UNIVERSITY OF Southampton

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

UNIVERSITY OF SOUTHAMPTON

Faculty of Medicine
School of Cancer Sciences

Using RNA sequencing to investigate therapeutic strategies in paediatric cancer

by

Christina Maria Putnam

BSc, MSc

ORCiD: 0009-0005-8936-9481

A thesis for the degree of
Doctor of Philosophy

February 2026

University of Southampton

Abstract

Faculty of Medicine
School of Cancer Sciences

Doctor of Philosophy

Using RNA sequencing to investigate therapeutic strategies in paediatric cancer

by Christina Maria Putnam

Neuroblastoma and rhabdomyosarcoma are two paediatric cancers characterised by poor prognosis and high relapse rates in high-risk cases. There is a need to understand treatment resistance/relapse and find more effective therapeutic options for these patients. Neuroblastoma arises from neural crest cells of the sympathetic nervous system that show impaired differentiation, this may involve promoting differentiation in minimal residual disease to reduce the risk of relapse. Two single agent treatments have been shown to promote neuroblastoma differentiation; EZH2, a histone methyltransferase that controls essential cellular processes, and retinoic acid, currently used in standard of care treatment for neuroblastoma. Whether the combination of both treatments could enhance differentiation has not yet been explored. This research aimed to investigate mechanisms of actions of the combination of EZH2 inhibitors with retinoic acid through the analysis of bulk RNA sequencing data to determine whether the combination therapy might further promote neuroblastoma differentiation, reduce minimal residual disease and risk of relapse.
Rhabdomyosarcoma is the most common soft tissue sarcoma in children. While most tumours initially respond to treatment, relapse and acquired resistance are common, leading to poor survival outcomes. Understanding why this resistance to therapy occurs may help to predict patient response, identify targets in resistant cells and ultimately prevent relapse. This research aimed to identify and validate a molecular signature of therapy resistance for resistant rhabdomyosarcoma. This was attempted through the analysis of RNA sequencing data from models of intrinsic and acquired chemotherapy resistance, as well as using a machine learning approach. Ultimately, these signatures may be applied to patient samples to predict treatment response, to identify tumours at high risk of recurrence, and to select effective treatments for these patients.

Contents

List of Figures

List of Tables

Acknowledgements

Words cannot express my gratitude to Zoë Walters and Will Tapper for their invaluable expertise, guidance and feedback. I could not have asked for better supervisors. Their constant support, openness to questions and kindness made this PhD genuinely enjoyable. I never felt alone in the process and I feel truly fortunate to have had such thoughtful and encouraging mentors throughout this journey.

I am also very grateful to all involved in the generation of the data; this thesis would not have been possible without your hard work. Many thanks to Ian Reddin, Jane Gibson and Ian Davies for their direction, teaching and bioinformatics troubleshooting.

Thanks must also go to all the individuals I have worked with and everyone in the ITRG team, including Jack Harrington, Ben Sharpe, Debbie Glenncross, Carmen Velasco Martinez, Rosa Gomes Alves Martins, Alex Look and many, many more! I feel incredibly lucky to have worked with such a friendly and amazing group of people. Thank you for your help, advice and funny conversations- you've made this experience truly memorable.

Lastly, thank you to my friends and family for all your encouragement, particularly my mum, dad, brother and Andy. A special thanks to Pet, for your patience and for bearing with me these last few years both financially and emotionally (mostly financially, let's be honest). I couldn't have done this without your support. And of course, I would like to thank my daughter, Lily, for making me laugh throughout this process, especially with her ghastly Biorender creations. I would be remiss not to mention my dog, Cakey, for being such a great hot water bottle during long days working from home.

Chapter 0

Abbreviations

ADCC Antibody-Dependent Cellular Cytotoxicity

ARMS Alveolar Rhabdomyosarcoma

ASCT Autologous Stem Cell Transplant

ATRA All-Trans Retinoic Acid

AUC Area Under the Curve

BCV Biological Coefficient of Variation

BET Bromodomain and Extra Terminal

C13 Clone 13

C7 Clone 7

CAF Cancer-Associated Fibroblast

cDNA Complementary DNA

CDK Cyclin-Dependent Kinase

COG Children's Oncology Group

CPM Counts Per Million

CRC Core Regulatory Circuit

CSC Cancer Stem Cell

DGE Differential Gene Expression

DEG Differentially Expressed Gene

ECM Extracellular Matrix

EMT Epithelial-Mesenchymal Transition

EpSSG European Paediatric Soft tissue sarcoma Study Group

ERMS Embryonal Rhabdomyosarcoma

EZH2 Enhancer of zeste homolog 2

EZH2i EZH2 Inhibitor

FCS Functional Class Sorting

FFS Failure-Free Survival

FN Fusion Negative

FP Fusion Positive

GD2 Disialoganglioside

GCL Glutamate–Cysteine Ligase

GSEA Gene Set Enrichment Analysis

GSVA Gene Set Variation Analysis

GST Glutathione-S-Transferases

GTF Gene Transfer Format

HD High Dose

HDAC Histone Deacetylase

HRR Homologous Recombination Repair

IFO Ifosfamide

IRS Intergroup Rhabdomyosarcoma Study Group

KNN k-Nearest Neighbours

LFC Log Fold Change

mAB Monoclonal Antibodies

MCTS Multicellular Tumour Spheroid

MDR Multi-Drug Resistance

MoA Mechanism of Action

mRNA Messenger RNA

NB Neuroblastoma

NCC Neural Crest Cells

ncRNA Noncoding RNA

NER Nucleotide Excision Repair

NK Natural Killer

NKTs Natural Killer T cells

NPV Negative Predictive Value

OS Overall Survival

P Parental cell line

PDX Patient-Derived Xenograft

PCA Principal Component Analysis

PC Principal Component

PFS Progression-Free Survival

P-gp P-glycoprotein

PPV Positive Predictive Value

PPI Protein-Protein Interaction

pre-mRNA Precursor messenger RNA

RA Retinoic Acid

RFE Recursive Feature Elimination

RG Risk Group

RIN RNA Integrity Number

RMA Robust Multichip Average

RMS Rhabdomyosarcoma

RNA-seq RNA-sequencing

ROC Receiver Operating Characteristic

ROS Reactive Oxygen Species

SOC Standard Of Care

STS Soft tissue sarcoma

TAZ Tazemetostat

TME Tumour Microenvironment

VCR Vincristine

WCSS within-Cluster Sum of Squares

WGCNA Weighted Gene Co-expression Network Analysis

XGBoost eXtreme Gradient Boosting

rRNA Ribosomal RNA

Chapter 1

Introduction

1.1 Paediatric cancers

1.1.1 Prevalence

Paediatric cancers account for the second leading cause of death in children and young adolescents aged 1-14 and is the largest contributor to disease-driven deaths in children [1, 2]. In the UK, there are around 1,800 new children's cancer cases every year, equivalent to around 5 every day [3]. Incidence rates have increased from the 1990's, but have stabilised in the last decade [3]. Leukemia is the most common childhood cancer, followed by brain and spinal tumours and lymphomas, which together account for ∼two thirds of cases in the UK [3]. Soft tissue sarcomas (STSs) and NB/other peripheral nervous cell tumours are the fourth and fifth most common paediatric cancers [3]. There are approximately 80 cases of NB and other peripheral nervous cell tumours in children (aged 0-4) every year in the UK [3]. RMS is the most common STS in children, with around 80 cases in the UK each year [4].

1.1.2 Genetics

Driver mutations can be defined as mutations that drive cancer progression and confer a selective advantage to the cancer cells [5]. Driver genes contain a driver mutation that confer a selective growth advantage to cells, contributing directly to the initiation and progression of cancer [6]. Driver genes have been identified in a proportion of paediatric cancers, with main drivers including germline mutations (present in around ∼8% of children) [7] as well as copy number and structural alterations (making up 62% of somatic drivers) [8]. However, the genome of paediatric cancers are generally considered quiet compared to adult tumours, as there are significantly fewer driver mutations [9, 7]. Somatic mutations can occur due to ageing

and environmental exposure to carcinogens, both of which are reduced in children which may be an explanation as to why children have fewer somatic mutations. In many paediatric cancer cases there is no identifiable driver gene or pathway [10].

1.1.3 Outcome and prognosis

In general, the Overall Survival (OS) for children after cancer treatment is 85% [1, 11]. Overall, the death rate of paediatric cancers has declined from 6.3 in 1975 to 1.9 per 100,000 persons in 2022 [1]. This is likely due to the 84% decline in mortality of leukemia, with remission rates now between 90%–100% for childhood acute lymphoblastic leukemia [1]. However, not all paediatric cancers have seen this improvement, with acute myeloid leukaemia, high-risk NB, metastatic sarcomas and some brain tumours having an extremely poor outcomes [12]. A recent study found that RMS has the lowest survival rate in children of 71% [1], although the survival rate of relapsed and metastatic RMS is even more dismal ∼20-30% [13, 14, 15]. This PhD project focuses on two paediatric cancers, NB and RMS, where high-risk disease has a particularly poor prognosis [16, 17].

1.2 Neuroblastoma

NB is the most common extracranial solid tumour in children and accounts for around 15% of all cancer-related childhood deaths [18]. NB arises from neural crest cells of the sympathetic nervous system that show impaired differentiation and are unable to develop into mature cells. Although it can arise anywhere in the sympathetic nervous system, the most common site is the adrenal gland [19]. NB tumours are highly heterogeneous, both at the clinical and molecular level, some regress without treatment whilst others metastasise and are resistant to therapy [20]. Due to the clinical heterogeneity in NB, patients are stratified into risk groups to inform treatment decision. The Children's Oncology Group stratifies patients based on age, tumour stage and histology, *MYCN* status and tumour cell ploidy into low, intermediate and high-risk groups [21]. High-risk NB includes patients with *MYCN* amplification (except a small subset of patients with localised *MYCN* amplified disease that is completely resected) and age above 18 months with metastatic disease [22]. Approximately half of patients have high-risk disease, which has a poor prognosis. 5-year OS and relapse rates for high-risk disease are around 50% [22] compared to low-risk NB which have a 5-year OS rate of ∼98% [16].

Neural Crest Cells (NCC) are formed during vertebrate embryogenesis, where initially they are indistinguishable from other neural epithelial cells [23]. Through signal inductions, they embark in Epithelial-Mesenchymal Transition (EMT) where

they migrate to different locations in the embryo and form a wide range of different tissues and cell types including neurons, bone, cartilage and connective tissues [24]. In NB, the ability of NCC to differentiate is impaired and they fail to develop into mature cells [23] (Figure 1.1).

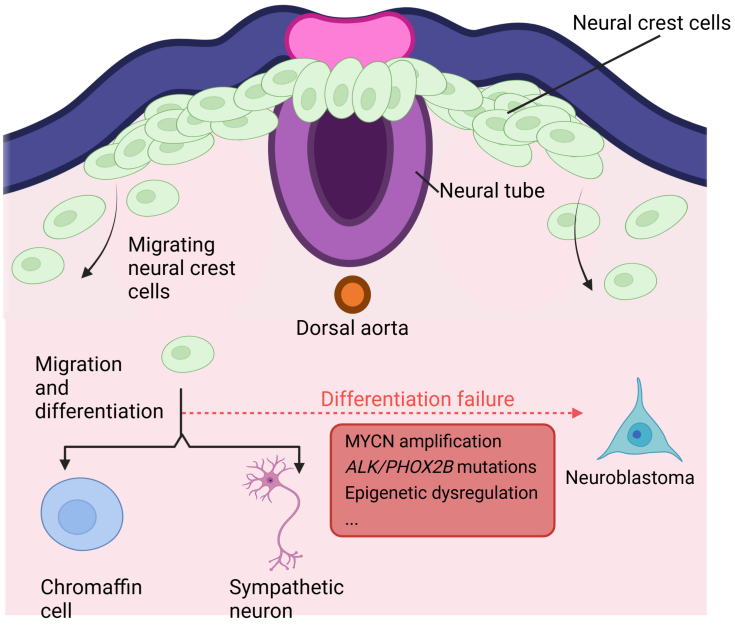


FIGURE 1.1: Development of NB from neural crest cells. Neural crest cells embark on epithelial to mesenchymal transition towards the dorsal aorta. Here they are committed to the sympathoadrenal lineage and can differentiate into chromaffin cells or sympathetic neurons. Failure to differentiate can result in the development of neuroblastoma. In the red box are the most frequent molecular factors involved in neuroblastoma development. Created with BioRender.com.

1.2.1 Underlying genetics

Familial NB makes up around 1-2% of NB cases, with the most common predisposition genes being *ALK* or *PHOX2B* but mutations in these can also occur in sporadic NB [25]. In sporadic cases, chromosomal aberrations including hemizygous deletions of chromosomes 11q and amplification of *MYCN* on 2p are common. 11q deletions are more common than *MYCN* amplifications and the two are almost mutually exclusive [26]. 11q deletions are present in between 35-45% of cases and are associated with later stage disease [26, 27]. Either the entire chromosome may be deleted (∼45% of cases), or there may be partial loss of the chromosome (∼55% of cases) [28]. Several genes have been proposed as contributing to tumourigenesis, including *CADM1*, *ATM* and *H2AFX*; many of these candidate genes are involved in cell growth regulation and DNA damage repair [27]. Amplifications of the *MYCN* gene occurs in around 25% of NB cases and is associated with high-risk NB and poor prognosis. *MYCN* encodes for a nuclear DNA binding protein that acts as a transcription factor, regulating target genes involved in the cell cycle. *MYCN*

amplification in NB tumourigenesis is thought to contribute to apoptosis resistance, suppression of differentiation, immune evasion and metastasis [29, 30]. Aside from these genetic aberrations, there are few recurrently mutated genes in NB, of these, *PTPN11*, *ALK* and *ATRX* are the most common, making up around 3%, 10% and 10% of cases respectively [31]. Out of these genes, *ALK* was the only one found to be associated with clinical outcome, with mutated *ALK* associated with decreased OS probability [31]. *ATRX* mutations were found to be mutually exclusive with *MYCN* amplification [31].

1.2.2 Standard of care treatments in neuroblastoma

Treatment for high-risk NB currently consists of induction, consolidation, and maintenance therapy [32] (Figure 1.2). The induction phase involves multiagent chemotherapy, surgical resection and stem cell collection and aims to reduce tumour burden [32]. In North America, chemotherapy consists of 5 cycles; topotecan/cyclophosphamide for the first two cycles, cisplatin/etoposide for the 3rd and 5th cycle and Vincristine (VCR)/doxorubicin/cyclophosphamide for cycle 4 [32]. In Europe, patients receive a treatment regime of rapid cisplatin, VCR, carboplatin, etoposide, and cyclophosphamide (COJEC). This is where VCR/carboplatin/etoposide, VCR/cisplatin, and VCR/etoposide/cyclophosphamide is given every 10 days [32]. Resection of the primary tumour is carried out near the end of chemotherapy [33]. Stem cells are taken ready for Autologous Stem Cell Transplant (ASCT), that occurs in the consolidation phase. Consolidation phase aims to limit remaining minimal residual disease and involves high-dose chemotherapy (e.g. busulfan/melphalan), ASCT followed by radiotherapy [32]. Any residual disease that remains is then targeted by maintenance therapy, which includes the differentiating agent isotretinoin, or 13-cis-Retinoic Acid (RA), and anti-Disialoganglioside (GD2) immunotherapy [32]. Isotretinoin is a vitamin A derivative that promotes differentiation of neuroblasts to mature neurons in NB, resulting in cell cycle arrest and suppressing tumour growth [34]. Treatment with isotretinoin is designed to promote differentiation of any remaining residual disease [34]. Clinical trials found that in patients without progressive disease, the addition of isotretinoin to the treatment regime significantly improved 3-year Failure-Free Survival (FFS) in comparison to patients who received no further treatment [35]. Anti-GD2 immunotherapy involves targeting disialoganglioside with monoclonal antibodies (Figure 1.3). Anti-GD2 acts by numerous mechanisms, including infiltration by macrophages resulting in activation of complement and cell lysis, Antibody-Dependent Cellular Cytotoxicity (ADCC) involving natural killer cells [36].

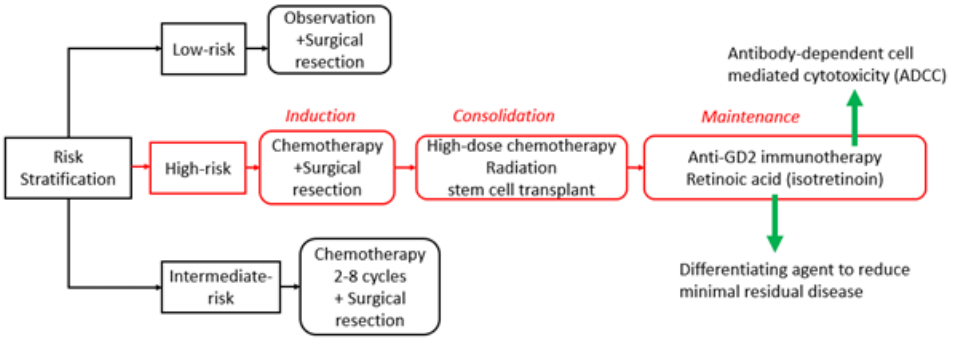


FIGURE 1.2: Standard of care treatment for neuroblastoma for low-risk, intermediate-risk and high-risk patients. Treatment for high-risk patients consists of 3 stages; induction, consolidation and maintenance. The consolidation and maintenance phases aim to remove any remaining residual disease that remains after prior treatment. The green arrows show how each aspect of the maintenance therapy achieves this. Figure provided by Dr Andy Gao.

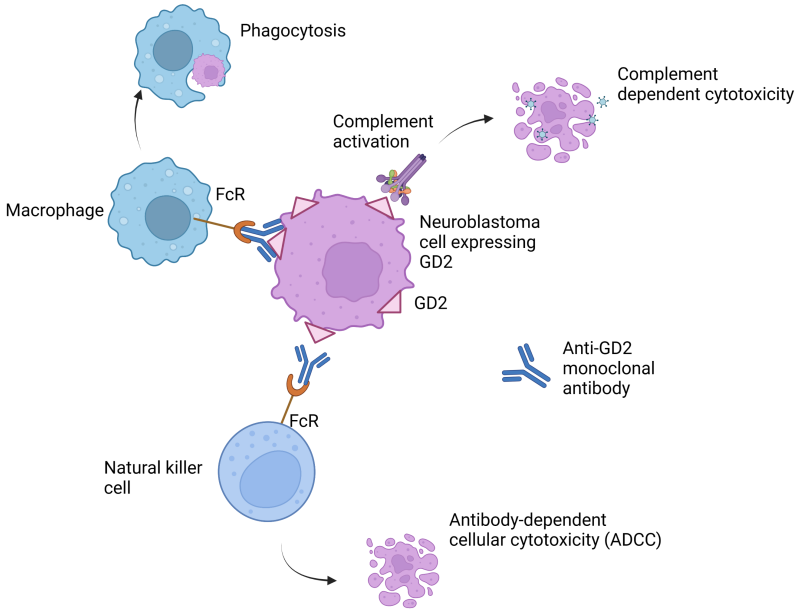


FIGURE 1.3: Mechanisms of anti-GD2 therapy in neuroblastoma. Showing activation of macrophages resulting in phagocytosis of neuroblastoma cells, complement dependent cytotoxicity and ADCC by natural killer cells. Created with BioRender.com.

1.2.3 Prognosis

The prognosis for high-risk NB is poor, with 5-year FFS around 50% [16]. Despite multimodal therapy, relapse is common for high-risk NB patients (∼50%) with most relapses occurring within two years of diagnosis [37]. Only around 20% of patients show complete response to induction therapy [38]. 10-15% will progress during induction therapy, with another 40% progressing even after initial response [38]. High-risk NB survivors may have lifelong effects from intensive treatment. The

majority of patients treated with multimodal therapy experience endocrine-related late effects including thyroid dysfunction, poor linear growth and impaired fertility [39]. Other common adverse conditions include hearing loss, pulmonary dysfunction and an increased risk for subsequent malignant neoplasms [39]. Due to high tumour heterogeneity, poor outcomes, resistance to therapy and the toxicity of current treatments, targeted therapies in NB are being explored to overcome this [39].

1.2.4 Differentiation and neuroblastoma

NB tumours can be histologically classified into three subtypes depending on their degree of differentiation (Figure 1.4). The undifferentiated subtype is rare, considered unfavourable in all patients and is often associated with other poor prognostic factors. FFS for the undifferentiated subtype is 50% [40]. Poorly differentiated NB is the most common subtype, where cells show varying degrees of neurites [41]. This subtype is considered unfavourable for individuals older than 18 months but may be favourable in children younger than 18 months old [21]. The differentiating NB subtype shows lots of neurite production and the majority of these tumours are favourable [21].

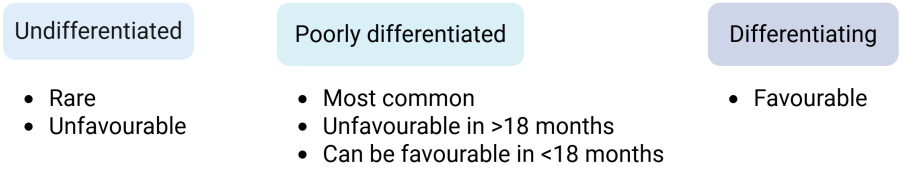


FIGURE 1.4: Summary of different neuroblastoma subtypes; undifferentiated, poorly differentiated and differentiating. Figure adapted from Shimada and Hiroyuki [41]. Created with BioRender.com.

NB cells can be also categorised into two distinct lineage states (Figure 1.5). These are (i) adrenergic cells that are more differentiated, show a more neuronal sympathetic identity and express markers including PHOX2B, TH, and DBH or (ii) mesenchymal cells; undifferentiated neural crest cell-like that express markers including vimentin or fibronectin [42]. These two NB cell identities are able to inter-convert through epigenetic regulation [42]. NB tumours were found to contain both mesenchymal and adrenergic cells, with a slight bias towards the adrenergic phenotype [42]. It has been proposed that this plasticity of NB can drive therapeutic resistance and intratumoural heterogeneity [43]. The mesenchymal phenotype has been shown to be more resistant to Standard Of Care (SOC) chemotherapy agents than adrenergic cells *in vitro* and were found to be enriched in post-chemotherapy and relapse samples [42]. This suggests there may be a link between the mesenchymal phenotype and recurrent disease. There is also evidence that cells with mesenchymal lineage may show resistance to differentiating agents, as they do not undergo differentiation in response to the differentiating agent All-Trans Retinoic Acid (ATRA) (the active metabolite of

Vitamin A) in the same way as adrenergic NB cells [20]. The mesenchymal phenotype has also been linked to low GD2 expression and a poor response to anti-GD2 immunotherapy [44]. However, mesenchymal NB cells have been shown to be more immunogenic, promoting T cell infiltration by secreting inflammatory cytokines increasing MHC-I expression [45]. They were found to be targeted by cytotoxic T and natural killer cells [45].

MYCN is has also been shown to play a role in NB differentiation, with reduced expression associated with a more differentiated phenotype [46]. *MYCN* knockdown in neural crest stem cells has been associated with an increase in key sympathetic neuron differentiation transcription factors [47]. Additional studies have shown that treatment with isotretinoin results in reduced *MYCN* expression [48], with a recent study showing in both *MYCN*-amplified and non-amplified cell lines treatment with isotretinoin significantly reduced the expression of *MYCN* [49].

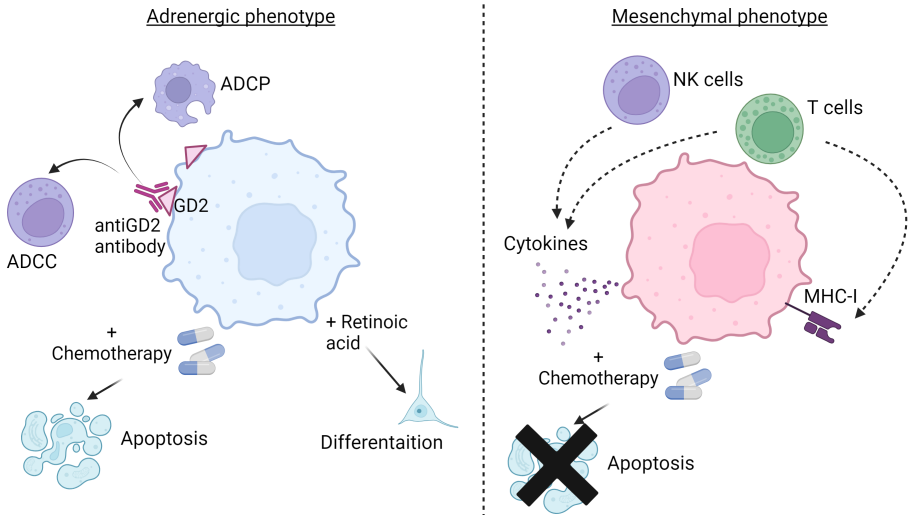


FIGURE 1.5: Differences between the adrenergic and mesenchymal phenotype in NB cells. Adrenergic cells are reported to be more responsive to chemotherapy and differentiate in response to retinoic acid derivatives. They respond to anti-GD2 immunotherapy and undergo antibody-dependent cell cytotoxicity or antibody-dependent cellular phagocytosis. Mesenchymal cells are more chemoresistant and do not undergo differentiation in response to retinoic acid derivatives. However, they are more immunogenic, showing increased MHC-I expression and can secrete inflammatory cytokines for NK and T cell infiltration and killing. Created with BioRender.com.

1.2.5 Targeted therapies in neuroblastoma

1.2.5.1 Targeting MYCN in neuroblastoma

The role of MYCN in NB tumourigenesis is well established and there have been many attempts to inhibit the action of MYCN both directly and indirectly through its downstream targets [50]. The expression of cell cycle protein AURKA has been linked

to poor prognosis in NB [51]. AURKA has been shown to form a complex with MYCN and stabilise the protein, disruption of this complex leads to MYCN degradation [52]. AURKA inhibitors have been shown to inhibit tumour growth in NB mouse models and showed good response rates in phase 1 clinical trials in combination with irinotecan and temozolomide, but not as a single-agent therapy [53]. Targeting epigenetic factors including Bromodomain and Extra Terminal (BET) transcription factors, is another approach that has been investigated to target MYCN. BET inhibitors suppress the transcription of *MYCN* and have been shown to suppress NB tumour growth *in vivo* but have not yet been explored in clinical trials for younger children [54]. Another approach tried is to target polyamines, which maintain NB phenotype by supporting MYCN activities. Ornithine decarboxylase 1 (ODC1) is the enzyme that is the rate limiting step in polyamine synthesis, and thus inhibiting this protein can help to prevent the activities of MYCN [55]. In NB mouse models, ODC1 inhibitors delayed tumour initiation and showed enhanced effects in combination with chemotherapy, increasing survival [56].

1.2.5.2 Targeting immune therapies in NB

With anti-GD2 therapy in clinical use, there is ongoing research to attempt to maximise the effectiveness of this therapy as well as reducing it's toxicity. Combination therapies with anti-GD2 have been investigated to improve it's effectiveness, including a study by Theruvath et al. who found combination of anti-GD2 and anti-CD47 significantly reduced tumour burden and extended survival in xenograft mouse models [57]. Mabe et al. focused on targeting mesenchymal cells that show low expression of GD2, showing they can be sensitised to anti-GD2 antibody using Enhancer of zeste homolog 2 (EZH2) inhibitors in mice models, to shift cells from a mesenchymal to adrenergic state [44]. The effectiveness of anti-GD2 in combination with chemotherapy has also been investigated in clinical trials and showed promising results [58]. Another approach being investigated is augmentation of anti-GD2 with other immune cells. Heczey et al. looked at enhancing the anti-tumour properties of Natural Killer T cells (NKTs) with chimeric antigen receptors which is currently undergoing clinical trials using NKTs co-expressing a GD2-specific chimeric antigen receptor with interleukin 15 [59]. Clinical trials are also looking at improving Natural Killer (NK) function by transplantation of stem cells from haploidentical family donors [60]. In addition to augmenting anti-GD2 therapy, other research by Evers et al. has focused on minimising the effect of pain from anti-GD2 therapy, by using an alternative isotype instead of the IgG1 isotype currently used [61]. The reformatted GD2 antibody, IgA1, was found to reduce pain in mice as well as improving neutrophil-mediated lysis.

Targeted immunotherapies, including immune checkpoint therapy, have also been explored in NB. PD-L1 binds to its receptor PD1 to inhibit T cell response (Figure 1.6). Blockade of this interaction with anti-PD-1/anti-PD-L1 Monoclonal Antibodies (mAB) restores T cell reactivity and is an effective treatment in many cancer types [62]. In NB, monotherapy with anti-PD1/anti-PD-L1 has shown poor efficacy, however in combination with other therapies results are more promising [63]. Combination of anti-PD-1/anti-PD-L1 with an anti-CD4 mAB showed a synergistic effect in NB mouse models [63]. Furthermore, dual immune checkpoint blockade of anti-PD1 and anti-CTLA-4 activates T cells and tumour associated macrophages *in vivo*. If dual immune checkpoint blockade was treated after induction chemotherapy, there was a significant survival benefit in NB mouse models [64]. Currently there is a clinical trial looking at the effect of combination of anti-PD1 with anti-GD2 in generating anti-neuroblastoma immunity [65].

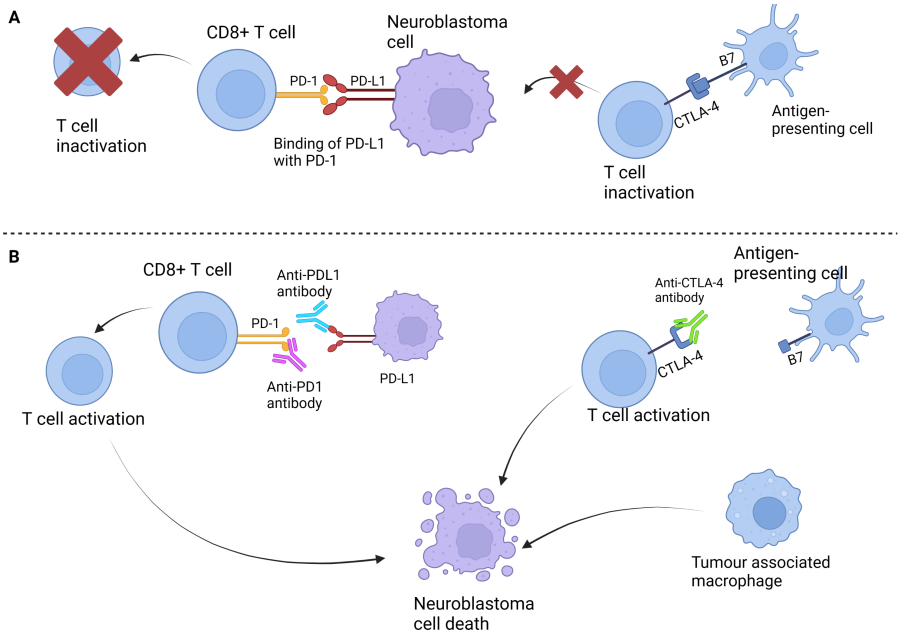


FIGURE 1.6: Anti-PD-1/anti-PD-L1 and anti-CD4 therapies in neuroblastoma. A) Binding of PD-L1 with PD-1 results in the T cell inactivation. Binding of CTLA-4 with B7 results in T cell inactivation. B) Binding of anti-PDL1 antibody with PDL-1 or anti-PD1 antibody with PD-1 restores T cell activation leading to NB cell death. Binding of CTLA-4 antibody blocks binding with B7, resulting in T cell activation, recruitment of tumour associated macrophages and NB cell death. Created with BioRender.com.

1.3 Rhabdomyosarcoma

Soft tissue sarcomas (STSs) are a group of rare cancers that resemble cells from mesenchymal origin including, fat, muscle, nerves, fibrous tissues or blood vessels. STSs represent about 1% of all malignant tumours and around 7% of cancer in children [66, 67]. RMS is the most common soft tissue sarcoma in children under 10

years of age, making up around 50% of cases [15]. RMS resembles skeletal muscle myoblasts, varying in degrees of differentiation depending on the subtype [68]. Clinically, it presents as an expanding mass with the location of the tumour varying by subtype [69]. The World Health Organization classification currently recognises four distinct subtypes of RMS [70]. ERMS typically affects younger children, lacks presence of the fusion gene and has a better prognosis while ARMS is common in patients aged 0-19 years, encompassing both children and adolescents, and is more aggressive with poorer outcomes [71]. ERMS is the most common subtype followed by ARMS, accounting for approximately 60-70% and 30-40% of RMS respectively [71, 72]. Spindle cell/sclerosing and pleomorphic are rarer subtypes. Spindle cell/sclerosing RMS was previously categorised as a subtype of ERMS, but is now recognised as a distinct subtype [73]. It is almost exclusive in children under 5 years old, mainly presenting in infants under 3 years of age and makes up around 10% of cases in infants [74, 75]. Pleomorphic RMS occurs predominantly in adults ($\sim$20% of adult cases) with the majority of cases diagnosed after the age of 40 [76, 77]. The different subtypes of RMS show variation in genetic aberrations, histology, prognosis, risk of metastasis and location of tumour [78]. ARMS can be further divided by the presence or absence of the fusion protein, into either FP- or FN- RMS (Figure 1.7). A chromosomal rearrangement involving either *PAX3* or *PAX7* and *FOXO1*, which fuses the DNA binding domain of PAX and the transactivation domain in FOXO, is found in the majority ($\sim$70%) of ARMS cases [79, 80]. FN-ARMS clinically and molecularly resemble ERMS tumours which are FN [81]. RMS is now classified based on fusion gene status rather than histology, as fusion status provides a more accurate prognostic marker [72]. ERMS typically arises in the head, neck and genitourinary tract and ARMS mainly arises in the extremities; but RMS can arise at virtually any site [15]. RMS is more common in children than adults, estimated to make up around 3-4% of paediatric cancer whereas it only accounts for less than 1% of adult malignancies [82]. It has an incidence rate of 4.4 cases per one million [83].

1.3.1 Risk Stratification

The European Paediatric Soft tissue sarcoma Study Group (EpSSG) stratifies RMS into four risk groups (low, standard, high and very high) based on TNM pre-treatment staging system, post-surgical Intergroup Rhabdomyosarcoma Study Group (IRS) clinical group and a pathological grouping system [84]. One study found that when looking at all subtypes, 10- year OS was $\sim$65% and 10-year FFS was $\sim$53% [85].

The TNM cancer staging system describes the primary tumour size and site, whether there is regional lymph node involvement and the presence or absence of metastasis. Increasing tumour size was found to be a prognostic factor for poor survival in RMS, with tumours $>$5 cm having a poor prognosis [85]. Regional nodal status has also

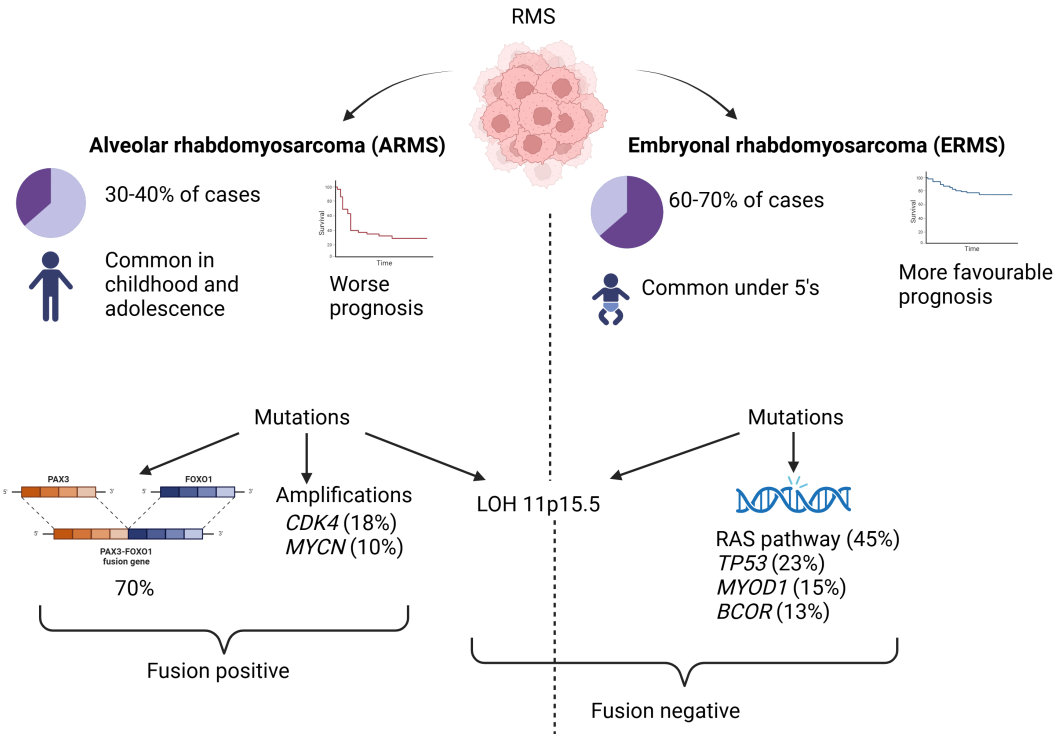


FIGURE 1.7: Summary of the key differences between alveolar and embryonal RMS. ARMS is the less common subtype but is associated with a worse prognosis. The majority of ARMS harbour the PAX3-FOXO1 fusion gene, known as fusion positive RMS, but otherwise have a quiet genomic landscape. If the fusion gene is not present this is known as fusion negative RMS and is more genetically more similar to ERMS. Created with BioRender.com.

been shown to have prognostic significance [86]. Favourable tumour sites include the orbit, non-bladder prostate genitourinary tumours and liver whereas unfavourable sites include bladder, prostate and extremities [87]. Just over half of tumours occur at favourable sites [71]. There is a significant association between absence or presence of metastasis and outcome [88]. In one study, absence of metastasis at diagnosis was shown to have an OS of 82.7% whilst 77.4% of individuals who had metastasis died [89]. 3-year OS and FFS of metastatic RMS have been reported around 34% and 27% respectively [90], but a more recent study found that 5-year OS and FFS rates were 30% and 42% [17]. ARMS is associated with higher risk of metastasis; 54.5% of patients with alveolar subtype and other subtypes had metastasis at the time of diagnosis, while only 30.9% of embryonal did [90].

IRS clinical group is determined after surgical resection to describe the completeness of the tumour resection and evidence of lymph node involvement and metastatic disease after examination of specimens. If complete resection is achieved at all sites, survival is improved [85]. IRS group has also been found to be a prognostic indicator, with higher IRS groups associated with significantly worse 10-year OS [85].

Previously, histologic subtype of RMS was used to stratify risk groups, however

evidence suggests fusion status is a better prognostic factor and EpSSG guidelines suggest that future stratification should be based on this [91]. As well as this, they also proposed DNA sequencing of high risk genes. This includes *MYOD1*, encoding for a key protein involved in muscle development and the tumour suppressor gene *TP53*. Although rare, the L122R mutation in *MYOD1* was found to enable a MYC-driven transcription program and was associated with aggressive tumours and poor outcome. This mutation is also associated with the spindle cell/sclerosing subtype of RMS [92]. Identification of this mutation would allow consideration to escalate treatment intensity. They proposed that screening of germline TP53 mutations to identify children with an increased risk of developing RMS. ERMS has a more favourable prognosis than other subtypes [93, 94]. Age is another prognostic factor used to stratify risk groups. Patients aged 1-9 years have a better prognosis than those over 10 years of age [85].

Guidelines developed by the Children's Oncology Group (COG) differ slightly to EpSSG, classifying into 3 groups (low, intermediate and high) instead of four [95]. It considers IRS post-surgical clinical group, TNM stage and fusion status. Unlike EpSSG classification, an overall number is assigned to TNM stage combining information on the tumour size and site, node involvement and metastasis, instead of them being treated as individual components. Another difference between classifications is that node involvement and fusion-positive RMS may be classed as intermediate risk in COG guidelines, whereas in EpSSG any node positivity or ARMS is considered high-risk.

1.3.2 Standard of care treatment

Treatment for RMS consists of a combination of chemotherapy, surgical resection and radiation therapy and is based on risk stratification. Surgical resection is performed if the tumour can be completely excised without causing significant organ or functional impairment [96]. In North America, standard chemotherapy is a combination of VCR, actinomycin-D and cyclophosphamide whereas in Europe, it is Ifosfamide (IFO), VCR and actinomycin-D [96] (Figure 1.8). No significant differences have been reported in outcomes between different chemotherapy treatment regimes [15]. VCR is a vinca alkaloid that inhibits mitosis by blocking polymerisation of mitotic spindle microtubules [97]. Actinomycin-D inhibits RNA synthesis by binding to DNA and inhibiting RNA polymerase [98]. IFO and cyclophosphamide are both prodrug alkylating agents that are metabolised into their cytotoxic species, IFO mustard and phosphoramide mustard [99]. They act by forming crosslinks between and within DNA strands which then induces apoptosis [99]. Cyclophosphamide acts through inhibiting DNA replication and RNA synthesis [100].

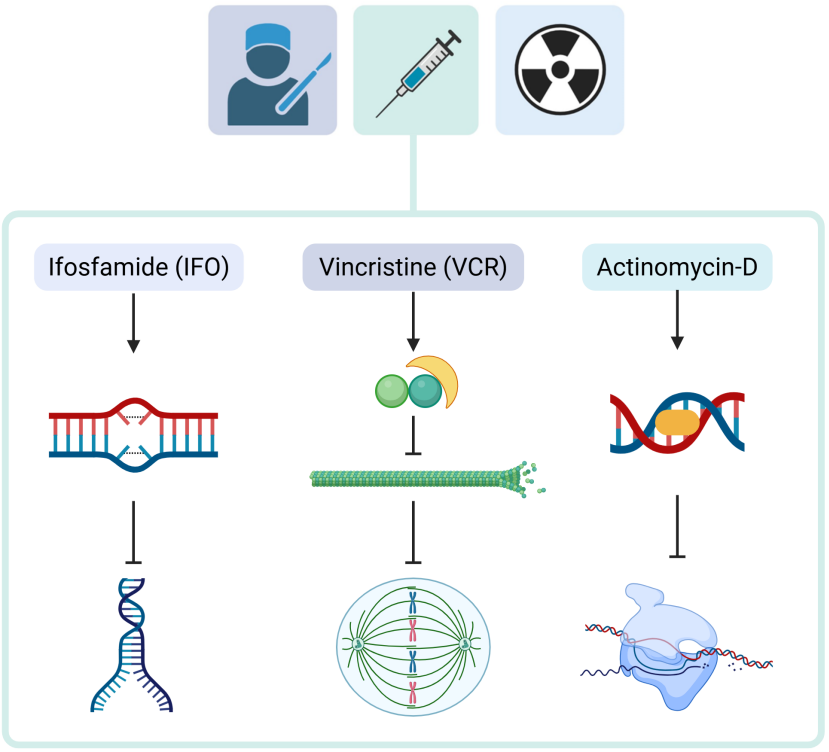


FIGURE 1.8: Standard of care treatment for RMS with chemotherapeutic agents used and their mode of action. The active form of ifosfamide alkylates DNA resulting in crosslinks. These crosslinks distrupt DNA replication and transcription. Vincristine binds to tubulin dimers, preventing the polymerisation of tubulin and formation of the mitotic spindle. Actinomycin-D binds to DNA and blocks the progression of RNA polymerase, inhibiting RNA synthesis. Created with BioRender.com.

1.3.3 Prognosis of high-risk RMS

Outcome of RMS has improved over the past 10 years, with the current combination treatment approach of chemotherapy, radiotherapy and surgery resulting in around 70% of individuals achieving long-term survival [101]. However, despite research into new chemotherapeutic agents and therapies in high-risk RMS, there has been no major advances in the treatment of high-risk disease [102]. Intensifying the dose of chemotherapy was shown to have no improvement on patient outcome for metastatic RMS [103]. Recently, the addition of low-dose maintenance chemotherapy to standard treatment improved patient outcomes in localised high-risk RMS, increasing 5-year OS from 74% without maintenance therapy to 87% [104]. Although the majority of tumours initially respond to treatment (∼85%) [105], relapse is common [106]. Around one third of localised and two thirds of metastatic RMS relapse [107] and post-relapse outcome is extremely poor (∼20%) [13, 14]. Resistance to therapy in RMS is responsible for a large part in the failure of SOC treatments [108].

1.3.4 Underlying genetics

The fusion gene encodes a chimeric transcription factor which has increased transcriptional power and alters expression of downstream transcriptional targets [109]. There are many transcriptional targets which appear to be involved in myogenic differentiation, myogenic signalling and mesodermal development [109]. FP tumours have a significantly lower mutational burden compared to FN tumours [110]. The most common genetic alterations in FP tumours aside from the fusion gene are amplifications in *MYCN*, *CDK4* and *MIR17HG* [110]. Loss of heterozygosity at 11p15.5 is seen in both FP and FN tumours (∼50%), where the *IGF2* gene lies. It is present in around 70% of ARMS and is associated with a poor prognosis [81].

FN tumours have a distinct genomic landscape compared to FP-RMS [111]. The majority of mutations are in pathways, with only a few single gene mutations at a low frequency. Alterations affecting the receptor tyrosine kinase/RAS/PIK3CA axis were found to be the most common mutations in FN-RMS, with 93% of mutations affecting this axis [110]. Activating mutations in RAS pathway are present in approximately 45% of cases, including *NRAS*, *KRAS*, *HRAS*, *NF1* and *PIK3CA* [110]. The most common single gene mutations in FN tumours include the transcriptional repressor *BCOR* (7.4% of FN tumours), *FBXW7* (7.4%) and *TP53* (5.3%) [110]. The most common copy number alterations in FN-RMS are *CDKN2A* (15%) and *CDKN2B* (13%) [111].

1.4 Therapy resistance in cancer

Resistance to treatment in cancer is responsible for up to 90% of cancer-related deaths [112]. Drug resistance in cancer is when the tumour becomes tolerant of treatment and promotes drug degradation and inhibition [113]. Resistance may be intrinsic, acquired or a combination of both. Intrinsic resistance can be defined as pre-existing resistance in the tumour before therapy and can be a result of genetic mutations present in the tumour, absence or low expression of the target, intratumoural heterogeneity that is selected for during treatment or activation of protective pathways [114]. Acquired resistance is when a tumour initially responds to treatment but then becomes insensitive, due to events including activation of a second proto-oncogene, altered expression of the drug target or changes in the Tumour Microenvironment (TME) [115].

1.4.1 Mechanisms of therapy resistance

1.4.1.1 Cancer stem cells

The existence of intratumoural genetic heterogeneity leads to variation in tumour cells sensitivity to therapy due to differences in genomic, transcriptomic, proteomic and epigenetics make-up [116]. Tumour cells that are resistant can be selected for and proliferate after treatment, leading to disease recurrence. This includes Cancer Stem Cells (CSCs)- a population of cells that exist within the tumour and have special mechanisms that promote drug resistance. They also show properties similar to normal stem cells, including self-renewal, the ability to differentiate and promote tumourigenicity [117]. These populations may remain stable during treatment or metastasise to another location and cause recurrence.

1.4.1.2 Altered drug metabolism and transport

Drug metabolism, including uptake, efflux and detoxification, is one mode of therapy resistance in cancer (Figure 1.9). Modified activity or reduced expression of surface transporters can reduce the effects of the drug by preventing entry to the cell or preventing the transmission of drug effect [112]. Elevated drug efflux is thought to be one of the major causes of chemotherapy resistance, with increased expression of members of the ABC transporter superfamily the most common cause [118]. P-glycoprotein (P-gp), encoded by the *ABCB1* gene, functions as a drug efflux transporter for several chemotherapeutic agents, including vinblastine, doxorubicin and VCR. Elevated expression of P-gp has been associated with the development of drug resistance in several cancers [119]. Although some studies have reported high expression of P-gp in RMS [120, 121] and NB [122], other studies have reported conflicting results [123, 121]. Certain drugs require metabolic activation to become therapeutically effective; consequently, mutations or downregulation of proteins involved in these pathways can cause drug inactivation, representing another mechanism of acquired resistance.

1.4.1.3 Impaired drug activation

Whilst targeted therapies are more selective than chemotherapies, there is still a risk of resistance by alterations in the target in both. This may be achieved by downregulation of the drug target or mutations that result in an altered target [124, 125]. In cases where the target is a component of a pathway, mutations in other parts of the pathway may result in activation of the pathway and cause resistance [126].

1.4.1.4 Enhanced DNA damage repair

Many cancer therapies exert their effects by causing DNA damage, either directly or indirectly, to induce cell death [113]. Tumour cells may show enhanced DNA damage response to overcome the damage inflicted by cancer therapies, eventually leading to resistance. For example, cisplatin induces DNA crosslinking damage to promote apoptosis [127]. DNA damage repair systems, including the Nucleotide Excision Repair (NER) system and Homologous Recombination Repair (HRR) mechanisms, can lead to resistance to this drug. The NER system is able to repair intrastrand DNA crosslinks induced by some chemotherapy drugs, that would otherwise result in inhibition of replication and apoptosis. HRR can repair double strand breaks induced by platinum-based chemotherapy agents. Efficacy of cisplatin is dependent on the inadequacy of DNA damage response mechanisms in the tumour cells.

1.4.1.5 Evasion of apoptosis

The majority of cancer therapies rely on the induction of apoptosis to eliminate malignant cells. Apoptosis can occur through two different pathways; intrinsic and extrinsic. The intrinsic pathway mediated by mitochondria, involves anti-apoptotic proteins BCL-2 and AKT and pro-apoptotic proteins BAX, BAK and caspase-9, whereas the extrinsic pathway involves death receptors such as TNF recptors and TRAIL.[128]. Both pathways result in the activation of caspase-3 which causes apoptosis. Upregulation of anti-apoptotic genes and downregulation of pro-apoptotic genes can allow tumour cells to avoid apoptosis and resist therapies.

1.4.1.6 Tumour microenviroment and resistance

The TME is the extracellular environment around the tumour consisting of the Extracellular Matrix (ECM), immune cells, blood vessels, cancer-associated fibroblasts and signalling molecules. Nearly all components of the TME interact with the tumour and can affect survival, heterogeneity and drug resistance [129]. For example, tumour-associated macrophages, tumour-associated neutrophils and myeloid-derived suppressor cells are some types of myeloid cells that may contribute to therapeutic resistance [130]. tumour-associated macrophages are recruited to the tumour by chemokines and are predominately differentiated into M2 macrophages [131]. They can induce EMT by releasing TGF-β, TNF-α and proteases including matrix metalloproteinases to activate EMT signalling pathways [131]. Tumour-associated macrophages may also release immunosuppressive factors that may induce therapeutic resistance. Cancer-Associated Fibroblasts (CAFs) are a sub-population of cells in the tumour stroma and activation of these promotes oncogeneic signals and

resistance in many cancer types including colorectal cancer [132], gastric [133], and bladder cancer [134]. When prostate tumour cells were grown in co-culture with CAFs, a weakened response to doxorubicin was observed due to obstruction of DNA damage [135]. The vasculature system is also a way that the TME may contribute to resistance by helping to create an acidic and hypoxic environment through insufficient supply of nutrients and oxygen [129]. Hypoxic zones can inhibit tumour cell proliferation and obstruct the delivery of chemotherapy drugs, contributing to therapy resistance and recurrence [136].

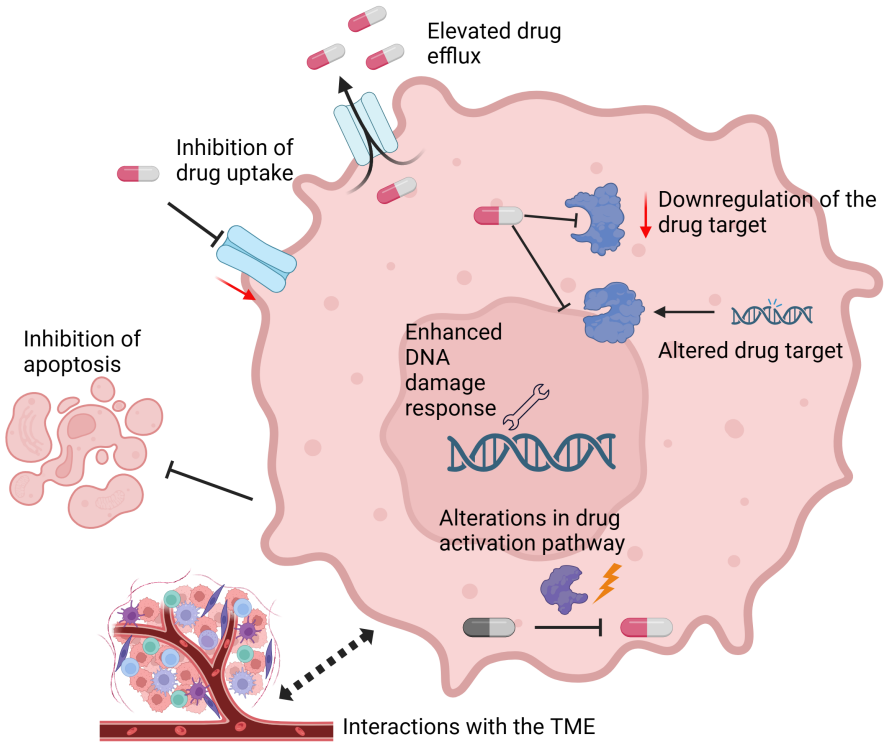


FIGURE 1.9: Summary of mechanisms of drug resistance in cancer. This includes alterations in drug metabolism such as inhibition of drug intake and elevated drug efflux; alterations in the drug target through mutations or downregulation of protein expression; enhanced DNA damage response; alterations in the drug activation pathway; inhibition of apoptosis and interactions with the TME. Created with BioRender.com.

1.4.2 Resistance in paediatric cancer

One of the main challenges in paediatric cancer is resistance to therapy [137]. Whilst the majority of paediatric cancers initially respond to therapy, relapse is frequent and post-relapse survival is extremely poor [138]. In addition, the application of immunotherapies, which have become a promising therapeutic option for adult cancers, has progressed more slowly in paediatric malignancies [139, 140]. The efficacy of immunotherapies in paediatric cancer treatment is highly varied and the underlying reasons for their success or failure remain poorly understood [141, 142].

Gaining deeper insight into the mechanisms of resistance and overcoming them is critical to improving patient outcomes in paediatric cancers.

1.4.2.1 Therapy resistance in NB

In NB, over 50% of patients with high-risk disease are either inherently resistant to chemotherapy, or relapse after treatment [143]. One possible mechanism contributing to this is Multi-Drug Resistance (MDR) via moderations in drug export through the ABC transporter superfamily; proteins that are involved in the transport of substances across the membrane. The multidrug resistance–associated protein MRP1 is an ATP efflux pump for chemotherapeutic agents VCR, doxorubicin, and etoposide. High expression of *MRP1* is associated with poor prognosis in NB and MRP1 reversal agents have been shown to sensitise cells to chemotherapy [144]. Another study showed that *ABCC1* and *ABCC4* are transcriptionally controlled by MYCN and high *ABCC1* expression is associated with poor prognosis in NB [145, 146]. Although elevated levels of the drug efflux protein P-gp have been reported by some studies [122], other studies have found its prognostic relevance unclear [123]. Together, these findings suggest that ABC transporters may play a role in MDR to chemotherapy in NB by elevating drug efflux.

Another mechanism that has been explored as possibly contributing to chemoresistance in NB is autophagy. One study showed that autophagy levels increased after chemotherapy *in vitro* and *in vivo*, and inhibition of autophagy sensitised cells to chemotherapy [147]. Another study also found that a different autophagy inhibitor was able to sensitise NB cells to chemotherapy [148].

Undifferentiated CSCs are thought to be an important mechanism of resistance and recurrence in NB. A study by Hansford et al. found evidence for potential NB CSCs existing in the bone marrow from NB tumours in remission as well as relapsed samples [149]. These cells showed neural crest progenitor markers, were capable of self-renewal, could differentiate into neurons and form tumours in immunocompromised mice. Undifferentiated NB CSCs were found to have a unique gene expression profile with high expression of *TRKB* and *LNGFR* possibly contributing to chemotherapy resistance [150]. These genes encode for proteins that mediate neurotrophins, proteins that play a role in the proliferation and differentiation of neuroepithelial precursors including neural crest cells [150].

MYCN amplification has been shown to protect NB cells against oxidative damage [151]. *MYCN* was found to transcriptionally upregulate Glutamate–Cysteine Ligase (GCL), the enzyme responsible for the rate-limiting step in the detoxification of Reactive Oxygen Species (ROS) by glutathione biosynthesis. Upregulation of GCL

enhances the glutathione biosynthesis, resulting in increased detoxification and preventing apoptosis induced by oxidative stress.

In addition to chemotherapy resistance, many patients do not respond to anti-GD2 treatment (45.2% to 71.4%) [152] and still relapse despite initial response to this immunotherapy [36] with one study reporting a 5 year FFS of 48.3% [153]. Mabe et al. showed that transition to a mesenchymal phenotype reduced GD2 expression and promoted anti-GD2 therapy resistance [44]. This was through downregulation of ST8SIA1, an enzyme involved in GD2 synthesis [44]. Another study found that YAP, a protein also found to be involved in chemoresistance, contributes to anti-GD2 resistance through ST8SIA1 expression [154]. The authors suggested this mechanism is a downstream effector of mesenchymal GD2 resistance.

1.4.2.2 Therapy resistance in RMS

Approximately a third of RMS patients relapse after completion of chemotherapy, with this rising to two thirds in patients with metastatic disease [155, 107, 17]. There have been a number of mechanisms proposed that may contribute to this chemotherapy resistance.

A recent study found that ERMS tumours transition through stages of myogenesis, from an immature mesoderm state to highly proliferative myoblast stage [156] (Figure 1.10A). The authors found that chemotherapy selected for cells in the mesoderm state in ERMS tumours, by killing off proliferative myoblast-like cells (Figure 1.10B). Cells in the mesoderm state were chemotherapy resistant and after treatment were able to expand and repopulate. They showed higher EGFR signalling activity compared to myoblasts and myocytes, inhibition of which enhanced therapeutic efficacy in organoids.

As paediatric RMS has a more favourable prognosis after treatment than adult RMS, it was investigated whether this could be explained by a decrease in sensitivity due to chemotherapy in adults by looking at proteins associated with multidrug resistance [121]. The expression of multidrug resistant proteins P-gp, MRP1 and LRP were assessed by IHC in RMS patients. It was found that LRP was more highly expressed in pleomorphic and embryonal adult RMS than paediatric RMS. Low expression of LRP was observed in ARMS, suggesting that other mechanisms may be responsible for resistant in this subtype. No difference was observed in expression of P-gp and MRP1 between children and adult cases, suggesting other mechanisms of MDR may play a role in resistance in RMS. However, another study found evidence for the upregulation of *ABCB1*, the gene encoding for P-gp, through the transcriptional activator GLI1 [120]. VCR-resistant RMS cell lines showed significantly higher expression of transcriptional activator GLI1 compared to parental cells. Other

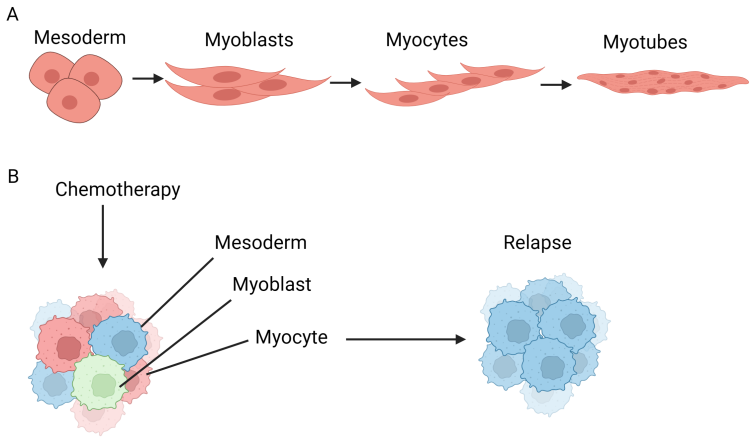


FIGURE 1.10: Clonal selection mechanism of resistance in ERMS tumours proposed by Patel et al. [156]. A) ERMS tumours were found to progress through the stages of myogenesis B) Selection of cells in the mesoderm state as a proposed mechanism of chemotherapy resistance in ERMS. Created with BioRender.com.

significantly up-regulated genes were major vault protein *MVP*, a transcriptional target of GLI1 and has also been implicated in multidrug resistance. Treatment of VCR-resistant cells with GLI1 inhibitor GANT61 increased sensitivity to VCR in RMS. GLI1 may contribute to VCR resistance in RMS.

A study on xenograft mouse models found 5 members of the Glutathione-S-Transferases (GST) network, involved in detoxification, had increased expression after treatment with VCR [157]. Increased protein expression and activity of GST was observed in ERMS and ARMS after treatment with topotecan, VCR and doxorubicin. Combination treatment of chemotherapy and GST inhibitors reduced cell viability, dependant on RMS subtype. This may provide some evidence for the role of GST in MDR in RMS.

Targeting cell death pathways is another approach that has been investigated in overcoming resistance in RMS. The protein IGF2BP1, expressed in patient-derived RMS cell lines, was shown to drive translation of cIAP1, a regulator of caspase-8-mediated cell death [158]. Reducing cIAP1 sensitised RMS cells to cell death, delayed tumour growth and improved survival in mouse models. IGF2BP1 may be a regulator of apoptosis resistance in RMS.

The BCL-2 family members control cell death by binding interactions that regulate mitochondrial outer membrane permeabilization, releasing pro-apoptogenic factors that result in caspase activation and apoptosis [128]. BCL-2 members are split into anti-apoptotic members for example BCL-2, BCL-XL and BCL-W, pro-apoptotic members that form membrane pores (BAX, BAK and BOK) and pro-apoptotic proteins that only contain the BH3 domain such as BAD, BID, BIK, NOXA, and PUMA. Research into overcoming anti-apoptotic mechanisms in treatment resistance RMS to sensitise chemotherapy has been explored.

Recently, a study looked at using anti-apoptotic inhibitors in combination with chemotherapy to reduce dosage, toxicity and side effects [159]. RMS cells were found to have acquired VCR resistance through enhanced binding of proteins BAK and BID to anti-apoptotic protein MCL-1 after treatment, inhibiting apoptosis. A combination of VCR and anti-apoptotic inhibitor MCL-1 inhibitor S63845 reduced tumour growth in xenograft models. Additionally, they identified a combination of doxorubicin with the same MCL-1 inhibitor or the BCL-XL inhibitor A-1331852 allowed decreased dosage of doxorubicin and maintained high cytotoxicity. These inhibitors could be used to increase sensitivity of standard RMS treatment.

A drug screen on patient-derived xenografts was conducted to attempt to identify drugs that, in combination with first line chemotherapy, would reduce cell viability [160]. They found ABT-263, an inhibitor of anti-apoptotic proteins BCL-2/BCL-XL/BCL-W, reduced viability in combination with etoposide, doxorubicin or VCR. Cells treated with ABT-263 in combination with another chemotherapy agent compared to a single-agent showed elevated caspase activity. ABT-263 is thought to act neutralising the pro-survival function of BCL-XL, inducing apoptosis. In mouse models, ABT-263 delayed tumour growth and prolonged animal survival, although did not lead to tumour regression. This provides further evidence that BCL-XL may play a role in anti-apoptotic mechanisms associated with resistance in RMS.

The PAX3-FOXO1 fusion gene, present in the majority of ARMS, encodes for a potent transcription factor that may partly contribute to therapy resistance in several ways. One way in which it may do this is by adaptation of cell checkpoints [161]. Cell cycle checkpoints prevent the transmission of DNA damage to daughter cells by arresting cell cycle progression and repairing DNA where possible or promoting cell death where there is unrepairable damage. In some cases, cells can overcome this arrest and continue to divide with the presence of unrepairable DNA, in a mechanism known as checkpoint adaptation. In cells exposed to radiation, there was a higher fraction of DNA breaks in cells that expressed PAX3-FOXO1 compared to knockdown and a higher number of dividing cells which could be explained by PAX3-FOXO1 overcoming G2 or M checkpoint arrest after radiation. More cells transitioned from G2 to M phase without initiating apoptosis compared to PAX3-FOXO1 knockdown, also supporting the involvement of PAX3-FOXO1 in checkpoint adaptation. It could also play a role in resistance to chemotherapy, as the G2/M is an important cell cycle checkpoint after DNA double strand break induced by chemotherapy.

A summary of mechanisms of chemoresistance in RMS is shown in Figure 1.11.

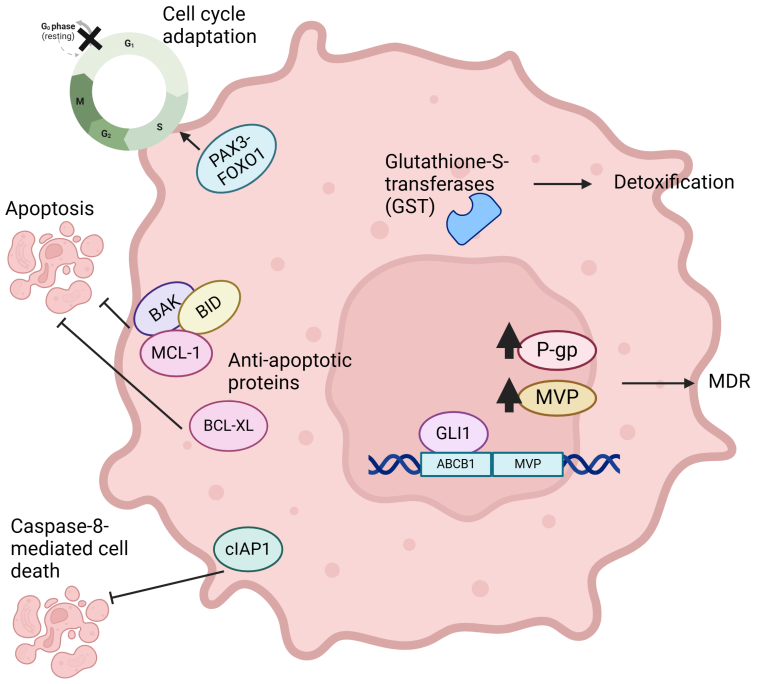


FIGURE 1.11: Mechanisms of chemotherapy resistance in RMS. Includes detoxification of chemotherapeutic agents by GST; upregulation of MDR proteins P-gp and MVP; inhibition of cell death by cIAP1; increased binding of BAK and BID to the anti-apoptotic protein MCL-1; inhibition of apoptosis by BCL-XL and adaptation of cell checkpoints by the PAX3-FOXO1 fusion protein. Created with BioRender.com.

1.5 RNA sequencing

1.5.1 Overview of RNA sequencing and analysis

Transcription, the process of copying a segment of DNA into Messenger RNA (mRNA), is the first step in transferring information from a gene to a protein. The transcriptome encompasses all the RNA molecules and can provide insight into the cause and development of disease. RNA-seq, is a high-throughput method that relies on the sequencing of Complementary DNA (cDNA) to analyse RNA and measure gene expression. RNA-seq can provide other information in addition to gene expression. It can be used for detecting splicing events, a form of gene regulation where multiple mRNA isoforms are produced from a single gene, and novel isoforms [162, 163]. At higher sequencing depths, RNA-seq can be used for variant calling, such as detecting gene fusions and single nucleotide variants [164]. RNA-seq also gives better detection of lowly expressed genes/transcripts compared to microarray [165].

Typically, an RNA-seq workflow consists of RNA extraction, mRNA isolation by either poly-A selection or ribosomal depletion and the conversion to cDNA by reverse transcriptase. Sequencing adapters are ligated to the cDNA fragments and amplification by PCR, before sequencing on a high-throughput platform (Figure 1.12).

1.5.1.1 RNA-sequencing workflow

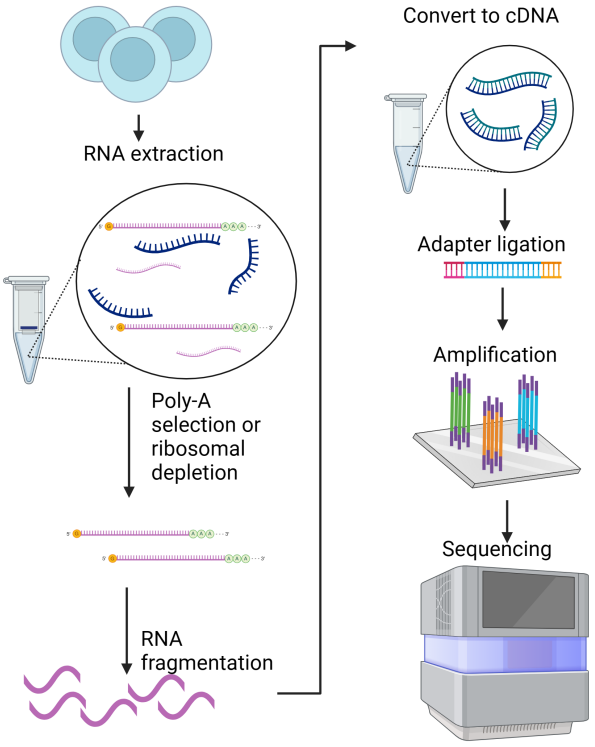


FIGURE 1.12: Bulk RNA sequencing workflow. Total RNA is extracted and mRNA is isolated through poly-A selection or ribosomal depletion. The mRNA is fragmented and converted to cDNA by reverse transcriptase. Adapters containing a sequencer binding site, index and a primer binding sites are ligated to the cDNA. The cDNA library is then amplified by PCR and sequenced on a high-throughput platform. Created with BioRender.com.

The RNA pool contains a number of different types of RNA, including Ribosomal RNA (rRNA) (80-90% of RNA) [166], Precursor messenger RNA (pre-mRNA), mRNA (∼4%) [167] and Noncoding RNA (ncRNA). For the sequencing of mRNA, rRNA molecules must be removed before library construction otherwise the majority of reads will consist of rRNA transcripts. If working with blood-derived RNA, haemoglobin RNA is also sometimes depleted to allow the detection of lower-level transcripts [168]. Poly-A selection and ribosomal depletion are the two main techniques for selecting mRNA. Poly-A selection uses oligo (dT) primers to select for 3'poly-A tail of mRNA molecules whilst ribosomal depletion relies on the removal of rRNA by hybridization capture followed by magnetic bead separation [168]. Poly-A selection selects a smaller proportion of the genome than ribosomal depletion, as ribosomal depletion does not exclude pre-mRNA (including introns) and therefore targets a bigger proportion of genome [169]. This means that poly-A selection results in a higher average read depth compared to ribosomal depletion for same amount of sequencing but can result in bias of coverage towards the 3'end of the transcripts in degraded RNA, so ribosomal depletion is favoured for lower quality RNA samples

[170]. Ribosomal depletion can aid the quantification pre-mRNA so is favoured in splicing studies [168].

After this, RNA is fragmented, end repaired and converted to cDNA using random hexamer priming to reverse transcribe selected mRNA. This process can result in stranded or unstranded sequencing. In unstranded sequencing the information about the strand of origin of the transcripts is lost when libraries are derived from cDNA. However, this information can be retained in more advanced protocols which distinguish sense from antisense strands, which is useful for overlapping transcripts. This can be done by incorporating deoxy-UTP during synthesis of the second cDNA strand and removing it during library preparation [171]. Next, oligonucleotide adapters are ligated to the ends of the cDNA fragments. These adapters each contain a sequencer binding site, that allow fragments to attach to the sequencing flow cell, an index that acts as a sample identifier if there are multiple samples, and a primer binding sites for the initiation of sequencing. Fragments can be sequencing from one direction, single-end sequencing, or both directions called paired-end sequencing [172]. For paired-end sequencing, adapters are ligated to both ends of the fragments, and the fragments are sequenced from both ends. Paired-end sequencing improves the accuracy of alignment, detection of structural rearrangements and improves the detection of insertion- deletion variants [173].

The library is then amplified by PCR and sequenced on a high-throughput platform to produce short sequence reads of 50-150bp in length [174]. The base call data is stored in a Binary Base Call file, and then converted to a FASTQ file, a text-based file that contains both sequencing data and quality scores. If multiple samples were sequenced, then samples are clustered based on the index sequence. One FASTQ file is generated for each sample per flow cell lane, meaning for single-end sequencing one FASTQ file is produced for each sample (Read 1) and for paired-end, two FASTQ files are generated for each sample (Read 1 and Read 2).

1.5.1.2 Analysis of RNA-sequencing data

Computational analysis of RNA-seq starts with the FASTQ files, which are assessed for quality to discard low-quality reads, detect any contaminating sequences, identify any issues with the sequencing and trim adaptor sequences if required. Next, reads are mapped to a reference genome and the number of reads that map to each gene is quantified. Alternatively, the reads can be assembled into transcripts, mapped to a reference transcriptome and transcripts can be used to quantify expression levels. This can be either alignment-based, where after alignment the abundance of transcripts is estimated by the alignments at annotated gene loci, or alignment-free which utilises k-mer-based counting [175]. k-mer based counting is much quicker than alignment to the reference genome and also has a smaller data footprint (as no bam file is created),

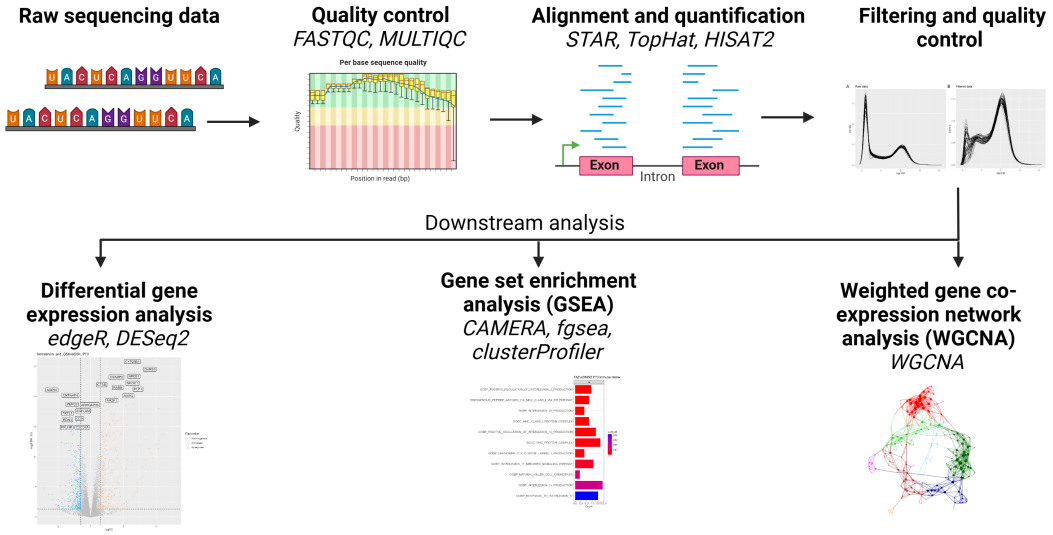


FIGURE 1.13: Bulk RNA sequencing analysis workflow. Raw sequencing data is assessed for quality control and aligned to the reference genome. The number of genes that aligned to each gene is quantified and filtered to remove genes with low expression. Available bioinformatic tools for each workflow step are shown in italics. Created with BioRender.com.

but there is less accurate quantification of rare transcripts and it is limited to only known genes [175].

After alignment, the number of reads that map to each gene is quantified. The raw counts are filtered to remove genes with low expression and undergo further quality control. The data is then ready for downstream analysis, which could include Differential Gene Expression (DGE) analysis, Gene Set Enrichment Analysis (GSEA), Weighted Gene Co-expression Network Analysis (WGCNA) and clustering (Figure 1.13).

1.5.2 Applications of RNA-sequencing in cancer research

RNA-seq is frequently used in cancer research and has many applications, including the study of molecular mechanisms of tumourigenesis, the Mechanism of Action (MoA) of drugs and in studying treatment resistance.

Understanding tumour biology between heterogenous tumours can be achieved by RNA-seq. For example, comparison of RNA-seq data from *MYCN* amplified NB versus non-amplified tumours revealed differential expression of 223 genes, and pathway analysis suggested mTOR-related genes were upregulated in MYCN amplified tumours [176]. mTOR now represents a potential novel pathway in MYCN-mediated tumourigenesis that could be targeted. Further preclinical work has showed a synergistic effect of inhibiting mTOR and MYCN (through BET inhibitors), resulting in suppression of NB cell growth and survival [177].

RNA-seq can be used to identify potential drug MoA by detection of drug-induced gene expression changes. For instance, in NB the WDR5 protein has been identified as a coactivator in MYCN–regulated transcriptional activation and tumourigenesis in NB [178] and the molecular MoA of WDR5 inhibitors was investigated using RNA-seq [179]. Four NB cell lines were treated with two WDR5 inhibitors that target different sites of WDR5 to investigate the differences in MoA. GSEA comparing the inhibitors to control revealed overlap in some pathways between the two inhibitors, including the ribosome pathway, p53 signalling, inhibition of cell cycle progression and DNA replication. However, the two inhibitors had different effects on ribosome genes, with one inhibitor upregulating the majority of ribosome genes, and the other resulting in more than half of the genes being downregulated. This highlights how RNA-seq can be used to investigate the different MoA of drugs. Despite the difference in MoA of these drugs on ribosomal genes, the two drugs showed a synergistic effect.

As well as aiding the understanding MoA of drugs, RNA-seq can identify genes associated with drug resistance and potential mechanisms that drive drug resistance, which may lead to the development of new therapeutic approaches. In NB mouse models, transcriptome sequencing of treatment-naïve versus chemo resistant tumours revealed differences in only a small number of genes, including genes that are markers of high-risk NB. [180]. GSEA showed significant enrichment of the JAK-STAT pathway in resistant tumours compared to treatment-naïve tumours, inhibition of which led to cell cycle arrest and reduced tumour growth [180].

Another use of RNA-seq is for the recognition of unique gene expression patterns in specific conditions, such as prognostic groups and resistant/non-resistant tumours for the development of gene signatures [181]. A study by Shen et al. derived a gene signature predictive of resistance to radiotherapy in breast cancer [182]. Using publicly available data, they conducted Weighted Gene Co-expression Network Analysis (WGCNA) and identified modules correlated with radiotherapy response using Progression-Free Survival (PFS). Using the genes from these modules, they then used a Cox regression model to identify survival related genes. Their 11 gene signature was shown to predict OS independently of clinical factors and had a better predictive power than clinical factors. Jemaà et al. also utilised RNA-seq to validate a gene signature of acquired chemotherapy resistance in NB [183]. The authors used publicly available RNA-seq and microarray data to cluster patients by expression of a previously identified gene signature of chemoresistance. They found that the gene signature could be used to predict OS and FFS.

1.6 Research motivation, aims and objectives

This study will use RNA sequencing data to investigate both the potential mode of action of drugs and also the emergence of treatment resistance in high-risk groups of NB and RMS. The first part will focus on understanding the MoA of SOC and a targeted therapy in NB. The second part will focus on understanding the molecular mechanisms underpinning resistance in RMS with a view to generate a predictive signature and derive novel therapeutic targets.

1.6.1 Research motivation/ Rationale

NB and RMS are two paediatric cancers where high-risk disease has a poor prognosis and high relapse rates. There is a need to understand the mechanisms underlying treatment resistance and relapse and to develop more effective therapeutic strategies.

Approximately half of NB patients are classified as high-risk and receive intensive multimodal therapies. Despite this treatment, the 5-year FFS rates for these patients remain around 50%, relapse rates are between 50-60%, and post-relapse survival is particularly poor [184]. Current treatments are intensive and associated with long-term adverse effects on survivors' quality of life. Isotretinoin is part of current maintenance therapy aimed to induce differentiation in residual NB cells and to reduce recurrence and drug resistance. Previous research has implicated EZH2 in NB tumourigenesis and shown that EZH2 inhibition can reduce tumour growth, enhance anti-GD2 therapy and induce differentiation [185, 186, 44]. Whether EZH2is could be used to further promote differentiation of NB cells in combination with isotretinoin has not yet been explored. This may enhance current maintenance therapy by further differentiating any remaining residual disease and mitigate risk of relapse. RNA sequencing is one approach that help elucidate underlying synergy between these two therapeutic strategies and also explain the MoA.

This study investigates the combination of EZH2is with isotretinoin as a potential therapeutic strategy in neuroblastoma. RNA-seq was employed to uncover the molecular mechanisms and phenotypic effects of single-agent (isotretinoin and two EZH2is) and combination treatments. It was hypothesised that the combination of EZH2is and isotretinoin would result in more differentiated cells compared to single-agent treatment. The mesenchymal/adrenergic cell lineage state was also investigated to determine if treatment impacts the mesenchymal/adrenergic phenotype of the cells. Based on previous work by Mabe et al. [44], it was hypothesised that treatment with EZH2is would shift cells to a more adrenergic state.

In RMS, outcomes for patients with high-risk disease is poor, with 5-year OS rates between 20-30% [15]. Although most tumours initially respond to therapy, relapse is

common and post-relapse survival is extremely poor [107]. Resistance, either intrinsic or acquired, is a significant contributor to treatment failure [108]. Understanding these resistance mechanisms is essential for predicting patient response and and identifying novel therapeutic targets.

To investigate resistance in RMS, RNA-seq was performed on models representing both intrinsic and acquired resistance. Intrinsic resistance was modelled using single-cell clones derived from heterogeneous RH30 cells that were screened for VCR resistance [187]. Acquired resistance was modelled by treating RMSYM and RH4 cell lines with increasing doses of VCR or IFO to generate resistant clones. DGE, GSEA, and WGCNA were used to identify resistance associated signatures. These signatures were evaluated for prognostic relevance using publicly available datasets. It was hypothesised that higher resistant signature scores would correlate with worse clinical outcomes. Additionally, a machine learning model was developed to predict clinical events using pre-treatment microarray and clinical data.

Together, these studies aim to improve our understanding of therapy resistance and recurrence in NB and RMS and to identify new therapeutic strategies.

1.7 Thesis aims and objectives

Hypothesis 1: Treatment of NB spheroids with combined EZH2is and isotretinoin will result in a more transcriptionally differentiated cell phenotype compared to single-agent treatment.

Aim 1: Investigate molecular MoA of targeted therapeutics in high-risk neuroblastoma.

Objective 1.1: Analyse RNA sequencing data from untreated NB cells and cells treated with EZH2is, isotretinoin and combination therapy. Perform DGE using the edgeR package and test for enrichment of biological functions and biochemical pathways using the CAMERA package with Reactome, Hallmark and GO databases.

Objective 1.2: Calculate the mesenchymal/adrenergic cell lineage scores of the samples to assess treatment-induced phenotypic transitions.

Objective 1.3: Analyse RNA sequencing data from NB treated versus treated cells. Perform DGE using the edgeR package and test for enrichment of biological functions and biochemical pathways using the CAMERA package with Reactome, Hallmark and GO databases.

Hypothesis 2: Chemotherapy-resistant rhabdomyosarcoma cells models exhibit a distinct gene expression signature compared to controls.

Aim 2: To identify a molecular signature of chemotherapy resistance in resistant rhabdomyosarcoma cell models.

Objective 2.1: Analyse RMS cell line RNA sequencing data from resistant and matched clones of three cell lines (RH30, RH4, RMSYM). Perform DGE using the edgeR package and test for enrichment of biological functions and biochemical pathways using the CAMERA package with Reactome, Hallmark and GO databases.

Objective 2.2: Create a profile of resistance to chemotherapy using WGCNA and Protein-Protein Interaction (PPI) networks to identify modules with a significant correlation with resistance.

Hypothesis 3: GSVA enrichment scores for chemotherapy-resistant gene signatures are significantly higher in patients with aggressive clinical outcomes.

Aim 3: Identify if the resistance signature is present in publicly available patient data and test for association with clinical traits.

Objective 3.1: Process raw publicly available patient microarray data using the Robust Multichip Average (RMA) package for background correction, normalisation and summarisation. Assign patient samples scores based on the gene signatures of resistance.

Objective 3.2: Use T test or Wilcox Rank Sum test to calculate pairwise comparisons between levels with corrections for multiple testing to determine if there is an association between gene signature scores and worse clinical outcome.

Hypothesis 4: Machine learning-derived gene expression signatures can predict clinical events in rhabdomyosarcoma patients.

Aim 4: Use a machine learning approach to identify genes that are able to predict events from patient data.

Objective 4.1: Train machine learning models using patient gene expression microarray data and clinical data to predict events (relapse or progression) in patients.

Objective 4.2: Assess and compare the performance of machine learning models on a test dataset using metrics including Area Under the Curve (AUC), sensitivity and specificity.

Chapter 2

Materials and methods

2.1 Analysis of RNA-sequencing data

2.1.1 Quality control of the raw data

RNA integrity was assessed using Bioanalyzer to provide a RNA Integrity Number (RIN). RIN is used to score the quality of RNA between 1 and 10, with a score of 10 being the highest quality and showing least degradation. It is determined by looking at the electrophoretic trace of the RNA samples using various measures of rRNA to determine RNA degradation levels. In humans, two bands are typically shown that represent 28S and 18S rRNA [188]. If the ratio of 28S:18S is 2:1 then the RNA is considered high quality and this corresponds to a RIN score of 10 [188]. Decreasing ratios correspond to a lower RIN score. Other indicators of low quality RNA on a electropherogram include the presence of additional peaks below the ribosomal bands, a decrease in the overall signal and a shift towards shorter fragments [188]. Samples with low RIN score can affect the sequencing results and so only high quality RNA should be used. Raw sequencing data was assessed for quality using FASTQC (v 0.11.9) and MULTIQC (v 1.0.0) [189]. These tools have been developed to provide an overview of quality control metrics on raw sequencing data from high throughput sequencing pipelines. FASTQC conducts these checks on individual samples and assess metrics including per base sequence quality, per sequence GC content and checking for overrepresented sequences. MULTIQC combines the output from multiple FASTQC files and compiles a HTML report to summarise the results and aid comparison between samples.

2.1.2 Alignment

After quality control the reads are then mapped to the reference genome. This step aims to determine the location in the genome where the read originates from. Alignment tools for RNA sequencing data differ in performance, the most commonly used are "splicing-aware" aligners that can discriminate between a read that aligns across an exon–intron boundary and a read with a short insertion. Some of the most popular aligners are STAR [190], TopHat [191], HISAT2 [192] and MapSplice [193]. After alignment, the number of reads that map to each gene are then counted to determine the expression level. The splice-aware aligner STAR (v 2.7.10a) was used to align fastq files to the UCSC human reference genome GRCh38. This was selected as it has a high mapping speed whilst also retaining high accuracy [190]. The –quantMode GeneCounts option was selected which counts the number of reads that maps to each gene while mapping. Default parameters were used for all other options. This requires a Gene Transfer Format (GTF) file that corresponds to the genome assembly used, in this case UCSC. While alternative annotations like Ensembl or RefSeq are also available, the UCSC transcriptome provides a well-supported, highly compatible and widely adopted resource for robust RNA-seq analysis. The GTF file defines the genomic location of all genes and exons that will be tested for differential gene expression.

2.1.3 Assessment of alignment

Assessment of alignment can indicate sequencing accuracy and highlight if low quality RNA was used in the starting samples. The percentage of mapped reads should typically be around 70-90% [194]. A low percentage of mapped reads may indicate issues such as contamination of the sample, poor quality reads or insufficient depletion of ribosomal RNA (for libraries generated using ribosomal depletion) [190]. The read coverage should be uniform without bias, if poly-A selected samples show bias at the 3' end this indicate significant degradation during RNA collection. RSeQC was used to assess gene body coverage with geneBody_coverage.py [195], with the USCS housekeeping gene inputted as the bed file.

2.1.4 Filtering

Filtering is applied to remove genes with low expression and reduce false positives, as a small difference in expression levels between conditions can result in significant differences. It is assumed that a gene must be expressed at some level before it is translated into a protein or to be biologically important. It is recommended to filter reads based on the observed data rather than pre-defined thresholds [196]. Filtering

should be done on normalised counts to account for differences in library size between samples. Raw read counts need to be normalised before comparison due to differences in the total number or reads per sample, differences in read depth between genes and sequencing bias as these factors could contribute to artificial differences in expression rather than true differences in expression between samples [197]. Reads were converted into Counts Per Million (CPM), a gene expression unit that normalises for sequencing depth. Following the edgeR manual [198], genes were kept if they had a CPM corresponding to $\sim$10 counts in the minimum number of samples (smallest group size). Density plots were used to visualise reads before and after filtering.

2.1.5 Quality control of the processed data

Further quality control checks can assess the intra- and intergroup sample variability and identify outliers. Plots of raw and normalised library sizes (log2 CPM) were generated to identify outliers and assess the data. Principal Component Analysis (PCA), an unsupervised tool to look at variation in the data, was also used to identify potential outliers using biplots of PC1 vs PC2 to identify samples that did not cluster with the rest of their experimental group. A more detailed description of PCA can be found in section 2.1.8.

It is recommended to have at least three biological replicates in RNA sequencing experiments [194]. Replicates are used to estimate the biological variation between samples and improve the accuracy of estimating expression levels [199]. Biological replicates of the same condition should cluster together on a PCA as there should be little variation between these samples. This visualisation can aid the identification of technical or biological outliers. The Biological Coefficient of Variation (BCV) is another way to measure the biological variation within a particular condition, representing variation that would remain between biological replicates if sequencing depth could be increased indefinitely [200]. BCV is the square root of the common dispersion and values should be between 0.2 and 0.4 for human data, indicating that expression levels vary by 20-40% between replicates [200]. As there is less genetic variation in cell lines, BCV is usually lower at around 0.1. BCV values higher than this may result in less DEGs detected.

2.1.6 Differential Gene Expression (DGE) analysis

The most common analysis for RNA-seq data is to identify genes that are differentially expressed across conditions, called Differential Gene Expression (DGE). A variety of tools are available for DGE analysis, using different statistical models and data normalisation methods. edgeR was used for differential gene expression analysis [198], as it is recommended over DESeq for data that has $<$12 replicates [201]. Samples

were normalised using edgeR (v 3.40.1) by the trimmed mean of M-values (TMM) method. This normalisation method calculates M values, where M is the log-2 fold change of a gene between two samples. Extreme M values are trimmed and scaling factors are calculated by taking a weighted trimmed mean of M [202]. TMM was selected over other normalisation methods as it is straightforward to apply through the edgeR pipeline and has been shown to perform better than other methods [203, 204].

Read counts were analysed for differential gene expression using edgeR glmQLFit to determine Differentially Expressed Genes (DEGs) as recommended by the edgeR manual [198]. glmQLFit uses the trended dispersion to fit a generalised linear model and uses quasi-likelihood F-test for determining differential expression. DEGs were visualised in volcano plots, a type of scatterplot showing significance as log transformed adjusted p-values vs log fold change. Genes with a negative fold change are presented on the left of the plot and positive fold changes on the right. Significant genes with a large fold change will be presented towards the top outer corners of the scatterplot. Heatmaps were also used to visualise gene expression using the R package ComplexHeatmaps [205], with Z scoring applied to CPM-normalised RNA-seq counts. Heatmaps can be used to present top differentially expressed genes or genes of interest and were combined with hierarchical clustering to group genes and samples with similar expression levels. This can be used to identify genes that are commonly regulated and biological signatures associated with experimental groups.

2.1.7 Gene Set Enrichment Analysis (GSEA)

GSEA is a common approach for the interpretation of gene expression data to identify significantly impacted biological processes/pathways between conditions. It determines whether a defined set of genes that share a biological function are differently expressed between two conditions [206]. These gene sets can be obtained from database collections that curate gene sets from online pathway databases and the biomedical literature. There are many tools available with different statistical approaches. The three main methods are over-representation based, Functional Class Sorting (FCS) and topology-based. Over-representation based methods determine whether a process/pathway is observed more than expected by chance in the list of DEGs compared to a background gene list. Since this method uses a binary cut-off (DEGs or non-DEGs) information about the strength of the over or under representation is lost [206]. FCS methods first calculates a score based on expression and/or significance for all genes analysed and ranks the list of genes based on this score [207]. Genes that deemed as more important (e.g. greater LFC or smaller p-value) will appear at the top of the list for upregulated genes and at the bottom of the list for downregulated genes. Then it is determined if genes in a gene set tend to

cluster at the top or bottom of the list and assigns significance of gene sets enriched. Topology-based methods are similar to FCS, but additionally incorporate gene-gene interactions [208]. R package tool CAMERA was used for GSEA as recommended by the edgeR manual [198]. CAMERA uses competitive gene set test accounting for inter-gene correlation and is a topology-based GSEA [209]. All genes from the edgeR output were input into GSEA for Reactome, Hallmark and Gene Ontology gene sets. The Reactome gene sets are from the Reactome pathway database- a curated and peer-reviewed knowledge-base of biomolecular pathways. GO gene sets are derived from Gene Ontology, a database which aims to provide representation of current scientific knowledge about the functions of genes. GO terms can be split into three main categories; biological process, cellular component and molecular function. Hallmark gene sets summarise well-defined biological states or processes. These gene sets were generated by a computational method that identified overlapping gene sets and retained genes that displayed coordinate expression [210]. The gene sets were refined and validated in microarray data. Gene sets with adjusted p value <0.05 (Benjamini-Hochberg) were defined as significant. In contrast to other GSEA tools, CAMERA uses all genes in the analysis to identify small but consistent expression fold-changes. This has the advantage that it may detect enriched gene sets that may otherwise be overlooked when using a FCS with a binary cut-off.

2.1.8 Principal Component Analysis (PCA)

PCA is an unsupervised technique that aims to reduce large sets of variables into smaller sets that are still contain most of the information in the larger set. It can be used to summarise information from large complex data, increase interpretability and identify patterns in the data. When considering gene expression data, samples that have similar patterns of gene expression will be grouped together, whilst samples that show variation in gene expression will be further apart. This can be used to explore the relationships between samples and identify outliers. The R package PCAtools was used for PCA [211]. Counts were normalised to log-2 CPM and scaled per gene to reduce the influence of highly expressed genes. The lower 10% of variables based on variance were removed to reduce the amount of ”unimportant” variables as per the PCAtools tutorial [212]. Although PC1 and PC2 are the two principal components that explain the majority of the variation in the data, other principal components can be investigated that explain smaller percentages of the variation in the data. They can be visualised in a Scree plot, where the x-axis depicts PCs and the y-axis shows % of variance explained. The optimum number of PCs to be retained was determined through the Elbow method and Horn’s parallel analysis [213]. The Elbow method determines the optimum number of PCs as the point before the eigenvalue (variance explained by a given principal component) drops dramatically in size and the additional PCs would add little value. Horn’s method compares the eigenvalues

produced from the PCA to the eigenvalues generated from random data sets of the same size. PCs are retained if they explain significantly more variance than expected by chance. Both the Elbow and Horn's methods were used and in cases where the methods did not agree on the optimum number of PCs to retain the largest number of PCs identified by either method was selected.

Chapter 3

Investigating the combination of EZH2 inhibitors with isotretinoin as a therapeutic strategy in neuroblastoma

3.1 Introduction

3.1.1 Targeting epigenetic regulators in neuroblastoma

Epigenetic regulation can be defined as changes in gene activity independent of DNA sequence changes [214]. It controls many important cellular processes including gene expression, cell differentiation, embryogenesis, imprinting and chromosomal stability and needs to be tightly controlled for normal embryonic development [215]. Disruption of epigenetic mechanisms can lead to dysregulation of developmental programs and result in diseases including cancer [216]. Understanding how epigenetic regulation may contribute to the initiation and progression of tumours has allowed the development of drugs targeting epigenetic mechanisms [217]. This may be a particularly attractive therapeutic approach for paediatric cancers like NB, where there is a lack of actionable somatic mutations [215]. In NB, several epigenetic factors have been investigated as potential therapeutic targets, including DNA and histone modifying enzymes (DNA methyltransferases, histone methyltransferases, histone acetyltransferases and histone deacetylases) as well as non-coding RNAs [215].

3.1.2 EZH2 in neuroblastoma

One epigenetic target being explored is Enhancer of zeste homolog 2 (EZH2), a histone methyltransferase that is aberrantly expressed in many cancers and thought to be involved in tumour progression [218]. EZH2 is a subunit of the polycomb repressive complex 2 (PRC2), a complex that catalyses the methylation of histone H3 at lysine 27, maintaining transcriptional repression.

In NB, high expression of EZH2 is associated with poor prognosis [185] and it's overexpression is thought to maintain cells in an undifferentiated state [185, 219]. *In vivo* studies have found that inhibition of EZH2 results in a significant decrease in tumour growth and proliferation [186]. The mechanisms behind how EZH2 inhibition exerts its anti-tumour effects in NB are still being investigated. EZH2 has been reported to maintain cells in an undifferentiated phenotype by repression of critical tumour suppressor genes [220]. This undifferentiated state is more likely to be resistant to therapy leading to recurrence or relapse [42]. Inhibition of EZH2 results in the de-repression of differentiation genes, promoting cell differentiation [185]. A study by Li et al. found that depletion and inhibition of EZH2 in NB cell lines promoted neurite extension and resulted in upregulation of neural differentiation markers GAP43, RARB and NF68 [185]. De-repression of NTRK1, a favourable clinical factor in NB, was also observed and was found to contribute to EZH2 regulated differentiation. Furthermore, there may also be an application for EZH2 inhibition in overcoming resistance to anti-GD2 immunotherapy. EZH2is were shown to shift NB cells from a mesenchymal to adrenergic cell state, upregulating adrenergic and neuronal differentiation genes [44]. This EZH2i also increased GD2 expression through increasing expression of ST8SIA1 (responsible for the bottleneck in GD2 production) and restored sensitivity to anti-GD2 therapy.

Maintenance therapy for NB patients aims to reduce the minimal residual disease that remains after induction and consolidation therapy and reduce the risk of relapse [22]. Isotretinoin is currently included in the maintenance therapy regime, after clinical research demonstrated the benefit of the addition of isotretinoin to SOC treatment in NB by improving FFS [35]. Although, many high-risk NB patients show MYCN-induced resistance and around 50% of patients that initially respond to isotretinoin therapy proceed to develop resistance [221, 222]. Despite its clinical benefit to some patients, the details of isotretinoin activity against NB lacks understanding [223]. Early preclinical studies showed isotretinoin was able to induce differentiation and prevent growth in NB cells [224, 225]. Multiple studies have attempted to unpick the action of isotretinoin in NB. A study by Halakos et al. used proteomics to identify proteins affected by isotretinoin in SK-N-SH cells [226]. They found that isotretinoin induced neurites and increased proteins involved in cell adhesion, with GO analysis showing enrichment of ECM synthesis and organisation.

Zimmerman et al. looked at the effect of isotretinoin on NB Core Regulatory Circuits (CRCs) (a group of transcription factors that regulate their own expression) [227]. Isotretinoin was found to alter the enhancer landscape and reprogram the adrenergic CRC, leading to downregulation of *MYCN* expression, increased cell differentiation and inhibition of cell proliferation [227].

Both isotretinoin and EZH2 have been shown to induce differentiation in NB as single agent treatments, but whether the combination of both treatments could result in enhanced differentiation has not yet been explored. Enhancing differentiation of NB cells may further reduce minimal residual disease and risk of relapse. This chapter aimed to explore the possible therapeutic benefit of combining EZH2is with isotretinoin and understand underlying mechanisms of isotretinoin, EZH2i and the combination using Multicellular Tumour Spheroid (MCTS) models.

Cell culture is a widely used tool for understanding cell biology, tissue morphology, mechanisms of diseases, drug action in cancer biology [228]. 2D cell culture, where cells are grown in a monolayer on a flat surface such as a culture flask or petri dish, has been the most predominantly used method for *in vitro* biological research. Although 2D models are inexpensive and easily maintained, they have many limitations and do not accurately represent tissues or tumours. Three-dimensional (3D) MCTSs are becoming more frequently used in tumour biology research. These 3D tumour cell clusters are formed from growing cells in single-cell suspensions either with or without other cell types [229]. Spheroids more closely mimic the 3D tumour structure and gradients of nutrients, oxygen and pH [229]. Tumour cells first show exponential expansion, which then slows and an increase in quiescent cells are seen [230]. Proliferating cells are seen close to blood vessels where they receive oxygen and nutrients, whilst quiescent and necrotic cells lie further from the blood vessels. Like tumours, the outer layer of spheroids cells are proliferative cells that have access to oxygen and nutrients, whilst quiescent cells are further towards the core [231] (Figure 3.1). At the centre of the spheroids, a necrotic core is formed where cells don't have access to oxygen or nutrients. MCTSs are a better model of solid tumours than 2D cell culture, as they show similarities to in growth, metabolic rates and resistance [229]. This study used MCTSs as human neuroblastoma models to investigate the effects of isotretinoin, EZH2i and the combination therapy (isotretinoin + EZH2i).

3.2 Methods

3.2.1 Spheroid generation, treatment and sequencing

All laboratory work was carried out by Dr Andy Gao and I conducted the computational analysis from the point of the raw RNA-sequencing data. MCTSs were

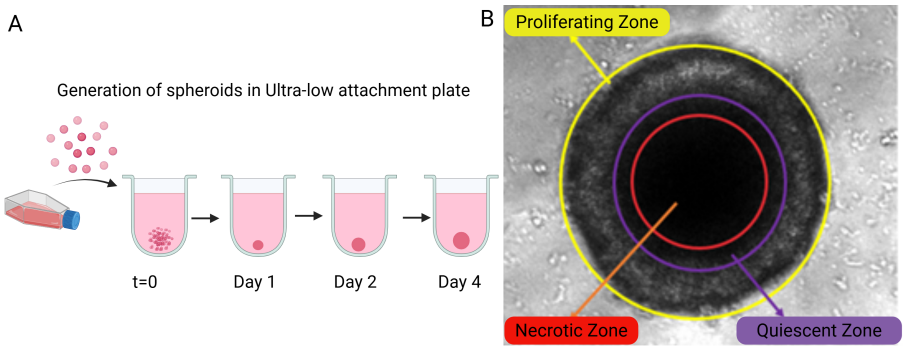


FIGURE 3.1: A) Spheroid generation using ultra-low attachment plates B) structure of multicellular tumour spheroids showing the outer proliferating layer, quiescent zone and necrotic core. Created with BioRender.com.

TABLE 3.1: Study design for the neuroblastoma spheroids. Spheroids generated from the NB cell line Kelly were treated with DMSO, isotretinoin, TAZ, GSK, isotretinoin+TAZ or isotretinoin+GSK. There were three technical replicates for each treatment condition. RNA was extracted from spheroids at three timepoints; post-treatment day 3, post-treatment day 7 and post-treatment day 10.

Treatment	Timepoint		
	Post-treatment day 3	Post-treatment day 7	Post-treatment day 10
DMSO	n=3	n=3	n=3
Isotretinion	n=3	n=3	n=3
TAZ	n=3	n=3	n=3
GSK	n=3	n=3	n=3
Isotretinoin+TAZ	n=3	n=3	n=3
Isotretinoin+GSK	n=3	n=3	n=3

generated from the high-risk NB cell line Kelly. This cell line is *MYCN* amplified, *ALK* and *p53* mutated [232]. It is considered a more adrenergic NB cell line than mesenchymal [233]. Spheroids were generated using a 96-well ultralow attachment plate (ThermoFisher Scientific), the spheroids formed after 3 to 4 days and the treatments were applied on day 4 and re-treated every 3 days. Treatments were; EZH2is GSK126 (GSK) (Biotechne-Tocris) and Tazemetostat (TAZ) (Selleckchem, Munich, Germany); isotretinoin (Sigma); isotretinoin + TAZ, isotretinoin + GSK and DMSO (Sigma). Concentrations were as follows; 2µM isotretinoin; 5µM TAZ; 5µM GSK. For vehicle controls equivalent volumes of 0.5% DMSO were added to cells. Three attachment plates were used to generate 3 replicates for each treatment and the DMSO control (a total of 18 spheroids). These were treated as biological replicates in order to carry out the analysis, similar to previous studies [234, 235]. The RNA was extracted from the spheroids at post-treatment day 3 (PT3), day 7 (PT7) and day 10 (PT10) using RNeasy mini kits (QIAGEN;74104). The study design can be seen in Table 3.1. A western blot was conducted by Dr Andy Gao to confirm a decrease in the H3K27 mark with EZH2i treatment and total RNA was extracted and submitted for RNA integrity number (RIN) assessment.(Figure 3.2).

Extracted RNA was sent to Novogene for library preparation and sequencing. After

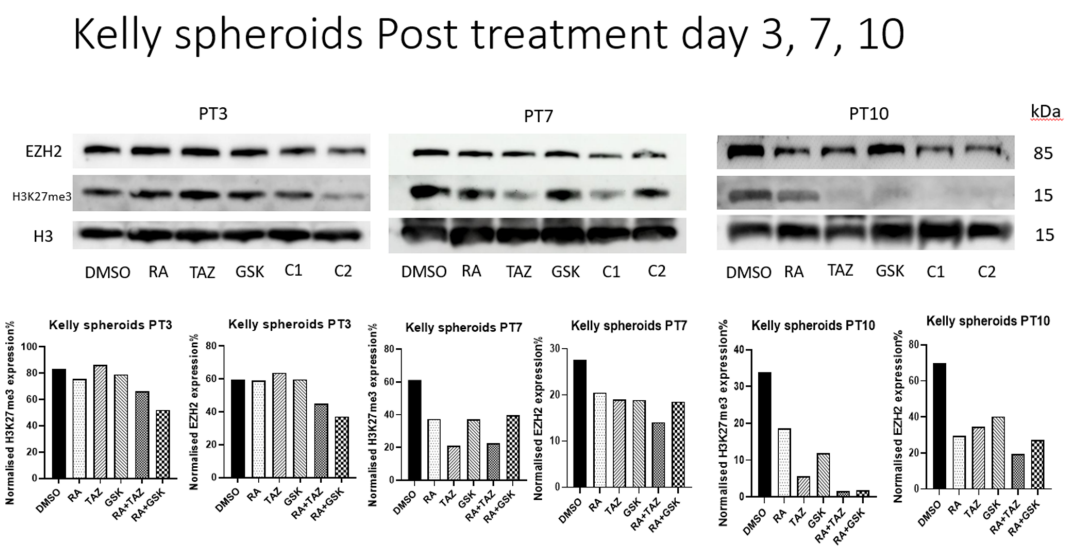


FIGURE 3.2: Western blots for the NB RNA-sequencing data at PT3, PT7 and PT10 after treatment with DMSO, isotretinoin (RA), TAZ, GSK, isotretinoin + TAZ (C1) and isotretinoin + GSK (C2). Western blots were carried out by Dr Andy Gao.

RNA extraction, the RNA is isolated and its quality assessed. Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads for poly-A selection. After fragmentation, the first strand cDNA was synthesized using psuedo random hexamer primers followed by the second strand cDNA synthesis. The library was ready after end repair, A-tailing, adapter ligation, size selection, amplification and purification. The library was checked with Qubit and real-time PCR for quantification, bioanalyzer for size distribution detection and nanodrop for RNA purity. Quantified libraries were sequenced on Illumina platforms for 150 paired-end, unstranded sequencing with Q30 $\geq$85%.

3.2.2 RNA sequencing analysis

All code is available in the Supplementary Files (Supplementary Files/Chapter 3 Neuroblastoma/Code). The analysis of the RNA sequencing data is described in Chapter 2. A CPM of 0.4 was determined as the threshold for excluding lowly expressed genes, corresponding to a read count of between 10-15 (Figure 3.3). Genes were kept if they had a CPM greater than 0.4 in at least 3 samples, as this was the lowest number of replicates for each condition. Read counts were analysed for DGE using edgeR glmQLFit [198] to determine DEGs between treatment conditions for each timepoint. Significant DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg). A LFC threshold of 1.5 was selected to identify genes with more substantial expression changes. Known targets of EZH2is and isotretinoin were used as sense check genes to identify whether changes in expression of those genes were as expected. In addition to GSEA for GO terms,

Reactome and Hallmarks as described in Chapter 2, a 59 gene neuroblastoma differentiation signature was tested for using CAMERA, developed by Frumm et al. [236]. For comparison, enrichment of 59 random genes were also tested for enrichment. These genes were randomly selected after filtering to remove lowly expressed genes using the R function 'sample'. The NB differentiation signature can be found in Supplementary Files (Supplementary Files/Chapter 3 Neuroblastoma/NB differentiation signatures).

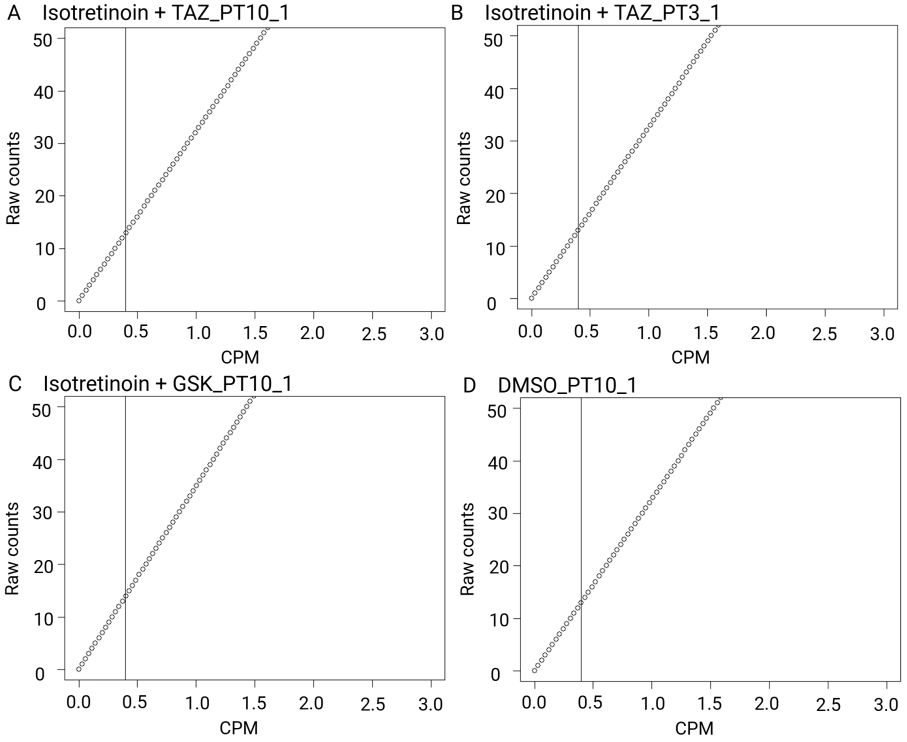


FIGURE 3.3: Plot showing count per million (CPM) vs raw count for samples. A) Isotretinoin + TAZ_PT10_1 B) Isotretinoin+TAZ_PT3_1 C) Isotretinoin + GSK_PT10_1 D) DMSO_PT10_1. Vertical line is added at a CPM of 0.4 which roughly corresponds to 10 counts.

In addition to DGE and GSEA analysis, genes that were significantly differentially expressed in at least one treatment group were used in subsequent k-means clustering analysis. K-means clustering can be used to identify similar groups in the data in an unsupervised way. The normalised gene expression (CPM) for all samples was subject to z-score scaling per gene, as in the paper by Ferguson et al. [237]. Data needs to be scaled otherwise all of the highly expressed genes will cluster together, even if they have different patterns among the samples.

In k-means clustering, each object in the data is partitioned into k number of clusters, where k is defined by the user. The optimal number of clusters was determined using the Elbow method and silhouette method. The elbow method iterates from k=1 to k=n (where n is maximum number of clusters to be considered) and calculates the within-Cluster Sum of Squares (WCSS) for each value of k. The default maximum k

value was used (10). The WCSS is the sum of the square distance between points in a cluster and the cluster centroid (the centre point of the cluster). The WCSS for each value of k can be plotted, where k=1 has the highest WCSS and increasing values of k should have a lower WCSS value. The optimum k value is selected where the graph starts to form a straight line; where increasing the value of k is having little effect on the WCSS value. The silhouette method looks at the separation distances between clusters. The optimal k number is determined as the k value with the maximum silhouette score.

K-means clustering was performed using hierarchical k-means clustering, a hybrid clustering method that aims to reduce the sensitivity of the initial random selection of cluster centres that is seen in k-means clustering. The agglomerative hierarchical clustering method selected was 'complete'. The resulting clusters were plotted using the R package ComplexHeatmap [205]. The genes in each cluster were analysed for over-representation based GSEA for GO terms, Reactome and Hallmarks using enrichGO, enrichPathway and enrichR from the R package clusterProfiler to identify enriched terms in these clusters [238]. An over-representation based method was used as this type of GSEA looks for enriched terms in a list of genes, rather than FCS or topology-based methods which look for enriched terms between samples. Significantly enriched pathways were defined as having an adjusted p value <0.05 (Benjamini-Hochberg).

This study also looked for additive genes (that had a greater LFC in combination therapy compared to either single agent treatment) for isotretinoin + TAZ at PT10. This timepoint was selected for further investigation due to the enrichment of gene sets related to DNA replication. All genes were separated into upregulated and downregulated. Genes were selected that had a greater LFC in isotretinoin + TAZ, than isotretinoin and TAZ alone. The resulting genes were input for over-representation based enrichment analysis as in k-means clustering.

To determine whether treatment affected the lineage differentiation stage, the adrenergic and mesenchymal score, developed by Mabe et al., was determined for each sample [44]. Genes for the mesenchymal and adrenergic gene sets (Supplementary Files/Chapter 3 Neuroblastoma/Mesenchymal and adrenergic score), based on the paper by Groningen et al. [42], were scored using the R package GSVA (version 2.0.7) [239] with default parameters to assign each sample a score based on the expression of these genes.

Shapiro-Wilk test was used to assess the normality of the data. Since the data was non-parametric, a Kruskal-Wallis test was used for comparisons between treatment at PT3, PT7 and PT10.

3.3 Results

3.3.1 Quality Control of raw sequencing data

A total of 54 samples with a read length of 150 bp were sequenced and analysed. All samples had a RIN score of 10, except for GSK-PT3-1 which had a RIN score of 7.8 (Supplementary Table 1). 108 FASTQ files were analysed (two FASTQ files per sample; for read 1 and read 2) using FASTQC and MULTIQC to evaluate the raw data quality and need for pre alignment filtering. All samples passed the quality control checks for sequence quality, GC content, N content, sequence length distribution and overrepresented sequences (Figure 3.4). The sample GSK-PT3-1 was not identified as an outlier in other quality control checks with FASTQC.

'Sequence Duplication Levels' and 'Per Base Sequence Content' failed which is expected for RNA sequencing data. Per base sequence content looks at the proportion of each base (A/G/C/T) at each base pair position along the reads. RNA-sequencing libraries use random hexamers (a mixture of single-stranded oligonucleotides representing all possible sequences) to anneal to RNA fragments which are reverse transcribed to cDNA [240]. This non-random binding results in bias in the nucleotide composition at the start of the DNA fragments for the first 1-15 base pairs, resulting in the data failing this QC check.

Sequence duplication levels assess the degree of duplication for all sequences in the library. High levels of duplication can indicate enrichment bias such as PCR overamplification. In the sequence duplication plot, high levels of duplication in the library can be seen represented by the peaks with sequence duplication levels >10 and >100 (Figure 3.4E). MULTIQC report shows that all samples failed on the sequence duplication levels. However, in RNA sequencing libraries large sets of duplicates are expected as highly expressed genes will have high coverage and may represent a large proportion of the sequenced reads [241]. The full MULTIQC report can be found in 'Supplementary Files/Chapter 3 Neuroblastoma/multiqc_report.html'.

To further investigate the source of sequence duplication, adapter content identifies whether a significant proportion of the library contains adapter sequences. The MULTIQC report showed that six samples had a warning (Isotretinoin + TAZ-PT3-3-1, Isotretinoin + TAZ-PT3-3-2, Isotretinoin + TAZ-PT7-3-1, Isotretinoin + TAZ-PT7-3-2, TAZ-PT7-3-1, TAZ-PT7-3-2) showing that an adapter sequence was present in more than 5% of reads (Figure 3.4F). Preprocessing trimming tools are available that remove adapter sequences and low quality reads, improving alignment to the reference genome in low quality datasets [242]. In high-quality datasets, trimming was shown to have a limited effect, with around a 1.5% gain in unique alignments [242]. More recently it was suggested that read trimming is not necessary, as a similar or slightly

better accuracy of gene expression quantification can be achieved without trimming as adapter contamination can be dealt with by the aligner [243]. Several aligners, including STAR, automatically use soft-clipping to trim the ends of reads with high mismatches. For this reason it was decided that the reads would not be trimmed.

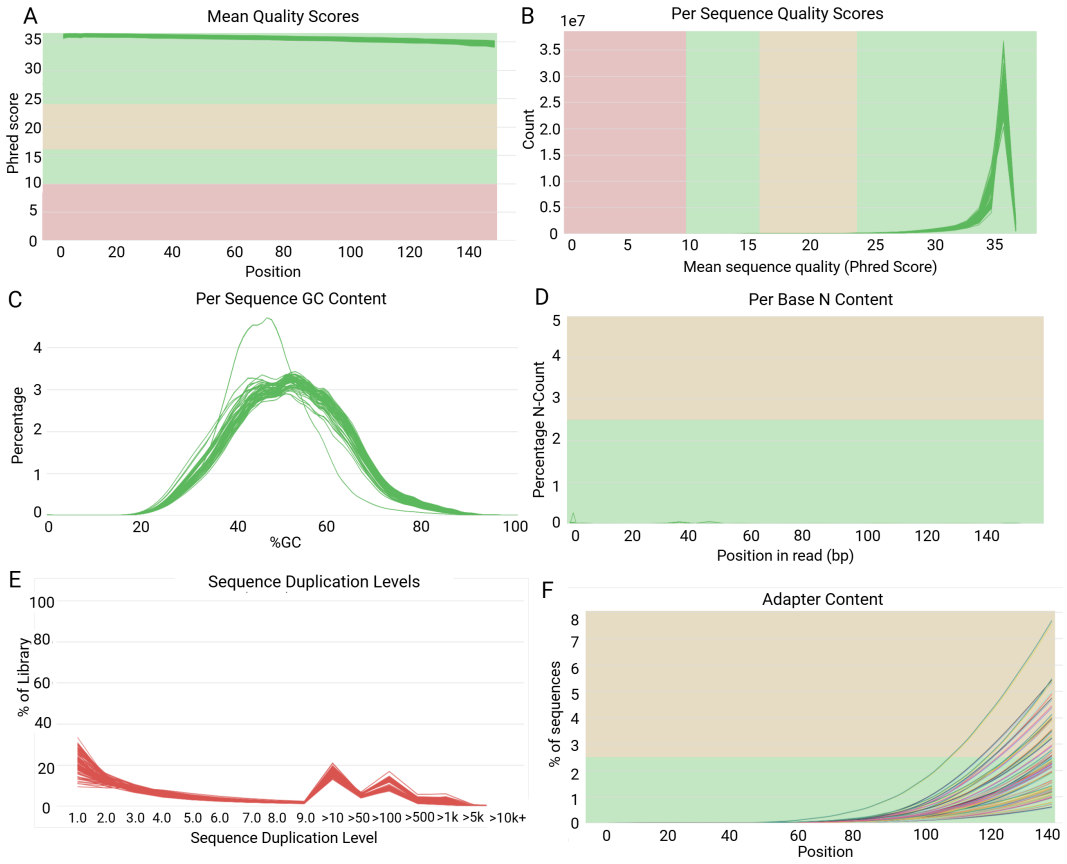


FIGURE 3.4: MULTIQC report for the neuroblastoma spheroid data. A) Mean quality scores showing the phred score for each bp in the read B) Per sequence quality scores C) Per sequence GC content D) Per base N content, where N represents where the sequencer is unable to make a base call with sufficient confidence E) Sequence duplication levels F) Adapter content. N=54 with 6 treatment conditions at 3 timepoints with 3 technical replicates each.

3.3.2 Quality Control of alignment

For all samples over 70% of reads mapped to the reference genome, except for one sample (isotretinoin + GSK-PT7-2), which had only 16% (Supplementary Table 2). This indicates that adapter contamination was dealt with appropriately by the aligner and the sample GSK-PT3-1 that had a lower RIN score gave similar results to the other samples. The sample isotretinoin + GSK-PT7-2 also had a similar percentage of unmapped reads due to mismatch compared to the rest of the sample, but a high percentage (81.93%) of unmapped reads due to reads being too short (Supplementary Table 3). The FASTQC report did not highlight any obvious errors and RIN score for this sample was very high (10). This sample was flagged as a possible outlier, but was

not excluded until further quality control checks were done. The samples with warnings for adapter content had similar percentage of uniquely aligned reads to the other samples. They did not appear to have a higher percentage of unmapped reads due to mismatch reads (Supplementary Table 3). Isotretinoin + TAZ-PT7-3 had a slightly high percentage of unmapped reads due to reads being too short (16.56%), although isotretinoin + TAZ-PT3-2 had a similar percentage to this (17.5%) and did not show a warning for adapter content, which could indicate it may be unrelated to this.

The read coverage over the gene body showed no 3' bias (Supplementary Figure 1). One sample, 'GSK-PT3-1', was slightly skewed towards 3' compared to the rest of the samples but did not show significant 3' bias. This was likely due to the lower RIN score of this sample (7.8) compared to the other samples that all had a RIN score of 10 (Supplementary Table 1).

3.3.3 Quality Control of read counts

After filtering to remove lowly expressed genes (genes that did not have a CPM of 0.4 in at least 3 samples), 16,787 genes remained. Density plots were used to visualise the effect of filtering on the data. When calculating logCPM, pseudo count is added to genes to enable calculation of logCPM for genes with zero counts (as log 0 is undefined). For genes with 0 counts, logCPM is negative. Before filtering, there is a sharp peak in density at a logCPM of around -4, from the genes a with raw read counts of 0 (Figure 3.5A). Post-filtering, the peak has been greatly reduced (Figure 3.5B). A peak at around -4 remains in the filtered data because these genes are sufficiently expressed in other samples.

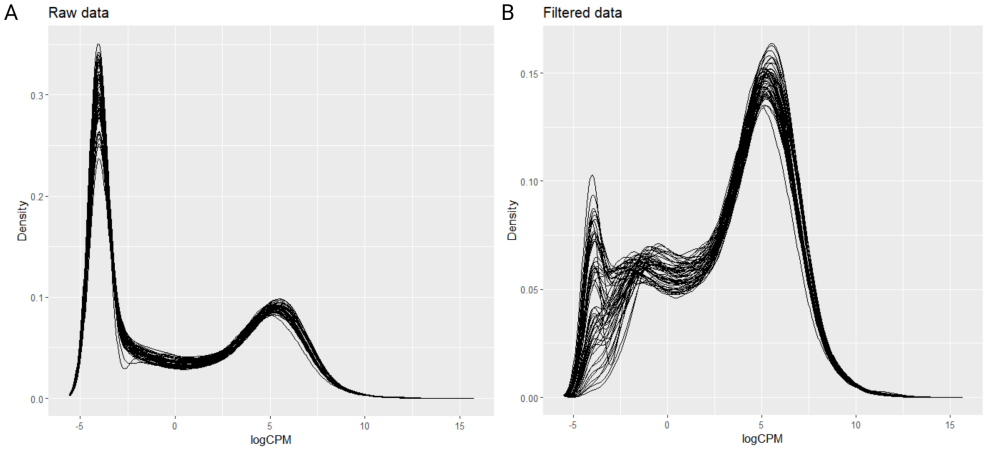


FIGURE 3.5: CPM density plots before and after filtering. A) Density plot of the raw data. N= 28,277. B) Density plot after filtering to remove genes that don't have a CPM of 0.4 in at least 3 samples. n= 16,787.

Filtered data was assessed for quality control before normalisation. No major discrepancies in library size were observed between experimental groups but normalisation is still needed to correct for individual differences (Supplementary Figure 3A). One outlier, also an outlier for the alignment (isotretinoin + GSK-PT7-2), is highlighted with a low number of reads in comparison with the rest of the samples. After TMM normalisation, there is less variation between samples (Supplementary Figure 2A).

PCA was used to explore variation in the data in an unsupervised way (Figure 3.6). The variation in PC1 (representing around 27% of total variation in the data) was driven by replicate 3 but replicate 3 shows a similar pattern to replicates 1 and 2 in PC2 (Figure 3.6B). The variation in PC2, representing ∼19% of the variation in the data, was mainly driven by time. DMSO (black) forms a cluster with relatively little variation at different timepoints compared to the treated samples. This is expected as the treatment is likely to increasingly alter gene expression over time. EZH2is and retinoids require time to alter chromatin states or transcription factor activity. These changes can lead to progressive shifts in gene expression as downstream targets are gradually activated or repressed. Cells may undergo differentiation, cell cycle arrest, or apoptosis in response to treatment, which can take hours to days to manifest at the transcriptomic level. Isotretinoin clusters nearby DMSO and separates into timepoints. EZH2is also cluster into timepoints, but there are slight differences between the two EZH2is. TAZ and isotretinoin+TAZ are further along the y-axis than GSK and isotretinoin+GSK, indicating that are more different compared to the other samples. Since the outlier with a smaller library size (isotretinoin + GSK-PT7-2) did not appear discrepant after normalisation of the library sizes (Figure 2B) and it clustered near the other replicates in the PCA biplot (Figure 3.6), it was decided the sample would not be excluded.

Other principal components that explain smaller percentages of the variation in the data were also investigated. A Scree plot was used to visualise the variance explained by each PC and Horn's method and the Elbow method were used to determine the optimum number of PCs to retain. The optimum number of PCs was 5 according to Horn's method and 7 according to the Elbow method (Supplementary Figure 4). The larger number of PCs were retained (7).

Biplots comparing PC1 through PC7 were generated to explore the additional PCs (Figure 3.7). PC3 (accounting for ∼13% of the variation in the data) shows the variation driven by isotretinoin, as DMSO, TAZ and GSK separate from isotretinoin, isotretinoin+TAZ and isotretinoin+GSK (seen along the x-axis of the plots in the column PC3) (Figure 3.7). PC5 showed variation between isotretinion and DMSO from the rest of the samples.

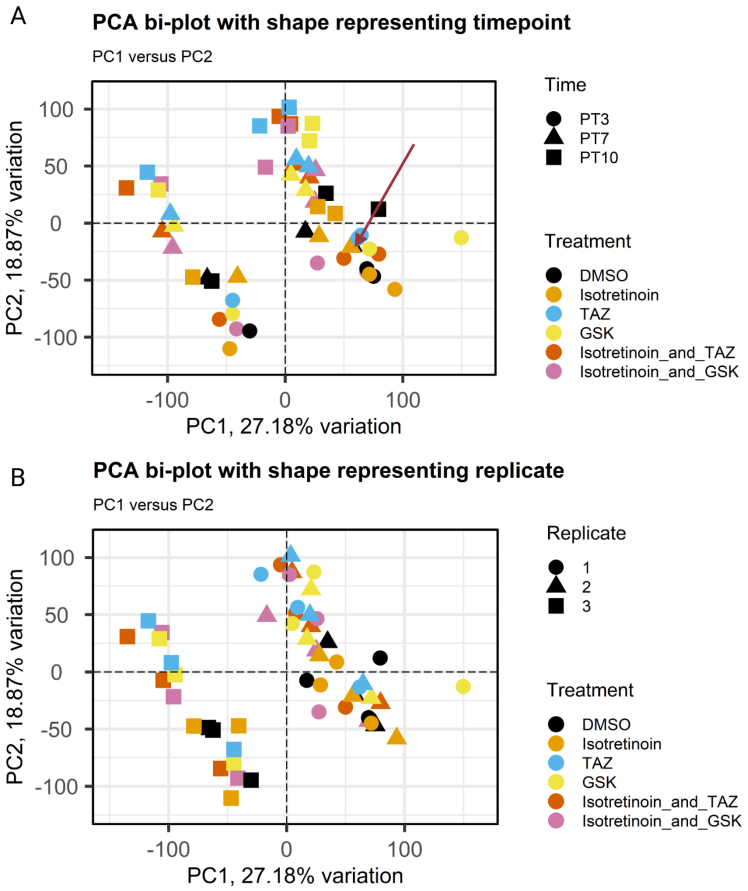


FIGURE 3.6: Biplot showing PC1 vs PC2 for the neuroblastoma RNA sequencing data. A) shape represent timepoint (PT3= post-treatment day 3, PT7= post-treatment day 7, PT10= post-treatment day 10) B) shape represents replicate. Plot shows the first 2 principal components, with the x-axis as PC1 and y-axis as PC2. Potential outlier isotretinoin + GSK-PT7-2 is highlighted by the red arrow. N=54 (6 treatment conditions at 3 timepoints with 3 technical replicates each).

Since replicate 3 was responsible for around 27% of the variation in the data (Figure 3.6B), there was a need to check if a batch effect for replicate 3 should be included in the design matrix. To further investigate if a batch effect was necessary, testing for differential expression between replicate 1/2 and replicate 3 was conducted, as recommended in the edgeR manual [198]. There is considerable differential expression, suggesting a need to adjust for batch effects (Supplementary Table 5). Although all spheroids were created and treated at the same time, other factors could contribute to a batch effect in replicate three. Factors such as the sequencing machine used, run session, flow cell ID and lane were investigated as potential sources of bias. This information was pulled from the FASTQ files and manually inspected but no evidence was revealed that could explain the source of the batch effect. The BCV without including a batch effect was around 0.31 (Figure 3.8A), which is quite high for cell lines which has an expected BCV of around 0.1. After including a batch effect in the design, BCV was reduced to around 0.15 (Figure 3.8 B).

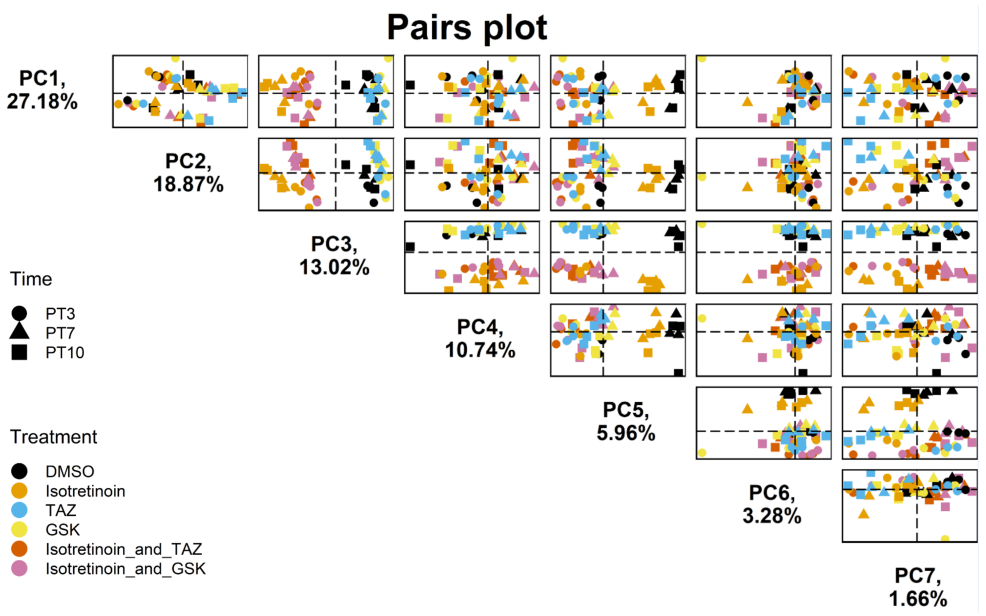


FIGURE 3.7: Pairs plot showing biplots for each PC (PC1 through PC7) for the neuroblastoma RNA sequencing data. Plots in each column represent a different PC along the x-axis. Plots in each row represent a different PC along the y-axis. Shape represents timepoint (PT3= post-treatment day 3, PT7= post-treatment day 7, PT10= post-treatment day 10). N=54 (6 treatment conditions at 3 timepoints with 3 technical replicates each).

For this reason, replicate 1/2 were grouped together as batch 1 and replicate 3 was treated as a separate batch when estimating the variation in expression between all replicates. In DGE analysis, BCV enables adjustment for differences in sequencing depth and other technical factors, allowing for more accurate identification of DEGs.

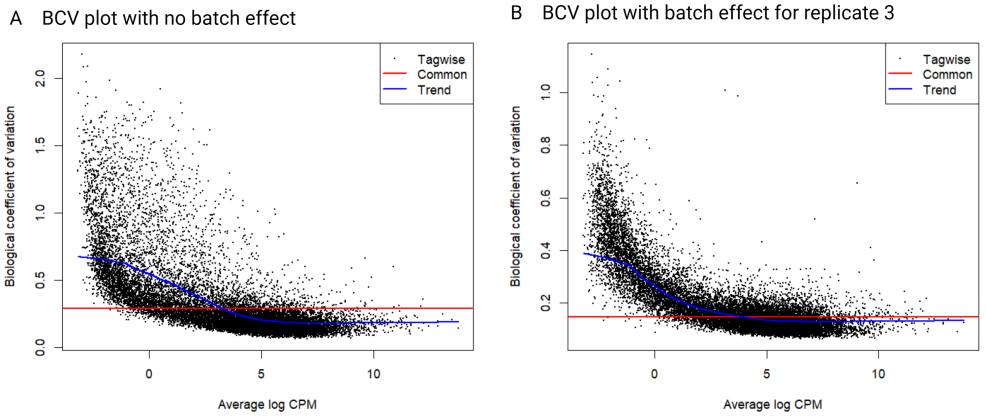


FIGURE 3.8: Biological coefficient of variation (BCV) plots against gene abundance with estimates of the common, trended and tagwise dispersions with and without batch effect for replicate 3. A) no batch effect included B) including a batch effect for replicate 3.

3.3.4 Sense check genes

Several genes that are known targets of EZH2is or isotretinoin in neuroblastoma were selected to sense check results. *RARB* encodes for retinoic acid receptor beta, which binds to retinoic acid and should be upregulated in samples treated with isotretinoin [244]. Another study found that *MYCN* amplified NB cell lines treated with isotretinoin rewires the core regulatory circuit and drives high expression of *SOX4* [227]. EZH2 has been shown to regulate NTRK1 in NB, and inhibition of EZH2 induced transcription of NTRK1 in NB cell lines [185]. Normalised counts were plotted for each sample to check gene expression levels. *RARB* and *SOX4* both showed higher expression in samples that had been treated with isotretinoin, with expression increasing from PT3 to PT10 (Figure 3.9A and B). *NTRK1* showed higher expression in samples treated with EZH2is (Figure 3.9C). These results were consistent with the literature.

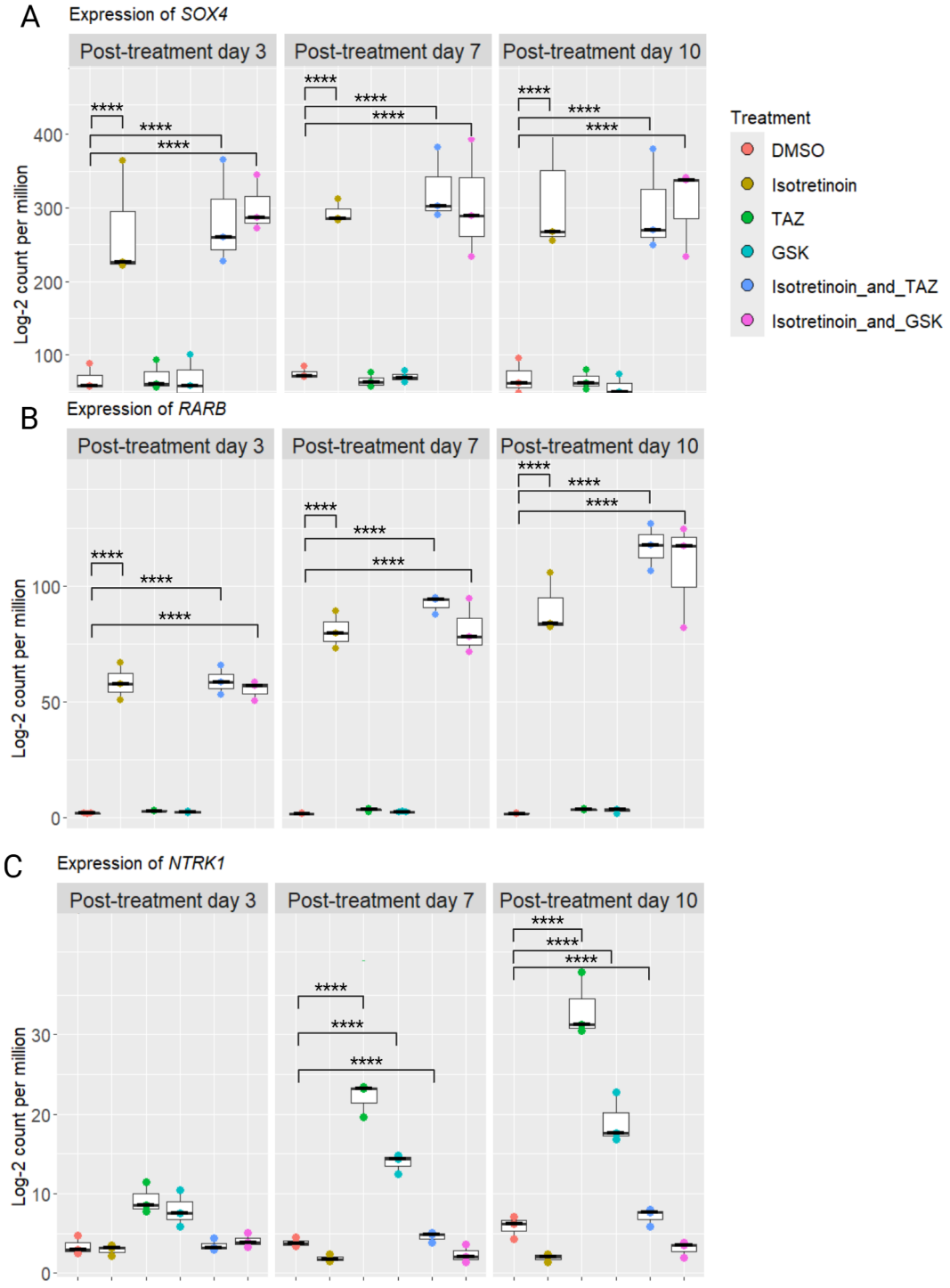


FIGURE 3.9: Normalised expression of sense check genes. A) *SOX4* B) *RARB* and C) *NTRK1* at post-treatment day 3, post-treatment day 7 and post-treatment day 10. Normalised expression was plotted as log-2 CPM. Dots represent samples. Significance is based on Benjamini-Hochberg adjusted p-values from edgeR. p >0.05, *: p ≤0.05, **: p ≤0.01, ***: p ≤0.001, ****: p ≤0.0001.

3.3.5 Number of differentially expressed genes

The number of upregulated and downregulated DEGs (LFC >1.5 or LFC <-1.5 and adjusted p value <0.05) for each treatment compared to the DMSO control was

compared (Figure 3.10 and Table 3.2).

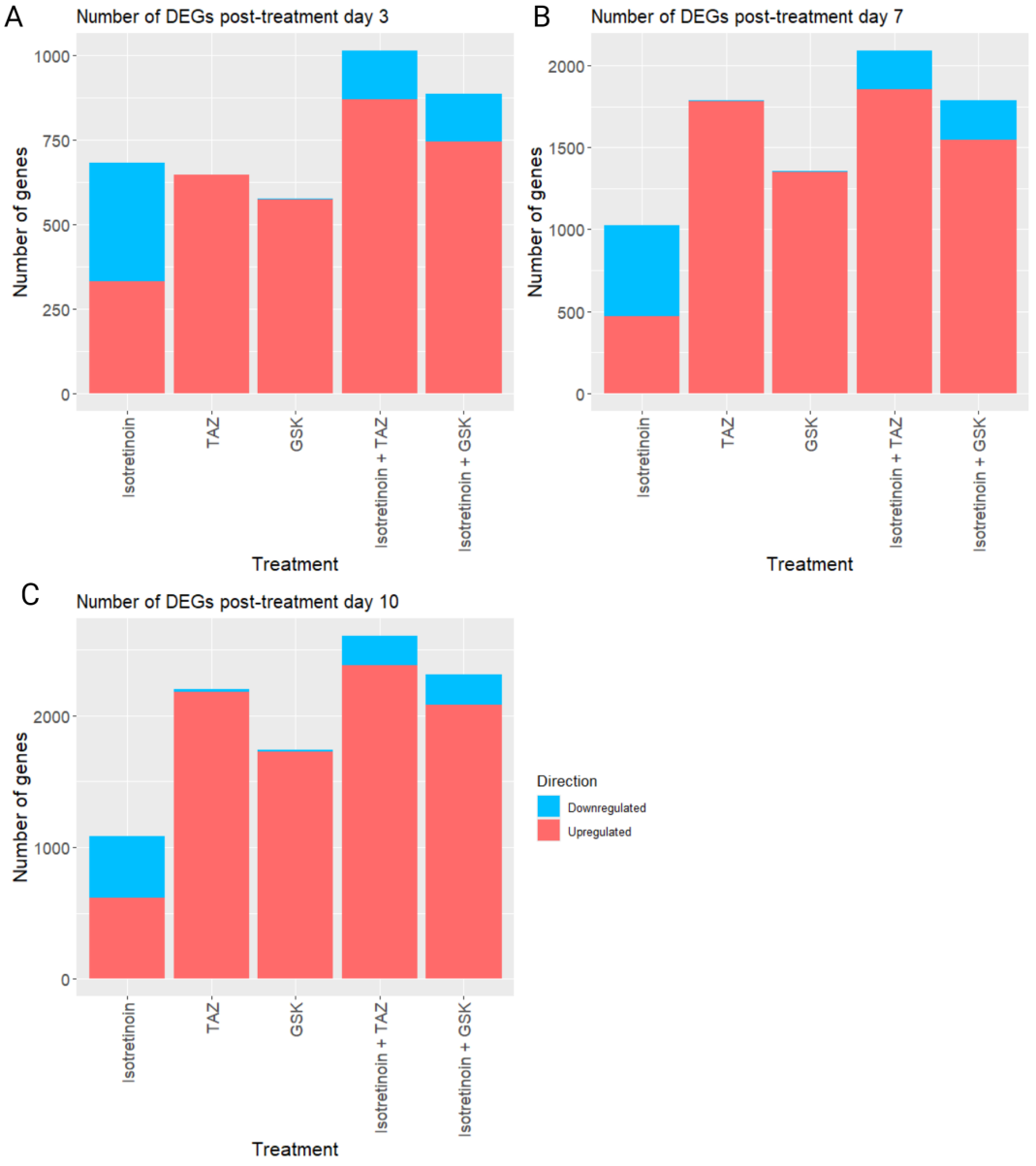


FIGURE 3.10: Bar plot showing the number of DEGs for each treatment compared to DMSO control at timepoints. A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg).

The number of DEGs increased from PT3 to PT10 for all treatments. Treatment with isotretinoin resulted in a roughly equal amount of upregulated and downregulated genes with a slight skew towards downregulated genes at PT3 and PT7 (333 upregulated and 349 downregulated at PT3, 473 upregulated and 553 downregulated at PT7) whilst more genes were upregulated than downregulated at PT10 (615 upregulated and 465 downregulated).

The majority of genes differentially expressed in EZH2i treated samples were upregulated, as expected given it's role in transcriptional repression and silencing

TABLE 3.2: Table summarising the number of significantly DEGs for each treatment condition vs DMSO control. Significant DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg).

	PT3		PT7		PT10	
Treatment condition	Upregulated	Downregulated	Upregulated	Downregulated	Upregulated	Downregulated
Isotretinoin	333	349	473	553	615	465
TAZ	647	0	1781	5	2177	26
GSK	575	1	1350	3	1726	17
Isotretinoin + TAZ	869	146	1854	237	2381	228
Isotretinoin + GSK	745	141	1544	245	2085	228

gene expression. TAZ resulted in 647 upregulated and 0 downregulated at PT3, 1781 upregulated and 5 downregulated at PT7 and 2177 upregulated and 26 downregulated at PT10. Treatment with GSK resulted in less DEGs than treatment with TAZ, with the smallest difference at PT3 and greater difference at PT7 and PT10 (575 upregulated and 1 downregulated at PT3, 1350 upregulated and 3 downregulated at PT7 and 1726 upregulated and 17 downregulated at PT10). This trend was also observed in the combination therapy, with isotretinoin + TAZ resulting in more DEGs (869 upregulated and 146 downregulated at PT3, 1854 upregulated and 237 downregulated at PT7, 2381 upregulated and 228 downregulated at PT10) than isotretinoin + GSK at each timepoint (745 upregulated and 141 downregulated at PT3, 1544 upregulated and 245 downregulated at PT7, 2085 upregulated and 228 downregulated at PT10). There was a lower number of downregulated genes in combination therapy compared to isotretinoin alone.

3.3.6 Overlap in differentially expressed genes

3.3.6.1 Venn Diagrams

Venn diagrams were plotted showing the number of overlapping DEGs for each treatment vs DMSO. The overlapping DEGs can be found in the Supplementary Files (Supplementary Files/Chapter 3 Neuroblastoma/SIGNIFICANT_DEGs). At PT3, there was overlap in DEGs between the EZH2is and isotretinoin, with 99 DEGs overlapping for TAZ and 87 for GSK (Figure 3.11A and B). There was a high number of DEGs shared between EZH2is, but around 100 DEGs exclusive to GSK and 170 to TAZ (Figure 3.11C). Since TAZ had a higher number of DEGs, this would account for some of the DEGs exclusive to TAZ.

When looking at the DEGs shared between isotretinoin, EZH2is and the combination therapy there was around 250 DEGs exclusive to combination therapy (Figure 3.11D and E). These are genes differentially expressed in isotretinoin + EZH2i, but not differentially expressed in single agent treatment with isotretinoin or EZH2i.

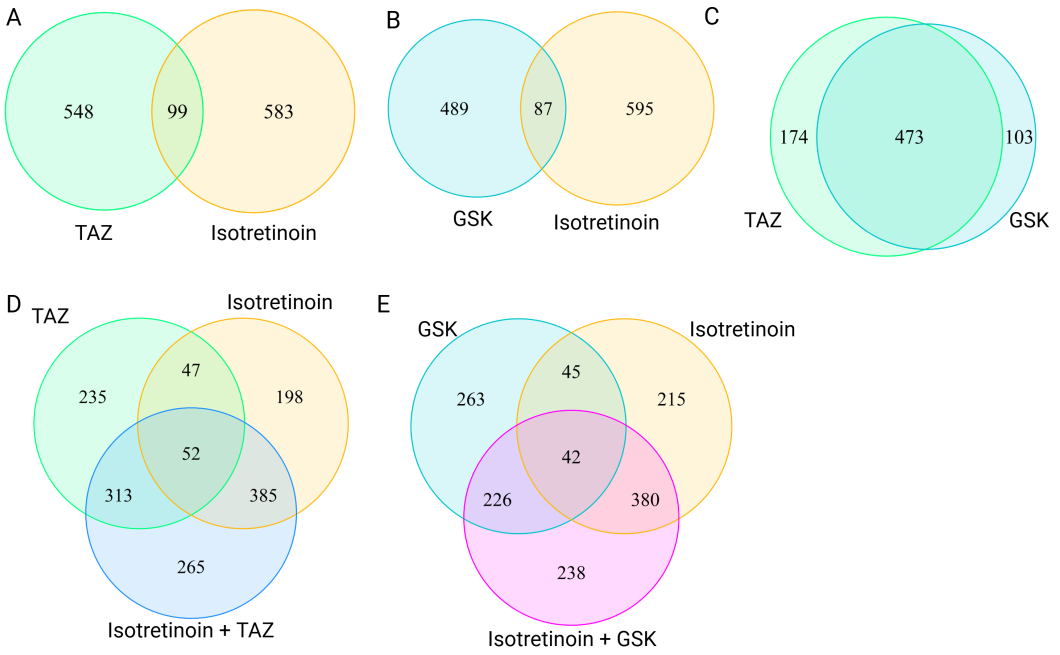


FIGURE 3.11: Venn diagrams showing overlapping DEGs (treatment vs DMSO) at post-treatment day 3 for A) isotretinoin and TAZ B) isotretinoin and GSK C) TAZ and GSK D) isotretinoin, TAZ and isotretinoin + TAZ E) isotretinoin, GSK and isotretinoin + GSK. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg)

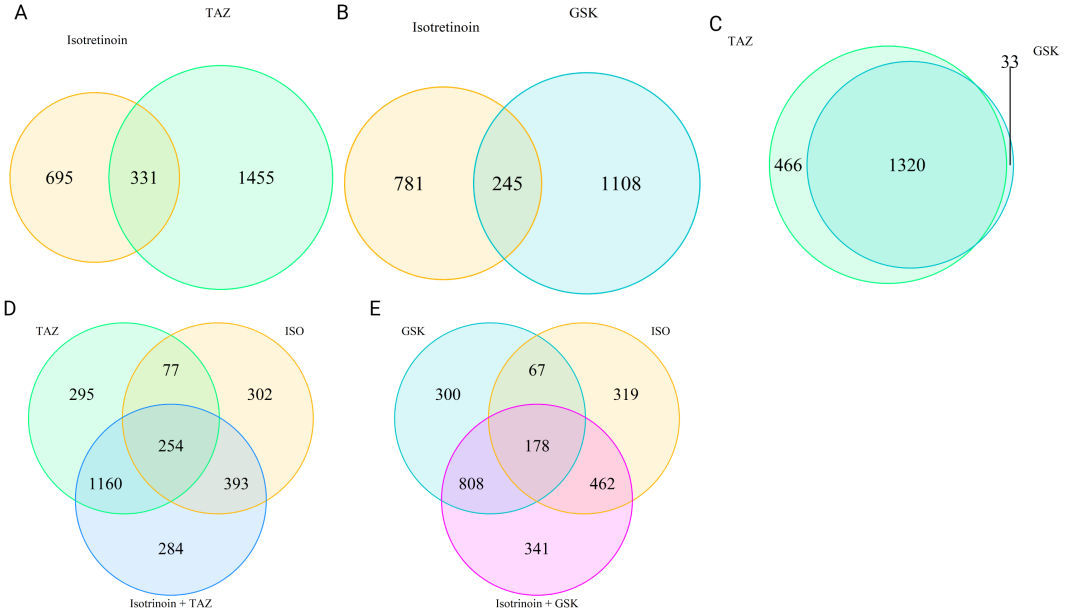


FIGURE 3.12: Venn diagrams showing overlapping DEGs (treatment vs DMSO) at post-treatment day 7 for A) isotretinoin and TAZ B) isotretinoin and GSK C) TAZ and GSK D) isotretinoin, TAZ and isotretinoin + TAZ E) isotretinoin, GSK and isotretinoin + GSK. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg)

The trends in overlapping DEGs at PT7 and PT10 were similar to PT3 just with increased numbers of DEGs (Figure 3.12 and Figure 3.13). One noticeable difference at

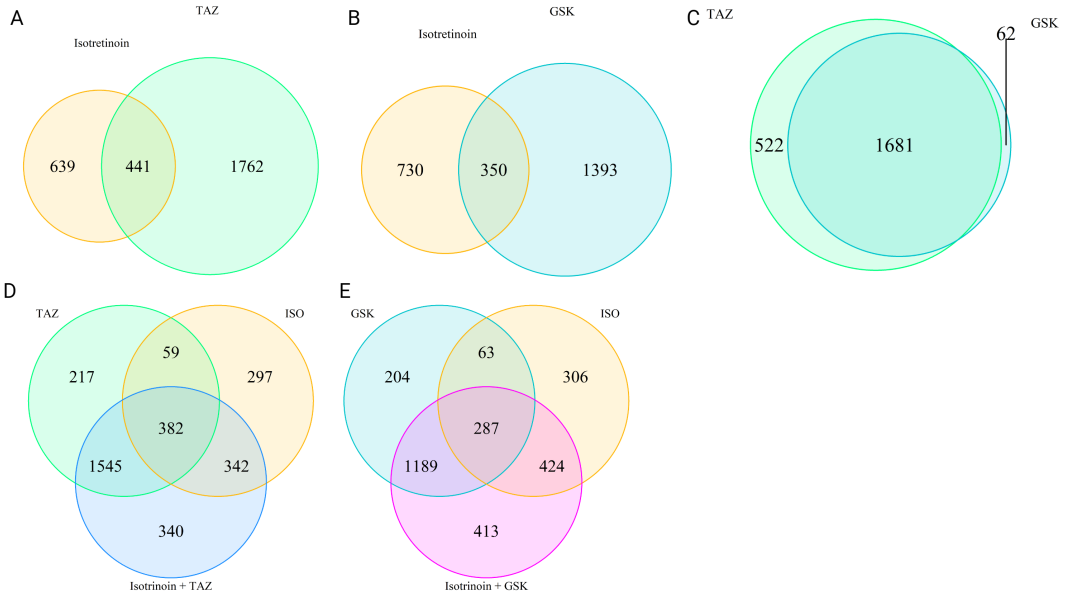


FIGURE 3.13: Venn diagrams showing overlapping DEGs (treatment vs DMSO) at post-treatment day 10 for A) isotretinoin and TAZ B) isotretinoin and GSK C) TAZ and GSK D) isotretinoin, TAZ and isotretinoin + TAZ E) isotretinoin, GSK and isotretinoin + GSK. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg)

PT7 was that there were fewer genes only differentially expressed by GSK and not by TAZ (Figure 3.12). At PT3 103 genes were expressed by GSK and not by TAZ but this reduced to 33 at PT7. These genes must be no longer significantly differentially expressed by GSK at PT7 or they are not differentially expressed by TAZ at PT3 but are at PT7.

3.3.6.2 Scatterplots

To further investigate the relationship between treatments and their DEGs, pairwise scatterplots comparing the LFC of genes for different treatments were plotted. Only pairwise comparisons of interest were made, including isotretinoin and TAZ, as well as GSK and TAZ. There was a weak positive correlation between the LFC of genes when comparing isotretinoin and TAZ (Figure 3.14A). Although the correlation is weak it is highly significant and present at all timepoints which may reflect the large number of genes. This correlation became stronger over time (r= 0.084, r= 0.09, r= 0.2) (Figure 3.14A, B, C) and the number of genes differentially expressed by both TAZ and isotretinoin also increased (99, 331, 441) (Figure 3.11A, Figure 3.12A, Figure 3.13A). Nearly all yellow points (indicating differentially expressed in both treatments) were to the right of the y-axis, showing they were upregulated by TAZ. This is expected as TAZ had very few downregulated DEGs. About half of these genes were also upregulated by isotretinoin (yellow points in the top right of the plots) whilst around half were downregulated by isotretinoin (yellow points at the bottom right of the plots).

When looking at the differential expression of genes for TAZ vs DMSO and GSK vs DMSO, there is a strong significant positive correlation (Figure 3.14D, E and F). The majority of genes upregulated by TAZ are also upregulated by GSK, visualised as points to the upper right of the y-axis. At PT3, the number of genes exclusively upregulated by TAZ (green) is similar to the number of genes exclusively upregulated by GSK (turquoise) (Figure 3.14D). At later timepoints, there are more genes exclusively upregulated by TAZ than GSK (shown in turquoise) (Figure 3.14E and F). Most of these exclusive genes meet the LFC threshold in TAZ but not in GSK. A western blot shows that treatment with TAZ shows a greater reduction in the H3K27me3 mark than GSK at PT7 and PT10 (Figure 3.2). This likely explains why TAZ results in more DEGs than GSK.

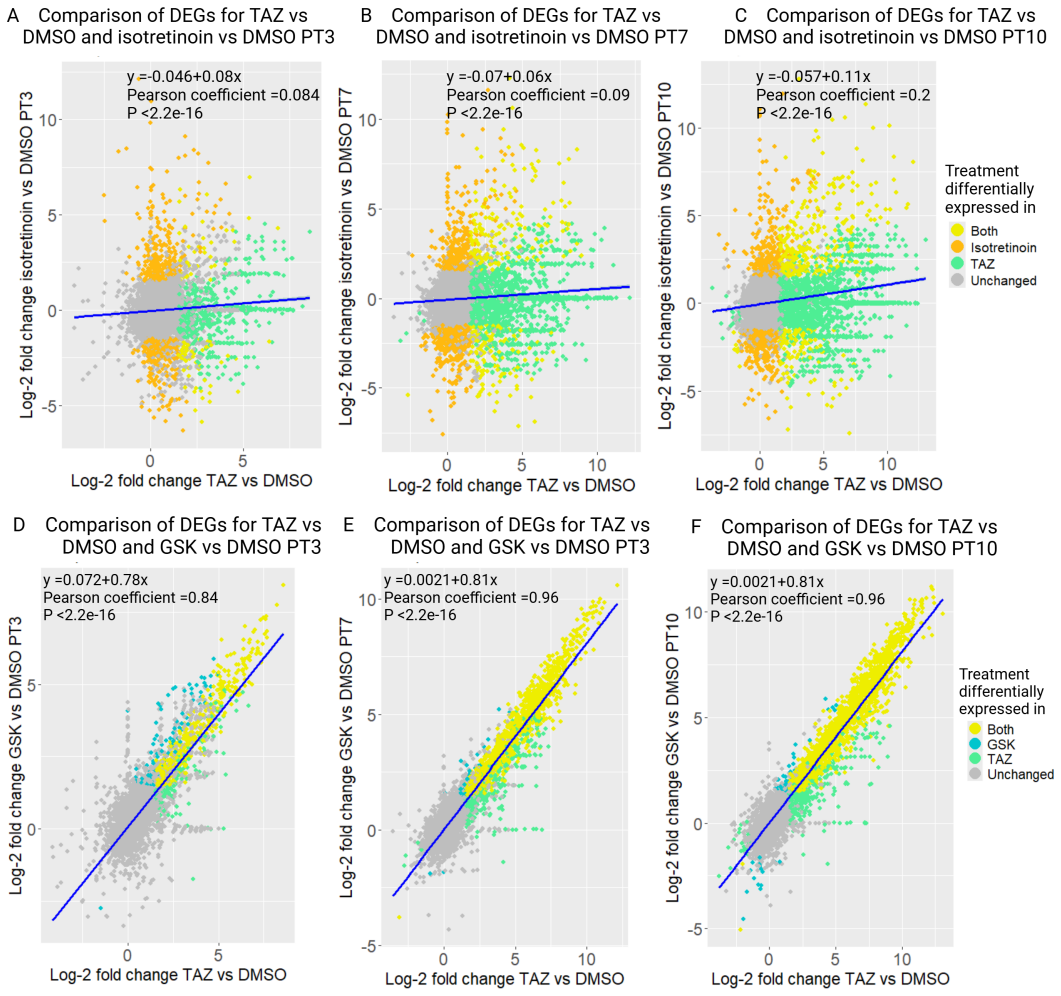


FIGURE 3.14: Scatterplot of the LFC of all genes comparing isotretinoin vs DMSO with TAZ vs DMSO A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Comparing GSK vs DMSO and TAZ vs DMSO D) post-treatment day 3 E) post-treatment day 7 and F) post-treatment day 10. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg). DEGs are highlighted in colour, corresponding to the treatment group in which they are differentially expressed. N=16,787

3.3.7 Visualisation of differentially expressed genes

Isotretinoin resulted in both upregulated and downregulated DEGs, the with number of DEGs increasing from PT3 to PT7 and PT10 (Figure 3.15A, B and C). Some of the most significant DEGs were retinoic response genes including *RARB*, *RET*, *DHRS3*, *CYP26B1* and *RXRG*.

Most DEGs were upregulated with EZH2is treatment, with there being no downregulated DEGs at PT3 by TAZ and only one gene was downregulated by GSK (*TRIML1*) (Figure 3.16A and D). At later timepoints, TAZ had more downregulated DEGs than GSK (Figure 3.16B, C, E and F). The increase in the number of DEGs from PT3 through to PT10 can be clearly seen in the volcano plots (Figure 3.16A, B and C).

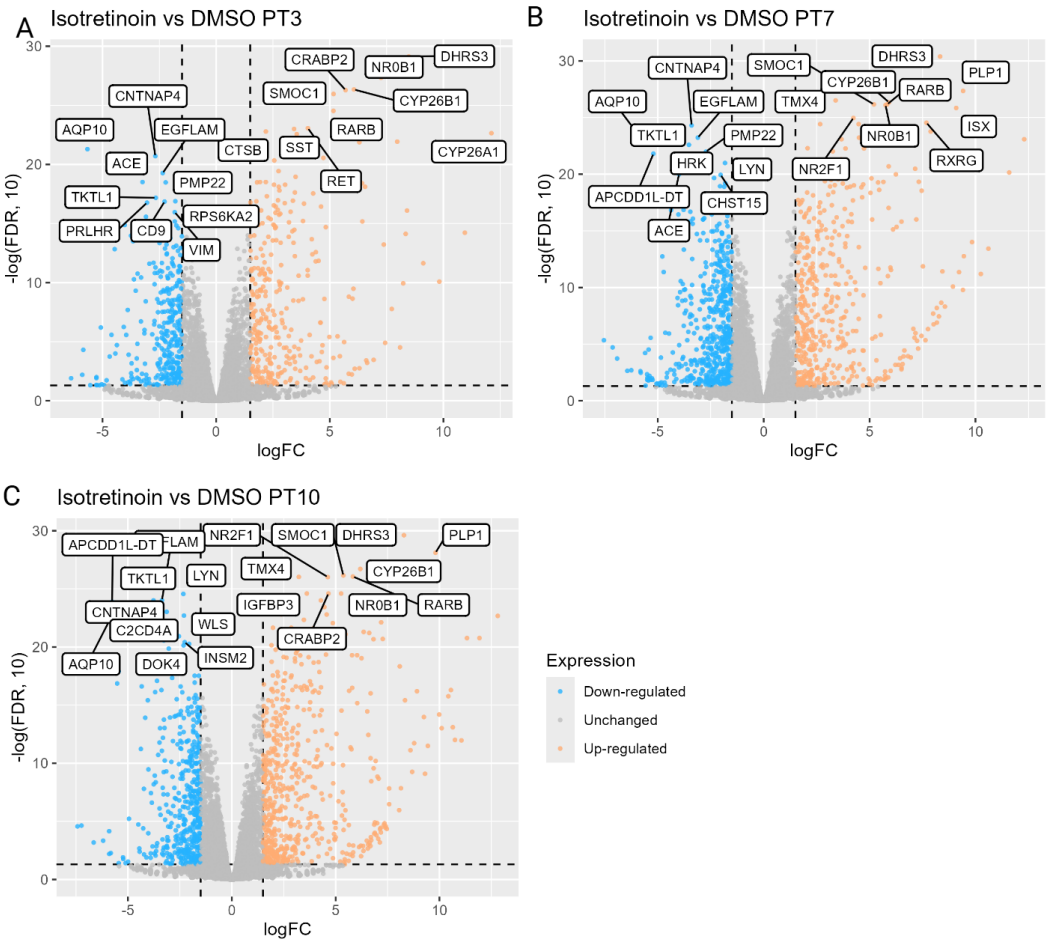


FIGURE 3.15: Volcano plots of DEGs for isotretinoin compared to control at A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg). Top ten upregulated and downregulated genes with the lowest p.adjust values are labelled.

Overall, the expression pattern of DEGs is very similar between the two EZH2is. Three of the top ten DEGs included genes encoding for proteins involved in microtubule dynamics and cytoskeletal organisation, *TUBB4A*, *WASF3* and *COTL1* [245, 246, 247]. Other genes were involved in DNA and RNA binding, including RADX, a single strand DNA binding protein that regulates double-strand break repair [248] and YBX3, a RNA binding protein that regulates amino acid levels [249].

Combination therapy resulted in both upregulated and downregulated DEGs, with more genes being upregulated than downregulated. The expression pattern was similar between isotretinoin + TAZ and isotretinoin + GSK (Supplementary Figure 5). Number of DEGs increased from PT3 through to PT10.

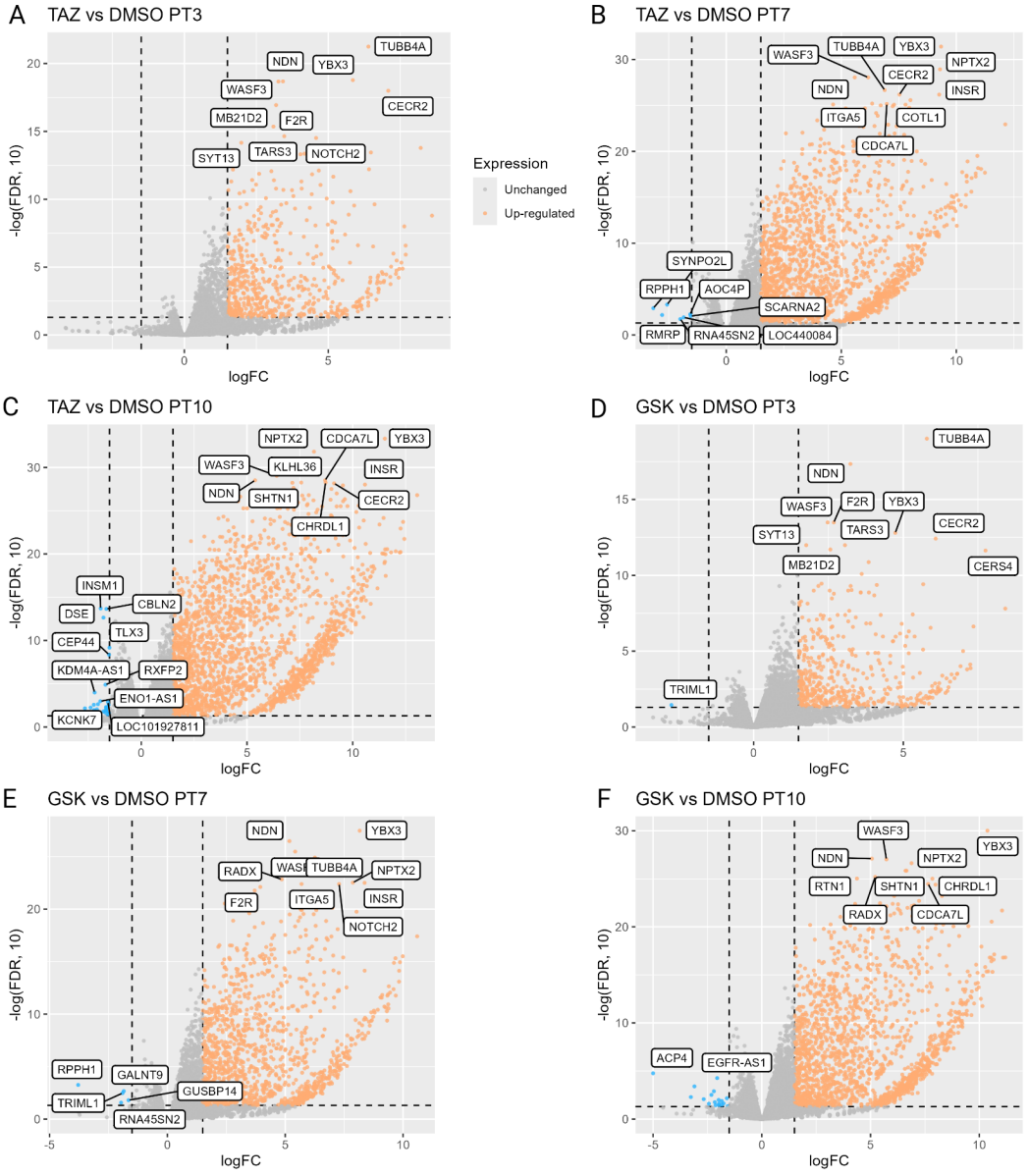


FIGURE 3.16: Volcano plots of DEGs for EZH2is TAZ and GSK compared to control. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg). TAZ at A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. ; GSK at D) post-treatment day 3 E) post-treatment day 7 and F) post-treatment day 10. Top ten upregulated and downregulated genes with the lowest p.adjust values are labelled.

3.3.8 Identification of biological functions and biochemical pathways enriched with treatment

All GSEA results for Reactome, GO terms and Hallmarks can be found in the Supplementary Files (Supplementary Files/Chapter 3 Neuroblastoma/GSEA GO terms and Supplementary Files/Chapter 3 Neuroblastoma/GSEA HALLMARK).

3.3.8.1 Isotretinoin

At PT3 the majority of Reactome gene sets and largest GO gene sets (containing the most genes enriched) in isotretinoin treated samples were retinoic acid-related (Supplementary Figure 6A, B and C and Figure 3.17). These included terms related to the retinoid cycle, 'retinoic acid biosynthesis' (∼15 genes) and 'signalling by retinoic acid' (∼35 genes) (Supplementary Figure 6A, B and C). There was upregulation of developmental and cell differentiation, such as 'enteric nervous system development', 'limb bud formation' and 'regulation of neuron maturation', with gene set sizes of around 8-13 genes (Figure 3.17A). GSEA results were very similar at PT7 and PT10 (Figure 3.17B and C), however at PT10 the largest GO gene set changed to 'proximal distal pattern formation' instead of a retinoic acid-related term (Figure 3.17C). Downregulated gene sets included terms relating to adrenergic receptor activity (receptors involving noradrenaline and adrenaline), laminin, meiotic spindle and glycerol transport Figure 3.17A, B and C).

No Hallmark gene sets were enriched at any timepoint. This might be expected as there are fewer Hallmark gene sets (50) than GO terms (10461) or Reactome (1692). Manually looking at the Hallmark gene sets, no relevant gene sets (i.e. retinoic acid or differentiation-related) were included, which may also explain why none were enriched.

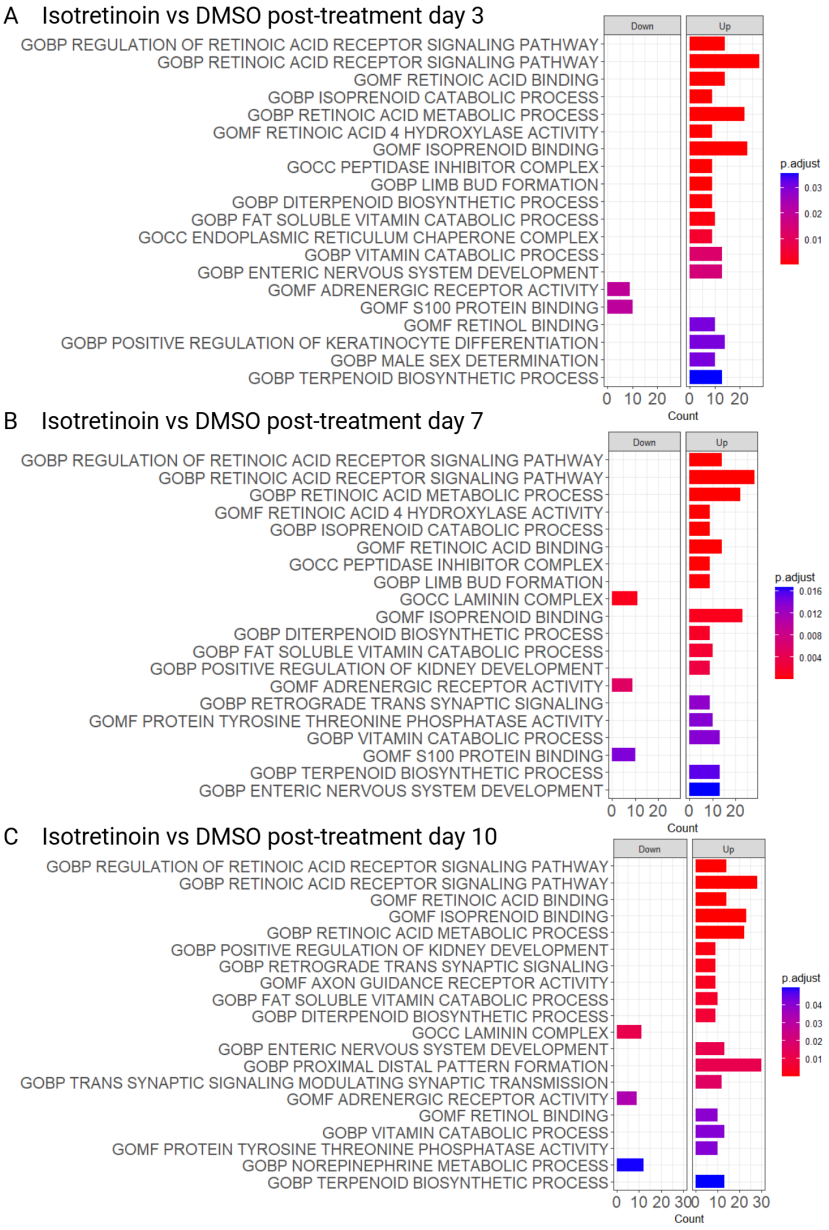


FIGURE 3.17: GSEA results from CAMERA showing enriched GO terms in isotretinoin compared to control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

3.3.8.2 TAZ

All GO gene sets enriched in samples treated with EZH2i TAZ were upregulated (Figure 3.18).

At PT3 the largest GO gene sets related to receptor protein kinase activity and there was enrichment of differentiation and development-related pathways (Figure 3.18A). The largest upregulated differentiation/development-related GO gene sets included

'bone morphogenetic protein (BMP) signalling' (which regulates embryonic patterning), 'renal tubule development' and 'cardiac atrium morphogenesis' (Figure 3.18A). These gene sets contained around 30 genes. There were also upregulated immune-related gene sets; 'interleukin 13 production', 'MHC protein complex', 'response to interleukin 17' and 'MHC class 1 protein complex' (Figure 7A).

No Hallmark gene sets were enriched at PT3 or PT7. At PT7, immune-related gene sets seemed to become more prominent. The GO terms 'immune receptor activity' and 'cytokine receptor activity' were the largest upregulated gene sets (66 genes and 57 respectively) and were not present at PT3 (Supplementary Figure 7A and B). The Reactome gene set 'immunoregulatory interactions between a lymphoid and non-lymphoid cell' was also upregulated in PT7 but not PT3. This gene set contains genes involved in modifying the response of lymphoid cells (e.g. B-cells, T-cells and NK cells) to self and tumour antigens. There were also more differentiation gene sets that were not present at PT3 and some from PT3 showed an increase in size, for example BMP signalling went from 34 genes to 89 genes (Figure 3.18B). At PT10, in addition to immune and differentiation gene sets, terms linked to the ECM and growth factors were among the largest gene sets; such as 'extracellular matrix structural constituent', 'growth factor binding' and 'collagen chain trimerization' (Figure 3.18C). Other large immune gene sets were related to cytokines and the Hallmark 'inflammatory response' (Supplementary Figure 7C). The Hallmark 'KRAS signalling up' was also upregulated (Supplementary Files/Chapter 3 Neuroblastoma/GSEA HALLMARK/Hallmarks_TAZvsDMSO_PT10.csv). To summarise, gene sets enriched by TAZ at PT3 appeared to be related to differentiation, kinase activity and immune gene sets. At PT7, immune-related gene sets and differentiation became more prominent. At PT10, in addition to differentiation and immune gene sets, there were large ECM gene sets.

FIGURE 3.18: GSEA results from CAMERA showing enriched GO terms for TAZ vs control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

3.3.8.3 GSK

The largest enriched GO gene sets in treatment group GSK at PT3 were linked to ribosomes, which were not upregulated in TAZ (Supplementary Figure 8A). Other

upregulated GO terms were immune-related terms. Multiple Reactome gene sets linked to translation were upregulated, and these were not observed in treatment with TAZ (Supplementary Figure 9A). No Hallmark gene sets were upregulated at PT3 or PT7. Like TAZ at PT7, GSK resulted in more differentiation and immune-related gene sets than at PT3 and these consisted of a large number of genes. Other GO terms upregulated were related to kinase activity, similar to TAZ at PT3 (Supplementary Figure 8B). At PT10 the largest gene sets were related to kinase activity, cytokine activity, differentiation and other immune genesets (Supplementary Figure 8C). The size of the largest gene sets were considerably smaller than TAZ at PT10, with the largest gene set being 77 genes whilst TAZ had 334. ECM terms were upregulated by GSK at PT10, but these contained fewer genes than TAZ; 48 whilst TAZ had 125. Noticeably, there were no large gene sets involved in growth factor signalling as seen in TAZ at PT10. Overall, enriched gene sets by GSK were quite different to TAZ at PT3 but became more similar at PT7 and PT10. There were shared themes at the later timepoints, including differentiation, immune-related gene sets and ECM, although overall fewer enriched gene sets containing fewer genes.

3.3.8.4 Isotretinoin + TAZ

For isotretinoin + TAZ, the majority of upregulated GO gene sets at PT3 were related to differentiation and development, including morphogenesis, 'proximal distal pattern formation', 'cranial nerve morphogenesis' and nervous system development (Supplementary Figure 10A). A number of retinoic acid terms were also upregulated. Reactome pathways enriched at PT3 were isotretinoin driven (Supplementary Figure 11A). No Hallmark gene sets were enriched at PT3 or PT7. PT7 was similar to PT3, but gene sets contained a greater number of genes. Immune gene sets were upregulated, but fewer than single agent treatment with TAZ and the large immune gene sets upregulated by TAZ, such as 'immune receptor activity', were not enriched. Enriched Reactome pathways at PT7 were retinoic acid and fibrin related (Supplementary Figure 11B). PT10 also showed upregulation of differentiation terms. At PT10, there was downregulation of terms involved in DNA replication consisting of large numbers of genes (Figure 3.19C). These included GO terms; 'DNA unwinding involved in DNA replication', 'DNA replication initiation' and 'positive regulation of chromosome segregation'. Reactome pathways; 'DNA strand elongation', 'unwinding of DNA' and 'activation of the pre-replicative complex' and Hallmark; 'G2M checkpoint'. Some of these terms were not seen in single-agent treatment.

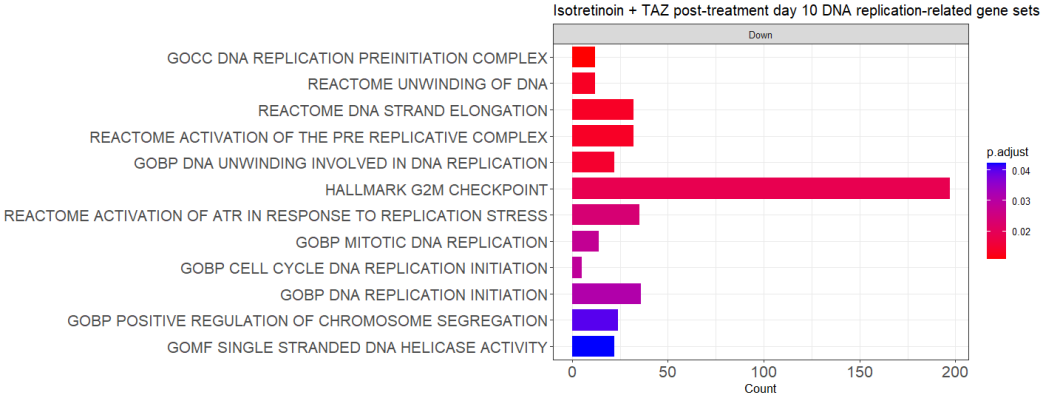


FIGURE 3.19: GSEA results from CAMERA showing enriched DNA-replication-related gene sets in isotretinoin + TAZ compared to control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

3.3.8.5 Isotretinoin + GSK

Isotretinoin + GSK showed very similar enriched gene sets to isotretinoin + TAZ at PT3 and PT7, consisting mainly of differentiation and retinoic acid gene sets (Supplementary Figure 12A and B). Like isotretinoin + TAZ at PT3, upregulated Reactome pathways were retinoic acid related (Supplementary Figure 12A). However at PT7 Reactome pathways involving fibrin were not upregulated as they were in

isotretinoin + TAZ, only at PT10 (Supplementary Figure 12B and C). Immune-related gene sets were upregulated, but the larger immune gene sets seen in single agent EZH2i treatment were not enriched in combination therapy. PT10 consisted mainly of differentiation terms, but unlike isotretinoin + TAZ, DNA replication-related terms were not downregulated in isotretinoin + GSK. No Hallmark gene sets were enriched at any timepoint.

To further investigate the impact of treatments on the immune response, 66 genes from the GO term 'immune response' were visualised in a heatmap (Figure 3.20). TAZ PT10 samples show the highest expression of these immune genes. The expression of these genes separates into two main clusters with hierarchical clustering. Higher expression is seen in EZH2i-treated samples at PT7 and PT10, as well as isotretinoin + TAZ at PT10 and lower expression is seen in DMSO, isotretinoin, isotretinoin + GSK, EZH2is at PT3 and isotretinoin + TAZ at PT3 and PT7.

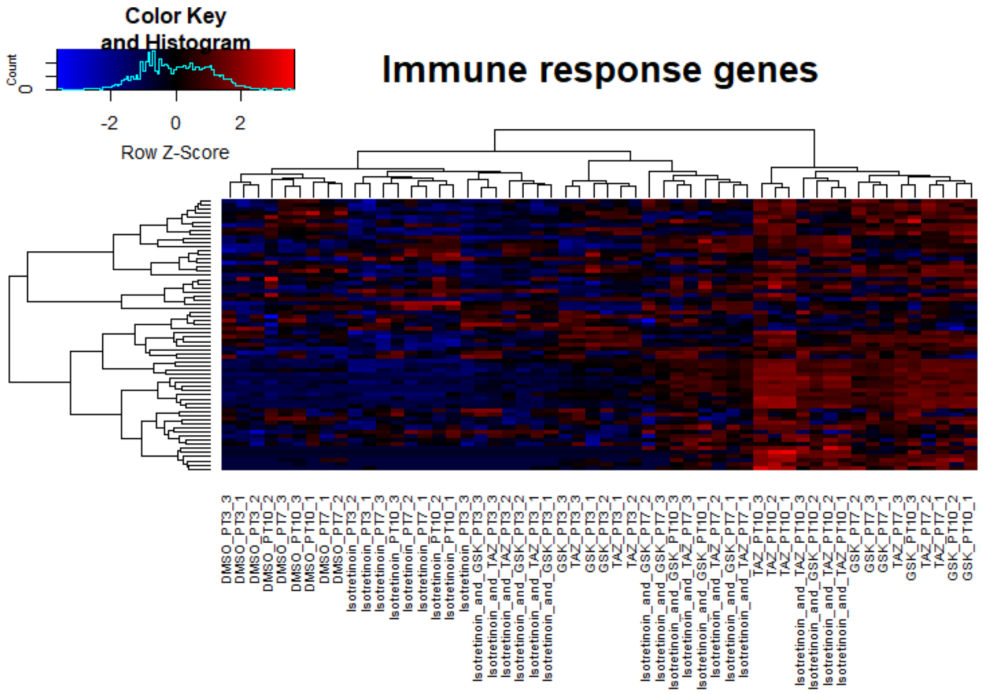


FIGURE 3.20: Heatmap of the 66 genes from the GO term 'immune response' in TAZ-treated samples at PT10. Z score has been applied to CPM-normalised RNA-seq counts.

3.3.9 Additive genes

Additive genes, that show a greater LFC in combination therapy compared to either single agent, may contribute to the enhanced effect of combination therapy compared to single agent treatment in reducing cell viability seen in cell models.
Over-representation-based enrichment analysis was used to investigate the function of additive genes. This was only done for isotretinoin + TAZ at PT10 to determine

whether DNA replication and cell cycle genes were indeed additive genes as suggested by their presence in GSEA at isotretinoin + TAZ at PT10 but not either single agent treatment. Additive upregulated and downregulated gene lists can be found in the Supplementary material (Supplementary Files/Chapter 3 Neuroblastoma/Additive genes).

There were 2,112 upregulated additive genes. On average, the difference in LFC of upregulated additive genes compared to single agent treatment was 1.05. The smallest difference in LFC was 0.00023, and the largest was 11.46. Enriched gene sets for upregulated additive genes were neuronal-related (Figure 3.21A and B). 2,650 downregulated genes showed a greater LFC in combination therapy compared to single agent treatment. Downregulated additive genes had a smaller difference in LFC between single agent treatment compared to upregulated additive genes, with an average difference in LFC of -0.27. The greatest difference in LFC was -5.11 and smallest difference was -0.000041. Enriched gene sets for downregulated additive were predominately DNA replication and cell cycle based (Figure 3.21C and D).

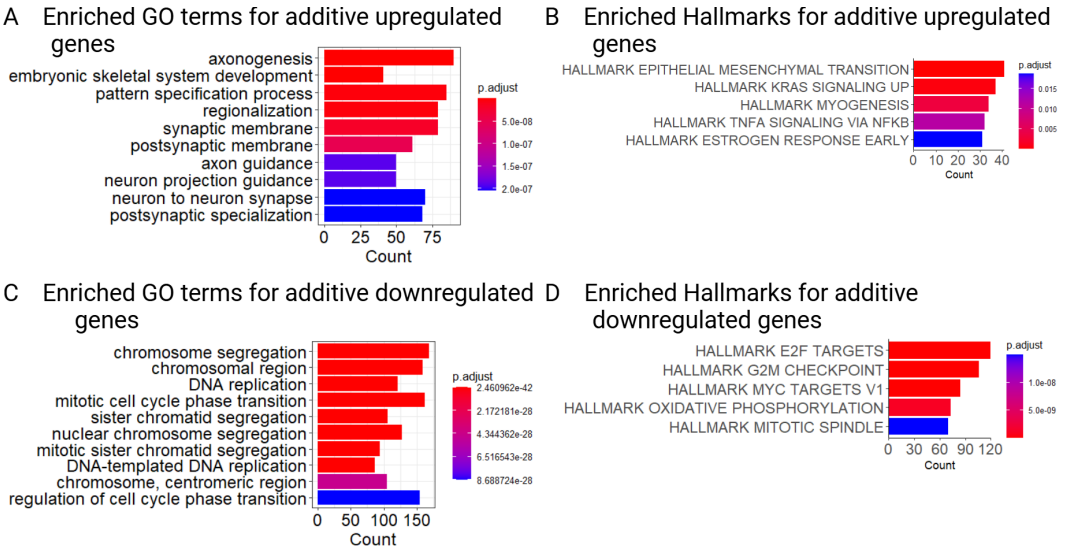


FIGURE 3.21: Over-representation based enrichment analysis of upregulated and downregulated additive genes. Additive genes were defined as genes which showed a greater LFC in combination therapy compared to single agent treatment. A) Enriched GO terms for upregulated additive genes B) enriched Hallmarks for upregulated additive genes C) enriched GO terms for downregulated additive genes D) enriched Hallmarks for updownregulated additive genes. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

3.3.10 Enrichment of neuroblastoma differentiation gene signature in isotretinoin and combination therapy

To further support the GSEA results suggesting that combination therapy resulted in differentiation, a neuroblastoma differentiation signature consisting of 59 genes was

tested for enrichment [236]. The signature was found to be significantly enriched in isotretinoin, isotretinoin + TAZ and isotretinoin + GSK compared to control at all timepoints (Supplementary Table 4). A random 59 gene signature was also selected for comparison, with no samples showing significant enrichment (Pvalue >0.05) (Supplementary Table 6).

3.3.11 Unsupervised identification of sets of genes with similar gene expression patterns through k-means clustering

Genes that were significantly differentially expressed in at least one treatment condition were input for k-means clustering analysis. This unsupervised analysis was used to identify sets of genes with similar patterns in gene expression. The elbow method suggested the optimum k value (number of clusters) as 3, whilst the silhouette method suggested the optimum value as 2 (Supplementary Figure 13). The optimum k value was selected as 3 as this was the optimum value using the Elbow method and second optimum value using the Silhouette method. The genes from each cluster were plotted in a heatmap (Figure 3.22) and analysed for over-representation-based GSEA. The genes and all GSEA results for each cluster can be found in the Supplementary Files (Supplementary Files/Chapter 3 Neuroblastoma/k-means clustering results). There were 474 genes in cluster 1, 1058 in cluster 2 and 211 in cluster 3.

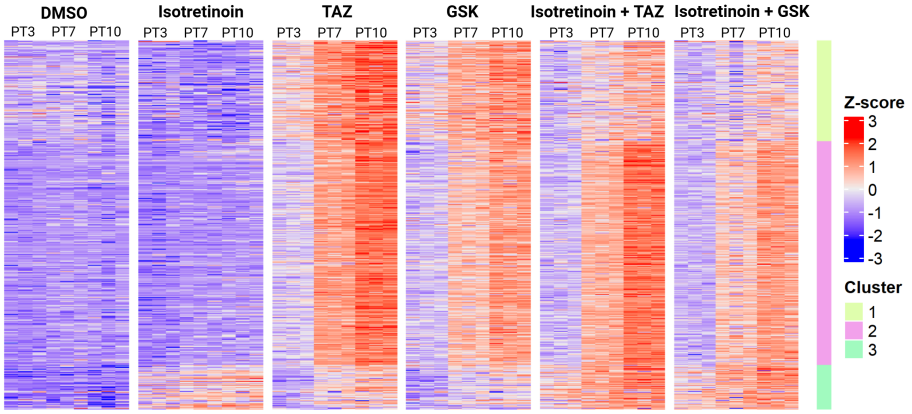


FIGURE 3.22: Heatmap of genes significantly differentially expressed in at least one comparison. Z score has been applied to CPM-normalised RNA-seq counts. Clusters were assigned using k-means clustering on the scaled data. Significant DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg).

Cluster 1 showed very high expression in EZH2i treated samples, lower expression in combination therapy and the lowest expression in DMSO and isotretinoin samples (Figure 3.22). These could represent genes upregulated by EZH2is but expression is reduced in combination with isotretinoin. In EZH2i and combination therapy samples, expression increased with time. Results from enrichment analysis showed that these genes were enriched for GO terms involving growth factor binding, DNA

transcription and ECM (Figure 3.23A). Enriched Hallmarks were 'inflammatory response' and EMT (Figure 3.23B).

Cluster 2 genes showed around equal expression in EZH2is and combination therapy, with low expression in isotretinoin and DMSO (Figure 3.22). Again, in EZH2i and combination therapy samples, expression increased with time. These genes were enriched for GO terms involving the ECM and neuronal differentiation (Figure 3.23C). Enriched Hallmarks included those enriched in cluster 1, as well as interferon response and KRAS signalling (Figure 3.23D).

Genes in cluster 3 showed low expression in DMSO, high expression in isotretinoin and EZH2i and were more highly expressed in the combination therapy samples (Figure 3.22). These genes were enriched for neuronal and differentiation GO terms (Figure 3.23E). Hallmarks enriched were EMT and 'angiogenesis' (Figure 3.23F).

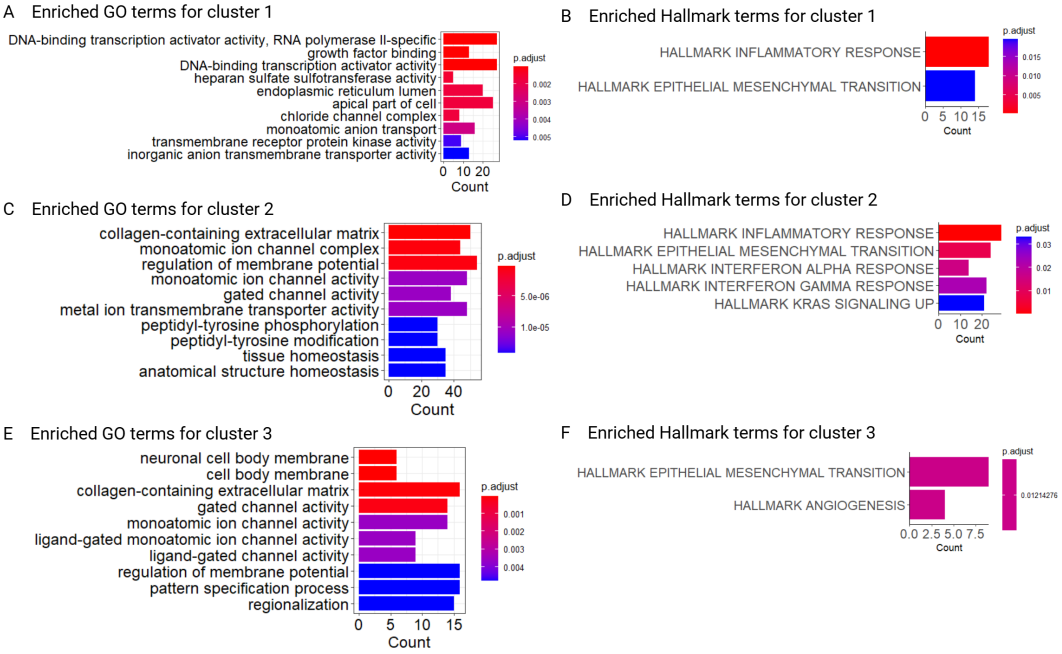


FIGURE 3.23: Over-representation analysis results for cluster 1 genes for A) GO terms B) Hallmark and C) Reactome. Significantly enriched pathways were defined as having an adjusted p value (Benjamini-Hochberg) <0.05. Top 20 most significant GO terms were shown if more than 20 terms were enriched.

3.3.12 Treatment vs treatment comparisons

To determine if combination therapy resulted in further differentiation of NB cells compared to single agent treatment, pairwise treatment vs treatment comparisons were made. This had been partially investigated through looking at additive genes as well as cluster 3 in k-means clustering analysis. Comparing combination therapy to isotretinoin will likely reflect genes differentially expressed by EZH2is, but may also

show changes in gene expression resulting from the interactions between isotretinoin and EZH2is in combination.

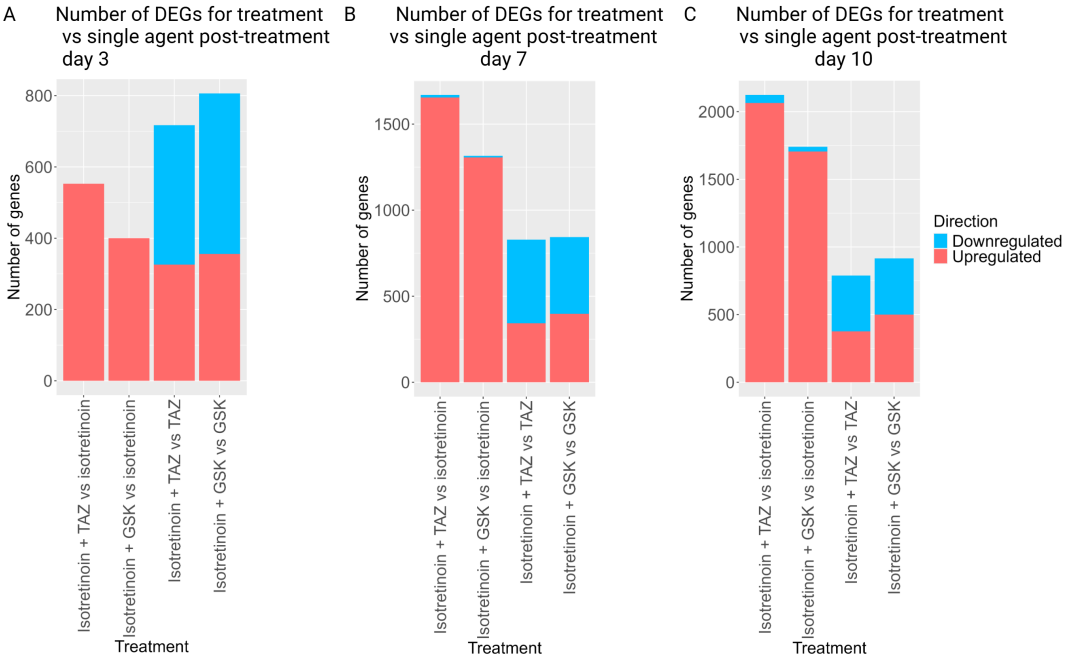


FIGURE 3.24: Over-representation analysis results for cluster 3 genes for A) GO terms B) Hallmark and C) Reactome. Significantly enriched pathways were defined as having an adjusted p value (Benjamini-Hochberg) <0.05. Top 20 most significant GO terms were shown if more than 20 terms were enriched.

At PT3 combination therapy vs isotretinoin resulted in only upregulated genes, with isotretinoin + TAZ resulting in around 550 DEGs and isotretinoin + GSK resulting in around 400 DEGs (Figure 3.24A). This is expected as these are likely genes differentially expressed by EZH2is. Similar to treatment vs control comparisons, TAZ resulted in more DEGs than GSK. The number of DEGs for combination therapy vs isotretinoin was less than the number of genes differentially expressed by TAZ and GSK vs control (∼650 and 575 respectively), which supports evidence from the scatterplots that isotretinoin might reduce the expression of some genes upregulated by EZH2is. Similar trends were shown at PT7 and PT10, with the majority of DEGs for combination therapy vs isotretinoin being upregulated with the number of DEGs increasing from PT3 through to PT10 (Figure 3.24A, B and C). Again, isotretinoin + TAZ vs isotretinoin resulted in more DEGs than isotretinoin + GSK vs isotretinion. When looking at combination therapy vs EZH2is, genes are both upregulated and downregulated (likely reflecting genes differentially expressed by isotretinoin), with number of DEGs increasing from PT3 to PT10 (Figure 3.24A, B and C).

GSEA was performed for treatment vs treatment comparisons. All GSEA results can be found in the Supplementary Files (Supplementary Files/Chapter 3 Neuroblastoma/Treatment vs treatment GSEA).

3.3.12.1 Combination therapy vs isotretinoin

Reactome pathways enriched in isotretinoin + TAZ compared to isotretinoin at PT3 were similar to terms enriched in TAZ vs DMSO, with 4 out of 9 terms also enriched in TAZ vs DMSO at PT3 (Supplementary Figure 14A). At PT10, DNA replication-related terms were downregulated including 'unwinding of DNA', 'DNA strand elongation' and 'activation of the pre replicative complex' (Supplementary Figure 14C). This is further evidence to support that combination therapy but not single agent may be downregulating the DNA damage response or cell proliferation. Also enriched at PT10 is the Reactome pathway 'immunoregulatory interactions between a lymphoid and non-lymphoid cell'. This could suggest that the addition of TAZ could be promoting the immune response when compared to isotretinoin alone.

Enriched GO terms for isotretinoin + TAZ vs isotretinoin at PT3 and PT7 were related to development, differentiation and neuronal differentiation, including terms such as 'nerve development', 'cranial nerve morphogenesis' and 'forebrain regionalisation' (Figure 3.25A and B). At PT10 there were more enriched terms that suggested further neuronal differentiation, including 'gamma-aminobutyric acid signalling pathway' and 'astrocyte differentiation' (Figure 3.25C). As seen for the Reactome pathways, at PT10 'DNA unwinding involved in DNA replication' was downregulated. The only enriched Hallmark at any timepoint for isotretinoin + TAZ vs isotretinoin was upregulation of 'E2F targets', transcription targets involved in cell cycle regulation and DNA replication. Isotretinoin + GSK vs isotretinoin showed very similar enriched GO terms and Reactome pathways. In conclusion, isotretinoin + TAZ may result in further neuronal differentiation compared to isotretinoin alone, with cells becoming more differentiated with time. There is also some evidence for isotretinoin + TAZ downregulating DNA damage response/ DNA replication compared to isotretinoin alone.

3.3.12.2 Combination therapy vs EZH2i

To further investigate the role of isotretinoin in combination therapy, GSEA was performed comparing combination therapy to EZH2is. Enriched Reactome pathways at PT3, PT7 and PT10 were related to retinoic acid, with the majority of these terms also seen enriched in GSEA of isotretinoin vs DMSO. When looking at enriched GO terms for isotretinoin + TAZ vs TAZ at PT3,10/15 top pathways were related to retinoic acid signalling (Figure 3.26A). At PT7 and PT10, terms were related to retinoic acid signalling and there were terms to suggest cells were more differentiated, including 'proximal distal pattern formation', 'embryonic skeletal system morphogenesis' and 'serotonin transport' (Figure 3.26B and C). These results might suggest that isotretinoin is also responsible for some of the differentiation seen in

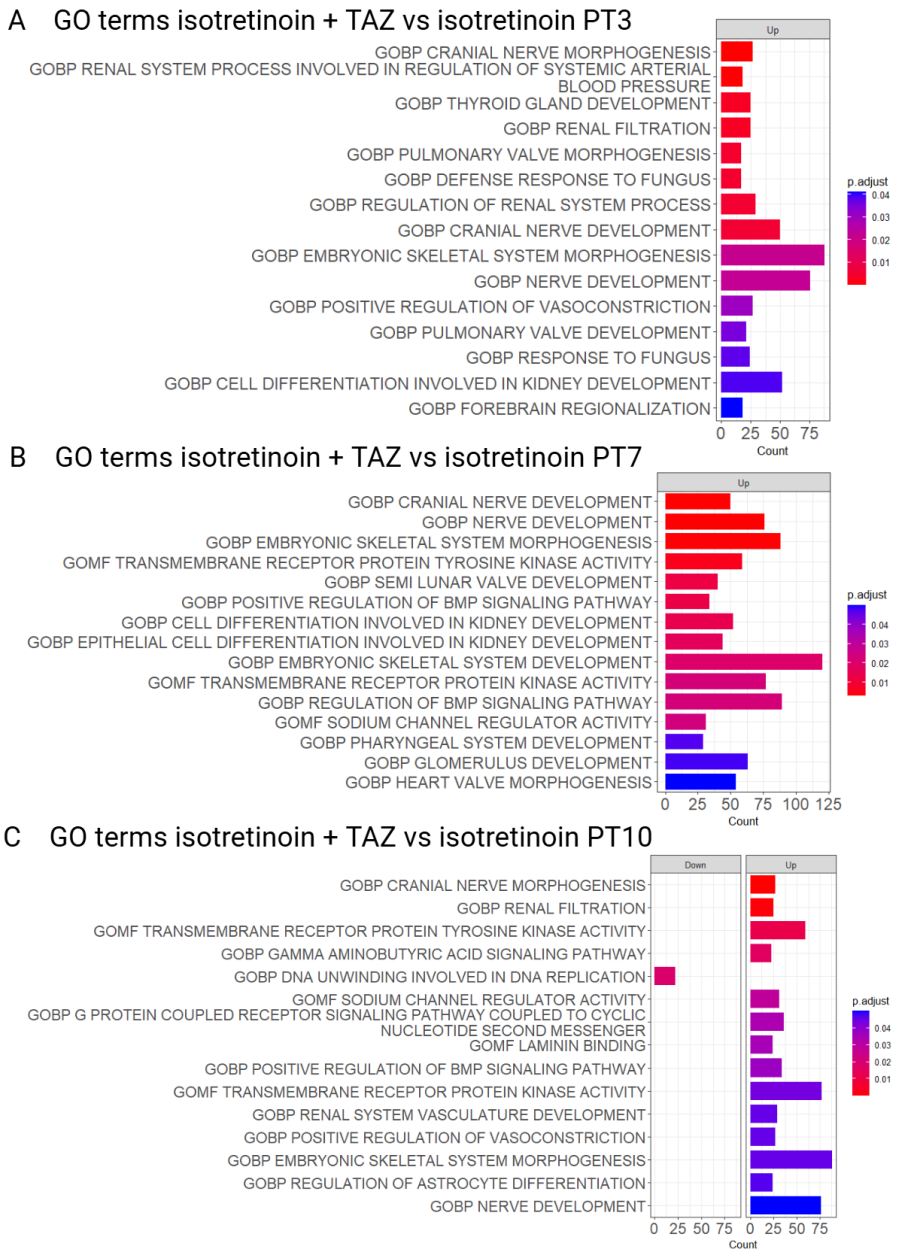


FIGURE 3.25: GSEA results from CAMERA showing enriched GO terms for isotretinoin + TAZ vs isotretinoin for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If more than 15 gene sets were significantly enriched, the top 15 largest gene sets were shown.

combination therapy. Again, isotretinoin + GSK vs GSK showed very similar enriched terms to isotretinoin + TAZ vs TAZ, with enrichment of retinoic acid and differentiation-related terms.

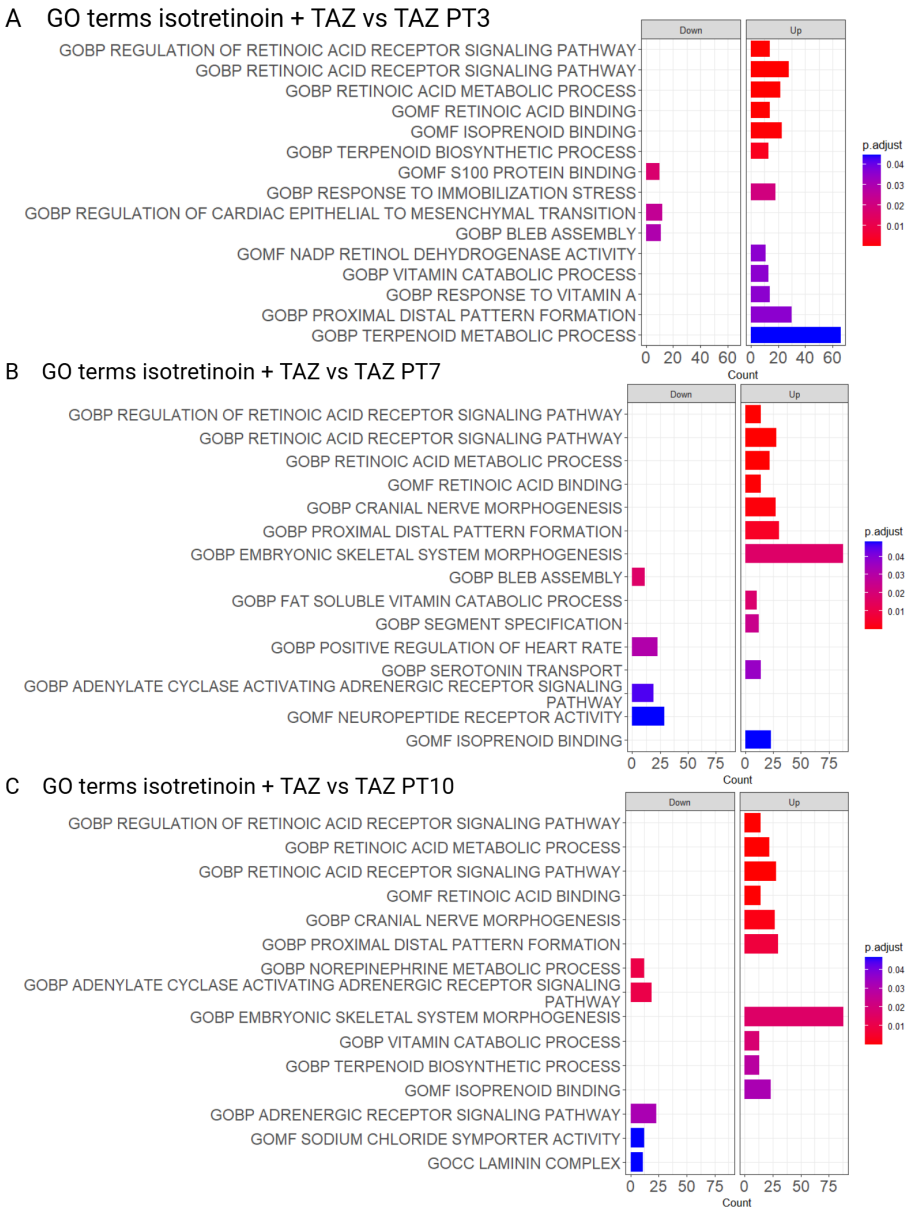


FIGURE 3.26: GSEA results from CAMERA showing enriched GO terms for isotretinoin + TAZ vs TAZ for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If more than 15 gene sets were significantly enriched, the top 15 largest gene sets were shown.

3.3.13 Adrenergic and mesenchymal score

To determine the lineage differentiation stages in each treatment group, GSVA was used to score samples based on the adrenergic and mesenchymal gene lists (Figure 3.27). A Shapiro Wilk test indicated the data was normally distributed (p =0.05137 for adrenergic scores and p =0.3163 for mesenchymal scores), allowing for pairwise comparisons using a T test. Treatment with EZH2is TAZ and GSK resulted in significantly lower adrenergic scores (Figure 3.27A) and higher mesenchymal scores

compared to DMSO (Figure 3.27B), suggesting treatment with EZH2is pushed cells towards the mesenchymal lineage. Isotretinoin appeared to have a slightly higher adrenergic score compared to DMSO, but this was not significant (Figure 3.27A). Treatment with isotretinoin resulted in a significantly lower mesenchymal score compared to DMSO (Figure 3.27B). GSVA scores for combination therapy were inbetween scores for isotretinoin and EZH2is. There was no significant difference between adrenergic or mesenchymal scores and DMSO for combination therapy (Figure 3.27A and B).

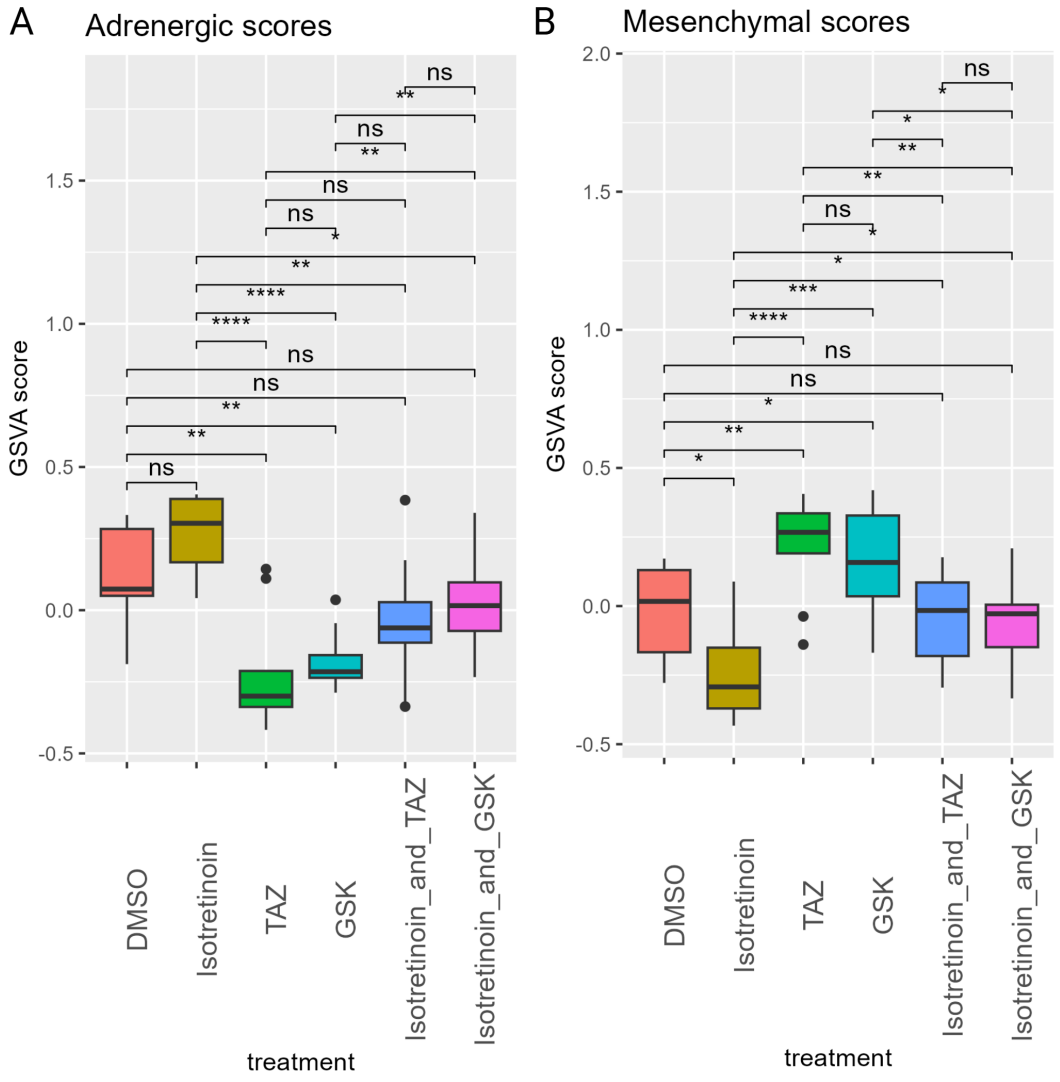


FIGURE 3.27: Lineage GSVA scores for A) adrenergic genes B) mesenchymal genes. Adrenergic and mesenchymal gene lists were from Groningen et al. [42]. PT= post-treatment day. Pairwise comparisons between treatments were made using a T test.

3.4 Discussion

This work aimed to investigate the MoA of isotretinoin, EZH2is and the combination therapy in NB through the analysis of RNA sequencing data. RNA sequencing results suggest that treatment with isotretinoin, EZH2is and combination therapy results in NB differentiation. There was evidence to suggest that combination therapy promoted further differentiation of NB cells compared to single agent treatment. Whilst some underlying mechanisms of combination therapy appear to be the same as single agent treatment, downregulation of DNA replication and cell cycle gene sets were exclusive to combination therapy.

3.4.1 Effects of isotretinoin

Treatment with isotretinoin resulted in around equal numbers of upregulated and downregulated genes. Some of the most significant DEGs were known retinoic acid response genes including *CYP26B1*, *DHRS3*, *CRABP2*, *SMOC1*, *RARB* and *RET* [250, 251, 252, 253, 254]. CYP26B1 is a member of the cytochrome P450 family and is involved in switching off RA signalling in the embryo [250]. It has previously been found to be upregulated by isotretinoin in neuroblastoma cell lines [251, 255]. Similarly, DHRS3 has been found to attenuate RA signalling during embryonic development and has previously been found to be induced in NB cell lines [256, 257]. High expression of DHRS3 has been associated with favourable outcome in NB, where it is thought to interact with the all-trans-retinol pathway and drive differentiation and senescence [253].

Treatment with isotretinoin resulted in enrichment of retinoic acid and differentiation-related gene sets at all timepoints. Differentiation gene sets included 'enteric nervous system development' and 'limb bud formation'. Neural crest cells differentiate to form the enteric nervous system and generate species-specific pattern formation including axial orientation, such as proximal distal pattern formation [258, 259]. Research has shown that formation of limb buds is controlled by retinoic acid, supporting this finding from the GSEA results [260]. The addition of the term involving synaptic transmission at PT10 suggests that cells are becoming more differentiated and neuron-like as time since treatment increases. This was supported by the observation that a NB differentiation gene set was significantly enriched in isotretinoin compared to control samples. Downregulated gene sets included 'adrenoreceptors', 'CD22 mediated BCR regulation' and 'laminins'. Recently a study showed that pharmacological inhibition or knockout of the adrenergic receptor α1B sensitised NB cells to treatment with isotretinoin [261]. This supports the GSEA results that there could be interplay between isotretinoin and adrenoreceptors. Laminins are glycoproteins of the extracellular matrix that can influence differentiation, migration,

and adhesion. Treatment with retinoic acid has previously been shown to modulate laminins in NB cells [262].

3.4.2 Effects of EZH2 inhibitors

The majority of DEGs from EZH2is were upregulated and very few DEGs were downregulated. This is expected as EZH2 is a transcriptional repressor, and its inhibition would result in upregulation of its target genes [263]. The most downregulated gene in response to TAZ at PT10 was *INSM1*. Overexpression of this gene has been associated with poor clinical outcome in NB [264].

Immune gene sets were upregulated PT7 at PT10 in EZH2i samples compared to control. These included terms relating to cytokines, T-cells and interactions between a lymphoid and non-lymphoid cell. There has been research into the role of EZH2 in regulating immune cells, including tumor-infiltrating lymphocytes [265], CD8+ T cells [266, 267], CD4+ T cells [268], B cells [269, 270] and NK cells [271, 272]. However, the potential role EZH2 plays in the immune response in NB is not well studied. A study by Seirer et al. found that EZH2 activity contributes to the T-cell poor cold immune phenotype in *MYCN* amplified NB [273]. They found EZH2 activity shows a negative correlation with T cell content and Th1-type chemokines. This could support the GSEA findings that inhibition of EZH2 upregulated gene sets involved in cytokine signalling and T cells. There is also some evidence that PCR2 silences interferon-induced MHC-I antigen presentation, resulting in evasion of T cell-mediated immunity in MHC-I-deficient tumours, such as NB [274]. Inhibition of EZH2 was able to restore cell surface MHC-I NB cell lines [274]. This study may support the results showing upregulation of MHC-related gene sets. These results suggest that treatment with TAZ could modulate some aspects of the immune response in NB. A summary of potential immuno-modulatory activities of EZH2 in NB based on the GSEA results can be seen in Figure 3.28. However, a study by Mabe et al. found treatment with TAZ had little effect on macrophage polarization or myeloid cell populations in NB mice models, although immunodeficient mice were used [44]. To study the role of EZH2 in the immune response in NB it would be more appropriate to use mice with an intact immune system such as TH-MYCN mouse models.

Integrin gene sets were also upregulated at PT7 and PT10 by TAZ. Integrins play a role in how a cell interactions with its microenvironment and can influence a wide range of cell activities including migration, proliferation, survival, differentiation, cytoskeletal organisation and immune response [275]. Previous studies have shown EZH2 to regulate integrin signalling in other cell types, supporting this finding [276, 277]. Upregulation of integrin in NB has been previously identified as a marker of differentiation as it correlates with neurite extension [278], which could explain the

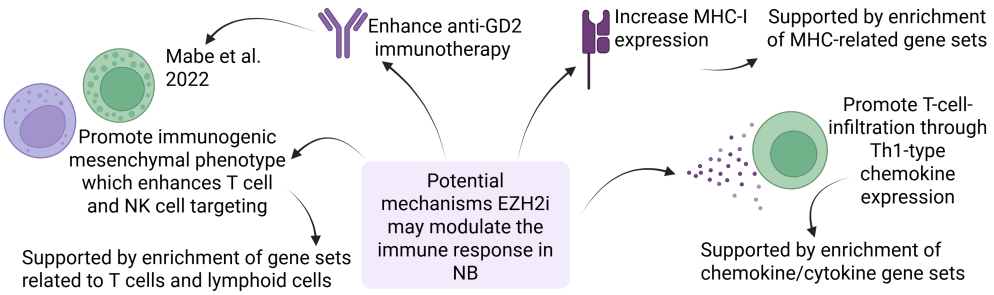


FIGURE 3.28: Summary of potential immuno-modulatory activities of EZH2 in NB, including promoting a immunogenic phenotype, enchancing anti-GD2 immunotherapy, increasing MHC-I expression and promoting T cell infiltration. Created with BioRender.com.

upregulation of integrin gene sets by TAZ. BMP signalling and growth factor gene sets also contained a large number of genes at PT7 and PT10. Growth factors are involved in regulating the neuronal cytoskeleton through cytoskeletal regulators such as BMPs [279]. BMPs can be involved in neurite growth by inducing dendrite formation [280], and this could explain the upregulation of BMP and growth factor gene sets by TAZ which could be a sign of neuronal differentiation. Further supporting this, at PT10 there were large gene sets related to neuronal differentiation. Although GSEA results seemed to suggest cells were becoming more differentiated, the NB differentiation gene set was not significantly enriched by TAZ at any timepoint. However, the NB differentiation gene set created by Frumm et al. was developed by comparing drug induced-differentiated NB cells to undifferentiated cells, with isotretinoin as one of the drugs. They confirmed the signature by testing for enrichment in ATRA and isotretinoin treated samples, which showed enrichment for the signature. The NB differentiation signature likely reflects genes differentially expressed by retinoic acid and thus may not reflect NB differentiation induced by other drugs. This is a limitation of using this signature to verify NB differentiation. These results could be confirmed by checking for markers of neuronal differentiation using a western blot.

3.4.3 Comparison of EZH2 inhibitors

Although the two different EZH2is largely resulted in similar gene expression changes and showed a strong correlation in DEGs, there were some differences. TAZ resulted in more DEGs than GSK, just as Isotretinoin + TAZ resulted in a greater number of significant DEGs than Isotretinoin + GSK. Most DEGs upregulated by TAZ but not GSK did not meet the LFC threshold in GSK. This could suggest that TAZ has a greater or quicker effect on expression changes than GSK. In addition to this, TAZ and isotretinoin+TAZ showed greater variation from the other samples in the PCA plot which could again supporting that the expression changes in GSK may be slower than TAZ. While both of these EZH2is are competitive inhibitors with

S-adenosylmethionine (SAM), which is essential for EZH2 function, they may have differences in their MoA [281]. Another study looking at the effects of EZH2is in colorectal cancer found that TAZ had more of an inhibitory effect than GSK on tumours *In vivo* [282]. This could support the hypothesis that TAZ has a greater effect on gene expression changes than GSK in addition to western blot that shows that treatment with TAZ shows a greater reduction in the H3K27me3 mark than GSK at PT7 and PT10 and likely explains why TAZ results in more DEGs than GSK. There has been some evidence that GSK results in off-target effects in B-non-Hodgkin lymphoma [283].

GSK-treated cells resulted in enrichment of ribosomal and translation related terms at PT3, not seen in TAZ-treated cells. This may also account for large number of DEGs expressed by GSK but not TAZ at PT3. Only one other study could be found where treatment with EZH2is resulted in enrichment in ribosomal gene sets, with the EZH2i GSK [284]. Terms that were enriched in TAZ-treated cells at PT10, such as ECM, were not by GSK. As well as this, gene set sizes were a lot smaller than TAZ. These findings support the hypothesis that TAZ exerts a stronger or quicker effect on gene expression than GSK.

3.4.4 Effects of combination therapy

Combination treatment predominately showed differentiation-related gene sets upregulated, with neuronal differentiation gene sets upregulated at PT10. Some of the gene sets enriched in the upregulated additive genes were also involved in neuronal differentiation. This suggests that combination therapy results in more differentiated cells than single agent treatment.

In combination therapy at PT7 and PT10, the large immune, ECM, growth factor, cytokine and neuronal differentiation gene sets seen upregulated in EZH2i treated samples were not present. This could suggest that treatment with isotretinoin dampens the possible immune-modulatory activity of EZH2is. When looking at the genes from the GO term 'immune response' enriched by TAZ at PT10, clustering shows separation of EZH2i-treated samples at later timepoints, as well as isotretinoin + TAZ at PT10 from other samples. This lead to the hypothesis that the addition of isotretinoin to EZH2is may be reducing the expression of immune genes, since combination therapy shows lower expression of these genes. Except for isotretinoin + TAZ at PT10, which clusters with the higher expressing samples. Further validation at the protein level is needed to confirm functional relevance. There is little research into the effect of retinoic acid on the immune response in NB. One study found that treatment with ATRA resulted in stabilisation of GD2 expression and enhanced ADCC [285]. Retinoic acid has also been found to induce IL-8, IL-18 and ICAM-1 responses in NB [286, 287, 288]. Of these, *ICAM1* was the only gene that was significantly

upregulated by isotretinoin in the RNA-seq data. In addition to this, a study by Vertuani et al. found that retinoic acid derivatives may play a role in the MHC-I presentation pathway and can sensitise cells to cytotoxic lymphocytes in NB [289]. More research is needed into the role of retinoic acid and the immune response in NB.

Another noticeable difference in combination therapy compared to single agent treatments was large DNA replication/cell cycle gene sets downregulated in isotretinoin + TAZ. The Hallmark 'G2M checkpoint' (197 genes) was downregulated in isotretinoin + TAZ at PT10. This cell cycle checkpoint is involved in preventing cells with DNA damage from entering mitosis [290]. Depletion and inhibition of EZH2 has previously been found to regulate the DNA damage response, abrogating the G2M checkpoint and drive cells towards apoptosis [291]. In addition to this, inhibition of EZH2 has been found to sensitise multiple myeloma cells to DNA-damaging agents and promoted apoptosis [292]. Another study in non-small cell lung cancer also found that EZH2 inhibition induced G2M arrest [293]. Furthermore, treatment with retinoic acid has been implicated in abrogating the G2M checkpoint [294, 295, 296]. Perhaps both single agent treatments moderate the DNA damage response in NB cells, but in combination this effect is amplified. Indeed, when looking at the enriched terms for downregulated additive genes, they are predominately involved in DNA replication and cell cycle, including 'mitotic prometaphase', 'M phase' and 'mitotic metaphase and anaphase'. Both isotretinoin and TAZ appear to downregulate mitotic pathways, with an additive effect observed in combination therapy. Whether this is due to G2M arrest and DNA damage is unclear, although the term 'damaged DNA binding' was enriched in these additive genes and Reactome term 'activation of ATR in response to replication stress' was enriched in isotretinoin + TAZ treated samples, which suggest a possible mechanism. Alternatively, these results could be a marker of decreased proliferation. Whilst GSEA can be used to aid interpretation of results, these results are based on annotations and should not be taken definitively as causing the enriched term. It would be useful to further look into the expression of the genes in this enriched set and markers associated with these gene sets. To further investigate this, work at the protein level could be conducted such as a western blot of markers of DNA damage in TAZ-treated samples. It may also be useful to look at the cell cycle stage of isotretinoin + TAZ treated cells, to check for any evidence of G2M arrest. An info graphic summarising the main themes identified in the GSEA results can be seen (Figure 3.29).

3.4.5 Effect of treatment on cell phenotype

NB cells have been shown to be plastic and able to shift between the adrenergic and mesenchymal identities due to external cues of the environment [297]. The lineage differentiation scores suggest that EZH2i treatment results in more mesenchymal

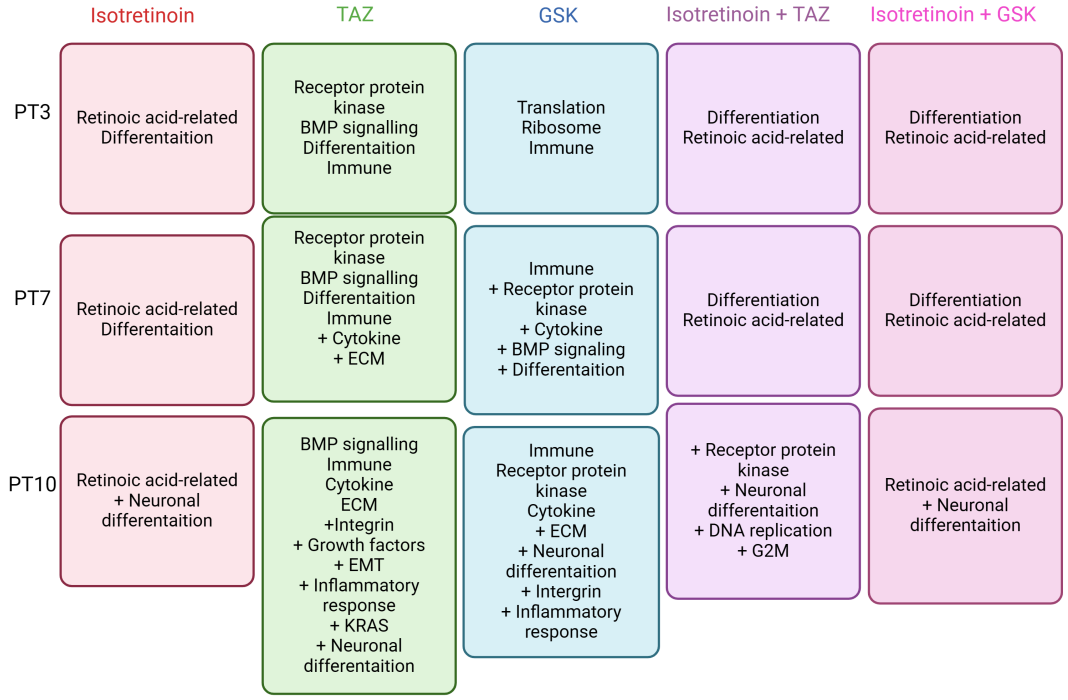


FIGURE 3.29: Summary of the main themes identified from GSEA for isotretinoin, TAZ, GSK, isotretinoin + TAZ and isotretinoin + GSK treated NB samples. + indicates new terms that were not present in the previous timepoint. Created with BioRender.com.

scores compared to DMSO. Mesenchymal NB cells have been shown to be more immunogenic and can promote T cell infiltration by secreting cytokines and be targeted by cytotoxic T cells and NK cells [45]. They also found there was upregulation of MHC-1 and antigen processing and presentation. This could support the GSEA results that suggest EZH2is may be affecting the immune response, including T cell gene sets. In addition to this, the gene *PRRX1* was found to be able to convert NB cells from an adrenergic to mesenchymal state, was also significantly upregulated by TAZ. However, these results contradict the results from Mabe et al., who found that treatment with EZH2is reprogrammed cells to a more adrenergic state [44]. Mabe et al. used the non-*MYCN*-amplified cell line SK-N-AS, whilst this study used the *MYCN*-amplified cell line Kelly, which could explain these discrepancies. The cell line Kelly also harbors a *p53* mutation whilst SK-N-AS is *p53* wild-type, which may also explain these contradictory findings. As well as this, GSEA terms upregulated with EZH2is seemed to suggest differentiation, which contradicts the mesenchymal undifferentiated phenotype. Treatment with isotretinoin resulted in lineage differentiation scores less mesenchymal than control. Treatment with ATRA has been shown to induce differentiation in adrenergic NB cell lines (such as Kelly) but not mesenchymal cell lines [45, 298].

3.4.6 Limitations

Some limitations of this work include a potential batch effect in replicate 3. This was included in the design for differential expression analysis to attempt to minimise the batch effect, although the source of the batch effect has not been determined. Another limitation of this analysis was treating biological replicates as technical replicates. This may explain why the BCV value was higher than expected, as the biological replicates would show more variation than technical replicates. A higher BCV will reduce statistical power for DGE and can affect MDS plots. In addition to this, the data is only from one cell line and it would have been beneficial to have biological replicates from other high-risk NB cell lines, including adrenergic cell lines KP-N-S19s and SK-N-BE-2.

In summary, this work investigated possible Mechanism of Action (MoA) of EZH2is, isotretinoin and combination therapy in NB. Treatment with isotretinoin resulted in a more adrenergic phenotype. Enriched GSEA results were mainly differentiation-related, with results suggesting cells become more differentiated and neuronal-like as time increases. EZH2is resulted in similar DEGs, with some slight differences between the two EZH2is. Treatment with the EZH2i TAZ showed enrichment of gene sets involved in differentiation, the immune response and ECM. The differentiation status of EZH2i-treated cells were contradictory, with GSEA suggesting neuronal differentiation whilst adrenergic/mesenchymal scores suggested a shift towards a more mesenchymal phenotype and the enrichment of immune-related gene sets also suggest a more immunogenic mesenchymal phenotype. Further work is required to determine the differentiation phenotype in EZH2i-treated cells. Treatment with isotretinoin + TAZ again suggested differentiation. Immune-related gene sets seen in treatment with EZH2is were not enriched in combination therapy, with isotretinoin possibly dampening the immune-modulatory effect of EZH2is. K-means clustering identified a subset of DEGs that were more highly expressed in EZH2i-treated samples than combination therapy, and enrichment analysis revealed these genes were enriched for inflammatory response and chemokines. There was enrichment of DNA replication and cell cycle related gene sets in downregulated additive genes, which could be a possible MoA that is enhanced in combination therapy. K-means clustering identified a subset of genes with higher expression in combination therapy than single agent treatment, which showed enrichment of neuronal differentiation and ECM. Additive upregulated genes appeared to be involved in neuronal differentiation. These results could suggest combination therapy is also enhancing differentiation.

Chapter 4

Unravelling mechanisms of therapy resistance in rhabdomyosarcoma

4.1 Introduction

RMS is a highly heterogeneous disease in terms of genetic background, presentation, histology and prognosis [299]. For patients with low-risk disease outcome is good with a 3-year OS around 98% [300]. However, metastatic and recurrent RMS have a dismal prognosis [90, 14, 107]. Although the majority of tumours initially respond to chemotherapy [105], relapse occurs in around two-thirds of patients with metastatic disease [107] and post-relapse survival is dismal.

To better understand and predict treatment outcomes in RMS, researchers have explored the use of gene expression signatures. These signatures are defined as a list of differentially expressed genes associated with a biological phenomenon and are typically created using transcriptomic data such as RNA-seq or microarray [301]. They can be predictive of prognosis, disease type/subtype, treatment response and resistance. Gene signatures have previously been developed for predicting subtype and prognosis in RMS. An earlier study by Wachtel et al. identified gene signatures using microarray data from patient samples that could distinguish ERMS, FP-ARMS and FN-ARMS samples [302]. The ARMS signature was used on an additional ARMS case, identifying it as FP leading to the identification of a novel translocation that existed between PAX3 and NCOA1 [302]. Another gene signature, developed by Davicioni et al., consisted of 34 genes and was found to be prognostic of COG risk group [303]. Although in an independent dataset the signature did not predict patient outcome when PAX-FOXO1 fusion gene status was considered [72]. However, the authors developed another gene signature using this dataset which consisted of 5 genes. The 5 gene signature was significantly associated with survival in the FN subtype, but not FP [72] and was validated in an independent cohort [304]. Despite

the existence of these prognostic signatures, there are none that predict chemotherapy resistance in RMS. This has been done in many other cancer types, such as gastric [305], breast cancer [306], NB [183]. Some studies derive these gene signature from patient microarray data whilst others use data from cell models of resistance. For example, a study by Jemaà et al. used VCR-resistant NB cell lines to look for expression changes related to chemotherapy resistance [183]. They identified 24 genes that showed significant differential expression in the VCR-resistance cells to be used as a chemoresistance signature. In patient microarray data, samples that were chemoresistant according to the signature had features associated with worse clinical outcome including *MYCN* amplification and worse overall survival.

Gene signatures can be created from RNA-seq data in multiple ways and several bioinformatic frameworks have been developed. In other cancer types, signatures have been derived from DEGs between resistant and non-resistant samples [307, 308, 309]. Although this approach is simpler, it relies on a binary cut-off and may overlook genes with smaller but consistent changes in expression. Mallik and Zhao use k-means clustering on DEGs and identify the best cluster by computing the average Spearman's Correlation for each gene in the cluster in a pairwise manner [181]. The best cluster is then treated as the signature. However, as a recently developed technique, it has not yet been widely adopted with only one citation to-date [310]. In addition to this, it has not been validated in an independent cohort, with validation instead using a statistical tool to classify the performance of the signature [181]. WGCNA is a tool used to identify modules of co-expressed genes, which are groups of genes with similar expression patterns across samples. These modules are often functionally related as genes with coordinated expression are more likely to participate in shared biological processes or pathways [311]. WGCNA is widely used to derive gene expression signatures, by identifying modules of co-expressed genes and then correlating module eigengenes with trait information, such as survival or response to treatment. The resulting modules associated with the trait can then be filtered using information from PPIs [312, 313, 314], cross-referencing with DEGs, or using Least Absolute Shrinkage and Selection Operator (LASSO) with multi-cox regression [315, 316] to reduce the gene signature to key genes associated with the trait of interest. WGCNA has been used to derive gene signatures of chemotherapy resistance in other cancers. For example, Zhang et al., applied WGCNA to transcriptomic data from ovarian cancer patients divided into chemoresistant and chemosensitive groups. Through this analysis they identified gene co-expression modules associated with chemotherapy resistance [317]. Several genes in the resistant modules were found to be upregulated in chemoresistant samples and were verified as prognostic markers in an independent dataset.

RMS tumours have a relatively quiet genome, with few genomic alterations aside from the fusion gene in FP-RMS or *RAS* pathway mutations in FN-RMS [110]. There is

evidence that heterogeneity at the transcriptional level may play a role in resistance to chemotherapy in RMS [156, 318]. Due to the lack of sample availability, cell models are an important resource in RMS research. Multiple studies have shown evidence of heterogeneity within RMS cell lines. Tonelli et al. reported the RH30 cell line had heterogenous *MYCN* amplification [187]. Another study compared RH30 cell lines from two different sources [319]. They found that although the cell lines from different sources matched in STR profile, one subclone was PAX7 positive and the other PAX7 negative. The subclones also differed in morphology, proliferation rate, migration activity and chemotactic abilities [319].

The ERMS cell line RD has been shown to have heterogenous expression of MYOD1 and NOG, with proteins only being expressed in a subset of RD cells [320]. Recently, single-cell RNA-seq studies have shown RMS cell lines contain different human skeletal muscle development cell states, for example the RH4 cell line contains both myogenic progenitors and myoblast-myocytes [318]. The heterogeneity present in RMS cell lines provides some utility in modelling chemotherapy resistance.

This work aims to identify a gene signature of chemotherapy resistance using RNA-seq data generated from models that represent intrinsic and acquired RMS resistance. The intrinsic resistance model used single-cell clones derived from RH30 cells that were screened for VCR resistance compared to parent cells. The clones that showed the highest resistance compared to the parental cells were selected and RNA-seq was performed on the treatment-naive clones. To generate models of acquired resistance, cell lines RMSYM and RH4 were treated with VCR or IFO to generate resistant cells. The intrinsic and acquired resistance data was analysed for DGE analysis, GSEA and WGCNA to create a network of co-expressed genes. Modules correlated with resistance were identified and a PPI network was generated from these modules. A gene signature was derived from the hub PPI network proteins which may have potential clinical utility in predicting chemotherapy resistance and relapse in RMS patients. These signatures will be validated in Chapter 5 using publicly available RMS microarray data.

4.2 Methods

4.2.1 Generation and sequencing of intrinsic resistance models

The laboratory work for the intrinsic resistance model was carried out by Dr Eleanor O'Brien and Dr Ian Tracy. Two methods were employed to select VCR resistant RH30 clones; single cell sorting, and drug selection.

4.2.1.1 Single cell clone selection

Single cells from the heterogeneous RMS cell line RH30, were sorted into 96-well plates using Fluorescence-activated cell sorting (FACS). RH30 is an alveolar RMS cell line with the *FOXO1-PAX3* fusion gene as well as mutations in *RARA* and *TP53*. These single cell clones were allowed to establish in culture over 28 days at which point viable, proliferating clones were selected for further testing (Figure 4.1). Given that VCR induces apoptosis via cell cycle disruption, the relative proliferation rates of the single-cell clones was determined over a 96-hour period. Per clone, one thousand cells were seeded in to each well of a 96-well tissue culture plates and cultured for 96-hours. CyQuantTM fluorescence assay was used to calculate cell numbers at 0, 24, 48, 72, and 96 hours and the growth curves plotted using Graphpad prism software. The amount of time required for three cell doublings was calculated for each clone and recorded. Drug sensitivity of the parent RH30 cell line was determined by calculating the dose of vincristine required to cause 50% of the maximal growth rate (GI50). Briefly, one thousand cells were seeded per well in to 96-well tissue culture plates and allowed to adhere for 6 hours. Media was then replaced with fresh media containing VCR diluted across a 7 step 2-fold concentration series (2.5nM – 39pM), plus a vehicle control and positive control well (1mM VCR). 6 technical replicates were performed for each condition using 6 separate wells from the 96 well plate. The parent cell line was incubated with drug until their calculated time for 3 cell doublings at which point media was removed and CyQuantTM assay used to determine relative cell count. The GI50 was calculated using Graphpad Prism statistical analysis software to be 187pM (95% CI = 153.3-218.8, R2 = 0.9523).

Next, each clone was cultured with 156pM VCR for three cell doublings, alongside an untreated control. CyQuantTM assay was used to determine the relative cell counts (treated versus control) for each clone and the parent cell line. The parent cell line was sensitive to this dose, as expected. Two clones, C7 and C13, were selected for RNA-seq as both clones showed no significant difference in growth between the treated and control conditions.

4.2.1.2 High dose sample generation

Another method to select resistant cells from the parent RH30 cell line involved reanalysing the dose response data of the parent RH30 cell line to determine the vincristine concentration required to inhibit growth in 90% of cells (GI90). This was calculated to be 625pM. The parental RH30 cell line was therefore cultured in complete media containing vincristine at the GI90 dose for two days, after which the cells were cultured in complete media containing the GI50 dose of vincristine (156pM). This was maintained for 30 days during which time the cells continued to grow and required

passaging every 4-5 days. A dose response was performed after 30 days and the GI50 was found to have increased to 834.5pM (95% CI = 755.9-940.4, R2=0.9548) in the high dose treated RH30 cells, in comparison to the parent GI50 of 214pM (95% CI=167.2-260.7, R2=0.8607). This sample, herein referred to as high dose (HD) was then prepared for sequencing along with RH30 parental cells (P) and single cell derived clones, C7 and C13. A summary of the study design is presented in Table 4.1.

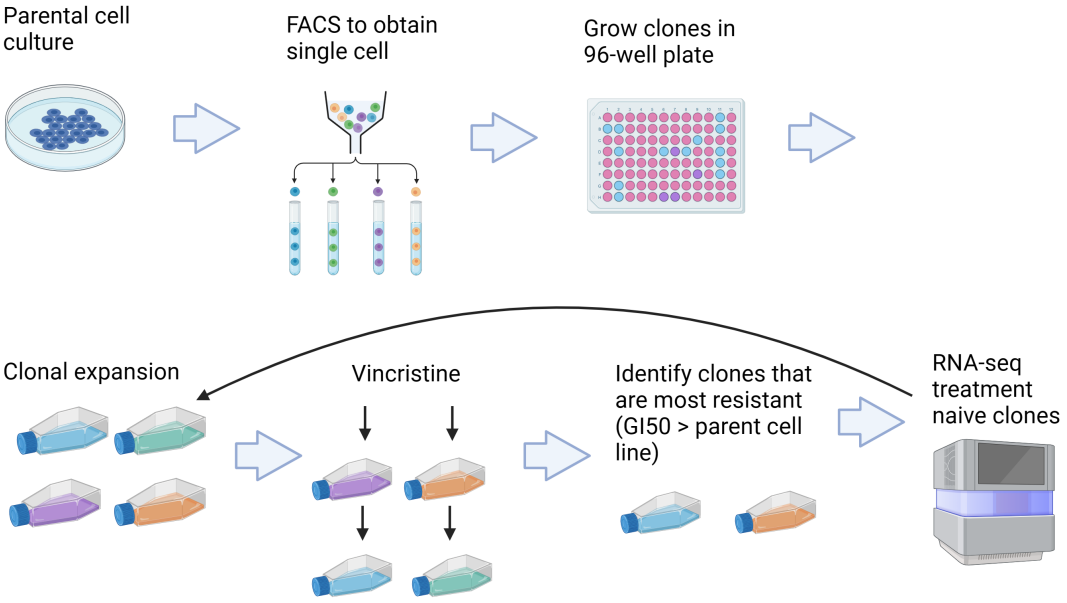


FIGURE 4.1: Generation of single cell clones by FACs of RH30 RMS cell line. Clones were treated with VCR and the most resistant clones selected. RNA-sequencing was performed on the treatment-naïve clones. Created with BioRender.com.

4.2.1.3 RNA sequencing protocol

According to Novogene methods, RNA degradation and contamination was monitored on 1% agarose gels. RNA purity was checked using the NanoPhotometer spectrophotometer (IMPLEN, CA, USA). RNA integrity and quantitation were assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). A total amount of 1µg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were prepared using the NEBNext Ultra™ RNA Library Prep Kit for Illumina, following the manufacturer's protocol. mRNA was enriched using poly-T oligo-attached magnetic beads, fragmented, and reverse-transcribed into cDNA. Libraries were size-selected (∼300 bp), indexed, and quality-checked using the Agilent Bioanalyzer 2100 system. Paired-end sequencing (150 bp) was performed on an Illumina platform, achieving a Q30 score $\geq$85%.

TABLE 4.1: Study design for the intrinsic resistance RMS models. The parent cell line, resistant C7, resistant C13 and HD were sent for RNA sequencing. There were 6 replicates for each treatment condition.

Cell line	Treatment condition	Intrinsic resistance model
RH30	Parent cell line (P)	n=6
	Clone 7 (C7)	n=6
	Clone 13 (C13)	n=6
	High dose (HD)	n=6

4.2.2 Generation and sequencing of acquired resistance model

4.2.2.1 Acquired resistance model generation

Laboratory work for the acquired model of resistance was conducted at the Institute of Cancer Research by Dr Lucile Delespaul. Two RMS cell lines, RH4 and RMSYM, were subjected to treatment with chemotherapy agents VCR or IFO. RH4 is an ARMS cell line with the *FOXO1-PAX3* fusion gene and *TP53* mutation [321]. The RMSYM cell line is ERMS with a *BCOR* mutation. Cells were treated with the drug for 72 hours before a medium change. The dose of drug was increased each time, until Cmax was reached or cells were no longer tolerant to the increased concentration (Figure 4.2). The acquisition of resistance ability was observed using cell viability assay by comparing the control and resistant cell lines. IC50 values for RH4 VCR were 0.003 for control and 0.088 for resistant. IC50 values for RH4 IFO were 2.1 for control and 6.3 for resistant. For RMSYM VCR, IC50 values were 0.06 for control and 3.3 for resistant. For RMSYM IFO, 1C50 values were 3.1 for control and 5.3 for resistant. For controlling the growth rate effect on cell viability after drug treatment, growth rates were adjusted for the cell doubling time. There were six replicates for each treatment. The established resistant cells were sent to Novogene for RNA sequencing. A summary of the study design can be seen in Table 4.2.

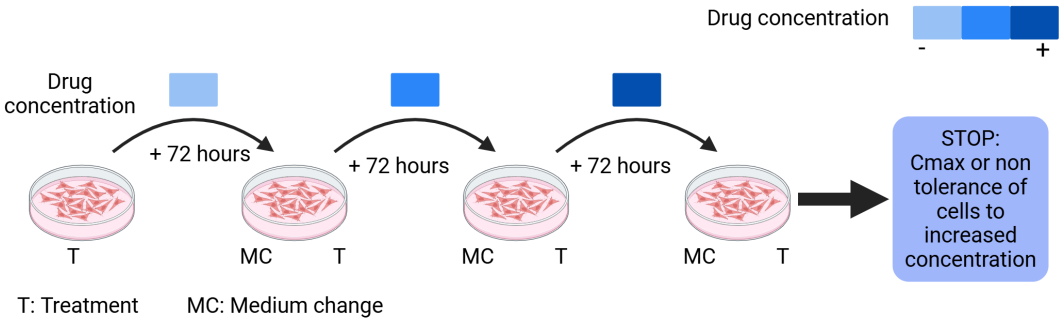


FIGURE 4.2: Generation of acquired RMS resistance models to VCR and IFO. Cells were treated with increasing concentrations of VCR or IFO, with a medium change and break between treatments, until Cmax was reached or cells were no longer tolerant to the concentration. Created with BioRender.com.

TABLE 4.2: Study design for the acquired resistance RMS models. There was a total of 6 treatment conditions; RH4 control, RH4 VCR-resistant, RH4 IFO-resistant, RMSYM control, RMSYM VCR-resistant and RMSYM IFO-resistant. There were 6 replicates for each treatment condition.

Cell line	Treatment condition	Acquired resistance model
RH4	Control	n=6
	VCR-resistant	n=6
	IFO-resistant	n=6
RMSYM	Control	n=6
	VCR-resistant	n=6
	IFO-resistant	n=6

4.2.2.2 RNA sequencing protocol

Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesized using random hexamer primers, followed by the second strand cDNA synthesis using dTTP. The library was ready after end repair, A-tailing, adapter ligation, size selection, amplification, and purification. The library was checked with Qubit and real-time PCR for quantification and bioanalyzer for size distribution. Quantified libraries were sequenced on Illumina platforms for 150 paired-end, unstranded sequencing with Q30 $\geq$85%.

4.2.3 Differential gene expression analysis

The analysis of the RNA sequencing data is described in Chapter 2. All code is available in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Code).

First, lowly expressed genes were removed as they often generate false positive results. For the intrinsic resistance dataset a CPM of 0.35 was used as the threshold, corresponding to a count of around 10-15 as identified through visual inspection (Supplementary Figure 15). Genes were kept if they had a CPM greater than 0.35 in at least 6 samples, as this was the lowest number of replicates for each condition and thus retained genes that were expressed in at least one condition. Read counts were analysed for DGE using edgeR glmQLFit [198] to determine DEGs between the resistant clones (C7, C13 and HD) compared to the Parental cell line (P). Significant DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg). This LFC threshold was selected to capture moderate change in expression whilst avoiding the noise inherent in RNA-seq data. As well as this, the LFC of DEGs in the clones (C7 and C13) were smaller and using a higher threshold would result an insufficient number of DEGs for downstream analyses. In addition to gene sets from GO, Reactome and Hallmarks, a cell proliferation gene signature developed by Locard-Paulet et al. [322] was also tested for enrichment using

CAMERA. This proliferation gene signature can be found in Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Proliferation_gene_signature.xlsx).

For the acquired resistance dataset, a CPM of 0.25 was used as the threshold, corresponding to a count of around 10-15. Genes were kept if they had a CPM greater than 0.25 in at least 6 samples, as this was the lowest number of replicates for each condition. Read counts were analysed for DGE using edgeR glmQLFit [198] to determine DEGs between RH4 IFO-resistant vs RH4 control, RH4 VCR-resistant vs RH4 control, RMSYM IFO-resistant vs RMSYM control and RMSYM VCR-resistant vs RMSYM control. Significant DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

4.2.4 Weighted gene co-expression network analysis

The R package WGCNA was used for this analysis [311]. Counts were normalised to log-2 CPM, checked for missing values and input as a matrix of gene expression values. A sample dendogram was generated using hierarchical clustering to identify and exclude potential outlier samples based on expression similarity. WGCNA allows the incorporation of phenotypic data with the expression data, which can be variables such as survival, tumour location, drug response, and age. The phenotypic data used for this analysis was binary trait data, with the trait being 'resistant' or 'non-resistant', similar to a previous study by Zhang et al. [317].

To first create the network, all genes are connected to one another, and a correlation threshold is determined to remove an edge if the genes are under the correlation threshold. Instead of a fixed threshold being used, a soft threshold power is used to enhance the differences between strong and weak correlations. To determine the correction factor, a plot of soft threshold powers are plotted. WGCNA recommends picking a soft threshold power with a signed R^2 value of at least 0.8 to reduce noise. The type of network is also defined at this point, as either unsigned or signed. An unsigned network considers connections regardless of direction of the correlation where as a signed network only considers connections that show the same direction of correlation. The type of network was 'signed', as recommended by WGCNA to preserve the directionality of gene co-expression, where the network adjacency scales the correlation to lie between 0 and 1, then raises it to the soft threshold power. The method used to calculate correlations was Kendall. A Topological Overlapping Matrix, which considers the number of neighbours the genes share, is created from the adjacency matrix. The corresponding dissimilarity can be calculated from the overlap (which is a measure of similarity) by subtracting it from 1. This information is stored in the dissimilarity matrix. The minimum module size was set to 40 to remove modules with a small number of genes that may represent noise. Genes with more

similar pattern of expression are clustered and grouped together into modules of highly connected genes. Similar modules were merged together using the dissimilarity of module eigengenes (the first principal component of a module that is representative of the gene expression profiles). The dissimilarity of module eigengenes threshold was set to 0.25.

Modules were then correlated with phenotypic data (whether samples were resistant or non-resistant), using the mean of the correlation value between each gene of the module and the phenotypic measure. The 'cor' function from WGCNA package was used to calculate the weighted Pearson correlation between genes in the module and phenotypic information. The function 'corPvalueStudent' was used to calculate the Student asymptotic p-value for the given correlations. Modules were identified as significantly correlated with resistance if they had a Student asymptotic p-value for correlation <0.05 and were selected for further analysis.

Module membership is a measure of the Pearson correlation between a gene expression profile with the module eigengene, and can be used to infer how well a gene fits a module. For example, a low module membership score suggests the gene does not fit the module well as its expression is not similar to the eigengene of that module. Gene significance is a measure of association between a gene and the trait. A high gene significance score shows that gene expression is correlated strongly with the trait. Gene significance and module membership scores can be plotted to investigate the relationship between the two. A threshold of gene significance and module membership was used to select for hub genes (important genes) that show a strong correlation with the trait and the module. For modules positively correlated with resistance, hub genes were defined as having a module membership and gene significance score >0.6. For modules negatively correlated with resistance, hub genes were defined as having a module membership >0.6 and gene significance $<$-0.6. Hub genes were than analysed for gene set enrichment analysis for Reactome, Hallmarks, KEGG and GO terms using the R package clusterProfiler [238]. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg).

For the intrinsic resistance dataset, a total of 18 samples were included for the WGCNA from the parental (P) and resistant (C7 and C13) groups. HD was not included for this analysis due to the different method of generation (described in detail in section 4.3.1.5). The raw data was reprocessed just including these samples and CPM thresholds determined. A CPM threshold of 0.25 was used, corresponding to a count of around 10-15. Genes were kept if they had a CPM greater than 0.25 in at least 6 samples, as this was the lowest number of replicates for each condition. A soft threshold power of 10 was selected.

For the acquired resistance data, two co-expression networks were generated; one using FP (FOXO1-PAX3) ARMS samples and one using FN ERMS samples. The

co-expression network of FP samples had a total of 18 samples from the RH4 cell line (6 control, 6 VCR-resistant and 6 IFO-resistant). The data was refiltered to remove genes that were expressed in RMSYM but not RH4. Genes were kept if they had a CPM greater than 0.25 in at least 6 samples, as this was the lowest number of replicates for each condition. A soft threshold power of 12 was selected. The co-expression network generated using FN ERMS samples from the RMSYM cell line included 6 control samples, 6 VCR-resistant samples and 6 IFO-resistant samples giving a total of 18 samples using in the analysis. The data had to be refiltered to remove genes that were expressed in RH4 but not RMSYM. Genes were kept if they had a CPM greater than 0.25 in at least 6 samples. A soft threshold power of 18 was used.

4.2.5 Selection of a gene signature with weighted gene co-expression network analysis and protein-protein interaction networks

A popular approach to derive a gene signature from WGCNA results is to use a PPI network [314, 313, 323]. PPI networks can be used to identify protein-protein interactions and highly interacting proteins can be identified, called hubs. This method uses WGCNA to identify groups of co-expressed genes related to the trait (that may have a common function related to this trait) and then selects the important members of this co-expression network by choosing proteins that have the highest interactions. A PPI network is constructed and filtered to select for interactions between genes of interest (e.g. DEGs , hub genes, module genes). Hub genes from the PPI network are selected and cross-referenced with hub genes from WGCNA. First a PPI network was constructed using information from the STRING database [324]. This database provides information on protein interactions from high-throughput experiments, co-expression data, literature data and genomic context. Each interaction is assigned a confidence score that reflects how likely STRING judges an interaction to be true based on the evidence. Confidence scores are between 0 and 999, with 999 being the highest confidence. The minimum required interaction score puts a threshold on the confidence score, so only interactions above this score are included in the network. The default threshold of 400 was used, which is classed as 'medium confidence'.

For the intrinsic resistance dataset, the PPI network was filtered to select for interactions between all module genes for each resistant module (modules with a significant positive correlation with the trait resistant and with positive correlation between gene significance and module membership sscores). This created two PPI networks (one network built from all 'darkgreen' module genes and one network for all 'darkgrey' module genes). After the PPI network was created, PPI hub proteins were selected based on the centrality score (the number of links to a given node). There is no pre-defined centrality threshold as this metric can vary significantly for

each network and instead it is suggested to select a centrality threshold based on the individual data. A centrality threshold of 8$\geq$ was used to select for PPI hub proteins. The PPI hub proteins from each module were merged and cross-referenced with the WGCNA hub genes to be used for the intrinsic chemo-resistance gene signature.

The same approach was used for the FP-acquired resistance dataset, the PPI network was filtered to select for interactions between all module genes for each resistant module, creating two PPI networks (one network built from all 'orange' module genes and one network for all 'lightgreen' module genes). A centrality threshold of 8$\geq$ was used to select for PPI hub proteins. The PPI hub proteins from each module were merged and cross-referenced with the WGCNA hub genes to be used for the FP acquired chemo-resistance gene signature.

For the FN-acquired resistance dataset, the PPI network was filtered to select for interactions between all module genes for the resistant module 'darkolivegreen' and for the 'green' resistant module. A centrality threshold of 8$\geq$ was used to select for PPI hub proteins. The PPI hub proteins from each module were merged and cross-referenced with the WGCNA hub genes to be used for the FN acquired chemo-resistance gene signature.

4.2.6 Quantification of the gene signatures

The gene signatures were then quantified using the R package GSVA with default parameters to assign a score to samples based on the expression of the genes in the signature [239]. This method is used in the literature to quantify scores for gene signatures or custom gene sets between samples [325, 326, 327]. To test if the GSVA scores were significantly different between resistant and non-resistant conditions, the Shapiro-Wilk was used to test for normality, and T test (parametric) or Pairwise Wilcox Rank Sum test (non-parametric) was used to calculate pairwise comparisons between group levels with corrections for multiple testing. P values were adjusted using the Benjamini-Hochberg method. GSVA gene signature scores were significantly different if adjusted p value $<$0.05.

4.3 Results

4.3.1 Analysis of the intrinsic resistance data

4.3.1.1 Quality Control of the data

All samples passed the quality control checks for sequence quality, GC contamination, N content, sequence length distribution, overrepresented sequences and adapter

content (Figure 4.3). 'Sequence Duplication Levels' and 'Per Base Sequence Content' did not meet standard thresholds which is expected for RNA sequencing data due to transcript abundance and library complexity. The full MULTIQC report can be found in Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Intrinsic resistance/intrinsic_resistance_multiqc_report.html).

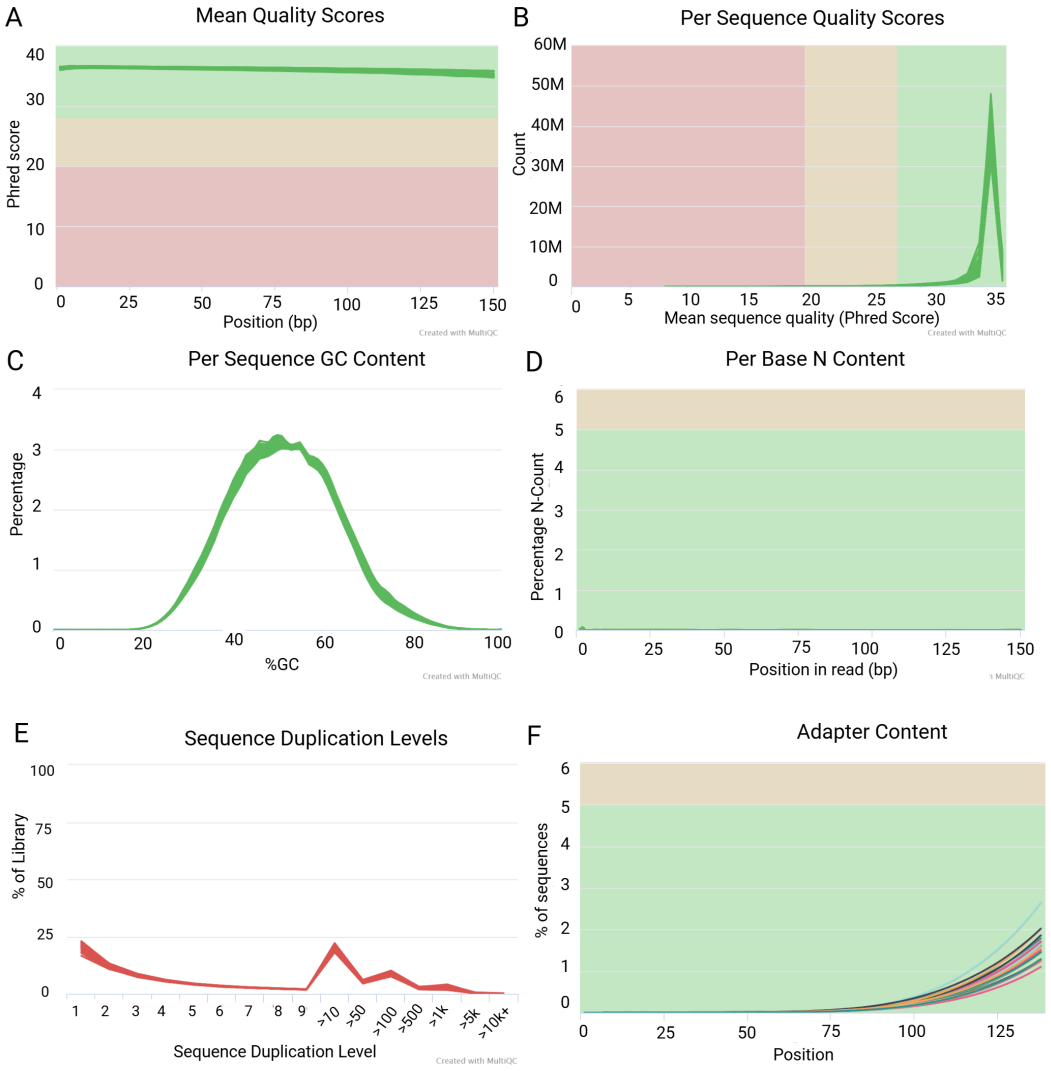


FIGURE 4.3: MULTIQC report for the intrinsic resistance data. A) Mean quality scores showing the phred score for each bp in the read B) Per sequence quality scores C) Per sequence GC content D) Per base N content, where N represents where the sequencer is unable to make a base call with sufficient confidence E) Sequence duplication levels F) Adapter content.

All samples had a high percentage of uniquely mapped reads >90%, showing that the majority of reads mapped to a single location in the reference genome (Supplementary Table 7). This agrees with expected proportions as discussed in Chapter 2.

After filtering to remove lowly expressed genes, 14,276 genes remained. Filtering reduced the number of genes with a count of 0, as shown in the density plots (Supplementary Figure 16), thereby improving data quality for downstream analysis.

Filtered data was assessed for quality before and after normalisation. The distribution of library sizes before and after TMM normalisation is similar between samples and there are no outlying samples (Supplementary Figure 17A and B). The median expression level of each sample is similar to the overall median (blue line) indicating consistent normalisation across samples. Library sizes vary slightly between samples and there are no obvious outliers (Supplementary Figure 17C). The BCV for the intrinsic data was 0.04 (Supplementary Figure 18), indicating that expression levels vary by 4% between replicates, which is slightly lower than expected value of 0.1 for cell lines.

The elbow method and Horn's method were used to determine the optimum number of PCs to retain to investigate additional PCs. Of these two values, the larger number of PCs to retain was chosen (4) (Supplementary Figure 19).

PCA biplots showed that the replicates of each treatment condition cluster tightly together, suggesting there is very little variation between replicates (Figure 4.4). The majority of the variation is driven by HD as it separates from the other samples on PC1 (∼40% of the total variation in the data). The variation on PC2 is driven by C7 and C13 and explains around 18% of the variation in the data. PC3, representing approximately 11% of the total variation in the data, showed a separation of the parent cell line from C7, C13 and HD. There was no clear inferences that could be drawn from PC4.

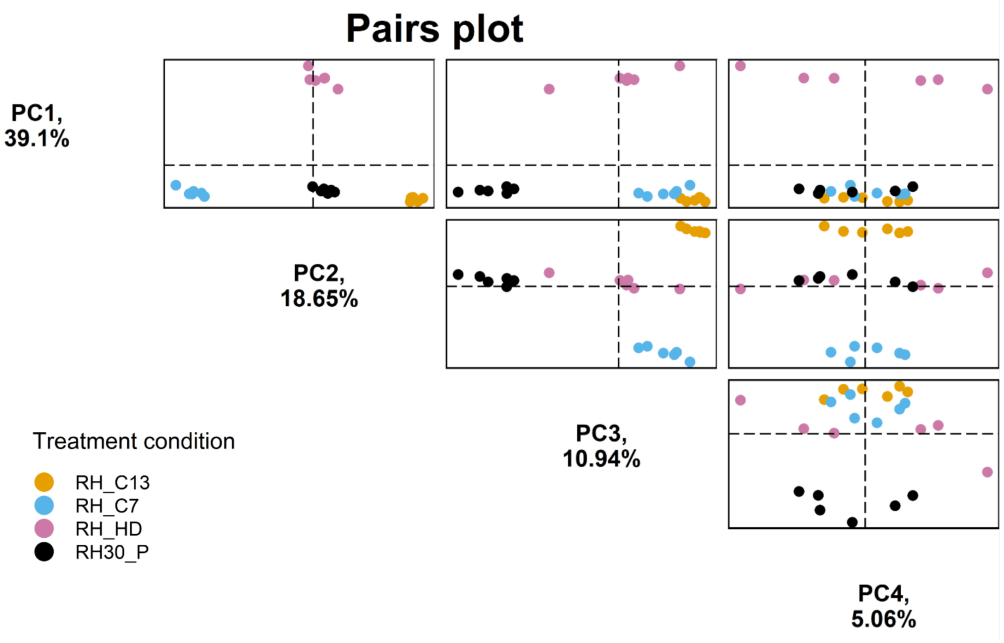


FIGURE 4.4: Pairs plot showing biplots for each PC (PC1 through PC4) for the intrinsic resistance RNA sequencing data. Plots in each column represent a different PC along the x-axis. Plots in each row represent a different PC along the y-axis. n = 24 (4 treatment conditions with 6 replicates each).

4.3.1.2 Differentially expressed genes in intrinsic resistant samples

The number of DEGs between the two clones were very similar, with around 80 genes upregulated and 350 downregulated (Figure 4.5). Unlike the clones, the HD sample had more upregulated DEGs than downregulated and a higher number of DEGs than the clones (Figure 4.5).

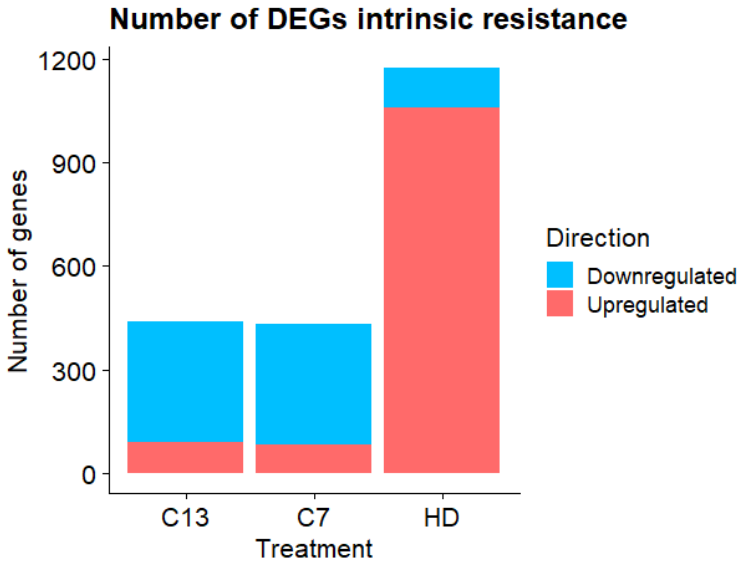


FIGURE 4.5: Bar plot showing the number of DEGs for clones and HD compared to parental control. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

Venn diagrams were plotted showing the number of overlapping DEGs for C7, C13 and HD vs P (Figure 4.6). Despite the clones showing nearly identical numbers of DEGs, there was little overlap between them, with only five (2.9%) upregulated and 109 (15%) downregulated genes in common (Figure 4.6B and C). However, when direction was ignored, there were 170 overlapping DEGs between C7 and C13. This indicates that 56 genes were differentially expressed in both clones but in opposite directions (Figure 4.6A). There was shared genes between the clones and HD, with a greater number of shared upregulated genes than downregulated genes. Only 4 genes were significantly upregulated in all conditions, these were *CASP10*, *BLK*, *CD33* and *TMC8*. 5 genes were downregulated in all conditions, three of which were long intergenic non-coding RNAs; *LINC00619*, *LINC01497*, *LINC00951* as well as *DCX* (involved in neurogenesis [328]) and *LHFPL1*. Downregulation of LINC00619 has been observed in osteosarcoma and was shown to regulate hepatocyte growth factor [329]. Upregulation of LINC00619 resulted in the inhibition of proliferation, migration and invasion of osteosarcoma cells and promoted apoptosis through inactivating hepatocyte growth factor signalling [329].

To further investigate the relationship between DEGs in the clones, a scatterplot comparing the LFC of genes for C7 vs P and C13 vs P was created (Figure 4.7). There is

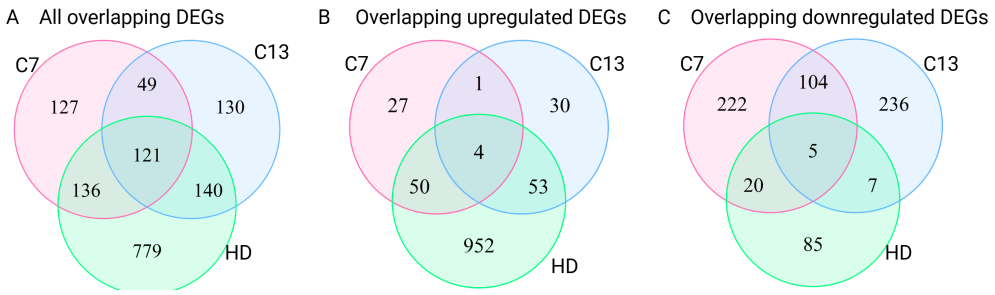


FIGURE 4.6: Venn diagram showing overlapping DEGs for C7 vs P, C13 vs P and HD vs P. A) All overlapping DEGs B) overlapping upregulated DEGs C) overlapping downregulated DEGs . DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

a weak but significant positive correlation in the expression pattern of genes between the clones (p<2.2 e-16, r=0.27). As seen in the Venn diagrams, there is very little overlap in upregulated DEGs and more overlap in downregulated DEGs. The majority of genes are only differentially expressed in one of the clones and not the other.

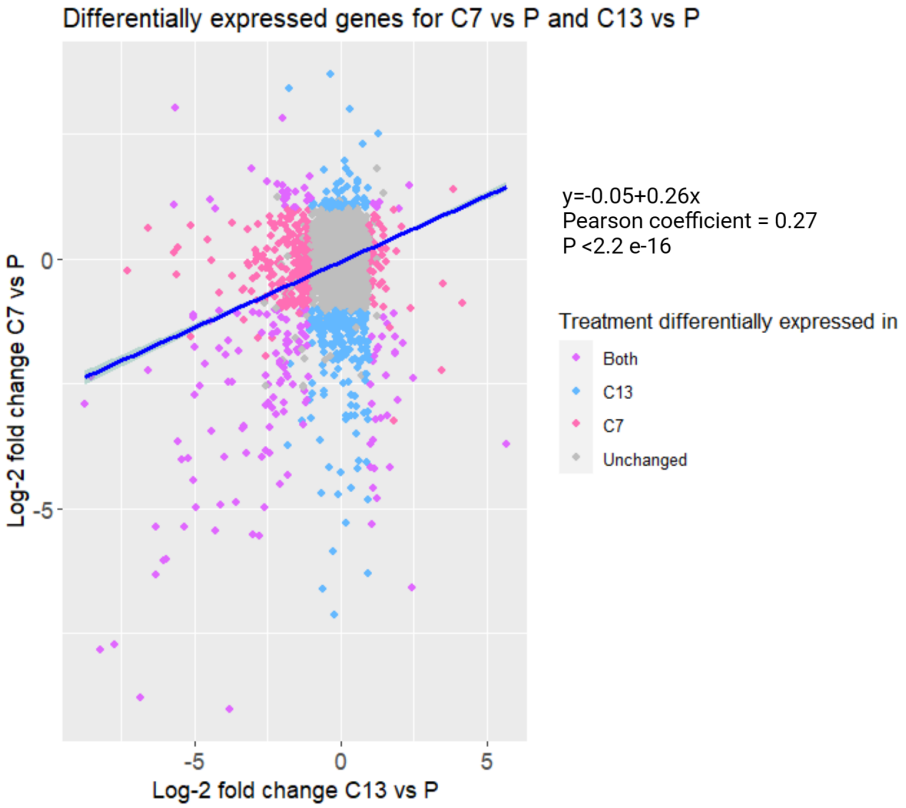


FIGURE 4.7: Scatterplot showing the LFC of all genes for C7 vs P and C13 vs P. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg). DEGs are highlighted in colour, corresponding to the group in which they are differentially expressed (C7, C13, both C7 and C13 or none (unchanged)).

4.3.1.3 Expression of multidrug-resistant genes and previously identified chemotherapy-resistant genes in rhabdomyosarcoma

As discussed in 1, previous research has found evidence for the involvement of Multi-Drug Resistant (MDR) genes in resistance to chemotherapy in RMS [121]. The expression of known MDR genes from the MDR/TAP subfamily was investigated, including *ABCB1*, *ABCB4*, *ABCC2* and *ABCB5*. *ABCB1* was downregulated in C7 (Supplementary Figure 20A). *ABCB1* and *ABCB4* were upregulated in HD (Supplementary Figure 20C).

4.3.1.4 Biological functions and biochemical pathways enriched in resistant samples

A summary of the GSEA is presented below. All GSEA tables results can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Intrinsic resistance/GSEA results).

Upregulated GO terms for C7 were linked to collagen, elastic fibres as well as 'skeletal muscle satellite cell differentiation' (Figure 4.8). Satellite cells are quiescent muscle precursor cells responsible for the postnatal growth, repair and maintenance of skeletal muscle. The GO term 'protein localisation to axon' was also upregulated, which could be related to a more neuronal phenotype. Downregulated gene sets were related to development and morphogenesis.

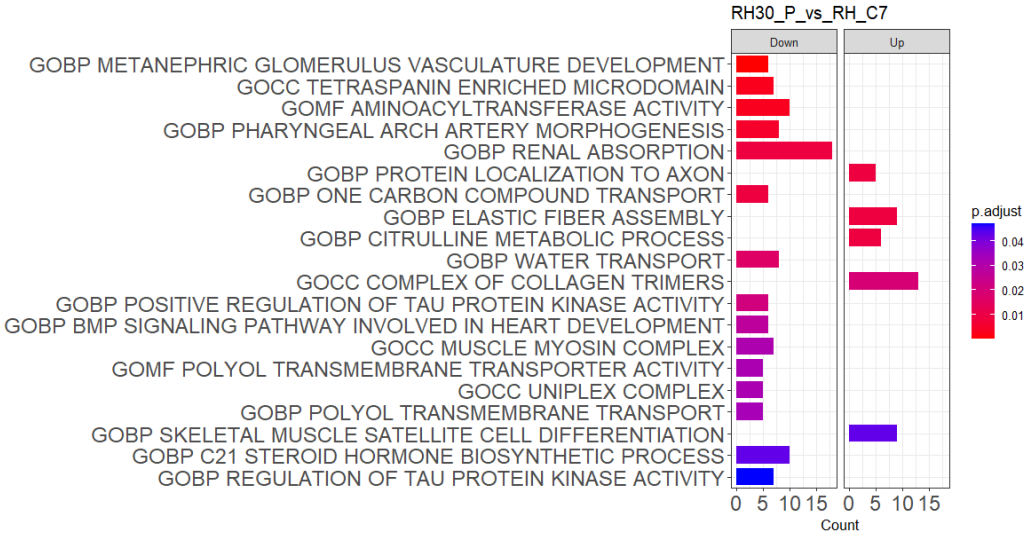


FIGURE 4.8: GSEA results from CAMERA showing enriched GO terms in C7 compared to P. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

The majority of the top 20 gene sets for C13 were downregulated and, like C7, included terms involved in differentiation and development (Figure 4.9). Immune

related gene sets linked to chemokines/chemotaxis were also downregulated. One Hallmark, 'angiogenesis', was downregulated which contained 24 genes.

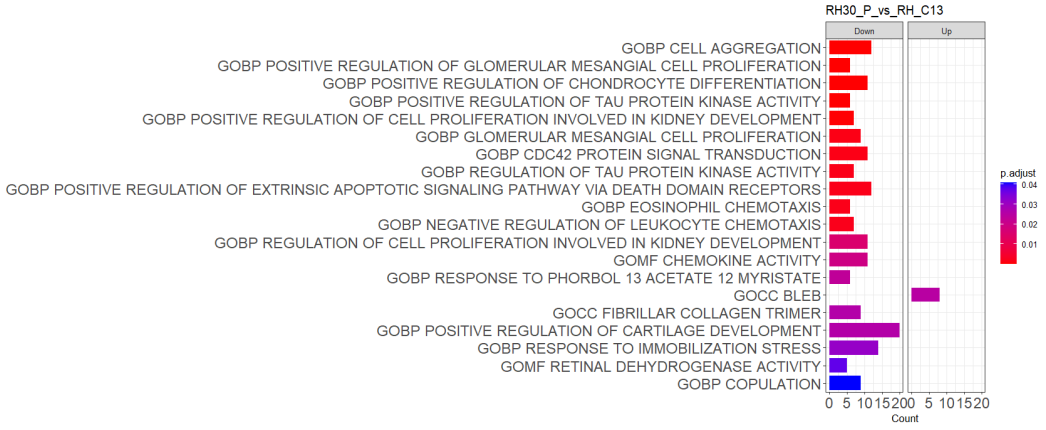


FIGURE 4.9: GSEA results from CAMERA showing enriched GO terms in C13 compared to P. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

HD had significantly more enriched gene sets than the clones. Multiple gene sets related to ribosomes were downregulated, including the largest GO term 'ribonucleoprotein complex biogenesis' (428 genes) which relates to the synthesis of ribonucleoproteins (complexes of protein and RNA) (Supplementary Figure 21). Other downregulated terms included mitochondrial and DNA replication-related terms. 'Collagen containing extracellular matrix', 'integrin binding', 'extracellular matrix structural constituent' and growth factor signalling terms were upregulated. Approximately 20% of all gene sets were immune-related. These were all upregulated, including 'response to interferon gamma', 'negative regulation of immune effector process' (64 genes, 'acute inflammatory response', 'negative regulation of T cell proliferation' and 'macrophage migration'.

Interferon and inflammatory response gene sets were also upregulated in the Hallmarks, as well as complement, KRAS signalling, hypoxia and apoptosis (Figure 4.10). Like GO terms, downregulated Hallmarks were related to DNA replication, specifically the DNA damage checkpoint G2M, as well as MYC and E2F targets.

A cell proliferation gene signature was also tested for enrichment using GSEA, to identify if resistance could be explained by a decrease in cell proliferation. Results can be found in Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Intrinsic resistance/proliferation_signature_results). Only HD showed a significant decrease in the proliferation gene signature compared to the parent cell line (p=3.51e-7).

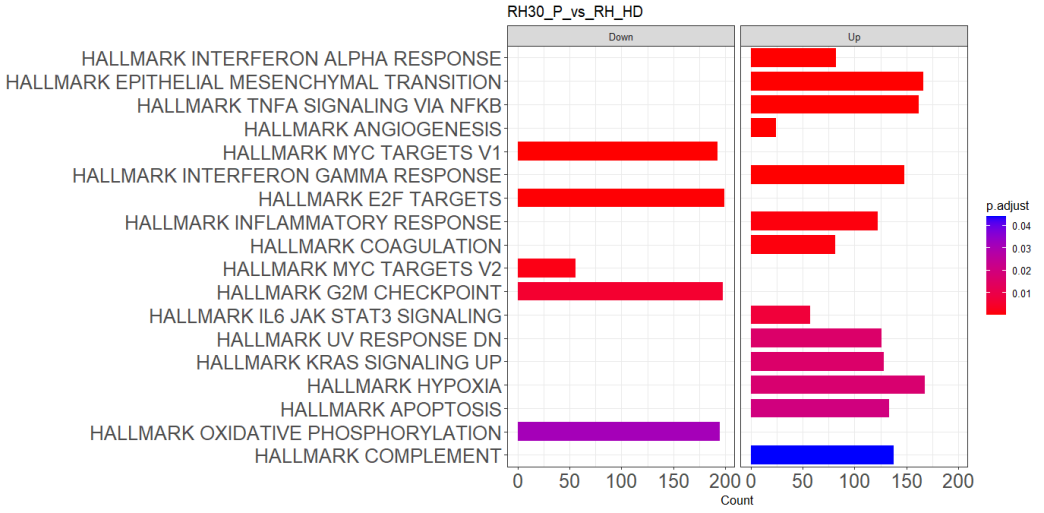


FIGURE 4.10: GSEA results from CAMERA showing enriched Hallmarks for HD compared to P. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

4.3.1.5 Rationale for excluding the high dose sample when deriving gene signatures

When deriving a gene signature of intrinsic resistance, it was decided to exclude the HD samples. This decision was based on the difference in how the HD model was created by prolonged exposure to high VCR concentrations compared to the single-cell clones which had not been exposed to drug. This rationale is supported by the PCA results which showed that HD was responsible for the majority of the variation in the data. Therefore, the intrinsic resistance signature was derived using only the clone samples which are considered to more closely reflect untreated pre-existing mechanisms of resistance.

4.3.1.6 Using weighted gene co-expression network analysis to identify modules of co-expressed genes correlated with intrinsic resistance

WGCNA was selected as the method to derive a gene signature by identifying modules of co-expressed genes correlated with intrinsic resistance. This approach avoids the use of arbitrary binary cut-offs for LFC, allowing detection of genes with modest but coordinated expression changes that may be overlooked when only considering DEGs. For example, there may be smaller LFC in many genes that share a biological process. Another approach considered for selection of a gene signature was to select overlapping DEGs between the C7 and C13. This was ruled out as there was minimal overlap in DEGs and 56 genes were significantly differentially expressed in both C7 and C13 but in opposite directions, highlighting significant differences between the clones.

WGCNA was conducted using samples from C7, C13 and Parent (P), which represented the non-resistant samples. The raw data were re-filtered to remove genes with low expression as the CPM threshold may be different since the HD samples were removed. A total of 14,179 genes remained after removing genes with a CPM of less than 0.25 in at least 6 samples.

Hierarchical clustering was used to create a dendrogram to visualise the similarities and differences between the samples and identify outliers (Figure 4.11A). Splits in the dendrogram indicate differences, with greater differences represented by greater heights. The first split was C13 from P and C7, suggesting that C13 was different to P and C7. The second split in the dendrogram was P from C7 suggesting there are differences in these groups. There is little difference in height between replicates in each group, indicating that the replicates are very similar to one another and there are no obvious outliers.

The soft threshold power was determined using a scale independence plot. A soft-thresholding power was selected based on the scale-free topology criterion, aiming for a value close to the inflection point with a high scale independence R^2. For this reason, a power of 10 was selected as the soft threshold (Figure 4.11B). There were 28 modules in the network, reduced to 8 after merging similar modules (modules below the threshold in Figure 4.11C) (Figure 4.11D). All module information can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Intrinsic resistance/WGCNA results/geneInfo_clones.csv).

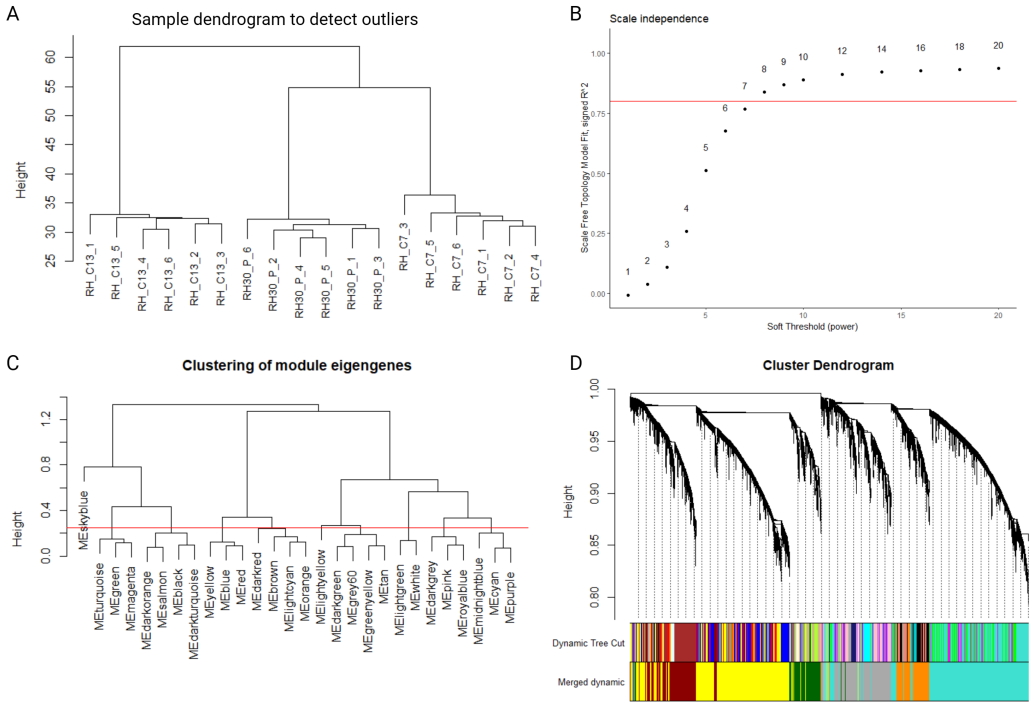


FIGURE 4.11: Outlier detection, selection of optimal soft threshold power and visualisation of modules for the gene co-expression network using C7, C13 and P samples. A) dendrogram showing C7, C13 and P samples B) scale independence plot for selection of optimal soft threshold power C) dendrogram of modules with red line showing dissimilarity threshold of 0.25 D) Cluster dendrogram of modules showing module colour before merging and after merging.

The modules were then correlated with the trait 'resistance'. Two modules had a strong significant positive correlation with resistance; 'darkgreen' and 'darkgrey' (Supplementary Figure 22). The 'darkgreen' module mainly showed high expression in C13 samples while the 'darkgrey' module showed high expression in both C7 and C13 (Figure 4.12).

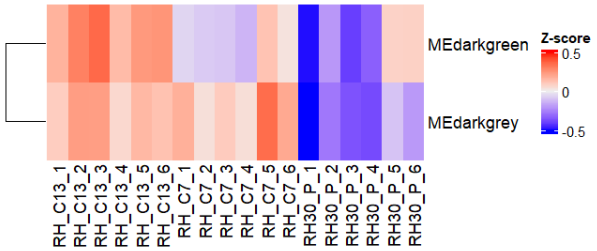


FIGURE 4.12: Heatmap of sample contribution to modules that had a significant positive correlation with intrinsic resistance.

The smallest module was the 'skyblue' module with 43 genes and the largest module was 'turquoise' with 4,211 genes (Figure 4.13). The average module size was 1,772. The 'darkgreen' module consisted of 1,159 genes and the 'darkgrey' module had 1,832 genes, both close to the average module size.

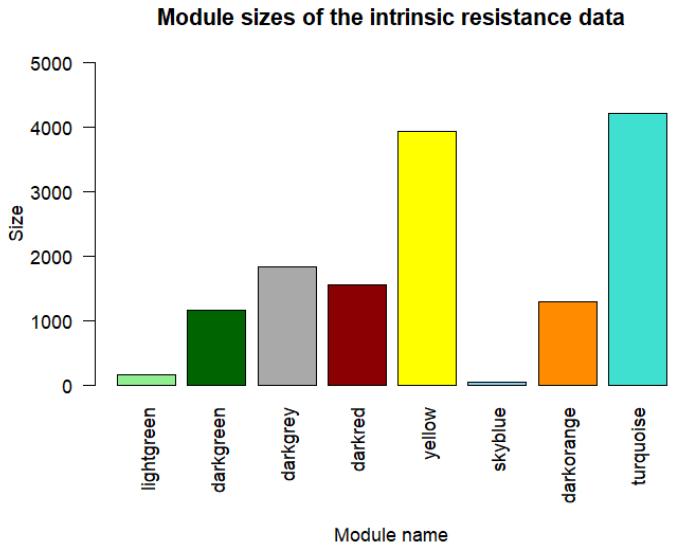


FIGURE 4.13: Module sizes of the merged modules from the weighted gene co-expression network using the C7, C13 and P samples from the intrinsic resistance dataset.

The correlation between gene significance and module membership for genes in the resistant modules was plotted (Figure 4.14). Gene significance is a correlation metric between 0 and 1 representing correlation between the gene expression and the binary trait 'resistance'. Module membership score is a correlation metric between 0 and 1 representing correlation with the gene expression and the module eigengene (the first principal component of a module that is representative of the gene expression profiles). Both modules showed a strong and statistically significant correlation between gene significance and module membership, indicating that genes most strongly associated with resistance are also central to their respective modules. A threshold of gene significance >0.6 and module membership >0.6 was used to filter for these important genes in the module called 'hub genes'. The 'darkgreen' module had 556 hub genes and 'darkgrey' module had 854 hub genes (Supplementary Table 8).

4.3.1.7 Gene set enrichment analysis on module hub genes

After hub genes were selected in the two resistant modules, over-representation-based GSEA was used to investigate the biological function of these hub genes. All GSEA can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Intrinsic resistance/WGCNA results/GSEA on hub genes from resistant modules). Hub genes from the 'darkgrey' module were significantly enriched in histone terms (Figure 4.15A) whilst the 'darkgreen' module was enriched for terms involving cell-ECM binding including 'cell cortex',

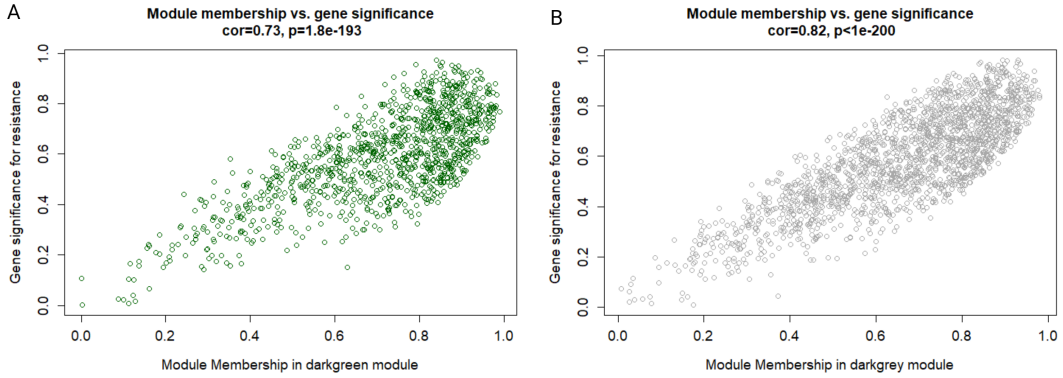


FIGURE 4.14: Gene significance and module membership plots of modules significantly correlated with intrinsic resistance. A) 'darkgreen' module and B) 'darkgrey' module. Each point represents gene. The Pearson correlation and Student asymptotic p-value for correlation for gene significance and module membership is shown.

'cell-substrate junction' and 'focal adhesion' (Figure 4.15C). Hub genes from both modules were also enriched for the Hallmark 'mitotic spindle' (Figure 4.15B and D).

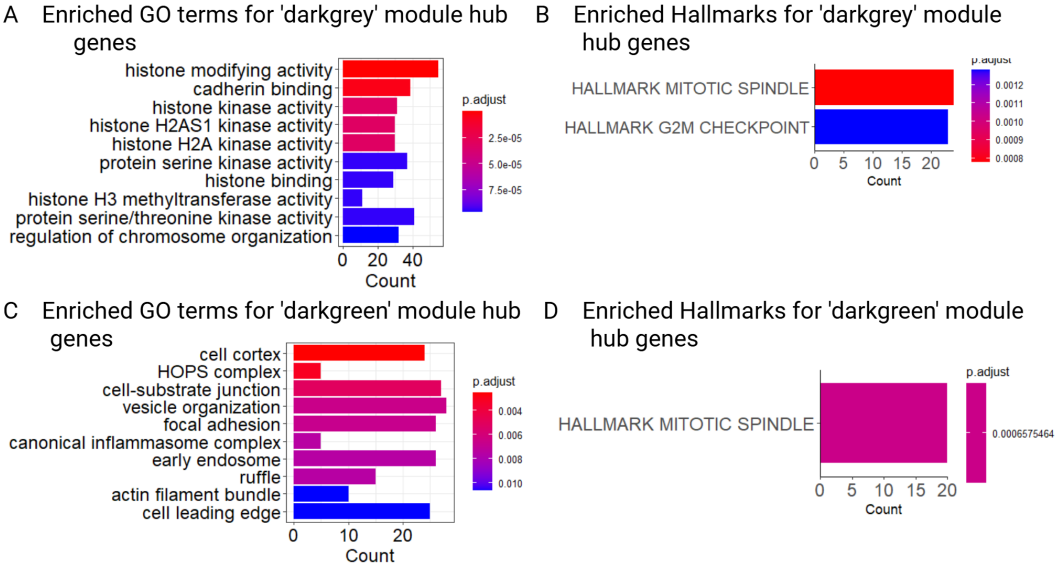


FIGURE 4.15: Enriched gene sets for hub genes in modules associated with intrinsic resistance. A) Enriched GO terms for 'darkgrey' B) enriched Hallmarks for 'darkgrey' C) enriched GO terms for 'darkgreen' D) enriched Hallmarks for 'darkgreen'. Hub genes defined as gene significance and module membership >0.6. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). If more than 10 terms were enriched, the ten terms with the smallest adjusted p values were displayed.

To further investigate the 'mitotic spindle' geneset, a heatmap containing the enriched mitotic spindle genes was plotted (Figure 4.16). The heatmap revealed higher expression of these genes in the resistant clones (indicated in red) and lower expression in the P samples (indicated in blue). Hierarchical clustering shows the first split is the P samples from the clones and the second split is C7 from C13 showing

these are the most similar. All replicates from each treatment condition cluster together increasing confidence in the reliability of the results.

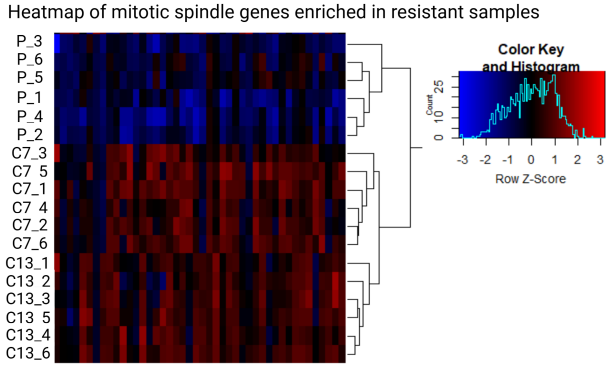


FIGURE 4.16: Heatmap of the 44 genes from the Hallmark 'mitotic spindle' in the 'darkgrey' and 'darkgreen' resistant modules from WGCNA. Z score has been applied to CPM-normalised RNA-seq counts.

4.3.1.8 Selection of intrinsic resistance signature genes from resistant modules using protein-protein interaction networks

After modules correlated with intrinsic resistance were identified, PPI networks were constructed to identify highly interacting proteins from genes in the modules for inclusion in the gene signature. PPI networks were constructed using STRING from the intrinsic resistant modules ('darkgreen' and 'darkgrey'). Alternative PPI networks were constructed using DEGs and WGCNA hub genes from resistant modules, but these networks had a minimal number of protein-interactions and did not form hub-like structures due to the paucity of genes (Supplementary Figure 23).

The interactive networks can be viewed in the link in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Intrinsic resistance/PPI). There were 102 proteins in the 'darkgreen' PPI network and 207 in the 'darkgrey' network (Supplementary Table 8 and Supplementary Figure 24). Highly interactive proteins (PPI hub proteins) were identified, resulting in 41 PPI hub proteins (Supplementary Table 9).

The hub proteins were cross-referenced with hub genes from the WGCNA to identify interacting proteins that were highly correlated with the trait resistance and important to the module (high gene significance and module membership). This ensures the exclusion of highly interacting proteins that were not correlated with resistance. This resulted in 7 genes from the 'darkgreen' module and 11 from the 'darkgrey' module (Table 4.3). Combined, there was a total of 18 genes to be used as the intrinsic resistance signature (Table 4.3). These genes did not show large LFCs in differential expression between C7, C13 and the parental samples (Table 4.3), which is consistent with the WGCNA approach that prioritises co-expression patterns over DEGs.

TABLE 4.3: Genes from the intrinsic resistance signature and their information from WGCNA, PPI and differential expression. P values were adjusted using Benjamini-Hochberg method.

Gene	Module	Gene significance	Module membership	PPI centrality score	LFC C7 vs P	P.adjust	LFC C13 vs P	P.adjust
SQSTM1	Darkgreen	0.85	0.96	12.00	0.14	0.00	0.29	0.00
USP10	Darkgreen	0.79	0.77	10.00	0.07	0.02	0.10	0.00
CDC16	Darkgreen	0.73	0.84	8.00	0.06	0.06	0.14	0.00
NR3C1	Darkgreen	0.70	0.87	8.00	0.06	0.13	0.16	0.00
BRCC3	Darkgreen	0.67	0.77	8.00	0.08	0.05	0.18	0.00
UBC	Darkgreen	0.65	0.94	36.00	0.03	0.17	0.18	0.00
UVRAG	Darkgreen	0.65	0.70	8.00	0.07	0.07	0.11	0.00
UPF1	Darkgrey	0.89	0.91	10.00	0.29	0.00	0.25	0.00
NCBP1	Darkgrey	0.85	0.78	8.00	0.08	0.00	0.08	0.00
RBBP5	Darkgrey	0.81	0.77	16.00	0.14	0.00	0.12	0.00
EIF3B	Darkgrey	0.79	0.95	8.00	0.13	0.00	0.06	0.02
SRRM2	Darkgrey	0.78	0.90	8.00	0.33	0.00	0.19	0.01
POLR2A	Darkgrey	0.74	0.71	14.00	0.18	0.00	0.21	0.00
XIAP	Darkgrey	0.73	0.71	8.00	0.12	0.00	0.13	0.00
CHD4	Darkgrey	0.71	0.76	8.00	0.12	0.00	0.09	0.02
MAPK1	Darkgrey	0.68	0.78	10.00	0.06	0.03	0.29	0.00
NUP153	Darkgrey	0.63	0.79	8.00	0.13	0.00	0.05	0.08
SETD1A	Darkgrey	0.61	0.83	8.00	0.17	0.00	0.07	0.10

4.3.1.9 Scoring samples based on the expression of genes in the intrinsic resistance signature

GSVA was used to calculate enrichment scores for each sample in the intrinsic and acquired models of resistance based on the expression of genes in the respective signatures. The Shapiro-Wilk normality test indicated that the data were not normally distributed (p value <0.05). Consequently, a Kruskal-Wallis rank sum test (non-parametric ANOVA) was used to test for a significant difference between groups, indicating a significant difference (p value <0.05). Pairwise comparisons between groups were performed using the Wilcoxon rank sum exact test. The GSVA score for the intrinsically resistant clones C7 and C13 were significantly different from P and HD (Figure 4.17). No significant difference in GSVA score was observed between the HD and P samples. For the acquired resistance dataset, there was no significant difference between the chemo-resistant samples and the control in either ARMS or ERMS cell lines (Figure 4.17).

4.3.1.10 Investigating modules negatively correlated with intrinsic resistance

There were two modules that had a significant negative correlation with resistance (Supplementary Figure 22). GSEA of GO terms was performed on the module hub genes to investigate mechanisms that may be protective against resistance in the intrinsic cell model. Both of the modules showed enrichment of ribosomal-related terms (Figure 4.18A and B).

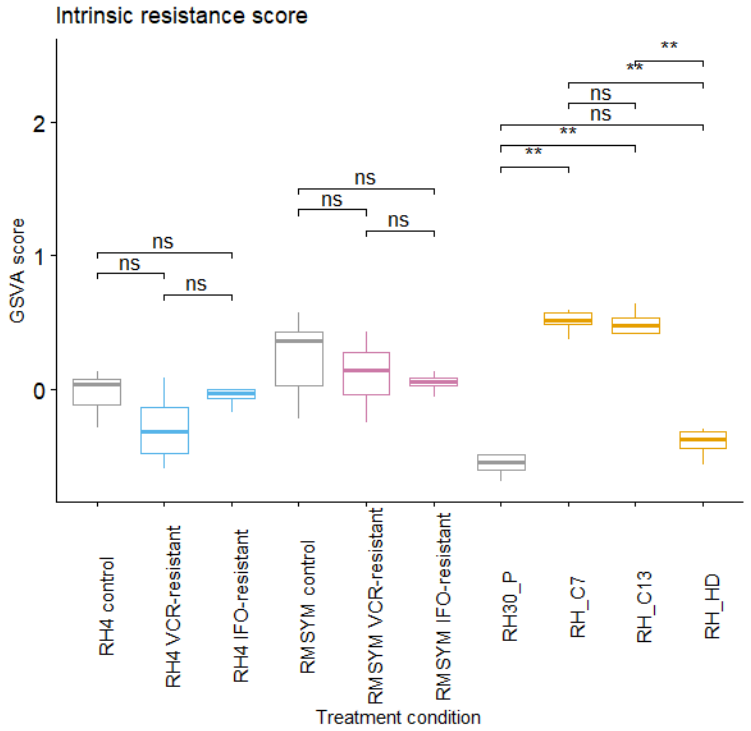


FIGURE 4.17: GSVA scores of the intrinsic resistance signature. Pairwise comparisons were made using the Wilcoxon rank sum exact test. P values were adjusted using Benjamini-Hochberg method. ns: p >0.05, *: p ≤0.05, **: p ≤0.01, ***: p ≤0.001, ****: p ≤0.0001.

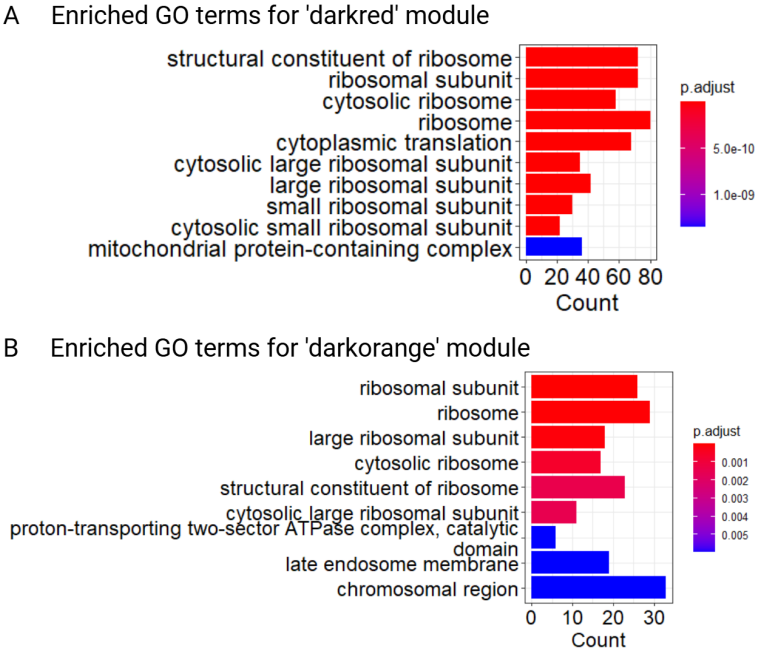


FIGURE 4.18: Enriched GO terms for modules with a significant negative correlation with intrinsic resistance. A) 'darkred' module B) 'darkorange' module. P values were adjusted using Benjamini-Hochberg method. Significant terms were defined as p adjust <0.05. Top ten GO terms with the smallest adjusted p values are shown.

4.3.2 Analysis of the acquired resistance data

The analysis was repeated for data from the acquired resistant cell models, where the fusion-positive cell line RH4 and fusion-negative cell line RMSYM were treated with VCR or IFO to generate resistant cells.

4.3.2.1 Quality control of the data

The full MULTIQC report can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/multiqc_report.html). All samples passed quality control checks for sequence quality, N content, sequence length distribution and overrepresented sequences (Figure 4.19). GC content had a warning for samples RMSYMifo_2_1, RMSYMifo_5_1 and RMSYMvcr_5_1. The MULTIQC manual states that warnings for this QC is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads. If there are sharp peaks this may represent a specific contaminant whilst broader peaks could be contamination with a different species. The samples that were flagged with a warning were not flagged in any other QC checks so they were flagged as potential outliers but not excluded to retain as many samples as possible. Adapter content had warnings for samples RH4vcr_1_1 and RH4vcr_1_2. 'Sequence Duplication Levels' and 'Per Base Sequence Content' failed, which is expected for RNA sequencing data.

All samples had a high percentage of uniquely mapped reads >85%, showing the majority of reads did not map to multiple locations in the reference genome (Supplementary Table 10).

To determine the CPM threshold to be used for filtering, the CPM vs raw counts for the first sample were plotted. A CPM of around 0.25 corresponded to a count of around 10-15, so this was selected as the CPM filter threshold (Supplementary Figure 25). This resulted in 16,865 after filtering to remove low read counts. Density plots were used to visualise the effect of filtering on the data (Supplementary Figure 26). After filtering, the first peak (corresponding to counts of 0) has been greatly reduced, while the second peak was retained (Supplementary Figure 25).

Quality control of the data was assessed before and after normalisation. Boxplots of the log-2 CPM identified no outliers (Supplementary Figure 27A and B). A barplot of the library sizes showed no obvious differences between treatment conditions and was similar for all samples (Supplementary Figure 27C).

The elbow method and Horn's method were used to determine the optimum number of PC to retain. Of these two values, the larger number of PC's to retain was chosen (5) (Supplementary Figure 28). PC1 variation was driven cell line (ARMS or ERMS)

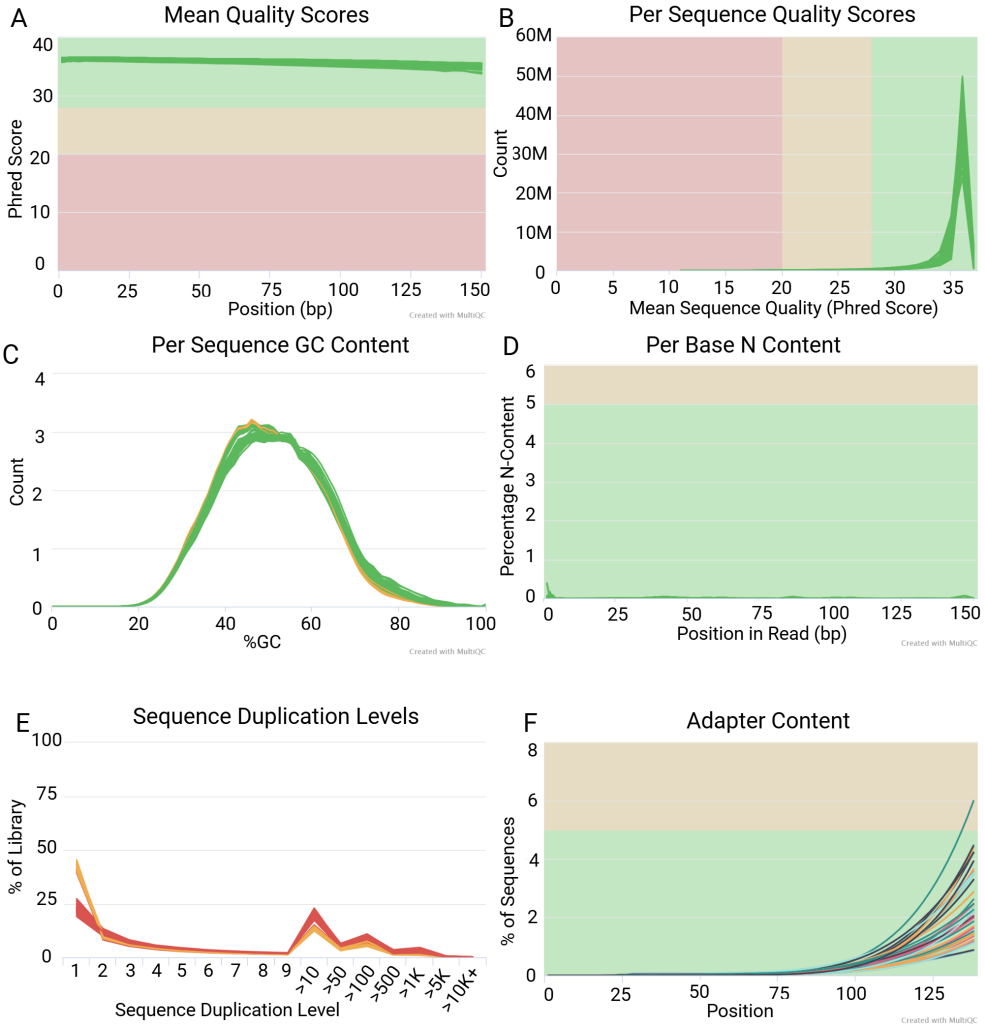


FIGURE 4.19: MULTIQC report for the acquired resistance data. A) Mean quality scores showing the phred score for each bp in the read B) Per sequence quality scores C) Per sequence GC content D) Per base N content, where N represents where the sequencer is unable to make a base call with sufficient confidence E) Sequence duplication levels F) Adapter content.

which accounted for around 50% of the total variation in the data. All RMSYM samples clustered together showing they are similar. RH4 IFO-resistant was driving the variation on PC2, which represented 10% of the total variation in the data and separated RH4 IFO-resistant from RH4 VCR-resistant and RH4 control. This showed that RH4 VCR-resistant and RH4 control were more similar and RH4 IFO-resistant was different. Replicates clustered together tightly, indicating there was very little variation between replicates. PC3 shows the variation between the VCR-resistant and IFO-resistant samples, which accounts for around 9% of the total variation in the data. No clear inferences could be drawn from PC4. PC5 shows around 4% of the total variation is between the resistant and the control samples.

The common dispersion value was 0.012 and BCV was 0.11 which is in line with the expected BCV value for cell lines (Supplementary Figure 29).

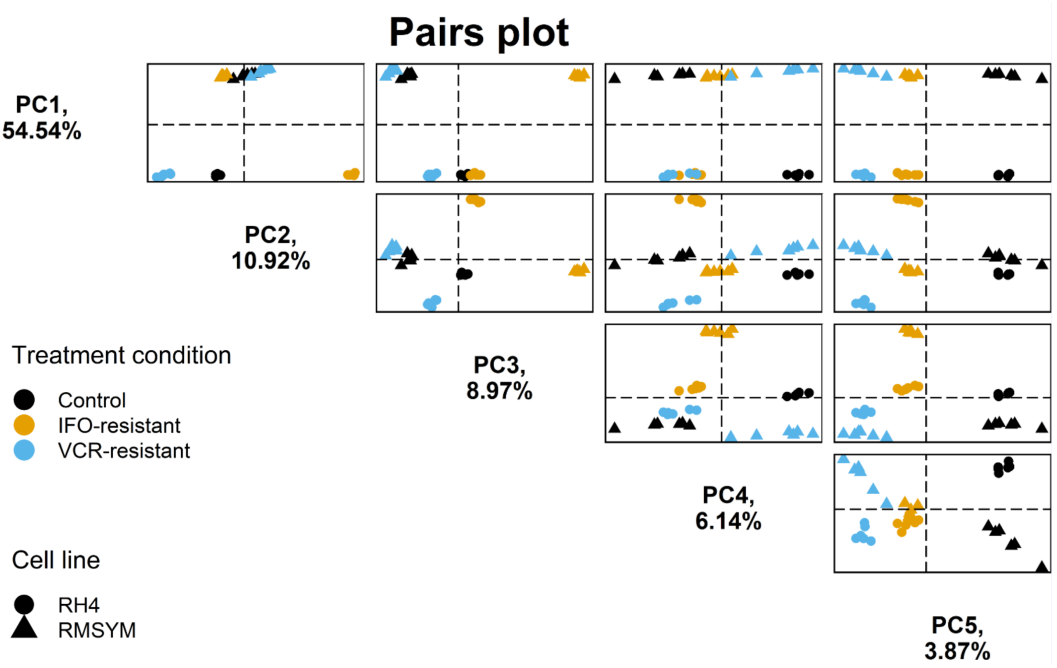


FIGURE 4.20: Pairs plot showing biplots for each PC (PC1 through PC5) for the acquired resistance RNA sequencing data. Plots in each column represent a different PC along the x-axis. Plots in each row represent a different PC along the y-axis.n = 36 (6 treatment conditions with 6 replicates each).

4.3.2.2 Differentially expressed genes in acquired resistant samples

In both ERMS and ARMS cell lines, VCR-resistant samples resulted in fewer DEGs than IFO-resistant samples (Figure 4.21). RH4 IFO-resistant, RMSYM IFO-resistant and RMSYM VCR-resistant all resulted in more downregulated than upregulated genes with only RH4 VCR-resistant resulting in more upregulated than downregulated genes.

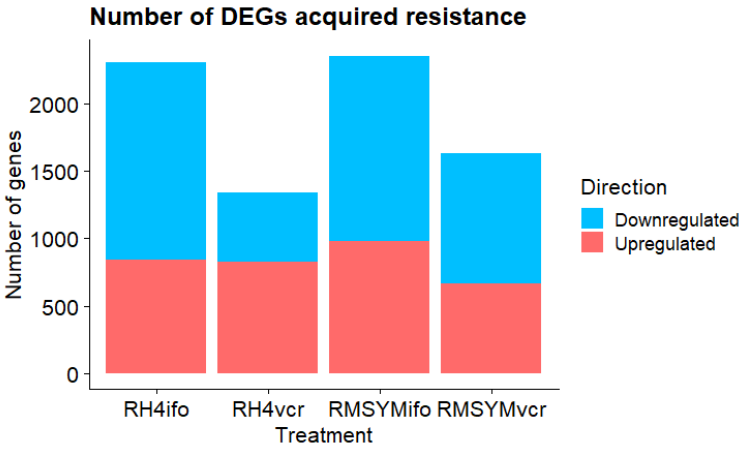


FIGURE 4.21: Bar plot showing the number of DEGs for clones and HD compared to parental control. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

Venn diagrams were used to compare overlap in DEGs . First, the overlap in DEGs between ARMS and ERMS IFO-resistant cell lines was looked to identify genes that may be common to the chemotherapy agent (Figure 4.22A and C). There were 108 overlapping upregulated DEGs (6%) and 220 overlapping downregulated DEGs (8%). VCR-resistant ARMS and ERMS cell lines showed less overlap in DEGs than IFO-resistant, with only 49 (3%) upregulated DEGs shared and 59 (4%) downregulated (Figure 4.22B and D).

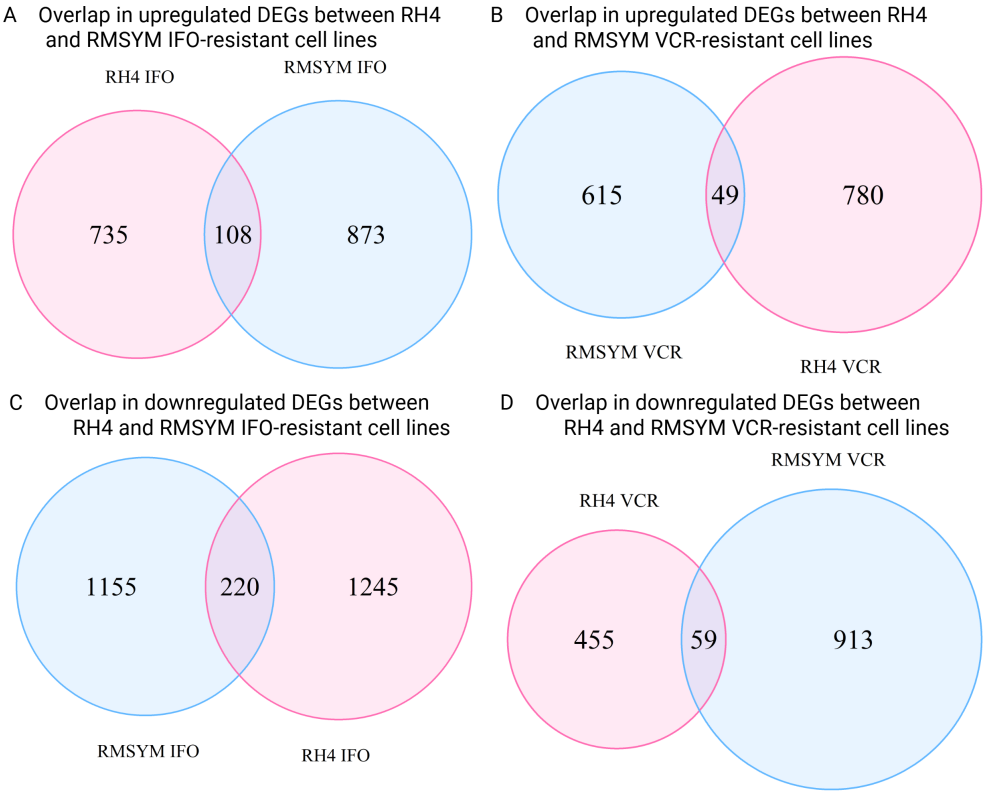


FIGURE 4.22: Venn diagrams showing the overlap in DEGs between chemo-resistant ARMS and ERMS cells. A) Upregulated DEGs between RH4 and RMSYM IFO-resistant cell lines B) Upregulated DEGs between RH4 and RMSYM VCR-resistant cell lines C) Downregulated DEGs between RH4 and RMSYM IFO-resistant cell lines D) Downregulated DEGs between RH4 and RMSYM VCR-resistant cell lines. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

The overlap in DEGs between chemotherapy agents in cell lines representing the different RMS subtypes was also investigated, to identify cells that may be common to resistance in the RMS subtype rather than the specific chemotherapy agent. RH4 IFO-resistant and RH4 VCR-resistant cell lines were compared, to identify genes that may be common to ARMS resistant cells. There were 133 (8%) shared upregulated DEGs , around 2% more than the overlap in ARMS and ERMS IFO-resistant cell lines and 5% more than VCR-resistant ERMS and ARMS cell lines. There were 188 (10%) overlapping downregulated DEGs . When comparing RMSYM IFO-resistant and VCR-resistant cells, there were 201 (12%) shared upregulated genes and 335 (14%)

shared downregulated genes. There was more overlap in DEGs between IFO and VCR-resistant ARMS cells than there was between ERMS and ARMS IFO-resistant cells and VCR-resistant cells. These results could indicate that RMS subtype is driving gene expression changes.

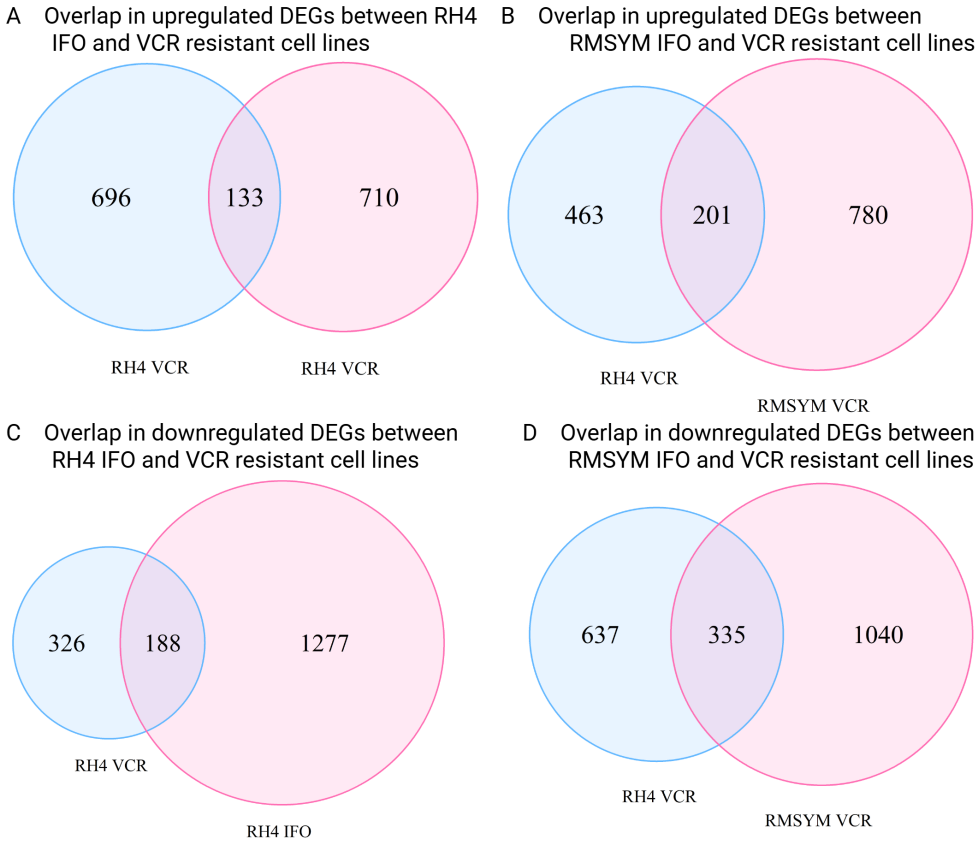


FIGURE 4.23: Venn diagrams showing the overlap in DEGs between VCR-resistant and IFO-resistant cell lines representing ARMS and ERMS subtypes. A) Upregulated DEGs between IFO-resistant and VCR-resistant RH4 cell lines B) Upregulated DEGs between IFO-resistant and VCR-resistant RMSYM cell lines C) Downregulated DEGs between IFO-resistant and VCR-resistant RH4 cell lines D) Downregulated DEGs between IFO-resistant and VCR-resistant RMSYM cell lines. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

Scatterplots were used to further investigate the relationship between DEGs in different resistant samples. First, the relationship between RH4 IFO-resistant and RH4 VCR-resistant samples was explored. There was a very weak but significant positive correlation between DEG expression in RH4 IFO-resistant and RH4 VCR-resistant samples (Figure 4.24A). Although there were genes that were differentially expressed in the same direction, there were a significant number of genes that were expressed in opposite directions in the two conditions. When comparing the expression of RMSYM IFO-resistant and RMSYM VCR-resistant DEGs, there was a weak significant positive correlation (Figure 4.24B). As seen in the Venn diagrams, there was a large number of genes that is both downregulated in RMSYM IFO-resistant and RMSYM VCR-resistant samples. Like in RH4, there were still genes differentially expressed in

opposite directions between samples. When comparing DEGs for chemotherapy agent between RMS subtype, both chemotherapy agents showed a significant very weak negative correlation (Figure 4.24C and D). This again may indicate that RMS subtype shows more similar gene expression patterns than chemotherapy agent.

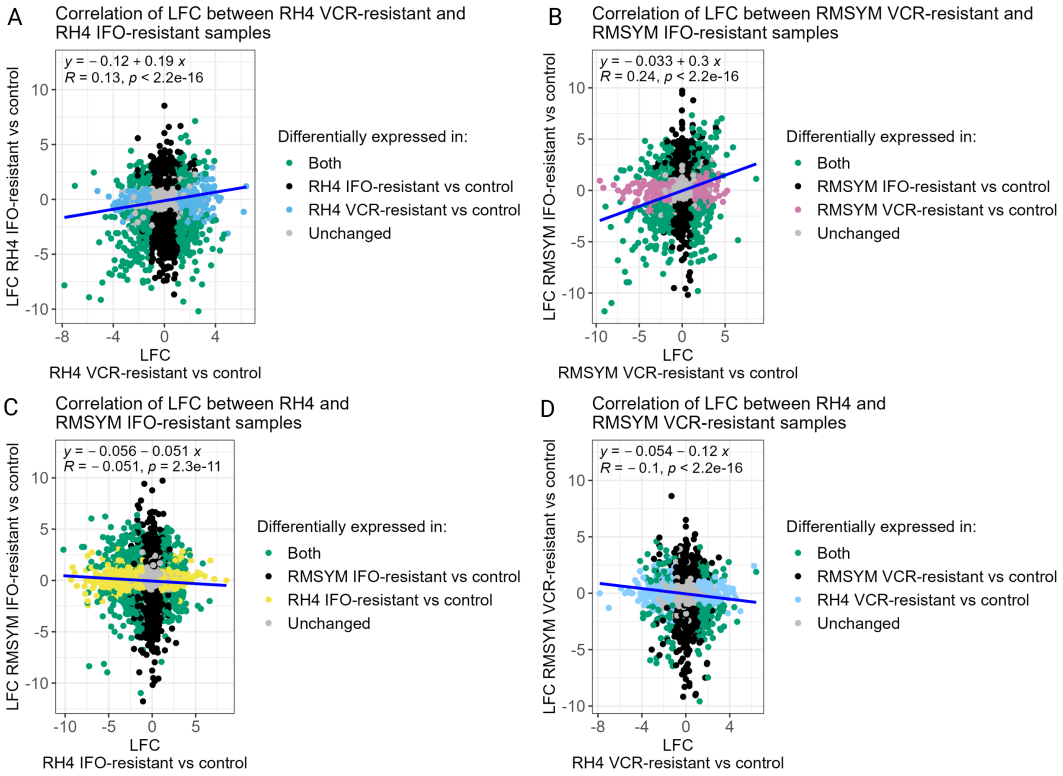


FIGURE 4.24: Scatterplots showing the correlation between the LFC of genes for the acquired resistance data. A) RH4 VCR-resistant vs control and RH4 IFO-resistant vs control B) RMSYM VCR-resistant vs control and RMSYM IFO-resistant vs control C) RH4 IFO-resistant vs control and RMSYM IFO-resistant vs control D) RH4 VCR-resistant vs control and RMSYM VCR-resistant vs control. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg). DEGs are highlighted in colour, corresponding to the condition in which they are differentially expressed in.

4.3.2.3 Biological functions and biochemical pathways enriched in acquired resistant samples

The full GSEA results can be found in Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/GSEA results). A summary of the GSEA is presented below. There were no enriched Hallmarks and few enriched Reactome gene sets. Gene sets that were enriched were small in size (∼5-10 genes).

RH4 VCR-resistant samples showed downregulation of terms related to the immune response (Figure 4.25) including 'immune response inhibiting cell surface receptor signalling' and 'immune response inhibiting signal transduction'. For RMSYM VCR-resistant samples, there was downregulation of neuronal-related terms (Figure 4.26). RMSYM IFO-resistant samples showed downregulation of gene sets linked to mesenchymal cells as well as a GO term related to the humoral immune response (Figure 4.27).

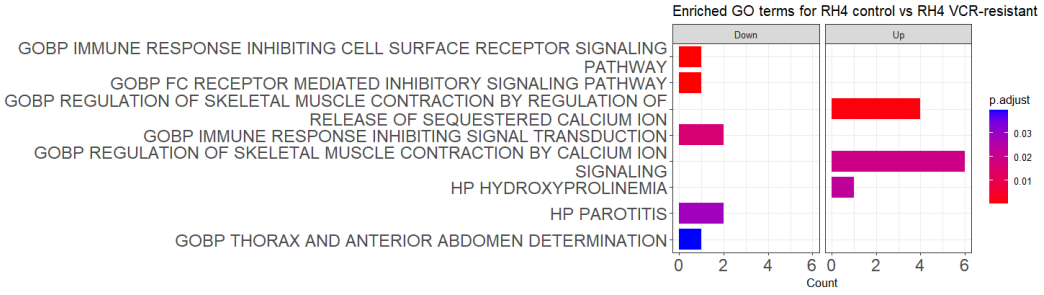


FIGURE 4.25: GSEA results from CAMERA showing enriched GO terms for RH4 VCR-resistant vs control cells. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

4.3.2.4 Expression of multidrug-resistant genes and previously identified chemotherapy-resistant genes in acquired resistance samples

None of the MDR genes were significantly differentially expressed in the RMSYM IFO-resistant cells (Supplementary Figure 30A). In RMSYM VCR-resistant cells, *ABCB1* and *ABCB4* were significantly differentially expressed (Supplementary Figure 30B). No MDR genes were significantly differentially expressed in RH4 cells (Supplementary Figure 30C and D). *GLI1* was not significantly differentially expressed in any VCR-resistant treatment conditions.

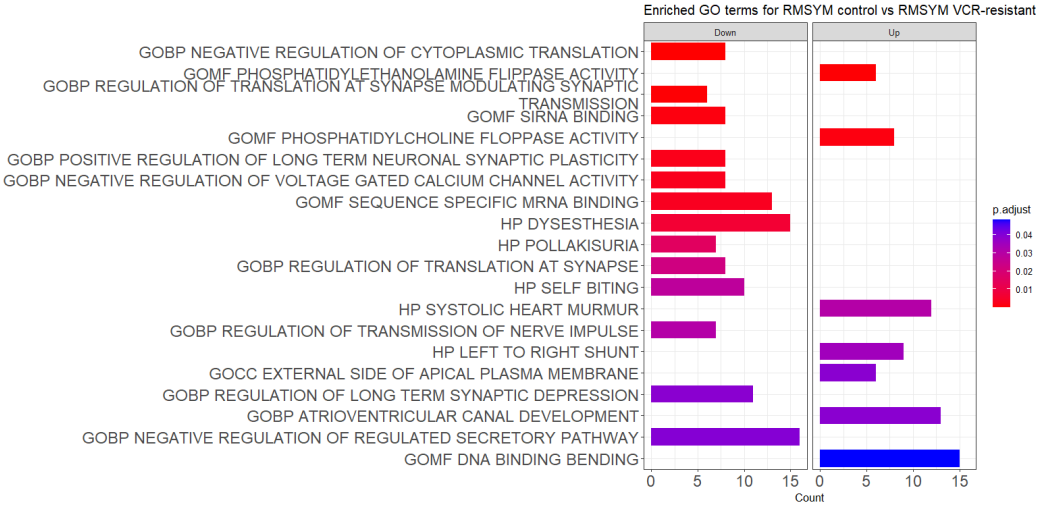


FIGURE 4.26: GSEA results from CAMERA showing enriched GO terms for RMSYM VCR-resistant vs control cells. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

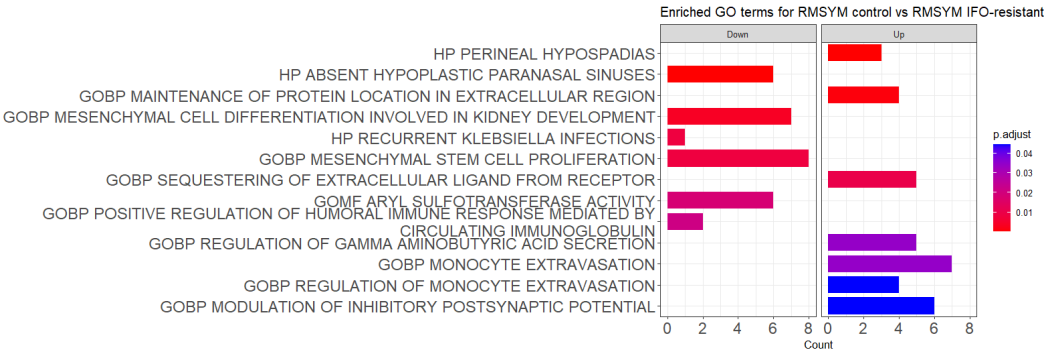


FIGURE 4.27: GSEA results from CAMERA showing enriched GO terms for RMSYM IFO-resistant vs control cells. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

4.3.2.5 Construction of a weighted gene co-expression network using all samples

First, WGCNA was performed using all acquired resistance samples from both ARMS and ERMS IFO-resistant and VCR-resistant samples (Supplementary Figure 31). However, no modules had a strong significant positive correlation with resistance (Supplementary Figure 32).

4.3.2.6 Rationale for the separation of fusion-positive and fusion-negative samples for weighted gene co-expression network analysis

One possible explanation for the lack of a significant correlation is the high degree of variation in the data, primarily driven by the different cell lines. A PCA biplot showed

that the "cell line" (RH4 or RMSYM) was driving the majority of the variation in the data and resistant / control samples did not separate on PC1 or PC2. It is unclear whether this variation is due to cell line specific effects or difference between RMS subtypes. To further investigate whether separating the data by cell line may help distinguish resistant from control samples, PCA plots were created for the RH4 and RMSYM samples separately. In the RH4 samples, IFO-resistant samples accounted for the majority of variation in PC1 (Figure 4.28A). However, it can be seen that the control sample separates from the resistant samples in PC2. This is also observed for RMSYM samples (Figure 4.28B). Since FP and FN-RMS differ in genetics, prognosis and histology, they may also differ in mechanisms of resistance. Previous research has identified different mechanisms of resistance in ARMS and ERMS [156, 161]. In addition, Venn diagrams suggested that the expression patterns were more strongly influenced by RMS subtype since these were more similar than the chemotherapy agent. It is possible that modules of co-expressed genes correlate with resistance in one RMS subtype but not the other, resulting in a weak and non-significant correlation with resistance when both subtypes are analysed together. For this reason, it was decided to repeat the WGCNA and create two networks for each RMS subtype.

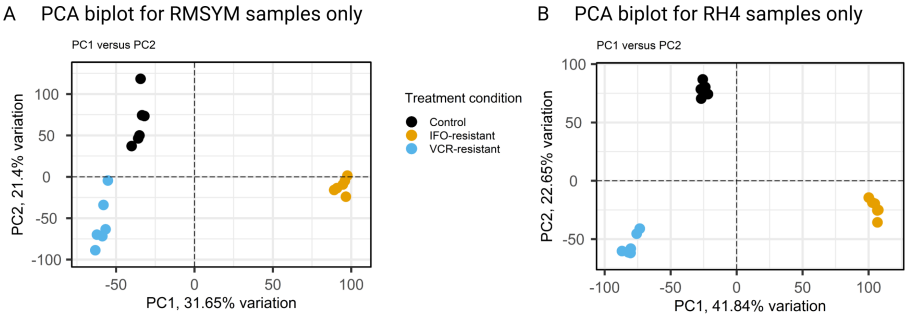


FIGURE 4.28: PCA biplot showing PC1 vs PC2 for A) RH4 samples B) RMSYM samples. Lower 10% of variables based on variance were removed. Ellipses show 95% confidence levels.

4.3.2.7 Identification of modules associated with acquired resistance in fusion-positive rhabdomyosarcoma using weighted gene co-expression network analysis

Initial filtering of the raw count data would have kept genes that were expressed in the RMSYM samples but not expressed in any RH4 samples. These genes have to be removed when constructing the RH4 WGCNA, as they will show a count of 0 for the RH4 samples. The raw count data was refiltered to remove these genes. After filtering to remove genes with low expression, 15,469 genes remained. No outliers were identified in the dendrogram (Figure 4.29A). A soft threshold power of 12 was selected (Figure 4.29B). There were 36 modules before merging and 7 modules after merging similar modules (Figure 4.29C and D).

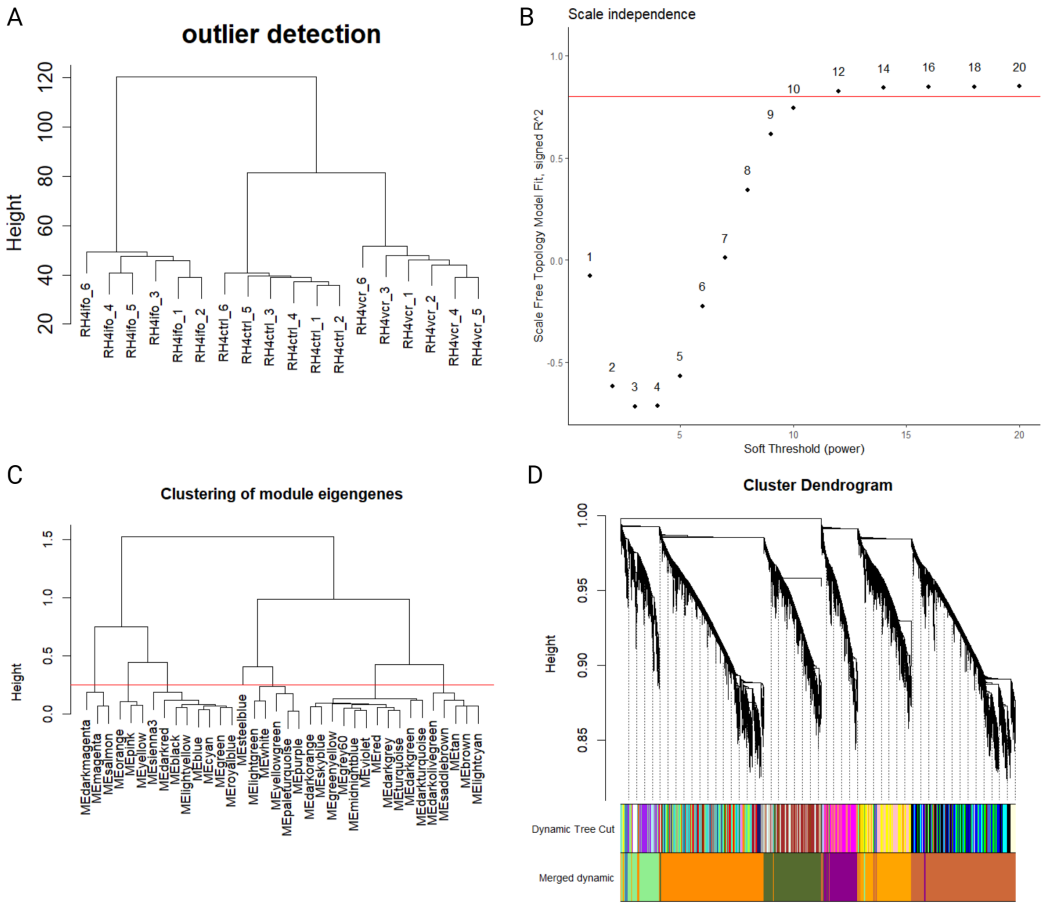


FIGURE 4.29: Outlier detection, selection of optimal soft threshold power and visualisation of modules for the gene co-expression network for WGCNA using FP-acquired resistance samples. A) dendrogram of samples B) scale independence plot for selection of optimal soft threshold power C) dendrogram of modules with red line showing dissimilarity threshold of 0.25 D) Cluster dendrogram of modules showing module colour before and after merging.

All module information can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/RH4/RH4_geneInfo.csv). The average module size was 2,209 genes. The largest modules were the 'darkorange' and 'sienna3', consisting of around 4,500 genes, while 'steelblue' was the smallest module with only around 100 genes (Supplementary Figure 33).

Two modules had a strong significant positive correlation with resistance; 'lightgreen' and 'orange' (Figure 35). 'Steelblue' module was significantly positively correlated with resistance, but the correlation was not as strong (0.59). The 'lightgreen' module consisted of 1,249 genes, the 'orange' module 1,948 genes and the 'steelblue' had 111 genes (Supplementary Table 11). The 'orange' module showed highest expression in the IFO-resistant samples while the 'lightgreen' module showed highest expression in the VCR-resistant samples (Figure 4.30). The 'steelblue' module showed high

expression in both IFO and VCR-resistant samples, but this expression was not consistent in all replicates (Figure 4.30).

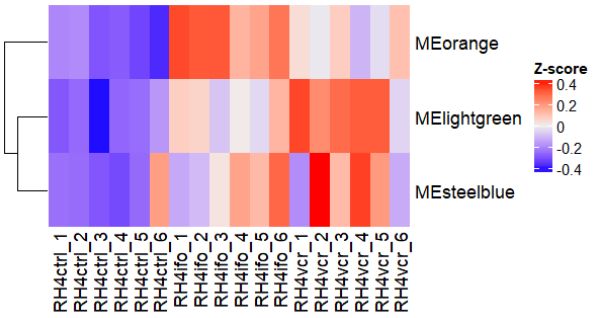


FIGURE 4.30: Heatmap showing sample contribution to modules with a significant positive correlation with FP-acquired resistance.

The 'lightgreen' and 'orange' modules showed a significant strong positive correlation between gene significance and module membership scores, whilst the 'steelblue' module was a weak significant positive correlation (Figure 4.31). This means that for the 'steelblue' module, genes that are contributing to resistance may not be similar to the module eigengene and vice versa. For this reason, the 'steelblue' module was not investigated for further analysis.

Hub genes (with a gene significance and module membership >0.6) were selected from the 'orange' and 'lightgreen' modules. There were 826 hub genes in the 'lightgreen' module and 1,142 in the 'orange' module (Supplementary Table 11).

4.3.2.8 Gene set enrichment analysis on fusion-positive module hub genes

All GSEA results can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/RH4/GSEA). Similarly to intrinsic resistant modules, FP-acquired resistant modules were enriched for gene sets related to cell-ECM adhesion and the mitotic spindle (Figure 4.32). Neuronal-related terms were also enriched including 'dendritic spine' and 'neuronal spine' (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/RH4/GSEA/GO_hub_genes_orange.png).

4.3.2.9 Selection of fusion-positive acquired resistance signature genes from resistant modules using protein-protein interaction networks

The same approach was used for the acquired resistance dataset as was used for the intrinsic resistance data. 21 PPI hub proteins from 'orange' PPI network (Table 4.4 and Supplementary Figure 34A) and 13 from 'lightgreen' (Table 4.4 and Supplementary

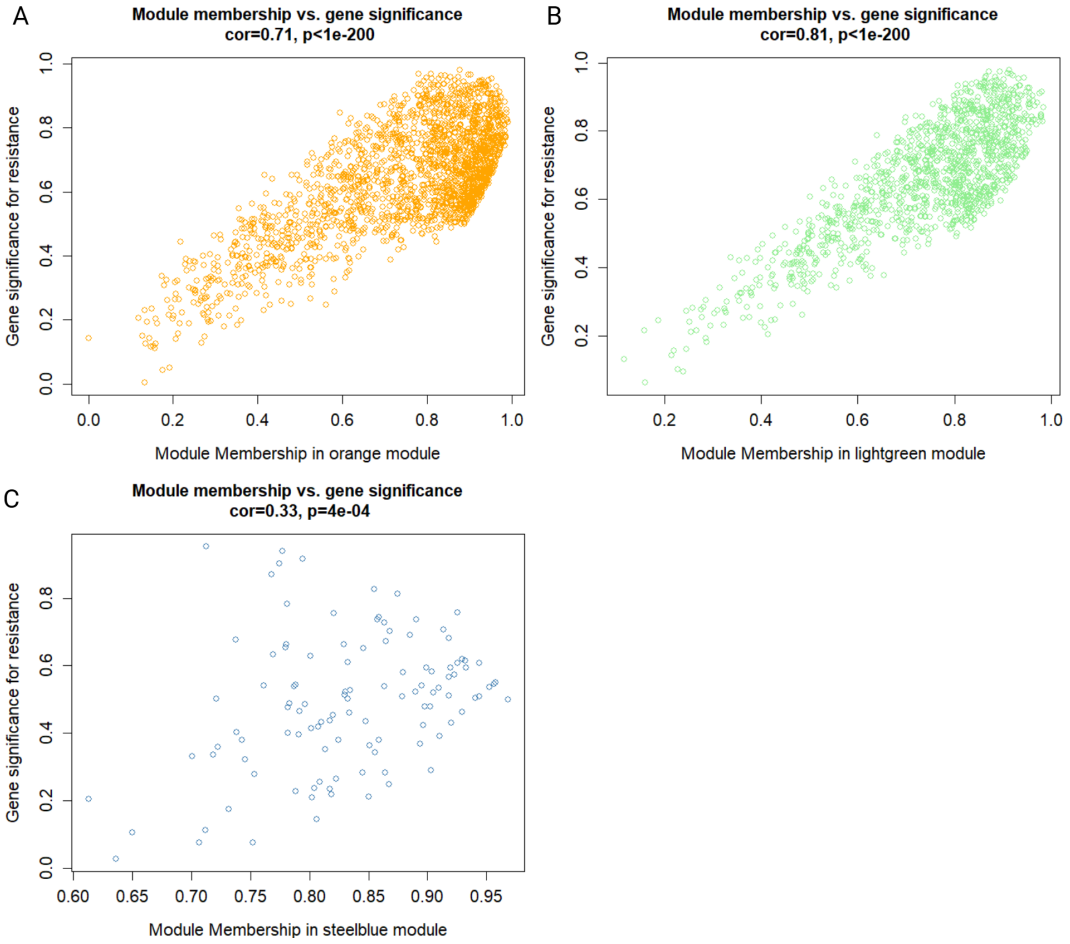


FIGURE 4.31: Gene significance and module membership plots of modules correlated with FP-acquired resistance. A) 'orange' module B) 'lightgreen' module C) 'steelblue' module. Each point represents a gene in the modules. The Pearson correlation and Student asymptotic p-value for correlation for gene significance and module membership is shown.

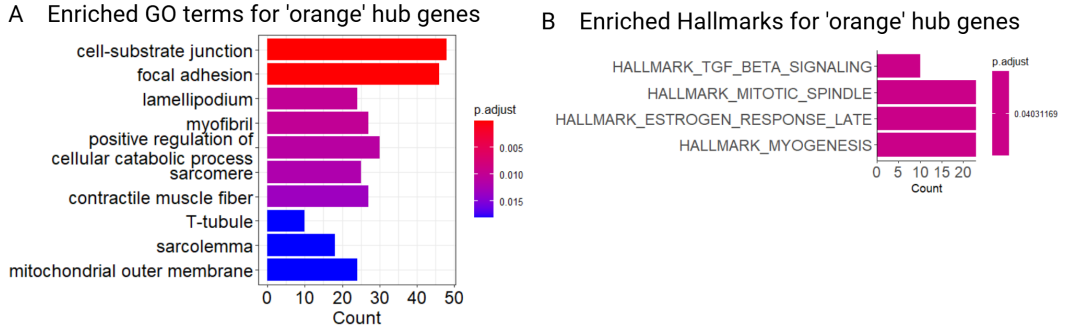


FIGURE 4.32: Enriched gene sets for orange hub genes (gene significance and module membership >0.6). A) Enriched GO terms B) enriched Hallmarks. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). If more than 10 terms were enriched, the ten terms with the smallest adjusted p values were displayed.

Figure 34B) were cross-referenced with WGCNA hub genes, resulting in 25 genes for the FP-acquired resistance signature (Table 4.4).

TABLE 4.4: FP-acquired resistance signature genes. Identified from overlapping PPI hub proteins (centrality score ≥8) and WGCNA hub genes (GS and MM >0.6) from the 'orange' and 'lightgreen' resistant modules. GS= gene significance, MM= module membership. LFC= log-2 fold change.

Gene	Module	GS	MM	PPI score	LFC VCR-res vs ctrl	P.adjust	LFC IFO-res vs ctrl	P.adjust
RPS13	lightgreen	0.86	0.86	24.00	0.37	0.00	0.31	0.00
EIF3G	lightgreen	0.84	0.70	8.00	0.16	0.02	0.16	0.02
RPL27A	lightgreen	0.81	0.78	20.00	0.21	0.00	0.20	0.00
RPL6	lightgreen	0.79	0.83	20.00	0.21	0.00	0.12	0.04
RPLP2	lightgreen	0.78	0.93	20.00	0.53	0.00	0.23	0.00
RPL31	lightgreen	0.71	0.72	20.00	0.27	0.00	0.24	0.00
RPL29	lightgreen	0.70	0.90	20.00	0.41	0.00	0.12	0.05
RPS20	lightgreen	0.67	0.80	24.00	0.21	0.00	0.11	0.06
EIF3L	lightgreen	0.66	0.70	8.00	0.24	0.01	0.14	0.10
RPL8	lightgreen	0.65	0.89	20.00	0.61	0.00	0.13	0.03
RPS28	lightgreen	0.65	0.69	24.00	0.16	0.08	0.10	0.28
MAPK1	orange	0.92	0.80	12.00	0.18	0.00	0.19	0.00
RPL10	orange	0.91	0.90	10.00	0.33	0.00	0.47	0.00
CDK1	orange	0.84	0.89	14.00	0.20	0.02	0.31	0.00
IKBKG	orange	0.82	0.94	10.00	0.23	0.00	0.45	0.00
GRB2	orange	0.74	0.92	14.00	0.11	0.03	0.27	0.00
XIAP	orange	0.74	0.69	10.00	0.15	0.06	0.17	0.02
ATF2	orange	0.73	0.64	10.00	0.16	0.05	0.16	0.04
USP14	orange	0.72	0.86	8.00	0.12	0.06	0.31	0.00
FAU	orange	0.71	0.91	14.00	0.11	0.12	0.28	0.00
EIF1	orange	0.67	0.94	8.00	0.13	0.06	0.55	0.00
EIF2S3	orange	0.66	0.95	8.00	0.17	0.00	0.82	0.00
MAP3K7	orange	0.66	0.94	10.00	0.11	0.06	0.51	0.00
BECN1	orange	0.61	0.73	8.00	0.06	0.31	0.12	0.02
RPS5	orange	0.60	0.79	16.00	0.15	0.14	0.40	0.00

4.3.2.10 Scoring samples based on the expression of genes in the FP-acquired resistance signature

GSVA was used to assign a value to the treatment condition based on the expression of the 25 gene FP-acquired resistance signature. Scores ranged from -0.8 to 0.5 (Figure 4.33). A Shapiro-Wilk normality test was performed which showed that the data was not normally distributed. A non-parametric Kruskal-Wallis was performed to determine if there are statistically significant differences between two or more groups. This was significant, so pairwise Wilcoxon Rank Sum Tests were performed between treatment conditions for each of the datasets.

For the acquired resistance data, there was a significant difference between the RH4 control samples and the chemotherapy-resistant samples (both VCR-resistant and IFO-resistant) (Figure 4.33). This is reassuring as it indicated the FP-acquired resistance signature worked well at distinguishing between resistant and non-resistant samples in the dataset from which it was derived. When looking at the RMSYM cell line in the acquired resistance dataset, there was a significant difference between the control and VCR-resistant samples but the VCR-resistant samples had a lower score. There was no significant difference between the control and IFO-resistant RMSYM samples. This suggests the signature is not able to distinguish between resistant and non-resistant samples in the RMSYM ERMS cell line.

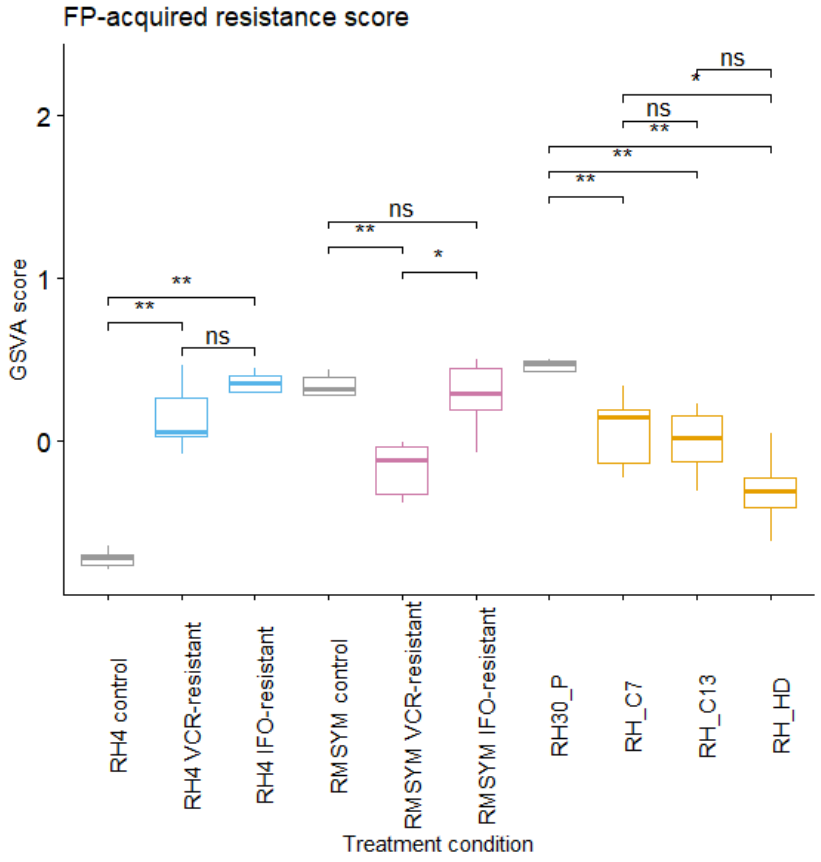


FIGURE 4.33: FP-acquired resistance score using GSVA. Pairwise comparisons were made using the Wilcoxon rank sum exact test. P values were adjusted using Benjamini-Hochberg method. ns: p >0.05, *: p ≤0.05, **: p ≤0.01, ***: p ≤0.001, ****: p ≤0.0001.

When looking at the FP-acquired resistance scores in the intrinsic resistance dataset, the resistant samples score significantly lower than the parental sample. This could be explained by the resistant clones being treatment-naive. Whilst the HD sample has been treated with chemotherapy, the method by which the HD samples were generated were very different from the VCR and IFO-resistant samples which might explain the difference in expression of the genes in the gene signature.

4.3.2.11 Investigating modules negatively correlated with fusion-positive acquired resistance

Two modules had a significant negative correlation with FP-acquired resistance in WGCNA. These were 'darkmagenta' and 'darkolivegreen' (Supplementary Figure 35). The 'darkmagenta' module was seen to be enriched for ribosomal-related gene sets and the 'darkolivegreen' module enriched for mitochondrial-related gene sets (Figure 4.34A and C). Both modules were enriched for genes involved in oxidative phosphorylation (Figure 4.34B and A).

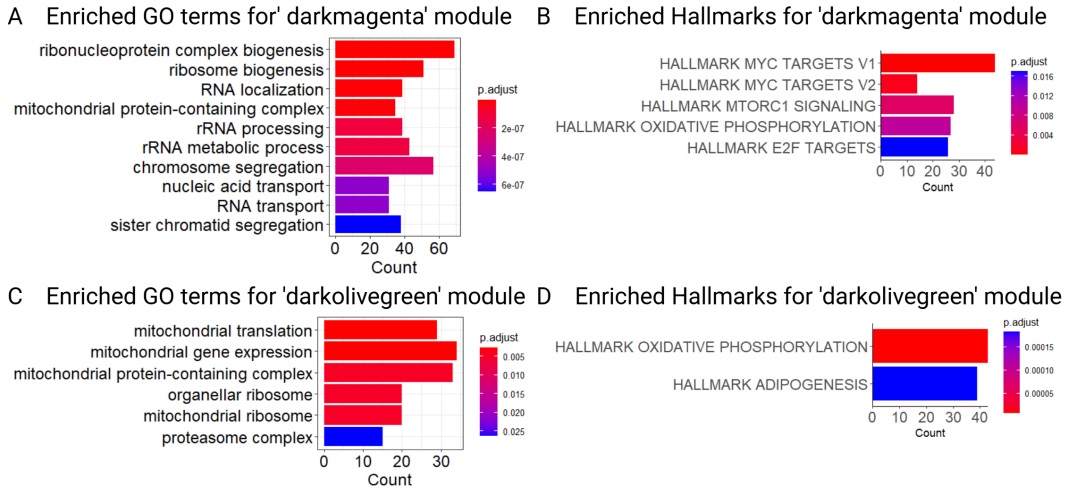


FIGURE 4.34: Enriched GO terms for hub genes from modules with a significant negative correlation with FP-acquired resistance. A) GO terms for 'darkmagenta' module B) Hallmarks for 'darkmagenta' module. Hub genes had a gene significance and module membership score >0.6. P values were adjusted using Benjamini-Hochberg method. Significant terms were defined as p adjust <0.05. Top ten GO terms with the smallest adjusted p values are shown.

4.3.2.12 Identification of modules associated with acquired resistance in fusion-negative rhabdomyosarcoma using weighted gene co-expression network analysis

The data had to be reprocessed to remove genes that were expressed in RH4 samples but not RMSYM. After filtering to remove genes with low expression, 15,347 genes remained. No outliers were identified in the dendrogram (Figure 4.35A). A soft threshold power of 18 was selected (Figure 4.35B). There were 77 modules before merging and 20 modules after merging similar modules (Figure 4.35C and D).

All modules information can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/RMSYM/geneInfo.csv). The average module size was 767 genes. The 'green' module had the largest number of genes, consisting of around 2700 genes, and 'grey' had the smallest (Supplementary Figure 36).

Five modules were significantly positively correlated with resistance, with 'brown4' module having the strongest correlation, followed by 'green', 'darkolivegreen', 'darkorange2' and 'lightcoral' (Supplementary Figure 37). The 'brown4', 'lightcoral' and 'darkorange2' module showed high expression in both VCR and IFO-resistant samples, but the expression was less consistent between replicates in the 'lightcoral' and 'darkorange2' modules (Figure 4.36). The 'green' module showed high expression in IFO-resistant samples while the 'darkolivegreen' showed high expression in VCR-resistant samples (Figure 4.36).

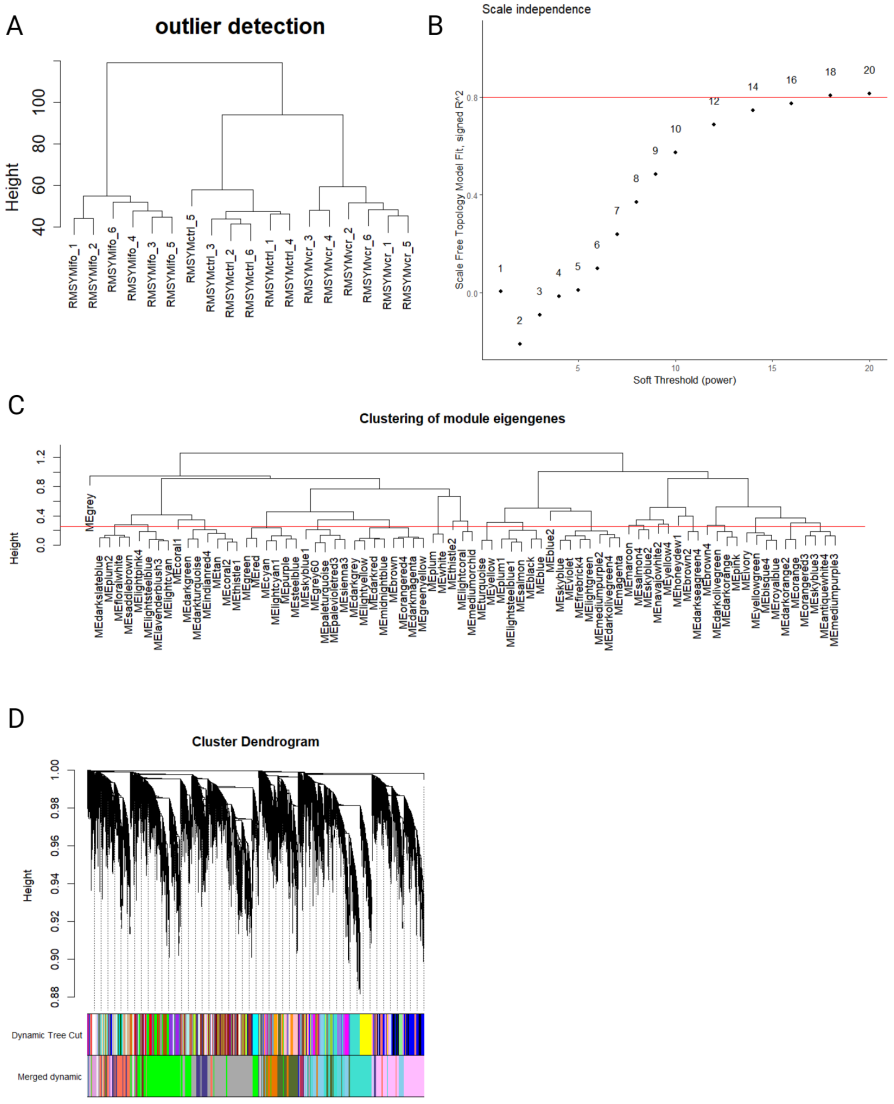


FIGURE 4.35: Outlier detection, selection of optimal soft threshold power and visualisation of modules for the gene co-expression network for WGCNA using FN-acquired resistance samples. A) dendrogram of samples B) scale independence plot for selection of optimal soft threshold power C) dendrogram of modules with red line showing dissimilarity threshold of 0.25 D) Cluster dendrogram of modules showing module colour before and after merging.

The relationship between the module membership and gene significance for these modules was assessed. 'Darkolivegreen', 'green', 'lightcoral' and 'brown4' all showed significant strong correlations between module membership and gene significance values (Supplementary Figure 38A, B, C and D). 'Darkorange2' showed a significant but weak correlation, so this module was not included for further analysis (Supplementary Figure 38E).

The hub genes (genes with a gene significance and module membership >0.6) were selected from each of the four resistant modules. The two largest modules, 'green' and 'darkolivegreen', also had the largest number of hub genes (Table 4.5). The 'lightcoral' and 'brown4' modules had a similar number of genes in the module, but 'brown4' had

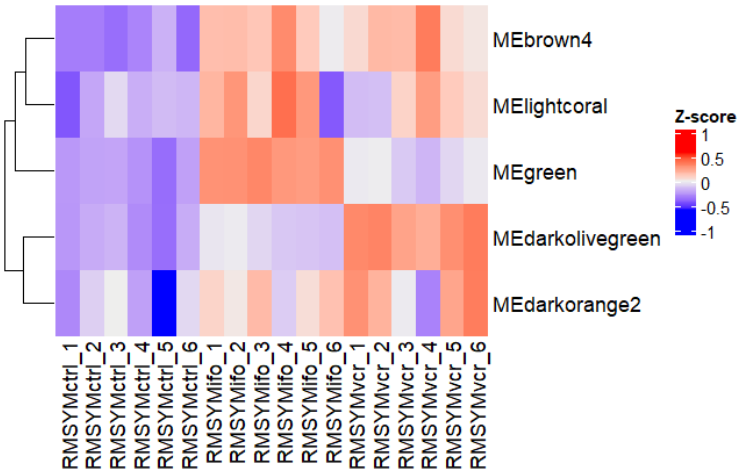


FIGURE 4.36: Heatmap showing sample contribution to modules with a significant positive correlation with FN-acquired resistance.

TABLE 4.5: Number of genes and proteins in the four FN-acquired resistant modules. Resistant modules were significantly positively correlated with the binary trait resistance. WGCNA module hub genes were defined as having a module membership and gene significance value >0.6. hub genes were defined as having a centrality score ≥8.

Resistant modules	Number of genes in the module	Number of hub genes	Number of proteins in PPI network	Number of PPI hub proteins	Number of overlapping PPI hub proteins and WGCNA hub genes
Brown4	93	82	0	0	0
Lightcoral	98	14	2	0	0
Darkolive green	1314	426	95	5	2
Green	2769	947	368	76	24

more hub genes as it had a stronger gene significance and module membership correlation (Table 4.5).

4.3.2.13 Gene set enrichment analysis on fusion-negative module hub genes

Resistant modules were enriched for terms related to cell division including 'chromosome segregation', 'mitotic nuclear division' and 'replication fork processing' (Supplementary Figure 39, Supplementary Figure 40 and Supplementary Figure 41). Again, the Hallmark 'mitotic spindle' was enriched for in the FN-acquired resistant modules (Supplementary Figure 39 and Supplementary Figure 41). The 'brown4' module had no enriched gene sets.

TABLE 4.6: FN-acquired resistance signature genes. Identified from overlapping PPI hub proteins (centrality score ≥8) and WGCNA hub genes (GS and MM >0.6) from the 'green' and 'darkolivegreen' resistant modules. GS= gene significance, MM= module membership.

Gene	Module	GS	MM	PPI score	LFC VCR-res vs ctrl	P.adjust	LFC IFO-res vs ctrl	P.adjust
JUN	darkolivegreen	0.74	0.76	10.00	0.72	0.00	0.38	0.00
BLM	darkolivegreen	0.61	0.88	8.00	0.45	0.00	0.13	0.07
VIM	green	0.94	0.77	8.00	0.54	0.00	0.60	0.00
CWF19L2	green	0.90	0.72	10.00	0.36	0.00	0.38	0.00
MRPL27	green	0.88	0.87	38.00	0.29	0.00	0.39	0.00
RNF2	green	0.87	0.92	8.00	0.38	0.00	0.62	0.00
RPS15	green	0.87	0.71	12.00	0.31	0.00	0.34	0.00
MALSU1	green	0.86	0.91	28.00	0.40	0.00	0.67	0.00
TRAF2	green	0.75	0.70	10.00	0.24	0.00	0.28	0.00
XAB2	green	0.72	0.92	20.00	0.13	0.00	0.33	0.00
TMCO1	green	0.71	0.86	8.00	0.13	0.01	0.25	0.00
PSMD6	green	0.69	0.84	8.00	0.22	0.01	0.38	0.00
PPIE	green	0.69	0.68	16.00	0.18	0.01	0.27	0.00
RBM22	green	0.68	0.97	18.00	0.22	0.00	0.71	0.00
CHCHD1	green	0.68	0.83	36.00	0.12	0.10	0.24	0.00
CWC15	green	0.67	0.92	16.00	0.15	0.03	0.40	0.00
RBL1	green	0.66	0.93	8.00	0.16	0.03	0.43	0.00
PPP2CA	green	0.66	0.98	16.00	0.32	0.00	1.43	0.00
CDC20	green	0.66	0.90	20.00	0.20	0.01	0.47	0.00
PCNA	green	0.64	0.73	24.00	0.28	0.01	0.44	0.00
HSD17B12	green	0.63	0.98	8.00	0.26	0.00	1.30	0.00
G3BP1	green	0.63	0.95	10.00	0.18	0.02	0.71	0.00
EFTUD2	green	0.61	0.72	32.00	0.14	0.02	0.20	0.00
CDC16	green	0.60	0.90	8.00	0.07	0.16	0.31	0.00
MRPS31	green	0.60	0.92	36.00	0.13	0.13	0.42	0.00
MRPL22	green	0.60	0.97	40.00	0.13	0.04	0.85	0.00

4.3.2.14 Selection of fusion-negative acquired resistance signature genes from resistant modules using protein-protein interaction networks

PPI networks could not be generated using genes from the 'brown4' and 'lightcoral' modules as there was no evidence for gene interactions in the 'brown4' module, and only two interacting genes in the 'lightcoral' module (Table 4.5). This is likely due to the small number of genes in these modules. All interactive PPI networks can be found in the Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/RMSYM/PPI). Overlapping hub proteins from the PPI network and WGCNA hub genes were selected as components of the resistance gene signature. The FN-acquired resistant signature consisted of 26 genes in total including 2 from the 'darkolivegreen' module and 21 from the 'green' module (Table 4.6).

4.3.2.15 Scoring samples based on the expression of genes in the FN-acquired resistance signature

GSVA was used to assign a score to the treatment conditions based on the expression of the 26 gene FN-acquired resistance signature. Scores ranged from -0.7 to 0.7 (Figure 4.37). The Shaprio-Wilk test indicated the data was not normally distributed as p value <0.05.

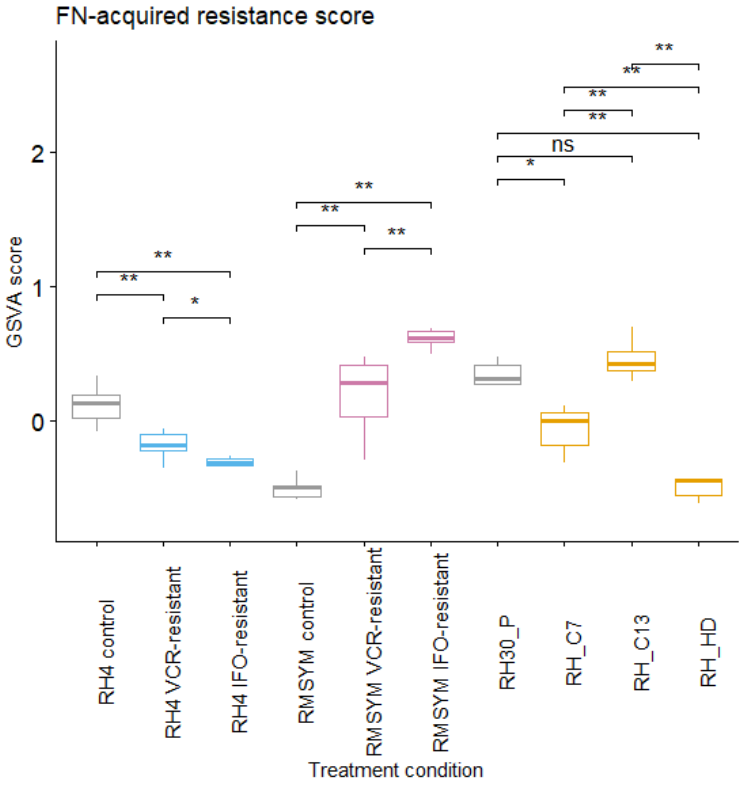


FIGURE 4.37: FN-acquired resistance score using GSVA. Pairwise comparisons were made using the Wilcoxon rank sum exact test. P values were adjusted using Benjamini-Hochberg method. ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$

For the acquired resistance dataset, the chemotherapy resistant FN samples had a significantly higher score than the control (Figure 4.37). This indicates the gene signature is able to distinguish between resistant and non-resistant samples in the dataset from which it was derived. However, in the FP samples, the resistant conditions had a significantly lower score than the control suggesting the signature may be unable to distinguish between resistant and non-resistant samples in the FP cell line. Additionally, the C7 and HD samples from the intrinsic resistance dataset had significantly lower scores compared to the P cell line, and C13 showed no significant difference. It may be expected that the signature does not work well in this dataset as the intrinsic resistance model used an ARMS cell line.

4.3.2.16 Investigating modules negatively correlated with fusion-negative acquired resistance

There were 5 modules that showed a significant negative correlation with FN-acquired resistance; 'coral1', 'yellow4', 'plum1', 'skyblue' and 'darkslateblue' (Supplementary Figure 37). All GSEA can be found in Supplementary Files (Supplementary Files/Chapter 4 Rhabdomyosarcoma cell models/Acquired resistance/Neg modules/GSEA). Similar to the modules negatively correlated with resistance in the FP and intrinsic data, the 'skyblue' module showed enrichment of ribosomal related gene sets (Figure 4.38).

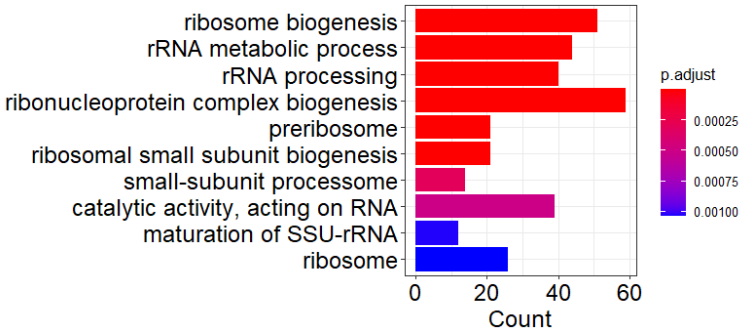


FIGURE 4.38: Enriched GO terms for the 'skyblue' module that had a significant negative correlation with intrinsic resistance. P values were adjusted using Benjamini-Hochberg method. Significant terms were defined as p adjust <0.05. Top ten GO terms with the smallest adjusted p values are shown.

4.4 Discussion

This work aimed to identify gene signatures of resistance in RMS using RNA-sequencing data from models of intrinsic and acquired resistance. An 18 gene signature of intrinsic resistance was identified as well as two acquired resistance signatures for each RMS subtype (FP and FN). Identification of these three distinct signatures of resistance is consistent with the expectation that mechanisms of resistance may vary depending on the type of resistance, subtype of RMS and chemotherapy treatment. This work will be externally validated in Chapter 5 using publicly available patient samples. These gene signatures may have potential clinical utility in distinguishing tumour relapse or recurrence in patient samples. An association with other clinical variables may also be evident, including overall survival and event-free survival, as these factors are lower in patients with relapse/recurrence and thus the signatures may also correlate with these.

Although it was hypothesised that the intrinsically resistance clones may be similar, results from DGE analysis suggested they show differences in gene expression. Another study found that subclones of the RH30 cell line differed in terms of adhesion molecules and myogenic transcription factors [319], suggesting that clones show different mechanisms of intrinsic resistance. GSEA results could provide some insight into possible different mechanisms of intrinsic resistance in C7 and C13. For example, C13 showed downregulation of the death receptor apoptotic signalling pathway. Aspects of this pathway such as the death receptor ligand TRAIL have previously been targeted in RMS however their effectiveness was limited due to resistance to TRAIL inhibitors [330]. C7 showed upregulation of ECM-related terms. There has been some evidence that RMS shows upregulation of collagen and collagen-modifying enzymes to modulate the ECM and could be implicated in tumour growth and metastasis [331]. There were some common themes in the GSEA results between the clones, mainly downregulation of terms related to the regulation of microtubule-associated protein TAU and renal-related terms.

WGCNA identified two modules significantly correlated with resistance for the intrinsic dataset. Possible mechanisms of intrinsic resistance were investigated through GSEA of these modules. There was enrichment of cadherin binding, proteins involved in cell-cell adhesion and organization of the actin cytoskeleton. Loss of epithelial-cadherin (E-cadherin), which prevents cells from dissociating from one another, is linked to EMT. The PAX3-FOXO1 protein has been found to cause EMT through the overexpression of the transcription factor SNAIL, resulting in repression of E-cadherin [332]. The enrichment of 'cadherin binding' in the resistant module could be a reflection of this. The intrinsic resistance signature consisted of genes involved in DNA-damage response, ubiquitin-related proteins, nonsense-mediated decay and mRNA splicing.

For the acquired resistance dataset, the data was separated into FP and FN cell lines as no modules were significantly correlated with resistance in the combined data. This could possibly be due to different mechanisms of resistance in FP and FN RMS, which has previously been seen in the literature. Patel et al. reported differences in the mechanisms of treatment resistance between ERMS and ARMS [156]. In addition, Seitz et al. showed there were differences in the expression of the multi-drug resistance protein ABCB1 between ERMS and ARMS tumours [333].

Interestingly there was some overlap in GSEA results from the FP-acquired and intrinsic resistant modules. Both intrinsic and acquired resistant FP modules also showed enrichment in RHO GTPases, part of the Ras superfamily of GTPases that are involved in actin dynamics and have been linked to chemo-resistance in many cancers [334]. RHO GTPases have been found to promote tumour propagating cells implicated in relapse in RMS [335]. Given that the mode of action of VCR is to bind to microtubules and disrupt the formation of the mitotic spindle, the regulation of actin by RHO GTPases could be a possible mechanism of intrinsic resistance. Multiple studies have found evidence for members of the RHO GTPase family to be involved in RMS progression and metastasis [336, 337, 335, 338].

The FP-acquired resistance signature consisted of many ribosomal proteins although there is minimal research into ribosomal proteins and RMS. Several genes were present in both the intrinsic and acquired FP resistance signatures, such as *MAPK1* and *XIAP*, which again could highlight the importance of RMS subtype in resistance. MAPK1 is a component of the MAP kinase signal transduction pathway, where inhibition of MAPK1 and PI3K in RMS was found to inhibit growth *in vitro* and *in vivo* [339]. XIAP has been shown to inhibit tumour cell apoptosis and has previously been used as a reporter to examine the effect of drugs on human RMS cells [340]. Research investigating the role of XIAP in paediatric cancers found that downregulation of XIAP with antisense oligonucleotides sensitised osteosarcoma cells to doxorubicin, etoposide and vincristine, suggesting it may play a role in chemotherapy resistance [341]. Ribosomal-related gene sets were also enriched for in modules negatively correlated with resistance in all models. This could also support the findings of ribosomal genes playing a role in resistance in RMS.

GSEA of FN-acquired resistant samples suggested more cells were more differentiated. Whereas Patel et al. found that populations of immature mesoderm cells were responsible for chemotherapy resistance in ERMS [156], perhaps populations of more differentiated cells may also contribute. Future work would look to compare the differentiation status of the resistant and non-resistant samples and deconvolute these resistant datasets using single-cell RNA-seq data. For the FN-acquired resistance signature, many of the genes encode for ribosomal proteins or proteins involved in mRNA splicing. Other genes were cell cycle genes. No evidence

could be found in the literature for the involvement of these genes in RMS chemotherapy resistance.

All resistant modules in both FP and FN showed enrichment of the gene set 'mitotic spindle'. Perhaps this is a common mechanism of resistance in RMS. Unfortunately for the acquired resistance data it is not determined whether the genes from the module are from the VCR-resistant or IFO-resistant samples. Since VCR disrupts the formation of the mitotic spindle, perhaps higher expression of these mitotic spindle genes could contribute to cells being more resistant to VCR. Recently a study identified a group of genes related to mitotic spindle organisation that are were essential to response to vincristine treatment in diffuse large B-cell lymphoma [342]. The role of mitotic spindle genes in chemotherapy resistance in RMS could be explored by looking at the expression of these genes at a protein level. Another explanation for this increase in expression of mitotic spindle genes in resistant samples may be due to an increase in cell proliferation. Whilst RMSYM IFO-resistant and VCR-resistant samples showed an increased in a cell proliferation gene signature compared to control, C7, C13 and RH4 resistant samples did not, suggesting in there is no increase in cell proliferation in these samples.

FN VCR-resistant cells from the acquired resistance dataset and FP HD from the intrinsic resistance dataset both showed upregulation of MDR genes *ABCB1* and *ABCB4*. ABCB1 has previously been associated with chemotherapy resistance in RMS [343] and an ABCB1 inhibitor was found to sensitise RH30 VCR-resistant cells to VCR [344]. *GLI1* expression, which has previously been found to be upregulated in VCR-resistant RMS [120], was not significantly differentially expressed in any VCR-resistant treatment conditions. Although Yoon et al. found upregulation of *GLI1* in RH30 VCR-resistant cell lines, VCR-resistant cell lines were established using a different method. The authors created VCR-resistance by exposing cells to serially increasing concentrations of VCR as opposed to the RH30 VCR-resistant samples in this work, which were treatment-naive intrinsic resistant clones or HD.

A major limitation of using WGCNA in this experimental design is the low number of biological replicates and high number of technical replicates. This may have resulted in the identification of modules driven by consistent expression patterns across technical replicates rather than robust biological co-expression. Consequently, the use of this approach to derive a gene signature of resistance is limited, as identified genes may reflect technical variation rather than underlying biological resistance mechanisms. Therefore, WGCNA results should be interpreted with caution and considered alongside complementary analyses, including DGE and GSEA.

The genes from the resistant WGCNA modules were able to create a PPI networks with hub-like structures, in contrast to PPI networks created from DEGs or only resistant hub genes. Using PPI to select for signature genes may induce bias in

selecting genes for which more information is known of their protein interactions, and other less known genes may be overlooked. It may also bias towards genes that their proteins are involved in pathways with a larger interaction network even if this may not be relevant to the trait. For example, the 'green' module from the FN-acquired resistance PPI showed a tight interaction network between mitochondrial ribosomal proteins, giving these proteins high hub scores. It is unknown whether these mitochondrial ribosomal proteins are indeed important hub proteins in resistance or if they have only been selected as they are part of a larger tight interaction network. The cross-referencing of PPI hub proteins with WGCNA hub genes hopefully limits the effect of this, as genes with a low gene significance and module membership score will be removed and genes important in resistance will be retained. Another limitation of this study using WGCNA and PPI for selection of a gene signature instead of a method like LASSO with multi-cox regression is that only correlation with a binary trait is measured and this may not translate to clinical feature such as relapse/survival.

WGCNA and integration with a PPI network identified gene signatures of resistant. External validation of this signature will be carried out in Chapter 5. The HD showed extensive variation in gene expression compared to the intrinsic resistant clones. This could suggest that treatment with drug significantly changes the gene expression profile from the intrinsic resistant clones. It may be that the intrinsic resistant signature may be masked and difficult to detect in samples treated with drug, however this will be further investigated in the external validation. Ideally to test the intrinsic resistant signature, it would useful to look at pre-treatment samples of tumours that relapse vs tumours that don't. Currently, there are no gene signatures that are able to predict relapse/recurrence in RMS. It is hoped that these gene signatures may be able to distinguish between relapse/recurrence in patient samples and provide novel insights into potential therapeutic strategies. This work also provided some insight into possible mechanisms of chemotherapy resistance in RMS which currently lacks understanding and highlighted the importance of RMS subtype in resistance.

Chapter 5

Validation of chemotherapy resistance signature in rhabdomyosarcoma using publicly available patient data

5.1 Introduction

RMS patients show variation in their response to chemotherapy, with some patients achieving long term failure-free survival (FFS) whilst others relapse during treatment [107]. Several factors have been recognised that influence the risk of relapse. Alveolar histology was found to be associated with an increased risk of relapse [345], although this likely reflects FP status. Non-alveolar tumours that arise from the orbit and genitourinary system have been found to have the lowest risk of relapse [346] while patients with metastatic disease, unfavorable primary tumour site and tumour stage 4 have a higher risk of relapse [107]. In addition to these clinical factors identified, it is likely that the cells responsible for relapse show differences in gene expression and this may be a result of intrinsic or acquired resistance. Identifying these key changes in gene expression of resistant cells might be useful for predicting relapse and finding therapeutic targets.

Gene signatures have been used for predicting resistance/relapse in many cancers [347, 183, 348, 349]. After gene signatures have been derived, they are validated in an external dataset, often historical microarray data with clinical annotation. Microarray is a technology that can be used to detect gene expression through cDNA. RNA is extracted from samples and converted into cDNA by reverse transcriptase, as this is more stable than mRNA [350] (Figure 5.1). Fluorescent markers are added to the

cDNA and the cDNA is added to the microarray chip. The microarray chip contains single stranded DNA bound to the slide, called probes. These probes match particular mRNA transcripts based on NCBI sequences [351]. Fluorescently labelled cDNA that hybridises to complementary probes on the chip indicates that the gene is expressed and those that don't bind are washed off. The microarray is scanned by a laser to detect the intensity of the fluorescently-labelled cDNA, thus intensity values of the probes correlate with gene expression.

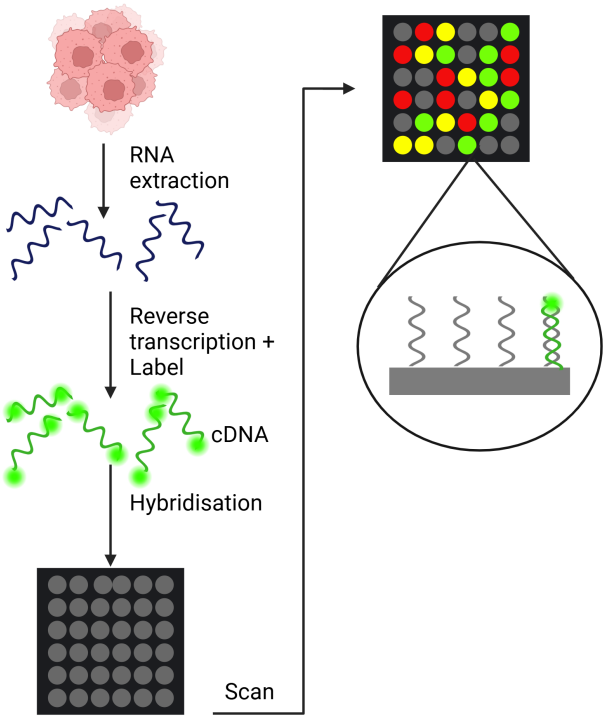


FIGURE 5.1: Microarray workflow. RNA is extracted from samples and converted to cDNA by reverse transcriptase. Fluorescent labels are added to the cDNA. Genes that are expressed are hybridised to complimentary probes on the microarray chip and their intensity is measured by a laser. Created with BioRender.com.

Resistance to chemotherapy is a considerable challenge in RMS. Most patients initially respond to chemotherapy [105], but relapse rates are high [107] and post-relapse survival is dismal [13, 14]. No gene expression signature currently exists that identifies patients at diagnosis who are likely to experience resistance/relapse. Chapter 4 aimed to derive gene signatures of resistance from cell models. WGCNA was used to find modules of co-expressed genes in the resistant samples. From these modules, important (hub) genes were selected and cross-referenced with hub genes from PPI networks, resulting in a short list of genes for the signatures. Three gene signatures representing intrinsic, FP-acquired and FN-acquired resistance were identified. In this chapter, the gene signatures identified in Chapter 4 are validated using publicly available patient microarray data. Since the signatures were derived from cell models, it is important to investigate if they are relevant in patients. Using two clinically well-annotated microarray datasets [352, 81], patient samples were scored based on

the expression of the previously identified gene signatures. Differences in scores between clinical traits were statistically tested. It was hypothesised that scores will be significantly higher in patients with worse clinical outcomes. In addition to the patient microarray data, signatures were also tested for in pre- and mid-treatment patient RNA-seq data. This is more similar to the cell models of acquired resistance, as it is comparing chemotherapy-treated samples with pre-treatment samples.

5.2 Materials and methods

5.2.1 Datasets

Three publicly available patient datasets were used for the validation of the RMS chemotherapy resistance signature. Two of these were microarray datasets from samples before chemotherapy treatment, which were used to test if the gene signatures were associated with any clinical traits which may indicate an association with resistance. The first dataset (here on referred to as the Triche) consisted of 186 RMS patient samples processed on an Affymetrix GeneChip Human U133A Expression Array [352] (accession GSE92689). Clinical annotation for this dataset included; sex; age; disease stage; grade; tumour location; histology; translocation status; overall survival time; event; FFS; Risk Group (RG) ; size; lymph node involvement and invasion. The second dataset (Williamson), consisted of 101 patient samples from Affymetrix GeneChip expression analysis performed using the HGU133plus2 array [81] (accession E-TABM-1202). Clinical annotation for this dataset included; tumour location, sex, histology, metastasis at diagnosis, IRS group, SIOP stage, tumour size, chromosomal translocation status,RG, age, complete remission achieved, first event occurred, first event type, survival time and lymph node involvement.

The third dataset was bulk RNA-sequencing data from patients [318] (accession GSE240287). Samples were from 9 matched patient biopsies and mid-treatment delayed resections (18 samples in total). This dataset was used to address whether gene signatures scores were higher in the mid-treatment resection compared the biopsies, as this was more similar to how the cell models were generated.

5.2.2 Microarray analysis

The raw data (.CEL files) and metadata were imported into R using the GEOquery package (version 2.66.0) [353] for the Triche dataset and the ArrayExpress package [354] for the Williamson dataset. Quality control of the raw data was assessed using histograms and boxplots of the log2 transformed data. Data was processed using the

RMA package [355] for background correction, quantile normalisation and summarisation. RMA was selected as the package for processing the microarray data as it is the most widely used method for microarray analysis due to due to it's ability to generate expression values that are precise and have low noise [356, 357]. Microarray data contains background noise from probes that produce signal unrelated to gene-specific cDNA hybridization [358]. The background correction step in the processing of microarray data aims to remove this background noise. Quantile normalisation is performed to minimise the effect of technical variability such as target preparation and hybridization. Since there are multiple probes targeting one gene, the summarisation step is required to summarise multiple probe intensities into a single value per probeset. RMA does this using a median-polish algorithm. Probesets may map to multiple genes so Jetset (version 3.4.0) was used to select one probe set for each gene [359]. Jetset selects one probe per gene based on an overall score which takes into account specificity of the probe to the gene, splice isoform coverage and robustness against transcript degradation. Genes with missing values or that were duplicated were removed. Samples that were not ERMS or ARMS were removed.

5.2.3 RNA sequencing analysis

The raw count matrix and clinical data was obtained from GEO and gene symbols mapped to Ensembl IDs. Data was analysed as described in Chapter 2. A design matrix was created comparing mid to pre-treatment samples, while adjusting for inter-patient variability. Data was normalised using the TMM method and PCA used to identify outliers. Significant DEGs were defined as having a LFC >1.5 or <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg).

5.2.4 Testing for association of the gene signatures with clinical traits

As in Chapter 4, samples were assigned a score based on the expression of the genes in the signature using the GSVA package [239]. For the intrinsic and FP-acquired gene signatures, only FP samples were used. For the FN-acquired gene signature, only FN samples were used.

In addition to the three gene signatures, GSVA scores were also calculated based on DEGs from the cell models, to determine whether DEGs may be a better alternative for the gene signatures. This was done for intrinsic DEGs (overlapping significant upregulated DEGs from C7 and C13), FP-acquired DEGs (overlapping significant upregulated DEGs from RH4 IFO-resistant and RH4 VCR-resistant) and FN-acquired (overlapping significant upregulated DEGs from RMSYM IFO-resistant and RMSYM VCR-resistant). As in Chapter 4, significant upregulated DEGs were defined as having a LFC >1 and adjusted p value <0.05 (Benjamini-Hochberg). Again, GSVA scores

were calculated only in the corresponding RMS subtype (e.g. for FN-acquired DEGs only FN samples were used).

To test if the GSVA scores were significantly different between levels of a clinical trait, the Shapiro-Wilk was used to test for normality. Either a T test or Wilcox Rank Sum test was used to calculate pairwise comparisons between levels with corrections for multiple testing. For example, for lymph node status, comparisons were made between GSVA scores of N-0 vs N-1. Since this research wanted to test for significant differences in GSVA scores and clinical outcome, only clinical traits relevant to clinical outcome were tested for. For the Triche dataset, this included disease stage, grade, survival time, event, FFS, RG, size; lymph node involvement and invasion. For the Williamson dataset, this included metastasis at diagnosis, IRS stage, tumour size, risk group, complete remission achieved, first event occurred, survival time and lymph node involvement. P values were adjusted using the Benjamini-Hochberg method. Differences in GSVA gene signature scores were considered statistically significant at an adjusted p value (Benjamini-Hochberg) <0.05.

5.3 Results

5.3.1 Analysis of the Triche microarray dataset

5.3.1.1 Quality control and processing of the raw microarray data

The raw microarray data contained 22,283 probe features and 186 samples. Intensity values for each sample were assessed after background correction, normalisation and summarisation.

Boxplots show the greatest variation in intensity between samples in the raw data (Supplementary Figure 42A). After background correction, the variation in intensity between samples was reduced but some variation remained (Supplementary Figure 42B). Normalisation further reduced the variation so that all samples showed similar distributions of intensity (Supplementary Figure 42C) which is maintained after summarisation (Supplementary Figure 42D).

Density plots were used to examine the distribution of intensities across individual samples in more detail. In the raw data, distributions differ between samples (Figure 5.2A). After background correction, samples show a more similar distribution except for one sample which was flagged as a potential outlier (Figure 5.2B). Normalisation results in all samples having a very similar distribution, including the outlier (Figure 5.2C). After summarisation, samples show slightly more variation but the overall distributions are similar (Figure 5.2D).

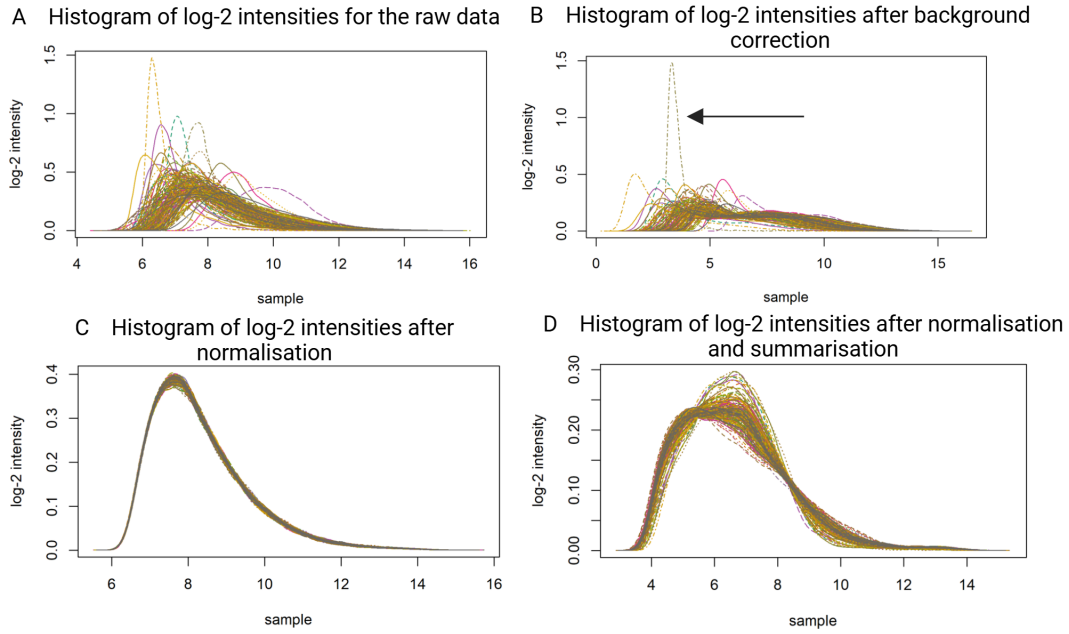


FIGURE 5.2: Histograms of intensity distributions for the Triche dataset.A) Raw data B) After background correction with RMA with potential outlier highlighted by the arrow C) After normalisation D) After summarisation. Intensity values are log-2 normalised.

After one to one mapping of genes to probeset using Jetset, there were 12,180 genes. Three genes with missing values and seven duplicates were removed leaving 12,170 genes for analysis.

5.3.1.2 Characteristics of the clinical data

The clinical data was assessed for missingness. 13 samples with incomplete clinical annotation were removed. Samples that were not classified as ARMS or ERMS (e.g. botryoid, other) were excluded (48 samples). In total, 125 samples were included.

Age ranged from 0-20 years, which is typical of RMS [15], with the most common ages being 2 and 5 years of age (Figure 5.3A). The most common tumour grade was IV closely followed by III (Figure 5.3B) and most common disease stage was 3 (Figure 5.3C). 59 tumours were >5cm and 35 <5cm (Figure 5.3D). 68% of samples had no nodal involvement (N-0) (Figure 5.3E) and the most common T stage was T-2 (Figure 5.3F). The cohort included more males than females, consistent with the known higher prevalence of RMS in males [15] (Figure 5.3G). There were 64 samples annotated as ARMS, 4 mixed ARMS/ERMS and 57 ERMS (Figure 5.3H). Clinically, ERMS has a higher incidence than ARMS (around 60-70% of cases are ERMS and 30-40% ARMS) [71, 72], but this was not reflected in the data from the original study [352]. 53 samples were FP (38 PAX3 and 15 PAX7) whilst 72 samples were FN (15 ARMS and 57 ERMS) (Figure 5.3I). Most samples were annotated as intermediate risk,

with around equal numbers or low and high risk (Figure 5.3J). Around 34% of patients had an event (Figure 5.3K), similar to the percentage reported in the literature [107]. The most common survival time was 2 years (15%), followed by 1 year (14%) (Figure 5.3L). The most common FFS time was 2 years followed by 1 year (Figure 5.3M). Tumour location also reflects what is seen in the literature, with common tumour locations being the head and neck, genitourinary system and extremities (Figure 5.3O).

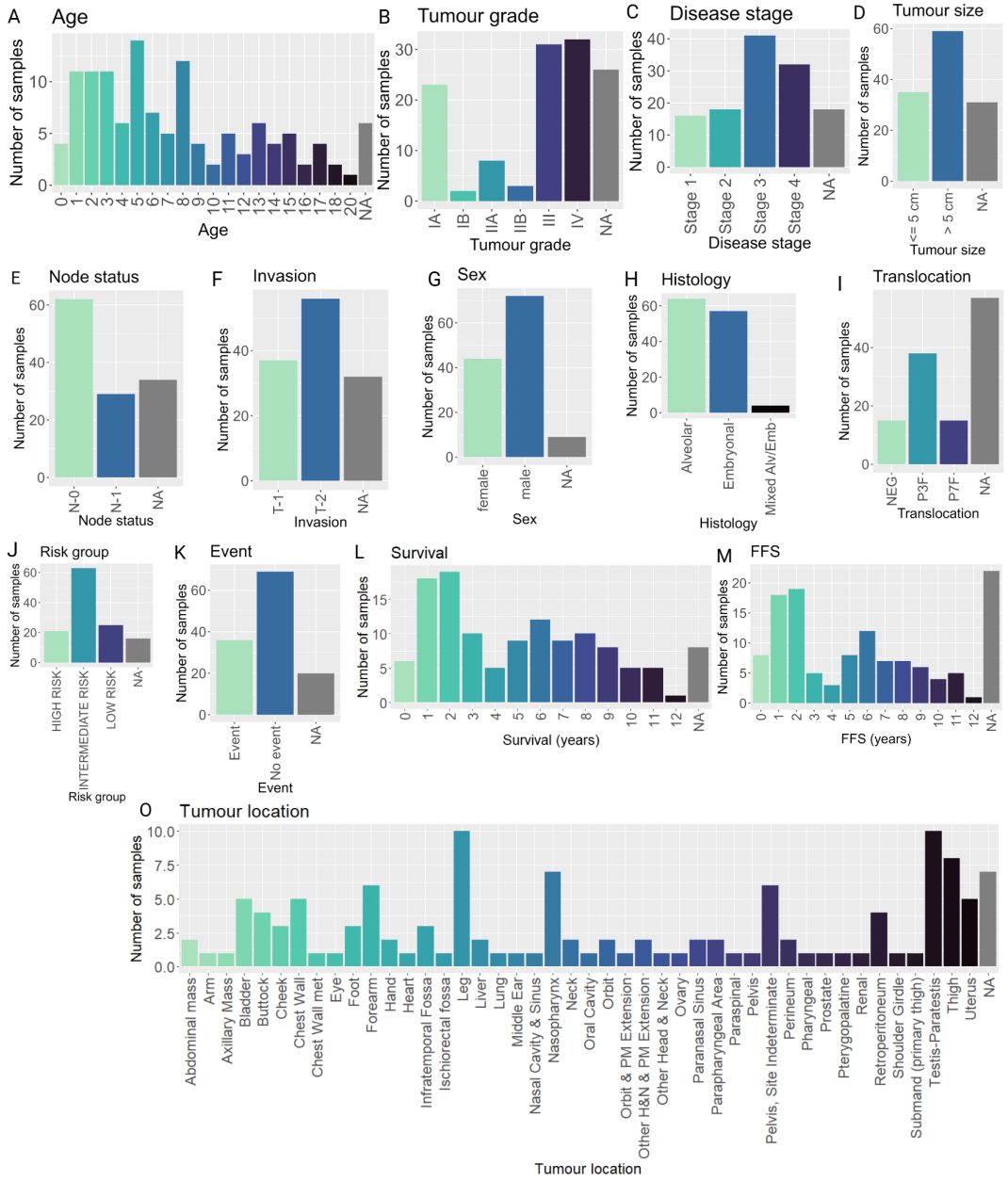


FIGURE 5.3: Clinical characteristics for the Triche dataset. A) Age B) Tumour grade C) Disease stage D) Tumour size E) Node status F) Invasion G) Sex H) Histology I) Translocation J) COG RG K) Event L) Survival M) FFS O) Tumour location. N=125

5.3.1.3 Investigating variation in the data with principal component analysis

Principal component analysis was used to investigate variation in the data that might be due to clinical traits. Some variation in the data was caused by histology as samples broadly divided by histology on PC2 (10% of the total variation) (Figure 5.4A). Annotating sample by chromosomal translocation shows that the samples are better divided by the presence or absence of the fusion gene (Figure 5.4B). ARMS samples that clustered with ERMS in Figure 5.4A are FN (Figure 5.4C). These data reflect what is seen in the literature; FN ARMS is similar to ERMS [110]. RG again shows a division by histology, with FP samples annotated as high or intermediate risk whereas FN samples could be low risk (Figure 5.4D). Node involvement also showed separation in the data, with more FP samples annotated as N-1 (regional nodal involvement) (Figure 5.4E). This is also observed clinically, with a study investigating lymph node involvement finding that N-1 were more likely to have alveolar histology [360]. Annotating samples by disease grade, sex, tumour size, age, survival, invasion (T status) and disease stage does not show any clear patterns in the data (Supplementary Figure 43).

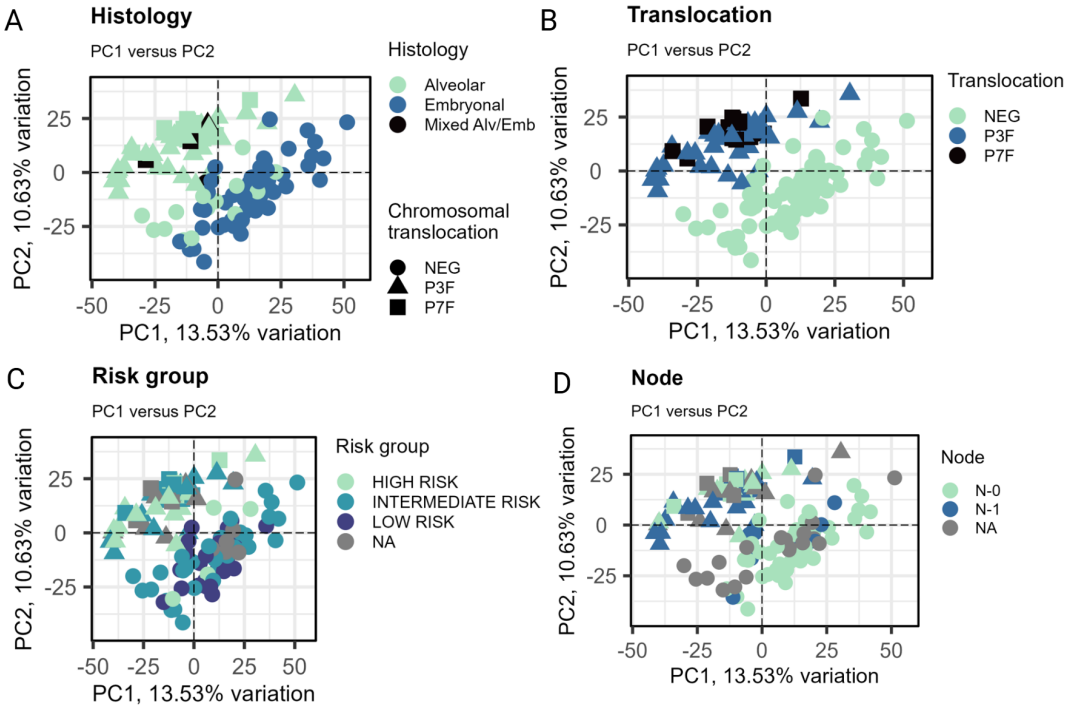


FIGURE 5.4: Principal component analysis biplots annotated with clinical traits for the Triche dataset. A) Histology B) Chromosomal translocation C) Disease stage D) Risk Group E) Node status. N=125

5.3.1.4 Testing for associations between gene signature scores and clinical traits in the Triche microarray data

Samples were scored for the intrinsic, FP-acquired and FN-acquired gene signatures. T tests were performed to assess for significant differences in GSVA scores between clinical traits (e.g. M-0 vs M-1). All results can be found in the Supplementary Files (Supplementary Files/Chapter 5 Signature validation/GSVA score based on DEGs from cell models).

For the intrinsic, FP-acquired or FN-acquired resistance signatures there were no significant differences in GSVA scores and levels of any categorical clinical trait. Looking at the relationship between the GSVA score and continuous variables, GSVA scores for the FP-acquired gene signature showed a significant but weak positive correlation (P=0.04, r=$\sim$0.3) with and survival (Figure 5.5A). However, this relationship is the opposite of what was expected, with higher GSVA scores associated with longer FFS/survival.

5.3.1.5 Testing for associations between differentially expressed gene signature scores and clinical traits in the Triche microarray data

In addition to the gene signatures, GSVA scores were also calculated based on expression of significant upregulated genes from cell models to determine whether DEGs may be better as a gene signature. Genes in the intrinsic, FP-acquired and FN-acquired DEGs can be found in Supplementary Files (Supplementary Files/Chapter 5 Signature validation/Overlapping upregulated DEGs from cell models). To get intrinsic DEGs, genes that were significantly upregulated in both C7 and C13 were identified. This resulted in 5 genes; *CASP10*, *BLK*, *CD33*, *BANK1* and *TMC8*. FP-acquired DEGs consisted of 133 genes that were upregulated in both RH4 IFO-resistant and RH4 VCR-resistant samples. FN-acquired DEGs consisted of 201 genes that were upregulated in both RMSYM IFO-resistant and RMSYM VCR-resistant samples.

There was a significant difference in the GSVA scores based on expression of FN-acquired DEGs between tumour grades. Grade IIB had significantly lower GSVA scores than grade IA and IIA (Figure 5.5B). Grade III had significantly lower scores than IIA (Figure 5.5B). These findings suggest that lower GSVA scores are associated with higher tumour grade which is contrary to the hypothesis in section 1.7.

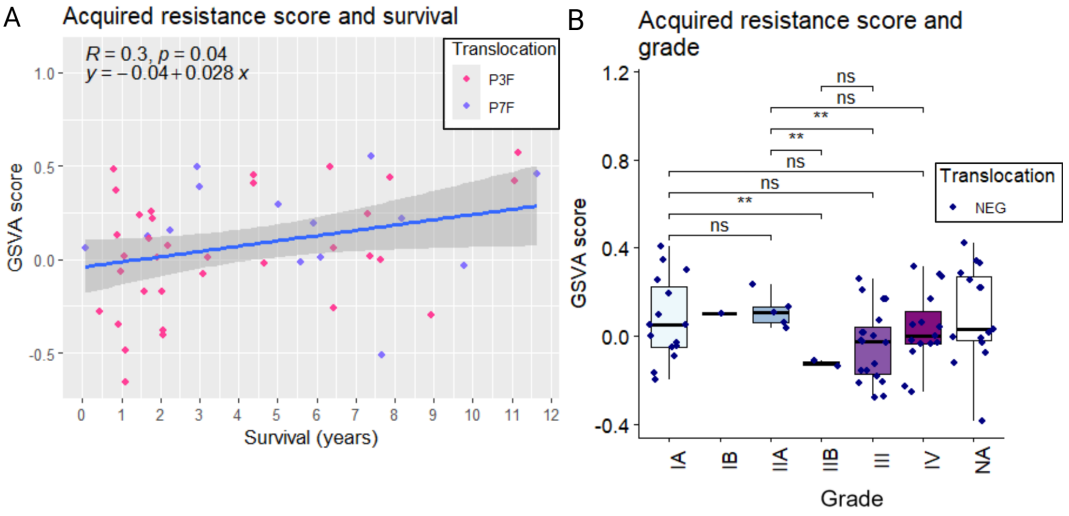


FIGURE 5.5: GSVA scores in the Triche dataset. A) FP-acquired gene signature for survival. 95% confidence interval is shaded. B) FN-acquired DEGs and tumour grade. FN-acquired DEGs consisted of 201 genes that were upregulated in both RMSYM IFO-resistant and RMSYM VCR-resistant samples. T-test. P values were adjusted using Benjamini-Hochberg method. ns: p >0.05, *: p ≤0.05, **: p ≤0.01, ***: p ≤0.001, ****: p ≤0.0001. N=68

5.3.2 Analysis of the Williamson microarray dataset

5.3.2.1 Quality control and processing of the raw microarray data

As in the Triche dataset, the quality of the raw and processed data was assessed to identify any outliers and potential issues with the data. No outliers or issues were detected in the raw or processed data (Supplementary Figure 44 and Figure 45). After one to one mapping between gene and probeset with Jetset, there were 19,968 genes across 101 samples.

5.3.2.2 Characteristics of the clinical data

The dataset contained a higher proportion of ARMS samples than typically seen in clinical settings, with 61% of samples ARMS and 39% ERMS (Figure 5.6A). Although more samples were alveolar than embryonal, there were more FN samples than FP (56 FN and 45 FP)(Figure 5.6B).

Similar to the Triche dataset, there was a slight skew towards males (Figure 5.6C) and more tumours greater than 5cm (Figure 5.6D). The dataset had poor prognostic features with the majority of patients classified as high or very high EpSSG risk group (Figure 5.6E) and high IRS clinical stage (Figure 5.6F). Around 28% of patients had metastasis at diagnosis (Figure 5.6G), similar to the number in a study by Huang et al [361]. Most patient did not have lymph node involvement (∼48%) (Figure 5.6H). 80%

of patients had complete remission (Figure 5.6I), similar to rates seen in other studies [105]. The most common SIOP stage was 2, followed by 1 (Figure 5.6J). As seen in the clinical setting, common tumour locations were the parameningeal (PM) area, limb and orbital (Figure 5.6K). Around half of patients had a first event (Figure 5.6L) with the most common event being local relapse (∼17%) (Figure 5.6M). The most common survival duration was 2 years (26%), followed by 3 years (22%) and 1 year (16%) (Figure 5.6N). Age ranged from 0-22, with the most common age being 6, followed by 2, 4 and 5 years of age (Figure 5.6O).

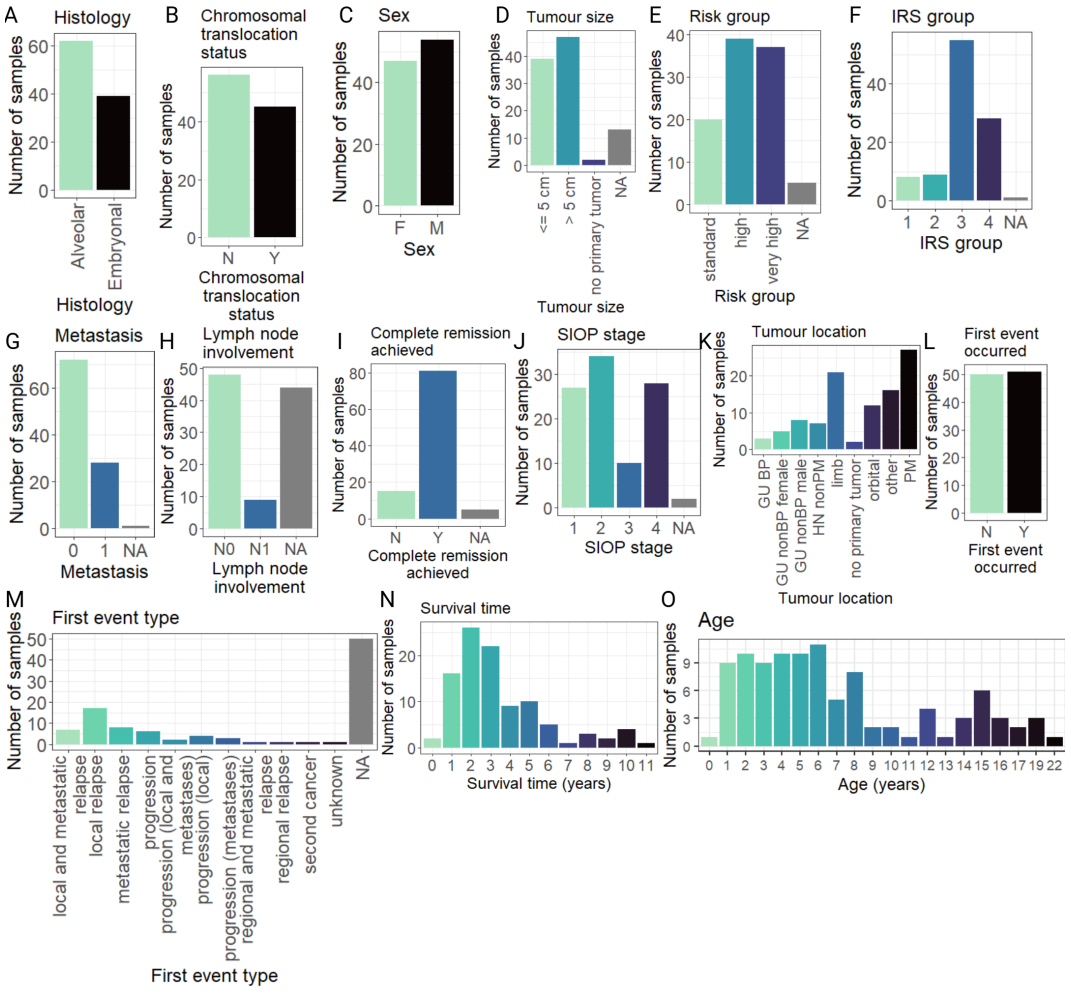


FIGURE 5.6: Clinical characteristics for the Williamson dataset. A) Histology B) Chromosomal translocation C) Sex D) Tumour size E) Risk Group F) IRS group G) Metastasis at diagnosis H) Lymph node involvement I) Complete remission achieved J) SIOP stage K) Tumour location L) First event occurred M) First event type N) Survival N) Age. N=101

5.3.2.3 Investigating variation in the data with principal component analysis

As expected, fusion gene status was responsible for the majority of variation in the data, dividing samples on PC1 (15% of the total variation in the data)(Figure 5.7A). This correlated strongly with histology, with most FP samples annotated as ARMS and FN ARMS clustered with ERMS (Figure 5.7A). It also highlighted 3 samples that were annotated with both ERMS histology and FP gene status (Figure 5.7A). It's possible that the histology of these samples were misclassified. In the PCA bi-plot looking at whether a first event has occurred, it is apparent that more FP samples have had a first event (Figure 5.7B). It can also be seen that FN-RMS is more common in children and less common in adolescents (Figure 5.7C). This has been seen in other studies, which show ERMS incidence peaks around the ages of 0-4 years of age [362]. The majority of FP cases were also annotated as either high risk or very high risk, with only two samples annotated as standard risk (Figure 5.7D). Again this is expected, as under current guidelines FP-RMS is either classified as high or very high risk [363]. More FP patients had metastasis at diagnosis (Figure 5.7E) and did not achieve complete remission compared to FN patients (Figure 5.7F). No clear inferences could be drawn from PCA biplots of other clinical traits (Supplementary Figure 46).

5.3.2.4 Testing for associations between gene signature scores and clinical traits in the Williamson microarray data

Samples were scored based on the gene signature derived from cell models (intrinsic, FP-acquired and FN-acquired). Significant differences were tested between levels of each relevant clinical trait.

For the intrinsic resistance signature, patients with a SIOP stage of 1 had a significantly higher GSVA score than patients with SIOP stage 2 ($p \leq 0.01$) and 4 ($p \leq 0.05$) (Figure 5.8). This finding does not support the hypothesis (section 1.7), which predicted that higher SIOP stages would be associated with higher GSVA score.

5.3.2.5 Testing for associations between differentially expressed gene signature scores and clinical traits in the Williamson microarray data

No significant associations were observed between clinical traits and GSVA scores derived from DEGs in the cell models.

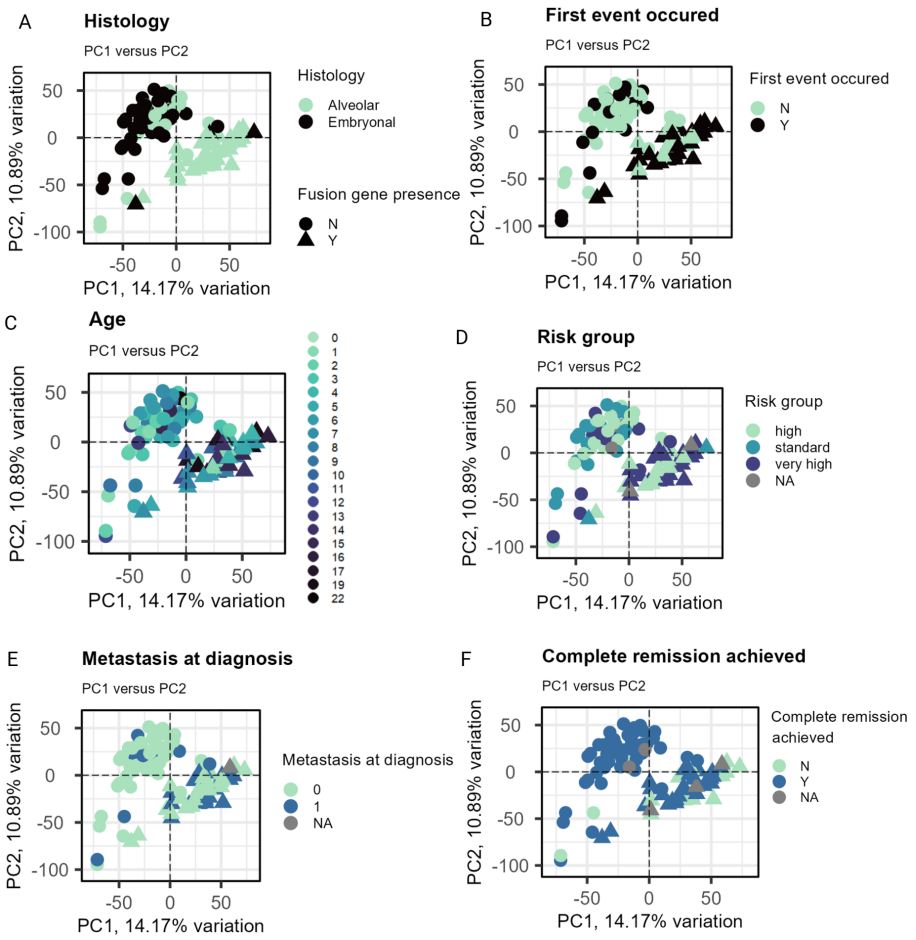


FIGURE 5.7: Principal component analysis biplots annotated with clinical traits for the Williamson dataset. A) Histology B) First event occurred C) Age D) Risk group E) Metastasis at diagnosis F) Complete remission achieved. N=101

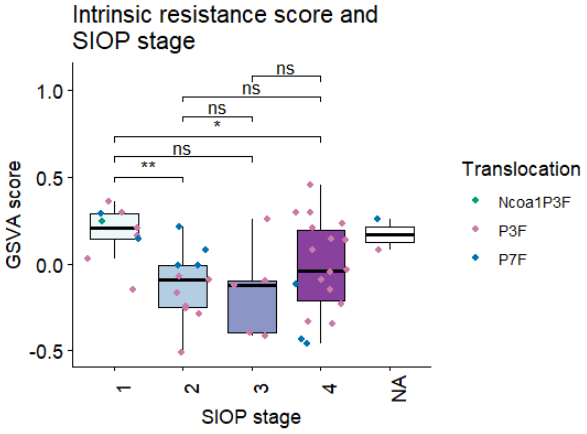


FIGURE 5.8: GSVA scores based on overlapping intrinsic DEGs for SIOP stage for the Williamson dataset. T test. P values were adjusted using Benjamini-Hochberg method. ns: p >0.05, *: p ≤0.05, **: p ≤0.01, ***: p ≤0.001, ****: p ≤0.0001. N=45.

5.3.3 Investigating resistance in RNA-sequencing data from pre- and mid-treatment patient samples

5.3.3.1 Quality control of the data

The raw dataset contained 60,754 Ensembl IDs. After mapping gene IDs to Ensembl IDs the dataset was reduced to 17,062 genes. As in RNA-sequencing analysis in

previous chapters, library sizes of the raw data were assessed. After normalisation, the variation in library sizes was reduced and all samples showed a similar distribution of library sizes (Supplementary Figure 47 and Supplementary Figure 48). No outliers were identified.

PCA was used to investigate sources of variation in the data in an unsupervised way. On PC1 (representing around 30% of the total variation in the data), samples broadly separated into pre- or mid-treatment (Figure 5.9A and B). The mid-treatment samples for patients 5 and patient 25 did not cluster with the other mid-treatment samples and exhibited minimal divergence from their corresponding pre-treatment samples (Figure 5.9A). These were flagged as potential outliers, but were retained for further analysis at this point to increase the sample size. On PC2 (representing around 12% of the total variation in the data), samples divided into FP and FN (Figure 5.9B).

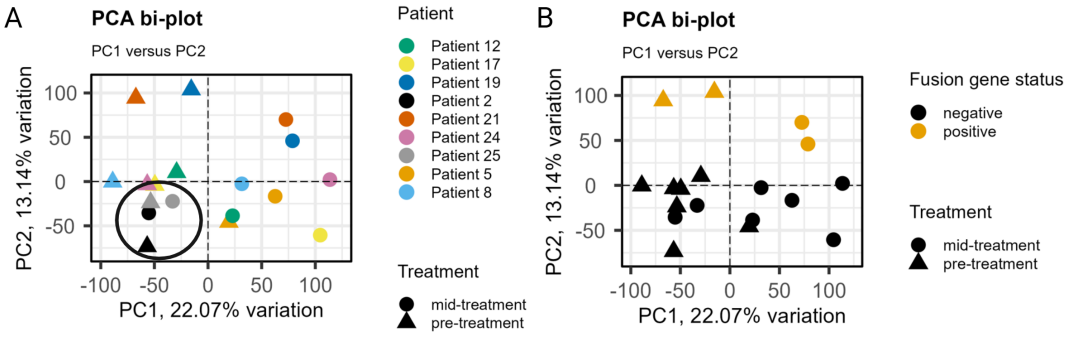


FIGURE 5.9: Biplot showing PC1 vs PC2 for the pre- and mid-treatment patient RNA seq data. A) Colour representing patient B) Colour representing RMS fusion gene status FP or FN. Black circle highlights potential outliers. N=18.

5.3.3.2 Correlation of gene signature scores with treatment

GSVA was used to score samples based on the gene signatures and T tests were performed to compare GSVA scores between pre and mid-treatment samples. There were no significant differences in GSVA scores between treatment stage (pre- and mid-treatment) for any gene signatures (Figure 5.10A, B and C). However, for the intrinsic resistance signature, mid-treatment sample do appear to have a higher score than pre-treatment samples but this is not significant likely due to the small sample size (Figure 5.10A). A post-hoc power analysis was performed using the observed effect size (Cohen's d = 4.30) from the paired t-test. This resulted in a statistical power of 36.7% (α = 0.05, two-tailed), indicating a high likelihood of a false negative result. For the acquired resistance signatures, GSVA scores pre- and mid-treatment were highly variable between patients (Figure 5.10B and C). GSVA scores of patient 5 and 25 (previously flagged as potential outliers) are not abnormal and therefore these samples were not removed from the analysis (Figure 5.10C).

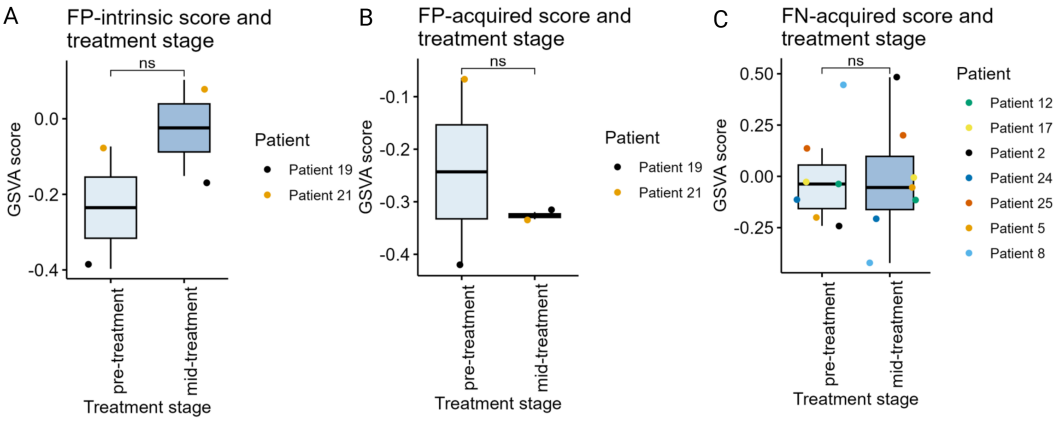


FIGURE 5.10: GSVA scores of the gene signatures for pre- and mid-treatment samples. A) Intrinsic gene signature (N=4) B) FP-acquired gene signature (N=4) C) FN-acquired gene signature (N=14). Significant differences in GSVA scores between treatment stages was assessed using paired T test. P values were adjusted using the Benjamini-Hochberg method. GSVA gene signature scores were significantly different if adjusted p value <0.05.

5.3.3.3 Correlation of differentially expressed genes with treatment

In addition to the gene signatures, GSVA scores were calculated based on significant upregulated genes from cell models. GSVA scores for intrinsic DEGs and FP-acquired DEGs were higher in the mid-treatment samples compared to pre-treatment, although the difference was not statistically significant, likely due to the small number of samples (Figure 5.11A and B). GSVA scores of DEGs from FN-acquired cell models had a significantly higher score mid-treatment compared to pre-treatment (Figure 5.11C). These findings suggest that DEGs may serve as more effective gene signatures of resistance compared to those derived from WGCNA.

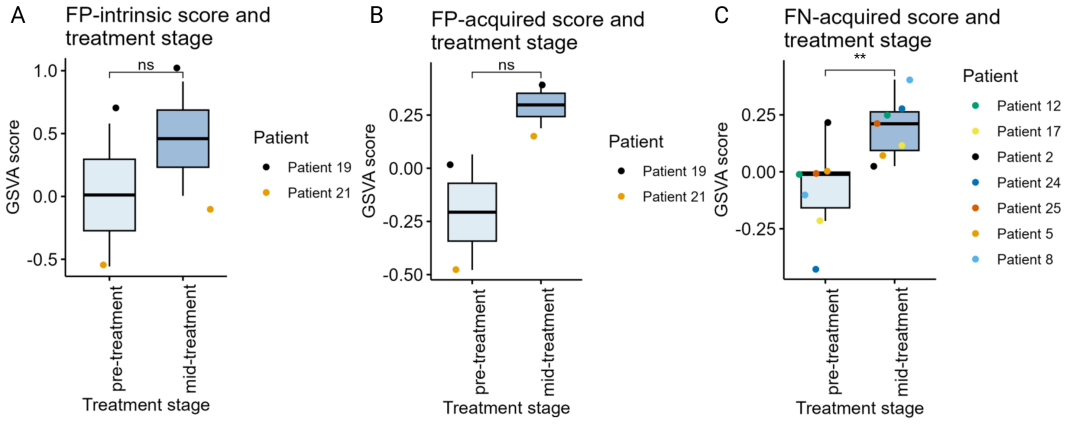


FIGURE 5.11: GSVA scores of DEGs from cell models in pre- and mid-treatment samples. A) Intrinsic DEGs (N=4) B) FP-acquired DEGs (N=4) C) FN-acquired DEGs (N=14). Significant differences in GSVA scores between treatment stages was assessed using paired T test. P values were adjusted using the Benjamini-Hochberg method. GSVA gene signature scores were significantly different if adjusted p value <0.05.

5.3.3.4 Identifying enriched gene sets in mid- vs pre-treatment

As in previous RNA-sequencing analysis, GSEA was conducted for GO terms, Hallmarks and Reactome gene sets. All enriched gene sets can be found in the Supplementary files (Supplementary Files/Chapter 5 Signature validation/Pre- and mid-treatment patient RNA seq data/GSEA). Enriched gene sets were visually compared with those from the cell models to determine if the cell models are a good representation of treatment in patient tumours. Of all the cell models, the HD model was the most similar to mid-treatment patient samples. Out of the 18 enriched Hallmarks in HD, 14 of them were also enriched in the same direction as mid-treatment samples (Figure 5.12A and B).

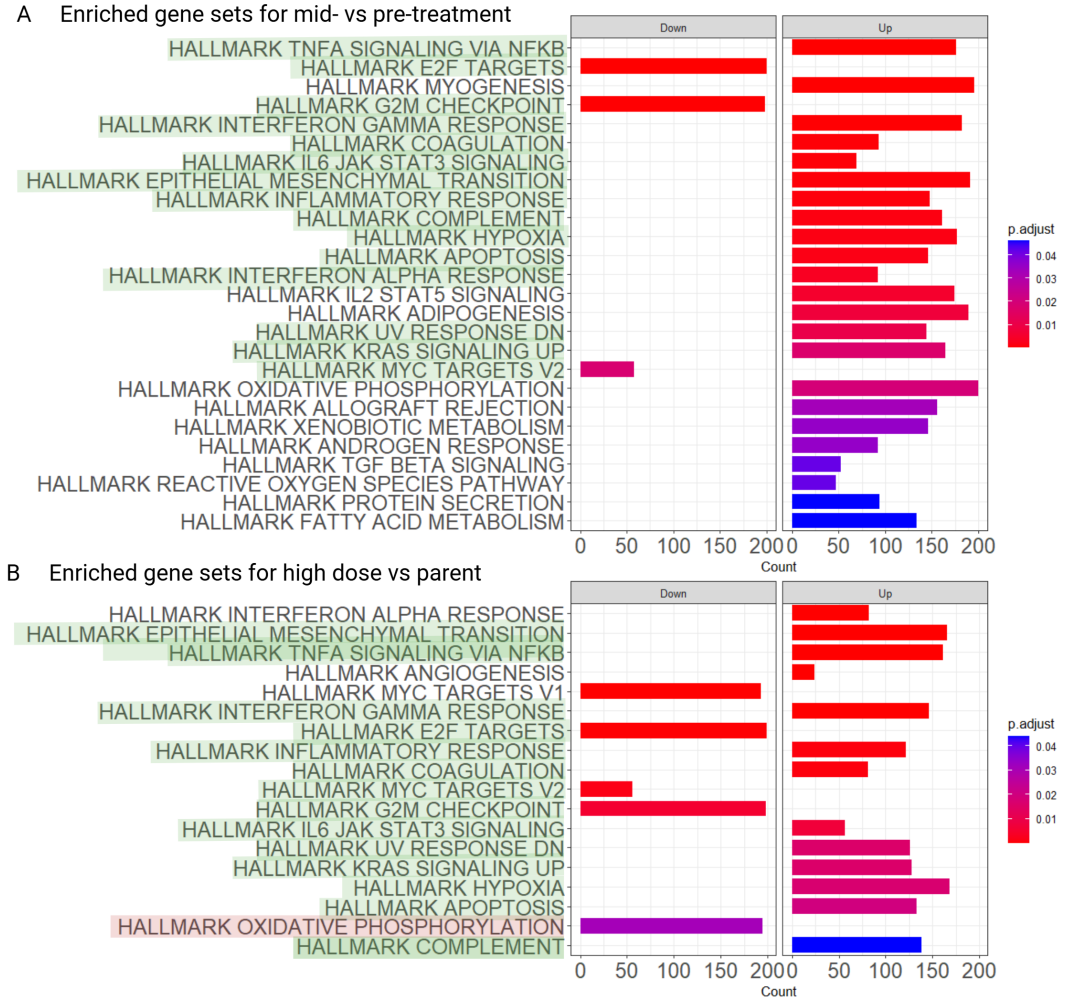


FIGURE 5.12: GSEA results from CAMERA showing enriched gene sets in mid-treatment and HD from cell models. A) Enriched Hallmarks for mid- vs pre-treatment samples B) Enriched Hallmarks for high dose vs parent from intrinsic resistance cell models. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. Gene sets highlighted in green are enriched in both conditions. Gene sets highlighted in red are enriched but in opposite directions.

Further evidence to support that the high dose model is most similar to mid-treatment patient samples is scatterplots of gene expression. Of all the cell models of resistance, the high dose model showed the strongest correlation between the LFC of HD vs P and mid- vs pre-treatment (r=∼3) (Figure 5.13A), suggesting it is the model that is most similar to mid-treatment patient samples. The modest strength of this correlation may relate to differences in the sample type (i.e comparing cell lines to patient tumours) or differences in RMS subtype (patient data contained predominately ERMS samples whereas the high dose model is in an ARMS cell line) . The clones and acquired resistance cell models showed no correlation in the LFC of genes with mid-treatment samples (Supplementary Figure 49).

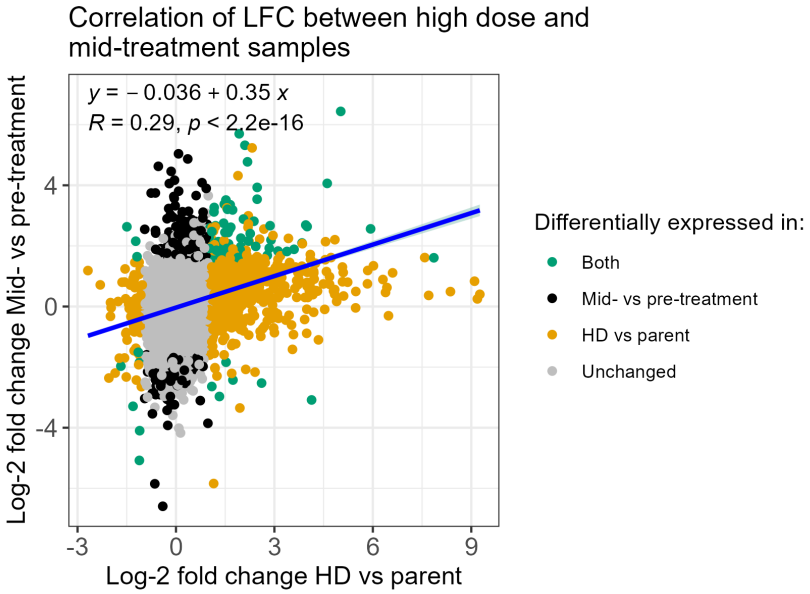


FIGURE 5.13: Scatterplot showing the LFC of genes for high dose cell model and mid-treatment samples. DEGs were defined as having an adjusted p value <0.05 (Benjamini-Hochberg) and LFC >1 or LFC <-1 for HD and LFC >1.5 or LFC <-1.5 for mid- vs pre-treatment.DEGs are highlighted in colour, corresponding to the group in which they are differentially expressed. 95% confidence interval is shaded.

5.4 Discussion

This chapter aimed to determine whether the gene signatures from cell models could distinguish patients with worse clinical outcomes. This was tested on two different microarray datasets from patients before treatment and one RNA-sequencing dataset with pre- and mid-treatment samples. Statistical tests were performed to identify patients with worse clinical outcomes had a significantly higher GSVA score based on the expression of gene signatures. GSVA scores based on the gene signatures were significantly different in some conditions, however the observed relationships were opposite to those predicted with high GSVA being associated with better clinical outcomes. Specifically, the FP-acquired signature was associated with longer survival and the FN-acquired signature was associated with lower tumour grade in the Triche dataset. The results from this chapter do not support the hypothesis, as higher gene signature scores did not correlate with worse clinical outcomes.

There may be multiple reasons as to why the gene signatures from cell models are not able to differentiate samples with a worse clinical outcome. Using WGCNA to derive a gene signature may have resulted in genes that showed consistent but smaller changes in gene expression in the cell models, which could make it challenging to detect these differences in the patient data. To explore this, gene signatures based on DEGs from the cell models were also tested. In one of the microarray datasets, GSVA scores based on the intrinsic DEG signature were significantly higher in patients with SIOP stage 1 compared to 2 and 4. This is the opposite to what was predicted, as it would be expected that GSVA scores would be higher in higher SIOP stages. DEG signatures were more promising in the pre- and mid-treatment patient RNA-sequencing data, with higher GSVA scores in mid-treatment samples (although not significant in FP). Alternatively, as discussed in the previous chapter, WGCNA may have identified genes that reflect technical variation rather than biological co-expression. This could explain why the gene signatures that were identified could not be validated in the independent patient datasets. The observation that FN-DEG signature was significantly in the pre- and mid-treatment samples but not the WGCNA signature could support this hypothesis.

Differences due to data type (i.e comparing gene signatures derived from RNA-sequencing to microarray) could be another potential cause of why the signatures are not seen in patient data. There is a wider quantitative range of expression level changes in RNA-sequencing data compared to microarray [364], which may affect the scoring of gene signatures as there may be less variation in gene expression in the microarray data. Indeed, for the microarray data of the Triche and Williamson datasets, PC1 represents around 13% of the variability in the data, whereas from the RNA-sequencing data of the clones PC1 represents around 45% of the total variation. As well as this, correlations in gene expression between

RNA-sequencing and microarray data are not perfect (r values of around 0.7 have been reported) [364], which may also contribute to why the gene signatures do not correlate with clinical traits in the patient microarray data. However, signatures were not higher in mid-treatment samples from patient RNA-sequencing data, which suggests differences in data type is not the single factor that is responsible for the gene signatures not being present.

Another factor that may explain why the signatures could not be validated in patient data is due to differences in the sample type. The gene signatures were derived from cell models but these genes may show different patterns of gene expression in patients. In the literature, there is some evidence that RNA-sequencing data from RMS cell lines varies from that of primary tumours [365]. RMS cell lines have also been shown to mainly represent the proliferative component of tumours but lack neuronal-like cells seen in FP patient tumours [318]. However, other studies have shown good concordance between cell lines and primary tumours [366, 367]. In the data from this chapter, there was no correlation in gene expression between most of the cell models and the patient RNA-sequencing data, with the exception of the HD cell model which had a weak correlation. This could suggest a discordance in gene expression between cell lines and patient data and could contribute to why the gene signatures do not correlate with clinical outcomes in the patient data.

A limitation of this work is the dataset the gene signatures were tested in. Gene signatures that would be successful in identifying patients with worse clinical outcomes in the patient microarray data would reflect differences in gene expression that already exist within patient tumours before treatment. However the acquired resistance signatures do not represent this. Instead, they represent changes in gene expression after chemotherapy so it may be expected that they do not correlate with clinical outcomes from pre-treatment data. These gene signatures are more likely to be present in chemotherapy-treated samples. Ideally, these signatures would be validated in RNA-seq data comparing pre-treatment and relapsed samples, however no publicly available data was available. The closest data that was available was from pre- and mid-treatment patient RNA-sequencing data, although this also showed no significant differences in acquired resistant signature scores. This dataset had a very small sample size, which is a limitation. It's possible that acquired resistance models in research do not reflect resistance as it presents clinically. This is supported by the observation that the high dose cell model most closely resembled mid-treatment patient samples and suggests other models of resistance may be more clinically relevant for RMS. Although the intrinsic resistance signature is closer to what is being tested for in the microarray data (as it intends to represent differences in gene expression that already exists within resistant cells), the signature represents individual cells and it is likely that in a tumour there is only a small number of these cells. Therefore, patients who had a worse clinical outcome may have shown changes in expression of the intrinsic

gene signature but this might not have been seen as these changes may only be present in a few cells. To address this, the intrinsic gene signature could be investigated using patient single-cell RNA-sequencing data to determine if there are cells types within the tumour that exhibit the intrinsic resistance signature; and if present, whether these cells are enriched in patients experiencing relapse.

In conclusion, the gene signatures derived from cell models could not be validated in pre-treatment patient microarray data or mid-treatment patient RNA-sequencing data. Potential contributing factors include differences in data type and sample type, biological variability in patient data, the method of deriving a gene signature and treatment status at the time of sampling. It was found that most resistant cell models did not reflect treated patient samples. The high dose cell model was the most similar to treated patient samples and may be a better model of resistance. The next chapter employed machine learning models to derive a resistance gene signature from the patient microarray data.

Chapter 6

Deriving a gene signature of resistance from patient data using machine learning models

6.1 Introduction

The application of machine learning in healthcare and medical research has significantly increased over the last 10 years [368]. Machine learning is a branch of artificial intelligence and refers to the ability of algorithms or statistical models to analyse and identify informative patterns within data [369]. Using variables (also called features) from input data, machine learning can be used to build models that can make predictions or decisions on new datasets. Input data can be either labelled, meaning it includes the true outcome or target variable, or unlabelled, where the true outcome is not provided. If the input data provided to the model is labelled this is called supervised machine learning, whereas unsupervised learning refers to a model that identifies patterns in unlabelled data [369]. Both unsupervised and supervised frameworks can be used for the identification of prognostic signatures, however this chapter will focus on the supervised framework as this will be used to generate the machine learning model. In the generation of prognostic signatures, gene expression or clinical data are input as continuous or categorical features (Figure 6.1). The input datasets also contain information on patient survival or FFS, which is the outcome variable, thus making this a form of supervised machine learning (Figure 6.1).

Prognostic gene signatures have been identified through machine learning in many cancer types, including kidney [370], gastric [371], colorectal [372] and hepatocellular carcinoma [373]. In RMS, most machine learning research has been applied to histopathology, particularly its use in diagnosis [374], identifying molecular subtypes [375] and predicting survival [376]. To date, no studies have applied machine learning

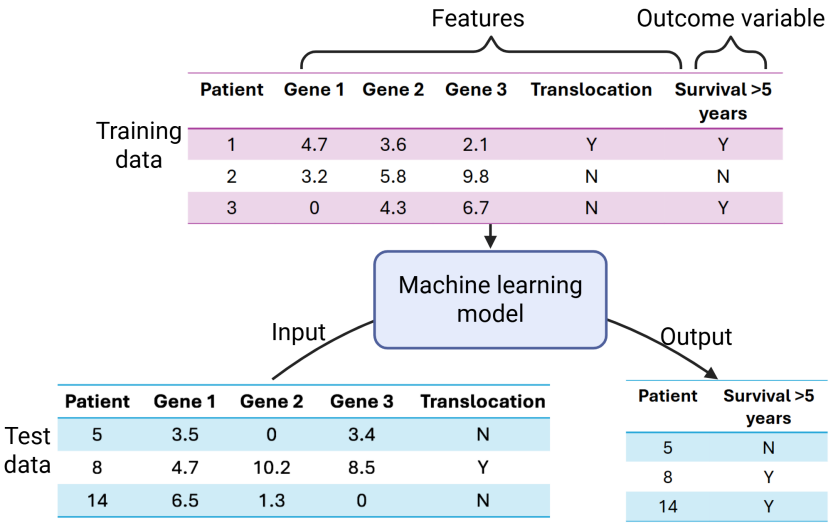


FIGURE 6.1: Example of a supervised machine learning model for identifying prognostic signatures. Input data is gene expression and clinical data, where features are gene expression values and clinical variables. Input data is labelled, as it contains information on the outcome variable (survival >5 years). Input data is used to train a machine learning model that can then predict the outcome variable on a new dataset (test dataset) based on the features it was trained on.

to predict survival or relapse in RMS using gene expression and clinical data. Using machine learning models to predict relapse/progression is a novel approach that could uncover genes, mechanisms and clinical traits associated with chemotherapy failure in RMS. This has been done in other cancer types, such as oesophageal squamous cell carcinoma, where a machine learning model identified a prognostic signature and determined the genes in the signature were related to immunological and metabolic pathways [377]. Additionally, gene signatures may aid in identifying potential therapeutic targets. For example, Al-Bzour et al. identified S100A9 and PBK as potential therapeutic targets in hepatocellular carcinoma [373].

Previous work has been conducted investigating clinical features associated with relapse in RMS, with multiple studies looking at the association between clinical features and FFS. These studies have provided some insight into which clinical variables may be important in relapse/metastasis. One study conducted univariate cox-regression analysis and found that FP-RMS, IRS group 3, unfavourable tumour site and T2 classification for tumour invasiveness were associated with a decreased FFS [86]. However, only fusion status and tumour invasiveness remained as independent prognostic factors after multivariate analysis [86]. Another study identified factors associated with a worse FFS through multivariate cox-regression analysis [378]. They found that age at diagnosis interacting with tumour size, primary tumour site, IRS group and fusion status were independent prognostic factors.

Other studies have focused on prognostic factors with FFS in subsets of RMS patients. Oberlin et al. looked at prognostic factors with FFS in metastatic RMS [90]. The

authors found that unfavourable age ($\leq$1 or $\geq$10), unfavourable tumour site, bone or bone marrow involvement and having 3 or more metastatic sites were associated with a worse FFS [90]. Walterhouse et al. found that age category, tumour size, IRS and T stage were prognostic in in patients with localised paratesticular RMS [379].

This chapter aims to create a machine learning model capable of predicting whether a patient will experience an event (relapse, progression or second cancer) using pre-treatment patient microarray data and clinical data. It is hoped that this machine learning model could be applied to new patient data to predict which patients are likely to relapse so they can be prioritised for clinical trials. This novel machine learning approach may provide insight into mechanisms of resistance to chemotherapy and could be used to identify genes to be investigated as potential therapeutic targets in RMS.

6.2 Methods

6.2.1 Datasets

The Triche and Williamson RMS microarray datasets from Chapter 5 were used for generating a machine learning model that predicts resistance. These were the largest publicly available datasets from pre-treatment samples where the outcome variable of interest (whether a patient had an event after chemotherapy) was available. 105 samples from the Triche dataset and 101 samples from the Williamson dataset had information on the outcome variable and were included in the analysis. Information on the type of event was not available in the Triche dataset. In the Williamson dataset, types of event included relapse, progression, second cancer and unknown. The single-cell dataset from Danielli et al. [318] was also considered, however there was only 13 pre-treatment patient samples which is insufficient to train and test a machine learning model.

The processed expression data from Chapter 5 (RMA normalised and one-to-one mapped probes) was used. In each dataset, genes that were duplicated after one-to-one mapping were collapsed to give the average expression, as described in [380]. A summary of the workflow for the input data can be found in Figure 6.2.

6.2.1.1 Imputation of missing clinical values

Since the input data for machine learning models cannot contain missing values, missing clinical variables were imputed rather than removed to retain sample size. This involves estimating missing values based on the complete dataset. The R package

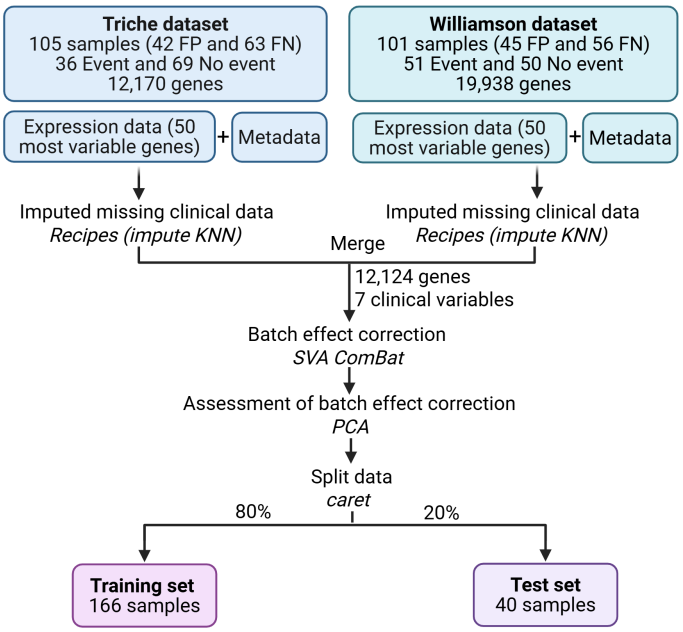


FIGURE 6.2: Workflow of the input data for the machine learning models. For each of the two microarray datasets, missing clinical data was imputed with the R package Recipes using KNN method based on the expression of top 50 most variable genes and clinical data. Datasets were merged based on common genes and 8 selected clinical variables (age, histology, translocation, fusion status, tumour size, node status, sex and batch). Batch effects were corrected for with SVA ComBat and assessed using PCA. Data was split into training and test set using the caret package function createDataPartition. Created with BioRender.com.

Recipes (version 1.3.1) [381] was used for imputation based using the k-Nearest Neighbours (KNN) algorithm. KNN was selected as the imputation method as it has been shown to outperform simpler methods [382] and less computationally expensive compared to other methods such as bagged trees.

In high-dimensional datasets, such as gene expression data, performing imputation using all available variables can decrease accuracy due to the curse of dimensionality and the inclusion of noisy features. For this reason, each dataset was imputed based on the gene expression of the 50 most variable genes and the clinical variables. For the Triche dataset, clinical variables were sex, age, disease stage, disease grade, tumour location, histology, translocation, overall survival time, event, FFS, RG, tumour size, lymph node involvement, invasion and fusion status. This gave a total of 65 variables. For the Williamson dataset, clinical variables were tumour location, sex, histology, metastasis at diagnosis, IRS group, SIOP stage, tumour size, fusion gene status, translocation, RG, age, complete remission achieved, first event occurred, first event type, overall survival time and lymph node involvement. This gave a total of 66 variables.

6.2.1.2 Merging expression data and adjusting for batch effects

As both datasets were similar in size, they were merged and then split into a larger training set and smaller test set. This approach enabled the use of a more substantial training set which is likely to improve model performance, while retaining enough samples for model validation.

In addition to expression data, clinical variables such as age, fusion status and tumour size were also input as predictors. Integrating clinical and expression data has been shown to increase model performance [383] and is frequently used in prediction models [384, 385]. Seven common clinical variables (age, histology, translocation, fusion status, tumour size, node status and sex) that were present in both datasets were selected to include as predictors. Other clinical variables were dropped as they either were not present in both datasets or included post-diagnostic information (such as survival time).

The Triche and Williamson datasets were merged based on common genes and the seven selected clinical variables. The R package Surrogate Variable Analysis (SVA) [386] and ComBat function based on methodology described in Johnson et al. [387] was used to adjust for batch effects in the expression data. This function is applicable where the batch covariate is known, in this case whether data was from the Triche or Williamson dataset. Other covariates that do not want to be adjusted for can be inputted, which included the 7 clinical variables. Event was the clinical variable treated as the outcome of interest in the model. This resulted in a single merged dataset of 206 samples and 12,133 features (input variables that will be used to make predictions), consisting of 12,124 genes and 8 clinical variables (age, histology, translocation, fusion status, tumour size, node status, sex and batch). Batch was retained as a clinical feature to investigate if it was an important feature in the models, as this would indicate that a batch effect still remained.

6.2.1.3 Assessing batch effect correction with PCA

After the datasets were merged and adjusted for batch effects, PCAtools (version 2.18.0) [211] was used to assess if batch effects were removed and biological variation retained.

6.2.1.4 Splitting data into train and test sets

The dataset was then split 80/20% into a train and test set using the createDataPartition function from the caret package (version 7.0-1) [388]. Event was

used as the outcome of interest when splitting the data, to ensure that the training and test datasets were balanced for this variable.

6.2.2 Feature preselection

High-dimensional data where the number of features significantly exceeds the number of samples, such as gene expression data, can cause issues for machine learning models. High-dimensional data can increase the risk of overfitting, where the model fits too closely to the training data, capturing noise and outliers, resulting in models that have poor generalizability and performance on new data [389]. There can also be challenges computationally, with a high number of features requiring more computational power. It is essential to reduce the number of features used to train the model. One approach to address this is feature preselection, where features are filtered before training the model to remove unimportant features and retain important ones. There are multiple approaches for this, including selecting genes that are differentially expressed [390, 391], show highly variable expression [392], have known biological relevance [373] or are associated with survival or another clinical trait [390, 393, 394, 395]. Since there were no significant differentially expressed genes between event vs no event and there is little literature on resistance in RMS for knowledge-based filtering, it was decided to filter for genes that are associated with the clinical trait event using logistic regression.

6.2.2.1 Feature preselection with logistic regression

Each of the 12,124 genes and 8 clinical variables was tested in a univariate logistic regression model using 'Event' as the binary outcome. Features with a p value <0.01 were selected. A less stringent cutoff was chosen to ensure that potentially important genes were included in the analysis as unimportant features would be excluded with further feature filtering. Over-representation-based GSEA for GO terms using the R package clusterProfiler (version 4.14.6) [238] was conducted. This was done separately for genes associated with either an increased or decreased risk of event. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). A summary workflow of training and testing the machine learning models can be seen in Figure 6.3.

6.2.2.2 Filtering features with recursive feature elimination

Features associated with event status were analysed using RFE through the caret package [388], a feature selection method that iteratively removes features and evaluates model performance to determine the optimum features. A random forest

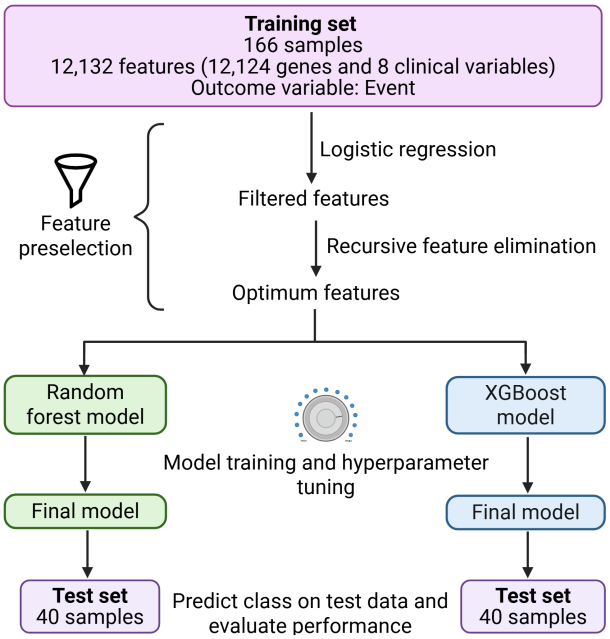


FIGURE 6.3: Workflow of the training and testing of machine learning models. From the training data, features were filtered by logistic regression and recursive feature elimination to get the optimum features. Random forest and XGBoost models were trained on the optimum features and underwent hyperparameter tuning. The final models were used to predict class ('Event'/'No event') on the test data and model performance was assessed. Created with BioRender.com.

model was used with 5-fold cross-validation repeated 5 times. In this approach, the dataset is split into 5 equal partitions, and the model is trained on 4 partitions and performance tested on the other partition. This cross-validation process was repeated 5 times to ensure robustness. The optimum features were identified through RFE results.

6.2.3 Training the models

The training data was filtered for the features determined through RFE. Two models were trained using the caret package [388]; Random forest and eXtreme Gradient Boosting (XGBoost). The random forest model trains multiple decision trees on different subsets of features. The results from these decision trees are combined and classifications are made based on the majority vote (the category predicted by most trees will be the predicted outcome). XGBoost models are also a tree-based model that uses boosting; where weak learners (ie models with limited predictive power) are sequentially trained based on the errors of the previous learner and combined to form a single stronger learner. With XGBoost, the first tree is the weakest learner and new trees are added based on the mistakes of this tree. Each tree that is added minimises the loss function (a measure of model performance) from the previous tree. These

models were selected as they are generally robust with high-dimensional data and are popular models for deriving gene signatures [371, 396, 397, 398, 399].

Models were trained using the caret package [388] with 5 cross-fold validation repeated 10 times. The parameter summaryFunction was set to twoClassSummary to compute performance metrics. The performance metric selected to optimise in the model was area under the Receiver Operating Characteristic (ROC) curve, as this may improve model performance for imbalanced data. For XGBoost, the data was pre-processed (centred and scaled) as, although not required, this can improve model performance. For XGBoost, L2 regularisation was applied with the parameter lambda=1.

6.2.3.1 Sampling methods

For datasets where the outcome variable is imbalanced, sampling techniques that adjust for class imbalances can improve model performance. Upsampling is a sampling technique where data is added to the minority class from the original samples until the classes are balanced. Downsampling removes samples from the majority class until the classes are balanced. For each model, standard sampling, upsampling and downsampling were attempted to determine which sampling technique resulted in the best model performance.

6.2.3.2 Optimisation of hyperparameters

Hyperparameters are parameters that control the models learning and are set before training of the model. They include things such as the maximum depth of decision trees, number of boosting rounds and number of features used per tree. For each model there are a range of different hyperparameter values that can be used that may influence model performance. Tuning grids can be used to provide a range of hyperparameters to pass to the model, for which then optimum hyperparameters can be selected.

For random forest models, the hyperparameter mtry is available for tuning with the caret package. mtry is the number of features randomly selected when building each decision tree. The default value for mtry is the square root of the number of samples ($\sim$13). The values for mtry selected to try were 5, 10, 15, 20, 25.

For the XGBoost model, hyperparameters included the number of boosting rounds, the maximum depth of trees, learning rate (the rate at which the boosting algorithm learns from each iteration), gamma (a regularisation parameter), percentage of features used for each tree, minimum child weight (number of samples required in a child node for a split to be made) and the percentage of samples used per boosting round.

The range of values selected for these parameters aimed to reduce overfitting of the model due to the high dimensionality of the data. This included making shallower trees, slower learning rates, increasing gamma, a lower number of features used per tree, lower number of samples used for boosting and larger minimum child weight.

6.2.4 Evaluation of model performance on the test data

The models were used to make predictions on the test data, with each sample being assigned a prediction probability for 'Event' and 'No event'. The threshold for classification ('Event' or 'No event') was optimised through ROC curves based on Youden's J (sensitivity + specificity - 1). Model performance was evaluated based on the true outcomes.

6.3 Results

6.3.1 Assessing batch effect correction with PCA

After merging the Triche and Williamson datasets and adjusting for batch effects, PCA was used to assess if the batch effects had been corrected for. Before correction, batch was responsible for the majority of the variation ($\sim$70%) in the data since the Triche and Williamson datasets formed separate clusters on the biplot Figure 6.4A). After batch correction, this variation due to batch effect was no longer apparent as samples did not cluster by batch in the PCA biplot Figure 6.4B). In Chapter 5, it was shown that chromosomal translocation (FN or FP) was responsible for most of the variation in the microarray datasets. Translocation status was therefore used to assess if the data retained biological variation after adjusting for batch effects. After correction, samples continued to cluster by translocation on PC1 (Figure 6.4B), suggesting that biological variation was still retained.

6.3.2 Filtering features to train the model using univariate logistic regression

After performing univariate logistic regression for each feature, there were 608 features significantly associated with the clinical trait event (p value $<$0.01). Of these, 336 features were associated with an increased (indicated by an odds ratio $>$1) and 272 features showed a decreased risk (indicated by an odds ratio $<$1). The average expression of genes significantly associated with event was 5.94 and expression ranged from 2.42 to 13.62. The clinical variables that were significantly associated with

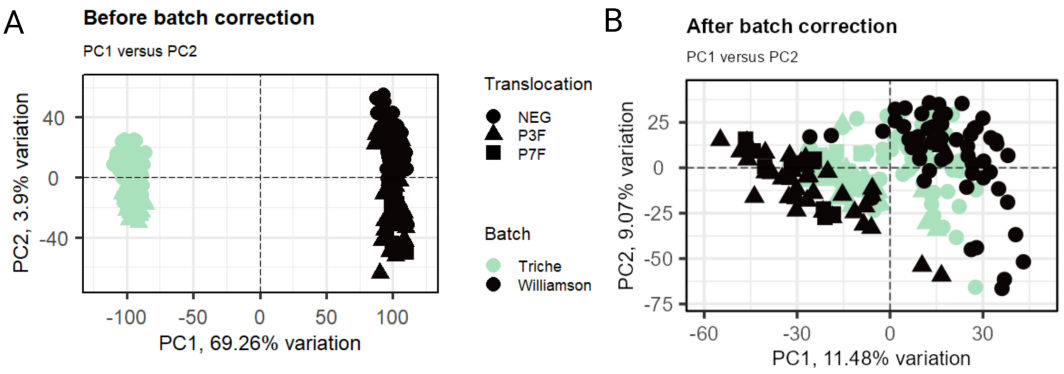


FIGURE 6.4: PCA biplots before and after batch effect correction. A) Before batch correction B) After batch correction. Batch effects were corrected for with the SVA package using ComBat function. Event was the outcome variable of interest and covariates were analysis age, histology, translocation, fusion status, tumour size, node status and sex. N= 166.

event were translocation, fusion status and histology, with *PAX3*, *PAX7*, FP and alveolar histology all associated with an increased risk of event.

The top ten features with the highest odds ratios and the top ten features with the lowest odds ratios were plotted. Features associated with a decreased risk had large error bars, indicating variability in their associated risk with event (Figure 6.5). The feature with the highest odds ratio was *PTK6*, a tyrosine kinase that has been shown to be involved in progression and metastasis of multiple cancers including breast and colorectal [400].

FIGURE 6.5: Odds ratio plot from logistic regression analysis using 'Event' as the outcome variable. Top 10 features with the highest odds ratio and top 10 features with the lowest odds ratio were plotted. Features with a significant association with 'Event' were defined as having a p value <0.01. Error bars represent 95% confidence intervals (CI). N=166.

GSEA was conducted on the genes significantly associated with an increased or decreased risk of event. Results for genes associated with an increased risk of event included neuronal related terms (Figure 6.6A), suggesting higher expression of

neuronal-related genes is associated with an increased risk of event. Other terms may be related to cell movement/cell-ECM binding, including 'cell cortex', 'cell substrate junction' and 'focal adhesion' (Figure 6.6A). Terms enriched in the genes associated with a decreased risk of event were ribosomal-related (Figure 6.6B), indicating that higher expression of ribosomal-related genes is associated with a decreased risk of event.

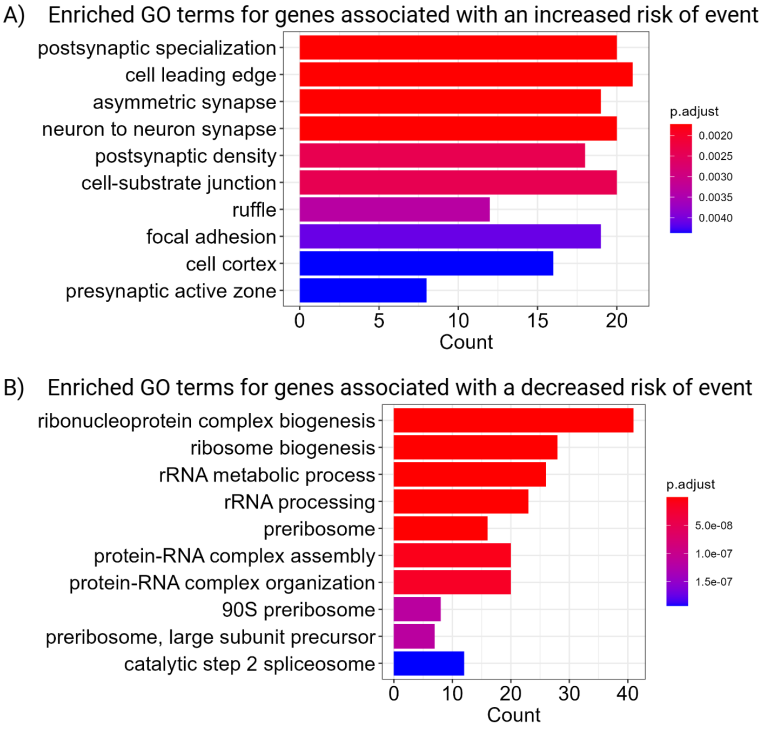


FIGURE 6.6: GSEA of features associated with event in univariate logistic regression. Enriched GO terms for genes associated with A) An increased risk of event B) A decreased risk of event. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). Top 10 gene sets with the lowest adjusted p values are displayed. Features with a significant association with event were defined as having a p value <0.01. N=166.

6.3.3 Identifying the optimum features with recursive feature elimination

RFE was performed on the 608 features to filter the features used to train the machine learning models. The full table of RFE results can be found in Supplementary Files (Supplementary Files/Chapter 6 Machine learning/rfe_results.csv). The model accuracy increases significantly from 0 to ~60 features after which accuracy fluctuates within the range of 0.65 to 0.72 (Figure 6.7). Model performance decreases slightly after around 60 features and then begins to steadily increase. The number of features with the highest accuracy was 518 features with an accuracy of 0.72 (Figure 6.7). To reduce the risk of overfitting, it is recommended to select fewer features than the number of samples. Therefore, 45 variables were selected to train the model, as this

offers a more appropriate feature-to-sample ratio while maintaining strong model performance (accuracy = 0.67, kappa = 0.31).

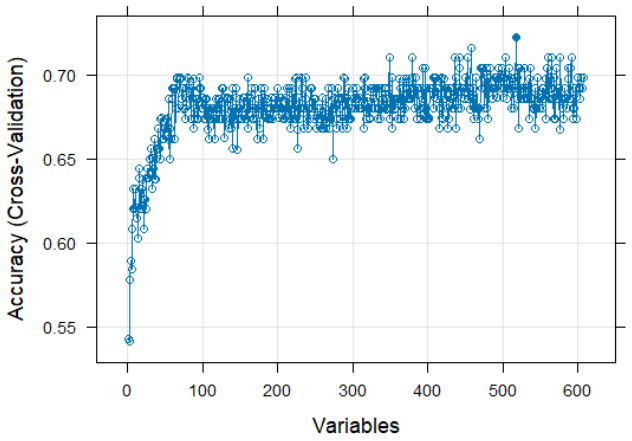


FIGURE 6.7: RFE plot of number of features and model accuracy. Random forest model with 5-fold cross validation was used for RFE with 608 features.

6.3.4 Training the random forest model

A random forest model was trained using the 45 features identified through RFE. Three sampling methods were used; upsampling, downsampling and standard sampling. The number for mtry that resulted in the best model accuracy was selected (Supplementary Figure 50).

ROC curves of the models based on their 5-fold cross-validation results were plotted and their performance was assessed using the AUC. A higher AUC represents better model performance, with a value of 1 being perfect and 0.5 being random chance. An AUC above 0.8 is considered a benchmark for clinical utility [401]. The AUC was 0.809 for standard sampling, and 0.811 for upsampling and 0.807 for downsampling (Figure 6.8A). Downsampling had the highest sensitivity at 0.71 while standard sampling had the highest specificity at 0.84 (Figure 6.8B).

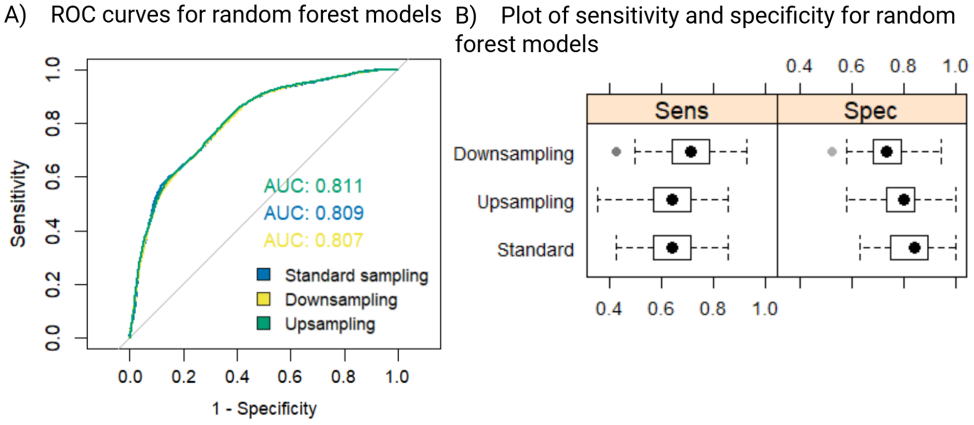


FIGURE 6.8: Random forest model performance from the 5-fold cross validation results in model training of the different sampling methods. Random forest models were generated using standard sampling, downsampling and upsampling. A) ROC curve B) Sensitivity and specificity plots. Models were trained on 45 features using 166 samples.

6.3.5 Training the XGBoost model

As with the random forest model, an XGBoost model was trained on the 45 features using standard sampling, upsampling and downsampling. The optimal hyperparameters were selected. The XGBoost model that used upsampling had the best AUC of 0.750 (Figure 6.9A). AUC of the XGBoost models for standard sampling and downsampling were lower at 0.718 and 0.711 respectively (Figure 6.9A). The AUC values for all XGBoost models were lower than the AUC for random forest models. Upsampling and downsampling achieved the highest sensitivity at 0.71 and standard sampling had the highest specificity at ∼0.84 (Figure 6.9B).

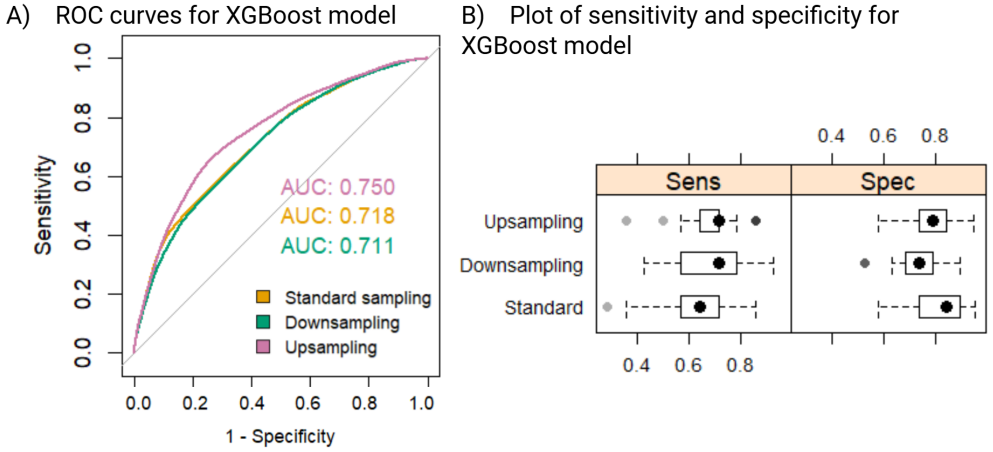


FIGURE 6.9: XGBoost model performance from the 5-fold cross validation results in model training of the different sampling methods. XGBoost models were generated using standard sampling, downsampling and upsampling. A) ROC curve B) Plots of sensitivity and specificity. Models were trained on 45 features using 166 samples.

6.3.6 Evaluating model performance on the test data

Model performance was evaluated on the test data; a subset of 40 samples from the original data that had been withheld from model training. The models were used to predict the probability of class 'Event' and 'No event' for the test data. The classification threshold for a sample to be classified as an event was optimised using ROC curves for Youden's J (sensitivity + specificity - 1) (Table 6.1).

For the random forest models, all models had an AUC considered 'fair' [402]. Upsampling had the the best AUC out of all the models at 0.788, followed by standard sampling at 0.783 and downsampling at 0.777 (Figure 6.10A). The XGBoost model with downsampling had the highest AUC of all the XGBoost models at 0.777 (Figure 6.10B). This was followed by upsampling and standard sampling, with AUCs of 0.765 and 0.734 respectively (Figure 6.10B). All XGBoost models has an AUC considered 'fair' [402].

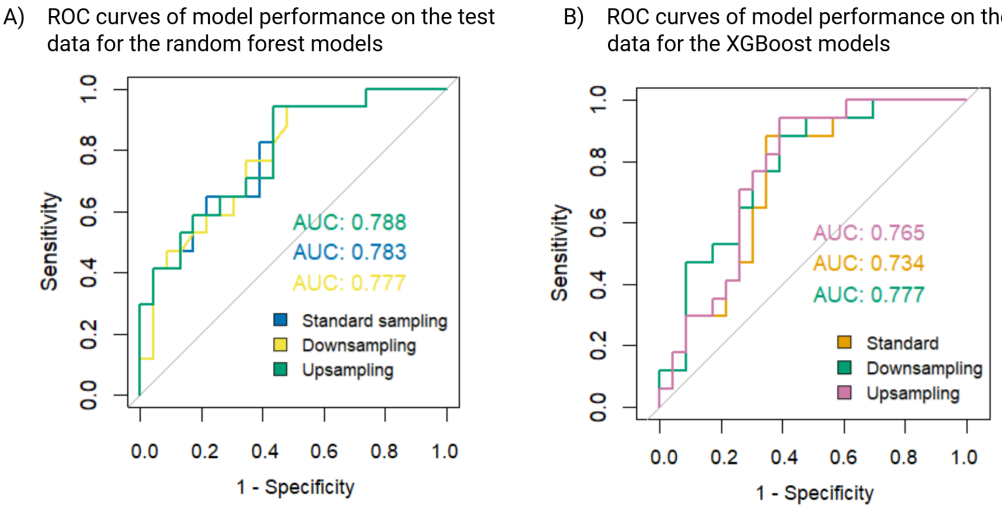


FIGURE 6.10: ROC curves showing model performance on the test dataset. Models were generated using standard sampling, downsampling and upsampling. A) ROC curves for random forest models B) ROC curves for XGBoost models. Models were trained on 45 features using 166 samples and tested on 40 samples.

In addition to AUC, other performance metrics were used to evaluate model performance. The balanced accuracy is the average of the true positive rate (sensitivity) and true negative rate (specificity), and provides a better measure compared to standard accuracy when there are class imbalances in the data. The XGBoost model with standard and up-sampling had the best balanced accuracy of 0.77 (Table 6.1). Kappa refers to Cohen's kappa, a statistic to measure agreement of classification that factors out agreement that may occur due to chance. The highest Kappa value was XGBoost model with upsampling (Table 6.1), with a kappa of 0.52 which corresponds to a moderate agreement. McNemar's test is used to assess whether there is a statistically significant difference in the types of classification errors made by a model (whether it tends to misclassify one class more often than the other). A p-value <0.05 indicates that the error rates for the two types of misclassification are significantly different. All of the random forest models, as well as the XGBoost upsampling model was biased toward better prediction of one class over the other (Table 6.1). All of these models better predict 'No event', as indicated by having a higher Negative Predictive Value (NPV) than Positive Predictive Value (PPV). Random forest models and XGBoost upsampling had the highest sensitivity of 0.94, and XGBoost standard sampling had the highest specificity of 0.65 (Table 6.1). Taking the results together, the XGBoost model with upsampling is the best performing model as it has the highest balanced accuracy, kappa, sensitivity and NPV.

A confusion matrix can be used to compare model predictions with true labels. From the confusion matrix for the XGBoost upsampling model, it can be seen that the model correctly predicted 16 events and incorrectly predicted 9 events (Table 6.2). There was only 1 event misclassified as no event and 14 no events correctly predicted (Table 6.2).

TABLE 6.1: Model performance metrics on the test data for the XGBoost and random forest models. Thres= classification threshold, acc= accuracy, sens= sensitivity, PPV= positive predictive value, NPV= negative predictive value, prev= prevalence, dect rate = detection rate, detect prev= detection prevalence.

Model	Sampling type	Thres	Balanced acc	Kappa	Mcnemar's P-Value	Sens	Spec	PPV	NPV	Prev	Detect rate	Detect prev
Random Forest	Standard	0.32	0.75	0.47	0.02	0.94	0.57	0.62	0.93	0.43	0.40	0.65
	Down	0.39	0.73	0.43	0.01	0.94	0.52	0.59	0.92	0.43	0.40	0.68
	Up	0.36	0.75	0.47	0.02	0.94	0.57	0.62	0.93	0.43	0.40	0.65
XGBoost	Standard	0.34	0.77	0.51	0.11	0.88	0.65	0.65	0.88	0.43	0.38	0.58
	Down	0.39	0.75	0.47	0.07	0.88	0.61	0.63	0.88	0.43	0.38	0.60
	Up	0.34	0.77	0.52	0.03	0.94	0.61	0.64	0.93	0.43	0.40	0.63

TABLE 6.2: Confusion matrix comparing XGBoost model predictions with true labels. Reference represents the true labels and prediction represents predictions from the model.

		Reference	
		Fail	Not fail
Prediction	Fail	16	9
	Not fail	1	14

6.3.7 Feature importance

The features that contribute most to the XGBoost model with upsampling were identified. The full list of features and their importance values can be found in Supplementary Files (Supplementary Files/Chapter 6 Machine learning/important_features.csv). The most important feature in the model was *CYTH2*, encoding for a protein involved in protein sorting and membrane trafficking. The second most important variable was *EFNA1*, which encodes for a receptor protein-tyrosine kinase involved in embryonic development.

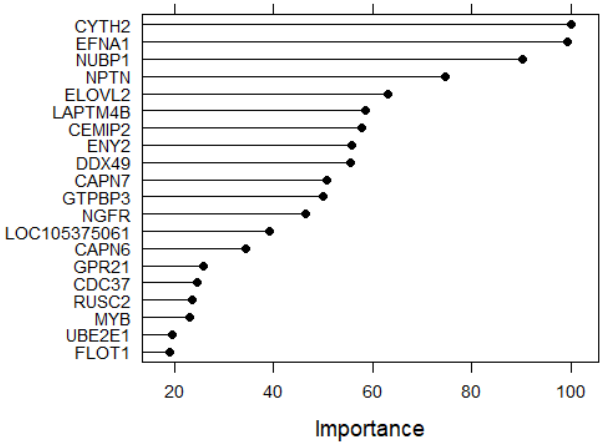


FIGURE 6.11: Feature importance plot from the XGBoost model for the top 20 most important features. Importance values have been scaled from 0 to 100.

The top 20 features were plotted in an odds ratio plot to determine whether the features were associated with an increased or decreased risk of event (Figure 6.12). *GPR21*, encoding for a G-protein-coupled receptor, had the highest odds ratio at ∼6 (Figure 6.12). *LOC105375061* had the lowest odds ratio, followed by *DDX49*, encoding for a protein that regulates mRNA export and pre-ribosomal RNA levels [403].

FIGURE 6.12: Odds ratio plot from logistic regression analysis using 'Event' as the outcome variable, showing the top 20 most important features identified by the XGBoost model.

6.4 Discussion

Using patient microarray data and clinical data to train machine learning models, this work created an XGBoost model that can predict whether a patient will have an event based on 45 features representing gene expression profiles and clinical risk factors. The model was used to classify 40 samples that had been withheld from model training to assess the model performance on an independent dataset. The XGBoost model could predict whether the patient had an event with ∼78% accuracy, which is 2% below the threshold clinical use [401]. This work suggests that machine learning can identify patients at high-risk of an event occurring even from pre-treatment data. This could prioritise high-risk patients for clinical trials and alternative treatment options. Based on the genes used to train this model and important features in the model, novel insights into potential biological pathways and processes important in relapse in RMS have also been identified. These insights may lead to the identification of drugs that can target these processes, reducing the risk of relapse.

Logistic regression analysis on the gene expression data and clinical variables resulted in features associated with the risk of an event occurring. Of these seven features, FP-RMS, alveolar histology and *PAX3/PAX7-FOXO1* translocation were the only clinical features associated an event. Association with translocations is expected as

multiple studies have shown FP-RMS have a worse FFS than FN-RMS [86, 404, 405]. Previous studies have found ARMS are more likely to relapse than ERMS [345]. Other clinical variables that have been found to be independently associated with FFS include tumour invasiveness, fusion status, IRS group, primary tumour size and age at diagnosis interacting with tumour size [86, 378], tumour invasiveness [86], tumour stage [379] and primary tumour site [86, 378], could not be input as features in the model. This is due to multiple reasons, such as unclear labelling of clinical variable (e.g. unclear if stage refers to T stage or SIOP stage), the clinical variable being missing from one of the datasets or the variables being coded differently in datasets (e.g. tumour location coded as 'Forearm' in the Triche dataset would be coded as 'limb' in the Williamson dataset). The latter could be addressed by the harmonisation of clinical coding, however this would have been impractical due to the labour-intentive and time consuming nature of the process for 206 samples. This is a limitation of the model, as clinical variables potentially important in relapse have not been included. If these additional clinical variables could be input as predictors in the model, it could improve model performance. This highlights a challenge in using clinical data to train machine learning models, as much clinical information is lost when trying to combine these datasets. To improve this there should be consistency in the type, labelling and encoding of clinical variables, to allow for unimpaired merging of clinical datasets.

Another limitation relating to the clinical data was addressing missing values. To include a clinical variable for RFE or as a predictor in the training of a machine learning model with the caret package, there must be no missing values. To resolve this, imputation was used to assign missing values an estimated value. This is a limitation, as the imputed values are only estimates and may not reflect what the true value should be. If these imputed values are wrong, clinical variables that should be significant after logistic regression may not be significant and would not be input as predictors for model training. This may adversely affect model performance. However, the alternative to imputation would be to remove samples with missing values. Alternatively, a different package could be used for model training instead of caret that is able to handle missing values. For XGBoost, other R packages exist [406], however for random forest models often other packages also rely on imputation [407, 408].

Findings from the logistic regression analysis complement results from previous chapters and the literature. Cell-ECM binding terms including 'focal adhesion', 'cell cortex' and 'cell substrate junction' were significantly enriched among genes associated with an increased risk of event. These gene sets were also enriched in modules associated with intrinsic resistance and FP-acquired resistance from Chapter 4. D'Agostino et al. showed *in vitro* and *in vivo* that RMS cells are directly involved in ECM remodelling [409]. It is possible that enhanced cell-ECM binding may be a mechanism of relapse in RMS, particularly FP-RMS.

A machine learning approach may be useful for identifying novel therapeutic targets by investigating features that show high importance within the model. CEMIP2 is involved in the degradation of hyaluronic acid, a non-protein component of the ECM. CEMIP2 regulates cell adhesion and migration through the degradation of hyaluronic acid at focal adhesion sites [410]. Another study has also found upregulation of focal adhesion-related GSEA terms in glioma patients expressing CEMIP2 [411]. Since these terms were also enriched in the genes associated with an increased risk of event, perhaps high *CEMIP2* expression may play a role in relapse through the regulation of RMS cell adhesion and migration. Studies have identified *CEMIP2* to be associated with resistance and prognosis in other cancer types. *CEMIP2* has previously been reported to confer resistance to telozamide in glioblastoma patients [412]. Expression of *CEMIP2* has also been seen to have prognostic value in pancreatic ductal adenocarcinoma [413]. The role of *CEMIP2* in RMS has not been investigated, but the results from this chapter suggest *CEMIP2* could be a potential target for reducing risk of relapse. Recent research by Srivastava et al. have identified a potential small molecule inhibitor of CEMIP2 [414]. Other potential targets include *EFNA1*, identified as one of the top-ranked features in the model that also has a potential small molecule inhibitor [415]. Upregulation of EFNA1 has been shown to be involved in the migration and invasiveness of ERMS [416].

The results from this chapter suggest that a neuronal phenotype may be associated with relapse events in RMS. Genes associated with an increased risk of relapse were enriched for neuronal-related terms, and some of the most important features had functions relating to neurite extension, including *CYTH2*, *EFNA1* and *NPTN*. Similar neuronal-related terms were also enriched in FP resistant cell models described in 4. In the intrinsic resistant model, C7 showed upregulation of the term 'protein localisation to axon' compared to the parent cell line. Neuronal gene sets were enriched for in FP-acquired resistance modules. A neuronal cell state has also been observed in FP tumours from single-cell RNA-seq data and was reported to be significantly enriched for in chemotherapy-treated patients [318]. This neuronal cell state was seen to be missing or extremely rare in Patient-Derived Xenograft (PDX) and cell line data [318]. These findings suggest that these neuronal-like RMS cells may be present in tumours before chemotherapy treatment and could be linked to relapse. To validate these findings, the single-cell RNA sequencing data could be used to deconvolute the resistant cell models to identify the different cell states in the cell lines. Statistical comparisons could then be made between the resistant and control samples, to assess whether the resistant samples have a higher proportion of these neuronal-like RMS cells.

Genes associated with a decreased risk of event showed enrichment of ribosomal-related terms, suggesting higher expression of ribosomal-related genes may be protective again relapse in RMS. This is supported from Chapter 4, where modules

negatively correlated with resistance showed enrichment of ribosomal gene sets. This was found in both intrinsic, FP and FN resistance models. Further work should be conducted to provide more evidence in this area. One possibility would be to analyse publicly available data to compare cells from patients experiencing relapse/metastasise with those who do not. Although this is challenging due to the lack of datasets that contain clinical information on relapse/metastasis in RMS.

While other feature selection methods could be used to identify predictors for the models, for this analysis logistic regression was selected. Regardless of the feature selection method, there is always a risk of omitting important genes. The results from this chapter support the use of logistic regression for feature selection, as the resulting genes associated with event are inline with previous work in resistant models as well as the literature.

A major limitation of this work is that the test data used to assess model performance had been altered when adjusting for batch effects. This may have impacted the model performance, either positively or negatively. Ideally, the test data would be a completely independent dataset. An alternative approach would have been to treat the datasets independently using one to train the model and the other to test. This would have resulted in equal sized training and test datasets each containing around 100 samples. When machine learning models were trained using this reduced number of samples, the models showed significantly poorer performance. In machine learning, it is common to have a training and test ratio of 80:20 [417], therefore it was decided to combine the datasets and split using this ratio.

The machine learning model with the best performance is appropriate for its intended purpose, to identify patients at risk of relapse. In the context of relapse, the cost of a false negative (predicting someone won't relapse when they do) is much higher than the cost of a false positive (predicting relapse when they don't). This should be reflected in the model, with the model favouring sensitivity over specificity. The XGBoost model with upsampling had a sensitivity of 0.94, indicating it correctly identified 94% of patients that had an event. However, there was one false negative, indicating that the model misclassified one patient who experienced failure as not having failed. To ensure clinical utility, the model must reduce false negatives to zero, as it is critical that no patients at risk of failure are misclassified as not failing. This could potentially be achieved by training the model on a larger dataset or adjusting the models classification threshold.

In summary, this chapter has created a machine learning model capable of identifying patients at risk of relapse from pre-treatment expression and clinical data. Testing the model performance on a test dataset suggests the model is suitable for it's purpose as it is able to identify 94% of patients that have an event correctly and has a low number of false negatives. This work provides further insight into genes and mechanisms

potentially involved in RMS relapse. The genes identified through this machine learning model should be investigated as potential targets that could reduce the risk of relapse/progression in RMS. These findings complement results from previous chapters and the literature, with neuronal-related genes found to be associated with an increased risk of an event and ribosomal-related genes associated with a decreased risk. Further studies are needed to validate the findings presented in this chapter.

Chapter 7

Discussion

Although rare, patients with high-risk NB and RMS have extremely poor outcomes highlighting the need for further research to improve patient outcomes. This thesis explored the use of RNA-seq analysis to investigate therapeutic strategies in these high-risk paediatric cancers.

High-risk NB, mainly characterised by the presence of *MYCN* amplification or metastatic disease, has a 5-year OS rate of ∼50%[22]. New targeted therapies for NB are urgently needed. One class of drugs currently being explored is EZH2 inhibitors (EZH2is), which have been shown to promote differentiation in NB [185, 219, 220, 44] and restore sensitivity to anti-GD2 immunotherapy [44]. However, the underlying MoAs of EZH2is and how they might interact with drugs in clinical use has not been fully explored. Chapter 3 used RNA-seq analysis to investigate the MoAs of current and potential targeted therapies (EZH2is and isotretinoin), both alone and in combination.

The second part of this thesis focused on chemotherapy-resistance in RMS. In RMS, most patients initially respond to chemotherapy [105] however a significant proportion experience relapse or metastasis [106], after which the tumour no longer responds to chemotherapy and survival rates are poor. The reasons for chemotherapy failure remain poorly understood. Chapter 4 aimed to gain further understanding of chemotherapy resistance in RMS through the use of resistant cell models that represent acquired and intrinsic resistance. In addition, this thesis aimed to identify a gene signature of chemotherapy resistance from the cell models using WGCNA and PPI networks. The gene signatures were tested for association with clinical traits in patient microarray data and a machine learning approach was applied to generate a model that could predict relapse.

7.0.1 Uncovering mechanisms of action of targeted therapeutics in neuroblastoma

Chapter 3 uncovered potential mechanisms of action of targeted therapies, including EZH2is, isotretinoin and combination therapy in NB. Consistent with the literature [185, 219, 220, 44], results support the hypothesis that EZH2is promote differentiation of NB cells. In addition, there was also evidence to suggest EZH2is may upregulate the immune response. Findings from the literature on the effect of EZH2is on the immune response in NB are conflicting [44, 273, 274]. The work from Chapter 3 may provide some further clarity to this research area and support studies that have observed an immune modulatory effect of EZH2 on the immune system in NB. This may reveal opportunities for EZH2is to be used to prime NB tumours with immunotherapy treatment, which have been unsuccessful so far [418].

Results from Chapter 3 showing immune gene sets upregulated by EZH2is are absent in combination therapy suggest isotretinoin may dampen this potential immune activation effect of EZH2is. However, since this work is based on MCTSs that lack immune cells, limited conclusions can be drawn from these results. To gain further understanding of the effects of these targeted therapies on the immune TME in NB, models that encompass both tumour and immune cells are essential. This could include the use of co-culture models of tumour and immune cells which can be developed in 2D or 3D, which have previously been used to investigate the effects of anti-GD2 immunotherapy [419]. Alternatively, murine models with a competent immune system could be used, such as the TH-MYCN murine model or humanised mice. The TH-MYCN transgenic murine model overexpress MYC-N and have a murine immune system whilst humanised mice have been engineered to have a human immune system. Of these, humanised mice models are most representative of patients as they have a human immune system. Although, both models lack tumour stromal cells such as CAFs, which have been shown to be involved in NB development and growth [420, 421]. This is a major limitation of these models. Conducting further work exploring the effect of EZH2is and isotretinoin on the immune response in NB using co-culture or immune-competent models would be essential to give greater confidence in the results from the MCTS models in Chapter 3.

Despite MCTS models lacking immune cells, RNA-seq data may still provide insight into tumour-cell intrinsic immune suppression. This can include mechanisms such as decreasing the expression of ligands involved in immune recognition, increasing the expression of immune inhibitory ligands, secreting immune suppressive factors and altering antigen processing and presentation [422]. The results from Chapter 3 suggested several of these mechanisms may be involved in NB cells response to targetted therapies. Treatment with EZH2is showed an enrichment in terms related to MHC-I signalling, antigen presentation, cytokine production and immune cell

chemotaxis, suggesting a reduction in tumour-cell intrinsic immune suppression. Conversely, treatment with isotretinoin showed mechanisms related to tumour-cell intrinsic immune suppression including downregulation of chemokines, antigen processing and presentation and tolerance induction to self-antigens. Although this can provide some insight into the effect of treatments on the immune response via tumour cells, immune-competent models would be required to get a comprehensive understanding of treatment effects on the immune response.

Beyond model selection, fully investigating the immune effects of these therapies requires alternative technology. With bulk RNA-seq it is not possible to determine the cell type the transcripts originate from. To enable a more detailed analysis, single-cell RNA-seq could be used. This technology would allow the characterisation and quantification of immune cell populations that could be compared between control and treated conditions. Alternatively, spatial transcriptomics could provide additional detail to single-cell RNA-seq, as spatial information of immune cell populations would also be observed. Although this technology is expensive, primary Formalin-Fixed Paraffin-Embedded (FFPE) RMS samples required for the analysis could be obtained through the VIVO biobank. This could provide information such as cell-cell interactions and immune cell infiltration that may be more relevant to patient outcomes. For example, single-cell RNA-seq analysis may show an increase in the number of immune cells with EZH2i treatment suggesting treatment enhances the immune response, but this may not necessarily correlate with a better outcome if these immune cells are unable to infiltrate into the tumour. This may be particularly relevant in NB, as research suggests high-risk NB tumours are immune cold, meaning immune cells are excluded from the tumour [423]. Using this technology in combination with models that recapitulate the immune TME of NB would likely provide a more comprehensive understanding of the role these therapeutics play in the immune response. Investigating the effect of isotretinoin on the immune response in NB should be highlighted as a key area of research, as this could have direct implications due to the current clinical use of isotretinoin.

Although treatment with isotretinoin aims to promote differentiation of remaining residual disease [34], relapse rates remain high [16], suggesting it is not completely effective. This has led research to focus on enhancing the effects of differentiation therapy through combination with other drugs. One therapeutic option being explored is the combination of retinoic acids with Histone Deacetylase (HDAC) inhibitors. The binding of retinoic acids to retinoic acid receptors recruits co-activator complexes via histone acetyltransferase activity, resulting in transcriptional activation [424]. It was hypothesised that HDAC inhibitors would enhance this effect by preventing the removal of acetyl groups [425]. HDAC inhibitors and ATRA were shown to be synergistic in *in vitro* and *in vivo* xenograft models [425, 426] and progressed to phase I clinical trials in NB [427]. In addition to this, Cyclin-Dependent

Kinase (CDK) inhibitors have also been explored as a therapeutic option to enhance NB differentiation. Recently the CDK inhibitor palbociclib was shown to promote neuronal differentiation of adrenergic neuroblastoma cell lines and inhibit tumour growth in TH-MYCN mice [237]. It was proposed that for NB cells to differentiate, they require resetting of the oncogenic core regulatory circuit (CRC) and cell cycle arrest [237]. Although palbociclib was shown to induce differentiation, results suggested complete cell-cycle arrest was not achieved and expression levels of adrenergic CRC transcription factors were not impacted. It was hypothesised that the combination of CDK inhibitors with retinoic acid, which have been shown to reset the adrenergic CRC [227], would drive cells into a more differentiated state [237]. The authors saw further inhibition of cell cycling, reduced proliferation and enhanced differentiation [237]. Together, this research shows the potential to enhance differentiation therapy in NB by understanding its mechanism of action and combining it with other therapeutic strategies.

A limitation of using bulk RNA-seq in Chapter 3 is that this sequencing technology does not accurately enable the identification of alternatively spliced transcripts. Alternative splicing, where a gene can produce multiple mRNA transcripts and proteins, can play an important role in cancer development and progression [428]. Reads from bulk RNA-seq are short (around 100-150 base pairs in length) and can align to multiple isoforms of the same gene, making it difficult to accurately detect alternatively spliced transcripts. Several bioinformatic tools exist that aim to predict alternative splicing events using statistical methods [429, 430, 431]. These tools may provide some insights into potential alternatively spliced transcripts, however they are only predictive and may not be accurate. Long-read sequencing can be used to directly detect alternative splicing as the increased length of reads allows for accurate mapping of reads to different gene isoforms. The detection of alternative splicing events may be particularly important in the context of this work, as EZH2 is a transcriptional repressor that may influence alternative splicing. In chronic myeloid leukaemia, EZH2 was found to exert part of its oncogenic effect through modulating mRNA splicing [432]. Potential impacts of EZH2is on alternative splicing in NB may have been missed in the analysis conducted in Chapter 3. This could be resolved through the use of bioinformatic tools to predict splicing events or through using an alternative sequencing technology such as long-read sequencing.

7.0.2 Using cell models to investigate chemotherapy-resistance in rhabdomyosarcoma

The work in Chapter 4 showed some of the challenges in access to availability of data in RMS research. As discussed in Chapter 4, the ideal datasets for testing whether the gene signatures from cell models correlated with clinical traits were not used due to

limited availability of data. For example, when checking for association between the acquired resistance signatures and clinical traits, a post-chemotherapy dataset would have been ideal. This difference made it difficult to determine if the lack of association between the cell model gene signatures and clinical traits was due to biological differences in resistance mechanisms of cell models and clinical resistance in patients or suitability of the validation dataset. Ideally, further work would be performed to identify cells that show high expression of the intrinsic resistance signature in patient single-cell RNA-seq data and assess their correlation with relapse or metastasise. Suitable datasets for this would be from Danielli et al. [318] or Demartino et al. [433]. The acquired resistance signatures would ideally be compared between primary and relapse samples, such as the dataset from Danielli et al. which contains diagnostic and recurrent samples [318]. Another challenge is that, due to the rarity of RMS, available datasets often have small sample sizes, which can limit statistical power and result in non-significant results. For example, in Chapter 4, FP gene signature scores appeared higher in mid-treatment patient samples but these results were not statistically significant as there were only two patients. This highlights how RMS research could be improved by the sharing of available datasets that have full clinical annotation.

One of the limitations of using RNA-seq for the cell models was that this only investigated transcriptional variation that could contribute to resistance and could not provide information at the genetic level. There is evidence that the RH30 cell line is heterogenous for *MYCN* amplification [187], and since the clones were derived from single cells of the RH30 cell line they may or may not have *MYCN* amplification. *MYCN* amplification, exclusive to FP-RMS, has been reported to be associated with worse survival [434]. It is possible that *MYCN* amplification could have contributed to the resistance phenotype seen in the clones rather than the transcriptional differences that were identified through bulk RNA-seq. Fluorescence in situ hybridization, whole genome sequencing or whole exome sequencing could be used to determine whether the clones have *MYCN* amplification. There could also be genetic variation in other genes that could contribute to resistance, as well as variation in the expression of the fusion gene.

A key consideration highlighted from this thesis is whether cell models of resistance are clinically relevant. In rare cancers such as RMS, patient samples can be difficult to obtain and thus cell models play an important role in research. Several cell models of resistance were used in this research and all provided some understanding of resistance in different contexts. Although the acquired and intrinsic resistance models did not show a strong correlation in gene expression with mid-treatment patient samples, some shared mechanisms with relapse in the patient microarray data were observed. The high-dose model more closely represented mid-treatment patient samples than any of the other models, suggesting it is a useful model in this context and could suggest the presence of an intrinsically resistant clone. Although the work

conducted in the cell models revealed some possible mechanisms of resistance, it is difficult to attribute whether intrinsic, acquired resistance or both are ultimately responsible for resistance in RMS. Since the high-dose model was the most similar to mid-treatment patient samples and relapse could be predicted from pre-treatment data, this could suggest that resistance could be intrinsic. Deconvolution of the bulk RNA-seq data from resistant cell models for RMS metaprograms (identified by Danielli et al. [318]) could inform whether cell lines recapitulate the transcriptional states that exist within RMS tumours. If the three metaprograms are present in the cell lines, this could suggest that transcriptionally they resemble patient tumours and may be suitable models for RMS.

Some of the findings from the intrinsic and acquired cell models overlapped with results from the analysis of patient data. These similarities were seen between pre-treatment microarray data, intrinsic resistance models and chemotherapy-treated resistant models, indicating there may be some mechanisms of resistance in RMS that are more universal and can be detected regardless of resistance mode (intrinsic or acquired) or sample type (cell model or patient). This could suggest that cell models are able to identify some mechanisms of resistance that relate to resistance as it presents clinically in RMS. This should be investigated by identifying whether these mechanisms and phenotypes can also be seen in post-chemotherapy or relapse patient samples. This also raises a key question as to whether these resistant cells exist in the same state before, during and after chemotherapy or whether they are able to shift between phenotypes. Some of these points could be further investigated through the analysis of spatial transcriptomic data from pre-treatment, mid-treatment and relapse patient samples.

The results of Chapter 6 provided some information on potential resistance mechanisms and were found to overlap with the findings of previous chapters. Interestingly, cell-ECM binding and neuronal terms were found to be enriched for in intrinsic and FP-acquired resistant models as well as being enriched for in genes that were associated with an increased risk of an event from Chapter 6. Further work should be conducted investigating these as potential mechanisms of resistance, particularly the neuronal phenotype which has been seen in another study and linked to chemotherapy treatment [318]. One option for further analysis would be to use the single-cell RNA-seq data that characterises RMS cell phenotypes from the Danielli et al. study [318] to deconvolute the resistant cell models. It is hypothesised that the resistant models would have a higher proportion of cells that have the neuronal phenotype when compared to the non-resistant cells. This may provide some further understanding on the association between the neuronal phenotype and resistance in RMS. To more comprehensively investigate whether and how the neuronal cell state is involved in RMS chemotherapy resistance, spatial transcriptomics could be used. Since Danielli et al. [318] have already characterised the neuronal cell state in

single-cell RNA-seq data, this could be applied to spatial transcriptomic data to identify populations of cells that display this neuronal cell phenotype. Ideally, spatial transcriptomics would be conducted on pre-treatment, mid-treatment and relapse patient samples. This could be used to determine whether cells that have the neuronal phenotype are present in patients prior to treatment (as suggested by results from Chapter 6) and whether they are present in all patient or only those who relapse. It may also provide insights into their interactions with immune cells and this could be compared between patients with and without relapse to investigate whether interactions with immune cells contribute to chemotherapy resistance. This approach could be used to observe the direct effects of chemotherapy on RMS cells and investigate whether neuronal-like RMS cells remain after treatment. This technology may enable identification of dependant pathways or mechanisms in these cells that could be used to target them.

7.0.3 Using machine learning models to predicting events in patient microarray data

The final results chapter of this thesis created a machine learning model to predict which patients will have an event based on pre-treatment patient microarray and clinical data. A model was developed that had an AUC of 0.78, close to the considered benchmark for clinical utility. Although, as this model was developed and validated on a small sample size and the validation was not conducted in a completely independent dataset, further validation in a larger well annotated cohort would be required before it's clinical application. This work has provided a basis on which further research can be conducted for the potential clinical application of this model to identify patients at risk of relapse so they can be prioritised for clinical trials.

Although the model already has good accuracy, it was developed using microarray data which has been reported to be more noisy than other sequencing technologies such as RNA-seq [165]. This machine learning approach could easily be translated into use with an alternative training dataset, such as bulk RNA-seq data, that may result in better model performance compared to the current model. Although single-cell RNA-seq would provide gene expression for individual cell types and cell type abundance information that may be useful for predicting prognosis, developing a machine learning model would be difficult due to the high-dimensionality of the data. To generate a dataset with a suitable number of samples, bulk RNA-seq may be a more practical approach, as it is more cost-effective and produces data that is less high-dimensional than single-cell RNA-seq. Ideally, a machine learning model would be trained on a large bulk RNA-seq dataset that contains detailed clinical annotation. This has potential to accurately predict which patients will relapse and could support findings of potential mechanisms and targets from Chapter 6.

A key point raised as a result of this work was the lack of consistency in the clinical data recorded from different sources. This led to many clinical variables being excluded from machine learning model training, that have previously been reported to be important prognostically in RMS. This could be resolved through the standardisation of the variables recorded in clinical RMS data and how they are encoded. For rare cancers such as RMS, data is collected across Europe within different studies and clinical trials. This can cause challenges in the cross-comparison of datasets, as different studies may record different variables depending on their research focus. Large clinical studies such as the Frontline and Relapsed Rhabdomyosarcoma (FaR-RMS) trial [84] and the European Proof-of-Concept Therapeutic Stratification Trial of Molecular Anomalies in Relapsed or Refractory Tumours (ESMART) trial [435] will help to address this issues. These studies will ensure the harmonised collection of data across Europe and conduct sequential sampling. This will likely provide a huge benefit for researchers analysing data and further improve the capabilities of machine learning models.

The ability of machine learning models to predict which patients relapse from pre-treatment gene expression data could suggest that intrinsic resistance may play a role in the failure of chemotherapy in RMS. Intrinsic resistance in RMS remains a largely unexplored area of research. Further research using different datasets may support the findings from the machine learning work. This area of research could reveal novel mechanisms and targets that could reduce the risk of relapse in RMS.

The work performed in Chapter 6 highlights the ability of machine learning to identify patterns in gene expression data that would be difficult or impossible to detect using standard analysis. When differential gene expression analysis was conducted comparing conditions 'Event' and 'No event', no differentially expressed genes were found. In addition to this, WGCNA revealed no modules that were significantly associated with event. Without machine learning, it is unlikely that a gene signature associated with relapse could be identified. Future research should utilise the ability of machine learning where standard analysis has not been successful. Ideally, this method would be applied to bulk RNA-seq data from a large cohort of patient samples with well detailed clinical annotation.

7.0.4 Conclusion

In conclusion, this thesis has highlighted the ability and limitations of using RNA-seq data for the investigative analysis of therapeutic strategies in the paediatric cancers NB and RMS. Some of the key findings from this thesis include the identification of enhancement of the immune response as a possible MoAs of EZH2is in NB. This provides some further information on this area of research that has shown conflicting results and may open opportunities for the use of EZH2is to be used in combination

with immunotherapies which have failed so far in NB. Another key finding from this work was the potential dampening of this immune activation by isotretinoin. This may have implications for a drug that is currently used clinically and warrants further investigation to either support or deny the results from this work. Further work in this area using different technologies in combination with models that recapitulate the immune microenvironment would allow for a more comprehensive understanding of the role EZH2is and isotretinoin might play in the immune response in NB. Results provided further support for the use of EZH2is in combination with isotretinoin for the differentiation of minimal residual disease in NB due to the increased neuronal differentiation observed in combination treatment compared to single-agent treatments. These findings support current research investigating EZH2i as potential therapeutic targets in NB and give information on how they may interact with other drugs in current clinical use. This thesis also investigated the use of traditional and more novel resistance cell models in modelling resistance in RMS. This has provided insights into whether cell models are useful to represent resistance as it presents clinically in RMS and which cell models may be used to represent resistance in different contexts. Potential mechanisms of resistance were identified that were observed in the analysis of cell models and patient data, including ECM-binding and a more resistant neuronal cell phenotype. This should be highlighted as a key area of future research to further investigate as targeting these cells may be a possible option to reduce relapse in RMS patients. It also discovered a potential enrichment of ribosomal genes with chemotherapy sensitivity that again was observed in both cell model and patient data. This could provide some new information on why some patients do not relapse. A machine learning approach identified key genes that could be used to predict relapse in pre-treatment patient microarray data that could be potential targets for preventing relapse in RMS. This work showed that machine learning models can be used to predict which patients will relapse with a good accuracy, which, with improvements, could be applied for clinical use.

References

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA Cancer J Clin*, vol. 73, no. 1, pp. 17–48, 2023.

[2] D. Shanmugavadivel, J. F. Liu, A. Ball-Gamble, A. Polanco, K. Vedhara, D. Walker, and S. Ojha, "The childhood cancer diagnosis (ccd) study: a uk observational study to describe referral pathways and quantify diagnostic intervals in children and young people with cancer," *BMJ Open*, vol. 12, no. 2, p. e058744, 2022.

[3] C. R. UK, "Cancer research uk," 2025.

[4] S. UK, "Soft tissue sarcoma incidence," 2025.

[5] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, and N. Lopez-Bigas, "A compendium of mutational cancer driver genes," *Nat Rev Cancer*, vol. 20, no. 10, pp. 555–572, 2020.

[6] J. R. Pon and M. A. Marra, "Driver and passenger mutations in cancer," *Annu Rev Pathol*, vol. 10, pp. 25–50, 2015.

[7] S. N. Gröbner, B. C. Worst, J. Weischenfeldt, I. Buchhalter, K. Kleinheinz, V. A. Rudneva, P. D. Johann, G. P. Balasubramanian, M. Segura-Wang, S. Brabetz, S. Bender, B. Hutter, D. Sturm, E. Pfaff, D. Hübschmann, G. Zipprich, M. Heinold, J. Eils, C. Lawerenz, S. Erkek, S. Lambo, S. Waszak, C. Blattmann, A. Borkhardt, M. Kuhlen, A. Eggert, S. Fulda, M. Gessler, J. Wegert, R. Kappler, D. Baumhoer, S. Burdach, R. Kirschner-Schwabe, U. Kontny, A. E. Kulozik, D. Lohmann, S. Hettmer, C. Eckert, S. Bielack, M. Nathrath, C. Niemeyer, G. H. Richter, J. Schulte, R. Siebert, F. Westermann, J. J. Molenaar, G. Vassal, H. Witt, B. Burkhardt, C. P. Kratz, O. Witt, C. M. van Tilburg, C. M. Kramm, G. Fleischhack, U. Dirksen, S. Rutkowski, M. Frühwald, K. von Hoff, S. Wolf, T. Klingebiel, E. Koscielniak, P. Landgraf, J. Koster, A. C. Resnick, J. Zhang, Y. Liu, X. Zhou, A. J. Waanders, D. A. Zwijnenburg, P. Raman, B. Brors, U. D. Weber, P. A. Northcott, K. W. Pajtler, M. Kool, R. M. Piro, J. O. Korbel, M. Schlesner, R. Eils, D. T. W. Jones, P. Lichter, L. Chavez, M. Zapatka, S. M.

Pfister, I. P.-S. Project, and I. M.-S. Project, "The landscape of genomic alterations across childhood cancers," *Nature*, vol. 555, no. 7696, pp. 321–327, 2018.

[8] X. Ma, Y. Liu, L. B. Alexandrov, M. N. Edmonson, C. Gawad, X. Zhou, Y. Li, M. C. Rusch, J. Easton, R. Huether, V. Gonzalez-Pena, M. R. Wilkinson, L. C. Hermida, S. Davis, E. Sioson, S. Pounds, X. Cao, R. E. Ries, Z. Wang, X. Chen, L. Dong, S. J. Diskin, M. A. Smith, J. M. Guidry Auvil, P. S. Meltzer, C. C. Lau, E. J. Perlman, J. M. Maris, S. Meshinchi, S. P. Hunger, D. S. Gerhard, and J. Zhang, "Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours," *Nature*, vol. 555, no. 7696, pp. 371–376, 2018.

[9] B. B. Campbell, N. Light, D. Fabrizio, M. Zatzman, F. Fuligni, R. de Borja, S. Davidson, M. Edwards, J. A. Elvin, K. P. Hodel, W. J. Zahurancik, Z. Suo, T. Lipman, K. Wimmer, C. P. Kratz, D. C. Bowers, T. W. Laetsch, G. P. Dunn, T. M. Johanns, M. R. Grimmer, I. V. Smirnov, V. Larouche, D. Samuel, A. Bronsema, M. Osborn, D. Stearns, P. Raman, K. A. Cole, P. B. Storm, M. Yalon, E. Opocher, G. Mason, G. A. Thomas, M. Sabel, B. George, D. S. Ziegler, S. Lindhorst, V. M. Issai, S. Constantini, H. Toledano, R. Elhasid, R. Farah, R. Dvir, P. Dirks, A. Huang, M. A. Galati, J. Chung, V. Ramaswamy, M. S. Irwin, M. Aronson, C. Durno, M. D. Taylor, G. Rechavi, J. M. Maris, E. Bouffet, C. Hawkins, J. F. Costello, M. S. Meyn, Z. F. Pursell, D. Malkin, U. Tabori, and A. Shlien, "Comprehensive analysis of hypermutation in human cancer," *Cell*, vol. 171, no. 5, pp. 1042–1056.e10, 2017.

[10] E. A. Sweet-Cordero and J. A. Biegel, "The genomic landscape of pediatric cancers: Implications for diagnosis and treatment," *Science*, vol. 363, no. 6432, pp. 1170–1175, 2019.

[11] A. V. Desai, A. L. Gilman, M. F. Ozkaynak, A. Naranjo, W. B. London, S. C. Tenney, M. Diccianni, J. A. Hank, M. T. Parisi, B. L. Shulkin, M. Smith, J. A. Moscow, H. Shimada, K. K. Matthay, S. L. Cohn, J. M. Maris, R. Bagatell, P. M. Sondel, J. R. Park, and A. L. Yu, "Outcomes following gd2-directed postconsolidation therapy for neuroblastoma after cessation of random assignment on anbl0032: A report from the children's oncology group," *J Clin Oncol*, vol. 40, no. 35, pp. 4107–4118, 2022.

[12] N. Shah, "Dodging the bullet: therapeutic resistance mechanisms in pediatric cancers," *Cancer Drug Resist*, vol. 2, no. 3, pp. 428–446, 2019.

[13] T. M. Dantonello, C. Int-Veen, A. Schuck, G. Seitz, I. Leuschner, M. Nathrath, P. G. Schlegel, U. Kontny, W. Behnisch, I. Veit-Friedrich, S. Kube, E. Hallmen, B. Kazanowska, R. Ladenstein, M. Paulussen, G. Ljungman, S. S. Bielack, T. Klingebiel, E. Koscielniak, and C. W. Studiengruppe, "Survival following disease recurrence of primary localized alveolar rhabdomyosarcoma," *Pediatr Blood Cancer*, vol. 60, no. 8, pp. 1267–73, 2013.

[14] L. Mascarenhas, E. R. Lyden, P. P. Breitfeld, D. O. Walterhouse, S. S. Donaldson, D. A. Rodeberg, D. M. Parham, J. R. Anderson, W. H. Meyer, and D. S. Hawkins, "Risk-based treatment for patients with first relapse or progression of rhabdomyosarcoma: A report from the children's oncology group," *Cancer*, vol. 125, no. 15, pp. 2602–2609, 2019.

[15] S. X. Skapek, A. Ferrari, A. A. Gupta, P. J. Lupo, E. Butler, J. Shipley, F. G. Barr, and D. S. Hawkins, "Rhabdomyosarcoma," *Nat Rev Dis Primers*, vol. 5, no. 1, p. 1, 2019.

[16] M. S. Irwin, A. Naranjo, F. F. Zhang, S. L. Cohn, W. B. London, J. M. Gastier-Foster, N. C. Ramirez, R. Pfau, S. Reshmi, E. Wagner, J. Nuchtern, S. Asgharzadeh, H. Shimada, J. M. Maris, R. Bagatell, J. R. Park, and M. D. Hogarty, "Revised neuroblastoma risk classification system: A report from the children's oncology group," *J Clin Oncol*, vol. 39, no. 29, pp. 3229–3241, 2021.

[17] E. Hibbitts, Y. Y. Chi, D. S. Hawkins, F. G. Barr, J. A. Bradley, R. Dasgupta, W. H. Meyer, D. A. Rodeberg, E. R. Rudzinski, S. L. Spunt, S. X. Skapek, S. L. Wolden, and C. A. S. Arndt, "Refinement of risk stratification for childhood rhabdomyosarcoma using foxo1 fusion status in addition to established clinical outcome predictors: A report from the children's oncology group," *Cancer Med*, vol. 8, no. 14, pp. 6437–6448, 2019.

[18] S. Agarwal, "Pediatric cancers: Insights and novel therapeutic approaches," *Cancers (Basel)*, vol. 15, no. 14, 2023.

[19] L. E. Matthyssens, J. G. Nuchtern, C. P. Van De Ven, H. O. S. Gabra, K. Bjornland, S. Irtan, J. Stenman, L. Pio, K. M. Cross, S. Avanzini, A. Inserra, J. G. Chacon, P. Dall'igna, D. Von Schweinitz, K. Holmes, J. Fuchs, R. Squire, D. Valteau-Couanet, J. R. Park, A. Eggert, P. D. Losty, M. P. La Quaglia, S. Sarnacki, Surgical, C. O. G. Medical Committees of SIOPEN, and GPOH, "A novel standard for systematic reporting of neuroblastoma surgery: The international neuroblastoma surgical report form (insrf): A joint initiative by the pediatric oncological cooperative groups siopen, cog, and gpoh," *Ann Surg*, vol. 275, no. 3, pp. e575–e585, 2022.

[20] R. L. Gomez, S. Ibragimova, R. Ramachandran, A. Philpott, and F. R. Ali, "Tumoral heterogeneity in neuroblastoma," *Biochim Biophys Acta Rev Cancer*, vol. 1877, no. 6, p. 188805, 2022.

[21] E. Sokol and A. V. Desai, "The evolution of risk classification for neuroblastoma," *Children (Basel)*, vol. 6, no. 2, 2019.

[22] V. Smith and J. Foster, "High-risk neuroblastoma treatment review," *Children (Basel)*, vol. 5, no. 9, 2018.

[23] C. U. Louis and J. M. Shohet, "Neuroblastoma: molecular pathogenesis and therapy," *Annu Rev Med*, vol. 66, pp. 49–63, 2015.

[24] M. Ponzoni, T. Bachetti, M. V. Corrias, C. Brignole, F. Pastorino, E. Calarco, V. Bensa, E. Giusto, I. Ceccherini, and P. Perri, "Recent advances in the developmental origin of neuroblastoma: an overview," *J Exp Clin Cancer Res*, vol. 41, no. 1, p. 92, 2022.

[25] E. K. Barr and M. A. Applebaum, "Genetic predisposition to neuroblastoma," *Children (Basel)*, vol. 5, no. 9, 2018.

[26] V. Mlakar, S. Jurkovic Mlakar, G. Lopez, J. M. Maris, M. Ansari, and F. Gumy-Pause, "11q deletion in neuroblastoma: a review of biological and clinical implications," *Mol Cancer*, vol. 16, no. 1, p. 114, 2017.

[27] J. Guan, B. Hallberg, and R. H. Palmer, "Chromosome imbalances in neuroblastoma-recent molecular insight into chromosome 1p-deletion, 2p-gain, and 11q-deletion identifies new friends and foes for the future," *Cancers (Basel)*, vol. 13, no. 23, 2021.

[28] J. T. Siaw, N. Javanmardi, J. Van den Eynden, D. E. Lind, S. Fransson, A. Martinez-Monleon, A. Djos, R. M. Sjöberg, M. Östensson, H. Carén, G. Trøen, K. Beiske, A. P. Berbegall, R. Noguera, W. Y. Lai, P. Kogner, R. H. Palmer, B. Hallberg, and T. Martinsson, "11q deletion or alk activity curbs dlg2 expression to maintain an undifferentiated state in neuroblastoma," *Cell Rep*, vol. 32, no. 12, p. 108171, 2020.

[29] D. Bartolucci, L. Montemurro, S. Raieli, S. Lampis, A. Pession, P. Hrelia, and R. Tonelli, "Mycn impact on high-risk neuroblastoma: From diagnosis and prognosis to targeted treatment," *Cancers (Basel)*, vol. 14, no. 18, 2022.

[30] S. Lampis, S. Raieli, L. Montemurro, D. Bartolucci, C. Amadesi, S. Bortolotti, S. Angelucci, A. L. Scardovi, G. Nieddu, L. Cerisoli, F. Paganelli, S. Valente, M. Fischer, A. M. Martelli, G. Pasquinelli, A. Pession, P. Hrelia, and R. Tonelli, "The mycn inhibitor bga002 restores the retinoic acid response leading to differentiation or apoptosis by the mtor block in mycn-amplified neuroblastoma," *J Exp Clin Cancer Res*, vol. 41, no. 1, p. 160, 2022.

[31] T. J. Pugh, O. Morozova, E. F. Attiyeh, S. Asgharzadeh, J. S. Wei, D. Auclair, S. L. Carter, K. Cibulskis, M. Hanna, A. Kiezun, J. Kim, M. S. Lawrence, L. Lichenstein, A. McKenna, C. S. Pedamallu, A. H. Ramos, E. Shefler, A. Sivachenko, C. Sougnez, C. Stewart, A. Ally, I. Birol, R. Chiu, R. D. Corbett, M. Hirst, S. D. Jackman, B. Kamoh, A. H. Khodabakshi, M. Krzywinski, A. Lo, R. A. Moore, K. L. Mungall, J. Qian, A. Tam, N. Thiessen, Y. Zhao, K. A. Cole, M. Diamond, S. J. Diskin, Y. P. Mosse, A. C. Wood, L. Ji, R. Sposto, T. Badgett,

W. B. London, Y. Moyer, J. M. Gastier-Foster, M. A. Smith, J. M. Guidry Auvil, D. S. Gerhard, M. D. Hogarty, S. J. Jones, E. S. Lander, S. B. Gabriel, G. Getz, R. C. Seeger, J. Khan, M. A. Marra, M. Meyerson, and J. M. Maris, "The genetic landscape of high-risk neuroblastoma," *Nat Genet*, vol. 45, no. 3, pp. 279–84, 2013.

[32] S. G. DuBois, M. E. Macy, and T. O. Henderson, "High-risk and relapsed neuroblastoma: Toward more cures and better outcomes," *Am Soc Clin Oncol Educ Book*, vol. 42, pp. 1–13, 2022.

[33] J. C. Jacobson, R. A. Clark, and D. H. Chung, "High-risk neuroblastoma: A surgical perspective," *Children (Basel)*, vol. 10, no. 2, 2023.

[34] N. Bayeva, E. Coll, and O. Piskareva, "Differentiating neuroblastoma: A systematic review of the retinoic acid, its derivatives, and synergistic interactions," *J Pers Med*, vol. 11, no. 3, 2021.

[35] K. K. Matthay, J. G. Villablanca, R. C. Seeger, D. O. Stram, R. E. Harris, N. K. Ramsay, P. Swift, H. Shimada, C. T. Black, G. M. Brodeur, R. B. Gerbing, and C. P. Reynolds, "Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. children's cancer group," *N Engl J Med*, vol. 341, no. 16, pp. 1165–73, 1999.

[36] G. C. Chan and C. M. Chan, "Anti-gd2 directed immunotherapy for high-risk and metastatic neuroblastoma," *Biomolecules*, vol. 12, no. 3, 2022.

[37] F. Herd, N. O. Basta, R. J. Q. McNally, and D. A. Tweddle, "A systematic review of re-induction chemotherapy for children with relapsed high-risk neuroblastoma," *Eur J Cancer*, vol. 111, pp. 50–58, 2019.

[38] N. Pinto, A. Naranjo, E. Hibbitts, S. G. Kreissman, M. M. Granger, M. S. Irwin, R. Bagatell, W. B. London, E. G. Greengard, J. R. Park, and S. G. DuBois, "Predictors of differential response to induction therapy in high-risk neuroblastoma: A report from the children's oncology group (cog)," *Eur J Cancer*, vol. 112, pp. 66–79, 2019.

[39] D. N. Friedman and T. O. Henderson, "Late effects and survivorship issues in patients with neuroblastoma," *Children (Basel)*, vol. 5, no. 8, 2018.

[40] L. L. Wang, R. Suganuma, N. Ikegaki, X. Tang, A. Naranjo, P. McGrady, W. B. London, M. D. Hogarty, J. M. Gastier-Foster, A. T. Look, J. R. Park, J. M. Maris, S. L. Cohn, R. C. Seeger, and H. Shimada, "Neuroblastoma of undifferentiated subtype, prognostic significance of prominent nucleolar formation, and myc/mycn protein expression: a report from the children's oncology group," *Cancer*, vol. 119, no. 20, pp. 3718–26, 2013.

[41] H. Shimada and N. Ikegaki, *Neuroblastoma Pathology and Classification for Precision Prognosis and Therapy Stratification*, book section 1, pp. 1–22. Academic Press, 2019.

[42] T. van Groningen, J. Koster, L. J. Valentijn, D. A. Zwijnenburg, N. Akogul, N. E. Hasselt, M. Broekmans, F. Haneveld, N. E. Nowakowska, J. Bras, C. J. M. van Noesel, A. Jongejan, A. H. van Kampen, L. Koster, F. Baas, L. van Dijk-Kerkhoven, M. Huizer-Smit, M. C. Lecca, A. Chan, A. Lakeman, P. Molenaar, R. Volckmann, E. M. Westerhout, M. Hamdi, P. G. van Sluis, M. E. Ebus, J. J. Molenaar, G. A. Tytgat, B. A. Westerman, J. van Nes, and R. Versteeg, "Neuroblastoma is composed of two super-enhancer-associated differentiation states," *Nat Genet*, vol. 49, no. 8, pp. 1261–1266, 2017.

[43] S. D'Amico, P. Tempora, P. Gragera, K. Król, O. Melaiu, M. A. De Ioris, F. Locatelli, and D. Fruci, "Two bullets in the gun: combining immunotherapy with chemotherapy to defeat neuroblastoma by targeting adrenergic-mesenchymal plasticity," *Front Immunol*, vol. 14, p. 1268645, 2023.

[44] N. W. Mabe, M. Huang, G. N. Dalton, G. Alexe, D. A. Schaefer, A. C. Geraghty, A. L. Robichaud, A. S. Conway, D. Khalid, M. M. Mader, J. A. Belk, K. N. Ross, M. Sheffer, M. H. Linde, N. Ly, W. Yao, M. C. Rotiroti, B. A. H. Smith, M. Wernig, C. R. Bertozzi, M. Monje, C. S. Mitsiades, R. Majeti, A. T. Satpathy, K. Stegmaier, and R. G. Majzner, "Transition to a mesenchymal state in neuroblastoma confers resistance to anti-gd2 antibody via reduced expression of st8sia1," *Nat Cancer*, vol. 3, no. 8, pp. 976–993, 2022.

[45] S. Sengupta, S. Das, A. C. Crespo, A. M. Cornel, A. G. Patel, N. R. Mahadevan, M. Campisi, A. K. Ali, B. Sharma, J. H. Rowe, H. Huang, D. N. Debruyne, E. D. Cerda, M. Krajewska, R. Dries, M. Chen, S. Zhang, L. Soriano, M. A. Cohen, R. Versteeg, R. Jaenisch, S. Spranger, R. Romee, B. C. Miller, D. A. Barbie, S. Nierkens, M. A. Dyer, J. Lieberman, and R. E. George, "Mesenchymal and adrenergic cell lineage states in neuroblastoma possess distinct immunogenic phenotypes," *Nat Cancer*, vol. 3, no. 10, pp. 1228–1246, 2022.

[46] A. Negroni, S. Scarpa, A. Romeo, S. Ferrari, A. Modesti, and G. Raschellà, "Decrease of proliferation rate and induction of differentiation by a mycn antisense dna oligomer in a human neuroblastoma cell line," *Cell Growth Differ*, vol. 2, no. 10, pp. 511–8, 1991.

[47] J. T. Zhang, Z. H. Weng, K. S. Tsang, L. L. Tsang, H. C. Chan, and X. H. Jiang, "Mycn is critical for the maintenance of human embryonic stem cell-derived neural crest stem cells," *PLoS One*, vol. 11, no. 1, p. e0148062, 2016.

[48] T. T. Amatruda, N. Sidell, J. Ranyard, and H. P. Koeffler, "Retinoic acid treatment of human neuroblastoma cells is associated with decreased n-myc expression," *Biochem Biophys Res Commun*, vol. 126, no. 3, pp. 1189–95, 1985.

[49] P. E. Zage, Y. Huo, D. Subramonian, C. Le Clorennec, P. Ghosh, and D. Sahoo, "Identification of a novel gene signature for neuroblastoma differentiation using a boolean implication network," *Genes Chromosomes Cancer*, vol. 62, no. 6, pp. 313–331, 2023.

[50] H. Yoda, T. Inoue, Y. Shinozaki, J. Lin, T. Watanabe, N. Koshikawa, A. Takatori, and H. Nagase, "Direct targeting of mycn gene amplification by site-specific dna alkylation in neuroblastoma," *Cancer Res*, vol. 79, no. 4, pp. 830–840, 2019.

[51] P. Ramani, R. Nash, and C. A. Rogers, "Aurora kinase a is superior to ki67 as a prognostic indicator of survival in neuroblastoma," *Histopathology*, vol. 66, no. 3, pp. 370–9, 2015.

[52] T. Otto, S. Horn, M. Brockmann, U. Eilers, L. Schüttrumpf, N. Popov, A. M. Kenney, J. H. Schulte, R. Beijersbergen, H. Christiansen, B. Berwanger, and M. Eilers, "Stabilization of n-myc is a critical function of aurora a in human neuroblastoma," *Cancer Cell*, vol. 15, no. 1, pp. 67–78, 2009.

[53] A. Faisal, L. Vaughan, V. Bavetsias, C. Sun, B. Atrash, S. Avery, Y. Jamin, S. P. Robinson, P. Workman, J. Blagg, F. I. Raynaud, S. A. Eccles, L. Chesler, and S. Linardopoulos, "The aurora kinase inhibitor cct137690 downregulates mycn and sensitizes mycn-amplified neuroblastoma in vivo," *Mol Cancer Ther*, vol. 10, no. 11, pp. 2115–23, 2011.

[54] E. G. Greengard, "Molecularly targeted therapy for neuroblastoma," *Children (Basel)*, vol. 5, no. 10, 2018.

[55] Z. Liu, S. S. Chen, S. Clarke, V. Veschi, and C. J. Thiele, "Targeting mycn in pediatric and adult cancers," *Front Oncol*, vol. 10, p. 623679, 2020.

[56] M. D. Hogarty, M. D. Norris, K. Davis, X. Liu, N. F. Evageliou, C. S. Hayes, B. Pawel, R. Guo, H. Zhao, E. Sekyere, J. Keating, W. Thomas, N. C. Cheng, J. Murray, J. Smith, R. Sutton, N. Venn, W. B. London, A. Buxton, S. K. Gilmour, G. M. Marshall, and M. Haber, "Odc1 is a critical determinant of mycn oncogenesis and a therapeutic target in neuroblastoma," *Cancer Res*, vol. 68, no. 23, pp. 9735–45, 2008.

[57] J. Theruvath, M. Menard, B. A. H. Smith, M. H. Linde, G. L. Coles, G. N. Dalton, W. Wu, L. Kiru, A. Delaidelli, E. Sotillo, J. L. Silberstein, A. C. Geraghty, A. Banuelos, M. T. Radosevich, S. Dhingra, S. Heitzeneder, A. Tousley, J. Lattin, P. Xu, J. Huang, N. Nasholm, A. He, T. C. Kuo, E. R. B. Sangalang, J. Pons, A. Barkal, R. E. Brewer, K. D. Marjon, J. G. Vilches-Moure, P. L. Marshall,

R. Fernandes, M. Monje, J. R. Cochran, P. H. Sorensen, H. E. Daldrup-Link, I. L. Weissman, J. Sage, R. Majeti, C. R. Bertozzi, W. A. Weiss, C. L. Mackall, and R. G. Majzner, "Anti-gd2 synergizes with cd47 blockade to mediate tumor eradication," *Nat Med*, vol. 28, no. 2, pp. 333–344, 2022.

[58] J. Gray, L. Moreno, R. Weston, G. Barone, A. Rubio, G. Makin, S. Vaidya, A. Ng, V. Castel, K. Nysom, G. Laureys, N. Van Eijkelenburg, C. Owens, M. Gambart, A. D. Pearson, J. Laidler, P. Kearns, and K. Wheatley, "Beacon-immuno: Results of the dinutuximab beta (db) randomization of the beacon-neuroblastoma phase 2 trial—a european innovative therapies for children with cancer (itcc–international society of paediatric oncology europe neuroblastoma group (siopen) trial," *Pediatric Oncology*, vol. 40, 2022.

[59] A. Heczey, X. Xu, A. N. Courtney, G. Tian, G. A. Barragan, L. Guo, C. M. Amador, N. Ghatwai, P. Rathi, M. S. Wood, Y. Li, C. Zhang, T. Demberg, E. J. Di Pierro, A. C. Sher, H. Zhang, B. Mehta, S. G. Thakkar, B. Grilley, T. Wang, B. D. Weiss, A. Montalbano, M. Subramaniam, C. Xu, C. Sachar, D. K. Wells, G. Dotti, and L. S. Metelitsa, "Anti-gd2 car-nkt cells in relapsed or refractory neuroblastoma: updated phase 1 trial interim results," *Nat Med*, vol. 29, no. 6, pp. 1379–1388, 2023.

[60] T. Flaadt, R. L. Ladenstein, M. Ebinger, H. N. Lode, H. B. Arnardóttir, U. Poetschger, W. Schwinger, R. Meisel, F. R. Schuster, M. Döring, P. F. Ambros, M. Queudeville, J. Fuchs, S. W. Warmann, J. Schäfer, C. Seitz, P. Schlegel, I. B. Brecht, U. Holzer, T. Feuchtinger, T. Simon, J. H. Schulte, A. Eggert, H. M. Teltschik, T. Illhardt, R. Handgretinger, and P. Lang, "Anti-gd2 antibody dinutuximab beta and low-dose interleukin 2 after haploidentical stem-cell transplantation in patients with relapsed neuroblastoma: A multicenter, phase i/ii trial," *J Clin Oncol*, vol. 41, no. 17, pp. 3135–3148, 2023.

[61] M. Evers, M. Stip, K. Keller, H. Willemen, M. Nederend, M. Jansen, C. Chan, K. Budding, S. Nierkens, T. Valerius, F. Meyer-Wentrup, N. Eijkelkamp, and J. Leusen, "Anti-gd2 iga kills tumors by neutrophils without antibody-associated pain in the preclinical treatment of high-risk neuroblastoma," *J Immunother Cancer*, vol. 9, no. 10, 2021.

[62] X. Shen and B. Zhao, "Efficacy of pd-1 or pd-l1 inhibitors and pd-l1 expression status in cancer: meta-analysis," *BMJ*, vol. 362, p. k3529, 2018.

[63] V. Rigo, L. Emionite, A. Daga, S. Astigiano, M. V. Corrias, C. Quintarelli, F. Locatelli, S. Ferrini, and M. Croce, "Combined immunotherapy with anti-pdl-1/pd-1 and anti-cd4 antibodies cures syngeneic disseminated neuroblastoma," *Sci Rep*, vol. 7, no. 1, p. 14049, 2017.

[64] S. Shirinbak, R. Y. Chan, S. Shahani, S. Muthugounder, R. Kennedy, L. T. Hung, G. E. Fernandez, M. D. Hadjidaniel, B. Moghimi, M. A. Sheard, A. L. Epstein, M. Fabbri, H. Shimada, and S. Asgharzadeh, "Combined immune checkpoint blockade increases cd8+cd28+pd-1+ effector t cells and provides a therapeutic strategy for patients with neuroblastoma," *Oncoimmunology*, vol. 10, no. 1, p. 1838140, 2021.

[65] U. H. S. N. F. Trust, U. C. L. Hospitals, M. University of Wisconsin, U. H. Greifswald, S. K. C. US/EU, J. A. C. i. Kids, and T. B. o. Parents, "Phase i study of 131-i mibg followed by nivolumab and dinutuximab beta antibodies in children with relapsed/refractory neuroblastoma," May 24 2018.

[66] K. Bourcier, A. Le Cesne, L. Tselikas, J. Adam, O. Mir, C. Honore, and T. de Baere, "Basic knowledge in soft tissue sarcoma," *Cardiovasc Intervent Radiol*, vol. 42, no. 9, pp. 1255–1261, 2019.

[67] D. S. Hawkins, S. L. Spunt, S. X. Skapek, and C. O. G. S. T. S. Committee, "Children's oncology group's 2013 blueprint for research: Soft tissue sarcomas," *Pediatr Blood Cancer*, vol. 60, no. 6, pp. 1001–8, 2013.

[68] E. R. Rudzinski, A. Kelsey, C. Vokuhl, C. M. Linardic, J. Shipley, S. Hettmer, E. Koscielniak, D. S. Hawkins, and G. Bisogno, "Pathology of childhood rhabdomyosarcoma: A consensus opinion document from the children's oncology group, european paediatric soft tissue sarcoma study group, and the cooperative weichteilsarkom studiengruppe," *Pediatr Blood Cancer*, vol. 68, no. 3, p. e28798, 2020.

[69] N. Jawad and K. McHugh, "The clinical and radiologic features of paediatric rhabdomyosarcoma," *Pediatr Radiol*, vol. 49, no. 11, pp. 1516–1523, 2019.

[70] F. Ramadan, A. Fahs, S. E. Ghayad, and R. Saab, "Signaling pathways in rhabdomyosarcoma invasion and metastasis," *Cancer Metastasis Rev*, vol. 39, no. 1, pp. 287–301, 2020.

[71] E. A. Perez, N. Kassira, M. C. Cheung, L. G. Koniaris, H. L. Neville, and J. E. Sola, "Rhabdomyosarcoma in children: a seer population based study," *J Surg Res*, vol. 170, no. 2, pp. e243–51, 2011.

[72] E. Missiaglia, D. Williamson, J. Chisholm, P. Wirapati, G. Pierron, F. Petel, J. P. Concordet, K. Thway, O. Oberlin, K. Pritchard-Jones, O. Delattre, M. Delorenzi, and J. Shipley, "Pax3/foxo1 fusion gene status is the key prognostic molecular marker in rhabdomyosarcoma and significantly improves current risk stratification," *J Clin Oncol*, vol. 30, no. 14, pp. 1670–7, 2012.

[73] J. W. Tsai, Y. C. ChangChien, J. C. Lee, Y. C. Kao, W. S. Li, C. W. Liang, I. C. Liao, Y. M. Chang, J. C. Wang, C. F. Tsao, S. C. Yu, and H. Y. Huang, "The expanding

morphological and genetic spectrum of myod1-mutant spindle cell/sclerosing rhabdomyosarcomas: a clinicopathological and molecular comparison of mutated and non-mutated cases," *Histopathology*, vol. 74, no. 6, pp. 933–943, 2019.

[74] S. Watson, V. Perrin, D. Guillemot, S. Reynaud, J. M. Coindre, M. Karanian, J. M. Guinebretiere, P. Freneaux, F. Le Loarer, M. Bouvet, L. Galmiche-Rolland, F. Larousserie, E. Longchampt, D. Ranchere-Vince, G. Pierron, O. Delattre, and F. Tirode, "Transcriptomic definition of molecular subgroups of small round cell sarcomas," *J Pathol*, vol. 245, no. 1, pp. 29–40, 2018.

[75] S. Whittle, R. Venkatramani, A. Schönstein, S. D. Pack, R. Alaggio, C. Vokuhl, E. R. Rudzinski, A. L. Wulf, A. Zin, J. R. Gruver, M. A. Arnold, J. H. M. Merks, S. Hettmer, E. Koscielniak, F. G. Barr, D. S. Hawkins, G. Bisogno, and M. Sparber-Sauer, "Congenital spindle cell rhabdomyosarcoma: An international cooperative analysis," *Eur J Cancer*, vol. 168, pp. 56–64, 2022.

[76] P. Q. Deb, R. J. Chokshi, S. Li, and D. I. Suster, "Pleomorphic rhabdomyosarcoma: A systematic review with outcome analysis and report of a rare abdominal wall lesion," *Int J Surg Pathol*, pp. 548–556, 2022.

[77] F. Taza, A. Kanwal, M. Zulty, and S. Mustafa, "High-grade pleomorphic rhabdomyosarcoma in a 60-year-old male: a case report and review of the literature," *J Community Hosp Intern Med Perspect*, vol. 10, no. 3, pp. 287–289, 2020.

[78] N. P. Agaram, "Evolving classification of rhabdomyosarcoma," *Histopathology*, vol. 80, no. 1, pp. 98–108, 2022.

[79] A. D. Marshall and G. C. Grosveld, "Alveolar rhabdomyosarcoma - the molecular drivers of pax3/7-foxo1-induced tumorigenesis," *Skelet Muscle*, vol. 2, no. 1, p. 25, 2012.

[80] S. X. Skapek, J. Anderson, F. G. Barr, J. A. Bridge, J. M. Gastier-Foster, D. M. Parham, E. R. Rudzinski, T. Triche, and D. S. Hawkins, "Pax-foxo1 fusion status drives unfavorable outcome for children with rhabdomyosarcoma: a children's oncology group report," *Pediatr Blood Cancer*, vol. 60, no. 9, pp. 1411–7, 2013.

[81] D. Williamson, E. Missiaglia, A. de Reynies, G. Pierron, B. Thuille, G. Palenzuela, K. Thway, D. Orbach, M. Lae, P. Freneaux, K. Pritchard-Jones, O. Oberlin, J. Shipley, and O. Delattre, "Fusion gene-negative alveolar rhabdomyosarcoma is clinically and molecularly indistinguishable from embryonal rhabdomyosarcoma," *J Clin Oncol*, vol. 28, no. 13, pp. 2151–8, 2010.

[82] R. Dasgupta and D. A. Rodeberg, "Update on rhabdomyosarcoma," *Semin Pediatr Surg*, vol. 21, no. 1, pp. 68–78, 2012.

[83] K. M. Amer, J. E. Thomson, D. Congiusta, A. Dobitsch, A. Chaudhry, M. Li, A. Chaudhry, A. Bozzo, B. Siracuse, M. N. Aytekin, M. Ghert, and K. S. Beebe, "Epidemiology, incidence, and survival of rhabdomyosarcoma subtypes: Seer and ices database analysis," *J Orthop Res*, vol. 37, no. 10, pp. 2226–2230, 2019.

[84] J. Chisholm, H. Mandeville, M. Adams, V. Minard-Collin, T. Rogers, A. Kelsey, J. Shipley, R. R. van Rijn, I. de Vries, R. van Ewijk, B. de Keizer, S. A. Gatz, M. Casanova, L. L. Hjalgrim, C. Firth, K. Wheatley, P. Kearns, W. Liu, A. Kirkham, H. Rees, G. Bisogno, A. Wasti, S. Wakeling, D. Heenen, D. A. Tweddle, J. H. M. Merks, and M. Jenney, "Frontline and relapsed rhabdomyosarcoma (far-rms) clinical trial: A report from the european paediatric soft tissue sarcoma study group (epssg)," *Cancers (Basel)*, vol. 16, no. 5, 2024.

[85] X. Ma, D. Huang, W. Zhao, L. Sun, H. Xiong, Y. Zhang, M. Jin, D. Zhang, C. Huang, H. Wang, W. Zhang, N. Sun, L. He, and J. Tang, "Clinical characteristics and prognosis of childhood rhabdomyosarcoma: a ten-year retrospective multicenter study," *Int J Clin Exp Med*, vol. 8, no. 10, pp. 17196–205, 2015.

[86] S. Gallego, I. Zanetti, D. Orbach, D. Ranchère, J. Shipley, A. Zin, C. Bergeron, G. L. de Salvo, J. Chisholm, A. Ferrari, M. Jenney, H. C. Mandeville, T. Rogers, J. H. M. Merks, P. Mudry, H. Glosli, G. M. Milano, S. Ferman, G. Bisogno, and E. P. S. T. S. S. G. (EpSSG), "Fusion status in patients with lymph node-positive (n1) alveolar rhabdomyosarcoma is a powerful predictor of prognosis: Experience of the european paediatric soft tissue sarcoma study group (epssg)," *Cancer*, vol. 124, no. 15, pp. 3201–3209, 2018.

[87] H. P. McDowell, "Update on childhood rhabdomyosarcoma," *Arch Dis Child*, vol. 88, no. 4, pp. 354–7, 2003.

[88] R. A. Schoot, J. C. Chisholm, M. Casanova, V. Minard-Colin, B. Geoerger, A. L. Cameron, B. Coppadoro, I. Zanetti, D. Orbach, A. Kelsey, T. Rogers, C. Guizani, M. Elze, M. Ben-Arush, K. McHugh, R. R. van Rijn, S. Ferman, S. Gallego, A. Ferrari, M. Jenney, G. Bisogno, and J. H. M. Merks, "Metastatic rhabdomyosarcoma: Results of the european," *J Clin Oncol*, vol. 40, no. 32, pp. 3730–3740, 2022.

[89] X. Wang, J. Feng, Z. Li, X. Zhang, J. Chen, and G. Feng, "Characteristics and prognosis of embryonal rhabdomyosarcoma in children and adolescents: An analysis of 464 cases from the seer database," *Pediatr Investig*, vol. 4, no. 4, pp. 242–249, 2020.

[90] O. Oberlin, A. Rey, E. Lyden, G. Bisogno, M. C. Stevens, W. H. Meyer, M. Carli, and J. R. Anderson, "Prognostic factors in metastatic rhabdomyosarcomas:

results of a pooled analysis from united states and european cooperative groups," *J Clin Oncol*, vol. 26, no. 14, pp. 2384–9, 2008.

[91] S. Hettmer, C. M. Linardic, A. Kelsey, E. R. Rudzinski, C. Vokuhl, J. Selfe, O. Ruhen, J. F. Shern, J. Khan, A. R. Kovach, P. J. Lupo, S. A. Gatz, B. W. Schafer, S. Volchenboum, V. Minard-Colin, E. Koscielniak, D. S. Hawkins, G. Bisogno, M. Sparber-Sauer, R. Venkatramani, J. H. M. Merks, and J. Shipley, "Molecular testing of rhabdomyosarcoma in clinical trials to improve risk stratification and outcome: A consensus view from european paediatric soft tissue sarcoma study group, children's oncology group and cooperative weichteilsarkom-studiengruppe," *Eur J Cancer*, vol. 172, pp. 367–386, 2022.

[92] N. P. Agaram, M. P. LaQuaglia, R. Alaggio, L. Zhang, Y. Fujisawa, M. Ladanyi, L. H. Wexler, and C. R. Antonescu, "Myod1-mutant spindle cell and sclerosing rhabdomyosarcoma: an aggressive subtype irrespective of age. a reappraisal for molecular classification and risk stratification," *Mod Pathol*, vol. 32, no. 1, pp. 27–36, 2019.

[93] J. C. Breneman, E. Lyden, A. S. Pappo, M. P. Link, J. R. Anderson, D. M. Parham, S. J. Qualman, M. D. Wharam, S. S. Donaldson, H. M. Maurer, W. H. Meyer, K. S. Baker, C. N. Paidas, and W. M. Crist, "Prognostic factors and clinical outcomes in children and adolescents with metastatic rhabdomyosarcoma–a report from the intergroup rhabdomyosarcoma study iv," *J Clin Oncol*, vol. 21, no. 1, pp. 78–84, 2003.

[94] D. Han, C. Li, X. Li, Q. Huang, F. Xu, S. Zheng, H. Wang, and J. Lyu, "Prognostic factors in patients with rhabdomyosarcoma using competing-risks analysis: A study of cases in the seer database," *J Oncol*, vol. 2020, p. 2635486, 2020.

[95] J. N. Crane, W. Xue, A. Qumseya, Z. Gao, C. A. S. Arndt, S. S. Donaldson, D. J. Harrison, D. S. Hawkins, C. M. Linardic, L. Mascarenhas, W. H. Meyer, D. A. Rodeberg, E. R. Rudzinski, B. L. Shulkin, D. O. Walterhouse, R. Venkatramani, and A. R. Weiss, "Clinical group and modified tnm stage for rhabdomyosarcoma: A review from the children's oncology group," *Pediatr Blood Cancer*, vol. 69, no. 6, p. e29644, 2022.

[96] R. L. Yechieli, H. C. Mandeville, S. M. Hiniker, V. Bernier-Chastagner, S. McGovern, G. Scarzello, S. Wolden, A. Cameron, J. Breneman, R. D. Fajardo, and S. S. Donaldson, "Rhabdomyosarcoma," *Pediatr Blood Cancer*, vol. 68 Suppl 2, p. e28254, 2021.

[97] C. Dumontet, "Mechanisms of action and resistance to tubulin-binding agents," *Expert Opin Investig Drugs*, vol. 9, no. 4, pp. 779–88, 2000.

[98] J. Blasiak, "Dna-damaging anticancer drugs - a perspective for dna repair-oriented therapy," *Curr Med Chem*, vol. 24, no. 15, pp. 1488–1503, 2017.

[99] M. Gangireddy and V. Nookala, *Ifosfamide*. 2022.

[100] B. Sprangers, L. Cosmai, and C. Porta, *Conventional chemotherapy*, vol. e11 of *Onco-Nephrology*. Elsevier, 2020.

[101] S. Malempati and D. S. Hawkins, "Rhabdomyosarcoma: review of the children's oncology group (cog) soft-tissue sarcoma committee experience and rationale for current cog studies," *Pediatr Blood Cancer*, vol. 59, no. 1, pp. 5–10, 2012.

[102] B. J. Weigel, E. Lyden, J. R. Anderson, W. H. Meyer, D. M. Parham, D. A. Rodeberg, J. M. Michalski, D. S. Hawkins, and C. A. Arndt, "Intensive multiagent therapy, including dose-compressed cycles of ifosfamide/etoposide and vincristine/doxorubicin/cyclophosphamide, irinotecan, and radiation, in patients with high-risk rhabdomyosarcoma: A report from the children's oncology group," *J Clin Oncol*, vol. 34, no. 2, pp. 117–22, 2016.

[103] M. Carli, R. Colombatti, O. Oberlin, G. Bisogno, J. Treuner, E. Koscielniak, G. Tridello, A. Garaventa, R. Pinkerton, and M. Stevens, "European intergroup studies (mmt4-89 and mmt4-91) on childhood metastatic rhabdomyosarcoma: final results and analysis of prognostic factors," *J Clin Oncol*, vol. 22, no. 23, pp. 4787–94, 2004.

[104] G. Bisogno, G. L. De Salvo, C. Bergeron, S. Gallego Melcon, J. H. Merks, A. Kelsey, H. Martelli, V. Minard-Colin, D. Orbach, H. Glosli, J. Chisholm, M. Casanova, I. Zanetti, C. Devalck, M. Ben-Arush, P. Mudry, S. Ferman, M. Jenney, A. Ferrari, and G. European paediatric Soft tissue sarcoma Study, "Vinorelbine and continuous low-dose cyclophosphamide as maintenance chemotherapy in patients with high-risk rhabdomyosarcoma (rms 2005): a multicentre, open-label, randomised, phase 3 trial," *Lancet Oncol*, vol. 20, no. 11, pp. 1566–1575, 2019.

[105] A. R. Rosenberg, J. R. Anderson, E. Lyden, D. A. Rodeberg, S. L. Wolden, S. C. Kao, D. M. Parham, C. Arndt, and D. S. Hawkins, "Early response as assessed by anatomic imaging does not predict failure-free survival among patients with group iii rhabdomyosarcoma: a report from the children's oncology group," *Eur J Cancer*, vol. 50, no. 4, pp. 816–23, 2014.

[106] A. Ferrari, J. C. Chisholm, M. Jenney, V. Minard-Colin, D. Orbach, M. Casanova, G. Guillen, H. Glosli, R. R. van Rijn, R. A. Schoot, A. L. Cameron, T. Rogers, R. Alaggio, M. Ben-Arush, H. C. Mandeville, C. Devalck, A. S. Defachelles, B. Coppadoro, G. Bisogno, and J. H. M. Merks, "Adolescents and young adults with rhabdomyosarcoma treated in the european paediatric soft tissue sarcoma study group (epssg) protocols: a cohort study," *Lancet Child Adolesc Health*, vol. 6, no. 8, pp. 545–554, 2022.

[107] C. M. Heske and L. Mascarenhas, "Relapsed rhabdomyosarcoma," *J Clin Med*, vol. 10, no. 4, 2021.

[108] S. Ghilu, C. L. Morton, A. V. Vaseva, S. Zheng, R. T. Kurmasheva, and P. J. Houghton, "Approaches to identifying drug resistance mechanisms to clinically relevant treatments in childhood rhabdomyosarcoma," *Cancer Drug Resist*, vol. 5, no. 1, pp. 80–89, 2022.

[109] C. M. Linardic, "Pax3-foxo1 fusion gene in rhabdomyosarcoma," *Cancer Lett*, vol. 270, no. 1, pp. 10–8, 2008.

[110] J. F. Shern, L. Chen, J. Chmielecki, J. S. Wei, R. Patidar, M. Rosenberg, L. Ambrogio, D. Auclair, J. Wang, Y. K. Song, C. Tolman, L. Hurd, H. Liao, S. Zhang, D. Bogen, A. S. Brohl, S. Sindiri, D. Catchpoole, T. Badgett, G. Getz, J. Mora, J. R. Anderson, S. X. Skapek, F. G. Barr, M. Meyerson, D. S. Hawkins, and J. Khan, "Comprehensive genomic analysis of rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors," *Cancer Discov*, vol. 4, no. 2, pp. 216–31, 2014.

[111] D. L. Casey, L. H. Wexler, K. L. Pitter, R. M. Samstein, E. K. Slotkin, and S. L. Wolden, "Genomic determinants of clinical outcomes in rhabdomyosarcoma," *Clin Cancer Res*, vol. 26, no. 5, pp. 1135–1140, 2020.

[112] X. Wang, H. Zhang, and X. Chen, "Drug resistance and combating drug resistance in cancer," *Cancer Drug Resist*, vol. 2, pp. 141–160, 2019.

[113] G. Housman, S. Byler, S. Heerboth, K. Lapinska, M. Longacre, N. Snyder, and S. Sarkar, "Drug resistance in cancer: an overview," *Cancers (Basel)*, vol. 6, no. 3, pp. 1769–92, 2014.

[114] S. Kelderman, T. N. Schumacher, and J. B. Haanen, "Acquired and intrinsic resistance in cancer immunotherapy," *Mol Oncol*, vol. 8, no. 6, pp. 1132–9, 2014.

[115] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–74, 2011.

[116] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nat Rev Clin Oncol*, vol. 15, no. 2, pp. 81–94, 2018.

[117] Z. Yu, T. G. Pestell, M. P. Lisanti, and R. G. Pestell, "Cancer stem cells," *Int J Biochem Cell Biol*, vol. 44, no. 12, pp. 2144–51, 2012.

[118] H. Zahreddine and K. L. Borden, "Mechanisms and insights into drug resistance in cancer," *Front Pharmacol*, vol. 4, p. 28, 2013.

[119] A. K. Nanayakkara, C. A. Follit, G. Chen, N. S. Williams, P. D. Vogel, and J. G. Wise, "Targeted inhibitors of p-glycoprotein increase chemotherapeutic-induced mortality of multidrug resistant tumor cells," *Sci Rep*, vol. 8, no. 1, p. 967, 2018.

[120] J. W. Yoon, M. Lamm, C. Chandler, P. Iannaccone, and D. Walterhouse, "Up-regulation of gli1 in vincristine-resistant rhabdomyosarcoma and ewing sarcoma," *BMC Cancer*, vol. 20, no. 1, p. 511, 2020.

[121] R. Komdeur, J. Klunder, W. T. van der Graaf, E. van den Berg, E. S. de Bont, H. J. Hoekstra, and W. M. Molenaar, "Multidrug resistance proteins in rhabdomyosarcomas: comparison between children and adults," *Cancer*, vol. 97, no. 8, pp. 1999–2005, 2003.

[122] H. S. Chan, G. Haddad, P. S. Thorner, G. DeBoer, Y. P. Lin, N. Ondrusek, H. Yeger, and V. Ling, "P-glycoprotein expression as a predictor of the outcome of therapy for neuroblastoma," *N Engl J Med*, vol. 325, no. 23, pp. 1608–14, 1991.

[123] C. R. Dhooge, B. M. De Moerloose, Y. C. Benoit, N. Van Roy, Philippé, and G. G. Laureys, "Expression of the mdr1 gene product p-glycoprotein in childhood neuroblastoma," *Cancer*, vol. 80, no. 7, pp. 1250–7, 1997.

[124] F. Di Nicolantonio, S. J. Mercer, L. A. Knight, F. G. Gabriel, P. A. Whitehouse, S. Sharma, A. Fernando, S. Glaysher, S. Di Palma, P. Johnson, S. S. Somers, S. Toh, B. Higgins, A. Lamont, T. Gulliford, J. Hurren, C. Yiangou, and I. A. Cree, "Cancer cell adaptation to chemotherapy," *BMC Cancer*, vol. 5, p. 78, 2005.

[125] R. Pagliarini, W. Shao, and W. R. Sellers, "Oncogene addiction: pathways of therapeutic response, resistance, and road maps toward a cure," *EMBO Rep*, vol. 16, no. 3, pp. 280–96, 2015.

[126] I. A. Cree and P. Charlton, "Molecular chess? hallmarks of anti-cancer drug resistance," *BMC Cancer*, vol. 17, no. 1, p. 10, 2017.

[127] S. Dasari and P. B. Tchounwou, "Cisplatin in cancer therapy: molecular mechanisms of action," *Eur J Pharmacol*, vol. 740, pp. 364–78, 2014.

[128] P. E. Czabotar, G. Lessene, A. Strasser, and J. M. Adams, "Control of apoptosis by the bcl-2 protein family: implications for physiology and therapy," *Nat Rev Mol Cell Biol*, vol. 15, no. 1, pp. 49–63, 2014.

[129] P. Wu, W. Gao, M. Su, E. C. Nice, W. Zhang, J. Lin, and N. Xie, "Adaptive mechanisms of tumor therapy resistance driven by tumor microenvironment," *Front Cell Dev Biol*, vol. 9, p. 641469, 2021.

[130] B. Son, S. Lee, H. Youn, E. Kim, W. Kim, and B. Youn, "The role of tumor microenvironment in therapeutic resistance," *Oncotarget*, vol. 8, no. 3, pp. 3933–3945, 2017.

[131] S. Wang, J. Wang, Z. Chen, J. Luo, W. Guo, L. Sun, and L. Lin, "Targeting m2-like tumor-associated macrophages is a potential therapeutic approach to overcome antitumor drug resistance," *NPJ Precis Oncol*, vol. 8, no. 1, p. 31, 2024.

[132] J. L. Hu, W. Wang, X. L. Lan, Z. C. Zeng, Y. S. Liang, Y. R. Yan, F. Y. Song, F. F. Wang, X. H. Zhu, W. J. Liao, W. T. Liao, Y. Q. Ding, and L. Liang, "Cafs secreted exosomes promote metastasis and chemotherapy resistance by enhancing cell stemness and epithelial-mesenchymal transition in colorectal cancer," *Mol Cancer*, vol. 18, no. 1, p. 91, 2019.

[133] H. Zhang, T. Deng, R. Liu, T. Ning, H. Yang, D. Liu, Q. Zhang, D. Lin, S. Ge, M. Bai, X. Wang, L. Zhang, H. Li, Y. Yang, Z. Ji, H. Wang, G. Ying, and Y. Ba, "Caf secreted mir-522 suppresses ferroptosis and promotes acquired chemo-resistance in gastric cancer," *Mol Cancer*, vol. 19, no. 1, p. 43, 2020.

[134] X. Long, W. Xiong, X. Zeng, L. Qi, Y. Cai, M. Mo, H. Jiang, B. Zhu, Z. Chen, and Y. Li, "Cancer-associated fibroblasts promote cisplatin resistance in bladder cancer cells by increasing igf-1/erb/bcl-2 signalling," *Cell Death Dis*, vol. 10, no. 5, p. 375, 2019.

[135] E. H. Cheteh2017, M. Augsten, H. Rundqvist, J. Bianchi, V. Sarne, L. Egevad, V. J. Bykov, A. Östman, and K. G. Wiman, "Human cancer-associated fibroblasts enhance glutathione levels and antagonize drug-induced prostate cancer cell death," *Cell Death Dis*, vol. 8, no. 6, p. e2848, 2017.

[136] Z. Chen, F. Han, Y. Du, H. Shi, and W. Zhou, "Hypoxic microenvironment in cancer: molecular mechanisms and therapeutic interventions," *Signal Transduct Target Ther*, vol. 8, no. 1, p. 70, 2023.

[137] L. Bo, Y. Wang, Y. Li, J. N. D. Wurpel, Z. Huang, and Z. S. Chen, "The battlefield of chemotherapy in pediatric cancers," *Cancers (Basel)*, vol. 15, no. 7, 2023.

[138] E. Ward, C. DeSantis, A. Robbins, B. Kohler, and A. Jemal, "Childhood and adolescent cancer statistics, 2014," *CA Cancer J Clin*, vol. 64, no. 2, pp. 83–103, 2014.

[139] M. F. Wedekind, N. L. Denton, C. Y. Chen, and T. P. Cripe, "Pediatric cancer immunotherapy: Opportunities and challenges," *Paediatr Drugs*, vol. 20, no. 5, pp. 395–408, 2018.

[140] B. Hutzen, S. N. Paudel, M. Naeimi Kararoudi, K. A. Cassady, D. A. Lee, and T. P. Cripe, "Immunotherapies for pediatric cancer: current landscape and future perspectives," *Cancer Metastasis Rev*, vol. 38, no. 4, pp. 573–594, 2019.

[141] A. H. Long, D. A. Morgenstern, A. Leruste, F. Bourdeaut, and K. L. Davis, "Checkpoint immunotherapy in pediatrics: Here, gone, and back again," *Am Soc Clin Oncol Educ Book*, vol. 42, pp. 1–14, 2022.

[142] B. Hutzen, M. Ghonime, J. Lee, E. R. Mardis, R. Wang, D. A. Lee, M. S. Cairo, R. D. Roberts, T. P. Cripe, and K. A. Cassady, "Immunotherapeutic challenges for pediatric cancers," *Mol Ther Oncolytics*, vol. 15, pp. 38–48, 2019.

[143] J. Wang, W. Yao, and K. Li, "Applications and prospects of targeted therapy for neuroblastoma," *World J Pediatr Surg*, vol. 3, no. 2, p. e000164, 2020.

[144] C. A. Burkhart, F. Watt, J. Murray, M. Pajic, A. Prokvolit, C. Xue, C. Flemming, J. Smith, A. Purmal, N. Isachenko, P. G. Komarov, K. V. Gurova, A. C. Sartorelli, G. M. Marshall, M. D. Norris, A. V. Gudkov, and M. Haber, "Small-molecule multidrug resistance-associated protein 1 inhibitor reversan increases the therapeutic index of chemotherapy in mouse models of neuroblastoma," *Cancer Res*, vol. 69, no. 16, pp. 6573–80, 2009.

[145] M. Haber, J. Smith, S. B. Bordow, C. Flemming, S. L. Cohn, W. B. London, G. M. Marshall, and M. D. Norris, "Association of high-level mrp1 expression with poor clinical outcome in a large prospective study of primary neuroblastoma," *J Clin Oncol*, vol. 24, no. 10, pp. 1546–53, 2006.

[146] M. D. Norris, J. Smith, K. Tanabe, P. Tobin, C. Flemming, G. L. Scheffer, P. Wielinga, S. L. Cohn, W. B. London, G. M. Marshall, J. D. Allen, and M. Haber, "Expression of multidrug transporter mrp4/abcc4 is a marker of poor prognosis in neuroblastoma and confers resistance to irinotecan in vitro," *Mol Cancer Ther*, vol. 4, no. 4, pp. 547–53, 2005.

[147] A. Belounis, C. Nyalendo, R. Le Gall, T. V. Imbriglio, M. Mahma, P. Teira, M. Beaunoyer, S. Cournoyer, E. Haddad, G. Vassal, and H. Sartelet, "Autophagy is associated with chemoresistance in neuroblastoma," *BMC Cancer*, vol. 16, no. 1, p. 891, 2016.

[148] T. Chen, C. Zeng, Z. Li, J. Wang, F. Sun, J. Huang, S. Lu, J. Zhu, Y. Zhang, X. Sun, and Z. Zhen, "Investigation of chemoresistance to first-line chemotherapy and its possible association with autophagy in high-risk neuroblastoma," *Front Oncol*, vol. 12, p. 1019106, 2022.

[149] L. M. Hansford, A. E. McKee, L. Zhang, R. E. George, J. T. Gerstle, P. S. Thorner, K. M. Smith, A. T. Look, H. Yeger, F. D. Miller, M. S. Irwin, C. J. Thiele, and D. R. Kaplan, "Neuroblastoma cells isolated from bone marrow metastases contain a naturally enriched tumor-initiating cell," *Cancer Res*, vol. 67, no. 23, pp. 11234–43, 2007.

[150] R. A. Ross, J. D. Walton, D. Han, H. F. Guo, and N. K. Cheung, "A distinct gene expression signature characterizes human neuroblastoma cancer stem cells," *Stem Cell Res*, vol. 15, no. 2, pp. 419–26, 2015.

[151] M. Veas-Perez de Tudela, M. Delgado-Esteban, J. Cuende, J. P. Bolaños, and A. Almeida, "Human neuroblastoma cells with mycn amplification are selectively resistant to oxidative stress by transcriptionally up-regulating glutamate cysteine ligase," *J Neurochem*, vol. 113, no. 4, pp. 819–25, 2010.

[152] A. Wieczorek, K. Śladowska, and H. N. Lode, "Efficacy and safety of anti-gd2 immunotherapy with dinutuximab beta in the treatment of relapsed/refractory high-risk neuroblastoma," *Target Oncol*, vol. 20, no. 4, pp. 551–568, 2025.

[153] A. L. Yu, A. L. Gilman, M. F. Ozkaynak, A. Naranjo, M. B. Diccianni, J. Gan, J. A. Hank, A. Batova, W. B. London, S. C. Tenney, M. Smith, B. L. Shulkin, M. Parisi, K. K. Matthay, S. L. Cohn, J. M. Maris, R. Bagatell, J. R. Park, and P. M. Sondel, "Long-term follow-up of a phase iii study of ch14.18 (dinutuximab) + cytokine immunotherapy in children with high-risk neuroblastoma: Cog study anbl0032," *Clin Cancer Res*, vol. 27, no. 8, pp. 2179–2189, 2021.

[154] A. A. Pilgrim, H. C. Jonus, A. Ho, A. C. Cole, J. Shim, and K. C. Goldsmith, "The yes-associated protein (yap) is associated with resistance to anti-gd2 immunotherapy in neuroblastoma through downregulation of," *Oncoimmunology*, vol. 12, no. 1, p. 2240678, 2023.

[155] C. Melguizo, J. Prados, A. R. Rama, R. Ortiz, P. J. Alvarez, J. E. Fernandez, and A. Aranega, "Multidrug resistance and rhabdomyosarcoma (review)," *Oncol Rep*, vol. 26, no. 4, pp. 755–61, 2011.

[156] A. G. Patel, X. Chen, X. Huang, M. R. Clay, N. Komorova, M. J. Krasin, A. Pappo, H. Tillman, B. A. Orr, J. McEvoy, B. Gordon, K. Blankenship, C. Reilly, X. Zhou, J. L. Norrie, A. Karlstrom, J. Yu, D. Wodarz, E. Stewart, and M. A. Dyer, "The myogenesis program drives clonal selection and drug resistance in rhabdomyosarcoma," *Dev Cell*, vol. 57, no. 10, pp. 1226–1240 e8, 2022.

[157] M. Pasello, M. C. Manara, F. Michelacci, M. Fanelli, C. M. Hattinger, G. Nicoletti, L. Landuzzi, P. L. Lollini, A. Caccuri, P. Picci, K. Scotlandi, and M. Serra, "Targeting glutathione-s transferase enzymes in musculoskeletal sarcomas: a promising therapeutic strategy," *Anal Cell Pathol (Amst)*, vol. 34, no. 3, pp. 131–45, 2011.

[158] M. D. Faye, S. T. Beug, T. E. Graber, N. Earl, X. Xiang, B. Wild, S. Langlois, J. Michaud, K. N. Cowan, R. G. Korneluk, and M. Holcik, "Igf2bp1 controls cell death and drug resistance in rhabdomyosarcomas by regulating translation of ciap1," *Oncogene*, vol. 34, no. 12, pp. 1532–41, 2015.

[159] C. Alcon, A. Manzano-Munoz, E. Prada, J. Mora, A. Soriano, G. Guillen, S. Gallego, J. Roma, J. Samitier, A. Villanueva, and J. Montero, "Sequential combinations of chemotherapeutic agents with bh3 mimetics to treat rhabdomyosarcoma and avoid resistance," *Cell Death Dis*, vol. 11, no. 8, p. 634, 2020.

[160] G. Manzella, D. C. Moonamale, M. Rommele, P. Bode, M. Wachtel, and B. W. Schafer, "A combinatorial drug screen in pdx-derived primary

rhabdomyosarcoma cells identifies the noxa - bcl-xl/mcl-1 balance as target for re-sensitization to first-line therapy in recurrent tumors," *Neoplasia*, vol. 23, no. 9, pp. 929–938, 2021.

[161] K. Kikuchi, S. Hettmer, M. I. Aslam, J. E. Michalek, W. Laub, B. A. Wilky, D. M. Loeb, B. P. Rubin, A. J. Wagers, and C. Keller, "Cell-cycle dependent expression of a translocation-mediated fusion oncogene mediates checkpoint adaptation in rhabdomyosarcoma," *PLoS Genet*, vol. 10, no. 1, p. e1004107, 2014.

[162] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nat Methods*, vol. 5, no. 7, pp. 621–8, 2008.

[163] C. Mertes, I. F. Scheller, V. A. Yépez, M. H. Çelik, Y. Liang, L. S. Kremer, M. Gusic, H. Prokisch, and J. Gagneur, "Detection of aberrant splicing events in rna-seq data using fraser," *Nat Commun*, vol. 12, no. 1, p. 529, 2021.

[164] R. Piskol, G. Ramaswami, and J. B. Li, "Reliable identification of genomic variants from rna-seq data," *Am J Hum Genet*, vol. 93, no. 4, pp. 641–51, 2013.

[165] S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of rna-seq and microarray in transcriptome profiling of activated t cells," *PLoS One*, vol. 9, no. 1, p. e78644, 2014.

[166] D. O'Neil2013, H. Glowatz, and M. Schlumpberger, "Ribosomal rna depletion for efficient use of rna-seq capacity," *Curr Protoc Mol Biol*, vol. Chapter 4, p. Unit 4.19, 2013.

[167] J. Wu2014, J. Xiao, Z. Zhang, X. Wang, S. Hu, and J. Yu, "Ribogenomics: the science and knowledge of rna," *Genomics Proteomics Bioinformatics*, vol. 12, no. 2, pp. 57–63, 2014.

[168] S. Zhao, Y. Zhang, R. Gamini, B. Zhang, and D. von Schack, "Evaluation of two main rna-seq approaches for gene quantification in clinical rna sequencing: polya+ selection versus rrna depletion," *Sci Rep*, vol. 8, no. 1, p. 4781, 2018.

[169] S. Zhao, Y. Zhang, R. Gamini, B. Zhang, and D. von Schack, "Evaluation of two main rna-seq approaches for gene quantification in clinical rna sequencing: polya+ selection versus rrna depletion," *Sci Rep*, vol. 8, no. 1, p. 4781, 2018.

[170] S. Schuierer, W. Carbone, J. Knehr, V. Petitjean, A. Fernandez, M. Sultan, and G. Roma, "A comprehensive assessment of rna-seq protocols for degraded and low-quantity samples," *BMC Genomics*, vol. 18, no. 1, p. 442, 2017.

[171] A. R. Hesketh, "Rna sequencing best practices: Experimental protocol and data analysis," *Methods Mol Biol*, vol. 2049, pp. 113–129, 2019.

[172] M. Hong, S. Tao, L. Zhang, L. T. Diao, X. Huang, S. Huang, S. J. Xie, Z. D. Xiao, and H. Zhang, "Rna sequencing: new technologies and applications in cancer research," *J Hematol Oncol*, vol. 13, no. 1, p. 166, 2020.

[173] T. Nakazato, T. Ohta, and H. Bono, "Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive," *PLoS One*, vol. 8, no. 10, p. e77910, 2013.

[174] R. Stark, M. Grzelak, and J. Hadfield, "Rna sequencing: the teenage years," *Nat Rev Genet*, vol. 20, no. 11, pp. 631–656, 2019.

[175] D. C. Wu, J. Yao, K. S. Ho, A. M. Lambowitz, and C. O. Wilke, "Limitations of alignment-free tools in total rna-seq quantification," *BMC Genomics*, vol. 19, no. 1, p. 510, 2018.

[176] A. Schramm, J. Köster, T. Marschall, M. Martin, M. Schwermer, K. Fielitz, G. Büchel, M. Barann, D. Esser, P. Rosenstiel, S. Rahmann, A. Eggert, and J. H. Schulte, "Next-generation rna sequencing reveals differential expression of mycn target genes and suggests the mtor pathway as a promising therapy target in mycn-amplified neuroblastoma," *Int J Cancer*, vol. 132, no. 3, pp. E106–15, 2013.

[177] M. J. Kling, C. N. Griggs, E. M. McIntyre, G. Alexander, S. Ray, K. B. Challagundla, S. S. Joshi, D. W. Coulter, and N. K. Chaturvedi, "Synergistic efficacy of inhibiting mycn and mtor signaling against neuroblastoma," *BMC Cancer*, vol. 21, no. 1, p. 1061, 2021.

[178] Y. Sun, J. L. Bell, D. Carter, S. Gherardi, R. C. Poulos, G. Milazzo, J. W. Wong, R. Al-Awar, A. E. Tee, P. Y. Liu, B. Liu, B. Atmadibrata, M. Wong, T. Trahair, Q. Zhao, J. M. Shohet, Y. Haupt, J. H. Schulte, P. J. Brown, C. H. Arrowsmith, M. Vedadi, K. L. MacKenzie, S. Hüttelmaier, G. Perini, G. M. Marshall, A. Braithwaite, and T. Liu, "Wdr5 supports an n-myc transcriptional complex that drives a protumorigenic gene expression signature in neuroblastoma," *Cancer Res*, vol. 75, no. 23, pp. 5143–54, 2015.

[179] Q. L. Han, X. L. Zhang, P. X. Ren, L. H. Mei, W. H. Lin, L. Wang, Y. Cao, K. Li, and F. Bai, "Discovery, evaluation and mechanism study of wdr5-targeted small molecular inhibitors for neuroblastoma," *Acta Pharmacol Sin*, vol. 44, no. 4, pp. 877–887, 2023.

[180] O. Yogev, G. S. Almeida, K. T. Barker, S. L. George, C. Kwok, J. Campbell, M. Zarowiecki, D. Kleftogiannis, L. M. Smith, A. Hallsworth, P. Berry, T. Möcklinghoff, H. T. Webber, L. S. Danielson, B. Buttery, E. A. Calton, B. M. da Costa, E. Poon, Y. Jamin, S. Lise, G. J. Veal, N. Sebire, S. P. Robinson, J. Anderson, and L. Chesler, "In vivo modeling of chemoresistant

neuroblastoma provides new insights into chemorefractory disease and metastasis," *Cancer Res*, vol. 79, no. 20, pp. 5382–5393, 2019.

[181] S. Mallik and Z. Zhao, "Identification of gene signatures from rna-seq data using pareto-optimal cluster algorithm," *BMC Syst Biol*, vol. 12, no. Suppl 8, p. 126, 2018.

[182] J. Shen, D. Yan, L. Bai, R. Geng, X. Zhao, H. Li, Y. Dong, J. Cao, Z. Tang, and S. B. Liu, "An 11-gene signature based on treatment responsiveness predicts radiation therapy survival benefit among breast cancer patients," *Front Oncol*, vol. 11, p. 816053, 2021.

[183] M. Jemaà, W. Sime, Y. Abassi, V. A. Lasorsa, J. Bonne Køhler, M. Michaelis, J. Cinatl, M. Capasso, and R. Massoumi, "Gene expression signature of acquired chemoresistance in neuroblastoma cells," *Int J Mol Sci*, vol. 21, no. 18, 2020.

[184] F. Li, W. Zhang, H. Hu, Y. Zhang, J. Li, and D. Huang, "Factors of recurrence after complete response in children with neuroblastoma: A 16-year retrospective study of 179 cases," *Cancer Manag Res*, vol. 14, pp. 107–122, 2022.

[185] Z. Li, H. Takenobu, A. N. Setyawati, N. Akita, M. Haruta, S. Satoh, Y. Shinno, K. Chikaraishi, K. Mukae, J. Akter, R. P. Sugino, A. Nakazawa, A. Nakagawara, H. Aburatani, M. Ohira, and T. Kamijo, "Ezh2 regulates neuroblastoma cell differentiation via ntrk1 promoter epigenetic modifications," *Oncogene*, vol. 37, no. 20, pp. 2714–2727, 2018.

[186] L. V. Bownes, A. P. Williams, R. Marayati, L. L. Stafman, H. Markert, C. H. Quinn, N. Wadhwani, J. M. Aye, J. E. Stewart, K. J. Yoon, E. Mroczek-Musulman, and E. A. Beierle, "Ezh2 inhibition decreases neuroblastoma proliferation and in vivo tumor growth," *PLoS One*, vol. 16, no. 3, p. e0246244, 2021.

[187] R. Tonelli, A. McIntyre, C. Camerin, Z. S. Walters, K. Di Leo, J. Selfe, S. Purgato, E. Missiaglia, A. Tortori, J. Renshaw, A. Astolfi, K. R. Taylor, S. Serravalle, R. Bishop, C. Nanni, L. J. Valentijn, A. Faccini, I. Leuschner, S. Formica, J. S. Reis-Filho, V. Ambrosini, K. Thway, M. Franzoni, B. Summersgill, R. Marchelli, P. Hrelia, G. Cantelli-Forti, S. Fanti, R. Corradini, A. Pession, and J. Shipley, "Antitumor activity of sustained n-myc reduction in rhabdomyosarcomas and transcriptional block by antigene therapy," *Clin Cancer Res*, vol. 18, no. 3, pp. 796–807, 2012.

[188] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, "The rin: an rna integrity number for assigning integrity values to rna measurements," *BMC Mol Biol*, vol. 7, p. 3, 2006.

[189] F. Bewicke-Copley, E. Arjun Kumar, G. Palladino, K. Korfi, and J. Wang, "Applications and analysis of targeted genomic sequencing in cancer studies," *Comput Struct Biotechnol J*, vol. 17, pp. 1348–1359, 2019.

[190] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[191] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with rna-seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–11, 2009.

[192] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with hisat2 and hisat-genotype," *Nat Biotechnol*, vol. 37, no. 8, pp. 907–915, 2019.

[193] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu, "Mapsplice: accurate mapping of rna-seq reads for splice junction discovery," *Nucleic Acids Res*, vol. 38, no. 18, p. e178, 2010.

[194] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for rna-seq data analysis," *Genome Biol*, vol. 17, p. 13, 2016.

[195] L. Wang, S. Wang, and W. Li, "Rseqc: quality control of rna-seq experiments," *Bioinformatics*, vol. 28, no. 16, pp. 2184–5, 2012.

[196] I. V. Deyneko, O. N. Mustafaev, A. Tyurin, K. V. Zhukova, A. Varzari, and I. V. Goldenkova-Pavlova, "Modeling and cleaning rna-seq data significantly improve detection of differentially expressed genes," *BMC Bioinformatics*, vol. 23, no. 1, p. 488, 2022.

[197] Y. Zhao, M. C. Li, M. M. Konaté, L. Chen, B. Das, C. Karlovich, P. M. Williams, Y. A. Evrard, J. H. Doroshow, and L. M. McShane, "Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository," *J Transl Med*, vol. 19, no. 1, p. 269, 2021.

[198] Y. Chen, L. Chen, A. T. L. Lun, P. L. Baldoni, and G. K. Smyth, "edger v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets," *Nucleic Acids Res*, vol. 53, no. 2, 2025.

[199] Y. Liu, J. Zhou, and K. P. White, "Rna-seq differential expression studies: more sequence or more replication?," *Bioinformatics*, vol. 30, no. 3, pp. 301–4, 2014.

[200] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor rna-seq experiments with respect to biological variation," *Nucleic Acids Res*, vol. 40, no. 10, pp. 4288–97, 2012.

[201] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton, "How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use?," *RNA*, vol. 22, no. 6, pp. 839–51, 2016.

[202] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of rna-seq data," *Genome Biol*, vol. 11, no. 3, p. R25, 2010.

[203] V. M. Kvam, P. Liu, and Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from rna-seq data," *Am J Bot*, vol. 99, no. 2, pp. 248–56, 2012.

[204] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and F. S. Consortium, "A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis," *Brief Bioinform*, vol. 14, no. 6, pp. 671–83, 2013.

[205] Z. Gu, R. Eils, and M. Schlesner, "Complex heatmaps reveal patterns and correlations in multidimensional genomic data," *Bioinformatics*, vol. 32, no. 18, pp. 2847–9, 2016.

[206] C. Simillion, R. Liechti, H. E. Lischer, V. Ioannidis, and R. Bruggmann, "Avoiding the pitfalls of gene set enrichment analysis with setrank," *BMC Bioinformatics*, vol. 18, no. 1, p. 151, 2017.

[207] J. Zyla, M. Marczyk, J. Weiner, and J. Polanska, "Ranking metrics in gene set enrichment analysis: do they matter?," *BMC Bioinformatics*, vol. 18, no. 1, p. 256, 2017.

[208] J. Ma, A. Shojaie, and G. Michailidis, "A comparative study of topology-based pathway enrichment analysis methods," *BMC Bioinformatics*, vol. 20, no. 1, p. 546, 2019.

[209] D. Wu and G. K. Smyth, "Camera: a competitive gene set test accounting for inter-gene correlation," *Nucleic Acids Res*, vol. 40, no. 17, p. e133, 2012.

[210] A. Liberzon2015, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The molecular signatures database (msigdb) hallmark gene set collection," *Cell Syst*, vol. 1, no. 6, pp. 417–425, 2015.

[211] K. Blighe and A. Lun, "Pcatools: Everything principal components analysis," 2023.

[212] K. Blighe and A. Lun, "Pcatools: everything principal component analysis," 2024.

[213] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, pp. 179–85, 1965.

[214] V. Davalos and M. Esteller, "Cancer epigenetics in clinical practice," *CA Cancer J Clin*, vol. 73, no. 4, pp. 376–424, 2023.

[215] I. S. Fetahu and S. Taschner-Mandl, "Neuroblastoma and the epigenome," *Cancer Metastasis Rev*, vol. 40, no. 1, pp. 173–189, 2021.

[216] E. R. Lawlor and C. J. Thiele, "Epigenetic changes in pediatric solid tumors: promising new targets," *Clin Cancer Res*, vol. 18, no. 10, pp. 2768–79, 2012.

[217] J. E. Lee and M. Y. Kim, "Cancer epigenetics: Past, present and future," *Semin Cancer Biol*, vol. 83, pp. 4–14, 2022.

[218] R. Duan, W. Du, and W. Guo, "Ezh2: a novel target for cancer treatment," *J Hematol Oncol*, vol. 13, no. 1, p. 104, 2020.

[219] L. Chen, G. Alexe, N. V. Dharia, L. Ross, A. B. Iniguez, A. S. Conway, E. J. Wang, V. Veschi, N. Lam, J. Qi, W. C. Gustafson, N. Nasholm, F. Vazquez, B. A. Weir, G. S. Cowley, L. D. Ali, S. Pantel, G. Jiang, W. F. Harrington, Y. Lee, A. Goodale, R. Lubonja, J. M. Krill-Burger, R. M. Meyers, A. Tsherniak, D. E. Root, J. E. Bradner, T. R. Golub, C. W. Roberts, W. C. Hahn, W. A. Weiss, C. J. Thiele, and K. Stegmaier, "Crispr-cas9 screen reveals a mycn-amplified neuroblastoma dependency on ezh2," *J Clin Invest*, vol. 128, no. 1, pp. 446–462, 2018.

[220] C. Wang, Z. Liu, C. W. Woo, Z. Li, L. Wang, J. S. Wei, V. E. Marquez, S. E. Bates, Q. Jin, J. Khan, K. Ge, and C. J. Thiele, "Ezh2 mediates epigenetic silencing of neuroblastoma suppressor genes casz1, clu, runx3, and ngfr," *Cancer Res*, vol. 72, no. 1, pp. 315–24, 2012.

[221] C. P. Reynolds, K. K. Matthay, J. G. Villablanca, and B. J. Maurer, "Retinoid therapy of high-risk neuroblastoma," *Cancer Lett*, vol. 197, no. 1-2, pp. 185–92, 2003.

[222] R. Masetti, C. Biagi, D. Zama, F. Vendemini, A. Martoni, W. Morello, P. Gasperini, and A. Pession, "Retinoids in pediatric onco-hematology: the model of acute promyelocytic leukemia and neuroblastoma," *Adv Ther*, vol. 29, no. 9, pp. 747–62, 2012.

[223] A. Makimoto, H. Fujisaki, K. Matsumoto, Y. Takahashi, Y. Cho, Y. Morikawa, Y. Yuza, T. Tajiri, and T. Iehara, "Retinoid therapy for neuroblastoma: Historical overview, regulatory challenges, and prospects," *Cancers (Basel)*, vol. 16, no. 3, 2024.

[224] C. P. Reynolds, D. J. Kane, P. A. Einhorn, K. K. Matthay, V. L. Crouse, J. R. Wilbur, S. B. Shurin, and R. C. Seeger, "Response of neuroblastoma to retinoic acid in vitro and in vivo," *Prog Clin Biol Res*, vol. 366, pp. 203–11, 1991.

[225] C. P. Reynolds, P. F. Schindler, D. M. Jones, J. L. Gentile, R. T. Proffitt, and P. A. Einhorn, "Comparison of 13-cis-retinoic acid to trans-retinoic acid using human neuroblastoma cell lines," *Prog Clin Biol Res*, vol. 385, pp. 237–44, 1994.

[226] E. G. Halakos, A. J. Connell, L. Glazewski, S. Wei, and R. W. Mason, "Bottom up proteomics reveals novel differentiation proteins in neuroblastoma cells treated with 13-cis retinoic acid," *J Proteomics*, vol. 209, p. 103491, 2019.

[227] M. W. Zimmerman, A. D. Durbin, S. He, F. Oppel, H. Shi, T. Tao, Z. Li, A. Berezovskaya, Y. Liu, J. Zhang, R. A. Young, B. J. Abraham, and A. T. Look, "Retinoic acid rewires the adrenergic core regulatory circuitry of childhood neuroblastoma," *Sci Adv*, vol. 7, no. 43, p. eabe0834, 2021.

[228] M. Kapalczynska, T. Kolenda, W. Przybyla, M. Zajaczkowska, A. Teresiak, V. Filas, M. Ibbs, R. Blizniak, L. Luczewski, and K. Lamperska, "2d and 3d cell cultures - a comparison of different types of cancer cell cultures," *Arch Med Sci*, vol. 14, no. 4, pp. 910–919, 2018.

[229] S. J. Han, S. Kwon, and K. S. Kim, "Challenges of applying multicellular tumor spheroids in preclinical phase," *Cancer Cell Int*, vol. 21, no. 1, p. 152, 2021.

[230] H. Lu and M. H. Stenzel, "Multicellular tumor spheroids (mcts) as a 3d in vitro evaluation tool of nanoparticles," *Small*, vol. 14, no. 13, p. e1702858, 2018.

[231] D. I. Wallace and X. Guo, "Properties of tumor spheroid growth exhibited by simple mathematical models," *Front Oncol*, vol. 3, p. 51, 2013.

[232] J. L. Harenza, M. A. Diamond, R. N. Adams, M. M. Song, H. L. Davidson, L. S. Hart, M. H. Dent, P. Fortina, C. P. Reynolds, and J. M. Maris, "Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines," *Sci Data*, vol. 4, p. 170033, 2017.

[233] D. Mercatelli, N. Balboni, A. Palma, E. Aleo, P. P. Sanna, G. Perini, and F. M. Giorgi, "Single-cell gene network analysis and transcriptional landscape of mycn-amplified neuroblastoma cell lines," *Biomolecules*, vol. 11, no. 2, 2021.

[234] D. J. Duffy, A. Krstic, M. Halasz, T. Schwarzl, A. Konietzny, K. Iljin, D. G. Higgins, and W. Kolch, "Retinoic acid and tgf-b signalling cooperate to

overcome mycn-induced retinoid resistance," *Genome Med*, vol. 9, no. 1, p. 15, 2017.

[235] J. Frost, A. Ciulli, and S. Rocha, "Rna-seq analysis of phd and vhl inhibitors reveals differences and similarities to the hypoxia response," *Wellcome Open Res*, vol. 4, p. 17, 2019.

[236] S. M. Frumm, Z. P. Fan, K. N. Ross, J. R. Duvall, S. Gupta, L. VerPlank, B. C. Suh, E. Holson, F. F. Wagner, W. B. Smith, R. M. Paranal, C. F. Bassil, J. Qi, G. Roti, A. L. Kung, J. E. Bradner, N. Tolliday, and K. Stegmaier, "Selective hdac1/hdac2 inhibitors induce neuroblastoma differentiation," *Chem Biol*, vol. 20, no. 5, pp. 713–25, 2013.

[237] K. M. Ferguson, S. L. Gillen, L. Chaytor, E. Poon, D. Marcos, R. L. Gomez, L. M. Woods, L. Mykhaylechko, L. Elfari, B. Martins da Costa, Y. Jamin, J. S. Carroll, L. Chesler, F. R. Ali, and A. Philpott, "Palbociclib releases the latent differentiation capacity of neuroblastoma cells," *Dev Cell*, vol. 58, no. 19, pp. 1967–1982.e8, 2023.

[238] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *OMICS*, vol. 16, no. 5, pp. 284–7, 2012.

[239] S. Hänzelmann, R. Castelo, and J. Guinney, "Gsva: gene set variation analysis for microarray and rna-seq data," *BMC Bioinformatics*, vol. 14, p. 7, 2013.

[240] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Res*, vol. 38, no. 12, p. e131, 2010.

[241] V. Bansal, "A computational method for estimating the pcr duplication rate in dna and rna-seq experiments," *BMC Bioinformatics*, vol. 18, no. Suppl 3, p. 43, 2017.

[242] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–20, 2014.

[243] Y. Liao and W. Shi, "Read trimming is not required for mapping and quantification of rna-seq reads at the gene level," *NAR Genom Bioinform*, vol. 2, no. 3, p. lqaa068, 2020.

[244] J. E. Balmer and R. Blomhoff, "Gene expression regulation by retinoic acid," *J Lipid Res*, vol. 43, no. 11, pp. 1773–808, 2002.

[245] R. Romaniello, F. Arrigoni, A. E. Fry, M. T. Bassi, M. I. Rees, R. Borgatti, D. T. Pilz, and T. D. Cushion, "Tubulin genes and malformations of cortical development," *Eur J Med Genet*, vol. 61, no. 12, pp. 744–754, 2018.

[246] R. Loveless and Y. Teng, "Targeting wasf3 signaling in metastatic cancer," *Int J Mol Sci*, vol. 22, no. 2, 2021.

[247] G. Li, Y. Yin, J. Chen, Y. Fan, J. Ma, Y. Huang, C. Chen, P. Dai, S. Chen, and S. Zhao, "Coactosin-like protein 1 inhibits neuronal migration during mouse corticogenesis," *J Vet Sci*, vol. 19, no. 1, pp. 21–26, 2018.

[248] H. Dungrawala, K. P. Bhat, R. Le Meur, W. J. Chazin, X. Ding, S. K. Sharan, S. R. Wessel, A. A. Sathe, R. Zhao, and D. Cortez, "Radx promotes genome stability and modulates chemosensitivity by regulating rad51 at replication forks," *Mol Cell*, vol. 67, no. 3, pp. 374–386.e5, 2017.

[249] A. Cooke, T. Schwarzl, I. Huppertz, G. Kramer, P. Mantas, A. M. Alleaume, W. Huber, J. Krijgsveld, and M. W. Hentze, "The rna-binding protein ybx3 controls amino acid levels by regulating slc mrna abundance," *Cell Rep*, vol. 27, no. 11, pp. 3097–3106.e5, 2019.

[250] A. C. Ross and R. Zolfaghari, "Cytochrome p450s in the regulation of cellular retinoic acid metabolism," *Annu Rev Nutr*, vol. 31, pp. 65–87, 2011.

[251] P. N. Stoney, Y. D. Fragoso, R. B. Saeed, A. Ashton, T. Goodman, C. Simons, M. S. Gomaa, A. Sementilli, L. Sementilli, A. W. Ross, P. J. Morgan, and P. J. McCaffery, "Expression of the retinoic acid catabolic enzyme cyp26b1 in the human brain to maintain signaling homeostasis," *Brain Struct Funct*, vol. 221, no. 6, pp. 3315–26, 2016.

[252] R. J. Sessler and N. Noy, "A ligand-activated nuclear localization signal in cellular retinoic acid binding protein-ii," *Mol Cell*, vol. 18, no. 3, pp. 343–53, 2005.

[253] Y. Hiyama, E. Yamaoka, T. Fukazawa, M. Kojima, Y. Sotomaru, and E. Hiyama, "In vitro transfection of up-regulated genes identified in favorable-outcome neuroblastoma into cell lines," *Cells*, vol. 11, no. 19, 2022.

[254] O. Oppenheimer, N. K. Cheung, and W. L. Gerald, "The ret oncogene is a critical component of transcriptional programs associated with retinoic acid-induced differentiation in neuroblastoma," *Mol Cancer Ther*, vol. 6, no. 4, pp. 1300–9, 2007.

[255] I. Neumann, J. L. Foell, M. Bremer, I. Volkmer, D. Korholz, S. Burdach, and M. S. Staege, "Retinoic acid enhances sensitivity of neuroblastoma cells for imatinib mesylate," *Pediatr Blood Cancer*, vol. 55, no. 3, pp. 464–70, 2010.

[256] R. K. Kam, W. Shi, S. O. Chan, Y. Chen, G. Xu, C. B. Lau, K. P. Fung, W. Y. Chan, and H. Zhao, "Dhrs3 protein attenuates retinoic acid signaling and is required for early embryonic patterning," *J Biol Chem*, vol. 288, no. 44, pp. 31477–87, 2013.

[257] F. Cerignoli, X. Guo, B. Cardinali, C. Rinaldi, J. Casaletto, L. Frati, I. Screpanti, L. J. Gudas, A. Gulino, C. J. Thiele, and G. Giannini, "retsdr1, a short-chain retinol dehydrogenase/reductase, is retinoic acid-inducible and frequently deleted in human neuroblastoma cell lines," *Cancer Res*, vol. 62, no. 4, pp. 1196–204, 2002.

[258] N. Nagy and A. M. Goldstein, "Enteric nervous system development: A crest cell's journey from neural tube to colon," *Semin Cell Dev Biol*, vol. 66, pp. 94–106, 2017.

[259] R. A. Schneider, "Neural crest and the origin of species-specific pattern," *Genesis*, vol. 56, no. 6-7, p. e23219, 2018.

[260] E. Feneck and M. Logan, "The role of retinoic acid in establishing the early limb bud," *Biomolecules*, vol. 10, no. 2, 2020.

[261] F. Broso, P. Gatto, V. Sidarovich, C. Ambrosini, V. De Sanctis, R. Bertorelli, E. Zaccheroni, B. Ricci, E. Destefanis, S. Longhi, E. Sebastiani, T. Tebaldi, V. Adami, and A. Quattrone, "Alpha-1 adrenergic antagonists sensitize neuroblastoma to therapeutic differentiation," *Cancer Res*, vol. 83, no. 16, pp. 2733–2749, 2023.

[262] S. Scarpa, G. D'Orazi, M. Ragano-Caracciolo, P. Cardelli, L. Masuelli, and A. Modesti, "Modulation of laminin synthesis in human neuroblastoma cells during retinoic acid induced differentiation," *Cancer Lett*, vol. 64, no. 1, pp. 31–7, 1992.

[263] Y. Liao, C. H. Chen, T. Xiao, B. de la Peña Avalos, E. V. Dray, C. Cai, S. Gao, N. Shah, Z. Zhang, A. Feit, P. Xue, Z. Liu, M. Yang, J. H. Lee, H. Xu, W. Li, S. Mei, R. S. Pierre, S. Shu, T. Fei, M. Duarte, J. Zhao, J. E. Bradner, K. Polyak, P. W. Kantoff, H. Long, S. P. Balk, X. S. Liu, M. Brown, and K. Xu, "Inhibition of ezh2 transactivation function sensitizes solid tumors to genotoxic stress," *Proc Natl Acad Sci U S A*, vol. 119, no. 3, 2022.

[264] J. Metovic, F. Napoli, S. Osella-Abate, L. Bertero, C. Tampieri, G. Orlando, M. Bianchi, D. Carli, F. Fagioli, M. Volante, and M. Papotti, "Overexpression of insm1, notch1, neurod1, and yap1 genes is associated with adverse clinical outcome in pediatric neuroblastoma," *Virchows Arch*, vol. 481, no. 6, pp. 925–933, 2022.

[265] G. Zhou, D. Sprengers, P. P. C. Boor, M. Doukas, H. Schutz, S. Mancham, A. Pedroza-Gonzalez, W. G. Polak, J. de Jonge, M. Gaspersz, H. Dong, K. Thielemans, Q. Pan, J. N. M. IJzermans, M. J. Bruno, and J. Kwekkeboom, "Antibodies against immune checkpoint molecules restore functions of tumor-infiltrating t cells in hepatocellular carcinomas," *Gastroenterology*, vol. 153, no. 4, pp. 1107–1119.e10, 2017.

[266] B. Kakaradov, J. Arsenio, C. E. Widjaja, Z. He, S. Aigner, P. J. Metz, B. Yu, E. J. Wehrens, J. Lopez, S. H. Kim, E. I. Zuniga, A. W. Goldrath, J. T. Chang, and G. W. Yeo, "Early transcriptional and epigenetic regulation of cd8," *Nat Immunol*, vol. 18, no. 4, pp. 422–432, 2017.

[267] J. Huang, J. Zhang, Z. Guo, C. Li, Z. Tan, J. Wang, J. Yang, and L. Xue, "Easy or not-the advances of ezh2 in regulating t cell development, differentiation, and activation in antitumor immunity," *Front Immunol*, vol. 12, p. 741302, 2021.

[268] X. Chen, G. Cao, J. Wu, X. Wang, Z. Pan, J. Gao, Q. Tian, L. Xu, Z. Li, Y. Hao, Q. Huang, P. Wang, M. Xiao, L. Xie, S. Tang, Z. Liu, L. Hu, J. Tang, R. He, L. Wang, X. Zhou, Y. Wu, M. Chen, B. Sun, B. Zhu, J. Huang, and L. Ye, "The histone methyltransferase ezh2 primes the early differentiation of follicular helper t cells during acute viral infection," *Cell Mol Immunol*, vol. 17, no. 3, pp. 247–260, 2020.

[269] I. H. Su, A. Basavaraj, A. N. Krutchinsky, O. Hobert, A. Ullrich, B. T. Chait, and A. Tarakhovsky, "Ezh2 controls b cell development through histone h3 methylation and igh rearrangement," *Nat Immunol*, vol. 4, no. 2, pp. 124–31, 2003.

[270] M. Guo, M. J. Price, D. G. Patterson, B. G. Barwick, R. R. Haines, A. K. Kania, J. E. Bradley, T. D. Randall, J. M. Boss, and C. D. Scharer, "Ezh2 represses the b cell transcriptional program and regulates antibody-secreting cell metabolism and antibody production," *J Immunol*, vol. 200, no. 3, pp. 1039–1052, 2018.

[271] J. Yin, J. W. Leavenworth, Y. Li, Q. Luo, H. Xie, X. Liu, S. Huang, H. Yan, Z. Fu, L. Y. Zhang, L. Zhang, J. Hao, X. Wu, X. Deng, C. W. Roberts, S. H. Orkin, H. Cantor, and X. Wang, "Ezh2 regulates differentiation and function of natural killer cells through histone methyltransferase activity," *Proc Natl Acad Sci U S A*, vol. 112, no. 52, pp. 15988–93, 2015.

[272] S. Bugide, M. R. Green, and N. Wajapeyee, "Inhibition of enhancer of zeste homolog 2 (ezh2) induces natural killer cell-mediated eradication of hepatocellular carcinoma cells," *Proc Natl Acad Sci U S A*, vol. 115, no. 15, pp. E3509–E3518, 2018.

[273] J. A. Seier, J. Reinhardt, K. Saraf, S. S. Ng, J. P. Layer, D. Corvino, K. Althoff, F. A. Giordano, A. Schramm, M. Fischer, and M. Hölzel, "Druggable epigenetic suppression of interferon-induced chemokine expression linked to mycn amplification in neuroblastoma," *J Immunother Cancer*, vol. 9, 2021.

[274] M. L. Burr, C. E. Sparbier, K. L. Chan, Y. C. Chan, A. Kersbergen, E. Y. N. Lam, E. Azidis-Yates, D. Vassiliadis, C. C. Bell, O. Gilan, S. Jackson, L. Tan, S. Q. Wong, S. Hollizeck, E. M. Michalak, H. V. Siddle, M. T. McCabe, R. K. Prinjha,

G. R. Guerra, B. J. Solomon, S. Sandhu, S. J. Dawson, P. A. Beavis, R. W. Tothill, C. Cullinane, P. J. Lehner, K. D. Sutherland, and M. A. Dawson, "An evolutionarily conserved function of polycomb silences the mhc class i antigen presentation pathway and enables immune evasion in cancer," *Cancer Cell*, vol. 36, no. 4, pp. 385–401.e8, 2019.

[275] L. R. Anderson, T. W. Owens, and M. J. Naylor, "Structural and mechanical functions of integrins," *Biophys Rev*, vol. 6, no. 2, pp. 203–213, 2014.

[276] M. Gunawan, N. Venkatesan, J. T. Loh, J. F. Wong, H. Berger, W. H. Neo, L. Y. Li, M. K. La Win, Y. H. Yau, T. Guo, P. C. See, S. Yamazaki, K. C. Chin, A. R. Gingras, S. G. Shochat, L. G. Ng, S. K. Sze, F. Ginhoux, and I. H. Su, "The methyltransferase ezh2 controls cell adhesion and migration through direct methylation of the extranuclear regulatory protein talin," *Nat Immunol*, vol. 16, no. 5, pp. 505–16, 2015.

[277] L. Zhang, J. Qu, Y. Qi, Y. Duan, Y. W. Huang, Z. Zhou, P. Li, J. Yao, B. Huang, S. Zhang, and D. Yu, "Ezh2 engages tgfb signaling to promote breast cancer bone metastasis via integrin b1-fak activation," *Nat Commun*, vol. 13, no. 1, p. 2543, 2022.

[278] C. Rozzo, V. Chiesa, and M. Ponzoni, "Integrin up-regulation as marker of neuroblastoma cell differentiation: correlation with neurite extension," *Cell Death Differ*, vol. 4, no. 8, pp. 713–24, 1997.

[279] M. Podkowa, T. Christova, X. Zhao, Y. Jian, and L. Attisano, "p21-activated kinase (pak) is required for bone morphogenetic protein (bmp)-induced dendritogenesis in cortical neurons," *Mol Cell Neurosci*, vol. 57, pp. 83–92, 2013.

[280] A. Majdazari, J. Stubbusch, C. M. Müller, M. Hennchen, M. Weber, C. X. Deng, Y. Mishina, G. Schütz, T. Deller, and H. Rohrer, "Dendrite complexity of sympathetic neurons is controlled during postnatal development by bmp signaling," *J Neurosci*, vol. 33, no. 38, pp. 15132–44, 2013.

[281] N. Gulati, W. Béguelin, and L. Giulino-Roth, "Enhancer of zeste homolog 2 (ezh2) inhibitors," *Leuk Lymphoma*, vol. 59, no. 7, pp. 1574–1585, 2018.

[282] C. Li, J. Song, Z. Guo, Y. Gong, T. Zhang, J. Huang, R. Cheng, X. Yu, Y. Li, L. Chen, X. Ma, Y. Sun, Y. Wang, and L. Xue, "Ezh2 inhibitors suppress colorectal cancer by regulating macrophage polarization in the tumor microenvironment," *Front Immunol*, vol. 13, p. 857808, 2022.

[283] H. Quentmeier, C. Pommerenke, V. Hauer, C. C. Uphoff, M. Zaborski, and H. G. Drexler, "Ezh2-activating mutation: no reliable indicator for efficacy of methyltransferase inhibitors," *Leuk Lymphoma*, vol. 61, no. 12, pp. 2885–2893, 2020.

[284] A. Lenard, H. M. Xie, T. Pastuer, T. Shank, C. Libbrecht, M. Kingsley, S. S. Riedel, Z. F. Yuan, N. Zhu, T. Neff, and K. M. Bernt, "Epigenetic regulation of protein translation in kmt2a-rearranged aml," *Exp Hematol*, vol. 85, pp. 57–69, 2020.

[285] R. Nguyen, J. Houston, W. K. Chan, D. Finkelstein, and M. A. Dyer, "The role of interleukin-2, all-trans retinoic acid, and natural killer cells: surveillance mechanisms in anti-gd2 antibody therapy in neuroblastoma," *Cancer Immunol Immunother*, vol. 67, no. 4, pp. 615–626, 2018.

[286] H. Sallmon, V. Hoene, S. C. Weber, and C. Dame, "Differentiation of human sh-sy5y neuroblastoma cells by all-trans retinoic acid activates the interleukin-18 system," *J Interferon Cytokine Res*, vol. 30, no. 2, pp. 55–8, 2010.

[287] K. D. Yang, S. N. Cheng, N. C. Wu, and M. F. Shaio, "Induction of interleukin-8 expression in neuroblastoma cells by retinoic acid: implication of leukocyte chemotaxis and activation," *Pediatr Res*, vol. 34, no. 6, pp. 720–4, 1993.

[288] M. Bouillon and M. Audette, "Transduction of retinoic acid and gamma-interferon signal for intercellular adhesion molecule-1 expression on human tumor cell lines: evidence for the late-acting involvement of protein kinase c inactivation," *Cancer Res*, vol. 53, no. 4, pp. 826–32, 1993.

[289] S. Vertuani, A. De Geer, V. Levitsky, P. Kogner, R. Kiessling, and J. Levitskaya, "Retinoids act as multistep modulators of the major histocompatibility class i presentation pathway and sensitize neuroblastomas to cytotoxic lymphocytes," *Cancer Res*, vol. 63, no. 22, pp. 8006–13, 2003.

[290] G. R. Stark and W. R. Taylor, "Analyzing the g2/m checkpoint," *Methods Mol Biol*, vol. 280, pp. 51–82, 2004.

[291] Z. Wu, S. T. Lee, Y. Qiao, Z. Li, P. L. Lee, Y. J. Lee, X. Jiang, J. Tan, M. Aau, C. Z. Lim, and Q. Yu, "Polycomb protein ezh2 regulates cancer cell fate decision in response to dna damage," *Cell Death Differ*, vol. 18, no. 11, pp. 1771–9, 2011.

[292] L. Xu, H. Tang, K. Wang, Y. Zheng, J. Feng, H. Dong, Y. Jin, C. Cao, X. Chen, and G. Gao, "Pharmacological inhibition of ezh2 combined with dna-damaging agents interferes with the dna damage response in mm cells," *Mol Med Rep*, vol. 19, no. 5, pp. 4249–4255, 2019.

[293] H. Xia, W. Zhang, Y. Li, N. Guo, and C. Yu, "Ezh2 silencing with rna interference induces g2/m arrest in human lung cancer cells in vitro," *Biomed Res Int*, vol. 2014, p. 348728, 2014.

[294] L. Guarrera, M. Kurosaki, S. K. Garattini, M. Gianni', G. Fasola, L. Rossit, M. Prisciandaro, M. Di Bartolomeo, M. Bolis, P. Rizzo, C. Nastasi, M. Foglia, A. Zanetti, G. Paroni, M. Terao, and E. Garattini, "Anti-tumor activity of

all-trans retinoic acid in gastric-cancer: gene-networks and molecular mechanisms," *J Exp Clin Cancer Res*, vol. 42, no. 1, p. 298, 2023.

[295] R. J. Zhou, X. Q. Yang, D. Wang, Q. Zhou, L. Xia, M. X. Li, L. L. Zeng, G. Wang, and Z. Z. Yang, "Anti-tumor effects of all-trans retinoic acid are enhanced by genistein," *Cell Biochem Biophys*, vol. 62, no. 1, pp. 177–84, 2012.

[296] X. Wang, V. C. Lui, R. T. Poon, P. Lu, and R. Y. Poon, "Dna damage mediated s and g(2) checkpoints in human embryonal carcinoma cells," *Stem Cells*, vol. 27, no. 3, pp. 568–76, 2009.

[297] C. Thirant, A. Peltier, S. Durand, A. Kramdi, C. Louis-Brennetot, C. Pierre-Eugène, M. Gautier, A. Costa, A. Grelier, S. Zaïdi, N. Gruel, I. Jimenez, E. Lapouble, G. Pierron, D. Sitbon, H. J. Brisse, A. Gauthier, P. Fréneaux, S. Grossetête, L. G. Baudrin, V. Raynal, S. Baulande, A. Bellini, J. Bhalshankar, A. M. Carcaboso, B. Geoerger, H. Rohrer, D. Surdez, V. Boeva, G. Schleiermacher, O. Delattre, and I. Janoueix-Lerosey, "Reversible transitions between noradrenergic and mesenchymal tumor identities define cell plasticity in neuroblastoma," *Nat Commun*, vol. 14, no. 1, p. 2575, 2023.

[298] R. L. Gomez, L. M. Woods, R. Ramachandran, A. N. Abou Tayoun, A. Philpott, and F. R. Ali, "Super-enhancer associated core regulatory circuits mediate susceptibility to retinoic acid in neuroblastoma cells," *Front Cell Dev Biol*, vol. 10, p. 943924, 2022.

[299] S. Hettmer, Z. Li, A. N. Billin, F. G. Barr, D. D. Cornelison, A. R. Ehrlich, D. C. Guttridge, A. Hayes-Jordan, L. J. Helman, P. J. Houghton, J. Khan, D. M. Langenau, C. M. Linardic, R. Pal, T. A. Partridge, G. K. Pavlath, R. Rota, B. W. Schäfer, J. Shipley, B. Stillman, L. H. Wexler, A. J. Wagers, and C. Keller, "Rhabdomyosarcoma: current challenges and their implications for developing therapies," *Cold Spring Harb Perspect Med*, vol. 4, no. 11, p. a025650, 2014.

[300] D. O. Walterhouse, A. S. Pappo, J. L. Meza, J. C. Breneman, A. A. Hayes-Jordan, D. M. Parham, T. P. Cripe, J. R. Anderson, W. H. Meyer, and D. S. Hawkins, "Shorter-duration therapy using vincristine, dactinomycin, and lower-dose cyclophosphamide with or without radiotherapy for patients with newly diagnosed low-risk rhabdomyosarcoma: a report from the soft tissue sarcoma committee of the children's oncology group," *J Clin Oncol*, vol. 32, no. 31, pp. 3547–52, 2014.

[301] M. Li, J. Zhao, X. Li, Y. Chen, C. Feng, F. Qian, Y. Liu, J. Zhang, J. He, B. Ai, Z. Ning, W. Liu, X. Bai, X. Han, Z. Wu, X. Xu, Z. Tang, Q. Pan, L. Xu, C. Li, Q. Wang, and E. Li, "Hifresp: A novel high-frequency sub-pathway mining approach to identify robust prognostic gene signatures," *Brief Bioinform*, vol. 21, no. 4, pp. 1411–1424, 2020.

[302] M. Wachtel, M. Dettling, E. Koscielniak, S. Stegmaier, J. Treuner, K. Simon-Klingenstein, P. Bühlmann, F. K. Niggli, and B. W. Schäfer, "Gene expression signatures identify rhabdomyosarcoma subtypes and detect a novel t(2;2)(q35;p23) translocation fusing pax3 to ncoa1," *Cancer Res*, vol. 64, no. 16, pp. 5539–45, 2004.

[303] E. Davicioni, J. R. Anderson, J. D. Buckley, W. H. Meyer, and T. J. Triche, "Gene expression profiling for survival prediction in pediatric rhabdomyosarcomas: a report from the children's oncology group," *J Clin Oncol*, vol. 28, no. 7, pp. 1240–6, 2010.

[304] P. Hingorani, E. Missiaglia, J. Shipley, J. R. Anderson, T. J. Triche, M. Delorenzi, J. Gastier-Foster, M. Wing, D. S. Hawkins, and S. X. Skapek, "Clinical application of prognostic gene expression signature in fusion gene-negative rhabdomyosarcoma: A report from the children's oncology group," *Clin Cancer Res*, vol. 21, no. 20, pp. 4733–9, 2015.

[305] J. Sun, J. Zhao, Z. Yang, Z. Zhou, and P. Lu, "Identification of gene signatures and potential therapeutic targets for acquired chemotherapy resistance in gastric cancer patients," *J Gastrointest Oncol*, vol. 12, no. 2, pp. 407–422, 2021.

[306] J. Wu, Y. Tian, W. Liu, H. Zheng, Y. Xi, Y. Yan, Y. Hu, B. Liao, M. Wang, and P. Tang, "A novel twelve-gene signature to predict neoadjuvant chemotherapy response and prognosis in breast cancer," *Front Immunol*, vol. 13, p. 1035667, 2022.

[307] C. A. Barrón-Gallardo, M. Garcia-Chagollán, A. J. Morán-Mendoza, R. Delgadillo-Cristerna, M. G. Martínez-Silva, M. M. Villaseñor-García, A. Aguilar-Lemarroy, and L. F. Jave-Suárez, "A gene expression signature in her2+ breast cancer patients related to neoadjuvant chemotherapy resistance, overall survival, and disease-free survival," *Front Genet*, vol. 13, p. 991706, 2022.

[308] P. Modarres, F. Mohamadi Farsani, A. A. Nekouie, and S. Vallian, "Meta-analysis of gene signatures and key pathways indicates suppression of jnk pathway as a regulator of chemo-resistance in aml," *Sci Rep*, vol. 11, no. 1, p. 12485, 2021.

[309] A. A. Emran, D. M. Marzese, D. R. Menon, H. Hammerlindl, F. Ahmed, E. Richtig, P. Duijf, D. S. Hoon, and H. Schaider, "Commonly integrated epigenetic modifications of differentially expressed genes lead to adaptive resistance in cancer," *Epigenomics*, vol. 11, no. 7, pp. 732–737, 2019.

[310] T. Kenarangi, E. Bakhshi, K. InanlooRahatloo, and A. Biglarian, "Identification of gene signature in rna-seq hepatocellular carcinoma data by pareto-optimal cluster algorithm," *Gastroenterol Hepatol Bed Bench*, vol. 15, no. 4, pp. 387–394, 2022.

[311] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, 2008.

[312] A. S. Nangraj, G. Selvaraj, S. Kaliamurthi, A. C. Kaushik, W. C. Cho, and D. Q. Wei, "Integrated ppi- and wgcna-retrieval of hub gene signatures shared between barrett's esophagus and esophageal adenocarcinoma," *Front Pharmacol*, vol. 11, p. 881, 2020.

[313] M. Xu, T. Ouyang, K. Lv, and X. Ma, "Integrated wgcna and ppi network to screen hub genes signatures for infantile hemangioma," *Front Genet*, vol. 11, p. 614195, 2020.

[314] Y. Zhao, T. Ma, and D. Zou, "Identification of unique transcriptomic signatures and hub genes through rna sequencing and integrated wgcna and ppi network analysis in nonerosive reflux disease," *J Inflamm Res*, vol. 14, pp. 6143–6156, 2021.

[315] Q. Zhou, L. Q. Zhou, S. H. Li, Y. W. Yuan, L. Liu, J. L. Wang, D. Z. Wu, Y. Wu, and L. Xin, "Identification of subtype-specific genes signature by wgcna for prognostic prediction in diffuse type gastric cancer," *Aging (Albany NY)*, vol. 12, no. 17, pp. 17418–17435, 2020.

[316] N. Daneshafrooz, M. Bagherzadeh Cham, M. Majidi, and B. Panahi, "Identification of potentially functional modules and diagnostic genes related to amyotrophic lateral sclerosis based on the wgcna and lasso algorithms," *Sci Rep*, vol. 12, no. 1, p. 20144, 2022.

[317] L. Zhang, X. Zhang, S. Fan, and Z. Zhang, "Identification of modules and hub genes associated with platinum-based chemotherapy resistance and treatment response in ovarian cancer by weighted gene co-expression network analysis," *Medicine (Baltimore)*, vol. 98, no. 44, p. e17803, 2019.

[318] S. G. Danielli, Y. Wei, M. A. Dyer, E. Stewart, H. Sheppard, M. Wachtel, B. W. Schäfer, A. G. Patel, and D. M. Langenau, "Single cell transcriptomic profiling identifies tumor-acquired and therapy-resistant cell states in pediatric rhabdomyosarcoma," *Nat Commun*, vol. 15, no. 1, p. 6307, 2024.

[319] K. Skrzypek, G. Adamek, M. Kot, B. Badyra, and M. Majka, "Progression and differentiation of alveolar rhabdomyosarcoma is regulated by pax7 transcription factor-significance of tumor subclones," *Cells*, vol. 10, no. 8, 2021.

[320] L. E. Dawson, L. D'Agostino, A. A. Hakim, R. D. Lackman, S. A. Brown, R. B. Sensenig, Z. A. Antonello, and I. I. Kuzin, "Induction of myogenic differentiation improves chemosensitivity of chemoresistant cells in soft-tissue sarcoma cell lines," *Sarcoma*, vol. 2020, p. 8647981, 2020.

[321] A. R. Hinson, R. Jones, L. E. Crose, B. C. Belyea, F. G. Barr, and C. M. Linardic, "Human rhabdomyosarcoma cell lines for rhabdomyosarcoma research: utility and pitfalls," *Front Oncol*, vol. 3, p. 183, 2013.

[322] M. Locard-Paulet, O. Palasca, and L. J. Jensen, "Identifying the genes impacted by cell proliferation in proteomics and transcriptomics studies," *PLoS Comput Biol*, vol. 18, no. 10, p. e1010604, 2022.

[323] F. L. Zhang, W. D. Li, G. Zhang, M. Zhang, Z. J. Liu, K. X. Zhu, Q. C. Liu, S. E. Zhang, W. Shen, and X. F. Zhang, "Identification of unique transcriptomic signatures through integrated multispecies comparative analysis and wgcna in bovine oocyte development," *BMC Genomics*, vol. 24, no. 1, p. 265, 2023.

[324] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering, "The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Res*, vol. 49, no. D1, pp. D605–D612, 2021.

[325] Y. Shou, L. Yang, Y. Yang, X. Zhu, F. Li, and J. Xu, "Identification of signatures of prognosis prediction for melanoma using a hypoxia score," *Front Genet*, vol. 11, p. 570530, 2020.

[326] M. D. Catalina, P. Bachali, N. S. Geraci, A. C. Grammer, and P. E. Lipsky, "Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus," *Commun Biol*, vol. 2, p. 140, 2019.

[327] X. Ling, L. Zhang, C. Fang, H. Liang, J. Zhu, and J. Ma, "Development of a cuproptosis-related signature for prognosis prediction in lung adenocarcinoma based on wgcna," *Transl Lung Cancer Res*, vol. 12, no. 4, pp. 754–769, 2023.

[328] A. A. Ayanlaja, Y. Xiong, Y. Gao, G. Ji, C. Tang, Z. Abdikani Abdullah, and D. Gao, "Distinct features of doublecortin as a marker of neuronal migration and its implications in cancer cell mobility," *Front Mol Neurosci*, vol. 10, p. 199, 2017.

[329] X. Zi, G. Zhang, and S. Qiu, "Up-regulation of linc00619 promotes apoptosis and inhibits proliferation, migration and invasion while promoting apoptosis of osteosarcoma cells through inactivation of the hgf-mediated pi3k-akt signalling pathway," *Epigenetics*, vol. 17, no. 2, pp. 147–160, 2022.

[330] S. Fulda, "Cell death pathways as therapeutic targets in rhabdomyosarcoma," *Sarcoma*, vol. 2012, p. 326210, 2012.

[331] X. Lian, J. S. Bond, N. Bharathy, S. P. Boudko, E. Pokidysheva, J. F. Shern, M. Lathara, T. Sasaki, T. Settelmeyer, M. M. Cleary, A. Bajwa, G. Srinivasa, C. P. Hartley, H. P. Bächinger, A. Mansoor, S. H. Gultekin, N. E. Berlow, and C. Keller,

"Defining the extracellular matrix of rhabdomyosarcoma," *Front Oncol*, vol. 11, p. 601957, 2021.

[332] B. Bhushan, R. Iranpour, A. Eshtiaghi, S. C. da Silva Rosa, B. W. Lindsey, J. W. Gordon, and S. Ghavami, "Transforming growth factor beta and alveolar rhabdomyosarcoma: A challenge of tumor differentiation and chemotherapy response," *Int J Mol Sci*, vol. 25, no. 5, 2024.

[333] G. Seitz, S. W. Warmann, C. O. Vokuhl, H. Heitmann, C. Treuner, I. Leuschner, and J. Fuchs, "Effects of standard chemotherapy on tumor growth and regulation of multidrug resistance genes and proteins in childhood rhabdomyosarcoma," *Pediatr Surg Int*, vol. 23, no. 5, pp. 431–9, 2007.

[334] C. W. Zheng, R. J. Zeng, L. Y. Xu, and E. M. Li, "Rho gtpases: Promising candidates for overcoming chemotherapeutic resistance," *Cancer Lett*, vol. 475, pp. 65–78, 2020.

[335] M. N. Hayes, K. McCarthy, A. Jin, M. L. Oliveira, S. Iyer, S. P. Garcia, S. Sindiri, B. Gryder, Z. Motala, G. P. Nielsen, J. P. Borg, M. van de Rijn, D. Malkin, J. Khan, M. S. Ignatius, and D. M. Langenau, "Vangl2/rhoa signaling pathway regulates stem cell self-renewal programs and growth in rhabdomyosarcoma," *Cell Stem Cell*, vol. 22, no. 3, pp. 414–427.e6, 2018.

[336] C. Liu, L. Zhang, W. Cui, J. Du, Z. Li, Y. Pang, Q. Liu, H. Shang, L. Meng, W. Li, L. Song, P. Wang, Y. Xie, Y. Wang, Y. Liu, J. Hu, W. Zhang, and F. Li, "Epigenetically upregulated geft-derived invasion and metastasis of rhabdomyosarcoma via epithelial mesenchymal transition promoted by the rac1/cdc42-pak signalling pathway," *EBioMedicine*, vol. 50, pp. 122–134, 2019.

[337] S. Thuault, F. Comunale, J. Hasna, M. Fortier, D. Planchon, N. Elarouci, A. De Reynies, S. Bodin, A. Blangy, and C. Gauthier-Rouvière, "The rhoe/rock/arhgap25 signaling pathway controls cell invasion by inhibition of rac activity," *Mol Biol Cell*, vol. 27, no. 17, pp. 2653–61, 2016.

[338] T. Q. Pham, K. Robinson, L. Xu, M. N. Pavlova, S. X. Skapek, and E. Y. Chen, "Hdac6 promotes growth, migration/invasion, and self-renewal of rhabdomyosarcoma," *Oncogene*, vol. 40, no. 3, pp. 578–591, 2021.

[339] A. Jahangiri and W. A. Weiss, "It takes two to tango: Dual inhibition of pi3k and mapk in rhabdomyosarcoma," *Clin Cancer Res*, vol. 19, no. 21, pp. 5811–3, 2013.

[340] L. Li, T. Xue, W. Xu, and B. Zhou, "Effect of matrine combined with cisplatin on the expression of xiap in human rhabdomyosarcoma rd cells," *Oncol Lett*, vol. 12, no. 5, pp. 3793–3798, 2016.

[341] S. V. Holt, K. E. Brookes, C. Dive, and G. W. Makin, "Down-regulation of xiap by aeg35156 in paediatric tumour cells induces apoptosis and sensitises cells to cytotoxic agents," *Oncol Rep*, vol. 25, no. 4, pp. 1177–81, 2011.

[342] A. B. Rovsing, E. A. Thomsen, I. Nielsen, T. W. Skov, Y. Luo, K. Dybkaer, and J. G. Mikkelsen, "Resistance to vincristine in dlbcl by disruption of p53-induced cell cycle arrest and apoptosis mediated by kif18b and usp28," *Br J Haematol*, vol. 202, no. 4, pp. 825–839, 2023.

[343] A. Zarrabi, D. Perrin, M. Kavoosi, M. Sommer, S. Sezen, P. Mehrbod, B. Bhushan, F. Machaj, J. Rosik, P. Kawalec, S. Afifi, S. M. Bolandi, P. Koleini, M. Taheri, T. Madrakian, M. J. Łos, B. Lindsey, N. Cakir, A. Zarepour, K. Hushmandi, A. Fallah, B. Koc, A. Khosravi, M. Ahmadi, S. Logue, G. Orive, S. Pecic, J. W. Gordon, and S. Ghavami, "Rhabdomyosarcoma: Current therapy, challenges, and future approaches to treatment strategies," *Cancers (Basel)*, vol. 15, no. 21, 2023.

[344] M. Michaelis, F. Rothweiler, N. Löschmann, M. Sharifi, T. Ghafourian, and J. Cinatl, "Enzastaurin inhibits abcb1-mediated drug efflux independently of effects on protein kinase c signalling and the cellular p53 status," *Oncotarget*, vol. 6, no. 19, pp. 17605–20, 2015.

[345] A. S. Pappo, J. R. Anderson, W. M. Crist, M. D. Wharam, P. P. Breitfeld, D. Hawkins, R. B. Raney, R. B. Womer, D. M. Parham, S. J. Qualman, and H. E. Grier, "Survival after relapse in children and adolescents with rhabdomyosarcoma: A report from the intergroup rhabdomyosarcoma study group," *J Clin Oncol*, vol. 17, no. 11, pp. 3487–93, 1999.

[346] R. B. Raney, W. M. Crist, H. M. Maurer, and M. A. Foulkes, "Prognosis of children with soft tissue sarcoma who relapse after achieving a complete response. a report from the intergroup rhabdomyosarcoma study i," *Cancer*, vol. 52, no. 1, pp. 44–50, 1983.

[347] A. K. Mitra, T. Harding, U. K. Mukherjee, J. S. Jang, Y. Li, R. HongZheng, J. Jen, P. Sonneveld, S. Kumar, W. M. Kuehl, V. Rajkumar, and B. Van Ness, "A gene expression signature distinguishes innate response and resistance to proteasome inhibitors in multiple myeloma," *Blood Cancer J*, vol. 7, no. 6, p. e581, 2017.

[348] X. Zhu, X. Tian, L. Ji, X. Zhang, Y. Cao, C. Shen, Y. Hu, J. W. H. Wong, J. Y. Fang, J. Hong, and H. Chen, "A tumor microenvironment-specific gene expression signature predicts chemotherapy resistance in colorectal cancer patients," *NPJ Precis Oncol*, vol. 5, no. 1, p. 7, 2021.

[349] D. X. He, Y. D. Xia, X. T. Gu, J. Jin, and X. Ma, "A transcription/translation-based gene signature predicts resistance to chemotherapy in breast cancer," *J Pharm Biomed Anal*, vol. 102, pp. 500–8, 2015.

[350] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, "Microarray and its applications," *J Pharm Bioallied Sci*, vol. 4, no. Suppl 2, pp. S310–2, 2012.

[351] H. Liu, I. Bebu, and X. Li, "Microarray probes and probe sets," *Front Biosci (Elite Ed)*, vol. 2, no. 1, pp. 325–38, 2010.

[352] E. Davicioni, F. G. Finckenstein, V. Shahbazian, J. D. Buckley, T. J. Triche, and M. J. Anderson, "Identification of a pax-fkhr gene expression signature that defines molecular classes and determines the prognosis of alveolar rhabdomyosarcomas," *Cancer Res*, vol. 66, no. 14, pp. 6936–46, 2006.

[353] S. Davis and P. S. Meltzer, "Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–7, 2007.

[354] A. Kauffmann, T. F. Rayner, H. Parkinson, M. Kapushesky, M. Lukk, A. Brazma, and W. Huber, "Importing arrayexpress datasets into r/bioconductor," *Bioinformatics*, vol. 25, no. 16, pp. 2092–4, 2009.

[355] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–64, 2003.

[356] M. N. McCall, B. M. Bolstad, and R. A. Irizarry, "Frozen robust multiarray analysis (frma)," *Biostatistics*, vol. 11, no. 2, pp. 242–53, 2010.

[357] E. A. Welsh, S. A. Eschrich, A. E. Berglund, and D. A. Fenstermacher, "Iterative rank-order normalization of gene expression microarray data," *BMC Bioinformatics*, vol. 14, p. 153, 2013.

[358] Y. Sui, X. Zhao, T. P. Speed, and Z. Wu, "Background adjustment for dna microarrays using a database of microarray experiments," *J Comput Biol*, vol. 16, no. 11, pp. 1501–15, 2009.

[359] Q. Li, N. J. Birkbak, B. Gyorffy, Z. Szallasi, and A. C. Eklund, "Jetset: selecting the optimal microarray probe set to represent a gene," *BMC Bioinformatics*, vol. 12, p. 474, 2011.

[360] D. A. Rodeberg, N. Garcia-Henriquez, E. R. Lyden, E. Davicioni, D. M. Parham, S. X. Skapek, A. A. Hayes-Jordan, S. S. Donaldson, K. L. Brown, T. J. Triche, W. H. Meyer, and D. S. Hawkins, "Prognostic significance and tumor biology of regional lymph node disease in patients with rhabdomyosarcoma: a report from the children's oncology group," *J Clin Oncol*, vol. 29, no. 10, pp. 1304–11, 2011.

[361] D. Huang, P. Watal, D. Drehner, D. Dhar, and T. Chandra, "Rhabdomyosarcoma with diffuse bone marrow metastases," *Cureus*, vol. 14, no. 2, p. e21863, 2022.

[362] S. Ognjanovic, A. M. Linabery, B. Charbonneau, and J. A. Ross, "Trends in childhood rhabdomyosarcoma incidence and survival in the united states, 1975-2005," *Cancer*, vol. 115, no. 18, pp. 4218–26, 2009.

[363] G. Bisogno, V. Minard-Colin, J. Haduong, I. Zanetti, A. Ferrari, J. Chisholm, C. M. Heske, R. Hladun, M. Jenney, J. H. M. Merks, and R. Venkatramani, "Implications of implementing children's oncology group risk stratification to patients with rhabdomyosarcoma treated on european paediatric soft tissue sarcoma study group clinical trial," *Pediatr Blood Cancer*, p. e31436, 2024.

[364] M. S. Rao, T. R. Van Vleet, R. Ciurlionis, W. R. Buck, S. W. Mittelstadt, E. A. G. Blomme, and M. J. Liguori, "Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies," *Front Genet*, vol. 9, p. 636, 2018.

[365] Y. Hu, Z. He2024, S. Liu, W. Ying, Y. Chen, M. Zhao, M. He, X. Wu, Y. Tang, W. Gu, M. Ying, J. Wang, and T. Tao, "Patient-derived rhabdomyosarcoma cells recapitulate the genetic and transcriptomic landscapes of primary tumors," *iScience*, vol. 27, no. 10, p. 110862, 2024.

[366] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, and P. S. Meltzer, "Gene expression profiling of alveolar rhabdomyosarcoma with cdna microarrays," *Cancer Res*, vol. 58, no. 22, pp. 5009–13, 1998.

[367] C. C. Whiteford, S. Bilke, B. T. Greer, Q. Chen, T. A. Braunschweig, N. Cenacchi, J. S. Wei, M. A. Smith, P. Houghton, C. Morton, C. P. Reynolds, R. Lock, R. Gorlick, C. Khanna, C. J. Thiele, M. Takikita, D. Catchpoole, S. M. Hewitt, and J. Khan, "Credentialing preclinical pediatric xenograft models using gene expression and tissue microarray analysis," *Cancer Res*, vol. 67, no. 1, pp. 32–40, 2007.

[368] E. H. Weissler, T. Naumann, T. Andersson, R. Ranganath, O. Elemento, Y. Luo, D. F. Freitag, J. Benoit, M. C. Hughes, F. Khan, P. Slater, K. Shameer, M. Roe, E. Hutchison, S. H. Kollins, U. Broedl, Z. Meng, J. L. Wong, L. Curtis, E. Huang, and M. Ghassemi, "The role of machine learning in clinical research: transforming the future of evidence generation," *Trials*, vol. 22, no. 1, p. 537, 2021.

[369] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," *Nat Rev Mol Cell Biol*, vol. 23, no. 1, pp. 40–55, 2022.

[370] P. Terrematte, D. S. Andrade, J. Justino, B. Stransky, D. S. A. de Araújo, and A. D. Dória Neto, "A novel machine learning 13-gene signature: Improving risk analysis and survival prediction for clear cell renal cell carcinoma patients," *Cancers (Basel)*, vol. 14, no. 9, 2022.

[371] X. Liu, P. Tao, H. Su, and Y. Li, "Machine learning-random forest model was used to construct gene signature associated with cuproptosis to predict the prognosis of gastric cancer," *Sci Rep*, vol. 15, no. 1, p. 4170, 2025.

[372] A. Altaf, J. Kawashima, M. Khalil, H. Stecko, Z. Rashid, M. Kalady, and T. M. Pawlik, "Identification of a gene signature and prediction of overall survival of patients with stage iv colorectal cancer using a novel machine learning approach," *Eur J Surg Oncol*, vol. 51, no. 5, p. 109718, 2025.

[373] N. N. Al-Bzour, A. N. Al-Bzour, A. Qasaymeh, A. Saeed, and L. Chen, "Machine learning approach identifies inflammatory gene signature for predicting survival outcomes in hepatocellular carcinoma," *Sci Rep*, vol. 14, no. 1, p. 30328, 2024.

[374] D. J. Ho, N. P. Agaram, A. O. Frankel, M. Lathara, D. Catchpoole, C. Keller, and M. R. Hameed, "Toward deploying a deep learning model for diagnosis of rhabdomyosarcoma," *Mod Pathol*, vol. 37, no. 3, p. 100421, 2024.

[375] A. O. Frankel, M. Lathara, C. Y. Shaw, O. Wogmon, J. M. Jackson, M. M. Clark, N. Eshraghi, S. E. Keenen, A. D. Woods, R. Purohit, Y. Ishi, N. Moran, M. Eguchi, F. U. A. Ahmed, S. Khan, M. Ioannou, K. Perivoliotis, P. Li, H. Zhou, A. Alkhaledi, E. J. Davis, D. Galipeau, R. L. Randall, A. Wozniak, P. Schoffski, C. J. Lee, P. H. Huang, R. L. Jones, B. P. Rubin, M. Darrow, G. Srinivasa, E. R. Rudzinski, S. Chen, N. E. Berlow, and C. Keller, "Machine learning for rhabdomyosarcoma histopathology," *Mod Pathol*, vol. 35, no. 9, pp. 1193–1203, 2022.

[376] X. Zhang, S. Wang, E. R. Rudzinski, S. Agarwal, R. Rong, D. A. Barkauskas, O. Daescu, L. Furman Cline, R. Venkatramani, Y. Xie, G. Xiao, and P. Leavey, "Deep learning of rhabdomyosarcoma pathology images for classification and survival outcome prediction," *Am J Pathol*, vol. 192, no. 6, pp. 917–925, 2022.

[377] P. Tang, B. Li, Z. Zhou, H. Wang, M. Ma, L. Gong, Y. Qiao, P. Ren, and H. Zhang, "Integrated machine learning developed a prognosis-related gene signature to predict prognosis in oesophageal squamous cell carcinoma," *J Cell Mol Med*, vol. 28, no. 21, p. e70171, 2024.

[378] G. L. De Salvo, P. Del Bianco, V. Minard-Colin, J. Chisholm, M. Jenney, G. Guillen, C. Devalck, R. Van Rijn, J. Shipley, D. Orbach, A. Kelsey, T. Rogers, F. Guerin, G. Scarzello, A. Ferrari, M. Cesen Mazic, J. H. M. Merks, G. Bisogno, and E. P. S. T. S. S. Group, "Reappraisal of prognostic factors used in the european pediatric soft tissue sarcoma study group rms 2005 study for localized rhabdomyosarcoma to optimize risk stratification and generate a prognostic nomogram," *Cancer*, vol. 130, no. 13, pp. 2351–2360, 2024.

[379] D. O. Walterhouse, D. A. Barkauskas, D. Hall, A. Ferrari, G. L. De Salvo, E. Koscielniak, M. C. G. Stevens, H. Martelli, G. Seitz, D. A. Rodeberg, M. Shnorhavorian, R. Dasgupta, J. C. Breneman, J. R. Anderson, C. Bergeron, G. Bisogno, W. H. Meyer, D. S. Hawkins, and V. Minard-Colin, "Demographic and treatment variables influencing outcome for localized paratesticular rhabdomyosarcoma: Results from a pooled analysis of north american and european cooperative groups," *J Clin Oncol*, vol. 36, no. 35, p. JCO2018789388, 2018.

[380] S. Park and G. Yi, "Development of gene expression-based random forest model for predicting neoadjuvant chemotherapy response in triple-negative breast cancer," *Cancers (Basel)*, vol. 14, no. 4, 2022.

[381] M. Kuhn, H. Wickham, and E. Hvitfeldt, *recipes: Preprocessing and Feature Engineering Steps for Modeling*, 2025. R package version 1.3.0.

[382] J. Li, S. Guo, R. Ma, J. He, X. Zhang, D. Rui, Y. Ding, Y. Li, L. Jian, J. Cheng, and H. Guo, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Med Res Methodol*, vol. 24, no. 1, p. 41, 2024.

[383] Y. Tang, S. Li, L. Zhu, L. Yao, J. Li, X. Sun, Y. Liu, Y. Zhang, and X. Fu, "Improve clinical feature-based bladder cancer survival prediction models through integration with gene expression profiles and machine learning techniques," *Heliyon*, vol. 10, no. 20, p. e38242, 2024.

[384] A. Mosquera Orgueira, M. S. Gonzalez Perez, J. Diaz Arias, B. Antelo Rodriguez, N. Alonso Vence, A. Bendana Lopez, A. Abuin Blanco, L. Bao Perez, A. Peleteiro Raindo, M. Cid Lopez, M. M. Perez Encinas, J. L. Bello Lopez, and M. V. Mateos Manteca, "Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data," *Leukemia*, vol. 35, no. 10, pp. 2924–2935, 2021.

[385] Y. Choi, J. Lee, K. Shin, J. W. Lee, J. W. Kim, S. Lee, Y. J. Choi, K. H. Park, and J. H. Kim, "Integrated clinical and genomic models using machine-learning methods to predict the efficacy of paclitaxel-based chemotherapy in patients with advanced gastric cancer," *BMC Cancer*, vol. 24, no. 1, p. 502, 2024.

[386] J. Leek, W. Johnson, H. Parker, E. Fertig, A. Jaffe, Y. Zhang, J. Storey, and L. Torres, *sva: Surrogate Variable Analysis*, 2025. R package version 3.56.0.

[387] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–27, 2007.

[388] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, p. 1–26, 2008.

[389] E. A. Clark and N. V. Giltiay, "Cd22: A regulator of innate and adaptive b cell responses and autoimmunity," *Front Immunol*, vol. 9, p. 2235, 2018.

[390] G. M. Liu, H. D. Zeng, C. Y. Zhang, and J. W. Xu, "Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma," *Cancer Cell Int*, vol. 19, p. 138, 2019.

[391] C. J. Walker, K. Mrózek, H. G. Ozer, D. Nicolet, J. Kohlschmidt, D. Papaioannou, L. K. Genutis, M. Bill, B. L. Powell, G. L. Uy, J. E. Kolitz, A. J. Carroll, R. M. Stone, R. Garzon, J. C. Byrd, A. K. Eisfeld, A. de la Chapelle, and C. D. Bloomfield, "Gene expression signature predicts relapse in adult patients with cytogenetically normal acute myeloid leukemia," *Blood Adv*, vol. 5, no. 5, pp. 1474–1482, 2021.

[392] Z. Zhao, J. Zobolas, M. Zucknick, and T. Aittokallio, "Tutorial on survival modeling with applications to omics data," *Bioinformatics*, vol. 40, no. 3, 2024.

[393] S. R. Carr, H. Wang, R. Hudlikar, X. Lu, M. R. Zhang, C. D. Hoang, F. Yan, and D. S. Schrump, "A unique gene signature predicting recurrence free survival in stage ia lung adenocarcinoma," *J Thorac Cardiovasc Surg*, vol. 165, no. 4, pp. 1554–1564, 2023.

[394] X. Zhou, C. Liu, H. Zeng, D. Wu, and L. Liu, "Identification of a thirteen-gene signature predicting overall survival for hepatocellular carcinoma," *Biosci Rep*, vol. 41, no. 4, 2021.

[395] L. Zhou, Y. Yu, R. Wen, K. Zheng, S. Jiang, X. Zhu, J. Sui, H. Gong, Z. Lou, L. Hao, G. Yu, and W. Zhang, "Development and validation of an 8-gene signature to improve survival prediction of colorectal cancer," *Front Oncol*, vol. 12, p. 863094, 2022.

[396] B. Aybey, S. Zhao, B. Brors, and E. Staub, "Immune cell type signature discovery and random forest classification for analysis of single cell gene expression datasets," *Front Immunol*, vol. 14, p. 1194745, 2023.

[397] R. Sundar, N. Barr Kumarakulasinghe, Y. Huak Chan, K. Yoshida, T. Yoshikawa, Y. Miyagi, Y. Rino, M. Masuda, J. Guan, J. Sakamoto, S. Tanaka, A. L. Tan, M. M. Hoppe, A. D. Jeyasekharan, C. C. Y. Ng, M. De Simone, H. I. Grabsch, J. Lee, T. Oshima, A. Tsuburaya, and P. Tan, "Machine-learning model derived gene signature predictive of paclitaxel survival benefit in gastric cancer: results from the randomised phase iii samit trial," *Gut*, vol. 71, no. 4, pp. 676–685, 2022.

[398] Q. Li, H. Yang, P. Wang, X. Liu, K. Lv, and M. Ye, "Xgboost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer," *J Transl Med*, vol. 20, no. 1, p. 177, 2022.

[399] S. Yu, M. Zhang, Z. Ye, Y. Wang, X. Wang, and Y. G. Chen, "Development of a 32-gene signature using machine learning for accurate prediction of inflammatory bowel disease," *Cell Regen*, vol. 12, no. 1, p. 8, 2023.

[400] M. B. Gilic and A. L. Tyner, "Targeting protein tyrosine kinase 6 in cancer," *Biochim Biophys Acta Rev Cancer*, vol. 1874, no. 2, p. 188432, 2020.

[401] S. Corbacioglu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," *Turk J Emerg Med*, vol. 23, no. 4, pp. 195–198, 2023.

[402] N. White, R. Parsons, G. Collins, and A. Barnett, "Evidence of questionable research practices in clinical prediction models," *BMC Med*, vol. 21, no. 1, p. 339, 2023.

[403] S. Awasthi, M. Verma, A. Mahesh, M. I. K Khan, G. Govindaraju, A. Rajavelu, P. L. Chavali, S. Chavali, and A. Dhayalan, "Ddx49 is an rna helicase that affects translation by regulating mrna export and the levels of pre-ribosomal rna," *Nucleic Acids Res*, vol. 46, no. 12, pp. 6304–6317, 2018.

[404] M. A. Arnold, J. R. Anderson, J. M. Gastier-Foster, F. G. Barr, S. X. Skapek, D. S. Hawkins, R. B. Raney, D. M. Parham, L. A. Teot, E. R. Rudzinski, and D. O. Walterhouse, "Histology, fusion status, and outcome in alveolar rhabdomyosarcoma with low-risk clinical features: A report from the children's oncology group," *Pediatr Blood Cancer*, vol. 63, no. 4, pp. 634–9, 2016.

[405] E. R. Rudzinski2017, J. R. Anderson, Y. Y. Chi, J. M. Gastier-Foster, C. Astbury, F. G. Barr, S. X. Skapek, D. S. Hawkins, B. J. Weigel, A. Pappo, W. H. Meyer, M. A. Arnold, L. A. Teot, and D. M. Parham, "Histology, fusion status, and outcome in metastatic rhabdomyosarcoma: A report from the children's oncology group," *Pediatr Blood Cancer*, vol. 64, no. 12, 2017.

[406] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, J. Yuan, and D. Cortes, *xgboost: Extreme Gradient Boosting*, 2025. R package version 3.0.2.1.

[407] D. Stekhoven and P. Buehlmann, "Missforest - non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[408] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[409] S. D'Agostino, L. Tombolan, M. Saggioro, C. Frasson, E. Rampazzo, S. Pellegrini, F. Favaretto, C. Biz, P. Ruggieri, P. Gamba, P. Bonvini, S. Aveic, R. Giovannoni, and M. Pozzobon, "Rhabdomyosarcoma cells produce their own extracellular matrix with minimal involvement of cancer-associated fibroblasts: A preliminary study," *Front Oncol*, vol. 10, p. 600980, 2020.

[410] F. Irie, Y. Tobisawa, A. Murao, H. Yamamoto, C. Ohyama, and Y. Yamaguchi, "The cell surface hyaluronidase tmem2 regulates cell adhesion and migration via degradation of hyaluronan at focal adhesion sites," *J Biol Chem*, vol. 296, p. 100481, 2021.

[411] Q. Jiang, X. Wang, Q. Yang, and H. Zhang, "Combined with," *DNA Cell Biol*, vol. 40, no. 11, pp. 1381–1395, 2021.

[412] L. Gao, S. Tong, J. Liu, J. Cai, Z. Ye, L. Zhou, P. Song, Z. Li, P. Lei, H. Wei, Q. Hua, D. Tian, and Q. Cai, "Tmem2 induces epithelial-mesenchymal transition and promotes resistance to temozolomide in gbm cells," *Heliyon*, vol. 9, no. 6, p. e16559, 2023.

[413] Y. Kudo, N. Sato, Y. Adachi, T. Amaike, A. Koga, S. Kohi, H. Noguchi, T. Nakayama, and K. Hirata, "Overexpression of transmembrane protein 2 (tmem2), a novel hyaluronidase, predicts poor prognosis in pancreatic ductal adenocarcinoma," *Pancreatology*, vol. 20, no. 7, pp. 1479–1485, 2020.

[414] P. Srivastava, V. K. Yadav, T. H. Chang, E. C. Su, B. Lawal, A. T. Wu, and H. S. Huang, "In-silico analysis of tmem2 as a pancreatic adenocarcinoma and cancer-associated fibroblast biomarker, and functional characterization of nsc777201, for targeted drug development," *Am J Cancer Res*, vol. 14, no. 6, pp. 3010–3035, 2024.

[415] H. Jiang, S. Wang, Y. Liu, C. Zheng, L. Chen, K. Zheng, Z. Xu, Y. Dai, H. Jin, Z. Cheng, C. Zou, L. Fu, K. Liu, and X. Ma, "Targeting efna1 suppresses tumor progression via the cmyc-modulated cell cycle and autophagy in esophageal squamous cell carcinoma," *Discov Oncol*, vol. 14, no. 1, p. 64, 2023.

[416] S. Chiappalupi, F. Riuzzi, S. Fulle, R. Donato, and G. Sorci, "Defective rage activity in embryonal rhabdomyosarcoma cells results in high pax7 levels that sustain migration and invasiveness," *Carcinogenesis*, vol. 35, no. 10, pp. 2382–92, 2014.

[417] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput Sci*, vol. 10, p. e2245, 2024.

[418] P. T. Kennedy, D. Zannoupa, M. H. Son, L. N. Dahal, and J. F. Woolley, "Neuroblastoma: an ongoing cold front for cancer immunotherapy," *J Immunother Cancer*, vol. 11, no. 11, 2023.

[419] W. M Kholosy, M. Derieppe, F. van den Ham, K. Ober, Y. Su, L. Custers, L. Schild, L. M J van Zogchel, L. M Wellens, H. R Ariese, C. L. Szanto, J. Wienke, M. P. Dierselhuis, D. van Vuurden, E. M. Dolman, and J. J. Molenaar, "Neuroblastoma and dipg organoid coculture system for personalized assessment of novel anticancer immunotherapies," *J Pers Med*, vol. 11, no. 9, 2021.

[420] L. Borriello, R. Nakata, M. A. Sheard, G. E. Fernandez, R. Sposto, J. Malvar, L. Blavier, H. Shimada, S. Asgharzadeh, R. C. Seeger, and Y. A. DeClerck, "Cancer-associated fibroblasts share characteristics and protumorigenic activity with mesenchymal stromal cells," *Cancer Res*, vol. 77, no. 18, pp. 5142–5157, 2017.

[421] O. Hashimoto, M. Yoshida, Y. Koma, T. Yanai, D. Hasegawa, Y. Kosaka, N. Nishimura, and H. Yokozaki, "Collaboration of cancer-associated fibroblasts and tumour-associated macrophages for neuroblastoma development," *J Pathol*, vol. 240, no. 2, pp. 211–23, 2016.

[422] E. Ghorani, C. Swanton, and S. A. Quezada, "Cancer cell-intrinsic mechanisms driving acquired immune tolerance," *Immunity*, vol. 56, no. 10, pp. 2270–2295, 2023.

[423] K. E. Masih, J. S. Wei, D. Milewski, and J. Khan, "Exploring and targeting the tumor immune microenvironment of neuroblastoma," *J Cell Immunol*, vol. 3, no. 5, pp. 305–316, 2021.

[424] L. Nagy, H. Y. Kao, D. Chakravarti, R. J. Lin, C. A. Hassig, D. E. Ayer, S. L. Schreiber, and R. M. Evans, "Nuclear receptor repression mediated by a complex containing smrt, msin3a, and histone deacetylase," *Cell*, vol. 89, no. 3, pp. 373–80, 1997.

[425] D. C. Coffey, M. C. Kutko, R. D. Glick, S. L. Swendeman, L. Butler, R. Rifkind, P. A. Marks, V. M. Richon, and M. P. LaQuaglia, "Histone deacetylase inhibitors and retinoic acids inhibit growth of human neuroblastoma in vitro," *Med Pediatr Oncol*, vol. 35, no. 6, pp. 577–81, 2000.

[426] D. C. Coffey, M. C. Kutko, R. D. Glick, L. M. Butler, G. Heller, R. A. Rifkind, P. A. Marks, V. M. Richon, and M. P. La Quaglia, "The histone deacetylase inhibitor, cbha, inhibits growth of human neuroblastoma xenografts in vivo, alone and synergistically with all-trans retinoic acid," *Cancer Res*, vol. 61, no. 9, pp. 3591–4, 2001.

[427] N. Pinto, S. G. DuBois, A. Marachelian, S. J. Diede, A. Taraseviciute, J. L. Glade Bender, D. Tsao-Wei, S. G. Groshen, J. M. Reid, D. A. Haas-Kogan, C. P. Reynolds, M. H. Kang, M. S. Irwin, M. E. Macy, J. G. Villablanca, K. K. Matthay, and J. R. Park, "Phase i study of vorinostat in combination with isotretinoin in

patients with refractory/recurrent neuroblastoma: A new approaches to neuroblastoma therapy (nant) trial," *Pediatr Blood Cancer*, vol. 65, no. 7, p. e27023, 2018.

[428] J. Ouyang, Y. Zhang, F. Xiong, S. Zhang, Z. Gong, Q. Yan, Y. He, F. Wei, W. Zhang, M. Zhou, B. Xiang, F. Wang, X. Li, Y. Li, G. Li, Z. Zeng, C. Guo, and W. Xiong, "The role of alternative splicing in human cancer progression," *Am J Cancer Res*, vol. 11, no. 10, pp. 4642–4667, 2021.

[429] R. F. Halperin, A. Hegde, J. D. Lang, E. A. Raupach, C. Legendre, W. S. Liang, P. M. LoRusso, A. Sekulic, J. A. Sosman, J. M. Trent, S. Rangasamy, P. Pirrotte, N. J. Schork, and C. R. Group, "Improved methods for rnaseq-based alternative splicing analysis," *Sci Rep*, vol. 11, no. 1, p. 10740, 2021.

[430] A. C. H. Wong, J. J. Wong, J. E. J. Rasko, and U. Schmitz, "Splicewiz: interactive analysis and visualization of alternative splicing in r," *Brief Bioinform*, vol. 25, no. 1, 2023.

[431] S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing, "rmats: robust and flexible detection of differential alternative splicing from replicate rna-seq data," *Proc Natl Acad Sci U S A*, vol. 111, no. 51, pp. E5593–601, 2014.

[432] R. Brunmeir, L. Ying, J. Yan, Y. T. Hee, B. Lin, H. Kaur, Q. Z. Leong, W. W. Teo, G. Choong, W. Y. Jen, L. P. Koh, L. K. Tan, E. Chan, M. Ooi, H. Yang, and W. J. Chng, "Ezh2 modulates mrna splicing and exerts part of its oncogenic function through repression of splicing factors in cml," *Leukemia*, vol. 39, no. 3, pp. 650–662, 2025.

[433] J. DeMartino, M. T. Meister, L. L. Visser, M. Brok, M. J. A. Groot Koerkamp, A. K. L. Wezenaar, L. S. Hiemcke-Jiwa, T. de Souza, J. H. M. Merks, A. C. Rios, F. C. P. Holstege, T. Margaritis, and J. Drost, "Single-cell transcriptomics reveals immune suppression and cell states predictive of patient outcomes in rhabdomyosarcoma," *Nat Commun*, vol. 14, no. 1, p. 3074, 2023.

[434] M. Zobeck, J. Khan, R. Venkatramani, M. F. Okcu, M. E. Scheurer, and P. J. Lupo, "Improving individualized rhabdomyosarcoma prognosis predictions using somatic molecular biomarkers," *JCO Precis Oncol*, vol. 9, p. e2400556, 2025.

[435] B. Geoerger, F. Bautista, N. André, P. Berlanga, S. A. Gatz, L. V. Marshall, J. Rubino, B. Archambaud, A. Marchais, A. Rubio-San-Simón, S. Ducassou, C. M. Zwaan, M. Casanova, K. Nysom, S. Pellegrino, N. Hoog-Labouret, A. Buzyn, P. Blanc, X. Paoletti, and G. Vassal, "Precision cancer medicine platform trials: Concepts and design of acsé-esmart," *Eur J Cancer*, vol. 208, p. 114201, 2024.

Supplementary figures

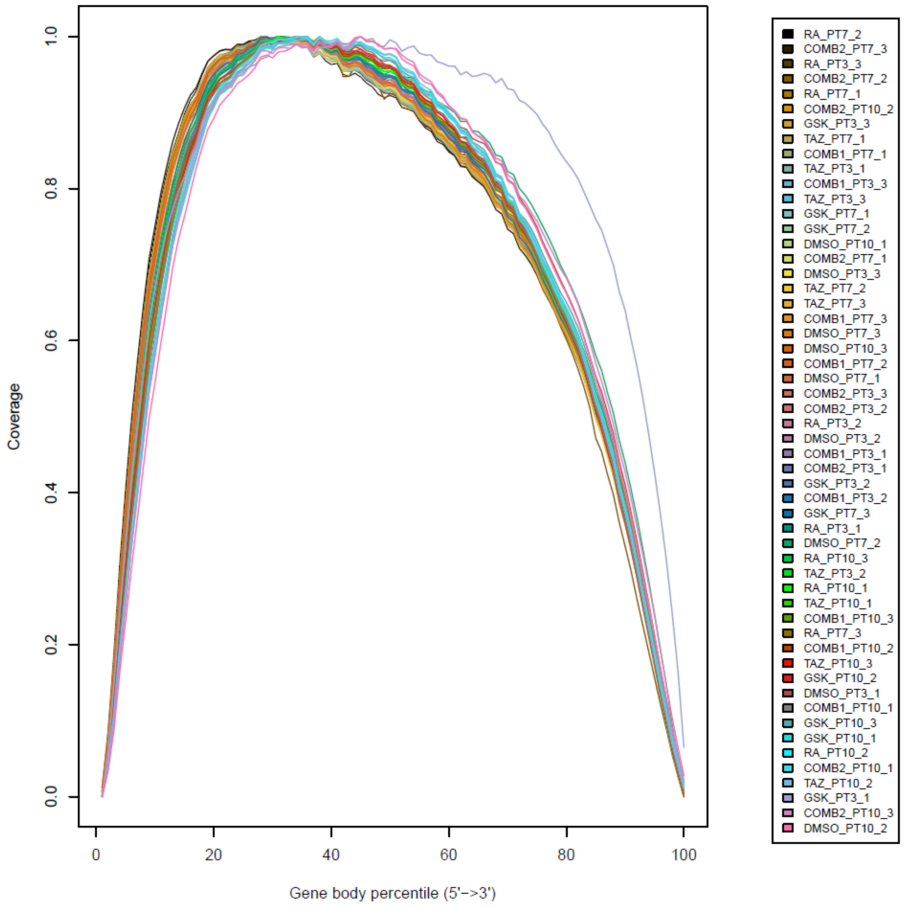


FIGURE 1: Gene body coverage plot of the NB RNA sequencing data showing RNA-seq read coverage for each sample over the gene body from 5′ to 3′. N=54 with 6 treatment conditions at 3 timepoints with 3 technical replicates each.

TABLE 1: Quality control results summary of the neuroblastoma Kelly spheroids sequenced by Novogene. RIN = RNA integrity score. COMB1 represents isotretinoin + TAZ and COMB2 represents isotretinoin + GSK. PT=Post-treatment day.

Sample Name	Concentration(ng/ul)	Volume(ul)	Total amount(ug)	RIN
DMSO_PT3_1	87.00	22	1.91400	10
RA_PT3_1	53.00	22	1.16600	10
TAZ_PT3_1	118.00	22	2.59600	10
GSK_PT3_1	46.00	23	1.05800	7.8
COMB1_PT3_1	23.00	22	0.50600	10
COMB2_PT3_1	83.00	21	1.74300	10
DMSO_PT3_2	163.00	22	3.58600	10
RA_PT3_2	122.00	19	2.31800	10
TAZ_PT3_2	137.00	21	2.87700	10
GSK_PT3_2	86.00	33	2.83800	10
COMB1_PT3_2	108.00	22	2.37600	10
COMB2_PT3_2	108.00	22	2.37600	10
DMSO_PT3_3	183.00	19	3.47700	10
RA_PT3_3	139.00	19	2.64100	10
TAZ_PT3_3	17.00	22	0.37400	10
GSK_PT3_3	88.00	22	1.93600	10
COMB1_PT3_3	28.00	28	0.78400	10
COMB2_PT3_3	113.00	21	2.37300	10
DMSO_PT7_1	314.00	21	6.59400	10
RA_PT7_1	139.00	23	3.19700	10
TAZ_PT7_1	129.00	22	2.83800	10
GSK_PT7_1	100.00	21	2.10000	10
COMB1_PT7_1	110.00	21	2.31000	10
COMB2_PT7_1	83.00	20	1.66000	10
DMSO_PT7_2	247.00	21	5.18700	10
RA_PT7_2	97.00	22	2.13400	10
TAZ_PT7_2	168.00	22	3.69600	10
GSK_PT7_2	178.00	23	4.09400	10
COMB1_PT7_2	174.00	23	4.00200	10
COMB2_PT7_2	146.00	23	3.35800	10
DMSO_PT7_3	227.00	18	4.08600	10
RA_PT7_3	18.00	20	0.36000	10
TAZ_PT7_3	135.00	20	2.70000	10
GSK_PT7_3	187.00	21	3.92700	10
COMB1_PT7_3	57.00	22	1.25400	10
COMB2_PT7_3	150.00	20	3.00000	10
DMSO_PT10_1	282.00	21	5.92200	10
RA_PT10_1	94.00	39	3.66600	10
TAZ_PT10_1	168.00	23	3.86400	10
GSK_PT10_1	166.00	23	3.81800	10
COMB1_PT10_1	48.00	24	1.15200	10
COMB2_PT10_1	233.00	22	5.12600	10
DMSO_PT10_2	248.00	22	5.45600	10
RA_PT10_2	79.00	22	1.73800	10
TAZ_PT10_2	183.00	23	4.20900	10
GSK_PT10_2	230.00	22	5.06000	10
COMB1_PT10_2	163.00	22	3.58600	10
COMB2_PT10_2	208.00	24	4.99200	10
DMSO_PT10_3	189.00	25	4.72500	10
RA_PT10_3	182.00	22	4.00400	10
TAZ_PT10_3	236.00	23	5.42800	10
GSK_PT10_3	265.00	21	5.56500	10
COMB1_PT10_3	145.00	34	4.93000	10
COMB2_PT10_3	82.00	27	2.21400	10

TABLE 2: Percentage of reads aligned for each sample in the NB sequencing data. COMB1 represents isotretinoin + TAZ and COMB2 represents isotretinoin + GSK. Post-treatment day is represented as PT. N=54

Sample	Percentage of reads aligned (%)
COMB1_PT10_1	85.85
COMB1_PT10_2	87.88
COMB1_PT10_3	84.63
COMB1_PT3_1	87.98
COMB1_PT3_2	77.69
COMB1_PT3_3	80.80
COMB1_PT7_1	84.77
COMB1_PT7_2	86.44
COMB1_PT7_3	77.82
COMB2_PT10_1	89.08
COMB2_PT10_2	85.63
COMB2_PT10_3	83.70
COMB2_PT3_1	88.53
COMB2_PT3_2	86.97
COMB2_PT3_3	83.51
COMB2_PT7_1	86.76
COMB2_PT7_2	16.58
COMB2_PT7_3	83.49
DMSO_PT10_1	89.72
DMSO_PT10_2	82.54
DMSO_PT10_3	88.29
DMSO_PT3_1	90.67
DMSO_PT3_2	89.75
DMSO_PT3_3	87.23
DMSO_PT7_1	86.92
DMSO_PT7_2	87.01
DMSO_PT7_3	86.40
GSK_PT10_1	90.72
GSK_PT10_2	90.24
GSK_PT10_3	88.07
GSK_PT3_1	90.47
GSK_PT3_2	89.43
GSK_PT3_3	85.16
GSK_PT7_1	88.22
GSK_PT7_2	88.96
GSK_PT7_3	86.39
RA_PT10_1	86.94
RA_PT10_2	85.71
RA_PT10_3	83.59
RA_PT3_1	89.33
RA_PT3_2	80.91
RA_PT3_3	84.07
RA_PT7_1	85.95
RA_PT7_2	85.70
RA_PT7_3	86.69
TAZ_PT10_1	89.11
TAZ_PT10_2	89.62
TAZ_PT10_3	87.42
TAZ_PT3_1	90.31
TAZ_PT3_2	90.38
TAZ_PT3_3	84.45
TAZ_PT7_1	89.77
TAZ_PT7_2	87.21
TAZ_PT7_3	84.48

TABLE 3: Percentage of unmapped reads for each sample in the NB sequencing data. COMB1 represents isotretinoin + TAZ and COMB2 represents isotretinoin + GSK. Post-treatment day is represented as PT. N=54. Highlighted are the 3 samples with adapter content warning (COMB1_PT3_3, COMB1_PT7_3, TAZ_PT7_3) and flagged outlier (COMB2_PT7_2).

Sample	Percentage of reads unmapped: mismatch reads	Percentage of reads unmapped: reads too short
COMB1_PT10_1	0.03%	7.78%
COMB1_PT10_2	0.03%	5.42%
COMB1_PT10_3	0.04%	10.13%
COMB1_PT3_1	0.01%	6.37%
COMB1_PT3_2	0.01%	17.50%
COMB1_PT3_3	0.02%	13.62%
COMB1_PT7_1	0.02%	8.22%
COMB1_PT7_2	0.02%	6.19%
COMB1_PT7_3	0.04%	16.56%
COMB2_PT10_1	0.02%	5.93%
COMB2_PT10_2	0.02%	7.74%
COMB2_PT10_3	0.02%	11.69%
COMB2_PT3_1	0.01%	6.79%
COMB2_PT3_2	0.01%	5.92%
COMB2_PT3_3	0.02%	10.86%
COMB2_PT7_1	0.02%	6.60%
COMB2_PT7_2	0.01%	81.93%
COMB2_PT7_3	0.03%	10.36%
DMSO_PT10_1	0.01%	4.18%
DMSO_PT10_2	0.01%	12.10%
DMSO_PT10_3	0.02%	5.96%
DMSO_PT3_1	0.01%	3.48%
DMSO_PT3_2	0.01%	3.77%
DMSO_PT3_3	0.02%	6.62%
DMSO_PT7_1	0.01%	3.62%
DMSO_PT7_2	0.01%	4.79%
DMSO_PT7_3	0.02%	7.75%
GSK_PT10_1	0.02%	3.33%
GSK_PT10_2	0.02%	3.30%
GSK_PT10_3	0.02%	6.85%
GSK_PT3_1	0.01%	2.08%
GSK_PT3_2	0.01%	3.55%
GSK_PT3_3	0.02%	8.65%
GSK_PT7_1	0.02%	6.25%
GSK_PT7_2	0.02%	3.91%
GSK_PT7_3	0.03%	7.66%
RA_PT10_1	0.01%	5.92%
RA_PT10_2	0.01%	7.90%
RA_PT10_3	0.02%	11.15%
RA_PT3_1	0.01%	5.61%
RA_PT3_2	0.01%	12.95%
RA_PT3_3	0.02%	9.63%
RA_PT7_1	0.01%	8.94%
RA_PT7_2	0.01%	7.70%
RA_PT7_3	0.02%	7.90%
TAZ_PT10_1	0.03%	4.06%
TAZ_PT10_2	0.03%	3.75%
TAZ_PT10_3	0.04%	6.83%
TAZ_PT3_1	0.02%	4.01%
TAZ_PT3_2	0.01%	3.67%
TAZ_PT3_3	0.02%	9.44%
TAZ_PT7_1	0.02%	4.88%
TAZ_PT7_2	0.03%	4.50%
TAZ_PT7_3	0.03%	9.63%

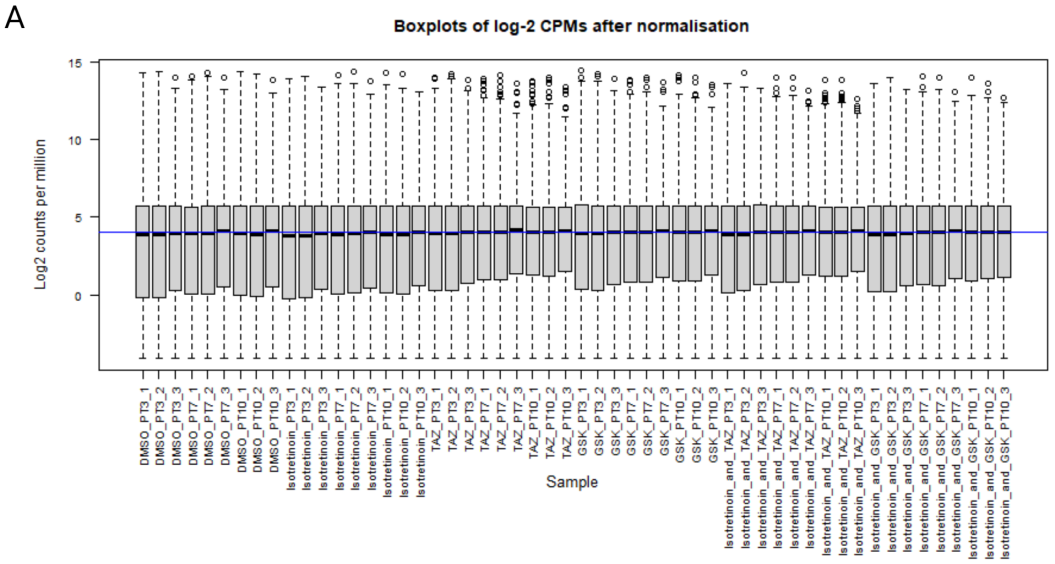


FIGURE 2: Boxplots of filtered log-2 CPMs before and after TMM normalisation for the NB RNA-seq data. Genes were filtered to remove genes that did not have a CPM of 0.4 in at least 3 samples. A) Boxplots of log-2 CPMs before TMM normalisation. B) Boxplots of log-2 CPMs after TMM normalisation. n=54 (6 treatment conditions at 3 timepoints with 3 technical replicates each).

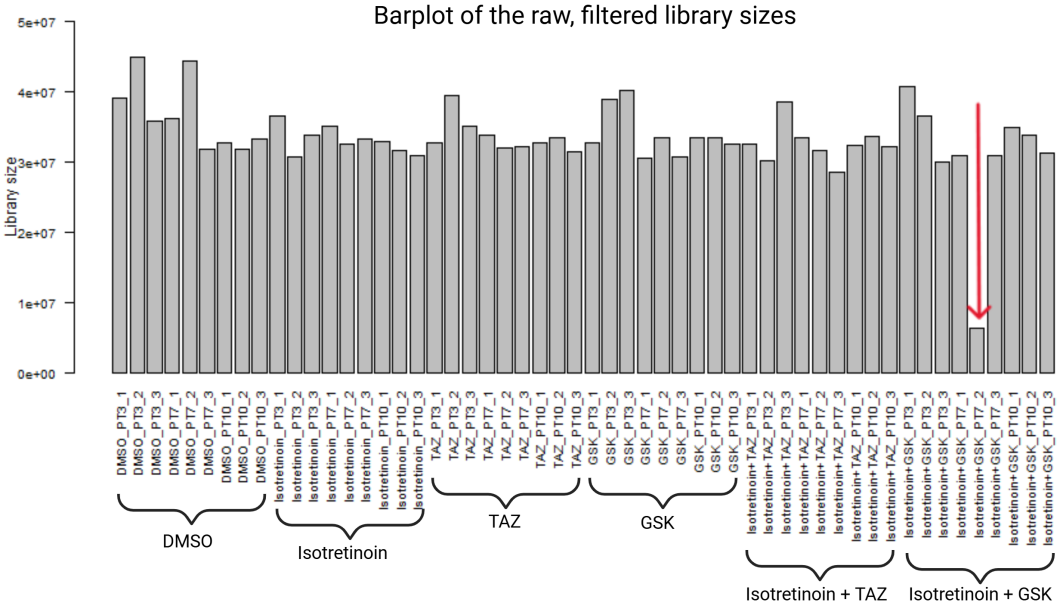


FIGURE 3: Barplot of the raw, filtered library sizes. Genes were filtered to remove genes that did not have a CPM of 0.4 in at least 3 samples. Outlier isotretinoin + GSK-PT7-2 is highlighted. N=54 with 6 treatment conditions at 3 timepoints with 3 technical replicates each.

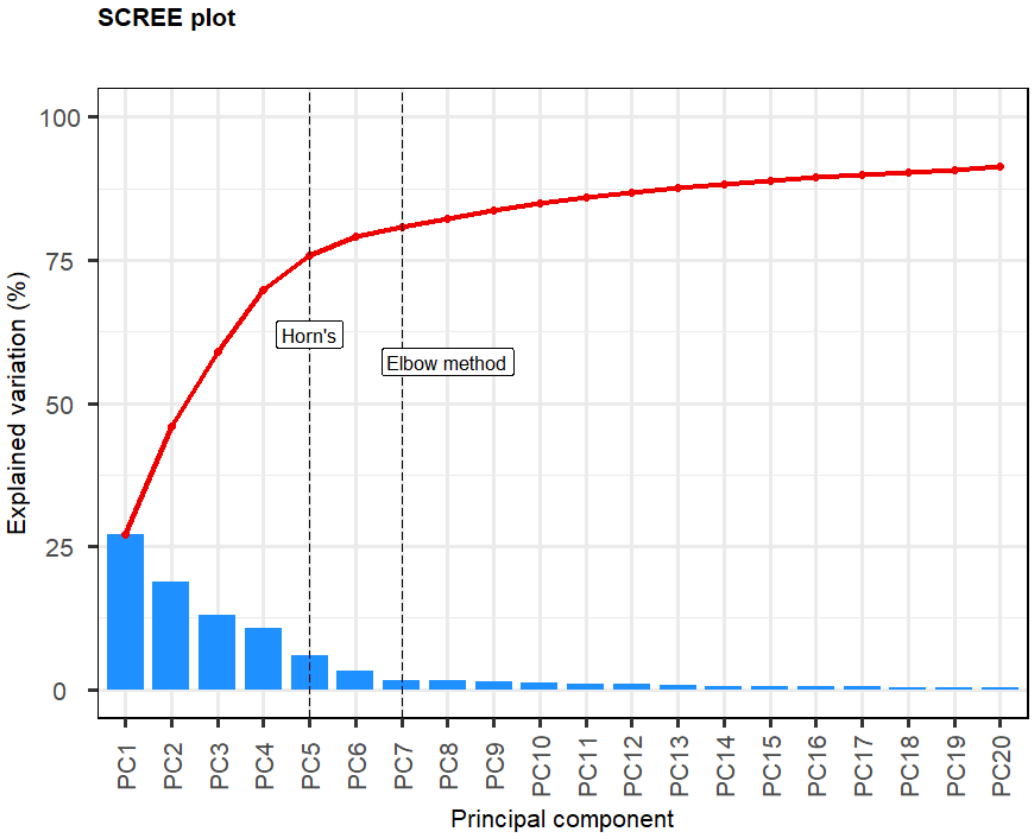


FIGURE 4: Scree plot to visualise the variance explained by each principal component from PCA of the NB RNA-seq data. The number of optimum principal components to retains are labelled determined by Horn's method and Elbow method.

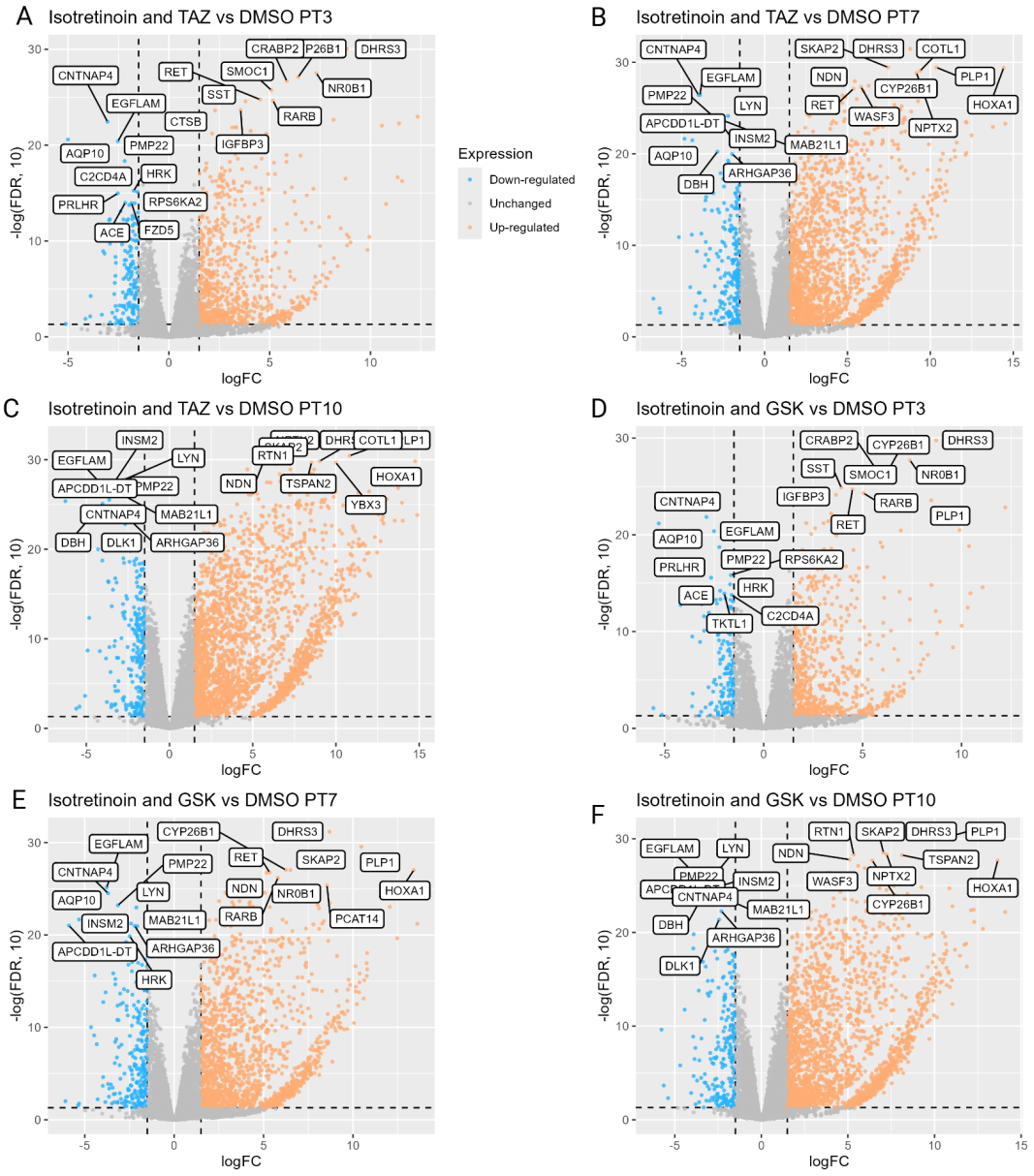


FIGURE 5: Volcano plots of DEGs for isotretinoin + TAZ and isotretinoin + GSK compared to control. DEGs were defined as having a LFC >1.5 or LFC <-1.5 and adjusted p value <0.05 (Benjamini-Hochberg). Isotretinoin + TAZ at A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. ; isotretinoin + GSK at D) post-treatment day 3 E) post-treatment day 7 and F) post-treatment day 10. Top ten upregulated and downregulated genes with the lowest p.adjust values are labelled.

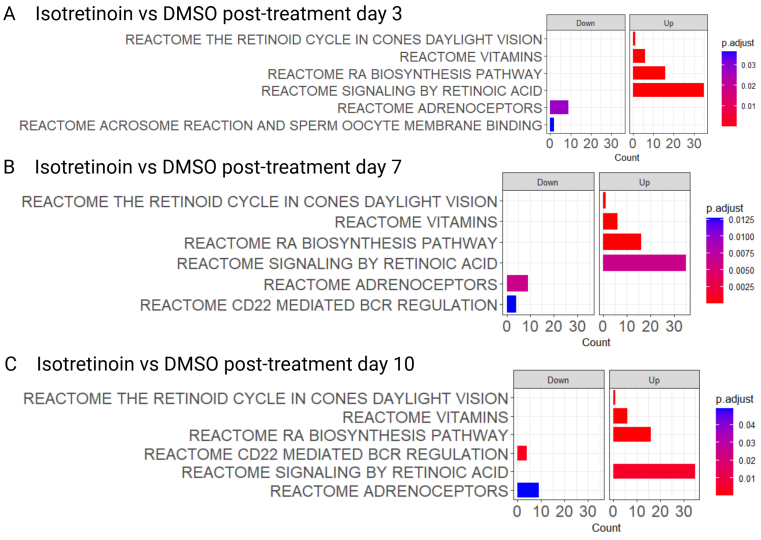


FIGURE 6: GSEA results from CAMERA showing enriched Reactome pathways in isotretinoin compared to control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

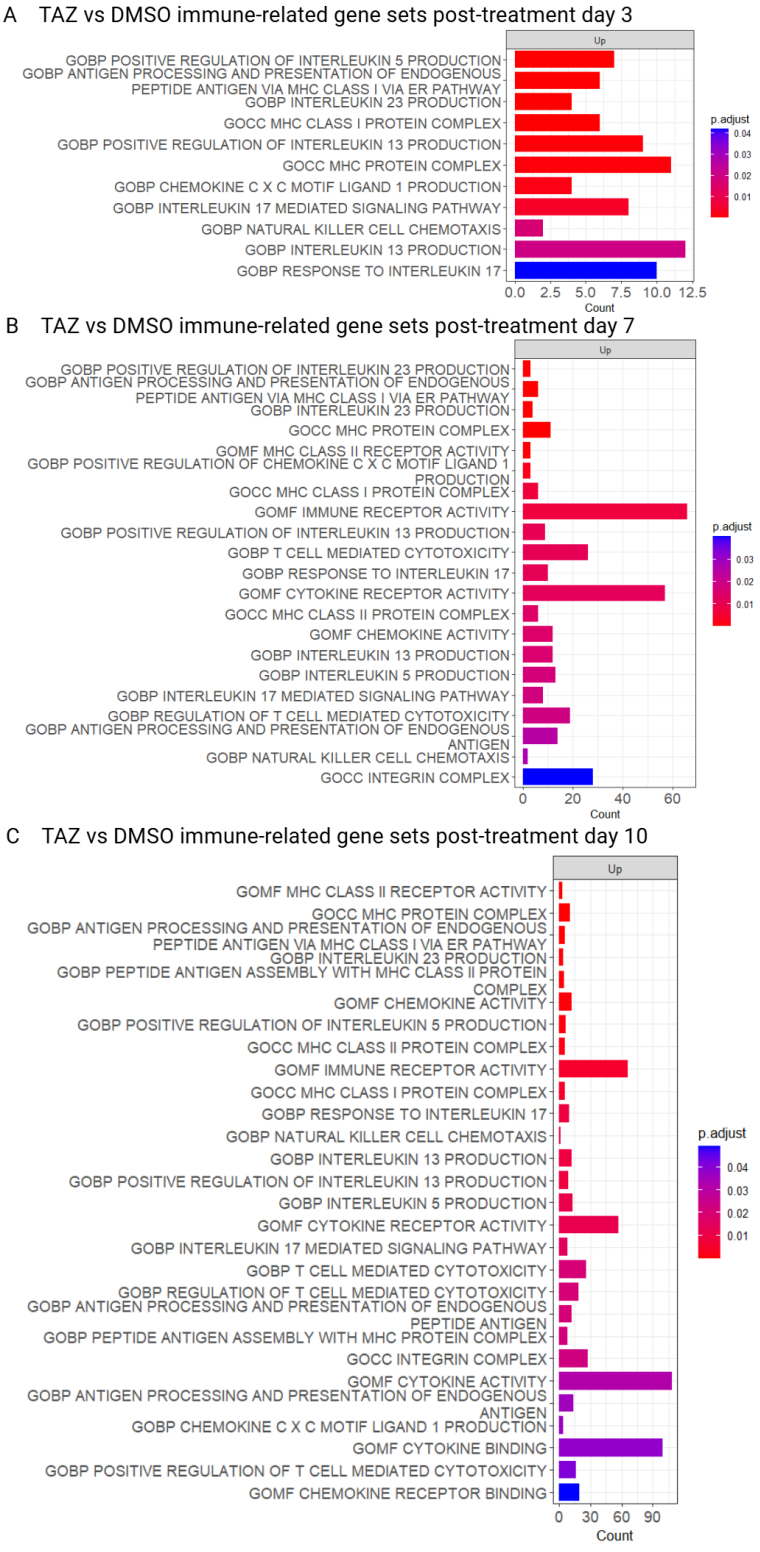


FIGURE 7: GSEA results from CAMERA showing enriched immune-related GO terms in TAZ compared to control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

TABLE 4: GSEA results of the 59 gene NB gene signature developed by [236]. Enrichment of the NB gene signature was testing used CAMERA. Significantly enriched samples are shown (pvalue <0.05).

Sample	Number of genes enriched	Direction	Pvalue
Isotretinoin_and_TAZvsDMSO_PT3	48	Up	4.91E-06
Isotretinoin_and_TAZvsDMSO_PT7	48	Up	0.000599
Isotretinoin_and_TAZvsDMSO_PT10	48	Up	0.003014
Isotretinoin_and_GSKvsDMSO_PT3	48	Up	1.76E-06
Isotretinoin_and_GSKvsDMSO_PT7	48	Up	0.000183
Isotretinoin_and_GSKvsDMSO_PT10	48	Up	0.001146
IsotretinoinvsDMSO_PT3	48	Up	1.48E-06
IsotretinoinvsDMSO_PT7	48	Up	6.56E-06
IsotretinoinvsDMSO_PT10	48	Up	3.04E-07

TABLE 5: Table showing the top 10 most significant DEGs when comparing replicate 1/2 to replicate 3 for all samples in the NB RNA-seq data.

Genes	LogFC replicate 1/2	LogFC replicate 3	logCPM	F	PValue	FDR
SRP68	-12.73585289	0.014345674	7.220168	6713.9	3.48E-55	2.62E-51
UBE2L3	-12.57100589	-0.16903386	7.519056	6689.99	3.76E-55	2.62E-51
FAM104A	-14.58181849	0.231909722	5.175271	6526.819	6.45E-55	2.62E-51
CAPZB	-12.94505281	-0.20546026	7.043792	6460.089	8.07E-55	2.62E-51
EFTUD2	-11.48466026	-0.27207266	8.356976	6456.407	8.17E-55	2.62E-51
FBXO7	-12.51946805	-0.24565835	7.310227	6361.419	1.13E-54	2.62E-51
IGFBPL1	-10.19876353	-0.37148744	9.574792	6345.432	1.19E-54	2.62E-51
CAPNS1	-12.30234623	-0.49241131	7.32573	6286.151	1.46E-54	2.62E-51
EIF2B5	-13.32347233	0.216355269	6.813436	6247.614	1.67E-54	2.62E-51
SPOP	-13.78734037	-0.055301	6.25468	6245.137	1.69E-54	2.62E-51

TABLE 6: GSEA analysis results from 59 random gene signature in the NB RNA-seq data. Genes were randomly selected after filtering to remove lowly expressed genes. GSEA was conducted using CAMERA.

Sample	Number of genes	Direction	Pvalue
Isotretinoin_and_TAZvsDMSO_PT3	59	Up	0.263169
Isotretinoin_and_TAZvsDMSO_PT7	59	Up	0.167883
Isotretinoin_and_TAZvsDMSO_PT10	59	Up	0.124745
Isotretinoin_and_GSKvsDMSO_PT3	59	Up	0.257275
Isotretinoin_and_GSKvsDMSO_PT7	59	Up	0.222015
Isotretinoin_and_GSKvsDMSO_PT10	59	Up	0.100277
GSKvsDMSO_PT3	59	Up	0.228264
GSKvsDMSO_PT7	59	Up	0.17696
GSKvsDMSO_PT10	59	Up	0.105216
IsotretinoinvsDMSO_PT3	59	Up	0.724429
IsotretinoinvsDMSO_PT7	59	Up	0.971232
IsotretinoinvsDMSO_PT10	59	Up	0.40851
TAZvsDMSO_PT3	59	Up	0.083094
TAZvsDMSO_PT7	59	Up	0.193607
TAZvsDMSO_PT10	59	Up	0.128452

TABLE 7: Percentage of reads aligned for each sample in the intrinsic resistance sequencing data. C7 = clone 7, C13 = clone 13, HD = high dose and P = parental. Numbers represent replicate number. N=24

Sample	Percentage of uniquely mapped reads
RH30_P_1	92.79%
RH30_P_2	92.97%
RH30_P_3	93.23%
RH30_P_4	93.05%
RH30_P_5	93.13%
RH30_P_6	93.26%
RH_C13_1	93.56%
RH_C13_2	93.29%
RH_C13_3	93.21%
RH_C13_4	92.65%
RH_C13_5	93.12%
RH_C13_6	93.01%
RH_C7_1	93.61%
RH_C7_2	93.30%
RH_C7_3	93.58%
RH_C7_4	93.46%
RH_C7_5	93.58%
RH_C7_6	93.41%
RH_HD_1	93.23%
RH_HD_2	93.11%
RH_HD_3	92.83%
RH_HD_4	93.10%
RH_HD_5	93.49%
RH_HD_6	93.38%

TABLE 8: WGCNA and PPI information on the intrinsic resistant modules 'darkgrey' and 'darkgreen'. Resistant modules were significantly positively correlated with the binary trait resistance. WGCNA module hub genes were defined as having a module membership and gene significance value >0.6. Hub proteins were defined as having a centrality score ≥8.

Resistant modules	Number of genes in the module	Number of hub genes	Number of proteins in PPI network	Number of PPI hub proteins	Number of overlapping PPI hub proteins and WGCNA hub genes
Darkgrey	1159	854	207	27	11
Darkgreen	1832	556	102	14	7

TABLE 9: Hub genes from the PPI's created from genes in the resistant modules 'dark-green' and 'darkgrey'. PPI hub genes were selected as a centrality score ≥8.

Gene	Centrality score	Module
UBC	36	Darkgreen
BRCA1	24	Darkgreen
UBE3A	12	Darkgreen
SQSTM1	12	Darkgreen
CCNB1	10	Darkgreen
HSPA4	10	Darkgreen
FZR1	10	Darkgreen
USP10	10	Darkgreen
NR3C1	8	Darkgreen
UVRAG	8	Darkgreen
RAC1	8	Darkgreen
CDC16	8	Darkgreen
BRCC3	8	Darkgreen
CDC27	8	Darkgreen
RBBP5	16	Darkgrey
YWHAH	14	Darkgrey
CBL	14	Darkgrey
DDX6	14	Darkgrey
POLR2A	14	Darkgrey
CNOT1	12	Darkgrey
XAB2	12	Darkgrey
MAPK1	10	Darkgrey
PRPF3	10	Darkgrey
UPF1	10	Darkgrey
RPS5	10	Darkgrey
SETD1A	8	Darkgrey
HDAC4	8	Darkgrey
CASC3	8	Darkgrey
ERBB2	8	Darkgrey
SRRM2	8	Darkgrey
PRPF31	8	Darkgrey
CHD4	8	Darkgrey
EIF3B	8	Darkgrey
XIAP	8	Darkgrey
NCBP1	8	Darkgrey
EIF4ENIF1	8	Darkgrey
RBM4	8	Darkgrey
CXXC1	8	Darkgrey
NUP153	8	Darkgrey
IKBKG	8	Darkgrey
CPSF1	8	Darkgrey

TABLE 10: Percentage of reads aligned for each sample in the acquired resistance sequencing data. Vcr = vincristine-resistant, ifo = ifosfamide-resistant, ctrl = control. Numbers represent replicate number. N=36

Sample	Uniquely mapped reads (%)
RH4ctrl_1	92.77
RH4ctrl_2	93.17
RH4ctrl_3	93.03
RH4ctrl_4	93.2
RH4ctrl_5	92.72
RH4ctrl_6	91.01
RH4ifo_1	93.15
RH4ifo_2	93.1
RH4ifo_3	86.92
RH4ifo_4	91.5
RH4ifo_5	91.74
RH4ifo_6	89.81
RH4vcr_1	91.89
RH4vcr_2	89.38
RH4vcr_3	92.11
RH4vcr_4	89.94
RH4vcr_5	91.07
RH4vcr_6	87.13
RMSYMctrl_1	92.86
RMSYMctrl_2	91.78
RMSYMctrl_3	91.33
RMSYMctrl_4	91.6
RMSYMctrl_5	92.33
RMSYMctrl_6	92.62
RMSYMifo_1	91.21
RMSYMifo_2	91.3
RMSYMifo_3	92.81
RMSYMifo_4	90.6
RMSYMifo_5	92.97
RMSYMifo_6	86.55
RMSYMvcr_1	91.16
RMSYMvcr_2	93.32
RMSYMvcr_3	92.95
RMSYMvcr_4	92.66
RMSYMvcr_5	92.6
RMSYMvcr_6	91.21

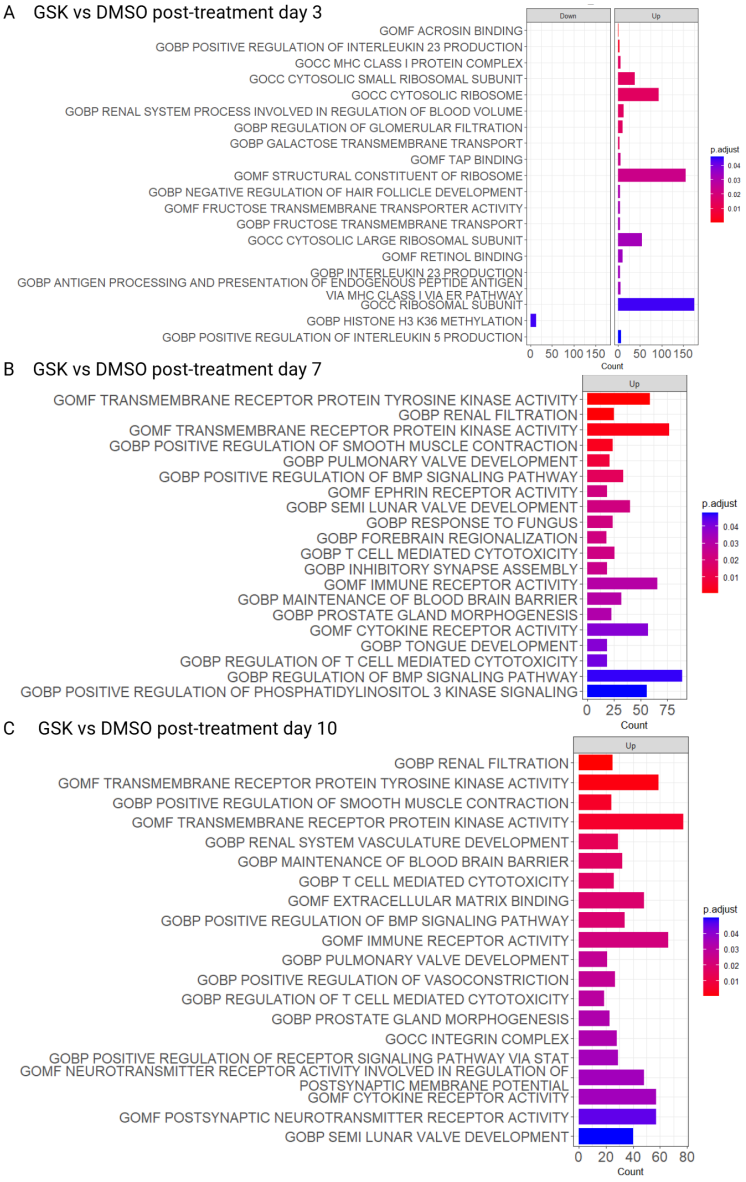


FIGURE 8: GSEA results from CAMERA showing enriched GO terms for GSK vs control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

TABLE 11: Number of genes and proteins in the two FP-acquired resistant modules. Resistant modules were significantly positively correlated with the binary trait resistance. WGCNA module hub genes were defined as having a module membership and gene significance value >0.6. Hub proteins were defined as having a centrality score ≥8.

Resistant module	WGCNA module size	Number of WGCNA hub genes	Number of interacting proteins in the PPI network	Number of hub proteins	Number of overlapping PPI hub proteins and WGCNA hub genes
Lightgreen	1249	826	43	13	11
Steelblue	111	31	0	0	0
Orange	1948	1142	175	21	14

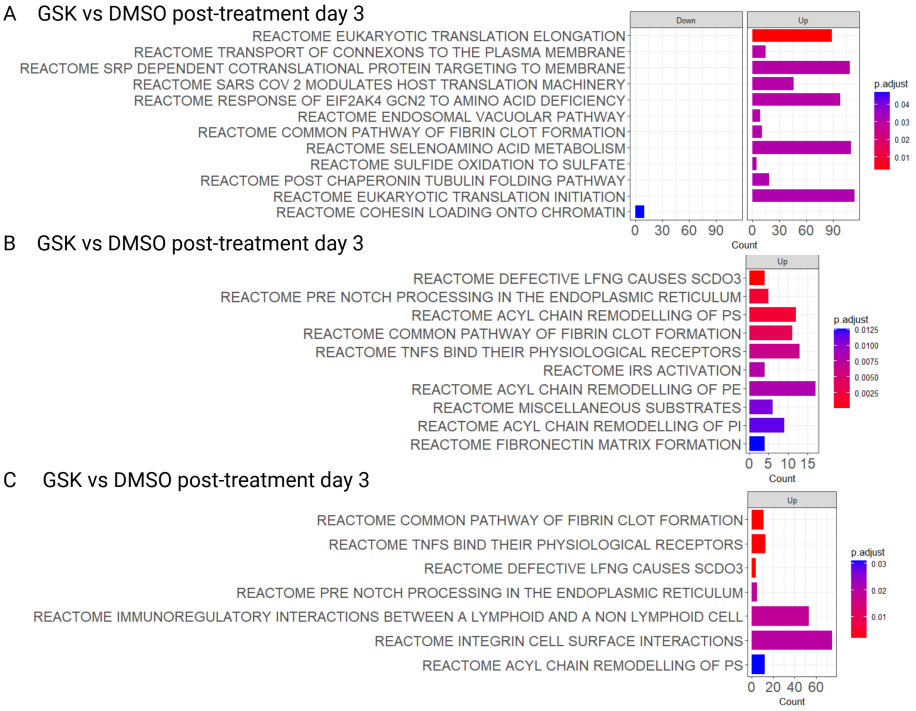


FIGURE 9: GSEA results from CAMERA showing enriched Reactome gene sets for GSK vs control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

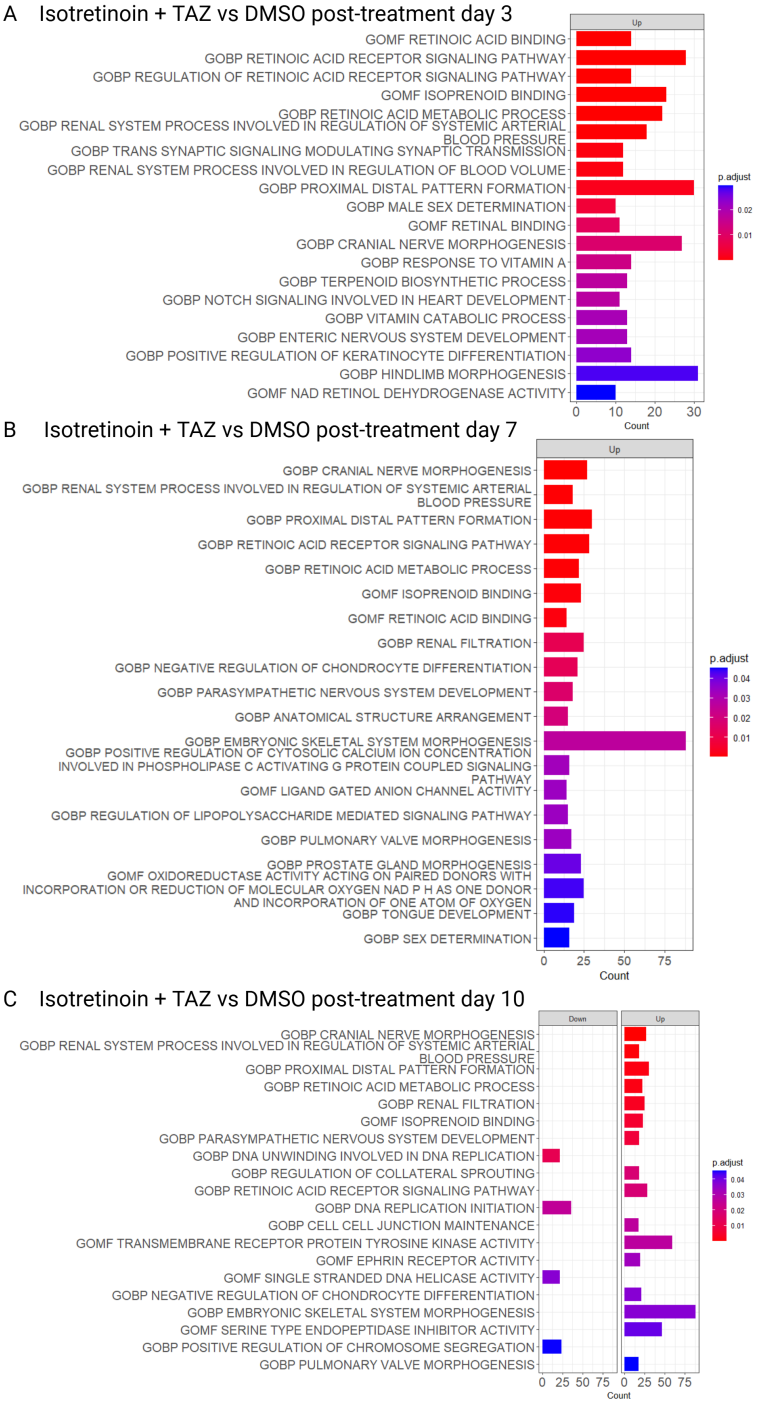


FIGURE 10: GSEA results from CAMERA showing enriched GO terms for isotretinoin + TAZ vs control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

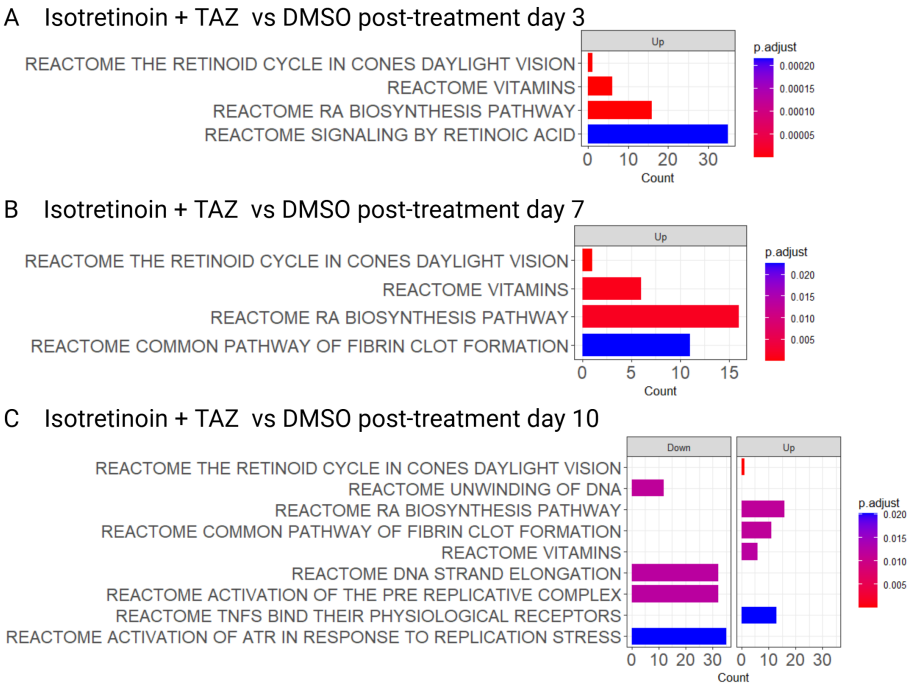


FIGURE 11: GSEA results from CAMERA showing enriched Reactome gene sets for isotretinion + TAZ vs control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

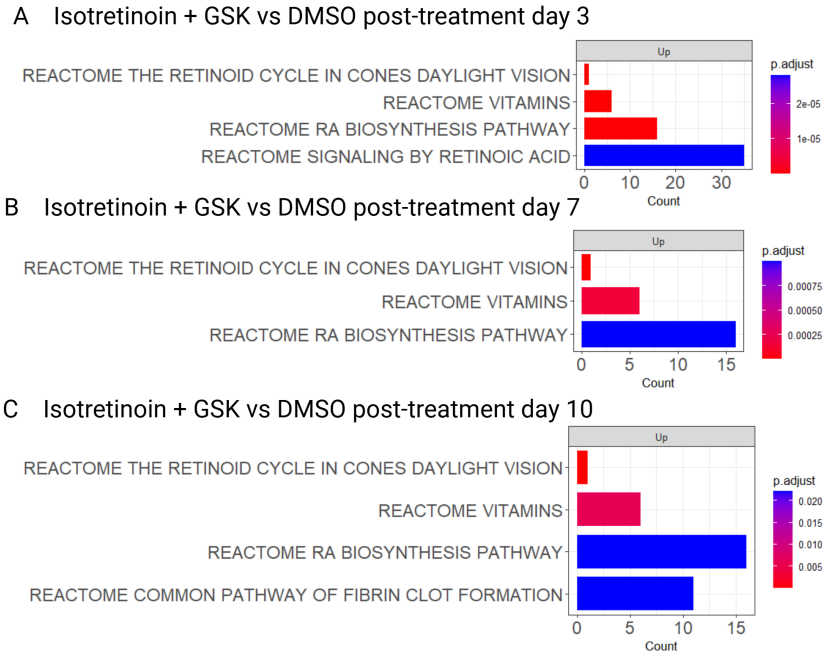


FIGURE 12: GSEA results from CAMERA showing enriched Reactome gene sets for Isotretinoin + GSK vs control, for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

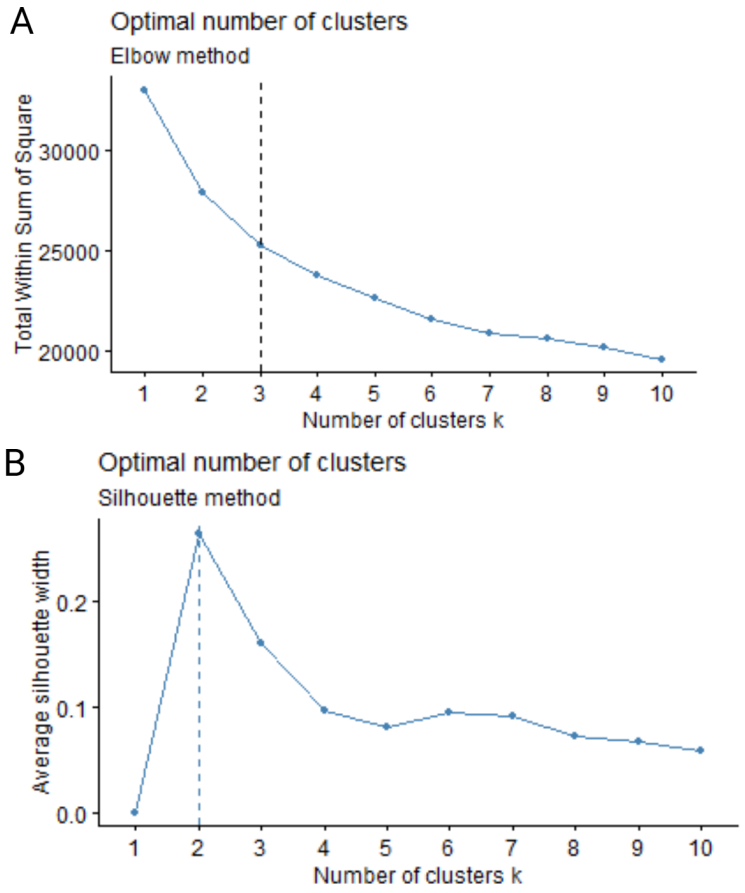


FIGURE 13: Determining the optimum number of clusters for k-means clustering using the Elbow method and Silhouette method in the NB RNA-seq data. A)Elbow plot showing the number of clusters and within sum of square for each cluster. B) Silhouette plot showing the average silhouette width for each cluster. k.max=10

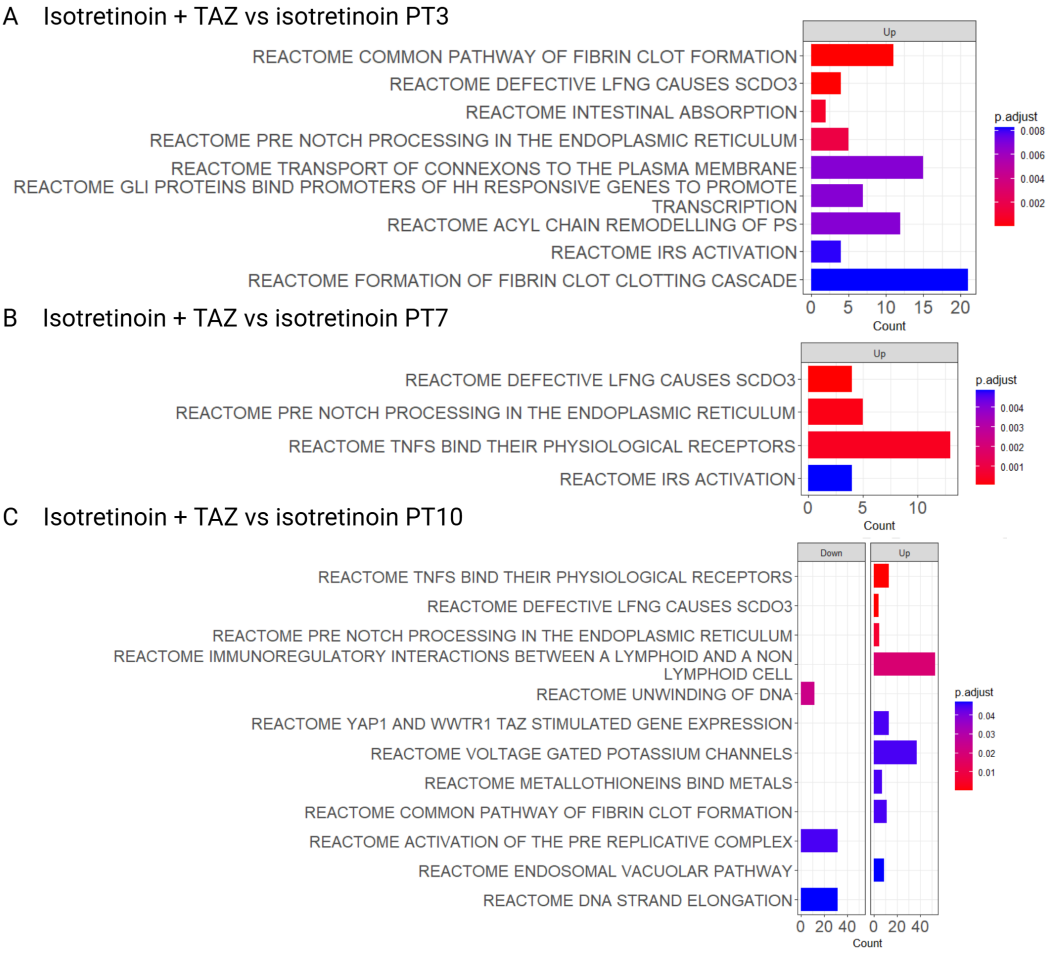


FIGURE 14: GSEA results from CAMERA showing enriched Reactome pathways for isotretinoin + TAZ vs isotretinoin for timepoints A) post-treatment day 3 B) post-treatment day 7 and C) post-treatment day 10. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05.

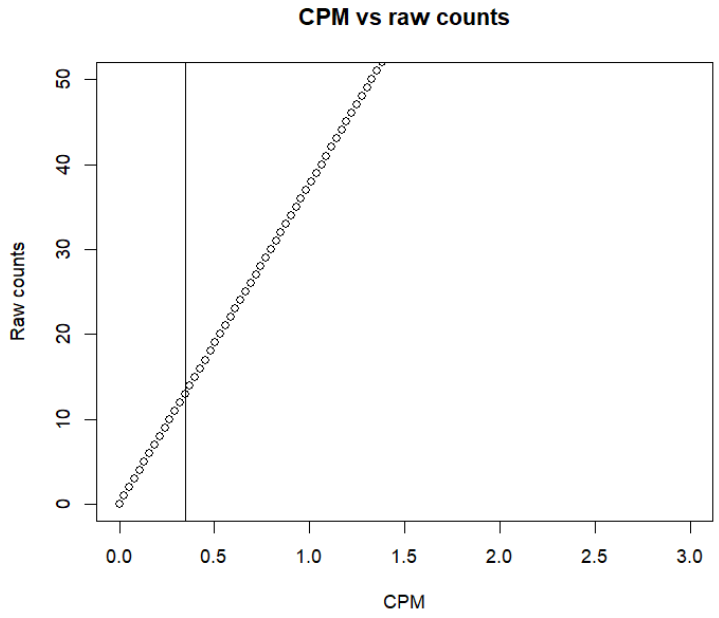


FIGURE 15: Plot showing count per million (CPM) vs raw count of sample RH-C13-1 (clone 13 replicate 1). Vertical line is added at a CPM of 0.35 which roughly corresponds to 10-15 counts.

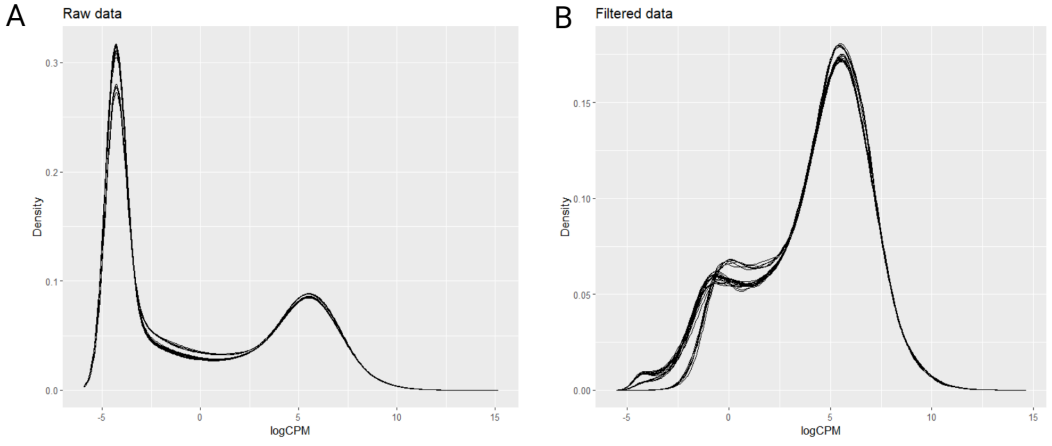


FIGURE 16: CPM density plots for 24 samples (each line represented a separate sample) before and after filtering of the intrinsic resistance data. A) Density plot of the raw data. N= 28277. B) Density plot after filtering to remove genes that don't have a CPM of 0.35 in at least 6 samples. N= 14,276.

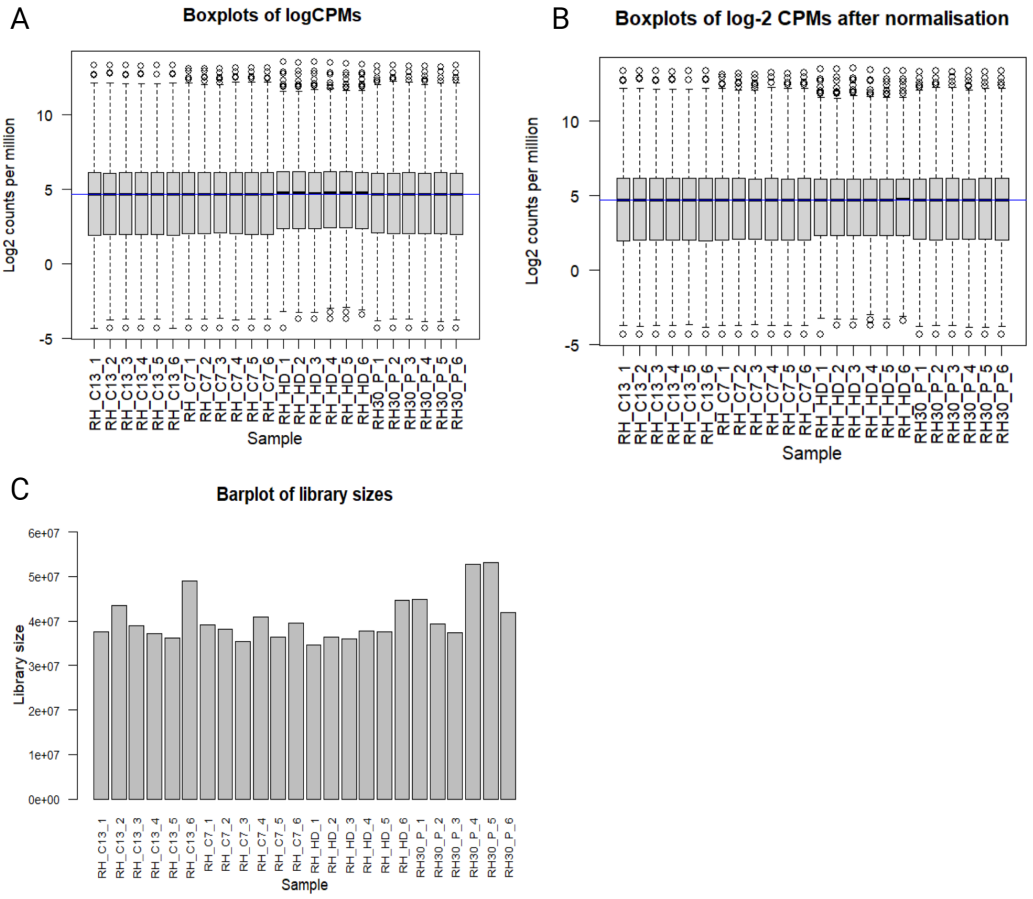


FIGURE 17: Quality control of the filtered data before and after TMM normalisation for the intrinsic resistant cell models.. A) Boxplot of filtered log-2 CPMs before TMM normalisation. B) Boxplots of log-2 CPMs after TMM normalisation C) Barplot of the raw, filtered library sizes. n= 24 (4 treatment conditions with 6 replicates each).

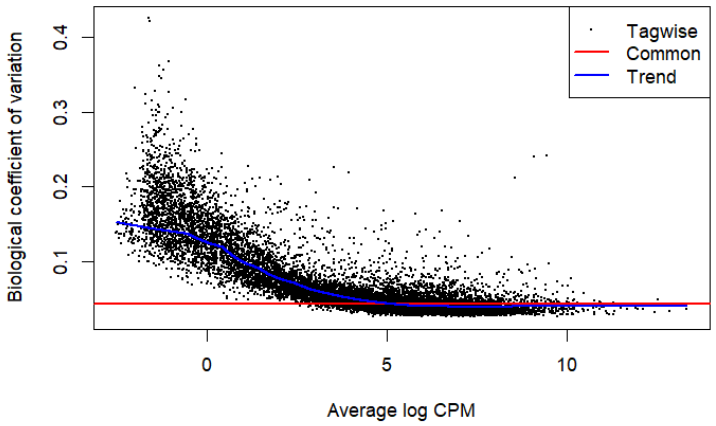


FIGURE 18: Biological coefficient of variation plot against gene abundance with estimates of the common, trended and tagwise dispersions for the intrinsic resistance data.

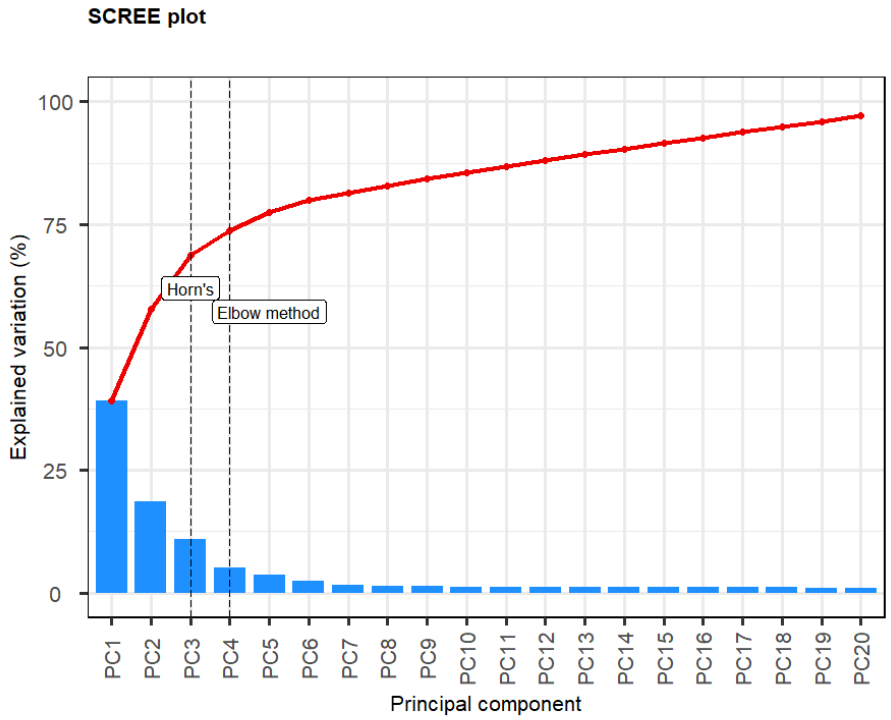


FIGURE 19: Scree plot to visualise the variance explained by each principal component from PCA of the intrinsic resistance RNA sequencing data. The number of optimum principal components to retains are labelled determined by Horn's method and Elbow method.

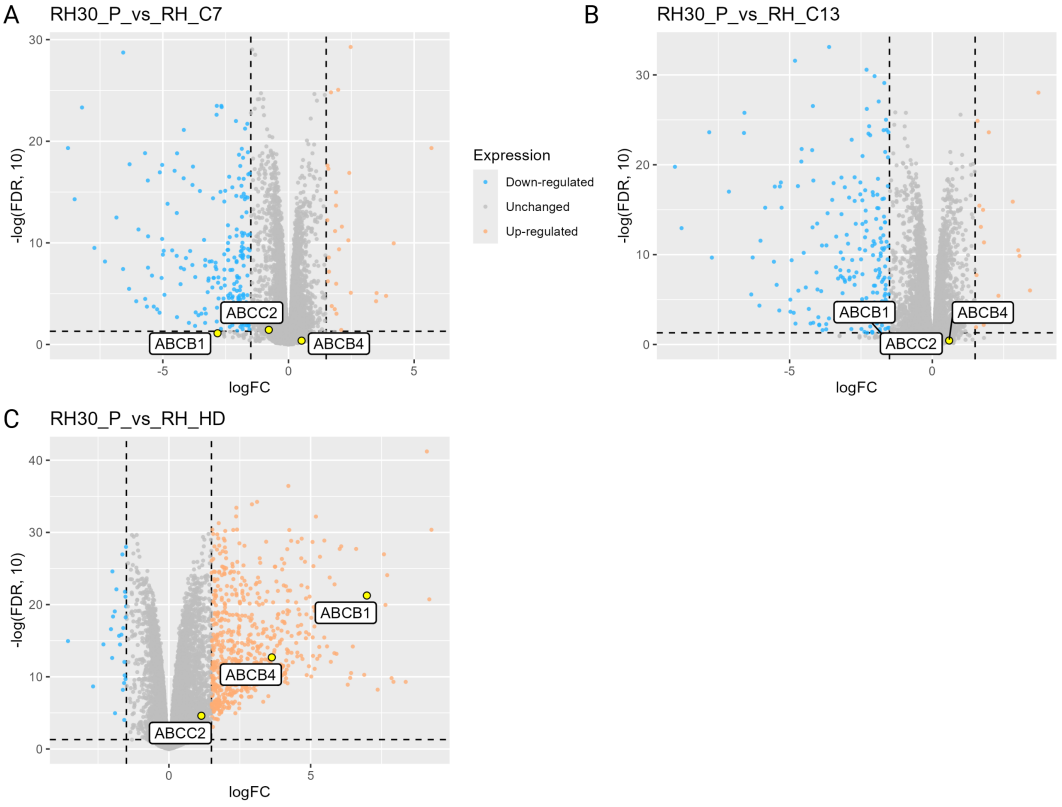


FIGURE 20: Volcano plots with MDR genes highlighted for A) C7 vs P B) C13 vs P C) HD vs P. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

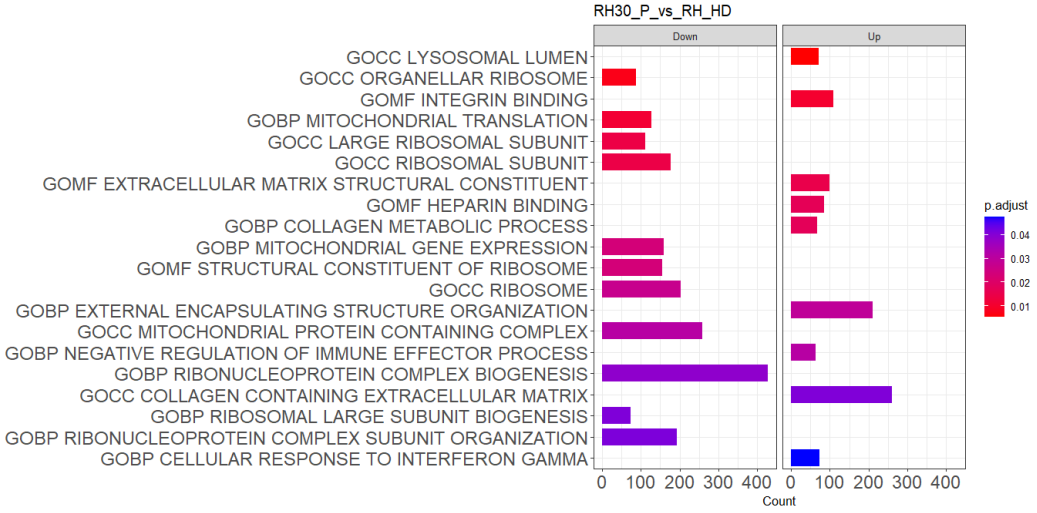


FIGURE 21: GSEA results from CAMERA showing enriched GO terms in HD compared to P. Enriched pathways are defined as adjusted p value (Benjamini-Hochberg) <0.05. If 20 or more gene sets were enriched the 20 largest gene sets were plotted.

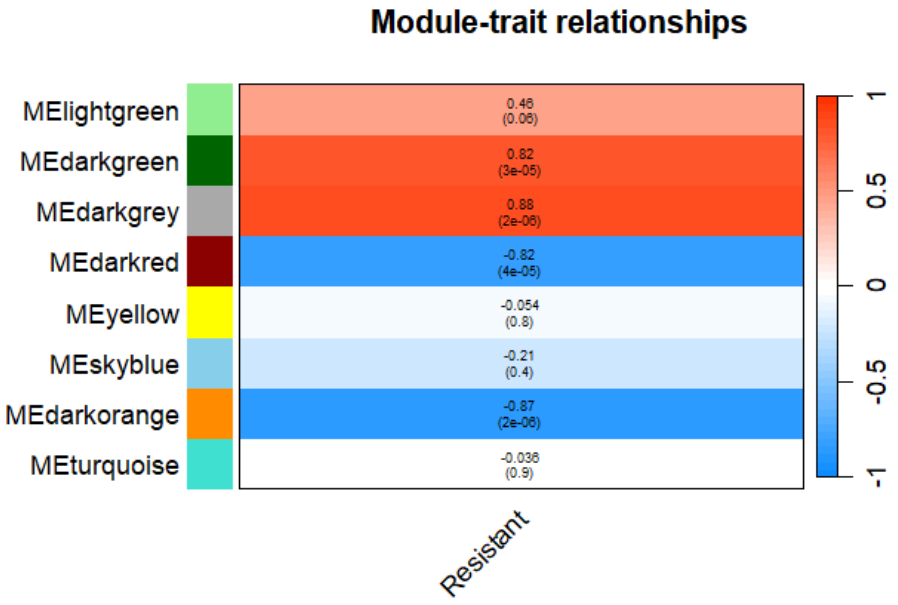


FIGURE 22: Heatmap showing module-trait correlation for the trait resistance. Modules are from the weighted gene co-expression network generated from C7, C13 and P samples from the intrinsic resistance data. For each module the Pearson correlation and Student asymptotic p-value for correlation are shown. Colour represents Pearson correlation.

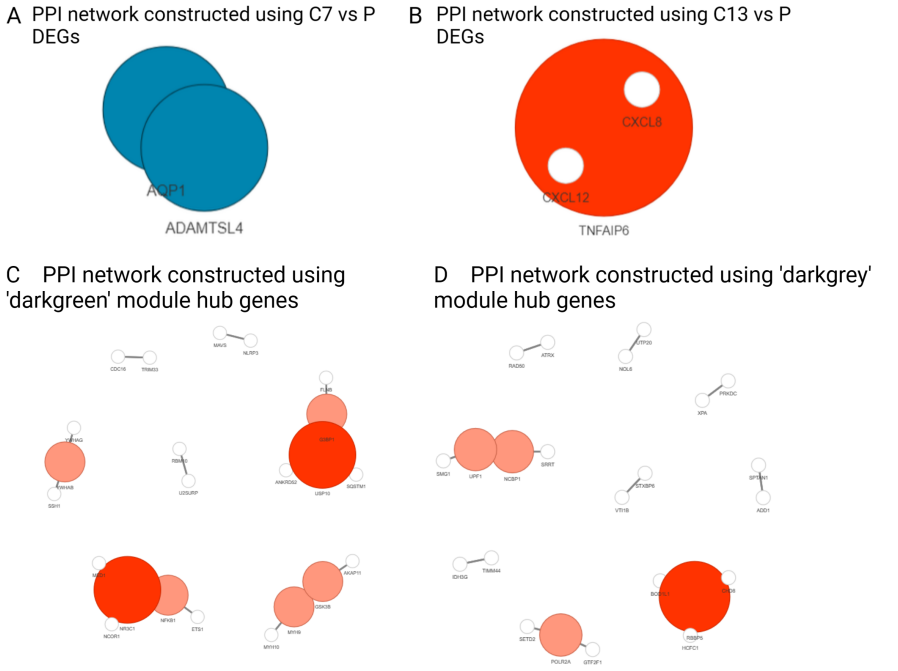


FIGURE 23: Visualisation of the PPI networks constructed using DEGs from A) C7 vs P and B) C13 vs P C) WGCNA hub genes from the 'darkgreen' module and D) WGCNA hub genes from the 'darkgrey' module. Differentially expressed genes were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg). Nodes are coloured by centrality score with red indicating a higher centrality score and white indicating a lower centrality score. Equal centrality scores are shown in blue. Hub genes were defined has having a module membership and gene significance score >0.6.

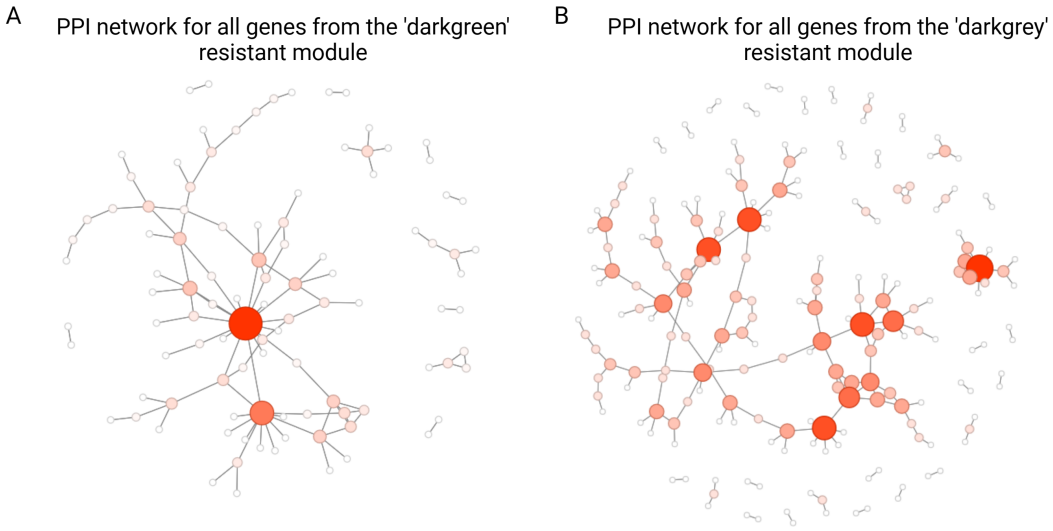


FIGURE 24: Visualisation of the PPI networks constructed from modules significantly correlated with intrinsic resistance. A) all genes from the resistant 'darkgreen' module and B) all genes from the resistant 'darkgrey' module. Nodes are coloured by centrality score with red indicating a higher centrality score and white indicating a lower centrality score.

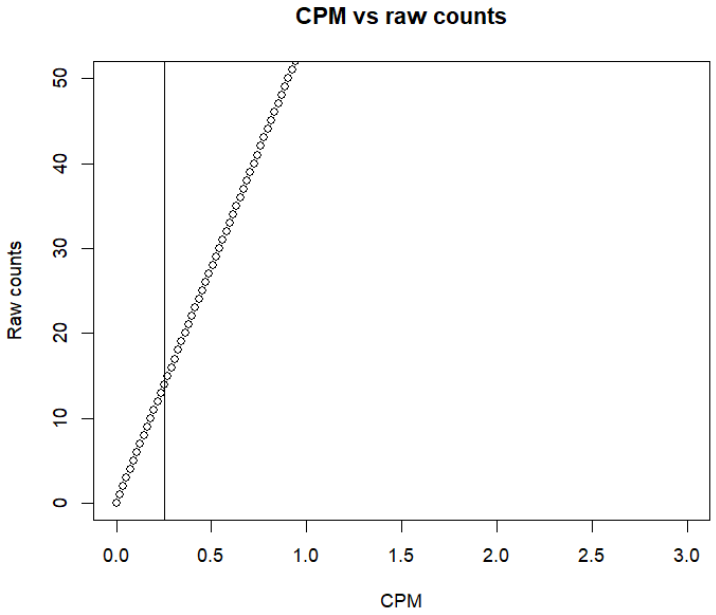


FIGURE 25: Plot showing count per million (CPM) vs raw count of sample RH4ctrl_1. Vertical line is added at a CPM of 0.25 which roughly corresponds to 10-15 counts.

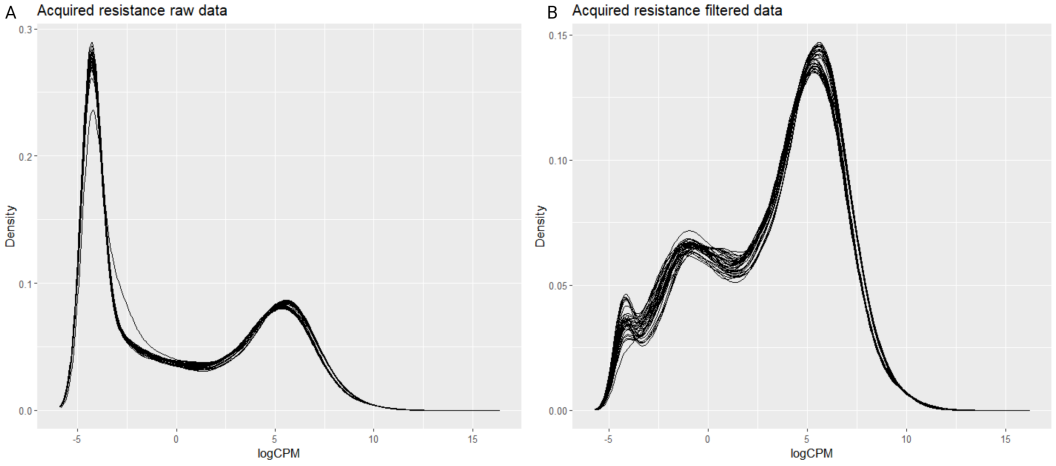


FIGURE 26: CPM density plots before and after filtering of the acquired resistance data. A) Density plot of the raw data. N= 28277. B) Density plot after filtering to remove genes that don't have a CPM of 0.25 in at least 6 samples. N= 16,865.

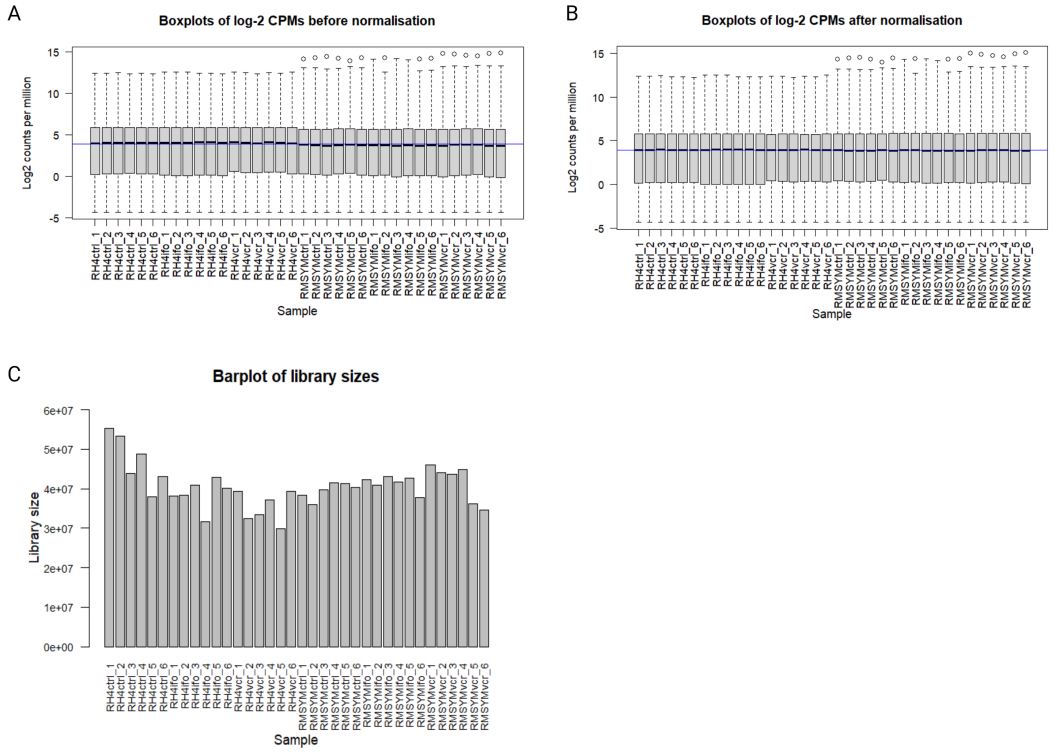


FIGURE 27: Quality control of the filtered acquired resistance data before and after TMM normalisation for the acquired resistance cell models. A) Boxplot of filtered log-2 CPMs before TMM normalisation. B) Boxplots of log-2 CPMs after TMM normalisation C) Barplot of the raw, filtered library sizes. n = 36 (6 treatment conditions with 6 replicates each).

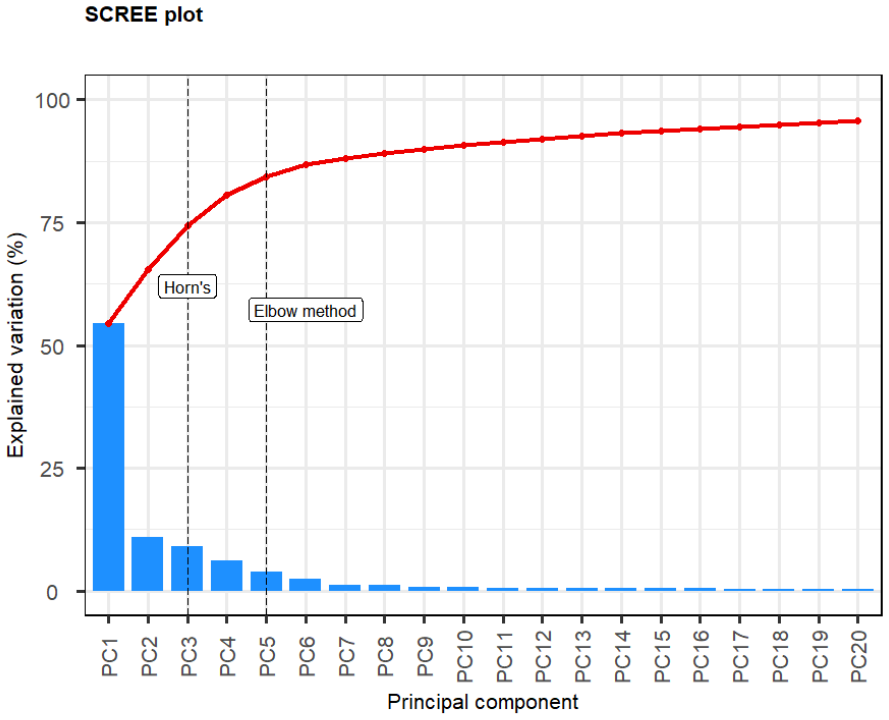


FIGURE 28: Scree plot to visualise the variance explained by each principal component from PCA of the acquired resistance RNA sequencing data. The number of optimum principal components to retains are labelled determined by Horn's method and Elbow method.

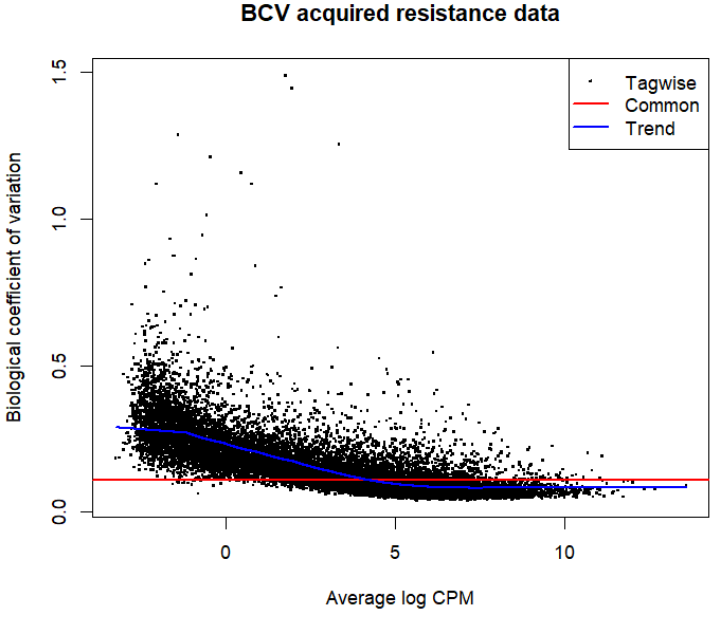


FIGURE 29: Biological coefficient of variation (BCV) plot against gene abundance with estimates of the common, trended and tagwise dispersions for the acquired resistance data.

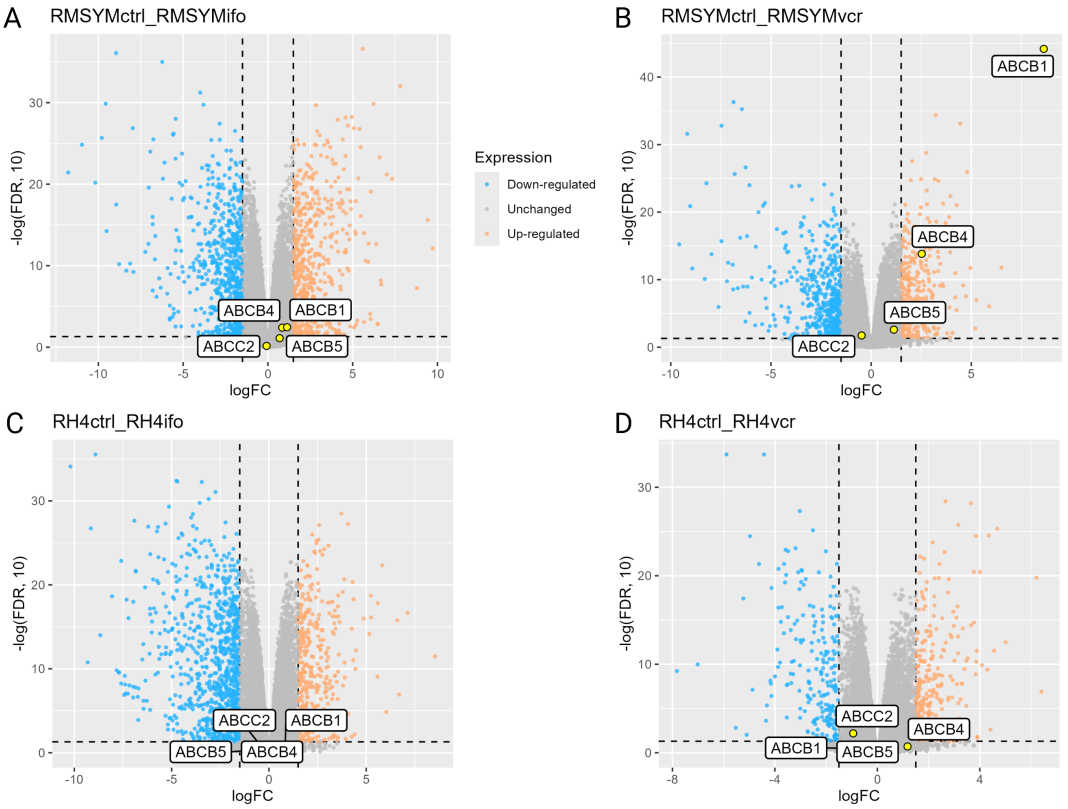


FIGURE 30: Volcano plots with highlighted MDR genes for A) RMSYM IFO-resistant vs RMSYM control and B) RMSYM VCR-resistant vs RMSYM control C) RH4 IFO-resistant vs RH4 control and D) RH4 VCR-resistant vs RH4 control. DEGs were defined as having a LFC >1 or LFC <-1 and adjusted p value <0.05 (Benjamini-Hochberg).

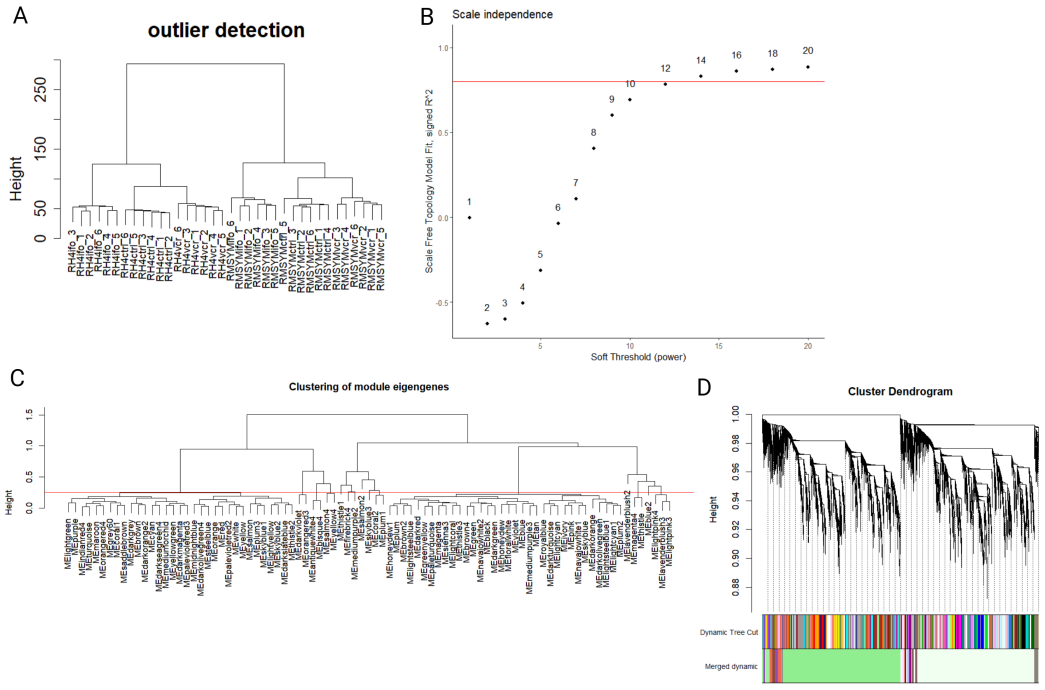


FIGURE 31: Outlier detection, selection of optimal soft threshold power and visualisation of modules for the gene co-expression network for WGCNA using all acquired resistance samples. A) dendrogram of samples B) scale independence plot for selection of optimal soft threshold power C) dendrogram of modules with red line showing dissimilarity threshold of 0.25 D) Cluster dendrogram of modules showing module colour before and after merging.

FIGURE 32: Heatmap showing module-trait correlation for the trait resistance from the acquired resistance data WGCNA generated using all samples. For each module the Pearson correlation and Student asymptotic p-value for correlation are shown.

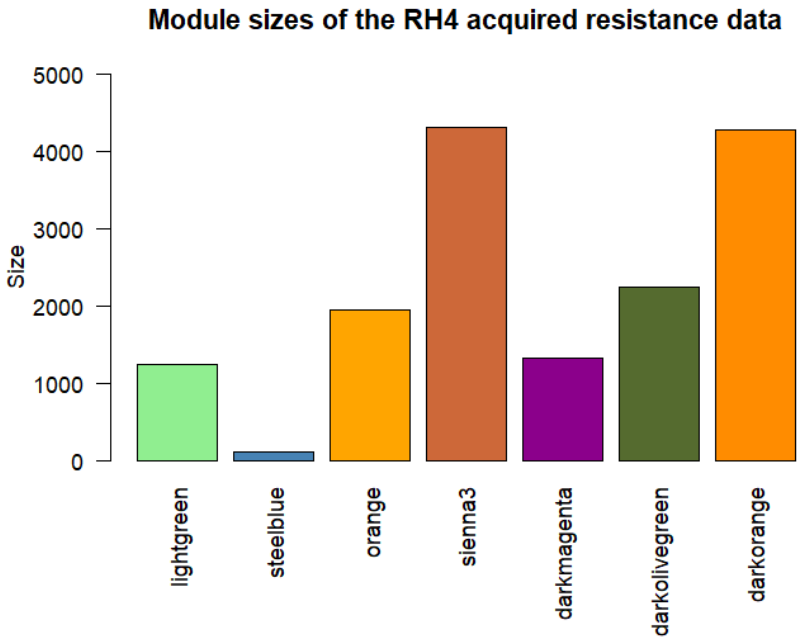


FIGURE 33: Module sizes of the FP weighted gene co-expression network generated from the acquired resistance data after merging modules.

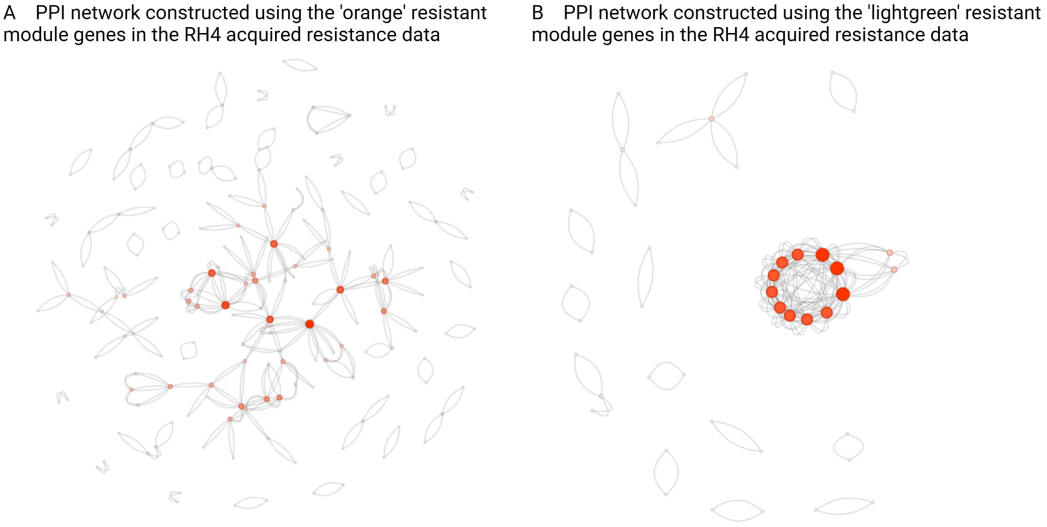


FIGURE 34: Visualisation of the PPI networks constructed from the FP-acquired resistance data. A) all genes from the resistant 'orange' module and B) all genes from the resistant 'lightgreen' module. Nodes are coloured by centrality score with red indicating a higher centrality score and white indicating a lower centrality score. Equal centrality scores are shown in blue.

TABLE 12: FP-acquired resistance hub proteins from the 'orange' and 'lightgreen' resistant modules. PPI hub proteins were defined as having a centrality score ≥8.

RH4 hub proteins	Centrality score	Module
RPS20	24	Lightgreen
RPS13	24	Lightgreen
RPS28	24	Lightgreen
RPL8	20	Lightgreen
RPLP2	20	Lightgreen
RPL27A	20	Lightgreen
RPL13A	20	Lightgreen
RPL31	20	Lightgreen
RPL6	20	Lightgreen
RPL29	20	Lightgreen
RPLP0	20	Lightgreen
EIF3G	8	Lightgreen
EIF3L	8	Lightgreen
BRCA1	18	Orange
RPS5	16	Orange
GRB2	14	Orange
CDK1	14	Orange
ESR1	14	Orange
FAU	14	Orange
MAPK1	12	Orange
ATF2	10	Orange
MAP3K7	10	Orange
XIAP	10	Orange
RPL10	10	Orange
IKBKG	10	Orange
EIF2S3	8	Orange
USP14	8	Orange
CEBPB	8	Orange
GSK3B	8	Orange
BECN1	8	Orange
EIF3I	8	Orange
SQSTM1	8	Orange
FZR1	8	Orange
EIF1	8	Orange

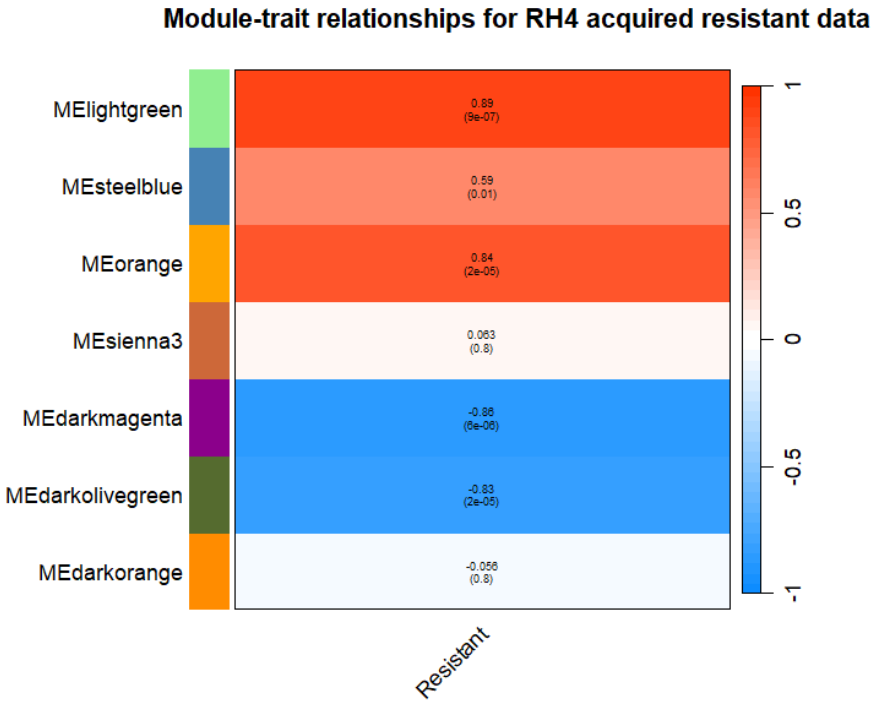


FIGURE 35: Heatmap showing module-trait correlation for the trait resistance from the acquired resistance data WGCNA generated using only RH4 samples. For each module the Pearson correlation and Student asymptotic p-value for correlation are shown.

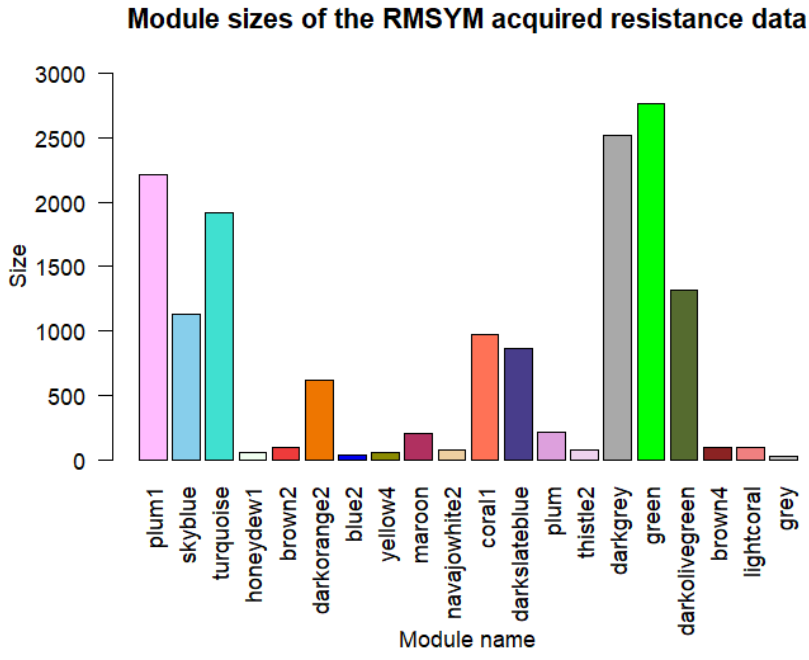


FIGURE 36: Module sizes of the FN-acquired resistance data after merging modules.

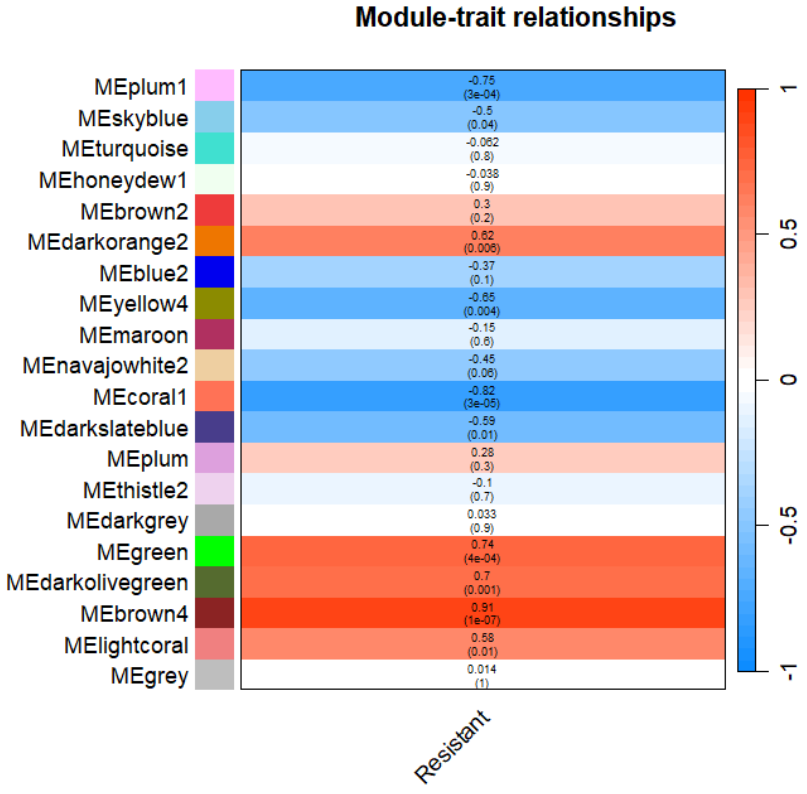


FIGURE 37: Heatmap showing module-trait correlation for the trait resistance from the acquired resistance data WGCNA generated using only FN samples. For each module the Pearson correlation and Student asymptotic p-value for correlation are shown.

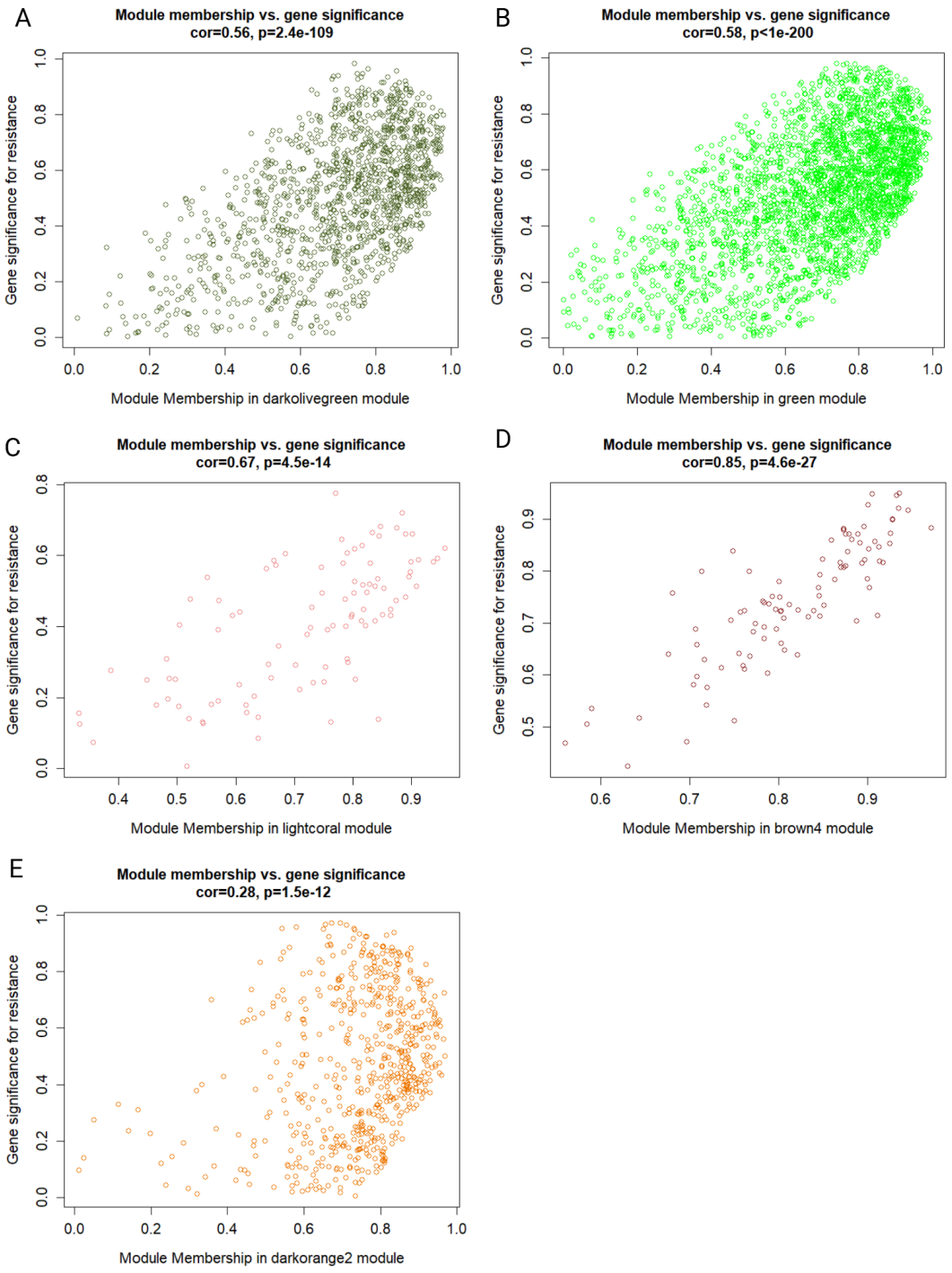


FIGURE 38: Gene significance and module membership plots of the FN WGCNA modules significantly positively correlated with acquired resistance. A) 'darkolivegreen' module B) 'green' module C) 'lightcoral' D) 'brown4' module E) 'darkorange2' module. Each point represents a gene in the modules. The Pearson correlation and Student asymptotic p-value for correlation for gene significance and module membership is shown.

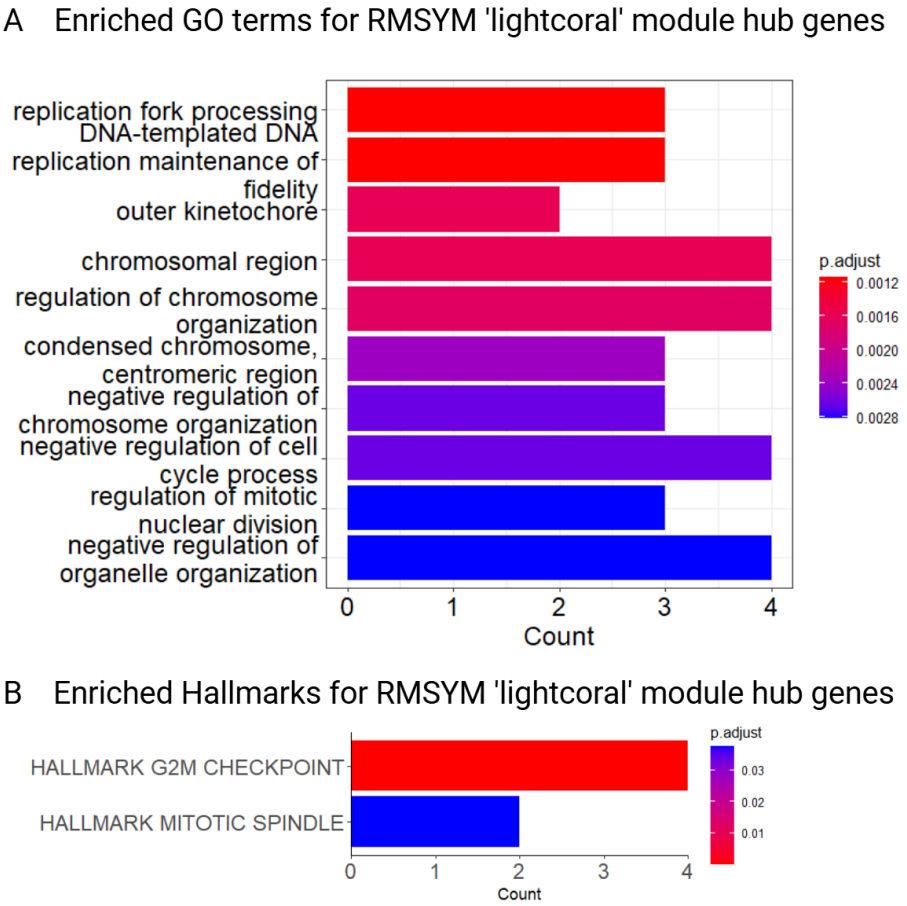


FIGURE 39: Enriched gene sets for 'lightcoral' hub genes from the FN weighted gene co-expression network. Hub genes had a gene significance and module membership >0.6. A) Enriched GO terms B) enriched Hallmarks. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). If more than 10 terms were enriched, the ten terms with the smallest adjusted p values were displayed.

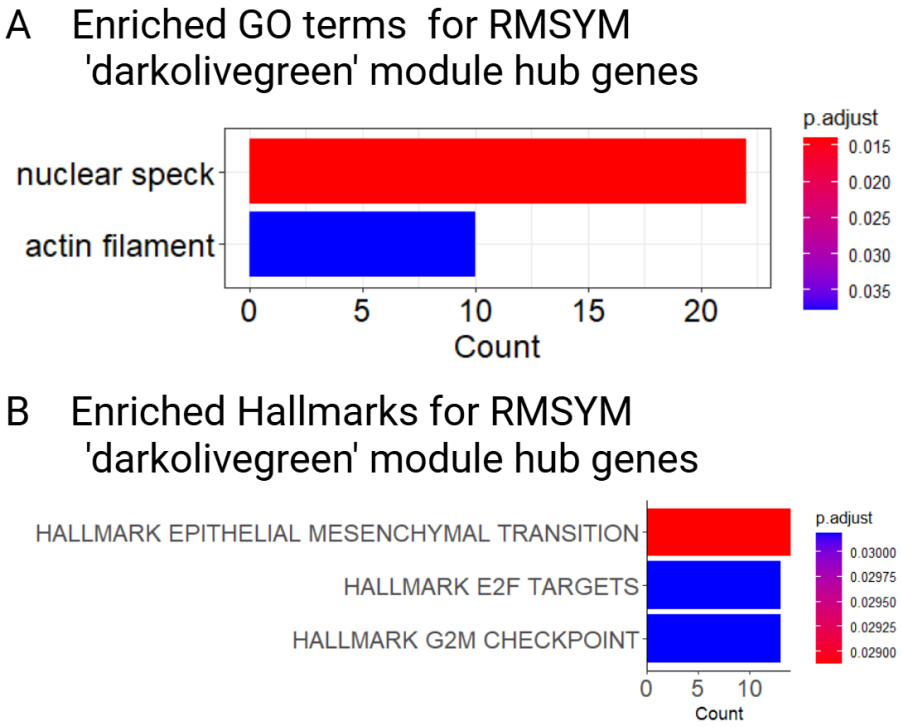


FIGURE 40: Enriched gene sets for 'darkolivegreen' hub genes from the FN weighted gene co-expression network. Hub genes had a gene significance and module membership >0.6. A) Enriched GO terms B) enriched Hallmarks. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). If more than 10 terms were enriched, the ten terms with the smallest adjusted p values were displayed.

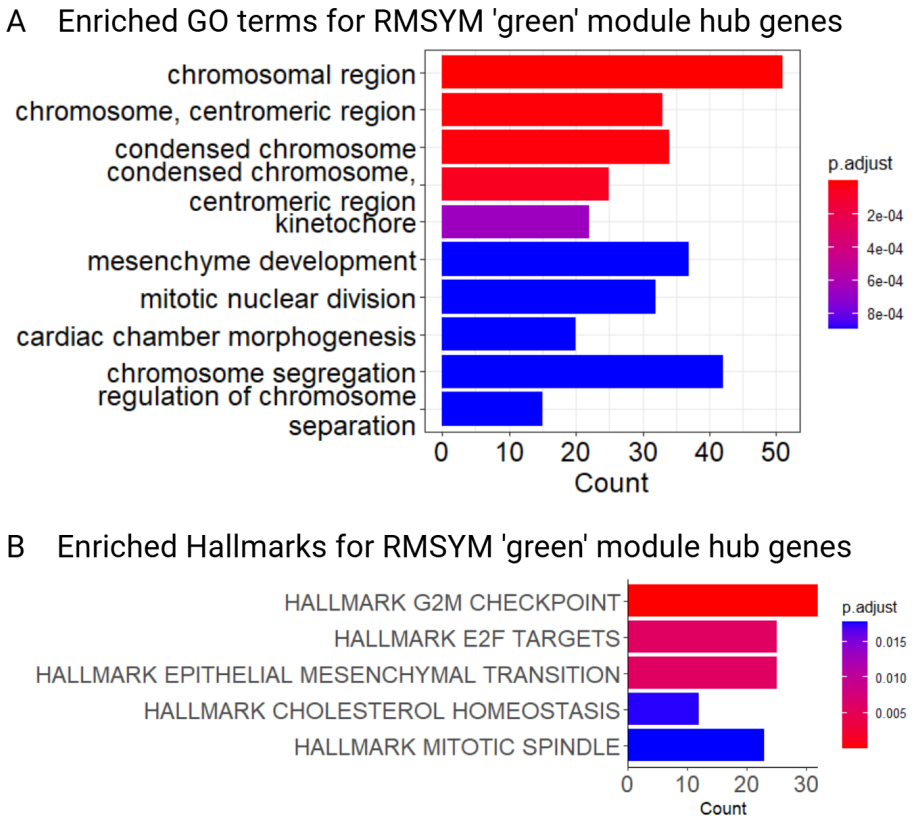


FIGURE 41: Enriched gene sets for 'green' hub genes from the FN weighted gene co-expression network. Hub genes had a gene significance and module membership >0.6. A) Enriched GO terms B) enriched Reactome gene sets C) enriched Hallmarks D) enriched KEGG pathways. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). If more than 10 terms were enriched, the ten terms with the smallest adjusted p values were displayed.

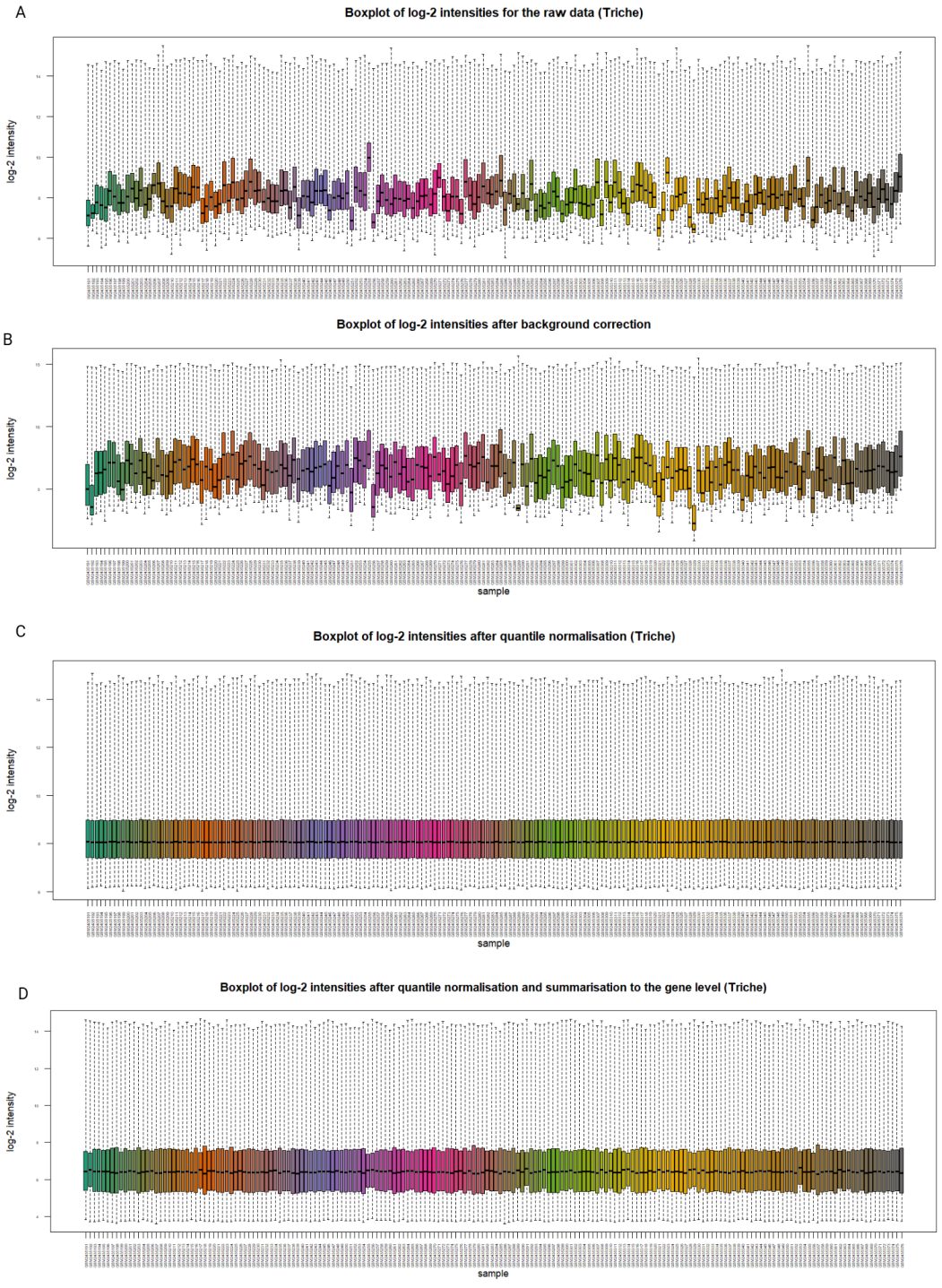


FIGURE 42: Boxplots of intensity distributions for the Triche dataset. A) Raw data B) After background correction with RMA C) After normalisation D) After summarisation. Intensity values are log-2 normalised.

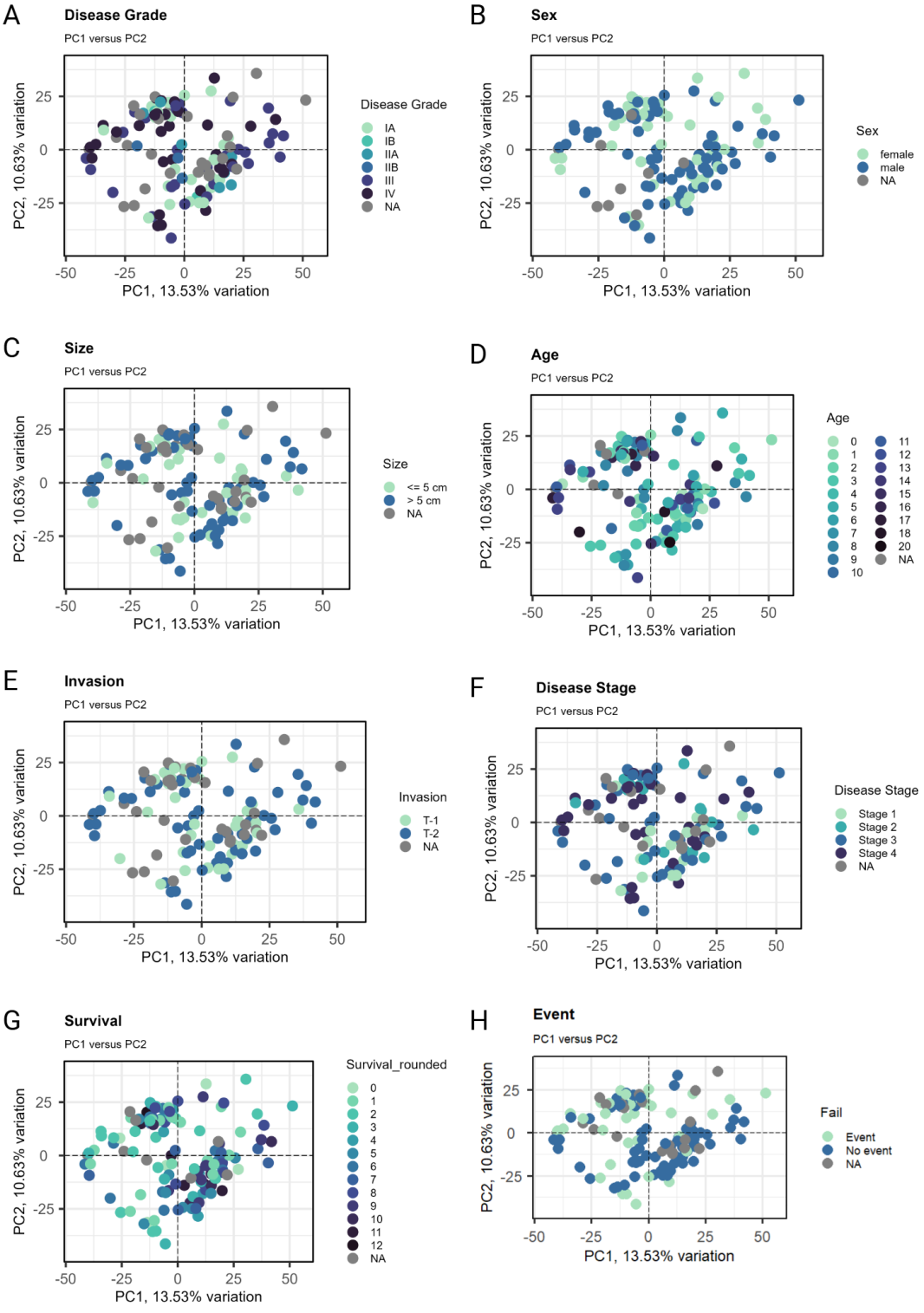


FIGURE 43: Principal component analysis biplots annotated with clinical traits for the Triche dataset. A) Disease grade B) Sex C) Tumour size D) Age E) Invasion F) Disease stage G)Survival H) Event. N=125

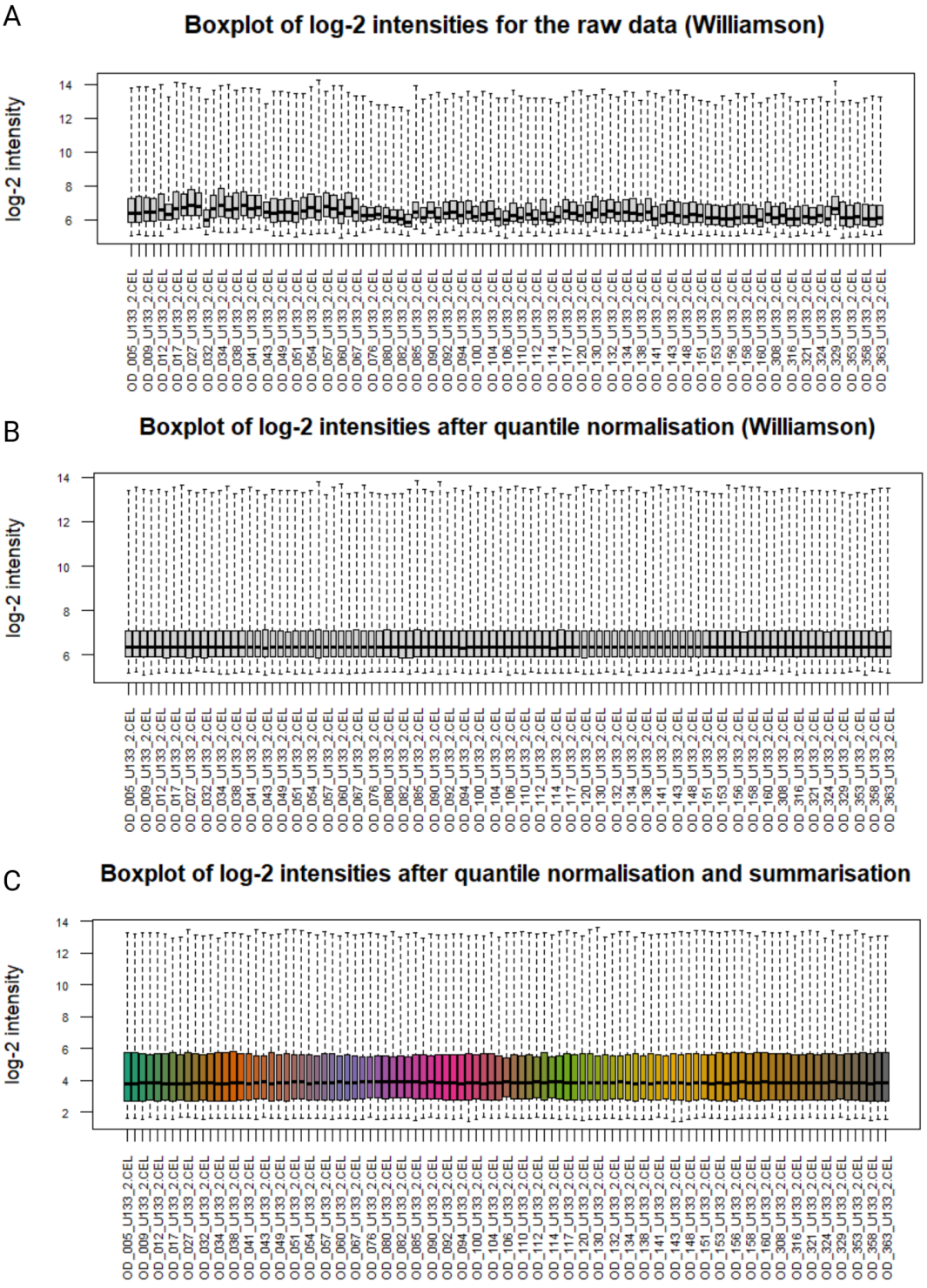


FIGURE 44: Boxplots of intensity distributions for the Williamson dataset. A) Raw data B) After normalisation C) After summarisation. Intensity values are log-2 normalised.

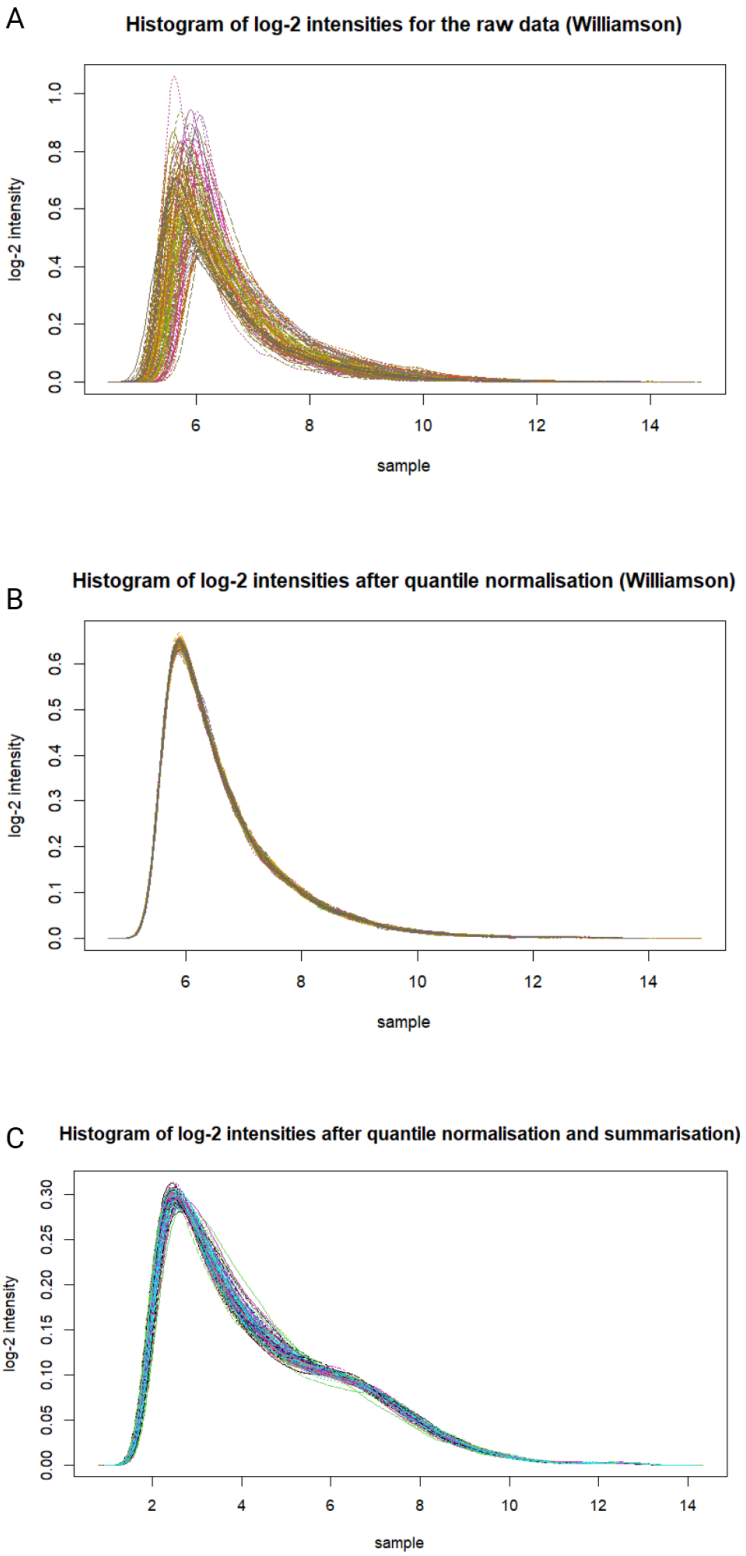


FIGURE 45: Histograms of intensity distributions for the Williamson dataset.A) Raw data B) After normalisation C) After summarisation. Intensity values are log-2 normalised.

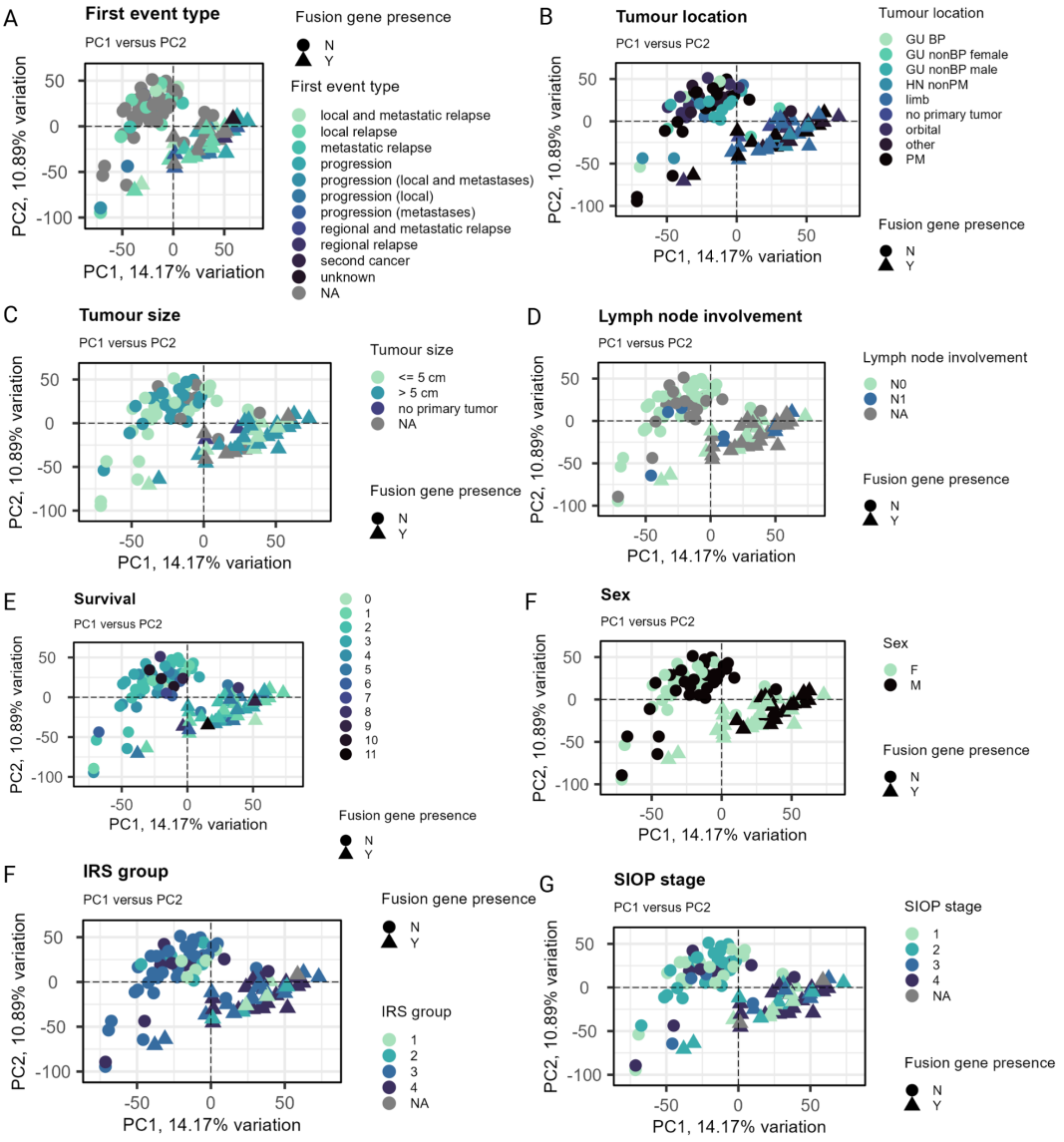


FIGURE 46: Principal component analysis biplots annotated with clinical traits for the Williamson dataset. A) First event type B) Tumour location C) Tumour size D) Lymph node involvement E) Survival F) Sex G) Risk group H) SIOP stage. N=125

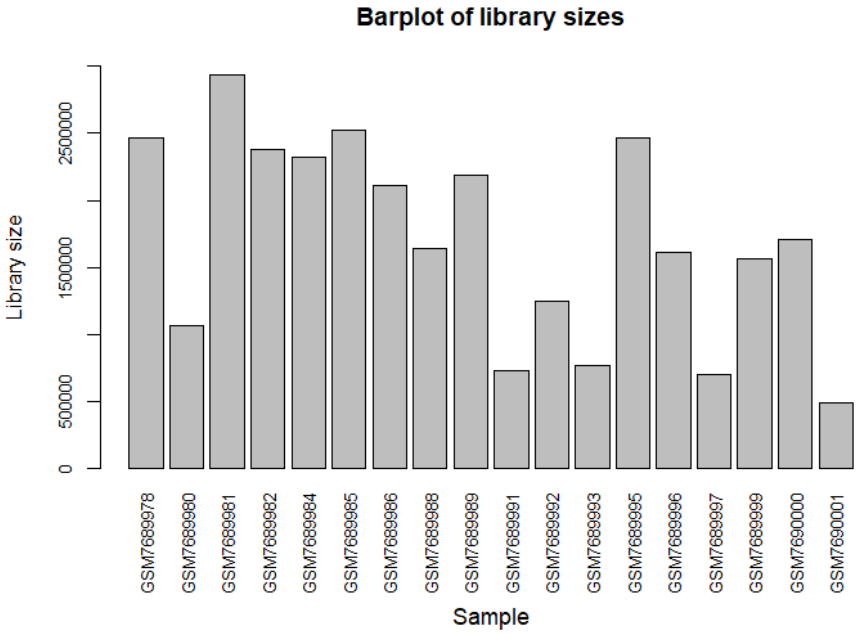


FIGURE 47: Barplot of the raw library sizes for the pre- and mid-treatment RNA seq data. N= 18.

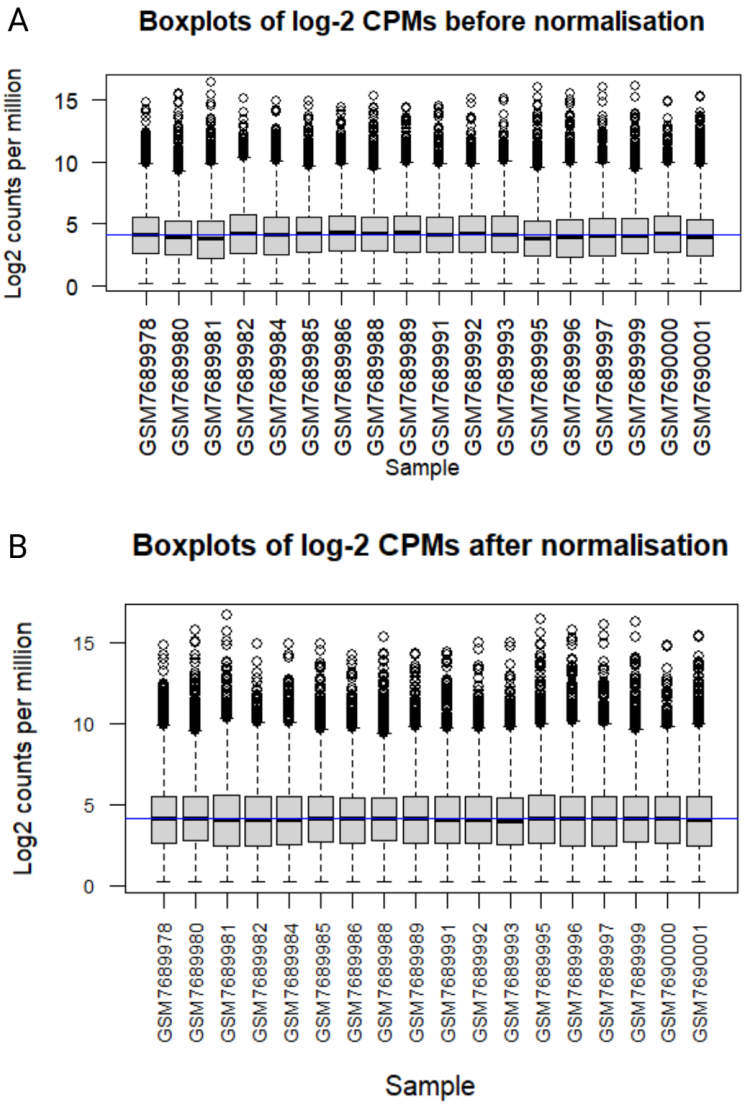


FIGURE 48: Boxplots of raw and normalised RNA-seq data from pre and mid-treatment patient samples. A) Raw data B) After normalisation. N=18.

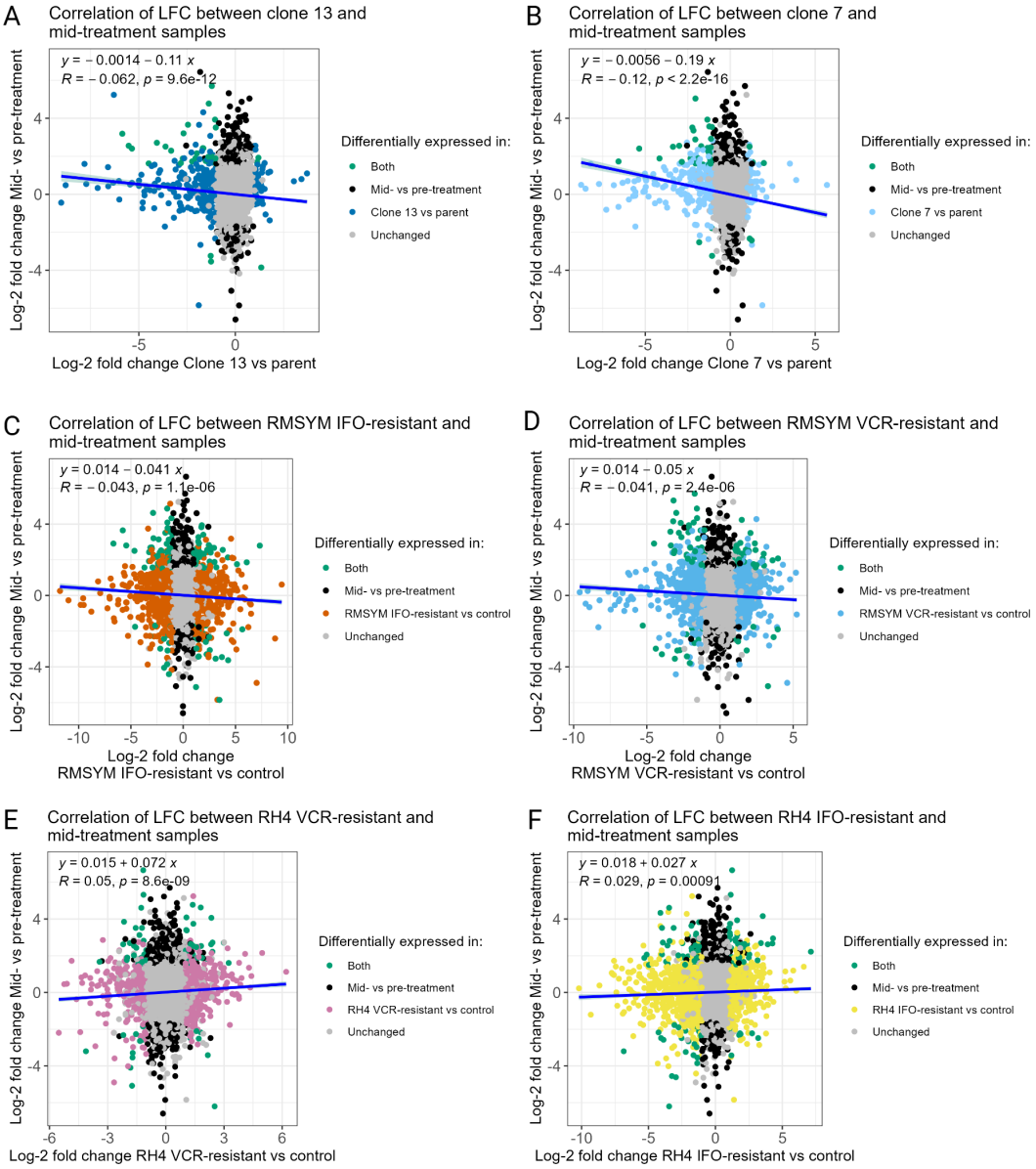
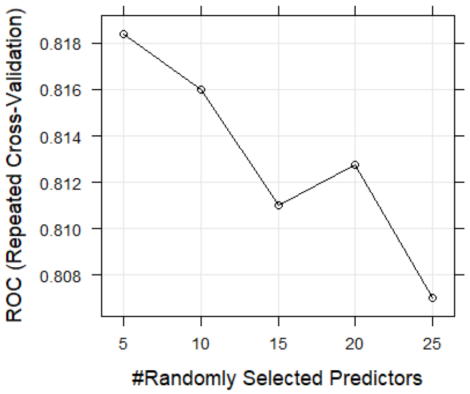
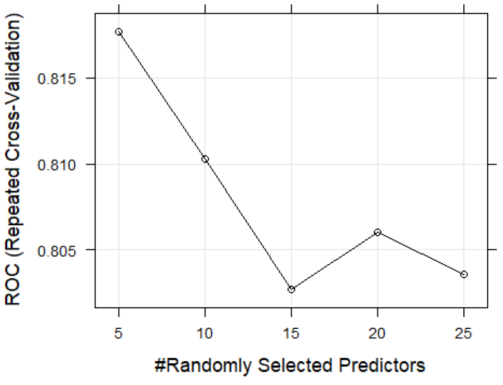


FIGURE 49: Scatterplot showing the LFC of genes for cell models and mid-treatment samples. A) clone 13 vs parent B) clone 7 vs control C) RMSYM IFO-resistant vs control D) RMSYM VCR-resistant vs control E) RH4 VCR-resistant F) RH4 IFO-resistant. DEGs were defined as having an adjusted p value <0.05 (Benjamini-Hochberg) and LFC >1 or LFC <-1.DEGs are highlighted in colour, corresponding to the group in which they are differentially expressed.

A) Plot of optimisation of mtry hyperparameter for standard sampling



B) Plot of optimisation of mtry hyperparameter for downsampling



C) Plot of optimisation of mtry hyperparameter for upsampling

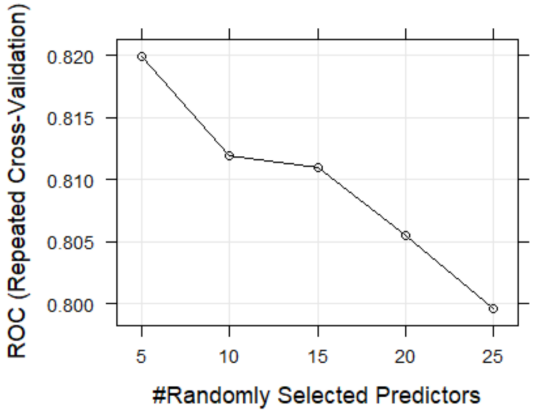


FIGURE 50: Model accuracy with different mtry hyperparameter values for random forest models using different sampling methods. A) Standard sampling B) Downsampling C) Upsampling.

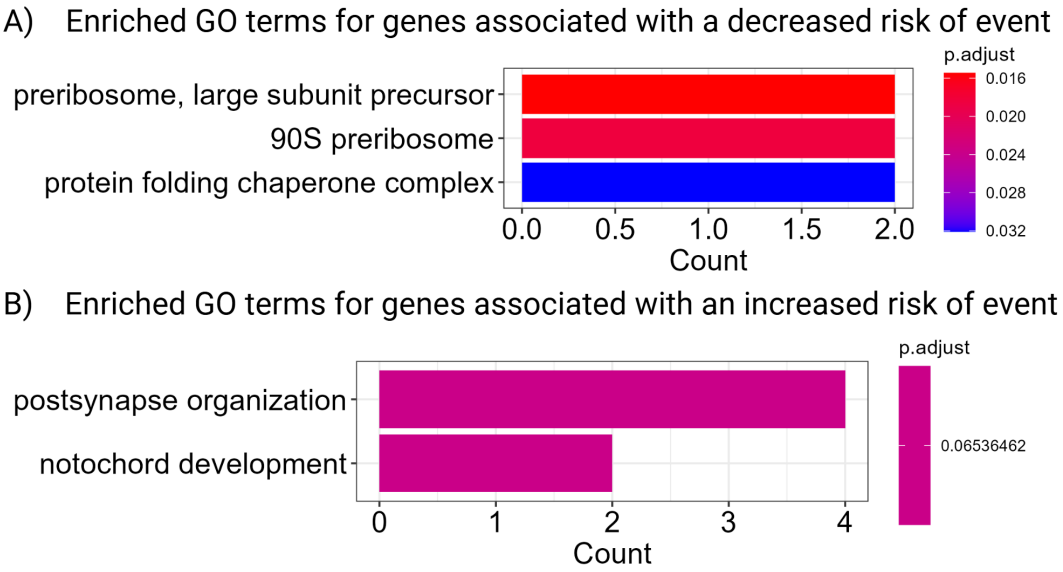


FIGURE 51: GSEA of the 42 features used to train the models. Enriched GO terms for genes associated with A) A decreased risk of event. Significantly enriched gene sets were defined as having an adjusted p value <0.05 (Benjamini-Hochberg). B) An increased risk of event. The significance threshold was reduced to an adjusted p value <0.1 (Benjamini-Hochberg), to provide a general indication of the biological processes the genes might be involved in.