Using a multi-omics approach to identify novel therapeutics in multiple myeloma capable of overcoming drug resistance

Anna James-Bott

St Hilda's College

Nuffield Department of Orthopaedics,
Rheumatology and Musculoskeletal Sciences

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2022

Acknowledgements

I would like to thank my supervisors Dr Adam Cribbs, Professor Udo Oppermann and Dr Sarah Gooding etc etc...

GSK... DTC... Family... Friends...

Abstract

Multiple myeloma (MM) is an incurable cancer of plasma cells, with an average five-year survival rate of approximately 50%. Newer therapeutics, namely proteasome inhibitors (PI) and immunomodulatory imide drugs, have almost doubled median survival time of MM patients. However, most patients relapse and become resistant to drugs they previously have been treated with. Acquired anti-cancer drug resistance remains one of the biggest barriers in the treatment of myeloma. Therefore, identifying novel therapeutics effective against MM, which are capable of overcoming drug resistance is of the utmost importance. Recently the prolyl-tRNA synthetase inhibitor, Halofuginone, has been shown to be effective against cancer, including one study demonstrating effectiveness against MM.

Contents

List of Figures

List of Tables

List of Abbreviations

BCR . . . . . B-cell receptor

Ig . . . . . Immunoglobulin

NK . . . . . Natural killer

TCR . . . . . T cell receptor

T_H . . . . . Helper T (cell)

T_{reg} . . . . . Regulatory T (cell)

T_c . . . . . Cytotoxic T (cell)

PBMC . . . . Peripheral blood mononuclear cell

DC . . . . . Dendritic cell

BM . . . . . Bone marrow

MM . . . . . Multiple Myeloma

RRMM . . . . Relapsed and refractory MM

ER . . . . . Endoplasmic reticulum

MM . . . . . Multiple Myeloma

RRMM . . . . Relapsed and refractory MM

HSC . . . . . Hematopoietic stem cell

MHC . . . . . Major histocompatibility complex

V(D)J . . . . . Variable/diversity/joining

IL . . . . . Interleukin

PC . . . . . Plasma cell

MGUS . . . . Monoclonal gammopathy of unknown significance

SMM . . . . . Smoldering multiple myeloma

PI . . . . . Proteasome inhibitor

IMiDs . . . . . Immunomodulatory imide drugs

UPS . . . . . Ubiquitin proteasome system

UPR . . . . . Unfolded protein response

DNA . . . . . Deoxyribonucleic acid

cDNA . . . . . Complementary DNA

RNA . . . . . Ribonucleic acid

mRNA . . . . Messenger RNA

tRNA . . . . . Transfer RNA

miRNA . . . . Micro RNA

rRNA . . . . . Ribosomal RNA

PCR . . . . . Polymerase chain reaction

NGS . . . . . Next generation sequencing

TGS . . . . . Third generation sequencing

ONT . . . . . Oxford Nanopore Technologies

WGS . . . . . Whole genome sequencing

RNA-seq . . . Ribonucleic acid sequencing

scRNA-seq . . Single cell RNA-Seq

dscRNA-seq . Droplet-based scRNA-Seq

ChIP-seq . . . Chromatin immunoprecipitation sequencing

ATAC-seq . . Assay for transposase-accessible chromatin sequencing

CB . . . . . Cellular barcode

UMI . . . . . Unique molecular identifier

LC-MS/MS . Liquid chromatography with tandem mass spectrometry

PCA . . . . . Principle component analysis

DMSO . . . . . Dimethyl sulfoxide

UMAP . . . . Uniform Manifold Approximation and Projection

tSNE . . . . . t-distributed Stochastic Neighbor Embedding

BUS . . . . . Barcode, UMI, set

QC . . . . . Quality control

SCE . . . . . Single cell experiment

GTF . . . . . Gene transfer format

CGAT . . . . . Computational genomics analysis toolkit

GEO . . . . . Gene expression omnibus

aaRS . . . . . Aminoacyl tRNA synthetase

EPRS . . . . . Glutamyl-prolyl-tRNA synthetase

ProRS . . . . . Prolyl-tRNA synthetase

MSC . . . . . Multi-tRNA synthetase

AIMP . . . . . aaRS interacting multifunctional proteins

FF . . . . . Febrifugine

HF . . . . . Halofuginone

BTZ . . . . . Bortezomib

CFZ . . . . . Carfilzomib

CFZr . . . . . CFZ resistant

FDA . . . . . Food drug administration

WT . . . . . Wild type

1 Introduction

1.1 Overview

Multiple myeloma (MM) is an incurable cancer of plasma cells. Multiple myeloma accounts for 1-2% of all cancers and 10% of all hematological malignancies[1]. In 2020, it was estimated that the total worldwide incidence of MM was 160,000[2]. Over the last two decades, several new classes of drugs have been approved for treatment in MM, including proteasome inhibitors (PIs), Immunomodulatory imide drugs (IMiDs) and monoclonal antibodies. Subsequently a drastic improvement in survival rates has been seen— median survival time for newly diagnosed MM patients has almost doubled in the last 10 years[3]. However, MM still remains an incurable disease, with patients eventually relapsing and becoming resistant to drugs they have previously been treated with. Drug resistance is one of the biggest barriers in the treatment of MM. Therefore, the generation of novel compounds effective against MM is incredibly important to extend progression-free survival of patients and work towards finding a cure to MM. This thesis aims to identify novel drugs effective against MM, which are capable of overcoming drug resistance to conventional MM therapies. It also aims to characterise the transcriptional landscape of MM, following treatment with these novel compounds, to ascertain the mechanism of action in MM and identify any potential side effects of treatment. Additionally, MM is very heterogenous disease with prominent interactions with the surrounding immune microenvironment, therefore MM cells must be examined

at the single-level along with other cells residing in the bone marrow niche, to fully capture the heterogenity and immune interactions we see in myeloma.

This introduction will give a brief overview of the immune system, multiple myeloma (MM), and current MM treatments. The genome, epigenome, transcriptome and proteome will be introduced, followed by a review of the technological advancements and informatics tools used to decode them. Chapter 2 introduces the enzyme class: aminoacyl tRNA synthetases (aaRS) and reviews literature regarding their role in disease and the application of therapeutics targeting enzymes in this class. Next, literature surrounding the drug Halofuginone, a Prolyl-tRNA synthetase (ProRS) inhibitor, is reviewed and its application in MM is considered.

1.2 The immune system

Humans are exposed to millions of potential pathogens every day and would die without defences to be able to protect themselves from infection. These defences can be innate or adaptive. An example of an innate defence is the skin acting as a physical barrier between the outside world and the body. Innate responses are non-specific and relied upon as the first line of defence, however sometimes a more sophisticated, specialised response is required- called the adaptive immune response[4]. Both innate and adaptive immune responses rely on cells produced by the hematopoetic system. Hematopoiesis is the process in which the body produces blood[5]. Postnatally the major hematopoetic organs are bone marrow, lymph nodes, the spleen and the thymus.

1.2.1 Hematopoietic stem cell differentiation

All blood cells originate from multipotent hematopoietic stem cells (HSCs), located in the bone marrow (BM). Stem cells are precursor cells which can give rise to at least one type of differentiated (mature) cell, with the capability of indefinite self-renewal. HSCs divide to either generate more HSCs (self-renewal) or committed progenitor cells. HSCs are thought to go through stages of progressive restrictive commitments, until they eventually become fully differentiated blood cells such as NK cells, T

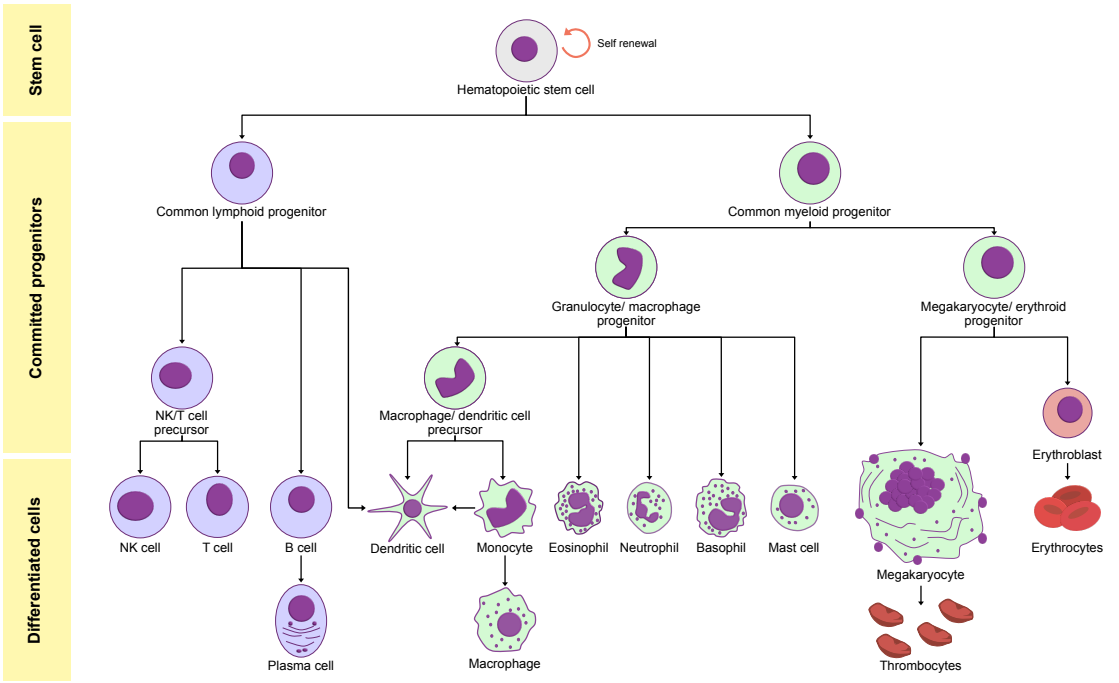


Figure 1.1: Hematopoietic stem cell (HSC) cell differentiation. Multipotent HSCs divide and either self-renew or commit to common myeloid or common lymphoid progenitor cells. The committed progenitor cells then further commit to progressive restrictive progenitors, then finally differentiated cells such as B cells, T cells, macrophages and erythrocytes. In adult mammals, all cells shown develop in the bone marrow, except for T cells, which develop in the thymus, and macrophages, which develop from blood monocytes. Numerous dendritic cell phenotypes exist, they can be derived from lymphoid and myeloid progenitors, as well as monocytes.

cells and macrophages (Figure 1.1). Firstly, HSCs lose self-renewal capacity, they then commit to either a myeloid or lymphoid fate. Myeloid progenitors give rise to monocytes, macrophages, granulocytes (neutrophils, eosinophils, basophils), mast cells, megakaryocytes and erythrocytes (red blood cells). Lymphoid progenitors give rise to natural killer (NK) cells, T cells and B cells. Both myeloid progenitors and lymphoid progenitors can produce dendritic cells (DCs)[5]. Each stage of commitment correlates with changes in expression of specific genes and transcriptional regulators, for the given subset of blood cells[6].

1.2.2 The innate immune response

The innate immune response, or 'non-specific immunity', is the first defence a host has to prevent infection from invading pathogens[4]. Innate immunity is

found in all species, from early multi-cellular organisms to humans[7]. Innate immunity comprises physical barriers, such as tight junctions in the skin and epithelial/mucous membranes, as well as defensive cell-coordinated responses. Once foreign pathogens enter the host, numerous signalling cascades are initiated. Host cells are able to recognise foreign microbials by characteristic features, known as pathogen-associated molecular patterns (PAMPs). Once PAMPs are recognised by host pattern recognition receptors (PRRs), inflammatory responses are triggered and the invading pathogen is targeted for phagocytosis. Phagocytic cells, for example neutrophils, monocytes and macrophages (Figure 1.1), non-specifically engulf foreign pathogens, and use a combination of toxic molecules, degradative enzymes and antimicrobial peptides to kill the invaders. Additionally, various cytokines and inflammatory mediators are released. NK cells directly kill host cells infected by a virus[4] and an array of tumour cells[7].

In addition to its role preventing infection and eliminating invading pathogens, the innate immune response also stimulates the acquired immune response. Dendritic cells act as a functional link between innate immunity and the adaptive immune response. DCs become activated after they phagocytose microbes, they then cleave the microbial proteins, which bind to MHC proteins and together are transported to their cell surface. The activated DCs then carry the peptide-MHC complexes to lymphoid organs, which in combination with DC-secreted cytokines and expression of co-stimulatory proteins and cell-cell adhesion molecules, activate T cells to initiate the adaptive immune response. Innate immunity is a fast response (minutes to hours), whilst adaptive immunity takes longer (days to weeks)[4].

1.2.3 The adaptive immune response

During evolution, vertebrates have also developed adaptive immune responses, in addition to innate responses[7]. Adaptive immune responses are specific to the pathogen that induced the response. Adaptive immunity is dependent on millions of B cells and T cells clones. Each clone shares a unique cell-surface receptor that can bind to a specific pathogen antigen. These receptors are incredibly antigen selective,

they can distinguish between optical isomers of the same molecule, and proteins differing by one single amino acid. Two classes of adaptive immune responses exist: antibody responses, co-ordinated by B cells, and cell mediated immune responses, co-ordinated by T cells[4].

T cells

T cells are predominantly produced in the thymus (hence their name). T cells are activated by fragments of partially proteolysed antigens displayed on the surface of antigen presenting cells (APCs), for example dendritic cells or macrophages. T cells recognise these fragments with their cell-surface T cell receptors (TCRs). All effector T cells interact with other host cells in the body, therefore have a short range of effect[4].

There are three main classes of T cells: cytotoxic (T_C) T cells, helper T (T_H) cells and regulatory T (T_{reg}) cells. Effector cytotoxic T cells express CD8 co-receptors ($CD8^+$ T cells) and act like NK cells by directly killing infected host cells, protecting against intracellular pathogens. Effector helper and regulatory T cells express CD4 co-receptors ($CD4^+$ T cells)[8]. Helper T cells work by activating other host cells in the innate and adaptive immune system, for example, by activating B cells to secrete antibodies and undergo class switching, by helping activate macrophages to destroy intracellular pathogens, recruiting neutrophils and promoting pro-inflammatory cytokine production, and inducing naive cytotoxic T cells to become effector cells. Helper T cells can be further divided into T_H1, T_H2, T_{FH} and T_H17 subclasses[4, 8]. Regulatory T cells have the opposite function to helper T cells. T_{reg}s prevent an excessive immune response that could be potentially harmful to the body. The development, activation and function of other immune cell types is suppressed by T_{reg}s by secreting suppressive cytokines such as TGFβ and interleukin (IL)-10, and inhibitory cell-surface proteins. This means that the response to harmless ingested or inhaled antigens is limited[4].

B cells

B cells and plasma cells produce immunoglobulins (Igs). Immunoglobulins (Ig) are typically large Y-shaped proteins. The function of immunoglobulin molecules is to recognise and bind specific foreign antigens on pathogens. Binding of immunoglobulins to antigens renders the virus or microbial toxin inactive as it blocks their ability to bind to host cells. Additionally, Ig binding makes it easier for phagocytic cells to ingest the pathogen[4].

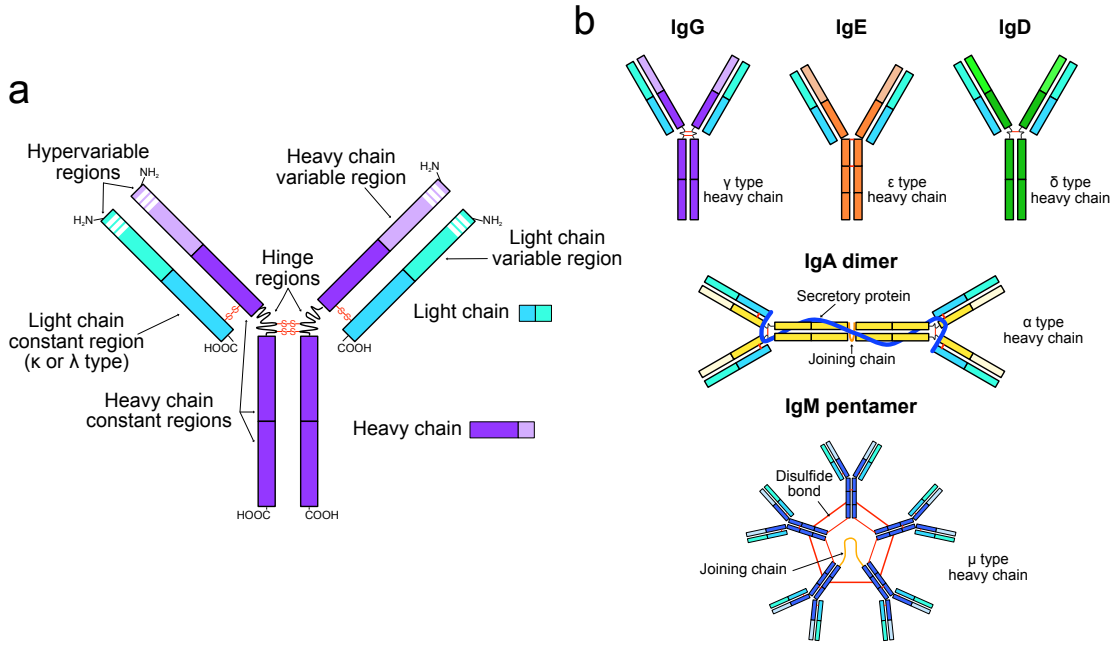


Figure 1.2: Immunoglobulin (Ig) diagram. a) A bivalent immunoglobulin molecule (specifically IgG). Igs are comprised of two identical heavy chains and two identical light chains. Each chain has a constant region (κ or λ for light chains, and γ, ϵ, δ, α or μ for heavy chains), and a variable region. Hypervariable regions are domains on heavy and light chains that come in direct contact with antigens. These regions are frequently mutated to allow binding of diverse antigens. The two heavy chains are covalently linked by disulfide bonds in their hinge regions. Hinge regions offer flexibility which improves antibody-antigen binding efficiency. Heavy and light chains are held together by a mixture of covalent disulfide bonds (red) and non-covalent bonds. b) The five major classes of Igs in humans: IgG, IgE, IgD, IgA and IgM, in their main secretory (antibody) form. IgE and IgM have no hinge regions and are composed of four heavy chain constant regions, not three. IgA and IgM are shown in their dimer and pentamer polymer formations, however other polymer formations are also found in the body.

Some Igs in circulation are found in polymer formations of multiple Ig units, such as as dimers or pentamers. A single Ig unit consists of two identical heavy

chains and two identical light chains (Figure 1.2a). Each heavy and light chain has a constant region and a variable region. Heavy chains consist of one variable domain, and three or four constant domains. Light chains consist of one variable domain and one constant domain. Constant regions are found at the C-terminal end of Igs, and are a constant sequence for each given chain type. Variable regions are found at the N-terminal end of Igs. The variables region of the heavy and light chain forms the antigen-binding sites of Igs. The amino acid sequence variation of variable domains confer diversity for antigen-binding. Both chains also contain hypervariable regions, which come in direct contact with antigens. These domains are frequently mutated to allow for even more diversity of antigen binding[9].

Humans produce five main isotypes of immunoglobulin: IgM, IgD, IgG, IgA and IgE (named after the greek letter of their respective heavy chain; Figure 1.2b). IgG can be divided into 4 subclasses, IgG1, IgG2, IgG3, and IgG4, and IgA can be divided into IgA1 and IgA2[9]. Igs can either be membrane-bound (mIgs) on the cell-surface or secreted (sIgs, also called antibodies). Naive B cells only produce mIgs– they do not secrete antibodies. IgM molecules are the first immunoglobulins produced by newly formed B cells. Membrane IgMs (mIgMs) associate with two signalling proteins Ig-α and Ig-β; these complexes are known as B-cell antigen receptors (BCRs)[10, 11].

Each B cell's individual Ig antigen-binding site is determined by a process called V(D)J recombination, whereby separate Ig gene segments are joined together. Separate loci, on three separate chromosomes code for κ light chain (chromosome 2), λ light chain (chromosome 22) and the heavy chains (chromosome 14). Each light chain locus consists of one or more coding sequence for constant (C) regions, and a set of variable (V) and joining (J) sequences. The heavy chain locus contains these elements plus an added diversity (D) gene segment. These gene segments are rearranged and brought together by site-specific recombination, aided by V(D)J recombinase enzyme. The RNA that codes for the whole Ig polypetide chain is produced by co-transcribing the constant regions and variable regions together. From all the potential combinations of V, D and J gene segments, in theory, over 1.5

million different antigen-binding sites could be made. Diversity of antigen-binding sites is further increased by junctional diversification, whereby variable numbers of nucleotides are lost or randomly inserted at joining sites. This can increase diversity of variable coding sequences by 10^8 fold. However, this process can also shift the reading frame to produce non-functional genes. B cells that produce non-functioning Igs subsequently die in the bone marrow. Additionally, B cells that produce BCRs that bind too strongly to self-antigens in the BM undergo another round of V(D)J recombination to alter specificity; if this fails again then B cells die by apoptosis. Once a B cell can produce functional Igs, the V(D)J recombination process is switched off, and it only makes Igs with the same antigen specificity[4].

1.2.4 B cell maturation

Most B cells die in the bone marrow soon after developing, however some will develop in the bone marrow, where initial stages of maturation occur and then migrate to secondary lymphoid organs, such as the spleen. Within secondary lymphoid organs, numerous critical decisions on B cell fate are made, involving complex transcriptional networks, cell interactions, gene rearrangements, and mutations[12, 13]. As B cells develop and leave the bone marrow to populate peripheral lymphoid tissues, they undergo alternative mRNA splicing and start to express IgD BCRs in addition to IgM; they are henceforth known as mature naive B cells. IgM and IgD BCRs on a single mature naive B cell all contain identical antigen-binding sites. To fully differentiate, B cells require antigen stimulation. Upon antigen stimulation, mature naive B cells are activated to differentiate into either Ig-secreting effector B cells (plasma cells) or memory B cells. Terminally differentiated plasma cells are the final effectors of the B cell lineage. Plasma cell differentiation involves the loss of expression of many genes, such as CD19, CD20 and increased expression of plasma cell markers, such as CD138 and CD38. Plasma cells have an extensive rough endoplasmic reticulum (ER), and have numerous genes involved in antibody secretion upregulated, including *XBP-1* and *CHOP*, to enable the production of copious amounts of antibody[14]. Antibodies recognise and bind to the specific

foreign antigen on the pathogen which stimulated their production. Binding of antibodies to antigens renders the virus or microbial toxin inactive as it blocks their ability to bind to host cells. Additionally, antibody binding makes it easier for phagocytic cells to ingest the pathogen[4].

Initially, effector B cells only secrete primary Igs (IgM and IgD) that have low binding affinity for antigens. However, further antigenic-stimulation, helper T cell activation, various cytokine secretions and activation-induced-deaminase (AID) enzyme activity, leads to processes which increase binding affinity of antibodies (affinity maturation) and diversify Ig isotype production (class switching). The main stages of B cell maturation can be seen in Figure 1.3.

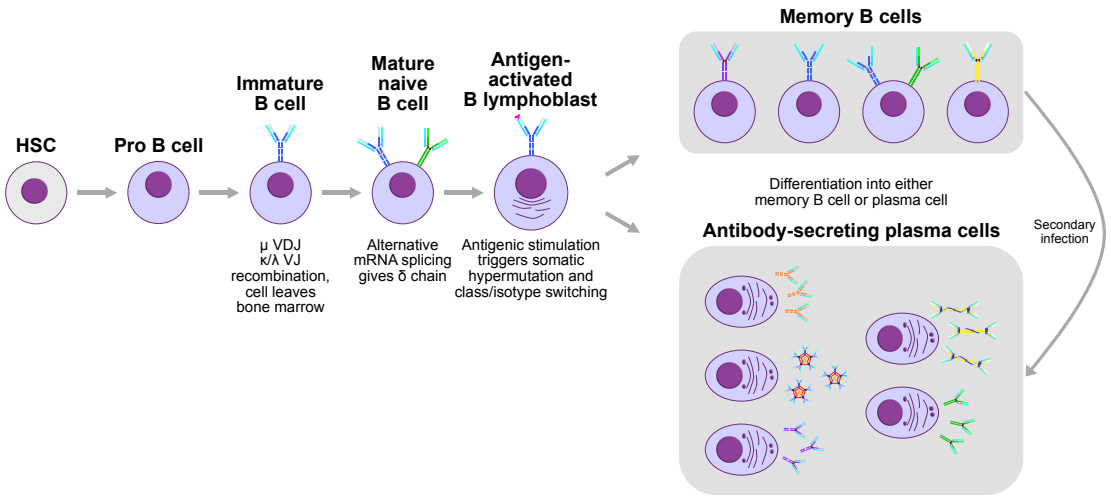


Figure 1.3: A simplified diagram of B cell maturation. Once hematopoietic stem cells (HSC) commit to the B cell lineage, the cells undergo a number of changes in the bone marrow and secondary lymphoid tissues until they eventually commit to either becoming antibody-secreting plasma cells or memory B cells.

Affinity maturation

Affinity maturation occurs through a process known as somatic hypermutation, whereby activated B cells form germinal centers and their variable regions mutate at a rate approximately a million times greater than spontaneous mutations. The enzyme AID is required for this process. Not many of the Igs that undergo hypermutation will have a stronger affinity for the antigen present. However, B cells produce

BCRs and antibodies, both of which contain the same antigen-binding site. The B cells with a higher affinity antigen-binding site for the given antigen, will be preferentially stimulated at their BCRs. This is in turn means that clones with increased affinity will preferentially proliferate and survive over other B cells with worse affinity (similar to how evolution and natural selection works in population genetics). Most other B cells in the germinal center will undergo apoptosis and die. After numerous repeated cycles of mutation and selection, the remaining effector B cell clones will have improved Ig affinity 10–5000 fold for the antigen during the course of the immune response[15].

Class switching

Class switching often occurs after somatic hypermutation has taken place, however they are independent processes, and B cells can class switch without having already undergone somatic hypermutation[16]. Various cytokines produced by activated T helper cells, such as IFNγ produced by T_H1 cells, IL-4 and IL-5 produced by T_H2 cells and IL-21 produced by T_{FH}, can promote B cell class/isotype switching. Class switching allows effector B cells to produce the secondary class of Igs: IgG, IgA and IgE. Class switching is dependent on the mechanism 'switch recombination', where the constant heavy chain region undergoes a number of DNA cutting/rejoining events that bring any of the secondary Igs constant heavy chain exons adjacent to heavy VDJ exon[16]. The mechanism is not yet fully understood, however AID is necessary to facilitate class switching. Daughter cells from the same effector B cell will then be able to produce different isotypes of Igs, but all retaining their original specific antigen-binding site. In the early stages of an immune response, IgM is the main circulating antibody. After a lag period, IgG becomes the main circulating antibody in the blood[4].

The Ig isotypes have distinct immune functionality and tissue distribution. For example, the tail of IgG molecules can bind to macrophages and neutrophils, aiding in phagocytosis of microorganisms, it is also the only Ig that can cross the placenta; IgE molecules can bind to mast cells and basophils, and play a large role in allergic

reactions and parasitic infections; IgA molecules are the main antibody in bodily secretions[4]. By each Ig class retaining the same antigen-binding site, this means that antigen-specificity can be distributed amongst the various isotypes. Therefore, a pathogen can be fought using an array of each isotype's biological properties[17].

Immunological memory

Following antigenic-stimulation, B cells proliferate and differentiate into either antibody-secreting plasma cells or memory B cells. Memory B cells are long-lived, quiescent cells that possess the ability to 'remember' past pathogens and respond quickly to the antigen[18]. Memory B cells express BCRs of various Ig isotypes. Plasma cells appear to consist of two distinct categories: short-lived plasma cells, which have life-spans of several months and are located in extrafollicular locales such as in medullary chords of lymph nodes or the red pulp of the spleen, and long-lived plasma cells (LLPCs), which have a much longer life-span, ranging from months to decades[19, 20]. LLPCs are mainly found in the bone marrow[19, 20]. At the end of a primary immune response (first exposure to an antigen), the majority of antigen-specific B cell clones undergo apoptosis, leaving memory cells and LLPCs as the only remaining cells from the B cell lineage. Upon second exposure to an antigen, these cell types speed up the secondary immune response and reduce the lag period seen in the primary immune response. Existing memory B cells, which already possess high-affinity receptors for the given antigen, are expanded for immediate differentiation into short-lived plasma cells that can secrete high affinity antibodies. Human memory B cells can be split up into three main categories: IgM^+ only, IgM^+IgD^+, and class switched IgG^+ and IgA^+[18]. IgG^+ memory B cells are more likely to differentiate into plasma cells following re-infection. Immunological memory is the basis for the success of vaccinations, where previous exposure to a weakened/dead microorganism or its protein/toxins stimulates an immune response and 'readies' the body for if it comes in contact with the full disease.

1.3 Multiple myeloma

1.3.1 Multiple myeloma cells

Multiple myeloma (MM) is a malignancy of terminally differentiated plasma cells. It is characterised by aberrant proliferation of clonal, long-lived plasma cells in the bone marrow[21]. The large accumulation of MM cells in the bone marrow, crowd out healthy cells. Under normal conditions, plasma cells produce antibodies that fight infection as part of the adaptive immune system. However malignant plasma cells (MM cells) produce large amounts of abnormal antibodies that are unable to fight infection, coined 'paraproteins' or 'M proteins' (Figure 1.4). Only one type of

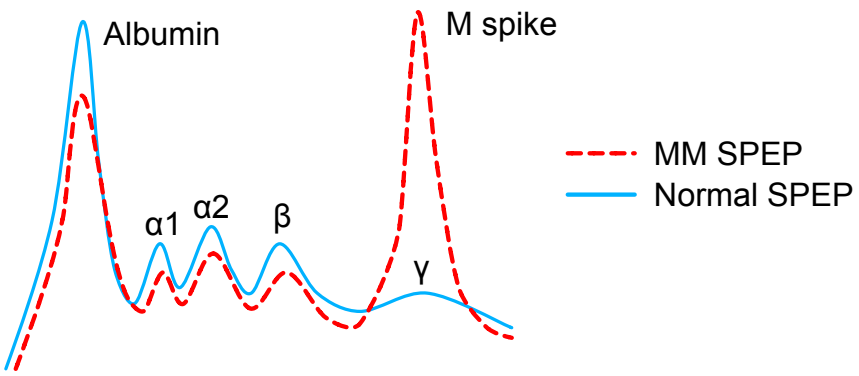


Figure 1.4: Diagram of serum protein electrophoresis (SPEP) for a normal individual and for a multiple myeloma (MM) patient. Based on their electrical charge, SPEP separates all the proteins in the blood. A large peak is recorded for albumin (the most abundant protein in the blood), followed by lower levels of the other proteins, grouped into areas labelled $\alpha 1$ and $\alpha 2$, then a β region, and then a γ region, which represents where antibodies lie on the graph. The large quantities of a single type of antibody (M protein) produced by MM cells cause a distinct 'M spike' in the antibody protein region of the graph (γ region).

Ig is overproduced by a MM patient. The type of Ig overproduced varies patient to patient. Paraproteins can either be whole Igs or just the light chain (κ or λ) part of an Ig (Figure 1.2a). A patient's myeloma type is named after the abnormal Ig isotype they are making. IgG is the most common type, followed by IgA and light-chain-only myeloma. IgM, IgD and IgE are very rare myeloma types[22].

Malignant MM cells vary from healthy plasma cells considerably. The transformation from healthy plasma cell to MM cell is a complex multi-step process, involving numerous genomic, epigenomic, transcriptomic, proteomic and metabolic

changes[23–26]. Examples of such changes include, a rearrangement of the 14q32 locus (where the Ig heavy chain is located)[27] in approximately 60% of patients, as well as rearrangements in c-MYC, cyclin D1, FGFR3 and cyclin D3[23]. Deletions of chromosome 13 and mutations in NRAS and KRAS are also frequently seen in MM. Gene expression changes in MM cells include loss of CD19 expression and aberrant CD56 expression. DNA methylation and histone modifications have also been indicated as crucial regulators of MM pathogenesis. Fluorescence in situ hybridization (FISH) testing is often performed on upon MM diagnosis and at time of relapse, to detect certain genetic alterations and aid in risk stratification of patients[28].

1.3.2 MM microenvironment

MM cells typically grow within the bone marrow (BM). This is often referred to as their BM niche or BM microenvironment. The surrounding microenvironment plays a large role in supporting the progression of MM. The BM microenvironment is comprised of a cellular compartment (immune cells, endothelial cells, osteoblasts, osteoclasts and stromal cells) and a non-cellular compartment (the extracellular matrix (ECM), cytokines, chemokines and growth factors)[29, 30]. MM cells interact with their surrounding environment. These interactions influence the migration, differentiation, survival, proliferation, and response to therapies of MM cells.

1.3.3 Epidemiology

Multiple myeloma accounts for 1-2% of all cancers and has the second highest incidence of hematological malignancies, after non-Hodgkin's lymphoma[1]. MM is rare in individuals under the age of 40, with a median age of 70 at time of diagnosis[31, 32]. MM is more prevalent in males than females and is around twice as common in black populations than in Caucasian or Asian populations[33]. The average incidence rate is approximately 1-6 cases per 100,000 individuals[31, 32, 34], with the highest age-standardised incidence rates in the regions of Australasia, North America, and

Western Europe[35]. Five-year survival rate of MM patients is approximately 49%, whilst approximately a third of MM patients survive ten years or greater[36, 37].

1.3.4 Presentation

Precursor states

All cases of MM are preceded by asymptomatic precursor states, monoclonal gammopathy of unknown significance (MGUS) and smoldering multiple myeloma (SMM). However, only some patients with SMM or MGUS progress to active MM.

MGUS is a pre-malignant condition where patients have the presence of monoclonal immunoglobulins in their blood or urine, <10% clonal plasma cells in their bone marrow, but lack any myeloma-related end-organ damage[38]. Patients with SMM have between 10 and 60% clonal plasma cells in their bone marrow, serum monoclonal immunoglobulin of ≥3 g/dL, and like MGUS, have no signs of end-organ damage[39]. Progression risk of MGUS into symptomatic MM is about 1% per year, whilst progression risk of SMM to MM is higher, at around 10% per year for the first 5 years, after which it decreases[40, 41].

Active MM

There are multiple classifications of active MM. The International Myeloma Working Group's definition[42] is as follows: Greater than 10% clonal plasma cells located in the bone marrow and one or more myeloma-defining event or biomarker of malignancy. Myeloma defining events consist of evidence of end-organ damage that can be attributed to the surplus of M protein and clonal plasma cells, namely the CRAB features:

- Hypercalcemia
  - Serum calcium 1 mg/dL higher than the upper limit of normal, or
  - Serum calcium > 11 mg/dL
- Renal insufficiency
  - Creatinine clearance < 40 mL per min, or

  - Serum creatine > 2 mg/dL

- Anemia
  - Hemoglobin value of 20 g/L below the lower limit of normal, or
  - Hemoglobin value < 100 g/L

- Bone lesions
  - One or more osteolytic lesions on skeletal radiography, computerized tomography (CT) or positron emission tomography (PET)-CT scans

Biomarkers of malignancy include greater than or equal to 60% clonal plasma cells in the BM, an involved-to-uninvolved serum free light chain ratio greater than or equal to 100, and more than one focal lesion on an MRI study[42].

It is currently unclear what causes the malignant transformation between precursor states and active MM. However certain factors have been identified as risk factors, including point mutations, chromosomal copy number variations, a large array of up-regulated transcription factors, and numerous immune events. The disease undergoes multiple stages of transformations and the genetic landscape changes considerably over the course of the disease.

1.3.5 Treatment of multiple myeloma

Multiple myeloma may be an incurable disease, however it is treatable. In fact, in the last decade median survival time for newly diagnosed MM patients has almost doubled[3]. Novel therapeutic advances have contributed to this improvement (Table 1.1). Myeloma is usually treated with a combination of drugs, often comprising a corticosteriod, a proteasome inhibitor, and an immunomodulatory drug (IMiD). A common regimen, approved in the USA, European Union and UK for untreated myeloma is the triplet VRd regimen. This consists of the proteasome inhibitor bortezomib (brand name Velcade), the IMiD lenalidomide (brand name Revlimid), and the corticosteroid dexamethasone.

Year	Treatment	Usage	Ref
1958	Melphalan	The alkylating agent melphalan was first used in plasma cell myeloma in 1958.	[43]
1960s	Corticosteroids	Placebo-controlled double-blind trial of prednisone in multiple myeloma. Combinations of prednisone and melphalan showed an increased survival over melphalan alone. Dexamethasone and prednisone have become a cornerstone in the treatment of multiple myeloma.	[44, 45]
1980s	Stem-cell transplantations	Numerous successful allogenic and autologous bone marrow transplantations in patients with multiple myeloma	[46–49]
2003	Proteasome inhibitors	Bortezomib (BTZ), a first-in-class proteasome inhibitor (PI), was first approved by the FDA for use in RRMM. In 2008 it was approved for patients with no prior treatment. Carfilzomib (CFZ) was approved in 2012 for advanced MM and later in 2015 for treatment of relapsed MM. The oral PI, ixazomib, was approved as a combination treatment with lenalidomide and dexamethasone in 2016 for people who have received at least one previous treatment.	[50–52]
2006	IMiDs	The anti-tumour activity of thalidomide was demonstrated in 1999, this led to the development of lenalidomide, the first approved immunomodulatory imide drug (IMiD) for use in multiple myeloma. Currently, thalidomide, lenalidomide and pomalidomide are approved for use in multiple myeloma	[53–55]
2015	Monoclonal antibodies	In 2015, daratumumab (anti-CD38) and elotuzumab (anti-SLAMF7) monoclonal antibodies, were approved for MM treatment.	[56, 57]
2020	Antibody-drug conjugates	In 2020, Belantamab mafodotin-blmf (Blenrep), was approved by the FDA for use in MM patients who have already received 4 other treatments. Blenrep is an anti-BCMA monoclonal antibody linked to a chemotherapy drug.	[58, 59]
2020	Nuclear export inhibitors	In 2020, Selinexor was approved by the FDA for use in combination with BTZ and dexamethasone for MM patients who have had at least one prior therapy.	[60, 61]

Table 1.1: Timeline of treatment options for multiple myeloma. Listed by first usage or FDA approval for MM.

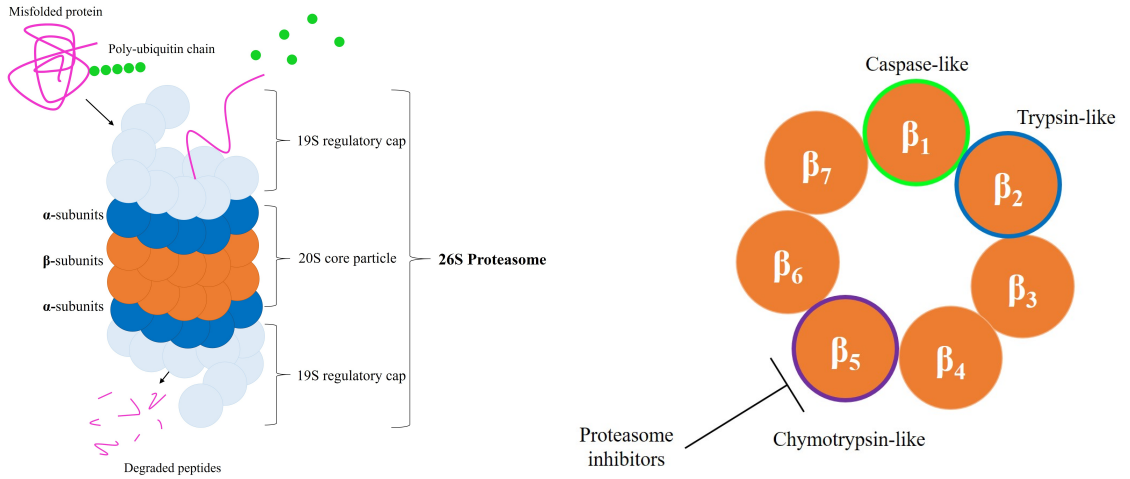
Proteasome inhibitors

Proteasome inhibitors have contributed greatly to the improved prognosis of MM since their introduction into treatment regimes. The first-in-class proteasome inhibitor bortezomib (Velcade®) was approved by the FDA in 2003 as a single-agent for injection of relapsed MM[50]. Since then it has been approved for use in combination therapies. Bortezomib in combination with melphalan-prednisone proved to be superior to the previous standard of care for patients ineligible for HDT-ASCT of melphalan-prednisone alone, increasing time until tumour progression[62]. The combination of bortezomib, dexamethasone and thalidomide was also shown to be superior to previous standard of care for patients prior to ASCT[63]. In 2010, bortezomib was approved as a frontline therapy for treatment-naive MM patients. Since then, two more proteasome inhibitors have been approved, carfilzomib and ixazomib. Carfilzomib is structurally and mechanistically different to bortezomib and shows activity on bortezomib resistant primary MM cells[63]; it is approved for relapsed or refractory MM.

The ubiquitin-proteasome system

Proteasome inhibitors work by blocking the action of the proteasome in the cell. Misfolded proteins can be harmful to a cell, so the combined activity of molecular chaperones, which aid in protein folding, and the ubiquitin-proteasome system (UPS), which acts to digest misfolded proteins, is needed to prevent massive protein aggregation. Unneeded, misfolded or damaged proteins are tagged with lysine-48-linked poly-ubiquitin chains, marking them for degradation by the proteasome (Figure 1.5a). The proteasome is sometimes described as a complex 'protein destruction machine'. The proteasome consists of the 20S core particle, a central hollow cylinder, and the 19S regulatory caps associated with each end of the cylinder. The 19S regulatory caps perform substrate recognition, deubiquitination, unfolding and threading of the protein substrate into the 20S core. The core is made up of four stacked heptameric ring structures. The outer rings are responsible for docking to the 19S cap and for acting as a gate to the inner rings. The inner rings consist

of seven β subunits, containing inward-facing protease active sites for degrading proteins[64, 65] (Figures 1.5a and 1.5b).



(a) 26S proteasome

(b) β-subunits of an inner ring of the 20S core particle

Figure 1.5: Structure of the proteasome. 1.5a shows the structure of the 26S proteasome, comprised of the 19S regulatory caps and 20S core particle. A misfolded protein tagged with a poly-ubiquitin chain is recognised by the 19s regulatory cap, which cleaves the ubiquitins from the protein and threads the protein through to the core, where it is degraded into small peptides. The 20S core particle is made up of two outer rings of α-subunits and two inner rings of β-subunits. 1.5b shows the β-subunit arrangement in one of the inner rings of the 20s particle. β1 (caspase-like), β2 (trypsin-like) and β5 (chymotrypsin-like) are the proteolytically active subunits. Proteasome inhibitors are designed to primarily inhibit β5.

PI Mechanism of action

Of the seven proteasome β subunits, only β1, β3 and β5 are proteolytically active (Figure 1.5b). Proteasome inhibitors are designed to target β5 as it has been shown as the rate limiting protease for proteasomal protein turnover[66]. Bortezomib reversibly co-inhibits β5 and β1 subunits, whilst carfilzomib irreversibly binds to β5, with greater selectivity than bortezomib, and at higher doses binds to β2 as well[66].

The precise downstream effects of β subunit proteasome inhibition are not fully understood, however the unfolded protein response (UPR), NF-κB signalling, JNK signalling, apoptotic factors and p53 are thought to be involved in the anti-MM effects[67]. Specifically, the action of the UPR has been demonstrated as an important mechanism in the anti-MM effect of PIs. MM cells secrete large amounts

of monoclonal protein, leading to the rapid accumulation of misfolded proteins within the endoplasmic-reticulum (ER) lumen. This results in heightened ER stress, which is compensated by the UPR by reducing global protein translation and up-regulating UPS machinery[68]. Therefore, by inhibiting the proteasome, fewer ubiquitin tagged proteins are degraded and more misfolded proteins accumulate in the ER lumen. ER stress is then further increased, causing the UPR to switch from a homeostatic, pro-survival system to a pro-apoptotic pathway[67, 68].

Another important mechanism for PI is the attenuation of NF-κB signalling. IκBα, a specific endogenous inhibitor of the transcription factor NFκB, is a protein degraded by the proteasome. Inhibition of the proteasome increases levels of IκBα, thereby abolishing NFκB signalling. NFκB is a key transcription factor in many cancers, contributing to overall tumour growth and chemoresistance. NFκB has been shown to promote tumour cell proliferation, anti-apoptotic and angiogenic factors[69].

1.3.6 Drug resistance in multiple myeloma

Although newer therapeutics are extremely effective at killing MM cells initially, long-term treatment inevitably results in a drug-resistant relapse. Drug resistance is one of the biggest barriers in the treatment of MM. Patients follow a pattern of peaks and troughs of treatment cycles, remission and relapse, until all therapies have little effect (Figure 1.6).

Minimal residual disease

Minimal residual disease (MRD) is when a small number of cancer cells remain in the body after treatment. These surviving cells are often undetectable and cause no symptoms, but they have the potential to proliferate and cause patients to relapse. MRD status after treatment is used as a prognostic factor and a marker of therapeutic success. Persistent MRD after treatment indicates that MM cells are not eliminated completely and the patient is expected to relapse in the near future[70]. Clinical MRD negativity is defined as no cancer cells being detected in the background of at

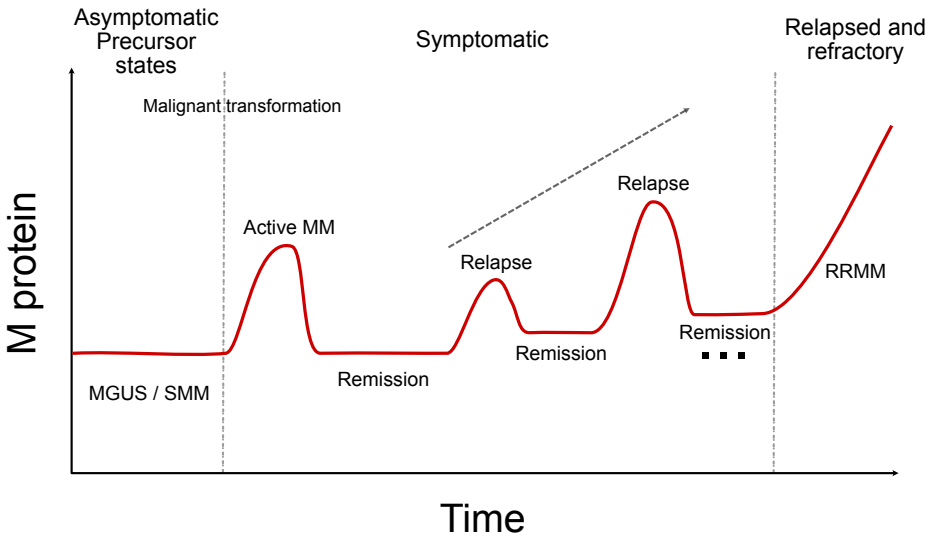


Figure 1.6: MM treatment cycles and disease progression over time. All MM patients begin with precursor states Monoclonal Gammopathy of Unknown Significance (MGUS) and/or smoldering multiple myeloma (SMM) prior to a malignant transformation to symptomatic/active MM. Patients undergo cycles of treatment, remission and relapse until eventually becoming relapsed and refractory (RRMM) and no longer responsive to treatment.

least 10^5 normal cells by multi-parameteric flow cytometry (MFC) or next-generation sequencing (NGS) technology in MM patients who have achieved complete remission (CR)[71]. Many studies have shown MRD negativity is positively associated with prolonged progression-free survival (PFS) and overall-survival (OS) in MM[72–74].

Clonal evolution

MM was originally thought to follow a linear pattern of evolution over time. Whereas MM actually displays clonal evolution and intra-clonal heterogeneity, similar to 'Darwinian' Natural selection: individual clones compete for the same microenvironmental 'niche' and limited resources, with anti-MM therapies acting as environmental selective pressure. Three seminal papers from 2012 elucidated the clonal element of MM disease progression using time-course analyses of primary MM samples[75–77]. These studies revealed substantial heterogeneity within tumor clones from the same patient (intra-patient heterogeneity). Egan et al. (2012) collected DNA from a single MM patient over a five-year period, covering their initial diagnosis, first relapse, second relapse, and end-stage secondary plasma cell

leukemia, and performed whole genome sequencing (WGS) on the samples[75]. The group observed genomic sequence variants that were only detectable at alternating time points, suggesting the 'waxing and waning' of different independent clones over time and treament-courses, which rose and fell in dominance[75]. Walker et al. (2012) further corroborated this finding and demonstrated intra-patient heterogeneity by identifying clonal and subclonal RAS mutations using whole exome sequencing (WES); this was also confirmed at the single-cell level[77]. It has since been shown that intra-patient heterogeneity is present at premalignant phases of MM (MGUS and SMM). On average, three to six major subclones are uniformly present at each stage of MM[78].

Keats et al. (2012) described three distinct patterns for clonal evolution in MM: stable, linear evolution, and heterogeneous clonal mixtures with shifting predominant clones (branching evolution)[76]. For clonally stable myeloma, there are few to no mutations over the disease course; for linear evolution, new genetic aberrations are gained on top of existing mutations; whereas in branching evolution new clones with a new set of genetic aberrations appear over time.

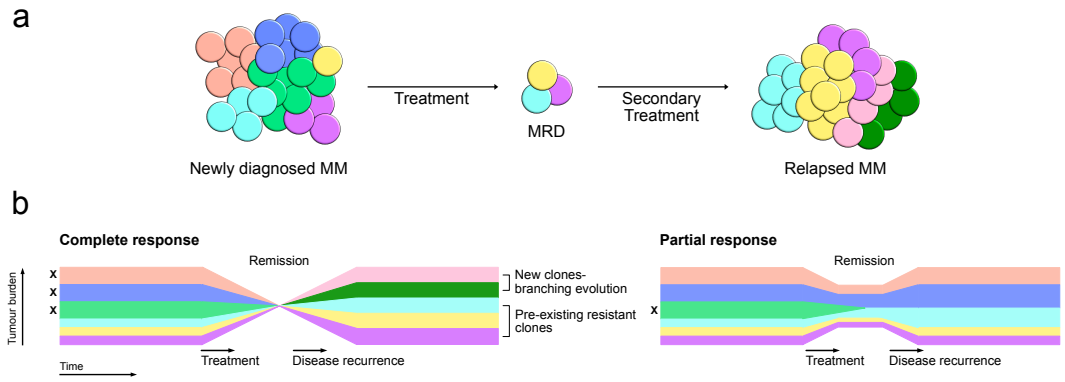


Figure 1.7: Minimal residual disease (MRD) and clonal evolution of MM following treatment. a) Initial treatment of newly-diagnosed MM kills the majority of MM clones (sensitive subclones), resistant subclones survive and proliferate. Pre-existing resistant subclones and new subclones form which are resistant to subsequent treatments, until MM patients are relapsed and refractory. b) Clonal evolution of MM following treatment, leading to a complete response- which gives rise to mainly branching evolution, and a partial response, which gives more stable clonal patterns. Figure adapted from[79].

Drug treatment plays an important role in clonal evolution. Therapeutics place selective pressure on the MM clonal architecture, whereby more sensitive clones are

completely eradicated by drug treatment and only highly-malignant drug-resistant clones remain, outcompeting other clones[78] (Figure 1.7a). These resistant clones are thought to affect the rate of disease progression and cause relapse of MM. Clonal evolution and diversity are causes of acquired drug resistance in MM. Moreover, anti-MM therapy response effectiveness has been shown to affect the pattern of clonal evolution. Jones et al. (2019) demonstrated that patients achieving a CR to treatment followed branching and linear evolution patterns leading to relapse, whilst patients with a partial response to treatment often maintained a more stable subclonal structure[79] (Figure 1.7b).

The clonal heterogeneity seen in MM makes treatment complex. Each individual clone has potentially different clinical behaviour, for example some clones will be more proliferative than others, and associated with an early relapse. Additionally, some clones may have different sensitivity to individual therapeutics to other clones. This highlights the rationale of using a combinatorial approach for treatment. Some agents may be more effective on certain subclones than others, and other agents more effective on other subclones. Therefore, agents are combined to try and achieve a deeper treatment response and reduce residual disease. However, this approach also means that the clones that remain are more likely to be the most resistant clones, capable of surviving a triplet combination of drugs. Combining a higher number of drugs could result in a more complete remission, however the added toxicity of the agents limits this. Clonal dominance should be assessed after each relapse, as the dominance of clones is constantly changing. Agents that were effective in a previous treatment cycle may become effective again, if the resistant clone has become less dominant in the balance of clones[80].

Overcoming drug resistance

In order to overcome resistance and increase overall survival of MM patients, the molecular mechanisms of resistance to existing anti-MM agents need to be better understood. This will aid in the design of novel therapies and inform better use of existing therapies. Previous studies on proteasome drug resistance have

been performed and certain mechanisms and genes have been identified. For example, point mutations have been noted in the *PSMB5* gene (coding for the $\beta 5$ subunit of the proteasome), as well as and over-expression of the $\beta 5$ subunit[81]. Other upregulated genes have been identified, for example *ABCB1*, coding for P-glycoprotein, responsible for pumping various substrates out of the cell, also referred to as multidrug resistant protein 1. *XBP1*, involved in the UPR, has been seen to be downregulated in PI resistance[81]. Although many genes have been identified to be differentially expressed in drug resistant MM, the mechanism is not fully elucidated and further research is imperative in the progression of treatment for multiple myeloma.

Another avenue to improve MM survival, would be to identify novel therapeutics effective against MM, which are capable of overcoming acquired anti-cancer drug resistance. By increasing the arsenal of drugs available to treat MM, the time until patients become refractory and no longer respond to treatment could be prolonged. Therapeutics with a novel mechanism of action are likely to be more effective for MM patients who have undergone several treatment cycles than drugs belonging to the same class as drugs they have previously been treated with. Novel therapeutics would have less overlap of mechanism of action with existing drugs, therefore there would be less cross-resistance.

A future aim for the treatment of MM is to implement a personalised, targeted approach to an individual patient's MM in the clinic. In theory, this would be achieved by characterising a patient's clonal architecture using NGS technologies, and then predicting the most effective treatment strategy for the individual. This could be achieved by targeting a highly conserved mutation (early in branching events) or a mutation of a specifically resistant/malignant clone, or by assessing many other factors. For example, the Horizon 2020 project aims to clinically validate MMPredict, a microarray-test that can genetic subtype MM patients and predict their treatment response based on gene-expression profiling[82]. Personalised patient treatment is quite far away from the clinic in MM, but the prospect is

very exciting. The inter- and intra-patient heterogeneity of MM makes it a very clear candidate for personalised medicine.

1.4 Transcriptomics, proteomics and epigenomics

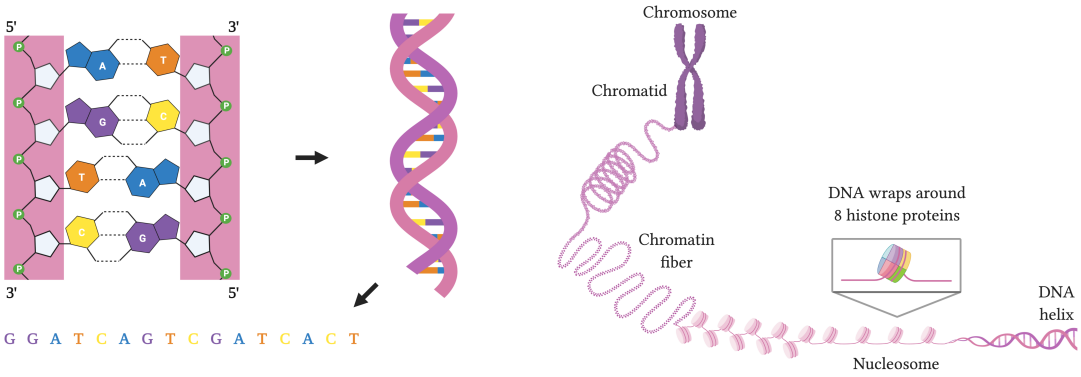
It has been shown that changes in the genome, transcriptome, epigenome and proteome all contribute to disease progression and drug resistance in MM. Therefore, to sufficiently investigate the multiple layers driving MM and to assess the effectiveness of new therapeutics, a multi-omics approach must be employed.

1.4.1 DNA and the genome

The genome is the genetic material of an organism, it consists of deoxyribonucleic acid (DNA). DNA consists of two polynucleotide chains (or strands), running anti-parallel to each other, held together in a double helix structure by hydrogen bonds. Nucleotides are composed of a five-carbon sugar (deoxyribose for DNA), attached to one or more phosphate group (a single phosphate group in the case of DNA) and a nitrogenous base. Nucleotides are covalently linked to form an alternating sugar-phosphate backbone, with bases extending from each sugar towards the inside of the double helix. Nucleotides contain four different types of bases: adenine (A), cytosine (C), guanine (G) and thymine (T). The two DNA chains are held together by hydrogen bonds via complementary base pairing between the bases of the strands, A pairing with T and G pairing with C. Often sections of DNA are denoted as their sequence of A, C, T and Gs (in order reading from the 5' to 3' direction).

Every individual has approximately 6 billion base pairs of DNA per cell, which would amount to about 2 metres of DNA if laid end-to-end. The nucleus of a human cell is approximately 6µm in diameter, therefore chromosomal DNA must be folded tightly to fit. DNA packaging is a complex task involving numerous speciliased proteins. Negatively charged DNA is complexed with an octomer of positively charged proteins called histones to form nucleosomes. The histone core is made up of eight subunits, two copies of H2A, H2B, H3 and H4 subunits. DNA wraps tightly around the histone core 1.65 times. Linker DNA connects adjacent

nucleosomes, to resemble 'beads on a string'. Nucleosomes fold tightly to form 30nm chromatin fibre, which in turns forms loops averaging 300nm in length. This fibre is folded and compressed again to form fiber 250nm in width with loops of 700nm in length. Tight coiling of this fiber forms the single chromatids of chromosomes[83, 84]. Human cells contain 23 pairs of chromosomes.



(a) Nucleotide and DNA double helix structure **(b)** DNA packaging

Figure 1.8: 1.8a shows the DNA nucleotides and the DNA double helix structure. DNA consists of two polynucleotide chains. Nucleotides are covalently linked to one another, forming a sugar-phosphate backbone. They contain one of four bases adenine (A), cytosine (C), guanine (G) and thymine (T). DNA strands are held together by hydrogen bonds between complementary base pairs, A pairing with T and G pairing with C. Sections of DNA are are often read by their sequence of bases from the 5' direction to the 3' direction. 1.8b shows how chromosomal DNA is packaged in the cell. DNA wraps 1.65 times around an octomer of histone proteins, to form a stucture called a nucleosome. Nucleosomes are linked by linker DNA to form a structure that resembles 'beads on a string'. Nucleosomes fold to create chromatin fiber. This is turn forms loops and coils tighter and tighter until it makes up the single chromatids of chromosomes.

The complete genome is made up of coding DNA (genes), non-coding DNA, as well as mitochondrial DNA and ribosomal DNA. An alteration in the nucleotide sequence of the genome is called a mutation. There are a number of types of mutations, including insertions, deletions, inversions, substitutions and duplications. A technique called whole genome sequencing (WGS) can be used to determine the sequence of nucleotides in an individual's DNA and therefore it can be used to determine any variations in the genome.

1.4.2 The epigenome

Epigenetics is the study of any heritable phenotypic changes that do not involve alterations of the DNA sequence itself. These changes occur at the chromatin level. Epigenetic changes include histone modifications, DNA methylation and chromatin remodelling.

DNA methylation is the addition of methyl groups to the C5 position of cytosines in DNA. This happens extensively at CpG sites (cytosine followed by a guanine). Stretches of DNA with a high CpG ratio (CpG islands) are often found in the promoter region of genes. Increased DNA methylation at CpG islands results in transcriptional silencing of those genes. Genome wide DNA methylation is often examined using DNA-methylation-seq or DNA methylation microarrays.

DNA wraps tightly around histones (Section 1.4.1), they contribute to the tight packaging of DNA. Histone modifications are post-translational modifications. They include methylation, acetylation, phosphorylation, ubiquitination and sumoylation. Histone modifications affect transcriptional activity either by directly influencing the structure of chromatin and DNA accessibility or by regulating binding of effector molecules to 'read' histone marks to mediate downstream biological effects. Histone modifications also regulate DNA processes, such as repair, replication and recombination[85]. Chromatin immunoprecipitation sequencing (ChIP-seq) can be used to investigate and measure various post-translational histone modifications.

Chromatin remodelling is the process of modifying chromatin architecture to regulate the accessibility of DNA. Gene expression is regulated by allowing certain gene regions better access to transcription machinery. This is achieved by ATP-dependent chromatin-remodelling complexes moving, ejecting or restructuring nucleosomes. Assay for transposase-accessible chromatin sequencing (ATAC-seq) can be used to identify accessible DNA regions.

1.4.3 The transcriptome

Transcription is the first of many steps in gene expression. During transcription, the enzyme RNA polymerase reads a DNA sequence and produces an anti-parallel,

complementary ribonucleic acid (RNA) strand. The transcriptome is the set of all RNA transcripts of an individual. RNA is a nucleic acid similar to DNA. Like DNA it has a sugar-phosphate backbone and 4 different types of bases attached to each sugar. However, unlike DNA, RNA is single-stranded, it contains the sugar ribose in place of deoxyribose, and the nucleotide uracil (U) inplace of thymine (T).

Despite the chemical differences between DNA and RNA, they are essentially written in the 'same language' and one-to-one mapping of nucleotides can be performed. Transcription begins with the unwinding and opening of a small part of the DNA double helix, so bases are exposed. One strand of DNA acts as a template and the RNA chain is formed by complementary base pairing with the template. RNA polymerases catalyse the reaction of forming phosphodiester bonds between nucleotides, forming the RNA chain. The RNA polymerase moves stepwise along the DNA chain, unwinding the chain just ahead exposing a new region of the template strand. Just behind the region where ribonucleotides are being added, the DNA helix reforms.

The genes in a cell's DNA that specify the amino acid sequence and result in protein synthesis are called messenger RNA (mRNA) molecules. Genes that produce the RNA molecule itself are called non-coding RNAs, because they do not code for proteins. There are many other types of RNA, such as transfer RNA (tRNA), ribosomal RNA (rRNA) and micro RNA (miRNA).

Traditionally microarrays were used to measure gene expression. Now RNA-seq (outlined in section 1.4.6) is more commonly used to study gene expression and the transcriptome. Depending on the library preparation, different types of RNAs can be selected for or excluded, to study different RNA molecules.

1.4.4 The proteome

The proteome is the entire set of proteins that is or can be expressed by an organism. mRNAs are translated into protein molecules. mRNA is made up of only four different nucleotides, but proteins are made up of 20 amino acids, therefore a direct one-to-one function matching nucleotides to amino acids is impossible. Instead,

the sequence of mRNA is read in groups of three consecutive nucleotides, called a codon. The three positions of a codon, and each position with four possible base options (A, C, U and G), gives a total of 64 different permutations (i.e. 4^3). Therefore, some combinations map to the same amino acid (many-to-one function), or signal to terminate translation of the current protein, named a stop codon. This genetic code directs the translation from mRNA to protein. Translation takes place in the ribosome.

Codons on mRNA do not directly recognize their given amino acid, they require tRNA molecules that bind to both the codon on mRNA and the correct amino acid. tRNAs possess an anticodon, a set of three nucleotides complementary to a given codon. Firstly, tRNAs are coupled to their cognate amino acid. This reaction is catalysed by aminoacyl-tRNA synthetase (aaRS) enzymes. aaRSs attach amino acids to the 3' end of tRNA. Most cells have a specific aaRS for each amino acid. Once the tRNA is charged with the correct amino acid, the tRNA molecule binds to its complementary codon on mRNA. Subsequent aminoacylated tRNA molecules bind to mRNA codons. A polypeptide chain grows by stepwise addition of amino acid to the C-terminal end. The formation of the new peptide bond between amino acids releases the tRNA molecule. The peptide chain grows until a stop codon is reached and synthesis of the current protein is complete.

Protein translation is partly regulated by availability of mRNAs, but it also depends on other factors such as RNA silencing and post-transcriptional modifications. Proteins have a large array of functions, such as transporting small molecules, catalysing reactions, cell-cell signalling and providing structural support.

Proteomics is the study of the proteome. CyTOF and LC-MS/MS are techniques often employed to examine the proteome.

1.4.5 Sequencing

DNA is often referred to as the 'genetic master code', therefore the ability to decode the order of nucelic acids has been seen as a highly desirable feat since its discovery. Sequencing is the process of determining the sequence of nucleotides of nucleic acid

residues. Over the last half-century, large numbers of researchers and vast sums of money have been applied to facilitate the techniques and technologies to decode DNA and RNA molecules' nucleotide sequence[86]. Over this time, massive technological innovations have been developed. The most commonly used 'first-generation' DNA-sequencing technology, Sanger sequencing, was developed in 1977[87]. In Sanger sequencing, or the 'chain termination method', chain-termination PCR is performed with a mix of regular nucleotides (deoxynucleotide triphosphates; dNTPs) and fluorescently-labelled, chain-terminating dNTPs (dideoxynucleotide triphosphates; ddNTPs), using the DNA of interest as a template. During the extension step of PCR, when DNA polymerase incorporates a ddNTP randomly, extension ceases. This creates greater than a million copies of the DNA sequence of interest, all terminated at random lengths by the ddNTPs. Capillary gel electrophoresis is then used to separate the extension products by size. The gel is then read to determine the sequence. By reading the gel bands from smallest to largest, the 5' to 3' sequence of the original DNA strand can be inferred. Sanger sequencing was used by the Human Genome Project, an international research effort to determine the DNA sequence of the entire human genome[88]. The whole project took approximately 13 years to complete, and was reported to have cost $3 billion.

Sanger sequencing dominated the sequencing world for over 30 years, until the advent of 'Next-generation sequencing' (NGS; now being dubbed 'second-generation sequencing', with the advent of newer technologies). NGS differs from its predecessors in that it is highly scalable and massively parallel. With NGS you can rapidly sequence the entire human genome in one day, for just under $5000. It is quicker and cheaper than traditional Sanger sequencing, and has progressed data output from the kilobase range up to potentially multiple terabases per run. Today, the largest and most commonly used NGS sequencing technology platform is Illumina, who own around 80% of the global sequencing market. NGS can be used to study the transcriptome, using RNA-seq techniques (Section 1.4.6); the epigenome, using techniques such as ChIP-seq or ATAC-seq; or the genome using techniques such as whole genome sequencing (WGS).

Recently, long-read technologies are becoming more prominent in the sequencing field. Illumina short-read sequencing limits read length to between 50 and 300 base pairs (bp). This read length is too short to detect more than 70% of human genome structural variation. Moreover, some of the most mutation-prone regions of the genome are inaccessible due to repeating or GC-heavy content, limiting sequence coverage and leaving large regions of the genome critically understudied[89]. Long-read technologies are capable of generating continuous sequences ranging from 10 kilobases to several megabases, enabling the sequencing of full-length transcripts. The two main emerging long-read technologies are PacBio single-molecule real-time (SMRT) sequencing[90] and Oxford Nanopore Technologies (ONT) sequencing[91, 92], together marking the birth of 'third-generation sequencing' (TGS). Both PacBio and ONT sequencing sequence single DNA molecules, rather than a pool of PCR-amplified fragments. PacBio sequencing employs a sequencing-by-synthesis strategy (similar to Illumina sequencing), with circular DNA templates to improve accuracy. ONT sequences a native linear single-stranded DNA molecule by measuring current changes as bases are threaded through a nanopore[93]. Whilst TGS technologies have the advantage of longer read length, the methods are also plagued by higher base-calling error rates and lower throughput than NGS technologies[93, 94]. Therefore, this makes applying TGS to single-cell RNA-seq especially challenging. Due to the lower through-put, fewer cells can be reported on at a comparable read-depth to NGS technologies; and due to the high base-calling error rate, inaccurate cellular and molecular barcode assignment can complicate associating mRNA reads with their cell of origin.

1.4.6 RNA-seq

Most modern RNA sequencing (RNA-seq) implements NGS technology to analyse RNA across the transcriptome of a biological sample and allows for the quantification of gene expression.

Bulk RNA-seq

Bulk RNA-seq measures the average expression across a sample. Creating a bulk RNA-seq library involves isolating RNA from a biological sample, filtering for a specific type of RNA (most commonly mRNA), fragmentation of RNA into fragments, reverse transcription of the fragments to generate a complementary DNA (cDNA) library, end repair and adaptor ligation of the cDNA library, followed by PCR amplification ready for sequencing.

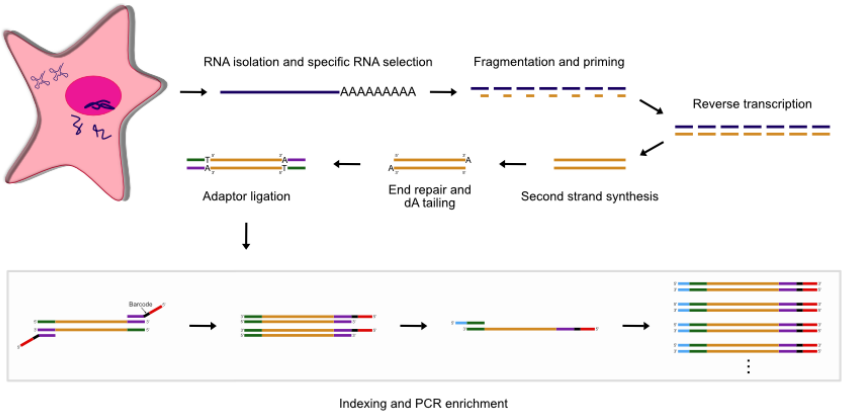


Figure 1.9: Outline of bulk RNA-sequencing library prep. Cells are lysed and RNA is extracted. The specific RNA of interest is selected and enriched, for example selecting for mRNA using polyA selection or ribo-depletion. The mRNA is fragmented into smaller pieces of RNA. First and second stranded cDNA are reverse transcribed from the RNA fragments using random primers. The ends of the cDNA are repaired and dAMP (dA) tails are added to the 3' end of the DNA. Adaptors are ligated to the 3' and 5' end of the cDNA. These adaptors contain complementary sequences that allow the fragments to hybridize to the flow cell during sequencing. Universal (P5/i5) and index (P7/i7) primers are added to the adaptor ligated DNA. The libraries are then amplified using PCR and cleaned-up, ready for sequencing.

Single-cell RNA-seq

Single-cell RNA-seq (scRNA-seq) measures gene expression for each individual cell across a population of cells and therefore provides information on clonal diversity that may be lost when pooling cells into bulk samples. Since its inception in 2009[95], there have been numerous scRNA-seq techniques developed, such as SMART-seq2[96], Drop-seq[97], STRT[98], scCOLOR-seq[94] and inDrops[99]. scRNA-seq library preparation shares many steps with bulk RNA-seq workflow, however preliminary

steps are required to isolate single cells and barcode reads that originated from the cell. For droplet-based scRNA-seq (dscRNA-seq) methods, single cells are isolated

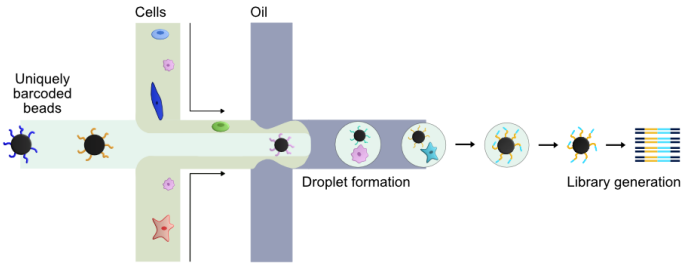


Figure 1.10: Outline of Drop-seq, a droplet-based scRNA-seq method. A microfluidic device combines two aqueous flows, one containing cells and the other containing barcoded primer beads suspended in lysis buffer. The two aqueous channels flow across an oil channel to form aqueous droplets surrounded by oil. Relatively few droplets contain both a cell and a bead. Following droplet formation, the cell is lysed and its mRNAs are released, which then hybridise to the primers on the bead surface. A reagent is added to break up the droplets and the beads are collected and washed. The mRNAs are reverse-transcribed into cDNAs, generating a set of "STAMPS" (single-cell transcriptomes attached to microparticles) and template switching is used to introduce a PCR handle. The barcoded STAMPS can then be amplified using PCR.

using microfluidic devices by individually encapsulating them in aqueous droplets contained in oil. Below, a dscRNA-seq method, Drop-seq, is outlined (Figure 1.10).

1.5 Thesis aims and chapter outline

This thesis aims to identify novel therapeutics with anti-MM properties, capable of overcoming acquired drug resistance in MM.

Chapter 1: General introduction of the adaptive immune system, multiple myeloma (MM) and treatment of MM, as well as an introduction into the multiple layers of information underpinning life: the genome, transcriptome, epigenome and proteome, and the different multi-omic techniques that can be employed to investigate them.

Chapter 2: Literature review introducing aminoacyl-tRNA synthetases (aaRS), the roles they play in disease and therapeutics targeting aaRSs, focusing on the prolyl-tRNA synthetase inhibitor, Halofuginone (HF) and its applications, particularly in multiple myeloma.

Chapter 3: Experimental materials and methodology used in this work.

Chapter 4: Outline of computational methods generated to support experimental work and benchmark validations of their effectiveness.

Chapter 5: Investigation of the use of ProRS inhibitors in PI-sensitive and PI-resistant MM cell lines. Bulk-RNA seq is employed and the transcriptional landscape following ProRS treatment is characterised.

Chapter 6: Exploration of ProRS inhibitor treatment of primary BM samples from MM patients at the single cell level. Their effectiveness against newly-diagnosed and relapsed MM patient tissue is investigated.

2
Literature review: aminoacyl tRNA synthetases and halofuginone

2.1 Introduction

Aminoacyl tRNA synthetases (aaRS) are a highly-conserved family of enzymes, responsible for "charging" tRNAs with their cognate amino acid[100]. Human cytoplasmic aaRSs are either "free" as individual species or bound in a macromolecular complex, comprised of eight aaRSs and three auxiliary proteins, known as the multi-tRNA synthetase complex (MSC)[100]. On top of their canonical enzymatic role, aaRSs also engage in non-enzymatic functions in numerous pathways, including angiogenesis, inflammation and metabolism. Often species are released from the MSC to regulate these non-canonical activities[101]. aaRSs have been shown to involved in numerous diseases, including cancer. Initially, due to their high fidelity and complex evolution over millennia, aaRSs were seen as an attractive drug target for antimicrobials, to enable specifically targeting microbial aaRSs with minimal effects on human cells. The mechanism of action of febrifugine (FF), a quinazoline alkaloid that has long been used as an antimalarial remedy, has recently been revealed; it acts as a competitive inhibitor of ProRS (part of the bifunctional glutamyl-prolyl-tRNA synthetase enzyme; EPRS), responsible for charging $tRNA^{pro}$ with proline. Although it has potent antimalarial effects, FF exhibits high liver and gastrointestinal (GI) toxicity, so cannot be used as a widespread drug, therefore several analogues of FF were developed in the hope of minimizing toxicity to

the host's cells. One such analogue, halofuginone (HF) was synthesized and was shown to have the most potent antimalarial properties of all the derivatives, with lower toxicity to the host than FF, but still some liver toxicity and GI side effects remain[102]. Halofuginone has been applied to and showed promise in many other non-parasitic diseases too. It has received orphan drug status for scleroderma and HIV-Related Kaposi's Sarcoma. Recently, halofuginone's application in various cancers has become of great interest, including but not limited to: metastatic brain tumours, bladder carcinomas, prostate cancer, renal carcinomas, hepatocellular carcinomas, lung cancer, breast cancer and multiple myeloma.

This review will introduce the structure and function of aminoacyl tRNA synthetases, provide an insight into their role in pathology and potential as therapeutic targets. aaRSs inhibitors and their application in disease will be explored, focusing on the usage of the Prolyl tRNA sythetase inhibitor, halofuginone, in multiple myeloma.

2.2 Function and structure of aminoacyl tRNA synthetases

aaRSs are an ancient family of ubiquitous enzymes, conserved across three major domains of life (but not present in viruses). They can be traced back prior to the "Last Universal Common Ancestor" (LUCA)[103]. aaRSs are essential for protein biosynthesis, and catalyse the first step in translation (see Section 1.4.4). aaRSs catalyse the charging of tRNAs with their cognate amino acid. This is a two-step process. Firstly, aaRSs catalyse the formation of an aminoacyl-adenylate (activated amino acid) from their corresponding amino acid and an ATP molecule, releasing an inorganic pyrophosphate. Next, aaRSs catalyse the reaction between the aminoacyl-adenylate and their cognate tRNA to release an AMP molecule and generate an aminoacyl (charged)-tRNA, ready to be used by the ribosome to decode mRNA (see Equation 2.1). An example of this process would be prolyl-tRNA synthetase (abbreviated to ProRS) charging $tRNA^{pro}$ with proline.

$$\begin{aligned} aa + ATP &\rightleftharpoons aa.AMP + PPi \\ aa.AMP + tRNA^{aa} &\longrightarrow aa.tRNA^{aa} + AMP \end{aligned} \tag{2.1}$$

Eukaryotes have 20 cytoplasmic aaRS and 20 nuclear-encoded mitochondrial aaRS. These are localised in distinct cellular compartments. aaRSs are often denoted by their one letter amino acid symbol, followed by ARS and either 1 (indicating they are cytoplasmic) or 2 (indicating they are mitochondrial), for example PARS1 for cytoplasmic ProRS. This review will focus on cytoplasmic aaRS enzymes. aaRS can be divided into two distinct classes based on the structure of the fold of their catalytic domains. Class I aaRS enzymes are functional monomers that contain a dinucleotide or Rossman fold (RF) of alternating alpha-helices and parallel beta-sheets. This fold is where ATP and amino-acid binding takes place and therefore facilitates the aminoacylation reaction. The active site of class I aaRS is marked by the signature motifs “HIGH” (His-Ile-Gly-His) and “KMSKS” (Lys-Met-Ser-Lys-Ser). Within the first half of the RF the HIGH motif helps to correctly position the adenine base of ATP and interacts with the phosphates. The second K of the KMSKS motif is thought to be involved in stabilising the transition state for the primary step of aminoacylation[104]. Amino acid recognition and binding takes place in the catalytic site when the KMSKS motif is open. The KMSKS loop closes after the aaRS binds ATP and the aminoacyl-adenylate is formed[105].

Class II aaRS enzymes are functional dimers or tetramers with an uncommon catalytic core, comprising seven anti-parallel beta-sheets, flanked by alpha-helices. Class II aaRS enzymes are defined by three conserved sequence motifs. Motif 1 is located at the interface of the dimer and enables oligomerization. Motifs 2 and 3 comprise part of the aminoacylation active site and facilitate amino acid/ ATP binding and adenylate formation. Motif 3 binds ATP, and motif 2 is involved in coupling ATP and the amino acid and then transferring the amino acid to the 3‘-tRNA[105]. The distinct active-site structures of class I and II enzymes confer markedly different binding mechanisms for the aminoacylation reaction. For example, class I aaRSs bind the tRNA acceptor stem via the minor groove side and bind ATP in an extended conformation, whilst class II aaRSs bind the tRNA acceptor

stem from the major groove side and bind ATP in a bent conformation. The two classes of aaRSs split the twenty amino acids into two groups. Val, Leu, Ile, Met, Glu, Gln, Trp, Tyr, Arg and Cys are activated by their cognate class I aaRS; and Gly, Pro, Ala, Thr, Ser, Hist, Asp, Asm, Lys and Phe are activated by their cognate class II aaRS. Class I and class II can be further divided into different sub-groups, however that is beyond the scope of this review. The structural diversity of aaRSs is likely attributable to the need to exclude similar non-cognate amino acids and to discriminate the correct tRNA isoacceptor.

Both class I and class II aaRSs are multi-domain proteins- in addition to their catalytic domains, they include other domains such as their anti-codon recognition domain or an editing domain. The editing domain found in some aaRSs is to ensure that the essential step of aminoacylation in protein biosynthesis is as accurate as possible, so incorrect amino acids can be removed from aminoacyl-adenylates or mischarged tRNAs[105]. Theoretically, it was estimated that mistranslation rate should be approximately 1 in 200 for amino acids differing by just a methyl group (such as valine and isoleucine)[106], however in-vivo work demonstrated that the error frequency is closer to 3 in 10,000 (approximately 1 in 3000)[107]. This suggested the existence of proof-reading capabilities of aaRSs, to account for the difference between observed and predicted error rates. Editing capability has since been shown to be of high functional importance to some aaRSs. For example, a study in mice in which there was a missense in the editing domain of AlaRS. The impaired proof reading activity of the enzyme lead to an accumulation of misfolded proteins, resulting in the activation of the unfolded protein response and substantial neurodegeneration[108]. Not all aaRSs possess editing activity, only about half do, however the high specificity of the active site of those aaRSs is enough to alleviate proofreading need.

2.2.1 Multi-tRNA synthetase complexes

Higher eukaryotes contain macromolecular complexes, which consist of nine enzymes and three auxiliary proteins, known as multi-tRNA synthetase complexes (MSC).

The 11 cytoplasmic aaRSs not located in the MSC remain free as individual species. The nine cytoplasmic aaRS enzymes of the MSC are GluRS (EARS1), ProRS (PARS1), IsoRS (IARS1), MetRS (MARS1), GlnRS (QARS1), LysRS (KARS1), ArgRS (RARS1), AspRS (DARS1), LeuRS (LARS1). GluRS and ProRS are covalently fused via triple repeats of WHEP domains to form a bifunctional enzyme, EPRS1. The non-enzyme component of the MSC consists of three aminoacyl-tRNA synthetase-interacting multifunctional proteins (AIMP), AIMP1, AIMP2 and AIMP3. Human MSCs contain more class II aaRS enzymes than other species, namely DARS1, KARS1, and PARS1, they also contain more auxiliary proteins. Human MSC components have several additional domains or motifs (Figure 2.1A), for instance GST-homology domains in EPRS, MetRS, AIMP1 and 2, and WHEP domains in EPRS and MetRS[101, 109, 110].

The structure of human MSC has not been fully elucidated, however some sub MSC-complex structures have been revealed. LysRS forms a homodimer and is anchored to the N-terminal peptide region of AIMP2 within the main body of the MSC. MetRS, AIMP3, EPRS1 and AIMP2 are compactly linked through their GST-homology domains. ArgRS, GlnRS and AIMP1 assemble into a heterotrimeric complex[101, 109, 110]. A proposed bisymmetrical model of the human MSC, via homodimerization of AspRS and ProRS, is shown in Figure 2.1B, based on subcomplex and interaction data[111–113]. This hypothesis proposes that the MSC is a super-complex of two identical, symmetrically arranged subunits (symmetrical along the y-axis in Figure 2.1B), each containing one copy of the constituent elements, except for LysRS which is present as a dimer in each subunit.

The function of the MSC was originally thought to be to increase efficiency of protein biosynthesis by localising aaRSs. Another proposed function of the MSC was to increase stability of its components. It has been shown using systematic depletion analysis that some of the components are in fact intrinsically less stable in isolation and dependent on their neighbours for stability[114]. More recently, examples have emerged where the MSC seems to work as a ‘molecular reservoir’ which can control the release of its components. The release of components from the

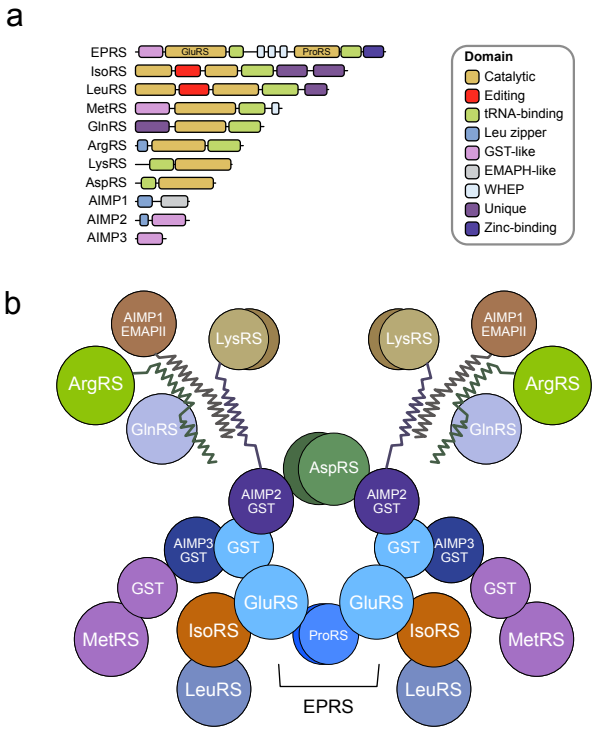


Figure 2.1: The human multi-tRNA synthetase (MSC) and its components. **A)** The domains of the aminoacyl tRNA synthetases and auxiliary proteins (AIMP1, 2 and 3) making up the human multi-tRNA synthetase (MSC). The bifunctional enzyme EPRS1 is made up of the class 1 enzyme GluRS and class II enzyme ProRS (dimer) covalently linked by three WHEP domains. **B)** Cartoon representation of a proposed bisymmetrical model structure of the human multi-tRNA synthetase complex (MSC). An adaption of a figure created by Myung Hee Kim and Sunghoon Kim[110].

MSC has been linked to numerous non-canonical pathways, including cell signalling, metabolism, inflammation and angiogenesis.

Higher eukaryotes usually have extra-domains at the N- or C- terminus of aaRS enzymes compared with lower eukaryotes and prokaryotes, which may partly contribute to MSC assembly. Most human cytoplasmic aaRS enzymes have at least one new sequence extension or domain, most of which are dispensable for enzymatic activity, suggesting they may contribute to the non-canonical roles of aaRS. Additionally, aaRSs are often found in the nucleus of cells, where protein biosynthesis does not occur. The additional evolutionary complexity in human aaRSs and MSC seems to explain the increased physiological complexity and their functionality in non-enzymatic processes.

Examples of non-canonical MSC functionality include— LARS1 translocating from the MSC to lysosomes, facilitating mTORC1 activation[115]; KARS1 translocating to the nucleus upon immune activation and activating MITF-dependent gene expression in mast cells[116]. Another example is EPRS1 release from the MSC in myeloid cells upon IFN-γ stimulation[117]. IFN-γ induces a network of kinase events (Cdk5, mTORC1 and S6K1 activation) which causes a two-step phosphorylation of two serines in the linker region of human EPRS, and causes its release from the MSC. EPRS1 combines with other proteins (namely NSAP1, L13a and GAPDH) to form the cytosolic IFN-γ activated inhibitor of translation (GAIT) complex, which represses translation of numerous inflammatory-related transcripts, including VEGFA and ceruloplasmin[118].

In addition to the enzymatic components of the MSC, the auxiliary proteins AIMP1, 2 and 3 are also involved in fundamental biological processes. AIMPs exhibit non-canonical functions aside from their roles as scaffolds in the MSC. AIMPs have been linked to numerous biological processes, including involvement in immune regulation, nervous system functions, viral replication, genome stability, angiogenesis, and cancer. AIMP1 interacts with RARS1 and facilitates incoming tRNA substrates to its catalytic site to enhance its enzymatic activity[119]. In addition to improving amino-acyl synthetase activity, secreted AIMP1 has also been shown to be involved in angiogenesis, inflammation induction, wound closure, and maintaining glucose homeostasis[120]. TGF-β and the DNA damage response have both been shown to cause phosphorylation of AIMP2 and disassociation from the MSC. Released AIMP2 has been shown to act as a pro-apoptotic mediator and tumorigenesis suppressor via various pathways[121]. AIMP3 largely interacts with MARS1, and under conditions initiating the DNA damage response, MARS1 undergoes a conformational change that releases AIMP3 from the MSC[122]. Released AIMP3 acts as a tumour suppressor, translocating to the nucleus and upregulating expression of the tumour suppressor gene p53.

The functional and structural complexity of the MSC is still being revealed. The canonical and non-canonical functionality of MSC components promises an

unexplored rich source of potential therapeutic targets, but also lends itself to associated pathology.

2.3 aaRSs in disease

With the diversity of functionality in human aaRSs comes an increase in functionality that can be associated pathologically with human disease. Structural and functional variations in aaRSSs' enzymatic and non-enzymatic activities have been linked to various human diseases. Changes in gene expression, copy number, mutations and genetic variations of aaRSs have been documented in relation to disease[105].

Charcot Marie Tooth (CMT) is a genetically and clinically-presenting heterogeneous group of hereditary peripheral neuropathies. CMT is characterised by progressive degeneration of distal sensory and motor neuron function[123]. Six aaRSs have been linked to CMT through dominant mono-allelic mutations, including GARS1 and YARS1, which are among numerous genetic-loci to have been linked causally to CMT. Drosophila models of CMT have demonstrated that CMT-causing YARS1 mutations lead to a conformational change in YARS1, leading to aberrant interactions with transcriptional regulators in the cell nucleus and aberrant expression of certain transcription factors[124].

Self-targeting of aaRSs as autoantigens has been implicated in autoimmune diseases. "Anti-synthetase syndrome" (ASS) is a heterogeneous group of autoimmune diseases, including interstitial lung disease (ILD), arthritis, idiopathic inflammatory myopathies, myositis and Reynaud's phenomenon[125]. Autoimmune antibodies against histidyl-, threonyl-, alanyl-, isoleucyl-, phenylalanyl-, glycyl-, tyrosyl-, asparaginyl-tRNA synthetase have been found in approximately 30% of all autoimmune patients[125]. Dysregulation of aaRS has also been noted in other autoimmune diseases, for example multiple sclerosis and immune thrombocytopenia[126].

aaRSs have been linked to viral and bacterial infection. For example, it has been shown that viral infection leads to the phosphorylation of EPRS and dissociation from the MSC, ultimately blocking PCBP2-mediated mitochondrial antiviral signalling (MAVS) ubiquitination and inhibiting viral replication[127]. Additionally,

HIV-1 infection leads to KRS release from the MSC, which is partially transported to the nucleus. Blocking this release reduced the infectivity of progeny virions, implying that HIV-1 utilizes a dynamic MSC for enhanced viral replication[128]. WARS1 was shown to be increased approximately 27-fold in sepsis patients with a bacterial infection compared with healthy controls. Following a range of infections by various pathogens, host monocytes were shown to rapidly secret WRS. The secreted WRS increased cell surface levels of CD40, CD80 and CD86, markers of macrophage activation[129].

2.3.1 aaRSs in cancer

A growing number of studies have implicated aaRSs and MSC components in tumorigenesis. Firstly, aaRS enzymatic activity is essential to sustain tumour growth. In cancer metabolism, biosynthesis of aminoacyl-tRNAs has been shown to be highly up-regulated[130]. In cancer, we see often see dramatic rapid cell growth, this demands an intense increase in overall protein synthesis. To keep up with this demand, the canonical aminoacylation role of aaRSs is crucial as the first step in protein synthesis.

On top of the enzymatic role of aaRSs, their non-canonical functionality has also been associated with both promoting and inhibiting cancer. The hallmarks of cancer- enhanced growth signalling and proliferation, vascularization, metastasis, altered metabolism, and immune/tumour microenvironment invasion, all have links to tRNA synthetase function. Cancer cells require enhanced growth signalling and proliferation to maintain their rapid growth beyond the capacity of normal cells, several aaRSs have been linked to this aberrant growth signalling. GlyRS has been shown to be integral for cancer-promoting neddylation (where ubiquitin-like protein NEDD8 is conjugated to its target proteins) to occur, and reduced MetRS expression resulted in reduced tumorigenicity in p16INK4a-negative breast cancer cells in vivo[131–133]. For tumours to grow and metastasize they need to hijack existing vasculature to get blood flow to growing area, or make new vessels by promoting angiogenesis. Endothelial cells (EC) exposed to TNF-α or VEGF secrete

ThrRS. ThrRS promotes EC migration and angiogenesis. Inhibition of ThrRS was shown to inhibit angiogenesis, with and without inducing the uncharged tRNA response[134, 135]. LysRS has been shown to support metastasis by increasing migration. Following phosphorylation by the MAPK pathway, LysRS binds to the 67kDa membrane bound laminin receptor protein (67LR), preventing its degradation and sustaining laminin-dependent migration. Once bound to LysRS, 67LR also binds integrin $\alpha 6\beta 1$, which initiates ERK and paxillin signalling, increasing migration by altering cell-cell and cell-ECM adhesion.

aaRSs have also been linked to altering metabolism in cancer. To make rampant growth feasible, cancer cells adjust metabolism to meet energy demands and provide building blocks for biosynthesis. LeuRS activates the mTORC1 pathway, which controls translation and autophagy. Cancer cells utilize the mTORC1 pathway to proliferate more efficiently. The mTORC1 pathway also causes phosphorylation of EPRS and the release of it from the MSC. In adipocytes, released EPRS interacts with FATP1 and directs it to the plasma membrane. Inhibition of FATP1 leads to increased cell viability in breast cancer cell lines, and its expression correlates with decreased patient survival in triple negative breast cancer[136].

2.3.2 AIMPs in cancer

As well as the association between aaRSs and cancer, AIMPs have also been shown to play a role in signalling pathways relevant to numerous cancers. The MSC-bound aaRSs seem to predominantly promote tumorigenic functions when released from the MSC. In contrast, the AIMPs bound with them seem to have more tumour-suppressive effects. AIMP2 has been shown to be a potent tumour suppressor, working via key regualtors in the p53, c-Myc, Wnt, TGF-β and TNF-α signalling pathways. Loss of a single allele of AIMP2 in mice resulted in a far higher susceptibility to tumour formation[137]. AIMP1 has also demonstrated tumour-suppressive effects. In mouse xenograft models, administered AIMP1 was found to reduce tumour volume[138, 139]. AIMP1 has been shown to induce apoptosis of endothelial cells, such that it supresses tumour vascularization[140]; it also

stimulates anti-tumour immune responses, for example activating natural killer (NK) cells via macrophages, dramatically reducing lung metastasis of melanoma cells[141]. AIMP3 activates the tumour-suppressor gene p53 following DNA damage or oncogenic stress. Loss of an AIMP3 allele results in higher susceptibility to spontaneous tumour formation[142].

2.4 aaRSs as therapeutic targets

aaRSs are considered very attractive drug-targets. Initially the interest in aaRSs as therapeutic targets arose with the detection of differences between prokaryotic and eukaryotic aaRSs. Thus, enabling specific targeting of microbial aaRSs with minimal effect on the homologous human aaRSs, making aaRS inhibitors attractive anti-microbial candidates.

2.4.1 Antibacterials and antifungals

In the 1990s, mupirocin (brand name Bactroban) was approved as an antibiotic for the topical treatment of bacterial skin infections. Mupirocin selectively inhibits bacterial IleRS, by simultaneously occupying the isoleucine and AMP binding sites and inhibiting aminoacylation[143]. Mupirocin has shown high selectivity for bacterial IleRS over mammalian IleRS (greater than 8000 fold)[144]. This conferred selectivity seems to be due to only a two-amino acid residue difference in the active site of eukaryotic and prokaryotic IleRS[145]. Another example is Kerydin (Tavaborole or AN2690), an anti-fungal used to treat onychomycosis (a fungal infection of the nail)[146]. Kerydin targets the editing site of fungal LeuRS. Kerydin contains boron (Benzoxaborole)[146]. The boron atom of Kerydin binds to the terminal $tRNA^{leu}$ ribose, trapping $tRNA^{leu}$ in the editing site and causes a non-productive enzyme conformation, which inhibits protein biosynthesis.

2.4.2 Anti-parasitics

On top of the success of the druggability of aaRS enzymes for bacterial and fungal infections, aaRSs have also showed promise as an anti-parasitic target. Much like

cancer, parasites are extremely reliant on protein synthesis to keep up with rapid cell growth and continuous proliferation, so are likely to be more sensitive to disruptions to aminoacylation. Additionally, the evolutionary distance between parasitic aaRSs and human aaRSs is quite large, in fact several parasites have bacterial-like protein translation pathways, not shared by humans[147]. Numerous aaRSs have previously shown promise as targets for anti-parasitic agents. Several naturally occurring compounds target the AsnRS site of parasites, such as *Brugia malayia*, a nematode which causes Lympathic Filariasis. *Trypanosoma brucei* has also been shown to be susceptible to aaRS inhibition, for example by Benzoxaboroles targeting LeuRS, or by Aminoquinoles and Benzimidazoles targeting MetRs. The parasite *Plasmodium falciparum* has been shown to be affected by numerous aaRS inhibitors, including mupirocin, cladosporin and febrifugine derivatives.

2.4.3 Febrifugine and its derivatives

Dichroa febrifuga has been used for centuries in Chinese medicine as an antimalarial remedy, it is considered one of the 50 fundamental herbs. In 1948, two quinazoline alkaloids, named febrifugine (FF) and Isofebrifugine, were first isolated from the plant Dichroa febrifuga (Figure 2.2)[148], as part of a directive to find new anti-malarials from plant sources. Although febrifugine has excellent anti-parasitic activity, it also has strong liver and gastrointestinal toxicity, limiting its use as a widespread therapeutic. This motivated the generation of febrifugine derivatives with the hope of reducing off-target toxicity. The medical applications of the long-used traditional anti-parasitic agent febrifugine and its derivatives have recently attracted much attention. Febrifugine derivatives have been used to treat malaria, fibrosis, inflammatory disease and cancer.

2.5 Halofuginone

One such analogue, a synthetic racemic halogenated derivative of febrifugine, halofuginone (HF; Figure 2.2), was synthesized in 1967 by American Cyanamid Company[102]. Halofuginone was found to have the most potent anti-malarial

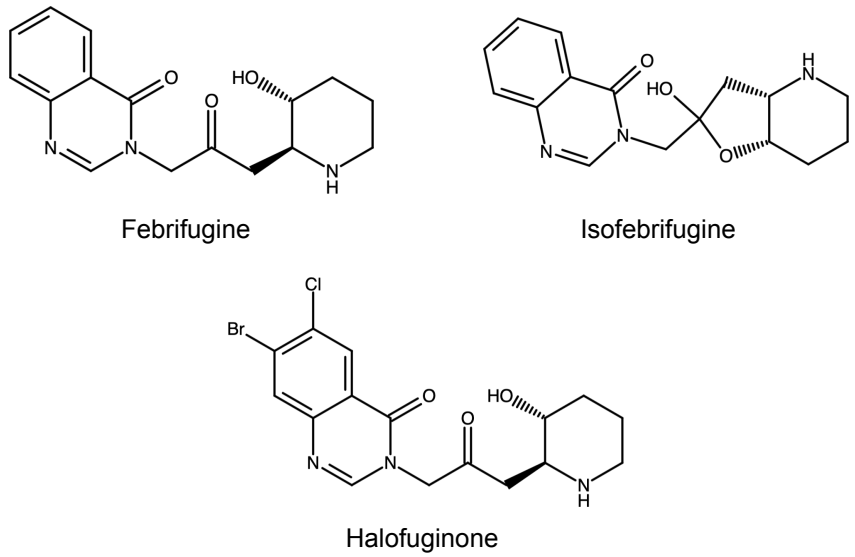


Figure 2.2: Chemical structures of prolyl-tRNA synthetase (ProRS/PARS1) inhibitors. Febrifugine and isofebrifugine were first isolated from Dichroa febrifuga in 1948. Halofuginone is a derivative of febrifugine, first synthesized in 1967.

activity of the FF analogues in vitro and affected all three stages of *P. falciparum* (ring stages, trophozoites and schizonts) with equal speed, unlike many other chemicals with antimalarial effects. The addition of bromine on the quinazoline ring in HF was found not to affect its antimalarial properties, whilst lowering the cytotoxicity for host cells compared to FF. However, HF does still demonstrate some toxicity to the liver, among other side effects, including diarrhoea and vomiting[149]. In an attempt to reduce the side effects of HF and increase the therapeutic window, trans-enantiomers (2R,3S / +) and (2S,3R/ -) of HF have been prepared. Although (-)-HF was found to have lower toxicity than its optical antipode, it was also found to be less efficacious than (+)-HF[150, 151]. This suggests that the biological activity and mammalian toxicity of HF reside with the same enantiomer, therefore there is no advantage to using a specific enantiomer over the racemic mixture.

Recently halofuginone has been researched extensively in association with its applications to non-parasitic diseases. HF is FDA-approved as a feed additive for poultry to prevent coccidiosis from the protozoa *coccidian*. HF has also received orphan drug status for scleroderma and Duchenne muscular dystrophy (in which

fibrosis is the main complication). HF has undergone clinical trials as a potential therapeutic in a number of conditions, including cancer[152, 153].

2.5.1 Halofuginone's antifibrotic properties

Fibrosis (or fibrotic scarring) is a pathological feature of most chronic inflammatory diseases, which can be induced by a variety of stimuli[154]. Fibrosis is defined by the accumulation of excess extracellular matrix (ECM) components, especially collagen type I. If highly progressive, fibrosis can eventually lead to organ malfunction and death[154]. ECM turnover is altered in most pathological states associated with fibrosis. Transforming growth factor β (TGFβ), the tissue inhibitor of metalloproteinases (TIMPs), and matrix metalloproteinases (MMPs) play an essential role in the regulation of the ECM turnover. Inflammatory cells are the main source of TGFβ, which induces collagens gene expression and is one of the leading candidates thought to elicit overproduction of ECM proteins in various fibrotic conditions.

Targeting components of the ECM has proved challenging, limiting the success of fibrosis treatment. HF has been found to have antifibrotic properties and the capability to elicit resolution of established, pre-existing fibrosis, not only act preemptively[155]. HF has been shown to reduce collagen synthesis, a hallmark of the disease[156]. HF is thought to regulate downstream effectors of the TGFβ signalling pathway by inhibiting Smad3 phosphorylation, which in turn causes a reduction in fibroblast differentiation and levels of ECM proteins[149].

2.5.2 Halofuginone and the amino acid starvation response

Until the last decade, the mechanism of action of halofuginone was unclear, until two papers authored by the same group in 2009[157] and 2012[158] elucidated HF's target and downstream effects. In the 2009 paper, the group demonstrated using mouse T_H17 cells that HF activates the amino acid response (AAR) pathway. Mouse T_H17 cells were treated with HF or an inactive derivative, MAZ1310, for 3 or 6 hours and microarray analyses were performed. ATF4 target genes were found to be activated by HF expression, including Asns, Chop, eIF4Ebp, Gpt2, as

well as amino acid transport genes, such as Slc6a9 and Slc7a3, both patterns that correspond with activation of the AAR. Using western blots, the group also showed that GCN2 autophosphorylation was activated by HF treatment, further indicating HF activates the AAR pathway. This effect was not limited to T_H17 cells, the AAR pathway was also activated by HF treatment in fibroblasts and epithelial cells[157]. However, this paper did not reveal how HF activated the AAR.

In 2012, the group identified HF's target protein and demonstrated that HF and FF activate the AAR by competing with proline as potent inhibitors of $tRNA^{pro}$ charging activity of EPRS. Rabbit reticulocyte lysate (RRL) was used as an in-vitro translation system. Following supplementation with excess amino acids, only proline was shown to restore translation inhibited by HF in the RRL system. Moreover, HF-derivatives that were shown to be inactive in functional cell-based assays, such as MAZ1320, also lacked activity in the RRL assay. Together, this suggests that HF functionality is linked to blocking proline utilization. To further demonstrate that HF and FF affect proline utilization, the group synthesized DNAs encoding two epitope-tagged polypeptides, one encoding a proline-dipeptide (ProPep), the second encoding a proline-free peptide (NoProPep). HF and FF treatment prevented translation of ProPep, but had no effect on NoProPep translation[158].

Next the group investigated the effect of HF on prolyl-tRNA charging and the bifunctional enzyme EPRS1 (comprised of GluRS and ProRS fused together). The addition of EPRS from purified-rat-liver reduced the sensitivity of RRL to HF. They then investigated the inverse using siRNA-mediated knockdown to reduce EPRS levels in lung fibroblasts. Lung fibroblasts have high levels of EPRS endogenously, so are quite resistant to HF treatment. The reduction of EPRS levels sensitized the cells to HF treatment and AAR pathway activation— GCN2 autophosphorylation was induced as well as ATF4 response genes, such as CHOP and ASNS. Together this established for the first time that EPRS is a critical target of inhibition for HF and FF, through which the compounds elicit AAR activation. The group demonstrated that HF inhibits EPRS in a competitive fashion with proline at the prolyl-tRNA synthetase active site. HF binding is an ATP-dependent process. ATP

directly locks onto and positions HF onto human ProRS so that one part of HF mimics bound proline and the other mimics the 3' end of bound $tRNA^{pro}$[159]. Excess proline addition was shown to abrogate AAR activation and reversed the biological effects of HF[158].

By binding the active site of ProRS, HF blocks proline from binding and inhibits ProRS enzymatic activity. This results in an intracellular build-up of unaminoacylated (uncharged) $tRNA^{pro}$s, mimicking the cellular state of proline deficiency, thus triggering the amino acid starvation response. Uncharged tRNAs bind to the protein kinase GCN2 and stimulates its dimerization and autophosphorylation. Activated GCN2 phosphorylates eukaryotic translation initiation factor 2A (eIF2α), this leads to a reduction in most protein synthesis, whilst increasing translation of ATF4. ATF4 is a transcription factor of the cAMP response element binding protein (CREB) and induces the expression of many genes involved in the integrated stress response (for example DDIT3/CHOP), amino acid synthetases and transporters, aminoacyl tRNA synthetases, and autophagy regulators (figure 2.3)[157, 160].

Proline is abundantly incorporated into collagen– together with hydroxyproline, it constitutes more than 25% of residues in collagen, which is the predominant protein (80%) in the ECM[161]. Both hydroxyproline and proline are essential for collagen biosynthesis, structure, and strength[162]. So perhaps the anti-fibrotic properties of Halofuginone could also be in part because of the proline-richness of collagen, which inhibition of the canonical function of $tRNA^{pro}$ charging interferes with.

2.5.3 Halofuginone and cancer

HF has exhibited anti-cancer effects in numerous studies and different cancers, including metastatic brain tumours, bladder carcinomas, prostate cancer, renal carcinomas, pheochromocytomas, hepatocellular carcinomas, esophageal squamous carcinomas, lung cancer and breast cancer[163–171]. HF has been shown to exert anti-cancer effects in numerous manners, including reducing tumour growth, reducing angiogenesis, activating autophagy and apoptosis, and disrupting the collagen network of tumours, among other mechanisms.

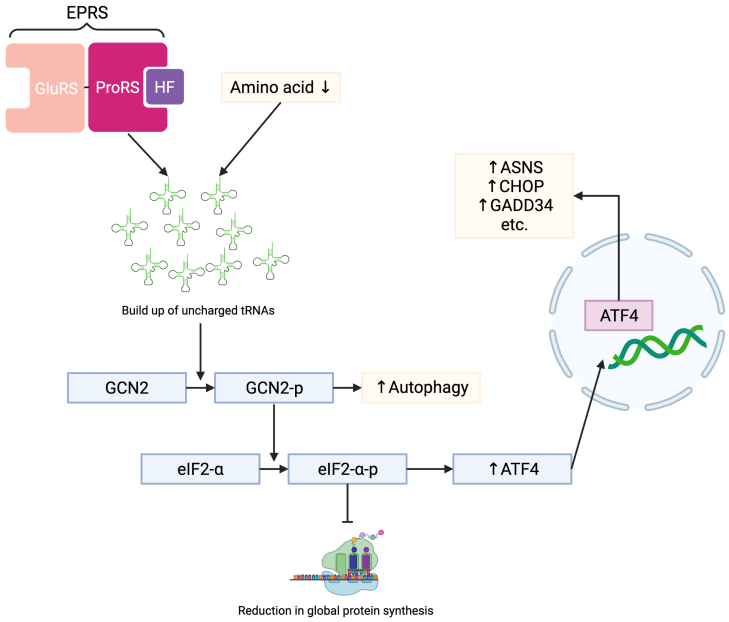


Figure 2.3: A diagram of halofuginone's (HF) mechanism of action and relationship with the amino acid starvation response (AAR). HF binding with the catalytic site of prolyl-tRNA synthetase (ProRS) of the bifunctional aminoacyl-tRNA synthetase, EPRS, causes an accumulation of uncharged tRNAs, mimicking the same cellular environment as if the cell were amino acid deprived. Uncharged tRNAs bind to the cellular sensor GCN2 and cause it to autophosphorylate and activate. Activated GCN2 then phosphorylates eIF2-α. eIF2-α-p reduces global protein synthesis, except for mRNAs containing an upstream ORF cluster in their 5' untranslated region (UTR) which are efficiently translated upon eIF2-*alpha* phosphorylation[160], including the transcription factor ATF4. Upregulated ATF4 results in increased expression of many genes involved in stress responses (e.g. CHOP/DDIT3), amino acid metabolism, amino acid synthetases (e.g. ASNS) and aminoacyl tRNA synthetases.

Halofuginone and multiple myeloma

As mentioned in Section 1.3, multiple myeloma (MM) is an incurable cancer of plasma cells. Drug resistance is a massive problem in MM, with patients becoming resistant to drugs they've previously been treated with, cycling through treatment and relapse cycles. Therefore, identifying novel therapeutics for the treatment of MM is of critical importance.

Following the success of HF treatment in numerous preclinical cancer studies and the phase II study of HIV-related Kaposi's sarcoma[172], Leiba et al. (2012) investigated the treatment of HF in multiple myeloma, both in-vitro and in-vivo[173]. 17 MM cells lines were treated for 48 hours with a range of HF concentrations.

HF was shown to induce a reduction in cell viability in a dose-dependent manner across all 17 MM cell lines, with an IC_{50} of approximately 100nM in most cell lines. The effect of HF on primary cells was then investigated. CD138+ cells from BM samples from five MM patients and PBMCs from two healthy donors were treated with a range of HF concentrations. A greater dose-dependent reduction in cell viability was seen in the primary MM cells compared to the healthy PBMCs, with an IC_{50} ranging from 101-253nM for the MM cells. Demonstrating that, at this concentration range, HF specifically inhibits the viability of MM cells while having no significant effect on normal cells; this also gave a therapeutic window for HF in MM. Next, the group demonstrated that HF induces apoptosis in MM—HF treatment triggered caspase 3, 8 and 9 activities in MM cell lines in a dose dependent manner; it increased the quantity of apoptotic cells (Annexin V-FITC apoptosis assay); it caused an accumulation of cells in sub G1 phase of the cell cycle, associated with DNA fragmentation; and it elevated expression of the heat shock protein Hsp-90. They also showed that exogenous IL-6 and IGF-1, which are central for MM growth and survival, did not rescue HF-induced cytotoxic effects on MM cell lines, indicating that paracrine MM cell growth and the BM environment are unlikely to reverse the biological effects of HF.

The group also exhibited the anti-MM effects of HF in-vivo, using in a xenograft model of SCID mice injected with MM.1S cells. Once tumours reached sufficient size, mice were treated with either PBS or HF for five days a week for the duration of the experiment. HF treatment was found to inhibit tumour growth and increase overall survival compared to the control mice.

Synergy of HF with existing MM drugs was investigated[173]. Cells were cultured for 48 hours with HF (25, 50 and 100nM) in combination with 5nM Bortezomib, 25uM lenalidomide, 500nM dexamethasone, or 500nM doxorubicin. Cells were cultured for 24 and 72 hours with HF (25, 50 and 100nM) in combination with 10uM melphalan. CalcuSyn software (Biosoft, Ferguson, MO, USA) was used to evaluate synergism. Lenalidomide, dexamethasone and doxorubicin were found to be synergistic or additive with HF in all MM cell lines. HF showed moderate

antagonism in combination with Bortezomib. However, only one concentration value was used for each of the established MM agents, and a small range of concentrations were used for HF treatment. A larger range of concentrations of both HF and the other agents would be required to gain a greater insight into the drugs interactions with one another.

From this study, it is clear HF is effective against MM, and could show promise as a potential line of therapy. However, Leiba et al. (2012) did not show how HF was exerting its effect, or if the AAR was activated. AAR activation results in upregulated levels of the transcription factor ATF4. It would be interesting to explore the transcriptional landscape of MM cells and the tumour microenvironment following HF treatment, to see how AAR activation affects this. The group used MM cell lines, mouse models and primary BM samples from MM patients. The primary BM samples were compared against healthy donors' PBMCs and not MM patients' own non-myeloma cells, so only limited conclusions can be drawn about HF's specificity for MM cells over normal cells. Moreover, MM cells are known to interact substantially with their microenvironment. In this study, the cells were studied in isolation, the effect of HF on the immune microenvironment was not investigated. The paper states that BM samples were taken from five MM patients, however it is not stated what stage of disease progression the patients were in, or if they were a mixed group of patients in various disease stages. Therefore, conclusions cannot be drawn whether HF is equally or preferentially effective against newly-diagnosed and relapsed MM patients.

2.6 Discussion

Considering aaRSs are such a highly conserved and ancient family of enzymes, it is surprising how much about their structure and function is still unknown. Concerted efforts are being made to elucidate the full structure of the MSC, with this knowledge, the full functionality of the MSC might be more clearly understood too. More non-canonical functions of aaRSs are being revealed, and with it, associated pathologies. This also presents unexplored potential for aaRS therapeutics.

Drugs targeting aaRSs have showed effectiveness in a wide range of clinical settings including numerous cancers. aaRS inhibitors are an exciting drug in the application of cancer, as tumours often have such a high protein biosynthesis burden, especially MM with the overproduction of large quantities of paraprotein. Therefore, targeting the first step in protein biosynthesis is highly attractive. Halofuginone, an inhibitor of the ProRS active site of EPRS, has been researched extensively in recent years. HF has shown anti-MM effects against MM cell lines, in-vivo mouse models and primary patient BM samples.

HF has been shown to activate the amino acid response in rabbit reticulocyte lysate and lung fibroblasts. HF's effects are abrogated by excess proline supplementation. HF also demonstrates substantial liver and GI toxicity, so may not be able to be used as a widespread drug in cases without orphan drug status. A ProRS inhibitor which was less toxic, with a wider therapeutic range and whose effects could not be overcome by excess proline would likely be a much more effective anti-cancer agent therapeutically.

The mechanism of action of HF in MM has not been clarified and the transcriptional changes of HF treatment in MM have not yet been described. MM patient cells have only been studied in isolation– MM patients' immune microenvironment following HF treatment has not been investigated. The specificity of HF for MM cells has also not been demonstrated fully. MM patients' transcriptome, epigenome and genome evolves greatly during disease progression, and the changes are not limited to MM cells and plasma cells. Therefore, the effect of HF treatment on MM cells must be compared to the patient's own non-myeloma cells to be able to assess specificity, rather than healthy donor's PBMCs, as they are so different from the PBMCs of MM patients. It would be hugely pertinent to employ single-cell sequencing of BM samples following HF treatment. scRNA-seq could capture transcriptional changes of the MM cells and their surrounding immune microenvironment, and would also allow composition analysis, so that proportional changes of each cell type could be quantified, and HF's specificity for MM cells evaluated.

So far, only BM samples from MM patients of unknown disease progression have been treated with HF. Therefore, no comment can be made on whether HF works on relapsed patients' MM cells. Various MM cell lines have been treated with HF, however the resistant variants were resistant to traditional chemotherapy agents (Doxorubicin, Mitoxantrone and Melphalan) or Dexamethasone. MM is conventionally treated with a three-drug regimen (as discussed in Section 1.3.5), comprising a corticosteroid (e.g. Dexamethasone), a proteasome inhibitor (PI; e.g. Bortezomib or Carfilzomib), and an immunomodulatory drug (IMiD; e.g. Lenalidomide). MM patients eventually accrue resistance to all three drugs in the regimen. HF's anti-MM effects were not demonstrated against either PI-resistant or IMiD-resistant cell lines or relapsed patients. It would be interesting to see if HF's anti-MM effects are maintained in PI-resistant MM cells. Proteasome inhibition leads to an accumulation of misfolded, damaged or unneeded proteins, which activates the unfolded protein response (UPR), which in part contributes to the anti-MM effects of PIs. The UPR and AAR share many joint effectors, such as ATF4 and CHOP, both contributing to ER stress and leading to apoptosis. As there is some overlap with HF's and PIs' mechanisms of action, and that Leiba et al. (2012) reported mild antagonism of HF in combination with the PI bortezomib[173], you may not expect HF to be effective against PI-resistant MM. Many new MM agents are approved for relapsed MM initially, rather than as first-line treatments; therefore, it would be critical for ProRS inhibitors' success in myeloma that their anti-MM effects extend to PI-resistant and relapsed MM. This is crucial question that must be answered.

Another thing that would be of interest to investigate, is the interaction of HF treatment and EPRS's non-canonical functionality. Following IFN-γ stimulation, the EPRS dissociates from the human MSC to participate in the GAIT complex in myeloid cells. The GAIT complex represses translation of inflammatory-related genes, including VEGFA. It would be interesting to see if HF has any effect on this non-canonical function of EPRS, or in fact, if the GAIT complex impacts HF treatment in myeloid cells.

aaRSs are a very exciting area of research, particularly as drug targets. So far, the application of aaRS inhibitors in disease has only scratched the surface of their potential as therapeutics. Much more work is required to fully understand their mechanism of action and breadth of potential applications in disease, particularly MM.

3
Methods

3.1 Cell culture

3.1.1 AMO-1 cells

AMO-1 cells, plasma cells from a 64-year old female myeloma patient, were used as a model cell-line for multiple myeloma. Proteasome inhibitor-sensitive AMO-1 cells are referred to as WT cells. Bortezomib resistant (aBTZ) and carfilzomib resistant cells (aCFZ), believed to be AMO-1 cells were generated and gifted by the Driessen lab[174]. After typing these cells, they were found to be a mix of AMO-1 cells and L363 cells. AMO-1 cells were cultivated in RPMI-1640 medium (Thermofisher, UK), supplemented with 10% fetal bovine serum (FBS) and 2mM L-glutamine (Invitrogen, UK). Cells were passaged when they reached approximately 1.5-2 million cells per ml. AMO-1 cells are suspension cells and were split twice a week to approximately 0.5 million cells per ml. All media was replaced with fresh media every two to three weeks, or prior to performing experiments with the cells.

3.1.2 L363 cells

After typing the cells gifted by the Driessen lab, they were found to be a mix of AMO-1 MM cells and L363 MM cells. In-house PI-resistant cell lines were produced by Dr James Dunford by continual and escalating drug exposure of drug-sensitive (WT) AMO-1 cells. However after these cells were typed, they were found to be L363 cells. This was due to the drug exposure selecting the L363 contaminate population

over the AMO-1 cells, due to their natural increased resistance to PI, compared to AMO-1 cells. Once this mistake made by our collaborators was noticed, WT L363 cells were purchased. WT, aCFZ and aBTZ cells were cultivated in RPMI-1640 medium (Thermofisher, UK), supplemented with 10% fetal bovine serum (FBS) and 2mM L-glutamine (Invitrogen, UK), and kept in 100nM of their respective proteasome inhibitor. Cells were passaged when they reached approximately 1.5-2 million cells per ml. L363 cells are suspension cells and were split twice a week to approximately 0.5 million cells per ml. All media was replaced with fresh media every two to three weeks, or prior to performing experiments with the cells.

3.2 Compounds

3.2.1 Proteasome inhibitors

Stock concentrations of proteasome inhibitors, bortezomib (BTZ) and carfilzomib (CFZ) were purchased from Thermofisher ??? ask jim

3.2.2 ProRS inhibitors

The synthesis of NCP22 (T-3767758) is detailed in [175]. NCP26 was synthesised by collaborators as a novel inhibitor of ProRS enzymes, synthesis detailed in [176] (currently under review). ProRS inhibitors: halofuginone (MAZ1392; HF) and halofuginol (MAZ1805) were purchased from.. ProRS inhibitors NCP26 and NCP22 were kindly donated by Swiss lab... NAME ProSA WHERE OBTAINED??

3.3 Assays

3.3.1 Cell viability assays

10X presto blue was added in a 1:10 ratio to cells in suspension and incubated at 37°C for two to three hours. Plates were read [DETAILS OF MACHINE AND PROTOCOL, e.g. wavelength]

3.3.2 Dose response curves

90µl of cells in fresh media were seeded into 96-well plates a day prior to treatment with compound. A total of 20,000 cells were seeded into each well. No cells were placed in edge wells, to avoid edge effects. The following day, media 0% viability controls were placed in the first and last row. Drug concentrations were made up 1000x the desired final concentration. Drugs were diluted once in media (usually 1 in 100), then into the final plate with seeded cells (usually in 10), depending on the experiment. All drug concentrations/combinations were performed in triplicate. Cells were treated with DMSO in triplicate as 100% viability controls.

3.4 Bulk RNA-seq

3.4.1 RNA extraction

RNA was isolated and purified using the Direct-Zol RNA MiniPrep kit (Zymo, USA), following the manufacturer's protocol. In brief, for each sample, approximately 100,000 cells were lysed in 300µl of TRIzol and the lysate was transferred to a microcentrifuge tube. 300µl of ethanol was added to the lysed samples and vortexed. The mixture was transferred to miniPrep columns and centrifuged at 13,000g for 30 seconds. The column was washed twice with 400µl of Direct-Zol pre-wash and once with 700µl of RNA wash buffer The column was transferred to an RNase-free tube and eluted with 50µl of nuclease-free water and centrifuged.

The RNA concentration was quantified using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA), and samples were stored at -80°C.

3.4.2 RNA library preparation

Samples were normalised to 100ng and made up to 50µl with nuclease-free water Poly-adenylated RNA was enriched using the NEBNext Poly(A) mRNA magnetic isolation module (NEB, USA), following the manufacturer's protocol. In short, ...

NEBNext® Ultra II directional RNA library prep kit for Illumina® with TruSeq indexes was used to prepare RNA libraries, following the manufacturer's protocol.

3.4.3 Pre-sequencing preparation

The molarities of the libraries were determined by electrophoresis on a TapeStation (Agilent, USA). The samples were then pooled according to their peak molarity. The pooled library was then denatured and diluted ready for sequencing. In short, 10µl of the approximately 2nM library was denatured with 0.2N 10µl NaOH. The mixture was vortexed briefly, centrifuged at 300g for 1 minute, and then incubated at room temperature for 5 minutes. 10µl 200mM Tris-HCl (pH 7.0) was then added, vortexed and spun as above. The denatured library was then diluted to 20pM. 970µl chilled HT1 buffer was added, vortexed and spun. 117µl of the 20pM library was then mixed with 1183µl chilled HT1 buffer to give a 1.8pM library ready for sequencing on the Illumina NextSeq platform.

Sequencing of the resultant libraries was performed on the NextSeq 500 (Illumina, USA) platform using a paired-end run, according to the manufacturer's instructions.

3.5 Single-cell RNA-seq

3.5.1 Drop-Seq

Cell encapsulation

The Drop-Seq protocol[97] was followed for single-cell RNA-seq sample preparation. Cells were loaded into a microfluidics cartridge. Nadia, an automated microfluidics device (Dolomite Bio, UK), performed cell capture, cell lysis and reverse transcription. Reverse transcription reactions were performed using ChemGene beads.

Library preparation

Beads were collected from the device and cDNA amplification was performed. The beads were treated with Exo-I prior to PCR. The amplified, purified cDNA then underwent tagmentation reactions. A TapeStation (Agilent, USA) was used to assess library quality. The samples were pooled together and split across multiple sequencing runs.

3.5.2 10X Chromium V3

Bone marrow samples were collected from two newly diagnosed multiple myeloma patients and two relapsed multiple myeloma patients; anonymised human tissue samples used in this project were obtained with informed consent by the Haem-Bio Tissue Bank (REC reference: 17/SC/0572). After Ficoll gradient separation, mononuclear bone marrow cells were diluted to 500,000 cells/ml in RPMI media supplemented with 2mM L-glutamine and 10% FBS and 1ml was added to 15ml polypropylene tubes. Compounds were dissolved in DMSO, and 1µl of compound solution was added to achieve a final concentration of 1µM and incubated for 24 hours. Cells were counted and single-cell RNA-seq library preparation was performed using the Chromium Next GEM Single Cell 3' GEM, Library and Gel Bead Kit v3.1 according to the manufacturer's instructions. Indexed libraries were quantitated by TapeStation, pooled and sequenced on an Illumina NovaSeq 6000 (Novogene, UK).

3.6 QuantM tRNA-seq

Approximately 1 million cells were collected per sample and seeded overnight in six-well plates. Cells were treated with 700nM NCP26, 300nM Halofuginone or 2µM ProSA. Controls were treated with equal volumes of DMSO. Cells were collected at time = 0, time = 3 hours or time = 6 hours. Samples were centrifuged at 300g for 5 minutes, the supernatant was discarded and the pellets were resuspended in 300µl of Trizol. RNA was extracted as above (section 3.4.1) and quantified using a nanodrop. Concentrations ranged from 189.7ng/µl to 398.85ng/µl.

QuantM tRNA-seq as outlined in [177] was used for library preparation with significant adaptations made to the protocol. Each sample was normalised to 500ng total RNA in 2.65µl of nuclease-free water. Samples were deacylated with deacylation buffer (Tris-HCl pH9.0; final concentration 20mM) and incubated at 37°C for 45 minutes. Samples were not demethylated.

3.6.1 Annealing and ligating adapters

The samples were transferred to LoBind PCR plates, and 10pM of 3‘ adaptor and 2.5pM of each 5’ adaptor (A, U, C and G) were added (1µl of mix that is 10µM for 3‘ and 2.5µM for 5‘). The plate was incubated at 95°C for 2 minutes in a thermocycler. 1µl of 5x annealing buffer (table 3.1) was added to each sample and the plate was incubated at 37°C for 15 minutes. Ligataion of the adapters to the tRNA was

Annealing buffer	**Volume (µl)**
1M Tris-HCl (pH 8.0)	250
0.5M EDTA (pH 8.0)	50
1M $MgCl_2$	400
Nuclease-free water	9200
Total	**10,000 (10ml)**

Table 3.1: Annealing buffer recipe

performed. 0.5µl T4 RNA ligase 2 (NEB), 1µl 10x reaction buffer and 3.2µl nuclease-free water was added to each 5.3µl of sample, to total a final volume of 10µl (0.5U/µl). The plate was placed in a thermocycler and incubated at 37°C for an hour, and then 4°C for an hour. The ligated samples were then transferred to 1.5ml eppendorfs.

3.6.2 RNA precipitation

1.5µl GlycoBlue was added to each tube. Each sample was made up to 100µl with nuclease-free water. 10µl of 3M sodium acetate (pH 5.2) and 250µl 100% ethanol was added to each tube and vortexed. Samples were precipitated overnight at -80°C. The following morning, tubes were centrifuged at >12,000g at 4°C for 30 minutes to form a pellet. 2 washes were performed with ice-cold, freshly prepared 75% ethanol, spinning for 10 minutes at 12,000g. All ethanol was removed with an extra 10 second top speed spin, and 10 minutes of air drying with the tube cap off.

3.6.3 Hybridization of RT primer

Samples were resuspended in 10µl nuclease-free water and transferred to a PCR plate. 1µl 10µM RT primer <PRIMER TABLE of sequences ref AT BOTTOM>,

1µl 10µM dNTP mix, and 1µl nuclease free water was added to each sample. The PCR plate was placed in a thermocyler and incubated at 70°C for 2 minutes.

3.6.4 cDNA synthesis

cDNA was synthesised using SuperScript IV Reverse Transcriptase (Invitrogen), following the manufacturer's instructions. 4µl 5x SuperScript IV buffer, 1µl 100mM DTT, 1µl SuperScript IV Reverse Transcriptase and 0.25µl RNase Nxgen inhibitor (Lucigen) was added to each sample (totalling 19.25µl). The plate was heated in a thermocycler at 55°C for an hour. 19.25µl 0.2N NaOH (final concentration 0.1N) was added to each sample, and heated in a thermocycler at 98°C for 20 minutes to hydrolyze RNA. The samples were then transferred to 1.5ml eppendorfs. Ethanol precipitation was performed overnight (as in section 3.6.2), and nucleic acids were resuspended in 12µl.

3.6.5 Separating cDNA libraries

Two 18-well 10% Criterion TBE-Urea Polyacrylamide Gels (Bio-Rad) were used to separate cDNA libraries, following the manufacturer's instructions. 1X TBE (89mM Tris, 89mM boric acid, 2mM EDTA) was used as running buffer. 5x sample buffer (89mM Tris, 89mM boric acid, 2mM, 12% Ficoll 400, 0.01% bromophenol blue, 0.02% xylene cyanole FF, 7M urea) was made up, and 3µl was added to each sample. 20µl 1xTBE, 2µl PCR marker (N3234; NEB) and 5.5µl 5x sample buffer was mixed and used as a ladder for each gel. cDNA libraries and ladders were pipetted into their corresponding well and the gels were ran for approximately an hour at 90V. Gels were removed and placed back into their plastic tray for staining. Gels were covered in excess 1xTBE and 3µl SYBR gold (Invitrogen) was added to each tray and stained for 15 minutes on a mixing tray. Gels were excised on 300 (????)nM UV light. A clean scalpel was used to cut out gel bands between 75nt and 300nt (the region representing tRNAs). Gel fragments were then placed in labelled eppendorfs, trying to limit contamination. The eppendorfs were placed in the fridge overnight.

A 25-gauge needle was used to pierce holes at the bottom of 500µl eppendorfs. The pierced 500µl eppendorfs were nested inside 1.5ml eppendorfs, and gel pieces were transferred into their corresponding, labelled nested tube. The nested eppendorfs were centrifuged at 18,500g for 5 minutes, until the gel was completely sheared into the bottom tube. The 500µl eppendorfs were discarded. 400µl of DNA extraction buffer (table 3.2) was added to each 1.5ml eppendorf and the tubes were frozen on dry ice for 30 minutes. The tubes were placed on a rotator overnight at room temperature.

DNA extraction buffer	**Volume (µl)**
4M NaCl	1500
1M Tris-HCl (pH 8.0)	200
0.5M EDTA	40
Deionised water	18260
Total	**20000 (20ml)**

Table 3.2: Extraction buffer recipe

UltraFree MC-VV centrifugal filters with a 0.1µM pore were used to remove small gel pieces. The filters were pre-wet with 7µl DNA extraction buffer, and the gel/ extraction buffer slurry was transferred to them. Tubes were spun at 20,000g for 3 minutes. The filter columns were discarded. 1.5µl GlycoBlue and 500µl propan-2-ol was added to each tube and precipitated on dry ice for 30 minutes. The samples were centrifuged at 18,500g at 4°C for 30 minutes, and washed once with 70% ethanol for 5 minutes. Ethanol was removed completely and air dried for 10 minutes. Samples were resuspended in 12µl nuclease-free water and transferred to a PCR plate.

3.6.6 Circularization

The cDNA libraries were circularized using CircLigase II (Lucigen). 1.5µl 10X reaction buffer, 0.75µl 50mM $MnCl_2$ and 0.75µl CircLigase II ssDNA Ligase was added to each sample. The plate was placed in a thermocycler and incubated at 60°C for an hour, followed by 80°C for 20 minutes. The libraries were transferred to eppendorfs and ethanol precipitation was performed (section 3.6.2). Samples were resuspended in 15µl nuclease-free water and transferred to a LoBind PCR plate.

Stage	Temperature	Time	Cycles
Initial Denaturation	98°C	30 seconds	1
Denaturation	98°C	10 seconds	12
Annealing/extension	65°C	75 seconds	
Final extension	65°C	5 minutes	1
Hold	4°C	Inf	-

Table 3.3: tRNA libraries PCR amplification thermocycling conditions

3.6.7 Amplification

The circularized cDNA libaries were amplified using NEBnext Ultra II Q5 master mix and custom i5 and i7 primers. 5µl 10µM PCR primer, 5µl 10µM specific itRNA primer, and 25µl Q5 master mix was added to each sample (a master mix was made up of PCR primer and Q5 master mix). The PCR plate was placed in a thermocycler and amplified following the thermocycling conditions in table 3.3.

3.6.8 Library purification

The amplified libaries were purified using 2% agarose gels stained with SYBR gold. 20µl 10,000x SYBR gold was added to 200ml 2% agarose gel. 5µl 10x bluejuice gel loading buffer (Invitrogen) was added to each sample. 4µl PCR marker, 36µl nuclease-free water and 4µl loading buffer were combined to form the ladders of the gel. The gel was ran at 120V for an hour. Bands corresponding to cDNA libraries (100bp-250bp) were excised on a UV light box with a clean scalpel and transferred to 2ml eppendorfs. Gel extraction was performed using the GeneJET gel extraction kit, following the manufacturer's protocol. In short, approximately 700µl binding buffer was added to each tube and the gel mixtures were incubated at 55°C until dissolved. The solubilized gel solutions were added to purification columns and centrifuged at 12,000g for 1 min and the through-flow discarded. 100µl of binding buffer was added to each column and centrifuged again. Two washes with 700µl of wash buffer were performed, followed by an additional spin and air dry to remove any residual ethanol from the columns. The collection tubes were discarded and the columns were nested in 1.5ml eppendorfs. 50µl of elution buffer was added to each column and centrifuged for 2 minutes. The columns were discarded and the eppendorfs stored at -20°C.

The libraries were quantified on a tape station (Agilent) and pooled according to their peak molarity. The pooled library was then denatured and diluted ready for sequencing (as above in Section 3.4.3).

Sequencing

Sequencing of the resultant libraries was performed on the NextSeq 500 (Illumina, USA) platform using a single-end run, according to the manufacturer's instructions.

3.7 Data Processing

3.7.1 Bulk RNA-seq

Fasta files were processed using a CGAT-flow[178] pipeline, the workflow can be found at: `https://github.com/cgat-developers/cgat-flow/blob/master/cgatpipelines/tools/pipeline_rnaseqdiffexpression.py`. The pseudo-alignment tool, Kallisto[179], was implemented to pseudo-align reads to the reference human genome sequence (GRCH38 (hg38) assembly) and to construct a counts matrix of samples against transcripts. DESeq2[180] was used for differential expression analysis of counts matrices (using negative binomial generalized linear models) within the R statistical framework (v3.5.1). XGR[181], Reactome[182] and KEGG[183] were used to perform pathway analysis, within R. Org.Hs.eg.db[184], AnnotationDbi[185] and biomaRt[186] were used for converting between Ensembl IDs, HGNC symbols and ENTREZ IDs.

3.7.2 Single-cell RNA-seq

The computational pipeline outlined in Section 4.2 was used to process scRNA-seq data. Downstream analysis was performed in Jupyter lab notebooks[187] using R kernels.

4

Computational method development

4.1 Introduction

4.1.1 Reproducible workflows

In data analysis, particularly in bioinformatics, many users often create simple bash or R scripts to execute the specific task at hand. However, if this is done frequently, the user will have an accumulation of these single-use scripts, which are often named uninformatively and never used again. This may mean the user creates numerous scripts which perform the same function. Another example of a bad practice is using the command line alone to perform tasks. This means that exactly how the analysis was performed is not recorded and may be lost or difficult to find later. These are bad practices in terms of efficiency and reproducibility. It is much better practice to create well-documented, generalised workflows which can then be applied to multiple different experiments. This enables the user to reuse their code more easily and reproduce results, if need be. This also allows other researchers to reproduce results or apply the code to their own research.

In addition to creating generalised, reproducible workflows, it can be beneficial to create more extensive computational pipelines for jobs which require multiple tasks or actions to be performed sequentially.

4.1.2 Computational pipelines

A computational pipeline consists of a series of manipulations and transformations, where the output of one element is the input of the next. Often these elements are executed in parallel. Pipelining 'omics' data-processing means that tasks that are not interdependent can be executed simultaneously. Additionally, multiple samples can be processed in parallel, thereby reducing run time. There are many available pipelining frameworks, for example Snakemake[188], Luigi and Ruffus[189].

For this work, a series of computational pipelines and workflows were generated. Ruffus and CGAT-core[190] were used as the backbone for the pipelines developed.

4.2 scRNA-Seq pseudoalignment pipeline

Fewer pipelines exist for single-cell RNA-Seq compared to bulk RNA-Seq. For the Chromium 10X Genomics platform, most of the processing and analysis is automated by Cell Ranger; however for other technologies, the workflow is not as well defined. A single-cell analysis pipeline was constructed with the aim to produce an easy-to-use, robust and reproducible workflow that works for Drop-Seq as well as 10X technology, which utilises pseudoalignment rather than traditional mapping methods.

4.2.1 Psuedoaligment

Traditional mapping techniques such as Tophat[191] or STAR[192], rely on aligning each read to a reference genome. This is generally very time consuming and computationally expensive. Another challenge that arises with traditional mapping is the occurrence of multi-mapping, whereby a read cannot be uniquely aligned as it could map equally well to multiple sites in the genome[193]. More recently, a series of methods called pseudoaligners have been developed that overcome some of the issues associated with traditional mapping approaches. Pseudoalignment (sometimes referred to as quasi-mapping) methods provide a lightweight, alignment-free alternative to traditional mapping. It has been shown that information on where exactly inside transcripts sequencing reads may have originated is not required for

accurate quantification of transcript abundances[194]. Rather, only which transcript the read could have originated from is needed and transcript abundances are calculated by computing the compatibility of reads with different transcripts. This negates the need for alignment to a reference genome, alleviating the issue of multi-mapping and reducing the computational load. Pseudoaligners have been shown to complete data processing of RNA-seq datasets up to 250-times faster than traditional alignment and quantification approaches[179]. Kallisto[179] and Salmon[195] are tools which implement pseudoalignment. They have similar speed and accuracy for bulk RNA-seq data[1].

Pseudoalignment of scRNA-seq

Pseudoalignment tools have recently been developed for droplet-based scRNA-seq analysis (dscRNA-seq). Additional challenges come with dscRNA-seq data processing, having the extra complication of cellular barcodes (CBs) and unique molecular identifiers (UMIs). These tools must handle transcript abundance estimation, as with bulk RNA-seq analysis, but also perform CB detection, collapsing of UMIs (arising from PCR duplication of molecules) and barcode error correction. Kallisto BUS[196] has been developed as an analysis tool and file format specifically for single-cell analysis, alongside BUStools, for processing of the resultant BUS file[197]. Salmon Alevin[198] has also been developed for single-cell RNA-seq analysis.

Pipeline outline

Kallisto BUS or Salmon Alevin performs pseudoalignment and generation of a cell-by-gene expression counts matrix. Quality control is performed using Scater[199] and alevinQC. Clustering is performed using Seurat3[200] and Monocle[201]. Clusters are projected onto tSNE and UMAP plots. Differentially expressed genes are identified by performing non-parametric Wilcoxon tests on $\log_2 TPM$ expression values and Fisher's exact test for comparing expressing cell frequency, these p values combined using Fisher's method. Multiple comparisons are accounted for by performing the Benjamini-Hochberg correction to adjust the false discovery rate.

[1] `https://liorpachter.wordpress.com/2017/09/02/a-rebuttal/`

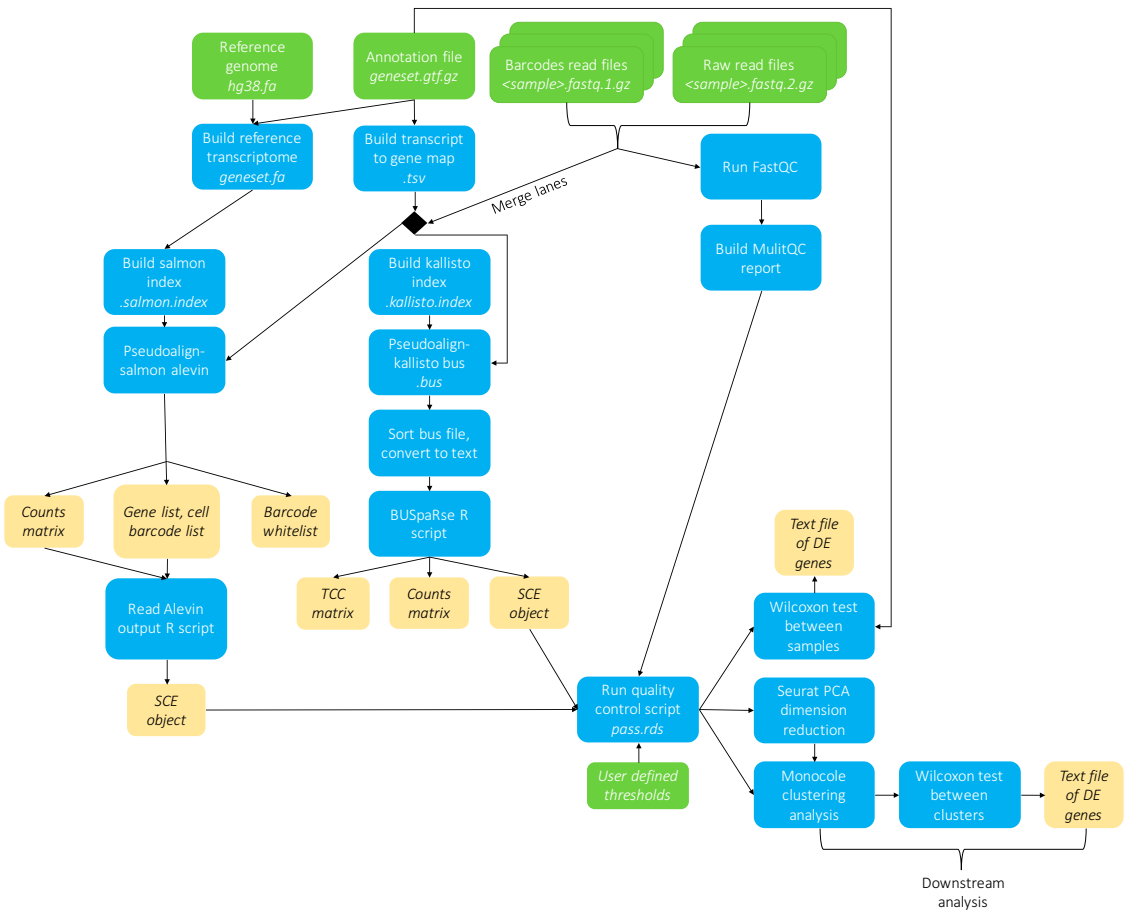


Figure 4.1: Flowchart outlining scRNA-Seq pseudoalignment pipeline- PLACEHOLDER- remake figure

4.2.2 Benchmark

Benchmarking measures the performance of a method/software relative to other methods available. Run time and the accuracy of results are often the factors considered in a benchmark. To be able to calculate the accuracy of results, the 'true' results must be known. This is difficult in scRNA-seq analysis as no gold standard analysis protocol exists. Instead, methods are compared against simulated results which act as the underlying 'ground truth'.

A benchmark was conducted using simulated data, utilising the pseudoalignment pipeline outlined above. Kallisto BUS/BUStools and Salmon Alevin pseudoalignment methods were both implemented and their performance compared to one another. Both pseudoalignment tools have previously been compared to traditional

mapping tools[196, 198] and both showed comparable accuracy levels, therefore this benchmark does not include the performance of traditional mapping methods.

Simulated data

Simulated reads with a know ground truth counts matrix were generated as follows: 10X (version 2) fastq files of 4k PBMCs from a healthy human donor were downloaded [2]. These sequencing files were processed using Salmon Alevin. The resulting Alevin output folder was used as input for Minnow, using Minnow's alevin-mode. Minnow generates droplet-based scRNA-seq simulated reads, working backwards from a known counts matrix to generating raw sequencing files from which the counts matrix could have originated. The valid cell barcode list (whitelist) for 10X chemistry was used (*737K-august-2016.txt*[3]). Minnow was ran with an error rate of 0.001 and with 12 simulated PCR cycles. Minnow accounts for core experimental dscRNA-seq characteristics, such as PCR amplification bias, barcode sequencing errors, the presence of doublets and ambiguously mapped reads, to try and emulate a realistic set of sequencing reads consistent with the provided counts matrix.

The ground-truth counts matrix was converted to a Single Cell Experiment object (SCE) and the simulated reads were used as input for the scRNA-Seq pseudoalignment pipeline. The resulting count matrices outputted by Salmon Alevin and Kallisto BUS were converted into SCEs, subset and reordered so that they all contained the same cells and genes, in the same order. The Salmon Alevin and Kallisto BUS produced SCEs could then be compared to the ground truth SCE.

Run time

The simulated reads consisted of 434 million reads. Running Salmon Alevin and creating an SCE object took approximately 64 minutes; running Kallisto BUS, sorting and creating an SCE object took approximately 24 minutes. Using the bustools 'count' command to create a counts matrix may have further reduced run time, however more time would be needed to parse it into R and create an SCE object.

[2] `https://support.10xgenomics.com/single-cell-gene-expression/datasets`
[3] `https://github.com/COMBINE-lab/minnow/blob/master/data/737K-august-2016.txt`

Cell barcode handling

The ground-truth data contained 4340 cells. Alevin determined a threshold for the initial whitelist (a set of CBs that likely represent non-empty droplets) by finding a 'knee' in the knee plot shown in Figure 4.2. This initial whitelist contained 5261 cell barcodes, each observed at least 191 times. Following barcode error correction, the final whitelist contained 4340 cells, all of which corresponded to the same CBs as the ground-truth data.

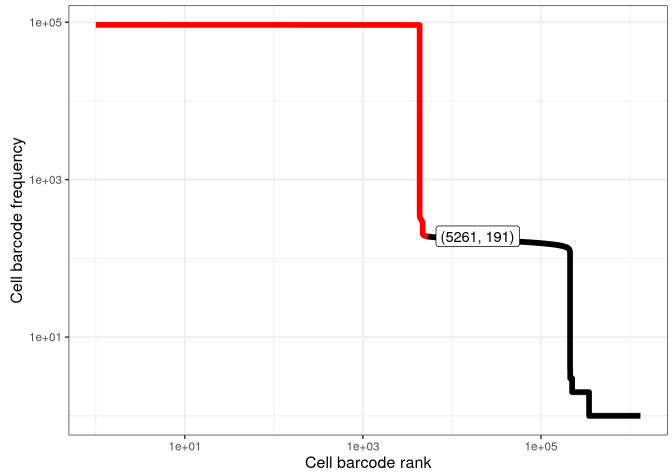


Figure 4.2: Alevin knee plot. This plot displays the number of times each cell barcode is observed, in decreasing order. Finding a 'knee' in this plot determines a threshold for the initial whitelist of CBs, which are unlikely to be empty droplets.

For Kallisto BUS, valid cell barcodes were determined using either emptyDrops (DropletUtils) or by using barcodeRanks and calculating the inflection point of a rotated knee plot (where the x- and y- axis are transposed; Figure 4.3). The inflection point method, gave a whitelist of 4339 cell barcodes (one fewer than the ground truth number), but all 4339 CBs corresponded to ground truth CBs. emptyDrops gave a total cell number of 12037, only 3746 of which were in the ground truth list of 4340 CBs. This was a large overestimate of number of cells present and the whitelist did not contain all of the valid CBs. Therefore, using the inflection point of the rotated knee plot was found to be the preferred method of filtering cell barcodes.

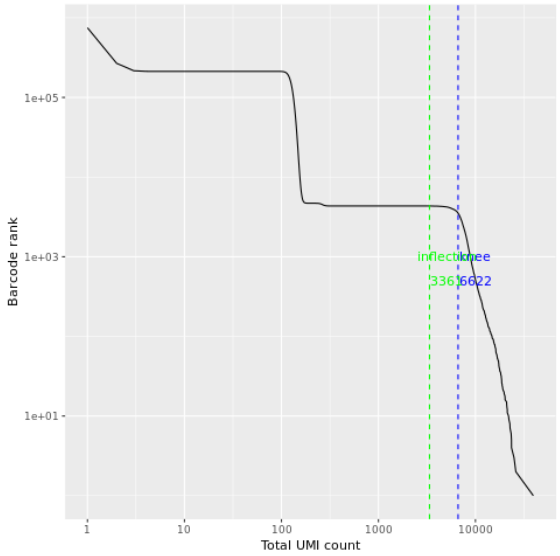


Figure 4.3: Kallisto BUS rotated knee plot. This plot shows the number of distinct UMIs against the rank of the barcode. The Pachter lab transpose the x- and y-axis on their knee plot, so that the x-axis displays distinct UMIs and the y-axis displays ranked cell barcodes, according to the number of corresponding UMIs to each CB. This is supposed to be more intuitive, having the number of distinct UMIs as the independent variable rather than cell barcode rank, as number of UMIs determine the cell barcode rank.

Gene expression predictive accuracy

To quantify each tool's accuracy of gene expression, precision, recall and an F1 score were calculated for each gene. The F1 score is a measure of a test's accuracy, it is the harmonic mean of precision and recall:

$$\begin{aligned} \text{precision} &= \frac{tp}{tp + fp} \\ \text{recall} &= \frac{tp}{tp + fn} \\ F_1 &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{precision + recall} \end{aligned} \tag{4.1}$$

Where for each gene: tp = number of true positives, fp = number of false positives, fn = number of false negatives.

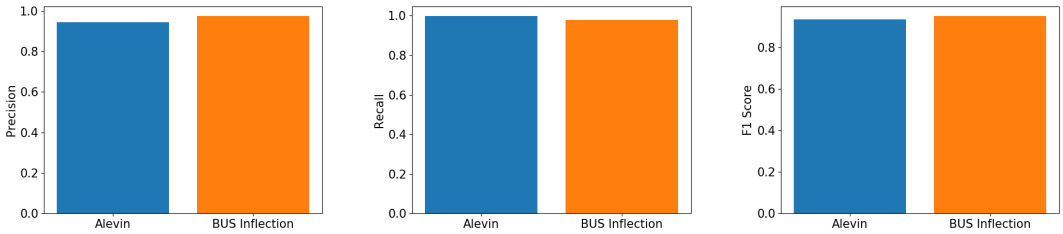
No expression was denoted by 0, and expression by 1. When recall or precision was undefined, i.e. a gene in Alevin/BUS matrix or the ground-truth matrix was not expressed by any cell, F score was defined as 0.

The mean F1 scores for Alevin and BUS processed data (Figure 4.4) were extremely similar to each other with scores of 0.93 and 0.95, this was due to the

		Ground truth	
		Expressed	**Not Expressed**
Alevin /BUS	**Expressed**	True positive	False positive
	Not Expressed	False negative	True negative

Table 4.1: Carol diagram of true/false positives/negatives based on expression between predicted values by Alevin/BUS and the ground truth matrix.

large number of F1 scores equal to 1. Figure 4.5 shows the distribution of F1 scores more clearly. Alevin seemed to produce more lower F1 scores than BUS.

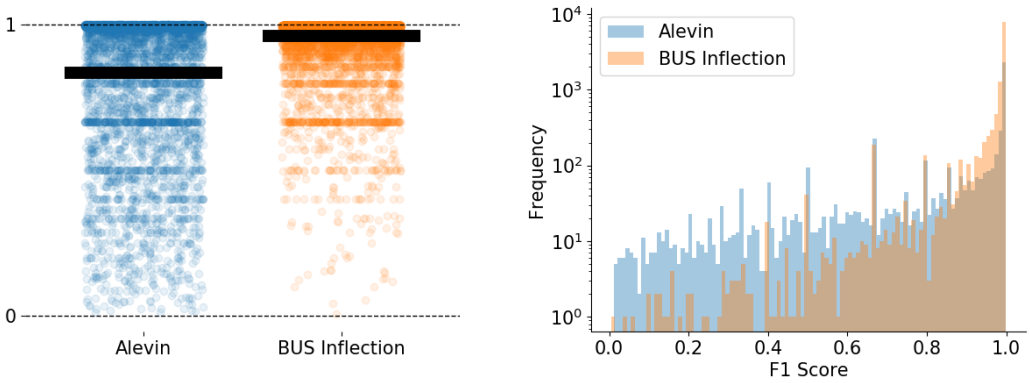


(a) Precision **(b)** Recall **(c)** F1 Score

Figure 4.4: F1 score. Two times the product of precision and recall divided by the sum of precision and recall. Measure of accuracy for the tools ability to predict gene expression. Expression classified by 0 or 1. Undefined scores have been removed. F1 scores were calculated for each gene across each cell.

Clustering

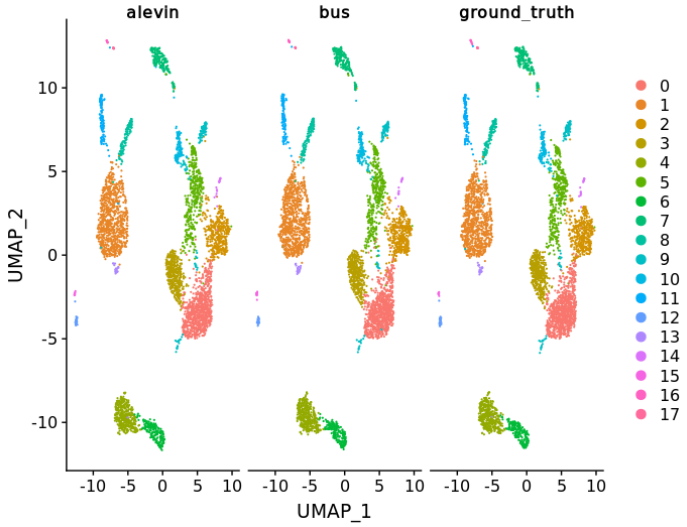
Clustering analysis was performed to visualise how well the tools processed the single-cell data and how clusters compared to ground-truth data. Seurat3 integrative analysis was performed so that the clusters of each sample could be directly compared. Figure 4.6 shows clustering of Alevin, BUS and ground-truth clustered data, using UMAP and tSNE dimension reductions. 18 clusters are present in all three of the data sets. Visual analysis suggests that the two dscRNA-seq quantification tools compare well to the ground-truth and capture most aspects of the data. From the benchmark it seems as if both tools are fit for purpose and can accurately quantify gene expression and correctly handle CBs and UMIs.



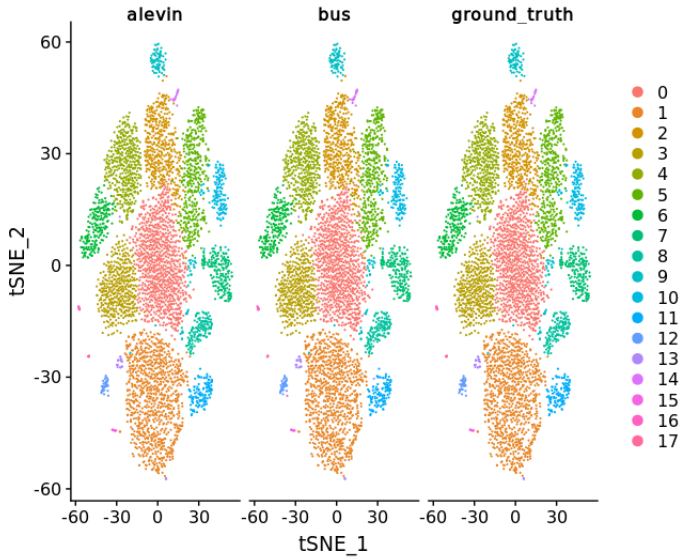
(a) F1 Score scatter graph

(b) F1 score histogram

Figure 4.5: F1 score distributions. 4.5a shows the F1 score for each gene expressed across all 4339 cells. The black bar denotes the mean F1 score for each cell. F1 scores of 0 have been removed.



(a) UMAP



(b) tSNE

Figure 4.6: Clustering analysis of the simulated data. 18 clusters are present in the ground truth data and Alevin and BUS processed data. Integrated clustering was performed using Seruat3[200], using both Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (tSNE) dimension reduction techniques.

4.2.3 Updated scRNA-seq pipeline

Following the bench-mark it was decided that Kallisto BUS and BUStools would be used to analyse single-cell data. This was due to its faster run-time and higher F1 scores. The analysis pipeline has been updated continually throughout the project. The updated workflow is outlined in Figure 4.7. Each major task has been split into separate lightweight pipelines, each containing multiple minor tasks. The black arrows in Figure 4.7 denote a separate processing pipeline. This allows the user to analyse the output after each step, and make changes to parameters as they see fit.

Following sequencing, raw FASTQ files are inputted into Kallisto bus. A bus file is generated along with corresponding information about equivalence classes and transcript names. Kallisto bustools is used to generate a cell by gene count matrix for each sample. These matrices are loaded into R and converted to Single cell experiment (SCE) and Seurat objects. Next, quality control (QC) is performed. Poor quality cells are removed based on numerous parameters. Cells with fewer than a set UMI number (the default is 500), cells with a very low or very high gene count (the default minimum and maximum is 300 and 6000, respectively), and cells with mitochondrial content over a certain ratio (the default is 0.1). This is a user-supervised process, which requires fine-tuning after inspecting graphs and quality metrics. Parameters can be altered and QC performed again. Following quality control, each sample is clustered individually by Seurat, then all samples are integrated together, by either Harmony or Seurat's SCTransform functionality. The integrated dataset is then annotated by cell type by the packages scClassfiy, singleR or clustfyr, in combination with a reference dataset or model. Often manual annotation by the user (using known biological markers) is required for finer annotation. The user can then perform further downstream analysis on the annotated integrated dataset, for example differential expression analysis and composition analysis.

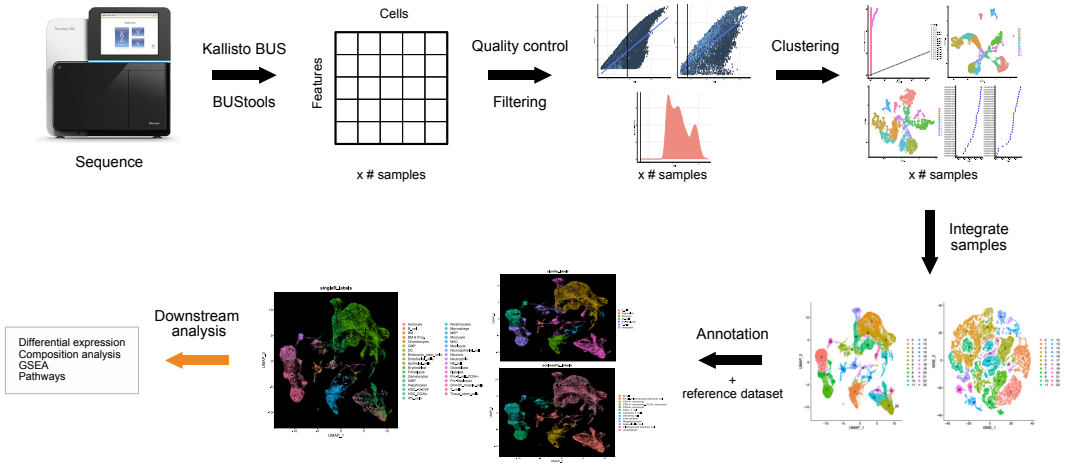


Figure 4.7: Outline of updated scRNA-seq workflow. Each black arrow denotes an independent pipeline. Following sequencing raw FASTQ files are processed by Kallisto bus and bustools and cell x gene count matrices are produced for each sample. Quality control (QC) is performed, based on a number of parameter thresholds. QC often requires user supervision to fun-tune the process. Seurat is then used to perform dimensionality reduction techniques and cluster samples individually. Harmony and Seurat are implemented to perform sample integration. The integrated dataset can then be annotated by cell type using packages cClassfiy, singleR or clustfyr, in combination with a reference dataset or model. The user can then perform further downstream analysis on the dataset, for example differential expression analysis.

4.3 scRNA-Seq velocity analysis pipeline

4.3.1 RNA velocity

scRNA-seq gene expression analyses capture only a static snapshot of the transcriptome in time. In a seminal paper by La Manno et al. (2018), RNA velocity methods were introduced with the aim of revealing the rate and direction of change of the transcriptome during dynamic processes, such as embryonic development or cellular dynamics following drug treatment[202]. In-line with previous observations, the authors found that between 15 and 25% of scRNA-seq reads contain unspliced intronic sequences. Using this knowledge, they found that the balance between unspliced and spliced mRNAs can be predictive of cellular state progression, and can be used to directly estimate the time derivative of gene expression on a timescale of hours. Therefore, using RNA velocity in combination with clustering analysis, the trajectory of a single cell can be tracked. This paper introduced Velocyto for RNA velocity analysis. Since its inception in 2018, numerous tools have been developed for RNA velocity analysis[197, 202–204].

4.3.2 Pipeline outline

A modular RNA velocity analysis workflow was constructed based around the kallisto BUStools (kb-python wrapper) velocity workflow[197]. The user requires an Ensembl reference GTF file and DNA fasta file for the species of interest, and their raw fastq read files as input. Firstly, an RNA velocity index is generated from the GTF and fasta file. Kb count is then used to generate spliced and unspliced transcript count matrices and a loom file. The resulting loom files are imported into R, and after some cell barcode manipulation, combined with existing gene-expression-based Seurat objects. This gives a Seurat object containing gene-expression data, UMAP/tSNE embeddings and spliced/unspliced RNA velocity counts. If the RNA workflow is combined with an annotated seurat object, the

trajectory of individuals cells can then be mapped onto UMAP plots with predefined cell-type annotations. The pipeline is deposited on GitHub and can be found here: `https://github.com/annajbott/pipeline_kb_velo`.

4.4 tRNA-seq analysis pipeline

4.4.1 Introduction

Transfer RNAs (tRNAs) are non-coding RNAs that transport amino acids to ribosomes during translation, to implement the genetic code. Most genomes contain distinct tRNAs for all 61 codons and these are encoded across multiple sites throughout the genome. Sequencing tRNAs is challenging both experimentally and computationally. The main experimental challenges arise from a stable secondary and tertiary structure, making library preparation difficult[205]. Therefore, efficient library preparation methods must be employed to overcome the ridged structure of tRNA, which usually limits the use of standard library prep methods for sequencing[206]. Computationally, the challenges come from overcoming the reverse transcription errors introduced by chemical modifications and accurately mapping reads to tRNA genomic regions, given their multiple genomic loci[207]. Typically, most mapping strategies for gene expression analysis only report read alignments with unique best matches and thus discard reads mapping to tRNA altogether. As a consequence, specialist mapping strategies to accurately map tRNAs have been proposed[208–210]. Specifically, Hoffmann et al (2018) proposed a two pass mapping strategy that first maps reads to a tRNA masked genome then secondly these unmapped reads are aligned directly to merged tRNA clusters[207].

Computational workflows for small RNA-seq data analysis have been developed previously[211]. However, there are currently a limited number specifically focused on tRNA data analysis. While there have been low-throughput implementations to aid tRNA analysis, such as tDRmapper[208], tRF2Cancer[212], SPORTS1.0[213] and MINTmap[209], there is now a significant unmet requirement for high-throughput approaches to analyse tRNA data. Firstly, it is desirable to have a single pipeline with the flexibility to perform read quality control, mapping to both general genomic

and tRNA features, and the ability to perform qualitative and quantitative analyses on tRNAs within a sample. Secondly, with decreasing sequencing costs, it is now common for small RNA-seq libraries to consist of many biological and technical repeats and be sequenced at a much greater depth than mRNA-seq libraries. Thirdly, an appropriate level of detailed reporting output is critical for biological interpretation, which should include appropriate visualisation and publication quality figures.

4.4.2 Pipeline outline

A tRNA analysis pipeline (tRNAnalysis) was constructed with the aim to meet the demands of tRNA-seq analysis. tRNAnalysis is written using the Computational Genomics Analysis Toolkit (CGAT) core workflow manager[190], therefore all input fastq files can be processed simultaneously, a detailed log is generated, and the pipeline can be locally run or executed in parallel across a high-performance cluster. tRNAnalysis implements best practice mapping strategies to allow accurate mapping of tRNA reads[207]. tRNAnalysis can be installed using Conda and a Docker image is also provided with all of the software and packages installed. Users can therefore use the pipeline without having to install numerous dependencies. Finally, tRNAnalysis provides a user-friendly html report to visualise qualitative and quantitative outputs. Given that the report is written in R using Rmarkdown, the report is easily extensible and customisable, thus providing a user-friendly approach for visualising the analysis.

The workflow is written predominantly in Python and R, using Ruffus decorators[189], and the CGAT-core workflow manager[190], allowing for automatic cluster submission and parallelisation. The pipeline runs via a single command line interface, providing appropriate default settings, with the option to customise configuration parameters and job resources as required.
The main steps in the analysis are:

- Read pre-processing and quality control

- Mapping of reads
- tRNA quantitative and qualitative analysis
- Downstream analyses and visualisation

Read pre-processing and quality control

tRNAnalysis accepts raw sequencing data (single-end fastq files) as input and integrates several tools for read quality checking and filtering, including CGAT tools[178], FastQC[214] and FastQ Screen[215]. Given the short length nature of small RNA library preparation, Trimmomatic[216] is used for adapter removal and to filter reads that fall short of quality thresholds. The quality metrics for each sample are then summarised with MultiQC[217].

Mapping of reads

The pre-processed fastq files and a list of gencode annotations are used as the input for mapping. The gencode annotations are supplemented with automatically downloaded RNA repeats (including RNA, tRNA, rRNA, snRNA, srpRNA) from the UCSC database[218]. Firstly, mapping is performed against the genome using Bowtie[219] to obtain a global representation of RNA types. For effective tRNA mapping, tRNAnalysis implements the best-practice mapping strategy proposed by Hoffmann et al (2018), in which tRNA loci are masked from the genome and instead, intron-less tRNA precursor sequences are appended as artificial chromosomes[207]. In first-pass mapping, reads that overlap the boundaries of mature tRNAs are extracted. In a subsequent round of mapping, the remaining reads are mapped to a tRNA-masked target genome that is augmented by representative mature tRNA sequences.

tRNA qualitative and quantitative analysis

tRNAnalysis provides RNA profile analysis that summarises the output of the read alignments mapping to various RNA sequences (e.g. miRNA, piRNA, snoRNA, lncRNA) by counting reads to features with featureCounts[220]. Plots are then generated for each sample, where the positional coverage counts are plotted relative

to the exon, upstream and downstream regions of the tRNAs. These plots can then be used to infer the levels of tRNA fragments within a sample. Using the nomenclature first proposed by Selitsky et al. (2015), if the primary tRNA sequence is < 41 nts and ≥ 28 nts, then it is defined as a tRNA-half, while if it is > 14 nts and < 28 nts then it is defined as a tRNA-fragment[221]. The frequency of read end site relative to the tRNA length is calculated and plotted as bar graphs for each tRNA cluster type. For quantitative measurement of tRNA differences between groups of samples, tRNAnalysis implements DESeq2 to perform differential expression analysis[180]. Finally, given there can be large sequence variations between tRNAs from different tissues as a consequence of RNA modifications[207, 222, 223], any nucleotide misincorporations in the mapped reads are determined. In order to accurately distinguish sequencing errors for true mismatches, *samtools mpileup* is employed to collate summary information in the mapped bam file and then likelihoods of misincorporation are calculated. This information is stored in a bcf file that is then parsed by *bcftools call*, which performs variant calling for each tRNA sequence[224]. The output is then stored as a vcf file and normalised for indels, then filtered for sequencing depth.

Downstream analyse and visualisation

DESeq2 differential expression is performed for sample groups inputted by the user. In order to visualise the output of tRNAnalysis, summary statistics are generated and then reports of these summaries are rendered in html format for visualisation using the Rmarkdown framework. Publication ready reports include figures such as coverage plots, volcano plots of differential expressed tRNAs, tables detailing tRNA modifications and bar graphs of tRNA frequencies. Once these standard reports have been generated, exploratory data analysis can be performed by modifying the Rmarkdown code, allowing for bespoke analysis, tweaking some parameters or to generate more specific publication-quality images.

4.4.3 Simulated data

The accuracy of tRNAnalysis was assessed using simulated read data. Splatter[225] was used to generate a simulated counts matrix, sampling over a negative binomial distribution with two conditions and three replicates. Biostrings[226] was used to generate a fasta file containing sequences from miRNAs, full-length coding mRNAs and tRNA clusters. The simulated counts matrix from splatter and the mixed RNA-species fasta file were used as input for Polyester[227]. Polyester was ran with an error rate per base of 0.003 to generate simulated RNA-seq read fasta files. Next, CGAT-apps[178] was used to convert the output fasta files into fastq files with uniform quality scores. The simulated fastq files were then used as input for tRNAnalysis to assess its performance.

Performance metrics

The accuracy of tRNAnalysis was evaluated using the simulated data above. Receiver operating characteristic (ROC) curves, precision, accuracy and F1 scores were generated, comparing differential expression between the tRNAnalysis processed simulated data to the known ground truth count matrices. The area under the curve (AUC) was found to be 0.91. The precision was 0.92, recall 0.64 and F1 score 0.76 using a p-value of 0.05, demonstrating a high level of accuracy (Figure 4.8).

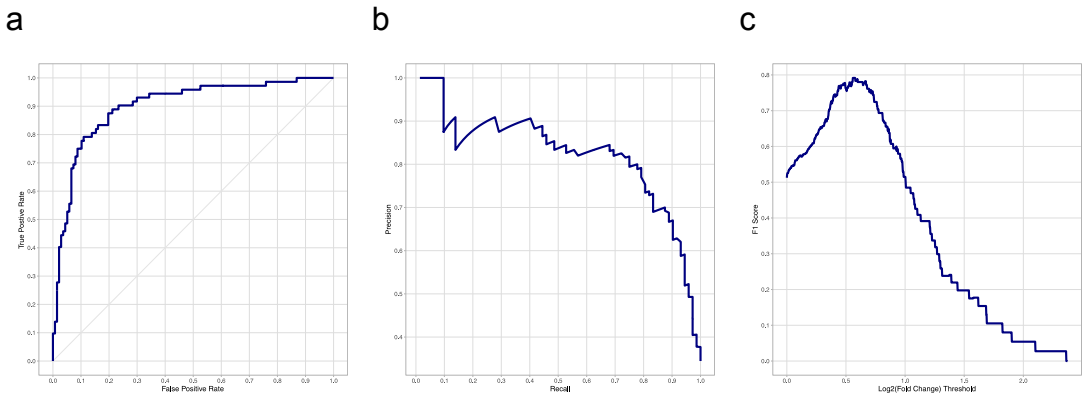


Figure 4.8: tRNAnalysis performance metrics. Simulated data ground truth counts matrix compared to tRNAnalysis-processed data. Classifying for whether a feature is differentially expressed or not by varying the $\log_2 FC$ threshold. a) Receiver Operator Curve (ROC) showing the accuracy of tRNAnalysis to perform differential expression. (Area Under the Curve (AUC), 0.91). b: Recall and precision. c) F1 score.

4.4.4 Reproduce published tRNA-seq analysis

Next, tRNAnalysis was applied to real data originating from a published paper. tRNA-seq data from Chiou et al. (2018) was used to illustrate the functionality of our pipeline[228].

This dataset comprises samples isolated from activated T cells and their associated extracellular vesicles. Using tRNAnalysis, we were able to confirm the main findings of Chiou et al. (2018), mainly that specific tRNA fragments are enriched in extracellular vesicles and released by activated T cells (Figure 2). Furthermore, we were able to present a more comprehensive evaluation of the different tRNA types within the extracellular vesicles (Figure 2). For each set of input files, interactive plots were created to show the coverage of reads over different types of tRNA fragments. For example, we plot coverage of our reads collapsed across the full length of all tRNAs (Figure 2A), which is further divided into coverage across codons (not shown) and amino acids (Figure 2B). Furthermore, we also report the relative proportion of tRNA fragments (Figure 2C) and tRNA halves (Figure 2D) in each of the input samples. Graphical comparisons between samples and relative controls are shown wherever possible. For example, the differential expression analysis is fully customisable by the user and generates comparative reports depending on the specific contrast supplied. < REPRODUCE DATA>.

4.5 Myeloma bone marrow classifier

4.5.1 Introduction

To be able to assess myeloma cells from MM patients, bone marrow samples are usually taken from the back of the hip bone. BM aspirates contain red blood cells (erythrocytes), white blood cells (myeloid cells, NK cells, T cells and B cells) and platelets. Traditionally flow cytometry immunophenotyping is used identify the malignant myeloma cells. Flow cytometry (FCM) immunophenotyping uses monoclonal antibodies against an array of antigens expressed on plasma cells, such as CD19, CD20, CD27, CD33, CD38, CD45, CD56, CD117, and CD138[229]. Many

scRNA-seq studies use biological markers, FCM immunophenotyping and bead enrichment to sort MM cells and immune cells prior to sequencing. Cells can also be identified post-hoc using computational methods. For example, Cohen et al. (2021) sorted MM cells prior to sequencing, they isolated MM cells by magnetic $CD138^+$ bead enrichment and stained $CD138^+/CD38^+$ cells with antibodies[230]. Another MM study examined the immune microenvironment of MM only by isolating $CD138^-$ and $CD45^+$ BM cell fractions, to exclude MM cells[231]. Cells can also be identified after sequencing using computational methods. Numerous automated packages exist to label cells, however a reference dataset is required to be able to draw comparisons to. Currently no MM classifiers or labelled references exist to automate MM scRNA-seq cell identification.

In Chapter 6, two scRNA-seq experiments were performed. For these experiments, no cell sorting took place and the whole bone marrow (WBM) niche was sequenced. Therefore, cell types had to be determined computationally. For annotation of these datasets, references based on healthy tissue were used to inform R annotation packages: clustifyr, scClassify and singleR. As the references originated from healthy tissue, they were unable to label the pathological myeloma cells, and MM cells had to be identified manually using expression of known biological markers. This took considerable time and required significant biological knowledge. An MM classifier or model that could automate cell-type annotation of MM patient BM scRNA-seq samples would save time for researchers and could encourage studies where the whole BM niche is sequenced, ensuring clonal MM populations lacking CD138 expression are not missed by traditional cell sorting techniques. This could also help remove some of the ambiguity of defining cell clusters.

4.5.2 Classifier building

Using the two scRNA-seq datasets from Chapter 6, cell classifiers were constructed for MM patient BM samples. scClassify's function `train_scClassify` was used to train reference models for scClassify annotation. Log-normalised expression data and a vector corresponding to each cell's manually defined cell-type annotation was used as

input for model training. Three models were generated. The first model was trained using the newly-diagnosed MM patient dataset, the second using the relapsed MM dataset, and the third was trained using both datasets. These models will be referred to as 'naive', 'relapsed' and 'joint'. The manual annotations used to train the models were much broader than the detailed annotations given in Chapter 6. This is to minimise the number of cells being labelled as 'intermediate' where significant overlap exists between cell subtypes The broad training annotations can be seen in Figure 4.9.

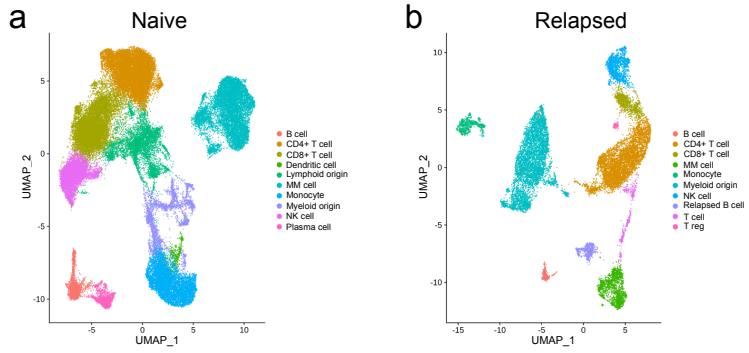


Figure 4.9: Cell type annotation for classifier building. Cell type classifications were made broader than the detailed annotation in given in Chapter 6. a) Newly-diagnosed (naive) MM dataset. b) Relapsed MM dataset.

4.5.3 Classifier testing

In order to to test the performance of the MM classifiers, publically available MM scRNA-seq data was downloaded from GEO. The test data comprises one PBMC sample from a relapsed and refractory MM (RRMM) patient, which contains both MM and immune cells (GEO accession number *GSE188632*). The deposited counts matrix was processed using the clustering and annotation modules of the scRNA-seq analysis workflow outlined in Section 4.2.3.

16 distinct cell clusters were identified using Seurat (Figure 4.10a). Firstly, the data was annotated using scClassify and clustifyr with healthy tissue references. A PBMC model reference was used with scClassify and a HumanCellAtlas (HCA) reference with clustifyr (Figures 4.10b and 4.10c). Both packages labelled the majority of cell clusters as monocytes or other myeloid cell types. Clustifyr-HCA

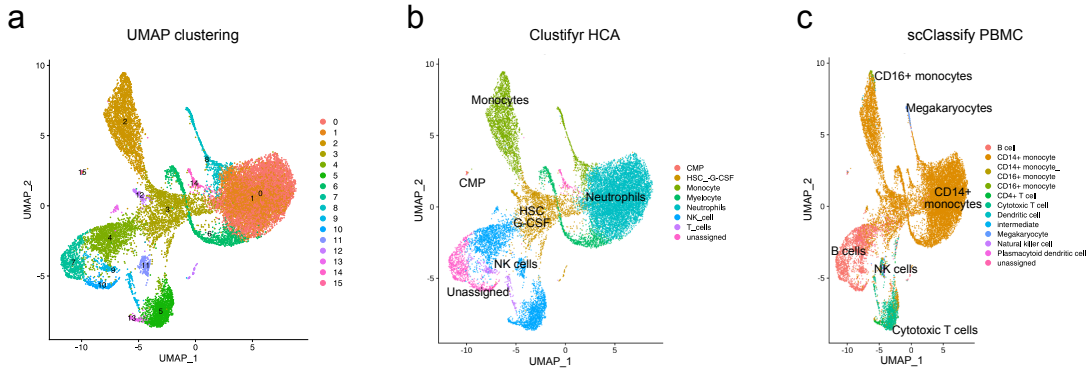


Figure 4.10: Public scRNA-seq data (GEO accession number *GSE188632*) clustering and annotation using references: from healthy tissue. a) UMAP clustering. b) Clustifyr annotation performed with the HumanCellAtlas (HCA) reference. c) scClassify annotation performed with the joint PBMC model as training data.

assigned clusters 4, 5, 11 and 13 as NK cells, whilst scClassify-PBMC assigned clusters 5 and 13 as T cells and cluster 11 as NK cells. Clusters 7 and 10 were left unassigned using clustifyr-HCA annotation, whilst clusters 4, 7, 9 and 10 were labelled as B cells by scClassify-PBMC. From the combination of the two annotations it seems like clusters 5 and 13 are T cells, cluster 11 is NK cells, and clusters 0, 1, 2, 3, 6, 8, 12 and 14 are myeloid cells. However, from these annotations alone, it is unclear which clusters are the MM cell population.

Next, the scRNA-seq data was annotated using the MM dataset-trained model classifiers generated above (Figure 4.11). Using each model, scClassify identified

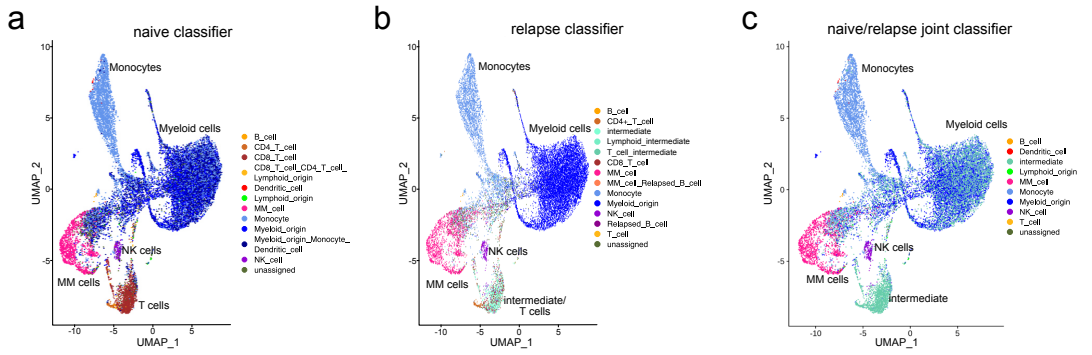


Figure 4.11: scClassify annotation performed on public scRNA-seq data (GEO accession number *GSE188632*) using MM classifier models generated above. a) Newly-diagnosed (naive) MM patient dataset. b) Relapsed MM patient dataset. c) Joint newly-diagnosed/relapsed MM patient dataset.

a number of cells as MM cells (coloured pink in Figure 4.11). The naive model

annotation identified the majority of cells in clusters 7, 9 and 10 as MM cells. The relapsed model annotation identified cells in clusters 4, 7, 9 and 10 as MM cells. The joint model annotation mainly identified clusters 7, 9, and 10 as MM cells, and some cells in cluster 4 as MM cells. Additionally, all three models seem to agree with each other and scClassify-PBMC's assignment of non-MM cells, with clusters 5 and 13 being assigned as T cells, cluster 11 as NK cells and the rest of the non-MM cells clusters as myeloid cells.

To assess if the models' MM cell classifications were correct, MM marker expression was examined (Figure 4.12). As discussed in Section 6.2.2, MM cells tend

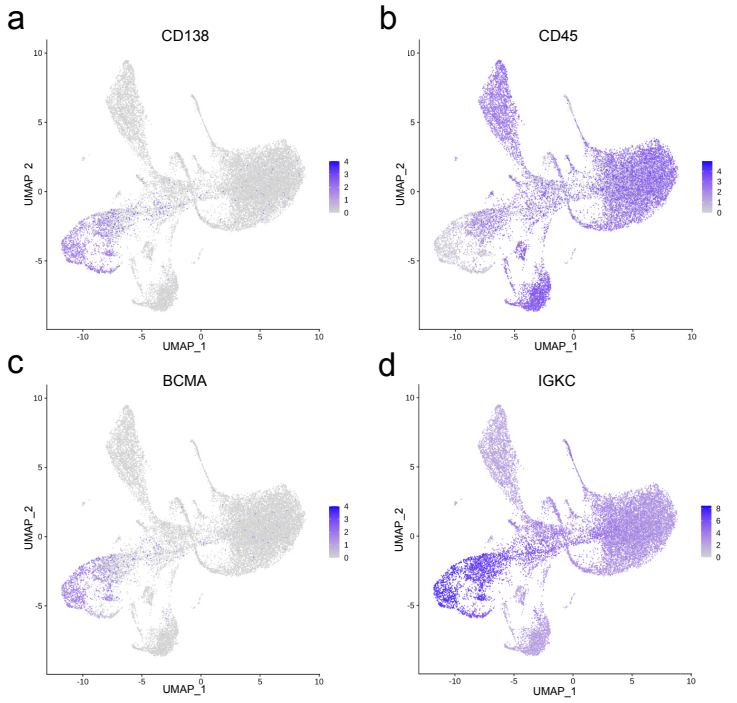


Figure 4.12: Public scRNA-seq data (GEO accession number *GSE188632*) biological MM markers featureplots. a) *CD138* expression- cluster 4, 7, 9 and 10 mainly expressing *CD138*. b) *CD45* expression- clusters 7, 9 and 10 not expressing *CD45*. c) *BCMA* expression- clusters 4, 7, and 9 mainly expressing *BCMA*. d) *IGKC* expression- clusters 4, 7, 9 and 10 have the highest expression of *IGKC*. Expression of *CD19* and *CD20* was not detected in this dataset. There was minimal expression of *SLAMF7* and *IGLC2*.

to express or over-express *CD138*, *BCMA*, *IGKC*/*IGLC2*, *SLAMF7*, and not express or under-express *CD20*, *CD19* and *CD45*. *CD20* and *CD19* were not expressed in this dataset, and *SLAMF7* and *IGLC2* were not expressed very highly in this

dataset. *CD138* and *BCMA* were shown to be expressed by clusters 4, 7, 9 and 10. *IGKC* was highly expressed by clusters 4, 7, 9 and 10. Clusters 7, 9 and 10 did not express *CD45*, whilst cluster 4 expressed *CD45*, but to a lesser degree than other clusters. Therefore, it can be concluded that clusters 4, 7, 9 and 10 are likely to be the MM cell population. Clusters 7, 9 and 10 are mainly $CD45^-$ MM cells, and cluster 4 is mainly comprised of $CD45^+$ MM cells. Therefore, all three MM model annotations were correct in assigning clusters 7, 9 and 10 as MM cells. Therefore there is high confidence, based on the MM gene marker expression knowledge and literature, that the automated annotation using the three MM models correctly assigned clusters 7, 9 and 10 as MM cells. Based on gene expression knowledge, it also seems highly likely that cluster 4 is also comprised of MM cells. However the naive and joint models were unable to label many of the cells in this cluster as MM cells, instead labelling them as intermediate cell-types or as myeloid cells.

Next, performance was evaluated quantitatively using ground-truth data. Knowledge on MM gene expression was used to inform the ground truth data, where cells in clusters 4, 7, 9 and 10 were labelled as MM cells, and other clusters as non-MM cells. To quantify each model's accuracy of classifying MM cells, precision, recall and F1 scores were calculated for each of the three models' assignment of cells (as MM vs non-MM) compared to the ground truth results (Figure 4.13). All three models had very high precision values: 0.998, 0.965, 0.987, for the naive, relapsed and joint models, respectively. This reflects a very small false positive rate- the models did not label many non-MM cells as MM cells. The recall score of the models was not as high, indicating there were more false negatives- the models assigned numerous cells in the MM clusters (4, 7, 9 and 10) as non-MM cells types. The model with the highest F1 score (0.778) was the model generated using the relapsed MM dataset only. This is a good score, indicating that the relapsed model was fairly accurate at classifying MM cells in the test dataset, and demonstrates the effectiveness of using the MM dataset trained models to annotate further MM patient BM scRNA-seq samples.

The GEO dataset originated from a RRMM patient, therefore it is logical that the relapsed model was the most accurate at labelling MM cells. MM patients'

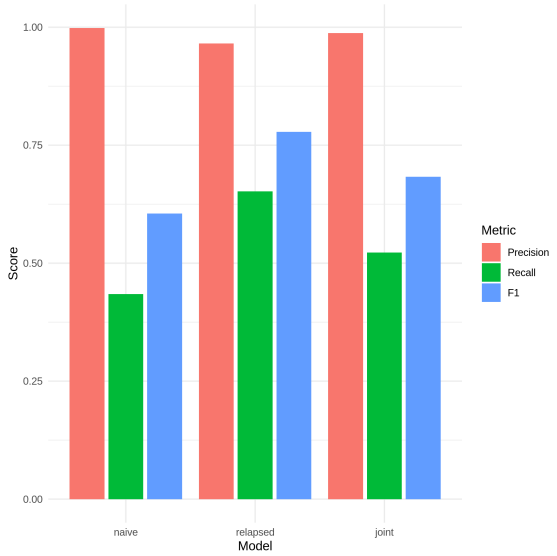


Figure 4.13: MM cell classifier performance. Precision, recall and F1 scores for the three models ability to accurately label MM cells; where true positive indicates a cell from the MM clusters (4, 7, 9 and 10) being labelled as 'MM_cell' by the model, false positive- a cell from any non-MM cluster being assigned the 'MM_cell' label, false negative- a cell from one of the MM clusters being assigned any label other than 'MM_cell'.

transcriptomes change considerably throughout the course of disease progression and treatment cycles, hence why the naive dataset (although being extremely cell rich- with over 60,000 cells) had the lowest recall score and was not able to classify MM cells as well as the other models. This should be considered when using the MM classifiers in future, applying the correct classifier to a dataset, such that the naive model is applied to newly-diagnosed MM patient datasets and the relapsed model to RRMM patient datasets. In addition, as shown in the joint model, scClassify models can be trained using multiple patients or scRNA-seq experiments. As more MM scRNA-seq experiments (including the full BM niche) are performed and the data made publically available, more comprehensive models can be generated, with further MM scRNA-seq annotations added to the training set. More patients samples would make the classifiers more accurately represent the heterogeneity seen in MM, and therefore more accurately annotate future MM datasets.

4.6 Discussion

Before the advent of NGS, the bottleneck in omics experiments was the technology available, which severely limited data acquisition. The duration of time for the completion of sequencing projects was very long (for example the 13 years taken to sequence the human genome). Since the use of NGS and TGS technologies– with their higher throughput compared to traditional first-gen techniques– the bottleneck of omics experiments has shifted to data processing and data discovery. Over the last two decades, numerous computational methods and tools have been invented to relieve the strain of data processing times.

The ever-increasing volumes of omics-data require data-processing pipelines/workflows that are robust, quick and efficient, whilst possessing the scientific rigour that allows for consistent and fully reproducible analysis of results. Robust and reproducible genomic workflows allow for confidence in interpreting results in meaningful biological context.

The CGAT framework was used to construct computational pipelines for data processing and extraction of biologically meaningful statistical analyses, summaries and visualisations from input data. The CGAT-framework was chosen as the base for the computational pipelines in this work because: CGAT-core can handle parallelisation across HPC clusters, it is integrated with the use of Conda environments, it allows for parameterisation of pipelines, it is open-source, and it records detailed logs of progress. Additionally, significant CGAT toolkits are available for specialised bioinformatics tasks. The large size of omics data means that many computational biologists use HPC clusters to analyse their data, to reduce computing time. Moreover, most cluster users do not have sudo access rights, therefore must use package and virtual environment managers. Conda integration was an important factor for the pipeline-framework, as other package managers, such as Pip, are mainly for python package installation, whereas Conda is a cross-platform environment manager. Many bioinformatic tools are written in R and are easily available on the Bioconda channel of Conda. Logging of pipeline progress allows for easy reproducibility of results. CGAT pipelines are built on top of Ruffus

workflow management. Compared with other workflow managers, Ruffus has been rated manager with the highest ease of development, such that it requires the least effort to compose new workflows[232]. This is further adds to its appeal, as it requires less time and effort to create new workflows, but also means that other users should be easily able to debug errors themselves and adapt the workflows as they see fit.

Using the CGAT framework, analysis pipelines were constructed for dscRNA-seq pseudoalignment analysis, scRNA velocity analysis and tRNA-seq analysis.

4.6.1 Benchmarks

Benchmarks were conducted to assess the performance of these pipelines. A combination of simulated and real data was used to assess performance. Simulated data allows comparison of the output of a given method to a known 'ground-truth'. This is a key strength of using simulated data as you can directly compare to known scores or values, and quantitatively assess the accuracy of a method. For transcriptomic data, often the known ground truth data is a counts matrix, either gene x sample for bulk methods or gene x cell for single-cell methods. Additionally, simulated data can be used as lightweight data for testing, as opposed to using large real-world data. However, using simulated data poses several limitations. The model under which the simulated data were generated can alter the results of the benchmark and favour certain methods over others. For example, a benchmark was performed for 12 different scRNA-seq simulation methods, across 35 different datasets[233]. The study observed differing results for each method across four evaluative categories (data property estimation, biological signals, scalability and applicability[233], indicating that the method of simulating data can affect the performance of a model in a benchmark. Often authors of a new method or tool perform a benchmark themselves, boasting its superiority over previous methods, without using a systematic assessment procedure (trialling numerous simulated datasets, with varying parameter values), this can lead to biases in their published results– this is known as the self-assessment trap[234].

Secondly, simulated data is unable to capture true experimental variability, and will always be less complex than real-world data[234]. It has been shown that various scRNA-seq data simulation methods are unable to reflect the level of heterogeneity seen in a population due to inter-patient variability[233]. Considering the pipeline generated in this work will be primarily used to analyse MM data, the benchmark has done little to support that it will be appropriate to infer results from heterogeneous MM patients.

Thirdly, transcriptome simulation methods are unable to simulate reads from intronic and intergenic regions of the genome, or transcripts that are not part of the annotation used to generate the reads. This complexity is present in experimental datasets.

4.6.2 scRNA-seq analysis benchmark

MM is an extremely heterogeneous disease, with numerous interactions with the surrounding immune microenvironment. Therefore, to best examine MM and the effects of therapeutics, single-cell techniques are required, including scRNA-seq to study the transcriptome at the single cell level. To be able to interpret results from scRNA-seq data, a robust and reproducible scRNA-seq analysis pipeline was required.

It has previously been shown using simulated data, that pseudoalignment/ lightweight mapping strategies demonstrate similar accuracy to traditional mapping strategies such as STAR and Bowtie2, yet provide massive improvements in computing time and memory usage. Therefore, it was decided that only lightweight mapping strategies would be implemented in the scRNA-seq analysis pipeline. The pipeline was written with the ability for the user too choose either Kallisto BUStools or Salmon Alevin for pseudoalignment. However, for consistency in the lab's analysis workflow, only one tool would be used for all future scRNA-seq data. To determine which lightweight mapping method to use, a benchmark was conducted using simulated scRNA seq reads. From this benchmark Kallisto BUS/BUStools was

determined more accurate and faster than Salmon Alevin. This result has been confirmed by subsequent benchmarks[197, 235, 236].

More recently, the Patro lab have introduced a reimplementation of Salmon Alevin, called Salmon-Alevin-Fry (SAF)[237]. SAF outputs a file similar to the BUS file, with the aim of quicker compute time and less memory usage. The Pachter lab (the lab responsible for Kallisto) performed a benchmark of the new SAF framework against Kallisto BUStools[238]. To avoid the self-assessment trap, both programmes were ran using the developer recommended settings, and compared over many datasets, across a variety of organisms and tissues. The group used real-world data to compare the methods. They achieved this by processing the same experimental reads with both SAF and Kallisto BUStools and then combining the SAF processed cells and Kallisto BUStools processed cells together into one dataset. After dimensionality reduction, the distance between identical cells either processed by SAF or BUS was examined and compared to the next nearest cell. The group found that gene expression differences between SAF and Kallisto were negligible for clustering analysis and irrelevant for downstream analysis. They also found that even with the updated SAF framework, Salmon-Alevin-Fry remained significantly slower and required more memory to run than Kallisto BUStools. This confirms the results of the benchmark conducted in this work and the subsequent decision to use Kallisto BUStools as the lightweight mapper of choice for scRNA-seq analysis.

The benchmark performed in this work used precision, recall and F1 scores to assess accuracy. These scores use binary classifications, therefore the benchmark only looked at the accuracy of 'Expressed' vs 'Not expressed' for every gene, across every cell. This was considered an appropriate measure of accuracy, as the fraction of zeros (cells not expressing a given gene) in a scRNA-seq experiment is extremely high, sometimes exceeding 90%[239]. However, the benchmark does not give any information about abundance quantification accuracy. As such, one of the methods outputting a count value of 1 vs another method outputting a count of 1000 would be considered as the same by the benchmark (i.e. expressed), even though they are clearly very different counts biologically. Srivastava et al. (2020)

investigated the transcript abundance quantification performance of traditional mapping and alignment strategies Bowtie2 and STAR compared to the quasi-mapping (pseudoalignment) method Salmon[240]. The group calculated Spearman correlation of quantification estimates by each alignment method compared to Polyester simulated scRNA-seq ground-truth data.

To demonstrate the application of the pipeline and its effectiveness on MM data, a more comprehensive benchmark could be performed by implementing some of the strategies summarised above. For simulated data: as well as assessing accuracy of expression vs no expression, gene abundance quantification accuracy should be examined using Spearman correlation statistics. Real world data should also be included, to assess how methods perform with added experimental complexity and heterogeneity. Since no ground-truth data is available, it becomes more challenging to analyse accuracy. The strategy implemented in [238] could be generalised to include all traditional mapping and lightweight mapping technique. This could be performed by processing a real-world dataset using the numerous mapping/alignment methods, then combining them together into one counts dataset, so that for each biological cell there are numerous 'doppelganger' cells each processed by a different method. Following dimensionality reduction of this matrix, you would expect the 'doppelganger' cells originating from the same biological cell to cluster together. L1 distances from the centroid of these clusters could then be calculated for all of the methods, across every cell in the dataset, to give an overall score for each method and to see how much each method agrees with one another. This strategy could be implemented for real-world MM data, to ensure the alignment methods chosen are appropriate to analyse MM patient data.

4.6.3 MM classifier

The MM classifier generated in this chapter allows for automated identification of MM cells in scRNA-seq datasets, and removes part of the data processing bottleneck of data analysis. The MM classifiers implement a supervised machine learning (ML) algorithm (k-nearest neighbour model), to classify cell types. The

classifiers performed well on test MM experimental scRNA-seq data. There was high agreement between classifier MM identification and gene expression-based cell type identification. An independent dataset was used for testing, originating from a different patient, and performed in a separate experiment. However, the 'ground-truth' data was defined by me based on cell type classifications using gene expression data of the test data. The classifiers were trained on MM scRNA-seq data, also labelled by me using gene expression data. Therefore, this might influence the agreement between the classifier outcome and ground truth data.

The performance of the classifier is dependent on the data used to train the model. The classifier generated in this work is based on three experiments, consisting of BM samples from four patients, two of whom were newly-diagnosed MM patients, and two relapsed patients, totalling approximately 80,000 cells (Chapter 6- Table 6.1). These are very cell-rich datasets, however due to the heterogeneity seen in MM, the classifiers would probably benefit from more MM scRNA-seq datasets being incorporated in the training of the models. Additionally, the labels outputted by the classifiers are dependent on the initial assignment of cell types on the training data by the user. Therefore, accurate cell type assignment must be made on the initial training set, to ensure correct identification by the classifier on the query dataset.

Cell sorting vs whole bone marrow niche

During normal B cell development into plasma cells, cells start expressing CD138. In the bone marrow, CD138 is a specific surface antigen for plasma cells and MM cells[241]. The typical approach for isolating plasma cells is to use $CD138^+$ magnetic selection; this can enrich plasma cells around 100 times[242]. This is useful to offset the high costs of single-cell sequencing, so that a high cell count and read depth can be achieved for MM cells, without sequencing other immune cells, which are often of low importance to researchers. However, previous studies of MM cell lines and clinical MM samples have shown a small population of MM cells lack CD138 expression[243], therefore anti-CD138 antibodies will recognise only a sub-population of MM cells. CD138 also has a fast turnover on the cell membrane,

and is constitutively shed on cultured cells and apoptotic cells[242]. Passaging in cell culture, sample processing and drug treatment can induce rapid apoptosis in primary myeloma cells, leading to CD138 shedding. Furthermore, $CD138^-$ MM cells have been reported to have more proliferative potential than $CD138^+$ MM cells and that they play an important role in regulating bone marrow stromal cells[244, 245]. Matsui et al. (2004) reported the existence of $CD138^-$ MM 'stem cells', which had greater clonogenic potential than $CD138^+$ MM cells and phenotypically resembled postgerminal center B cells[243]. A decrease in CD138 expression has also been observed during the course of treatment in some MM patients[241]. Therefore, by separating cells based on CD138 expression, information on a whole subpopulation of MM cells is lost. Results will be skewed towards $CD138^+$ MM cells. With the large heterogeneity seen in MM and acquired-drug resistance theories regarding MRD and clonal evolution, the entire MM cell population must be investigated to get an accurate picture of the whole MM landscape.

Bansal et al. (2021) highlighted issues with sorting PCs using CD138, beyond the scope of losing heterogeneous $CD138^-$ MM populations. The group investigated the impact of $CD138^+$ magnetic bead-based selection compared to WBM processing on BM plasma cell surface markers using FCM[242]. They found that $CD138^+$ selection of PCs appears to change important markers on the PCs, such as substantial loss of expression of CD71, CD11b, CD11a, CD69 and CD49e, and also almost a complete loss of $CD45^+$ cells in half of cases. From the study it was not clear if this phenotypic change of PCs was due to antigen loss from the process of selection or if subsets of cells were eliminated/ preferentially selected. For DNA and RNA-based studies (like scRNA-seq) it is very concerning that $CD138^+$ selection could preferentially select for a subset of cells, as this would mean further heterogeneity is being lost and results are being skewed to a subset of MM cells. $CD138^+$ selection vs WBM methodology should be investigated using scRNA-seq to ascertain if the phenotypic changes of PCs after $CD138^+$ selection seen in flow cytometry reflect antigen loss or selection of a subset of cells.

Considering the above limitations regarding CD138 expression and selection, together with the established interactions between MM cells and their surrounding immune microenvironment, it seems sequencing the whole bone marrow niche should be the preferred method for MM scRNA-seq studies. Despite this, the vast majority of MM scRNA-seq studies sort cells prior to library preparation. Sequencing the WBM niche is more expensive than sequencing isolated PCs (if maintaining a good read depth and high cell count), it also requires more time, effort and skill to process the data– for example assigning cell types to clusters. The MM classifier created in this work automates cell type classification and helps address the additional computational effort required for MM WBM scRNA-seq. All future MM scRNA-seq data produced by the Oppermann lab will use the MM classifiers to assign cell types to clusters. Hopefully more external MM researchers will start adopting the WBM method for scRNA-seq. As more researchers perform MM WBM scRNA-seq experiments and make the resulting data publically available, MM classifiers can be improved further with more data from heterogeneous patients. The relapsed MM classifier created in this work is hosted publically on Github (`https://github.com/annajbott/MM_classifier`), along with code to generate new classifiers, so that other researchers can use this tool.

Further, this approach could be expanded to create MM classifiers not only to identify and label MM and immune cell types, but to also predict clinical characteristics in datasets, such as risk, response to certain therapeutics or drug resistance status. Previously, machine learning (ML) approaches have been applied to bulk RNA sequencing profiles from 53 MM patients to create a classifier to predict good and poor responders to certain treatment regimens[246]. ML classifiers based on scRNA-seq datasets would likely have more power and stronger predictive capabilities than classifiers built using bulk RNA-seq data, due to the underlying heterogeneity of the patient data and the high dimensionality of scRNA-seq data. In recent years there have been applications of scRNA-seq based ML predictions and classifications in cancer. scRNA-seq datasets have been used for binary and pan cancer classification of certain breast and skin cancers[247]. ML techniques have

been applied to scRNA-seq datasets to predict drug resistance and characterise inter- and intra-tumour heterogeneity in MM[248, 249]. However, these studies looked at CD138$^+$ PCs only and did not consider the immune microenvironment. I expect there to be numerous applications of ML and deep learning methods to MM scRNA-seq data in the coming years, as the price of single-cell sequencing lowers and more computational biologists move into the field. However, I do believe the WBM niche should be included in MM classifiers and predictive tools to fully utilise all knowledge available and capture as many MM features as possible to inform the underlying models.

5
Bulk RNA-seq analysis of ProRS inhibitors

5.1 Introduction

Although MM treatment has improved significantly in the last 20 years, MM remains an incurable disease. MM patients relapse and become resistant to drugs they have previously been treated with. Therefore, research into novel therapeutics that can overcome multi-drug resistance and can be used to treat relapsed patients is of great importance.

New analogues of the drug Febrifugine are being actively researched for the treatment of many diseases, including numerous cancers[152, 153]. Febrifugine is the biologically active component of the herb *Dichroa febrifuga*, which is considered one of the fundamental herbs in traditional Chinese medicine[148]. Febrifugine is a quinazolinone alkaloid, which has been shown to possess strong anti-malarial properties. One such Febrifugine derivative, halofuginone (HF), has previously been shown to inhibit T Helper 17 (TH17) cell differentiation, by activating the amino acid response (AAR)[157]. Halofuginone inhibits the enzyme glutamyl-prolyl tRNA synthetase (EPRS)[158]. EPRS is a bifunctional aminoacyl-tRNA synthetase (AARS) and catalyses the the aminoacylation of glutamic acid and proline tRNA species (such that it charges its cognate tRNAs with glutamic acid and proline). Halofuginone and Febrifugine compete with proline at the prolyl-tRNA synthetase active site of EPRS, specifically targeting utilisation of proline during

translation[158]. This results in an accumulation of uncharged $tRNA^{pro}$s, giving the same cellular environment as if the cell were proline deficient, triggering the AAR to respond to the apparent proline deprived state.

AARSs are essential in protein synthesis, aiding in building chains of amino acids. Human cancer cells often have an increased rate of protein synthesis, this is especially true in multiple myeloma, creating huge amounts of non-functional paraprotein, therefore are more reliant on aaRSs. As discussed in Chapter 2, HF has previously shown anti-MM activity in-vitro and in-vivo[173]. HF induced cytotoxicity and apoptosis in numerous MM cell-lines and primary MM cells. HF was also shown to inhibit MM growth and prolong survival in a mouse xenograft MM model. However, the mechanism by which HF exerted its affect was not elucidated, and it is not clear if the AAR plays a role in HF's effectiveness in MM.

It has also been shown that HF's anti-MM effect can be reduced in the presence of excessive proline[173]. Proline is very abundant in many tumours, which may result from upregulated matrix metalloproteinases (MMPs) degrading collagen in the extracellular matrix (ECM), and/or increased conversion of glutamine to proline compared to normal cells[161]. Elevated proline levels have been noted in the BM of MM patients[250]. This increased proline concentration means that HF has a very narrow therapeutic window, and exhibits many side effects.

Recently, The Mazitschek group have synthesized numerous other compounds which target the ProRS site of EPRS. One such example, NCP26 (Figure 5.1b), does not compete with proline for the active site of ProRS, unlike febrifugine and halofuginone. NCP26 binds to the ATP binding site of ProRS, inhibiting utilization of ATP. Aminoacylation is an ATP-dependent process, therefore blocking ATP binding inhibits this process, and also leads to an accumulation of uncharged $tRNA^{pro}$s. NCP26 will hopefully alleviate some of the issues associated with HF treatment. More ProRS inhibitors have been synthesized by the group, including NCP22.

This chapter uses MM cell-line models to assess the effectiveness of these ProRS inhibitors and uses bulk-RNA-seq to capture the transcriptional landscape following treatment, and determine their mechanism of action in MM.

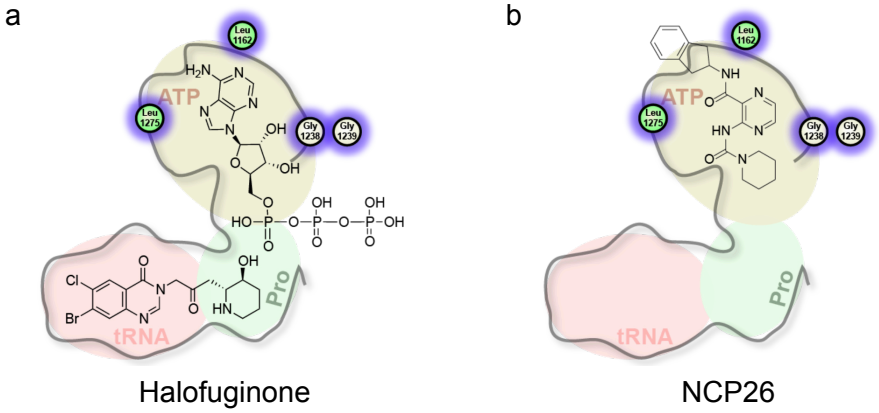


Figure 5.1: Diagrams of halofuginone/MAZ1392 (a) and NCP26 (b) and their chemical structures. Halofuginone is an ATP dependent, proline and tRNA competitive ProRS inhibitor. NCP26 is an ATP competitive and proline uncompetitive ProRS inhibitor. Aminoacylation is an ATP-dependent process, requiring ATP to activate amino acids. Figure by Ralph Mazitschek.

5.2 Cell-based assay results

5.2.1 Cell line validation

In this chapter, two MM cell lines are used. Low-passage-number AMO-1 (henceforth referred to as wild type; WT) cells, which are sensitive to various MM treatments, and carfilzomib-resistant L363 (henceforth referred to as CFZr) cells. Dose response curves for the two cell lines treated with proteasome inhibitors (PIs), carfilzomib and bortezomib, were generated to confirm the PI-resistance of CFZr cells and PI-sensitivity of WT cells (Figure 5.2). The IC_{50}s for AMO-1 (WT) cells were

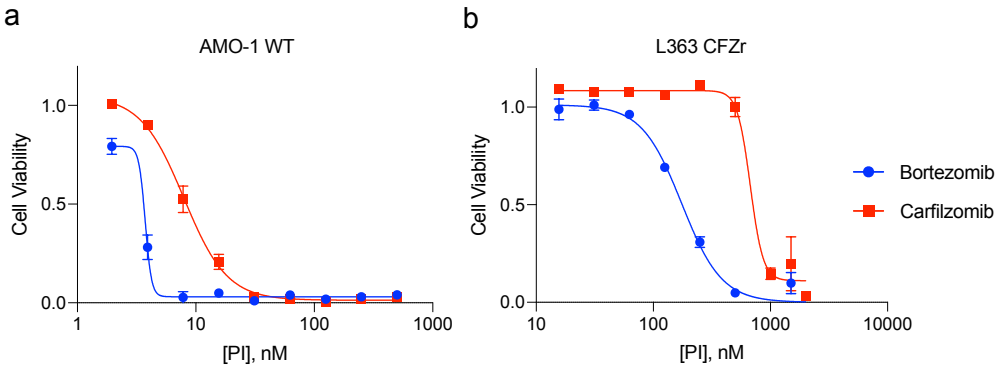


Figure 5.2: Multiple myeloma (MM) cell lines, AMO-1 (WT) and Carfilzomib-resistant L363 CFZr, treated with carfilzomib (CFZ) and bortezomib (BTZ). MM cell lines were treated for 48 hours with a range of concentrations (approximately 1nM-1µM) of proteasome inhibitors (PI). a) AMO-1 (WT cells). b) Carfilzomib-resistant L363 cells (CFZr).

approximately 3.74nM and 7.93nM for bortezomib and carfilzomib, respectively. For L363 CFZr cells, the IC_{50}s were approximately 174.8nM and 673nM for bortezomib and carfilzomib, respectively. This confirms that WT cells are very sensitive to proteasome inhibition, with IC_{50}s in the low nanomolar range. It also demonstrates that CFZr cells are much more resistant to proteasome inhibition, with the IC_{50} of CFZr cells almost 85 times higher than WT cells for carfilzomib.

As expected, CFZr cells are more resistant to carfilzomib treatment than bortezomib treatment, as their resistance was developed by exposure to increasing carfilzomib concentrations. However, it is clear that some cross-resistance to bortezomib is also acquired (the IC_{50} for BTZ treatment is 46 times greater in CFZr

cells than WT cells.) This is likely due to the similar mechanism of actions of the two drugs, conferring resistance to bortezomib treatment too. It could also be in part due to increased expression of general multi-drug resistant genes, such as ATP Binding Cassette Subfamily B Member 1 (*ABCB1*). This mimics clinical MM well, as patients are treated with carfilzomib as a second-line treatment, once they have already been treated with bortezomib, developed some resistance to it and relapsed.

Together, this validates that AMO-1 (WT) cells and L363 CFZr cells are sufficient cell models for PI-sensitive and PI-resistant MM.

5.2.2 Halofuginone and NCP26 are cytotoxic to drug sensitive and drug resistant MM cell lines in a dose-dependent manner

The effect of the ProRS inhibitors NCP22, NCP26, MAZ1805 (Halofuginol) and MAZ1392 (Halofuginone; HF) on cell viability was investigated using the MM cell lines AMO-1 and L363 CFZ-r. The cell lines were treated with a range of concentrations of each compound (3.9nM- 1µM). Cell viability was assessed using presto-blue assays (section 3.3.2), and dose response curves were generated (Figures 5.3a and 5.3c). Halofuginone and NCP26 reduced cell viability of PI-sensitive AMO-1 cells and carfilzomib resistant L363 cells in a dose-dependent fashion. For this concentration range, MAZ1805 and NCP22 seemed to have little effect on cell viability of WT or CFZ-r cells, and IC_{50} values were unable to be calculated. Halofuginone was found to be more potent than NCP26.

For WT AMO-1 cells, halofuginone had an IC_{50} of 141.8nM and NCP26 an IC_{50} of 502nM. For CFZ-r cells, halofuginone had an IC_{50} of 1185nM and NCP26's IC_{50} was ambiguous, a higher stock concentration would be required for calculation. Figure 5.3b and 5.3d show the proportion of viable cells following 48 hours of treatment of the ProRS inhibitors. WT AMO-1 cells were found to be more sensitive to NCP26 and halofuginone treatment than carfilzomib resistant L363 cells. This may indicate some acquired cross-resistance built up from carfilzomib exposure or inherent resistance in the L363 cell line over the AMO-1 cell line.

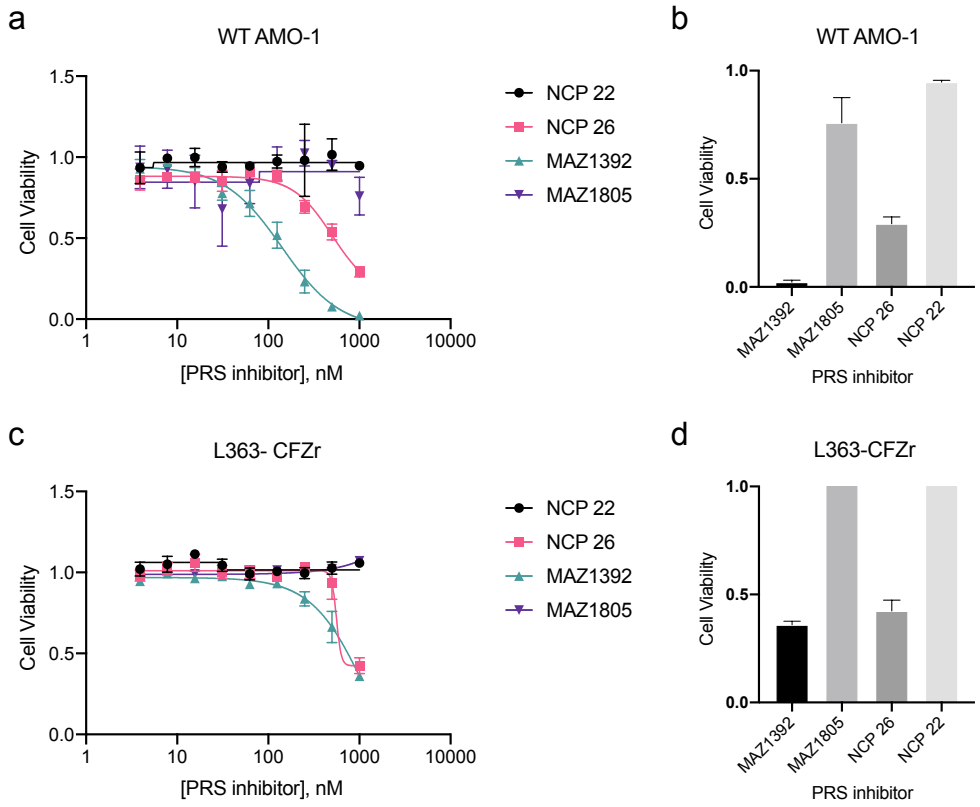


Figure 5.3: Multiple myeloma (MM) cell lines treated with ProRS inhibitors- MAZ1392 (Halofuginone), MAZ1805, NCP26 and NCP22. MM cell lines were treated for 48 hours with a range of concentrations (3.91nM-1µM) of ProRS inhibitors. a) and c) Dose response curves. a and b) WT AMO-1 cells. c) and d) Carfilzomib resistant L363 cells. b) and d) Proportion of cells viable after 48 hours of 1µM treatment with each agent.

5.2.3 Carfilzomib and NCP26 have an additive or mild antagonistic effect together

Drug combinations have proved effective in MM in recent years, for example the combination of bortezomib, lenalidomide, and dexamethasone (VRd) is used extensively for newly diagnosed MM patients. Drugs are often used in combination so that outcomes are improved (synergistic efficacy) or to reduce off-target effects and toxicity by minimizing the doses of the drugs (synergistic potency)[251].

To assess if NCP26 and carfilzomib work together synergistically, AMO-1 cells were treated with varying concentrations of NCP26 and Carfilzomib for 72 hours, then presto blue assays were performed to determine cell viability. As shown by dose response curves in Figure 5.4a and the response matrix in Figure 5.4b, NCP26

and carfilzomib elicit a stronger cytoxic effect together than each agent individually. SynergyFinder[252] was used to calculate the compounds' synergy scores (-4.66 ZIP; -4.18 Loewe; -5.53 Bliss). From these values it is unlikely that NCP26 and Carfilzomib work together synergistically. NCP26 and carfilzomib seem to have an additive effect together, or slight antagonistic effect. This reflects a previous result

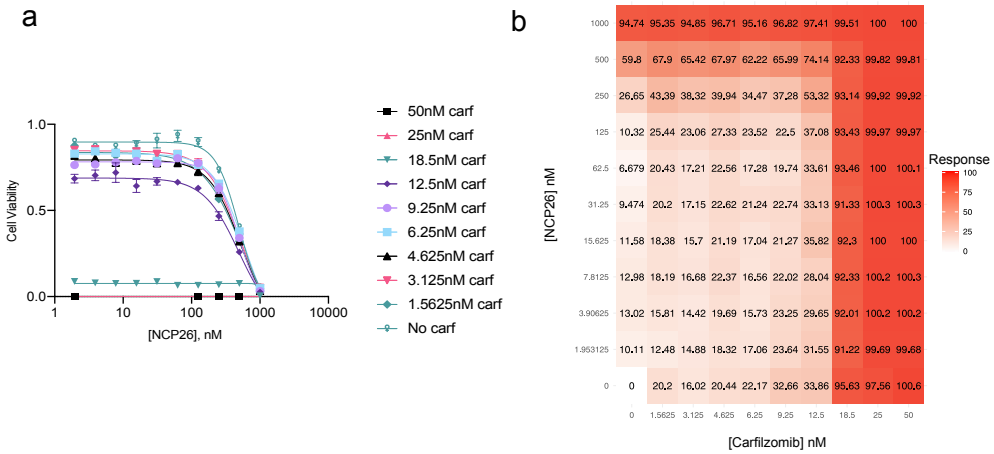


Figure 5.4: Investigating potential synergy between NCP26 and carfilzomib. AMO-1 cells were treated with varying concentration combinations of NCP26 and carfilzomib for 72 hours, then cell viability was determined using presto blue assays. **A)** Dose response curves for NCP26 with different carfilzomib concentrations. **B)** Matrix view of NCP26 and Carfilzomib concentration responses.

where HF demonstrated moderate antagonism with the bortezomib[173]. This may indicate that HF and NCP26 may not work very well in combination with proteasome inhibitors in MM. Although, the group only used a single concentration of BTZ (5nM) and 3 concentrations of HF (25, 50 and 100nM). This limits their ability to determine synergy, as only very few concentration combinations were tested. In the same study, HF was shown to exhibit moderate synergistic cytotoxicity with other anti-MM agents, including the IMiD lenalidomide and the corticosteroiod dexamethasone. This might demonstrate the potential of using ProRS inhibitors in combination with IMiDs and corticosteroids, perhaps in place of proteasome inhibitors, once PIs become less effective for MM patients. Indeed, with the promising results in Figure 5.3d, HF and NCP26 have shown cytotoxcity against PI-resistant cells and therefore hint towards being an effective agent against relapsed and PI-resistant MM.

5.3 Bulk RNA-seq

5.3.1 Experiment overview

The transcriptome of drug-sensitive and PI-resistant MM cells following treatment with four ProRS inhibitors was investigated using bulk RNA-seq. PI-sensitive WT AMO-1 cells and carfilzomib resistant L363 cells (CFZr) were used. Cells were treated for 6 and 24 hours with a DMSO control, or 1µM of a ProRS inhibitor (MAZ1392 (Halofuginone), NCP26, NCP22 and MAZ1805 (Halofuginol)), or 100nM carfilzomib (AMO-1 WT cells only). CFZr cells were treated in the presence of 100nM carfilzomib. Samples were prepped for sequencing as in Section 3.4. The computational workflow for bulk RNA-seq analysis is outlined in Section 3.7.1.

5.3.2 Clustering

Figure 5.5 shows clustering analysis. Figure 5.5a and 5.5b show the samples distinctly separate into respective cell types, and this makes up the majority (69%) of the variance in the dataset. Therefore, the different cell types (WT and CFZr) were separated and analysed individually (Figures 5.5 c-f).

The less active compounds from the ProRS inhibitor dose response curves, MAZ1805 and NCP22, cluster closely with the DMSO-treated controls. The more active inhibitors, NCP26 and halofuginone, cluster separately from DMSO controls and less active inhibitors, indicating they have elicited a stronger transcriptional response. At 6 hours NCP26 and halofuginone cluster closely together. At 24 hours NCP26 and halofuginone separate more. This may suggest a distinction in their mechanism of action, or this could just be reflecting the differences in their potency.

In WT cells, carfilzomib clusters separately from DMSO and the less active inhibitors. At 6 hours, carfilzomib samples are very separate from the NCP26 and halofuginone cluster, but they cluster together more closely at 24 hours. This could suggest an initial difference in mechanism of action and transcriptional response to the ProRS inhibitors, but culminating in a similar response as time progresses, such as cell stress and cell death pathways.

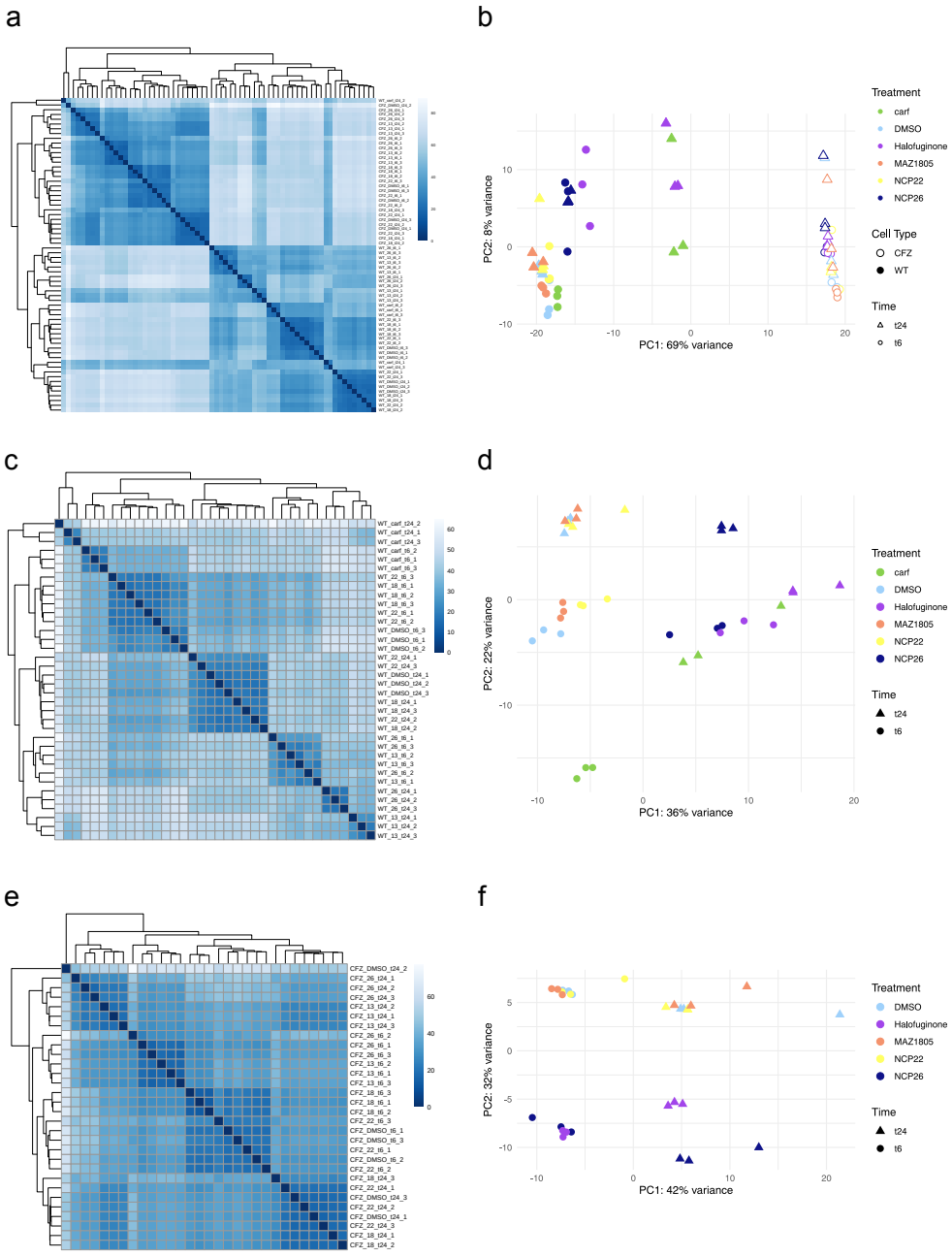
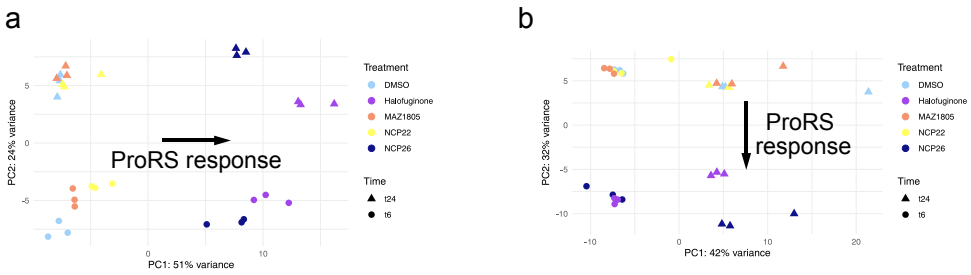


Figure 5.5: Bulk RNA-seq sample clustering. WT AMO-1 cells and CFZr L363 cells treated for 6 or 24 hours with a DMSO control, 100nM carfilzomib or 1uM of a ProRS inhibitor (MAZ1392/Halofuginone, MAZ1805/Halofuginol, NCP26 and NCP22). Clustering analysis of sample-sample distances (a, c and e) and principal component analysis (PCA; b, d and f). a) and b) both cell types (WT and CFZr) displayed; c) and d) WT AMO-1 only; e) and f) CFZr samples only.



WT AMO-1:

PC1 enriched pathways	FDR	Up/down?
ATF4 activates genes in response to endoplasmic reticulum stress	5.44E-6	Up
Response of EIF2AK4 (GCN2) to amino acid deficiency	7.31E-6	Up
Cellular response to stress	7.31E-6	Up
RNA Polymerase I Promoter Opening	5.44E-6	Up
Mitotic G1 phase and G1/S transition	1.38E-8	Down
Glycolysis	4.54E-6	Down
Cell cycle	2.19E-4	Down
Transcriptional regulation by TP53	1.29E-3	Down

CFZr L363:

PC2 enriched pathways	FDR	Up/down?
PERK regulates gene expression	2.74E-9	Up
ATF4 activates genes in response to endoplasmic reticulum stress	3.6E-8	Up
Response of EIF2AK4 (GCN2) to amino acid deficiency	1.23E-8	Up
Cellular response to starvation	1.53E-7	Up
Cytosolic tRNA aminoacylation	1.41E-5	Up
Unfolded protein response	3.58E-4	Up
Gap junction trafficking	6.81E-3	Down
HSF1 activation/ attenuation phase	3.59E-10	Down

Figure 5.6: PCA pathway enrichment. Pathway enrichment analysis performed using REACTOME of top contributing genes from principal component analysis (PCA). A) WT AMO-1 dataset with carfilzomib-treated samples removed. PC1 seems to account for the separation between DMSO controls/ less active compounds and the more active ProRS inhibitors (NCP26 and halofuginone). Genes contributing positively towards PC1 are upregulated in NCP26 and halofuginone compared to DMSO controls and less active compounds. Enriched pathways from top genes in PC1 shown beneath PCA plot. B) CFZr L363 dataset. PC2 seems to account for the separation between DMSO controls/ less active compounds and the more active ProRS inhibitors (NCP26 and halofuginone). Genes contributing negatively towards PC2 (down arrow), are upregulated in NCP26/ halofuginone compared to controls. Pathways enriched from top genes in PC2 shown beneath PCA plot.

The top principal components of the WT AMO-1 and CFZr datasets were examined more closely. Carfilzomib-treated samples were removed from the WT dataset, to ensure that the difference between controls and active ProRS inhibitors was captured in PC1 or PC2 following dimensionality reduction Differential expression, variance stabilising transformation and PCA was re-performed to the CFZ-less dataset. Pathway enrichment analysis was performed for the top genes from the principal component accounting for the difference between controls and halofuginone/NCP26 treatment (PC1 for WT cells, PC2 for CFZr cells; Figure

5.6). The pathways 'ATF4 activates genes in response to ER stress', 'Response of GCN2 to amino acid deficiency' and 'cytosolic tRNA aminoacylation' were all enriched, suggesting that the amino acid response is activated. The 'unfolded protein response' (UPR) was also enriched, as well as 'PERK regulates gene expression'. The UPR is a member of the integrated stress response (ISR), so many genes involved in the UPR overlap with genes involved in the AAR. *PERK* regulates the translation response of the UPR. Additionally, genes involved in the cell cycle and G1/S transition are negatively enriched. Previously, HF has been shown to induce the accumulation of cells in the G_0/G_1 cell cycle[173].

5.3.3 Drug sensitive MM

Differential expression

For AMO-1 WT cells at 6 hours, 2119 genes were differentially expressed ($|log_2FC| > 0.5$ and $p_{adj}<0.05$; DE) for NCP26 treated-samples, 3019 DE genes (DEGs) for halofuginone, 33 DEGs for MAZ1805, 218 DEGs for NCP22, and 983 DEGs for carfilzomib-treated samples compared to DMSO controls. At 24 hours, 3323 DEGs for NCP26-treated samples, 3426 DEGs for halofuginone, 2 DEGs for MAZ1805 and 2260 DEGs for carfilzomibtreated samples compared to DMSO controls. DEGs for NCP26 and halofuginone treatment are shown in Figure 5.7. Genes highly differentially expressed by the ProRS inhibitors are coloured and annotated with their gene symbol. Numerous genes involved in stress response pathways can be seen to be upregulated following NCP26 and halofuginone treatment, including *TRIB3*, *JUN* and *ATF3*. Additionally, various histone genes are upregulated. Some genes involved in cell cycle progression, such as *CDKN1A*, are downregulated following ProRS inhibition.

Transcriptional changes for 6 hours of exposure to NCP26 were compared with proteomic changes for the same treatment condition (Figure 5.8). Proteomic data was supplied by collaborators <COLLABORATORs names here>. Consistent with a mechanism of a global reduction of protein synthesis (with the exception of preferential translation of *ATF4*) upon ISR induction and eIF2α phosphorylation,

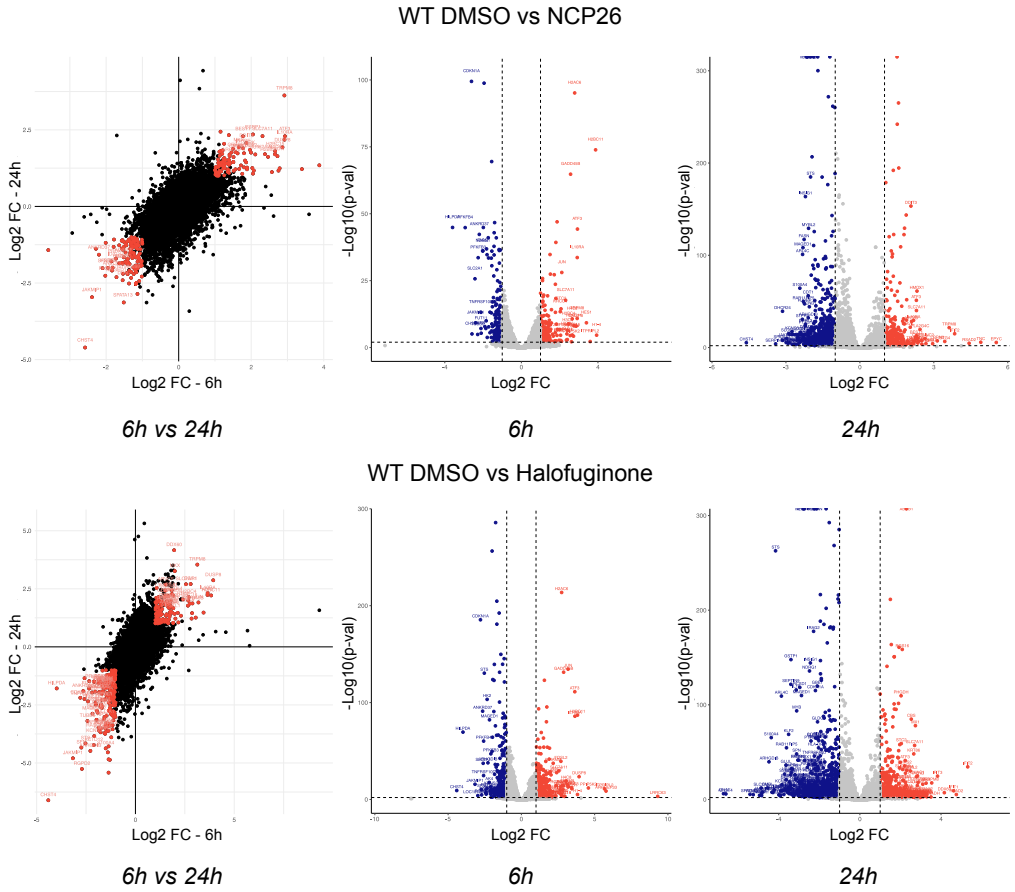


Figure 5.7: Differentially expressed genes (DEGs) for halofuginone and NCP26 treated WT AMO-1 cells at 6 and 24 hours. Scatter plot of genes for 6 hours treated vs 24 hours treated. Red points indicate genes which are differentially expressed ($p_{adj} < 0.01$) at both 6 and 24 hours. Volcano plots are also shown. Blue points indicate downregulated DEGs ($p_{adj} < 0.01$ & $\log_2 FC < -1$). Red points indicate upregulated DEGs ($p_{adj} < 0.01$ & $\log_2 FC > 1$). Top DEGs are annotated with HGNC symbols.

very few proteins with an increased abundance were identified (52 proteins with a $\log_2$ fold-change between 0.2 and 0.64). Additionally, a larger shift of proteins with lower abundance in NCP26 samples compared to DMSO was found. This correlates with transcriptional data where more DEGs were downregulated than upregulated for all ProRS treatment conditions. Also, in-fitting with this mechanism, selective *ATF4* target genes are upregulated, such as *TRIB3* and *INHBE* (as seen highlighted in Figure 5.8). This shows a dominant role of the integrated stress response at 6 hours on the transcriptomic and proteomic level.

The effects of NCP26 and halofuginone treatment on the amino acid starvation response were examined in more detail. AAR genesets 'GOBP response to amino acid

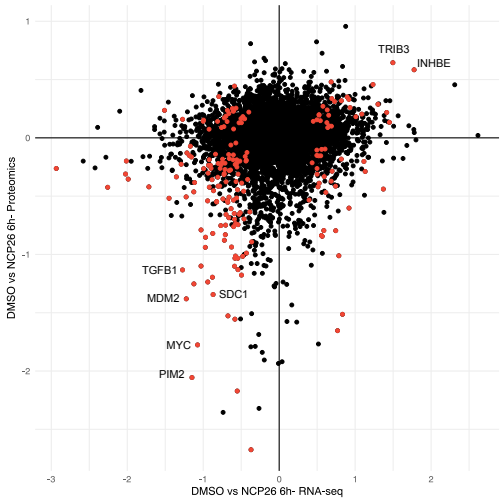


Figure 5.8: Scatterplot of proteomic and RNA-seq datasets depicting changes after 6 hr NCP26 exposure to AMO-1 cells. Red points indicate genes which are differentially expressed ($p_{adj} < 0.01$) in both RNA-seq and proteomic datasets. Folllowing NCP26 treatment, TRIB3 and INHBE (both *ATF4* targets) were the proteins with the highest increase in abundance ($log_2FC = 0.64$ and INHBE $log_2FC = 0.58$) in the proteomic dataset and both were significantly upregulated in the RNA-seq dataset. I analysed proteomic data supplied by <COLLABORATORS>.

starvation', 'REACTOME response of EIF2AK4/GCN2 to amino acid deficiency' and 'KRIGE amino acid deprivation' were collated from the Molecular Signatures Database (MSigDb), making up a list of 166 unique genes.

This list of AAR genes was intersected with DEGs for ProRS inhibitor-treatment (i.e NCP26 vs DMSO and halofuginone vs DMSO at 6 or 24 hours) and a heatmap was constructed (Figure 5.9). The AAR transcription factors *ATF3*, *DDIT3* (CHOP), *CEBPB* and *CEBPG* are all markedly upregulated following NCP26 and halofuginone treatment. Genes coding for aminoacyl tRNA synthetases are also shown to be upregulated, such as *WARS1*, *SARS1* and *CARS1*. Amino acid transporters, such as *SLC7A11*, were also upregulated following ProRS treatment.

A list of 287 unique genes activated by *ATF4* was compiled from the genesets: 'REACTOME ATF4 activates genes in response to endoplasmic reticulum stress' and 'ATF4 Q2' (genes having at least one occurrence of the transcription factor binding site V.ATF4 Q2 in the regions spanning up to 4 kb around their transcription starting sites). Figure 5.10 shows a heatmap for ATF4 activated genes and genes

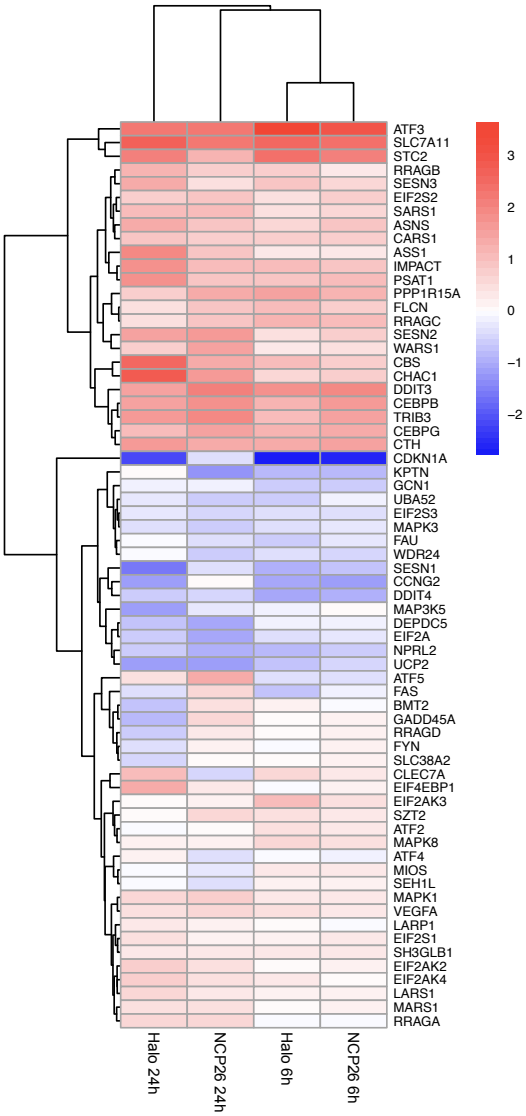


Figure 5.9: Amino acid starvation response (AAR) genes heatmap for WT cells. Differentially expressed genes (DEGs) from WT AMO-1 cells intersected with genes involved in the AAR. A list of known AAR genes was compiled by collating AAR genesets from the Molecular Signatures Database (MSigDb). The colour scale shows $\log_2$ fold change of expression for each treated sample, compared to its DMSO time control. Red indicates an upregulated gene and blue a downregulated gene.

DE by NCP26 and halofuginone. Numerous *ATF4* targets can be seen to be differentially expressed following ProRS inhibition with NCP26 and halofuginone. Of the 287 genes in the list, 165 genes were differentially expressed with NCP26 or halofuginone treatment.

Taken together, it is clear that following HF and NCP26 treatment of AMO-1 MM cells, the amino acid starvation response is activated, mediated via the transcription factor *ATF4* and culminates in ER stress and downstream apoptotic mechanisms.

ProRS inhibitors vs carfilzomib

Since their first use in MM, over 20 years ago, the mechanism of action of proteome inhibitors has been extensively studied and well described by researchers[253]. Figure 5.5d demonstrates similarities at 24 hours between the transcriptional effects of ProRS inhibitors and the proteasome inhibitor carfilzomib on AMO-1 cells. It also demonstrates ProRS and carfilzomib-treated samples separation at 6 hours, highlighting differences in their initial mechanism of action.

The similarities and differences between ProRS inhibitors and carfilzomib were studied in more detail. As seen by x-y trends of scatter plots and number of overlapping genes in venn diagrams (Figure 5.11a and b), ProRS inhibitors and carfilzomib treatment are more similar at 24 hours, and share more DEGs (compared to DMSO controls), than at 6 hours. Somewhat obviously, NCP26 and HF share more overlapping DEGs with eachother than with CFZ. HF shares more overlapping DEGs with CFZ than NCP26. This is likely to do with potency and dosing, where both HF and CFZ were used at approximately 10 times their IC_{50} value.

Carfilzomib treatment of AMO-1 cells resulted in a pronounced induction of the heat shock response, changes to ubiquitin mediated processes and protein folding, in line with the well-defined cellular changes of proteasome inhibition. At 6 hours, some ISR effectors, such as *ATF3* and *JUN*, are upregulated by both carfilzomib and NCP26/Halofuginone treatment.

At 24 hours, CFZ and ProRS inhibitor treatment seem to culminate in similar end-stage stress, cell cycle changes and apoptotic mechanisms. Both ProRS inhibition

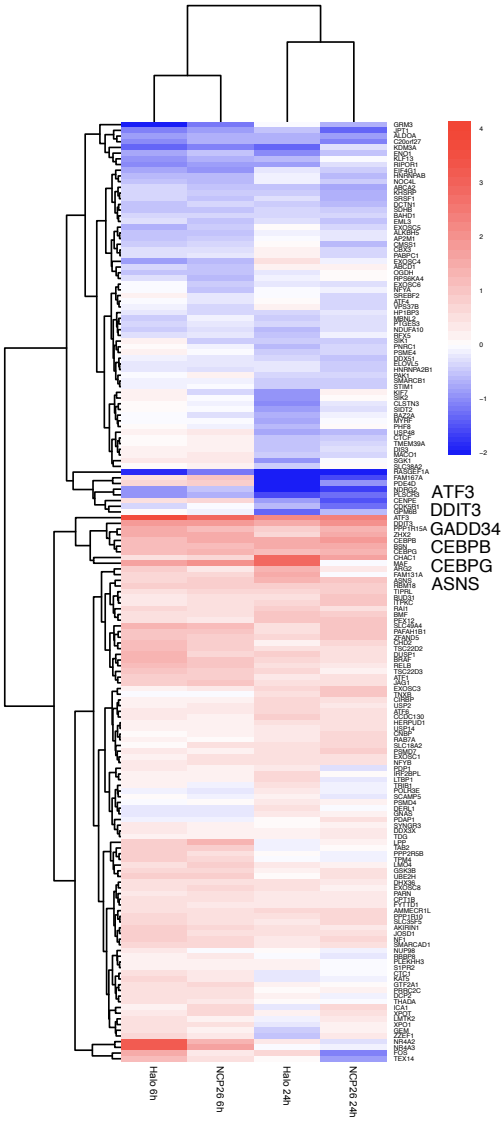


Figure 5.10: Heatmap of *ATF4* activated genes for WT cells treated with NCP26 and halofuginone. Differentially expressed genes (DEGs) from WT AMO-1 cells intersected with *ATF4* activated genes. A list of genes activated by the transcription factor *ATF4* was compiled by collating genesets from the Molecular Signatures Database (MSigDb). The colour scale shows $\log_2$ fold change of expression for each treated sample, compared to its DMSO time control. Red indicates an upregulated gene and blue a downregulated gene.

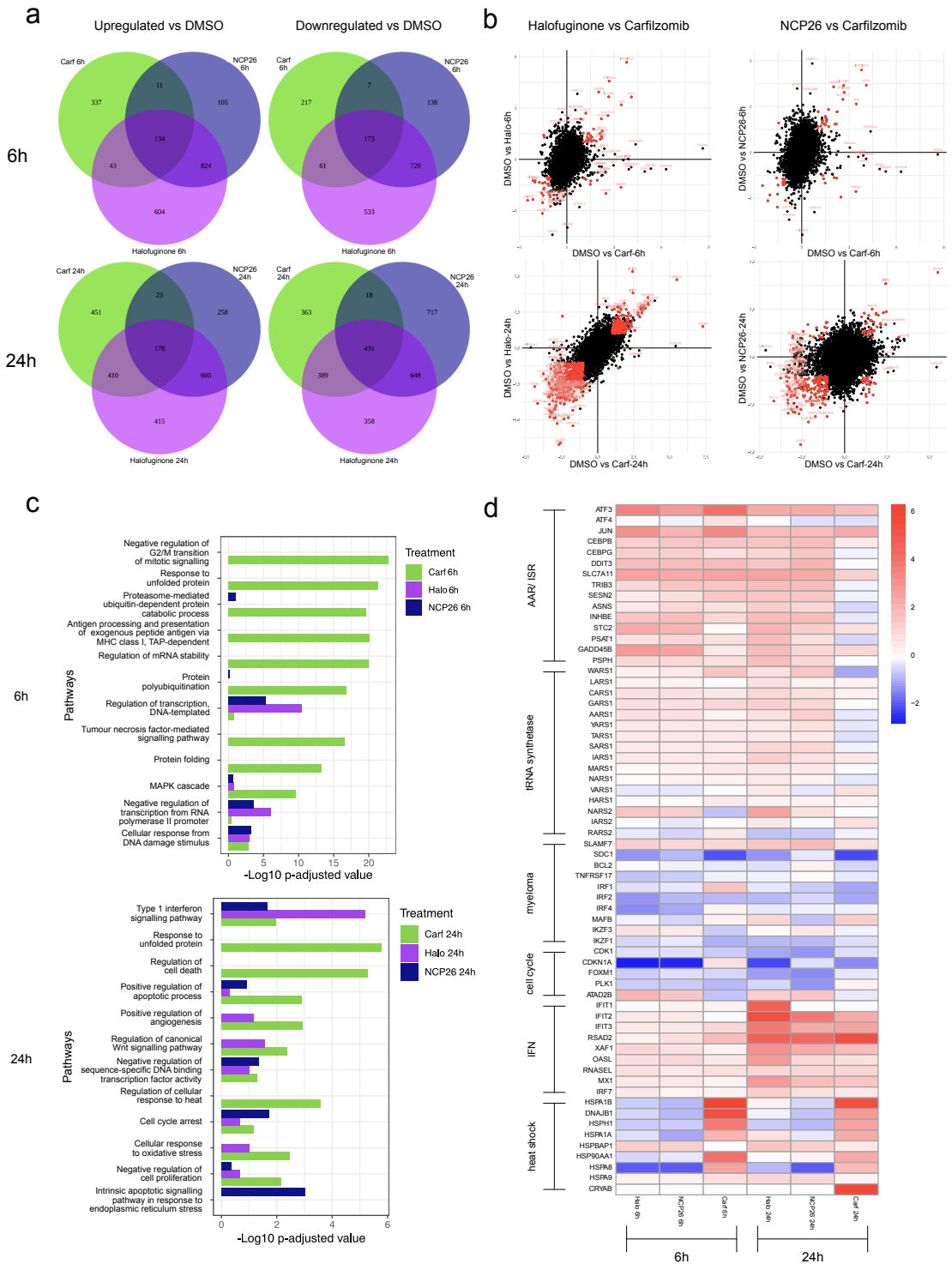


Figure 5.11: ProRS inhibitors compared with Carfilzomib's mechanism of action. a) Venn diagrams showing overlapping differentially expressed genes (DEGs; upregulated or downregulated following treatment)at 6 and 24 hours. b) Scatter plots showing correlation of carfilzomib DEGs against halofuginone or NCP26 DEGs. c) Pathway analysis (Gene ontology biological processes; GOBP) for top upregulated genes. d) Heatmap of selected differentially expressed genes upon carfilzomib, halofuginone and NCP26 treatment.

and CFZ treatment result in pathway enrichment of 'cell cycle arrest', 'negative regulation of cell proliferation', 'positive regulation of apoptosis process' and 'type 1 interferon signalling pathway'. Figure 5.11d shows a heatmap of NCP26, halofuginone and CFZ for $\log_2 FC$ compared to DMSO controls at 6 and 24 hours for selected genes, belonging to various pathways/ classes. NCP26 and halofuginone demonstrate similar effects on myeloma markers as CFZ, such as downregulating the MM pathological marker *SDC1/ CD138*. CFZ is an established anti-MM therapy approved in the clinic, therefore this is promising for the effectiveness of ProRS inhibitors in MM.

5.3.4 Carfilzomib-resistant cells

Differential expression

For CFZr L363 cells at 6 hours, 1165 genes were DE for NCP26 treated-samples, 2424 DEGs for halofuginone, 222 DEGs for MAZ1805, and 0 DE genes for NCP22-treated samples compared to DMSO control samples. At 24 hours, 852 DEGs for NCP26-treated samples, 256 DEGs for halofuginone, no genes were DE for MAZ1805 and NCP22 compared to DMSO controls. DEGs for NCP26 and halofuginone treatment are shown in Figure 5.12. Numerous genes involved in the AAR can be seen to be upregulated following NCP26 and halofuginone treatment, including *TRIB3*, *ATF3*, *CHAC1*, *INHBE*, *PSAT1*, *ASNS* and *SESN2*; Amino acid transporters *SLC7A11* and *SLC6A9*, and tRNA aminoacyl synthetase *WARS1* are also upregulated.

Pathway enrichment analysis

Pathway enrichment analysis was performed for the top DE genes for NCP26 and halofuginone treated samples compared to DMSO controls (Figure 5.13). Multiple pathways relating to endoplasmic reticulum stress and apoptosis were enriched following ProRS inhibition. 'Cellular repsonse to amino acid starvation' was enriched, along with 'amino acid transport', indicating that NCP26/Halofuginone likely activate the amino acid starvation response in the CFZ resistant MM cells. Additionally the pathway 'response to unfolded protein' was enriched. The unfolded

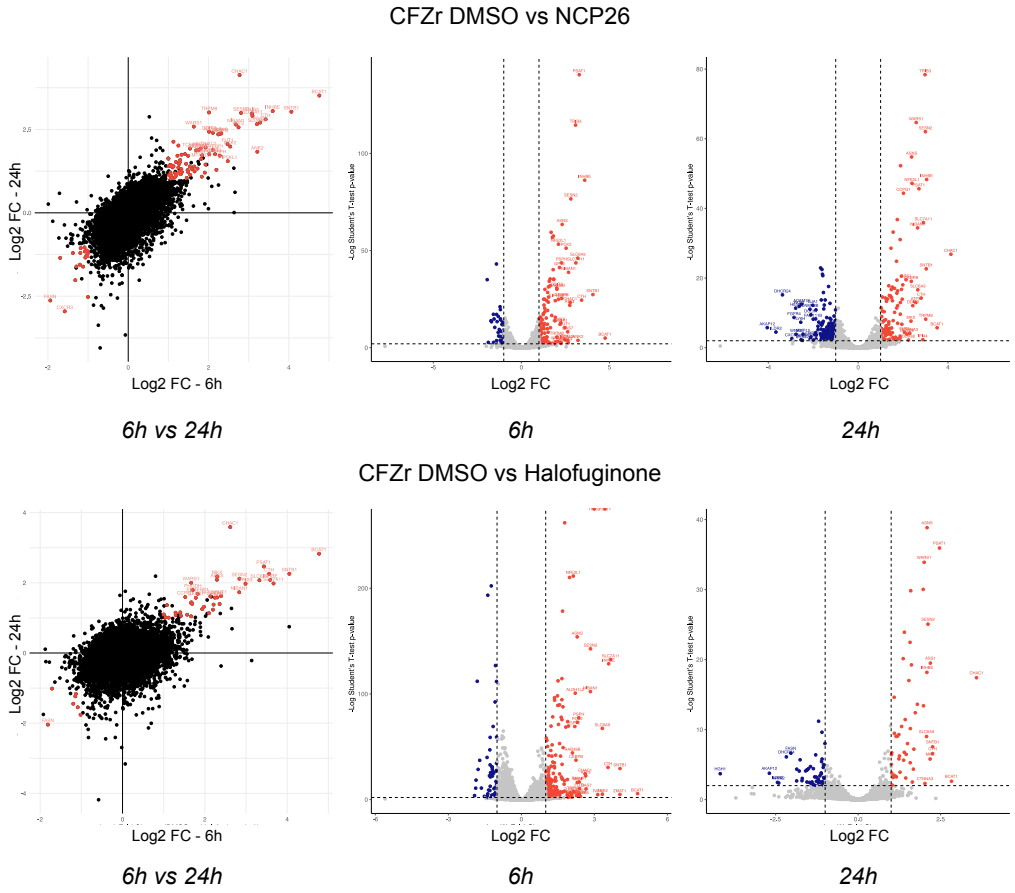


Figure 5.12: Differentially expressed genes (DEGs) for halofuginone and NCP26 treated CFZr L363 cells at 6 and 24 hours. Scatter plot of genes for 6 hours treated vs 24 hours treated. Red points indicate genes which are differentially expressed ($p_{adj} < 0.01$) at both 6 and 24 hours. Volcano plots are also shown. Blue points indicate downregulated DEGs ($p_{adj} < 0.01$ & $\log_2 FC < -1$). Red points indicate upregulated DEGs ($p_{adj} < 0.01$ & $\log_2 FC > 1$). Top DEGs are annotated with HGNC symbols.

protein response is part of the integrated stress response so shares many of the same effectors as the AAR, such as *DDIT3/CHOP*.

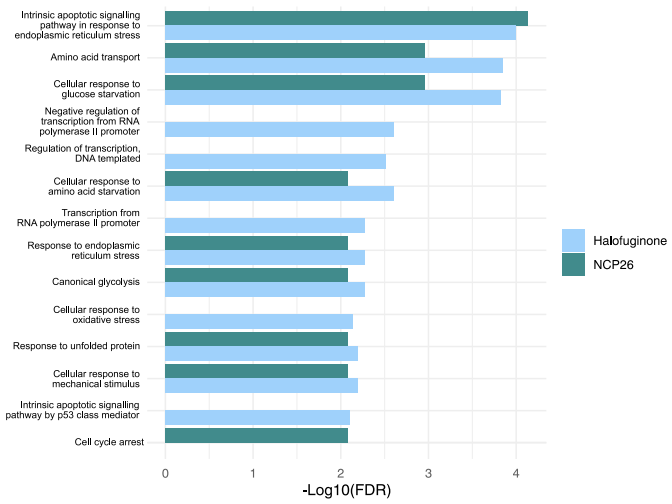


Figure 5.13: Pathway analysis for ProRS treated CFZr cells at 24 hours. Gene ontology biological processes (GOBP) was performed for the top DE genes.

The amino acid response was investigated more closely for the PI-resistant cell line. Figure 5.14 shows a heatmap of DEGs for NCP26/Halofuginone treated CFZr cells at 6 or 24 hours, which are known members of the AAR (collated from MSigDbl- as above).

A strong elicitation of the AAR can be seen following NCP26 and halofuginone treatment. The transcription factors *CEBPG*, *CEBPB*, *ATF3*, *DDIT3* are strongly upregulated, along with amino acid transporter *SLC7A11* and tRNA aminoacyl synthetase *WARS1*. Other AAR genes are strongly upregulated with ProRS inhibitor treatment, such as *ASS1*, *ASNS*, *CHAC1*, *PSAT1*, *PPP1R15A/GADD34* and *TRIB3*.

A list of 287 unique genes activated by *ATF4* was compiled from the genesets: 'REACTOME ATF4 activates genes in response to endoplasmic reticulum stress' and 'ATF4 Q2' (genes having at least one occurrence of the transcription factor binding site V.ATF4 Q2 in the regions spanning up to 4 kb around their transcription starting sites). Figure 5.15 shows a heatmap for ATF4 activated genes and genes DE by NCP26 and halofuginone in CFZr cells. Numerous *ATF4* targets can be seen to be differentially expressed following ProRS inhibition with NCP26 and halofuginone. Of the 287 genes activated by *ATF4* in the list, 157 genes were differentially expressed

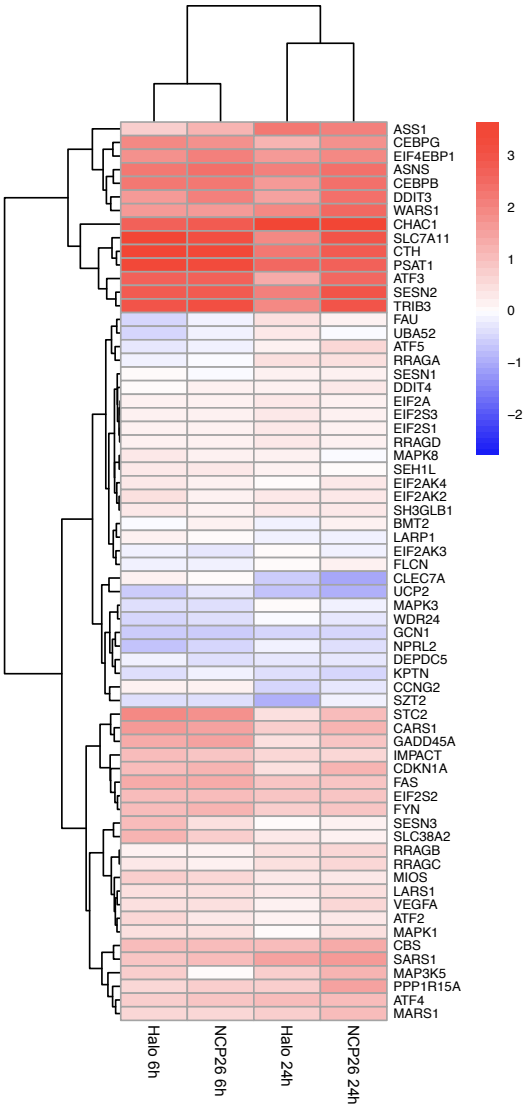


Figure 5.14: Amino acid starvation response (AAR) genes heatmap for CFZr cells. Differentially expressed genes (DEGs) from CFZr AMO-1 cells intersected with genes involved in the AAR. A list of known AAR genes was compiled by collating AAR genesets from the Molecular Signatures Database (MSigDb).

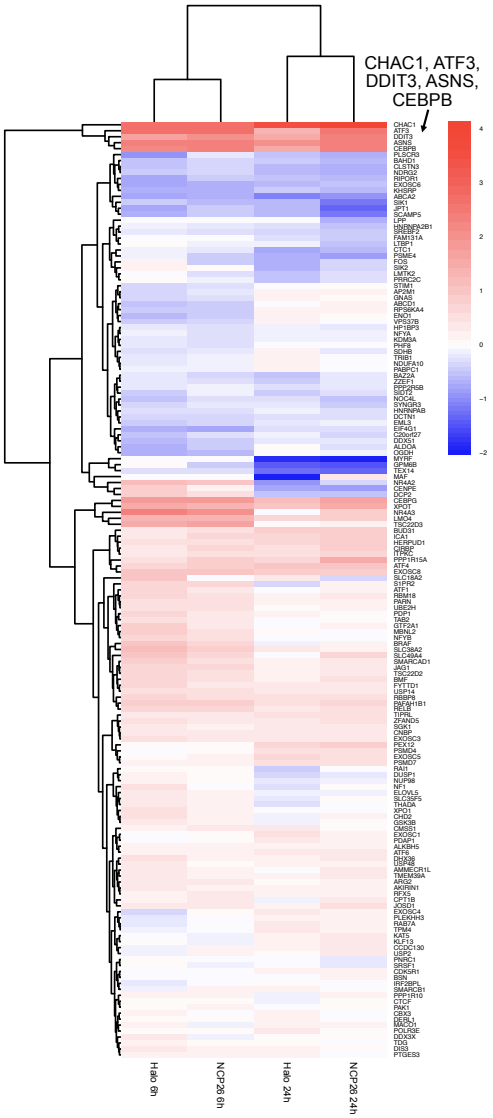


Figure 5.15: Heatmap of *ATF4* activated genes for CFZr L363 cells treated with NCP26 and halofuginone. *ATF4* activated genes, which are differentially expressed following NCP26 or halofuginone treatment at 6 or 24 hours. A list of genes activated by the transcription factor *ATF4* was compiled by collating genesets from the Molecular Signatures Database (MSigDb). The colour scale shows $\log_2$ fold change of expression for each treated sample, compared to the corresponding DMSO control. Red indicates an upregulated gene and blue a downregulated gene.

with NCP26 or halofuginone treatment. This indicates that many *ATF4* downstream targets are being mediated, indicating that *ATF4* has high transcriptional activity.

CFZ-resistant L363 cells show a similar engagement of the AAR and activation of *ATF4*-mediated genes as PI-sensitive AMO-1 cells, following NCP26/HF treatment. Additionally many downstream apoptotic mechanisms are enriched, indicating the initiation of cell death by AAR activation. Taken with the cytotoxicity dose response data, this demonstrates that ProRS inhibitors are capable of killing PI-resistant MM cell lines. Thus targeting the ProRS could be a potential effective strategy in overcoming PI drug resistance in clinical MM.

5.4 Summary

NCP26 and halofuginone have been shown to reduce cell viability of drug sensitive and PI-resistant MM cell lines in a dose-dependent manner. It has been shown that NCP26 and carfilzomib are not synergistically cytotoxic together. This reflects a previous study where HF and bortezomib were found to be moderately antagonistic in combination. Therefore, administering a PI and ProRS inhibitor in combination would likely offer no increased benefit to patients. This is likely due to similar mechanisms of activating ER stress pathways. However from the dose response data, it is clear that ProRS inhibitors are cytotoxic against PI-resistant MM cell lines. Indicating that their mechanism of actions are still distinct enough that NCP26 and HF have an effect on CFZ resistant cells. Together with the previous result of HF interacting synergistically with lenalidomide and dexamethasone[173], this could support a case for using ProRS inhibitors as a replacement for PIs once patients have relapsed and stop responding to proteasome inhibition.

From omics data, a general shift towards lower protein abundance and downregulation of gene expression has been demonstrated following exposure of MM cells to ProRS inhibitors. This indicates activation of components of the ISR to halt global protein synthesis (with the exception of *ATF4* target genes). *ATF4* target genes were shown to be upregulated and larger in abundance following NCP26 and halofuginone treatment. *ATF4* is the master regulator of amino acid metabolism. It is activated by amino acid deprivation. Halofuginone and NCP26 treatment caused increased expression of *ATF4* in both drug sensitive and PI-resistant MM cell lines. Expression of numerous genes involved in the AAR and ISR were markedly increased following halofuginone and NCP26 treatment. *DDIT3* and other downstream pro-apoptotic genes were over expressed following NCP26 and HF treatment. Western blot data from collaborators has demonstrated that NCP26 and halofuginone elicit canonical ISR activation with GCN2 and eIF2α phosphorylation in a dose-dependent manner (Figure A.1). Together this data shows that the ProRS inhibitors NCP26 and halofuginone activate the amino acid starvation response in MM cell lines. It also demonstrates that apoptotic pathways are activated following AAR activation,

indicating that the cytotoxic effects of NCP26 and halofuginone in MM cell lines are attributable (in part) to AAR activation and its downstream apoptotic mechanisms.

5.5 Discussion

Other therapeutics..

In a co-authored manuscript currently under review, it has been shown that numerous inhibitors targeting ProRS, as well as other aaRSs induce significant anti-proliferative effects across multiple MM cell lines (Figure A.2). Borrelidin, a natural threonyl-tRNA synthetase (ThrRS) inhibitor and CysSA, an amino acyl-adenylate analogue (which closely mimics the corresponding amino acyl-AMP intermediate) both reduced proliferative activity of MM cell lines. This shows that the anti-MM effects of Halofuginone and NCP26 may not be specific to only ProRS inhibitors, and this may extend to other aaRS inhibitors too.

Significant anti-proliferative effects across all MM cell lines inhibitors targeting ProRS Other aaRS inhibitors or proline specifically? A.2 Proline metabolism in cancer (https://www.intechopen.com/chapters/42308) Proline metabolism as a target in MM (https://jeccr.biomedcentral.com/articles/10.1186/s13046-022-02250-3) This study identifies PYCR1 as a novel target in bortezomib-based combination therapies for MM.

6
Single-cell RNA-seq analysis of ProRS inhibitors

6.1 Introduction

MM cells grow within the bone marrow and are supported as they grow by their microenvironment. The MM microenvironment comprises a cellular compartment (composed of immune cells, endothelial cells, osteoblasts, osteoclasts and stromal cells) and a non-cellular compartment (composed of the extracellular matrix (ECM), cytokines, chemokines and growth factors)[29, 30]. There are interactions between malignant plasma cells and the surrounding microenvironment. The bone marrow microenvironment has been indicated to play a supportive role in migration, proliferation, differentiation and drug resistance of malignant plasma cells. There is evidence linking the tumour microenvironment to progression of MGUS to active MM, for example significant matrix remodelling has been seen between the bone marrow of healthy individuals, MGUS and MM patients[30]. Therefore, to get an accurate picture of MM, information must be acquired about the surrounding niche.

Historically, the tumour environment has been investigated following the isolation of populations of cells sorted from the tumour and then sequenced using traditional microarray or bulk RNA-seq techniques. Bulk techniques measure the average expression across a sample, which is the sum of cell type specific expression weighted by cell type proportions. Single-cell techniques measure expression for each individual cell and therefore provide information on clonal diversity that may

be lost when pooling cells into bulk samples. Furthermore, multiple myeloma is an extremely heterogeneous disease, this is seen both between patients and within an individual's own tissue. Applying single-cell techniques to capture the inter- and intra-individual heterogeneity is fundamental to identifying molecular and cellular signatures that define MM.

The advent of single-cell technologies has led to a better understanding of the complexity and diversity of the tumour microenvironment. Seminal papers from Melnekoff et al. (2017)[254] and Ledergor et al. (2018)[255] use scRNA-seq to reveal clonal transcriptomic heterogeneity in MM samples. Melnekoff et al. (2017) demonstrated the clonal heterogeneity within MM using samples that were collected from eight relapsed MM patients. The group performed t-SNE clustering analysis and the samples separated into eight transcriptionally distinct clones, each corresponding to a different patient. This highlights the inter-patient differences of MM. Ledergor et al. (2018) performed a similar study to evaluate clonal heterogeneity within MM but also had a set of controls with which to compare the MM group. They found that MM patients have greater inter-individual transcriptional variation, where each MM patient possessed a unique and individual plasma cell transcriptional program. They also showed substantial intra-tumour heterogeneity (subclonal structures) of plasma cells in a third of their MM patient cohort. These papers established the importance of using single-cell techniques to study MM, as to not miss the underlying clonal heterogeneity. However both of these papers focussed soley on plasma cells and did not look at the surrounding bone marrow microenvironment. To truly understand the complexities of MM and treatment of MM, interactions between plasma cells and the bone marrow niche must also be explored using single-cell techniques.

This chapter uses MM patient-derived bone marrow (BM), from both newly-diagnosed and relapsed patients, to investigate the transcriptional effects of ProRS on MM cells and their surrounding immune microenvironment.

6.1.1 Experiment overviews

Three single-cell experiments, comprising samples from four MM patients, were performed to explore the effect of various compounds (including Halofuginone and NCP26) on MM patient tissue. The BM samples for experiments 1 and 2 were obtained from two treatment-naive, newly-diagnosed MM patients. Experiment 3 comprised samples from two relapsed MM patients, therefore both presenting with a degree of acquired anti-cancer drug resistance. For experiment 1, BM samples were treated for 24 hours with 1µM Casin, GSK-J4, Halofuginone, NCP26, SGC-GAK, Verteporfin or a DMSO control, totalling 7 samples. For experiment 2, BM samples were treated with with 1µM CAMKK2, CLK or CSNK2 for 24 hours; 1µM SGC-GAK, Halofuginone, NCP26 or a DMSO control for 24 and 48 hours, totalling 11 samples. For experiment 3, BM samples from patient 3 and 4 were treated for 24 hours with either a DMSO control, 1µM Halofuginone, 1µM NCP26 or 5µM NCP26, totalling 8 samples.

Following compound treatment, scRNA-seq library preparation was performed by Dr Martin Philpott as outlined in section 3.5.2. <GENEWIZ experiment 3 >

6.2 Data processing

Initially all four patient samples were processed and integrated together. However, integration was found to be poor between treatment-naive patients and the relapsed patients. This was expected as MM patients' transcriptome has been shown to change considerably following successive rounds of treatment cycles. Therefore, samples from experiments 1 and 2 (treatment-naive patients 1 and 2) were integrated together and samples from experiment 3 (relapsed patients 3 and 4) were integrated together.

Experiments 1 and 2 contained numerous samples that were not of interest to this project. Yet, all samples originated from MM patient tissue, therefore all 18 samples were included in the analysis pipeline, and were included in integration and annotation steps. These irrelevant samples were included to increase the granularity

of the data, and allow for easier annotation of clusters. Downstream analyses was only performed for DMSO, Halofuginone and NCP26 samples.

6.2.1 Analysis overview

The single cell analysis pipeline outlined in Section 4.2.3 was used to process the scRNA-seq data. Kallisto BUS/ BUStools was used to pseudoalign reads and quantify gene expression. Next, quality control and filtering of the samples was performed. Poor quality cells are likely to have a low number of genes and UMIs per cell. So any cells with fewer than 500 UMIs were removed. Cells with a gene count below 300 or above 6000 were also removed. Cells with a mitochondrial ratio higher than 0.1 were removed (a high proportion of mitochondrial genes indicates mitochondrial contamination from dead or dying cells).

After quality control and filtering of the data, clustering was performed using Seurat v3, followed by integration of all samples, using Seurat and Harmony. Using the Seurat SCTransform normalisation method of integration, the two experiments were too large to integrate across all samples. Therefore, a reference-based approach was taken, whereby a subset of samples were selected (based on their cell richness and relevance to the research question) and listed as 'reference datasets' for SCTransform normalisation. Harmony integration was also implemented, using 'Patient Number' and 'Experiment Number' as additional covariates.

Cell type annotation was performed with several automated packages (singleR, clustifyr and scClassify), and then fine-tuning manually using a list of known biological markers. The HumanCellAtlas (HCA) database was used to inform singleR and clustifyr annotation. scClassify was performed using a pre-trained scClassify model, based on seven peripheral blood mononuclear cell (PBMC) scRNA-seq datasets (including data from 10Xv2, 10Xv3, smartSeq, celSeq, dropSeq and inDrops). As the reference datasets were based on healthy tissue, they were unable to label pathological cells, like MM cells. Myeloma cells were identified manually using a range of known markers, for example: *CD38*, *CD138*, *SLAMF7* and *BCMA* (see Table A.1).

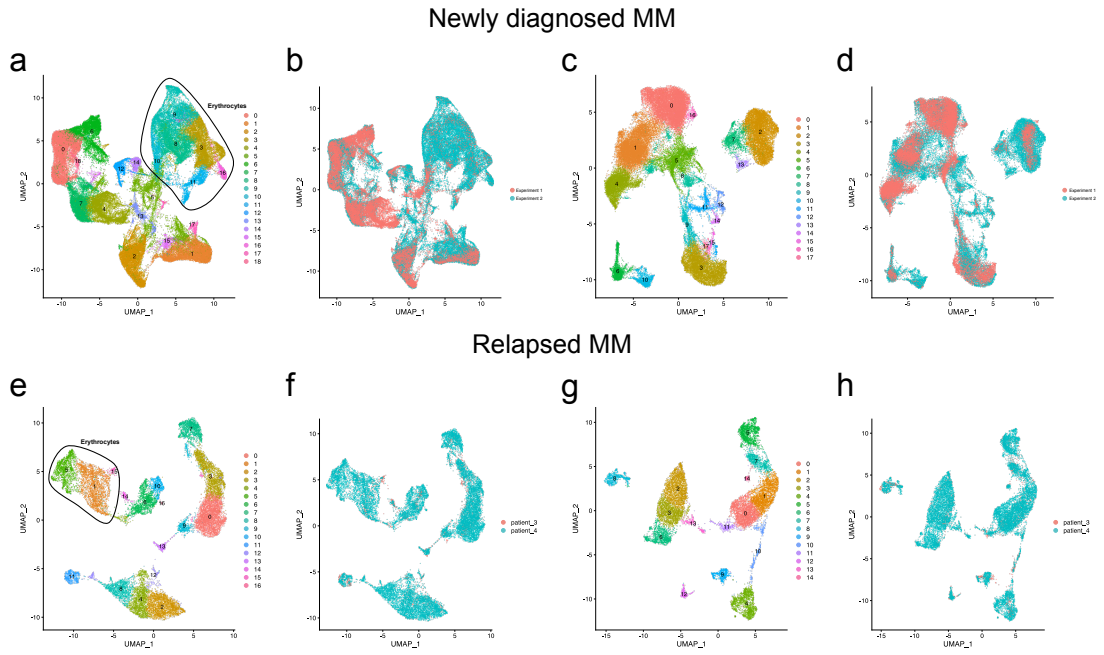


Figure 6.1: UMAP dimension plots following integration of samples from experiment 1 and 2 (treatment naive patients), and samples from patients 3 and 4 in experiment 3 (relapsed MM). [a-d] Experiment 1 and 2-newly diagnosed MM patients. [e-h] Experiment 3, patients 3 and 4- relapsed MM patients. [a, b, e, f] Integrated UMAP plots before erythrocyte cell and gene removal. Erythrocyte clusters are circled in a) and e). [c, d, g, h] UMAP plots following removal of erythrocyte cell clusters and genes and re-integration of samples. [b, d, f, h] show the composition of each dataset by experiment or patient.

Experiment 2 was found to have an extremely high erythrocyte population (Figure 6.1a and b). In addition, many other cell populations were expressing erythrocyte-specific genes, where we would not expect to see them expressed, for example MM cells expressing numerous haemoglobin subunit genes. Many of the variable features that Seurat uses for clustering and dimension reduction were made up of these erythrocyte-specific and haemoglobin genes. The high expression of these genes was affecting the integration of the two experiments together. A theory for the presence of the large number of erythrocytes and un-localised erythrocyte gene expression is that perhaps the BM sample taken for experiment 2 was one of the later samples taken from the patient and contained a large amount of blood. Library prep clean-up may have missed many of these cells, leading to ambient erythrocyte RNA being present within many droplets. Other MM scRNA-seq studies have observed similar contamination of erythrocyte specific genes, such as *HBB*, *HBA1* and *HBA2*,

in non-blood cell populations, despite performing a red blood cell lysis step[256].

It was decided to remove the erythrocyte clusters (clusters 3, 8, 9, 10, 11 and 16 in newly-diagnosed; clusters 1, 5 and 15 in relapsed) and haemoglobin-related genes or erythrocyte-specific genes that were dominating expression in the integrated dataset. After the integrated Seurat object had the erythrocyte genes and cells removed, it was split back up into separate Seurat objects for each sample, and integration and clustering was performed again. Seurat's SCTransform with reference datasets and Harmony (using a multi-covariate model, accounting for each different sample and the two different experiments) were used to re-integrate all samples. Harmony integration was found to integrate clusters across patients and experiments better than using Seurat's SCTransform. The Harmony-integrated datasets were taken forwards and used for cell type annotation and further downstream analyses. Better integration was achieved after removing erythrocytes and erythrocyte-specific genes (see Figure 6.1d).

A large erythrocyte component was also found for patient 4 in the relapsed dataset (Figures 6.1e and 6.1f). The same analysis workflow outlined above was applied to experiment 3, removing the erythrocyte cluster and erythrocyte-specific genes and re-integrating using Harmony with samples and patients as covariates.

The total number of cells after each data processing step for each patient and experiment is summarised in Table 6.1. The number of cells originating from

Experiment	**Patient**	**Total cells**	**Cells passing filter**	**Cell number after erythrocyte removal**
Experiment 1	Patient 1	112452	25779	23915
Experiment 2	Patient 2	462560	61059	37161
Experiment 3	Patient 3	4894	2625	2257
	Patient 4	21682	18674	14934

Table 6.1: Total cells, the number of cells passing filter, and the number of cells passing filter once erythrocyte clusters were removed across all samples for each patient.

patient 3 after filtering might be a barrier to interpreting meaningful results. Patients 1, 2 and 4 have sufficient cell numbers to perform downstream analysis, for example differential expression.

6.2.2 Annotation of re-integrated data

R packages clustifyr and scClassify were used to aid in cell type annotation of the integrated datasets, the result of this annotation can be seen in Figure 6.2.

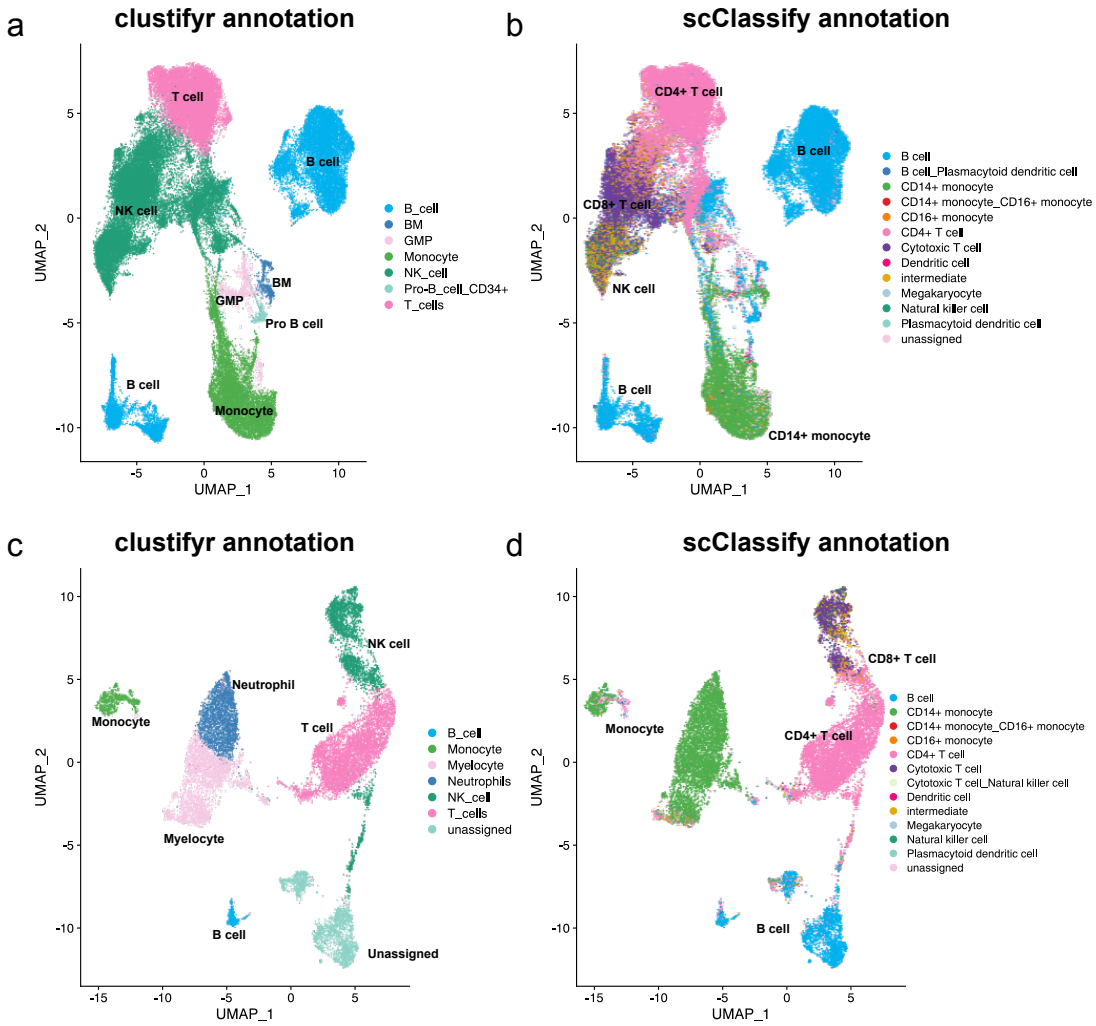


Figure 6.2: Automated annotation of scRNA-seq cell clusters, using the R packages clustifyr and scClassify in combination with reference datasets. [a, b] newly-diagnosed MM, [c, d] relapsed MM. The output of automated packages clustifyr and scClassify is used to aid cell type annotation. Clustifyr assigns a cell type to each cell cluster, whilst scClassify assigns a cell type to each individual cell. Clustifyr was used in conjunction with the HumanCellAtlas reference. scClassify was ran using a pre-trained model trained on seven PBMC single cell RNA-seq datasets. Both references originate from healthy data, therefore neither are able to identify the myeloma cell population.

Automated annotation using computational methods gives a good starting point for more detailed manual annotation. Currently, these methods are not a complete

substitute for manual annotation using biological knowledge of cell markers. Both cell-type references originate from healthy tissue, therefore the myeloma cluster could not be identified using clustifyr, scClassify and these healthy references alone. Both packages either incorrectly labelled the MM cluster as B cells, or were unable to assign any cell type to the myeloma cluster with any confidence and labelled it unassigned. Thus, the MM clusters had to be identified manually using MM biological knowledge. Section 4.5 constructs a MM classifier reference to overcome this problem for future MM scRNA-seq datasets.

Figures 6.3 and 6.4 illustrate MM cell identification for the naive and relapsed datasets, respectively, using gene expression of certain MM markers. For MM cells, you would expect to see: expression of *CD38*, but lower expression than in normal plasma cells; high expression of *CD138* in MM cells, as well as high expression of *SLAMF7*, *BCMA*, *KRAS*, *IGKC* and *IGL2*, however these are not exclusive to the MM cluster. You would expect to see little or no expression of *CD45* and *CD19* in the MM cluster and reduced expression of *CD20* in MM cells compared to normal B cells. You may also expect to see expression of *CD56* in abnormal plasma cells, however this is not always detected at the RNA level. Other common markers used to identify MM cells include, *CD81*, *B2M* and *CD117*. Using the expression of a combination of these markers, clusters 2, 7 and 13 were identified as the MM cell population in the newly diagnosed dataset and cluster 4 was identified as MM cells in the relapsed MM dataset.

However, from marker expression alone, the identity of clusters 9 and 12 in the relapsed dataset were unclear. Both clusters 9 and 12 seem to belong to a B-cell lineage. Cluster 12 shows increased *CD20* expression, indicating they could be mature B cells. However, clusters 4 and 9 have an increased expression of *IGKC*, compared to non-B cell clusters. Indicating that perhaps, all B-cell clusters (4, 9 and 12) in the relapsed dataset, may be in fact MM cells. It has previously been shown that MM possess a level of plasticity, that could allow de-differentiation into other cell lineages[257], this may account for the *CD20* expression of cluster 12.

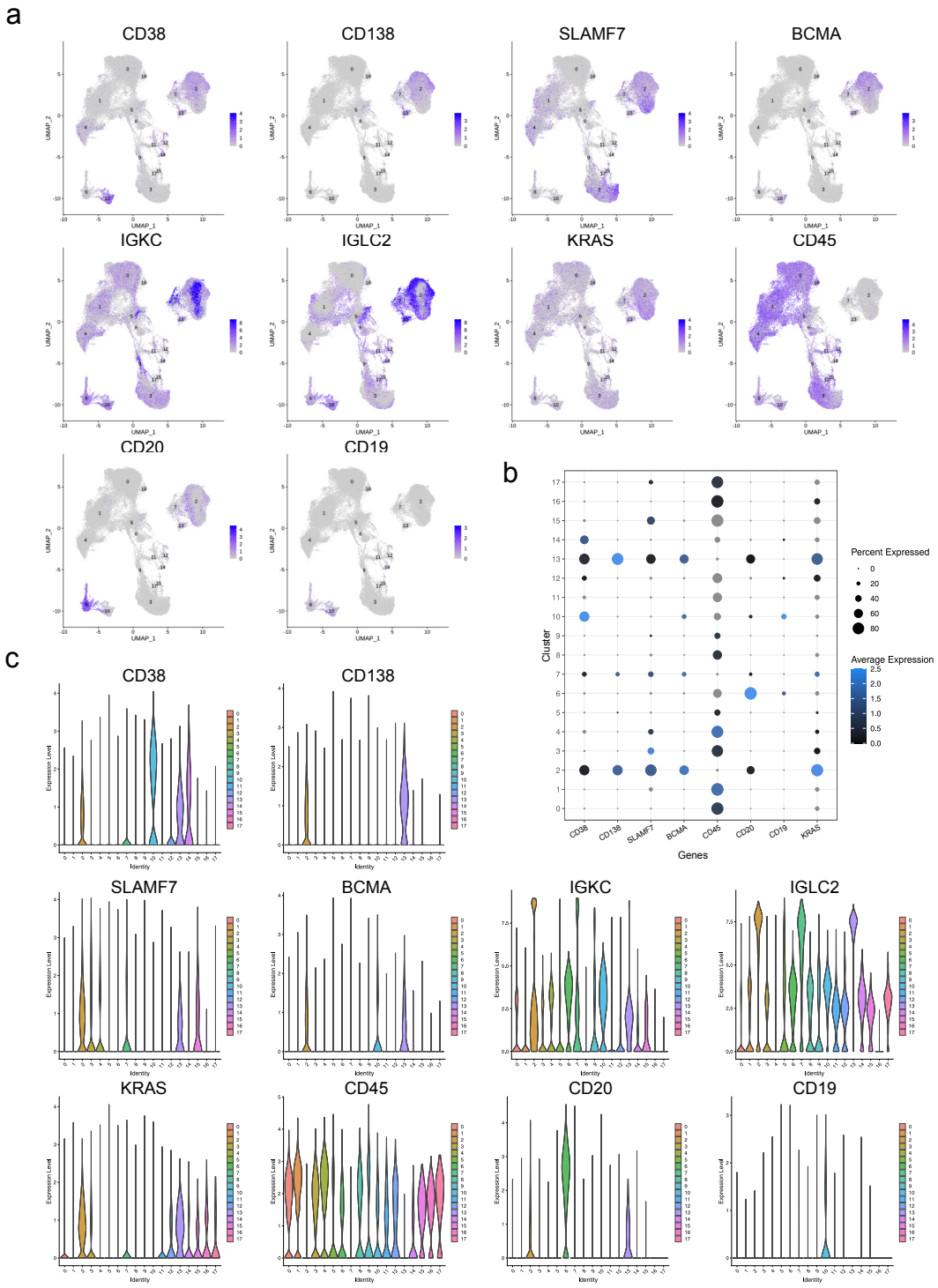


Figure 6.3: Manual annotation of multiple myeloma (MM) cell clusters in experiment 1 and 2 (newly-diagnosed patients) using known MM biological markers. a) UMAP feature plots of MM marker expression. Grey indicates no expression and purple indicates expression. b) A dot plot of the percentage of cells expressing a given marker and the average expression (for each cluster). c) Violin plots of gene expression for each cluster.

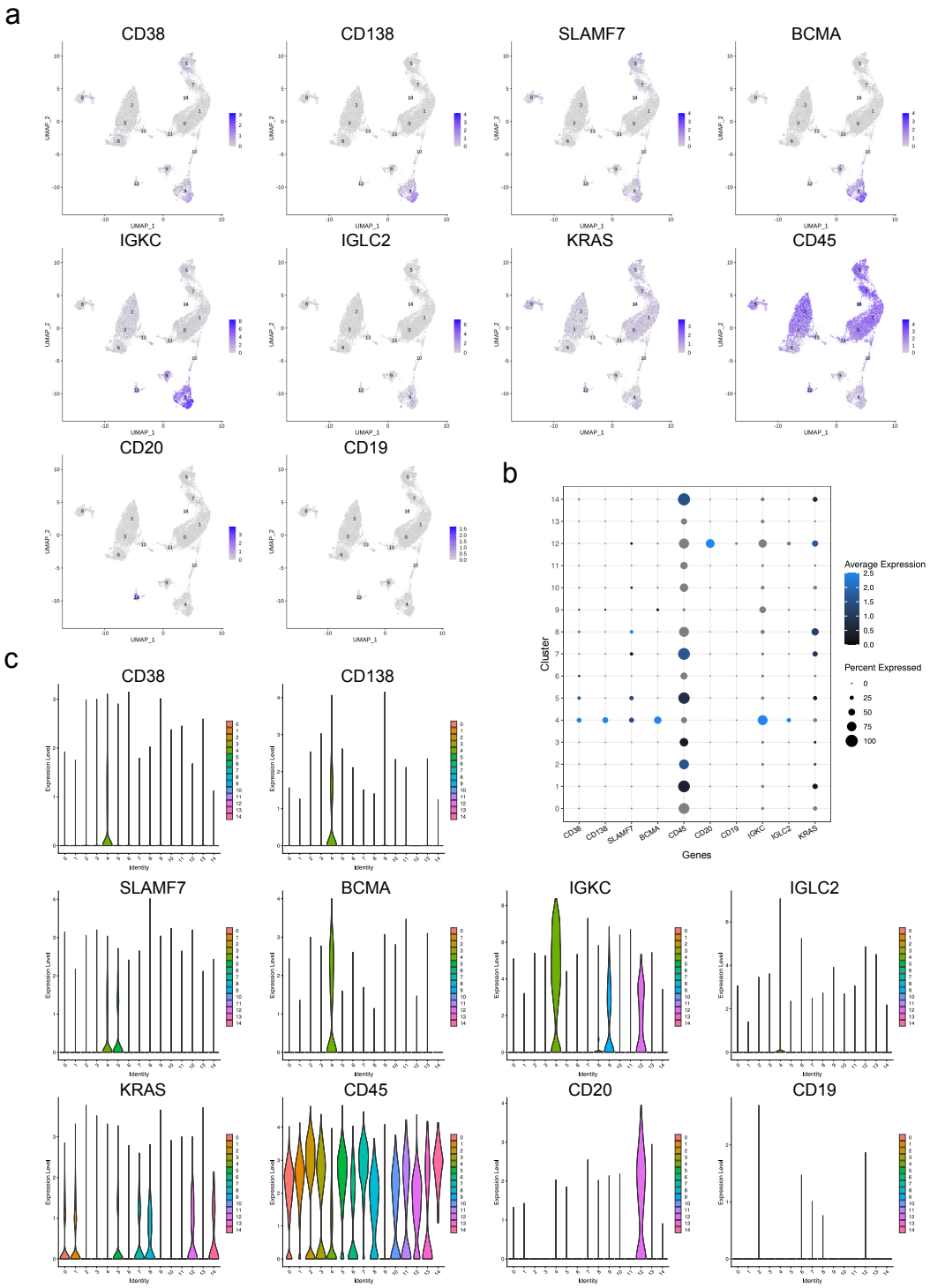


Figure 6.4: Manual annotation of MM cell clusters in experiment 3 (relapsed MM) using known biological markers. a) UMAP feature plots of MM marker expression. b) A dot plot of the percentage of cells expressing a given marker and the average expression (for each cluster). c) Violin plots of gene expression for each cluster.

Due to the uncertainty of cell-type identity in the relapsed dataset based on MM marker expression alone, inferCNV[258, 259] was employed to detect large-scale chromosomal copy number variations (CNVs) and help inform annotation. The raw counts expression matrix of the relapsed, integrated Seurat object was used as input for inferCNV. 'Normal' cells (that is: myeloid cells, T cells and NK cells) were used as references for for the relative expression intensity of the suspected malignant cell clusters (clusters 4, 9 and 12). A Cutoff value of 0.1 was used for the analysis, to account for the sparsity of droplet-based 3' scRNA-sequencing. Figure

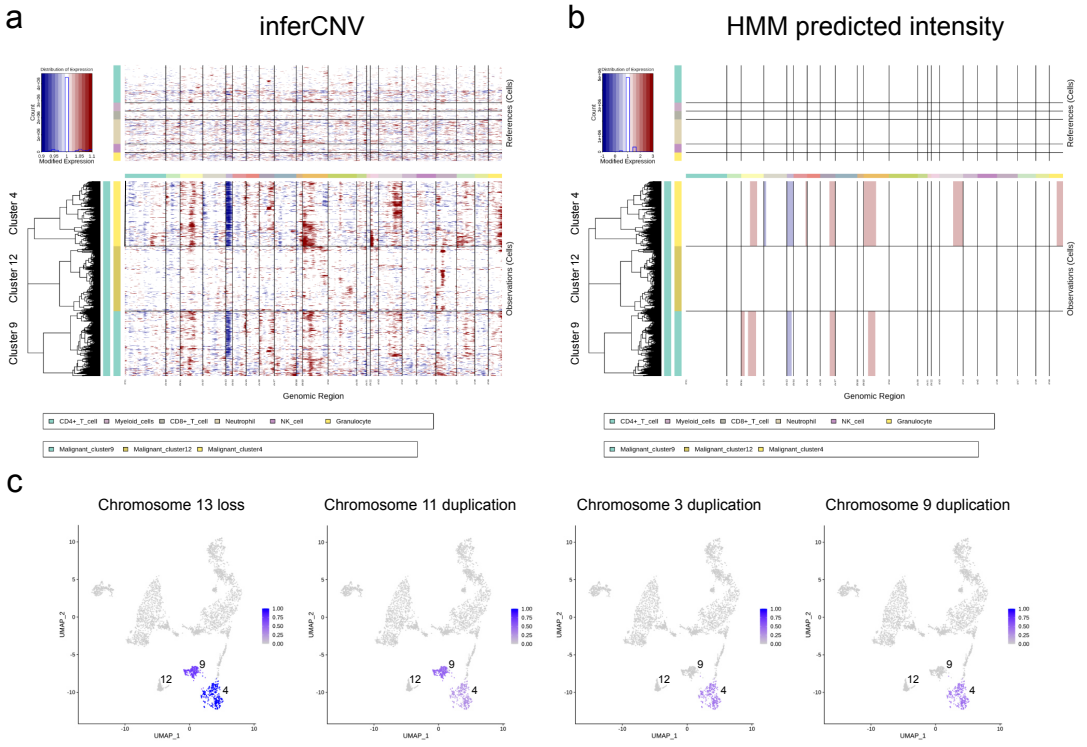


Figure 6.5: InferCNV results for the relapsed MM dataset. [a and b] InferCNV heatmaps. The top panel shows expression values for the reference 'normal' cells. The bottom panel shows expression values for the suspected malignant cells (clusters 4, 9 and 12). Red indicates chromosomal region amplifications and blue indicates chromosomal region deletions. a) De-noised inferCNV results. b) Hidden Markov-Model (HMM) copy number variation (CNV) region predictions. Chromosomal deletions predicted for chromosome 13 and chromosomal duplications/gains predicted in chromosomes 11 and 19 for clusters 4 and 9. Gains also predicted in chromosome 3 and 9 for cluster 4. c) UMAP featureplots demonstrating chromosome 13 loss and chromosomes 3, 9 and 11 duplications.

6.5 shows the inferCNV results for the relapsed dataset. Myeloma-defining CNVs were seen in both cluster 4 and cluster 9, including chromosome 13 deletion and

duplications of chromosomes 11 and 19. The inferCNV results also demonstrate that there are different genetic MM subclones within the population, as cluster 4 alone shows regional gain of chromosomes 3, 7, 9 and 16. This fits with the <FISH - ask martin for ORBID> data for the relapsed patients, whereby they had chromosome 13 deletions... Therefore, it can be concluded that clusters 4 and 9 are subclonal MM populations.

InferCNV was also used on the naive dataset (Figure A.3) to check for CNVs. The inferCNV results confirmed the gene expression-based identification of clusters 2, 7 and 13 as the MM cell population.

6.3 Results

6.3.1 Newly-diagnosed MM

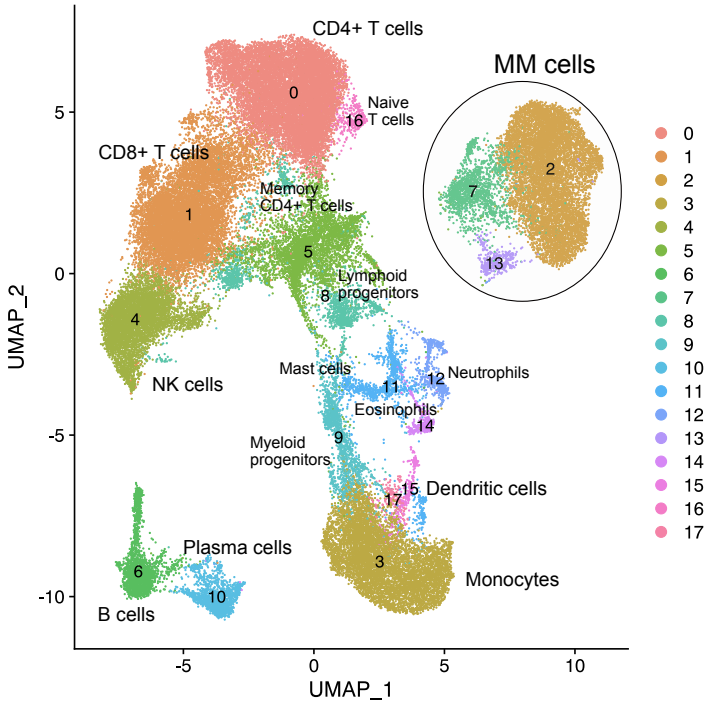


Figure 6.6: Fully annotated UMAP clustering analysis of two newly-diagnosed multiple myeloma (MM) patients. The MM cell population (circled), consists of three distinct clusters.

Figure 6.6 shows final cell-type annotation for the newly-diagnosed MM integrated dataset. 18 distinct clusters were identified using Seurat embeddings. The expected major immune clusters were identified (such as B cells, T cells, NK cells and myeloid cells.) Using the established MM biological markers (shown in Figure 6.3), three distinct MM clusters (2, 7 and 13) were identified.

Composition

Cluster composition analysis by treatment condition is shown in Figure 6.7. Halofuginone treatment at both 24 and 48 hours reduced the proportion of cells in the MM cluster ($p<0.00001$) compared to DMSO, for both experiment 1 and 2. NCP26 treatment at 24 hours reduced the proportion of cells in the MM cluster ($p<0.00001$), for both experiment 1 and 2. HF and NCP26 treatment increased the proportion

of cells in the CD4+ T cell, CD8+ T cell and B cell clusters. Together with dose response curves and cell death assays, this suggests that Halofuginone and NCP26 are selectively killing MM cells to a higher degree than other cell types.

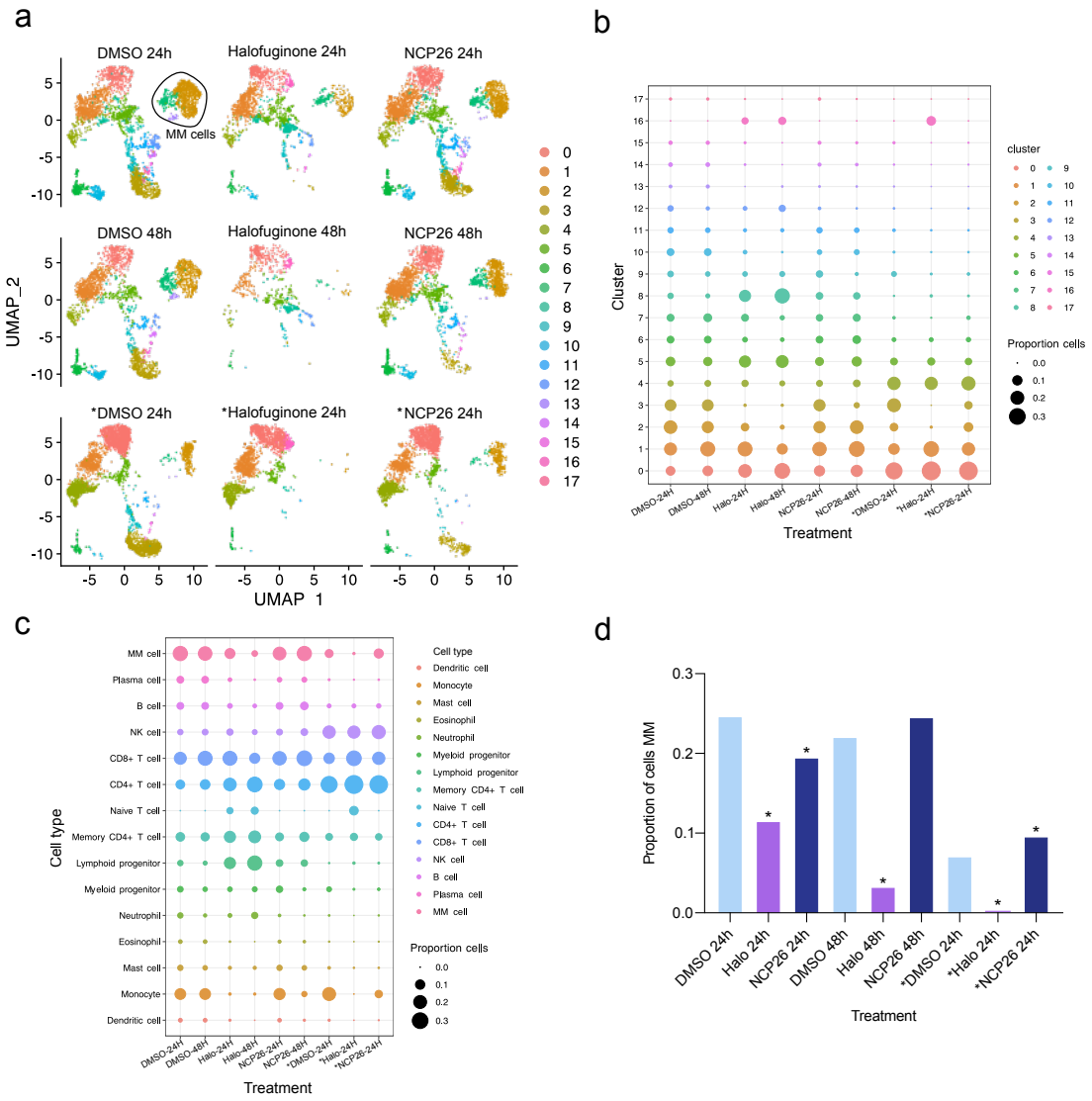


Figure 6.7: Composition analysis of newly diagnosed MM. a) UMAP cell composition plots separated by treatment condition. b) Dot plot showing proportion of cells in each cluster for each sample. c) Dot plot showing proportion of cells in each cell class for each sample (as labelled in Figure 6.6). d) The proportion of cells in the MM cluster only (stars above bars indicate significant at $p<0.01$ compared to corresponding control). Sample names starting with asterisks originate from experiment 1, no asterisk indicates experiment 2 origin.

Halofuginone treatment and 48-hour NCP26 treatment were also found to reduce the proportion of cells in the monocyte and neutrophil cluster ($p<0.00001$),

indicating they may have some off-target effects on myeloid cells.

All of the compounds were used at a concentration of 1µM. This is approximately 10 times the concentration of Halofuginone's IC_{50} value on MM cell lines and two times NCP26's IC_{50} value. As you can see, large number of cells were killed by Halofuginone treatment at both 24 and 48 hours. Therefore, the effects seen on the myeloid cells could be due to high dosing of the ProRS inhibitors.

Differential expression

Next, differential gene expression was investigated using Seurat's FindMarkers function. At the 48 hour time point substantial cell death is seen, therefore the 24-hour time point will be mainly focussed on.

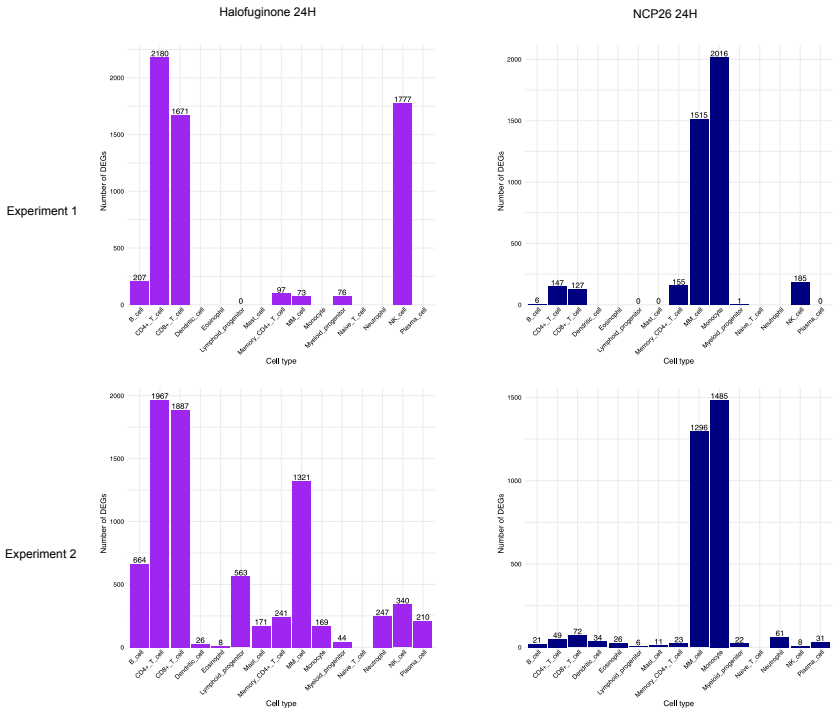


Figure 6.8: Number of differentially expressed genes (DEGs; $p_{adj} < 0.05$) broken down by cell type for two newly-diagnosed MM patients treated with ProRS inhibitors (Halofuginone and NCP26) for 24 hours. Cell type annotation corresponding to Figure 6.6. Experiment 1 and experiment 2 denote separate experiments, each containing BM samples from different newly-diagnosed MM patients.

Following 24-hour 1µM NCP26 treatment, 1515 genes were differentially expressed (DE; $p_{adj} < 0.05$) in the myeloma cell population in experiment 1, and

1294 genes in experiment 2, compared to DMSO treatment. Figure 6.8 shows the breakdown of DEGs per cell type for NCP26 and HF treatment. 24 hour 1µM NCP26 treatment had very little transcriptional effect on many of the other immune cell types, for example T, B and NK cells, where the number of DEGs is an order of magnitude less than for MM cells. This corroborates the composition analysis, where MM cells seem more sensitive to ProRS inhibition than other immune cell types. However, 2016 and 1485 genes were DE in the monocyte cluster following NCP26 treatment in experiment 1 and 2, respectively. This reflects the composition analysis, where the proportion of cells in the monocyte cluster was markedly reduced. This could indicate that NCP26 is not selective for MM cells over myeloid cells.

In MM cells, 73 genes were DE following Halofuginone treatment in experiment 1, and 1318 genes in experiment 2. This reflects results seen in the composition analysis (Figure 6.7), whereby we saw very few cells remaining in the MM cluster for experiment 1. Due to the low cell number, there is less statistical power and diversity across the cells, therefore you would likely see fewer statistically significant DEGs for this cluster. Unlike NCP26 treatment, for 24 hour 1µM HF treatment we see a large number of DEGs in B, T and NK cell clusters. This likely reflects the differences in potency between NCP26 and HF. HF is approximately five times more potent than NCP26. This fits with composition analysis, where HF showed a great deal of cell killing in the myeloid and MM cluster.

At lower doses (i.e. NCP26 at 1µM) of ProRS inhibition we see clear evidence of greater transcriptional effects on MM cells and monocytes over other immune subtypes. However at higher doses (i.e. Halofuginone at 1µM) we see substantial cell killing of MM cells and monocytes and larger transcriptional effects on other immune subtypes. This together with the composition analysis, demonstrates that NCP26 and Halofuginone are selective for MM cells over most immune cells, except for myeloid cells. However, this could be a problem with the doses used for treatment. 1µM is almost 10 times greater than Halofuginone's IC_{50} and two times greater than NCP26's IC_{50}. This experiment should be performed at a lower concentration to ascertain if NCP26 and Halofuginone are more selective for MM

cells or monocytes. Ideally, the experiment would be performed over the course of three days at a lower concentration, as in our cell line studies. However, human bone marrow samples do not last very well in extended culture, therefore acute treatment was the preferred method.

Pathway analysis

Pathway enrichment analysis was performed for the top 200 upregulated and top 200 downregulated DEGs in the MM cluster, following Halofuginone and NCP26 treatment (Figure 6.9).

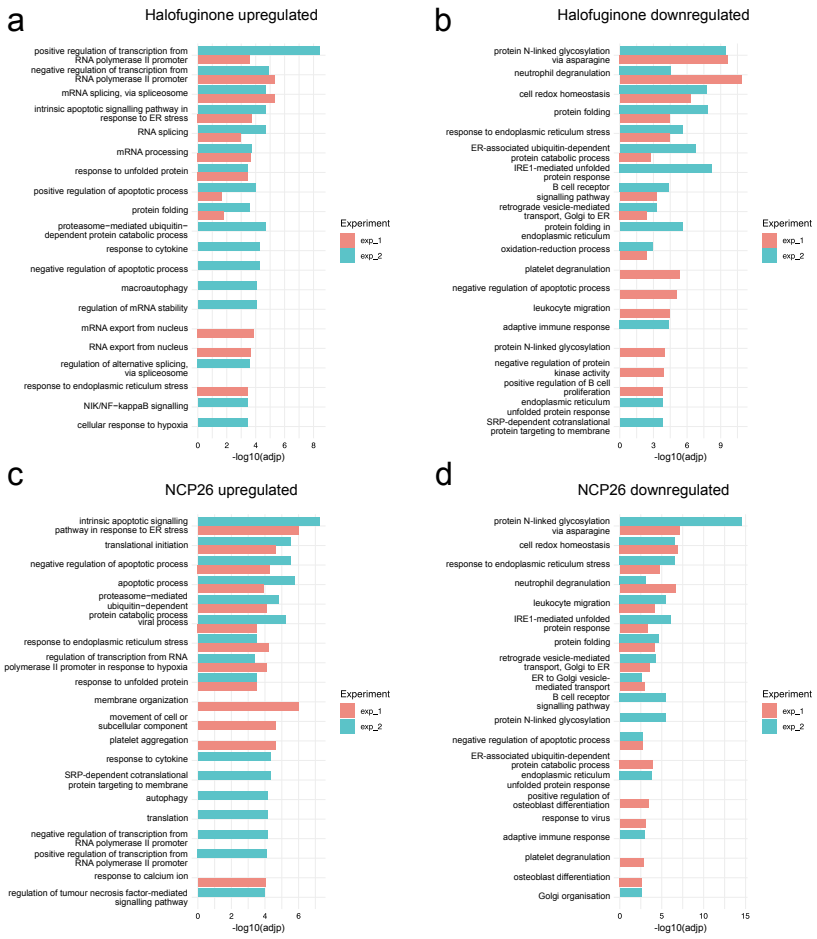


Figure 6.9: Pathway analysis (Gene ontology biological processes; GOBP) of the MM cluster. a) and b) Halofuginone treatment. c) and d) NCP26 treatment. a) and c) GOBP pathway analysis performed using top 200 upregulated DEGs ($p_{adj} < 0.05$ and $\log_{2FC} > 0$) ranked by fold change. b) and d) GOBP pathway analysis performed using top 200 downregulated DEGs ($p_{adj} < 0.05$ and $\log_{2FC} < 0$) ranked by fold change.

For NCP26 and Halofuginone treatment of MM cells, pathways related to ER-stress and apoptosis are enriched. The unfolded protein response- another member of the ISR, which shares many effectors with AAR- is also enriched. This supports our bulk RNA-seq results that the AAR response is activated following NCP26 and Halofuginone treatment of MM cells.

Figure 6.10 shows a more detailed exploration of AAR genes and their expression in MM cells, as well as the immune microenvironment. Figure 6.10a shows increased

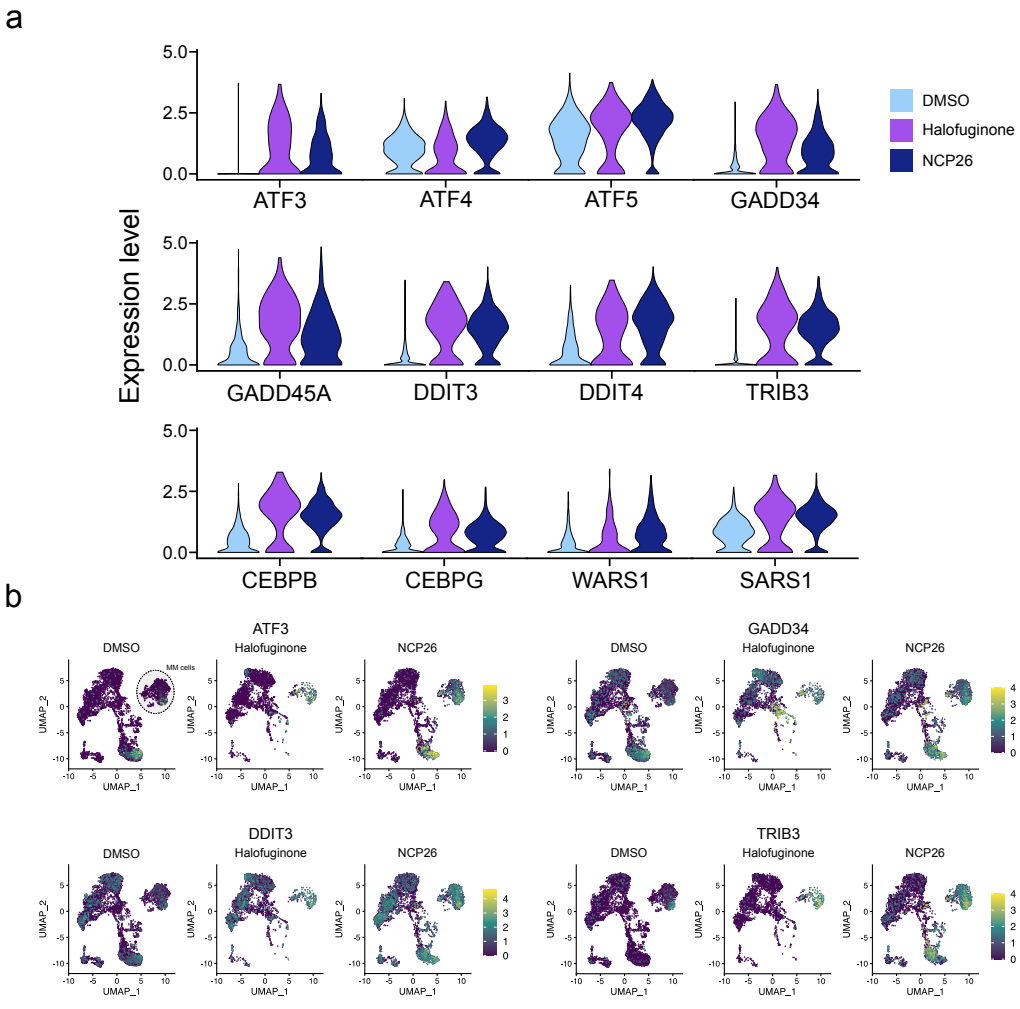


Figure 6.10: A selection of differentially expressed amino acid starvation response (AAR) genes following 24 hour treatment with ProRS inhibitors (newly-diagnosed MM patients). a) Violin plots showing expression of selected AAR genes in the MM population following DMSO, Halofuginone and NCP26 treatment. b) Feature plots of UMAP clustering, showing gene expression of AAR genes *ATF3*, *GADD34*, *DDIT3* and *TRIB3*. Feature plots are split into three panels based on treatment condition (DMSO, Halofuginone and NCP26). The myeloma cell population is circled in the first panel.

expression of *ATF4* target genes, such as *ATF3*, *DDIT3* and *GADD34*, following HF and NCP26 treatment. tRNA aminoacyl synthetase genes (*WARS1* and *SARS1*) are also over-expressed in MM cells following ProRS inhibitor treatment. Figure 6.10b shows selected AAR genes' expression for the entire UMAP clustering plot, separated by treatment condition. Markedly increased expression of *ATF3* and *TRIB3* can be seen localised to MM cells and monocytes for HF and NCP26 treatment. Increased expression of *DDIT3* and *GADD34* can be seen across a few different clusters, however it is most pronounced in the MM and monocyte clusters. Additionally, the apoptotic mediator *TNFRSF10B*, positively regulated by DDIT3, is upregulated in MM cells and monocytes (Figure 6.11b). Indicating that, although aaRSs are ubiquitous enzymes and ProRS inhibitors cause some degree of activation of the AAR in all cells, they are cytotoxically more selective for MM cells, and lead to an apoptotic cascade of events, preferentially over other cell types. *CDKN1A* is also overexpressed in MM cells following NCP26 and HF treatment. *CDKN1A* is a target gene of *ATF4* and arrests cell cycle progression by inhibiting the activity of cyclin-dependent kinases. HF has previously been show to cause cell cycle arrest and accumulating cells in the G_0/G_1 phase of cell cycle.

Moreover, Figure 6.11 demonstrates that NCP26 and HF have profound anti-MM effects on MM cells. NCP26 and HF treatment markedly reduce expression of MM pathological marker *CD138* in MM cells. Additionally NCP26 and HF treatment reduced expression of *IGKC* and *IGLC2*, the genes coding for the constant regions of immunoglobulin light chains, Kappa and Lambda. Previously these genes have been implicated in MM outcome.

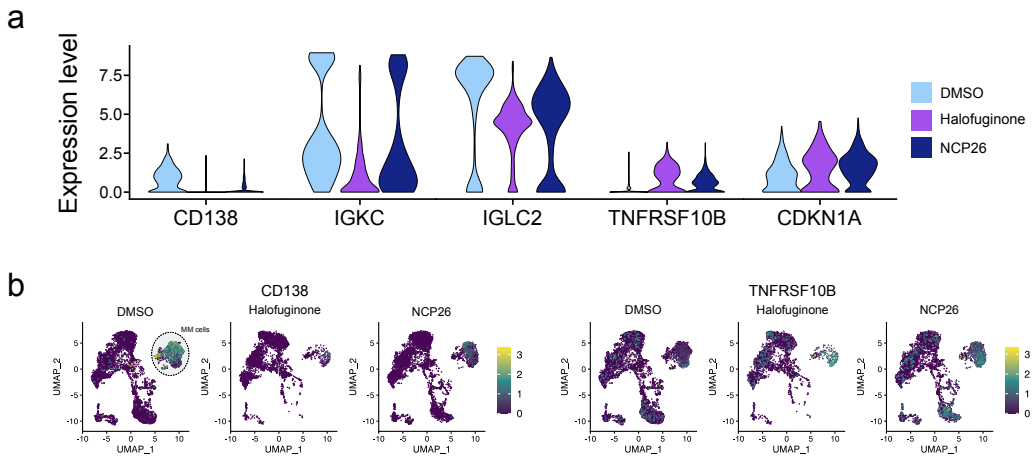


Figure 6.11: A selection of differentially expressed multiple myeloma (MM) and cell cycle/apoptotic markers in newly diagnosed patients treated with ProRS inhibitors for 24 hours. a) Violin plots showing expression of selected genes in the MM population following DMSO, Halofuginone and NCP26 treatment. b) Feature plots of UMAP clustering, showing gene expression of MM pathological marker *CD138* and apoptotic marker *TNFRSF10B*. The myeloma cell population is circled in the first panel. Halofugione and NCP26 treatment reduce *CD138* expression and increase *TNFRSF10B* expression in the MM clusters.

6.3.2 Relapsed MM

Figure 6.12 shows final cell-type annotation for the integrated relapsed MM patients. The relapsed patients demonstrate substantial transcriptional differences to treatment-naive patients, in both their myeloma cells and their immune microenvironment. 15 distinct clusters were identified. The expected major immune clusters

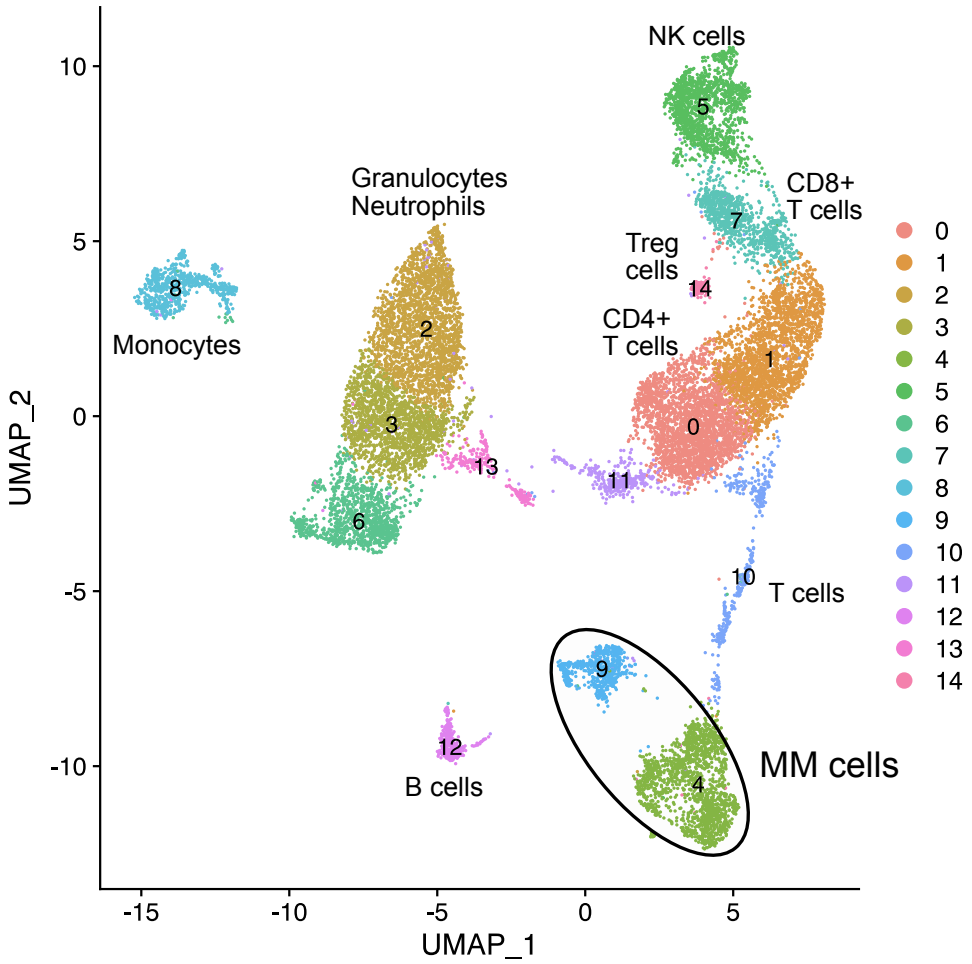


Figure 6.12: Fully annotated UMAP clustering analysis of two relapsed multiple myeloma (MM) patients. 15 distinct clusters were identified. Two subclones of MM were identified (circled; clusters 4 and 9).

were identified (such as B cells, NK cells, T cells and myeloid cells.) Using the established MM biological markers (shown in Figure 6.4), and inferCNV results, two distinct MM subclones (cluster 4 and cluster 9) were identified. MM cells in cluster 4 were seen to express *CD138* (Figure 6.4). However, MM cells in cluster 9

showed little or no *CD138* expression. As mentioned in Chapter 4, these *CD138*$^{-}$ MM cells are likely to have higher proliferative potential than *CD138*$^{+}$ MM cells. They are also sometimes regarded to have 'stem-cell' like clonogenic properties and to be more resistant to chemotherapy than their *CD138*$^{+}$ counterparts[260]. This *CD138*$^{-}$ subclone of MM cells would have been removed if *CD138*$^{+}$ enrichment techniques were used prior to sequencing.

Composition

Cluster composition analysis for relapsed MM by treatment is shown in Figure 6.13. 1µM and 5µM NCP26 treatment reduced the proportion of cells in MM cluster 4 compared to the DMSO control. There is insufficient evidence ($p>0.05$) to conclude if Halofuginone treatment affected the proportion of cells in MM cluster 4. However, Halofuginone significantly reduced the proportion of cells in the myeloid and B cell clusters. This perhaps indicates that in relapsed MM, Halofuginone is not selective for myeloma cells over other immune cells. This may indicate an advantage of NCP26 and proline non-competitive ProRS inhibitors in relapsed myeloma. Neither ProRS inhibitors reduced the proportion of cells in MM cluster 9.

Differential expression

Next, differential gene expression was investigated for the relapsed MM patients. 1073 genes were DE in MM cluster 4 following 24-hour 1µM HF treatment. 157 and 801 genes were DE in the MM cell cluster following 24-hour 1µM and 5µM NCP26 treatment, respectively. The breakdown of DEGs per cell type for NCP26 and HF treatment is shown in Figure 6.14.

4, 1, and 3 genes were DE in MM cluster 9 by Halofuginone, 1µM NCP26 and 5µM NCP26 treatment, respectively. Together with the composition analysis, this indicates that the cluster 9 MM subclone may be more resistant to ProRS inhibition than the other myeloma subclone.

In contrast to the newly-diagnosed MM data, many of the other immune subtypes are differentially expressed to similar degrees as MM cells by ProRS inhibition. Additionally, the myeloid cell cluster has more DEGs than MM cells for 1µM and

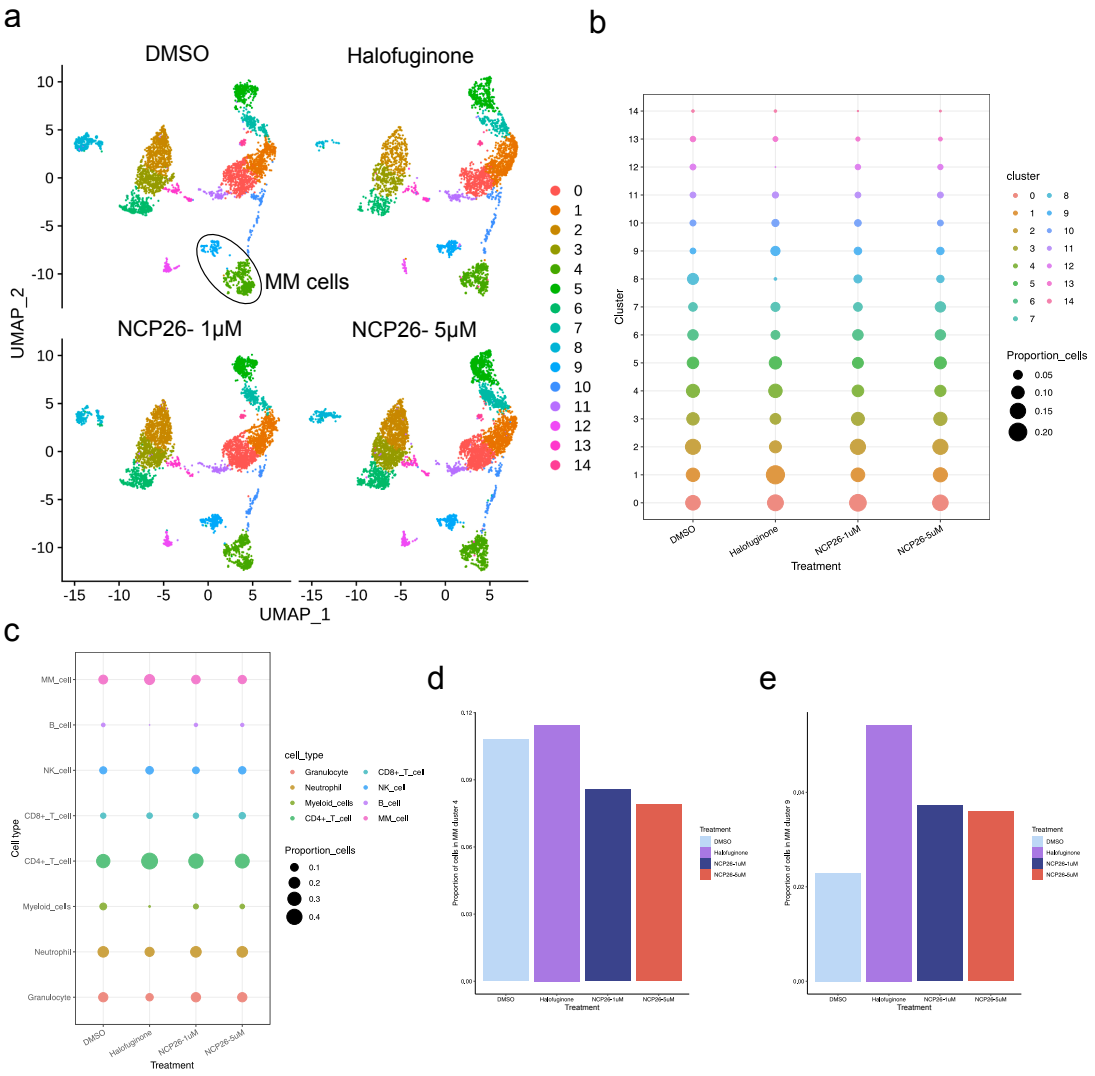


Figure 6.13: Composition analysis of relapsed MM cells treated for 24 hours with ProRS inhibitors. a) UMAP cell composition plots separated by treatment condition. b) Dot plot showing proportion of cells in each cluster for each sample. c) Dot plot showing proportion of cells in each cell class for each sample (as labelled in Figure 6.12). d) The proportion of cells in the MM cluster only (stars above bars indicate significant at $p<0.01$ compared to DMSO control). NCP26 treatment reduces the proportion of cells in the MM cluster ($p<0.01$).

5μM NCP26 treatment. Myeloid cells have fewer DEGs than cluster 4 MM cells for HF treatment, however Figure 6.13 demonstrates a substantial lower proportion of cells in the myeloid cells for the HF-treated sample, therefore you may not expect as many statistically significant DEGs for the few cells remaining in the cluster. This could indicate that NCP26 and HF are not as selective for MM cells in relapsed

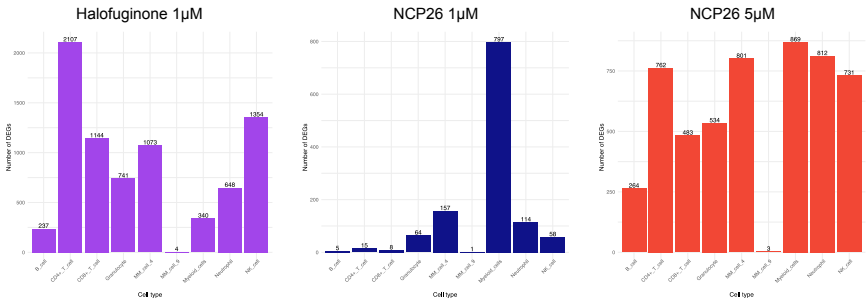


Figure 6.14: Number of differentially expressed genes (DEGs; $p_{adj} < 0.05$) broken down by cell type for a relapsed MM patient (patient 4 only) treated with ProRS inhibitors (Halofuginone and NCP26) for 24 hours. Cell type annotation corresponding to Figure 6.12.

MM as they are for newly-diagnosed MM, and that the ProRS inhibitors may not be very effective against some resistant myeloma subclones.

Pathway analysis

Pathway enrichment analysis was performed for the top 200 upregulated and top 200 downregulated DEGs in MM cluster 4, following Halofuginone and NCP26 treatment (Figure 6.15). Pathway analysis could not be performed for MM cluster 9, as there were too few DEGs. Genes involved in translation pathways are downregulated

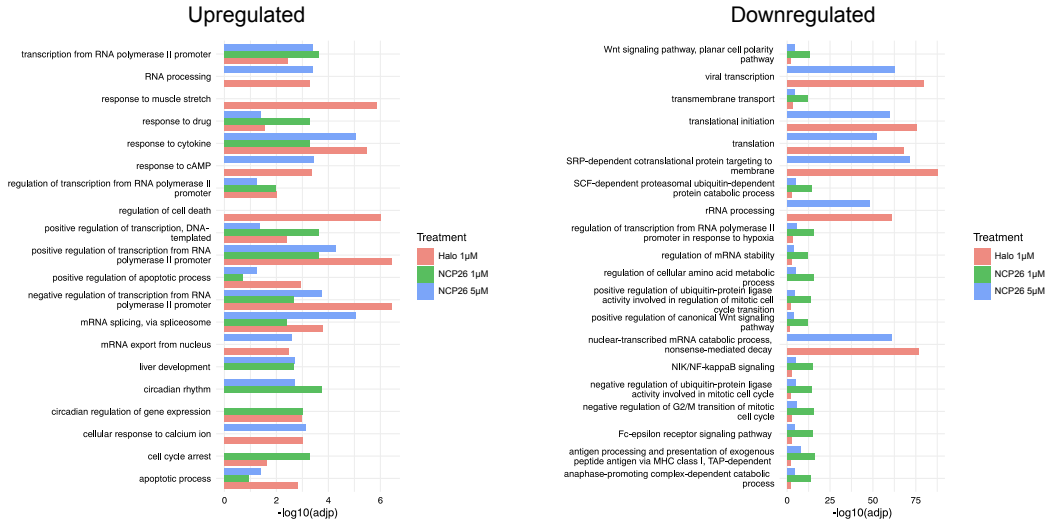


Figure 6.15: Pathway analysis (Gene ontology biological processes; GOBP) of MM cluster 4 in RRMM dataset. GOBP pathway analysis performed using top 200 upregulated or downregulated DEGs for ProRS inhibitor (1μM Halofuginone and 1μM or 5μM NCP26) treatment compared to DMSO control treatment for MM cluster 4.

following ProRS inhibition, as well as cell cycle pathways. Cell death, apoptotic and cell cycle pathways are enriched following ProRS inhibition. This follows literature and previous sections, whereby AAR activation triggers cell cycle changes, cell death and a global reduction in the translational machinery. Interestingly, for all three treatments, mRNA splicing was enriched, indicating a potential spliceosomal mechanism for these inhibitors.

Figure 6.16 shows AAR gene expression in relapsed myeloma cells, as well as the surrounding immune cells. AAR gene expression in MM cluster 4 follows

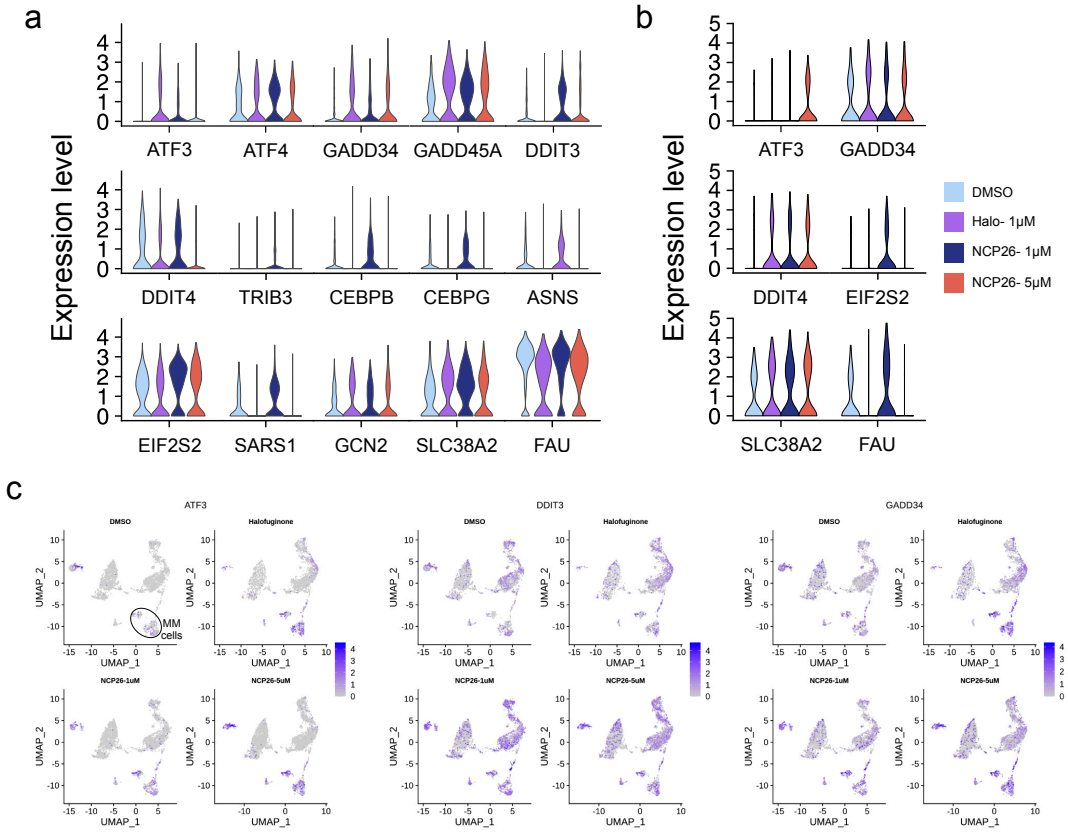


Figure 6.16: A selection of differentially expressed amino acid starvation response (AAR) genes following 24 hour treatment with ProRS inhibitors (newly-diagnosed MM patients). a) Violin plots showing AAR gene expression in MM cell cluster 4. a) Violin plots showing AAR gene expression in MM cell cluster 9. b) Feature plots of UMAP clustering, showing gene expression of AAR genes *ATF3*, *DDIT3* and *GADD34*. Feature plots are split into four panels based on treatment condition. The myeloma subclones are circled in the first panel.

the typical patterns of ProRS inhibition, as seen in the naive dataset and bulk RNA-seq data, increased expression of *ATF3*, *GADD34*, *GADD45A*, *DDIT3* and

amino acid transporters, and decreased expression of some AAR-related genes such as *DDIT4*. However, for the other myeloma subclonal population (cluster 9), AAR transcriptional changes following ProRS inhibitor treatment do not seem to fit the typical pattern seen in previous data. For example, *ATF3*, *GADD34* and *SLC38A2* are upregulated in cluster 9 following NCP26 treatment, however many other AAR genes' expression levels remain unaffected by treatment. Additionally, some AAR gene expression levels change inversely to how they behave in previous data, for example ProRS inhibition causes upregulation of *DDIT4* and *FAU* in cluster 9, where previously HF and NCP26 treatment has been shown to cause downregulation of these genes in myeloma cells.

6.3.3 ProRS inhibitor effect on myeloid cells

From differential expression and composition analysis, it seems that Halofuginone and NCP26 have a large transcriptional effect on myeloid cells, especially monocytes. Feature plots and violin plots demonstrate that amino acid starvation response effector genes, such as *DDIT3* and *ATF3* are enriched in monocytes, as well as MM cells. However, myeloid cells are non-malignant. ProRS inhibitors were originally used as cancer treatments because tumour cells produce large amounts of protein, and are therefore heavily dependent on translational machinery, like tRNA-synthetase enzymes. This is especially true for myeloma cells, secreting huge amounts of M protein. Additionally, tumours are usually extremely proline-rich. Therefore, it is surprising that ProRS inhibitors are non-selective for MM cells over MM patients' non-malignant myeloid cells.

Previously, Leiba et al. (2012) demonstrated HF's selectivity for MM patients' $CD138^+$ tumour cells over PBMCs from healthy donors[173] using a cytotoxicity assay at the range of 25-200nM. This is a significantly lower concentration of HF than used in the scRNA-seq experiments (1µM), so this could represent dosing differences. However, it must also be noted that Leiba et al. (2012) did not separate PBMCs into their respective subtypes prior to performing the cytotoxicity assays. As previously demonstrated, HF treatment had very little effect on B, T and

NK cells. Monocytes only make up approximately 5-10% of the cells found in PBMCs- therefore, the effect of HF on monocytes could have been missed due to the minimal effects on the other cell types making up PBMCs. Furthermore, the control PBMCs originated from healthy donors and not from MM patients' immune microenvironment. The immune microenvironment of MM patients has been shown to be substantially altered and impaired compared to healthy individuals[261], meaning that using healthy PBMCs as a control for the selectivity of HF is not appropriate. To truly determine the cytotoxic selectivity of ProRS inhibition for MM cells and myeloid cells, cytotoxic assays must be performed for MM cells and myeloid cells originating from MM patients.

Wang et al. (2020) investigated the use of Halofuginone to treat the inflammatory disease chronic periodontitis[262]. Using a mouse model, they demonstrated that HF treatment significantly reduced the expression levels of pro-inflammatory cytokines, for example IL-1β, IL-6 and TNF-α. HF treatment was also shown to reduce the total percent of infiltrating immune cells and myeloid cells in gingival tissues when compared with those in control, PBS-treated mice. The percentage of myeloid cells making up the $CD45^+$ fraction of cells in gingival tissue was reduced by approximately 40%, following HF treatment. The authors demonstrated that HF treatment did not affect the cell viability of BM-derived osteoclast precursors cells (that is monocytes and macrophages). However, cell viability was investigated with concentrations of HF ranging from 10pM to 100nM, so perhaps cell myeloid death is reserved for higher doses of HF treatment. The scRNA-seq experiments used 1µM HF, so this may be the concentration levels required to observe ProRS inhibitory effects on myeloid cells. Additionally, the monocytes/macrophages orignated from mice not humans, and without the immune microenvironment remodelling seen in MM.

Pathway analysis

Inflammation etc. etc. Non-canonical functionality etc. etc.

Velocity analysis

6.4 Summary

6.5 Discussion

Stem cell CD138- MM cells, resistant to drugs. More research include whole BM niche, not just CD138+ fraction. MM classifier can simplify process of MM identification.

Drugs showed little effect on cluster 9. Could be key to MRD and clonogenic. Naive patient showed large CD138+ MM expression. Difference between naive and relapse Argument for sequencing whole BM niche. Also if just sequenced MM cells (CD138+) would miss effects on myeloid cells.

Myeloid stuff, off target non-canonical functionality

Alternative splicing...

https://www.mdpi.com/1422-0067/19/2/545/htm Borrelidin ThrRS threonyl-tRNA synthetase In colon tumor cells, the spliceosome-associated protein FBP21 (formin binding protein 21) was the target of borrelidin

Appendices

A
Supplementary figures

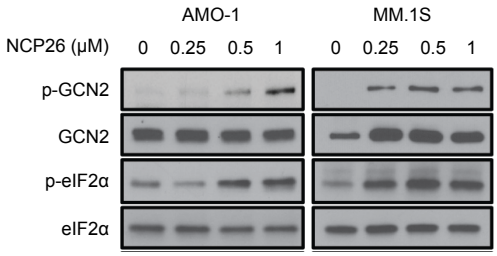


Figure A.1: Western blot demonstrating dose-dependent NCP26 canonical ISR activation with GCN2 and eIF2α phosphorylation. AMO-1 and MM.1S myeloma cell lines used. Experiment performed by international collaborators on a multiple-lab collaborative research paper[263].

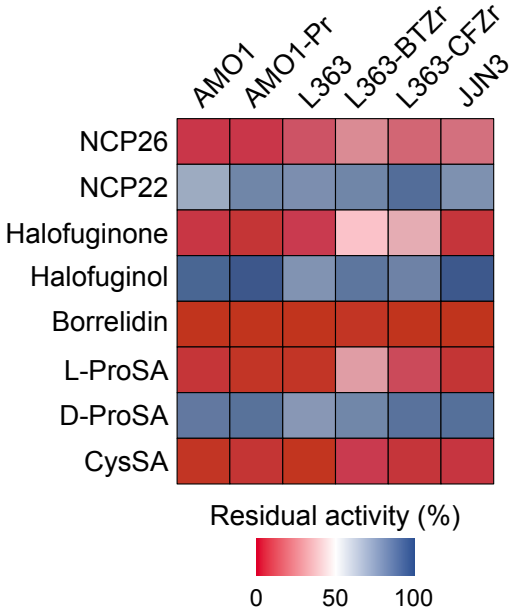


Figure A.2: Anti-proliferative activities of aaRS inhibitors in MM cell lines. Experiment performed by international collaborators on a multiple-lab collaborative research paper[263]. 1µM NCP26, NCP22, halofuginone, halofuginol (MAZ1805), borrelidin (threonyl-tRNA synthetase/ThrRS inhibitor), or 5 µM L-ProSA, D-ProSA and CysSA (amino acid analogues). 72 hour MTT assay, n=2-5 independent experiments in triplicate technical repeats).

Cell type	**Markers**
Multiple myeloma cells	CD138, CD38 (lower than plasma cells), SLAMF7, BCMA, KRAS, IGKC, IGCL2. Reduced/ no CD20, CD19, CD45 expression.
Normal plasma cells	CD38, CD19, some BCMA
B cells	CD20, some CD19
T cells	TRAC, CD3D
Cytotoxic cells	GZMH, GZMB, GZMA, PRF1
CD4+ T cells	CCR7, SELL, TCF7 and T cell markers
CD8+ T cells	CD8A, cytotoxic markers and T cell markers
NK cells	KLRB1, KLRC1, KLRF1, CD16 and cytotoxic markers
Dendritic cells	CD1C, FCER1A
Monocytes	CD14/CD16, CD68

Table A.1: Manual annotation markers for cell types originating from transcriptomic profiles of bone marrow samples. SLAMF7, BCMA, KRAS, IGKC and IGCL2 are very highly expressed by MM cells, but are not exclusive to this cluster. MM patient $CD45^+$ immune cells scRNA-seq marker annotation can be found [264].

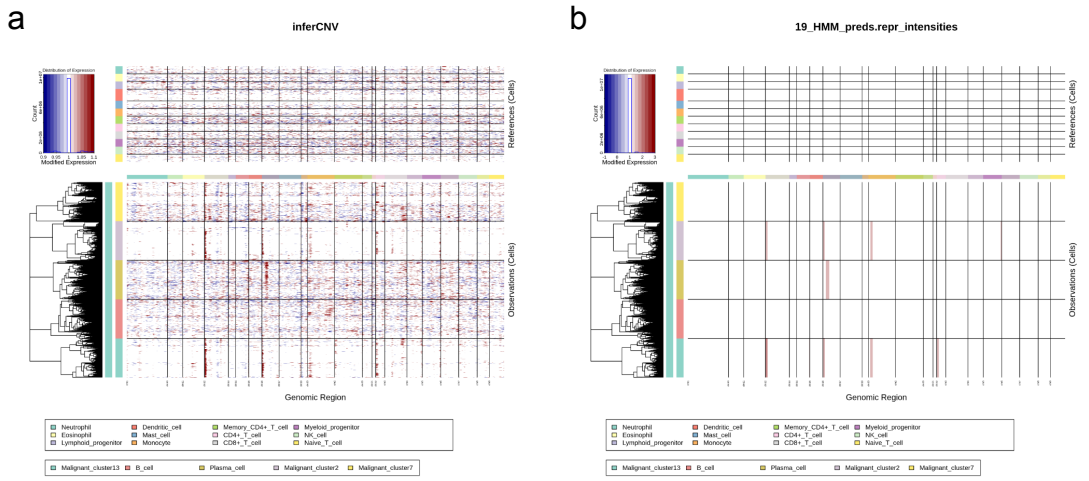


Figure A.3: InferCNV results for the newly-diagnosed MM dataset. [a and b] InferCNV heatmaps. The top panel shows expression values for the reference 'normal' cells. The bottom panel shows expression values for the suspected malignant cells (clusters 2, 7 and 13), and other B-cell lineages (B cells and plasma cells). Red indicates chromosomal region amplifications and blue indicates chromosomal region deletions. a) De-noised inferCNV results. b) Hidden Markov-Model (HMM) copy number variation (CNV) region predictions. Only some chromosomal region gains predicted in MM clusters 2 and 13, and the plasma cell cluster.

References

[1] International Myeloma Working Group. "Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group". In: *British journal of haematology* 121.5 (2003), pp. 749–757.

[2] Heinz Ludwig et al. "Multiple myeloma incidence and mortality around the globe; Interrelations between health access and quality, economic resources, and patient empowerment". In: *The Oncologist* 25.9 (2020), e1406–e1413.

[3] Dickran Kazandjian and Ola Landgren. "A look backward and forward in the regulatory and treatment history of multiple myeloma: approval of novel-novel agents, new drug development, and longer patient survival". In: *Seminars in oncology*. Vol. 43. 6. Elsevier. 2016, pp. 682–689.

[4] Bruce Alberts et al. "The innate and adaptive immune systems". In: *Molecular Biology of the Cell. 6th edition*. 6th ed. Garland science, 2014, pp. 1297–1339.

[5] Bruce Alberts et al. "Stem cells and tissue renewal". In: *Molecular Biology of the Cell. 6th edition*. 6th ed. Garland science, 2014, pp. 1239–1247.

[6] Jun Seita and Irving L Weissman. "Hematopoietic stem cell: self-renewal versus differentiation". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2.6 (2010), pp. 640–653.

[7] Jules Hoffmann, Shizuo Akira, and J Hoffmann. "Innate immunity editorial overview". In: *Curr Opin Immunol [Internet]* 25.1 (2013), pp. 1–3.

[8] Rishi Vishal Luckheeram et al. "CD4+ T cells: differentiation and functions". In: *Clinical and developmental immunology* 2012 (2012).

[9] Harry W Schroeder Jr and Lisa Cavacini. "Structure and function of immunoglobulins". In: *Journal of Allergy and Clinical Immunology* 125.2 (2010), S41–S52.

[10] Mario D Friess, Kristyna Pluhackova, and Rainer A Böckmann. "Structural model of the mIgM B-cell receptor transmembrane domain from self-association molecular dynamics simulations". In: *Frontiers in immunology* 9 (2018), p. 2947.

[11] Janis Dylke et al. "Role of the extracellular and transmembrane domain of Ig-α/β in assembly of the B cell antigen receptor (BCR)". In: *Immunology letters* 112.1 (2007), pp. 47–57.

[12] Katrin Roth et al. "Tracking plasma cell differentiation and survival". In: *Cytometry Part A* 85.1 (2014), pp. 15–24.

[13] Michel Jourdan et al. "Characterization of a transitional preplasmablast population in the process of human B cell to plasma cell differentiation". In: *The Journal of Immunology* 187.8 (2011), pp. 3931–3941.

[14] Miriam Shapiro-Shelef and Kathryn Calame. “Plasma cell differentiation and multiple myeloma”. In: *Current opinion in immunology* 16.2 (2004), pp. 226–234.

[15] Arjun K Mishra and Roy A Mariuzza. “Insights into the structural basis of antibody affinity maturation from next-generation sequencing”. In: *Frontiers in immunology* 9 (2018), p. 117.

[16] TW Mak, ME Saunders, and BD Jett. “B cell development, activation and effector functions”. In: *Primer to the Immune Response* 2 (2014), pp. 111–142.

[17] Zhenming Xu et al. “Immunoglobulin class-switch DNA recombination: induction, targeting and beyond”. In: *Nature Reviews Immunology* 12.7 (2012), pp. 517–531.

[18] Anna-Karin E Palm and Carole Henry. “Remembrance of things past: long-term B cell memory after infection and vaccination”. In: *Frontiers in immunology* 10 (2019), p. 1787.

[19] Alexandra Bortnick and David Allman. “What Is and What Should Always Have Been: Long-Lived Plasma Cells Induced by T Cell–Independent Antigens”. In: *The Journal of Immunology* 190.12 (2013), pp. 5913–5918.

[20] Mathieu Andraud et al. “Living on three time scales: the dynamics of plasma cell and antibody populations illustrated for hepatitis a virus”. In: *PLoS Comput Biol* 8.3 (2012), e1002418.

[21] Kenneth C Anderson and Ruben D Carrasco. “Pathogenesis of myeloma”. In: *Annual Review of Pathology: Mechanisms of Disease* 6 (2011), pp. 249–274.

[22] Cancer Research UK. *Types of myeloma*. `https://www.cancerresearchuk.org/about-cancer/myeloma/types`. Accessed: 02-2022.

[23] John De Vos et al. “Comparison of gene expression profiling between malignant and normal plasma cells with oligonucleotide arrays”. In: *Oncogene* 21.44 (2002), pp. 6848–6857.

[24] Cinzia Caprio et al. “Epigenetic aberrations in multiple myeloma”. In: *Cancers* 12.10 (2020), p. 2996.

[25] Venkatesh Chanukuppa et al. “Proteomic alterations in multiple myeloma: A comprehensive study using bone marrow interstitial fluid and serum samples”. In: *Frontiers in Oncology* (2021), p. 3164.

[26] Chaima El Arfani et al. “Metabolic features of multiple myeloma”. In: *International journal of molecular sciences* 19.4 (2018), p. 1200.

[27] Kazuhiro Nishida et al. “The Ig heavy chain gene is frequently involved in chromosomal translocations in multiple myeloma and plasma cell leukemia as detected by in situ hybridization”. In: *Blood, The Journal of the American Society of Hematology* 90.2 (1997), pp. 526–534.

[28] Steven H Swerdlow et al. *WHO classification of tumours of haematopoietic and lymphoid tissues*. Vol. 2. International agency for research on cancer Lyon, 2008.

[29] S Manier et al. “Bone marrow microenvironment in multiple myeloma progression”. In: *Journal of Biomedicine and Biotechnology* 2012 (2012).

[30] Yawara Kawano et al. “Targeting the bone marrow microenvironment in multiple myeloma”. In: *Immunological reviews* 263.1 (2015), pp. 160–172.

[31] Matthew Tsang et al. “Multiple myeloma epidemiology and patient geographic distribution in Canada: a population study”. In: *Cancer* (2019).

[32] Antonio Palumbo and Kenneth Anderson. “Multiple Myeloma”. In: *New England Journal of Medicine* 364.11 (2011), pp. 1046–1060.

[33] NHS UK. *Multiple Myeloma*. `https://www.nhs.uk/conditions/multiple-myeloma/`. Accessed: 06-2019.

[34] Lauren R Teras et al. “2016 US lymphoid malignancy statistics by World Health Organization subtypes”. In: *CA: a cancer journal for clinicians* 66.6 (2016), pp. 443–459.

[35] Andrew J Cowan et al. “Global burden of multiple myeloma: a systematic analysis for the Global Burden of Disease Study 2016”. In: *JAMA oncology* 4.9 (2018), pp. 1221–1227.

[36] Cancer Research UK. *Myeloma Survival Statistics*. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/myeloma/survival`. Accessed: 06-2019.

[37] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. “Cancer statistics, 2016”. In: *CA: a cancer journal for clinicians* 66.1 (2016), pp. 7–30.

[38] Niels van Nieuwenhuijzen et al. “From MGUS to multiple myeloma, a paradigm for clonal evolution of premalignant cells”. In: *Cancer research* 78.10 (2018), pp. 2449–2456.

[39] S Vincent Rajkumar, Ola Landgren, and Maria-Victoria Mateos. “Smoldering multiple myeloma”. In: *Blood, The Journal of the American Society of Hematology* 125.20 (2015), pp. 3069–3075.

[40] Neha Korde, Sigurdur Y Kristinsson, and Ola Landgren. “Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): novel biological insights and development of early treatment strategies”. In: *Blood* 117.21 (2011), pp. 5573–5581.

[41] Robert A Kyle et al. “Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma”. In: *New England Journal of Medicine* 356.25 (2007), pp. 2582–2590.

[42] S Vincent Rajkumar et al. “International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma”. In: *The lancet oncology* 15.12 (2014), e538–e548.

[43] N Blokhin et al. “Clinical experiences with sarcolysin in neoplastic diseases”. In: *Annals of the New York Academy of Sciences* 68.3 (1958), pp. 1128–1132.

[44] ROBERT E MASS. “A comparison of the effect of prednisone and a placebo in the treatment of multiple myeloma.” In: *Cancer chemotherapy reports* 16 (1962), p. 257.

[45] Raymond Alexanian et al. “Treatment for multiple myeloma: combination chemotherapy with different melphalan dose regimens”. In: *Jama* 208.9 (1969), pp. 1680–1685.

[46] TJ McElwain and RL Powles. “High-dose intravenous melphalan for plasma-cell leukaemia and myeloma”. In: *The Lancet* 322.8354 (1983), pp. 822–824.

[47] Elliott F Osserman et al. “Identical twin marrow transplantation in multiple myeloma”. In: *Acta haematologica* 68.3 (1982), pp. 215–223.

[48] Alexander Fefer, Martin A Cheever, and Philip D Greenberg. “Identical-twin (syngeneic) marrow transplantation for hematologic cancers”. In: *Journal of the National Cancer Institute* 76.6 (1986), pp. 1269–1273.

[49] G Gahrton et al. “Bone marrow transplantation in multiple myeloma: report from the European Cooperative Group for Bone Marrow Transplantation”. In: *Blood* 69.4 (1987), pp. 1262–1264.

[50] Robert C Kane et al. “Velcade: US FDA approval for the treatment of multiple myeloma progressing on prior therapy”. In: *The oncologist* 8.6 (2003), pp. 508–513.

[51] Paul G Richardson et al. “A phase 2 study of bortezomib in relapsed, refractory myeloma”. In: *New England Journal of Medicine* 348.26 (2003), pp. 2609–2617.

[52] Alla Katsnelson. *Next-generation proteasome inhibitor approved in multiple myeloma*. 2012.

[53] Seema Singhal et al. “Antitumor activity of thalidomide in refractory multiple myeloma”. In: *New England Journal of Medicine* 341.21 (1999), pp. 1565–1571.

[54] FDA Label. “Revlimid-lenalidomide capsule”. In: *For Multiple Myeloma Myelodysplastic Syndrome and Mantle Cell Lymphoma* 47 ().

[55] Jesus San Miguel et al. “Pomalidomide plus low-dose dexamethasone versus high-dose dexamethasone alone for patients with relapsed and refractory multiple myeloma (MM-003): a randomised, open-label, phase 3 trial”. In: *The lancet oncology* 14.11 (2013), pp. 1055–1066.

[56] Henk M Lokhorst et al. “Targeting CD38 with daratumumab monotherapy in multiple myeloma”. In: *New England Journal of Medicine* 373.13 (2015), pp. 1207–1219.

[57] Sagar Lonial et al. “Elotuzumab therapy for relapsed or refractory multiple myeloma”. In: *New England Journal of Medicine* 373.7 (2015), pp. 621–631.

[58] US FDA. *FDA granted accelerated approval to belantamab mafodotin-blmf for multiple myeloma*. `https://www.fda.gov/drugs/resources-information-approved-drugs/fda-granted-accelerated-approval-belantamab-mafodotin-blmf-multiple-myeloma`. Accessed: 02-02-2022.

[59] Anthony Markham. “Belantamab mafodotin: first approval”. In: *Drugs* (2020), pp. 1–7.

[60] US FDA. *FDA approves selinexor for refractory or relapsed multiple myeloma*. `https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-selinexor-refractory-or-relapsed-multiple-myeloma`. Accessed: 02-02-2022.

[61] Klaus Podar et al. “Selinexor for the treatment of multiple myeloma”. In: *Expert opinion on pharmacotherapy* 21.4 (2020), pp. 399–408.

[62] Jesus F San Miguel et al. “Bortezomib plus melphalan and prednisone for initial treatment of multiple myeloma”. In: *New England Journal of Medicine* 359.9 (2008), pp. 906–917.

[63] Philippe Moreau et al. “Proteasome inhibitors in multiple myeloma: 10 years later”. In: *Blood* 120.5 (2012), pp. 947–959.

[64] Gary Kleiger and Thibault Mayor. “Perilous journey: a tour of the ubiquitin–proteasome system”. In: *Trends in cell biology* 24.6 (2014), pp. 352–359.

[65] Bruce Alberts et al. *Molecular Biology of the Cell*. 6th ed. Garland science, Dec. 2014, pp. 356–359.

[66] Andrej Besse et al. “Proteasome Inhibition in Multiple Myeloma: Head-to-Head Comparison of Currently Available Proteasome Inhibitors”. In: *Cell chemical biology* 26.3 (2019), pp. 340–351.

[67] Lenka Kubiczkova et al. “Proteasome inhibitors–molecular basis and current perspectives in multiple myeloma”. In: *Journal of cellular and molecular medicine* 18.6 (2014), pp. 947–961.

[68] Craig T Wallington-Beddoe et al. “Resistance to proteasome inhibitors and other targeted therapies in myeloma”. In: *British journal of haematology* 182.1 (2018), pp. 11–28.

[69] Andrew J Kale and Bradley S Moore. “Molecular mechanisms of acquired proteasome inhibitor resistance”. In: *Journal of medicinal chemistry* 55.23 (2012), pp. 10317–10327.

[70] Hong Ding et al. “Minimal residual disease in multiple myeloma: current status”. In: *Biomarker research* 9.1 (2021), pp. 1–10.

[71] Shaji Kumar et al. “International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma”. In: *The lancet oncology* 17.8 (2016), e328–e346.

[72] Bruno Paiva et al. “Multiparameter flow cytometric remission is the most relevant prognostic factor for multiple myeloma patients who undergo autologous stem cell transplantation”. In: *Blood, The Journal of the American Society of Hematology* 112.10 (2008), pp. 4017–4023.

[73] Joaquin Martinez-Lopez et al. “Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma”. In: *Blood, The Journal of the American Society of Hematology* 123.20 (2014), pp. 3073–3079.

[74] Nikhil C Munshi et al. “A large meta-analysis establishes the role of MRD negativity in long-term survival outcomes in patients with multiple myeloma”. In: *Blood advances* 4.23 (2020), pp. 5988–5999.

[75] Jan B Egan et al. “Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides”. In: *Blood, The Journal of the American Society of Hematology* 120.5 (2012), pp. 1060–1066.

[76] Jonathan J Keats et al. “Clonal competition with alternating dominance in multiple myeloma”. In: *Blood, The Journal of the American Society of Hematology* 120.5 (2012), pp. 1067–1076.

[77] Brian A Walker et al. “Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t (4; 14) and t (11; 14) myeloma”. In: *Blood, The Journal of the American Society of Hematology* 120.5 (2012), pp. 1077–1086.

[78] Yusuke Furukawa and Jiro Kikuchi. “Molecular basis of clonal evolution in multiple myeloma”. In: *International Journal of Hematology* 111.4 (2020), pp. 496–511.

[79] John R Jones et al. “Clonal evolution in myeloma: the impact of maintenance lenalidomide and depth of response on the genetics and sub-clonal structure of relapsed disease in uniformly treated newly diagnosed patients”. In: *Haematologica* 104.7 (2019), p. 1440.

[80] Annamaria Brioli et al. “The impact of intra-clonal heterogeneity on the treatment of multiple myeloma”. In: *British journal of haematology* 165.4 (2014), pp. 441–454.

[81] Pawel Robak et al. “Drug resistance in multiple myeloma”. In: *Cancer treatment reviews* (2018).

[82] Myeloma patients europe. *MMPredict*. `https://www.mpeurope.org/what-we-do/projects/european-commission-projects-horizon-2020/mmpredict`. Accessed: 03-2022.

[83] A Annunziato. “DNA packaging: nucleosomes and chromatin”. In: *Nature Education* 1.1 (2008), p. 26.

[84] Bruce Alberts et al. “Chromosomal DNA and its packaging in the chromatin fiber”. In: *Molecular Biology of the Cell. 4th edition*. Garland science, 2002.

[85] Andrew J Bannister and Tony Kouzarides. “Regulation of chromatin by histone modifications”. In: *Cell research* 21.3 (2011), pp. 381–395.

[86] James M Heather and Benjamin Chain. “The sequence of sequencers: The history of sequencing DNA”. In: *Genomics* 107.1 (2016), pp. 1–8.

[87] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[88] Elizabeth Pennisi. *The human genome*. 2001.

[89] Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. “Long-read human genome sequencing and its applications”. In: *Nature Reviews Genetics* 21.10 (2020), pp. 597–614.

[90] Aaron M Wenger et al. “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome”. In: *Nature biotechnology* 37.10 (2019), pp. 1155–1162.

[91] Camilla LC Ip et al. “MinION Analysis and Reference Consortium: Phase 1 data release and analysis”. In: *F1000Research* 4 (2015).

[92] Miten Jain et al. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. In: *Genome biology* 17.1 (2016), pp. 1–11.

[93] Jason L Weirather et al. “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”. In: *F1000Research* 6 (2017).

[94] Martin Philpott et al. “Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq”. In: *Nature biotechnology* 39.12 (2021), pp. 1517–1520.

[95] Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature methods* 6.5 (2009), pp. 377–382.

[96] Simone Picelli et al. “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”. In: *Nature methods* 10.11 (2013), pp. 1096–1098.

[97] Evan Z Macosko et al. “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets”. In: *Cell* 161.5 (2015), pp. 1202–1214.

[98] Saiful Islam et al. “Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq”. In: *Genome research* 21.7 (2011), pp. 1160–1167.

[99] Allon M Klein et al. “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells”. In: *Cell* 161.5 (2015), pp. 1187–1201.

[100] Laurie S. Kaguni Lluis Ribas de Pouplana. *The Enzymes: Biology of Aminoacyl-tRNA Synthetases*. Vol. 48. Elsevier, 2020.

[101] Jong Hyun Kim et al. “Evolution of the multi-tRNA synthetase complex and its role in cancer”. In: *Journal of Biological Chemistry* 294.14 (2019), pp. 5340–5351.

[102] Junren Zhang, Qizheng Yao, and Zuliang Liu. “A novel synthesis of the efficient anti-coccidial drug halofuginone hydrobromide”. In: *Molecules* 22.7 (2017), p. 1086.

[103] Lluis Ribas de Pouplana. “The evolution of aminoacyl-tRNA synthetases: From dawn to LUCA”. In: *Biology of Aminoacyl-tRNA Synthetases* 48 (2020), p. 11.

[104] Kate J Newberry, Ya-Ming Hou, and John J Perona. “Structural origins of amino acid selection without editing by cysteinyl-tRNA synthetase”. In: *The EMBO Journal* 21.11 (2002), pp. 2778–2787.

[105] Nam Hoon Kwon, Paul L Fox, and Sunghoon Kim. “Aminoacyl-tRNA synthetases as therapeutic targets”. In: *Nature reviews Drug discovery* 18.8 (2019), pp. 629–650.

[106] L Pauling. *Festschrift fur Prof. Dr. Arthur Stoll*. 1958.

[107] ROBERT B Loftfield and DOROTHY Vanderjagt. “The frequency of errors in protein biosynthesis.” In: *Biochemical Journal* 128.5 (1972), p. 1353.

[108] Jeong Woong Lee et al. “Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration”. In: *Nature* 443.7107 (2006), pp. 50–55.

[109] Krishnendu Khan et al. “3-Dimensional architecture of the human multi-tRNA synthetase complex”. In: *Nucleic acids research* 48.15 (2020), pp. 8740–8754.

[110] Myung Hee Kim and Sunghoon Kim. “Structures and functions of multi-tRNA synthetase complexes”. In: *Biology of Aminoacyl-tRNA Synthetases* 48 (2020), p. 149.

[111] Ha Yeon Cho et al. “Assembly of multi-tRNA synthetase complex via heterotetrameric glutathione transferase-homology domains”. In: *Journal of Biological Chemistry* 290.49 (2015), pp. 29313–29328.

[112] Monika Kaminska et al. “Dissection of the structural organization of the aminoacyl-tRNA synthetase complex”. In: *Journal of Biological Chemistry* 284.10 (2009), pp. 6053–6060.

[113] Marc Mirande. "The aminoacyl-tRNA synthetase complex". In: *Macromolecular Protein Complexes* (2017), pp. 505–522.

[114] Jung Min Han et al. "Hierarchical network between the components of the multi-tRNA synthetase complex: implications for complex formation". In: *Journal of Biological Chemistry* 281.50 (2006), pp. 38663–38667.

[115] Jung Min Han et al. "Leucyl-tRNA synthetase is an intracellular leucine sensor for the mTORC1-signaling pathway". In: *Cell* 149.2 (2012), pp. 410–424.

[116] Nurit Yannay-Cohen et al. "LysRS serves as a key signaling molecule in the immune response by regulating gene expression". In: *Molecular cell* 34.5 (2009), pp. 603–611.

[117] Abul Arif et al. "Two-site phosphorylation of EPRS coordinates multimodal regulation of noncanonical translational control activity". In: *Molecular cell* 35.2 (2009), pp. 164–180.

[118] Abul Arif et al. "The GAIT translational control system". In: *Wiley Interdisciplinary Reviews: RNA* 9.2 (2018), e1441.

[119] Sang Gyu Park et al. "Precursor of pro-apoptotic cytokine modulates aminoacylation activity of tRNA synthetase". In: *Journal of Biological Chemistry* 274.24 (1999), pp. 16673–16676.

[120] Sang Gyu Park et al. "Hormonal activity of AIMP1/p43 for glucose homeostasis". In: *Proceedings of the National Academy of Sciences* 103.40 (2006), pp. 14913–14918.

[121] Zheng Zhou et al. "Roles of aminoacyl-tRNA synthetase-interacting multi-functional proteins in physiology and cancer". In: *Cell Death & Disease* 11.7 (2020), pp. 1–14.

[122] Nam Hoon Kwon et al. "Dual role of methionyl-tRNA synthetase in the regulation of translation and tumor suppressor activity of aminoacyl-tRNA synthetase-interacting multifunctional protein-3". In: *Proceedings of the National Academy of Sciences* 108.49 (2011), pp. 19635–19640.

[123] Peng Yao and Paul L Fox. "Aminoacyl-tRNA synthetases in medicine and disease". In: *EMBO molecular medicine* 5.3 (2013), pp. 332–343.

[124] Sven Bervoets et al. "Transcriptional dysregulation by a nucleus-localized aminoacyl-tRNA synthetase associated with Charcot-Marie-Tooth neuropathy". In: *Nature communications* 10.1 (2019), pp. 1–14.

[125] Sang Gyu Park, Paul Schimmel, and Sunghoon Kim. "Aminoacyl tRNA synthetases and their connections to disease". In: *Proceedings of the National Academy of Sciences* 105.32 (2008), pp. 11043–11049.

[126] Anzheng Nie et al. "Roles of aminoacyl-tRNA synthetases in immune regulation and immune diseases". In: *Cell death & disease* 10.12 (2019), pp. 1–14.

[127] Eun-Young Lee et al. "Infection-specific phosphorylation of glutamyl-prolyl tRNA synthetase induces antiviral immunity". In: *Nature immunology* 17.11 (2016), pp. 1252–1262.

[128] Alice A Duchon et al. "HIV-1 exploits a dynamic multi-aminoacyl-tRNA synthetase complex to enhance viral replication". In: *Journal of virology* 91.21 (2017), e01240–17.

[129] Young Ha Ahn et al. “Secreted tryptophanyl-tRNA synthetase as a primary defence system against infection”. In: *Nature microbiology* 2.1 (2016), pp. 1–13.

[130] Jie Hu et al. “Heterogeneity of tumor-induced gene expression changes in the human metabolic network”. In: *Nature biotechnology* 31.6 (2013), pp. 522–529.

[131] Zhongying Mo et al. “Neddylation requires glycyl-tRNA synthetase to protect activated E2”. In: *Nature structural & molecular biology* 23.8 (2016), pp. 730–737.

[132] Lu Deng et al. “The role of ubiquitination in tumorigenesis and targeted drug discovery”. In: *Signal transduction and targeted therapy* 5.1 (2020), pp. 1–28.

[133] Nam Hoon Kwon et al. “Stabilization of cyclin-dependent kinase 4 by methionyl-tRNA synthetase in p16INK4a-negative cancer”. In: *ACS pharmacology & translational science* 1.1 (2018), pp. 21–31.

[134] Tamara F Williams et al. “Secreted Threonyl-tRNA synthetase stimulates endothelial cell migration and angiogenesis”. In: *Scientific reports* 3.1 (2013), pp. 1–7.

[135] Adam C Mirando et al. “Aminoacyl-tRNA synthetase dependent angiogenesis revealed by a bioengineered macrolide inhibitor”. In: *Scientific reports* 5.1 (2015), pp. 1–17.

[136] Cindy Mendes et al. “Unraveling FATP1, regulated by ER-β, as a targeted breast cancer innovative therapy”. In: *Scientific reports* 9.1 (2019), pp. 1–15.

[137] Jin Woo Choi et al. “Multidirectional tumor-suppressive activity of AIMP2/p38 and the enhanced susceptibility of AIMP2 heterozygous mice to carcinogenesis”. In: *Carcinogenesis* 30.9 (2009), pp. 1638–1644.

[138] Jung Min Han, Heejoon Myung, and Sunghoon Kim. “Antitumor activity and pharmacokinetic properties of ARS-interacting multi-functional protein 1 (AIMP1/p43)”. In: *Cancer letters* 287.2 (2010), pp. 157–164.

[139] Yeon-Sook Lee et al. “Antitumor activity of the novel human cytokine AIMP1 in an in vivo tumor model.” In: *Molecules & Cells (Springer Science & Business Media BV)* 21.2 (2006).

[140] Sang Gyu Park et al. “Dose-dependent biphasic activity of tRNA synthetase-associating factor, p43, in angiogenesis”. In: *Journal of Biological Chemistry* 277.47 (2002), pp. 45243–45248.

[141] Myun Soo Kim et al. “Aminoacyl tRNA synthetase–interacting multifunctional protein 1 activates NK cells via macrophages in vitro and in vivo”. In: *The Journal of Immunology* 198.10 (2017), pp. 4140–4147.

[142] Bum-Joon Park et al. “The haploinsufficient tumor suppressor p18 upregulates p53 via interactions with ATM/ATR”. In: *Cell* 120.2 (2005), pp. 209–221.

[143] Julian Gregston Hurdle, Alexander John O’Neill, and Ian Chopra. “Prospects for aminoacyl-tRNA synthetase inhibitors as new antimicrobial agents”. In: *Antimicrobial agents and chemotherapy* 49.12 (2005), pp. 4821–4833.

[144] Julia Hughes and Graham Mellows. “Interaction of pseudomonic acid A with Escherichia coli B isoleucyl-tRNA synthetase”. In: *Biochemical Journal* 191.1 (1980), pp. 209–219.

[145] Takashi Nakama, Osamu Nureki, and Shigeyuki Yokoyama. "Structural basis for the recognition of isoleucyl-adenylate and an antibiotic, mupirocin, by isoleucyl-tRNA synthetase". In: *Journal of Biological Chemistry* 276.50 (2001), pp. 47387–47393.

[146] Guilherme Felipe Santos Fernandes, William Alexander Denny, and Jean Leandro Dos Santos. "Boron in drug design: Recent advances in the development of new therapeutic agents". In: *European Journal of Medicinal Chemistry* 179 (2019), pp. 791–804.

[147] James S Pham et al. "Aminoacyl-tRNA synthetases as drug targets in eukaryotic parasites". In: *International Journal for Parasitology: Drugs and Drug Resistance* 4.1 (2014), pp. 1–13.

[148] JB Koepfli, JF Mead, and John A Brockman. "Alkaloids of Dichroa febrifuga. I. Isolation and degradative studies". In: *Journal of the American Chemical Society* 71.3 (1949), pp. 1048–1054.

[149] Mark Pines and Itai Spector. "Halofuginone—the multifaceted molecule". In: *Molecules* 20.1 (2015), pp. 573–594.

[150] Sharon Mordechay et al. "Differential Effects of Halofuginone Enantiomers on Muscle Fibrosis and Histopathology in Duchenne Muscular Dystrophy". In: *International Journal of Molecular Sciences* 22.13 (2021), p. 7063.

[151] Michael R Linder et al. "(2R, 3S)-(+)-and (2S, 3R)-(-)-Halofuginone lactate: Synthesis, absolute configuration, and activity against Cryptosporidium parvum". In: *Bioorganic & medicinal chemistry letters* 17.15 (2007), pp. 4140–4143.

[152] European Organisation for Research and Treatment of Cancer - EORTC. *Halofuginone Hydrobromide in Treating Patients With Progressive Advanced Solid Tumors*. Accessed: 2021-10-27. 2012.

[153] National Cancer Institute (NCI). *Halofuginone Hydrobromide in Treating Patients With HIV-Related Kaposi's Sarcoma*. Accessed: 2021-10-27. 2013.

[154] Thomas A Wynn and Thirumalai R Ramalingam. "Mechanisms of fibrosis: therapeutic translation for fibrotic disease". In: *Nature medicine* 18.7 (2012), pp. 1028–1040.

[155] M Pines and A Nagler. "Halofuginone: a novel antifibrotic therapy". In: *General Pharmacology: The Vascular System* 30.4 (1998), pp. 445–450.

[156] Mark Pines et al. "Reduction in dermal fibrosis in the tight-skin (Tsk) mouse after local application of halofuginone". In: *Biochemical pharmacology* 62.9 (2001), pp. 1221–1227.

[157] Mark S Sundrud et al. "Halofuginone inhibits TH17 cell differentiation by activating the amino acid starvation response". In: *Science* 324.5932 (2009), pp. 1334–1338.

[158] Tracy L Keller et al. "Halofuginone and other febrifugine derivatives inhibit prolyl-tRNA synthetase". In: *Nature chemical biology* 8.3 (2012), pp. 311–317.

[159] Huihao Zhou et al. "ATP-directed capture of bioactive herbal-based medicine on human tRNA synthetase". In: *Nature* 494.7435 (2013), pp. 121–124.

[160] Jiangbin Ye et al. “GCN2 sustains mTORC1 suppression upon amino acid deprivation by inducing Sestrin2”. In: *Genes & development* 29.22 (2015), pp. 2331–2336.

[161] Wei Liu and James M Phang. “MiRNA and Proline Metabolism in Cancer”. In: *Oncogene and Cancer-From Bench to Clinic*. IntechOpen, 2013, pp. 360–390.

[162] Vance L Albaugh, Kaushik Mukherjee, and Adrian Barbul. “Proline precursors and collagen synthesis: biochemical challenges of nutrient supplementation and wound healing”. In: *The Journal of nutrition* 147.11 (2017), pp. 2011–2017.

[163] Rinat Abramovitch et al. “Halofuginone inhibits angiogenesis and growth in implanted metastatic rat brain tumor model-an MRI study”. In: *Neoplasia* 6.5 (2004), pp. 480–489.

[164] Michael Elkin et al. “Inhibition of bladder carcinoma angiogenesis, stromal support, and tumor growth by halofuginone”. In: *Cancer research* 59.16 (1999), pp. 4111–4118.

[165] Zohar Gavish et al. “Growth inhibition of prostate cancer xenografts by halofuginone”. In: *The Prostate* 51.2 (2002), pp. 73–83.

[166] Olga Genin et al. “Myofibroblasts in pulmonary and brain metastases of alveolar soft-part sarcoma: a novel target for treatment?” In: *Neoplasia* 10.9 (2008), pp. 940–948.

[167] David J Gross et al. “Treatment with halofuginone results in marked growth inhibition of a von Hippel-Lindau pheochromocytoma in vivo”. In: *Clinical cancer research* 9.10 (2003), pp. 3788–3793.

[168] Arnon Nagler et al. “Suppression of hepatocellular carcinoma growth in mice by the alkaloid coccidiostat halofuginone”. In: *European Journal of Cancer* 40.9 (2004), pp. 1397–1403.

[169] Yaqiu Wang, Zhihui Xie, and Hong Lu. “Significance of halofuginone in esophageal squamous carcinoma cell apoptosis through HIF-1α-FOXO3a pathway”. In: *Life Sciences* 257 (2020), p. 118104.

[170] Asuman Demiroglu-Zergeroglu et al. “Anticarcinogenic effects of halofuginone on lung-derived cancer cells”. In: *Cell Biology International* 44.9 (2020), pp. 1934–1944.

[171] Xiaojing Xia et al. “Halofuginone-induced autophagy suppresses the migration and invasion of MCF-7 cells via regulation of STMN1 and p53”. In: *Journal of cellular biochemistry* 119.5 (2018), pp. 4009–4020.

[172] Henry B Koon et al. “PHASE II AIDS MALIGNANCY CONSORTIUM TRIAL OF TOPICAL HALOFUGINONE IN AIDS-RELATED KAPOSI’S SARCOMA”. In: *Journal of acquired immune deficiency syndromes (1999)* 56.1 (2011), p. 64.

[173] Merav Leiba et al. “Halofuginone inhibits multiple myeloma growth in vitro and in vivo and enhances cytotoxicity of conventional and novel agents”. In: *British journal of haematology* 157.6 (2012), pp. 718–731.

[174] GP Soriano et al. “Proteasome inhibitor-adapted myeloma cells are largely independent from proteasome activity and show complex proteomic changes, in particular in redox and energy metabolism”. In: *Leukemia* 30.11 (2016), pp. 2198–2207.

[175] Ryutaro Adachi et al. "Discovery of a novel prolyl-tRNA synthetase inhibitor and elucidation of its binding mode to the ATP site in complex with l-proline". In: *Biochemical and Biophysical Research Communications* 488.2 (2017), pp. 393–399.

[176] M Tye et al. *Discovery of triple-site inhibitors for human and Plasmodium prolyl-tRNA synthetases*. 2021.

[177] Otis Pinkard et al. "Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation". In: *Nature communications* 11.1 (2020), pp. 1–15.

[178] David Sims et al. "CGAT: computational genomics analysis toolkit". In: *Bioinformatics* 30.9 (2014), pp. 1290–1291.

[179] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature biotechnology* 34.5 (2016), p. 525.

[180] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[181] Hai Fang et al. "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits". In: *Genome medicine* 8.1 (2016), pp. 1–20.

[182] Antonio Fabregat et al. "Reactome pathway analysis: a high-performance in-memory approach". In: *BMC bioinformatics* 18.1 (2017), p. 142.

[183] Minoru Kanehisa et al. "KEGG: new perspectives on genomes, pathways, diseases and drugs". In: *Nucleic acids research* 45.D1 (2017), pp. D353–D361.

[184] M Carlson. *org. Hs. eg. db: Genome Wide Annotation for Human. R package version 3.2. 3*. 2019.

[185] H Pagès et al. "AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor". In: *Bioconductor version: Release (3.10)* (2020).

[186] Steffen Durinck et al. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". In: *Nature protocols* 4.8 (2009), p. 1184.

[187] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90.

[188] Johannes Köster and Sven Rahmann. "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19 (2012), pp. 2520–2522.

[189] Leo Goodstadt. "Ruffus: a lightweight Python library for computational pipelines". In: *Bioinformatics* 26.21 (2010), pp. 2778–2779.

[190] Adam P Cribbs et al. "CGAT-core: a python framework for building scalable, reproducible computational biology workflows". In: *F1000Research* 8 (2019).

[191] Cole Trapnell, Lior Pachter, and Steven L Salzberg. "TopHat: discovering splice junctions with RNA-Seq". In: *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[192] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[193] Ali Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature methods* 5.7 (2008), p. 621.

[194] Marius Nicolae et al. “Estimation of alternative splicing isoform frequencies from RNA-Seq data”. In: *International Workshop on Algorithms in Bioinformatics*. Springer. 2010, pp. 202–214.

[195] Rob Patro et al. “Salmon provides fast and bias-aware quantification of transcript expression”. In: *Nature methods* 14.4 (2017), p. 417.

[196] Páll Melsted, Vasilis Ntranos, and Lior Pachter. “The barcode, UMI, set format and BUStools”. In: *bioRxiv* (2018), p. 472571.

[197] Páll Melsted et al. “Modular and efficient pre-processing of single-cell RNA-seq”. In: *BioRxiv* (2019), p. 673285.

[198] Avi Srivastava et al. “Alevin efficiently estimates accurate gene abundances from dscRNA-seq data”. In: *Genome biology* 20.1 (2019), p. 65.

[199] Davis J McCarthy et al. “Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R”. In: *Bioinformatics* 33.8 (2017), pp. 1179–1186.

[200] Tim Stuart et al. “Comprehensive Integration of Single-Cell Data”. In: *Cell* (2019).

[201] Cole Trapnell et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. In: *Nature biotechnology* 32.4 (2014), p. 381.

[202] Gioele La Manno et al. “RNA velocity of single cells”. In: *Nature* 560.7719 (2018), p. 494.

[203] Guangzheng Weng, Junil Kim, and Kyoung Jae Won. “VeTra: a tool for trajectory inference based on RNA velocity”. In: *Bioinformatics* 37.20 (2021), pp. 3509–3513.

[204] Volker Bergen et al. “Generalizing RNA velocity to transient cell states through dynamical modeling”. In: *Nature biotechnology* 38.12 (2020), pp. 1408–1414.

[205] Sung-Hou Kim et al. “Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain”. In: *Science* 179.4070 (1973), pp. 285–288.

[206] Megumi Shigematsu et al. “YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs”. In: *Nucleic acids research* 45.9 (2017), e70–e70.

[207] Anne Hoffmann et al. “Accurate mapping of tRNA reads”. In: *Bioinformatics* 34.7 (2018), pp. 1116–1124.

[208] Sara R Selitsky and Praveen Sethupathy. “tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data”. In: *BMC bioinformatics* 16.1 (2015), pp. 1–13.

[209] Phillipe Loher, Aristeidis G Telonis, and Isidore Rigoutsos. “MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data”. In: *Scientific reports* 7.1 (2017), pp. 1–20.

[210] Daniel Gebert, Charlotte Hewel, and David Rosenkranz. “unitas: the universal tool for annotation of small RNAs”. In: *Bmc Genomics* 18.1 (2017), pp. 1–14.

[211] Xiaogang Wu et al. “sRNAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline”. In: *Nucleic acids research* 45.21 (2017), pp. 12140–12151.

[212] Ling-Ling Zheng et al. “tRF2Cancer: a web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers”. In: *Nucleic acids research* 44.W1 (2016), W185–W193.

[213] Junchao Shi et al. “SPORTS1. 0: a tool for annotating and profiling non-coding RNAs optimized for rRNA-and tRNA-derived small RNAs”. In: *Genomics, proteomics & bioinformatics* 16.2 (2018), pp. 144–151.

[214] S Andrew. *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. Accessed: 2022-01-12. 2010.

[215] Steven W Wingett and Simon Andrews. “FastQ Screen: A tool for multi-genome mapping and quality control”. In: *F1000Research* 7 (2018).

[216] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[217] Philip Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–3048.

[218] Donna Karolchik et al. “The UCSC Table Browser data retrieval tool”. In: *Nucleic acids research* 32.suppl_1 (2004), pp. D493–D496.

[219] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome biology* 10.3 (2009), pp. 1–10.

[220] Yang Liao, Gordon K Smyth, and Wei Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2014), pp. 923–930.

[221] Sara R Selitsky et al. “Small tRNA-derived RNAs are increased and more abundant than microRNAs in chronic hepatitis B and C”. In: *Scientific reports* 5.1 (2015), pp. 1–7.

[222] Tiina Vilmi et al. “Sequence variation in the tRNA genes of human mitochondrial DNA”. In: *Journal of molecular evolution* 60.5 (2005), pp. 587–597.

[223] Brian Y Lin, Patricia P Chan, and Todd M Lowe. “tRNAviz: explore and visualize tRNA sequence features”. In: *Nucleic acids research* 47.W1 (2019), W542–W547.

[224] Heng Li et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[225] Luke Zappia, Belinda Phipson, and Alicia Oshlack. “Splatter: simulation of single-cell RNA sequencing data”. In: *Genome biology* 18.1 (2017), p. 174.

[226] H Pages et al. “Package ‘Biostrings’”. In: (2013).

[227] Alyssa C Frazee et al. “Polyester: simulating RNA-seq datasets with differential transcript expression”. In: *Bioinformatics* 31.17 (2015), pp. 2778–2784.

[228] Ni-Ting Chiou, Robin Kageyama, and K Mark Ansel. “Selective export into extracellular vesicles and function of tRNA fragments during T cell activation”. In: *Cell reports* 25.12 (2018), pp. 3356–3370.

[229] Tae-Dong Jeong et al. “Simplified flow cytometric immunophenotyping panel for multiple myeloma, CD56/CD19/CD138 (CD38)/CD45, to differentiate neoplastic myeloma cells from reactive plasma cells”. In: *The Korean Journal of Hematology* 47.4 (2012), pp. 260–266.

[230] Yael C Cohen et al. “Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing”. In: *Nature medicine* 27.3 (2021), pp. 491–503.

[231] Oksana Zavidij et al. “Single-cell RNA sequencing reveals compromised immune microenvironment in precursor stages of multiple myeloma”. In: *Nature cancer* 1.5 (2020), pp. 493–506.

[232] Jeremy Leipzig. “A review of bioinformatic pipeline frameworks”. In: *Briefings in bioinformatics* 18.3 (2017), pp. 530–536.

[233] Yue Cao, Pengyi Yang, and Jean Yee Hwa Yang. “A benchmark study of simulation methods for single-cell RNA sequencing data”. In: *Nature communications* 12.1 (2021), pp. 1–12.

[234] Serghei Mangul et al. “Systematic benchmarking of omics computational tools”. In: *Nature communications* 10.1 (2019), pp. 1–11.

[235] Yue You et al. “Benchmarking UMI-based single-cell RNA-seq preprocessing workflows”. In: *Genome biology* 22.1 (2021), pp. 1–32.

[236] Páll Melsted, Vasilis Ntranos, and Lior Pachter. “The barcode, UMI, set format and BUStools”. In: *Bioinformatics* 35.21 (2019), pp. 4472–4473.

[237] Hirak Sarkar et al. “Accurate, Efficient, and Uncertainty-Aware Expression Quantification of Single-Cell RNA-Seq Data”. In: (2020).

[238] A Booeshaghi and Lior Pachter. “Benchmarking of lightweight-mapping based single-cell RNA-seq pre-processing”. In: (2021).

[239] George C Linderman et al. “Zero-preserving imputation of single-cell RNA-seq data”. In: *Nature Communications* 13.1 (2022), pp. 1–11.

[240] Avi Srivastava et al. “Alignment and mapping methodology influence transcript abundance estimation”. In: *Genome biology* 21.1 (2020), pp. 1–29.

[241] Yawara Kawano et al. “Multiple myeloma cells expressing low levels of CD138 have an immature phenotype and reduced sensitivity to lenalidomide”. In: *International journal of oncology* 41.3 (2012), pp. 876–884.

[242] Radhika Bansal et al. “Impact of CD138 magnetic bead-based positive selection on bone marrow plasma cell surface markers”. In: *Clinical Lymphoma Myeloma and Leukemia* 21.1 (2021), e48–e51.

[243] William Matsui et al. “Characterization of clonogenic multiple myeloma cells”. In: *Blood* 103.6 (2004), pp. 2332–2336.

[244] Samantha Reid et al. “Characterisation and relevance of CD138-negative plasma cells in plasma cell myeloma”. In: *International journal of laboratory hematology* 32.6p1 (2010), e190–e196.

[245] Dan Wu et al. “CD138-negative myeloma cells regulate mechanical properties of bone marrow stromal cells through SDF-1/CXCR4/AKT signaling pathway”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1853.2 (2015), pp. 338–347.

[246] Nicolas Borisov et al. “Machine learning applicability for classification of PAD/VCD chemotherapy response using 53 multiple myeloma RNA sequencing profiles”. In: *Frontiers in oncology* (2021), p. 1124.

[247] Kazi Ferdous Mahin et al. “PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning”. In: *Genomics* (2022).

[248] Muhammad Kashif, Evren Alici, and Hareth Nahi. *Predicting Drug Resistance by Single-Cell RNASeq in Patients with Multiple Myeloma*. 2021.

[249] Travis S Johnson et al. “Development of a Novel Deep Transfer Learning Framework to Characterize Inter-and Intra-Tumor Heterogeneity in Myeloma Patients”. In: *Blood* 134 (2019), p. 3075.

[250] Fei Fei et al. “Metabolic markers for diagnosis and risk-prediction of multiple myeloma”. In: *Life Sciences* 265 (2021), p. 118852.

[251] Christian T Meyer et al. “Quantifying drug combination synergy along potency and efficacy axes”. In: *Cell systems* 8.2 (2019), pp. 97–108.

[252] Shuyu Zheng et al. “SynergyFinder Plus: towards a better interpretation and annotation of drug combination screening datasets”. In: *bioRxiv* (2021).

[253] Ana T Nunes and Christina M Annunziata. “Proteasome inhibitors: structure and function”. In: *Seminars in oncology*. Vol. 44. 6. Elsevier. 2017, pp. 377–380.

[254] David Melnekoff et al. “Single-cell RNA sequencing reveals distinct transcriptomic profiles of multiple myeloma with implications for personalized medicine”. In: *Blood* 130.Supplement 1 (2017), pp. 62–62.

[255] Guy Ledergor et al. “Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma”. In: *Nature medicine* 24.12 (2018), pp. 1867–1876.

[256] Duojiao Chen et al. “Cryopreservation Preserves Cell-Type Composition and Gene Expression Profiles in Bone Marrow Aspirates From Multiple Myeloma Patients”. In: *Frontiers in genetics* 12 (2021), p. 583.

[257] Pavel P Kotouček and Alberto Orfao. “Myeloma stem cell concepts, heterogeneity and plasticity of multiple myeloma.” In: *British journal of haematology* 166.3 (2014), pp. 466–467.

[258] Anoop P Patel et al. “Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma”. In: *Science* 344.6190 (2014), pp. 1396–1401.

[259] *inferCNV of the Trinity CTAT Project*. `https://github.com/broadinstitute/inferCNV`. Accessed: 2022-01-26.

[260] Audi Francesca Setiadi and Yuri Sheikine. “CD138-negative plasma cell myeloma: a diagnostic challenge and a unique entity”. In: *BMJ Case Reports CP* 12.11 (2019), e232233.

[261] Roberto J Pessoa de Magalhães et al. “Analysis of the immune system of multiple myeloma patients achieving long-term disease control by multidimensional flow cytometry”. In: *Haematologica* 98.1 (2013), p. 79.

[262] Jiang Wang et al. “Halofuginone functions as a therapeutic drug for chronic periodontitis in a mouse model”. In: *International Journal of Immunopathology and Pharmacology* 34 (2020), p. 2058738420974893.

[263] Keiji Kurata et al. *Pre-clinical validation of prolyl-tRNA synthetase as a novel therapeutic target in multiple myeloma*. 2022.

[264] Oksana Zavidij et al. “Single-cell RNA sequencing reveals compromised immune microenvironment in precursor stages of multiple myeloma”. In: *Clinical Lymphoma, Myeloma and Leukemia* 19.10 (2019), e27.