and a mean of 33.2 with a standard deviation of 32.77.

Compounds that reduced the luminescent signal by >50% of the reduction achieved by the 1xCohen competitor were considered hits. This totalled 67 compounds that were potentially inhibitors of YB1 binding to both the \( E2F1 \) promoter and the 3xCohen ssDNA oligonucleotide. These were subsequently screened to eliminate the false positives.

7.4.2. Performing the TruHits screen to eliminate false positives

The 67 hit compounds were analysed using the TruHits kit. Compounds were considered to be false positives if they reduced the luminescent signal of the TruHits reaction more than 2 standard deviations below the signal of the control reactions (treated with DMSO only). From this, 59 compounds were eliminated, leaving 8 compounds as “true” positives. The percentages of signal reduction for each of the 8 hit compounds, in each of the assay screens are shown in Table 7-2 below.

Table 7-2 Percentage reduction in luminescent signal during screening assay on 8 “true” positive compounds.
The percentage reduction in luminescent signal was calculated in each assay by considering the luminescent signal shown in reactions belonging to two groups of controls. In each reaction, the control group in which inhibition by the 1xCohen competitor oligonucleotide was absent was considered to represent the total luminescent signal with 0% percentage decrease, while reactions in which the 1xCohen competitor oligonucleotide inhibited YB-1 activity were considered to represent 100% signal decrease.

<table>
<thead>
<tr>
<th>Compound</th>
<th>The percentage of luminescent signal decrease in the initial luciferase reporter gene assay primary screen</th>
<th>The percentage of luminescent signal decrease in the follow-up luciferase reporter gene assay screen</th>
<th>The percentage of luminescent signal decrease in the AlphaScreen assay secondary screen</th>
<th>The percentage of luminescent signal decrease in the TruHits screen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>103</td>
<td>37</td>
<td>90</td>
<td>-9</td>
</tr>
<tr>
<td>2</td>
<td>103</td>
<td>74</td>
<td>58</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>104</td>
<td>102</td>
<td>81</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>101</td>
<td>107</td>
<td>148</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>96</td>
<td>75</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>99</td>
<td>98</td>
<td>67</td>
<td>-24</td>
</tr>
<tr>
<td>7</td>
<td>76</td>
<td>109</td>
<td>86</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>87</td>
<td>59</td>
<td>13</td>
</tr>
</tbody>
</table>

This screening is summarised graphically below in Figure 7-7.
7,360 small molecule compounds were screened using a luciferase reporter gene assay primary screen. Compounds from reactions that exhibited a large decrease in luminescent signal were re-tested, confirming that 251 compounds were causing a decrease in luminescent signal. These 251 compounds were testing in a secondary AlphaScreen assay screen using purified YB-1 binding to ssDNA. Of those 251 compounds, 67 compounds also decreased luminescent signal during the secondary screen. Those 67 compounds were tested using the TruHits assay to eliminate false positive compounds. 8 compounds remained after screening with the TruHits assay.

7.5. Computational filtering

Although the 8 hit compounds were screened for false positives detectable by the TruHits assay, this assay cannot remove all false positive compounds. For example, a compound that interferes with the anti-YB-1 antibodies in the AlphaScreen assay would not be detected by the TruHits assay; and may have also interfered with the luciferase reporter gene assay. In order to eliminate any remaining false positives, and also any compounds not suitable for
drug development, the structures of the 8 hit compounds were obtained and put through computational filters.

The first filter eliminated compounds containing substructures identified as pan-assay interference compounds (PAINS)⁵⁷, that would indicate they are likely false positives. All 8 compounds passed this PAINS filter. It is possible that testing with the TruHits kit had already successfully eliminated any false positives.

Five increasingly stringent filters were applied to eliminate groups unfavourable for drug development, such as groups with toxicity, poor pharmacokinetic behaviour, or that are highly electrophilic. Those filters, from least stringent to most stringent, were: WEHI_93K, Baell 2013 Filters 1, 2 and 3, and the CTX filter¹⁵⁸. The three least stringent filters passed 7/8 hit compounds, and eliminated one compound as overly chemically reactive. However, none of the compounds passed the two most stringent filters, Baell 2013 Filter 3 and CTX. Baell 2013 Filter 3 represents a progressive tightening of the criteria used in Filters 1 and 2. The high stringency of Filter 3 and CTX means that they should not absolutely exclude compounds from development into drug leads. While a compound that does pass these filters would be an especially good candidate for drug development, compounds that do not pass may still be suitable for development. It may be possible to find more drug-like, but still active, analogues of these compounds.

Additionally, the structures of the 8 hit compounds were assessed by a medicinal chemist from the Auckland Cancer Society Research Centre, Gordon Rewcastle (https://unidirectory.auckland.ac.nz/profile/g-rewcastle). The compound that did not pass the WEHI_93K and Baell 2013 Filters 1 and 2 computational filters was dismissed as containing a nitro group that is too chemically reactive. Three compounds were recognised as belonging to classes known to undergo colloidal aggregation, suggesting they may still be false positives despite passing the TruHits assay and the PAINS filter. One compound was identified as containing a bond vulnerable to hydrolysis, requiring modification if the compound is to progress as a drug lead. The remaining three compounds were assessed as not being likely to be residual false positives, and having structures sufficiently drug-like to warrant continued investigation as putative drug leads. The next stage will be to order stocks of these compounds and re-test them in the two YB-1 assays described here, in Auckland. If these results are reproducible, they will be tested in a broad range of in vitro assays, including general proliferation assays and YB-1-centric assays.
7.6. Summary

Variation in the quality of AlphaScreen acceptor beads increased the cost per reaction. Therefore, the in vitro cell-based luciferase assay was selected for use as the primary screen, over the AlphaScreen assay, due to its lower cost. The cell-based luciferase assay was optimised to increase signal ratio between controls, and reduce the variation within control groups. Following optimisation, 7,360 compounds were screening with hits being validated during re-testing using the cell-based luciferase assay. 251 compounds proceeded into secondary screening using the AlphaScreen and TruHits assays. The 8 compounds remaining after the secondary screen then had their structures subjected to computational filtering, and were assessed by an experienced medicinal chemist, in order to identify those suitable for further development. 3 compounds were identified as leads warranting further investigation.
Chapter 8 Concluding remarks

A large collection of melanoma genomic data of several types from published studies and publically available datasets were assembled into an easily utilised data matrix. A diverse range of data was integrated into the database, including data on drugs, clinical trials, biomarkers, druggability, literature relationships, as well as data on tumours from individual patients. Statistical analyses were performed to find associations between gene expression and patient disease-free survival in melanoma.

The integrated database was used to gain insights into melanoma obtained from an overview of the collected melanoma samples. Mutational and expression patterns of genes in signalling pathways were analysed and visualised, at both population and individual patient levels. A sub-set of this information was incorporated into a web-accessible SQL database, MelanomaDB, which can be used to perform molecular pathway and gene set analysis of melanoma genomic data.

The database was used to identify a short list of putative melanoma drug targets. The multifunctional transcription factor YB-1 was selected from that list. Two assays were developed to detect the binding of YB-1 to DNA: a cell-based luciferase reporter gene assay, and a novel AlphaScreen assay. The assays were optimised for use in high throughput screening. Screening led to the identification of eight small molecule compounds that appeared to inhibit the binding of YB-1 to its DNA binding sites in both assays. Three of these compounds look to be good prospects for development into lead compounds.

This project travelled an arc from data gathering and bioinformatic analysis at a computer to assay development and then screening at the lab bench. This project is an example of how biomedical research is becoming an information science. Increasingly, well-informed hypothesis generation requires consideration of a vast amount of available data. However, the project also revealed the challenges of integrating data from multiple sources over which incomplete quality control is possible. It forced consideration of the appropriate caution required when using data integrated from disparate sources, to avoid perpetuating errors inherent in subsets of the data. Finally, this project revealed how quickly biomedical science is moving. When the database generation in this project was first begun, only a fraction of the currently available on-line data integration tools were accessible by our research team and others in the scientific community. The melanoma database still remains unique and useful.
(the paper describing it has been downloaded over 600 times) and its content is currently being updated for a re-release in 2017. However, there are now several complementary tools as well as new sources of data that could be incorporated.

The three most promising small molecule compounds that appeared to inhibit the binding of YB-1 to its DNA binding sites in both HTS assays are currently being followed. Independent batches of re-synthesised compounds are being ordered to re-confirm in Auckland the results in the two assays that were performed in Shanghai. This will be followed by a range of additional assays of the effect of these compounds on YB-1 function using YB-1 transgenic and knock-out lines to assess specificity, as well as additional assays of non-specific toxicity, and RNaseq analysis of the effect of the compounds on cell lines. Should these assays suggest one or more of the compounds is favourable for further development, funding will be sought for preclinical studies such as xenograft and orthotopic tumour experiments. I recognise the rarity with which lead compounds like these pass through to the clinic, however I remain hopeful that this project may yield a clinically useful inhibitor of YB-1, or a compound that can act as a useful reagent in the laboratory to further study YB-1 function.
Appendix

R script used to associate gene expression with patient disease-free survival

```r
library(affy)
library(survival)
print("Do you wish to reuse a previously loaded data file - yes=(y)?")
Fl<-readLines(con = stdin(), n = 1, ok = TRUE)
if(Fl!="y") {
  print("CHOOSE FILE")
  FileChoice<-file.choose() # choose input file
  inputFile <- read.delim(FileChoice, skip=0, sep="\t", as.is=TRUE, header=FALSE)
  print("File is read")
  transposedInput<-t(inputFile)
  colnames(transposedInput)<-transposedInput[1,]
  TransposedInputNoHeaders<-transposedInput[-1,]
  CCC<-TransposedInputNoHeaders[,4:length(TransposedInputNoHeaders[1,])]
  time<-as.numeric(TransposedInputNoHeaders[,2])
  status<-as.numeric(TransposedInputNoHeaders[,3])
  rm(TransposedInputNoHeaders)
}
outt<-matrix(NA,ncol=length(CCC[1,]),nrow=12)
outt[1,]<-colnames(CCC)
ty<-c("10% quantile","20% quantile","30% quantile","40% quantile","50% quantile","60% quantile","70% quantile","80% quantile","90% quantile","CoxPR")
for (counter in 1:length(CCC[1,])) {
  print(counter)
  for (c2 in 1:9) {
    x<-CCC[,counter];x<-as.numeric(x)
    y<-data.frame(cbind(time,status,x))
    e<-survdiff(Surv(time, status) ~ x, data = y)
    outt[1+c2,counter]<-1-pchisq(e$chisq,1)
  }
  x<-CCC[,counter]; y<-data.frame(cbind(time,status,x))
  z<-as.matrix(y$time); y$time<-as.numeric(z)
  z<-as.matrix(y$status); y$status<-as.numeric(z)
  z<-as.matrix(y$x); y$x<-as.numeric(z)
  k<-coxph( Surv(time, status) ~ x, y)
  d<-summary(k);outt[11,counter]<-d$coef[5]
  outt[12,counter]<-ty[which.min(as.numeric(outt[c(2:11),counter]))]
}
out<-
outd<-rbind(c("gene or Principal Component","10% quantile ChiSq p val","20% quantile ChiSq p val","30% quantile ChiSq p val","40% quantile ChiSq p val","50% quantile ChiSq p val","60% quantile ChiSq p val","70% quantile ChiSq p val","80% quantile ChiSq p val","90% quantile ChiSq p val","Cox PH p val", "instance where p val is min"), t(data.frame(outt)))
write.table(outd,file="survival.txt", sep="\t", row.names=FALSE, col.names=FALSE, append=TRUE)
```

169
Perl script used to mine records from the Chembl web-site, and then extract protein database identifiers, domain accession identifiers and three scores of druggability for each domain

```
my ($outputfile) = "resultsMerged.txt";
open OUTFILE, "$outputfile" or die "Can't open $outputfile: $!";
use strict;
use warnings;
use Net::HTTP;
use LWP::Simple;
use Scalar::Util qw(looks_like_number);
use PDB;
my $ensembl;
my $successionArray;
my $succession;
my $i;
my $pdb;
my $domain;
my $drugline;
my $druglineList;
my $druglineList2;
my $druglineList3;
my $get;
my @successionArray = ("26009", "33403", "19410", "26346", "56735", "41137", "81534", "60216", "60206", "60205", "60205", "60203", "60203", "60375", "60375");
foreach ($i = 0; $i < @successionArray; $i++){
    my $succession = $successionArray[$i];
    print OUTFILE $succession;
    print OUTFILE "\n";
    $get = get("https://www.ebi.ac.uk/chembl/druggability/domain/$succession.json");
    my @pdb = $get =~ /"pdb_code":"(\d+)"/i;
    my @json = $get =~ /"domain_accessions":"(\d+)"/i;
    print OUTFILE "$domain",
    print OUTFILE "\n";
    print OUTFILE "$pdb",
    print OUTFILE "\n";
    my @drugline = $get =~ /"average_druggability_scores":\[(\d+)\]/i;
    my @druglineList = $get =~ /"domain_accessions":\[(\d+)\]/i;
    my @druglineList2 = $get =~ /"domain_accessions":\[(\d+)\]/i;
    my @druglineList3 = $get =~ /"domain_accessions":\[(\d+)\]/i;
    print OUTFILE "$druglineList",
    print OUTFILE "\n";
    print OUTFILE "$druglineList2",
    print OUTFILE "\n";
    print OUTFILE "$druglineList3",
    print OUTFILE "\n";
}
```

Perl script used to extract Ensembl gene identifiers from FASTA files

```
my $inputFile = "fa.rank3.en.hu";
my ($outputfile) = "SPDListEnsembl3.txt";
open INFILE, $inputFile or die "Can't open $inputfile: $!";
open OUTFILE, ">$outputfile" or die "Can't open $outputfile: $!";
my @rawFile = <INFILE>;
my $line;
my $label;
for ($i = 0; $i <= $#rawFile; $i++){
    $line = @rawFile[$i];
    ($label) = ($line =~ /gene:(\w+) /);
    if (defined $label){
        print OUTFILE $label;
        print OUTFILE "\n";
    }
}
```
Perl script used to extract drug identifiers from Clinicaltrials.gov records

```perl
my $inputFile = "inputForPer12.txt";
my $outputfile = "outputDrugNames.txt";
open INFILE, $inputFile or die "Can't open $inputfile: $!";
open OUTFILE, ">$outputfile" or die "Can't open $output: $!";

my @rawFile = <INFILE>;
my $line;
my @remain;
$/ = "SPLIT";

for($i = 0; $i<=$#rawFile; $i++){
    $line = @rawFile[$i];
    ($NCT,$remainder) = ($line =~ /(\S+)\t(.+)/);
    ($DrugRemainder) = ($remainder =~ /Drug: (.+)/);

    @remain = split(/Drug: /, $DrugRemainder);

    for($j = 0; $j<=$#remain; $j++){
        print OUTFILE $NCT;
        print OUTFILE "\t";

        print OUTFILE @remain[$j];
        print OUTFILE "\n";
    }
}
```

Overpage: Tables showing the trends in experimental objectives of experiments performed to optimise the conditions of the AlphaScreen assay. Indicated also are experiments that failed, and the suspected reasons for their failure.
<table>
<thead>
<tr>
<th>Experiment set</th>
<th>Experiment number</th>
<th>Oligo concentration</th>
<th>Protein concentration</th>
<th>Lysate concentration</th>
<th>Bead concentration</th>
<th>Buffer</th>
<th>Protocol/Addition order</th>
<th>Competitor</th>
<th>Experiment failed? Reason</th>
<th>Acceptor beads</th>
<th>Imbalance of components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment set</td>
<td>Experiment number</td>
<td>Oligo concentration</td>
<td>Protein concentration</td>
<td>Lysate concentration</td>
<td>Bead concentration</td>
<td>Buffer</td>
<td>Protocol/Addition order</td>
<td>Competitor</td>
<td>Experiment failed? Reason</td>
<td>Acceptor beads</td>
<td>Imbalance of components</td>
</tr>
<tr>
<td>Experiment set</td>
<td>Experiment number</td>
<td>Oligo concentration</td>
<td>Protein concentration</td>
<td>Lysate concentration</td>
<td>Bead concentration</td>
<td>Buffer</td>
<td>Protocol/Addition order</td>
<td>Competitor</td>
<td>Experiment failed? Reason</td>
<td>Acceptor beads</td>
<td>Imbalance of components</td>
</tr>
</tbody>
</table>

| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |
| Experiment set | Experiment number | Oligo concentration | Protein concentration | Lysate concentration | Bead concentration | Buffer | Protocol/Addition order | Competitor | Experiment failed? Reason | Acceptor beads | Imbalance of components |

172
Experiment failed? Reason: degraded aliquot of protein
Follow up on screen hits
Protocol/Addition order
Buffer
Lysate
Bead concentration
Oligo concentration
Protein concentration
Follow up on screen hits
Degraded aliquot of protein
Acceptor beads
Imbalance of components
Degraded aliquot of protein
Follow up on screen hits
Degraded aliquot of protein
Follow up on screen hits
References


195. Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. 2007;8:353.


266. Skakbina OV, Skabkin MA, Lyabin DN, Ovchinnikov LP. P50/YB-1, a major protein of cytoplasmic mRNPs, regulates its own synthesis. . 2004;395:93-95.


