Defining molecular vulnerabilities in childhood leukaemia through biological network analysis

Cosmin Tudose, MSci

20209354

This thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology.

UCD School of Medicine
Head of School: Professor Michael Keane

Supervisor: Professor Jonathan Bond
Co-supervisor: Professor Colm J. Ryan

Research Studies Panel: Professor Amanda McCann, Professor Desmond G. Higgins, Professor Christina Kiel

September 2024

Table of contents

List of Figures

List of Tables

Abstract

Cellular phenotype is largely governed by the regulation of gene expression that dictates critical biological processes such as cell differentiation, adaptation to stimuli, and metabolism. Gene regulation is typically altered in cancer, leading to dysregulation of networks involving multiple genes, transcription factors and genomic regulatory elements. Studying these gene regulatory networks (GRNs) can help us untangle the regulatory mechanisms that underpin disease biology and potentially identify new therapeutic targets. The aim of the research reported in this thesis was to analyse gene regulation in cancer in order to identify new vulnerabilities and understand mechanisms of gene regulation and resistance to current therapies.

Firstly, we took a systematic approach to evaluate the ability of GRNs to predict gene essentiality in cancer cell lines. We employed computational methods to infer the activity of regulatory genes in cell lines from ten different cancer types. We then tested the ability of GRN-inferred activity to predict the sensitivity of different cancer cell lines to gene inhibition, using genome-wide CRISPR screens from The Cancer Dependency Map. We found that while GRNs display some cancer specificity, GRN-inferred activity does not perform any better than gene expression at finding essential genes in any tumour type. Treating sensitivity to inhibition as a binary variable or assessing the ability of GRN-inferred activity to predict gene essentiality led to similar results, with gene expression performing better than inferred activity. Finally, stratifying GRNs by their size or number of unique targets did not improve predictions for GRN-inferred activity. Our results were concordant across multiple GRN sources and activity estimation methods.

Secondly, we took a focused approach to study genetic and epigenetic regulation in the blood cancer acute myeloid leukaemia (AML). To replicate PRC2 loss-of-function, which is present in 15% of paediatric AML and which is linked to chemoresistance, we used CRISPR/Cas9 editing to create isogenic cell line models of heterozygous *EZH2* loss. We found that *EZH2*+/- AML cells had altered gene expression, decreased genome-wide repressive chromatin marks, and notably had markedly increased chromatin accessibility. This altered regulatory landscape resulting from the depletion of a genome-wide epigenetic

transcriptional repressor led to partial activation of alternative lineage transcriptional programs, including overexpression of the fetal haematopoiesis gene *LIN28B.* The activation of a LIN28B-driven program included increased *CDK6* expression that correlated with decreased sensitivity to CDK6 inhibition in *EZH2*+/- cells. Interestingly, the 3D genome architecture was largely maintained upon *EZH2* depletion, with preferential retention and even gain of H3K27me3 at regions with high 3D contact frequency.

Overall, our work provides insights into approaches to studying gene regulation and applications of computational methods to understand this field. In addition, we provide a detailed characterisation of the complexities of genomic regulation in PRC2-depleted AML that has implications for understanding the aggressive disease biology in these cases.

Statement of Original Authorship

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

Thesis Format

This thesis was prepared as a “Collection of Papers” in accordance with the “Guidelines for Preparation, Submission, Examination and Dissemination of Research Degree Theses, Appendix 2: Production of Thesis Format”, available at: https://www.ucd.ie/graduatestudies/documentrepository/

The format has been discussed by the candidate with the supervisors and the Research Studies Panel.

Funding

This research was funded by Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214 and supported in part by the EU's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant H2020-MSCA-COFUND-2019-945385.

Experimental work and analysis was supported by additional funding (see Chapters 2 and 3 Funding)

List of Publications and manuscripts

Part of this thesis

- **Tudose, C.**, Bond, J. and Ryan, C.J., 2023. Gene essentiality in cancer is better predicted by mRNA abundance than by gene regulatory network-inferred activity. *NAR cancer*, *5*(4), p.zcad056. DOI: https://doi.org/10.1093/narcan/zcad056

Contributions: Performed all the computational data processing and analysis, interpreted the results, created the figures and drafted the article.

- **Tudose, C.**, Jones, L., Grosu, T.I., Fitzgerald, M.C., Maziak, N., Ling, R., Roy, A., Vazuerizas, J.M., Ryan, .C.J., Bond, J., EZH2 loss leads to priming and partial activation of alternative lineage transcriptional programs in acute myeloid leukaemia. *To be submitted for review*

Contributions: Performed all the computational data processing and analysis, interpreted the results, created the figures and drafted the article.

Not included in this thesis:

- Lefeivre, T., **Tudose, C.**, Grosu, T.I., Jones, L., León, T.E., Wynne, K., Oliviero, G., Smith, O.P., Trinquand, A., Mansour, M.R. and Ryan, C.J., 2023. Loss of Polycomb Repressive Complex 2 function causes asparaginase resistance in T-acute lymphoblastic leukemia through decreased WNT pathway activity. *bioRxiv*, pp.2023-08. DOI: https://doi.org/10.1101/2023.08.04.552014

Contributions: Analysed RNA-seq from cell lines and publicly available T-ALL patient data: performed differential expression analysis, GSEA and applied the PROGENy pipeline to find differentially responsive pathways.

- Grady, E., Biswas, S., Dias, T., McCarthy, P., **Tudose, C.**, Betts, D., Malone, A. and Bond, J., A novel TCF3::PIK3R1 fusion linked to decreased PI3K‐AKT signalling activity in paediatric B‐acute lymphoblastic leukaemia. *British Journal of Haematology*. DOI: https://doi.org/10.1111/bjh.19587

Contributions: Ran AlphaFold2 Multimer models of protein interactions between p85α (with and without the SH3 domain) and p110α (Figure 1F).

Acknowledgements

Firstly, I would like to thank my supervisors Prof Jonathan Bond and Prof Colm Ryan for all the support during my PhD. Thank you for being great supervisors, all the scientific input, all the advice, for always being available and for making me a researcher. It has been a great four years and I have learnt a lot from you.

I would like to thank SFI and the EU MSCA for the funding that made this project possible. Also thanks to the CRT Genomics Data Science for creating this incredible program.

Thanks to my research studies panel: Prof Amanda McCann, Prof Des Higgins and Prof Christina Kiel for the support and overseeing my progress.

Mulțumesc părinților mei pentru susținerea din toate punctele de vedere, întrebările constante despre știință și dedicarea pe care au avut-o în creșterea mea. Multumiri fratelui meu, Ionuț pentru ca e cel mai bun frate.

Big thanks to Nico for always being supportive and making me happier. I am very grateful for you always being here.

Thanks to my friends all over the globe for always being here, despite the distance: Anca, Ariff, Catalin, Ellie, Léa, Lucy, Mihael, Robert.

Thanks to all the past and present Bond and Ryan group members for the scientific discussions and the time spent together: Alanah, Amélie, Anjan, Barbara, Ciardha, Claire, Hamda, Inas, Krishn, Luke, Metin, Narod, Olivier, Peter, Sharmila, Sophie, Sutanu, Swathi, Tânia, Teerna, Theodora, Thomas. Thanks to Luke for all the support during my PhD and for always listening and answering my stupid questions. Thanks to Theodora for helping me not forget Romanian in SBI and all the endless epigenetics chats. Also thanks to everyone in SBI. Thanks to Luis for being able to answer any ML or GRN question.

Thanks to the Vaquerizas lab for making me feel welcome during my stay in London: Christos, Irina, Jahnavi, Lars, Maria, Mel, Noura, Sara, Srishti.

Big thanks to the CRT cohort 2 - made great memories on the Kylemore Abbey retreat and the nights out in Galway. Special thanks to my friends Anjan, Louise, Marina and Padraic (honorary cohort P) for the quality time spent together in and outside UCD.

I would also like to acknowledge all the dogs I would see during my walk to SBI - they made my early mornings happier.

List of Abbreviations

acronyms will be spelled out when first used in the text, but all are provided here for reference

2-HG	=	2-hydroxyglutarate
3C	=	Chromatin conformation capture
4DN	=	4D nucleome project
α-KG	=	α-ketoglutarate
ABC	=	Atlas of human blood
AEBP2	=	Adipocyte enhancer-binding protein 2
AFDN	=	Afadin
AL	=	Acute leukaemia
ALL	=	Acute lymphoblastic leukaemia
AML	=	Acute myeloid leukaemia
AP-1	=	Activator protein 1
ATAC-seq	=	Assay for Transposase-Accessible Chromatin with HTS
AUC ROC	=	Area under the receiver operating characteristic curve
AZU1	=	Azurocidin
B-ALL	=	B-acute lymphoblastic leukaemia
BCOR	=	BCL6 corepressor
bHLH	=	basic helix-loop-helix
BLCA	=	Bladder carcinoma
BMI1	=	B Lymphoma Mo-MLV Insertion Region 1 Homolog
BRCA	=	Breast carcinoma
bZIP	=	leucine zipper
Cas9	=	CRISPR associated protein 9
CCLE	=	Cancer cell line encyclopaedia
CD3/4/11b/14/34	=	Cluster of differentiation 3/4/11b/14/34
CDCA7	=	Cell division cycle associated 7
CDK4/6	=	Cyclin-dependent kinase 4/6
cDNA	=	Complementary DNA
CEBPA	=	CCAAT/enhancer-binding protein-alpha

CGC	=	Cancer gene census
CGI	=	CpG island
ChiP-seq	=	Chromatin immunoprecipitation with sequencing
CITE-seq sequencing	=	cellular indexing of transcriptomes and epitopes with
CL	=	Chronic leukaemia
CLES	=	Common language effect size
CML	=	Chronic myelogenous leukaemia
CMP	=	Common myeloid progenitor
CNA	=	Copy number alteration
COAD	=	Colon adenocarcinoma
COMPASS	=	Complex proteins associated with SET1
CPM	=	Counts per million
CPTAC	=	Clinical Proteomic Tumour Analysis Consortium
CTCF	=	CCCTC-binding factor
CTNNB1	=	Beta-catenin gene
CRE	=	Cis-regulatory element
CRISPR	=	Clustered regularly interspaced short palindromic repeats
CUT&RUN	=	Cleavage under targets and release using nuclease
CUT&TAG	=	Cleavage under targets and tagmentation
DBD	=	DNA-binding domain
DEGs	=	Differentially expressed genes
DepMap	=	The Cancer Dependency Map
DLBCL	=	Diffuse large B-cell lymphoma
DNA	=	Deoxyribonucleic acid
DNMT	=	DNA methyltransferase
DNMT1/3A	=	DNA methyltransferase 1/3A
DREAM Methods	=	Dialogue on Reverse Engineering Assessment and
DP	=	Double positive
EED	=	Embryonic ectoderm development
ENCODE	=	Encyclopedia of DNA Elements

EPOP Associated Protein	=	Elongin BC And Polycomb Repressive Complex 2
E-P	=	Enhancer-promoter loop
ETP	=	Early thymic progenitor
ETS1 Homolog-1	=	Avian erythroblastosis virus E26 (V-Ets) Oncogene
EZH1/2	=	Enhancer of zeste homologue 1/2
FAB	=	French-American-British
FDR	=	False discovery rate
FL	=	Fetal liver
FOXA1	=	Forkhead box A1
GATA1/2/3/6	=	GATA binding protein 1/2/3/6
GBM	=	Glioblastoma
GDSC	=	Genomics of Drug Sensitivity in Cancer
GMP	=	Granulocyte–monocyte progenitor
GO	=	Gene ontology
GP:BP	=	Gene ontology:Biological processes
GOF	=	Gain of function
GRN	=	Gene regulatory network
gRNA	=	guide-RNA
GSEA	=	Gene set enrichment analysis
H2A/2B/3/4	=	Histone 2A/2B/3/4
H2AK119Ub	=	Ubiquitylation of lysine 119 of histone H2A
H2BK120Ub	=	Ubiquitylation of lysine 120 of histone H2B
H3K4ac	=	Acetylation of lysine 4 of histone 3
H3K9ac	=	Acetylation of lysine 9 of histone 3
H3K14ac	=	Acetylation of lysine 14 of histone 3
H3K27ac	=	Acetylation of lysine 27 of histone 3
H3K27me	=	Methylation of lysine 27 of histone 3
H3K27me1/2/3	=	Mono-, di-, and tri-methylation of lysine 27 of histone 3
H3K36me3	=	Tri-methylation of lysine 36 of histone 3
H3K4me1/3	=	Mono-, tri-methylation of lysine 4 of histone 3
HAT	=	Histone acetylase

HDAC	=	Histone deacetylase
HDAC1/2/3	=	Histone deacetylase 1/2/3
HER2	=	Herstatin
Hg38	=	Human genome assembly 38
hMDP/cMoP monocytic progenitor	=	human monocytic-dendritic progenitor/common
HNSC	=	Head and neck squamous cell carcinoma
Hox	=	Homeotic genes
HOXA	=	Homeobox A cluster
HSPC	=	Haematopoietic stem and progenitor cell
HSC	=	Haematopoietic stem cell
IDH1/2	=	Isocitrate Dehydrogenase (NADP(+)) ½
IgG	=	Immunoglobulin G
ISP	=	Immature single-positive
ITGAM	=	Integrin Subunit Alpha M
ITGAV	=	Integrin Subunit Alpha V
HTS	=	High-throughput sequencing
JARID1A/2	=	Jumonji And AT-Rich Interaction Domain Containing 1A/2
JMML	=	Juvenile myelomonocytic leukaemia
KDM1A	=	Lysine demethylase 1A
KDM2B	=	Lysine demethylase 2B
KIRC	=	Kidney renal clear cell carcinoma
KMT2A/B/C/D/E	=	Lysine methyl-transferase 2 A/B/C/D/E
KMT2Ar	=	KMT2A rearrangement
KO	=	Knock-out
LCK	=	Lymphocyte cell-specific protein-tyrosine kinase
LMO2	=	LIM domain only 2
LOF	=	Loss-of-function
LT-HSC	=	Long-term haematopoietic stem cell
LUAD	=	Lung adenocarcinoma
m6A	=	N6-Methyladenosine
MAP2K1	=	Mitogen-Activated Protein Kinase Kinase 1
MAPK	=	Mitogen-activated protein kinase

MDS	=	Myelodysplastic syndrome
MECOM	=	MDS1 and EVI1 complex locus
MEF2C	=	Myocyte-specific enhancer factor 2C
MEIS1	=	Meis Homeobox 1
MEN1	=	Menin
MEP	=	Megakaryocyte–erythrocyte progenitor
miRNA	=	MicroRNA
MLL	=	Mixed-lineage leukaemia
MLLT3/9	=	Myeloid/Lymphoid Or Mixed-Lineage Leukemia Translocated To Chromosome 3/9
MLP	=	Multilymphoid progenitor
MOR	=	Mode of regulation
MPAL	=	Mixed phenotype acute leukaemia
MPO	=	Myeloperoxidase
MPP	=	Multipotent progenitor
mRNA	=	Messenger RNA
MTF2	=	Metal response element binding transcription factor 2
mtscATAC-seq	=	ATAC-seq combined with clone tracking via mitochondrial sequencing
NetBID	=	Network-based Bayesian inference of driver
NES	=	Normalised enrichment score
NFkB cells	=	Nuclear factor kappa-light-chain-enhancer of activated B
NK	=	Natural killer
NMI	=	Non-methylated island
NREP	=	Neuronal regeneration related protein
NUP98	=	Nucleoporin 98
NuRD	=	Nucleosome remodelling and deacetylase
NSD1	=	Nuclear Receptor Binding SET Domain Protein 1
OR	=	Odds ratio
PAAD	=	Pancreatic adenocarcinoma
PAD	=	Polycomb associated domain
PBMC	=	Peripheral blood mononuclear cell

PCA	=	Principal component analysis
PCDH0	=	Protocadherin-9
PcG	=	Polycomb group
PCGF	=	Polycomb group ring finger
PCGF1/2/3/4/5/6	=	Polycomb group ring finger 1/2/3/4/5/6
PCL1/2/3	=	Polycomb-Like Protein 1/2/3
PedDep	=	Paediatric cancer dependency map
Ph+	=	Philadelphia chromosome positive
PHF23	=	PHD Finger Protein 23
PPI	=	Protein-protein interaction
PRC1/2/2.1/2.2	=	Polycomb Repressive Complex 1/2/2.1/2.2
cPRC1	=	canonical PRC1
vPRC1	=	variant PRC1
Pre-B	=	Precursor B-cell
Pre-T	=	Precursor T-cell
Pro-B	=	Progenitor B-cell
Pro-T	=	Progenitor T-cell
PRO-seq	=	Precision nuclear run-on sequencing
PTK2	=	Protein Tyrosine Kinase 2
PTM	=	Post translational modification
QC	=	Quality control
RAB7B	=	Ras-related protein Rab-7
RB1	=	RB transcriptional corepressor 1
RBBP4/7	=	Retinoblastoma binding protein 4/7
RING1A/1B	=	Ring Finger Protein 1A/1B
RNA	=	Ribonucleic acid
RNAi	=	RNA interference
RPKM	=	Read per kilobase million
RUNX1/2	=	RUNX family transcription factor ½
S100A8/9/12	=	S100 Calcium Binding Protein A8/9/12
SEC	=	super elongation complex
shRNA	=	Short-hairpin RNA

SILAC	=	Stable isotope labelling by amino acids in cell culture
SKIDA1	=	SKI/DACH Domain Containing 1
SMARCA2/4	=	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 2/4
SMC1A/3	=	Structural maintenance of chromosomes 1A/3
SNAP	=	SILAC nucleosome affinity purification
SNV	=	Single nucleotide variant
SOX4	=	SRY-Box transcription factor 4
SP	=	Single positive
STAD	=	Stomach adenocarcinoma
STAG1/2	=	Stromal antigen ½
STAT3/5	=	Signal transducer and activator of transcription 3/5
ST-HSC	=	Short-term haematopoietic stem cell
SWI/SNF	=	Switch/sucrose non-fermentable
tRNA	=	Transfer ribonucleic acid
TAD	=	Topologically associated domain
TAL2	=	T-cell acute lymphocytic leukaemia 2 protein
T-ALL/LBL	=	T-acute lymphoblastic leukaemia/lymphoma
TARGET	=	Therapeutically applicable research to generate effective treatments
TBP	=	TATA-binding protein
TCGA	=	The Cancer Genome Atlas
TCR	=	T-cell receptor
TEAD1	=	TEA domain transcription factor 1
TET1/2/3	=	Ten-eleven-translocation 1/2/3
TF	=	Transcription factor
TFBS	=	Transcription factor binding site
TKI	=	Tyrosine kinase inhibitor
TLX1	=	T cell leukaemia homeobox 1
TMM	=	Trimmed means of M values
Tn5	=	Transposase 5
TP53	=	Tumour protein P53
TPM	=	Transcripts per million

TSS	=	Transcriptional start site
ULM	=	Univariate linear model
VCAN	=	Versican core protein
VIPER	=	Virtual inference of protein-activity by enriched regulon
W. Mean	=	Weighted Mean
W. Sum	=	Weighted Sum
WAPL	=	Wings apart-like protein
WES	=	Whole-exome sequencing
WGS	=	Whole-genome sequencing
WHO	=	World Health Organisation
WNT	=	Wingless and Int-1
ZNF43	=	Zinc finger protein 43

CHAPTER 1 - Introduction

Overview

Leukaemia is the most common paediatric malignancy and is characterised by the uncontrolled proliferation of haematopoietic progenitor cells with arrested differentiation. Thanks to improved standards of care, survival rates have improved in the past 50 years. Unfortunately, treatments still lack precision, leading to side effects and chemoresistance. Additionally, there are leukaemia subtypes with poorer prognosis that are more difficult to treat. With advances in high-throughput sequencing (HTS) and computational approaches, molecular characterisation of leukaemia has led to important discoveries about the biological processes that underpin blood cancer development. Molecular characterisation of cancer in general can help us understand the underlying mechanisms that govern aggressive disease. Additionally, it can aid us in finding new targets within the altered molecular programs that we can exploit via targeted therapies.

Epigenetics plays a major role in how cells develop and differentiate by controlling gene regulation. As haematopoietic differentiation is dictated by development stage-specific gene regulatory programs that are subverted in leukaemia, this introduction will first discuss gene regulation and the role of epigenetics in transcription. In the second and third sections, I will discuss high-throughput methods for molecular profiling, and will provide an overview of computational approaches to analyse their outputs. Finally, I will explain the major role epigenetic alterations have in leukaemia development, focusing on a particular epigenetic complex: polycomb repressive complex 2 (PRC2).

Section 1 - Gene regulation

1.1.1 Background

Gene regulation is the sum of processes that govern which genes are active in a cell. Early discoveries (Jacob & Monod 1961) demonstrated that environmental conditions could dictate transcriptional regulation in *Escherichia coli*. As a result of multicellularity, eukaryotes have evolved regulatory mechanisms that enable formation of different cell types, despite the fact that all cells contain the same DNA. Therefore, cell types are determined by which genes are “active” in a cell (Britten & Davidson 1969). Gene expression is regulated at all levels from DNA transcription to mRNA and mRNA translation to protein, ultimately affecting the phenotype (Figure 1.1). In this thesis I will focus on transcriptional regulation, which comprises the factors that influence the production of mRNA from DNA.

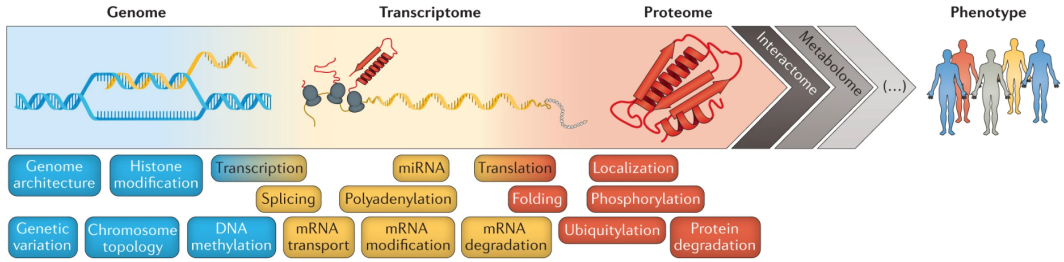


Figure 1.1 | Summary of processes that constitute layers of gene regulation from DNA sequence to phenotype. Figure from Buccitelli & Selbach (2020).

1.1.2 Transcriptional regulation

Transcriptional regulation controls the amount of mRNA that is produced in the cell. This flow of information is regulated at two levels. Firstly, regulation is determined by specific DNA sequences, which display affinity for the binding of proteins that initiate transcription, namely RNA polymerase, transcription factors (TFs) and co-factors (Summarised in Figure 1.2). These DNA sequences are known as cis-regulatory elements (CREs) and include promoters, enhancers and insulators (Jacob et al. 1964; Moreau et al. 1981; Peifer & Bender 1986). Secondly, the packaging of the DNA affects gene regulation independent of DNA sequence, altering the accessibility of genes and CREs (Lee & Young

2013) (Summarised in Figure 1.2). This can hinder or promote the binding of trans-acting regulators, including transcriptional machinery and TFs.

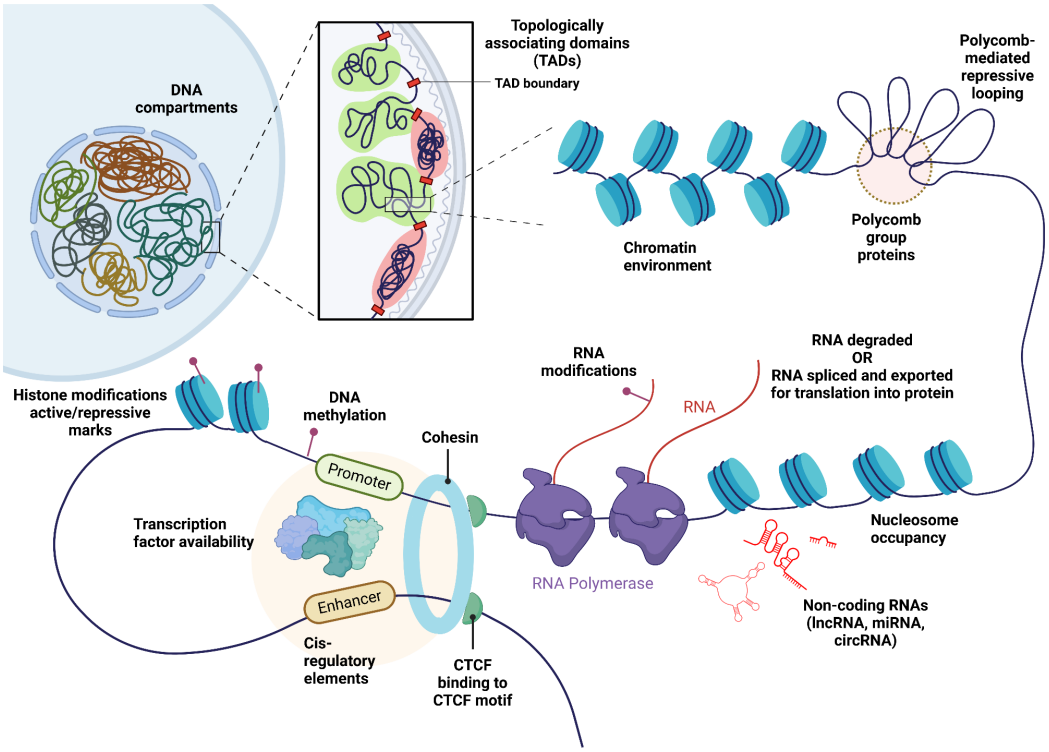


Figure 1.2 | Gene regulation at the DNA level focusing on chromatin and DNA-DNA interactions - Adapted from Nicolas et al. (2017); Lenstra et al. (2016); Cramer (2019); Matharu & Ahituv (2015); Zheng & Xie (2019). Created with Biorender.com

1.1.3 Chromatin - the packaging of DNA

Each human cell contains ~2m of DNA from end-to-end. All this DNA is packaged as chromatin inside the cells' nuclei, which is only ~6µm in diameter. Chromatin represents the structure containing DNA packaged by proteins (Kornberg 1974). The first layer of packaging is represented by the nucleosome, which has ~146bp of DNA wrapped around it (Figure 1.2) (Luger et al. 1997). The nucleosome is an octamer composed of two of each of the core histones H2A, H2B, H3 and H4, which can themselves undergo modifications that are associated with gene regulation (see subsection 1.1.10). The linker histone H1 is not part of the nucleosome, but stabilises the chromatin fibre by binding linker DNA between nucleosomes (Robinson & Rhodes 2006).

1.1.4 Cis-regulatory elements (CREs)

Promoters are found upstream of genes and determine where the transcriptional machinery assembles to initiate transcription. The precise region where mRNA production starts is called a transcriptional start site (TSS). The assembly of transcriptional machinery starts at the TATA box, found 30 base pairs from the TSS and equivalent to the prokaryotic Pribnow box (Pribnow 1975; Lifton et al. 1978). The general transcription factor TATA-binding protein (TBP) recognizes the TATA-box and recruits other proteins, and ultimately RNA polymerase II, which initiates RNA synthesis.

Enhancers are DNA sequences that can increase transcription. The first described enhancer was a 72 base pair sequence from the SV-40 virus that enhanced transcription of genes (Moreau et al. 1981; Banerji et al. 1981). More specifically, the two groups found independently that introduction of the viral enhancer sequence into cells enhanced the expression of T-antigen (Moreau et al. 1981), and the rabbit beta-globin gene (Banerji et al. 1981), respectively. Enhancers can regulate the transcription of proximal and distal genes, and enhancer activity can be regulated by a number of other factors (see subsections 1.1.8, 1.1.9 and 1.1.11). Like promoters, enhancers contain TF binding sites (TFBSs).

1.1.5 Transcription factors

Transcription factors (TFs) regulate gene expression by binding specific DNA motifs in enhancer or promoter regions, facilitating or impeding the binding of RNA polymerase. All TFs contain at least one DNA-binding domain (DBD), such as zinc finger, leucine zipper (bZIP) or basic helix-loop-helix (bHLH) domains (Siggers & Gordân 2013). DBDs have affinities for different DNA sequences, determining the sequences to which TFs can bind. For example, CTCF is a TF that contains 11 zinc-finger domains and binds a DNA sequence containing three regularly spaced repeats containing the core motif CCCTC (Lobanenkov et al. 1990). The GATA family of TFs also have zinc finger domains and bind the (T/A)GATA(A/G) sequence (Lowry & Atchley 2000). GATA1, GATA2 and GATA3 predominantly play roles in haematopoiesis (Simon 1995).

TFs activate specialised transcriptional programs involved in developmental patterning (Spitz & Furlong 2012), cell differentiation (Lee & Young 2013), signalling cascades (Weidemüller et al. 2021) or immune responses (Singh et al. 2014). Since a single TF can regulate hundreds of genes, some target genes will show strong co-expression patterns. Additionally, an assumption that is often made is that TFs linearly regulate their target genes, leading to a high correlation between TF expression and the expression of its target genes. Although this is true in specific cases, the correlation between TFs mRNA and the mRNA of their respective targets has been shown to be weak in a systematic analysis (Zaborowski & Walther 2020). Additionally, TF binding depends on the state of the chromatin at target loci. The accessibility to these loci can be hindered or enhanced by epigenetic marks, chromatin compaction, or abundance of pioneer factors.

1.1.6 Pioneer factors

Pioneer factors are TFs that can directly and independently bind condensed chromatin (Zaret & Carroll 2011). Pioneer factors are involved in the first steps leading to cell differentiation by de-repressing lineage-specific genes. For example the forkhead box (Fox) transcription factors are structurally similar to histone H1, but can displace H1 from DNA due to their higher DNA-binding affinity (Iwafuchi-Doi et al. 2016). Other pioneer factors bind to DNA to initiate transcription by recruiting cofactors such as chromatin remodelers (see subsection 1.3.6). For example the AP-1 pioneer complex recruits the chromatin remodelling complex SWI/SNF (Wolf et al. 2022). Furthermore, after mRNA is produced, it can also undergo post-transcriptional modifications such as polyadenylation, m6A methylation, pseudouridylation (Delaunay et al. 2023; Helm & Motorin 2017). These epitranscriptomic changes can have diverse effects on RNA function, splicing, degradation, and transport.

1.1.7 Chromatin as a layer of regulation

Chromatin state impacts whether a gene is transcribed or not: a compacted state (heterochromatin) is associated with gene repression, whilst euchromatin is transcriptionally active (Huisinga et al. 2006). Broadly, euchromatin and heterochromatin correspond respectively with A (active) and B (inactive) compartments identified by chromosome conformation capture analyses (Lieberman-Aiden et al. 2009). Heterochromatin is divided into two major types: constitutive and facultative heterochromatin. Typically, constitutive heterochromatin is found at pericentromeres and repetitive DNA elements and is conserved across cell types. On the contrary, facultative heterochromatin is present at silenced genes, and vastly differs between cell types and differentiation stages. The genome is further hierarchically compartmentalised in the nucleus into A (active) and B (inactive) compartments. The large compartments are further divided into topologically associated domains (TADs). TADs are self-interacting chromatin segments < 1Mb. TADs come into shape as a result of loop extrusion - DNA slides through the cohesin complex until a CTCF-bound region stops the process, creating a TAD boundary (Dixon et al. 2012; De Wit et al. 2015; Fudenberg et al. 2017). The function of TADs is still largely unknown and is an important area of further research. It has been hypothesised that TADs limit enhancer-promoter loops within their boundaries, and therefore control gene expression. Multiple studies have confirmed that the genome exhibits more intra- than inter-TAD contacts and that housekeeping genes are enriched at TAD-borders (Dixon et al. 2012; Rao et al. 2014). However, the location of TAD boundaries does not directly couple with gene expression (Ghavi-Helm et al. 2019).

1.1.8 DNA looping

TADs contain several loop domains, mediating DNA-DNA contacts (Rowley & Corces 2018; Bonev & Cavalli 2016). Loop domains are, in most cases, mediated by CCCTC-binding factor (CTCF) and formed by a loop extrusion process, through cohesin and its accessory proteins (Fudenberg et al. 2016). The cohesin complex is formed by core members SMC1A, SMC3, RAD21 and

STAG1/2 and accessory proteins WAPL and PDS5A/B (Peters et al. 2008). Loops can either be activating, via promoter-enhancer contacts, or repressive, where the enhancer-promoter contact is prevented (Figure 1.2, Figure 1.3). Additionally, high-order chromatin structures can form independently of cohesin and CTCF, such as R-loops (Petermann et al. 2022) and polycomb-mediated loops (Figure 1.3) (Eagen et al. 2017; Ogiyama et al. 2018).

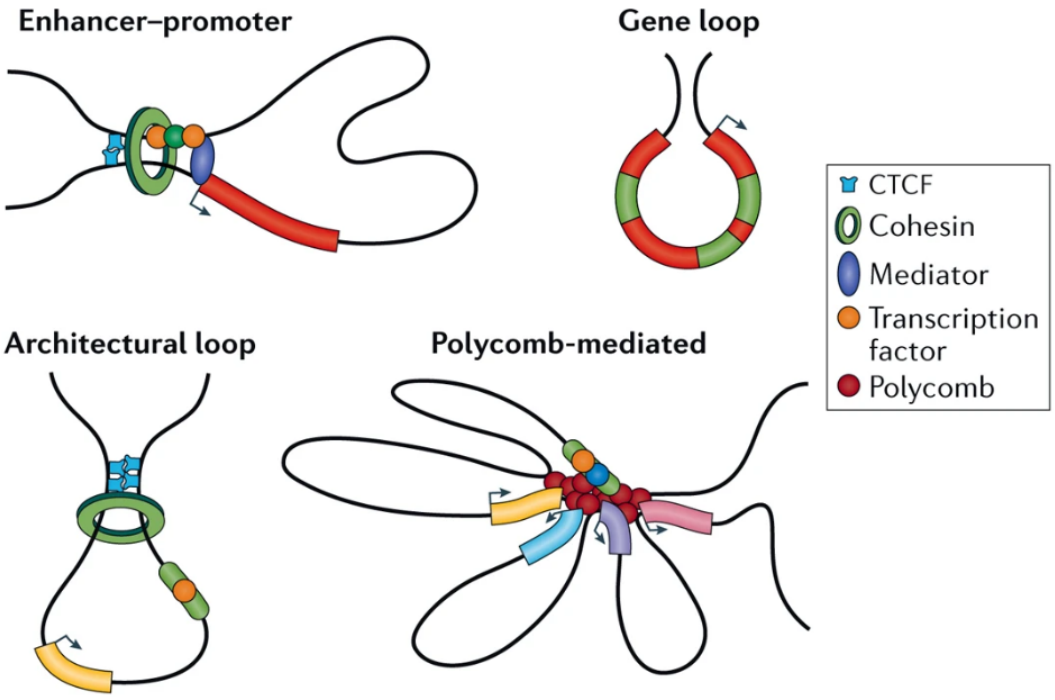


Figure 1.3 | Types of loops and the proteins that mediate the DNA-DNA contacts Adapted from Bonev & Cavalli (2016).

1.1.9 DNA methylation

Epigenetic marks act at the DNA level or at DNA-associated proteins to activate or inhibit transcription. Epigenetic marks can be placed on the DNA directly by DNA methyltransferases (DNMTs) at CpG islands (CGIs). CGIs are ~1,000 bp in length and are rich in G-C base pairs and are associated with 70% of gene promoters (Saxonov et al. 2006; Deaton & Bird 2011). CpG methylation at gene promoters leads to repression of gene transcription (Moore et al. 2013). Furthermore, methylation can also inhibit enhancer-mediated gene activation (Angeloni & Bogdanovic 2019). Non-methylated islands (NMIs) can be active or can be repressed by mutually exclusive mechanisms (see subsection 1.1.14).

1.1.10 Histone post-translational modifications

Another layer of epigenetic regulation occurs through chemical modification of histones. DNA is compacted by nucleosomes (composed of histones H2A, H2B, H3 and H4) and linker histone H1 to form chromatin. The histone proteins can undergo post-translational modifications (PTMs) which are associated with different levels of gene activity. It has been hypothesised that the pattern of histone modifications is part of a “histone code”, similar to the genetic code given by DNA codons corresponding to specific amino-acid tRNAs (Strahl & Allis 2000). The histone PTMs are altered by epigenetic factors, which can be split into three categories: writers (add PTMs), readers (bind to histone PTMs and interpret the code), and erasers (remove PTMs) (Figure 1.4A, B). Epigenetic factors do not act individually, but in protein complexes. Therefore, complexes can act in all three fashions, leading to crosstalk between histone modifications (i.e., if a histone PTM is recognised by the reader component of an epigenetic complex another PTM can be added or removed by the writer or eraser component, respectively) (Lee et al. 2010). For example, the protein complex COMPASS recognizes ubiquitylation of histone H2B at lysine residue 120 (H2BK120Ub), stimulating its methyltransferase component SET1 to deposit methyl groups on H3K4 (Kim et al. 2013).

Additionally, many chromatin regulators have multiple binding domains, suggesting the “histone code” is complex and combinatorial (Eustermann et al. 2011; Ruthenburg et al. 2011). By using SILAC (stable isotope labelling by amino acids in cell culture) nucleosome affinity purification (SNAP), Lukauskas et al. (2024) describe part of the “histone code”. Their approach finds proteins that are recruited or preferentially excluded by histone marks or combinations of histone marks by testing 55 different modified dinucleosomes, showing the complexity of histone PTM co-occurrences and mutual exclusivities. For example, they found that the INO80 complex recognizes a unique multivalent nucleosome signature formed by H2A.Z histone variant, H4ac and H3ac. Additionally, this study was able to confirm already known protein-protein interactions that have a role in histone PTM reading or writing (Lukauskas et al. 2024).

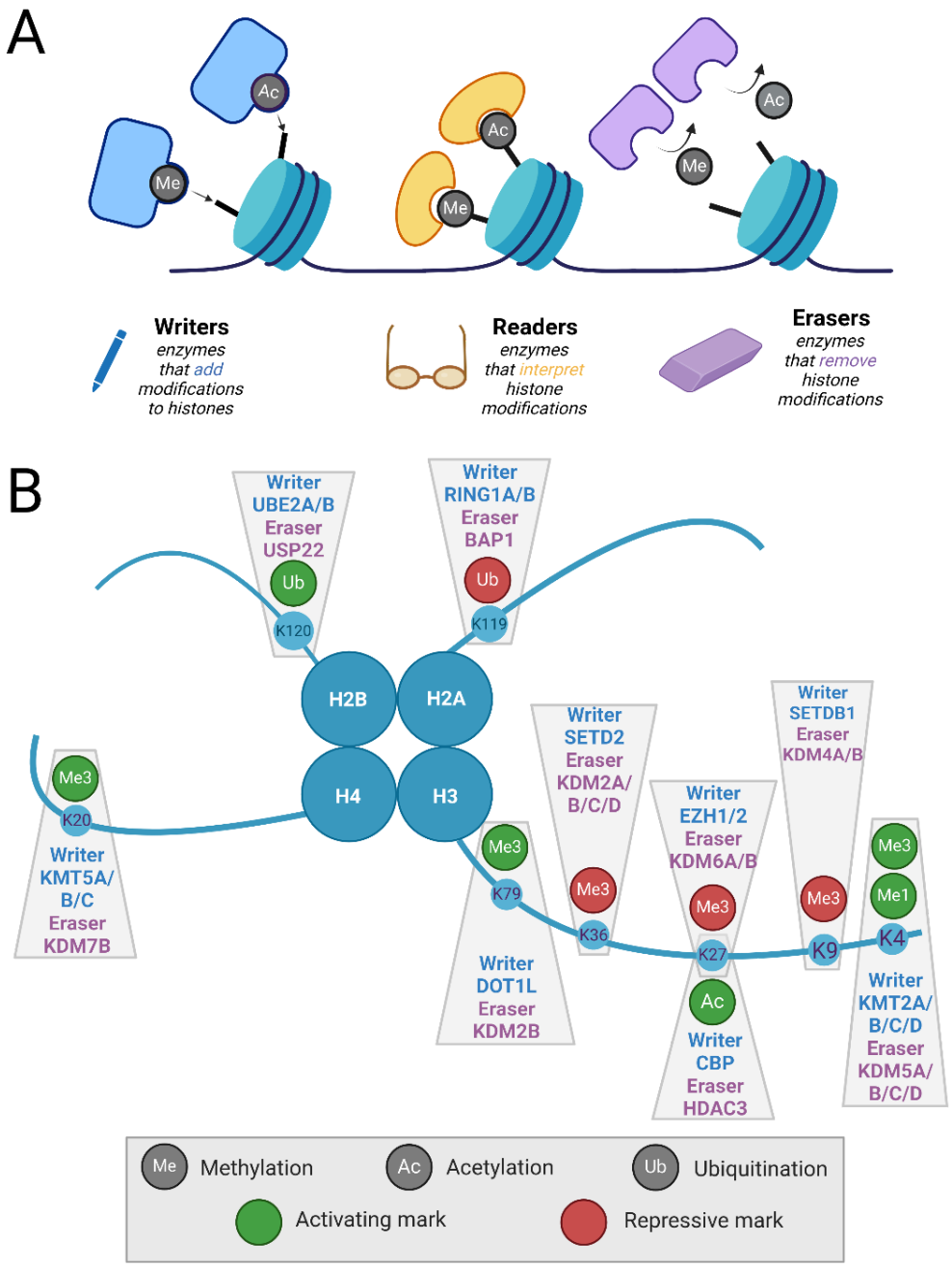


Figure 1.4 | Histone modifiers. A, Types of enzymes catalysing histone PTMs. **B,** Examples of important histone PTMs and the enzymes that write (in blue) and erase (in purple) each histone PTM. Created with Biorender.com.

1.1.11 Histone PTMs are associated with gene expression

Acetylation marks at lysine 27 or lysine 9 of H3 (H3K27ac and H3K9ac) are associated with gene or enhancer activity at marked loci. Monomethylation of

H3 at lysine 4 (H3K4me1) is also a mark associated with active and primed enhancers. However, experiments in mouse cells show that the marks are dispensable for enhancer activity, suggesting that presence of the epigenetic factors (CBP, and KMT2B/C, respectively) at these loci is directly responsible for enhancer activity (Dorighi et al. 2017; T. Zhang et al. 2020). Neither H3K27ac nor H3K4me1 appear to increase protein binding to enhancers, suggesting they may just have the role of preventing repressive factors from binding active regions (Lukauskas et al. 2024).

Trimethylation of H3 at lysine 4 (H3K4me3) is also associated with active transcription, but it is not necessary for most transcription to occur. The precise role of H3K4me3 is not clearly understood, as it promotes gene activation in a context-dependent manner (Cano-Rodriguez et al. 2016). In T-cell acute lymphoblastic leukaemia (T-ALL), H3K4me3 covering large domains is associated with driver oncogenes and T-ALL-specific essential genes (Belhocine et al. 2022). Additionally, H3K4me3 presence is mutually exclusive with CGI methylation (Hughes et al. 2020). Further, H3K4me3 preferentially excludes recruitment of PRC2 at nucleosomes (Lukauskas et al. 2024). Together, these suggest H3K4me3 acts as a “protector” of active genes from actively being repressed.

H3K36me3 is strongly correlated with active transcription due to the ability of the catalysing methyltransferase SETD2 to promote elongation by RNA polymerase II (Kizer et al. 2005; Millán-Zambrano et al. 2022). Additionally, H3K36me3 hinders polycomb protein EED from binding to the nucleosome and inhibiting gene expression at H3K36me3 marked loci. As a result, H3K36me3 and trimethylation of H3 at lysine 27 (H3K27me3) are mutually exclusive marks (Finogenova et al. 2020).

1.1.12 Polycomb proteins

Polycomb group proteins (PcGs) were initially discovered in *Drosophila melanogaster,* where they act as repressors of *Hox* genes, therefore fulfilling an essential role in homeotic gene expression and body plan specification (Kassis et al. 2017; Blackledge & Klose 2021). PcGs are conserved throughout the

animal kingdom, including in mammals, where they are involved in early development, cell renewal and lineage commitment through the repression of non lineage-specific genes (Piunti & Shilatifard 2021; Li et al. 2024). PcGs operate as part of complexes which have subfunction-dependent dynamic compositions. Of these, the Polycomb Repressive Complexes 1 and 2 (PRC1 and PRC2 respectively) are the best described (Figure 1.5A, B).

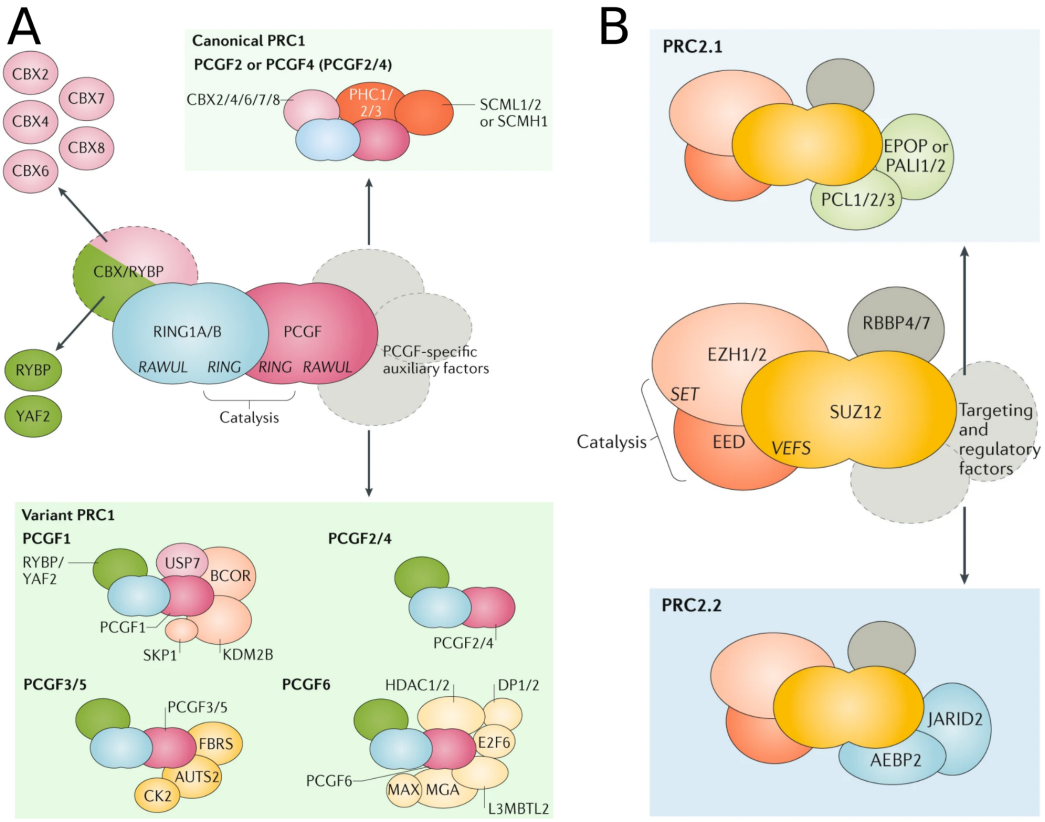


Figure 1.5 | Possible configurations of polycomb complexes. A, PRC1. **B,** PRC2. From Blackledge & Klose (2021).

1.1.13 PRC1

Polycomb repressive complex 1 (PRC1) always includes a RING-domain protein (RING1A/B) which mediates deposition of ubiquitin on histone H2A lysine 119 (Figure 1.5A, 6A, 6B). Additionally, PRC1 can incorporate one of six distinct paralogues PCGF1-6, which define the composition of PRC1 as following: canonical PRC1 (cPRC1) incorporates PCGF2/4, leading to recruitment of CBX proteins that can bind to H3K27me3. The other PCGF

paralogues can assemble variant PRC1 (vPRC1) and can also promote transcriptional silencing through PRC2-independent mechanisms (Blackledge & Klose 2021). vPRC1 is also critical for PRC2-dependent silencing (Fursova et al. 2019). By placing H2AK119Ub upstream of PRC2 and cPRC1 activity, vPRC1 initiates a cascade of repressive marks at polycomb loci (Figure 1.6A) (Bsteh et al. 2023; Moussa et al. 2019). PcGs typically act at NMIs (Figure 1.6A, B) (Wu et al. 2013).

1.1.14 PRC2

PRC2 has four core components: EZH1 or its paralogue EZH2, SUZ12, EED and RBBP4 or RBBP7 (Figure 1.5B). These core components are essential for the catalytic activity of the complex: depositing mono-, di- and tri- methylation on H3 at lysine 27 (H3K27me1/2/3) (Margueron et al. 2008; Shen et al. 2008) (Figure 1.6A, B). Additionally, there are accessory proteins which can form PRC2.1 (PCL1-3, EPOP, PALI1/2) or PRC2.2 (JARID2 and AEBP2) (Figure 1.5B). PRC2.1 is primarily involved in binding NMIs via PCL2 (also known at MTF2), whilst PRC2.2 starts the feedback loop with PRC1 following JARID2 binding to H2AK119Ub placed by vPRC1 (Healy et al. 2019; Glancy et al. 2023). EZH1/EZH2 catalyses the methylation reaction at H3K27, resulting in gene repression. EED interacts with H3K27, leading to allosteric activation and a “read-and-write” mechanism, with subsequent methylation of H3K27 at neighbouring nucleosomes (Hansen et al. 2008; Margueron et al. 2009; Lee et al. 2018) (Figure 1.6A, B). A mechanism of autostimulation has been proposed, whereby an EZH1/EZH2 autostimulation loop acts *in trans* as an allosteric activator for a neighbouring PRC2 complex, thereby activating it and causing further spreading H3K27me3 (Sauer et al. 2023).

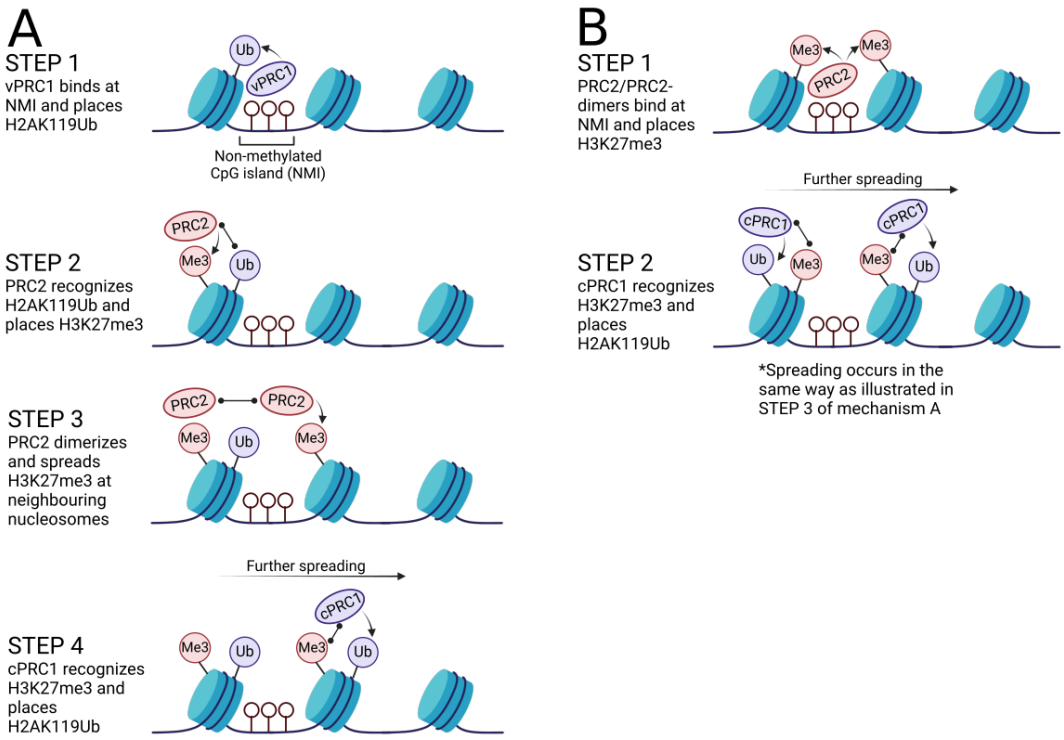


Figure 1.6 | Mechanism of polycomb repression at non-methylated CpG islands. A, Via NMI recognition by vPRC1. **B,** Via NMI recognition by PRC2. Ub = H2AK119Ub; Me3 = H3K27me3. (Created with Biorender.com)

EZH2 is essential for embryonic development, but EZH1 is dispensable, pointing towards functional differences between the two paralogues (O'Carroll et al. 2001). While EZH2 exhibits higher activity as a methyltransferase, EZH1 has higher affinity to chromatin binding and dimerizes more effectively than EZH2, suggesting it may be more effective at direct chromatin compaction (Margueron et al. 2008; Grau et al. 2021). Additionally, EZH2 is more highly expressed in earlier progenitor cells, decreasing in expression with differentiation stage, while EZH1 slightly increases in expression at later stages of differentiation (Margueron et al. 2008; Lee et al. 2022).

1.1.15 Polycomb-mediated looping

In addition to their roles as repressors of chromatin via H2AK119Ub and H3K27me3, PRC1 and PRC2 are able to form largely repressed regions, known as Polycomb bodies or Polycomb Associated Domains (PADs) (Schoenfelder et

al. 2015; Blackledge & Klose 2021; Boyle et al. 2020). These structures are seen as accumulations of polycomb proteins in confocal microscopy experiments, with these interactions being involved in the transcriptional repression of these regions and forming heterochromatin at pericentromeres (Saurin et al. 1998; Satijn et al. 1997). Interestingly, H3K27me3 is not necessary for PAD maintenance, but cPRC1 components RING1B and PHC proteins are essential for PAD integrity (Boyle et al. 2020; Bonev et al. 2017). However, the role of H3K27me3 may be to function as anchors for PADs and recruitment of cPRC1, as it is required for the initial establishment and re-establishment of PADs (Du et al. 2020).

1.1.16 Regulation after transcription

It is important to keep in mind that RNAs can be processed at later stages, undergoing post-transcriptional changes. Additionally, RNA can also be degraded and undergo nonsense-mediated decay. Even after mRNA translation, the resulting proteins can undergo post-translational changes or undergo degradation themselves. All these contribute to a modest mRNA-protein correlation in human cells, affected by both biological and technical factors (Zhang et al. 2016; Upadhya & Ryan 2022).

Section 2 - Functional and molecular profiling of cancer cells

1.2.1 Background

The discovery of targetable cancer type-specific alterations led to the proposition that molecular profiling of all tumours may provide a treatment solution for all cancers. For example, the discovery that the BRAF V600E mutation is present in more than >60% of melanomas led to work on inhibitors to target the mutated protein (Davies et al. 2002). Vemurafenib, a selective inhibitor of $BRAF^{V600E}$ was developed and approved for clinical use in 2011 (Yang et al. 2010; Kim & Cohen 2016).

Measuring gene expression was also identified at an early stage as a potential source of markers to characterise cancer subtypes and guide therapeutic development. For example, in 1999, the monoclonal antibody trastuzumab was approved as the first targeted therapy for breast cancers that overexpress the *HER2* gene (encoding herstatin) (Dillman 1999). With the explosion in high-throughput -omics technologies, we are now capable of performing detailed molecular profiling of tumours via HTS.

1.2.2 Whole-genome sequencing

We can use HTS to analyse the entire DNA sequence of a cancer cell through whole-genome sequencing (WGS). HTS can also be targeted to the coding portion of DNA by whole-exome sequencing (WES), which is typically performed at greater sequencing depth than WGS. WGS/WES can uncover single nucleotide variants (SNVs), insertions or deletions (indels) in protein-coding genes that may lead to truncated proteins (nonsense mutation) or an amino-acid change (missense) that significantly alters the protein folding leading to a dysfunctional (loss of function mutation - LOF) or enhanced activity protein (gain of function mutation - GOF). WGS data can also identify non-coding sequence variation in promoter sequences and CREs that play a role in cancer. Furthermore, genome sequencing can uncover copy number alterations (CNA), which represent either lost or gained copies of DNA that may

alter the amount of mRNA and protein in the cell for the lost/gained gene. Translocations are another alteration that are commonly seen in cancer, where a fragment of the genome breaks and is attached at a different genomic location. Translocations typically occur when two separate chromosome breaks happen to occur at the same time and are in close proximity within the nuclear space (Ramsden & Nussenzweig 2021). Translocations can lead to fusion genes containing two different gene sequences that ultimately produce chimeric proteins. The fusion protein is under the control of a tissue specific promoter or enhancer, resulting in aberrant expression. For example, the t(10;14)(q24;q11) results in the *TLX1* (*HOX11*) oncogene being aberrantly activated by the *TRA/D* enhancer in paediatric and adult T-ALL (Ferrando et al. 2002). Finally, mutations, translocations and CNA can also affect CREs, resulting in silencing or enhancement of neighbouring genes. CRE somatic mutations can also result in oncogenic activity, by activating oncogenes, such as *LMO2* and *TAL2* in T-ALL (Mansour et al. 2014; Rahman et al. 2017). In fact, non-coding mutations are increasingly recognised to be of high importance in T-ALL (O'Connor et al. 2023; Pölönen et al. 2024)

1.2.3 RNA-seq

RNA sequencing (RNA-seq) makes possible the quantification of gene products (i.e., transcripts) by sequencing mRNA or other types of non-coding RNA in a sample (Figure 1.7A) (Bainbridge et al. 2006). Comparing across RNA-seq samples can reveal changes in gene expression or isoform usage. RNA-seq is also commonly used to detect fusion genes (Heyer et al. 2019) and has more recently been employed to identify protein-coding mutations and CNAs (Umeda et al. 2024).

The rapid increase in WGS and RNA-seq data being produced has led to efforts such as The Cancer Genome Atlas (TCGA) or Therapeutically Applicable Research to Generate Effective Treatments (TARGET) (TARGET 2022; Cancer Genome Atlas Research Network 2008) which provide genome/exome sequences and RNA-seq profiles for thousands of tumours.

1.2.4 Functional characterisation

Furthermore, there are efforts to phenotypically characterise tumours or tumour cell lines, with a primary aim of identifying cancer-specific molecular vulnerabilities that might be exploited for targeted therapies. The Genomics of Drug Sensitivity in Cancer (GDSC) project generated molecular profiles for >1000 cell lines (Yang et al. 2013). The same cell lines were further characterised with drug screens (i.e., treating the cells with a wide-range of anti-cancer drugs and measuring cell viability) facilitating the identification of associations between molecular features and drug sensitivity. The same approach can be taken for *ex vivo* drug screening. For example, the first phase of the Biomarker-Based Treatment of AML (BEAT AML) study characterised 409 patient samples by WES, RNA-seq and drug screening (Burd et al. 2020).

CRISPR screening is another method commonly employed for the functional characterisation of tumour cell lines using HTS. CRISPR screens exploit the CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats and the Cas9 protein) genome editing technique (Jinek et al. 2012). For CRISPR editing, typically a guide RNA (gRNA) matching the targeted sequence is used to direct the Cas9 protein to its cutting locus (Barrangou & Doudna 2016). In a CRISPR screen, a library of gRNAs is introduced into a cell line, with every gRNA targeting a different genomic region (Bock et al. 2022). With a stable expression of the Cas9 enzyme, each gRNA results in a perturbation in the pool of cells. Upon selective pressure, the cells with genes essential for survival affected by perturbation are depleted. After a defined experimental period, gRNAs are quantified using HTS and essential genes are identified. The Cancer Dependency Map (DepMap) is an effort to characterise > 1,000 cancer cell lines using genome-wide CRISPR screening (Dempster et al. 2021; Dempster et al. 2019; Meyers et al. 2017; Pacini et al. 2021). Additionally the DepMap contains extensive molecular characterisation of the cell lines used for CRISPR screening, such as gene expression, proteomics, genomic data and drug screen information. Correspondingly, the paediatric cancer dependency map (PedDep) has performed CRISPR screens on 82 cell lines across 13 paediatric solid and brain tumour types (Dharia et al. 2021).

1.2.5 Epigenomics assays

Currently, there are more than 350 described HTS methods (according to Enseqlopedia, https://enseqlopedia.com/), many overlapping in approach, but tailored to answer different biological questions. However, I will focus on a few approaches that relate to gene regulation and more specifically the epigenome, taking into consideration the importance of these methods in profiling cancer.

Epigenomic assays profile epigenetic changes genome-wide and focus primarily on DNA methylation, histone modifications, DNA binding proteins and chromatin conformation.

The assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a method to profile accessible chromatin regions (Figure 1.7B). The Tn5 transposase binds open chromatin, which is then sequenced and aligned against the reference genome to map open chromatin, closed chromatin and nucleosome positioning.

Chromatin immunoprecipitation with sequencing (ChIP-seq) uses an antibody to bind histone modifications or DNA-binding proteins and precipitate the bound DNA regions, thereby mapping chromatin factor binding and/or chromatin marks across the genome. (Figure 1.7C). Cleavage Under Targets and Tagmentation (CUT&TAG) and cleavage under targets and release using nuclease (CUT&RUN) are newer, more efficient methods similar to ChIP-seq.

Chromatin conformation capture methods (3C, 4C, Hi-C, Micro-C and region capture Micro-C) entail crosslinking of DNA fragments that are spatially close and define DNA-DNA contacts and genomic domains within the spatial organisation of the genome within the nucleus (Figure 1.7D).

The development of these assays has made significant contributions to functional characterisation efforts. For example, the Encyclopedia of DNA Elements (ENCODE) consortium is an international effort to map the functional elements of the human genome: CREs, DNA methylation, histone PTMs, TFs, chromatin structure, etc (ENCODE Project Consortium 2012; Luo et al. 2020). Additionally, the 4D Nucleome (4DN) Project tries to address a limitation of ENCODE or other similar efforts: how do CREs exert their regulatory effects on the genome? By combining chromatin conformation capture methods and genetic and biophysical perturbation experiments, the 4DN project aims to

characterise the structure of the genome and link it to its function and role in biological processes related to development and disease (Dekker et al. 2017).

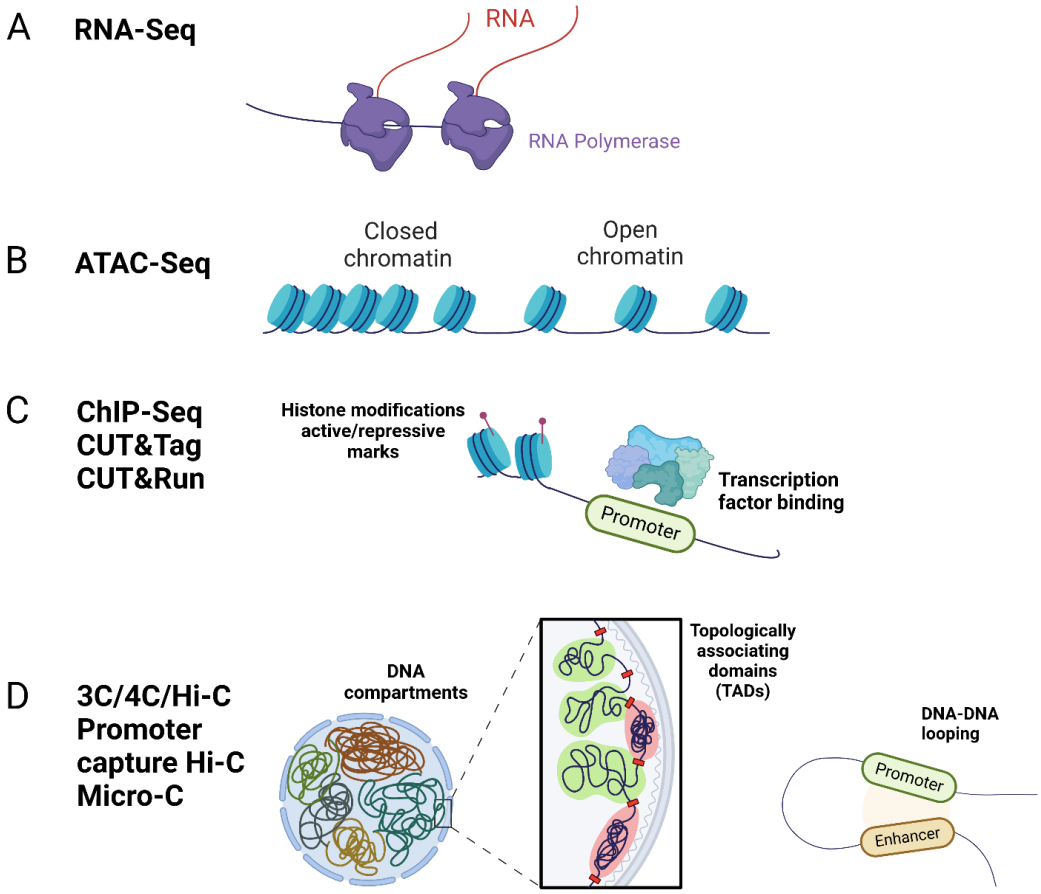


Figure 1.7 | Examples of epigenomic assays and information that can be extracted from each. A, RNA-seq. **B,** ATAC-seq. **C,** Methods for finding distribution of chromatin-associated proteins and histone PTMs. **D,** Chromosome conformation capture. Created with Biorender.com.

Whilst the systematic and functional profiling of tumours and tumour cell lines has led to numerous discoveries and improvements in cancer treatment, we have not been able to find a way to treat all cancer subtypes yet. Challenges remain due to undruggable targets, acquired resistance, tumour heterogeneity and complex biology which we do not fully understand with our current tools. An area that still faces many challenges is how do we best integrate the large amounts of information from all the profiling methods to make the most of it (i.e., find new drug targets, understand mechanisms of resistance, etc.)? This

question will be explored in the next part and will be further detailed in a leukaemia context in Part 4.

Section 3 - Network and integrative biology

1.3.1 Background

Cells function and respond to stimuli via the coordinated action of their molecular components (i.e., DNA, RNA, proteins, metabolites) (Chasman et al. 2016). Network biology studies interactions between biological entities, e.g. physical interactions between proteins within cells. It is part of a broader field, network science, that uses common computational and mathematical approaches to describe and analyse a variety of network types (e.g. computer networks, social networks, citation networks). Entities within networks are referred to as 'nodes', while the relationships between them are termed 'edges' (Barabási & Oltvai 2004). This implies that a given abnormality in one of the nodes can “spread” throughout the network to nodes that are not defective themselves. Numerous biological questions such as comparing biological conditions (i.e, mapping interactions and comparing the nodes and the relationships between the nodes between two conditions) (Ideker & Krogan 2012), interpreting genetic variants (Wong et al. 2021), or prioritising genes for experiments can be addressed with biological networks (Jiang 2015). Numerous methods have been developed to reconstruct biological networks that represent the relationships between nodes to better understand cellular and molecular mechanisms (Koutrouli et al. 2020; Lefebvre et al. 2012).

Network-based approaches have applications in both basic and translational biology (Barabási et al. 2010). The advancements in omics technologies have led to the generation of an unprecedented amount of information on the molecular landscape of cancer. These large amounts of data have led to both opportunities and challenges to develop network biology approaches that can accurately describe molecular mechanisms involved in cancer. For example, due to its large number of profiled samples (~11,000 patients across over 30 types of cancer) the TCGA is often employed for network-based analysis (Paull et al. 2021; Alvarez et al. 2016). Methods have

been developed to systematically integrate high-throughput omics data to understand disease, find drug resistance mechanisms and find novel genetic associations (Chasman et al. 2016).

1.3.2 Types of biological networks

Biological information can be summarised and interpreted through networks. Different data modalities will produce different types of networks (Summarised in Table 1). These networks, however, are not independent from each other and form a network of networks, that taken all together may describe a cell's behaviour (Barabási & Oltvai 2004). For example, a kinase may phosphorylate a TF, resulting in altered transcription of target genes that have accessible promoter binding sites. For the purpose of this thesis, I will primarily focus on transcriptional/gene regulatory networks.

1.3.3 Gene regulatory networks

Gene regulatory networks (GRNs) describe the regulation of gene products (i.e., mRNA). In a gene regulatory network the nodes are represented by source genes (regulators) and target genes, which form the regulon of a source gene (Table 1). Typically, the regulators are TFs. The edges represent the regulatory process (i.e., activation or inhibition) and these can be weighted or unweighted. The weights on edges determine the strength of interaction. Biologically speaking, the effect a single TF has on a target gene can vary, owing to a number of factors such as: the need of cofactors for the target gene's activation, chromatin architecture, RNA polymerase II stalling, presence of RNA degrading machinery (i.e., the rixosome complex).

Table 1.1 | Types of biological networks and their description.

Type of biological networks	Components (Nodes and edges)	Example omics used to extract information about network
Protein-protein interaction (PPI) networks	● Nodes = Proteins, ● Edges = Physical interactions (typically undirected)	Proteomics (Immunoprecipitation -Mass Spectrometry)
Transcriptional networks/gene regulatory networks	● Nodes = TFs, gene, regulatory RNAs or CREs ● Edges = directional or non-directional relationships	RNA-seq, ChIP-seq, ATAC-seq, Hi-C, PRO-seq
Metabolism networks	● Nodes = Enzymes and metabolites ● Edges = reactions	Metabolomics
Cell signalling networks	● Nodes = Cell receptors, ligands, kinases ● Edges = flow of information	Proteomics, Phosphoproteomics
Genetic networks	● Nodes = Genes ● Edges = Functional relationships	WGS, CRISPR screens
Cell-cell communication networks	● Nodes = cells ● Edges = intercellular flow of signal	scRNA-seq, Spatial omics

1.3.4 Computational approaches for GRN construction

Gene regulatory networks precede modern high-throughput computational biology, with Britten & Davidson (1969) describing theoretical models of how genes may be regulated in eukaryotes. They hypothesised the existence of gene batteries made of four classes of gene components: an integrator, a producer, a receptor site, and a sensor site. The sensor site receives a signal (e.g., from a hormone), which is transferred to the integrator, which in turn creates a product (RNA or protein). This product then directly activates the receptor site to activate transcription. Additionally, they proposed that gene batteries can be simultaneously regulated or overlapping and some genes can be part of hundreds of batteries (Britten & Davidson 1969). This becomes

particularly relevant in development, where previously inactive programs become simultaneously activated.

In the late 1990s the first expression microarray experiments were performed (Schena 1996). Multiple research groups have attempted to exploit the large amounts of microarray expression data produced and use computational approaches to reconstruct GRNs (Hughes et al. 2000; Husmeier 2003; Margolin et al. 2006).

1.3.5 GRN inference from RNA-seq

Nowadays, RNA-seq has superseded the gene expression microarray as the method of choice for transcriptomic profiling in biological research and multiple methods to infer GRNs from RNA-seq data have been developed, such as ARACNe, KBoost, GENIE3, etc. (Margolin et al. 2006; Iglesias-Martinez et al. 2021; Huynh-Thu et al. 2010). These methods use statistical modelling to find transcriptomic associations that define the GRN's architecture. GRNs are typically composed of three components: regulatory genes, their targets and the weights between.

The DREAM (Dialogue on Reverse Engineering Assessment and Methods) challenges are open collaborative computational biology competitions that have attempted to rigorously assess GRN inference methods. To do this the participating algorithms were benchmarked against synthetic gene expression data inferred from *in silico* gene networks in a double-blind fashion (i.e., the organisers were blind to the inference methods and the participants were blind to the networks) (Marbach et al. 2010; Marbach et al. 2016). Additionally, the algorithms were benchmarked against networks from *E. coli, Saccharomyces cerevisiae, and Staphylococcus aureus.* This is an important limitation because bacterial and yeast GRNs have been extensively validated, whilst a gold-standard human GRN does not exist yet (Salgado et al. 2006; Salgado et al. 2023).

Some GRN inference methods use prior data, such as TF binding information or SNPs associated with disease from genome-wide association studies (GWAS) (Dey et al. 2022; Aibar et al. 2017). For example, SCENIC first

uses gene expression correlations to construct networks, then cleans the networks against a database of TFBSs, pruning off interactions that are not present in the database (Aibar et al. 2017). SCENIC+ and GRaNIE take this approach one step further and also link CREs (inferred from chromatin accessibility or chromatin interactions data) to the TF-target interactions (Bravo González-Blas et al. 2023; Kamal et al. 2023).

1.3.6 Integration approaches for epigenomic assays

Integrating epigenetic information from HTS-based assays (see Section 2) with genomics can provide a full picture of how genes are regulated during normal development or when development is inhibited by an oncogenic event. We know that cells can change both their transcriptomic and chromatin accessibility landscape during development and oncogenic changes can alter chromatin accessibility. This was demonstrated in haematopoiesis by performing RNA-seq and ATAC-seq in healthy bone marrow and acute myeloid leukaemia (AML) and defining cell-type specific transcriptomic and chromatin accessibility profiles (Corces et al. 2016).

Integrating epigenetic information can provide mechanistic explanations for drug resistance. For example, rhabdoid tumours with mutations in components of the chromatin remodelling complex SWI/SNF are dependent on EZH2. Normally, SWI/SNF has the ability to evict PRC2 from nucleosomes and activate genes, but this function is altered in rhabdoid tumours, leading to genome wide repression of differentiation programs (Kadoch et al. 2016; Alver et al. 2017). However, these tumours become resistant to treatment with the EZH2 inhibitor tazemetostat (Wilson et al. 2010; Knutson et al. 2013). By performing a CRISPR-Cas9 screen, RNA-seq and CUT&RUN, Drosos et al. (2022) have elegantly shown that the absence of H3K36me2 written by the NSD1 enzyme is essential for polycomb to expand H3K27me3. Furthermore, they have shown that NSD1 cooperates with SWI/SNF and opposes PRC2, but mutations in *NSD1* co-occur with mutations in SWI/SNF, making the tumours polycomb-driven. Finally, they show that inhibition of H3K36me2 demethylase

KDM2A restores H3K36me2 and restores sensitivity to EZH2 inhibition (Drosos et al. 2022).

Information about looping and 3D chromatin architecture can also provide mechanistic explanations for drug treatments. For example, NOTCH1 inhibition in T-ALL can lead to altered enhancer-promoter interactions at NOTCH1-bound enhancers, specifically at shorter enhancers (Kloetgen et al. 2020). In this scenario, some enhancers are resistant to NOTCH1 inhibition and still exert an oncogenic role, due to other factors keeping them active. Additionally, CTCF and NOTCH1 can cooperate and induce CTCF binding at inappropriate loci in T-ALL, completely changing the transcriptional landscape determined by new promoter-enhancer loops (Fang et al. 2020).

Further, integration of epigenomic data with RNA-seq data can broaden the understanding of why some alterations lead to cancer and how they rewire the gene regulatory network. For example, Assi et al. (2019) have integrated AML-specific genomic alterations with epigenetic assays and RNA-seq to uncover subtype-specific chromatin conformation and transcription factor binding profiles. Another example is the usage of single-cell RNA-seq and ATAC-seq combined with clone tracking via mitochondrial sequencing (mtscATAC-seq). Using this technique, Nuno et al. (2024) have shown that relapsed AML can show features of convergent epigenetic evolution. This means that some patients with AML experiencing relapse, despite having different driver alterations, develop treatment resistance via changes in their epigenetic landscape without leukaemia cells' acquiring new mutations. Additionally, the different cell populations within each leukaemia converge towards a similar epigenetic landscape (Nuno et al. 2024).

Finally, integration of cancer epigenomic landscape information with mutations in epigenetic factors may uncover context-specific essential genes that can be targeted therapeutically in addition to causal mechanisms for disease development. For example, transcription factor footprinting and nucleosome analysis from ATAC-seq suggests that *ARID1A* is essential for B-cell development (Barisic et al. 2024). ARID1A is part of the nucleosome remodelling complex SWI/SNF and guides B-cell development by controlling binding of PU.1 and NF-κB. However, *ARID1A* haploinsufficient lymphomas are

driven towards a pre-memory B cell state, making the cells dependent on the SWI/SNF complex, and susceptible to inhibition of SMARCA2/4 (part of the SWI/SNF complex).

1.3.7 From network and integrative biology to targeted therapy

Most cancer subtypes are still treated using chemotherapy, which can lead to side-effects and toxicity due to non-specific activity. Molecular markers associated with cancer can help us identify and develop more precise therapies. For example, the HER2 antibody trastuzumab was developed to target *HER2*-overexpressing breast cancers (Dillman 1999). Another landmark example of targeted therapy is the treatment of chronic myelogenous leukaemia (CML) harbouring the Philadelphia chromosome (Ph+) (translocation between chromosomes 9 and 22) with imatinib, a tyrosine kinase inhibitor (TKI) that targets the *BCR::ABL1* fusion protein that causes the disease (Cohen et al. 2002). Similarly, acute lymphoblastic leukaemias can also harbour the Philadelphia chromosome, and are treated with TKIs. Interestingly, there are ALL subtypes with gene expression patterns similar to Ph+ ALL, but lacking the translocation (Roberts et al. 2014). These typically showcase other fusions involving ABL-class kinases and can also be targeted with TKIs due to their similarity to Ph+ ALL (Tanasi et al. 2019; Senapati et al. 2023).

Although there are precision therapies that can directly target an oncogenic protein, many cancer types may harbour untargetable oncogenic proteins. Additionally, tumour formation can be driven by loss of tumour suppressor genes, multiple oncogenic events or more complex genetic and epigenetic backgrounds. Therefore, to systematically find novel associations between drug sensitivity and transcriptomics, network approaches have been employed to find regulators driving sensitivity or resistance to drugs in cancer cells (Garcia-Alonso et al. 2018). Similarly, associations between pathway activity inferred from transcriptomics and drug sensitivity may provide robust markers for drug indication (Schubert et al. 2018). Additionally, Thatikonda et al. (2024) have shown that correlations between the inferred activity of TFs and gene dependencies from CRISPR screens may reveal new potential targets in

cancer. They confirm two of the associations (increased inferred *TEAD1* activity leads to increased dependency on ITGAV and PTK2) by performing validatory experiments

One question that has remained unanswered is on the human bias of inferred activity approaches: Do they provide more information than the expression of the regulator, or is their role as a hypothesis creation tool? For example, Trescher & Leser (2019) show through a systematic study that inferred TF activity does not provide robust insights in knock-down experiments in *E.coli* and in human cells. This question will be further explored in Chapter 2 (Tudose et al. 2023).

As detailed in Part 1, genes operate through biological networks, which are complex and include many steps of regulation (Figure 1.1). Therefore, efforts have been made to take into consideration these steps of regulation in finding novel appropriate therapeutic avenues. For example, methylation status at CGIs may be used as a biomarker for drug response in glioblastoma (Tuominen et al. 2015). And, as described above, chromatin accessibility and chromatin architecture can also be predictive or explain the mechanisms of drug resistance and unravel alternative targets (Drosos et al. 2022; Kloetgen et al. 2020; Fang et al. 2020). Integrative and network biology approaches used in haematopoiesis and leukaemia will be covered in the next part, with a focus on chromatin biology in acute myeloid leukaemia and particularly the PRC2 complex.

Section 4 - Leukaemia

1.4.1 Background

Leukaemia is the most common paediatric cancer worldwide, accounting for ~25% of cancer cases in children, amounting to approximately 45-50 cases per year in Ireland in < 16 year olds (Ries et al. 1999*; Irish National Children's Cancer Service, internal figures)*.

Leukaemia can be broadly classified by the degree of cell differentiation into chronic (CL - more mature cells) and acute leukaemia (AL - less

differentiated cells) (Khoury et al. 2022). Paediatric leukaemia is almost exclusively acute and is characterised by the disruption of developmental processes during haematopoietic lineage differentiation (Greaves 1997). The block in differentiation is due to subversion of haematopoietic TFs function caused by somatic mutations, deletions or translocations in TFs or key upstream epigenetic or signalling genes (Tenen et al. 1997).

In normal haematopoiesis, haematopoietic stem and progenitor cells (HSPCs) commit to a diverse array of possible cell types via the expression of lineage-specific transcriptional programs (Figure 1.8) (Laurenti & Göttgens 2018). The majority of post-natal haematopoiesis occurs in the bone marrow resulting in the generation of B-cells, natural killer (NK) cells, dendritic cells, monocytes, granulocytes, erythrocytes and platelets. In addition, T-lymphoid development takes place in the thymus (Rothenberg 2005; Ciofani & Zúñiga-Pflücker 2007). In leukaemia, these processes are disrupted, leading to clonal expansion of undifferentiated progenitors in the bone marrow and other organs.

1.4.2 Classification

Acute leukaemia is also classified based on phenotypic resemblance to normal blood cell precursors: lymphoid (ALL) or myeloid (AML). Typically, ALL arises from disruption of the lymphoid lineage and AML arises from myeloid progenitors. However, single-cell studies have shown that the line between lymphoid and myeloid development is blurred (Karamitros et al. 2017; Belluschi et al. 2018). Both early and mature normal progenitors retain the potential to differentiate along different lineages (Goardon et al. 2011; Luc et al. 2012).

There are a proportion of rare cases (<5%) that show mixed phenotype, known as mixed phenotype acute leukaemia (MPAL). MPALs are heterogeneous and display lineage ambiguity, expressing both myeloid and lymphoblastic phenotypic markers and transcriptional programs (Quesada et al. 2018; Granja et al. 2019; Bond et al. 2020). For example, T/Myeloid MPALs share characteristics of both myeloid and T-lymphoid leukaemias, such as *MPO* and *CD3* expression and have poor prognosis, leading to the suggestion that

they should be treated as a distinct diagnostic group and based on their mutational and transcriptomic profile (Gutierrez & Kentsis 2018; George et al. 2022).

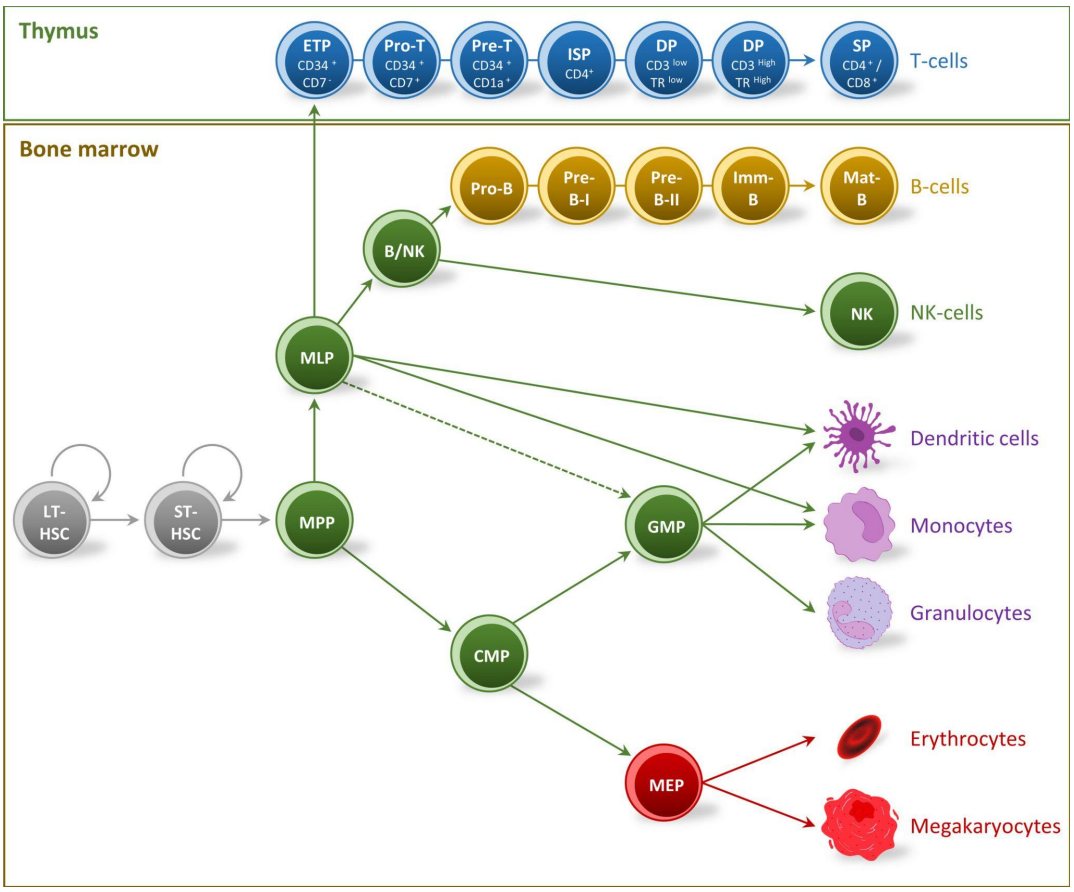


Figure 1.8 | Normal haematopoietic development in the bone marrow and T-cell maturation in the thymus. LT-HSC = long-term haematopoietic stem cell; ST-HSC = short-term haematopoietic stem cell; MPP = multipotent progenitor; MLP = multilymphoid progenitor; ETP = early thymic progenitor; B/NK = B-cell/natural killer cell; CMP = common myeloid progenitor; GMP = granulocyte–monocyte progenitor; MEP = megakaryocyte–erythrocyte progenitor; Pro-B = progenitor B-cell; Pre-B = precursor B-cell; Imm. B = immature B-cell; Mat. B = mature B-cell; Pro-T = progenitor T-cell; Pre-T = precursor T-cell; ISP = immature single-positive; DP = double positive; SP = single positive. Figure from Lefeivre et al. (2022).

1.4.3 Acute myeloid leukaemia

AML comprises 15-20% of childhood leukaemias (De Rooij et al. 2015). AML is a heterogeneous disease and historically was classified based on morphology according to the French-American-British (FAB) classification system. This was superseded by the World Health Organization (WHO) classification that

increasingly considers molecular alterations to define leukaemia subtypes (Khoury et al. 2022; Alaggio et al. 2022). AML incidence rises with age, with most cases occurring after 60 years of age. However, there is also a peak in incidence in patients <1 year old.

Paediatric leukaemias are significantly different from leukaemias occurring in older patients in terms of genetic aberrations. More than half of paediatric AML are driven by translocations, whereas only 26% of adult AML have any translocation (Bolouri et al. 2018). Adult AML tends to be more often driven by point mutations. Further, the repertoire of mutations observed in adult and paediatric AML is very different (see below).

1.4.4 Epigenetic alterations in AML

Normal haematopoiesis follows lineage commitment from HSCs to differentiated progenitors (Figure 1.8). The differentiation processes are tightly regulated by gene regulation through epigenetic mechanisms, such as DNA methylation, TF binding, chromatin remodelling, 3D chromatin architecture and histone PTMs (Cui et al. 2009; Mochizuki-Kashio et al. 2011; Corces et al. 2016; Li et al. 2017; Han et al. 2019; Izzo et al. 2020). Therefore, it is unsurprising that these processes are commonly altered during leukaemogenesis (Figure 1.9).

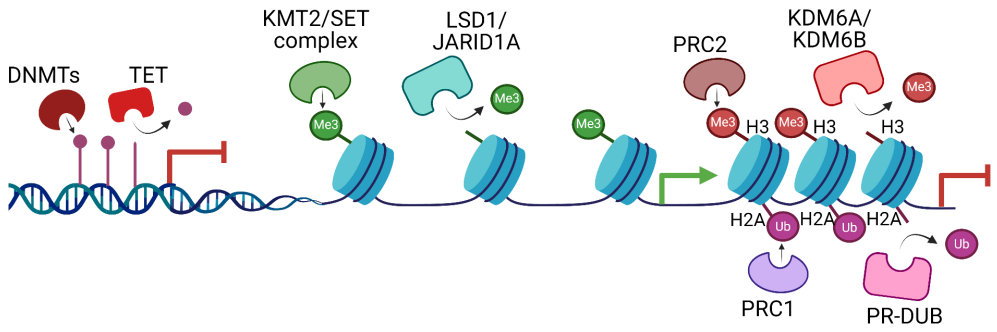


Figure 1.9 | Epigenetic processes that are commonly altered in childhood AML. Figure adapted from Jones et al. (2020). Created with Biorender.com.

1.4.5 *KMT2A* rearrangements

One of the most commonly occurring alterations in paediatric AML is *KMT2A* (*MLL*)-rearrangement (*KMT2A*r). In total, 160 different translocation partners

have been identified that *KMT2A* forms fusion transcripts with, the most common ones being *MLLT3*, *MLLT10* and *AFDN* (Meyer et al. 2018; Meyer et al. 2023). Many of the fusion partners are part of the super elongation complex (SEC), leading to aberrant recruitment of KMT2A with the SEC. KMT2A is a methyltransferase that deposits trimethyl on lysine 4 of histone 3 (H3K4), which is associated with gene activation. The KMT2Ar fusion proteins display aberrant localisation, leading to H3K4me3 deposition at inappropriate genomic regions and recruitment of other factors at these loci. For example, histone methyltransferase DOT1L is recruited by KMT2A fusion partners, leading to aberrant H3K79me at KMT2A target genes and active transcription (Bernt et al. 2011).

1.4.6 *DNMT3A* and *TET2* mutations

DNMT3A mutations are very common in adult patients ≥16 years old (20-22%), however they are only present in 1-2% of paediatric cases (Ley et al. 2010; Liang et al. 2013). DNMT3A is a DNA methyltransferase and loss-of-function mutations in *DNMT3A* lead to hypomethylated CGIs and active transcription at target loci (Jeong et al. 2014; Qu et al. 2014). Similarly, *TET2* mutations are rare in paediatric AML, but common in adults. TET2 is a key player in demethylating CGIs, by catalysing the conversion of methylcytosine to 5-hydroxymethylcytosine and de-repressing gene promoters (Liang et al. 2013).

1.4.7 IDH mutations

Isocitrate dehydrogenases (IDHs) are enzymes that catalyse the formation of α-ketoglutarate (α-KG) from isocitrate. Mutations in *IDH1* or *IDH2* increase with age, with 3.4% of paediatric (0-17 years old) samples harbouring an *IDH* mutation (Zarnegar-Lumley et al. 2023). IDH mutations result in the enzymes being able to catalyse formation of 2-hydroxyglutarate (2-HG) from α-KG (Ward et al. 2010). Downstream, this results in the inhibition of epigenetic factors that use α-KG during histone demethylation, such as TET2 and lysine demethylases (KDM) (Losman et al. 2013). IDH1 and IDH2 can be targeted with ivosidenib

and enasidenib, respectively. Enasidenib shows an overall response rate of 40.3% in a clinical trial on patients between 19-100 years old (median=67 years old) (Stein et al. 2017).

1.4.8 HDACs and RUNX1::RUNX1T1

Histone deacetylases (HDACs) have an important role in epigenetic regulation and by removing acetyl groups from histones at actively transcribed genomic regions. *HDAC* mutations are very rare in paediatric leukaemia. However, RUNX1::RUNX1T1 fusion proteins, which are common in childhood leukaemia, can recruit HDACs, leading to aberrant transcriptional patterns and interfering with differentiation (Swart & Heidenreich 2021).

1.4.9 NUP98

NUP98 fusions are also common drivers of leukaemia in paediatric cases and are associated with poor outcomes. *NUP98* can fuse with more than 28 different partner genes, including both *HOX* and non-HOX genes. Additionally, some of the non-HOX partners of *NUP98* include epigenetic regulators, such as *NSD1*, *JARID1A* and *PHF23*. The fusion protein typically upregulates the HOXA cluster (i.e., *HOXA7*, *HOXA9*, *HOXA10*) (Gough et al. 2011). Furthermore, Saw et al. (2013) show that HOX partner fusions have a greater effect on aberrant gene expression and greater self-renewal potential than non-HOX fusions in *NUP98*-rearranged AML.

1.4.10 Polycomb alterations in leukaemia

The polycomb complexes PRC1 and PRC2 have key roles in haematopoiesis. From murine experiments we know that depletion of PRC1 and PRC2 in early haematopoiesis can severely affect the HSC pool (Kamminga et al. 2006; Vidal & Starowicz 2017). Cell identity in later progenitors can be impaired by loss of both PRC1 and PRC2 components. For example, loss of BMI1 in HSPCs

results in early activation of lymphoid programs (Oguro et al. 2010). Similarly, EZH2 loss impairs cell cycle and B cell development (Su et al. 2002).

Alterations of PRC2 core components are associated with poor outcome and chemoresistance in paediatric AML and T-ALL (Bond et al. 2018; Ariës et al. 2018). Low expression of *EZH2* in patients has also been linked to chemoresistance (Göllner et al. 2017). Additionally, a poor-outcome epigenetic subgroup of AML which included, amongst other genes, *EZH2* mutations has been reported (Papaemmanuil et al. 2016). However, the mechanism underpinning aggressive disease biology is still poorly understood. In *RAS*-mutated myeloid leukaemias, EZH2 inactivation hyperactivates global RAS-signalling, making the cancer cells more sensitive to MEK (MAP2K1) inhibition (Berg et al. 2021). Low *EZH2* expression is also strongly correlated with evolution from paediatric myelodysplastic syndrome (MDS) into AML. *EZH2* can also act as an oncogene in some cancer subtypes, such as diffuse large B-cell lymphoma (DLBCL). In DLBCL, *EZH2* activating mutations lead to accumulation of genome-wide H3K27me3, a block in differentiation and aggressive disease (Zhou et al. 2015). Alterations of PRC2 components are also found in cases of juvenile myelomonocytic leukaemia (JMML), an early childhood myelodysplastic/myeloproliferative neoplasm (Caye et al. 2015).

1.4.11 3D chromatin in haematopoiesis and AML

3D chromatin architecture dynamics have emerged as important characteristics of development and lineage specification (Zheng & Xie 2019). Recently, a role for 3D chromatin organisation has been outlined in haematopoiesis (Kloetgen et al. 2019). In lymphopoiesis, findings from time course experiments suggest that changes in 3D genome organisation may restrict lineage commitment in T-cell development (Hu et al. 2018). These adjustments occur before gene expression changes, priming the cells for the activation of lineage-specification transcriptional programs. However, there are also clear reports of 3D genome changes directly affecting gene expression, via Pax5-mediated loop formation during B-cell differentiation, or via Bcl11b activity in T-cell differentiation (Isoda et al. 2017; Johanson et al. 2018). Additionally, the haematopoietic

differentiation TF CEBPA is able to bind chromatin, form long range loops and promote compartment switching, leading to altered expression in transdifferentiation between B-cells and macrophages (Christou-Kent et al. 2023).

Enhancer-Promoter (E-P) loops can be altered in leukaemia by alterations in the enhancer sequence. For example, the *MYC* proto-oncogene displays cell-type specific E-P looping in CD4+ CML and B-cell lymphoblasts (Mumbach et al. 2017). Additionally, E-P loops can be altered by genomic translocations (DNA sequence), epigenomic translocations (region with modified DNA sequence or histone PTMs), inversions or TAD-boundary disruptions (Ryan et al. 2015; Hnisz et al. 2016; Weischenfeldt et al. 2017; Mikulasova et al. 2022). These alterations may lead to “enhancer hijacking” events where an enhancer comes in contact with oncogene promoters. For example, In AMLs with inv(3)/t(3;3), an enhancer comes in contact with the promoter of the *EVI1* (*MECOM*) oncogene, activating it (Gröschel et al. 2014). Simultaneously, the enhancer loses contact with the *GATA2* promoter, conferring functional haploinsufficiency (i.e., one of the *GATA2* alleles is not expressed anymore) (Gröschel et al. 2014).

Proteins involved in the formation of 3D chromatin structures are also commonly altered in both T-ALL and AML (Mazumdar et al. 2015; Liu et al. 2017). Mutations in the cohesin complex members are present in more than 15% of AMLs in a mutually exclusive fashion (Kon et al. 2013). Mutations in *SMC1A*, *SMC3*, *STAG2* and *RAD21* lead to altered chromatin architecture and an increase in open regions at promoter sites of gene targets of key TFs involved in haematopoiesis, such as ERG, GATA2 and RUNX1 (Mazumdar et al. 2015). This leads to a blockage in the differentiation of HSPCs. Alterations in oncogenic TFs can also be associated with impaired 3D chromatin architecture. As described in the subsection 1.3.6, Notch contributes directly to looping in development. Petrovic et al. (2019) have described that in B-cell lymphoma, this can lead to large 3D clusters involving multiple enhancers activating promoters of oncogenes such as *MYC.*

The largest study interrogating 3D chromatin architecture changes in AML performed to date comes from Xu et al. (2022). They performed in-situ Hi-C and RNA-seq on samples from a cohort of 32 patients: 25 AML samples,

four normal HSPC samples and three normal peripheral blood mononuclear cells (PBMCs) samples. Additionally, they performed CTCF, H3K27Ac and H3K27me3 CUT&TAG, ATAC-seq and bisulfite sequencing (to identify methylated CGIs) in a subset of the samples. They found AML-specific looping involving promoters and enhancers of genes involved in haematopoiesis and myeloid transformation pathways (i.e., *MYCN*, *RUNX1*, *MEIS1*). Surprisingly, they found that >70% of promoter-silencer loops are also mediated by CTCF at one anchor, at least. Additionally, these loops were enriched for H3K27me3. This suggests that CTCF may play a role in polycomb loop formation. Finally, they showed that long-range promoter-silencer loops act as tumour suppressors in AML.

1.4.12 Targeted therapies in leukaemia

One of the first targeted cancer therapies approved for clinical use was imatinib for the treatment of Ph+ CML (Cohen et al. 2002). Imatinib is now also used for the treatment of paediatric Ph+ ALL (Schultz et al. 2014).

For AML there are many small-molecule inhibitors that have been approved for targeted therapy. A few examples include IDH inhibitors, used for treating AML with *IDH* mutations (Liu & Gong 2019), FLT3 inhibitors (Larrosa-Garcia & Baer 2017) and BCL2 inhibitors, such as venetoclax, with pro-apoptotic characteristics (Wei et al. 2020). Azacitidine and decitabine are hypomethylating agents used in first-line treatment in AML which act by blocking DNA methylation and inducing cellular differentiation (Totiger et al. 2023).

*KMT2A*r AMLs and ALLs have very aggressive disease and recently two potential targetable components of the KMT2A network have been uncovered. Firstly, the H3K79 methyltransferase DOT1L interacts with KMT2A fusion partners (see subsection 1.4.5) (Bernt et al. 2011) and can be targeted by pinometostat, which has been shown to be well tolerated in clinical trials, but only shows modest efficacy as a single agent (Shukla et al. 2016; Stein et al. 2018). Secondly, menin (encoded by *MEN1*) is a cofactor which is essential for TF recruitment at KMT2A loci. This interaction between menin and KMT2A

fusion proteins is at the *HOX* and *MEIS1* loci, which are responsible for leukaemogenesis in *KMT2A*r leukaemia (Caslini et al. 2007; Cierpicki & Grembecka 2014). Revumenib and ziftomenib are inhibitors that target the PPI between menin and KMT2A and currently in clinical trials (Issa et al. 2022; Erba et al. 2022).

Careful delineation of transcriptional programs may also identify targeted treatment approaches that may even be subclone-specific. For example, a network pharmacology approach named network-based Bayesian inference of drivers (NetBID) was applied to T-ALL transcriptional profiles (Gocho et al. 2021). NetBID analysis has shown that some patients can be sensitive to dasatinib (TKI) treatment due to stage-specific activation of preTCR-LCK signalling. Conversely, patients resistant to dasatinib show strong BCL2 activation and could potentially respond better to venetoclax treatment (Gocho et al. 2021). Similarly, NetBID analysis has shown intra-leukaemia heterogeneity in B-ALL (Huang et al. 2024). Clones exhibiting Pre-Pro-B-cell (precursors of progenitor B-cell) transcriptional programs were susceptible to venetoclax, whilst Pro-B cells were more sensitive to asparaginase (Huang et al. 2024).

With the rise in proteomic characterisation of tumours, including proteogenomic characterisation in leukaemia we can point towards additional resistance markers (Zhang et al. 2014; Li et al. 2023). Integration of proteomics, transcriptomics and *ex-vivo* drug-screen response data from the BEAT-AML cohort has shown that proteomic subtypes were able to complement mutational subtypes in drug response stratification (Bottomly et al. 2022; Pino et al. 2024).

Taken together, in this introductory chapter, I have summarised experimental and computational approaches to investigate gene regulation in cancer. In the next chapter I will present research work on a systematic approach to gene regulation on GRN-inferred regulatory gene activity and its ability to identify vulnerabilities in cancer. Finally, in the final results chapter I will present work that focuses on an epigenetic perspective, as we explore the role of PRC2 in gene regulation in AML via detailed epigenomic and transcriptomic characterisation of an *EZH2* loss of function model.

Section 5 - PhD research overview

1.5.1 Aims and objectives

1. Systematically assess the ability of GRN-inferred activity to uncover vulnerabilities in cancer cells by exploiting unbiased CRISPR screen data from the DepMap.
2. Determine whether GRN-inferred activity performs better than mRNA expression at uncovering vulnerabilities in cancer by using multiple GRN reconstruction and activity inference methods.
3. Perform a transcriptomic and epigenomic characterization of PRC2 depletion in an AML cell line model, focusing on the alternative lineage programs controlled by PRC2.
4. Identify alterations in the chromatin landscape associated with PRC2 depletion that may result in phenotype alterations and explain drug resistance in AML.

1.5.2 Primary hypotheses for research chapters

Chapter 2: Here I test the hypothesis that GRN-inferred activity outperforms mRNA abundance at predicting dependencies in cancer cell lines.

Chapter 3: Here I test the hypothesis that heterozygous PRC2-depletion leads to epigenomic and transcriptomics changes that mediate altered leukaemia biology and chemoresistance.

CHAPTER 2 - Gene essentiality in cancer is better predicted by mRNA abundance than by gene regulatory network-inferred activity

Cosmin Tudose[1,2,3], Jonathan Bond*[1,2,4], Colm J. Ryan*[1,5,6]

[1]Systems Biology Ireland, University College Dublin, Dublin 4, Ireland,

[2]School of Medicine, University College Dublin, Dublin 4, Ireland,

[3]The SFI Centre for Research Training in Genomics Data Science, Ireland,

[4]Children's Health Ireland at Crumlin, Dublin,

[5]School of Computer Science, University College Dublin, Dublin 4, Ireland,

[6]Conway Institute, University College Dublin, Dublin 4, Ireland

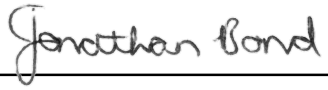
***Co-corresponding Authors:** Colm J. Ryan, E-mail: colm.ryan@ucd.ie, Jonathan Bond, E-mail: jonathan.bond@ucd.ie

Author's contributions:

Conceptualization, Formal analysis, Original Draft, Writing: Review & Editing, Visualisation: CT, CJR, JB. Methodology and Validation: CT. Supervision, Project administration, Funding Acquisition: CJR, JB.

2.1 Declaration of co-authorship collaboration

1. Examination Candidate Details
Examination Candidate Name: **Cosmin Tudose**
Examination Candidate UCD Student Number: **20209354**
Research Degree for which thesis is being submitted: **PhD**
Title of Research thesis: **Defining molecular vulnerabilities in childhood leukaemia through biological network analysis**
2. Details of the Paper
Title of Paper: **Gene essentiality in cancer is better predicted by mRNA abundance than by gene regulatory network-inferred activity**
Current Status of the Paper: **Published**
Name of Journal: **NAR Cancer**
Digital Object Identifier: **10.1093/narcan/zcad056**
AUTHOR CONTRIBUTIONS
Cosmin Tudose's contribution to the research work described in this chapter: ● Project conceptualisation. ● Formal analysis. ● Original draft. ● Review and editing of manuscript. ● Presentation of this research work at conferences.
Other Authors' contributions to the research work:
Jonathan Bond: ● Project conceptualisation. ● Primary supervision. ● Original draft. ● Review and editing of manuscript.

Colm J. Ryan: ● Project conceptualisation. ● Primary supervision. ● Original draft. ● Review and editing of manuscript.
Principal Supervisor: **Jonathan Bond**
Signature: 
Date: 10.12.2024
PhD candidate: **Cosmin Tudose**
Signature: 
Date: 10.12.2024

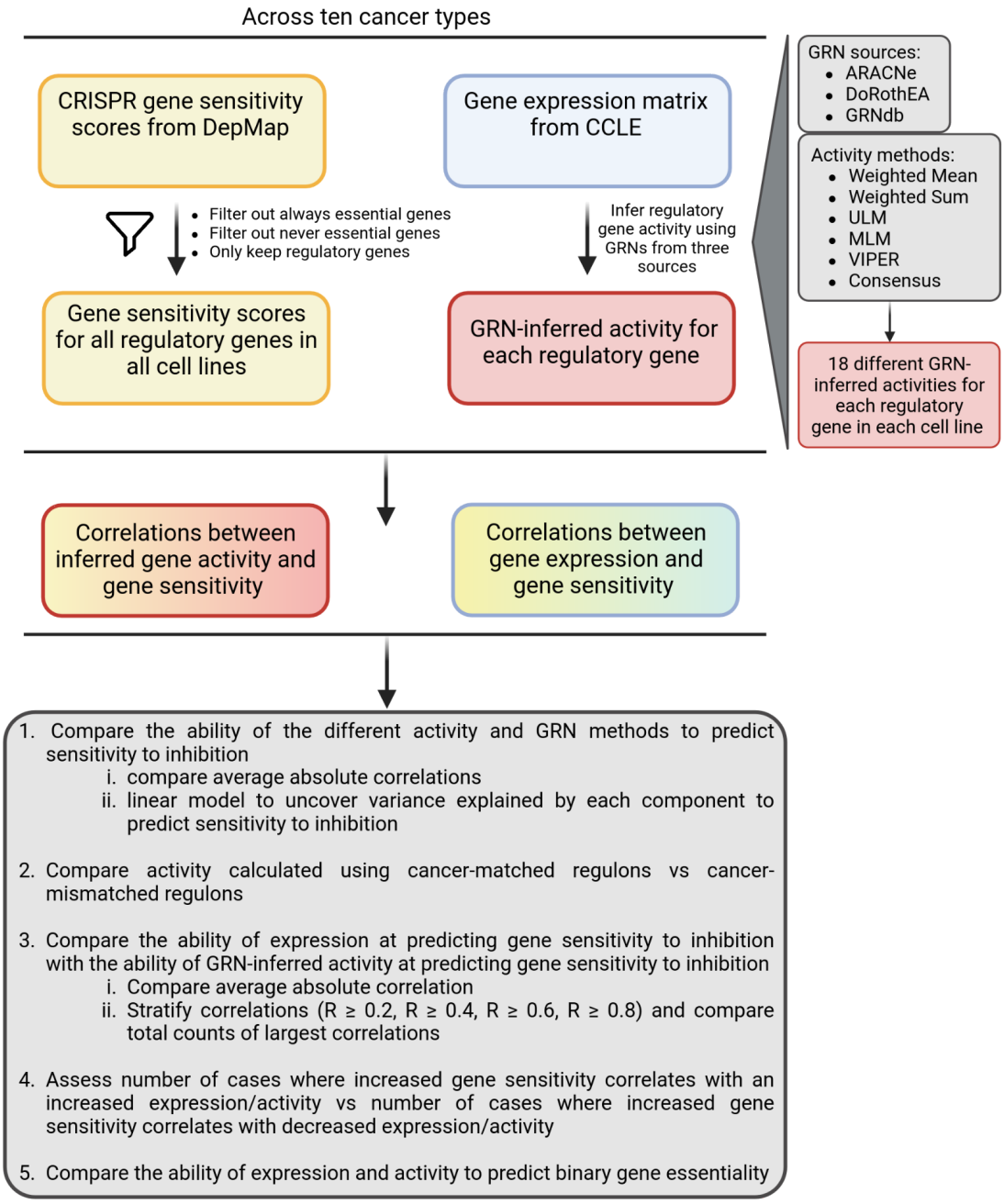


Figure 2.0 | Flowchart describing the analytical design of Chapter 2. Created with Biorender.com.

2.2 Abstract

Gene regulatory networks (GRNs) are often deregulated in tumour cells, resulting in altered transcriptional programs that facilitate tumour growth. These altered networks may make tumour cells vulnerable to the inhibition of specific regulatory proteins. Consequently, the reconstruction of GRNs in tumours is often proposed as a means to identify therapeutic targets. While there are examples of individual targets identified using GRNs, the extent to which GRNs can be used to predict sensitivity to targeted intervention in general remains unknown. Here we use the results of genome-wide CRISPR screens to systematically assess the ability of GRNs to predict sensitivity to gene inhibition in cancer cell lines. Using GRNs derived from multiple sources, including GRNs reconstructed from tumour transcriptomes and from curated databases, we infer regulatory gene activity in cancer cell lines from ten cancer types. We then ask, in each cancer type, if the inferred regulatory activity of each gene is predictive of sensitivity to CRISPR perturbation of that gene. We observe slight variation in the correlation between gene regulatory activity and gene sensitivity depending on the source of the GRN and the activity estimation method used. However, we find that there is consistently a stronger relationship between mRNA abundance and gene sensitivity than there is between regulatory gene activity and gene sensitivity. This is true both when gene sensitivity is treated as a binary and a quantitative property. Overall, our results suggest that gene sensitivity is better predicted by measured expression than by GRN-inferred activity.

2.3 Introduction

A large volume of cancer molecular profiles have become available through compendia such as The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al. 2019). Additionally, maps of cancer vulnerabilities have been generated using CRISPR and drug screens through efforts such as The Cancer Dependency Map (DepMap) and the Genomics of Drug Sensitivity in Cancer (GDSC) (Dempster et al. 2021; Dempster et al. 2019; Meyers et al. 2017; Pacini et al. 2021; Garnett et al. 2012; Iorio et al. 2016). A major outstanding challenge is to identify therapeutic targets for molecularly defined cohorts.

Many genetic alterations drive oncogenesis by altering transcriptional programs that govern critical cellular processes such as proliferation, cell cycle and apoptosis via gene regulatory networks (GRNs) (Bushweller 2019). In cancer, GRN perturbation disrupts key transcriptional programs, and can lead to changes in response or resistance to therapies (Alessandrini et al. 2018; Ohanian et al. 2019). Targeting GRNs to restore normal cell function is a clinically attractive idea. Currently, there are ongoing trials targeting molecular networks, such as STAT3/5 or menin in acute myeloid leukaemia (AML), estrogen receptor in ER+/HER2- breast cancer and MDM2 as part of the p53-MDM2 interaction in various cancer types (Henley & Koehler 2021).

Computational tools are often employed to reconstruct GRNs with a view to identifying therapeutic targets (Lefebvre et al. 2012; Barabási et al. 2010), e.g., ARACNe (Lachmann et al. 2016), GENIE3 (Huynh-Thu et al. 2010) and KBoost (Iglesias-Martinez et al. 2021) interpret correlations from transcriptomes to construct GRNs. Each GRN is composed of regulons and each regulon contains a regulatory gene, its targets, and the weights between. These GRNs, however, are *in-silico* inferred, and biological validation is not straightforward. This is particularly challenging in human cells, as a gold standard map of human GRNs does not yet exist.

Transcription factor (TF) activity, also referred to as protein activity (Alvarez et al. 2016) or regulon activity (Aibar et al. 2017), represents the inferred activity of a regulatory gene derived from the variance in transcript abundance of its targets, according to a predetermined regulon (Essaghir et al.

2010). Inferred activity has been used to investigate drug response (Gocho et al. 2021; Garcia-Alonso et al. 2018), uncover “hidden” drivers (Shaw et al. 2021) and showcase the role of “master regulators” in cancer (Alvarez et al. 2018; Wang et al. 2009; Alvarez et al. 2016). However, validation often involves assessing the impact of perturbing a small number of example genes and measuring the resulting transcriptional changes (Alvarez et al. 2016). Previous work has assessed the extent to which inferred activity associates with mutational status and copy number variation (Garcia-Alonso et al. 2018; Sousa et al. 2023). Somewhat surprisingly, in a large multi-omic dataset of tumor and cell line samples, Sousa et al. (2023) found limited correlation between copy number variation and the inferred activity of transcription factors. While the mutation of individual genes (e.g., *TP53*, *GATA6*) could in some cases be associated with altered activity of the encoded transcription factor, this was not the default. In general the authors observed a lower correlation between inferred activity and protein abundance than they observed between inferred activity and mRNA abundance (Sousa et al. 2023). This is somewhat surprising as one would assume that the protein abundance of a tumor is a better proxy for its activity than its mRNA abundance. Akin to GRN inference methods, there are many activity inference methods, with little consensus across them (Trescher et al. 2017).

Given that GRNs have been suggested to drive oncogenic processes, dysregulation of regulatory gene activity may lead to vulnerability to perturbation and dependency to regulatory genes (Bhagwat & Vakoc 2015; Bradner et al. 2017). However, this has not been assessed at a systematic level. Here, we used CRISPR screens as a precise method to validate whether GRN-inferred activity can predict sensitivity to inhibition. In CRISPR screens, each gene is perturbed with sgRNAs, and a gene’s sensitivity to inhibition is assigned a score calculated from cell growth and survival (Shalem et al. 2015). The DepMap project uses this approach to characterise the gene sensitivity profiles of more than 1,000 cell lines via genome-wide CRISPR screens (Dempster et al. 2021; Dempster et al. 2019; Meyers et al. 2017; Pacini et al. 2021). We inferred regulatory gene activity in these cell lines using both computationally derived and curated regulons (Garcia-Alonso et al. 2019). We then evaluated correlations between gene sensitivity and inferred activity across cell lines.

Additionally, in regulatory genes, we compared expression and activity in their ability to predict sensitivity to inhibition. Although gene essentiality is often discussed as a binary property, processed CRISPR screens typically report a quantitative score representing the sensitivity of each cell line to the inhibition of each gene. In this work, we analyze the ability of GRN-inferred activity, and mRNA expression, to predict sensitivity to gene inhibition as a quantitative property and also as a binary score (essential / non-essential). Overall, we found little evidence of activity estimation methods providing an advantage over measured mRNA abundance.

2.4 Materials and methods

2.4.1 ARACNe regulons processing

We loaded cancer type-specific ARACNe regulons from the aracne.networks 1.20.0 R package and transformed them into data frames using the 'reg2tibble' function from the binilib 0.2.0 R package. We converted Entrez IDs into gene symbols via the org.Hs.eg.db 3.14.0 R package. We calculated an updated mode of regulation (MOR) by multiplying the likelihood with the sign of the MOR for each interaction. MOR indicates the directionality of the interaction (i.e., -1 = inhibition; 1 = activation). 'Source' genes missing in the expression data were filtered out.

To infer ARACNe regulons from the BRCA and LUAD CCLE gene expression data, we ran ARACNe-AP with the default settings (Lachmann et al. 2016).

2.4.2 GRNdb regulons processing

We downloaded TCGA-inferred cancer type-specific regulons from http://www.grndb.com (Fang et al. 2021). We used the gene symbols provided. These regulons contain a weight calculated with GENIE3 (Huynh-Thu et al. 2010), but do not contain directionality (inhibition or activation), which the

ARACNe regulons do, by providing MOR. Therefore, we inferred MOR using the TCGA dataset, as it was used to build the regulons, as follows:

We downloaded log2(tpm+1) normalised RNA-seq from the UCSC Treehouse Public Data, v11 Public PolyA. We separated the transcriptomics into ten matrices, for each cancer type, using the “disease” column from the clinical data file. For each cancer type, we removed genes with 0s in more than half of the samples, whilst we imputed the others using “impute.knn” from the impute 1.68.0 R package. We inferred the MOR for each interaction in the GRNdb regulons by calculating the Spearman correlation between the expression of each regulatory gene and its target. We computed an updated MOR by multiplying the GENIE3 weight with the sign of the MOR from the previous step.

To build GRNdb-like regulons from the CCLE, we followed the instructions from Fang et al. (2021).

2.4.3 DoRothEA regulons processing

We loaded the human DoRothEA regulons from the dorothea 1.6.0 R package (Garcia-Alonso et al. 2019). For the downstream analysis we used the high confidence regulons: A, B and C.

2.4.4 Data wrangling

We downloaded CCLE gene expression, gene sensitivity data and cell line information from DepMap release 21Q4 (Dempster et al. 2021; Dempster et al. 2019; Meyers et al. 2017; Pacini et al. 2021). We filtered these for cell lines from cancer types present in Figure 2.1B and for the downstream analysis we kept only cell lines present in both gene expression and gene sensitivity datasets. Likewise, only genes present in both datasets were retained for each cancer type. For gene nomenclature we used HGNC symbols, discarding Entrez IDs. For each cancer individually, we dropped genes with more than 20% 0s across samples in the gene expression profile.

2.4.5 Filtering 'sometimes' essential genes

We defined genes as essential in a given cell line if they had a CHRONOS score < -0.6 in that cell line. Within each cancer type, we restricted our analysis to genes that were variably essential across cell lines from that cancer type (i.e., genes that were either essential in all cell lines or non-essential in all cell lines were filtered out). Therefore, we were left with sometimes-essential genes only, in order to study variation in sensitivity to inhibition across tumour cells. The CHRONOS score is a scoring system for quantifying and normalising outputs from CRISPR screens and aggregating the results of multiple gRNAs targeting the same gene into a single gene-level score. Full details about CHRONOS are available in Dempster et al. (2021).

For the analysis across all cancers (pan-cancer) we implemented three different thresholds: 1% (i.e., nine cell lines), 5% (i.e., 47 cell lines) and 10% (i.e., 97 cell lines). This means for a gene to be considered sometimes essential it had to be essential in at least 1% (or 5% or 10%) of cell lines and non-essential in at least 1% (or 5% or 10%, respectively) of cell lines.

2.4.6 Computing regulatory gene activity

We ran the 'decouple' function from the decoupleR (Badia-i-Mompel et al. 2022) 2.1.8 package individually on each cancer expression matrix paired with each regulon (Figure 2.1). DecoupleR infers activity via five different methods: Univariate Linear Model (ULM), Multivariate Linear Model (MLM), Virtual Inference of Protein-activity by Enriched Regulon (VIPER), Weighted Mean (W. Mean), Weighted Sum (W. Sum) activity. It then calculates a Consensus across all methods. We used ARACNe, GRNdb and DoRothEA regulons.

2.4.7 Correlation analysis

For each gene we calculated the Pearson's correlation between inferred regulatory gene activity/expression and gene sensitivity scores for sometimes essential genes.

We filtered for genes with a significant Pearson's correlation ($p < 0.05$) and grouped in categories based on absolute Pearson's R: 0.2, 0.4, 0.6, 0.8, 1 and based on the sign of R: positive or negative. We used the cor.test function from the R stats 4.1.2 package.

2.4.8 Regulon size stratification

To stratify our analysis based on regulon size we separated the regulons into three categories based on the number of targets the regulon regulates: small (≤20 targets), medium (>20 targets & ≤100 targets) and large regulons (>100 targets). This could only be performed on GRNdb and DoRothEA regulons, as ARACNe regulons are provided with edge values between every regulatory gene-target pair combination.

2.4.9 Regulon stratification based on the number of unique targets

To stratify our analysis based on the number of unique targets each regulon regulates we separated the regulons into three categories: No unique targets, ≤10% of targets are unique and >10% of targets are unique. A target is unique in a regulon if it is not present as a target gene in any other tested regulon. For the same reasons stated above, this could only be performed on GRNdb and DoRothEA regulons.

2.4.10 Enrichment analysis

We ran Gene Ontology (GO) term enrichment analysis using the WebGestalt R package, with FDR = 10%. For each cancer type and regulon source, we used all regulatory genes tested as a background list.

We used the genes marked as "oncogene" in the cancer gene census (CGC, https://cancer.sanger.ac.uk/census) (Sondka et al. 2018) to test for oncogene enrichment in the genes where activity is better correlated with

sensitivity, using a Fisher's exact test. Similarly, we tested for master regulator enrichment using the master regulator list from Paull et al. (2021) in Table S2.

We have performed GO enrichment analysis on genes with an absolute correlation between Consensus activity/expression and sensitivity to inhibition larger than 0.6. We have looked at the overlap of enriched GO terms between all GRN methods and expression.

2.4.11 Calculate per gene variance for each method

For each gene we calculated the variance across all samples, all regulon sources and cancer type + cancer type-matched regulons.

2.4.12 Comparing activity methods

For each possible cancer type + cancer type-matched regulon combination we calculated the mean Pearson's correlation across all genes (no p-value filtering) for each activity method and compared.

We fit linear models using the lm function from the R stats package (v4.1.0):

|R| ~ Cancer type + Regulon source + Activity method + No. cell lines + RNA-seq variance

|R| ~ Cancer type + Regulon source + Activity method + No. cell lines + No. Unique targets

|R| ~ Cancer type + Regulon source + Activity method + No. cell lines + Regulon Size

The former two linear models were fit only using GRNdb and DoRothEA regulons. This was because in ARACNe regulons all regulatory genes have an edge with every possible target in the genome, leading to all regulons being the same size and having no unique targets.

We estimated the percentage of the variance each term in the linear model explains using adjusted R-squared.

2.4.13 Comparing regulons

For each possible cancer type + regulon combination, we calculated the mean Pearson's correlation across all genes (no p-value filtering) for each activity method to investigate whether cancer type-matched regulons are more predictive of sensitivity than mismatched regulons. We plotted the absolute mean Pearson's R for each combination and assigned ranks 1-10 for each cancer. We conducted Unpaired Two-Samples Wilcoxon tests (wilcox.test function from the R 4.1.2 stats package) to compare the ranks of cancer type-matched regulons to the ranks of cancer type-mismatched regulons.

2.4.14 Common language effect size calculation

For each cancer type + cancer type-matched regulon combination, we calculated the common language effect size (CLES) across sometimes essential genes for each activity method. Here we considered genes essential in at least three cell lines, and non-essential in at least three cell lines as sometimes essential. We used the CLES function from the bmbstats v0.0.0.9001 R package to predict binary essentiality.

We filtered for significance based on the expression/activity difference in non-essential vs. essential genes (Wilcoxon unpaired test $p < 0.05$) and counted the number of genes for each method with a CLES > 0.7, > 0.8 and > 0.9, respectively.

R code to run analyses is available at:

https://github.com/cancergenetics/GRN_activity_corr_essentiality

2.5 Results

2.5.1 Variation in correlation between activity and gene inhibition sensitivity is driven more by cancer type than activity estimation method

Estimating regulatory gene activity requires a GRN (containing edges between regulatory genes and targets) and a gene expression matrix (quantifying the expression levels of all genes in a set of samples) (Figure 2.1A) (Badia-i-Mompel et al. 2022). Typically, activity estimation methods assign an activity score to a given regulatory gene such that higher expression of the gene's targets in a given sample is associated with a higher activity score for the regulatory gene in that sample. Here, rather than focusing on a single GRN or single activity estimation method, we assessed three GRN sources and six activity estimation methods (Figure 2.1A).

We selected ARACNe (Alvarez et al. 2018; Ding et al. 2021), GRNdb (Huynh-Thu et al. 2010; Fang et al. 2021), and DoRothEA (Garcia-Alonso et al. 2019) as representatives of different GRN reconstruction approaches – ARACNe is one of the longest-established methods and infers GRNs solely from transcriptomes; GRNdb is more recently developed and uses GENIE3 GRNs inferred from transcriptomes that are further refined with ChIP-seq data; while DoRothEA contains curated GRNs that incorporate cis-regulatory information from ChIP-seq peaks, literature curated resources, and TF binding motifs within promoters. For ARACNe and GRNdb, we obtained cancer type-specific GRNs, i.e., a breast cancer GRN was assembled from gene expression profiles of breast cancer samples, while for DoRothEA, only cancer type-agnostic GRNs were available.

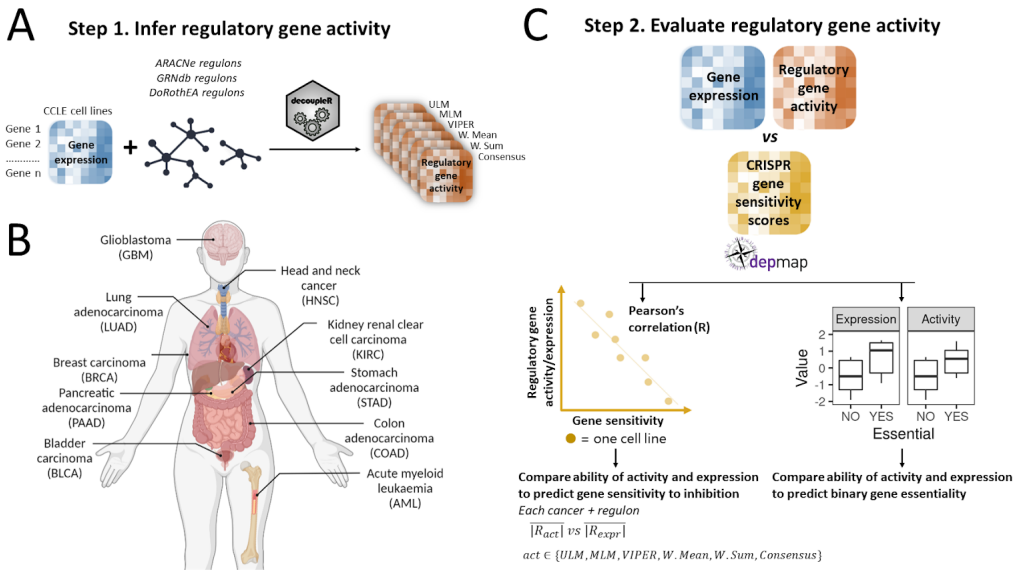


Figure 2.1 | Workflow for evaluating TF activity estimation using CRISPR gene sensitivity profiles from DepMap. A, The activity of regulatory genes was inferred in cancer cell lines using six methods from the DecoupleR package. Gene expression profiles from the CCLE were paired with cancer type-specific regulons from ARACNe, GRNdb and curated pan-cancer regulons from DoRothEA to infer regulatory gene activity. **B,** Ten different cancer types (TCGA abbreviations in brackets) with CCLE gene expression profiles and regulons were used for the analysis (generated using BioRender). **C,** Activity inferred using different GRNs and different activity estimation methods was compared with gene expression using two approaches 1. The Pearson's correlation between inferred activity and gene sensitivity was compared with the Pearson's correlation between expression and gene sensitivity. 2. Activity and expression were used to look at the degree of separation between essential and non-essential genes in a binary fashion using the common-language effect size (CLES) and a Wilcoxon test.

Using decoupleR, we estimated activity in the DepMap cancer cell lines using five different methods: ULM, MLM, VIPER, W. Mean, W. Sum as well as a consensus score calculated by decoupleR using all five scores. These methods work in similar ways: they estimate enrichment scores for each regulatory gene based on its number of targets, targets' expression, and MOR (i.e., inhibition or activation). We then determined the mean absolute Pearson's correlation ($\overline{|R|}$) between the regulatory gene activities and CRISPR gene sensitivity scores across cell lines from specific cancer types. Cancer types for which we had matched GRNs derived from relevant TCGA tumour samples were included in

this analysis, resulting in ten cancer types being assessed (Figure 2.2A). The number of cell lines for each cancer type ranged from 24 (kidney renal clear cell carcinoma - KIRC) to 51 (head and neck squamous cell carcinoma - HNSC) (Figure 2.2A). For comparison, we also included the $\overline{|R|}$ between mRNA abundance and gene sensitivity scores.

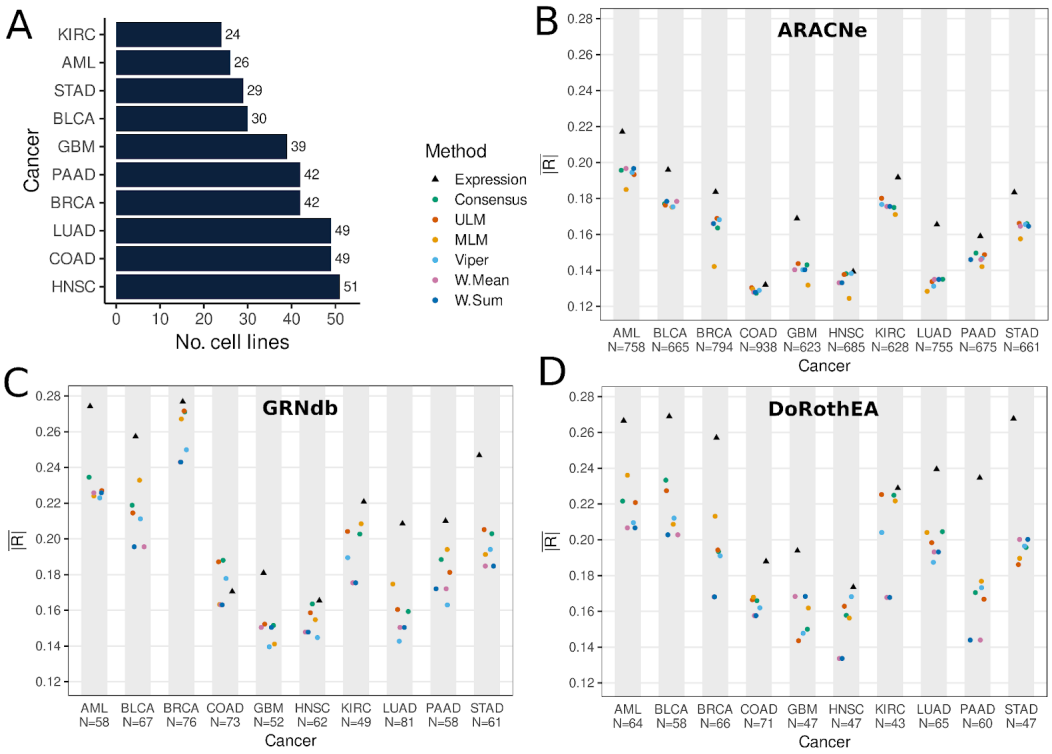


Figure 2.2 | Activity estimation methods have similar performance in predicting gene sensitivity. A, Number of cancer cell lines present in DepMap and CCLE for each cancer. **B, C, D,** Comparison between the different inferred activity methods (paired with cancer type-matched regulons) correlating with gene sensitivity and gene expression correlating with gene sensitivity. **B,** ARACNe. **C,** GRNdb. **D,** DoRothEA (N = number of regulatory genes used to generate $\overline{|R|}$ for each cancer type).

We used absolute correlation to assess the association between regulatory gene activity/mRNA abundance and sensitivity because we anticipated that both increased activity (e.g., resulting from amplification) and decreased activity (e.g., resulting from copy number loss) might result in increased sensitivity to inhibition. The former might occur with oncogene addiction-like effects, e.g. *MYC* amplification driving *MYC* sensitivity, while the latter might occur with haploinsufficiency-like effects, e.g. reduced copy number

or expression/activity of a gene making cells more sensitive to further perturbation of that gene (Nijhawan et al. 2012; Paolella et al. 2017).

Correlations were only calculated for genes that were 1) identified as regulatory genes in the GRN and 2) identified as essential in a subset of cell lines from the cancer type assessed i.e., after excluding genes that are always or never essential (see Methods). A consequence of these criteria is that different GRN methods were evaluated over different gene sets, because they include different regulatory genes (e.g., the ARACNe breast cancer network contains 6,054 regulatory genes, while DoRothEA only contains a total of 271 regulatory genes).

Across all cancer types, the six activity scores yielded an $\overline{|R|}$ between 0.12 and 0.28 for GRNdb, ARACNe and DoRothEA (Figure 2.2B, C, D). Across all cancer types and activity estimation methods, GRNdb shows the highest average correlation ($\overline{|R|}$ = 0.189), followed by DoRothEA (0.185) and ARACNe (0.156). However, the genes included in each GRN varied significantly – between 628 and 938 genes assessed for ARACNe, between 49 and 81 for GRNdb, and between 43 to 71 for DoRothEA (Figure 2.2B, C, D). Summarising over all regulon sources and cancer types, Consensus had the highest correlation with gene sensitivity, with an $\overline{|R|}$ of 0.181. W. Sum and W. Mean performed identically and were jointly the lowest performing of the methods ($\overline{|R|}$ = 0.171).

Although there are differences in $\overline{|R|}$ between the different activity estimation methods and the different GRN sources, visual inspection of the results in Figure 2.2 suggests that cancer type may have a much bigger influence than either activity method or GRN source. For instance, although there is variation between the $\overline{|R|}$ calculated with different activity estimation methods using ARACNe regulons in AML (range 0.18 - 0.2) there is a much bigger difference between the $\overline{|R|}$ calculated in AML and GBM (median 0.195 for AML; 0.14 for GBM). To understand the relative contributions of different factors to the variability in $\overline{|R|}$, we fit a linear model with three terms: cancer type, regulon source (ARACNe, GRNdb, DoRothEA) and activity method (Consensus, VIPER, etc.). Our results suggest that cancer type does indeed

explain 51% of the variance in the model whilst regulon source and activity estimation method explain much less of the variance: 21% and < 0.01%, respectively (Supplementary Figure S2.1). Thus, cancer type has a bigger influence than the regulon source, and the activity estimation method shows no consistent contribution.

It is reasonable to ask why cancer type has such a big influence – why would AML cell lines have higher average correlations between activity and gene sensitivity than HNSC cell lines? We reasoned that there might be two explanations: 1) the different numbers of cell lines for each cancer type (ranging from 24 to 51) may result in different distributions of correlations and 2) there may be more transcriptomic diversity in the cell lines from different cancer types. The latter might occur if the cancer in question has more intrinsic heterogeneity or simply if the cell lines available cover more diverse subtypes. We found that there is a strong correlation between $\overline{|R|}$ and the number of cell lines used for our analysis (Pearson's R = -0.89, p < 0.01). In fact, adding the number of cell lines as a variable in the linear model, shows that it explains 29% of the variance when cancer type is excluded, but does not explain any additional variance to cancer type. We find that variance in mRNA abundance together with number of cell lines explained 44% of the variance in the linear model, which is 86% of the variance explained by cancer type (Supplementary Figure S2.1). This suggests the majority of the variance explained by cancer type is in fact explained by the number of cell lines analysed and the variance in mRNA abundance of these cell lines, with unknown factors contributing ~7% in the model.

Regulons vary greatly in the number of targets they regulate: from 10 to 386 in DoRothEA regulons and between 1 and 2,103 in GRNdb regulons. Additionally, a target gene can be regulated by multiple regulons. These factors may affect the accuracy with which a regulatory gene's activity is calculated due to the increased complexity of a regulon. To address this problem, we stratified our analysis based on the regulon size and on the number of unique genes regulated by each regulatory gene. Our results suggest that the size of the regulons tested or the number of targets they regulate do not seem to be associated with higher or lower correlations with gene essentiality using either

the GRNdb or DoRothEA regulons (Supplementary Figure S2.2, S2.3). Including regulon size and number of unique targets in our linear model analysis shows that both variables explain very little variance compared to the other factors previously identified. Regulon size explains 6% of the variance and the percentage of unique targets that a regulon regulates does not explain any of variance in the model (Supplementary Figure S2.4A, B).

2.5.2 Regulons convey cancer type-specific information in relation to gene sensitivity to inhibition

As noted, GRNs for ARACNe and GRNdb are cancer type-specific. We wished to assess whether cancer type-matched GRNs were more informative for predicting gene sensitivity than cancer type-mismatched GRNs. For each cancer type, we ran decoupleR with the cancer type-matched regulons as well as with the nine regulons from the other cancer types (Figure 2.1A).

Our results suggest that, on average, for all regulon sources and activity estimation methods, except for MLM, cancer type-matched regulons result in a higher absolute correlation between activity and sensitivity than cancer type-mismatched ones (Unpaired Two-Samples Wilcoxon Test p-value < 0.01) (Figure 2.3A, B, Supplementary Table S2.1). Although cancer type-matched regulons were inferred from patient samples and tested in cell lines, our results suggest that tissue-specific regulon interactions are more relevant, as previously suggested (Garcia-Alonso et al. 2018), thereby improving inference of regulatory gene activity and correlation with gene sensitivity in the DepMap. However, despite cancer type-matched regulons performing better than cancer type-mismatched ones, the correlation between activity and gene sensitivity is still relatively poor on average ($\overline{|R|}$ < 0.28).

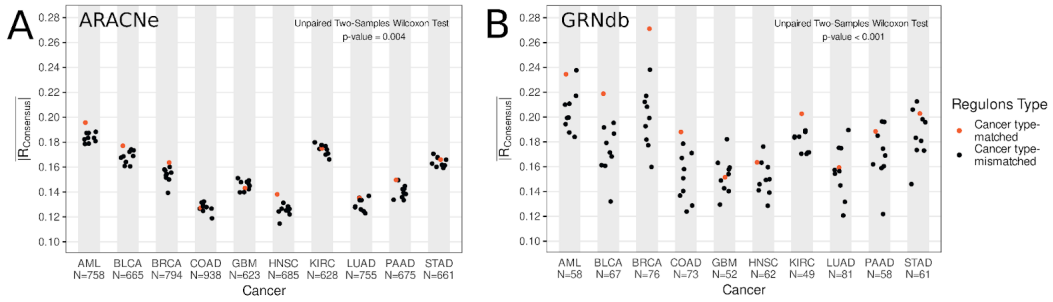


Figure 2.3 | Cancer type-matched regulons predict sensitivity to inhibition better than mismatched regulons. A, B, Absolute Pearson correlation between consensus activity and sensitivity for each cancer paired with every regulon. Each dot represents the average absolute Pearson's correlation coefficients between regulatory gene activity and gene sensitivity across all regulatory genes **A,** ARACNe. **B,** GRNdb. (N = number of regulatory genes used to generate $\overline{|R|}$ for each cancer type).

2.5.3 Gene sensitivity to inhibition is better predicted by mRNA abundance than by GRN-inferred activity

We have so far discussed the correlation between regulatory gene activity and gene sensitivity. We have slightly touched on the simpler approach of just using mRNA abundances to predict sensitivity to inhibition. Such a comparison is important for understanding whether the activity estimation methods provide an advantage for predicting gene sensitivity over plain transcript abundance.

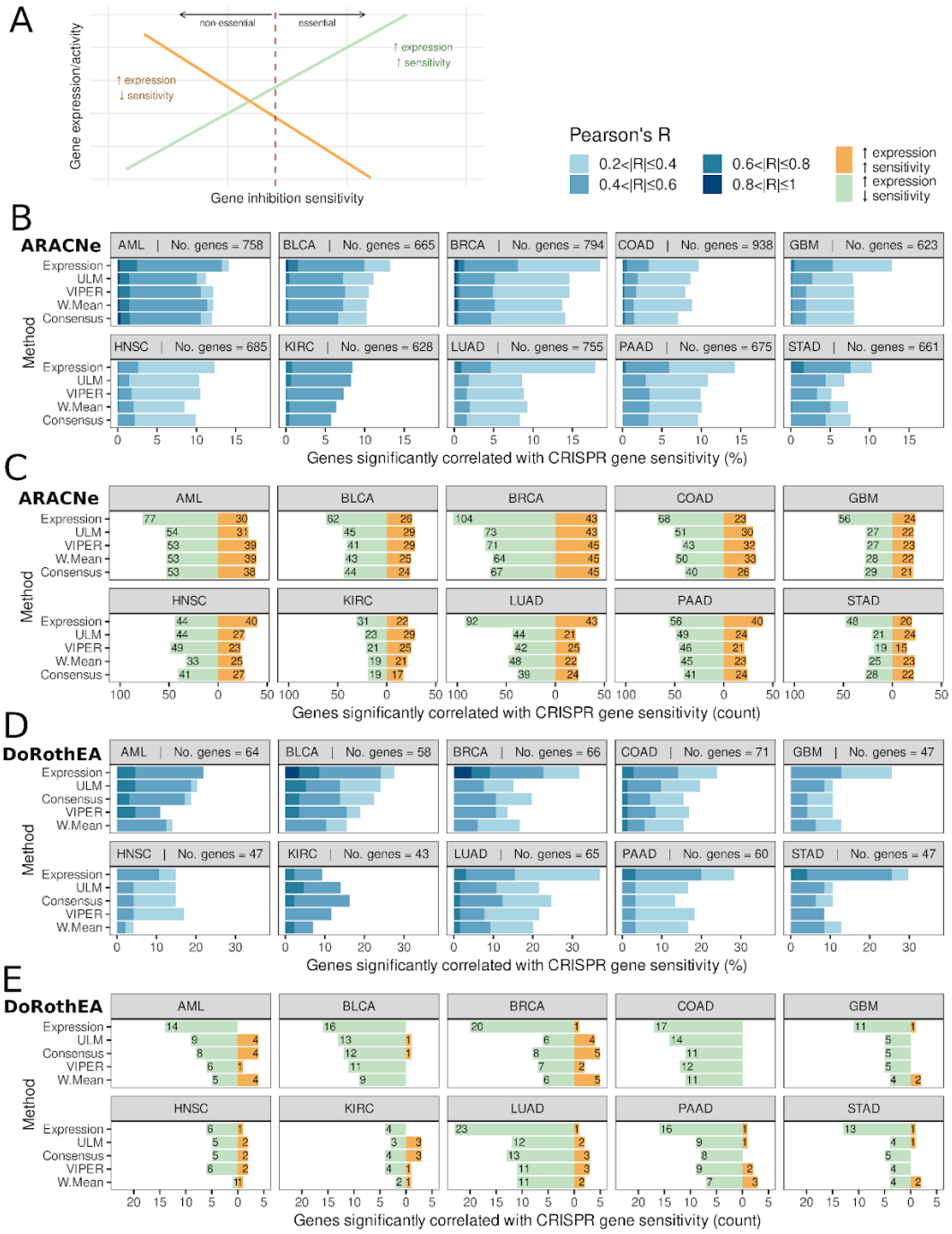
Visual inspection of Figure 2.2B, C, D suggests that mRNA abundance has a higher correlation with gene sensitivity to inhibition than any of the gene activity estimation methods. This is true across all cancer types analysed, across all activity estimation methods, and across all GRN sources (Figure 2.2B, C, D). While direct comparison of the correlations between regulons from different sources (e.g., ARACNe vs DoRothEA) is challenging due to coverage of different gene sets by each regulon source, this is not the case when comparing the activity estimation methods to mRNA abundance. When comparing the average correlation of activity estimation methods from ARACNe regulons to mRNA abundance we did so over the same set of genes.

We compared the $\overline{|R|}$ derived from mRNA abundance with that derived from Consensus (the best performing individual activity estimation method) across all regulon sources and all cancer types. We found that mRNA

abundance had a significantly higher $\overline{|R|}$ (Wilcoxon paired test p-value = 6.9×10^{-6}). Overall, this suggests that the average correlation between mRNA abundance and sensitivity to inhibition is higher than that for any of the inferred activity methods using any of the GRNs.

However, for the purpose of identifying new therapeutic targets, strong correlations are more important – those that are highly predictive of gene sensitivity. We therefore compared the proportion of genes that show significant correlations between sensitivity and activity to those with significant correlations between sensitivity and mRNA abundance. We found that strong correlations with gene sensitivity ($p < 0.05$, $|R| > 0.2$) were rare. Across all cancer types, neither expression, nor activity had strong correlations with more than 20% of genes (Figure 2.4B, D, Supplementary Figure S2.5A). For nine out of ten cancer types we studied, more genes had a strong correlation between their mRNA abundance and sensitivity than their activity and sensitivity (Figure 2.4B, D). KIRC, which has the fewest number of cell lines, was the only cancer type where inferred activity was comparable. Expression consistently had more high correlations than activity across all GRN sources and all activity estimation methods (Figure 2.4B, D, Supplementary Figure S2.5A). The same trend is evident if the threshold for strong correlations is set at ($p < 0.05$, $\overline{|R|} > 0.4$) or ($p < 0.05$, $\overline{|R|} > 0.6$). This suggests that, regardless of the exact threshold used to define a strong correlation, gene expression displays more strong correlations with gene sensitivity. Neither GRNdb-inferred activity, nor DoRothEA-inferred activity perform better than mRNA abundance, confirming the findings of ARACNe-inferred activity (Figure 2.4D, Supplementary Figure S2.5A). Additionally, we looked at genes with $|R| > 0.8$ and across all cancer types. At this threshold we discovered two genes whose sensitivity to inhibition is correlated with inferred activity, but not by mRNA abundance: *FOXA1* by both the ARACNe and GRNdb regulons in BRCA and *KLF1* by ARACNe regulons in AML. However, when we investigate the correlations between the expression of these genes and their sensitivity to inhibition, whilst the threshold of $|R| > 0.8$ is not reached, we see that their correlation is still extremely high ($|R| > 0.7$ in all cases) (Supplementary Table S2.2, Supplementary Figure S2.6B, C). Rather

than being a radically distinct predictor, this suggests that the activity estimation methods are only slightly better in these few instances.



However, we wished to know whether the genes found as significant by inferred activity and plain expression were the same or different across the different regulon sources. We investigated the overlaps for genes with |R| > 0.4 and genes with |R| > 0.6 for BRCA and COAD. We note that there are no genes that are common to the three GRN methods but not identified by expression at R > 0.4 and only one gene at R > 0.6 (Supplementary Figure S2.5C, D, E, F). Additionally, there are very few regulatory genes which overlap in any two of the three methods and are not also identified by expression.

Whilst expression correlated better with sensitivity to inhibition on average, there were specific cases where inferred activity was a better predictor of sensitivity to inhibition. For example, CDX2 activity performs better in COAD: $R_{Consensus}$ = -0.68, $R_{Expression}$ = -0.49 (Supplementary Table 2). However, we did not find any consistent pattern that explained why these genes were exceptions. For example, GO enrichment analysis did not reveal any specific functional enrichment in genes for which their sensitivity to inhibition is better correlated with activity than expression. Additionally, we found no enrichment for oncogenes among these cases, according to the CGC (Sondka et al. 2018), or for master regulators, as listed by Paull et al. (2021). Additionally, we have performed GO enrichment analysis on genes with an absolute correlation between Consensus activity/expression and sensitivity to inhibition greater than 0.6 (Supplementary Table S2.3). We find the largest

---

Figure 2.4 | Gene sensitivity to inhibition correlates better with expression than with inferred activity. A, Distinction between the two types of correlations between expression/activity and gene sensitivity score. An increase in expression can be correlated with higher sensitivity (in green). An increase in expression/activity can also be associated with a decrease in sensitivity (in orange). **B, D,** Pearson's correlation coefficients between activity/expression and gene sensitivity stratified incrementally from |R| = 0.2 to 1 to show the percentage of significant regulatory genes correlated with gene sensitivity ($p < 0.05$) after filtering out genes that are never essential and genes that are always essential in a cancer. Methods are sorted top-to-bottom in order of performance across all cancer types for the GRN-inferred method in cause. **B,** ARACNe regulons. **D,** DoRothEA regulons. **C, E,** Analysis of the positive and negative correlations between activity/expression and gene sensitivity shows there are more negative correlations, suggesting there are more cases where an increase in sensitivity is associated with increased expression/activity. Analysis also confirms expression is better correlated with gene sensitivity. **C,** ARACNe regulons. **E,** DoRothEA regulons.

number of GO terms enriched when analysing genes strongly correlated with gene expression. We note that there are no GO terms enriched for genes found using DoRothEA regulons. For the GRN reconstruction methods, only one unique term is enriched among genes inferred using ARACNe regulons and eight terms are inferred using GRNdb regulons. No enriched terms are common across the two GRN methods (Supplementary Figure S2.6A).

To investigate the correlation between activity/expression and sensitivity to inhibition independently of cancer type, we performed the same analysis at a pan-cancer level, using all cell lines (n = 973). We calculated activity based on DoRothEA regulons, as they are cancer type-agnostic (Supplementary Figure 2.S7A, C, E). We found a similar trend: ~50% of sometimes-essential genes having a correlation > 0.2 between mRNA abundance and sensitivity to inhibition (Supplementary Figure S2.7A). The activity inference methods have strong correlations with fewer genes (25/90 genes with |R| > 0.2 for Consensus vs 43/90 for mRNA abundance). Additionally, expression found two extremely high correlations (|R| > 0.8), whilst activity found none.

One potential explanation for gene expression having higher absolute correlations with gene sensitivity could be that expression measurements display higher variance than activity scores. However, comparing the per gene variance across our methods shows that gene activities, as determined by VIPER and ULM, have a comparable variance to expression, while activities determined by W. Sum have a significantly higher variance. W. Mean, MLM and Consensus have a slightly lower variance than expression (Supplementary Figure S2.8).

One limitation of testing the performance of regulons derived from patient samples is that these regulons may not be representative for tumour cells only as patient samples contain a mix of tumour and non-tumor cells. Genes expressed in other cell types, but not in the tumour cells, may be a confounder for the regulon inference. To assess the impact of this, we have performed the same analysis using regulons inferred only from tumour cell lines. To build the regulons, we used the expression matrix containing all cancer cell lines of a cancer type (i.e., not just the cell lines that have both gene expression and gene essentiality in DepMap). We built ARACNe and GRNdb-like regulons (see Methods) for BRCA, COAD, LUAD and PAAD, as they are the cancer types with

the largest number of cell lines (excluding HNSC, which is highly heterogeneous due to the large number of subtypes it encompasses). Our analysis suggests that regulons inferred from the CCLE are still not suited for the task of detecting molecular vulnerabilities when paired with activity inference methods. Using ARACNe-inferred regulons we still saw that gene inhibition sensitivity was better correlated with gene expression than with any activity estimation method for BRCA, LUAD and PAAD (Supplementary Figure S2.8A). For GRNdb-like regulons we noticed ULM and Consensus performing slightly better than expression for BRCA and COAD, but not for LUAD and PAAD (Supplementary Figure S2.8B). Additionally, using ARACNe regulons there is a larger number of high correlations ($|R| > 0.6$) between sensitivity to inhibition and expression than with any of the activity methods for all four cancer types (Supplementary Figure S2.8C). For BRCA GRNdb-like regulons, we saw Consensus identifying two very high correlations (>0.8): *CTNNB1* and *FOXA1*, whilst Expression finds *GATA3* with a correlation >0.8 (Supplementary Figure S2.8E). Taken together, these results suggest that even with GRNs inferred from tumor cell lines only, the performance of activity-estimation methods is not notably better than simply using mRNA abundance.

2.5.4 Increased sensitivity to gene inhibition is more commonly correlated with increased expression, rather than decreased expression

Thus far, we have focused on the analysis of absolute correlations between activity/expression and sensitivity to inhibition. As noted previously, this is because we anticipated there may be two distinct effect types associated with different genes – sometimes increased expression/activity may be associated with increased sensitivity to inhibition, as observed for oncogene addiction effects, while in other cases reduced expression/activity may be associated with increased sensitivity (Figure 2.4A). We sought to understand which type of effect was more common, and whether there were differences between inferred activity and gene expression. We found that for gene expression there were consistently more genes where higher expression was associated with

increased gene inhibition sensitivity, as previously suggested by other studies (Figure 2.4C, E, Supplementary Figure S2.5B) (Hart et al. 2014; Tim Wang et al. 2015). This strong skew towards increased expression – increased sensitivity correlations was less evident for the activity methods, e.g., in AML, using ARACNe regulons, 58% of significant genes (53/91) showed an increase in sensitivity with increased Consensus activity. Across the same gene set 72% of significant genes (77/107) showed an increase in sensitivity with increased gene expression.

There was some variation across the different GRN inference methods, ARACNe in general was associated with a much lower proportion of increased activity – increased sensitivity correlations than GRNdb (Figure 2.4C, Supplementary Figure S2.5B). However, across both regulon sources, increased expression was consistently associated with an increase in sensitivity.

Interestingly, the use of curated regulons from DoRothEA led to very few cases where an increase in expression/activity results in a decreased sensitivity (Figure 2.4E, Supplementary Figure S2.7B, D, F). This suggests that the TFs included in DoRothEA are skewed towards those for which increased activity/expression is associated with increased inhibition sensitivity.

2.5.5 Expression better predicts binary essentiality

So far, we have analysed gene inhibition sensitivity from CRISPR screens as a quantitative trait. However, in many cases the results of CRISPR screens are binarized, such that genes are deemed to be either essential or non-essential for survival (Hart et al. 2014; De Kegel & Ryan 2019; Vinceti et al. 2021). Genes which are essential in a specific context might then be considered as suitable therapeutic targets.

To assess the ability of gene activity and gene expression to predict binary essentiality, in each cell line we separated genes into two groups: essential and non-essential (see Methods) (Figure 2.1B). We then compared the ability of expression and inferred activity to separate the two groups using a Wilcoxon test and the CLES. The interpretation of the CLES is equivalent to the area under the receiver operating characteristic curve (AUC ROC) often used to

evaluate binary classifiers. The CLES represents the probability that a gene sampled at random from the essential group will have a higher gene activity/expression than a gene sampled at random from the non-essential group (McGraw & Wong 1996). We consider that a gene's essentiality can be well predicted by expression/activity if CLES > 0.7 and p-value < 0.05.

Our results suggest that, on average, gene expression better predicts binary essentiality, irrespective of whether ARACNe, GRNdb or DoRothEA regulons were used (Figure 2.5A, B, C, Supplementary Table S2.4). In 20 of 30 cases across all regulon types (ten cancer types × three regulon sources), more genes have a CLES > 0.7 when their essentiality is predicted using expression, rather than activity. The same is true for different thresholds – for CLES > 0.8 and > 0.9, expression still predicts more essential genes overall than any of the activity estimation methods using any of the regulon sources. Similarly, on the DoRothEA pan-cancer analysis we found that expression finds ~51% of sometimes-essential genes with a CLES > 0.7 and ~32% of genes with a CLES > 0.9. Consensus, the best performing activity method finds ~28% genes with a CLES > 0.7 (Supplementary Figure S2.9). Thus, gene expression, rather than inferred activity, is a better predictor of binary gene essentiality.

2.6 Discussion

Our systematic analysis suggests that gene expression performs better than GRN-inferred activity at predicting sensitivity to CRISPR gene inhibition in cancer. This is true regardless of the GRN source and activity inference method used and whether essentiality is treated as a binary or quantitative trait. Whilst extensively used to find "master regulators" of cancer (Garcia-Alonso et al. 2019), regulatory gene activity does not outperform gene expression for the task of predicting gene sensitivity to inhibition. Across ten cancer types and at a pan-cancer level, more genes are found to have a strong correlation between sensitivity and mRNA abundance than they do between sensitivity and inferred activity.

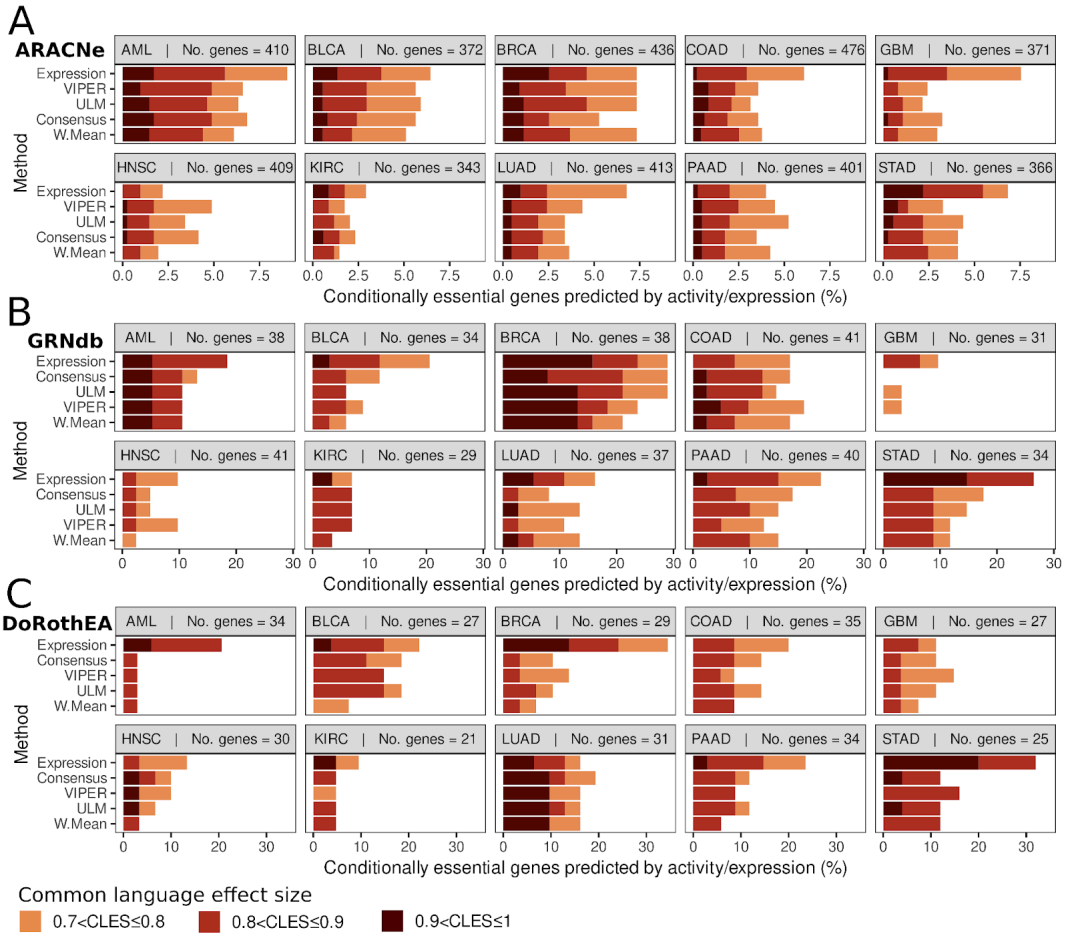


Figure 2.5 | Gene essentiality correlates better with expression than with inferred activity. A, B, C, CLES between activity/expression and binary gene essentiality stratified incrementally from CLES = 0.7 to 1 to show the percentage of significant conditionally essential genes predicted by activity/expression ($p < 0.05$) after filtering out genes that are not essential and genes that are essential in less than three cell lines in a cancer type. Methods are sorted top-to-bottom in order of performance across all cancer types for the GRN-inferred method in cause. **A,** ARACNe regulons. **B,** GRNdb regulons. **C,** DoRothEA regulons.

We find that matched regulons may be more accurate in describing the regulatory gene activity landscape of each cancer type than mismatched regulons. This suggests that there is value to using co-expression information from relevant patient tumours to build GRNs. Additionally, this suggests a degree of similarity between primary patient tumours and cell-line models that can be captured by GRNs. Our study also suggests that cancer type contributes more to the variance in average correlation with gene sensitivity than the GRN-building method or the activity-inference method. This may be at least

partially attributable to different cancer types having more variable transcriptomes. We find that there is no significant difference between activity and expression in finding correlations where an increase in sensitivity is associated with decreased expression/activity. However, in all cancer types, there are more genes where an increase in expression, rather than activity, is associated with an increase in sensitivity, in an oncogene addiction-like effect. This is expected to be the case in most cancer cells reliant on the activity of a TF for survival.

Both inferred and curated networks (Garcia-Alonso et al. 2019) have been used to infer the activity of TFs and to investigate the differential activity of TFs in different conditions (Du et al. 2018; Garcia-Alonso et al. 2018). The assumption behind these approaches is that the activity of a TF inferred from the expression profile of its targets can help uncover hidden potential therapeutic targets. Therefore, these methods have been used as hypothesis-creation tools to find novel targets (Garcia-Alonso et al. 2018) or associations with patient survival (Falco et al. 2016). A significant limitation of TF activity estimation approaches is that typically only a small number of candidate targets are selected for experimental testing (Alvarez et al. 2016). It is thus extremely challenging to understand how broadly useful these approaches are, and to estimate false positive or false negative rates.

We propose a computational approach based on CRISPR screen data to assess the ability of GRN-inferred activity to predict sensitivity to perturbation in tumor cell lines. The repository of cell lines being screened grows every year, offering more statistical power (Dempster et al. 2021; Dempster et al. 2019; Meyers et al. 2017; Pacini et al. 2021). A significant advantage of our approach is that it is unbiased, in the sense that all genes are evaluated, rather than one or two selected candidates. Evaluating only one or two candidates may lead to a false sense of accuracy of the approach, downplaying its limitations. A limitation of our approach is that we are evaluating a downstream use of inferred GRNs rather than the GRNs themselves, i.e., we have not evaluated the ability of the reconstructed networks to predict transcriptional changes, but rather their ability to predict therapeutic targets. However, the latter is a purpose for which they are often employed.

Our study is primarily limited by the availability of the data, as there is a limited number of cell lines with genome-wide CRISPR screens data. We tried to mitigate this by selecting cancer types with a large number of cell lines screened. However, we see the number of cell lines used contributes 29% to the variance in, suggesting the results from cancer types with a low number of cell lines available (i.e., STAD, AML) may be less reliable in our analysis. Furthermore, the gene sensitivity measurements we use are made in cell line models. These might not entirely reflect cancer cells within actual tumours, surrounded by the tumour microenvironment. Additionally, our results are reflective of cohorts of cell lines displaying a range of regulatory landscapes. A better approach might be to integrate multiple sources of data for model-specific GRNs, as done by Goode et al. (2016) and Assi et al. (2019). They create GRNs from multiple omics sources that are cell line-specific.

We have assessed the ability of GRNs to predict gene essentiality as measured in CRISPR screens. Other methodologies, such as RNA interference (RNAi), have also been used to assess gene essentiality in large numbers of cell lines (Tsherniak et al. 2017; Campbell et al. 2016). While in general CRISPR screens appear better at identifying essential genes than RNAi-based methods (Hart et al. 2015; Smith et al. 2017), there is also evidence that combining CRISPR-based gene essentiality scores with RNAi-based scores can be especially informative for understanding variation in pan-essential genes (Wang et al. 2019; Krill-Burger et al. 2023). Therefore, there may be value in assessing the ability of GRNs to predict gene essentiality derived from RNAi.

Ultimately, our study primarily looks at correlations between GRN-inferred activity and sensitivity and future work could explore indirect relationships between the GRNs and sensitivity to inhibition. For instance, Garcia-Alonso et al. (2018) explores the relationships between drug sensitivity and the activity of indirect targets, finding correlations with clinical significance. However, despite these caveats, expression consistently outperforms GRN-inferred activity in predicting sensitivity to CRISPR inhibition in a variety of cancer types. This work underlines the utility of sensitivity data from CRISPR screens in benchmarking the use of GRN-inferred activity methods for nominating therapeutic targets.

2.7 Acknowledgements

We thank Dr. Christina Kiel, Dr. Luis Iglesias Martinez and members of the Bond and Ryan labs for useful suggestions on the analysis. We thank Philip Cotter for technical support on maintaining the server on which many of the scripts were run.

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

2.8 Data availability

R code to run analyses is available at https://github.com/cancergenetics/GRN_activity_corr_essentiality and https://doi.org/10.5281/zenodo.8256519.

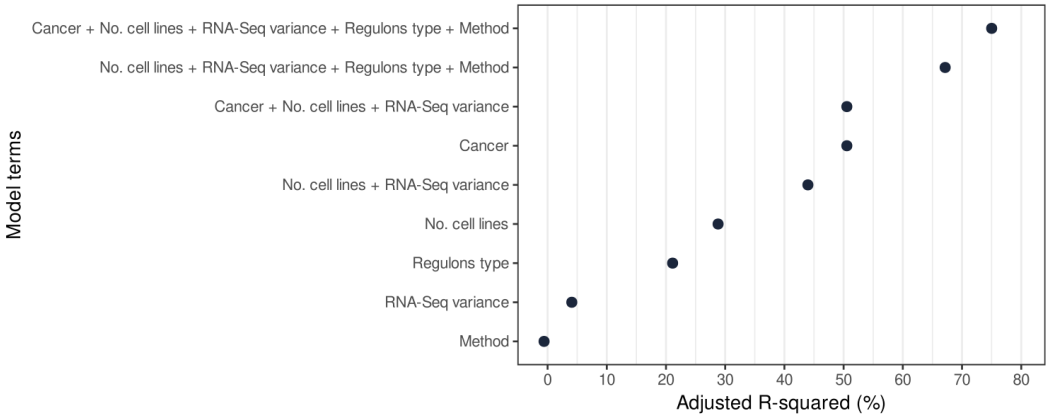
Correlations between gene essentiality and inferred activity/expression are available on figshare via DOI: 10.6084/m9.figshare.23858484.

2.9 Funding

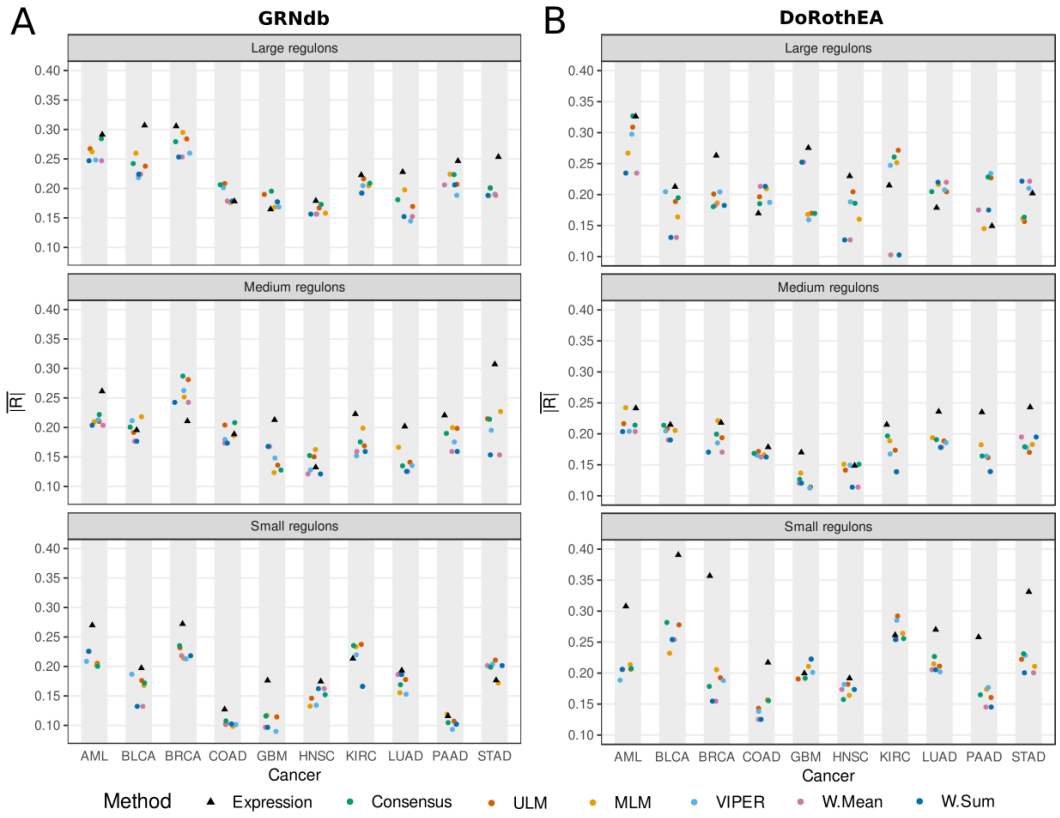
Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science [18/CRT/6214]; EU's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie [H2020-MSCA-COFUND-2019-945385, in part]; work in the Bond laboratory is supported by Science Foundation Ireland [20/FFP-P/8844, 18/SPP/3522]; the latter together with Children's Health Ireland; work in the Ryan laboratory is supported by Science Foundation Ireland [20/FFP-P/8641].

Conflict of interest statement. None declared.

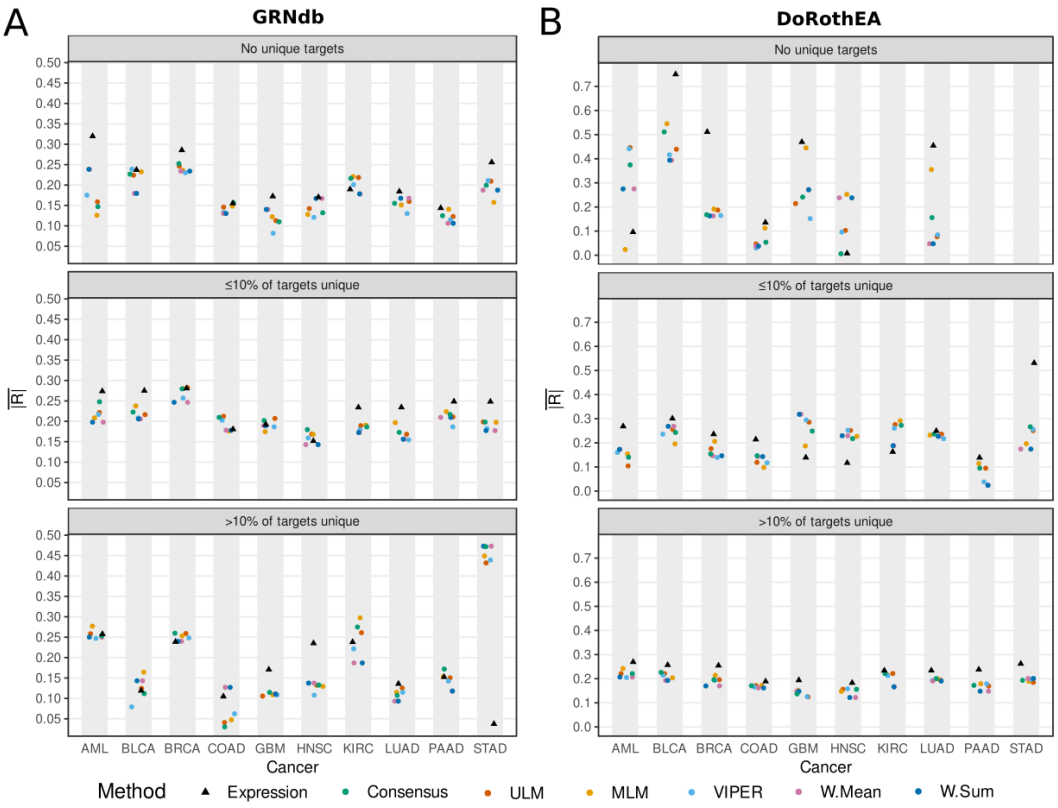
2.10 Supplementary Data



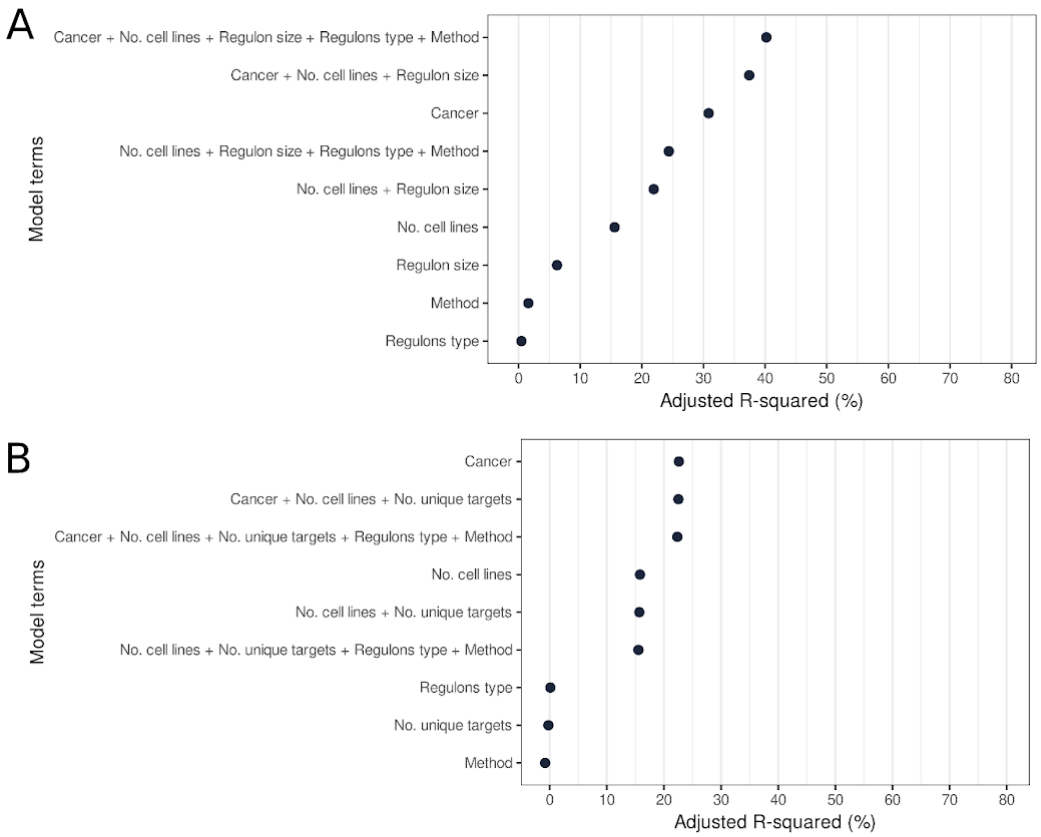
Supplementary Figure S2.1 | Each terms' contribution to linear model predicting |R|. Each dot represents the percentage of variance explained (Adjusted R-squared) by each variable in the linear model predicting the absolute correlation between essentiality and activity (|Pearson's R|).



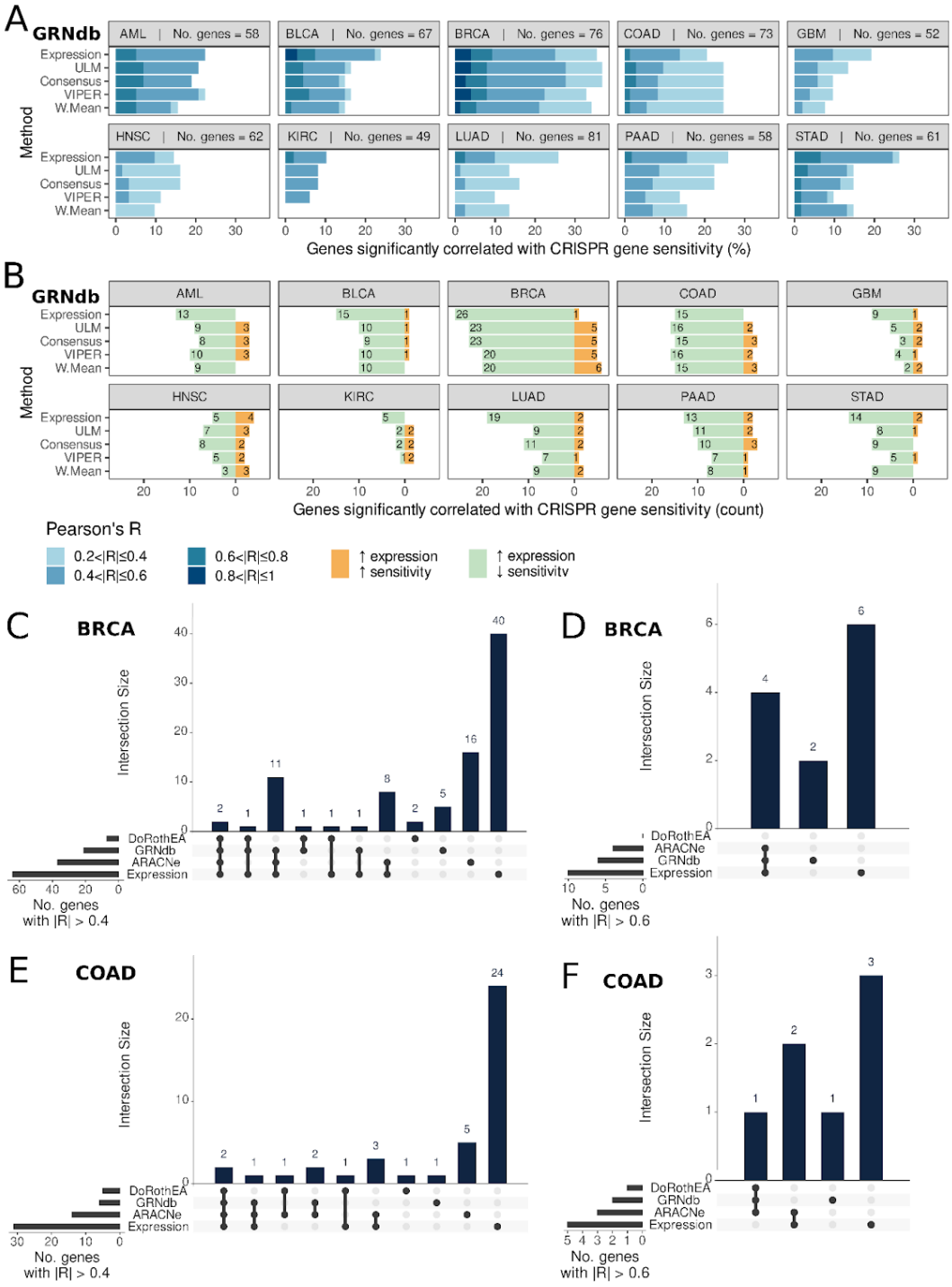
Supplementary Figure S2.2 | Differences between activity and expression correlations with gene sensitivity to inhibition are similar independent of regulon size A, B, Comparison between the different inferred activity methods (paired with cancer type-matched regulons) correlating with gene sensitivity and gene expression correlating with gene sensitivity, by regulon size. **A,** GRNdb. **B,** DoRothEA.



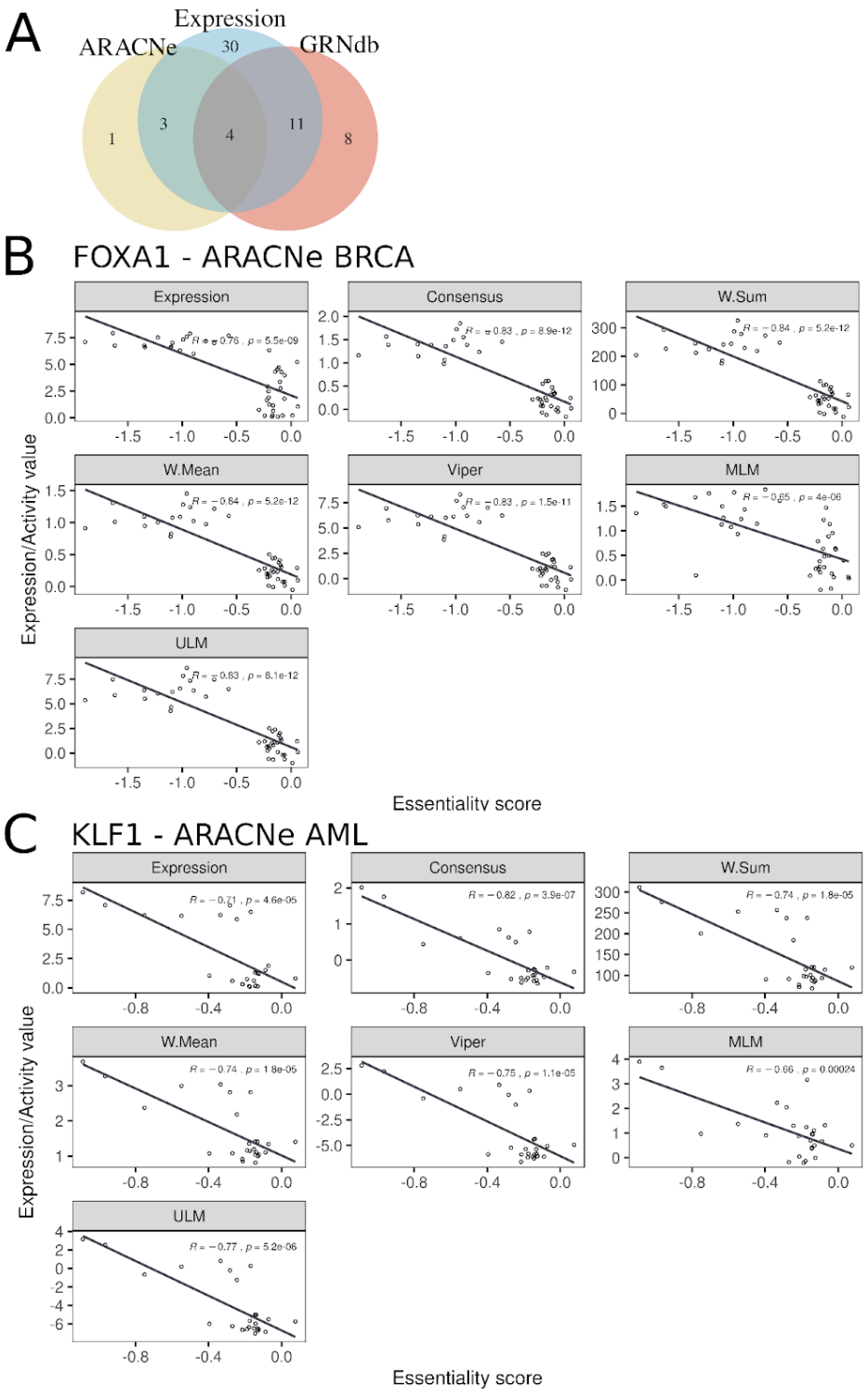
Supplementary Figure S2.3 | Differences between activity and expression correlations with gene sensitivity to inhibition are similar independent of the number of unique targets each regulon has. A, B, Comparison between the different inferred activity methods (paired with cancer type-matched regulons) correlating with gene sensitivity and gene expression correlating with gene sensitivity, by regulon size. **A,** GRNdb **B,** DoRothEA.



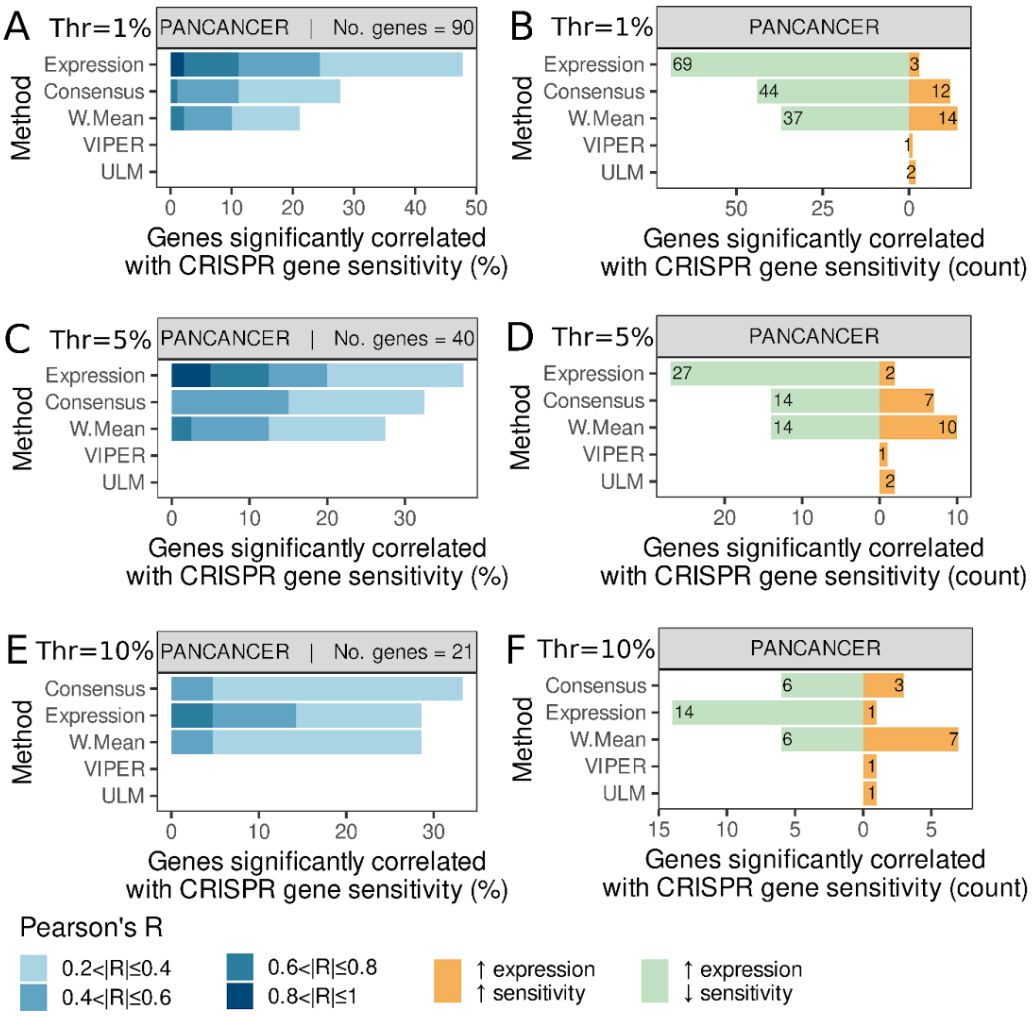
Supplementary Figure S2.4 | Each terms' contribution to linear model predicting |R|. A, Using regulon size to stratify. **B,** Using the number of unique targets each regulon regulates. Each dot represents the percentage of variance explained (Adjusted R-squared) by each variable in the linear model predicting the absolute correlation between essentiality and activity (|Pearson's R|). Note: the percentages are different from Figure 2.1A, as here the value for |R| is calculated over the stratification variable (i. e., regulon size or number of unique targets) and we only used GRNdb and DoRothEA regulons.


A
GRNdb
AML | No. genes = 58
BLCA | No. genes = 67
BRCA | No. genes = 76
COAD | No. genes = 73
GBM | No. genes = 52
HNSC | No. genes = 62
KIRC | No. genes = 49
LUAD | No. genes = 81
PAAD | No. genes = 58
STAD | No. genes = 61
Method
Expression
ULM
Consensus
VIPER
W.Mean
Genes significantly correlated with CRISPR gene sensitivity (%)
B
Genes significantly correlated with CRISPR gene sensitivity (count)
Pearson's R
0.2<|R|≤0.4
0.4<|R|≤0.6
0.6<|R|≤0.8
0.8<|R|≤1
↑ expression
↑ sensitivity
↑ expression
↓ sensitivity
C
BRCA
Intersection Size
DoRothEA
GRNdb
ARACNe
Expression
No. genes
with |R| > 0.4
D
BRCA
No. genes
with |R| > 0.6
E
COAD
No. genes
with |R| > 0.4
F
COAD
No. genes
with |R| > 0.6

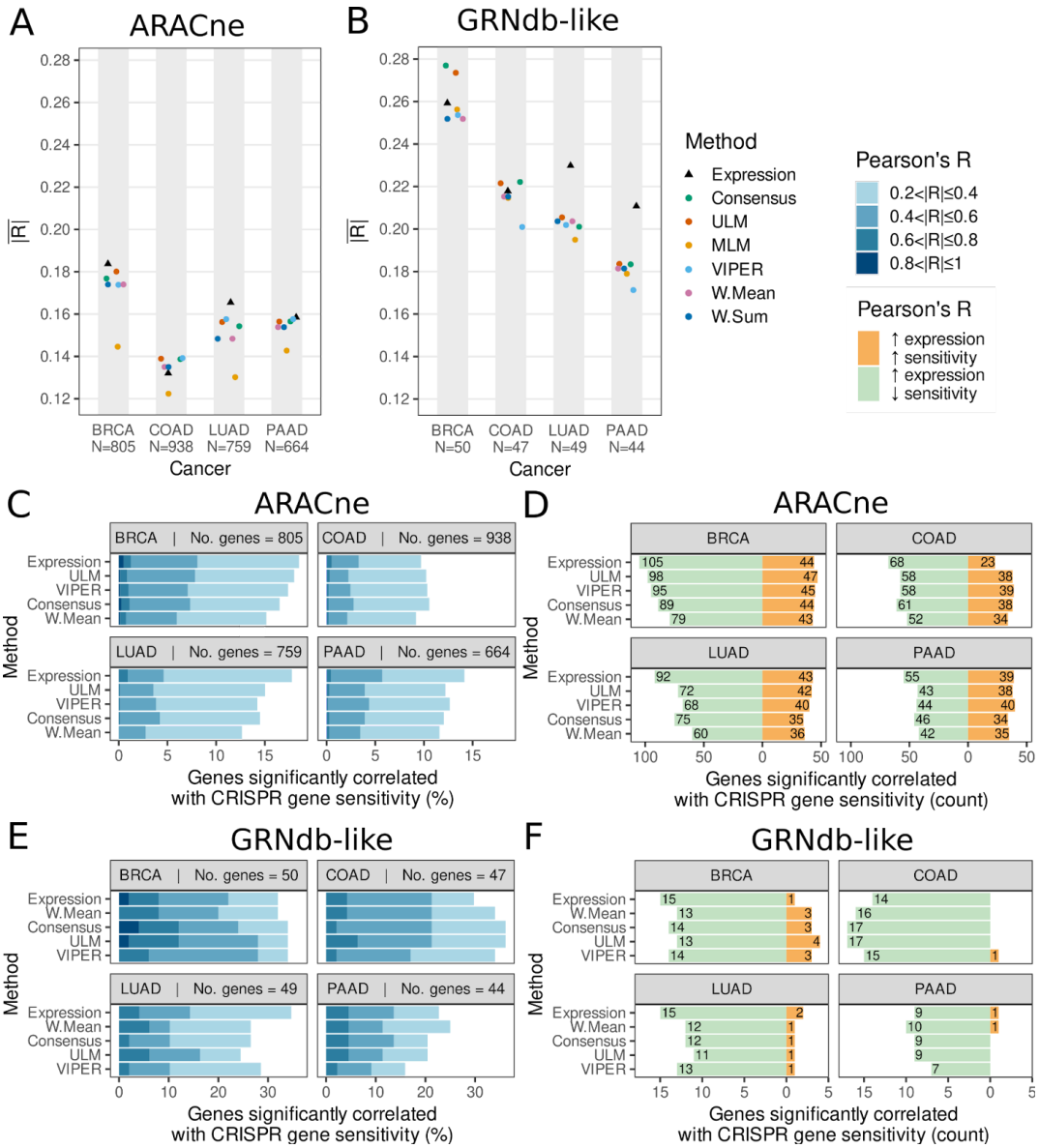
Supplementary Figure S2.5 | Gene essentiality correlates better with expression than with inferred activity using GRNdb regulons. A, Pearson's correlation coefficients between activity/expression and gene sensitivity stratified incrementally from |R| = 0.2 to 1 to show the percentage of significant regulatory genes correlated with gene sensitivity ($p < 0.05$) after filtering out genes that are never essential and genes that are always essential in a cancer type. Methods are sorted top-to-bottom in order of performance across all cancer types for the GRN-inferred method in cause. **B,** Analysis of the high expression – high sensitivity and high expression low sensitivity correlations between activity/expression and gene essentiality shows there are more cases where an increase in sensitivity is associated with increased expression/activity. **C, D,** Overlap of significantly correlated genes between expression/activity and sensitivity to inhibition using GRNdb, ARACNe and DoRothEA regulons in BRCA. **C,** |R| > 0.4, **D,** |R| > 0.6. **E, F,** Overlap of significantly correlated genes between expression/activity and sensitivity to inhibition using GRNdb, ARACNe and DoRothEA regulons in COAD. **E,** |R| > 0.4, F, |R| > 0.6.



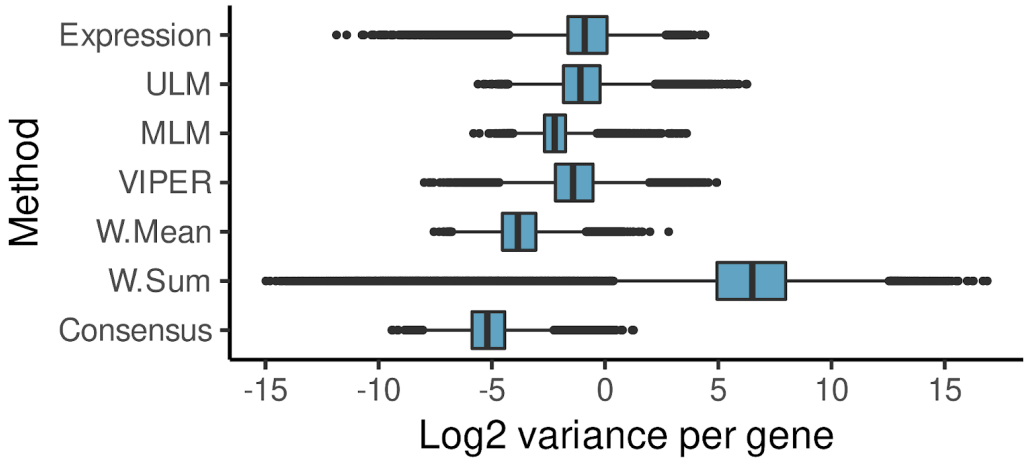
Supplementary Figure S2.6. | GO enrichment over GRN methods and individual correlations with |R| > 0.6. A, Overlap between GO terms found as enriched in genes with a correlation > 0.6 between Consensus activity/expression and sensitivity to inhibition. **B, C,** Scatter plots and correlations between inferred activity/expression for individual genes **B,** *FOXA1* using ARACNe GRNs in BRCA and **C,** *KLF1* using ARACNe GRNs in AML.



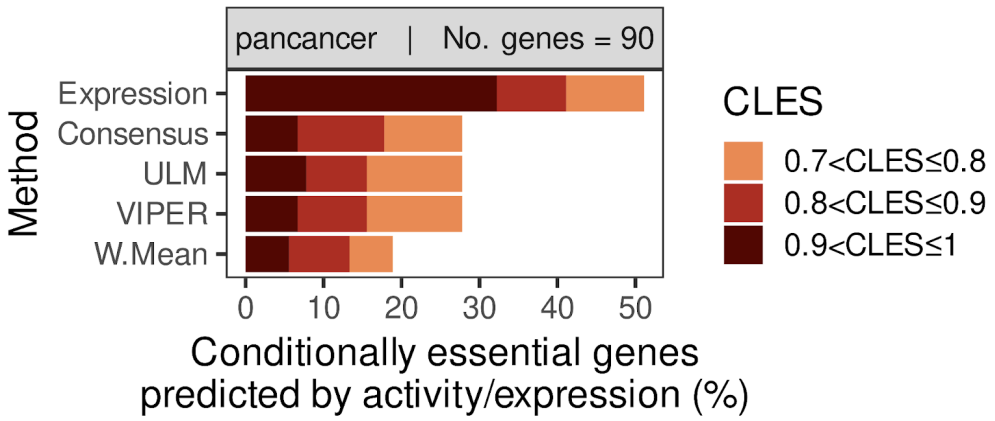
Supplementary Figure S2.7 | Gene essentiality correlates better with expression than with inferred activity using literature curated DorothEA regulons in a pan-cancer analysis regardless of the threshold used to call a gene 'sometimes' essential (see Methods). A, C, E, Pearson's correlation coefficients between activity/expression and gene sensitivity stratified incrementally from |R| = 0.2 to 1 to show the percentage of significant regulatory genes correlated with gene sensitivity ($p < 0.05$) after filtering out genes that are never essential and genes that are always essential in a cancer. Methods are sorted top-to-bottom in order of performance. A, Threshold = 1%. **C,** Threshold = 5%. **E,** Threshold = 10%. **B, D, F,** Analysis of the high expression – high sensitivity and high expression low sensitivity correlations between activity/expression and gene essentiality shows there are more cases where an increase in sensitivity is associated with increased expression/activity. **B,** Threshold = 1%. **D,** Threshold = 5%. **F,** Threshold = 10%.



Supplementary Figure S2.8 | Correlations between sensitivity to inhibition and GRN-inferred activity/mRNA abundance using regulons inferred from the CCLE. **A, B,** Comparison between the different inferred activity methods (paired with cancer type-matched regulons) correlating with gene sensitivity and gene expression correlating with gene sensitivity. **C, E,** Pearson's correlation coefficients between activity/expression and gene sensitivity stratified incrementally from |R| = 0.2 to 1 to show the percentage of significant regulatory genes correlated with gene sensitivity ($p < 0.05$) after filtering out genes that are never essential and genes that are always essential in a cancer. Methods are sorted top-to-bottom in order of performance across all cancer types for the GRN-inferred method in cause. **D, F,** Analysis of the high expression – high sensitivity and high expression low sensitivity correlations between activity/expression and gene essentiality shows there are more cases where an increase in sensitivity is associated with increased expression/activity. **A, C, E,** ARACNe regulons. **B, D, F,** GRNdb-like regulons.



Supplementary Figure S2.9 | Per gene variance for each method, across all regulon sources. Boxplots showing the log2 variance for each activity method and expression. Black line shows median; blue box represents the interquartile range; black dots show outliers.



Supplementary Figure S2.10 | Gene essentiality correlates better with expression than with inferred activity using DoRothEA regulons across in a pan-cancer analysis. CLES between activity/expression and binary gene essentiality stratified incrementally from CLES = 0.7 to 1 to show the percentage of significant conditionally essential genes predicted by activity/expression ($p < 0.05$) after filtering out genes that are not essential and genes that are essential in less than three cell lines in a cancer type. Methods are sorted top-to-bottom in order of performance across all cancer types for the GRN-inferred method in cause.

Supplementary Table S2.1 | P-value table of unpaired two-samples Wilcoxon test comparing the ranks of the activity vs essentiality absolute correlation between cancer type-matched and cancer type-mismatched regulons for each activity method. (Same as in Figure 2.3A, B).

Unpaired Two-Samples Wilcoxon Test p-values

Method	ARACNe	GRNdb
Consensus	0.004	<0.001
ULM	0.006	<0.001
MLM	0.206	0.002
W. Mean	0.008	<0.001
W.Sum	0.008	<0.001
VIPER	0.011	0.003

Supplementary Table S2.2 | Pearson's correlations coefficients and p-values for correlations between activity/expression and sensitivity to inhibition for ARACNe, DoRothEA and GRNdb, ARACNe CCLE and GRNdb-like.

Supplementary Table S2.3 | Enriched Gene Ontology analysis terms for genes with a correlation > 0.6 for ARACNe, GRNdb and mRNA expression.

Supplementary Table S2.4 | CLES coefficients and Wilcoxon test p-values for testing conditionally essential genes for ARACNe, GRNdb and DoRothEA.

Available at: https://bit.ly/suppltableschapter2

CHAPTER 3 - EZH2 loss leads to priming and partial activation of alternative lineage transcriptional programs in acute myeloid leukaemia

Cosmin Tudose[1,2,3], Luke Jones[1,2], Theodora Grosu[1,2], Marie-Claire Fitzgerald[1,2], Noura Maziak[4,5], Rebecca Ling[6], Anindita Roy[6,7,8], Juan M. Vaquerizas[4,5], Colm J. Ryan[1,2,9,10], Jonathan Bond[1,2,11]

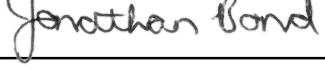
[1]Systems Biology Ireland, University College Dublin, Dublin, Ireland
[2]School of Medicine, University College Dublin, Dublin, Ireland
[3]The SFI Centre for Research Training in Genomics Data Science
[4]MRC London Institute of Medical Sciences, London, United Kingdom
[5]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital 8 Campus, London, United Kingdom
[6]Department of Paediatrics, University of Oxford, Oxford, United Kingdom
[7]MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom
[8]Department of Haematology, Great Ormond Street Hospital for Children, London, United Kingdom
[9]School of Computer Science, University College Dublin, Dublin, Ireland
[10]UCD Conway Institute, University College Dublin, Dublin, Ireland
[11]Children's Health Ireland at Crumlin, Dublin, Ireland
Corresponding Author: Jonathan Bond, E-mail: jonathan.bond@ucd.ie

Author's contributions:

Conceptualization: CT, LJ, CJR, JB. Methodology and Validation: CT, LJ, TG, MCF, RL. Formal analysis: CT, JMV, CJR, JB. Investigation: CT, LJ, TG, MCF, NM, RL. Resources: RL, AR. Data Curation: CT, LJ, TG, MCF, RL. Writing: Original Draft: CT, CJR, JB. Writing: Review & Editing: All authors. Visualisation: CT, CJR, JB. Supervision, Project administration, Funding Acquisition: CJR, JB.

3.1 Declaration of co-authorship collaboration

1. Examination Candidate Details
Examination Candidate Name: **Cosmin Tudose**
Examination Candidate UCD Student Number: **20209354**
Research Degree for which thesis is being submitted: **PhD**
Title of Research thesis: **Defining molecular vulnerabilities in childhood leukaemia through biological network analysis**
2. Details of the Paper
Title of Paper: **EZH2 loss leads to priming and partial activation of alternative lineage transcriptional programs in acute myeloid leukaemia**
Current Status of the Paper: **Drafted**
AUTHOR CONTRIBUTIONS
Cosmin Tudose's contribution to the research work described in this chapter: ● Project conceptualisation. ● Bioinformatics and statistics. ● Original draft. ● Review and editing of manuscript. ● Presentation of this research work at conferences.
Other Authors' contributions to the research work:
Luke Jones: ● Project conceptualisation. ● Data collection. ● Review and editing of manuscript.
Theodora Grosu: ● Data collection.

● Review and editing of manuscript.
Marie-Claire Fitzgerald: ● Data collection. ● Review and editing of manuscript.
Noura Maziak: ● Bioinformatics. ● Review and editing of manuscript.
Rebecca Ling: ● Data collection. ● Review and editing of manuscript.
Anindita Roy: ● Data collection. ● Review and editing of manuscript.
Juan M. Vaquerizas: ● Secondary supervision. ● Review and editing of manuscript.
Colm J. Ryan: ● Project conceptualisation. ● Primary supervision. ● Original draft. ● Review and editing of manuscript.
Jonathan Bond: ● Project conceptualisation. ● Primary supervision. ● Original draft. ● Review and editing of manuscript.
Principal Supervisor: **Jonathan Bond**
Signature: 

Date: 10.12.2024
PhD candidate: **Cosmin Tudose**
Signature: 
Date: 10.12.2024

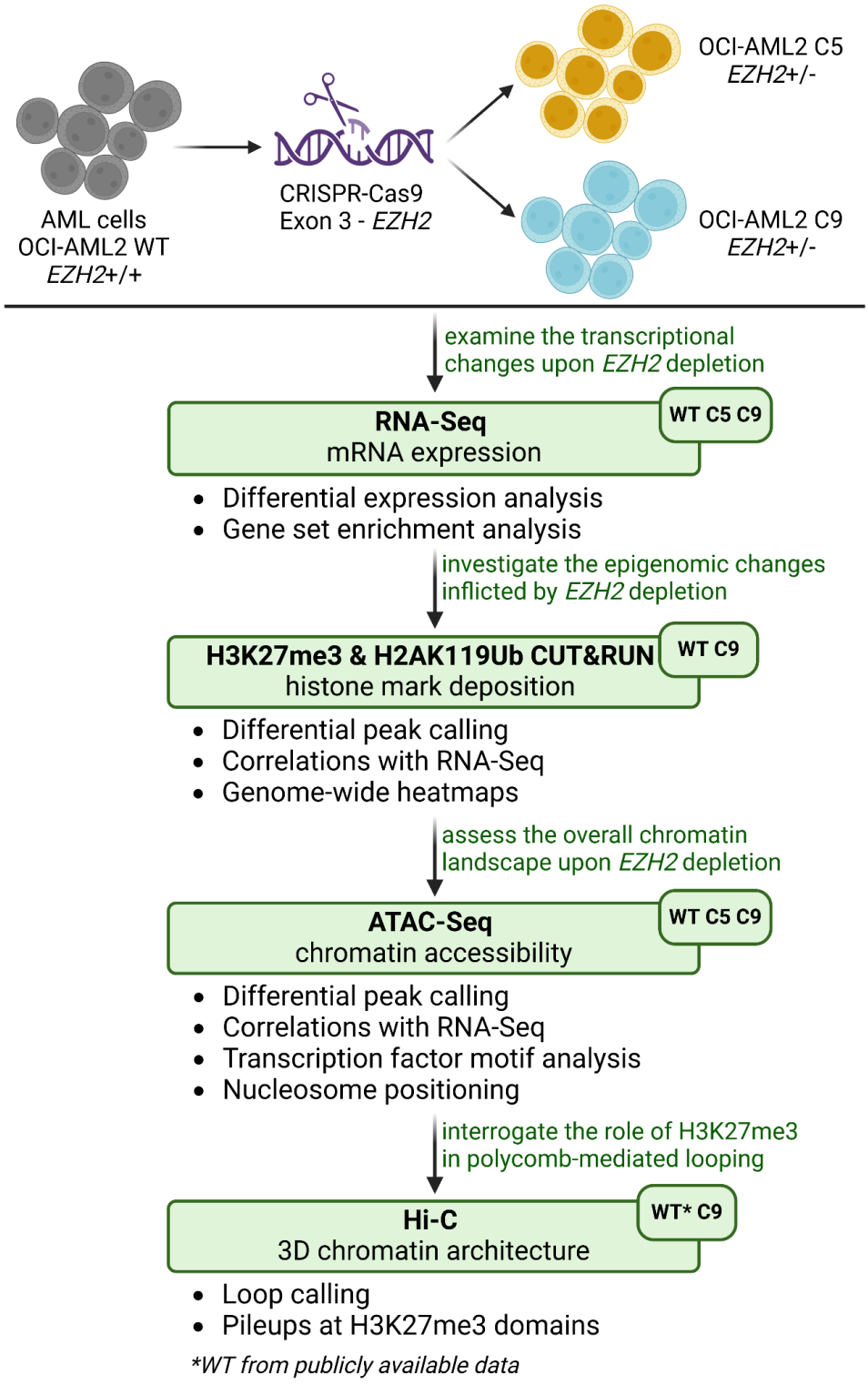


Figure 3.0 | Flowchart describing the experimental and analytical design of Chapter 3. Created with Biorender.com.

3.2 Abstract

Polycomb Repressive Complex 2 (PRC2) core components EZH2, SUZ12 and EED are frequently altered in acute myeloid leukaemia (AML), leading to decreased genome-wide H3K27me3. PRC2 haploinsufficiency correlates with poor therapeutic response in childhood AML, but the reasons for treatment resistance are poorly understood.

To understand the consequences of PRC2 loss-of-function in AML, we developed an isogenic model of PRC2 loss via CRISPR-Cas9 editing of *EZH2* in the OCI-AML2 cell line, and performed RNA-seq, H3K27me3 and H2AK119Ub CUT&RUN, ATAC-seq and Hi-C on *EZH2*+/+ and *EZH2*+/- cells.

In PRC2-depleted cells we observed transcriptomic changes associated with alternative lineage gene expression programs, especially along the monocytic development axis, including decreased *CD14* expression. PRC2 depletion resulted in marked increases in genome-wide chromatin accessibility, accompanied by decreased H3K27me3 and H2AK119Ub. Despite these generalised changes, 3D chromatin architecture was largely maintained, with H3K27me3 being preferentially lost in regions with low frequency of DNA-DNA contacts. Surprisingly, a subset of genomic regions gained broad H3K27me3 domains at heavily compacted chromatin. In *EZH2*+/- cells we saw genome compartmentalisation changes upstream of the fetal haematopoiesis gene *LIN28B*. These changes were accompanied by increased *LIN28B* expression and activation of LIN28B-specific transcriptional programs, including upregulation of the *CDK6* oncogene. These results correlated with phenotypic changes in *EZH2*+/- cells, which displayed decreased cell proliferation and increased resistance to the CDK6 inhibitor palbociclib.

Our findings suggest that PRC2 depletion has diverse effects on AML transcriptional regulation that directly impact cell phenotype and treatment responsiveness, at least partly by epigenetically priming the chromatin to activate alternative transcriptional programs.

3.3 Introduction

Although many patients with acute myeloid leukaemia (AML) respond well to conventional chemotherapy, cases that show poor prognosis and chemoresistance are still difficult to treat, in large part due to an incomplete mechanistic understanding of therapy resistance and aggressive leukaemia biology.

Alterations in genes coding for epigenetic factors such as *DNMT3A*, *TET2*, *KMT2A* and *EZH2* have been associated with poor prognosis in AML (Ribeiro et al. 2012; Chou et al. 2011; Meyer et al. 2018; Ernst et al. 2010; Bond et al. 2018). EZH2 or its paralogue EZH1 are mutually-exclusive catalytic components of the polycomb repressive complex 2 (PRC2). PRC2 and PRC1 are polycomb group protein (PcG) complexes that work in complementarity and maintain gene repression primarily through trimethylation at histone 3 lysine 27 (H3K27me3) and ubiquitination of histone 2A at lysine 119 (H2AK119Ub), respectively. PcGs have important roles in development and cell fate commitment (Loubiere et al. 2019), guiding cell identity by maintaining lineage-appropriate transcriptional programs. Depletion of PRC1 and PRC2 in early haematopoiesis has detrimental effects on the Haematopoietic Stem Cell (HSC) pool in murine experiments (Kamminga et al. 2006; Vidal & Starowicz 2017). In more mature progenitors cell identity can be altered through activation of alternative lineage specification programs upon loss of either PRC1 or PRC2 components (Su et al. 2002; Oguro et al. 2010).

Aside from their enzymatic activities, PRC2 and PRC1 shape the chromatin landscape by mediating long-range looping and by creating largely repressed polycomb-associated domains (PADs), also known as polycomb bodies (Du et al. 2020; Saurin et al. 1998; Doyle et al. 2022). In line with this long-range function, disruption of H3K27me3-mediated loop anchors can activate genes at megabase distances (Kraft et al. 2022). Whilst H3K27me3 is not required for PAD maintenance, it is indispensable for the initial establishment and re-establishment of PADs (Boyle et al. 2020; Bonev et al. 2017). Therefore, regions with H3K27me3 may function as anchors for PADs and recruitment of cPRC1. PRC1 is also important for correct nucleosome positioning at transcriptional start sites (TSSs) (King et al. 2018). In contrast, the

role of PRC2 in nucleosome positioning is still a subject of debate. While some evidence indicates that it has no role in the nucleosome landscape (King et al. 2018), EZH2 depletion has been shown to lead to altered nucleosome occupancy and increased chromatin accessibility at bivalent promoters (i.e., marked with H3K4me3 and H3K27me3) at certain loci (Prorok et al. 2023).

In AML, reduced EZH2 activity due to mutation, copy number loss, or reduced expression correlates with chemoresistance and poor outcomes (Göllner et al. 2017; Bond et al. 2018; Basheer et al. 2019; Kempf et al. 2021). A mechanistic understanding of therapy resistance in these cases is however lacking, and is further complicated by the pleiotropic effects of PcG proteins in leukaemia cells. For example, we also know that PRC2 activity is important in AML tumorigenicity and cross-talks with other frequently altered epigenetic components (Ren et al. 2022; Sparbier et al. 2023). In a *KMT2A*-rearranged (*KMT2A*r) context, EZH2 can also act as an oncogene, with cells that have *KMT2A*::*MLLT3* translocations being dependent on EZH2 activity (Neff et al. 2012; Tanaka et al. 2012). However, in KMT2Ar murine models, EZH2 acts as a tumour suppressor at disease induction, decreasing survival of mice with EZH2-depleted leukaemia (Basheer et al. 2019). Conversely, *EZH2* depletion on the same genetic background during AML maintenance results in better survival (Basheer et al. 2019). In normal haematopoiesis, EZH2 is essential for maintaining haematopoietic stem and progenitor cell (HSPC) identity by mediating H3K27me3 placement and 3D chromatin architecture (X. Zhang et al. 2020). Furthermore, EZH2 is crucial in repressing fetal hematopoiesis programs in adult blood cells by repressing a specific transcriptional program controlled by the Let-7 miRNA suppressor LIN28B (Oshima et al. 2016).

To gain insight into the epigenetic and transcriptional landscape of PRC2-altered leukaemia, we created isogenic models of EZH2 depletion in an AML cell line. Extensive epigenomic characterisation of this system reveals key roles for PRC2 in regulating chromatin accessibility and genome architecture in AML. We further report that these changes are linked to altered lineage gene expression programs that correlate with clinical transcriptional data, providing insight into how PRC2 dysfunction might alter AML biology in patients.

3.4 Materials and methods

3.4.1 CRISPR/Cas9 gene editing of AML cell lines

Electroporation: Two guide-RNAs (gRNA, protospacer adjacent motif or PAM sequence underlined, #1: CGGAAATCTTAAACCAAGAATG, #2: ACCAAGAATGGAAACAGCGAAG), were designed to specifically direct the Cas9 endonuclease to the third exon of our gene of interest, *EZH2* (ENST00000460911.5; chr7:148,543,562-148,543,689 Reverse strand, GRCh37/hg19). gRNAs were purchased from Integrated DNA Technologies.

To deliver both the Cas9 enzyme and *EZH2*-targeted gRNAs into the OCI-AML2 cell line (gift from Mills Laboratory, Queen's University Belfast), we utilised the Alt-R CRISPR-Cas9 System from Integrated DNA Technologies. This approach allowed incorporation of both components into the cells, following electroporation, as a ribonucleoprotein (RNP) complex. Electroporation was performed using the Cell Line Nucleofector™ Kit V (Lonza, cat. no. VCA-1003) coupled with the Amaxa® Nucleofector® System (Lonza, Nucleofector 2b) according to the manufacturer's instructions. Lyophilised gRNAs were reconstituted in IDTE buffer at a final concentration of 200 µM. To achieve the recommended total amount of guide RNAs (100 µM) using 2 separate gRNAs, a 1:1 mixture was made in which each gRNA was at a concentration of 50 µM. The gRNA mixture was combined with 104 pmol of recombinant Cas9 protein and incubated for 20 minutes at room temperature to allow formation of the RNP complex. Cells were rinsed with PBS, counted using trypan blue exclusion, centrifuged, and resuspended in 60 µL Nucleofector™ solution at a density of 16,500 cells/µL. 15 µL of the RNP complex solution was added to the cells, as well as 3 µL of the Alt-R Cas9 Electroporation Enhancer solution (previously resuspended in IDTE buffer at a concentration of 100 µM). Nucleofection was performed using the Amaxa Nucleofector 2b device, using the electroporation settings outlined above. Following electroporation, cells were transferred into 500 µL of pre-warmed media per well of a 24 well plate and stored at 37°C, 5% CO_2.

Selection of single-cell clones: 48 hours post-electroporation, the bulk transfected cells were isolated into single cells to allow growth of single cell

colonies. For this, cells were counted using trypan blue exclusion and diluted to a concentration of 5 cells/mL. The cell suspension was then distributed in 96 well plates, adding 100 µL of cell suspension in each well. Using an initial concentration of 5 cells/mL per well (0.5 cells per well) minimised the probability of seeding more than one cell per well.

Cells were maintained in RPMI-1640 media supplemented with 20% fetal bovine serum (FBS) and 2 mM L-Glutamine. Once the cells reached confluence, they were successively transferred to 24-, 12-, and then 6-well plates. When the cell number was sufficient, proteins were extracted, and the lysates were run on a 10% acrylamide gel and subjected to immunoblotting to assess EZH2 protein levels. DNA was then extracted from clones showing reduced EZH2 levels, using the DNeasy® Blood and Tissue Kit (Qiagen, Cat 69504). Genomic DNA was used to amplify Exon 3 of *EZH2* via PCR. Amplicons were then subjected to direct (Sanger) sequencing to detect potential insertions or deletions (indels) responsible for reduced EZH2 levels. Cell line identity was authenticated by single nucleotide polymorphism profiling using a commercial service (Eurofins) and all cell lines used in this study were tested regularly (at least every 3 months) for mycoplasma using Lonza's MycoAlert® Mycoplasma Detection Kit (LT07-710) according to the manufacturer's instructions.

Identification of indels induced by CRISPR/Cas9 editing: Sanger sequencing results for *EZH2*-WT and candidate *EZH2*-knock-out (KO) cells were analysed using the ICE (Inference of CRISPR Edits) CRISPR Analysis Tool (available at https://ice.synthego.com), which integrates gRNA sequences and DNA sequence electropherogram files (.ab1 format). The ICE tool provides a knock-out (KO) score, which is the proportion of cells that have a frameshift or an indel greater than 21 base-pairs in length. Identification of a frameshift or indel using this tool allows for the assessment of functional consequences to the gene and its protein product. We selected two heterozygous *EZH2*-depleted clones: clone 5 (C5) and clone 9 (C9) to further investigate.

3.4.2 RNA-seq

Library generation: RNA was extracted from 3-5 million cells using the RNAeasy kit (Qiagen, 74104). Three independent RNA extractions were performed at serial passages for all cell lines to provide technical replicates. RNA quality was assessed using the Agilent 2100 Bioanalyzer prior to submission for sequencing as a service provided by Novogene UK. Before library preparation, samples were first enriched using oligo(dT) beads, mRNA was then randomly fragmented (average length of reads: 150 bp) before cDNA synthesis by reverse transcriptase. Library preparation was then carried out and library QC analysis was performed prior to paired-end sequencing at a depth of 30 million reads, resulting in .fastq files for processing.

HTS data processing: We pseudoaligned the reads against the GRCh38 (hg38) human reference genome using kallisto 0.46.1 (Bray et al. 2016) to obtain gene-level read counts. Quality Control (QC) was performed using FastQC 0.11.9 and MultiQC 11.9 (Andrews et al. 2012; Ewels et al. 2016). Subsequent analyses were performed using R v4.4.0 (https://www.R-project.org/) and Bioconductor 3.19 (Huber et al. 2015).

Differential expression analysis: We normalised the read counts to log2 transcripts per million (TPM) and Trimmed Means of M values (TMM) using the edgeR package v4.2.0 (Robinson et al. 2010). Then, we filtered out lowly expressed genes (<1 counts per million (CPM) in more than three samples). We performed variance-stabilisation with the voom function from limma v3.60.3 (Ritchie et al. 2015). To identify differentially expressed genes (DEGs) we used a limma linear model and we adjusted p-values for multiple testing using the Benjamini-Hocheberg correction. We identified significantly up- and down-regulated genes based on a false discovery rate (FDR) of 10% and $|logFC| \geq 0.5$. Overlaps between DEGs were plotted using the VennDiagram v1.6.0 R package (Chen & Boutros 2011). Odds ratios and p-values for overlaps were calculated using the fisher.test function from the stats package.

GSEA: We performed gene set enrichment analysis (GSEA) on the ranked list of detectable transcripts (identified as described above) through the clusterProfiler package v4.12.0 (Yu et al. 2012) using the Hallmarks (H) gene sets from MSigDB (Liberzon et al. 2011). We identified significantly enriched

gene sets using an FDR of 5%. We performed GSEA using gene signature from the Atlas of Human Blood cells (Xie et al. 2021). We considered genes most highly expressed in each cell type as a separate transcriptional signature and identified significantly enriched cell type transcriptional signatures under an FDR of 10%. Similarly, we performed GSEA using two signatures from genes differentially expressed between CD34+ fetal liver (FL) cells and *LIN28B*-KD CD34+ fetal liver cells. We tested the up- and the down-regulated gene signatures, respectively in the AML2 *EZH2*+/- and AML2 *EZH2*+/+ comparison and applied a Bonferroni correction for multiple testing.

3.4.3 CUT&RUN

We performed Cleavage Under Targets & Release Using Nuclease (CUT&RUN) on WT and C9 for H3K27me3 and H2AK119Ub (see antibodies in Supplementary Table S3.1). Full details of the protocol are available in Supplementary Methods.

HTS library preparation: HTS libraries for Ilumina were prepared using the NEBNext UltraExpress kit (NEB, Cat. E3325S) following the manufacturer's instructions. In brief, 25 ng purified DNA were end prepped using the provided enzyme mix in a thermocycler (20°C for 15 minutes then 65°C for 15 minutes). The NEB adaptor (sequence) was then ligated to the end prep reaction mixture using the NEBNext Ligation Master mix and incubated for 15 minutes at 20°C. Following this incubation, 2 µL USER enzyme was added to each tube and allowed to incubate for a further 5 minutes at 37°C.

The adaptor-ligated product was then PCR amplified using the provided NEBNext MSTC High Yield Master Mix and indexed using a combination of the i7 and i5 Illumina primers. Following one cycle of initial denaturation (98°C for 30 seconds), the sample underwent further denaturation (98°C for 10 seconds) and annealing/extension (65°C for 75 seconds) for 8 cycles. Final extension was then performed for one cycle at 65°C for 5 minutes.

Lastly, PCR-amplified libraries were cleaned using the Beckman Coulter Agencourt AMPure XP beads (Thermo Scientific, Cat. 10136224). A volume of 0.7x resuspended beads were added to the PCR reaction. The solution was

incubated on the benchtop for 5 minutes, then placed on a magnetic rack. After the supernatant was discarded, 50 µL 0.1x TE buffer was used to resuspend the beads, to which 0.4x NEBNext Bead Reconstitution buffer was added. After a 5 minute incubation at room temperature, the tubes were placed on a magnetic rack and the supernatant was discarded. Beads were then washed twice with 200 µL 80% ethanol. DNA from the beads was ultimately eluted using 33 µL of 0.1x TE buffer.

DNA concentration was measured using the Qubit dsDNA High-Sensitivity kit (Invitrogen, Cat. Q32851), while the size distribution of the libraries was assessed using the Agilent Bioanalyzer High Sensitivity DNA chip following the manufacturer's instructions. Libraries consisting of uniquely indexed samples were combined at equimolar concentrations and were sequenced using paired-end reads.

HTS data processing: QC of the unaligned reads was performed using FastQC 0.11.9 and MultiQC 11.9 (Andrews et al. 2012; Ewels et al. 2016). We performed adapter (TrueSeq adapters) removal with Trimmomatic v0.39 on the paired reads with settings as recommended by the 4DN project (4DN, https://data.4dnucleome.org/; Bolger et al. 2014). We aligned the reads against the human reference genome hg38 using bowtie2 v2.3.5.1 --end-to-end with the following parameters: inclusion of dovetailed reads (--dovetail), only concordant (--no-discordant), paired reads only (--no-mixed), and with lengths between 10 and 700 bp (-I 10 -X 700). We marked and removed duplicates and multi mappers using Picard v3.1.1 and samtools v1.18 (-F 1280 parameter) (Li et al. 2009; Picard, https://broadinstitute.github.io/picard/). For visualisation purposes we created .bigwig files using bamCoverage using --binSize 30 --smoothLength 60 --normalizeUsing CPM --effectiveGenomeSize 2913022398 --extendReads. We used these .bigwig files for generating deeptools heatmaps and profile plots (Ramírez et al. 2014).

Peak calling and annotation: We performed peak calling following the recommendations from SEACR (Meers et al. 2019). Specifically, we used the .bam files to obtain .bedgraph files through bedtools genomecov and called stringent peaks for each condition using SEACR v1.3 normalised to IgG. For heatmaps and profile plots over .bigwig files, we used the computeMatrix and

plotHeatmap functions from deeptools v3.5.4 (Ramírez et al. 2014). We found unique and overlapping peaks across the two conditions and two histone marks using the vennCount function with maxgap=1000bp from the hicVennDiagram v1.2.0 R package. We annotated the called CUT&RUN peaks with ChIPseeker v1.5.1 with a TSS region spanning between -3kb and +3kb and the parameter overlap = "all" to limit bias towards TSS annotation (Yu et al. 2015).

Figures containing tracks of ATAC-seq, CUT&RUN or Hi-C were obtained using pyGenomeTracks v3.9 (Lopez-Delisle et al. 2020).

3.4.4 ATAC-seq

HTS library generation: Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) was performed using an Active Motif kit (Cat. no. 53150) according to the manufacturer's instructions. Full details of the protocol are available in Supplementary Methods.

HTS library preparation: The eluted DNA was subsequently amplified using a combination of indexed primers i7 and i5 (25 µM) in a thermocycler with the heated lid on with the following steps: 72°C for 5 minutes, 98°C for 30 seconds, followed by 10 cycles of 98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute. Amplified DNA was size-selected using 1.2x SPRI beads. These were then washed using 180 µL 80% ethanol twice and DNA was ultimately eluted using 25 µL elution buffer. The complexity of the library was assessed using the Agilent Bioanalyzer and DNA was quantified using the High-Sensitivity DNA Qubit kit (Cat no. Q32851, ThermoFisher Scientific).

HTS data processing: We applied the nf-core/atacseq v2.1.2 pipeline to perform initial QC, adapter trimming, duplicate removal and alignment of reads into .bam files. We used the peak calls from the pipeline to generate a PCA plot in order to check for similarity between replicates. Based on this output, we concluded that the replicates were similar (Supplementary Figure S3.4A) and used the merged .bam files to apply custom pipelines as described below.

Peak calling and annotation: We performed peak calling using HMMRATAC v1.2.10 using default settings (Tarbell & Liu 2019) and excluding the hg38 v2 blacklist from https://github.com/Boyle-Lab/Blacklist (Amemiya et

al. 2019). We found unique and overlapping peaks between the three conditions using the vennCount function with maxgap=50bp from the hicVennDiagram v1.2.0 R package. We annotated the called peaks with ChIPseeker v1.5.1 with a TSS region spanning between -3kb and +3kb and the parameter overlap = "all" to limit bias towards TSS annotation.

Heatmaps: Using alignmentSieve from deeptools v3.5.4, we separated the ATAC-seq data for each condition into two groups based on fragment size: nucleosome-free regions (NFRs) <100bp and mononucleosomal fragments between 180 and 250bp. We generated bigwig files from .bam files using bamCoverage and the following parameters: --binSize 1 --normalizeUsing RPGC --effectiveGenomeSize 2913022398. For heatmaps and profile plots over .bigwig files, we used the computeMatrix and plotHeatmap functions from deeptools v3.5.4 (Ramírez et al. 2014). All heatmaps for ATAC-seq were generated using NFR regions.

Enrichment analyses: To determine enriched TF motifs in each condition, we used the findMotifsGenome.pl function from Homer v5.0.1 with the parameter -size 50 (i.e., searching for motifs ±50 bp from the peak) (Heinz et al. 2010). We searched for known motifs around peaks only present in C5 and C9 against peaks only present in WT as background, and vice-versa. We performed genomic regions enrichment on the overlapping C5 and C9 open regions, using all called peaks as background. For this analysis we used the web-based tool GREAT with default parameters: http://great.stanford.edu/ (McLean et al. 2010)

Nucleosome positioning inference: We inferred nucleosome positions around NFRs using NucleoATAC 0.3.4 with default settings (Schep et al. 2015). For each nucleosome inferred, NucleoATAC provides an occupancy score and a fuzziness score. Median inter-dyad distances were compared across the three conditions using Wilcoxon unpaired tests. We stratified the inferred nucleosomes within ±500bp of TSSs into -2, -1, +1 and +2 nucleosomes. We did so by fitting a mixture model with four components to the inferred nucleosome positions using the mixtools R package and allocating each nucleosome to one of the preset positions based on its likelihood (Benaglia et al. 2009). After this, we used the average position of each nucleosome type (-2, -1, +1 and +2) to calculate the average distance from TSS. Median nucleosome

occupancy scores and nucleosome fuzziness were compared across the three conditions using Wilcoxon unpaired tests.

3.4.5 Hi-C

HTS library preparation: All steps for the generation of Hi-C samples for sequencing (sample processing and library preparation) were performed as a service by Active Motif. For this, 10×10^6 cells per cell line were washed in 1x PBS, pelleted and stored at -80^{o}C prior to submission.

The Active Motif Hi-C workflow used the Arima-HiC Kit (Arima Genomics) for performing chromatin conformation capture (crosslinking, digestion, biotinylation, ligation and fragmentation). The Active Motif workflow for library preparation used the KAPA Library Amplification Kit (Roche), according to manufacturer's instructions.

Hi-C libraries were diluted to 20 nM and submitted for sequencing to Novogene. Sequencing of the pre-made libraries was performed, generating 300 million reads (paired-end) per sample.

HTS data processing: We performed QC using FastQC v0.12.1 and aggregated the outputs using MultiQC (Andrews et al. 2012; Ewels et al. 2016). We aligned the raw reads against the reference human genome hg38 using BWA-MEM2 v2.2.1 (Vasimuddin et al. 2019). We then created pair files using pairtools v1.0.3 with a minimum mapping quality (mapq) of 3 and the parameter "--report-position outer", as recommended on the Pairtools Github https://github.com/open2c/pairtools (Open2C, 2024). We used pairtools to further split and sort the paired reads. We then used the FAN-C command "fanc pairs" with the "-restriction_enzyme" argument to filter for ligation errors (Kruse et al. 2020). For OCI-AML2 we used the *DpnII* restriction fragments as input, as *DpnII* was used to generate the data, as stated by Takayama et al. (2021). For the Hi-C data generated by us (OCI-AML3 and OCI-AML2 C9) we used the Arima-specific list of restriction fragments (i.e., Arima uses both *HinfI* and *DpnII*). After investigating ligation error statistics, we filtered the paired reads using fanc pairs with the following parameters: "--filter-unmappable --filter-multimapping --filter-inward 5000 --filter-outward 5000 --filter-self-ligations

--filter-pcr-duplicates 1". We used “fanc hic” to create a .hic matrix from the .pairs file and then applied ICE (iterative correction and eigenvector decomposition) normalisation.

Cooler files generation: To calculate the resolution of each .hic matrix, we used “fanc resolution”. More precisely, for bin sizes of 5kb, 10kb, 12kb, 15kb, 20kb, 50kb, 100kb, in each matrix, we calculated the percentage of bins with >1000 contacts. We set a threshold at 80% of bins having >1000 contacts and concluded that the optimal resolution for the .hic maps was 15kb. We created .hic maps at 7.5kb resolution for visual inspection purposes and used fanc to-cooler with the parameters “--uncorrected” and “--no-multi” to generate a .cool file. Further, we used the zoomify function from cooler 0.9.3 to generate an .mcool file with decreased resolutions (15kb, 30kb, 60kb, 120kb, 240kb etc.). We used the “--balance” argument for normalisation and “--mad-max 0 --max-iters 1000” as suggested on https://github.com/open2c/cooler.

Distance decay: We calculated distance decay for each condition using fanc expected. The distance-decay curve allows us to visualise the frequency of contacts as a function of genomic distance. The derivative of the distance-decay curve allows us to better visualise any change in directionality of the curve.

Loop calling: We performed loop calling at 15kb resolution using Mustache v.1.3.2 with σ_0=1.6 (Roayaei Ardakany et al. 2020). For determining condition-specific and overlapping loops, we used the hicVennDiagram v1.2.0 R package with a maxgap=20kb.

Pile-ups: We performed pile-ups of Mustache loop calls using coolpup.py v1.1.0 at 15kb resolution with 150kb flanking in each direction from the loop centre. Similarly, we performed pile ups on interactions between H3K27me3 regions. Due to the maximum resolution of the Hi-C data being 15kb, we decided to merge H3K27me3 loops into domains using bedtools, as follows: if the distance between two loops is <15kb, they are collapsed within a domain (bedtools merge -d 15000). Then we performed pileups at 15kb, 30kb, 60kb, 120kb and 240kb with 20x resolution as flanking. We observed that the domains were most likely larger than we expected, hence the 120kb resolution pileups looked the best (i.e, there was a visible central enrichment smaller than the size of the pileup window).

For visualisation of Hi-C maps we used HiGlass (Kerpedjiev et al. 2018) and pyGenomeTracks (Lopez-Delisle et al. 2020).

3.4.6 TARGET AML samples RNA-seq data analysis

We downloaded TARGET AML RNA-seq and clinical information from cBioportal (N=45) (See Supplementary Methods) (Cerami et al. 2012; Gao et al. 2013; De Bruijn et al. 2023). We stratified the whole cohort into two groups based on median *EZH2* mRNA expression (≤ median and > median). We performed the same stratification for the subset of samples with *KMT2A*r (N=7). For both cases (whole cohort and *KMT2A*r-restricted) we performed differential expression analysis (see differential expression analysis in RNA-seq HTS data processing). We also calculated Spearman correlations between log2(RPKM+1) of genes (e.g., *EZH2* and *CD14*) in the *KMT2A*r cohort.

3.5 Results

3.5.1 EZH2 depletion leads to activation of alternative lineage transcriptional programs

To model PRC2 haploinsufficiency encountered in patient leukaemia cells, we targeted the methyltransferase component *EZH2* using CRISPR-Cas9 editing of the AML cell line OCI-AML2. The chosen cell line is an acute myelomonocytic leukaemia harbouting a *KMT2A*::*AFDN* translocation and a *DNMT3A* R635W mutation. To ensure that experimental results were not clone-specific (Westermann et al. 2022), we generated two models of heterozygous EZH2 depletion, henceforth named clone 5 (C5) and clone 9 (C9) (Figure 3.1A, Supplementary Figure S3.1). We confirmed EZH2 protein decrease and H3K27me3 decrease by approximately 40-50% via immunoblotting (Figure 3.1B). While EZH2 depletion had no effects on cellular viability, EZH2-deficient cells had reduced proliferation rates equivalent to 61% (C5) and 53% (C9) of the WT cell line (Figure 3.1C).

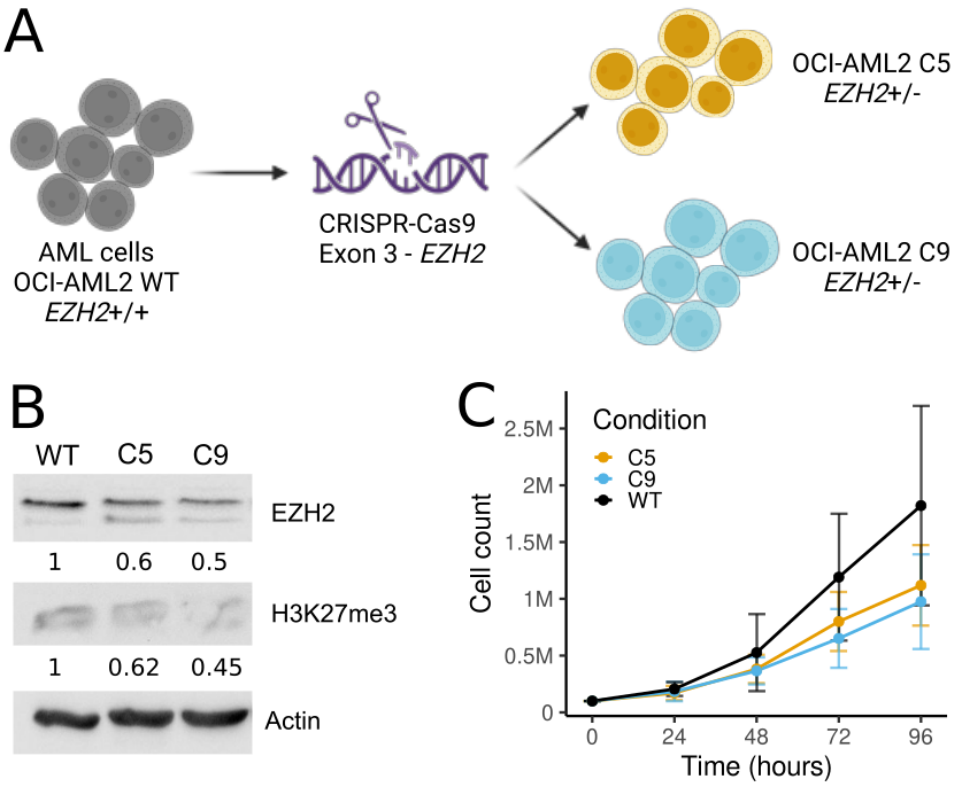


Figure 3.1 | Creation and characterisation of an AML cell line model of PRC2 depletion. A, Two heterozygous EZH2 loss models were generated using CRISPR-Cas9 editing of the OCI-AML2 cell line: C5 and C9 (See Methods). **B,** Immunoblotting of protein extracts from WT and EZH2-depleted cell lines using EZH2 and H3K27me3 antibodies, with β-actin as a loading control. Quantification performed using ImageJ (Schneider et al. 2012) **C,** Cell growth in WT, C5 and C9, as measured by cell counting with Trypan Blue exclusion (N = 7).

We first assessed the transcriptional differences in EZH2-deficient cells by performing RNA-seq on WT, C5 and C9 (Supplementary Figure S3.2A). Differential expression analysis uncovered a similar number of up- and downregulated genes in both C5 (619 genes upregulated and 845 genes downregulated) and C9 (369 genes upregulated and 366 genes downregulated) compared with WT transcription (Figure 3.2A, Supplementary Table 2). We found the overlap between C5 and C9 to be larger than expected by chance for both up- and down-regulated genes, with odds ratios (OR) of 10.275 and 7.005, respectively (Figure 3.2A). We performed gene set enrichment analysis (GSEA) on DEGs between *EZH2*+/- cells and *EZH2*+/+ using MSigDB Hallmarks to identify common signatures enriched in the two clones. We found very few statistically significant changes in Hallmark gene signatures, being limited to

negative enrichment of complement, NF-κB signalling and interferon gamma signalling, and positive enrichment of E2F targets in EZH2-depleted cells compared with WT (Figure 3.2B). Relationships between PcGs and E2F family of transcription factors have been previously reported, with PRC1 inhibiting E2F-dependent gene expression and E2F6 being part of the PRC1.6 complex (Stielow et al. 2018; Hanselmann et al. 2023)

As PRC2 is known to be critical for regulation of lineage-specific transcription at multiple stages of blood cell development, we were particularly interested in assessing expression changes in haematopoietic-related factors in more detail. To do so, we used publicly available scRNA-seq data from the Atlas of Human Blood cells (ABC), a large scale effort to transcriptionally characterise blood cell types (Xie et al. 2021). We generated gene sets of cell-specific transcriptional programs from the 32 ABC cell types and performed GSEA to test for transcriptional programs altered upon PRC2 depletion (Supplementary Figure S3.2B).

Notably, we observed a significant enrichment of human monocytic dendritic progenitors/common monocytic progenitors (hMDP/cMoP) genes in PRC2-depleted cells (NES = 2.18, adjusted p-value = 1.18×10^{-4}) (Figure 3.2C, Supplementary Figure S3.2B). Conversely, we also observed a negative enrichment of differentiated monocyte programs (i.e., classical monocytes: NES (Normalised enrichment score) = -1.92, adjusted p-value = 1.86×10^{-4} and non-classical monocytes and NES = -1.94, adjusted p-value = 7.57×10^{-5}) (Figure 3.2C, Supplementary Figure S3.2B). Together, these results suggest that PRC2 depletion results in partial activation of alternative transcriptional programs associated with more immature stages of differentiation in this model (Supplementary Figure S3.2B). The PRC2-depleted cells show increased expression of hMDP/cMoP genes such as *NREP*, *SOX4* and *CDCA7* (Supplementary Figure S3.2E). Additionally, monocytic genes such as *CD14* and *S100A9,* and *ITGAM* (encoding CD11b) are downregulated in PRC2-depleted cells (Figure 3.2E, Supplementary Figure S3.2F, Supplementary Table S3.2).

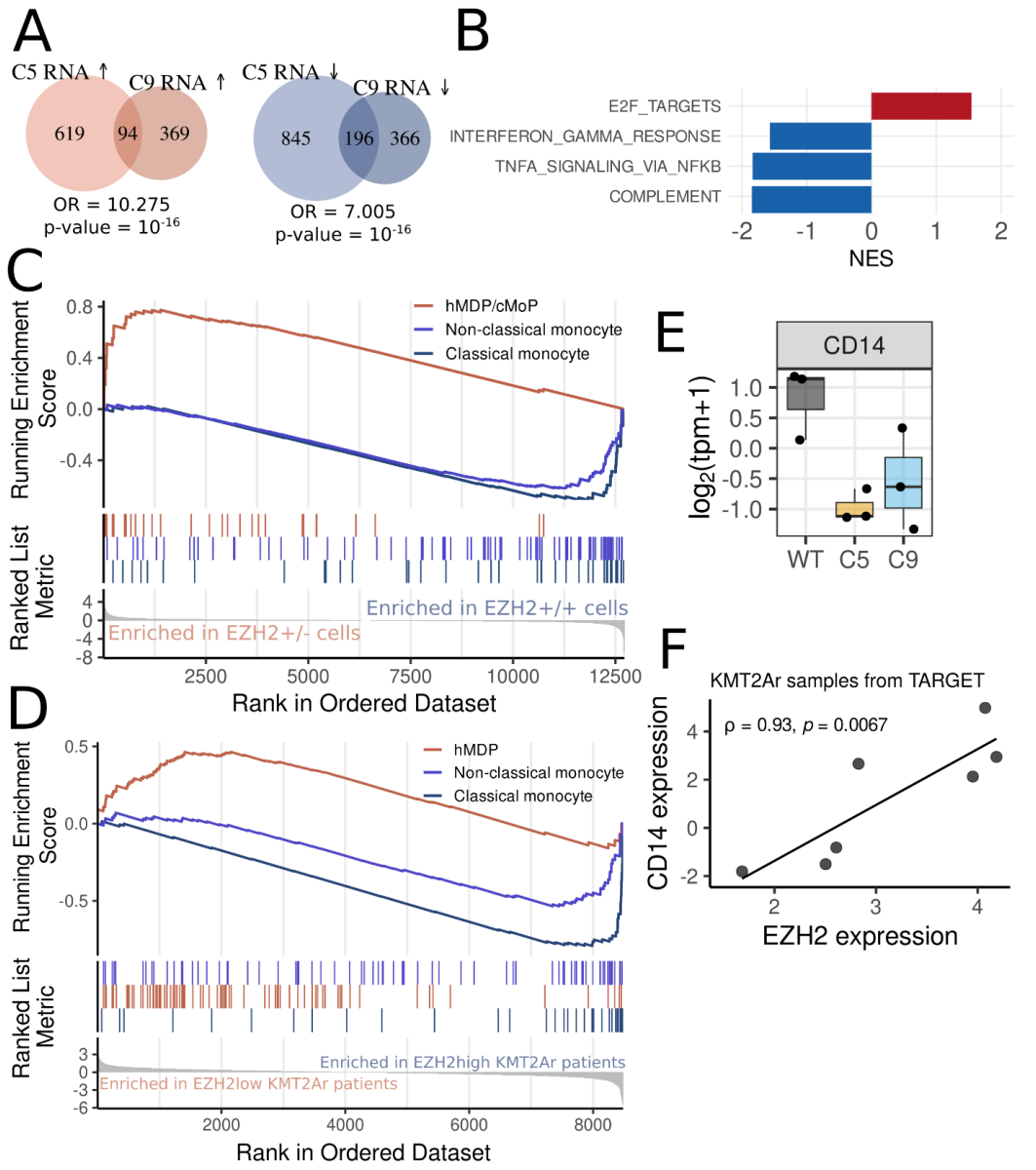


Figure 3.2 | PRC2 depletion leads to changes in gene expression related to cell differentiation. A, C5 and C9 DEGs from RNA-seq. Red = upregulated; blue = downregulated; OR = Odds ratio for overlap; p-value = Fisher's exact test p-value. **B,** Gene set enrichment of MSigDB hallmarks comparing *EZH2*+/- (C5 and C9) and *EZH2*+/+ (WT) RNA-seq. Red = Sets upregulated in *EZH2*+/- cells; blue = sets downregulated in *EZH2*+/- cells. **C,** Most highly enriched gene sets from the atlas of human blood cells in OCI-AML2 PRC2-depleted (red) and PRC2-WT (blue). **D,** Most highly enriched gene sets from the atlas of human blood cells in patients with low *EZH2* expression (red) and patients with high *EZH2* expression (blue). Data used was from patients with a *KMT2A* translocation background from TARGET (N=7). **E,** Monocyte marker *CD14* mRNA expression in WT, C5 and C9. **F,** Spearman correlation between *EZH2* and *CD14* mRNA expression in TARGET patients with *KMT2A*r.

To test whether these lineage genes are also altered in patient leukaemias, we tested the same gene sets in transcriptional data from the TARGET paediatric AML project. We split the cohort of 45 paediatric AML samples into two groups based on median *EZH2* expression: *EZH2*-low and *EZH2*-high. GSEA comparing these groups in the whole cohort showed enrichment of monocyte/dendritic signature in *EZH2*-low samples and a Pre-monocyte signature in *EZH2*-high samples (Supplementary Figure S3.2C). However, the OCI-AML2 cell-line harbours a *KMT2A* rearrangement (*KMT2A*r), which also affects epigenetic regulation in leukaemia. Therefore, we hypothesised that the difference in expression of lineage defining genes may be context-dependent. When we restricted our analysis to *KMT2A*r samples (N = 7), we found a significant enrichment of a hMDP signature in *EZH2*-low patients and monocytic signatures in *EZH2*-high patients, recapitulating the findings of the cell line model (Figure 3.2D, Supplementary Figure S3.2D). Furthermore, we observed a high positive correlation between *EZH2* expression and transcription of the monocyte cell surface marker *CD14* (Figure 3.2F). Additionally, we saw a negative correlation between *EZH2* expression and hMDP/cMoP signature genes such as *FLT3, NREP, SOX4* and *CDCA7* (Supplementary Figure S3.2H, Supplementary Table 3). We also assessed the expression patterns of the CD14+ monocyte-specific genes from Zeng et al. (2023) which were also present in the negatively enriched monocytic signatures. We confirmed a positive correlation between *EZH2* expression and the expression of these marker genes (i.e., *S100A9*, *S100A8*, *VCAN*, *S100A12*, *CD14*) in both cell-line models and patients (Figure 3.2E,F, Supplementary Figure S3.2F, G, Supplementary Table S3.3).

3.5.2 Heterozygous loss of EZH2 leads to significant decreases in genome-wide polycomb marks in AML

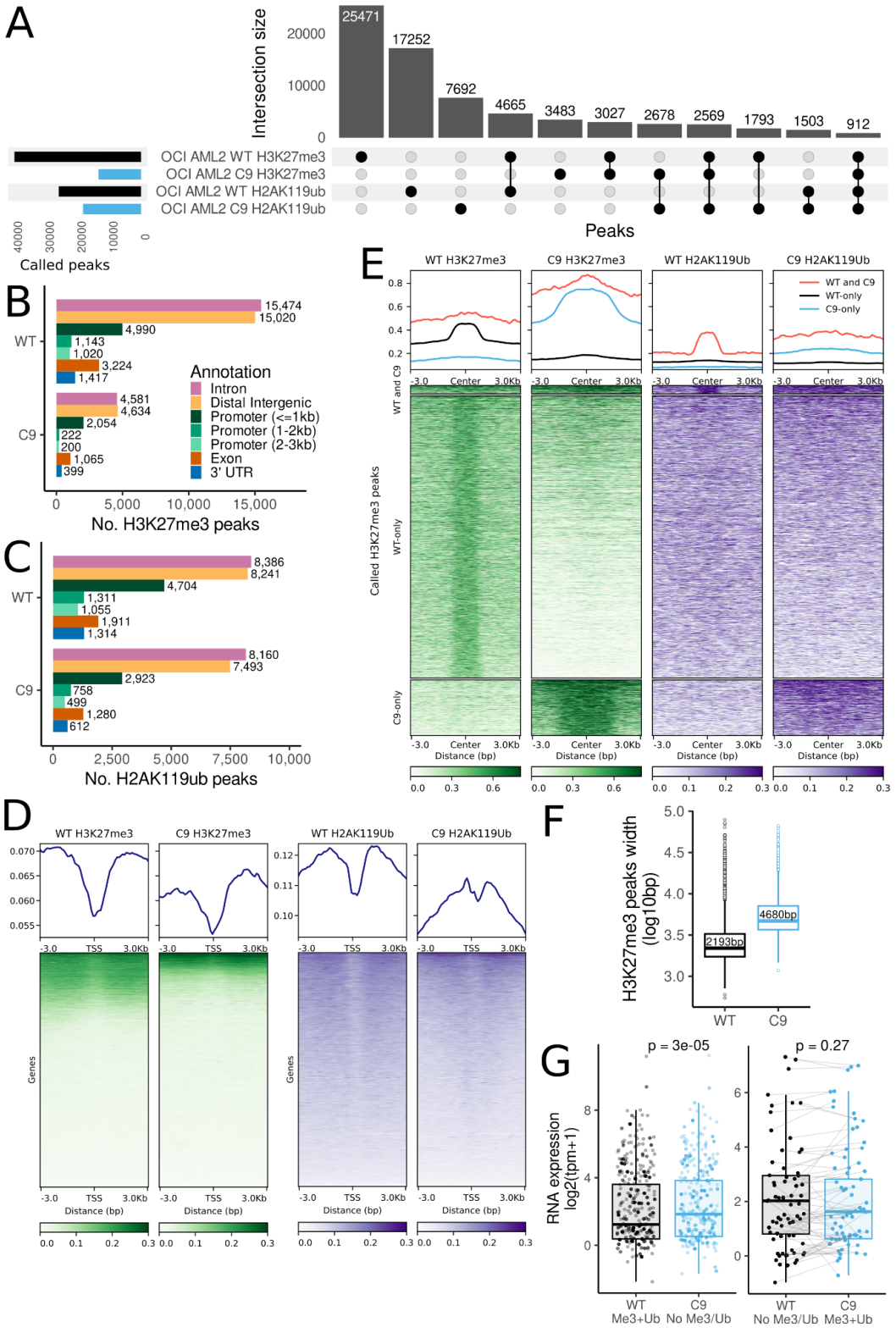
To understand the epigenetic implications of depleting *EZH2*, we next performed Cleavage Under Targets & Release Using Nuclease (CUT & RUN) (Skene & Henikoff 2017) in WT and EZH2-deficient C9 cells to assess genome-wide deposition of PRC2-placed H3K27me3 and PRC1-placed

H2AK119Ub. We first confirmed a strong concordance between H3K27me3 signal in our samples and publicly available H3K27me3 called peaks from two different AML cell lines: MOLM13 and HL60 (Supplementary Figure S3.3A), suggesting that PRC2-mediated gene regulation is broadly conserved across different AML subtypes.

Strikingly, we observed an ~70% reduction in the total number of H3K27me3 peaks in our cell line model (42,741 peaks called in the WT vs 13,277 peaks called in C9) (Figure 3.3A). We also observed a more moderate reduction in the number of H2AK119Ub peaks, from 27,345 peaks called in the WT to 21,981 peaks in C9 (Figure 3.3A). Whilst H3K27me3 peaks are lost across all types of genomic regions (Figure 3.3B), H2AK119Ub peaks are preferentially lost at promoter sites (Figure 3.3C, Supplementary Table S3.4). Investigating the average signal at genome-wide TSS, we observed a reduction in both H3K27me3 and H2AK119Ub (Figure 3.3D, Supplementary Figure S3.3B).

Interestingly, more than 4,500 loci had reductions in both H3K27me3 and H2AK119Ub upon EZH2 depletion, suggesting that H3K27me3 loss affects H2AK119ub placement at these regions (Figure 3.3A). In contrast, we also observed increases in PRC2- and PRC1-associated histone marks at certain loci in EZH2-depleted cells. A subset of loci in C9 had increases in H3K27me3 and H2AK119Ub both concomitantly and independently of each other (Figure 3.3A). In fact, we observed that some regions gained stronger H3K27me3 signal in C9. Additionally, these regions also became wider by more than two-fold, suggesting more internucleosomal H3K27me3 spread in C9 (Figure 3.3E, F). Furthermore, the regions with the strongest H2AK119Ub signal also had conserved trimethylation upon EZH2 depletion (Figure 3.3E), suggesting a role for PRC1 in maintaining full repression of these regions. Finally, we observed an increase in RNA expression upon EZH2 depletion for genes with loss of both H3K27me3 and H2AK119Ub at their promoters or within the gene body (Figure 3.3G). This included *CDCA7* from the hMDP/cMOP signature analysed above (Figure 3.2C, Supplementary Figure S3.2F, S3.3C). However, we observed no correlation between RNA expression and gain of H3K27me3 and H2AK119Ub upon EZH2-depletion, with some genes becoming

downregulated and others upregulated, despite deposition of repressive marks (Figure 3.3G).



3.5.3 Heterozygous loss of PRC2 methyltransferase EZH2 leads to significant increases in chromatin accessibility in AML

Having assessed the effects of EZH2 loss on Polycomb-mediated histone modification, we next wished to evaluate whether reduced PRC2 function also affected chromatin accessibility. To analyse this on a genome-wide basis, we performed ATAC-seq in our OCI-AML2 WT, C5 and C9 models. Principal component analysis showed high similarity between technical replicates (Supplementary Figure S3.4A). Therefore, we combined the three replicates for each cell line and performed peak calling using HMMRATAC (see Methods). We found a large increase in the total number of peaks called in both C5 and C9, with 14,585 and 33,476 accessible peaks being gained in C5 and C9 respectively (Figure 3.4A, Supplementary Figure S3.4B). 10,601 of these were common to both clones (Figure 3.4A, Supplementary Figure 3.S4B). This exceeds the number of lost accessible peaks by more than five-fold (Supplementary Figure S3.4B). Overall, these data suggest that heterozygous loss of EZH2 leads to a marked global increase in chromatin accessibility. Similar to the patterns seen for histone marks (Figures 3B and 3C), these changes were not specific to any class of genomic region (Figure 3.4B, Supplementary Table S3.5).

Figure 3.3 | PRC2 depletion leads to genome-wide decrease in H3K27me3 and H2AK119Ub. A, UpSet plot of H3K27me3 and H2AK119ub peaks called by SEACR in WT and C9 cells **B,** Annotation of H3K27me3 peaks called in WT and C9. **C,** Annotation of H2AK119Ub peaks called in WT and C9. **D,** Profile plots and heatmaps of H3K27me3 (green) and H2AK119Ub (purple) signal at transcriptional start sites (TSS) **E,** Profile plots and heatmaps of H3K27me3 (green) and H2AK119Ub (purple) signal at called H3K27me3 peaks stratified by sample in which the peaks were called. **F,** RNA expression in WT and C9 for genes that lose or gain both H3K27me3 and H2AK119Ub within the gene body or at the promoter upon PRC2 loss. P-values obtained upon comparisons using paired Wilcoxon tests. **G,** Width of H3K27me3 peaks from CUT&RUN in WT and C9.

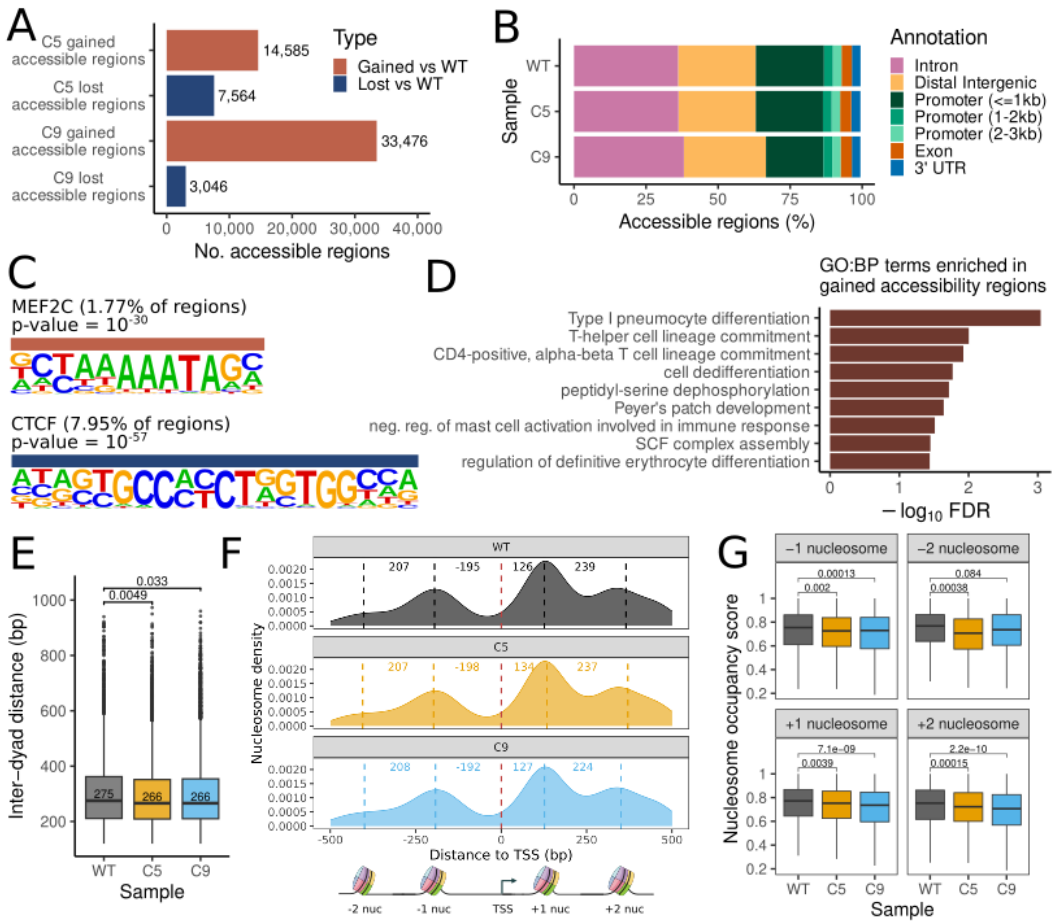


Figure 3.4 | Changes in chromatin accessibility correlate with changes in gene expression upon PRC2 depletion. A, Chromatin accessibility profiles from ATAC-seq summarised by comparing total number of accessible regions in C5 and C9 with total number of accessible regions in WT. **B,** Annotation by type of genomic regions for called accessible regions in the three conditions. **C,** Enrichment of known TFs in C5 and C9 accessible regions (red) and WT accessible regions (blue). **D,** GO enrichment of C5 and C9 accessible regions. **E,** Inter-dyad distance for nucleosomes called using nucleoatac near accessible regions at TSSs (Wilcoxon unpaired test p-value above brackets). **F,** Distribution of nucleosomes at genome-wide TSSs. Red dashed line = TSS. Black, yellow and blue lines = average nucleosome position. Values between dashed lines = average distance between TSS and nucleosome or between two neighbouring nucleosomes. **G,** Nucleosome occupancy score for -2, -1, +1 and +2 nucleosomes, respectively in WT, C5 and C9 (Wilcoxon unpaired test p-value above brackets).

3.5.4 Open chromatin regions in PRC2-depleted cells are associated with development and cell differentiation

To interrogate functional differences between genes located in the increased and decreased accessible regions we performed Hypergeometric Optimization of Motif EnRichment (Homer) TF analysis and Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al. 2010; Heinz et al. 2010).

Firstly, using Homer, we calculated the enrichment of known TFs in the regions with increased accessibility in C5 and C9 compared with regions with decreased accessibility, and vice versa (Supplementary Table S3.6). We found that CTCF motifs were enriched in regions that were less accessible upon PRC2 depletion (145/1,824 regions, 7.95%), suggesting a change in loop extrusion patterns at these loci in our model system (Figure 3.4C). In contrast, more accessible regions showed less specific enrichment (Supplementary Table 3.6), with MEF2C (Myocyte-specific enhancer factor 2C) exhibiting the highest enrichment overall (188/10,621, 1.77%) (Figure 3.4C). MEF2C is an essential TF in myogenesis and bone marrow B-lymphopoiesis (Dodou et al. 2003; Wang et al. 2016).

Gene ontology (GO) analysis using GREAT revealed no significant enrichment for any specific biological process (BP) in less accessible regions. In contrast, we found enrichment of terms associated with haematopoietic development (e.g., T-helper cell lineage commitment) and broad development (e.g., cell dedifferentiation) in regions with increased accessibility (Figure 3.4D). This supports the hypothesis that PRC2 loss leads to opening of chromatin associated with alternative lineages. However, differential gene expression analysis revealed that only 2/40 genes that drive the enriched GO terms are significantly upregulated in EZH2-deficient cells: *AZU1* and *CDK6* (Supplementary Table S3.2, S3.6), suggesting that the links between chromatin accessibility and transcriptional activity are incomplete in this case.

To explore the links to gene expression further, we directly mapped fragments associated with open chromatin, i.e. nucleosome free regions (< 100 bp) to transcriptional start sites (TSSs) of DEGs (see Methods). In both C5 and C9 we observed a modest increase of accessibility at TSSs of upregulated genes (Supplementary Figure S3.4A) and a slight decrease in accessibility at

TSSs of downregulated genes (Supplementary Figure S3.4B). Visual inspection of the heatmaps suggested that most TSSs were already accessible in WT cells. In line with this, of 74 genes upregulated in *EZH2*+/- cells, only 19 were not expressed at all in *EZH2*+/+ cells, suggesting the majority of changes were quantitative in nature, rather than qualitative.

We stratified genes based on increases/decreases in RNA and ATAC signals, and found that 134 and 117 genes had both increased promoter accessibility and increased RNA in C5 and C9, respectively (Supplementary Figure S3.5C, D). However, the vast majority of genes do not overlap, e.g., 1,517 genes have an accessible promoter in C5, without showing a significant upregulation at the RNA level (Supplementary Figure S3.5C). Investigating the overlap between the two clones, we identified 23 genes that increase both RNA and promoter accessibility in C5 and C9, a larger overlap than expected by chance (OR = 4.469; Fisher's exact test p-value = 10^{-7}) (Supplementary Figure S3.5E, G). Furthermore, we observe an overlap larger than expected by chance between C5 and C9 for genes that decrease in both RNA and accessibility (OR = 13.227; p-value = 5×10^{-10}) (Supplementary Figure S3.5F, G). Taken together our analysis suggests modest correlations between changes in mRNA abundance and changes in chromatin accessibility, with a limited, but significant number of genes following a pattern of increased promoter accessibility resulting in increased mRNA abundance.

3.5.5 Depletion of PRC2 alters genome-wide nucleosome positioning near TSSs

We hypothesised that EZH2 depletion may not only affect chromatin accessibility, but also positioning of nucleosomes at TSSs (Prorok et al. 2023). To infer nucleosome positioning near open chromatin regions we used NucleoATAC (Schep et al. 2015). We observed that there is a slight increase in inter-dyad distance between nucleosomes (i.e., the length of DNA between neighbouring nucleosomes) in *EZH2*+/- cells, suggesting alteration of the nucleosome landscape (Figure 3.4E). We plotted the genome-wide nucleosome density at TSSs and observed a multi-modal distribution with four peaks,

suggesting we can capture the -2, -1, +1 and +2 nucleosomes within ±500bp of the TSS (Figure 3.4F). We inferred the average positions of these nucleosomes by fitting a mixture model on each distribution (see Methods), and calculated the distance between each nucleosome and the TSS. Overall, we observed no difference in average nucleosome positions. However, nucleosome occupancy scores suggest a decrease in occupancy for all four inferred nucleosomes in *EZH2*+/- cells than in *EZH2*+/+ (Figure 3.4G). This may be caused by a larger spread of nucleosomal fragments at these nucleosomes, therefore nucleosomes are called with less confidence. This hypothesis is confirmed by comparing nucleosome fuzziness scores, which are significantly higher in C5 at all nucleosome positions, and significantly higher in C9 at the -2 position (Supplementary Figure S3.6).

3.5.6 H3K27me3 is preferentially maintained and gained at loci involved in 3D genome structure

As we know that PcGs can control long range interactions (X. Zhang et al. 2020; Kraft et al. 2022), we interrogated 3D chromatin architecture changes upon PRC2 depletion by performing Hi-C of the EZH2-depleted C9 cell line. We compared these results to publicly available OCI-AML2 Hi-C data (Takayama et al. 2021). OCI-AML2 has a highly disorganised genome, with many large scale translocations and inversions that introduce extraneous noise in Hi-C data. Therefore, we performed Hi-C on an additional control cell line OCI-AML3 to control for both unwanted variation and batch effects.

After QC analyses, we calculated a matrix resolution of 15 kb for the Hi-C matrices of three samples (i.e., more than 80% of 15 kb-sized bins contain more than 1,000 contacts) (Supplementary Figure S3.7A). Furthermore we saw no significant difference across the three samples on the distance-decay curve, which allows us to estimate contact frequency by genomic distance (Supplementary Figure S3.7B, C). We performed loop calling using Mustache (Roayaei Ardakany et al. 2020). As expected, the majority of loops (33%) were present in all three samples, suggesting general conservation of global chromatin architecture (Figure 3.5A). Genome-wide pileups of these loops

suggest a slight increase in looping intensity upon PRC2-depletion (Figure 3.5B). However, we do not see a clear change in the number of loops upon PRC2-depletion (Figure 3.5A).OBJ OBJ

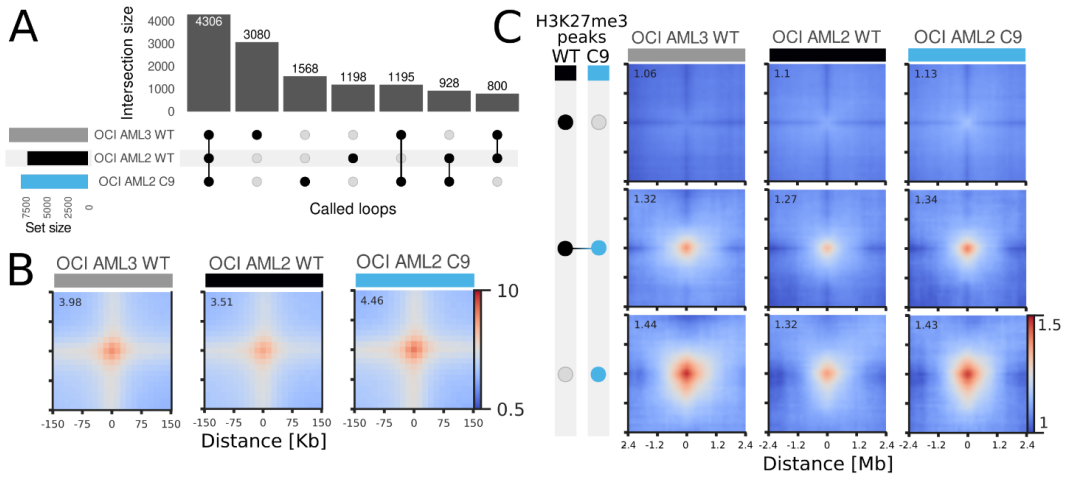


Figure 3.5 | The global chromatin looping landscape is maintained upon PRC2 depletion. A, UpSet plot showing overlaps of loops called via Mustache in OCI-AML3, OCI-AML2 WT and OCI-AML2 C9 Hi-C at 15kb resolution. **B,** Pileups of all called loops in the three conditions at 15kb resolution **C,** Pileups of contacts between domains of collapsed H3K27me3 peaks and stratified by: WT-specific peaks (black), C9-specific peaks (cyan) and WT and C9 overlapping (black+cyan) in the three cell lines at 120kb resolution.

We wanted to investigate the 3D chromatin changes at PRC2-bound loci in our model. As described above, we noticed that there is a subset of H3K27me3 regions which are gained upon EZH2 depletion (Figure 3.3D). These newly formed H3K27me3 peaks are wider than the peaks present in *EZH2*+/+ cells. In fact, comparing the H3K27me3 peak width between WT and C9, we observed a doubling of the median peak width upon EZH2 depletion, suggesting more spreading and formation of broader H3K27me3 domains (Figure 3.3E, F). To gain insights into whether PRC2 binding influences genome architecture at specific loci, we performed pile-up analysis of H3K27-enriched regions according to whether these were found only in WT, C9, or in both. As shown in Figure 3.5C, there were major differences in 3D chromatin interactions between the three categories . Firstly, the H3K27me3 regions that are lost upon EZH2 depletion do not seem to be involved in interactions. Secondly, the H3K27me regions that are maintained upon EZH2 depletion show strong DNA-DNA looping interactions. Finally, on average, the newly deposited

H3K27me3 marks are also enriched for 3D chromatin interactions, suggesting mechanisms of maintenance of genome architecture. Additionally, we observed the same patterns of looping in publicly available Micro-C from the chronic myelogenous leukaemia cell line K562 (Supplementary Figure S3.7D).

We identified some short-range loops lost upon EZH2 depletion, correlating with H3K27me3 deposition and gene expression, suggesting they are potentially mediated by PcGs (Supplementary Figure S3.8). For example, the *SKIDA1* gene loses H3K27me3 and H2AK119Ub upon EZH2 depletion, resulting in the loss of a loop between *SKIDA1* and the *BMI1* locus (Supplementary Figure S3.8A). Interestingly, *SKIDA1* is upregulated in C5 and C9, compared with WT, and its paralogue, *EPOP* is downregulated (Supplementary Figure S3.8C). Another example is related to *PCDH9* downregulation in C5 and C9. Upon *EZH2* depletion, H3K27me3 and H2AK119Ub are lost at the boundaries of an ~6Mb domain covering multiple long non-coding RNAs and *PCDH9* (Supplementary Figure S3.8B). Surprisingly, this leads to the formation of multiple loops within that domain and the downregulation of *PCDH9* (Supplementary Figure S3.8B, D).

3.5.7 Changes in chromatin organisation reveal activation of a LIN28B signature associated with CDK6 overexpression

We observed an increase in *LIN28B* expression upon PRC2 depletion in our clones (logFC = 0.95, adjusted p-value = 0.04) (Supplementary Table S3.2). LIN28B is a fetal lymphopoiesis marker and is rarely expressed in adult haematopoiesis (Yuan et al. 2012) and due to its previously reported associations with EZH2 depletion, we wanted to investigate further (Oshima et al. 2016; Jacobsen et al. 2020). Strikingly, at the 3D chromatin level we found potential changes in compartmentalisation near *LIN28B* that could correlate with increased transcriptional activity, reflected by a “bowtie” shape (Figure 3.6A). These changes were not directly due to H3K27me3 depletion, as the locus is already derepressed in OCI-AML2 and there is some level of *LIN28B* expression in this model according to our own results, and RNA-seq data from the Cancer Cell Line Encyclopaedia (CCLE) (Supplementary Figure S3.7E).

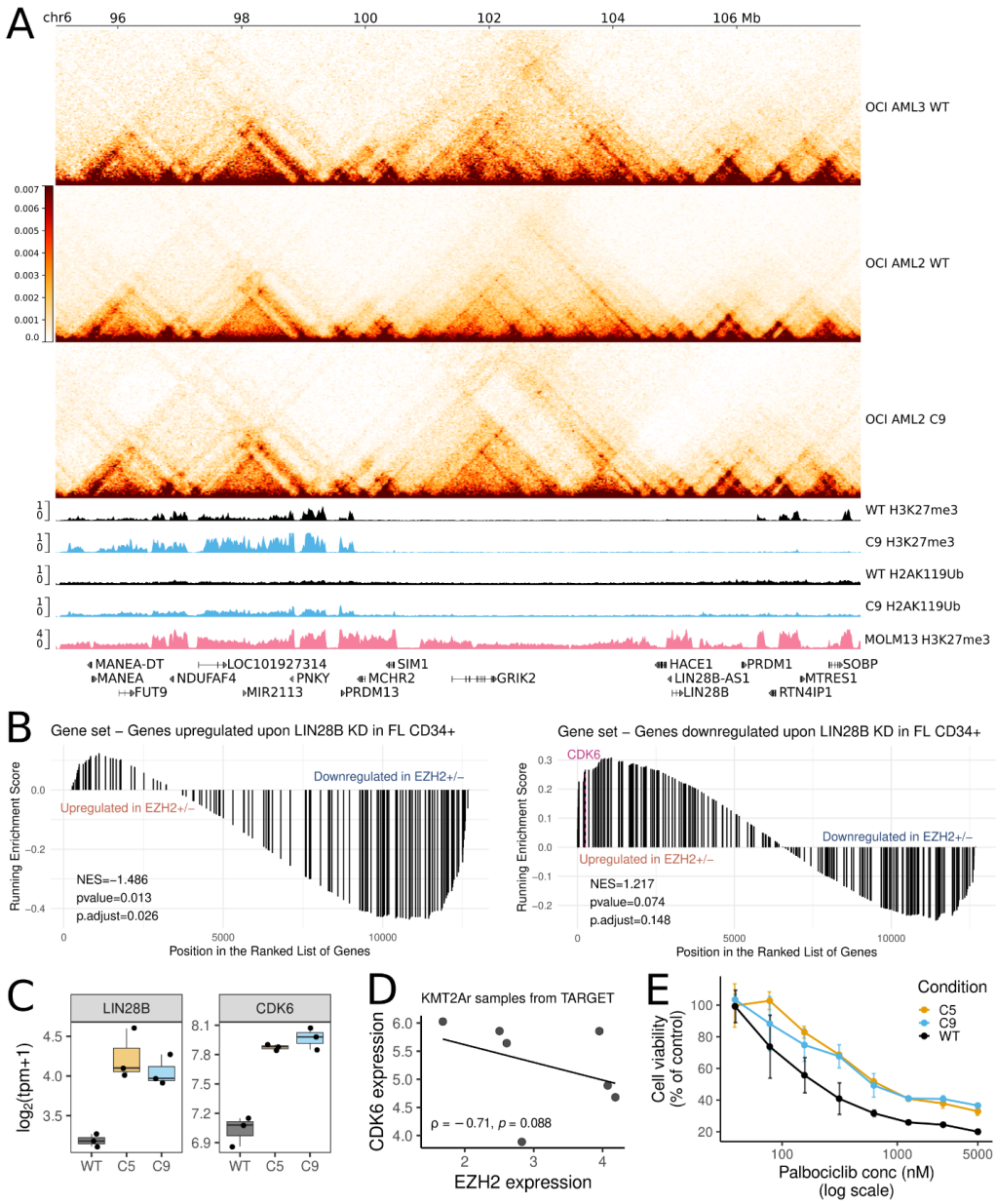


Figure 3.6 | Changes in chromatin organisation reveal a LIN28B signature contributing to drug resistance through *CDK6* overexpression. A, Hi-C matrices for OCI-AML3, OCI-AML2 WT and OCI-AML2 C9 at the *LIN28B* locus at 30 kb resolution. H3K27me3, H2AK119Ub tracks for WT (black) and C9 (cyan) and H3K27me3 track from MOLM13 publicly available CUT&RUN data (see Methods). **B,** Gene set enrichment analysis on logFC ranked genes comparing *EZH2*+/- (C5 and C9 together) against *EZH2*+/+ cells. Gene sets tested were: genes up- and down-regulated upon *LIN28B* KD in foetal liver CD34+ cells (See Methods). **C,** *LIN28B* and *CDK6* mRNA expression in OCI-AML2 WT, C5 and C9. **D,** Spearman correlation between mRNA expression of *CDK6* and *EZH2* in patients with *KMT2A* translocations from the TARGET study. **E,** Palbociclib treatment in OCI-AML2 WT, C5 and C9. F. Cell proliferation assay on OCI-AML2 WT, C5 and C9.

In contrast, the *LIN28B* locus is decorated with H3K27me3 in MOLM13, which is reflected at the RNA level (Figure 3.6A, Supplementary Figure S3.7E). In OCI-AML2 C9 we observed an increase in H3K27me3 and H2AK119Ub >5 Mb upstream of *LIN28B*, potentially contributing to the insulation of this large region.

To test whether *LIN28B* upregulation correlates with the activation of a LIN28B-regulated transcriptional program, we used RNA-seq data from CD34+ fetal liver (FL) cells perturbed with *LIN28B* shRNA. We derived two gene signatures from these data: genes upregulated upon *LIN28B* knock-down (KD) and genes downregulated upon *LIN28B* KD. We saw that genes upregulated upon *LIN28B* KD in CD34+ FL cells are downregulated in PRC2-depleted cells (NES = -1.486) (Figure 3.6B). Additionally, genes downregulated with LIN28B KD are upregulated in PRC2-depleted cells (NES = 1.217) (Figure 3.6B). We found upregulation of LIN28B target genes involved in alternative lineages such as *ETS1*, *RAB7B* and *ZNF43* (John et al. 2008; He et al. 2011; González-Lamuño et al. 2002). Furthermore, we found that *CDK6* was highly downregulated in *LIN28* KD CD34+ FL cells (logFC = -1.53; adjusted p-value = 3×10^{-39}). *CDK6* was also upregulated in PRC2-depleted cells, potentially due to *LIN28B* overexpression (logFC = 0.89, adjusted p-value = 0.003) and showed a strong, but statistically insignificant correlation with *EZH2* expression in KMT2Ar AML patients (Figure 3.6C, D). CDK6 is a cyclin-dependent kinase essential for the transition of cells from G1 to S phase and is essential for the progression and survival of many cancer types (Goel et al. 2022). Cell cycle genes, including *CDK6*, have been reported to be directly inhibited by let-7 miRNAs, which act as tumour suppressors, unless repressed by LIN28B (Johnson et al. 2007). We wished to know whether this altered expression of *CDK6* might lead to altered sensitivity to CDK6 inhibition. To do so, we used palbociclib, a CDK4/6 inhibitor currently approved for use in select breast cancer subtypes and in clinical trials for additional cancers, including AML (Alves et al. 2021; Fröhling et al. 2016; Kadia et al. 2018). We found that PRC2-depleted cells were significantly more resistant to palbociclib-induced cell death than WT (Figure 3.6E). This result is in line with an elevated CDK6 activity in this setting, and correlates with reports in other cancers where *CDK6* amplification has been reported to promote resistance to CDK6 inhibition (Yang et al. 2016).

3.6 Discussion

We provide a comprehensive epigenomic characterisation and mechanistic insights into the effects of PRC2 depletion in AML cells. Specifically, results were largely generated from the OCI-AML2 cell line model that harbours the *MLL*::*AFDN* translocation that is linked to poor treatment response. We notably find that heterozygous loss of *EZH2* causes large increases in genome-wide accessibility, decreased genome-wide H3K27me3 and H2AK119Ub, and changes in gene expression. The epigenomic and transcriptomic changes driven by *EZH2* depletion are in line with activation of alternative lineage signatures, including a LIN28B-driven program that leads to increased *CDK6* expression and decreased sensitivity to palbociclib-mediated death in *EZH2*-depleted cells.

While we observed strong correlations between ATAC-seq profiles and transcription, it is important to keep in mind that increased chromatin accessibility is not synonymous with increased gene expression, as these regions may just be 'primed' for activation of associated enhancers or transcription at neighbouring promoters. We do however see activation of a hMDP transcriptional signature in our cell line model and in patients with *KMT2A*r and low *EZH2* expression. This is accompanied by decreased expression of monocytic genes such as *CD14* in each setting. Taken together, these results suggest that although PRC2 depletion "opens up" the chromatin near genes involved in development, these cells may not be in the appropriate context to express all the alternative transcriptional programs. OCI-AML2 is a myelomonocytic cell line and so may not be able to fully activate alternative transcriptional programs, even if the PRC2 "block" is removed. In other contexts and in early development, it has been demonstrated that depletion of PcGs can activate lineage inappropriate genes and PcGs are thought to stabilise lineage specification (Illingworth et al. 2016; Pivetti et al. 2019; Hölzenspies et al. 2024). Taking into consideration the contrasting roles of KMT2A and PRC2 in gene regulation, caution should be taken in extrapolating these findings to non-*KMT2A*-rearranged contexts.

We also showed that PRC2 depletion may contribute to the modification of the nucleosomal landscape. In particular, we observed a decreased

nucleosome occupancy score and increased fuzziness for nucleosomes near TSS. This means that whilst nucleosomes remain, on average, at their normal positions, they are more disorganised and may have more “wiggle room" upon PRC2 depletion in AML cells. A role for PRC2 in nucleosome positioning has not been fully clarified. Whilst we know that PRC1 depletion drastically changes the nucleosome landscape, *SUZ12* depletion has been reported to have no effect on nucleosome occupancy (King et al. 2018). However, Prorok et al. (2023) show that *EZH2* depletion in mouse embryonic stem cells results in nucleosome repositioning at *HOX* gene promoters that are typically repressed by PcG proteins.

While the role of PRC2 in 3D chromatin architecture in normal cells has been extensively studied (Vieux-Rochas et al. 2015; Wani et al. 2016; Cai et al. 2021), a similar role in leukaemia has not been rigorously evaluated. Our analysis suggests that chromatin architecture is largely maintained upon heterozygous loss of *EZH2* in AML cells. This was despite some experimental limitations such as lack of replicates and integration of different data sources. Strikingly, H3K27me3 was maintained in regions with high contact density, and loss of H3K27me3 largely occurred in regions with no 3D chromatin looping. This suggests a structural role for H3K27me3 regions gained upon PRC2, with very little involvement in direct gene regulation at these loci, as supported by transcriptomics. The role of structural H3K27me3 upon heterozygous loss of EZH2 has not been extensively studied and may play a role in keeping intact 3D genome architecture. Experiments in mouse embryos and oocytes have shown that both PRC1 and PRC2 play an important role in the formation and maintenance of polycomb-associated domains in early development (Du et al. 2020). In haematopoiesis, H3K27me3-mediated very long range interactions are essential in HSPCs and treatment with an EZH2 inhibitor leads to relaxation of these loops and cell differentiation (X. Zhang et al. 2020). Furthermore, Kraft et al. (2022) show that polycomb-mediated 3D chromatin contacts are essential for H3K27me3 spreading, potentially explaining why upon heterozygous *EZH2* depletion the gained H3K27me3 marks in our model can be found at regions with high DNA-DNA contact frequency.

Additionally, we found a change in compartmentalisation upstream of the *LIN28B* gene. We observed *LIN28B* overexpression upon *EZH2* depletion,

leading to an activation of a LIN28B signature, as previously described in other blood cell models (Basheer et al. 2019; Oshima et al. 2016). LIN28B is an RNA-binding protein that mainly acts as an inhibitor of let-7 micro-RNAs (miRNA). Let-7 miRNA have tumour suppression functions and their inhibition leads to upregulation of oncogenes and cell-cycle genes such as *MYC*, *RAS, CDK6* and *HMGA2* (Balzeau et al. 2017; Tianzhen Wang et al. 2015). *LIN28B* is rarely expressed in adult haematopoiesis and is a fetal lymphopoiesis marker (Yuan et al. 2012). However, it is found activated in more than 20 cancer types, as summarised by Zhou et al. (2013) and it can be activated by multiple signalling pathways such as MAPK, WNT and NFkB (Tianzhen Wang et al. 2015). Upregulation of *LIN28B* upon PRC2 depletion further points towards the partial activation of alternative lineage transcriptional programs. Whilst we did not investigate the changes of let-7 miRNAs, we observed a strong *LIN28B* signature as defined by expression patterns in *LIN28B*-KD CD34+ fetal liver cells.

Part of the *LIN28B* regulated gene signature is the *CDK6* gene, which is upregulated upon PRC2 depletion. CDK6 is a cyclin dependent kinase important for G1/S transition that can be overexpressed in T- and B-ALL (Nebenfuehr et al. 2020). Furthermore, *CDK6* is often upregulated in *KMT2A*r AML and is a direct target of KMT2A fusion proteins (Placke et al. 2014). Therefore, CDK6 is an attractive target for *KMT2A*r AML, supported by data from Placke et al. (2014) showing that palbociclib treatment leads to decreased growth of cell lines harbouring *KMT2A*::*MLLT3* fusions. Furthermore, *KMT2A*r AMLs treated with palbociclib showed increased expression of myeloid differentiation markers such as CD11b (Placke et al. 2014). As we have shown, myeloid markers including CD11b are downregulated at the mRNA level upon PRC2 depletion in AML cell lines.

One drawback of our study is that our results are influenced by the genetic context of the OCI-AML2 cell line, which already harbours leukaemia driving alterations (i.e., *KMT2A*r and *DNMT3A* mutation). Our analyses of patient transcriptional data (Figure 3.2D, 2F) suggest that the presence of a *KMT2A* fusion heavily influences our results. As expected for model systems, we observed epigenetic and transcriptomic heterogeneity in the two *EZH2*+/- clones C5 and C9. Furthermore, we have only investigated the H3K27me3 and

H2AK119Ub landscape in C9. Further investigation of marks such as H3K27ac and H3K4me1 may unravel mechanisms of transcriptional activation through enhancers becoming active upon *EZH2* depletion. Furthermore, genome-wide H3K4me3 may offer insights into the role of bivalent promoters in gene regulation in a heterozygous *EZH2* depletion context. However, we observed consistent similarities in chromatin accessibility near development-related genes and transcriptional activation of alternative lineage genes, including *LIN28B*.

While our data shed light on the epigenetic consequences of *EZH2* depletion in leukaemia, several important questions remain. Key to understanding the role of EZH2 in leukaemia are the dynamics via which PcGs suppress alternative lineages during haematopoiesis, similarly to what has been recently described during endodermal differentiation (Hölzenspies et al. 2024). Single-cell studies investigating transcription, chromatin accessibility and response to perturbation may be able to disentangle the role of *EZH2* in lineage priming. Furthermore, inter-patient heterogeneity may be driven by the presence of other oncogenic alterations, such as *KMT2A* in the patient data we presented. There is a need to study the effects of alterations in different contexts and how the relationships between combinations of alterations drive leukaemogenesis. Finally, the role of EZH1 in this context has not been thoroughly explored, especially as we observed broad H3K27me3 domains gained at certain genomic loci, which may suggest higher EZH1-PRC2 inter-nucleosomal spreading.

3.7 Acknowledgements

This research was funded by Science Foundation Ireland through the SFI Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214 and supported in part by the EU's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant H2020-MSCA-COFUND-2019-945385.

Work in the Bond laboratory is supported by Science Foundation Ireland grants 20/FFP-P/8844 and 18/SPP/3522, the latter together with Children's Health

Ireland. Work in the Ryan laboratory is supported by Science Foundation Ireland grant 20/FFP-P/8641. Hi-C experiments were supported by a Haematology Association of Ireland Career Development Award to LJ.

We thank Dr. Eric Conway, Dr. Aleksandar Krstic and members of the Bond, Ryan and Vaquerizas labs for useful suggestions on experiments and analyses.

The results published here are in part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (https://www.cancer.gov/ccg/research/genome-sequencing/target) initiative, phs000465.

3.8 Code availability

Code to run RNA-seq, CUT&RUN, ATAC-seq, and part of Hi-C analyses is available at:
https://github.com/cosmintudose/PRC2_AML_chromatin/releases/tag/code_draft
Code to run Hi-C analyses was provided by the Vaquerizas lab and is partly based on:
https://github.com/vaquerizaslab/Ing-Simmons_et_al_dorsoventral_3D_genome

3.9 Supplementary Methods

3.9.1 Immunoblotting

Cell lysis: Cells were harvested (centrifugation at 350 xg) and pelleted. Protein lysates were obtained from a minimum of 1×10^6 cells lysed in a home-made lysis buffer (1M Tris pH7.5, 5M NaCl, and 0.5% (v/v) NP40, H_2O, filtered). For each cell lysis, an aliquot of the stock solution was supplemented with protease and phosphatase inhibitors (1 Tablet of each/10mL, COmplete Mini Protease Inhibitor Cocktail Tablets – Roche 11836153001; PhosSTOP Phosphatase Inhibitor Cocktail Tablets – Roche 4906837001). Briefly, cell pellets were resuspended with a minimum of 120 µL of lysis buffer. Tubes were briefly

vortexed and left on ice for 15 minutes before being vortexed again and centrifuged for 10 minutes at 18,000 xg at 4°C. The supernatant (i.e., cell lysate) was collected and was either used immediately or stored at -20°C. To assess the protein concentration in cell lysates, the Pierce™ BCA Protein Assay Kit (Thermo Scientific - cat number: 23227) was used according to the manufacturer's instructions..

Western blotting: 10 to 20 µg of protein/sample were mixed (4:1 v/v) with a mix of NuPAGE LDS Sample Buffer (1:4) containing 1M dithiothreitol (DTT), incubated for 5 minutes at 95°C and then resolved on 10% acrylamide gels via electrophoresis. Next, separated proteins were transferred to a methanol-activated PVDF membrane using a semi-dry transfer method (transfer buffer: 10% 20x Tris-glycine and 20% methanol in MiliQ water). The membrane was then blocked with 5% (w/v) skimmed milk powder dissolved in 1x TBS-T (Tris Buffered Saline-Tween) shaking for 1 hour at room temperature.

The membrane was then incubated overnight with primary antibody at 4°C (Supplementary Table S3.1). After washing with 1x TBS-T, the membrane was incubated for 1 hour with an HRP-conjugated secondary antibody. An in-house enhanced chemiluminescence (ECL, Solution 1: 1M Tris/HCl pH8.5, p-coumaric acid, luminol and H_2O; Solution 2: 1M Tris/HCl pH8.5, 30% Hydrogen peroxide (H_2O_2), and H_2O) solution (1:1) was added to the membrane to allow detection using the Advanced Molecular Vision Chemi Image Unit of the ChemoStar Imager (INTAS Science Imaging Instruments GmbH).

3.9.2 *In vitro* cytotoxicity assay

100,000 cells were seeded in 1 mL of complete media in 24-well plates and incubated at 37 ℃ in 5% O_2 in a humidified incubator. Viable cells were manually counted using a haemocytometer following trypan blue staining at 24, 48, 72 and 96 hours.

3.9.3 CUT&RUN

Binding cells to Concanavalin A beads and primary antibody: 500,000 OCI-AML2 WT, C5, and C9 cells per condition were harvested at 700 xg for 3 minutes and washed in 100 µL Digitonin-Wash buffer twice before resuspension in 200 µL Digitonin-Wash buffer. 10 µL activated-Concanavalin A beads per sample (Cell Signaling Technology, Cat. 93569S) were added to the resuspended cells and placed on the rotator to incubate for 10 minutes at room temperature.

Beads were isolated using a magnetic rack and resuspended in a 200 µL buffer containing the antibody of interest (Supplementary Table S3.1) in a 1:50 (v/v) final concentration. These tubes were rotated overnight at 4°C.

Binding of pAG-MNase and targeted chromatin digestion: Beads were isolated using a magnetic rack following antibody binding and resuspended in 200 µL Digitonin-Wash buffer. pAG-MNase was then added to the tube in a final concentration of 0.9 ng/µL and incubated on a rotator for 1 hour at 4°C.

Following incubation, the beads were isolated using a magnetic rack and resuspended in a 150 µL Digitonin-Wash buffer and allowed to chill to 0°C for 5 minutes. 3 µL 100 mM CaCl2 were added to each tube to activate the pAG-MNase. After an hour at 0°C incubation, 150 µL 2xSTOP-buffer was added to each tube and incubated at 37°C for 10 minutes to release soluble chromatin fragments.

DNA isolation and purification: Soluble chromatin fragment DNA was isolated using the MinElute PCR Purification kit (Qiagen, Cat. 28004) following the manufacturer's instructions. In short, 5 volumes of Buffer PB were added to 1 volume of digested samples, placed on the provided MinElute column, centrifuged at 17,900 xg for 1 minute, then washed with 750 µL Buffer PE, centrifuged twice for 1 minute each at 17,900 xg, then eluted in 21 µL Elution buffer. Using the Qubit dsDNA High-Sensitivity kit (Invitrogen, Cat. Q32851), 1 µL DNA was quantified on the Qubit fluorometer.

3.9.4 ATAC-seq

Harvesting cells and lysis: 100,000 cells from each of the conditions (OCI-AML2 WT, C5 and C9) were harvested at 500 xg at 4°C for 5 minutes and subsequently lysed in an ice-cold ATAC Lysis buffer, after which the samples were spun down at 1000 xg at 4°C for 10 minutes.

Tagmentation and DNA purification: The collected nuclei were then resuspended in 50 µL Tagmentation Master mix, containing assembled transposons, 10% Tween-20, and 1% Digitonin. Tagmentation of the samples was performed in a thermomixer set at 800 rpm for 30 minutes at 37°C. Tagmented DNA was purified using the provided DNA purification columns and finally eluted in 35 µL elution buffer.

3.9.5 *In vitro* efficacy of palbociclib against AML cell lines

Cells were plated at 8,000 cells in 100µL media/well in U-bottom tissue culture plates. Cells were equilibrated for 2 hours at 37 °C, 5% CO2 prior to drug treatment. Palbociclib (MedChemExpress) was serially diluted (1:2) in culture media and dilutions added in triplicate wells (final concentration range 0-5 µM, with control wells containing 0.1% DMSO). Following 72 h drug exposure, cell viability was assessed by mitochondrial activity assay (Resazurin cell viability assay). Resazurin solution (0.6 mmol/L Resazurin, 0.07 mmol/L Methylene Blue, 1 mmol/L potassium hexacyanoferrate (III), 1 mmol/L potassium hexacyanoferrate (II) trihydrate) was added to all wells and plates incubated for a further 4 h at 37 °C, 5% CO2. Luminescence was read using a SpectraMax M3 plate reader (Molecular Devices) with excitation at 560 nm and emission at 590 nm. Cell viability was calculated as the percentage of untreated controls. IC50 values were calculated from cumulative dose response curves.

3.9.6 Publicly available data

Gene signatures based on scRNA-seq from the Atlas of Blood Cells were downloaded from http://scrna.sklehabc.com/ on the 14th of July 2023. These data were published by Xie et al. (2021).

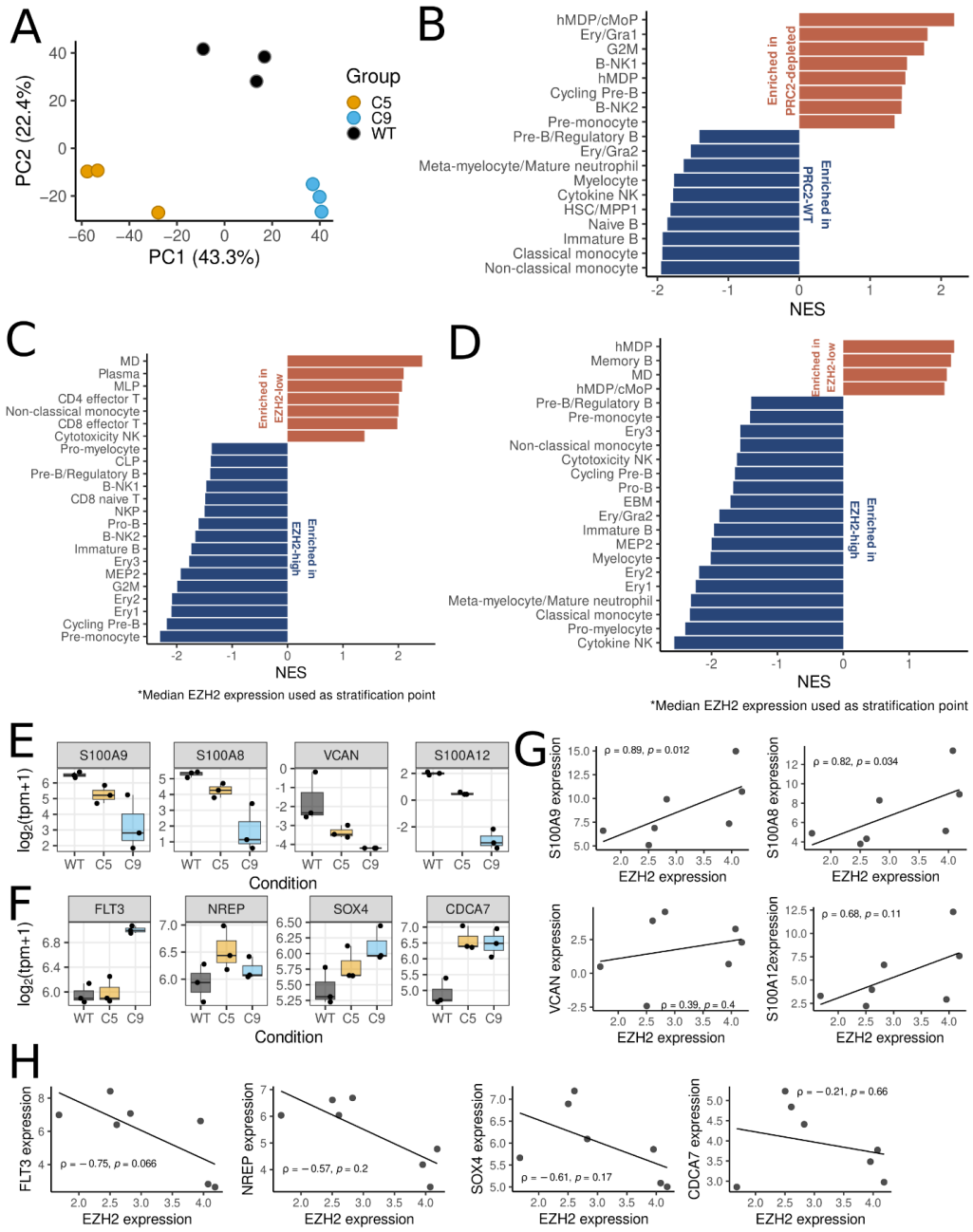
H3K27me3 CUT&RUN signal in .bigwig format and peak calls in .broadPeak format from the MOLM13 cell line were downloaded from GEO, accession number GSE221701, published by Agrawal-Singh et al. (2023). H3K27me3 ChIP-seq peaks in .bed format from the HL-60 cell line were downloaded from GEO, accession number GSE175082, published by ENCODE Project Consortium (2012).

K562 Micro-C data in .mcool format was downloaded from GEO, accession number GSE206131, published by Barshad et al. (2023). OCI-AML2 Hi-C data were downloaded in .fastq format from The European Genome-phenome Archive at the European Bioinformatics Institute, study ID EGAD00001006447, dataset ID EGAS00001004743, published by Takayama et al. (2021).

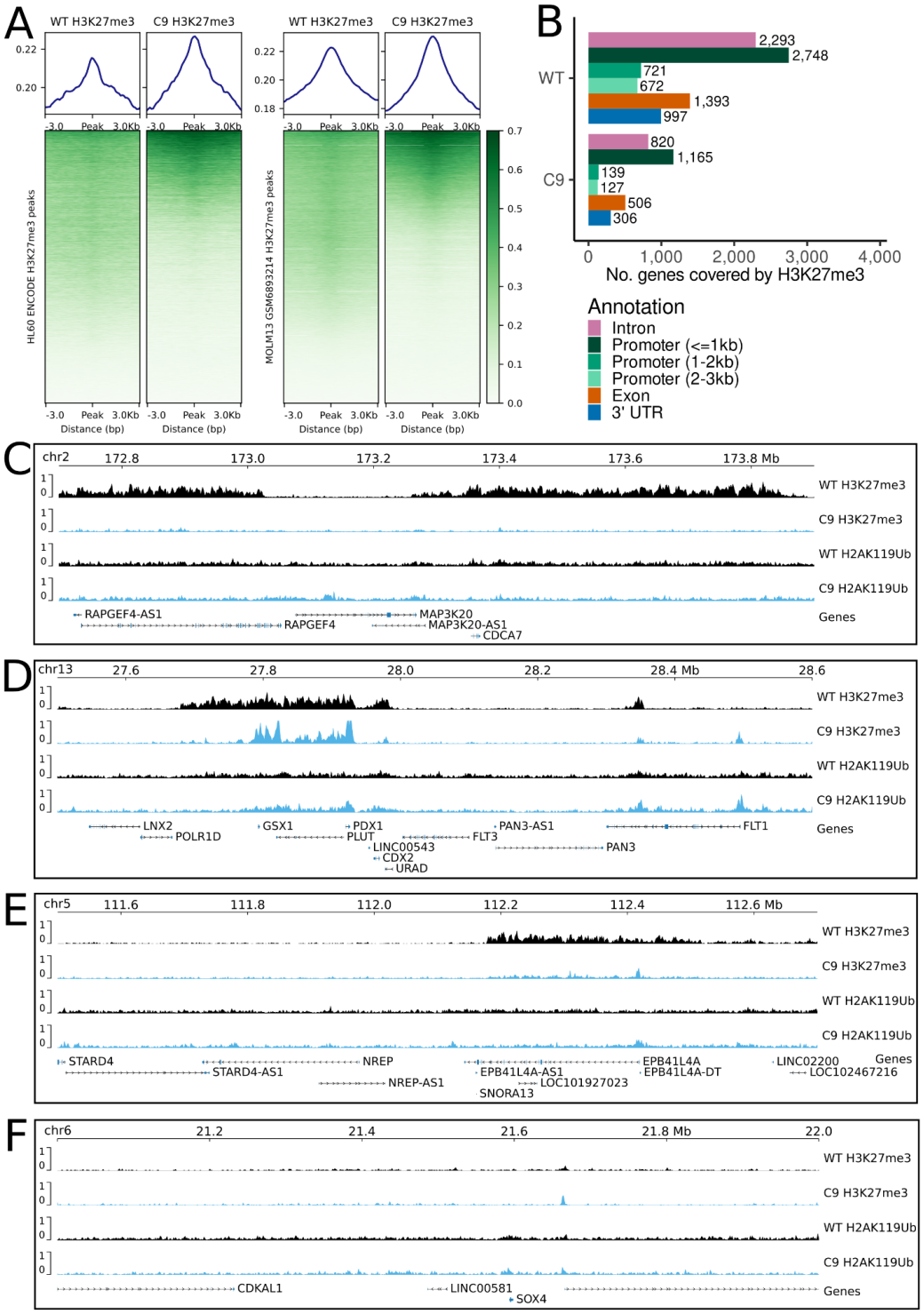
We downloaded RNA-seq and clinical data from the paediatric AML TARGET study from cBioportal (Cerami et al. 2012; Gao et al. 2013; De Bruijn et al. 2023).

3.10 Supplementary Data

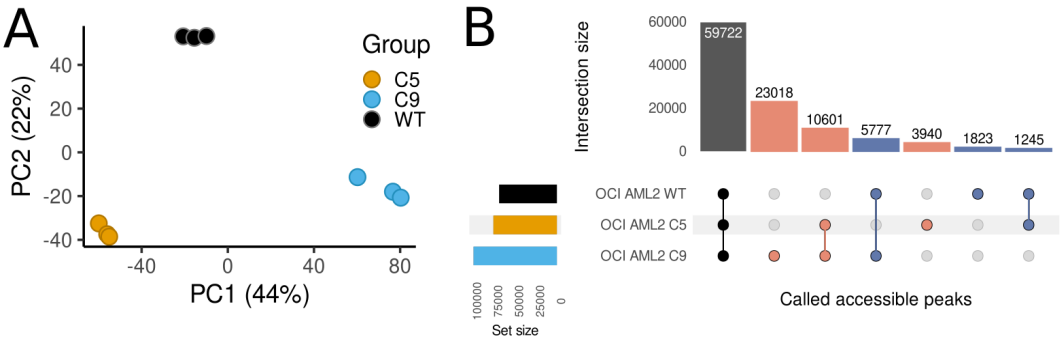
Supplementary Figure S3.1 | Sanger sequencing chromatogram showing CRISPR-Cas9 targeted region of *EZH2 (*exon 3) in WT, C5 and C9.



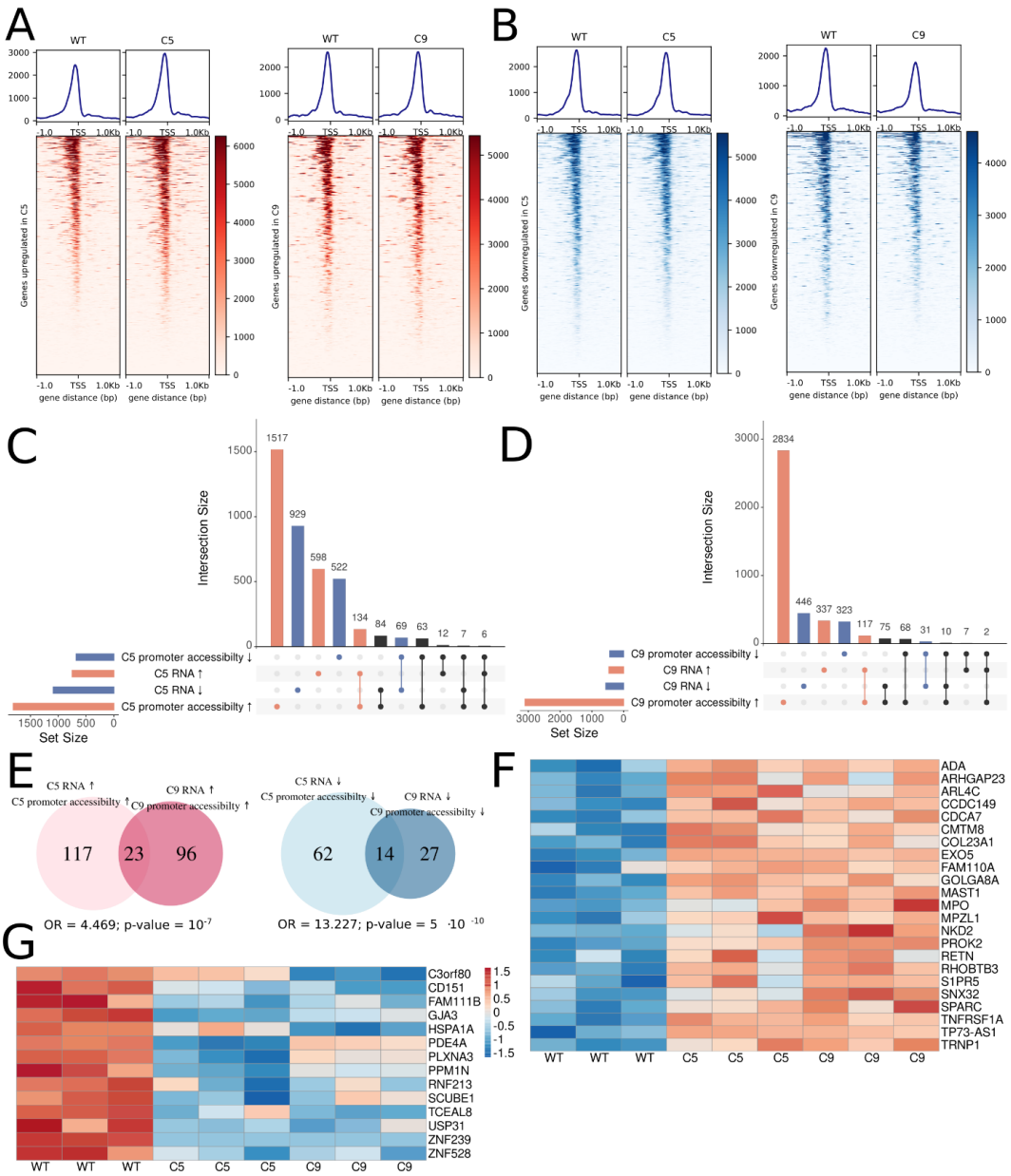
Supplementary Figure S3.2 | Correlations between EZH2 depletion and genes involved in monocytic differentiation. A, PCA of OCI-AML2 RNA-seq - all expressed genes. **B,** GSEA performed on ranked DEGs in OCI-AML2 *EZH2*+/- cells vs *EZH2*+/+ cells using gene sets from the Human Blood Atlas. FDR = 5%. **C, D,** GSEA performed on ranked DEGs in EZH2-low vs EZH2-high TARGET AML samples from patients using gene sets from the Human Blood Atlas. FDR = 5% **C,** All samples used for the analysis **D,** Only KMT2Ar cases. **E,** Monocytic gene expression in WT, C5 and C9. **F,** hMDP/cMOP genes expression in WT, C5 and C9. **G,** Correlations between monocytic genes and EZH2 expression in KMT2Ar TARGET samples. **H,** Correlations between hMDP/cMOP gene expression and *EZH2* expression in KMT2Ar TARGET samples.



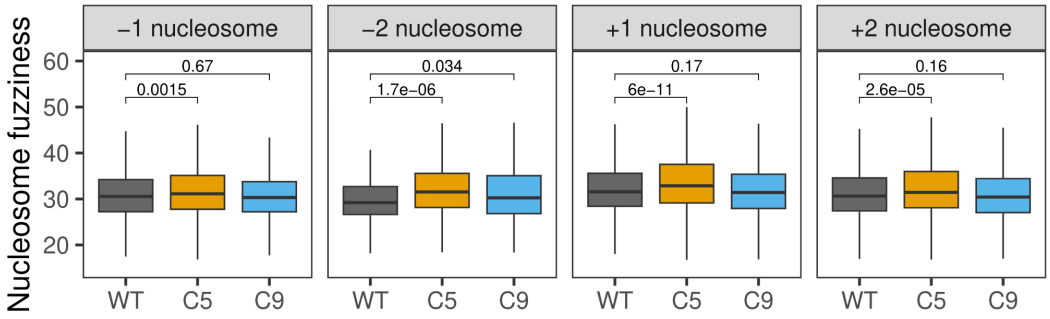
Supplementary Figure S3.3 | Differences in H3K27me3 and H2AK119Ub between WT and EZH2-deficient clone C9. A, WT and C9 H3K27me3 signal at H3K27me3 peaks called in HL60 and MOLM13 from publicly available data (see Methods). **B,** Number of genes where there is any called H3K27me3 peak in WT and C9, stratified by location within the gene body. **C, D, E, F,** CUT&RUN tracks for WT (black) and C9 (cyan) at genes enriched in the GSEA from Supplementary Figure S3.2B in the hMDP/cMoP signature **C,** *FLT3*. **D,** *CDCA7*. **E,** *NREP*. **F,** *SOX4*.



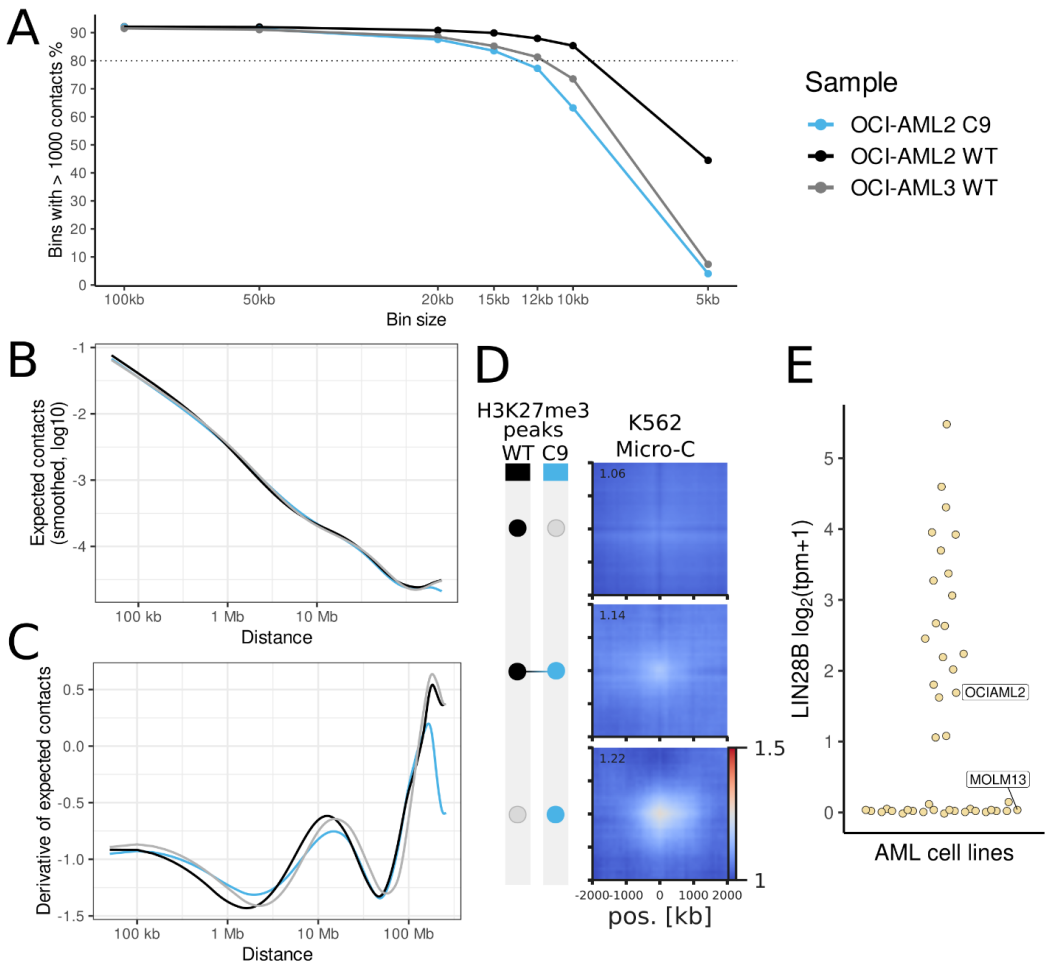
Supplementary Figure S3.4 | ATAC-seq reveals changes in chromatin accessibilityWT C5 and C9. A, PCA of ATAC peaks from MACS2. **B,** Peak overlaps of peaks called by HMMRATAC.



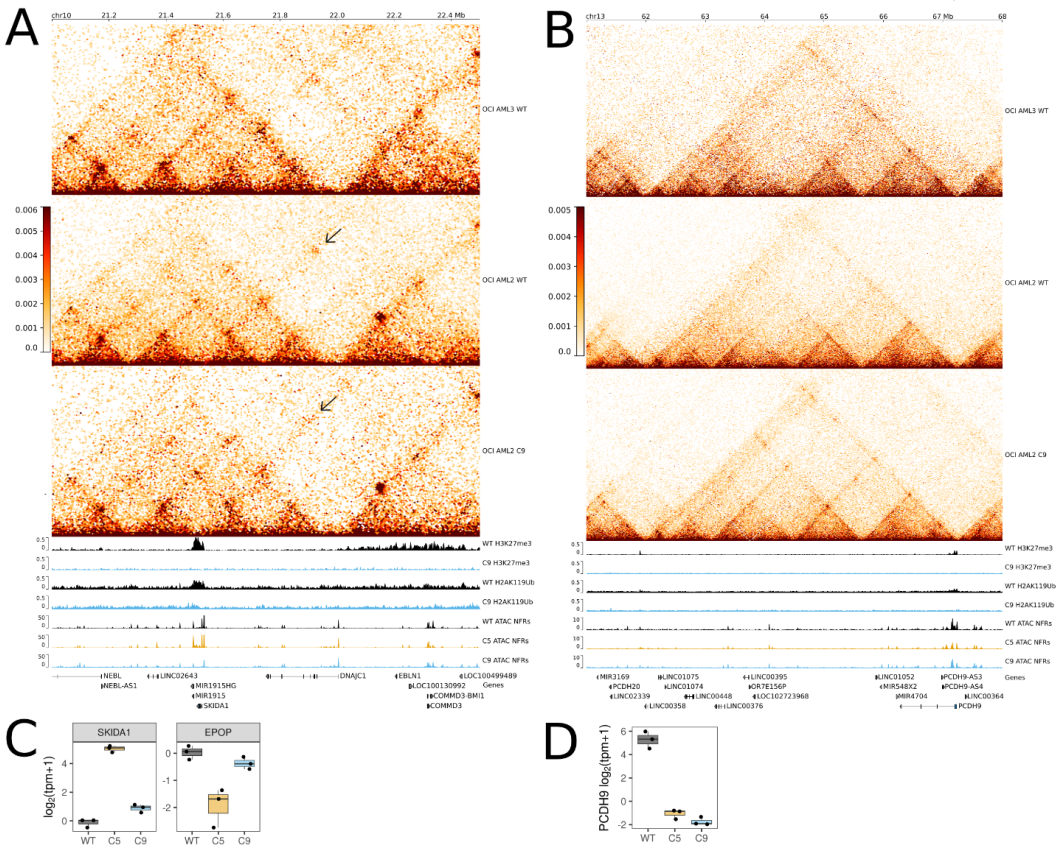
Supplementary Figure S3.5 | Correlation between chromatin accessibility and gene expression. A, B, Nucleosome free region (NFR) signal at TSS of genes upregulated (**A**) and downregulated (**B**) in C5 and C9, respectively, compared with WT. **C, D,** Overlaps between genes with changed chromatin accessibility at the promoter and changed gene expression in C5 (**C**) and C9 (**D**) compared to WT. **E,** Overlaps of genes with increased promoter accessibility and increased RNA expression in C5 vs WT and C9 vs WT (red) and overlaps of genes with decreased promoter accessibility and decreased RNA expression in C5 vs WT and C9 vs WT. **F,** Genes with increased RNA expression and increased promoter accessibility in both C5 and C9 vs WT. **G,** Genes with decreased RNA expression and decreased promoter accessibility in both C5 and C9 vs WT.



Supplementary Figure S3.6 | Nucleosome fuzziness scores. Nucleosome fuzziness at -2, -1, +1 and +2 nucleosomes for WT, C5 and C9. Box = 1st and 3rd quartiles; middle black line = median; whiskers extend to 95% of data points. Values above brackets indicate p-values from Wilcoxon unpaired tests.



Supplementary Figure S3.7 | QC on Hi-C data and validation on Micro-C. A, Resolution of Hi-C data for each sample (see Methods). **B,** Distance-decay curve for each sample **C,** Derivative of the distance-decay curve for each sample **D,** Pileups of contacts between CUT&RUN H3K27me3 peaks in Micro-C data from K562 at 100kb resolution and 2Mb flanking regions. **E,** *LIN28B* expression in the Cancer Cell Line Encyclopaedia (CCLE) in AML cell lines (N = 43).



Supplementary Figure S3.8 | Integration of epigenomic analysis results at selected loci. **A, B,** Hi-C (OCI-AML3, OCI-AML2 WT and C9), H3K27me3 (WT - black and C9 - cyan), H2AK119Ub (WT - black and C9 - cyan) and ATAC-seq (WT - black, C5 - orange and C9 - cyan) tracks at **A,** *SKIDA1*. **B,** *PCDH9.* **C,** Expression of *SKIDA1* and *EPOP* in WT (black), C5 (orange) and C9 (blue). **D,** Expression of *PCDH9* in WT, C5 and C9

Supplementary Table S3.1 | List of antibodies

Protein	Supplier	Catalogue Number	Species	Dilution
EZH2 (D2C9) XP® Rabbit mAb #5246	Cell Signaling Technology	#5246	Rabbit	1:1000
Tri-Methyl-Histone H3 (Lys27) (C36B11) Rabbit mAb #9733	Cell Signaling Technology	#9733	Rabbit	1:1000 1:100
β-Actin (13E5) Rabbit mAb #4970	Cell Signaling Technology	#4970	Rabbit	1:1000
Ubiquityl-Histone H2A (Lys119) (D27C4) XP® Rabbit mAb	Cell Signaling Technology	#8240	Rabbit	1:100
IgG XP® Isotype Control (DA1E) XP® Rabbit mAb	Cell Signaling Technology	#3900S	Rabbit	1:100

Supplementary Table S3.2 | Differential expression analysis OCI-AML2 EZH2+/- vs *EZH2*+/+
Supplementary Table S3.3 | GSEA on cell lines and patient data using ABC gene sets
Supplementary Table S3.4 | Annotated H3K27me3 and H2AK119Ub called peaks in WT and C9
Supplementary Table S3.5 | Annotated ATAC open chromatin regions in WT, C5 and C9
Supplementary Table S3.6 | Homer analysis of TFs enriched at regions more accessible in clones, compared to WT
Supplementary Table S3.7 | GO:BP enrichment with GREAT tool on regions more accessible in clones, compared to WT

Supplementary tables available at: https://bit.ly/suppltableschapter3

CHAPTER 4 - General discussion

4.1 Summary of major findings

Using computational approaches, I have investigated gene regulation in cancer both systematically and in a specific context in which epigenetic factor function is altered in leukaemia.

In Chapter 2 we performed the first systematic analysis of whether GRN-inferred activity can effectively predict gene essentiality in cancer cells. Briefly, we have made the following advancements:

- GRNs maintain some level of cancer-type specificity, with cancer type -matched GRNs performing better than cancer type-mismatched GRNs at predicting gene sensitivity to inhibition.
- However, even cancer type-matched GRNs perform worse than gene expression at predicting gene sensitivity to inhibition.
- Increased sensitivity to gene inhibition is more commonly correlated with increased expression, rather than decreased expression, suggesting oncogene addiction-like effects.
- Binarising gene sensitivity to inhibition and separating genes into two groups: essential and non-essential yields similar results, with gene expression performing better than GRN-inferred activity.

In Chapter 3, we took a focused approach and investigated the role of the EZH2 methyltransferase in an AML context. We believe this is the most complete integrated analysis of the epigenomic and transcriptomic consequences of heterozygous *EZH2* loss in AML performed to date. Briefly, we made the following discoveries:

- *EZH2* depletion leads to “priming” of the chromatin and partial activation of alternative lineage transcriptional programs.
- Upon heterozygous loss of *EZH2,* genome-wide H3K27me3 marks decrease by ~70%, accompanied by a moderate decrease in the complementary H2AK119Ub repressive mark and marked genome-wide increases in chromatin accessibility.

- H3K27me3 was preferentially retained at genomic regions with high looping frequency upon *EZH2* depletion.
- 3D chromatin architecture changes included altered contacts at the *LIN28B* locus, with transcriptional and functional evidence of activation of a LIN28B transcriptional program that includes increased *CDK6* expression and decreased sensitivity to CDK6 inhibition.

Both systematic and focused approaches to study gene regulation make use of computational approaches to extract information from the large amounts of data that molecular and functional profiling can generate. Both approaches have advantages and limitations that I will summarise in the following sections, alongside a discussion of potential future directions in the field.

4.2 Untangling gene regulation

The complexity and the many layers of gene regulation in multicellular organisms, as visualised in Figure 1.1, may contribute to why GRN inference methods struggle to accurately reconstruct functional relationships between genes. In addition, tissue-specific differences in gene expression make it difficult to extrapolate GRNs between different cancer types. In our analyses, whilst GRNs were poor at finding sensitivities to inhibition, cancer-type matched GRNs performed slightly better than cancer mismatched GRNs, suggesting they may provide some tissue-specific information.

Whilst overall we observed that plain mRNA abundance uncovers more correlations with gene sensitivity to inhibition, there were a few examples where GRN-inferred activity better correlated with gene sensitivity to inhibition. However, there is no reason to believe that genes with higher correlation between activity and essentiality are more important than genes with higher correlation between expression and essentiality without a systematic analysis, and cherry picking results from GRN methods is a pitfall that must be avoided. In fact, a recent paper argued that transcriptomic data may not provide enough total information to enable accurate GRN reconstruction (Kernfeld et al. 2024). However, in Chapter 3 we have shown that methods that incorporate TF binding motif information such as GRNdb do not perform better than methods that only

use transcriptomics at predicting gene sensitivity to inhibition. Perhaps GRNs can be improved by taking advantage of multiome approaches (see subsection 4.3), as showcased by Dictys and CellOracle (Wang et al. 2023; Kamimoto et al. 2023). The two approaches attempt to reconstruct context-dependent GRNs on a developmental continuum rather than a static GRN. This approach has theoretical appeal in cancers that exhibit differentiation arrest, such as leukaemia.

Given the key role that epigenetic factors play in gene regulation, integration of information from epigenomic assays should help to improve GRN performance. However, the results in the third chapter, in which we experimentally perturb a single epigenetic factor, exemplify the challenges involved in predicting cancer-associated GRN dysregulation. Whilst depleting a transcriptional repressor would make us expect more gene upregulation, the observed effects were more complex than predicted. This can be explained by a number of potential causes. Firstly, it is difficult to distinguish between primary effects of transcriptional derepression due to decreased K27me3, and secondary effects of PRC2 loss such as upregulation of factors that inhibit gene expression at other loci. Secondly, EZH1, the EZH2 paralogue, may partially compensate for the loss of EZH2. Whilst EZH1 has a weaker methyltransferase activity, it has a stronger dimerisation activity and can more efficiently spread polycomb groups on neighbouring nucleosomes (Sauer et al. 2023). The possibility of EZH1 compensating is also suggested by the increased width of H3K27me3 peaks, accompanied by the loss of initial-event peaks and the increase in median peak width suggesting increased H3K27me3 spreading. Finally, EZH2 also has non-canonical functions, such as androgen-receptor (AR) binding, which may be affected by its depletion (Wang et al. 2022). These possibilities are currently being explored experimentally by our group.

Overall, we do not see a strong correlation between chromatin accessibility and active transcription. These findings are concordant with Kiani et al. (2022). They found that perturbing single factors results in some genes having concordant RNA and ATAC signals, i.e., increased expression and increased chromatin accessibility, or vice versa, whilst other genes were discordant. In the two experiments performed, only ~33% and ~54% of genes with increased RNA also had increased chromatin accessibility. These results

suggested that altered expression may happen independently of chromatin accessibility. Some potential explanations for this finding are that the chromatin is already accessible at these loci, or that there may be changes in TF binding patterns altering expression or binding of repressive factors. Furthermore, the authors of this report found stronger correlations along the haematopoietic developmental trajectory, where chromatin accessibility guides transcriptional patterns more stringently. Taking together the conclusions of Kiani et al. (2022) and our findings, we believe the relationship between chromatin accessibility and transcriptional upregulation is not linear and shows variability across the AML genome. Other studies have reported similar discordances between chromatin accessibility and gene expression in other cellular contexts, and have found that TF presence and activating histone marks are more predictive of active transcription than accessibility alone (Chen et al. 2021; Tu et al. 2023). Additionally, other factors may have to be taken into consideration to build better models, such as TF binding, mRNA degradation machinery (i.e., the rixosome complex) or RNA-pol II stalling. A systematic study that, in addition to chromatin accessibility, investigates the contribution of all these factors to activating transcription may shed light on how to best interpret chromatin accessibility.

Additionally, we are not taking into consideration the regulation that happens after mRNA production (Figure 1.1). We know that the correlation between mRNA abundance and protein abundance is low at ~0.2–0.5 (Zhang et al. 2014; Zhang et al. 2016; Mertins et al. 2016). Proteomic measurements have in some cases been shown to outperform transcriptomic measurements at predicting sensitivity to inhibition in cancer cell lines (Gonçalves et al. 2022). Additionally, it seems that GRNs are biased towards the data type they are inferred from, with GRN-inferred activity being better correlated with mRNA abundance than with protein abundance (Sousa et al. 2023).

4.3 Intra-tumour heterogeneity

In Chapter 3, to control for potential heterogeneity within the OCI-AML2 cell line, we generated two separate isogenic clones to model heterozygous *EZH2* loss.

In line with clonal variability in CRISPR-generated stable knockout models (Westermann et al. 2022), we observed some epigenetic and transcriptomic differences between the two clones.

To account for clonal variation, a systematic approach to GRN analysis may be more helpful (i.e., studying EZH2 loss in a large sample of tumours - which may be, again, limited by inter-tumour heterogeneity - see subsection 4.4). While PRC2 haploinsufficiency is relatively common (15%) in childhood AML, the rarity of this disease makes it difficult to study the effects of PRC2 depletion systematically in a patient cohort. Advances in single cell approaches have made it possible to study gene regulation in different cell populations within the same tumour via using scRNA-seq (transcriptome) or scATAC-seq (chromatin accessibility). For example, scRNA-seq of glioblastoma patient samples unravelled intra-tumour heterogeneity within the samples, with a tumour-driving subpopulation being characterised by *EZH2* overexpression (Chen et al. 2022).

Performing multiple assays at the single cell level in the same cells may also aid in untangling intra-tumour heterogeneity. Several methods are being developed to tackle this issue. For example, single cell multiome sequencing allows simultaneous scATAC-seq and scRNA-seq within the same single cells, and allows linking of gene expression with epigenetic characterisation (Belhocine et al. 2021). This method has been proven to be useful at linking disease-associated SNPs from genome-wide association studies to putative enhancers that alter the expression of disease-causing genes (Mitra et al. 2024). An important limitation of scRNA-seq methods is that only the transcriptome is used to infer cell identity, and having both protein and mRNA measurements gives a more precise cellular profile. This is possible when studying haematopoiesis and haematological diseases, with the development of CITE-seq (cellular indexing of transcriptomes and epitopes with sequencing). CITE-seq uses DNA-barcoded antibodies to identify cell type-specific surface markers to identify cell populations in an unbiased manner (Stoeckius et al. 2017). Using an extensive CITE-seq based bone-marrow atlas, Zhang et al. (2024) were able to directly map leukaemic stem cells to distinct healthy multilineage progenitors and better characterise the stages of differentiation at which disease may occur. In MPAL, transcriptional programs from multiple

hematopoietic lineages may be active (see subsection 1.4.2). Again, by using CITE-seq combined with scATAC-seq, links connecting TFs, CREs and marker genes were uncovered as potential molecular mechanisms of disease (Granja et al. 2019).

Furthermore, new methods are constantly being developed. Multiome Perturb-seq and CUT&Flow (coupling cleavage under target and tagmentation with flow cytometry) have recently been developed, with the caveat that they are only feasible in tumour cell lines. Multiome Perturb-seq allows the quantification of transcript and chromatin accessibility in a pool of cells perturbed with CRISPRi, allowing integrative analysis of perturbations on cell state (Metzner et al. 2024; Veronezi & Ramachandran 2024). CUT&Flow allows the mapping of genome-wide chromatin across cell types based on cell surface markers. In the original publication it was used to identify H3K27me3 patterns at different points within the cell cycle (Veronezi & Ramachandran 2024). There is potential in the future to use CUT&Flow coupled with different cell surface markers to study multiple cell types within the same tumour.

In evaluating the molecular consequences of epigenetic alterations, a further consideration is that these factors have cell type-specific activities. This applies to the effects of EZH2 loss, as described by Mochizuki-Kashio et al. (2015) and Basheer et al. (2019). The former paper reported that the stage at which *EZH2* depletion occurs dictates AML prognosis in murine models. In these studies, mice transplanted with AML cells with KMT2A::MLLT3 or *AML1::ETO9a* backgrounds accompanied by homozygous *EZH2* loss had lower survival than mice with WT *EZH2* at induction phase. However, at maintenance, mice with *EZH2* loss showed improved survival in both genetic backgrounds. Other unaddressed questions relate to the timing and context in which driver alterations occur. For example, *KMT2A* rearrangements in infant cases may occur in fetal progenitors (Rice et al. 2021; Khabirova et al. 2022), whilst in adult cases they may occur after the accumulation of other mutations or at relapse.

All these concepts link to the key question of how alterations in epigenetic factors may alter cell development in the long term, and even whether transient changes in activity may affect oncogenesis. Very recently, a pioneering paper on this topic has suggested that transient loss of PcGs may be enough to permanently induce a cancer fate, even in the absence of other

cancer-causing mutations (Parreno et al. 2024). This appeared to occur due to a persistent depression of genes involved in tumourigenesis, including JAK/STAT pathway components. The experiments were performed in *Drosophila melanogaster* and it will be very interesting to follow whether these findings can be reproduced in mammalian cells.

4.4 Inter-tumour heterogeneity

In Chapter 2 we mainly focused on major cancer categories i.e., based on tissue type, rather than specific cancer subtypes. Within a given cancer type, distinct subtypes can have different survival responses and distinct responses to therapy. For example, paediatric patients carrying *FLT3*-ITD alterations have different prognosis depending on the co-occurring mutations, with *FLT3*-ITD+ AMLs accompanied by *WT1* mutations and NUP98-NSD1 fusions having the worst prognosis (Tarlock et al. 2018). However, patients with *FLT3*-ITD AMLs accompanied by *NPM1* mutations have better survival than any other *FLT3*-ITD altered AML (Tarlock et al. 2018). If we were to investigate specific subtypes of cancer based on genotype or more granular phenotype (e.g. *KMT2A*r AML instead of all AMLs in one group), we would have come across issues with statistical power. One can envision that a more granular approach may allow for finding more specific dependencies for different subtype backgrounds. However, our goal was to test whether GRN-inferred activity may be able to uncover dependencies within ten major cancer types.

We observed potential inter-tumour heterogeneity effects in our patient data analysis in Chapter 3 as well. When analysing a larger TARGET cohort (N = 45), we were not able to reproduce our main findings from *EZH2*-depleted AML cell lines. However, restricting the analysis to *KMT2A*r samples (N = 7) yielded similar results to our cell line model. This suggests that there are significant differences across this paediatric AML cohort and that variation in *EZH2* expression has distinct consequences across different genetic backgrounds, with multiple potential transcriptional programs being “primed” for activation upon its loss. Variable effects have been observed in *EZH2*-depleted AML, depending on co-occurring alterations, with different transcriptional

programs being activated in *KMT2A*::*MLLT3* and *AML1*::*ETO9a* AML when *EZH2* was depleted (Basheer et al. 2019).

4.5 From cell lines to patients

As alluded to in the previous section, extrapolating results from cell lines to patients can be tricky. Taking into consideration confounding factors such as sex, genetic background and co-occurring mutations would be ideal, but patient cohorts are limited in size and current cell line collections do not adequately represent all subtypes observed in patients (Boehm & Golub 2015). Further, cancer-associated alterations often exhibit statistical collinearity with patient-associated factors e.g., *DNMT3A* mutations and age.

In Chapter 2, we do not see a clear difference in the performance of GRNs inferred from patient data and GRNs inferred from cell lines. In both cases, we concluded that GRN-inferred activity is no better at predicting sensitivity to inhibition than mRNA abundance.

Many assays are only possible in cell lines, therefore taking appropriate measures to modelling diseases found in patients is important. For example, some cell lines may not be appropriate models for commonly studied cancer subtypes (Najgebauer et al. 2020; McCabe et al. 2023) and tools such as Cellector and Celligner attempt to control for this and give predictions of best suited cell line models for studying cohorts of patient samples (Najgebauer et al. 2020; Warren et al. 2021).

It is important that we validate our findings in as many appropriate models as possible and patient-derived xenografts (PDXs) offer great opportunities for extensive testing. PDXs transplanted into mice offer the advantages of *in vivo* testing of tumour, larger material output for testing (e.g., genomic and epigenomic assays) and closer resemblance to tumours in patients (Wang et al. 2017; Richter-Pechańska et al. 2018; Rokita et al. 2019).

Due to the limitations given by cell line models (i.e., media growth, no tumour microenvironment, adaptation to *in vitro* survival), an approach where patient samples are directly tested for drug sensitivity may be attractive (Mirabelli et al. 2019). This will be further discussed in the next section.

4.6 Potential avenues to uncover sensitivities to inhibition

CRISPR screens are an unbiased way to find sensitivities to inhibition by gRNA in cancer cells (see subsection 1.2.4). We have exploited the DepMap, a large resource of CRISPR screens for our Chapter 2 analysis. However, there are other potential approaches that can be taken for systematic screening of cancer sensitivities. The BEAT-AML studies (Tyner et al. 2018; Bottomly et al. 2022) tested *ex vivo* ~400 drugs on samples from patients with AML. One of their main findings was that leukaemia differentiation state plays an important role in response to drug sensitivity. Additionally, there are numerous other studies which perform *ex vivo* drug screening on ALL samples to screen for sensitivities (Frismantas et al. 2017; Lee et al. 2023).

Whilst we studied the transcriptional and epigenetic consequences of *EZH2* depletion, we have not made any direct enquiry about the changes in genes that may be essential for cell survival or that may promote cell proliferation in an *EZH2*-depleted background. Whilst we have found *CDK6* as a differential vulnerability, there may be many more to be revealed. This could be systematically assessed by performing CRISPR screens in AML cell lines with and without depleted *EZH2*. This is currently being performed in the Bond group and will help link the transcriptomic changes to gained vulnerabilities. These investigations could provide insights into the relationships between PRC2 and other epigenetic regulators, similar to discoveries on the cooperativity of NSD1 and SWI/SNF or ARID1A and PU.1 (see subsection 1.3.6). The results from these assays could also be used to test whether GRNs inferred in our engineered cell lines can predict which genes become essential for survival in our *EZH2* loss models. A further test would be to see if incorporating the extensive epigenomic characterisation information from these cell lines would lead to reconstructed GRNs that are better able to predict essential genes.

5. References

4DN, CUT&RUN processing pipeline. Available at: https://data.4dnucleome.org/resources/data-analysis/cut-and-run-pipeline [Accessed August 24, 2024].

Agrawal-Singh, S. et al., 2023. HOXA9 forms a repressive complex with nuclear matrix-associated protein SAFB to maintain acute myeloid leukemia. *Blood*, 141(14), pp.1737–1754.

Aibar, S. et al., 2017. SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, 14(11), pp.1083–1086.

Alaggio, R. et al., 2022. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia*, 36(7), pp.1720–1748.

Alessandrini, F. et al., 2018. ETV7-Mediated DNAJC15 Repression Leads to Doxorubicin Resistance in Breast Cancer Cells. *Neoplasia*, 20(8), pp.857–870.

Alvarez, M.J. et al., 2018. A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nature genetics*, 50(7), pp.979–989.

Alvarez, M.J. et al., 2016. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature genetics*, 48(8), pp.838–847.

Alver, B.H. et al., 2017. The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. *Nature communications*, 8, p.14648.

Alves, C.L. et al., 2021. Co-targeting CDK4/6 and AKT with endocrine therapy prevents progression in CDK4/6 inhibitor and endocrine therapy-resistant breast cancer. *Nature communications*, 12(1), p.5112.

Amemiya, H.M., Kundaje, A. & Boyle, A.P., 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific reports*, 9(1), pp.1–5.

Andrews, S. et al., 2012. FastQC. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed December 3, 2024].

Angeloni, A. & Bogdanovic, O., 2019. Enhancer DNA methylation: implications for gene regulation. *Essays in biochemistry*, 63(6), pp.707–715.

Ariës, I.M. et al., 2018. PRC2 loss induces chemoresistance by repressing apoptosis in T cell acute lymphoblastic leukemia. *The Journal of experimental medicine*, 215(12), pp.3094–3114.

Assi, S.A. et al., 2019. Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nature genetics*, 51(1), pp.151–162.

Badia-i-Mompel, P. et al., 2022. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1), pp.1–3.

Bainbridge, M.N. et al., 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC genomics*, 7(1), p.246.

Balzeau, J. et al., 2017. The LIN28/let-7 Pathway in Cancer. *Frontiers in genetics*, 8, p.31.

Banerji, J., Rusconi, S. & Schaffner, W., 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1), pp.299–308.

Barabási, A.-L., Gulbahce, N. & Loscalzo, J., 2010. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1), pp.56–68.

Barabási, A.-L. & Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2), pp.101–113.

Barisic, D. et al., 2024. ARID1A orchestrates SWI/SNF-mediated sequential binding of transcription factors with ARID1A loss driving pre-memory B cell fate and lymphomagenesis. *Cancer cell*. Available at: http://dx.doi.org/10.1016/j.ccell.2024.02.010.

Barrangou, R. & Doudna, J.A., 2016. Applications of CRISPR technologies in research and beyond. *Nature biotechnology*, 34(9), pp.933–941.

Barshad, G. et al., 2023. RNA polymerase II dynamics shape enhancer-promoter interactions. *Nature genetics*, 55(8), pp.1370–1380.

Basheer, F. et al., 2019. Contrasting requirements during disease evolution identify EZH2 as a therapeutic target in AML. *The Journal of experimental medicine*, 216(4), pp.966–981.

Belhocine, K., DeMare, L. & Habern, O., 2021. Single-Cell Multiomics: Simultaneous Epigenetic and Transcriptional Profiling: 10x Genomics shares experimental planning and sample preparation tips for the Chromium Single Cell Multiome ATAC + Gene Expression system. *Genetic engineering & biotechnology news*, 41(1), pp.66–68.

Belhocine, M. et al., 2022. Dynamics of broad H3K4me3 domains uncover an epigenetic switch between cell identity and cancer-related genes. *Genome research*, 32(7), pp.1328–1342.

Belluschi, S. et al., 2018. Myelo-lymphoid lineage restriction occurs in the human haematopoietic stem cell compartment before lymphoid-primed

multipotent progenitors. *Nature communications*, 9(1), pp.1–15.

Benaglia, T. et al., 2009. mixtools: AnRPackage for Analyzing Finite Mixture Models. *Journal of statistical software*, 32(6), pp.1–29.

Berg, J.L. et al., 2021. EZH2 inactivation in RAS-driven myeloid neoplasms hyperactivates RAS-signaling and increases MEK inhibitor sensitivity. *Leukemia*, 35(5), pp.1521–1526.

Bernt, K.M. et al., 2011. MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. *Cancer cell*, 20(1), pp.66–78.

Bhagwat, A.S. & Vakoc, C.R., 2015. Targeting Transcription Factors in Cancer. *Trends in cancer research*, 1(1), pp.53–65.

Blackledge, N.P. & Klose, R.J., 2021. The molecular principles of gene regulation by Polycomb repressive complexes. *Nature reviews. Molecular cell biology*, 22(12), pp.815–833.

Bock, C. et al., 2022. High-content CRISPR screening. *Nature reviews. Methods primers*, 2(1).

Boehm, J.S. & Golub, T.R., 2015. An ecosystem of cancer cell line factories to support a cancer dependency map. *Nature reviews. Genetics*, 16(7), pp.373–374.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15), pp.2114–2120.

Bolouri, H. et al., 2018. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nature medicine*, 24(1), pp.103–112.

Bond, J. et al., 2020. A transcriptomic continuum of differentiation arrest identifies myeloid interface acute leukemias with poor prognosis. *Leukemia*, 35(3), pp.724–736.

Bond, J. et al., 2018. Polycomb repressive complex 2 haploinsufficiency identifies a high-risk subgroup of pediatric acute myeloid leukemia. *Leukemia*, 32(8), pp.1878–1882.

Bonev, B. et al., 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3), pp.557–572.e24.

Bonev, B. & Cavalli, G., 2016. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11), pp.661–678.

Bottomly, D. et al., 2022. Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer cell*, 40(8), pp.850–864.e9.

Boyle, S. et al., 2020. A central role for canonical PRC1 in shaping the 3D

nuclear landscape. *Genes and Development*, 34(13-14).

Bradner, J.E., Hnisz, D. & Young, R.A., 2017. Transcriptional Addiction in Cancer. *Cell*, 168(4), pp.629–643.

Bravo González-Blas, C. et al., 2023. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature methods*, 20(9), pp.1355–1367.

Bray, N.L. et al., 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), pp.525–527.

Britten, R.J. & Davidson, E.H., 1969. Gene regulation for higher cells: A theory. *Science*, 165(3891), pp.349–357.

Bsteh, D. et al., 2023. Loss of cohesin regulator PDS5A reveals repressive role of Polycomb loops. *Nature communications*, 14(1), pp.1–16.

Buccitelli, C. & Selbach, M., 2020. mRNAs, proteins and the emerging principles of gene expression control. *Nature reviews. Genetics*, 21(10), pp.630–644.

Burd, A. et al., 2020. Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial. *Nature medicine*, 26(12), pp.1852–1858.

Bushweller, J.H., 2019. Targeting transcription factors in cancer — from undruggable to reality. *Nature reviews. Cancer*, 19(11), pp.611–624.

Cai, Y. et al., 2021. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nature communications*, 12(1), pp.1–22.

Campbell, J. et al., 2016. Large-Scale Profiling of Kinase Dependencies in Cancer Cell Lines. *Cell reports*, 14(10), pp.2490–2501.

Cancer Genome Atlas Research Network, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216), pp.1061–1068.

Cano-Rodriguez, D. et al., 2016. Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner. *Nature communications*, 7, p.12284.

Caslini, C. et al., 2007. Interaction of MLL amino terminal sequences with menin is required for transformation. *Cancer research*, 67(15), pp.7275–7283.

Caye, A. et al., 2015. Juvenile myelomonocytic leukemia displays mutations in components of the RAS pathway and the PRC2 network. *Nature genetics*, 47(11), pp.1334–1340.

Cerami, E. et al., 2012. The cBio cancer genomics portal: an open platform for

exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5), pp.401–404.

Chasman, D., Fotuhi Siahpirani, A. & Roy, S., 2016. Network-based approaches for analysis of complex biological systems. *Current opinion in biotechnology*, 39, pp.157–166.

Chen, D. et al., 2021. Nonlinear relationship between chromatin accessibility and estradiol-regulated gene expression. *Oncogene*, 40(7), pp.1332–1346.

Chen, H. & Boutros, P.C., 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC bioinformatics*, 12(1), p.35.

Chen, X. et al., 2022. Single-cell RNA sequencing reveals intra-tumoral heterogeneity of glioblastoma and a pro-tumor subset of tumor-associated macrophages characterized by EZH2 overexpression. *Biochimica et biophysica acta. Molecular basis of disease*, 1868(12), p.166534.

Chou, W.-C. et al., 2011. TET2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics. *Blood*, 118(14), pp.3803–3810.

Christou-Kent, M. et al., 2023. CEBPA phase separation links transcriptional activity and 3D chromatin hubs. *Cell reports*, 42(8), p.112897.

Cierpicki, T. & Grembecka, J., 2014. Challenges and Opportunities in Targeting the Menin–MLL Interaction. *Future medicinal chemistry*, 6(4), pp.447–462.

Ciofani, M. & Zúñiga-Pflücker, J.C., 2007. The thymus as an inductive site for T lymphopoiesis. *Annual review of cell and developmental biology*, 23, pp.463–493.

Cohen, M.H. et al., 2002. Approval summary for imatinib mesylate capsules in the treatment of chronic myelogenous leukemia. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 8(5), pp.935–942.

Corces, M.R. et al., 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10), pp.1193–1203.

Cramer, P., 2019. Organization and regulation of gene transcription. *Nature*, 573(7772), pp.45–54.

Cui, K. et al., 2009. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell stem cell*, 4(1), pp.80–93.

Davies, H. et al., 2002. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892), pp.949–954.

De Bruijn, I. et al., 2023. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer research*, 83(23), pp.3861–3867.

De Kegel, B. & Ryan, C.J., 2019. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. *PLoS genetics*, 15(10), pp.1–25.

De Rooij, J.D.E., Zwaan, C.M. & van den Heuvel-Eibrink, M., 2015. Pediatric AML: From Biology to Clinical Management. *Journal of clinical medicine research*, 4(1), pp.127–149.

De Wit, E. et al., 2015. CTCF Binding Polarity Determines Chromatin Looping. *Molecular cell*, 60(4), pp.676–684.

Deaton, A.M. & Bird, A., 2011. CpG islands and the regulation of transcription. *Genes & development*, 25(10), pp.1010–1022.

Dekker, J. et al., 2017. The 4D nucleome project. *Nature*, 549(7671), pp.219–226.

Delaunay, S., Helm, M. & Frye, M., 2023. RNA modifications in physiology and disease: towards clinical applications. *Nature reviews. Genetics*, 25(2), pp.104–122.

Dempster, J.M. et al., 2021. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome biology*, 22(1), pp.1–23.

Dempster, J.M. et al., 2019. Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv*. Available at: http://dx.doi.org/10.1101/720243.

Dey, K.K. et al., 2022. SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell genomics*, 2(7).

Dharia, N.V. et al., 2021. A first-generation pediatric cancer dependency map. *Nature genetics*, 53(4), pp.529–538.

Dillman, R.O., 1999. Perceptions of Herceptin: a monoclonal antibody for the treatment of breast cancer. *Cancer biotherapy & radiopharmaceuticals*, 14(1), pp.5–10.

Ding, Y.Y. et al., 2021. Network analysis reveals synergistic genetic dependencies for rational combination therapy in Philadelphia chromosome-like acute lymphoblastic leukemia. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 27(18), pp.5109–5122.

Dixon, J.R. et al., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp.376–380.

Dodou, E., Xu, S.-M. & Black, B.L., 2003. mef2c is activated directly by myogenic basic helix-loop-helix proteins during skeletal muscle development in vivo. *Mechanisms of development*, 120(9), pp.1021–1032.

Dorighi, K.M. et al., 2017. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular cell*, 66(4), pp.568–576.e4.

Doyle, E.J., Morey, L. & Conway, E., 2022. Know when to fold 'em: Polycomb complexes in oncogenic 3D genome regulation. *Frontiers in cell and developmental biology*, 10, p.986319.

Drosos, Y. et al., 2022. NSD1 mediates antagonism between SWI/SNF and polycomb complexes and is required for transcriptional activation upon EZH2 inhibition. *Molecular cell*, 82(13), pp.2472–2489.e8.

Du, X. et al., 2018. Hippo/Mst signalling couples metabolic state and immune function of CD8α+ dendritic cells. *Nature*, 558(7708), pp.141–145.

Du, Z. et al., 2020. Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Molecular cell*, 77(4), pp.825–839.e7.

Eagen, K.P., Aiden, E.L. & Kornberg, R.D., 2017. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences of the United States of America*, 114(33), pp.8764–8769.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Enseqlopedia, https://enseqlopedia.com/. *Enseqlopedia*. Available at: http://enseqlopedia.com/ [Accessed July 17, 2024].

Erba, H.P. et al., 2022. Update on a Phase 1/2 First-in-Human Study of the Menin-KMT2A (MLL) Inhibitor Ziftomenib (KO-539) in Patients with Relapsed or Refractory Acute Myeloid Leukemia. *Blood*, 140(Supplement 1), pp.153–156.

Ernst, T. et al., 2010. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature genetics*, 42(8), pp.722–726.

Essaghir, A. et al., 2010. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic acids research*, 38(11).

Eustermann, S. et al., 2011. Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. *Nature structural & molecular biology*, 18(7), pp.777–782.

Ewels, P. et al., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, 32(19),

pp.3047–3048.

Falco, M.M. et al., 2016. The pan-cancer pathological regulatory landscape. *Scientific reports*, 6(July), pp.1–13.

Fang, C. et al., 2020. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *Genome biology*, 21(1), p.247.

Fang, L. et al., 2021. GRNdb: Decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic acids research*, 49(D1), pp.D97–D103.

Ferrando, A.A. et al., 2002. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer cell*, 1(1), pp.75–87.

Finogenova, K. et al., 2020. Structural basis for PRC2 decoding of active histone methylation marks H3K36me2/3. *eLife*, 9, p.e61964.

Frismantas, V. et al., 2017. Ex vivo drug response profiling detects recurrent sensitivity patterns in drug-resistant acute lymphoblastic leukemia. *Blood*, 129(11), pp.e26–e37.

Fröhling, S. et al., 2016. CDK4/6 inhibitor palbociclib for treatment of KMT2A-rearranged acute myeloid leukemia: Interim analysis of the AMLSG 23-14 trial. *Blood*, 128(22), pp.1608–1608.

Fudenberg, G. et al., 2017. Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harbor symposia on quantitative biology*, 82, pp.45–55.

Fudenberg, G. et al., 2016. Formation of Chromosomal Domains by Loop Extrusion. *Cell reports*, 15(9), pp.2038–2049.

Fursova, N.A. et al., 2019. Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression. *Molecular cell*, 74(5), pp.1020–1036.e8.

Gao, J. et al., 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269), p.l1.

Garcia-Alonso, L. et al., 2019. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, 29(8), pp.1363–1375.

Garcia-Alonso, L. et al., 2018. Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer research*, 78(3), pp.769–780.

Garnett, M.J. et al., 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391), pp.570–575.

George, B.S. et al., 2022. Mixed-Phenotype Acute Leukemia: Clinical Diagnosis

and Therapeutic Strategies. *Biomedicines*, 10(8). Available at: http://dx.doi.org/10.3390/biomedicines10081974.

Ghandi, M. et al., 2019. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757), pp.503–508.

Ghavi-Helm, Y. et al., 2019. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature genetics*, 51(8), pp.1272–1282.

Glancy, E. et al., 2023. PRC2.1- and PRC2.2-specific accessory proteins drive recruitment of different forms of canonical PRC1. *Molecular cell*, 83(9), pp.1393–1411.e7.

Goardon, N. et al., 2011. Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer cell*, 19(1), pp.138–152.

Gocho, Y. et al., 2021. Network-based systems pharmacology reveals heterogeneity in LCK and BCL2 signaling and therapeutic sensitivity of T-cell acute lymphoblastic leukemia. *Nature Cancer*, 2(3), pp.284–299.

Goel, S., Bergholz, J.S. & Zhao, J.J., 2022. Targeting CDK4 and CDK6 in cancer. *Nature reviews. Cancer*, 22(6), pp.356–372.

Göllner, S. et al., 2017. Loss of the histone methyltransferase EZH2 induces resistance to multiple drugs in acute myeloid leukemia. *Nature medicine*, 23(1), pp.69–78.

Gonçalves, E. et al., 2022. Pan-cancer proteomic map of 949 human cell lines. *Cancer cell*, 40(8), pp.835–849.e8.

González-Lamuño, D. et al., 2002. Expression and regulation of the transcriptional repressor ZNF43 in Ewing sarcoma cells. *Pediatric pathology & molecular medicine*, 21(6), pp.531–540.

Goode, D.K. et al., 2016. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Developmental cell*, 36(5), pp.572–587.

Gough, S.M., Slape, C.I. & Aplan, P.D., 2011. NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. *Blood*, 118(24), pp.6247–6257.

Granja, J.M. et al., 2019. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature biotechnology*, 37(12), pp.1458–1465.

Grau, D. et al., 2021. Structures of monomeric and dimeric PRC2:EZH1 reveal flexible modules involved in chromatin compaction. *Nature communications*, 12(1), pp.1–12.

Greaves, M.F., 1997. Aetiology of acute leukaemia. *Lancet*, 349(9048),

pp.344–349.

Gröschel, S. et al., 2014. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*, 157(2), pp.369–381.

Gutierrez, A. & Kentsis, A., 2018. Acute myeloid/T-lymphoblastic leukaemia (AMTL): a distinct category of acute leukaemias with common pathogenesis in need of improved therapy. *British journal of haematology*, 180(6), pp.919–924.

Han, L. et al., 2019. Chromatin remodeling mediated by ARID1A is indispensable for normal hematopoiesis in mice. *Leukemia*, 33(9), pp.2291–2305.

Hanselmann, S. et al., 2023. Expression of the cytokinesis regulator PRC1 results in p53-pathway activation in A549 cells but does not directly regulate gene expression in the nucleus. *Cell cycle (Georgetown, Tex.)*, 22(4), pp.419–432.

Hansen, K.H. et al., 2008. A model for transmission of the H3K27me3 epigenetic mark. *Nature cell biology*, 10(11), pp.1291–1300.

Hart, T. et al., 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6), pp.1515–1526.

Hart, T. et al., 2014. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular systems biology*, 10(7), p.733.

Healy, E. et al., 2019. PRC2.1 and PRC2.2 Synergize to Coordinate H3K27 Trimethylation. *Molecular cell*, 76(3), pp.437–452.e6.

He, D. et al., 2011. Small Rab GTPase Rab7b promotes megakaryocytic differentiation by enhancing IL-6 production and STAT3-GATA-1 association. *Journal of molecular medicine*, 89(2), pp.137–150.

Heinz, S. et al., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4), pp.576–589.

Helm, M. & Motorin, Y., 2017. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature reviews. Genetics*, 18(5), pp.275–291.

Henley, M.J. & Koehler, A.N., 2021. Advances in targeting “undruggable” transcription factors with small molecules. *Nature reviews. Drug discovery*, 20(9), pp.669–688.

Heyer, E.E. et al., 2019. Diagnosis of fusion genes using targeted RNA sequencing. *Nature communications*, 10(1), p.1388.

Hnisz, D. et al., 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280), pp.1454–1458.

Hölzenspies, J.J. et al., 2024. PRC2 Promotes Canalisation During Endodermal Differentiation. *bioRxiv*, p.2024.04.23.590736.

Huang, X. et al., 2024. Single-cell systems pharmacology identifies development-driven drug response and combination therapy in B cell acute lymphoblastic leukemia. *Cancer cell*, 42(4), pp.552–567.e6.

Huber, W. et al., 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2), pp.115–121.

Hu, G. et al., 2018. Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. *Immunity*, 48(2), pp.227–242.e8.

Hughes, A.L., Kelley, J.R. & Klose, R.J., 2020. Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(8), p.194567.

Hughes, T.R. et al., 2000. Functional discovery via a compendium of expression profiles. *Cell*, 102(1), pp.109–126.

Huisinga, K.L., Brower-Toland, B. & Elgin, S.C.R., 2006. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*, 115(2), pp.110–122.

Husmeier, D., 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* , 19(17), pp.2271–2282.

Huynh-Thu, V.A. et al., 2010. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9), pp.1–10.

Ideker, T. & Krogan, N.J., 2012. Differential network biology. *Molecular systems biology*, 8, p.565.

Iglesias-Martinez, L.F., De Kegel, B. & Kolch, W., 2021. KBoost: a new method to infer gene regulatory networks from gene expression data. *Scientific reports*, 11(1), pp.1–13.

Illingworth, R.S. et al., 2016. Polycomb enables primitive endoderm lineage priming in embryonic stem cells. *eLife*, 5, p.e14926.

Iorio, F. et al., 2016. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3), pp.740–754.

Isoda, T. et al., 2017. Non-coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. *Cell*, 171(1), pp.103–119.e18.

Issa, G.C. et al., 2022. The Menin Inhibitor SNDX-5613 (revumenib) Leads to Durable Responses in Patients (Pts) with KMT2A-Rearranged or NPM1 Mutant AML: Updated Results of a Phase (Ph) 1 Study. *Blood*, 140, pp.150–152.

Iwafuchi-Doi, M. et al., 2016. The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Molecular cell*, 62(1), pp.79–91.

Izzo, F. et al., 2020. DNA methylation disruption reshapes the hematopoietic differentiation landscape. *Nature genetics*, 52(4), pp.378–387.

Jacob, F. & Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3, pp.318–356.

Jacob, F., Ullman, A. & Monod, J., 1964. THE PROMOTOR, A GENETIC ELEMENT NECESSARY TO THE EXPRESSION OF AN OPERON. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, 258, pp.3125–3128.

Jacobsen, J.A. et al., 2020. Ezh2 represses transcription of innate lymphoid genes in B lymphocyte progenitors and maintains the B-2 cell fate. *The journal of immunology*, 204(7), pp.1760–1769.

Jeong, M. et al., 2014. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature genetics*, 46(1), pp.17–23.

Jiang, R., 2015. Walking on multiple disease-gene networks to prioritize candidate genes. *Journal of molecular cell biology*, 7(3), pp.214–230.

Jinek, M. et al., 2012. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096), pp.816–821.

Johanson, T.M. et al., 2018. Transcription-factor-mediated supervision of global genome architecture maintains B cell identity. *Nature immunology*, 19(11), pp.1257–1264.

John, S.A. et al., 2008. Ets-1 regulates plasma cell differentiation by interfering with the activity of the transcription factor Blimp-1. *The journal of biological chemistry*, 283(2), pp.951–962.

Johnson, C.D. et al., 2007. The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer research*, 67(16), pp.7713–7722.

Jones, L., McCarthy, P. & Bond, J., 2020. Epigenetics of paediatric acute myeloid leukaemia. *British journal of haematology*, 188(1), pp.63–76.

Kadia, T.M. et al., 2018. Phase I study of palbociclib alone and in combination in patients with relapsed and refractory (R/R) leukemias. *Blood*, 132(Supplement 1), pp.4057–4057.

Kadoch, C. et al., 2016. Dynamics of BAF–Polycomb complex opposition on

heterochromatin in normal and oncogenic states. *Nature genetics*, 49(2), pp.213–222.

Kamal, A. et al., 2023. GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Molecular systems biology*, 19(6), p.e11627.

Kamimoto, K. et al., 2023. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949), pp.742–751.

Kamminga, L.M. et al., 2006. The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood*, 107(5), pp.2170–2179.

Karamitros, D. et al., 2017. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nature immunology*, 19(1), pp.85–97.

Kassis, J.A., Kennison, J.A. & Tamkun, J.W., 2017. Polycomb and Trithorax Group Genes in Drosophila. *Genetics*, 206(4), pp.1699–1725.

Kempf, J.M. et al., 2021. Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML. *Scientific reports*, 11(1), pp.1–13.

Kernfeld, E. et al., 2024. Transcriptome data are insufficient to control false discoveries in regulatory network inference. *Cell systems*, 15(8), pp.709–724.e13.

Kerpedjiev, P. et al., 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome biology*, 19(1), p.125.

Khabirova, E. et al., 2022. Single-cell transcriptomics reveals a distinct developmental state of KMT2A-rearranged infant B-cell acute lymphoblastic leukemia. *Nature medicine*, 28(4), pp.743–751.

Khoury, J.D. et al., 2022. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*, 36(7), pp.1703–1719.

Kiani, K. et al., 2022. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. *Molecular systems biology*, 18(9).

Kim, A. & Cohen, M.S., 2016. The discovery of vemurafenib for the treatment of BRAF-mutated metastatic melanoma. *Expert opinion on drug discovery*, 11(9), pp.907–916.

Kim, J. et al., 2013. The n-SET domain of Set1 regulates H2B ubiquitylation-dependent H3K4 methylation. *Molecular cell*, 49(6), pp.1121–1133.

King, H.W. et al., 2018. Polycomb repressive complex 1 shapes the nucleosome landscape but not accessibility at target genes. *Genome*

research, 28(10), pp.1494–1507.

Kizer, K.O. et al., 2005. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Molecular and cellular biology*, 25(8), pp.3305–3316.

Kloetgen, A. et al., 2019. 3D Chromosomal Landscapes in Hematopoiesis and Immunity. *Trends in immunology*, 40(9), pp.809–824.

Kloetgen, A. et al., 2020. Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nature genetics*, 52(4), pp.388–400.

Knutson, S.K. et al., 2013. Durable tumor regression in genetically altered malignant rhabdoid tumors by inhibition of methyltransferase EZH2. *Proceedings of the National Academy of Sciences of the United States of America*, 110(19), pp.7922–7927.

Kon, A. et al., 2013. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nature genetics*, 45(10), pp.1232–1237.

Kornberg, R.D., 1974. Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139), pp.868–871.

Koutrouli, M. et al., 2020. A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in bioengineering and biotechnology*, 8, p.34.

Kraft, K. et al., 2022. Polycomb-mediated genome architecture enables long-range spreading of H3K27 methylation. *Proceedings of the National Academy of Sciences*, 119(22), p.e2201883119.

Krill-Burger, J.M. et al., 2023. Partial gene suppression improves identification of cancer vulnerabilities when CRISPR-Cas9 knockout is pan-lethal. *Genome biology*, 24(1), pp.1–26.

Kruse, K., Hug, C.B. & Vaquerizas, J.M., 2020. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome biology*, 21(1), p.303.

Lachmann, A. et al., 2016. ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* , 32(14), pp.2233–2235.

Larrosa-Garcia, M. & Baer, M.R., 2017. FLT3 Inhibitors in Acute Myeloid Leukemia: Current Status and Future Directions. *Molecular cancer therapeutics*, 16(6), pp.991–1001.

Laurenti, E. & Göttgens, B., 2018. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689), pp.418–426.

Lee, C.-H. et al., 2018. Allosteric Activation Dictates PRC2 Activity Independent of Its Recruitment to Chromatin. *Molecular cell*, 70(3), pp.422–434.e6.

Lee, J.-S., Smith, E. & Shilatifard, A., 2010. The language of histone crosstalk. *Cell*, 142(5), pp.682–685.

Lee, S.H. et al., 2022. The role of EZH1 and EZH2 in development and cancer. *BMB reports*, 55(12), pp.595–601.

Lee, S.H.R. et al., 2023. Pharmacotypes across the genomic landscape of pediatric acute lymphoblastic leukemia and impact on treatment response. *Nature medicine*, 29(1), pp.170–179.

Lee, T.I. & Young, R.A., 2013. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6), pp.1237–1251.

Lefebvre, C., Rieckhof, G. & Califano, A., 2012. Reverse-engineering human regulatory networks. *Wiley interdisciplinary reviews. Systems biology and medicine*, 4(4), pp.311–325.

Lefeivre, T. et al., 2022. Immature acute leukaemias: lessons from the haematopoietic roadmap. *The FEBS journal*, 289(15), pp.4355–4370.

Lenstra, T.L. et al., 2016. Transcription Dynamics in Living Cells. *Annual review of biophysics*, 45, pp.25–47.

Ley, T.J. et al., 2010. DNMT3A mutations in acute myeloid leukemia. *The New England journal of medicine*, 363(25), pp.2424–2433.

Liang, D.-C. et al., 2013. Cooperating gene mutations in childhood acute myeloid leukemia with special reference on mutations of ASXL1, TET2, IDH1, IDH2, and DNMT3A. *Blood*, 121(15), pp.2988–2995.

Liberzon, A. et al., 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* , 27(12), pp.1739–1740.

Lieberman-Aiden, E. et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950), pp.289–293.

Lifton, R.P. et al., 1978. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harbor symposia on quantitative biology*, 42 Pt 2, pp.1047–1051.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–2079.

Liu, X. & Gong, Y., 2019. Isocitrate dehydrogenase inhibitors in acute myeloid leukemia. *Biomarker research*, 7, p.22.

Liu, Y. et al., 2017. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nature genetics*, 49(8), pp.1211–1218.

Li, Y. et al., 2023. Proteogenomic data and resources for pan-cancer analysis.

Cancer cell, 41(8), pp.1397–1406.

Li, Y. et al., 2024. Research advances of polycomb group proteins in regulating mammalian development. *Frontiers in cell and developmental biology*, 12, p.1383200.

Li, Z. et al., 2017. ASXL1 interacts with the cohesin complex to maintain chromatid separation and gene expression for normal hematopoiesis. *Science advances*, 3(1), p.e1601602.

Lobanenkov, V.V. et al., 1990. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, 5(12), pp.1743–1753.

Lopez-Delisle, L. et al., 2020. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* , 37(3), pp.422–423.

Losman, J.-A. et al., 2013. (R)-2-hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. *Science*, 339(6127), pp.1621–1625.

Loubiere, V., Martinez, A.-M. & Cavalli, G., 2019. Cell fate and developmental regulation dynamics by Polycomb proteins and 3D genome architecture. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 41(3), p.e1800222.

Lowry, J.A. & Atchley, W.R., 2000. Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. *Journal of molecular evolution*, 50(2), pp.103–115.

Luc, S. et al., 2012. The earliest thymic T cell progenitors sustain B cell and myeloid lineage potential. *Nature immunology*, 13(4), pp.412–419.

Luger, K. et al., 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648), pp.251–260.

Lukauskas, S. et al., 2024. Decoding chromatin states by proteomic profiling of nucleosome readers. *Nature*, pp.1–9.

Luo, Y. et al., 2020. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research*, 48(D1), pp.D882–D889.

Mansour, M.R. et al., 2014. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, 346(6215), pp.1373–1377.

Marbach, D. et al., 2010. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14), pp.6286–6291.

Marbach, D. et al., 2016. Wisdom of crowds for robust gene network inference

the DREAM5 Consortium HHS Public Access. *Nature methods*, 9(8), pp.796–804.

Margolin, A.A. et al., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1(Suppl 1), p.S7.

Margueron, R. et al., 2008. Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Molecular cell*, 32(4), pp.503–518.

Margueron, R. et al., 2009. Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature*, 461(7265), pp.762–767.

Matharu, N. & Ahituv, N., 2015. Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS genetics*, 11(12), p.e1005640.

Mazumdar, C. et al., 2015. Leukemia-Associated Cohesin Mutants Dominantly Enforce Stem Cell Programs and Impair Human Hematopoietic Progenitor Differentiation. *Cell stem cell*, 17(6), pp.675–688.

McCabe, A. et al., 2023. Investigating the suitability of in vitro cell lines as models for the major subtypes of epithelial ovarian cancer. *Frontiers in Cell and Developmental Biology*, 11.

McGraw, K.O. & Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1), p.30.

McLean, C.Y. et al., 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5), pp.495–501.

Meers, M.P., Tenenbaum, D. & Henikoff, S., 2019. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics & chromatin*, 12(1), p.42.

Mertins, P. et al., 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605), pp.55–62.

Metzner, E., Southard, K.M. & Norman, T.M., 2024. Multiome Perturb-seq unlocks scalable discovery of integrated perturbation effects on the transcriptome and epigenome. *bioRxiv.org: the preprint server for biology*, p.2024.07.26.605307.

Meyer, C. et al., 2023. The KMT2A recombinome of acute leukemias in 2023. *Leukemia*, 37(5), pp.988–1005.

Meyer, C. et al., 2018. The MLL recombinome of acute leukemias in 2017. *Leukemia*, 32(2), pp.273–284.

Meyers, R.M. et al., 2017. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells.

Nature genetics, 49(12), pp.1779–1784.

Mikulasova, A. et al., 2022. Epigenomic translocation of H3K4me3 broad domains over oncogenes following hijacking of super-enhancers. *Genome research*, 32(7), pp.1343–1354.

Millán-Zambrano, G. et al., 2022. Histone post-translational modifications — cause and consequence of genome function. *Nature reviews. Genetics*, 23(9), pp.563–580.

Mirabelli, P., Coppola, L. & Salvatore, M., 2019. Cancer cell lines are useful model systems for medical research. *Cancers*, 11(8), p.1098.

Mitra, S. et al., 2024. Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nature genetics*, 56(4), pp.627–636.

Mochizuki-Kashio, M. et al., 2011. Dependency on the polycomb gene Ezh2 distinguishes fetal from adult hematopoietic stem cells. *Blood*, 118(25), pp.6553–6561.

Mochizuki-Kashio, M. et al., 2015. Ezh2 loss in hematopoietic stem cells predisposes mice to develop heterogeneous malignancies in an Ezh1-dependent manner. *Blood*, 126(10), pp.1172–1183.

Moore, L.D., Le, T. & Fan, G., 2013. DNA methylation and its basic function. *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, 38(1), pp.23–38.

Moreau, P. et al., 1981. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic acids research*, 9(22), pp.6047–6068.

Moussa, H.F. et al., 2019. Canonical PRC1 controls sequence-independent propagation of Polycomb-mediated gene silencing. *Nature communications*, 10(1), pp.1–12.

Mumbach, M.R. et al., 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature genetics*, 49(11), pp.1602–1612.

Najgebauer, H. et al., 2020. CELLector: Genomics-guided selection of cancer in vitro models. *Cell systems*, 10(5), pp.424–432.e6.

Nebenfuehr, S., Kollmann, K. & Sexl, V., 2020. The role of CDK6 in cancer. *International journal of cancer. Journal international du cancer*, 147(11), pp.2988–2995.

Neff, T. et al., 2012. Polycomb repressive complex 2 is required for MLL-AF9 leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13), pp.5028–5033.

Nicolas, D., Phillips, N.E. & Naef, F., 2017. What shapes eukaryotic transcriptional bursting? *Molecular bioSystems*, 13(7), pp.1280–1290.

Nijhawan, D. et al., 2012. Cancer vulnerabilities unveiled by genomic loss. *Cell*, 150(4), pp.842–854.

Nuno, K. et al., 2024. Convergent epigenetic evolution drives relapse in acute myeloid leukemia. *eLife*, 13, p.e93019.

O'Carroll, D. et al., 2001. The polycomb-group gene Ezh2 is required for early mouse development. *Molecular and cellular biology*, 21(13), pp.4330–4336.

O'Connor, D. et al., 2023. The Clinicogenomic Landscape of Induction Failure in Childhood and Young Adult T-Cell Acute Lymphoblastic Leukemia. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 41(19), pp.3545–3556.

Ogiyama, Y. et al., 2018. Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. *Molecular cell*, 71(1), pp.73–88.e5.

Oguro, H. et al., 2010. Poised lineage specification in multipotential hematopoietic stem and progenitor cells by the polycomb protein Bmi1. *Cell stem cell*, 6(3), pp.279–286.

Ohanian, M. et al., 2019. MYC protein expression is an important prognostic factor in acute myeloid leukemia. *Leukemia and Lymphoma*, 60(1), pp.37–48.

Open2C, 2024. Pairtools: From sequencing data to chromosome contacts. *PLoS computational biology*, 20(5), p.e1012164.

Oshima, M. et al., 2016. Ezh2 regulates the Lin28/let-7 pathway to restrict activation of fetal gene signature in adult hematopoietic stem cells. *Experimental hematology*, 44(4), pp.282–96.e3.

Pacini, C. et al., 2021. Integrated cross-study datasets of genetic dependencies in cancer. *Nature communications*, 12(1).

Paolella, B.R. et al., 2017. Copy-number and gene dependency analysis reveals partial copy loss of wild-type SF3B1 as a novel cancer vulnerability. *eLife*, 6, pp.1–35.

Papaemmanuil, E. et al., 2016. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England journal of medicine*, 374(23), pp.2209–2221.

Parreno, V. et al., 2024. Transient loss of Polycomb components induces an epigenetic cancer fate. *Nature*, 629(8012), pp.688–696.

Paull, E.O. et al., 2021. A modular master regulator landscape controls cancer transcriptional identity. *Cell*, pp.1–18.

Peifer, M. & Bender, W., 1986. The anterobithorax and bithorax mutations of the bithorax complex. *The EMBO journal*, 5(9), pp.2293–2303.

Petermann, E., Lan, L. & Zou, L., 2022. Sources, resolution and physiological relevance of R-loops and RNA-DNA hybrids. *Nature reviews. Molecular cell biology*, 23(8), pp.521–540.

Peters, J.-M., Tedeschi, A. & Schmitz, J., 2008. The cohesin complex and its roles in chromosome biology. *Genes & development*, 22(22), pp.3089–3114.

Petrovic, J. et al., 2019. Oncogenic Notch Promotes Long-Range Regulatory Interactions within Hyperconnected 3D Cliques. *Molecular cell*, 73(6), pp.1174–1190.e12.

Picard, https://broadinstitute.github.io/picard/. Available at: https://broadinstitute.github.io/picard/ [Accessed December 3, 2024].

Pino, J.C. et al., 2024. Mapping the proteogenomic landscape enables prediction of drug response in acute myeloid leukemia. *Cell Reports Medicine*, 5(1), p.101359.

Piunti, A. & Shilatifard, A., 2021. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nature reviews. Molecular cell biology*, 22(5), pp.326–345.

Pivetti, S. et al., 2019. Loss of PRC1 activity in different stem cell compartments activates a common transcriptional program with cell type-dependent outcomes. *Science advances*, 5(5), p.eaav1594.

Placke, T. et al., 2014. Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. *Blood*, 124(1), pp.13–23.

Pölönen, P. et al., 2024. The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature*, pp.1–10.

Pribnow, D., 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America*, 72(3), pp.784–788.

Prorok, P. et al., 2023. Loss of Ezh2 function remodels the DNA replication initiation landscape. *Cell reports*, 42(4), p.112280.

Quesada, A.E. et al., 2018. Mixed phenotype acute leukemia contains heterogeneous genetic mutations by next-generation sequencing. *Oncotarget*, 9(9), pp.8441–8449.

Qu, Y. et al., 2014. Differential methylation in CN-AML preferentially targets non-CGI regions and is dictated by DNMT3A mutational status and associated with predominant hypomethylation of HOX genes. *Epigenetics: official journal of the DNA Methylation Society*, 9(8), pp.1108–1119.

Rahman, S. et al., 2017. Activation of the LMO2 oncogene through a somatically acquired neomorphic promoter in T-cell acute lymphoblastic leukemia. *Blood*, 129(24), pp.3221–3226.

Ramírez, F. et al., 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42, pp.W187–91.

Ramsden, D.A. & Nussenzweig, A., 2021. Mechanisms driving chromosomal translocations: lost in time and space. *Oncogene*, 40(25), pp.4263–4270.

Rao, S.S.P. et al., 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), pp.1665–1680.

Ren, Z. et al., 2022. A PRC2-Kdm5b axis sustains tumorigenicity of acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 119(9).

Ribeiro, A.F.T. et al., 2012. Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. *Blood*, 119(24), pp.5824–5831.

Rice, S. et al., 2021. A human fetal liver-derived infant MLL-AF4 acute lymphoblastic leukemia model reveals a distinct fetal gene expression program. *Nature communications*, 12(1), p.6905.

Richter-Pechańska, P. et al., 2018. PDX models recapitulate the genetic and epigenetic landscape of pediatric T-cell leukemia. *EMBO molecular medicine*, 10(12), p.e9443.

Ries, L.A.G. et al., 1999. *Cancer incidence and survival among children and adolescents: United States SEER Program 1975-1995* L. A. G. Ries et al., eds., Bethesda: National Cancer Institute (NCI).

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), p.e47.

Roayaei Ardakany, A. et al., 2020. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome biology*, 21(1), p.256.

Roberts, K.G. et al., 2014. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *The New England journal of medicine*, 371(11), pp.1005–1015.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , 26(1), pp.139–140.

Robinson, P.J.J. & Rhodes, D., 2006. Structure of the “30nm” chromatin fibre: A key role for the linker histone. *Current opinion in structural biology*, 16(3), pp.336–343.

Rokita, J.L. et al., 2019. Genomic profiling of childhood tumor patient-derived xenograft models to enable rational clinical trial design. *Cell reports*, 29(6), pp.1675–1689.e9.

Rothenberg, E.V., 2005. Thymic Regulation--Hidden in Plain Sight. *Science*, 307(5711), pp.858–859.

Rowley, M.J. & Corces, V.G., 2018. Organizational principles of 3D genome architecture. *Nature reviews. Genetics*, 19(12), pp.789–800.

Ruthenburg, A.J. et al., 2011. Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions. *Cell*, 145(5), pp.692–706.

Ryan, R.J.H. et al., 2015. Detection of Enhancer-Associated Rearrangements Reveals Mechanisms of Oncogene Dysregulation in B-cell Lymphoma. *Cancer discovery*, 5(10), pp.1058–1071.

Salgado, H. et al., 2023. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic acids research*, 52(D1), pp.D255–D264.

Salgado, H. et al., 2006. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic acids research*, 34(Database issue), pp.D394–7.

Satijn, D.P. et al., 1997. RING1 is associated with the polycomb group protein complex and acts as a transcriptional repressor. *Molecular and cellular biology*, 17(7), pp.4105–4113.

Sauer, P.V. et al., 2023. Activation of automethylated PRC2 by dimerization on chromatin. *bioRxiv : the preprint server for biology*. Available at: http://dx.doi.org/10.1101/2023.10.12.562141.

Saurin, A.J. et al., 1998. The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. *The Journal of cell biology*, 142(4), pp.887–898.

Saw, J. et al., 2013. The fusion partner specifies the oncogenic potential of NUP98 fusion proteins. *Leukemia research*, 37(12), pp.1668–1673.

Saxonov, S., Berg, P. & Brutlag, D.L., 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5), pp.1412–1417.

Schena, M., 1996. Genome analysis with gene expression microarrays. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 18(5), pp.427–431.

Schep, A.N. et al., 2015. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions.

Genome research, 25(11), pp.1757–1770.

Schneider, C.A., Rasband, W.S. & Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 9(7), pp.671–675.

Schoenfelder, S. et al., 2015. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature genetics*, 47(10), pp.1179–1186.

Schubert, M. et al., 2018. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature communications*, 9(1), pp.1–11.

Schultz, K.R. et al., 2014. Long-term follow-up of imatinib in pediatric Philadelphia chromosome-positive acute lymphoblastic leukemia: Children's Oncology Group study AALL0031. *Leukemia*, 28(7), pp.1467–1471.

Senapati, J. et al., 2023. Philadelphia-Like Genetic Rearrangements in Adults With B-Cell ALL: Refractoriness to Chemotherapy and Response to Tyrosine Kinase Inhibitor in ABL Class Rearrangements. *JCO precision oncology*, 7, p.e2200707.

Shalem, O., Sanjana, N.E. & Zhang, F., 2015. High-throughput functional genomics using CRISPR-Cas9. *Nature reviews. Genetics*, 16(5), pp.299–311.

Shaw, T.I. et al., 2021. Integrative network analysis reveals USP7 haploinsufficiency inhibits E-protein activity in pediatric T-lineage acute lymphoblastic leukemia (T-ALL). *Scientific reports*, 11(1), pp.1–12.

Shen, X. et al., 2008. EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Molecular cell*, 32(4), pp.491–502.

Shukla, N. et al., 2016. Final Report of Phase 1 Study of the DOT1L Inhibitor, Pinometostat (EPZ-5676), in Children with Relapsed or Refractory MLL-r Acute Leukemia. *Blood*, 128(22), pp.2780–2780.

Siggers, T. & Gordân, R., 2013. Protein–DNA binding: complexities and multi-protein codes. *Nucleic acids research*, 42(4), pp.2099–2111.

Simon, M.C., 1995. Gotta have GATA. *Nature genetics*, 11(1), pp.9–11.

Singh, H., Khan, A.A. & Dinner, A.R., 2014. Gene regulatory networks in the immune system. *Trends in immunology*, 35(5), pp.211–218.

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

Smith, I. et al., 2017. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS*

biology, 15(11), pp.1–23.

Sondka, Z. et al., 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature reviews. Cancer*, 18(11), pp.696–705.

Sousa, A. et al., 2023. Pan-Cancer landscape of protein activities identifies drivers of signalling dysregulation and patient survival. *Molecular systems biology*.

Sparbier, C.E. et al., 2023. Targeting Menin disrupts the KMT2A/B and polycomb balance to paradoxically activate bivalent genes. *Nature cell biology*, 25(2), pp.258–272.

Spitz, F. & Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics*, 13(9), pp.613–626.

Stein, E.M. et al., 2017. Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood*, 130(6), pp.722–731.

Stein, E.M. et al., 2018. The DOT1L inhibitor pinometostat reduces H3K79 methylation and has modest clinical activity in adult acute leukemia. *Blood*, 131(24), pp.2661–2669.

Stielow, B. et al., 2018. MGA, L3MBTL2 and E2F6 determine genomic binding of the non-canonical Polycomb repressive complex PRC1.6. *PLoS genetics*, 14(1), p.e1007193.

Stoeckius, M. et al., 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9), pp.865–868.

Strahl, B.D. & Allis, C.D., 2000. The language of covalent histone modifications. *Nature*, 403(6765), pp.41–45.

Su, I.-H. et al., 2002. Ezh2 controls B cell development through histone H3 methylation and Igh rearrangement. *Nature immunology*, 4(2), pp.124–131.

Swart, L.E. & Heidenreich, O., 2021. The RUNX1/RUNX1T1 network: translating insights into therapeutic options. *Experimental hematology*, 94, pp.1–10.

Takayama, N. et al., 2021. The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. *Cell stem cell*, 28(3), pp.488–501.e10.

Tanaka, S. et al., 2012. Ezh2 augments leukemogenicity by reinforcing differentiation blockage in acute myeloid leukemia. *Blood*, 120(5), pp.1107–1117.

Tanasi, I. et al., 2019. Efficacy of tyrosine kinase inhibitors in Ph-like acute lymphoblastic leukemia harboring ABL-class rearrangements. *Blood*, 134(16), pp.1351–1355.

Tarbell, E.D. & Liu, T., 2019. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic acids research*, 47(16), p.e91.

TARGET, 2022. Therapeutically Applicable Research to Generate Effective Treatments (TARGET). Available at: https://www.cancer.gov/ccg/research/genome-sequencing/target [Accessed July 29, 2024].

Tarlock, K. et al., 2018. Distinct co-occurring mutational profiles in acute myeloid leukemia confers prognostic significance in children and young adults with FLT3/ITD mutations. *Blood*, 132, pp.443–443.

Tenen, D.G. et al., 1997. Transcription factors, normal myeloid development, and leukemia. *Blood*, 90(2), pp.489–519.

Thatikonda, V. et al., 2024. Genetic dependencies associated with transcription factor activities in human cancer cell lines. *Cell reports*, 43(5), p.114175.

Totiger, T.M. et al., 2023. Targeted Therapy Development in Acute Myeloid Leukemia. *Biomedicines*, 11(2).

Trescher, S. & Leser, U., 2019. Estimation of Transcription Factor Activity in Knockdown Studies. *Scientific reports*, 9(1), pp.1–11.

Trescher, S., Münchmeyer, J. & Leser, U., 2017. Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. *BMC systems biology*, 11(1), pp.1–18.

Tsherniak, A. et al., 2017. Defining a Cancer Dependency Map. *Cell*, 170(3), pp.564–576.e16.

Tudose, C., Bond, J. & Ryan, C.J., 2023. Gene essentiality in cancer is better predicted by mRNA abundance than by gene regulatory network-inferred activity. *NAR cancer*, 5(4), p.zcad056.

Tuominen, R. et al., 2015. MGMT promoter methylation is associated with temozolomide response and prolonged progression-free survival in disseminated cutaneous melanoma. *International journal of cancer. Journal international du cancer*, 136(12), pp.2844–2853.

Tu, Z. et al., 2023. Discordance between chromatin accessibility and transcriptional activity during the human primed-to-naïve pluripotency transition process. *Cell regeneration (London, England)*, 12(1), p.35.

Tyner, J.W. et al., 2018. Functional genomic landscape of acute myeloid leukaemia. *Nature*, 562(7728), pp.526–531.

Umeda, M. et al., 2024. A new genomic framework to categorize pediatric acute myeloid leukemia. *Nature genetics*, 56(2), pp.281–293.

Upadhya, S.R. & Ryan, C.J., 2022. Experimental reproducibility limits the correlation between mRNA and protein abundances in tumor proteomic

profiles. *Cell reports methods*, 2(9), p.100288.

Vasimuddin, M. et al., 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, pp. 314–324.

Veronezi, G.M.B. & Ramachandran, S., 2024. Nucleation and spreading maintain Polycomb domains every cell cycle. *Cell reports*, 43(4), p.114090.

Vidal, M. & Starowicz, K., 2017. Polycomb complexes PRC1 and their function in hematopoiesis. *Experimental hematology*, 48, pp.12–31.

Vieux-Rochas, M. et al., 2015. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proceedings of the National Academy of Sciences of the United States of America*, 112(15), pp.4672–4677.

Vinceti, A. et al., 2021. CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. *BMC genomics*, 22(1), pp.1–16.

Wang, J. et al., 2022. A cryptic transactivation domain of EZH2 binds AR and AR's splice variant, promoting oncogene activation and tumorous transformation. *Nucleic acids research*, 50(19), pp.10929–10946.

Wang, K. et al., 2009. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature biotechnology*, 27(9), pp.829–837.

Wang, K. et al., 2017. Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. *Leukemia*, 31(1), pp.151–158.

Wang, L. et al., 2023. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nature methods*, 20(9), pp.1368–1378.

Wang, T. et al., 2015. Aberrant regulation of the LIN28A/LIN28B and let-7 loop in human malignant tumors and its effects on the hallmarks of cancer. *Molecular cancer*, 14(1), p.125.

Wang, T. et al., 2015. Identification and characterization of essential genes in the human genome. , 350(6264), pp.1096–1101.

Wang, W. et al., 2019. Combined gene essentiality scoring improves the prediction of cancer dependency maps. *EBioMedicine*, 50, pp.67–80.

Wang, W. et al., 2016. MEF2C protects bone marrow B-lymphoid progenitors during stress haematopoiesis. *Nature communications*, 7(1), pp.1–15.

Wani, A.H. et al., 2016. Chromatin topology is coupled to Polycomb group protein subnuclear organization. *Nature communications*, 7, p.10291.

Ward, P.S. et al., 2010. The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer cell*, 17(3), pp.225–234.

Warren, A. et al., 2021. Global computational alignment of tumor and cell line transcriptional profiles. *Nature communications*, 12(1), p.22.

Wei, A.H. et al., 2020. Venetoclax plus LDAC for newly diagnosed AML ineligible for intensive chemotherapy: a phase 3 randomized placebo-controlled trial. *Blood*, 135(24), pp.2137–2145.

Weidemüller, P. et al., 2021. Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics*, 21(23-24), p.e2000034.

Weischenfeldt, J. et al., 2017. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nature genetics*, 49(1), pp.65–74.

Westermann, L. et al., 2022. Wildtype heterogeneity contributes to clonal variability in genome edited cells. *Scientific reports*, 12(1), p.18211.

Wilson, B.G. et al., 2010. Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer cell*, 18(4), pp.316–328.

Wolf, B.K. et al., 2022. Cooperation of chromatin remodeling SWI/SNF complex and pioneer factor AP-1 shapes 3D enhancer landscapes. *Nature structural & molecular biology*, 30(1), pp.10–21.

Wong, A.K. et al., 2021. Decoding disease: from genomes to networks to phenotypes. *Nature reviews. Genetics*, 22(12), pp.774–790.

Wu, X., Johansen, J.V. & Helin, K., 2013. Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation. *Molecular cell*, 49(6), pp.1134–1146.

Xie, X. et al., 2021. Single-cell transcriptomic landscape of human blood cells. *National science review*, 8(3), p.nwaa180.

Xu, J. et al., 2022. Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature*, 611(7935), pp.387–398.

Yang, C. et al., 2016. Acquired CDK6 amplification promotes breast cancer resistance to CDK4/6 inhibitors and loss of ER signaling and dependence. *Oncogene*, 36(16), pp.2255–2264.

Yang, H. et al., 2010. RG7204 (PLX4032), a selective BRAFV600E inhibitor, displays potent antitumor activity in preclinical melanoma models. *Cancer research*, 70(13), pp.5518–5527.

Yang, W. et al., 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids*

research, 41(Database issue), pp.D955–61.

Yuan, J. et al., 2012. Lin28b reprograms adult bone marrow hematopoietic progenitors to mediate fetal-like lymphopoiesis. *Science*, 335(6073), pp.1195–1200.

Yu, G. et al., 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5), pp.284–287.

Yu, G., Wang, L.-G. & He, Q.-Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics (Oxford, England)*, 31(14), pp.2382–2383.

Zaborowski, A.B. & Walther, D., 2020. Determinants of correlated expression of transcription factors and their target genes. *Nucleic acids research*, 48(20), pp.11347–11369.

Zaret, K.S. & Carroll, J.S., 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25(21), pp.2227–2241.

Zarnegar-Lumley, S. et al., 2023. Characteristics and prognostic impact of IDH mutations in AML: a COG, SWOG, and ECOG analysis. *Blood advances*, 7(19), pp.5941–5953.

Zeng, A.G.X. et al., 2023. Precise single-cell transcriptomic mapping of normal and leukemic cell states reveals unconventional lineage priming in acute myeloid leukemia. *bioRxiv : the preprint server for biology*. Available at: http://dx.doi.org/10.1101/2023.12.26.573390.

Zhang, B. et al., 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518), pp.382–387.

Zhang, H. et al., 2016. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, 166(3), pp.755–765.

Zhang, T. et al., 2020. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome biology*, 21(1), p.45.

Zhang, X. et al., 2024. An immunophenotype-coupled transcriptomic atlas of human hematopoietic progenitors. *Nature immunology*, 25(4), pp.703–715.

Zhang, X. et al., 2020. Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity. *Molecular cell*, 78(3), pp.506–521.e6.

Zheng, H. & Xie, W., 2019. The role of 3D genome organization in development and cell differentiation. *Nature reviews. Molecular cell biology*, 20(9), pp.535–550.

Zhou, J., Ng, S.-B. & Chng, W.-J., 2013. LIN28/LIN28B: an emerging oncogenic

driver in cancer stem cells. *The international journal of biochemistry & cell biology*, 45(5), pp.973–978.

Zhou, Z. et al., 2015. Strong expression of EZH2 and accumulation of trimethylated H3K27 in diffuse large B-cell lymphoma independent of cell of origin and EZH2 codon 641 mutation. *Leukemia & lymphoma*, 56(10), pp.2895–2901.