Investigating Transcriptional Dysregulation in Acute Myeloid Leukemia to Inform Novel Therapeutic Strategies

---

By

William M Yashar

A Dissertation

Presented to the Department of Biomedical Engineering

of the Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements for the degree

of Doctor of Philosophy

February 2024

Acknowledgments

First and foremost, I want to extend my deepest appreciation to my mentorship team: Drs. Brian Druker, Emek Demir, and Ted Braun. I am grateful for their unwavering support throughout my training, particularly during the early stages of my PhD amidst the COVID-19 pandemic. I am thankful to Brian for emphasizing the importance of remaining focused on clinical issues in my research. I appreciate Emek for imparting invaluable lessons on addressing problems as a computational biologist while remaining flexible as my interests evolved. I am deeply appreciative of Ted's care, time, and dedication to my training. Our extensive discussions about our research have profoundly influenced my perspective as a physician-scientist. Each of them has set exemplary standards for mentorship, which I aspire to emulate in the future.

I want to express my sincere gratitude to my DAC members: Drs. Laura Heiser, Naoki Oshimori, and Andrew Adey. I appreciate Laura for her role as the DAC chair, ensuring my research progress remained on track while providing pivotal guidance to my work. I want to thank Naoki for his thoughtful questions and insights that helped shape my research trajectory. I am thankful to Andrew for the invaluable support provided by him and his lab in single-cell sequencing, along with his insightful perspectives on my DAC. I am fortunate to have such a dedicated DAC and am deeply grateful for their significant contributions to my academic endeavors.

I am sincerely thankful to the members of the Braun, Maxson, and Pathway & Omics labs for their invaluable support during my PhD training, which started amid the COVID-19 pandemic. The efforts of each lab member to integrate me into their groups eased the isolation of that time, for which I am deeply grateful. Throughout my graduate training, I have cherished their steadfast support, critical insights, and camaraderie. I

want to extend special recognition to Brittany Curtiss for her indispensable role in several projects presented in this dissertation. Additionally, I want to acknowledge Joey Estabrook for his critical contributions to the development of Priori as well as his unwavering kindness and guidance in computational biology.

I am deeply appreciative of our small yet rapidly expanding community of MD/PhD trainees at OHSU. I extend my gratitude to the former director of the MD/PhD program, Dr. David Jacoby, for his significant impact on my physician-scientist training. Dr. Jacoby consistently offered thoughtful advice and unwavering support, and I admire his ability to provide kindness and levity, especially during challenging times. I want to express my heartfelt gratitude to the former and current MD/PhD program coordinators, Alexis Young and Amanda Asbrock. Their essential and often thankless work is the backbone of our program, and I truly appreciate their unwavering commitment. I want to especially thank my MD/PhD cohort members, Rick Mathews, Natasha Warikoo, and Ashley Anderson, for six years of enduring friendship. The support and joy they have brought to this journey have been invaluable.

I want to express my heartfelt gratitude to my future in-laws, Drs. Sue Palmiter and Mark Bajorek, for their incredible warmth and support. I will always cherish our Friday pizza and game nights, which have become a symbol of shared joy and connection. Their willingness to celebrate victories and lend a listening ear during challenging times has been a source of comfort and strength. I am grateful for the love and camaraderie they bring to my life, making the journey ahead even more special.

I am deeply grateful to my immediate family—Mom, Dad, Jake, and Drew—for their enduring connection. Despite the physical distance, I am thankful that we remain close through video calls, annual summer visits to Portland during the berry season, and holiday gatherings in sunny San Diego. I want to especially thank my mom who introduced me to a career as a physician-scientist and fostered my passion for research.

My family’s constant love, encouragement, and shared moments have been the pillars of my journey, and I am immensely thankful for our bond that enriches my life.

I want to express my deepest appreciation to my wonderful fiancée, Lynn, whose unwavering love and support have been my guiding lights throughout my training. The MD/PhD journey is undoubtedly challenging, but Lynn's continuous encouragement and understanding make it infinitely easier and more enjoyable. I am profoundly grateful for her willingness to engage with and learn about my work. Additionally, I appreciate her gentle reminders to savor the joys of life beyond research. I am excited and grateful to continue building our life together, knowing that her steadfast presence adds depth and meaning to every endeavor.

List of Abbreviations

Ac-DEVD-AMC	N-Acetyl-Asp-Glu-Val-Asp-7-amino-4-methylcoumarin
AML	Acute myeloid leukemia
AOF2	Amine oxidase domain-containing protein 2
ARACNe	Algorithm for the Reconstruction of Accurate Cellular Networks
ASO	Antisense oligonucleotide
ATAC-seq	Assay for transposase-accessible chromatin with sequencing
AUC	Area under the curve
AUPRC	Area under the precision recall curve
AUROC	Area under the receiver operator curve
BCL2	B-cell lymphoma 2
BENC	Blood enhancer complex
BIDC	Breast invasive ductal carcinoma
Cas9	Clustered regularly interspaced short palindromic repeats-associated protein 9
cDC	Classical dendritic cell
CDK	Cyclin dependent kinase
CHIP	Clonal hematopoiesis of indeterminant potential
ChIP-seq	Chromatin immunoprecipitation with sequencing
CLP	Common lymphoid progenitor
CML	Chronic myelogenous leukemia
CMP	Common myeloid progenitor
CMP LMPP	Common myeloid progenitor and lympho-myeloid primed progenitor
CoREST	Corepressor for element-1-silencing transcription factor
Corr	Corrected

CR	Complete response
CRISPR	Clustered regularly interspaced short palindromic repeats
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
CUT&Tag	Cleavage Under Targets and Tagmentation
DMSO	Dimethyl sulfoxide
Dox	Doxycycline
eGFP	Enhanced green fluorescent protein
ELN	European LeukemiaNet
ENCODE	The Encyclopedia of Deoxyribonucleic acid Elements
FAB	French, American, and British classification system
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
FGSEA	Fast Gene Set Enrichment
FLT3	Fms-like tyrosine kinase 3
FLT3-ITD	Fms-like tyrosine kinase 3 internal tandem duplication
FOXA1	Forkhead box A1
FOXO1	Forkhead box O1
GFI1	Growth factor independent 1
GMP	Granulocyte-monocyte progenitor
GVSA	Gene Set Variation Analysis
H3K27ac	Histone 3 lysine 27 acetylation
H3K27me3	Histone 3 lysine 27 trimethylation
H3K4me1	Histone 3 lysine 4 monomethylation
H3K4me2	Histone 3 lysine 4 dimethylation
H3K4me3	Histone 3 lysine 4 trimethylation
H3K9me1	Histone 3 lysine 9 monomethylation
H3K9me2	Histone 3 lysine 9 dimethylation

HDAC	Histone deacetylase
HSC	Hematopoietic stem cell
KDM1A	Lysine demethylase 1A
LMPP	Lympho-myeloid primed progenitor
LSC	Leukemia stem cell
LSD1	Lysine specific demethylase 1
MACS2	Model-based analysis of ChIP-seq version 2
MAX	MYC-associated protein X
MCL1	Myeloid cell leukemia sequence 1
MDT	Multivariate Decision Trees
MLM	Multivariate Linear Model
NFE2L2	Nuclear factor erythroid-derived 2-like 2
Norm	Normalized
NRF1	Nuclear respiratory factor 1
ORA	Over Representation Analysis
OS	Overall survival
pDC	Plasmacytoid dendritic cell
PRC	Precision recall curve
PROTAC	Proteolysis-targeting chimaera
PS	Phospholipid phosphatidylserine
RNA Pol II	Ribonucleic acid polymerase II
RNA-seq	Ribonucleic acid sequencing
RNAi	Ribonucleic acid interference
ROC	Receiver operator curve
RPB1	Ribonucleic acid polymerase II subunit 1
RPPA	Reverse-phase protein array
RUNX1	Runt-related transcription factor 1

s3ATAC-seq	Symmetrical strand sci assay for transposase-accessible chromatin with sequencing
scATAC-seq	Single-cell assay for transposase-accessible chromatin with sequencing
scRNA-seq	Single-cell ribonucleic acid sequencing
SEACR	Sparse Enrichment Analysis
shRNA	Short hairpin ribonucleic acid
STAT	Signal transducer and activator of transcription
TCGA	The Cancer Genome Atlas
TF	Transcription factor
UDT	Univariate Decision Tree
ULM	Univariate Linear Model
VIPER	Virtual Inference of Protein-activity by Enriched Regulon analysis
WMEAN	Weighted Mean
WSUM	Weighted Sum

Table of Contents

List of Tables

List of Figures

Abstract

Gene transcription is fundamental in establishing and maintaining cell identity and function. All normal cells are encoded with the same DNA, so transcription is essential for regulating many biological processes, including blood cell formation. Dysregulation of the complex transcriptional mechanisms governing blood cell formation can result in the development of hematologic malignancies, including acute myeloid leukemia (AML). AML is an aggressive blood cancer characterized by the rapid outgrowth of immature blood cells. Despite advances in treatment, AML remains an often fatal disease with a 5-year survival rate of less than 30%. This dissertation aims to uncover mechanisms of transcriptional dysregulation in AML, nominating novel molecular targets that can be used to design more effective treatment strategies.

One primary goal of this dissertation was to deepen responses to targeted disruption of a common mutational driver of AML, internal tandem duplication (ITD) of Fms-like tyrosine kinase 3 (FLT3), by simultaneously disrupting an epigenetic regulator associated with leukemogenesis. Clinical evaluation of FLT3 inhibitors has demonstrated higher overall survival and remission rates in patients with refractory FLT3-ITD AML compared to salvage chemotherapy. However, FLT3 inhibitors are rarely curative and only induce a transient decrease in AML cells. Prior work has shown that lysine specific methylase 1 (LSD1) inhibitors enhance kinase inhibitor activity in AML. In our studies, we show that combined LSD1 and FLT3 inhibition induces synergistic cell death in FLT3-ITD AML. To understand the mechanism of the drug combination, we investigated its impact on genome-wide histone modifications using CUT&Tag. Since no computational method had been developed for this application, we created GoPeaks. We found that drug combination disrupted histone modifications at regions associated

with the MYC proto-oncoprotein. To nominate the transcriptional regulators of *MYC* expression in AML, we developed Priori, a computational method that uses literature-supported gene regulatory networks to infer transcription factor activity. Using the findings from Priori as well as genome-wide profiling of accessible chromatin and transcription factor binding, we found that the drug combination disrupts STAT5, LSD1, and GFI1 binding at the MYC blood super-enhancer. We validated these findings in 72 primary AML samples, with nearly every sample demonstrating synergistic responses to the drug combination. Collectively, these studies reveal how an understanding of transcriptional dysregulation in AML can inform novel therapeutic strategies.

The second goal of this dissertation is to understand the transcriptional adaptation to venetoclax, a small molecule inhibitor that disrupts the anti-apoptotic activity of BCL2. Venetoclax, in combination with azacitidine, is the standard of care for patients unfit for intensive chemotherapy. While this strategy results in a remission rate of nearly 70%, the majority of patients ultimately relapse. Frequently, the relapsed disease is composed of differentiated, monocyte-like AML cells. Given the importance of BCL2-inhibitor-based therapeutic strategies in AML, we investigated the differentiation state dynamics of AML cells following BCL2 inhibition. In our studies, we found that BCL2 inhibition promotes the production of differentiated leukemic cells. Single-cell chromatin accessibility sequencing revealed that short-term BCL2 inhibition primes immature AML cells for myeloid differentiation. Long-term BCL2 inhibition in a PDX model of FLT3-ITD/NPM1-mutant AML resulted in a depletion of HSC-like AML cells and outgrowth of progenitor-like cells. These findings support the role of differentiation state plasticity in BLC2 inhibitor resistance.

Chapter 1: Introduction

1.1 Transcriptional control of genes

Gene transcription, the process of creating an RNA copy of a gene's DNA sequence, is a dynamic regulatory mechanism within the cell. Although proteins execute most cellular molecular functions, the blueprints of proteins are encoded in genes. As a result, transcription plays a vital upstream role in regulating protein expression. Understanding transcription illuminates fundamental mechanisms governing cellular behavior and function.

DNA is stored in the nucleus as chromatin, a complex of DNA tightly wound around histone proteins (Figure 1.1). Compact DNA-histone complexes limit the access of transcriptional machinery to DNA and are associated with low gene transcription[1]. Several mechanisms, including post-translational modification of histone proteins, loosen these DNA-histone complexes and facilitate binding of transcriptional regulators, including RNA Polymerases and transcription factors. These regulators modulate gene expression at regions of accessible DNA by binding to gene regulatory elements like promoters and enhancers. The intricate interplay between chromatin structure and chromatin binding proteins regulates cellular gene expression patterns.

1.2 Deciphering transcriptional regulation with next-generation sequencing

Recent next-generation sequencing advances have enabled the high-throughput investigation of transcriptional regulatory mechanisms. These techniques have been

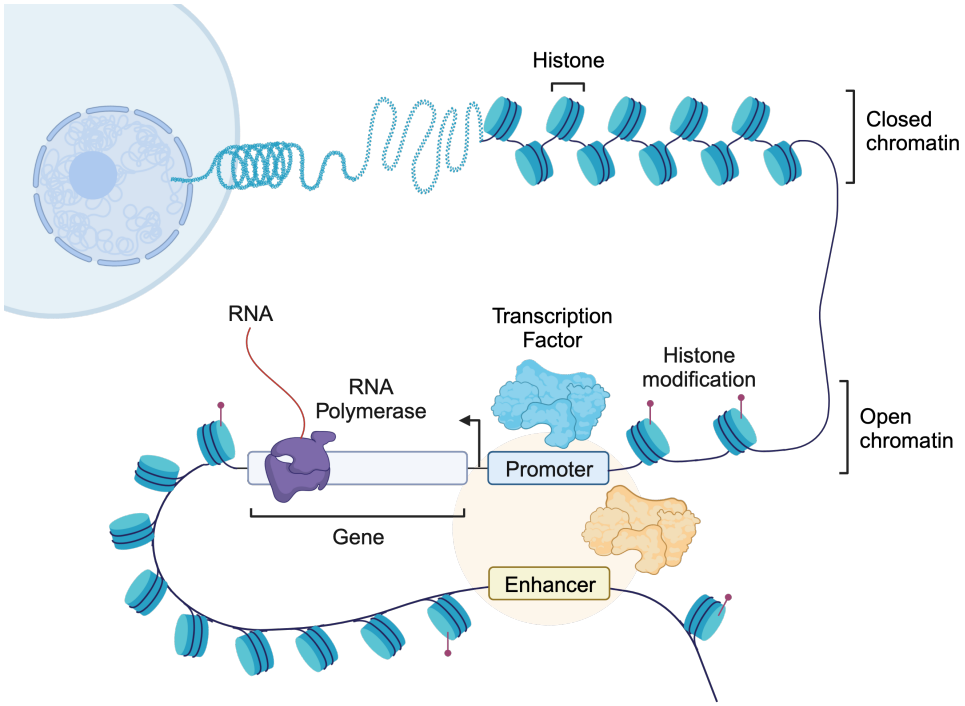

Figure 1.1: The structure of genes regulates its expression.
This figure was created with BioRender.com.

developed for multiple modalities, allowing for an understanding of epigenetic regulation at the level of chromatin but also its downstream effects on gene expression. Sequencing methods, such as assay for transposase-accessible chromatin with sequencing (ATAC-seq), have been developed to identify regions of open chromatin that are accessible for transcription. Other techniques, like chromatin immunoprecipitation with sequencing (ChIP-seq), enable genome-wide profiling of chromatin binding proteins and chemical modifications of histone proteins. The downstream effects of chromatin regulation can be interrogated by measuring the abundance of RNA transcripts using RNA sequencing (RNA-seq). In this sub-section, I will review next-generation sequencing technologies to profile transcriptional regulation and computational methods to analyze them.

1.2.1 Profiling genome-wide chromatin accessibility

Sequencing techniques such as ATAC-seq enable the high-throughput characterization of accessible chromatin[2,3]. By employing a hyperactive Tn5 transposase, ATAC-seq selectively cleaves open regions of DNA and adds oligonucleotide adaptors for subsequent sequencing[4]. This method provides researchers with valuable insights into the dynamic interplay between open and closed chromatin regions.

Conventional ATAC-seq characterizes chromatin accessibility of pooled cell populations. While these “bulk” sequencing techniques enable a high-throughput understanding of cellular chromatin accessibility profiles, many tissues are highly heterogeneous. Moreover, without pre-sequencing enrichment techniques, it is exceedingly difficult to decipher the chromatin accessibility profiles of rare populations. Pioneering work by Eberwine et al. and Isocove et al. has enabled the sequencing of many molecular modalities at a single-cell resolution, including chromatin accessibility (scATAC-seq)[5,6]. The most popular single-cell technology provider, 10x Genomics, uses microfluidic partitioning and encapsulation to capture single cells and prepare barcoded, next-generation sequencing cDNA libraries. These advancements allow for a deeper understanding of rare cellular populations and provide a more nuanced investigation of chromatin dynamics at the individual cell level.

While 10x Genomics has revolutionized single-cell genomics with its innovative technology, there are key disadvantages associated with its platforms. The reagents and equipment required for 10x Genomics experiments are expensive, often costing thousands of dollars per reaction. Moreover, the 10x Genomics and other single-cell technologies often suffer from high levels of read dropout and data sparsity[7]. In our lab, we have deployed an adapter-switching strategy, symmetrical strand sci assay for

transposase-accessible chromatin with sequencing (s3ATAC-seq), capable of producing six- to 13-fold improvement in usable reads per cell[8]. s3ATAC-seq leverages sequencing adaptors to uniquely identify reads from individual cells. Unlike other sci assays, however, the s3 transposome complex contains a uracil base deposited before the sequencing barcode. Using a uracil-intolerant polymerase prevents extension into the barcode, ensuring that forward and reverse primer sequences are incorporated at each end of the molecule[8]. s3ATAC-seq is a scATAC-seq technology that allows for high-fidelity, single-cell, genome-wide profiling of chromatin accessibility at a reasonable cost.

1.2.2 Identifying genomic regions bound by chromatin-interacting proteins

Modification of histone proteins is a key regulator of gene transcription (Figure 1.1)[9]. Histones package eukaryotic DNA into chromatin and control DNA conformation and organization[10]. Post-translational chemical modification histone proteins alters the chromatin structure, enabling the recruitment of nuclear proteins to genomic features[11,12]. For example, trimethylation of histone three lysine 4 (H3K4me3) facilitates the binding of positive transcriptional regulators to transcription start sites[13–15]. Similarly, H3K27ac neutralizes the positive charge of the histone tail and loosens the interaction between nucleosomes and DNA, allowing access of transcription factors to DNA regulatory sequences[16,17].

ChIP-seq, which couples antibodies that recognize DNA-associated proteins with next-generation sequencing technology, has enabled genome-wide profiling of histone modifications and transcription factors[13,18–20]. Although widely used for epigenetic profiling, ChIP-seq is prone to high background, artificial enrichment of highly expressed genes, and often requires a prohibitively large number of cells per experiment[21,22].

Enzyme-tethering strategies, including Cleavage Under Targets and Tagmentation (CUT&Tag) and Cleavage Under Targets and Release Using Nuclease (CUT&RUN), have been developed to overcome these issues and perform epigenetic profiling with a low number of cells and minimal background[23–26].

Epigenetic studies require mapping multiple histone modifications to comprehensively understand transcriptional regulation. Detecting regions bound to H3K4me1 or H3K4me3 aids with the identification of promoters and enhancers, respectively, throughout the genome[13,27–30]. The co-localization of H3K4me1 and H3K4me3 with H3K27ac is characteristic of activated genomic features[31,32]. Histone modifications can also be associated with heterochromatic regions of the genome. Localization of trimethylation of histone three lysine 27 (H3K27me3) is associated with silenced gene bodies and promoters[13,33–36]. Regions of modified histones in ChIP-seq and CUT&Tag are identified as stacks of aligned reads; such regions are called peaks. The peak profiles of common histone modifications are highly variable, so algorithms that identify histone modification peaks need to robustly detect a range of peak profiles. While H3K4me3 peaks tend to be sharply localized, H3K4me1 and H3K27me3 peaks span a wider domain[34–38]. Moreover, H3K27ac can mark large domains such as super-enhancers and smaller, discrete regions such as promoters, thus having both wide and narrow characteristics[29,36–40]. To extract meaning for epigenetic studies reliant on histone modification CUT&Tag datasets, peak calling algorithms need to be flexible to identify narrow and broad peak characteristics.

Peak calling algorithms have been developed to identify genome-wide enrichment of aligned reads and distinguish peaks of modified histones from noise and artifacts. Model-based Analysis of ChIP-seq version 2 (MACS2), a widely-used peak calling algorithm for ChIP-seq, and other ChIP-seq peak calling methods are designed to address the high rate of background in ChIP-seq. As a result, they are vulnerable to

mistaking background signal as peaks particularly when the background is low[41–43]. Sparse Enrichment Analysis (SEACR) has been developed to perform peak calling from CUT&RUN data that, like CUT&Tag data, is characterized by low background. However, no peak calling algorithms have been designed to address the low background and peak profile variability characteristic of histone modification CUT&Tag data[43]. These problems form the basis of our studies, detailed in Chapter 2.

1.2.3 Predicting transcription factor activity from gene expression data

The coordinated expression and activity of transcription factors are fundamental mechanisms in establishing and maintaining cell identity and function (Figure 1.2A). Transcription factors are epigenetic regulatory proteins that bind to *cis*-regulatory DNA sequences, including promoters and enhancers, and modulate gene transcription[44,45]. Dysregulation of these normal transcription factor functions frequently contributes to the development of a pathogenic cell phenotype (Figure 1.2B)[46,47]. Abnormal transcription factor activity can result from mutations in the putative DNA binding sequences or the transcription factors themselves. Recent studies have highlighted the importance of aberrant expression of pathogenetic transcription factors as drivers of disease[47]. *MYC*, which is essential for cellular growth and proliferation, is the most frequently amplified oncogene. Elevated levels of *MYC* have been shown to promote tumorigenesis in various tissue types[47]. In tumor cells expressing high levels of *MYC*, the transcription factor accumulates in *cis*-regulatory regions of genes associated with cellular proliferation and growth, resulting in transcriptional amplification of MYC's gene regulatory network and, subsequently, abnormal cellular proliferation[48,49]. Therefore,

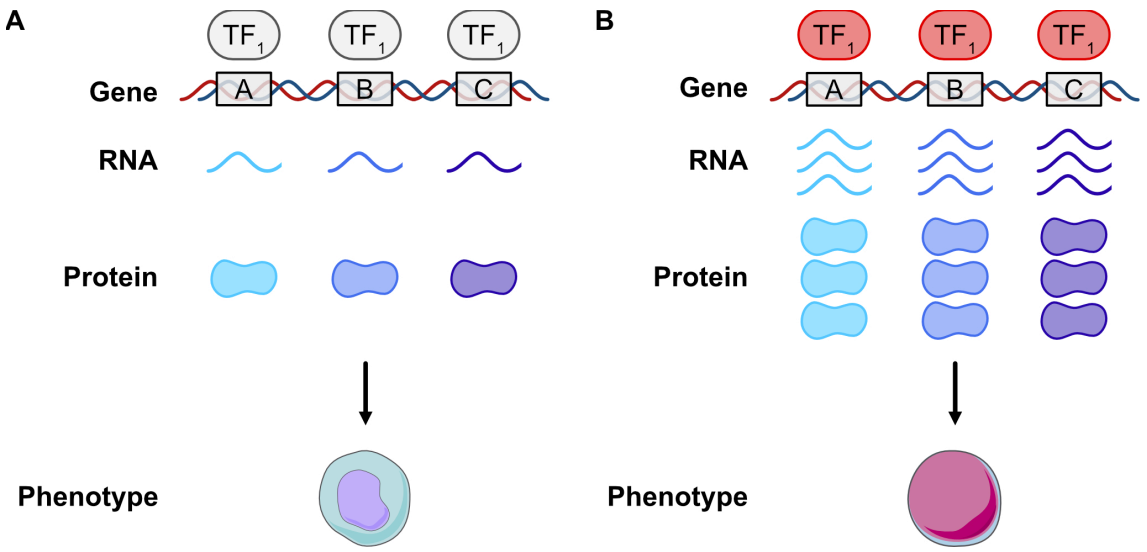

Figure 1.2: Transcription factors are often central nodes of disease

This figure was made using Servier Medical Art.

detecting abnormal transcription factor activity is valuable for better understanding the mechanisms underlying disease pathogenesis.

Gene expression profiling, including RNA sequencing (RNA-seq), is commonly used to monitor dynamic changes in transcription factors and their gene regulatory networks. Initial studies to infer transcription factor activity only used transcription factor gene expression as a proxy for activity[50–52]. However, this approach has several shortcomings. Gene expression is only an indirect measurement of protein activity due to the complex mechanisms controlling protein synthesis and degradation[53–55]. Feedback loops may alter the expression of transcription factors in response to their regulatory activity[47,56–58]. Therefore, reliable predictions of transcription factor activity cannot be limited to evaluating transcription factor expression alone.

An alternative approach to inferring transcription factor activity is to assess the expression of their downstream target genes[50–52]. This approach has two major benefits. First, evaluating hundreds or thousands of downstream targets instead of a single transcription factor likely improves prediction robustness. While some of these targets

may be context-specific, analyzing them in aggregate likely improves the prediction generalizability across many contexts. Second, as target gene expression is downstream of transcription factor control, these signatures are expected to reflect the actual transcriptional impact more accurately. Therefore, accounting for the downstream impact of transcription factors on its gene regulatory networks is essential for activity inference.

Multiple methods have been developed to quantify transcription factor activity from gene expression data. These approaches can be grouped based on how they select gene expression features (Figure 1.3). Methods like Univariate Linear Model (ULM) and Multivariate Linear Model (MLM) use every gene in a dataset, nominating transcription factors as a covariate that best estimates the expression of all other genes[59]. However, these methods develop activity signatures using genes that may not have a genuine biological relationship to the transcription factor of interest. Gene set approaches like Over Representation Analysis (ORA), Fast Gene Set Enrichment (FGSEA), Gene Set Variation Analysis (GVSA), and AUCell infer activity using sets of

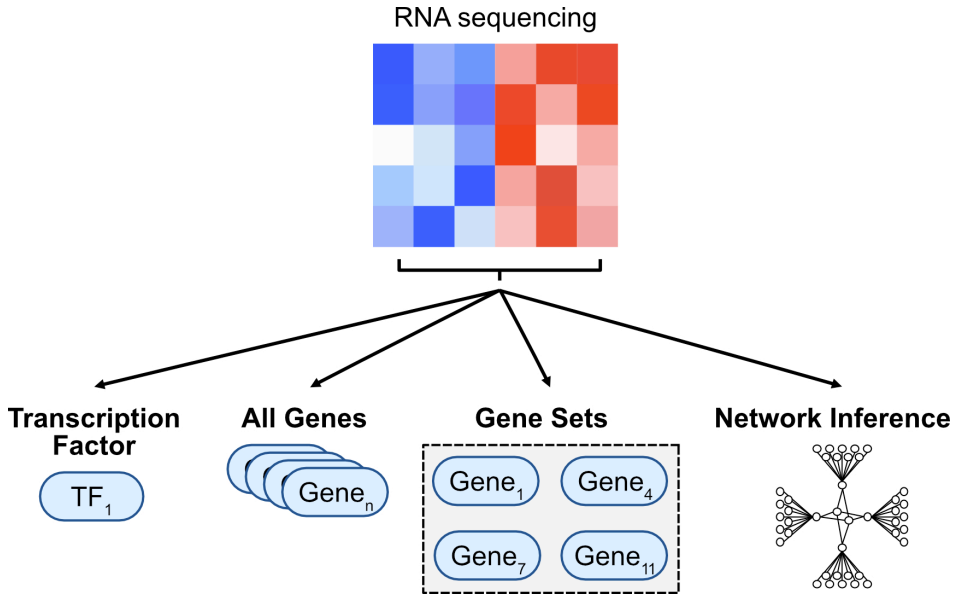

Figure 1.3: Transcription factor activity methods can be grouped based on how they select features.

published transcription factor target genes or target genes curated by experts[60–63]. While gene set methodologies are simple and popular, they are susceptible to the quality and comparability of gene set signatures[64]. Network inference approaches, including Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe), infer gene regulatory networks based on the covariance of transcription factors and their putative targets[65,66]. However, this process is not entirely unsupervised, as ARACNe requires a user-defined list of transcription factors to infer gene regulatory networks[65,66]. The same group that developed ARACNe also created Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) to infer transcription factor activity from ARACNe gene expression signatures[67]. The challenge with these approaches is deconvoluting combinatorial regulation, where multiple transcription factors control the expression of a target gene. While some of these methods, including VIPER, have an option to correct this, it remains challenging to infer transcriptional networks as many possible solutions can explain the underlying data[50,68]. While these methods deploy various techniques to generate activity scores using the expression of downstream target genes, most do not select their target gene features from literature-supported transcriptional relationships.

Recent studies have highlighted that grounding predictions using transcription factor activity methods remain challenging[68–72]. A rigorous evaluation of widely used transcription factor algorithms demonstrated that most methods do not robustly detect perturbed transcription factors[63]. Despite this, there are precision medicine clinical studies using inferred transcription activity from bulk RNA-seq as a marker to guide clinical decisions. While there is an increasing number of single-cell and spatial -omic modalities available to clinical researchers, these studies and many larger cohorts and clinical trials most commonly use markers identified from bulk sequencing of RNA or DNA. Therefore, it is critical to develop methods that can robustly detect aberrant

transcription factor activity from primary patient bulk RNA-seq data. These questions informed our studies detailed in Chapter 3.

1.3 Transcriptional control of hematopoiesis

Next-generation sequencing technologies have helped investigate transcriptional mechanisms that govern biological processes, including blood cell formation. Hematopoiesis, the process of blood cell formation, must balance enormous production needs (the average person produces more than 500 billion blood cells daily) while simultaneously regulating the equilibrium of blood cell types in circulation[73]. Most hematopoiesis occurs in the bone marrow and is derived from a limited number of hematopoietic stem cells (HSCs). HSCs are multipotent progenitor cells that are capable of extensive self-renewal. However, HSCs also divide asymmetrically to produce unipotent progenitor cells, which can create cells committed to one of two hematopoietic lineages: the lymphoid and myeloid lineages. Lymphoid cells, including B, T, and natural killer cells, are involved in innate and adaptive immunity. While some myeloid cells are also involved in immunity, including granulocytes, monocytes, and dendritic cells, myeloid cells also have functions relating to tissue oxygenation (erythrocytes) and blood coagulation (thrombocytes). Hematopoiesis is a critical function of the body that maintains the production and differentiation of lymphoid and myeloid cells.

The coordinated action of transcription factors orchestrates hematopoiesis. Previous studies have identified transcription factors that regulate HSC formation and function, such as RUNX1 and GFI1, or myeloid lineage determination, including PU.1 and GATA1 (Figure 1.4)[74]. Lineage-determining factors promote the differentiation of their lineages and counteract the actions of other factors. For instance, GATA1 and PU.1 drive erythroid/megakaryocytic and myeloid differentiation, respectively[75,76]. Moreover,

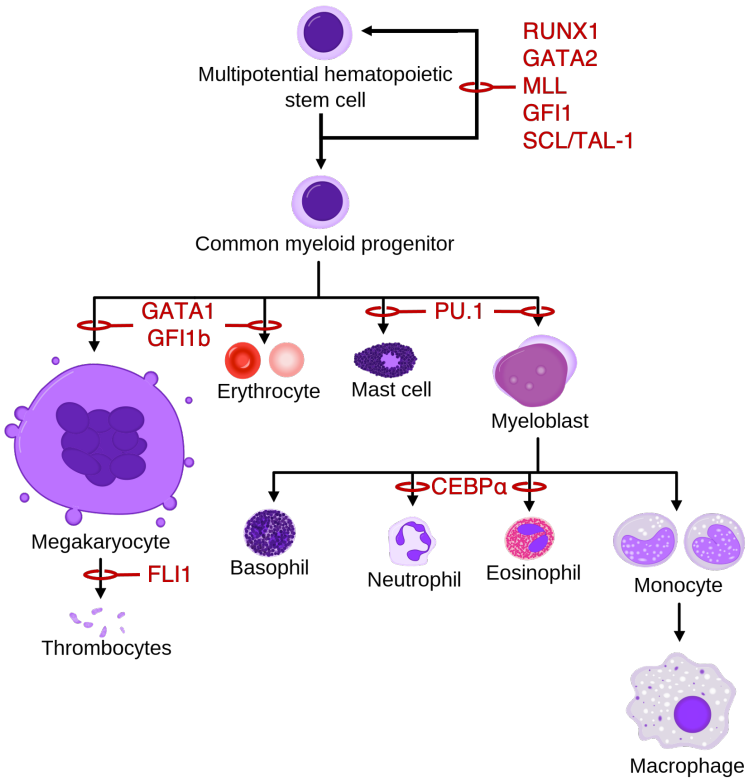

Figure 1.4: Interplay of transcription factors that regulate HSC formation or function or regulate myeloid lineage determination.
This figure was adapted from A. Rad and M. Häggström and Orkin and Zon Cell (2008)[74].

these proteins physically interact and inhibit each other's functions. Downregulating *GATA1* expression biases hematopoietic progenitors toward a myeloid fate, whereas inhibiting *PU.1* expression elicits the opposite effect. Myeloid progenitors can even be programmed by forced expression of lineage-determining factors. Introducing *GATA1* into granulocyte-monocyte progenitors (GMPs) redirects their commitment toward megakaryocytic/erythroid progenitors[77]. Transcription factors are important for modulating gene expression to ensure the proper balance of blood cell types and functions.

1.4 AML: a disease of dysregulated myeloid hematopoiesis

Dysregulation of normal myeloid hematopoiesis is associated with a continuum of myeloid malignancies. Although genetic aberrations, particularly cytogenetic abnormalities and oncogene mutations, play an integral role in the development of myeloid malignancies, transcriptional processes regulating hematopoiesis are also key drivers of their pathophysiology[78]. Alterations in DNA methylation, histone acetylation, histone methylation, and transcription factor activity, often occurring in HSCs, can increase the proliferation and/or impair the differentiation of myeloid cells[79,80]. One such condition, acute myeloid leukemia (AML), is characterized by the uncontrolled proliferation of differentiation-arrested myeloid cells[81]. These malignant blood cells, termed blasts, rapidly accumulate in the bone marrow and halt the production of normal blood cells. The disease process of AML is often rapid, aggressive, and deadly, with only one in three patients expected to survive five years after diagnosis[82]. AML is an often fatal myeloid malignancy that is linked to the disruption of normal myeloid hematopoiesis.

While AML is thought to arise from differentiation-arrested hematopoietic stem and progenitor cells, the bulk disease does not always resemble these populations. In the 1970s, French, American, and British leukemia experts grouped AML into subtypes based on their cellular morphology (Table 1.1)[83]. According to the French, American, and British (FAB) classification system, AML blasts with morphological characteristics similar to hematopoietic stem and progenitor cells are assigned scores of M0 or M1. In contrast, AML blasts resembling differentiated myeloid cells, like monocytes, are assigned M4, M4 eos, or M5 scores. The advent of single-cell RNA sequencing has more deeply characterized the differentiation state heterogeneity reflected in AML blast

morphology. A seminal study led by Dr. van Galen revealed substantial inter- and intra-patient heterogeneity in the AML blast differentiation state[84]. They identified that AML blasts have transcriptional characteristics of hematopoietic stem and progenitor cells and differentiated myeloid cells, including monocytes. Although AML likely arises from malignant stem and progenitor cells, the differentiation state of the bulk disease is highly heterogeneous.

Table 1.1: The FAB classification system groups AMLs on their morphological resemblance to normal blood cells.

FAB subtype	Name
M0	Undifferentiated acute myeloblastic leukemia
M1	Acute myeloblastic leukemia with minimal maturation
M2	Acute myeloblastic leukemia with maturation
M3	Acute promyelocytic leukemia
M4	Acute myelomonocytic leukemia
M4 eos	Acute myelomonocytic leukemia with eosinophilia
M5	Acute monocytic leukemia
M6	Acute erythroid leukemia
M7	Acute megakaryoblastic leukemia

1.5 Targeting the genetic drivers of AML

Controlling and, when possible, eradicating AML has primarily relied on chemotherapeutic regimens. Anthracyclines and cytarabine, small molecules that poison Topoisomerase II and intercalate with DNA, remain the backbone of intensive chemotherapy[85]. The 2022 European LeukemiaNet (ELN) guidelines recommend that newly diagnosed patients undergo a cycle of “7+3” induction chemotherapy. This

regimen consists of IV daunorubicin, idarubicin, or mitoxantrone for days 1-3 and continuous IV cytarabine for days 1-7[85]. After complete response (CR) is attained, the guidelines recommend consolidating patients with 3-4 cycles of intermediate-dose cytarabine[86]. While these regimens have demonstrated improved disease management, they are often associated with substantial adverse events. In a survey of 1,182 patients with AML, 89% of respondents reported that they had experienced short-term (nausea/vomiting, diarrhea, hair loss, mouth sores, infection, rash) and long-term (organ dysfunction, chemobrain, fatigue, neuropathy) effects while undergoing chemotherapy[87]. The highest likelihood of achieving long-lasting remission in AML is an allogeneic stem cell or bone marrow transplant, which produces cure rates of up to 60% in patients with intermediate-risk AML[88]. However, the intensive chemoradiotherapy conditioning and subsequent immunosuppressive treatment required for allogeneic transplant limits its use[85,89]. There is a severe unmet need for treatments that improve survival while limiting substantial toxicities in AML.

Treatment regimens may be tailored to the genetic profile of the bulk disease. Advances in DNA sequencing technologies have illuminated recurrent mutations associated with AML pathogenesis[89–91]. These mutations range from amplifications, deletions, rearrangements, and point mutations in proteins associated with various functional classes (Table 1.2). Notably, many of these genetic aberrations carry prognostic implications, independently or in cooperation with co-occurring mutations. For example, the 2022 ELN guidelines classified *NPM1*-mutant AML as favorable risk. However, the presence of internal tandem duplications (ITD) in Fms-like tyrosine kinase 3 (FLT3) with mutant *NPM1* is classified as intermediate risk[85]. Efforts to investigate these genetic drivers deepen our understanding of AML and pave the way for targeted therapeutic strategies.

Table 1.2: Common AML mutations organized by functional class

Functional class	**Specific example mutations**
Signaling and kinase pathway	*FLT3, KIT, KRAS, NF1, NRAS,* and *PTPN11*
Nucleophosmin	*NPM1*
Transcription factors	*CEBPA, GATA2,* and *RUNX1*
Tumor suppressors	*TP53*
Spliceosome complex	*SF3B1, SRSF2, U2AF1,* and *ZRSR2*
Cohesin complex	*RAD21, SMC1A, SMC3,* and *STAG1/2*
Epigenetic modifiers	*ASXL1, DNMT3A, EZH2, IDH1/2, MLL/KMT2A,* and *TET2*

Mutations in *FLT3* are the most common in AML, occurring in approximately 30% of all cases[92,93]. FLT3 is a transmembrane receptor tyrosine kinase expressed by hematopoietic stem and progenitor cells. FLT3 receptor binding to its extracellular ligand induces receptor dimerization and autophosphorylation, activating downstream signaling pathways associated with cell survival, differentiation, and proliferation (Figure 1.5)[94]. Mutations in the *FLT3* receptor commonly occur in one of two locations: near the activation loop or in the tyrosine kinase domain (TKD). Mutations near the activation loop, such as FLT3-ITD, allow the receptor to activate downstream FLT3 signaling without receptor dimerization. TKD mutations, on the other hand, facilitate the dimerization and autophosphorylation of FLT3 receptors bound to FLT3 ligands. Both sets of mutations result in constitutive activation of the FLT3 signaling pathway, resulting in abnormal proliferation and differentiation[92]. Mutations in common hematopoietic signaling pathways, including *FLT3*, are genetic drivers of AML.

Several small molecule inhibitors have been developed to target mutant *FLT3*, two of which are FDA-approved for AML[92,95]. FLT3 inhibitors disrupt mutant *FLT3* activity

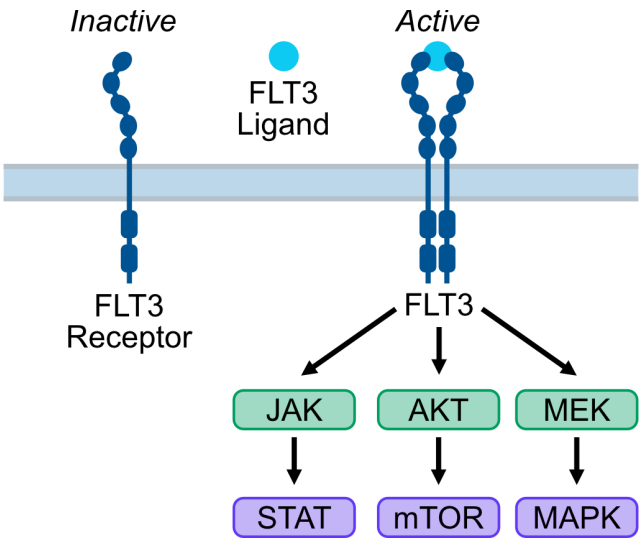

Figure 1.5: FLT3 dimerization results in the activation of downstream signaling pathways.

in two ways[92] (Figure 1.6). Type II FLT3 inhibitors bind the FLT3 receptor in the inactive conformation to a region adjacent to the ATP-binding domain. Quizartinib, a Type II FLT3 inhibitor, was highly active in phase 2 and 3 trials[96,97]. In the randomized, double-blind, phase 3 study QuANTUM-First, quizartinib was evaluated with standard cytarabine and anthracycline induction and as maintenance monotherapy for treating adult patients with newly diagnosed FLT3-ITD AML. Treatment with quizartinib resulted in CR rates of 55% and significantly improved overall survival (OS) compared to the placebo arm[98]. Type I FLT3 inhibitors, on the other hand, bind the FLT3 receptor in the active conformation, either near the activation loop or the ATP-binding pocket. As a result, Type I inhibitors such as gilteritinib are active against FLT3-ITD and -TKD mutations. The randomized, phase 3 clinical trial ADMIRAL showed that gilteritinib monotherapy in patients with relapsed/refractory FLT3-ITD AML results in a CR rate of 34% and a significantly increased OS as compared to salvage chemotherapy[99]. Collectively, these studies show that therapeutic targeting of FLT3 improves survival in patients with *FLT3*-mutant AML.

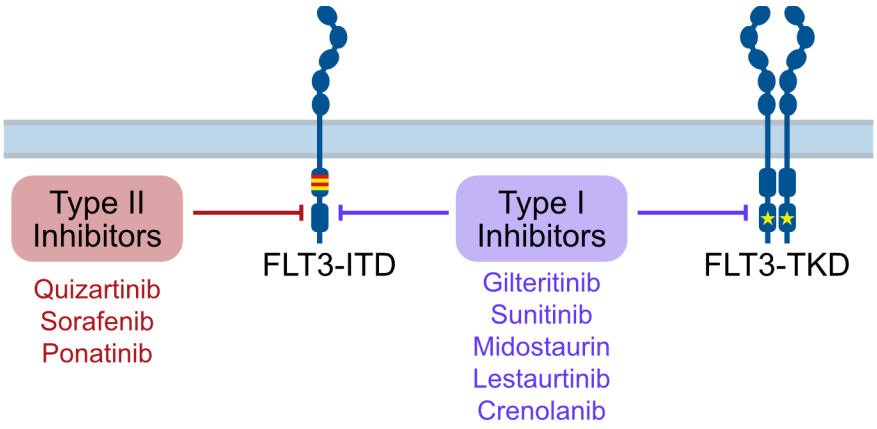

Figure 1.6: FLT3 inhibitors are activate against ITD and TKD mutations.
This figure was adapted from Daver et al. Leukemia (2019)[92].

While FLT3 inhibitors have demonstrated substantial clinical benefits, responses are short-lived. For example, findings from the ADMIRAL phase 3 clinical trial showed that gilteritinib monotherapy only increases OS by 3.7 months[99]. There have been many proposed inherent and acquired mechanisms attributed to FLT3 inhibitor resistance[92]. The persistence of wildtype FLT3 ligand, which is relatively resistant to FLT3 inhibition, activates downstream signaling and provides survival signals to leukemic blasts[100,101]. Other inherent factors, including the bone microenvironment and insufficient drug concentrations in the plasma, are also associated with FLT3 inhibitor resistance[102]. Leukemic blasts may also confer drug resistance by secondary resistance mechanisms. Research has identified various *FLT3* receptor mutations that arise in the presence of and confer resistance to FLT3 inhibitor treatment[103]. Moreover, leukemic blasts may evade FLT3 inhibition by activating parallel signaling pathway mutations, including PI3K/AKT or RAS/MEK/MAPK pathways[104]. While FLT3 inhibitors are effective, there is an unmet need to improve their effectiveness in *FLT3*-mutant AML.

1.6 Transcriptional control of *MYC,* a downstream mediator of aberrant FLT3 signaling

A proposed strategy to deepen responses to FLT3 inhibitors is to target downstream mediators of mutant *FLT3*. Aberrant FLT3-ITD signaling constitutively activates regulators of cellular proliferation, including the MYC proto-oncoprotein[105–107]. MYC coordinates gene expression by recruiting transcriptional regulatory proteins including histone deacetylates to enhancer box sequences and by stimulating the release of RNA polymerase II (RNA Pol II) from its paused state[108]. Through these actions, MYC directly and indirectly regulates the expression of thousands of genes involved in a vast range of biological events, including proliferation, differentiation, survival, programmed cell death, and immune regulation[109,110]. However, MYC is activated in most cancers, and its overexpression has been causally linked to tumorigenesis[111–113]. In the context of FLT3-ITD AML, previous work has revealed an enrichment of MYC-related genes in FLT3-ITD AML cells as compared to FLT3-WT AML, indicating that aberrant FLT3 signaling activates MYC[105]. Moreover, it has been shown that downregulation of MYC activity is required for FLT3 inhibitor efficacy, and reactivation of MYC-controlled oncogenic networks promotes FLT3 inhibitor resistance[114–116]. Strategies that target downstream effectors of FLT3-ITD signaling, including MYC, will likely improve responses to FLT3 inhibitors.

Direct inhibition of MYC has been an objective of anti-cancer therapeutic development for nearly the last three decades[117]. However, MYC has been considered undruggable due to its intrinsically disordered nature with no binding pockets or specific enzymatic activity[117,118]. Nevertheless, various strategies to inhibit MYC, directly and indirectly, are being explored[110]. Most direct approaches do not target MYC itself but rather its interface with MYC-associated protein X (MAX). MYC recruits transcription

factor binding to target gene promoters through heterodimerization with MAX[119]. Small molecules have been developed to inhibit MYC–MAX dimerization, sequester MAX via homodimer stabilization, or interfere with MYC–MAX binding to target DNA sequences[110]. However, most strategies to disrupt MYC activity have focused on indirect approaches, such as interfering with the expression of MYC at the DNA, RNA, or protein levels. Two strategies that have gained prominence are proteolysis-targeting chimaeras (PROTACs), which target transcription factors for proteasomal degradation, and antisense oligonucleotides (ASOs), which are synthetic, single-stranded oligodeoxynucleotides that reduce the stability of *MYC* mRNA[120,121]. Despite technological advancements to target MYC, no clinically approved methods exist to inhibit its activity.

Another way to therapeutically target MYC is to disrupt the mechanisms that direct *MYC* expression. Despite its broad expression pattern, the tissue-specific expression of *MYC* is controlled by discrete regulatory elements. In blood cells, recent work has characterized a blood-specific super-enhancer (BENC), an evolutionarily conserved region located 1.7 megabases downstream of the *MYC* gene body, which is obligatory for regulating *MYC* expression[122]. The MYC BENC is a super-enhancer, a putative enhancer region that forms chromosome loops to facilitate transcription factor binding to promoter regions. Unlike typical enhancer regions, super-enhancers are associated with high levels of activating histone modifications, including monomethylation of histone H3 at lysine 4 (H3K4me1) or acetylation of histone H3 at lysine 27 (H3K27ac)[123–125]. Targeted deletion of the MYC BENC in mice inhibits MLL-driven AML formation, indicating this locus' functional relevance in AML[122]. The MYC BENC is a blood-specific enhancer that controls the expression of *MYC* and is an enticing indirect target to disrupt MYC activity.

The MYC BENC is comprised of multiple enhancer modules that recruit a compendium of transcription factors and global chromatin activators (Figure 1.7A). These chromatin binding proteins, which include BRD4, p300, and members of the Mediator kinase complex, exhibit diverse binding patterns across distinct enhancer modules[122,126]. Disruption of different MYC BENC modules in mice results in loss of specific blood lineages, arguing that other elements control *MYC* expression in different tissues[122]. Moreover, investigation of this locus has revealed that specific MYC BENC modules are also focally amplified in a subset of AML (Figure 1.7B)[122]. The initial excitement surrounding the targeting of activating chromatin complexes bound to the MYC BENC stemmed from the discovery of BRD4 inhibitors. Studies in AML cell lines showed that BRD4 inhibitors resulted in a loss of *MYC* expression and leukemia cell death[127–130]. However, initial clinical trials yielded disappointing results, with only modest clinical activity observed, no discernible correlation between *MYC* gene expression and clinical response, and notable toxicity[131]. While targeting the *MYC* super-enhancer remains promising, new therapeutic approaches are imperative to achieve meaningful clinical benefits.

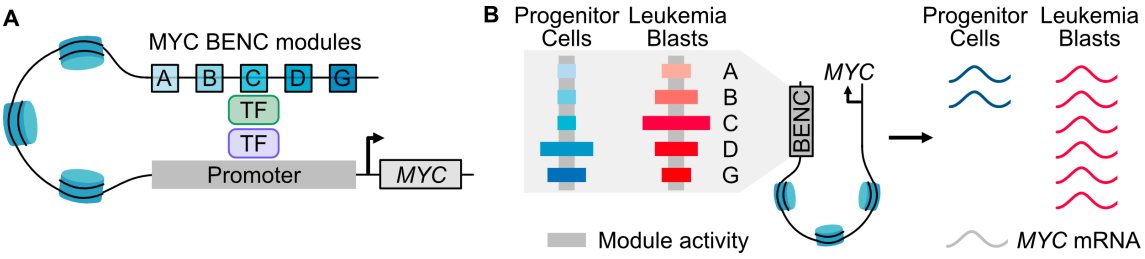

Figure 1.7: The MYC BENC regulates MYC expression in blood cells.
A. The MYC blood-specific super-enhancer cluster (BENC) is composed of individual enhancer modules. The activity of these modules is regulated by transcription factor (TF) binding. **B.** Individual module activity is cell-type specific and is dysregulated in leukemia.

1.7 Enhancing tyrosine kinase inhibitor activity by co-targeting LSD1

Since MYC BENC activity depends on the binding of multiple regulatory proteins, an alternative approach is to simultaneously target two or more factors that regulate *MYC* gene expression. Small molecule inhibitors of lysine specific demethylase 1 (LSD1), a histone demethylase, decrease MYC expression and activity in AML cell lines and primary samples[132–134]. While it has been shown that LSD1 aids MYC transcription by demethylating MYC target gene promoters, it is not well understood how LSD1 regulates *MYC* expression[135,136]. Regardless, these studies suggest that simultaneous targeting of LSD1 and FLT3 represents a promising avenue to disrupt *MYC* expression and, therefore, aberrant FLT3 signaling in FLT3-ITD AML.

LSD1, also known as KDM1A (lysine demethylase 1A) or AOF2 (amine oxidase domain-containing protein 2), is an epigenetic regulatory protein that controls gene expression in two manners. LSD1 is a canonical transcriptional repressor that enzymatically removes activating mono- and di-methylated histone three lysine 4 (H3K4me1/2) marks[137]. LSD1 coordinates hematopoietic differentiation by demethylating H3K4me1/2 and repressing hematopoietic stem and progenitor cell gene expression programs[138]. However, LSD1 can also act as a transcriptional activator by removing repressive mono- and di-methylated histone three lysine 9 (H3K9me1/2) marks. In prostate cancer, LSD1 forms a chromatin-associated complex with androgen receptor that demethylates H3K9 and de-represses androgen receptor target genes[139]. In neuronal cells, an LSD1 isoform, LSD1+8a, complexes with supervillain to regulate neuronal differentiation by demethylating H3K9me2[140]. LSD1 orchestrates gene expression by removing activating H3K4me1/2 and repressive H3K9me1/2 marks.

LSD1 also regulates gene expression as a scaffolding protein, recruiting transcriptional repressor complexes to gene regulatory regions[141,142]. LSD1 H3K4me1/2 demethylation was initially linked to transcriptional repression by its association with the protein complex corepressor for element-1-silencing transcription factor (CoREST)[141]. The CoREST complex is comprised of several transcriptional repressive factors, including histone deacetylases 1/2 (HDAC1/2) and the zinc finger protein growth factor independent 1 (GFI1)[143,144]. GFI1 not only serves as a transcriptional regulator of blood cell proliferation, differentiation, and survival, GFI1 is a molecular hook for LSD1 and its cofactors[145,146]. GFI1 contains a histone 3-mimicking motif, which exhibits a high affinity for LSD1[147]. The CoREST/LSD1 complex is a transcriptional repressor that facilitates GFI1 binding to chromatin.

LSD1-mediated epigenetic dysregulation is an known driver of leukemogenesis. Studies have shown that *LSD1* is overexpressed in various hematological malignancies, including AML[148]. In MLL-driven leukemia, LSD1 sustains expression of oncogenic programs in leukemia stem cells, thus preventing differentiation and apoptosis[149]. Subsequent studies have shown that LSD1 inhibition dislodges the LSD1/CoREST complex from chromatin, resulting in blast differentiation[150,151]. Specifically, LSD1 inhibitors disrupt LSD1-GFI1 binding rather than interfering with LSD1 H3K4me1/2 demethylase activity and reactivate PU.1-bound enhancers. PU.1, encoded by *SPI1*, is a pioneer transcription factor critical for coordinating myeloid differentiation[152]. Overall, studies have demonstrated that LSD1 inhibitors reactivate PU.1-bound enhancers and induce leukemic blast differentiation.

Our group and others have shown that LSD1 inhibitors are not only effective as a monotherapy but also enhance the activity of kinase inhibitors in AML[133,134,153]. In an unbiased chemical screen, our group discovered that LSD1 inhibitors synergize with the JAK/STAT inhibitor ruxolitinib in *CSF3R/CEBPA*-mutant AML. We found that dual

JAK/STAT and LSD1 inhibition resulted in maturation of immature leukemia cells and improved survival in mouse models in a synergistic manner[153]. Our group also investigated the mechanism of combined kinase and LSD1 inhibition in another molecular subtype of AML, *KIT*-mutant AML. In this study, we evaluated the KIT inhibitor avapritinib combined with the LSD1 inhibitor ORY-1001, which has been shown pre-clinically to induce blast differentiation and reduce leukemic stem cell capacity in AML. ORY-1001 is also currently being evaluated in phase I and II clinical trials[154,155]. We found that dual KIT and LSD1 inhibition evicted MYC and PU.1 from chromatin and drove synergistic leukemic cell death[156]. A study led by Dr. Pedicona suggests that LSD1 inhibitors prime leukemia cells for kinase inhibition by activating MEK kinase activity while simultaneously suppressing the activity of other kinase networks[133]. These studies suggest that LSD1 inhibition enhances kinase inhibitor activity in AML and provide a rationale for our research investigating the efficacy and mechanism of combined FLT3 and LSD1 inhibition in Chapter 4.

1.8 Transcriptional adaptation to therapeutic pressure

Transcriptional mechanisms can be used not only to deepen treatment responses but also to understand drug resistance. Disease relapse has been traditionally thought to emerge from intrinsically resistant cells that confer a survival advantage under therapeutic pressure. However, increasing evidence highlights the importance of epigenetic resistance mechanisms to cancer therapy[157]. In this model of drug resistance, therapeutic pressure induces changes in chromatin structure and function, facilitating the expression of transcriptional programs that confer drug resistance. Studies in AML have shown that drug-naïve AML blasts can adopt a

transcriptional state that allows them to remain tolerant to the drug and actively proliferate in the presence of therapeutic pressure[158].

In 2018, the FDA first approved venetoclax in combination with azacitidine, decitabine, or low-dose cytarabine for treating newly diagnosed AML in adults 75 years of age or older. This drug combination established a new standard of care for older and other patients with AML unfit for intensive chemotherapy. With its favorable side effect profile, the addition of venetoclax to azacitidine improved CR (66.4% vs 28.3%) compared to azacitidine alone[159]. Given its importance to the clinical management of AML, venetoclax has garnered much attention from the research community. Venetoclax (ABT-199) is a small molecule inhibitor of B-cell lymphoma 2 (BCL2)[160]. The BCL2 family proteins, such as BLC2, BCL-XL, and MCL1, are anti-apoptotic proteins tightly regulated in the intrinsic pathway of apoptosis[161]. Pro-apoptotic proteins, such as BAX and BAK, induce apoptosis by triggering mitochondrial outer membrane permeabilization. Activation of these pro-apoptotic proteins is inhibited by BCL2 homology 3 (BH3) proteins. By sequestering their pro-apoptotic counterparts, the BCL2 family proteins promote cell survival[162]. Venetoclax is a BH3 mimetic that restores cellular intrinsic apoptosis by blocking the activity of BCL2, which is overexpressed in AML and other malignancies[162]. Venetoclax, which putatively restores intrinsic apoptosis in AML cells, is the basis of standard-of-care therapy for elderly and unfit patients with AML.

A study by Pei et al. revealed an intriguing pattern of venetoclax sensitivity among patients with AML[163]. They found that the AML differentiation state was a critical determinant of response to venetoclax-based therapy. Their studies showed that monocyte-like AML cells were resistant to venetoclax, whereas progenitor-like cells were acutely sensitive. Moreover, this group demonstrated that patients treated with venetoclax-based treatment regimens tend to relapse with monocyte-like disease. These

findings reveal an essential relationship between AML differentiation state and venetoclax response.

Several studies have shown that GMP-like as well as common myeloid and lympho-myeloid primed progenitor (CMP LMPP)-like AMLs, which are sensitive to venetoclax, highly express *BCL2*. Differentiated AMLs, conversely, are resistant to BCL2 inhibition and express high levels of *MCL1* and *BCL-XL*[163,164]. These findings support a model where venetoclax treatment selects monocytic cells that overexpress *MCL1* and *BCL-XL*, evading the initiation of intrinsic apoptosis. However, the study from Pei et al. also showed that patients with dominant, differentiated AML clones achieved a long-lasting response with venetoclax-based therapies. Notably, the authors highlighted a patient with a dominant, differentiated, $NRAS^{Q61K}$-mutant clone that achieved CR for 355 days following venetoclax and azacitidine treatment[163]. This finding casts doubt on a purely selective model of intrinsic resistance with venetoclax, forming the basis of our work described in Chapter 5.

1.9 Dissertation overview

In this work, I outline my investigation of transcriptional dysregulation in AML and how these findings inform novel therapeutic strategies. In Chapters 2 and 3, I describe two computational methods we developed to elucidate transcriptional regulatory mechanisms from next-generation sequencing data. In these chapters, I describe our novel techniques, GoPeaks, which identifies genome-wide histone modifications from CUT&Tag data, and Priori, which infers transcription factor activity from bulk RNA-seq data. In Chapter 4, I detail how we applied these tools to understand the mechanism of a novel drug combination, dual FLT3 and LSD1 inhibition. In Chapter 5, I describe our findings investigating the transcriptional dynamics associated with resistance to the

standard of care therapy for AML, venetoclax. Finally, in Chapter 6, I review the conclusions from these studies and discuss future directions.

Chapter 2: GoPeaks: histone modification peak calling for CUT&Tag

---

- We developed GoPeaks, a peak calling method specifically designed for histone modification CUT&Tag data.
- We compared the performance of GoPeaks against commonly used peak calling algorithms to detect histone modifications that display a range of peak profiles and are frequently used in epigenetic studies.
- We found that GoPeaks robustly detects genome-wide histone modifications and, notably, identifies a substantial number of H3K27ac peaks with improved sensitivity compared to other standard algorithms.

This work has been published in *BMC Genome Biology*:

Yashar WM, Kong G, VanCampen J, Curtiss BM, Coleman DJ, Carbone L, Yardimci GG, Maxson JE, Braun TP. GoPeaks: histone modification peak calling for CUT&Tag. *BMC Genome Biology* 2022 July 4[165]

2.1 Abstract

Genome-wide mapping of histone modifications is critical to understanding transcriptional regulation. CUT&Tag is a new method for profiling histone modifications, offering improved sensitivity and decreased cost compared with ChIP-seq. Here, we present GoPeaks, a peak calling method specifically designed for histone modification CUT&Tag data. We compare the performance of GoPeaks against commonly used peak calling algorithms to detect histone modifications that display a range of peak profiles and are frequently used in epigenetic studies. We find that GoPeaks robustly detects genome-wide histone modifications and, notably, identifies a substantial number of H3K27ac peaks with improved sensitivity compared to other standard algorithms.

2.2 Background

Modification of histone proteins is a key mechanism of transcriptional regulation[9]. Histones package eukaryotic DNA into chromatin and control DNA conformation and organization[10]. Modification of the histone protein results in alteration of chromatin structure and recruitment of nuclear proteins to genomic features[11,12]. For example, trimethylation of histone 3 lysine 4 (H3K4me3) facilitates the binding of positive transcriptional regulators to transcription start sites[13–15]. Similarly, acetylation of histone 3 lysine 27 (H3K27ac) neutralizes the positive charge of the histone tail and loosens the interaction between nucleosomes and DNA, allowing access of transcription factors to DNA regulatory sequences[16,17]. Transcription factors are canonical regulators of transcription whose binding is strongly correlated with histone modifications[12,166–168]. Large-scale studies have demonstrated that transcription factor-binding profiles can be

used to predict histone modifications[169]. An understanding of genome-wide histone modifications is crucial to the understanding of transcriptional regulation.

Chromatin immunoprecipitation with sequencing (ChIP-seq), which couples antibodies that recognize DNA-associated proteins with next generation sequencing technology, has enabled genome-wide profiling of histone modifications and transcription factors[13,18–20]. Although widely used for epigenetic profiling, ChIP-seq is prone to high background, artificial enrichment of highly expressed genes, and often requires prohibitively large number of cells per experiment[21,22]. Enzyme-tethering strategies including Cleavage Under Targets and Tagmentation (CUT&Tag) and Cleavage Under Targets and Release Using Nuclease (CUT&RUN) have been developed to overcome these issues and perform epigenetic profiling with a low number of cells and minimal background[23–26].

Epigenetic studies require mapping multiple histone modifications for a comprehensive understanding of transcriptional regulation. Detecting regions bound to lysine 4 residues on histone 3 that are mono- (H3K4me1) or trimethylated aids with the identification of promoters and enhancers, respectively, throughout the genome[13,27–30]. Co-localization of H3K4me1 and H3K4me3 with H3K27ac is characteristic of activated genomic features[31,32]. Histone modifications can also be associated with heterochromatic regions of the genome where gene transcription is repressed. Localization of trimethylation of histone 3 lysine 27 (H3K27me3), for example, is associated with promoters and gene bodies of silenced genes[13,33–36]. Regions of modified histones in ChIP-seq and CUT&Tag are identified as stacks of aligned reads; such regions are called peaks. The peak profiles of common histone modifications are highly variable, so algorithms that identify histone modification peaks need to robustly detect a range of peak profiles. While H3K4me3 peaks tend to be sharply localized, H3K4me1 and H3K27me3 peaks span a broader domain (Figure 2.1)[34–38]. Moreover,

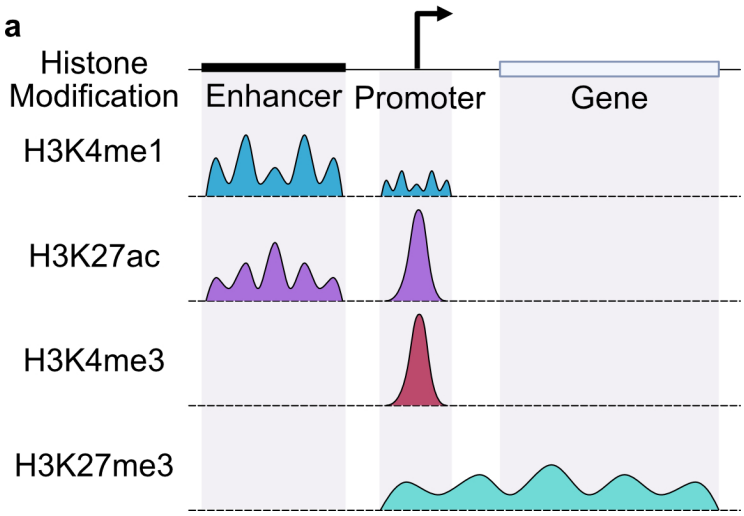

Figure 2.1: Histone modifications exhibit a range of peak profiles. Representative peak profiles for H3K4me1, H3K27ac, H3K4me3, and H3K27me3 histone modifications. Black box indicates an enhancer region, and the gray box indicates the gene body.

H3K27ac can mark large domains such as super-enhancers as well as smaller, discrete regions such as promoters, thus having both broad and narrow characteristics[29,36–40]. In order to extract meaning for epigenetic studies reliant on histone modification CUT&Tag datasets, peak calling algorithms need to be flexible to identifying narrow and broad peak characteristics.

Peak calling algorithms have been developed to not only identify genome-wide enrichment of aligned reads, but also to distinguish peaks of modified histones from noise and artifacts. Model-based Analysis of ChIP-seq version 2 (MACS2), a widely-used peak calling algorithm for ChIP-seq, and other ChIP-seq peak calling methods are designed to address the high rate of background in ChIP-seq and are vulnerable to mistaking background signal as peaks particularly when the background is low[41–43]. Sparse Enrichment Analysis (SEACR) has been developed to perform peak calling from CUT&RUN data that, like CUT&Tag data, is characterized by low background. However,

no peak calling algorithms have been designed to address both the low background and peak profile variability that is characteristic of histone modification CUT&Tag data[43].

Here, we present GoPeaks, a peak calling algorithm designed for histone modification CUT&Tag data. We compared the performance of GoPeaks against other widely used peak calling algorithms to detect H3K4me3, H3K4me1, H3K27me3, and H3K27ac peaks from CUT&Tag data. We demonstrate that GoPeaks robustly detects genome-wide histone modifications and notably, identifies H3K27ac with improved sensitivity compared to other peak callers. Moreover, we showed that GoPeaks can be used to detect transcription factor and chromatin accessibility peaks from ChIP-seq, CUT&RUN, and ATAC-seq data.

2.3 Results

2.3.1 Peak Calling with a Binomial Distribution and a Minimum Count Threshold

GoPeaks performs genome-wide peak identification of histone modification binding from CUT&Tag data in five general steps (Figure 2.2a). First, GoPeaks bins the genome into small intervals. Users can control the width of each bin with the “step” parameter and the width of bin overlap with the “slide” parameter. GoPeaks then quantifies the number of aligned reads contained within each bin. In bins greater than 15 counts (set by the “minreads” parameter), GoPeaks uses a Binomial distribution to determine whether the bin counts are significantly different from the genome-wide distribution of aligned reads. Bins with a significantly large number of counts are retained (p-value less than 0.05 before Benjamini-Hochberg correction by default). Finally, significant bins are merged into peaks if they are located within 150 bp of each other,

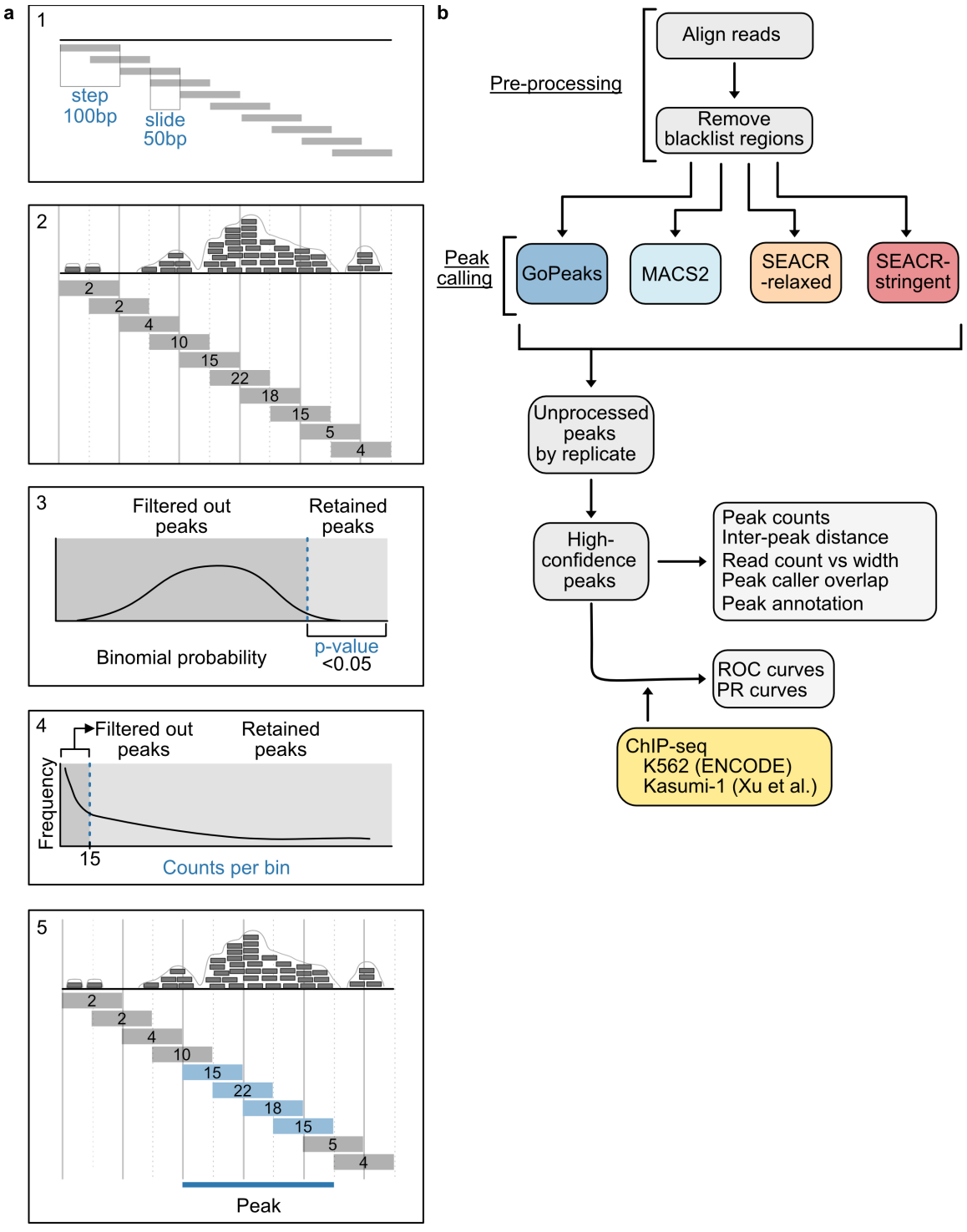

Figure 2.2: Overview of the GoPeaks methodology and benchmarking workflow.

a. Five general steps of the GoPeaks peak calling methodology. Each panel (a1-a5) represents a separate step. (a1) Step indicates the bin width and slide, the width of the bin overlap. (a2) Counting the number of aligned reads per bin.

(a3) Example of a Binomial probability test distribution and threshold to retain significantly different peaks. (a4) Filtering out bins with less than 15 counts. (a5) Retained bins within 150 bp are merged and identified as a peak. **b.** Schematic overview of the benchmarking workflow. All CUT&Tag datasets undergo the same pre-processing and are separately analyzed by the peak calling methods. The unprocessed peaks are extracted for sub-analyses. High-confidence peaks were defined by taking the union of statistically significant peaks from all replicates and retaining the peaks present in at least two biological replicates. Generation of the receiver operator characteristic (ROC) and precision-recall (PR) curves require ChIP-seq standards.

which can be adjusted with the "mdist" parameter. In contrast, MACS2 slides across the genome using an empirically-derived window size and deploys a dynamic Poisson distribution to evaluate the likelihood that aligned reads within a given window region are statistically significant[41]. Following peak p-value correction using the Benjamini-Hochberg procedure, MACS2 then merges overlapping significant regions into a peak. SEACR, on the other hand, bins the genome by regions with contiguous, non-zero signal blocks and calls regions with counts greater than an empirically-derived threshold based on the global distribution of background counts as peaks[43].

We developed a computational workflow to compare the performance of GoPeaks against MACS2 and SEACR to identify histone modification peaks from CUT&Tag data (Figure 2.2b). We evaluated both SEACR threshold parameters, "SEACR-relaxed" and "SEACR-stringent", as well as MACS2 "narrowPeaks" unless otherwise stated. A false discovery rate (FDR) threshold of 0.05 for GoPeaks and MACS2 as well as SEACR's standard empirical FDR threshold was used[43]. We evaluated each peak callers' ability to identify peaks from CUT&Tag sequencing using publicly available H3K4me1, H3K4me3, and H3K27me3 CUT&Tag data in K562 cells, a cell line model of blast-phase chronic myeloid leukemia (CML), and our H3K27ac CUT&Tag data in Kasumi-1 cells, an acute myeloid leukemia (AML) cell line. Each

CUT&Tag dataset was aligned to the GRCh38 genome and the ENCODE blacklist regions were removed[170]. The unprocessed peaks and the high-confidence peaks, defined by taking the union of statistically significant peaks from all replicates and retaining the peaks present in at least two biological replicates, from each CUT&Tag dataset were used to quantify peak characteristics detected by each peak caller. We measured the sensitivity and specificity of each peak caller by their ability to recall peaks from publicly available ChIP-seq standards. We filtered the ChIP-seq standards for peaks with $-\log_{10}$(p-value) > 10 and merged adjacent peaks within 1,000 bp.

2.3.2 Identification of Narrow H3K4me3 Peaks

We first compared the peak counts and characteristics detected by each peak calling algorithm from H3K4me3 CUT&Tag data. GoPeaks and MACS2 identified the greatest number of H3K4me3 peaks (Figure 2.3a). To assess the characteristics of the peaks called by each algorithm, we calculated the average distance to the next nearest peak. This measurement indicates whether peak calling algorithms are splitting up enriched regions and inflating the peak count total. We found that the peaks called by GoPeaks were similar distances apart as MACS2 and SEACR-relaxed (Figure 2.3b). MACS2, however, detected a small population of peaks less than 1,000 bp apart. The peaks called by SEACR-stringent were noticeably farther apart than the other methods. To directly measure peak sizes identified by each peak calling method, we quantified the number of counts in each peak and the peak width. We found that GoPeaks and MACS2 called peaks across a range of widths (Figure 2.3c). Both SEACR-relaxed and SEACR-stringent did not identify any peaks with a width less than 100 bp, potentially missing or aggregating important regions. As an example, all peak callers recognized a peak overlapping the *CBX3* and *HNRNPA2B1* promoters that is approximately 8,500 bp wide (Figure 2.3d). However, only GoPeaks identified a peak located at the promoter of

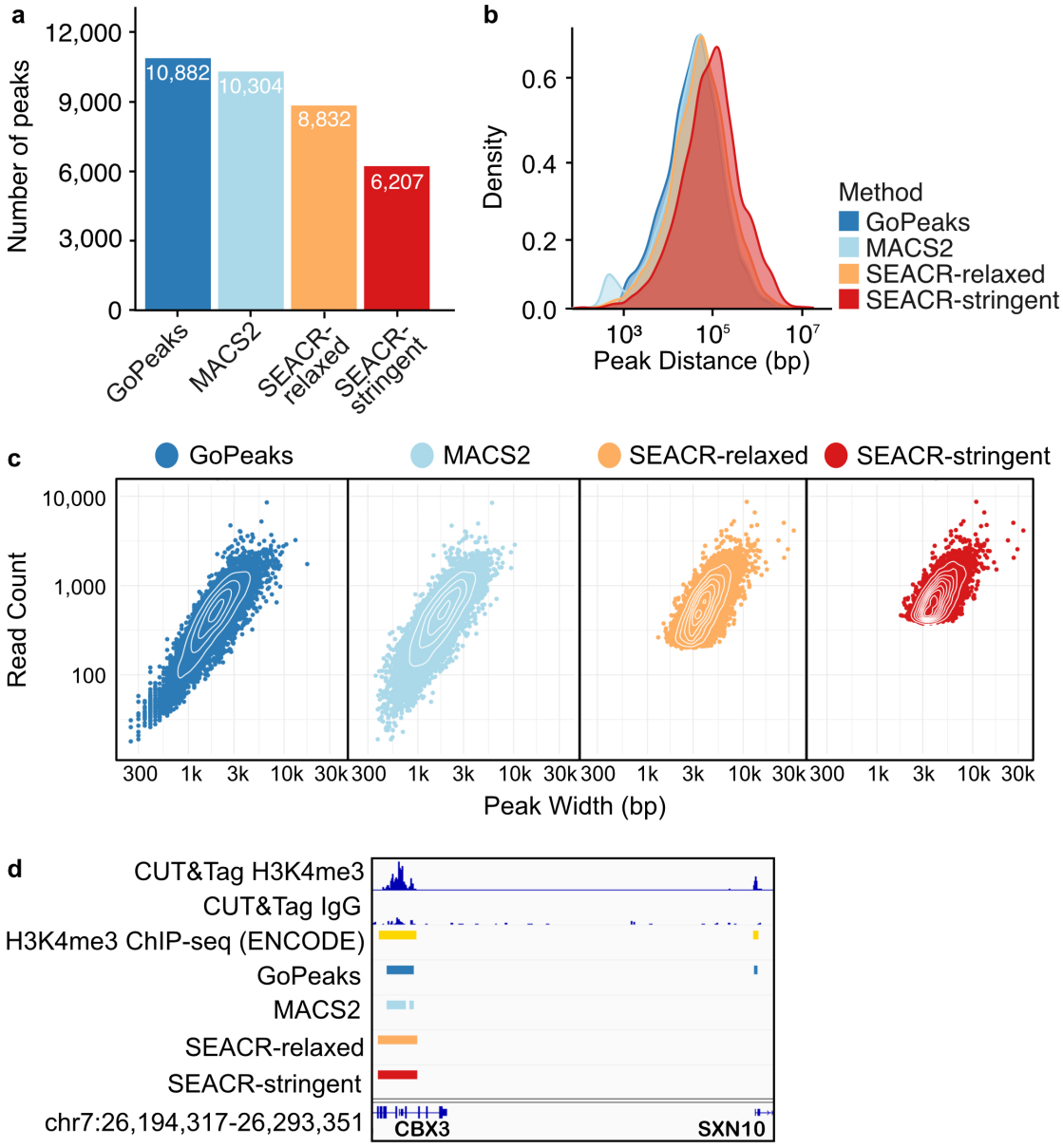

Figure 2.3: GoPeaks and MACS2 perform better than SEACR at identifying a range of H3K4me3 peak sizes.

a. Number of high-confidence peaks identified from H3K4me3 CUT&Tag data in K562 cells per peak calling method. Colors indicate the peak calling method. **b.** Distribution of the distances to the next nearest peak. **c.** Distribution of read counts by peak width. Each dot represents the read count and peak width of a single peak. **d.** Example peaks at the *CBX3* and *SNX10* genes. IgG replicates are negative controls. Consensus peak calls for each method are shown. Tracks are CPM normalized and are scaled to the range [0-5.10] by IGV. Tracks are depicted on the GRCh38 genome assembly.

SNX10 nearly 1,450 bp wide. Together, these results demonstrate GoPeaks' ability to identify H3K4me3 peaks across a range of sizes.

2.3.3 Sensitivity and Specificity of Detecting Narrow H3K4me3 Peaks

While both GoPeaks and MACS2 identify more H3K4me3 peaks than the SEACR peak calling methods, it is unclear whether some of these peaks may be false positives. To understand the sensitivity and specificity of each peak caller for H3K4me3 marks, we compared the peaks identified from publicly available K562 CUT&Tag data to those identified by ChIP-seq on the same cell line from the ENCODE Project[23,171]. We created receiver operating characteristic (ROC) curves, which maps the true positive rate, or recall, against the false positive rate. A true positive is defined as a peak identified in the K562 CUT&Tag data, which is also present in the ENCODE K562 ChIP-seq data. ROC curves along with precision-recall (PR) curves, which instead quantify the relationship between precision and recall, were used to characterize peak caller sensitivity and specificity. GoPeaks and MACS2 demonstrated a greater degree of peak recall than the SEACR methods for a given false positive rate (Figure 2.4a). GoPeaks and MACS2 had comparable area under the ROC curve (AUROC), which is a measurement of how well the peak callers detect CUT&Tag peaks that are also present in the ChIP-seq standard (Figure S2.1a). Every method demonstrated a similar ability to identify peaks with high precision across a range of recall values (Figure 2.4b; Figure S2.1b). It is unlikely to observe perfect concordance due to the technical differences between the CUT&Tag and ChIP-seq assays. Indeed, given the sensitivity of CUT&Tag, it is likely that CUT&Tag will identify regions enriched with aligned reads that are not evident in ChIP-seq data.

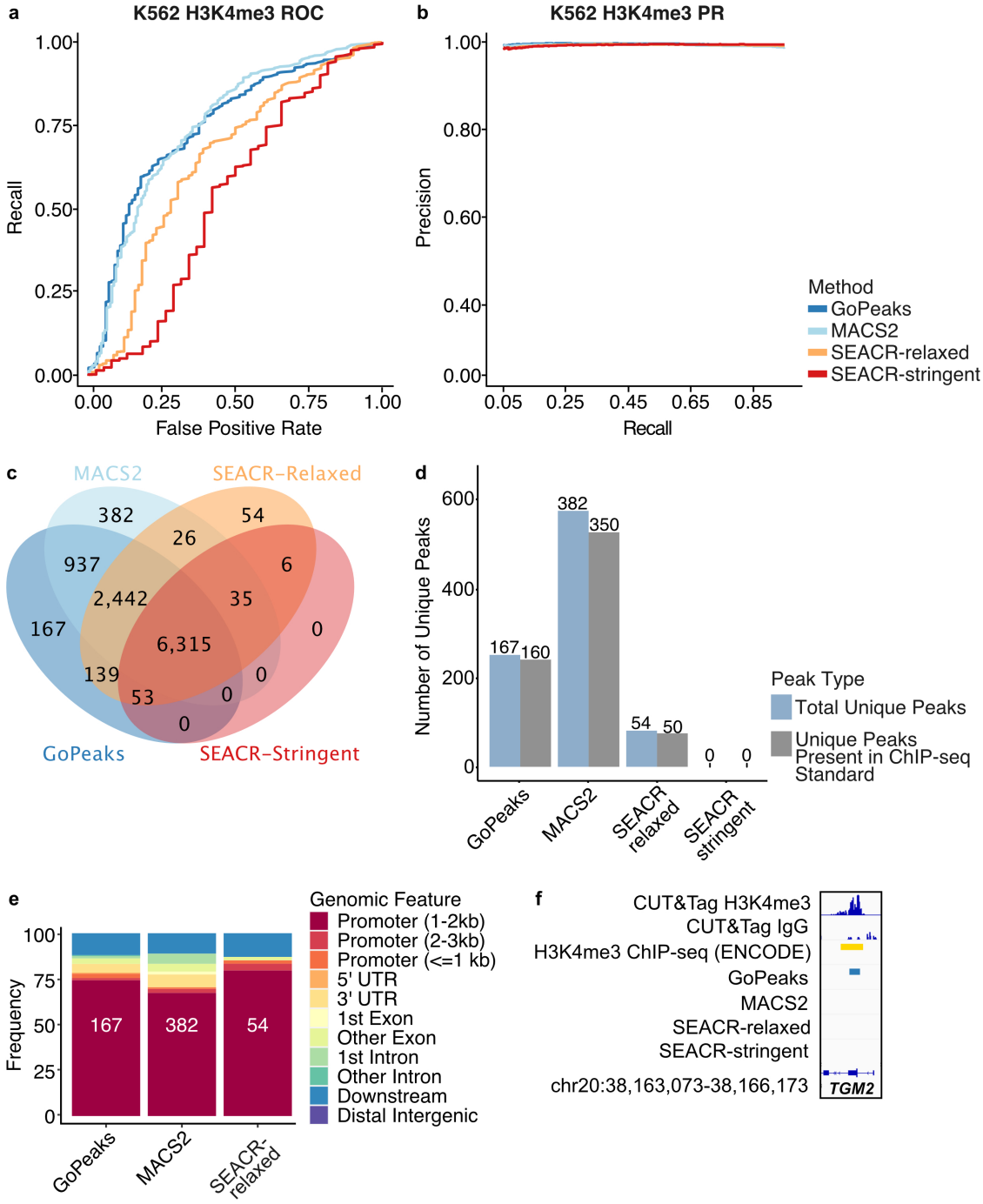

Figure 2.4: GoPeaks has a favorable specificity and sensitivity for narrow H3K4me3 CUT&Tag peaks.

a. ROC curves quantifying the recall and false positive rates and **b.** PR curves quantifying the precision and recall rates of H3K4me3 CUT&Tag data from H3K4me3 ChIP-seq data. Both ChIP-seq and CUT&Tag datasets were generated in K562 cells. Colors indicate the peak calling method.

c. Overlap of high-confidence peaks identified by each peak caller. **d.** Comparison of unique peaks that are identified by each peak calling algorithm and are also present in the ChIP-seq standard. Each bar is labeled by the number of peaks it represents. Colors indicate the peak type. **e.** Annotation of unique peaks identified by each peak caller. Each bar is labeled by the number of unique peaks. Colors indicate the genomic feature. Downstream is at least 300 bp towards 3' end of DNA strand. **f.** Example peaks at the *TGM2* gene. IgG replicates are the negative controls. Consensus peak calls for each method are shown. Tracks are CPM normalized and are scaled to the range [0-1.46] by IGV. Tracks are depicted on the GRCh38 genome assembly. UTR = untranslated region.

GoPeaks' high sensitivity and specificity is likely due in part to its ability to identify peaks not captured by the other peak calling algorithms. To assess what may have distinguished each peak callers' performance, we studied the overlap of the high-confidence peaks. GoPeaks and MACS2 called the majority of peaks detected by both SEACR-stringent and SEACR-relaxed (Figure 2.4c). In addition, GoPeaks identified 167 peaks not detected by any other peak caller, which likely contributed to GoPeaks' high sensitivity and specificity as 95.8% (160) of GoPeaks' unique peaks were also present in the ChIP-seq standard (Figure 2.4d; Figure S2.1c). Since H3K4me3 peaks are associated with promoters, we annotated each peak set to the nearest genomic feature[13,27,28]. The unique peaks identified by GoPeaks, MACS2, and SEACR-stringent were mostly associated with promoters (77.8%, 70.2%, and 85.2%, respectfully), consistent with the established biology of H3K4me3 (Figure 2.4e). SEACR-stringent did not identify any unique peaks. Consistent with these findings, GoPeaks detected a peak in both CUT&Tag replicates located at the promoter of *TGM2* that was not identified by the other peak calling methods (Figure 2.4f). Collectively, these results reveal that GoPeaks has favorable sensitivity and specificity for H3K4me3 data when compared with ChIP-seq, enabling the identification of an increased number of true positive peaks with a minimal false positive rate and at high precision.

2.3.4 Sensitivity and Specificity of Detecting Broad H3K4me1 Peaks

While GoPeaks was highly sensitive and specific to narrow H3K4me3 peaks, we wanted to evaluate its performance to detecting broad H3K4me1 peaks. To compare the performance of each peak caller to detect H3K4me1 CUT&Tag peaks from K562 cells, we measured their sensitivity and specificity against ENCODE H3K4me1 ChIP-seq on the same cell line[23,171]. Overall, GoPeaks demonstrated comparable sensitivity and specificity across both H3K4me1 replicates (Figure 2.5a, b; Figure S2.2a, b).

GoPeaks' sensitivity and specificity may be due to its ability to detect H3K4me1 marks in intronic and intergenic regions. We therefore evaluated the overlap of the high-confidence peaks identified by each peak calling method. SEACR-relaxed identified the greatest number of unique H3K4me1 peaks among the four peak calling algorithms (Figure 2.5c; Figure S2.2c). GoPeaks still identified 1,142 unique peaks, 85.5% (976) of which were also present in the ChIP-seq standard (Figure S2.2d). To confirm the genomic features associated with the H3K4me1 peaks, we annotated each peak set to the nearest genomic feature. The unique peaks identified by GoPeaks and MACS2 were primarily associated with intronic and intergenic regions (76.8% and 85.1%, respectively) whereas SEACR-relaxed were mostly associated with promoters (62.7%; Figure 2.5d). While H3K4me1 is found at active promoters, it displays the greatest enrichment at enhancers[13,27,28]. As an example, GoPeaks was able to identify a unique peak in the intronic region of *FOXO3* and the center of another H3K4me1 peak with greatest density of H3K4me1 signal (Figure 2.5e). Both SEACR methods, on the other hand, called regions with widths greater than region annotated by the ChIP-seq standard. In regions where the intronic region is much smaller, like in *FTCD*, resolving the center of peaks is crucial (Figure 2.5f). Both SEACR methods identify genomic features that include the

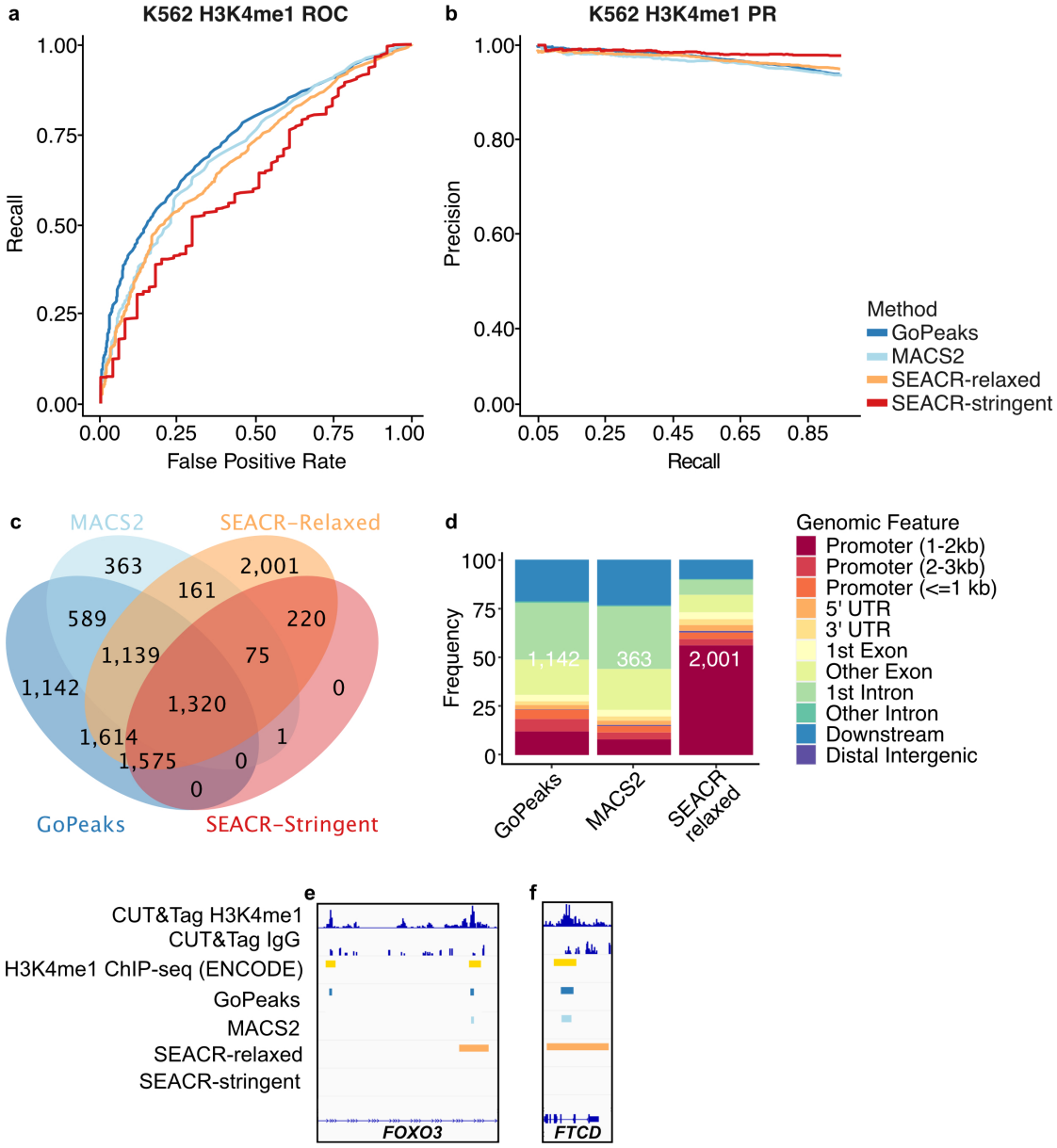

Figure 2.5: GoPeaks has a favorable specificity and sensitivity for broad H3K4me1 CUT&Tag peaks.

a. ROC curves quantifying the recall and false positive rates and **b**. PR curves quantifying the precision and recall rates of H3K4me1 CUT&Tag data from H3K4me1 ChIP-seq data. Both ChIP-seq and CUT&Tag datasets were generated in K562 cells. Colors indicate the peak calling method. **c**. Overlap of high-confidence peaks identified by each peak caller. **d**. Annotation of unique peaks identified by each peak caller. Each bar is labeled by the number of unique peaks. Colors indicate the genomic feature.

Downstream is at least 300 bp towards 3' end of DNA strand. **e, f**. Example peaks at the e. FOXO3 gene and f. FTCD genes. IgG replicates are the negative controls. Consensus peak calls for each method are shown. Tracks are CPM normalized and are scaled to the range [0-1.31] for D and [0-1.33] for E by IGV. Tracks are depicted on the GRCh38 genome assembly. UTR = untranslated region.

promoter, exonic, and intronic regions of the *FTCD* gene, in contrast to what is annotated in the ChIP-seq standard as well as detected by GoPeaks. Together, these findings reveal that GoPeaks has favorable sensitivity and specificity in identifying H3K4me1 peaks while simultaneously calling sufficiently narrow peaks to separate promoter and non-promoter regulatory regions.

2.3.5 Sensitivity and Specificity of Detecting Broad H3K27me3 Peaks

Although we demonstrated that GoPeaks robustly detects peaks of histone modifications associated with euchromatic regions of the genome, we wanted to understand the ability of GoPeaks to identify H3K27me3 peaks, which are associated with regions of repressed gene transcription[13,33–36]. Similar to the peak profiles of H3K4me1, H3K27me3 peaks tend to span broad regions. However, H3K27me3 signal can be detected across entire bodies of silence genes[13,33–36]. To address this, we developed a "--broad" flag for GoPeaks to alter the width of genome bins to 5,000 bp and bin overlap to 1,000 bp, enabling the capture of peaks with broad domains. While MACS2 also has a broad peaks feature that increases the gap length to merge nearby enriched peaks, the SEACR methods do not[41,43].

We compared the sensitivity and specificity of the GoPeaks and MACS2 broad methods and the standard SEACR methods to detect H3K27me3 CUT&Tag peaks from K562 cells, using ENCODE H3K27me3 ChIP-seq data from the same cell line as the

standard[23,171]. While GoPeaks and MACS2 had enhanced AUROC as compared to the SEACR methods, MACS2 had a considerably lower AUPRC (Figure 2.6a, b; Figure S2.3a, b). As precision is dependent on the number of true positive peaks, we measured the number of peaks identified by each method. MACS2 detected 26,819 peaks, more than double the peak count of the next closest method (Figure 2.6c). However, 94.3% (25,285) of these peaks overlapped peaks identified by the other methods (Figure 2.6d; Figure S2.3c). Moreover, of the 1,534 peaks uniquely identified by MACS2, only 38.6% (592) were present in the standard (Figure S2.3d). In contrast, GoPeaks detected a comparable number of unique peaks (1,303), the majority of which were present in the standard (60.9%; 793). We hypothesized that MACS2 may be splitting the broad domains of enriched H3K27me3 signal into small peaks, so we measured the distribution of peak sizes and distances to the next nearest peak. Indeed, we identified a large population of MACS2 peaks that contained less than 100 counts (Figure S2.3e) and that MACS2 peaks were substantially closer together than the other methods (Figure 2.6e). As an example, GoPeaks and the SEACR methods preserved the broad domain of enriched H3K27me3 signal spanning the gene bodies of *FIBCD1*, *LAMC3*, and *AIF1L* (Figure 2.6f). Even with the "—broad" flag, MACS2 breaks up these domains into small peaks. These findings demonstrate that GoPeaks detects broad domain peaks within heterochromatic regions with high sensitivity and specificity.

2.3.6 Sensitivity and Specificity of Detecting Broad & Narrow H3K27ac Peaks

H3K27ac marks are crucial for defining active genomic features and can have characteristics of broad and narrow peaks[31,32]. To evaluate the performance of GoPeaks on H3K27ac, we performed CUT&Tag sequencing for H3K27ac on Kasumi-1

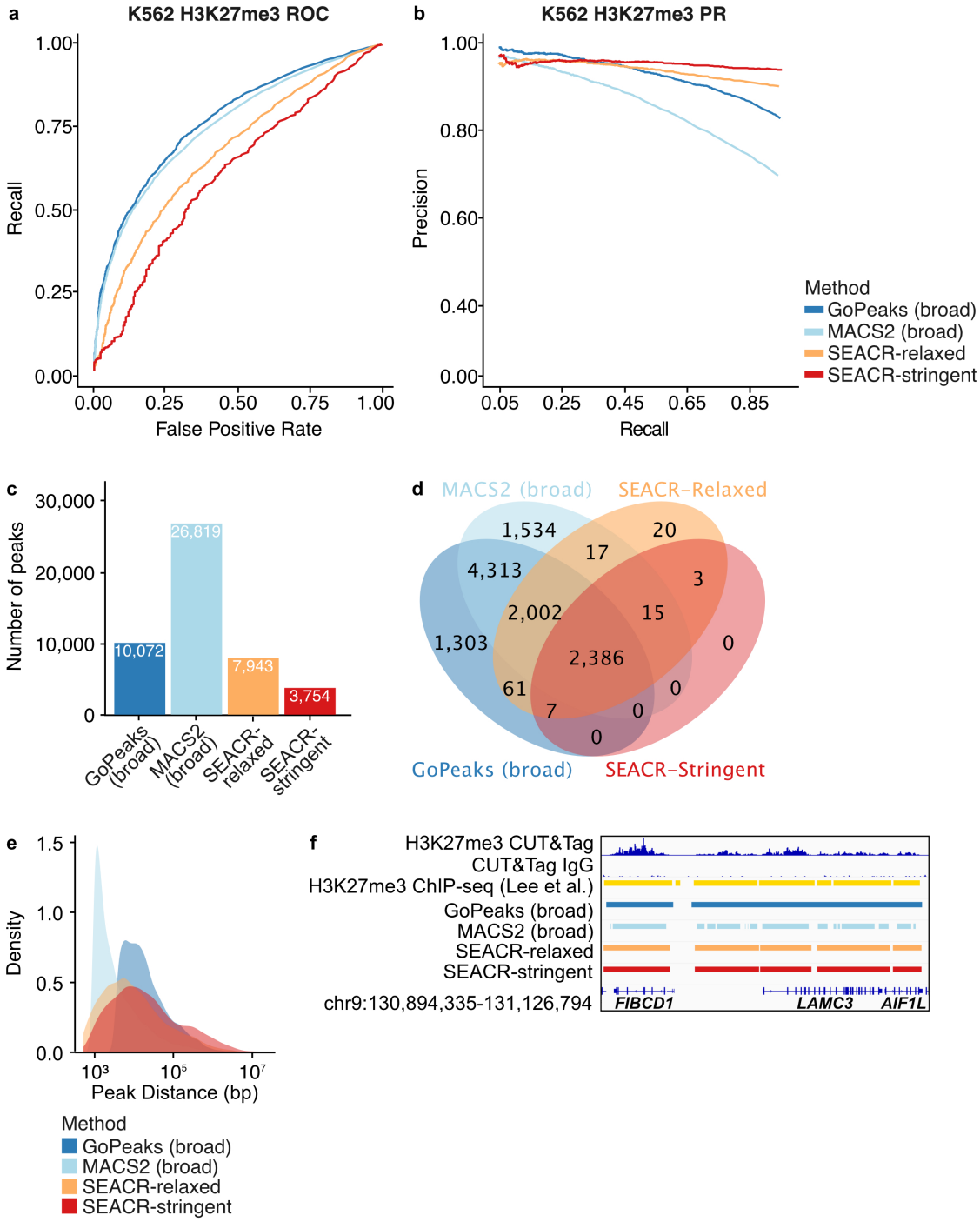

Figure 2.6: GoPeaks captures the broad peak profiles of H3K27me3 CUT&Tag peaks.

a. ROC curves quantifying the recall and false positive rates and **b.** PR curves quantifying the precision and recall rates of H3K27me3 CUT&Tag data from H3K27me3 ChIP-seq data.

Both ChIP-seq and CUT&Tag datasets were generated in K562 cells. Colors indicate the peak calling method. GoPeaks and MACS2 "--broad" flags were used. **c.** Number of high-confidence peaks identified from H3K4me3 CUT&Tag data per peak calling method. **d.** Overlap of high-confidence peaks identified by each peak caller. **e.** Distribution of the distances to the next nearest peak. **f.** Example peaks at *FIBCD1*, *LAMC3*, and *AIF1L* genes. IgG replicates are the negative controls. Consensus peak calls for each method are shown. Tracks are CPM normalized and are scaled to the range [0-18] by IGV. Tracks are depicted on the GRCh38 genome assembly.

cells. We again measured the number and characteristic of peaks called in CUT&Tag data as compared to ChIP-seq. Since ENCODE does not have H3K27ac ChIP-seq data for Kasumi-1 cells, we identified the high-confidence peaks of published H3K27ac ChIP-seq data on the same cell line[172]. While GoPeaks showed an enhanced ability to recall peaks across a range of false positive rates, this may have been at the expense of its precision and recall (Figure 2.7a, b; Figure S2.4a, b). Since precision is dependent on the number of true positives identified by each caller, we once again measured the number of peaks identified by each method. Indeed, GoPeaks detected a total of 9,760 high-confidence peaks, which is nearly 3,000 more peaks than what was detected by the closest peak caller, as well as 2,817 peaks that were not detected by any other peak caller (Figure 2.7c, d; Figure S2.4c). While only 1,078 peaks of the unique peaks were present in the standard (Figure S2.4d), GoPeaks identified 69.4% of all peaks present in the ChIP-seq standard (Figure 2.7e). In contrast, SEACR-stringent, which demonstrated improved precision and recall over the other peak calling methods, identified the least number of total peaks (3,743). Moreover, SEACR-stringent did not detect any unique peaks and only identified 33.0% of the peaks present in the standard. Overall, GoPeaks identified a substantial number of H3K27ac peaks that were present in the ChIP-seq standard with some trade-off to its precision.

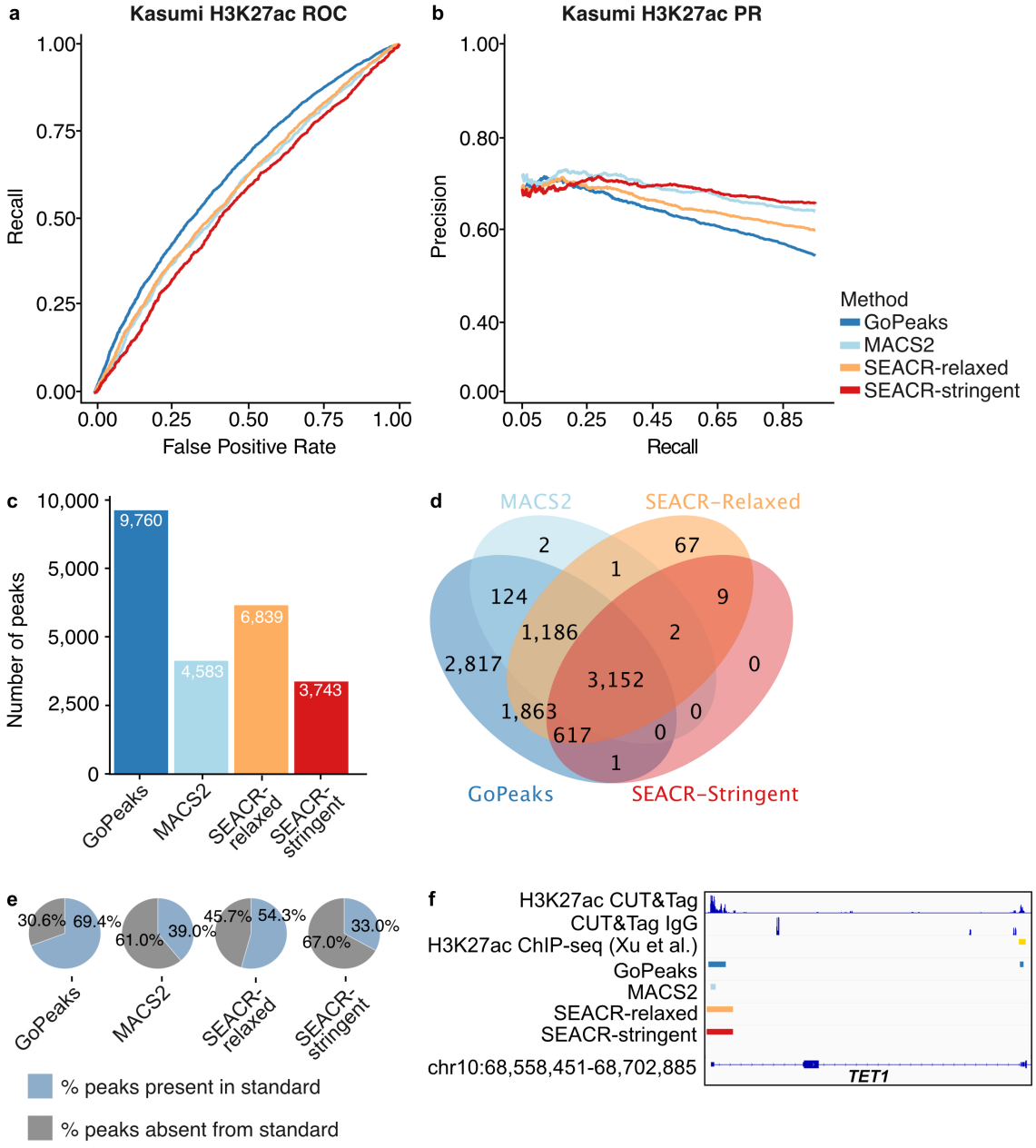

Figure 2.7: GoPeaks has higher specificity and sensitivity for H3K27ac CUT&Tag peaks with broad and narrow peak shapes.

a. ROC curves quantifying the recall and false positive rates and **b.** PR curves quantifying the precision and recall rates of H3K27ac CUT&Tag data from H3K27ac ChIP-seq data. Both ChIP-seq and CUT&Tag datasets were generated in Kasumi-1 cells. Colors indicate the peak calling method. **c.** Number of high-confidence peaks identified from H3K4me3 CUT&Tag data per peak calling method. **d.** Overlap of high-confidence peaks identified by each peak caller.

e. Percent of total H3K27ac ChIP-seq standard peaks that are identified by each peak caller. **f.** Example peaks near at the *TET1* gene. IgG replicates are the negative controls. Consensus peak calls for each method are shown. Tracks are CPM normalized and are scaled to the range [0-4.26] by IGV. Tracks are depicted on the GRCh38 genome assembly.

GoPeaks was able to identify peaks across a range of widths, which is crucial for H3K27ac peak detection (Figure S2.4e). As an example, GoPeaks identified a peak with both narrow and broad characteristics at the promoter of *TET1* (Figure 2.7f). While GoPeaks and SEACR identified the whole peak, MACS2 only identified the narrow portion of the peak. Additionally, GoPeaks identified another narrow peak in an exonic region of *TET1*, which is also present in the ChIP-seq standard. Together, this data highlights GoPeaks' dynamic range to identify both the narrow and broad peaks, which are characteristic of H3K27ac marks.

2.3.7 Identification of Transcription Factor and Chromatin Accessibility Peaks Using GoPeaks

While GoPeaks was designed for histone modification CUT&Tag data, the GoPeaks peak calling methodology can be applied to any epigenetic profiling data with read pileups aligned to a genome. Therefore, we wanted to determine the applicability of GoPeaks to other epigenetic profiling techniques, including ChIP-seq, CUT&RUN, and Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq). Moreover, surveying these methods allowed us to understand whether GoPeaks is capable of identifying peaks associated with transcription factors and chromatin accessibility. We first assessed the ability of GoPeaks to detect transcription factor binding events from RUNX1 ChIP-seq from K562 cells and Sox2 CUT&RUN from H1 human embryonic stem cells (hESC)[171,173]. For the analysis of ChIP-seq data, we compared the characteristics of peaks detected by GoPeaks only to MACS2 as SEACR was developed for CUT&RUN

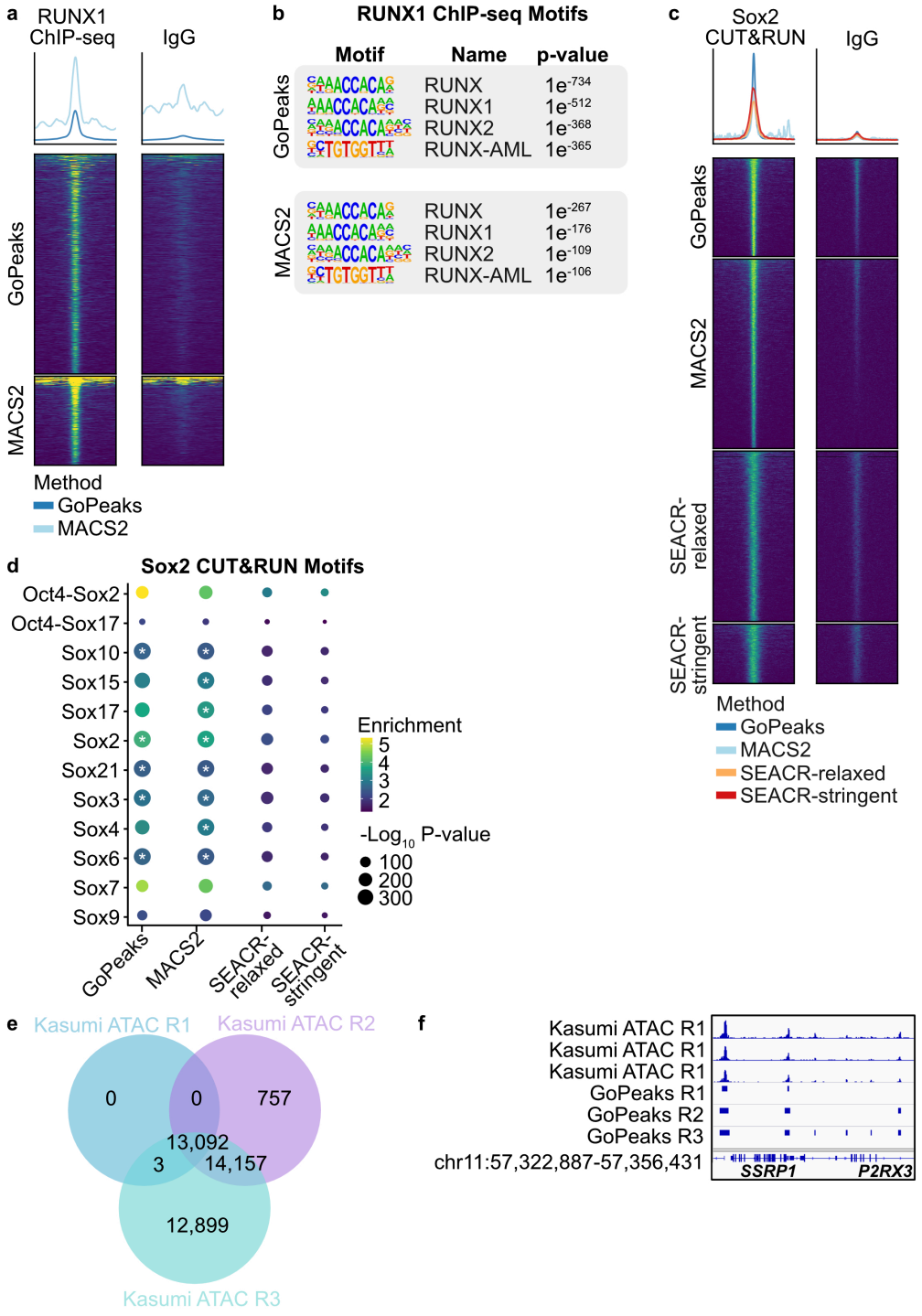

Figure 2.8: GoPeaks detects peaks from other epigenetic profiling techniques.

a. Heatmap of global RUNX1 and IgG ChIP-seq signal from K562 cells per peak calling method. **b.** Transcription factor motif enrichment for regions of global

RUNX1 ChIP-seq signal. Top four known motifs are shown. **c.** Heatmap of global Sox2 and IgG CUT&RUN signal from H1 hESC cells per peak calling method. **d.** Transcription factor motif enrichment for regions of global Sox2 CUT&RUN signal. Dot color represents the binomial fold enrichment and color represents the $-\log_{10}$(P-value) of the motif. The asterisk indicates the motif p-value is less than R's smallest representable number ($1x10^{-324}$). **e.** ATAC-seq was performed on Kasumi cells. Overlap of peaks identified in each biological replicate by GoPeaks. **f.** Example peaks at *SSRP1* and *P2RX3* genes. Peak calls for individual biological replicates are shown. Tracks are CPM normalized and are scaled to the range [0-3.11] by IGV. Tracks are depicted on the GRCh38 genome assembly.

data[43]. We found that both GoPeaks and MACS2 detected peaks from the K562 RUNX1 ChIP-seq data that were enriched for RUNX transcription factor binding motifs (Figure 2.8a, b)[124]. Moreover, both GoPeaks and MACS2 detected the enrichment of Sox bindings motifs with high confidence from the H1 hESC Sox2 CUT&RUN data (Figure 2.8c, d)[124]. These findings show that GoPeaks can identify transcription factor binding events from ChIP-seq and CUT&RUN data. To understand whether GoPeaks detects chromatin accessibility peaks, we performed ATAC-seq in Kasumi cells. GoPeaks identified 13,092 overlapping peaks in the three biological replicates (Figure 2.8e). GoPeaks detected the most peaks in the R3 biological replicate with an additional 14,157 peaks that were shared by the R2 replicate as well as 12,899 peaks only present in the R3 replicate. This was evident at the *SSRP1* and *P2RX3* genes where GoPeaks identified enriched regions of chromatin accessibility in all three replicates, but also detected additional low count peaks in the R3 replicate (Figure 2.8f). Collectively, these findings demonstrate the applicability of GoPeaks to call transcription factor and chromatin accessibility peaks.

2.4 Discussion

GoPeaks was designed to address the low background and peak profile variability characteristic of histone modification CUT&Tag data. GoPeaks demonstrated a favorable ability to call peaks and were highly sensitive and specific to calling peaks across a range of histone modification CUT&Tag data. These results were particularly encouraging for H3K27ac peaks, which exhibit both narrow and broad characteristics. Since H3K27ac is a marker of active promoters and enhancers, it is crucial to pinpoint active regulatory non-coding elements. Moreover, we demonstrated that GoPeaks can identify transcription factor and chromatin accessibility peaks from ChIP-seq, CUT&RUN, and ATAC-seq data.

MACS2 and SEACR both demonstrated biases towards the identification of narrow or broad peaks, respectively. MACS2 performed particularly well in analyzing H3K4me3 CUT&Tag data, in which peaks tend to be sharply localized. MACS2 identified a comparable amount of H3K4me3 CUT&Tag peaks as GoPeaks with similar sensitivity and specificity. The bias of MACS2 for narrow peaks was also evident when calling peaks from H3K27me3 CUT&Tag data. Even with the MACS2 broad feature, MACS2 fragmented enriched H3K27me3 domains that spanned the entire gene bodies into small peaks. As MACS2 was designed to identify narrow transcription factor peaks in ChIP-seq data, its bias for narrow peaks is unsurprising. SEACR, in contrast, demonstrated a favorable ability in detecting broad H3K4me1 peaks[42]. SEACR-relaxed identified the most unique H3K4me1 peaks, but with comparable precision and recall to GoPeaks. The reason for SEACR's bias for broad marks may be due to its empirical segmentation of the genome with contiguous, non-zero signal blocks[43]. Since the signal blocks are not fixed, peaks may contain excess regions with low counts that are adjacent to a true peak (Figure 2.5f). GoPeaks avoids this potential problem as each bin has a fixed width and is

evaluated for significance before merging. GoPeaks' simple but flexible framework is more amendable to the identification of both broad and narrow peaks, like those present in H3K27ac data.

SEACR was more conservative in the identification of peaks across all marks. SEACR-stringent, in particular, consistently detected less peaks than the other methods. This strategy seems to be beneficial for SEACR's precision and recall, notably in the detection of H3K27ac peaks, and may be more appropriate for researchers that have a low threshold for calling false positives. GoPeaks, on the other hand, may be best suited for researchers that are interested in discovering new peaks with a small trade-off in precision. However, GoPeaks largely performed at comparable precision and improved recall over the other peak callers. Our analysis demonstrates GoPeaks detects a substantial number of histone modification peaks at high sensitivity and specificity.

There are important limitations to consider in this analysis. We only evaluated the sensitivity and specificity of each peak calling method using four CUT&Tag histone modification datasets. Although the modifications studied are likely important for epigenetic studies, the peak profiles cover a broad range that will be likely encountered by other marks. We encourage users to test GoPeaks on other histone modification datasets as well as transcription factor and chromatin accessibility datasets. While we demonstrated that GoPeaks can call peaks from these modalities, more work is needed to characterize its performance. Additionally, we only tested these peak calling methods in three different cell lines. We cannot confidently rule out that GoPeaks may have a biological bias for K562, Kasumi-1, or H1 hESC although this is unlikely. The epigenetic profiles of these cells are publicly available, which served as an important comparator for the ROC studies. Lastly, there were no CUT&Tag standards for the ROC studies. Although it would have been preferable to compare peaks between CUT&Tag datasets, the CUT&Tag technique is still new, and few datasets exist in the public domain. While

ChIP-seq served as a useful comparator given the abundance of publicly available data, the fundamental differences in the ChIP-seq and CUT&Tag methods as well as the histone modification antibodies used to generate each dataset should be considered when comparing the results of these two methods. Notably, the antibodies used for the H3K4me1 and H3K27me3 CUT&Tag and ChIP-seq experiments were the same, but the antibodies for H3K4me3 and H3K27ac were not. However, our analysis indicates GoPeaks' ability to extract biological meaning from CUT&Tag data. Overall, GoPeaks demonstrated to be a robust peak calling method across a range of histone modification CUT&Tag data.

2.5 Conclusions

GoPeaks is a peak calling algorithm designed for histone modification CUT&Tag data. We showed that GoPeaks detects peaks of histone modifications that are frequently used in epigenetic studies with high sensitivity and specificity. Notably, GoPeaks demonstrated an improved ability to identify H3K27ac peaks, which are critical to localizing active regulatory non-coding elements throughout the genome, over other standard peak calling algorithms. Moreover, we showed that GoPeaks is able to detect peaks from other epigenetic profiling techniques, including ChIP-seq, CUT&RUN, and ATAC-seq.

2.6 Methods

2.6.1 GoPeaks Algorithm

GoPeaks detects peaks from aligned, paired-end sequencing reads by calculating read coverage in genome bins ("step" 100 bp and "slide" 50 bp by default).

Reads from each sample are normalized to counts per million (CPM) and, when a negative control experiment is provided (e.g., IgG, input), then scaled per pin using a custom scaling function:

$$sample_{bin} = \begin{cases} sample_{bin} > negativeControl_{bin}, & sample_{bin} * \left(1 - \frac{CPM(negativeControl_{bin})}{CPM(sample_{bin})}\right) \\ sample_{bin} < negativeControl_{bin}, & 0 \end{cases}$$

GoPeaks then collects important parameters “n” and “p” to model the Binomial distribution. “n” is equal to the number of reads and “p” represents the probability of success, where success is defined as choosing “x” reads in a genome bin from “n”. “p” is estimated as the average read depth in all non-zero coverage bins over the number of non-zero bins, divided by “n”. Modeling read counts using a Binomial distribution was originally inspired by works from the Regulatory Genomics Toolbox[174]. GoPeaks then traverses the genome and tests bins greater than “minreads” (15 by default), followed by multiple hypothesis correction using Benjamini-Hochberg. Bins are filtered by their adjusted p-values with “-p”. Adjacent bins are merged with the “--mdist” parameter (150 bp by default) into peaks, where they are filtered using a minimum peak width flag with “--minwidth” (150 bp by default).

2.6.2 Method Comparison Workflow

2.6.2.1 Pre-processing

K562 H3K4me3 (GEO accession GSM3536516), H3K4me1 (GEO accession GSM3536518), H3K27me3 (GEO accession GSM3560261), and IgG (GEO accession GSM3560264) CUT&Tag data from Kaya-Okur et al. 2019 as well as H1 hESC Sox2 (GEO accession SRR8748855 and SRR8748856) and IgG (GEO accession

SRR8748845) CUT&RUN data from Meers et al. 2019 were downloaded through National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO)[23,173,175]. K562 RUNX1 (ENCODE ID ENCSR414TYY) and the IgG input control (ENCODE ID ENCSR173USI) ChIP-seq data were downloaded from the ENCODE portal[171,176]. All data was aligned to the GRCh38 genome with Bowtie2[13] with the following options "--local --very-sensitive-local --no-unal --no-mixed --no-discordant --phred33 -I 10 -X 700".

2.6.2.2 Peak Calling

GoPeaks (v1.0.0) used the optional flag "--mdist 1000" to merge peaks within 1,000 bp. For H3K27me3 CUT&Tag, the flags "--mdist 3000" and "--broad" were used to adjust the step to 5,000 bp and the slide to 1,000 bp. MACS2 (v 2.2.6) used the "--format BAMPE" flag with a genome size of 2.7e9 and the standard FDR threshold of 0.05[41]. MACS2 "--broad" was used for H3K27me3 CUT&Tag data and "narrowPeak" was used for all other data. SEACR (v 1.4) used the "norm" flag when treatment and IgG samples were used in addition to using the relaxed and stringent mode[43]. SEACR uses an empirical false discovery rate (FDR) calculated by quantifying the percentage of control signal blocks remaining out of the total above the threshold[43].

2.6.2.3 Post-processing

After peaks were called for each method, high-confidence peaks were defined by taking the union of statistically significant peaks from all replicates and retaining the peaks present in at least two biological replicates within a study's data set via a custom script. The purpose of finding high-confidence peaks is to reduce spurious peaks called in only one replicate and focus on peaks that consistently appear in multiple replicates.

Intervene was used on the high-confidence peak sets to find common and exclusive peaks across peak callers[178].

2.6.2.4 Peak Characterization

Peak counting was done in base R. ChIPseeker was used to annotate peaks to the nearest transcription start site[179]. The read count at high-confidence peak intervals was tallied with BEDtools intersect –C to yield read depth density distributions, and peak-peak distances were calculated with GRanges[180,181]. Enrichment of transcription factor motifs in RUNX1 ChIP-seq and Sox2 CUT&RUN peaks were identified using HOMER[124]. Data cleaning and visualization were mainly facilitated using data.table, ggplot2, and deeptools[182,183]**.** Tracks were normalized by CPM and visualized using Integrative Genomics Viewer (IGV)[184].

2.6.2.5 Receiver Operating Characteristic and Precision-Recall Curves

In the receiver operating characteristic (ROC) and precision-recall (PR) analyses, the ranking metrics for each peak calling algorithm was counts at high-confidence peaks (obtained through BEDtools intersect -C). The outputs of MACS2, SEACR, and GoPeaks high-confidence peak counts were the input for ROC and PR analyses. The high-confidence counts from CUT&Tag data were compared to publicly available ChIP-seq standards downloaded from the ENCODE portal and ChIP-Atlas[171,176,185]. K562 H3K4me3 (ENCODE ID ENCFF885FQN), H3K4me1 (ENCODE ID ENCFF759NWD), and H3K27me3 (ENCODE ID ENCFF795ZOS) ChIP-seq data was accessed from the ENCODE portal[171,176]. Kasumi-1 H3K27ac (ChIP-Atlas SRX ID SRX4143063 and

SRX4143067) ChIP-seq data was accessed from ChIP-Atlas[172,185]. The standards were filtered for peaks with $-\log_{10}$(p-value) > 10 and adjacent peaks were merged if they were within 1,000 bp.

Custom scripts were used to threshold over unique values of the ranking metric to define predicted truth and false, which were intersected with the ChIP-seq standards to fill out the confusion matrix. True negatives are defined as peaks that did not meet the threshold for significance and were not annotated in the ChIP-seq standard. Secondary properties such as precision, recall, and FPR, were calculated from the primary properties. ROC curves were made by plotting precision versus false positive rate; PR curves were made by plotting precision versus recall. The area under the AUROC and AUPR curves were approximated with Riemann Sums using trapezoids.

2.6.3 Cell Culture and CUT&Tag Methods

2.6.3.1Cell Lines

Kasumi-1 cells (ATCC) were cultured in RPMI (Gibco) supplemented with 20% fetal calf serum (FCS, HyClone), 2 mM GlutaMAX (Gibco), 100 units/mL Penicillin, and 100 ug/mL Streptomycin (Gibco). Cells were cultured at 5% CO_2 and 37°C. Cell lines were tested monthly for mycoplasma contamination.

2.6.3.2 CUT&Tag

Benchtop CUT&Tag was performed as previously described[8]. In brief, Kasumi-1 cells were counted, harvested, and centrifuged for 5 min at 300xg at room temperature. Cells were washed two times in 1.5 mL wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail). Concanavalin A magnetic

coated beads (Bangs Laboratories) were activated in binding buffer by washing two times (20 mM HEPES pH 7.5, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$). Washed cells were separated into 100,000 cell aliquots and 10 ul of activated beads were added to each sample. Samples rotated at room temperature end over end for 7 minutes. Beads were separated with a magnetic and supernatant was removed. 1 µl of primary antibody was diluted 1:50 in antibody buffer (20 mM HEPES pH 7.5, 150mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% digitonin, 2 mM EDTA, 0.1% BSA). The primary antibodies used were: H3K27ac (ab4729, Abcam) and Normal Rabbit IgG (#2729, CST). Cells were incubated overnight at 4°C on a nutator. Primary antibody was replaced with a guinea-pig anti rabbit secondary antibody (Antibodies Online, cat. no. ABIN101961) diluted to 1:100 in wash buffer. Samples were incubated for 45 minutes at room temperature on nutator. Secondary antibody was removed, and samples were washed 2X in dig-wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% Digitonin). pA-Tn5 transposase, prepared and loaded with adaptors as previously described[23], was diluted 1:250 in dig-300 buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.01% digitonin) and added to samples. Samples incubated for 1 hour at room temperature on nutator. Samples were washed 2X with dig-300 buffer then resuspended in tagmentation buffer (dig-300 buffer with 10 mM $MgCl_2$). Samples were incubated at 37°C for 1 hour. DNA was extracted with a DNA Clean & Concentrator-5 kit (ZYMO). Samples were amplified by PCR using custom Nextera primers at 400 nM and NEBNext HiFi 2x PCR Master Mix (New England Biolabs)[186]. PCR conditions were set to: 72°C for 5 minutes, 98°C for 30 seconds, 14-27 cycles of 98°C for 10 sec, 63°C for 10 sec, and 72°C for 1 minute. Libraries were purified with AMPure Beads (Beckman) and sequenced on a NextSeq 500 sequencer (Illumina) using 37 BP PE sequencing by

Massive Parallel Sequencing Shared Resource at Oregon Health and Science University.

2.6.3.3 ATAC-seq

Samples were prepared as previously described[187]. In brief, cells were resuspended tagmentation master mix (25 ul of 2× tagmentation buffer, 2.5 ul of TDE1 [Illumina], 0.5 ul of 1% digitonin; 2x tagmentation buffer: 66 mM Tris-Acetate, pH 7.8, 132 mM potassium acetate, 20 mM magnesium acetate, 32% v/v N,N-Dimethylformamide). Samples were incubated at 37°C for 30 minutes. DNA was purified using Zymo Clean and Concentrator 5 Kit (Zymo). Transposed DNA was amplified and purified as described previously with adapted primers[188,189]. Samples were quantified using Qubit dsDNA HS Assay Kit (Invitrogen), pooled, and sequenced by Genewiz with a HiSeq-X (Illumina) using 75 BP PE sequencing.

2.7 Declarations

2.7.1 Ethics approval and consent to participate

Not applicable.

2.7.2Consent for publication

Not applicable.

2.7.3 Availability of data and materials

GoPeaks is free to use and is publicly accessible on GitHub (https://github.com/maxsonBraunLab/gopeaks)[190]. A stable version of the GitHub repository is available through Zenodo (https://zenodo.org/record/6413077#.YkydK5PMKAI)[191]. Custom scripts used to compare the peak calling algorithms are available in the gopeaks-compare repository (https://github.com/maxsonBraunLab/gopeaks-compare)[192]. The raw and processed sequencing datasets generated during the current study are available in the NCBI Gene Expression Omnibus (GEO) under accession number GSE190793[175,193]. The K562 H3K4me3 (GEO accession GSM3536516), H3K4me1 (GEO accession GSM3536518), H3K27me3 (GEO accession GSM3560261), and IgG (GEO accession GSM3560264) CUT&Tag datasets from Kaya-Okur et al. 2019 analyzed during the current study are available in the NCBI GEO[23,175,194–197]. The K562 H3K4me3 (ENCODE ID ENCFF246IEW), H3K4me1 (ENCODE ID ENCFF590NGQ), and H3K27me3 (ENCODE ID ENCFF795ZOS) ChIP-seq datasets analyzed during the current study are available in ENCODE[171,176,198–201]. The Kasumi-1 H3K27ac (ChIP-Atlas SRX ID SRX4143063 and SRX4143067) ChIP-seq data analyzed during the current study is available in ChIP-Atlas[172,185,202,203].

2.7.4 Competing interests

WMY potential competing interests – WMY is a former employee of Abreos Biosciences, Inc. and was compensated in part with common stock options. Pursuant to the merger and reorganization agreement between Abreos Biosciences, Inc. and Fimafeng, Inc., WMY surrendered all of his common stock options in 03/2021. JEM --

SAB: Ionis pharmaceuticals, Research Funding: Gilead Sciences. The other authors do not have competing interests, financial or otherwise.

2.7.5 Funding

Funding was provided by an American Society of Hematology Research Restart Award, an American Society of Hematology Scholar Award and 1 K08 CA245224 from NCI awarded to TPB. WMY was supported by NIH T32 GM109835: Medical Scientist Training Program of Oregon Health and Science University.

2.7.6 Authors' contributions

BMC and DJC identified the need for a new peak caller. JV quantitatively described the problem and wrote GoPeaks. WMY, GLK, JV, BMC, and TPB designed the research plan. BMC performed sample and library preparation. GLK designed and executed the benchmarking workflow. WMY, GLK, JV, BMC, GGY, and TPB analyzed data. WMY wrote the manuscript. WMY, BMC, TPB, LC, GGY, and JEM reviewed and edited the manuscript. TPB and JEM acquired funding and supervised the project. All co-first authors may identify themselves as lead authors in their respective CVs.

2.8 Supplementary figures

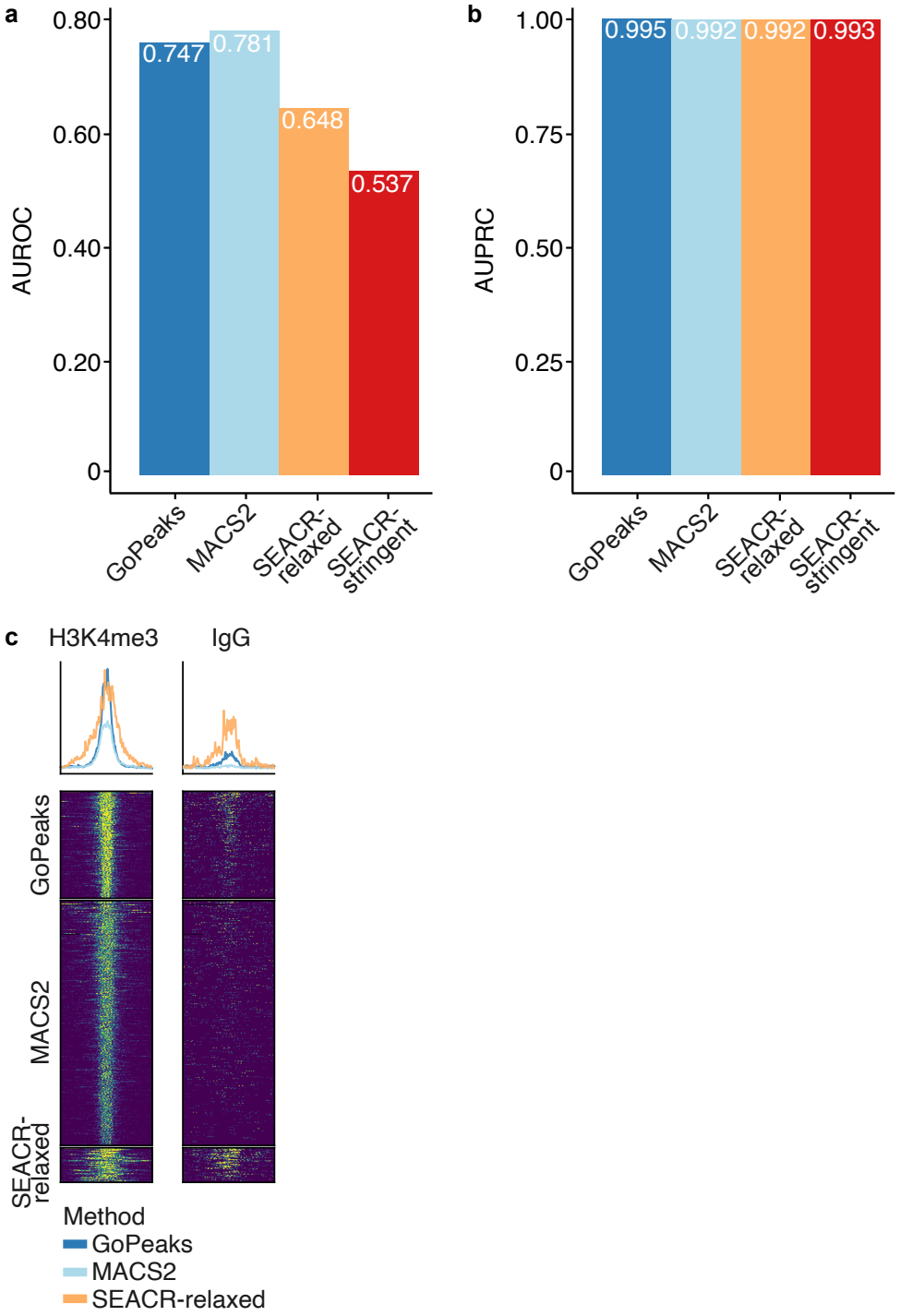

Figure S2.1: GoPeaks demonstrates comparable sensitivity and specificity in identifying H3K4me3 ChIP-seq standard peaks from CUT&Tag data.

a, b. Area under the curve (AUC) of a. ROC and b. PRC from Figure 2.4 for each peak calling method. Each bar is labeled by the value it represents. Colors indicate the peak calling method. **c.** Heatmap of global signal from unique H3K4me3 CUT&Tag peaks identified by each method in Figure 2.4.

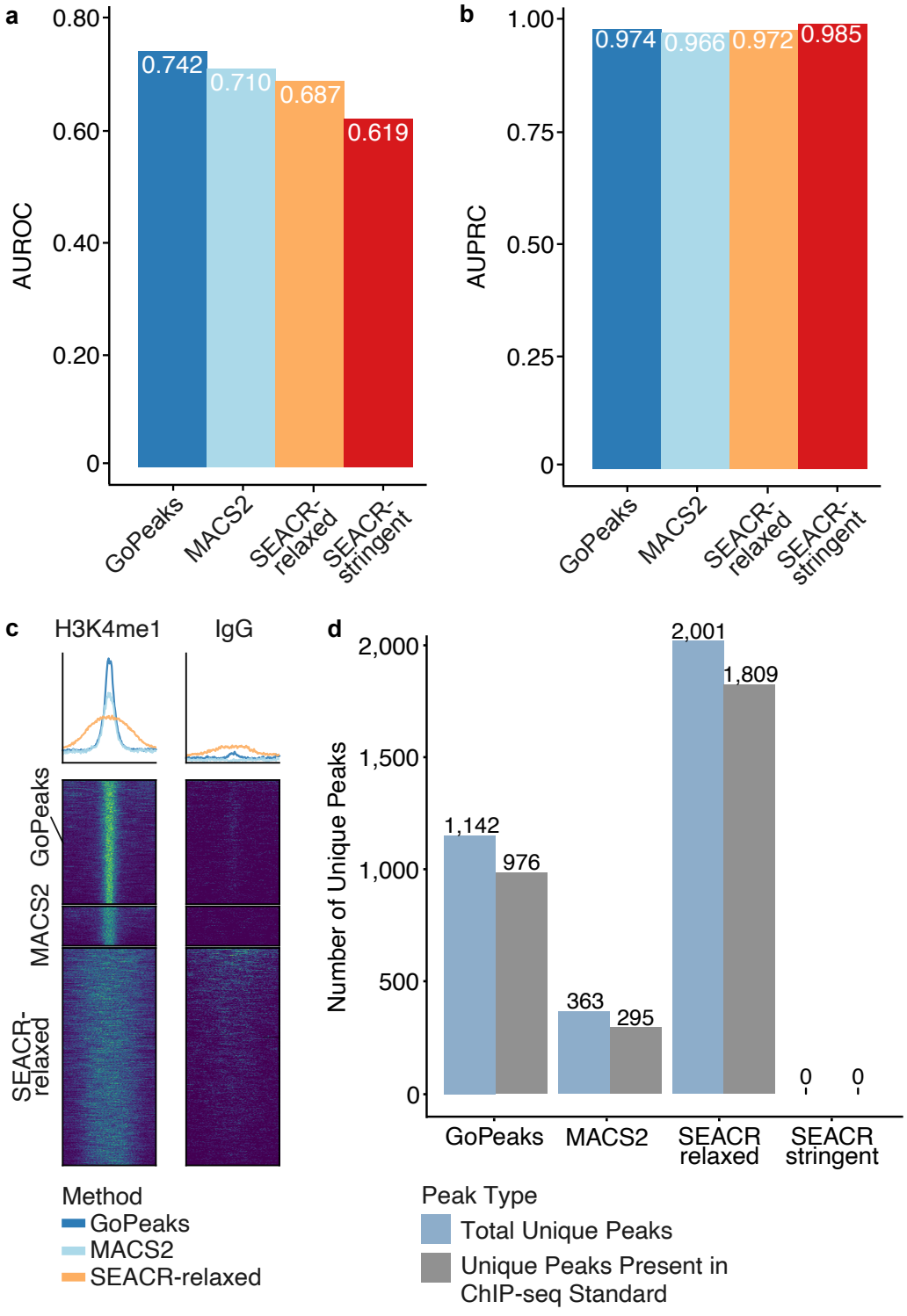

Figure S2.2: GoPeaks demonstrates comparable sensitivity and specificity in identifying H3K4me1 ChIP-seq standard peaks from CUT&Tag data.

a, b. AUC of a. ROC and b. PRC from Figure 2.5 for each peak calling method. Each bar is labeled by the value it represents. Colors indicate the peak calling method. **c.** Heatmap of global signal from unique H3K4me1 CUT&Tag peaks identified by each method in Figure 2.5. **d.** Comparison of unique peaks that are identified by each peak calling algorithm and are also present in the ChIP-seq standard. Each bar is labeled by the number of peaks it represents. Colors indicate the peak type.

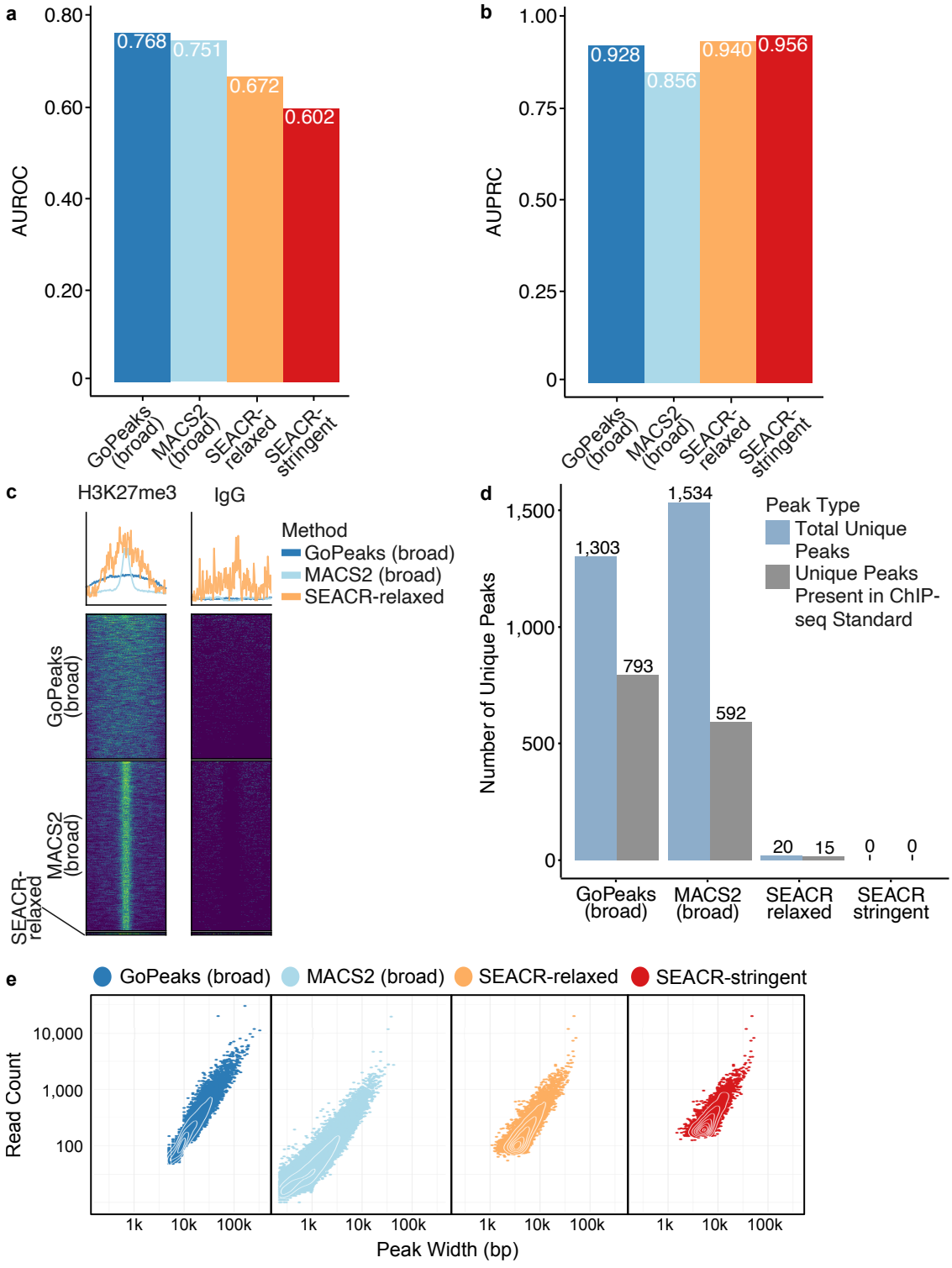

Figure S2.3: GoPeaks demonstrates improved sensitivity and specificity in identifying H3K27me3 ChIP-seq standard peaks from CUT&Tag data.

a, b. AUC of a. ROC and b. PRC from Figure 2.6 for each peak calling method. Each bar is labeled by the value it represents. Colors indicate the peak calling method. **c.** Heatmap of global signal from unique H3K27me3 CUT&Tag peaks identified by each method in Figure 2.6. **d.** Distribution of read counts by peak width. Each dot represents the read count and peak width of a single detected peak. Colors indicate the peak calling method. **e.** Comparison of unique peaks identified by each peak calling algorithm and how many are also present in the ChIP-seq standard. Each bar is labeled by the number of peaks it represents. Colors indicate the peak type.

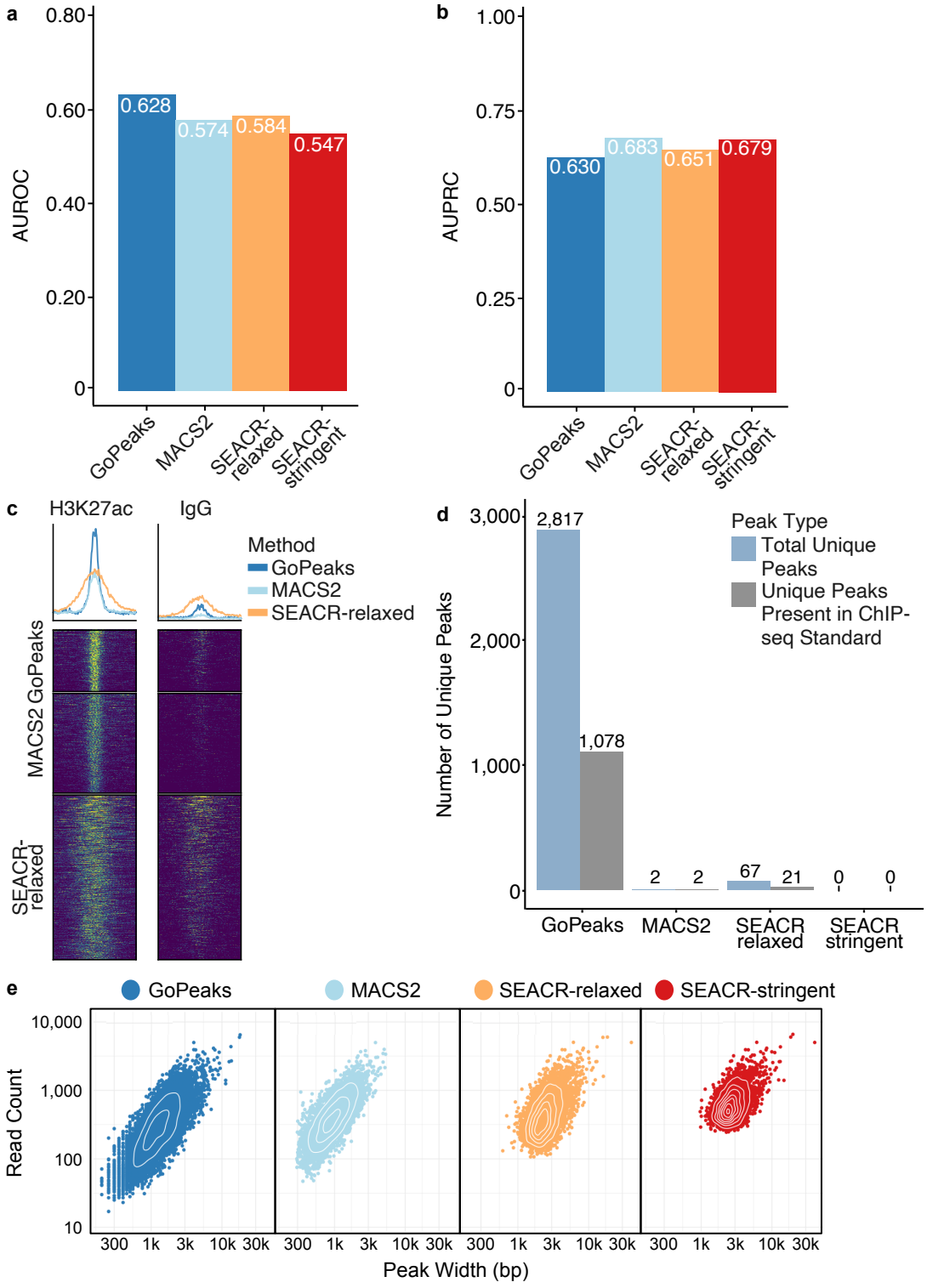

Figure S2.4: GoPeaks demonstrates improved sensitivity and specificity in identifying H3K27ac ChIP-seq standard peaks from CUT&Tag data.

a, b. AUC of a. ROC and b. PRC from Figure 2.7 for each peak calling method. Each bar is labeled by the value it represents. Colors indicate the peak calling method. **c.** Heatmap of global signal from unique H3K27ac CUT&Tag peaks identified by each method in Figure 2.7. **d.** Comparison of unique peaks identified by each peak calling algorithm and how many are also present in the ChIP-seq standard. Each bar is labeled by the number of peaks it represents. Colors indicate the peak type. **e.** Distribution of read counts by peak width. Each dot represents the read count and peak width of a single detected peak. Colors indicate the peak calling method.

Chapter 3: Predicting transcription factor activity using prior biological information

---

- We have developed Priori, a method to predict transcription factor activity from RNA sequencing data.
- Priori utilizes literature-supported regulatory information to identify transcription factor-target relationships. It then applies linear models to determine the impact of transcription factor regulation on the expression of its target genes.
- Results from a third-party benchmarking pipeline reveals that Priori detects aberrant activity from 124 gene perturbation experiments with higher sensitivity and specificity than 11 other methods.
- Our work demonstrates that Priori uniquely discovered significant determinants of survival in breast cancer and identified mediators of drug response in leukemia.

This work has been published in *Cell iScience*:

Yashar WM, Estabrook J, Holly HD, Somers J, Nikolova O, Babur Ö, Braun TP, Demir E. Predicting transcription factor activity using prior biological information. *Cell iScience* 2024 January 31[204]

3.1 Abstract

Dysregulation of normal transcription factor activity is a common driver of disease. Therefore, it is important to detect aberrant transcription factor activity to better understand disease pathogenesis. We have developed Priori, a method to predict transcription factor activity from RNA sequencing data. Priori has two key advantages over existing methods. First, Priori utilizes literature-supported regulatory information to identify transcription factor-target relationships. It then applies linear models to determine the impact of transcription factor regulation on the expression of its target genes. Second, results from a third-party benchmarking pipeline reveals that Priori detects aberrant activity from 124 gene perturbation experiments with higher sensitivity and specificity than 11 other methods. We applied Priori and other top-performing methods to predict transcription factor activity from two large primary patient datasets. Our work demonstrates that Priori uniquely discovered significant determinants of survival in breast cancer and identified mediators of drug response in leukemia.

3.2 Introduction

The coordinated expression and activity of transcription factors are fundamental mechanisms in establishing and maintaining cell identity and function. Transcription factors are key regulatory proteins that bind to *cis*-regulatory DNA sequences, including promoters and enhancers, and modulate gene transcription[44,45]. Dysregulation of these normal transcription factor functions frequently contributes to the development of a pathogenic cell phenotype[46,47]. Abnormal transcription factor activity can result from mutations in the putative *cis*-regulatory DNA binding sequences or in the transcription factors themselves. Recent studies have highlighted the importance of aberrant

expression of pathogenetic transcription factors as drivers of disease[47]. *MYC*, which is important for cellular growth and proliferation, is the most frequently amplified oncogene. Elevated levels of *MYC* has been shown to promote tumorigenesis in a variety of tissue types[47]. In tumor cells expressing high levels of *MYC*, the transcription factor accumulates in *cis*-regulatory regions of genes associated with cellular proliferation and growth, resulting in transcriptional amplification of MYC's gene regulatory network and, subsequently, abnormal cellular proliferation[48,49]. Therefore, detection of abnormal transcription factor activity is valuable for better understanding the mechanisms that underly disease pathogenesis.

Gene expression profiling, including RNA sequencing (RNA-seq), is commonly used to monitor dynamic changes in transcription factors and their gene regulatory networks. Initial studies to infer transcription factor activity only used transcription factor gene expression as a proxy for activity[50–52]. However, this approach has several shortcomings. Gene expression is only an indirect measurement of protein activity due to the complex mechanisms controlling protein synthesis and degradation[53–55]. Feedback loops may alter the expression of transcription factors in response to their regulatory activity[47,56–58]. Reliable predictions of transcription factor activity, therefore, cannot be limited to evaluating transcription factor expression alone.

An alternative approach to inferring transcription factor activity is to assess the expression of their downstream target genes[50–52]. This approach has two major benefits. First, evaluation of hundreds or thousands of downstream targets instead of a single transcription factor likely improves the prediction robustness. While some of these targets may be context-specific, analyzing them in aggregate likely improves the prediction generalizability across many contexts. Second, as target gene expression is downstream of transcription factor control, these signatures are expected to reflect the actual transcriptional impact more accurately. Therefore, accounting for the downstream

impact of transcription factors on its gene regulatory networks is important for activity inference.

Multiple methods have been developed to quantify transcription factor activity from gene expression data. These approaches can be grouped based on how they select gene expression features. Methods like Univariate Linear Model (ULM) and Multivariate Linear Model (MLM) use every gene in a dataset, nominating transcription factors as a covariate that best estimates the expression of all other genes[59]. However, these methods develop activity signatures using genes that may not have a true biological relationship to the transcription factor of interest. Gene set approaches like Over Representation Analysis (ORA), Fast Gene Set Enrichment (FGSEA), Gene Set Variation Analysis (GVSA), and AUCell infer activity using sets of published transcription factor target genes or target genes curated by experts[60–63]. While gene set methodologies are simple and popular, they are susceptible to the quality and comparability of gene set signatures[64]. Network inference approaches, including Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe), infer gene regulatory networks based on the covariance of transcription factors and its putative targets[65,66]. This process is not completely unsupervised, however, as ARACNe requires a user-defined list of transcription factors in order to infer gene regulatory networks[65,66]. The same group that developed ARACNe also created Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) to infer transcription factor activity from ARACNe gene expression signatures[67]. The challenge with these approaches is deconvoluting combinatorial regulation, where the expression of a target gene is controlled by multiple transcription factors. While some of these methods, including VIPER, have an option to correct for this, it remains difficult to infer transcriptional networks as there are many possible solutions that can explain the underlying data[50,68]. While these methods deploy various techniques to generate activity scores using the

expression of downstream target genes, most do not select their target gene features from literature-supported transcriptional relationships.

Recent studies have highlighted that grounding predictions using transcription factor activity methods remains challenging[68–72]. A rigorous evaluation of widely-used transcription factor algorithms demonstrated that most methods do not robustly detect perturbed transcription factors[63]. Despite this, there are precision medicine clinical studies that use inferred transcription activity from bulk RNA-seq as a marker to guide clinical decisions. While there is an increasing number of single-cell and spatial -omic modalities available to clinical researchers, these studies as well as many larger cohorts and clinical trials most commonly use markers identified from bulk sequencing of RNA or DNA. Therefore, it is critical to develop methods that can robustly detect aberrant transcription factor activity from primary patient bulk RNA-seq data.

Here, we propose an approach that uses prior, peer-reviewed biological information to infer transcription factor activity called Priori. Our method has two major advantages over the existing methods. First, Priori identifies transcription factor target genes using carefully extracted transcriptional regulatory networks from Pathway Commons[205,206]. This resource continually collects information on biological pathways including molecular interactions, signaling pathways, regulatory networks, and DNA binding. Pathway Commons currently contains data from 22 high-quality databases with over 5,700 detailed pathways and 2.4 million interactions. Using the transcriptional relationships from Pathway Commons, Priori fits linear models to the expression of transcription factors and their target genes. These models allow Priori to understand the impact and direction of transcription factor regulation on its known target genes. Second, comparison with a third-party benchmarking workflow reveals that Priori detects aberrant transcription factor activity from 124 gene perturbation experiments with higher sensitivity and specificity than 11 other methods. We applied Priori and three other

methods nominated from the benchmarking workflow to generate activity scores from two large primary patient datasets, TCGA BRCA and Beat AML. We demonstrate that Priori can be deployed to discover significant predictors of survival in breast cancer as well as identify mediators of drug response in leukemia from primary patient samples that were not robustly detected using the other methods.

3.3 Results

3.3.1 Priori uses prior biological information to infer transcription factor activity

For each transcription factor in an RNA-seq dataset, Priori generates an activity score (Figure 3.1A). The activity score is a weighted, aggregate statistic that reflects the impact and direction of transcription factor regulation on its target genes. Priori first identifies the known target genes for each transcription factor in a dataset from Pathway Commons (or another network provided by the user)[205,206]. Priori then assigns weights to each target gene by correlating the target gene expression to its transcription factor (Equation 5). To identify the targets that are most impacted by transcription factor regulation, Priori separates the up- and down-regulated genes by their transcription factor-target gene weights and ranks their expression (Equation 6). These ranks are subsequently scaled by multiplying it with the transcription factor-target gene weights. The activity score for each transcription factor is then calculated by summing the weighted ranks (Equation 7). With this single-component model, researchers can use Priori to predict transcription factor activity from RNA-seq data.

In order to compare how well Priori detects aberrant transcription factor activity to other methods, we used a third-party benchmarking workflow called decoupleR (Figure

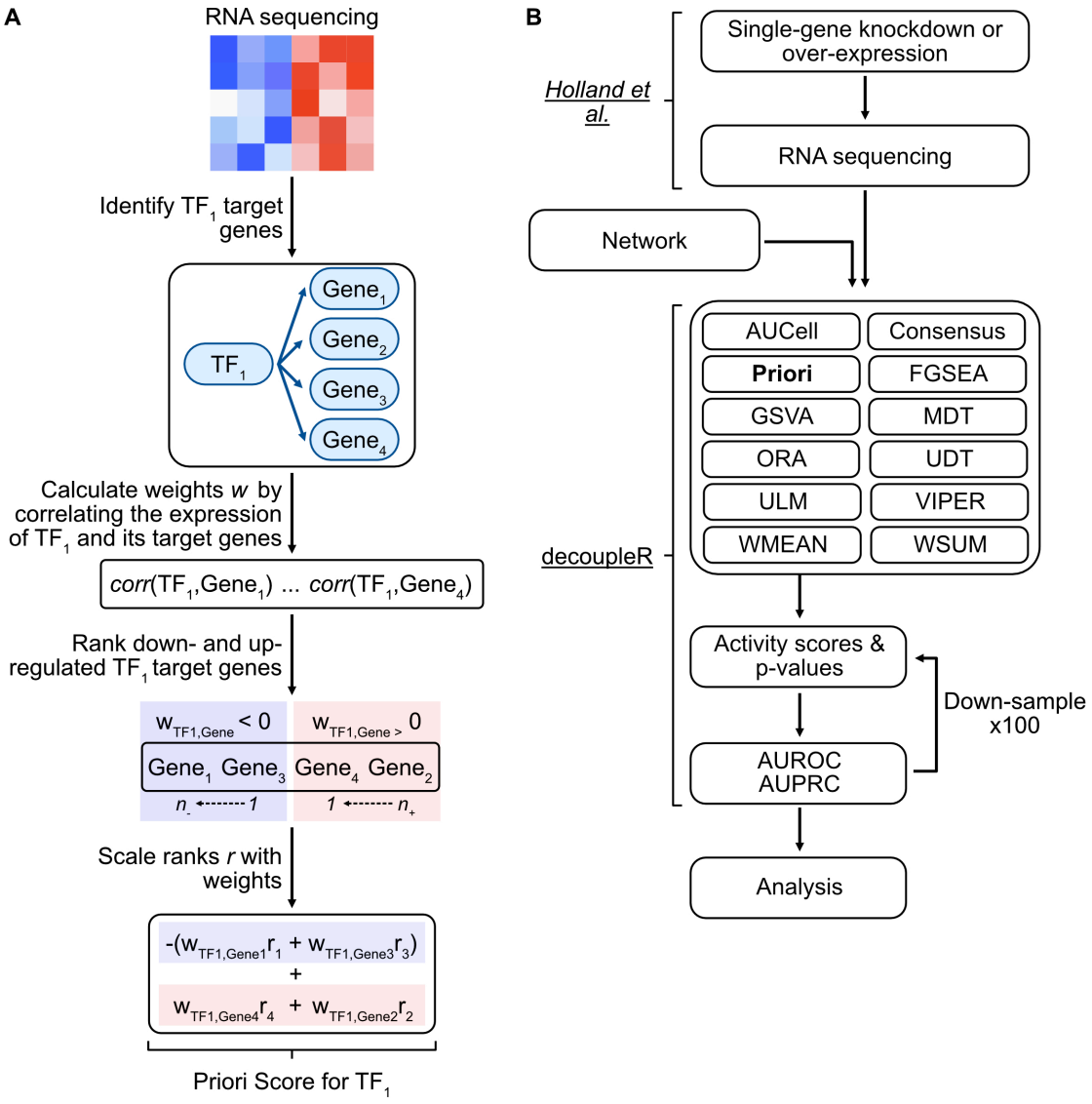

Figure 3.1: Overview of the Priori methodology and benchmarking workflow.

(A) Priori generates an activity score for each transcription factor in an RNA-seq dataset. Priori first extracts the downstream target genes for each transcription factor from Pathway Commons. Priori then calculates weights for each target gene by correlating the expression of each transcription factor to its target genes. Priori then ranks the absolute expression of all genes in the dataset and scales these ranks by the transcription factor-target gene weights. The summation of the weighted target gene ranks is the transcription factor activity score. (B) Schematic overview of the benchmarking workflow. We generated transcription factor activity scores for each method using normalized RNA-seq counts following single-gene knockdown or over-expression.

Priori, along with other methods that use prior information, generated activity scores with transcriptional relationships from Pathway Commons. AUROC and AUPRC values were calculated for each down-sampling permutation. 100 down-sampling permutations were performed to compare an equal number of perturbed and unperturbed genes.

3.1B)[63]. The decoupleR workflow facilitated an unbiased, common evaluation scheme to determine how often each method correctly identified transcription factors that have been knocked-down or over-expressed from RNA-seq data, resembling the pathologic disruption of normal transcription factor regulation[207]. Using this workflow, we generated transcription factor activity scores for 16 methods, including Priori, using the default parameters (Table 3.1). Several of these methods were developed by the authors of the decoupleR workflow, including the standard, normalized (Norm), and corrected (Corr) versions of Weighted Mean (WMEAN), Weighted Sum (WSUM), Univariate Decision Tree (UDT), and Multivariate Decision Trees (MDT)[63]. These authors also developed a normalized version of FGSEA. For methods that use prior information, we generated activity scores using the Pathway Commons transcriptional relationships. We ranked the transcription factor activity scores for each experiment and compared the ranks of perturbed and unperturbed transcription factors. We evaluated how often the perturbed transcription factor activity score was among the top activity scores in each experiment. The authors of the decoupleR pipeline defined this threshold by the number of perturbed transcription factors in the dataset. Since the number of unperturbed transcription factors vastly outnumbered the perturbed transcription factors, we implemented a down-sampling strategy. For each down-sampling permutation, we calculated area under the precision recall curve (AUPRC) and receiver operating characteristic (AUROC) metrics. The results of this third-party benchmarking workflow allowed us to objectively compare the sensitivity and specificity of Priori to detect perturbed transcription factor activity to 11 other methods.

Table 3.1: Transcription factor activity methods evaluated using the decoupleR benchmarking workflow.

Feature Selection	Method	Acronym	Normalized Method	Corrected Method	Reference
All features	Univariate Linear Model	ULM	-	-	Teschendorff and Wang *npj genomic medicine* (2020)[59]
	Multivariate Linear Model	MLM	-	-	Teschendorff and Wang *npj genomic medicine* (2020)[59]
Gene set	Over Representation Analysis	ORA	-	-	Badia-i-Mompel et al. *Bioinformatic Advances* (2022)[63]
	Fast Gene Set Enrichment	FGSEA	Norm FGSEA	-	Korotkevich et al. *bioRxiv* (2021)[60]
	Gene Set Variation Analysis	GVSA	-	-	Hänzelmann et al. *BMC Bioinformatics* (2013)[61]
	AUCell	-	-	-	Aibar et al. *Nature*

					Methods (2017)[62]
Inferred networks	Virtual Inference of Protein-activity by Enriched Regulon analysis	VIPER	-	-	Alvarez et al. *Nature Genetics* (2016)[67]
Curated gene networks	Weighted Mean	WMEAN	Norm WMEAN	Corr WMEAN	Badia-i-Mompel et al. *Bioinformatic Advances* (2022)[63]
	Weighted Sum	WSUM	Norm WSUM	Corr WSUM	Badia-i-Mompel et al. *Bioinformatic Advances* (2022)[63]
	Univariate Decision Tree	UDT	-	-	Badia-i-Mompel et al. *Bioinformatic Advances* (2022)[63]
	Multivariate Decision Trees	MDT	-	-	Badia-i-Mompel et al. *Bioinformatic Advances* (2022)[63]

ID: Norm = normalized; Corr = corrected.

3.3.2 Priori identifies perturbed transcription factors with greater sensitivity and specificity than other methods

The decoupleR workflow evaluates transcription factor activity methods using a curated dataset from Holland et al.[207] This dataset includes RNA-seq data from 94 knockdown and 30 over-expression experiments of 62 different transcription factors. Before using this dataset to compare transcription factor activity methods, we wanted to evaluate whether transcription factor knockdown resulted in decreased normalized expression and over-expression resulted in increased normalized expression. Moreover, since Priori and other methods use the transcription factor expression to generate activity scores, we wanted to understand the extent to which transcription factor expression alone could predict gene perturbation status. To assess this, we compared the normalized expression of perturbed transcription factors to unperturbed genes. We found that the normalized expression of knocked-down transcription factors was significantly less than unperturbed genes (Figure S3.1A). While over-expressed transcription factors were associated with a higher normalized expression, it was not significantly different than unperturbed genes (Figure S3.1B). This data indicates that while transcription factor expression is a reasonable indicator of gene perturbation, other features are needed to confidently predict aberrant activity.

Using the decoupleR workflow, we generated activity scores for 16 transcription factor activity methods using RNA-seq data from the 124 gene perturbation experiments. To understand the patterns in predicted activity across the methods, we correlated the activity scores (Figure 3.2A). The Priori activity scores were dissimilar from the other methods, indicating that Priori identified a unique pattern of transcription factor activity. To assess how well the methods detected perturbed transcription factors, we calculated

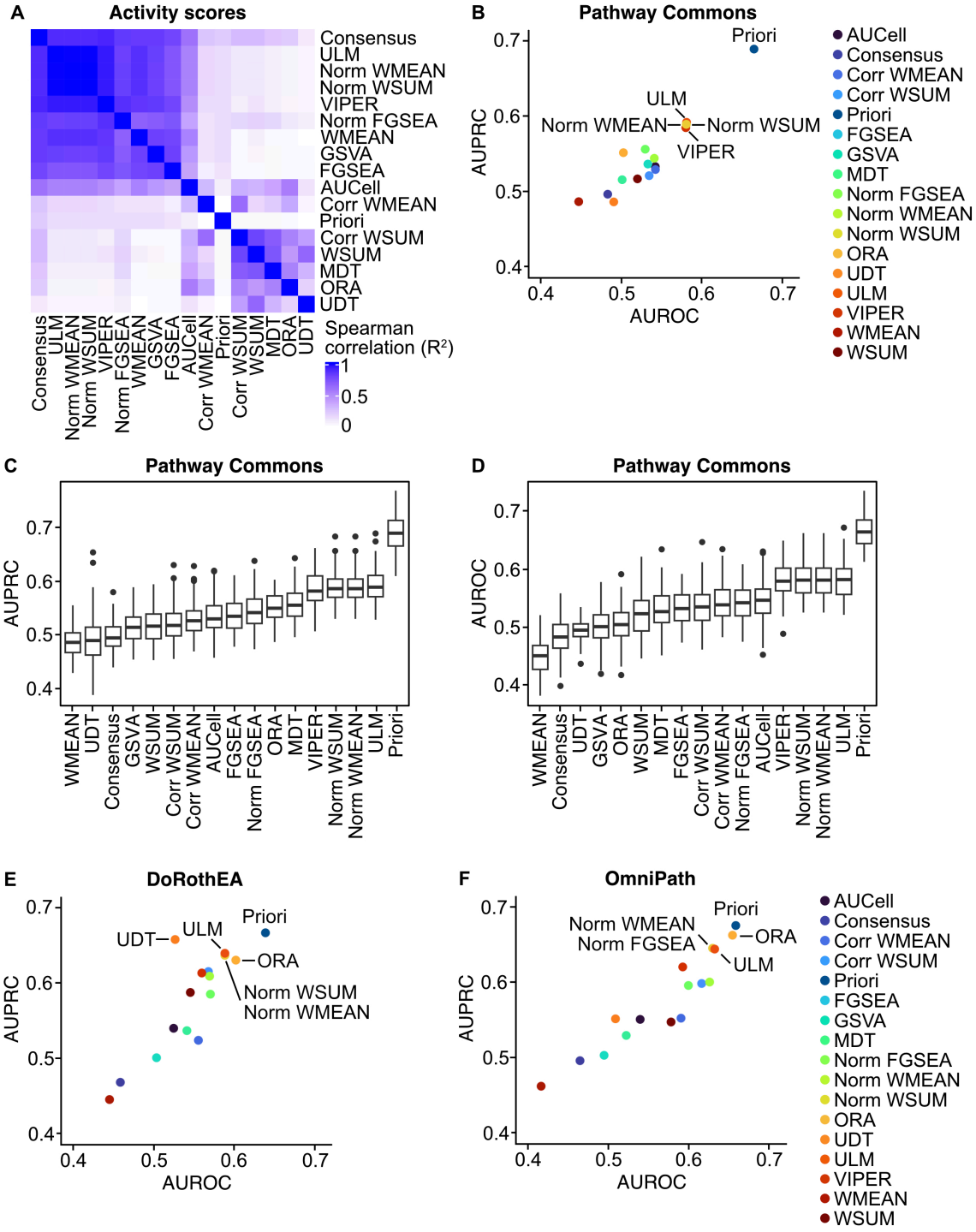

Figure 3.2: Priori detects aberrant transcription factor activity with improved sensitivity and specificity.

(A) Using the decoupleR workflow, transcription factor activity scores were generated using the perturbation dataset. Spearman correlation of the activity scores for each method.

(B) Activity scores were generated using Pathway Commons transcriptional relationships. Mean AUPRC and AUROC values across the 100 down-sampling permutations for each method. (C and D) The distribution of (C) AUPRC and (D) AUROC values across the 100 down-sampling permutations from (B). (E and F) Activity scores were generated using (E) DoRothEA or (F) OmniPath transcriptional relationships. Mean AUPRC and AUROC values across the 100 down-sampling permutations for each method.

AUPRC and AUROC metrics for each down-sampling permutation (Figure 3.2B-D). Priori had a greater AUPRC and AUROC values than the other methods across all experiments. ULM, Norm WMEAN, Norm WSUM, and VIPER were the next closest methods by AUPRC and AUROC. Collectively, these experiments show that Priori detects perturbed transcription factors with greater sensitivity and specificity than other methods.

While our analyses showed that Priori can accurately identify perturbed transcription factors, we wanted to understand the extent to which prior information from other datasets impacted its performance. Pathway Commons is one of several large databases that curates transcriptional relationships. DoRothEA is a comprehensive resource, which assembles transcription factor-target gene relationships from ChIP-seq peaks, inferred regulatory networks (including ARACNe), transcription factor binding motifs, and literature-curated resources[208]. OmniPath, on the other hand, integrates intra- and inter-cellular signaling networks in addition to transcriptional relationships from 100 different resources, including DoRothEA and Pathway Commons[209]. To evaluate the impact of prior information on each method, we used the decoupleR workflow to generate activity scores using transcription factor-target gene interactions from DoRothEA or OmniPath instead of Pathway Commons. We used the default parameters for each method, including a lower limit of 15 downstream targets for the Priori analyses. We found that in both instances, Priori had higher AUPRC and AUROC values than the other methods (Figure 3.2E and F; Figure S3.1C-F). Priori exhibited similar AUPRC and

AUROC values when using transcriptional relationships from DoRothEA, OmniPath, or Pathway Commons. However, Priori had the highest mean AUPRC and AUROC values using the Pathway Commons transcriptional relationships. Our analyses demonstrated that regardless of the prior transcriptional relationship information, Priori detected perturbed transcription factor activity with improved sensitivity and specificity.

We evaluated how transcriptional relationships from Pathway Commons, DoRothEA, and OmniPath affected the performance of each method. However, these databases do not provide the appropriate information for methods, including VIPER, that were designed to generate activity scores from networks with signed edges that are weighted by likelihood. While decoupleR does not allow for networks with signed edges, we wanted to understand how VIPER's performance changed with ARACNe-generated networks with likelihood-weighted edges[63]. The authors of ARACNe have published cancer-type-specific networks trained on TCGA RNA-seq datasets[66]. We identified the cancer types and organ sites associated with each cell line tested in the perturbation dataset (Figure S3.2A-C). With decoupleR, we generated VIPER activity scores using any TCGA ARACNe network that was trained on a cancer type that was also evaluated in the perturbation dataset. We observed that VIPER had greater AUPRC and AUROC values when using the transcriptional relationships from Pathway Commons than those from the TCGA ARACNe networks (Figure S3.2D). ARACNe can also be used to reverse engineer gene regulatory networks from a RNA-seq dataset[66]. Using the new implementation of ARACNe, ARACNe-AP, we inferred transcriptional relationships using the perturbation RNA-seq data. We observed improved AUPRC and AUROC values when VIPER used ARACNe-AP transcriptional relationships as prior information (Figure S3.2E). However, these AUPRC and AUROC values were still not greater than Priori using prior transcriptional relationships from Pathway Commons.

3.3.3 Priori's predictions are robust to noise

We demonstrated that Priori detects perturbed transcription factors with improved sensitivity and specificity, particularly when it used Pathway Commons transcriptional relationships as prior information. However, we wanted to understand how robust these predictions are to noise. For methods that use prior information, the major sources of noise are introduced in the gene expression data and the prior transcriptional relationships. First, to evaluate the effect of noisy gene expression data, we introduced increasing amounts of zero-centered, Gaussian-distributed noise to the Holland et al. perturbation dataset. We evaluated the accuracy of perturbed transcription factor predictions using the decoupleR pipeline as described above. We observed a drop in AUPRC and AUROC in most methods when more than one standard deviation of gaussian-centered noise was introduced (Figure S3A and B). Notably, Priori could still accurately identify perturbed transcription factors even when five standard deviations of noise was added. Second, to understand the impact of noisy prior information, we gradually removed the number of transcription factor-target gene relationships from the prior Pathway Commons network. We observed that Priori and other methods that use prior information had similar AUPRC and AUROC values as long as 20% of the network was retained (Figure S3C and D). It was unclear, however, whether this pattern was due to the diminished number of transcriptional relationships or to the masking of true relationships. To test this, we randomized the target genes in the Pathway Commons by sampling with replacement any feature in the gene expression dataset. We observed a drop in prediction accuracy when more than 60% of the target genes were randomized (Figure S3E and F). These analyses demonstrate that Priori is robust to artificial noise introduced to the gene expression data and can accurately identify perturbed transcription factors as long as 60% of the Pathway Commons network is retained.

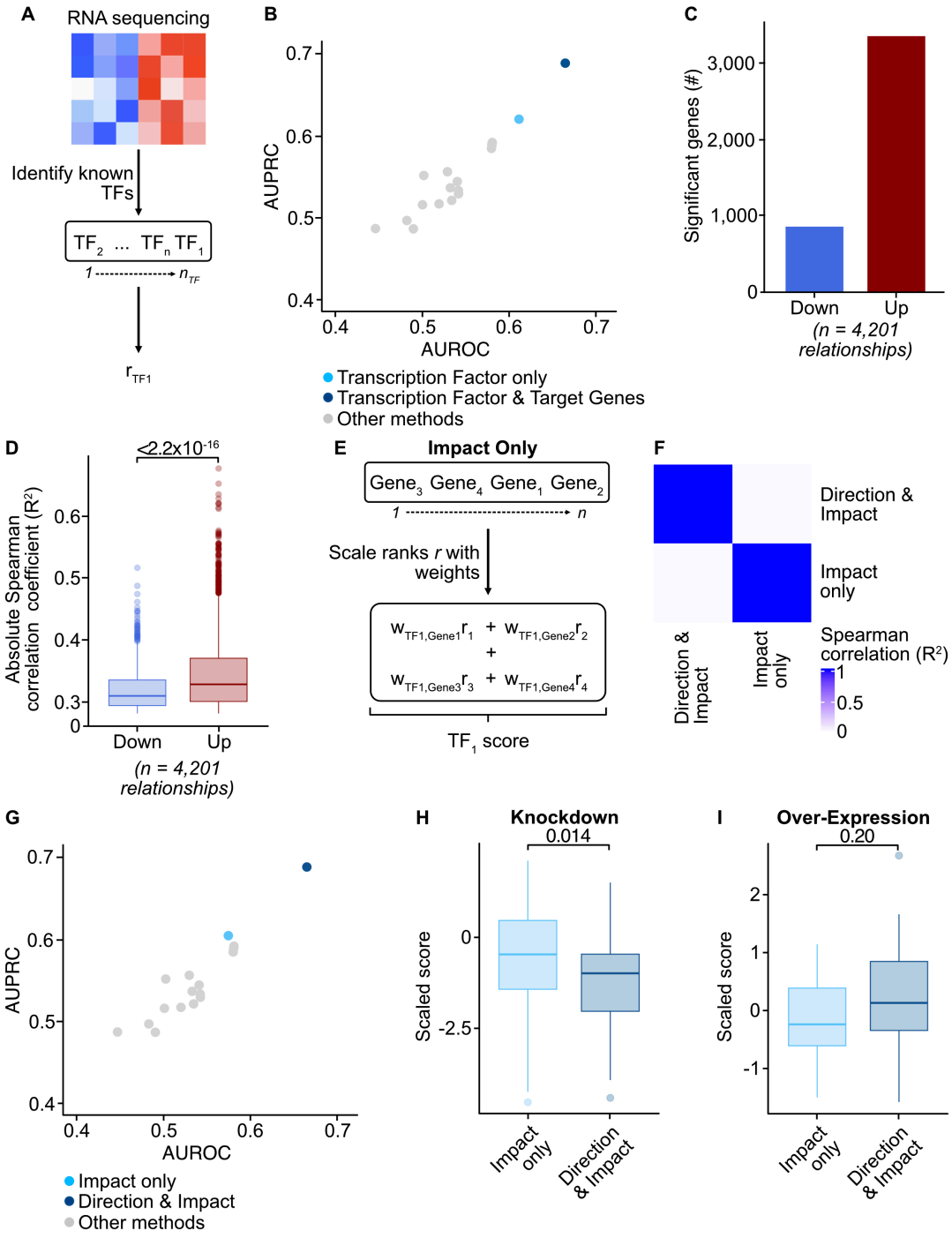

Figure 3.3: Evaluation of the direction and impact of transcriptional regulation is critical for Priori to detect aberrant transcription factor activity.

(A) Schematic showing how transcription factor activity scores were generated using transcription factor expression only (in contrast to both transcription factor and target gene expression as shown in Figure 3.1A).

The alternative method first identified transcription factors in the perturbation dataset from Pathway Commons. The method then ranks the transcription factors by expression and reports the normalized rank as the activity score. (B) Activity scores were generated using the method outlined in (A). Transcriptional relationships from Pathway Commons were used as prior information. Mean AUPRC and AUROC values across the 100 down-sampling permutations. Mean AUPRC and AUROC values from the methods in Figure 3.2B are also shown. (C) Priori identified transcription factor target genes in the perturbation dataset using Pathway Commons transcriptional relationships. The expression of transcription factors and their target genes were evaluated using Spearman correlation. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. The Spearman correlation coefficient was used to determine down-regulated ($R^2 < 0$) and up-regulated target genes ($R^2 > 0$). (D) Absolute Spearman correlation coefficient of the expression of transcription factors and their down-regulated or up-regulated target genes. Statistical significance was determined by a two-sided Student's t-test. (E) Schematic showing how Priori was altered to assess only the impact of transcriptional regulation (in contrast to both direction and impact of regulation as shown in Figure 3.1A). (F) Using the decoupleR workflow, transcription factor activity scores were generated using the perturbation dataset. Spearman correlation of the activity scores for each method.

3.3.4 Priori's improved performance is due to evaluating the direction of transcriptional regulation

We evaluated how Priori's prediction accuracy was affected by artificial noise. However, it was unclear what enabled Priori to detect aberrant activity better than other methods. We previously demonstrated that transcription factor expression is a reasonable indicator of gene perturbation. We wanted to understand the extent to which assessment of transcription factor expression enabled Priori to detect perturbed transcription factors. We designed a variant of Priori that only uses transcription factor expression to infer transcription factor activity (Figure 3.3A). Like Priori, this method first identifies known transcription factors in an RNA-seq dataset from Pathway Commons.

While this alternative method also deploys rank-based analyses, the activity score is the normalized rank of the expression of a given transcription factor relative to others in a dataset (Equation 8). We used the decoupleR benchmarking pipeline to evaluate how well this alternative method detects aberrant transcription factor activity using known transcription factors from Pathway Commons. While this alternative method had a higher AUROC and AUPRC than the other methods, Priori still detected the perturbed transcription factors with higher sensitivity and specificity (Figure 3.3B). This data implies that while assessment of transcription factor expression is important for detection of aberrant activity, performance is improved when target gene expression is evaluated as well.

While Pathway Commons does not provide signed relationships, we designed Priori to infer the impact and direction of transcription factor regulation on its target genes. Priori correlates the expression of transcription factors and their known target genes. Priori subsequently assigns the impact and direction of transcriptional regulation as the sign and coefficient of the Spearman correlation, respectively. Using the sign of the Spearman correlation, we observed that Priori inferred 2,507 more significant up-regulated targets than down-regulated target genes in the perturbation dataset (Figure 3.3C). Using the coefficient of the Spearman correlation, we observed that Priori predicted that transcription factors had a significantly greater impact on its up-regulated targets than its down-regulated target genes (Figure 3.3D). These analyses show that there are important directions of transcription factor regulation on their target genes.

To understand how assessment of the direction of transcriptional regulation is important for detecting aberrant transcription factor activity, we created another variant of Priori that only evaluates the absolute impact of transcriptional regulation (Figure 3.3E). Like Priori, this method deploys a rank-based analysis to evaluate the expression of transcription factors and their target genes (Equation 9). However, this alternative

method does not delineate target genes by their direction of transcriptional regulation, which is indicated by the sign of the Spearman correlation. We used the decoupleR pipeline to generate activity scores using the Pathway Commons transcriptional relationships as prior information. These activity scores were highly dissimilar to the scores generated by Priori that evaluated both the direction and impact of transcriptional regulation ($R^2 = 0.013$; Figure 3.3F). Moreover, when Priori only uses the impact of transcriptional regulation to detect perturbed transcription factors, its AUPRC and AUROC values are less than when it uses both the direction and impact (Figure 3.3G). Accounting for direction of transcriptional regulation likely allows Priori to detect knocked-down and over-expressed transcription factors. We observed an expected decrease in scaled scores across all knockdown experiments when Priori evaluates both the direction and impact of transcriptional regulation (Figure 3.3H). Scaled scores less than zero indicate that the activity of a transcription factor is down-regulated compared to all other perturbed and unperturbed transcription factors in the dataset (and vice versa). These scaled scores are significantly less than the impact-only method. Consistently, Priori predicted an expected increase in activity to over-expressed transcription factors (Figure 3.3I). While these scores are not statistically different to the scores from the impact-only method, the mean predicted activity from the impact-only scores was less than zero. Overall, these analyses demonstrate that assessment of the direction and impact of transcriptional regulation allows Priori to detect aberrant transcription factor activity with improved sensitivity and specificity.

3.3.5 FOXA1 transcription factor activity is a significant determinant of survival among patients with breast cancer

Since we demonstrated that Priori identifies perturbed regulators with greater sensitivity and specificity than other methods, we sought to determine whether Priori could be used to understand transcription factor drivers of disease. We used Priori to generate transcription factor activity scores for 637 patients with invasive breast ductal carcinoma (BIDC) from the TCGA BRCA cohort[210]. To understand the impact of prior information on these scores, we generated scores using transcriptional relationships not only from Pathway Commons, but from DoRothEA and Omnipath as well. In addition, we compared these predicted scores to activity scores from the three top methods identified in the decoupleR benchmark analysis: VIPER, ORA, and Norm WMEAN. Regardless of the prior information, the Priori scores from these patients clustered by breast cancer subtypes (Figure 3.4A; Figure S3.4A-C). Like in the Holland et al. perturbation dataset, we observed more up-regulated than down-regulated target genes (Figure S3.4D). However, we did not observe a clear separation of breast cancer subtypes when using activity scores from the other methods (Figure S3.4E-G). BIDC is classified into three molecular subtypes: luminal, HER2, and basal cancers[211–215]. The most common types, luminal and basal breast cancers, are distinguished by hormone receptor expression. Luminal breast cancers express estrogen and progesterone receptors, whereas basal breast cancers do not. Unsupervised clustering of the Priori Pathway Commons scores revealed that the predicted activity of transcription factors that regulate the expression of estrogen receptors (ESR1) and progesterone receptors (NR3C1) were decreased in the basal breast cancer cluster (Cluster 2; Figure 3.4B). Moreover, Cluster 2 had decreased predicted GATA3 activity, which is critical to luminal cell specification, as well as

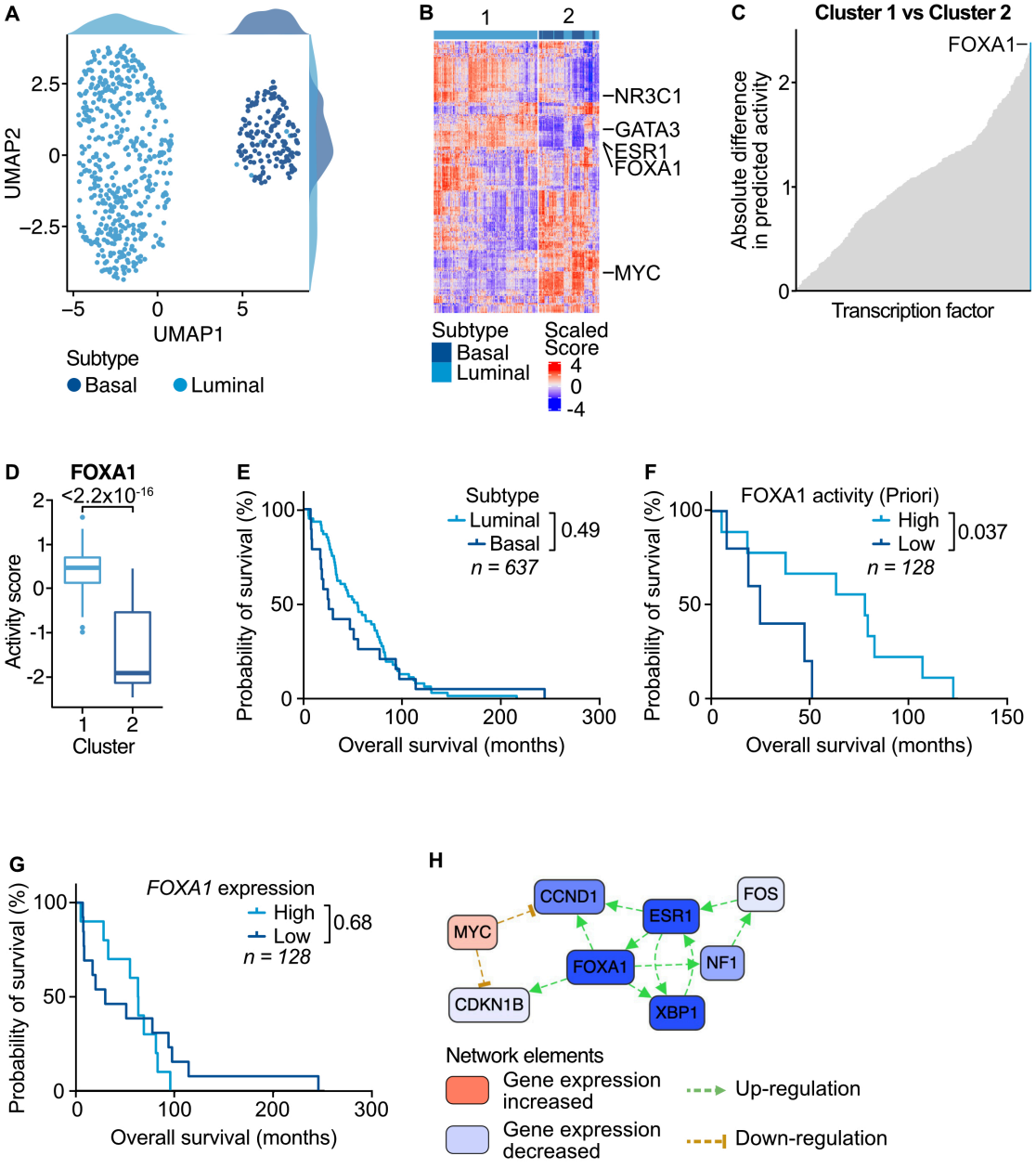

Figure 3.4: FOXA1 transcription factor activity is a significant determinant of survival for patients with BIDC.

(A) Priori scores were generated from RNA-seq of 637 patients with BIDC. UMAP dimensional reduction and projection of Priori scores. Dots are colored by the breast cancer molecular subtype. (B) Unsupervised hierarchical clustering of Priori scores generated in (A). (C) Mean absolute difference of Priori scores from patients in the clusters 1 and 2 defined in (B). (D) Distribution of FOXA1 Priori scores among patients in clusters 1 and 2 defined in (B). Statistical significance was determined by a two-sided Student's t-test.

(E-G) Kaplan-Meier survival analysis of patients grouped by (E) molecular subtype, (F) FOXA1 Priori scores, or (G) *FOXA1* normalized gene expression counts. Patients among the top 90% of Priori scores or counts were grouped into "High" and those in the bottom 10% were grouped into "Low". Statistical significance was determined by a log-rank Mantel-Cox test. (H) Differential gene expression network enrichment between clusters defined in (B). Select significantly enriched nodes are shown.

increased predicted MYC activity, which is consistent with previous studies[211,216–218]. These analyses show that Priori transcription factor activity scores delineated primary BIDC patients by their molecular subtype.

While the transcription factors associated with luminal and basal breast cancer are likely important in distinguishing BIDC patient samples by their molecular subtype, the greatest difference in predicted Priori activity between the two clusters was FOXA1 (Figure 3.4C and D; Table 3.2). FOXA1 is a forkhead protein associated with mammary gland development. While the difference in FOXA1 activity was significantly different between the patient clusters defined by the DoRothEA and OmniPath Priori scores, its activity was much less pronounced between the clusters defined by the other methods (Figure S3.5A-E; Table 3.2). Evaluation of the greatest difference in predicted activity between the other networks and methods nominated transcription factors other than FOXA1. While FOXA1 was the second greatest difference in predicted activity between the Priori DoRothEA and OmniPath scores, these methods instead nominated ESR1 (Figure S3.5F and G). Notably, FOXA1 is a known regulator of ESR1[219]. The other methods nominated different transcription factors, including OR10H2 by VIPER, ACTL6A by ORA, and TGFβ2 by Norm WMEAN (Figure S3.5H-J; Table 3.2). Together, these analyses show that we were able nominate drivers of basal and luminal breast cancer with Priori, identifying that luminal cancer samples are associated with high predicted FOXA1 activity.

Table 3.2: The top 5 differences in predicted transcription factor activity between BIDC primary patient sample clusters.

Method	Network	Transcription Factor	Absolute Difference in Activity (Cluster 1 vs 2)
Priori	Pathway Commons	FOXA1	2.382
		ESR1	2.324
		PAX2	2.235
		GATA3	2.226
		XBP1	2.173
	DoRothEA	ESR1	2.209
		FOXA1	2.099
		GATA3	2.008
		HOXB13	2.007
		MYOD1	2.002
	OmniPath	ESR1	2.279
		FOXA1	2.271
		GATA2	2.236
		GATA3	2.199
		TFAP2C	2.027
VIPER	ARACNe BRCA	OR10H2	3.2
		PBRM1	3.136
		SAP130	2.979
		ELOB	2.79

		ZIC2	2.783
ORA	Pathway Commons	ACTL6A	1.394
		RUVBL1	1.394
		RUVBL2	1.394
		KAT5	1.154
		TRRAP	1.048
Norm WMEAN	Pathway Commons	TGFB2	62.63
		RHOA	48.07
		EDNRA	46.81
		PTHLH	32.89
		LEP	22.77

To understand the clinical impact of transcription factor activity in breast cancer, we evaluated survival differences among the patients in the BIDC cohort. Patients grouped by their molecular subtypes demonstrated no significant difference in survival (Figure 3.4E). When we grouped patients by predicted FOXA1 Priori activity scores, patients with low FOXA1 activity had a significantly decreased chance of survival, which is consistent with previous reports (Figure 3.4F)[220]. We did not observe a survival difference when patients were instead grouped by *FOXA1* expression or by the FOXA1 activity scores generated by Priori using DoRothEA or OmniPath prior networks or the other methods (Figure 3.4G; Figure S3.6). Additionally, we did not see a survival difference in the transcription factor drivers nominated by Priori using DoRothEA or OmniPath prior networks or the other methods (Figure S3.7A-E). ORA predicted that ACTL6A activity was either 1.91 or 3.30 in the vast majority of samples (52.7% and 42.7%, respectively; Figure S3.7F). As a result, normal quantiles could not be

calculated, so all samples were included in the survival analysis. These data suggest that high FOXA1 transcription factor activity is protective of survival in BIDC.

In order to understand how FOXA1 activity may mediate survival in BIDC, we generated a differential gene regulatory network between the two patient clusters identified by Priori using Pathway Commons transcriptional relationships (Figure 3.4H)[205,206,221]. This analysis suggests two molecular mechanisms that distinguish luminal samples in Cluster 1 and basal samples in Cluster 2. First, we observed down-regulation of a positive feedback loop of *ESR1*, *FOXA1, XBP1, NF1,* and *FOS* in the basal Cluster 2 samples[222]. This is consistent with basal breast cancer, which is characterized by repression of estrogen receptor encoded by *ESR1*. Additionally, this analysis also shows that MYC downregulates FOXA1 cell cycle target genes, *CCND1* and *CDKN1B*, in basal breast cancer[223]. Downregulation of cell proliferation is a known mechanism of chemotherapy resistance in basal breast cancer[224]. These analyses nominate putative targets in the FOXA1 network that may regulate survival in BIDC.

3.3.6 FOXO1 transcription factor activity mediates venetoclax resistance in leukemia

Aberrant transcription factor activity is also an important regulator of drug resistance in multiple tumor types[46,225–228]. As we have shown that Priori nominates transcription factor regulation associated with breast cancer survival, we wanted to understand whether Priori could also be used to identify mediators of drug sensitivity. Since *ex vivo* drug screening data is not available in the TCGA datasets, we calculated Priori scores for 859 patients with acute myeloid leukemia (AML) from the Beat AML cohort[229,230]. This dataset provides paired baseline RNA-seq as well as *ex vivo* drug sensitivity data. Once again, we generated activity scores with Priori using transcriptional

relationships from Pathway Commons, DoRothEA, and Omnipath as well three of the top methods from the decoupleR benchmark analysis (VIPER, ORA, and Norm WMEAN). Consistent with the Holland et al. perturbation and TCGA BRCA datasets, we observed more up-regulated than down-regulated target genes (Figure S3.8A). To nominate transcription factors mediators of drug sensitivity, we correlated predicted transcription factor activity scores from each method to drug response. We found 11,075 significant inhibitor-transcription factor activity relationships using Priori, 2,934 of which were also identified when using Priori scores that were generated from DoRothEA or OmniPath prior networks (Figure 3.5A; Figure S3.8B). In contrast, only 192 of Priori Pathway Commons relationships were identified by VIPER, ORA, and Norm WMEAN (Figure S3.8C). VIPER likely identified the most significant inhibitor-transcription factor activity relationships (29,682) because it generated scores for more transcription factors than the other methods (1,363; Figure S3.8D). Among the strongest correlations from Priori was predicted FOXO1 transcription factor activity with venetoclax resistance (R^2 = -0.5895; Figure 3.5B). This relationship was the 33rd, 16th, and 10th highest correlations among Priori scores that used Pathway Commons, DoRothEA, or OmniPath as prior networks, respectively (Figure S3.8E and F). Venetoclax resistance is more highly correlated with predicted FOXO1 activity than *FOXO1* expression alone (R^2 = -0.499; Figure 3.5C). Notably, while FOXO1 Norm WMEAN activity scores were directly proportional to venetoclax resistance, the VIPER and ORA activity scores were anti-correlated to venetoclax resistance (Figure S3.8G-I). These findings nominate FOXO1 activity as a mediator of venetoclax activity in AML.

Venetoclax induces cancer cell death by restoration of intrinsic mitochondrial apoptosis. Venetoclax blocks BCL2 from sequestering factors that activate pro-apoptotic BCL2 family proteins, such as BAX[231]. In mantle cell lymphoma (MCL), it has been reported that genomic regions of BAX and multiple other pro-apoptotic BCL2 family

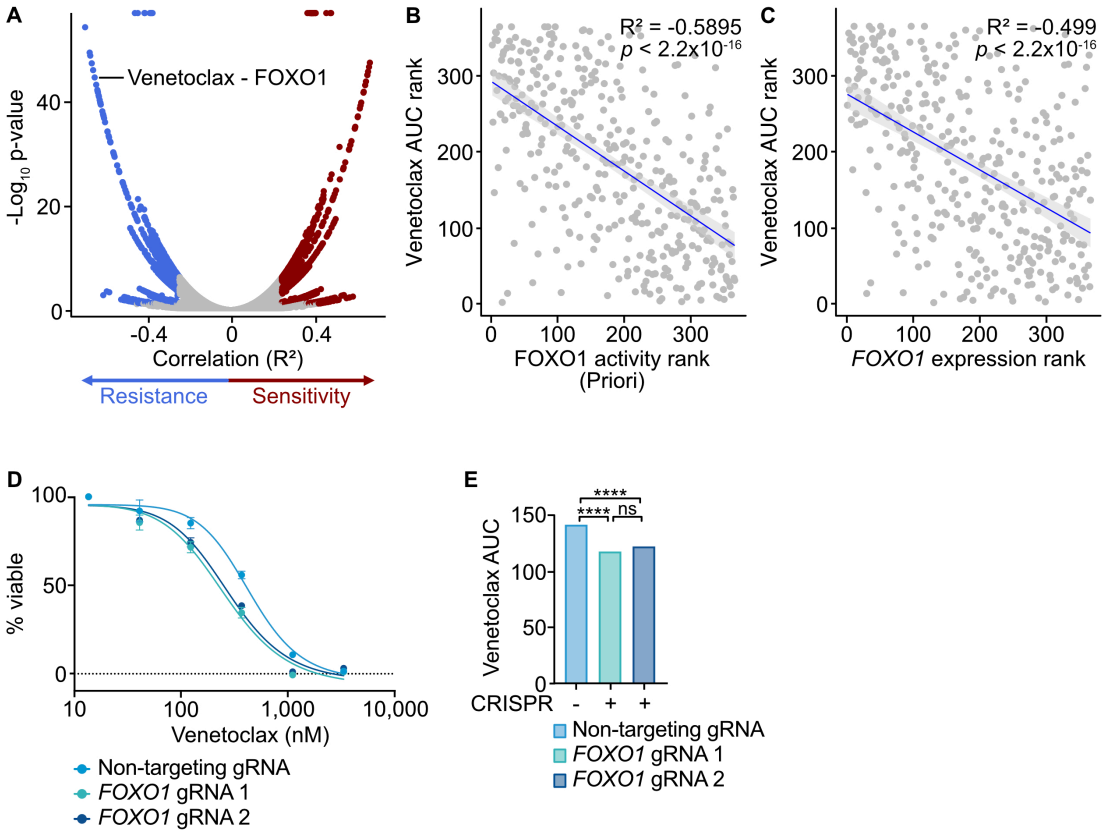

Figure 3.5: FOXO1 is a critical mediator of response to venetoclax in AML. (A) Priori scores generated from RNA-seq of 859 patients with AML. Spearman correlation of Priori scores and *ex vivo* drug response AUC data. (B and C) Spearman correlation of ranked venetoclax AUC and ranked (B) FOXO1 Priori scores or (C) *FOXO1* normalized counts. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. (D and E) THP-1 cells were transduced with lentiviral particles harboring expression cassettes for hSpCas9 and a non-targeting or *FOXO1* guide RNA. Cells were cultured for 3 days along a 7-point curve with venetoclax. Cell viability was assessed by CellTiter Aqueous colorimetric assay. ns = not significant; * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001.

proteins are bound by FOXO1[232]. The authors further demonstrated that disruption of FOXO1 activity sensitized MCL cell lines to venetoclax. However, the relationship between FOXO1 activity and venetoclax resistance has not been investigated in AML. Prior work has shown that monocytic AML is intrinsically resistant to venetoclax-based

therapy[163]. Given the results from our analysis of transcription factor activity in AML patients, we wanted to understand the extent to which *FOXO1* knockdown could sensitize monocytic AML to venetoclax treatment. We used CRISPR-Cas9 to knock-out *FOXO1* in THP-1 cells, a cell line model of monocytic leukemia (Figure S3.9). Both *FOXO1* CRISPR guides significantly increased venetoclax sensitivity in this cell line model of monocytic AML, suggesting *FOXO1* is an important mediator of venetoclax sensitivity (Figure 3.5D and E). These findings were consistent with the predictions from Priori (regardless of prior network) and Norm WMEAN, demonstrating how Priori can be used to detect transcription factor mediators of drug resistance.

3.4 Discussion

We have developed Priori, a computational algorithm that infers transcription factor activity using prior biological knowledge. The results from a third-party, unbiased benchmarking workflow demonstrate that Priori detects perturbed transcription factors with higher sensitivity and specificity than 11 other methods. Our analyses show that while accounting for transcription factor expression aids Priori in detecting perturbed transcription factors, Priori's improved performance over other methods is likely due to assessment of the direction and impact of transcription factor regulation on their target genes. Using Priori, we identified FOXA1 activity as a regulator of survival in BIDC and nominated important downstream targets that may contribute to this survival difference[220]. Importantly, there was no significant survival difference among patients that were stratified using FOXA1 scores generated by VIPER, ORA, Norm WMEAN, or even Priori that used DoRothEA or OmniPath as prior information. Moreover, we used Priori to nominate transcription factor regulators of drug sensitivity in AML. We found that predicted FOXO1 activity by Priori is highly associated with venetoclax resistance. We

validated these findings in a cell line model of monocytic AML, which is resistant to venetoclax[163]. While Priori (regardless of prior information) and Norm WMEAN predicted that FOXO1 activity was associated with venetoclax resistance, VIPER and ORA strongly associated FOXO1 activity with venetoclax sensitivity.

Priori leverages the Pathway Commons resource to identify known gene regulatory networks in gene expression data. While using the relationships from Pathway Commons enables Priori to ground its findings in peer-reviewed literature, this may limit the discovery of novel transcriptional relationships. Analyses from other groups suggest that prior-based method tend to replicate prior information[233]. Indeed, the analysis of BIDC patient samples revealed that Priori was able to identify known transcription factor drivers of BIDC from several biological datasets, including Pathway Commons, DoRothEA, and OmniPath. Our findings also showed that Priori detected a relationship between FOXA1 activity and BIDC survival. While it has been shown that high *FOXA1* expression is associated with improved outcomes in patients with estrogen receptor-positive disease, these findings reveal a novel relationship between FOXA1 activity in estrogen receptor-positive and receptor-negative disease[220,234–236]. Our analyses provide a novel mechanistic hypothesis that is ready for experimental investigation.

Pathway Commons integrates publicly available RNA, DNA, and protein data sourced from a variety of tissue types. However, Pathway Commons is not designed to curate tissue-specific transcription factor gene regulatory networks. Cell context influences regulatory interactions between transcription factors and their downstream target genes[47,237]. The single-gene perturbation experiments, whose gene expression data we used to evaluate the transcription factor activity methods, were performed in numerous cell types. The small size of this dataset precluded a definitive evaluation of tissue-context as a determinant of method performance. However, we designed Priori to

allow researchers to include their own regulatory networks for context-specific evaluation of their experiments. Overall, Priori should generate robust predictions that are generalizable across many cellular contexts.

In our study, we used RNA-seq data from large clinical cohorts to evaluate Priori. While we demonstrated that Priori can be used to identify determinants of survival and mediators of drug response in this context, more investigation is needed to understand Priori's ability to predict transcription factor activity from smaller scale experiments. Notably, in a different study, we have applied a preliminary version of Priori to investigate the mechanism of combined FLT3 and LSD1 inhibition in FLT3-ITD AML[238]. We used Priori to identify important determinants of the drug combination response and showed that predicted activity of a putative drug combination target, *MYC*, decreased in six patient samples. Since Priori scores are normalized across all samples analyzed within the same run, we expect that Priori has more power to identify differences in transcription factor activity among larger cohorts of patients. Researchers can contextualize Priori scores from smaller scale experiments by generating scores from the large and small cohort RNA-seq data in the same run. Of course, this is only possible if the RNA-seq data has been consistently normalized and batch effects have been mitigated. We encourage researchers to only compare Priori scores that have been generated in the same run.

In conclusion, results from this study showed that our transcription factor activity method, Priori, detects perturbed transcription factors with improved sensitivity and specificity. Using Priori, we found that predicted FOXA1 activity is a significant determinant of survival in BIDC. We nominated putative FOXA1 targets that may be important for this survival difference. Lastly, we found that predicted FOXO1 activity is highly correlated with venetoclax resistance. We validated these findings in a cell line model of AML that is intrinsically resistant to venetoclax.

3.5 Limitations of the study

The decoupleR benchmarking workflow facilitated a robust and unbiased comparison of numerous transcription factor activity methods[63]. Employing the dataset curated by Holland et al., the workflow reported average AUROC and AUPRC values across all down-sampling runs as well as the transcription factor activity scores generated for each run[207]. However, the workflow did not report the positively identified perturbed transcription factors for each run. Consequently, our ability to investigate the distinctive characteristics of experiments (e.g., cell type, functional genetic technique, perturbed transcription factor) that might have influenced each method's predictions was constrained. Further efforts are warranted to include these reporting metrics, enabling a more comprehensive understanding of the factors underlying Priori's proficiency in identifying perturbed transcription factor activity.

3.6 Methods

3.6.1 Key resources table

REGENT OR RESOURCE	**SOURCE**	**IDENTIFIER**
Deposited data		
Holland et al. perturbation dataset (RNA-seq)	Holland et al. 2020[207]	Used in this study: https://zenodo.org/records/8368697/files/holland_rna_expr.tsv

		Original source: https://zenodo.org/record/5645208/files/rna_expr.rds?download=1
Holland et al. perturbation dataset (metadata)	Holland et al. 2020[207]	Used in this study: https://zenodo.org/records/8368697/files/holland_rna_meta.tsv Original source: https://zenodo.org/record/5645208/files/rna_expr.rds?download=1
TCGA PanCancer Atlas Breast Invasive Carcinoma (RNA-seq)	TCGA[210]	Used in this study: https://zenodo.org/records/8368697/files/tcga_brca_normalized_counts.tsv Original source: https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018
TCGA PanCancer Atlas Breast Invasive Carcinoma (metadata)	TCGA[210]	Used in this study: https://zenodo.org/records/8368697/files/tcga_brca_metadata.tsv Original source: https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018
Beat AML (RNA-seq)	Bottomly et al. 2022[230]	Used in this study: https://zenodo.org/records/8368697/files/beataml_rna_expr.tsv Original source: https://github.com/biodev/beataml2.0_data/raw/main/beataml_waves1to4_norm_exp_dbgap.txt

Beat AML (Inhibitor AUC values)	Bottomly et al. 2022[230]	Used in this study: https://zenodo.org/records/8368697/files/beataml_inhibitor_auc.tsv Original source: https://github.com/biodev/beataml2.0_data/raw/main/beataml_probit_curve_fits_v4_dbgap.txt
Software and algorithms		
Priori	This study	https://github.com/ohsu-comp-bio/regulon-enrichment
decoupleR version 2.0.0 (fork for this study)	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleR
decoupleRBench version 0.1.0 (fork for this study)	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleRBench
decoupleR benchmarking workflow	This study	https://github.com/ohsu-comp-bio/decoupler_workflow
Pathway Commons	Rodchenkov et al. 2020	https://www.pathwaycommons.org/
Patterns	Babur et al. 2014	https://code.google.com/archive/p/biopax-pattern/
Univariate Linear Model	Teschendorff and Wang 2020[59]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-ulm.R

(ULM) version 1.0.0		
Multivariate Linear Model (MLM) version 1.0.0	Teschendorff and Wang 2020[59]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-mlm.R
Over Representation Analysis (ORA) version 1.0.0	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-ora.R
Fast Gene Set Enrichment (FGSEA) version 1.20.0	Korotkevich et al. 2021[60]	https://github.com/ctlab/fgsea/
Gene Set Variation Analysis (GVSA) version 1.42.0	Hänzelmann et al. 2013[61]	https://bioconductor.org/packages/3.17/bioc/html/GSVA.html
AUCell version 1.16.0	Aibar et al. 2017[62]	https://github.com/aertslab/AUCell
Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) version 1.28.0	Alvarez et al. 2016[67]	https://www.bioconductor.org/packages/release/bioc/html/viper.html

Weighted Mean (WMEAN) version 1.0.0	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-wmean.R
Weighted Sum (WSUM) version 1.0.0	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-wsum.R
Univariate Decision Tree (UDT) version 1.0.0	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-udt.R
Multivariate Decision Trees (MDT) version 1.0.0	Badia-i-Mompel et al. 2022[63]	https://github.com/ohsu-comp-bio/decoupleR/blob/master/R/statistic-mdt.R
OmnipathR version 3.6.0	Turei et al. 2021[209]	https://www.bioconductor.org/packages/release/bioc/html/OmnipathR.html
dorothea version 1.10.0	Garcia-Alonso et al. 2019[208]	https://bioconductor.org/packages/release/data/experiment/html/dorothea.html
ARACNE-AP version 1.0	Lachmann et al. 2016[66]	https://github.com/califano-lab/ARACNe-AP
arcane.networks version 1.18.0	Lachmann et al. 2016[66]	https://bioconductor.org/packages/release/data/experiment/html/aracne.networks.html
Seurat version 4.4.0	Hafemeister, Satija, et al. 2019[239]	https://github.com/satijalab/seurat
CausalPath version 1.8.0	Babur et al. 2021[221]	https://github.com/PathwayAndDataAnalysis/causalpath
Geneious Prime	Dotmatics	https://www.geneious.com/

ICE CRISPR Analysis Tool	Synthego	https://ice.synthego.com/#/
Experimental models: Cell lines		
THP-1	ATCC	Catalog #TIB-202
Lenti-X 293T cells	Clontech	Catalog #632180
Chemicals, peptides, and recombinant proteins		
RPMI 1640 Medium	Gibco	Catalog #11875093
HyClone Characterized FBS	Cytiva Life Sciences	Catalog #SH30071.04
GlutaMAX	Gibco	Catalog #35050079
2-Mercaptoethanol	Sigma Aldrich	Catalog #M6250
HighPrep PCR Clean-up System	MagBio Genomics	Catalog #AC-60001
Recombinant DNA		
FOXO1 gRNA in pLentiCRISPR v2 backbone (#1)	Genscript	Sequence: GCTCGTCCCGCCGCAACGCG
FOXO1 gRNA in	Genscript	Sequence: ACAGGTTGCCCCACGCGTTG

pLentiCRISPR v2 backbone (#2)		
Non-targeting pLentiCRISPR v2	Addgene	Catalog #169795
psPAX2	Addgene	Catalog #12260

3.6.2 Resource availability

3.6.2.1 Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Emek Demir (demire@ohsu.edu).

3.6.2.2 Materials availability

This study did not generate new unique reagents.

3.6.2.3 Data and code availability

All original and forked code have been deposited at GitHub and are publicly available as of the date of publication. The sources of the datasets supporting the current study are presented in the Key resources table and the Method Details section. Any additional information required to re-analyze the data reported in this paper or reproduce the results is available from the lead contact upon request.

3.6.3 Method details

3.6.3.1 Priori Algorithm

3.6.3.1.1 Pre-Processing

Priori normalizes and scales the input RNA-seq data prior to downstream analysis. Priori first filters out counts with a standard deviation less than 0.1 (controlled by the *thresh_filter* parameter). Priori linearly shifts the remaining counts by the minimum value and then $\log_2$ normalizes them (x_{gene}). Priori then scales the normalized counts (z_{gene}) using one of four methods: “standard”

$$z_{gene} = \frac{(x_{gene} - \mu_X)}{\sigma_X}$$

(Equation 1)

(where the mean and standard deviation of all normalized counts are indicated by μ_X and σ_X, respectively), “robust”

$$z_{gene} = \frac{(x_{gene} - X_{median})}{IQR}$$

(Equation 2)

(where the median normalized counts are indicated by X_{median} and the interquartile range is indicated by IQR), “minmax”

$$z_{gene} = \frac{(x_{gene} - X_{min})}{(X_{max} - X_{min})}$$

(Equation 3)

(where the minimum and maximum normalized counts are indicated by X_{min} and X_{max}, respectively), or “quant” where the normalized counts are scaled using the inverse of the cumulative distribution function, *F(x)*[240].

$$F(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

(Equation 4)

Priori defaults to the “standard” scaling function.

3.6.3.1.2 Network

Priori uses known gene regulatory networks to predict transcription factor activity from RNA-seq data. By default, Priori extracts transcriptional relationships from the Pathway Commons database to generate activity scores[205,206]. Users can also generate Priori scores using other gene regulatory networks with the *regulon* parameter. The user-defined network must specify the transcription factor (*Regulator*) and their downstream target genes (*Target*). Pathway relationships in Pathway Commons are represented with the BioPAX language[205,241]. BioPAX abstracts major pathway relationships, including gene regulatory networks, into a standardized format. However, BioPAX representations cannot be interpreted directly. In order to identify the gene regulatory networks encoded in Pathway Commons, we extracted transcription factors and their primary targets using Patterns[242]. Using the extracted network, Priori removes transcription factors with less than 15 downstream targets. Users can control the number of targets with the *regulon_size* parameter.

3.6.3.1.3 Activity Scores

Once the network is prepared, Priori generates an activity score for each transcription factor in an RNA-seq dataset. To calculate the activity score for a

transcription factor (TF_i), Priori first calculates weights for its target genes ($w_{TF_i,target_j}$). The target gene weight is the product of the F-statistic ($F_{TF_i,target_j}$) and the Spearman correlation coefficient ($\rho_{TF_i,target_j}$) of the transcription factor TF_i and its target gene $target_j$ expression:

$$w_{TF_i,target_j} = F_{TF_i,target_j} * \rho_{TF_i,target_j}$$

$$w_{TF_i,target_j} = \frac{\left(z_{TF_i} - \mu_{TFs}\right) * \left(z_{target_j} - \mu_{targets}\right)}{\sigma_{TF_i} * \sigma_{targets}} * \frac{cov\left(r_{TF_i}, r_{target_j}\right)}{\sigma_{r_{TF_i}} * \sigma_{r_{target_j}}}$$

(Equation 5)

where *i* represents the range of transcription factors in a dataset, *j* represents the range of target genes for a given transcription factor TF_i, and *r* represents the rank of the scaled counts relative to all features in the dataset. The non-negative, $\log_2$-normalized counts are used to calculate the F-statistic.

Priori first uses the transcription factor-target gene weights to determine the direction of regulation. Priori defines *k* down-regulated targets among all *j* target genes as those with $w_{TF_i,target_j} < 0$. Priori also identifies *l* up-regulated targets among all *j* target genes as those with $w_{TF_i,target_j} > 0$. Priori then ranks the scaled target counts for each transcription factor grouped by their direction of regulation ($r_{target_j,targets_k}$ or $r_{target_j,targets_l}$). Priori weighs the ranks of the scaled counts using the target gene weight $w_{TF_i,target_j}$:

$$r'_{target_j,targets_k} = w_{TF_i,target_j} * r_{target_j,targets_k}$$

$$r'_{target_j,targets_l} = w_{TF_i,target_j} * r_{target_j,targets_l}$$

(Equation 6)

The resulting activity score for a given transcription factor ($Activity_{TF_i}$) is the summation of the weighted ranks:

$$Activity_{TF_i} = -1\left(\sum_{k}^{down-regulated} r'_{target_j,targets_k}\right) + \left(\sum_{l}^{up-regulated} r'_{target_j,targets_l}\right)$$

(Equation 7)

where downregulated genes are scaled by -1 prior to summation. Finally, the activity scores for each transcription factor are z-transformed relative to all other transcription factors and then again to all other samples in the dataset.

3.6.3.2 Alternative Method: Evaluation of Transcription Factor Expression Only

We wanted to understand the extent to which assessment of transcription factor expression enabled Priori to detect perturbed transcription factors. We designed an alternative method that only uses the transcription factor expression to infer transcription factor activity. To infer the activity for a transcription factor TF_i, this alternative method first ranks the transcription factor scaled counts relative to other transcription factors in the dataset, $r_{TF_i,TFs}$. The metho then scales these ranks by the total number of transcription factors, n_{TF}, and normalizes them using a normal distribution:

$$r_{TF_i} = \frac{1}{\sqrt{2\pi}} e^{-\frac{\frac{r_{TF_i,TFs}}{n_{TF}+1}^2}{2}}$$

(Equation 8)

3.6.3.3 Alternative Method: Impact of Transcriptional Regulation Only

We wanted to understand the extent to which assessment of transcriptional regulation on target genes enabled Priori to detect perturbed transcription factors. We designed an alternative method that only uses the impact of transcriptional regulation to infer transcription factor activity. The alternative method uses the same values of the weighted ranks ($r'_{target_j,targets_k}$ or $r'_{target_j,targets_l}$) as calculated above. However, the resulting activity score for a given transcription factor ($Activity_{TF_i}$) is the summation of the weighted ranks, but the down-regulated genes are not scaled by -1 prior to summation:

$$Activity_{TF_i} = \left(\sum_{k}^{down-regulated} r'_{target_j,targets_k} \right) + \left(\sum_{l}^{up-regulated} r'_{target_j,targets_l} \right)$$

(Equation 9)

The activity scores for each transcription factor are z-transformed relative to all other transcription factors and then again to all other samples in the dataset.

3.6.3.4 Benchmarking Workflow

3.6.3.4.1 Pre-Processing

The decoupleR benchmarking workflow has been previously described[63]. The Holland et al. normalized gene expression counts and metadata for were downloaded Zenodo (https://zenodo.org/record/5645208)[63]. The normalized RNA-seq data was linearly shifted by the minimum value so all values were non-negative[207].

3.6.3.4.2 Transcription factor activity scores and p-values

With the decoupleR workflow (version 2.0.0), we generated transcription factor activity scores for 11 methods including Priori: AUCell (version 1.16.0), FGSEA (version 1.20.0), GSVA (version 1.42.0), MDT (version 1.0.0), ORA (version 1.0.0), UDT (version 1.0.0), ULM (version 1.0.0), VIPER (version 1.28.0), WMEAN (version 1.0.0), and WSUM (version 1.0.0). decoupleR also generated normalized transcription factor activity scores for FGSEA, WMEAN, and WSUM as well as corrected scores for WMEAN and WSUM[63]. We generated transcription factor activity scores for each method using the default parameters. OmniPath (downloaded using OmipathR package version 3.6.0) and DoRothEA (downloaded using Dorothea package version 1.10.0), which assign confidence scores to their transcriptional relationships, were filtered for high confidence relationships (A, B, or C). We also calculated p-values for the Priori activity scores using a Student's two-sided t-test with an FDR post-test correction.

3.6.3.4.3 Area under the receiver operating characteristic and precision recall curves

Using the decoupleRBench package (version 0.1.0), we generated AUROC and AUPRC values. Briefly, we ranked the absolute value of the activity scores from the decoupleR workflow for each experiment[63]. The activity scores were ranked separately for each method. We determined whether the perturbed transcription factors were among the top "n" scores. "n" was defined as the number of unique perturbed transcription factors in the dataset. There are 62 unique transcription factors in the Holland et al. perturbation dataset, which is the dataset that we used to compare the prediction accuracy of Priori to other methods[207]. Therefore, a true positive is assigned to a method whose perturbed transcription factor ranks among the top 62 activity scores.

For methods that use the Pathway Commons transcriptional relationships (a total of 610 transcription factors), this is the top 10.2% of features. As the number of unperturbed transcription factors in the dataset substantially outnumbered the perturbed factors, we deployed a down-sampling strategy to compare an equal number. We calculated AUROC and AUPRC values for 100 down-sampling permutations.

3.6.3.4.4 ARACNe Network

In order to better understand whether Priori's performance advantages depend on the design of its input transcription factor network, we generated VIPER activity scores using transcriptional relationships from TCGA ARACNe networks and an alternative network using ARACNe-AP[66]. We downloaded the TCGA ARACNe networks using the arcane.networks package (version 1.18.0). For the ARACNe-AP network, we computed it from the Holland et al. perturbation dataset[207]. ARACNe-AP requires a list of transcription factors in order to generate a gene regulatory network. We used a list of transcription factors from the Alvarez et al. 2016 publication that was provided by Dr. Mariano Alvarez on September 25, 2019[67]. We excluded transcription factors that were not present in the Holland et al. dataset, resulting in an input list of 1,726 transcription factors. We ran ARACNe-AP (version 1.0, created with java 1.8.0_171-b11) with 100 bootstraps, *--p-value = 1E-8,* and *–random seeds = TRUE*[66]. The consolidated interactome included 1,726 transcription factors and 302,444 interactions.

3.6.3.4.5 Noise

In order to investigate the robustness transcription factor activity scores to noise, we artificially altered the input data, including the gene expression data from the Holland et al. perturbation dataset and the transcriptional relationships from Pathway Commons, to each method. First, we added zero-centered, Gaussian-distributed noise to the gene

expression data using the stats package (R base version 4.1.3). We increased the amount of noise by altering the standard deviation of the Gaussian distribution. Next, we evaluated activity scores by artificially altering the Pathway Commons transcriptional relationships. We tested this in two ways. First, we randomly removed transcription factor-target gene pairs in the Pathway Commons prior network using the stats package (R base version 4.1.3). Second, we randomized the transcription factor-target gene pairs by sampling genes in the Holland et al. dataset with replacement. We replaced these selected genes as the target genes of randomly selected transcription factors using the stats package (R base version 4.1.3).

3.6.3.5 Data Analysis

3.6.3.5.1 Benchmarking Workflow

Custom scripts were used to evaluate the results of the decoupleR benchmarking workflow. In order to compare the scores across the different methods, we z-transformed the transcription factor activity scores. We calculated the Spearman correlation between the activity scores and p-values of each method.

3.6.3.5.2 TCGA BRCA

The normalized gene expression counts and metadata from the Breast Invasive Carcinoma (TCGA, PanCancer Atlas) study were downloaded from cBioPortal (https://www.cbioportal.org)[210,243]. Priori scores were generated from the normalized gene expression counts linearly shifted by the minimum value using the same Pathway Commons, DoRothEA, or OmniPath relationships that were evaluated in the decoupleR benchmarking workflow. VIPER scores were generated using the ARACNe BRCA network, *--pleiotropy* = *TRUE,* and *--eset.filter* = FALSE[66,67]. Norm WMEAN scores were

generated using Pathway Commons transcriptional relationships, *--times* = *100*, *--sparse* = *TRUE*, and *--randomize_type* = *rows*[63]. ORA scores were generated using Pathway Commons transcriptional relationships, *--n_up* = *300*, *--n_down* = *300*, *--n_background* = *20000*, and *--with_ties* = *TRUE*. Custom scripts were used to exclude patients without basal or luminal BIDC. The Seurat package (version 4.4.0) was used to perform dimensional reduction on the Priori scores by PCA and UMAP[239]. Survival data was analyzed for significance using a log-rank Mantel-Cox test. We assigned patients to "High" or "Low" groups depending on whether they were among the top 90% or bottom 10% of the value of interest. Differential gene expression networks were generated from the normalized gene expression data using CausalPath (version 1.18.0)[221]. An FDR threshold of 0.001 was used to evaluate significant relationships following 100 permutations.

3.6.3.5.3 Beat AML

The normalized gene expression counts and inhibitor AUC values from the Beat AML study were downloaded from GitHub (https://github.com/biodev/beataml2.0_data)[229,230]. Priori scores were generated from the normalized gene expression counts of baseline AML patient samples from the Beat AML cohort using the same Pathway Commons, DoRothEA, or OmniPath relationships that were evaluated in the decoupleR benchmarking workflow. These counts were also linearly shifted by the minimum value. VIPER scores were generated using a trained ARACNe-AP network, *--pleiotropy* = *TRUE,* and *--eset.filter* = FALSE[66,67]. For the ARACNe-AP network, we computed it from the normalized RNA-seq data from Beat AML[229,230]. ARACNe-AP requires a list of transcription factors in order to generate a gene regulatory network. We used a list of transcription factors from the Alvarez et al. 2016 publication that was provided by Dr. Mariano Alvarez on September 25, 2019[67].

We excluded transcription factors that were not present in the Holland et al. dataset, resulting in an input list of 1,408 transcription factors. We ran ARACNe-AP (version 1.0, created with java 1.8.0_171-b11) with 100 bootstraps, *--p-value = 1E-8,* and *--random seeds = TRUE*[66]. The consolidated interactome included 1,402 transcription factors and 320,632 interactions. Norm WMEAN scores were generated using Pathway Commons transcriptional relationships, *--times = 100*, *--sparse = TRUE*, and *--randomize_type = rows*[63]. ORA scores were generated using Pathway Commons transcriptional relationships, *--n_up = 300*, *--n_down = 300*, *--n_background = 20000*, and *--with_ties = TRUE*. Custom scripts were used to exclude patients without a diagnosis of AML or those with a prior myeloproliferative neoplasm. Priori scores and single-inhibitor drug AUC values on the same patient sample were evaluated using a Spearman correlation. Significant correlations were those with an FDR < 0.05.

3.6.3.6 Cell Culture

3.6.3.6.1 Cell Lines

THP-1 cells (DSMZ) were cultured in RPMI (Gibco) supplemented with 10% fetal bovine serum (FBS, HyClone), 2 mM GlutaMAX (Gibco), 100 units/mL Penicillin, 100 µg/mL Streptomycin (Gibco), and 0.05 mM 2-Mercaptoethanol (Sigma Aldrich). All cells were cultured at 5% CO_2 and 37°C. Cell lines were tested monthly for mycoplasma contamination.

3.6.3.6.2 CRISPR

Two *FOXO1* CRISPR guide RNAs in a pLentiCRISPR v2 backbone were obtained from GenScript[244]. The target sequence for guide RNA (gRNA) 1 is GCTCGTCCCGCCGCAACGCG and the sequence for gRNA 2 is

ACAGGTTGCCCCACGCGTTG. Additionally, a non-targeting CRISPR guide RNA (target sequence CCTGGGTTAGAGCTACCGCA) generated by scrambling the target sequence of LentiCRISPRv2-ACTB-C1 in a pLentiCRISPR v2 backbone was obtained from Addgene (catalog #169795). Lentivirus was produced by transfecting Lenti-X 293T cells (Clontech) with the SMARTvector transfer plasmid and packaging/pseudotyping plasmids. psPAX2 was a gift from Didier Trono (Addgene plasmid #12260; http://n2t.net/addgene:12260; RRID:Addgene_12260). The supernatant containing lentivirus was collected after 48 hours of culture and filtered with a 0.45 µm filter. THP-1 cells were transduced with virus via spinnoculation in the presence of polybrene. Transduced cells were selected with 1 µg/mL puromycin to produce a stable cell line.

3.6.3.6.3 CRISPR Validation

FOXO1 knockdown was validated using Tracking of Indels by Decomposition (TIDE)[245]. Briefly, cellular DNA was PCR-amplified using primers upstream (sequence AAGTAGGGCACGCTCTTGAC) and downstream (sequence CGTTCCCCCAAATCTCGGAC) of the *FOXO1* gRNA target sequences. The primers were designed in Geneious Prime and synthesized by Integrated DNA Technologies. Paramagnetic beads were used to purify the PCR DNA fragments (MagBio Ref #AC-60001) and subsequently sequenced by EuroFins. Inference of CRISPR Edits (ICE) was performed using the Synthego web tool (https://ice.synthego.com/#/).

3.6.3.6.4 Drug Sensitivity Assay

Cells were cultured for 72 hours along a 7-point dose curve with venetoclax. Cell viability was assessed by CellTiter Aqueous colorimetric assay.

3.6.4 Quantification and statistical analysis

Values are represented as the mean and error bars are the SEM unless otherwise stated. Python and R were used to perform statistical analyses. Significance was tested using two-sided Student's t-test. Correlations were performed using Spearman's rank method. Statistical significance in the survival analyses was determined by a log-rank Mantel-Cox test. Where appropriate, p values were adjusted for repeated testing using the Benjamini–Hochberg method.

3.7 Acknowledgements

E Demir received funding from the National Institutes of Health (5U2CCA233280-03, 5U01CA224012-02, U24CA264007) and Ö Babur received funding from the NIH National Heart, Lung, and Blood Institute (R01HL146549) for this work. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank all of the patients for their precious time and donation of samples supporting this research. We appreciate the OHSU core facilities ExaCloud Cluster Computational Resource and the Advanced Computing Center for their assistance. We would also like to thank Michal R. Grzadkowski for his advice in the early stages of Priori's development.

3.8 Author contributions

WM Yashar: conceptualization, software, formal analysis, validation, investigation, visualization, methodology, writing-original draft, writing-review and editing. J Estabrook: conceptualization, software, formal analysis, validation, investigation, visualization, methodology, writing-review and editing. H Holly: conceptualization, formal

analysis, investigation, writing-review and editing. J Somers: formal analysis, investigation, writing-review and editing. O Nikolova: conceptualization, investigation, writing-review and editing. Ö. Babur: conceptualization, investigation, writing-review and editing. TP Braun: conceptualization, resources, supervision, funding acquisition, validation, investigation, writing-review and editing. E Demir: conceptualization, resources, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing-review and editing. The co-first authors may identify themselves as lead authors in their respective CVs.

3.9 Declaration of interests

WMY is a former employee of Abreos Biosciences, Inc. and was compensated in part with common stock options. Pursuant to the merger and reorganization agreement between Abreos Biosciences, Inc. and Fimafeng, Inc., WMY surrendered all of his common stock options in 03/2021. TPB has received research support from AstraZeneca, Blueprint Medicines as well as Gilead Sciences and is the institutional PI on the FRIDA trial sponsored by Oryzon Genomics. The authors certify that all compounds tested in this study were chosen without input from any of our industry partners. The other authors do not have competing interests, financial or otherwise.

3.10 Supplementary figures

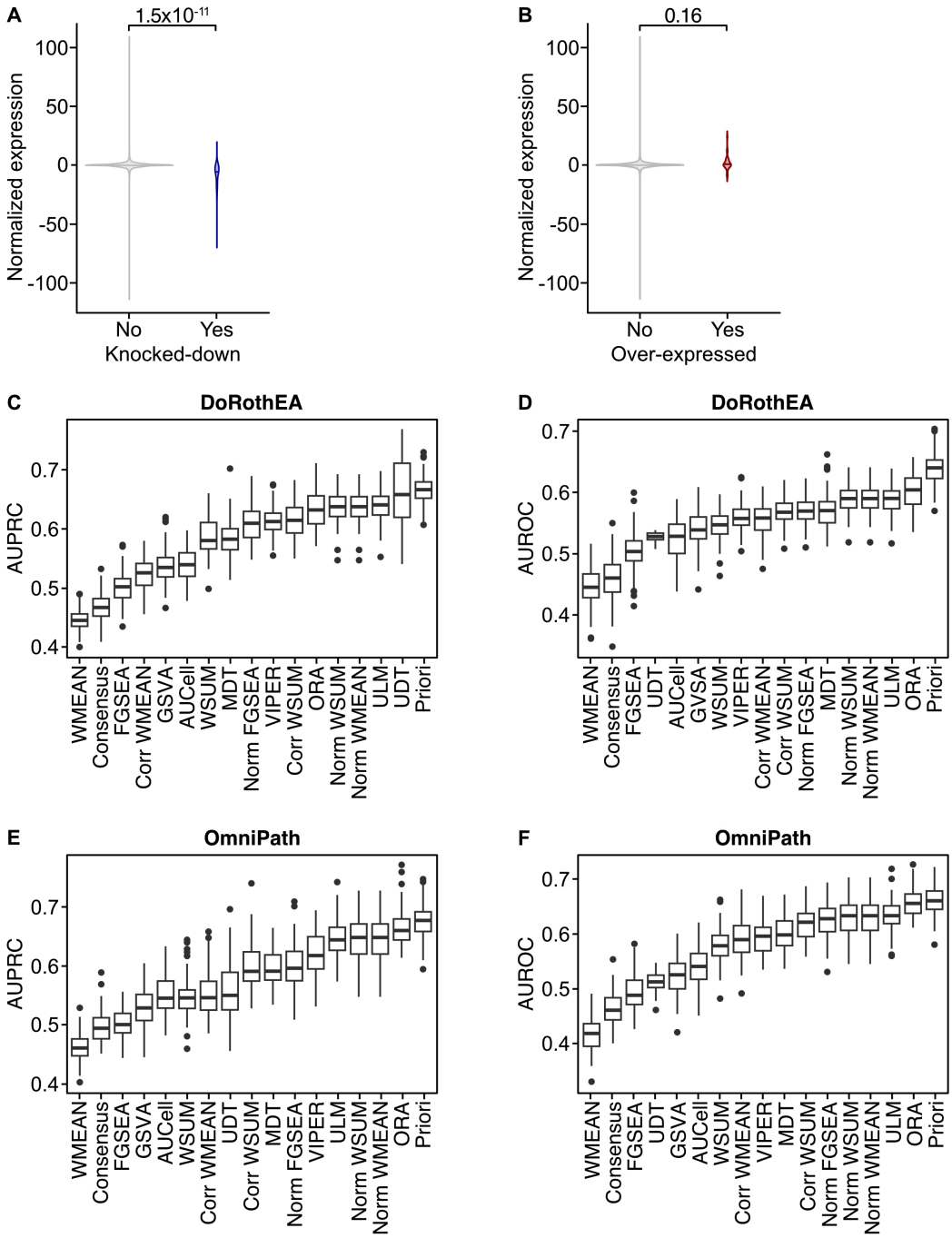

Figure S3.1: Priori demonstrated improved sensitivity and specificity when using transcriptional relationships from DoRoTHEA or OmniPath.

(A and B) Transcription factors in the perturbation dataset were identified using Pathway Commons transcriptional relationships. Transcription factors were delineated by whether they were (A) knocked-down or (B) over-expressed prior to RNA-seq. Statistical significance was determined by a two-sided Student's t-test. (C and D) Activity scores were generated using DoRothEA transcriptional relationships. The distribution of (C) AUPRC and (D) AUROC values across the 100 down-sampling permutations. (E and F) Activity scores were generated using OmniPath transcriptional relationships. The distribution of (E) AUPRC and (F) AUROC values across the 100 down-sampling permutations.

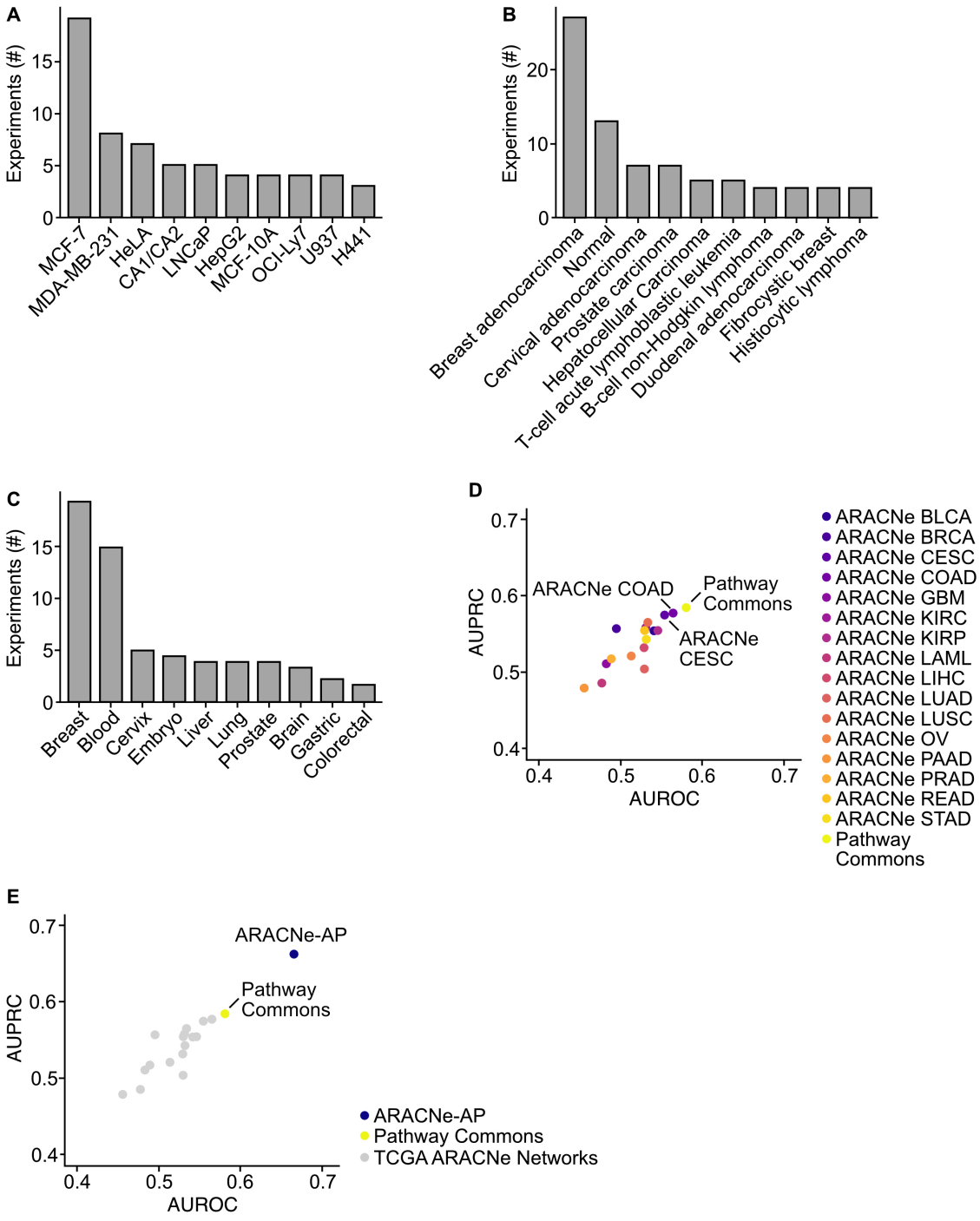

Figure S3.2: ARACNe-generated networks are important for VIPER performance.

(A-C) Top 10 represented (A) cell lines and their associated (B) diseases and (C) organ sites used in the perturbation dataset. (D and E) VIPER activity scores were generated using transcriptional relationships from (D) TCGA ARACNe networks or (E) an ARACNe-AP network. The ARACNe-AP interactome was computed from the perturbation dataset. Mean AUPRC and AUROC values across the 100 down-sampling permutations for each network.

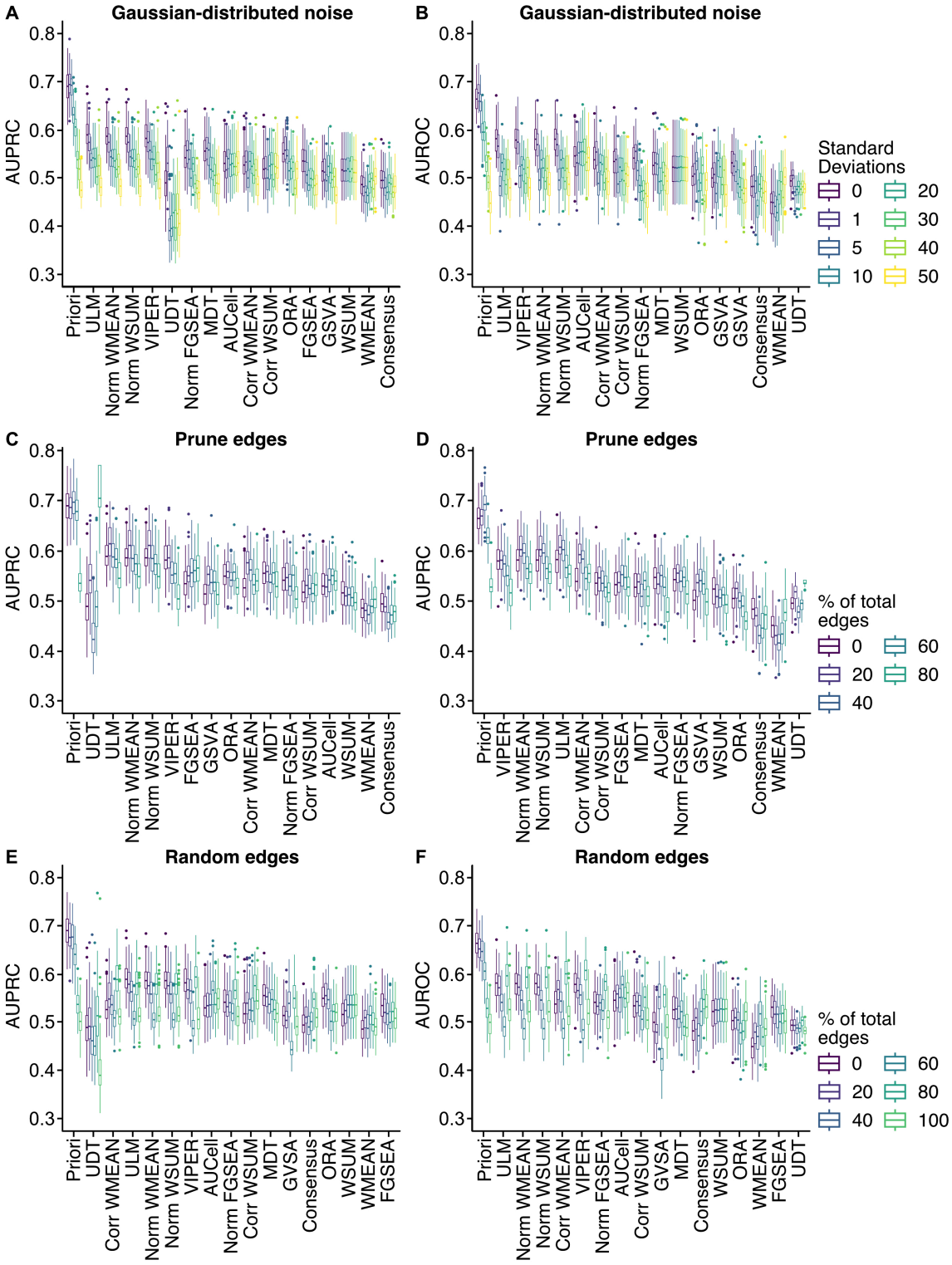

Figure S3.3: Priori's predictions are robust to noise.

(A and B) Zero-centered, Gaussian-distributed noise was introduced to the Holland et al. perturbation gene expression data. The amount of noise was controlled by altering the standard deviation of the Gaussian distribution. Activity scores were subsequently generated using Pathway Commons transcriptional relationships. The distribution of (A) AUPRC and (B) AUROC values across the 100 down-sampling permutations. (C and D) Transcription factor-target gene pairs were randomly removed from the prior Pathway Commons network. Activity scores were subsequently generated from the pruned networks. The distribution of (C) AUPRC and (D) AUROC values across the 100 down-sampling permutations. (E and F) Transcription factor target genes were randomized by sampling with replacement from genes sequenced in the Holland et al. perturbation dataset. Activity scores were subsequently generated from the randomized networks. The distribution of (E) AUPRC and (F) AUROC values across the 100 down-sampling permutations.

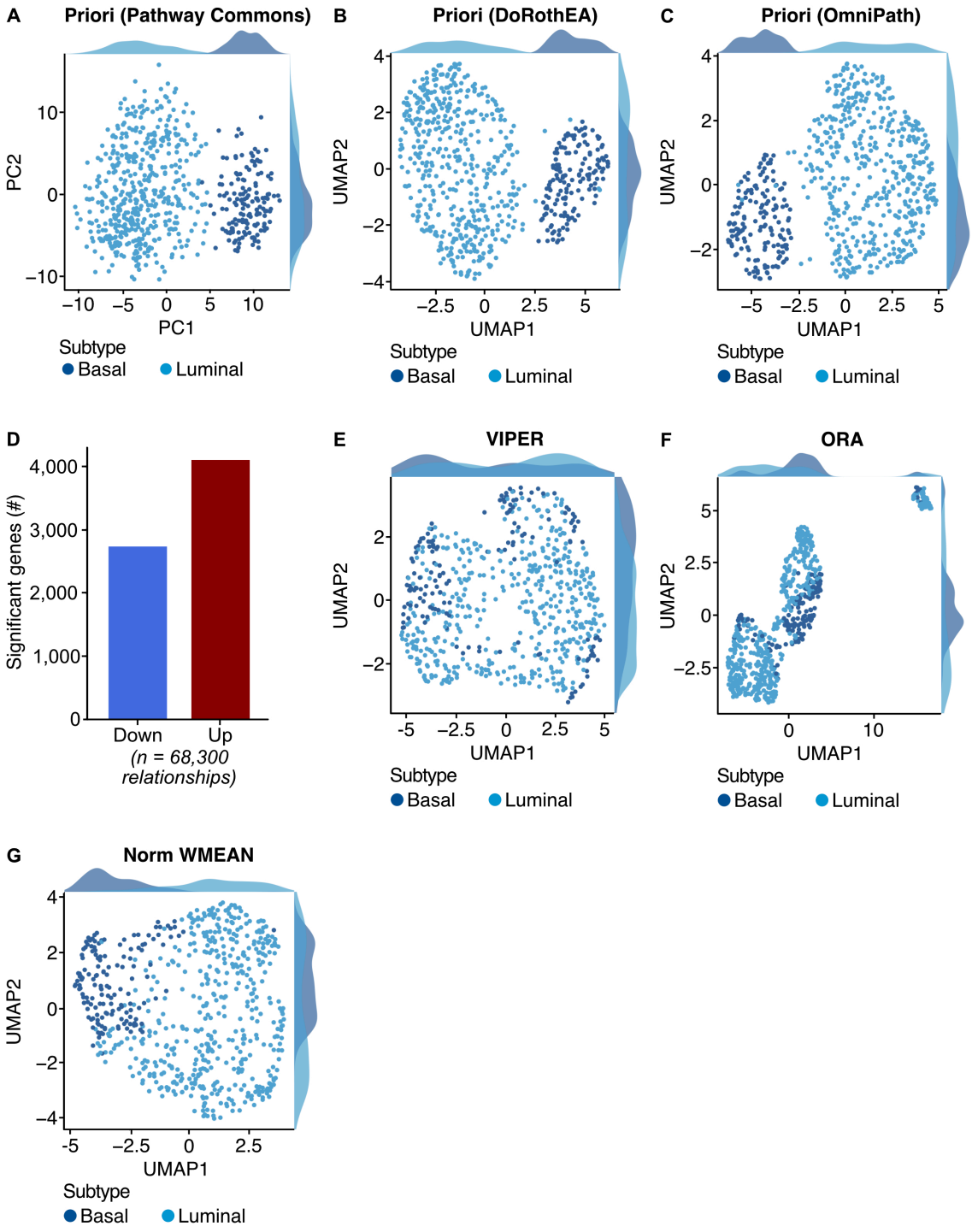

Figure S3.4: Prior transcription factor activity score, regardless of prior network, cluster by BIDC molecular subtype.

(A) Priori scores were generated from RNA-seq of 637 patients with BIDC using Pathway Commons transcriptional relationships. PCA dimensional reduction and projection of Priori. Dots are colored by the breast cancer molecular subtype. (B and C) Priori scores were generated from RNA-seq of 637 patients with BIDC using (B) DoRothEA or (C) OmniPath transcriptional relationships. UMAP dimensional reduction and projection of activity scores. Dots are colored by the breast cancer molecular subtype. (D) Priori identified transcription factor target genes using Pathway Commons transcriptional relationships. The expression of transcription factors and their target genes were evaluated using Spearman correlation. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. The Spearman correlation coefficient was used to determine down-regulated ($R^2 < 0$) and up-regulated target genes ($R^2 > 0$). (E-G) Activity scores were generated using (E) VIPER, (F) ORA, and (G) Norm WMEAN from RNA-seq of 637 patients with BIDC. UMAP dimensional reduction and projection of activity scores. Dots are colored by the breast cancer molecular subtype.

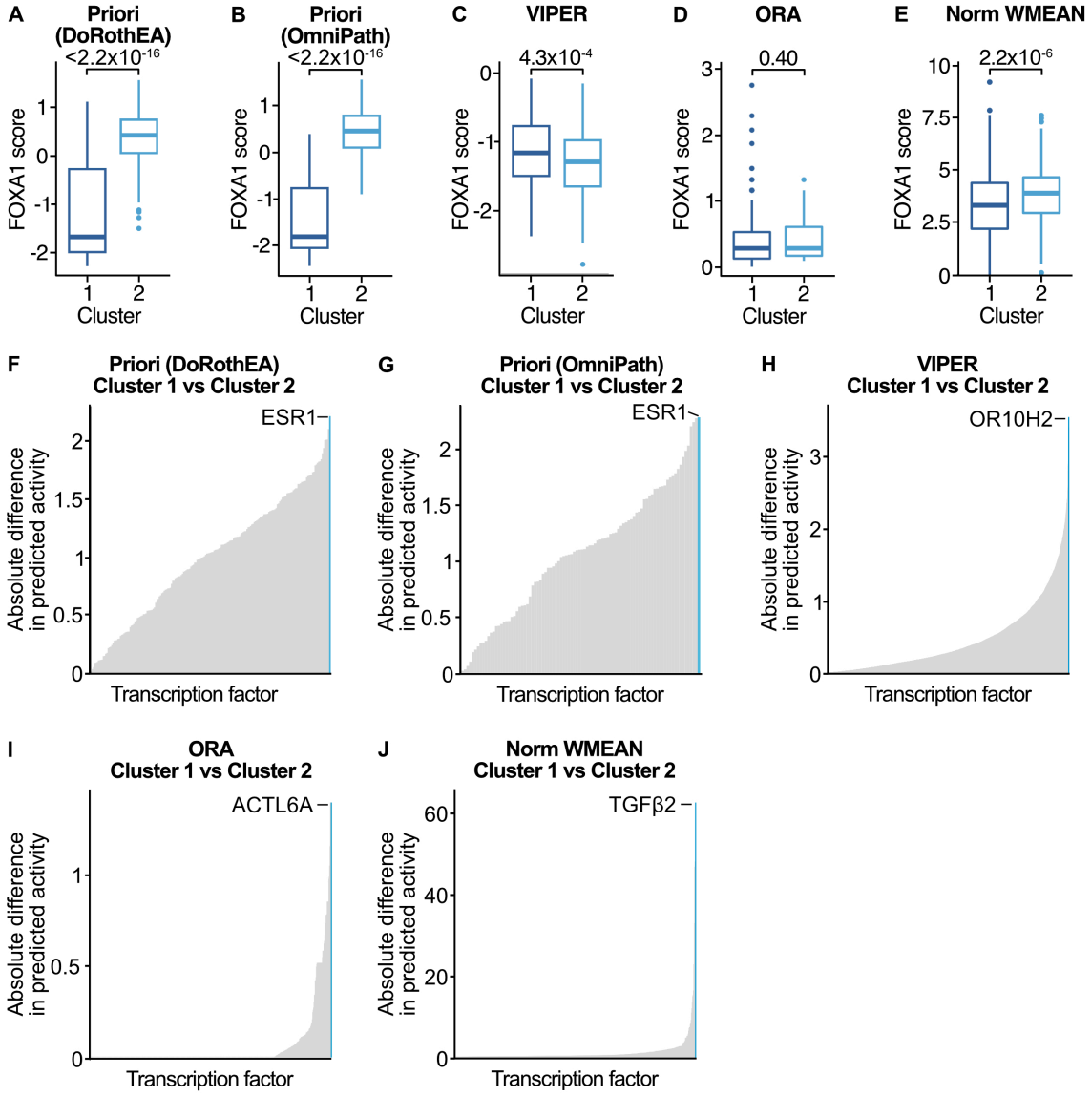

Figure S3.5: VIPER, ORA, and Norm WMEAN nominated distinct transcription factor regulators of BIDC pathogenesis.

(A and B) Unsupervised hierarchical clustering was performed using activity scores in Figure S3.4C and D. Distribution of FOXA1 Priori scores generated using (A) DoRothEA or (B) OmniPath transcriptional relationships among patients in clusters 1 and 2 (defined separately for each method). (C-E) Unsupervised hierarchical clustering was performed using activity scores in Figure S3.4E-G. Distribution of FOXA1 (C) VIPER, (D) ORA, and (E) Norm WMEAN scores among patients in clusters 1 and 2 (defined separately for each method). (F and G) Mean absolute difference of FOXA1 Priori scores generated using (F) DoRothEA or (G) OmniPath transcriptional from patients in clusters 1 and 2 (defined separately for each method). Transcription factor with the greatest absolute difference in activity scores between the two clusters is highlighted. (H-J) Mean absolute difference of (H) VIPER, (I) ORA, and (J) Norm WMEAN activity scores from patients in clusters 1 and 2 (defined separately for each method). Transcription factor with the greatest absolute difference in activity scores between the two clusters is highlighted.

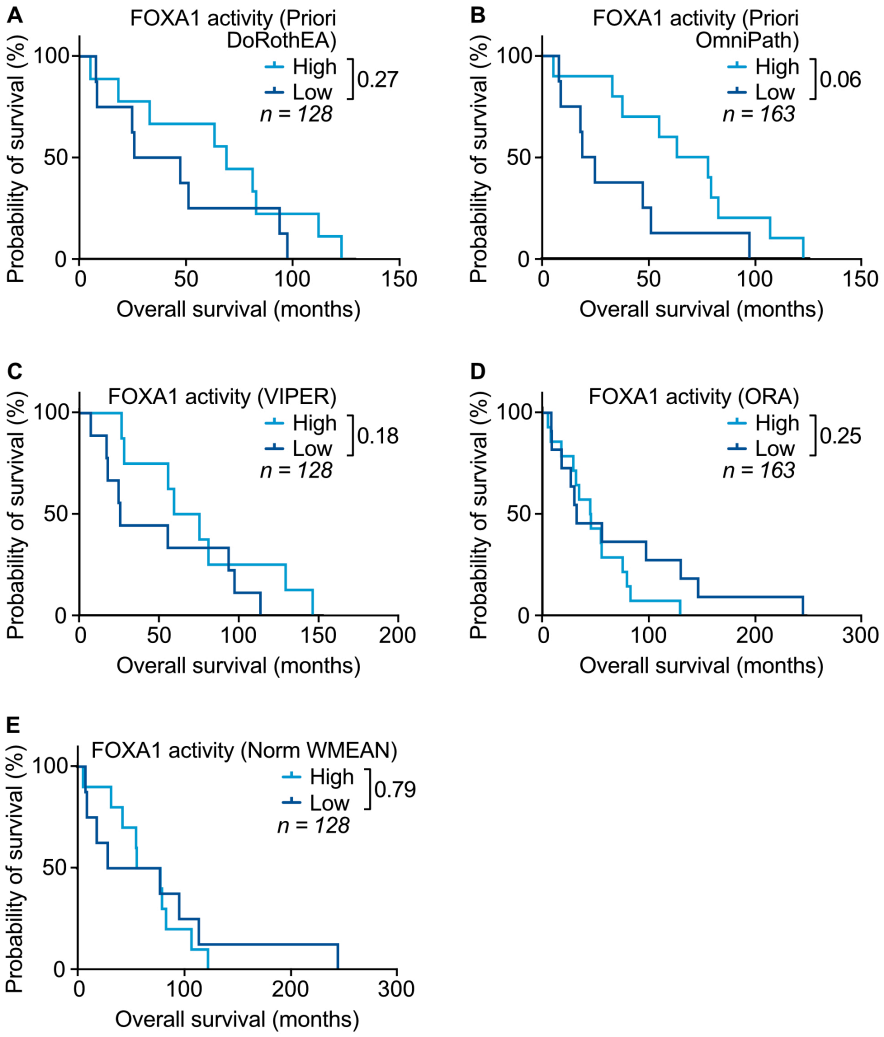

Figure S3.6: There is no survival difference when patients with BIDC are stratified FOXA1 activity scores generated by alternative methods.
(A and B) Kaplan-Meier survival analysis of patients grouped by FOXA1 Priori scores generated using (A) DoRothEA or (B) OmniPath transcriptional relationships. Patients among the top 90% of scores were grouped into "High" and those in the bottom 10% were grouped into "Low". Statistical significance was determined by a log-rank Mantel-Cox test. (C-E) Kaplan-Meier survival analysis of patients grouped by FOXA1 (C) VIPER, (D) ORA, and (E) Norm WMEAN scores. Patients among the top 90% of scores were grouped into "High" and those in the bottom 10% were grouped into "Low". Statistical significance was determined by a log-rank Mantel-Cox test.

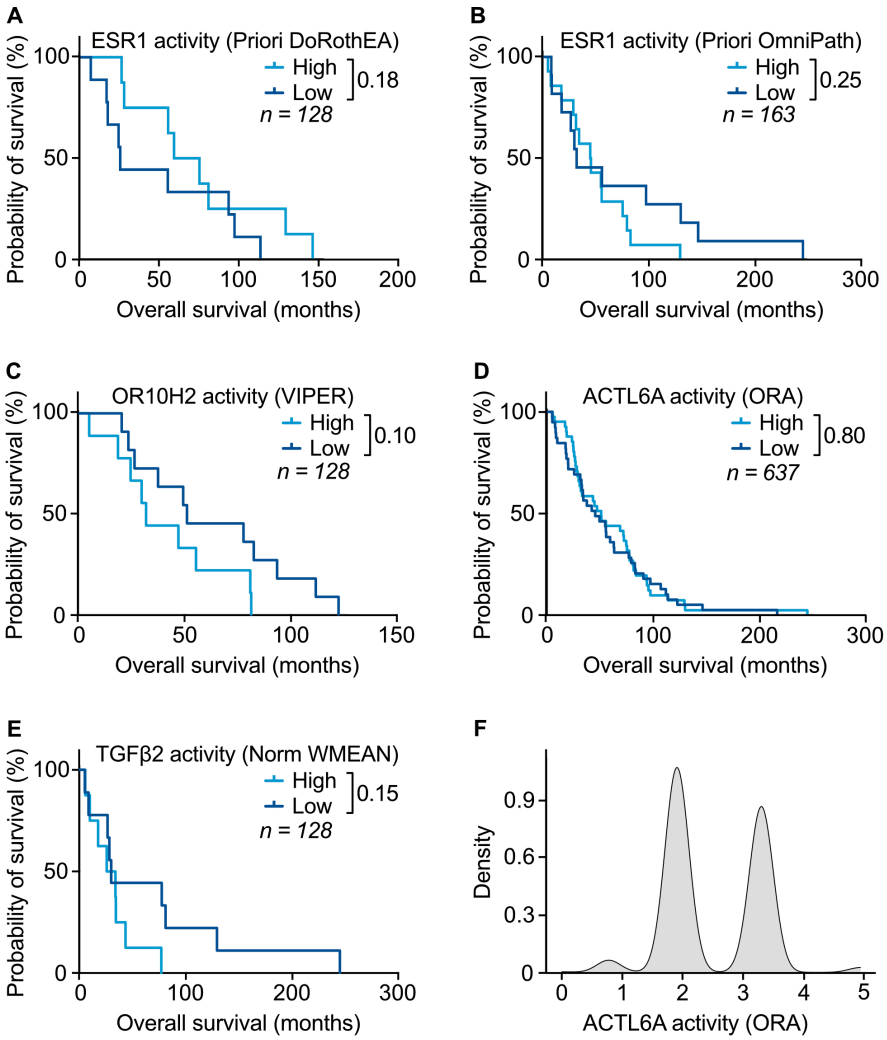

Figure S3.7: The alternative methods do not identify survival differences among patients with BIDC.

(A and B) Kaplan-Meier survival analysis of patients grouped by ESR1 Priori scores generated using (A) DoRothEA or (B) OmniPath transcriptional relationships. Patients among the top 90% of scores were grouped into "High" and those in the bottom 10% were grouped into "Low". Statistical significance was determined by a log-rank Mantel-Cox test. (C-E) Kaplan-Meier survival analysis of patients grouped by (C) OR10H2 VIPER scores, (D) ACTL6A ORA scores, or (E) TGFβ2 Norm WMEAN scores. Patients among the top 90% of scores were grouped into "High" and those in the bottom 10% were grouped into "Low". Statistical significance was determined by a log-rank Mantel-Cox test. (F) Distribution of ORA ACTL6A activity scores.

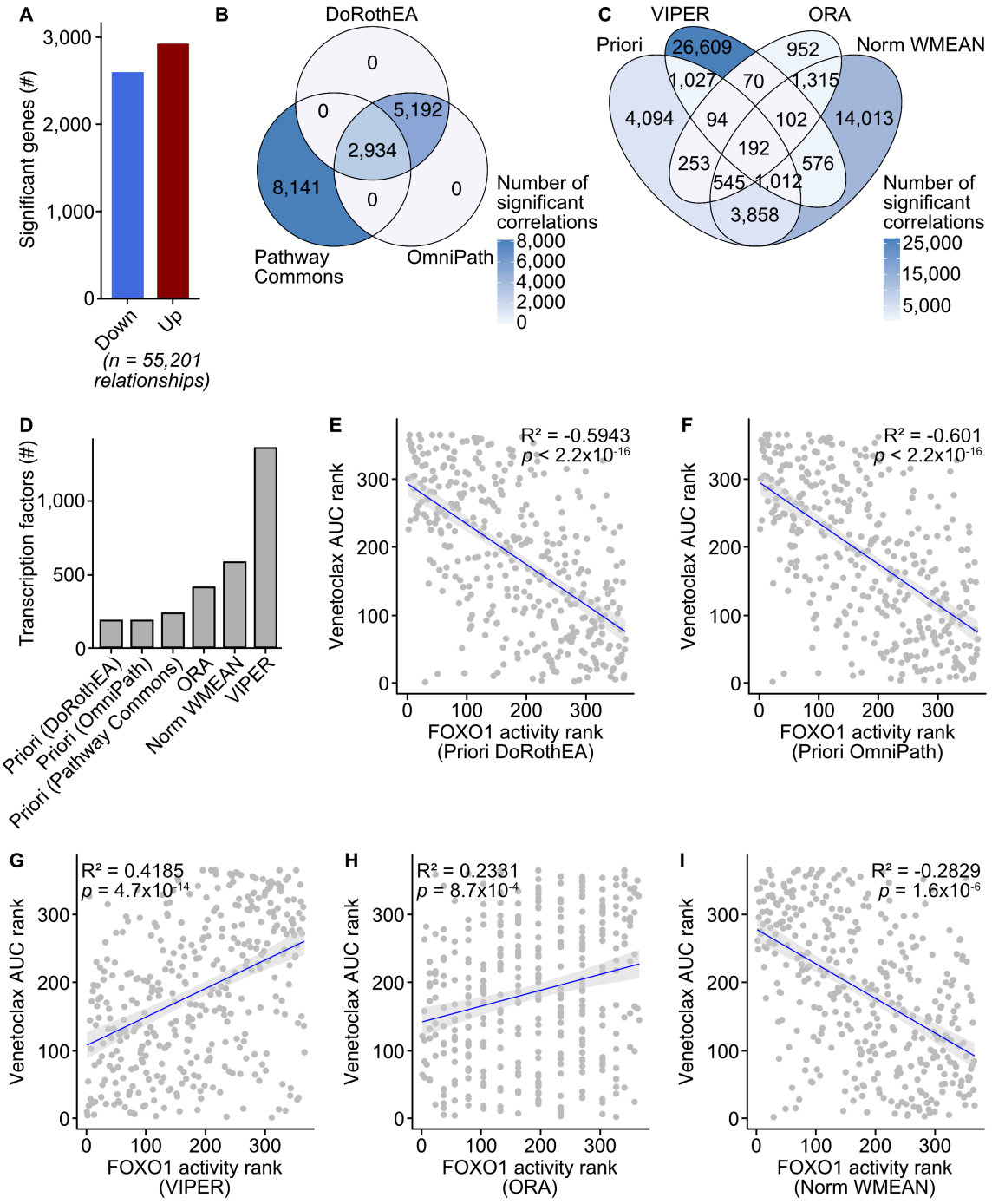

Figure S3.8: The alternative methods have different predictions whether FOXO1 activity correlates with venetoclax sensitivity or resistance.

(A) Priori identified transcription factor target genes using Pathway Commons transcriptional relationships. The expression of transcription factors and their target genes were evaluated using Spearman correlation. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. The Spearman correlation coefficient was used to determine down-regulated (R2 < 0) and up-regulated target genes (R2 > 0). (B) Priori scores were generated from RNA-seq of 859 patients with AML using Pathway Commons, DoRothEA, or OmniPath transcriptional relationships. Activity scores for each method and ex vivo drug response AUC data were evaluated using Spearman correlation. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. Overlap of significant activity score-inhibitor correlations for each method (including the significant Priori correlations from Figure 3.5A). (C) Activity scores were generated from RNA-seq of 859 patients with AML using VIPER, ORA, and Norm WMEAN. Activity scores for each method and ex vivo drug response AUC data were evaluated using Spearman correlation. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. Overlap of significant activity score-inhibitor correlations for each method (including the significant Priori correlations from Figure 3.5A). (D) Unique transcription factors associated with activity scores generated by Priori (using Pathway Commons transcriptional relationships), VIPER, ORA, and Norm WMEAN. (E and F) Spearman correlation of ranked venetoclax AUC and ranked Priori scores generated using (E) DoRothEA or (F) OmniPath transcriptional relationships. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction. (G-I) Spearman correlation of ranked venetoclax AUC and ranked FOXO1 (G) VIPER, (H) ORA, or (I) Norm WMEAN scores. Statistical significance was determined using the Spearman correlation p-value with an FDR post-test correction.

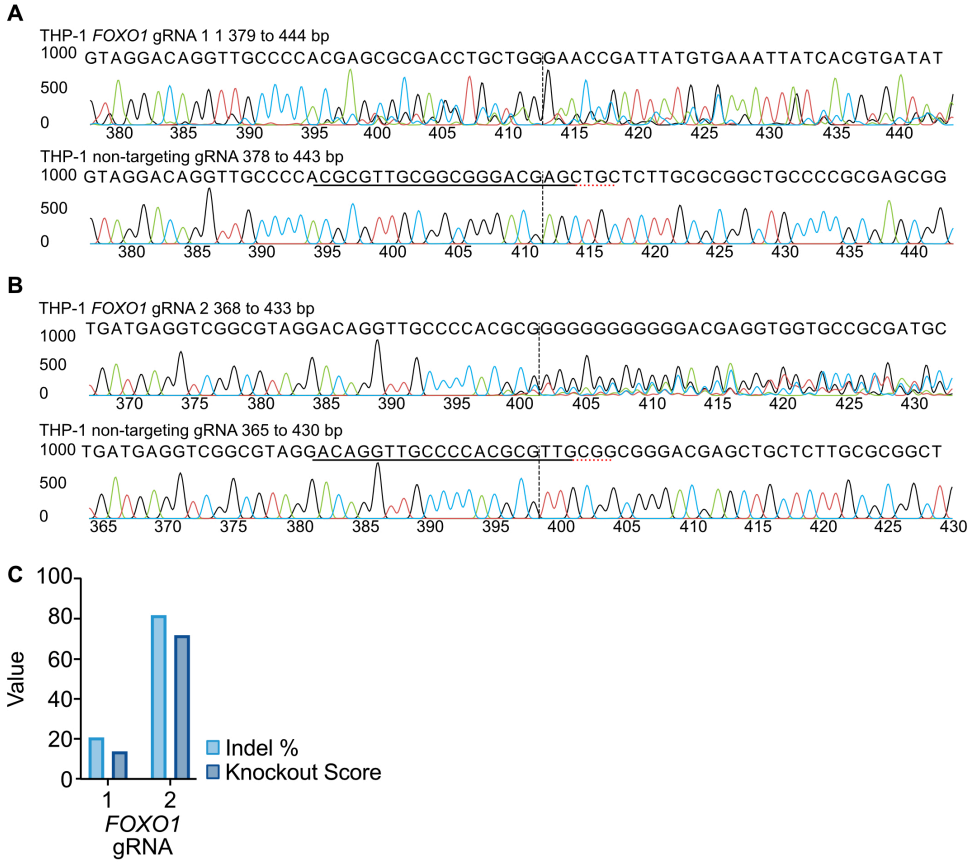

Figure S3.9: Validation of *FOXO1* knockdown by CRISPR.

(A and B) THP-1 cells were transduced with lentiviral particles harboring expression cassettes for hSpCas9 as well as a non-targeting or *FOXO1* gRNA. Sanger sequencing traces showing edited and control (cells with non-targeting gRNA) sequences in the region around the guide sequence. The horizontal black underlined region represents the guide sequence. The horizontal red underline represents the PAM site. The vertical black dotted line represents the cut site. (C) CRIPSR-Cas9-edited THP-1 cells were evaluated by the percentage of indels and knockout score, which is the proportion of indels that are a frameshift or are greater than 21 bp in length.

Chapter 4: Disruption of the MYC super-enhancer complex by dual targeting of FLT3 and LSD1 in acute myeloid leukemia

---

- Combined LSD1 and FLT3 inhibition induces synergistic cell death in FLT3-ITD AML.
- The drug combination disrupts STAT5, LSD1, and GFI1 binding at the MYC blood super-enhancer, suppressing super-enhancer accessibility as well as MYC expression and activity.
- The drug combination results in the accumulation of repressive H3K9me1 methylation, an LSD1 substrate, at MYC target genes.
- We validated these findings in 72 primary AML samples with the nearly every sample demonstrating synergistic responses to the drug combination.

This work has been published in *Molecular Cancer Research*:

Yashar WM, Curtiss BM, Coleman DJ, VanCampen J, Kong G, Macaraeg J, Estabrook J, Demir E, Long N, Bottomly D, McWeeney SK, Tyner JW, Druker BJ, Maxson JE, Braun TP. Disruption of the MYC Superenhancer Complex by Dual Targeting of FLT3 and LSD1 in Acute Myeloid Leukemia. *Molecular Cancer Research* 2023 Jul 5[238]

4.1 Abstract

Mutations in Fms-like tyrosine kinase 3 (FLT3) are common drivers in acute myeloid leukemia (AML) yet FLT3 inhibitors only provide modest clinical benefit. Prior work has shown that inhibitors of lysine-specific demethylase 1 (LSD1) enhance kinase inhibitor activity in AML. Here we show that combined LSD1 and FLT3 inhibition induces synergistic cell death in FLT3-mutant AML. Multi-omic profiling revealed that the drug combination disrupts STAT5, LSD1, and GFI1 binding at the MYC blood super-enhancer, suppressing super-enhancer accessibility as well as MYC expression and activity. The drug combination simultaneously results in the accumulation of repressive H3K9me1 methylation, an LSD1 substrate, at MYC target genes. We validated these findings in 72 primary AML samples with the nearly every sample demonstrating synergistic responses to the drug combination. Collectively, these studies reveal how epigenetic therapies augment the activity of kinase inhibitors in FLT3-ITD AML.

4.2 Introduction

Mutations in Fms-like tyrosine kinase 3 (FLT3) occur in nearly a third of all patients with acute myeloid leukemia (AML) and are associated with an inferior overall survival[246]. The most frequent mutation in FLT3 is the internal tandem duplication (ITD) of the juxta-membrane domain[247]. While small molecule inhibitors of FLT3 kinase produce higher overall response rates and improved survival compared to salvage chemotherapy in patients with relapsed or refractory FLT3-ITD positive AML, FLT3 inhibitor monotherapy is rarely curative and responses are short-lived[248–250]. There is a clinical need for approaches to deepen the initial response to FLT3 inhibitors, enabling longer-lasting clinical responses.

An approach to improving responses to FLT3 inhibitors in AML is to simultaneously target aberrant FLT3 activity and its downstream mediators. A major driver of mutant-FLT3-dependent oncogenesis is the MYC proto-oncoprotein (6–8)[105,106]. MYC, a critical regulator of proliferation and differentiation, is over-expressed in the vast majority of patients with AML[251]. Reactivation of MYC-controlled oncogenic networks by the bone marrow microenvironment promotes FLT3 inhibitor resistance[116,251]. These findings suggest that improved responses to FLT3 inhibitors may be achieved with combination strategies that target MYC-dependent proliferative programs.

Direct inhibition of MYC has been an objective of anti-cancer therapeutic development for over the last twenty years. However, MYC has been considered undruggable due to its intrinsically disordered nature and lack of enzymatic activity[252]. Another approach is to instead disrupt the molecular mechanisms that drive *MYC* over-expression. In blood cells, *MYC* expression is regulated by a blood-specific super-enhancer complex (BENC), which is bound by numerous transcription factors and global chromatin activators[122,253]. Recent studies in AML cell lines have demonstrated that small molecule inhibitors targeting these activating chromatin complexes, including BRD4, resulted in a loss of *MYC* expression and leukemia cell death[127–130]. However, initial clinical trials have only shown modest clinical activity and substantial toxicity[254].

An alternate approach is to simultaneously target two factors that regulate *MYC* gene expression. The chromatin regulatory protein lysine specific demethylase 1 (LSD1) is a well-established regulator of *MYC* gene expression[132–134]. LSD1 regulates gene expression by removing activating methylation marks on lysine 4 of histone 3 (H3K4) and repressive methylation marks on lysine 9 of histone 3 (H3K9) or by recruiting repressive complexes to gene promoters[141,142]. Inhibitors of LSD1 have been shown to decrease MYC abundance and activity in AML cell lines and primary samples[132–134]. Our prior work and that of other groups shows that LSD1 inhibition augments the efficacy of

kinase inhibitors in AML[133,134,153]. However, the extent to which synergy exists between LSD1 and FLT3 inhibition and the underlying mechanism of drug synergy has not been investigated.

Here we report *ex vivo* drug screening data on a cell line model of FLT3-ITD positive AML and primary FLT3-ITD positive AML samples demonstrating that LSD1 inhibition potentiates the efficacy of FLT3 inhibition. Using high-sensitivity epigenetic profiling, we establish that dual FLT3/LSD1 inhibition disrupts regulatory factor binding at the MYC BENC, resulting in a loss of *MYC* expression. Using short-term *ex vivo* culture, we confirm that these transcriptional and epigenetic responses to combined FLT3/LSD1 inhibition occur in primary FLT3-ITD positive leukemic blasts. Collectively, this data reveals how epigenetic therapies augment the activity of kinase inhibitors in FLT3-ITD AML.

4.3 Results

4.3.1 Combined FLT3/LSD1 inhibition synergistically represses MYC transcriptional programs, while activating PU.1 transcriptional programs

Prior work from our lab and others suggest that the combination of kinase and LSD1 inhibition may be an effective therapeutic strategy in multiple molecular subtypes of AML[132–134,153]. To establish whether this approach is effective for FLT3-ITD AML, we treated FLT3-ITD-positive (MOLM13 and MV4;11) and FLT3-ITD-negative (K562) cell lines with multiple FLT3/LSD1 inhibitor combinations. We observed potent synergy between the FLT3 inhibitors and LSD1 inhibitors in the FLT3-ITD-positive cell lines but not in the FLT3-ITD-negative cell lines, suggesting that this drug combination has

specificity for FLT3-ITD-positive AML (Figure 4.1A-B; Figure S4.1). In addition, we observed that the drug combination increased both early (Annexin V+/PI-) and late (Annexin V+/PI+) apoptosis populations with minimal toxicity to healthy CD34+ cells (Figure S4.2A-C). These results indicate that synergy exists between FLT3/LSD1 inhibition in FLT3-ITD-positive AML.

To understand the mechanism of synergy, we performed RNA-seq on MOLM13 cells treated with quizartinib, GSK-2879552, or the combination for 24 hours. The majority of the 1,716 differentially expressed genes (80%; 1,374 of 1,716 genes) following the drug combination were not differentially expressed following single agent therapies (Figure S4.2D). Unsupervised clustering of differentially expressed genes revealed clusters of genes that were either up-regulated or down-regulated by the drug combination (Figure S4.2E). Transcription factor target gene enrichment analysis of the down-regulated genes revealed an enrichment of MYC target genes (Figure 4.1C; Figure S4.2F). To understand the potential impact of these down-regulated genes on cell viability, we investigated essential genes nominated from genome-wide CRISPR dropout screens performed in MOLM13 cells[255]. There were 74 genes, including *MYC,* that were differentially down-regulated by the drug combination and identified as depleting genes (Figure S4.2G). Notably, these genes were also enriched for MYC target genes (Figure S4.2H). In contrast, the genes up-regulated by the drug combination were enriched for SPI1/PU.1 targets. The *SPI1* gene encodes the transcription factor PU.1, which is critical for coordinating myeloid differentiation[152]. To further corroborate these gene expression profiles, gene set enrichment analysis (GSEA) was performed, which revealed depletion of MYC target genes in cells treated with the drug combination (Figure 4.1D and E). Collectively, we observed that the combination activates an anti-proliferative and pro-differentiative transcriptional program with repression of MYC target genes and activation of PU.1 target genes.

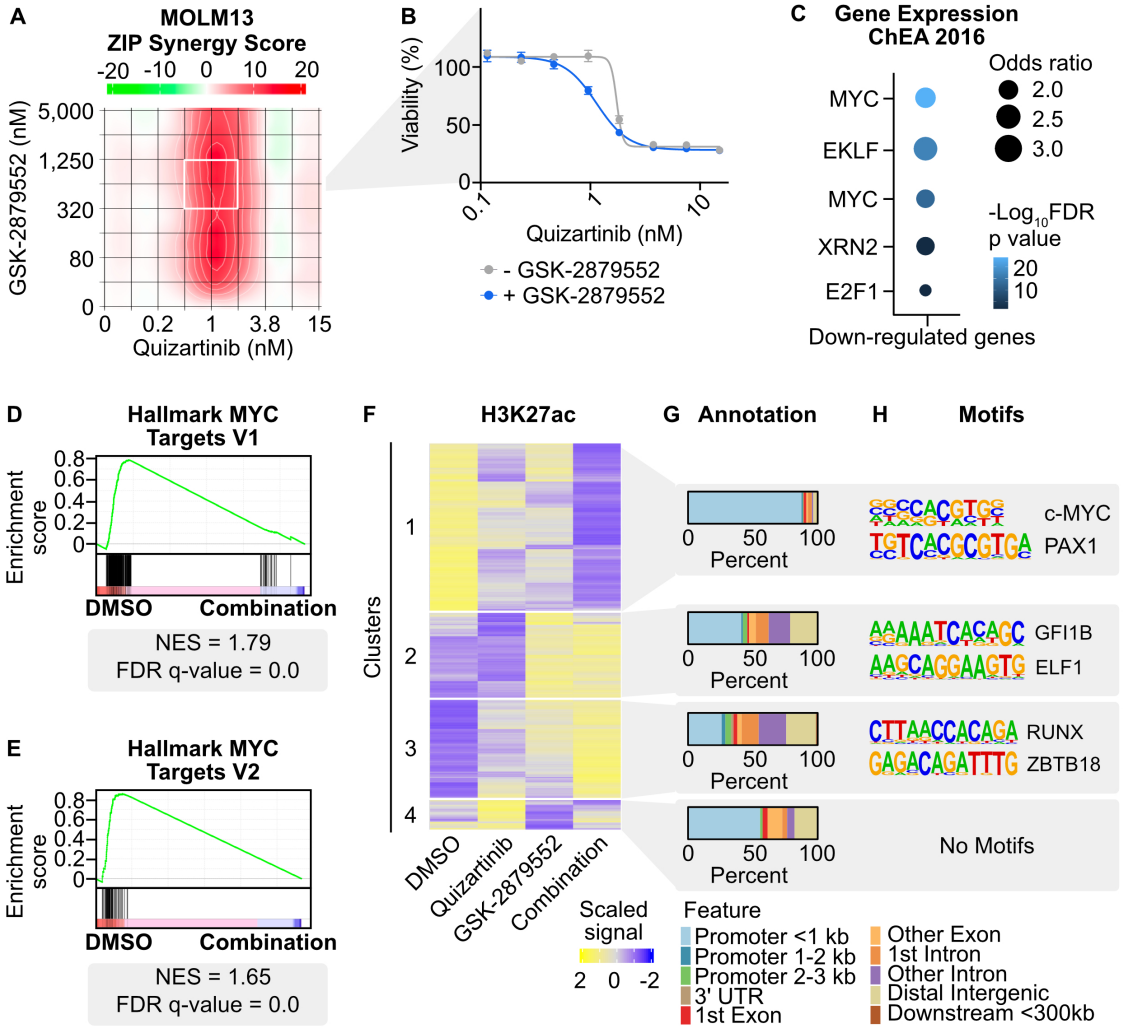

Figure 4.1: Transcriptional and chromatin dynamics in response to combined FLT3/LSD1 inhibition in FLT3-ITD-positive AML.

A, MOLM13 cells were treated in triplicate with an 8x8 dose matrix of quizartinib and GSK-2879552 for 72 hours prior to viability assessment by CellTiter Aqueous colorimetric assay. Zero interaction potency (ZIP) synergy scores were calculated on the average values for each drug dose. The white box indicates the quizartinib and GSK-2879552 concentrations corresponding to maximal synergy. **B,** Quizartinib response curves with and without GSK-2879552 (638 nM, which is the concentration corresponding to maximal synergy in (A)). The GSK-2879552 response curve with and without quizartinib is shown in Figure S4.1. **C,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO vehicle for 24 hours prior to RNA sequencing. Analysis was performed on genes with decreased expression with the drug combination relative to DMSO.

D, E, GSEA was performed comparing the drug combination to DMSO. **F,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO vehicle for 2 hours prior to CUT&Tag for H3K27ac. Unsupervised hierarchical clustering of regions with differential signal following drug treatment. **G,** Annotation of regions in clusters from (F). **H,** Motif enrichment of regions with differential H3K27ac signal. Top two de novo motifs with p-value <10-12 are shown.

4.3.2 Combined FLT3/LSD1 inhibition disrupts chromatin dynamics at distinct genomic loci

A key component of LSD1 inhibitor activity has been ascribed to displacement of GFI1/CoREST from chromatin and re-activation of enhancers associated with differentiation[256]. Therefore, to characterize the early chromatin dynamics following combined FLT3/LSD1 inhibition, we utilized cleavage under targets and tagmentation (CUT&Tag)[31]. We used CUT&Tag to assess changes in acetylation of histone 3 lysine 27 (H3K27ac), a marker of transcriptional activation, in MOLM13 cells 2 hours following drug treatment. Unsupervised clustering of the regions with differential H3K27ac signal revealed four clusters (Figure 4.1F). The regions in cluster 1 were associated with repressed H3K27ac signal by the drug combination and were primarily localized to promoters (Figure 4.1G). The cluster 1 regions were also enriched for MYC motifs, consistent with the findings of decreased gene expression of MYC target genes (Figure 4.1H). An example of a down-regulated region is observed at the PVT1 promoter, a known regulator of *MYC* expression (Figure S4.2I)[257]. Cluster 2 contained regions with increased H3K27ac signal largely driven by LSD1 inhibition. Cluster 2 regions were nearly equally distributed at promoter and non-promoter elements and were enriched for GFI1/GFI1B motifs. Cluster 3 regions were also localized at promoter and non-promoter elements and showed enrichment for RUNX motifs. RUNX1 is a critical regulator of myeloid differentiation and potentiates the transcriptional activation activity of PU.1[258].

An example of up-regulated non-promoter regions from clusters 2 and 3 were observed upstream of the lysozyme promoter, which is expressed in mature granulocytes (Figure S4.2J). These data collectively show that the drug combination alters the chromatin landscape at distinct genomic loci. Furthermore, based on pathway analysis and motif enrichment, MYC, GFI1, RUNX1, and PU.1 transcription factors are candidate regulators of these chromatin dynamics.

4.3.3 Chromatin segmentation reveals that MYC-, STAT5-, and PU.1-driven molecular programs underlie the response to combined FLT3/LSD1 inhibition

Our CUT&Tag results revealed substantial changes in histone acetylation at both promoters and outside promoters in response to combined FLT3/LSD1 inhibition, arguing that both types of regulatory elements have distinct roles in the drug response. We therefore profiled a series of covalent histone marks in MOLM13 cells 6 hours following drug treatment. These marks enabled the segmentation of chromatin into promoters and enhancers. Trimethylation of histone 3 lysine 4 (H3K4me3) is primarily localized at promoters, whereas monomethylation of histone 3 lysine 4 (H3K4me1) is predominantly at enhancers[28]. We also profiled H3K27ac at this same time point to understand transcriptional activation at prompters and enhancers[31]. Following LSD1 inhibition, we observed regions of increased H3K4me3 and H3K4me1 signal consistent with the known demethylase activity of LSD1 for H3K4me1/3 (Figure S4.3A-D)[137,256]. Unsupervised clustering of regions with differential H3K27ac signal at promoters and enhancers revealed multiple patterns of regulation. Similar to the global 2-hour acetylation CUT&Tag data, we observed a large cluster of repressed H3K27ac signal at

promoters (cluster P2) that were enriched for MYC motifs as well as a cluster of increased H3K27ac signal (cluster P1) enriched for GFI1/GFI1B motifs (Figure 4.2A and B). GO analysis revealed that cluster P2 regions are associated with cell cycle and proliferation (Figure S4.3E). At enhancers, we identified a cluster of suppressed H3K27ac signal (cluster E3) associated with STAT5 motifs along with a cluster of increased H3K27ac signal (cluster E4) enriched for SPI1/PU.1 motifs (Figure 4.2C and D; Figure S4.3F). This data suggests that combined FLT3/LSD1 inhibition simultaneously activates GFI1/GFI1B-bound promoters and represses the activation of MYC-bound promoters. In parallel, the drug combination suppresses STAT5-bound enhancers and activates PU.1-bound enhancers.

To better characterize the transcription factors driving the response to combined FLT3/LSD1 inhibition, we profiled the genome-wide binding of multiple candidate transcription factors following single or dual drug treatment using chromatin immunoprecipitation sequencing (ChIP-seq), cleavage under targets and release using nuclease (CUT&RUN) and CUT&Tag. We then examined the signal of these factors at each cluster, identifying promoters and enhancers with differential H3K27ac signal following drug treatment (Figure 4.2E and F). We observed modest enrichment of LSD1 at all acetylated promoter and enhancer regions. MYC signal was most pronounced at promoters (cluster P2) and enhancers (cluster E3) associated with H3K27ac signal suppressed by the combination. In both clusters, a greater loss of MYC signal was observed with combined FLT3/LSD1 inhibition compared with no drug or single drug controls. STAT5 binding was localized to enhancers with differential H3K27ac signal that were down-regulated by quizartinib and/or enriched for STAT5 motifs (clusters E1 and E3), which is consistent with studies demonstrating that STAT5 is a primary downstream target of FLT3 inhibitors[259]. Notably, combined FLT3/LSD1 inhibition resulted in a global depletion of STAT5 signal, including at clusters E1 and E3 (Figure S4.3G). PU.1 and

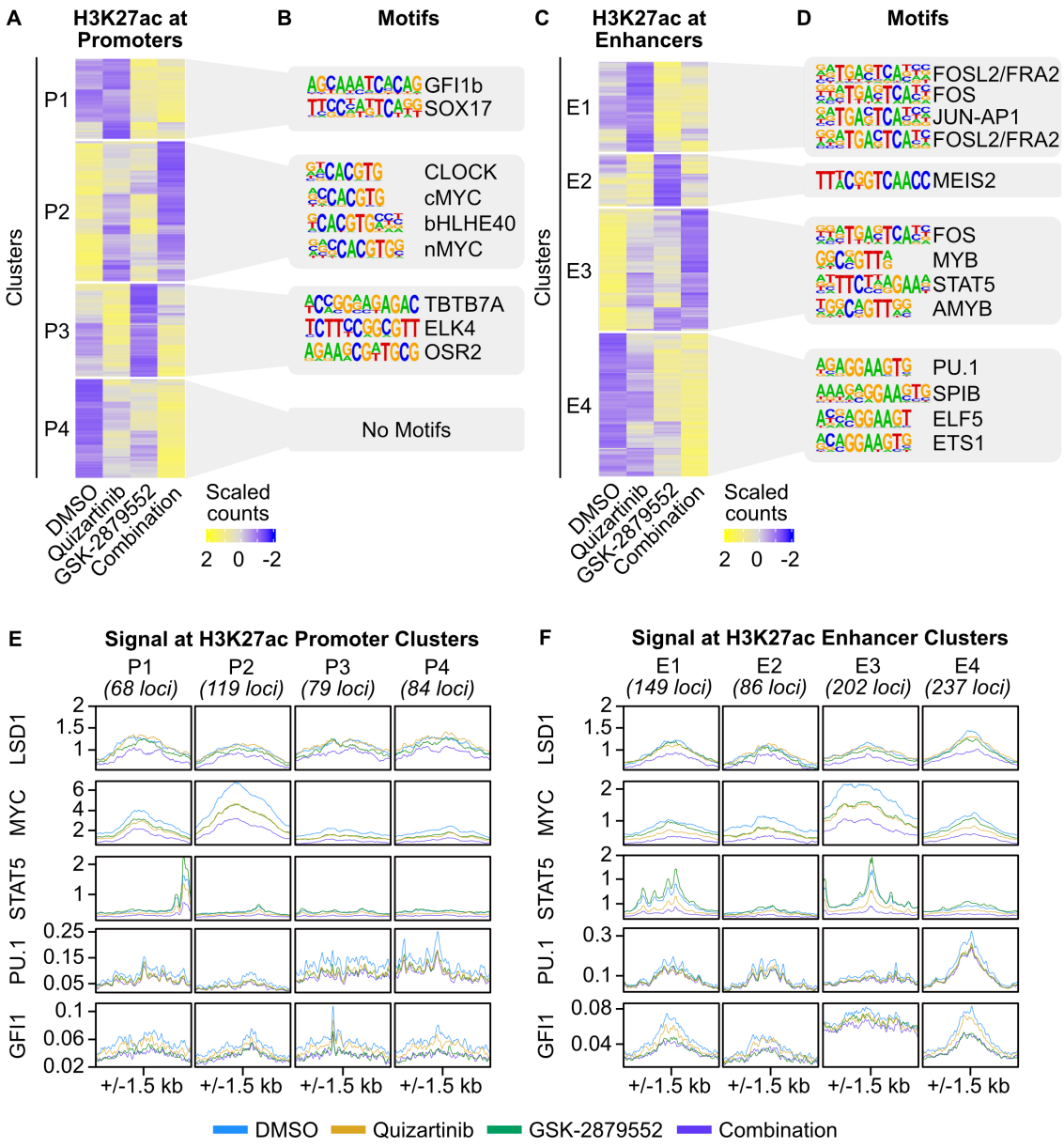

Figure 4.2: Discrete components of the response to FLT3/LSD1 inhibition are mediated by promoters and enhancers.

A, MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO vehicle for 6 hours prior to CUT&Tag for H3K27ac, H3K4me1, and H3K4me3. On the basis of these marks, chromatin was segmented into promoters and enhancers. Unsupervised hierarchical clustering of differential H3K27ac signal at promoters. **B,** Motif enrichment of promoters with differential H3K27ac signal. Top four de novo motifs with p-value <10-12 are shown.

C, D, Same analyses as (A) and (B) were performed at enhancers. **E,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 6 hours. LSD1, MYC, and STAT5 binding was assessed by ChIP-seq. PU.1 and GFI1 binding was assessed by CUT&RUN. Transcription factor binding profiles at promoters with differential H3K27ac identified in (A). **F,** Transcription factor profiles at enhancers with differential H3K27ac identified in (C).

GFI1 showed specific enrichment at enhancers with increased differential H3K27ac signal following LSD1 inhibition (clusters E1 and E4). LSD1 inhibition led to a loss GFI1 signal at these clusters, consistent with previously reported displacement of GFI1 from chromatin upon LSD1 inhibition[256]. While RUNX1 and CEBPA were enriched at both promoters and enhancers with differential H3K27ac signal, they did not demonstrate appreciable changes in signal following drug treatment (Figure S4.3H and I). Collectively these results implicate MYC, STAT5, PU.1, and GFI1 in the synergistic cytotoxicity of combined FLT3/LSD1 inhibition.

4.3.4 Loss of *MYC* expression is critical for the response to combined FLT3/LSD1 inhibition

Our transcriptomic and epigenetic analyses nominated MYC as a key driver of the molecular responses to combined FLT3/LSD1 inhibition. We observed that combined FLT3/LSD1 inhibition results in the suppression of *MYC* transcript abundance and in a genome-wide decrease in MYC binding (Figure 4.3A and B). We confirmed that the transcriptional suppression of *MYC* was associated with a decrease in MYC protein abundance (Figure S4.4A). Modulation of MYC target gene expression can be influenced both by changes in MYC gene abundance and activity. MYC regulates the transcription of cell cycle proteins through the recruitment of pause-released factors to poised RNA Polymerase II (RNA PolII)[260]. We observed that the drug combination increased RNA PolII signal at MYC-bound gene transcription start site (TSS; Figure

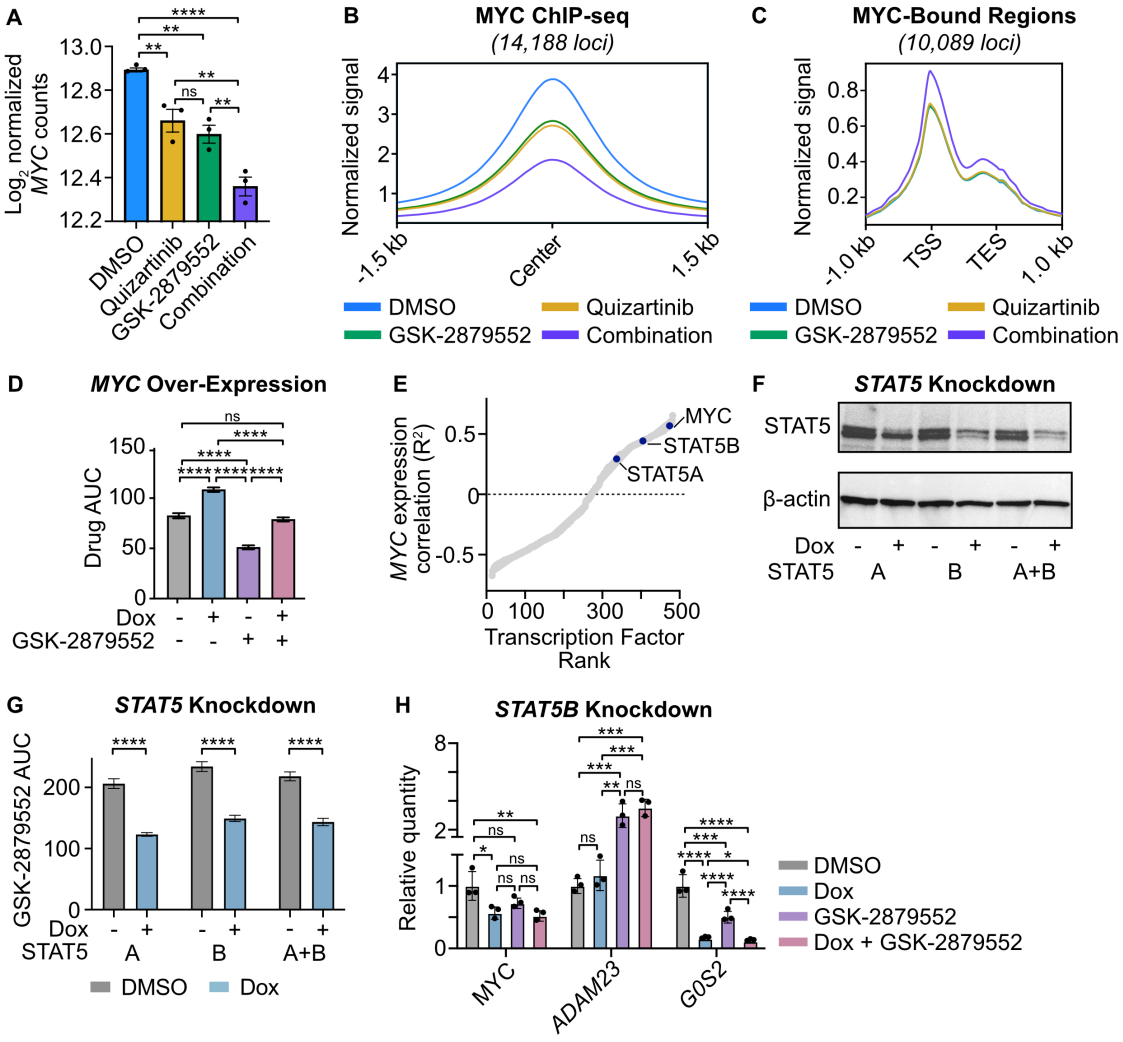

Figure 4.3: *MYC* expression is suppressed by combined FLT3/LSD1 inhibition and is associated with STAT5 regulatory activity.

A, Normalized *MYC* counts from RNA-seq presented in Figure 4.1. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **B,** MYC binding profile at consensus peaks from MOLM13 ChIP-seq presented in Figure 4.2. **C,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO vehicle for 6 hours prior to CUT&Tag for RBP1. RBP1 binding profile at RBP1 and MYC co-bound regions. **D,** MOLM13 cells were transduced with lentiviral particles harboring a doxycycline-inducible *MYC* expression vector. Cells were treated with doxycycline (1 µg/mL) or DMSO for 48 hours and then plated in an 8x8 matrix of quizartinib and GSK-2879552 for 72 hours prior to viability assessment by CellTiter Aqueous colorimetric assay.

AUC data from the 311 nM GSK-2979552 isoline (the concentration corresponding to maximal synergy in the *MYC* over-expressed MOLM13 cells) is shown. Dose responses and synergy over the entire drug matrix is shown in Figure S4. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **E,** Spearman's correlation of normalized *MYC* gene counts and predicted transcription factor activity scores. Activity scores were inferred from baseline RNA-seq performed on patients in the Beat AML cohort. Transcription factors are ranked by goodness of fit (R^2). **F,** MOLM13 cells were transduced with lentiviral particles harboring a doxycycline-inducible *STAT5* short hairpin RNA (shRNA) knockdown vector. Western blot for STAT5 and β-actin following treatment with doxycycline (1 μg/mL) or DMSO for 48 hours. **G,** GSK-2879552 AUC of MOLM13 *STAT5* knockdown cells treated with doxycycline (1 μg/mL) or DMSO for 72 hours. The GSK-2879552 response curves are shown in Figure S4.5. Statistical significance was determined by Student's t-test. **H,** qPCR assessment of gene expression in MOLM13 cells expressing a doxycycline-inducible *STAT5B* shRNA. Cells were treated with doxycycline (1 μg/mL) for 48 hours prior to the addition of GSK-2879552 (100 nM) for 24 hours. Expression was normalized to *GUSB* as an endogenous control. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$, TSS = transcription start site, TES = transcription end site.

4.3C). Although the drug combination also results in increased RNA PolII signal at gene bodies, the RNA-PolII pause index shows that the accumulation of paused RNA PolII at gene TSS exceeds the amount present at gene bodies (Figure S4.4B and C). Overall, this data indicates that the drug combination disrupts the ability of MYC to promote RNA PolII pause release at its target. In addition, we observed an increase in TP53 protein levels, and an enrichment of a phosphoprotein network controlled by TP53 following the drug combination, consistent with repression of MYC-dependent cell cycle regulation (Figure S4.4D). Together, these findings suggest a mechanism of combined FLT3/LSD1 that suppresses MYC expression and activity.

To evaluate the importance of *MYC* expression to the mechanism of dual FLT3/LSD1 inhibition, we derived a MOLM13 cell line with a doxycycline-inducible *MYC* expression construct (Figure S4.4E and F). *MYC* over-expression resulted in decreased sensitivity to the drug combination (Figure 4.3D; Figure S4.4G-I). In addition, *MYC* over-expression attenuated the induction of apoptosis by the drug combination (Figure S4.4J). These data suggest that forced expression of *MYC* partially abrogates the effect of combined FLT3/LSD1 inhibition.

To identify potential regulators of *MYC* gene expression in AML that may mediate the response to the combination, we analyzed transcription factor activity from RNA-seq on 681 primary AML samples. We generated scores for each sample that reflected the predicted activity of 468 different transcription factors. Correlation of these transcription factor activity scores with *MYC* gene expression revealed a strong positive correlation with STAT5 transcription factor activity (Figure 4.3E). As FLT3 is a known activator of STAT5, we generated MOLM13 cell lines with perturbed STAT5 activity to evaluate its role in the response to combined FLT3/LSD1 inhibition[259]. Knockdown of *STAT5A* and/or *STAT5B* resulted in increased sensitivity to GSK-2879552 (Figure 4.3F and G; Figure S4.5A-F). Furthermore, we observed synergy between *STAT5* knockdown and GSK-2879552, demonstrating that a loss of STAT5 activity is sufficient to recapitulate a portion of the quizartinib effect (Figure S4.5G-J). *STAT5* knockdown resulted in reduced expression of *MYC* as well as dysregulated expression of drug combination response markers, *ADAM23* and *G0S2*, which were nominated from our RNA-seq analysis (Figure 4.3H; Figure S4.5K and L). To evaluate whether a decreased STAT5 activity is necessary for drug combination efficacy, we created MOLM13 cells stably expressing *Stat5a1*6* (Figure S4.6A and B). Mutant Stat5a1*6 harbors two point mutations, rendering it constitutively active[261,262]. *Stat5a1*6* expression markedly attenuated synergy between FLT3 and LSD1 inhibition, demonstrating that STAT5 is a critical target

of the drug combination (Figure S4.6C-F). Collectively, these results show that the STAT5-MYC axis plays a major role in the response to dual FLT3/LSD1 inhibition.

4.3.5 FLT3 inhibition suppresses STAT5 binding to the MYC blood super-enhancer

To identify the mechanism by which STAT5 regulates *MYC* expression in FLT3-ITD AML, we examined our STAT5 ChIP-seq data. While we did not identify a STAT5 binding event at the MYC promoter, ranking STAT5 peaks by normalized signal revealed a strong binding event at the MYC BENC consistent with previous findings (Figure 4.4A)[122,263]. Notably, the MYC BENC was among the cluster of enhancers associated with suppressed H3K27ac signal and STAT5 signal depletion following combined FLT3/LSD1 inhibition (cluster E3). STAT5-bound elements within the MYC BENC showed a significant decrease in H3K27ac signal after treatment with quizartinib or the drug combination (Figure 4.4B). To characterize the changes in accessibility of the MYC BENC we performed assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)[257]. This analysis revealed a loss of accessibility across all modules in response to drug combination treatment (Figure 4.4C-E). Evaluation of STAT5 binding and H3K27ac signal revealed several sub-modules that also display dynamic behavior in response to drug treatment but have not been previously characterized (A0.1-0.5, C1, G1). Collectively, our findings nominate the MYC BENC as a crucial locus for down-regulation of *MYC* gene expression by dual FLT3/LSD1 inhibition.

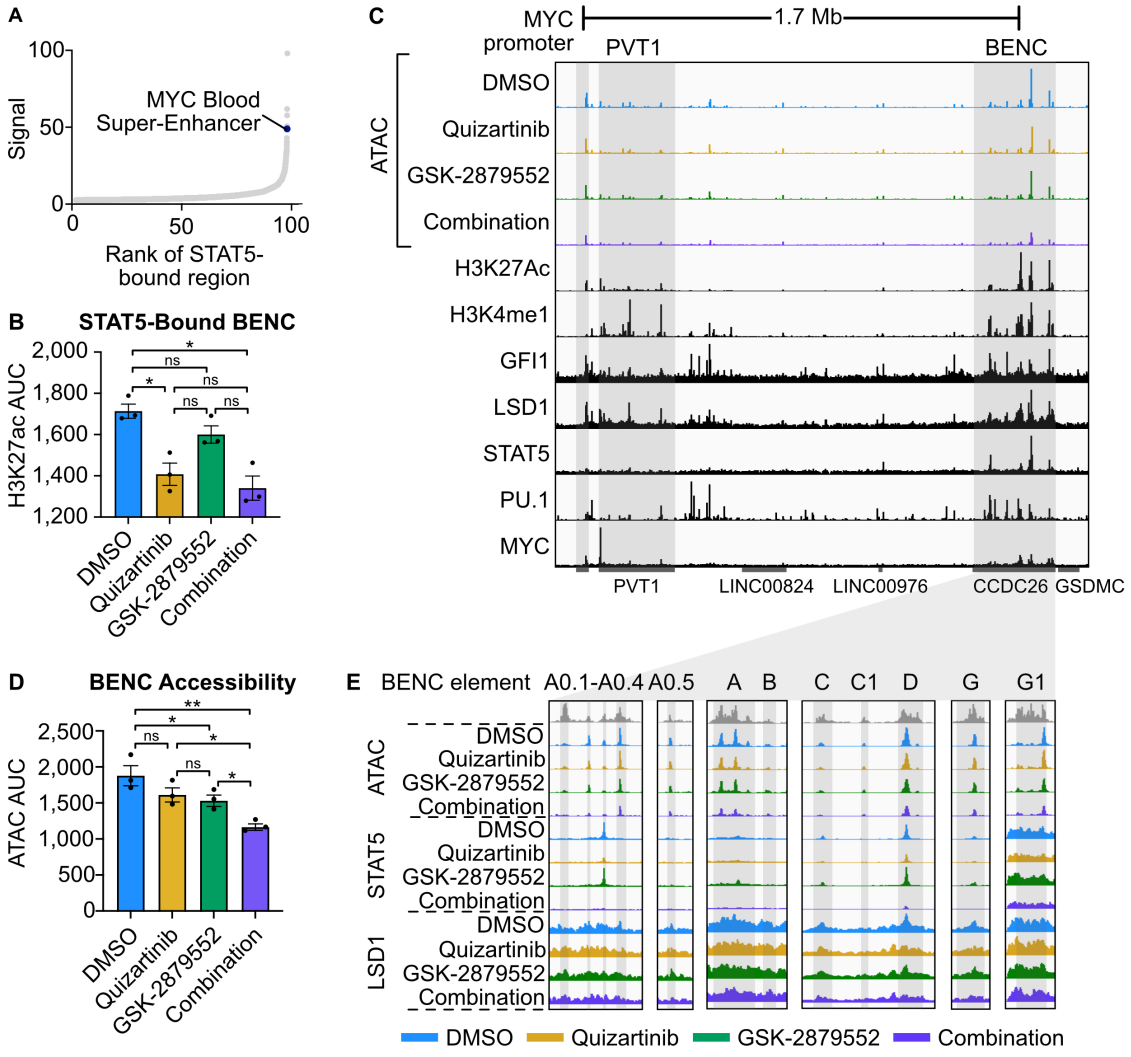

Figure 4.4: FLT3-Inhibition represses *MYC* expression through a loss of STAT5 binding to the MYC blood super-enhancer cluster.

A, STAT5-bound regions from Figure 4.2 ranked by ChIP-seq signal. **B,** H3K27ac CUT&Tag signal AUC at STAT5-bound BENC elements. H3K27ac signal data is from MOLM13 cells in Figure 4.1 whereas the STAT5 signal data is from MOLM13 cells in Figure 4.2. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **C,** ATAC-seq was performed on MOLM13 cells quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 24 hours. Representative histone modification and transcription factor tracks (from DMSO conditions in Figure 4.2) shown at the extended *MYC* locus. **D,** ATAC signal AUC at all MYC BENC modules. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction.

E, ATAC, STAT5 ChIP-seq, and LSD1 ChIP-seq signal at twelve BENC modules. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

4.3.6 LSD1 inhibition represses the expression of MYC and its target genes by altering GFI1 and histone modification dynamics

Our data demonstrates that dual FLT3/LSD1 inhibition suppresses *MYC* gene expression by displacement of STAT5 from the MYC BENC. However, the MYC BENC is bound by many other transcription factors, indicating that drug combination efficacy may be dependent on interruption of MYC BENC-bound factors in addition to STAT5. Prior studies have shown that LSD1 inhibitor monotherapy decreases *MYC* expression[132–134]. A critical component of LSD1-inhibitor efficacy is the disruption of LSD1 scaffolding of GFI1 from the CoREST transcription repressor complex[256]. Examination of our GFI1 CUT&RUN data confirmed that GFI1 is bound to the MYC BENC but is disrupted by the drug combination at module C (Figure 4.5A; Figure S4.7A-E). To evaluate the importance of GFI1 in the response to combined FLT3/LSD1 inhibition, we generated MOLM13 cell lines with doxycycline-inducible knockdown of *GFI1* (Figure S4.7F and G). We found that *GFI1* knockdown increased sensitivity to FLT3 inhibition and enhanced FLT3-inhibitor-dependent repression of *MYC* and its target genes (Figure 4.5B and C; Figure S4.7H and I). Collectively, this data indicates that displacement of GFI1 binding by LSD1 inhibition reduces *MYC* expression and is important to combined FLT3/LSD1 inhibitor response.

Our data demonstrates that dual FLT3/LSD1 inhibition exerts a portion of its activity via a STAT5- and GFI1-dependent decrease in *MYC* gene expression. However, *MYC* over-expression only partially reverses the impact of combination treatment,

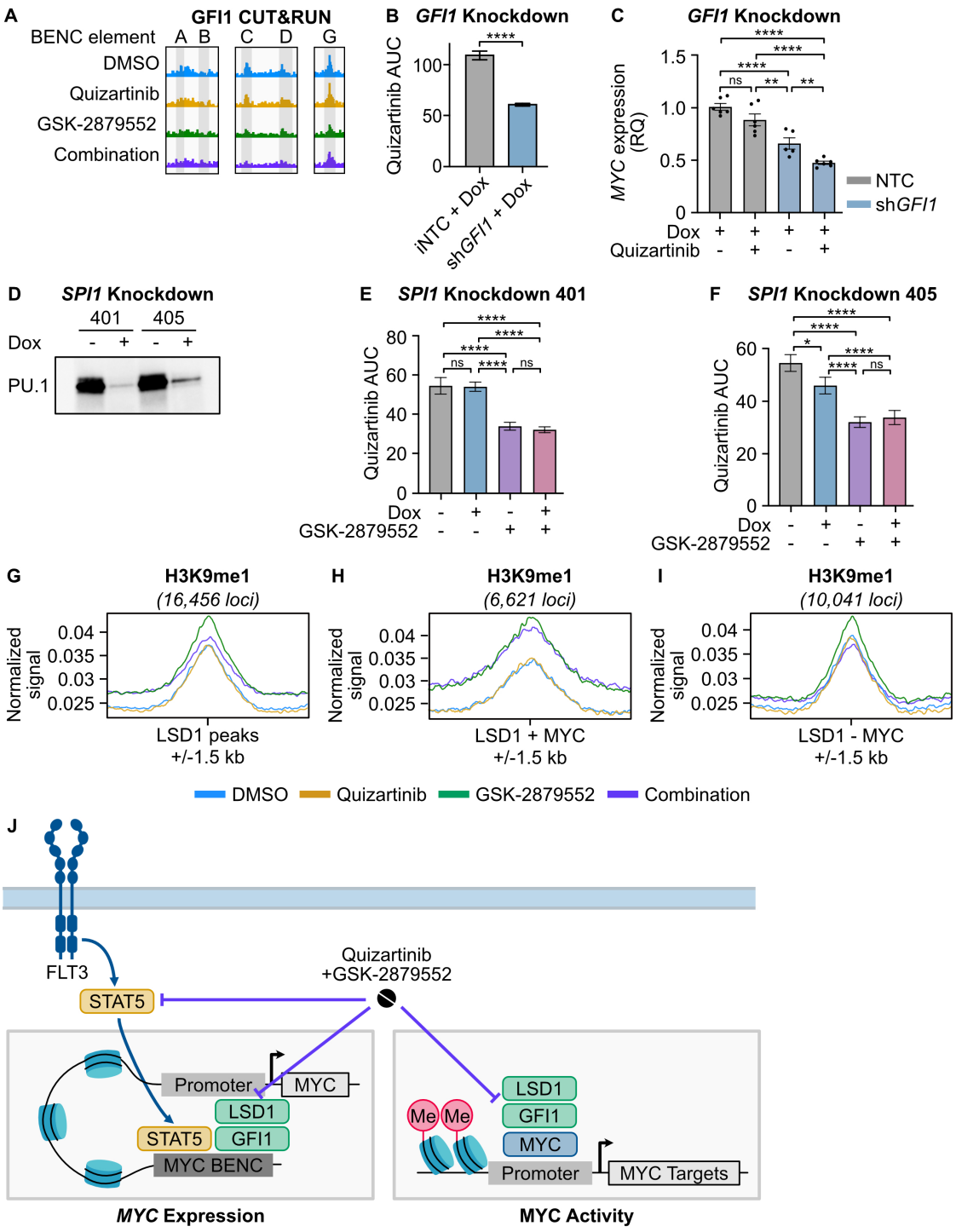

Figure 4.5: LSD1 inhibition disrupts GFI1 binding at the MYC BENC and induces a gain of H3K9me1 binding at MYC-bound promoters.

A, GFI1 CUT&RUN signal from Figure 4.2 at five BENC modules. **B,** MOLM13 cells were transduced with lentiviral particles harboring a doxycycline-inducible non-targeting codon (NTC) or *GFI1* shRNA knockdown vector.

Quizartinib AUC of cells treated with doxycycline (1 μg/mL) or DMSO for 72 hours. Substantial knockdown was observed in the absence of doxycycline treatment, so only doxycycline-treated samples were compared. The quizartinib response curves are shown in Figure S4.7. Statistical significance was determined by Student's t-test. **C,** qPCR assessment of gene expression in cells treated with doxycycline (1 μg/mL) for 48 hours prior to the addition of quizartinib (1 nM) for 24 hours. Expression was normalized to *GUSB* as an endogenous control. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **D,** MOLM13 cells were transduced with lentiviral particles harboring a doxycycline-inducible *SPI1* shRNA knockdown vector. Western blot for PU.1, which is encoded by *SPI1,* in cells treated with doxycycline (1 μg/ml) or an equivalent volume of DMSO for 48 hours. **E, F,** Cells were treated with doxycycline (1 μg/mL) or DMSO for 48 hours and then plated in an 8x8 matrix of quizartinib and GSK-2879552 for 72 hours prior to viability assessment. AUC data from the 311 nM GSK-2979552 isoline (the concentration corresponding to maximal synergy in the *SPI1* knockdown MOLM13 cells) is shown. Dose responses and synergy over the entire drug matrix is shown in Figure S4.9. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **G-I,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), or the combination for 6 hours prior to CUT&Tag for H3K9me1. Normalized signal for H3K9me1 at LSD1-bound regions, LSD1 and MYC co-bound regions, and at regions bound by LSD1 but not MYC. **J,** Schematic describing the drug combination mechanism. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

suggesting the involvement of additional mechanisms. Prior work has shown that LSD1 inhibitor efficacy is also dependent on activation of PU.1-bound enhancers and subsequent induction of myeloid differentiation[150]. Our transcriptional and epigenetic data shows that combined FLT3/LSD1 inhibition results in the activation of PU.1 target genes and acetylated enhancers enriched for PU.1 motifs. Therefore, we evaluated whether this transcriptional signal resulted in immunophenotypic differentiation of AML blasts. Drug treatment did result in a modest increase in CD11b expression, however the majority of treated cells remained CD11b negative (Figure S4.8). To evaluate whether PU.1-dependent activation of this differentiation-associated gene expression program

was necessary for drug effect, we evaluated the impact of *SPI1* (gene coding for PU.1) knockdown. PU.1-deficient cells demonstrated no significant reduction in drug synergy, arguing that the PU.1-driven pro-differentiation gene expression program is dispensable for the cytotoxic drug effect (Figure 4.5D-F; Figure S4.9A-F).

In other cell types, LSD1 plays a role in gene activation via removal of repressive mono- and demethylation of histone 3 lysine 9 (H3K9me1/2)[139]. To evaluate this possible mechanism, we profiled the genome-wide distribution of H3K9me1 using CUT&Tag . LSD1 inhibition, with or without FLT3 inhibition, resulted in an accumulation of H3K9me1 at MYC target genes co-bound with LSD1 (Figure 4.5G-I). This was accompanied by a loss of the reciprocal activating mark acetylated histone 3 lysine 9 (H3K9ac), consistent with the observed decrease in the expression of MYC target genes (Figure S4.9G). These findings suggest that dual inhibition of FLT3/LSD1 exerts locus-specific effects on the chromatin landscape by interrupting STAT5 and GFI1/CoREST transcriptional regulation as well as altering the balance of repressive H3K9 marks at MYC-bound promoters (Figure 4.5J).

4.3.7 Efficacy of combined FLT3/LSD1 inhibition in primary AML samples

To understand the activity of dual FLT3/LSD1 inhibition in primary AML patients, we performed a 3-day *ex vivo* drug assay on 72 primary AML samples. Nearly every sample (94%; 68 of 72 samples) demonstrated a synergistic increase in efficacy of dual agent therapy over single agents alone (Figure 4.6A and B; Figure S4.10A). Although synergy was observed regardless of FLT3 mutation status, the mean quizartinib AUC was lower in FLT3-ITD-positive samples (140.1) as compared to FLT3-wildtype samples (172.4) or samples harboring a FLT3 mutation other than FLT3-ITD (168.2; Figure

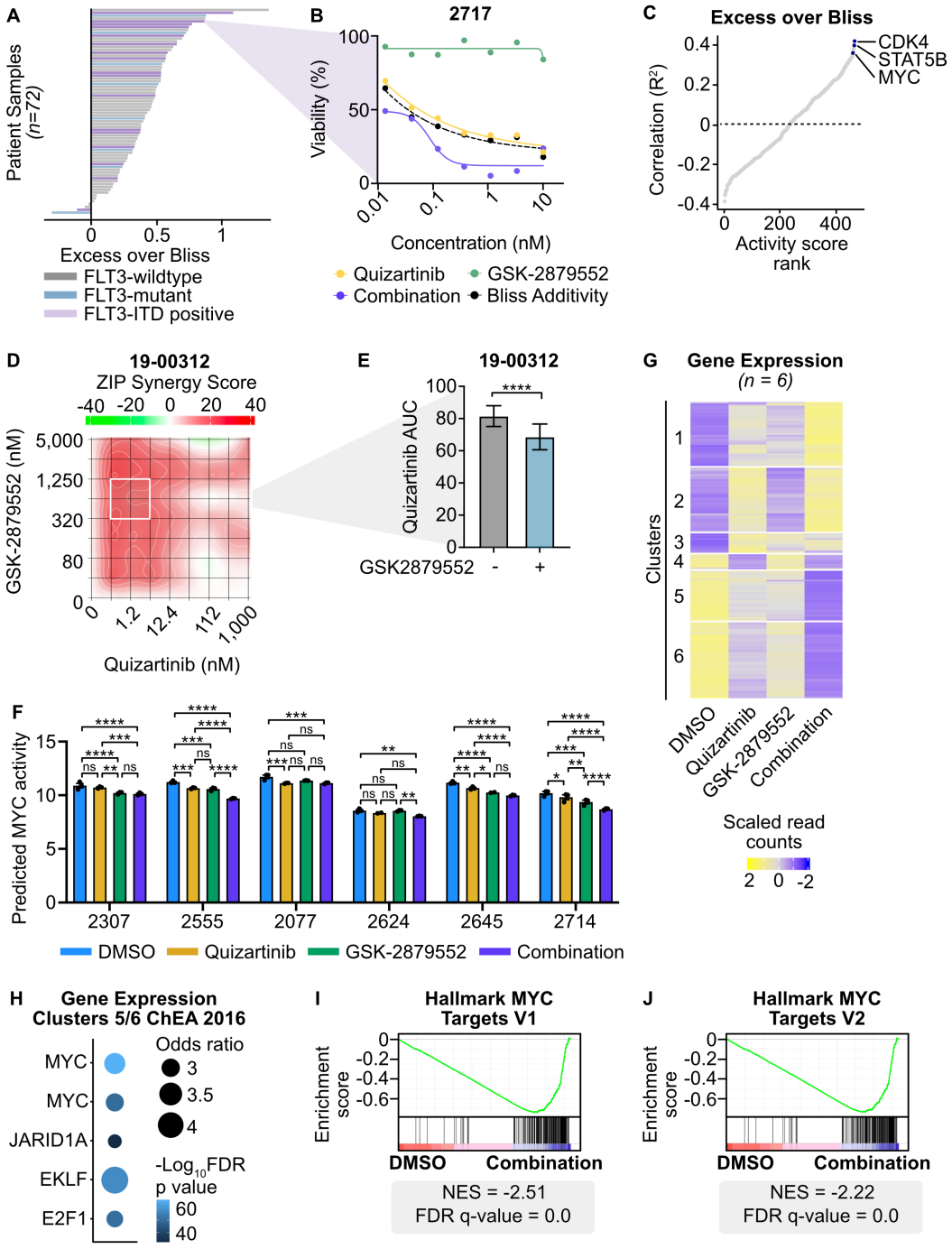

Figure 4.6: Combined FLT3/LSD1 inhibition drives synergistic cell death by repressing a MYC-dependent transcriptional network in primary AML blasts.

A, Primary AML blasts from 72 total samples (18 FLT3-ITD-positive) were cultured for 72 hours along a 7-point curve with either quizartinib, GSK-2879552, or equimolar amounts of the drug combination.

Cell viability was assessed by CellTiter Aqueous colorimetric assay. Excess over Bliss was calculated using cell viability at corresponding drug concentrations. Each bar represents the mean excess over Bliss across all concentrations. Bar color indicates FLT3 mutation status. **B,** Dose response curves for quizartinib, GSK-2879552, and the drug combination in a FLT3-ITD-positive AML sample from (**A**). **C,** Spearman's correlation of excess over Bliss and predicted transcription factor activity. Transcription factors were ranked by goodness of fit (R^2). **D,** Primary blasts from a FLT3-ITD-positive AML sample were treated in triplicate with an 8x8 dose matrix of quizartinib and GSK-2879552 for 72 hours prior to viability assessment by CellTiter Aqueous colorimetric assay. ZIP synergy scores were calculated on the average values for each drug dose. **E,** AUC data from the 628 nM GSK-2979552 isoline (the concentration corresponding to maximal synergy in (**D**)) is shown. Statistical significance was determined by Student's t-test. **F,** Bulk RNA-seq was performed on six FLT3-ITD-positive patient samples treated in triplicate with 500 nM quizartinib, 500 nM GSK-2879552, both drugs in combination, or an equivalent volume of DMSO for 24 hours. MYC transcription factor activity was inferred from RNA-seq. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **G,** Unsupervised hierarchical clustering of differentially expressed genes following drug treatment. **H,** Transcription factor target enrichment from clusters in (**G**). **I, J,** GSEA was performed comparing the drug combination to DMSO. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

S4.10B). To characterize the determinants of response to combined FLT3/LSD1 inhibition, we generated transcription factor activity scores on baseline RNA sequencing performed on the cohort. Predicted MYC and STAT5B transcription factor activity were among the strongest correlates with the degree of combination synergy (Figure 4.6C). CDK4, a known transcriptional target of MYC, was the strongest correlate[264]. These findings reveal that AML samples with high baseline MYC activity have the greatest sensitivity to combined FLT3/LSD1 inhibition.

To characterize the response to combined FLT3/LSD1 inhibition in patient samples, we performed drug sensitivity studies and RNA-seq on 6 FLT3-ITD-positive patient samples treated with single or dual agent therapy. Similar to the pattern of

synergy observed in cell lines, we found drug synergy across a broad range of doses (Figure 4.6D and E). In addition, the drug combination synergistically induced apoptosis in primary AML blasts (Figure S4.10C and D). MYC expression and predicted activity was down-regulated in all samples by the drug combination, although differing patterns of individual drug effect were observed (Figure 4.6F; Figure S4.10E). Unsupervised hierarchical clustering of differentially expressed genes across all 6 samples revealed a similar pattern to what was observed in MOLM13 cells (Figure 4.6G). Nearly a quarter of the differentially expressed genes in the patient samples were also identified in the MOLM13 cells (26.0% of up-regulated genes and 18.4% of down-regulated genes; Figure S4.10F and G). Transcription factor target analysis revealed suppression of MYC target genes and activation of SPI1/PU.1 target genes following combined FLT3/LSD1 inhibition (Figure 4.6H; Figure S4.10H). Finally, GSEA revealed the drug combination decreased expression of MYC target genes and increased expression of differentiation-associated genes (Figure 4.6I and J; Figure S4.10I). Collectively, this data confirms the findings of MYC gene expression and transcription factor activity from AML cell lines in primary AML blasts.

4.3.8 Dual FLT3/LSD1 inhibition suppresses MYC super-enhancer accessibility in primary AML blasts

Our initial investigations in AML cell lines demonstrate that suppression of chromatin accessibility at the MYC BENC is an important mechanism of cytotoxicity produced by combined FLT3/LSD1 inhibition. To validate this mechanism in primary AML, we performed single-cell ATAC-seq on primary AML blasts from three FLT3-ITD-positive patients 24 hours after treatment with combined quizartinib and GSK-2879552 or DMSO. Cell type identification using a reference dataset and predicted *CD34*

expression revealed that the samples were largely of GMP-like myeloid blasts (Figure S4.11A-F). Clustering revealed population shifts in response to treatment (Figure 4.7A-F; Figure S4.11G-I). Assessment of accessibility at the MYC BENC showed that these population shifts were associated with decreased accessibility at most BENC modules (Figure 4.7G-J). However, each sample showed a distinct pattern of change. Sample 2684 showed a decrease in the majority of modules, 2645 demonstrated the most prominent decrease in modules A and G, and sample 2603 exhibited the strongest decrease in modules F and G. To understand the regions that are associated decreased accessibility following drug combination treatment, we calculated the peak score fold change among peaks identified in each condition. In sample 2684, ranking by $\log_2$ peak score fold change revealed that MYC BENC module C was amongst the regions with the greatest decrease in accessibility (Figure 4.7K). Collectively, these results demonstrate that suppression of MYC BENC accessibility is a conserved feature of the response to dual FLT3/LSD1 inhibition in primary AML samples. However, patient-to-patient heterogeneity does exist, suggesting diversity in the regulatory factors that sustain MYC BENC accessibility in primary AML samples.

4.3.9 Combined FLT3/LSD1 inhibition disrupts human LSC colony formation

Our single-cell ATAC-seq data indicates that dual FLT3/LSD1 inhibition targets GMP-like leukemic blasts. Previous studies have established that leukemic blasts originate from a small number of leukemia stem cells (LSCs)[265]. It has been suggested that to fully eradicate the bulk disease, new therapeutic strategies must not only target leukemic blasts, but also eliminate residual LSCs[266]. Therefore, we wanted to understand the impact of combined FLT3/LSD1 inhibition on primary patient LSCs.

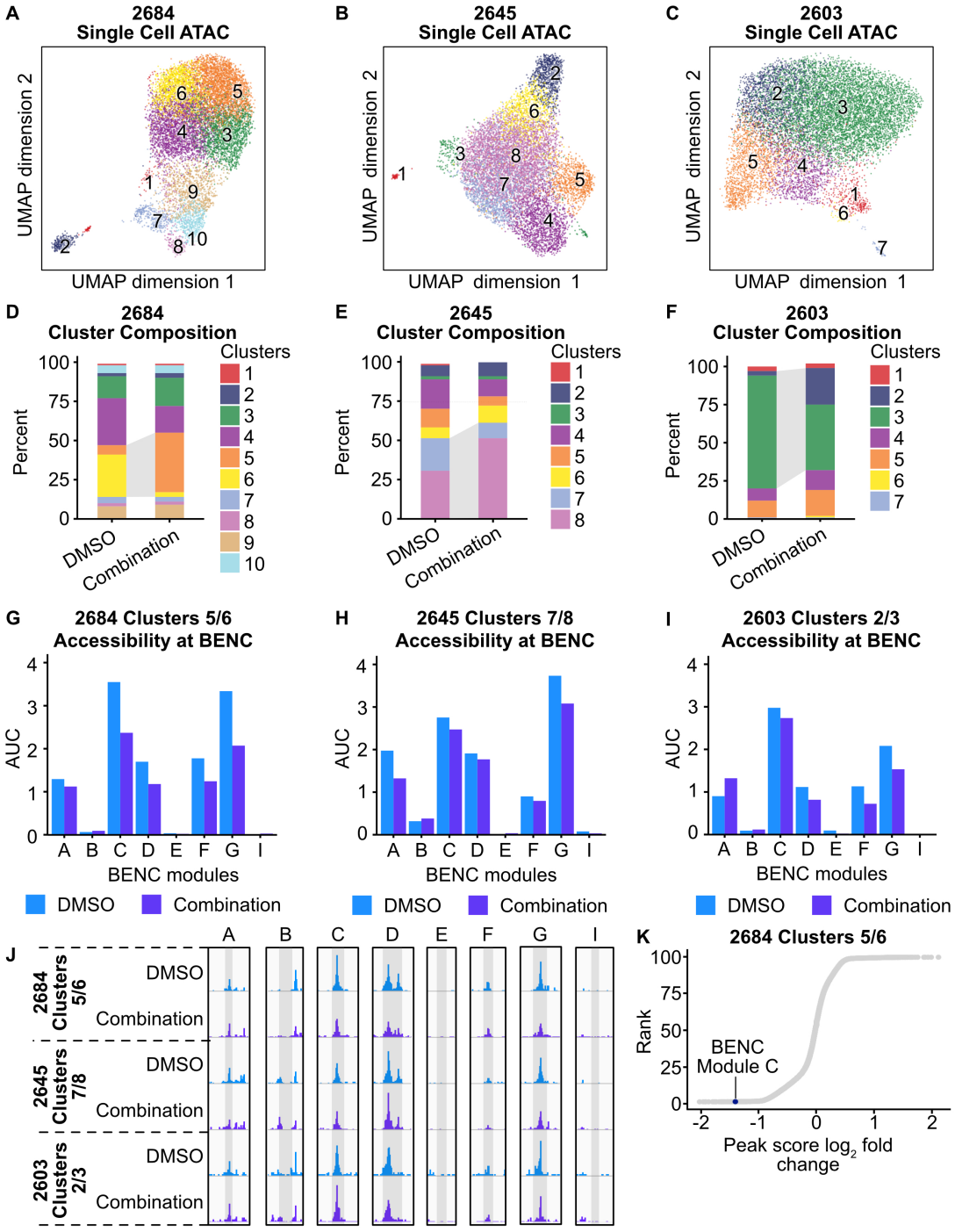

Figure 4.7: Dual FLT3/LSD1 inhibition results in a shift from a MYC super-enhancer-high to a MYC super-enhancer-low cell state in primary AML blasts.

A-C, Single-cell ATAC-seq was performed on three AML patient samples following treatment with quizartinib (500 nM) and GSK-2879552 (500 nM) or an equivalent volume of DMSO for 24 hours.

UMAP of DMSO-treated and drug-treated cells colored by cluster. D-F, Percent of cells assigned to each cluster. Dynamic clusters were identified as the populations that shift between DMSO-treated and drug-treated conditions. Dynamic clusters are highlighted with gray shading between bars. G-I, AUC of accessibility at each BENC module. J, Pseudo-bulked accessibility at the MYC BENC modules separated by treatment condition. K, Peak score fold change was calculated between peaks in DMSO-treated and combination-treated cells within dynamic clusters. Peaks are ranked by log2(peak score fold change).

LSCs have been immunophenotypically characterized using cell surface markers. We utilized fluorescence-activated cell sorting (FACS) to enrich LSCs from three primary patient samples. Specifically, we isolated hematopoietic stem and progenitor cells (CD34+CD38-) expressing at least one previously identified LSC marker (IL1RAP, CD123, CD45RA; Figure S4.12A)[267–270]. We sorted 40,000 enriched LSCs and performed colony forming assays in the presence of drug (Figure S4.12B). We observed a significant decrease in the number of colonies from LSCs treated with combined FLT3/LSD1 as compared to those exposed to vehicle (Figure S4.12C and D). This data suggests that combined FLT3/LSD1 disrupts primary patient clonogenicity.

4.4 Discussion

Activating mutations in FLT3 are amongst the most common molecular events in AML[246]. While FLT3 inhibitors are clinically available, they produce only modest improvements in survival[249,250]. Here, we demonstrated that LSD1 inhibition potentiates the efficacy of FLT3 inhibition in FLT3-ITD AML cell lines and primary cell blasts. High-resolution transcriptomic and epigenetic profiling revealed that the mechanism of synergy is in part due to depletion of regulatory transcription factor binding, STAT5 and GFI1, at the MYC BENC. Moreover, we identified additional evidence that dual FLT3/LSD1 inhibition results in the accumulation repressive H3K9me1 marks at MYC-

controlled proliferation genes. These findings reveal how epigenetic therapies augment the activity of kinase inhibitors in FLT3-ITD AML.

A crucial component to the mechanism of FLT3/LSD1 inhibitor synergy was altering *MYC* expression through regulation of the MYC BENC (Figure 4 and Figure 4.7). While others have demonstrated that *MYC* transcription can by altered by inhibiting general chromatin regulators, disruption of MYC BENC activity by combined epigenetic modulatory drugs and kinase inhibitors is a novel approach to targeting this central oncogenic regulator[127–130]. Our single-cell ATAC-seq analysis revealed substantial variation in the pattern of MYC BENC module utilization between AML samples at baseline and in response to drug treatment. Indeed, other studies have suggested that each BENC module is bound by a distinct set of transcription factors and regulates *MYC* expression in specific blood cell lineages[122]. Understanding MYC BENC module utilization between molecularly-defined AML subtypes and its impact drug responses is an important question for future studies.

Prior work on LSD1 inhibitors has largely implicated the pro-differentiation effects of these drugs as the central mechanism of cytotoxicity. Our work here shows that LSD1 inhibition activates enhancers that are associated with PU.1 (Figure 4.2). Other groups have shown that suppression of *SPI1* expression results in a block in LSD1-inhibitor-induced differentiation and decreased cytotoxicity[150]. While our work confirmed the role of PU.1 as a putative mediator of LSD1-inhibitor responses, we found that *SPI1* knockdown had little effect on the transcriptional or cytotoxic response to dual FLT3/LSD1 inhibition (Figure 4.5). Moreover, combined FLT3/LSD1 inhibition did not seem to have a major effect on myeloid differentiation (Figure S4.8). It is unclear whether the PU.1-associated transcriptional effects observed in our study are important to the drug effect. Investigation of the pro-differentiation effects of dual FLT3/LSD1 inhibition will be an important question for future investigation.

Previous studies of LSD1 inhibitors have also demonstrated that drug efficacy is dependent on the interruption of LSD1 scaffolding activity rather than its demethylation activity[132,151]. Our work confirmed that a critical component of LSD1 inhibitor activity is the disruption of LSD1 binding to GFI1/CoREST (Figure 4.5). However, LSD1 inhibition also resulted in the accumulation of repressive H3K9me1 marks at the promoters of MYC target genes. While LSD1 canonically demethylates activating H3K4 marks, alternative LSD1 complexes remove repressive H3K9 methylation marks in cells from other tissues, resulting in transcriptional activation[139,140]. In prostate cancer, LSD1 forms a chromatin-associated complex with androgen receptor that demethylates H3K9 and de-represses androgen receptor target genes. In neuronal cells, on the other hand, an LSD1 isoform, LSD1+8a, complexes with supervillain and demethylates H3K9me2 to regulate neuronal differentiation. Interestingly, the H3K9 demethylation activity of LSD1 may be slightly antagonized by FLT3 inhibition as H3K9me1 signal in cells treated with the drug combination was lower than in those only treated with LSD1 inhibition. Further work needs to be done to nominate binding factors with LSD1 or LSD1 isoforms that, as a complex, functions as a transcriptional activator by H3K9 demethylation and how these complexes are affected by FLT3 inhibitors.

Collectively, our work demonstrates that LSD1 inhibition enhances the activity of FLT3 inhibition in FLT3-ITD AML. The efficacy of the drug combination is dependent on the simultaneous disruption of STAT5 and GFI1 from the MYC blood super-enhancer complex, resulting in repressed *MYC* expression, as well as the accumulation of repressive H3K9me1 at MYC target genes.

4.5 Methods

4.5.1 Cell and Patient Sample Culture

4.5.1.1 Cell Lines

MOLM13 cells (DSMZ) were cultured in RPMI (Gibco) supplemented with 10% fetal bovine serum (FBS, HyClone), 2 mM GlutaMAX (Gibco), 100 units/mL Penicillin, and 100 µg/mL Streptomycin (Gibco). MV4;11 and K562 cells (ATCC) were cultured in IMDM supplemented with 20% fetal bovine serum (FBS, HyClone), 2 mM GlutaMAX (Gibco), 100 units/mL Penicillin, and 100 µg/mL Streptomycin (Gibco). All cells were cultured at 5% CO_2 and 37°C. Cell lines were tested for mycoplasma concentration at the time of freezing as well as monthly for any cell lines in culture. Cell lines were authenticated by the OHSU Cell Line Authentication service, which uses Promega's GenePrint 10 system to confirm the identity of human cell lines by short tandem repeat analysis. All cell clines were maintained below fifteen passages.

4.5.1.2 Patient Samples

All patients gave written informed consent to participate in this study, which was conducted in accordance with the Declaration of Helsinki and had the approval and guidance of the institutional review boards at Oregon Health & Science University (OHSU), University of Utah, University of Texas Medical Center (UT Southwestern), Stanford University, University of Miami, University of Colorado, University of Florida, National Institutes of Health (NIH), Fox Chase Cancer Center and University of Kansas. Samples were sent to the coordinating center (OHSU; IRB#9570; #4422; NCT01728402) where they were coded and processed. Mononuclear cells were isolated by Ficoll gradient centrifugation from freshly obtained bone marrow aspirates or

peripheral blood draws. Clinical, prognostic, genetic, cytogenetic, and pathologic lab values as well as treatment and outcome data were manually curated from patient electronic medical records. Genetic characterization of the leukemia samples included results of a clinical deep-sequencing panel of genes commonly mutated in hematologic malignancies (Sequenome and GeneTrails [OHSU]; Foundation Medicine [UT Southwestern]; Genoptix; and Illumina). Patient samples were cultured in RPMI with 10% FBS and 10% HS-5 conditioned media (ATCC) or SFEMII supplemented with 1x StemSpan CD34+ Expansion Media and 1 μM UM729 (StemCell Technologies).

4.5.1.3 Colony Assay

Whole bone marrow was obtained (AllCells) and CD34+ cells were selected using CD34 MicroBead Kit (Miltenyi Biotec) according to manufacturer's instructions. For the colony assay, 500 CD34+ cells were used per replicate and plated in MethoCult™ H4435 Enriched (StemCell Technologies). The four groups were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or DMSO. Plates were incubated for 14 days in 5% CO_2 and 37°C. Samples were imaged using STEMvision (StemCell Technologies) and blinded prior to counting by another investigator by assigning letters randomly. ImageJ (NIH) was used to count colonies after blinding.

4.5.1.4 Drug Synergy

Drug synergy was assessed using an 8 x 8 matrix of drug concentrations. Cells were treated for 72 hours prior to MTS assay to evaluate viability. Cell viability was used to calculate drug synergy with SynergyFinder based on the ZIP reference model[271].

4.5.1.5 RNA Interference

Two SMARTvector Inducible short hairpin RNAs (shRNAs) for Human *SPI1* (V3IHSHER_10431275, V3IHSHER_10642739), two for *STAT5A* (V3IHSHEG_6691183, V3IHSHEG_4988581) two for *STAT5B* (V3IHSHER_4778243, V3IHSHER_6411380) and two for *GFI1* (V3IHSHER_5266412 and V3IHSHER_5697821) in a hEF1a-TurboRFP or hEF1a-TurboGFP (*STAT5A*) backbone were obtained from Horizon Discovery. Both *SPI1* shRNA constructs showed effective knockdown of *SPI1*. *STAT5A* V3IHSHEG_6691183 was specific for *STAT5A*, *STAT5A* V3IHSHEG_4988581 was ineffective against either *STAT5A* or *STAT5B*, *STAT5B* V3IHSHER_6411380 was selective for *STAT5B*, and V3IHSHER_4778243 knocked down both *STAT5A* and *STAT5B*. *GFI1* V3IHSHER_5266412 produced effective *GFI1* knockdown while V3IHSHER_5697821 was ineffective. Lentivirus was produced by transfecting Lenti-X 293T cells (Clontech) with FuGENE (Promega #E2311) and Opti-MEM (ThermoFisher #31985062) as well as the SMARTvector transfer plasmid and packaging/pseudotyping plasmids. psPAX2 was a gift from Didier Trono (Addgene plasmid #12260) and pMD2.G was a gift from Didier Trono (Addgene plasmid #12259). The supernatants containing lentivirus was collected after 48 hours of culture and filtered with a 0.45 um filter. MOLM13 cells were transduced with virus via spinnoculation in the presence of polybrene. Transduced cells were selected with 1 µg/mL puromycin to produce a stable cell line.

4.5.1.6 MYC Over-expression

For human *MYC* over-expression, pDONR223_MYC_WT was a gift from Jesse Boehm & Matthew Meyerson & David Root (Addgene plasmid #82927) and cloned into pCW57.1, a gift from David Root (Addgene plasmid #41393). Lentiviral particles were

generated as above and MOLM13 cells were selected after viral transduction with 1 µg/mL puromycin. After selection, cells were treated with 1 µg/mL doxycycline for 48 hours prior to experiments.

*4.5.1.7 Stat5a1*6 Over-expression*

For *Stat5a1*6* over-expression, pMXs-IRES-Puro (pMX Empty) was acquired from Cell Biolabs Inc. (#RTV-014) and pBABE-Stat5a1*6 was acquired from Addgene (#130668). Retroviral particles were generated by transfecting 293T17 cells (ATCC Number #CRL-3216) with FuGENE (Promega #E2311) and Opti-MEM (ThermoFisher #31985062) as well as pCMB-VSV-G (Addgene plasmid #8454), pUMCV (Addgene plasmid #8449), and the appropriate transfer plasmid. The supernatants containing retrovirus was collected after 48 hours of culture and filtered with a 0.45 um filter. MOLM13 cells were transduced with virus via spinnoculation in the presence of polybrene. Transduced cells were selected with 1 µg/mL puromycin to produce a stable cell line.

4.5.1.8 LSC FACS

CD34+ cells were isolated from patient samples using CD34 MicroBead Kit (Miltenyi Biotec) according to manufacturer's instructions. Cells were subsequently stained with Calcein Violet (BioLegend #425203) as well as the following antibodies: CD45RA FITC (BD Pharmigen #55488), IL1-RAP PE (R&D Systems #FAB676P), CD123 PE (BD Pharmigen #554529), CD38 APC (BD Pharmigen #555462), and CD34 APC-Cy7 (BioLegend #343513). Stained cells were analyzed and sorted into cytokine-enriched methylcellulose (MethoCult StemCell #H4435) using a Sony SH800S. 40,000 cells were sorted from each patient sample per replicate (n=3) per drug condition (n=2). The methylcellulose was treated with 500 nM quizartinib and 500 nM GSK-2879552 or

an equivalent volume of DMSO. Plates were incubated for 10 days in 5% CO_2 and 37°C. Samples were imaged using STEMvision (StemCell Technologies) and blinded prior to counting by another investigator by assigning letters randomly. FIJI (NIH) was used to count colonies after blinding.

4.5.2 Sequencing Methods

4.5.2.1 Bulk RNA-Seq

MOLM13 cells were treated with 1 nM quizartinib, 100 nM GSK-2879552, the combination, or equal volume of DMSO for 24h. Total RNA was isolated using a RNeasy Plus Mini Kit (Qiagen). BGI performed the library preparation and sequencing with 50 base pair (bp) single-end (SE) sequencing. Patient samples were cultured in 10% HS-5 CM/RPMI with 20% FBS and treated with 500 nM quizartinib, 500 nM GSK-2879552, the combination, or the equivalent volume of DMSO for 24 hours. Total RNA was isolated with RNeasy Micro kit (Qiagen) according to the manufacturer's instructions. Libraries were prepared using the NEBNext Low Input RNA Library Prep Kit for Illumina (NEB) according to the manufacturer's instructions. Libraries were sequenced by the OHSU Massively Parallel Sequencing Shared Resource (MPSSR) using 100 bp SE sequencing on an Ilumina NovaSeq S1 flow cell.

4.5.2.2 Bulk ATAC-Seq

MOLM13 cells were treated with 1 nM quizartinib, 100 nM GSK-2879552, the combination, or an equivalent volume of DMSO for 24 hours. After treatment, 50,000 cells per replicate were harvested for Fast-ATAC sequencing performed as previously described[272]. In brief, cells were resuspended in cold PBS and tagmentation master mix (25 µL of 2x tagmentation buffer, 2.5 µL of TDE1 [Illumina], 0.5 µL of 1% digitonin; 2x

tagmentation buffer: 66 mM Tris-Acetate, pH 7.8, 132 mM potassium acetate, 20 mM magnesium acetate, 32% v/v N,N-Dimethylformamide) was added. Samples were incubated at 37°C for 30 minutes. DNA was purified using Zymo Clean and Concentrator 5 Kit (Zymo). Transposed DNA was amplified and purified as described previously with adapted primers[273,274]. Samples were quantified using Qubit dsDNA HS Assay Kit (Invitrogen), pooled, and sequenced by BGI using 50 bp paired-end (PE) sequencing.

4.5.2.3 CUT&Tag

MOLM13 cells were treated with 1 nM quizartinib, 100 nM GSK-2979552, the combination, or an equal volume of DMSO for 2 or 6 hours. Benchtop CUT&Tag was performed as previously described[275]. In brief, cells were counted, harvested, and centrifuged for 5 min at 300xg at room temperature. Cells were washed 2X in 1.5 mL wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1x Protease inhibitor cocktail). Concanavalin A magnetic coated beads (Bangs Laboratories) were activated in binding buffer by washing 2X (20 mM HEPES pH 7.5, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$). Washed cells were separated into 100,000 cell aliquots and 10 µL of activated beads were added to each sample. Samples were rotated end-over-end for 7 minutes at room temperature. A magnetic stand was used to separate beads and the supernatant was removed. Primary antibody was diluted 1:50 in antibody buffer (20 mM HEPES pH 7.5, 150mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% digitonin, 2 mM EDTA, 0.1% BSA). The following antibodies were diluted 1:100 in dig-300 buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.01% digitonin) and added to samples as previously described: H3K27ac (Abcam #ab4729), H3K4me1 (CST #5326), H3K4me3 (CST #9751), RBP1 (CST #2629), H3K9me1 (Diagenode #C15410065), H3K9ac (Diagenode #C15410004), CEBPA (CST #8178), and Normal Rabbit IgG (CST #2729)[275]. Samples incubated for 1

hour at room temperature on nutator. Samples were washed 2X with dig-300 buffer then resuspended in tagmentation buffer (dig-300 buffer with 1 mM $MgCl_2$). Samples were incubated at 37°C for 1 hour. DNA was extracted with phenol:chloroform extraction. Samples were amplified by PCR using custom Nextera primers at 400 nM and NEBNext[276]. PCR conditions were set to: 72°C for 5 minutes, 98°C for 30 seconds, 14 cycles of 98°C for 10 sec, 63°C for 10 sec, and 72°C for 1 minute. Libraries were purified with AMPure Beads (Beckman) and sequenced by the OHSU MPSSR on an Ilumina NovaSeq using 50 bp SE sequencing or NextSeq 500 using 37 bp PE sequencing.

4.5.2.4 CUT&RUN

MOLM13 cells were treated with 1 nM quizartinib, 100 nM GSK-2979552, the combination, or an equal volume of DMSO 24 hours. CUT&RUN was performed as previously described[277]. Briefly, concanavalin A magnetic coated beads (Bangs Laboratories) were washed 2x in binding buffer (20 mM HEPES pH 7.5, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$). 500,000 cells per replicate were washed 2x with wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1x Protease inhibitor cocktail). Cells were bound to beads by nutating for 10 minutes at room temperature. Cells were permeabilized and incubated overnight at 4°C on nutator with primary antibody in antibody buffer (wash buffer, 0.001% digitonin, 3 mM EDTA). The following antibodies were used at 1:50 PU.1 (Invitrogen #MA5-15064), GFI1 (Abcam #ab21061), and normal rabbit IgG (CST #2729). Bead slurry was washed 2x with dig wash buffer (wash buffer, 0.001% dig) and resuspended with dig wash buffer and 1x pAG-MNase (Epicypher). Cells were incubated for 10 minutes on nutator at room temperature then washed 2x with dig wash buffer followed by resuspension in pAG-MNase reaction mix (dig wash buffer, 2 mM $CaCl_2$). Bead slurry was incubated for 2 hours at 4°C on nutator. STOP buffer (340 mM NaCl, 20 mM EDTA, 4 mM EGTA, 50 µg/mL RNase A, 50 µg/mL

glycogen, 0.02% dig) was then added, then tubes were incubated at 37°C for 10 minutes. DNA was extracted using phenol:cholorform extraction. Libraries were prepared using NEBNext Ultra II DNA Library Prep Kit (NEB), modified for CUT&RUN as previously described[278]. After adapter ligation fragments were cleaned up with 1.75x AMPure beads (Beckman). Following PCR amplification, libraries were purified 2x with 1.2x AMPure beads to rid of adaptor fragments. Libraries were quantified on the 2100 Bioanalyzer instrument (Agilent) with the High Sensitivity DNA Analysis Kit (Agilent). Libraries were pooled and sequenced by MPSSR on a NextSeq 500 sequencer (Illumina) using 37 bp PE sequencing.

4.5.2.5 ChIP-Seq

ChIP-seq was performed using the SimpleChIP plus Enzymatic Chromatin IP Kit (Cell Signaling Technology). For each replicate, 20 million cells were fixed in 4% formaldehyde (Sigma-Aldrich) for 10 minutes at room temperature then quenched with glycine, washed and stored at -80^{o}C until use. Nuclei were extracted according to the manufactures instructions and treated with 1.25 µL MNase in 500 µL Buffer B at 37^{o}C for 20 minutes. Samples were sonicated on a Qsonic sonicator at 50% amplitude for 5 cycles of 15 sec on 15 sec off on ice. Crosslinks were reversed on a small aliquot of extracted chromatin quantified by OD_{260}. A total of 5 µg of chromatin was used for each immunoprecipitation. The following antibodies were used: LSD1 (Abcam #ab17721), MYC (CST #13987), RUNX1 (Abcam #ab23980), STAT5 (CST #94205S) and rabbit IgG (CST #2729). After overnight incubation, complexes were captured using protein G beads. Crosslinks were reversed and libraries prepped using an NEBNex Ultra II for DNA Library Prep kit. Libraries were sequenced by the OHSU MPSSR. The STAT5 libraries were sequenced by Genewiz using a HiSeqX and 150 bp PE sequencing.

4.5.2.6 Reverse Phase Protein Array (RPPA)

MOLM13 cells were treated for 24 hours with 1 nM quizartinib, 100 nM GSK-2979552, the combination. Cells were washed 2x in PBS then flash frozen. Cell pellets were lysed and processed by the University of Texas MD Anderson Cancer Center Functional Proteomics RPPA Core Facility.

4.5.2.7 Single Cell ATAC-Seq

Patient samples were treated with 500 nM quizartinib and 500 nM GSK-2879552 or an equal volume of DMSO for 24 hours. Nuclei were prepared using the demonstrated protocol for primary cell nuclei extraction from 10x Genomics. ATAC libraries were prepared using Chromium Single Cell ATAC Library and Gel Bead kit v1.1 (10x Genomics, 1000176). Libraries were sequenced with 50 bp PE sequencing by the OHSU MPSSR.

4.5.3 Primary AML Blast Dataset

Gene mutation, drug response, and gene expression data from primary AML blasts was accessed through the Beat AML database[246]. Samples from collected from patients who were in remission or had a history of myelodysplastic syndrome at the time of collection were excluded from downstream analysis. Samples with quizartinib/GSK-2979552 single and dual agent drug response data were selected and stratified by FLT3-ITD mutation status. Gene expression dataset was downloaded in the form of reads per kilobase of exon per million reads mapped (RPKM).

4.5.4 Cell and Patient Sample Culture

4.5.4.1 Flow Cytometry

To assess differentiation, MOLM13 cells were treated with 1 nM quizartinib, 100 nM GSK-2879552, the combination, or an equal volume of DMSO for 72 hours. Cells were stained with CD14 BV421 (BioLegend #301830) and CD11b PE-Cy7 (BD Pharmigen #557743). Stained cells were analyzed using a Sony SH800S. The flow cytometry data was subsequently analyzed using FlowJo.

4.5.4.2 Apoptosis

Apoptosis was assessed 48-72 hours after drug treatment by flow cytometry using an eBioscience Annexin V -APC apoptosis detection kit (ThermoFisher) according to the manufacturer's instructions.

4.5.4.3 Western Blot

For MYC abundance, MOLM13 cells were treated with 1 nM quizartinib, 100 nM GSK-2979552, the combination, or an equivalent volume of DMSO for 24 hours. Cells were lysed using cell lysis buffer (Cell Signaling Technologies [CST] #9803) containing protease inhibitor cocktail (Sigma-Aldrich #11697498001). Lysates were mixed with 3X SDS sample buffer (75 mmol/L Tris [pH 6.8], 3% SDS, 15% glycerol, 8% β-mercaptoethanol, and 0.1% bromophenol blue) then incubated for 5 minutes at 95°C. Samples were run on Criterion 4–15% Tris-HCl gradient gels (Bio-Rad). Gels were transferred to PVDF membranes, then blocked in Tris-buffered saline with 0.1% Tween (TBS-T) and 5% bovine serum albumin (Sigma-Aldrich #A3059-500G). Blots were probed with MYC (CST #13987), PU.1 (Invitrogen #MA5-15064), GFI1 (Abcam #ab21061), STAT5 (CST #94205), H3 (CST #4620), or β-actin (CST #8457) antibodies

at 1:1000 in TBS-T overnight at 4°C. Membranes were incubated for 1 hour at room temperature with HRP-conjugated secondary antibody (CST #7074S) at 1:1500. SuperSignalTM West Pico PLUS Chemiluminescent Substrate (ThermoFisher) was used to develop blots. Imaging was performed on BioRad ChemiDoc Imaging System and analyzed using Image Lab (Bio-Rad).

4.5.5 Data Analysis

4.5.5.1 Bulk RNA-Seq

Raw reads were trimmed with Trimmomatic and aligned with STAR[279,280]. Two MOLM13 replicates were identified as outliers during QC and excluded from downstream analysis. Differential expression analysis was performed using DESeq2[281]. Raw p values were adjusted for multiple comparisons using the Benjamini-Hochberg method. MOLM13-essential genes were identified from a genome-wide CRISPR/Cas9 dropout screen[255]. Significant genes were selected using a false discovery rate (FDR) cut-off of 0.05. Gene set enrichment analyses were performed using Enrichr and Gene Set Enrichment Analysis (GSEA)[282–284]. Enrichr compiles and queries annotated gene set libraries, including ChEA[283–285]. Enrichr, we reported the odds ratio (the likelihood that the gene set of interest is enriched relative to background gene sets) and the $-\log_{10}$ FDR-adjusted p-value. From GSEA, we included the enrichment score, normalized enrichment score, and FDR adjusted q-value. Transcription factor activity scores were generated using Priori[286].

4.5.5.2 Bulk ATAC-Seq Analysis

ATAC libraries were aligned to the human genome (hg38) using BWA-MEM and sorted using SAMtools[287,288]. Duplicates were marked with Sambamba, duplicates and

mitochondrial reads were removed[289]. Counts per million (CPM) normalized tracks were generated with deepTools[290]. Differential accessibility was assessed with DESeq2[281]. Gene ontology (GO) analysis was performed using Genomic Regions Enrichment of Annotations Tool (GREAT)[291]. CPM-normalized tracks were merged with bigWigMerge and visualized using Integrative Genomics Viewer[292].

4.5.5.3 CUT&Tag and CUT&RUN Analysis

CUT&Tag and CUT&RUN libraries were aligned to the human genome (hg38) using Bowtie2 and the following options --local --very-sensitive-local --no-unal --no-mixed --no-discordant --phred33 -I 10 -X 700[293]. Peaks were called using GoPeaks and the following option -mdist 1000[165]. High confidence peaks were defined as those present in at least two replicates. Consensus bed files were formed by merging the high confidence peaks from DMSO, quizartinib, GSK-2979552, and the combination using BEDTools[294]. Differential peaks were identified using DESeq2 with default parameters[281]. Heatmaps were produced using the ComplexHeatmap package from Bioconductor[295]. Peaks were annotated to the nearest feature using ChIPseeker[296]. GO analysis was performed using GREAT[291]. Counts tables for differential peaks were produced using multicov from BEDTools[294]. CPM-normalized tracks, global signal heatmaps, and plot profiles at specified regions were generated using deepTools[290]. Active promoters were defined by the presence of H3K4me3 and H3K27ac within 1000 bp of a TSS. Active enhancers were defined by the presence of H3K4me1 and H3K27ac beyond 1000 bp of a TSS. CPM-normalized tracks were merged with bigWigMerge and visualized using Integrative Genomics Viewer[292]. Signal within 1500 kb upstream and downstream of LSD1-bound peaks as well as those that include or exclude MYC and PU.1 was quantified using computeMatrix reference-point from deepTools[290]. Signal was scaled and visualized with a heatmap using the ComplexHeatmap package from

Bioconductor[295]. RNA polymerase II (RNA PolII) pause indices were calculated by first quantifying RBP1 signal within a -30 to +100 bp window around each transcription start site (TSS) and within the gene bodies. After normalizing the signal by region length, the pause index is calculated by dividing the TSS normalized signal by the gene body signal. Spearman correlation of transcription factor binding sites was assessed using multiBigwigSummary and plotCorrelation from deepTools[290].

4.5.5.4 ChIP-Seq

ChIP-seq libraries were aligned to the human genome (hg38) using bowtie using the following options -v 2 -m 1 --best –strata. Peaks were called using MACS2 callpeak with the -narrow option and IgG library used as the control[41]. Bigwigs for signal visualization were generated by sequentially converting alignments to bed files, big bed files then CPM-normalized bigwig files using UCSC Kent Utilities. Signal at enhancers and promoters was evaluated using deepTools[290]. The STAT5 super-enhancer-like analysis was performed using the HOMER findPeaks -super command[124].

4.5.5.5 Reverse Phase Protein Array (RPPA)

The University of Texas MD Anderson Cancer Center Functional Proteomics RPPA Core Facility linearized, normalized, and batch-corrected (Level 4) the phosphoproteomic data. The data was analyzed with CausalPath (v 1.2.0) using a significant change of mean value transformation, an FDR threshold of 0.2, and 100 permutations for significance[297]. The causal network was visualized on the CausalPath web server causalpath.org.

4.5.5.6 Drug Synergy

Since Beat AML provides cell viability values at set drug concentrations, it was not amendable to ZIP synergy analysis. Instead, we calculated excess over Bliss using the Bliss Independence conditional response curve in base R[298]. The R stats package (v 3.6.2) was used to calculate the spearman correlation between excess over Bliss and predicted transcription factor activity scores from Priori[286].

4.5.5.7 Single Cell ATAC-Seq

To understand sample cell type populations, single-cell ATAC-seq libraries were aligned to the human genome (hg19) using 10x Genomics Cell Ranger[299]. ATAC-seq fragment coverage was quantified in a common coordinate system using Signac FeatureMatrix()[300]. The common coordinate system was created by segmenting the hg19 genome into 5,000 bp bins. Signac was used to generate an integrated object from each fragment file, using a TSS enrichment of three and 1,000 unique fragments as cutoffs[300]. Sample and reference dataset transfer anchors were identified using Signac FindTransferAnchors() and then mapped with Signac MapQuery[300]. The Healthy Hematopoiesis dataset compiled by Granja et al. (https://github.com/GreenleafLab/MPAL-Single-Cell-2019) was the reference dataset used in this analysis. Cells were assigned cell type labels according to their mapping to the reference latent space. Predicted *CD34* expression was inferred using Signac GeneActivity()[300].

To analyze the drug effects on chromatin accessibility, single-cell ATAC-seq libraries were aligned to the human genome (hg38) using 10x Genomics Cell Ranger[299]. ArchR was used in R to generate Arrow files for each fragment file, using a TSS enrichment of four and 5,000 unique fragments as cutoffs[301]. An ArchRProject was

created for each patient sample. Latent semantic indexing (LSI) was preformed using the addIterativeLSI function and clusters were added using the addClusters function. The UMAP for each patient sample, with both drug conditions integrated, was generated. The percent of cells from each treatment condition was used to identify the dynamic clusters. Pseudo-bulk replicates from the dynamic clusters were normalized to TSS enrichment and used to generate bigwigs. The bigwigs were used to produce deepTools matrices to assess accessibility at the MYC BENC modules[290]. The bar plots were generated using ggplot2. Peaks were called on pseudo-bulk replicates from the dynamic clusters using the ArchR addReproduciblePeakSet command and MACS2[41,301]. Dynamic cluster peaksets from the same patient were combined using BEDTools intersect –loj command[294]. Peaks were ranked by their peak score and the $\log_2$(peak score fold change) between dynamic clusters was calculated using a custom script.

4.5.6 Quantification and Statistical Analysis

Values are represented as the mean and error bars are the SEM unless otherwise stated. Prism software (version 9.1; Prism Software Corp.) or R was used to perform statistical analyses. Significance was tested using Student's t-test or two-way ANOVA followed by Holm-Šidák post-test correction unless otherwise stated. For differential analysis of RNA-seq, CUT&Tag, CUT&RUN, and ATAC-seq p-values were adjusted for repeated testing using the Benjamini-Hochberg method.

4.5.7 Data Availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE190785.

4.6 Authors' contributions

W.M. Yashar: conceptualization, software, formal analysis, validation, investigation, visualization, methodology, writing-original draft, writing-review and editing. **B.M. Smith**: conceptualization, software, formal analysis, validation, investigation, visualization, methodology, writing-review and editing. **D.J. Coleman**: conceptualization, formal analysis, validation, investigation, visualization, methodology. **J. VanCampen**: software. **G. Kong**: software. **J. Macaraeg**: validation, investigation, methodology. **J. Estabrook**: software. **E. Demir**: software. **N. Long**: data curation, writing-review and editing. **D. Bottomly**: data curation. **S.K. McWeeney**: data curation. **J.W. Tyner**: resources, data curation, supervision, funding acquisition, writing-review and editing. **B.J. Druker**: conceptualization, resources, supervision, funding acquisition, writing-review and editing. **J.E. Maxson**: conceptualization, resources, supervision, funding acquisition, writing-review and editing. **T.P. Braun:** conceptualization, resources, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing-original draft, project administration, writing-review and editing. The co-first authors may identify themselves as lead authors in their respective CVs.

4.7 Acknowledgements

We would like to thank all of the Beat AML patients for their precious time and donation of samples supporting this research. We appreciate the following OHSU core facilities for their assistance: Advanced Light Microscopy, Flow Cytometry Shared Resource, Massive Parallel Sequencing Shared Resource, ExaCloud Cluster Computational Resource, and the Advanced Computing Center. This research was funded by the National Cancer Institute (K08 CA245224 to T.P. Braun; U54 CA224019

to S.K. McWeeney, J.W. Tyner, B.J. Druker, and J.E. Maxson; 5 R01 CA065823-25 to B.J. Druker; 5 R01 CA247943-02 to J.E. Maxson), the National Institutes of Health (T32 GM109835 and 1F30CA278500-01A1 to W.M. Yashar), the American Society of Hematology (American Society of Hematology Fellow Scholar Award (no number assigned) to T.P. Braun; American Society of Hematology Research Restart Award (no number assigned) to T.P. Braun), the American Cancer Society (P30 CA016672-40 to T.P. Braun), and the Leukemia & Lymphoma Society (Leukemia & Lymphoma Society Scholar Award (no number assigned) to J.E. Maxson).

4.8 Supplementary figures

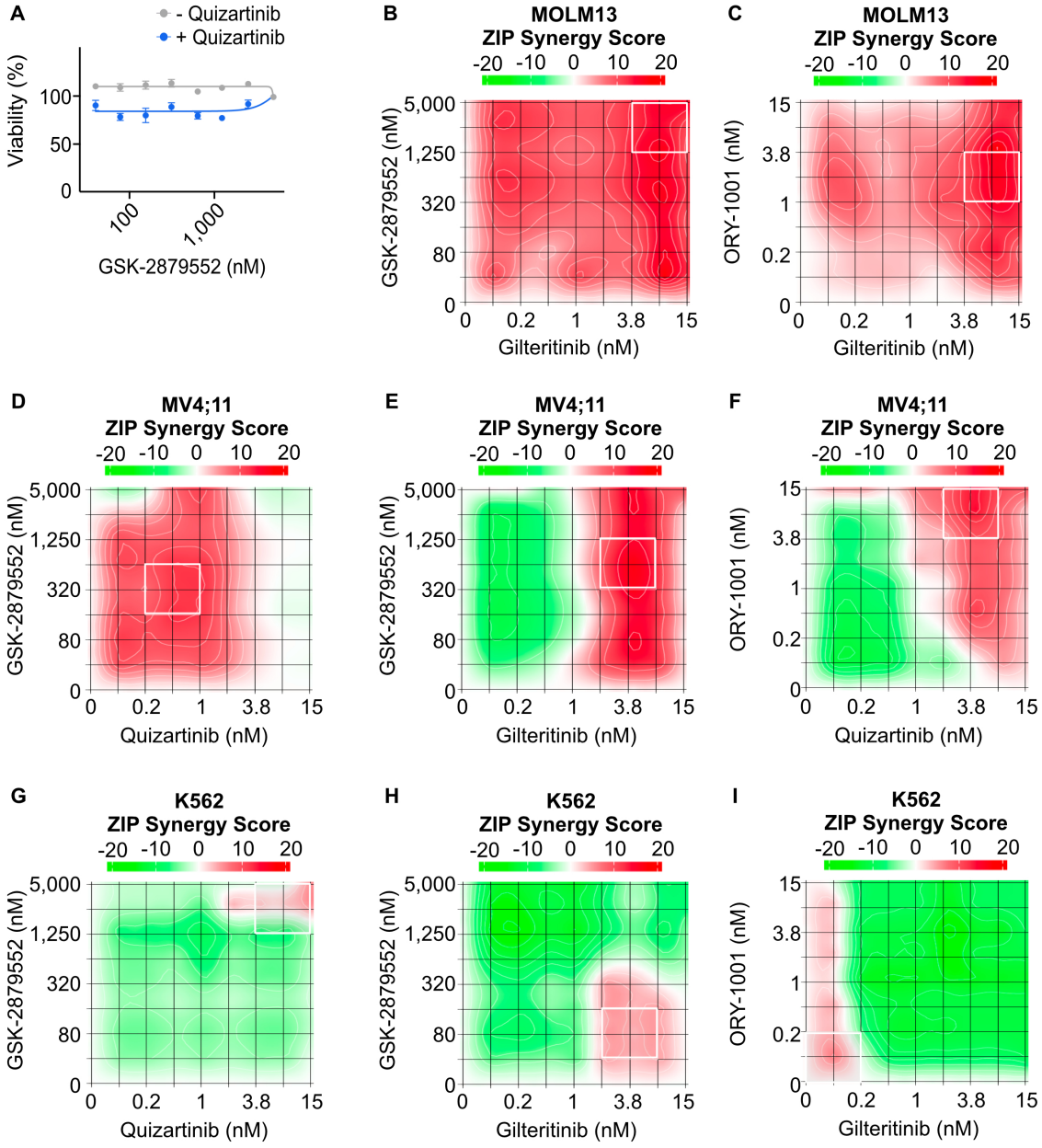

Figure S4.1: Drug synergy between FLT3 and LSD1 inhibitors in FLT3-ITD and FLT3-wildtype cell lines.

A, Dose response curves for GSK-2879552 with and without quizartinib (1 nM, which is the concentration corresponding to maximal synergy in Figure 4.1). **B,** MOLM13 (FLT3-ITD-positive) cells were treated with an 8x8 matrix of gilteritinib and GSK-2879552 or **(C**) gilteritinib and ORY-1001 in triplicate. **D,** MV4;11 (FLT3-ITD-positive) cells were treated with an 8x8 matrix of quizartinib and GSK-2879552, **(E**) gilteritinib and GSK-2879552, or (**F)** quizartinib and ORY-1001 in triplicate. **G,** K562 (FLT3-wildtype) cells were treated with an 8x8 matrix of quizartinib and GSK-2879552, (**H**) gilteritinib and GSK-2879552, or (**I**) gilteritinib and ORY-1001 in triplicate. Cell viability was assessed after 72 hours of culture using CellTiter Aqueous colorimetric assay. Synergy was assessed using the ZIP method.

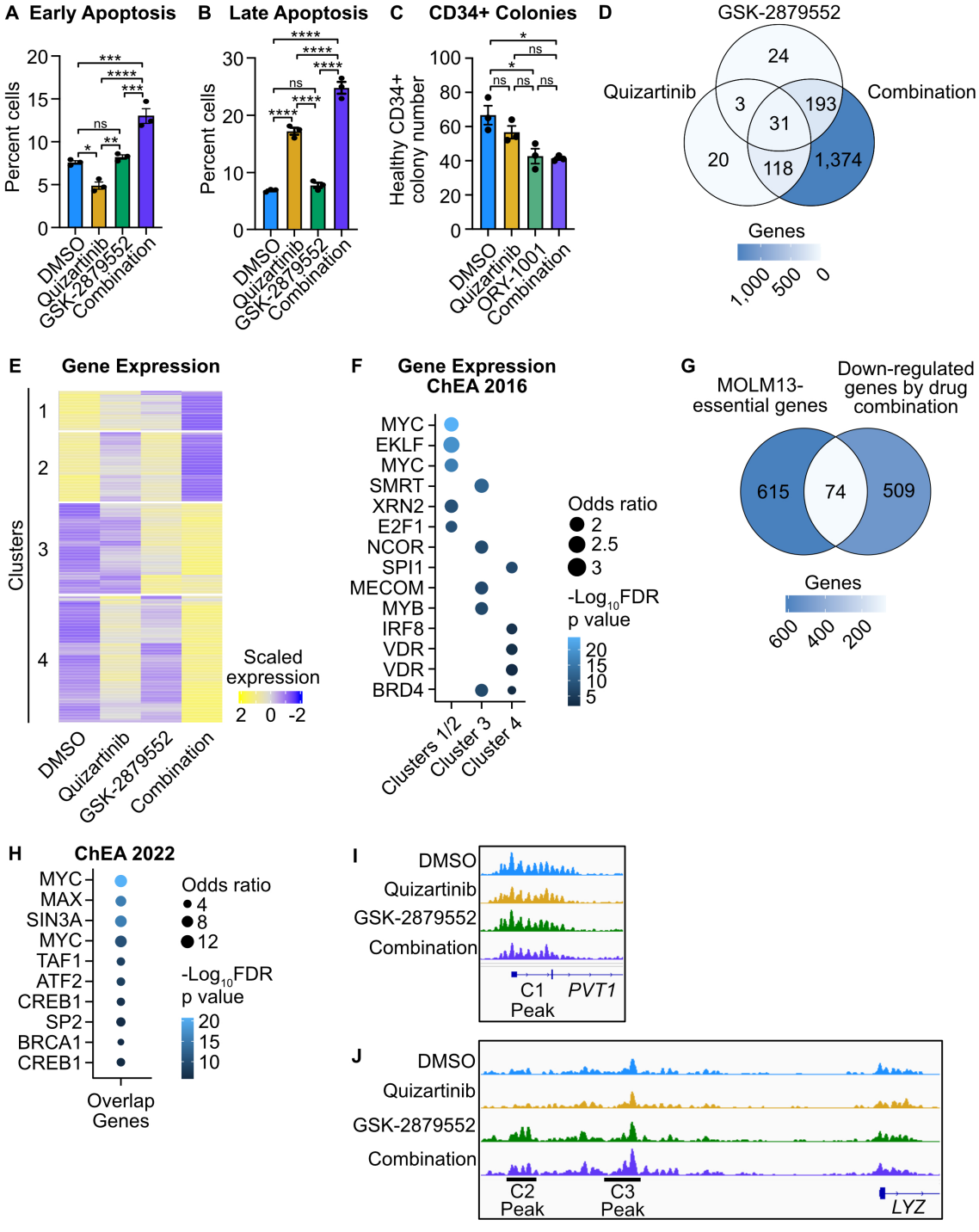

Figure S4.2: Efficacy of dual FLT3/LSD1 inhibition in MOLM13 cells.

A, B, MOLM13 cells were treated quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO vehicle for 48 hours prior to assessment of apoptosis by flow cytometry. Early apoptosis is Annexin V positive and propidium iodide negative. Late apoptosis is Annexin V positive and propidium iodide positive. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **C,** CD34+ cells from healthy donor marrow were plated in complete human methocult media along with quizartinib (1 nM), GSK-2879552 (100 nM) the combination, or an equal volume of DMSO. Colony number was assessed after 14 days of growth. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **D,** Overlap of differentially expressed genes following single-agent or dual-agent therapy. **E,** Unsupervised hierarchical clustering of differentially expressed genes and **F,** complete transcription factor target enrichment from RNA-seq data in Figure 4.1. **G,** Overlap of differentially down-regulated genes by the drug combination and MOLM13-essential genes identified in a genome-wide CRISPR/Cas9 screen. **H,** Transcription factor target enrichment of shared genes in (**G**). **I, J,** Example tracks from H3K27ac CUT&Tag described in Figure 4.1. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

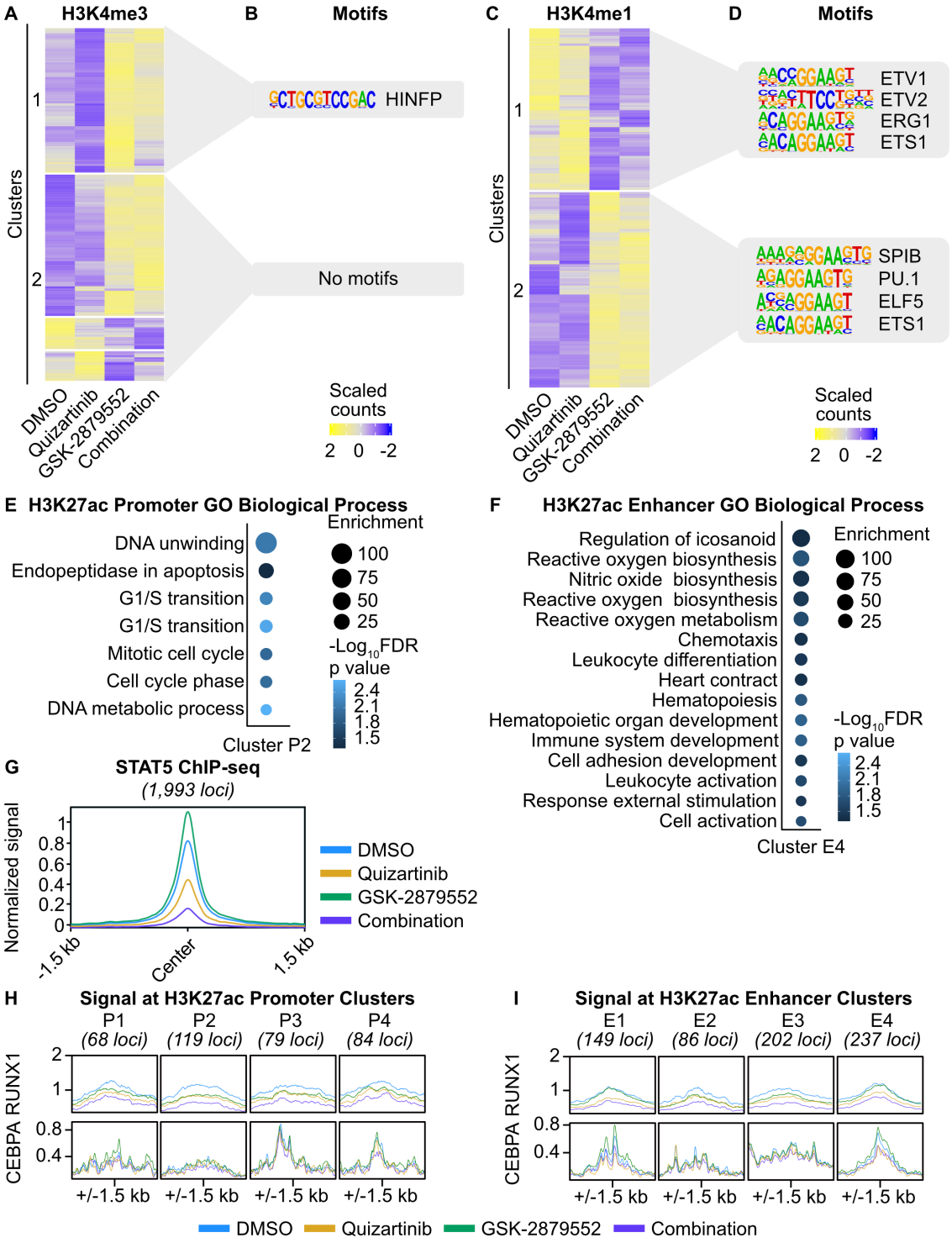

Figure S4.3: Epigenetic impact of dual FLT3/LSD1 inhibition.

A, Unsupervised hierarchical clustering of regions with differential H3K4me3 signal at promoters assessed by CUT&Tag from MOLM13 cells treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 6 hours. **B,** Motif enrichment of promoters with differential H3K27ac signal. Top four *de novo* motifs with p-value $<10^{-12}$ are shown. **C,** Unsupervised hierarchical clustering of regions with differential H3K4me1 signal at enhancers assessed by CUT&Tag from MOLM13 cells treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 6 hours. **D,** Motif enrichment of regions with differential H3K4me1 signal at enhancers. Top 4 enriched transcription factor motifs are shown. **E, F,** GO analysis of differential H3K27ac signal at (**E**) promoters and (**F**) enhancers from analysis in Figure 4.2. Dot size represents the binomial fold enrichment and color represents the $-\log_{10}$(FDR p-value) of the GO Biological Process term. **G,** STAT5 binding profile at consensus peaks from MOLM13 ChIP-seq presented in Figure 4.2. **H,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 6 hours. RUNX1 binding was assessed by ChIP-seq and CEBPA binding was assessed by CUT&Tag. Transcription factor binding profiles at promoters with differential H3K27ac signal identified in Figure 4.2. **I,** Transcription factor profiles at enhancers with differential H3K27ac signal identified in Figure 4.2.

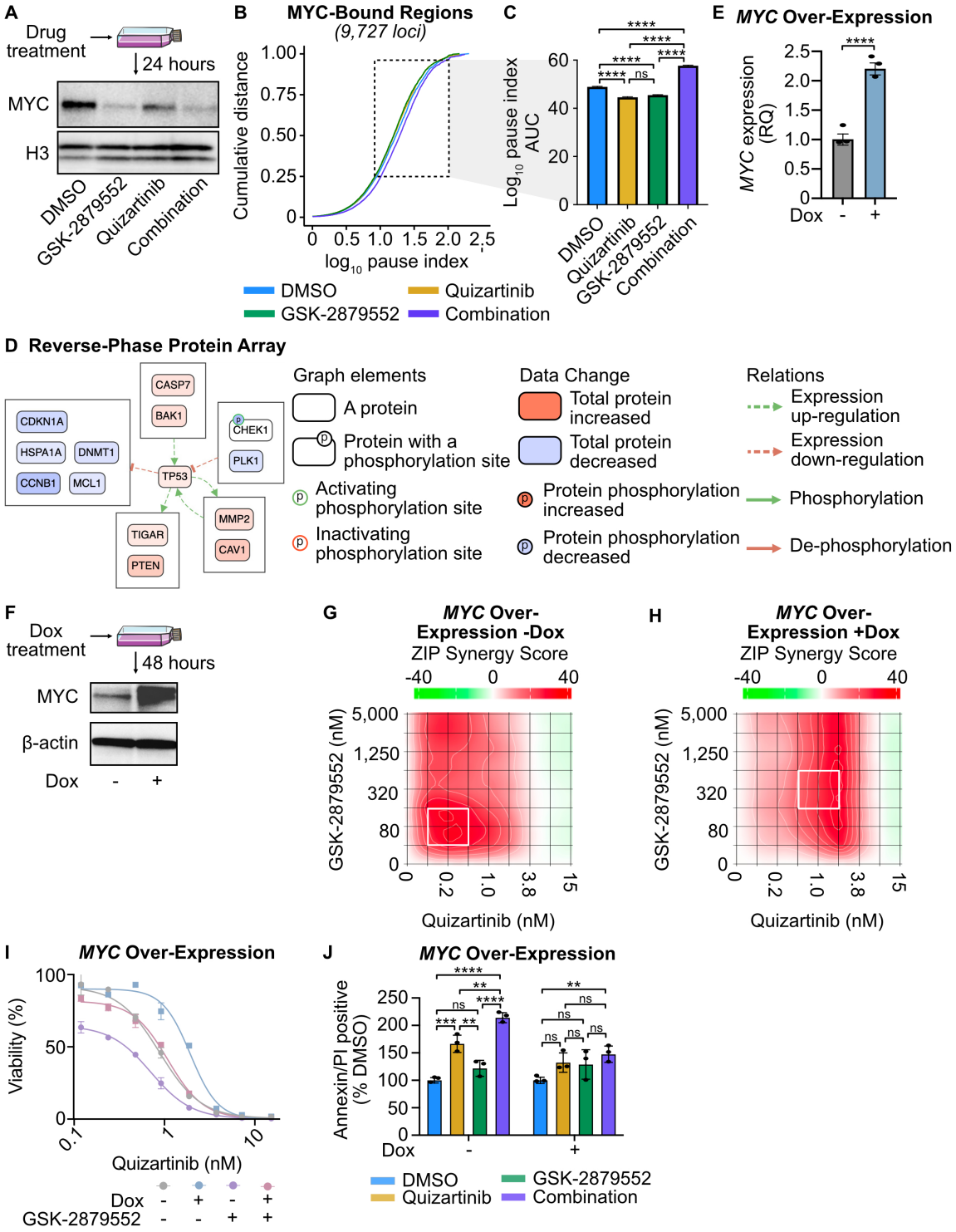

Figure S4.4: Dual FLT3/LSD1 inhibition suppresses *MYC* expression and activity.

A, Western blot for MYC in MOLM13 cells treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equivalent volume of DMSO for 24 hours. **B,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO vehicle for 6 hours prior to CUT&Tag for RBP1. Cumulative distribution of RNA PolII pause-indices at RBP1 and MYC co-bound regions. **C,** AUC of RNA PolII pause-indices within the 25-95% cumulative distribution. Statistical significance was determined by Kolmogorov-Smirnov test with an FDR post-test correction. **D,** Differential phosphoprotein network enrichment in MOLM13 cells treated for 24 hours with 1 nM quizartinib and 100 nM GS-K2979552 or an equivalent volume of DMSO. **E,** MOLM13 cells were transduced with lentiviral particles harboring a doxycycline-inducible *MYC* expression vector. *MYC* expression was assessed 48 hours after the addition of doxycycline (1 µg/mL) and normalized to *GAPDH* as an endogenous control. Statistical significance was determined by Student's t-test. **F,** Western blot for MYC and β-actin in MOLM13 cells harboring a doxycycline-inducible *MYC* expression vector following treatment with doxycycline (1 µg/mL) or DMSO for 48 hours. **G, H,** Cells were treated doxycycline (1 µg/mL) or an equivalent volume of DMSO for 48 hours then plated in triplicate in an 8x8 matrix of concentrations of quizartinib and GSK-2879552 for 72 hours. Cell viability was measured using CellTiter Aqueous colorimetric assay. Synergy was assessed using the ZIP method. **I,** Quizartinib response curves in MOLM13 cells harboring a doxycycline-inducible *MYC* expression construct treated with and without doxycycline (1 µg/mL) and/or GSK-2979552 (311 nM, which is the concentration corresponding to maximal synergy in the *MYC* over-expressed MOLM13 cells in (**G**)). **J,** Cells were treated with doxycycline (1 µg/mL) for 48 hours then with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 48 hours. Apoptosis was assessed using flow cytometry for Annexin V and PI. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

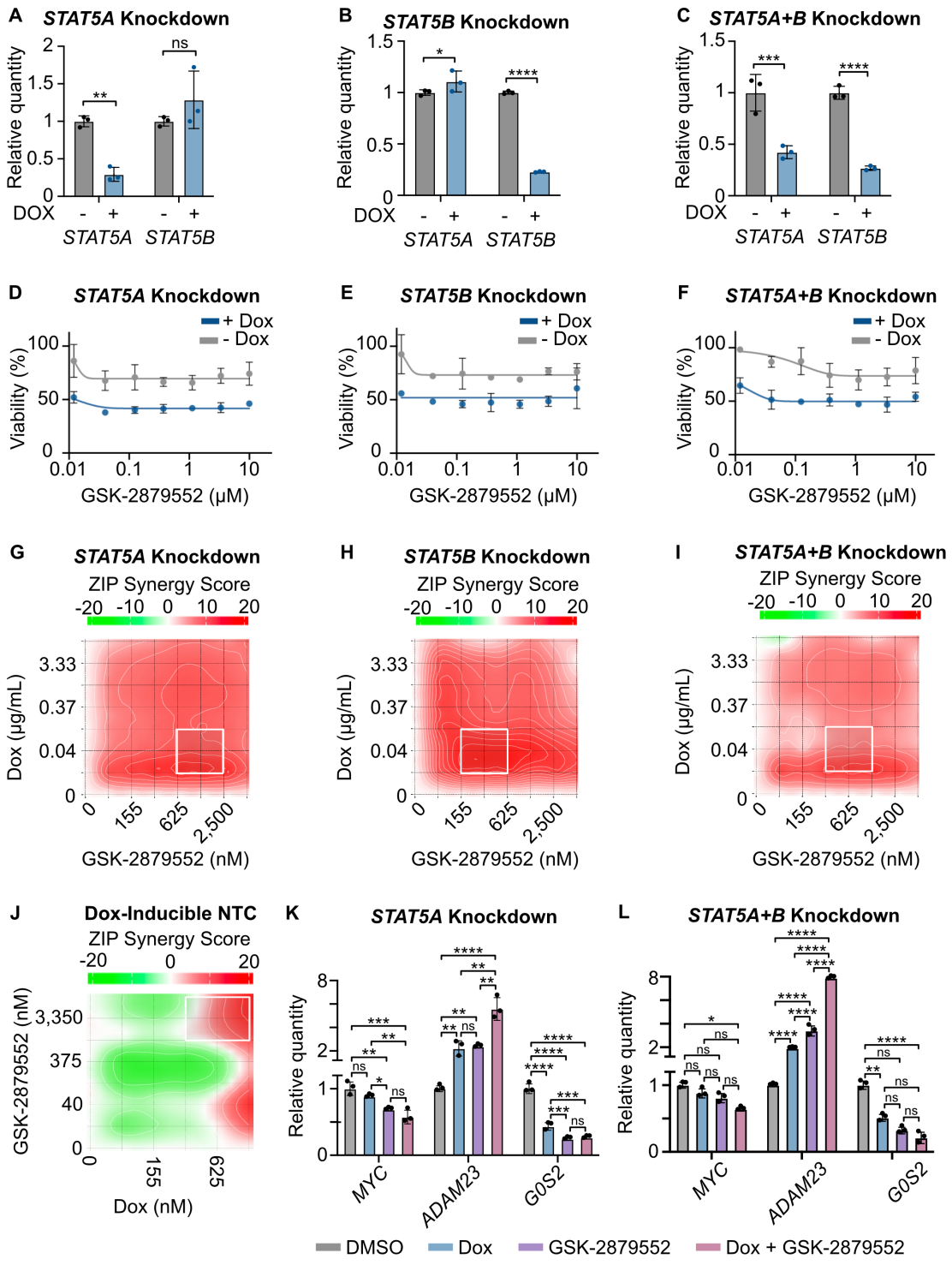

Figure S4.5: STAT5 and MYC play a key role in the response to FLT3/LSD1 inhibition.

A-C, MOLM13 cells were transduced with lentiviral particles harboring a doxycycline-inducible *STAT5* knockdown. Cells were induced with doxycycline (1 µg/mL) for 48 hours. Validation of knockdown of *STAT5A* and *STA5B* by qPCR relative to *GUSB* endogenous control. Significance was determined by Student's t-test. **D-F,** Dose response curves for GSK-2879552 with and without doxycycline (1 µg/mL). **G, H,** 8x8 drug synergy matrices between GSK-2879552 and (**G**) *STAT5A*, (**H**) *STAT5B*, or (**I**) *STAT5A* and *STAT5B* knockdown (induced by doxycycline) were performed in triplicate in MOLM13 cells with viability assessed by CellTiter Aqueous colorimetric assay after 72 hours of drug exposure. ZIP synergy scores were calculated on the average values for each drug dose. **J,** MOLM13 cells harboring a doxycycline-inducible non-targeting control (NTC) short hairpin RNA (shRNA) construct were treated doxycycline (1 µg/mL) or an equivalent volume of DMSO for 48 hours then plated in triplicate in an 8x8 matrix of concentrations of GSK-2879552 and doxycycline for 72 hours. Cell viability was measured using CellTiter Aqueous colorimetric assay. Synergy was assessed using the ZIP method. **K, L,** qPCR assessment of gene expression in MOLM13 cells expressing a doxycycline-inducible (**K**) *STAT5A* or (**L**) *STAT5A+B* shRNA. Cells were treated with doxycycline (1 µg/mL) or an equivalent volume of DMSO for 48 hours prior to the addition of GSK-2879552 (100 nM) for 24 hours. Expression was normalized to *GUSB* as an endogenous control. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

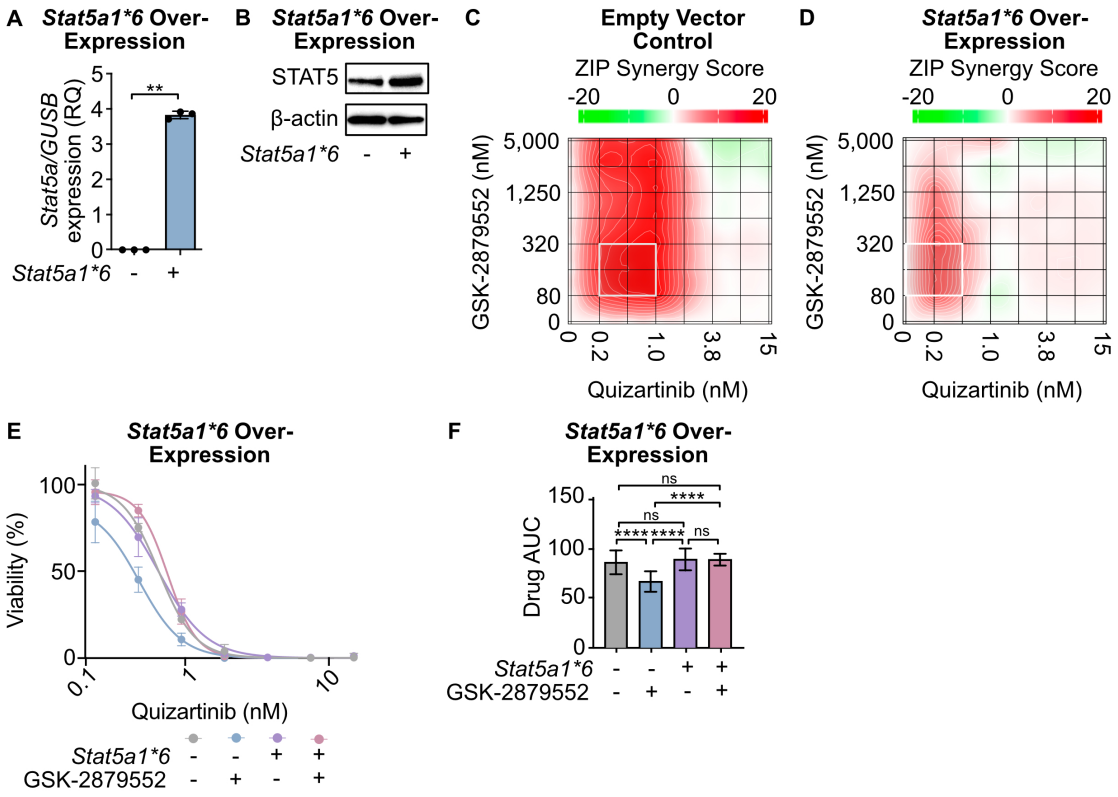

Figure S4.6: *Stat5* over-expression diminishes the efficacy of combined FLT3/LSD1 inhibition.

A, MOLM13 cells were transduced with retroviral particles harboring a *Stat5a1*6* expression construct. Validation of *Stat5a* over-expression by qPCR normalized to *GUSB* expression. Significance was determined by a two-tailed Wilcoxon test. **B,** Western blot for STAT5 and β-actin in MOLM13 cells harboring an empty or a *Stat5a1*6* expression vector. **C, D,** Cells were plated in triplicate in an 8x8 matrix of concentrations of quizartinib and GSK-2879552 for 72 hours. Cell viability was measured using CellTiter Aqueous colorimetric assay. Synergy was assessed using the ZIP method. **E,** Quizartinib response curves in cells treated with and without GSK-2979552 (168 nM, which is the concentration corresponding to maximal synergy in the *Stat5a1*6* over-expressed MOLM13 cells in (**D**)). **F,** AUC data from the 168 nM GSK-2979552 isoline. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

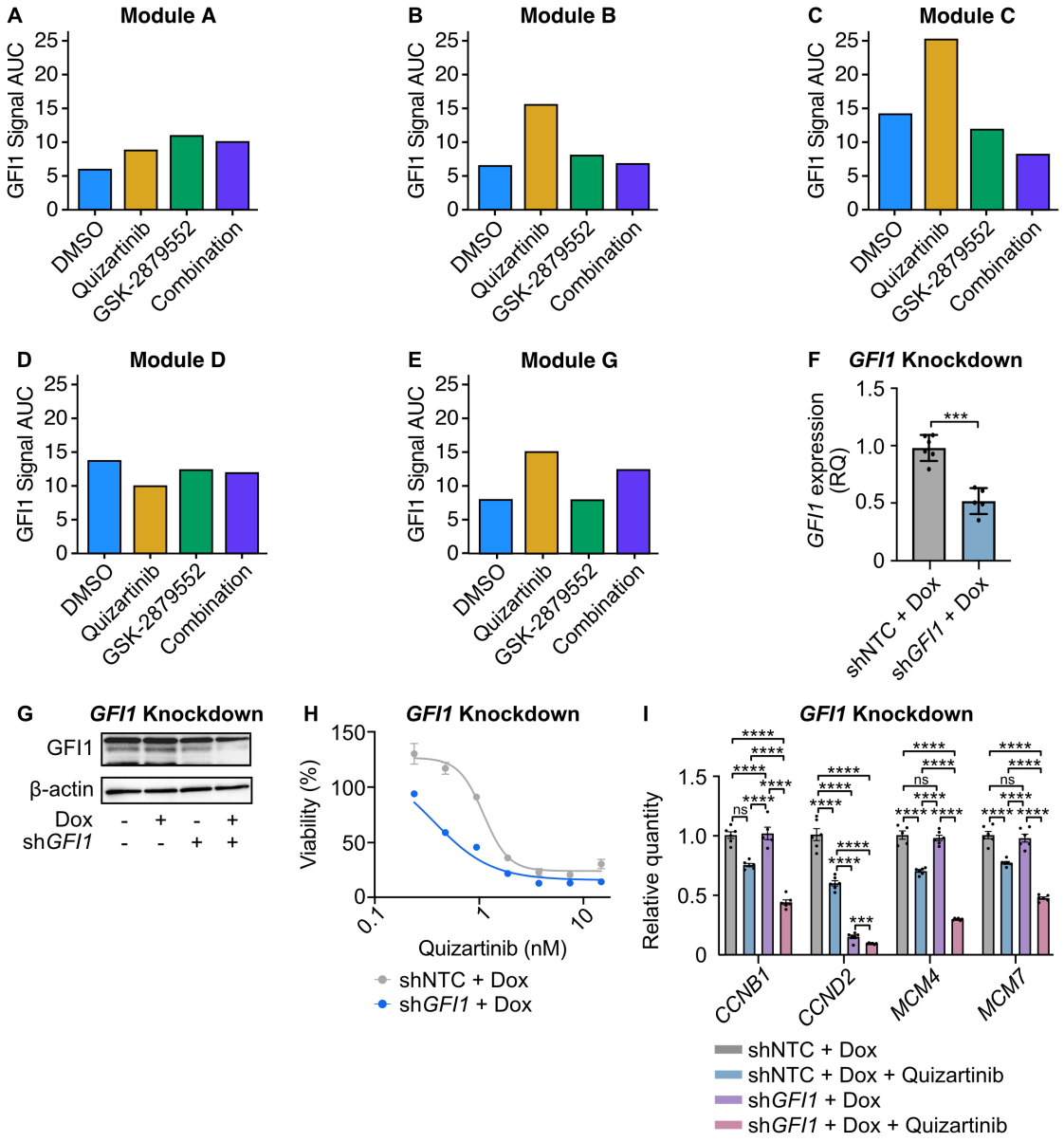

Figure S4.7: *GFI1* knockdown weakens the effects of dual FLT3/LSD1 inhibition.

A-E, AUC of GFI1 CUT&RUN signal from Figure 4.2 at five BENC modules. **F,** Validation of *GFI1* knockdown by qPCR relative to *GUSB* endogenous control. Statistical significance was determined by Student's t-test. **G,** MOLM13 cells harboring doxycycline-inducible NTC or *GFI1* shRNA were treated with doxycycline (1 μg/mL) for 48 hours and then probed for GFI1 and β-actin. **H,** Dose response curves for MOLM13 cells harboring a doxycycline-inducible NTC or *GFI1* shRNA with doxycycline (1 μg/mL). **I,** qPCR assessment of gene expression in MOLM13 cells expressing a doxycycline-inducible *GFI1* shRNA. Cells were treated with doxycycline (1 μg/mL) for 48 hours prior to the addition of quizartinib (1 nM) for 24 hours. Substantial knockdown was observed in the absence of doxycycline treatment, so only doxycycline-treated samples were compared. Expression was normalized to *GUSB* as an endogenous control. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

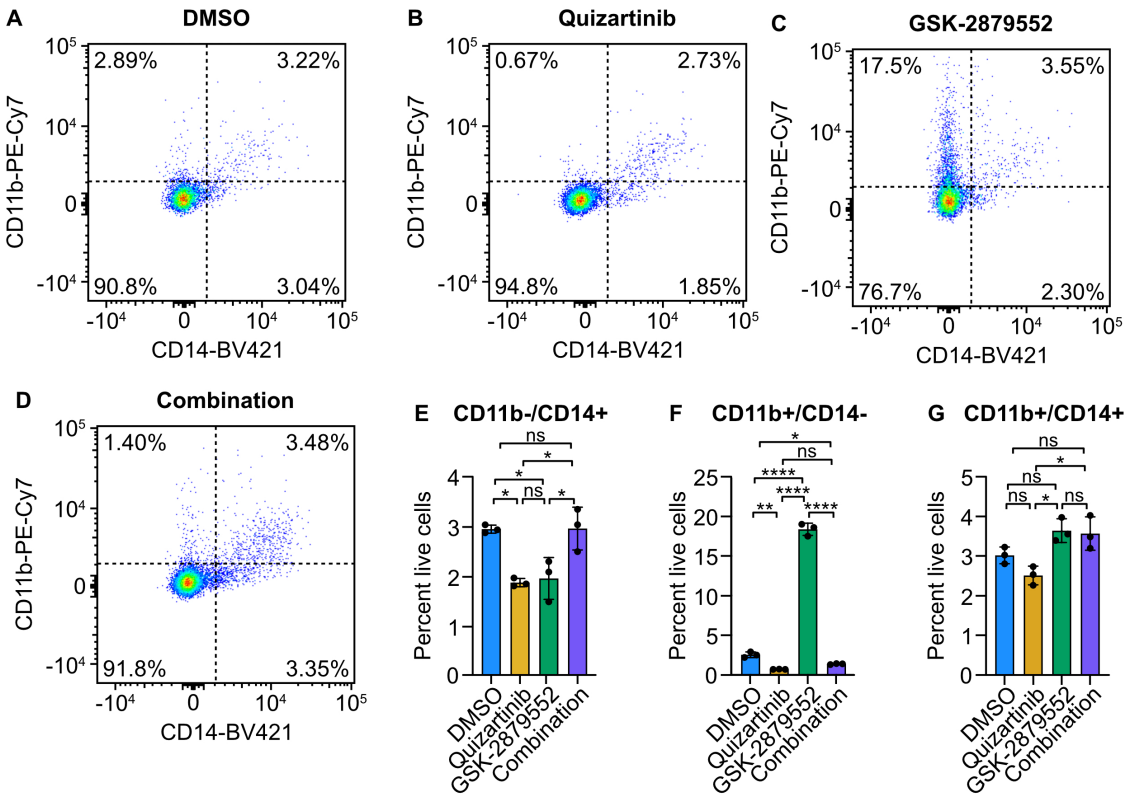

Figure S4.8: Combined FLT3/LSD1 inhibition does not substantially alter myeloid differentiation.

A-D, Flow cytometry assessment of CD11b and CD14 surface markers in MOLM13 cells treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 72 hours. Representative histograms of surface marker signal with the dotted lines indicating positive signal cut-off. **E-G,** Percent of cells expressing cell surface markers as defined by the gates in (**A-D**). Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

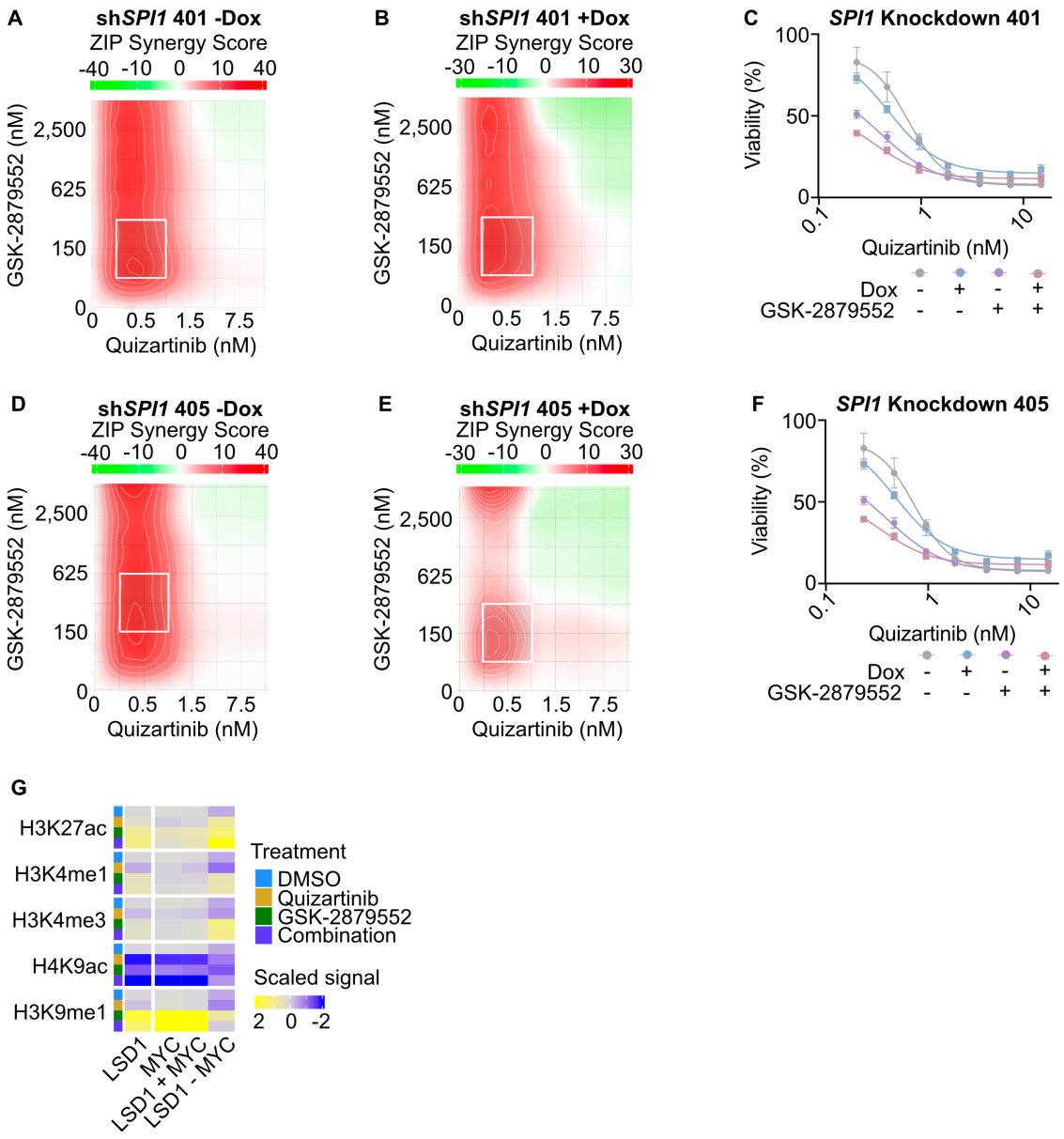

Figure S4.9: *SPI1* knockdown does not disrupt dual FLT3/LSD1 inhibition synergy.

A, B, MOLM13 cells were transduced with retroviral particles harboring a doxycycline-inducible *SPI1* vector (shRNA construct 401). Cells were plated in triplicate in an 8x8 matrix of concentrations of quizartinib and GSK-2879552 for 72 hours. Cell viability was measured using CellTiter Aqueous colorimetric assay. **C,** Dose response curves in MOLM13 cells harboring a doxycycline-inducible *SPI1* shRNA construct 401 with quizartinib. Cells were treated with and without doxycycline (1 µg/mL) and/or GSK-2979552 (311 nM, which is the concentration corresponding to maximal synergy in the *SPI1* knockdown MOLM13 cells following doxycycline in (**B**)). **D, E,** Same as (**A**) and (**B**) except with the *SPI1* shRNA construct 405. **F,** Same as (**C**) except with the *SPI1* shRNA construct 405. **G,** MOLM13 cells were treated with quizartinib (1 nM), GSK-2879552 (100 nM), the combination, or an equal volume of DMSO for 6 hours prior to CUT&Tag for H3K9ac or H3K9me1. Normalized histone modification CUT&Tag signal at LSD1-bound regions, MYC-bound regions, LSD1 and MYC co-bound regions, and LSD1-bound regions not bound by MYC. H3K27ac, H3K4me1, and H3K4me3 CUT&Tag signal and LSD1 and MYC ChIP-seq peaks from Figure 4.2 were examined. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

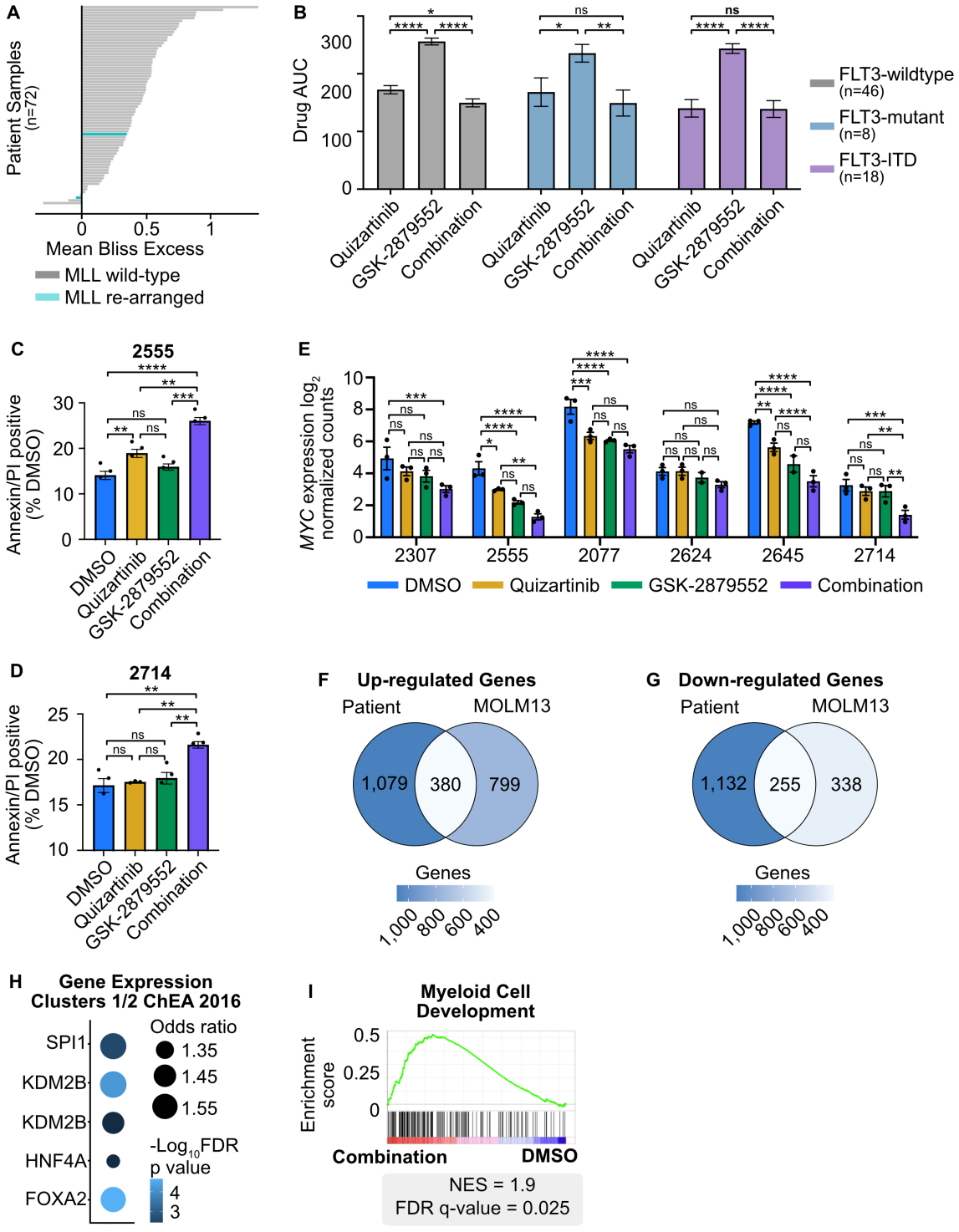

Figure S4.10: Efficacy of dual FLT3/LSD1 inhibition in primary AML samples.

A, Primary AML blasts from 72 total samples (18 FLT3-ITD-positive) were cultured for 72 hours along a 7-point curve with either quizartinib, GSK-2879552, or equimolar amounts of the drug combination. Cell viability was assessed by CellTiter Aqueous colorimetric assay. Excess over Bliss was calculated using cell viability at corresponding drug concentrations. Each bar represents the mean excess over Bliss across all concentrations. Bar color indicates MLL re-arrangement status. **B,** Mean drug AUC in patient samples grouped by FLT3 mutation status. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **C, D,** Primary AML blasts were treated with quizartinib (500 nM), GSK-2879552 (500 nM), the combination, or an equivalent volume of DMSO for 24 hours. Apoptosis was assessed by flow cytometry for Annexin V and PI after 48 hours of drug exposure. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **E,** RNA-seq was performed on six independent, FLT3-ITD-positive patient samples treated in triplicate with 500 nM quizartinib, 500 nM GSK-2879552, both drugs in combination, or an equivalent volume of DMSO for 24 hours. *MYC* gene expression in each individual sample is shown. Statistical significance was determined by two-way ANOVA with a Holm-Šidák post-test correction. **F, G,** Overlap of differentially up-regulated and down-regulated genes following drug combination treatment in primary AML blasts from (**E**) and in MOLM13 cells from Figure S4.2. **H,** Transcription factor target enrichment from clusters in Figure 4.6G. **I,** Select GSEA was performed comparing the drug combination to DMSO. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

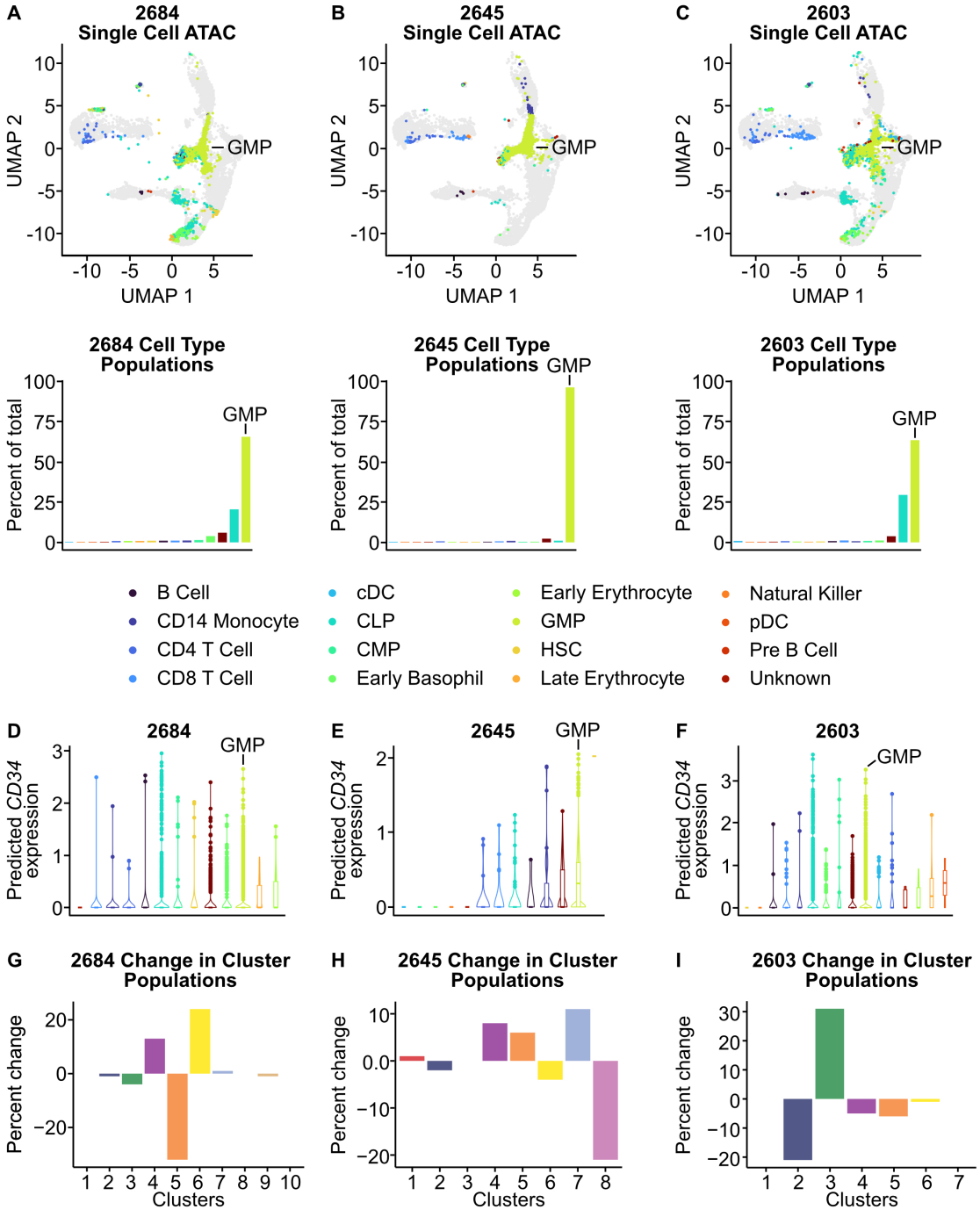

Figure S4.11: Single-cell ATAC-seq cell population dynamics in primary patient samples following treatment with dual FLT3/LSD1 inhibition.

A-C, Primary AML blasts from three FLT3-ITD-positive samples were treated with quizartinib (500 nM) and GSK-2879552 (500 nM) or an equal volume of DMSO vehicle for 24 hours prior to single-cell ATAC sequencing. Cells were mapped to a latent UMAP space created from a reference healthy hematopoiesis dataset. Cells were assigned cell type labels according to their mapping to the reference UMAP**. D-F,** Predicted *CD34* expression calculated from chromatin accessibility for each cell type population. **G-I,** Percent change in cluster populations defined in Figure 4.7 between DMSO and combination treatments.

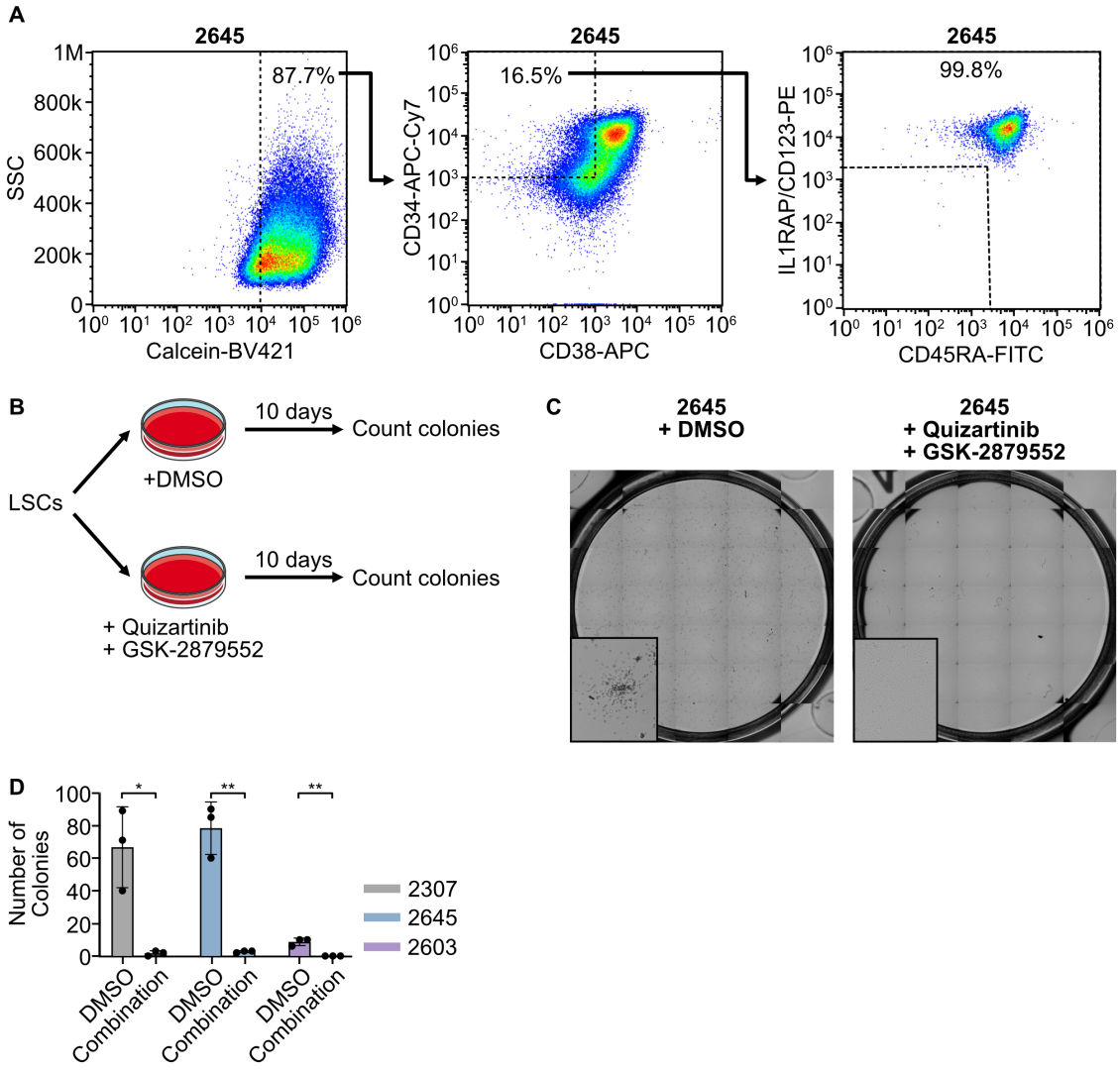

Figure S4.12: Dual FLT3/LSD1 inhibition disrupts LSC clonogenicity.

A, FACS strategy to isolate LSCs from CD34-enriched primary patient samples. Cells were sorted based on fluorescent cell surface markers and plated in triplicate into cytokine-enriched methylcellulose media. Representative histograms of surface marker signal with the dotted lines indicating positive signal cut-off. **B,** Schema of LSC sorting into colony assays. **C,** Representative images of colony assays at day 10. **D,** Quantification of the colony phenotype shown in (**C**). Statistical significance was determined by Student's t-test. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

Chapter 5: Differentiation state plasticity as a mechanism of BCL2 inhibitor resistance in acute myeloid leukemia

---

- BLC2 inhibition promotes the production of differentiated $CD34^{-}CD38^{+}$ AML cells.
- BCL2 inhibition was not associated with increased apoptosis of immature $CD34^{+}CD38^{-}$ cells, suggesting that the outgrowth of differentiated $CD34^{-}CD38^{+}$ cells was likely not due to selective killing of immature cells.
- Single-cell chromatin accessibility sequencing revealed that short-term BCL2 inhibition primes immature AML cells for myeloid differentiation.
- Long-term BCL2 inhibition in a PDX model of FLT3-ITD/NPM1-mutant AML resulted in a depletion of HSC-like cells and an outgrowth of progenitor-like cells.

An abstract of this work has been published in *Blood*:

Yashar WM, Pacentine IV, Tsuchiya M, Macaraeg J, Taherinasab A, Evans-Dutson S, O'Connell B, Lusardi T, Szczepanski N, Yardimci GG, Adey AC, Maxson JE, Braun TP. Differentiation State Plasticity as a Mechanism of BCL2 Inhibitor Resistance in Acute Myeloid Leukemia. *Blood*, 2023 November 28[302]

5.1 Introduction

Acute myeloid leukemia (AML) is composed of a heterogeneous population exhibiting characteristics of normal myeloid cells, including immature, HSC-like cells and differentiated, monocyte-like cells[84]. The balance of immature and differentiated AML cells within a patient is important clinically. Immature cells are acutely sensitive to treatment with venetoclax, a BH3 mimetic that blocks the anti-apoptotic activity of BCL2[163]. BCL2-inhibitor-based therapeutic strategies are the standard of care among patients with untreated or relapsed/refractory AML unfit for chemotherapy. While these strategies result in a remission rate of nearly 70%, most patients who do achieve remission ultimately relapse[159]. Frequently, the relapsed disease is composed of differentiated, monocyte-like leukemic cells[163]. Given the importance of BCL2 inhibitors to the clinical management of AML, it is critical to understand the mechanisms by which BCL2 inhibitor treatment results in monocytic relapse.

Disease relapse has been traditionally thought to emerge from intrinsically resistant cells that confer a survival advantage under therapeutic pressure[157]. Several studies have suggested that differentiated AML cells have a survival advantage in the setting of BCL2 inhibition due to a reliance on anti-apoptotic proteins other than BCL2[163,164]. Immature AML cells express high levels of *BCL2*, whereas cells primed for lympho-myeloid or granulocytic-monocytic progenitor cell fates express high levels of *MCL-1* and *BCL-XL*, respectively. However, these studies do not explain why patients with dominant, differentiated subclones frequently achieve long-lasting remission with BCL2-inhibitor-based strategies[163]. Therefore, the intrinsic resistance of differentiated AML cells appears to be an insufficient model to describe BCL2 inhibitor resistance. An alternative model of therapeutic resistance is that cancer cells transcriptionally adapt to a resistant phenotype and acquire a survival advantage under therapeutic pressure[157]. In

the context of BCL2 inhibitor resistance, the extent to which the emergence of differentiated cells results from selective pressure or adaptive changes to the cell differentiation state is unclear.

In our work, we have investigated the response to BCL2 inhibition in AML. We found that BLC2 inhibition resulted in a dose-dependent decrease in immature AML cells and a simultaneous increase in differentiated AML cells. Notably, this reduction of immature cells was not due to the induction of apoptosis, suggesting that BCL2 inhibition did not selectively kill immature AML cells. Single-cell chromatin accessibility revealed that short-term BCL2 inhibition primed leukemic cells for a shift in their differentiation state. Following long-term BCL2 inhibition in an AML patient-derived xenograft (PDX) model, we observed a depletion of hematopoietic stem cell (HSC)-like AML cells and an emergence of hematopoietic progenitor-like AML cells. These findings support the role of differentiation state plasticity in BLC2 inhibitor resistance.

5.2 Results

5.2.1 A leukemia stem cell model stably produces immature and differentiated cells

Investigating the association between venetoclax sensitivity and AML cell differentiation state is difficult since functional studies in primary AML cells are technically challenging. Moreover, most immortalized human AML cell lines are homogenous and do not recapitulate the hierarchical organization of hematopoiesis. To model the functional hematopoietic hierarchy *in vitro*, we sought an indefinitely growing AML culture system (OCI-AML8227) that produces clonogenic progenitors and terminally differentiated cells[303]. To characterize the differentiation state heterogeneity of

OCI-AML8227 cells, we investigated the frequency cell surface markers associated with differentiation, including CD34 (hematopoietic stem and progenitor cells), CD38 (committed hematopoietic progenitors but not stem cells), and CD14 (monocyte/macrophage). We found a small population of immature $CD34^{+}CD38^{-}$ cells and a larger population of differentiated $CD34^{-}CD38^{+}$ and $CD34^{-}CD38^{-}$ cells (Figure 5.1a). To understand whether the immature AML cells produced the differentiated cells, we immunomagnetically fractioned the OCI-AML8227 cells by CD34 expression. We then monitored the surface marker expression over seven days. We found that the relative fraction of immature $CD34^{+}CD38^{-}$ cells significantly declined over this period (Figure 5.1b-d). Moreover, we observed that the fraction of $CD34^{+}CD38^{-}CD14^{+}$ cells increased over this same period. These findings suggest that immature $CD34^{+}$ OCI-AML8227 cells produce a stable population of differentiated cells.

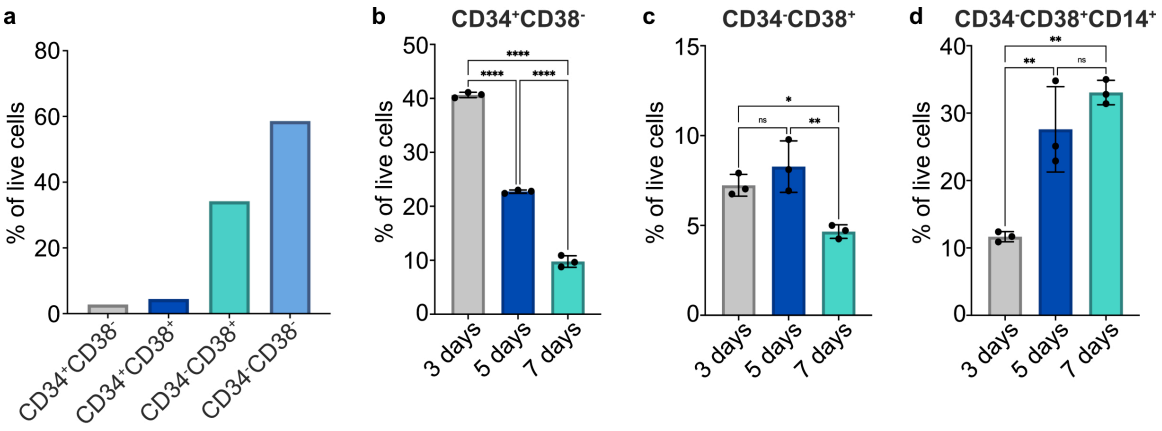

Figure 5.1: Immature $CD34^{+}$ OCI-AML8227 cells maintain a heterogenous population of immature and differentiated AML cells.

a. OCI-AML8227 cells were cultured for three days before assessment of cell surface markers by flow cytometry. **b-d.** OCI-AML8227 cells were immunomagnetically fractioned by CD34 surface expression. These cells were subsequently cultured for 3, 5, and 7 days before flow cytometry. P-values were calculated by one-way ANOVA followed by Holm–Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

5.2.2 BCL2 inhibition promotes the production of differentiated AML cells

Previous studies have shown that differentiation state is an key determinant of venetoclax sensitivity[163]. To understand whether the OCI-AML8227 system was appropriate to investigate this relationship, we treated OCI-AML8227 cells with venetoclax after immunomagnetic fractionation by CD34 expression. Consistent with previous studies in AML patient samples, we found that immature $CD34^+$ AML cells were significantly more sensitive to venetoclax than differentiated $CD34^-$ cells (Figure 5.2a, b). These findings indicate that OCI-AML8227s are an appropriate model to study the relationship between differentiation state and venetoclax sensitivity.

Patients with AML who relapse on venetoclax treatment often present with a differentiated, monocyte-like disease[163]. To understand how venetoclax impacts AML differentiation state, we studied the cell surface marker expression in OCI-AML8227 cells exposed to venetoclax for 72 hours. We observed a significant dose-dependent decrease in the number of live $CD34^+CD38^-$ cells (Figure 5.2c). Conversely, we observed a significant increase in the number of differentiated $CD34^-CD38^+$ cells with venetoclax treatment (Figure 5.2d). While venetoclax increased the number of apoptotic cells, no substantial differences were observed between mature and immature fractions (Figure 5.2e, f). These findings indicate that BCL2 inhibition is promoting the differentiation of AML cells.

5.2.3 Short-term BCL2 inhibition primes the induction of differentiation in AML cells

To understand the transcriptional dynamics following exposure to venetoclax, we evaluated genome-wide single-cell chromatin accessibility in OCI-AML8227 cells 24

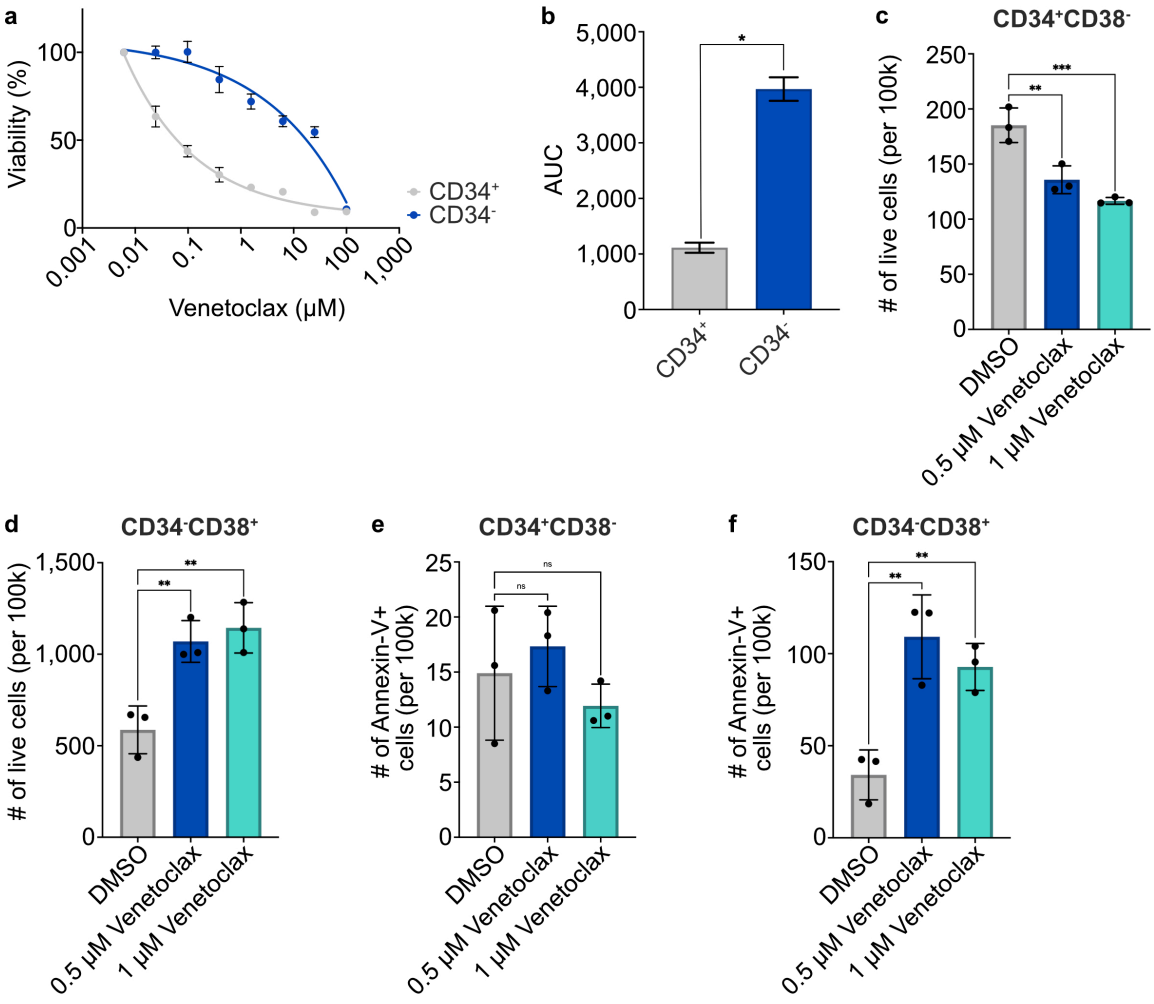

Figure 5.2: BCL2 inhibition results in the outgrowth of differentiated AML cells.

a. OCI-AML8227 cells were immunomagnetically fractioned by CD34 surface expression and then cultured for three days along a 7-point curve in venetoclax. **b.** Area under the curve (AUC) of drug curves in (a). **c-f.** Unfractionated OCI-AML8227 cells were cultured for three days in DMSO or venetoclax before assessment of cell surface markers and Annexin-V by flow cytometry. P-values were calculated by (b) Student's t- test and (c-f) one-way ANOVA followed by Holm–Šidák post-test correction. ns = not significant, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

hours after drug treatment. We found that cells from both conditions had chromatin accessibility profiles resembling normal myeloid cells (Figure 5.3a, b). We observed that the CMP LMPP-like cells associated with venetoclax treatment resided farther up the

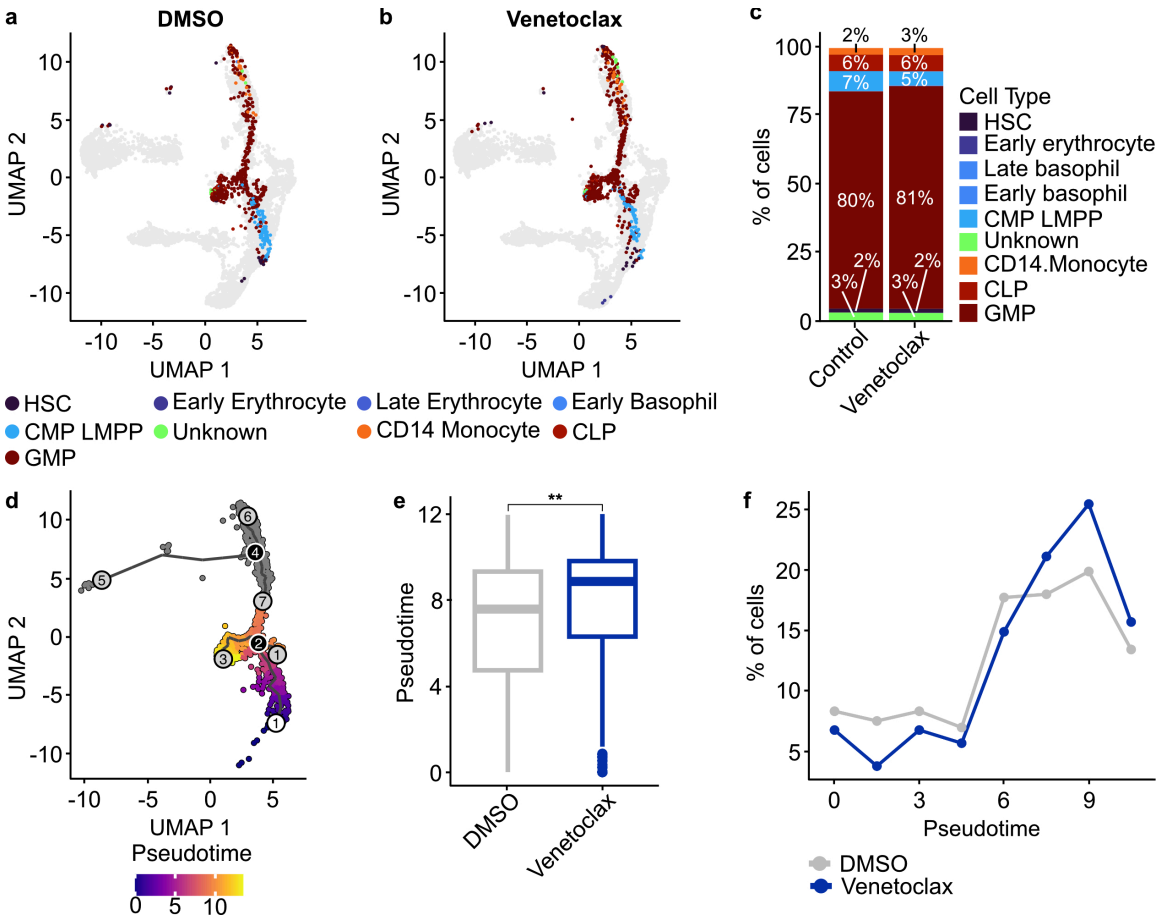

Figure 5.3: Short-term BLC2 inhibition promotes the progression of immature AML cells into the myeloid lineage.

a-c. OCI-AML8227 cells were cultured for 24 hours in DMSO or venetoclax before single-cell ATAC sequencing (scATAC-seq). Cells were mapped to a latent UMAP space created from a reference healthy hematopoiesis dataset (grey). Cells were assigned cell type labels according to their mapping to the reference latent space. **d, e.** Pseudotime was inferred using k=10 from DMSO- and venetoclax-treated cells. *P*-value was calculated using Wilcoxon t-test. **f.** Pseudotime was binned into intervals of 1.5. ns = not significant, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001.

myeloid lineage than the DMSO-treated CMP LMPP-like cells. Despite this, short-term venetoclax therapy only resulted in a minor increase in the number of AML cells resembling GMPs and CD14 monocytes (Figure 5.3c). To further characterize the effects of venetoclax on differentiation, we performed a pseudotime trajectory analysis,

where increasing pseudotime values are associated with more differentiation (Figure 5.3d). Using this approach, we found that venetoclax-treated cells were associated with significantly greater pseudotime values than the DMSO-treated cells (Figure 5.3e, f). These findings show that differentiation increases rapidly after exposure to venetoclax.

5.2.4 Long-term BCL2 inhibition drives a differentiation shift in a mouse PDX model of AML

Our findings suggest that short-term inhibition of BLC2 primes immature AML cells for myeloid differentiation. We wanted to understand how long-term BLC2 inhibition shifted the AML differentiation state in a clinically relevant model. We evaluated the effects of long-term BLC2 inhibition in a PDX model of FLT3-ITD/NPM1-mutant AML (Figure 5.4a). Since venetoclax is not typically used as a monotherapy, we evaluated the effects of combined venetoclax and azacitidine. Fifteen weeks post-engraftment, we randomized the mice into control and treatment groups based on their peripheral blood chimerism levels (Figure 5.4b). We performed single-cell RNA-seq (scRNA-seq) from harvested bone marrow in the control and treatment groups. Similar to our scATAC-seq analysis, we identified cell types using a reference dataset (Figure 5.4c, d). We found that venetoclax/azacitidine treatment was associated with a depletion of HSC-like AML cells, along with a subsequent increase in the CMP LMPP-like and early erythrocyte-like population. These findings provide pre-clinical evidence supporting the role of differentiation state plasticity following venetoclax treatment.

5.2.5 Conclusions

In our work, we have investigated the response to BCL2 inhibition in AML. We used an AML cell model (OCI-AML8227) that stably produces a heterogeneous population of

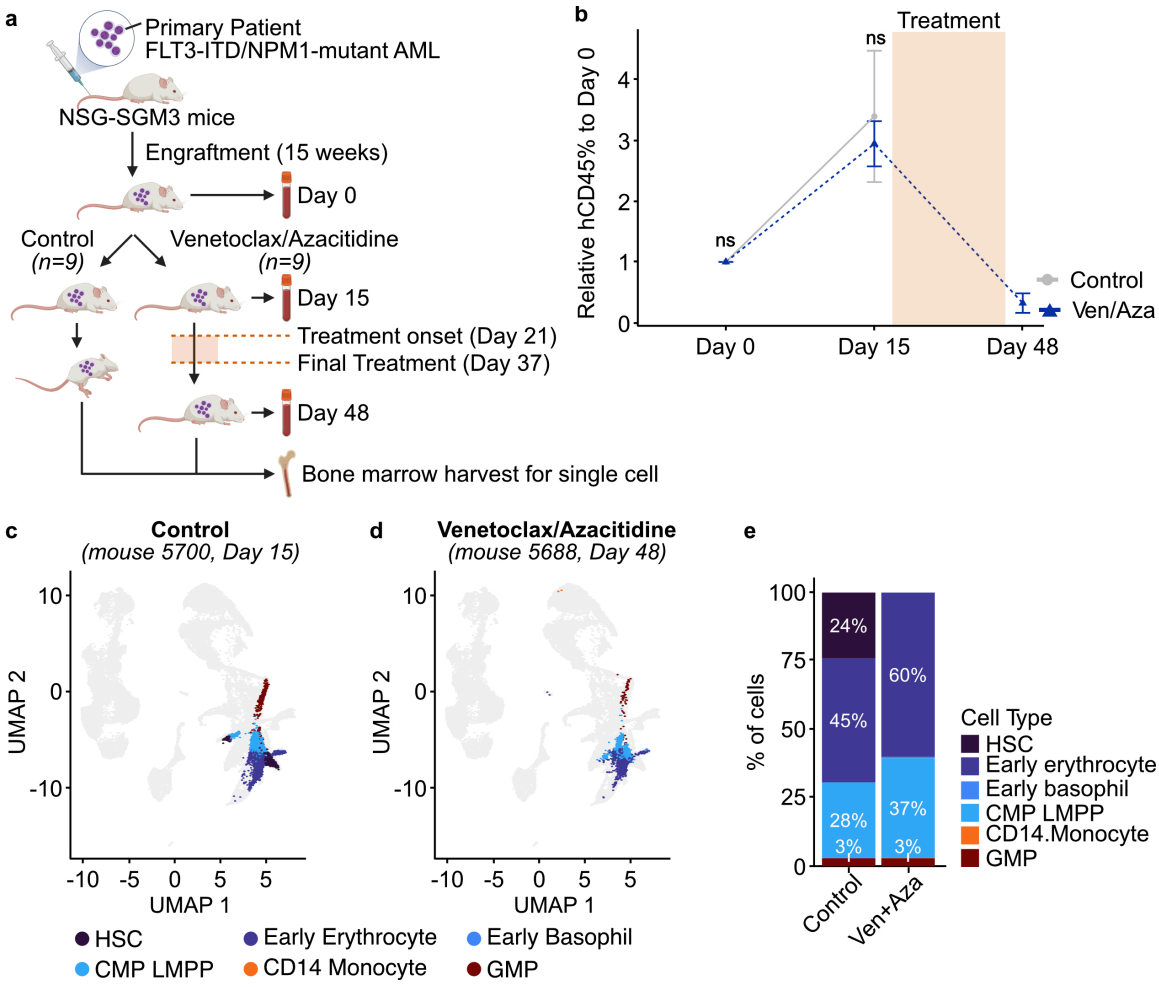

Figure 5.4: Long-term BCL2 inhibition is associated with an emergence of differentiated AML cells.

a. Study schematic (created using BioRender). **b.** Mean ± SD is shown. *P*-values were calculated using Student's t-test comparing groups on the same day. **c-e.** scRNA-seq was performed from harvested bone marrow. Cell types were identified as described in Figure 3a-c, except with a scRNA-seq dataset (grey). ns = not significant, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001.

immature and differentiated AML cells. Consistent with previous studies in AML patient samples, we found that immature $CD34^+$ cells were significantly more sensitive to BCL2 inhibition than differentiated $CD34^-$ cells. BLC2 inhibition resulted in a dose-dependent decrease in immature $CD34^+CD38^-$ cells and a simultaneous increase in differentiated $CD34^-CD38^+$ cells. Notably, this reduction of immature cells was likely not due to the

induction of apoptosis. Single-cell chromatin accessibility profiling revealed that while short-term BCL2 inhibition resulted in a minor increase in the number of differentiated AML cells, it primed cells for a shift in their differentiation state. Following long-term BCL2 inhibition in an AML PDX model, we observed a depletion of HSC-like cells and a subsequent emergence of CMP LMPP-like and early erythrocyte-like cells. These findings support the role of differentiation state plasticity in BLC2 inhibitor resistance.

5.3 Methods and materials

5.3.1 Cell culture

5.3.1.1 OCI-AML8227 cell culture

OCI-AML8227 cells were provided by Dr. John Dick (University of Toronto) under a materials transfer agreement. These cells were shipped frozen on dry ice and then thawed in a water bath set to 37°C. A thaw media consisting of 49.5% FBS (HyClone), 49.5% X-VIVO 10 (Lonza #04-380Q), and 1% DNAse I (Roche #11284932001) was slowly added drop-wise to the cells. They were incubated for 10 minutes at room temperature and then washed twice with DPBS (Gibco). After washing, the cells were transferred into the culture media consisting of sterile filtered (0.22 um, EMD Milipore #SE1M179M6) 20% BIT (StemCell Techologies #9500) and 80% X-VIVO 10 supplemented with IL-3 (10 ng/mL; Peprotech #200-03), SCT (100 ng/mL; Peprotech #300-07), IL-6 (20 ng/mL; Peprotech #200-06), Flt3-Ligand (50 ng/mL; Peprotech #300-19), G-CSF (10 ng/mL; Peprotech #300-23), and TPO (25 ng/mL; Peprotech #300-18). Unless otherwise stated, OCI-AML8227 cells were cultured in the inner eight wells of a 24-well plate with the outer wells filled with DBPS. Additionally, OCI-AML8227 cells were

seeded at a concentration of 4e5 cells per well and then incubated for 48 hours in 5% CO_2 and 37°C before experiments.

5.3.1.2 Immunomagnetic fractionation

OCI-AML8227 were immunomagnetically fractioned for CD34 surface expression using CD34 Microbeads (Miltenyi Biotec #130-046-702). Cells were resuspended in sterile filtered (0.22 um) MACS buffer (0.5% BSA and 2 mM EDTA). Cells were incubated with CD34 Microbeads for 30 minutes at 4°C. The beads were washed from the media, resuspended in MACS buffer, and then passed through a MACS Cell Separator LS column (Miltenyi Biotec #130-042-401) with a pre-separation cell strainer (Miltenyi Biotec #130-101-812). The column was washed three times with MACS buffer. $CD34^+$ OCI-AML8227 cells were isolated from the LS column flow through off of the MACS separator.

5.3.1.3 Flow cytometry

OCI-AML8227 cells were first counted using a Bio-RAD TC20 Automated Cell Counter to assess differentiation. An equal number of cells in each condition were stained with CD34 FITC (BD Pharmigen Cat #555821), CD38 APC (BD Pharmigen Cat #555462), and CD14 BV421 (BioLegend #301830). Stained cells were analyzed using a Sony SH800S. The flow cytometry data was subsequently analyzed using FlowJo. The number of cells was calculated by multiplying the fraction of cells in a quadrant by the number of cells counted before flow cytometry was performed.

5.3.1.4 Apoptosis

Apoptosis was assessed 72 hours after venetoclax treatment by flow cytometry using an eBioscience Annexin V-FITC apoptosis detection kit (ThermoFisher #88-8102-72) according to the manufacturer's instructions.

5.3.1.5 Drug Sensitivity

OCI-AML8227 cells were cultured for 72 hours along a 7-point curve of venetoclax (ABT-199; Selleckchem #S8048). Cell viability was assessed by CellTiter Aqueous colorimetric assay.

5.3.2 Mouse xenograft model

5.3.2.1 Mice

NSG-SGM3 mice (Strain #013062) were obtained from The Jackson Laboratory. All animals were maintained on a standard 12-hour light and 12-hour dark cycle and provided *ad libitum* access to water and food (Purina Mills laboratory rodent diet 5001). 7-week-old male mice were used for experimentation. All experiments were conducted in accordance with the NIH *Guide for the Care and Use of Laboratory Animals* and approved by the Institutional Animal Care and Use Committee of Oregon Health & Science University[304].

5.3.2.2 Mouse xenograft model

Splenic cells from an FLT3-ITD, NPM1-mutant AML, patient-derived murine model were obtained from the Dana Farber Cancer Institute (Model #CPDM_1061X). Non-irradiated donor mice were engrafted with 1e6 cells by tail vein injection.

5.3.2.3 Drug preparation and treatment

Venetoclax (MedChemExpress #HY-15531) and azacitidine (MedChemExpress #HY-10586) were reconstituted with dimethyl sulfoxide (DMSO), sterile filtered (0.2 μM), and stored in aliquots at −80°C. On the day of treatment, venetoclax was reconstituted with PEG300 (Selleckchem #S6704), Tween-80 (MedChemExpress #HY-Y1891), and ddH20. Azacitidine was reconstituted with PEG300 and ddH20. Venetoclax was administered by oral gavage for five days on and two days off for a total of 21 days. Azacitidine was administered every day for 21 days by intraperitoneal injection. Weights were measured daily during treatment, and doses were recalculated to ensure the mice received a consistent dose of 50 mg/kg venetoclax and 2.5 mg/kg azacitidine[305,306].

5.3.2.4 Mouse xenograft study

Complete blood counts were obtained weekly starting 15 weeks following engraftment (Day 0) using an automated counter (Scil Vet abc). Flow cytometry was used to monitor for human blood chimerisms and was performed on BD LSRFortessa with human CD45 FITC (Invitrogen #11-0459-42) and mouse CD45 APC (Invitrogen #17-0451-82) antibodies. Mice were assigned into control and treatment groups based on their Day 0 engraftment levels. Daily treatment administration was maintained between Days 21-37. Blood and bone marrow smears were prepared from mice found deceased or at the time of sacrifice (Day 48).

5.3.3 Sequencing

5.3.3.1 Single-cell ATAC-seq

OCI-AML8227 cells were treated with 1 μM venetoclax for 24 hours before assessment of chromatin accessibility. Single-cell ATAC-seq was performed using the

s3ATAC-seq protocol as previously described[8]. Briefly, nuclei were isolated from OCI-AML8227 cells using nuclei isolation buffer (10 mM HEPES-KOH pH 7.2 [Fisher Scientific #BP310-500 and Sigma Aldrich #1050121000], 10 mM NaCl [Fisher Scientific #S271-3], 3 mM $MgCl_2$ [Fisher Scientific #AC223210010], 0.1% IGEPAL-CA630 [Sigma Aldrich #I3021], 0.1% Tween-20 [Sigma Aldrich #P-7949], and EDTA-free Pierce protease inhibitor [ThermoFisher #A32955]). Nuclei were tagmented in tagmentation buffer (132 mM TAPS [[tris(hydroxymethyl)methylamino]propanesulfonic acid] pH 8.5 [Sigma Aldrich #T0647], 264 mM potassium acetate [Sigma Aldrich #P1190], 40 mM magnesium acetate tetrahydrate [Sigma Aldrich #M5661], 12 mM D-Glucosamine sulfate [Fisher Scientific #AAJ6627106], and NN-Dimethylformamide [Sigma Aldrich #D4551]) using Tn5 transposase (Scale Biosciences). During tagmentation, nuclei were incubated for 30 minutes at 42°C while shaking at 300 RPM on an Eppendorf ThermoMixer C. Tagmented nuclei were pooled, stained with DAPI (Cell Signaling Technologies #4083), and sorted into a 96-well plate using a Sony SH800S cell sorter. Nuclear proteins were denatured by incubation at 55°C for 15 minutes on an Eppendorf Mastercycler nexus with the lid heated to 90°C. Linear extension and adapter switching were performed in nuclei resuspended in 300 nM LNA oligo, 1.5% Triton-X100 (Sigma Aldrich #T8787), and VeraSeq PCR Mix (Qiagen #P7610L) on an Eppendorf Mastercycler nexus. qPCR was performed using VeraSeq Buffer II (Qiagen #B7102), 10 mM dNTPs (ThermoFisher #R0192), and VeraSeq Ultra Polymerase (Qiagen #P7520L). Libraries were cleaned using the DNA Clean & Concentrator-5 kit (Zymo Research #D4013) and then were size-selected using MagBind TotalPure NGS Binding Beads (Omega Biotech #M1378-01). Cleaned libraries were quantified using an Invitrogen Qubit and an Agilent TapeStation. Libraries were diluted to 600 pM and then sequenced with 100 bp PE on an Illumina NextSeq 2000.

5.3.3.2 Single-cell RNA-seq

Nuclei were extracted from harvested bone marrow and then prepared using the demonstrated primary cell nuclei extraction protocol from 10x Genomics. RNA libraries were prepared using Chromium Single Cell RNA Library and Dual Index Next GEM Single Cell 3' Reagent Kits v3.1 (10x Genomics #PN-1000268). Libraries were sequenced with 100 bp PE sequencing by the OHSU MPSSR.

5.3.4 Analysis

5.3.4.1 Single-cell ATAC-seq

Single-cell ATAC-seq libraries were demultiplexed with the Tn5, i5, and i7 barcodes and the Tn5 tagmentation plate position using a custom script. The demultiplexed fastqs were aligned to the human genome (hg19) using bwa align[288]. Aligned fastqs were filtered for chromosomes 1-22, X, Y, and M, and insert sizes between 20-10,000 bp using SAMtools[307]. Duplicates were removed using a custom script. ATAC-seq fragment coverage was quantified in a common coordinate system with Signac FeatureMatrix()Field 299 for each deduplicated BAM. The common coordinate system was created by segmenting the hg19 genome into 5,000 bp bins. Signac was used to generate an integrated object from each fragment file, using a TSS enrichment of two and 1,000 unique fragments as cutoffs[300]. Sample and reference dataset transfer anchors were identified using Signac FindTransferAnchors() and then mapped with Signac MapQuery[300]. The Healthy Hematopoiesis scATAC-seq dataset compiled by Granja et al. (https://github.com/GreenleafLab/MPAL-Single-Cell-2019) was the reference dataset used in this analysis[308]. Cells were assigned cell-type labels according to their mapping to the reference latent space. Pseudotime was calculated on an integrated object containing DMSO- and venetoclax-treated cells using monocle3[309].

5.3.4.2 Single-cell RNA-seq

Single-cell RNA-seq libraries were aligned to the murine genome (mm10) using 10x Genomics Cell Ranger[299]. Seurat was used to generate an integrated object from each library, which included features present in at least 10 cells, cells with at least 500 features and at most 8,000 features, and cells with at most 15% of features associated with mitochondrial genome[310]. Cell cycle genes were regressed during gene count normalization and scaling with Seurat. Sample and reference dataset transfer anchors were identified using Seurat FindTransferAnchors() and then mapped with Seurat MapQuery[310]. The Healthy Hematopoiesis scRNA-seq dataset compiled by Granja et al. (https://github.com/GreenleafLab/MPAL-Single-Cell-2019) was the reference dataset used in this analysis[308]. Cells were assigned cell-type labels according to their mapping to the reference latent space.

5.3.5 Quantification and Statistical Analysis

Values are represented as the mean, and error bars are the SEM unless otherwise stated. Prism (version 9.1; Prism Software Corp.) or R was used for statistical analyses. Significance was tested using Student's t-test or one-way ANOVA followed by Holm-Šidák post-test correction unless otherwise stated.

5.4 Authors' contributions

WM Yashar: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing-original draft, writing-review and editing, visualization. **IV Pacentine**: software, formal analysis, investigation. **M Tsuchiya**: investigation, writing-review and editing. **J Macaraeg**: investigation, writing-review and

editing. **A Taherinasab**: methodology. **S Evans-Dutson**: methodology. **B O'Connell**: methodology, resources. **T Lusardi**: software, data curation, writing-review and editing. **N Szczepanski**: software, data curation. **GG Yardimci**: conceptualization, resources, writing-review and editing, supervision. **AC Adey**: conceptualization, resources, supervision, funding acquisition. **JE Maxson**: conceptualization, resources, writing-review and editing, supervision, funding acquisition. **TP Braun:** conceptualization, methodology, writing-review and editing, supervision, project administration, funding acquisition.

5.5 Acknowledgements

We thank the patients who donated their precious time and samples to support this research. We appreciate the OHSU ExaCloud Cluster Computational Resource and the Advanced Computing Center for their assistance. The OCI-AML8227 cells were generously provided by Dr. John Dick (University of Toronto). This work was funded by the National Institutes of Health (K08 CA245224 and 1R01 CA282133-01 to TP Braun), the American Cancer Society (P30 CA016672-40 to TP Braun), and the American Society of Hematology Research Restart and Scholar Awards to TP Braun.

Chapter 6: Conclusions and future directions

6.1 Identification of histone modification peaks from CUT&Tag data using GoPeaks

6.1.1 Conclusions

Histone modifications regulate gene transcription and are associated with genetic regulatory elements, including enhancers and promoters. Regions of modified histones can be identified using next-generation sequencing techniques, including ChIP-seq, CUT&Tag, and CUT&RUN, as stacks of aligned reads[13,18–20]. Algorithms have been developed to identify genome-wide enrichment of aligned reads, also known as peaks, and distinguish peaks of modified histones from noise and artifacts. ChIP-seq peak calling methods, including MACS2, were designed to address the high background rate in ChIP-seq. As a result, they are vulnerable to mistaking background signal as peaks, particularly when the background is low[41–43]. While SEACR was developed to perform transcription factor peak calling from low background CUT&RUN data, no method has been designed to call peaks from histone modification CUT&Tag data[43]. In our work, we developed GoPeaks and compared its performance against commonly used peak calling algorithms to detect histone modifications frequently used in epigenetic studies. We showed that GoPeaks detects peaks of histone modifications that are frequently used in epigenetic studies, including H3K4me3, H3K4me1, H3K27me3, and H3K27ac with high sensitivity and specificity. Notably, GoPeaks demonstrated an improved ability to identify H3K27ac peaks, which are critical to localizing active regulatory non-coding elements

throughout the genome, over other standard peak calling algorithms. Moreover, we showed that GoPeaks is able to detect peaks from other epigenetic profiling techniques, including ChIP-seq, CUT&RUN, and ATAC-seq.

6.1.2 Future directions

GoPeaks was developed to identify peaks from bulk CUT&Tag data. However, peak calling is not limited to ChIP-seq and CUT&Tag and is essential for analyzing various epigenetic profiling data. Therefore, we investigated whether GoPeaks identifies transcription factor peaks from CUT&RUN and chromatin accessibility peaks from ATAC-seq. We showed that GoPeaks was able to robustly identify peaks in both of these modalities. While these findings reassured us, the field is rapidly moving away from bulk sequencing techniques in favor of single-cell sequencing[189,311]. To keep up with the pace of single-cell sequencing technologies, future studies need to adapt GoPeaks to identify peaks from scATAC-seq and scCUT&Tag data. Data sparsity is a well-documented limitation of most single-cell modalities[7]. Since GoPeaks was designed to identify regions of enriched signal in the context of low background, it is well-positioned to be applied to single-cell data. However, it is unclear whether GoPeaks will need to deploy common single-cell strategies, such as pseudobulking, to improve its robustness and reduce the false discovery rate in the setting of single-cell sequencing data[312]. Future studies should investigate the ability of GoPeaks to call peaks from single-cell sequencing epigenetic profiling data.

6.2 Predicting transcription factor activity using prior biological information

6.2.1 Conclusions

Dysregulation of normal transcription factor activity is a common driver of disease. Many methods have been developed to infer transcription factor activity from RNA-seq data[50–52]. These approaches differ based on how they select gene expression features, either using the expression of the transcription factors themselves, every gene in the dataset, genes from expert-curated sets, or genes nominated from network inference analysis[59–63,65,66]. Despite advances in these methods, however, recent studies have highlighted that grounding predictions using transcription factor activity methods remains challenging[68–72]. Despite this, several precision medicine clinical studies use inferred transcription activity from bulk RNA-seq as a marker to guide clinical decisions. While there is an increasing number of single-cell and spatial -omic modalities available to clinical researchers, many larger cohorts most commonly use markers identified from bulk RNA or DNA sequencing. Therefore, it is critical to develop methods that can robustly detect aberrant transcription factor activity from primary patient bulk RNA-seq data.

In our studies, we developed a computational method called Priori that uses prior, peer-reviewed biological information to infer transcription factor activity. Our method has two major advantages over the existing methods. First, Priori identifies transcription factor target genes using carefully extracted transcriptional regulatory networks from Pathway Commons[205,206]. Using these transcriptional relationships, Priori fits linear models to the expression of transcription factors and their target genes. Second, comparison with a third-party benchmarking workflow reveals that Priori detects

aberrant transcription factor activity from 124 gene perturbation experiments with higher sensitivity and specificity than 11 other methods. We applied Priori and three different methods nominated from the benchmarking workflow to generate activity scores from two large primary patient datasets, TCGA BRCA and Beat AML. We demonstrate that Priori can be deployed to discover significant predictors of survival in breast cancer and identify mediators of drug response in leukemia from primary patient samples that were not robustly detected using the other methods.

6.2.2 Future directions

Priori generates transcription factor activity scores by fitting linear models to gene expression data. Linear models have many benefits for biologists, including ease of interpretation. For example, Priori generates individual target gene weights, allowing users to understand the contribution of each target gene to the overall activity score. This will enable researchers to triage targets that may be important drivers of transcriptional regulation. However, computational biology is rapidly adopting non-linear models. Conceptually, models such as random forest, support vector regression, k-nearest neighbors, or artificial neural networks can capture non-linear relationships between transcription factors and their target genes. While we were encouraged to see that Priori identified perturbed transcription factor activity better than other methods, Priori's predictions still have room for improvement. It would be worthwhile to empirically test other model structures, separately or in combination with linear models, to understand which structure best captures transcriptional relationships.

Priori uses transcriptional relationships from Pathway Commons as prior information. The Pathway Commons relationships have many benefits, including grounding findings in literature-supported biological information from a vast dataset (22

high-quality databases with over 5,700 detailed pathways and 2.4 million interactions)[205,206]. However, Pathway Commons is designed to curate and generalize transcriptional relationships across many tissue types. Transcription factor regulation is likely context-dependent. MAX, for example, functions as an inhibitor when binding to DNA as a heterodimer with MNT or MXD1 and as an activator when binding as a heterodimer with MYC[313]. While the generalized transcriptional relationships from Pathway Commons are likely tuned by Priori when generating the target gene weights, Priori's performance may be improved with refined prior data. It may be more appropriate to generate Priori activity scores using disease-specific networks. Other groups have deployed a similar approach, including the authors of VIPER. They encourage researchers to generate activity scores using their pre-trained networks inferred from TCGA datasets[65]. Moreover, single-cell sequencing may allow researchers to generate context-dependent networks. There are major efforts to characterize normal and diseased tissue at single-cell resolution. One such effort, Tabula Sapiens, is compiling a multi-organ, single-cell transcriptomic atlas of humans. Resources like this will provide an unprecedented ability to study the regulatory dynamics of transcription factors within different cell types. Since transcription factors and target genes have a lead-lag relationship, mathematical models, including Granger causality, can infer transcription factor and target gene relationships[314]. These inferred transcriptional relationships could be used to refine the generalized Pathway Commons relationships for each dataset. The utility of context-dependent prior networks grounded in literature-supported relationships to predict transcription factor activity should be further explored.

Our work identified 11,075 significant inhibitor-transcription factor activity relationships from 859 patients with AML using Priori. Previous findings showed that disruption of FOXO1 activity sensitized MCL cell lines to venetoclax[232]. As this

relationship had not been tested in leukemia, we evaluated it using a CRISPR-Cas9 knockdown of *FOXO1* in a cell line model of monocytic AML. We found that *FOXO1* knockdown sensitized these cells to venetoclax, consistent with Priori's prediction. While we were encouraged by these findings, it would be beneficial to systematically test these significant inhibitor-transcription factor activity relationships using an arrayed CRISPR screen. Guide RNA libraries comprised of Cas9 single guide RNAs can be developed to target one gene per well in a multi-well plate[315]. *Cas9*-expressing monocytic AML cells can be therapeutically perturbed with small molecule compounds, such as venetoclax, and then evaluated for cell viability. This technique would identify transcription factors whose knockdown sensitizes monocytic AML to venetoclax. This experiment would not only allow us to evaluate the robustness of Priori's predictions but also design rationale combination therapies with venetoclax.

6.3 Disruption of the MYC super-enhancer with dual FLT3/LSD1 inhibition in FLT3-ITD AML

6.3.1 Conclusions

Mutations in FLT3 are among the most common mutations in AML and confer a poor prognosis. The use of FLT3 inhibitors has demonstrated improved clinical benefits in the setting of relapsed/refractory FLT3-ITD AML compared to salvage chemotherapy. However, FLT3 inhibitors are rarely curative, and patient remissions are short-lived, limiting their use as adjuncts to conventional therapy. Thus, there is a need for new therapeutic drug combinations to treat this disease.

LSD1 inhibitors have shown promise to improve clinical responses to FLT3 inhibitors. My lab revealed in an unbiased drug screen that LSD1 inhibition enhances the

efficacy of a JAK/STAT kinase inhibitor in JAK/STAT-driven AML[153]. As the JAK/STAT signaling pathway is activated downstream from FLT3, we reasoned that LSD1 inhibition might also improve the activity of FLT3 inhibitors in FLT3-ITD AML[316–320]. In our studies, we showed that LSD1 enhanced the activity of FLT3 inhibitors in FLT3-mutant AML. Multi-omic profiling revealed that the drug combination disrupts STAT5, LSD1, and GFI1 binding at the MYC BENC, suppressing super-enhancer accessibility as well as MYC expression and activity. Combined LSD1 and FLT3 results in the accumulation of repressive H3K9me1 methylation, an LSD1 substrate, at MYC target genes. We validated these findings in 72 primary AML samples, with nearly every sample demonstrating synergistic responses to the drug combination. Collectively, these studies reveal how epigenetic therapies augment the activity of kinase inhibitors in FLT3-ITD AML.

6.3.2 Future directions

A crucial component of the FLT3/LSD1 inhibitor synergy mechanism was altering *MYC* expression by regulating the MYC BENC. While others have demonstrated that *MYC* transcription can be altered by inhibiting general chromatin regulators, disruption of MYC BENC activity by combined epigenetic modulatory drugs and kinase inhibitors is a novel approach to targeting this central oncogenic regulator[127–130]. Our scATAC-seq analysis revealed substantial variation in the MYC BENC module utilization pattern between AML samples at baseline and in response to drug treatment. Indeed, other studies have suggested that a distinct set of transcription factors binds each BENC module and regulates *MYC* expression in specific blood cell lineages[122]. Understanding MYC BENC module utilization between molecularly defined AML subtypes and its impact on drug responses is crucial for future studies.

Prior work on LSD1 inhibitors has broadly implicated the pro-differentiation effects of these drugs as the central mechanism of cytotoxicity. Our work here shows that LSD1 inhibition activates enhancers associated with PU.1. Other groups have demonstrated that suppression of *SPI1* expression results in a block in LSD1-inhibitor-induced differentiation and decreased cytotoxicity[150]. While our work confirmed the role of PU.1 as a putative mediator of LSD1-inhibitor responses, we found that *SPI1* knockdown had little effect on the transcriptional or cytotoxic response to dual FLT3/LSD1 inhibition. Moreover, combined FLT3/LSD1 inhibition did not significantly affect myeloid differentiation. It is unclear whether the PU.1-associated transcriptional effects observed in our study are crucial to the drug effect. Investigation of the pro-differentiation effects of dual FLT3/LSD1 inhibition will be a necessary question for future investigation.

Previous studies of LSD1 inhibitors have also demonstrated that drug efficacy is dependent on the interruption of LSD1 scaffolding activity rather than its demethylation activity[132,151]. Our work confirmed that a critical component of LSD1 inhibitor activity is the disruption of LSD1 binding to GFI1/CoREST. However, LSD1 inhibition also resulted in the accumulation of repressive H3K9me1 marks at the promoters of MYC target genes. While LSD1 canonically demethylates activating H3K4 marks, alternative LSD1 complexes remove repressive H3K9 methylation marks in cells from other tissues, resulting in transcriptional activation[139,140]. In prostate cancer, LSD1 forms a chromatin-associated complex with androgen receptor that demethylates H3K9 and de-represses androgen receptor target genes. In neuronal cells, on the other hand, an LSD1 isoform, LSD1+8a, complexes with supervillain and demethylates H3K9me2 to regulate neuronal differentiation. Interestingly, the H3K9 demethylation activity of LSD1 may be slightly antagonized by FLT3 inhibition as the H3K9me1 signal in cells treated with the drug

combination was lower than in those only treated with LSD1 inhibition. Further work needs to be done to nominate binding factors with LSD1 or LSD1 isoforms that, as a complex, function as a transcriptional activator by H3K9 demethylation and how these complexes are affected by FLT3 inhibitors.

We demonstrated that combined FLT3 and LSD1 inhibition effectively induces leukemic cell death in FLT3-ITD AML cell line models and primary patient samples. While *ex vivo* treatment of cell lines and primary patients is useful for its ease of use, this system may only partially reflect *in vivo* disease. Our work provides a preclinical rationale for using combined FLT3 and LSD1 inhibition in FLT3-ITD AML, which needs to be validated clinically. Since our findings have been released, a Phase I clinical trial investigating the efficacy of combined iadademstat and gilteritinib in patients with relapsed/refractory FLT3-ITD AML has been initiated (FRIDA Trial; ID #NCT05546580). This trial will not only evaluate the safety, tolerability, pharmacokinetics, and pharmacodynamics of combined FLT3 and LSD1 inhibition but also to understand the *in vivo* mechanism of the drug combination. As a part of this trial, we will have access to pre- and post-treatment bone marrow and peripheral blood samples from these patients. Using scATAC-seq, we can understand how the drug combination disrupts transcriptional regulation at the MYC BENC. The FRIDA trial and subsequent clinical investigations will be necessary to understand the efficacy of combined FLT3 and LSD1 inhibition among patients with FLT3-ITD AML.

6.4 Differentiation state plasticity following BCL2 inhibition in AML

6.4.1 Conclusions

Frequently, the relapsed disease following venetoclax-based-therapeutic strategies is composed of differentiated, monocyte-like cells[163]. Several studies have suggested that differentiated AML cells have a survival advantage in the presence of BCL2 inhibition due to a reliance on alternative anti-apoptotic proteins[163,164]. However, these studies do not explain why patients with dominant, differentiated subclones frequently achieve long-lasting remission with BCL2-inhibitor-based strategies[163]. Therefore, simple selection of differentiated cells appears insufficient to describe BCL2 inhibitor resistance. Increasing evidence shows that therapeutic pressure can facilitate transcriptional programs that confer drug resistance. In the context of BCL2 inhibitor resistance, the extent to which the emergence of differentiated AML cells is the result of intrinsic resistance or transcriptional adaptation is unclear.

In our studies, we found that BCL2 inhibition promotes the differentiation of AML cells. Using an AML cell line model (OCI-AML8227) that recapitulates the functional hematopoietic hierarchy, we showed that venetoclax decreases the number of immature $CD34^{+}CD38^{-}$ cells and simultaneously increases the number of differentiated $CD34^{-}CD38^{+}$ cells. To understand the transcriptional dynamics following venetoclax treatment, we investigated the single-cell chromatin accessibility profiles of OCI-AML8227 cells treated with venetoclax for 24 hours. This short-term venetoclax treatment primed immature HSC-like and CMP LMPP-like AML cells for myeloid differentiation. To understand the effects of long-term venetoclax treatment in a clinically relevant model, we performed a venetoclax drug study in a PDX mouse model of FLT3-ITD/NPM1-

mutant AML. We found that venetoclax depleted HSC-like AML cells and increased the production of early erythrocyte-like and CMP LMPP-like cells. These findings highlight the role of differentiation state plasticity as a mechanism of resistance to venetoclax.

6.4.2 Future directions

Our studies investigating the mechanism of BCL2 inhibition have relied on temporal snapshots using single-cell sequencing data. While this data has allowed for a high-throughput dissection of the heterogeneous response to BCL2 inhibition, it does not provide a complete view of the drug effect. Single-cell sequencing cannot link the detailed states of cells to their ultimate fate, as cells are destroyed in the process of measurement. Moreover, given the high-dimensional nature of single-cell sequencing data, pseudotime trajectory inference can nominate several cell fate trajectories from the same data[321]. Single-cell barcoding methods have been developed to link the whole transcriptome of the cell to its long-term dynamic behavior. One system, called Lineage and RNA recovery (LARRY), labels cells with lentivirally-delivered DNA barcodes[322]. Each barcode consists of a 28 randomer in the 3′ untranslated region of an enhanced green fluorescent protein transgene (eGFP) controlled by a ubiquitous EF1α promoter. Barcoded eGFP transcripts can be captured using standard single-cell sequencing methodologies, including 10x Genomics. We have created a concentrated lentivirus containing the LARRY plasmids and transduced OCI-AML8227 cells. We have generated single-cell 10x Genomics libraries from LARRY-barcoded OCI-AML8227 cells 24 hours and 72 hours following venetoclax treatment. In future studies, we need to amplify the LARRY barcodes of these libraries, which we have already demonstrated is feasible in a pilot experiment. Once we have sequenced these libraries, we will be able to trace the trajectories of the immature and differentiated AML cell populations and nominate putative regulators of transcriptional adaptation to venetoclax. With an

understanding of the adaptive mechanism of resistance to BCL2 inhibition, we can select alternative targets to prevent the outgrowth of resistant populations.

Our investigation of BCL2 inhibitor resistance showed that venetoclax decreased the number of immature $CD34^{+}CD38^{-}$ OCI-AML8227 cells without changing the apoptosis rate. This is a crucial finding as the putative mechanism of BCL2 inhibition is the reactivation of intrinsic apoptosis in AML cells[160]. To understand the apoptosis rate in drug-treated cells, we measured levels of Annexin-V, a phospholipid-binding protein with a high affinity for phospholipid phosphatidylserine (PS)[323]. In normal healthy cells, PS is located on the cytoplasmic surface of the plasma membrane. During apoptosis, PS is translocated to the extracellular side of the plasma membrane. In our study, we measured cell surface markers associated with differentiation and Annexin-V using flow cytometry. Notably, we excluded cellular debris from our flow cytometry analysis, potentially excluding apoptotic cells from our analysis. This is necessary to correct for cellular auto-fluorescence. While we could still observe a significant increase in the number of apoptotic, differentiated $CD34^{-}CD38^{+}$ OCI-AML8227 cells at 72 hours, this limitation highlights the importance of orthogonal measurements. During apoptosis, caspase-3 is responsible for the proteolytic cleavage of many essential proteins[324]. Activated caspase-3 cleaves the substrate between fluorogenic substrates N-Acetyl-Asp-Glu-Val-Asp-7-amino-4-methylcoumarin (Ac-DEVD-AMC), generating highly fluorescent AMC. As cleavage of DEVD-AMC only occurs in apoptotic cells, the amount of AMC produced is proportional to the number of apoptotic cells in a sample. Future studies should compare the rates of caspase-3 activity following venetoclax treatment to better understand its effects on apoptosis in immature and differentiated AML cells.

We found that short-term treatment with BCL2 inhibition primed HSC-like and CMP LMPP-like AML cells for myeloid differentiation. While these early time points are

important to understand the transcriptional dynamics following venetoclax treatment, we have only observed changes in differentiation-associated cell surface markers 72 hours post-therapy. While it is likely that we would observe an increase in the number of differentiated AML cells 72 hours following venetoclax using scATAC-seq, this study has yet to be completed. Moreover, this scATAC-seq time point would help with the nomination of molecular regulators mediating the differentiation shift. These studies would help us understand the molecular mechanism underlying the differentiation shift induced by BCL2 inhibition.

While our findings show that BCL2 inhibition promotes the differentiation of immature AML cells, the mechanistic underpinning of this phenomenon needs to be clarified. To nominate putative regulators of the differentiation shift induced by venetoclax, we analyzed the short-term venetoclax scATAC-seq data of OCI-AML8227 cells. Differential peak analysis comparing the venetoclax- and DMSO-treated HSC-like and CMP LMPP-like AML cells revealed 184 differentially downregulated and 44 differentially upregulated peaks. Transcription factor motif analysis of the differentially downregulated peaks nominated several transcription factors. In particular, nuclear respiratory factor 1 (NRF1) motifs were significantly enriched ($p = 1x10^{-12}$). NRF1, along with nuclear factor erythroid-derived 2-like 2 (NFE2L2 also known as NRF2), are Cap`n'Collar family transcription factors, which have a documented role in mitochondrial biogenesis[325,326]. To understand the importance of NRF1/2 in AML, I generated Priori transcription factor activity scores from baseline RNA-seq data of 707 AML patients[229]. I correlated these scores to the van Galen AML subtypes and *ex vivo* venetoclax sensitivity. I found that predicted NRF1 and NRF2 activity was significantly associated with Monocyte-like AML cells ($R^2 = 0.346$ and 0.770, respectively), suggesting that these factors are active in differentiated cells. Moreover, I found that the predicted activity of

NRF1 and NRF2 are significantly associated with venetoclax resistance (R^2 = 0.334 and 0.625, respectively). Notably, predicted NRF2 activity is the 19th highest correlate with venetoclax activity out of 233 significant relationships. The relationship between NRF1/2 activity and the pro-differentiative effects of venetoclax need to be further interrogated. While the short-term venetoclax scATAC-seq data provided insights into the impact of venetoclax, this library was only sequenced to 100 million reads. Deeper sequencing, as well as an understanding of the chromatin accessibility profiles at 72 hours following venetoclax treatment, will help determine the role of NRF1/2 or other putative regulators. Once these transcription factor(s) are identified, their causal relationships can be interrogated with functional genetic studies, including CRISPR-Cas9 knockdown. Identifying a molecular mechanism of venetoclax differentiation will nominate new targets and inform more effective, rational BLC2-inhibitor-based therapeutic strategies.

Our investigation of BCL2 inhibitor resistance has primarily relied on a leukemia stem cell line, OCI-AML8227. This model is helpful as it maintains a functional hematopoietic hierarchy and stably produces immature, differentiated AML cells. However, it is vital to understand the generalizability of the pro-differentiative effect of venetoclax in models other than OCI-AML8227. Several other leukemia stem cell models could help validate the findings from OCI-AML8227, including MUTZ3 or OCI-AML22[327,328]. More importantly, this hypothesis should be directly tested in primary patient samples. These samples can be obtained from the Beat AML initiative, which has collected frozen viables from over 900 patients with AML. There is a ready supply of samples available to evaluate our findings.

6.5 Summary

The work in this dissertation reveals how the investigation of transcriptional dysregulation in AML unveils novel therapeutic strategies. Two major findings emerged from this work. First, we found that combined FLT3 and LSD1 inhibition induced synergistic cell death in FLT3-ITD AML. Utilizing computational methods that we developed, GoPeaks and Priori, we showed that this drug combination suppressed transcriptional regulators of the MYC blood super-enhancer complex and its activity. These findings propose a novel therapeutic strategy for patients with FLT3-ITD AML, which is currently under investigation in a Phase I clinical trial. Second, we determined that BCL2 inhibition induces the differentiation of immature AML cells. These findings revealed a novel resistance mechanism to BCL2 inhibition, which can be targeted to prevent the outgrowth of resistant populations.

References

1. Huisinga, K. L., Brower-Toland, B. & Elgin, S. C. R. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma* **115**, 110–122 (2006).
2. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).
3. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**, 1518–1552 (2022).
4. Goryshin, I. Y. & Reznikoff, W. S. Tn5 in vitro transposition. *J Biol Chem* **273**, 7367–7374 (1998).
5. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 1–14 (2018).
6. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010–3014 (1992).
7. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).
8. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat Biotechnol* **39**, 1574–1580 (2021).
9. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381–395 (2011).

10. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).

11. Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. *Cell* **128**, 707–719 (2007).

12. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

13. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

14. Buratowski, S. & Kim, T. The Role of Cotranscriptional Histone Methylations. *Cold Spring Harb Symp Quant Biol* **75**, 95–102 (2010).

15. Pinskaya, M. & Morillon, A. Histone H3 Lysine 4 di-methylation: A novel mark for transcriptional fidelity? *Epigenetics* **4**, 302–306 (2009).

16. Carrozza, M. J., Utley, R. T., Workman, J. L. & Côté, J. The diverse functions of histone acetyltransferase complexes. *Trends in Genetics* **19**, 321–329 (2003).

17. Yang, X.-J. & Seto, E. HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene* **26**, 5310–5318 (2007).

18. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).

19. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651–657 (2007).

20. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).

21. Policastro, R. A. & Zentner, G. E. Enzymatic methods for genome-wide profiling of protein binding sites. *Brief Funct Genomics* **17**, 138–145 (2018).

22. Teytelman, L., Thurtle, D. M., Rine, J. & Oudenaarden, A. van. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *PNAS* **110**, 18602–18607 (2013).

23. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).

24. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

25. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc* **13**, 1006–1019 (2018).

26. Hainer, S. J. & Fazzio, T. G. High-Resolution Chromatin Profiling Using CUT&RUN. *Curr Protoc Mol Biol* **126**, e85 (2019).

27. Pundhir, S., Poirazi, P. & Gorodkin, J. Emerging applications of read profiles towards the functional annotation of the genome. *Front Genet* **6**, 188 (2015).

28. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

29. Roh, T.-Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).

30. Roh, T., Wei, G., Farrell, C. M. & Zhao, K. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.* **17**, 74–81 (2007).

31. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931–21936 (2010).

32. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**, 897–903 (2008).

33. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).

34. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349–353 (2006).

35. Lee, T. I. *et al.* Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. *Cell* **125**, 301–313 (2006).

36. Roh, T.-Y., Cuddapah, S., Cui, K. & Zhao, K. The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences* **103**, 15782–15787 (2006).

37. Bernstein, B. E. *et al.* Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell* **120**, 169–181 (2005).

38. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).

39. Fox, S. *et al.* Hyperacetylated chromatin domains mark cell type-specific genes and suggest distinct modes of enhancer function. *Nat Commun* **11**, 4544 (2020).

40. Zhang, T., Zhang, Z., Dong, Q., Xiong, J. & Zhu, B. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology* **21**, 45 (2020).

41. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

42. Thomas, R., Thomas, S., Holloway, A. K. & Pollard, K. S. Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics* **18**, 441–450 (2017).

43. Meers, M. P., Tenenbaum, D. & Henikoff, S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics & Chromatin* **12**, 42 (2019).

44. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).

45. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

46. Bushweller, J. H. Targeting transcription factors in cancer — from undruggable to reality. *Nat Rev Cancer* **19**, 611–624 (2019).

47. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).

48. Lin, C. Y. *et al.* Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell* **151**, 56–67 (2012).

49. Nie, Z. *et al.* c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**, 68–79 (2012).

50. Khatri, P., Sirota, M. & Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology* **8**, e1002375 (2012).

51. Nguyen, T.-M., Shafi, A., Nguyen, T. & Draghici, S. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol* **20**, 203 (2019).

52. Essaghir, A. *et al.* Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Research* **38**, e120 (2010).

53. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5**, 1512–1526 (2009).

54. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227–232 (2012).

55. Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J. & Smith, V. A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci Rep* **5**, 10775 (2015).

56. Kiełbasa, S. M. & Vingron, M. Transcriptional Autoregulatory Loops Are Highly Conserved in Vertebrate Evolution. *PLOS ONE* **3**, e3210 (2008).

57. Benito, J., Zheng, H., Ng, F. S. & Hardin, P. E. Transcriptional feedback loop regulation, function and ontogeny in Drosophila. *Cold Spring Harb Symp Quant Biol* **72**, 437–444 (2007).

58. Bornstein, C. *et al.* A negative feedback loop of transcription factors specifies alternative dendritic cell chromatin states. *Mol Cell* **56**, 749–762 (2014).

59. Teschendorff, A. E. & Wang, N. Improved detection of tumor suppressor events in single-cell RNA-Seq data. *npj Genom. Med.* **5**, 1–14 (2020).

60. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).

61. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 1–15 (2013).

62. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).

63. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* **2**, vbac016 (2022).

64. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z. & DeLisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics* **13**, 281–291 (2012).

65. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**, 1–15 (2006).

66. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).

67. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838–847 (2016).

68. Olsen, C. *et al.* Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* **103**, 329–336 (2014).

69. Walhout, A. J. M. What does biologically meaningful mean? A perspective on gene regulatory network validation. *Genome Biol* **12**, 109 (2011).

70. Barbuti, R., Gori, R., Milazzo, P. & Nasti, L. A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bioinspired models. *J Membr Comput* **2**, 207–226 (2020).

71. Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. Bioinformatics challenges for personalized medicine. *Bioinformatics* **27**, 1741–1748 (2011).

72. Yngvadottir, B., Macarthur, D. G., Jin, H. & Tyler-Smith, C. The promise and reality of personal genomics. *Genome Biol* **10**, 237 (2009).

73. Fliedner, T. M., Graessle, D., Paulsen, C. & Reimers, K. Structure and function of bone marrow hemopoiesis: mechanisms of response to ionizing radiation exposure. *Cancer Biother Radiopharm* **17**, 405–426 (2002).

74. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644 (2008).

75. Rhodes, J. *et al.* Interplay of Pu.1 and Gata1 Determines Myelo-Erythroid Progenitor Cell Fate in Zebrafish. *Developmental Cell* **8**, 97–108 (2005).

76. Galloway, J. L., Wingert, R. A., Thisse, C., Thisse, B. & Zon, L. I. Loss of Gata1 but Not Gata2 Converts Erythropoiesis to Myelopoiesis in Zebrafish Embryos. *Developmental Cell* **8**, 109–116 (2005).

77. Iwasaki, H. & Akashi, K. Myeloid Lineage Commitment from the Hematopoietic Stem Cell. *Immunity* **26**, 726–740 (2007).

78. Goldman, S. L. *et al.* Epigenetic Modifications in Acute Myeloid Leukemia: Prognosis, Treatment, and Heterogeneity. *Front Genet* **10**, 133 (2019).

79. Deininger, M. W. N., Tyner, J. W. & Solary, E. Turning the tide in myelodysplastic/myeloproliferative neoplasms. *Nat Rev Cancer* **17**, 425–440 (2017).

80. Zhao, A., Zhou, H., Yang, J., Li, M. & Niu, T. Epigenetic regulation in hematopoiesis and its implications in the targeted therapy of hematologic malignancies. *Sig Transduct Target Ther* **8**, 1–40 (2023).

81. Vincent T. DeVita, Theodore S. Lawrence, & Steven A. Rosenberg. Chapter 111: Myelodysplastic syndromes. in *DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology* (Lippincott Williams & Wilkins, 2015).

82. National Cancer Institute, DCCPS, Surveillance Research Program. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Mortality - All COD, Aggregated With State, Total U.S. (1969-2020). (2022).

83. Schiffer, C. A. & Stone, R. M. Morphologic Classification and Clinical and Laboratory Correlates. in *Holland-Frei Cancer Medicine. 6th edition* (BC Decker, 2003).

84. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e24 (2019).

85. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood* **140**, 1345–1377 (2022).

86. Löwenberg, B. Sense and nonsense of high-dose cytarabine for acute myeloid leukemia. *Blood* **121**, 26–28 (2013).

87. Crossnohere, N. L. *et al.* Side effects from acute myeloid leukemia treatment: results from a national survey. *Current Medical Research and Opinion* **35**, 1965–1970 (2019).

88. Li, Y. & Zhou, F. Efficacy of bone marrow transplantation in treating acute myeloid leukemia: a systematic review and meta-analysis. *Am J Transl Res* **15**, 1–12 (2023).

89. Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N Engl J Med* **373**, 1136–1152 (2015).

90. Makkar, H. *et al.* Acute myeloid leukemia: novel mutations and their clinical implications. *Am J Blood Res* **13**, 12–27 (2023).

91. DiNardo, C. D. & Cortes, J. E. Mutations in AML: prognostic and therapeutic implications. *Hematology Am Soc Hematol Educ Program* **2016**, 348–355 (2016).

92. Daver, N., Schlenk, R. F., Russell, N. H. & Levis, M. J. Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia* **33**, 299–312 (2019).

93. O'Donnell, M. R. *et al.* Acute Myeloid Leukemia, Version 3.2017, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* **15**, 926–957 (2017).

94. Grafone, T., Palmisano, M., Nicci, C. & Storti, S. An overview on the role of FLT3-tyrosine kinase receptor in acute myeloid leukemia: biology and treatment. *Oncol Rev* **6**, e8 (2012).

95. Smith, C. C. The growing landscape of FLT3 inhibition in AML. *Hematology* **2019**, 539–547 (2019).

96. Cortes, J. *et al.* Quizartinib, an FLT3 inhibitor, as monotherapy in patients with relapsed or refractory acute myeloid leukaemia: an open-label, multicentre, single-arm, phase 2 trial. *Lancet Oncol* **19**, 889–903 (2018).

97. Cortes, J. E. *et al.* Phase 2b study of 2 dosing regimens of quizartinib monotherapy in FLT3-ITD-mutated, relapsed or refractory AML. *Blood* **132**, 598–607 (2018).

98. Erba, H. P. *et al.* Quizartinib plus chemotherapy in newly diagnosed patients with FLT3-internal-tandem-duplication-positive acute myeloid leukaemia (QuANTUM-First): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet* **401**, 1571–1583 (2023).

99. Perl, A. E. *et al.* Gilteritinib or Chemotherapy for Relapsed or Refractory FLT3-Mutated AML. *N Engl J Med* **381**, 1728–1740 (2019).

100. Sato, T. *et al.* FLT3 ligand impedes the efficacy of FLT3 inhibitors in vitro and in vivo. *Blood* **117**, 3286–3293 (2011).

101. Yang, X., Sexauer, A. & Levis, M. Bone marrow stroma-mediated resistance to FLT3 inhibitors in FLT3-ITD AML is mediated by persistent activation of extracellular regulated kinase. *Br J Haematol* **164**, 61–72 (2014).

102. Dutreix, C., Munarini, F., Lorenzo, S., Roesel, J. & Wang, Y. Investigation into CYP3A4-mediated drug-drug interactions on midostaurin in healthy volunteers. *Cancer Chemother Pharmacol* **72**, 1223–1234 (2013).

103. von Bubnoff, N. *et al.* FMS-Like Tyrosine Kinase 3–Internal Tandem Duplication Tyrosine Kinase Inhibitors Display a Nonoverlapping Profile of Resistance Mutations In vitro. *Cancer Research* **69**, 3032–3041 (2009).

104. Piloto, O. *et al.* Prolonged exposure to FLT3 inhibitors leads to resistance via activation of parallel signaling pathways. *Blood* **109**, 1643–1652 (2007).

105. Kim, K.-T. *et al.* Constitutive Fms-like tyrosine kinase 3 activation results in specific changes in gene expression in myeloid leukaemic cells. *British Journal of Haematology* **138**, 603–615 (2007).

106. Basit, F., Andersson, M. & Hultquist, A. The Myc/Max/Mxd Network Is a Target of Mutated Flt3 Signaling in Hematopoietic Stem Cells in Flt3-ITD-Induced Myeloproliferative Disease. *Stem Cells International* **2018**, e3286949 (2018).

107. Tickenbrock, L. *et al.* Flt3 tandem duplication mutations cooperate with Wnt signaling in leukemic signal transduction. *Blood* **105**, 3699–3706 (2005).

108. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141**, 432–445 (2010).

109. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nat Rev Cancer* **8**, 976–990 (2008).

110. Dhanasekaran, R. *et al.* The MYC oncogene — the grand orchestrator of cancer growth and immune evasion. *Nat Rev Clin Oncol* **19**, 23–36 (2022).

111. Dalla-Favera, R. *et al.* Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proc Natl Acad Sci U S A* **79**, 7824–7827 (1982).

112. Gurel, B. *et al.* Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis. *Mod Pathol* **21**, 1156–1167 (2008).

113. Stock, C., Kager, L., Fink, F. M., Gadner, H. & Ambros, P. F. Chromosomal regions involved in the pathogenesis of osteosarcomas. *Genes Chromosomes Cancer* **28**, 329–336 (2000).

114. Zhang, X. *et al.* MYC-dependent downregulation of telomerase by FLT3 inhibitors is required for their therapeutic efficacy on acute myeloid leukemia. *Ann Hematol* **97**, 63–72 (2018).

115. Li, L. *et al.* SIRT1 activation by a c-MYC oncogenic network promotes the maintenance and drug resistance of human FLT3-ITD acute myeloid leukemia stem cells. *Cell Stem Cell* **15**, 431–446 (2014).

116. Park, H. J. *et al.* Therapeutic resistance to FLT3 inhibition is driven by a novel ATM/mTOR pathway regulating oxidative phosphorylation. 2022.02.14.480468 Preprint at https://doi.org/10.1101/2022.02.14.480468 (2022).

117. Massó-Vallés, D. & Soucek, L. Blocking Myc to Treat Cancer: Reflecting on Two Decades of Omomyc. *Cells* **9**, 883 (2020).

118. Wolf, E. & Eilers, M. Targeting MYC Proteins for Tumor Therapy. *Annual Review of Cancer Biology* **4**, 61–75 (2020).

119. Cascón, A. & Robledo, M. MAX and MYC: A Heritable Breakup. *Cancer Research* **72**, 3119–3124 (2012).

120. Winter, G. E. *et al.* DRUG DEVELOPMENT. Phthalimide conjugation as a strategy for in vivo target protein degradation. *Science* **348**, 1376–1381 (2015).

121. Koh, C. M., Sabò, A. & Guccione, E. Targeting MYC in cancer therapy: RNA processing offers new opportunities. *Bioessays* **38**, 266–275 (2016).

122. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515–520 (2018).

123. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).

124. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–589 (2010).

125. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat Genet* **47**, 8–12 (2015).

126. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev.* (2013) doi:10.1101/gad.232710.113.

127. Roe, J.-S., Mercan, F., Rivera, K., Pappin, D. J. & Vakoc, C. R. BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia. *Molecular Cell* **58**, 1028–1039 (2015).

128. Pelish, H. E. *et al.* Mediator kinase inhibition further activates super-enhancer-associated genes in AML. *Nature* **526**, 273–276 (2015).

129. Bhagwat, A. S., Lu, B. & Vakoc, C. R. Enhancer dysfunction in leukemia. *Blood* **131**, 1795–1804 (2018).

130. Zhao, Y. *et al.* High-Resolution Mapping of RNA Polymerases Identifies Mechanisms of Sensitivity and Resistance to BET Inhibitors in t(8;21) AML. *Cell Rep* **16**, 2003–2016 (2016).

131. Shorstova, T., Foulkes, W. D. & Witcher, M. Achieving clinical success with BET inhibitors as anti-cancer agents. *Br J Cancer* **124**, 1478–1490 (2021).

132. Fiskus, W. *et al.* Superior efficacy of co-targeting GFI1/KDM1A and BRD4 against AML and post-MPN secondary AML cells. *Blood Cancer J.* **11**, 1–16 (2021).

133. Pedicona, F. *et al.* Targeting the lysine-specific demethylase 1 rewires kinase networks and primes leukemia cells for kinase inhibitor treatment. *Science Signaling* **15**, eabl7989.

134. Smith, B. M. *et al.* PU.1 and MYC transcriptional network defines synergistic drug responses to KIT and LSD1 inhibition in acute myeloid leukemia. Preprint at https://doi.org/10.1101/2021.08.24.456354 (2021).

135. Amente, S. *et al.* LSD1-mediated demethylation of histone H3 lysine 4 triggers Myc-induced transcription. *Oncogene* **29**, 3691–3702 (2010).

136. Poole, C. J. & Van Riggelen, J. MYC—Master Regulator of the Cancer Epigenome and Transcriptome. *Genes* **8**, 142 (2017).

137. Shi, Y. *et al.* Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell* **119**, 941–953 (2004).

138. Kerenyi, M. A. *et al.* Histone demethylase Lsd1 represses hematopoietic stem and progenitor cell signatures during blood cell maturation. *eLife* **2**, e00633 (2013).

139. Metzger, E. *et al.* LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* **437**, 436–439 (2005).

140. Laurent, B. *et al.* A specific LSD1/KDM1A isoform regulates neuronal differentiation through H3K9 demethylation. *Mol Cell* **57**, 957–970 (2015).

141. Shi, Y.-J. *et al.* Regulation of LSD1 histone demethylase activity by its associated factors. *Mol Cell* **19**, 857–864 (2005).

142. Yang, M. *et al.* Structural Basis for CoREST-Dependent Demethylation of Nucleosomes by the Human LSD1 Histone Demethylase. *Molecular Cell* **23**, 377–387 (2006).

143. Song, Y. *et al.* Mechanism of Crosstalk between the LSD1 Demethylase and HDAC1 Deacetylase in the CoREST Complex. *Cell Rep* **30**, 2699-2711.e8 (2020).

144. Grimes, H. L., Chan, T. O., Zweidler-McKay, P. A., Tong, B. & Tsichlis, P. N. The Gfi-1 proto-oncoprotein contains a novel transcriptional repressor domain, SNAG, and inhibits G1 arrest induced by interleukin-2 withdrawal. *Mol Cell Biol* **16**, 6263–6272 (1996).

145. Saleque, S., Kim, J., Rooke, H. M. & Orkin, S. H. Epigenetic Regulation of Hematopoietic Differentiation by Gfi-1 and Gfi-1b Is Mediated by the Cofactors CoREST and LSD1. *Molecular Cell* **27**, 562–572 (2007).

146. van der Meer, L. T., Jansen, J. H. & van der Reijden, B. A. Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia* **24**, 1834–1843 (2010).

147. Lin, Y. *et al.* The SNAG domain of Snail1 functions as a molecular hook for recruiting lysine-specific demethylase 1. *EMBO J* **29**, 1803–1816 (2010).

148. Niebel, D. *et al.* Lysine-specific demethylase 1 (LSD1) in hematopoietic and lymphoid neoplasms. *Blood* **124**, 151–152 (2014).

149. Harris, W. J. *et al.* The histone demethylase KDM1A sustains the oncogenic potential of MLL-AF9 leukemia stem cells. *Cancer Cell* **21**, 473–487 (2012).

150. Cusan, M. *et al.* LSD1 inhibition exerts its antileukemic effect by recommissioning PU.1- and C/EBPα-dependent enhancers in AML. *Blood* **131**, 1730–1742 (2018).

151. Maiques-Diaz, A. *et al.* Enhancer Activation by Pharmacologic Displacement of LSD1 from GFI1 Induces Differentiation in Acute Myeloid Leukemia. *Cell Reports* **22**, 3641–3659 (2018).

152. Nerlov, C. & Graf, T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev* **12**, 2403–2412 (1998).

153. Braun, T. P. *et al.* Combined inhibition of JAK/STAT pathway and lysine-specific demethylase 1 as a therapeutic strategy in CSF3R/CEBPA mutant acute myeloid leukemia. *PNAS* **117**, 13670–13679 (2020).

154. Maes, T. *et al.* ORY-1001, a Potent and Selective Covalent KDM1A Inhibitor, for the Treatment of Acute Leukemia. *Cancer Cell* **33**, 495-511.e12 (2018).

155. Salamero, O. *et al.* First-in-Human Phase I Study of Iadademstat (ORY-1001): A First-in-Class Lysine-Specific Histone Demethylase 1A Inhibitor, in Relapsed or Refractory Acute Myeloid Leukemia. *JCO* JCO.19.03250 (2020) doi:10/ghf3fx.

156. Curtiss, B. M. *et al.* PU.1 and MYC transcriptional network defines synergistic drug responses to KIT and LSD1 inhibition in acute myeloid leukemia. *Leukemia* **36**, 1781–1793 (2022).

157. Marine, J.-C., Dawson, S.-J. & Dawson, M. A. *Non-Genetic Mechanisms of Therapeutic Resistance in Cancer*. *Nature Reviews Cancer* www.nature.com/nrc doi:10.1038/s41568-020-00302-4.

158. Bell, C. C. *et al.* Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia. *Nature Communications* **10**, 2723 (2019).

159. DiNardo, C. D. *et al.* Azacitidine and Venetoclax in Previously Untreated Acute Myeloid Leukemia. *N Engl J Med* **383**, 617–629 (2020).

160. Lasica, M. & Anderson, M. A. Review of Venetoclax in CLL, AML and Multiple Myeloma. *J Pers Med* **11**, 463 (2021).

161. Delbridge, A. R. D., Grabow, S., Strasser, A. & Vaux, D. L. Thirty years of BCL-2: translating cell death discoveries into novel cancer therapies. *Nat Rev Cancer* **16**, 99–109 (2016).

162. Letai, A. G. Diagnosing and exploiting cancer's addiction to blocks in apoptosis. *Nat Rev Cancer* **8**, 121–132 (2008).

163. Pei, S. *et al.* Monocytic Subclones Confer Resistance to Venetoclax-Based Therapy in Patients with Acute Myeloid Leukemia. *Cancer Discov* **10**, 536–551 (2020).

164. Waclawiczek, A. *et al.* Combinatorial BCL2 Family Expression in Acute Myeloid Leukemia Stem Cells Predicts Clinical Response to Azacitidine/Venetoclax. *Cancer Discovery* **13**, 1408–1427 (2023).

165. Yashar, W. M. *et al.* GoPeaks: histone modification peak calling for CUT&Tag. *Genome Biology* **23**, 144 (2022).

166. Liu, L., Jin, G. & Zhou, X. Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Research* **43**, 3873–3885 (2015).

167. Berger, S. L. Histone modifications in transcriptional regulation. *Current Opinion in Genetics & Development* **12**, 142–148 (2002).

168. Cedar, H. & Bergman, Y. Programming of DNA Methylation Patterns. *Annual Review of Biochemistry* **81**, 97–117 (2012).

169. Benveniste, D., Sonntag, H.-J., Sanguinetti, G. & Sproul, D. Transcription factor binding predicts histone modifications in human cell lines. *PNAS* **111**, 13367–13372 (2014).

170. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

171. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

172. Xu, Y. *et al.* TAF1 plays a critical role in AML1-ETO driven leukemogenesis. *Nat Commun* **10**, 4925 (2019).

173. Meers, M. P., Janssens, D. H. & Henikoff, S. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Molecular Cell* **75**, 562-575.e5 (2019).

174. Gusmao, E. G., Dieterich, C., Zenke, M. & Costa, I. G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* **30**, 3143–3151 (2014).

175. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2013).

176. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794–D801 (2018).

177. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

178. Khan, A. & Mathelier, A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics* **18**, 287 (2017).

179. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

180. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

181. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**, e1003118 (2013).

182. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag https://doi.org/10.1007/978-0-387-98141-3 (2009).

183. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).

184. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192 (2013).

185. Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* **19**, e46255 (2018).

186. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

187. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).

188. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21.29.1-21.29.9 (2015).

189. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

190. Yashar, W. *et al.* GoPeaks: Histone Modification Peak Calling for CUT&Tag (GitHub). https://doi.org/10.1101/2022.01.10.475735 (2022).

191. Yashar, W. *et al.* GoPeaks: Histone Modification Peak Calling for CUT&Tag (Zenodo). https://doi.org/10.1101/2022.01.10.475735 (2022).

192. Yashar, W. *et al.* GoPeaks Compare (GitHub). https://doi.org/10.1101/2022.01.10.475735 (2022).

193. Braun TP, Yashar WM, Kong G, VanCampen J, & Curtiss BM. GoPeaks: histone modification peak calling for CUT&Tag. *Gene Expression Omnibus* (2022) doi:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE190793.

194. Henikoff J. K562_K4me3_(20180919_HK_Hs_K5l_K4me3_0912). *Gene Expression Omnibus* (2019) doi:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3536518.

195. Henikoff J. K562_K4me1_(20180919_HK_Hs_K5l_K4me1_0912). *Gene Expression Omnibus* (2019) doi:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3536516.

196. Henikoff J. K562_H3K27me3_(20190104_SH_Hs_M30_20181224). *Gene Expression Omnibus* (2019) doi:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3560261.

197. Henikoff J. K562_IgG_(20180919_HK_Hs_K5_IgG_0912). *Gene Expression Omnibus* (2019) doi:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3560264.

198. Lee, D., Zhang, J., Liu, J. & Gerstein, M. Epigenome-based splicing prediction using a recurrent neural network. *PLoS Comput Biol* **16**, e1008006 (2020).

199. Farnham P. H3K4me3 ChIP-seq on human K562. *ENCODE* (2011) doi:https://doi.org/doi:10.17989%2FENCSR000EWA.

200. Bernstein B. H3K4me1 ChIP-seq on human K562. *ENCODE* (2011) doi:https://doi.org/doi:10.17989%2FENCSR000AKS.

201. Farnham P. H3K27me3 ChIP-seq on human K562. *ENCODE* (2011) doi:https://doi.org/doi:10.17989%2FENCSR000EWB.

202. Nimer S. Kasumi1-WT-H3K27ac_rep1. *ChIP-Atlas* (2019) doi:https://chip-atlas.org/view?id=SRX4143063.

203. Nimer S. Kasumi1-WT-H3K27ac_rep2. *ChIP-Atlas* (2019) doi:https://chip-atlas.org/view?id=SRX4143067.

204. Yashar, W. M. *et al.* Predicting transcription factor activity using prior biological information. *iScience* 109124 (2024) doi:10.1016/j.isci.2024.109124.

205. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685–D690 (2011).

206. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* **48**, D489–D497 (2020).

207. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol* **21**, 36 (2020).

208. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).

209. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology* **17**, e9923 (2021).

210. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690-705.e9 (2018).

211. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–10874 (2001).

212. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* **96**, 9212–9217 (1999).

213. Brenton, J. D., Carey, L. A., Ahmed, A. A. & Caldas, C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J Clin Oncol* **23**, 7350–7360 (2005).

214. Tamimi, R. M. *et al.* Comparison of molecular phenotypes of ductal carcinoma in situand invasive breast cancer. *Breast Cancer Res* **10**, 1–9 (2008).

215. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418–8423 (2003).

216. Kouros-Mehr, H. & Werb, Z. Candidate regulators of mammary branching morphogenesis identified by genome-wide transcript analysis. *Dev Dyn* **235**, 3404–3412 (2006).

217. Kouros-Mehr, H., Slorach, E. M., Sternlicht, M. D. & Werb, Z. GATA-3 Maintains the Differentiation of the Luminal Cell Fate in the Mammary Gland. *Cell* **127**, 1041–1055 (2006).

218. Sotiriou, C. *et al.* Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* **100**, 10393–10398 (2003).

219. Seachrist, D. D., Anstine, L. J. & Keri, R. A. FOXA1: A Pioneer of Nuclear Receptor Action in Breast Cancer. *Cancers (Basel)* **13**, 5205 (2021).

220. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).

221. Babur, Ö. *et al.* Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *PATTER* **0**, (2021).

222. Chen, X. *et al.* XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature* **508**, 103–107 (2014).

223. Green, A. R. *et al.* MYC functions are specific in biological subtypes of breast cancer and confers resistance to endocrine therapy in luminal tumours. *Br J Cancer* **114**, 917–928 (2016).

224. Nedeljković, M. & Damjanović, A. Mechanisms of Chemotherapy Resistance in Triple-Negative Breast Cancer—How We Can Rise to the Challenge. *Cells* **8**, 957 (2019).

225. Garcia-Alonso, L. *et al.* Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Research* **78**, 769–780 (2018).

226. Alessandrini, F., Pezzè, L., Menendez, D., Resnick, M. A. & Ciribilli, Y. ETV7-Mediated DNAJC15 Repression Leads to Doxorubicin Resistance in Breast Cancer Cells. *Neoplasia* **20**, 857–870 (2018).

227. Neel, D. S. & Bivona, T. G. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. *NPJ Precis Oncol* **1**, 3 (2017).

228. Matkar, S. *et al.* An Epigenetic Pathway Regulates Sensitivity of Breast Cancer Cells to HER2 Inhibition via FOXO/c-Myc Axis. *Cancer Cell* **28**, 472–485 (2015).

229. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

230. Bottomly, D. *et al.* Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer Cell* **40**, 850-864.e9 (2022).

231. Mihalyova, J. *et al.* Venetoclax: A new wave in hematooncology. *Experimental Hematology* **61**, 10–25 (2018).

232. Brown, F. *et al.* PRMT5 Inhibition Promotes FOXO1 Tumor Suppressor Activity to Drive a Pro-Apoptotic Program That Creates Vulnerability to Combination Treatment with Venetoclax in Mantle Cell Lymphoma. *Blood* **138**, 681 (2021).

233. Hawe, J. S. *et al.* Network reconstruction for trans acting genetic loci using multi-omics data and prior information. *Genome Med* **14**, 1–21 (2022).

234. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

235. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958–970 (2008).

236. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**, 27–33 (2011).

237. MacNeil, L. T. *et al.* Transcription Factor Activity Mapping of a Tissue-Specific In Vivo Gene Regulatory Network. *cels* **1**, 152–162 (2015).

238. Yashar, W. M. *et al.* Disruption of the MYC Super-Enhancer Complex by Dual Targeting of FLT3 and LSD1 in Acute Myeloid Leukemia. *Molecular Cancer Research* MCR-22-0745 (2023) doi:10.1158/1541-7786.MCR-22-0745.

239. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).

240. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

241. Demir, E. *et al.* BioPAX – A community standard for pathway data sharing. *Nat Biotechnol* **28**, 935–942 (2010).

242. Babur, Ö. *et al.* Pattern search in BioPAX models. *Bioinformatics* **30**, 139–140 (2014).

243. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404 (2012).

244. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783–784 (2014).

245. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Research* **42**, e168 (2014).

246. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

247. Kennedy, V. E. & Smith, C. C. FLT3 Mutations in Acute Myeloid Leukemia: Key Concepts and Emerging Controversies. *Front Oncol* **10**, (2020).

248. Herold, T. *et al.* Validation and refinement of the revised 2017 European LeukemiaNet genetic risk stratification of acute myeloid leukemia. *Leukemia* **34**, 3161–3172 (2020).

249. Stone, R. M. *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *New England Journal of Medicine* **377**, 454–464 (2017).

250. Perl, A. E. *et al.* Gilteritinib or Chemotherapy for Relapsed or Refractory FLT3-Mutated AML. *New England Journal of Medicine* **381**, 1728–1740 (2019).

251. Ohanian, M. *et al.* Clinical significance of MYC expression in acute myeloid leukemia. *JCO* **32**, 7094–7094 (2014).

252. Wang, C. *et al.* Alternative approaches to target Myc for cancer treatment. *Sig Transduct Target Ther* **6**, 1–14 (2021).

253. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev* **27**, 2648–2662 (2013).

254. Berthon, C. *et al.* Bromodomain inhibitor OTX015 in patients with acute leukaemia: a dose-escalation, phase 1 study. *Lancet Haematol* **3**, e186-195 (2016).

255. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Reports* **17**, 1193–1205 (2016).

256. Maiques-Diaz, A. *et al.* Enhancer Activation by Pharmacologic Displacement of LSD1 from GFI1 Induces Differentiation in Acute Myeloid Leukemia. *Cell Rep.* **22**, 3641–3659 (2018).

257. Shigeyasu, K. *et al.* The PVT1 lncRNA is a novel epigenetic enhancer of MYC, and a promising risk-stratification biomarker in colorectal cancer. *Mol Cancer* **19**, 155 (2020).

258. Hu, Z. *et al.* RUNX1 regulates corepressor interactions of PU.1. *Blood* **117**, 6498–6508 (2011).

259. Choudhary, C. *et al.* Activation mechanisms of STAT5 by oncogenic Flt3-ITD. *Blood* **110**, 370–374 (2007).

260. Price, D. H. Regulation of RNA Polymerase II Elongation by c-Myc. *Cell* **141**, 399–400 (2010).

261. Ariyoshi, K. *et al.* Constitutive Activation of STAT5 by a Point Mutation in the SH2 Domain*. *Journal of Biological Chemistry* **275**, 24407–24413 (2000).

262. Mallette, F. A., Gaumont-Leclerc, M.-F., Huot, G. & Ferbeyre, G. Myc Down-regulation as a Mechanism to Activate the Rb Pathway in STAT5A-induced Senescence. *Journal of Biological Chemistry* **282**, 34938–34944 (2007).

263. Pinz, S., Unser, S. & Rascle, A. Signal transducer and activator of transcription STAT5 is recruited to c-Myc super-enhancer. *BMC Molecular Biology* **17**, 10 (2016).

264. Hermeking, H. *et al.* Identification of CDK4 as a target of c-MYC. *Proceedings of the National Academy of Sciences* **97**, 2229–2234 (2000).

265. Thomas, D. & Majeti, R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood* **129**, 1577–1585 (2017).

266. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).

267. Chen, J. *et al.* Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level. *Nat Med* **25**, 103–110 (2019).

268. Jin, L. *et al.* Monoclonal antibody-mediated targeting of CD123, IL-3 receptor alpha chain, eliminates human acute myeloid leukemic stem cells. *Cell Stem Cell* **5**, 31–42 (2009).

269. Askmyr, M. *et al.* Selective killing of candidate AML stem cells by antibody targeting of IL1RAP. *Blood* **121**, 3709–3713 (2013).

270. Ågerstam, H. *et al.* Antibodies targeting human IL1RAP (IL1R3) show therapeutic effects in xenograft models of acute myeloid leukemia. *Proc Natl Acad Sci U S A* **112**, 10786–10791 (2015).

271. Ianevski, A., Giri, A. K. & Aittokallio, T. SynergyFinder 2.0: visual analytics of multi-drug combination synergies. *Nucleic Acids Research* **48**, W488–W493 (2020).

272. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193–1203 (2016).

273. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology* **109**, 21.29.1-21.29.9 (2015).

274. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

275. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications* **10**, 1–10 (2019).

276. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

277. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

278. Liu, N. *et al.* Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch. *Cell* **173**, 430-442.e17 (2018).

279. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

280. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

281. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

282. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**, 15545–15550 (2005).

283. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90–W97 (2016).

284. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

285. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).

286. Estabrook, J. *et al. Predicting Transcription Factor Activity Using Prior Biological Information*. http://biorxiv.org/lookup/doi/10.1101/2022.12.16.520295 (2022) doi:10.1101/2022.12.16.520295.

287. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

288. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

289. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

290. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**, W160–W165 (2016).

291. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495–501 (2010).

292. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nat Biotechnol* **29**, 24–26 (2011).

293. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

294. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

295. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

296. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

297. Babur, Ö. *et al.* Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns (N Y)* **2**, 100257 (2021).

298. Lederer, S., Dijkstra, T. M. H. & Heskes, T. Additive Dose Response Models: Defining Synergy. *Frontiers in Pharmacology* **10**, (2019).

299. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925–936 (2019).

300. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).

301. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403–411 (2021).

302. Yashar, W. M. *et al.* Differentiation State Plasticity As a Mechanism of BCL2 Inhibitor Resistance in Acute Myeloid Leukemia. *Blood* **142**, 2786 (2023).

303. Lechman, E. R. *et al.* miR-126 Regulates Distinct Self-Renewal Outcomes in Normal and Malignant Hematopoietic Stem Cells. *Cancer Cell* **29**, 214–228 (2016).

304. Council, N. R., Studies, D. on E. and L., Research, I. for L. A. & Animals, C. for the U. of the G. for the C. and U. of L. *Guide for the Care and Use of Laboratory Animals: Eighth Edition*. (National Academies Press, 2010).

305. Zhang, Q. *et al.* Dual Targeting of Mitochondrial Vulnerability Using Complex I Inhibitor Iacs-010759 with Bcl-2 Inhibitor Venetoclax and Azacitidine in Pre-Clinical Acute Myeloid Leukemia (AML) Models. *Blood* **136**, 13–14 (2020).

306. Wunderlich, M. *et al.* AML cells are differentially sensitive to chemotherapy treatment in a human xenograft model. *Blood* **121**, e90–e97 (2013).

307. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

308. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458–1465 (2019).

309. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).

310. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

311. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol* **39**, 825–835 (2021).

312. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).

313. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

314. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Rep* **38**, 110333 (2022).

315. Bock, C. *et al.* High-content CRISPR screening. *Nat Rev Methods Primers* **2**, 1–23 (2022).

316. Hayakawa, F. *et al.* Tandem-duplicated Flt3 constitutively activates STAT5 and MAP kinase and introduces autonomous cell growth in IL-3-dependent cell lines. *Oncogene* **19**, 624–631 (2000).

317. Takahashi, S., Harigae, H., Kaku, M., Sasaki, T. & Licht, J. D. Flt3 mutation activates p21WAF1/CIP1 gene expression through the action of STAT5. *Biochem Biophys Res Commun* **316**, 85–92 (2004).

318. Nosaka, T. *et al.* STAT5 as a molecular regulator of proliferation, differentiation and apoptosis in hematopoietic cells. *EMBO J* **18**, 4754–4765 (1999).

319. Mizuki, M. *et al.* Flt3 mutations from patients with acute myeloid leukemia induce transformation of 32D cells mediated by the Ras and STAT5 pathways. *Blood* **96**, 3907–3914 (2000).

320. Spiekermann, K., Bagrintseva, K., Schwab, R., Schmieja, K. & Hiddemann, W. Overexpression and constitutive activation of FLT3 induces STAT5 activation in primary acute myeloid leukemia blast cells. *Clin Cancer Res* **9**, 2140–2150 (2003).

321. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci U S A* **115**, E2467–E2476 (2018).

322. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).

323. Demchenko, A. P. Beyond annexin V: fluorescence response of cellular membranes to apoptosis. *Cytotechnology* **65**, 157–172 (2013).

324. Fernandes-Alnemri, T., Litwack, G. & Alnemri, E. S. CPP32, a novel human apoptotic protein with homology to Caenorhabditis elegans cell death protein Ced-3 and mammalian interleukin-1 beta-converting enzyme. *J Biol Chem* **269**, 30761–30764 (1994).

325. Merry, T. L. & Ristow, M. Nuclear factor erythroid-derived 2-like 2 (NFE2L2, Nrf2) mediates exercise-induced mitochondrial biogenesis and the anti-oxidant response in mice. *J Physiol* **594**, 5195–5207 (2016).

326. Hwang, J.-H. *et al.* TAZ links exercise to mitochondrial biogenesis via mitochondrial transcription factor A. *Nat Commun* **13**, 653 (2022).

327. Masterson, A. J. *et al.* MUTZ-3, a human cell line model for the cytokine-induced differentiation of dendritic cells from CD34+ precursors. *Blood* **100**, 701–703 (2002).

328. Boutzen, H. *et al.* A primary hierarchically organized patient-derived model enables in depth interrogation of stemness driven by the coding and non-coding genome. *Leukemia* **36**, 2690–2704 (2022).