Using a multi-omics approach to understand the drivers of proteasome inhibitor resistance in multiple myeloma, and using epigenetic inhibitors to identify mechanisms of reversing the 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

Trinity 2021

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%. More novel 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. Recently, epigenetic mechanisms have been implicated in both the onset of MM and in the development of drug resistance. This thesis aims to investigate the changes that drive proteasome inhibitor drug resistance and to identify epigenetic compounds capable of reversing the resistance phenotype, and characterise their mechanism of action.

Following an epigenetic compound library screen and bulk RNA-seq, a dual TRIM24/BRPF inhibitor (TRIM24i) was selected to be investigated further as it was shown to kill carfilzomib-resistant AMO-1 cells (aCFZ) in the presence of carfilzomib but had little effect on PI-sensitive (WT) AMO-1 cells, demonstrating that it is capable of re-sensitizing carfilzomib-resistant AMO-1 cells to carfilzomib. Transcriptomic, epigenomic and proteomic changes were studied using an array of 'omic' techniques, including bulk and single-cell RNA-Seq, phosphoproteomics, ubiquitinomics, total proteomics, CyTOF and ChIP-Seq (PROBABLY will at some point).

Contents

List of Figures

List of Tables

List of Abbreviations

MM . . . . . Multiple Myeloma

BM . . . . . Bone marrow

MGUS . . . . Monoclonal gammopathy of unknown significance

SMM . . . . . Smoldering multiple myeloma

PI . . . . . Proteasome inhibitor

IMiDs . . . . . Immunomodulatory imide drugs

ER . . . . . Endoplasmic reticulum

UPS . . . . . Ubiquitin proteasome system

UPR . . . . . Unfolded protein response

DNA . . . . . Deoxyribonucleic acid

RNA . . . . . Ribonucleic acid

NGS . . . . . Next generation sequencing

WGS . . . . . Whole genome sequencing

RNA-Seq . . . Ribonucleic acid sequencing

scRNA-Seq . . Single cell RNA-Seq

dscRNA-Seq . Droplet-based scRNA-Seq

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

1 Introduction

1.1 Overview

...

1.2 The adaptive immune system

Humans are exposed to millions of potential pathogens every day and therefore require defences to be able to protect themselves against 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. Another example of an innate defence is non-specific engulfing (phagocytosis) of foreign pathogens by macrophages (a type of white blood cell). Innate responses are relied upon as the first line of defence, however sometimes a more sophisticated, specialised response is required- called the adaptive immune response. (REF-mol biology of the cell).

Adaptive immune responses are specific to the pathogen that induced the response and are dependent on B cells and T cells, two major classes of lymphocytes (a class of white blood cell). Two classes of adaptive immune responses exist: antibody responses, co-ordinated by B cells, and cell mediated immune responses, co-ordinated by T cells. T-cell-mediated immune responses recognise foreign antigens (antibody generators; substances capable of eliciting an immune response by stimulating B or T cell activation) on the surface of cells and can either kill the pathogen-infected cells or stimulate B cells or phagocytes to help eliminate the pathogen.

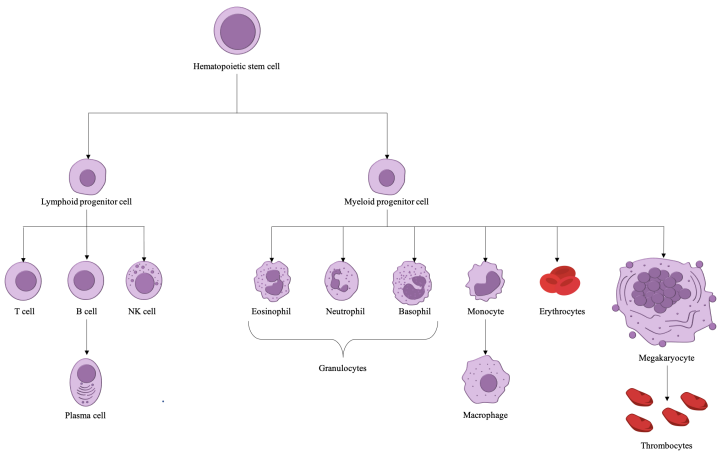


Figure 1.1: Hematopoietic stem cell (HSC) cell differentiation. HSCs divide into myeloid or lymphoid progenitor cells. Dendritic cells and a number of precursor states have been ommitted.

In antibody responses, B cells and plasma cells secrete antibodies, also known as immunoglobulins. Immunoglobulins are large Y-shaped proteins, which recognise and bind to the specific foreign antigen on the pathogen which stimulated their production. Binding of immunoglobulins 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.

1.2.1 Plasma cells

Plasma cell development

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. Hematopoietic stem cells (HSC) are stem cells that give rise to all the cells of the hematopoietic system. Two predominant cell populations are produced by HSCs: the common myeloid progenitor (CMP) and the common lymphocyte progenitor (CLP). CMP differentiation produces erythrocytes (red blood cells), mast cells, monocytes, macrophages, neutrophils, eosinophils, basophils and myeloid dendritic cells. CLP differentiation results in B cells, T cells, natural killer (NK) cells and lymphoid dendritic cells.

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[1, 2]. Upon antigenic-stimulation, naive B cells differentiate into memory B cells or plasma cells. Terminally differentiated plasma cells are the final effectors of the B cell lineage, each dedicated to secreting large amounts of a single type of antibody. Plasma cells have an extensive rough endoplasmic reticulum (ER), and have numerous genes involved in immunoglobulin secretion upregulated, including *XBP-1* and *CHOP*[3], to enable the production of copious amounts of antibody. 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, which have life-spans of decades and are mainly found in the bone marrow[4, 5].

1.3 Multiple myeloma

1.3.1 Multiple myeloma cells

Multiple myeloma is a malignancy of terminally differentiated plasma cells. It is characterised by aberrant proliferation of clonal, long-lived plasma cells in the bone marrow[6].

1.3.2 Epidemiology

Multiple myeloma accounts for 1-2% of all cancers and has the second highest incidence of hematological malignancies, after non-Hodgkin's lymphoma[7]. MM is rare in individuals under the age of 40, with the average age at time of diagnosis centering around 70[8, 9]. MM is more prevalent in males than females and is around twice as common in black populations than in Caucasian or Asian populations[10]. The average incidence rate is approximately 1-6 cases per 100000 individuals[8, 9,

11], with the highest age-standardised incidence rates in the regions of Australasia, North America, and Western Europe[12]. Five-year survival rate of MM patients is approximately 49%, whilst approximately a third of MM patients survive ten years or greater[13, 14].

1.3.3 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[15]. 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[16]. 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[17, 18].

Active MM

There are multiple classifications of active MM. The International Myeloma Working Group's definition[19] 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, CT or PET-CT

Biomarkers of malignancy include greater than or equal to 60% clonal plasma cells in the bone marrow, an involved:uninvolved serum free light chain ratio greater than or equal to 100, and more than one focal lesion on an MRI study[19].

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, a large array of up-regulated transcription factors, and numerous immune events.

1.3.4 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[20]. Novel therapeutic advances have contributed to this improvement (Table1.1).

1.3.5 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[28]. Since then it has been approved for use

Year	Treatment	Usage	Ref
1958	Melphalan	The alkylating agent melphalan was first used in plasma cell myeloma in 1958.	[21]
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.	[22, 23]
1980s	Stem-cell transplantations	Numerous successful allogenic and autologous bone marrow transplantations in patients with multiple myeloma	[24–27]
2003	Proteasome inhibitors	Bortezomib, a first-in-class proteasome inhibitor, was first approved by the FDA for use in relapsed and refractory multiple myeloma. In 2008 it was approved for patients with no prior treatment. Carfilzomib was approved in 2012 for advanced MM and later in 2015 for treatment of relapsed MM. The oral proteasome inhibitor, ixazomib, was approved as a combination treatment with lenalidomide and dexamethasone in 2016 for people who have received at least one previous treatment.	[28–30]
2006	IMiDs	The antitumour 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	[31–33]
2015	Monoclonal antibodies	In 2015, daratumumab, an anti-CD38 monoclonal antibody and elotuzumab, an anti-SLAMF7 monoclonal antibody, were approved for MM treatment.	[34, 35]

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

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[36]. The combination of bortezomib, dexamethasone and thalidomide was also shown to

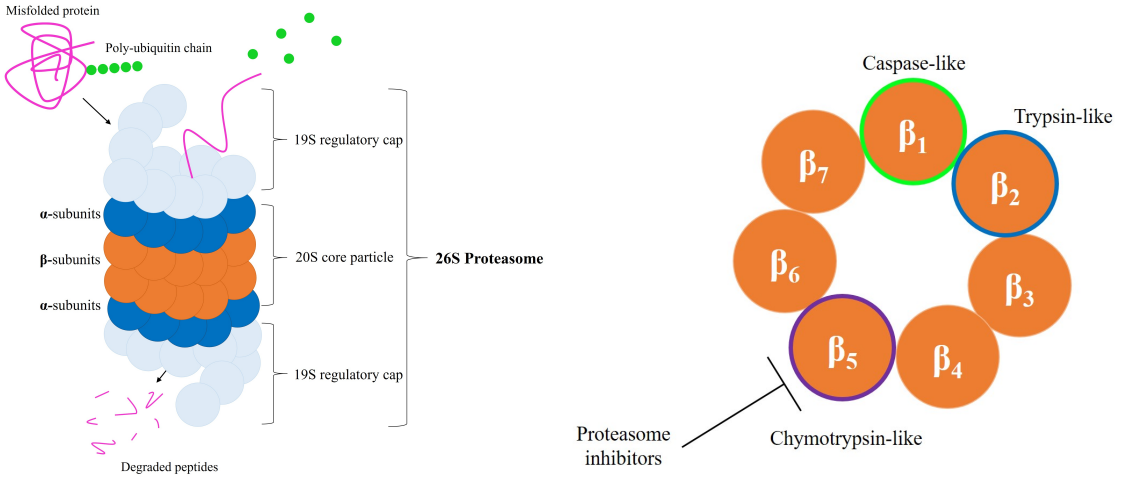
be superior to previous standard of care for patients prior to ASCT[37]. In 2010, bortezomib was approved as a frontline therapy for treatment-naive MM patients. Since then, two more proteasome inhibitors have been approved, carfilzomib and ixazomib. Carfilzomib is structurally and mechanistically different to bortezomib and shows activity on bortezomib resistant primary MM cells[37]; 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 1.2a). 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[38, 39] (Figures 1.2a and 1.2b).

1.4 Drug resistance in multiple myeloma

... Lit review??



(a) 26S proteasome

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

Figure 1.2: Structure of the proteasome. 1.2a 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. 1.2b shows the β-subunit arrangement in one of the inner rings of the 20s particle. $\beta 1$ (caspase-like), $\beta 2$ (trypsin-like) and $\beta 5$ (chymotrypsin-like) are the proteolytically active subunits. Proteasome inhibitors are designed to primarily inhibit $\beta 5$.

1.5 Transcriptomics, proteomics and epigenomics

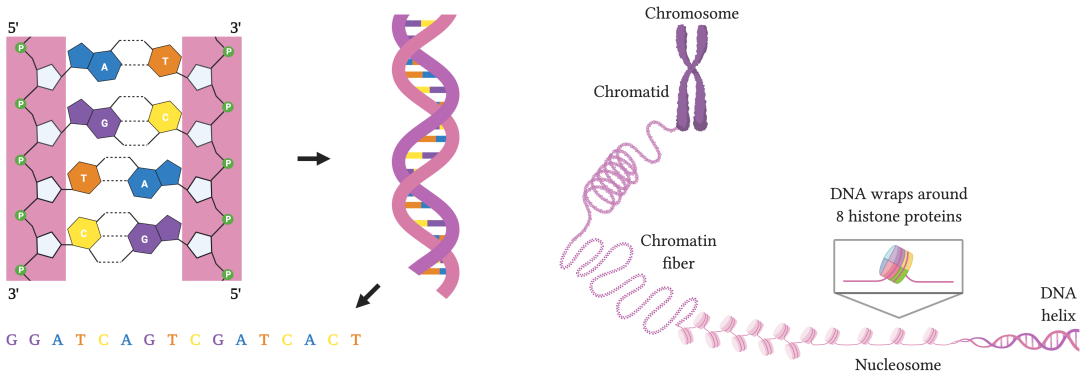
It has been shown that changes in the genome, transcriptome, epigenome and proteome all contribute to acquired-drug resistance in myeloma. Therefore, to sufficiently investigate the multiple layers driving this development of resistance, a multi-omics approach must be employed.

1.5.1 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 tasking 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 [40, 41]. Human cells contain 23 pairs of chromosomes.



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

Figure 1.3: 1.3a 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. 1.3b 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.
Created with BioRender.com.

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.

1.5.2 The epigenome

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

chromatin level and include DNA methylation, histone modification, and chromatin remodelling. These epigenetic changes are described in more detail below:

1.5.3 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 inplace of deoxyribose, and the nucleotide uracil (U) inplace of thymine (T). There are many types of RNA, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). RNA-seq is frequently used to study the transcriptome (outlined in section 1.5.5).

1.5.4 The proteome

Translation...

CyTOF and LC-MS/MS are often used to examine the proteome (section 1.5.9 and section 1.5.10).

1.5.5 RNA-seq

Modern RNA sequencing (RNA-seq) implements next generation sequencing (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

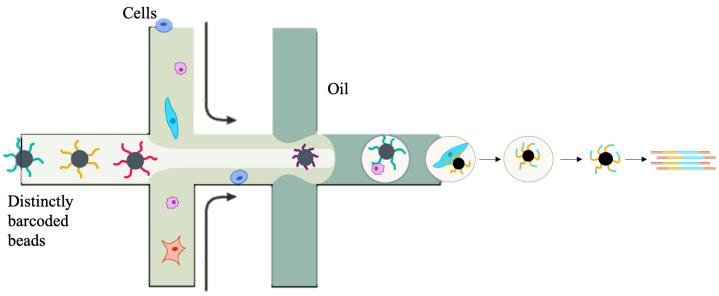


Figure 1.4: Outline of Drop-seq, a droplet-based scRNA-seq method. Figure adapted from Macosko et al. (2015) [44]. A microfluidic device combines two aqueous flows, one containing cells and the other containing barcoded primer beads suspended in lysis buffer. The two aqueous channels flow across an oil 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.

DNA (cDNA) library, end repair and adaptor ligation of the cDNA library, followed by PCR amplification 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[42], there have been numerous scRNA-seq techniques, such as SMART-seq2[43], Drop-seq[44], STRT[45] and inDrops[46]. scRNA-seq library preparation shares many steps with bulk RNA-seq workflow, however preliminary steps are required to isolate single cells and track them (??/ barcode) individually.

For droplet-based scRNA-seq (dscRNA-seq) methods, single cells are isolated using microfluidic devices by individually encapsulating them in aqueous droplets contained in oil. Below, a dscRNA-seq method, Drop-seq, is outlined (Figure 1.4).

1.5.6 ATAC-seq

...

1.5.7 ChIP-Seq

Chromatin immunoprecipitation sequencing (ChIP-seq) is used to analyse protein interactions with DNA. At a base-pair resolution, it is used to map DNA-binding proteins and histone modifications. Therefore ChIP-seq is often used to determine the the mechanisms of gene regulation of transcription factors, and to study epigenetic mechanisms in detail.

1.5.8 Next generation sequencing

Next generation sequencing (NGS), differs from its predecessors in that it is highly scalable and massively parallel. With NGS you can rapidly sequence the entire genome if desired. It is quicker and cheaper than traditional Sanger sequencing, and progressed data output from the kilobase range up to potentially multiple terabases per run.

1.5.9 CyTOF

...

1.5.10 Liquid chromatography with tandem mass spectrometry

Liquid chromatography with tandem mass spectrometry (LC-MS/MS) based proteomics is a popular analytical technique to measure the protein abundance of a sample. The general steps for LC-MS/MS-based proteomics include: cell lysis, protein extraction, protein digestion using an enzyme to cleave proteins into peptides, peptide purification, and analysis by mass spectrometry. The resultant data includes mass and charge (m/z) information and peak intensities. Software is then employed which performs database searches and calculates the most likely peptide for each peak. From this data, protein abundance can then be calculated and normalised.

LC-MS/MS-based proteomics can also be used to search for specific proteins within the proteome. For example, immobilized metal affinity chromatography (IMAC) can be used to enrich for phosphorylated peptides (phosphoproteomics),

and anti-ubiquitin antibodies can be used to enrich for ubiquitinated peptides (ubiquitinomics).

1.6 Summary

1.7 Aims

This thesis aims to characterise the changes driving proteasome-inhibitor resistance in multiple myeloma and identify possible mechanisms of reversing resistance. Chapter 2 gives background to the project, it reviews the literature surrounding drug resistance in MM and outlines preliminary work performed in the lab, that birthed this project. Chapter 3 outlines the methodology used in this work. Chapter 4 outlines the computational workflows generated to support this work and gives the results of a benchmark conducted to analyse the effectiveness of a computational pipeline developed. Chapter **??** describes how a lead epigenetic compounds was selected, analysing results from bulk RNA-seq data and validating previous compound screen results with dose response curves. Chapter **??** delves deeper into the drivers of changes in drug resistance and the mechanism of action of TRIM24i on resistant cells by looking at an array of multi-omics data.

2
Background

2.1 Drug resistance in MM

2.1.1 Genomic changes in drug resistant MM

2.1.2 Epigenetic changes in drug resistant MM

2.2 Preliminary work

2.2.1 Epigenetic compound library screen

REMAKE FIGURE— LOADS OF SPELLING MISTAKES AND LOOKS AWFUL!!! Previously in the Oppermann group laboratory, Dr James Dunford performed a compound screen against WT and proteasome inhibitor-resistant AMO-1 cells with an epigenetic compound library (figure 2.1) and cell viability assays.

The results from the screen can be seen in figure 2.2. Six compounds, shown circled in red, were identified as compounds of interest. SGC-CBP30, TRIM24i, OF1 and GSK959 are bromodomain inhibitors, TMP269 is a HDAC inhibitor and TDOSI000054a (T54) is a methyl lysine binder. These compounds were of interest as they had little to no effect on WT cells and decreased cell viability of resistant cells in the presence of their respective proteasome inhibitor. This indicated that the compounds were not just killing cells with their own distinct mechanism of action, but instead seemed to be reversing proteasome inhibitor resistance and re-sensitising resistant AMO-1 cells to proteasome inhibitors. These compounds were then taken

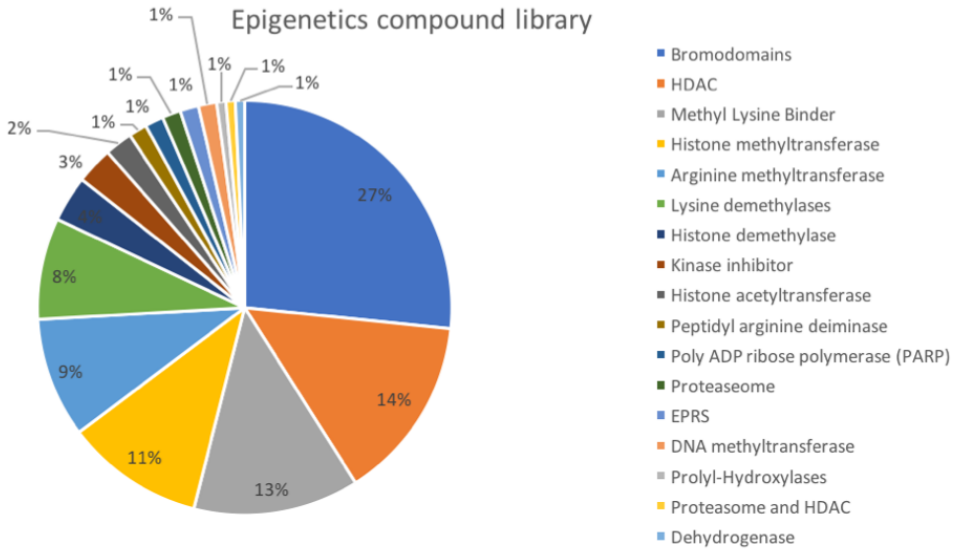


Figure 2.1: The Oppermann group epigenetic compound library. Proportions of targets in the 144-compound library.

forwards and used to treat carfilzomib resistant AMO-1 cells (aCFZ) and bulk RNA-seq perform to try and understand their mechanism of action.

Bromodomains (BRDs) are conserved modular protein-protein interaction domains. Primarily, BRDs recognise acetylated lysine (Lys) residues in histone tails. Proteins containing bromodomains are epigenetic regulators and regulate gene expression (on their own or as part of a larger complex) via chromatin remodelling, histone modification, histone recognition and transcriptional machinery regulation [REPHRASE]. Proteins containing BRDs have frequently been seen to be dysregulated in cancer. The compound SGC-TRIM24 is a dual inhibitor that targets the bromodomains of TRIM24 and BRPF.

(cite: https://www.nature.com/articles/nrm.2016.143)

Bromodomains are “readers” that bind acetylated lysines in histone tails

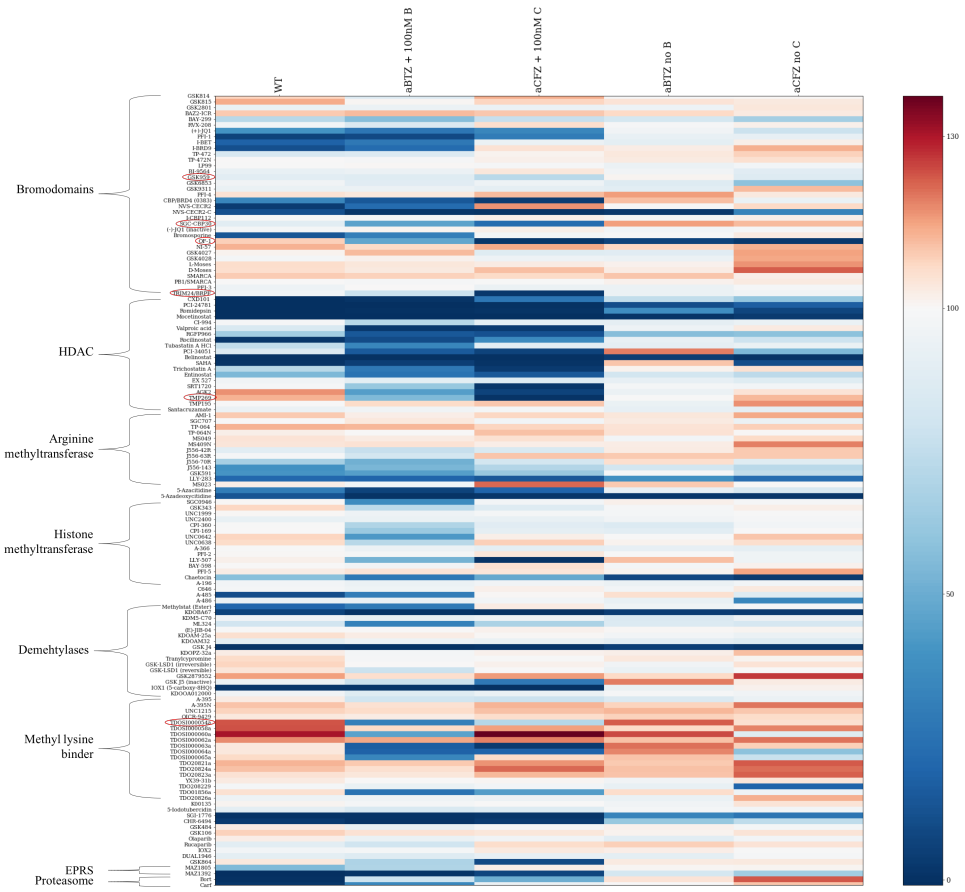


Figure 2.2: Preliminary epigenetic compound library screen, previously performed in the lab by Dr James Dunford. 138 epigenetic compounds were screened against proteasome inhibitor sensitive (WT) AMO-1 cells, carfilzomib resistant (aCFZ) and bortezomib resistant AMO-1 cells, in the presence of their respective proteasome inhibitor (+ 100nM B/C) and without (no B/C). Six compounds of interest were identified, circled in red in the figure.

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 (WT), bortezomib resistant AMO-1 cells (aBTZ) and carfilzomib resistant AMO-1 cells (aCFZ) were kindly gifted by the Driessen lab[47]. Bortezomib resistant, carfilzomib resistant, pomalidomide resistant (aPom) and dual bortezomib-pomalidomide resistant (aBTZPom) AMO-1 cells were also generated by Dr James Dunford by continual and escalating drug exposure of drug-sensitive (WT) cells. The Driessen lab AMO-1 resistant cells are commonly referred to as 'batch 1' in this work, and the in-house resistant AMO-1 cells as 'batch 2'. aCFZ and aBTZ cells were kept in 100nM of their respective proteasome inhibitor and aPom cells were kept in 6µM pomalidomide. AMO-1 cells were cultivated in RPMI-1640 medium (Thermofisher, UK), supplemented with 10% fetal bovine serum (FBS), 100µg ml^{-1} streptomycin and 100 U/ml penicillin (P/S) 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.

3.2 Compounds

3.2.1 Proteasome inhibitors

<WHERE were they obtained> etc etc.

3.2.2 Epigenetic inhibitors

The Oppermann group has an epigenetic compound screening library, consisting of 144 compounds. The compounds were obtained XYZ <where did Jim get compounds> SGC???? A dual TRIM24/BRPF inhibitor was identified as a possible candidate to reverse drug-resistance in AMO-1 cells. The structure of the inhibitor is shown below in figure 3.1.

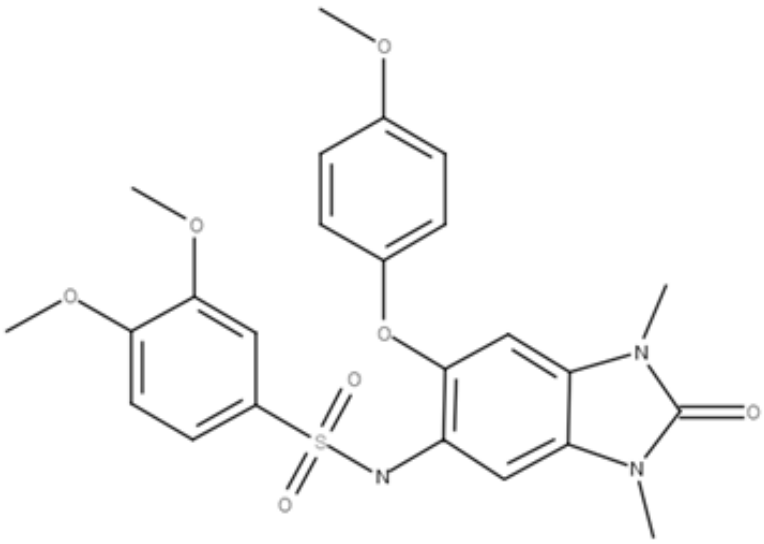


Figure 3.1: TRIM24 inhibitor chemical structure

3.3 Assays

3.3.1 Cell viability assays

10X presto blue (alamar??) was added in a 1:10 ratio to cells in suspension and incubated at 37°Cfor 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 10,000 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 in eppendorfs. Drugs were diluted 1 in 100 in 96 well round bottom plates with media and then 10µlwas added to the 90µlof seeded cells 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 extracted 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 10,000-16,000g 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. Samples were normalised to 100ng with nuclease-free water.

3.4.2 RNA library preparation

NEBNext® Ultra II directional RNA library prep kit for Illumina® with TruSeq indexes was used to prepare RNA libraries, following the manufacturer's protocol. RNA concentration was normalised to 100ng with nuclease-free water, made up

to 50µl. The NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, USA) was used to enrich poly-adenylated RNA. READ booklet in lab

The molarities of the libraries were determined by electrophoresis on a TapeStation (Agilent, USA).

3.5 Single-cell RNA-seq

3.5.1 Cell encapsulation

The Drop-Seq protocol[44] 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 or (ATDBio beads 2020 onwards!!!!! might need to change if reperform).

3.5.2 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.

3.6 ATAC-seq

3.6.1 Cell lysis

Approximately 2 million cells were collected in 15ml falcon tubes for each condition. The cells were centrifuged at 300g for 5 minutes at 4°C and the supernatant was discarded. The cell pellets were resuspended in 1ml of cold PBS and centrifuged at 300g for 5 minutes at 4°C, the supernatant was then discarded. Fresh lysis buffer was prepared (see tables **??**) with occasional gentle flicking. The falcons were then centrifuged at 500g for 10 minutes at 4°C. The supernatant (cytoplasm) was discarded, leaving the nuclei pellet.

Resuspension buffer	Volume (µl)
1M Tris-HCl (pH 7.5)	500
5M NaCl	100
1M $MgCl_2$	150
Nuclease-free water	49,250
Total	**50,000 (50ml)**

Table 3.1: Resuspension buffer recipe

Lysis buffer	Volume (µl)
Resuspension buffer	940
10% non-iodet P40	50
10% tween 20	10
Total	**1000 (1ml)**

Table 3.2: Lysis buffer recipe

3.6.2 Transposition

Pellets were resuspended in 890µl transposition mix (500µl 2X TD buffer, 330µl 1X PBS, 10µl 10% Tween-20, 10µl 5% Digitonin, 40µl nuclease-free water). For each condition, 176µl was taken in triplicate and transferred to LoBind 1.5ml eppendorfs (Eppendorf, UK). 4µl Tn5 enzyme was added to each eppendorf. The samples were then incubated at 37°C for an hour at 500rpm.

3.6.3 DNA purification

Magic bead clean-ups were performed to purify the DNA. 220µl of magic beads was added to each tube (1.2X), vortexed, centrifuged for 1-2 seconds and incubated at room temperature for 5 minutes. Tubes were placed on a magnetic rack for 2 minutes, until the solution was clear. The liquid from the tubes was aspirated away, leaving about 10µl of liquid remaining. 200µl of 80% ethanol was dispensed over the beads, the tubes were vortexed, spun and placed back on the magnetic rack until the solution was clear and then the ethanol was aspirated away. This wash was repeated for a total of two ethanol washes. Following aspiration on the 2^{nd}wash, an additional spin was performed and the tubes were placed back on the magnetic rack and any remaining liquid was aspirated away, to ensure all ethanol was removed. The beads were left to air dry for 3-5 minutes on the magnetic rack with the lids of

the tubes open. The tubes were removed from the magnetic rack and eluted with 26µl 0.1X TE buffer (Zymo Research, UK). The tubes were vortexed, spun and left to incubate for 5 minutes at room temperature, before being placed back on the magnetic rack. The eluant was transferred to fresh LoBind tubes. The purified DNA was then stored at -20°C until PCR amplification was ready to be performed.

3.6.4 PCR amplification

20µl of purified DNA from each sample was mixed with 20µl nuclease-free water, 5µl ATAC-seq universal primer, 50µl Nebnext high fidelity 2X master mix and 5µl unique ATAC-seq index primer, and split across two PCR tubes. The PCR tubes were put in a thermocycler with a lid temperature of 103.5°C, they were heated to 72°C for 5 minutes, 98°C for 30 seconds, and then cycled at 98°C for 10 seconds, 63°C for 30 seconds and 72°C for 1 minute, 13 times. Samples were then held at 4°C. The paired PCR tubes for each sample were then combined into single 1.5ml LoBind eppendorfs. Magic bead clean-up (as above) was performed, with 110µl magic beads (1.1X). The purified amplified DNA was eluted in 20µl 0.1X TE buffer and transferred to new LoBind tubes. D1000 high sensitivity screen tapes and 2200 TapeStation (Agilent, USA) were used to quantify libraries.

3.7 Pooling, denaturing and diluting libraries

Libraries were then denatured and diluted, folowing the NextSeq denature and dilute libraries guide, ready for sequencing.

3.8 Sequencing

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.

3.9 Phosphoproteomics

3.9.1 Collecting cell pellets

Greater than 20 million cells for each condition (in triplicate) was taken. The cell suspension was centrifuged at 1500g for five minutes. The supernatant was removed, the pellet was re-suspended in 500µl of ice-cold PBS, transferred to a 1.5ml eppendorf and centrifuged for a further five minutes. The supernatant was removed using a pipette and the pellet was stored at -80°C.

3.9.2 Cell lysis

300µl of fresh lysis buffer (10ml RIPA buffer, 3µl benzonase, 1 tablet phos stop) was added to each pellet, vortexed and left for 10 minutes on ice and then sonicated. The supernatant was transferred to a fresh tube.

3.9.3 Protein quantification

Protein concentrations were determined by BCA protein assay (Thermofisher, UK). 400µg of protein was taken from each sample. Samples were made up to a volume of 200µl with MilliQ-H_2O.

3.9.4 Protein Digestion

Kessler lab protocols were followed (`https://www.tdi.ox.ac.uk/research/research/tdi-mass-spectrometry-laboratory/mass-spectrometry/protocols-and-tools`). The lysed samples were reduced with 5µl of 200mM DTT in 0.1 M Tris buffer and incubated for 40 minutes at room temperature. The reduced samples were alkylated with 20µl of 200mM iodoacetamide in 0.1M Tris buffer, vortexed and then incubated for 45 minutes in the dark at room temperature. The protein was precipitated using methanol/chloroform extraction. The alkylated samples were transferred to 2ml eppendorfs. 600µl of methanol was added to each sample, followed by 150µl of chloroform and then vortexed gently. 450µl of MilliQ-H_2O was then added and vortexed gently. The samples were centrifuged at maximum speed on a table top centrifuge for one minute. The upper aqueous phase was removed, without disturbing

the precipitate at the interface. 450µl of methanol was added to each sample, without disturbing the disc and centrifuged for two minutes. Protein pellets were resuspended, one sample at a time: the supernatant was removed and 100µl of 6M urea in 0.1M Tris buffer was added. The samples were vortexed and then sonicated (???). Samples were diluted with 500µl MilliQ-H_2O, to ensure the final urea concentration was below 1M. Porcine trypsin (Sequencing Grade Modified Trypsin; Promega, USA) was added in a 1:50 ratio of enzyme:total protein content of sample, such that 40µl of trypsin solution containing 8µg trypsin in 0.1M Tris buffer was added to each sample. Samples were left to digest overnight at 37°C in an incubator shaker.

3.9.5 Peptide purification

The following day, the reaction was stopped, acidifying samples to 1% Trifluoroacetic acid (TFA). Samples were desalted and concentrated using 1ml C-18 Sep-Pak (Waters) cartridges. Two reagents were used: solution A (98% MilliQ-H_2O, 2% Acetonitrile (CH_3CN) and 0.1% TFA) for washing and solution B (65% Acetonitrile, 35% MilliQ-H_2O and 0.1% TFA) for activation and elution. The columns were flushed with 1ml of solution B and then washed with 1ml of solution A. The digested samples were added to the columns and vacuumed through slowly. Two 1ml washes with solution A were performed. Fresh, labelled eppendorfs were placed beneath the columns and peptides were eluted with 500µl of solution B. For phosphopeptide-enrichment, 90% of the peptides were removed for Immobilized Metal Affinity Chromatography (IMAC) on a Bravo Automated Liquid Handling Platform (Agilent). 10% of the peptides were used for total proteome analysis. Eluted peptides were dried using a vacuum concentrator (Speedvac, Eppendorf) and stored at -20°C until analysis by mass spectrometry (MS). Prior to MS analysis, dried peptides were resuspended in solution A.

3.10 Ubiquitinomics

3.10.1 Collecting cell pellets

100 million cells were taken for each condition in triplicate. The cell suspension was centrifuged at 1500g for five minutes. The supernatant was removed, the pellet was re-suspended in 500µl of ice-cold PBS and centrifuged for a further five minutes. The supernatant was removed and the pellet was stored at -80°C.

3.10.2 Cell lysis

PMTScan Ubiquitin Remnant Motif Kit (K-ε-GG; Cell signalling, USA) was used, following the manufacturer's protocol (REF). Pellets were solubilized and denatured in 4ml urea lysis buffer (20mM HEPES, pH 8.0, 9M urea, 1mM sodium orthovanadate, 2.5mM sodium pyrophosphate, 1mM β-glycerophosphate). The lysates were sonicated on ice, with two bursts of 15 seconds with a one minute break in-between.

3.10.3 Protein quantification

Protein concentrations were determined by BCA protein assay (Thermofisher, UK). All samples were found to contain between 10mg and 20mg of protein, so all of the available protein was used, with no normalisation.

3.10.4 Protein digestion

Lysates were reduced using dithiothreitol (DTT) at a final concentration of 4.5 mM for 30 minutes at room temperature. The reduced samples were alkylated using iodoacetamide (100mM final) for 15 minutes in the dark at room temperature. The alkylated samples were diluted four-fold with 20mM HEPES (pH 8.0) and digested with 400µl trypsin solution, containing $1\,\mathrm{mg\,ml^{-1}}$ trypsin-TPCK (Worthington, LS003744) in 1mM HCl. Samples were left to digest overnight at room temperature on a rotator.

3.10.5 Peptide purification

The following day, the reaction was stopped, acidifying samples to 1% Trifluoroacetic acid (TFA). Samples were desalted and concentrated using 10ml C-18 Sep-Pak (Waters) cartridges. The columns were activated using 5ml of solution B, washed with 10ml of solution A. The samples were added to the columns and ran through slowly. The peptides were washed with 10ml of solution A. The cartridges were then removed from the vacuum and the peptides were eluted into fresh falcon tubes with 6ml of solution B, using the plunger of the syringes. 20µg of digested protein was removed from each sample for matching total proteome analysis. The eluate was kept at -80°C overnight. The frozen peptide solutions were lyophilized for two days and then stored at -80°C.

3.10.6 Immunoaffinity purification

10x immunoaffinity purification (IAP) buffer provided with PTMScan Kit was diluted to 1x concentration with MilliQ-H_2O. Purified peptides pellets were resuspended in 1.4ml of IAP buffer by pipetting up and down and transferred to 1.7ml eppendorfs. The samples were centrifuged at 4°C for 5 minutes at 10000xg and kept on ice whilst preparing antibody beads. The anti-body bead slurry was centrifuged (30 seconds at 2000 g) and 1ml of PBS was added and then centrifuged. The supernatant was removed and the antibody beads were washed a further four times with PBS and resuspended in 40µl of PBS. The peptide solution was transferred to the antibody vial and the solution was incubated on a rotator for two hours at 4°C. The samples were centrifuged, put on ice and the supernatant was removed. The beads were washed twice with 1ml IAP, followed by three washes with 1ml chilled HLPC water. Immunoprecipitated material was eluted at room temperature in 55µl and 50µl 0.15% TFA in water, letting the sample stand for 10 minutes after each elution, with gentle mixing every two-three minutes. The eluates were centrifuged and the supernatant was transferred to new tubes. Peptide material was desalted and concentrated using 1ml C-18 Sep-Pak cartridges as above. Prior to mass spectrometry

analysis, purified GlyGly-modified peptide eluates and matching proteome material were dried by vacuum centrifugation, and re-suspended in solution A.

3.11 Liquid-chromatography-tandem mass spectrometry

Liquid-chromatography-tandem mass spectrometry (LC-MS/MS) analysis was performed using a Dionex Ultimate 3000 nano-ultra high pressure reverse-phase chromatography coupled on-line to an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific) (REF: adan's 3-5 dropbox). In brief, samples were separated on an EASY-Spray PepMap RSLC C18 column (500mm × 75µm, 2µm particle size; Thermo Scientific) over a 60 min (120 min in the case of the matching proteome) gradient of 2–35% acetonitrile in 5% dimethyl sulfoxide (DMSO), 0.1% formic acid at 250nl min^{-1}. MS1 scans were acquired at a resolution of 60000 at m/z 200 and the top 12 most abundant precursor ions were selected for high collision dissociation (HCD) fragmentation.

3.12 CyTOF

Get data off ADAM

3.12.1 CyTOF stuff

3.13 Data Processing

3.13.1 Bulk RNA-seq

Fasta files were processed using a CGAT-flow[48] 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[49], 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[50] was used for differential expression analysis of counts matrices (using negative binomial generalized linear models) within

the R statistical framework (v3.5.1). XGR[51], Reactome[52] and KEGG[53] were used to perform pathway analysis, within R. Org.Hs.eg.db[54], AnnotationDbi[55] and biomaRt[56] were used for converting between Ensembl IDs, HGNC symbols and ENTREZ IDs.

3.13.2 ATAC-seq

Raw ATAC reads (in fasta file format) were mapped to the GRCh38 reference genome using the CGAT-flow mapping pipeline (`https://github.com/cgat-developers/cgat-flow/blob/master/cgatpipelines/tools/pipeline_mapping.py`), using the mapper Bowtie. The mapped bam files were then used as input for the CGAT-flow peak calling pipeline ((`https://github.com/cgat-developers/cgat-flow/blob/master/cgatpipelines/tools/pipeline_peakcalling.py`)). Filtering was performed to filter out [what is filterd out!!] and peak calling was implemented using macs2 (v2.2.7)[57].

3.13.3 Single-cell RNA-seq

The computational pipeline outlined in section 4.2 was used to process scRNA data.

3.13.4 LC-MS/MS

Mass-spectrometry raw data were searched against the UniProtKB human sequence data base and label-free quantitation (LFQ) was performed using MaxQuant Software (v1.5.5.1). Digestion was set to trypsin/P. Search parameters were set to include carbamidomethyl (C) as a fixed modification, oxidation (M), deamidation (NQ), and phosphorylation (STY) as variable modifications. A maximum of 2 missed cleavages were allowed for phosphoproteome analysis and 3 for the GlyGly peptidome analysis, with matching between runs. LFQ quantitation was performed using unique peptides only. Label-free interaction data analysis was performed using Perseus (v1.6.0.2). Results were exported to Microsoft Office Excel and imported into the R statistical framework (v3.5.1) for further analysis.

4
Workflow Generation

4.1 Introduction

4.1.1 Reproducible workflows

In data analysis, particularly in bioinformatics, many users create simple bash or R scripts to execute the specific task at hand. However, if this is done often, the user can have an accumulation of these single-use scripts, which are often named uninformatively and never used again. Subsequently, the user may create scripts which perform the same function numerous times. Additionally, users may just use the command line alone to perform tasks. This means that exactly how they performed the analysis is difficult to find or not recorded. 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[58], Luigi and Ruffus[59].

For this work, a series of computational pipelines and workflows were generated. Ruffus and CGAT-core[60] 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[61] or STAR[62], 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[63]. 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[64]. 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[49]. Kallisto[49] and Salmon[65] 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[66] has been developed as an analysis tool and file format specifically for single-cell analysis, alongside BUStools, for processing of the resultant BUS file[67]. Salmon Alevin[68] 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[69] and alevinQC. Clustering is performed using Seurat3[70] and Monocle[71]. 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/`

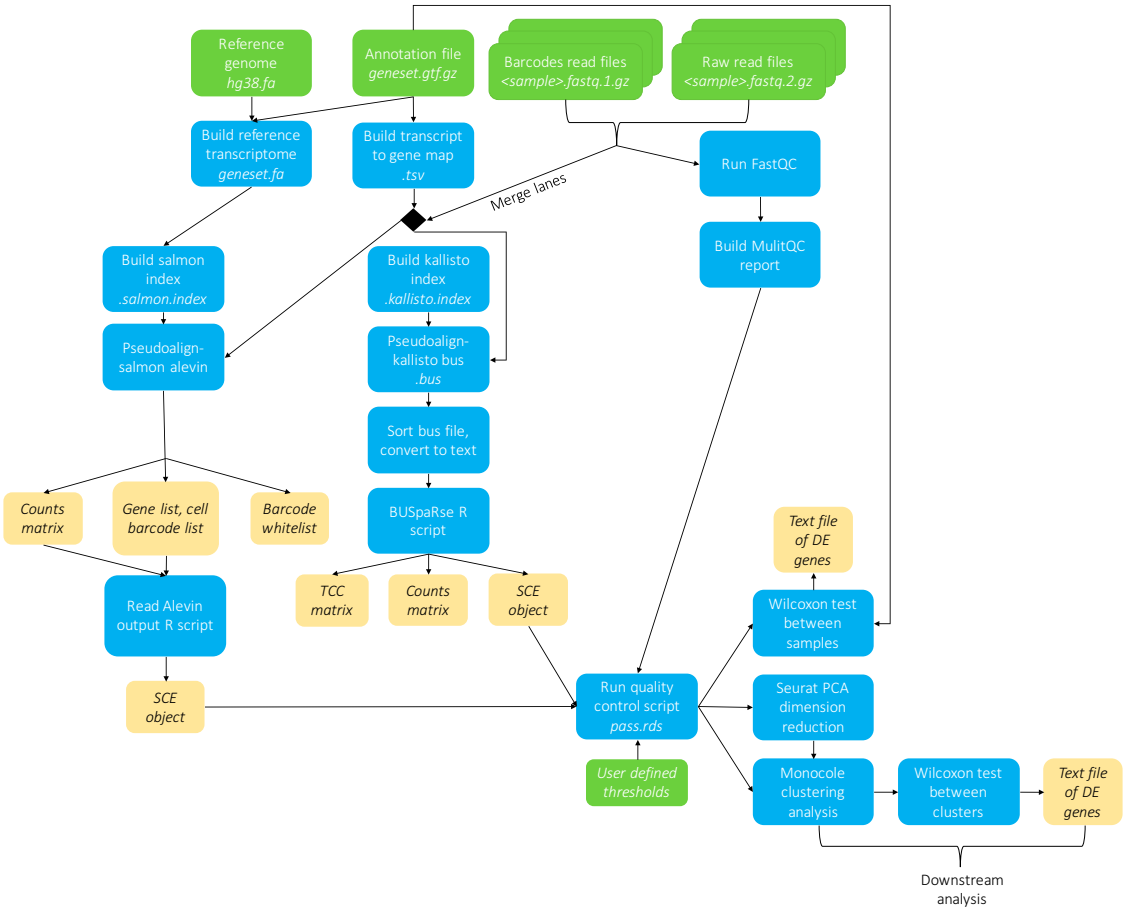


Figure 4.1: Flowchart outlining scRNA-Seq pseudoalignment pipeline- PLACEHOLDER- remake figure

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'.

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.

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

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

'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.

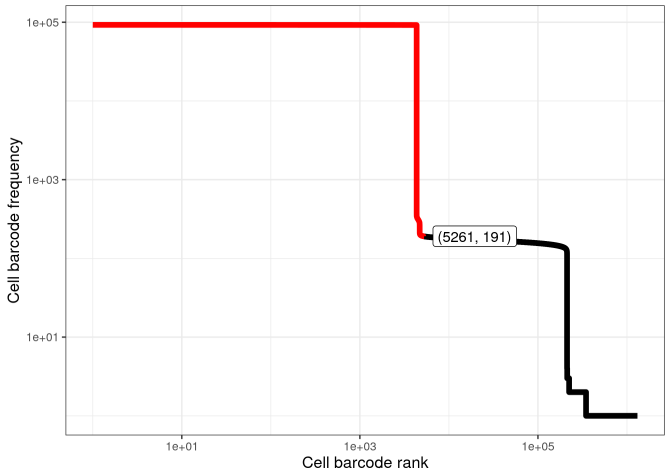


Figure 4.2: Alevin knee plot. This plot displays the number of times each cell barcode is observed, in decreasing order. Finding a 'knee' in this plot determines a threshold for the initial whitelist of CBs, which are unlikely to be empty droplets.

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.

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,

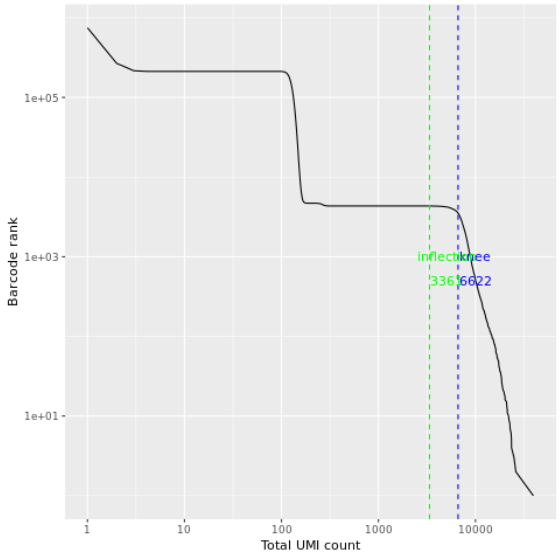


Figure 4.3: Kallisto BUS rotated knee plot. This plot shows the number of distinct UMIs against the rank of the barcode. The Pachter lab transpose the x- and y-axis on their knee plot, so that the x-axis displays distinct UMIs and the y-axis displays ranked cell barcodes, according to the number of corresponding 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.

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.

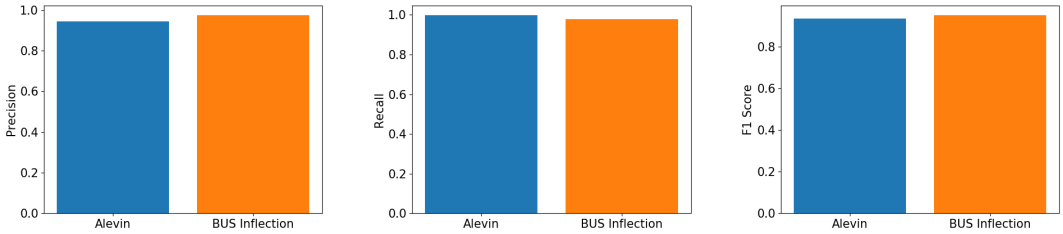
		Ground truth	
		Expressed	**Not Expressed**
Alevin /BUS	**Expressed**	True positive	False positive
	Not Expressed	False negative	True negative

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

No expression was denoted by 0, and expression by 1. When recall or precision

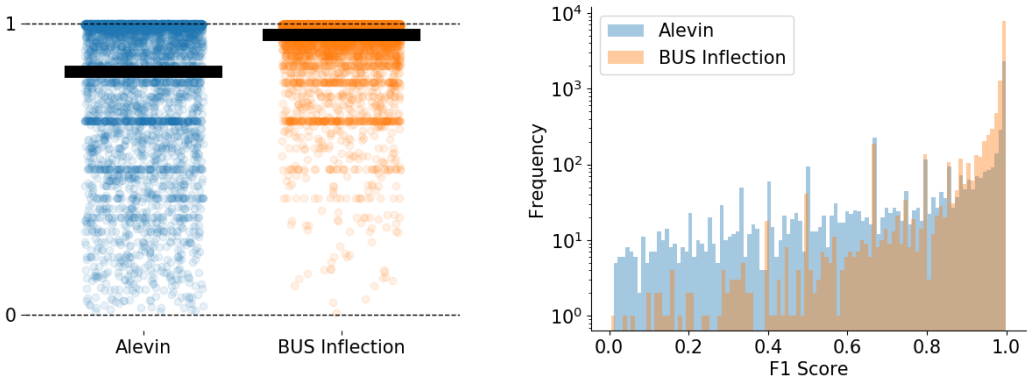
was undefined, i.e. a gene in Alevin/BUS matrix or the ground-truth matrix was not expressed by any cell, F score was defined as 0.

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 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.

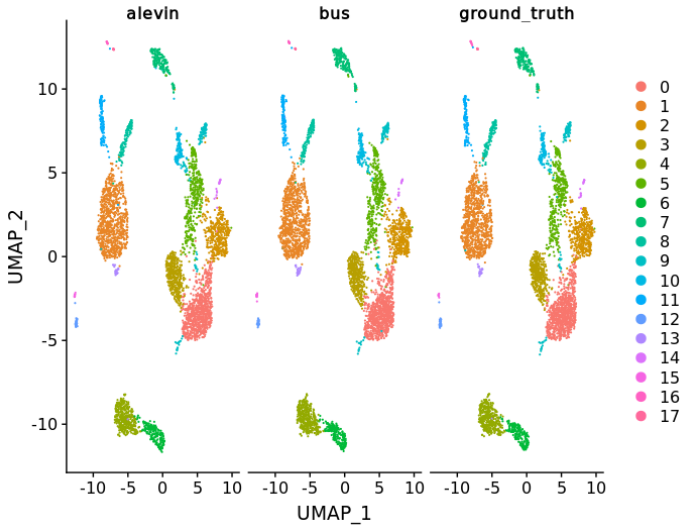


(a) F1 Score scatter graph **(b)** F1 score histogram

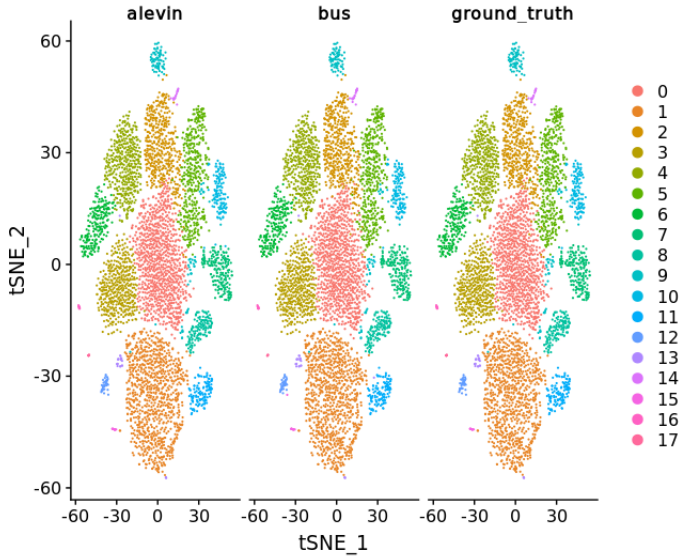
Figure 4.5: F1 score distributions. 4.5a shows the F1 score for each gene expressed across all 4339 cells. The black bar denotes the mean F1 score for each cell. F1 scores of 0 have been removed.

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



(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[70], using both Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (tSNE) dimension reduction techniques.

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.

4.3 scRNA-Seq velocity analysis pipeline

4.3.1 RNA velocity

“RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours"[72]. In combination with clustering analysis, the trajectory of a single-cell can be tracked.

approaches/system_approaches

reversal/trim24i_reversal

Appendices

A
Epigenetic compound screen

A compound screen consisting of approximately 140 epigenetic inhibitors was performed for AMO-1 cells.

References

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

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

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

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

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

[33] 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.

[34] 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.

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

[36] Jesús 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.

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

[38] 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.

[39] Bruce Alberts et al. *Molecular biology of the cell*. 6th ed. Garland science, Dec. 2014, pp. 356–359.

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

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

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

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

[44] 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.

[45] 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.

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

[47] 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.

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

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

[50] 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.

[51] 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.

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

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

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

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

[56] 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.

[57] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), pp. 1–9.

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

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

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

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

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

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

[64] 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.

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

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

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

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

[69] 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.

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

[71] 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.

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