Investigating Disease Progression and Therapeutic Targets in Multiple Myeloma Using Single-cell Technologies

Lijun Yao
Investigating Disease Progression and Therapeutic Targets in Multiple Myeloma
Using Single-cell Technologies
by
Lijun Yao

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2023
St. Louis, Missouri
# Table of Contents

List of Figures ......................................................................................................................... v

Acknowledgments .................................................................................................................... vii

Abstract of the Dissertation ...................................................................................................... x

Chapter 1: Introduction ............................................................................................................. 1

1.1 Multiple myeloma ............................................................................................................... 1

  1.1.1 Epidemiology and pathogenesis .................................................................................. 1

  1.1.2 Cytogenetic abnormalities ........................................................................................... 2

  1.1.3 Tumor microenvironment ............................................................................................. 3

  1.1.4 Heterogeneity in symptoms and prognosis .................................................................. 4

  1.1.5 Treatments .................................................................................................................... 5

1.2 Single-cell technologies ...................................................................................................... 6

  1.2.1 Limitations of bulk sequencing and advantages of single-cell techniques in myeloma research ........................................................................................................ 6

  1.2.2 CyTOF .......................................................................................................................... 7

  1.2.3 scRNA-seq ................................................................................................................... 8

  1.2.4 CITE-seq ..................................................................................................................... 8

1.3 Shortcomings in current multiple myeloma single-cell studies ........................................ 9

Chapter 2: Characterize immune microenvironment using multiple single cell technologies .. 11

2.1 Summary ........................................................................................................................... 12

2.2 Introduction ....................................................................................................................... 12

2.3 Results ............................................................................................................................. 14

  2.3.1 Patient characteristics and overview of CD45+ immune cells measured by scRNA-seq, CyTOF and CITE-seq ................................................................................... 14

  2.3.2 Advantages of CITE-seq in distinguishing T cell subtypes in multiple myeloma ...... 15

  2.3.3 Data reproducibility and comparisons of cell populations measured by the same technologies across different centers ................................................................. 16

  2.3.4 Comparisons of cell type abundances and correlations of cell type marker expression across the three techniques .................................................................................. 18

  2.3.5 Decreased ratio of CD4+/CD8+T cells from ISS stage 2 to ISS stage 3 patients and fast progression-related gene signatures ................................................................. 20

2.4 Discussion .......................................................................................................................... 21

2.5 Materials and Methods ..................................................................................................... 23

  scRNA-seq data quantification preprocessing ......................................................................... 27

  scRNA-seq cell type annotation ............................................................................................ 27
Our work has been accepted for publication: Yao et al. Single-Cell Discovery and Multi-Omic Characterization of Therapeutic Targets in Multiple Myeloma. Cancer Res 2023 Feb 13:CAN-22-1769; Contribution: I led the data analysis, results interpretation, figure generation, and manuscript writing as a first author of the paper.
4.3.2 Plasma cells subpopulations are associated with genomic alterations and disease progression ............................................................... 119
4.3.3 Delining B cells to plasma cells transition by trajectory analysis and mutation mapping ........................................................................ 121
4.3.4 Compositional alterations in lymphoid lineage during MM progression and associated gene sets ................................................................ 123
4.3.5 Compositional alterations in myeloid lineage during MM progression and associated gene sets ................................................................ 126
4.3.6 Cellular communications changes between baseline and relapse .......................................................... 128

4.4 Discussion .......................................................................................................................... 129

4.5 Methods ............................................................................................................................... 133
   Receptor–ligand interactions ............................................................................................ 138
   Monocle trajectory analysis .......................................................................................... 138

4.6 Figures ................................................................................................................................. 140

Chapter 5: Conclusions and future directions ................................................................. 155

5.1 Discussion ............................................................................................................................... 155

5.2 Future directions .................................................................................................................. 157
   5.2.1 Test preclinical efficacy of myeloma targets and explore potential bispecific CAR-T partners ........................................................................ 157
   5.2.2 Expand tumor cell associated targets discovery to multiple cancer types ............. 158
   5.2.3 Incorporate proteomics for characterizing multiple myeloma immune microenvironment ........................................................................ 159

References ............................................................................................................................. 161
List of Figures

Chapter 1
Figure 1. Overview of cell populations of 18 multiple myeloma patient samples subject to scRNA-seq, CyTOF, and CITE-seq. .................................................................36
Figure 2. Comparison of cell subset frequencies and correlations of expression of canonical cell type markers across different modalities. .........................................................38
Figure 3. Ratio of CD4+ T/CD8+ T of patients in different ISS stages and markers associated with ISS disease stages and multiple myeloma progression. ........................................39
Fig. S1. Expression of cell type markers in CITE-seq and comparison of cell subset abundance between technical replicates and across different centers. ............................................41
Fig. S2. Correlation of expression of canonical cell type markers across different modalities. ..42

Chapter 2
Figure 1. Myeloma cell associated therapeutic protein discovery workflow, datasets and overview. .........................................................................................................................86
Figure 2. Pathway overrepresentation analysis of plasma and B cell marker genes using the Kegg and Reactome databases. ......................................................................................88
Figure 3. Characterization of myeloma cell associated surface proteins. ............................89
Figure 4. Expression correlation of myeloma cell associated genes encoding surface proteins.92
Figure 5. Characterization of myeloma cell associated intracellular proteins. .......................94
Figure 6. Expression of myeloma cell associated therapeutic targets in relation to clinical features. .........................................................................................................................96
Figure 7. Comparison of scRNA-seq, bulk RNA-seq, and bulk proteomic expression of candidate targets. ..............................................................................................................99
Figure S1. Related to Figure 2 .................................................................101
Figure S2. Related to Figure 3 .................................................................102
Figure S3. Related to Figure 3 .................................................................105
Figure S4. Related to Figure 4 .................................................................107
Figure S5. Related to Figure 4 .................................................................109
Figure S6. Related to Figure 6 .................................................................109
Figure S7. Related to Figure 7 .................................................................111

Chapter 3
Figure 1. Datasets, genomic landscapes and immune profiles. ................................................142
Figure 2. Characteristics of plasma cells and their connection to disease progression. ..........144
Figure 3. The interaction between B cells and plasma, and the main dynamics involved in disease development. .............................................................................................146
Figure 4. Overview of T cell and NK cell lineages and characterization of fast progression enriched population. ..........................................................................................147
Figure 5. Characterization of T cell and NK cell populations associated with MM relapse. .....150
Figure 6. Overview of myeloid cell lineages and characterization of populations enriched for relapse samples. .............................................................. 152
Figure 7. Overview of cell-cell interactions in relation to disease progression. ................. 153
Acknowledgments

I would not have been able to complete my PhD without the support of many people in my life. My family, friends, and mentors have all played a crucial role in shaping who I am today, both personally and scientifically.

First, I want to express my gratitude to my PhD advisor, Li, who has been guiding me through the last few years of graduate school. She has given me the freedom to explore various directions in my research projects, allowing me to become an independent thinker. Her example of balancing multiple projects while minimizing stress has been inspiring. Beyond her academic mentorship, Li has been a great personal support to me.

I also want to thank my thesis committee, including Daniel Link, Todd A. Fehniger, Ravi Vij, and Malachi Griffith, for their valuable insight and feedback. John Dipersio was also very instrumental in my scientific growth and giving guidance when we characterized therapeutic targets in MM. In addition, I would like to express my gratitude to my Master’s program research advisor, Peggy Farnham, who introduced me to the field of cancer research and genomics, which sparked my interest and led me to pursue a PhD in the same field. Ting Wang, who helped me decide to pursue a PhD at WashU and gave me tremendous support and encouragement during the PhD.

I am grateful to my lab colleagues for creating a fun and stimulating work environment. My rotation mentor, Steven Foltz, taught me programming skills, statistical analysis and communication skills. Qingsong Gao, Amila Weerasinghe, and Sunantha Sethuraman, who were
postdocs in the lab, have challenged me to think critically and have taught me a lot about data analysis, result interpretation and developing projects based on existing literature. I also want to thank Reyka and Julia, whom I have been working closely with for several years since we were co-first authors on the MM therapeutic targets paper and collaborating on MMRF projects. Furthermore, I would like to express my gratitude to Yizhe, I-Ling, Song, Michael Iglesia, Andre, Yanyan, Dan Cui, Yige Wu, Liang-Bo Wang, Erik, Fernanda, Alla, Yize, Ruiyang, Matt W, Simon, Siqi, Ilya, Wagma, Clara, Omar, Austin, Rita, Hua, Kuan, Sohini, Wen-Wei Liang, Mike Wendl, and Mike M. for the enjoyable moments, laughter, and sharing of knowledge.

I want to acknowledge my close friends in graduate school, including Yujie, Yiqiao, Jiayang, Lingzhen, my HSG program mentor Lei Chen, my climbing friends including Yiran, Wenjun, Miwei, Tayyab, my close friends from Master’s program, including Yishu, Mengyao and Yuting, and many others. They have been an important source of support and companionship throughout my time in graduate school, and I have enjoyed sharing new experiences and traveling with them.

I would like to thank DBBS and WashU for providing me with this wonderful opportunity. I would like to thank NCI and NHGRI for providing funding for research in the Ding Lab.

Lastly, I would like to thank my family for always constantly inspiring me to reach my fullest potential and for their unconditional love and support. I am grateful for my grandparents, Zhengui Kang and Fengxian Shi, who raised me with love and care especially before my primary school. Their diligence and kindness helped me to become the person I am today. I am deeply grateful for my parents, Junfeng Kang and Bin Yao. I am always inspired by their curiosity to knowledge, passion for life, resilience, self-discipline and courtesy. Their unwavering support and
encouragement helped me achieve my goals. In addition, I am lucky to have my sister, Liping Yao, who played with me when I was little girl and goes shopping with me when I grow up. I hope all my family, friends and mentors can lead a happy and healthy life in the future.

Lijun Yao

Washington University in St. Louis

August 2023
Multiple Myeloma (MM) is a highly heterogeneous disease characterized by uncontrolled clonal expansion of plasma cells. Single-cell techniques are advantageous in providing a more granular understanding of inter- and intratumoral genomics and surrounding microenvironments. The high relapse rate and intrinsic complexity of MM make the application of single-cell technologies particularly beneficial. Understanding the concordance of the measurements across single cell techniques in MM is of great interest. In this dissertation, we first integrated three single-cell technologies, namely scRNA-seq, CyTOF, and CITE-seq, to characterize MM immune microenvironment and assess their concordances of measurement. Overall, cell type abundances were relatively consistent, while variations were observed in T cells, macrophages, and monocytes. In addition to immune profiling, we sought to discover tumor specific markers based on single-cell transcriptomic profiling. With better understanding of single cell technologies, we then leveraged a number of scRNA-seq datasets and developed a robust scRNA-seq driven tumor-marker discovery pipeline. In total, we identified 20 MM marker genes encoding cell-surface
proteins that are not yet under clinical study. The findings were cross-validated using different methods, including bulk RNA sequencing, flow cytometry, and proteomic mass spectrometry, on both MM cell lines and patient bone marrows. We also used both transcriptomic and immuno-imaging techniques to examine target dynamics and heterogeneity to identify potential combinatorial target partners. Lastly, we further characterized tumor heterogeneity, malignant B cell to plasma cell transitions, lineage compositional changes, and signature genes associated with MM progression by utilizing single-cell RNA sequencing of 361 samples from 263 MM patients in the Multiple Myeloma Research Foundation CoMMpass study. Interestingly, we identified B cell subpopulations as precancerous given their higher mutation burden. Additionally, we observed compositional alterations of immune subsets from baseline to relapse stages and identified differentially expressed genes associated with MM progression. Overall, this dissertation provides a comprehensive interrogation of tumor and the immune microenvironment in MM using single-cell technologies and proteomics, which deepens our understanding of MM disease onset and clinical outcomes and potentially provides novel targets for immunotherapies.
Chapter 1: Introduction

1.1 Multiple myeloma

1.1.1 Epidemiology and pathogenesis

Multiple myeloma, representing 10% of hematologic malignancies, is the second most common hematologic cancer in the United States\(^1\). Hematologic malignancies are the types of cancer beginning in blood-forming tissues or in the cells of the immune system. The three major types of hematologic malignancies are leukemia, lymphoma, and multiple myeloma (MM). New cases of these three types of cancers represented about 9.9% of all new cancer cases diagnosed in the US in 2020. In 2023, an estimated 35,730 adults in the US will be diagnosed with MM, representing ~0.76% of all cancer diagnoses. The overall 5-year survival rate for people with multiple myeloma was 54% in 2020\(^2\). There are clear gender and race disparities in MM incidences\(^3\). Incidence rates are higher among males than females, with 57% of cases in men and 43% in women. The age-standardized incidence rate of African-Americans is 2-fold higher than that of Caucasians. Incidence rates are similar for white Americans, Canadians, and in most European countries\(^4\).

MM is a disease characterized by clonal proliferation of malignant plasma cells (PCs), sometimes manifesting clinically with anemia, renal impairment, and pathologic bone fractures. Precursor stages of MM include monoclonal gammopathy of undetermined significance (MGUS) and smoldering myeloma (SMM), both of which are asymptomatic. As disease progresses, asymptomatic MM finally becomes symptomatic myeloma, leading to bone marrow infiltration
and osteolytic lesions. MGUS is characterized by the presence of a serum monoclonal proteins (M proteins) but M proteins are usually <3 g/dL. SMM, the more advanced disease stage than MGUS, is defined by serum M-protein > 3 g/dL and/or 10-60% clonal plasma cells, while multiple myeloma has ≥ 60% bone marrow plasma cells.

Over the past three decades, novel therapies, such as autologous hematopoietic cell transplantation, proteasome inhibitors (PIs), immunomodulatory drugs (IMiDs), and targeted monoclonal antibodies, have greatly improved the quality and length of life in patients with MM. Despite these advances, the disease remains incurable for most patients as it progresses even after complete remission and becomes resistant to these treatments, pointing to the urgent need for developing novel therapies.

1.1.2 Cytogenetic abnormalities

Cytogenetic abnormalities are present in most MM patients and have been revealed to be associated with disease characteristics and treatment response. Based on a study that investigated 2,027 patients, only 6% of patients do not have abnormalities detected by fluorescence in situ hybridization (FISH). Hyperdiploid (HD) or translocations of the immunoglobulin heavy chain (IgH) locus are commonly found in MM tumors. The most frequently observed primary IgH translocation is t(11;14), found in 19% patients, followed by t(4;14) and t(14;16), with t(6;14) and t(14;20) each found in ~1% of patients. Some recurrent translocations and copy number variations (CNVs), such as del17p, t(4;14) and t(14;16), have prognostic values and might guide treatment. t(4;14),
t(14;16), t(6;14), and t(14;20) are related to high-risk disease characteristics, including anemia and high bone marrow plasma cell content (50%). Notably, t(4;14) was associated with plasma cell proliferation and higher serum M-protein. Overall response rate to immunomodulatory drug (IMiD)-based treatment was higher for trisomies relative to IGH translocations, but was higher for IGH translocations for proteasome inhibitor (PI)-based treatment\textsuperscript{11}.

Most oncogenic mutations (63%) are accounted for just nine driver genes, including \textit{KRAS}, \textit{NRAS}, \textit{TP53}, \textit{FAM46C}, \textit{BRAF}, \textit{DIS3}, \textit{TRAFF}, \textit{SP140}, and \textit{IRF4}. 64\% of cases have a mutation in at least one of these nine genes. Of note, \textit{CRBM} and \textit{IKZF1} have been revealed to be associated with drug resistance\textsuperscript{12}.

\textbf{1.1.3 Tumor microenvironment}

The tumor microenvironment (TME) is a complex and dynamic ecosystem of cells and molecules that surround tumors, including fibroblasts, blood vessels, immune cells, and signaling molecules. TME has been widely implicated in cancer development, progression, and control\textsuperscript{13}. In other words, effective tumor surveillance by the host immune system protects against disease, but chronic inflammation and tumor “immunoediting” have also been implicated in disease development and progression\textsuperscript{14}. Accordingly, reactivation and maintenance of appropriate antitumor responses within the tumor microenvironment correlate with a good prognosis in cancer patients. Since MM mainly progresses in the bone marrow, bone marrow TME harbors tumor cells that interact with surrounding cells through the circulatory and lymphatic systems.

Myeloid cells are one of the major compartments in the bone marrow (BM) TME. Myeloid cells include monocytes, macrophages, dendritic cells (DCs), neutrophils, eosinophils, basophils, mast
cells, and platelets. A specific macrophage subset, M1 macrophages, plays an antitumoral role by secreting pro-inflammatory cytokines\textsuperscript{15}. By contrast, M2 macrophages have an immunosuppressive function that could promote MM progression\textsuperscript{16}. Apart from M2 macrophages, Myeloid-derived suppressor cells (MDSCs) also play an immunosuppressive role\textsuperscript{17}. They can be classified into granulocytic (G-MDSC) and monocytic (M-MDSC) and the frequency of M-MDSCs is associated with the abundance of M proteins\textsuperscript{18}. High-density neutrophils (HDNs) have been shown to be associated with reduction of T cell proliferation, suggesting their immunosuppressive effect in MM TME\textsuperscript{19}.

In addition to myeloid populations, T cells and Natural Killer (NK) cells are involved in the disease progression in several ways. Bryant et al. showed expansion of Th17 in Long-Term Survival (LTS)-MM subjects\textsuperscript{20} and observed that the Treg/Th17 ratio increased in MM patients, but was reduced in LTS-MM patients. Suen et al. showed an increase of CD8\textsuperscript{+} T cell clones, which have senescent characteristics but express low levels of exhaustion markers, has been related to improved prognosis in MM\textsuperscript{21}. Natural Killer (NK) cells become active in immune surveillance against viral infections and play an important role in cancer\textsuperscript{22}. The reported proportion of NK cells in MM is controversial. Some studies showed a decreased population of NK cells in MM patients, while other data showed the expansion of NK cells in MM\textsuperscript{23}.

1.1.4 Heterogeneity in symptoms and prognosis

MM is also known as a genetically complex and clinically heterogeneous disease, reflected by diverse symptoms and large variance of survival outcomes\textsuperscript{24}. Some patients experience extensive bone disease, while other patients experience cytopenias. A striking difference is also observed in progression-free survival (PFS). On both sides of the spectrum, some myeloma patients could be
functionally cured with PFS of more than 15 years\textsuperscript{25}, while about 15\% of patients have a median survival of less than 2 years\textsuperscript{26}. This clinical heterogeneity can be driven by heterogeneous tumor genetics and epigenetics, as well as by some extrinsic factors, such as the tumor microenvironment, disease stage, and host immune response, which make the disease very difficult to target therapeutically\textsuperscript{27,28,29,30,31,32,33}.

1.1.5 Treatments

The treatment strategies chosen depend on patients’ symptoms and disease stages. For asymptomatic MM, such as SMM, surveillance and doctor checkups might be sufficient. For symptomatic MM, major options include bone marrow stem cell transplantation, targeted therapies, radiation therapy, chemotherapy, immunomodulatory drugs, steroids, bone-modifying drugs, and immunotherapies.

Bone marrow transplant, also called stem cell transplant (SCT), is commonly used in MM. There are two different SCTs, autologous transplants and allogeneic transplants. In allogeneic SCTs, patients obtain stem cells from another person, whereas autologous transplants use patients’ own stem cells. In MM, most patients are treated by autologous SCTs. Due to the risk of toxic and even severe complications, SCT is not available to every patient. Transplant eligibility depends on patients’ age and overall health conditions. For example, SCT is primarily offered to patients under 65 years of age in most European countries, although there is no strict age restriction for transplants in the United States\textsuperscript{34}. 
Drug medications often precede SCTs to eliminate myeloma cells. Chemotherapy aims to destroy or slow down the growth of cancer cells. In recent years, chemotherapy has become less important since several newer types of drugs have become available. A few categories of these new drugs include:

1) Immunomodulatory drugs (IMiDs) enhance the activation and clonal expansion of T cells by stimulating both CD4+ and CD8+ T cells through phosphorylation of CD28. They also have the ability to induce cell death and hinder the function of cell adhesion molecules\(^{35}\). IMiDs include lenalidomide (brand name: Revlimid), pomalidomide (brand name: Pomalex), and thalidomide (brand name: Thalamid).

2) Proteasome inhibitors. Multiple myeloma arises from proliferation of malignant plasma cells generating antibodies. Proteasome inhibitors could inhibit tumor growth and lead to myeloma cell death by inducing accumulation of unfolded and misfolded proteins, turning on apoptosis pathways, and suppressing tumor survival pathways\(^{36}\). They include bortezomib, carfilzomib, and ixazomib.

3) Monoclonal antibodies. By targeting specific antigens on the surface of myeloma cells, monoclonal antibodies stimulate immunotherapeutic responses to kill tumor cells\(^{37}\).

1.2 Single-cell technologies

1.2.1 Limitations of bulk sequencing and advantages of single-cell techniques in myeloma research
Several landmark genomic investigations have used bulk sequencing technologies to broadly describe initiation and progression of multiple myeloma and where possible these abnormalities have been linked to disease prognosis. For example, to characterize diverse and complex myeloma tumor genetics, Skerget et al. have analyzed whole genome, exome, and transcriptome data of tumor samples collected at baseline and progression stages in the MMRF CoMMpass study to identify high-risk genetic subtypes and expression subtypes. However, bulk tumor sequencing can only address questions to the tumor populations, but not clonal genetic events within the tumor cells themselves. This is particularly true when tumor samples contain DNA from contaminating normal cells. Single-cell techniques allow for a more granular understanding of inter- and intratumoral genomics. Furthermore, tumor and its microenvironment could be obtained in parallel without first biasing our research by prior sorting. Moreover, single cell technologies are especially beneficial to mechanistic studies where both functional states and phenotypic traits are of interest.

1.2.2 CyTOF

Before the broad application of various sequencing technologies, Cytometry by time of flight (CyTOF) was widely used to identify different cell subsets within a complex population. Different from flow cytometry in which antibodies are labeled with fluorochromes, CyTOF uses antibodies conjugated to rare heavy metal isotopes that are not found in biological samples. Target expression is then quantified by mass spectrometric detection of these isotope labels. CyTOF has been widely used for biomarker discovery and immune profiling due to its ability to identify more than 40 parameters simultaneously, which exceeds the limit of multiplexing capability of FACS. CyTOF empowers obtaining comprehensive immunological information from small sample sizes. Previous studies have used CyTOF to characterize leukemia heterogeneity in relation...
to clinical features and to identify myeloid and lymphoid cell subtypes\textsuperscript{43}. Therefore, CyTOF has advantages to guide biomarker discovery and therapy outcome prediction by identifying disease associated microenvironment changes in cancer\textsuperscript{40}. However, CyTOF is limited by the number of metal isotopes (~50) conjugated to monoclonal antibodies. Therefore, it is not able to simultaneously examine the entire proteome that exists in a single cell.

1.2.3 scRNA-seq

In contrast to CyTOF, scRNA-seq provides whole transcriptomic profiles of each individual cell without constraints from marker selection, panel design, or antibody specificity\textsuperscript{44}. This is especially helpful to identify rare cell populations by assessing transcriptional differences between single cells. scRNA-seq also enables examination of highly diverse B-cell / T-cell receptors expressed by individual B or T lymphocytes\textsuperscript{45}, which is of great importance given that MM is a B-cell malignancy. Moreover, scRNA-seq is also ideal to be used in lineage tracing\textsuperscript{46} and identification of transitional cell states during cancer development and evolution\textsuperscript{47}. Importantly, in addition to dissecting cellular heterogeneity, scRNA-seq allows investigation of co-regulated gene modules and dynamics of gene-regulatory networks in a cell-type specific manner during treatment responses\textsuperscript{48}. Yet, scRNA-seq cannot distinguish different cell states with similar transcriptome profiles, highlighting the need for an alternative approach to delineate more cell types while comparing their transcriptome at the same time.

1.2.4 CITE-seq

By replacing fluorophore tagged antibodies with antibody-oligonucleotide conjugates, Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq) takes advantage of both
CyTOF and single-cell RNA sequencing (scRNA-seq). It allows simultaneous quantification of single-cell transcriptome and surface proteins. Since CITE-seq uses sequencing as a readout, there is no theoretical maximum number of protein markers that can be measured simultaneously. Up until now, over 200 markers have been utilized in applications\(^49\). Although CITE-seq resolved the restriction on the number of protein markers by utilizing sequencing based approaches, it has a few technical limitations. One major limitation is the current inability of CITE-seq to quantify intracellular epitopes. In addition, surface protein detection depends on availability of manufactured antibodies.

1.3 Shortcomings in current multiple myeloma single-cell studies

Taken together, because of the high relapse rate, various clinical features, and intrinsically complex genetics in MM, there is a strong need to better understand myeloma heterogeneity, disease progression, and to develop efficient therapeutic approaches. Application of single-cell technologies in MM is especially beneficial given MM is a highly heterogeneous disease with a complex immune environment. scRNA-seq has been used in several studies to understand this hematopoietic malignancy in both tumor and immune populations\(^50,51-55\) and CyTOF has been used to identify the expansion of novel memory B cells in MM\(^56\). CITE-seq is a relatively new technique and researchers have applied it to characterizing tumor heterogeneity in solid tumors\(^57\). However, there are no published studies using CITE-seq to characterize cell subsets in MM and the correlation of cell type population identified by these approaches is under investigation in MM. Moreover, several published myeloma single cell studies were focused on precursor stages of MM
rather than symptomatic MM progression\textsuperscript{51,52}, which will be comprehensively characterized with a dataset of unprecedented size in Chapter 4.

To evaluate the relationship between the cell surface proteome and cell intrinsic transcriptome, in Chapter 2 we assess bone marrow microenvironment by integrating three single-cell modalities. We examine the concordance and the correlation of cell type marker gene expression and compare the measurement of the immune compositions in multiple myeloma using the three techniques. Furthermore, we identify markers predicted to be significantly associated with the rapid progression of MM.

To identify potential myeloma-specific target antigens, in Chapter 3, we develop a pipeline to search for genes with specific expression in plasma and/or B cells using scRNA-seq data from 53 bone marrow mononuclear cells (BMMC) samples. Using Pearson correlation analysis for sample-level average expression between surface targets, we identify several co-expressed target pairs and mutually exclusive target pairs which could serve as dual targeting partners to improve efficacy. Finally, we validate candidate therapeutic targets by cross-referencing bulk RNA-seq, flow cytometry, and mass spectrometry data.

Finally, we investigate myeloma disease progression by comprehensively characterizing the CD138 negative fraction in BMMC and activity of key immune cell populations in 361 scRNA-seq data from myeloma patients in Chapter 4. We delineated the tumor heterogeneity and their transformation from B cells to plasma cells during the progression of Multiple Myeloma by
trajectory analysis. Furthermore, we investigate tumor microenvironment changes in relation to disease progression and related gene sets. We hope this study will deepen our understanding of myeloma disease progression, provide novel targets for immunotherapies, and clinically stratify patients and determine strategies for early intervention in the clinic.

Chapter 2: Characterize immune microenvironment using multiple single cell technologies

Our work has been accepted for publication: Yao et al. Comprehensive characterization of the multiple myeloma immune microenvironment using integrated scRNA-seq, CyTOF, and CITE-seq analysis. Cancer Research Communications 2023 Feb 13:CAN-22-1769; Contribution: I led the data processing and analysis of the entire project and coordinated the collaborations with other
researchers as a first author of the paper. I played a major role in preparing the results section by planning, analyzing the data, creating figures, and writing the manuscript.

2.1 Summary

As part of the Multiple Myeloma Research Foundation (MMRF) immune atlas pilot project, we compared immune cells of Multiple Myeloma (MM) bone marrow samples from 18 patients assessed by single-cell RNA-seq (scRNA-seq), mass cytometry (CyTOF), and Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) to understand the concordance of measurements among single-cell techniques. Cell type abundances are relatively consistent across the three approaches, while variations are observed in T cells, macrophages, and monocytes. Concordance and correlation analysis of cell type marker gene expression across different modalities highlighted the importance of choosing cell type marker genes best suited to particular modalities. By integrating data from these three assays, we found International Staging System (ISS) stage 3 patients exhibited decreased CD4+ T/ CD8+ T cells ratio. Moreover, we observed upregulation of RAC2 and PSMB9, in NK cells of fast progressors (FP) compared to those of non-progressors (NP), as revealed by both scRNA-seq and CITE-seq RNA measurement. This detailed examination of the immune microenvironment in MM using multiple single cell technologies revealed markers associated with MM rapid progression which will be further characterized by the full-scale immune atlas project.

2.2 Introduction

Single-cell sequencing technologies offer advantages over traditional bulk methods in cancer genomics research for evaluating cellular heterogeneity and investigating evolution of cellular
subpopulations between the tumor and its microenvironment. For example, single-cell methods have been extensively applied to Multiple Myeloma (MM), a highly heterogeneous disease marked by uncontrolled clonal expansion of plasma cells. Single-cell RNA sequencing (scRNA-seq) has been used to examine tumor and immune cell populations\(^{51,58}\) and mass cytometry (CyTOF) to evaluate the impact of drugs on immune populations in MM\(^{59}\). The third technology, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), is a more recent, multimodal approach with simultaneous quantification of single-cell transcriptomes and surface proteins. All three approaches enable identification of cell types, cell states, and characterization of cellular heterogeneity at transcriptomic and/or protein levels. Consequently, understanding their concordances across technologies is of great practical interest.

In addition, the bone marrow microenvironment plays an important role in the evolution of premalignant MM, MM progression and treatment response. Single-cell transcriptomics analysis of the tumor microenvironment revealed compositional alterations begin at the Monoclonal gammopathy of undetermined significance (MGUS) stage, including enrichment of T cells, NK cells and CD16+ monocytes\(^{51}\). Specifically, the percentage of CD4+ T cells was significantly reduced in bone marrow of MM patients, leading to altered CD4+ T/CD8+ T ratio\(^{60}\). When comparing the clinical status, the ratio decreased in International Staging System (ISS) stage 3 patients compared to stage 1 patients\(^{61}\). With respect to treatment, the proportion of CD3+ T cells was lower in treated patients compared to chemo-naive MM patients\(^{62}\). Further work is needed to expand initial findings using various assays and reveal candidate markers for characterizing clinical features of MM patients and optimizing treatment.

Combining the timeliness of the technology concordance question with furtherance of MM research, we subjected bone marrow samples from 18 MM patients to scRNA-seq, CyTOF, and
CITE-seq, examining the similarities across the aforementioned single cell techniques. We used the results to investigate the relationship between immune population compositional alterations and disease stages and revealed a set of markers associated with MM rapid progression.

2.3 Results

2.3.1 Patient characteristics and overview of CD45+ immune cells measured by scRNA-seq, CyTOF and CITE-seq

We used eighteen cryopreserved Multiple Myeloma (MM) samples of CD138-negative “immune cell” fractions from patients enrolled in the Multiple Myeloma Research Foundation (MMRF) CoMMpass study (NCT01454297). Nine were fast progressors (FP, progressed within 6 months) and nine were non-progressors (NP, progressed >6 months but within 5 years) with patient ages ranging from 37 to 83 years. Twelve patients were in the International Staging System (ISS) stage III, eight underwent Autologous Stem Cell Transplantation (ASCT), eleven were females and fifteen were Caucasions (Fig. 1A). Each sample was subjected to scRNA-seq, CyTOF, and CITE-seq at three different respective academic research centers, namely Washington University in St. Louis (WUSTL), Icahn School of Medicine at Mount Sinai (ISMMS), and Beth Israel Deaconess Medical Center (BIDMC). All sites received aliquots from the same sample and technical replicates were conducted for 2 samples for each assay (Fig. 1A).

To assess immune cell composition of MM patients, bone marrow (BM) baseline samples (collected at the initial diagnosis) from these eighteen patients were subjected to scRNA-seq, with immune cells clustered based on their transcriptome profiles using the Louvain clustering algorithm implemented by Seurat63,64 (Fig. 1B). We then investigated immune cells of these same
samples by mass cytometry (CyTOF) using a 39-marker panel. Cell populations were characterized by expression of markers, clustered by the flowsom algorithm\textsuperscript{65}, and visualized with vi-SNE in the Cytobank\textsuperscript{66} platform (Fig. 1C). Given the discordance between RNA expression and protein expression that is known to exist\textsuperscript{67}, it is informative to characterize cell populations by measuring RNA and protein at the same time. Finally, we utilized CITE-seq with antibody-oligonucleotide conjugates and 29 protein markers to simultaneously quantify single-cell transcriptomes and surface proteins. Following standard scRNA-seq quality filtering protocols, immune cells were clustered based on integrated multi-omic profiles by the similarity network fusion (SNF) integration algorithm in CiteFuse\textsuperscript{68} (Fig. 1D). From CD138-negative BM aliquots, we detected, on average, 1,051 immune cells/sample using scRNA-seq, >64K CD45+ cells/sample using CyTOF, and 718 immune cells/sample using CITE-seq.

2.3.2 Advantages of CITE-seq in distinguishing T cell subtypes in multiple myeloma

To assess the potential advantages of simultaneous quantification of RNA and protein expression in CITE-seq as compared to standard scRNA-seq, we labeled immune cell identities determined by integrated transcriptome and protein expression, but clustered cells by transcriptional profiles alone (Fig. 1E). Interestingly, most cell types, including B cells, monocytes, macrophages, neutrophils, and plasmacytoid dendritic cells (pDCs), formed distinct clusters, while T cell subtypes mixed together. To further understand the difference of cell type marker expression between the RNA and protein levels, we visualized the expression of some canonical markers in Uniform Manifold Approximation and Projection (UMAP) and investigated the concordance of
the sample-level average expression of the 29 CITE-seq protein markers between RNA level and Antibody-Derived Tags (ADT) level (Fig. 1F, G; Supplementary Fig. S1A). As expected, expression levels of markers are generally concordant \((R =0.72, p < 10^{-4})\), with some exceptions where protein-level expression is higher than RNA-level expression and vice versa. One impressive example is CD4 (Fig. 1F, G), which is highly expressed at ADT measurement, but minimally expressed at the RNA level, mainly because mRNAs are produced at much lower rates and have much shorter half-lives than proteins\(^{69}\). This observation is consistent with previous studies showing low CD4 mRNA expression compared to surface CD4 protein\(^{70}\). Last, since Naive CD8+ T cells were clustered together with CD4+ T cells based on transcriptome profiles (Fig. 1E), we investigated whether reclustering T cells alone could help to distinguish subtypes at the RNA level. Due to the high similarities of transcriptional profiles among T cells\(^{70}\) and different surface protein markers could be encoded by the same gene\(^{71}\), reclustering CD4+ and naive CD8+ T cells did not provide additional resolution of T cell subtypes (Fig. 1H). Consistent with a published study about renal T subtype identification using CITE-seq\(^{72}\), our observation emphasizes the advantage of integrating protein-level expression of cell type markers for MM T cell subtype identification in CITE-seq as compared to standard scRNA-seq.

### 2.3.3 Data reproducibility and comparisons of cell populations measured by the same technologies across different centers

To examine data reproducibility, percentages of cell subsets in CD45+ populations were compared between technical replicates for 2 samples in each assay. The technical replicate pairs are strongly correlated in all 3 assays (average Pearson correlation coefficient \(r=0.94\) in scRNA-seq, 0.89 in...
CyTOF, and 0.92 in CITE-seq) (Supplementary Fig. S1B-D). Next, to examine the consistency of immune cell populations measured by the same techniques at different sites, we evaluated the percentage of immune populations captured by three centers using four samples. scRNA-seq data was generated in ISMMS, WUSTL and BIDMC using aliquots of the same samples and CyTOF data was generated in ISMMS, Mayo Clinic and Emory University. BIDMC scRNA-seq data is from CITE-seq data analyzed with RNA signal alone. (Supplementary Fig. S1E). We observed that the percentages of B cells, pre-B cells, NK cells, pDCs, monocytes and macrophages are generally consistent, while the T cell subset varies across centers in scRNA-seq measurement (Supplementary Fig. S1F). This suggests that T cell composition could vary by aliquots and potential sample processing differences across centers while other cell types are more similar in scRNA-seq measurement. The cell type abundance measured by CyTOF is less variable than that measured by scRNA-seq, with smaller differences observed in T cell subsets across centers (Supplementary Fig. S1G, mean difference calculated by Bland-Altman analysis). Moreover, cell subset abundances of ISMMS samples tend to have less variation likely due to the benefit of barcoding samples (Methods). The cell type frequencies calculated by one center (Emory) tend to be lower overall compared to other centers in CyTOF, probably because wide bore injector assembly with cell acquisition solution was not used to maintain cell integrity (Methods). It is worthwhile noting that including reference samples in CyTOF is very helpful for identifying potential artifacts. For example, we observed a big proportion of CD66b/CD3+ cells in patient samples while these were absent in the reference sample from a healthy donor (data not shown). We hypothesized that this CD66b staining artifact (CD66b is not expressed on CD3+ T cells) was likely due to non-specific staining from dead cells. Indeed, the percentage of CD66b/CD3+ cells dropped dramatically after dead cell depletion. Lastly, to evaluate the similarity of expression
profiles across different samples and centers, we calculated the Pearson correlation coefficient of expression of the B cell markers between populations detected from different centers using scRNA-seq (Supplementary Fig. S1H). We observed that B cells clustered according to patients instead of centers, suggesting patient-dependence of B cell transcriptome profiles, likely because B cells are potential reservoirs of plasma cells. Overall, we observed that cell type abundances are generally consistent across centers for most cell types and that similarity of transcriptome profiles of immune populations is center-independent, suggesting absence of strong batch effects across centers. These observations imply that our cross-technique comparisons should be valid.

2.3.4 Comparisons of cell type abundances and correlations of cell type marker expression across the three techniques

To evaluate the concordance of cell type composition determined by the three methods, we calculated the cell subset frequency of each immune population relative to the CD45+ populations (Fig. 2A). Overall, all three approaches were concordant, though there is somewhat stronger concordance between scRNA-seq and CITE-seq for all cell types except NK cells (mean difference calculated by Bland-Altman analysis). Cell type abundance is especially consistent for B cells, plasmacytoid dendritic cells (pDC), and neutrophils. Interestingly, the cell frequency decreased and increased for T cells and macrophages/monocytes, respectively, in CyTOF as compared with scRNA-seq and CITE-seq. The mean differences between CyTOF and CITE-seq were -13.6% (95% CI: -24.02% to -3.11%) for T cells and 11.07% (95% CI: 3.19% to 18.95%) for macrophages/monocytes. This finding is consistent with a previous study where fewer T cells were detected in CyTOF compared to scRNA-seq in healthy bone marrow samples. To further investigate which subpopulations were discordant, the frequencies of T cell subsets, monocytes,
and macrophages were evaluated (Fig. 2B, mean difference calculated by Bland-Altman analysis). Interestingly, CITE-seq detected far more CD4+ T cells compared to CyTOF and scRNA-seq, while CyTOF detected far fewer CD8+ T cells compared to the other two techniques. In terms of T cell subtypes, regulatory T (Treg) cell frequency increased and memory CD8+ T cells reduced in scRNA-seq, as compared to CyTOF. In addition, scRNA-seq detected far more macrophages than the other 2 methods, while monocyte frequency was the lowest in CyTOF.

To further evaluate concordance between scRNA-seq and CITE-seq, we examined expression of cell type marker genes, including both the RNA and ADT levels. Average expressions of each marker gene at the transcriptional level (blue dots) between scRNA-seq and CITE-seq are generally concordant (Fig. 2C). By contrast, we observed drastic differences of some marker genes between RNA and ADT expression in CITE-seq, probably due to the RNA dropout and shorter half-lives of mRNAs versus proteins. For example, expression of CD4_adt is higher than that of transcriptional CD4, whereas CD127/IL7R tends to be highly expressed at the transcriptional level. This dynamic explains why IL7R is often differentially expressed in CD4+ T cell population, while CD4 is weakly expressed in scRNA-seq. Taken together, these observations highlight the importance of choosing cell type marker genes best suited to particular modalities.

We also correlated expressions of marker genes among scRNA-seq, CyTOF, and CITE-seq. The vast majority are positively correlated in protein-protein comparison (Fig. 2D) and RNA-RNA comparison (Fig. 2E). Next, we investigated the correlations of expressions of marker genes between the transcriptome and protein levels (Fig. 2F, G; Supplementary Fig. S2A, B). As expected, the overall correlation between different modalities is lower than that of the same
modalities. We observed significant correlation for some markers, including *CCR7* in CD4+ naive T cells, *IL7R* in CD4+ memory T cells, and *FCGR3A* in NK cells, between RNA and protein level of CITE-seq, while no markers are significantly correlated between scRNA-seq and CyTOF (Fig. 2G). We also found that *FCGR3A* in macrophages has a strong correlation, while some markers are significantly anti-correlated between CITE-seq transcriptional level and CyTOF, such as *CD3D, CD3G, IL7R, CD8A*, etc. (Supplementary Fig. S2A, C).

**2.3.5 Decreased ratio of CD4+/CD8+ T cells from ISS stage 2 to ISS stage 3 patients and fast progression-related gene signatures**

Further, we sought to investigate the relationship between clinical features and immune cell composition of MM patients by examining the ratio of CD4+/CD8+ T cells of patients at different disease stages. A previous study used flow cytometry to reveal that this ratio was significantly lower in peripheral blood mononuclear cells (PBMC) of MM patients as compared to that of normal controls and the ratio decreased with the MM progression. By integrating 3 assays, we found the ratio tends to decrease from ISS stage 2 to ISS stage 3 patients (Fig. 3A). Further, CITE-seq and CyTOF analyses revealed significant downregulation of CD45RA in stage 3 patients, suggesting that CD8+ T cells tend to be activated rather than naive in stage 3 patients (Fig. 3B). In addition, we then identified several DEGs of NK cells from FPs relative to NPs, including *ARPC5, XAF1, RAC2* and *PSMB9*, as revealed by both scRNA-seq and CITE-seq assays (Fig. 3C). *ARPC5*, Actin-Related Protein 2/3 Complex Subunit 5, has been revealed to be highly expressed in patients with poor overall survival and could be treated as an independent biomarker for patients with MM, consistent with our observations. A previous microarray-based study found that *RAC2*, Rac Family Small GTPase 2, is significantly upregulated in MM as compared to MGUS. One
subunit of the proteasome (PSMB9), was remarkably highly expressed in cell groups with t(4;14) translocations versus cells from MGUS\textsuperscript{78}. In summary, previous studies indicated \textit{RAC2} and \textit{PSMB9} are associated with disease development from MGUS to MM and our analysis suggested that they might also be related to MM progression. Taken together, we observed the ratio of CD4\textsuperscript{+} T/CD8\textsuperscript{+} T cells decreased in stage 3 patients relative to stage 2 patients, suggesting an increased population of CD8\textsuperscript{+} T cells in bone marrow microenvironment (BMME) of patients in stage 3. We also found that \textit{RAC2} and \textit{PSMB9} are upregulated in NK cells in FPs relative to NPs at transcriptional level, which could potentially serve as MM progression markers.

\textbf{2.4 Discussion}

Single-cell sequencing technologies have been widely used in studying tissue heterogeneity, tumorigenesis and metastasis given their advantages of being able to depict genome, transcriptome, proteome, and other multi-omics profiles of single cells\textsuperscript{79}. However, the similarities of measurements across the various single cell techniques remains to be fully elucidated. Herein, we integrated scRNA-seq, CyTOF, and CITE-seq to perform a detailed comparison of their measurements for MM BM microenvironment. From CD138-negative BM aliquots of 20 samples from 18 patients, we detected, on average, 1,051 immune cells/sample using scRNA-seq, >64K CD45\textsuperscript{+} cells/sample using CyTOF, and 718 immune cells/sample using CITE-seq. By clustering cells with or without protein profiles in CITE-seq, we showed the advantages of multimodal measurement over transcriptional measurement alone of cell type markers when characterizing T cell subtypes in MM (Fig. 1E, H). This observation is in line with a study to investigate renal T cell subtypes by CITE-seq\textsuperscript{72}. 
Next, to examine the consistency of cell populations measured by the same techniques at different sites, we evaluated the cell subset abundances captured by three centers using four samples. Cross-center comparisons (Supplementary Fig. S1F, G) suggested no strong batch effect across centers and there are some important factors to consider in order to obtain reproducible and reliable results: 1) It is important to include reference samples in CyTOF to help identify marker non-specific staining artifacts; 2) Barcoding samples, sample delivery mechanism, and using lyophilized panels is important in CyTOF experiments. Further, cross-technique comparisons revealed that the percentages of immune populations measured by scRNA-seq, CyTOF, and CITE-seq are generally concordant, except some variations in T cells, macrophages, and monocytes (Fig. 2A, B). Analysis revealed relatively high correlations of most markers between the same modalities, though some markers are negatively correlated. (Fig. 2C-G). This observation highlighted the importance of choosing marker genes best suited to particular modalities.

Previous studies have found MM patients have lower CD4+ T/CD8+ T ratios relative to healthy donors and these ratios are further decreased in ISS stage 3 versus ISS stage 1 patients. Here, we confirmed this trend using 3 single cell technologies, finding that this ratio tends to decrease even in stage 3 versus stage 2 patients (Fig. 3A). We also observed the decreased ratio in stage 2 compared to stage 1 patients based on CyTOF and CITE-seq measurement but not in scRNA-seq, probably due to the limited number of patients in stage 1. Future study could further investigate how immune cell composition changes along with ISS stages with expanded sample size. In addition, we observed upregulation of ARPC5, XAF1, RAC2, and PSMB9 in NK cells of FPs compared to those of NPs, as suggested by both scRNA-seq and CITE-seq RNA measurements (Fig. 3C). RAC2 and PSMB9 have been revealed to be associated with disease development from...
MGUS to MM\textsuperscript{77,78} and our analysis suggested that they might also be related to MM rapid progression, supported by both scRNA-seq and CITE-seq. Due to the limited number of protein markers in CITE-seq, we were unable to evaluate the protein-level expression of these MM progression-related genes identified from RNA measurement, which requires further validation. It would also be interesting to investigate MM progression-related markers after controlling for treatments in future studies.

This analysis is just a small sampling of the larger work being conducted by the MMRF and their associated academic research centers to provide a sufficiently broad, deep, and technologically diverse vast dataset for accurately characterizing BMME and to help interrogate MM tumor microenvironment (TME) using different single-cell technologies. We hope this study will help researchers refine cell population characterization strategies and provide insights to those considering integrating multiple single-cell methods to comprehensively address biological questions.

2.5 Materials and Methods

Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the MMRF research committee. These samples provided by MMRF were all from the MMRF’s CoMMpass clinical trial (NCT NCT01454297). Written informed patient consent was obtained from all patients for the collection and analysis of their samples by the MMRF. The CoMMpass study was conducted in accordance with recognized ethical guidelines in
the US and EU. The Institutional Review Board at each participating center approved the study protocol.

**Ammonium-chloride-potassium (ACK) Lysis of Bone Marrow Aspirates (BMA)**

BMA samples obtained from subjects enrolled in the MMRF CoMMpass study (NCT01454297). Any blood clots were removed from BMA samples via passage through 70 mM cell strainer. BMA samples were aliquoted into 5mL aliquots in 50mL conical tubes and 45mL of 22mM-filtered ACK lysing buffer (155mM Ammonium Chloride/10mM Potassium Bicarbonate/0.1mM EDTA/pH7.4) was added to each 5 mL aliquot and the tube gently inverted several times to mix. Tubes were then centrifuged at 400xg for 5 minutes. The supernatant was removed and the cell pellet resuspended with 5 mL of RPMI-1640 and transferred to a clean tube. All aliquots of ACK-lysed BMA aliquots were combined into 1x 50 mL tube, the volume adjusted to 50 mL with RPMI-1640. The cells were then mixed by gentle inversion and the tube centrifuged at 400xg for 5 minutes. The supernatant was then removed by aspiration. Depending on the size of the BMA cell pellet, the cell pellet resuspended in 1-10mL of EasySep buffer (Phosphate-buffered saline (PBS) containing 2%FBS (v/v) and 1mM EDTA (PBS/FCS/EDTA buffer). 25mL of cell suspension was removed for cell counting.

**Isolation of CD138-positive and CD138-negative cells from BMA**

CD138-negative immune cell mononuclear (CD138-) cells in bone marrow aspirates from subjects enrolled in the MMRF CoMMpass study (NCT NCT01454297) were isolated via negative selection from CD138-positive (CD138+) myeloma cells using the EasySepTM immunomagnetic bead technology (EasySepTM Human CD138 Positive Selection Kit: Stem Cell Technologies) in
accordance with the manufacturers protocol. Briefly, 100x10^6 cell/mL bone marrow mononuclear cells (MNC) in a sterile 17x100mm (14mL) tube were gently mixed and incubated with 100mL/mL CD138 selection antibody cocktail for 15 minutes at room temperature. 50mL/mL of EasySep magnetic nanoparticles was then added to the cell suspension, gently mixed, and incubated for a further 10 minutes at room temperature. The volume of the cell suspension was then adjusted to 8mL with phosphate-buffered saline (PBS) containing 2%FBS (v/v) and 1mM EDTA (PBS/FCS/EDTA buffer) and the cell suspension mixed by gentle pipetting (2-3x). The tube was then placed in the magnetic separator. After 5 minutes incubation at room temperature, the magnet and tube were carefully inverted to pour off the supernatant into a sterile 50mL conical tube. This supernatant contains the heterogeneous CD138-negative immune cell mononuclear population (MNC). The tube was then removed from the magnet and an additional 8mL of PBS/FCS/EDTA added, gently mixed, and returned to the magnetic separator. Again, after 5 minutes incubation in the magnetic separator, the tube and magnet were carefully inverted to pour of the supernatant into the 50mL collection tube. This PBS/FCS/EDTA ‘wash’ step was repeated once more resulting in ~24mL suspension of CD138-negative bone marrow MNC cells. CD138- MNC cells were then pelleted by centrifugation at 400xg for 5 minutes and the supernatant removed by aspiration. The CD138-MNC cell pellet was resuspended in freezing medium (90%FCS/10%DMSO) at a concentration of ~8-10x10^6 cells/mL prior to cryogenic storage in liquid nitrogen.

Processing of BMMC and library prep from MMRF CoMMpass study for scRNA-seq at WUSTL

WUSTL Cell Thawing: Multiple Myeloma bone marrow mononuclear cells (BMMC) aliquots were thawed in 37°C water bath. Cells were then pelleted by centrifugation at 300g for 5 min and
all supernatant was removed. To prepare cells for the Miltenyi Dead Cell Removal Kit, cells were resuspended in 100 µL of beads and incubated at room temperature for 15 minutes. Dead cells were depleted using the autoMACS®Pro Separator. Live cells were pelleted by centrifugation at 450g for 5 minutes. Cells were finally resuspended in ice cold phosphate buffer saline (PBS) and 0.5% BSA and loaded onto the 10x Genomics Chromium Controller and using the Chromium Next GEM Single Cell 3’ GEM, Library & Gel Bead Kit v3.3. Utilizing the 10x Genomics Chromium Single Cell 3’v3 Library Kit and Chromium instrument, approximately 16,500 to 20,000 cells were partitioned into nanoliter droplets to achieve single cell resolution for a maximum of 10,000 individual cells per sample. The resulting cDNA was tagged with a common 16nt cell barcode and 10nt Unique Molecular Identifier during the RT reaction. Full length cDNA from poly-A mRNA transcripts was enzymatically fragmented and size selected to optimize the cDNA amplicon size (approximately 400 bp) for library construction (10x Genomics). The concentration of the 10x single cell library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the HiSeq 4000 or NovaSeq 6000 platform (Illumina). 26x98bp (3'v2 libraries) sequence data were generated targeting between 25K-50K read pairs/cell, which provided digital gene expression profiles for each individual cell.

**ISMMS BMMC processing differences from WUSTL:** BMMC aliquots were partially thawed in 37°C water bath. 1mL of warm thawing media (RPMI + 10% FBS) was added to the partially thawed BMMC aliquot and the entire volume was transferred to a 15 mL conical tube containing 10 mL of warm thawing media. The empty BMMC tube was rinsed with another 1 mL of thawing media which was then also transferred to the 15 mL conical tube. Cells were processed using the EasySep Dead Cell Removal (Annexin V) Kit (StemCell Technologies, Cat# 17899).
**scRNA-seq data quantification preprocessing**

For single cell RNA-seq analysis, the proprietary software tool Cell Ranger v3.0.0 from 10x Genomics was used for de-multiplexing sequence data into FASTQ files, aligning reads to the human genome (GRCh38), and generating gene-by-cell UMI count matrix.

Seurat v3.0.0\(^{63, 64}\) was used for all subsequent analysis. First, a series of quality filters was applied to the data to remove those barcodes which fell into any one of these categories recommended by Seurat: too few total transcript counts (< 300); possible debris with too few genes expressed (< 200) and too few UMIs (< 1,000); possible more than one cell with too many genes expressed (> 50,000) and too many UMIs (> 10,000); possible dead cell or a sign of cellular stress and apoptosis with too high proportion of mitochondrial gene expression over the total transcript counts (> 20%). Finally, predicted doublets were also removed using scrublet V0.2.3.

We constructed a Seurat object using the unfiltered feature-barcode matrix for each sample. Each sample was scaled and normalized using Seurat’s ‘SCTransform’ function to correct for batch effects (with parameters: \text{vars.to.regress} = \text{c("nCount\_RNA", "percent.mito")}, \text{return.only.var.genes} = \text{F}).

**scRNA-seq cell type annotation**

Cell types were assigned to each cluster by manually reviewing the expression of marker genes. The marker genes for main cell types were $CD79A$, $CD79B$, $MS4A1$ (B cells); $CD8A$, $CD8B$, $CD7$, $CD3E$ (CD8+ T cells); $CD4$, $IL7R$, $CD7$, $CD3E$ (CD4+ T cells); $NKG7$, $GNLY$, $KLRD1$, $NCAM1$ (NK cells); $MZB1$, $SDC1$, $IGHG1$ (Plasma cells); $CLEC4C$, $IL3RA$, $IRF8$, $GZMB$ (Dendritic cells);
FCGR3A (Macrophages); CD14, LYZ, S100A8, S100A9 (Monocytes); AZU1, ELANE, MPO (Neutrophils); COL1A1, COL3A1, TNC, S100A4 (Fibroblasts); and AHSP1, HBA, HBB (Erythrocytes). All cells that were labeled as erythrocytes and plasma cells were removed from subsequent analysis.

**Processing of BMMC from MMRF CoMMpass study for CITE-seq**

Samples were thawed in the water bath at 37°C for 2-3 min and the cell concentration, viability were determined using a Bio-Rad T20 Cell Counter (Cat# 145-0102). Samples were blocked by incubation with TruStain fcX (Biolegend, cat# 422301) in a 50 µL cell labeling buffer. Next, samples were labeled with Total-seq antibodies (Biolegend) for 30 minutes. Cells were washed and resuspended to obtain a cell concentration of 700-1,200 cells/µl and gently pipette mix using a regular-bore pipette tip until a single cell suspension is achieved. We then proceed immediately to Single cell Gene Expression Library (3’GEX) construction using 10X Chromium Single Cell 3' Reagent Kits v3 (Cat# 1000075) and Chromium i7 Sample Index Plate with Barcoding technology for Cell Surface Protein. For each sample, 5000 cells were injected for CITE-Seq. The libraries were sequenced on NovaSeq S4 platform in pair end sequencing and a single index with at least 50,000 read pairs per cell.

**CITE-seq data quantification preprocessing**

We used Cell Ranger to demultiplex, map to the human reference genome (grch38), and count UMIs in the mRNA libraries, and CITE-seq-Count to count UMIs in the ADT libraries. We filtered out cells with more than 10% UMIs from mitochondrially-encoded genes or less than 1,200 mRNA UMIs in total. We then constructed a Seurat object using the feature-barcode matrix for each
sample (Seurat v3.0.0). Each sample was scaled and normalized using Seurat’s ‘SCTransform’ function to correct for batch effects (with parameters: vars.to.regress = c("nCount_RNA", "percent.mito"), return.only.var.genes = F). Next, the protein expression levels were added to the Seurat object, followed by normalization and scaling for ADT assay.

**CITE-seq data multi-modal integration and cell type annotation**

Using Citefuse v1.2.0, expression was normalized by function normaliseExprs(sce, altExp_name = "ADT", transform = "log"). We then integrated RNA and ADT matrix by an integration algorithm called similarity network fusion (SNF) and clustered cells by louvain clustering. Then, cell types were assigned to each cluster by manually reviewing the expression of marker genes at RNA levels (same as scRNA-seq) and ADT levels (if available). All cells that were labeled as erythrocytes and plasma cells were removed from subsequent analysis.

**Processing of BMMC from MMRF CoMMpass study for CyTOF at ISMMS**

BMMC aliquots were thawed in a 37°C water bath and immediately transferred into RPMI+10% FBS. Cells were pelleted by centrifugation at 300g for 5 minutes and all supernatant was removed. Cells were then incubated for 20 minutes in a 37°C water bath with Cell-ID Rh103 Intercalator (Fluidigm, Cat# 201103A) to label non-viable cells. Samples were then blocked with Fc receptor blocking solution (Biolegend, Cat# 422302) and stained with a cocktail of surface antibodies for 30 minutes on ice. All antibodies were either conjugated in-house using Fluidigm's ×8 polymer conjugation kits or purchased commercially from Fluidigm. Next, samples were fixed and barcoded using Fluidigm’s 20-Plex Pd barcoding kit (Cat# 201060) and pooled into a single tube. The pooled sample was then fixed and permeabilated using BD's Cytofix/Cytoperm
Fixation/Permeabilization Kit (Cat# 554714), blocked with heparin at a concentration of 100U/ml to prevent non-specific staining of eosinophils and stained with a cocktail of intracellular antibodies. Finally, the sample was re-fixed with freshly diluted 2.4% formaldehyde in PBS containing 0.02% saponin and Cell-ID Intercalator-Ir (Fluidigm, Cat# 201192A) to label nucleated cells. The sample was then stored as a pellet in PBS until acquisition.

Immediately prior to acquisition, the pooled sample was washed with Cell Staining Buffer and Cell Acquisition Solution (Fluidigm, Cat# 201240) and resuspended in Cell Acquisition Solution at a concentration of 1 million cells per ml containing a 1:20 dilution of EQ normalization beads (Fluidigm, Cat# 201078). The sample was acquired on the Fluidigm Helios mass cytometer using the wide bore injector configuration at an acquisition speed of < 400 cells per second.

**Processing of BMMC from MMRF CoMMpass study for CyTOF at Mayo**

BMMC aliquots were thawed in a 37°C water bath and immediately transferred into 15ml tubes and slowly diluted with 10 mL of pre warmed RPMI+10% FBS+25U/ml Benzonase(Sigma-Aldrich; Catalog Number-E1014-5KU; 250U/mL). Cells were pelleted by centrifugation (all spins at 500g for 5 minutes) and supernatant was removed. Cells were then incubated for 1hr in a 37°C water bath in 10mL of RPMI+10% FBS. Cells were counted and 3-4 million cells were aliquoted into microfuge 2mL conical tubes, pelleted and washed 2X with 2mL CSB Maxpar® Cell Staining Buffer (Fluidigm; Catalog number-201068; 500 mL) and resuspended in 300uL of Cell-ID™ Cisplatin (Fluidigm; Catalog Number: 201064) 5 min/RT, to label dead cells. Immediately quenched with 1.5 mL CSB, pelleted, and washed with CSB 2X.

For staining, the cell pellet was gently resuspended in 50uL CSB and the addition of an equal volume of diluted surface antibody cocktail, for a final staining volume of 100uL. The staining
reaction was incubated on a rocker platform for 45 min at RT. 1 mL of CSB was used to wash and pellet the cells 2X. Cell pellet was resuspended in the residual volume and then gently resuspended in 500 μL of 1X PBS. An equal volume of 4% PFA in PBS was added to fix cells for a minimum of 20 minutes at a final concentration of 2% PFA in PBS. The sample was labeled overnight at 4°C on a rocker platform with Cell-ID Intercalator-Ir (Fluidigm, Cat# 201192A) in Maxpar Fix and Perm Buffer (Fluidigm; Catalog Number-201067; 100 mL) to label nucleated cells.

The following day the sample was washed 1X with CSB (all cell pelleting performed at 800g for 5 minutes after fixation) and twice with Cell Acquisition Solution (Fluidigm, Cat# 201240). Final resuspension was in Cell Acquisition Solution at a concentration of 0.7 million cells per mL containing a 1:10 dilution of EQ normalization beads (Fluidigm, Cat# 201078). The sample was acquired on the Fluidigm Helios mass cytometer using the wide bore injector configuration at a targeted acquisition speed of 300 events per second. A cryopreserved specimen of 3-4 million Ficoll enriched PBMC derived from a pool of 4 anonymous platelet donors was included with every batch of MMRF samples. This sample was treated and analyzed in parallel throughout the entire experiment as a process control.

**Processing of BMMC from MMRF CoMMpass study for CyTOF at Emory**

BMMC aliquots were thawed in a 37°C water bath and immediately transferred into RPMI+10% FBS. Cells were pelleted by centrifugation at 300g for 5 minutes and all supernatant was removed. Cells were then incubated for 20 minutes in a 37°C incubator. Cells were pelleted by centrifugation at 300g for 5 minutes and all supernatant was removed. Cells were resuspended in PBS and incubated with cisplatin for 1 min (Fluidigm, Cat# 201195) to label non-viable cells. Samples were
washed with Maxpar cell staining buffer (Fluidigm, Cat# 201068) and stained with a cocktail of surface antibodies for 15 minutes at room temperature. All antibodies were either conjugated in-house using Fluidigm's X8 polymer conjugation kits or purchased commercially from Fluidigm. Next, samples were washed and fixed and perm'd with TF Fix/Perm and Perm/Wash Kit (BD Pharmigen, Cat# 51-9008100 and # 51-9008102) using manufacturer’s recommendations. Permeabilized samples were incubated for 30 min in Perm/Wash with a cocktail of intracellular antibodies. After washing and centrifugation at 800g for 5 minutes, the sample was re-fixed with Maxpar Fix I buffer (Fluidigm, Cat# 201065) and Cell-ID Intercalator-Ir (Fluidigm, Cat# 201192A) to label nucleated cells. The sample was then stored as a pellet in PBS until acquisition. Immediately prior to acquisition, the sample was washed with Cell Staining Buffer and Maxpar Water (Fluidigm, Cat# 201069) and resuspended in Maxpar Water at a concentration of 1 million cells per ml containing a 1:10 dilution of EQ normalization beads (Fluidigm, Cat# 201078). The sample was acquired on the Fluidigm Helios mass cytometer using the HT injector configuration at an acquisition speed of < 500 cells per second.

**CyTOF data preprocessing**

The resulting FCS files were normalized and concatenated using Fluidigm's CyTOF software and then de-multiplexed using the Zunder lab single-cell debarocder (https://github.com/zunderlab/single-cell-debarocder). The FCS files were further cleaned on Cytobank by removing EQ beads, low DNA debris, and gaussian multiplets. Barcoding multiplets were also removed based on the Mahalanobis distance and barcode separation distance parameters provided by the Zunder lab debarocoder.
CyTOF cell type annotation and expression normalization

Gating and data analysis were done using WUSTL Cytobank. Live, single cells are selected by gating out cells/debris with outlier cisplatin and DNA intercalator staining. Cell populations were determined based on gating of cell type marker expression. **ISMMS**: CD3+CD19-CD56-CD33- (T cells); CD3-CD19-CD56-CD33-CD123+HLA_DR+CD11c+ (pDC); CD3-CD19+CD56-CD33- (B cells); CD56+CD3-CD19-CD33- (NK cells); CD33+CD3-CD19-CD14+ (Monocytes); CD33+CD3-CD19-CD14-CD16+ (Macrophages). **Mayo**: CD3+CD19- (T cells); CD3-CD19+CD56- (B cells); CD56+CD3-CD16+HLA_DRC/CD56+CD3-CD16-CD123-CD11c- (NK cells); CD3-CD19-CD20-CD14+ (Monocytes); CD3-CD19-CD20-CD14-CD16+ (Macrophages); CD3-CD19-CD20-CD123+ (pDC); **Emory**: CD3+CD19- (T cells); CD3-CD19+ (B cells); CD3-CD19-CD14+ (Monocytes); CD3-CD19-CD14-CD16+ (Macrophages). For T cell subtypes, ISMMS and Mayo used the same gating strategy: CD4+CD8- (CD4+ T cells); CD8+CD4- (CD8+ T cells); CD4+CD8-CD25+CD127- (Treg); CD45RA+CCR7+ (Naive T cells); CD45RA+CCR7- (EMRA T cells); CD45RA-CCR7+ (Central memory T cells), CD45RA-CCR7- (Effector memory T cells), Emory: CD4+CD8- (CD4+ T cells); CD8+CD4- (CD8+ T cells); CD45RO-CCR7+ (Naive T cells). Next, we performed t-SNE analysis for 18 samples from ISMMS. We used the scaled expression of markers, including CD57, CD11c, Ki67, CD19, CD45RA, KLRG1, CD4, CD8, ICOS, CD16, CD127, CD1c, CD123, CD66b, TIGIT, TIM3, CD27, PD-L1, CD33, CD14, CD56, NKG2A, CD5, CD45RO, NKG2D, CD25, CCR7, CD3, Tbet, CD38, CD39, CD28, DNAM1, HLA-DR, PD-1, GranzymeB, CD11b. For expression normalization in CyTOF analysis, we followed instructions from Cytobank and used transformed ratios itself compared to its control, which is the table’s minimum of median of channel (described here

**Bland-Altman analysis**

R package Blandr (v0.5.3) was used to calculate mean difference and 95% confidence interval in Bland-Altman analyses\(^1\). Parameter sig.level = 0.95.

**Differential expression analysis**

Differential expression analysis was performed using the default test (Wilcoxon Rank Sum test) of function FindMarkers (from the Seurat package) with the specified parameters: min.pct=0.25, logfc.threshold = 0.25, and only.pos = T.
2.6 Figures
Figure 1. Overview of cell populations of 18 multiple myeloma patient samples subject to scRNA-seq, CyTOF, and CITE-seq.

A, Patient characteristics and single-cell data collection. FP and NP denote fast progressors and nonprogressors, respectively. ISS = International Staging System. ASCT = Autologous Stem Cell Transplantation. B, UMAP projection of integrated scRNA-seq data, with cells colored by immune cell types. C, t-SNE projection of integrated CyTOF data, with cells colored by immune cell types. D, UMAP projection of integrated CITE-seq data, with cells clustered by integrated RNA and ADT expression, colored by immune cell types. E, UMAP projection of integrated CITE-seq data, with cells clustered by transcriptional level alone, colored by immune cell identities from D. F, Comparison of canonical cell type marker gene expressions between protein level (ADT, top) and transcriptional level (RNA, bottom). Cells are colored by normalized expression. G, Concordance of sample-level average expressions of CITE-seq protein markers measured at RNA level and ADT level. The gray shaded area represents the 95% confidence interval around the line of best fit. R = Pearson correlation coefficient. H, UMAP projection of CD4+ T cells and naïve CD8+ T cells, which is the subset of integrated data in E, with cells clustered by transcriptional level alone, colored by immune cell identities from D and E.
Figure 2. Comparison of cell subset frequencies and correlations of expression of canonical cell type markers across different modalities.

A, Main immune cell population (CD45+) frequencies observed by CITE-seq, CyTOF, and scRNA-seq. Each boxplot is colored by assay. CITE-seq populations are determined on the basis of integrated RNA and ADT expressions. B, Immune cell subtype frequencies for CITE-seq, CyTOF, and scRNA-seq. Each boxplot is colored by assay. CITE-seq populations are determined on the basis of integrated RNA and ADT expressions. C, Concordance of sample-level average expressions of canonical cell type markers in main cell subsets between scRNA-seq and CITE-seq. CITE-seq RNA and protein (ADT) level expressions are represented by blue and red dots, respectively. D, Spearman correlation coefficients of protein level expressions of cell type markers between CyTOF and CITE-seq. Each dot represents a marker gene and the color of the dot represents the $P$ value of correlation. Markers are highlighted with an outer circle if the $P$ value is less than 0.05. E, Spearman correlation coefficients of transcriptional level expressions of cell type markers between scRNA-seq and CITE-seq. Each dot represents a marker gene and the color of the dot represents the $p$ value of correlation. Markers are highlighted with an outer circle if the $P$ value is less than 0.05. F, Spearman correlation coefficients of cell type markers between transcriptional level and protein level expressions in CITE-seq. Each dot represents a marker gene and the color of the dot represents the $P$ value of correlation. Markers are highlighted with an outer circle if the $P$ value is less than 0.05. G, Spearman correlation coefficients of cell type markers between transcriptional level expressions from scRNA-seq and protein level expressions from CyTOF. Each dot represents a marker gene and the color of the dot represents the $P$ value of correlation. Markers are highlighted with an outer circle if the $P$ value is less than 0.05.
Figure 3. Ratio of CD4+ T/CD8+ T of patients in different ISS stages and markers associated with ISS disease stages and multiple myeloma progression.

A, Violin plots showing the ratio of CD4+ T/CD8+ T of patients in ISS stage 2 and 3 in scRNA-seq, CyTOF, and CITE-seq. Horizontal lines indicate the median of data points in each group. B, Violin plots showing single cell-level normalized expression of CD45RA in CITE-seq ADT measurement and CyTOF. The difference is significant at $P \leq 0.0001$ based on Wilcoxon rank–sum test. C, Heatmaps showing DEGs of NK cells of FP versus NP patients in CITE-seq RNA measurement (left) and scRNA-seq measurement (right). The samples are ordered on the basis of hierarchical clustering of expression profiles of these genes in CITE-seq RNA.
measurement. Expression values are scaled such that for each gene, the average of the scaled expression is 0 and the SD is 1. Adjusted $P$ values and log fold change in CITE-seq and scRNA-seq were shown on the left and right side of DEGs, respectively. FC = fold change.
**Fig. S1. Expression of cell type markers in CITE-seq and comparison of cell subset abundance between technical replicates and across different centers.**

**A,** Gene and protein expressions of canonical cell type markers for CITE-seq (related to Fig. 1F). Cells are colored by normalized expression. **B,** Reproducibility of technical replicates for scRNA-seq in terms of cell subset frequency. Linear regression line is shown in grey. Dots are colored by immune cell types and shaped by samples. R = Pearson correlation coefficient. **C,** Reproducibility of technical replicates for CyTOF. Dots are colored by immune cell types and shaped by samples. R = Pearson correlation coefficient. **D,** Reproducibility of technical replicates for CITE-seq. CITE-seq populations are determined by integrated RNA and ADT expressions. Dots are colored by immune cell types and shaped by samples. R = Pearson correlation coefficient. **E,** Overview of datasets used in cross-center comparisons. * denotes that BIDMC scRNA-seq data is from CITE-seq with cells analyzed using RNA signal alone. **F,** Immune cell population (CD45+) frequencies observed by scRNA-seq in three different centers. Boxplot is colored by center. **G,** Immune cell population (CD45+) frequencies observed by CyTOF in three different centers. Boxplot is colored by center. **H,** A heatmap showing Pearson correlation coefficient of expression profiles of B cell marker genes in B cells between populations detected from different centers using scRNA-seq. B cell markers include $CD79A$, $CD79B$, $MS4A1$ and pre-B cell markers.
Fig. S2. Correlation of expression of canonical cell type markers across different modalities.

A, Spearman correlation coefficients of cell type markers between transcriptional level expression from CITE-seq and protein level expression from CyTOF. Each dot represents a marker gene and the color of the dot represents the p value of correlation. Markers are highlighted with an outer circle if the p value is less than 0.05. B, Spearman correlation coefficients of cell type markers between transcriptional level expression from scRNA-seq and protein level expression from CITE-seq. Each dot represents a marker gene and the color of the dot represents the p value of correlation. Markers are highlighted with an outer circle if the p value is less than 0.05. C, Scatter plots showing
examples of negative correlation of expression of cell type marker genes between transcriptional level expression from CITE-seq and protein level expression from CyTOF. MMRF_1505 and MMRF_2251 were excluded due to lack of cell populations of interest based on CITE-seq measurement.
Chapter 3: Therapeutic targets discovery

Our work has been accepted for publication: Yao et al. Single-Cell Discovery and Multi-Omic Characterization of Therapeutic Targets in Multiple Myeloma. Cancer Res 2023 Feb 13: CAN-22-1769; Contribution: I led the data analysis, results interpretation, figure generation, and manuscript writing as a first author of the paper.

3.1 Summary

Multiple myeloma (MM) is a highly refractory hematological cancer. Targeted immunotherapy has shown promise in MM but remains hindered by the challenge of identifying specific yet broadly representative tumor markers. We analyzed 53 bone marrow (BM) aspirates from 41 MM patients using an unbiased, high-throughput pipeline for therapeutic target discovery via single cell transcriptomic profiling, yielding 38 MM marker genes encoding cell-surface proteins and 15 encoding intracellular proteins. Of these, 20 candidate genes were highlighted that are not yet under clinical study, 11 of which are previously uncharacterized as therapeutic targets. The findings were cross-validated using bulk RNA-sequencing, flow cytometry, and proteomic mass spectrometry of MM cell lines, and patient BM, demonstrating high overall concordance across data types. Independent discovery using bulk RNA-sequencing reiterated top candidates, further affirming the ability of single cell transcriptomics to accurately capture marker expression despite limitations in sample size or sequencing depth. Target dynamics and heterogeneity were further examined using both transcriptomic and immuno-imaging methods. In summary, this study
presents a robust and broadly applicable strategy for identifying tumor markers to better inform the development of targeted cancer therapy.

3.2 Introduction

Multiple myeloma (MM) is an adult hematological malignancy, characterized by the uncontrolled proliferation of bone marrow plasma cells (PCs), that are highly prone to drug-resistant relapse. Targeted immunotherapies\(^8\), which may reduce toxicity towards healthy tissues, have shown promise in treating MM\(^3\). Immunotherapies currently in development include monoclonal antibodies, bi-specific T-cell engagers (BiTEs), antibody drug conjugates (ADC), and adoptive cellular therapies such as CAR-T, CAR-Natural killer (NK), and TCR-T\(^3\), which have demonstrated substantial efficacy in combating liquid tumors, including MM\(^4\),\(^8\),\(^6\).

Despite these advances, major limitations in the development of targeted immunotherapies lie in both the discovery of optimal candidate targets with sufficient tissue specificity and tumor coverage, and in the risk of evolutionary downregulation of target proteins once subjected to engineered immune surveillance\(^3\). Recent single cell studies characterize the inherent subclonal heterogeneity and longitudinal evolution of MM tumors\(^5\),\(^8\), further underlining the need for a nuanced understanding of myeloma tumor epitopes. Unlike bulk strategies that have been used to date as the rationale for immunotherapeutic strategies against MM\(^8\), single-cell approaches minimally obscure tumor architecture and do not rely on selection strategies that may bias resulting tumor profiles. Applications of single-cell RNA-sequencing (scRNA-seq) in diseases have demonstrated the utility of single-cell data in overcoming tumor heterogeneity when optimizing treatment regimens\(^8\). In this study, we build upon these established principles by directly profiling
the MM patient population and tailoring our approach to inform targeted immunotherapy against MM.

To systematically identify myeloma markers of potential therapeutic relevance, we used scRNA-seq to perform an unbiased search for genes with specific expression in plasma cells and/or B cells (relative to other detected cell types) on 53 bone marrow (BM) samples taken from 41 patients. We collectively analyzed 146,725 single-cells from these patients, including 40,177 PCs, 7,050 B cells, and 99,498 other immune cells. We then used publicly available databases to annotate plasma/B-cell specific genes with known tissue expression and the predicted cellular localization of their protein products. After filtering candidates' expressions, their prevalences across our cohort, and tissue specificities, we examined their expression correlations and co-expression patterns across patients. Consistent with previous studies, we found PCs from each patient to be transcriptionally distinct, with unique sets of highly-expressed genes driving patient-specific cluster resolution in ways not seen in other BM cell types. Additionally, we observed that certain targets were expressed dynamically during disease development within the same patient and that the preferential expression of certain targets may be related to disease progression. Finally, we cross-validated the expression of candidate targets using bulk RNA-seq and protein quantification. This study represents the first effort using high-throughput scRNA-seq for the systematic discovery of myeloma-associated antigens with the potential for therapeutic applications, paving the way for expanding targeted immunotherapy options in MM.
3.3 Results

3.3.1 Identification of myeloma targets using scRNA-seq-driven strategies with three independent datasets

We developed a single-cell RNA-sequencing based strategy for identifying MM markers (Fig. 1A) using samples collected from 3 independent studies: the Immune Atlas Pilot study led by the Multiple Myeloma Research Foundation (MMRF) and 2 internal studies at Washington University (WashU Cohort 1 and WashU Cohort 2). In total, 53 bone marrow samples were analyzed from 41 patients (patient 83942 was included in both WashU cohort 1 and WashU cohort 2) using 3’ scRNA-seq (Methods). Of these, 18 samples belonging to the MMRF Immune Atlas Pilot were CD138-depleted fractions taken at diagnosis from patients in the MMRF CoMMpass study (CD138+ PC fractions having been used for earlier bulk sequencing studies by the same name). Despite prior sorting, PCs were still detectable in remaining fractions. The remaining 35 samples were sourced from internal studies and represent 23 individuals with varied disease stage and treatment history. The proportions of captured PC and immune cell types varied across patients (Fig. 1B), with PC proportions ranging from 0.5% to 94.3% and B cell proportions ranging from 0.03% to 29.6%. scRNA-seq of 8 BM samples taken from healthy donors were included as a control group (Fig. 1B).

To identify MM markers, we first defined PCs, which make up the myeloma neoplasm, and B-cells, which developmentally precede PCs and may harbor premalignant subsets, as our lineages of interest. We began by comparing patient PCs against all other cells detected in each scRNA
sample. From this analysis, we identified 142 genes with a positive fold-change in PCs of at least 90% of samples and reaching significance in at least 75% of samples (adjusted p<0.05) (**Fig. 1C**). We then compared combined PCs and B cells against other lineages (including T, NK, and myeloid cells), identifying 120 genes that are overexpressed by this compound group (subject to the same prevalence and significance criteria). We then annotated these genes and their encoded proteins with predicted cellular localization and known tissue specificity using information from Gene Ontology (GO)**, Cell Surface Protein Atlas (CSPA)**, Human Protein Atlas (HPA)**, and Genotype-Tissue Expression (GTEx) (Methods, **Fig. 1A**). Following these first-pass statistical filtering, we undertook extensive manual revision to eliminate tissue-specificity false positives and refine the protein localization assignment (Methods). Lastly, since relative (rather than absolute) expression values were used to identify our genes of interest, we subsequently compared patient PC/B cells to healthy donor PC/B cells, eliminating any genes whose absolute expressions were in fact lower than normal (by 2-sided t-test, FDR <0.05) (**Fig. 1A**). These filtering and annotation steps ultimately led to the retention of the following: 136 genes encoding intracellular or secreted proteins and 38 encoding proteins that localize to the cell surface (**Fig. 1C**); 15 surface-protein-encoding genes specific to PC/B cells but not expressed by tissues outside of the lymphoid system; 23 surface-protein-encoding genes specific to PC/B cells within the bone marrow but expressed in other tissues; and 15 genes encoding intracellular or secreted products that are highly specific to both lymphoid tissues and the PC/B cells within them. Bulk RNA-seq and bulk global proteomics were then utilized to cross-validate the expression of potential targets.

Of the 38 MM associated surface-protein-encoding genes, 32 (84%) were discovered in all three cohorts, suggesting high concordance in differential expression among the three datasets (**Fig. 1D**).
This gene set included the top 20 candidate genes as ranked by magnitude of positive fold-change in the lineages of interest. Importantly, our strategy recapitulated several of the most promising MM therapeutic targets currently under pre-clinical and clinical evaluation, including TNFRSF17, SLAMF7, CD38, FCRL5, GPRC5D, and SDC1, which we employ as benchmarks of accuracy. We also identified less well-known targets (Fig. 1D) that might have potential for CAR targeting.

3.3.2 Candidate targets are enriched in immune signaling and protein-processing related pathways

To understand the biological functions of myeloma-specific genes, we used pathway enrichment analysis, finding significant overrepresentation of cellular processes related to hematopoiesis and protein synthesis, modification, and secretion. In particular, protein processing in the endoplasmic reticulum (ER; \( q = 8.13 \times 10^{-5} \)) and protein export (\( q = 1.81 \times 10^{-2} \)) are highly enriched, which is consistent with current understanding of active antibody production in malignant PCs (Fig. 2A). Enrichment of N-glycan trimming (\( q = 0.03 \)) and glucagon-like peptide-1 (GLP-1) synthesis (\( q = 0.02 \)) may reflect changes in glycoprotein-mediated immune signaling in the myeloma microenvironment, as well as serum glycosylation changes in MM (Fig. 2A). Several genes, including CD19, MS4A1 (CD20), CD22, CD38, CD40, and IL5RA, also enrich for hematopoietic cell lineage (\( q = 2.81 \times 10^{-4} \), Fig. 2B), as they represent canonical B cell markers with important lineage-determining functions. Consistent with observations that B cell receptor signaling is enriched in Monoclonal gammopathy of undetermined significance (MGUS), BCR signaling (\( q = 0.01 \)) is also represented in our findings by high expression of RASGRP3, CD79A/B, CD19, and CD22. Notably, we see evidence that the NF-κB pathway, which has been shown in multiple
studies to play a critical role in the proliferation and drug-resistance of myeloma cells\textsuperscript{99,100}, may be affected in myeloma cells due to enrichment of non-canonical TNF-mediated NF-κB signaling (q =1.81x10$^{-2}$). Overall, we find that many myeloma-associated intracellular markers, including \textit{DERL3, HERPUD1, HSP90B1, SEC11C, MZB1, SPCS2,} and \textit{SSR4}, coordinate protein transport and metabolism. Compared to intracellular markers, which have high expression fold change relative to other cells, genes encoding myeloma-associated surface proteins often have relatively smaller fold differences, and are often related to signal transduction and hematopoiesis. Such members include \textit{CD19, CD22, CD24, CD38, CD40, SDC1, TNFRSF17} (BCMA), \textit{TNFRSF13B} (TACI) and \textit{TNFRSF13C} (Fig. 2B). Gene Ontology (GO) enrichment of myeloma-associated genes likewise shows overrepresentation of protein metabolism, cell proliferation, and molecular signaling processes (Supplementary Fig. S1A).

To assess the functional importance of proposed myeloma markers, we compared reported dependency scores from the DepMap CRISPR-Cas9 gene knockout dataset\textsuperscript{101}. Roughly half of our marker genes were consistently found to have a detrimental loss-of-function effect across 21 plasma cell lines (Supplementary Fig. S1B). Most strikingly, \textit{PIM2} and \textit{POU2AF1}, which encode intracellular proteins, reached dependency scores similar to or exceeding those of pan-essential genes in nearly all plasma cell lines.

In considering how upregulated PC/B functions may be disrupted for anti-tumor treatment, we next examined the utility of targeting chemokine systems to enhance tumor access by CAR-T or other immune cell modalities. Cancers often rely on altered chemokine signaling channels for growth and metastasis, and receptors that modulate bone architecture or immune surveillance play
important roles in MM pathogenesis\textsuperscript{102}. For cell-based immunotherapies, taking advantage of existing communication patterns between tumor and the immune microenvironment may improve tumor penetration. In our cohort, we found that \textit{CCL3}, \textit{CXCL10}, and \textit{CXCL12} are upregulated in PCs in a few patients (Supplementary Fig. S1B). Interestingly, \textit{CXCL12} is strikingly highly expressed in 27522_2, the only remission sample in our datasets. Some chemokine receptors (\textit{CXCR6} and \textit{CCR4}) are found to be specifically expressed in T cells in most patients (Supplementary Fig. S1C). Although our findings are inconclusive regarding the role of chemokines or their receptors in this cohort, this analysis may provide insights toward enhancing tumor homing by taking advantage of chemotaxis gradients in the diseased bone marrow.

### 3.3.3 Characterization of surface proteins’ potentials as antigenic targets of CAR-T cells

Ideal target antigens for CAR-T therapy are plasma membrane-localized proteins with accessible extracellular domains that have both unique and abundant expression in tumor cells of most patients\textsuperscript{103}. We further examined the 38 surface-protein-encoding genes identified from our discovery pipeline to assess expression specificity and magnitude across individual samples. Although certain genes, including \textit{CD79A}, \textit{EDNRB}, \textit{CD40}, and \textit{MS4A1} (CD20), show sample-specific overexpression, most have relatively uniform overexpression across samples (Fig. 3A). As expected, known myeloma markers, including \textit{TNFRSF17} (BCMA), \textit{SDC1} (CD138), \textit{FCRL5}, \textit{CD38}, \textit{SLAMF7} (CS1), and \textit{CD79A}, were identified as top-ranking genes upregulated in PCs relative to other immune populations, with high average expression in PCs as well (Fig. 3A).
To avoid non-specific toxicity, ideal target antigens must have limited expression on critical normal tissues. We stratified surface proteins into primary (specific) and secondary targets (nonspecific or unknown) based on tissue expression as reported in GTEx\textsuperscript{104} and HPA\textsuperscript{105} (Fig. 3B and C). We found further support for TNFRSF17 as being exclusively high in lymphocytes (GTEx; fold change $\sim$100, Fig. 3B). CD79A, a receptor involved in adaptive immunity that is restricted to lymphoid tissues, has the second highest average expression in PCs among primary targets and is overexpressed in the combined PC/Bs of 46/53 samples (Fig. 2B and 3B). However, CD79A is also highly expressed in normal PCs and B cells due to its critical role in B cell development and activation\textsuperscript{106}. FCRL5, a member of immunoglobulin receptor superfamily, ranks third in PC-specific overexpression and, with the exception of spleen, appears exclusively expressed in lymphocytes according to GTEx (Fold Change (FC) $\sim$50) (Fig. 3B and Supplementary Fig. S2). SLAMF7, a primary target already in clinical application\textsuperscript{86,107,108}, ranks fourth; however, its expression in other immune cells, such as Dendritic cells (DC), Natural Killer (NK) cells and T cells, is relatively high (Fig. 3B and Supplementary Fig. S2). Other primary targets, exhibiting lower average expression in PCs than the aforementioned candidates but still meeting our specificity criteria, include CD79B, CD40, MS4A1, LAX1, P2RX5, TNFRSF13C (BAFF-R), HVCN1, FCER2, CD19, and FCRL1. (Fig. 2B and 3B). Of these, CD79B, TNFRSF13C, and FCER2 are strikingly overexpressed in B cells and may be therapeutically relevant for eradicating the B developmental reservoir of MM PCs (Fig. 3B). The preclinical efficacy of CD19/CD20 bi-specific CARs in limiting MM antigen escape\textsuperscript{109}, the success of BAFF-inhibitory drugs\textsuperscript{110}, and the evidence that FCER2, which canonically encodes a B-cell-specific antigen, is strongly associated with chromosomal abnormalities in myeloma\textsuperscript{111}, all indicate B-cell involvement in MM. CD79B,
which encodes part of the B cell receptor (BCR) signaling complex, has been previously studied as a therapeutic target in B-cell lymphoma\textsuperscript{112}, but its role in myeloma is not yet extensively studied. While \textit{LAX1} and \textit{HVCN1} are expressed in multiple bone marrow cell types, they may still be considered as MM targets in their comparability to \textit{SLAMF7}, which is also expressed in non-PC/B immune cells (Fig. 3B and Supplementary Fig. S2). Taken together, \textit{CD79A}, \textit{P2RX5}, \textit{FCRL1}, \textit{LAX1}, and \textit{HVCN1}, for which there are not yet any published pre-clinical studies in the context of MM, merit further investigation.

Secondary targets, which have either unknown protein-level specificity or have been reported in other tissues, remain relevant considerations for dual-targeted CAR-T-cell immunotherapy approaches currently under development\textsuperscript{113}. It must also be noted that database-reported expression in non-lymphoid compartments does not necessarily indicate toxicity, as MM PC expression levels may far exceed those of normal tissues. For example, \textit{SDC1} (CD138) and \textit{CD38} have been extensively investigated as targets for anti-MM monoclonal antibodies\textsuperscript{114,115} despite showing abundant expression in non-hematopoietic tissues in HPA (Fig. 3C). Like the aforementioned \textit{TNFRSF17}, \textit{TNFRSF13B} (TACI) belongs to the tumor necrosis factor (TNF) superfamily and is expressed primarily in lymphocytes (GTEx; FC ~200). In our data, 46/53 samples exhibited a logFC > 1.5 of \textit{TNFRSF17} in PCs versus non-PCs, while \textit{TNFRSF13B} showed similar PC fold elevation in only 14/53 samples. In pre-clinical MM models, dual targeting of BCMA and TACI successfully maintained tumor control in the absence of BCMA due antigen escape\textsuperscript{116}. \textit{GPRC5D}, a member of G protein-coupled receptor family, is another secondary target currently in clinical trials as both an anti-tumor CAR-T target and as a GPRC5D/CD3 bispecific antibody\textsuperscript{117}. In addition to these known targets, our study identified \textit{PLPP5}, \textit{CADMI}, \textit{CAVI},
GPR160, KCNN3, EDNRB, LSR, FCRL2, and several other genes as novel secondary targets with potential therapeutic utility (Fig. 3C and D).

Taking into consideration the fact that protein expression data reported in GTEx and HPA may be incomplete or misleading, we present 20 candidate genes not yet under clinical investigation whose PC and/or B cell-specific overexpression can at least be confirmed in bone marrow populations by our scRNA-seq dataset. Of these genes, 9 are supported by existing publications as myeloma markers of possible clinical interest, and 11 are novel targets whose expression in myeloma have to date been largely uncharacterized. Novel candidates include KCNN3, LSR, PERP, FCRL2, GPR160, and IL5RA. Our candidate genes, highlighted in Fig. 3D, may provide therapeutic opportunities for patients refractory to current therapies. While further investigation is of course necessary to determine protein abundance and off-target toxicity, our analyses greatly narrow the search space in identifying tumor markers. Flow cytometry of an independent cohort including 11 MM patients and one Smoldering multiple myeloma (SMM) patient indicated that protein-level expression of TNFRSF17 (BCMA), FCRL5 (CD307), MS4A1 (CD20), CD19, SDC1 (CD138), and CD38 are concordant with our RNA-level findings and even suggest underestimation of sample prevalence in scRNA-seq (Fig. 3E, F, and Supplementary Fig. S3), further supporting the utility and robustness of our discovery approach.

3.3.4 Potential combinatorial targeting partners revealed by correlation analysis
Simultaneously targeting multiple tumor markers may mitigate antigen escape and enhance efficacy. To evaluate target co-expression, we aggregated samples using an integration method that corrects for batch-effect-induced biases (Methods)\(^\text{118}\). We observed that most immune lineages (eg. T cells, monocytes) from different samples clustered together by cell type, whereas PCs formed 49 clusters with unique patient origin, indicating intrinsic transcriptomic diversity of tumors (Supplementary Fig. S4A). Samples taken from the same individual tended to cluster together, as in the cases of 81012, 58408, and 47491 (Fig. 4A). To reduce noise from potentially irrelevant differences in gene expression, we re-clustered PCs using candidate target expression instead of global expression (Methods). We then observed that most samples cluster together, while cells from nineteen patients remained distinct. This suggests that most patients may in fact share common target profiles, while others will require more personalized targeted therapy (Supplementary Fig. S4B).

Using Pearson correlation analysis for sample-level average expression in PCs, we identified co-expressed (\(r>0.30\)) and mutually exclusive target pairs (\(r<0.05\)) (Fig. 4B). The r-value thresholds were determined based on the distribution of correlation of 2000 random genes, and by visualizing expression of gene pairs in Uniform Manifold Approximation and Projection (UMAP) (Supplementary Fig. S4C; Methods). To account for potential cohort-related biases, we performed this correlation analysis within each cohort, finding that 66% (25/38) of genes have co-expression partners and 97% (37/38) of genes have at least one mutually-exclusive partner in all three cohorts (Fig. 4B, Supplementary Fig. S4D). We observed generally high concordance between the cohort-specific and pooled correlation analyses, meaning that genes with the greatest number of co-
expression or mutual-exclusion partners in the pooled analyses show similar trends in each individual cohort.

While many candidate genes were simultaneously expressed at the sample level, we saw evidence for distinct groups of closely-correlated target genes (Fig. 4B; indicated by triangular outlines). One major group includes SDC1, LAX1, SLC1A4, and KCNN3; another includes HVCI, FCRL5, CD40, and TNFRSF17; and a third includes SLAMF7, CD38, PLPP5, and TNFRSF13B. Consistent with sample-level analysis, SLAMF7 and CD38 expression are largely concordant in UMAP visualization of individual cells (Fig. 4C; “Co-expression” row), as are TNFRSF13B, CD38, and PLPP5. GPRC5D, MS4A1 and CD79A are co-expressed as well (Supplementary Fig. S5A). Conversely, CD40 and SLAMF7 are very rarely simultaneously expressed by individual tumors (Fig. 4C; “Mutually Exclusive Expression” row). Other mutually-exclusive pairs include FCRL5/GPRC5D, CD79A/CCR10, SLC1A4/KCNN3 and HVCI/CD79B (Fig. 4C, Supplementary Fig. S5A). These findings suggest that certain tumor phenotypes may respond best to targeting against specific markers: for example, FCRL5 may be a more broadly applicable alternative than CS1 (SLAMF7) for patients who also qualify for anti-BCMA (TNFRSF17) therapies but may not be appropriate for tumors expressing GPRC5D. As clinical reagents are in development for several of these targets, our data indicates that further study is warranted to characterize their interchangeability. As yet, however, our interpretation is limited by low expression of some targets and the possibly insufficient coverage or sample size for thoroughly interrogating these relationships in our scRNA cohort.
To address these weaknesses of scale and depth, we next examined co-expression across the much larger CD138+ sorted CoMMpass bulk RNA-seq dataset, wherein we see two distinct groups of highly correlated surface targets (Supplementary Fig. S4E). The first group includes the well-characterized myeloma markers SDC1, SLAMF7, TNFRSF17, and CD38, along with LAX1, CAV1, and SLC1A4. The distinct correlation patterns between these targets, which we see in scRNA-based co-expression analysis, are largely obscured in this dataset. This may be due to the limitations discussed above, but the CD138-selection of tumors for the CoMMpass dataset may also contribute to this discrepancy. The second highly-correlated gene group contains certain known B-cell-associated genes such as MS4A1, CD19, TNFRSF13C, and CD79A, which are also highly expressed in PCs from our scRNA-seq dataset. Interestingly, these two gene groups are negatively correlated with each other, suggesting that most samples preferentially express either one set or the other, and that there may be a distinct “B-cell-like” phenotype among myeloma tumors58. However, we do not see resolution of a distinct B-like co-expression group in our scRNA data, and due to the imperfect isolation of PCs and the aggregate nature of bulk studies, it remains inconclusive whether this apparent mutual exclusivity is actually an artifact of B-cell contamination.

Given potential differences in mRNA and protein half-life, as well as possible dropout error in scRNA-seq, we next assessed co-expression of well-characterized myeloma markers using flow cytometry of 12 unsorted patient samples (Fig. 4D). PCs were identified by CD138+/CD38+ gating. Flow cytometry confirmed overall CS1 (SLAMF7), BCMA (TNFRSF17), FCRL5, CD38, and CD138 co-expression as observed in scRNA-seq (Fig. 4D and E); among these markers, FCRL5 expression was the least prevalent. In our flow cytometry results, CD19 and CD20 (MS4A1)
expression appeared restricted to CD138-/BCMA- cells, which we categorized as non-plasma B cells for lack of better evidence. Whether a CD19+/CD20+ subset of PCs do exist in a subset of patients, as suggested by our scRNA-seq findings, remains to be further investigated.

Finally, we note the observed sub-cluster heterogeneity of marker expression, particularly the single-positive expression of several gene pairs including the canonical MM markers\textsuperscript{119} TNFRSF17 (BCMA) and SDC1 (CD138), SLAMF7 (CS1/CD319) and SDC1, and FCRL5 (CD307e) and SDC1 (Supplementary Fig. S5B). The relative abundance of single-positives varies between individuals. For example, PCs from patient MMY83942 exhibit slight preference for either TNFRSF17 or SDC1 with low-mid expression of both genes, whereas PCs from MMY22933 highly express one or the other. Several PC subpopulations exhibit high SLAMF7 or FCRL5 expression and low or no SDC1 expression (and vice-versa). When comparing RNA expression of SLAMF7 and FCRL5 to TNFRSF17 instead of to SDC1, we again see subpopulations that skew towards one over the other, with relative proportions differing between patients (Supplementary Fig. S5B). The proportion of FCRL5+/SDC1-/TNFRSF17- plasma cells in MMY83942, MMY80649 and 77570 are 45%, 48%, 33% respectively. As CD138 selection is frequently used to isolate MM PCs for study, our results suggest that current MM cell isolation techniques are likely not sufficient to capture all MM clones within a sample and that additional techniques should be considered. To see whether this may be explained by technical drop-out or differential mRNA metabolism irrespective of protein expression, we next performed immunofluorescence (IF) co-staining of these pairs (Supplementary Fig. S5C). The presence of both co-expressing and single positive cells (expressing either CD138 or one of the other three genes) are corroborated by IF in patient s10-10686 (newly diagnosed sample; BCMA vs. CD138, FCRL5 vs. CD138, and SLAMF7
vs. CD138 in Supplementary Fig. S5C). A mounting body of evidence that CD138 may be rapidly shedding from the PC surface\textsuperscript{120,121} renders inconclusive whether BCMA+, CS1+, or FCRL5+ PCs are truly distinct from CD138+ PCs, but an alternative explanation may lie in the stochastic fluctuation of either marker in individual cells, which scRNA-seq and imaging of fixed tissues cannot detect. These findings demonstrate both the potential difficulty in capturing MM tumors with any single marker and the importance of selecting the right set of markers when designing combinatorial strategies.

### 3.3.5 Characterization of intracellular or secreted proteins with potential relevance for MHC-based therapies

As the majority of our detected PC/B-cell-specific genes encode non-surface proteins, we next evaluated these genes for potential utility as therapeutic targets. Cytosolic proteins are presented as peptides on the cell surface by the major histocompatibility complex (MHC), and intracellular tumor markers may therefore be targeted via T-cell receptor (TCR) recognition of specific MHC-bound peptides\textsuperscript{122}. Of the 136 genes encoding non-surface targets highly expressed in PC and/or B cell populations, 11 were specifically upregulated in PCs relative to other BM immune populations in nearly all samples, namely \textit{MZB1, SSR4, DERL3, SEC11C, HSP90B1, FKBP2, SPCS2, HERPUD1, FKBP11, PRDX4 and MYDGF} (Fig. 5A). In contrast to the signaling and developmental pathways enriched by surface protein targets, 6 of these 11 genes (\textit{SSR4, DERL3, SEC11C, HSP90B1, SPCS2, HERPUD1}) are involved in protein transport and metabolism (Fig. 2B). Of note, \textit{MZB1} (marginal zone B and B1 cell specific protein), and \textit{SSR4} (signal sequence receptor subunit delta, an ER-targeted protein) are extremely highly expressed in PCs. These
observations are consistent with previous reports characterizing MZB1 as a candidate marker of MM in a functional proteomics study\textsuperscript{123} and reaffirm the important role of ER stress in MM\textsuperscript{124}.

Considering reported gene expression in non-BM tissues, we retained 15/136 genes showing specific expression in lymphocytes or lymphoid tissues only (Fig. 5B). We then compared expression in patient PC/B cells to those from normal BM, as true tumor antigens have more therapeutic potential than proteins native to healthy tissue in a TCR-based approach. MZB1 meets key criteria for a good target antigen, namely significant high expression in patient PC/Bs and lymphocyte/lymphoid tissue-restricted expression. SEC11C, which encodes another ER-related protein restricted to lymphoid tissues, has the second-highest average normalized expression in PCs among the 15 genes (Fig. 5B, C and D, note the scale). Similarly, PIM2, which encodes a serine/threonine kinase, is differentially expressed in PCs in 83\% of samples and specific to lymphocytes in GTEx (FDR \( \sim 10^{-10} \), Fig. 5B). Studies have shown that PIM2 is upregulated and required for MM proliferation, suggesting the potential clinical efficacy of targeting PIM2 despite this limitation\textsuperscript{125}. POU2AF1 (POU Class 2 Homeobox Associating Factor 1) may be another promising target, as it is highly expressed in PCs in our cohort in 49/53 of samples (Fig. 5C) and has been found to promote MM cell growth by direct transactivation of TNFRSF17\textsuperscript{126}. As with MZB1, the expression levels of POU2AF1, TEX9, BLK, BACH2, and SWAP70 were significantly elevated in patient PCs and/or B cells compared to those from normal BM. Given their specific and abundant expression in PCs, MZB1, SEC11C, PIM2, POU2AF1, and TEX9 are highlighted as priority tissue-specific intracellular targets (Fig. 5D).
We next analyzed candidate target protein sequences using NetMHC\textsuperscript{127} to predict high-affinity peptide binding for the five most frequently occurring MHC class-I HLA-A, -B, and -C alleles in the US population\textsuperscript{128} (Fig. 5E). Predicted binding affinity of 8-11mer peptides were compared to a set of 400,000 random natural peptides to determine strong binding. Overall, 4,921 peptides obtained from these 15 genes (Fig. 5E) were predicted to bind with high affinity (defined as the top 0.5\% of predicted affinity values; see Methods) to the 15 HLA alleles. At least one potentially high affinity binding peptide was obtained from each gene for 14 out of the 15 HLA alleles examined (Fig. 5E). 57 total predicted high affinity peptides from MZB1 cover alleles including A*01:01, A*02:01, A*03:01, A*11:01, and A*24:02, each of which represent significant (10-40\%) fractions of the US population\textsuperscript{129}. With respect to its high prevalence and tumor specificity, MZB1 may be a model candidate for TCR-based targeting. Certain genes (\textit{POU2AF1}, \textit{HLA-DOB}, \textit{TEX9}, and \textit{AIM2}) appear to have distinct preferential binding for a subset of HLA alleles, whereas others (\textit{BACH2}, \textit{SWAP70}, \textit{BLK}, and \textit{BANK1}) are predicted to have high affinity for several common HLA alleles. As MHC presentation is required for TCR-based adoptive cell therapy, HLA compatibility is an important factor in choosing targets. However, high native immunogenicity may present a potential complication as tumors may be under selective pressure to downregulate highly expressed and highly presented proteins\textsuperscript{130}. Further study is necessary to fully evaluate the utility of our proposed candidates in a TCR-based approach in MM.

### 3.3.6 Expression of candidate targets in relation to clinical characteristics

To better understand how target persistence may be affected by cycles of treatment and relapse, we next compared PCs sampled longitudinally from 6 individuals (Fig. 6A). Notably, certain genes
appear more stably expressed across timepoints than others. For example, *TNFRSF17, CD79A, MZB1*, and *SEC11C* expression fluctuate minimally, whereas *SLAMF7, LAX1, PLPP5*, and *PIM2* show evidence of increased expression in later timepoints. In patient 27522, *PIM2* is significantly downregulated at remission (27522_2 and 27522_5). This evidence aligns well with previous studies identifying Pim2 as a key modulator of MMPC proliferation\(^\text{131}\). Baseline heterogeneity between individuals was also evident; for instance, *CD79A* and *GPRC5D* are highly expressed in patient 47491, and *MS4A1* in 56203 compared to other patients at all timepoints (Fig. 6A).

Upon integrating longitudinal samples from patient 56203, we see that the primary tumor (56203_1) forms two distinct clusters (0 and 1 in Fig. 6B), while the remission tumor (56203_2) forms a third. Several markers appear upregulated in cluster 0 relative to cluster 1, including *SLAMF7, SDC1, EDNRB, LSR, POU2AF1*, and *FCRLA*; conversely, *TNFRSF17* and *SEC11C* are lower (Fig. 6C). In the relapse cluster 2, upregulation of marker genes, including *MS4A1, SDC1, CAV1, EDNRB*, and *FCRLA*, become increasingly pronounced. This pattern of overall higher marker expression at relapse can be found in patients 59114 and 60329 as well, though to lesser degrees (Fig. 6A). Conversely, cluster 2 expression of *TNFRSF13B* and *POU2AF1* are lower than cluster 0 and similar to cluster 1 levels. These observations suggest that cluster 0 may be an intermediate tumor phenotype for 56203, and that escalating levels of tumor marker expression may in general be indicative of advancing disease.

Our most complete longitudinal profile was of patient 27522, for whom we sampled from six timepoints representing primary tumor (timepoint 1), first remission (2), first and second relapses (3, 4), second remission (5), and third relapse (6; Fig. 6D). *TNFRSF17* is lost at the second
remission (5) and recovered at the third relapse (6); \textit{CD79A} is absent from the primary tumor but appears in the first remission (2) as well as in relapse timepoints 3 and 4; \textit{FCRL2} and \textit{TEX9} are absent from the first remission (2) but reappear in later time points (Fig. 6E). \textit{SLAMF7} is lost in an isolated subcluster at the third relapse (timepoint 6), which also lacks \textit{TEX9} expression. Even within a given time point, we see transcriptomic heterogeneity. In contrast, \textit{FCRL5}, \textit{SDC1}, and \textit{TNFRSF13B} are expressed fairly uniformly at all time points. Further studies may elucidate whether these fluctuations reflect physiological changes of the tumor in response to specific treatment regimens.

We further investigated the evolution of immunogenic mutations in patient 27522 by comparing predicted immunogenic peptides across three time points using bulk whole exome sequencing (WES) and RNA-seq, which were derived from a previous study\textsuperscript{58} (Supplementary Fig. S6A). Twenty putative neoantigens were detected in the primary MM sample and retained during MM progression, including potentially immunogenic peptides associated with the driver mutation \textit{NRAS} G13R. Seven putative neoantigens (\textit{BRD2}-P243Q, \textit{XRN1}-W1305C, \textit{SOCS6}-S106F, \textit{EVPL}-T1677S, \textit{TP63}-F181Y, \textit{PHLPP1}-N666S, and \textit{NIPBL}-A1250S) were lost during MM progression, suggesting potential immune-mediated clearance of subclones harboring these mutations in the primary sample. We then expanded our analysis to 34 patient samples from CoMMpass bulk RNA-seq dataset and WES data, which confirmed that the landscape of immunogenic peptides evolves appreciably as disease progresses. Only an average of 40\% of predicted immunogenic mutations overlapped across successive time points (Supplementary Fig. S6B). In MMRF patient 1931, new immunogenic mutations, such as \textit{KRAS}-G12R, \textit{SAMDI}-G309V, and \textit{CDK13}-L1011F, arose at the later time point, while others, such as \textit{UTRN}-M394I, and \textit{CABIN1}-N1252K, had receded
(Supplementary Fig. S6C). These evolutionary dynamics suggest that a tumor’s ability to initiate antigenic immune responses is time-dependent. Interestingly, we found four recurrent mutations with common predicted neoantigen across patients. They are from four genes, including KRAS, PABPC1, KMT2C, and RHPN2. As there are a limited number of predicted neoantigens overall, this strategy may be challenging for MM.

We next sought to shed light on expression patterns associated with specific clinical features. Given the diverse treatment background in WashU cohorts compared to the relatively uniform MMRF cohort, we restricted this analysis to MMRF samples and integrated all PCs from these 18 patients (Fig. 6F). Nine subjects were rapid progressors (Progression-Free Survival (PFS) < 18 months), and nine subjects were non-progressors (PFS not reached) (Fig. 6G). TNFRSF13C (BAFF-R) and RASGRP3 are significantly highly expressed in rapid progressors compared to non-progressors (Fig. 6H; logistic regression test; p = 0.02 and 0.03 respectively). RASGRP3 is involved in B-cell receptor signaling and is critical in B cell receptor-mediated Ras activation\(^{132,133}\). While the role of RASGRP3 in MM is not well understood, our observations suggest that TNFRSF13C and RASGRP3 may serve as indicators of MM progression.

### 3.3.7 Cross-referencing bulk RNA-seq and bulk protein expression for multifaceted characterization of myeloma associated markers

To better assess expression patterns across larger sample sizes, we first analyzed average normalized expression and logFC of candidate genes in integrated PCs from our combined discovery cohorts (Fig. 7A; integrated PCs shown in Fig. 4A). We then cross-referenced gene expression data with 907 bulk RNA-sequencing samples, including 892 CD138+ sorted samples
from the MMRF CoMMpass Study and 15 unsorted samples with matching single-cell data from WashU Cohort 1. When comparing PC gene expression from scRNA-seq to matching bulk CD138+ RNA-seq across the MMRF Immune Atlas pilot cohort, we see high positive correlation for a majority of genes (Pearson r ranging between 0.60 and 0.99; Fig. 7B). Notably, TNFRSF13C (r=0.80), CD79A (r=0.80), TNFRSF13B (r=0.76), CD38 (r=0.74), and TNFRSF17 (r=0.72), were among the genes with highest positive correlation, whereas CD24 (r=-0.08), PLPP5 (r=-0.05), CD79B (r= 0.05), and SLAMF7 (r=0.23) were among genes with the lowest. When comparing PCs in scRNA-seq to bulk, unsorted RNA-seq from WashU Cohort 1, we see that bulk expression of proposed target genes are highest in samples predicted by in silico deconvolution to have high plasma cell content, such as 83942, 27522_1, 47499, and 25183 (Fig. 7B).

Looking at overall trends in MMRF CoMMpass data, we see that average non-zero bulk expression levels of target genes largely agree with normalized expression in PCs from our scRNA-seq cohort. Twenty-nine out of 38 genes encoding candidate surface targets are detected at levels higher than global median non-zero gene expression in >80% of the entire MMRF bulk RNA-seq cohort (Fig. 7C). These genes may represent potentially effective targets for large portions of the myeloma patient population. Among them, TNFRSF17, SLAMF7, FCRL5, CD79B, P2RX5, LAX1, CD40, HVCN1, and CD79A are primary candidates with high tissue specificity. Primary surface candidates with lesser prevalence include TNFRSF13C (higher expression than global median in 67% of samples), FCER2 (52%), MS4A1 (42%), and CD19 (30%). 9 of 15 tissue-specific intracellular targets also have higher-than-median expression with >80% prevalence, including EAF2, PIM2, FCRLA, AIM2, and BANK1.
Due to the limitations of scRNA-seq in both cohort size and sequencing depth, and having established the general concordance of both data types, we next utilized the CoMMpass bulk RNA-seq dataset as a supplementary cohort for target discovery. We selected 1,261 genes with expression above the 95 percentile in at least 75% of 892 CD138+ sorted samples. Next, we stratified candidate genes based on their tissue specificity (evaluated via GTEx and HPA) into categories previously described: primary and secondary surface targets, and tissue-specific intracellular targets. We also examined the expression of these genes across over 1000 cancer cell lines in the Cancer Cell Line Encyclopedia (CCLE), retaining 21 genes with significantly higher average expression across 25 MM cell lines compared to those of other tissues (FDR<0.1, Supplementary Fig. S7A). Top candidate genes identified by this method (including SLAMF7, TNFRSF17, FCRL5, CD79B, and GPRC5D) had been identified in scRNA-seq cohorts as well, reaffirming the validity of our scRNA-seq-based target discovery strategy. Other targets, such as CD53 (a possible growth regulator of hematopoietic cells134), CD48 (which encodes a protein involved in lymphocyte adhesion and activation135), DNAJC1, TXNDC11, IRF4, and CPNE5, were not captured by our scRNA-seq discovery. Among these, however, only TXNDC11, a thioredoxin that localizes to the endoplasmic reticulum (ER)136, is both restricted to MM cell lines (CCLE; Supplementary Fig. S7B) and specifically overexpressed in PCs/B cells in our scRNA-seq data (Supplementary Fig. S7C). Its omission from our scRNA-seq-driven discovery may be due to sampling variation, as it failed to meet our criteria for prevalence of overexpression. Other targets discovered only via bulk RNA-seq were not specifically expressed by PCs and/or B cells in our scRNA data. While sampling variation may affect these results, our findings demonstrate the imprecision of sort-based lineage isolation upon which bulk sequencing strategies rely.
To assess protein-level expression of candidate genes, we performed global proteomic analysis via label-free quantitation (LFQ) and tandem mass tag (TMT) mass spectrometry. 4 MM cell lines (TIB.U266, RPMI8226, OPM2, and MM1ST) and 4 patient samples (Fig. 7D) were analyzed by LFQ. The subsequent TMT-Pro assay was comprised of 12 patient samples (Fig. 7E), 6 of which were also divided for concurrent bulk RNA-seq (Fig. 7F). For each sample with paired bulk RNA and proteomic data, we see overall positive correlation across 40 target and 3 housekeeping genes (Fig. 7F). Across these 6 samples, 23 target genes had positive protein-vs-RNA correlation, 14 with r>0.5 and 2 (CD38, LSR) reaching significance. Other genes with high positive correlation include CD53 (r=0.75), CADMI (r=0.74), MZB1 (r=0.73), and CD40 (r=0.72). Overall, trends in protein expression appear consistent across samples and between batches; however, we see clear differential expression between patients for CD40 and GPRC5D in the second batch (Fig. 7E). Genes with highly negative protein-vs-RNA correlation include DPEP1 (r=-0.65), TNFRSF13B (r=-0.59), and KCNN3 (r=-0.52). As low protein detection may be due to protease incompatibility during peptide extraction or other technical artifacts, these negative correlations remain inconclusive. Taken together, in our two proteomic datasets 57% (30/53) of targets discovered via scRNA-seq and 90% (19/21) of targets discovered via bulk RNA had detectable protein expression (Fig. 7D and E). 5 genes (CD38, DNAJC1, MZB1, SEC11C, and IRF4) were expressed at levels surpassing the detected global median in nearly all 20 samples profiled by proteomic quantitation, and 8 more (SLAMF7, CD53, CD48, SDC1, SLC1A4, ICAM3, POU2AF1, and IL16; Supplementary Fig. S7D) surpassed global medians in a majority of samples in one of the two batches. Although further functional studies are required, our proteomics data serve as an important starting reference that demonstrates the overall concordance between RNA and protein...
levels for candidate targets, as well as the heterogeneity between patients in both RNA and protein expression.

3.4 Discussion

We have described a strategy for high-throughput discovery of MM therapeutic targets that captures known CAR-T targets and identifies previously understudied MM markers. Our analysis strategy has 3 key advantages: (1) systematic search for genes specifically overexpressed in PC/B cells without relying on prior sorting; (2) direct comparison of cell types simultaneously captured by scRNA-seq that minimizes technical biases; (3) versatility and ease of application towards identifying tumor markers in other cancer types.

Through pathway enrichment analysis, we offered a “broad-strokes” view of how upregulated genes in PC/B lineages may affect protein synthesis and immune signaling. The power of database-driven functional analysis lies in its ability to narrow the search space for relevant biological processes. However, further study is needed to elucidate how individual genes promote or inhibit these processes before therapeutic strategies can be built around disrupting their function.

We proceeded to examine tumor heterogeneity at both inter-patient and sub-tumor levels using multiple data types profiling multiple patient cohorts. Taken together, our data reinforce the notion that patient heterogeneity restricts target selection when designing treatment strategies. Consistent co-expression may constitute a particular disease profile, wherein intratumor subpopulations
exhibit heterogeneous expression of a limited group of markers, but are unlikely to express an altogether different set of markers. Alternatively, pairing targets from consistently exclusive marker groups may be more effective in “catching” tumor cells that undergo antigen escape as they transition into different disease phenotypes.

Our findings suggest the existence of discrete target co-expression phenotypes, as well as the potential stochastic variation in consistently co-occurring targets within otherwise highly uniform tumor clusters. This latter finding requires further study, as distinguishing consistent phenotypical differences from dynamic expression changes is critical to solving the challenges of antigen escape and antigen coverage. To cross-validate our proposed targets, we examined marker gene expression in bulk RNA-seq from a large cohort of CD138+ sorted MM PCs, and quantified encoded protein levels in MM cell lines and primary samples using immunophenotyping and mass spectrometry. While limited by technical incompatibilities, we demonstrated overall concordance between data types for most targets.

Our strategy can be improved in several ways. Ideal discovery datasets would include patients exposed to uniform treatment regimens and controlled for clinical demographics. The three independent cohorts in our study were sequenced using two different library preparation chemistries, and although we were largely able to remove batch effects, subtle technical biases may persist. Due to our limited cohort of normal BM, we were unable to deeply investigate therapeutic opportunities in MM precursors, and due to the low coverage of scRNA-seq data, we are yet unable to perform expression correlation at single-cell resolution. Our analysis has revealed intratumor heterogeneity of tumor markers (Fig. 6B-E), and further investigation may elucidate
the clinical or pathological significance of MM subpopulations. We were unable to confirm cellular localization of all proposed targets due to limited antibody availability for conclusive immunophenotyping, and further experimental validation is therefore required. Future studies functionally characterizing our proposed MM markers, and assessing the toxicity and efficacy of targeting them, will ultimately decide their clinical utility.

We are mindful that the full collection of MM-associated antigens may still be incomplete; however, this study represents the most systematic search of myeloma antigens to date, and our findings provide researchers with a reference list of candidate targets to aid the development of immunotherapy and other targeted therapies against MM.

3.5 Materials and Methods

**Ethics approval and consent to participate**

All procedures performed in studies involving human participants were in accordance with ICH guidelines, applicable regulations and guidelines governing clinical study conduct, and the ethical principles that have their origin in the Declaration of Helsinki. Written informed patient consent was obtained from all patients for the collection and analysis of their samples. The CoMMpass study was conducted in accordance with recognized ethical guidelines in the US and EU and the Institutional Review Board (IRB) at each participating center provided protocol oversight. All work conducted at WashU was performed under the oversight of the IRB at Washington University in St. Louis.
**Processing of BMMCs prior to library preparation**

WashU Cohort 1, WashU Cohort 2, and MMRF Immune Atlas Pilot Cohort bone marrow aspirates were collected at varying disease timepoints. Bone marrow mononuclear cells (BMMCs) were isolated using Ficoll-Paque purification and cryopreserved in a 1:10 mixture of dimethyl sulfoxide and fetal bovine serum. Upon thawing in 37°C water baths, whole BMMCs from WashU cohort 1 were loaded onto the 10X Genomics Chromium Controller using 10x Genomics Chromium Single Cell 3’v2 Library Kits.

BMMCs aliquots from WashU Cohort 2, MMRF Pilot, and healthy donors were centrifuged upon thawing at 300G for 5 min to pellet cells. All supernatant was removed. To prepare cells for the Miltenyi Dead Cell Removal Kit, cells were resuspended in 100 uL of beads and incubated at room temperature for 15 minutes. Cells were then run through the DepeleS selection using the autoMACS® Pro Separator. The negative fraction (live cells) were pelleted by centrifugation at 450G for 5 minutes. Cells were finally resuspended in ice cold phosphate buffer saline (PBS) and 0.5% BSA and loaded onto the 10x Genomics Chromium Controller. WashU Cohort 2 and healthy donor samples were loaded using 10x Genomics Chromium Next GEM Single Cell 3’ GEM, Library & Gel Bead Kit v3.2. MMRF Pilot samples were loaded using 10x Genomics Chromium Controller and using the Chromium Next GEM Single Cell 3’ GEM, Library & Gel Bead Kit v3.3.

**Single cell library prep and sequencing**

Using the 10x Genomics Chromium Single Cell and Chromium instrument, approximately 16,500 to 20,000 cells were partitioned into nanoliter droplets to achieve single cell resolution for a
maximum of 10,000 individual cells per sample. The resulting cDNA was tagged with a common 16nt cell barcode and 10nt Unique Molecular Identifier during reverse transcription. Full length cDNA from poly-A mRNA transcripts was enzymatically fragmented and size selected for ~400bp cDNA amplicons optimal for library construction (10x Genomics). The concentration of the 10x single cell library was determined through qPCR (Kapa Biosystems) or automated electrophoresis (Agilent TapeStation) to produce cluster counts appropriate for the HiSeq 4000 or NovaSeq 6000 platform (Illumina). 26x98bp sequence data were generated targeting between 25K-50K read pairs/cell, which provided digital gene expression profiles for each individual cell.

scRNA-seq data quantification preprocessing

For single cell RNA sequence data was de-multiplexed to FASTQ, aligned to the human reference genome (GRCh38), and transposed into gene-by-cell UMI count matrices using the proprietary software tool Cell Ranger (v3.0.0) from 10x Genomics.

Seurat v3.0.0 was used for all subsequent analysis. First, all data underwent quality filtering (as recommended by Seurat) to remove barcodes with: too few total transcript counts (<300); too few genes expressed (<200) or too few UMIs (<1,000), indicating possible debris; too many genes expressed (>50,000) and too many UMIs (>10,000), indicating the presence of multiplet; or too high proportion of mitochondrial gene expression over the total transcript counts (>20%), indicating a dying cell. The median number of genes detected per plasma cell is about 1,500 in our datasets.
Seurat objects were constructed for each sample using its unfiltered feature-barcode matrix. Each sample was scaled and normalized using Seurat’s ‘SCTransform’ function to correct for batch effects (with parameters: `vars.to.regress = c("nCount_RNA", "percent.mito"), return.only.var.genes = F).

**scRNA-seq cell type annotation**

Cell types were assigned to each cluster by manually reviewing the expression of marker genes. The marker genes used were CD79A, CD79B, MS4A1 (B cells); CD8A, CD8B, CD7, CD3E (CD8+ T cells); CD4, IL7R, CD7, CD3E (CD4+ T cells); NKG7, GNLY (NK cells); MZB1, SDC1, IGHG1 (Plasma cells); FCGR3A (Macrophages); CD14, LYZ (Monocytes); FCER1A, CLEC10A (Dendritic cells); AHSP1, HBA, HBB (Erythrocytes); CLEC4C, LILRA4 (pDC); and AZU1, MPO, ELANE (Neutrophils).

**Differential expression analysis**

Differential expression analysis was performed using the default test (Wilcoxon Rank Sum test) of function `FindMarkers` (from the Seurat package) with the specified parameters: `min.pct=0.25`, `logfc.threshold = 0.25`, and `only.pos = T`.

**scRNA-seq driven tumor cell associated marker discovery**

Potential tumor-specific marker discovery was done in Seurat by comparing gene expression between tumor cells and non-tumor cells in patient samples. A gene is determined as tumor cell specific if both the following criteria are satisfied: 1) the average expression of the gene is higher in tumor cells compared with any other cell type, respectively, for at least one sample, and that all
the differences are of statistical significance (logFold Change >0; adjusted p value<0.05); 2) the average expression of the gene is higher in tumor cells compared with non-tumor cells (as a combined population) for 90% of the samples and that such differences are found to be statistically significant in at least 75% of the samples. Here, all p values were adjusted stringently by Bonferroni correction.

**Tumor cell associated marker subcellular location annotation**

To find potential antigens, we further annotated tumor cell specific genes by their subcellular location and tissue specificity. We used 3 databases to curate the subcellular location information: 1) Gene Ontology Term 0005886; 2) Mass Spectrometric-Derived Cell Surface Protein Atlas (CSPA)\textsuperscript{92}; 3) The Human Protein Atlas (HPA) subcellular location data based on HPA version 19.3 and Ensembl version 92.38.

**GTEx tissue expression specificity analysis**

Expression data and metadata were downloaded from the GTEx portal (gtexportal.org). We then designed a more focused analysis than the ~3500 hypothesis tests implied by a would-be round-robin approach for the 53 GTEx tissue type/sub-type categories and plasma and/or B cell-specific genes in Fig. 1. First, we distilled GTEx categorizations into 31 major tissue types, generally corresponding to the first word in the GTEx SMTSD tissue type designation. For each gene, we then identified good hypothesis test candidates, i.e. those showing high average value of Transcripts Per Million as evaluated over the major tissue types, using Peirce’s criterion\textsuperscript{137}. Each candidate combination of tissue and gene was then tested against the null hypothesis that its average expression did not exceed the grand average of the remaining tissue types in that gene, the
specific test being a standard difference-of-means t-test for unequal variances. Then, p-values were corrected for multiple tests using the standard Benjamini-Hochberg FDR.

**HPA tissue expression specificity analysis**

Normal tissue protein expression data were downloaded from The Human Protein Atlas (HPA) portal (proteinatlas.org). These data are based on The Human Protein Atlas version 19.3 and Ensembl version 92.38. We sought to identify proteins that are statistically significantly highly expressed in bone marrow and lymphoid tissues. For each gene, we tested expression for each tissue against the remaining tissue expressions as a background (“take one out” approach) using standard t-testing. The resulting p-values were corrected for multiple tests using the standard Benjamini-Hochberg FDR. In addition, tissue expression specificity analysis was also performed based on RNA consensus tissue gene data downloaded from HPA.

**Manual revision to refine tissue-specificity annotation and protein localization assignment**

To remove false positives from GTEx tissue expression specificity statistical analysis and HPA tissue expression specificity statistical analysis, we manually reviewed expression of candidate targets across tissue on GTEx webpage ([https://www.gtexportal.org/home/](https://www.gtexportal.org/home/)) and HPA webpage ([https://www.proteinatlas.org/](https://www.proteinatlas.org/)). Candidate targets with high and specific expression in lymphocytes (according to GTEx) as well as bone marrow and lymphoid tissues (according to HPA) are defined as tissue-specific targets. In terms of protein localization assignment, in addition to subcellular location annotated by 3 databases, namely Gene Ontology, CSPA and HPA, we refined the protein cellular location annotation by reviewing literature support.
Comparison of gene expression between tumors and normals

Single cell level normalized expression was used to compare expression of myeloma cell associated genes between 53 tumor bone marrows and 8 normal bone marrows, again using standard t-testing and FDR correction.

Pathway Analysis

We performed pathway overrepresentation analysis on the gene list output from our filtering pipeline using ConsensusPathDB\textsuperscript{138} against pathways defined by the Kegg\textsuperscript{139} and Reactome databases\textsuperscript{140}, with a 0.05 p-value cutoff. 42 pathways were identified as significantly overrepresented by our gene list; we then categorized each pathway under a broad biological function using descriptive information provided by the database sources. Gene Ontology enrichment analysis was also performed using ConsensusPathDB, using the “Biological Function” (B) level 3 search filter with enrichment cutoff FDR (q) <0.05.

Essential Genes Analysis

Functional dependency scores of proposed myeloma markers were obtained from the DepMap database via the depmap R package (\url{https://depmap.org/portal/}).

Flow validation

Antibodies used (clone and source in parenthesis): CD38-BUV395 (HB7; BD Biosciences), CD8-BUV496 (RPA-T8; BD Biosciences), HLA-DR-BUV661 (G46-6; BD Biosciences), CD56-BUV737 (NCAM16.2; BD Biosciences), CD4-BV510 (SK3; BD Biosciences), CD3-BV650 (UCHT1; BD Biosciences), CD20-BV785 (2H7; BioLegend), CD138-VioBrightB515 (44F9;
Miltenyi Biotec), gamma/delta TCR-BB700 (11F2; BD Biosciences), CD33-PECF594 (WM53; BD Biosciences), CD19-PE/Cyanine7 (HIB19; BioLegend), CD279-APC (REA315; Miltenyi Biotec), CD45-R718 (HI30; BD Biosciences), CD14-APC-Vio770 (REA599; Miltenyi Biotec). Sytox AADvanced (Invitrogen) was used for viability. The FCRL5 antibody (clone 952) was generated for us by Alloy Therapeutics by immunizing humanized Gx mice with human domain 9 protein. We conjugated the FCRL5 antibody to PE using a lightning-link kit (Innova Biosciences Ltd). Purified CS1 antibody (Luc90) was obtained from Creative Biolabs and used with a BV421 IgG secondary antibody. Cryopreserved unsorted bone marrow aspirates from myeloma patients (Washington University) and healthy PBMC obtained from leftover platelet apheresis products were thawed and run on a ZE5 Cell Analyzer (Bio-Rad) and data analyzed using FCS Express flow cytometry software (De Novo Software).

**Clustering plasma cells by candidate target expression profiles**

We used customized features, which are candidate targets, to compute PCA on. seurat_object <- RunPCA(seurat_object, npcs = 30, features=candidate_targets).

**MHC peptide binding prediction**

MHC binding prediction of protein sequences was performed using NetMHC (v4.0). Fasta sequences of myeloma-associated intracellular proteins were parsed into 8-11mer peptides and predicted binding affinity for frequently occurring MHC class-I HLA-A, -B, and -C alleles in the US population were calculated and compared to a set of 400,000 random natural peptides to determine strong binding (default top 0.5% of affinity values used as strong binder threshold).
**Bulk RNA-seq data generation for WashU internal samples**

RNA extraction for bulk sequencing is done using the RNeasy Mini Kit (Qiagen). Following library construction, samples are sequenced to a targeted depth of 80 million reads/sample using NovaSeq 6000 (Illumina).

**MMRF Whole-Exome Sequencing and bulk RNA-seq data downloading**

We obtained Whole-Exome Sequencing (WES) and bulk RNA-seq data from The Multiple Myeloma Research Foundation (MMRF) CoMMpass (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) Study (NCT01454297). dbGaP Study Accession: phs000748.

**Expression analysis and deconvolution of bulk RNA-seq data**

Gene expression was estimated using Kallisto (v0.43.1). The abundance of each cell type was inferred by XCell on sample-by-gene TPM matrices for each cohort, which performed the cell type enrichment analysis from gene expression data for 64 immune and stromal cell types (default xCell signature). xCell is a gene signatures-based method learned from thousands of pure cell types from various sources. We input the FPKM-UQ expression matrix of this study in xCell using the expression levels ranking.

**Bulk RNA-seq driven plasma cell associated marker discovery**

First, we selected 1,261 genes with expression above 95 percentile across genes in at least 75% samples of 892 CD138+ sorted samples from MMRF CoMMpass Study. We then identified genes with significant high expression in MM cell lines compared to other cancer cell lines in CCLE.
(n=1061; CCLE accessed March 23rd, 2021). For each gene, we identified good hypothesis test candidates, i.e. those showing high average value of Transcripts Per Million as evaluated over the major malignant cell lines, using Peirce’s criterion. Each candidate combination of cell line and gene was then tested against the null hypothesis that its average expression did not exceed the grand average of the remaining cell lines in that gene, the specific test being a standard difference-of-means t-test for unequal variances. Then, p-values were corrected for multiple tests using the standard Benjamini-Hochberg FDR.

Proteomics data generation and analysis

Protein Digestion

Viably frozen BMMC aliquots were thawed and incubated with Miltenyi CD138 microbeads prior to PosseID selection using the autoMACS®Pro Separator. Retained positive fractions were washed by ice-cold PBS and resuspended in cell lysis buffer (50 mM tetraethylammonium bicarbonate, TEABC, pH 8.0, 8 M urea, 1% protease and phosphatase inhibitor, pH 8.0). Protein concentrations were measured with a Pierce BCA protein assay (Thermo Fisher Scientific). Proteins were reduced with 10 mM dithiothreitol (DTT) for 30 mins at 25 °C and subsequently alkylated with 50 mM iodoacetamide (IAM) for 30 min at 25 °C in the dark. The concentration of urea was then diluted to 2 M by 50 mM TEABC for enzymatic digestion. Proteins in the sorted cells were digested with 0.1 mg Lys-C (Wako) and 0.1 mg sequencing-grade modified trypsin (Promega, V5117) at 25 °C for 4h and 16 h, respectively. For boosting samples, cell pellets from 3 cell lines (OPM, MM1S and Jurkat) were washed and digested similarly except for a 1:10 enzyme-to-substrate ratio for both Lys-C and trypsin digestion. Digested peptides were desalted by C18 solid-phase extraction (SPE) extraction. Resulting peptides from sorted cells were
dissolved in 5 mL 200 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES, pH = 8.5) and further digested by 0.1 mg Lys-C and 0.1 mg sequencing-grade modified trypsin at 25 °C to reduce the percentage of missed cleavage.

TMT Labeling

Digested peptides (in 200 mM HEPES) were mixed with 2 mL TMT-16 reagents (20 mg/mL) dissolved in 100% acetonitrile (ACN) and allowed to react for 1 h. An optimized ratio of TMT to peptide amount of 1:1 (w/w) was used (i.e. 100 mg of peptides labeled by 100 mg of TMT reagent). The labeling reactions were stopped by adding 5% hydroxylamine (final concentration is 0.5%) for 15 min and then acidified with trifluoroacetic acid (TFA; final concentration is 0.5%). Peptides labeled with different TMT tags were mixed in the same tube, after which the ACN concentration was adjusted to below 5% (v/v) and the samples were desalted by C18 SPE.

Peptide Fractionation by Basic Reversed-Phase Liquid Chromatography (bRPLC)

50µg of TMT-labeled peptide were dissolved in 0.1% TFA with 2% ACN (1 mg/ mL) and fractionated by the high-pressure, high-resolution separations coupled with intelligent selection and multiplexing (PRISM) fractionation method. A nanoACQUITY UPLC® system (Waters Corporation) equipped with a reversed-phase capillary liquid chromatography (LC) column (3 mm Jupiter C18 bonded particles packed in 200 mm i.d. × 50 cm) and an autosampler were used for fraction collection. Separations were performed by bRPLC fractionation at flow rate of 2.2 mL/min on the binary pump systems using 10 mM ammonium formate (pH 10) in water as mobile phase A and 10 mM ammonium formate (pH 10) in 90% ACN as mobile phase B. 45mL of sample was loaded onto the reversed-phase capillary LC column and separated using a 190-min gradient of (min:%B): 35:1, 37:10, 52:15, 87:25, 112:35, 125:45, 150:90, 156:1. A total of 96 fractions was
collected at equal time intervals and subsequently concatenated into 24 fractions for global proteome analysis.

LC-MS/MS Analysis and Protein Identification

The LC-MS/MS setting was similar to a previous study. Lyophilized peptides were reconstituted in 12 mL (TMT labeling) or 30 mL (Label-free) of 0.1% formic acid (FA) with 2% ACN and separated by a nanoACQUITY UPLC system (Waters) (buffer A: 0.1% FA with 3% ACN and buffer B: 0.1% FA in 90% ACN) as previously described. Peptides were separated by a gradient mixture with an analytical column (75 mm i.d. × 20 cm) packed using 1.9-mm ReproSil C18 and with a column heater set at 50 °C. Peptides were separated by an LC gradient: 2-6% buffer B in 1 min, 6-30% buffer B in 84 min, 30-60% buffer B in 9 min, 60-90% buffer B in 1 min, and finally 90% buffer B for 5 min at 200 nL/min. For TMT-labeled samples, data were acquired by Orbitrap Eclipse Tribrid mass spectrometer (Thermo Scientific) in a data-dependent mode with a full MS scan (m/z 400-1600) at a resolution of 120K with automatic gain control (AGC) setting set to 1×10^6 and maximum ion injection period set to 246 ms. The isolation window for MS/MS was set at 0.7 m/z and optimal higher-energy C-trap dissociation (HCD) fragmentation was performed at a normalized collision energy of 32% with AGC set as 1×10^6 and a maximum ion injection time of 246 ms. The MS/MS spectra were acquired at a resolution of 120K. For label-free analysis, data were acquired by Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific) in a data-dependent mode with a full MS scan (m/z 350-1800) at a resolution of 60K with AGC setting set to 4×10^5 and maximum ion injection period set to 50 ms. The isolation window for MS/MS was set at 2 m/z and optimal HCD fragmentation was performed at a normalized collision energy of 30% with AGC set at 1×10^5 and a maximum ion injection time of 105 ms. The MS/MS spectra were acquired at a resolution of 50K.
Data Analysis

MaxQuant\textsuperscript{145,146} was used to process the raw MS/MS data. MS/MS spectra were searched against a human UniProt database (fasta file dated April 12, 2017 with 20,198 sequences) or specific fasta file generated by Uniprot search for genes of interest. The default setting (4.5 ppm for peptides tolerance and 20 ppm for MS/MS match tolerance) was used for mass tolerance for precursor ions and fragment ions. The search type was set to “Reporter ion MS2” for isobaric label measurements. A peptide search was performed with Trypsin/P and allowed a maximum of two missed cleavages. Carbamidomethyl (C) was set as a fixed modification; acetylation (protein N-term) and oxidation (M) were set as variable modifications for global proteome analysis. The false discovery rate (FDR) was set to 1% at the level of proteins, peptides, and modifications; no additional filtering was performed. The intensities of all ten TMT reporter ions and LFQ values were extracted from MaxQuant outputs and analyzed by Perseus\textsuperscript{147} for statistical analyses.

\textbf{scRNA-seq data integration}

Cells from multiple samples were merged in Seurat, followed by scaling and normalization. All cells were then clustered using the original Louvain algorithm and top 30 PCA dimensions via ‘FindNeighbors’ and ‘FindClusters’ (with parameters: resolution = 0.5) functions. The resulting merged and normalized matrix was used for the subsequent analysis.

\textbf{Co-expression and mutually exclusive expression analysis using scRNA-seq data}

Sample-level average normalized expression in the all sample merged Seurat object was used to perform pearson correlation analysis. To determine the threshold of correlation coefficient(r) for co-expressed gene pairs and mutually exclusively expressed gene pairs, we sampled 2,000 random genes and plotted the distribution of their correlation coefficients. Also, we manually reviewed the
expression of some gene pairs in UMAP and decided to use 0.05 and 0.30 as cutoffs for mutually exclusive expression and co-expression of a gene pair respectively.

Co-expression and mutually exclusive expression analysis using bulk RNA-seq data

Transcript quantification was performed using Kallisto v0.46.2\textsuperscript{141}, against the Ensembl transcript reference (release 95, GRCh38). Subsequent analysis was performed using a python script to aggregate transcript level data to the gene level. Then, Pearson correlation analysis was performed on TPM estimates of target gene candidates.

Immunofluorescent imaging of bone marrow biopsies

5mm sections of formalin-fixed paraffin-embedded (FFPE) core biopsies of patient bone marrow were incubated at 55°C to prevent lifting and rehydrated using xylene followed by successive washes in ethanol at decreasing concentrations. Antigen retrieval was done using 10mM sodium citrate (prewarmed to 80-90°C) for 25 minutes, then blocked using 100mM glycine (2 rounds of 10 minute incubations) once cooled to room temperature and washed with 1X PBS. Sections were then incubated for 1 hour in blocking solution (10% normal donkey serum and 1% BSA in PBS) and rinsed twice in 1X PBS. Primary antibodies (CD138: Akoya #4450008; BCMA: Biolegend #357502; CS1: Biolegend #331802; FCRL5: Biolegend #340302) were diluted in blocking serum (all at 1:50) and incubated on sections overnight at 4°C. Sections were then washed using 1X PBS and incubated for 1 hour with secondary antibodies (Alexa Fluor® 647 AffiniPure F(ab')\textsubscript{2} Fragment Donkey Anti-Mouse IgG (H+L) and green 2nd antibody (1:1000): Alexa Fluor® 488 AffiniPure Donkey Anti-Rabbit IgG, both from Jackson ImmunoResearch) diluted to 1:1000 in blocking serum. Following secondary staining, sections were incubated with Hoescht 33342
staining (Thermofisher; 1:2000 in PBS), then washed in 1X PBS. Slides were mounted with VECTASHIELD (Vector Labs) mounting medium and analyzed on a Leica DMi8.

**Somatic Mutation detection from Whole-Exome Sequencing**

Somatic variants were called using our SomaticWrapper pipeline, which integrates the established bioinformatic tools Strelka, Mutect, VarScan2 (2.3.83), and Pindel (0.2.54)\textsuperscript{148–151}. SNVs identified by any 2 callers among Mutect, VarScan, and Strelka and INDELs identified by any 2 callers among VarScan, Strelka, and Pindel were retained. We then applied coverage cut-offs for these merged SNVs and INDELs of 14X and 8X for tumor and normal, respectively. We also filtered SNVs and INDELs with a high-pass variant allele fraction (VAF) of 0.05 in tumor and a low-pass VAF of 0.02 in normal. The SomaticWrapper pipeline is freely available from GitHub at https://github.com/ding-lab/somaticwrapper. We also applied this pipeline to MMRF samples.

**Neoantigen prediction**

The wild-type protein sequences are obtained from Ensembl Database. We constructed different epitope lengths (8mer, 9mer, 10mer and 11mer) from the translated protein sequence. Each sample’s HLA type was predicted by OptiType\textsuperscript{152} based on bulk RNA-Seq data. We predicted the binding affinity between epitopes and the major histocompatibility complex (MHC) using NetMHC4\textsuperscript{127}. Novel epitopes with binding affinity 500nM which are also not present in the wild-type transcript are reported as neoantigens.
3.6 Data availability statement

Single-cell RNA-seq and bulk RNAseq expression matrices are available at NCBI GEO under accessions numbers GSE223060 and GSE223061, respectively.

3.7 Figures
Figure 1. Myeloma cell associated therapeutic protein discovery workflow, datasets and overview.
A, Tumor cell associated therapeutic protein discovery pipeline. B, Stacked bar plot of cell type fractions for each sample, including both tumor bone marrow samples for discovery and normal bone marrow samples for validation. Bar segments are colored by cell types, grouped by datasets.

C, Overview of plasma-specific genes and combined plasma-and-B specific genes, intersections of the two as well as their cellular localization. D, Stacked bar charts showing number of samples with differentially expressed surface proteins in plasma and/or B cells compared to other cell types, colored by datasets.
Figure 2. Pathway overrepresentation analysis of plasma and B cell marker genes using the Kegg and Reactome databases.

A, Bubble plot of significantly enriched pathways identified via overrepresentation analysis. Pathways are binned into functional categories and plotted by significance (-log10 of FDR, q); bubble size indicates number of genes called in the enrichment result. Yellow horizontal line indicates a 5% significance threshold. DEG= differentially expressed genes; ER = endoplasmic reticulum; GIP = gastric inhibitory polypeptide; GLP = glucagon-like peptide; BCR = B-cell receptor. B, Bubble plot of fold change and cellular localization of differentially expressed plasma and B cell genes identified as functionally relevant by pathway enrichment. Bubble size indicates fold change (ln) of gene expression in either plasma cells or combined plasma and B cells (as indicated by gene name color) relative to other bone marrow cells; color denotes predicted cellular localization of the encoded protein.
Figure 3. Characterization of myeloma cell associated surface proteins.
A, Average gene expression in plasma cells (top - box and whisker plot) and hierarchical clustering of z-score scaled log fold change between plasma and non-plasma cells across samples (bottom - heatmap). B, Summarized characterization of primary candidate targets expressed on the cell surface with specific expression in lymphoid tissues. Genes are ordered by average expression in PCs. Top 4 genes are top-tier primary targets, highlighted in red. Average normalized expression in plasma and B cells are shown in the first 2 sections respectively. Average expression log fold change between PCs and other cells is shown in the third section followed by expression log fold change of combined plasma and B cells compared to other cells shown in the fourth section (FC: fold change). Each dot represents a sample, colored by its corresponding cohort. Sample frequency is the proportion of samples having genes differentially expressed in PCs or in combined plasma and B cell population if the gene is only expressed in B cells. In the 6th section, triangle and circle denote whether genes are significantly highly expressed in tumors relative to normal BMs. Cell types with target expression are annotated in the 7th section. In the GTEx tissue expression specificity analysis section, only gene-tissue type relationships with a significant FDR are plotted. Size corresponds to fold difference and colored by expression. HPA protein expression was shown in the last section with color indicating expression level. NA= not available; TPM= transcript per million; BM= bone marrow. C, Summarized characterization of secondary candidate targets expressed on the cell surface without specific expression in lymphoid tissues. Figure layout is the same as panel B. D, Heatmap showing the z-score scaled average expression of known and novel targets across cell populations in scRNA-seq. Columns are ordered based on hierarchical clustering of target expressions; rows are ordered by average normalized expression in plasma cells. Left annotation indicates novelty level: 1) currently under clinical study as CAR-based therapy; 2) currently under clinical study as antibody-based therapy; 3) potential therapeutic utility is
supported by existing literature; 4) previously uncharacterized in its capacity as a myeloma marker.

E, Heatmap showing the average percentage of cells with positive expression of surface antigens across cell populations from 12 patient samples (11 MM and 1 SMM patients) in flow cytometry.

F, Box plot showing mean fluorescence intensity (MFI) of BCMA in patient BMMC (n=12, including 11 MM and 1 SMM) or peripheral blood mononuclear cell (PBMC) from healthy donors (n=3) across cell populations in flow cytometry.
Figure 4. Expression correlation of myeloma cell associated genes encoding surface proteins.
A, UMAP showing plasma cells from 53 patients reveal tumor-specific clusters, colored by patient. 

B, Dot plots showing correlation of gene expression averaged by sample in all 53 samples, colored by pearson correlation coefficient (r value). Only predicted co-expressed gene pairs (r>0.30) and mutually exclusively expressed gene pairs (r<0.05) are shown. Co-expressed gene groups are highlighted in triangles with visually confirmed ones listed on the right. Bar charts showing number of predicted gene pairs (r<0.05 or r>0.30) in 2 datasets or 3 datasets. C, Gene pairs showing mutually exclusive expressions (top row) or co-expressions (bottom row). The first box of each row shows an example of expression of gene A (first UMAP), expression of gene B (second UMAP) and dual expression of gene A and gene B (third UMAP). D, Bar chart showing percentage of plasma cells from 12 patient samples (11 MM and 1 SMM) with co-expression of surface proteins detected by flow cytometry. E, Pearson correlation plots of gene pair (same genes shown in panel D) expression in scRNA-seq.
Figure 5. Characterization of myeloma cell associated intracellular proteins.

A, Average gene expression in plasma cells (PCs) shown in box and whisker plot (top) and hierarchical clustering of z-score scaled log fold change (bottom) between plasma and non-plasma cells across samples shown in heatmap. B, Summarized characterization of intracellular targets specifically expressed in lymphoid tissues. Genes are ordered by average expression in PCs. Average normalized expression in plasma and B cells are shown in the first 2 sections respectively. Average expression log fold change between PCs and other cells is shown in the third section followed by expression log fold change of combined plasma and B cells compared to other cells.
shown in the fourth section (FC: fold change). Each dot represents a sample, colored by its corresponding cohort. Sample frequency is the proportion of samples having genes differentially expressed in PCs or in combined plasma and B cell population if the gene is only expressed in B cells. In the 6th section, triangle and circle denote whether genes are significantly highly expressed in tumors relative to normal BMs. Secreted proteins in blue. Unknown protein localization in gray. Cell types with target expression are annotated in the 7th section. In the GTEx tissue expression specificity analysis section, only gene-tissue type relationships with a significant FDR are plotted. Size corresponds to fold difference and colored by expression. HPA protein expression was shown in the last section with color indicating expression level. NA= not available; TPM= transcript per million; BM= bone marrow. C, Bubble plots showing the normalized expression of SEC11C and POU2AF1 averaged by sample and cell type. D, Heatmap showing the z-score scaled average expression of promising intracellular targets across cell populations in scRNA-seq. Columns are ordered based on hierarchical clustering of target expressions. E, Number of peptides for each candidate gene product predicted to have high-affinity binding to common MHC class-I alleles in the US population as determined by NetMHC.
Figure 6. Expression of myeloma cell associated therapeutic targets in relation to clinical features.
A, Heatmap showing normalized expression of candidate targets across longitudinal samples. Different disease stages are annotated in different colors on the right of the heatmap. B, UMAP showing plasma cells from samples taken at two different stages of patient 56203, colored by seurat clusters. C, Heatmap showing normalized expression of candidate targets across three plasma populations in patient 56203. Genes detected in fewer than 5 cells are not considered during normalization and thus were not included in the heatmap. D, UMAP showing plasma cells from samples taken at six different time points of patient 27522, colored by sample. E, Expression of typical myeloma cell associated therapeutic targets reveals diverse and dynamic change of gene expression along the disease progression in the same patient. F, UMAP showing plasma cells from samples in MMRF Immune Atlas Pilot study, colored by patient. G, UMAP showing plasma cells from samples in MMRF Immune Atlas Pilot study, colored by patients’ progression features. H, Violin plots showing expression of TNFRSF13C and RASGRP3 are significantly elevated in rapid progressors compared to non-progressors (p<0.05).
Figure 7. Comparison of scRNA-seq, bulk RNA-seq, and bulk proteomic expression of candidate targets.

A, Average normalized expression (ln) and average fold change (ln) of target genes across scRNA-seq discovery cohorts. HK: Housekeeping genes. B, Unsorted bulk RNA-seq expression of WashU Cohort 1 and CD138+ sorted bulk RNA-seq of the MMRF Immune Atlas Pilot Cohort. XCell deconvolution of plasma and B cell relative abundance shown along bottom. Plasma cell percentages from matching scRNA-seq data is shown for WashU Cohort 1. Correlations between plasma cell expression (scRNA) and bulk CD138+ expression of each gene for MMRF pilot samples are shown in the rightmost column. C, Average non-zero expression and expression prevalence across MMRF CoMMpass bulk RNA-seq. Expression prevalence is defined as % of all samples wherein target expression is higher than global median non-zero gene expression. D, Bulk protein expression of target gene products assayed via label-free proteome; LFQ values were analyzed by MaxQuant, Log2 transformed and subsequently internally normalized with missing values imputed. E, Bulk protein expression of target gene products assayed via TMT Pro; TMT intensities were analyzed by MaxQuant, Ln transformed, and internally normalized. F, Bulk RNA expression of 6 samples (names identified in purple) concurrently processed via TMT Pro. Sample-wise RNA-vs-protein correlations across all target genes shown is displayed along Y-axis; gene-wise RNA-vs-protein correlations across samples is displayed along X-axis.
Figure S1. Related to Figure 2

A, Gene ontology (GO) gene-set enrichment analysis of plasma and B-cell marker genes. GO terms from “Biological Function”, level 3 with enrichment FDR (q) <0.05 are shown for individual genes. Bubble size indicates fold change (ln) of plasma and B-cell expression relative to that of other BM cells; color indicates predicted cellular localization of gene products. B, Heatmap showing DepMap dependency scores, as assessed by CRISPR-Cas9 knockout, of myeloma markers across 21 plasma cell lines. As described by DepMap, a score of 0 indicates a non-essential gene, and -1 is the median dependency score of common essential genes. C, Bubble plots showing the normalized expression of chemokine and chemokine receptors, averaged by sample and cell type. Column corresponds to samples, colored by cohorts.
Figure S2. Related to Figure 3
Bubble plots showing the normalized expression of target candidates, averaged by sample and cell type. Column corresponds to samples, colored by cohorts. All genes are annotated as having cell surface expression.
Figure S3. Related to Figure 3

Box plots showing percentage of target positive cells (left) and MFI of targets (right) in MM BMMC or PBMC from healthy donors across cell populations in flow cytometry. Viable cell populations were gated on the following: CD45loCD319+CD38+CD138+ PCs; CD45+CD20+CD19+ B cells; CD45+CD14+CD33+ Monocytes; CD45+CD3+ T cells; CD45+CD3+gdTCR+ gdT; CD45+CD56+CD3+ NKTs; CD45+CD56+CD3- NK cells. Plasma cells were not detected in sufficient quantity for comparison in healthy donor PBMCs.
Figure S4. Related to Figure 4.
A, UMAP showing all cells from 53 samples, colored by cell type. Compared to immune populations, plasma populations formed much more dispersed clusters, indicating the transcriptomic heterogeneity of tumor cells across patients. B, UMAP showing plasma cells from 53 patients, clustered by candidate targets, colored by samples. Most samples are clustered around the center while some samples form their individual clusters. C, Distribution of correlation coefficient of 2000 random genes and candidate targets. Red box shows the cutoff of identification of co-expressions and the blue box shows the cutoff of identification of mutually exclusive expressions. D, Co-expressed or mutually exclusively expressed gene pairs found in all three cohorts. Gene pairs with $0.05 < r < 0.3$ were not shown. E, Pearson correlation heatmap of CoMMpass study bulk RNA-seq expression values. Gene groups with prominent positive correlation are boxed. Notable gene group containing canonical tumor markers have bolded names; a second group containing B-cell markers is labeled in purple.
Figure S5. Related to Figure 4.

A, In addition to Fig. 4C, more examples of gene pairs showing mutually exclusive expressions (left two) or co-expressions (right two). B, UMAPs of PCs from 4 patients colored by dual expression of gene pairs. C, Representative immunofluorescence (IF) co-staining of target pairs in bone marrow biopsies from patient s10-10686; 63x magnification; scale bar 100µm. Three channel overlay (Hoescht staining of nuclei in blue) enlarged, single channel violet/red and green images shown on right.

Figure S6. Related to Figure 6.

A, Variant allele frequency (VAF) and expression pattern of neoantigens from patient 27522 across three time points. B, Expressed immunogenic peptides across timepoints along with disease
progression. WES, Whole Exome sequencing. C, Predicted neoantigens for MMRF 1931 and the associated RNA and WES VAF from two time points.
Figure S7. Related to Figure 7.
A, Bar chart showing average expression of candidate markers in CD138+ sorted MM BM samples in the MMRF CoMMpass study. Color of bars indicates the proportion of samples having genes with expression above 95 percentile across all genes. Heatmap showing averaged mRNA expression of candidate markers across all profiled cell lines for each tissue type (n=1061; CCLE accessed March 23rd, 2021). DLBCL, diffuse large B cell lymphoma; AML, acute myeloid leukemia; CML, chronic myeloid leukemia; ALL, acute lymphoblastic leukemia; NSC, non–small cell. B, mRNA expression of TXNDC11 in malignant cell lines in CCLE. C, Bubble plot showing the normalized expression of TXNDC11 averaged by sample and cell type. Row corresponds to samples, colored by cohorts. D, Bar plot showing number of samples in which target protein expression surpasses the detected global median in that sample. Bars are colored by batch chemistry.
Chapter 4: Landscape of tumor and the immune environment in multiple myeloma onset and progression

4.1 Summary

Multiple Myeloma (MM) is a hematologic cancer marked by uncontrolled clonal expansion of plasma cells. Bone marrow microenvironment (BMME) could promote myeloma cell proliferation, survival, and drug resistance by direct and indirect interactions with MM cells. Therefore, characterizing how tumor and tumor microenvironment interplay myeloma disease onset and progression is of great interest. By leveraging 361 scRNA-seq from 263 MM patients enrolled in the Multiple Myeloma Research Foundation (MMRF) CoMMpass study, we delineated tumor heterogeneity and malignant B cells to plasma cell transition, depicted lineage compositional changes and their associated gene sets during MM progression. To note, we observed that a tumor subpopulation had significant enrichment of baseline plasma cells, while relapsed plasma cells with DIS3 mutations were specifically abundant in one small plasma subpopulation. We identified B cell subpopulations as precancer populations given their higher mutation burden. We next characterized the immune landscape in relation to MM progression. We observed compositional alterations of immune subsets from baseline to relapse stages. Interestingly, relapse samples have
significant enrichment for earlier B subtypes, cytotoxic and interferon-responsive CD8 T cells, activated and interferon-responsive CD4 T cells, and CD56<sup>bright</sup> and bone marrow resident NK cells, while mature memory B cells, naive T cells, gamma delta T cells, CD16<sup>+</sup>monocytes, and IFN responsive CD14<sup>+</sup> monocytes are enriched at baseline. CD74, LTB, and IFN-stimulated genes are universally differentially expressed between baseline and relapse in multiple lineages and might be associated with MM progression. In summary, our study provides a systematic and comprehensive interrogation of tumor and immune microenvironment during myeloma onset and progression, which could deepen our understanding of myeloma disease progression and potentially provide novel targets for immunotherapies.

4.2 Introduction

Multiple myeloma, characterized by aberrant clonal plasma cell growth in the bone marrow, is the second most common hematologic malignancy and affects approximately 588,161 patients worldwide annually<sup>153,154,155</sup>. Much progress has been made over the years on multiple myeloma, including elucidating genetics of the cancer, recognizing contributions of the tumor microenvironment, identifying prognostic factors, and improving efficacy of therapeutics<sup>155–157</sup>. Despite these advances, however, treatment of the disorder remains challenging, as many patients relapse<sup>155</sup>.

Prediction of disease progression and relapse of multiple myeloma has been of much clinical interest, as there is a paucity of relevant tools. The revised International Staging System, which utilizes the International Staging System (ISS) with chromosomal abnormalities (CA) to predict relative risk to their survival<sup>157</sup>, is essentially the only commonly used prognostication method in
the clinical setting. However, studies have shown that 1q gain, which was not included in the R-ISS, proved to be a poor prognostic factor\textsuperscript{158}. This suggests that more work is necessary for elucidation of factors driving disease progression of multiple myeloma. Some of these factors include genetic factors and influences of the tumor microenvironment underlying progression of disease.

Previous research led by O Zavidij\textsuperscript{51} has revealed natural killer (NK) cell abundance is enhanced in the early stages and associated with altered chemokine receptor expression during the evolution from Monoclonal gammopathy of undetermined significance (MGUS) to smoldering MM (SMM) to MM. Tirier et al. examined interplay between tumor and bone marrow environment before and after treatment for Relapsed/refractory multiple myeloma (RRMM)\textsuperscript{159}. Up until now, however, there is no published research comprehensively characterizing the landscape of tumor and immune microenvironment in MM onset and progression in a large cohort.

In this study, we subjected Bone Marrow Mononuclear cells (BMMC) samples to CD138 sorting. CD138-positive sorted fraction was subjected to WES, WGS, and bulk RNA-seq for tumor genomic landscape profiling, while CD138-negative sorted fraction was subjected to scRNA-seq for immune environment profiling.

We collectively analyzed 1,149,344 single cells from 263 patients, including 97,291 plasma cells and 1,052,053 immune cells. Deeper dissection of plasma cells and B cells reveals that there are two distinct sets of early and immature B cells that link pro-B cells with either memory B cells or plasma cells. This suggests that transcriptomic differences accompanying B cell differentiation
into these two terminal states could be merging as early as the pre-B stage. By comparing abundances of immune subsets from relapse samples to those from baselines samples, we observed significant enrichment for earlier B subtypes, cytotoxic and interferon-responsive CD8 T cells, activated and interferon-responsive CD4 T cells, and also CD56<sup>bright</sup> and bone marrow resident NK cells. On the contrary, baseline samples have significant enrichment of mature memory B cells, naive T cells, gamma delta T cells, CD16+ monocytes, and IFN responsive CD14+ monocytes. Differential expression gene analysis revealed that CD74, LTB, and IFN-stimulated genes might be associated with MM progression. In summary, our study provides a systematic and comprehensive interrogation of the landscape of tumor and immune microenvironment during myeloma disease onset and progression.

4.3 Results

4.3.1 Overview of datasets, tumor and the immune environment landscape in MM bone marrow

To understand the genomic alterations and characterize immune profiling of MM patients, CD138-positive sorted BMMC was subjected to bulk sequencing and CD138-negative sorted BMMC was subjected to scRNA-seq. Overall, we performed bulk RNA-seq on 297 samples, WES on 381 paired tumor-normal samples from 325 patients, WGS on 379 samples from 325 patients, and
scRNA-seq on 361 CD138-negative sorted Bone Marrow Mononuclear cells (BMMC) samples from 263 patients enrolled in the Multiple Myeloma Research Foundation (MMRF) CoMMpass study (NCT01454297) (**Figure 1A**).

We have performed somatic mutation calling, Copy Number Variation (CNV) calling from the whole exome sequencing (WES) data and structural variant calling from the whole genome sequencing (WGS) data. Utilizing the bulk analysis including structural variants, somatic mutations, CNV and clinical lab testing, we hope to see a distinct pattern between patients that progressed fast (Progression Free Survival <18 months) and patients that didn’t have disease progression within 18 months of first diagnosis.

The most prevalent mutations include *KRAS* and *NRAS*, which are altered in 23% samples. Following that, *DIS3* and *BRAF* mutations are present in 10% of samples (**Figure 1B**). We were able to identify individual risk factors for disease progression, for instance, race, sex, ISS stage and several genetic alterations at bulk level were associated with early progression from Cox proportional hazards regression models (**Figure 1C and 1D**). We tried to build a risk prediction model for multiple myeloma progression utilizing bulk data and clinical testing. However, even with all of these characteristics, it is difficult to accurately estimate a patient's prognosis. This motivated us to try to take advantage of scRNA-seq to dissect the progression mechanism of multiple myeloma. At the single cell level, we thought we might be able to identify characteristics that might potentially explain how the disease progresses.

scRNA-seq data was generated and analyzed at four different academic research centers, namely Beth Israel Deaconess Medical Center (BIDMC), Mayo Clinic, Emory University, Icahn School of Medicine at Mount Sinai (ISMMMS) and Washington University in St. Louis. Patients were
stratified into 2 groups based on progression: 173 NON-progressors (NPs) (PFS>18 years), and 67 FAST-progressors (FPs) (PFS<=18 months). 204 patients had both Proteinase inhibitors and IMIDs in their first line treatment and 129 underwent bone marrow transplant in their first line treatment. Patient ages range from 27 to 93 years. 65 of them were the International Staging System (ISS) stage III, 97 were females, 32 were African Americans and 159 were Caucasians. 261 samples were collected at primary diagnosis, 32 at remission, and 57 at Relapse/Progression.

We showed that the log normalization method significantly outperforms Seurat SCTransform approach for integrating this large multi-center single cell dataset. This is probably due to the majority of reported expression levels being zeros in scRNA-seq, which is closely intertwined with the total number of UMIs. Furthermore, we performed batch correction using Harmony, where each batch from each site was considered a separate batch. To note, given the variance of plasma cells were mostly from immunoglobulin genes according to principle component analysis, each sample was considered a separate batch for harmony correction in plasma populations. In addition, we removed empty droplets identified by DropletUtils (version 1.14.2) with FDR <0.001 (0.1%) and ambient RNA using the latest version of cellBender.

Overall, we were able to profile 1,149,344 cells from 361 samples with about 3,200 cells per sample. To be specific, we identified 331,986 (29%) CD4+T cells; 202,320 (18%) CD8+ T cells; 95,571 (8%) NK cells; 100,070 (9%) B cells; 107,510 (9%) CD14+ Monocytes; 14,880 (1%) CD16+ Monocytes; 39,785 (3%) Neutrophil; 20,559 (2%) Erythrocytes; 22,130 (2%) pre/pro-B/mature B cells. (Figure 1E). To note, despite scRNA-seq was performed on CD138-negative sorted bone marrow samples, there are 323 plasma cells on average in baseline samples (ranging from 2% to 90%) and 269 plasma cells in relapse samples (ranging from 3% to 77%) (Figure 1F).
4.3.2 Plasma cells subpopulations are associated with genomic alterations and disease progression

From 361 samples, we isolated 92,275 plasma cells, grouped them into 21 clusters (plasma C1 - C21), and numbered each cluster according to the decreasing total cell counts inside it. We linked somatic mutations of 276 matching samples from WES data to scRNA data. (Figure 2A). In plasma cells, IGL, IGK, and CCND1 genes were the most frequently altered genes; IGLV3 was altered in 9331 cells, IGLV2 was altered in 2250 cells, and IGKV3 was altered in 612 plasma cells, respectively. In 612, 361 and 321 plasma cells, respectively, IGKV3, IGKV2D and IGKV1D mutations were found. Histone 1 (H1) was mutated in 390 cells while CCND1 was mutated in 536 plasma cells. As opposed to the majority of mutant genes discovered in bulk WES data, we only discovered 50 NRAS mutated cells (23% in bulk), 23 mutated KRAS mutated cells, 23 mutated DIS3 mutated cells, 19 IRF4 mutated cells, and 12 mutated FGFR3 cells (Figure 2B). This could be attributed to the fact that scRNA-seq data has lower coverage compared to WES data, but also could suggest that we may underestimate the mutational landscape of multiple myeloma at bulk level.

With particular mutations that were enriched in that cluster, we were able to locate multiple different clusters. Most mutations were found in Plasma C1, which also had a higher percentage of non-driver mutations (64%). Over 20% of the cells in plasma C6, C7, C8, C14, C17, C18 and C20 exhibited non-driver mutations. The majority of non-driver mutations are caused by genetic changes in immunoglobulin genes (Figure 2B). CCND1 is the most prevalent driver mutation and was present in 38.9% of the cells in plasma C16, suggesting that plasma C16 may be driven by
CCND1 mutations. In addition, when we looked at 6 patients who had CCND1 mutations in 12 plasma cell clusters at baseline sampling, we found that none of them had a fast progression over the course of 18 months.

We observed that particular clusters were related to various tumor growth phases. Relapsed samples with DIS3 mutations were more abundant in plasma C9 (Figure 2C). It's interesting to note that 5 out of 6 clusters showed considerably greater mutation frequency in baseline samples than that in relapse samples from plasma C1 to plasma C6 (Figure 2D). Plasma C1 was primarily enriched for samples taken at the baseline and fast progressors. Plasma C2 was composed of samples taken from all phases of disease, although patients with those without fast progression have diverse mutational landscapes. In Plasma C3, the cell lineage abundance and mutation frequency were higher at baseline and then fell throughout relapse (Figure 2D, Figure 2E). What was more intriguing was that, at the sample level, these clusters were mutually exclusive. According to Figure 2F, plasma C1, plasma C2, and plasma C3 were virtually always drawn from three different patient groups. This illustrates the intra-tumor heterogeneity patterns seen in various clusters and samples.

To further characterize the tumor heterogeneity and examine where the malignant plasma cells came from. We attempted to connect the B cell population with plasma cells and discovered the transition lineage from B cell to plasma cells in the following section, as well as further explain which plasma cluster might have a crucial contribution to patients prognosis.
4.3.3 Delineting B cells to plasma cells transition by trajectory analysis and mutation mapping

Researchers have been searching for early B cell lineage cells that are clonally associated with malignant plasma cells and may indicate the growth and evolution of the tumor. The majority of the B cell compartment's cells (10 out of 15 clusters) were grouped apart from the plasma cells when the B cell and plasma cell compartments were combined for clustering. We have revealed the rest of 5 clusters to be tightly associated with plasma cells from the 2 dimensional UMAP reduction map (Figure 3A, Figure 2A, Figure 3D). In addition to identifying the cell subtypes along the B cell trajectories, our goal is to identify pre-cancerous B cells that may eventually transform into malignant plasma cells using a variety of methods.

We were able to classify B cells into their subtypes using various sets of immune markers except for the clusters overlapped with plasma cells or have markers expressed by plasma cells (Figure 3A). We designated 2 plasma associated clusters as B/immature B switched to plasma because of their relationship to immature B cells, and the other cluster was designated as malignant B switched to plasma because of the higher expression of BCMA. We hypothesized that these three clusters represented potential precancer B cell populations.

We validated our hypothesis by linking somatic mutations to individual cells. We observed the clusters we marked as precancer populations had higher mutation burden than other B cell lineages (Figure 3B and Figure 3C). These clusters were primarily observed in baseline samples (Figure 3C) which also imply that they transit into plasma cells at early disease developmental stages.
Trajectory inferred using monocle3 has shown two trajectories from early pro B cells. We built pseudo-time trajectories based on the transcriptome of each cell (Figure 3D). We found memory B lineage and plasma lineage as the two main branches that originate from pro B cells. Interestingly, it appears that there were two very different sets of early and immature B cells that connecting pro-B cells with either memory B cells or plasma cells. This implies that cell fate that accompany the development of B cells into these two terminal states may already be merging at the pro B stage.

We compared cell lineage abundance to validate our hypothesis that the cell state has been determined at as early as pro B stage (Figure 3E and Figure 3F). We found that the abundances of pro B cells, immature B cells, and pre B cells were enriched in relapse samples when compared to baseline samples. Additionally, in relapse samples, the lineage abundance of memory B cells was reduced and the abundance of B cells that switched to plasma was enriched (Figure 3E and Figure 3F). Together, we seem to discover a pattern whereby the plasma cell trajectory was augmented where patients were headed to a route with rapid progression, whereas the memory B trajectory was decreased throughout disease progression. In summary, the most intriguing implication of our findings is that we might be able to anticipate a patient's prognosis by evaluating the interaction between B cells and plasma cells.
4.3.4 Compositional alterations in lymphoid lineage during MM progression and associated gene sets

Multiple single cell studies have revealed that the progression from MGUS to active MM is associated with T cells and NK cells dysfunction\textsuperscript{163}. To better characterize different T/NK cell lineages, we extracted and reclusted T/NK cells from Figure 1C and captured six CD4+ cell types (CD4+ naive T cells, CD4+ activated T cells, CD4+ IFN responsive T cells, regulatory T cells, and gamma delta T cells, CD4+ cytotoxic T cells), six CD8+ cell types (CD8+ naive T cells, CD8+ effector memory T cells, CD8+ central memory T cells, CD8+ cytotoxic T cells, CD8+ IFN responsive T cells, CD8+ exhausted T cells), NKT cells, and four NK cell types (CMV adaptive-like NK cells, CD56\textsuperscript{bright} NK cells, CD56\textsuperscript{dim} NK cells, and bone marrow resident NK cells) (Figure 4A). To note, we were able to identify rare lymphoid populations such as CD4+ cytotoxic T cells and CD4+ IFN responsive T cells with markers shown in Figure 4B.

To better understand how these different cell subsets interact, we performed regulatory gene networks analysis using a machine-learning based tool Ecotyper\textsuperscript{164}. We identified several discrete cell states even within each cell type described above. By examining the co-association of cell states, Ecotyper identified seven co-associated cell states (ecotypes). We didn’t observe certain ecotypes associated with progression features, sample collection timepoints or treatments (Figure 4D). Interestingly, cell states of CD4+ cytotoxic T cells, CD8+ cytotoxic T cells, BM resident NK cells, and CMV-adaptive NK cells are co-associated in ecotype 1 (E1). Cell states of IFN responsive CD4+ T cells and IFN responsive CD8+ T cells and CD4+ Tregs are co-associated in ecotype 7 (E7).
Zavidij et al. has revealed increases of NK, T, CD16+ cell abundance in diseased bone marrows compared to normal bone marrows\textsuperscript{51}, however, few studies have investigated how lymphoid populations are associated with progression features of MM patients. By subsetting cells from Figure 4A for only baseline samples collected from patients treated by both PI and IMID and coloring cells by progression-free survival (PFS), we observed that CD8+ cytotoxic T cells are enriched for patients with shorter FPS (Figure 4E). Further, CD8+ cytotoxic T cells are statistically significantly more abundant in fast progressor patients who progressed within 18 months compared to those non-progressors who progressed after 4 years (Figure 4F). Differential expression analysis revealed translocator protein (TSPO) and FK binding protein 5 (FKBP5) are upregulated in CD8+ cytotoxic T cells from fast progressors (FP) (Figure 4G). TSPO, the most significantly overexpressed gene, has been revealed to be frequently overexpressed in glioblastoma (GBM) and its expression level has been positively correlated to poor prognosis of patients\textsuperscript{165}. Yet, the role of The translocator protein (TSPO) in MM is under-investigated. FKBP5 encodes a molecular co-chaperone of the glucocorticoid receptor complex and has been revealed to be increased at the mRNA and protein level post-dexamethasone exposure\textsuperscript{166}. It plays an important role in the regulation of multiple signaling pathways and in tumorigenesis and chemoresistance\textsuperscript{167}.

To investigate how lymphoid lineages change from baseline to relapse, we calculated the ratio of observed number of cells to random expectation\textsuperscript{168} followed by paired t tests. With FDR cutoff 0.1, we observed that CD4+ naive T cells, CD8+ naive T cells and gamma delta T cells are enriched in baseline samples. CD4+ regulatory T cells, CD4+ activated T cells, IFN responsive
CD4+ T/CD8+ T cells, CD8+ cytotoxic T cells, bone marrow resident NK cells and CD56bright NK cells are enriched in relapse samples. Interestingly, we observed that Naive T cells including both CD4 and CD8 Naïve T cells are enriched in baseline samples, indicating abundance of Naïve T cells reduced during disease progression. Given the immune-suppressive function of Treg, expansion of Treg could contribute to myeloma escape from the immune system in relapse. The reduction or expansion of populations were then visualized by UMAPs. For example, gamma delta T cells are reduced whereas the immature, CD56bright NK cells expanded from baseline to relapse. (Figure 5B).

Interestingly, CD74 and LTB are upregulated at relapse/progression in all T cell subsets enriched for relapse samples. CD74 has been revealed to be expressed in relapse and refractory and anti-CD74 has shown promising clinical results169. LTB is encoded by lymphotoxin-beta (LT-beta), which is also known as tumor necrosis factor C (TNF-C) protein. Lymphotoxin b receptor (LTβR) overexpression promotes plasma cell accumulation170 and its immune interaction helps tumor growth by inducing angiogenesis171. NEAT1 is highly expressed at relapse/progression in three CD4+ T cell populations enriched for relapse samples, including Tregs, activated CD4+ T cells and IFN responsive CD4+ T cells. It has been reported to promote cell proliferation in MM via activation of the PI3K/AKT pathway172. tumor necrosis factor (TNF), which could induce myeloma cells into cell cycles173, is significantly upregulated at relapse in IFN responsive CD4+T/CD8+T cells and cytotoxic CD8+T cells. Interferon-gamma (IFNG) is overexpressed at relapse in IFN responsive CD4+T and cytotoxic CD8+T cells. In addition, AP-1 transcription factors are related to myeloma cell proliferation, survival and drug resistance174. We observed that
AP-1 complex family members, *JUN* and *FOS*, are upregulated at relapse in cytotoxic CD8+T cells.

The increase in CD56\textsuperscript{bright} NK cells was associated with upregulation at relapse/progression previously associated with poor outcomes including *CD8A\textsuperscript{175}* and *LGALS3* (Galectin-3), and transcription factors associated with depreciated anti-tumor functionality, *NR4A2\textsuperscript{176}* and *HIF1A\textsuperscript{177}* in NK and T cells (Figure 5C). Surprisingly, CD56\textsuperscript{bright} NK cells upregulated *GZMB*, associated with cytotoxic function, while downregulating Granzyme M (*GZMM*), and upregulating *IL18*, a pro-inflammatory cytokine, previously not associated with intrinsic production by NK cells. Due to the large number of samples, we were able to dissect the BM resident NK cell population. At relapse/progression, BM resident NK cells had decreased expression of *MTRNR2L12*, *CCL3* (MIP-1\textalpha{}), *TNFRSF18* (GITR), *CD52*, *CXCR4*, implicated in NK cell trafficking to the BM\textsuperscript{178}, and increased *GZMB* and *FCGR3A* (CD16). Meanwhile, the more mature CD56\textsuperscript{dim} NK cell population expressed genes at Relapse/Progression associated with decreased maturation, including increased *KLRC1* (NKG2A), *SELL* (CD62L), *CD3E*, and *FCGR3A\textsuperscript{179}*

### 4.3.5 Compositional alterations in myeloid lineage during MM progression and associated gene sets

We isolated myeloid cell compartment from Figure 1C and clustered cells to identify cell states. We identified CD14+monocytes, CD16+ monocytes, neutrophils, Dendritic cells (DC), and GMP/CMP. We also captured rare immuno suppressive populations including M2 macrophages and myeloid-derived suppressor cells (MDSCs) (Figure 6A). Among CD14+ monocytes, we identified two understudied populations: myeloid-associated suppressor cells (MAM), a
hyperinflammatory population featured by high expression of IL1B, CCL3, CCL4 (Figure 6B) and IL1R2+ monocytes, featured with high expression of IL1R2, CD163, and FKBP5.

By performing regulatory gene networks analysis using Ecotyper, we identified six co-associated ecotypes. Similar to lymphoid lineages, we didn’t observe certain ecotypes associated with progression features, sample collection timepoints or treatments (Figure 6C). Interestingly, three different cell states of MDSC are co-associated in ecotype 6 (E6). Five distinct cell states of neutrophil and all cell states of GMP/CMP are co-associated in ecotype 2 (E2).

To investigate how population abundance changes from baseline to relapse in myeloid populations, we calculated the ratio of observed number of cells to random expectation. With FDR cutoff 0.1, we observed that abundance of IFN responsive CD14+ monocytes and CD16+ monocytes are reduced in relapse samples (FDR adjusted p value < 0.1, Figure 6D and 6E). Interestingly, deduction of IFN responsive CD14+ monocytes was associated with downregulation of IFN stimulated genes (Figure 6F), including IFI44L, ISG15, LY6E, IFITM1, IFITM3, IFI6 and TYMP. Downregulation of IFI44L, ISG15, LY6E, IFITM1 is also observed in CD16+ monocytes. Since type I IFN signaling is needed to trigger anti-tumor immunity, the downregulation of IFN-stimulated genes at relapse might be related to MM treatment resistance mechanism. Besides, FCN1, a macrophage marker gene, and SOX4, which cooperates with CREB in myeloid transformation, are downregulated at relapse in CD16+ monocytes as well.

To further characterize CD14+/CD16+ monocytes and MDSCs, we reclustered these cells (Figure 6G) and observed that MDSC formed a distinct cluster. Although there is no significant difference
of MDSC abundance based on paired t-test potentially due to limited number of paired samples with limited number of cells, MDSC is highly enriched with relapse samples (enhancement ratio=24.4). Since there are 2 subtypes of MDSCs, including monocytic (M-MDSC) and granulocytic (PMN-MDSC) MDSCs, we wondered which subtype of MDSC is in this population. Based on marker expression, $CD11b$ (ITGAM$^+$), $CD66b$ (CEACAM8$^+$), $CD15$ (FUT4$^+$), HLA-DRA$^-$, CD14-, this MDSC population are Granulocytic MDSCs (Figure 6H). Unfortunately, we didn’t observe the transcriptomic expression of CD33. To further confirm the identity of this rare population, we performed differential expression gene (DEG) analysis. Notably, $ARG1$ is highly specially expressed in the MDSC cluster. Other DEGs include $STAT1$, $STAT3$, $IRF1$, and $IL4R$ with adjusted p values 4.74e-12, 3.54e-54, 1.64e-44, and 2.08e-15, affirming that this cluster is MDSC. In MM, given MDSC is a rare population, current studies with relative small sample size was not able to evaluate its association with disease progression and other features. By leveraging a big dataset, we were able to identify a subtype of MDSC and reveal its enrichment in relapse samples. This observation highlights the immunosuppressive role of MDSC in MM disease relapse.

### 4.3.6 Cellular communications changes between baseline and relapse

Myeloma plasma cells and their surrounding microenvironments are complex communities of cells that could affect immune response and cellular functions. Understanding how cell-cell communication changes along the disease progression could shed light on developing myeloma immunotherapies. Overall, we observed that cell cell interaction profiles between baseline and relapse are similar (Figure 7A and 7B). Myeloid lineages tend to have more interactions while plasma cells, while B cells and naive T cells tend to have less interactions. Erythrocytes and erythroblasts have the least number of interactions with other lineages. Interestingly, the number of predicted cell-cell interactions of IL1R2+ CD14+ monocytes decreased from baseline to relapse.
Moreover, plasma/B cells tend to have more predicted interactions with CD14+ monocytes, CD16+ monocytes, IFN responsive CD14+ monocytes, M2 macrophages and MAM at relapse than at baseline (Figure 7A and 7B). We predicted cell-cell interactions involving cell types enriched in either baseline or relapse samples to identify key interactions characterizing these states. Many cell-cell interactions were predicted in multiple cell types and in both baseline and relapse samples. Top interactions that were identified in baseline samples include MIF on DCs and TNFRSF14 on several T cell subsets, and FFAR2 on monocytes interacting with TNF on multiple CD4 T cell subsets. Top interactions that were identified in relapse samples include CCL3 on IFN responsive CD14+ monocytes interacting with CCR5 on gamma delta T cells, and IL1B on several myeloid cell types interacting with ADRB2 on CD8+ T cells or gamma delta T cells.

4.4 Discussion

In this study, by integrating genomic alterations based on bulk sequencing performed on CD138 positive sorted BMMCs and immune profiling based on scRNA-seq performed on CD138 negative fraction of 361 MM samples, we comprehensively interrogated the landscape of tumor and immune microenvironment during myeloma disease onset and progression. These patients were enrolled in the Multiple Myeloma Research Foundation (MMRF) CoMMpass study (NCT01454297).

With cross-center collaboration coordinated by MMRF, we were able to profile 1,149,344 single cells from 361 samples taken from 263 MM patients. Although scRNA-seq data were obtained
from CD138-negative sorted BMMC, we were able to capture a decent amount of transcriptomic level SDC1 (CD138) positive plasma cells (8%). By mapping mutations detected from WES to scRNA-seq, we illustrated tumor heterogeneity in relation to disease progression. We identified a tumor subpopulation, plasma cluster 1 with 64% of non-driver mutations, enriched for baseline samples. Furthermore, we found that CCND1 mutations are enriched in plasma cluster 16 (39% cells with CCND1 mutations). Given CCND1 mutation is associated with t(11;14), the most common translocation in MM\textsuperscript{182}, this result illustrates that plasma cell population clustering could be driven by genomic alterations.

We have identified the precancer populations and two trajectories in B cell lineages. One is from early pro B to malignant B, another one is from early pro B to memory B. Therefore we hypothesized that the subpopulations of pro B and pre B cells play essential roles in determining the fate of B cell development. We have found that the majority of pre-cancer populations showed high somatic mutation rates in the light chain genes, but its precursors showed low somatic mutation rates, suggesting that variation due to somatic mutations might be expected in the B cell malignant processes. Blimp-1, IRF4, and XBP-1 transcription factors are known to be essential for the differentiation of mature B cells into plasma cells. B lymphocyte-induced maturation protein-1 (Blimp-1) is a critical transcription regulator of plasma cell formation, which mainly functions as a transcription repressor\textsuperscript{183}. Blimp-1 is expressed at low levels during the plasmablast stage and at high levels in mature plasma cells\textsuperscript{184}. Deficiency in Blimp-1 impaired plasma cell differentiation, but did not affect B cell development\textsuperscript{185}. XBP-1 functions as a transcription regulator that is essential for Ig secretion and remodeling of the endoplasmic reticulum in plasma cells\textsuperscript{186}. In addition to Blimp-1 and XBP-1, plasma cell formation requires IRF4. IRF4 is highly expressed in
B cells and plasma cells where it plays essential roles in controlling B cell to plasma cell differentiation and immunoglobulin class switching. Loss of IRF4 leads to impaired antibody production, but overexpression of IRF4 is found in MM patients' derived cells. We were able to observe the expression of Blimp-1 and XBP-1 in the pre cancer populations we have identified. XBP-1 was also expressed in early pro B cells and late pro b cells. We were also able to see the IRF4 expression in the pre cancer populations as well as Immature B cluster1 and Pre B cluster 1. This suggested that the pre-cancer populations might come from Immature B cluster1 and Pre B cluster1. CD79a is present in the cytoplasm of both precursor B-lymphocytes and mature, differentiated plasma cells, and plays a critical role in B-lymphocyte antigen signal transduction and overall B-lymphocyte development and stabilization. Its aberrant loss of expression has been noted in certain samples of plasma cell neoplasms. The pre-cancer populations we found in our study had shown the loss of expression of CD79a. We also found that immature B cluster 1 had shown the loss of expression of CD79a. In summary, we further validated that the pre-cancer populations had higher expression of several plasma cell makers. Those markers were also found in its precursors.

With the large sample size, we were able to dissect cytotoxic CD4 T cells, interferon responsive CD4 T and CD8 T cells and bone marrow resident NK cells population. To examine how the immune population changes along the disease progression, we compared the abundance of cell subsets in longitudinal samples. We observed that CD4+ naive T cells, CD8+ naive T cells, and gamma delta T cells are enriched in baseline samples. CD4+ regulatory T cells, CD4+ activated T cells, IFN responsive CD4+T/CD8+T cells, CD8+ cytotoxic T cells, bone marrow resident NK
cells, and CD56bright NK cells are enriched in relapse samples (Figure 6A). We found that CD74 and LTB are overexpressed at relapse/progression in all T cell subsets enriched for relapse samples (Figure 6C). Moreover, Interferon-gamma (IFNG) is upregulated at relapse in IFN responsive CD4+ T and cytotoxic CD8+T cells. AP-1 complex family members, JUN and FOS, are upregulated at relapse in cytotoxic CD8+T cells. The increase of CD56bright NK cells abundance was associated with upregulation of genes previously associated with poor outcomes and transcription factors associated with depreciated anti-tumor functionality. Surprisingly, CD56bright NK cells upregulated GZMB, associated with cytotoxic function. Furthermore, we are also interested in differences of immune subsets in baseline samples between patients with distinct progression features. Surprisingly, cytotoxic CD8+T cells are enriched for patients who progressed within 18 months compared to those who progressed after 4 years. Contradictory to prior work revealing that long-term survival is associated with proliferative cytotoxic T-cells 20, this observation remains to be further investigated.

In myeloid lineages, we found that abundances of IFN responsive CD14+ monocytes and CD16+ monocytes are reduced in relapse samples (Figure 6D), which is associated with downregulation of IFN stimulated genes, including IFI44L, ISG15, LY6E, and IFITM1 (Figure 6F). This highlights the role of type I IFN response in MM relapse and cancer therapy. To note, by leveraging a big dataset, we were able to identify rare populations, including MDSC and M2 macrophages. We showed that MDSC is enriched in relapse samples, indicating the immunosuppressive role of MDSC in MM disease relapse as revealed in previous research190.
Cellular communication analysis revealed that myeloid lineages tend to have more interactions with other cell subsets while plasma cells, B cells and naive T cells tend to have less interactions in general. CD74 on multiple myeloid cell types interact with MIF/COPA on T cells in both baseline and relapse samples. In addition, we identified several baseline-specific top interactions including MIF on DCs and TNFRSF14 on several T cell subsets, and FFAR2 on monocytes interacting with TNF on multiple CD4 T cell subsets. Relapse-specific interactions include CCL3 on IFN responsive CD14+ monocytes interacting with CCR5 on gamma delta T cells, and IL1B on several myeloid cell types interacting with ADRB2 on CD8+ T cells or gamma delta T cells.

We hope this study deepens our understanding of myeloma disease onset and progression and provides potential novel targets for immunotherapies. In addition to scRNA-seq, future studies could utilize protein-level assays including CyTOF, CITE-seq, CODEX and even single-cell proteomics to validate our findings and aid the development of MM immunotherapies for clinical utility.

4.5 Methods

Ammonium-chloride-potassium (ACK) Lysis of Bone Marrow Aspirates (BMA)

BMA samples obtained from subjects enrolled in the MMRF CoMMpass study (NCT01454297). Any blood clots were removed from BMA samples via passage through 70 mM cell strainer. BMA samples were aliquoted into 5mL aliquots in 50mL conical tubes and 45mL of 22mM-filtered ACK lysing buffer (155mM Ammonium Chloride/10mM Potassium Bicarbonate/0.1mM EDTA/pH7.4) was added to each 5 mL aliquot and the tune gently inverted several times to mix. Tubes were then centrifuged at 400xg for 5 minutes. The supernatant was removed and the cell pellet resuspended
with 5 mL of RPMI-1640 and transferred to a clean tube. All aliquots of ACK-lysed BMA aliquots were combined into 1x 50 mL tube, the volume adjusted to 50 mL with RPMI-1640. The cells were then mixed by gentle inversion and the tube centrifuged at 400xg for 5 minutes. The supernatant was then removed by aspiration. Depending on the size of the BMA cell pellet, the cell pellet resuspended in 1-10 mL of EasySep buffer (Phosphate-buffered saline (PBS) containing 2%FBS (v/v) and 1 mM EDTA (PBS/FCS/EDTA buffer). 25 mL of cell suspension was removed for cell counting.

**Isolation of CD138-positive and CD138-negative cells from BMA**

CD138-negative immune cell mononuclear (CD138-) cells in bone marrow aspirates from subjects enrolled in the MMRF CoMMpass study (NCT NCT01454297) were isolated via negative selection from CD138-positive (CD138+) myeloma cells using the EasySep™ immunomagnetic bead technology (EasySep™ Human CD138 Positive Selection Kit: Stem Cell Technologies) in accordance with the manufacturers protocol. Briefly, 100 x 10^6 cell/mL bone marrow mononuclear cells (MNC) in a sterile 17 x 100 mm (14 mL) tube were gently mixed and incubated with 100 mL/mL CD138 selection antibody cocktail for 15 minutes at room temperature. 50 mL/mL of EasySep magnetic nanoparticles was then added to the cell suspension, gently mixed, and incubated for a further 10 minutes at room temperature. The volume of the cell suspension was then adjusted to 8 mL with phosphate-buffered saline (PBS) containing 2%FBS (v/v) and 1 mM EDTA (PBS/FCS/EDTA buffer) and the cell suspension mixed by gentle pipetting (2-3x). The tube was then placed in the magnetic separator. After 5 minutes incubation at room temperature, the magnet and tube were carefully inverted to pour off the supernatant into a sterile 50 mL conical tube. This supernatant contains the heterogeneous CD138-negative immune cell mononuclear
population (MNC). The tube was then removed from the magnet and an additional 8mL of PBS/FCS/EDTA added, gently mixed, and returned to the magnetic separator. Again, after 5 minutes incubation in the magnetic separator, the tube and magnet were carefully inverted to pour of the supernatant into the 50mL collection tube. This PBS/FCS/EDTA ‘wash’ step was repeated once more resulting in ~24mL suspension of CD138-negative bone marrow MNC cells. CD138-MNC cells were then pelleted by centrifugation at 400xg for 5 minutes and the supernatant removed by aspiration. The CD138-MNC cell pellet was resuspended in freezing medium (90%FBS/10%DMSO) at a concentration of ~8-10x10^6 cells/mL prior to cryogenic storage in liquid nitrogen.

**Processing of BMMC and library prep from MMRF CoMMpass study for scRNA-seq at WUSTL**

WUSTL Cell Thawing: Multiple Myeloma bone marrow mononuclear cells (BMMC) aliquots were thawed in 37°C water bath. Cells were then pelleted by centrifugation at 300g for 5 min and all supernatant was removed. To prepare cells for the Miltenyi Dead Cell Removal Kit, cells were resuspended in 100 uL of beads and incubated at room temperature for 15 minutes. Dead cells were depleted using the autoMACS®Pro Separator. Live cells were pelleted by centrifugation at 450g for 5 minutes. Cells were finally resuspended in ice cold phosphate buffer saline (PBS) and 0.5% BSA and loaded onto the 10x Genomics Chromium Controller and using the Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.3. Utilizing the 10x Genomics Chromium Single Cell 3’v3 Library Kit and Chromium instrument, approximately 15,000 cells were partitioned into nanoliter droplets to achieve single cell resolution for a maximum of 10,000 individual cells per sample. The resulting cDNA was tagged with a common 16nt cell barcode and
10nt Unique Molecular Identifier during the RT reaction. Full length cDNA from poly-A mRNA transcripts was enzymatically fragmented and size selected to optimize the cDNA amplicon size (approximately 400 bp) for library construction (10x Genomics). The concentration of the 10x single cell library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the HiSeq 4000 or NovaSeq 6000 platform (Illumina). 26x98bp (3′v2 libraries) sequence data were generated targeting between 25K—50K read pairs/cell, which provided digital gene expression profiles for each individual cell.

**Cellranger alignment**

For single-cell RNA-seq analysis, the proprietary software tool Cell Ranger (V6.0.1) from 10x Genomics was used for demultiplexing sequence data into FASTQ files, aligning reads to the human genome (GRCh38), and generating gene-by-cell UMI count matrix. The R package Seurat (v4.0) was used for all subsequent analysis

**Empty Droplet removal**

Empty droplets were identified by DropletUtils (version 1.14.2) with FDR <0.001 (0.1%). Only putative cells were kept for downstream analysis.

**Cellbender**
This latest version of CellBender remove-background cleans up count matrices using a principled model of noise generation in scRNA-Seq. Input was “raw_feature_bc_matrix.h5” from cellranger. The false positive rate parameter “fpr” was set to 0.01. For downstream analyses we used the ‘FPR_0.01_filtered.h5’ file.

**Quality control**

We filtered out cells/nuclei with fewer than 1000 nCount, 200 nfeature, or greater than 20% of UMIs mapped to mitochondrial genes.

**Mouse cell removal**

To identify mouse cells, we first merged samples with spike-in mouse cells with reads mapped to the human and mouse combined reference genome. We removed clusters with more than 80% of cells (cluster 17, 100%; cluster 15, 98.88%; cluster 23, 81.68%; cluster 42, 80.18%) having less than 95% reads mapped to the human reference genome and removed 2 samples with 65% cells being mouse cells. Then, we removed these mouse cell barcodes from the all-sample merged object, which derive from reads mapped to the human reference genome.

**Harmony batch effect correction**
Raw counts from Seurat object with mouse cells removed were lognormalized with scale.factor = 10000. Then, data dimensionality was reduced to 25 top principal components by PCA, followed by Harmony for batch correction, where each batch from each site was considered a separate batch.

**Differential expression analysis**

Differential expression analysis was performed using the default test (Wilcoxon Rank Sum test) of function FindMarkers (from the Seurat package) with the specified parameters: min.pct=0.1, logfc.threshold = 0.1, and only.pos = T.

**Cell Identification**

NK cell subsets were identified using the following markers: CD56\textsuperscript{bright} (TCF7, GZMK), CD56\textsuperscript{dim} (FGFBP2, SPON2, FCGR3A, KIR2DL1, KIR3 genes), BM resident NK cells (Up: CXCR6, CD69, EOMES\textsuperscript{hi}, IKZF3\textsuperscript{hi}, CCR5, NKG7; Down: TCF7, TBX21\textsuperscript{lo}, GNLY) (PMID: 30186282, PMID: 27226093), CMV Adaptive-like NK cells (Up: KLRC2; Down: ZBTB16, FCER1G, SH2D1B, SYK) (PMID: 25786176).

**Receptor–ligand interactions**

We used the CellPhoneDB tool\textsuperscript{102} to detect significant pairs of receptor–ligand interactions between cell types. This comparison was done at the sample level using default parameters between tumor and lymphocyte cell types.

**Monocle trajectory analysis**
We used the Monocle3 tool (https://cole-trapnell-lab.github.io/monocle3/) to infer cell-type transition states among acinar, transitional, PanIN and normal ductal populations. Objects and trajectory mapping were obtained by following tutorials outlined by developers (https://cole-trapnell-lab.github.io/monocle3/).
4.6 Figures
A

Bone marrow
CoMMpass study
CD138 sorting
CD138+
RNA-seq
n=297
WES
n=381
WGS
n=379
scRNA-seq
n=361
Gene expression
Fusions
mutations
CNVs
SVs
Immune Profiling

Kallisto
STAR-Fusion, EriScript, Integrate
Somatic Wrapper
GATK4
Manta
Sourat

B

C

Survival probability
NRAS, ZNF208
Pe=0.00138, HR=3.52

Survival probability
HI-4, CSMD3
Pe=0.00329, HR=3.51

D

Hazard ratio

E

F

Percentage of Cells

0 50 75 100
0 1 2 3 4 5
baseline remission relapse

Erythrocyte
NK
CD14+Mono
CD16+Mono
CD2+T
B
Plasma
pDC
Neutrophil
Erythroblast
HSC
cDC
Granulocyte
pDC
pre-B
prepro-B
Fibroblast
NA
Figure 1. Datasets, genomic landscapes and immune profiles.

A. Sample type, technology and analysis tools.

B. Heatmap showing the mutational landscape of multiple myeloma from the MMRF cohort

C. Kaplan meier survival curve showing genes associated with poor survival

D. Hazard Ratio of sex, race and tumor stage(R_ISS) in relation to patient’s survival

E. UMAP showing main cell types in CD138 negative sorted BMMC of 361 MM samples.

F. Stacked bar chart showing percentage of cell types in 361 samples. Samples were ordered by sample collection timepoints, including primary, remission, relapse/progression and other timepoints. For the samples collected at the same timepoint, samples are ordered based on the percentage of plasma cells.
Figure 2. Characteristics of plasma cells and their connection to disease progression.

A. UMAP showing plasma clusters in CD138 negative sorted BMMC of 361 MM samples.

B. Barplot showing mutation frequency across different plasma clusters

C. UMAP showing the mutation landscape differences in plasmas collected at baseline vs relapse and fast progression vs non-fast progression

D. Box plot showing the significant differences in mutation frequencies in plasma cluster from baseline vs relapse samples

E. Box plot showing the differences of plasma cell type enrichment in baseline-relapse paired samples

F. Scatter plot showing the mutual exclusivity of plasma cluster 1-3
Figure 3. The interaction between B cells and plasma, and the main dynamics involved in disease development.

A. UMAP showing B cell subtypes in CD138 negative sorted BMMC of 361 MM samples.
B. Barplot showing mutation frequency across different B cell clusters
C. UMAP showing the mutation differences in B cell population collected at baseline vs relapse and fast progression vs non-fast progression
D. UMAP showing the trajectories of B cell and plasma cell lineages
E. Box plot showing the abundance differences of B cell subtypes in baseline-relapse paired samples
F. Box plot showing the abundance differences of B cell subtypes in fast progression and non-fast progression samples
Figure 4. Overview of T cell and NK cell lineages and characterization of fast progression enriched population.

A. UMAP showing cell types of T cell and NK cell lineages. Each dot is a cell, colored by cell types.
B. Dotplot showing marker gene expression in each lineage. The size of the dot represents the percentage of cells expressing the marker gene, while the color represents the average expression across all cells.

C. Heatmap showing expression of top five differentially expressed genes (DEGs) in random 500 cells in each cell type

D. Heatmap showing cell state abundances across the samples assigned to ecotypes. Rows correspond to the cell states forming ecotypes, while columns correspond to the samples assigned to ecotypes.

E. UMAP showing subset cells from panel A, selected from baseline samples from patients treated by both PI and IMID. Each dot is a cell, colored by progression-free survival (PFS) of patients.

F. Box plot showing abundance of cell subset in each sample in fast progressors (FP, PFS< 18 months) and non-progressors (NP, PFS> 4 years) followed by t test. Abundance of CD8+ cytotoxic T cells in each sample = number of cells in CD8+ cytotoxic T cells / number of total lymphoid cells in each sample

G. Volcano plot showing upregulated and downregulated DEGs of CD8+ cytotoxic T cells in FP compared to that in NP.
Figure 5. Characterization of T cell and NK cell populations associated with MM relapse.

A. Box plots showing the ratio of observed cell numbers to random expectation calculated by chi-square (RO/E), which was used to adjust cell-sampling biases for each patient. y axis: average RO/E across all patients. RO/E > 1 (above the dashed line) indicates enrichment. Paired t test was performed to examine enrichment between baseline and relapse. FDR adjusted p values are shown on top of box plots.

B. UMAPs showing T/NK cells from 48 pairs of baseline samples (left, n=99,036 cells) and relapse samples (right, n=77,713 cells). Cell subsets with statistical significant enrichment in either baseline or relapse are labeled on the UMAP.

C. Violin plots depicting normalized expression of select differentially expressed genes between baseline and relapse/progression within CD56\textsuperscript{bright} NK cells, bone marrow resident NK cells, and CD56\textsuperscript{dim} NK cells. Wilcoxon Rank-Sum Test, adjusted p-value < 0.05 for all.
Figure 6. Overview of myeloid cell lineages and characterization of populations enriched for relapse samples.

A. UMAP showing cell types of myeloid cell lineages. Each dot is a cell, colored by cell types.

B. Dotplot showing marker gene expression in each lineage. The size of the dot represents the percentage of cells expressing the marker gene, while the color represents the average expression across all cells.

C. Heatmap showing cell state abundances across the samples assigned to ecotypes. Rows correspond to the cell states forming ecotypes, while columns correspond to the samples assigned to ecotypes.

D. UMAPs showing myeloid cells from 48 pairs of baseline samples (left, n=18,996 cells) and relapse samples (right, n=29,594 cells).

E. Box plots showing the ratio of observed cell numbers to random expectation calculated by chi-square (RO/E), which was used to adjust cell-sampling biases for each patient. y axis: average RO/E across all patients. RO/E > 1 (above the dashed line) indicates enrichment. Paired t test was performed to examine enrichment between baseline and relapse. FDR adjusted p values are shown on top of box plots.

F. Violin plots depicting normalized expression of selected differentially expressed genes between baseline and relapse/progression within IFN responsive CD14+ monocytes and CD16+ monocytes

G. UMAPs showing CD14+ monocytes, CD16+ monocytes, and MDSCs from 48 pairs of baseline samples (left, n=12,357 cells) and relapse samples (right, n=20,397 cells).

H. Expression of MDSC marker genes on UMAP shown in panel F.
Figure 7. Overview of cell-cell interactions in relation to disease progression.

A. Heatmap showing number of potential ligand-receptor pairs across cell types at baseline.

B. Heatmap showing number of potential ligand-receptor pairs across cell types at relapse.

C. Dot plot showing top predicted cell-cell interactions in baseline samples (downsampled, n = 41,165 cells). Size of dot indicates mean expression of interacting gene partners in their respective cell types and color indicates p-value.
D. Dot plot showing top predicted cell-cell interactions in relapse samples (downsampled, n = 37,037 cells). Size of dot indicates mean expression of interacting gene partners in their respective cell types and color indicates p-value.
Chapter 5: Conclusions and future directions

5.1 Discussion

This dissertation presents a series of advancements in our understanding of MM and its microenvironment by utilizing various cutting edge technologies. Compared with traditional bulk sequencing technologies, single-cell technologies have advantages for evaluating cellular heterogeneity and investigating the evolution of cellular subpopulations from the tumor and microenvironment. Application of single-cell sequencing is particularly advantageous in the case of MM because of its significant heterogeneity and complicated immune microenvironment.

Since scRNA-seq, CyTOF, and CITE-seq enable the identification of cell types, cell states and characterization of cellular heterogeneity at transcriptomic and/or protein levels, understanding the concordance of the measurements among these three modalities is of great interest. In Chapter 2, comparisons between different techniques revealed that the percentages of immune populations measured by scRNA-seq, CyTOF, and CITE-seq are generally concordant, except some variations in T cells, macrophages, and monocytes. Of note, we found CD4+ T/CD8+ T ratios tend to decrease in stage 3 versus stage 2 patients as suggested by all three techniques. This work provides the field with reference data sets and shows more detailed examination of T cells, macrophages, and monocytes is needed when handling single cell sequencing with different modalities.

In the past few years, targeted immunotherapies have emerged as a promising treatment approach by using the immune system's ability to specifically eliminate tumor cells that express certain
proteins. However, a significant challenge in developing these therapies is identifying suitable targets that are highly expressed in tumor cells and have strict tissue specificity. In Chapter 3, we performed an unbiased search for genes with specific expression in plasma and/or B cells using single-cell RNA-sequencing (scRNAseq). We were able to identify 20 candidate genes, which have not yet been studied clinically. Among these, 11 genes have not been previously recognized as targets for therapy. To support our scRNAseq-based findings, we validated the gene expression data by comparing with bulk RNA-sequencing samples and bulk global proteomics data. Finally, target dynamics and heterogeneity were further examined by utilizing both transcriptomic and immuno-imaging methods. Our experimental approach and analytical methods facilitated the discovery of potential therapeutic targets associated with myeloma in an effective manner. This is the first time MM immunotherapy targets were discovered using single-cell techniques. The research work demonstrates a robust and broadly applicable approach to recognizing tumor markers, which can assist in enhancing the development of targeted therapies for cancer.

Prediction of disease progression and relapse of multiple myeloma has been of much clinical interest, as there is a paucity of relevant tools. The revised International Staging System, which utilizes the International Staging System (ISS) with chromosomal abnormalities (CA) to predict relative risk to their survival\textsuperscript{157}, is essentially the only commonly used prognostication method in the clinical setting. However, studies have shown that 1q gain, which was not included in the R-ISS, proved to be independent of poor prognostic factors\textsuperscript{158}. This suggests that more work is necessary for elucidation of factors driving disease progression of multiple myeloma. In Chapter 4, we observed some cell subsets are enriched in relapse samples and we were able to identify rare populations including MDSC and M2 macrophages by leveraging big datasets. We showed that
MDSC is enriched in relapse samples, indicating the immunosuppressive role of MDSC in MM disease relapse as revealed in previous research\textsuperscript{190}. Our studies expand our knowledge of the development and advancement of myeloma disease and offer new possibilities for immunotherapeutic treatments.

5.2 Future directions

5.2.1 Test preclinical efficacy of myeloma targets and explore potential bispecific CAR-T partners

We identified 20 candidate genes for MM clinical study in Chapter 3, out of which 11 have not been previously studied for their therapeutic potential. We were not able to validate these targets due to the scope of the study in Chapter 3. Future studies could evaluate the preclinical efficacy of these targets by in-vitro and in-vivo experiments. For instance, CAR-T cells could be created by isolating human T cells from healthy donors' peripheral blood and using viral transduction to introduce CAR constructs into the cells. Then, activation, proliferation, cytokine secretion, and cytotoxicity of CAR T cells should be examined. If in-vitro experiments show promising results, in vivo activity of CAR-T cells could be further examined in terms of myeloma cell eradication and improvement of survival\textsuperscript{191}. Specifically, the efficacy of the CARs could be tested in a mouse xenograft model using the same control CARs. Tumor burden will be monitored using bioluminescent imaging, and CAR efficacy will be measured by both survival outcomes and tumor mass reduction.
So far, a number of bispecific CAR-T constructs have been developed and evaluated in preclinical models. Although bispecific targeting is not a new concept, there is still room for exploration in terms of selecting the target and designing the construct. Basically, a CAR is made up of single-chain variable fragments (scFV) that bind to antigens, linked to an intracellular region containing co-stimulatory domains that enhance T cell activation. The configuration of scFVs and linkers can greatly affect binding kinetics and anti-tumor efficacy, as can the combination of costimulatory domains used. Therefore, testing different constructs is important.

Currently, BCMA is the most effective target for MM. Single-target anti-BCMA CAR-T could be used as the positive control for their downstream experiments. For the bispecific construct, future studies could use BCMA as one of the targets, and choose the second target using one of several possible strategies. These include selecting a target that is highly correlated with BCMA and expressed on target cells to potentially increase the magnitude of the cytotoxic response; selecting a target that is negatively or un-correlated with BCMA to increase tumor "coverage" in case of antigen escape; or selecting a target that is highly expressed but not specific to myeloma cells, to test whether one highly specific binder can compensate for a promiscuous partner while still benefiting from increased antigen surveillance.

5.2.2 Expand tumor cell associated targets discovery to multiple cancer types

The scRNA-seq driven myeloma specific target discovery pipeline could be widely applied to identify targets in other cancer types. For example, we were able to identify tumor cell markers in clear cell renal cell carcinoma (ccRCC) using the same pipeline in an internal project revealing key markers and pathways in ccRCC tumorigenesis. Compared to the traditional target discovery approach, our high-throughput method allows for systematic search for tumor-specific
genes without relying on prior sorting. This is especially beneficial given 90% of adult human cancers are solid tumors, which makes sorting impossible. Furthermore, this strategy enables direct comparison of cancer cells and non-cancerous cells to identify highly expressed genes in tumor cells. Once pan-cancer tumor-specific markers are identified from scRNA-seq, we can further validate their expression in bulk RNA-seq and proteomic mass spectrometry by TCGA and CPTAC datasets. The result will provide a valuable list of potential targets for researchers to use in the development of immunotherapy and other targeted treatments in the cancer research field.

5.2.3 Incorporate proteomics for characterizing multiple myeloma immune microenvironment

Much of this dissertation revolves around single-cell transcriptomics for characterizing MM immune microenvironment. However, protein level expression is not always concordant with transcriptome level expression profiles and understanding protein-level immune profiles is critical for immunotherapy development. Therefore, MMRF immune atlas teams have subjected 485 MM BMMC samples to CyTOF, 120 of which have scRNA-seq data available (Chapter 4). In Chapter 3, we have shown various cell populations enriched for relapse MM based on single-cell transcriptomics. By integrating scRNA-seq and CyTOF, we could validate population abundance differences between baseline and relapse patients. Besides, we could also examine marker expression at protein level in relation to clinical features if their protein panels are available. However, given the limited number of protein panels in CyTOF, we will not be able to comprehensively characterize protein level expression of genes associated with progression, which calls for other high-throughput proteomics techniques. Single-cell proteomics (scProteomics) offers detailed functional profiles, but studying single cells using conventional proteomics
workflows presents challenges such as non-specific binding of proteins to container surfaces and low protein concentrations, which reduce the efficiency of analyzing proteins by LC-MS. To address these issues, we will collaborate with PNNL which developed the nanoPOTS technology. Compared to other available scProteomics systems, their nanoPOTS-iBASIL platform offers significantly better proteome coverage and quantitation quality, as demonstrated in various studies\textsuperscript{195–197}.

By combining scRNA-seq and scProteomics data, we can create a comprehensive, multi-dimensional view of tumor and immune cells. Our goal is to identify lineage-specific patterns of protein and transcript expression, which could provide new insights into how gene expression is regulated at different levels in various cell types. This is particularly important as we study the communication between immune populations in the multiple myeloma tumor microenvironment since changes in certain pathways may not be apparent from RNA levels alone. We plan to identify and validate protein markers that indicate the progression of multiple myeloma at both the gene and protein levels.

These studies could help discover new targets for MM immunotherapies and to identify patients who are at a high risk of rapid disease progression and poor survival so that they could receive early intervention in the clinical setting.
References


61. Koike, M. *et al.* Relationship between CD4(+)CD8(+) T cell ratio and T cell activation in


72. Krebs, C. F. *et al.* Pathogen-induced tissue-resident memory TH17 (TRM17) cells amplify


83. Franssen, L. E., Mutis, T., Lokhorst, H. M. & van de Donk, N. W. C. J. Immunotherapy in


125. Lu, J. et al. Pim2 is required for maintaining multiple myeloma cell growth through


177. Ni, J. *et al.* Single-Cell RNA Sequencing of Tumor-Infiltrating NK Cells Reveals that
Inhibition of Transcription Factor HIF-1α Unleashes NK Cell Activity. *Immunity* **52**, 1075–1087.e8 (2020).


187. Mittrücker, H. W. *et al.* Requirement for the transcription factor LSIRF/IRF4 for mature B


