DEVELOPMENT AND APPLICATION OF SOFTWARE TO UNDERSTAND 3D CHROMATIN STRUCTURE AND GENE REGULATION

Eric Scott Davis

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill
in partial fulfilment of the requirements for the degree of Doctor of Philosophy
in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2023

Approved by:

Douglas H. Phanstiel

Terrence S. Furey

Karen L. Mohlke

Michael I. Love

Hyejung Won

Daniel Dominguez

ABSTRACT

Eric Scott Davis: Development and Application of Software to Understand 3D Chromatin Structure and Gene Regulation
(Under the direction of Douglas H. Phanstiel)

Nearly all cells contain the same 2 meters of DNA that must be systematically organized into their nucleus for timely access to genes in response to stimuli. Proteins and biomolecular condensates make this possible by dynamically shaping chromatin into 3D structures that connect regulators to their genes. Chromatin loops are structures that are partly responsible for forming these connections and can result in disease when disrupted or aberrantly formed.

In this work, I describe three studies centered on using 3D chromatin structure to understand gene regulation. Using multi-omic data from a macrophage activation time course, we show that regulation temporally precedes gene expression and that chromatin loops play a key role in connecting enhancers to their target genes. In the next study, we investigated the role of biomolecular condensates in loop formation by mapping 3D chromatin structure in cell lines before and after disruption of NUP98-HOXA9 condensate formation. Differential analysis revealed evidence of CTCF-independent loop formation sensitive to condensate disruption. In the last study, we used 3D chromatin structure and multi-omic data in chondrocytes to link variant-gene pairs associated with Osteoarthritis (OA). Computational analysis suggests that a specific variant may disrupt transcription factor binding and misregulate inflammatory pathways in OA.

To carry out these analyses I built computational pipelines and two R/Bioconductor packages to support the processing and analysis of genomic data. The *nullranges* package contains functions for performing covariate-matched subsampling to generate null-hypothesis genomic data and mitigate the effects of confounding. The *mariner* package is designed for working with large chromatin contact data. It extends existing Bioconductor tools to allow fast and efficient extraction and manipulation of chromatin interactions for better understanding 3D chromatin structure and its impact on gene regulation.

ACKNOWLEDGEMENTS

My work was supported by funding from the NIH National Institute of General Medical Sciences T32 training grant (T32-GM067553), the UNC Graduate Student Transportation Grant, and an Essential Open Source Software award from the Chan Zuckerberg Initiative. None of this would be possible without the incredible network of support from my friends and family and guidance from my many mentors.

To my parents, Dave and Tammy Davis, I couldn't have done this without you – I mean, literally, I wouldn't exist! But in all seriousness, you've always encouraged my curiosity and instilled me with the confidence that I can do anything I set my mind to. You've been excellent role-models for hard-work and lifelong learning. There is nothing like a weekend with the two of you to put life into perspective and motivate me to be my very best. You've taught me so many things like how to renovate a house, how to make buckeyes and other delicious food, and especially how not to build a bike ramp! Even when you don't have all the answers, you've made sure to teach me the importance of searching for them myself. You've always been my strongest advocates and know how to encourage me when life gets hard. I look forward to many more years of fun family game nights with silly nicknames, comedy nights out, and road trips with far too many GPS running at once! To my sister Jennifer and brother-in-law Matt, thanks for the support and friendship all these years. You both are so kind, thoughtful and a true power couple. I am excited to witness the amazing life you are building together. Our Sunday running days have been a fun, exhausting, and cathartic. There's nothing like running to your limits to push other concerns away. Even though I'll never catch up to you guys, I'll never stop trying!

Thanks also to my extended families. To my family-in-law (or as we like to say "family-in-love"), Mark and September Stallings and Aunt Carolyn and Uncle Jr., I'm so lucky that you've loved and accepted me into your family with open arms and am very thankful for the great relationship we have. You've all been wonderful confidants, listening and offering excellent advice before and throughout graduate school. I appreciate that you never stop trying to understand exactly what Bioinformatics is. I look forward to many more years to excellent smoked food, trips to the beach and finding lemons in my backpack (P.S. check

behind your bathroom cabinet). To my extended families in Ohio and New York, I am thankful to have been raised in a loving community of supportive grandparents, aunts and uncles, and cousins.

To my friends, thanks for your support and patience with my research endeavors. Only true friends would be willing to wait in lab for 4 hours while I performed experiments after promising I was almost done. I'm glad I have you all to grab a drink with and rant about the stresses of life and debate about nonsense. Thanks especially to Preston for motivating my interest in Biology and research and being an amazing best friend since 5th grade. I'm glad you were able to get past my many annoying qualities and continue to hang out to this day. Thanks also to my UNC racquetball buddies – the sport is super fun and great for stress-relief but I never thought I'd make such great friends along the way.

To the Phanstiel lab members past and present, I can't imagine a better group of people to work with. Gradschool would have been 10 times harder without your friendship and support. Thanks especially to my "comp-cave" and "comp-cabana" people! We've had a lot of fun (probably too much fun) and it truly makes it a joy to come into lab. I'm sad that we missed a few years during the pandemic, but it feels like old times again when were finally able to reunite. Katie, I am so thankful that you were the first gradstudent in the lab. You created an awesome environment and are always have the best advice. I know you are doing amazing work at the NIH, but I wish you could have stayed around at least until I graduated. Nicole you are my "BCB-twin" and have been with me through every step of this program. Thanks for being a kind ear for me to complain to and a truly good friend. You are a smart, talented individual and you deserve all the recognition for your hard work. Katie and Nicole, one day I hope to be as good of a scientist as you both are! Sarah you are a remarkable person. I am amazed by all the neat projects you do and by the skill of your presentations! You bring life to the lab and I'll certainly miss our ice-cream Fridays and comp-cave quote quizzes. Marielle, it's been a blast hanging out and working with you. You are super talented in both the wet lab and dry lab and truly keep the lab going. I have no doubt that you will become a future leading PI if that's the path you choose. JP ("Japes") and Jess, we got so lucky that you both joined the lab. Jess you are incredibly talented, and I know you ensure the computational strength of the lab long after Nicole and I leave. Japes, you bring a fun, youthful energy to the group and have such a passion for science. It's been a privilege to mentor and work with you and am excited to see you lead the NIH one day and make a huge difference in the scientific community. Yoseli, Eliza, Andrea, Susan, and Zack, we couldn't have done

anything without your hard work. You all skillfully develop and adapt protocols, troubleshoot, and produce the highest quality data. You are also excellent scientists, writers, and awesome mentors to help me navigate gradschool.

To my scientific mentors – thank you for teaching me how to do science. Rob and Flori, thank you for teaching me everything I know about lab work starting from how to pipette water. I will always be grateful that you advocated for me and provided the guidance necessary to get into UNC's graduate program. To my committee (Terry, Karen, Mike, Hyejung, and Dan) and the BCB program managers (John, Cara, and Will) thank you for making my gradschool experience as smooth as possible. I've learned so many things from you all whether it be through 1 on 1 interactions or during seminars. Special thanks to Mike (and the Bioconductor community) for teaching me how to develop better software! Doug, thanks for training me as a scientist. You are among the best mentors I've had, and I hope I am half as understanding, ethical, and easy-going as you are. Thanks for giving me the opportunity to learn so many skills, guiding me when I struggled, and for your patience when I was stubborn. I look forward to the amazing research your lab will continue to produce and know that the scientific community is improved by your mentorship.

Most importantly of all, thanks to my smart, amazing, loving, kind and beautiful wife Amelia. You are truly a bright spot in my life and are the perfect partner for me. I couldn't have done any of this without your love and support. I am constantly amazed by your kind, caring nature and can't imagine how my life would be without you. I am especially grateful for your patience during these past few months during the stress of graduation, always knowing when I need a break even when I don't know it myself. Thanks for putting up with my oddities and making life so much fun! You are strong where I am weak, and you constantly inspire me to be a better person. I am excited to continue building a great life with you. I love you, always.

PREFACE

Much of the work presented here has been previously published. While I was the primary author and lead developer for each of the software chapters (chapters 2, 3, and 4), I was second author for chapter 5, and co-first author for chapters 6 and 7 where I performed computational analysis. This work would not have been possible without the contributions of many talented scientists as outlined below.

Chapter 2 is not currently published in a journal but is available on GitHub and under the following Zenodo DOI: https://doi.org/10.5281/zenodo.7514346. Though I was the primary developer, Kathleen S. M. Reed and Marielle L. Bond contributed code and Douglas H. Phanstiel acquired funding, conceptualized the project, and supervised development. This work was supported by grants from the NIH (E.S.D., T32-GM067553) and the Chan Zuckerberg Initiative (M.I.L., Essential Open Source Software for Science Round 3 Award).

Chapter 3 is currently under review at *Bioinformatics* and is distributed as an R/Bioconductor package under the name *nullranges*. I was the lead developer for the covariate-matching functionality, wrote vignettes, and wrote the manuscript. Other contributors to this work include Wancen Mu (lead developer for the block bootstrapping functionality and contributed vignettes), Stuart Lee (contributed code, design ideas, and portions of vignettes), Mikhail G. Dozmorov (contributed design ideas), Michael I. Love (acquired funding, conceptualized the project, contribute code and vignettes, and aided in writing), and Douglas H. Phanstiel (acquired funding, conceptualized the project, and aided in writing). The Bioconductor slack channel, and Tim Triche and Kasper Hanson in particular, provided feedback and advice. This work was supported by grants from the NIH (T32-GM067553 to E.S.D., R35-GM128645 to D.H.P., and R01-HG009937 to M.I.L) and an Essential Open Source Software award from the Chan Zuckerberg Initiative.

Chapter 4 is not currently published in a journal but is available on GitHub and under the following Zenodo DOI: https://doi.org/10.5281/zenodo.7514395. I conceptualized the project, was the lead developer, wrote vignettes, and the manuscript. Other contributors to this work include Manjari Kiran (contributed code and design ideas), Nicole Kramer (design ideas), Sarah Parker (design ideas), and

Douglas H. Phanstiel (acquired funding and supervised the project). This work was supported by grands from the NIH (T32-GM067553 to E.S.D.).

Chapter 5 has been previously published with the following citation: Ahn, Jeong Hyun, Eric S. Davis, Timothy A. Daugird, Shuai Zhao, Ivana Yoseli Quiroga, Hidetaka Uryu, Jie Li, et al. 2021. “Phase Separation Drives Aberrant Chromatin Looping and Cancer Development.” *Nature* 595 (7868): 591–95. J.H.A. designed the research, performed experiments, interpreted data and wrote the manuscript. J.H.A., Y-H. T., H.U., J. L., L. C., D.Z. and G.G.W. performed genomic data analysis. J.H.A. and S.Z. performed in vitro phase separation assays. D.P.K. conducted imaging quantification analysis. A.J.S., S.G.M., R.D.E. and S.D.B. performed proteomic analysis under the supervision of A.J.T. T.A.D. and J.H.A. performed single molecule tracking studies under the supervision of W.R.L. J.H.A. and J.L. performed murine leukaemia assays. E.S.D., I.Y.Q. and J.H.A. performed Hi-C mapping, data analysis and interpretation under the supervision of D.H.P. G.G.W. conceived the idea, supervised and designed the research, interpreted data, and wrote the manuscript with the inputs from all authors. This work was supported by NIH grants (R01-CA215284 and R01-CA218600 to G.G.W.; R35-GM128645 to D.H.P.; DP2GM136653 to W.R.L.; P20GM121293, R24GM137786, R01CA236209, S10OD018445, and TL1TR003109 to A.J.T; R01HL148128 and R01HL153920 to D.Z.), a Kimmel Scholar Award (to G.G.W.), Gabrielle’s Angel Foundation for Cancer Research (to G.G.W.), Gilead Sciences Research Scholars Program in haematology/oncology (to G.G.W.), When Everyone Survives (WES) Leukemia Research Foundation (to G.G.W.) and UNC Lineberger Stimulus Awards (to D.H.P. and to L.C.). E.S.D. was supported by the NIH-NIGMS training grant T32-GM067553. W.R.L. is a Searle Scholar, a Beckman Foundation Young Investigator, and a Packard Fellow for Science and Engineering. G.G.W. is an American Cancer Society (ACS) Research Scholar, an American Society of Hematology (ASH) Scholar in basic science, and a Leukemia and Lymphoma Society (LLS) Scholar.

Chapter 6 has been previously published with the following citation: Reed, Kathleen S. M., Eric S. Davis, Marielle L. Bond, Alan Cabrera, Eliza Thulson, Ivana Yoseli Quiroga, Shannon Cassel, et al. 2022. “Temporal Analysis Suggests a Reciprocal Relationship between 3D Chromatin Structure and Transcription.” *Cell Reports* 41 (5): 111567. I co-led this work with Kathleen S. M. Reed. K.S.M.R. designed and performed the majority of experiments, performed computational analysis, and wrote the paper. E.S.D.

developed software and performed computational analyses. M.L.B. performed some cell culture and genomic library preparation experiments. A.C. performed some cell culture experiments. E.T. performed some cell culture experiments and assisted with ChIP-seq. I.Y.Q. prepared ATAC-seq libraries. S.C. helped prepare ATAC-seq libraries and ChIP-seq cross-linking. K.T.W. assisted with cell culture experiments. H.W. supervised and helped interpret genomics data. I.H. oversaw the planning, supervision, and interpretation of some experiments. M.I.L. supervised computational analyses and software development. D.H.P. acquired funding, conceptualized the project, supervised experiments and data analyses, and helped write the paper. This work was supported by NIH grants (R35-GM128645 to D.H.P., R00HG008662 to D.H.P., R35GM143532 to I.H., and R01AG066871 to H.W.) and multiple NIH training grants (T32-GM067553 to E.S.D., T32 GM007092 to K.S.M.R. and E.T., and T32 GM135128 to M.L.B.). I.Y.Q. was supported by a BrightFocus Foundation postdoctoral fellowship. I.H. was supported by an award from the Cancer Prevention & Research Institute of Texas (RR170030). E.S.D. and M.I.L. were supported by a CZI Essential Open Source Software for Science (EOSS) Round 3 award.

Chapter 7 has been previously published with the following citation: Thulson, Eliza, Eric S. Davis, Susan D'Costa, Philip R. Coryell, Nicole E. Kramer, Karen L. Mohlke, Richard F. Loeser, Brian O. Diekman, and Douglas H. Phanstiel. 2022. "3D Chromatin Structure in Chondrocytes Identifies Putative Osteoarthritis Risk Genes." *Genetics* 222 (4). https://doi.org/10.1093/genetics/iyac141. I co-led this work with Eliza Thulson and Susan D'Costa. E.T. performed genomic experiments and wrote the manuscript. E.S.D performed computational analysis and wrote the manuscript. S.D. performed CRISPR-KO experiments and wrote the manuscript. P.R.C collected samples, N.E.K prepared some computational data, K.L.M supervised, R.F.L acquired funding and supervised. B.O.D and D.H.P acquired funding, co-supervised the work, and helped write the manuscript. This work was supported by NIH grants (R35-GM128645 to DHP, R37-AR049003 to RFL, and R56-AG066911 to BOD) and multiple NIH training grants (T32-GM067553 for ESD and NEK and T32-GM007092 for ET). The project was also supported by the National Center for Advancing Translational Sciences (NCATS) through NIH Grant UL1TR002489 and by the UNC Thurston Arthritis Research Center through a pilot and feasibility grant. ET was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2040435.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

3C	Chromatin confirmation capture
ABC	Activity by contact
AML	Acute myeloid leukemia
APA	Aggregate peak analysis
ATAC	Assay for transposase-accessible chromatin
BED	Browser extensible data format
BEDPE	Browser extensible data paired-end format
bp	Base pair(s)
ChIP	Chromatin immunoprecipitation
CTCF	CCCTC-binding factor
DBD	DNA binding domain
DBP	DNA binding protein
DNA	Deoxyribonucleic acid
DSL	Domain-specific language
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait loci
FC	Fold-change
FG	Phenylalanine-Glycine
FN-f	Fibronectin fragment
FS	F to S mutant
GC	Guanine-Cytosine (nucleotides)
GEO	Gene expression omnibus
GFP	Green fluorescent protein
GLEBS	GLE2-binding sequence
GREGOR	Genomic regulatory elements and GWAS overlap algorithm
GRO	global run-on
GWAS	Genome-wide association studies
H3K27ac	Histone H3 lysine 27 acetylation
HDF5	Hierarchical data format version 5
Hi-C	High-throughput sequencing of chromosome confirmation capture
HSPC	Hematopoietic stem and progenitor cell
IDR	Intrinsically disordered region

IFN_{γ}	Interferon-gamma
IGV	Integrative genomics viewer
ITS	Inferred transcription score
Kb/kb	Kilobase(s)
KR	Knight-Ruiz
LD	Linkage disequilibrium
LFC	Log2 fold-change
LLPS	Liquid-liquid phase separation
LPS	Lipopolysaccharide
Mb	Megabase(s)
MSCV	Murine stem cell virus
NHA9	NUP98-HOXA9
OA	Osteoarthritis
PCR	Polymerase chain reaction
PRO	Precision nuclear run-on
qPCR	Quantitative polymerase chain reaction
QTL	Quantitative trait loci
RNA	Ribonucleic acid
SE	Super enhancer
SEM	SNP effect matrix
SEMpl	SEM pipeline
SIP	Significant interaction peak caller
SNP	Single nucleotide polymorphism
SOCS2	Suppressor of cytokine signaling 2
TAD	Topologically associating domain
TF	Transcription factor
TPM	Transcripts per million
VEP	Variant effect predictor
WT	Wildtype

Chapter 1: Introduction

One of the most fascinating aspects of human biology is how nearly every cell within the body contains the same genome. It is remarkable that the same genetic blueprint gives rise to a wide variety of cells and functions. Couple this with the fact that only 1-2% of the human genome codes for proteins. The rest of the non-coding space contains epigenetic regulatory regions, non-coding gene components (introns), and repetitive sequences. It is partly due to the epigenetic regulation of genes that we develop into complex organisms. However, gene regulation is a delicate dance. Genes must be expressed, or transcribed, in the correct amount and at the appropriate time to maintain proper cellular function. Misregulation caused by any number of internal or external factors can have detrimental effects on development and result in disease. This raises some of the most fundamental and interesting questions about human biology. How do regulators traverse a genome of 3 billion base-pairs to find their target genes? How do cells coordinate these changes in response to changing environmental stimuli? How can disruptions in these processes lead to disease? My dissertation explores the interplay between 3D chromatin structure and gene transcription using custom software developed to address these outstanding questions.

1.1. Epigenetic regulation coordinates gene expression

Most of our genome is non-coding space. Historically believed to be non-functional “junk” DNA, the non-coding space is now understood to be essential for regulating our protein-coding genome (ENCODE Project Consortium 2012). The non-coding genome contains regulatory regions such as enhancers, silencers, and insulators which recruit transcription factors (TFs) and DNA-binding proteins (DBPs) that modulate gene transcription. Changes to the accessibility of chromatin through histone modification can affect the ability of TFs to bind to these elements. For example, acetylation of lysine 27 of histone H3 (H3K27ac) is associated with active enhancers, while trimethylation of lysine 9 of histone H3 (H3K9me3) is associated with silencers (Moyra Lawrence, Daujat, and Schneider 2016).

These and other marks have been characterized genome-wide in many cell and tissue types through the work of large consortiums such as the Encyclopedia of DNA Elements (ENCODE) and the Roadmap Epigenomics project (ENCODE Project Consortium 2012; Bernstein et al. 2010). TFs and chromatin accessibility can be mapped with high-throughput sequencing assays such as chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) and assay for transposase-accessible chromatin followed by high-throughput sequencing (ATAC-seq) (Buenrostro et al. 2013). These assays have been instrumental in identifying the location of epigenetic regulators and the proteins bound to their sequences.

Epigenetic regulators can modify gene expression, but a major challenge is identifying which gene or genes they effect. Regulatory regions are scattered across the non-coding space. They can be found near the genes they regulate as well as millions of base-pairs away. Some regulators are even found in the introns of other genes. So how does the cell coordinate the connections between enhancers and their target genes?

1.2. 3D Chromatin structure connects regulatory elements to target genes

The three-dimensional (3D) arrangement of chromatin within the cell is thought to play a major role in bringing regulatory elements to their target genes. Since DNA is over 2 meters long, it must be packaged efficiently to allow access to genes and regulatory regions for proper cellular function. This is accomplished through the formation of domains such as compartments, topologically associating domains (TADs), and loops, which dictate the accessibility of genomic regions and how they interact.

Compartments are structures that generally separate active and inactive chromatin. The genome can be separated into two large compartments – termed “A” and “B” compartments (Lieberman-Aiden et al. 2009). The “A” compartment contains actively transcribed regions that preferentially interact and is located towards the center of the nucleus. The “B” compartment is mostly composed of inactive heterochromatin and forms around the periphery of the nucleus. Initially thought to only occur at megabase scales, recent work by ENCODE suggests that fine-scale compartments as small as kilobases connect enhancers throughout the genome (H. Gu et al. 2021).

TADs and loops are structures that are typically bordered by the CCCTC-binding factor (CTCF), an insulating protein that helps maintain 3D chromatin structure. TADs are 0.2-1 Mb sized regions that

preferentially interact with themselves, contain actively transcribed genes, and create insulated boundaries. Loop domains are more punctate than TADs, forming stable interactions between two genomic loci that can be millions of base-pairs away. Cells contain tens of thousands of loops – most of which are conserved across cell types and species (Rao et al. 2014). However, some loops are cell type- and context- specific and are thought to be important for connecting regulatory regions to their target genes (Phanstiel et al. 2017; Ahn et al. 2021).

In situ Hi-C is a chromatin conformation capture method coupled with high-throughput sequencing that measures the pair-wise interactions between all genomic loci (Rao et al. 2014). In short, Hi-C uses a chemical fixation agent to “freeze” the genome in its 3D conformation. The DNA is then digested with restriction enzymes into fragments that are ligated by proximity – resulting in chimeric DNA connecting linearly distant regions that were interacting in 3D space. After sequencing, these fragments are computationally mapped back to their linear positions while recording which loci were connected. Downstream software can then be used to identify compartments, TADs, and loops from the Hi-C contact maps. However, Hi-C experiments often require billions of reads to identify loops and fine-compartments and processing complex experiments can be difficult to orchestrate. Chapter 2 describes dietJuicer, a Hi-C processing pipeline that is optimized to process Hi-C data efficiently and coordinate processing of large, complex Hi-C experiments like those described in chapters 5, 6, and 7.

In 2014 Rao et al. mapped 3D chromatin structure in nine cell types, showing that chromatin loops frequently link enhancers to target genes (Rao et al. 2014); this suggests that looping is important for cell-type specific regulation. However, a few years later the same group degraded cohesin (one of the proteins responsible for loop formation) and eliminated all looping domains (Rao et al. 2017). Despite the complete loss of loop structures, gene expression was largely unaffected (Rao et al. 2017).

Chapter 6 explores these topics in more depth using a multi-omic macrophage inflammatory activation time course to disentangle the complex relationship between 3D chromatin structure and transcription. In this chapter we find that loops linking enhancers and genes show stronger concordance of H3K27ac and expression than can be explained by distance or contact frequency alone. Our time course data allows us to place these events in temporal order where we find that regulation precedes expression

among enhancer-gene pairs. While 3D chromatin structure may promote transcription, we find that transcription itself may impact loop formation.

1.3. Phase separation, looping, and cancer

In humans, loops are formed in a process called “loop extrusion” where chromatin is pulled through a ring-like protein complex called cohesin until it collides with two convergently oriented CTCFs on either end (Davidson and Peters 2021; Sanborn et al. 2015). The loop extrusion model is thought to explain how most loops in the genome form and maintain structure. Recently, there has been growing evidence that interactions between chromatin, RNA, and protein can form biomolecular condensates that not only regulate gene transcription, but also shape 3D chromatin structure (Tong et al. 2022; Banani et al. 2017). Biomolecular condensates are formed in a process known as liquid-liquid phase separation (LLPS), where weak multivalent interactions form a dense phase of concentrated molecules. Proteins containing intrinsically disordered regions (IDRs) caused by amino acid repeats result in unstructured domains which associate and form LLPS condensates. Nearly 30% of protein-coding regions are disordered and many proteins, including TFs, DNA- and RNA- binding proteins, contain IDR tails suggesting that LLPS may be a wide-spread phenomenon (Ruff and Pappu 2021; Boija et al. 2018). When these proteins bind to their targets, they form condensates which concentrate TFs, mediator proteins, and even RNA Pol II (Guo et al. 2019). For example, clusters of enhancers, called super enhancers (SEs), recruit large numbers of IDR-containing TFs which form LLPS SE condensates that concentrate the molecules needed for high levels of transcription (Sabari 2018).

In addition to regulating transcription, LLPS is used for many biological processes such as chromatin organization and intracellular signaling. Changes to DNA or protein through post translational modifications (PTMs), for example, can trigger condensates to form and sequester chromatin into euchromatic or heterochromatic states (Shin 2018). Yes-associated protein (YAP) has been shown to form LLPS condensates in response to hyperosmotic and mechanical stress and is involved in the tumor suppressor Hippo signaling pathway (D. Cai et al. 2019; G. Zhu et al. 2021). Despite these cellular roles, little is known about how condensates form and interact. Some condensates require cofactors or scaffolds composed of RNA or protein to form while others can form independently (Roden and Gladfelter 2021). Interestingly, condensates appear to have specificity, maintaining distinct molecular compositions.

Disease can result when LLPS-controlled biological processes are disrupted or aberrant LLPS occurs. For example, translocations that result in fusions between DNA-binding domains (DBDs) and IDRs can lead to aberrant LLPS that results in a severe subtype of acute myeloid leukemia (AML) (Alberti and Hyman 2021; Gough et al. 2014). We describe this in more detail in chapter 5, where we explore the effect that a NUP98-HOXA9 fusion has on cancer progression and 3D chromatin structure. Here we provide some of the first direct evidence of aberrant LLPS-looping that results in the expression of protooncogenes and drives AML.

1.4. Loops connect variants to target genes

Since loops form connections between genes and their regulators, it stands to reason that they can be useful for identifying mutations that disrupt regulatory networks in the non-coding space. Approximately 10% of single nucleotide polymorphisms (SNPs), or genetic variants, fall into coding regions and result in missense mutations that disrupt protein synthesis (Edwards et al. 2013). Most diseases are complex and polymorphic, meaning that combinations of mutations predispose individuals to disease. Mutations in regulators are likely responsible for polymorphic disease, but since these occur in the non-coding space it can be challenging to identify genes affected by these variants.

Genome-wide association studies (GWAS) identify variants that are statistically associated with a phenotype (such as a disease) among individuals. Because large chunks of the genome are inherited in blocks, a phenomenon known as linkage-disequilibrium (LD), benign mutations hitch-hike along with the disease-associated variant or variants. Furthermore, the strongest association may not be the causal signal and LD can obscure the contributions of weaker effects.

Narrowing down these putative variants is a technique known as fine-mapping. There are a host of fine-mapping approaches such as using allelic imbalance to identify quantitative trait loci (QTL), overlap with regulatory elements, and finding disrupted TF binding site motifs (Kumasaka, Knights, and Gaffney 2016; Farh et al. 2015; Nishizaki et al. 2020). These approaches can be used in combination with 3D chromatin structure data to link putative variants to target genes. For example, Hi-C-coupled multimarker analysis of genomic annotation (H-MAGMA) uses chromatin interaction profiles to assign SNPs to genes (Sey et al. 2020). Gene prioritization through expression QTL (eQTL) can be used to identify genes associated with variants at a locus (Giambartolomei et al. 2014). However, since many QTL are context-

specific, having data in the right cell-type or stimulation conditions is necessary to find causal variant-gene pairs (Umans, Battle, and Gilad 2021; Alasoo et al. 2018).

In chapter 7 we use a combination of these approaches to identify causal variants and prioritize affected genes in Osteoarthritis (OA). After using epigenetic profiles to identify chondrocytes as a likely cellular context, we use chromatin loops to link variants overlapping regulatory regions to genes that are responsive to a previously published model for OA progression. Focusing on one variant-gene pair with evidence of TF motif disruption, we validate the gene's affect on inflammation with CRISPR-KO experiments to explain its role in OA.

1.5. Bioconductor ecosystem for genomic data analysis

As the cost of sequencing drops, the amount of genomic data is growing drastically (Lander et al. 2001; Papageorgiou et al. 2018). Furthermore, new sequencing assays are constantly being developed to improve our understanding of genetics, molecular biology, and human health. As these technologies progress, software must be developed to keep up with the needs of biologists. Genomic-oriented software has been developed ad hoc by a wide array of groups to support the analysis of growing data. Such practices have led to duplication of effort each with slightly different implementations and results, inconsistent interfaces, and a codebase written in a variety of programming languages for different tasks. This is not only inefficient, but it also creates confusion as biological results are different depending on the software used. Biologists must be programming polyglots as a single analysis might require tools not available in all languages. Consequently, there is a lack of reproducibility and biological discovery is impeded.

Bioconductor provides a platform for publishing biologically oriented software packages written in the R programming language (Gentleman et al. 2004). Packages submitted to Bioconductor must conform to specific formatting requirements, contain unit tests for accuracy, and be designed in a modular fashion. Bioconductor also provides data-structures and classes for representing common types of genomic data and encourages reuse of classes and methods across packages. This standardization leads to interoperability between packages written by different groups, easier adoption as similar conventions are used across methods, and reusing existing components for new packages. Finally, Bioconductor is open source and freely available, ensuring access and collaboration for the entire community.

Chapters 3 and 4 describe two R/Bioconductor software packages. Chapter 3 introduces *matchRanges*, a function for generating covariate-matched subsets of ranged or paired genomic data. *matchRanges* is part of the *nullranges* package which contains additional methods for generating genomic null-hypothesis sets. *matchRanges* enables users to disentangle causal effects from genomic studies and has been used in many projects, including chapter 6. Chapter 4 extends and improves existing tools for exploring Hi-C data. It describes *mariner,* software that improves the infrastructure for performing analysis on Hi-C data like those carried out in chapter 5. Together, these software packages expand the Bioconductor ecosystem and provide tools for improving genomic analysis to help understand the role of 3D chromatin organization and gene regulation.

Chapter 2: dietJuicer: A Snakemake pipeline for processing Hi-C data

2.1. Introduction

3D chromatin structure is important for gene regulation, human development, and disease. It can be mapped by a genomic method called high-throughput sequencing of chromosome conformation capture (Hi-C); however, Hi-C generates massive datasets that are computationally challenging to process due to their large size and intricate analysis steps. Existing software for processing Hi-C data is not optimized for speed or file storage requirements and is difficult to adapt to local computational environments. To address these issues, we created dietJuicer, a streamlined pipeline for efficient, accurate, and reproducible processing of Hi-C datasets.

Bioinformatic pipelining software is used to automate the processing and analysis of biological data. These tools are essential for researchers working with large genomic datasets, like Hi-C, that require running time-consuming processing steps with many software programs. Pipelining tools simplify these steps by mapping inputs to outputs in a modularized fashion that increases both the efficiency and reproducibility of data processing.

Snakemake and Nextflow are the two most popular open-source workflow management systems for processing biological data. Both tools boast the ability to create reproducible and scalable analyses that can be deployed seamlessly on computing clusters and cloud environments without modifying the pipeline steps (Mölder et al. 2021). Snakemake follows an output-first approach with workflows written in a python-like syntax that define how to create output files from input files. Nexflow is based on the bash scripting language and defines modular processes that are chained together into workflows. While both tools require learning their unique domain-specific language (DSL), the flexibility and maintainability afforded outweighs the alternative method of writing and maintaining custom scripts.

Hi-C and its variants are widely used to map pair-wise interactions between genomic regions. Hi-C data is generated by sequencing pairs of DNA fragments that are nearby in 3D space. These data are invaluable in understanding how 3D chromatin interactions contribute to modulating gene transcription in

human development and disease. The major steps of Hi-C processing are alignment, filtering, matrix aggregation and normalization ("Hi-C Processing Pipeline" n.d.). During alignment, Hi-C read pairs in FASTQ format are mapped to a reference genome such as the GRCh38 human genome (Cock et al. 2010). The mapped reads are then filtered for valid Hi-C contacts by sorting paired reads according to their chromosome and position, merging across sequencing replicates (replicates within an experiment), and removing duplicate reads. Duplicates are removed after merging because reads resulting from polymerase chain reaction (PCR) duplication may be present in both sequencing replicates. The final steps are matrix aggregation and normalization, where Hi-C contacts are binned into kilo-base (Kb) resolutions and balanced for technical and biological biases such as fragment length and mappabiliity (Servant et al. 2015; Hu et al. 2012; Imakaev et al. 2012). This results in matrix files (typically `.hic` or `.mcool` formats) that contain all pair-wise contacts between binned genomic loci (Durand, Shamim, et al. 2016; Abdennur and Mirny 2020).

HiC-Pro and Juicer are two of the most widely used tools for processing Hi-C data from raw reads to normalized contact maps (Durand, Shamim, et al. 2016; Servant et al. 2015). Neither pipeline utilizes workflow management tools, instead using Make (HiC-Pro) or bash scripting (Juicer) to assemble their steps. While both tools provide the ability to run on a variety of computing cluster systems and cloud environments, there is a huge maintainability burden since each additional platform requires modification of workflows. Additionally, HiC-Pro and Juicer are executed from the command line on individual Hi-C libraries which makes complex experimental designs laborious to launch and errors difficult to track.

To improve Hi-C processing pipelines we have created dietJuicer, a Snakemake pipeline for processing Hi-C data and its variants. dietJuicer uses Juicer's scripts for Hi-C processing but manages and deploys these steps across compute clusters and environments with Snakemake. dietJuicer uses configuration files and a user-created sample sheet to launch the pipeline. This greatly simplifies deployment of complex experimental setups and requires no programming expertise. Due to Snakemake's management capabilities, dietJuicer minimizes runtime by allowing non-dependent steps to execute in parallel. Similarly, it conserves storage space by removing intermediate files as they are no longer needed. Tracking progress and recovering from errors is greatly simplified with log and benchmarks files that document each step in the pipeline. Finally, processing terabyte-sized data is straightforward with

configuration files that define compute resources for each step. In the following section, we describe the design and usage of dietJuicer.

2.2. Design and usage

There are two major phases to the `dietJuicer` pipeline: `dietJuicerCore` and `dietJuicerMerge`. The first phase, `dietJuicerCore`, performs the main pipeline steps of alignment, filtering, and matrix aggregation/normalization on each Hi-C sample. The second phase, `dietJuicerMerge`, uses output files produced by the core workflow to merge data across samples to create deeper, aggregated Hi-C maps. This is particularly useful for combining biological replicates to visualize contact maps of experimental conditions. It is important to note that biological and technical replicates (i.e., everything except sequencing replicates) should first be processed individually with the `dietJuicerCore` workflow before being combined with `dietJuicerMerge`, since the core workflow includes a PCR duplicate removal step that is not included in the merge workflow. The following sections describe in detail how to use each phase of the pipeline.

2.2.1. Initial processing with `dietJuicerCore`

The core pipeline can be run in four steps:

1. Clone the repository into the desired working directory with the following bash command:

```
git clone https://github.com/EricSDavis/dietJuicer.git <DIRECTORY>
```

2. Edit the tab-separated `samplesheet.txt` file with sample information. Required columns include “Read1”, “Read2”, and “Sequencing_Directory” where “Read1” and “Read2” list the filenames for each FASTQ file and “Sequencing_Directory” contains the full path to the directory containing each file. No naming convention is needed for FASTQ files, but they must be compressed with `gzip`. At least one additional column is needed to determine which files should be merged as a group (i.e., sequencing replicates). Optional columns can be included as desired to capture metadata about each sample. Common columns to include might be “Project”, “Cell_Type”, “Genotype”, “Bio_Rep”, “Tech_Rep”, and “Seq_Rep”.

3. Edit the `config/config.yaml` file with 1) the path to `samplesheet.txt`, 2) A list of columns from the samplesheet.txt to use for grouping, 3) paths to genome build-specific reference parameters, and 4) restriction enzyme site information used in the Hi-C experiment (for Micro-C use “none”). The column names listed in the `groupBy` parameter are used to create prefixes for the output files. The following code chunk shows an example of the `config/config.yaml` file:

```yaml
## Path to sample sheet
samplesheet: 'samplesheet.txt'

## Group columns (missing columns will become part of the same group)
groupBy: ['Project', 'Cell_Type', 'Genotype', 'Bio_Rep', 'Tech_Rep']

## Genome-specific reference parameters
fasta: '/path/to/hg38/BWAIndex/genome.fa'
chromSizes: 'path/to/hg38_chromSizes.txt'

## Restriction site information
site: "MboI" # or 'none' for Micro-C
site_file: 'restriction_sites/hg38_MboI_chr.txt'
ligation: "GATCGATC"

## Set splitsize for parallel processing
splitsize: 200000000

## Java memory for buildHiC (hic/norm rules)
javaMem: "250880"

## Additional options
mapq0_reads_included: 0
```

While most datasets should complete processing with the default parameters, large datasets may require more compute resources. Advanced users can edit the `config/cluster.yaml` file to adjust the default or rule-specific memory and runtime requirements.

4. A submission script for the SLURM job scheduler is included with `dietJuicer`. Use the following command to launch a long-running, low-resource job that will spawn other jobs in the

pipeline as dependencies are fulfilled: `sbatch dietJuicerCore.sh`. To use other job schedulers, edit `dietJuicierCore.sh` file as appropriate.

Successful completion of the pipeline will result in the directory structure shown below where `{jobid}` is the SLURM-assigned job identifier, `{group}` is the sample name created by joining columns listed in the `groupBy` parameter in the `config/config.yaml` file, `{rule}` is the name of step in the pipeline, and `{splitName}` refers to split files created for parallel processing.

```
slurm-{jobid}.out
output/
├── logs_slurm
└── {group}
    ├── benchmarks
    │   ├── {group}_{rule}_split{splitName}.tsv
    │   └── ...
    ├── logs
    │   ├── {group}_{rule}_split{splitName}.err
    │   └── ...
    ├── {group}_dedup_merged_nodups.txt.gz
    ├── {group}_inter_30.hic
    ├── {group}_inter_30_hists.m
    ├── {group}_inter_30.txt
    ├── {group}_inter.hic
    ├── {group}_inter_hists.m
    ├── {group}_inter.txt
    ├── {group}_splitR1_done.txt
    └── {group}_splitR2_done.txt
```

2.2.2. Creating merged Hi-C maps with `dietJuicerMerge`

After running the core workflow on the biological and technical replicates (if applicable), the resulting Hi-C maps can be merged by following the four steps below:

1. If the `dietJuicerMerge` workflow is being run in the same directory as `dietJuicerCore`, then this step can be skipped. Otherwise, clone the repository into the desired working directory with the following bash command:

```
1  git clone https://github.com/EricSDavis/dietJuicer.git <DIRECTORY>
```

2. Edit the tab-separated `samplesheet.txt` file to include paths to `dietJuicerCore` output files. Columns “merged_nodups”, “inter”, and “inter30” should be added to the sample sheet with full paths to their corresponding output files. A python script, `scripts/addPaths.py`, is included to automatically add these columns and file paths to an existing sample sheet from the `output` directory of `dietJuicerCore`. A minimum of one additional column is required to determine which combination of samples to combine into a group (see `groupBy` parameter in step 3). Optional columns can be included to capture sample metadata.
3. Edit the `config/config.yaml` file to set reference files and define file merging. Merging is controlled by the `groupBy` parameter which uses a list of column names from the sample sheet to create the group identifier used as a prefix for all output files. Different merge-levels can be achieved by altering columns included in `groupBy`. For example, `groupBy = [“Project”, “Cell”, “Condition”]` would produce prefixes “Project_Cell_Condition” that correspond to each unique combination of values in these columns. A sample sheet with one project, two cells, and two conditions would result in four groups: “Project_Cell1_Condition1”, “Project_Cell1_Condition2”, “Project_Cell2_Condition1”, and “Project_Cell2_Condition2”. Only columns listed are used for grouping; if “condition” is not listed (i.e., `groupBy = [“Project”, “Cell”]`) then two groups will result with prefixes, “Project_Cell1”, and “Project_Cell2”. For complex grouping or custom naming, a single column may be provided with the desired output file names.
4. A submission script for the SLURM job scheduler is included with `dietJuicer`. Use the following command to launch a long-running, low-resource job that will spawn other jobs in the pipeline as dependencies are fulfilled: `sbatch dietJuicerMerge.sh`. To use other job schedulers, edit `dietJuicierMerge.sh` file as appropriate.

Successful completion of the pipeline will result in new folders in the output directory for each merged group with `*.hic` files, raw contact data (`*merged_nodups.txt.gz`), and library statistics (`*inter` and `*inter_30` files).

2.3. Discussion and conclusion

dietJuicer has improved the experience of processing Hi-C data over traditional Juicer by simplifying and managing the execution of deeply sequenced datasets with complex experimental designs (Kathleen S. M. Reed et al. 2022; Bond et al. 2022; Kelly et al. 2022). For example, dietJuicer was recently used to process approximately 24.5 billion reads in a macrophage activation time course, one of the deepest sequenced Hi-C datasets to date (Reed et al. 2022). Similarly, a megakaryocyte differentiation time course with three timepoints across four biological replicates totaling over 18 billion reads was also recently processed with dietJuicer (Bond et al. 2022). Both systems benefited from the sample sheet-driven execution of the pipeline during creation of multiple timepoint-aggregated contact maps.

dietJuicer is based on the Juicer pipeline and follows the recommendations for processing Hi-C data set forth by the Encyclopedia of DNA Elements (ENCODE) research consortium (*Hic-Pipeline: HiC Uniform Processing Pipeline* n.d.; Durand, Robinson, et al. 2016; Durand, Shamim, et al. 2016). Since the initial development of Juicer, new command line tools and data format standards have emerged. For example, the pairtools suite provides a consistent interface to perform many Hi-C pipeline steps and returns outputs in the `.pairs` format recommended by the 4D Nucleome Consortium (Goloborodko et al. 2018). Future directions for dietJuicer could be to replace its Juicer scripts with pairtools to benefit from the latest speed and memory updates.

One limitation of dietJuicer is that it cannot easily run individual steps. dietJuicer is intended to be launched from a sample sheet to reduce the programming expertise required. While this allows dietJuicer to be run by a wider audience, it cannot easily execute individual steps of the pipeline. Furthermore, to keep configuration files concise, not all parameters are abstracted from the workflow. To address these limitations, future development could be aimed at further modularizing the dietJuicer by re-writing it with Nextflow.

dietJuicer is a modern, flexible Hi-C processing pipeline. By using the workflow management of Snakemake with a sample sheet-driven interface, dietJuicer offers a simple and efficient way to process billions of Hi-C reads. Broadly, these processing benefits reduce the burden of time-consuming and complex nature of Hi-C processing to accelerate biological discovery.

Chapter 3: matchRanges: Generating null hypothesis genomic ranges via covariate-matched sampling[1]

3.1. Introduction

Genome-wide analyses can provide valuable insights into biological systems and human disease by revealing patterns of features that may be missed by interrogation of individual loci. Determining if observed trends are statistically significant, however, commonly requires comparing attributes between a focal and a null set of genomic loci. Accurate inference requires that null sets exhibit similar distributions of covariates observed in the focal set, to mitigate interpretability issues due to confounding. This can be challenging since many common covariates (e.g., GC content, gene density, histone acetylation, chromatin accessibility, etc.) are not uniformly distributed throughout the genome and must therefore be explicitly controlled when selecting null sets of loci (Bickel et al. 2010). Propensity score-matching is a computational method that allows for the selection of covariate-matched sets and several packages implement it within the R programming language (Ho et al. 2011; Sekhon 2011). However, these packages can be slow for genome-scale data sets and are not well-integrated into genomic analysis platforms such as Bioconductor making them difficult to incorporate into genomic workflows.

To address this problem, we developed *matchRanges*, an efficient and convenient tool for generating covariate-matched sets of genomic ranges from a pool of background ranges. *matchRanges* computes for each range a propensity score, the probability of assigning a range to focal or background groups, given a chosen set of covariates. It provides three methods including nearest-neighbor matching, rejection sampling, and stratified sampling for null set selection (Ho et al. 2007). Additionally, *matchRanges* provides utilities for accessing matched data, assessing matching quality, and visualizing covariate distributions. The code has been optimized to accommodate genome scale data sets, such that most matchRanges functions can efficiently process sets of millions of loci in seconds on a single core (**Fig 3.S1**). *matchRanges* accepts and returns common Bioconductor objects, such as *GRanges* and *GInteractions* for

[1] Parts of the work in this chapter are currently under review at *Bioinformatics*.

seamless integration with existing workflows (Gentleman et al. 2004; Michael Lawrence et al. 2013; Lun, Perry, and Ing-Simmons 2016) (**Fig 3.S2**). matchRanges is distributed as part of the *nullranges* package, with multiple software vignettes. *matchRanges* is ideally suited to cases in which feature covariates are known and differ between focal and pool sets. If controlling for local genomic context is of interest, the sister function *bootRanges* may be more appropriate.

3.2. The matchRanges workflow

To generate a covariate-matched set of ranges, users can provide *data.frame*, *GRanges* or *GInteractions* R objects annotated with columns describing one or more potentially confounding covariates (Dowle and Srinivasan 2021; Michael Lawrence et al. 2013; Lun, Perry, and Ing-Simmons 2016). The *matchRanges* function takes as input a "focal" set of data to be matched and a "pool" set of background ranges to select from. *matchRanges* performs subset selection based on the provided covariates and returns a null set of ranges with distributions of covariates that approximately match those of the focal set (**Fig 3.1A**). Users should ensure that focal and pool sets share features across all strata being matched to obtain an adequately matched set (Westreich and Cole 2010; Y. Zhu et al. 2021). This allows for an unbiased comparison between features of interest in the focal and matched sets without confounding by matched covariates. As the returned matched sample object is the same class as the inputs, it can be easily incorporated into new or existing Bioconductor workflows (S. Lee, Cook, and Lawrence 2019).

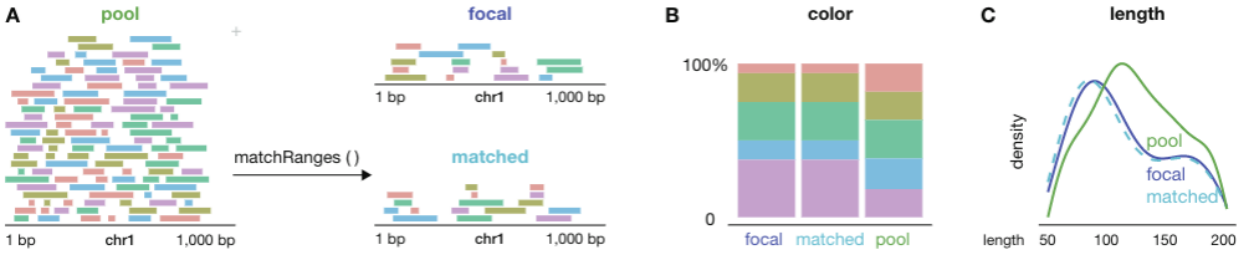


Figure 3.1. matchRanges workflow. (A) A schematic demonstrating how the *matchRanges* function can be used to select a set of *GRanges* matched for covariate features of color and length. **(B and C)** Example visualization of covariate distributions for assessing matching quality and covariate balance. Figure generated with the *plotgardener* R/Bioconductor package (Kramer et al. 2022).

A key aspect of inference based on covariate matching is visual inspection of the results. We provide several functions to assess the overall quality of matching, including plots of the distribution of covariates amongst the "focal", "pool", and "matched" sets (**Fig 3.1B and C**). Accessor functions allow users

to easily extract data for further inspection or integration with covariate balance packages, such as *cobalt* (Greifer 2020) (**Fig 3.S3**). Since matching is a pre-processing step, multiple matching methods can be tried and assessed before downstream analyses.

Detailed documentation on how to use *matchRanges* and when to use each matching method is available at an accompanying website (https://nullranges.github.io/nullranges), which contains step-by-step tutorials and biological case studies demonstrating the power of *matchRanges*.

3.3. Usage and implementation

In this section, we highlight the features of `matchRanges()` and its associated functions using a simulated dataset generated by the `makeExampleMatchedDataSet()` utility function. We also provide some guidance for choosing among the supported matching methods, describe how to assess covariate balance with *matchRanges* and the *cobalt* R package, and explain some implementation details for future developers. To see *matchRanges* used in real biological examples, visit the *Case study I: CTCF occupancy*, and *Case study II: CTCF orientation* vignettes included with the *nullranges* package and on the online documentation (https://nullranges.github.io/nullranges/articles).

3.3.1. Terminology

matchRanges references four sets of data: “focal”, “pool”, “matched” and “unmatched”. The focal set contains the outcome of interest (`Y=1`) while the pool set contains all other observations (`Y=0`). *matchRanges* generates the matched set, which is a subset of the pool that is matched for provided covariates (via the `covar` argument) but does not contain the outcome of interest (i.e., `Y=0`). Finally, the unmatched set contains the remaining unselected elements from the pool. The diagram below depicts the relationships between the four sets **(Fig. 3.2)**.

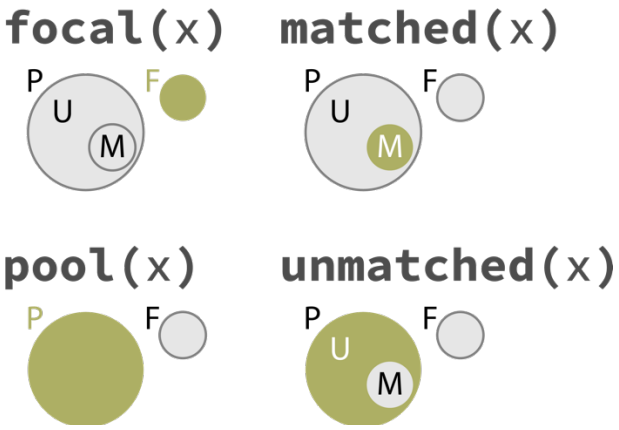


Figure 3.2. Sets used in matchRanges. Visual representation of the relationships between the four sets, “focal” (F), “pool” (P), “matched” (M), and “unmatched” (U), used in *matchRanges*. The focal and pool sets are distinct, while the matched and unmatched sets are subsets of the pool. The matched set is the same size and has the same distribution of covariates as the focal set. For subset selection to work adequately, the pool must be much larger than the focal set and both sets must share features across all strata being matched.

3.3.2. Methodology

matchRanges uses propensity scores to perform subset selection on the pool set such that the resulting matched set contains similar distributions of covariates to that of the focal set. Briefly, a propensity score is the conditional probability of assigning an element (in our case, a genomic range) to a particular outcome (Y) given a set of covariates. Propensity scores are estimated using a logistic regression model where the outcome Y=1 for focal and Y=0 for pool, over the provided covariates. The resulting propensity scores are used to select matches using one of three available matching options: “nearest”, “rejection”, or “stratified” with or without replacement. For more information see the section on *Choosing the method parameter* below.

3.3.3. Matching with matchRanges

We will use a simulated data set to demonstrate matching across covarying features:

```
library(nullranges)
set.seed(123)
x <- makeExampleMatchedDataSet(type = 'GRanges')
x
```

```
## GRanges object with 10500 ranges and 3 metadata columns:
##           seqnames      ranges strand |  feature1  feature2    feature3
##              <Rle>   <IRanges>  <Rle> | <logical> <numeric> <character>
##       [1]     chr1       1-100      * |      TRUE   2.87905           c
##       [2]     chr1       2-101      * |      TRUE   3.53965           c
##       [3]     chr1       3-102      * |      TRUE   7.11742           c
##       [4]     chr1       4-103      * |      TRUE   4.14102           a
##       [5]     chr1       5-104      * |      TRUE   4.25858           c
##       ...      ...         ...    ... .       ...       ...         ...
##   [10496]     chr1 10496-10595      * |     FALSE   1.23578           b
##   [10497]     chr1 10497-10596      * |     FALSE   1.69671           a
##   [10498]     chr1 10498-10597      * |     FALSE   6.11140           a
##   [10499]     chr1 10499-10598      * |     FALSE   2.21657           d
##   [10500]     chr1 10500-10599      * |     FALSE   5.33003           b
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

The simulated dataset has 3 features: logical “feature1”, numeric “feature2”, and character/factor “feature3”. We can use `matchRanges()` to compare ranges where feature1 is `TRUE` to ranges where feature1 is `FALSE`, matched by feature2 and/or feature3:

```
set.seed(123)
mgr <- matchRanges(focal = x[x$feature1],
                   pool = x[!x$feature1],
                   covar = ~feature2 + feature3)
mgr
```

```
## MatchedGRanges object with 500 ranges and 3 metadata columns:
##         seqnames      ranges strand |  feature1  feature2    feature3
##            <Rle>   <IRanges>  <Rle> | <logical> <numeric> <character>
##     [1]     chr1   4373-4472      * |     FALSE  8.959578           d
##     [2]     chr1   9740-9839      * |     FALSE  0.959336           e
##     [3]     chr1   7755-7854      * |     FALSE  2.107003           c
##     [4]     chr1   8266-8365      * |     FALSE  6.231860           d
##     [5]     chr1   4298-4397      * |     FALSE  6.955316           c
##      ...      ...         ...    ... .       ...       ...         ...
##   [496]     chr1   2443-2542      * |     FALSE   1.12276           b
##   [497]     chr1   2455-2554      * |     FALSE   3.38518           c
##   [498]     chr1   1285-1384      * |     FALSE   1.58546           c
##   [499]     chr1 10137-10236      * |     FALSE   9.39272           c
##   [500]     chr1   6119-6218      * |     FALSE  10.22412           c
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

The resulting `MatchedGRanges` object is a set of null hypothesis ranges selected from our pool of options that is the same length as our input focal ranges and matched for `covar` features 2 and 3. These matched ranges print and behave as regular `GRanges` object. For example, we can use the sort method that is defined for a `GRanges` object on this `MatchedGRanges` object:

```
10  library(GenomicRanges)
11  sort(mgr)
```

```
## MatchedGRanges object with 500 ranges and 3 metadata columns:
##         seqnames      ranges strand |  feature1  feature2    feature3
##            <Rle>   <IRanges>  <Rle> | <logical> <numeric> <character>
##     [1]     chr1     511-610      * |     FALSE  5.545186           c
##     [2]     chr1     513-612      * |     FALSE  2.221684           b
##     [3]     chr1     534-633      * |     FALSE  1.563458           b
##     [4]     chr1     565-664      * |     FALSE  0.932659           c
##     [5]     chr1     577-676      * |     FALSE  3.256908           c
##      ...      ...         ...    ... .       ...       ...         ...
##   [496]     chr1 10377-10476      * |     FALSE  0.795032           c
##   [497]     chr1 10380-10479      * |     FALSE  0.977984           b
##   [498]     chr1 10409-10508      * |     FALSE  3.662119           c
##   [499]     chr1 10455-10554      * |     FALSE  6.815473           c
##   [500]     chr1 10483-10582      * |     FALSE  3.724147           c
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

The type argument of `makeExampleMatchedDataSet()` can be changed to generate `data.frames`, `data.tables`, `DataFrames`, `GRanges` and `GInteractions` objects - all of which work as inputs for `matchRanges()`. These produce either `MatchedDataFrame`, `MatchedGRanges`, or

`MatchedGInteractions` objects. More information about the `Matched` class structure and available methods is available in the *Implementation Details* section below or in the help documentation for each class accessible by entering `?MatchedDataFrame`, `?MatchedGRanges`, or `?MatchedGInteractions` in an R terminal.

3.3.4. Accessing matched data

Several accessor functions are included to allow users to easily access data contained in the `Matched` objects. All covariate and propensity score data can be extracted by set with the `matchedData()` function:

```
matchedData(mgr)
```

```
##        id feature2 feature3         ps       set
##     1:  1 2.879049        c 0.21095908     focal
##     2:  1 3.539645        c 0.19210984     focal
##     3:  1 7.117417        c 0.11193396     focal
##     4:  1 4.141017        a 0.01771986     focal
##     5:  1 4.258575        c 0.17308581     focal
##    ---
## 20496:  0 1.235781        b 0.08945367 unmatched
## 20497:  0 1.696712        a 0.02707977 unmatched
## 20498:  0 6.111404        a 0.01255772 unmatched
## 20499:  0 2.216575        d 0.07578989 unmatched
## 20500:  0 5.330029        b 0.04535856 unmatched
```

Attributes of the matched set:

```
covariates(mgr)
method(mgr)
withReplacement(mgr)
```

```
## [1] "feature2" "feature3"
## [1] "rejection"
## [1] FALSE
```

The `GRanges` object for each set can be extracted with the following functions:

```
summary(focal(mgr))
summary(pool(mgr))
summary(matched(mgr))
summary(unmatched(mgr))
```

```
## [1] "GRanges object with 500 ranges and 3 metadata columns"
## [1] "GRanges object with 10000 ranges and 3 metadata columns"
## [1] "GRanges object with 500 ranges and 3 metadata columns"
## [1] "GRanges object with 9500 ranges and 3 metadata columns"
```

The indices for each set can be obtained with the `indices()` function. For example, `indices(x, set="matched")` supplies the indices from the pool set that corresponds to the matched set. In fact, `matched(x)` is a convenient wrapper around `pool(x)[indices(x, set='matched')]`:

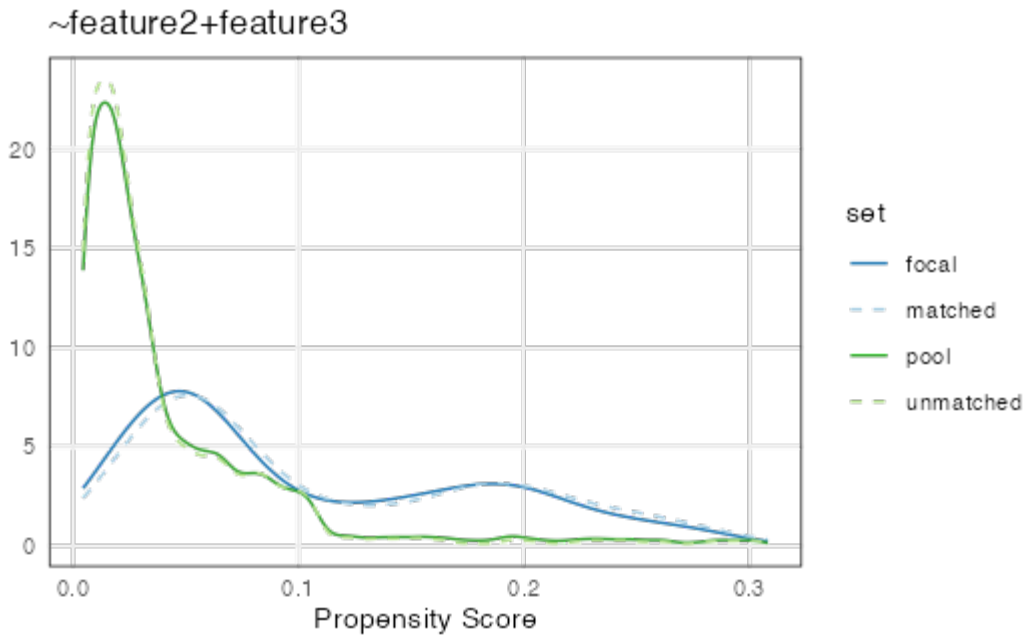
```
20  identical(matched(mgr), pool(mgr)[indices(mgr, set = 'matched')])
```

```
## [1] TRUE
```

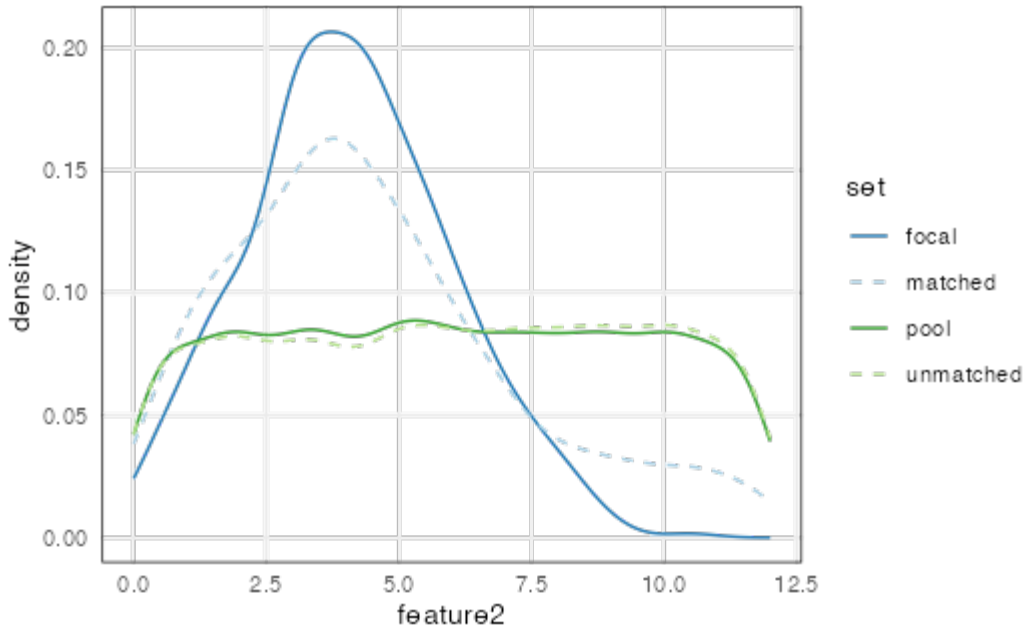
3.3.5. Assessing covariate balance

An important part of propensity-score matching is assessing similarity, or balance, between covariates of the focal and matched sets (Ho et al. 2007; Ali et al. 2015; Greifer 2020). One way to do this is to visually examine the distributions of covariates among sets. The `Matched` class provides `plotPropensity()` and `plotCovariate()` for visualizing propensity score and covariate distributions, respectively.

```
21  plotPropensity(mgr)
```

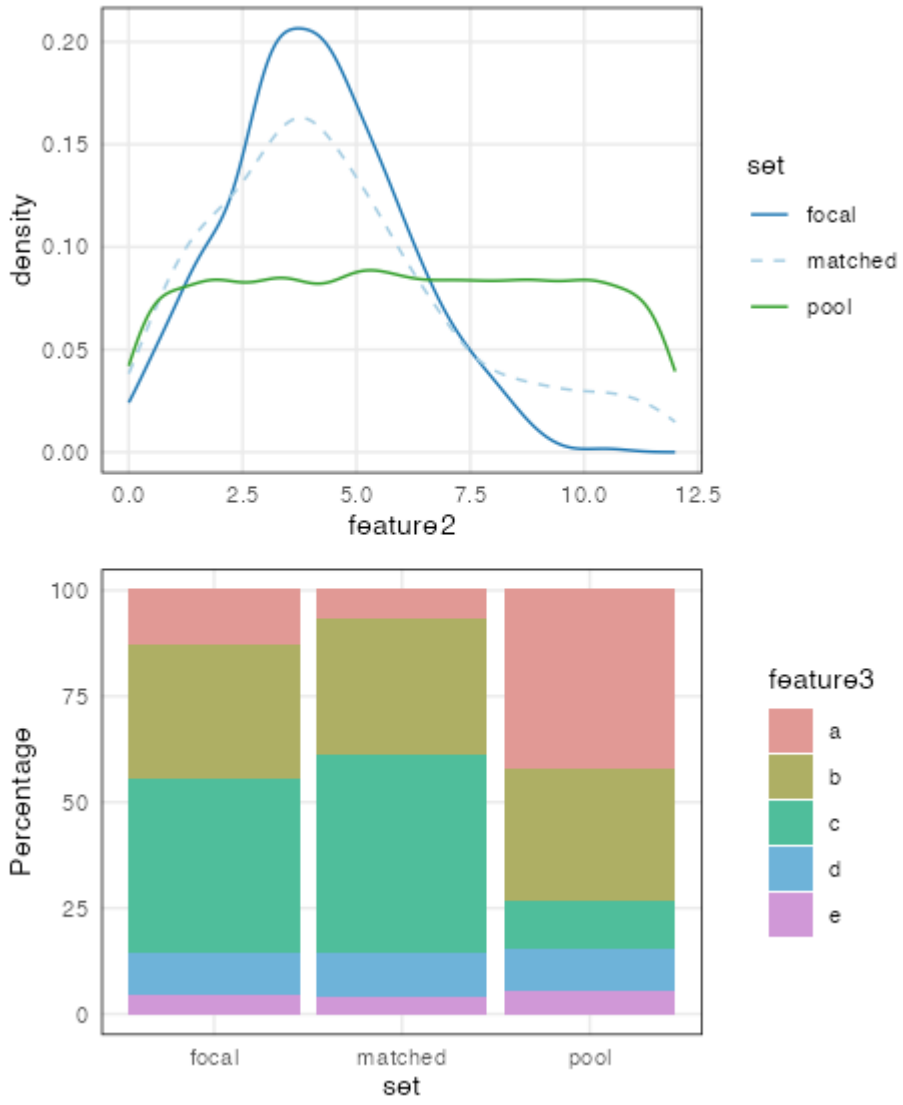


```
plotCovariate(mgr)
```



Since these functions return `ggplot` objects, the *patchwork* R package can be used to visualize all covariates together:

```
library(patchwork)
plots <- lapply(covariates(mgr), plotCovariate, x=mgr, sets = c('f', 'm', 'p'))
Reduce('/', plots)
```



By default, continuous features are plotted as density line plots while categorical features are plotted as stacked bar plots. All sets are also shown by default but can be overridden by setting the “type” and “sets” arguments.

The `overview()` function provides a summary of covariate distributions before and after matching by reporting the mean and standard deviation for covariates and propensity scores of the focal, pool, matched, and unmatched sets. For factor, character, or logical covariates (e.g., categorical covariates) the N per set (frequency) is returned. The mean difference between focal and matched sets is also reported.

```
26  overview(mgr)
```

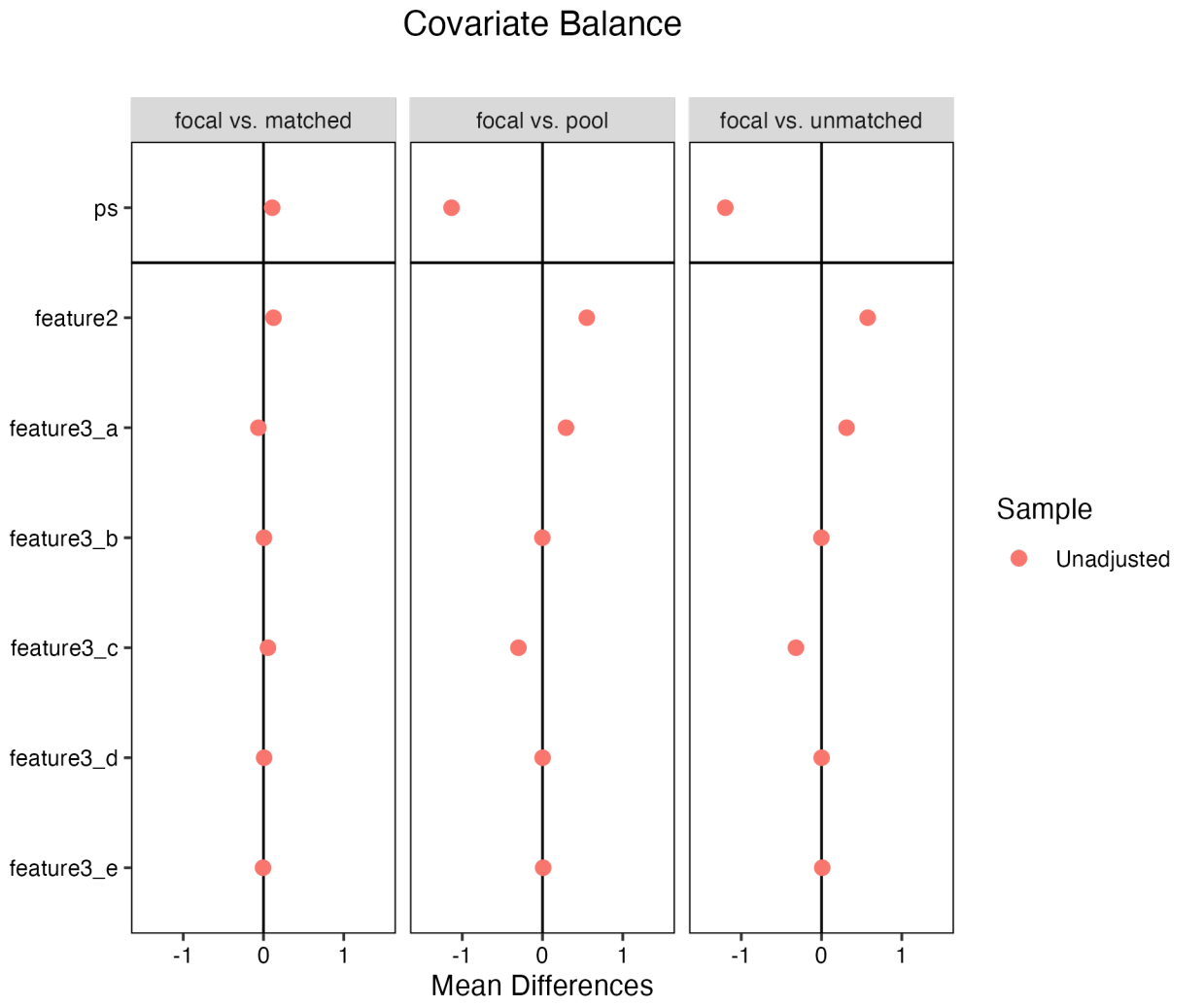
```
## MatchedGRanges object:
##        set     N feature2.mean feature2.sd feature3.a feature3.b feature3.c
##      focal   500           4.1         1.9         66        157        206
##    matched   500           4.5         2.7         34        160        234
##       pool 10000           6.0         3.4       4248       3121       1117
##  unmatched  9500           6.1         3.5       4214       2961        883
##  feature3.d feature3.e ps.mean ps.sd
##          49         22   0.100 0.076
##          53         19   0.110 0.078
##         992        522   0.045 0.051
##         939        503   0.041 0.047
## --------
## focal - matched:
##  feature2.mean feature2.sd feature3.a feature3.b feature3.c feature3.d
##          -0.42       -0.84         32         -3        -28         -4
##  feature3.e ps.mean   ps.sd
##           3 -0.0057 -0.0019
```

While an in-depth assessment of covariate balance is outside the scope of this package, the R package *cobalt* includes detailed documentation on this topic (see `vignette("cobalt", package="cobalt")` in R). Below, we demonstrate how to use *matchRanges* and *cobalt* (v4.4.1) to calculate the standardized mean differences and visualize these statistics with a love plot.

```
library(cobalt)
res <- bal.tab(f.build("set", covariates(mgr)),
               data = matchedData(mgr),
               distance = "ps", # name of column containing propensity score
               focal = "focal", # name of focal group in set column
               which.treat = "focal", # compare everything to focal
               s.d.denom = "all") # how to adjust standard deviation
res
```

```
## Balance by treatment pair
##
##  - - - focal (0) vs. matched (1) - - -
## Balance Measures
##                Type Diff.Un
## ps         Distance  0.1088
## feature2    Contin.  0.1242
## feature3_a   Binary -0.0640
## feature3_b   Binary  0.0060
## feature3_c   Binary  0.0560
## feature3_d   Binary  0.0080
## feature3_e   Binary -0.0060
##
## Sample sizes
##     focal matched
## All   500     500
##
##  - - - focal (0) vs. pool (1) - - -
## Balance Measures
##                Type Diff.Un
## ps         Distance -1.1340
## feature2    Contin.  0.5520
## feature3_a   Binary  0.2928
## feature3_b   Binary -0.0019
## feature3_c   Binary -0.3003
## feature3_d   Binary  0.0012
## feature3_e   Binary  0.0082
##
## Sample sizes
##     focal  pool
## All   500 10000
##
##  - - - focal (0) vs. unmatched (1) - - -
## Balance Measures
##                Type Diff.Un
## ps         Distance -1.1994
## feature2    Contin.  0.5745
## feature3_a   Binary  0.3116
## feature3_b   Binary -0.0023
## feature3_c   Binary -0.3191
## feature3_d   Binary  0.0008
## feature3_e   Binary  0.0089
##
## Sample sizes
##     focal unmatched
## All   500      9500
##  - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

```r
library(ggplot2)
love.plot(res) + xlim(c(-1.5, 1.5))
```

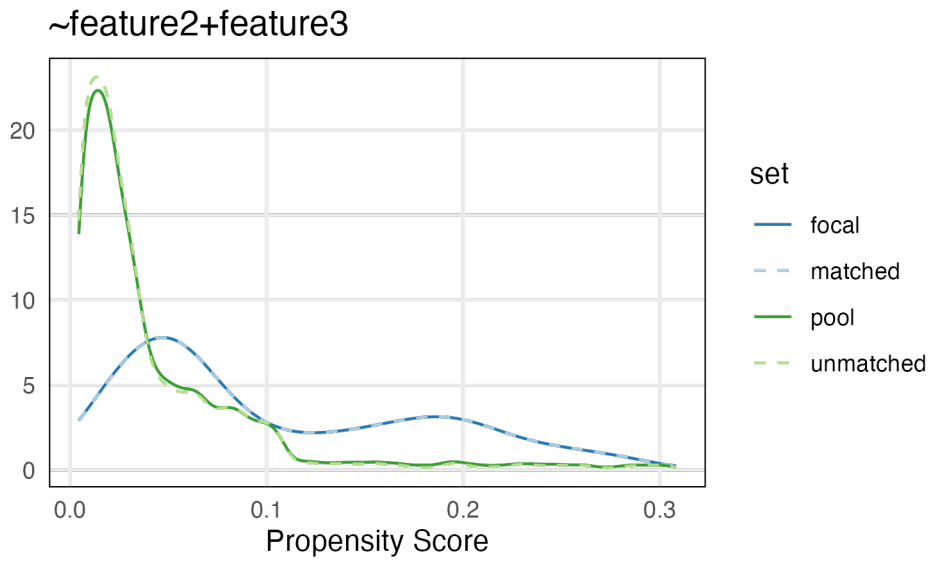


3.3.6. Choosing the method parameter

There are currently 3 methods available for selecting a matched set: Nearest-neighbor matching with replacement, Rejection sampling with/without replacement, and Stratified sampling with/without replacement. Currently, nearest-neighbor matching without replacement is not implemented in *matchRanges*, but stratified sampling without replacement is a suitable substitute.

Nearest-neighbor matching uses rolling-join method from the *data.table* package to find the closest value between propensity scores of the focal and pool sets. The following code and plot show how to use `method="nearest"` and visualize the resulting propensity scores:

```r
set.seed(123)
mgr <- matchRanges(focal = x[x$feature1],
                   pool = x[!x$feature1],
                   covar = ~feature2 + feature3,
                   method = 'nearest',
                   replace = TRUE)
nn <- overview(mgr)
plotPropensity(mgr)
```



The nearest method is best for very large datasets because it is the fastest matching method **(Fig. 3.S1)**. However, because sampling is done with replacement the user should be careful to assess the number of duplicate ranges pulled, especially among strata with few candidate ranges. This can be done using the `indices()` function, as shown below, to ensure there is limited re-use of ranges.

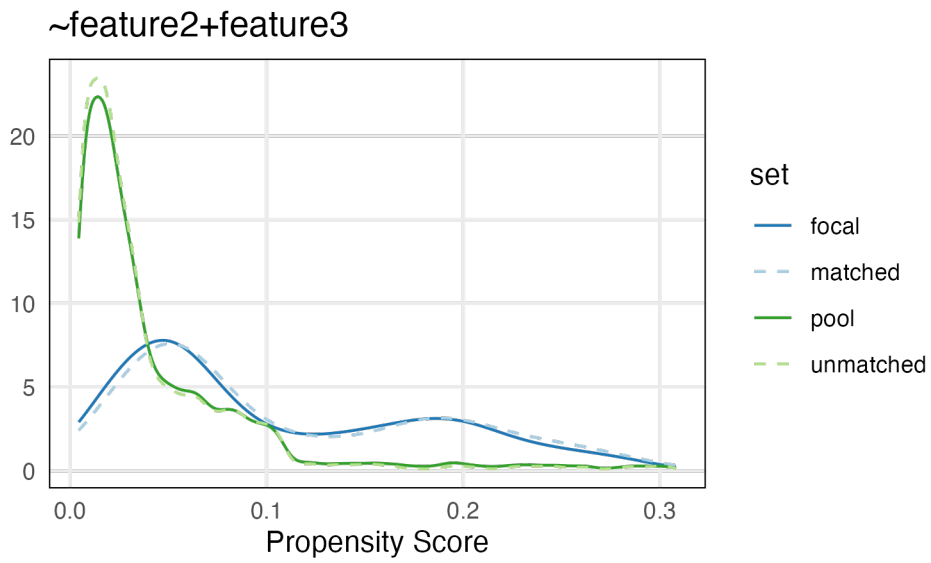
```r
## Total number of duplicated indices
length(which(duplicated(indices(mgr))))

sum(table(indices(mgr)) > 1) # used more than once
sum(table(indices(mgr)) > 2) # used more than twice
sum(table(indices(mgr)) > 3) # used more than thrice
```

```
## [1] 59
## [1] 51
## [1] 8
## [1] 0
```

Rejection sampling uses a probability-based approach to select options in the pool that distributionally match the focal set based on their propensity scores. Briefly, the rejection sampling method first generates kernel-density estimates for both the focal and pool sets. Next, a scale factor is determined by finding the point at which the difference in focal and pool densities is maximized. This scale factor is applied such that the pool distribution covers the focal distribution at all points. Random sampling is conducted, with the probability of accepting a pool range into the matched set given by the ratio between the height of the density and the scaled (covering) density. The following code and plot show how to use `method="rejection"` and visualize the resulting propensity scores:

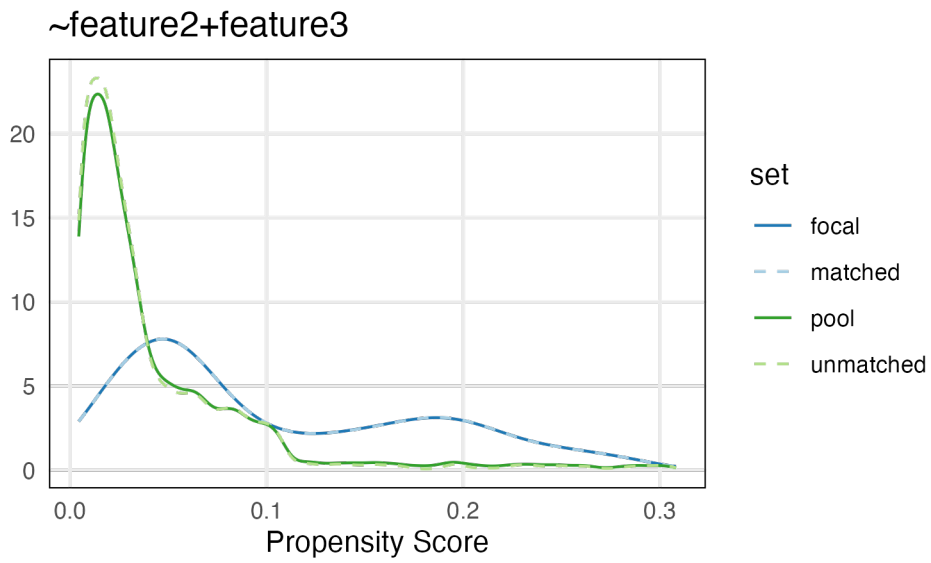
```
set.seed(123)
mgr <- matchRanges(focal = x[x$feature1],
                   pool = x[!x$feature1],
                   covar = ~feature2 + feature3,
                   method = 'rejection',
                   replace = FALSE)
rs <- overview(mgr)
plotPropensity(mgr)
```



Rejection sampling is the fastest available matching method for sampling without replacement **(Fig. 3.S1)**. Therefore, it is ideal to use on large datasets when sampling without replacement is important. However, when the density of the pool that covers the focal set is low then probability-based sampling is poor which results in instability. In those cases, the best alternative method to use is stratified sampling.

The stratified matching method performs iterative sampling on increasingly large bins of data. Briefly, focal and pool propensity scores are binned by their value with high granularity; options are randomly selected (with or without replacement) within each bin and subsequently removed from the pool of available options. This procedure is repeated, decreasing the number of bins (and increasing bin size) until the number of selected matches is equal to the focal set. While matches are being found in each bin the bins stay small. However, as the number of bins with no matches increases the algorithm expands bin size faster, which maintains matching quality while decreasing run time. The following code and plot show how to use `method="stratified"` and visualize the resulting propensity scores:

```
set.seed(123)
mgr <- matchRanges(focal = x[x$feature1],
                   pool = x[!x$feature1],
                   covar = ~feature2 + feature3,
                   method = 'stratified',
                   replace = FALSE)
ss <- overview(mgr)
plotPropensity(mgr)
```



For very large data sets this method produces longer run times compared to the other methods **(Fig. 3.S1)**. However, compared to nearest-neighbor matching with *MatchIt*, the stratified method runs faster and with similar results in terms of covariate balance **(Fig. 3.S1 and Fig. 3.S3)**. The stratified method tends to work very well for discrete data, and often produces the best matches even on continuous data. For example, the code below extracts the "quality" score produced by the "overview" function for each

matching method used above (lines 43, 57, and 65) to show that stratified produces the smallest difference between focal and matched sets.

```
67  fmps <- sapply(c(nn, rs, ss), `[[`, "quality")
68  c('nearest', 'rejection', 'stratified')[which.min(fmps)]
```

```
    ## [1] "stratified"
```

3.3.7. Implementation details

This section briefly discusses the implementation structure of the *matchRanges* class structure for future developers of this package or other packages that extend the `GRanges` or `GInteractions` Bioconductor classes. `matchRanges()` acts as a constructor, combining a `Matched` superclass - which contains the matching results - with either a `DataFrame` (or `data.frame`/`data.table`), `GRanges`, or `GInteractions` superclass. This results in the `MatchedDataFrame`, `MatchedGRanges`, or `MatchedGInteractions` subclasses. Internally, each `Matched` subclass uses a "delegate" object of the same type to assign its slots. The delegate object used is the matched set. Therefore, the resulting `Matched*` object behaves as a combination of both its superclasses and has access to methods from both. For example, using `matchRanges()` on `GRanges` objects assigns a `GRanges` delegate object which is used to populate `GRanges`-specific slots. This results in a `MatchedGRanges` object, with access to both Matched functions (e.g., `plotCovariate`) as well as normal `GRanges` methods (e.g., `seqnames`, `resize`, `sort`, etc.). **Figure 3.S2** shows a diagram outlining this implementation.

3.4. Conclusion

matchRanges is a collection of R functions for generating covariate matched ranges to test associations between sets of genomic ranges. Distributed as part of the *nullranges* R package, *matchRanges* uses a propensity score-based method to perform subset selection on genomic ranges, allowing fair comparisons between two sets of interest while avoiding problems with confounding by nuisance covariates. The package provides functions for assessing, visualizing, and extracting matched data that integrates seamlessly into existing Bioconductor workflows. *matchRanges* will be useful to genomic researchers from all disciplines and will help accelerate scientific progress by improving the accuracy and rigor of genomic analyses.

3.5. Supplementary Figures

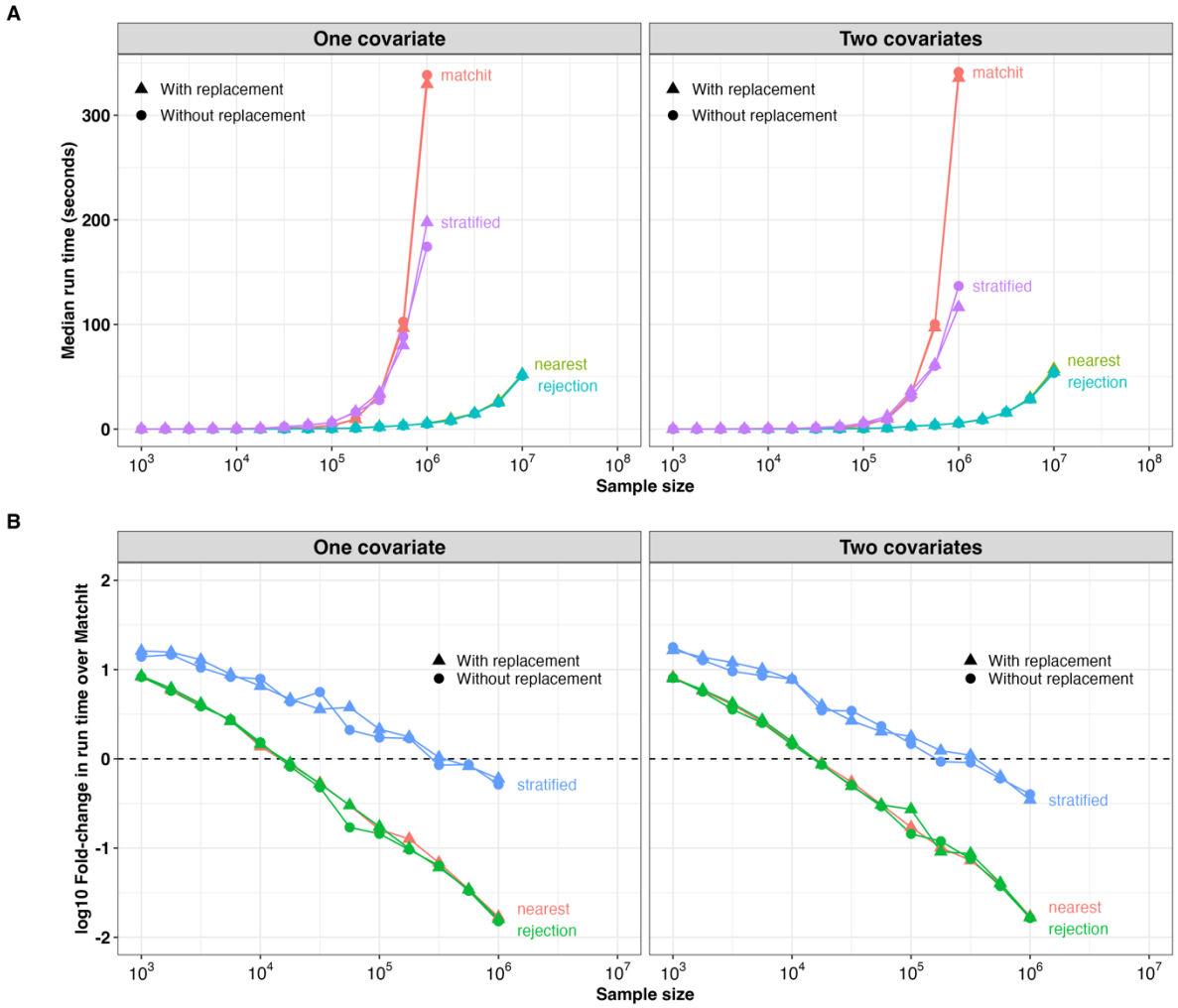


Figure 3.S1. matchRanges run time. Runtime analysis for *matchRanges and MatchIt* applied to simulated data. Data were matched for one or two continuous features for each matching method and replacement option. Sample size contains 95% values as pool and 5% as focal. *MatchIt* (v4.5.0) was run with default parameters using method="nearest". Benchmarking was performed 10 times using the *microbenchmark* R package on a single core machine (2.50 GHz Intel processor, 16Gb of memory). **(A)** Median runtime in seconds for evenly spaced datasets spanning 10^3 to 10^7 in size. Runtimes for "matchit" and "stratified" exceeded 90 and 5 minutes, respectively, at sample sizes of $10^{6.25}$ and greater. **(B)** Log10 fold-change in median runtime for each matching method over "matchit" (method="nearest", with and without replacement).

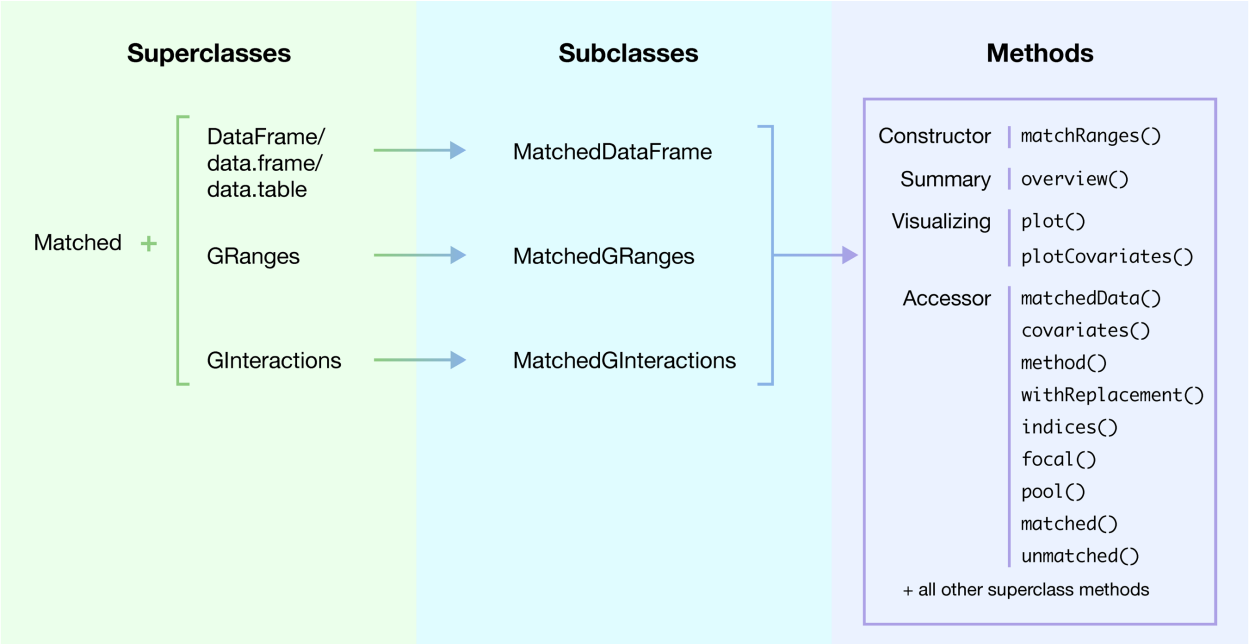


Figure 3.S2. matchRanges class structure. Overview of the *matchRanges* class structure and methods. The *Matched* class is combined with either the *DataFrame, data.frame, data.table, GRanges,* or *GInteractions* classes (*left panel*) to create the *MatchedDataFrame*, *MatchedGRanges*, or *MatchedGInteractions* subclasses (*middle panel*). Each subclass behaves as a combination of both its superclasses - with access to both methods of the *Matched* class (*right panel*) and each respective class' methods.

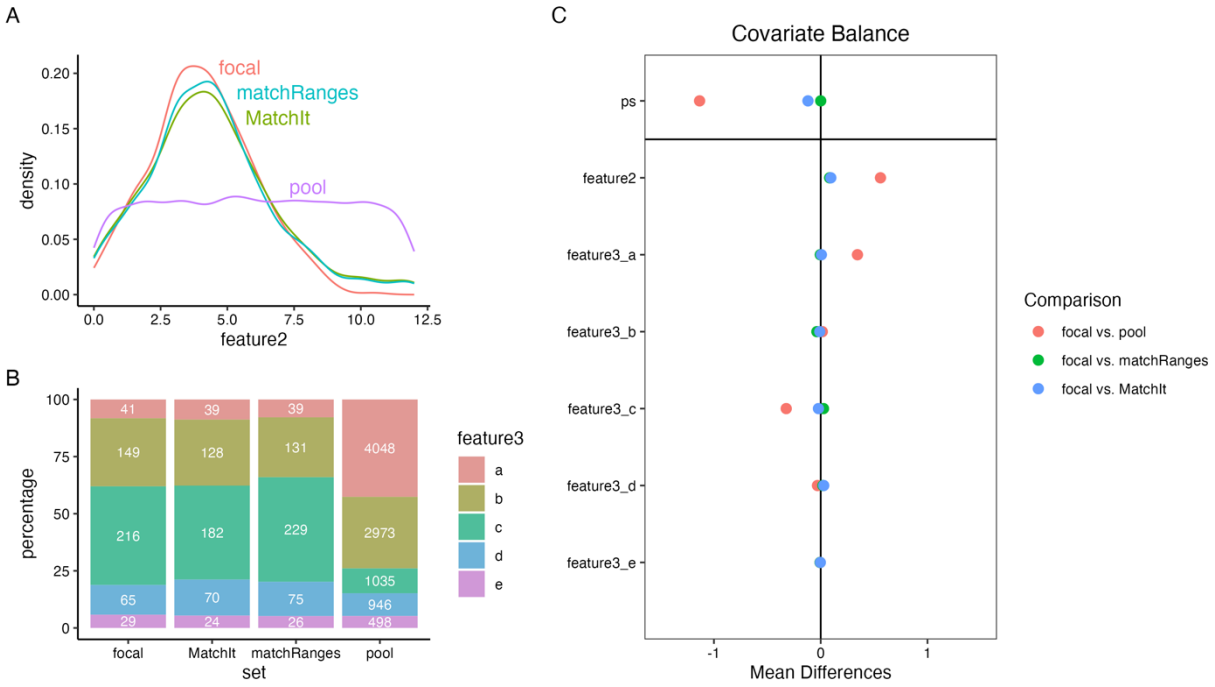


Figure 3.S3. Assessing covariate balance with matchRanges and cobalt. A simulated dataset containing 1e4 data points (500 focal, 9500 pool) was matched for continuous (feature2) and discrete (feature3) covariates with *MatchIt* and *matchRanges* using 'nearest' matching with replacement. **(A)** Density plots comparing distributions of the continuous "feature2" among the focal, pool, *MatchIt*-matched (n=443), and *matchRanges*-matched (n=500) sets. **(B)** Stacked bar plots comparing proportions of the discrete "feature3" among the focal, pool, *MatchIt*-matched (n=443), and *matchRanges*-matched (n=500) sets. (C) Love plot assessing covariate balance by comparing the mean standardized differences ("ps" and "feature2") or mean differences ("feature3" strata) when comparing the unadjusted (focal vs. pool) to adjusted (focal vs. *matchRanges* and focal vs. *MatchIt*) sets. "ps" denotes the propensity scores. Values calculated with the "bal.tab()" function from *cobalt*.

Chapter 4: Mariner: Explore the Hi-Cs

4.1. Introduction

3D genome organization plays an important role in the regulation of gene expression during human development and disease. Chromatin features such as loops, topologically associating domains (TADs), and compartments shape the regulatory landscape by bringing actively transcribed regions into proximity with their linearly distant regulators. However, the formation of these structures and their impact on transcription is not completely understood. This gap in knowledge is due, in part, to insufficient tools to analyze 3D chromatin data.

Hi-C and Micro-C are chromosome conformation capture methods that produce a snapshot of the combined 3D chromatin structure of a cell population. Many tools exist for processing these data into sparse contact frequency matrices and identifying chromatin features (Durand, Shamim, et al. 2016; Servant et al. 2015; Davis 2023). Several studies have used Hi-C to map changes in chromatin structure during cellular transitions and following experimental perturbations to explore the mechanisms of chromatin feature formation and gene regulation (Phanstiel et al. 2017; Ahn et al. 2021; Kathleen S. M. Reed et al. 2022).

To address these questions biologists must query contact maps, extract submatrices, and perform a host of custom analyses such as differential and pileup analyses. Until very recently, few software tools existed for performing these tasks. While several tools have been developed in the past year to address this gap, many lack the flexibility necessary for analyzing Hi-C data (Flyamer, Illingworth, and Bickmore 2020; Sahin et al. 2021; Open2C et al. 2022; Chang et al. 2022). Furthermore, these tools lack a unified framework, and require knowledge of multiple programming languages.

Here we introduce *mariner*, a complete suite of tools for exploring Hi-C data in R. *mariner* combines existing and novel functionality into an efficient and easy to use Bioconductor package. Bioconductor provides an infrastructure of classes for genomic data types which allows interoperability between software packages (Gentleman et al. 2004). *mariner* extends the Bioconductor infrastructure with classes for storing Hi-C submatrices with hierarchical data format 5 (HDF5) and block-processing methods for efficiently

operating on them. These tools are completely flexible and highly modular, enabling full customization of analyses and facilitating extension by future developers. *mariner* forms a software ecosystem with several existing Bioconductor packages which enables the complete analysis and visualization of Hi-C data without leaving R. *mariner's* vignettes demonstrate how its modular functions can be combined into workflows for tasks such as differential and pileup analyses of chromatin features. Together, these tools will empower biologists to explore Hi-C data to uncover the mysteries of chromatin features and how they impact gene regulation.

4.2. Key Features

Processed Hi-C data is stored in compressed `.hic` files and contains the contact frequency between all pairs of genomic regions, genome-wide. Chromatin loops are a feature of 3D chromatin structure that represent paired genomic regions with enriched contact frequency. The *InteractionSet* Bioconductor package provides classes, or data structures, for representing paired-range data like chromatin loops or Hi-C contacts (Lun, Perry, and Ing-Simmons 2016). *InteractionSet* enables common operations such as overlapping, linking, and computing on paired ranges. *mariner* extends this infrastructure with functions for more flexible manipulation of paired ranges and additional classes for storing interaction data extracted from `.hic` files.

mariner provides several key functions for manipulating, clustering, and merging paired ranges. The *GInteractions* class from *InteractionSet* stores the chromosome, start, end, and strand of each pair of genomic ranges along with any associated metadata. *mariner* provides the `as_ginteractions()` function for converting to this class from the more general *data.frame* class in R. This is particularly helpful when importing BEDPE files with chromatin features such as loop calls. Additionally, the `seqnames1()`, `start1()`, `end1()`, `seqnames2()`, `start2()`, and `end2()` functions provides convenient access to chromosomes, starts, and ends for each pair in *GInteractions* objects.

Another common task with paired-range data is to assign them to evenly spaced genomic bins. The `binPairs()`, `binRanges()`, and `snapToBins()` functions allow users to assign paired-range or single-range data to a desired bin size or "snap" them to their nearest bin **(Fig. 4.1)**. The `shiftRanges()` function improves the generic `shift()` function by making it strand-aware for accurate up or downstream shifting for stranded ranges.

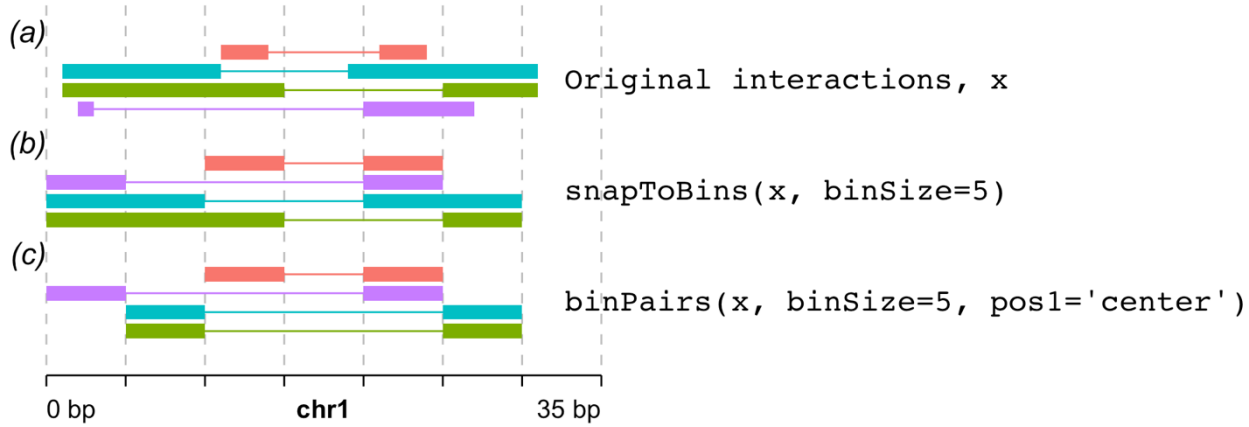


Figure 4.1. Binning *GInteractions* objects with *mariner* functions. Each pair of a *GInteractions* object was transformed by adjusting the starts and ends into 5 base-pair (bp) bins (vertical dashed lines denote bin boundaries). The color corresponds to the same pair before and after each transformation. Visualization created with *plotgardener* (Kramer et al. 2022) *(a)* Interactions before transformation, denoted as "x". *(b)* Interactions transformed with "snapToBins" function. Starts and ends for each pair of ranges are "snapped" to their nearest bin boundary. *(c)* Interactions transformed with "binPairs" function. The "pos1" and "pos2" (not shown) arguments determine which point along the first and second ranges to use for assigning into bins. The default value is "center". Ranges larger than a bin are contracted to the bin containing the center of the range while ranges smaller than a bin are expanded.

When chromatin loops are identified in different cell types or conditions, it is often necessary to merge them together for differential analysis. *mariner* adds the `mergePairs()` function which uses the DBscan algorithm (Hahsler, Piekenbrock, and Doran 2019) to cluster and merges lists of *GInteractions* objects, such as loops **(Fig. 4.2)**. The result is a *MergedGInteractions* object which contains the final merged set of *GInteractions* and information about the clusters. Several accessor functions can be used to aggregate metadata columns of clustered interactions (`aggPairMcols`), return interactions specific to each cluster of interactions (`getPairClusters`), or return the interactions unique to each data source or combination of sources (`subsetBySource`). These functions allow users to distinguish *de novo*, transient, and static loops from their data.

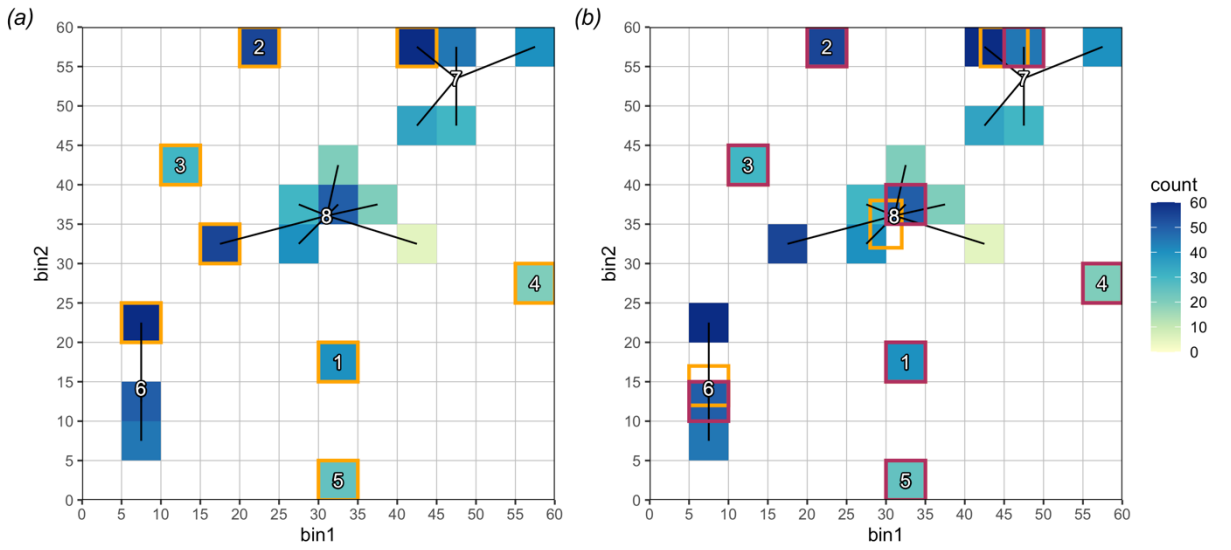


Figure 4.2. Clustering and merging interactions. Simulated dataset of 20 randomly generated interactions on a 60x60 unit grid divided into 5 unit bins. Each interaction is colored by its "count" feature. Numbers indicate the cluster membership after running the "mergePairs" function. *(a)* Orange outlines represent the selected interactions after running "mergePairs" with `radius=10`, `method="manhattan"`, `column="count"`, `selectMax=TRUE`, and `pos="center"`. This results in selecting the interaction with the highest "count" in each cluster. *(b)* Orange outlines represent the selected interactions after running "mergePairs" with `radius=10`, `method="manhattan"`, and `pos="center"`. When the "column" argument is not provided interactions are selected by calculating the mean of the mode for each pair of ranges. This can result in interactions falling outside of bin boundaries. The maroon outlines represent the result after using "binPairs" with `binSize=5` to assign interactions to bins.

A crucial question in biology is how often two genomic regions, such as enhancers and promoters, anchors of chromatin loops, or binding sites of proteins, contact each other. While Hi-C data contains this information, it can be difficult to extract efficiently and store for flexible downstream analysis. *mariner* adds two generic functions, "pullHicPixels" and "pullHicMatrices", for extracting interaction data from `.hic` files. Pixels are defined as interactions that span exactly one bin. The "pullHicPixels" function extracts pixels from a list of provided `.hic` files and returns an *InteractionMatrix* object containing a *DelayedMatrix* of Hi-C interactions (rows) for each `.hic` file (columns) **(Fig. 4.3a)**. The "pullHicMatrices" function is used to extract spans of pixels (i.e., submatrices) and returns an *InteractionArray* object containing a 4-dimensional *DelayedArray* of submatrices for each interaction range and `.hic` file **(Fig. 4.3b)**.

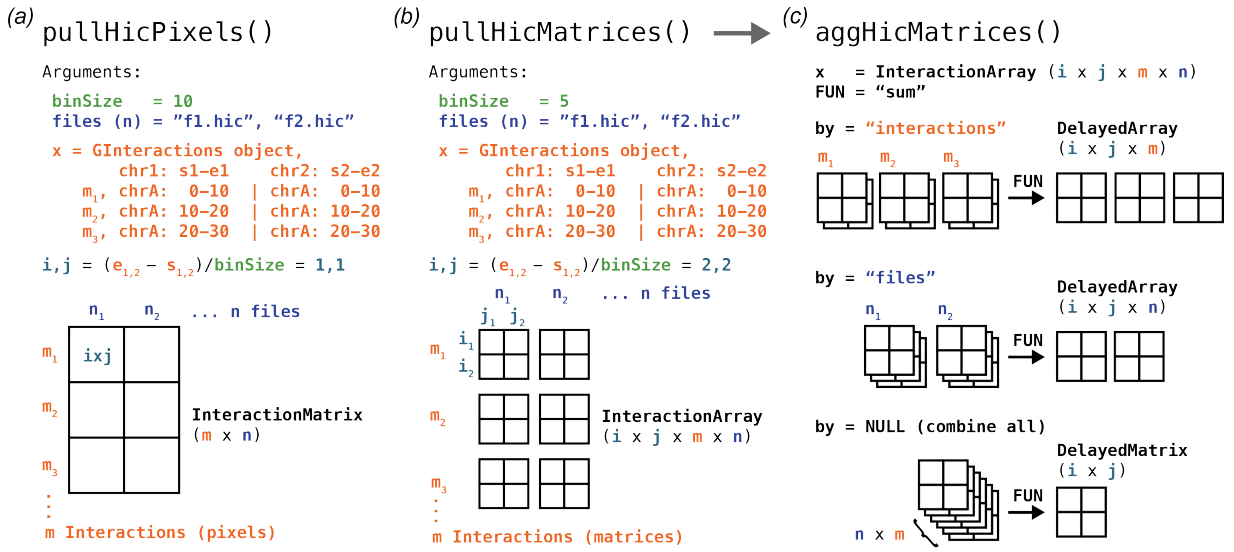


Figure 4.3. Overview of pullHic and aggHic functions. *(a)* An overview of the "pullHicPixels" function using a *GInteractions* object (x) with three interactions (m_{1-3}), two `.hic` files (n_{1-2}), and "binSize" of 10 units. The "pullHicPixels" function is used when both pairs (chr1: s1-e1 and chr2: s2-e2) have a width equal to "binSize" (i, j) which defines a pixel interaction. The result is an *InteractionMatrix* with dimensions of (m, n). *(b)* An overview of the "pullHicMatrices" function using a *GInteractions* object (x) with three interactions (m_{1-3}), two `.hic` files (n_{1-2}), and "binSize" of 5 units. The "pullHicMatrices" function is used when one or both pairs (chr1: s1-e1 and chr2: s2-e2) have a width less than "binSize" (i, j) which define the rows (i) and columns (j) of the extracted matrices. The result is an *InteractionArray* with dimensions of (i, j, m, n). *(c)* An overview of the "aggHicMatrices" function using the example *InteractionArray* from *(b)* along with the summary "FUN = 'sum'" and different values of "by". When "by = 'interactions'", matrices are aggregated across Hi-C files, producing an (i, j) matrix for each interaction (m). This results in a *DelayedArray* of dimensions (i, j, m). When "by = 'files'", matrices are aggregated across interactions, producing an (i, j) matrix for each Hi-C file (n). This results in a *DelayedArray* of dimensions (i, j, n). When "by = NULL", matrices are aggregated across interactions (m) and Hi-C files (n), producing a single *DelayedMatrix* of dimensions (i, j).

These Hi-C extraction functions use HDF5 to store interactions directly on-disk, allowing users to extract much larger data than can be held in computer memory. Furthermore, the *HDF5Array* and *DelayedArray* Bioconductor packages used in *mariner* allow users to interact with these contact matrices as if they were ordinary R matrices held in memory (Hervé Pagès 2020; H. Pagès, Hickey, and Lun 2021). *mariner* also provides an "aggregate" function for applying summary or custom functions across .hic files, interactions, or subsets of interactions **(Fig. 4.3c)**. This function supports block and parallel processing to allow users to perform aggregation on large data quickly. These tools allow users to perform aggregate peak analysis, use Hi-C signal to characterize chromatin features, and identify differential interactions between conditions.

4.3. Workflows

The following sections demonstrate a few of the possible workflows that can be conducted with *mariner*. First we demonstrate how *mariner* allows flexibly clustering and merging of paired features by identifying *de novo* and transient looping using looping data from a eight-point time course of macrophage activation (Reed et al. 2022). Next we demonstrate Hi-C pixel extraction with *mariner* and *DESeq2* to show how differential interactions can be identified from looping and Hi-C data using a phase-separated chromatin looping study (Love, Huber, and Anders 2014; Ahn et al. 2021). Finally, with the same study we show Hi-C matrix extraction, aggregation, and visualization with *mariner* and *plotgardener* to visualize genome-wide trends in a pileup analysis (Kramer et al. 2022).

4.3.1. Identifying transient and *de novo* chromatin looping

Reed et al. performed a Hi-C time course to understand changes in 3D chromatin structure during macrophage activation. In short, human THP-1 macrophage cell lines were treated with lipopolysaccharide (LPS) and interferon-gamma (IFNγ) for 0, 30, 60, 90, 120, 240, 360, or 1440 minutes before performing Hi-C and other genomic assays (Reed et al. 2022). For each of the eight time points, loops were identified with Significant Interaction Peak (SIP) caller (version 1.6.1) at 5-Kb resolution (Rowley et al. 2020). Loop files are accessible at the Gene Expression Omnibus (GEO) under the SuperSeries, “GSE201376”. The following section demonstrates *mariner’s* functions for clustering and merging interactions to identify transient and *de novo* chromatin looping in this macrophage activation dataset.

To begin, loop calls for each time point are read into R. mariner includes these BEDPE files with the package so they can be loaded with the `system.file()` command as shown below:

```
library(mariner)
loopFiles <-
    system.file("extdata/lima_loops",
                package = "mariner") |>
    list.files(full.names = TRUE)

basename(loopFiles)
```

```
## [1] "LIMA_0000.bedpe" "LIMA_0030.bedpe" "LIMA_0060.bedpe" "LIMA_0090.bedpe"
## [5] "LIMA_0120.bedpe" "LIMA_0240.bedpe" "LIMA_0360.bedpe" "LIMA_1440.bedpe"
```

Each file is read in, converted to a *GInteractions* object, and named with its corresponding time point to create a list of *GInteractions* objects.

```
library(data.table)
giList <-
    lapply(loopFiles, fread) |>
    lapply(as_ginteractions) |>
    setNames(gsub(".*LIMA_([0-9]+).*", "\\1", loopFiles))

lapply(giList, summary)
```

```
## $`0000`
## [1] "GInteractions object of length 17401 with 9 metadata columns"
##
## $`0030`
## [1] "GInteractions object of length 15221 with 9 metadata columns"
##
## $`0060`
## [1] "GInteractions object of length 15895 with 9 metadata columns"
##
## $`0090`
## [1] "GInteractions object of length 15275 with 9 metadata columns"
##
## $`0120`
## [1] "GInteractions object of length 15489 with 9 metadata columns"
##
## $`0240`
## [1] "GInteractions object of length 16636 with 9 metadata columns"
##
## $`0360`
## [1] "GInteractions object of length 16189 with 9 metadata columns"
##
## $`1440`
## [1] "GInteractions object of length 17505 with 9 metadata columns"
```

Each file contains the following metadata columns:

```
library(S4Vectors)
colnames(mcols(giList[[1]]))
```

```
## [1] "color"                    "APScoreAvg"
## [3] "ProbabilityofEnrichment" "RegAPScoreAvg"
## [5] "Avg_diffMaxNeihgboor_1"  "Avg_diffMaxNeihgboor_2"
## [7] "avg"                      "std"
## [9] "value"
```

Next, the *GInteractions* list is passed to the "mergePairs" function to cluster and merge loops using a 10-Kb manhattan distance radius. This produces a *MergedGInteractions* object which extends the *GInteractions* class with information about the clusters of interactions.

```
mgi <- mergePairs(x = giList,
                  radius = 10e3,
                  method = "manhattan")
mgi
```

```
## MergedGInteractions object with 43736 interactions and 0 metadata columns:
##           seqnames1           ranges1     seqnames2           ranges2
##               <Rle>         <IRanges>         <Rle>         <IRanges>
##       [1]        22 30435000-30440000 ---        22 30475000-30480000
##       [2]        22 46955000-46960000 ---        22 47080000-47085000
##       [3]        22 18940000-18945000 ---        22 19625000-19630000
##       [4]        22 17105000-17110000 ---        22 17215000-17220000
##       [5]        22 19165000-19170000 ---        22 19625000-19630000
##       ...       ...               ... ...       ...               ...
##   [43732]        20 31485000-31490000 ---        20 31885000-31890000
##   [43733]        20 30692500-30697500 ---        20 31242500-31247500
##   [43734]        21 44772500-44777500 ---        21 44817500-44822500
##   [43735]        21 26412500-26417500 ---        21 26755000-26760000
##   [43736]        21 46500000-46505000 ---        21 46545000-46550000
##   -------
##   regions: 63774 ranges and 0 metadata columns
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

Since no "column" was provided, the default method calculates the mean of modes for each pair range to construct a new merged pair (see also **Fig. 4.2**).

```
selectionMethod(mgi)
```

```
## [1] "Mean of modes"
```

No metadata is shown because it is not clear which loop's metadata should represent the cluster. The "aggPairMcols" function allows users to aggregate and summarize the values of one or more specified

metadata columns which are then appended as metadata column of the *MergedGInteractions* object. For example, the following code calculates the mean of the “APScoreAvg” column for each cluster of loops.

```
aggPairMcols(mgi, columns = "APScoreAvg", funs = "mean")
```

```
## MergedGInteractions object with 43736 interactions and 1 metadata column:
##           seqnames1           ranges1     seqnames2           ranges2 |
##               <Rle>         <IRanges>         <Rle>         <IRanges> |
##       [1]        22 30435000-30440000 ---        22 30475000-30480000 |
##       [2]        22 46955000-46960000 ---        22 47080000-47085000 |
##       [3]        22 18940000-18945000 ---        22 19625000-19630000 |
##       [4]        22 17105000-17110000 ---        22 17215000-17220000 |
##       [5]        22 19165000-19170000 ---        22 19625000-19630000 |
##       ...       ...               ... ...       ...               ... .
##   [43732]        20 31485000-31490000 ---        20 31885000-31890000 |
##   [43733]        20 30692500-30697500 ---        20 31242500-31247500 |
##   [43734]        21 44772500-44777500 ---        21 44817500-44822500 |
##   [43735]        21 26412500-26417500 ---        21 26755000-26760000 |
##   [43736]        21 46500000-46505000 ---        21 46545000-46550000 |
##           mean.APScoreAvg
##                 <numeric>
##       [1]         1.96029
##       [2]         3.59065
##       [3]         4.23653
##       [4]         2.87979
##       [5]         2.52364
##       ...             ...
##   [43732]         3.15526
##   [43733]         2.68770
##   [43734]         1.78600
##   [43735]         3.17716
##   [43736]         2.27946
##   -------
##   regions: 63774 ranges and 0 metadata columns
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

An advantage to the clustering and merging approach is that it eliminates many redundant loops called in multiple datasets. For example, the following code chunk compares the combined number of loops found (n=129,611), the number of unique loops (n=92,790), and the number after using `mergePairs()` (n=43,736):

```r
## Number of all loops combined
vapply(giList, length, integer(1L)) |>
    sum()

## Number of unique loops
lapply(loopFiles, fread) |>
    lapply(as_ginteractions, keep.extra.columns = FALSE) |>
    do.call(what = c, args = _) |>
    unique() |>
    length()

## Number of loops from merging
length(mgi)
```

```
# [1] 129611
# [1] 92790
# [1] 43736
```

Taking the unique set of loops only removes duplicates when coordinates are exactly equal. However, due to either biological or technical variability in Hi-C data, loop callers may shift the exact loop pixel called in datasets. Therefore, clustering and merging is important for eliminating loops that are near duplicates. In this example, more than half the loops were reduced using `mergePairs()` compared with using unique coordinates.

The *MergedGInteractions* object and its "getPairClusters" accessor makes it easy to retrieve the loops belonging to each cluster. The following code shows an example of a merged loop that was present in all eight time points.

```r
i <- 22559
mgi[i] # Merged loop
```

```
## MergedGInteractions object with 1 interaction and 0 metadata columns:
##       seqnames1           ranges1     seqnames2           ranges2
##           <Rle>         <IRanges>         <Rle>         <IRanges>
##   [1]        22 18310000-18315000 ---        22 18555000-18560000
##   -------
##   regions: 63774 ranges and 0 metadata columns
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

```r
getPairClusters(mgi[i]) # Loops in this cluster
```

```
## [[1]]
##    seqnames1   start1     end1 width1 strand1 seqnames2   start2     end2
## 1:        22 18310000 18315000   5001       *        22 18555000 18560000
## 2:        22 18310000 18315000   5001       *        22 18555000 18560000
## 3:        22 18310000 18315000   5001       *        22 18555000 18560000
## 4:        22 18310000 18315000   5001       *        22 18555000 18560000
## 5:        22 18310000 18315000   5001       *        22 18555000 18560000
## 6:        22 18310000 18315000   5001       *        22 18555000 18560000
## 7:        22 18310000 18315000   5001       *        22 18555000 18560000
## 8:        22 18310000 18315000   5001       *        22 18555000 18560000
##    width2 strand2 color APScoreAvg ProbabilityofEnrichment RegAPScoreAvg
## 1:   5001       * 0,0,0   4.676819               0.9994674      2.609090
## 2:   5001       * 0,0,0   4.896639               0.9995357      2.695787
## 3:   5001       * 0,0,0   4.471948               0.9987885      2.704470
## 4:   5001       * 0,0,0   4.485403               0.9998216      2.718797
## 5:   5001       * 0,0,0   4.362114               0.9998372      2.626260
## 6:   5001       * 0,0,0   3.802124               0.9965332      2.454063
## 7:   5001       * 0,0,0   4.155105               0.9977696      2.663298
## 8:   5001       * 0,0,0   5.510427               0.9998780      2.897078
##    Avg_diffMaxNeihgboor_1 Avg_diffMaxNeihgboor_2      avg      std     value
## 1:               3.093862               5.313139 7.180269 1.374978  9.930369
## 2:               3.610100               5.274895 6.872339 1.829796 10.081317
## 3:               2.673441               4.234599 6.645153 1.273942  9.021544
## 4:               2.497944               4.063483 6.412110 1.393734  8.632504
## 5:               2.750544               4.049063 6.278961 1.211849  8.723889
## 6:               1.581288               3.555524 6.439600 1.312741  7.845190
## 7:               2.214143               3.572022 6.372737 1.222640  8.340864
## 8:               4.092073               6.506835 7.904566 1.962919 11.541966
##     src
## 1: 0000
## 2: 0030
## 3: 0060
## 4: 0090
## 5: 0120
## 6: 0240
## 7: 0360
## 8: 1440
```

With this function it is possible to identify how many times loops occupy their own cluster, clusters of two, three, four, and more. The example below shows that there are 22,553 single clusters (clusters containing only one loop), and 5,150 clusters of eight loops together. Here the top row represents the number of loops clustered and the bottom row shows how often those clusters are found.

```
getPairClusters(mgi) |>
    vapply(nrow, integer(1L)) |>
    table()
```

```
##     1     2     3     4     5     6     7     8
## 22553  4672  2664  2122  1933  2125  2517  5150
```

The *MergedGInteraction* object also stores the file sources of the input loops which are accessible with the “sources” function. `mergePairs()` uses the names of the *GInterations* list provided as the source names.

```
sources(mgi)
```

```
## [1] "0000" "0030" "0060" "0090" "0120" "0240" "0360" "1440"
```

mariner includes a “subsetBySource” function which returns the subset of merged loops belonging to specific source files. Optional “include” and “exclude” arguments modulate the behavior of “subsetBySource”. For example, “include” requires that pairs be present in specific sources, while 'exclude' requires that pairs be absent from specific sources. Sources not listed in either “include” or “exclude” are ignored and they may or may not be present in the returned *MergedGInteractions* object. “include” and “exclude” can be used separately or in combination to return every possible set. When neither “include” or “exclude” are provided, the default behavior is to return a list of loops unique to each source. The code below shows the number of unique loops (bottom row) in each source time point (top row) after merging.

```
subsetBySource(mgi) |>
    vapply(length, integer(1L))
```

```
## 0000 0030 0060 0090 0120 0240 0360 1440
## 3097 2681 2504 2497 2662 2924 2676 3512
```

Transient loops, or loops that appear only in the middle time points, can be found by including all of the middle time points and excluding the first and last time point. For example, the 53 loops shown below are transient and do not appear in the first or last time point.

```
subsetBySource(x = mgi,
               include = sources(mgi)[2:7],
               exclude = sources(mgi)[c(1,8)])
```

```
## MergedGInteractions object with 53 interactions and 0 metadata columns:
##        seqnames1             ranges1     seqnames2             ranges2
##            <Rle>           <IRanges>         <Rle>           <IRanges>
##    [1]         X   75775000-75780000 ---         X   76635000-76640000
##    [2]         X 106255000-106260000 ---         X 107020000-107025000
##    [3]         X   46507500-46512500 ---         X   46652500-46657500
##    [4]        10 101320000-101325000 ---        10 101365000-101370000
##    [5]        10 101195000-101200000 ---        10 101305000-101310000
##    ...       ...                 ... ...       ...                 ...
##   [49]         9   35790000-35795000 ---         9   35880000-35885000
##   [50]         9   28780000-28785000 ---         9   29210000-29215000
##   [51]        20   29900000-29905000 ---        20   30087500-30092500
##   [52]        20   55665000-55670000 ---        20   56075000-56080000
##   [53]        20   11990000-11995000 ---        20   12190000-12195000
##   -------
##   regions: 63774 ranges and 0 metadata columns
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

De novo loops, which we define as loops that were not called in the first time point but appear in any of the later timepoints can be found by including the last time point but excluding the first.

```
47 subsetBySource(x = mgi,
48                include = "1440",
49                exclude = "0000")
```

```
## MergedGInteractions object with 6933 interactions and 0 metadata columns:
##          seqnames1           ranges1     seqnames2           ranges2
##              <Rle>         <IRanges>         <Rle>         <IRanges>
##      [1]        22 39740000-39745000 ---        22 39910000-39915000
##      [2]        22 42355000-42360000 ---        22 42850000-42855000
##      [3]        22 24730000-24735000 ---        22 24790000-24795000
##      [4]        22 45500000-45505000 ---        22 45655000-45660000
##      [5]        22 28270000-28275000 ---        22 28315000-28320000
##      ...       ...                 ... ...       ...                 ...
##   [6929]        20 31485000-31490000 ---        20 31885000-31890000
##   [6930]        20 30692500-30697500 ---        20 31242500-31247500
##   [6931]        21 44772500-44777500 ---        21 44817500-44822500
##   [6932]        21 26412500-26417500 ---        21 26755000-26760000
##   [6933]        21 46500000-46505000 ---        21 46545000-46550000
##   -------
##   regions: 63774 ranges and 0 metadata columns
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

By not explicitly including or excluding the intervening time points, the result includes 6,933 loops that appear any time after (but not including) the “000” time point.

4.3.2. Identifying differential chromatin interactions

This section demonstrates the workflow for finding differential interactions between two experimental conditions using *mariner* and *DESeq2*. The data comes from our recently published study on phase-separated chromatin looping involving a rare fusion protein in acute myeloid leukemia (AML) (Ahn et al. 2021) which is available on GEO at GSE143465 and GSE143465. The fusion protein, NUP98-HOXA9 (NHA9), contains a DNA-binding domain fused to an intrinsically disordered region (IDR) which forms phase-separated condensates and leads to changes in 3D chromatin structure and deregulated gene expression. To explore how phase separation leads to changes in chromatin structure the IDR of NHA9 was rendered incapable of phase separation by mutating phenylalanine (F) amino acid residues to serine (S), then expressed in HEK293T cells to make the "FS" mutant cell line. The naturally occurring NHA9 was also expressed in HEK293T cells to make the wildtype "WT" cell line. Hi-C was performed in both WT and FS cell lines to compare changes in chromatin structure induced by phase separation. Here *mariner* and *DESeq2* are used to identify differential chromatin interactions between the WT and FS conditions.

Hi-C data were processed into 10 total `.hic` files – four biological replicates per condition (4x2) and one merged file per condition (1x2). SIP was used to call loops in the condition-merged contact maps at 5-Kb resolution and are included with *mariner*.

The first step is to cluster and merge loops (as described in the previous section) to remove duplicate and near-duplicate loops from both datasets. Define loop paths:

```
library(mariner)
loopFiles <-
    system.file("extdata", package = "mariner") |>
    list.files(pattern = "(WT|FS)",
               full.names = TRUE)

basename(loopFiles)
```

```
## [1] "FS_5kbLoops.txt" "WT_5kbLoops.txt"
```

Load as a list of *GInteractions* objects:

```
library(data.table)
giList <-
    lapply(loopFiles, fread) |>
    lapply(as_ginteractions) |>
    setNames(gsub(".*(WT|FS).*", "\\1", loopFiles))

lapply(giList, summary)
```

```
## $FS
## [1] "GInteractions object of length 8566 with 9 metadata columns"
##
## $WT
## [1] "GInteractions object of length 12095 with 9 metadata columns"
```

Cluster and merge loops with a 10-Kb manhattan distance radius followed by assigning loops to 10-Kb bins:

```
## Merge loops and bin to 10-Kb
loops <-
    mergePairs(x = giList, radius = 10e3) |>
    binPairs(binSize = 10e3)
loops
```

```
## MergedGInteractions object with 16716 interactions and 0 metadata columns:
##           seqnames1             ranges1     seqnames2             ranges2
##               <Rle>           <IRanges>         <Rle>           <IRanges>
##       [1]      chr9 118645000-118650000 ---      chr9 119330000-119335000
##       [2]      chr9   15280000-15285000 ---      chr9   15405000-15410000
##       [3]      chr9 110180000-110185000 ---      chr9 111520000-111525000
##       [4]      chr9   80375000-80380000 ---      chr9   80650000-80655000
##       [5]      chr9 108380000-108385000 ---      chr9 108475000-108480000
##       ...       ...                 ... ...       ...                 ...
##   [16712]     chr17   71422500-71427500 ---     chr17   72170000-72175000
##   [16713]     chr17   28450000-28455000 ---     chr17   28657500-28662500
##   [16714]     chr17   65375000-65380000 ---     chr17   65685000-65690000
##   [16715]     chr17   77712500-77717500 ---     chr17   77965000-77970000
##   [16716]     chr17   46880000-46885000 ---     chr17   46955000-46960000
##   -------
##   regions: 28189 ranges and 0 metadata columns
##   seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

These 16,716 merged loops are the interactions that will be tested. Replicates are required to identify differential interactions with *DESeq2*. Next, we define the `.hic` file paths to the eight replicates (four per condition) and remove redundant information from their names. While it is possible to extract

counts from remote files, it is not recommended as it is much faster to download the file and operate on it locally.

```
parentPath <- "https://ftp.ncbi.nlm.nih.gov/geo/samples/GSM4259nnn"
replicates <- c(
    "GSM4259896/suppl/GSM4259896_HEK_HiC_NUP_IDR_WT_A9_1_1_inter_30.hic",
    "GSM4259897/suppl/GSM4259897_HEK_HiC_NUP_IDR_WT_A9_1_2_inter_30.hic",
    "GSM4259898/suppl/GSM4259898_HEK_HiC_NUP_IDR_WT_A9_2_1_inter_30.hic",
    "GSM4259899/suppl/GSM4259899_HEK_HiC_NUP_IDR_WT_A9_2_2_inter_30.hic",
    "GSM4259900/suppl/GSM4259900_HEK_HiC_NUP_IDR_FS_A9_1_1_inter_30.hic",
    "GSM4259901/suppl/GSM4259901_HEK_HiC_NUP_IDR_FS_A9_1_2_inter_30.hic",
    "GSM4259902/suppl/GSM4259902_HEK_HiC_NUP_IDR_FS_A9_2_1_inter_30.hic",
    "GSM4259903/suppl/GSM4259903_HEK_HiC_NUP_IDR_FS_A9_2_2_inter_30.hic"
)
hicFiles <- file.path(parentPath, replicates)

names(hicFiles) <-
    gsub(pattern = ".*(WT|FS).*(1|2)_(1|2).*",
         replacement = "\\1_\\2_\\3",
         x = hicFiles)
names(hicFiles)
```

```
## [1] "WT_1_1" "WT_1_2" "WT_2_1" "WT_2_2" "FS_1_1" "FS_1_2" "FS_2_1" "FS_2_2"
```

It is important to check that the sequence (chromosome) names are the same in both the `.hic` files and the loop files. Functions from *strawr* and *GenomeInfoDb* show the chromosome names from the `.hic` files and loop files, respectively (Cherniavsky Durand and Shamim 2022; Arora et al., n.d.).

```
strawr::readHicChroms(hicFiles[1])$name
```

```
##  [1] "1"   "10"  "11"  "12"  "13"  "14"  "15"  "16"  "17"  "18"  "19"  "2"
## [13] "20"  "21"  "22"  "3"   "4"   "5"   "6"   "7"   "8"   "9"   "ALL" "MT"
## [25] "X"   "Y"
```

```
GenomeInfoDb::seqlevels(loops)
```

```
##  [1] "chr1"  "chr2"  "chr3"  "chr4"  "chr5"  "chr6"  "chr7"  "chr8"  "chr9"
## [10] "chr10" "chr11" "chr12" "chr13" "chr14" "chr15" "chr16" "chr17" "chr18"
## [19] "chr19" "chr20" "chr21" "chr22" "chrX"
```

The “seqnames” in the loops must be the same as those in the `.hic` files to ensure that the correct chromosomes are accessed. The “chr” prefix can be removed by changing the sequence levels style:

```
GenomeInfoDb::seqlevelsStyle(loops) <- "ENSEMBL"
GenomeInfoDb::seqlevels(loops)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
## [17] "15" "16" "17" "18" "19" "20" "21" "22" "X"
```

The "pullHicPixels" function can be used to extract interactions from a list of `.hic` files at a specified `binSize`, or resolution. Pixels are interactions where the width of each pair is equal to the `binSize` **(Fig. 4.3a)**. The "half" argument determines whether to pull the "upper" triangular, "lower" triangular, or "both" from the `.hic` count matrix **(Fig. 4.S1)**. The "norm" and "matrix" arguments determine which normalized values to apply and whether to pull the observed, expected, or "oe" (`observed / expected`) counts. The "pullHic" functions use *strawr* to extract blocks of data from `.hic` files before matching them to interactions. By default, whole chromosomes are extracted as blocks because this is faster than making repeated calls to *strawr*. The "blockSize" parameter limits the size of blocks (in base pairs) to process large .hic files that are too large to fit into computer memory. The following code shows how to extract 10-Kb pixels using `loops` and `hicFiles`:

```
pixels <- pullHicPixels(x = loops,
                        files = hicFiles,
                        binSize = 10e3,
                        half = "both",
                        norm = "NONE",
                        matrix = "observed",
                        blockSize = 10e6)
pixels
```

```
## class: InteractionMatrix
## dim: count matrix with 16716 interactions and 8 file(s)
## metadata(3): binSize norm matrix
## assays(1): counts
## rownames: NULL
## rowData names(0):
## colnames(8): WT_1_1 WT_1_2 ... FS_2_1 FS_2_2
## colData names(2): files fileNames
## type: MergedGInteractions
## regions: 24041
```

The result is an *InteractionMatrix* that contains interactions, metadata, column data (`colData`), row data (`rowData`), and an interaction-by-files matrix of pixel counts. The following code shows how to access each of these values from the *InteractionMatrix* object:

```
50  InteractionSet::interactions(pixels)
```

```
## MergedGInteractions object with 16716 interactions and 0 metadata columns:
##           seqnames1             ranges1     seqnames2             ranges2
##               <Rle>           <IRanges>         <Rle>           <IRanges>
##       [1]         9 118640000-118650000 ---         9 119330000-119340000
##       [2]         9   15280000-15290000 ---         9   15400000-15410000
##       [3]         9 110180000-110190000 ---         9 111520000-111530000
##       [4]         9   80370000-80380000 ---         9   80650000-80660000
##       [5]         9 108380000-108390000 ---         9 108470000-108480000
##       ...       ...                 ... ...       ...                 ...
##   [16712]        17   71420000-71430000 ---        17   72170000-72180000
##   [16713]        17   28450000-28460000 ---        17   28660000-28670000
##   [16714]        17   65370000-65380000 ---        17   65680000-65690000
##   [16715]        17   77710000-77720000 ---        17   77960000-77970000
##   [16716]        17   46880000-46890000 ---        17   46950000-46960000
```

```
51  S4Vectors::metadata(pixels)
```

```
## $binSize
## [1] 10000
##
## $norm
## [1] "NONE"
##
## $matrix
## [1] "observed"
```

```
52  SummarizedExperiment::colData(pixels)
```

```
## DataFrame with 8 rows and 2 columns
##                          files              fileNames
##                    <character>            <character>
## WT_1_1 https://ftp.ncbi.nlm.. GSM4259896_HEK_HiC_N..
## WT_1_2 https://ftp.ncbi.nlm.. GSM4259897_HEK_HiC_N..
## WT_2_1 https://ftp.ncbi.nlm.. GSM4259898_HEK_HiC_N..
## WT_2_2 https://ftp.ncbi.nlm.. GSM4259899_HEK_HiC_N..
## FS_1_1 https://ftp.ncbi.nlm.. GSM4259900_HEK_HiC_N..
## FS_1_2 https://ftp.ncbi.nlm.. GSM4259901_HEK_HiC_N..
## FS_2_1 https://ftp.ncbi.nlm.. GSM4259902_HEK_HiC_N..
## FS_2_2 https://ftp.ncbi.nlm.. GSM4259903_HEK_HiC_N..
```

```
53 SummarizedExperiment::rowData(pixels)
```

```
## DataFrame with 16716 rows and 0 columns
```

```
54 counts(pixels)
```

```
## <16716 x 8> matrix of class DelayedMatrix and type "double":
##          WT_1_1 WT_1_2 WT_2_1 ... FS_2_1 FS_2_2
##     [1,]      9      1     10   .     10     12
##     [2,]     10     11     10   .     12     11
##     [3,]      6      7     15   .      9     13
##     [4,]      9      8      7   .      5     14
##     [5,]     18     16     17   .     22     13
##      ...      .      .      .   .      .      .
## [16712,]     15     16     16   .     15     10
## [16713,]     38     34     34   .     22     23
## [16714,]     54     53     59   .     41     50
## [16715,]     25     34     42   .     30     34
## [16716,]     73    103     65   .     91    113
```

Before calculating differential interactions, condition and replicate information for each .hic file is added to the sample-specific column metadata:

```
55 library(SummarizedExperiment)
56 colData(pixels)$condition <- factor(c(rep("WT", 4), rep("FS", 4)))
57 colData(pixels)$replicate <- factor(rep(1:4, 2))
58 colData(pixels)
```

```
## DataFrame with 8 rows and 4 columns
##                          files              fileNames condition replicate
##                    <character>            <character>  <factor>  <factor>
## WT_1_1 https://ftp.ncbi.nlm.. GSM4259896_HEK_HiC_N..        WT         1
## WT_1_2 https://ftp.ncbi.nlm.. GSM4259897_HEK_HiC_N..        WT         2
## WT_2_1 https://ftp.ncbi.nlm.. GSM4259898_HEK_HiC_N..        WT         3
## WT_2_2 https://ftp.ncbi.nlm.. GSM4259899_HEK_HiC_N..        WT         4
## FS_1_1 https://ftp.ncbi.nlm.. GSM4259900_HEK_HiC_N..        FS         1
## FS_1_2 https://ftp.ncbi.nlm.. GSM4259901_HEK_HiC_N..        FS         2
## FS_2_1 https://ftp.ncbi.nlm.. GSM4259902_HEK_HiC_N..        FS         3
## FS_2_2 https://ftp.ncbi.nlm.. GSM4259903_HEK_HiC_N..        FS         4
```

Since both *InteractionMatrix* and *DESeqDataSet* are subclasses of *SummarizedExperiment*, converting between them requires very little modification with the "DESeqDataSet" function (Morgan et al., n.d.):

```
## Realize DelayedMatrix to an R matrix
counts(pixels) <- as.matrix(counts(pixels))

## Build DEseq data set
library(DESeq2)
dds <- DESeqDataSet(se = pixels, design = ~replicate + condition)
dds
```

```
## class: DESeqDataSet
## dim: 16716 8
## metadata(4): binSize norm matrix version
## assays(1): counts
## rownames: NULL
## rowData names(0):
## colnames(8): WT_1_1 WT_1_2 ... FS_2_1 FS_2_2
## colData names(4): files fileNames condition replicate
```

Then *DESeq2* can be used to find differential interactions between "WT" and "FS":

```
res <-
    DESeq(dds) |>
    lfcShrink(coef = "condition_WT_vs_FS", type = "apeglm")

summary(res)
```

```
## out of 16716 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 338, 2%
## LFC < 0 (down)     : 107, 0.64%
## outliers [1]       : 0, 0%
## low counts [2]     : 1621, 9.7%
## (mean count < 5)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

The results can be added to the *InteractionMatrix* object:

```
library(InteractionSet)
rowData(pixels) <- res
diffLoops <- interactions(pixels)
diffLoops
```

```
## MergedGInteractions object with 16716 interactions and 5 metadata columns:
##           seqnames1             ranges1     seqnames2             ranges2 |
##               <Rle>           <IRanges>         <Rle>           <IRanges> |
##       [1]         9 118640000-118650000 ---         9 119330000-119340000 |
##       [2]         9   15280000-15290000 ---         9   15400000-15410000 |
##       [3]         9 110180000-110190000 ---         9 111520000-111530000 |
##       [4]         9   80370000-80380000 ---         9   80650000-80660000 |
##       [5]         9 108380000-108390000 ---         9 108470000-108480000 |
##       ...       ...                 ... ...       ...                 ... .
##   [16712]        17   71420000-71430000 ---        17   72170000-72180000 |
##   [16713]        17   28450000-28460000 ---        17   28660000-28670000 |
##   [16714]        17   65370000-65380000 ---        17   65680000-65690000 |
##   [16715]        17   77710000-77720000 ---        17   77960000-77970000 |
##   [16716]        17   46880000-46890000 ---        17   46950000-46960000 |
##            baseMean log2FoldChange     lfcSE    pvalue      padj
##           <numeric>      <numeric> <numeric> <numeric> <numeric>
##       [1]   7.43441     -0.0411289  0.217238  0.523631  0.859060
##       [2]   9.78755     -0.0872578  0.226545  0.299526  0.750448
##       [3]  10.82937     -0.0585465  0.214457  0.474826  0.835271
##       [4]   7.52569     -0.0250846  0.211547  0.722496  0.927315
##       [5]  16.99661     -0.0753053  0.207614  0.446470  0.822384
##       ...       ...            ...       ...       ...       ...
##   [16712]   13.8058      0.0604038  0.208170 0.5147400  0.854539
##   [16713]   29.6061      0.1840567  0.238798 0.1465524  0.625448
##   [16714]   48.1686      0.1255563  0.184890 0.2970630  0.748859
##   [16715]   28.7744      0.0906467  0.195974 0.4171083  0.808143
##   [16716]   88.5492     -0.2452487  0.186228 0.0637452  0.483205
##   -------
##   regions: 24041 ranges and 0 metadata columns
##   seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

These interactions are then annotated as “WT-specific”, “FS-specific”, or “static” (i.e., non-differential) for downstream analysis.

```
diffLoops$diff <- "static"
diffLoops$diff[which(diffLoops$padj <= 0.01 &
    diffLoops$log2FoldChange > 0)] <- "WT-specific"
diffLoops$diff[which(diffLoops$padj <= 0.01 &
    diffLoops$log2FoldChange < 0)] <- "FS-specific"

table(diffLoops$diff)
```

```
## FS-specific      static WT-specific
##          21       16521         174
```

The next section demonstrates how to perform an aggregate peak analysis for these differential loops.

4.3.3. Extracting, aggregating, and visualizing matrices

Now that we have identified differential loops between “WT” and “FS” conditions we can perform an aggregate (or average) peak analysis (APA) to visualize the quality of loop calls. APAs are pileup plots of contact frequency matrices from Hi-C data surrounding the central loop pixels. *mariner’s* “pullHicMatrices” and “aggHicMatrices” functions are used to extract count matrices and aggregate them.

The condition-merged .hic files are used to extract count matrices. The following code shows the GEO paths for accessing these files. Downloading these locally will speed up performance.

```
parentPath <-
"https://ftp.ncbi.nlm.nih.gov/geo/series/GSE143nnn/GSE143465/suppl"
conditions <- c(
    "GSE143465_HEK_HiC_NUP_IDR_WT_A9_megaMap_inter_30.hic",
    "GSE143465_HEK_HiC_NUP_IDR_FS_A9_megaMap_inter_30.hic"
)
hicFiles <- file.path(parentPath, conditions)
names(hicFiles) <- gsub(".*(WT|FS).*", "\\1", hicFiles)
basename(hicFiles)
```

```
## [1] "GSE143465_HEK_HiC_NUP_IDR_WT_A9_megaMap_inter_30.hic"
## [2] "GSE143465_HEK_HiC_NUP_IDR_FS_A9_megaMap_inter_30.hic"
```

The “pixelsToMatrices” function expands pixel interactions to square submatrices such that there is a “buffer” of surrounding pixels. In the example below, each 10-Kb interaction is expanded with `buffer=10` to produce 21x21-Kb region:

```
regions <- pixelsToMatrices(x = diffLoops, buffer = 10)
width(regions) |> lapply(unique)
```

```
## $first
## [1] 210001
##
## $second
## [1] 210001
```

The “pullHicMatrices” function is used to extract matrices from .hic files. The dimensions of extracted matrices are defined by the width of each pair divided by the `binSize` **(Fig. 4.3b)**. In the example below each range in the interaction is 21-Kb and the `binSize` is 10-Kb resulting in a 21x21 matrix.

```
matrices <-
    pullHicMatrices(x = regions,
                    files = hicFiles,
                    binSize = 10e3,
                    half = "upper",
                    norm = "SCALE",
                    matrix = "observed")
matrices
```

```
## class: InteractionArray
## dim: 16716 interaction(s), 2 file(s), 21x21 count matrix(es)
## metadata(3): binSize norm matrix
## assays(3): counts rownames colnames
## rownames: NULL
## rowData names(6): baseMean log2FoldChange ... padj diff
## colnames(2): FS WT
## colData names(2): files fileNames
## type: MergedGInteractions
## regions: 24041
```

The result is an *InteractionArray* object that contains the interactions, metadata, Hi-C file metadata (`colData`), interaction metadata (`rowData`), and count matrices for all interactions and `.hic` files. These count matrices are stored on-disk as HDF5 files and use the *HDF5Array* and *DelayedArray* packages to make them available directly in R. Count matrices can be accessed with the “counts” accessor:

```
counts(matrices)
```

```
## <21 x 21 x 16716 x 2> array of class DelayedArray and type "double":
## ,,1,FS
##            [,1]      [,2]      [,3] ...     [,20]     [,21]
##  [1,] 14.386778 11.680648 10.228218   . 15.029434  6.118511
##  [2,] 10.849279 12.688503  7.070478   . 14.167397 10.150929
##   ...         .         .         .   .         .         .
## [20,]  7.692675  8.412569  9.822011   .  5.740219 17.479666
## [21,] 15.632150 13.296138 10.672594   . 16.330437  6.500415
##
## ...
##
## ,,16716,WT
##           [,1]     [,2]     [,3] ...     [,20]     [,21]
##  [1,] 155.2766 143.2167 108.6274   .  35.76799  34.24961
##  [2,] 142.4421 159.5958 127.7207   .  29.75465  35.06651
##   ...        .        .        .   .         .         .
## [20,]       NA       NA       NA   .  86.76485 108.90497
## [21,]       NA       NA       NA   . 124.06264 118.42393
```

An optional “showDimnames” argument is available for showing the row names and column names of the count matrices. These correspond to the first and second ranges of the interactions, respectively.

```
102  counts(matrices, showDimnames = TRUE)
```

```
## <21 x 21 x 16716 x 2> array of class DelayedArray and type "double":
## ,,1,FS
##           119230000 119240000 119250000 ... 119420000 119430000
## 118540000 14.386778 11.680648 10.228218   . 15.029434  6.118511
## 118550000 10.849279 12.688503  7.070478   . 14.167397 10.150929
## 118560000  9.255283 10.121409  5.908575   . 11.049962  3.711230
## 118570000 15.165666  6.046576 10.589430   .  5.941168  4.434211
## 118580000  8.802512  7.487096 13.112227   . 14.304454  9.151010
##       ...         .         .         .   .         .         .
## 118700000  5.560346  9.458856  9.318031   .  8.713089  7.803653
## 118710000 11.445317  9.734970 13.852273   . 10.960197  8.031450
## 118720000  5.951754  3.037407 11.968744   . 16.165810  8.909839
## 118730000  7.692675  8.412569  9.822011   .  5.740219 17.479666
## 118740000 15.632150 13.296138 10.672594   . 16.330437  6.500415
##
## ...
##
## ,,16716,WT
##           46850000 46860000 46870000 ...  47040000  47050000
## 46780000 155.2766 143.2167 108.6274   .  35.76799  34.24961
## 46790000 142.4421 159.5958 127.7207   .  29.75465  35.06651
## 46800000 191.6443 219.0644 140.4027   .  27.17860  35.75507
## 46810000 212.0027 208.6717 122.4590   .  20.39452  34.15556
## 46820000 278.5381 229.4422 145.5619   .  34.80447  28.56599
##      ...        .         .         .   .         .         .
## 46940000       NA        NA        NA   .  57.37274  48.09729
## 46950000       NA        NA        NA   .  46.20938  46.08050
## 46960000       NA        NA        NA   .  63.08284  66.23421
## 46970000       NA        NA        NA   .  86.76485 108.90497
## 46980000       NA        NA        NA   . 124.06264 118.42393
```

Notice that when the row names are less than the column names, the corresponding values are “NA”. These portions are from the lower-triangular of the contact matrix because they cross the Hi-C diagonal. Since `half="upper"`, only upper-triangular values are returned. The code below removes any interactions with portions that cross the diagonal by removing contact matrices with any “NA” values.

```
keep <- apply(counts(matrices), 3, \(x) {
    !(anyNA(x[,,"WT"]) | anyNA(x[,,"FS"]))
})
matrices <- matrices[keep]
dim(matrices)
```

```
## [1] 12047     2
```

Next, the interactions can be grouped into “static”, “WT-specific”, and “FS-specific”:

```
groups <- split(seq_along(matrices), rowData(matrices)$diff)
lapply(groups, head)
```

```
## $`FS-specific`
## [1] 1309 2330 2351 2834 3147 6213
##
## $static
## [1] 1 2 3 4 5 6
##
## $`WT-specific`
## [1] 3354 3380 3627 3648 3714 3821
```

Then the “aggHicMatrices” function can be used to aggregate each group by “files”, producing a list of 21x21x2 arrays. Since the number of interactions varies widely between groups, arrays are then normalized to the number of interactions per group so that values represent the contact frequency per interaction.

```
apa <- lapply(groups, \(i){
    aggHicMatrices(x = matrices[i],
                   by = "files",
                   FUN = sum)
})

## Normalize to interaction per loop
apa <- Map("/", apa, lapply(groups, length))
apa
```

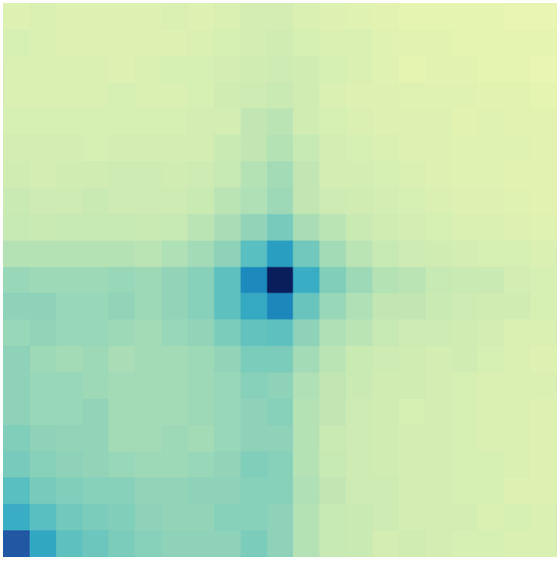
```
## $`FS-specific`
## <21 x 21 x 2> array of class DelayedArray and type "double":
## ,,FS
##            [,1]      [,2]      [,3] ...    [,20]    [,21]
##  [1,]  22.31005  20.40253  16.97262   . 14.11318 14.38912
##  [2,]  20.27343  19.42099  18.94848   . 15.34504 18.71077
##   ...         .         .         .   .        .        .
## [20,]  86.53564  65.81916  61.57496   . 23.52775 17.15086
## [21,] 114.64741  85.05710  65.66801   . 22.58858 20.51914
##
## ,,WT
##            [,1]      [,2]      [,3] ...    [,20]    [,21]
##  [1,]  21.35625  21.14789  20.45077   . 18.44598 19.52502
##  [2,]  23.87447  22.49654  22.13853   . 18.29684 17.32092
##   ...         .         .         .   .        .        .
## [20,]  90.41072  72.52456  54.60084   . 21.40589 22.19293
## [21,] 115.35828  87.80839  62.19818   . 22.58564 21.89970
```

```
## $static
## <21 x 21 x 2> array of class DelayedArray and type "double":
## ,,FS
##           [,1]     [,2]     [,3] ...     [,20]     [,21]
##  [1,] 14.47249 14.37464 14.31745   . 10.010353  9.774367
##  [2,] 15.08028 14.91651 14.83672   . 10.365193 10.091617
##   ...        .        .        .   .         .         .
## [20,] 50.59064 39.77130 35.13720   .  14.83853  14.33831
## [21,] 80.18885 50.57224 39.99249   .  15.06547  14.43662
##
## ,,WT
##           [,1]     [,2]     [,3] ...    [,20]    [,21]
##  [1,] 16.55757 16.35118 16.37003   . 11.90753 11.53139
##  [2,] 17.01740 17.00389 16.85605   . 12.24321 11.87514
##   ...        .        .        .   .        .        .
## [20,] 53.61567 42.27641 37.75946   . 16.90056 16.37085
## [21,] 88.14842 53.59007 42.52097   . 17.05335 16.44842
```

```
## $`WT-specific`
## <21 x 21 x 2> array of class DelayedArray and type "double":
## ,,FS
##           [,1]     [,2]     [,3] ...    [,20]    [,21]
##  [1,] 17.18696 17.74875 18.38293   . 13.47539 13.52509
##  [2,] 18.83570 18.44648 16.74037   . 14.84407 14.39754
##   ...        .        .        .   .        .        .
## [20,] 48.95613 41.84279 35.51618   . 19.06738 18.72985
## [21,] 71.47380 50.81453 42.01940   . 19.16286 19.69601
##
## ,,WT
##           [,1]     [,2]     [,3] ...    [,20]    [,21]
##  [1,] 18.78090 19.59817 18.90748   . 14.31281 14.73756
##  [2,] 20.59297 19.82596 18.97023   . 15.23358 15.69385
##   ...        .        .        .   .        .        .
## [20,] 56.39098 47.84704 42.74671   . 20.88606 20.06353
## [21,] 83.18273 58.69284 46.82731   . 19.94355 19.78126
```

mariner includes the “plotMatrix” function for visualizing *DelayedMatrix* or *matrix* objects:

```
119  plotMatrix(data = apa[["WT-specific"]][,,"WT"])
```



This function is compatible with the R/Bioconductor package, *plotgardener*, which allows customized multi-panel figure plotting. The following code demonstrates how to visualize the APA results with *plotgardener*.

```
## Visualize with plotgardener
library(plotgardener)

## Initiate plotgardener page
pageCreate(width = 4.25, height = 3, showGuides = FALSE)

## Define shared parameters
p <- pgParams(
    x = 0.5,
    y = 0.5,
    width = 1,
    height = 1,
    space = 0.075,
    zrange = c(0, max(vapply(apa, max, numeric(1L))))
)
```

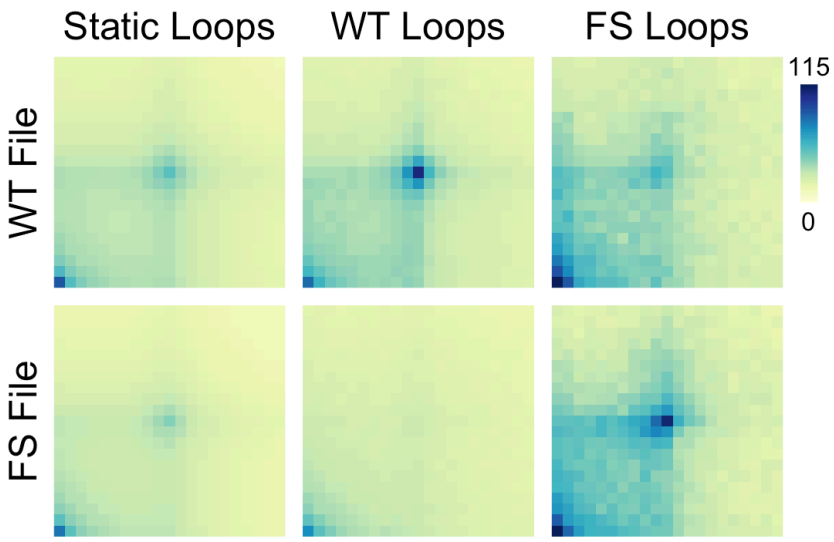
```r
## Define grid of coordinate positions & values
grid <- expand.grid(
    rows = pageLayoutRow(p$y, p$height, p$space, 2) |> as.vector(),
    cols = pageLayoutCol(p$x, p$width, p$space, 3) |> as.vector()
)
data <- expand.grid(
    cell = c("WT", "FS"),
    loop = c("static", "WT-specific", "FS-specific"),
    stringsAsFactors = FALSE
)
layout <- cbind(grid, data)
```

```r
## Plot APAs
apas <- lapply(seq_len(nrow(layout)), \(i) {
    x <- layout[i,]
    plotMatrix(data = apa[[x$loop]][,,x$cell],
               params = p,
               x = x$cols,
               y = x$rows)
})
```

```r
## Add legend
annoHeatmapLegend(plot = apas[[1]],
                  x = p$x + (p$width + p$space)*3,
                  y = grid$rows[1],
                  width = p$space,
                  height = p$height*0.75,
                  fontcolor = 'black')

## Add text labels
plotText(label = c("Static Loops", "WT Loops", "FS Loops"),
         x = unique(grid$cols) + p$width / 2,
         y = grid$rows[1] - p$space,
         just = c('center', 'bottom'))

plotText(label = c("WT File", "FS File"),
         x = grid$cols[1] - p$space,
         y = grid$rows[1:2] + p$height / 2,
         rot = 90,
         just = c('center', 'bottom'))
```



4.4. Conclusion

mariner provides flexible and efficient tools for working with large Hi-C data in the R/Bioconductor ecosystem. However, there are several areas where development would more completely support Hi-C analysis. The *InteractionMatrix* and *InteractionArray* classes allow users to work with large Hi-C data in R; yet they are limited to regular arrays. In other words, the count matrices must be the same dimensions for each interaction. Support for irregular/jagged arrays with an *InteractionJaggedArray* class would allow users to extract and operate on different sized regions, such as TADs, stripes, compartments, and loops at different resolutions. While it is possible to extract separate *InteractionArrays* for these features, a dedicated class would improve speed and usability.

Additional tools for generating interactions from single range data are being developed for the next major release of *mariner*. These would allow users to assess the contact frequency between genomic regions such as protein binding sites, open chromatin regions, or other epigenetic features. Support might include generating interactions within sliding windows or within a specified range of another feature.

While *mariner's* tools are incredibly flexible, its complexity may be challenging for some users. To address this, we are developing wrapper functions that combine *mariner* tools to accomplish more complex tasks. For example, the "calcLoopEnrichment" function combines the "pullHicMatrices" and "aggHicMatrices" functions to calculate the enrichment, or strength, of loops over their local background.

Functions for calculating and visualizing multiple APAs would also simplify *mariner's* usage, while allowing customization when necessary.

4.5. Supplementary figures

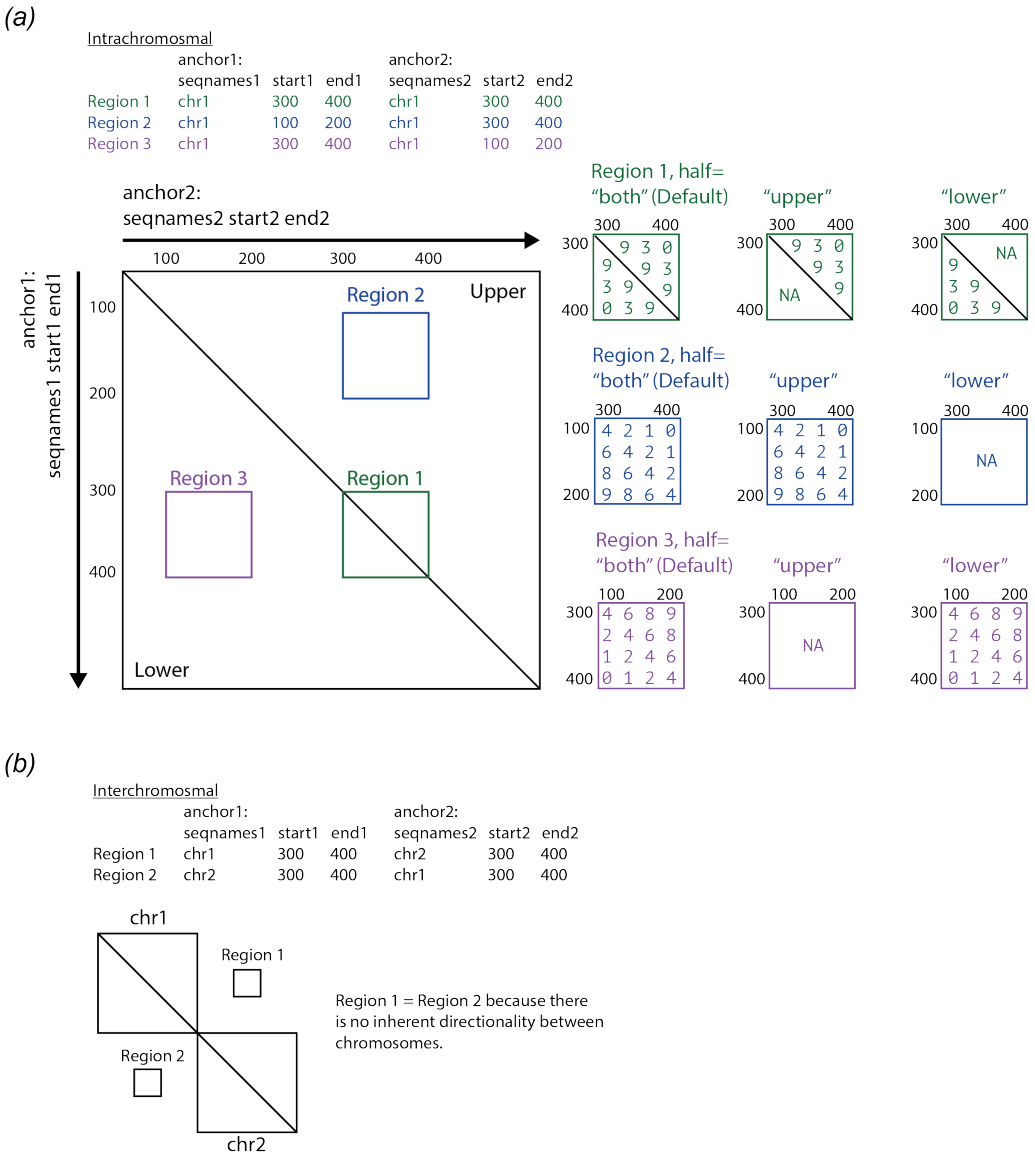


Figure 4.S1. Upper and lower triangular for Hi-C contact matrices. Example interactions demonstrating the results returned when using different values for the "half" argument with "pullHicPixels" or "pullHicMatrices". *(a)* Intrachromosomal interactions for three regions (green, blue, purple), their locations on a Hi-C contact map (lower left), and the results returned for different values of the "half" argument (lower right). Region 1 (green) is located on the Hi-C diagonal and returns mirrored output values when half="both",and values in the corresponding triangular when half="upper" or half="lower". Region 2 (blue) is located in the upper triangular of the Hi-C contact matrix and returns values when half="both" or half="upper", but "NA" when half="lower". Region 3 (purple) is located in the lower triangular of the Hi-C contact matrix and returns values when half="both" or half="lower", but "NA" when half="upper". *(b)* Interchromosomal interactions for two regions. Since there is no inherent directionality between chromosomes, Region 1 and Region 2 will always return the same values.

Chapter 5: Phase separation drives aberrant chromatin looping and cancer development[1]

5.1. Introduction

The development of cancer is intimately associated with genetic abnormalities that target proteins with intrinsically disordered regions (IDRs). In human haematological malignancies, recurrent chromosomal translocation of nucleoporin (NUP98 or NUP214) generates an aberrant chimera that invariably retains the nucleoporin IDR—tandemly dispersed repeats of phenylalanine and glycine residues (Gough, Slape, and Aplan 2011; Mendes and Fahrenkrog 2019). However, how unstructured IDRs contribute to oncogenesis remains unclear. Here we show that IDRs contained within NUP98–HOXA9, a homeodomain-containing transcription factor chimera recurrently detected in leukaemias (Gough, Slape, and Aplan 2011; Mendes and Fahrenkrog 2019), are essential for establishing liquid–liquid phase separation (LLPS) puncta of chimera and for inducing leukaemic transformation. Notably, LLPS of NUP98–HOXA9 not only promotes chromatin occupancy of chimera transcription factors, but also is required for the formation of a broad 'super-enhancer'-like binding pattern typically seen at leukaemogenic genes, which potentiates transcriptional activation. An artificial HOX chimera, created by replacing the phenylalanine and glycine repeats of NUP98 with an unrelated LLPS-forming IDR of the FUS protein (Murray et al. 2017; Alberti and Hyman 2021), had similar enhancing effects on the genome-wide binding and target gene activation of the chimera. Deeply sequenced Hi-C revealed that phase-separated NUP98–HOXA9 induces CTCF-independent chromatin loops that are enriched at proto-oncogenes. Together, this report describes a proof-of-principle example in which cancer acquires mutation to establish oncogenic transcription factor condensates via phase separation, which simultaneously enhances their genomic targeting and induces organization of aberrant three-dimensional chromatin structure during tumourous transformation. As LLPS-competent molecules are frequently implicated in diseases (Gough, Slape, and Aplan 2011; Mendes and Fahrenkrog 2019; Murray et al. 2017; Alberti and Hyman 2021; Boija, Klein, and Young 2021; Wan et al.

[1] The work in this chapter has been previously published. The citation is: Ahn, Jeong Hyun, Eric S. Davis, Timothy A. Daugird, Shuai Zhao, Ivana Yoseli Quiroga, Hidetaka Uryu, Jie Li, et al. 2021. "Phase Separation Drives Aberrant Chromatin Looping and Cancer Development." *Nature* 595 (7868): 591–95.

2020; Kovar 2011), this mechanism can potentially be generalized to many malignant and pathological settings.

5.2. Main

IDRs within various proteins—including transcription factors, chromatin modulators and RNA-binding proteins—form liquid droplets via phase separation, which affects myriad biological processes ranging from organelle formation and stress tolerance to gene transcription (Alberti and Hyman 2021; Boija, Klein, and Young 2021; Sabari 2018; Nair 2019; Chong 2018). Notably, many cancers are characterized by recurrent fusions between genes encoding IDR-containing and chromatin-binding proteins. For instance, a subset of leukaemias that display poor prognosis carry a characteristic chromosomal translocation that produces a gene fusion between an IDR-containing segment of nucleoporin and a chromatin/DNA-binding factor (Gough, Slape, and Aplan 2011; Mendes and Fahrenkrog 2019; G. G. Wang 2009; Jankovic 2008). Similarly, in Ewing's sarcoma, aberrant fusion occurs between transcription factors and the IDR of RNA-binding proteins (Kovar 2011). Both chromatin-binding and IDR-containing domains were previously shown to be essential for tumorigenicity, which supports chromatin deregulation as a general mechanism (Gough, Slape, and Aplan 2011; G. G. Wang 2009; Jankovic 2008). However, how IDRs contribute to gene mis-regulation and oncogenesis is unclear.

5.3. Results

5.3.1. IDRs induce transcription factor phase separation

We aimed to define the role for IDR and potentially phase separation in tumorigenicity by characterizing the NUP98–HOXA9 protein fusion, which shares similarity with other NUP98–transcription factor chimeras identified from various leukaemia subtypes (Gough, Slape, and Aplan 2011; Mendes and Fahrenkrog 2019). NUP98–HOXA9 contains two protein motifs from NUP98—dispersed phenylalaine and glycine (FG) repeats and a GLE2-binding sequence (GLEBS) **(Extended Data Fig. 1a)**. Deletion of GLEBS did not interfere with NUP98–HOXA9-mediated transformation of primary haematopoietic stem and progenitor cells (HSPCs) **(Extended Data Fig. 1b, c)**. Normally, NUP98 is mainly localized at the nuclear periphery. Live-cell imaging showed that full-length and GLEBS-deleted NUP98–HOXA9 displayed a pattern of nucleoplasmic puncta **(Extended Data Fig. 1d)**. Immunoblotting showed that the levels of NUP98 and NUP98–HOXA9 were comparable **(Extended Data Fig. 1d)**. Thus, NUP98–HOXA9-mediated HSPC

transformation and condensate formation are GLEBS-independent. To investigate the role for the NUP98 IDR in leukaemogenesis, we mainly used GLEBS-deleted NUP98–HOXA9 (hereafter referred to as N-IDR$_{WT}$/A9) **(Fig. 1a, b)**.

To determine whether N-IDR$_{WT}$/A9 puncta are established via LLPS, we used several approaches (Sabari 2018; Nair 2019; Chong 2018; Pak 2016). First, we found that N-IDR$_{WT}$/A9 puncta were sensitive to treatment with 1,6-hexanediol, a chemical used to disrupt phase-separated condensates (Sabari 2018; Nair 2019; Chong 2018; Pak 2016) **(Fig. 1c)**. Second, the purified NUP98 IDR (N-IDR) proteins formed liquid condensates in vitro (38xFG) **(Fig. 1d)**. To further assess concentration dependency and importance of multivalency conferred by FG-repeats for condensate formation, we generated recombinant N-IDR proteins that contained a varying number of FG repeats **(Extended Data Fig. 1e, f)**. While N-IDR containing 38x or 36xFG repeats formed liquid droplets in a concentration-dependent fashion **(Fig. 1d)**, those with 27x or 11xFG repeats were unable to phase separate under the same conditions (not shown). Only with the assistance of a crowding agent and at higher concentrations was the 27xFG-repeat-containing N-IDR able to establish condensates in vitro **(Fig. 1d)**. However, when mixed with N-IDR proteins that contained 38xFG repeats, those with 11x or 27xFG repeats were readily incorporated into formed condensates in vitro **(Fig. 1e)**. Imaging of cells expressing N-IDR/A9 with the varying FG-repeat number corroborated in vitro findings—compared with chimeras with 38x or 36xFG repeats, those with fewer FG repeats formed fewer condensates in cells (27x) or could not at all (11x), which is similar to that seen with the HOXA9 fusion segment alone **(Fig. 1f, Extended Data Fig. 1g)**. In addition, DNA binding is dispensable for forming LLPS-like NUP98–HOXA9 puncta. Relative to N-IDR$_{WT}$/A9, its DNA-binding-defective form (carrying an N51S homeodomain mutation (LaRonde-LeBlanc and Wolberger 2003; Calvo et al. 2000)) formed considerably fewer but larger puncta **(Extended Data Fig. 1a, d, h)**, which were also readily detected as droplet-like nuclear structures even under the phase-contrast microscope **(Fig. 1g)**. This indicates that chromatin binding of NUP98–HOXA9 may spatially restrict condensates from further coalescence, which occurs more readily with the N51S-mutant puncta. Condensates of NUP98–HOXA9(N51S) were also sensitive to 1,6-hexanediol treatment **(Extended Data Fig. 1h)**. Notably, live-cell imaging after induction of GFP–NUP98-HOXA9(N51S) showed events of coalescence in which several small condensates collided producing a

larger one **(Fig. 1h)**, which is a characteristic of liquid condensates (Pak 2016). Together, IDR within NUP98–HOXA9 establishes LLPS in a valency-dependent and concentration-dependent manner.

5.3.2. IDRs in transcription factors drive oncogenesis

To investigate the roles for IDR and LLPS in leukaemogenesis, we mutated phenylalanine in the FG repeats of chimeras to serine **(Fig.1a)**—a mutation previously shown to disable hydrogel formation by FG repeats in vitro (Frey, Richter, and Görlich 2006). Such Phe-to-Ser mutations did not affect the protein stability but abolished the nucleoplasmic droplet formation by N-IDR$_{WT}$/A9 carrying either wild-type or N51S-mutated homeodomain, which supports a crucial requirement of FG repeats for LLPS in cells **(Fig.1b, c, Extended Data Figs. 1h, 2a, b)**. NUP98–HOXA9 was reported to interact, either directly or indirectly, with coactivators such as CBP–p300 (Kasper 1999) and MLL–NSL complexes (H. Xu 2016). We next queried whether Phe-to-Ser mutations perturbed such interaction networks by using BioID and found that most N-IDR$_{WT}$/A9- and N-IDR$_{FS}$/A9-interacting proteins were shared, including all reported interactors and many general transcriptional machinery proteins **(Extended Data Fig. 2c, Supplementary Table 1[2])**. To examine the relationship between IDR-mediated LLPS and leukaemogenesis further, we performed the retrovirus-mediated oncogene transduction and transformation assays with mouse HSPCs, and found that, unlike N-IDR$_{WT}$/A9 that efficiently formed nuclear condensates and had a potent HSPC-transforming capacity as previously described (Kroon et al. 2001), the Phe-to-Ser mutant was unable to establish puncta in HSPCs, did not transform HSPCs in vitro, and was unable to induce leukaemia in vivo **(Fig. 1i–k, Extended Data Fig. 2d–g)**. We further assessed the involvement of IDR and LLPS in leukaemogenesis with an artificial chimera termed F-IDR$_{WT}$/A9 by fusing the homeodomain of HOXA9 to an unrelated IDR of the FUS protein that can phase separate (J. Wang 2018; Qamar 2018) **(Fig. 1a, b)**. As expected, F-IDR$_{WT}$/A9 formed puncta in cells, a process that was suppressed by treatment with 1,6-hexanediol or a condensate-disrupting mutation (J. Wang 2018) (F-IDR$_{YS}$/A9) **(Fig. 1a, b, Extended Data Fig. 2h)**. Consistent with NUP98–HOXA9, only the IDR-intact and not the Tyr-to-Ser mutant form of F-IDR/A9 caused leukaemic transformation in vitro and in vivo **(Fig. 1i, j, Extended Data Fig.2g, i)**. Altogether, LLPS-forming IDRs retained within chimeric transcription factors are essential for cancerous transformation.

[2] Supplementary material is available online at https://doi.org/10.1038/s41586-021-03662-5.

5.3.3. IDRs enhance genomic binding of chimeras

NUP98–HOXA9 binds DNA via the homeodomain, causing gene deregulation during leukaemogenesis. Next, we assessed the effect of IDR-mediated phase separation on chromatin targeting of NUP98–HOXA9 by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) to map genome-wide binding of LLPS-competent N-IDR$_{WT}$/A9 versus LLPS-incompetent N-IDR$_{FS}$/A9 in their corresponding stable expression cells. Here, 293FT cells provide a system for assessing direct gene-regulatory effects of NUP98–HOXA9, because its cellular state is relatively stable and not apparently altered after transduction of the chimera, in contrast to what was observed in HSPCs such as differentiation arrest (H. Xu 2016; Kroon et al. 2001) **(Extended Data Fig. 2e, f)**. ChIP–seq using antibodies of different tags attached to N-IDR/A9 produced robust, highly correlated signals, whereas ChIP–seq with non-tagged cells generated almost no binding **(Extended Data Fig. 3a–c)**. Both N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 showed preferential binding to intergenic and intronic enhancers, with binding most enriched in expected motifs of HOX-related transcription factors **(Extended Data Fig. 3d–g)**. Despite shared features seen for their targeting, N-IDR$_{WT}$/A9 had a notably enhanced genomic occupancy relative to N-IDR$_{FS}$/A9, irrespective of peak subclasses defined by unsupervised clustering **(Fig. 2a)**. Also, the broad and dense super-enhancer-like peaks are unique to N-IDR$_{WT}$/A9 **(Supplementary Table 2[2])** and enriched at development- and leukaemia-associated genes **(Extended Data Fig. 3h)**, exemplified by *HOX*, *PBX3* and *MEIS1* **(Fig. 2b, c, Extended Data Fig. 4a–e)**. Super-enhancer calling by N-IDR$_{WT}$/A9 or H3K27ac verified their dense binding at proto-oncogenes **(Extended Data Fig. 5a–c)**.

To further assess the role for IDR-induced LLPS in chromatin targeting of chimeras, we used several additional strategies. First, the treatment of 1,6-hexanediol markedly decreased chromatin occupancy of N-IDR$_{WT}$/A9, whereas it had minimal effects on the overall binding of N-IDR$_{FS}$/A9 **(Fig. 2d, Extended Data Fig. 5d, e)**. Treatment with 1,6-hexanediol also suppressed the formation of a vast majority of broad N-IDR$_{WT}$/A9 peaks **(Extended Data Fig. 5f, Supplementary Table 2[2])**. As a result, overall binding of N-IDR$_{WT}$/A9 after treatment with 1,6-hexanediol more closely resembled that of LLPS-incompetent N-IDR$_{FS}$/A9, compared with N-IDR$_{WT}$/A9 without treatment **(Extended Data Fig. 5g)**. Second, we turned to F-IDR/A9 and tested whether the FUS IDR is sufficient to enhance genomic binding of the chimera. ChIP–seq analysis revealed that these two chimeras carrying unrelated LLPS-competent IDRs showed similar

binding patterns—F-IDR$_{WT}$/A9 shows significantly enhanced genomic targeting and broad binding at AML-related oncogenes, in contrast to F-IDR$_{YS}$/A9 **(Fig. 3a, b, Extended Data Figs. 4, 6a, b Supplementary Table 3[2])**. ChIP–seq for N-IDR$_{WT}$/A9 in mouse leukaemias uncovered similar super-enhancer-like peaks at oncogenes, which overlapped those found in 293FT cells **(Extended Data Fig. 6c–e)**. ChIP combined with quantitative PCR (ChIP–qPCR) verified the enhanced enrichment of N-IDR$_{WT}$/A9 and F-IDR$_{WT}$/A9, relative to their corresponding IDR mutant, and suppressive effect by 1,6-hexanediol on binding of N-IDR$_{WT}$/A9, but not its LLPS-defective mutant, to the tested loci **(Extended Data Fig. 6f, g)**. Third, we used cells that expressed NUP98–HOXA9 with varied numbers of FG repeats, which were either LLPS-competent or LLPS-incompetent, and ChIP–qPCR detected significantly enhanced enrichment of LLPS-competent and not LLPS-incompetent fusions at loci that show broad N-IDR$_{WT}$/A9 binding **(Fig. 3c)**, which indicates a crucial FG-repeat number required for establishing LLPS and intensified binding of chimeras. Lastly, we conducted single-molecule imaging studies to evaluate chromatin occupancy of N-IDR$_{WT}$/A9 relative to N-IDR$_{FS}$/A9. Measurements of single-molecule speed and track displacement showed N-IDR$_{WT}$/A9 to be significantly less mobile than N-IDR$_{FS}$/A9 **(Extended Data Fig. 7)**. Two-state kinetic modelling of single-molecule trajectories (Hansen 2018) showed that, compared with N-IDR$_{FS}$/A9, N-IDR$_{WT}$/A9 had a greater fraction of molecules in the low-diffusion bound state and had slower diffusion coefficients **(Fig. 3d, Extended Data Fig. 7f, g)**, which suggests that assemblies of transcription factors, confined within phase-separated puncta, engage target DNA sequences more tightly and generally display slower diffusion, compared with LLPS-defective transcription factors. Collectively, using both genetic and pharmacological approaches, we have demonstrated a causal role for IDR-mediated LLPS in establishing enhanced targeting of chimeric transcription factors, particularly those seen at super-enhancer-like peaks.

5.3.4. IDRs potentiate target gene activation

To assess the relationship between NUP98–HOXA9 binding and gene activation, we conducted histone 3 Lys27 acetylation (H3K27ac) ChIP–seq and observed that increased chimera transcription factor binding is correlated with increased H3K27ac **(Fig. 2a–c, Extended Data Fig. 4)**. Immunofluorescence also revealed co-localization of N-IDR$_{WT}$/A9 ‘dots’ with H3K27ac, in comparison to H3K9me3 **(Extended Data Fig. 8a, b)**. To define the role for IDR in target gene regulation further, we performed RNA sequencing (RNA-seq) analysis in 293FT cells with stable chimera expression and identified 303 differentially

expressed genes that were significantly upregulated by N-IDR$_{WT}$/A9, compared with mock treatment and N-IDR$_{FS}$/A9 **(Fig. 3e, Supplementary Table 4[2])**, the effect confirmed by quantitative PCR with reverse transcription (RT–qPCR) **(Extended Data Fig. 8c)**. IDR-dependent gene activation was also observed in 293FT cells with expression of F-IDR$_{WT}$/A9 versus F-IDR$_{YS}$/A9 **(Extended Data Fig. 8d, Supplementary Table 5[2])**, albeit gene activation of F-IDR$_{WT}$/A9 is less than that of N-IDR$_{WT}$/A9 **(Fig. 3f)**, in agreement with a relatively less oncogenic potency by the former in vivo **(Fig. 1i, j)**. In addition, RNA-seq of mouse HSPCs transduced with fusion relative to mock control corroborated that N-IDR$_{WT}$/A9, but not N-IDR$_{FS}$/A9, sustains oncogenic gene-expression programs, which again include *HOX*, *MEIS* and *PBX* family genes and other signatures related to leukaemia and HSPCs **(Fig. 3g, Extended Data Fig. 8e, f, Supplementary Table 6[2])**; as expected, differentiation-related gene sets were suppressed in the N-IDR$_{WT}$/A9 sample **(Extended Data Fig. 8f)**. Gene-regulatory effects of the artificial chimera F-IDR$_{WT}$/A9 were similar to those of N-IDR$_{WT}$/A9 in HSPCs **(Extended Data Fig. 8g, Supplementary Table 7[2])**. Furthermore, a reduction in the FG-repeat number, which decreased LLPS competence, also significantly decreased the effects of the chimera on oncogene transcription and HSPC transformation **(Fig. 3h, i)**. Thus, genomic profiling of independent models strongly supports a crucial role for IDRs in activating proto-oncogenes, many of which carry super-enhancer-like elements bound by chimeric transcription factors and H3K27ac.

5.3.5. IDRs and LLPS induce chromatin looping

Increasing evidence suggests that the phase separation of chromatin-associated factors can modulate gene transcription via alterations to three-dimensional chromatin structure (Nair 2019; Strom 2017; L. Wang 2019; Gibson 2019; Shin 2018). However, so far there is little direct evidence that phase separation can form DNA loops similar to those created by CTCF and cohesin, nor that such phase separation-driven loops have a causal role in human disease. To test the ability of NUP98–HOXA9 to form chromatin loops via LLPS, we generated Hi-C profiles of 293FT cells that expressed either N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9, which revealed 6,615 DNA loops **(Fig. 4a)** and high correlation between replicates **(Extended Data Fig. 9a, b)**. To determine the effect of N-IDR$_{WT}$/A9 on Hi-C contact frequency, we aggregated the interaction counts between the 500 most strongly N-IDR$_{WT}$/A9-occupied sites for both N-IDR$_{WT}$/A9- and N-IDR$_{FS}$/A9-expressing cells. Regions with high occupancy of N-IDR$_{WT}$/A9 exhibited increased interaction frequencies, even between binding sites separated by great distances (greater than 2 Mb) or on different

chromosomes entirely **(Fig. 4b)**. Increased interaction frequencies were not observed between the same loci in cells expressing N-IDR$_{FS}$/A9 **(Fig. 4b)**. Differential analysis revealed 232 loops specific to N-IDR$_{WT}$/A9 and 52 specific to N-IDR$_{FS}$/A9 (DESeq2, $P < 0.01$) **(Fig. 4a, c–e)**. Most (91%) N-IDR$_{WT}$/A9-specific-loop anchors overlapped N-IDR$_{WT}$/A9 binding, whereas only 31% overlapped a CTCF-binding site **(Fig. 4f, Extended Data Fig.9c)**. Thus, N-IDR$_{WT}$/A9 loops form in a largely CTCF-independent manner, consistent with a phase-separation-driven mechanism. Chromatin conformation capture (3C) followed by qPCR (3C–qPCR) after treatment with 1,6-hexanediol showed that the N-IDR$_{WT}$/A9-specific loop at *PBX3*, but not an unrelated CTCF loop, was significantly disrupted **(Extended Data Fig. 9d–g)**. The vast majority (82%) of N-IDR$_{WT}$/A9-specific-loop anchors overlapped H3K27ac, in contrast to only 31% observed for non-differential loop anchors **(Fig. 4f)**, which suggests that N-IDR$_{WT}$/A9-specific loops rewire connections between enhancers and target genes. Indeed, genes with promoters that overlapped N-IDR$_{WT}$/A9-specific-loop anchors exhibited increased expression in N-IDR$_{WT}$/A9-expressing cells, compared to those with N-IDR$_{FS}$/A9, which further supports a regulatory role of these loops **(Fig. 4g)**. The upregulated genes at N-IDR$_{WT}$/A9-specific-loop anchors include proto-oncogenes such as *HOX* and *PBX3* **(Fig. 4d, g, Extended Data Fig. 10a–c)**. These results support the idea that IDRs of chimeric transcription factors induce DNA looping between super-enhancer-like targeting sites and oncogenes via phase separation.

5.4. Discussion

In summary, we show that the LLPS-competent IDR contained within NUP98–HOXA9 is crucial for leukaemogenesis and activation of the oncogenic gene-expression program. These effects are mediated by the ability of the IDR to (1) enhance transcription factor binding to genomic targets, and/or (2) promote long-distance looping between enhancers and oncogene promoters **(Extended Data Fig. 11)**. We demonstrated these effects by both genetic (IDR mutagenesis or replacement with an unrelated one and changing the FG-repeats valency) and pharmacological methods. This study provides a proof-of-principle example of an oncogenic mutation that promotes LLPS-driven transcription factor binding and 3D chromatin reorganization during transformation of tumours. As a wide range of IDR-containing LLPS-competent molecules are implicated in diseases (Gough, Slape, and Aplan 2011; Mendes and Fahrenkrog 2019; Alberti and Hyman 2021; Boija, Klein, and Young 2021; Wan et al. 2020; Kovar 2011), this mechanism can potentially be generalized to many pathological settings.

5.5. Methods

5.5.1. Plasmid construction

The MSCV-based retroviral vector for expression of NUP98–HOXA9 fusion has previously been described (Calvo et al. 2002) and the mammalian expression constructs containing various tagged NUP98–HOXA9 (such as GFP–NUP98–HOXA9 in an inducible expression vector (Fahrenkrog 2016)) were gifts from M. Kamps, B. Fahrenkrog and J. Schwaller. The IDR (amino acids 1–215) of FUS can phase separate and is used for creating an artificial fusion of F-IDR/A9. To generate various chimera constructs of N-IDR/A9 or F-IDR/A9 fusions, we synthesized the gBlocks (IDT) that contain cDNA segments of both fusion partners fused in-frame, with a 3xHA-3xFlag tag added at the C terminus. Each gBlock fragment was cloned into the MSCV retroviral vector with a drug selection marker (Puro or Neo). For live-cell imaging studies, we replaced the 3xHA-3xFlag tag in fusion constructs with EGFP by subcloning. For generating a series of constructs with a varying number of NUP98 FG repeats, we used the following NUP98 portion as its fusion segment in the expression vector: amino acids 1–468 as 38 x FG repeats, 1–468(Δ132–224) as 36 x FG repeats, 65–468 (Δ132–224) as 27 x FG repeats and 357–468 as 11 x FG repeats. For bacterial expression of IDR, the same fragments with varying number of FG repeats were cloned into the pRSFDuet-1 vector (a gift from J. Song). For single-molecule tracking studies, we synthesized gBlocks (IDT) that contain cDNA segments of a HaloTag with flanking enzymatic sites of MluI and XhoI to replace the 3xHA-3xFlag tag described in the above expression vectors. All plasmids used were confirmed by sequencing before use and are listed in Supplementary Table 8[2].

5.5.2. Tissue culture and stable cell line generation

293FT (Thermo Fisher R70007), a fast-growing variant of the HEK293T cell line, and HeLa (ATCC CCL-2) cells were obtained from commercial vendors and maintained using recommended culture conditions. Authentication of cell identities, including those parental and derived lines, was ensured by the Tissue Culture Facility affiliated to UNC Lineberger Comprehensive Cancer Center with genetic signature profiling and fingerprinting analyses (Yu 2015). A routine examination for any possible mycoplasma contamination was performed every month with kits (Lonza). Cells in a passage of less than 10 were used. Retrovirus or lentivirus was packaged and produced in 293FT cells, and the stable cell lines were generated by viral infection followed by drug selection as previously performed (B. Xu 2015; L. Cai 2013). The 293FT

cell lines with stable expression of chimera carrying either wild-type or mutant IDRs were first examined by western blotting and immunofluorescence of the transgene, and the same sets of cells then used throughout this study for various assays such as live-cell imaging and genomic profiling (RNA-seq, ChIP–seq and Hi-C).

5.5.3. Antibodies and western blotting

Immunoblotting was performed as previously described (B. Xu 2015; L. Cai 2013). Affinity-purified antibodies against endogenous NUP98 (raised in rabbits against NUP98 amino acids 51–223 covering GLEBS) was a gift from J. M. van Deursen and used as previously described (Kasper 1999; G. G. Wang et al. 2007). The antibodies used (including the antibody source and dilution) are listed in Supplementary Table 8[2].

5.5.4. Fixed cell immunofluorescence

293FT cells were grown on polylysine-coated coverslips (Corning, 354085) for 24 h at a 37 °C incubator. For non-adherent mouse HPSCs, 0.1 million cells were added on top of polylysine-coated coverslips and centrifuged for 30 min at 1,600*g*. The cover slips were briefly washed with PBS and then fixed in 4% formaldehyde (Thermo Scientific, 28908) for 10 min at room temperature. Fixed cell samples were washed with cold PBS three times and incubated in PBS plus 0.1% Triton X-100 for 10 min, followed by washing with PBS for three times and incubation in blocking buffer (1% BSA in PBS plus 0.1% Tween-20) for 30 min. After discarding the blocking buffer, the fixed samples were incubated with a primary antibody diluted in the blocking buffer for 2 h at room temperature or overnight at 4 °C in a humidified chamber, and then washed with PBS plus 0.1% Tween-20 for three times (3 min each time). Lastly, the samples were incubated with the secondary antibody conjugated to appropriate fluorophores for 2 h at room temperature and washed three times with PBST before adding the mounting medium (Thermo Scientific, P36935). The slides were then dried overnight at dark before imaging on the Olympus FV1000 confocal microscope with a 100x/1.4NA Plan Apochromat oil immersion objective. DAPI was imaged with an excitation of 405nm and emission from 430–470nm, Alexa Fluor 488 was imaged with an excitation of 488nm and emission from 505–540nm, and Alexa Fluor 594 was imaged with an excitation of 559 nm and emission from 575–675nm.

5.5.5. Live-cell imaging

For live-cell imaging, cells were grown on 35-mm dish with 20-mm glass bottom well (Cellvis, D35-20-1.5-N) for 24 h before imaging. Live-cell imaging was conducted on Olympus FV1000 confocal microscope using 60x and 100x oil objectives. Three-dimensional lattice light sheet microscopy movies of fusion events were acquired on the lattice light sheet system as previously described (Chen 2014) using a square lattice excitation with numerical apertures of 0.5 (outer) and 0.42 (inner). Three-dimensional volumes of cells, acquired every 24 seconds, were imaged by scanning the coverslip along the sample-plane axis and consisted of 140 planes spaced 360 nm apart. Raw data was de-skewed and deconvolved via 10 iterations of Richardson Lucy deconvolution, using an experimentally measured point spread function prior to quantification. To capture the events of coalescence in which multiple small liquid condensates of chimera are fused into a single one, we used Hela cells with stable expression of doxycycline-inducible GFP-tagged NUP98-HOXA9^{N51S} for live-cell imaging upon chimera expression induction.

5.5.6. Chemical treatment

To test the sensitivity of protein aggregates to 1,6-hexanediol treatment, 10% of 1,6-hexanediol (Sigma-Aldrich, 240117) were prepared in PBS. Throughout this study, the 1,6-hexanediol treatment condition was 10% for 1 min. Such 1,6-hexanediol-treated cells, together with the vehicle-treated control cells, were used for various experiments such as immediate imaging or fixation with 1% formaldehyde for subsequent ChIP–seq experiments.

5.5.7. Recombinant protein purification

For bacterial expression of IDR proteins, the His6x tag-containing pRSFDuet-1 vector that contains NUP98 segment covering FG repeats was transformed into Rosetta 2 (DE3) competent cells (Sigma, 71397). Three litres of bacterial cultures were grown at 37 °C for 12 h and then added with a final concentration of 0.5 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) for overnight induction at 16 °C. Bacterial cells were spun down at 6,340*g* for 15 min, resuspended and lysed in 6 M guanidine hydrochloride added with 20 mM imidazole. After brief sonication, lysates were centrifuged 30,600*g* for 1 h at 4 °C, and supernatants were collected. Supernatants were run through Ni-column (Qiagen, 30250) and washed sequentially with the following buffers: 2 M guanidine hydrochloride with 20 mM imidazole, 2 M guanidine hydrochloride with 1 M NaCl, and 2 M guanidine hydrochloride with 20 mM imidazole. The His6x-tagged

target proteins were eluted in 2 M guanidium hydrochloride with 500 mM imidazole, with 50 µl of elution assessed by SDS–PAGE after ethanol precipitation. Then, protein samples were further purified on size exclusion column 10/300 SD75 (GE healthcare) using the AKTA purifier (GE Healthcare, AKTA pure 25) in SEC buffer (2 M guanidine hydrochloride). Fractions with purified target proteins were combined and concentrated using microcon-10 filter (Millipore, MRCPRT010) to reach the sample concentration ranging from 27 µM to 255 µM and kept at −80 °C for storage.

5.5.8. In vitro phase separation assay

We first carried out the labelling of recombinant protein with the Alexa Flour 488 and 594 protein labelling kit (ThermoFisher, A30006 and A3008) according to manufacturer's protocols. To set up the in vitro phase separation assays, the labelled proteins were mixed with unlabelled ones at a ratio of 1:20, and such a mixture further diluted to a desired concentration in the Eppendorf tubes with either TBS buffer alone (50mM Tris-HCl pH 7.5, 150 mM NaCl) or TBS plus a crowding agent such as 20% of polyethylene glycol (PEG) 3350 (ThermoFisher, NC0620958). Imaging was carried out immediately with samples transferred to a 35-mm dish with 20-mm glass bottom well (Cellvis, D35-20-1.5-N) using Olympus FV3000RS Confocal microscope with 100x oil objective. For fluorescence imaging studies with a mixture of two species of N-IDR recombinant proteins containing FG-repeats in different numbers, we used those with 38 x FG-repeats in the final concentration of 2.5 µM in the TBS buffer (labelled with Alexa Flour 488), which was mixed with those labelled with Alexa Flour 594, either carrying 27 x FG-repeats (a final concentration of 2.5 µM) or 11 x FG repeats (a final concentration of 6 µM).

5.5.9. Colocalization analysis

Colocalization analysis between fusion and H3K27ac or H3K9me3 was performed using the EzColocalization plugin in FIJI version 1.53 (Stauffer, Sheng, and Lim 2018). Colocalization was measured using the Pearson's correlation coefficient (PCC). An *a*-priori power analysis of pilot data was performed in G*Power (*z*-tests, two independent Pearson *r* values) and showed that a sample size of at least 388 cells would be required to determine significance at $P > 0.05$ given an effect size of 0.24. For analysis, nuclei were manually segmented by hand tracing with the polygon selection tool, then converted into binary masks used in the EzColocalization plugin to restrict colocalization analysis to the nuclei. PCC values for each cell were averaged and the calculated means were compared with an independent two-tailed Student's *t*-test.

5.5.10. Purification, transduction, and cultivation of primary mouse HSPCs

Primary bone marrow cells were obtained from femur and tibia of 10-week-old female Balb/C mice and then subject to a lineage-negative (Lin^-) enrichment protocol to remove differentiated cell populations as previously described (G. G. Wang et al. 2007; G. G. Wang 2006). Lin^- enriched HSPCs were first stimulated in the base medium (OptiMEM, Invitrogen, 31985) supplemented with 15% of FBS (Invitrogen, 16000-044), 1% of antibiotics, 50 μM of β-mercaptoethanol and a cytokine cocktail that contains 10 ng ml^{-1} each of mouse SCF (Peprotech), FLT3 ligand (FLT3L; Sigma), IL-3 (Peprotech) and IL-6 (Peprotech) for 4 days as previously described (G. G. Wang 2009; G. G. Wang et al. 2007; G. G. Wang 2006). Two days after infection with retrovirus, mouse HSPCs were subject to drug selection and then plated for assaying proliferation and differentiation in the same liquid base medium with SCF alone as previously described (G. G. Wang 2009; G. G. Wang et al. 2007; G. G. Wang 2006). These in vitro cultured HSPC cells were routinely monitored under microscopy and cellular morphology examined by Wright–Giemsa staining as previously described (G. G. Wang 2009; G. G. Wang et al. 2007; G. G. Wang 2006). For HSPCs transduced with a bicistronic GFP-containing retroviral construct, we also scored relative proliferation of GFP-positive HSPCs by FACS every 2–3 days after infection.

5.5.11. Flow cytometry (FACS) analysis

Cells were washed once in the cold FACS buffer (PBS with 1% of FBS added) and then resuspended and incubated in the FACS buffer added with the respective antibodies (1:100 dilution) for 30 min on ice. The cell pellets were washed with FACS buffer and the stained cells were subject to analysis with the FACS machine (Attune Nxt, Thermo Fisher; available in UNC Flow Cytometry Core Facility). Data were analysed using FlowJo software.

5.5.12. In vivo leukaemogenic assay

All animal experiments were approved by and performed in accord with the guidelines of Institutional Animal Care and Use Committee (IACUC) at the University of North Carolina (UNC) at Chapel Hill. Mice were purchased from the Jackson Laboratory and maintained by the Animal Studies Core, UNC Lineberger Comprehensive Cancer Center. Determination of potential leukaemogenic properties of the oncogene was carried out as previously described (G. G. Wang et al. 2007; G. G. Wang, Pasillas, and Kamps 2005), and no statistical method was used to determine size of cohorts, with investigators blinded

to allocation during assays. In brief, 0.5 million of freshly infected and selected murine HSPCs were transplanted to syngeneic 10-week-old female Balb/C mice (JAX lab, 000651) via tail vein injection (carried out by Animal Studies Core of UNC Cancer Center). Mice were regularly monitored with complete blood counting with the collected peripheral blood and abdomen palpation for early signs of leukaemia such as lethargy, increased white blood cell counts and enlarged spleen (B. Xu 2015). Mice exhibiting leukaemic phenotypes were euthanized followed by pathological and histological analyses as described (G. G. Wang et al. 2007; G. G. Wang, Pasillas, and Kamps 2005). Haematoxylin and eosin (H&E) staining of spleen sections was carried by UNC Pathology Core as previously described (R. Lu 2016).

5.5.13. BioID

A BirA cDNA sequence (a gift from B. Strahl) was inserted into N terminus of target protein in the MSCV based retroviral vector, followed by viral production and establishment of 293FT stable expression cells. Proximity-dependent labelling of interacting proteins or BioID was conducted as previously described (Roux, Kim, and Burke 2013; Roux et al. 2018; J. Li 2021). In brief, 293FT stable cells were collected from five 15-cm plates after treatment with 50 μM of biotin for 24 h, and then washed twice with cold PBS. The cell pellets were resuspended in 1 ml of RIPA lysis buffer (10% glycerol, 25mM Tris-HCl pH 8, 150mM NaCl, 2mM EDTA, 0.1% SDS, 1% NP-40, 0.2% sodium deoxycholate), and lysates were added with 1 μl of benzonase (Sigma-Aldrich, E1014) followed by incubation on ice for 1 h. After centrifugation at maximum speed for 30 min at 4 °C, the supernatant was collected and incubated with Neutravidin beads (Thermo Fisher, 29204) overnight at 4 °C. The Neutravidin beads were then washed twice with the RIPA buffer and TAP lysis buffer (10% glycerol, 350mM NaCl, 2 mM EDTA, 0.1% NP-40, 50 mM HEPES, pH 8) sequentially. Finally, the beads were washed three times with the ABC buffer (50 mM ammonium bicarbonate, pH 8) and subjected to mass spectrometry-based analysis.

5.5.14. Mass spectrometry-based protein identification

Proteins were eluted from beads by adding 50 μl 2 × Laemmli buffer (Boston Bioproducts) and heating at 95 °C for 5 min. A total of 50μl of each sample was resolved by SDS–PAGE using a 4–20% Tris-glycine wedge well gel (Invitrogen) and visualized by Coomassie staining. Each SDS–PAGE gel lane was sectioned into 12 segments of equal volume. Each segment was subjected to in-gel trypsin digestion as follows. Gel slices were destained in 50% methanol (Fisher), 50 mM ammonium bicarbonate (Sigma-

Aldrich), followed by reduction in 10 mM Tris [2-carboxyethyl] phosphine (Pierce) and alkylation in 50 mM iodoacetamide (Sigma-Aldrich). Gel slices were then dehydrated in acetonitrile (Fisher), followed by addition of 100 ng porcine sequencing grade modified trypsin (Promega) in 50 mM ammonium bicarbonate (Sigma-Aldrich) and incubation at 37 °C for 12–16 h. Peptide products were then acidified in 0.1% formic acid (Pierce). Tryptic peptides were separated by reverse phase XSelect CSH C18 2.5 um resin (Waters) on an in-line 150 x 0.075 mm column using a nanoAcquity UPLC system (Waters). Peptides were eluted using a 30 min gradient from 97:3 to 67:33 buffer A:B ratio (buffer A: 0.1% formic acid, 0.5% acetonitrile; buffer B: 0.1% formic acid, 99.9% acetonitrile). Eluted peptides were ionized by electrospray (2.15 kV) followed by MS/MS analysis using higher-energy collisional dissociation (HCD) on an Orbitrap Fusion Tribrid mass spectrometer (Thermo) in top-speed data-dependent mode. MS data were acquired using the FTMS analyser in profile mode at a resolution of 240,000 over a range of 375 to 1,500 *m*/*z*. Following HCD activation, MS/MS data were acquired using the ion trap analyser in centroid mode and normal mass range with precursor mass-dependent normalized collision energy between 28.0 and 31.0. Proteins were identified by searching the UniProtKB database restricted to *Homo Sapiens* using Mascot (Matrix Science) with a parent ion tolerance of 3 ppm and a fragment ion tolerance of 0.5 Da, fixed modifications for carbamidomethyl of cysteine, and variable modifications for oxidation on methionine and acetyl on N terminus. Scaffold (Proteome Software) was used to verify MS/MS-based peptide and protein identifications. Peptide identifications were accepted if they could be established with less than 1.0% false discovery by the Scaffold Local false discovery rate algorithm. Protein identifications were accepted if they could be established with less than 1.0% false discovery and contained at least two identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii et al. 2003). Proteins were filtered out if they had a spectral count <8 in all sample groups and the counts were normalized to $\log_2$-normalized spectral abundance factor (NSAF) values. Significant interacting proteins were defined with a cut-off of a $\log_2$-transformed fold change above 2 in the experimental versus control samples.

5.5.15. ChIP-seq

ChIP–seq was carried out as previously described (B. Xu 2015; L. Cai 2018). In brief, cells were fixed in 1% formaldehyde (Thermo Scientific, 28908) for 10 min, followed by quenching with 125 mM glycine for 5 min. Cells were then washed twice with cold PBS added with protease inhibitors (Sigma-Aldrich,

4693132001), and then subjected to resuspension and incubation in LB1 buffer (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100), LB2 buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), and LB3 buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine). The cell nuclei were collected for sonication using Bioruptor sonicator (Diagenode, B01020001; at high-energy setting for 45 cycles with 30 s on and 30 s off). After treatment with Triton X-100 (1% as a final concentration), the supernatant was collected after centrifugation (20,000*g* for 10 min at 4 °C) for incubation with the dynabeads (Invitrogen, 11204D) that are pre-bound with antibodies for around 8 h at 4 °C. After a series of wash, the chromatin–protein complexes bound to beads were eluted, subject to reverse crosslink overnight at 65 °C, and treated with RNase (Roche, 11119915001; 1 h at 37 °C) and then protease K (Roche, 03115828001; 2 h at 55 °C). The final DNA sample, as well as 1% of input chromatin, was recovered using PCR purification kit (Qiagen, 28106). The ChIP–seq library was prepared using NEBNext Ultra II kit (NEB, E7645L) following the manufacturer's instructions. ChIP–seq libraries were sequenced on the Nextseq 550 system using Nextseq 550 High Output Kit v2.5 (Illumina, 20024906). For ChIP–seq of HA-tagged N-IDR/A9 (with either wild-type or mutated IDRs), we used the matched input signals for signal normalization; for ChIP–seq of GFP-tagged N-IDR/A9 (with either wild-type or mutated IDRs), we used signals of *Drosophila* spike-in chromatin for normalization as previously described (Egan 2016) (Active Motif spike-in ChIP–seq reagents, 53083 and 61686).

5.5.16. ChIP-seq data analysis

ChIP–seq data alignment, filtration, peak calling and assignment, and cross-sample comparison were performed as previously described (B. Xu 2015; L. Cai 2018) with slight modifications. In brief, ChIP–seq reads were aligned to human genome build GRCh37/hg19 or to mouse genome build GRCm38/mm10 using STAR version 2.7.1a (Dobin 2013). The MACS2 software was used for peak identification with data from input as controls and default parameters (Yong Zhang et al. 2008). Homer (ver 4.10.0) 'annotatePeaks' and 'findMotifsGenome' functions were used to annotate the called peaks and to find enriched motifs in these called peaks. Alignment files in the bam format were also transformed into read coverage files (bigWig format) using DeepTools (Ramírez et al. 2016). Genomic binding profiles were generated using the deepTools 'bamCompare' functions with options [–operation ratio—pseudocount 1 -binSize 10–

extendReads 250] and normalized to the matched input. The resulting bigWig files were visualized in the Integrative Genome Viewer (IGV). Heat maps for ChIP–seq signals were generated using the deepTools 'computeMatrix' and 'plotHeatmap' functions. ROSE were used for defining super-enhancers (Lovén 2013), with input signals used as control for normalization and peaks at ±2.5 kb from the transcriptional start site excluded. Homer mergePeaks was used to determine overlap of ChIP–seq peaks with default settings.

5.5.17. RNA-seq and data analysis

RNA-seq was performed as previously described (L. Cai 2018; Z. Ren 2019). For 293FT cells, the same stable expression lines used for ChIP–seq were used. For mouse HPSCs, cells were collected for RNA isolation 7 days after viral transduction and drug selection in the OptiMEM medium supplemented with the HSPC-supporting cytokines. In brief, total RNAs were purified using RNeasy Plus kit (Qiagen, 74136) and further processed with Turbo DNA-free kit (Thermo Fisher, AM1907) to ensure the purity of RNA sample. For RNA-seq, the RNA samples were either sent to Novogene or processed using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490) and NEBNext Ultra II RNA library Prep kit (NEB, E7770) as per the manufacturer's instructions. The multiplexed RNA-seq libraries were subjected for deep sequencing using the Illumina NextSeq500 platform (available in the UNC Sequencing Facility) with the Nextseq 550 High Output Kit v2.5 (Illumina, 20024906). For data analysis, RNA-seq reads were mapped to the reference genome followed by differential gene expression analysis as previously described (L. Cai 2018; Z. Ren 2019). In brief, RNA-seq reads were mapped using MapSplice (K. Wang 2010) and quantified using RSEM(B. Li and Dewey 2011). Read counts were upper-quantile normalized and $\log_2$-transformed. Raw read counts were used for differential gene expression analysis by DESeq (Anders and Huber 2010). Gene Ontology analysis was done using the C5 gene set of Molecular Signature Database (MsigDB) collections available in GSEA website (Subramanian 2005).

5.5.18. ChIP-qPCR or RT-qPCR

ChIP–qPCR or RT–qPCR was performed as previously described (R. Lu 2016; L. Cai 2018). ChIP DNA was prepared as described above for ChIP–seq, whereas total RNA was used to generate cDNA with the iScript cDNA Synthesis kit (Biorad, 1708890) for qPCR.

5.5.19. Single molecule tracking, lattice light sheet microscopy, and data analysis

Three-dimensional lattice light sheet microscopy movies of cells were acquired on a modified version of the lattice light sheet system as previously described (Chen 2014) using a square lattice excitation with numerical apertures of 0.4 (outer) and 0.3 (inner). Time intervals and imaging duration are specified in the legends for each dataset presented. Single-molecule tracking was performed on the same system by focusing on a single plane within the nucleus of cells expressing Halo-tag protein fusions. Before imaging, cells were incubated win 1 nM of Halo Tag-Janelia Fluor 549 ligand for 20 min and then washed in PBS (Grimm 2015). After transferring to the microscope, single planes within the nucleus of each cells were imaged under the same lattice illumination parameters above for a total of 20,000–40,000 frames with 20 ms exposures. Before tracking, images were pre-processed with a rolling ball background subtraction and histogram equalization contrast enhancement using ImageJ. Single molecules were then tracked using the TrackMate plugin for ImageJ (Tinevez 2017). To account for variation in protein expression levels between cells and avoid potential tracking artefacts due to different densities of fluorescent molecules, Pandas software library for python (Virtanen 2020) was used to register single particle tracking datasets such that the number particles within a rolling 100 window was consistent both within and between conditions. We controlled for photobleaching and phototoxicity by confirming that mean molecular speeds within a single cell did not vary substantially throughout the course of the imaging experiment. Finally, molecular trajectories were fit to a two-state kinetic model using Spot-On (Hansen 2018) to estimate the mean diffusion coefficients and fraction of molecular populations for both the slow-diffusing/bound state and rapidly diffusing/free state.

5.5.20. In situ Hi-C

In situ Hi-C was performed as previously described (Rao et al. 2014). Five million cells were crosslinked in 1% formaldehyde for 10 min with stirring and quenched by adding 2.5 M glycine to a final concentration of 0.2 M for 5 min with rocking. Cells were pelleted by spinning at 300*g* for 5 min at 4 °C. The pellet was washed with cold PBS and spun again before freezing in liquid nitrogen. Cells were lysed with 10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630 and protease inhibitors (Sigma, P8340) for 15 min on ice. Cells were pelleted and washed once more using the same buffer. Pellets were resuspended in 50 µl of 0.5% SDS and incubated for 7 min at 62 °C. Next, reactions were quenched with 145 µl of water

and 25µl of 10% Triton X-100 (Sigma, 93443) at 37 °C for 15 min. Chromatin was digested overnight with 25 µl of 10X NEBuffer2 and 100 U of MboI at 37 °C with rotation. Reactions were incubated at 62 °C for 20 min to inactivate MboI and then cooled to room temperature. Fragment overhangs were repaired by adding 37.5 µl of 0.4 mM biotin-14-dATP, 1.5 µl of 10mM dCTP, 1.5 µl of 10mM dGTP, 1.5 µl of 10mM dTTP, and 8 µl of 5 U μl^{-1} DNA polymerase I, large (Klenow) fragment and incubating at 37 °C for 1.5 h. Ligation was performed by adding 667 µl of water, 120 µl of 10X NEB T4 DNA ligase buffer, 100 µl of 10% Triton X-100, 12 µl of 10 mg ml^{-1} BSA, and 1 µl of 2,000 U μl^{-1} T4 DNA ligase and incubating at room temperature for 4 h with slow rotation. Samples were pelleted at 2,500*g* and resuspended in 432 µl of water, 18 µl of 20 mg ml^{-1} proteinase K, 50 µl of 10% SDS, 46 µl of 5M NaCl and incubated for 30 min at 55 °C. The temperature was raised to 68 °C and incubated overnight. Samples were cooled to room temperature. Then, 874 µl of pure ethanol and 55 µl of 3 M sodium acetate pH 5.2 were added to each tube which were subsequently incubated for 15 min at −80 °C. Tubes were spun at maximum speed at 2 °C for 15 min and washed twice with 70% ethanol. The resulting pellet was resuspended in 130 µl of 10 mM Tris-HCl, pH 8, and incubated at 37 °C for 15 min. DNA was sheared using an LE220 Covaris Focused-ultrasonicator to a fragment size of 300–500 bp. Sheared DNA was size selected using AMPure XP beads. One hundred and ten µl of beads were added to each reaction and incubated for 5 min. Using a magnetic stand, supernatant was removed and added to a fresh tube. Then, 30 µl of fresh AMPure XP beads were added and incubated for 5 min. Beads were separated on a magnet and washed twice with 700 µl of 70% ethanol without mixing. Beads were left to dry and then sample was eluted using 300 µl of 10 mM Tris-HCl, pH 8. 150 of 10 mg ml^{-1} Dynabeads MyOne Streptavidin T1 beads were washed resuspended in 300 µl of 10 mM Tris HCl, pH 7.5. This solution was added to the samples and incubated for 15 min at room temperature. Beads were washed twice with 600 µl Tween Washing Buffer (TWB; 250 µl Tris-HCl, pH 7.5, 50 µl 0.5 M EDTA, 10 ml 5 M NaCl, 25 µl Tween-20, and 39.675 ml water) at 55 °C for 2 min with shaking. Sheared ends were repaired by adding 88 µl 1 x NEB T4 DNA ligase buffer with 1 mM ATP, 2 µl of 25 mM dNTP mix, 5 µl of 10 U μl^{-1} NEB T4 PNK, 4 µl of 3 U μl^{-1} NEB T4 DNA polymerase I, 1 µl of 5 U μl^{-1} NEB DNA polymerase I, large (Klenow) fragment and incubating at room temperature for 30 min. Beads were washed two more times with TWB for 2 min at 55 °C with shaking. Beads were washed once with 100 µl of 1x NEBuffer 2 and resuspended in 90 µl of 1x NEBuffer 2, 5 µl of 10 mM dATP, 5 µl of 5 U μl^{-1} NEB Klenow exo minus, and incubated at

37 °C for 30 min. Beads were washed two more times with TWB for 2 min at 55 °C with shaking. Beads were washed once in 50 µl of 1x Quick Ligation reaction buffer and resuspended in 50 µl of 1x Quick Ligation reaction buffer. Then, 2 µl of NEB DNA Quick ligase and 3 µl of an Illumina-indexed adaptor were added and the solution was incubated for 15 min at room temperature. Beads were reclaimed using the magnet and washed two more times with TWB for 2 min at 55 °C with shaking. Beads were washed once in 100 µl of 10 mM Tris-HCl, pH 8, and resuspended in 50 µl of 10 mM Tris-HCl, pH 8. Hi-C libraries were amplified for 7–12 cycles in 5 µl of PCR primer cocktail, 20 µl of Enhanced PCR mix, and 25 µl of DNA on beads. The PCR settings included 3 min of 95 °C followed by 7–12 cycles of 20 s at 98 °C, 15 s at 60 °C, and 30 s at 72 °C. Samples were then held at 72 °C for 5 min before lowering to 4 °C until samples were collected. Amplified samples were brought to 250 µl with 10 mM Tris-HCl, pH 8. Samples were separated on a magnet and supernatant was transferred to a new tube. One hundred and seventy-five µl of AMPure XP beads were added to each sample and incubated for 5 min. Beads were separated on a magnet and washed once with 700 µl of 70% ethanol. Supernatant was discarded. One hundred µl of 10 mM Tris-HCl and 70 µl of fresh AMPure XP beads were added and the solution was incubated for 5 min at room temperature. Beads were separated with a magnet and washed twice with 700 µl 70% ethanol. Beads were left to dry and DNA was eluted in 25 µl of Tris HCl, pH 8.0. The resulting libraries were next quantified by Qubit and Bioanalyzer. A low depth sequencing was performed first using the MiniSeq sequencer system (Illumina) and analysed using the Juicer pipeline (Durand, Shamim, et al. 2016) to assess quality control before deep sequencing (NovaSeq S4). Each Hi-C library was assessed in biological and technical duplicate achieving a total of 3 billion reads per cell line.

5.5.21. Hi-C data processing and analysis

In situ Hi-C datasets were processed using the Juicer Hi-C pipeline with default parameters as previously described (Durand, Shamim, et al. 2016). MboI was used as the restriction enzyme, and reads were aligned to the hg19 human reference genome with bwa (version 0.7.17). Data were processed for 3,058,370,530 Hi-C read pairs in N-IDR$_{WT}$/A9 cells, yielding 1,791,818,927 Hi-C contacts (58.59%) and 2,914,343,903 Hi-C read pairs in N-IDR$_{FS}$/A9 cells, yielding 1,708,441,327 Hi-C contacts (58.62%). Hi-C matrices were constructed for each individual replicate for downstream analysis. A Hi-C mega map was constructed by combining all replicates for each condition (that is, N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9). For

visualization, the resulting Hi-C contact matrices were normalized with a matrix balancing algorithm as previously described (Knight and Ruiz 2013) ('KR') to adjust for regional background differences in chromatin accessibility.

Loops were detected using HiCCUPS from the Juicer tools software (version 1.11.09) as previously described (Rao et al. 2014) via the following command: 'hiccups -m 2048 -c 2 -r 5000,10000,25000 -k KR -f 0.1,0.1,0.1 -p 4,2,1 -i 8,6,4 -t 0.2,1.5,1.5,1.75 -d 30000,30000,60000'. A total of 4,788 loops were identified in N-IDR$_{WT}$/A9 and 2,826 loops were identified in N-IDR$_{FS}$/A9 for a total of 7,616 loops at 10-kb resolution. After filtering out redundant loops, 6,615 combined loops remained. Unnormalized loop counts were extracted using the straw api (Durand, Shamim, et al. 2016) for all loops in each replicate (8 total). Differential loops between N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 were determined using DESeq2 (Love, Huber, and Anders 2014), including biological replicate and condition as covariates in the model. 232 N-IDR$_{WT}$/A9-specific loops and 52 N-IDR$_{FS}$/A9-specific loops were considered significantly differential at a Benjamini–Hochberg adjusted *P* value ≤ 0.01.

APA of N-IDR/A9 binding site interactions was conducted in R using straw. All unique, paired interactions between the 500 strongest N-IDR$_{WT}$/A9 ChIP–seq binding sites were categorized into (1) inter-chromosomal (*n* = 95,959), (2) long (≥ 2 Mb) intra-chromosomal (*n* = 6,298), or (3) short (< 2 Mb) intra-chromosomal (*n* = 574) interactions. Short interactions were filtered out such that the corner of the APA plot would not intersect the diagonal, reducing them from *n* = 574 to *n* = 309. Unnormalized pixel values ±10 surrounding pixels were extracted from N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 Hi-C files at 10-kb resolution for each interaction pair. Resulting 21×21, 10-kb pixel matrices were aggregated and normalized to the number of binding site pairs.

APA of differential loop calls was conducted in R using straw. APA was run for all loops (*n* = 6,615), N-IDR$_{WT}$/A9-specific loops (*n* = 232), and N-IDR$_{FS}$/A9-specific loops (*n* = 52) using both N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 Hi-C. Short interactions were filtered out as described above, reducing the number of interactions to *n* = 3,427, *n* = 121, and *n* = 24 for all, N-IDR$_{WT}$/A9-specific and N-IDR$_{FS}$/A9-specific loops, respectively. Unnormalized pixels were extracted with straw producing a 21×21 pixel matrix at 10-kb resolution that was aggregated and normalized by the number of loops per group.

All loops were partitioned as either N-IDR$_{WT}$/A9-specific loops (WT loops) or N-IDR$_{FS}$/A9-specific loops (FS loops) based on differential loop calling (as described above) and then split into separate loop anchors. Loop anchors were then intersected (bedtoolsr) with several features including ChIP–seq peaks for NUP98–HOXA9, CTCF, or H3K27Ac in both cell types (N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9) and with promoter regions (defined as 1,000 bp upstream of transcription start sites). Permutation testing was used to calculate *P* values for each feature's intersection with loop anchors. In short, the observed percentage of each feature present at wild-type or FS loop anchors was calculated. The expected percentage was determined by randomly sampling an equivalent number of loop anchors from all loop anchors called, then calculating the percentage overlap with each feature. This procedure was repeated 1,000 times to create a distribution of expected values. *P* values were determined by summing the number of expected values greater than (or less than if the observed value was less than the mean) the observed value for that feature.

All loops were partitioned as either N-IDR$_{WT}$/A9-specific loops (WT loops) or N-IDR$_{FS}$/A9-specific loops (FS loops) based on differential loop calling (as described above). Each loop was then intersected with 5-kb windows around the transcription start sites of genes using the bedtoolsr 'pairtobed' function with either end of the loop constituting an overlap. The $\log_2$-transformed fold change in expression value (WT/FS) of genes overlapping either end of a wild-type or FS differential loop were plotted along with the expression of all genes. A Dunn's multiple comparison test following a Kruskal–Wallis test showed a statistically significant difference in expression between wild-type-specific gene-loops and either FS-specific gene-loops ($P = 0.015$) or all genes ($P < 0.001$), after *P* value correction with the Benjamini–Hochberg procedure. In this study, wild-type-specific loops were present in the N-IDR$_{WT}$/A9-expressing cells and absent in N-IDR$_{FS}$/A9 cells whereas mutant-specific loops were absent in N-IDR$_{WT}$/A9 cells and present in N-IDR$_{FS}$/A9 cells, supporting accurate calling of differential loops.

5.5.22. 3C-qPCR

Cell samples were processed and analysed as previously described with slight modifications (R. Lu 2016). In brief, 10 million of cells were fixed in 1% formaldehyde at room temperature for 10 min, followed by quenching in 0.125 M glycine for 5 min. Fixed cells were washed in cold PBS and lysed in ice-cold lysis buffer (10 mM Tris-Cl, pH 8.0, 10 mM NaCl, 0.2% NP-40, 1x complete protease inhibitor cocktail) for 1 h at 4 °C. Nuclei were collected by centrifugation at 2,400*g* for 5 min and digested with 800 U of Bgl-II enzyme,

added with 0.3% of SDS and 1.8% of Triton X-100 in the molecular-grade water with respective enzyme digestion buffer (1.2x) for overnight at 37 °C. After inactivation at 65 °C for 20 min with 1.6% of SDS, digested chromatin was subjected to ligation by T4 ligase (NEB) with 1% Triton X-100 for overnight at 16 °C, followed by 30 min incubation at room temperature. Ligated chromatin was treated with protease K for overnight at 65 °C and then treated with RNase for 2 h at 37 °C, followed by DNA purification with the phenol–chloroform extraction protocol. For qPCR, the obtained DNA was diluted 50-fold and used as a template. Primers were designed for the respective genomic loci with chromatin loop as detected by Hi-C mapping experiment. All PCR products were sequenced to confirm that they are indeed correctly ligated products from two distant genomic loci where chromatin loop is expected to form between them. All the primers used for 3C–qPCR are listed in Supplementary Table 8[2].

5.5.23. Statistics and reproducibility

Experimental data are presented as the mean ± s.d. of three independent experiments unless otherwise noted. Statistical analysis was carried out with two-sided Student's *t*-test for comparing the two sets of data with assumed normal distribution. We used a log-rank test for the Kaplan–Meier survival curve to define statistical significance. A *P* value of less than 0.05 was considered to be significant. Statistical significance levels are denoted as follows: **P* < 0.05; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001.

Sample numbers are indicated in the figure legends. Results of images or staining **(shown in Figs. 1c–h, k and Extended Data Figs. 1d, f, h, 2b, d, e, h, 7a,8a)** and western blotting **(Fig. 1b and Extended Data Figs. 1b, d, f, g, 2a)** were reproducible with at least three independent experiments or prepared samples, with the representative ones shown in the figures.

5.6. Main figures

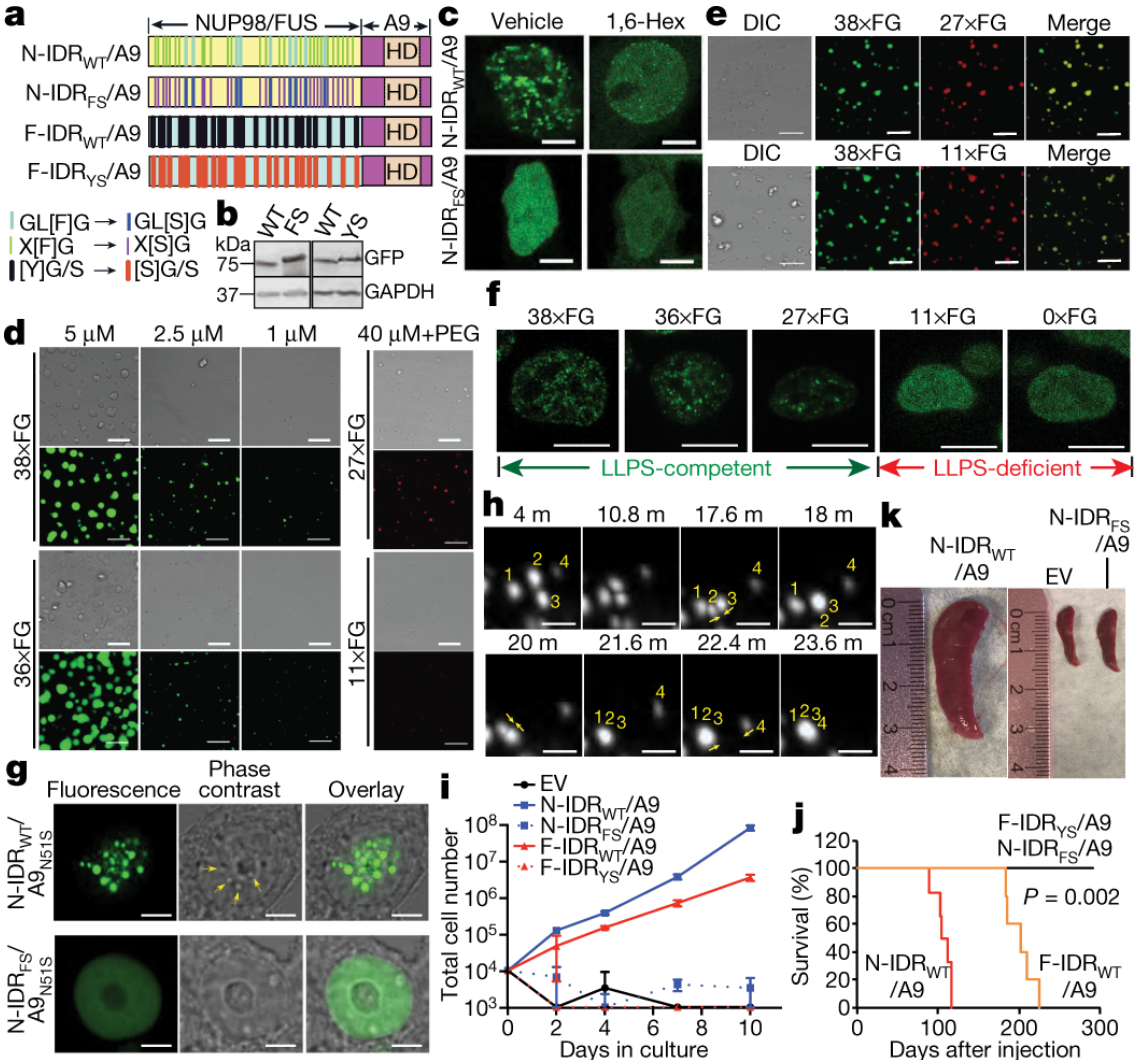


Fig. 5.1: IDRs within chimeric transcription factor oncoproteins establish phase-separated assemblies, inducing leukaemogenesis. a, Scheme for N-IDR/A9 and F-IDR/A9 chimera, with the Phe-to-Ser and Tyr-to-Ser mutations introduced to the NUP98 and FUS IDRs, respectively. HD, homeodomain. **b**, **c**, Immunoblotting (**b**; GADPH was used as a loading control) and live-cell fluorescence (**c**) for GFP-tagged chimera carrying the wild-type (WT) or mutant IDR in 293FT cells. 1,6-Hex, 1,6-hexanediol. Scale bars, 10 μm. For gel source data, see Supplementary Fig. 1[2]. **d**, **e**, Differential interference contrast (DIC) and concurrent fluorescence imaging (bottom) of N-IDR recombinant proteins that contain varying number of FG repeats, prepared at the indicated concentration with either single protein species (**d**) or a mixture of the two (**e**). PEG, polyethylene glycol-3350. Scale bars, 10 μm. **f**, Live-cell imaging of GFP-tagged N-IDR/A9 with the indicated number of FG repeats. Scale bars, 10 μm. **g**, Live-cell imaging (GFP) and concurrent phase-contrast imaging for N51S-mutated GFP–NUP98–HOXA9 with either wild-type (top) or Phe-to-Ser-mutated IDR (bottom). Arrows indicate droplet-like structures. Scale bars, 10 μm. **h**, Coalescence of GFP–NUP98–HOXA9 condensates (N51S-mutated). Scale bars, 2 μm. **i**, Proliferation of mouse HSPCs transduced with empty vector (EV) or the indicated chimera (n = 3 independent biological replicates; data are mean ± s.d.). **j**, Kaplan–Meier survival plot of mice after transplantation of HSPCs transduced with the indicated chimera (n = 5 mice per group). P values were calculated by two-sided log-rank test. k, Splenomegaly associated with N-IDRWT/A9-induced leukaemias, three months after transplantation of infected HSPCs into mice.

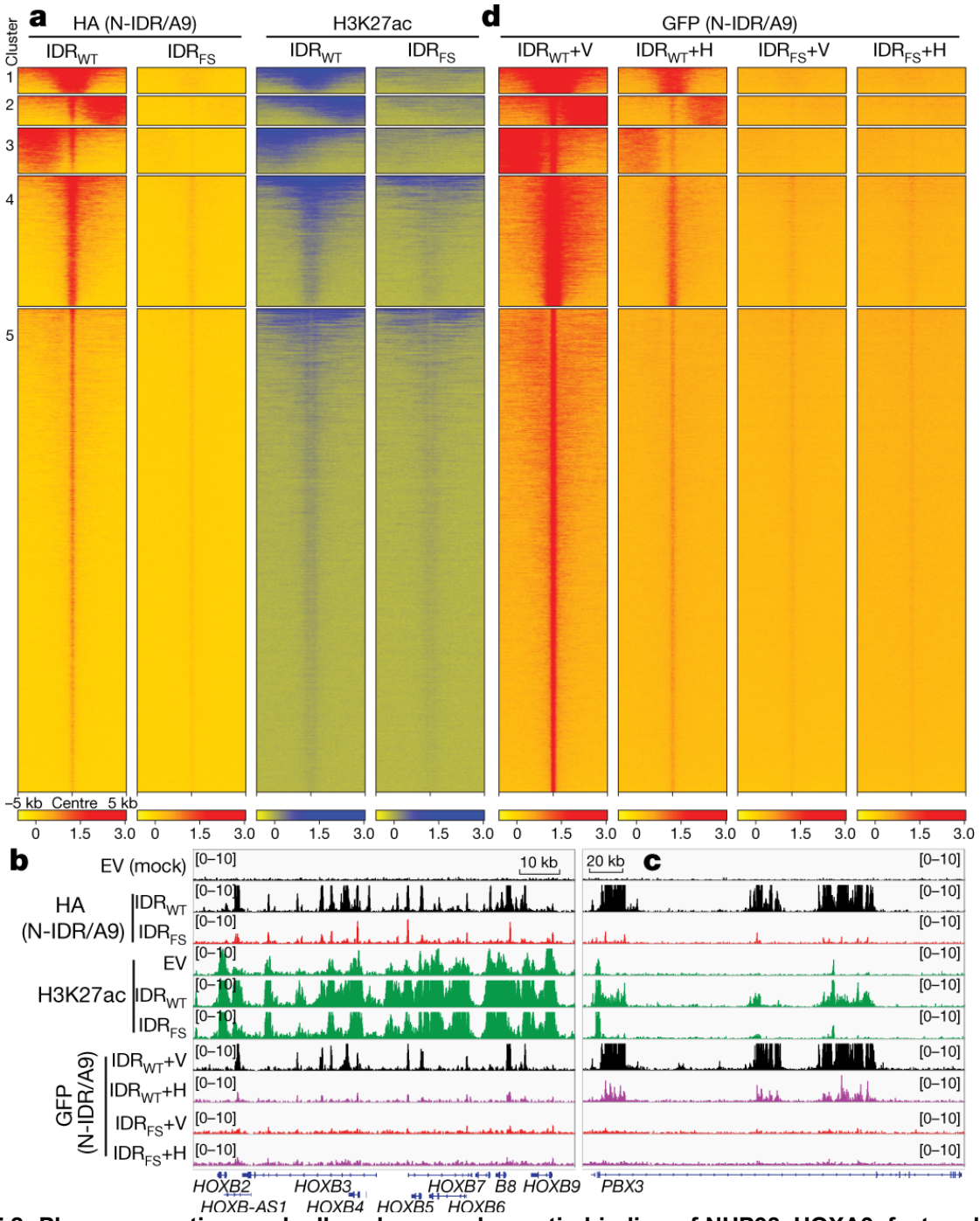


Fig. 5.2: Phase separation markedly enhances chromatin binding of NUP98–HOXA9, featured with broad, super-enhancer-like genomic occupancy. a, **d**, Heat maps for *k*-means clustering of ChIP–seq signals in 293FT cells that express haemagglutinin (HA)-tagged (**a**; input-normalized) or GFP-tagged (**a**; spike-in control normalized) N-IDR/A9 with either wild-type or Phe-to-Ser-mutated IDRs. Cells in **d** were treated with 10% of 1,6-hexanediol (+H), compared with vehicle (+V), for 1 min. Each row represents a peak called for wild-type samples (first column) ± 5 kb from peak centre). **b**, **c**, IGV tracks of the indicated ChIP–seq signals at *HOXB* (**b**) and *PBX3* (**c**) in 293FT cells. EV-transduced cells act as a ChIP control.

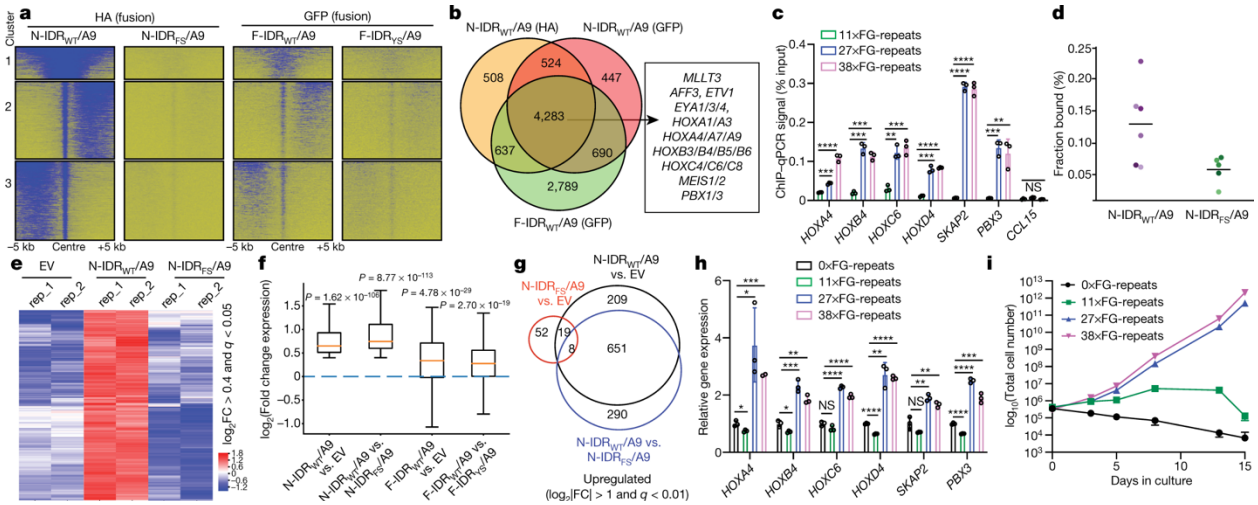


Fig. 5.3: Creation of an artificial F-IDR/A9 chimera and alteration of the FG-repeat valency in NUP98–HOXA9 demonstrate a role for IDR and LLPS in promoting target oncogene activation and cancerous transformation. a, ChIP–seq signal heat maps showing N-IDR/A9 (HA-tagged; left) and F-IDR/A9 (GFP-tagged; right), either wild-type or IDR-mutated (FS or YS), in 293FT cells. See also Extended Data Fig. 6a. **b**, Venn diagram using direct targets of N-IDR$_{WT}$/A9 or F-IDR$_{WT}$/A9 in 293FT cells, with a battery of leukaemia-related oncogenes highlighted. *MLLT3* is also known as *AF9*. **c**, ChIP–qPCR for binding of GFP-tagged N-IDR/A9 with the indicated number of FG repeats at examined loci in 293FT cells (n = 3 independent samples; data are mean ± s.d.). CCL15 acts as a negative control for ChIP. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001, two-sided t-test. **d**, Single-molecule imaging estimated the fraction of chromatin-bound N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 in 293FT stable cells. Presented are values based on two-state kinetic modelling (individual s.d. < 0.0003). Black bars denote mean values. **e**, Heat map of 303 genes upregulated in 293FT cells after transduction of N-IDR$_{WT}$/A9, compared with empty vector and N-IDR$_{FS}$/A9. **f**, Box plots showing relative expression of 303 N-IDR$_{WT}$/A9-activated genes in **e** among the indicated pairwise comparison of 293FT cells. Boxes extend from the first to the third quartile values of the dataset; line denotes median value; whiskers show the data range. P values were determined by two-sided t-test. **g**, Venn diagram using genes upregulated in mouse HPSCs after transduction of the indicated construct. **h**, RT–qPCR for oncogenes in 293FT cells expressing chimera with the indicated number of FG-repeats (n = 3 independent samples; data are mean ± s.d.). Expression was normalized to the 0 x FG-repeat sample. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001, two-sided t-test. NS, not significant. **i**, Proliferation of mouse HSPCs transduced with N-IDR fusion carrying the indicated number of FG repeats (n = 3 independent replicates; data are mean ± s.d.).

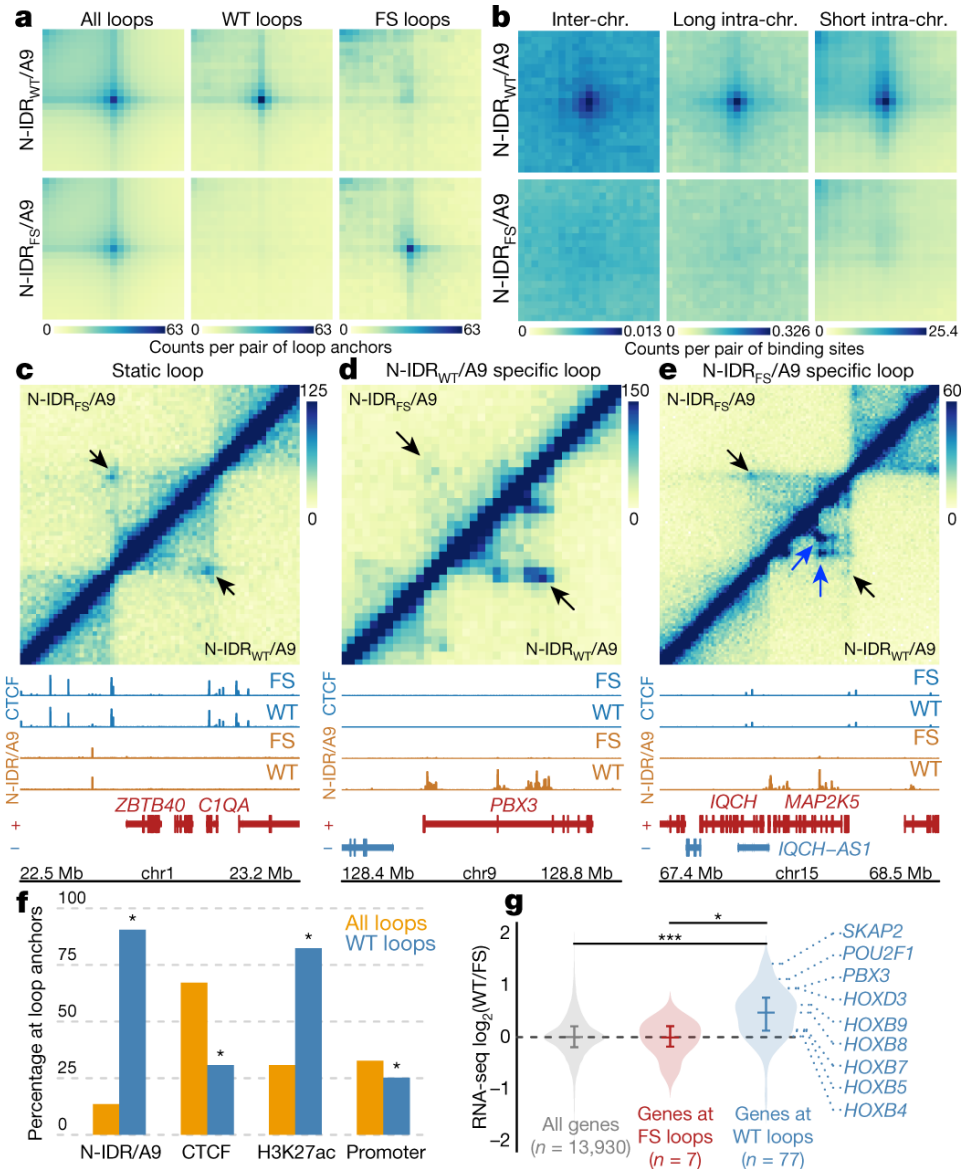
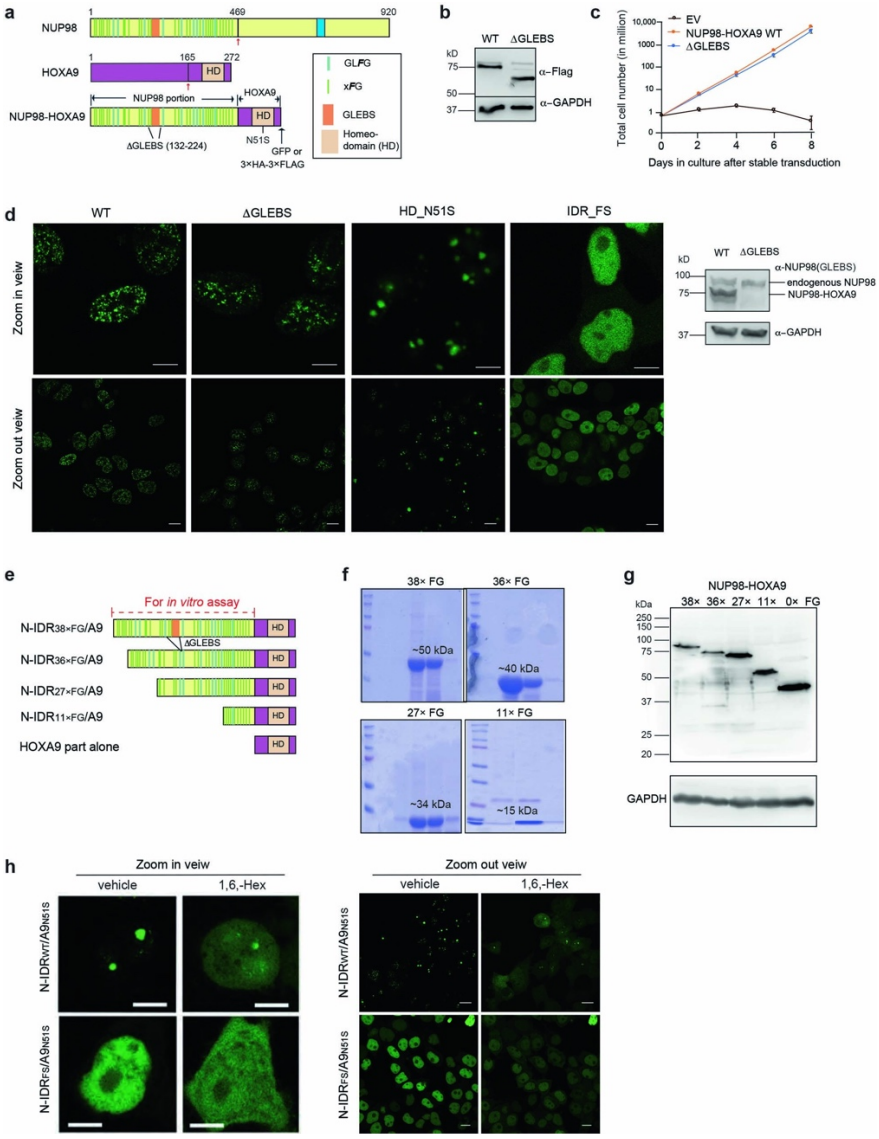


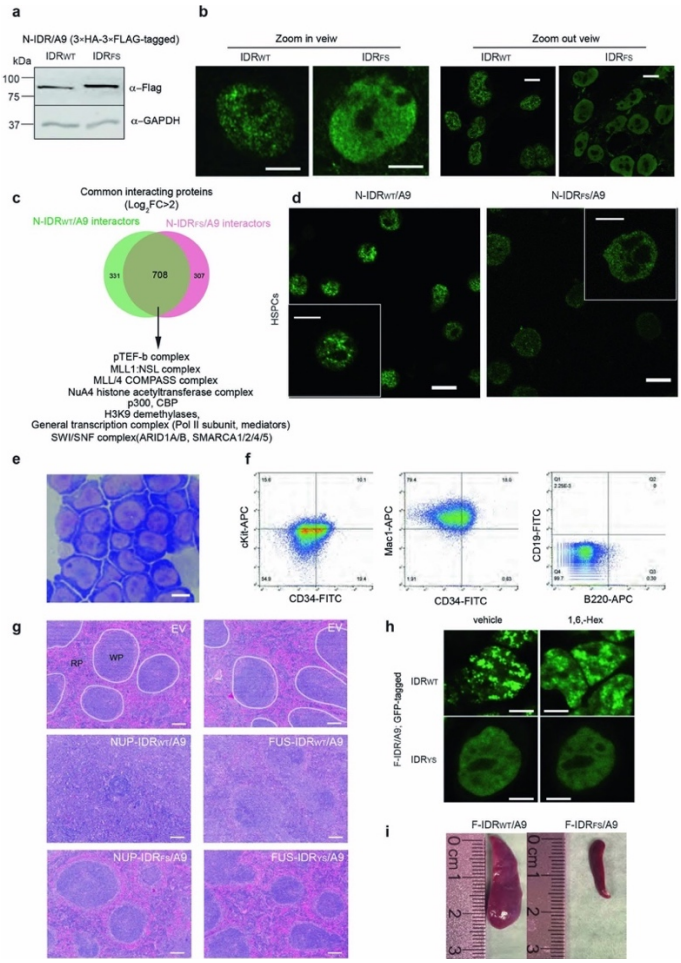
Fig. 5.4: Phase-separation-competent IDRs within NUP98–HOXA9 induce CTCF-independent looping at oncogenes. a, Aggregate peak analysis (APA) for all loops (*n* = 6,615), Wild-type-specific (*n* = 232) and FS-specific (*n* = 52) loops defined by Hi-C in 293FT cells expressing N-IDR$_{WT}$/A9 (top) or N-IDR$_{FS}$/A9 (bottom). Pixel colour represents the mean interaction counts per loop, plotted on a common scale. **b**, APA plots at 10-kb resolution for interactions between the 500 strongest N-IDR/A9 binding sites in cells with N-IDR$_{WT}$/A9 (top) or N-IDR$_{FS}$/A9 (bottom). Paired interactions were categorized as inter-chromosomal (*n* = 95,959), long (≥2 Mb) intra-chromosomal (*n* = 6,298), or short (<2 Mb) intra-chromosomal (*n* = 574). Pixel colour represents the mean interaction counts per pair of loci interrogated. Colour scale in each plot is adjusted to the maximum value. **c**–**e**, Non-differential static (**c**), N-IDR$_{WT}$/A9-specific (**d**; 'Gained in WT' at PBX3) and N-IDR$_{FS}$/A9-specific loop (**e**; 'Lost in WT') detected by Hi-C (arrowheads in top panel) with 293FT cells expressing N-IDR$_{WT}$/A9 (below diagonal) or N-IDR$_{FS}$/A9 (above diagonal). Bottom panels show CTCF (blue) and N-IDR/A9 (orange) ChIP–seq signals (gene tracks shown below) in same cells. **f**, Percentage of the indicated feature present at either all loops or WT-specific loops. **P* < 0.001, permutation test (Methods). **g**, Relative expression of genes associated with wild-type-specific (*n* = 77) and FS-specific loops (*n* = 7) in 293FT cells expressing N-IDR$_{WT}$/A9 versus N-IDR$_{FS}$/A9. **P* < 0.05; ****P* < 0.00001, Benjamini–Hochberg-adjustment method.

5.7. Extended data figures and tables

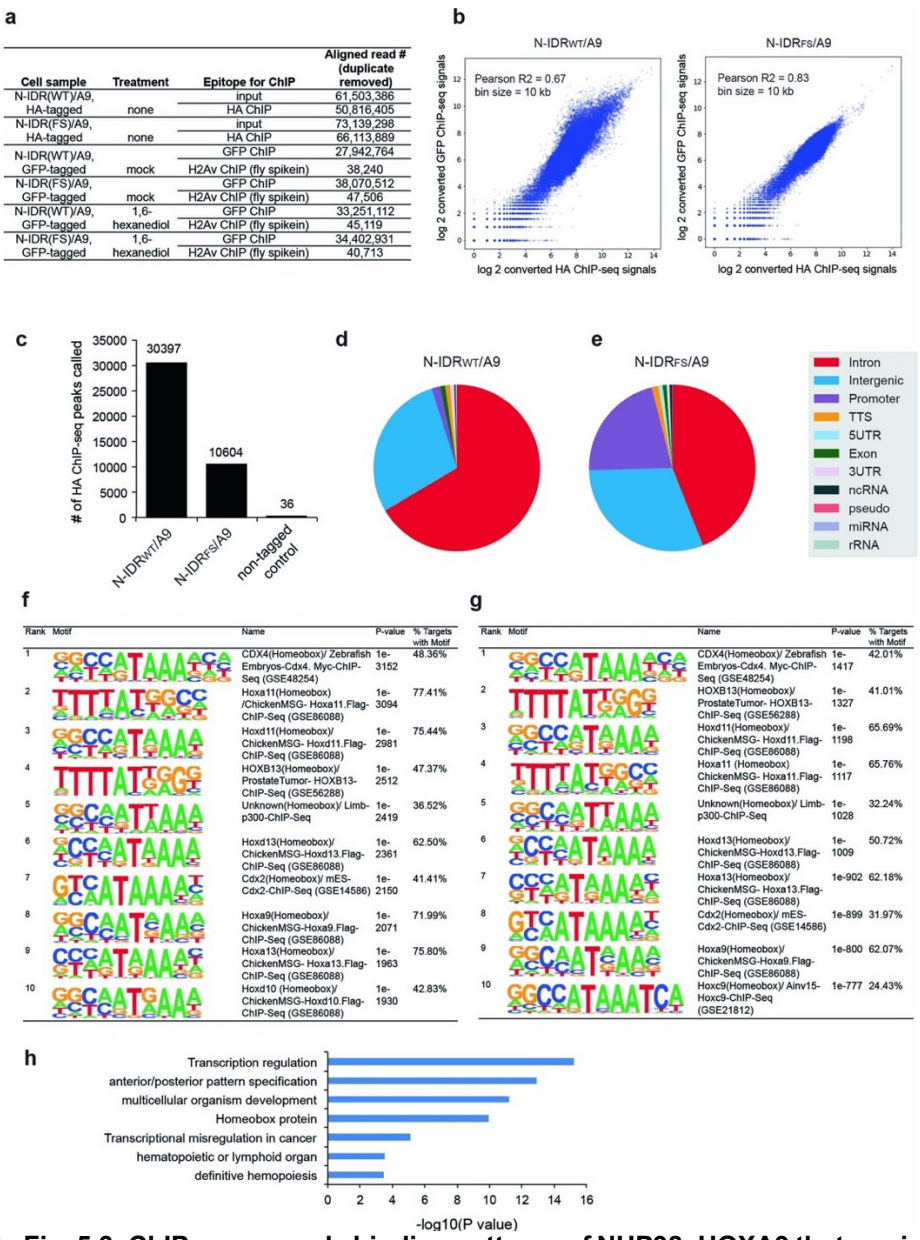


Caption for this figure continues on the next page.

Extended Data Fig. 5.1: IDR retained within the leukaemia-related NUP98–HOXA9 chimera forms phase-separated condensates in vitro and is essential for establishing phase-separated chimeric transcription factor assemblies in the nucleus. a, Schematic showing the domain architecture of normal NUP98 (top), normal HOXA9 (middle) and leukaemic NUP98–HOXA9 chimera (bottom; with either GFP or 3xHA–3xFlag tag fused to C terminus). The GLFG or non-GLFG (xFG) motif contents, which make up IDR, and other important domains are shown. GLEBS represents the GLE2-binding sequence, which directs the NUP98 interaction with GLE2 (also known as RAE1) for mRNA export when NUP98 acts as component of nuclear pore complex (Y. Ren et al. 2010). Red arrows indicate the common breakage point of NUP98 and HOXA9. **b**, Immunoblotting of NUP98–HOXA9, either full-length (WT) or with GLEBS deleted (Δ132-224; see **a**), as detected by the indicated antibodies after stable transduction into primary mouse HSPCs. For gel source data, see Supplementary Fig.1[2]. **c**, Mouse HSPCs stably transduced with wild-type or GLEBS-deleted NUP98–HOXA9 showed similar proliferation in liquid cultures (n = 3 independent cell cultures per group), in agreement to previous reports (Kasper 1999; Yung 2011). Empty vector (EV)-infected HSPCs served as a control. Data are mean ± s.d. **d**, Live-cell fluorescence imaging (GFP; with zoomed-in and zoomed-out views shown in the top and bottom panels, respectively) of 293FT cells with stable transduction of GFP-tagged NUP98–HOXA9, wild-type, GLEBS-deleted (also referred to as N-IDR$_{WT}$/A9; see Fig.1a) or carrying a DNA-binding-defective mutation in homeodomain (HD$_{N51S}$) or a Phe-to-Ser mutation that substitutes Phe residues within all FG repeats to Ser (IDR$_{FS}$, also referred to as N-IDR$_{FS}$/A9; see Fig.1a). The right panel shows immunoblotting of endogenous normal NUP98 in 293FT cells, as well as the stably transduced exogenous NUP98–HOXA9, either wild-type (lane 1) or GLEBS-deleted (lane 2), as detected by antibodies against GLEBS of NUP98 (Kasper 1999). For gel source data, see Supplementary Fig.1[2]. Scale bars, 10 μm. **e**, Schematic of the indicated N-IDR fusion domains with a varying number of FG repeats. The IDR portion used for in vitro assay in main Fig. 1d is indicated by a red dotted line. **f**, SDS–PAGE images showing recombinant N-IDR domain protein with the indicated varying number of FG repeats (His6x-tagged; see **e**), purified with Ni-column and an additional size exclusion column purification step. The protein size is labelled above the recombinant protein. **g**, Anti-GFP immunoblotting for GFP-tagged NUP98–HOXA9 chimera with the indicated varying number of FG repeats described in **e** after stable transduction in 293FT cells. For gel source data, see Supplementary Fig.1[2]. **h**, Live-cell fluorescence imaging for the N51S-mutated N-IDR/A9 (GFP-tagged) with either wild-type (top) or the Phe-to-Ser mutated IDR (bottom) in 293FT stable expression lines before (left) and after (right) treatment with 10% 1,6-hexanediol for 1 min. The left panels show zoomed-in images of a representative cell from the right panels of zoomed-out cell images. Scale bar, 10 μm.



Extended Data Fig. 5.2: IDR contained within chimeric transcription factor is required for leukaemic transformation of primary mouse HSPCs. a, **b**, Immunoblotting (**a**) and fixed cell immunostaining (**b**; anti-Flag) of the LLPS-competent N-IDR$_{WT}$/A9 and LLPS-incompetent N-IDR$_{FS}$/A9 after stable transduction in 293FT cells. The left panel of **b** shows a zoomed-in view on the right panel. Scale bars, 10 μm. For gel source data, see Supplementary Fig.1[2]. **c**, Venn diagram shows significant overlap between the N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 interactomes as detected by BioID, with the cut-off value set as the log$_2$-transformed fold change value above 2 compared with control. Examples of the detected interacting proteins are shown below. **d–f**, Immunostaining (d; anti-GFP), Wright–Giemsa staining (**e**) and FACS with the indicated surface markers (**f**) using mouse HSPCs 1 month after transduction of N-IDR$_{WT}$/A9 (GFP or 3xHA-3xFlag-tagged), which revealed a typical acute myeloid leukaemia phenotype (cKit$^+$, CD34+, Mac1^{high}, CD19$^-$, B220$^-$). The insert in **d** shows a zoomed-in view of the representative cell. Scale bars, 5 μm. For FACS gating strategy, see Supplementary Fig.1[2]. **g**, H&E-stained spleen section images for the indicated cohort at 10x magnification. White pulp (WP) is outlined with white line for the sample from mice transplanted with empty vector-infected HSPCs (top). Note that clear demarcation between white pulp and red pulp (RP), as observed in cohorts receiving either empty vector or the mutant forms of fusion (bottom), is lost in those with N-IDR$_{WT}$/A9 and F-IDR$_{WT}$/A9 (middle) due to an excessive expansion of transformed leukaemia cells that infiltrated into spleen, leading to splenomegaly observed in **i** and Fig. 1k. **h**, Live-cell fluorescence (GFP) imaging of 293FT cells with stable expression of an artificial HOXA9 chimera created by replacing the NUP98 FG repeats with an unrelated IDR of the RNA-binding protein FUS, either wild-type or Tyr-to-Ser mutated (hereafter referred to as the F-IDR$_{WT}$/A9 and F-IDR$_{YS}$/A9 fusion, respectively; see Fig. 1a), before and after treatment with 10% 1,6-hexanediol for 1 min. Scale bar, 10 μm. **i**, Representative image of spleen from mice 7 months after transplantation of mouse HPSCs stably transduced with either F-IDR$_{WT}$/A9 (left) or F-IDR$_{YS}$/A9 (right).

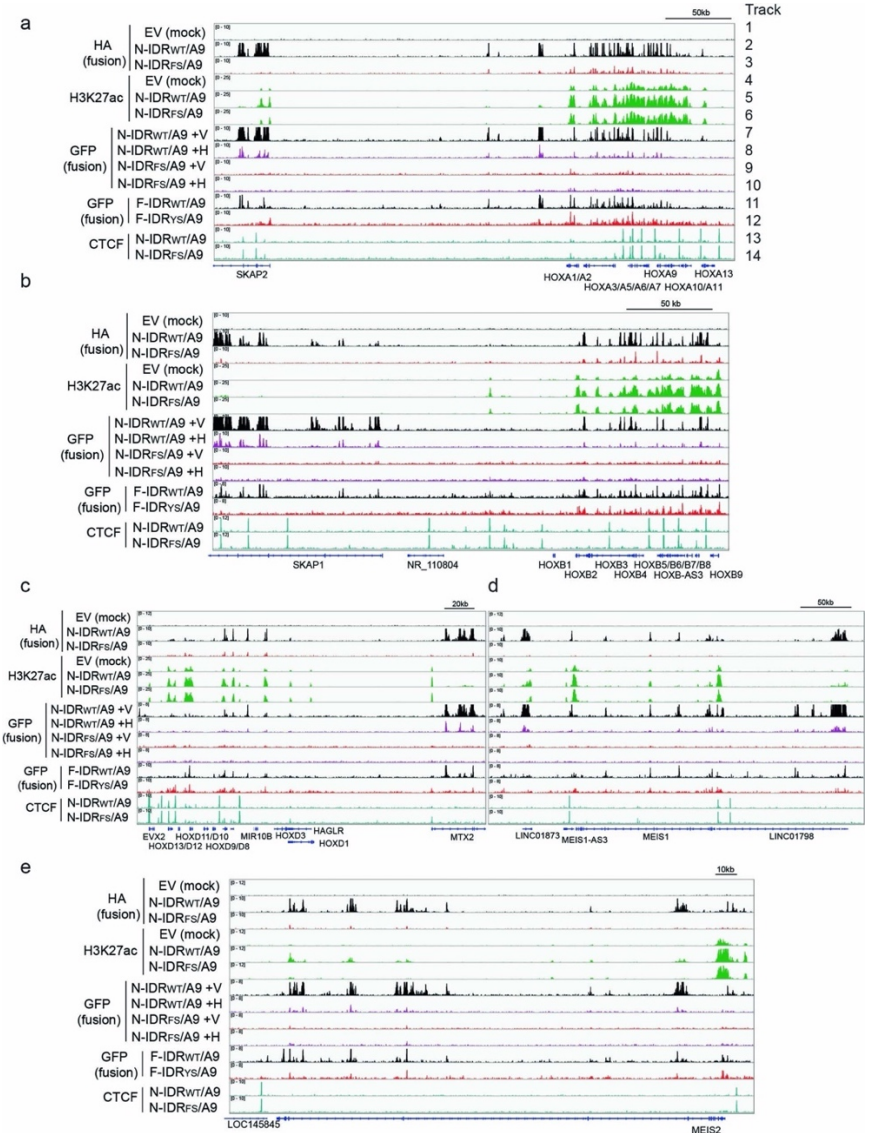


Cell sample	Treatment	Epitope for ChIP	Aligned read # (duplicate removed)
N-IDR(WT)/A9, HA-tagged	none	input	61,503,386
		HA ChIP	50,816,405
N-IDR(FS)/A9, HA-tagged	none	input	73,139,298
		HA ChIP	66,113,889
N-IDR(WT)/A9, GFP-tagged	mock	GFP ChIP	27,942,764
		H2Av ChIP (fly spikein)	38,240
N-IDR(FS)/A9, GFP-tagged	mock	GFP ChIP	38,070,512
		H2Av ChIP (fly spikein)	47,506
N-IDR(WT)/A9, GFP-tagged	1,6-hexanediol	GFP ChIP	33,251,112
		H2Av ChIP (fly spikein)	45,119
N-IDR(FS)/A9, GFP-tagged	1,6-hexanediol	GFP ChIP	34,402,931
		H2Av ChIP (fly spikein)	40,713

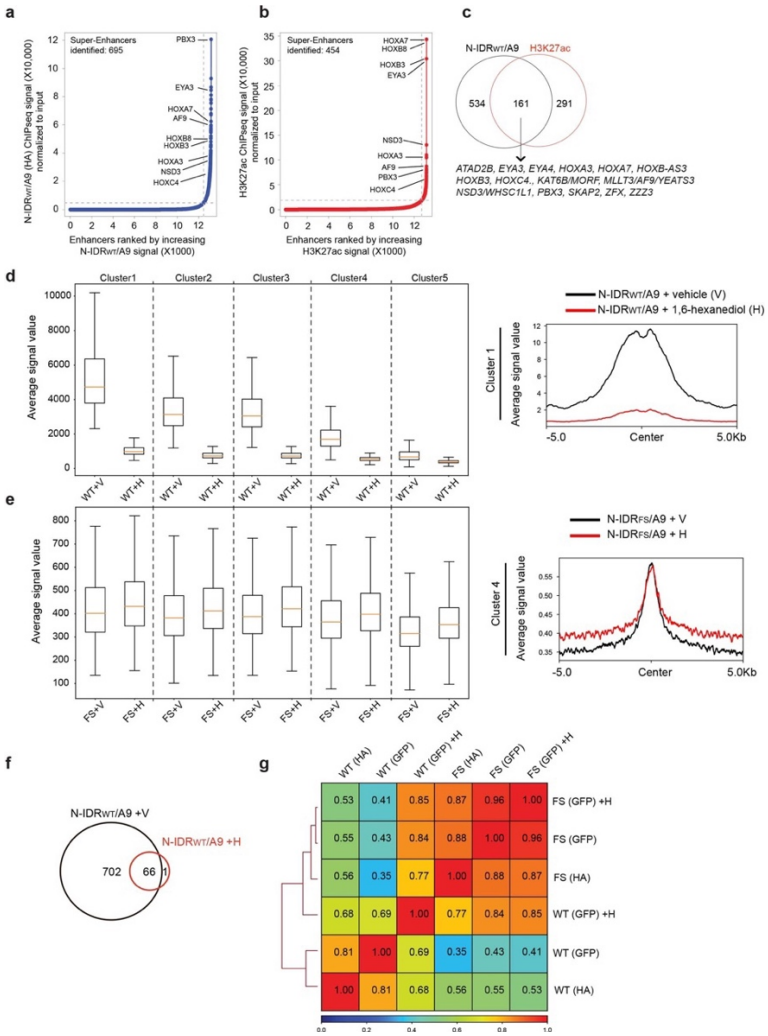
Rank	Motif	Name	P-value	% Targets with Motif
1		CDX4(Homeobox)/ Zebrafish Embryos-Cdx4. Myc-ChIP-Seq (GSE48254)	1e-3152	48.36%
2		Hoxa11(Homeobox) /ChickenMSG- Hoxa11.Flag-ChIP-Seq (GSE86088)	1e-3094	77.41%
3		Hoxd11(Homeobox)/ ChickenMSG- Hoxd11.Flag-ChIP-Seq (GSE86088)	1e-2981	75.44%
4		HOXB13(Homeobox)/ ProstateTumor- HOXB13-ChIP-Seq (GSE56288)	1e-2512	47.37%
5		Unknown(Homeobox)/ Limb-p300-ChIP-Seq	1e-2419	36.52%
6		Hoxd13(Homeobox)/ ChickenMSG-Hoxd13.Flag-ChIP-Seq (GSE86088)	1e-2361	62.50%
7		Cdx2(Homeobox)/ mES-Cdx2-ChIP-Seq (GSE14586)	1e-2150	41.41%
8		Hoxa9(Homeobox)/ ChickenMSG-Hoxa9.Flag-ChIP-Seq (GSE86088)	1e-2071	71.99%
9		Hoxa13(Homeobox)/ ChickenMSG- Hoxa13.Flag-ChIP-Seq (GSE86088)	1e-1963	75.80%
10		Hoxd10 (Homeobox)/ ChickenMSG-Hoxd10.Flag-ChIP-Seq (GSE86088)	1e-1930	42.83%

Rank	Motif	Name	P-value	% Targets with Motif
1		CDX4(Homeobox)/ Zebrafish Embryos-Cdx4. Myc-ChIP-Seq (GSE48254)	1e-1417	42.01%
2		HOXB13(Homeobox)/ ProstateTumor- HOXB13-ChIP-Seq (GSE56288)	1e-1327	41.01%
3		Hoxd11(Homeobox)/ ChickenMSG- Hoxd11.Flag-ChIP-Seq (GSE86088)	1e-1198	65.69%
4		Hoxa11 (Homeobox) ChickenMSG- Hoxa11.Flag-ChIP-Seq (GSE86088)	1e-1117	65.76%
5		Unknown(Homeobox)/ Limb-p300-ChIP-Seq	1e-1028	32.24%
6		Hoxd13(Homeobox)/ ChickenMSG-Hoxd13.Flag-ChIP-Seq (GSE86088)	1e-1009	50.72%
7		Hoxa13(Homeobox)/ ChickenMSG- Hoxa13.Flag-ChIP-Seq (GSE86088)	1e-902	62.18%
8		Cdx2(Homeobox)/ mES-Cdx2-ChIP-Seq (GSE14586)	1e-899	31.97%
9		Hoxa9(Homeobox)/ ChickenMSG-Hoxa9.Flag-ChIP-Seq (GSE86088)	1e-800	62.07%
10		Hoxc9(Homeobox)/ Ainv15-Hoxc9-ChIP-Seq (GSE21812)	1e-777	24.43%

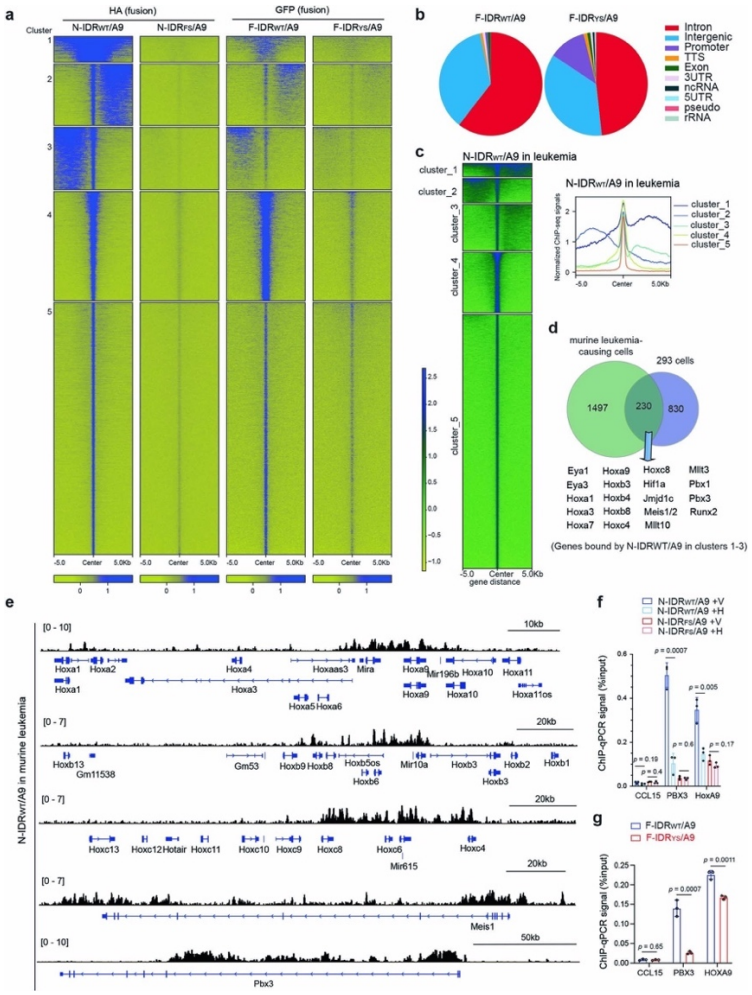
Extended Data Fig. 5.3: ChIP–seq reveals binding patterns of NUP98–HOXA9 that carries either wild-type or an Phe-to-Ser mutated IDR. a, Summary of the counts of ChIP–seq read tags for the indicated samples. **b**, Scatterplots showing correlation of global N-IDR$_{WT}$/A9 (left) or N-IDR$_{FS}$/A9 (right) ChIP–seq signals using either HA (*x* axis) or GFP (*y* axis) antibodies in two biological replicates of 293FT stable cells. Coefficient of determination (R^2) is determined by Pearson correlation. **c**, Total number of the called HA ChIP–seq peaks in stable 293FT cell lines expressing HA-tagged N-IDR$_{WT}$/A9 (left) or N-IDR$_{FS}$/A9 (middle) or empty vector control (right). **d**, **e**, Pie chart showing distribution of the indicated annotation feature among the called N-IDR$_{WT}$/A9 (**d**) or N-IDR$_{FS}$/A9 (**e**) ChIP–seq peaks in 293FT stable expression cells. **f**, **g**, Summary of the most enriched motifs identified within the called N-IDR$_{WT}$/A9 (**f**) or N-IDR$_{FS}$/A9 (**g**) ChIP–seq peaks in 293FT stable expression cells. Motif enrichment was statistically determined by ZOOPS scoring (zero or one occurrence per sequence) coupled with the hypergeometric enrichment calculations. **h**, Gene Ontology analysis of genes associated with broad super-enhancer-like peaks of N-IDR$_{WT}$/A9 as identified in 293FT stable cells. *P* values were determined by Fisher's exact test.



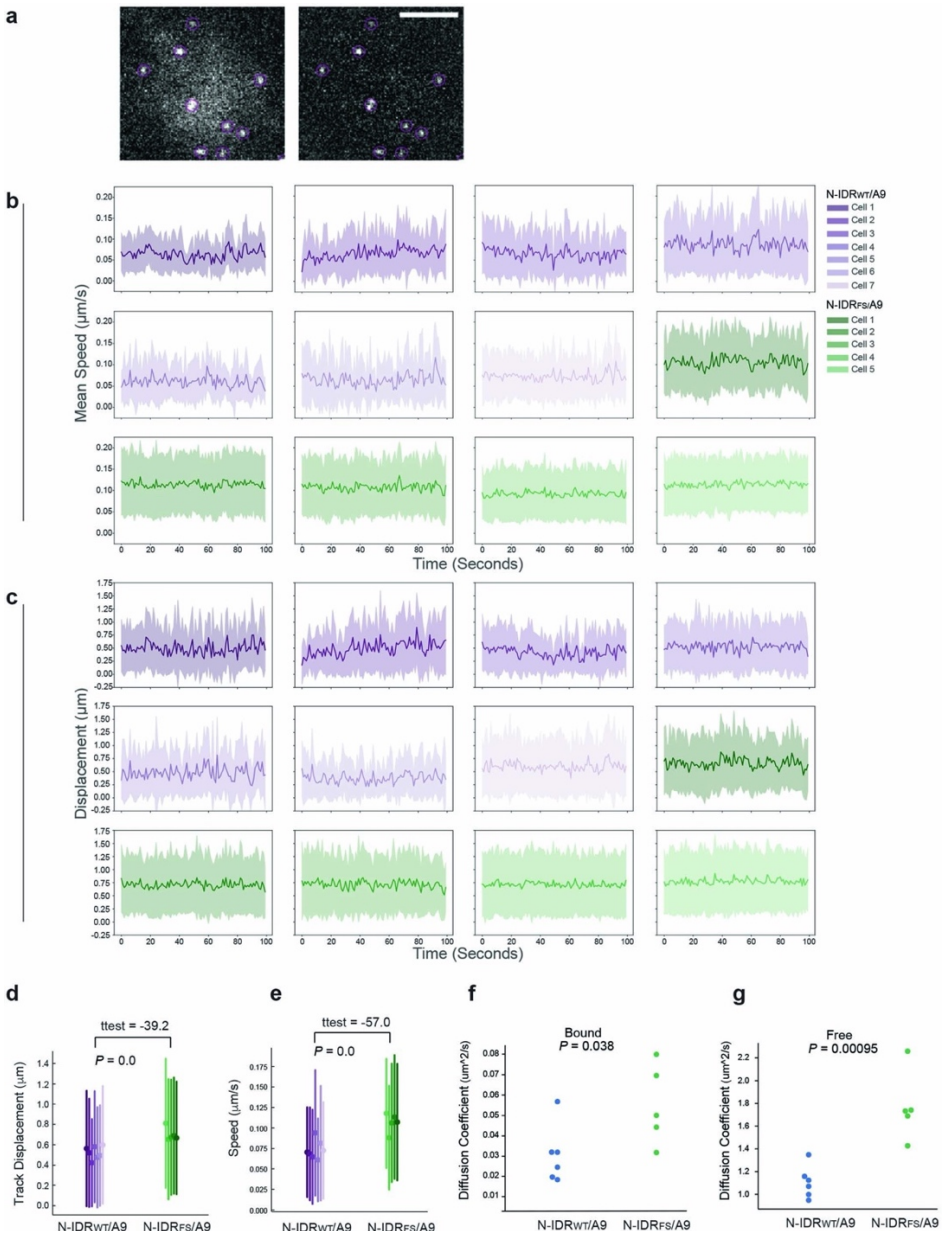
Extended Data Fig. 5.4: Enhanced chromatin occupancy, as well as a broad super-enhancer-like binding pattern typically seen at leukaemia-related genomic loci, is characteristic for the LLPS-competent NUP98–HOXA9 (N-IDR$_{WT}$/A9) and not its LLPS-incompetent IDR mutant (N-IDR$_{FS}$/A9). a–**e**, Integrative genomics viewer (IGV) views for the indicated ChIP–seq signal at the well-known leukaemia-associated loci such as the *HOXA* (**a**), *HOXB* (**b**) and *HOXD* (**c**) gene clusters, *MEIS1* (**d**) and *MEIS2* (**e**). Samples from top to bottom are HA (tracks 1–3) and H3K27ac (tracks 4–6) ChIP–seq signals in the 293FT cells stably expressed with either empty vector (tracks 1 and 4; EV in track 1 acts as a negative control for HA ChIP) or the HA-tagged N-IDR$_{WT}$/A9 (tracks 2 and 5) or N-IDR$_{FS}$/A9 (tracks 3 and 6), GFP ChIP–seq signals (tracks 7–12) in the 293FT cells stably expressed with GFP-tagged N-IDR$_{WT}$/A9 (tracks 7–8 represent samples after treatment with vehicle or 10% 1,6-hexanediol, respectively, for 1 min), N-IDR$_{FS}$/A9 (tracks 9–10 represent samples after treatment with vehicle or 1,6-hexanediol, respectively), F-IDR$_{WT}$/A9 (track 11) or F-IDR$_{YS}$/A9 (track 12), as well as CTCF ChIP–seq in 293FT cells with N-IDR$_{WT}$/A9 (track 13) or N-IDR$_{FS}$/A9 (track 14). HA and CTCF ChIP–seq signals were normalized to input signals, whereas GFP ChIP–seq, conducted in the spike-in controlled experiments, normalized to the spike-in *Drosophila* chromatin signals (those from antibody of a *Drosophila*-specific histone, H2Av).



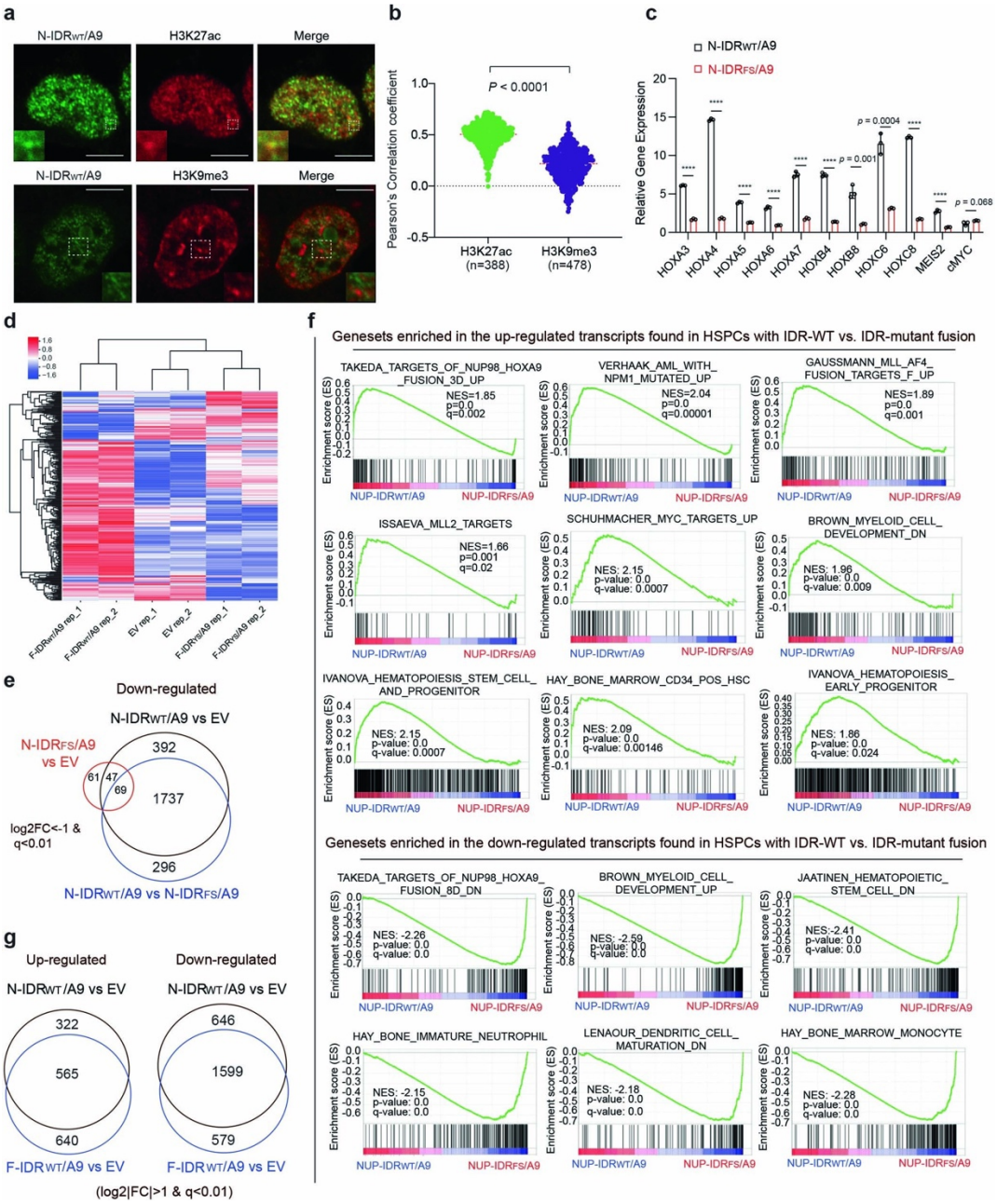
Extended Data Fig. 5.5: Formation of the enhanced and broad super-enhancer-like binding patterns of leukaemia-related chimera transcription factors requires an intact phase-separation-competent IDR. a, **b**, Hockey-stick plot shows distribution of the input-normalized ChIP–seq signals of N-IDR$_{WT}$/A9 (**a**) or H3K27ac (**b**) across all enhancers annotated by H3K27ac peaks (transcriptional start site ± 2.5 kb regions were excluded) in 293FT cells. Dotted line indicates the threshold level set by the ROSE algorithm to call super-enhancers. Relative rankings of super-enhancers associated with some example genes are shown. **c**, Venn diagram illustrates overlap among super-enhancers called based on N-IDR$_{WT}$/A9 and H3K27ac ChIP–seq signals. **d**, **e**, Box plots showing averaged ChIP–seq signals for *k*-means clustered peaks (see Fig. 2b) of the LLPS-competent N-IDR$_{WT}$/A9 (WT; **d**) show a marked reduction in binding after treatment of 293FT stable cells with 1,6-hexanediol (WT+H), relative to treatment with vehicle control (WT+V); this reduction is particularly significant for peak clusters 1–3 shown in Fig. 2b. By contrast, genomic binding of N-IDR$_{FS}$/A9 (FS; **e**) shows general insensitivity to the same treatment of 1,6-hexanediol (FS+H) in comparison to mock (FS+V). Right, averaged ChIP–seq signal distribution profiles are shown for N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 over a 10-kb region in the indicated peak cluster as an example. Box plots as defined in Fig. 3f. **f**, Venn diagram to compare genes associated with the broad super-enhancer-like peaks of N-IDR$_{WT}$/A9 after treatment with 1,6-hexanediol (+H), relative to vehicle control (+V), after treatment for 1 min. **g**, Hierarchical clustered heat maps for the pairwise correlation of ChIP–seq signals between each of the indicated sample. The coefficients were determined by Pearson correlation. HA and GFP represent ChIP–seq for HA-tagged and GFP-tagged chimera transcription factors, respectively.



Extended Data Fig. 5.6: The phase-separation-promoting property within F-IDR is sufficient to induce the enhanced binding of the chimeric transcription factor. a, Heat maps showing the *k*-mean clustering of ChIP–seq signals for chimeric transcription factors that contain the NUP98 IDR (N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9, two panels on the left) or FUS IDR (F-IDR$_{WT}$/A9 and F-IDR$_{YS}$/A9, two panels on the right) reveal a similarly enhanced binding for the LLPS-competent chimera that carries a wild-type form of IDR, relative to its LLPS-incompetent IDR mutant, in 293FT stable expression cells. Note that, although to a lesser degree, the artificially created F-IDR$_{WT}$/A9 fusion also displays a broad, super-enhancer-like binding pattern at the same sites observed with the N-IDR$_{WT}$/A9 fusion. **b**, Pie chart showing percentage distribution of the indicated genomic annotation feature among the ChIP–seq peaks of GFP-tagged F-IDR$_{WT}$/A9 (left) or F-IDR$_{YS}$/A9 (right) in the 293FT stable expression cells. **c**, Heat maps (left) and its averaged ChIP–seq signal distribution profiles (right) for *k*-mean clustered peaks of N-IDR$_{WT}$/A9 in the transformed mouse HPSCs. **d**, Venn diagram showing overlap between the annotated genes associated with the clusters 1–3 of N-IDR$_{WT}$/A9 ChIP–seq peaks detected in the transformed mouse HPSCs (left) and the 293FT stable expression cells (right). Examples of the shared oncogenes are shown below. **e**, IGV views of N-IDR$_{WT}$/A9 ChIP–seq signals (GFP-tagged) at the indicated loci in mouse HSPCs transformed by this chimera. **f**, ChIP–qPCR to assess the binding of GFP-tagged N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9 at *CCL15* (a negative control region), *PBX3* and *HOXA9* in the 293FT stable cells after treatment with 10% 1,6-hexanediol for 1 min (+H), relative to mock (+V). ChIP signals, normalized to those of input, are presented as mean ± s.d. of three replicate experiments. **g**, ChIP–qPCR to assess the binding of GFP-tagged F-IDR$_{WT}$/A9 or F-IDR$_{YS}$/A9 at *CCL15* (a negative control region), *PBX3* and *HOXA9* in the 293FT stable cells. ChIP signals, normalized to those of input, are presented as mean ± s.d. of three replicates.

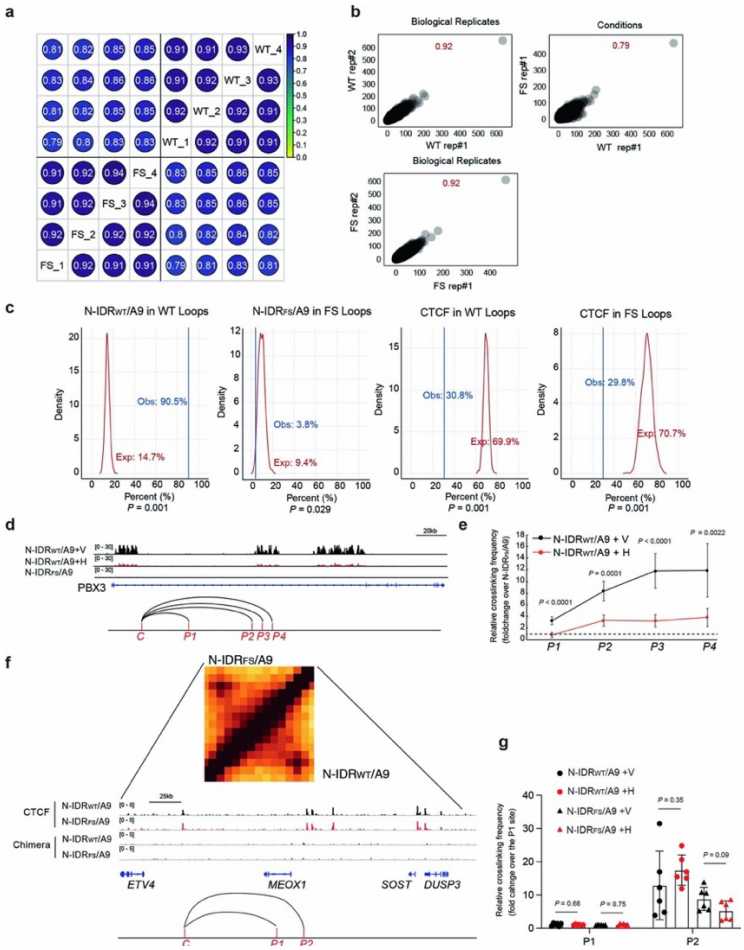


Extended Data Fig. 5.7: Single-molecule tracking shows that phase-separation-competent N-IDR$_{WT}$/A9 proteins behave with less dynamic characteristics, compared with phase-separation-incompetent N-IDR$_{FS}$/A9. a, Representative images of single-molecule particles identification in an N-IDR$_{WT}$/A9-expressing cell, either the original captured image (left) or after processing to remove background (right). Scale bars, 5 μm. **b**, **c**, Single-particle tracks for mean speed (**b**) and mean displacement (**c**) of either N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9 single molecules within the temporally registered reference frame binned into 1-s intervals. **d**, **e**, Displacement (**d**) and mean velocity (**e**) of single-particle tracks indicate that N-IDR$_{WT}$/A9 with the LLPS-competent IDR (WT) is less mobile and navigates nuclear space at a slower rate than its LLPS-incompetent IDR mutant (FS). Dots indicate mean values in a single cell. Line indicates one standard deviation. *P* values determined by two-sided *t*-test. **f**, **g**, The diffusion coefficient for chromatin-bound (**f**) and freely diffusing states (**g**) of N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9, calculated based on single-molecule tracking studies of its 293FT stable expression cells. *P* values determined by two-sided *t*-test.

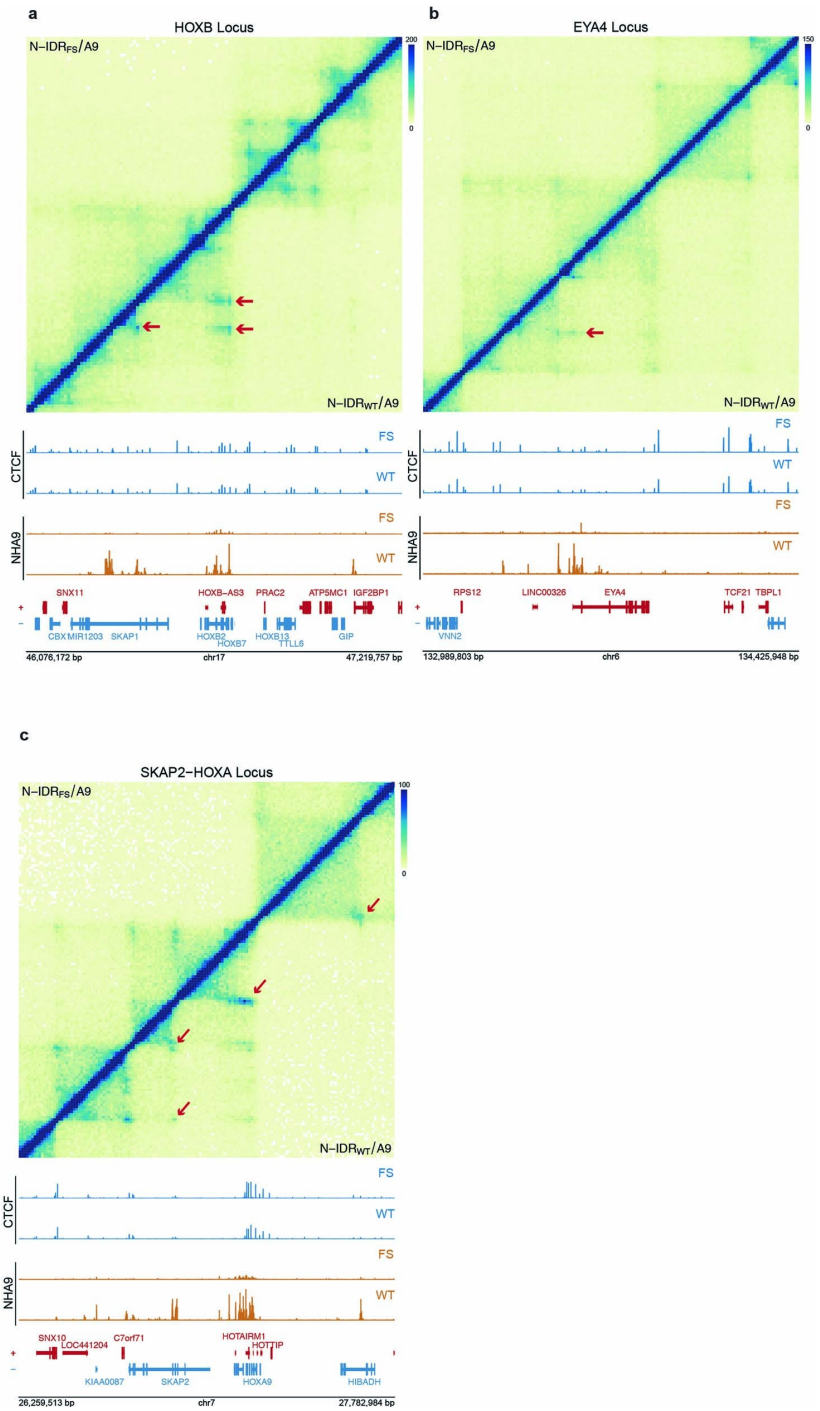


The caption for this figure continues on the next page.

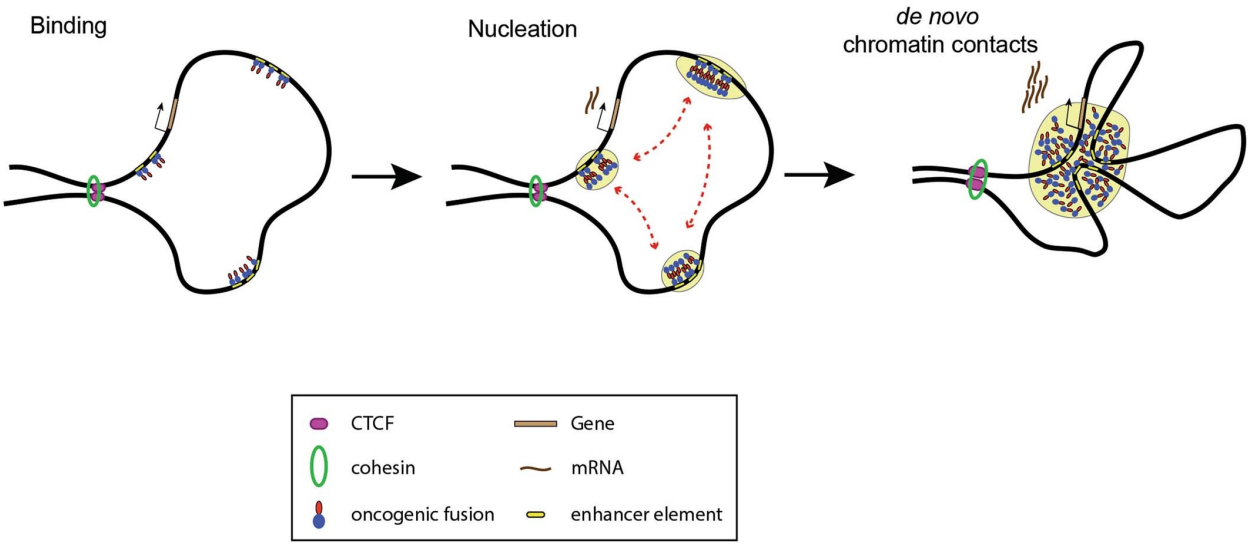
Extended Data Fig. 5.8: An LLPS-competent IDR within the leukaemia-related transcription factor chimera is essential for potentiating transcriptional activation of the downstream oncogenic gene-expression program. a, Fixed cell immunostaining for the 3×HA-3×Flag-tagged N-IDR$_{WT}$/A9 (left; anti-Flag) and the indicated histone modification (middle) in the 293FT stable expression cells. Top panels show the enlarged images of an example region within the white dotted box shown in the bottom panels, in which the transcription factor chimera is co-localized with H3K27ac (top) and not H3K9me3 (bottom). Scale bars, 10 μm. **b**, Pearson's correlation coefficient values between N-IDR$_{WT}$/A9 and the indicated histone modification. The red dotted line indicates the calculated average value of each plot. The calculated means (red dotted lines) were compared with an independent two-tailed Student's *t*-test. n, the number of cells analysed. **c**, RT–qPCR to assess the effect of phase separation in target gene expression in 293FT cells. All of the tested *HOX* and *MEIS2* genes are direct targets of both N-IDR$_{WT}$/A9 and N-IDR$_{FS}$/A9 based on ChIP–seq, whereas *MYC* is not and serves as a negative control. Note that LLPS-competent N-IDR$_{WT}$/A9 induces significantly more upregulation of target genes, relative to LLPS-incompetent N-IDR$_{FS}$/A9. PCR signals were normalized first to those of an internal control (18S RNA) and then to vector-expressing cells and presented as mean ± s.d. of three replicated experiments. ***P < 0.001; ****P < 0.0001; two-sided *t*-test. n.s., not significant. **d**, Heat map illustrating relative expression of the 374 genes that show significant upregulation post-transduction of F-IDR$_{WT}$/A9, compared to empty vector and its IDR-mutant form (F-IDR$_{YS}$/A9), in 293FT stable expression cells. **e**, Venn diagrams showing the overlap of the significantly downregulated genes identified 7 days after transduction of the indicated construct into mouse HPSCs. **f**, Gene set enrichment analysis (GSEA) shows that, compared with that of N-IDR$_{FS}$/A9, the expression N-IDR$_{WT}$/A9 in mouse HPSCs is positively correlated with the indicated leukaemia- or HSPC-related gene sets (top) and negatively correlated with the indicated differentiation-related gene sets (bottom). The P value was calculated by an empirical phenotype-based permutation test; the false discovery rate (q) is adjusted for gene set size and several hypotheses testing whereas the P value is not. **g**, Venn diagrams showing the overlap of the significantly upregulated (left) or downregulated (right) genes identified after transduction of the indicated construct into mouse HPSCs.



Extended Data Fig. 5.9: Hi-C mapping reveals that a phase-separation-competent IDR within NUP98–HOXA9 is required to induce formation of CTCF-independent chromatin loops at the leukaemia-related genomic loci. a, Matrix of Pearson correlation coefficients of loop counts among and between biological replicates of N-IDR$_{WT}$/A9 (WT; *n* = 4 replicates) or N-IDR$_{FS}$/A9 (FS; *n* = 4 replicates) conditions. Numbers following WT or FS indicate biological replicate for that condition. **b**, Example correlation plots of loop counts between biological replicates and conditions. **c**, All loops were partitioned into either WT- or FS-specific loops and split into separate loop anchors. Loop anchors were then intersected with ChIP–seq peaks of N-IDR/A9 or CTCF. The percentage of observed (Obs.) overlaps for each feature is shown as a vertical blue line. The red line shows the expected (Exp.) distribution of overlaps as determined by randomly sampling loop anchors and calculating the overlap of each feature 1,000 times. *P* values were determined by summing the number of expected values greater than (or less than if the observed value was less than the mean) the observed value for that feature. **d**–**g**, 3C–qPCR assays measuring the change in crosslinking frequency of either an N-IDR$_{WT}$/A9-specific loop at the *PBX3* locus (**d**, **e**) or a CTCF-dependent loop (**f**, **g**; at Chr17 (41604677–41883642)) after treatment of 293FT stable cells with 10% 1,6-hexanediol for 1 min (+H), relative to mock (+V). The IGV view panels at **d** and **f** show the indicated ChIP–seq signals, with positions of the used 3C–PCR primers labelled under IGV tracks. PCR was performed using the same constant forward primer (C) paired with a differently numbered reverse primer (P1 to P4) at each locus tested. Panels **e** and **g** are plotted with signals of 3C–qPCR measuring the relative crosslinking frequency at *PBX3* (**d**, **e**) or a Chr17 locus with CTCF loop (**f**, **g**) before (V) and after (H) treatment with 1,6-hexanediol. Signals in **e** are normalized to those of the N-IDR$_{FS}$/A9-expressing cells (*n* = 3 replicated experiments). *P* values were determined by two-sided *t*-test. Data are mean ± s.d. of three or six replicates.



Extended Data Fig. 5.10: Hi-C mapping reveals the chromatin loops specific to cells with the LLPS-competent NUP98–HOXA9, compared with the LLPS-competent mutant, at leukaemia-relevant gene loci. Views for Hi-C mapping, RNA-seq and ChIP–seq for CTCF, N-IDR/A9 and H3K27ac at the *HOXB* (**a**), *EYA4* (**b**), and *SKAP2-HOXA* loci (**c**) in 293FT stable cells expressing either N-IDR$_{WT}$/A9 (WT) or N-IDR$_{FS}$/A9 (FS). Hi-C mapping views (top) show results from the N-IDR$_{WT}$/A9 or N-IDR$_{FS}$/A9 expressing cells (bottom and top diagonal, respectively). Corresponding ChIP–seq and gene tracks are shown below each Hi-C plot. N-IDR$_{WT}$/A9 loops are indicated by red arrows.



Extended Data Fig. 5.11: Model illustrating requirement of LLPS-competent IDR within NUP98–HOXA9 for leukaemogenesis and activation of the oncogenic gene-expression program. The LLPS-competent IDR contained with NUP98–HOXA9 is crucial for promoting long-distance chromatin looping between proto-oncogene promoter and enhancers, which thus induces an oncogenic gene-expression program and malignant development.

Chapter 6: Temporal analysis suggests a reciprocal relationship between 3D chromatin structure and transcription[1]

6.1. Introduction

3D chromatin structure is thought to play a critical role in gene expression, cellular identity, and organismal development by modulating contact frequencies between gene promoters and distal regulatory elements such as enhancers (Dekker and Mirny 2016). Alterations in 3D chromatin architecture have been associated with developmental abnormalities and human disease (Spielmann, Lupiáñez, and Mundlos 2018; Akdemir et al. 2020; Johnstone et al. 2020; Rosencrance et al. 2020; Ahn et al. 2021). Despite growing knowledge regarding the proteins and molecules that govern 3D chromatin architecture, the relationship between 3D chromatin architecture and gene transcription is less certain. Although some functional connections between chromatin interactions and transcription have been established, the degree to which 3D chromatin structure shapes—or is shaped by—transcription remains unclear.

The continued development of chromatin conformation capture (3C)-based technologies has provided valuable insights into the mechanisms driving 3D chromatin structure (Dekker et al. 2002; Dostie et al. 2006; Simonis et al. 2006; Fullwood et al. 2009; Lieberman-Aiden et al. 2009; Rao et al. 2014; Mumbach et al. 2016). In particular, genome-wide approaches, including Hi-C, have revealed tens of thousands of loops throughout the human genome, many of which connect regulatory elements such as enhancers to gene promoters. With some notable exceptions (J. Lee et al. 2017; Monahan, Horta, and Lomvardas 2019; Ahn et al. 2021), the majority of loops are bound at each anchor by CCCTC-binding factor (CTCF) and are formed via loop extrusion by the cohesin complex (Heidari et al. 2014; Sanborn et al. 2015). Mapping these loops across cell types and biological conditions has revealed cell-type-specific looping events that often correlate with differences in gene transcription (Rao et al. 2014; Phanstiel et al. 2017; D'Ippolito et al. 2018; Winick-Ng et al. 2021).

[1] The work in this chapter has been previously published. The citation is: Reed, Kathleen S. M., Eric S. Davis, Marielle L. Bond, Alan Cabrera, Eliza Thulson, Ivana Yoseli Quiroga, Shannon Cassel, et al. 2022. "Temporal Analysis Suggests a Reciprocal Relationship between 3D Chromatin Structure and Transcription." *Cell Reports* 41 (5): 111567.

Despite these advances, the mechanisms and degree to which looping drives transcriptional changes are far less certain. A widely held hypothesis is that chromatin loops facilitate transcriptional activation by increasing the frequency of interactions between enhancers and gene promoters; however, studies that removed looping genome wide have produced conflicting results. Acute depletion of cohesin in a human cancer cell line was sufficient to eliminate cohesin-bound loops but had only a modest effect on transcription, casting doubt on the importance of DNA looping for transcriptional control (Rao et al. 2017). In contrast, deletion of the cohesin loading factor NIPBL (Nipped B-like protein) in mouse liver cells *in vivo* induced transcriptional changes of thousands of genes (Schwarzer et al. 2017). Depletion of cohesin and CTCF has been shown to significantly affect the ability of human and mouse macrophages to mount a proper transcriptional response to the endotoxin lipopolysaccharide (LPS) (Cuartero et al. 2018; Stik et al. 2020), which suggests that loops might be specifically important for regulating changes to (as opposed to maintenance of) transcriptional signatures.

Mounting evidence also suggests that transcription can shape 3D chromatin structure, although the exact relationship remains unclear. Several studies have shown that transcription can displace cohesin and condensin complexes (Lengronne et al. 2004; Busslinger et al. 2017; Brandão et al. 2019). For example, knocking down CTCF and the cohesin unloader WAPL (Wings apart-like protein homolog) causes cohesin to accumulate at the 3′ end of highly transcribed genes, suggesting that cohesin may be relocated by transcription in the absence of boundary elements (Busslinger et al. 2017). Other work has demonstrated that transcription-induced displacement of structural maintenance of chromsosome (SMC) complexes results in altered chromatin structure. Macrophages infected with influenza A, which inhibits transcription termination, show readthrough transcription that displaces cohesin at CTCF binding sites, repositioning it toward the 3′ end of genes and disrupting existing chromatin structure (Heinz et al. 2018). Fibroblasts undergoing senescence exhibit *de novo* transcription-dependent cohesin peaks at the 3′ end of select genes, resulting in newly formed loops (Olan et al. 2020). Alternatively, knockdown of RNA polymerase II (-RNAPII) causes new loops to form at CTCF anchors and many enhancer-promoter loops to be lost (S. Zhang et al. 2022). These findings are supported by *in vitro* experiments performed on DNA ‘‘curtains,’’ showing that RNAP or other translocases can push cohesin; however, more recent studies suggest that

molecules as large as 200 nm may be able to pass through SMC complexes (Davidson et al. 2016; Stigler et al. 2016; Pradhan et al. 2021).

One approach to dissect causal relationships between looping and transcription while circumventing genome-wide perturbations with potential knock-on effects is to quantify changes in looping, transcription, and other regulatory features across biological time courses. 3C-based time courses of biological transitions have produced valuable insights into the dynamics of 3D chromatin architecture (Bonev et al. 2017; Abramo et al. 2019; Bertero et al. 2019; H. Zhang et al. 2019; Yanxiao Zhang et al. 2019; Yang et al. 2020; Furlan-Magaril et al. 2021; Vilarrasa-Blasi et al. 2021). For example, D'Ippolito et al. (2018) characterized differential looping at 4 time points after glucocorticoid treatment and found that, on average, loops changed maximally at 4 h, whereas gene expression changed maximally at 9 h. This timing is consistent with a regulatory relationship, but the relatively broad spacing of time points made temporal ordering of individual pairs of loops and genes more difficult. Beagan et al. (2020) used chromosome conformation capture carbon copy (5C) to identify differential looping events in activated neurons in time frames as short as 20 min (Beagan et al. 2020); however, these studies focused on just a handful of genomic loci.

To characterize the temporal order of regulatory events and infer potential causal relationships, we mapped 3D chromatin architecture, histone H3K27 acetylation, chromatin accessibility, and gene expression across eight time points of macrophage activation. Narrowly spaced timepoints allowed correlation and temporal ordering of events at a locus-by-locus level. These analyses provided insights into the putative causal relationships between these events, which were consistent with a reciprocal relationship between chromatin looping and gene transcription.

6.2. Results

6.2.1. LPS + IFN_γ triggers genome-wide changes in chromatin looping, enhancer acetylation, and gene expression

To understand how chromatin loops and enhancers work together to regulate gene transcription in response to external stimuli, we used an eight-point time course of human macrophage activation **(Figure 1A)**. Human macrophages derived from the THP-1 monocytic cell line were stimulated with 10 ng/mL LPS and 20 ng/mL interferon-gamma (IFN_γ) and collected at eight time points (0, 0.5, 1, 1.5, 2, 4, 6, and 24 h). At each time point, we profiled 3D chromatin structure using *in situ* Hi-C (Rao et al. 2014), putative enhancer

activity using chromatin immunoprecipitation sequencing (ChIP-seq) targeting histone 3 lysine 27 acetylation (H3K27ac), chromatin accessibility using assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al. 2013), and gene expression using RNA sequencing (RNA-seq).

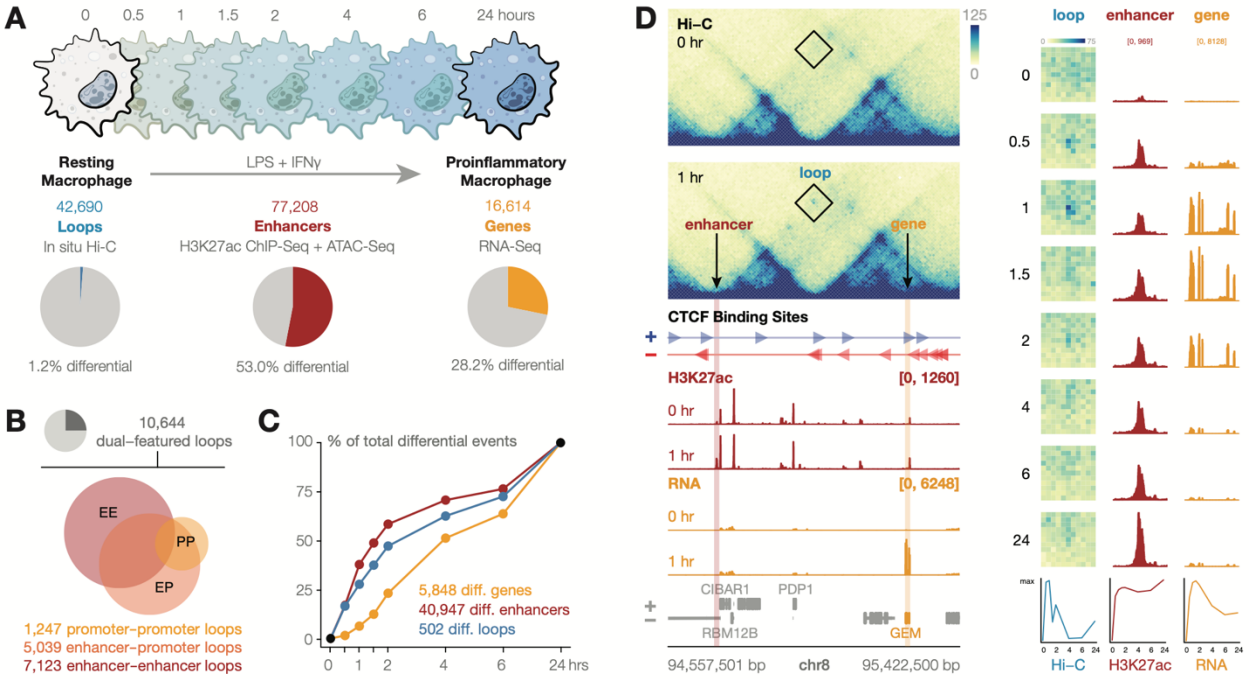


Figure 6.1. Multi-omics time course of macrophage activation physically and temporally connects regulatory events. (A) Experimental design to identify changes to 3D chromatin structure, enhancers, and gene expression across eight time points during macrophage proinflammatory activation. Differential chromatin loops were identified using Hi-C from 5 technical replicates across 3 biological replicates for a total of roughly 2 billion contacts per time point. Differential H3K27ac peaks and genes were identified from 2 biological replicates of H3K27ac ChIP-seq and RNA-seq, respectively. **(B)** Fraction and number of loops that connect two distal elements. **(C)** A cumulative sum of differential events identified by each time point reveals the relative timing of changes to genes, loops, and enhancers. **(D)** Intersecting differential chromatin loops, enhancers, and genes provides the regulatory context of transcriptional changes. Predicted CTCF binding sites obtained from the *CTCF* R package are also shown for this locus. At this region, a 570-kb loop connects the promoter of the *GEM* gene to a distal enhancer. The enhancer's activity peaks 30 min before gene expression but remains high throughout the treatment, whereas the loop connecting them fades alongside gene expression after 2 h. Line plots show the log2 fold change of Hi-C, enhancer H3K27ac, and *GEM* expression signal over the full time course. Hi-C is shown at 10-kb resolution for the full region and 5-kb resolution for the loop magnifications. Scales indicate read signal (RNA, ChIP) or KR-normalized counts (Hi-C).

With eight time points and roughly 2 billion contacts per Hi-C map, this is one of the most comprehensive characterizations of 3D chromatin changes to date (D'Ippolito et al. 2018; Rowley et al. 2020). To comprehensively catalog long-distance chromatin interactions, we combined our maps from each time point into a single, ultra-deep ''mega'' map comprising 24.5 billion reads and 15.6 billion chromatin contacts **(Figures S1A and S1B)**. This increased read depth provided the power to identify over 10,000

additional loops at 5-kb resolution that were undetectable at the resolution of individual time points **(Figure S1C)**. Loops from each timepoint as well as the Mega map were then merged, combining any loops with both anchors within 20 kb to provide 42,690 total loops for this study.

To identify potential regulatory connections among these loops, we classified putative enhancers (hereafter called enhancers) as loci with overlapping ATAC-seq and histone H3K27 acetylation peaks that did not overlap gene promoters **(STAR Methods)**. Intersecting these enhancers with chromatin loops revealed 5,039 enhancer-promoter loops **(Figure 1B)**. The regulatory activity of enhancers was inferred via quantification of histone H3K27ac at each enhancer. Finally, we used stranded rRNA-depleted RNA-seq at each time point to quantify the potential effects of these loops and enhancers on gene expression.

Differential analysis using the DESeq2 package (Love, Huber, and Anders 2014) identified statistically significant genome-wide alterations in DNA looping, enhancer activity, and gene expression at each time point **(Figures 1C, S2A, Tables S1, S2, and S3[2])**. The transcriptional changes we observed are consistent with previously established profiles of inflammatory activation **(Figures S2B and S2C)**. Only 1.2% (220 up, 282 down) of loops were detected as differential for at least one time point compared with 53.0% (21,858 up, 19,089 down) of enhancers and 28.2% (3,025 up, 2,823 down) of genes. Of these 502 differential loops, 79 were detected only at intermediate time points and were not visible at 0 or 24 h, highlighting the insights offered from this level of temporal resolution. On average, enhancers and loops changed faster than genes, with 58.4% of differential enhancers and 47.2% of differential loops changing significantly within the first 2 h of LPS + IFN_{γ} treatment compared with only 23.1% of genes **(Figure 1C)**. This temporal lag between changes in loops and enhancers compared with changes in gene expression is consistent with our understanding of loops and enhancers as regulators of gene transcription and highlights the power of using temporal analysis to generate hypotheses about causal relationships (Arner et al. 2015; Rowley and Corces 2018; Schoenfelder and Fraser 2019; Zheng and Xie 2019).

Integrating the resulting multi-omics data provided insights into gene-regulatory mechanisms of macrophage activation. An example of this concept can be seen at the *GEM* locus on chromosome 8 **(Figures 1D, S2D, and S2E)**. The *GEM* gene is transiently upregulated during LPS + IFN_{γ} treatment, with expression increasing at 30 min and peaking at 1.5 h. An enhancer 570 kb downstream of the *GEM*

[2] Supplementary tables are available online at https://doi.org/10.1016/j.celrep.2022.111567.

promoter becomes acetylated and physically looped to the promoter of *GEM* after only 30 min of treatment, coinciding with increased expression of GEM. Although the enhancer remains acetylated throughout the time course, a precipitous drop in contact frequency between the enhancer and promoter at 1.5 h is followed by a similar decrease in gene expression 30–60 min later. These data are consistent with a model in which contacts between active enhancers and gene promoters play a causal role in transcriptional changes. Throughout the rest of this paper, we explore these relationships quantitatively on a genome-wide scale.

6.2.2. Looped enhancer-promoter pairs exhibit ordered and correlated changes in acetylation and expression

The importance of chromatin looping for transcriptional regulation remains unclear; studies disrupting chromatin loops comprehensively throughout the genome have produced mixed results (Rao et al. 2017; Schwarzer et al. 2017; Cuartero et al. 2018; Stik et al. 2020). Ablation of loops in the human colorectal cancer cell line HCT-116 only altered the expression of a handful of genes (Rao et al. 2017). In contrast, loss of loops in murine liver cells and macrophages responding to LPS induced thousands of transcriptional changes (Schwarzer et al. 2017; Cuartero et al. 2018; Stik et al. 2020). Differences in biological systems, cellular contexts, and even the method of loop disruption could potentially explain the conflicting findings.

We investigated our data to see whether it supported a role of looping in gene regulation in response to external stimuli. Our analyses were based on the assumption that, if loops play a role in transcriptional control, then looped enhancer-gene pairs should exhibit correlated changes in histone H3K27ac and gene expression. Because only a small fraction of loops change over time, all loops were used to connect enhancers to promoters regardless of differential status. In total, this involved 5,039 enhancer-promoter loops featuring 4,093 unique genes, 1,483 of which were differential. In total, 25.4% of differential genes were connected to a distal enhancer via a chromatin loop. We investigated the temporal patterns of these looped enhancer-promoter pairs and compared them with sets of non-looped enhancer-promoter pairs of similar genomic distance (base pairs) or contact frequency (Hi-C measurements). These matched non-looped sets were identified using the “matchRanges” function available from the *nullranges* R/Bioconductor package **(Figure 2A)** (Davis et al. 2022). Because of the absence of long-range chromatin loops at these enhancer-promoter pairs, contact-matched pairs had Hi-C contact frequencies similar to looped pairs but were much closer in base pairs **(Figure 2B, dark gray)**. In contrast, distance-matched

pairs were separated by a similar number of base pairs as the looped pairs but had much lower Hi-C contact frequencies (Figure 2B, light gray). Comparisons between these non-looped matched sets and the looped enhancer-promoter pairs allowed us to isolate the effect of contact frequency and distance independently and investigate the additional effect of chromatin looping by adjusting for these features.

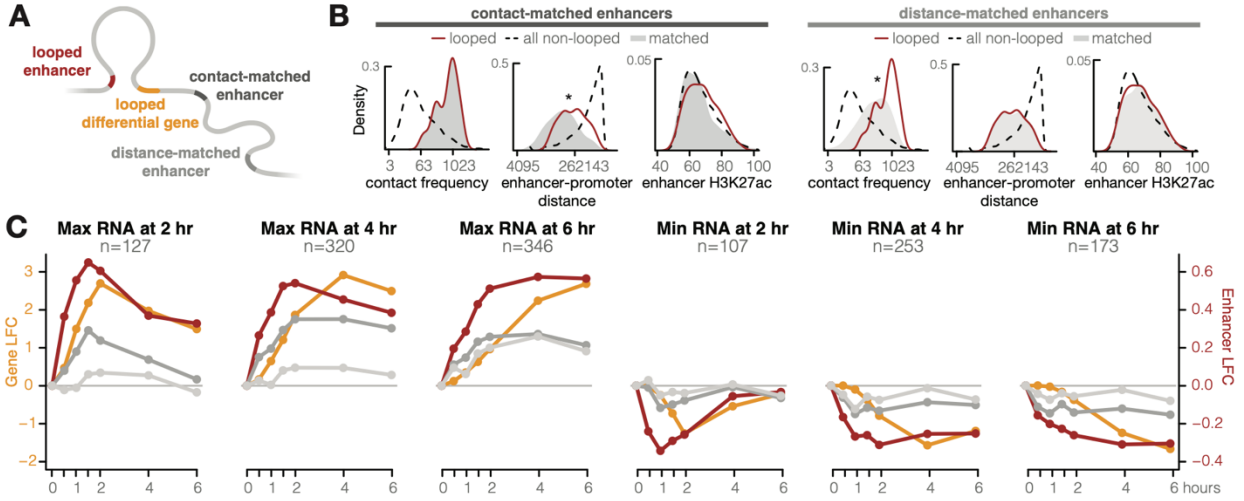


Figure 6.2. Enhancer acetylation and gene expression correlate most highly at looped enhancer-promoter pairs. **(A)** Distal enhancers looped to the promoters of differential genes were compared with matched enhancers of equal H3K27ac and contact frequency (dark gray) or distance (light gray). **(B)** Representative distributions of contact- and distance-matched enhancers compared with the pool of non-looped enhancer-promoter pairs and the looped subset. Compared with looped pairs, contact-matched enhancers are closer on average in base pairs (Wilcoxon rank-sum test, $p < 10^{-6}$), whereas distance-matched enhancers are in less frequent contact (Wilcoxon rank-sum test, $p < 10^{-11}$). Both sets of matched enhancers have H3K27ac levels similar to the looped pairs (Wilcoxon rank-sum test, p = 0.028, 0.86). **(C)** Average log2 fold change of gene expression (gold) for genes reaching minimum or maximum fold change at 2, 4, or 6 h compared with log2 fold change of the H3K27ac levels of their looped enhancers (red), contact-matched enhancers (dark gray), and distance-matched enhancers (light gray). Looped enhancers correlate significantly with changes in gene expression, to a larger extent than matched enhancers. Contact-matched enhancers tend to correlate better than distance-matched enhancers at upregulated genes. Changes in distal enhancer H3K27 acetylation precede changes in gene expression among all time scales and among up- and downregulated genes.

To explore the correlation of enhancers and genes over time, we clustered our differential genes based on the time point when they exhibited their maximal up- or downregulation with respect to the 0 h time point and plotted their average normalized expression **(Figure 2C, yellow lines)**. Only clusters peaking at intermediate time points and with more than 100 genes are shown. For each gene cluster, we identified enhancers that were connected to those genes via a chromatin loop and plotted their average normalized histone H3K27ac signal **(Figure 2C, red lines)**. All 6 clusters revealed a clear correlation between histone H3K27ac and gene expression at looped enhancer-promoter pairs, supporting the idea that looped pairs are functionally connected. Interestingly, the changes in acetylation preceded changes in

gene expression by 30–60 min. This lag is also seen in changes in promoter acetylation **(Figure S3A, black lines)**, and is consistent with enhancer activation causing changes to gene expression.

Despite this evidence, it is important to consider that chromatin loops occur over relatively short distances (median, ~390 kb), and at such short distances, even non-looped enhancers and promoters exhibit elevated chromatin contact frequencies compared with randomly selected enhancers and genes across the genome. Therefore, the correlation between looped enhancers and promoters we observe could be explained by genomic distance alone. To determine whether looped enhancer-promoter pairs exhibited higher correlation than expected given their genomic distance, we compared looped enhancer-promoter pairs with non-looped enhancer-promoter pairs that were matched for genomic distance **(Figure 2A)**. Distance-matched, non-looped enhancer-promoter pairs exhibited some degree of correlation **(Figure 2C, light gray lines)**; however, the correlation was weaker than that observed at looped enhancer-promoter pairs. Thus, distance alone does not account for the enhancer-promoter correlations observed at loop anchors and offers further support for the functional role of loops in enhancer-based gene regulation.

One explanation for how loops exhibit transcriptional control is by increasing contact frequencies between enhancers and their target genes. To determine whether looped enhancer-promoter pairs exhibited a higher correlation than expected given their contact frequency, we compared looped enhancer-promoter pairs to non-looped enhancer-promoter pairs that were matched for contact frequency. Surprisingly, although contact-matched pairs exhibited a stronger correlation than distance-matched pairs, the correlation was still weaker than that observed at looped enhancer-promoter pairs **(Figure 2C, dark gray lines)**. We confirmed these results using data from our previously published study of monocyte differentiation **(Figures S3B and S3C)** (Phanstiel et al. 2017). There too, looped enhancer-promoter pairs exhibited better correlation than enhancer-promoter pairs that were matched for distance or contact frequency. This was surprising and suggests that the presence of a chromatin loop may facilitate a functional regulatory connection through mechanisms beyond simply increasing their frequency of physical proximity. We explore some possible explanations for this in the discussion.

6.2.3. Changes in gene expression exhibit a directional bias at differential loop anchors

Given the correlation we observed between acetylation and gene expression at opposite ends of chromatin loops, we hypothesized that changes in looping would be associated with altered transcription of genes at loop anchors and that the directionality of changes in expression would match that of the changes in looping. To test this, we used k-means clustering to identify four categories of differential looping: gained early, gained late, lost early, and lost late. Generally, ‘‘early’’ differential loops changed within the first 2 h of treatment, whereas ‘‘late’’ differential loops did not change until 4 h or beyond, many unique to the 24-h time point. Examples of loops from each cluster are shown in **Figure 3A**. Differential loops spanned approximately 170–200 kb on average, with the exception of gained late loops, which were much larger with an average length of 610 kb **(Figure S4A)**. Compartmental analysis revealed that these gained late loops were also closer to the B compartment than other differential loop classes **(Figure S4B)**. Although all loops on average fall into the more transcriptionally active A compartment, differential loops shifted farther to the A compartment over the course of activation.

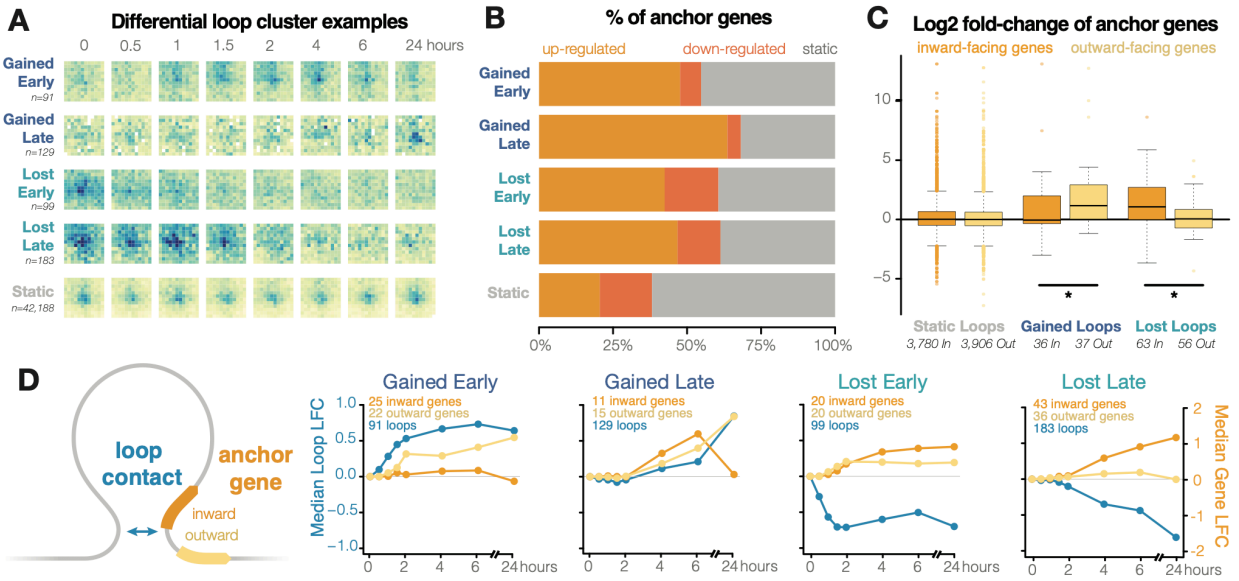


Figure 6.3. Upregulated genes anchored at differential loops exhibit directionality bias. (A) 502 differential loops were clustered (k-means) according to their timing and direction. Generally, ‘‘early’’ changes occurred within the first 2 h, and ‘‘late’’ changes occurred within 4 h and beyond. Representative loops are shown for each cluster (5-kb resolution). **(B)** The anchors in all differential loop clusters are enriched for upregulated genes. **(C)** Distributions of log2 fold changes of genes with promoters in the anchors of static and differential loops. Anchor genes were classified by whether they are oriented toward (inward, orange) or away from (outward, yellow) the center of the loop. Among genes at gained loop anchors, the fold change of outward-facing genes is significantly higher than of inward-facing genes, whereas the opposite trend is seen among genes at lost loops (Wilcoxon rank-sum test, $p < 0.05$). **(D)** Average log2 fold change of differential loops (blue) and inward- and outward-facing genes (orange, yellow) with promoters overlapping those loop anchors.

Next we calculated the percentage of genes at each set of loop anchors that were significantly up- or downregulated in response to LPS + IFN_γ treatment **(Figures 3B, S4C, and D)**. Anchor genes were defined by overlapping gene promoters with loop anchors **(STAR Methods)**. Gained loop anchors were enriched for the promoters of upregulated genes (permutation test, n = 10,000, $p < 0.05$) and shifted toward the A compartment during treatment **(Figure S4B)**. These observations are consistent with findings from previous work by our lab and others that have associated increased looping with increased transcription of anchor genes (Rao et al. 2014; Phanstiel et al. 2017) and generally support a causal role of looping in transcriptional control. However, lost loops were also associated with increased transcription of anchor genes and a shift toward the A compartment (permutation test, n = 10,000, $p < 0.05$). Although surprising, this is consistent with data from Rao et al. (2017) that showed that removal of DNA loops is not necessarily accompanied by decreased transcription of anchor genes (Rao et al. 2017).

To explore this further, we separately analyzed anchor gene expression based on whether genes were oriented toward or away from the center of the loop. Intriguingly, genes at the anchors of gained and lost loop classes exhibited different directional biases **(Figures 3C and S4E)**. At gained loops, outward-oriented anchor genes exhibited significantly more increased expression than inward-oriented genes (Wilcoxon rank-sum test, $p < 0.05$). In contrast, at lost loops, inward-oriented anchor genes exhibited more increased expression than outward-oriented genes (Wilcoxon rank-sum test, $p < 0.05$). Similar trends can be seen using differential loops from monocyte-to-macrophage differentiation **(Figure S4F)**. To investigate this further, we examined the temporal profiles of differential loops and anchor genes. For each loop cluster, we calculated the average fold change of inward- and outward-facing anchor genes **(Figure 3D)**. At gained loops, the contact frequency and transcription of anchor genes were positively correlated over time, particularly for loops oriented away from the center of the loop. The gained early loops exhibited increased contact frequency 30–60 min prior to the increased transcription of outward-oriented anchor genes, which is consistent with the notion of loops playing a causal role in gene expression. We see a similar lag between increases in compartment and transcription on a coarser level **(Figure S5)**. The gained late loops showed correlated changes in outward-oriented anchor gene expression, but because they changed most drastically between 6 and 24 h, the time points were not close enough to observe a temporal lag. In contrast,

at lost loops, contact frequency and transcription of anchor genes were inversely correlated, particularly for loops oriented toward the center of the loop.

6.2.4. Lost loops are associated with high levels of transcription within loop boundaries

One possible explanation for the directional biases we observe at differential loop anchors is that transcription may be antagonistic to loop extrusion and that high levels of transcription at loop anchors, or within the loop itself, may destabilize loop extrusion complexes. This would agree with several previous studies highlighting the ability of RNAP to push and/or displace cohesin (Busslinger et al. 2017; Heinz et al. 2018).

To determine whether antagonism between transcription and loop extrusion could explain the increased expression we observed at lost loop anchors, we explored the absolute and relative levels of transcription occurring within the boundaries of differential loops. As with anchor genes, the sets of genes within the bounds of loops of the various clusters were largely unique, with minimal overlap **(Figures S4G and S4H)**. Because the majority of transcription occurs at introns, which are generally not captured in our RNA-seq data, we devised an inferred transcription score (ITS) to roughly estimate the levels of transcription for every 10-kb bin in the genome using our RNA-seq data **(STAR Methods)**. Briefly, the transcript per million (TPM) value for each gene was assigned to every genomic bin covered by the gene body. Values were summed for bins that overlapped multiple genes.

Using our ITSs, we observed that gained loops have relatively low levels of internal transcription at all time points (average ITS ≤ 50; **Figure 4A**). In contrast, decreasing loops achieved much higher average levels of internal transcription during the time course (**Figure 4A**; ITS > 50 at most time points), and the amount of internal transcription is inversely correlated with changes in loop strength (R^2 is −0.59 for lost early and −0.99 for lost late loops). To determine how big of a change in ITS was required to observe a decrease in loop strength, we explored how the changes in ITS within a loop correlated with loop fold change. Loops with a mean increase in ITS of 10 or more exhibited a statistically significant decrease in loop strength (Wilcoxon signed-rank test, $p < 0.01$; **Figure 4B**). Visualizing transcription relative to the loop boundaries **(Figure 4C)** confirmed these findings. Transcription was enriched outside of the boundaries of gained loops (in the 50 kb upstream of the upstream anchors, or downstream from the downstream

anchors). In contrast, transcription was enriched between the anchors of lost loops, within the loop bounds. These data are consistent with a model in which high levels of transcription antagonize loop extrusion.

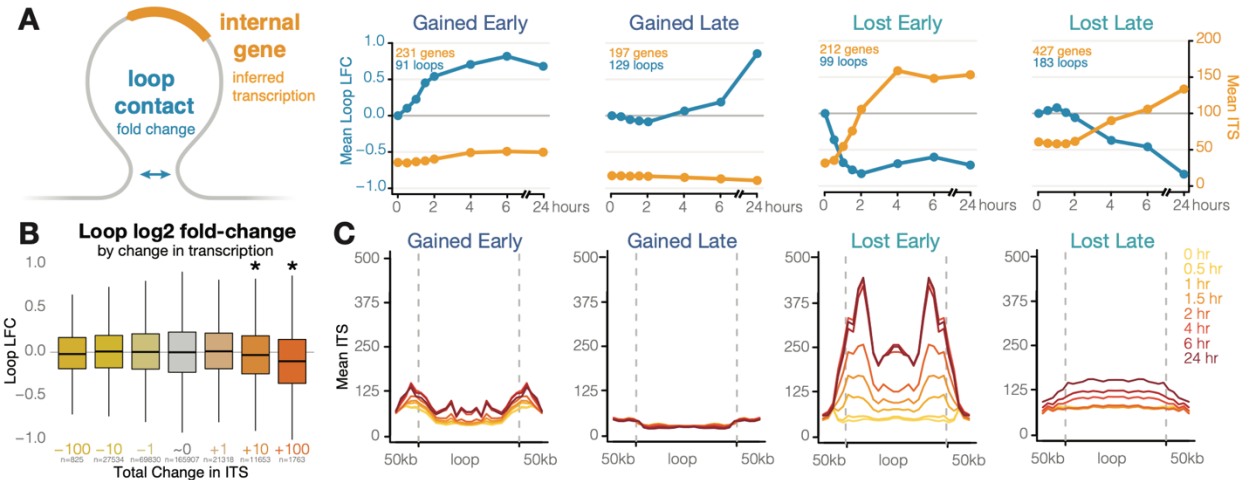


Figure 6.4. Lost loops are characterized by high levels of internal transcription. (A) Log2 fold change of differential loops (blue) and average internal inferred transcription score (ITS; gold) for loops of each cluster. Gained loops have lower levels of internal transcription than lost loops, and the temporal dynamics of changes in transcription are anti-correlated with changes in loop strength among lost loops. For each loop, ITS was calculated and averaged across 2 biological replicates. **(B)** Binning loops based on their change in internal transcription shows significant weakening of loops that gain 10 or more ITSs per 10 kb (Wilcoxon signed-rank test, $p < 10^{-20}$). **(C)** Average ITS within and 50 kb beyond loop boundaries. Transcription is highest at and beyond loop anchors in gained early loops, low among gained late loops, and localized within loop bounds in lost loops.

This suggests that the causal arrow between looping and transcription might point both ways: DNA loop formation may contribute to increased transcription of target genes, but very high levels of transcription could contribute to the weakening loops by antagonizing loop extrusion, as observed previously (Busslinger et al. 2017; Heinz et al. 2018; Brandão et al. 2019; B. Gu et al. 2020; Banigan et al. 2022; Leidescher et al. 2022). An example of these potential phenomena can be seen at the *GBP* locus (**Figure 5A**; see **Figure S6** for other examples). In untreated cells, seven *GBP* genes are encompassed by two large (370- and 470-kb) ‘‘structural’’ loops whose anchors do not overlap active gene promoters. Small loops start to form as early as 30 min after activation, connecting H3K27ac peaks to promoters. This is followed by increased expression of genes at the anchors of those loops. This increased expression is coupled with loss of the large structural loops that span this locus. Visualizing these changes via line plots **(Figures 5B–5E)** highlights the correlation between looping and anchor gene transcription as well as the inverse correlation between structural loops and internal transcription.

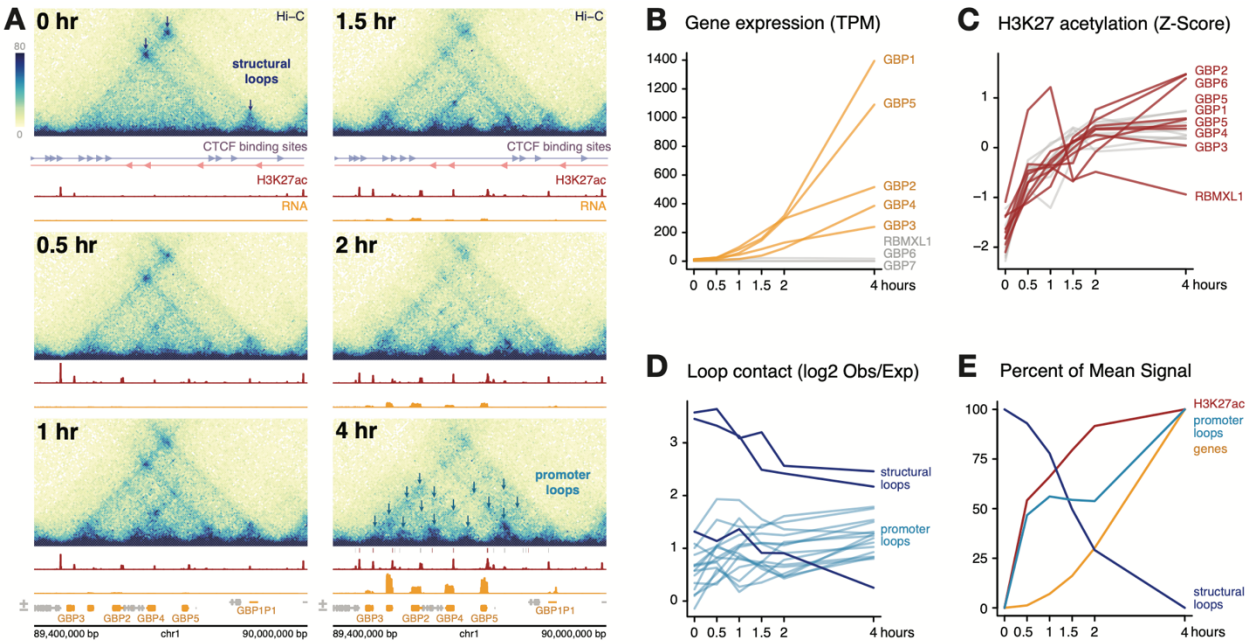


Figure 6.5. Long-distance loops are lost concurrently with increased internal transcription and restructuring at the *GBP* locus. (A) Chromatin structure (5-kb resolution), H3K27 acetylation, and gene transcription change drastically over the first 4 h at the *GBP* locus of chromosome 1. Predicted CTCF binding sites are also shown. Prior to treatment, two large ‘‘structural’’ loops (not connecting enhancers and promoters) encompass several *GBP* genes. After 30–60 min of LPS/IFN$_\gamma$ treatment, *GBP* promoters become acetylated. From 1 h onward, as acetylation increases, connections form between the *GBP* promoters. As genes become more highly expressed at 1.5 h and beyond, the original long-distance structural loops weaken in favor of shorter-range, transcription-correlated contacts. Hi-C scale is indicated in KR-normalized counts. **(B)** The TPM of each gene in the region, with the up-regulated genes highlighted in yellow (as in A). **(C)** The *Z* score-normalized change in H3K27ac at promoters (red) and putative enhancers (gray) in this region. The promoters and enhancers plotted are highlighted in the 4-h panel in (A). **(D)** The log-transformed ratio of observed to expected contact frequency of several points in the region. ‘‘Structural’’ loops (as in A, 0 h) are colored dark blue, and promoter-promoter and promoter-enhancer loops (as in A, 4 h) are colored light blue. **(E)** The mean TPM (for expressed genes), *Z* score (for H3K27ac), and log2 observed/expected ratio (for structural or promoter contacts) for the individual features highlighted in (B)–(D).

6.3. Discussion

6.3.1. Temporal analysis of macrophage activation

We used a fine-scale multi-omics time course of macrophage activation and quantified changes in DNA looping, enhancer acetylation, and gene expression. The high temporal resolution of the Hi-C data revealed changes in chromatin looping along short, transcriptionally relevant time scales that were undetected at time course endpoints. Combining the data across time points yielded Hi-C data at a depth of over 16 billion contacts, allowing sensitive and robust detection of macrophage chromatin loops. Integration of the data revealed several findings regarding the nature of DNA loops.

6.3.2. The influence of chromatin structure on transcription

The correlated changes we observe between looped enhancers and promoters are consistent with loops serving as a functional bridge between enhancers and their target genes, at least for genes regulated in response to external stimuli. This is supported by temporal ordering of events that reveal that loop formation and looped-enhancer activation occur prior to increases in anchor gene expression. This agrees with a previous study of cells responding to glucocorticoids in which maximal changes in loops were observed earlier than maximal changes in genes (D'Ippolito et al. 2018). These results are somewhat inconsistent with results from Rao et al. (2017), which showed very few changes in gene expression in response to global loop disruption via RAD21 degradation (Rao et al. 2017); however, in that study, the cells were grown in steady state and not responding to external stimuli. In experimental designs more comparable with ours, cohesin or CTCF depletion did disrupt the transcriptional response of macrophages to microbial stimuli (Cuartero et al. 2018; Stik et al. 2020). This suggests that loops likely play a critical role in mediating transcriptional changes in cellular response to stimuli.

Intriguingly, we found that non-looped enhancer-promoter pairs that were matched for contact frequency and histone H3K27ac levels did not exhibit the same level of temporal correlation as looped enhancer-promoter pairs. This suggests that loops may exert their regulatory control via mechanisms beyond merely increasing contact frequency between enhancers and promoters. One possible but speculative explanation is that activation of transcription by distal enhancers may require prolonged enhancer-promoter contact rather than overall contact frequency. Transcription factor binding is typically quite transient (F. Lu and Lionnet 2021), and prolonged contact might be required for proper formation of enhancer, polymerase, or mediator complexes that drive transcriptional activation. Recent work using 3D super-resolution live-cell imaging found that loops stabilized contact between anchors for 10–30 min (Gabriele et al. 2021). In the absence of a chromatin loop, such prolonged contact is unlikely even for non-looped enhancers and promoters that are separated by relatively short genomic distances. Closely spaced but non-looped enhancer-promoter pairs might participate in much more frequent but shorter-duration contacts that are insufficient for transcriptional activation. Hi-C data measure contact frequency but cannot differentiate between frequent short interactions and infrequent but prolonged interactions. Further

exploration is required to determine whether prolonged contacts do indeed account for these differences and, if so, what the exact mechanisms are.

These observations are consistent with many possible models of long-range enhancer activity, including those that require stable or transient contact (also called ‘‘hit-and-run’’ or ‘‘kissing’’ models) (Brandão, Gabriele, and Hansen 2021; Karr et al. 2022). We see gained transcription at not just *de novo* but also quantitatively strengthened loops, which is consistent with a model where modest enhancer contact coupled with futile cycles of promoter activity could result in large effects on transcription (Xiao, Hafner, and Boettiger 2021). This futile cycle model also predicts a lag between changes in loop strength and transcription but on a longer timescale than we observed. Other models, such as the transcription factor activity gradient (TAG) model, suggest that proximity rather than direct contact is required for long-range regulation (Karr et al. 2022). Although such a model could also support looped enhancers regulating distal promoters, it would require promoter-specific activities to explain how non-looped enhancers in equal contact frequency are not as strongly regulated.

6.3.3. The influence of transcription on chromatin structure

Several analyses from this paper support a model in which high levels of transcription could stall, displace, or generally antagonize loop extrusion complexes. First, we found that changes in gene expression exhibit a directional bias at differential loop anchors. The anchors of gained enhancer-promoter loops were associated with increased gene expression of genes oriented away from the center of the loop but not genes oriented toward the center of the loop. In contrast, lost enhancer-promoter loops were associated with increased expression of anchor genes oriented toward the center of the loop but not those oriented away from it. The temporal patterns of loop loss and internal transcription were anti-correlated, although, when averaging across all lost loops, the temporal order is not as obvious as the trends seen among gained loops, which more clearly precede transcriptional changes. These temporal analyses agree with previous work showing accumulation of cohesin at the 3′ ends of genes in a manner correlated with the amount of transcription and also sensitive to transcriptional inhibition (Busslinger et al. 2017; Heinz et al. 2018; Olan et al. 2020) and recent studies demonstrating that RNAP may act as a ‘‘moving barrier’’ to loop extrusion (Banigan et al. 2022). Transcription may also shape chromatin independent of cohesin. Recent high-resolution microscopy and Micro-C experiments have detected fine-scale cohesin-

independent structures between and within highly expressed genes, which could compete with or disrupt cohesin-mediated structures (Hsieh et al. 2020; Leidescher et al. 2022). Transcription inhibition interrupts these local structures but leaves intact broader loops, domains, and compartments. Virtually all transcriptionally active chromatin exhibits elevated contact frequency via a phenomenon called compartmentalization that does not require cohesin (Rao et al. 2017; Rowley and Corces 2018; Vian et al. 2018). It remains possible that such compartmentalization itself could disrupt loops surrounding highly expressed genes.

Finally, lost loops were associated with relatively high levels of internal transcription, and only very large changes in transcription were associated with decreased looping. This might reflect the fact that, in most cases, collisions between transcription and loop extrusion are rare. Indeed, transcription occurs in relatively infrequent bursts (Fukaya, Lim, and Levine 2016), and loops appear to spend at least some time in fully looped or fully non-looped states (Gabriele et al. 2021). So at low to moderate levels of transcription, collisions might be uncommon and are not a major driver of 3D chromatin structure. So perhaps it is only at extremely high levels of transcription where such collisions are frequent enough to lead to observable losses in loop-based contacts. Alternatively, it is possible that transcription only slightly impedes extrusion and that at low levels of transcription, changes in contact frequency are imperceptible. This agrees with studies showing that transcription briefly stalls condensin translocation but that it only measurably affects 3D chromatin structures at extremely highly expressed loci, such as at rRNA genes (Brandão et al. 2019).

6.3.4. Future directions

This fine-scale time course of looping in human macrophages provides insight into the temporal organization of regulatory events in human cells responding to external stimuli and a deeper understanding of the mechanisms driving transcriptional regulation in human cells. Some of the findings could be useful for predicting functional enhancer-promoter pairs. For example, the temporally coordinated changes observed at looped enhancer-promoter pairs could be employed to refine and potentially improve predictions made by the activity-by-contact model (Fulco et al. 2019).

This work supports a model in which loop extrusion and transcription participate in a coordinated dance and can influence each other in a reciprocal relationship. If this holds true, then it could have important implications for how genes are organized in the context of chromatin loops. For example, genes

oriented toward the center of a loop could be regulated by a negative feedback mechanism in which high levels of transcription might decrease looping between the promoter and a distal enhancer. Moving forward, incorporation of more data types into these time courses should reveal further insights into the mechanisms of 3D chromatin structure and gene regulation.

6.3.5. Limitations of the study

This paper aimed to place genomic events in temporal order, but several caveats should be considered based on the methods used. RNA-seq measures accumulation of transcribed RNA but does not directly measure the amount of transcription along the genome. Our ITS, calculated by expanding TPM counts across gene bodies including introns, was used as a proxy, but alternative methods, such as precision nuclear run-on sequencing (PRO-Seq) or global run-on sequencing (GRO-Seq), would be required to directly measure this. Use of bulk methods in independent cell populations provided high sample numbers and allowed us to view general trends, but it means that we are only able to detect how changes correlate on a population level and cannot determine how events are co-occurring in individual cells. Finally, additional time points would provide further insights into the changes observed here. For example, evenly spaced time points of every 30 min would reduce extrapolation and improve temporal correlation assessments. Similarly, more closely spaced timepoints could reveal even finer-scale lags in temporal trends but could be more logistically challenging for high-throughput methods. Other studies implementing single-cell techniques could complement these limitations nicely and would be interesting to compare with the trends observed here.

6.4. STAR methods

6.4.1. Cell lines

Human male THP-1 monocytes (RIID CVCL_0006, ATCC TIB-202) were grown and maintained in RPMI media with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS). Cells were routinely checked for mycoplasma using Genetic Cell Line Testing and confirmed negative. Cell lines were also authenticated using STR analysis via the UNC Vironomics Core.

6.4.2. Macrophage differentiation and activation

For differentiation into macrophages, THP-1 monocytes were transferred to 6-well plates (RNA-seq, ATAC-seq) or T-175 flasks (Hi-C, ChIP-seq) at a density of 2×10^5 cells/mL and treated with 25 nM

PMA for 24 h, over which time the cells become adherent. The media was then aspirated off, the flasks were washed gently with RPMI, and then fresh RPMI (10% FBS, 1% PS) and rested for 72 h.

The resting macrophages were then treated with a combination of 10 ng/mL lipopolysaccharide (LPS) and 20 ng/mL interferon gamma (IFN_γ) in fresh RPMI (10% FBS, 1% PS). Cells were harvested without treatment, or 0.5, 1, 1.5, 2, 4, 6, or 24 h after LPS and IFN_γ treatment.

During each treatment, extra 0- and 2-h samples were prepared simultaneously for RNA extraction, and qPCR was used to measure the regulation of FOS, IL1B and IL6 to confirm consistent treatment response.

For all library preparations, the differentiation and activation treatment was performed from freshly thawed THP-1 cells on two separate occasions, to achieve the closest approximation to two biological replicates using cultured cell types.

6.4.3. Crosslinking

For ChIP-seq and Hi-C, cells were grown in T-175 flasks, each containing 20×10^6 cells at a density of 2×10^5 cells/mL. Cells were crosslinked using 1% formaldehyde in RPMI for 10 min with gentle shaking. Crosslinking was then quenched with 10% 2.0 M cold glycine for 5 min. The media was then removed and cells were scraped into cold PBS. Each flask was divided into 4 tubes of approximately 5×10^6 cells each. Cells were spun down at 526 G for 5 m, resuspended in PBS and respun to wash away residual formaldehyde. Cells were then frozen in liquid nitrogen and stored at −80 for library preparation.

6.4.4. RNA-seq library preparation

RNA was extracted using the QIAGEN RNeasy Mini kit with DNase I treatment. RNA integrity numbers were confirmed using a Tapestation RNA screentape to be above 9.8, and a Qubit High Sensitivity assay was used to determine RNA concentration.

Ribosomal RNA was removed using the NEB rRNA Depletion Kit (Human/Mouse/Rat) using 500 ng of RNA as input. Following depletion, RNA-seq libraries were prepared using the NEB Ultra II Directional RNA Library Prep Kit for Illumina, and NEBNext Multiplex Oligos for Illumina. Library concentration and fragment size was determined using Qubit (dsDNA HS assay) and Tapestation (D1000 screentape). Libraries from each timepoint were pooled to a final DNA concentration of 15 nM, and 75-bp paired-end reads were sequenced on an Illumina NextSeq 500 using a High Output Kit.

6.4.5. ChIP-seq library preparation

Four frozen cell pellets (5×10^6 cells each) were used for each timepoint. Pellets were first rinsed in 10 mL rinse buffer 1 (50 mM HEPES pH 8, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, Triton-X), incubated on ice for 10 min, and then spun down at 2,400 G at 4°C for 5 min. Supernatant was removed and the pellets were rinsed again in rinse buffer 2 (10 mM Tris pH 8, 1 mM EDTA, 0.5 mM EGTA, 200 mM NaCl), and spun at the same settings. Supernatant was removed, and 5 mL of shearing buffer (10 mM Tris pH 8, 2% Triton-X, 1% SDS, 100 mM NaCl, 1 mM EDTA) was added to the tubes to wash out the rinse buffer. The samples were centrifuged at 2,400G at 4°C for 3 min, the shearing buffer was removed and this step was repeated. The cell pellets were then resuspended in 88 uL of shearing buffer, 2 uL of protease inhibitor cocktail (PIC), and 10 uL of nanodroplets (Triangle Biotechnology, Inc.) per 10 million cells (Kasoji et al. 2015; Marcel et al. 2021). Samples were aliquoted into 100 uL tubes and sheared using a Covaris E110 (intensity 6, 210 s). Cells were spun down at max speed for 2 min and the supernatant was retained.

In order to determine the concentration of chromatin, 10 uL was removed (while the rest of the sample was stored at −80°C), and crosslinking was reversed by adding 5 uL of 5 M NaCl, 125 uL of TE buffer (10 mM Tris pH 8, 1 mM EDTA) and 125 uL of elution buffer (1 M Tris pH 8, 10 mM EDTA, 1% SDS), vortexed, and incubated overnight at 65°C. Samples were spun down and added 7.5 uL of proteinase K and 3 uL of RNase A. DNA was extracted using the Zymo ChIP DNA Clean & Concentrator Kit, quantified using Qubit (dsDNA broad-range (BR) assay), and run on a gel to ensure fragment sizes of 100–300 bp and concentrations high enough to continue with library prep.

Immunoprecipitation of the remaining volume from each sheared sample was completed using the Active Motif ChIP-IT High Sensitivity kit, using 2.8 ug of chromatin from each timepoint (as determined by the lowest yield samples), and 4 ug of anti-H3K27ac antibody (AbCam ab4729). Following overnight antibody incubation and washing steps, crosslinking was reversed by adding 100 uL of elution buffer (described previously) and 4 uL of 5 M NaCl to 100 uL of the IP reactions, vortexing, and then incubating overnight at 65°C. DNA was purified using the Zymo ChIP DNA Clean & Concentrator kit, and quantified using Qubit (dsDNA high-sensitivity (HS) assay), as before.

Following the final dilution, libraries were prepared using the NEB Ultra II DNA Library Prep Kit with NEBNext Multiplex Oligos for Illumina with 0.88 ng of DNA as input. Libraries were analyzed using Qubit (dsDNA HS assay) and Tapestation (D1000 screentape) and pooled to a final concentration of 12 nM, and then 75-bp paired-end reads were sequenced on an Illumina NextSeq 500 using a High Output Kit.

6.4.6. *In situ* Hi-C library preparation

Three treatments (biological replicates) were conducted, and one or two frozen cell pellets ($5x10^6$ cells each) were used to generate separate libraries as technical replicates (1 technical replicate for first biological replicate; 2 technical replicates for second and third biological replicates). Libraries were prepared using the *in situ* Hi-C protocol as described in Rao et al. (2014) (Rao et al. 2014). In brief, crosslinked cells were lysed on ice, nuclei were isolated, and chromatin was digested overnight with the MboI restriction enzyme. Chromatin ends were biotinylated, proximity ligated, and crosslinking was reversed. Samples were sheared on a Covaris LE 220 (DF 25, PIP 500, 200 cycles/burst, 90 s), quantified using Qubit (dsDNA High Sensitivity (HS) assay), and a small sample was run on an agarose gel to ensure proper fragmentation. DNA sized 300–500 bp was selected for using AMPure XP beads, and then eluted. Biotinylated chromatin was then pulled down using streptavidin beads. Following removal of biotin from unligated ends and repair of sheared DNA ends, unique Illumina TruSeq Nano (Set A) indices were ligated onto the samples. Libraries were amplified off of streptavidin beads using 7–10 PCR cycles based on post-size selection concentrations, quantified again using a Qubit (dsDNA HS assay), and fragment length was determined using Tapestation (D1000 screentape). Libraries were pooled to 10 nM. Paired-end 150-bp reads were sequenced on one or two lanes of an Illumina NovaSeq S4.

6.4.7. ATAC-seq library preparation

ATAC-seq libraries were prepared using the Omni ATAC-seq protocol as described in Corces et all (2017). Adherent macrophages were washed once with PBS and lifted off of the plate with EDTA for 5 min. EDTA was quenched with RPMI, and library preparation was performed on 50,000 cells. 3.75 µL of Illumina Nextera XT indices were used in PCR and qPCR.

After performing the initial 5 cycles of PCR, 5% of the PCR reaction was used in qPCR to determine how many additional cycles were required. 4–7 cycles were determined to be sufficient for the final amplification. A 2-sided bead cleanup with AMPure XP beads was performed (0.5X, then 1.3X). Libraries

were quantified using Qubit (dsDNA HS Assay) and the KAPA Library Quantification kit. Libraries from each timepoint were pooled to a concentration of 8 nM or 10 nM for each biological replicate, and 75-bp paired-end reads were sequenced on an Illumina NextSeq 500 using a High Output Kit.

6.4.8. RNA-seq processing and gene quantification

Adaptors and low-quality reads were trimmed from paired-end reads using Trim Galore! (version 0.4.3) ("Babraham Bioinformatics" 2015). Salmon (version 1.4.0) (Patro et al. 2017) was used in quant mode to quantify reads to hg19 transcripts from GENCODE (version 19). For signal tracks, reads were aligned using HISAT2 (version 2.1.0) (D. Kim et al. 2019), indexed and replicates were merged with samtools (version 1.9) (Danecek et al. 2021), and converted to bigwigs using deeptools (version 3.0.1) (Ramírez et al. 2016). Reads were summarized to a gene level using tximport (R version 3.3.1, tximport version 1.2.0) (R Core Team 2022; Soneson, Love, and Robinson 2015), which was then used as input for differential analysis in DESeq2 (version 1.33.5) (Love, Huber, and Anders 2014). FastQC and MultiQC were used to assess library quality metrics (version 0.11.5; version 1.5) ("Babraham Bioinformatics" 2010; Ewels et al. 2016).

6.4.9. Inferred transcription score (ITS) calculations

Inferred transcription scores (ITS) were calculated in order to estimate the degree of transcription occurring throughout gene bodies, including introns, as extrapolated from the mature mRNA TPM levels. Gene-level TPM as quantified by Salmon and summarized by txImport (see RNA-seq processing methods). The genome was binned into 10-kb or 50-kb regions using bedtools makewindows (version 2.28), and then overlapped with gene bodies (Quinlan and Hall 2010). Gene TPM values were applied to each overlapping bin, adjusted based on the percentage of bin overlap. For example, a gene of TPM 50 with a TSS at position 100,000 bp and a TTS at position 115,000 bp would contribute an ITS of 50 to the bin of 100,000–110,000, and an ITS of 25 to the bin of 110,000–120,000. In bins with multiple genes, ITS scores were generated by summing the TPM contribution from each gene.

6.4.10. ATAC- and ChIP-seq processing and peak calling

Adaptors and low-quality reads were trimmed from paired-end reads using Trim Galore! (version 0.4.3) ("Babraham Bioinformatics" 2015). Reads were aligned using BWA mem (version 0.7.17) (H. Li 2013) and sorted using Samtools (version 1.9) (Danecek et al. 2021). Duplicates were removed with PicardTools

(version 2.10.3) (“Picard” n.d.) and for ATAC-seq libraries, mitochondrial reads were removed using Samtools idxstats. Samtools was also used to merge replicates for each timepoint, and index BAM files. Peaks were called from the merged alignments using MACS2 with the following settings: -f BAM -q 0.01 -g hs –nomodel –extsize 200 –keep-dup all -B –SPMR (version 2.1.1.20160309;) (Yong Zhang et al. 2008). ChIP-seq peaks used the MACS2 setting –shift 0, while ATAC-seq peaks used –shift 100. Peaks from all timepoints were then merged using bedtools (version 2.28) (Quinlan and Hall 2010), generating 118,344 ChIP-seq and 193,853 ATAC-seq peaks in total. For each replicate BAM file, ChIP-seq counts were extracted from ATAC-seq peak locations using bedtools multicov. Bedtools intersect was used to subset for ATAC-seq peaks that overlapped H3K27ac ChIP-seq peaks, and these 89,503 peaks were considered putative regulatory regions. Raw counts at these enhancers (8 timepoints, 2 replicates each) were used as input for differential analysis with DESeq2 (version 1.33.5) (Love, Huber, and Anders 2014). Signal tracks were made from alignments using deeptools (version 3.0.1) (Ramírez et al. 2016).

6.4.11. Enhancer and promoter definitions

Gene promoters were identified as regions 2,000 base pairs upstream and 200 base pairs downstream of gene transcriptional start sites (TSS). Promoter H3K27ac signal was calculated based on any overlapping H3K27ac and ATAC-seq peaks within promoter regions using bedtools intersect (version 2.28) (Quinlan and Hall 2010). Enhancers were identified as overlapping H3K27ac and ATAC-seq peaks that did not overlap with defined promoter regions.

6.4.12. Predicted CTCF binding sites

Directional predicted CTCF binding sites for hg19 were obtained from AnnotationHub ID AH95565 using the ‘‘CTCF’’ R package (Dozmorov et al. 2022). The specific track used here features data from UCSC Jaspar, and represents the hg19 coordinates of binding motif MA0139.1 as detected by FIMO.

6.4.13. Hi-C processing, loop and compartment calling

Hi-C data were processed using the Juicer pipeline as initially described in Rao et al. (version 1.5.6) (Rao et al. 2014). Hi-C maps were made at 5- and 10-kb resolution for each technical replicate (8 timepoints, each with 5 technical replicates across 3 biological replicates), as well as for each timepoint (all replicates combined). Additionally, a ‘‘Mega’’ map from all timepoints was made, also using Juicer.

Loops were identified at 5-kb using SIP (version 1.6.1) (Rowley et al. 2020). Loops were called from the individual timepoint maps using the settings ‘‘-g 2 -t 2000 -fdr 0.05’’, and from the Mega map with the settings ‘‘-g 1 -t 2000 -fdr 0.05’’. The loops were then extrapolated to 10-kb, concatenated, and merged in R using DBScan (version 1.1.8) (Hahsler, Piekenbrock, and Doran 2019) with an epsilon of 20 kb (manhattan distance), keeping the mean of modes for coordinates, resulting in 42,690 total loops. The counts for these loops were then extracted from the Hi-C files of each technical replicate (un-normalized, 10-kb resolution) using strawr (version 0.0.9) (Durand, Shamim, et al. 2016). These raw counts (8 timepoints, 3 biological replicates each, two with 2 technical replicates and one with 1 technical replicate each) were used as input for differential analysis with DESeq2 (Love, Huber, and Anders 2014).

Compartments were called for Hi-C biological replicates 2 and 3 using the EigenVector script at a 50-kb resolution (Olshansky 2021). Biological replicate 1 was not used for this analysis due to its decreased sequencing depth and the method of differential compartment calling (see “differential compartment and ITS bin analysis” section below). The A compartment was determined based on gene content overlap on a per-chromsome basis.

6.4.14. Differential gene and peak analysis

DESeq2 was used for differential analysis of genes, loops, and peaks (Love, Huber, and Anders 2014). Each analysis used a likelihood ratio test (LRT), with a full design of ‘‘~bioRep + time’’ and a reduced design of ‘‘~bioRep’’. Posterior log-2 fold changes (LFC) were estimated using apeglm (A. Zhu, Ibrahim, and Love 2019). Significant results were determined based on an absolute LFC greater than 1 and an adjusted p value below 0.01.

Raw counts were converted into Z-scores by first conducting a variance-stabilizing transformation across all features, and then centering and scaling the data in each feature based on standard deviations from the mean. Genes were categorized into up- and downregulated based on the signage of their Z-score at 0 h of LPS/IFNg treatment, and then sorted based on their timepoint of maximum Z-score.

6.4.15. Differential loop analysis and clustering

DESeq2 was also used for differential analysis of loops (Love, Huber, and Anders 2014). Differential analysis used a likelihood ratio test (LRT), with a full design of ‘‘~techRep + bioRep + time’’, and

a reduced design of ‘‘~techRep + bioRep’’. Significant results were determined based on an absolute LFC greater than 0.585 (fold-change of ±1.5) and an adjusted p value below 0.05.

Raw counts were converted into Z-scores by first conducting a variance-stabilizing transformation across all features, and then centering and scaling the data in each feature based on standard deviations from the mean. These Z-scores were then used to cluster loops using k-means clustering (k = 4). For the survey of loop contacts at the GBP locus, log2 observed/expected KR-normalized counts were extracted using strawr.

Compartment eigenvectors (EV) and ITS values were both binned at 50-kb for comparison and overlap. Differential EV and ITS bins were identified based on an average raw difference of 0.5 (EV) or a fold-change of 2 (ITS) between any timepoint and 0 h. False-positive rates of 13.4% for EV and 10.7% for ITS were calculated by finding the number of bins with differences beyond these thresholds between biological replicates (regarded as false positives) out of the total number of bins passing the threshold. ITS bins were further subset for bins containing a score of 5 or higher in any biological replicate. A total of 5,611 differential EV bins were found, and a 8,068 differential ITS bins were found, with 930 bins that were differential for both.

Differential compartment and ITS bins were then categorized according to their timing and direction of change. First, z-scores were calculated for each bin at each timepoint by meaning the eigenvector or ITS values between replicates, then centering and scaling the data. The overall direction (up or down) of each differential bin was determined based on the sign of their initial z-score. Based on these categorizations, 81.2% of overlapping differential bins have EV and ITS that changed in the same direction. The increasing and decreasing bins were then further categorized by the timepoint of minimum or maximum z-score (30 min–24 h), as for differential gene categorization.

6.4.16. Matched enhancer-promoter sets

Covariate-matched subset selection among non-looped enhancer-promoter pairs was performed using the *matchRanges* function from the *nullranges* package (Davis et al. 2022). Enhancer-promoter pair distance or total contact frequency were used as covariates. Total contact frequency was calculated from KR normalized counts from the combined Mega map, effectively a sum of contacts across all timepoints and replicates. Matching was done with the stratified matching method without replacement. Enhancer

strength, defined by the sum of H3K27ac variance-stabilized counts across all timepoints and replicates, was compared between the looped and matched non-looped sets.

6.5. Supplementary figures

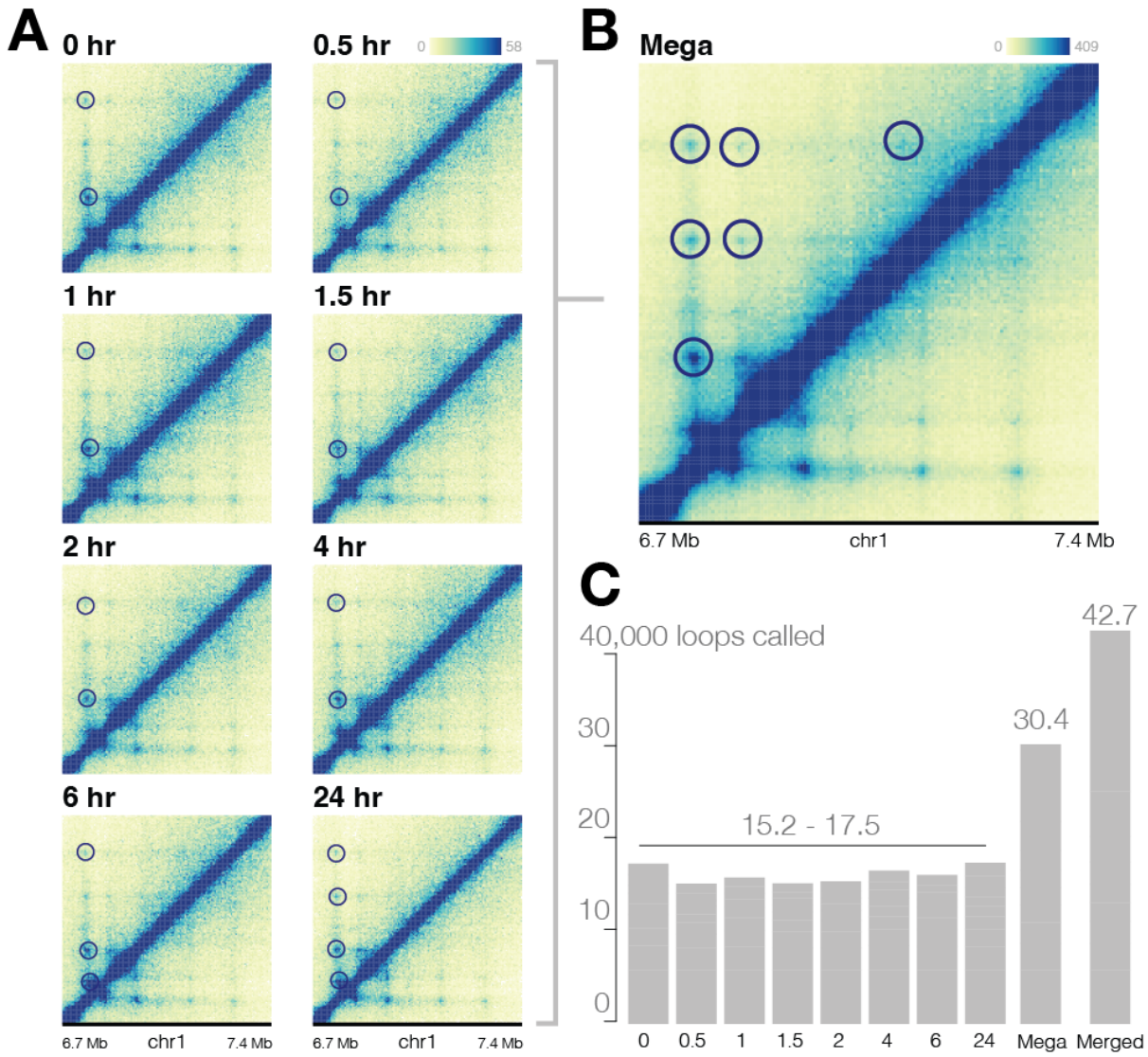


Figure 6.S1. Deeply sequenced in situ Hi-C sensitively identifies loops. (A) Hi-C maps of roughly 2 billion Hi-C contacts were generated for each timepoint (5-kb resolution; merged from 5 technical replicates across 3 biological replicates). Loop calls from each time point are circled in blue. **(B)** For added depth and sensitivity of loop detection, each timepoint was merged into a Mega map of 15.7 billion Hi-C contacts (5-kb resolution). Loop calls from the Mega map are circled in blue, showing higher sensitivity than individual timepoints. **(C)** Loops were called at 5-kb in each individual timepoint, as well as the Mega map, extrapolated to 10-kb, and loops with both anchors within 20 kb were merged. Nearly twice as many loops were called from the Mega map compared to individual timepoint maps, resulting in 42,690 total chromatin loops.

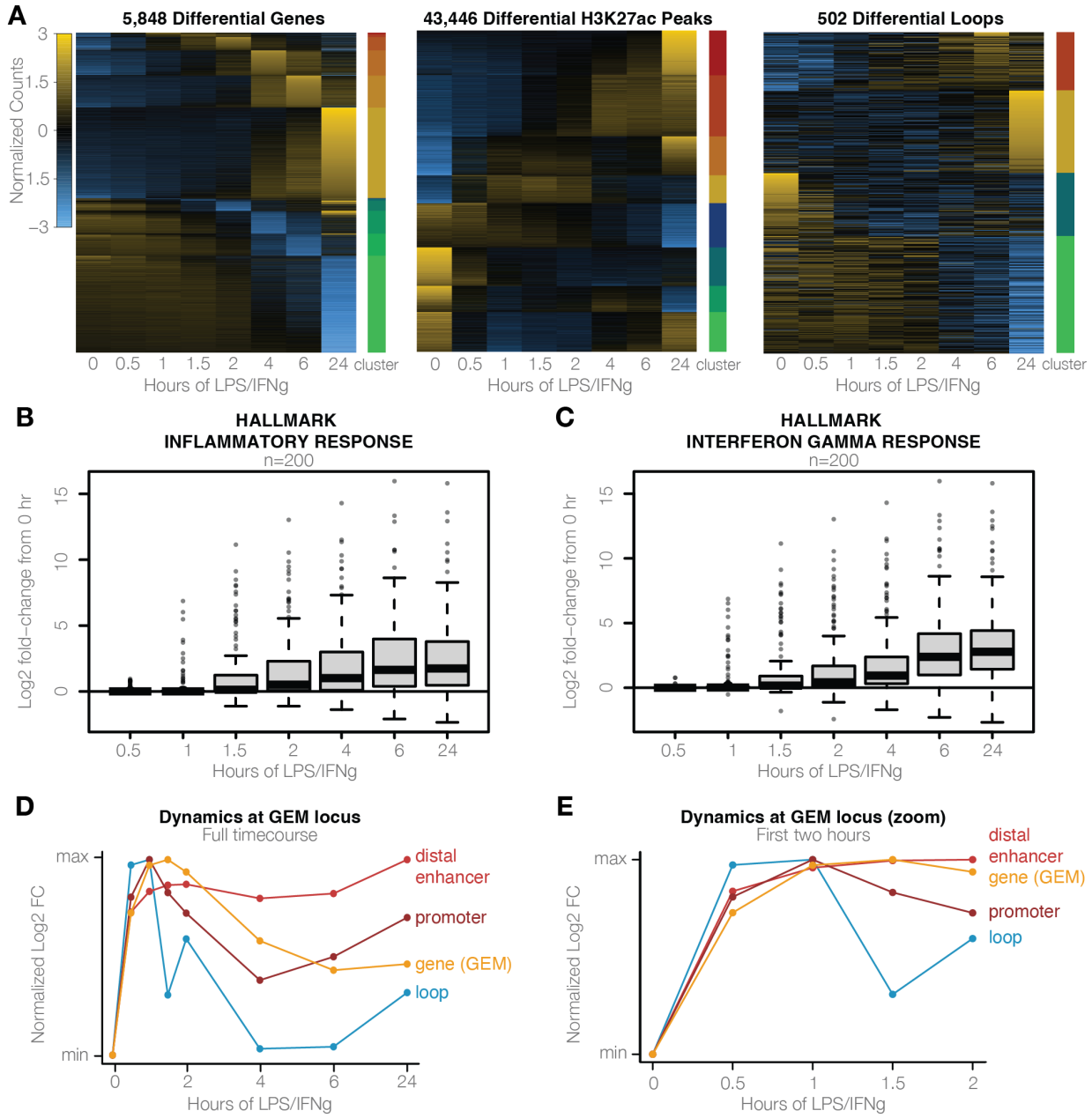


Figure 6.S2. Transcriptional profile consistent with inflammatory hallmarks. (A) Differential genes, overlapping ATAC-seq and H3K27ac ChIP-seq peaks, and loops occur at multiple timescales. Differential features were clustered independently to show diversity of temporal patterns in each. **(B)** Genes previously identified as part of the hallmark inflammatory response and **(C)** hallmark interferon gamma response were investigated in this system. Canonically upregulated genes exhibit positive fold-change in response to LPS/IFNg, especially at 4 hours and beyond. **(D)** Scaled log2 fold-changes in loop strength (blue, Hi-C contact), GEM gene expression (gold, RNA-seq z-score), and H3K27ac at the GEM promoter (dark red) and distal looped enhancer (light red, H3K27ac ChIP-seq) compared to the 0-hour time point. (E) The same information as in panel D, but showing only the first 2 hours of treatment. Log2 fold-changes are re-normalized for this subset of data.

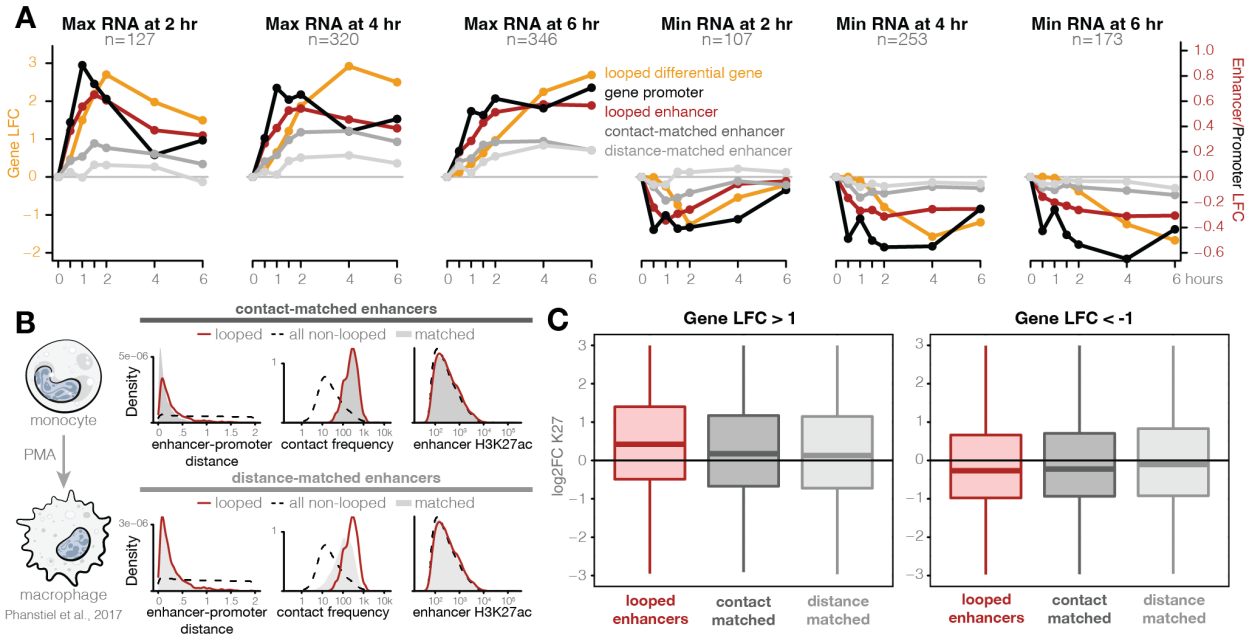


Figure 6.S3. Looped enhancer-promoter pairs correlate in other systems. (A) The same information as Figure 2C (average log2 fold-change of gene expression [gold], looped enhancer H3K27ac [red], and non looped matched enhancer H3K27ac as defined in Figure 2A-B [grey]) but with the gene promoter H3K27ac fold-changes also plotted along the same y-axis as the enhancers (black). Gene promoters exhibit the same temporal lag seen in the looped enhancers. **(B)** Contact- and distance-matched enhancer-promoter pairs were identified to compare against looped enhancer-promoter pairs seen in monocyte-to-macrophage differentiation (Phanstiel et al. 2017). **(C)** Enhancers looped to upregulated genes show a correlated increase in H3K27ac that is significantly higher than both contact- and distance-matched pairs (Wilcoxon rank-sum test, p-value < 10-5). Similarly, enhancers looped to downregulated genes show a correlated decrease in H3K27ac that is significantly higher than distance-matched pairs (Wilcoxon rank-sum test, p-value < 10-5).

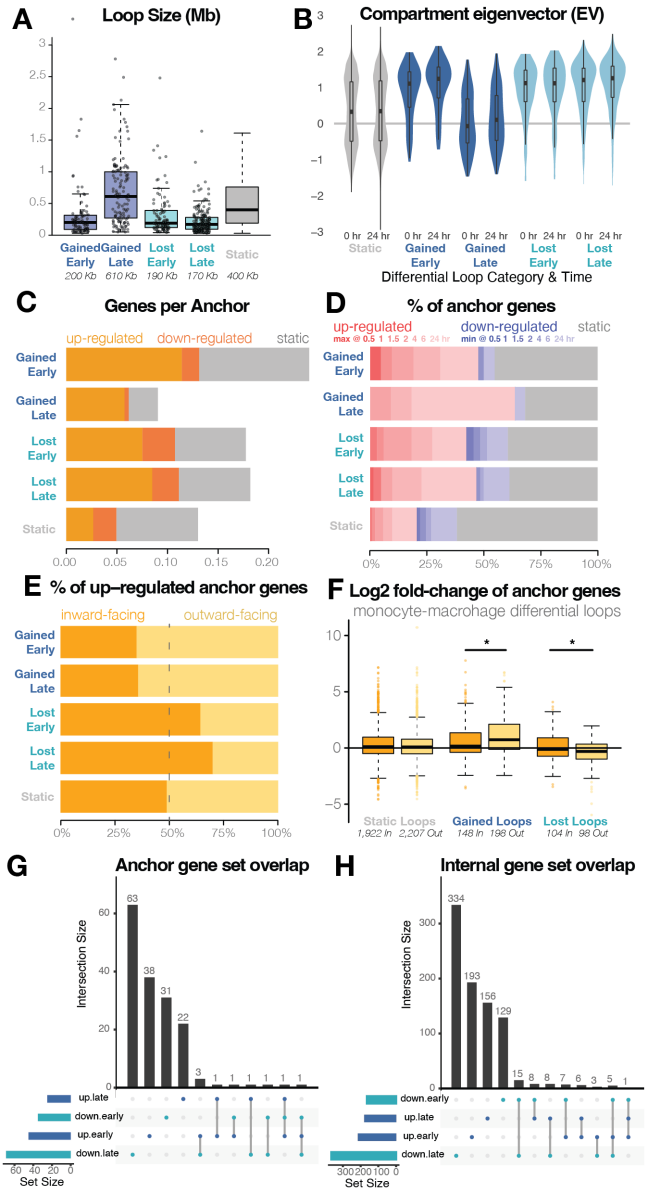


Figure 6.S4. Differential loop features. (A) The distribution of loop sizes based on differential cluster. **(B)** Most loops exist within the transcriptionally active A compartment, with gained late loops being the most B-like. All differential loop classes exhibit an overall shift towards the A compartment with time, while static loops remain unchanged on average. **(C)** The average number of genes per loop anchor for loops of each differential cluster, distinguished by differential status of anchor genes. **(D)** The percent of genes per loop anchor for loops of each differential cluster (as in Figure 3B), but with up- and down-regulated genes further separated into clusters based on their timepoint of maximum or minimum expression. **(E)** The percentage of upregulated anchor genes which are inward- or outward-facing among each class of differential loops. **(F)** Distributions of log2 fold-changes of genes with promoters in the anchors of static and differential loops from monocyte-macrophage differentiation. At gained loop anchors, the fold-change of outward-facing genes is significantly higher than inward-facing genes, while the opposite trend is seen among genes at lost loops (Wilcoxon rank-sum test, p-value < 0.05). **(G)** The amount of overlap between unique genes found at the anchors or **(H)** within the bounds of differential loops from different clusters. The sets of genes found at different types of loops is generally very unique to each loop set, meaning that individual genes are not frequently contributing to trends in multiple loop classes simultaneously.

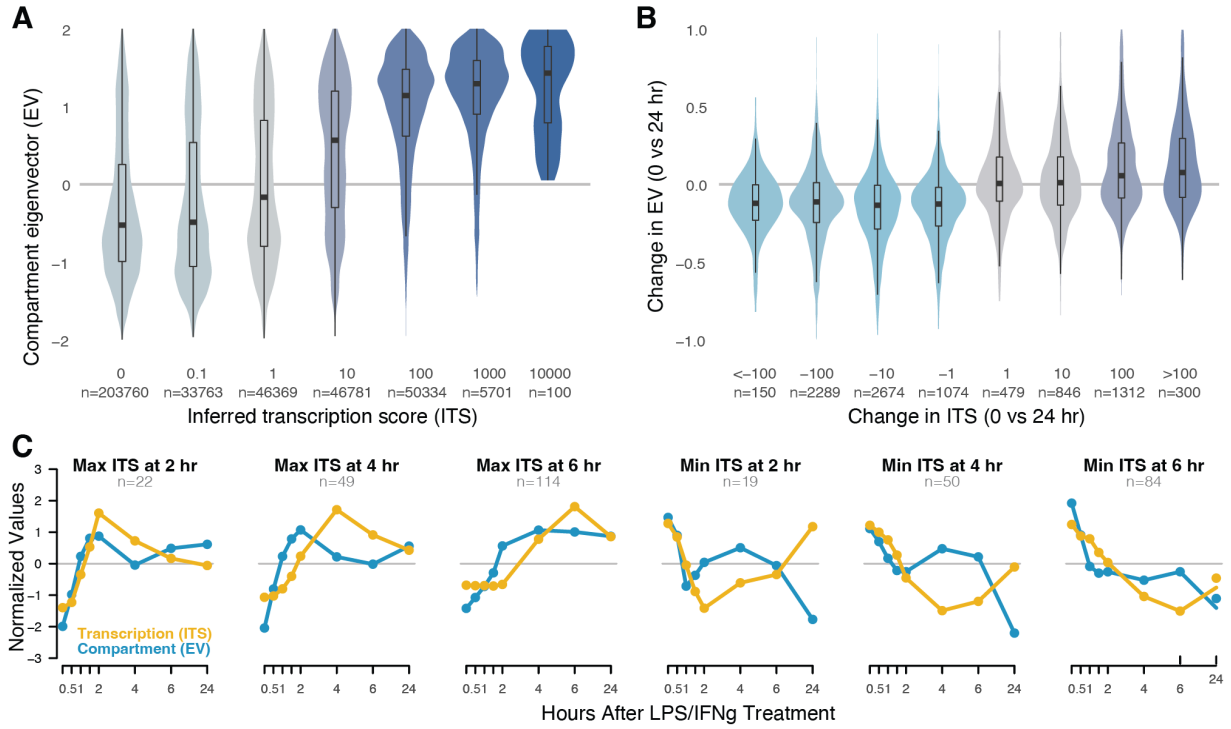


Figure 6.S5. Compartmental and transcriptional changes. In addition to differential loops, we also looked into the broader-scale structural changes by calling compartments at 50-kb resolution. For these figures we overlapped compartmental eigenvectors (EV) with inferred transcription scores (ITS), both at 50-kb resolution. Eigenvectors were calculated from and averaged across 2 biological replicates. To calculate ITS, the transcript per million (TPM) value for each gene was assigned to every genomic bin covered by the gene body. Values were summed for bins that overlapped multiple genes (see methods for details). ITS scores were then averaged between 2 biological replicates of RNA-Seq. **(A)** Consistent with the definition of the A compartment as transcriptionally active, compartment eigenvectors (EV) are correlated with the amount of transcription (ITS) occurring in the same 50- kb bin. High EV (A compartment) is associated with high transcription, while low EV (B compartment) is associated with lower amounts of transcription, with the threshold existing roughly between ITS scores of 0.1 and 10. Colors correspond to mean EV values. **(B)** To see how transcription and structure affected each other, we next looked at the per-bin changes in compartment EV compared to changes in ITS. Changes in ITS score across the timecourse correlate with changes in compartment EV, but in different ways depending on the direction of the change. Any loss of transcription (-1 to -100) appears to result in an equal decrease in EV (shift to B compartment), while gains in transcription seem to result in a proportional gain in EV (a +1 gain has a smaller effect than a +100 gain). Colors correspond to change in EV values. **(C)** To observe the temporal relationship between changes in ITS and EV, we identified bins that have both differential compartment EVs and differential ITS and organized them by the timepoint of minimum or maximum ITS. Increases in compartmentalization appear to precede changes in ITS, while the trends for bins with decreasing ITS are less clear.

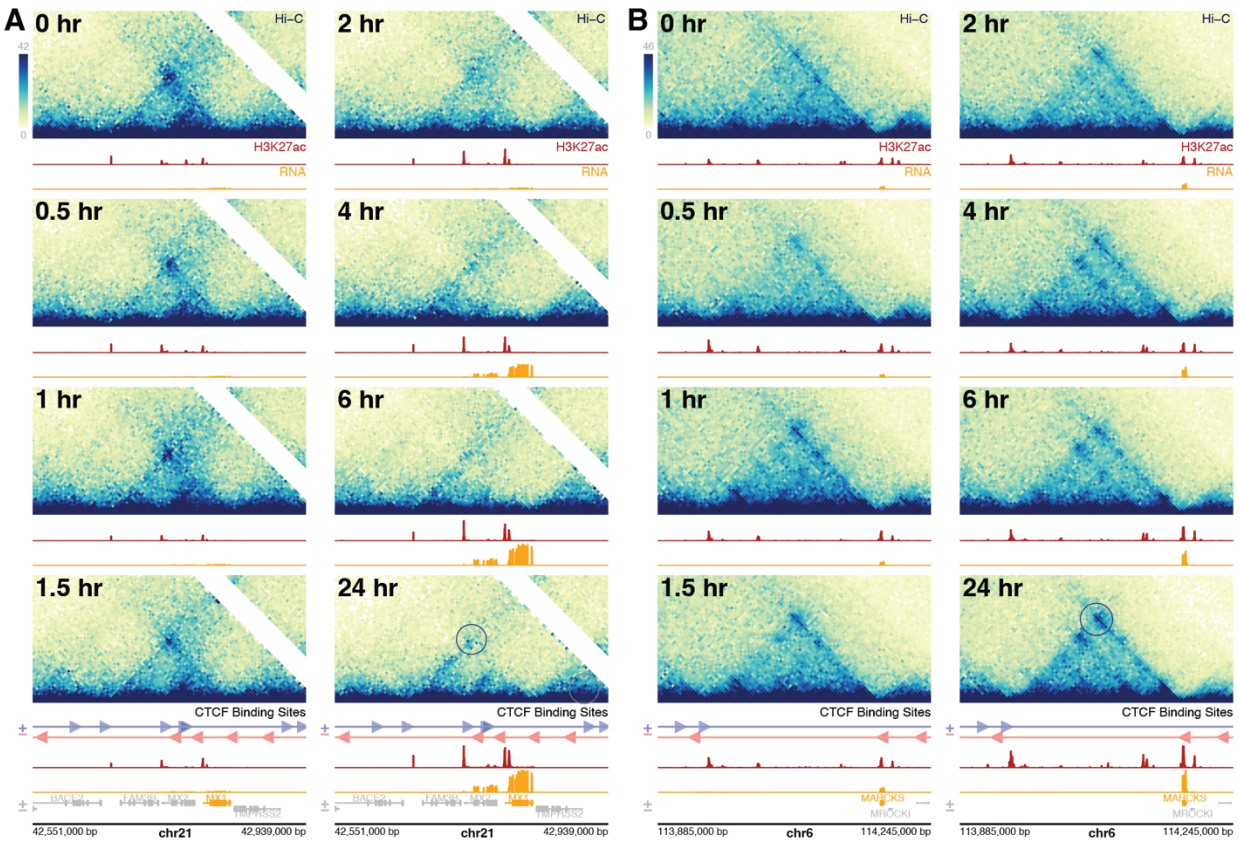


Figure 6.S6. Additional examples of gained and lost loops at differential genes. Two additional examples of **(A)** a lost loop containing highly upregulated genes within its boundaries, and **(B)** a gained enhancer-promoter loop with increased gene expression corresponding to increased contact frequency with a distal enhancer. Both regions are plotted at 5-kb resolution. Differential loops of interest are circled in the 24-hour map. Contact frequency (Hi-C), enhancer acetylation (ChIP-Seq), gene expression (RNA-Seq), and predicted CTCF binding sites ("CTCF" R library) are shown for each timepoint.

Chapter 7: 3D chromatin structure in chondrocytes identifies putative osteoarthritis risk genes[1]

7.1. Introduction

Osteoarthritis (OA) affects over 300 million people worldwide, yet treatment options are limited in large part because the mechanisms driving OA are not fully understood (Hunter and Bierma-Zeinstra 2019; Boer *et al.* 2021). Genome-wide association studies (GWAS) have identified over 100 loci associated with OA risk (Reynard and Barter 2020), but translating these broad loci into therapeutic targets has been challenging for several reasons. First, the effects of disease-associated variants are likely cell-type and context specific (Umans, Battle, and Gilad 2021); therefore, studying these variants in the correct system that mimics the OA phenotype is required. Second, linkage disequilibrium (LD) between nearby variants makes it difficult to identify the causal variant(s) at each locus. Finally, because the majority of OA risk variants occupy non-coding regions of the human genome and can regulate genes up to a million base pairs away, the genes impacted by most OA risk variants are unknown.

Several studies have successfully used genomic and bioinformatic techniques to identify the genes impacted by gene-distant non-coding GWAS variants for a variety of disease phenotypes (Claussnitzer *et al.* 2015; Won *et al.* 2016; Chesi *et al.* 2019; Laarman *et al.* 2019; Duan *et al.* 2021). Mapping regulatory loci using ChIP-seq, ATAC-seq, or CUT&RUN and intersecting the resulting data with disease-associated variants can identify a short list of putative causal variants. These variants can then be linked to potential target genes by quantifying 3D chromatin contacts using Hi-C or other chromatin conformation capture (3C) techniques. For example, chromatin interaction data was used to determine that an obesity-associated variant located in an intron of the *FTO* gene affects expression of the downstream genes *IRX3* and *IRX5*, which are involved in obesity-related biological processes (Claussnitzer *et al.* 2015). Likewise, Hi-C in

[1] The work in this chapter has been previously published. The citation is: Thulson, Eliza, Eric S. Davis, Susan D'Costa, Philip R. Coryell, Nicole E. Kramer, Karen L. Mohlke, Richard F. Loeser, Brian O. Diekman, and Douglas H. Phanstiel. 2022. "3D Chromatin Structure in Chondrocytes Identifies Putative Osteoarthritis Risk Genes." *Genetics* 222 (4). https://doi.org/10.1093/genetics/iyac141.

human cerebral cortex identified FOXG1 as a distal target of a schizophrenia GWAS variant, supporting its potential role as a schizophrenia risk gene (Won *et al.* 2016).

Because the effects of disease-associated variants are likely limited to particular biological states (Umans, Battle, and Gilad 2021), studies of their impact must be conducted in the correct cellular and biological context. Several pieces of evidence suggest that chondrocytes—particularly those responding to cartilage matrix damage—are one of the most likely cell types to be affected by OA risk variants. Cartilage breakdown and loss is a primary feature of OA. Chondrocytes are the only cell type found in cartilage and are responsible for maintaining the cartilage matrix. Osteoarthritic cartilage harbors activated chondrocytes that exhibit a proinflammatory phenotype thought to contribute to progressive cartilage degradation, which includes production of bioactive matrix fragments (Loeser 2014; van den Bosch *et al.* 2020). We have developed an *ex vivo* system that simulates the OA chondrocyte phenotype by treating primary human articular chondrocytes with fibronectin fragment (FN-f) (Forsyth *et al.* 2002; Pulai *et al.* 2005; Wood *et al.* 2016; Reed *et al.* 2021). Fibronectin is a ubiquitous extracellular matrix protein, and high levels of FN-f are present in cartilage and synovial fluid of OA joints (Xie *et al.* 1992; Homandberg *et al.* 1998). Subsequently, FN-f has been shown to be an OA mediator that recapitulates gene expression changes associated with OA (Homandberg 1999; Forsyth *et al.* 2002; Pulai *et al.* 2005; Reed *et al.* 2021). We have leveraged this model of OA for use in clonal populations of genome-edited primary human chondrocytes, allowing us to quantify the phenotypic impact of putative target genes of genomic variants in an appropriate disease context.

In this study, we generated CUT&RUN data in primary human chondrocytes and Hi-C data in a human chondrocyte cell line and intersected them with publicly available RNA-seq data from our *ex vivo* OA model, ChIP-seq data from the Roadmap Epigenomics project, and OA GWAS variants from Boer et al. In doing so, we identified 56 putative OA risk genes, including *SOCS2*, whose promoter loops to an OA GWAS variant ~174 Kb away. Deletion of *SOCS2* in primary human chondrocytes using CRISPR-Cas9 led to heightened expression of inflammatory markers in response to treatment with FN-f , providing a possible mechanism for influencing OA risk.

7.2. Results

7.2.1. OA risk variants are enriched in chondrocyte regulatory loci

One of the first steps in decoding GWAS variant mechanisms is to determine the cell types that are likely mediating genetic OA risk. While different risk variants may impact distinct cell types, one approach to help direct research is to determine the cell types which harbor regulatory loci (e.g., enhancers) that are enriched for risk variants. To accomplish this, we performed SNP enrichment analysis using the Genomic Regulatory Elements and Gwas Overlap algoRithm (GREGOR) (Schmidt *et al.* 2015). Publicly available H3K27ac, H3K4me1, and H3K4me3 ChIP-seq peaks from the NIH Roadmap Epigenomics Mapping Consortium (Roadmap) were merged to define regulatory elements for 98 cell types. GREGOR was used to determine each cell type's enrichment for 104 OA GWAS signals recently published in Boer et al. (Boer *et al.* 2021).

The regulatory elements of "Chondrocytes from Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells" (E049) exhibited a strong effect size and p-value of enrichment for OA risk variants (**Fig. 1A**), suggesting that many OA risk variants may impact regulatory events in chondrocytes. This is consistent with the known role of chondrocytes in maintaining joint homeostasis. Chondrocytes have been heavily implicated in OA, as activation of chondrocytes by mechanical and inflammatory stimuli triggers downstream inflammatory and catabolic response pathways in diseased tissue (Sandell and Aigner 2001; Pelletier *et al.* 2001; Loeser *et al.* 2012; Caron *et al.* 2015). An example of an OA risk variant that overlaps a chondrocyte-specific regulatory element (H3K27ac peak) is shown in **Figure 1B**. For comparison, **Figure 1C** shows a non-OA associated variant that overlaps a non-cell type specific, or ubiquitous, enhancer on chromosome 10 that is active in >90% of the 98 cell types evaluated. These examples underscore the importance of interpreting GWAS risk variants in light of the correct cellular context, as the variant-H3K27ac peak overlap shown in **Figure 1B** would not have been detected in any of the other cell types investigated. In addition to E049, IMR90 fetal fibroblasts (E017) and HSMM cell derived Skeletal Muscle Myotubes (E121) were also enriched, suggesting that OA risk variants may also contribute to disease risk through altering the function of fibroblasts and muscle. However, given the strong enrichment in chondrocytes and their documented role in OA biology, we chose to focus our investigation of OA GWAS variants in human chondrocytes.

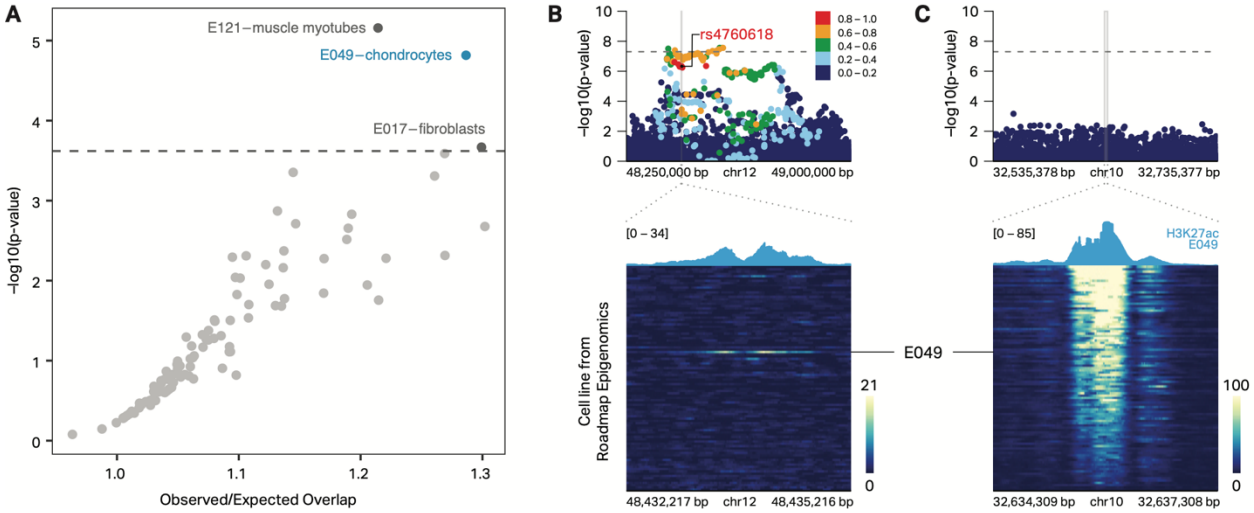


Figure 7.1. OA risk variants are enriched in chondrocyte regulatory elements. (A) Enrichment analysis of 98 cell types from the NIH Roadmap Epigenomics Mapping Consortium reveals that OA GWAS variants are enriched in the regulatory regions (H3K27ac, H3K4me1, or H3K4me3 ChIP-seq peaks) of chondrocytes, skeletal muscle myotubes, and fibroblasts. **(B)** Heatmap of H3K27ac signal from 98 cell types (bottom) highlights a chondrocyte-specific enhancer that overlaps Knee/Hip osteoarthritis risk variant(Boer *et al.* 2021) (rs4760618, circled) that is in high LD ($r^2 > 0.8$) with the lead variant (rs7967762, red diamond) for this locus (top). **(C)** Heatmap of H3K27ac signal from 98 cell types (bottom) highlights a ubiquitous enhancer (active in >90% of cell types) that does not overlap an OA GWAS variant (top).

7.2.2. Multi-omic integration identifies putative variant-gene associations in OA

Due to high LD between variants and the fact the most risk variants reside in non-coding sequences, determining the causal variants and genes they impact remains a major challenge. To address these issues, we generated novel maps of epigenetic features in human chondrocytes and integrated them with GWAS results and publicly available genomic datasets to identify putative variant-gene associations for OA.

First, we identified OA risk variants that are predicted to directly affect protein sequences. We used ENSEMBL's Variant Effect Predictor (VEP) tool to predict the consequences of 1,259 putative OA risk variants that were in high LD ($r^2 > 0.8$) with 104 OA GWAS signals from Boer et al. (Boer *et al.* 2021). VEP identified 29 variants at 19 loci predicted to affect the coding sequence of 24 unique genes (**Fig. 2A, top**). 18 of these variants encode a missense mutation impacting 17 genes, while 11 variants encode a synonymous mutation impacting 8 genes (**Fig. 2A, top**). Though synonymous variants do not impact the protein sequence directly, differences in transcription efficiency, tRNA availability, and mRNA stability introduced through these variants could contribute to the OA phenotype (Venetianer 2012; Zeng and Bromberg 2019). To identify genes most likely to impact OA risk, we incorporated our previously published

RNA-seq FN-f time course data to find genes that change expression in an OA-context. Of the 24 genes identified here, 6 exhibited differential expression in response to FN-f, (**Fig. 2A bottom**). Several of the genes identified have been previously implicated in OA, including Interleukin 11 (*IL11*), Solute Carrier Family 39 Member 8 (*SLC39A8*/*ZIP8*), and Serpin Family A Member 1 (*SERPINA1*). IL11 plays a role in bone turnover and is upregulated in subchondral bone and articular cartilage from OA tissue (Tuerlings *et al.* 2021). SLC39A8 is upregulated in OA chondrocytes and suppression of SLC39A8 in a mouse OA model significantly reduces cartilage degradation (Song *et al.* 2013). SERPINA1, a serine protease inhibitor with anti-inflammatory capabilities (Jain *et al.* 2011), is downregulated in OA (Boeuf *et al.* 2008; Wanner *et al.* 2013).

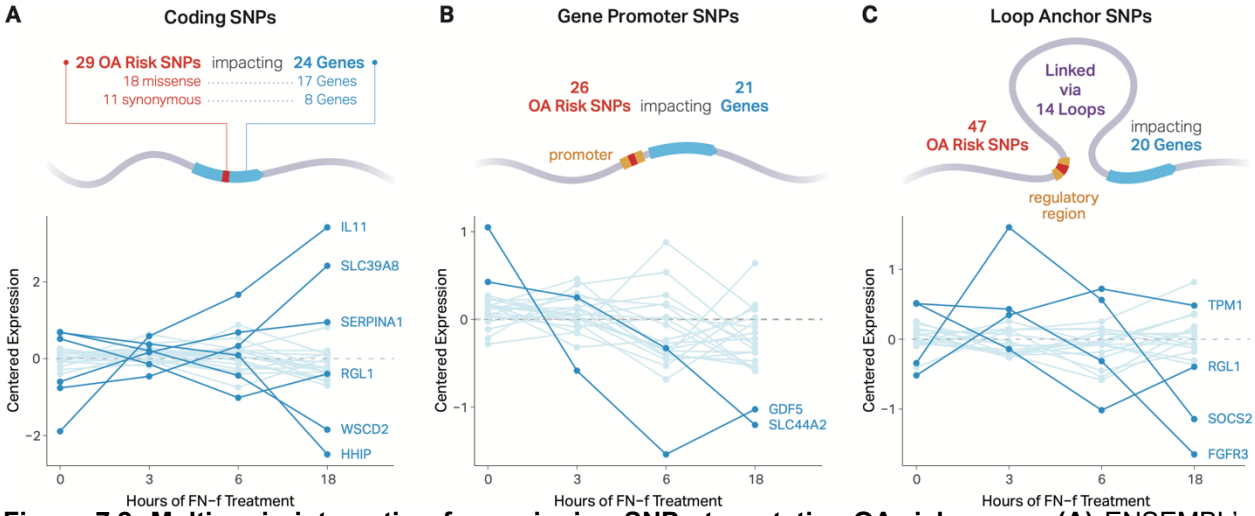


Figure 7.2. Multi-omic integration for assigning SNPs to putative OA risk genes. (A) ENSEMBL's Variant Effect Predictor tool identified 29 unique OA risk SNPs (18 missense and 11 synonymous) overlapping coding regions of 24 unique genes (17 missense, 8 synonymous). **(B)** 26 unique OA risk SNPs overlapped both a chondrocyte regulatory region (H3K27ac, H3K4me1, or H3K4me3 ChIP-seq peaks) and a gene promoter for 21 unique genes. **(C)** 47 unique SNPs overlapped chondrocyte regulatory regions connected to 20 unique gene promoters via 14 C-28/I2 chromatin loops. RNA-seq data from our *ex-vivo* OA model depicts how putative OA risk genes change in response to FN-f. Normalized expression of genes are shown below each category over an 18 hour time-course of fibronectin fragment (FN-f) treatment. Differential genes ($p \leq 0.01$, absolute $\log_2$ fold-change ≥ 1.25) are colored and labeled.

Next, we identified OA risk variants that could impart their phenotypic impact by altering promoter and/or enhancer activity. In order to define the most accurate regulatory regions in chondrocytes we used CUT&RUN to map histone H3K27 acetylation (H3K27ac)—a mark of active enhancers and promoters—in primary human chondrocytes isolated from the knees of 2 cadaveric donors. As expected, Roadmap chondrocyte (E049) H3K27ac peaks showed the highest degree of similarity by jaccard distance to H3K27ac peaks called in our primary human chondrocyte data (**Fig. S1**). Therefore, we merged our primary

chondrocyte H3K27ac peaks with all available marks (including active and repressive marks) from the E049 chondrocyte cell line from Roadmap Epigenomics to define a comprehensive set of chondrocyte regulatory elements. Integration of public and newly generated sources of human chondrocyte features allowed us to identify 507 plausible regulatory variants from 1,259 OA risk variants.

Intersecting these 507 plausible regulatory variants with gene annotations (UCSC) identified 26 unique variants that overlapped the promoters of 21 genes (**Fig. 2B**). Two of these genes were differentially expressed in response to FN-f, both of which have been previously implicated in OA. Growth and Differentiation Factor 5 (GDF5), a member of the TGF-beta family, has roles in skeletal and joint development (Francis-West *et al.* 1999) and has been identified as a major risk locus for OA (Miyamoto *et al.* 2007; Southam *et al.* 2007). Specifically, variants in the GDF5 enhancers *R4* and *GROW1* have been associated with altered anatomical features of the knee and hip, which are thought to confer an increased risk of OA (Capellini *et al.* 2017; Richard *et al.* 2020; Muthuirulan *et al.* 2021). Solute Carrier Family 44 Member 2 (SLC44A2, aka choline transporter-like protein 2) is a mitochondrial choline transporter that has been identified as an expression quantitative trait locus (eQTL) in OA tissue (Steinberg *et al.* 2021) that colocalizes with the OA GWAS signal rs1560707 (Steinberg *et al.* 2020).

In addition to direct regulation of genes by their promoters, long-range regulation of genes also occurs via enhancer-promoter interactions mediated by chromatin loops (Maurano *et al.* 2012). To identify such connections, we conducted deeply sequenced (~2.8 billion reads) in situ Hi-C in C-28/I2 chondrocyte cells and identified 9,271 chromatin loops with Significant Interaction Peak (SIP) caller at 5-Kb resolution which we expanded to 20-Kb for downstream analysis (Rowley *et al.* 2020). C-28/I2 cells were used because they could be expanded to easily provide the number of cells required for Hi-C analysis. To our knowledge, this is the first Hi-C map in a chondrocyte cell line, enabling us to discover novel OA-associated variant-gene connections. We performed four replicates which exhibited a high degree of reproducibility as measured by stratum adjusted correlation coefficient (SCC > 0.98) with the *HiCRep* package (**Fig. S2**) (Yang *et al.* 2017; Lin *et al.* 2021). Overlapping these data with OA risk variants identified 14 loops connecting 47 variants among 14 loci to 20 unique gene promoters (**Fig. 2C**). Four of these genes were differentially expressed in response to FN-f ($p \leq 0.01$, absolute log2 fold-change ≥ 1.25) and are visualized in Figures 2A and S3. Several of these genes have interesting implications for OA, including *FGFR3*

(Fibroblast Growth Factor Receptor 3), which plays a role in skeletal development. FGFR3 may have an important function in the maintenance of articular cartilage (S. Zhou et al. 2016; Tang et al. 2016; Okura et al. 2018), possibly through the Indian hedgehog signaling pathway, which plays a role in regulating chondrocyte hypertrophy and the expression of cartilage matrix-degrading enzymes (Lin *et al.* 2009). FGFR3 is also downregulated in OA tissues, further implicating its potential role in limiting articular cartilage degeneration (Li *et al.* 2012; Shu *et al.* 2016).

Altogether we identified 24 genes impacted by a coding variant, 21 genes with at least one regulatory variant in their promoters, and 20 genes that were connected to a regulatory variant via a chromatin loop. Since genes can fall into multiple categories, the number of total distinct genes identified is 56. All putative variant-gene associations are reported in **Supplementary Table 1**[2]. Boer et al. identified 637 putative effector genes and ranked them by the amount of evidence for association with OA signals (Boer *et al.* 2021). In general, genes with higher tiers of evidence as reported by Boer et al. were more likely to be supported by our analyses (**Fig. S4; Supplementary Table 2**[2]). For example, 67% genes that were supported by six tiers of evidence were also detected in our study, whereas only 2.5% of tier 1 genes were supported by our work. Interestingly, 42% of the genes we identified were not previously implicated by Boer et al. 54% of the genes unique to our study were supported by a chromatin loop compared to only 22% of genes implicated by both studies. This underscores the additional value our study provided by incorporating cell type specific Hi-C data.

7.2.3. Chondrocyte chromatin features identify SOCS2 as a putative regulator of OA

Our multi-omic analysis identified an association between rs7953280 and the promoter of Suppressor Of Cytokine Signaling 2 (*SOCS2).* rs7953280 is located in an intron of the *CRADD* gene, which is expressed at low levels in primary chondrocytes, does not change expression in response to FN-f, and lacks an obvious biological relevance to OA. However, rs7953280 overlaps a putative chondrocyte enhancer (i.e., histone H3K27ac peak), suggesting that it could alter the regulatory capacity of the enhancer and impact the expression of a proximal or distal gene. This enhancer is connected to the promoter of *SOCS2* via a 174 Kb chromatin loop (**Fig. 3A**). Unlike *CRADD*, *SOCS2's* expression changes in response to FN-f, peaking at 3 hours (**Fig. 2C** and **3A orange signal tracks**). Moreover, SOCS2 is known to play a

[2] Supplementary tables are available online at https://doi.org/10.1093/genetics/iyac141.

role in resolving inflammatory response through NFKB and is downregulated in knee OA tissues (de Andrés *et al.* 2011; Paul *et al.* 2017) , making it an intriguing candidate as an OA risk gene. No other SNPs from this locus can be assigned to genes using our integrated approach.

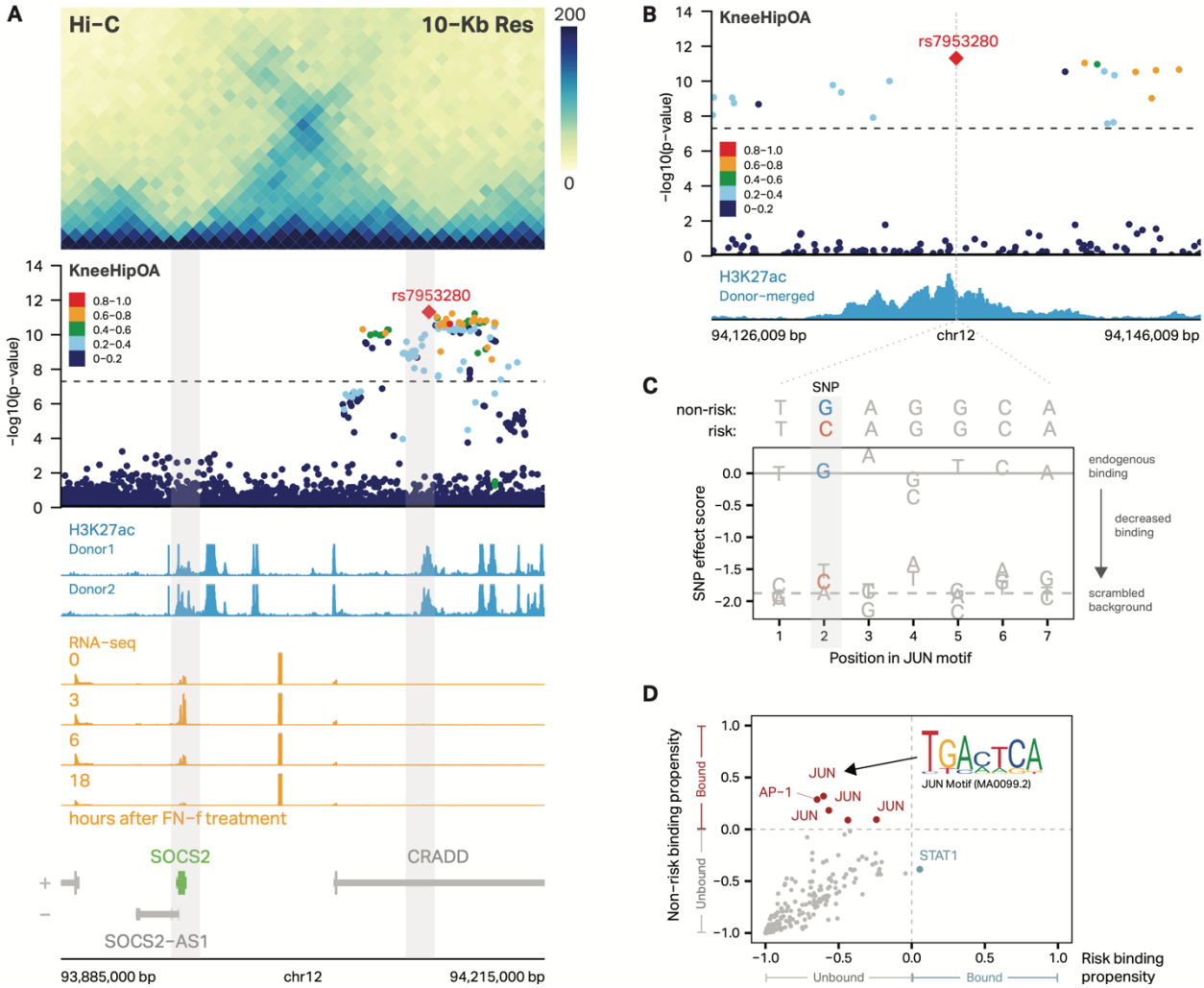


Figure 7.3. 3D chromatin interactions identify SOCS2 as a putative regulator of OA. **(A)** Hi-C performed in C-28/I2 cells reveals a chromatin loop connecting OA risk variant rs7953280 (right gray bar) to the promoter of *SOCS2* (left gray bar). rs7953280 is located in an intronic region of *CRADD* and overlaps an H3K27ac peak in primary articular human chondrocytes from two donors (blue signal tracks). *SOCS2* is differentially expressed in response to treatment with FN-f. Gene tracks are shown below with +/- indicating gene strand. **(B)** Zoom-in on rs7953280 shows that the SNP is located within an H3K27ac peak in primary articular human chondrocytes. **(C)** Motif analysis identifies a JUN binding site at rs7953280. SNP effect matrix (SEM) data predicts decreased binding at the JUN motif (JASPAR ID: MA0099.2) with a G to C polymorphism in the second position. **(D)** Motif analysis from 211 precomputed SEMs from SEMpl predicts that JUN/AP-1 motifs (red, upper left quadrant) bind to the non-risk but not the risk allele.

To further understand how rs7953280 may confer risk for OA, we examined the sequence surrounding rs7953280 to see if it overlaps and alters any transcription factor (TF) binding motifs. Motif comparison with Tomtom from the MEME suite identified FOS and JUN as matching target motifs (**Fig. 3C,**

Supplementary Table 3[2]). FOS and JUN are members of the Activator Protein 1 (AP-1) complex, which is upregulated in response to FN-f (Reed *et al.* 2021), and the inhibition of which prevents cartilage degradation in a model of OA (Motomura *et al.* 2018; Fisch *et al.* 2018; Gao *et al.* 2019). We then used SNP effect matrices (SEMs) generated by the SNP effect matrix pipeline (SEMpl) (Nishizaki *et al.* 2020) to assess the predicted consequence of the G (non-risk) to C (risk) variant on binding of JUN or any other of the 211 motifs included with SEMpl (**Fig. 3D**). Most TFs are predicted to be unbound at both alleles. However, multiple JUN/AP-1 motifs are predicted to bind to the non-risk, but not the OA-risk sequence (**Fig. 3D**) providing further evidence that the G->C mutation in rs7953280 may disrupt JUN/AP-1 binding. Our analysis also showed that STAT-1 was predicted to bind only to the OA-risk sequence, although the SEM score was very close to the cutoff for predicted binding. Nevertheless, since STAT-1 is an important mediator for inflammatory signaling, rs7953280 could influence inflammation during OA progression by modulating STAT-1 binding.

7.2.4. SOCS2 deletion increases proinflammatory gene expression in response to FN-f

To assess the functional role of SOCS2, we used CRISPR-Cas9 to knock out *SOCS2* in primary human chondrocytes isolated from three individual donors. After targeting the *SOCS2* gene with two guide RNAs that flank exon 2 (a constitutive exon that contains the translational start site), we used our previously developed method that employs PCR to screen single-cell-derived colonies (D'Costa *et al.* 2020). The screening primers generated a 1068 bp product if the region was intact and a novel 240 bp amplicon if the two guides successfully deleted the intended 828 bp region (**Fig 4A**). We saw efficient deletion in each of the three donors, with 31% of the colonies showing no deletion, 49% of the colonies showing heterozygous deletion, and 20% showing homozygous deletion (**Fig. 4B**). Sanger sequencing was used to confirm deletions, while qPCR and western blotting confirmed partial (heterozygous) or complete (homozygous) loss of SOCS2 expression (**Fig. S5**).

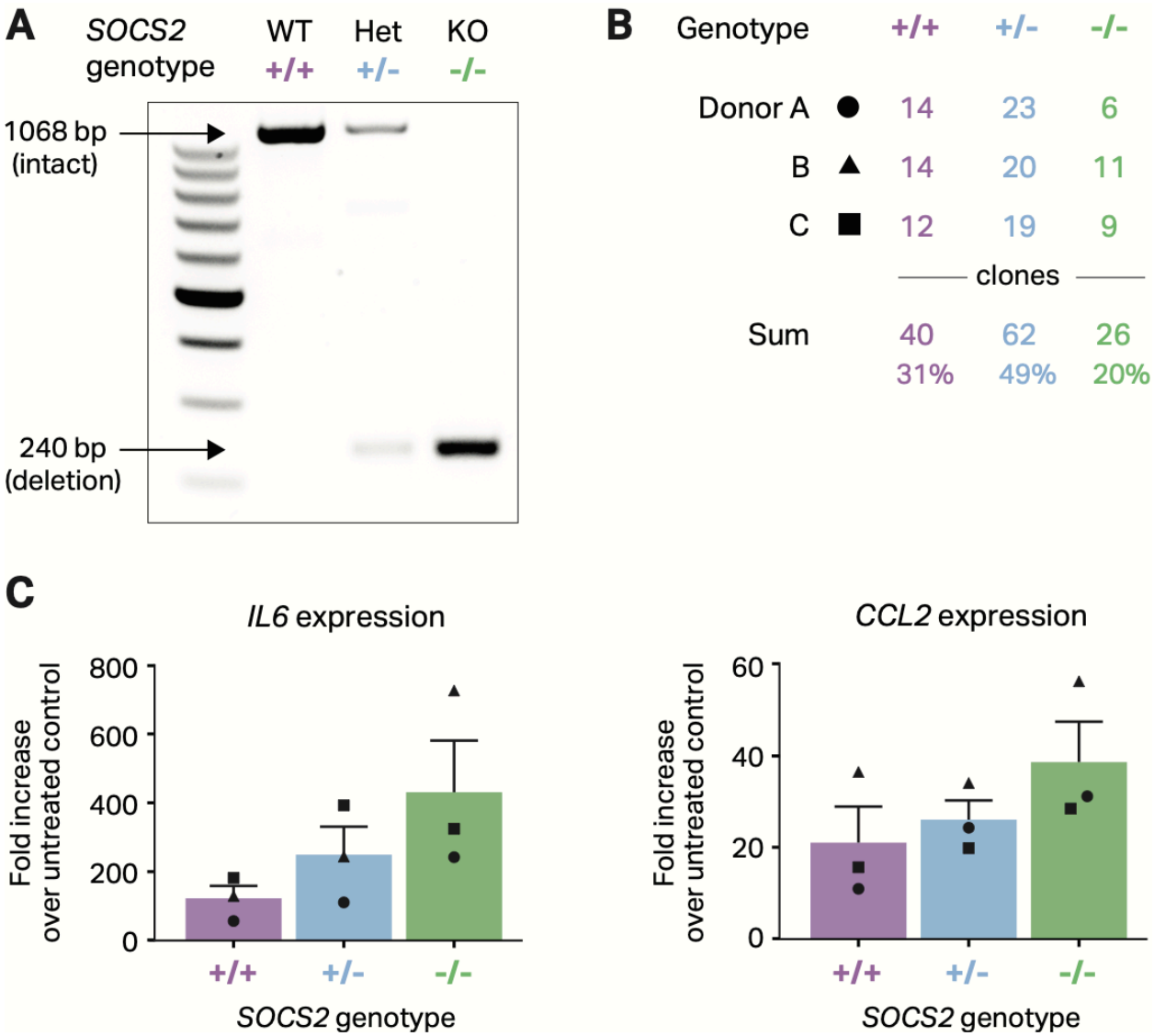


Figure 7.4. *SOCS2* deletion increases proinflammatory gene expression in response to FN-f. (A) PCR primers surrounding the intended *SOCS2* deletion were used to screen single-cell-derived colonies from three independent donors. WT - wildtype (+/+. purple); het - heterozygous deletion (+/-, blue); KO - homozygous knockout deletion (-/-, green). **(B)** Efficiency of deleting the intended *SOCS2* deletion in primary human chondrocytes from three independent donors. **(C)** qPCR at 18 hours after FN-f treatment revealed increased expression of the proinflammatory genes *IL6* and *CCL2* in *SOCS2* deletion colonies from three independent donors.

Because SOCS2 is a known negative regulator of the inflammatory response in other settings (Paul *et al.* 2017; Monti-Rocha *et al.* 2018), we hypothesized that *SOCS2* deletion would lead to an increased expression of inflammatory cytokines in chondrocytes during the response to FN-f. To test this hypothesis, we treated chondrocytes with defined genotypes (wild type, heterozygous, or homozygous knockout) with either FN-f or PBS for 18 hours and quantified the change in proinflammatory cytokines C-C Motif Chemokine Ligand 2 (*CCL2*) and Interleukin 6 (*IL6*) using qPCR. *IL6* and *CCL2* have previously been

shown to exhibit increased expression after 18 hours of FN-f treatment and are also implicated in OA (Wojdasiewicz *et al.* 2014; Wang and He 2018; van den Bosch *et al.* 2020; Reed *et al.* 2021). Deletion of *SOCS2* led to increased expression of both *IL6* and *CCL2* in response to FN-f treatment (**Fig. 4C**), and these increases were observed in a dose-dependent fashion, with greater increases observed in the homozygous compared to heterozygous genotypes. These results suggest that the loss of *SOCS2* may promote a heightened inflammatory response to FN-f stimulation, which is consistent with a potential role in OA.

7.3. Discussion

We used a multi-omic approach to identify putative causal SNPs and genes associated with OA risk. The efficacy of this approach was supported by the identification of previously known OA risk genes including *GDF5*, *SLC44A2*, and *IL11*. We generated the first maps of histone H3K27ac in primary human chondrocytes and integrated this dataset with publicly available genomic datasets to reduce thousands of OA risk GWAS variants to a small list of variants and genes for further study. By generating the first Hi-C contact map of human chondrocytes, we were able to uncover 73 previously unknown connections between OA risk variants and putative target genes. Most looped variant-gene pairs (71 of 73) skipped over the nearest gene, connecting variants to genes as far as 414 Kb away. DNA looping revealed 20 unique genes, 13 of which were not identified by recent fine mapping approaches (Boer *et al.* 2021) and could provide new avenues for therapeutic interventions for OA.

Among the genes identified with Hi-C, four were found to be differentially expressed in our OA model. FGFR3 and SOCS2 have previously been implicated in OA, while Tropomoyosin 1 (TPM1) and Ral Guanine Nucleotide Dissociation Stimulator Like 1 (RGL1) have not. However, TPM1, an actin-binding protein involved in the contractile system of muscle cells and the cytoskeleton of non-muscle cells, has been shown to play roles in an inflammatory response in various cell types, such as human primary coronary artery smooth muscle cells (Li *et al.* 2022) and rod bipolar and horizontal cells in the retina (Gagat *et al.* 2021). RGL1, which functions as a RAS effector protein that activates GTPase by stimulating nucleotide exchange, has also been shown to modulate immune response in both vascular and immune cells (Kirkby *et al.* 2014), and, interestingly, is downregulated in human articular chondrocytes upon

treatment with interleukin-1 and oncostatin-M (Yang *et al.* 2021). The functions of TPM1 and RGL1 in inflammatory responses may point to potentially undiscovered roles in osteoarthritis.

One especially intriguing gene was *SOCS2*, whose promoter is looped to an OA risk SNP within a histone H3K27ac peak ~170 Kb away. SOCS2 is known to inhibit the JAK/STAT pathway and is induced by various pro-inflammatory cytokines such as interleukin-6, growth hormone, and tumor necrosis factor-alpha (Starr *et al.* 1997; Metcalf *et al.* 2000; Santangelo *et al.* 2005). CRISPR-mediated deletion of *SOCS2* was associated with increased expression of *IL6* and *CCL2* in our *ex vivo* model of OA, suggesting that it may also play a role in mediating inflammation in response to cartilage matrix damage. These findings make SOCS2 a candidate for further studies and the activation of more robust SOCS2 expression could be a goal for future therapeutic development. The regulatory role of SOCS2 in chondrocytes is likely to be subtle, as *Socs2* knockout mice did not show altered OA development (Samvelyan *et al.* 2022). Because that study used a global germline deletion, other members of the inflammatory cascade may have compensated for *Socs2* loss, and it would be interesting to determine whether the inducible loss of *Socs2* in adult chondrocytes would generate a different result.

Expression quantitative trait locus analysis (GTEx Project v8) provides evidence that variation at rs7953280 is associated with *SOCS2* expression in fibroblasts. In that data set the C allele is associated with increased expression of *SOCS2*. This is contrary to what we predict for chondrocytes but could be explained by differences in cell type or condition (e.g., resting vs stimulated with FN-f). Mapping of QTLs in chondrocytes responding to FN-f could shed light on these differences.

While further work is needed to clarify the role of rs7953280 and SOCS2 in mediating OA risk, our multi-omic analysis suggests the following potential model. In cells harboring the non-risk variant, proinflammatory cytokines such as IL-6 and matrix damage products such as FN-f may activate AP-1 via the JAK/STAT pathway. AP-1 may then bind the enhancer at the rs7953280 locus, increase enhancer activity, and upregulate transcription of *SOCS2* via a chromatin loop between the enhancer and the *SOCS2* promoter. In cells harboring the risk allele, AP-1 binding would be decreased, impeding enhancer activation and proper upregulation of *SOCS2*. As a result, JAK/STAT signaling would remain high, resulting in prolonged or heightened inflammation and further cartilage degradation.

This model, while compelling, will require further experimental investigation and validation. One such experiment would be to use genome editing of noncoding sequences to directly test the effect of rs7953280 on *SOCS2*. While implementing single base changes using CRISPR/Cas9 and homology-directed repair donor oligos in chondrocytes is technically challenging, engineering isogenic chondrocytes with the risk and protective alleles will help validate the association between the variant and *SOCS2* expression and the observed inflammatory response. Moreover, future experiments are required to determine to the degree to which these findings translate from our *ex vivo* model into an *in vivo* system and/or if activation of *SOCS2* could provide therapeutic avenue for OA treatment.

We generated the first maps of histone H3K27ac in primary human chondrocytes, provided the first maps of 3D chromatin contacts in chondrocytes of any type, and identified 56 putative OA risk genes using multi-omic data integration. For each locus, we provide 0, 1 or multiple putative OA risk genes. While these analyses narrow the search space for the genes affected by OA risk variants and allow for the formation of new hypotheses, determining which genes are truly causal will require further experimental validation similar to the approach described here to investigate SOCS2. We chose to perform functional experiments on SOCS2 because it had the strongest genomic evidence for mediating OA risk of the genes looped to OA risk variants; however, many of the other genes implicated here (via looping or otherwise) may also influence disease risk and warrant further investigation. These putative risk genes and novel epigenetic datasets will provide a foundation for future studies to investigate the genetic variants responsible for OA risk and expedite our search for better prevention and treatment of OA.

7.4. Methods

7.4.1. Primary chondrocyte isolation and culture

Primary articular chondrocytes were isolated via enzymatic digestion from human talar cartilage obtained from tissue donors, without a history of arthritis, through the Gift of Hope Organ and Tissue Donor Network (Elmhurt, IL) as previously described (Loeser *et al.* 2003; Reed *et al.* 2021)**.** For Cut and Run, two million primary articular chondrocytes from two male donors, ages 39 and 63, were plated onto four 6cm plates in DMEM/F12 media supplemented with 10% fetal bovine serum, 1% penicillin streptomycin solution, 1% amphotericin B, and 0.04% gentamicin. For genome editing, primary chondrocytes from three male

donors ages 56, 59, and 64 years were cultured in 6 or 10cm dishes at a density of approximately 70,000 cells/cm^2 in DMEM/F12 media supplemented with 10% FBS and antibiotics.

7.4.2. Fibronectin fragment (FN-f) treatment

After serum starvation, cells were treated with either purified 42 kDa endotoxin-free recombinant FN-f (final concentration 1μM in PBS), prepared as previously described, or PBS as control (Wood *et al.* 2016). Cells were harvested and crosslinked after 90 min or 180 min and immediately subjected to Cut & Run, described below.

7.4.3. Hi-C

C-28/I2 cells were cultured in DMEM/F12 media with 10% fetal bovine serum, 1% penicillin streptomycin solution, 1% amphotericin B, and 0.04% gentamicin. Cells were treated with DMEM/F12 media with 1% ITS-Plus for 48 hours prior to experiments to promote the chondrocyte phenotype. Cells were then washed with 1X PBS and treated with trypsin-EDTA (0.25%) for 3 minutes. Trypsin was quenched and cells were pelleted at 4°C for 5 minutes at 300*g*. Cells were resuspended in 1mL DMEM/F12 per million cells and crosslinked in 1% formaldehyde for 10 min with rotation before quenching in a final concentration of 0.2M glycine for 5 min with rotation. Cells were pelleted at 300*g* for 5 min at 4°C. Pellets were washed with cold PBS and aliquoted into ~3 million cell aliquots. Pellets were flash frozen in liquid nitrogen and stored at -80°C.

In situ Hi-C was performed as previously described (Rao *et al.* 2014). Pellets were lysed in ice-cold Hi-C lysis buffer (10mM Tris-HCl pH 8.0, 10mM NaCl, 0.2% IGEPAL CA630) with 50μL of protease inhibitors for 15 min on ice. Cells were pelleted and washed using the same buffer. Pellets were resuspended in 50μL 0.5% SDS and incubated at 62°C for 7 min. Reactions were quenched with 145μL water and 25μL 10% Triton X-100 at 37°C for 15 min. Chromatin was digested overnight with 25μL 10X NEBuffer2 and 100U MboI at 37°C with rotation.

Reactions were incubated at 62°C for 20 min then cooled to RT. Fragment overhangs were repaired by adding 37.5μL 0.4mM biotin-14-dATP; 1.5μL each 10mM dCTP, dGTP, dTTP; 8μL 5U/μL DNA Polymerase I, Large (Klenow) Fragment and incubating at 37°C for 1.5 h with rotation. Ligation was performed by adding 673μL water, 120μL 10X NEB T4 DNA ligase buffer, 100μL 10% Triton X-100, 6μL 20mg/mL BSA, and 1μL 2000U/μL T4 DNA ligase and incubating at RT for 4 h with slow rotation. Samples

were pelleted at 2500g, resuspended in 432µL water, 18µL 20mg/mL proteinase K, 50µL 10% SDS, and 46µL 5M NaCl, incubated at 55°C for 30 min, and then transferred to 68°C overnight.

Samples were cooled to RT and 1.6x volumes of pure ethanol and 0.1x volumes of 3M sodium acetate pH 5.2 were added to each sample, prior to incubation at -80°C for over 4-6 h. Samples were spun at max speed at 2°C for 15 min and washed twice with 70% ethanol. Pellets were dissolved in 130µL 10mM Tris-HCl pH 8.0 and incubated at 37°C for 1-2 h. Samples were stored at 4°C overnight.

DNA was sheared using the Covaris LE220 (Covaris, Woburn, MA) to a fragment size of 300-500bp in a Covaris microTUBE. DNA was transferred to a fresh tube and the Covaris microTUBE was rinsed with 70µL of water and added to the sample. A 1:5 dilution of DNA was run on a 2% agarose gel to verify successful shearing.

Sheared DNA was size selected using AMPure XP beads. 0.55x volumes of 2X concentrated AMPure XP beads were added to each reaction and incubated at RT for 5 min. Beads were reclaimed on a magnet and the supernatant was transferred to a fresh tube. 30µL 2X concentrated AMPure XP beads were added and incubated for 5 min at RT. Beads were reclaimed on a magnet and washed with fresh 70% ethanol. Beads were dried for 5 min at RT prior to DNA elution in 300µL 10mM Tris-HCl pH 8. Undiluted DNA was run on a 2% agarose gel to verify successful size selection between 300-500 bp.

150µL 10mg/mL Dynabeads MyOne Streptavidin T1 beads were washed with 400µL 1X Tween washing buffer (TWB; 250µL Tris-HCl pH 7.5, 50µL 0.5M EDTA, 10mL 5M NaCl, 25µL Tween 20, 39.675µL water). Beads were then resuspended in 300µL 2X Binding Buffer (500µL Tris-HCl pH 7.5, 100µL 0.5M EDTA, 20mL 5M NaCl, 29.4mL water), added to the DNA sample, and incubated at RT for 15 min with rotation. DNA-bound beads were then washed twice with 600µL 1X TWB at 55°C for 2 min with shaking. Beads were resuspended in 100µL 1X NEBuffer T4 DNA ligase buffer, transferred to a new tube, and reclaimed.

Sheared ends were repaired by resuspending the beads in 88µL 1X NEB T4 DNA Ligase Buffer with 1mM ATP, 2µL 25mM dNTP mix, 5µL 10U/uL NEB T4 PNK, 4uL 3U/uL NEB T4 DNA polymerase I, and 1uL 5U/uL NEB DNA polymerase 1, large (Klenow) fragment and incubating at RT for 30 min. Beads were washed two more times with 1X TWB for 2 min at 55°C with shaking. Beads were washed once with 100uL 1X NEBuffer 2, transferred to a new tube, and resuspended in 90uL 1X NEBuffer 2, 5uL 10mM

dATP, and 5uL NEB Klenow exo minus, and incubated at 37°C for 30 min. Beads were washed two more times with 1X TWB for 2 min at 55°C with shaking. Beads were washed in 100uL 1X Quick Ligation Reaction Buffer, transferred to a new tube, reclaimed, and resuspended in 50uL 1X NEB Quick Ligation Reaction Buffer. 2uL NEB DNA Quick Ligase and 3uL of an appropriate Illumina indexed adapter (TruSeq nano) were added to each sample before incubating at RT for 15 minutes. Beads were reclaimed and washed twice with 1X TWB for 2 min at 55°C. Beads were washed in 100uL 10mM Tris-HCl pH 8, transferred to a new tube, reclaimed, and resuspended in 50uL 10mM Tris-HCl pH 8.

Hi-C libraries were amplified directly off T1 beads with 8 cycles in 5uL PCR primer cocktail, 20uL Enhanced PCR mix, and 25uL of DNA on beads. The PCR settings were as follows: 3 min at 95°C followed by 4-12 cycles of 20s 98°C, 15s at 60°C, and 30s at 72°C. Samples were held at 72°C for 5 min before holding at 4°C. Amplified samples were transferred to a new tube and brought to 250uL in 10mM Tris-HCl pH 8.

Beads were reclaimed and the supernatant containing the amplified library was transferred to a new tube. Beads were resuspended in 25uL 10mM Tris-HCl pH 8 and stored at -20°C. 0.7x volumes of warmed AMPure XP beads were added to the supernatant sample and incubated at RT for 5 min. Beads were reclaimed and washed with 70% ethanol without mixing. Ethanol was aspirated. Beads were resuspended in 100uL 10mM Tris-HCl pH 8, 70uL of fresh AMPure XP beads were added, and the solution was incubated for 5 min at RT. Beads were reclaimed and washed twice with 70% ethanol without mixing. Beads were left to dry and DNA was eluted in 25uL 10mM Tris-HCl pH 8. The resulting libraries were then quantified by Qubit and Tapestation. A low depth sequence was performed first using the Miniseq sequencer system (Illumina) and analyzed using the Juicer pipeline to assess quality. The resulting libraries underwent paired-end 2x150bp sequencing on an Illumina NovaSeq sequencer. Each replicate was sequenced to an approximate depth of 750 million reads. The full sequencing depth was 2.8 billion reads.

7.4.4. Hi-C data processing

In situ Hi-C datasets were processed using a modified version of the Juicer Hi-C pipeline (https://github.com/EricSDavis/dietJuicer) with default parameters as previously described (Durand, Shamim, et al. 2016). Reads were aligned to the hg19 human reference genome with bwa (v0.7.17) and MboI was used as the restriction enzyme. Four biological replicates were aligned and merged for a total of

2,779,816 Hi-C read pairs in C-28/I2 cells yielding 2,373,892,594 valid Hi-C contacts (85.40%). For visualization, the merged Hi-C contact matrix was normalized with the “KR” matrix balancing algorithm as previously described (Knight and Ruiz 2013) to adjust for regional background differences in chromatin accessibility.

Looping interactions were called at 5-Kb resolution with Significant Interaction Peak (SIP) caller (Rowley *et al.* 2020) (v1.6.2) and Juicer tools (v1.14.08) using the replicate-merged, mapq > 30 filtered hic file with the following parameters: “-norm KR -g 2.0 -min 2.0 -max 2.0 -mat 2000 -d 6 -res 5000 -sat 0.01 -t 2000 -nbZero 6 -factor 1 -fdr 0.05 -del true -cpu 1 -isDroso false”. Loop anchors were expanded to 20 Kb and loops with overlapping anchors were filtered out (14 loops). This resulted in 9,271 loops after filtering.

7.4.5. Cut and Run

Primary chondrocytes were washed with 1X PBS and treated with trypsin-EDTA (0.25%) for 5 minutes. Trypsin was quenched and cells were pelleted at 4°C for 5 minutes at 1000*g*. Cells were resuspended in 1mL plain DMEM per million cells and crosslinked in 1% formaldehyde for 10 min with rotation before quenching in a final concentration of 125 mM glycine for 5 min with rotation. Cells were pelleted by spinning at 1000*g* for 5 min at 4°C. Each 2 million cell pellet was washed in 1 mL cold PBS prior to flash freezing in liquid nitrogen. We performed Cut and Run following existing protocols (Skene and Henikoff 2017) but modified for crosslinked cells.

Following flash freezing, thawed pellets were resuspended in 1mL ice-cold nuclei isolation buffer (NE1 buffer; 20mM HEPES pH 7.5, 10mM KCl, 1mM MgCl2, 1mM DTT, 0.1% Triton X-100, 1X CPI added fresh) and incubated for 10 min at 4°C with rotation. Nuclear pellet was collected by centrifugation at 1000g for 5 min at 4°C, then resuspended in 1mL of ice-cold wash buffer (WB; 20mM HEPES pH 7.5, 0.2% Tween-20, 150mM NaCl, 150mM BSA, 0.5mM Spermidine, 10mM Na-Butyrate, 1X CPI added fresh). 10uL concanavalin A lectin beads washed and resuspended in binding buffer (BB; 20mM HEPES pH 7.5, 10mM KCl, 1mM CaCl2, 1mM MnCl2) were add- ed to each sample and incubated for 10 min at RT. Beads were reclaimed and resuspended in 50uL antibody buffer (AbB; WB supplemented with 0.1% Triton X-100 and 2mM EDTA). 0.01ug/uL H3K27ac antibody in AbB was added to each sample and samples were incubated overnight at 4°C with mixing at 1000xRPM.

Beads were reclaimed and washed with 1mL triton wash buffer (TwB; WB supplemented with 0.1% Triton X-100) without mixing. Beads were reclaimed and resuspended in 50uL AbB. 2.5uL CUTANA pAG-MNase (Epicypher, #15-1016) was added and samples were incubated for 1 h at 4°C on a metal block. Beads were reclaimed and washed twice with 1mL TwB before resuspension in 100uL TwB. To digest chromatin, 2uL 100mM CaCl2 was added and samples were incubated for 30 min at 4°C on a metal block. Digestion was halted by the addition of 100uL 2X STOP buffer (340mM NaCl, 20mM EDTA, 4mM EGTA, 0.1% Triton X-100, 50ug/mL RNAse A). Samples were incubated for 20 min at 37°C to release pA-MNase cleaved fragments from nuclei). Beads were placed on a magnet and the supernatant containing DNA fragments was transferred to a new tube. To reverse crosslinks, 2uL 10% SDS and 2uL 20mg/mL proteinase K were added to each sample and incubated for 1 h at 65°C. DNA was purified using the Zymo DNA Clean & Concentrator Kit according to manufacturer's protocols using 5 volumes of DNA binding buffer. DNA was eluted in 55uL water.

Sequencing libraries were prepared from CUT&RUN fragments using KAPA HyperPrep with library amplification kit (no. KK8504) following the manufacturer's instructions. Post-ligation bead cleanup was performed with two rounds of 1.2X volumes of beads and DNA was eluted in a final volume of 25uL 10mM Tris-HCl pH 8. Library amplification was performed with 20uL of the adapter ligated DNA with 12 PCR cycles. One round post amplification cleanup was performed with 1.2X volumes of beads. The resulting libraries were then quantified by Qubit and Tapestation. A low depth sequence was performed first using the Miniseq sequencer system (Illumina) and analyzed using the Juicer pipeline to assess quality control. The resulting libraries underwent paired-end 2x150bp sequencing on an Illumina NextSeq sequencer.

7.4.6. Cut and Run data processing and peak calling

Adaptors and low-quality reads were trimmed from paired-end reads using Trim Galore! (v0.4.3). Reads were aligned to the hg19 genome with BWA mem (v0.7.17) and sorted with Samtools (v1.9). Duplicates were removed with PicardTools (v2.10.3) and mitochondrial reads were removed with Samtools idxstats. Samtools was also used to merge donors, and index BAM files. Peaks were called from the merged alignments using MACS2 with the following settings: -f BAM -q 0.01 -g hs --nomodel --shift 100 --extsize 200 --keep-dup all -B --SPMR (v2.1.1.20160309). Peaks were then merged using bedtools (v2.26), and

multicov was used to extract counts from each replicate BAM file. Signal tracks were made from alignments using deeptools (v3.0.1).

7.4.7. Preparation of gRNA: Cas9 RNP complex

Two custom SOCS2 Alt-R crRNAs TGACAAGGGCCTATTCCCAC and TTACGCATTCCCAAGGACCC were synthesized by Integrated DNA technologies (IDT). Both sequences are written 5' to 3' and do not include PAM sequence. The first crRNA targets the plus strand and the second the minus strand. Ribonucleoprotein (RNP) complexes containing the Cas9 enzyme and sequence-targeting guide RNAs were prepared according to the manufacturer's recommendation. Briefly, Alt-R tracrRNA (1072533, IDT) and crRNA were resuspended in Tris-EDTA buffer to 100 µM concentration and equimolar concentration of crRNA and tracrRNA was combined, heated at 95 °C for 5 min and cooled to room temperature to produce the gRNA. Separate RNP complex for each guide was prepared by combining the gRNA (50 µM) with Alt-R® Cas9 Nuclease (61 µM) (1081058, IDT) and PBS at a ratio of 1:1.1:2 µl at room temperature for 15 min.

7.4.8. Transfection of primary human chondrocytes with RNP complex and single cell colony selection

Chondrocytes were trypsinized, washed with PBS and transfected with the RNP complex as previously described with modifications; volumes were scaled up for transfection of more cells in larger cuvettes (D'Costa *et al.* 2020). Two million cells were resuspended in 100 µl of P3 Primary Cell Nucleofector™ solution (V4XP-3024, Lonza). The RNP complex and Alt-R® Cas9 Electroporation Enhancer (1075916, IDT) was added to the cells. The mixture was gently pipetted up and down and transferred to 100 µl Nucleocuvette vessels (V4XP-3024, Lonza) and transfected using program ER-100 on a 4D-Nucleofector™ Core unit (Lonza). Cells were kept at room temperature for 8 minutes and then incubated in prewarmed antibiotic free media containing 20% FBS for recovery. An aliquot of the transfected cells was placed in a 96-well and used for DNA extraction and PCR. Following confirmation of editing, the transfected bulk cells were seeded at low cell density (200 cells per 6 cm^2 dish) for generation of single-cell colonies. Individual colonies were picked under a microscope (EVOS FL, ThermoFisher), the colony was disrupted by pipetting and split into 96- and 24-well plates for genetic analysis and continued expansion, respectively.

7.4.9. PCR screening of genome-edited bulk and single-cell derived colonies

DNA was extracted using QuickExtract™ DNA Extraction Solution (Lucigen), depending on confluency 25 to 100 µl of solution was added to the wells containing the cells for 15 minutes at 37C, cell suspension was transferred to tubes and vortexed for a minute. Samples were then heated at 65 °C for 6 minutes, and 98 °C for 2 minutes. The extracted DNA solution was stored at -20 °C. PCR amplification was performed by adding 4 or 5 µl of template DNA, 1 µM forward (SOCS2_F1: accaagtttgtgtgggtgct) and reverse (SOCS2_R1: cttccagcgtgctaagaagc) primers, and EconoTaq PLUS GREEN 2X Master Mix (Lucigen) in a 25 µl reaction. PCR conditions included an initial denaturation at 94 °C for 2 minutes, 35 cycles of denaturation at 94 °C for 30 seconds, annealing at 63 °C for 30 seconds, and extension at 72 °C for 65 seconds, followed by a final extension at 72 °C for 10 minutes. Following amplification of column purified genomic DNA, the PCR product was cleaned up and sequenced using the primers described above and the Bioedit software was used to visualize the chromatograms.

7.4.10. Fibronectin fragment (FN-f) treatment and qPCR analysis of genome edited samples

Single cell colonies in 24-well plates were passaged to 6-well plates for expansion. Following genotype confirmation, colonies with similar genotype were combined and seeded at 250,000 cells per well in a 12-well plate. Cultured cells were made serum free and treated with FN-f 1 µM or PBS. Following treatment, media was removed and cells were immediately lysed in the RLT buffer. RNA was isolated with RNeasy Plus columns (Qiagen) and reverse transcribed to cDNA using qScript™ XLT cDNA SuperMix (VWR) or iScript cDNA Synthesis Kit (1708891, Bio-Rad). DNase treatment was used for the second and third donors in order to confirm that detectable *SOCS2* signal in knockout cells was due to the presence of genomic DNA. To evaluate the effect of *SOCS2* editing on inflammatory gene response quantitative polymerase chain reaction (qPCR) was performed on a QuantStudio™ 6 Flex machine (Applied Biosystems) with TaqMan™ Universal Master Mix and TaqMan Gene Expression Assays for human *CCL2* (Hs00234140_m1), *IL6* (Hs00174131_m1), and housekeeping gene *TBP* (Hs00427620_m1). SOCS2 expression was assessed in pooled colonies with TaqMan Gene Expression Assay Hs00919620_m1.

7.4.11. Western Blot analysis

Following genotype identification by PCR, cells from a wildtype, heterozygous and knockout colony were expanded in chondrocyte media supplemented with 5 ng/ml bFGF and 1 ng/ml TGF-β1 (Life

technologies) for 11 days. Cells were lysed in standard cell lysis buffer (1X) (Cell signaling technology) containing phenylmethanesulfonyl fluoride (PMSF; Sigma-Aldrich) and phosphatase inhibitor mix. Protein (15 µg) was separated by SDS-PAGE and transferred to nitrocellulose membrane. After blocking in 5% nonfat milk in TBST, the blot was incubated with SOCS2 antibody (PA5-17219; 1:1000; Thermo Fisher) overnight at 4C and secondary antibody solution for 1 hour. The membrane was incubated in Radiance Plus Chemiluminescent Substrate (Azure Biosystems) and signal detected using the Azure c600 gel imaging system. The membrane was striped and incubated with the loading control beta tubulin antibody.

7.4.12. Osteoarthritis GWAS

Genome-wide association statistics for 11 osteoarthritis phenotypes and lead variants identified in Boer et al. (Boer *et al.* 2021) were obtained from the Musculoskeletal Knowledge Portal (Kiel *et al.* 2020).

7.4.13. Epigenome Roadmap Data

Consolidated reference human epigenomes for 98 cell/tissue types were obtained from the NIH Roadmap Epigenomics Project (Bernstein *et al.* 2010) and The Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium 2012). Processed narrowPeak files for H3K27ac, H3K4me1, and H3K4me3 and BigWig files for H3K27ac were used for each cell/tissue type. Additional narrowPeak files for H3K9ac, H3K9me3, H3K27me3, and H3K36me3 were obtained for mesenchymal stem cell derived chondrocyte cultured cells (E049).

7.4.14. RNA-seq time course of fibronectin fragment (FN-f) treatment

RNA-seq data from a prior study of FN-f treated human chondrocytes was obtained from KSM Reed et al. (Reed *et al.* 2021) and vst-normalized, centered, and replicate-combined. The 0-hour FN-f treatment time point was created by combining the 9 PBS-treated replicates. Genes were considered differential with a BH-adjusted p-value of 0.01 and a log2 fold-change threshold > 1.25 across any time point.

7.4.15. Cell type enrichment for OA risk variants

To identify the cell types that likely mediate genetic OA risk, we performed SNP enrichment analysis using GREGOR (Genomic Regulatory Elements and Gwas Overlap algoRithm) (Schmidt *et al.* 2015). Publicly available H3K27ac, H3K4me1, and H3K4me3 ChIP-seq narrowPeaks files from the NIH Roadmap Epigenomics Mapping Consortium were merged and sorted using bedtools (v2.29.2) (Quinlan and Hall

2010) to define regulatory loci for 98 cell types in hg19. GREGOR was used to determine each cell type's enrichment for 104 OA lead SNPs (Boer *et al.* 2021) by comparing the observed overlap between regulatory loci and SNPs with their expected overlap and evaluating significance. Expected overlap is determined using a matched control set of ~500 variants that control for the number of LD proxies, gene proximity and minor allele frequency. Reference data from 1000 Genomes Phase 1 version 2 EUR panel were used with GREGOR to control for LD proxies (1Mb, $r^2 > 0.7$) (1000 Genomes Project Consortium *et al.* 2010). Results were imported into R (v4.1.0) (R Core Team 2022) and visualized with ggplot2 (Wickham 2016) and plotgardener (Kramer *et al.* 2022).

7.4.16. Putative OA risk variants

LD proxies for 104 OA GWAS signals from Boer et al. were identified using the 1000 Genomes European reference panel since the GWAS data primarily analyzed individuals of European ancestry (11 of 13 cohorts are of European descent). r^2 values were calculated with the –ld function in PLINK 1.9 (Purcell *et al.* 2007; 1000 Genomes Project Consortium *et al.* 2010) using a window of 1 Mb for LD calculation. Putative OA risk variants were defined as those in high LD ($r^2 > 0.8$, n = 1,259) with lead variants.

7.4.17. Multi-omic integration for assigning SNPs to putative OA risk genes

We took a multi-omic approach to identify putative SNP-gene pairs implicated in OA. SNPs that 1) were predicted to affect coding regions of genes, 2) overlapped gene promoters, or 3) overlapped a regulatory peak looped to a gene's promoter were assigned to the "Coding gene", "Gene promoter", or "Loops to gene promoter" categories, respectively. Genes in each category that change in response to FN-f ($p \leq 0.01$ and LFC at any time point ≥ 1.25) were highlighted as putative OA risk genes.

Coding SNP-gene pairs were identified using ENSEMBL's Variant Effect Predictor (VEP) tool. Putative OA risk variants (n = 1,259) were annotated with their predicted consequence on coding sequence using VEP run with the GRCh37.p13 human genome and default parameters. SNPs with a predicted consequence of "missense" or "synonymous" were paired with their affected genes assigned to the "Coding gene" category.

Promoter regions were defined as 2000 bp upstream and 200 bp downstream of the TSS of transcripts obtained with the TxDb.Hsapiens.UCSC.hg19.knownGene Bioconductor package for a total of 8,2960 transcripts. Gene symbols were linked to transcript ranges using the OrganismDbi and

Homo.sapiens packages. Transcripts without gene symbols or those not present in the FN-f RNA-seq data were filtered out, leaving a total of 62,590 transcript promoters.

Chondrocyte regulatory regions were defined by combining Roadmap Epigenomics data with data from primary human articular chondrocytes. Specifically, H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, and H3K36me3 peaks from mesenchymal stem cell derived chondrocyte cultured cells (E049) obtained through AnnotationHub (v3.1.7, snapshot date 2021-10-20) (Morgan *et al.* 2017) were combined with Donor-merged H3K27ac peaks from primary human articular chondrocytes. OA SNPs were overlapped with chondrocyte regulatory regions resulting in 507 SNPs.

SNPs overlapping chondrocyte regulatory regions that also overlapped a promoter region were assigned to their affected gene and added to the “Gene promoter” category. SNPs overlapping chondrocyte regulatory regions were intersected with loop calls from Hi-C in the C-28/I2 chondrocyte cell line (see Methods on Hi-C processing and loop calling). The linkOverlaps function from the InteractionSet package was used to identify chondrocyte regulatory SNPs that are connected to promoters by loops. These SNP-gene pairs were assigned to the “Loops to gene promoter” category.

7.4.18. Motif Analysis

Tomtom (v5.4.1; release date: Sat Aug 21 19:23:23 2021 -0700) from the MEME suite was used to identify motif matches for sequences surrounding the rs7953280 variant (Gupta *et al.* 2007). All 7-mers surrounding rs7953280 ("GGCTTTG", "GCTTTGA", "CTTTGAG", "TTTGAGG", "TTGAGGC", "TGAGGCA", "GAGGCAT") and the entire 13 bp sequence ("GGCTTTGAGGCAT") were used to identify motif matches. Sequences were input into the online motif comparison tool and queried against the JASPAR2022_CORE_vertebrates_non-redundant_v2 and HOCOMOCOv11_core_HUMAN_mono_meme_format motif databases. Pearson correlation coefficient was used as the motif column comparison function and the significance threshold was set to an E-value < 10; no q-value threshold was set and reverse complementing of motifs was permitted. The following command summarizes the parameters used: “tomtom -no-ssc -oc . -verbosity 1 -min-overlap 5 -mi 1 -dist pearson -evalue -thresh 10.0 -time 300 query_motifs motif_databases."

7.4.19. Transcription factor (TF) motif binding propensity

We used SNP Effect Matrix scores (SEMs) to predict the TF binding propensity between risk and non-risk SNPs in OA. Pre-calculated SEMs for 211 TF motifs were obtained from SEMpl (https://github.com/Boyle-Lab/SEMpl) and used for scoring risk and non-risk SNP sequences (Nishizaki *et al.* 2020). Binding propensity scores were determined by generating frame-shifted K-mers covering each TF motif position for both risk and non-risk sequences. K-mers were scored against 211 TF SEMs using position-weight matrix (PWM) scoring functions from the Biostrings Bioconductor package (Pagès *et al.* 2021). The best scoring K-mer frame for each TF motif was used to select the binding score for risk and non-risk sequences. Scores were normalized by applying inverse-log transformation, subtracting the scrambled baseline provided with each SEM, and dividing the result by the absolute value of that baseline. TFs with positive scores are predicted to be bound while negative scores are predicted to be unbound.

7.5. Supplementary Figures

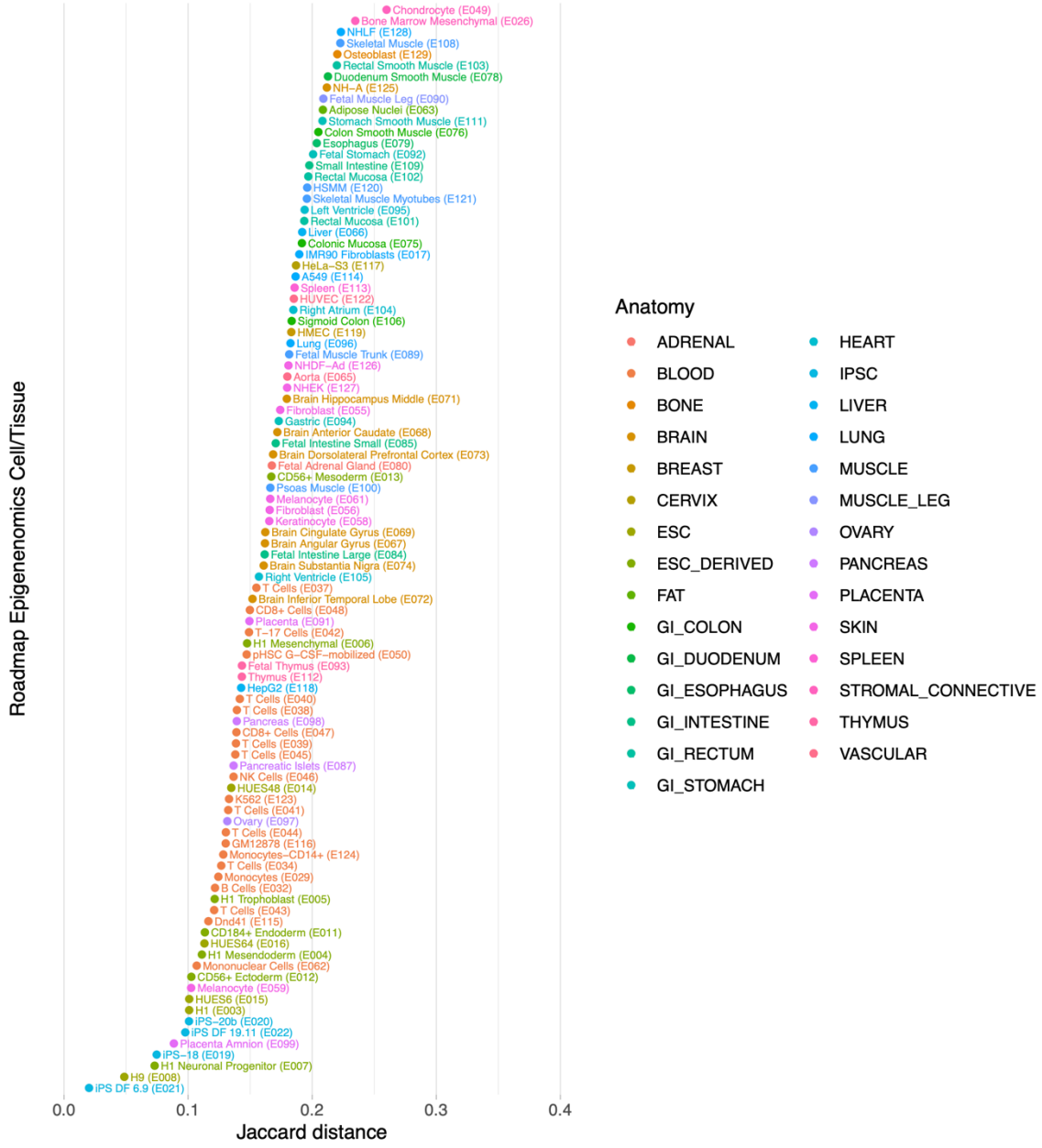


Figure 7.S1. Jaccard distance (similarity) between primary human chondrocytes and each cell type from the Roadmap Epigenomics Project. The cells with the highest similarity to primary human chondrocytes are "E049 - Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells". The Jaccard distance was calculated as intersection over union between each set of H3K27ac ChIP-seq peaks.

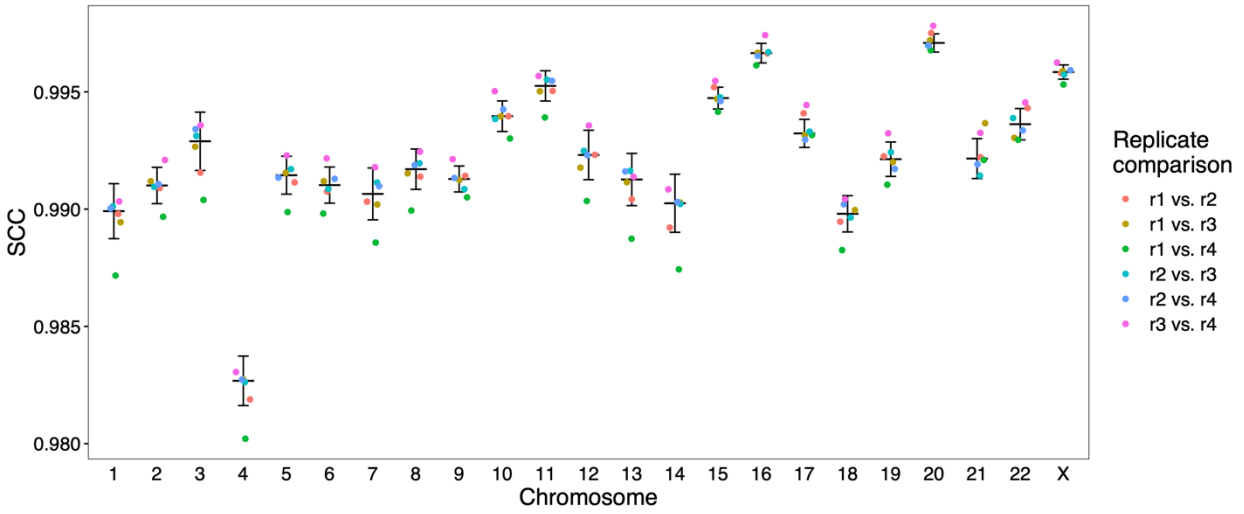


Figure 7.S2. Quantifying similarity of Hi-C replicates. Stratum adjusted correlation coefficient (SCC) comparing similarity between C-28/I2 Hi-C replicates for each chromosome (T. Yang et al. 2017). The python implementation of HiCRep was used to calculate SCC values at 10-kb resolution with a smoothing parameter of --h 20 and maximum distance of --dBPMax 5000000 (D. Lin, Sanders, and Noble 2021).

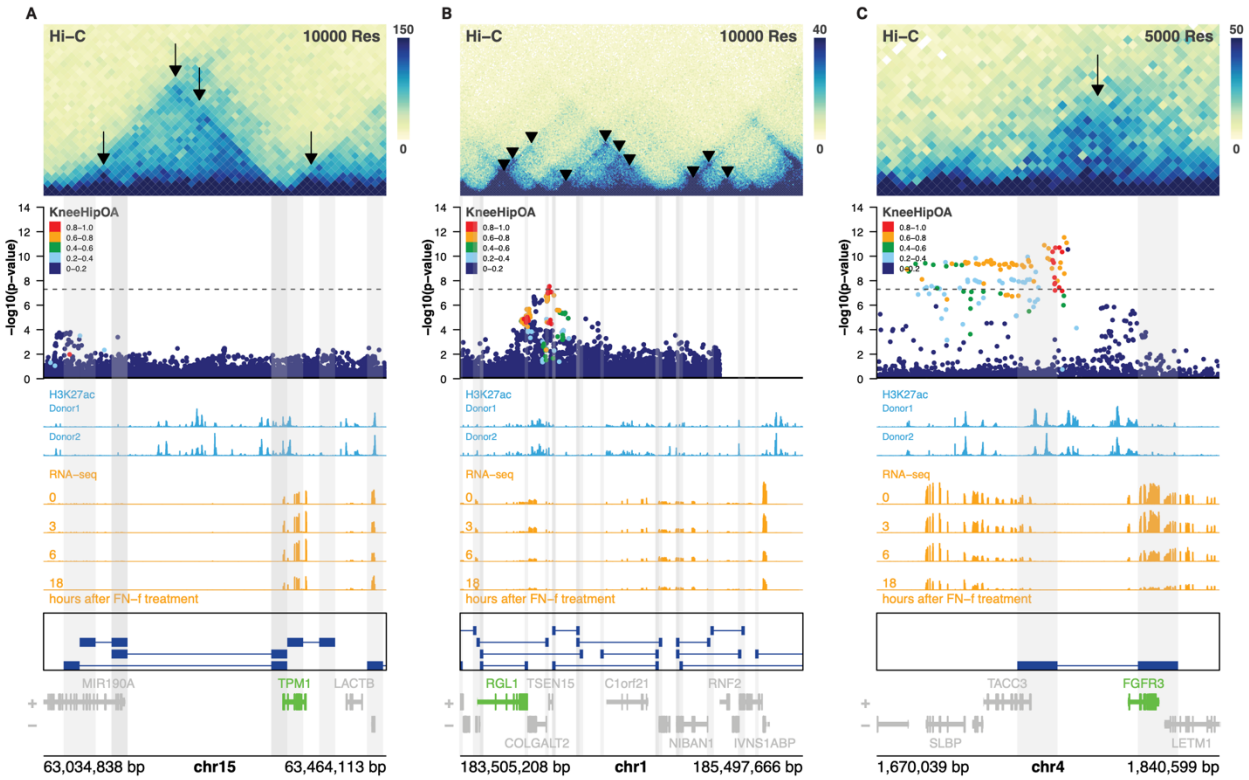


Figure 7.S3. Loci of looped variant-gene pairs identified as differentially expressed in response to FN-f. Hi-C in C-28/I2 cells (Hi-C heatmap) shows chromatin loop anchors (gray vertical highlights and blue bars) connecting OA GWAS variants (Manhattan plot) to promoters (bottom gene track) of differentially expressed genes in response to a timecourse of FN-f treatment (orange signal tracks). Blue signal tracks show CUT&RUN data for H3K27ac from two donors of primary human chondrocytes. Loci of variant-gene pairs shown are **(A)** rs746239049-TPM1, **(B)** rs10797938-RGL1, and **(C)** rs4535386/rs2896518-FGFR3.

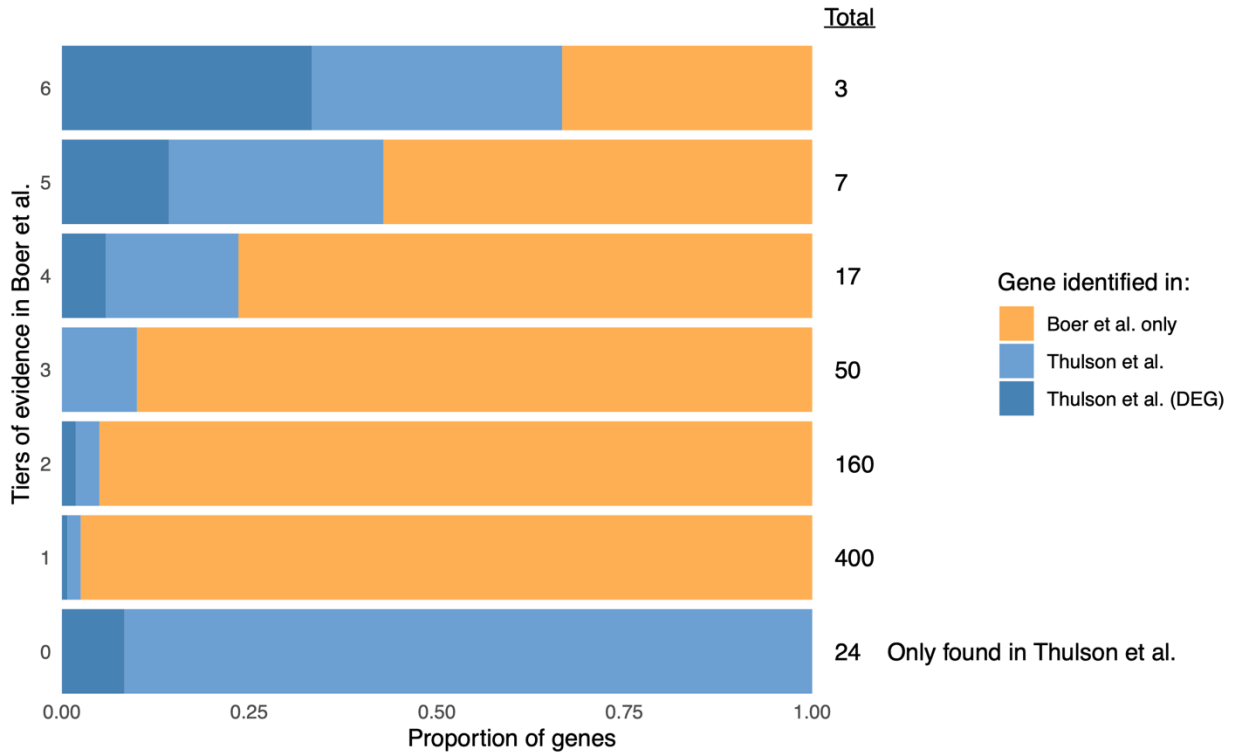


Figure 7.S4. Effector gene comparison between Boer et al. and Thulson et al. Proportion of genes identified in Thulson et al. by evidence score in Boer et al. Differential gene in Thulson et al. indicates the proportion of genes that changed significantly in response to FN-f from Reed et al. (see methods). Evidence level 0 indicate novel genes that were found in Thulson et al. but not in Boer et al.

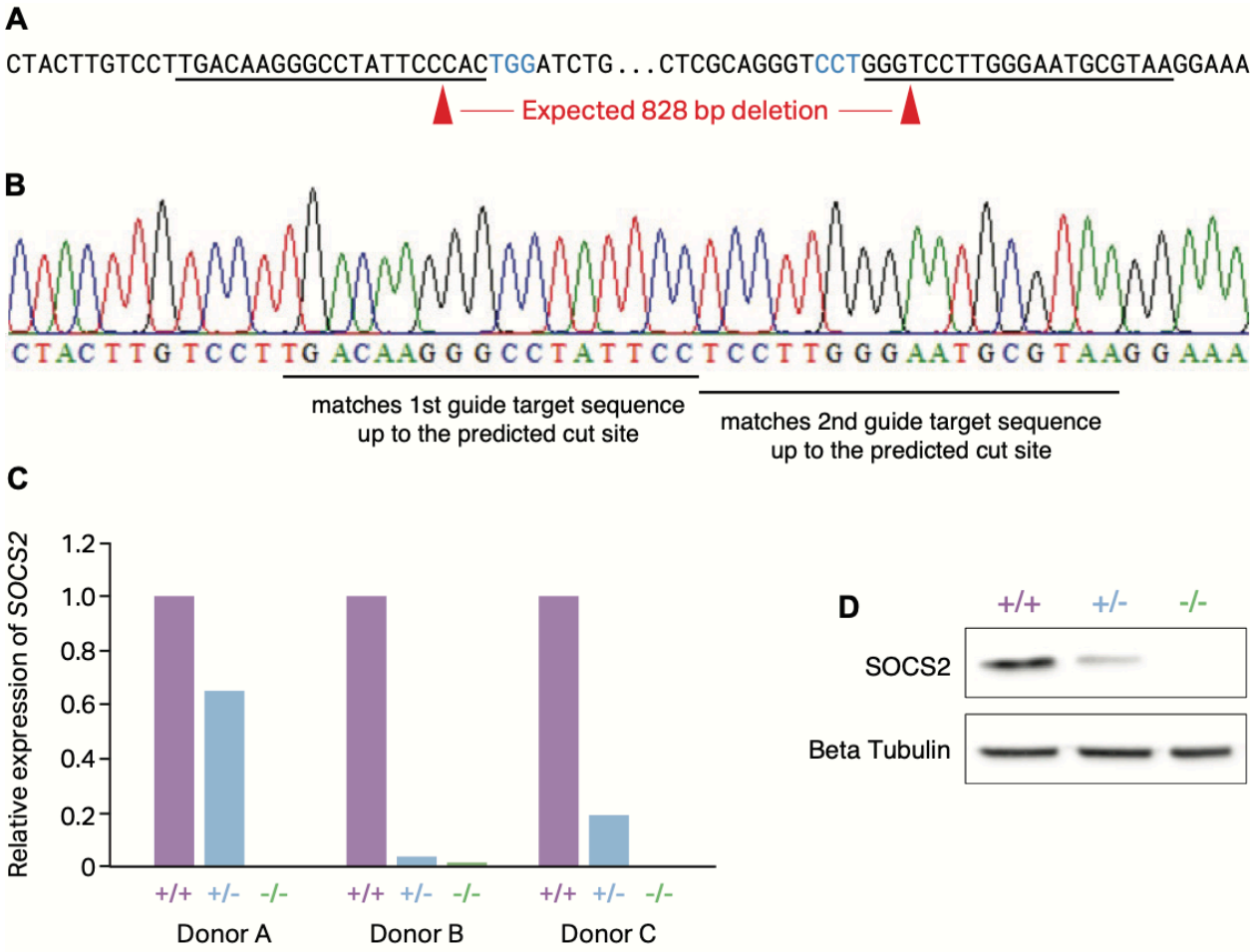


Figure 7.S5. Validation of SOCS2 knockout. (A) Region of SOCS2 targeted by editing, PAM sites in blue, guide RNA target sites underlined, expected cut sites depicted by red triangles. **(B)** Sanger sequencing of knockout colony confirmed deletion. **(C)** SOCS2 expression was analyzed in pooled colonies by qPCR, RNA from colonies of donors A and C were treated with DNase before qPCR to eliminate any background signal of SOCS2 from remaining genomic DNA. Expression was normalized to housekeeping gene TBP and relative expression calculated. **(D)** Western blot analysis confirmed decreased expression in heterozygous colony and loss of SOCS2 in knockout colony.

Chapter 8: Conclusions & Future Directions

8.1. The interplay of CTCF and LLPS looping

Biomolecular condensates play a key role in shaping 3D chromatin structure. The work presented in chapter 5 shows some of the first direct evidence of LLPS-driven aberrant looping during cancer development. Here, we show that the NHA9 fusion protein containing an IDR and DBD can bind to DNA and bring it into condensates with other distant binding sites. While most of the gained loops are bound by NHA9 at their anchors, CTCF-bound loops are also present and sometimes lost in favor of gained NHA9 loops. This raises interesting questions about the extent that LLPS- and extrusion-based looping interact to shape the chromatin landscape. Advancements in LLPS-based looping direction technology will enable further investigation on this phenomenon. Light-activated-dynamic-looping (LADL) uses an enzymatically dead Cas9 (dCas9) fused to a light sensitive CIBN protein to direct looping (J. H. Kim et al. 2019). For LLPS, an IDR fused to enzymatically dead Cas9 (dCas9) could allow directed, inducible looping. CasDrop uses this approach combined with optogenetics to inducibly seed condensates (Shin 2018). CasDrop or a similar technology could be used to direct LLPS looping at different positions in and around CTCF-loops. The sequencing depth required to call loops in Hi-C data for all these conditions would likely be cost-prohibitive. However, region-capture Hi-C techniques, such as Hi-C^2 or region-capture Micro-C would allow testing of multiple conditions at a fraction of the sequencing depth (Sanborn et al. 2015; Goel, Huseyin, and Hansen 2022). These techniques would be superior to other 3C technologies because it would be essential to know how the entire capture region changes.

Future directions could also be aimed at characterizing the mechanisms and which types of fusions can alter chromatin structure. For example, condensates containing BRD4-NUT fusions bind acetylated histones and recruit histone acetyltransferases (HATs) in a feed-forward loop that produces large chromatin subcompartments (Quiroga et al. 2022; Rosencrance et al. 2020; Alekseyenko et al. 2015). NUP98-PHD fusions which also form condensates that bind and methylate histones may also alter chromatin structure (Quiroga et al. 2022). Many cancers result from fusions proteins and screening different combinations of

IDR-DBDs for their ability to form condensates and form chromatin loops can provide further insight into disease mechanisms.

Understanding LLPS will not only help understand disease progression, but also potentially how to resolve it. For example, some cancer therapeutics such as cisplatin has been shown to concentrate into mediator 1 (MED1) condensates in vitro (I. A. Klein et al. 2020). This suggests that manipulating condensates might impact effective drug concentrations and pharmacodynamics, properties which could be exploited to improve drug design.

8.2. Improving temporal and cellular resolution

Chapters 6 and 7 use bulk Hi-C to link regulatory regions to their target genes. One limitation of bulk Hi-C is that it is unable to distinguish strong contacts in a small fraction of cells from weak contacts in a larger fraction of cells. Single-cell Hi-C (scHi-C) technology is still developing and could be used to distinguish these differences by clustering cells by their contact patterns (Nagano et al. 2013). Current approaches to scHi-C include coupling traditional Hi-C steps with fluorescence-activated cell/nucleus sorting (FACS/FANS) to isolate single cells or barcoding followed by computational demultiplexing (Galitsyna and Gelfand 2021). scHi-C is a relatively new in chromatin biology and the current sparse contact maps it produces makes it challenging to analyze and draw meaningful biological conclusions (J. Zhou et al. 2019). Fortunately, the systems used in these chapters were composed of cell lines with nearly identical cell populations hopefully limiting the effects of cell heterogeneity on Hi-C contacts.

Currently, a major challenge in the field of gene regulation is correctly assigning enhancer-promoter/gene (EP) pairs. Here and in other work, we have used Hi-C (via chromatin loops or increased contact frequency) to assign these connections with encouraging results (Kelly et al. 2022). While this is an improvement over assigning enhancers to their nearest genes, contact alone may not be sufficient to indicate EP pairs. A leading theory in gene regulation is that enhancer activity (measured by H3K27ac strength and accessibility through ATAC- or DNase-seq) should be incorporated with chromatin contact data. The activity by contact (ABC) model uses this approach and has performed better than distance or Hi-C alone (Fulco et al. 2019). However, ABC cannot capture dynamic EP pairs that form in response to stimuli. Our macrophage activation time course dataset is unique in that it measures enhancer activity and chromatin contact for multiple timepoints during macrophage activation – a prime dataset for applying and

expanding the ABC model to include dynamics (i.e., ABCD). Such a model would identify EP pairs with ABC scores that correlate with expression, albeit with a lag consistent with our findings. One of the limitations in fully realizing this model is our irregularly spaced timepoints. Imputation methods and incorporation of uncertainty between gaps in measurement is essential for developing an ABCD model that captures true EP pairs. Imputation methods are far more feasible than measuring more timepoints with multiple genomic assays. A major challenge for ABCD or any other EP prediction modeling is developing an appropriate validation set. QTLs in simulated macrophages (conditions very similar to our own) exist but are not very specific, with far too many false-positives to reliably validate findings (Alasoo et al. 2018). A massively parallel reporter assay (MPRA) could provide a more targeted validation, though some interpretability issues may arise as some chromatin context is lost (J. C. Klein et al. 2020). Ultimately, models for the spatiotemporal regulation of genes will benefit from validation sets.

8.3. Software interoperability for improved genomic workflows

The future of genomic software development is trending towards software that works well together. As increasing amounts of genomic data are created to understand the complexities of gene regulation, there has been a concordant shift in demand for computational analysis. Ad hoc software development was sufficient early on in genomics as it was a niche pioneered by a small set of individuals interested in applying computational tools to biology. However, this increased demand brings in biologists of all backgrounds and demands tools that are built accordingly. Ecosystems, like that of Bioconductor and Biopython for R and python respectively, ensure software operates in standardized and predictable ways and increases the accessibility of genomic data analysis (Gentleman et al. 2004; Cock et al. 2009). For example, the widely-used R/Bioconductor *DESeq2* package for analyzing differential bulk RNA-seq data has recently been ported to python increasing the reproducibility of differential gene expression analysis across platforms (Muzellec et al. 2022). Our group has brought the widely used *bedtools* package into R (Quinlan and Hall 2010; Patwardhan et al. 2019). Similarly, there has been a push to make genomic data more intuitive through natural language syntax (grammar of genomics) and data more standardized (so called “tidy” genomics) (S. Lee, Cook, and Lawrence 2019; L’Yi et al. 2022; Mangiola et al. 2021). For example, the plyranges package implements a grammar of genomics approach for manipulation of genomic data with simple verbiage for more readable workflows (S. Lee, Cook, and Lawrence 2019). All these features make

genomic data analysis more widely accessible, particularly for single range data types. Future work could extend additional grammar of genomics support for paired-end data.

For Hi-C analysis, the *HiC-DC+* package has wrapped many of the *Juicer* tools that are typically executed on the command line into R (Sahin et al. 2021; Durand, Shamim, et al. 2016). This allows users to identify many chromatin features without leaving the R platform. In chapter 4, we introduce *mariner* which further improves this infrastructure by providing modular interfaces for creating custom analysis of Hi-C as well as operating on data that is too large to store in memory. There are several areas where additional features could improve Hi-C data analysis. For example, a common way of visualizing individual loops and loop aggregates is with a bullseye transformation that reduces visual artifacts introduced when Hi-C data is represented as a heatmap (Rowley et al. 2020). This could be incorporated as a "type" argument to *mariner's* "plotMatrix" function. Another gap in Hi-C analysis is the lack of classes to efficiently contain irregular or jagged arrays of Hi-C matrices. The HDF5-based approach used in *mariner* could be extended to accommodate these data structures. This would allow users to store non-uniformly sized genomic features, such as TADs, in a single object.

The size and scale of genomic data can make it difficult to navigate. Software tools for visualizing and exploring multi-omic data are maturing. For example, *plotgardener* is an R/Bioconductor package that allows users to create publication-ready, multi-panel figures directly in R (Kramer et al. 2022). *mariner* is compatible with *plotgardener*, further extending this ecosystem. *Plotgardener* makes it easy to create genomic views with multi-omic data and programmatically explore them. Because some programming is required, plotgardener can be difficult for newer users to adopt. An interface for assembling these views would be bridge the gap between programmable visualizations and usability. Gosling is a tool for building dynamic, interactive views. Gosling is a javascript library that allows grammar-like syntax for assembling web-based genomic visualizations (L'Yi et al. 2022). Users can dynamically explore different genomic regions with an intuitive graphical user interface and allow customization. One draw-back of this dynamic approach is that the genome is so large it can be like looking for a needle in a haystack. While these tools create interactive data visualizations, they lack the programmatic surveying available in plotgardener. Future work in interactive data visualization should focus on combining these two approaches to create dynamic, programmable figures for genomics.

REFERENCES

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature* 467 (7319): 1061–73.

Abdennur, Nezar, and Leonid A. Mirny. 2020. “Cooler: Scalable Storage for Hi-C Data and Other Genomically Labeled Arrays.” *Bioinformatics* 36 (1): 311–16.

Abramo, Kristin, Anne-Laure Valton, Sergey V. Venev, Hakan Ozadam, A. Nicole Fox, and Job Dekker. 2019. “A Chromosome Folding Intermediate at the Condensin-to-Cohesin Transition during Telophase.” *Nature Cell Biology* 21 (11): 1393–1402.

Ahn, Jeong Hyun, Eric S. Davis, Timothy A. Daugird, Shuai Zhao, Ivana Yoseli Quiroga, Hidetaka Uryu, Jie Li, et al. 2021. “Phase Separation Drives Aberrant Chromatin Looping and Cancer Development.” *Nature* 595 (7868): 591–95.

Akdemir, Kadir C., Victoria T. Le, Sahaana Chandran, Yilong Li, Roel G. Verhaak, Rameen Beroukhim, Peter J. Campbell, et al. 2020. “Disruption of Chromatin Folding Domains by Somatic Genomic Rearrangements in Human Cancer.” *Nature Genetics* 52 (3): 294–305.

Alasoo, Kaur, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J. Knights, Alice L. Mann, Kousik Kundu, HIPSCI Consortium, Christine Hale, Gordon Dougan, and Daniel J. Gaffney. 2018. “Shared Genetic Effects on Chromatin and Gene Expression Indicate a Role for Enhancer Priming in Immune Response.” *Nature Genetics* 50 (3): 424–31.

Alberti, Simon, and Anthony A. Hyman. 2021. “Biomolecular Condensates at the Nexus of Cellular Stress, Protein Aggregation Disease and Ageing.” *Nature Reviews. Molecular Cell Biology* 22 (3): 196–213.

Alekseyenko, Artyom A., Erica M. Walsh, Xin Wang, Adlai R. Grayson, Peter T. Hsi, Peter V. Kharchenko, Mitzi I. Kuroda, and Christopher A. French. 2015. “The Oncogenic BRD4-NUT Chromatin Regulator Drives Aberrant Transcription within Large Topological Domains.” *Genes & Development* 29 (14): 1507–23.

Ali, M. Sanni, Rolf H. H. Groenwold, Svetlana V. Belitser, Wiebe R. Pestman, Arno W. Hoes, Kit C. B. Roes, Anthonius de Boer, and Olaf H. Klungel. 2015. “Reporting of Covariate Selection and Balance Assessment in Propensity Score Analysis Is Suboptimal: A Systematic Review.” *Journal of Clinical Epidemiology* 68 (2): 112–21.

Anders, S., and W. Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11. https://doi.org/10.1186/gb-2010-11-10-r106.

Andrés, María C. de, Kei Imagawa, Ko Hashimoto, Antonio Gonzalez, Mary B. Goldring, Helmtrud I. Roach, and Richard O. C. Oreffo. 2011. “Suppressors of Cytokine Signalling (SOCS) Are Reduced in Osteoarthritis.” *Biochemical and Biophysical Research Communications* 407 (1): 54–59.

Arner, Erik, Carsten O. Daub, Kristoffer Vitting-Seerup, Robin Andersson, Berit Lilje, Finn Drabløs, Andreas Lennartsson, et al. 2015. “Transcribed Enhancers Lead Waves of Coordinated Transcription in Transitioning Mammalian Cells.” *Science* 347 (6225): 1010–14.

Arora, S., M. Morgan, M. Carlson, and H. Pagès. n.d. “GenomeInfoDb: Utilities for Manipulating Chromosome Names, Including Modifying Them to Follow a Particular Naming Style.” *R Package Version*.

“Babraham Bioinformatics.” 2010. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

“———.” 2015. 2015. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Banani, Salman F., Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen. 2017. “Biomolecular Condensates: Organizers of Cellular Biochemistry.” *Nature Reviews. Molecular Cell Biology* 18 (5): 285–98.

Banigan, Edward J., Wen Tang, Aafke A. van den Berg, Roman R. Stocsits, Gordana Wutz, Hugo B. Brandão, Georg A. Busslinger, Jan-Michael Peters, and Leonid A. Mirny. 2022. “Transcription Shapes 3D Chromatin Organization by Interacting with Loop-Extruding Cohesin Complexes.” *BioRxiv*. https://doi.org/10.1101/2022.01.07.475367.

Beagan, Jonathan A., Elissa D. Pastuzyn, Lindsey R. Fernandez, Michael H. Guo, Kelly Feng, Katelyn R. Titus, Harshini Chandrashekar, Jason D. Shepherd, and Jennifer E. Phillips-Cremins. 2020. “Three-Dimensional Genome Restructuring across Timescales of Activity-Induced Neuronal Gene Expression.” *Nature Neuroscience* 23 (6): 707–17.

Bernstein, Bradley E., John A. Stamatoyannopoulos, Joseph F. Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, et al. 2010. “The NIH Roadmap Epigenomics Mapping Consortium.” *Nature Biotechnology* 28 (10): 1045–48.

Bertero, Alessandro, Paul A. Fields, Vijay Ramani, Giancarlo Bonora, Galip G. Yardimci, Hans Reinecke, Lil Pabon, William S. Noble, Jay Shendure, and Charles E. Murry. 2019. “Dynamics of Genome Reorganization during Human Cardiogenesis Reveal an RBM20-Dependent Splicing Factory.” *Nature Communications* 10 (1): 1538.

Bickel, Peter J., Nathan Boley, James B. Brown, Haiyan Huang, and Nancy R. Zhang. 2010. “Subsampling Methods for Genomic Inference.” *The Annals of Applied Statistics* 4 (4): 1660–97.

Boer, Cindy G., Konstantinos Hatzikotoulas, Lorraine Southam, Lilja Stefánsdóttir, Yanfei Zhang, Rodrigo Coutinho de Almeida, Tian T. Wu, et al. 2021. “Deciphering Osteoarthritis Genetics across 826,690 Individuals from 9 Populations.” *Cell* 184 (24): 6003–5.

Boeuf, S., E. Steck, K. Pelttari, T. Hennig, A. Buneb, K. Benz, D. Witte, H. Sültmann, A. Poustka, and W. Richter. 2008. “Subtractive Gene Expression Profiling of Articular Cartilage and Mesenchymal Stem Cells: Serpins as Cartilage-Relevant Differentiation Markers.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 16 (1): 48–60.

Boija, Ann, Isaac A. Klein, Benjamin R. Sabari, Alessandra Dall’Agnese, Eliot L. Coffey, Alicia V. Zamudio, Charles H. Li, et al. 2018. “Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains.” *Cell* 175 (7): 1842-1855.e16.

Boija, Ann, Isaac A. Klein, and Richard A. Young. 2021. “Biomolecular Condensates and Cancer.” *Cancer Cell* 39 (2): 174–92.

Bond, Marielle L., Eric S. Davis, Ivana Y. Quiroga, Michael I. Love, Hyejung Won, and Douglas H. Phanstiel. 2022. “Chromatin Loop Dynamics during Cellular Differentiation Are Associated with Changes to Both Anchor and Internal Regulatory Features.” *BioRxiv*. https://doi.org/10.1101/2022.10.31.514600.

Bonev, Boyan, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, et al. 2017. “Multiscale 3D Genome Rewiring during Mouse Neural Development.” *Cell* 171 (3): 557-572.e24.

Bosch, M. H. J. van den, P. L. E. M. van Lent, and P. M. van der Kraan. 2020. “Identifying Effector Molecules, Cells, and Cytokines of Innate Immunity in OA.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 28 (5): 532–43.

Brandão, Hugo B., Michele Gabriele, and Anders S. Hansen. 2021. “Tracking and Interpreting Long-Range Chromatin Interactions with Super-Resolution Live-Cell Imaging.” *Current Opinion in Cell Biology* 70: 18–26.

Brandão, Hugo B., Payel Paul, Aafke A. van den Berg, David Z. Rudner, Xindan Wang, and Leonid A. Mirny. 2019. “RNA Polymerases as Moving Barriers to Condensin Loop Extrusion.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (41): 20489–99.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18.

Busslinger, Georg A., Roman R. Stocsits, Petra van der Lelij, Elin Axelsson, Antonio Tedeschi, Niels Galjart, and Jan-Michael Peters. 2017. “Cohesin Is Positioned in Mammalian Genomes by Transcription, CTCF and Wapl.” *Nature* 544 (7651): 503–7.

Cai, Danfeng, Daniel Feliciano, Peng Dong, Eduardo Flores, Martin Gruebele, Natalie Porat-Shliom, Shahar Sukenik, Zhe Liu, and Jennifer Lippincott-Schwartz. 2019. “Phase Separation of YAP Reorganizes Genome Topology for Long-Term YAP Target Gene Expression.” *Nature Cell Biology* 21 (12): 1578–89.

Cai, L. 2013. “An H3K36 Methylation-Engaging Tudor Motif of Polycomb-like Proteins Mediates PRC2 Complex Targeting.” *Molecular Cell* 49. https://doi.org/10.1016/j.molcel.2012.11.026.

———. 2018. “ZFX Mediates Non-Canonical Oncogenic Functions of the Androgen Receptor Splice Variant 7 in Castrate-Resistant Prostate Cancer.” *Molecular Cell* 72. https://doi.org/10.1016/j.molcel.2018.08.029.

Calvo, K. R., D. B. Sykes, M. Pasillas, and M. P. Kamps. 2000. “Hoxa9 Immortalizes a Granulocyte-Macrophage Colony-Stimulating Factor-Dependent Promyelocyte Capable of Biphenotypic Differentiation to Neutrophils or Macrophages, Independent of Enforced Meis Expression.” *Molecular and Cellular Biology* 20. https://doi.org/10.1128/MCB.20.9.3274-3285.2000.

Calvo, K. R., D. B. Sykes, M. P. Pasillas, and M. P. Kamps. 2002. “Nup98-HoxA9 Immortalizes Myeloid Progenitors, Enforces Expression of Hoxa9, Hoxa7 and Meis1, and Alters Cytokine-Specific Responses in a Manner Similar to That Induced by Retroviral Co-Expression of Hoxa9 and Meis1.” *Oncogene* 21. https://doi.org/10.1038/sj.onc.1205516.

Capellini, Terence D., Hao Chen, Jiaxue Cao, Andrew C. Doxey, Ata M. Kiapour, Michael Schoor, and David M. Kingsley. 2017. “Ancient Selection for Derived Alleles at a GDF5 Enhancer Influencing Human Growth and Osteoarthritis Risk.” *Nature Genetics* 49 (8): 1202–10.

Caron, Marjolein M. J., Pieter J. Emans, Don A. M. Surtel, Peter M. van der Kraan, Lodewijk W. van Rhijn, and Tim J. M. Welting. 2015. “BAPX-1/NKX-3.2 Acts as a Chondrocyte Hypertrophy Molecular Switch in Osteoarthritis.” *Arthritis & Rheumatology*. https://doi.org/10.1002/art.39293.

Chang, Jia-Ming, Yi-Fu Weng, Wei-Ting Chang, Fu-An Lin, and Giacomo Cavalli. 2022. “HiCmapTools: A Tool to Access HiC Contact Maps.” *BMC Bioinformatics* 23 (1): 64.

Chen, B. -. C. 2014. “Lattice Light-Sheet Microscopy: Imaging Molecules to Embryos at High Spatiotemporal Resolution.” *Science* 346. https://doi.org/10.1126/science.1257998.

Cherniavsky Durand, Neva, and Muhammad Saad Shamim. 2022. “Strawr: Fast Implementation of Reading/Dump for .Hic Files.” https://github.com/aidenlab/straw/tree/master/R.

Chesi, Alessandra, Yadav Wagley, Matthew E. Johnson, Elisabetta Manduchi, Chun Su, Sumei Lu, Michelle E. Leonard, et al. 2019. “Genome-Scale Capture C Promoter Interactions Implicate Effector Genes at GWAS Loci for Bone Mineral Density.” *Nature Communications* 10 (1): 1260.

Chong, S. 2018. “Imaging Dynamic and Selective Low-Complexity Domain Interactions That Control Gene Transcription.” *Science* 361. https://doi.org/10.1126/science.aar2555.

Claussnitzer, Melina, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, et al. 2015. “FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.” *The New England Journal of Medicine* 373 (10): 895–907.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25 (11): 1422–23.

Cock, Peter J. A., Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. 2010. “The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants.” *Nucleic Acids Research* 38 (6): 1767–71.

Cuartero, Sergi, Felix D. Weiss, Gopuraja Dharmalingam, Ya Guo, Elizabeth Ing-Simmons, Silvia Masella, Irene Robles-Rebollo, et al. 2018. “Control of Inducible Gene Expression Links Cohesin to Hematopoietic Progenitor Self-Renewal and Differentiation.” *Nature Immunology* 19 (9): 932–41.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Davidson, Iain F., Daniela Goetz, Maciej P. Zaczek, Maxim I. Molodtsov, Pim J. Huis In ’t Veld, Florian Weissmann, Gabriele Litos, et al. 2016. “Rapid Movement and Transcriptional Re-Localization of Human Cohesin on DNA.” *The EMBO Journal* 35 (24): 2671–85.

Davidson, Iain F., and Jan-Michael Peters. 2021. “Genome Folding through Loop Extrusion by SMC Complexes.” *Nature Reviews. Molecular Cell Biology* 22 (7): 445–64.

Davis, Eric S. 2023. *DietJuicer: DietJuicer Is a Lighter-Weight, HPC Flexible Version of Juicer Written with Snakemake*. Github. https://github.com/EricSDavis/dietJuicer.

Davis, Eric S., Wancen Mu, Stuart Lee, Mikhail G. Dozmorov, Michael I. Love, and Douglas H. Phanstiel. 2022. “MatchRanges: Generating Null Hypothesis Genomic Ranges via Covariate-Matched Sampling.” *BioRxiv*. https://doi.org/10.1101/2022.08.05.502985.

D’Costa, Susan, Matthew J. Rich, and Brian O. Diekman. 2020. “Engineered Cartilage from Human Chondrocytes with Homozygous Knockout of Cell Cycle Inhibitor P21.” *Tissue Engineering. Part A* 26 (7–8): 441–49.

Dekker, Job, and Leonid Mirny. 2016. “The 3D Genome as Moderator of Chromosomal Communication.” *Cell* 164 (6): 1110–21.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. “Capturing Chromosome Conformation.” *Science* 295 (5558): 1306–11.

D’Ippolito, Anthony M., Ian C. McDowell, Alejandro Barrera, Linda K. Hong, Sarah M. Leichter, Luke C. Bartelt, Christopher M. Vockley, et al. 2018. “Pre-Established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids.” *Cell Syst* 7 (2): 146-160.e7.

Dobin, A. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29. https://doi.org/10.1093/bioinformatics/bts635.

Dostie, Josée, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, et al. 2006. “Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements.” *Genome Research* 16 (10): 1299–1309.

Dowle, Matt, and Arun Srinivasan. 2021. “Data.Table: Extension of `data.Frame`.” https://CRAN.R-project.org/package=data.table.

Dozmorov, Mikhail G., Eric Davis, Wancen Mu, Stuart Lee, Tim Triche, Douglas Phanstiel, and Michael Love. 2022. “CTCF.” https://github.com/mdozmorov/CTCF.

Duan, Aiping, Hong Wang, Yan Zhu, Qi Wang, Jing Zhang, Qing Hou, Yuexian Xing, et al. 2021. “Chromatin Architecture Reveals Cell Type-Specific Target Genes for Kidney Disease Risk Variants.” *BMC Biology* 19 (1): 38.

Durand, Neva C., James T. Robinson, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. 2016. “Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.” *Cell Systems* 3 (1): 99–101.

Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. “Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.” *Cell Systems* 3 (1): 95–98.

Edwards, Stacey L., Jonathan Beesley, Juliet D. French, and Alison M. Dunning. 2013. “Beyond GWASs: Illuminating the Dark Road from Association to Function.” *American Journal of Human Genetics* 93 (5): 779–97.

Egan, B. 2016. “An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon Ezh2 Inhibition.” *PloS One* 11. https://doi.org/10.1371/journal.pone.0166438.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32 (19): 3047–48.

Fahrenkrog, B. 2016. “Expression of Leukemia-Associated Nup98 Fusion Proteins Generates an Aberrant Nuclear Envelope Phenotype.” *PloS One* 11. https://doi.org/10.1371/journal.pone.0152321.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2015. “Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants.” *Nature* 518 (7539): 337–43.

Fisch, K. M., R. Gamini, O. Alvarez-Garcia, R. Akagi, M. Saito, Y. Muramatsu, T. Sasho, J. A. Koziol, A. I. Su, and M. K. Lotz. 2018. “Identification of Transcription Factors Responsible for Dysregulated Networks in Human Osteoarthritis Cartilage by Global Gene Expression Analysis.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 26 (11): 1531–38.

Flyamer, Ilya M., Robert S. Illingworth, and Wendy A. Bickmore. 2020. “Coolpup.Py: Versatile Pile-up Analysis of Hi-C Data.” *Bioinformatics* 36 (10): 2980–85.

Forsyth, Christopher B., Judit Pulai, and Richard F. Loeser. 2002. “Fibronectin Fragments and Blocking Antibodies to Alpha2beta1 and Alpha5beta1 Integrins Stimulate Mitogen-Activated Protein Kinase Signaling and Increase Collagenase 3 (Matrix Metalloproteinase 13) Production by Human Articular Chondrocytes.” *Arthritis and Rheumatism* 46 (9): 2368–76.

Francis-West, P. H., J. Parish, K. Lee, and C. W. Archer. 1999. “BMP/GDF-Signalling Interactions during Synovial Joint Development.” *Cell and Tissue Research* 296 (1): 111–19.

Frey, S., R. P. Richter, and D. Görlich. 2006. “FG-Rich Repeats of Nuclear Pore Proteins Form a Three-Dimensional Meshwork with Hydrogel-like Properties.” *Science* 314. https://doi.org/10.1126/science.1132516.

Fukaya, Takashi, Bomyi Lim, and Michael Levine. 2016. “Enhancer Control of Transcriptional Bursting.” *Cell* 166 (2): 358–68.

Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. “Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of CRISPR Perturbations.” *Nature Genetics* 51 (12): 1664–69.

Fullwood, Melissa J., Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, et al. 2009. “An Oestrogen-Receptor-Alpha-Bound Human Chromatin Interactome.” *Nature* 462 (7269): 58–64.

Furlan-Magaril, Mayra, Masami Ando-Kuri, Rodrigo G. Arzate-Mejía, Jörg Morf, Jonathan Cairns, Abraham Román-Figueroa, Luis Tenorio-Hernández, et al. 2021. “The Global and Promoter-Centric 3D Genome Organization Temporally Resolved during a Circadian Cycle.” *Genome Biology* 22 (1): 162.

Gabriele, Michele, Hugo B. Brandão, Simon Grosse-Holz, Asmita Jha, Gina M. Dailey, Claudia Cattoglio, Tsung-Han S. Hsieh, Leonid Mirny, Christoph Zechner, and Anders S. Hansen. 2021. “Dynamics of CTCF and Cohesin Mediated Chromatin Looping Revealed by Live-Cell Imaging.” *BioRxiv*. https://doi.org/10.1101/2021.12.12.472242.

Gagat, Maciej, Wioletta Zielińska, Klaudia Mikołajczyk, Jan Zabrzyński, Adrian Krajewski, Anna Klimaszewska-Wiśniewska, Dariusz Grzanka, and Alina Grzanka. 2021. “CRISPR-Based Activation of Endogenous Expression of TPM1 Inhibits Inflammatory Response of Primary Human Coronary Artery Endothelial and Smooth Muscle Cells Induced by Recombinant Human Tumor Necrosis Factor α.” *Frontiers in Cell and Developmental Biology*. https://doi.org/10.3389/fcell.2021.668032.

Galitsyna, Aleksandra A., and Mikhail S. Gelfand. 2021. “Single-Cell Hi-C Data Analysis: Safety in Numbers.” *Briefings in Bioinformatics* 22 (6). https://doi.org/10.1093/bib/bbab316.

Gao, Xiang, Yu Sun, and Xu Li. 2019. “Identification of Key Gene Modules and Transcription Factors for Human Osteoarthritis by Weighted Gene Co-Expression Network Analysis.” *Experimental and Therapeutic Medicine* 18 (4): 2479–90.

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. 2004. “Bioconductor: Open Software Development for Computational Biology and Bioinformatics.” *Genome Biology* 5 (10): R80.

Giambartolomei, Claudia, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. 2014. “Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics.” *PLoS Genetics* 10 (5): e1004383.

Gibson, B. A. 2019. “Organization of Chromatin by Intrinsic and Regulated Phase Separation.” *Cell* 179. https://doi.org/10.1016/j.cell.2019.08.037.

Goel, Viraat Y., Miles K. Huseyin, and Anders S. Hansen. 2022. “Region Capture Micro-C Reveals Coalescence of Enhancers and Promoters into Nested Microcompartments.” *BioRxiv*. https://doi.org/10.1101/2022.07.12.499637.

Goloborodko, Anton, Nezar Abdennur, Sergey Venev, hbbrandao, and gfudenberg. 2018. *Mirnylab/Pairtools: V0.2.0*. https://doi.org/10.5281/zenodo.1490831.

Gough, Sheryl M., Fan Lee, Fan Yang, Robert L. Walker, Yeulin J. Zhu, Marbin Pineda, Masahiro Onozawa, et al. 2014. “NUP98-PHF23 Is a Chromatin-Modifying Oncoprotein That Causes a Wide Array of Leukemias Sensitive to Inhibition of PHD Histone Reader Function.” *Cancer Discovery* 4 (5): 564–77.

Gough, Sheryl M., Christopher I. Slape, and Peter D. Aplan. 2011. “NUP98 Gene Fusions and Hematopoietic Malignancies: Common Themes and New Biologic Insights.” *Blood* 118 (24): 6247–57.

Greifer, Noah. 2020. “Covariate Balance Tables and Plots: A Guide to the Cobalt Package.” *Accessed March* 10: 2020.

Grimm, J. B. 2015. “A General Method to Improve Fluorophores for Live-Cell and Single-Molecule Microscopy.” *Nature Methods* 12. https://doi.org/10.1038/nmeth.3256.

Gu, Bo, Colin J. Comerci, Dannielle G. McCarthy, Saumya Saurabh, W. E. Moerner, and Joanna Wysocka. 2020. “Opposing Effects of Cohesin and Transcription on CTCF Organization Revealed by Super-Resolution Imaging.” *Molecular Cell* 80 (4): 699-711.e7.

Gu, Huiya, Hannah Harris, Moshe Olshansky, Yossi Eliaz, Akshay Krishna, Achyuth Kalluchi, Mozes Jacobs, et al. 2021. “Fine-Mapping of Nuclear Compartments Using Ultra-Deep Hi-C Shows That Active Promoter and Enhancer Elements Localize in the Active A Compartment Even When Adjacent Sequences Do Not.” *BioRxiv*. https://doi.org/10.1101/2021.10.03.462599.

Guo, Yang Eric, John C. Manteiga, Jonathan E. Henninger, Benjamin R. Sabari, Alessandra Dall’Agnese, Nancy M. Hannett, Jan-Hendrik Spille, et al. 2019. “Pol II Phosphorylation Regulates a Switch between Transcriptional and Splicing Condensates.” *Nature* 572 (7770): 543–48.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. 2007. “Quantifying Similarity between Motifs.” *Genome Biology* 8 (2): R24.

Hahsler, Michael, Matthew Piekenbrock, and Derek Doran. 2019. “Dbscan: Fast Density-Based Clustering with R.” *Journal of Statistical Software*. https://doi.org/10.18637/jss.v091.i01.

Hansen, A. S. 2018. “Robust Model-Based Analysis of Single-Particle Tracking Experiments with Spot-On.” *ELife* 7. https://doi.org/10.7554/eLife.33125.

Heidari, Nastaran, Douglas H. Phanstiel, Chao He, Fabian Grubert, Fereshteh Jahanbani, Maya Kasowski, Michael Q. Zhang, and Michael P. Snyder. 2014. “Genome-Wide Map of Regulatory Interactions in the Human Genome.” *Genome Research* 24 (12): 1905–17.

Heinz, Sven, Lorane Texari, Michael G. B. Hayes, Matthew Urbanowski, Max W. Chang, Ninvita Givarkes, Alexander Rialdi, et al. 2018. “Transcription Elongation Can Affect Genome 3D Structure.” *Cell* 174 (6): 1522-1536.e22.

“Hi-C Processing Pipeline.” n.d. Accessed December 15, 2022. https://data.4dnucleome.org/resources/data-analysis/hi_c-processing-pipeline.

Hic-Pipeline: HiC Uniform Processing Pipeline. n.d. Github. Accessed January 4, 2023. https://github.com/ENCODE-DCC/hic-pipeline.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 15 (3): 199–236.

———. 2011. “MatchIt: Nonparametric Preprocessing for Parametric Causal Inference.” *Journal of Statistical Software* 42 (8). https://doi.org/10.18637/jss.v042.i08.

Homandberg, G. A. 1999. “Potential Regulation of Cartilage Metabolism in Osteoarthritis by Fibronectin Fragments.” *Frontiers in Bioscience: A Journal and Virtual Library* 4 (October): D713-30.

Homandberg, G. A., C. Wen, and F. Hui. 1998. “Cartilage Damaging Activities of Fibronectin Fragments Derived from Cartilage and Synovial Fluid.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 6 (4): 231–44.

Hsieh, Tsung-Han S., Claudia Cattoglio, Elena Slobodyanyuk, Anders S. Hansen, Oliver J. Rando, Robert Tjian, and Xavier Darzacq. 2020. “Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding.” *Molecular Cell* 78 (3): 539-553.e8.

Hu, Ming, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S. Liu. 2012. “HiCNorm: Removing Biases in Hi-C Data via Poisson Regression.” *Bioinformatics* 28 (23): 3131–33.

Hunter, David J., and Sita Bierma-Zeinstra. 2019. “Osteoarthritis.” *The Lancet* 393 (10182): 1745–59.

Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. 2012. “Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization.” *Nature Methods* 9 (10): 999–1003.

Jain, Sachin, Vidhi Gautam, and Sania Naseem. 2011. “Acute-Phase Proteins: As Diagnostic Tool.” *Journal of Pharmacy & Bioallied Sciences* 3 (1): 118–27.

Jankovic, D. 2008. “Leukemogenic Mechanisms and Targets of a NUP98/HHEX Fusion in Acute Myeloid Leukemia.” *Blood* 111. https://doi.org/10.1182/blood-2007-09-108175.

Johnstone, Sarah E., Alejandro Reyes, Yifeng Qi, Carmen Adriaens, Esmat Hegazi, Karin Pelka, Jonathan H. Chen, et al. 2020. “Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer.” *Cell* 182 (6): 1474-1489.e23.

Karr, Jonathan P., John J. Ferrie, Robert Tjian, and Xavier Darzacq. 2022. “The Transcription Factor Activity Gradient (TAG) Model: Contemplating a Contact-Independent Mechanism for Enhancer-Promoter Communication.” *Genes & Development* 36 (1–2): 7–16.

Kasoji, Sandeep K., Samantha G. Pattenden, Ewa P. Malc, Chatura N. Jayakody, James K. Tsuruta, Piotr A. Mieczkowski, William P. Janzen, and Paul A. Dayton. 2015. “Cavitation Enhancing Nanodroplets Mediate Efficient DNA Fragmentation in a Bench Top Ultrasonic Water Bath.” *PloS One* 10 (7): e0133014.

Kasper, L. H. 1999. “CREB Binding Protein Interacts with Nucleoporin-Specific FG Repeats That Activate Transcription and Mediate NUP98-HOXA9 Oncogenicity.” *Molecular and Cellular Biology* 19. https://doi.org/10.1128/MCB.19.1.764.

Kelly, Michael R., Kamila Wisniewska, Matthew J. Regner, Michael W. Lewis, Andrea A. Perreault, Eric S. Davis, Douglas H. Phanstiel, Joel S. Parker, and Hector L. Franco. 2022. “A Multi-Omic Dissection of Super-Enhancer Driven Oncogenic Gene Expression Programs in Ovarian Cancer.” *Nature Communications* 13 (1): 4247.

Kiel, Douglas P., John P. Kemp, Fernando Rivadeneira, Jennifer J. Westendorf, David Karasik, Emma L. Duncan, Yuuki Imai, et al. 2020. “The Musculoskeletal Knowledge Portal: Making Omics Data Useful to the Broader Scientific Community.” *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research* 35 (9): 1626–33.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37 (8): 907–15.

Kim, Ji Hun, Mayuri Rege, Jacqueline Valeri, Margaret C. Dunagin, Aryeh Metzger, Katelyn R. Titus, Thomas G. Gilgenast, et al. 2019. “LADL: Light-Activated Dynamic Looping for Endogenous Gene Expression Control.” *Nature Methods* 16 (7): 633–39.

Kirkby, Nicholas S., Martina H. Lundberg, William R. Wright, Timothy D. Warner, Mark J. Paul-Clark, and Jane A. Mitchell. 2014. “COX-2 Protects against Atherosclerosis Independently of Local Vascular Prostacyclin: Identification of COX-2 Associated Pathways Implicate Rgl1 and Lymphocyte Networks.” *PloS One* 9 (6): e98165.

Klein, Isaac A., Ann Boija, Lena K. Afeyan, Susana Wilson Hawken, Mengyang Fan, Alessandra Dall'Agnese, Ozgur Oksuz, et al. 2020. "Partitioning of Cancer Therapeutics in Nuclear Condensates." *Science* 368 (6497): 1386–92.

Klein, Jason C., Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure. 2020. "A Systematic Evaluation of the Design and Context Dependencies of Massively Parallel Reporter Assays." *Nature Methods* 17 (11): 1083–91.

Knight, P. A., and D. Ruiz. 2013. "A Fast Algorithm for Matrix Balancing." *IMA Journal of Numerical Analysis* 33 (3): 1029–47.

Kovar, Heinrich. 2011. "Dr. Jekyll and Mr. Hyde: The Two Faces of the FUS/EWS/TAF15 Protein Family." *Sarcoma* 2011: 837474.

Kramer, Nicole E., Eric S. Davis, Craig D. Wenger, Erika M. Deoudes, Sarah M. Parker, Michael I. Love, and Douglas H. Phanstiel. 2022. "Plotgardener: Cultivating Precise Multi-Panel Figures in R." *Bioinformatics* , February. https://doi.org/10.1093/bioinformatics/btac057.

Kroon, E., U. Thorsteinsdottir, N. Mayotte, T. Nakamura, and G. Sauvageau. 2001. "NUP98-HOXA9 Expression in Hemopoietic Stem Cells Induces Chronic and Acute Myeloid Leukemias in Mice." *The EMBO Journal* 20. https://doi.org/10.1093/emboj/20.3.350.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. "Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq." *Nature Genetics* 48 (2): 206–13.

Laarman, Melanie D., Geert Geeven, Phil Barnett, Null Null, Gabriël J. E. Rinkel, Wouter de Laat, Ynte M. Ruigrok, and Jeroen Bakkers. 2019. "Chromatin Conformation Links Putative Enhancers in Intracranial Aneurysm–Associated Regions to Potential Candidate Genes." *Journal of the American Heart Association* 8 (9): e011201.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

LaRonde-LeBlanc, N. A., and C. Wolberger. 2003. "Structure of HoxA9 and Pbx1 Bound to DNA: Hox Hexapeptide and DNA Recognition Anterior to Posterior." *Genes & Development* 17. https://doi.org/10.1101/gad.1103303.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology* 9 (8): e1003118.

Lawrence, Moyra, Sylvain Daujat, and Robert Schneider. 2016. "Lateral Thinking: How Histone Modifications Regulate Gene Expression." *Trends in Genetics: TIG* 32 (1): 42–56.

Lee, Jongjoo, Ivan Krivega, Ryan K. Dale, and Ann Dean. 2017. "The LDB1 Complex Co-Opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions." *Cell Reports* 19 (12): 2490–2502.

Lee, Stuart, Dianne Cook, and Michael Lawrence. 2019. "Plyranges: A Grammar of Genomic Data Transformation." *Genome Biology* 20 (1): 4.

Leidescher, Susanne, Johannes Ribisel, Simon Ullrich, Yana Feodorova, Erica Hildebrand, Alexandra Galitsyna, Sebastian Bultmann, et al. 2022. "Spatial Organization of Transcribed Eukaryotic Genes." *Nature Cell Biology*, February. https://doi.org/10.1038/s41556-022-00847-6.

Lengronne, Armelle, Yuki Katou, Saori Mori, Shihori Yokobayashi, Gavin P. Kelly, Takehiko Itoh, Yoshinori Watanabe, Katsuhiko Shirahige, and Frank Uhlmann. 2004. "Cohesin Relocation from Sites of Chromosomal Loading to Places of Convergent Transcription." *Nature* 430 (6999): 573–78.

Li, B., and C. N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12. https://doi.org/10.1186/1471-2105-12-323.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv [q-Bio.GN]*. http://arxiv.org/abs/1303.3997.

Li, J. 2021. “ZMYND11-MBTD1 Induces Leukemogenesis through Hijacking NuA4/TIP60 Acetyltransferase Complex and a PWWP-Mediated Chromatin Association Mechanism.” *Nature Communications* 12. https://doi.org/10.1038/s41467-021-21357-3.

Li, Rong, Yuxiang Liang, and Bin Lin. 2022. “Accumulation of Systematic TPM1 Mediates Inflammation and Neuronal Remodeling by Phosphorylating PKA and Regulating the FABP5/NF-KB Signaling Pathway in the Retina of Aged Mice.” *Aging Cell* 21 (3): e13566.

Li, Xin, Michael B. Ellman, Jeffrey S. Kroin, Di Chen, Dongyao Yan, Katalin Mikecz, K. C. Ranjan, et al. 2012. “Species-Specific Biological Effects of FGF-2 in Articular Cartilage: Implication for Distinct Roles within the FGF Receptor Family.” *Journal of Cellular Biochemistry* 113 (7): 2532–42.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.

Lin, Alvin C., Brian L. Seeto, Justyna M. Bartoszko, Michael A. Khoury, Heather Whetstone, Louisa Ho, Claire Hsu, S. Amanda Ali, and Benjamin A. Alman. 2009. “Modulating Hedgehog Signaling Can Attenuate the Severity of Osteoarthritis.” *Nature Medicine*. https://doi.org/10.1038/nm.2055.

Lin, Dejun, Justin Sanders, and William Stafford Noble. 2021. “HiCRep.Py : Fast Comparison of Hi-C Contact Matrices in Python.” *Bioinformatics* , February. https://doi.org/10.1093/bioinformatics/btab097.

Loeser, Richard F. 2014. “Integrins and Chondrocyte-Matrix Interactions in Articular Cartilage.” *Matrix Biology: Journal of the International Society for Matrix Biology* 39 (October): 11–16.

Loeser, Richard F., Steven R. Goldring, Carla R. Scanzello, and Mary B. Goldring. 2012. “Osteoarthritis: A Disease of the Joint as an Organ.” *Arthritis and Rheumatism* 64 (6): 1697–1707.

Loeser, Richard F., Carol A. Pacione, and Susan Chubinskaya. 2003. “The Combination of Insulin-like Growth Factor 1 and Osteogenic Protein 1 Promotes Increased Survival of and Matrix Synthesis by Normal and Osteoarthritic Human Articular Chondrocytes.” *Arthritis and Rheumatism* 48 (8): 2188–96.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Lovén, J. 2013. “Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers.” *Cell* 153. https://doi.org/10.1016/j.cell.2013.03.036.

Lu, Feiyue, and Timothée Lionnet. 2021. “Transcription Factor Dynamics.” *Cold Spring Harbor Perspectives in Biology* 13 (11). https://doi.org/10.1101/cshperspect.a040949.

Lu, R. 2016. “Epigenetic Perturbations by Arg882-Mutated Dnmt3a Potentiate Aberrant Stem Cell Gene-Expression Program and Acute Leukemia Development.” *Cancer Cell* 30. https://doi.org/10.1016/j.ccell.2016.05.008.

Lun, Aaron T. L., Malcolm Perry, and Elizabeth Ing-Simmons. 2016. “Infrastructure for Genomic Interactions: Bioconductor Classes for Hi-C, ChIA-PET and Related Experiments.” *F1000Research* 5 (May): 950.

L'Yi, Sehi, Qianwen Wang, Fritz Lekschas, and Nils Gehlenborg. 2022. "Gosling: A Grammar-Based Toolkit for Scalable and Interactive Genomics Data Visualization." *IEEE Transactions on Visualization and Computer Graphics* 28 (1): 140–50.

Mangiola, Stefano, Ramyar Molania, Ruining Dong, Maria A. Doyle, and Anthony T. Papenfuss. 2021. "Tidybulk: An R Tidy Framework for Modular Transcriptomic Data Analysis." *Genome Biology* 22 (1): 42.

Marcel, Shelsa S., Austin L. Quimby, Melodie P. Noel, Oscar C. Jaimes, Marjan Mehrab-Mohseni, Suud A. Ashur, Brian Velasco, et al. 2021. "Genome-Wide Cancer-Specific Chromatin Accessibility Patterns Derived from Archival Processed Xenograft Tumors." *Genome Research*, November. https://doi.org/10.1101/gr.275219.121.

Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." *Science* 337 (6099): 1190–95.

Mendes, Adélia, and Birthe Fahrenkrog. 2019. "NUP214 in Leukemia: It's More than Transport." *Cells* 8 (1). https://doi.org/10.3390/cells8010076.

Metcalf, D., C. J. Greenhalgh, E. Viney, T. A. Willson, R. Starr, N. A. Nicola, D. J. Hilton, and W. S. Alexander. 2000. "Gigantism in Mice Lacking Suppressor of Cytokine Signalling-2." *Nature* 405 (6790): 1069–73.

Miyamoto, Yoshinari, Akihiko Mabuchi, Dongquan Shi, Toshikazu Kubo, Yoshio Takatori, Susumu Saito, Mikihiro Fujioka, et al. 2007. "A Functional Polymorphism in the 5' UTR of GDF5 Is Associated with Susceptibility to Osteoarthritis." *Nature Genetics* 39 (4): 529–33.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. 2021. "Sustainable Data Analysis with Snakemake." *F1000Research* 10 (January): 33.

Monahan, Kevin, Adan Horta, and Stavros Lomvardas. 2019. "LHX2- and LDB1-Mediated Trans Interactions Regulate Olfactory Receptor Choice." *Nature* 565 (7740): 448–53.

Monti-Rocha, Renata, Allysson Cramer, Paulo Gaio Leite, Maísa Mota Antunes, Rafaela Vaz Sousa Pereira, Andréia Barroso, Celso M. Queiroz-Junior, et al. 2018. "SOCS2 Is Critical for the Balancing of Immune Response and Oxidate Stress Protecting Against Acetaminophen-Induced Acute Liver Injury." *Frontiers in Immunology* 9: 3134.

Morgan, M., V. Obenchain, J. Hester, and H. Pagès. n.d. "SummarizedExperiment: SummarizedExperiment Container." *R Package Version*.

Morgan, Martin, Marc Carlson, Dan Tenenbaum, and Sonali Arora. 2017. "AnnotationHub: Client to Access AnnotationHub Resources." *R Package Version* 2 (1).

Motomura, Hiraku, Shoji Seki, Shunichi Shiozawa, Yukihiko Aikawa, Makiko Nogami, and Tomoatsu Kimura. 2018. "A Selective C-Fos/AP-1 Inhibitor Prevents Cartilage Destruction and Subsequent Osteophyte Formation." *Biochemical and Biophysical Research Communications* 497 (2): 756–61.

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. "HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture." *Nature Methods* 13 (11): 919–22.

Murray, Dylan T., Masato Kato, Yi Lin, Kent R. Thurber, Ivan Hung, Steven L. McKnight, and Robert Tycko. 2017. "Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains." *Cell* 171 (3): 615-627.e16.

Muthuirulan, Pushpanathan, Dewei Zhao, Mariel Young, Daniel Richard, Zun Liu, Alireza Emami, Gabriela Portilla, et al. 2021. “Joint Disease-Specificity at the Regulatory Base-Pair Level.” *Nature Communications* 12 (1): 4161.

Muzellec, Boris, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. 2022. “PyDESeq2: A Python Package for Bulk RNA-Seq Differential Expression Analysis.” *BioRxiv*. https://doi.org/10.1101/2022.12.14.520412.

Nagano, Takashi, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. 2013. “Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure.” *Nature* 502 (7469): 59–64.

Nair, S. J. 2019. “Phase Separation of Ligand-Activated Enhancers Licenses Cooperative Chromosomal Enhancer Assembly.” *Nature Structural & Molecular Biology* 26. https://doi.org/10.1038/s41594-019-0190-5.

Nesvizhskii, A. I., A. Keller, E. Kolker, and R. Aebersold. 2003. “A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry.” *Analytical Chemistry* 75. https://doi.org/10.1021/ac0341261.

Nishizaki, Sierra S., Natalie Ng, Shengcheng Dong, Robert S. Porter, Cody Morterud, Colten Williams, Courtney Asman, Jessica A. Switzenberg, and Alan P. Boyle. 2020. “Predicting the Effects of SNPs on Transcription Factor Binding Affinity.” *Bioinformatics* 36 (2): 364–72.

Okura, Toshiaki, Masaki Matsushita, Kenichi Mishima, Ryusaku Esaki, Taisuke Seki, Naoki Ishiguro, and Hiroshi Kitoh. 2018. “Activated FGFR3 Prevents Subchondral Bone Sclerosis during the Development of Osteoarthritis in Transgenic Mice with Achondroplasia.” *Journal of Orthopaedic Research: Official Publication of the Orthopaedic Research Society* 36 (1): 300–308.

Olan, Ioana, Aled J. Parry, Stefan Schoenfelder, Masako Narita, Yoko Ito, Adelyne S. L. Chan, Guy St C. Slater, et al. 2020. “Transcription-Dependent Cohesin Repositioning Rewires Chromatin Loops in Cellular Senescence.” *Nature Communications* 11 (1): 6049.

Olshansky, Moshe. 2021. “EigenVector.” August 1, 2021. https://github.com/moshe-olshansky/EigenVector.

Open2C, Nezar Abdennur, Sameer Abraham, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, Betul A. Oksuz, and Sergey V. Venev. 2022. “Cooltools: Enabling High-Resolution Hi-C Analysis in Python.” *BioRxiv*. https://doi.org/10.1101/2022.10.31.514564.

Pagès, H., P. Aboyoun, R. Gentleman, and S. DebRoy. 2021. “Biostrings: Efficient Manipulation of Biological Strings.” https://bioconductor.org/packages/Biostrings.

Pagès, H., Peter Hickey, and A. Lun. 2021. “DelayedArray: A Unified Framework for Working Transparently with on-Disk and in-Memory Array-like Datasets.” *R Package Version 0. 20. 0*.

Pagès, Hervé. 2020. “HDF5Array: HDF5 Backend for DelayedArray Objects.” R package version.

Pak, C. W. 2016. “Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein.” *Molecular Cell* 63. https://doi.org/10.1016/j.molcel.2016.05.042.

Papageorgiou, Louis, Picasi Eleni, Sofia Raftopoulou, Meropi Mantaiou, Vasileios Megalooikonomou, and Dimitrios Vlachakis. 2018. “Genomic Big Data Hitting the Storage Bottleneck.” *EMBnet.Journal* 24 (April). https://doi.org/10.14778/1687553.1687625.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Patwardhan, Mayura N., Craig D. Wenger, Eric S. Davis, and Douglas H. Phanstiel. 2019. “Bedtoolsr: An R Package for Genomic Data Analysis and Manipulation.” *Journal of Open Source Software* 4 (44). https://doi.org/10.21105/joss.01742.

Paul, Indranil, Tanveer S. Batth, Diego Iglesias-Gato, Amna Al-Araimi, Ibrahim Al-Haddabi, Amira Alkharusi, Gunnar Norstedt, Jesper V. Olsen, Fahad Zadjali, and Amilcar Flores-Morales. 2017. “The Ubiquitin Ligase Cullin5SOCS2 Regulates NDR1/STK38 Stability and NF-KB Transactivation.” *Scientific Reports* 7 (February): 42800.

Pelletier, J. P., J. Martel-Pelletier, and S. B. Abramson. 2001. “Osteoarthritis, an Inflammatory Disease: Potential Implication for the Selection of New Therapeutic Targets.” *Arthritis and Rheumatism* 44 (6): 1237–47.

Phanstiel, Douglas H., Kevin Van Bortle, Damek Spacek, Gaelen T. Hess, Muhammad Saad Shamim, Ido Machol, Michael I. Love, Erez Lieberman Aiden, Michael C. Bassik, and Michael P. Snyder. 2017. “Static and Dynamic DNA Loops Form AP-1-Bound Activation Hubs during Macrophage Development.” *Molecular Cell* 67 (6): 1037-1048.e6.

“Picard.” n.d. Accessed March 27, 2022. https://broadinstitute.github.io/picard/.

Pradhan, Biswajit, Roman Barth, Eugene Kim, Iain F. Davidson, Benedikt Bauer, Theo van Laar, Wayne Yang, et al. 2021. “SMC Complexes Can Traverse Physical Roadblocks Bigger than Their Ring Size.” *BioRxiv*. https://doi.org/10.1101/2021.07.15.452501.

Pulai, Judit I., Hong Chen, Hee-Jeong Im, Sanjay Kumar, Charles Hanning, Priti S. Hegde, and Richard F. Loeser. 2005. “NF-KB Mediates the Stimulation of Cytokine and Chemokine Expression by Human Articular Chondrocytes in Response to Fibronectin Fragments.” *The Journal of Immunology* 174 (9): 5781–88.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.

Qamar, S. 2018. “FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions.” *Cell* 173. https://doi.org/10.1016/j.cell.2018.03.056.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.

Quiroga, Ivana Y., Jeong Hyun Ahn, Gang Greg Wang, and Douglas Phanstiel. 2022. “Oncogenic Fusion Proteins and Their Role in Three-Dimensional Chromatin Structure, Phase Separation, and Cancer.” *Current Opinion in Genetics & Development* 74 (June): 101901.

R Core Team. 2022. “R: A Language and Environment for Statistical Computing.” https://www.R-project.org.

Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. 2016. “DeepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis.” *Nucleic Acids Research* 44 (W1): W160-5.

Rao, Suhas S. P., Su-Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong-Rim Kieffer-Kwon, Adrian L. Sanborn, et al. 2017. “Cohesin Loss Eliminates All Loop Domains.” *Cell* 171 (2): 305-320.e24.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.

Reed, K. S. M., V. Ulici, C. Kim, S. Chubinskaya, R. F. Loeser, and D. H. Phanstiel. 2021. “Transcriptional Response of Human Articular Chondrocytes Treated with Fibronectin Fragments: An in Vitro Model of the Osteoarthritis Phenotype.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 29 (2): 235–47.

Reed, Kathleen S. M., Eric S. Davis, Marielle L. Bond, Alan Cabrera, Eliza Thulson, Ivana Yoseli Quiroga, Shannon Cassel, et al. 2022. “Temporal Analysis Suggests a Reciprocal Relationship between 3D Chromatin Structure and Transcription.” *Cell Reports* 41 (5): 111567.

Ren, Y., H. -. S. Seo, G. Blobel, and A. Hoelz. 2010. “Structural and Functional Analysis of the Interaction between the Nucleoporin Nup98 and the MRNA Export Factor Rae1.” *Proceedings of the National Academy of Sciences of the United States of America* 107. https://doi.org/10.1073/pnas.1005389107.

Ren, Z. 2019. “PHF19 Promotes Multiple Myeloma Tumorigenicity through PRC2 Activation and Broad H3K27me3 Domain Formation.” *Blood* 134. https://doi.org/10.1182/blood.2019000578.

Reynard, L. N., and M. J. Barter. 2020. “Osteoarthritis Year in Review 2019: Genetics, Genomics and Epigenetics.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 28 (3): 275–84.

Richard, Daniel, Zun Liu, Jiaxue Cao, Ata M. Kiapour, Jessica Willen, Siddharth Yarlagadda, Evelyn Jagoda, et al. 2020. “Evolutionary Selection and Constraint on Human Knee Chondrocyte Regulation Impacts Osteoarthritis Risk.” *Cell* 181 (2): 362-381.e28.

Roden, Christine, and Amy S. Gladfelter. 2021. “RNA Contributions to the Form and Function of Biomolecular Condensates.” *Nature Reviews. Molecular Cell Biology* 22 (3): 183–95.

Rosencrance, Celeste D., Haneen N. Ammouri, Qi Yu, Tiffany Ge, Emily J. Rendleman, Stacy A. Marshall, and Kyle P. Eagen. 2020. “Chromatin Hyperacetylation Impacts Chromosome Folding by Forming a Nuclear Subcompartment.” *Molecular Cell* 78 (1): 112-126.e12.

Roux, K. J., D. I. Kim, and B. Burke. 2013. “BioID: A Screen for Protein–Protein Interactions.” *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.]* 74. https://doi.org/10.1002/0471140864.ps1923s74.

Roux, K. J., D. I. Kim, B. Burke, and D. G. May. 2018. “BioID: A Screen for Protein–Protein Interactions.” *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.]* 91. https://doi.org/10.1002/cpps.51.

Rowley, M. Jordan, and Victor G. Corces. 2018. “Organizational Principles of 3D Genome Architecture.” *Nature Reviews. Genetics* 19 (12): 789–800.

Rowley, M. Jordan, Axel Poulet, Michael H. Nichols, Brianna J. Bixler, Adrian L. Sanborn, Elizabeth A. Brouhard, Karen Hermetz, et al. 2020. “Analysis of Hi-C Data Using SIP Effectively Identifies Loops in Organisms from C. Elegans to Mammals.” *Genome Research* 30 (3): 447–58.

Ruff, Kiersten M., and Rohit V. Pappu. 2021. “AlphaFold and Implications for Intrinsically Disordered Proteins.” *Journal of Molecular Biology* 433 (20): 167208.

Sabari, B. R. 2018. “Coactivator Condensation at Super-Enhancers Links Phase Separation and Gene Control.” *Science* 361. https://doi.org/10.1126/science.aar3958.

Sahin, Merve, Wilfred Wong, Yingqian Zhan, Kinsey Van Deynze, Richard Koche, and Christina S. Leslie. 2021. “HiC-DC+ Enables Systematic 3D Interaction Calls and Differential Analysis for Hi-C and HiChIP.” *Nature Communications* 12 (1): 3366.

Samvelyan, Hasmik Jasmine, Carmen Huesa, Lin Cui, Colin Farquharson, and Katherine Ann Staines. 2022. “The Role of Accelerated Growth Plate Fusion in the Absence of SOCS2 on Osteoarthritis Vulnerability.” *Bone & Joint Research* 11 (3): 162–70.

Sanborn, Adrian L., Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, et al. 2015. “Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (47): E6456-65.

Sandell, Linda J., and Thomas Aigner. 2001. “Articular Cartilage and Changes in Arthritis: Cell Biology of Osteoarthritis.” *Arthritis Research & Therapy* 3 (2): 107.

Santangelo, Carmela, Angela Scipioni, Lorella Marselli, Piero Marchetti, and Francesco Dotta. 2005. “Suppressor of Cytokine Signaling Gene Expression in Human Pancreatic Islets: Modulation by Cytokines.” *European Journal of Endocrinology / European Federation of Endocrine Societies* 152 (3): 485–89.

Schmidt, Ellen M., Ji Zhang, Wei Zhou, Jin Chen, Karen L. Mohlke, Y. Eugene Chen, and Cristen J. Willer. 2015. “GREGOR: Evaluating Global Enrichment of Trait-Associated Variants in Epigenomic Features Using a Systematic, Data-Driven Approach.” *Bioinformatics* 31 (16): 2601–6.

Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer-Promoter Contacts in Gene Expression Control.” *Nature Reviews. Genetics* 20 (8): 437–55.

Schwarzer, Wibke, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, et al. 2017. “Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal.” *Nature* 551 (7678): 51–56.

Sekhon, Jasjeet S. 2011. “Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R.” *Journal of Statistical Software* 42 (June): 1–52.

Servant, Nicolas, Nelle Varoquaux, Bryan R. Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. 2015. “HiC-Pro: An Optimized and Flexible Pipeline for Hi-C Data Processing.” *Genome Biology* 16 (December): 259.

Sey, Nancy Y. A., Benxia Hu, Won Mah, Harper Fauni, Jessica Caitlin McAfee, Prashanth Rajarajan, Kristen J. Brennand, Schahram Akbarian, and Hyejung Won. 2020. “A Computational Tool (H-MAGMA) for Improved Prediction of Brain-Disorder Risk Genes by Incorporating Brain Chromatin Interaction Profiles.” *Nature Neuroscience* 23 (4): 583–93.

Shin, Y. 2018. “Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome.” *Cell* 175. https://doi.org/10.1016/j.cell.2018.10.057.

Shu, Cindy C., Miriam T. Jackson, Margaret M. Smith, Susan M. Smith, Steven Penm, Megan S. Lord, John M. Whitelock, Christopher B. Little, and James Melrose. 2016. “Ablation of Perlecan Domain 1 Heparan Sulfate Reduces Progressive Cartilage Degradation, Synovitis, and Osteophyte Size in a Preclinical Model of Posttraumatic Osteoarthritis.” *Arthritis & Rheumatology*. https://doi.org/10.1002/art.39529.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. 2006. “Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture–on-Chip (4C).” *Nature Genetics* 38 (11): 1348–54.

Skene, Peter J., and Steven Henikoff. 2017. “An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites.” *ELife* 6 (January). https://doi.org/10.7554/eLife.21856.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. “Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences.” *F1000Research* 4 (December): 1521.

Song, Jinsoo, Dongkyun Kim, Chang Hoon Lee, Myeung Su Lee, Churl-Hong Chun, and Eun-Jung Jin. 2013. “MicroRNA-488 Regulates Zinc Transporter SLC39A8/ZIP8 during Pathogenesis of Osteoarthritis.” *Journal of Biomedical Science* 20 (May): 31.

Southam, Lorraine, Julio Rodriguez-Lopez, James M. Wilkins, Manuel Pombo-Suarez, Sarah Snelling, Juan J. Gomez-Reino, Kay Chapman, Antonio Gonzalez, and John Loughlin. 2007. “An SNP in the 5’-UTR of GDF5 Is Associated with Osteoarthritis Susceptibility in Europeans and with in Vivo Differences in Allelic Expression in Articular Cartilage.” *Human Molecular Genetics* 16 (18): 2226–32.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. “Structural Variation in the 3D Genome.” *Nature Reviews. Genetics* 19 (7): 453–67.

Starr, R., T. A. Willson, E. M. Viney, L. J. Murray, J. R. Rayner, B. J. Jenkins, T. J. Gonda, et al. 1997. “A Family of Cytokine-Inducible Inhibitors of Signalling.” *Nature* 387 (6636): 917–21.

Stauffer, W., H. Sheng, and H. N. Lim. 2018. “EzColocalization: An ImageJ Plugin for Visualizing and Measuring Colocalization in Cells and Organisms.” *Scientific Reports* 8. https://doi.org/10.1038/s41598-018-33592-8.

Steinberg, Julia, Lorraine Southam, Natalie C. Butterfield, Theodoros I. Roumeliotis, Andreas Fontalis, Matthew J. Clark, Raveen L. Jayasuriya, et al. 2020. “Decoding the Genomic Basis of Osteoarthritis.” *BioRxiv*. https://doi.org/10.1101/835850.

Steinberg, Julia, Lorraine Southam, Theodoros I. Roumeliotis, Matthew J. Clark, Raveen L. Jayasuriya, Diane Swift, Karan M. Shah, et al. 2021. “A Molecular Quantitative Trait Locus Map for Osteoarthritis.” *Nature Communications* 12 (1): 1309.

Stigler, Johannes, Gamze Ö. Çamdere, Douglas E. Koshland, and Eric C. Greene. 2016. “Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin.” *Cell Reports*. https://doi.org/10.1016/j.celrep.2016.04.003.

Stik, Grégoire, Enrique Vidal, Mercedes Barrero, Sergi Cuartero, Maria Vila-Casadesús, Julen Mendieta-Esteban, Tian V. Tian, et al. 2020. “CTCF Is Dispensable for Immune Cell Transdifferentiation but Facilitates an Acute Inflammatory Response.” *Nature Genetics* 52 (7): 655–61.

Strom, A. R. 2017. “Phase Separation Drives Heterochromatin Domain Formation.” *Nature* 547. https://doi.org/10.1038/nature22989.

Subramanian, A. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102. https://doi.org/10.1073/pnas.0506580102.

Tang, Junzhou, Nan Su, Siru Zhou, Yangli Xie, Junlan Huang, Xuan Wen, Zuqiang Wang, et al. 2016. “Fibroblast Growth Factor Receptor 3 Inhibits Osteoarthritis Progression in the Knee Joints of Adult Mice.” *Arthritis & Rheumatology (Hoboken, N.J.)* 68 (10): 2432–43.

Tinevez, J. -. Y. 2017. “TrackMate: An Open and Extensible Platform for Single-Particle Tracking.” *Methods* 115. https://doi.org/10.1016/j.ymeth.2016.09.016.

Tong, Xuhui, Rong Tang, Jin Xu, Wei Wang, Yingjun Zhao, Xianjun Yu, and Si Shi. 2022. “Liquid-Liquid Phase Separation in Tumor Biology.” *Signal Transduction and Targeted Therapy* 7 (1): 221.

Tuerlings, Margo, Marcella Hoolwerff, Evelyn Houtman, Eka H. E. Suchiman, Nico Lakenberg, Hailiang Mei, Enrike H. M. Linden, et al. 2021. “RNA Sequencing Reveals Interacting Key Determinants of

Osteoarthritis Acting in Subchondral Bone and Articular Cartilage: Identification of *IL11* and *CHADL* as Attractive Treatment Targets." *Arthritis & Rheumatology*. https://doi.org/10.1002/art.41600.

Umans, Benjamin D., Alexis Battle, and Yoav Gilad. 2021. "Where Are the Disease-Associated EQTLs?" *Trends in Genetics: TIG* 37 (2): 109–24.

Venetianer, Pál. 2012. "Are Synonymous Codons Indeed Synonymous?" *Biomolecular Concepts* 3 (1): 21–28.

Vian, Laura, Aleksandra Pękowska, Suhas S. P. Rao, Kyong-Rim Kieffer-Kwon, Seolkyoung Jung, Laura Baranello, Su-Chen Huang, et al. 2018. "The Energetics and Physiological Impact of Cohesin Extrusion." *Cell* 173 (5): 1165-1178.e20.

Vilarrasa-Blasi, Roser, Paula Soler-Vila, Núria Verdaguer-Dot, Núria Russiñol, Marco Di Stefano, Vicente Chapaprieta, Guillem Clot, et al. 2021. "Dynamics of Genome Architecture and Chromatin Function during Human B Cell Differentiation and Neoplastic Transformation." *Nature Communications* 12 (1): 651.

Virtanen, P. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17. https://doi.org/10.1038/s41592-019-0686-2.

Wan, Liling, Shasha Chong, Fan Xuan, Angela Liang, Xiaodong Cui, Leah Gates, Thomas S. Carroll, et al. 2020. "Impaired Cell Fate through Gain-of-Function Mutations in a Chromatin Reader." *Nature* 577 (7788): 121–26.

Wang, G. G. 2006. "Quantitative Production of Macrophages or Neutrophils Ex Vivo Using Conditional Hoxb8." *Nature Methods* 3. https://doi.org/10.1038/nmeth865.

———. 2009. "Haematopoietic Malignancies Caused by Dysregulation of a Chromatin-Binding PHD Finger." *Nature* 459. https://doi.org/10.1038/nature08036.

Wang, G. G., L. Cai, M. P. Pasillas, and M. P. Kamps. 2007. "NUP98-NSD1 Links H3K36 Methylation to Hox-A Gene Activation and Leukaemogenesis." *Nature Cell Biology* 9. https://doi.org/10.1038/ncb1608.

Wang, G. G., M. P. Pasillas, and M. P. Kamps. 2005. "Meis1 Programs Transcription of FLT3 and Cancer Stem Cell Character, Using a Mechanism That Requires Interaction with Pbx and a Novel Function of the Meis1 C-Terminus." *Blood* 106. https://doi.org/10.1182/blood-2004-12-4664.

Wang, J. 2018. "A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins." *Cell* 174. https://doi.org/10.1016/j.cell.2018.06.006.

Wang, K. 2010. "MapSplice: Accurate Mapping of RNA-Seq Reads for Splice Junction Discovery." *Nucleic Acids Research* 38. https://doi.org/10.1093/nar/gkq622.

Wang, L. 2019. "Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism." *Molecular Cell* 76. https://doi.org/10.1016/j.molcel.2019.08.019.

Wang, Tiantian, and Chengqi He. 2018. "Pro-Inflammatory Cytokines: The Link between Obesity and Osteoarthritis." *Cytokine & Growth Factor Reviews* 44 (December): 38–50.

Wanner, John, Roopashree Subbaiah, Yelenna Skomorovska-Prokvolit, Yousef Shishani, Eric Boilard, Sujatha Mohan, Robert Gillespie, Masaru Miyagi, and Reuben Gobezie. 2013. "Proteomic Profiling and Functional Characterization of Early and Late Shoulder Osteoarthritis." *Arthritis Research & Therapy* 15 (6): R180.

Westreich, Daniel, and Stephen R. Cole. 2010. "Invited Commentary: Positivity in Practice." *American Journal of Epidemiology*.

Wickham, Hadley. 2016. “Ggplot2: Elegant Graphics for Data Analysis.” Springer-Verlag New York. https://ggplot2.tidyverse.org.

Winick-Ng, Warren, Alexander Kukalev, Izabela Harabula, Luna Zea-Redondo, Dominik Szabó, Mandy Meijer, Leonid Serebreni, et al. 2021. “Cell-Type Specialization Is Encoded by Specific Chromatin Topologies.” *Nature* 599 (7886): 684–91.

Wojdasiewicz, Piotr, Łukasz A. Poniatowski, and Dariusz Szukiewicz. 2014. “The Role of Inflammatory and Anti-Inflammatory Cytokines in the Pathogenesis of Osteoarthritis.” *Mediators of Inflammation* 2014 (April): 561459.

Won, Hyejung, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Jerry Huang, Carli K. Opland, Michael J. Gandal, et al. 2016. “Chromosome Conformation Elucidates Regulatory Relationships in Developing Human Brain.” *Nature* 538 (7626): 523–27.

Wood, Scott T., David L. Long, Julie A. Reisz, Raghunatha R. Yammani, Elizabeth A. Burke, Chananat Klomsiri, Leslie B. Poole, Cristina M. Furdui, and Richard F. Loeser. 2016. “Cysteine-Mediated Redox Regulation of Cell Signaling in Chondrocytes Stimulated with Fibronectin Fragments.” *Arthritis & Rheumatology* 68 (1): 117–26.

Xiao, Jordan Yupeng, Antonina Hafner, and Alistair N. Boettiger. 2021. “How Subtle Changes in 3D Structure Can Create Large Changes in Transcription.” *ELife* 10 (July). https://doi.org/10.7554/eLife.64320.

Xie, D. L., R. Meyers, and G. A. Homandberg. 1992. “Fibronectin Fragments in Osteoarthritic Synovial Fluid.” *The Journal of Rheumatology* 19 (9): 1448–52.

Xu, B. 2015. “Selective Inhibition of EZH2 and EZH1 Enzymatic Activity by a Small Molecule Suppresses MLL-Rearranged Leukemia.” *Blood* 125. https://doi.org/10.1182/blood-2014-06-581082.

Xu, H. 2016. “NUP98 Fusion Proteins Interact with the NSL and MLL1 Complexes to Drive Leukemogenesis.” *Cancer Cell* 30. https://doi.org/10.1016/j.ccell.2016.10.019.

Yang, Jing, Amanda McGovern, Paul Martin, Kate Duffus, Xiangyu Ge, Peyman Zarrineh, Andrew P. Morris, et al. 2020. “Analysis of Chromatin Organization and Gene Expression in T Cells Identifies Functional Genes for Rheumatoid Arthritis.” *Nature Communications* 11 (1): 4402.

Yang, Tao, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C. Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. 2017. “HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient.” *Genome Research* 27 (11): 1939–49.

Yang, Yute, Panyang Shen, Teng Yao, Jun Ma, Zizheng Chen, Jinjin Zhu, Zhe Gong, Shuying Shen, and Xiangqian Fang. 2021. “Novel Role of CircRSU1 in the Progression of Osteoarthritis by Adjusting Oxidative Stress.” *Theranostics* 11 (4): 1877–1900.

Yu, M. 2015. “A Resource for Cell Line Authentication, Annotation and Quality Control.” *Nature* 520. https://doi.org/10.1038/nature14397.

Yung, E. 2011. “Delineating Domains and Functions of NUP98 Contributing to the Leukemogenic Activity of NUP98-HOX Fusions.” *Leukemia Research* 35. https://doi.org/10.1016/j.leukres.2010.10.006.

Zeng, Zishuo, and Yana Bromberg. 2019. “Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives.” *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2019.00914.

Zhang, Haoyue, Daniel J. Emerson, Thomas G. Gilgenast, Katelyn R. Titus, Yemin Lan, Peng Huang, Di Zhang, et al. 2019. “Chromatin Structure Dynamics during the Mitosis-to-G1 Phase Transition.” *Nature* 576 (7785): 158–62.

Zhang, Shu, Nadine Übelmesser, Mariano Barbieri, and Argyris Papantonis. 2022. “Enhancer-Promoter Contact Formation Requires RNAPII and Antagonizes Loop Extrusion.” *BioRxiv*. https://doi.org/10.1101/2022.07.04.498738.

Zhang, Yanxiao, Ting Li, Sebastian Preissl, Maria Luisa Amaral, Jonathan D. Grinstein, Elie N. Farah, Eugin Destici, et al. 2019. “Transcriptionally Active HERV-H Retrotransposons Demarcate Topologically Associating Domains in Human Pluripotent Stem Cells.” *Nature Genetics* 51 (9): 1380–88.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137.

Zheng, Hui, and Wei Xie. 2019. “The Role of 3D Genome Organization in Development and Cell Differentiation.” *Nature Reviews. Molecular Cell Biology* 20 (9): 535–50.

Zhou, Jingtian, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J. Sejnowski, Jesse R. Dixon, and Joseph R. Ecker. 2019. “Robust Single-Cell Hi-C Clustering by Convolution- and Random-Walk-Based Imputation.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (28): 14011–18.

Zhou, Siru, Yangli Xie, Wei Li, Junlan Huang, Zuqiang Wang, Junzhou Tang, Wei Xu, et al. 2016. “Conditional Deletion of Fgfr3 in Chondrocytes Leads to Osteoarthritis-like Defects in Temporomandibular Joint of Adult Mice.” *Scientific Reports*. https://doi.org/10.1038/srep24039.

Zhu, Anqi, Joseph G. Ibrahim, and Michael I. Love. 2019. “Heavy-Tailed Prior Distributions for Sequence Count Data: Removing the Noise and Preserving Large Differences.” *Bioinformatics* 35 (12): 2084–92.

Zhu, Guangya, Jingjing Xie, Zhenzhen Fu, Mingliang Wang, Qichen Zhang, Hao He, Zijun Chen, Xin Guo, and Jidong Zhu. 2021. “Pharmacological Inhibition of SRC-1 Phase Separation Suppresses YAP Oncogenic Transcription Activity.” *Cell Research* 31 (9): 1028–31.

Zhu, Yaqian, Rebecca A. Hubbard, Jessica Chubak, Jason Roy, and Nandita Mitra. 2021. “Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches.” *Pharmacoepidemiology and Drug Safety* 30 (11): 1471–85.