Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

Isabelle Seufert, M.Sc.

Born in Achern, Germany

Oral examination on December 11th, 2024

Decoding principles of transcription regulation: A single-cell chromatin accessibility approach

Referees: Prof. Dr. Karsten Rippe

Prof. Dr. Oliver Stegle

Abstract

Eukaryotic cells precisely regulate gene expression programs in response to environmental or cellular stimuli, controlling the timing and output of thousands of genes. These transcriptional responses are regulated by a complex network of different molecular mechanisms and their spatial and temporal organization. Critical components of this regulation include local chromatin states and transcription factor (TF) binding kinetics at *cis*-regulatory elements (CREs) like promoters and enhancers, their long-range interactions, and the local enrichment of TFs and transcription machinery into nuclear subcompartments. Transcription regulation by TFs has been extensively studied using fluorescence microscopy-based assays, while chromatin topology is usually explored through next-generation sequencing (NGS) methods. These NGS techniques encompass chromosome conformation capture, chromatin immunoprecipitation, and chromatin accessibility assays. Data from single-cell chromatin accessibility sequencing (scATAC-seq) allows to identify CREs in accessible and active states as well as locus-specific TF binding activity at single-cell resolution. However, a model of transcription regulation that integrates genome-wide data on TF dynamics and chromatin topology is still lacking.

This thesis aimed to develop such a model of transcription regulation by collectively inferring local chromatin states, locus-specific TF binding activity, and global chromatin organization from scATAC-seq data. To achieve this, I addressed three specific objectives: (i) Advancing the experimental and computational analysis of scATAC-seq, (ii) developing a computational framework to dissect molecular mechanisms underlying chromatin co-accessibility, and (iii) identifying the structure-function relationship between different regulatory mechanisms and their transcriptional output.

In the first part of this thesis, I identified data sparsity as a key challenge in scATAC-seq data analysis. To address this, I introduced the TurboATAC protocol, which reduces data sparsity by optimizing the transposase reaction efficiency in scATAC-seq. Additionally, I developed a method for allele-aware quantification of scATAC-seq data. Together, these advances enabled me to distinguish true biological variability between cells from data sparsity at individual genomic loci.

In the second part, I developed the R package RWireX, a computational framework designed to resolve different layers of chromatin co-accessibility between multiple genomic loci. RWireX differentiates between different co-accessibility features: Autonomous links of co-accessibility (ACs) and domains of contiguous co-accessibility

(DCs). ACs represent spatial contacts between co-active distal chromatin sites, while DCs likely result from local enrichment of TF binding activity in nuclear subcompartments. Furthermore, RWireX revealed different types of ACs, driven either by targeted structural chromatin loops or by random spatial interactions within dynamic chromatin regions.
In the third part, I analyzed various human and mouse cellular systems under perturbation with RWireX to link genome-wide regulation mechanisms with their functional transcriptional output. These analyses revealed that promoters, ACs and DCs regulate with distinct transcriptional bursting kinetics. Promoters and DCs primarily regulate burst size, leading to a rapid transcriptional response. In contrast, ACs mainly regulate burst frequency, revealing slower transcriptional changes. Promoters induce a significantly stronger response compared to ACs and DCs. However, ACs and DCs can co-regulate multiple genes by either inducing co-expression or alternating patterns of expression in single cells.

Finally, I combined these findings to derive the AC/DC model of transcription regulation, which links promoter-mediated regulation, chromatin contacts (via targeted loops or stochastic interactions), and local subcompartments with enriched TF binding activity to their specific transcriptional effects. With this model, the thesis provides a novel approach to explain how mammalian systems precisely regulate the magnitude, direction, and temporal hierarchy of transcriptional responses to both external and internal stimuli.

Zusammenfassung

Eukaryotische Zellen regulieren spezifische Genexpressionsprogramme als Reaktion auf externe oder interne Stimuli. Dabei steuern sie präzise das Timing und die Aktivität von Tausenden von Genen. Diese transkriptionellen Antworten werden durch ein komplexes Netzwerk verschiedener molekularer Mechanismen sowie deren räumliche und zeitliche Organisation reguliert. Wesentliche Komponenten dieser Regulation umfassen lokale Chromatinzustände und die Bindungskinetik von Transkriptionsfaktoren (TFs) an cis-regulatorische Elemente (CREs) wie Promotoren und Enhancer. Hinzu kommen deren räumliche Interaktionen mit anderen, entfernten CREs, sowie die lokale Anreicherung von TFs und der Transkriptionsmaschinerie in nukleären Subkompartimenten. Die Regulation der Transkription durch TFs wurde mithilfe von fluoreszenzmikroskopischen Experimenten intensiv untersucht. Im Gegensatz dazu werden Chromatinzustände und seine dreidimensionale Organisation üblicherweise durch Next-Generation-Sequencing (NGS) erforscht. Diese NGS-Methoden umfassen die Chromosomen-Konformations-Analyse, Chromatin-Immunopräzipitation und Methoden zu Messung der Chromatin-Zugänglichkeit. Dabei ermöglichen es Daten der Einzelzell-Chromatin-Zugänglich-keitssequenzierung (scATAC-seq), CREs in zugänglichen und aktiven Zuständen sowie die lokusspezifische TF-Bindungsaktivität auf Einzelzellebene zu analysieren. Jedoch fehlt bislang ein Modell der Transkriptionsregulation, das genomweite Daten zur lokalen Dynamik von TFs, zu Chromatinzuständen sowie seine Organisation integriert.

Diese Dissertation hatte das Ziel, ein solches Modell der Transkriptionsregulation zu entwickeln. Hierfür wurden lokale Chromatinzustände, lokusspezifische TF-Bindungsaktivität und die globale Chromatinorganisation aus scATAC-seq Daten abgeleitet. Ich verfolgte drei konkrete Ziele um dies zu erreichen: (i) Verbesserung der experimentellen und computergestützten Analyse von scATAC-seq, (ii) Entwicklung einer Methode zur Untersuchung der zugrundeliegenden molekularen Mechanismen von Chromatin-Ko-Zugänglichkeit, und (iii) Identifizierung der Struktur-Funktions-Beziehung zwischen verschiedenen regulatorischen Mechanismen und ihrem transkriptionellen Output.

Im ersten Teil dieser Dissertation habe ich die hohe Anzahl an Datenlücken als zentrales Problem bei der Analyse von scATAC-seq Daten identifiziert. Um diese zu beheben, habe ich das TurboATAC-Protokoll eingeführt. Es reduziert fehlende Daten durch die Optimierung der Transposase-Reaktionseffizienz von scATAC-seq. Zudem habe ich eine

Methode zur Allel-basierten Quantifizierung von scATAC-seq Daten entwickelt. Diese neuen Methoden haben mir auf der Ebene einzelner genomischer Loci ermöglicht, wahre biologische Variabilität zwischen Zellen von Datenlücken zu unterscheiden.
Im zweiten Teil habe ich das R-Paket RWireX entwickelt. Die Methode kann verschiedene Ebenen der Chromatin-Ko-Zugänglichkeit zwischen mehreren genomischen Loci auflösen. Dabei differenziert RWireX zwischen verschiedenen Merkmalen der Ko-Zugänglichkeit: Autonome Links der Ko-Zugänglichkeit (ACs) und Domänen kontinuierlicher Ko-Zugänglichkeit (DCs). ACs repräsentieren räumliche Kontakte zwischen ko-aktiven entfernten Chromatin-Loci. DCs entstehen durch die lokale Anreicherung von TF-Bindungsaktivität in nukleären Subkompartimenten. Darüber hinaus zeigte RWireX verschiedene Typen von ACs, die entweder durch gezielte strukturelle Chromatin-Schleifen oder durch zufällige räumliche Interaktionen innerhalb dynamischer Chromatinregionen entstehen.
Im dritten Teil habe ich die Transkription verschiedener humaner und muriner Systeme nach externen oder internen Stimuli mit RWireX analysiert. Dadurch verknüpfte ich genomweite Regulationsmechanismen mit ihrem funktionellen transkriptionellen Output. Diese Analysen ergaben, dass Promotoren, ACs und DCs unterschiedliche Kinetiken des transkriptionellen Burstings regulieren. Promotoren und DCs regulieren hauptsächlich die Größe eines Bursts und führen zu einer schnellen transkriptionellen Antwort. ACs hingegen regulieren primär die Frequenz von Bursts und zeigen langsamere Transkriptionsänderungen auf. Des Weiteren induzieren Promotoren eine signifikant stärkere Antwort im Vergleich zu ACs und DCs. Dahingegen können ACs und DCs jedoch mehrere Gene entweder durch Koexpression oder alternierende Expressionsmuster in Einzelzellen ko-regulieren.

Abschließend kombinierte ich diese Erkenntnisse, um das sogenannte AC/DC-Modell der Transkriptionsregulation herzuleiten. Dieses Modell verknüpft die promotorvermittelte Regulation, Chromatinkontakte (über gezielte Schleifen oder stochastische Interaktionen) und lokale Subkompartimente mit angereicherter TF-Bindungsaktivität mit ihren spezifischen transkriptionellen Effekten. Mit diesem Modell bietet diese Dissertation einen neuartigen Ansatz, um zu erklären, wie Säugetiersysteme die Stärke, Richtung und zeitliche Hierarchie von transkriptionellen Antworten auf externe und interne Stimuli präzise regulieren.

Acknowledgements

First and foremost, I would like to thank Prof. Karsten Rippe for the opportunity to pursue my PhD in his group. Thank you, Karsten, for your unwavering scientific support, infectious enthusiasm and inspiring ideas throughout this journey. I am truly grateful for the countless hours we spent discussing ATAC, transcription regulation, and AC/DC. Thank you for trusting me with increasingly complex data sets and for teaching me that 'any result is a good result'. Over the years, you have consistently encouraged me to step out of my comfort zone, helping me grow both professionally and personally.

I am grateful to Prof. Oliver Stegle, Prof. Carl Herrmann and Jun.-Prof. Simon Anders for their scientific support and valuable feedback throughout my PhD. Furthermore, I would like to thank Jun.-Prof. Lauren Saunders and apl. Prof. Stefan Wiemann for participating in my thesis defense.

I am thankful for the incredible wet lab support I received throughout the years of my PhD: Irene, Kathi, Markus, Philipp and Sabrina – without you, my PhD thesis would consist of nothing but blank pages. A special thanks to Simon and Nick for your support and maintenance of our Curry Cluster. Additionally, I would like to thank my external collaborators, Prof. Argyris Papantonis, Vassiliki Varamogianni-Mamatsi, Dr. Philipp Roessner and Dr. Martina Seiffert for the engaging discussions, valuable insights, and successful collaborations.

Thank you to all the current and former members of the Rippe lab: Afzal, Alexandra, Anne, Arjun, Armin, Caro, Chloe, Claire, Emma, Ezgi, Fabian, Fabio, Irene, Jorge, Lara, Linda, Lukas, Markus, Mislav, Nick, Norbert, Robin, Sabrina, Simon, Sina, Sofie, Stephan. Thanks to you, my PhD years have been filled with laughter – whether from good or bad (dad) jokes – endless personality tests, culinary adventures I would have never dared to try on my own, game nights with werewolves and exploding kittens, and so much more. I am also very grateful to the amazing students that have worked with me: Anastasiya, Ezgi, Franziska and Stefanie. Your support has helped me shape this thesis.

Beyond the science, I am deeply grateful to my family and friends. Mom and Dad, thank you for your unwavering love, support and encouragement over the past 28 (not yet 30!) years. You have always supported me without hesitation, no matter how sensible – or less sensible – my ideas may have been. Felix, thank you for the constant flow of book, game

and music recommendations, and for being the best little brother anyone could ask for. I want to thank my amazing in-law family for all their love, support, and lively distractions when I needed them most. Thank you, Talisa, for our evenings of ‘educational’ television that kept me going through the hardest times. Special thanks to my dog Nala for our long walks (in case of no rain, adequate temperatures, and a moderate pace) and the most wonderful non-verbal emotional support throughout my PhD. Lastly, and most importantly: Julian, thank you for everything! You are simply the best.

Table of Contents

List of Publications

In the course of this thesis I contributed to the following publications:

First author publications

Seufert I, Gerosa I, Varamogianni-Mamatsi V, Vladimirova A, Sen E, Mantz S, Rademacher A, Schumacher S, Liakopoulos P, Kolovos P, Anders S, Mallm JP, Papantonis A, Rippe K. Two distinct chromatin modules regulate proinflammatory gene expression. bioRxiv (2024).

Seufert I, Sant P, Bauer K, Syed AP, Rippe K, Mallm JP. Enhancing sensitivity and versatility of Tn5-based single cell omics. Frontiers in Epigenetics and Epigenomics **1**, 1245879 (2023).

Wu Y*, Seufert I*, Al-Shaheri FN*, Kurilov R, Bauer AS, Manoochehri M, Moskalev EA, Brors B, Tjaden C, Giese NA, Hackert T**, Buchler MW**, Hoheisel JD**. DNA-methylation signature accurately differentiates pancreatic cancer from chronic pancreatitis in tissue and plasma. Gut **72**, 2344-2353 (2023).
*Shared first co-authors
** Shared last co-authors

Co-author publications

Roessner PM, Seufert I, Chapaprieta V, Jayabalan R, Briesch H, Massoni-Badosa R, Boskovic P, Beckendorff J, Roider T, Arseni L, Coelho M, Chakraborty S, Vaca A, Sivina M, Muckenhuber M, Rodriguez-Rodriguez S, Bonato A, Herbst SA, Zapatka M, Sun C, Kretzmer H, Naake T, Bruch PM, Czernilofsky F, Ten Hacken E, Schneider M, Helm D, Yosifov DY, Kauer J, Danilov AV, Bewarder M, Heyne K, Schneider C, Stilgenbauer S, Wiestner A, Mallm JP, Burger JA, Efremov DG, Lichter P, Dietrich S, Martin-Subero JI, Rippe K, Seiffert M. T-bet suppresses proliferation of malignant B cells in chronic lymphocytic leukemia. Blood **144**, 510-524 (2024).

Muckenhuber M, Seufert I, Müller-Ott K, Mallm JP, Klett LC, Knotz C, Hechler J, Kepper N, Erdel F, Rippe K. Epigenetic signals that direct cell type-specific interferon beta response in mouse cells. Life Science Alliance **6**, e202201823 (2023).

Poos AM*, Prokoph N*, Przybilla MJ*, Mallm JP, Steiger S, Seufert I, John L, Tirier SM, Bauer K, Baumann A, Rohleder J, Munawar U, Rasche L, Kortum K M, Giesen N, Reichert P, Huhn S, Müller-Tidow C, Goldschmidt H, Stegle O, Raab MS**, Rippe K**, Weinhold N**. Resolving therapy resistance mechanisms in multiple myeloma by multiomics subclone analysis. Blood **142**, 1633-1646 (2023).
*Shared first co-authors
** Shared last co-authors

Schuster LC, Syed AP, Tirier SM, Steiger S, Seufert I, Becker H, Duque-Afonso J, Ma T, Ogawa S, Mallm JP, Lubbert M, Rippe K. Progenitor like cell type of an MLL-EDC4 fusion in acute myeloid leukemia. Blood Advances **7**, 7079-7083 (2023).

Bundschuh C, Weidner N, Klein J, Rausch T, Azevedo N, Telzerow A, Mallm JP, Kim H, Steiger S, Seufert I, Borner K, Bauer K, Hubschmann D, Jost KL, Parthe S, Schnitzler P, Boutros M, Rippe K, Müller B, Bartenschlager R, Kräusslich HG, Benes V. Evolution of SARS-CoV-2 in the Rhine-Neckar/Heidelberg Region 01/2021 - 07/2023. Infection, Genetics and Evolution, 105577 (2024).

Groth C, Maric J, Garcés Lázaro I, Hofman T, Zhang Z, Ni Y, Keller F, Seufert I, Hofmann M, Neumann-Häfelin C, Sticht C, Rippe K, Urban S, Cerwenka A. Hepatitis D infection induces IFN-β-mediated NK cell activation and TRAIL-dependent cytotoxicity. Frontiers in Immunology **14**, 1287367 (2023).

Abbreviations

3C	Chromosome conformation capture
AC	Autonomous link of co-accessibility
ac	Acetylation
ACC-seq	Assay for chromatin-associated condensate sequencing
AML	Acute myeloid leukemia
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
ATF	Activating transcription factor
BATF	Basic leucine zipper ATF-like transcription factor
BMNC	Bone marrow mononuclear cell
BORIS	Brother of Regulator of Imprinted Sites
bp	Base pairs
CCA	Canonical correlation analysis
CDK	Cyclin-dependent kinase
CEBP	CCAAT enhancer binding proteins
ChIP-seq	Chromatin immunoprecipitation sequencing
CLL	Chronic lymphocytic leukemia
CpG	Cytosine-phosphate-Guanine
CRE	*Cis*-regulatory element
CTCF	CCCTC-binding factor
CUT&Tag-seq	Cleavage under targets and tagmentation sequencing
DBD	DNA-binding domain
DC	Domain of contiguous co-accessibility
DNase HS-seq	DNase hypersensitivity sequencing
ED	Effector domain
EDC4	Enhancer of the messenger RNA decapping 4

ESC	Embryonic stem cell
FAIRE-seq	Sequencing of formaldehyde-assisted isolation of regulatory elements
FDR	False discovery rate
FRIP	Fraction of reads in peaks
GEO	Gene Expression Omnibus
GFP	Green fluorescent protein
GRO-seq	Global run-on sequencing
H3	Histone 3
HC	Healthy control
HDV	Hepatitis D virus
HUVEC	Human umbilical vein endothelial cell
IDR	Intrinsically disordered region
IFN	Interferon
IFNAR	Interferon alpha receptor
IgG	Immunoglobulin G
IKK	IκB kinase
IRF	Interferon-regulatory factor
ISG	Interferon-stimulated gene
ISGF3	Interferon-stimulated gene factor 3
ISRE	Interferon response element
IκBα	Inhibitor of nuclear factor kappa B
JAK1	Janus kinase 1
K	Lysine
kb	Kilo base pairs
K_d	Dissociation konstant
KNN	K-nearest neighbor
k_{off}	Off-rate
k_{on}	On-rate

k_{syn}	Synthesis rate
LncRNA	Long non-coding RNA
Log2FC	Log2 fold change
LSI	Latent semantic indexing
M-CLL	CLL patients with mutated *IGHV* genes
MAPK	Mitogen-activated protein kinase
Mb	Mega base pairs
ME	Mosaic end
me1	Monomethylation
me3	Trimethylation
MEF	Mouse embryonic fibroblast
MLL	Mixed lineage leukemia
MM	Multiple myeloma
MS	Mass spectrometry
NA	Not assigned
NF-κB	Nuclear factor kappa B
NGS	Next-generation sequencing
NK cell	Natural killer cell
OCT	Octamer-binding transcription factor
ORCA	Optical reconstruction of chromatin architecture
PAC	Percent accessible cells
PBMC	Peripheral blood mononuclear cells
PCA	Principle component analysis
Perturb-ATAC	Single-cell ATAC sequencing on pooled genetic perturbation screens
Perturb-seq	Single-cell RNA sequencing on pooled genetic perturbation screens
phos-ME	Phosphate-5'-methyl ether
POU	Pituitary-specific, Octamer-binding, and Unc-86 transcription factors

PRDM1	PR domain zinc finger protein 1
PTM	Post-translational modification
PTO	Phosphorothioate
PU1	Purine-rich box 1
RIP	Death domain-containing Ser/Thr kinase receptor-interacting protein
RNA-seq	RNA sequencing
RNAP II	RNA polymerase II
ROI	Region of interest
RSE	Residual standard error
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
sc	Single-cell
sci	Single-cell combinatorial indexing
SMART-seq	Switching Mechanism at 5′ end of RNA Template sequencing
smFISH	Single-molecule fluorescence *in situ* hybridization
sn	Single-nucleus
SNN	Shared nearest neighbor
SPT	Single-particle tracking
STAT	Signal transducer and activator of transcription
SVD	Singular value decomposition
TAD	Topologically associating domain
TBX21 T-bet	T-box transcription factor 21 T-box expressed in T cells
TCL1	T-cell leukemia-1 oncogene
TF	Transcription factor
Tn5	Transposase 5
Tn5-H	In-house Tn5
Tn5-ILMN	Illumina TDE1 enzyme
Tn5-TXG	10x Genomics 10x
TNFR	TNF receptor

TNFα	Tumor necrosis factor alpha
TPM	Transcripts per kilobase million
TRADD	TNF receptor 1-associated death domain protein
TRAF2	TNF receptor-associated factor 2
TRG	TNFα-responsive gene
TSS	Transcription start site
TXGv2	10x Genomics scATAC-seq v2 protocol
TYK2	Tyrosine kinase 2
U-CLL	CLL patients with unmutated *IGHV* genes
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier

List of Figures

1. Introduction

2. Results

3. Discussion

List of Tables

1. Introduction

2. Results

3. Discussion

4. Materials and Methods

1. Introduction

1.1. Eukaryotic transcription and transcription regulation

1.1.1. Molecular mechanisms and kinetics of transcription

Establishing diverse gene expression programs with well-defined timing and transcriptional output of thousands of genes is a key feature of eukaryotic cells. These gene expression programs are often initiated by environmental or cellular stimuli and their underlying regulatory mechanisms are multilayered and complex. In eukaryotic cells, transcription is inherently stochastic, with genes alternating independently between active transcription and inactive states (Rodriguez & Larson, 2020). This process is commonly described by the two-state model of transcriptional bursting (**Figure 1.1A**), where a gene switches between transcriptionally active and inactive states based on its specific but variable on and off rates k_{on} and k_{off} (Ko, 1991). The so-called transcriptional bursts occur during the active state, defined by a certain synthesis rate k_{syn} (Ko, 1991) and are often characterized by their frequency and size (Raj *et al.*, 2006; Suter *et al.*, 2011; Mahat *et al.*, 2024). Burst size reflects the number of RNA molecules produced during a single burst and derived from k_{syn}/k_{off} (Ko, 1991). This is mainly determined by the number of RNAP II molecules that are recruited to the TSS and subsequently synthesize RNA (Bartman *et al.*, 2019). In contrast, burst frequency represents the on rate with which a gene switches from the inactive into the active state (Ko, 1991). It is governed by the promoter's state and its affinity for initiating transcription (Bartman *et al.*, 2019). Thus, by regulating transcription burst frequency and size, both the overall kinetics as well as the strength of transcription at a specific gene can be controlled.

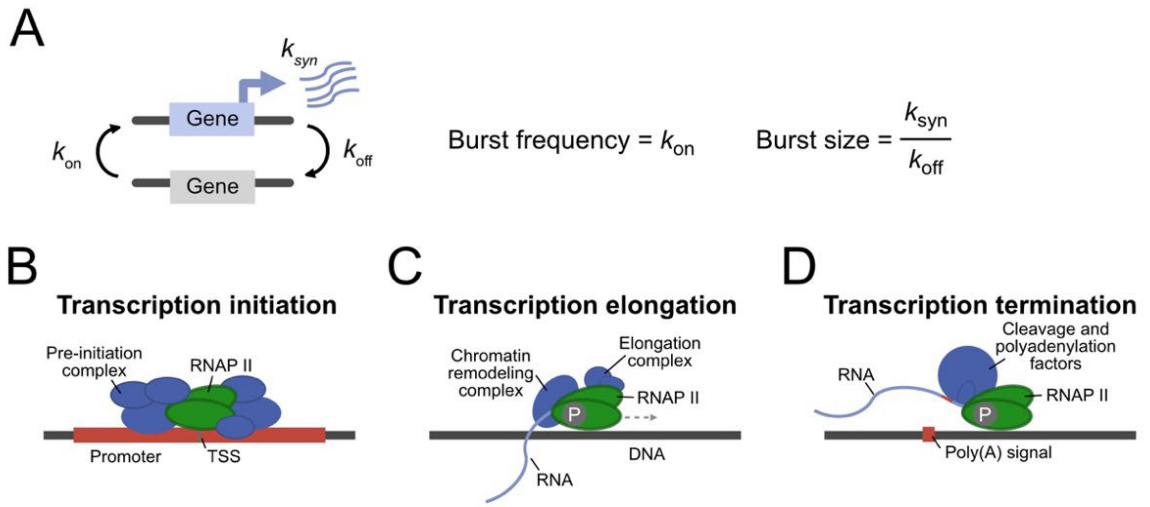


Figure 1.1 Eukaryotic transcription. A Two-state model of transcriptional bursting. Genes switch between their active and inactive states by on-rate (k_{on}) and off-rate (k_{off}). In the active state, RNA is transcribed at synthesis rate (k_{syn}). Adapted from Seufert *et al.* (2024). **B** Transcription initiation: The pre-initiation complex recruits RNA polymerase II (RNAP II) to the transcription start site (TSS) and unwinds the DNA, allowing RNAP II to start transcription. After Haberle & Stark (2018). **C** Transcription elongation: RNAP II dissociates from the promoter and synthesizes RNA. Elongation factors and chromatin remodeling complexes facilitate the movement along the gene. After Wang *et al.* (2023). **D** Transcription termination and release: RNAP II mostly transcribes beyond the poly(A) signal. The signal is recognized by RNA cleavage and polyadenylation factors, which cut the RNA and release both RNA and RNAP II from the DNA. After Porrua & Libri (2015).

Transcription is initiated through a multi-step process in which multiple transcription factors (TFs) assemble into the pre-initiation complex (**Figure 1.1B**) (Malik & Roeder, 2023). This complex promotes the recruitment of RNA polymerase II (RNAP II) to the promoter and DNA unwinding at the transcription start site (TSS), allowing RNAP II to start RNA synthesis (Haberle & Stark, 2018). RNAP II then dissociates from the promoter-bound TFs and moves along the gene, entering the elongation phase (**Figure 1.1C**) (Wang *et al.*, 2023). During elongation, RNAP II synthesizes RNA, aided by binding of multiple elongation factors (Kwak & Lis, 2013). Additionally, histone chaperones and chromatin remodeling complexes facilitate the advancement through the chromatin chain (Gamarra & Narlikar, 2021). For most protein-coding genes, RNAP II transcribes beyond the poly(A) signal, which is recognized by RNA cleavage and polyadenylation factors (Proudfoot, 2016). These factors cleave the RNA, releasing both the RNA molecule and RNAP II, thereby terminating transcription (**Figure 1.1D**) (Porrua & Libri, 2015). Taken together, the transcription of RNA is a highly complex process with a multitude of different molecules involved. Consequently, the two key parameters to describe the transcriptional output, burst size and frequency, are controlled by a complex network of molecular mechanisms and their precisely orchestrated spatial and temporal organization (Fukaya *et al.*, 2016; Cramer, 2019). However, the exact molecular mechanisms regulating burst size and frequency, respectively, remain unclear (Wang *et al.*, 2019).

The following sections will introduce and discuss two different perspectives on transcription regulation and their commonly used experimental techniques (**Table 1.1**). On the one hand, the chromatin topology-centric viewpoint focuses on both local chromatin states and global chromatin organization. The field commonly applies next-generation sequencing methods to investigate genome-wide chromatin topology in the genomic coordinate system. On the other hand, the TF-centric viewpoint revolves around soluble TFs in the nucleus, their distribution, movement and DNA binding kinetics. Transcription regulation by TFs is commonly studied using fluorescence microscopy approaches for specific reporters or genes of interest in the spatial nuclear coordinate system.

Table 1.1 Chromatin topology-centric and TF-centric perspectives on transcription regulation.

	Chromatin topology-centric	**TF-centric**
Perspective	Local chromatin states, global chromatin organization, and its interplay with soluble proteins and RNAs	Soluble TFs, their nuclear distribution, and DNA binding dynamics
Coordinate system	Genomic coordinate	Spatial nuclear coordinate
Experiments	Next-generation sequencing	Fluorescence microscopy
Generalizability	Genome-wide	Reporters / genes of interest

1.1.2. Chromatin topology-centric transcription regulation

Transcription is regulated by *cis*-regulatory elements (CREs) throughout the genome, their associated active or inactive local chromatin states, and the three-dimensional organization of the nucleosome chain (Lim *et al.*, 2010; Grosveld *et al.*, 2021; Wang *et al.*, 2021). These mechanisms are not independent of proteins, as chromatin is defined as the assembly of DNA around nucleosomes, which consist of eight histone proteins each (Hubner *et al.*, 2013). This so-called “beads on a string” chain of nucleosomes organizes into higher-order chromatin structures in the nucleus (Hubner *et al.*, 2013). In this context, loosely packed chromatin is termed euchromatin, leaving the DNA accessible to TFs, the transcription machinery, and other co-factors. In contrast, densely packed chromatin is referred to as heterochromatin, in which chromatin is in a silenced state and TF binding sites are less accessible (Morrison & Thakur, 2021). Although many proteins are involved

in the formation and organization of chromatin, this section will focus on CREs and their molecular mechanisms in regulating transcription, thereby adopting this chromatin topology-centric viewpoint on transcription regulation.

The regulatory active DNA regions, CREs, are evolutionary conserved, non-coding genomic regions that contain multiple binding sites for TFs (Kim & Wysocka, 2023). They can be located both close to and far from TSSs (Kim & Wysocka, 2023). TSS-proximal CREs are commonly called promoters (Juven-Gershon & Kadonaga, 2010), while TSS-distal CREs are variously referred to by their function, such as enhancers, silencers, or boundary control elements (Bulger & Groudine, 2011). To simplify, in this thesis, I only distinguish between promoters and *distal CREs*. While promoters regulate transcription by directly recruiting TFs and the transcription machinery to the TSS (Juven-Gershon & Kadonaga, 2010), *distal CREs* need to interact with promoters and TSSs to exert their regulatory effects (Grosveld *et al.*, 2021). Furthermore, transcription co-regulation of multiple distal genes has been observed via the formation of so-called transcription factories or active chromatin hubs, where multiple gene loci, *distal CREs*, RNAP II molecules and co-factors interact and actively promote simultaneous transcription of multiple genes (de Laat & Grosveld, 2003; Papantonis & Cook, 2013). Several models have been proposed to explain how the interaction between distal genomic sites occurs, including (i) RNAP II tracking, (ii) TF linking, (iii) relocation into spatial proximity forming a transcription factory or communicating by diffusion, (iv) transient contact to deposit TFs, and (v) stable contact via loop formation (**Figure 1.2**) (Grosveld *et al.*, 2021; Karr *et al.*, 2022). However, it remains unclear which exact molecular mechanisms, or a combination thereof, are present in eukaryotic nuclei (Ibrahim, 2024). Additionally, it is uncertain whether these interactions require direct chromatin contact or if spatial proximity between the two sites is sufficient (Friedman *et al.*, 2024).

The activity of CREs, meaning their potential to exert their regulatory effects on transcription, is largely influenced by their local chromatin state (Lim *et al.*, 2010). In this context, DNA methylation at the nucleic acid cytosine and various post-translational modifications (PTMs) on the N-terminal tail of histone 3 (H3) in nucleosomes help define different chromatin states (Jenuwein & Allis, 2001; Greenberg & Bourc'his, 2019). The most prominent PTMs include mono- or trimethylation (me1 and me3) and acetylation (ac) at lysine residues, namely lysine 4 (K4), lysine 9 (K9), lysine 27 (K27), and lysine 36 (K36) on the H3 tails (**Figure 1.3A**) (Lim *et al.*, 2010). These PTMs influence the direct recruitment of proteins to specific chromatin regions and determine the local chromatin state, making it either more or less accessible to TFs and the transcription machinery

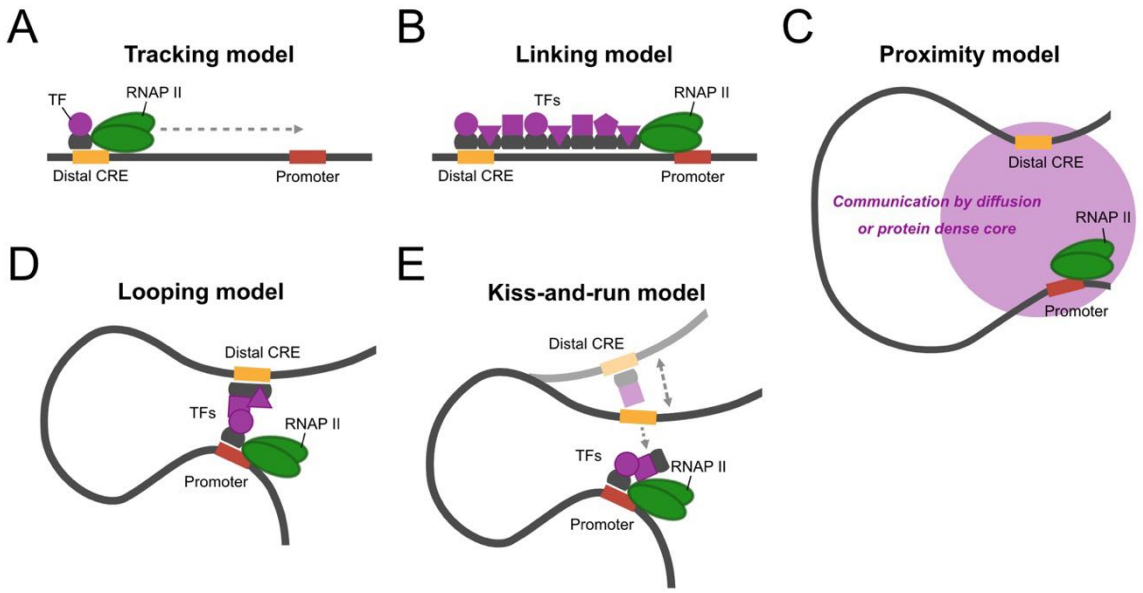


Figure 1.2 Interaction models between promoters and *distal CREs*. A Tracking model: RNAP II starts transcription at the *distal CRE* either pulling the *distal CRE* along or leaving it behind, while moving toward the promoter. **B** Linking model: Various TFs oligomerize between the *distal CRE* and promoter, initiating transcription at the promoter. **C** Proximity model: The *distal CRE* and promoter relocate into spatial proximity. Promoter activation may occur through a dense protein core between the sites or TF activation at the *distal CRE*, followed by diffusion to the promoter. **D** Looping model: A chromatin loop brings the *distal CRE* and promoter into stable spatial contact, enabling complex formation of TFs, co-factors and the transcription machinery. **E** Kiss-and-run model: Transient chromatin contacts transfer TFs from the *distal CRE* to the promoter. After Grosveld *et al.* (2021) and Karr *et al.* (2022).

(**Figure 1.3B**) (Kouzarides, 2007). The complex combinations of PTMs result in specific chromatin activity patterns: At promoters, H3K9ac, H3K4me3 and H3K27ac are strong indicators of activation (Santos-Rosa *et al.*, 2002; Ernst *et al.*, 2011; Karmodiya *et al.*, 2012), while the gene bodies of active genes are marked by H3K36me3 (**Figure 1.3A**, right) (Wagner & Carpenter, 2012). In contrast, repressed promoters often display H3K27me3, which, along with H3K9me3, is also present at the gene bodies of silenced genes (Morey & Helin, 2010). Similar patterns are observed for *distal CREs*: Active *distal CREs* display H3K4me1 and H3K27ac (Creyghton *et al.*, 2010; Kang *et al.*, 2021), whereas repressive marks like H3K9me3 and H3K27me3 predominate at inactive *distal CREs* (**Figure 1.3A**, left) (Morey & Helin, 2010). Within the nucleus, a multitude of so-called writers, readers, and erasers constantly modify these chromatin states, leading to a highly dynamic chromatin landscape (Gourisankar *et al.*, 2024). Together with many other potential PTMs, these epigenetic modifications regulate the local chromatin state, its accessibility to proteins, and consequently the activity of CREs (Bernstein *et al.*, 2005).

In addition to the local chromatin states at CREs, their activity and ability to regulate transcription are influenced by the three-dimensional, higher-order chromatin organization (**Figure 1.3C+D**) (Uyehara & Apostolou, 2023). During interphase, chromosomes occupy

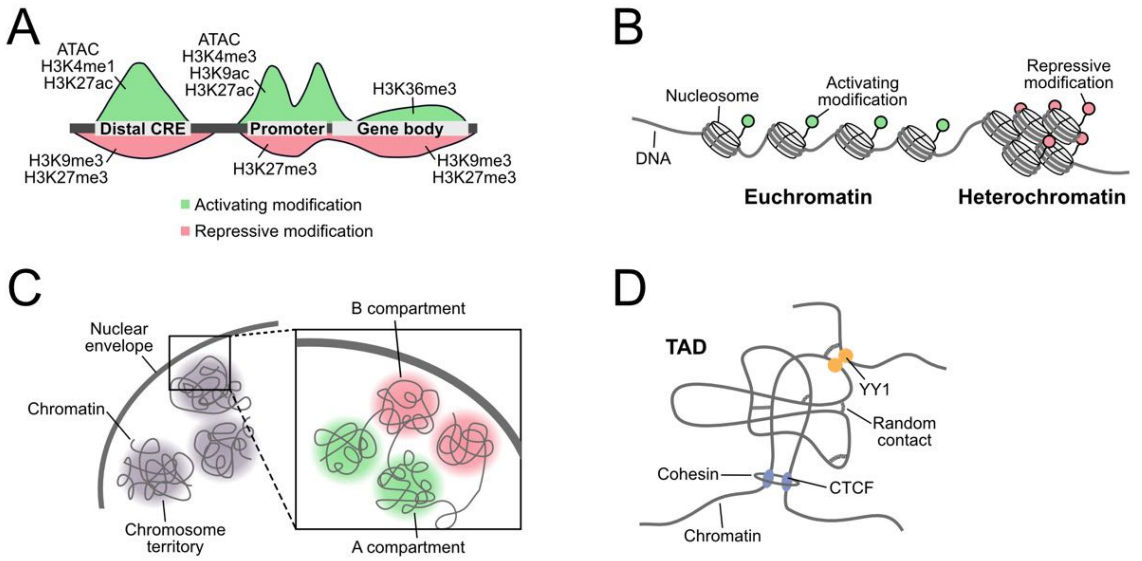


Figure 1.3 Chromatin state, organization and three-dimensional architecture. A Epigenetic histone PTMs determine local chromatin state. Activating (green) and repressive (red) histone PTMs at promoters, gene bodies, and *distal CREs* are shown. After Lim *et al.* (2010). **B** Local organization of the chromatin fibre into euchromatic and heterochromatic regions. **C** Nuclear chromatin organization into chromosome territories and A/B compartments. **D** Higher-order chromatin conformation, where targeted loops between distal chromatin sites determine topologically associating domains (TADs) with random contacts in a dynamic chromatin environment. B-D after Wang *et al.* (2021).

distinct territories within the nucleus, where the chromatin of each chromosome is highly intermingled (**Figure 1.3C**) (Hubner *et al.*, 2013). Within these chromosome territories, chromatin is organized into A and B compartments (Wang *et al.*, 2021). The A compartments, which are euchromatic and generally active, are typically located in the nuclear interior. In contrast, B compartments are largely composed of heterochromatin and found near the nuclear envelope (Wang *et al.*, 2021). In euchromatic A compartments, distal chromatin regions can contact through targeted chromatin loop formation or random interactions along the dynamic chromatin chain (**Figure 1.3D**) (Hubner *et al.*, 2013; Sood & Misteli, 2022; Bruckner *et al.*, 2023; Uyehara & Apostolou, 2023). Here, the proteins CTCF and cohesin often form targeted structural loops that enhance the likelihood of dynamic spatial contacts in between the stably linked chromatin regions (Mach *et al.*, 2022; Chan & Rubinstein, 2023), resulting in stochastically interacting domains, termed topologically associating domains (TADs) (Dixon *et al.*, 2012; Hansen *et al.*, 2018). Moreover, the mediator complex and specific TFs, such as YY1 or NANOG, can mediate targeted chromatin contacts (Weintraub *et al.*, 2017; Choi *et al.*, 2022; Ramasamy *et al.*, 2023). Thus, both targeted and random chromatin contacts shape higher-order chromatin organization, intricately regulating transcription by bringing multiple, otherwise distal CREs into close spatial proximity (Sood & Misteli, 2022; Uyehara & Apostolou, 2023).

1.1.3. TF-centric transcription regulation

As discussed previously, transcription initiation, elongation and termination are regulated by various soluble proteins, such as TFs, RNAP II and co-activators (Porrua & Libri, 2015; Haberle & Stark, 2018; Wang *et al.*, 2023). The regulation of transcription by these factors, along with other TFs not directly involved in transcription, depends on chromatin, as their DNA-binding affinity and nuclear distribution are strongly influenced by the local chromatin state and its higher-order organization (Liu *et al.*, 2015; Xin & Rohs, 2018). In this section, I will introduce a TF-centric perspective on transcription regulation, primarily focusing on the nuclear activity and distribution of TFs and their implications for transcription regulation.

TFs are commonly present at low concentrations in the nucleus (Ferrie *et al.*, 2022). They are composed of different protein domains, typically categorized into DNA-binding domains (DBDs) and effector domains (EDs) (Trojanowski & Rippe, 2022). The EDs regulate transcription through various mechanisms, such as modifying local chromatin state by writing or erasing histone PTMs, recruiting other TFs or co-factors through specific interactions, or forming weak and nonspecific multivalent interactions with other proteins via their intrinsically disordered regions (IDRs) (Garcia *et al.*, 2021; Rippe & Papantonis, 2022; Soto *et al.*, 2022). To target the ED's effects to specific CREs, DBDs recognize and bind TF-specific DNA motifs that are only 6-12 base pairs (bp) in length (Stormo, 2013). In this context, TFs bind to DNA with specific kinetic on and off rates k_{on} and k_{off} with the residence time τ_{res} in the bound state derived from $1/k_{off}$ (**Figure 1.4A**) (Trojanowski & Rippe, 2022). The affinity of a TF to a genomic site containing its motif, referred to as a TF binding site, is defined by its binding site-specific equilibrium dissociation constant K_d from k_{off}/k_{on} (Trojanowski & Rippe, 2022). A high degree of motif overlap with the DNA sequence corresponds to a high-affinity binding site with low K_d. Together with the overall TF concentration in the nucleus, these constants determine the overall TF occupancy at a binding site and thus the TFs impact on transcription regulation (Lu & Lionnet, 2021; Popp *et al.*, 2021).

The generally low nuclear concentrations of TFs raise the question how they efficiently locate their target binding sites within the large eukaryotic genomes and nuclei (Jana *et al.*, 2021; Mazzocca *et al.*, 2021). Fundamentally, TFs move through the nucleus by diffusion, following a so-called "random walk" in three-dimensional space (**Figure 1.4B**, left) (Woringer & Darzacq, 2018). Here, the likelihood of finding a potential binding site may be increased by facilitated diffusion, either along the one-dimensional chromatin fibre

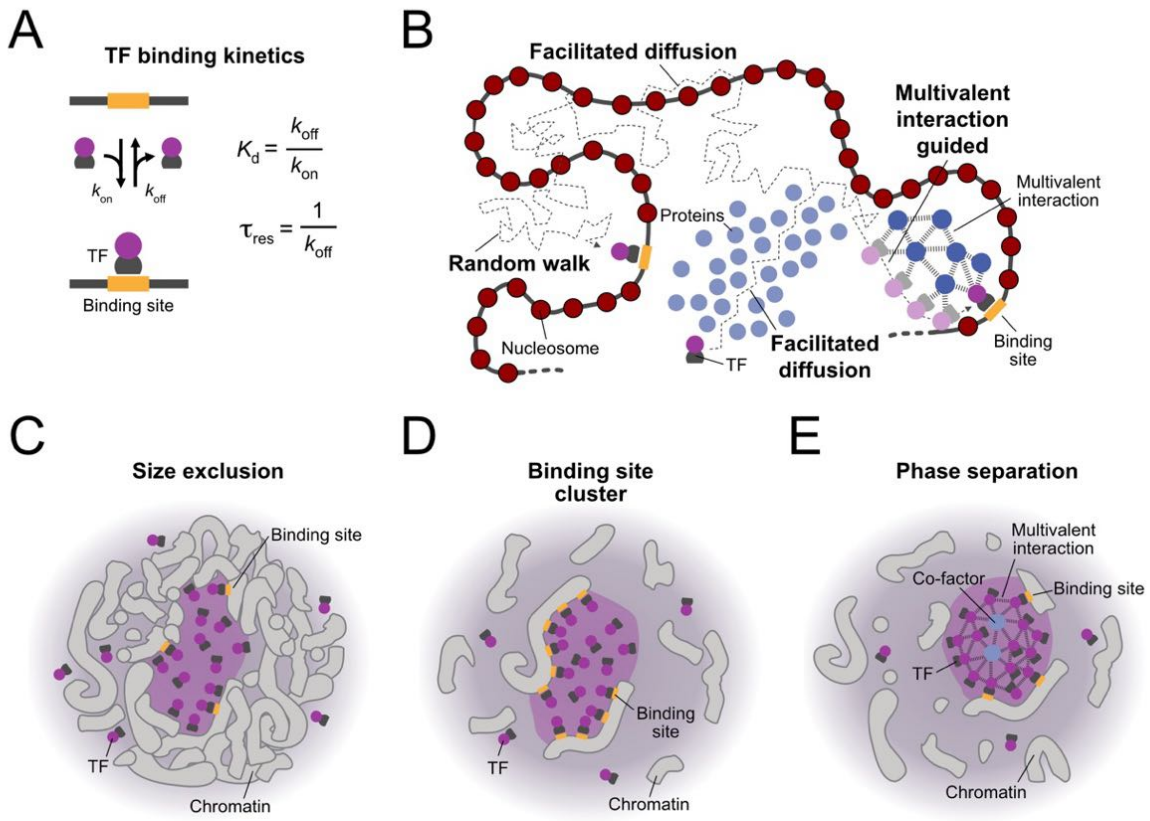


Figure 1.4 Principles of TF-mediated transcription regulation. A TF binding kinetics at genomic sites with TF-specific binding motifs. TFs bind to and dissociate from their binding site with a specific on- and off-rate (k_{on}, k_{off}), determining an individual equilibrium dissociation constant (K_d) for each binding site. τ_{res} describes the residence time of the TF at the binding site. **B** Models of TF movement through the nucleus. TFs move via diffusion in a random walk. The diffusion can be facilitated along the chromatin fibre or by molecular crowding. Multivalent interactions via IDRs might guide diffusion. **C** Local TF enrichment by size exclusion from densely packed surrounding chromatin. **D** Local TF enrichment by simultaneous binding of TFs to a binding site cluster. **E** Local TF enrichment by phase separation into liquid droplets. Adapted from Trojanowski & Rippe (2022).

or as directed by macromolecular crowding (**Figure 1.4B**, middle) (Berg & von Hippel, 1985; Esadze & Stivers, 2018). Another mechanism that might reduce TF search times involves the previously mentioned IDRs in TF effector domains (Trojanowski & Rippe, 2022). These IDRs can form nonspecific multivalent interactions with other locally enriched chromatin factors, guiding the TF toward its specific DNA motif (**Figure 1.4B**, right) (Brodsky *et al.*, 2020; Chen *et al.*, 2022). Furthermore, recent studies have observed nuclear subcompartments with increased concentrations of RNAP II, TFs, and co-factors (Rippe & Papantonis, 2021). Various models have been proposed to explain the formation of these nuclear subcompartments (Rippe, 2022): (i) Size exclusion from densely packed chromatin may confine TFs to specific, less occupied regions of the nucleus (**Figure 1.4C**) (Mazzocca *et al.*, 2023); (ii) local clustering of multiple binding sites for the same TF could increase its concentration through simultaneous binding and high residence times (**Figure 1.4D**) (Li *et al.*, 2020); and (iii) physicochemical phase separation, driven by multivalent

interactions of IDRs in TFs may lead to the formation of liquid droplets above critical concentrations (**Figure 1.4E**) (Hnisz *et al.*, 2017). Regardless of the underlying mechanism, the confinement and local increase in TF concentration reduce the TF search time for its binding sites and increase binding site occupancy within these subcompartments (Kent *et al.*, 2020; Garcia *et al.*, 2021). However, their overall effects on transcription regulation beyond this remain controversially discussed (Wei *et al.*, 2020; Chong *et al.*, 2022; Trojanowski *et al.*, 2022; Meeussen *et al.*, 2023).

1.2. Methods to study transcription regulation

While the nuclear distribution of molecules and TF kinetics are extensively studied using fluorescence microscopy-based assays (Hwang *et al.*, 2024), local chromatin state and global chromatin organization are commonly investigated with chromosome conformation capture or chromatin immunoprecipitation sequencing methods (**Table 1.1**) (Preissl *et al.*, 2023; van Mierlo *et al.*, 2023). In the following, I provide a brief overview of various microscopy- and sequencing-based methods used to study transcription and its regulation. I then focus on single-cell sequencing data of chromatin accessibility, which I employ to investigate chromatin topology-centric and TF-centric transcription regulation collectively.

1.2.1. Fluorescence microscopy

Fluorescence microscopy techniques can resolve nuclear substructures like chromosome territories and nucleoli (Nunes & Moretti, 2017). Additionally, they enable to study the spatial distribution and temporal dynamics of transcripts, as well as various regulatory components (Hwang *et al.*, 2024). For instance, real-time imaging of single RNA molecules, achieved by labeling RNAs of interest using synthetic RNA aptamers, provides both spatial resolution of nascent RNAs and direct observation of transcriptional kinetics (Bouhedda *et al.*, 2017). Conversely, single-molecule fluorescence *in situ* hybridization (smFISH) techniques applied to fixed cells or tissues lack temporal resolution but allow multiplexed detection of the spatial distribution of numerous RNAs (Young *et al.*, 2020). To study transcription regulation, methods such as single-particle tracking (SPT) have been employed to observe real-time kinetics, spatial movement, and local enrichment of multiple nuclear proteins (Dahal *et al.*, 2023). Additionally, SPT facilitates the investigation of target-search times and binding kinetics of individual TF molecules (Hwang *et al.*, 2024). At the chromatin level, immunofluorescence staining of specific histone PTMs reveals chromatin compartmentalization and the dynamics of chromatin remodelers in the

deposition or removal of PTMs (Hayashi-Takanaka *et al.*, 2009). Furthermore, advanced computational approaches like optical reconstruction of chromatin architecture (ORCA), when combined with RNA labeling, can reconstruct 3D chromatin organization and transcription in single cells (Mateo *et al.*, 2019).

However, despite the broad range of information these microscopy techniques provide, including nuclear distribution and temporal kinetics in transcription regulation, they are often limited by imaging duration, resolution, and noise, as well as the challenge of simultaneously observing multiple regulatory factors (Hwang *et al.*, 2024). Moreover, imaging data inherently lack information about genomic coordinates, restricting their insights to general nuclear observations or locus-specific findings, usually limited to single reporters or genes (van Mierlo *et al.*, 2023).

1.2.2. Next-generation sequencing

Next-generation sequencing (NGS) provides reliable, fast, and cost-efficient sequence information of DNA molecules at high depths (Pettersson *et al.*, 2009). Prior to the sequencing reaction, sequencing libraries are generated by fragmenting the DNA and attaching primers for sequencing and indices for sample identification in multiplexed sequencing runs to each DNA molecule (Hess *et al.*, 2020). These libraries are amplified by PCR, which enhances detection sensitivity but impairs quantitative analysis of the results (Kivioja *et al.*, 2011). Typically, NGS does not sequence DNA fragments at full length (Pettersson *et al.*, 2009). Instead, only 50 to 200 bp from one or both ends of the DNA molecule are sequenced, referred to as single-end and paired-end sequencing, respectively (Pettersson *et al.*, 2009). The obtained sequence information is stored in one or two paired sequencing reads for single-end and paired-end sequencing, respectively, which are then computationally mapped to a reference genome to obtain their genomic coordinates (Schbath *et al.*, 2012). By applying NGS to specific subsets of DNA molecules, a suite of complementary techniques has been developed to study transcription, local chromatin state, and global chromatin organization in the context of the genomic coordinate system (van Mierlo *et al.*, 2023).

Among this toolbox to study transcription regulation, *in situ* cross-linking combined with chromosome conformation capture (3C) methods map chromatin contacts (Dekker *et al.*, 2002; de Wit & de Laat, 2012). The further development of HiC-seq enabled the genome-wide detection of chromatin contacts, facilitating the study of genome topology and higher-order chromatin organization, such as A and B compartments or TADs (Lin *et al.*, 2018;

van Mierlo *et al.*, 2023). Beyond global chromatin organization, epigenomic assays, such as sequencing after bisulfite conversion, chromatin immunoprecipitation sequencing (ChIP-seq), and cleavage under targets and tagmentation sequencing (CUT&Tag-seq), can map the genomic positions of DNA methylation and various histone PTMs (Fraga & Esteller, 2002; Barski *et al.*, 2007; Kaya-Okur *et al.*, 2019). Additionally, ChIP-seq and CUT&Tag-seq are used to profile genomic sites of TF binding events (van Mierlo *et al.*, 2023). Collectively, these epigenomic techniques are employed to define CREs by describing local chromatin states. In parallel, genomic regions in generally active chromatin states can be identified using assays for accessible chromatin, such as DNase hypersensitivity sequencing (DNase HS-seq) or assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Song & Crawford, 2010; Buenrostro *et al.*, 2013). Lastly, RNA sequencing (RNA-seq) methods enable genome-wide, quantitative investigation of the entire transcriptome by converting RNA molecules into complementary DNA and utilizing unique molecular identifiers (UMIs) (Wang *et al.*, 2009; Kivioja *et al.*, 2011).

While these methods provide genome-wide snapshots of transcription and its regulation, they lack spatial and temporal resolution (van Mierlo *et al.*, 2023). Furthermore, these sequencing approaches typically analyze thousands of cells simultaneously, in so-called bulk sequencing, which results in the loss of information regarding single-cell heterogeneity and stochasticity in transcription regulation (Preissl *et al.*, 2023). To overcome this, single-cell sequencing techniques have been developed to resolve cell-to-cell differences. During sequencing library preparation, DNA molecules from individual cells are barcoded differently by isolating cells into separate reaction chambers (e.g., wells, tubes, or droplets) or through combinatorial indexing in iterative split-pool approaches (Hess *et al.*, 2020; Preissl *et al.*, 2023). In addition, simultaneous profiling of RNA and chromatin topology (including accessibility, epigenomic and 3C assays) from the same cell has been developed, allowing for genome-wide investigation of transcription and its immediate regulatory features (Baysoy *et al.*, 2023). However, single-cell sequencing data tend to be sparse, making it difficult to differentiate between technical noise and biological variation among cells (Preissl *et al.*, 2023). Furthermore, the integration of information on transcription, local chromatin state and global chromatin organization from these various complementary experiments at single-cell resolution is costly, time consuming, and computationally challenging.

1.2.3. Single-cell sequencing of chromatin accessibility

Different sequencing assays resolve accessible chromatin regions, such as DNase HS-seq, ATAC-seq, or sequencing of formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) (Song & Crawford, 2010; Simon *et al.*, 2012; Buenrostro *et al.*, 2013). In contrast to the other complex and multi-step protocols, ATAC-seq is a simple two-step process of adapter insertion by hyperactive transposase 5 (Tn5) and PCR amplification (Buenrostro *et al.*, 2013). During ATAC library preparation, the Tn5 binds to regions of accessible chromatin, cuts the DNA, and inserts sequencing adapters (**Figure 1.5A**). Following paired-end sequencing of the ATAC library, the genomic positions of these so-called Tn5 insertions can be inferred from the individual sequencing reads. Additionally, the DNA fragments between two paired reads provide information about nucleosome positioning, as the fragment sizes indicate the number of nucleosomes between the Tn5 insertions (**Figure 1.5B**). The pseudo-bulk distribution of these fragment sizes reveals distinct periodic maxima for each number of nucleosomes present. At higher genomic resolution, the ATAC-seq signal can also reveal TF binding sites (**Figure 1.5C+D**). In this

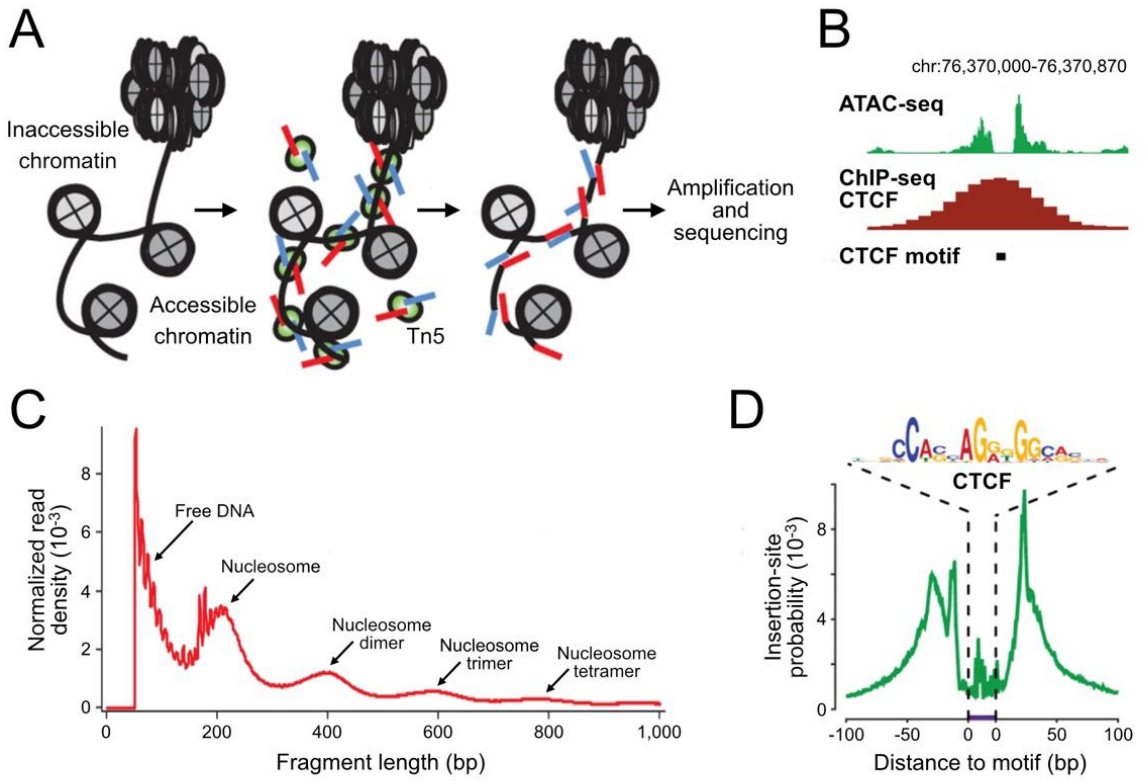


Figure 1.5 Investigating accessible chromatin with ATAC-seq. Data from human lymphoblastoid GM12878 cells is shown. **A** Experimental approach of ATAC-seq. Transposase 5 (Tn5) binds to accessible DNA and inserts sequencing adapters. The resulting ATAC sequencing library is amplified by PCR. **B** Distribution of fragment sizes from ATAC-seq. **C** Genomic signal tracks from ATAC-seq and CTCF ChIP-seq at CTCF motif position on chr1. **D** CTCF footprint from ATAC-seq across genome-wide CTCF binding sites. Adapted from Buenrostro *et al.* (2013).

context, DNA-bound TFs protect their binding sites from Tn5 insertion, rendering them inaccessible. Simultaneously, TF binding increases accessibility in the regions adjacent to the binding site, increasing the likelihood of nearby Tn5 insertions. This results in TF-specific footprints in the chromatin accessibility signal (**Figure 1.5D**). (Buenrostro *et al.*, 2013)

At lower genomic resolution, the ATAC-seq signal aligns with active histone PTMs, such as H3K4me3, H3K9ac, and H3K27ac, at genomic regions corresponding to CREs (**Figure 1.6A+B**). In contrast, no chromatin accessibility is detected at CREs marked by repressive histone PTMs, such as H3K4me1 (**Figure 1.6C**). Additionally, non-CRE genomic regions show no chromatin accessibility regardless of active histone PTMs like H3K4me1, H3K4me3, H3K9ac, and H3K36me3, for example in actively transcribed gene bodies (**Figure 1.6A+B**). (Muckenhuber *et al.*, 2023) Thus, the measurement of chromatin accessibility by ATAC-seq can serve as a reliable proxy to identify CREs with active local chromatin states, which are generally determined by epigenetic histone PTMs and DNA methylation (Jenuwein & Allis, 2001; Greenberg & Bourc'his, 2019), as described in **Section 1.1**.

Furthermore, the ATAC protocol can be integrated into single-cell sequencing techniques, which resolve cell-to-cell differences in chromatin accessibility profiles (Buenrostro *et al.*, 2015). The combined information in chromatin accessibility – encompassing active local chromatin states, nucleosome positioning, and TF binding – is also captured in this data from single-cell sequencing of chromatin accessibility (scATAC-seq). However, at the single-cell level, chromatin accessibility data is inherently sparse due to the limited number of template DNA molecules per genomic position and cell (with only two copies in a diploid genome). (Buenrostro *et al.*, 2015) To address this sparsity, advanced bioinformatic methods have been developed that exploit artificial pseudo-bulks from single cells for genomic feature identification, while maintaining single-cell resolution during quantification (Shi *et al.*, 2022).

The initial computational processing of scATAC-seq data involves several key steps. First, sequencing adapters are trimmed from the reads, followed by their filtering based on sequencing quality, demultiplexing of the single-cell barcodes, and alignment to the reference genome (Shi *et al.*, 2022). Next, the aligned read positions are corrected for the Tn5 insertion offset, duplicated reads from PCR are removed, and paired reads are identified to infer their intermediate fragments (Buenrostro *et al.*, 2013). The quality of scATAC-seq data is assessed using the pseudo-bulk distribution of fragment sizes (**Figure 1.5B**), the expected enrichment of reads at TSS compared to non-TSS regions, and the

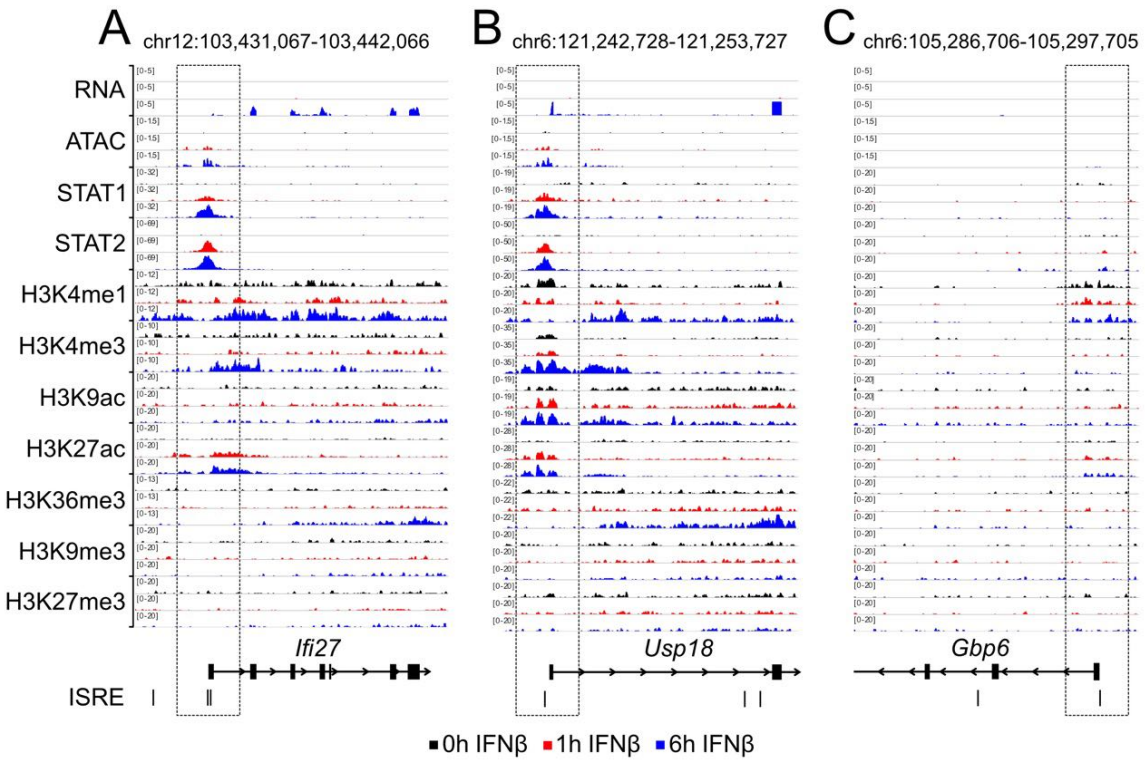


Figure 1.6 Genomic signal tracks from RNA-seq, ATAC-seq, and ChIP-seq of TF binding and histone PTMs in mouse embryonic stem cells. Basal transcription in ESCs (0 h IFNβ, black) was perturbed by 1 h (1 h IFNβ, red) and 6 h (6 h IFNβ, blue) of IFNβ treatment. Promoters are defined as TSS ± 500 bp and marked by dashed boxes. ChIP-seq of IFNβ-induced TFs STAT1 and STAT2 and their complex's "ISRE" binding motifs from Homer database are shown. ChIP-seq of most common histone PTMs H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K9me3, and H3K27me3 are depicted. **A** Genomic region around *Ifi27* gene. RNA: No *Ifi27* expression at 0 h and 1 h IFNβ, strong expression increase at 6 h IFNβ. ATAC: No chromatin accessibility at 0 h IFNβ, gradually increased promoter accessibility at 1 h and 6 h IFNβ. STAT1 and STAT2: No TF binding at 0 h IFNβ, promoter STAT1 and STAT2 binding at 1 h and 6 h IFNβ. Histone PTMs: Depleted H3K4me1 signal at promoter. Increasing H3K4me3 and H3K27ac signal at promoter upon IFNβ treatment. Increasing H3K4me1 and H3K36me3 signal in gene body upon IFNβ treatment. **B** Genomic region around *Usp18* gene. RNA: No *Usp18* expression at 0 h and 1 h IFNβ, strong expression increase at 6 h IFNβ. ATAC: No chromatin accessibility at 0 h IFNβ, gradually increased promoter accessibility at 1 h and 6 h IFNβ. STAT1 and STAT2: No TF binding at 0 h IFNβ, promoter STAT1 and STAT2 binding at 1 h and 6 h IFNβ. Histone PTMs: Depletion of H3K4me1 signal at promoter upon IFNβ treatment. Increasing H3K4me3, H3K9ac, and H3K27ac signal at promoter upon IFNβ treatment. Increasing H3K4me1, H3K4me3, H3K9ac, and H3K36me3 signal in gene body upon IFNβ treatment. **C** Genomic region around *Gbp6* gene. RNA: No *Gbp6* expression. ATAC: No chromatin accessibility. STAT1 and STAT2: No TF binding. Histone PTMs: Scattered promoter H3K4me1 signal at 0 h, 1 h, and 6 h IFNβ. No additional histone PTM signal. Adapted from Muckenhuber *et al.* (2023).

number of unique fragments per cell (Shi *et al.*, 2022). After removing low-quality cell barcodes, accessibility in the high-quality cells is quantified to obtain accessibility count matrices of genomic features in single cells. Unlike RNA-seq, where sequencing reads are typically quantified within well-annotated genes, different genomic features can be used for to generate count matrices for ATAC-seq (**Figure 1.7A**): (i) Genomic tiles of a defined size (e. g., 1 kb), (ii) peaks identified from the pseudo-bulk ATAC-seq signal, (iii) annotated genes (commonly used to infer so-called gene activity scores from scATAC-seq data), and

(iv) regions of interest from complementary data sets (Shi *et al.*, 2022). These different genomic features will result in different accessibility count matrices that strongly impact the genomic resolution and viewpoint of further analysis. Additionally, the accessibility count matrices can be generated by fragment-based or insertion based quantification of scATAC-seq data (**Figure 1.7B**) (Shi *et al.*, 2022). While fragment-based counting of accessibility within these genomic features is recommended by some methods to reduce data sparsity in single cells (Stuart *et al.*, 2021; Martens *et al.*, 2024), insertion-based counting provides a more direct measure of accessibility without capturing the intermediate nucleosome positions within longer fragments (Granja *et al.*, 2021; Miao & Kim, 2024).

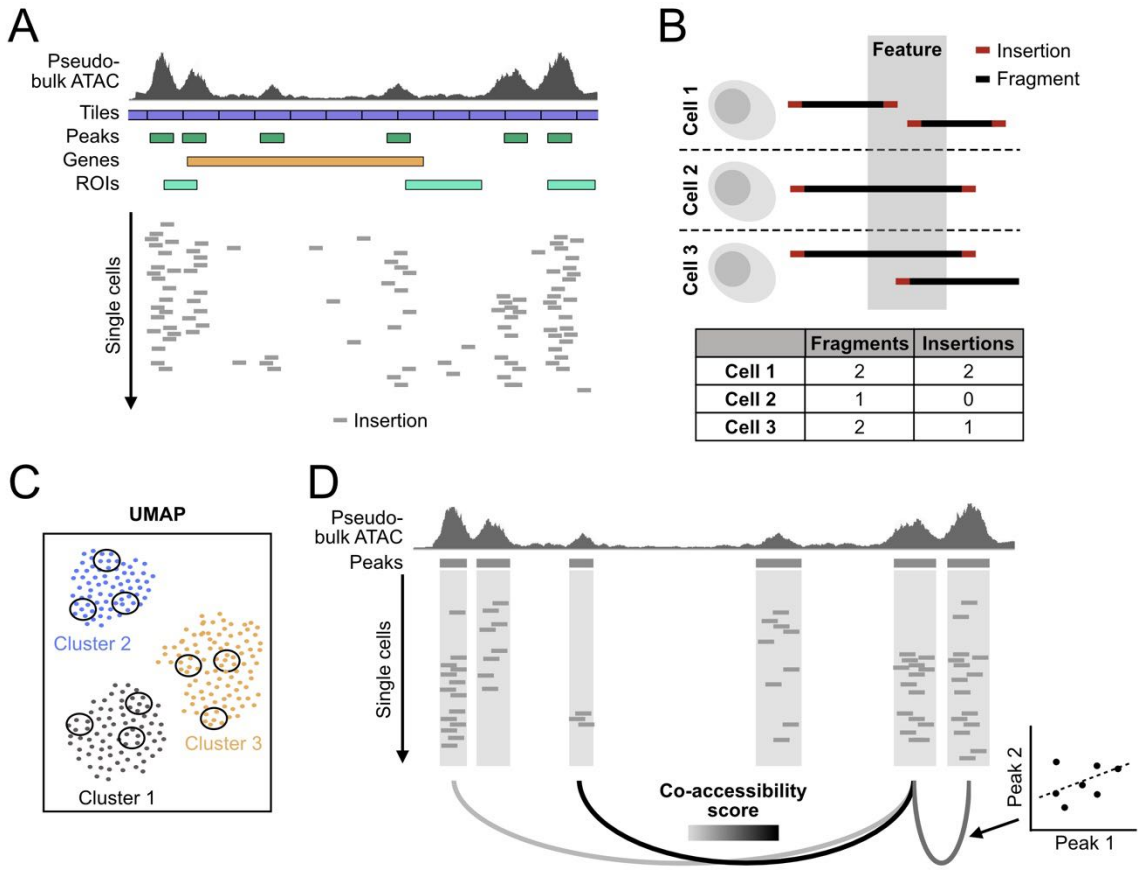


Figure 1.7 Analysis of scATAC-seq data. A Definition of genomic features to quantify chromatin accessibility in single cells. Genomic tiles, peaks from pseudo-bulk ATAC-seq signal, annotated genes, and regions of interest (ROIs) from complementary data set are indicated. **B** Fragment-based and insertion-based quantification of scATAC-seq data. **C** UMAP visualization of scATAC-seq data. Circles indicate cells that are aggregated into metacells. **D** Chromatin co-accessibility analysis identifies simultaneously accessible distal genomic sites.

After initial processing, the accessibility count matrix obtained from scATAC-seq data is used for dimensionality reduction, cell clustering, and subsequent visualization by low-dimensional embedding of the single cells (**Figure 1.7C**) (Shi *et al.*, 2022). For the three steps, techniques such as latent semantic indexing (LSI), k-means clustering, and uniform

manifold approximation and projection (UMAP) are commonly applied to explore local and global patterns among single cells (Cusanovich *et al.*, 2015; Becht *et al.*, 2019). These methods enable the identification of cell clusters, containing single cells with similar chromatin accessibility profiles (Shi *et al.*, 2022). The underlying cell types or states of these cell clusters can be identified using the activity of marker genes, integrating with scRNA-seq data, or mapping to an reference cell atlas (Berest & Tangherloni, 2023; Lotfollahi *et al.*, 2024). Once cell types, states or clusters are defined, a variety of computational tools can be utilized to explore different features of the scATAC-seq data between the different cell groups: Differential accessibility analysis on specific genomic features (Zhao *et al.*, 2024), inference of gene regulatory networks and their activity (Badia *et al.*, 2023), prediction of higher-order chromatin conformation (Duan *et al.*, 2024), genotyping in regions of accessible chromatin (Wiens *et al.*, 2024), or TF footprinting and binding predictions (Schep *et al.*, 2017). Many of these methods aggregate similar cells into metacells to mitigate data sparsity and allowing for more robust analyses (Persad *et al.*, 2023). However, this approach sacrifices the single-cell resolution, potentially losing true stochastic variations in accessibility between cells of the same type or state (**Figure 1.7C**, black circles).

In addition to these computational methods that analyze chromatin accessibility at individual genomic loci, so-called chromatin co-accessibility methods aim to infer the simultaneous accessibility of distal genomic sites along the linear genomic coordinate (**Figure 1.7D**) (Shi *et al.*, 2022). Several methods have been developed to identify co-accessible regions using different computational approaches and count matrices (**Table 1.2**). For instance, RWire calculates Pearson correlation between genomic tiles on accessibility counts from single cells (Mallm *et al.*, 2019). In contrast, Cicero identifies co-accessible peaks by calculating correlation in metacells and applying a distance penalty using graphical lasso (Pliner *et al.*, 2018). To address data sparsity and limited cell numbers, Cicero aggregates cells into metacells, allowing each cell to contribute to multiple metacells with a maximum overlap of 80 % between distinct metacells (Pliner *et al.*, 2018). Lastly, ArchR combines aspects of both methods, calculating Pearson correlation between peaks using Cicero's metacell approach (Granja *et al.*, 2021). Despite the differences in their genomic and cellular scales (**Table 1.2**), all methods report individual co-accessible links between distal genomic regions as potential regulatory links. Indeed, since accessible CREs are considered to be in active local chromatin states (**Figure 1.6**), co-accessible distal sites can be assumed co-active across cells or metacells. However, it remains uncertain at the molecular level whether this co-activity

implies a direct regulatory interaction, a spatial contact in 3D chromatin organization, or simply simultaneous TF binding at distal sites (Shi *et al.*, 2022).

Table 1.2 Key differences between chromatin co-accessibility methods. Cicero was developed by Pliner *et al.* (2018), RWire by Mallm *et al.* (2019), and ArchR by Granja *et al.* (2021).

	Cicero	**RWire**	**ArchR**
Genomic resolution	ATAC peaks	Genomic tiles, regions of interest	ATAC peaks
Cellular resolution	Metacells	Single cells	Metacells
Input cell population	Single cell states	Single cell types	Multiple cell types
Co-accessibility method	Correlation with graphical lasso distance penalty	Pearson correlation	Pearson correlation
Significance assessment	Manual cutoff	Local background model	Student's T statistics

1.3. Studying transcription regulation upon perturbations

To investigate transcription regulation, targeted perturbations of transcription in a model system provide a comparative analysis of different gene regulatory states. Additionally, the regulatory mechanisms can be directly linked to their transcriptional effects. In this context, transcription can be perturbed in a cell by various external or internal stimuli. For instance, cytokine treatment of cultured cells induces transcriptional responses through specific intracellular signaling pathways (Lawrence, 2009; Ivashkiv & Donlin, 2014). This external perturbation facilitates the study of time-resolved transcriptional changes upon a controlled stimulation time point (Bhatt *et al.*, 2012; Bolen *et al.*, 2014). Similarly, co-culture experiments of multiple cell types can externally perturb transcription via cell-cell interactions (Shannon *et al.*, 2021). Conversely, transcription can also be perturbed through genome editing, which induces a transcriptional response by creating a targeted internal stimulus (Doench, 2018). Additionally, naturally occurring genomic perturbations, such as those seen in cancer, offer insights into deregulated transcription by comparing distinct inter- or intratumor clones, each with specific genomic alterations (McGranahan & Swanton, 2017).

Overall, these perturbation experiments are powerful tools for investigating transcription regulatory mechanisms and linking them to their transcriptional effect. Several criteria can be applied to characterize the transcriptional impact of different regulatory mechanisms:

(i) Direction of regulation (up- or downregulation, or both)

(ii) Strength of regulation (magnitude of transcriptional change)

(iii) Response kinetics (fast, slow or persistent transcriptional response)

(iv) Transcriptional bursting kinetics (modification of burst size, burst frequency, or both)

(v) Gene co-regulation (single gene, co-induction or alternating induction of multiple genes)

In the following sections, I will outline the model systems and external or internal perturbations used in this thesis.

1.3.1. Interferon beta treatment of mouse cells

In the first model system, transcription in mouse embryonic stem cells (ESCs) and mouse embryonic fibroblasts (MEFs) was externally perturbed using interferon beta (IFNβ) treatment. The ESCs and MEFs were genetically identical and offer a model to investigate both shared and cell type-specific transcriptional responses. This model highlights the role of different epigenetic modifications and thus local chromatin states at CREs in an otherwise identical genetic environment. ESCs and MEFs were studied under unperturbed conditions, as well as after 1 h and 6 h of IFNβ treatment. This time-course analysis allowed exploration of the complex temporal hierarchy of interferon-stimulated gene (ISG) expression and its underlying regulatory mechanisms (Bolen *et al.*, 2014). The cellular response to interferons (IFNs) is a critical component of innate antiviral immunity and inflammatory responses (Ivashkiv & Donlin, 2014; Au-Yeung & Horvath, 2018). The molecular pathways involved in intracellular IFNβ signaling are outlined below.

Type I IFNs, including IFNα and IFNβ, bind to the extracellular domain of the IFNα receptor (IFNAR) (**Figure 1.8**) (Ivashkiv & Donlin, 2014). This binding activates the intracellular domains of its subunits IFNAR1 and IFNAR2, which in turn activate Janus kinase 1 (JAK1) and tyrosine kinase 2 (TYK2), initiating the JAK-STAT signaling cascade (Stark & Darnell, 2012). These kinases phosphorylate signal transducer and activator of transcription (STAT) transcription factors (Stark & Darnell, 2012) enabling them to dimerize and form specific complexes. STAT1 and STAT3 both form homodimers, while STAT2 only pairs with STAT1 to form the IFN-stimulated gene factor 3 (ISGF3) complexes, along with IFN-regulatory factor 9 (IRF9) (Ivashkiv & Donlin, 2014). Once formed, these STAT complexes

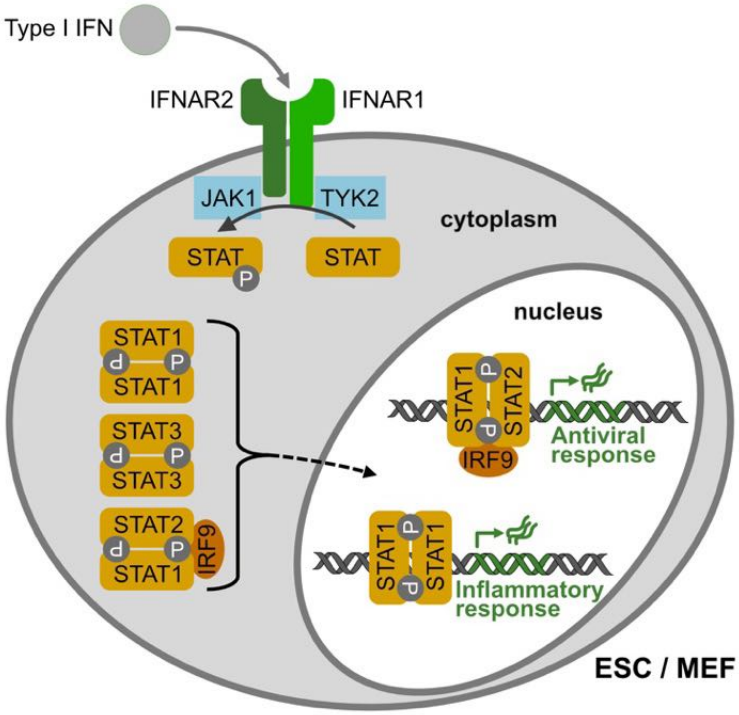


Figure 1.8 Type I IFN-stimulated gene expression. Type I IFNs bind to the extracellular domain of their receptor (IFNAR) and activate the cytoplasmic kinases JAK1 and TYK2. These phosphorylate and thereby activate STAT transcription factors. The phosphorylated STATs can dimerize into homodimer or heterodimer complexes and translocate to the nucleus. They bind their respective DNA motifs and induce antiviral or inflammatory ISGs. After Ivashkiv & Donlin (2014).

translocate to the nucleus, where they bind to specific DNA motifs, inducing the expression of ISGs (Au-Yeung & Horvath, 2018). ISGF3-driven ISGs are primarily involved in antiviral responses, while STAT1 and STAT3 homodimers induce ISGs related to inflammation (Schoggins *et al.*, 2011; Rusinova *et al.*, 2013). These tightly regulated ISGs in antiviral and inflammatory responses serve as excellent examples for studying different molecular mechanisms of transcription upregulation by a defined set of transcription factor complexes.

1.3.2. Tumor necrosis factor alpha treatment of human endothelial cells

In addition to the cytokine perturbation experiments conducted in mouse cells, I studied the external perturbation of transcription through tumor necrosis factor alpha (TNFα) treatment in human umbilical vein endothelial cells (HUVECs). HUVECs are non-immortalized, primary human cells and serve as a valuable model system to investigate transcriptional regulation in healthy human cells. The transcriptome of HUVECs was analyzed in unperturbed condition as well as after TNFα treatment for 30 min and 240 min.

The TNFα treatment induces a tightly regulated proinflammatory gene expression response (Smale, 2010), and the regulatory mechanisms underlying its specific temporal kinetics were studied across the multiple treatment time points (Bhatt *et al.*, 2012). Below, I outline the intracellular signaling cascade that transmits the external TNFα stimulation into a proinflammatory gene expression program.

At the HUVEC cell membrane, TNFα binds to two distinct receptors, namely TNF receptor 1 and 2 (TNFR1 and TNFR2), activating the canonical nuclear factor kappa B (NF-κB) signaling pathway (Legler *et al.*, 2003; Lawrence, 2009) (**Figure 1.9**). The intracellular domains of both receptors recruit IκB kinase (IKK) via the TNF-receptor-associated factor 2 (TRAF2), which is activated by the death domain-containing Ser/Thr kinase receptor-interacting protein (RIP) (Devin *et al.*, 2000). Although TNFα can interact with both receptors, it primarily binds to TNFR1 in most cell types to activate NF-κB. This activation is mediated through TNFR1-associated death domain protein (TRADD), which facilitates the recruitment of downstream factors (Hsu *et al.*, 1995). In addition to mediating NF-κB activation, TRADD also plays a role in inducing cell death through apoptosis (Rahman & McFadden, 2006). Once in a complex with TRAF2 and RIP, IKK dissociates the inhibitor of nuclear factor kappa B (IκBα) from NF-κB. Subsequently, the active transcription factor

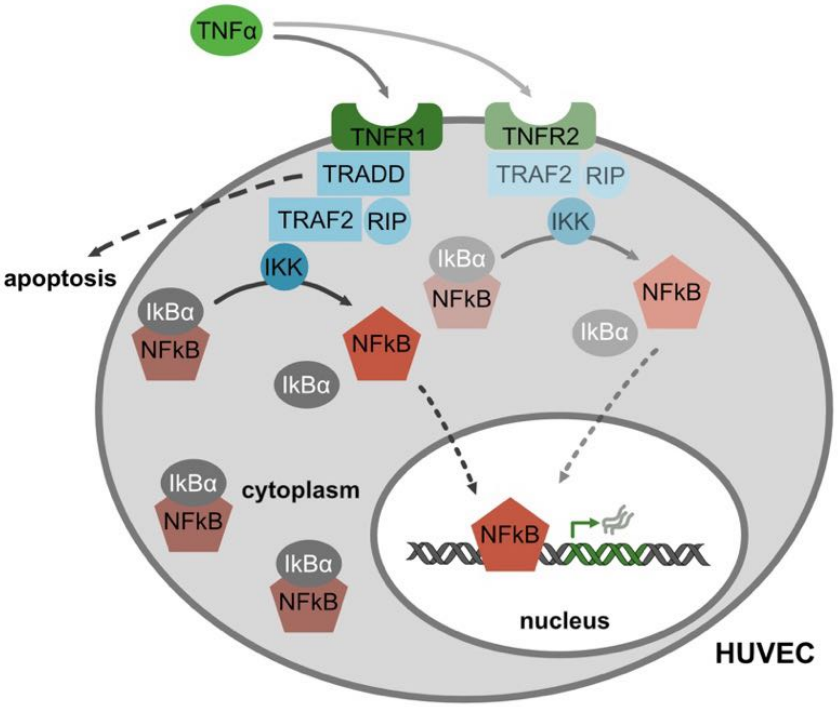


Figure 1.9 Intracellular signaling cascade upon TNFα stimulation. After binding to one of its receptors (TNFR1, TNFR2), TNFα activates IKK to dissociate IκBα from NF-κB. Consequently, the activated transcription factor NF-κB can translocate to the nucleus, bind to DNA and regulate transcription. After Rahman & McFadden (2006).

NF-κB translocates to the nucleus, where it binds specific DNA motifs to regulate gene expression (Karin & Ben-Neriah, 2000). Ultimately, this results in a tightly controlled transcriptional response of proinflammatory genes (Smale, 2010), whose specific temporal kinetics (Bhatt *et al.*, 2012) motivated my investigation of their underlying regulatory mechanisms.

1.3.3. TF knock-out in TCL1 mouse models for CLL

In addition to the previously discussed cytokine-induced transcription perturbations in human and mouse cells, I investigated internal transcription perturbation through genome editing in the Eμ-T-cell leukemia-1 oncogene (*TCL1*)-transgenic mouse model for chronic lymphocytic leukemia (CLL) (Bresin *et al.*, 2016). This mouse model closely mimics human CLL, as the overexpression of *TCL1* leads to CLL development and progression (Bresin *et al.*, 2016). In this system, transcription was perturbed by double knock-out of the *Tbx21* gene, which encodes the TF T-box expressed in T cells (T-bet). Transcriptomic regulation in *Tbx21* wild-type and knock-out TCL1 cells were studied and compared. While this model does not offer the temporal resolution seen in cytokine stimulation experiments, it serves as a suitable framework to examine both the direct and secondary effects of a complete loss of the typically highly expressed TF T-bet on transcription and the regulatory landscape. In the following, a brief overview of the importance of T-bet in CLL is provided.

In the context of human CLL, *TBX21* expression is significantly higher in malignant B cells compared to non-malignant B cells from healthy controls (**Figure 1.10A**). Consistently, elevated T-bet protein levels were observed in CLL samples (**Figure 1.10B**). Importantly, CLL patients with high *TBX21* expression levels showed a significantly longer overall survival compared to those with low *TBX21* expression (**Figure 1.10C**), suggesting a tumor-suppressive role for T-bet. The elevated *TBX21* expression was further associated with increased ATAC-seq and H3K27ac ChIP-seq signals at both the *TBX21* promoter and an intronic region in CLL cells (**Figure 1.10D**). To investigate the driver behind the higher *TBX21* expression in CLL, long-term cultures of CLL cells were co-cultured with different tumor microenvironment cells or treated with various cytokines. Significantly higher *TBX21* expression was observed in CLL cells that were co-cultured with activated autologous T cells (**Figure 1.10E**). Additionally, stimulating CLL peripheral blood mononuclear cells (PBMCs) with IFNγ, CpG oligonucleotides, and combinations of these with other cytokines led to significant increases in *TBX21* expression (**Figure 1.10F+G**). (Roessner *et al.*, 2024) Taken together, the high *TBX21* expression observed in CLL cells, its induction by

inflammatory signals, and the correlation with longer patient survival motivated us to further investigate its role in tumor suppression and transcription regulation.

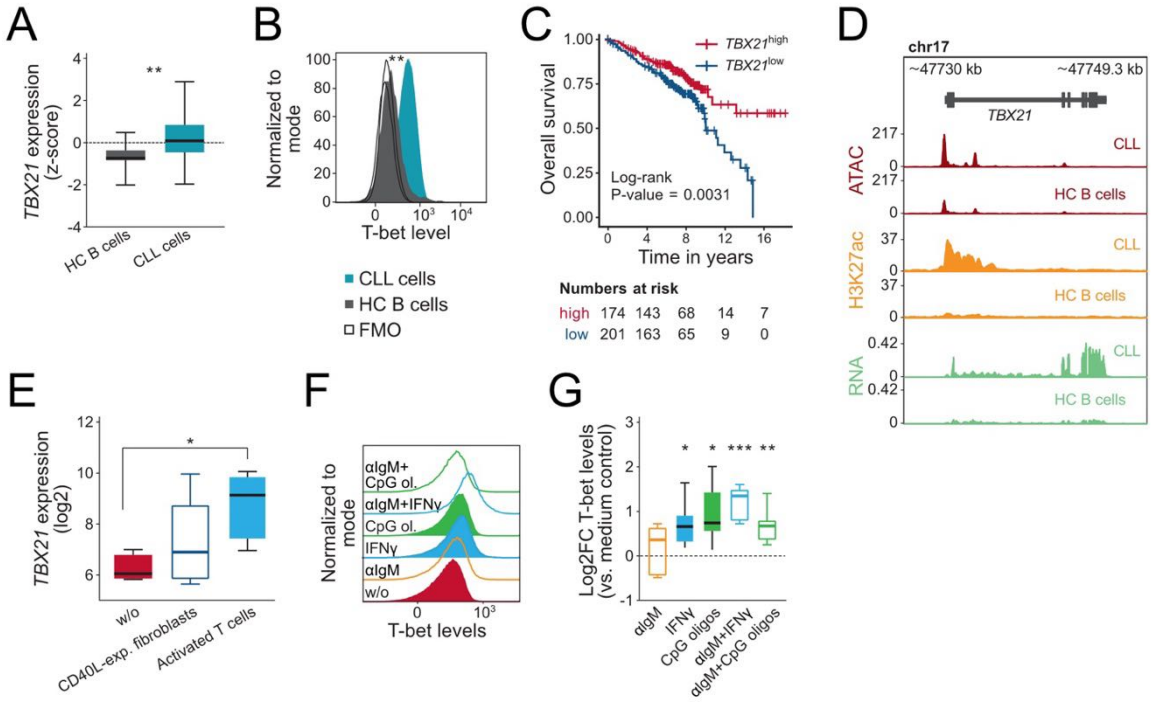


Figure 1.10 TF T-bet in chronic lymphocytic leukemia. A *TBX21* expression in malignant B cells from CLL patients (CLL cells) and untransformed B cells from healthy controls (HC B cells). Bulk RNA-seq data from 41 CLL patients and 11 age-matched healthy controls (HCs) are shown. P-values from unpaired t-test are indicated as *, P < 0.05; **, P < 0.01. **B** T-bet protein levels in CLL cells and HC B cells. Flow cytometry data with fluorescence minus one (FMO) controls from 20 CLL patients and 5 age-matched HCs are shown. P-value from unpaired t-test is indicated as *, P < 0.05; **, P < 0.01. **C** Overall survival of CLL patients with high (*TBX21*^{high}) and low (*TBX21*^{low}) *TBX21* expression in the ICGC cohort. **D** Genomic signal tracks for ATAC-seq (red), H3K27ac ChIP-seq (yellow), and RNA-seq (green) in CLL cells and HC B cells at the *TBX21* gene region. Median signals from 7 CLL patients and 4 healthy controls are shown. **E** *TBX21* expression in CLL cells cultured alone (w/o), co-cultured with CD40L-expressing fibroblasts, and with *in vitro*-activated autologous T cells. Data from 5 replicates per condition are shown. Significant p-values from one-way ANOVA with Benjamini-Hochberg correction are indicated as *, P < 0.05. **F** T-bet protein levels in CLL PBMCs without stimulation (medium control, w/o), and stimulated with CpG oligos, αIgM, IFNγ, and combinations thereof. Flow cytometry data from 7 replicates per condition are shown. **G** Log2FCs of T-bet protein levels in CLL PBMCs under stimulated conditions relative to medium controls. Significant p-values from Wilcoxon tests with Benjamini-Hochberg correction are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001. Adapted from Roessner *et al.* (2024).

1.4. Scope of the thesis

Eukaryotic transcription is regulated by a complex network of molecular mechanisms and their precisely orchestrated temporal and spatial nuclear organization. Today, two largely separate research areas study different aspects of transcription regulation: On the one hand, chromatin topology-centric transcription regulation primarily uses next-generation sequencing methods to study local chromatin states and global chromatin organization. On the other hand, TF-centric transcription regulation studies soluble TFs in the nucleus, their distribution, and DNA binding kinetics using mostly fluorescence microscopy-based approaches. However, an integrated investigation of regulatory mechanisms from both fields is currently lacking. Therefore, this thesis aimed to develop a model of transcription regulation that integrates genome-wide information on chromatin topology- and TF-mediated mechanisms using chromatin accessibility sequencing data at single-cell resolution. To achieve this, three specific objectives were addressed in this thesis:

(i) Advancing the experimental and computational analysis of scATAC-seq.

In the first part of the thesis, I identified data sparsity as the key challenge of scATAC-seq data analysis. I aimed to efficiently reduce data sparsity with an improved experimental protocol of scATAC-seq. Furthermore, I enhanced computational methods for quantifying chromatin accessibility in single cells. These methods were then used to differentiate between true biological variation between single cells and data sparsity at individual genomic loci.

(ii) Developing a computational framework to dissect the molecular mechanisms underlying chromatin co-accessibility.

In the second part, I aimed to design a computational framework to resolve different layers of chromatin co-accessibility between multiple genomic loci based on the insights gained from the previous analyses. I sought to link these layers of chromatin co-accessibility to their underlying molecular mechanisms. Additionally, the goal was to make this computational framework available to the scientific community as a user-friendly and well-documented R software package.

(iii) Identifying the structure-function relationship between different regulatory mechanisms and their transcriptional output.

In the third part of the thesis, I applied the newly developed computational framework termed RWireX for chromatin co-accessibility analysis to different mammalian systems under perturbation to study genome-wide mechanisms of transcription regulation using chromatin co-accessibility. Additionally, my goal was to examine how these chromatin co-

accessibility features are distributed across the genome in different model systems and assess their potential cooperation in forming a multi-layered transcription regulation network.

My thesis successfully addressed these three objectives. By bringing them together, I developed a novel model of transcription regulation, which reconciles previously diverging observations from chromatin topology- and TF-centric studies into a unified, genome-wide framework.

2. Results

2.1. Advancing the experimental and computational analysis of scATAC-seq

In this chapter, I address the first aim of this thesis: Advancing the experimental and computational analysis of scATAC-seq data. To achieve this, I used single-cell sequencing data from MEFs, which were analyzed under perturbation with IFNβ for 6 h (**Table 2.1**; see **Section 1.3.1**). Additionally, I used a more heterogeneous cell system of primary human PBMCs (**Table 2.1**) to explore the potential of chromatin accessibility in resolving different cell types and their specific chromatin footprints. Lastly, I utilized scRNA-seq data from patients with acute myeloid leukemia (AML; **Table 2.1**) to demonstrate how TF activity can be studied at single-cell resolution using expression data.

Table 2.1 Overview of sequencing data sets used for advancing the experimental and computational analysis of scATAC-seq.

Sample	Perturbation	Sequencing type	Assay type
MEF	IFNβ treatment	Single-cell	ATAC, TurboATAC
PBMC	Cell types	Single-cell	ATAC, TurboATAC, Multiome (RNA+ATAC), Multiome (RNA+TurboATAC)
AML patient samples	Mutational landscape	Single-cell	RNA

The majority of the presented results were published in Seufert *et al.* (2023) and Schuster *et al.* (2023). For MEFs, sequencing data acquisition was performed by Markus

Muckenhuber (formerly Division of Chromatin Networks, German Cancer Research Center, Germany). For PBMCs, sequencing data acquisition was conducted by Katharina Bauer and Jan-Philipp Mallm (both Single Cell Open Lab, German Cancer Research Center, Germany). My contribution comprised the computational analysis of sequencing data from MEFs and PBMCs. For AML patient samples, sequencing data acquisition was performed by Linda Schuster (formerly Division of Chromatin Networks, German Cancer Research Center, Germany). Computational analysis was conducted in collaboration with Linda Schuster, where I supported coding and conceptualization. The analysis of TF activity was carried out using my scripts.

2.1.1. Identifying technical biases in scATAC-seq data

Utilizing chromatin accessibility as a genome-wide measure of chromatin state has enhanced the understanding of chromatin organization and its impact on transcription. Investigating chromatin accessibility at single-cell resolution is necessary to resolve the transient nature of transcription and chromatin state. However, the analysis of scATAC-seq data introduces challenges. We acquired two biological replicates of scATAC-seq data from MEFs stimulated for 6 h with IFNβ to observe potential technical biases within and between samples at single-cell resolution and in pseudo-bulks (computational aggregation of single-cell information from one sample). Both replicates were acquired with the same protocol for scATAC-seq but varied in cell numbers used for library generation (20,000 cells for Rep1; 7,500 cells for Rep2).

Pseudo-bulk chromatin accessibility reveals overall data quality

The two replicates showed variations in the total number of sequenced read pairs (**Table 2.2**). Rep2 had twice as many sequenced read pairs as Rep1. However, the duplication rate of Rep1 was higher, and 15 % more read pairs were PCR duplicates that needed to be excluded. Consequently, this resulted in similar numbers of unique read pairs for the two replicates (less than a 1.3-fold difference). On pseudo-bulk level, the two replicates showed similar distributions of fragment sizes (**Figure 2.1A**). However, the number of insertions at TSSs was much lower in Rep1 than Rep2 (**Figure 2.1B**).

When examining the data at single-cell resolution, both replicates showed similar numbers of empty cell barcodes, identified by low numbers of unique fragments and low TSS enrichment scores (grey area in **Figure 2.1C**, **Table 2.3**). Rep1 showed a high number of

Table 2.2 Sequencing quality metrics of scATAC-seq replicates from 6 h IFNβ-treated MEFs.

	Rep1	Rep2
Sequenced read pairs	423,651,840	827,356,762
Percent mapped read pairs	92.5	92.2
Percent duplicated read pairs	55.6	70.1
Percent nucleosome-free fragments	36.5	40.6
Unique read pairs	188,101,417	247,379,672

cell-containing barcodes, of which 19,114 cells were selected as the high-quality cell population for further analysis (red rectangle in **Figure 2.1C**, left; **Table 2.3**). Rep2 showed a four-fold lower number of cell-containing barcodes than Rep1, of which 5,153 cells were selected as the high-quality cell population for further analysis (red rectangle in **Figure 2.1C**, right; **Table 2.3**). For both replicates, high-quality cells were chosen by defining minimal cutoffs for number of unique fragments and TSS enrichment score. These cutoffs were selected to obtain best possible normal distributions of these two quality criteria in the resulting cell populations since the samples contained homogeneous MEFs with no expected biological variation in cell size and chromatin state. For Rep1, this resulted in a minimal cutoff of $10^{3.1}$ for the number of unique fragments and 5 for TSS enrichment score. For Rep2, minimal cutoffs of $10^{3.8}$ for number of unique fragments and 5 for TSS enrichment score were selected.

Table 2.3 Quality metrics of scATAC-seq replicates from 6 h IFNβ-treated MEFs.

	Rep1	Rep2
Barcode number	76,917	67,933
High-quality cell number	19,114	5,153
Singlet number	17,827	4,426
Mean number of fragments	7,751	35,545
Mean TSS enrichment score	9.39	14.78
Mean nucleosome ratio	1.98	2.06

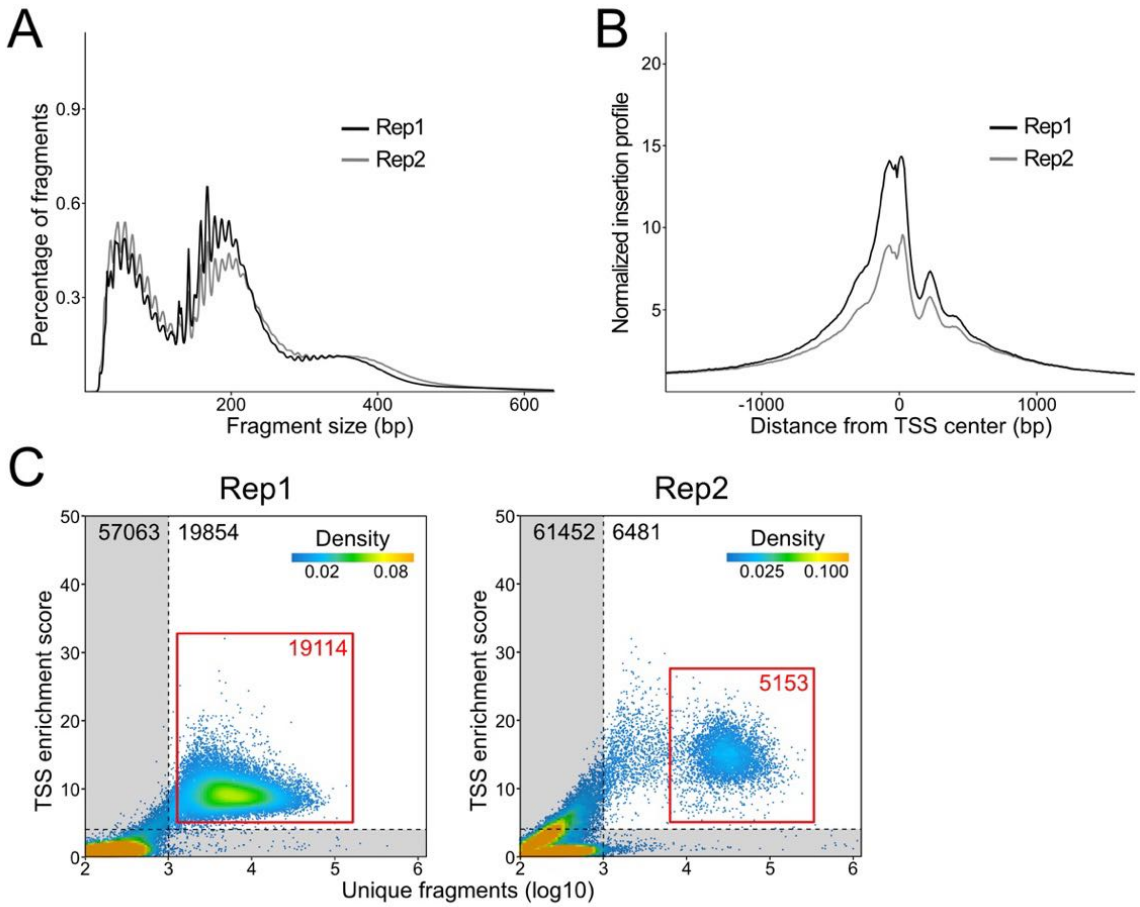


Figure 2.1 Sequencing data quality of scATAC-seq replicates from 6 h IFNβ-treated MEFs. A Fragment size distribution. **B** Number of insertions in 4 kb around TSSs. **C** TSS enrichment score against the number of unique fragments (log10) of cell barcodes in Rep1 (left) and Rep2 (right). The color of points reflects the density of cell barcodes. The grey area marks low-quality cell barcodes. The red rectangle marks selected high-quality cells.

Cell number and sequencing depth drive technical biases between samples

The filtered high-quality cells from Rep1 and Rep2 varied greatly in their numbers of unique fragments and TSS enrichment scores (**Figures 2.2A+B**). Rep1 showed on average almost 5-fold fewer unique fragments and 1.6-fold lower TSS enrichment scores per cell than Rep2 (**Table 2.3**). Nucleosome ratios were comparable between the cells of the two replicates (**Figure 2.2C**, **Table 2.3**).

Next, cell barcodes containing more than one cell, so-called doublets, were identified and filtered out. Doublets were identified by quantifying the number of more than two overlapping fragments at a genomic site per cell barcode. Cell barcodes with disproportionally high numbers of these polyploid overlaps were considered doublets, as healthy MEFs are expected to have a diploid set of chromosomes. The corresponding q-value of exceptionally high polyploid overlaps was closely linked to generally high numbers of unique fragments in cells (**Figure 2.2D**). Doublets were identified by significant q-value

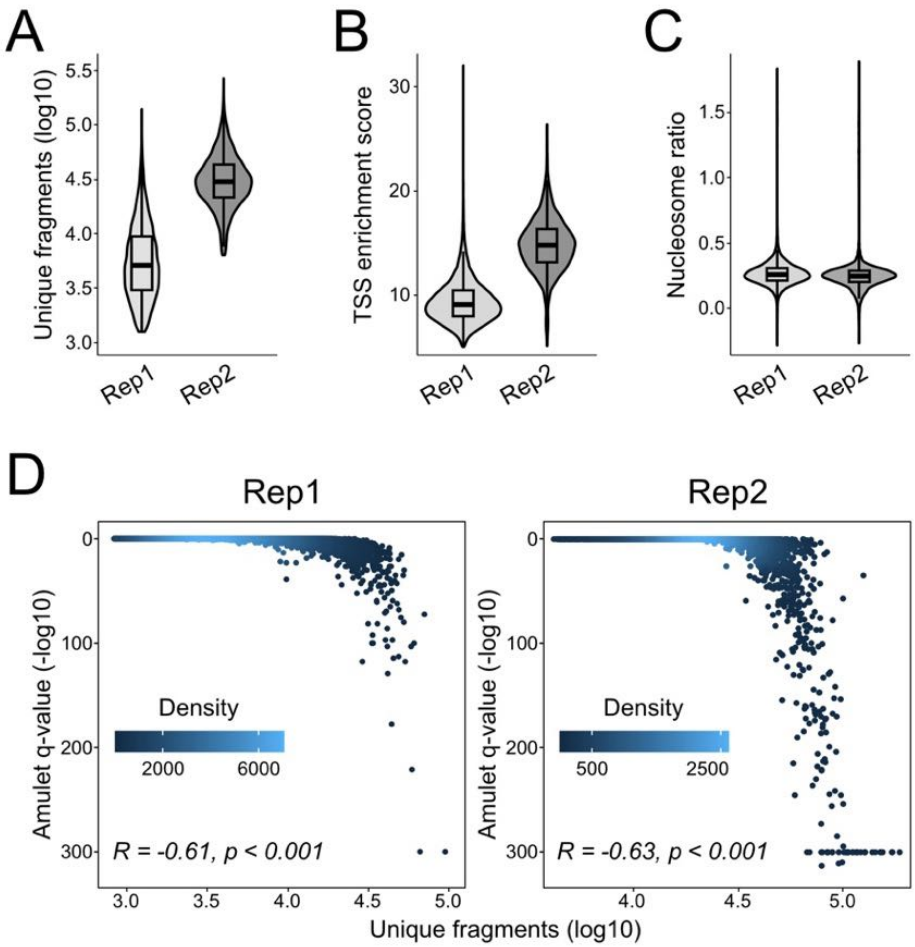


Figure 2.2 Cell quality of scATAC-seq replicates from 6 h IFNβ-treated MEFs. A Number of unique fragments (log10) per cell. **B** TSS enrichment score per cell. **C** Nucleosome ratio per cell. **D** Q-value of doublet probability (log10) against the number of unique fragments (log10) of cells in Rep1 (left) and Rep2 (right). The color of points reflects the density of cells.

of polyploid overlaps below 0.05. This identified 7 % and 14 % of high-quality cells as doublets for Rep1 and Rep2, respectively. Consequently, the doublet filtering resulted in 17,827 and 4,426 high-quality singlets for Rep1 and Rep2 (**Table 2.3**). Here, fewer cells were lost for Rep1 than for Rep2, although Rep1 contained more cells and therefor a higher likelihood of doublets. The lower number of unique fragments per cell in Rep1 appeared to underestimate the actual doublet probability (**Figure 2.2D**).

Low-dimensional embedding reveals technical biases within samples

In addition to the observed technical biases between samples, technical biases within samples were investigated by inspecting the high-quality singlets in low-dimensional embeddings. Dimensionality was reduced using iterative LSI, and relevant components were assessed by their singular value decomposition (SVD) deviation (**Figure 2.3A**). SVD deviation of iterative LSI components above 10 approached a plateau for both replicates, thereby not contributing further information to the single-cell embeddings. Furthermore,

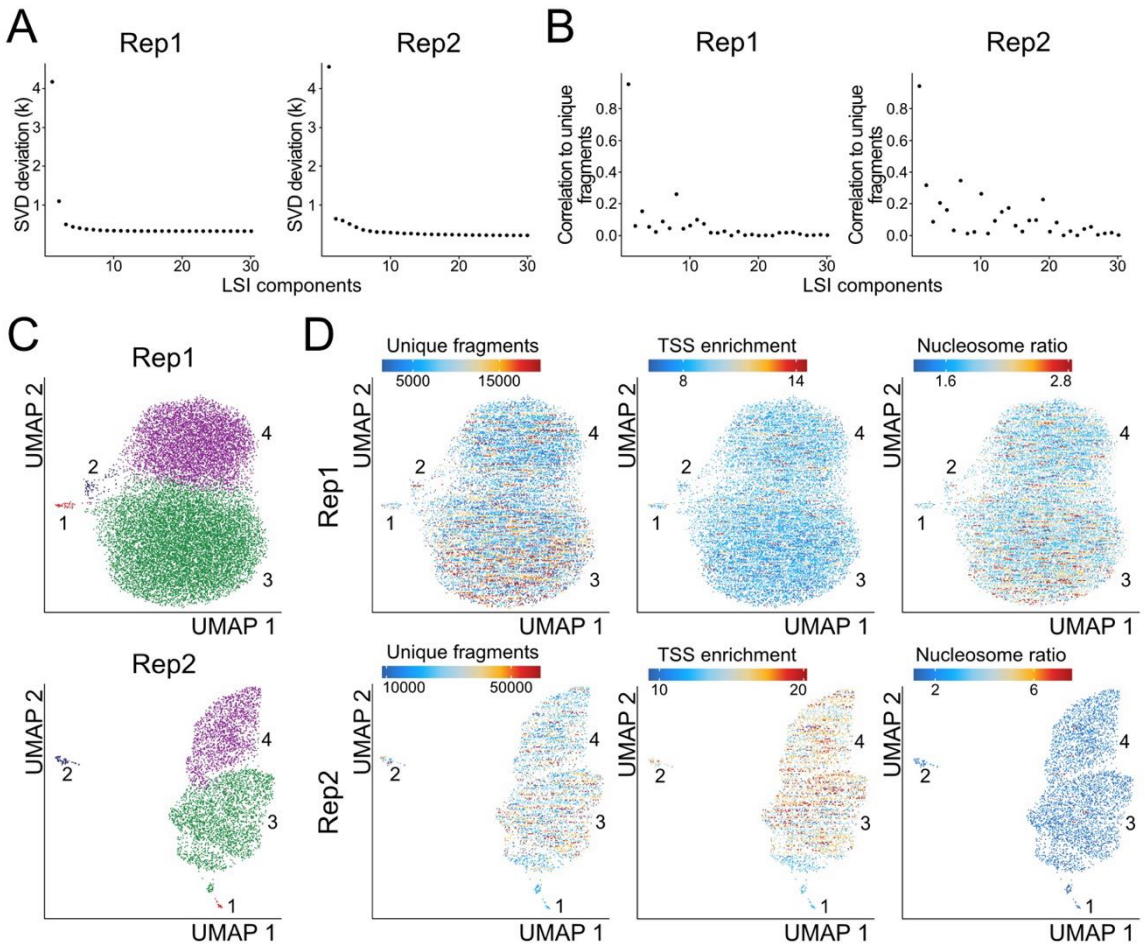


Figure 2.3 Low-dimensional embeddings of scATAC-seq replicates from 6 h IFNβ-treated MEFs. A SVD deviation of LSI components for Rep1 (left) and Rep2 (right). **B** Same as panel A showing LSI component correlation to the number of unique fragments. **C** UMAP embedding of Rep1 (top) and Rep2 (bottom). The color of points reflects the k-nearest neighbor cluster. **D** Same as panel C with color of points reflecting the number of unique fragments per cell (left), TSS enrichment score (middle) and nucleosome ratio (right).

correlation of iterative LSI components to the number of unique fragments was assessed (**Figure 2.3B**). For both replicates, the first iterative LSI component showed high correlation to sequencing depth above 0.8 and thus was removed. Consequently, iterative LSI components 2 to 10 from both replicates were used for low-dimensional embedding (**Figure 2.3C**). Clustering of cells revealed four distinct clusters in both replicates. For both, clusters C1 and C2 only contained few cells, while clusters C3 and C4 contained more than 98 % of the cell populations (**Table 2.4**).

In Rep1, cluster C1 showed lower numbers of unique fragments than clusters C2-4 (**Figures 2.3D**, top left; **2.4A**), while high TSS enrichment scores were enriched in cluster C2 (**Figures 2.3D**, top middle; **2.4B**). Clusters C3 and C4 showed consistent numbers of unique fragments, TSS enrichments scores, and nucleosome ratios (**Figures 2.3D**, top; **2.4**). In Rep2, clusters C1 and C2 showed lower numbers of unique fragments and TSS

enrichment scores than clusters C3 and C4 (**Figures 2.3D**, bottom left and middle; **2.4A+B**). Additionally, cluster C2 showed lower nucleosome ratios (**Figures 2.3D**, bottom right; **2.4C**). In contrast to Rep1, Rep2 clusters C3 and C4 showed small differences in their numbers of unique fragments, but TSS enrichment scores and nucleosome ratios were consistent (**Figures 2.3D**, bottom; **2.4**). For both replicates, clusters C1 and C2 were removed from further analyses, as the differences in chromatin accessibility patterns were predominantly driven by the previously mentioned technical biases within the samples.

Table 2.4 Cell numbers in clusters of scATAC-seq replicates from 6 h IFNβ-treated MEFs.

	C1	C2	C3	C4
Rep1	46	105	11,788	5,888
Rep2	22	62	2,637	1,705

When assessing the low-dimensional embeddings of Rep1 and Rep2 without the technically biased clusters C1 and C2, cluster C3 from both replicates comprised similar fractions of cells (**Figure 2.3C**). Here, Rep1 cluster C3 contained 66 % of all high-quality singlets, while Rep2 cluster C3 comprised 60 % of all high-quality singlets (**Tables 2.3**, **2.4**). Hence, both samples likely captured the same cell subpopulations of MEFs as

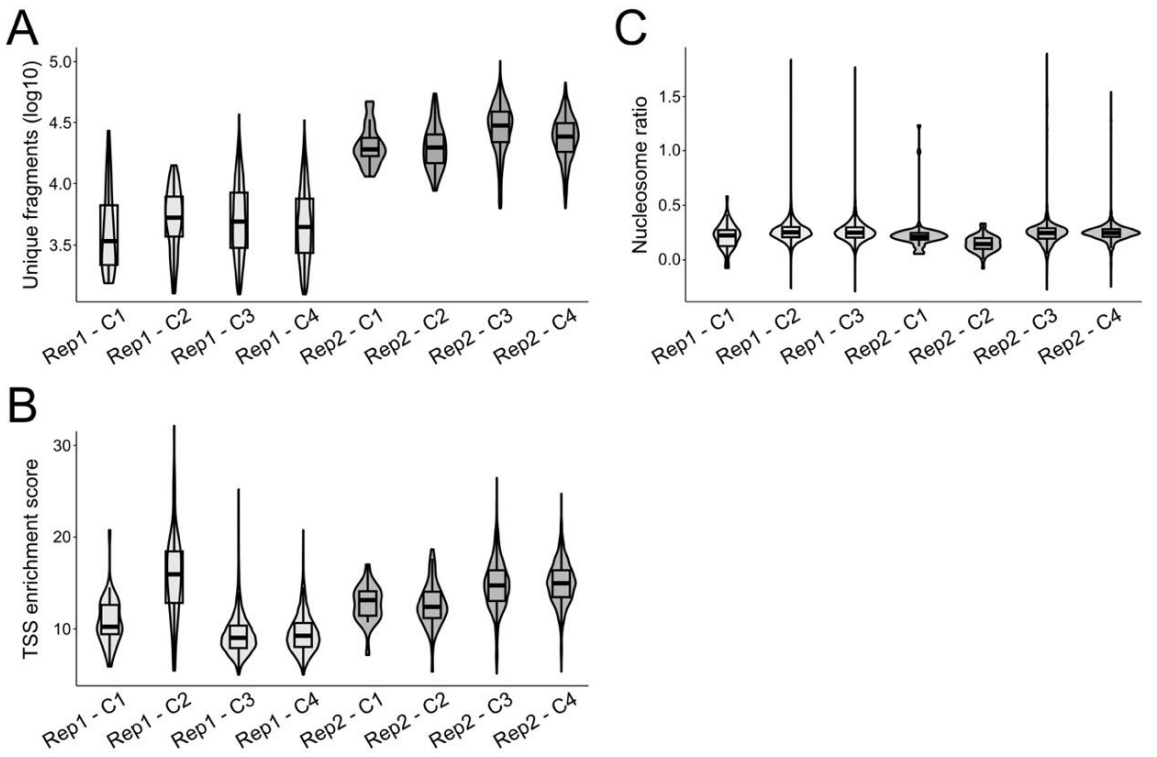


Figure 2.4 Cell quality in clusters of scATAC-seq replicates from 6 h IFNβ-treated MEFs. **A** Number of unique fragments (log10) per cell. **B** TSS enrichment score per cell. **C** Nucleosome ratio per cell.

separated clusters. Notably, Rep2 showed more heterogeneity within clusters in low-dimensional embedding than Rep1, potentially due to the higher single-cell sequencing depth of Rep2.

In summary, sequencing depth was identified as the key determinant of scATAC-seq data quality and strongly biased comparisons between and within samples. Further biases were observed by variations in TSS enrichment scores and, to a lesser extent, nucleosome ratios. Low single-cell sequencing depths resulted in random non-sequenced accessible sites (dropouts), which led to lower TSS enrichment, underestimation of doublets, and less detailed low-dimensional embeddings. This observation is also referred to as sparsity of scATAC-seq data.

2.1.2. Reducing sparsity of scATAC-seq data

Data sparsity is a key challenge in scATAC-seq data analysis, where low data sparsity enables detailed analyses at single-cell and high genomic resolution. Several parameters determine the degree of data sparsity: (i) sequencing depth, (ii) cell number, (iii) Tn5 activity, and (iv) Tn5 reaction buffer. High sequencing depth per single cell significantly reduces data sparsity and can be achieved by either increasing total sequencing depth or decreasing cell number. However, increasing sequencing depth is costly and inflates the rate of sequenced PCR duplicates, while decreasing of cell numbers hinders downstream analyses. Consequently, we aimed to reduce scATAC-seq data sparsity by optimizing Tn5 activity and the Tn5 reaction buffer in collaboration with the Single Cell Open Lab at DKFZ.

High Tn5 activity increases the number of detected accessible sites

We assessed the relative activities of in-house Tn5 (Tn5-H), 10x Genomics Tn5 (Tn5-TXG), and Illumina TDE1 enzyme (Tn5-ILMN) by measuring the fragmentation of lambda DNA using qPCR. Here, higher concentrations of Tn5-H showed increased fragmentation activity (**Figure 2.5A**). Different versions of Tn5-TXG varied significantly in their activity. The highest concentration of in-house Tn5 (Tn5-H100) showed 1.3- to 4-fold higher activity than the Tn5-TXGs. Additionally, we tested the effect of different buffer compositions on Tn5 activity. We observed significantly lower Tn5-H100 and Tn5-ILMN activity with standard Tn5 reaction (Tag) buffer than with the 10x Genomics (TXG) buffer, while the effect on Tn5-TXGv2 was not significant (**Figure 2.5B**).

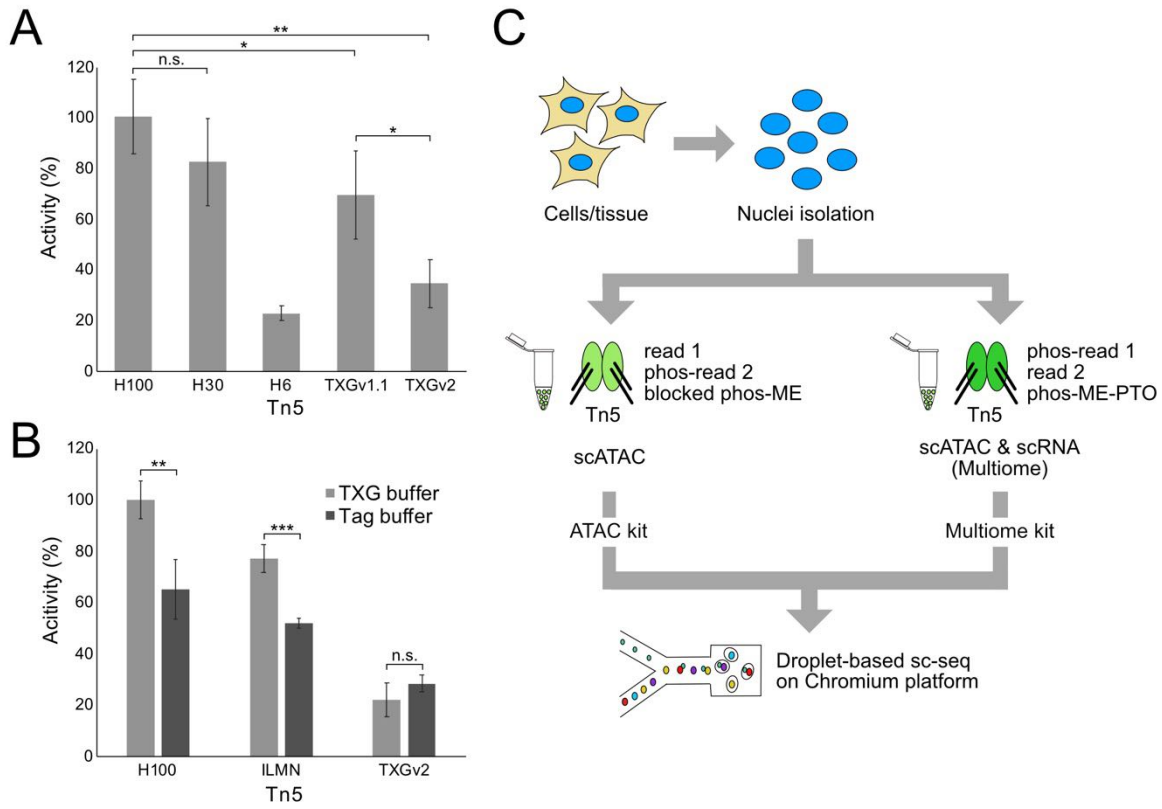


Figure 2.5 TurboATAC protocol for scATAC-seq experiments. A Tn5 activity calculated from qPCR measurement of fragmented lambda phage DNA. Comparison of in-house Tn5 at three concentrations (relative activity levels from highest to lowest: Tn5-H100, Tn5-H30, Tn5-H6) and Tn5 from two versions of 10x Genomics kits (Tn5-TXGv1.1, Tn5-TXGv2). Error bars represent the standard deviation from triplicates. P-values from two-sided, unpaired Student's t-test are indicated as *, P> 0.05; **, P < 0.01; ***, P > 0.001. **B** Same as panel A with Tn5-H100, Tn5-TXGv2 and Illumina TDE1 enzyme (Tn5-ILMN) in buffer from 10x Genomics (TXG buffer) and standard Tn5 reaction buffer (Tag buffer). **C** Experimental workflow of scATAC-seq as well as Multiome (scATAC-seq and scRNA-seq). Variable Tn5 preparations can be used. Adapted from Seufert *et al.* (2023).

We devised flexible experimental protocols for scATAC-seq and simultaneous Multiome scRNA-/scATAC-seq from the same cells that allowed the application of varying Tn5 preparations (**Figure 2.5C**). For both protocols, nuclei were isolated and incubated with Tn5. The adapter loading of the Tn5 varied for the two protocols. The scATAC protocol required blocking of the read 2 adapter phosphorylation with a phosphate-5'-methyl ether (phos-ME), while the Multiome-scATAC protocol utilized blocking of the read 1 adapter phosphorylation with a phos-ME-phosphorothioate (PTO). Subsequently, library generation of scATAC and Multiome-scATAC followed standard protocols.

Next, we tested Tn5-H100, Tn5-H30 and Tn5-TXGv1.1 in the scATAC protocol using MEFs treated with IFNβ for 6 h. While the mapping rate of pseudo-bulk sequencing data was highly comparable across scATAC protocols with different Tn5 preparations, the duplicate rate and nucleosome-free fragment rate showed strong variations (**Table 2.5**). scATAC Tn5-H samples had higher rates of nucleosome-free fragments, in line with the

previously observed higher activity of Tn5-H and consequently higher chromatin fragmentation. A lower duplicate rate in the Tn5-H100 sample indicated a higher complexity of the constructed ATAC library, resulting in a three-fold increase in unique read pairs. However, this was biased by the higher sequencing depth of Tn5-H100 compared to Tn5-TXGv1.1, which is why down-sampling of Tn5-H100 sequencing reads was performed. At comparable sequencing depths, Tn5-H100 showed a two-fold lower duplicate rate than Tn5-TXGv1.1 and Tn5-H30, as well as a two-fold higher number of unique read pairs (**Table 2.5**).

Table 2.5 Sequencing quality metrics of scATAC-seq and scTurboATAC-seq from 6 h IFNβ-treated MEFs. Adapted from Seufert *et al.* (2023).

	Tn5-TXGv1.1	Tn5-H30	Tn5-H100	Tn5-H30 down-sampled	Tn5-H100 down-sampled
Sequenced read pairs	827,356,762	1,221, 969,310	1,304, 430,940	827,400,971	827,404,289
Percent mapped read pairs	92.2	88.2	87.7	88.2	87.7
Percent duplicated read pairs	70.1	75.8	45.9	68.7	36.0
Percent nucleosome-free fragments	40.6	54.5	64.8	54.9	65.7
Unique read pairs	247,379,672	295,716,573	705,697,139	258,976,504	529,538,745
Cell barcode number	67,933	74,384	87,989	-	-

At single-cell resolution, the three samples showed similar numbers of cell-containing barcodes, identified by high numbers of unique fragments and high TSS enrichment scores (white area in **Figure 2.6A**). Moreover, all samples showed similar numbers of high-quality cells (red rectangles in **Figure 2.6A**). For Tn5-H100, this high-quality cell population shifted towards higher numbers of unique fragments and TSS enrichment scores (**Table 2.6**). Furthermore, Tn5-H100 exhibited a second barcode population of intermediate numbers of unique fragments below the cell cutoff but TSS enrichment scores above the cell cutoff. These barcodes had low fraction of reads in peaks (FRIP) scores of approximately 0.3 and were clearly separated from high-quality cells with FRIP scores of roughly 0.7 (**Figure 2.6B**, right). Tn5-TXGv1.1 and Tn5-H30 did not show this second barcode population with intermediate numbers of unique fragments and TSS enrichment scores but low FRIP scores (**Figure 2.6B**).

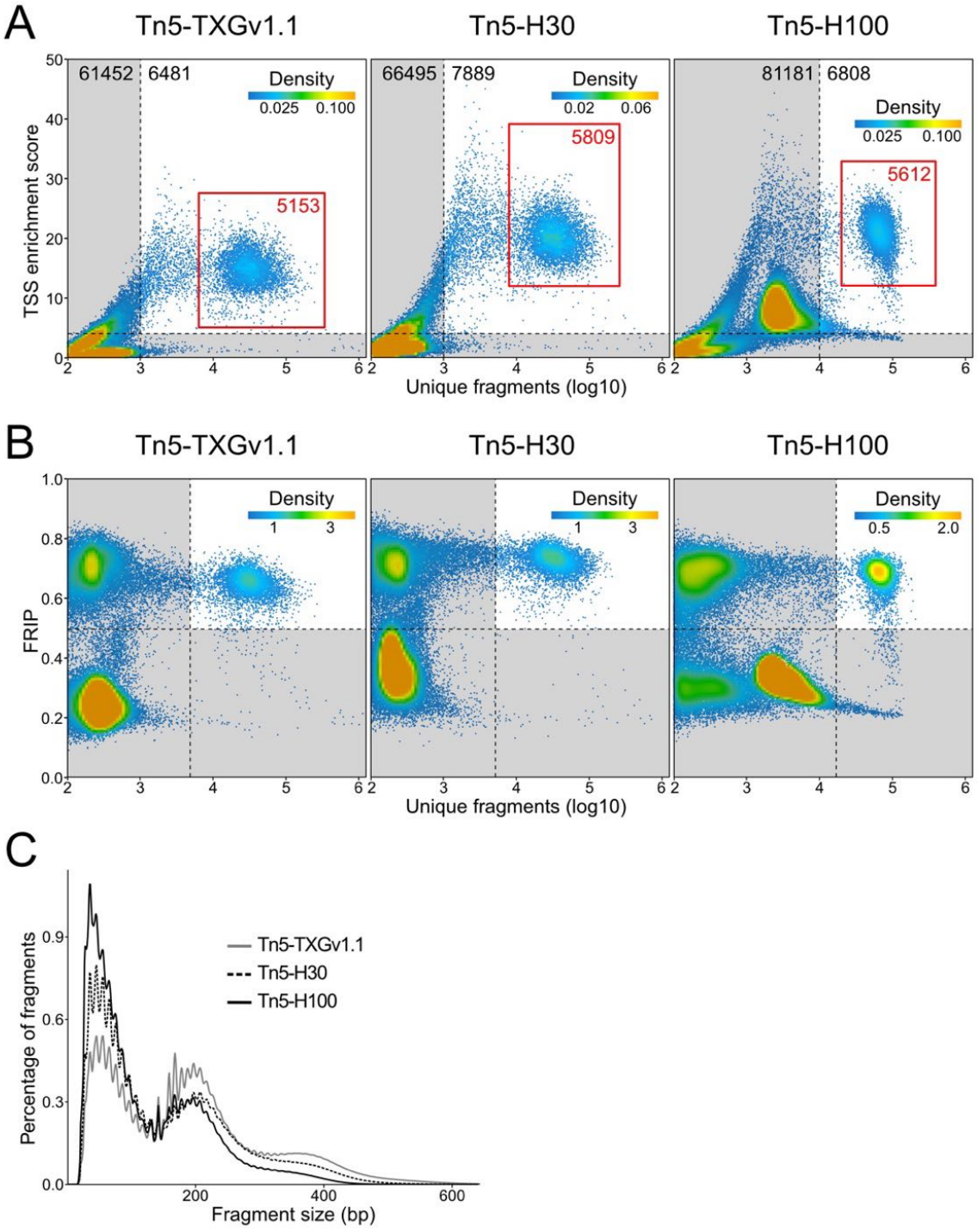


Figure 2.6 Quality of scATAC-seq experiments with different Tn5 preparations from 6 h IFNβ-treated MEFs. A TSS enrichment score against the number of unique fragments (log10) of cell barcodes in scATAC with Tn5-TXGv1.1 (left), Tn5-H30 (middle) and Tn5-H100 (right). The color of points reflects the density of cell barcodes. The grey area marks low-quality cell barcodes. The red rectangle marks selected high-quality cells. **B** Same as panel A with fraction of reads in peaks (FRIP) score against the number of unique fragments (log10). The grey area marks cell barcodes with numbers of unique fragments below respective high-quality cell cutoffs and FRIP scores below 0.5. **C** Fragment size distribution in scATAC with Tn5-TXGv1.1, Tn5-H30 and Tn5-H100. Adapted from Seufert *et al.* (2023).

Table 2.6 Quality metrics of scATAC-seq and scTurboATAC-seq from 6 h IFNβ-treated MEFs.

	Tn5-TXGv1.1	Tn5-H30	Tn5-H100
High-quality cell number	5,153	5,809	5,612
Mean number of unique fragments	30,200	31,623	64,565
Mean TSS enrichment score	14.77	20.02	20.94
Mean nucleosome ratio	1.76	0.91	0.64

The pseudo-bulk distribution of fragment sizes revealed an enrichment of fragments without nucleosomes in Tn5-H samples compared to Tn5-TXGv1.1 (**Figure 2.6C**, **Table 2.6**). We selected the Tn5-H100 protocol as our scTurboATAC protocol, as Tn5-H100 demonstrated the lowest duplicate rate, highest single-cell unique fragment numbers, high TSS enrichment, and high FRIP. In the following, the Tn5-TXGv1.1 sample will serve as the scATAC-seq reference.

scTurboATAC reduces data sparsity in MEFs

Low-dimensional embedding and clustering revealed four distinct clusters for scTurboATAC-seq and scATAC-seq data (**Figure 2.7A**). For scATAC-seq data, clusters C1 and C2 contained only a few cells (**Figure 2.7A**, left) with high apoptosis scores (**Figure 2.7B**, left), while clusters C3 and C4 included more than 98 % of the cell population. For scTurboATAC-seq data, cluster C1 only contained few cells (**Figure 2.7A**, right) with high apoptosis scores (**Figure 2.7B**, right), while clusters C2, C3, and C4 contained the majority of the cell population. Excluding the clusters with low cell numbers and high apoptosis scores, scTurboATAC resolved an additional cluster, likely due to the higher number of unique fragments per cell in scTurboATAC-seq data than in scATAC-seq data (**Figure 2.7C**). When down-sampling scTurboATAC-seq data to the same sequencing depth as scATAC-seq data, the three major cell clusters and the higher number of unique fragments per cell persisted (**Figure 2.7D+E**). This confirmed that higher Tn5 activity reduced data sparsity in MEFs, leading to higher resolution in low-dimensional embeddings.

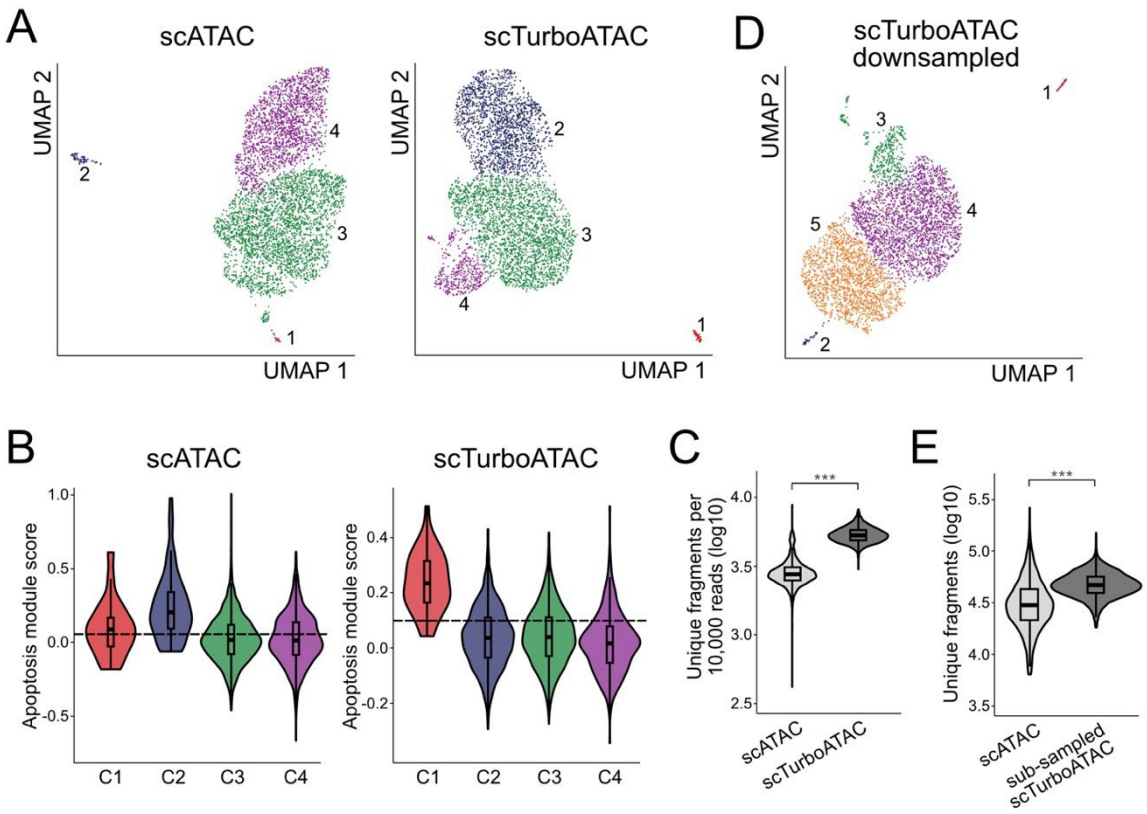


Figure 2.7 Data complexity of scATAC-seq and scTurboATAC-seq data from 6 h IFNβ-treated MEFs. A UMAP embedding of scATAC (left) and scTurboATAC (right). The color of points reflects the k-nearest neighbor cluster. **B** Module score of apoptosis genes per cell in clusters of scATAC (left) and scTurboATAC (right). **C** Number of unique fragments per 10,000 raw reads (log10) and cell for scATAC and scTurboATAC. P-values from two-sided, unpaired Student's t-test are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001. **D** Same as panel A for sub-sampled scTurboATAC. **E** Same as panel C with the number of unique fragments (log10) per cell for scATAC and sub-sampled scTurboATAC. Adapted from Seufert *et al.* (2023).

scTurboATAC reduces data sparsity in primary human PBMCs

Next, we aimed to test whether scTurboATAC-seq also reduces data sparsity in samples of primary cells, rather than cultured cell samples. Therefore, we performed scATAC using Tn5-H100 (scTurboATAC-seq) and Tn5-TXGv2 (scATAC-seq) on primary human PBMCs. Here, sequencing depth was comparable for the two protocols, with a slightly lower mapping rate for scTurboATAC-seq (**Table 2.7**). As with MEFs, scTurboATAC-seq data showed 15 % fewer duplicates, resulting in a 1.5-fold higher number of unique read pairs. The rates of nucleosome-free fragments were more comparable between scTurboATAC-seq and scATAC-seq data than for MEFs.

scTurboATAC-seq detected a higher number of total barcodes than scATAC-seq, but the number of cell-containing barcodes was comparable (grey and white area in **Figure 2.8A**). Moreover, both samples showed similar numbers of high-quality cells (red rectangles in **Figure 2.8A**). The high-quality cell population shifted towards higher TSS enrichment

scores for scATAC-seq data, but scTurboATAC-seq data showed significantly higher numbers of unique fragments (**Figure 2.8B**). Low-dimensional embedding and clustering revealed 10 and 14 distinct cell clusters in scATAC-seq and scTurboATAC-seq data, respectively (**Figure 2.8C**). Additionally, scTurboATAC-seq captured significantly more accessible peaks per single cell than scATAC-seq (**Figure 2.8D**). Finally, annotation of detected cell clusters to hematopoietic cell types by gene activity scores revealed improved detection and resolution by scTurboATAC-seq (**Figure 2.8E**). In scATAC-seq data, classical monocytes and dendritic cells shared one cell cluster (C9), which was resolved into separate clusters in scTurboATAC-seq data (classical monocytes in C3; dendritic cells in C6; **Figures 2.8C+E**). Furthermore, scTurboATAC-seq improved the detection of progenitor cells and showed a higher resolution of the B cell cluster, thrombocyte/granulocyte cluster, and naïve and mature $CD8^+$ T cell clusters (**Figure 2.8E**).

Table 2.7 Sequencing quality metrics of scATAC-seq and scTurboATAC-seq from human PBMCs. Adapted from Seufert *et al.* (2023).

	scATAC	scTurboATAC
Sequenced read pairs	1,466,934,091	1,468,516,560
Percent mapped read pairs	94.0	90.5
Percent duplicated read pairs	66.6	52.2
Percent nucleosome-free fragments	50.4	47.5
Unique read pairs	489,955,986	701,950,916
High-quality cell number	7,658	8,309

Integration of scTurboATAC into multi-omic assays improves data quality

Finally, we intended to assess whether the TurboATAC protocol also reduced data sparsity of Multiome scATAC-seq data, while preserving similar data quality of the respective Multiome scRNA-seq data. Consequently, we tested Tn5-H100, Tn5-H50 and Tn5-TXGv2 in the Multiome protocol (see **Figure 2.5C**) using primary human PBMCs. While sequencing depth and mapping rate of pseudo-bulk Multiome scATAC-seq data were highly comparable across protocols with different Tn5 preparations, the duplicate rate was considerably lower for the Tn5-H100 protocol (**Table 2.8**). This resulted in a higher number of unique Multiome-ATAC read pairs for the Tn5-H100 protocol. The corresponding Multiome scRNA-seq data showed comparable sequencing depths and duplicate rates. However, the mapping rate was approximately 10 % lower for Multiome-RNA from Tn5-

H50 and Tn5-H100 protocols. Regardless, the number of unique Multiome-RNA read pairs was highly comparable across all protocols with different Tn5 preparations.

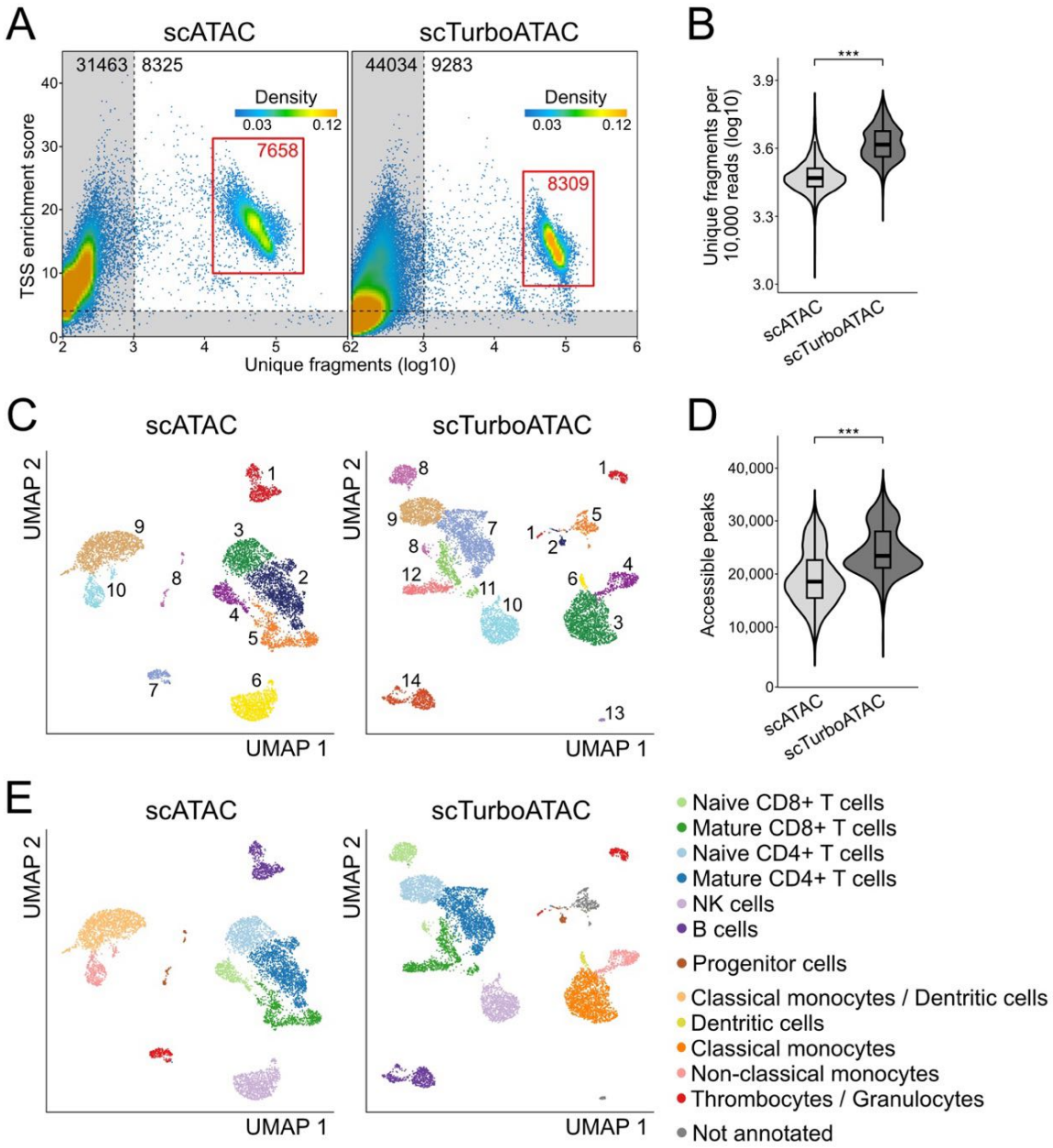


Figure 2.8 Quality of scATAC-seq and scTurboATAC-seq data from human PBMCs. A TSS enrichment score against the number of unique fragments (log10) of cell barcodes in scATAC (left) and scTurboATAC (right). The color of points reflects the density of cell barcodes. The grey area marks low-quality cell barcodes. The red rectangle marks selected high-quality cells. **B** Number of unique fragments per 10,000 raw reads (log10) and cell for scATAC and scTurboATAC. P-values from two-sided, unpaired Student's t-test are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001. **C** UMAP embedding of scATAC (left) and scTurboATAC (right). The color of points reflects the k-nearest neighbor cluster. **D** Number of accessible peaks per cell from the merged peak set of scATAC and scTurboATAC. P-values from two-sided, unpaired Student's t-test are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001. **E** Same as panel C with the color of points reflecting cell type annotation by marker gene activity scores. Adapted from Seufert *et al.* (2023).

Table 2.8 Quality metrics of Multiome scRNA-seq and scATAC-seq experiments with different Tn5 preparations from human PBMCs. Adapted from Seufert *et al.* (2023).

	Tn5-TXGv2	Tn5-H50	Tn5-H100
Sequenced ATAC read pairs	1,304,897,617	1,431,221,793	1,444,079,941
Percent mapped ATAC read pairs	92.5	91.0	91.2
Percent duplicated ATAC read pairs	68.5	67.2	57.3
Unique ATAC read pairs	411,042,749	469,440,748	616,622,135
Sequenced RNA read pairs	714,778,267	644,951,970	716,455,503
Percent mapped RNA read pairs	61.4	53.1	51.8
Percent duplicated RNA read pairs	92.3	91.8	92.2
Unique RNA read pairs	55,037,926	52,886,061	55,883,529
Cell number ATAC	4,563	7,405	6,690
Cell number RNA	4,587	7,916	7,101
High-quality cell number ATAC + RNA	3,620	5,999	5,077

At single-cell resolution, Multiome scATAC-seq data from Tn5-H protocols identified higher numbers of cell-containing barcodes (**Figures 2.9A+B**, **Table 2.8**). Specifically, Multiome scATAC-seq with Tn5-H100 showed numerous cells with high numbers of unique fragments but low TSS enrichment scores (grey area in **Figure 2.9A**, right). These cells exhibited high numbers of TSS reads, but their TSS read numbers appeared to plateau at high unique fragment numbers (**Figure 2.9B**, right). Consequently, cell-containing barcodes were identified by high numbers of unique fragments and high TSS read numbers, rather than TSS enrichment scores (white area in **Figure 2.9B**, **Table 2.8**). Cells from Multiome scATAC-seq data with Tn5-H100 contained significantly higher numbers of unique fragments than those with Tn5-TXGv2 and Tn5-H50 (**Figure 2.9C**, left). Additionally, the Multiome scRNA-seq data from the Tn5-TXGv2 protocol detected nearly half the number of cell-containing barcodes compared to the Tn5-H protocols (**Table 2.8**). Nevertheless, all protocols resulted in comparable numbers of Multiome RNA UMI counts per single cell (**Figure 2.9C**, right). Finally, high-quality cells from both Multiome scATAC-seq and scRNA-seq data were determined, revealing a higher number of multi-omic high-quality cells for Tn5-H protocols (**Table 2.8**). We selected the Tn5-H100 protocol as the Multiome scTurboATAC-seq protocol, as it showed the lowest pseudo-bulk duplicate rate,

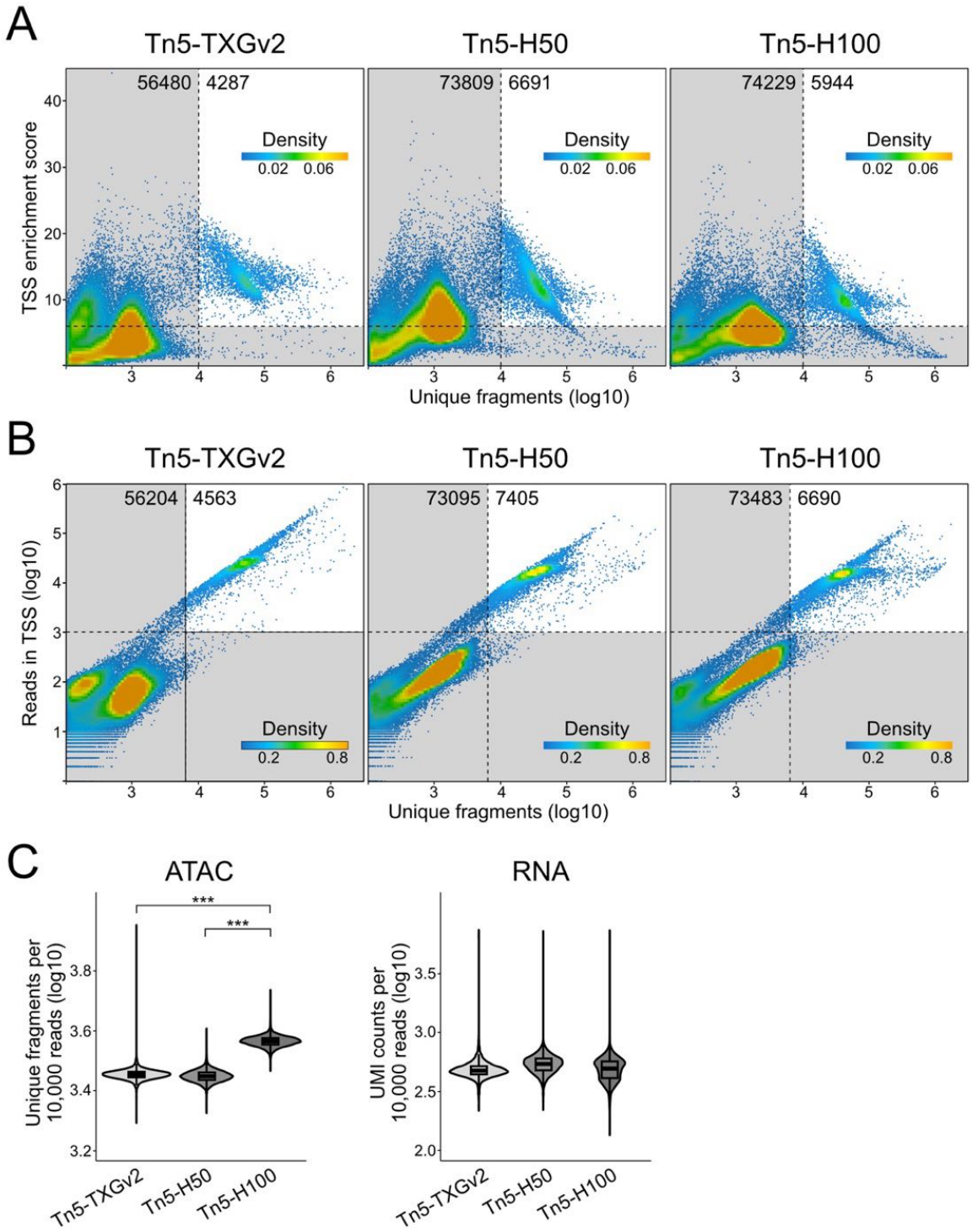


Figure 2.9 Quality of Multiome scRNA-seq and scATAC-seq experiments with different Tn5 preparations from human PBMCs. A TSS enrichment score against the number of unique fragments (log10) of cell barcodes in Multiome-scATAC with Tn5-TXGv2 (left), Tn5-H50 (middle) and Tn5-H100 (right). The color of points reflects the density of cell barcodes. The grey area marks low-quality cell barcodes. **B** Same as panel A with the number of reads in TSS (log10) against the number of unique fragments (log10). **C** Number of unique Multiome-scATAC fragments (left) and Multiome-scRNA UMI counts (right) per 10,000 raw reads (log10) and cell for Multiome with Tn5-TXGv2, Tn5-H50 and Tn5-H100. P-values from two-sided, unpaired Student's t-test are indicated as *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Adapted from Seufert *et al.* (2023).

highest single-cell unique fragment numbers, and high RNA data quality. The Tn5-TXGv2 sample was used as the Multiome scATAC reference.

Low-dimensional embedding and clustering revealed 11 cell clusters for Multiome scTurboATAC-seq and 9 clusters for Multiome scATAC-seq data (**Figures 2.10A+B**, left). For Multiome scRNA-seq data, 11 distinct clusters were resolved with both protocols (**Figures 2.10A+B**, middle). Lastly, the co-embedding of Multiome scATAC- and scRNA-seq resulted in 15 clusters for the TurboATAC protocol and 14 clusters for the standard protocol (**Figures 2.10A+B**, right). Consequently, the reduced data sparsity in the Multiome-scTurboATAC protocol resolved more cell clusters in only ATAC and combined ATAC+RNA embeddings, while only RNA embedding showed similar resolution than the Multiome-scATAC protocol.

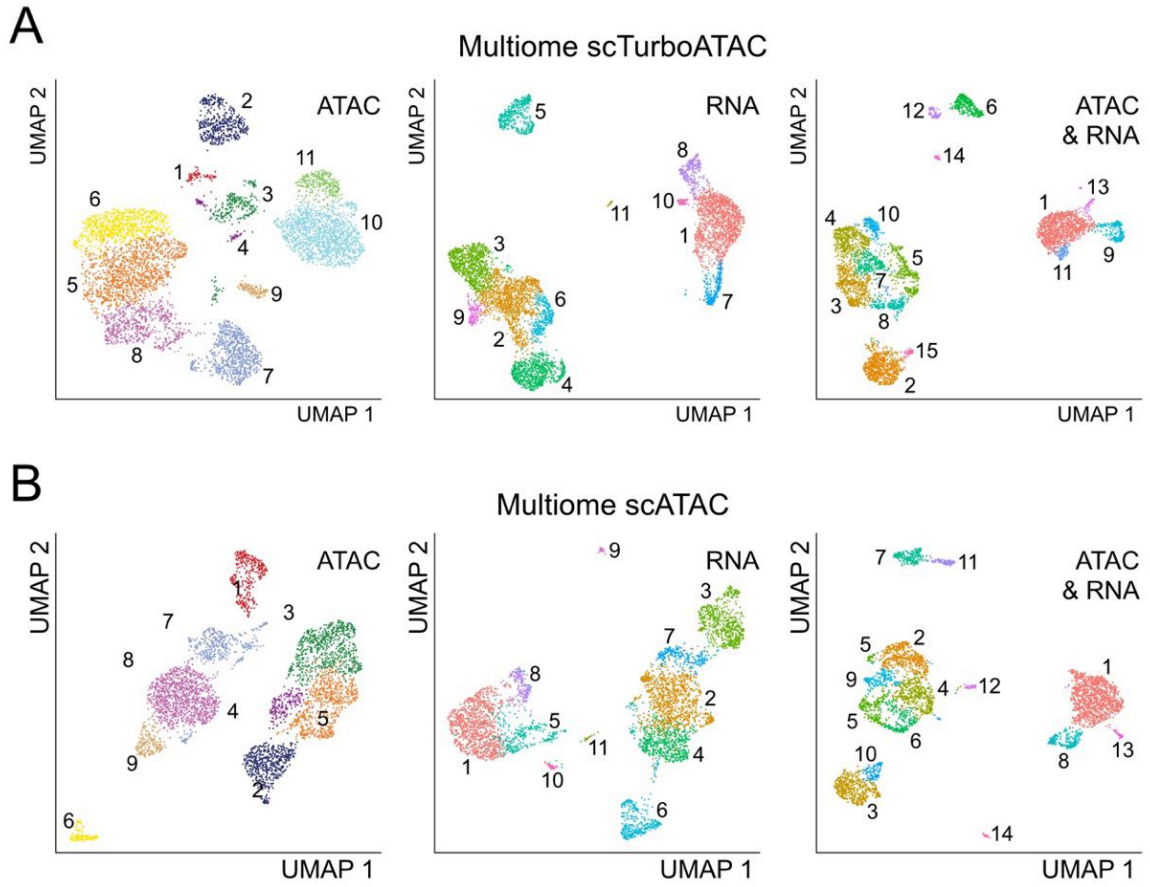


Figure 2.10 Data complexity of Multiome scATAC-seq and Multiome scTurboATAC-seq protocols from human PBMCs. A UMAP embedding of Multiome-scTurboATAC (left), corresponding Multiome-scRNA (middle) and co-embedding of Multiome-scTurboATAC and -scRNA (right). The color of points reflects the k-nearest neighbor cluster. **B** Same as panel A for Multiome-scATAC (left), corresponding Multiome-scRNA (middle) and co-embedding of Multiome-scATAC and -scRNA (right). Adapted from Seufert *et al.* (2023).

In summary, we reduced the sparsity of scATAC-seq data by introducing the TurboATAC protocol. The TurboATAC protocol increased library complexity by optimizing Tn5 activity and the selected Tn5 reaction buffer. Higher library complexity reduced the duplicate rate of sequencing reads, ultimately increasing the detected number of unique fragments per single cell. Consequently, the reduced sparsity in scTurboATAC-seq data resolved cell clusters at higher resolution for both single-omic and multi-omic protocols.

2.1.3. Stochasticity of single cell chromatin accessibility

After reducing scATAC-seq data sparsity, as the most challenging technical bias, I aimed to further characterize chromatin accessibility at single cell resolution. As shown before, I observed variability in chromatin accessibility between homogeneous cells (see **Figure 2.7A**). However, whether this was caused by persisting data sparsity or stochastic fluctuations of chromatin accessibility remained uncertain. In the following, I used scATAC- and scTurboATAC-seq data from MEFs treated with IFNβ for 6 h to assess whether variation in chromatin accessibility between homogeneous single cells was due to technical dropouts or reflected true biological variation.

Peaks of chromatin accessibility were called from the pseudo-bulks of scATAC- and scTurboATAC-seq data, separately (**Figure 2.11A**). Both samples yielded equal numbers of approx. 140,000 peaks and similar distributions across promoter, exonic, intronic and intergenic regions. For comparison, peaks of chromatin accessibility were called from two bulk ATAC replicates of MEFs treated with IFNβ for 6 h (**Figure 2.11A**). The number of total peaks detected was highly variable between the two bulk ATAC replicates. While bulkATAC_rep1 resulted in roughly 140,000 peaks and was similar to the single cell peak sets, bulkATAC_rep2 yielded almost twice as many peaks (roughly 230,000). Notably, both bulk ATAC replicates called comparable numbers of promoter and exonic peaks as the single cell samples. However, bulkATAC_rep2 showed much higher numbers of intronic and intergenic peaks. When intersecting the genomic positions of single cell and bulk peak sets, 60 % of single cell peaks overlapped with bulk peaks and 65 % vice versa.

To facilitate comparability when assessing peak coverage at single cell resolution, a consensus peak set from the two single cell peak sets and the scATAC-H30 peak set (see **Section 2.1.2**) was generated comprising 202,369 peaks of 2 kb each (peak summits extended by 1 kb in both directions). Data from scTurboATAC-seq showed significantly higher numbers of accessible peaks per cell than scATAC-seq data (**Figure 2.11B**). The number of accessible peaks per cell showed a strong relation to the number of unique

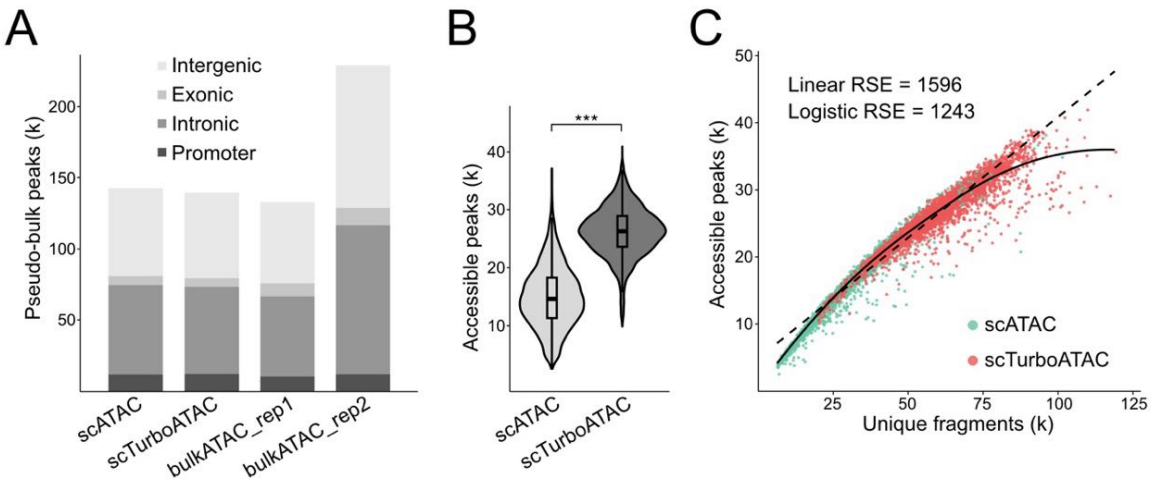


Figure 2.11 Peak coverage of accessibility signal in bulk ATAC-seq, scATAC-seq and scTurboATAC-seq data from 6h IFNβ-treated MEFs. A Number of peaks from pseudo-bulk scATAC and scTurboATAC as well as two replicates of bulk ATAC. Peaks are annotated by their genomic location. **B** Number of accessible peaks detected per cell in scATAC and scTurboATAC. Merged peak set of scATAC, scATAC Tn5-H30 and scTurboATAC was used. P-value from two-sided, unpaired Student's t-test is indicated as *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. **C** Number of accessible peaks against number of unique fragments of cells from scATAC (blue) and scTurboATAC (red). Linear and logistic regression lines across both samples are visualized as dashed and black lines, respectively. Residual standard errors (RSEs) of linear and logistic models are reported. Adapted from Seufert *et al.* (2023) and extended.

fragments per cell (**Figure 2.11C**). As expected, cells from scATAC-seq data showed less accessible peaks per cell due to lower numbers of unique fragments. Data points were fitted by linear and logistic regression (**Figure 2.11C**). The lower residual standard error (RSE) of the logistic regression indicated that the best fit of the data was by the logistic regression model. Its predicted plateau at approximately 35,000 accessible peaks per cell might be underestimated as single cells with sufficiently high numbers of unique fragments were missing from the model.

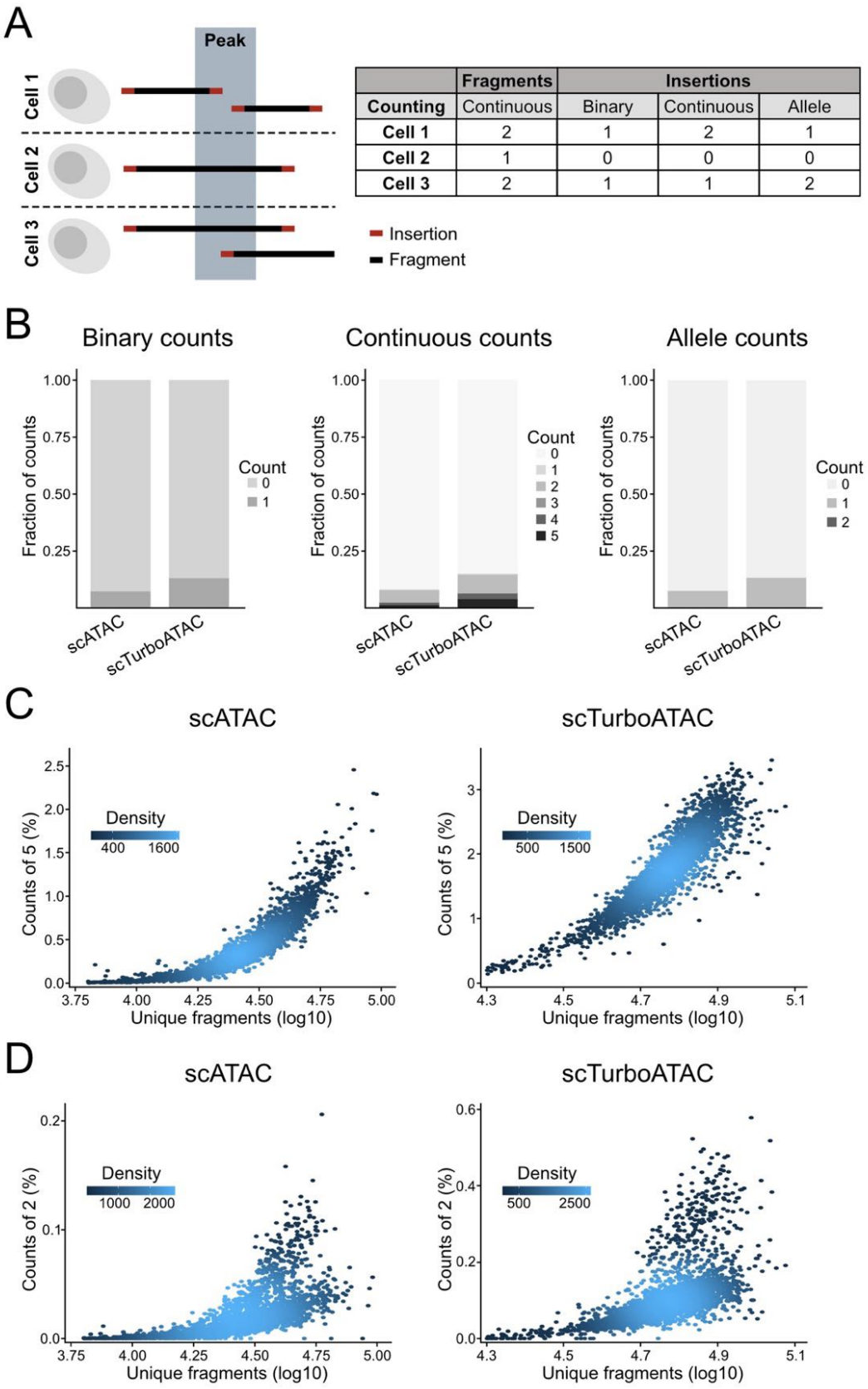
In summary, pseudo-bulk single-cell data recovered similar numbers of ATAC peaks as one of the bulk ATAC replicates. However, only 7.5-15 % of all pseudo-bulk peaks were simultaneously detected as accessible in the single cells. A logistic relation between accessible peaks per cell and number of unique fragments per cell was observed. This indicated that low numbers of accessible peaks per cell were only partially due to data sparsity and persisted even at high single cell sequencing depths, showing that not all pseudo-bulk peaks were simultaneously accessible in single cells. Consequently, chromatin accessibility appeared to be stochastic at single-cell level.

2.1.4. Quantification of scATAC-seq data

Chromatin accessibility is fundamentally binary, since chromatin can either be accessible or inaccessible to Tn5. However, due to the limited resolution of scATAC-seq data, chromatin accessibility is typically not analyzed at genomic base pair resolution. Instead, it is investigated and quantified in pre-defined regulatory regions, such as annotated promoters, enhancers, or data-driven peaks of chromatin accessibility. The quantification of Tn5 insertions in these regulatory regions enables to study their activity at single-cell resolution. In the following, I compare different methods of quantification and investigate to what extent non-binary information is encoded in scATAC-seq data. For this, I used again scATAC- and scTurboATAC-seq data from MEFs treated with IFNβ for 6 h to additionally assess the impact of data sparsity on quantification of scATAC-seq data.

Count matrix design resolves variable non-binary information

Essentially, scATAC-seq data can be quantified by either counting insertions or fragments (**Figure 2.12A**). On the one hand, insertion-based counting quantifies unique sequencing reads, which represent individual binding events of Tn5 molecules to accessible chromatin and their insertion of sequencing primers into the accessible chromatin regions. On the other hand, fragment-based counting quantifies whole fragments between paired sequencing reads These provide information on the distance between the two accessible sites on the same allele, but lack information on the actual chromatin accessibility state between the two sequencing reads. To quantify true accessibility, I created insertion-based count matrices for binary and continuous quantification of the previously established consensus peak set (see **Section 2.1.3**) in single cells. Binary count matrices quantified either no accessibility (0 count) or any accessibility (1 count) in peaks and cells, while continuous count matrices differentiated between no accessibility (0 count) and the actual number of insertions per peak and cell (> 0 count; **Figure 2.12A**). Furthermore, I devised an allele count matrix, which resolved no accessibility (0 count), mono-allelic accessibility (1 count), and bi-allelic accessibility (2 count) in peaks and cells. Here, the number of insertions per peak and cell distinguished inaccessibility (0 count) from accessibility (1 and 2 counts; **Figure 2.12A**). Additionally, overlapping fragments in peaks and cells differentiated mono-allelic (1 count, no overlapping fragments) and bi-allelic accessibility (2 count, overlapping fragments). Allele counting was validated using the X chromosome as a negative control. Here, X-inactivation theoretically allowed only mono-allelic counts. Indeed, all 4,909 peaks on the X chromosome did not show any bi-allelic counts.



	Fragments	Insertions		
Counting	Continuous	Binary	Continuous	Allele
Cell 1	2	1	2	1
Cell 2	1	0	0	0
Cell 3	2	1	1	2

Figure 2.12 Binary, continuous and allele counting of scATAC-seq and scTurboATAC-seq data in peaks from 6 h IFNβ-treated MEFs. Consensus peak set of scATAC, scATAC Tn5-H30, and scTurboATAC was used. **A** Scheme of exemplary fragments in peaks and cells. **B** Distribution of peak counts from binary (left), continuous (middle) and allele (right) counting of scATAC and sc-

Figure 2.12 (continued) TurboATAC. **C** Percentage of continuous peak counts of 5 or greater against number of unique fragments (log10) for cells from scATAC (left) and scTurboATAC (right). The color of points reflects the density of cells. **D** Same as panel C with percentage of bi-allelic peak counts.

All count matrices revealed twofold higher accessibility counts for scTurboATAC- than scATAC-seq data (**Figure 2.12B**). Additionally, continuous counting showed more accessibility counts above 1 for scTurboATAC- than scATAC-seq data. At single-cell resolution, the percentage of high insertion counts exhibited an exponential relationship to the total number of unique fragments per cell (**Figure 2.12C**). The allele count matrices showed only a few occurrences of bi-allelic counts (0.11 % in scTurboATAC-seq data; 0.02 % in scATAC-seq data) (**Figure 2.12B**, right). The percentage of bi-allelic counts also exhibited an exponential relationship to the total number of unique fragments per cell (**Figure 2.12D**). Compared to continuous counting, there were two populations of cells in allele counting. The majority showed a moderate increase in bi-allelic counts with increasing unique fragment numbers. However, a subset of cells displayed exceptionally high bi-allelic counts, following a second, steeper exponential distribution over unique fragment numbers for scATAC- as well as scTurboATAC-seq data.

Chromatin accessibility peaks vary in continuous counts

At peak level, continuous quantification of chromatin accessibility revealed differences between peaks. For both scATAC- and scTurboATAC-seq data, the majority of peaks showed high inaccessibility in more than 90 % of cells (**Figure 2.13A**). The remaining peaks were accessible in a higher fraction of cells. Here, accessibility counts of 1 or 2 as well as ≥3 increased simultaneously. Peaks never exceeded counts of 1 or 2 in more than 40 % of cells, but ≥3 counts were detected in up to 70 % (scATAC-seq) or 100 % (scTurboATAC-seq) of cells. Examining the abundance of ≥ 3 counts in peaks of different genomic annotations revealed two- to fourfold higher fractions of cells with ≥3 counts in promoter peaks compared to intergenic or gene body peaks (**Figure 2.13B**). Furthermore, the genomic sequence of peaks with high cell fractions of ≥ 5 counts showed higher GC content (**Figure 2.13C**).

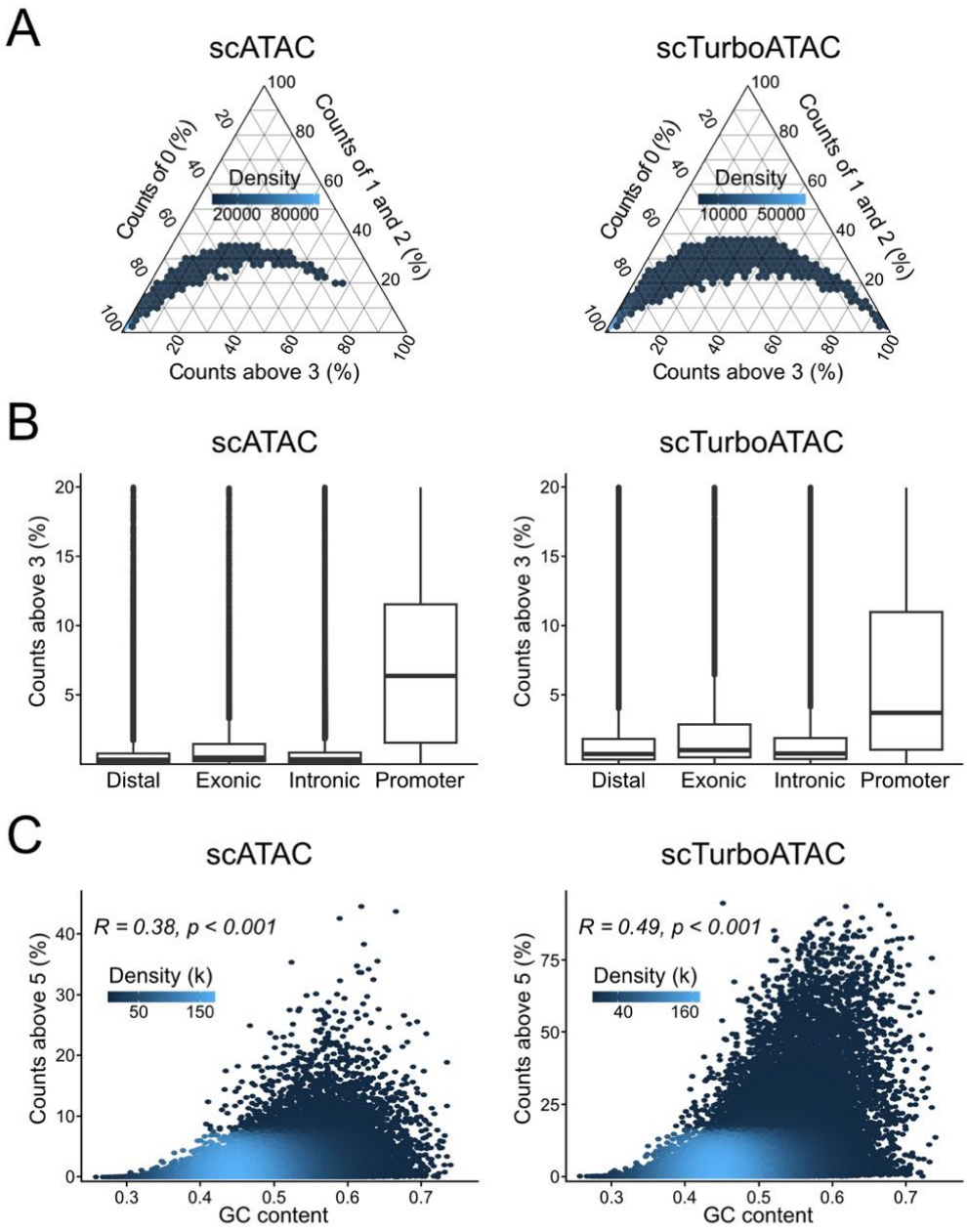


Figure 2.13 Peak investigation by continuous counts of scATAC-seq and scTurboATAC-seq data from 6 h IFNβ-treated MEFs. A Percent of continuous counts of 0 against percent of continuous counts of 1 and 2 and percent of continuous counts ≥ 3 per peak in scATAC (left) and scTurboATAC (right). Hexagons represent the distribution of peaks. The color of hexagons reflects the density of peaks. **B** Percent of continuous counts ≥ 3 in peaks at different genomic locations in scATAC (left) and scTurboATAC (right). **C** Percent of continuous counts ≥ 5 against GC content of peaks in scATAC (left) and scTurboATAC (right). The color of points reflects the density of peaks.

Allele counts reveal sequence-dependent biases in chromatin accessibility

Differences between peaks were less prominent for allelic quantification of chromatin accessibility. As with continuous counting, the majority of peaks showed inaccessibility in more than 90 % of cells for both scATAC- and scTurboATAC-seq data (**Figure 2.14A**). Again, the remaining peaks were accessible in a higher fraction of cells. Counts of mono-allelic accessibility increased in up to 80 % (scATAC-seq) and 100 % (scTurboATAC-seq)

of cells. In contrast, counts of bi-allelic accessibility did not exceed 20 %, except for a single peak in scTurboATAC-seq data with bi-allelic accessibility detected in more than 90 % of cells. The peak had a GC content of 0.45 and was located in an intergenic genomic region. When examining the relationship between bi-allelic counts per peak and its original peak calling scores, the identified peak with exceptionally high bi-allelic counts in scTurboATAC-seq data appeared as an upper outlier in peak calling scores with a score of 3,453 (**Figure 2.14B**). The exceptionally high peak calling score and high percentage of bi-allelic counts of the specific peak might represent a technical artifact and might originate from non-blacklisted repetitive genomic sequences. Regardless, like continuous accessibility counts, promoter peaks showed two- to threefold higher fractions of cells with bi-allelic accessibility than intergenic or gene body peaks (**Figure 2.14C**).

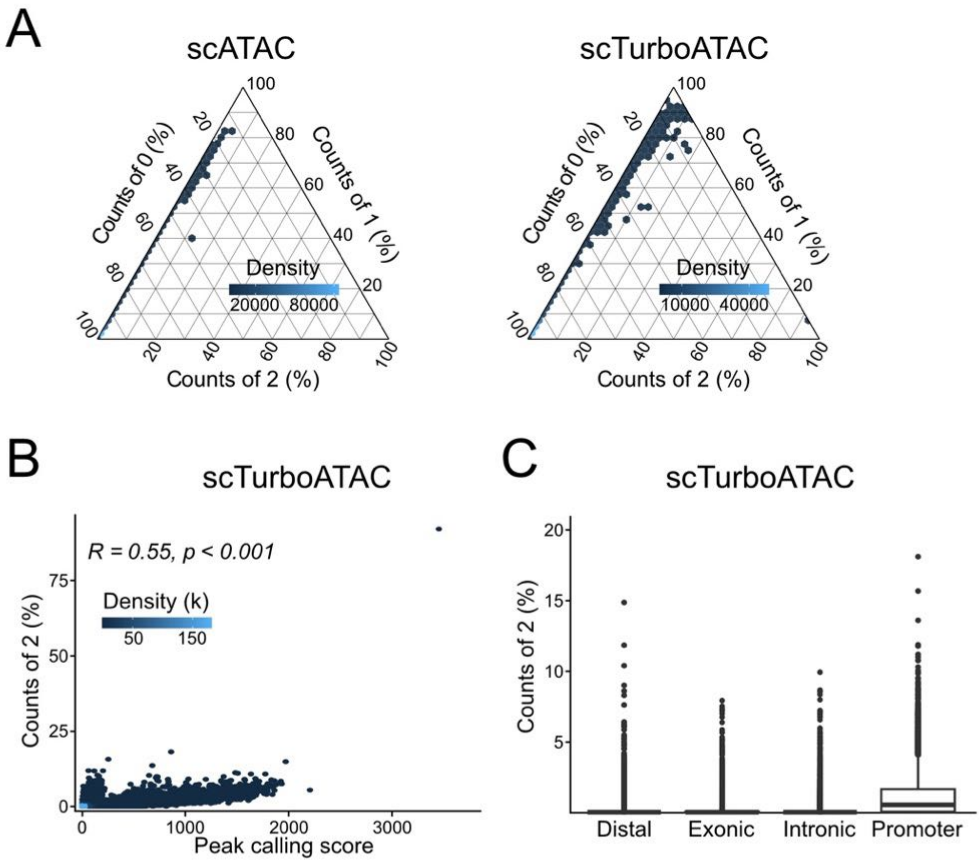


Figure 2.14 Peak investigation by allele counts of scATAC-seq and scTurboATAC-seq data from 6 h IFNβ-treated MEFs. **A** Percent of allele counts of 0 against percent of mono-allelic counts and percent of bi-allelic counts per peak in scATAC (left) and scTurboATAC (right). Hexagons represent the distribution of peaks. The color of hexagons reflects the density of peaks. **B** Percent of bi-allelic counts against peak calling scores of peaks in scTurboATAC. The color of points reflects the density of peaks. **C** Percent of bi-allelic counts in peaks at different genomic locations in scTurboATAC.

In summary, different quantification methods for chromatin accessibility resulted in highly different count matrices, indicating the method's strong impacts. Less sparsity in scTurboATAC-seq data led to higher counts for all methods, indicating improved data quality in scTurboATAC-seq for non-binary analyses. Continuous count matrices showed a clear enrichment of high counts in promoter peaks compared to intergenic and gene body peaks, indicating true biological information encoded in continuous counts. Additionally, high counts were associated with high GC content in the underlying peak sequences, potentially reflecting Tn5 insertion biases. Allele count matrices detected only low levels of bi-allelic counts. This appeared to be partially due to sequencing depth but also suggested that stochastic single-cell chromatin accessibility was unlikely to occur simultaneously on both alleles. Furthermore, there were two cell populations of lower and higher bi-allelic counts, suggesting that allele count matrices might contain additional information on cell state or technical factors such as doublets.

2.1.5. Inferring transcription factor activity at single-cell resolution

TFs play a key role in transcription regulation and varying levels of activity information can be obtained from different sequencing assays. Bulk or single-cell ChIP-seq using TF-specific antibodies can be used to determine the genomic coordinates of sample-specific TF binding sites (see **Section 1.2.2**). Alternatively, TF expression can be obtained from bulk or single-cell RNA-seq. In addition, TF activity can be inferred from the expression of previously annotated TF target genes. Finally, bulk ATAC-seq enables the inference of so-called TF footprints at known genomic TF motif positions. At both bulk and single-cell level, TF binding activity can be deduced from accessibility levels proximal to these known genomic TF motif positions. In the following section, I aimed to provide examples of TF activity assessment from scRNA- and scATAC-seq data (see **Table 2.1** for overview of data sets). I used scRNA-seq data from five AML patients to compare TF expression ,and TF activity, inferred from the expression of annotated TF target genes. Additionally, I used scATAC- and scTurboATAC-seq data from MEFs treated with IFNβ for 6 h to compute TF footprints at pseudo-bulk level and assess the impact of data sparsity. Finally, I used scATAC- and scTurboATAC-seq data from PBMCs to infer TF binding activity at single-cell resolution.

Inferring TF expression and activity from scRNA-seq data

All five examined AML patients possessed translocations at the genomic region of the mixed lineage leukemia (MLL) gene (Schuster *et al.*, 2023). Three patients had translocations resulting in a common MLL fusion with MLLT3 and were thus grouped as the *MLL-MLLT3* patients. Another patient exhibited a frequent fusion of MLL with ELL, referred to as the *MLL-ELL* patient. The final patient had a novel fusion of MLL with an enhancer of the messenger RNA decapping 4 gene (EDC4), termed *MLL-EDC4* patient. In the low-dimensional embedding of the combined scRNA-seq data of all patients, healthy cells from all patients clustered together, whereas AML tumor cells formed patient- and subclone-specific cell clusters (**Figure 2.15A**). Clusters of healthy cells from all patients were annotated as B cells, erythroblasts, monocytes, natural killer (NK) cells, and T cells by marker gene expression (**Figure 2.15B**). AML tumor cell clusters revealed one distinct subclone for patients *MLL-MLLT3#1* and *MLL-EDC4,* and two distinct tumor subclone clusters for patients *MLL-MLLT3#2*, *MLL-MLLT3#3* and *MLL-ELL* (**Figure 2.15B**). The expression of the most differentially expressed genes clearly separated *MLL-EDC4* tumor cells from all other tumor cells (**Figure 2.15C**). Similarly, TF activity of the most differentially active TFs separated the *MLL-EDC4* tumor cell cluster from other patient tumor cell clusters (**Figure 2.15D**). Here, TF activity was inferred from the expression of annotated TF target genes. Additionally, *MLL-ELL* tumor cell clusters exhibited divergent TF activities compared to *MLL-MLLT3* tumor cells, which was not evident from differential expression. Notably, from the most differentially active TFs only MYC and MYB were identified as part of the most differentially expressed genes (**Figures 2.15C+D**, black rectangles).

When comparing the expression and activity of the two exemplary TFs MYC and MYB, no direct relationship between TF expression and the expression of their target genes was observed. The TF MYC showed slightly higher relative expression in tumor cells of patient *MLL-EDC4* and particularly lower relative expression in tumor cells of patients *MLL-MLLT3#3* and *MLL-ELL* (**Figure 2.15C**, black rectangle). In contrast, its relative activity was much higher in tumor cells of patient *MLL-EDC4*, both high and low in tumor cell clusters of patient *MLL-ELL,* and reduced in all tumor cells of *MLL-MLLT3* patients (**Figure 2.15D**, black rectangle). Conversely, the TF MYB exhibited slightly higher relative expression in tumor cells of patient *MLL-EDC4* and subclone clusters 2 of patients *MLL-MLLT3 #2* and *#3* (**Figure 2.15C**, dashed rectangle). Consistently, relative MYB activity was higher in the same tumor cell clusters of patient *MLL-EDC4* and subclone clusters 2 of patients *MLL-MLLT3 #2* and *#3* (**Figure 2.15D**, dashed rectangle).

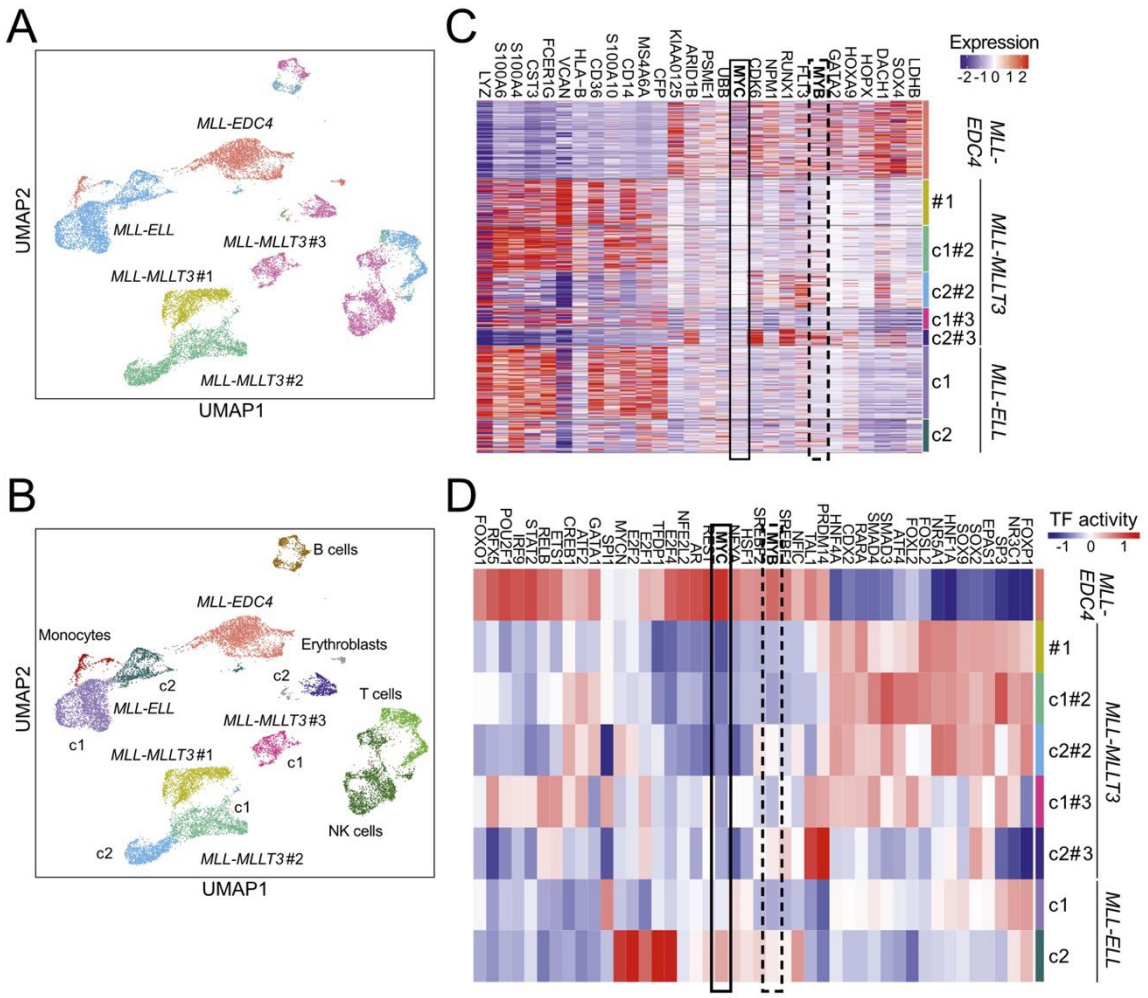


Figure 2.15 TF expression and activity in scRNA-seq data from AML patients. Data of five AML patients are shown. **A** UMAP embedding with the color of points reflecting patient sample. **B** Same as panel A with the color of points reflecting annotated k-nearest neighbor clusters. **C** Single-cell expression of most differentially expressed genes between AML cell clusters. **D** Pseudo-bulk activity of most differentially active TFs between AML cell clusters. TF activity was inferred from the expression of annotated TF target genes. Adapted from Schuster *et al.* (2023).

scTurboATAC-seq enhances the detection of TF binding at pseudo-bulk level

TF footprints were computed from the pseudo-bulks of scATAC- and scTurboATAC-seq data of 6 h IFNβ-stimulated MEFs. Insertions at TF binding motifs and their surrounding genomic sequences were averaged across all motif occurrences in ATAC peaks. The STAT1 and CCCTC-binding factor (CTCF) footprints showed strongly depleted accessibility signal at the motif positions in both scATAC- and scTurboATAC-seq data (**Figure 2.16**), indicating TF binding and thereby protecting the motifs from Tn5 insertions during ATAC-seq. Here, the less sparse scTurboATAC-seq data demonstrated slightly better enrichment and accessibility imprints at STAT1 and CTCF motifs. Moreover, accessibility was moderately increased in the surrounding 50-75 bp of STAT1 motifs, suggesting gradual repositioning of nucleosomes to facilitate STAT1 binding (**Figure 2.16**, left). In contrast, CTCF footprints showed increased accessibility in their surrounding 100

bps, reflecting the removal of one nucleosome to enable CTCF binding (**Figure 2.16**, right).

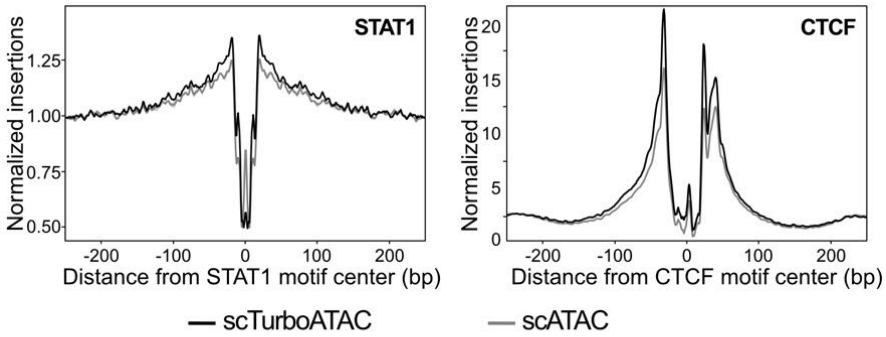


Figure 2.16 TF binding footprints in scATAC-seq and scTurboATAC-seq data from 6 h IFNβ-treated MEFs. Chromatin accessibility footprints at STAT1 (left) and CTCF binding motifs (right) from scATAC and scTurboATAC. Adapted from Seufert *et al.* (2023).

Inferring TF binding activity from scTurboATAC-seq data improves the resolution of cellular heterogeneity

Lastly, TF binding activity was computed at single-cell resolution from scATAC- and scTurboATAC-seq data of human PBMCs. Low-dimensional embeddings of the B cell subpopulations of PBMCs revealed three and four B cell clusters in scTurboATAC- and scATAC-seq data, respectively (**Figure 2.17A**). The integrated low-dimensional embedding showed that cluster C1 in scTurboATAC-seq data was not co-embedded with any scATAC-seq B cell cluster and therefore not resolved by scATAC-seq data (**Figure 2.17B**). TF binding activity was inferred at single-cell resolution to investigate variations in B cell state between the detected B cell clusters. Differential TF binding activities between the B cell clusters from both scTurboATAC- and scATAC-seq data were observed for purine-rich box 1 (PU.1) and IRF complexes, octamer-binding transcription factor 2 (OCT2) and activating transcription factor 3 (ATF3), among others (**Figures 2.17C+D**). B cell clusters C2 and C3 from scATAC, as well as corresponding clusters C3 and C4 from scTurboATAC-seq data, showed the highest relative TF binding activity of PU.1-IRF complex (**Figures 2.17C+D**, left). Conversely, B cell clusters C1 from scATAC- and C2 from scTurboATAC-seq data showed higher relative OCT2 binding activity than other B cell clusters (**Figures 2.17C+D**, middle). Furthermore, B cell cluster C1 from scTurboATAC-seq data showed higher relative ATF3 binding activity than other scTurboATAC- and scATAC-seq B cell clusters (**Figures 2.17C+D**, right). The analysis of TF binding activity revealed a specific B cell state characterized by high ATF3 binding activity and low PU.1-IRF and OCT2 binding activities in B cell cluster C1 from scTurboATAC-seq data.

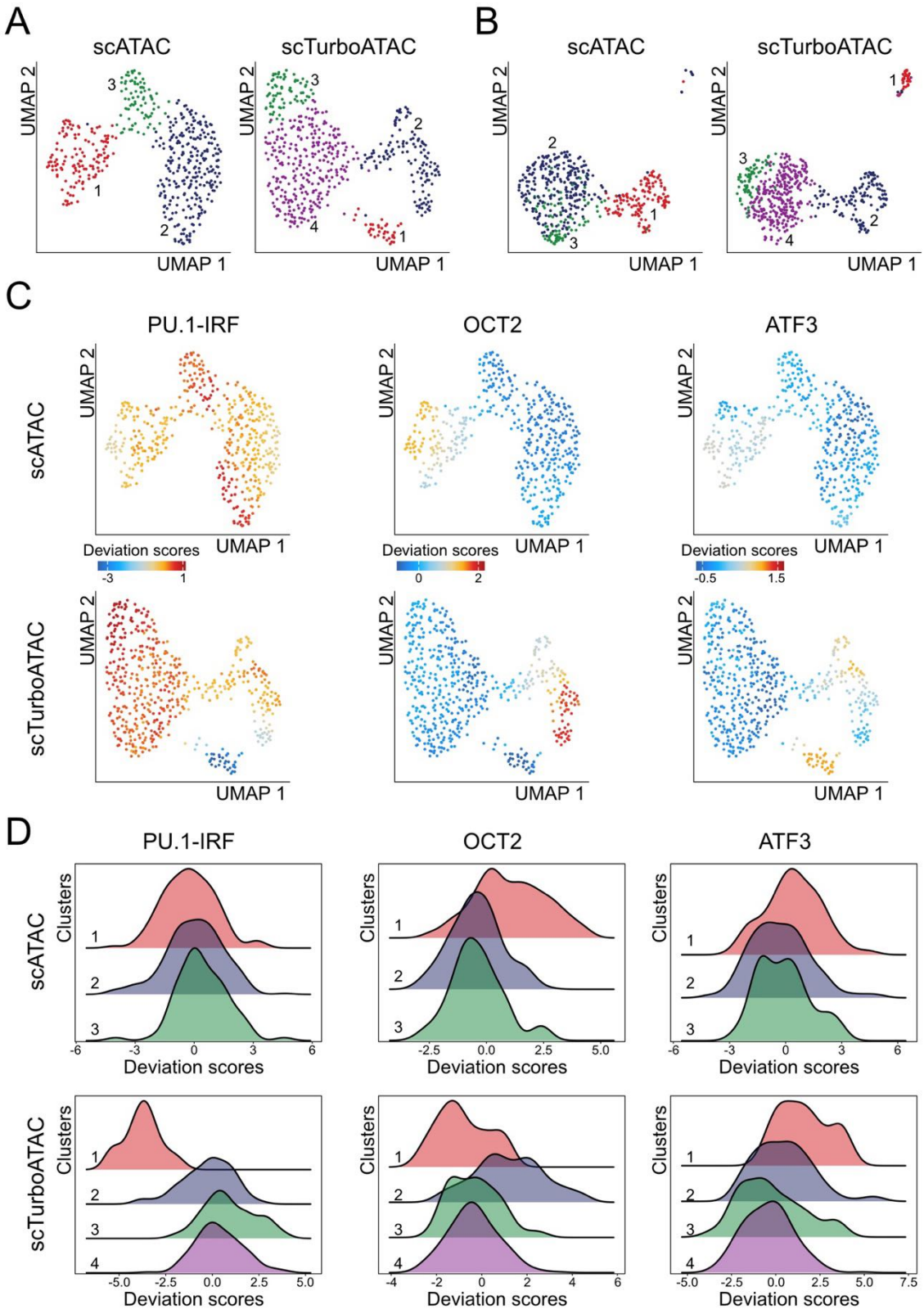


Figure 2.17 B cell heterogeneity in scATAC-seq and scTurboATAC-seq data from human PBMCs. A Low-dimensional embeddings of B cells from scATAC (left) and scTurboATAC (right). The color of points reflects the k-nearest neighbor cluster. **B** Integrated low-dimensional embedding of B cells split by scATAC (left) and scTurboATAC (right). The color of points reflects the sample-specific k-nearest neighbor cluster. **C** Same as panel A with color of points reflecting relative and imputed TF binding activity of PU.1-IRF (left), OCT2 (middle), and ATF3 (right) for scATAC (top) and scTurboATAC (bottom). Relative TF binding activity inferred from TF motif accessibility.

Figure 2.17 (continued) D Non-imputed relative TF binding activity of PU.1-IRF (left), OCT2 (middle), and ATF3 (right) in clusters of scATAC (top) and scTurboATAC (bottom). Relative TF binding activity inferred from TF motif accessibility. Adapted from Seufert *et al.* (2023).

To conclude, TF activity was studied using different sequencing assays at (pseudo-)bulk and single-cell resolution. TF activities varied in the information encoded using (i) the measurement of TF expression itself by RNA-seq, (ii) the inference of TF activity from measuring expression of annotated TF target genes by RNA-seq, (iii) the measurement of TF binding to chromatin by TF-specific ChIP-seq, and (iv) the inference of TF binding to accessible chromatin from ATAC-seq. All provided reasonable estimates of TF activity in the biological systems studied but showed that they resolve different molecular layers of TF activity.

In summary, in this chapter I identified data sparsity as the key challenge in scATAC-seq data analysis (**Section 2.1.1**) and overcame this challenge by introducing the TurboATAC protocol (**Section 2.1.2**). This allowed me to investigate the stochastic nature of chromatin accessibility among single cells (**Section 2.1.3**) and determine the non-binary information content encoded in this data (**Section 2.1.4**). Lastly, I studied the applicability of inferring TF activity from transcription and chromatin accessibility at single-cell resolution (**Section 2.1.5**). Overall, this led to an advancement of the experimental as well as computational analysis of scATAC-seq data.

2.2. Developing a computational framework to dissect the molecular mechanisms underlying chromatin co-accessibility

In this chapter, I address the second aim of this thesis: Developing a computational framework to compute, visualize and interpret chromatin co-accessibility, ultimately dissecting the molecular mechanisms causing chromatin co-accessibility. To achieve this, I used scTurboATAC-seq data from HUVECs, since they are primary, non-immortalized human cells and thus a suitable model system to study unaltered chromatin and transcription. HUVECs were analyzed untreated and upon treatment with TNFα for 30 or 240 min (**Table 2.9**). As described in **Section 1.3.2**, extracellular TNFα can bind to TNF receptors in HUVEC cell membranes, intracellularly activating the transcription factor NF-κB. Consequently, NF-κB translocates to the nucleus, binds to specific DNA motifs, and regulates gene expression. For all TNFα treatment time points, data were obtained from three independent biological replicates, facilitating an in-depth investigation of chromatin co-accessibility and the assessment of reproducibility in the co-accessibility analysis. Furthermore, the perturbation of HUVECs by TNFα treatment allowed to study induced changes in chromatin co-accessibility. In addition to the scTurboATAC-seq data of HUVECs in three biological replicates, I utilized scATAC- and scTurboATAC-seq data of 6 h IFNβ-treated MEFs from **Chapter 2.1** (see **Table 2.1**, **Figure 2.7**). This direct comparison of scATAC-seq and scTurboATAC-seq data allowed me to investigate the impact of data quality on various aspects of co-accessibility analysis.

Table 2.9 Overview of sequencing data sets used for developing a computational framework to dissect the molecular mechanisms underlying chromatin co-accessibility.

Sample	Perturbation	Sequencing type	Assay type
MEF	IFNβ treatment	Single-cell	ATAC, TurboATAC
HUVEC	TNFα treatment	Single-cell	TurboATAC
		Bulk	HiC, Histone ChIP (H3K27ac)

Most of the presented results were published in Seufert *et al.* (2023) and Seufert *et al.* (2024). Acquisition of scTurboATAC-seq data from HUVECs was conducted by Irene Gerosa and Sabrina Schumacher (both Division of Chromatin Networks, German Cancer Research Center, Germany). Bulk HiC-seq data from HUVECs were obtained from GEO (accession number: GSE63525) and computational analysis was conducted by Vassiliki

Varamogianni-Mamatsi (Institute of Pathology, University Medical Center Göttingen, Göttingen, Germany). Bulk ChIP-seq data from HUVECs were acquired by the labs of Argyris Papantonis (Institute of Pathology, University Medical Center Göttingen, Göttingen, Germany) and Petros Kolovos (Department of Molecular Biology & Genetics, Democritus University of Thrace, Greece). Computational analysis was conducted by Panagiotis Liakopoulos (Department of Molecular Biology & Genetics, Democritus University of Thrace, Greece). My contribution comprised the computational analysis of the scTurboATAC-seq data from HUVECs and the integration of bulk HiC- and ChIP-seq data. I conceptualized, developed and applied the co-accessibility analysis framework, termed RWireX. It is based on RWire (Mallm *et al.*, 2019) and implemented as an extension to the ArchR software package (Granja *et al.*, 2021) building on its existing functionalities. Coding and implementation were supported by Anastasiya Vladimirova (formerly Division of Chromatin Networks, German Cancer Research Center, Germany).

2.2.1. Inferring chromatin co-accessibility from scATAC-seq data with RWireX

Computing the co-accessibility of genomic regions has been widely used to infer regulatory interactions of distal chromatin sites. However, the understanding of the underlying molecular mechanisms leading to these co-accessibility patterns remains limited. Using insights gained from **Chapter 2.1**, I developed a computational framework for co-accessibility analysis to resolve different layers of co-accessibility (**Figure 2.18**). The framework comprises two workflows: The *single cell co-accessibility workflow* infers co-accessible regions from a snapshot of stochastic accessibility changes among uniform single cells. In contrast, the *metacell co-accessibility workflow* identifies broader domains of enriched co-accessibility from cell state-dependent accessibility changes of aggregated metacells. The design of the RWireX co-accessibility workflows is described in more detail below.

The *single cell co-accessibility workflow* computes Pearson correlation on continuous accessibility count matrices of ATAC peaks and single cells (**Figure 2.18**, left). It requires a homogeneous cell population as input to resolve stochastic co-accessibility changes, which would be dominated by cell state variations in heterogeneous cell populations. Pearson correlation coefficients are compared against a local background model from shuffled count matrices, revealing autonomous links of co-accessibility (ACs). The average fraction of accessible cells in the linked peaks of an AC determines the detection rate among the single cells. ACs are visualized by loops between the linked ATAC peaks

on the genomic coordinate, where the loop color reflects the strength of co-accessibility and the loop height indicates the AC's detection rate. Finally, RWireX computes AC activity scores per cell from the multiplied accessibility counts of the AC's start and end peaks. The activity scores can be used to identify sets of ACs that are active in the same cells.

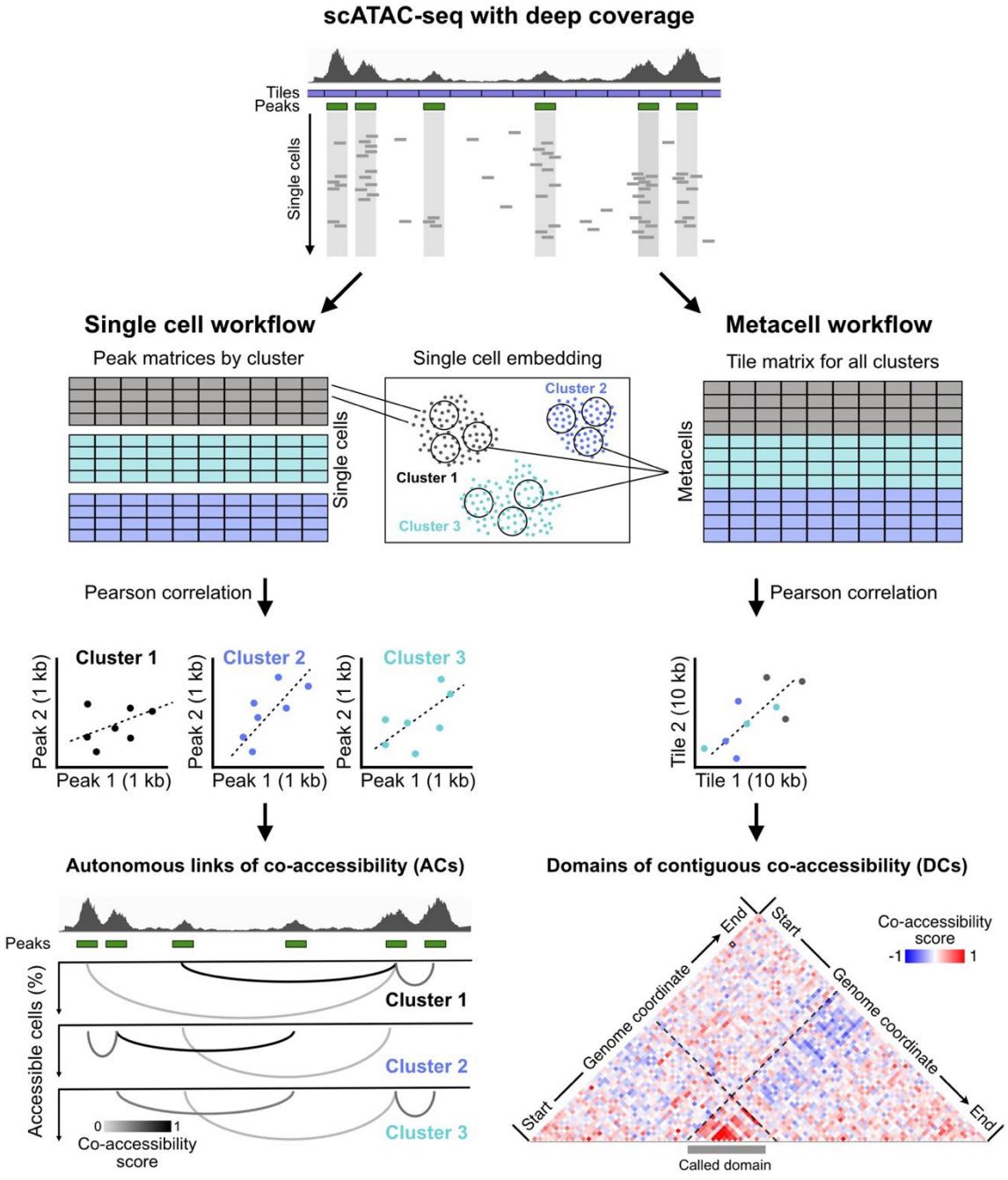


Figure 2.18 RWireX's co-accessibility workflows using scATAC-seq data. Chromatin accessibility is quantified in single cells by continuous insertion-based counts in ATAC peaks or genomic tiles (top). The *single cell workflow* computes Pearson correlation coefficients between ATAC peaks across single cells from a homogeneous cell cluster (middle left). It reveals autonomous links of co-accessibility (ACs) driven by stochastic accessibility changes between single cells (bottom left). The *metacell workflow* correlates accessibility in genomic tiles across metacells, which are formed of cells from multiple cell clusters (middle right). It resolves broad domains of increased and contiguous co-accessibility (DCs) driven by accessibility differences between cell states (bottom right). Adapted from Seufert *et al.* (2024).

The *metacell co-accessibility workflow* computes Pearson correlation on continuous accessibility count matrices of genomic tiles and metacells from aggregated profiles of similar cells (**Figure 2.18**, right). Metacells are generated by aggregating the chromatin accessibility profiles of similar cells from a heterogeneous cell population. The cell diversity in the selected population influences the co-accessibility information, such as varying accessibility patterns determined by different transcription factor activities among cells. *Metacell co-accessibility* is visualized by heatmap, where the color indicates the strength of co-accessibility (positive correlation in red; negative correlation in blue). Here, only half of the mirrored *metacell co-accessibility* matrices are visualized by triangular heatmap and the color at the intersection between two genomic regions depicts their co-accessibility. Furthermore, the *metacell co-accessibility* matrices allow the identification of broader co-accessible domains or patterns, and investigating locally increased co-accessibility identifies domains of contiguous co-accessibility (DCs).

In both co-accessibility workflows, RWireX computes Pearson correlation coefficients between continuous accessibility counts of two genomic regions. These Pearson correlation coefficients are used as so-called *co-accessibility scores* to describe the degree of co-accessibility between two genomic regions. Pearson correlation was selected instead of more complex machine learning-based regression models to maintain explainability, computational efficiency and robustness in this exploratory analysis. In other use cases, or as our understanding of chromatin co-accessibility improves, more complex models might be employed. For example, I applied a more advanced machine learning-based classification approach to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis based on multi-omic data (Wu *et al.*, 2023).

To make the computational framework available for the scientific community, I developed a software package termed RWireX, which includes both the *single cell* and *metacell co-accessibility workflows*. RWireX is available and maintained on Github (https://github.com/RippeLab/RWireX). The Github repository contains the R code, installation guidelines, test data, and vignettes for both the *single cell* and *metacell co-accessibility workflows*. The following paragraphs provide further information on critical features of RWireX's co-accessibility workflows.

Quantification of chromatin accessibility signal

The methods used to quantify accessibility signals in genomic regions strongly influence the information content and count matrix-based downstream analysis of scATAC-seq data, as demonstrated in **Section 2.1.4** for binary, continuous, and allelic quantification of insertions. Consequently, the selected quantification method is a crucial parameter for co-

accessibility analysis. Using the *single cell workflow*, I computed co-accessibility scores from binary, continuous and allelic count matrices of scTurboATAC-seq data to investigate the influence of quantification methods on co-accessibility analysis. Here, co-accessibility scores from a binary count matrix ranged between -0.15 and 0.55, while those from a continuous count matrix ranged from -0.1 to 1.0 (**Figure 2.19A**). These values showed high concordance (correlation coefficient of 0.79), with a fraction of data points lying on the diagonal, indicating high consensus between co-accessibility scores of the two count matrices. However, a significant fraction of data points was above the diagonal, showing distinctly higher co-accessibility scores from the continuous count matrix. For the allelic count matrix, co-accessibility scores ranged from -0.1 to 0.4 and showed high concordance (correlation coefficient of 0.79) with co-accessibility scores from the continuous count matrix (**Figure 2.19B**). Again, a fraction of data points precisely followed

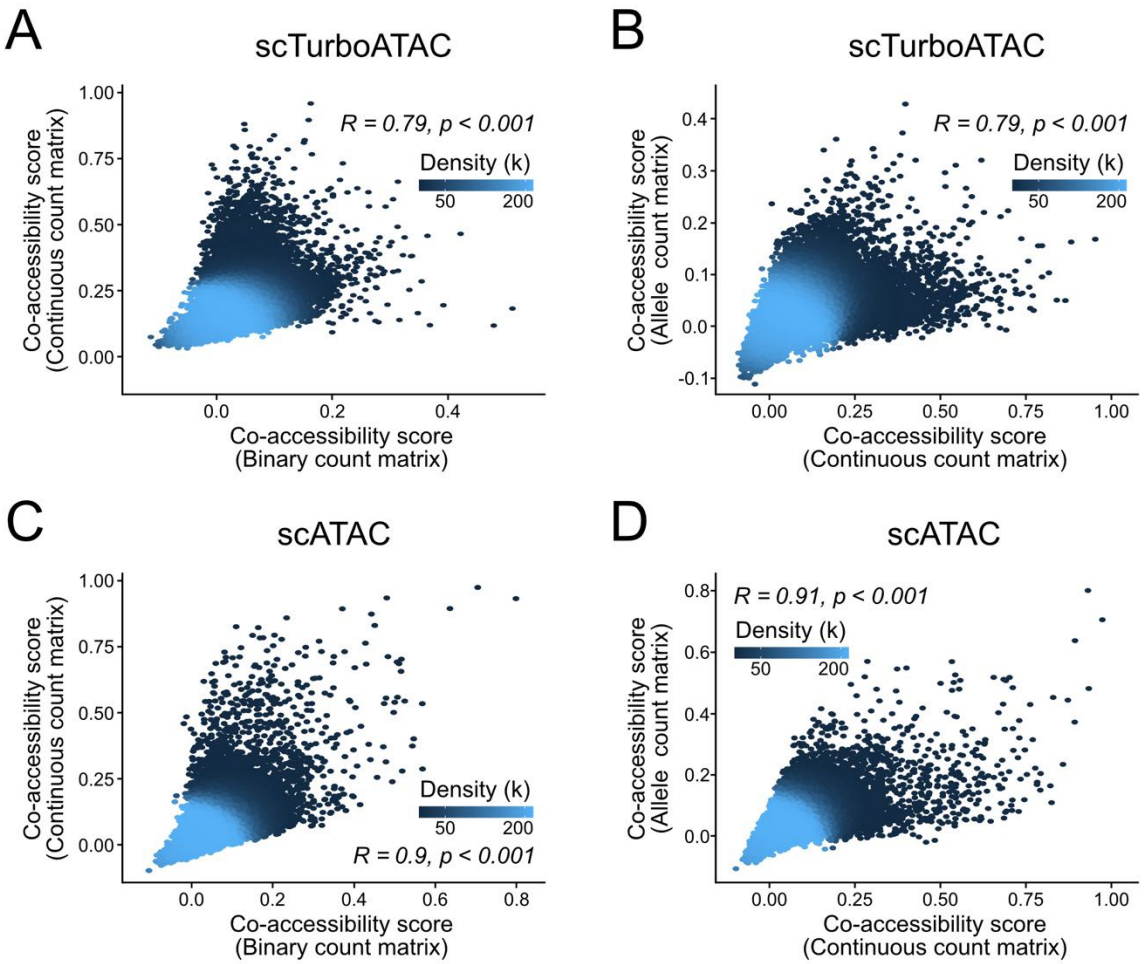


Figure 2.19 Co-accessibility analysis using different peak count matrices of scATAC-seq and scTurboATAC-seq data from 6 h IFNβ-treated MEFs. A Co-accessibility scores from continuous and binary count matrices of scTurboATAC-seq data. Color of points reflects the density. Pearson correlation is annotated. **B** Same as panel A for allelic and continuous count matrices. **C** Same as panel A for scATAC-seq data. **D** Same as panel C for allelic and continuous count matrices.

the diagonal, while the remaining data points showed higher co-accessibility scores for continuous counts.

Co-accessibility scores from binary, continuous, and allelic count matrices of scATAC-seq data revealed the same trends of higher scores from continuous count matrices (**Figures 2.19C+D**). Notably, lower quality of scATAC-seq data led to higher co-accessibility scores from binary and allelic count matrices (values up to 0.8 for both). This was likely due to higher false-negative zero counts (drop-outs), which might falsely increase co-accessibility. Additionally, scATAC-seq data showed a higher correlation of approximately 0.9 between co-accessibility scores from binary, continuous, and allelic count matrices. Consequently, the benefit of continuous quantification was present but less substantial in scATAC-seq data of lower quality.

In summary, the use of different methods for quantification of chromatin accessibility in genomic regions influences the results of co-accessibility analysis. Continuous quantification generally resulted in higher co-accessibility scores than other methods. Additionally, it revealed a fraction of data points with exceptionally higher co-accessibility scores. To resolve these specific co-accessible links, continuous quantification is essential. The influence of quantification methods is more substantial for higher quality data, as shown here and in **Section 2.1.4**. Consequently, I selected continuous quantification of scATAC-seq signal for the design of the co-accessibility workflows.

Compensation of potential biases between samples

Previously, single-cell sequencing depth and the resulting level of data sparsity were identified as key parameters for assessing the quality of scATAC-seq data (**Section 2.1.1**). To evaluate potential biases from data quality in the co-accessibility analysis, co-accessibility scores from continuous count matrices of scATAC- and scTurboATAC-seq data were compared (**Figure 2.20A**). Higher quality of scTurboATAC-seq data showed higher positive co-accessibility scores and a less pronounced peak at co-accessibility scores of zero. Consequently, I reduced biases in the number of cells and the number of unique fragments, as a measure of single-cell data sparsity, to facilitate the comparability of co-accessibility results between samples. For example, the nine scTurboATAC-seq samples from TNFα treatment of HUVECs showed high variability in data sparsity for two samples (**Figure 2.20B**). Additionally, cell numbers of the samples ranged between 1,577 and 6,152. I randomly selected 1,000 cells from the 240 min TNFα-treated HUVEC Rep1 sample as a reference, as it showed a considerably narrow distribution of unique fragments and was highly comparable to most other samples. Next, I selected 1,000 cells from each sample that showed similar numbers of unique fragments compared to the

reference cells (**Figure 2.20C**). This approach resolved differences in cell number and reduced variability in the numbers of unique fragments between samples. However, the approach was insufficient to compensate for extreme differences in data sparsity. For example, experimental issues from cell clogging during the 10x Genomics GEM formation occurred for the untreated Rep2 HUVEC sample, causing a distinctly lower cell number and higher single-cell sequencing depth.

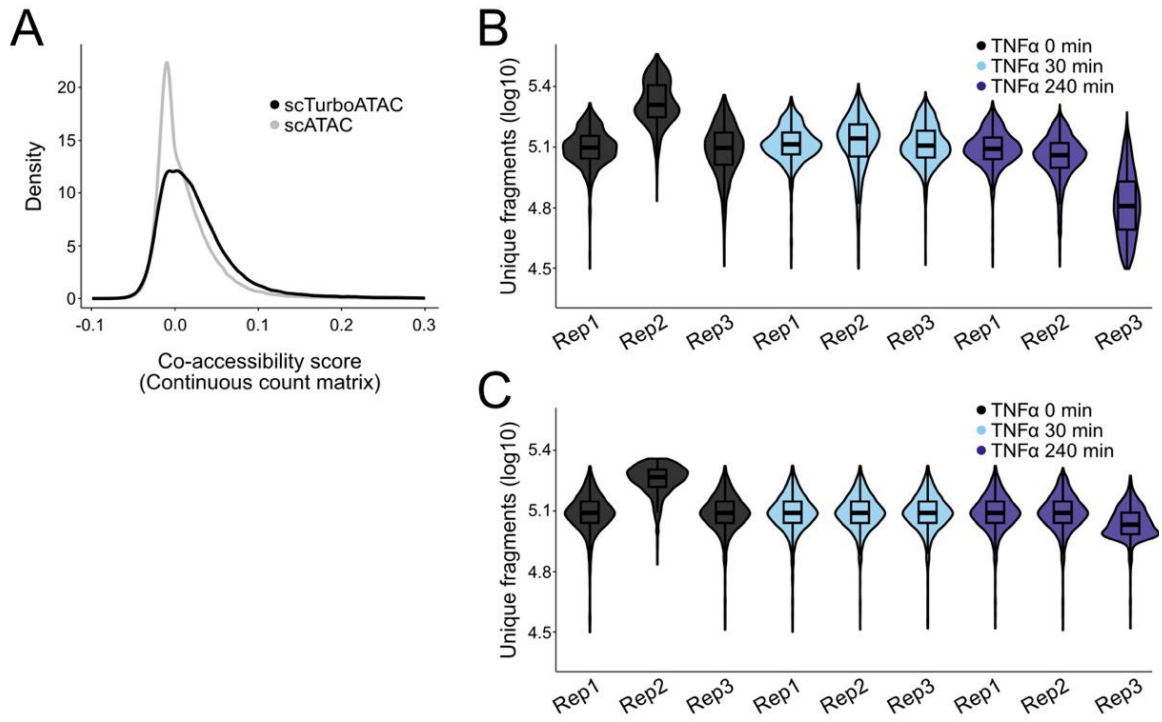


Figure 2.20 Bias compensation for single-cell sequencing depth in co-accessibility analysis. **A** Density distribution of co-accessibility scores from continuous count matrices of scATAC- (grey) and scTurboATAC-seq (black) data from 6 h IFNβ-stimulated MEFs. Limits of the x-axis are set to -0.1 and 0.3. **B** Number of unique fragments per cell for scTurboATAC-seq data from untreated, 30 min, and 240 min TNFα-treated HUVECs. **C** Same as panel B for 1,000 cells per sample selected for most similar sequencing depth distribution across samples.

Single cell and *metacell co-accessibility workflows*

A key feature of RWireX are the two distinct workflows to perform co-accessibility analysis, aimed at resolving different layers of variability in the accessibility signal. Variation in accessibility signal between cells can originate from various sources: (i) Differences in data quality/sparsity between cells (see **Section 2.1.1**); (ii) varying cell types or cell states, e. g. due to cell cycle or apoptosis, resulting in different patterns of accessible genomic regions (see **Section 2.1.2**); (iii) stochastic nature of chromatin accessibility with independent fluctuations at each genomic locus (see **Sections 2.1.3**, **2.1.4**); or (iv) external or internal perturbations inducing specific intracellular responses, which might influence global or locus-specific accessibility. I developed the *single cell workflow* to resolve co-accessibility from stochastic events and the *metacell workflow* to enrich for co-

accessibility derived from varying cell types or induced by perturbations, while reducing co-accessibility from technical biases.

The RWireX workflows require careful selection of input cells to best facilitate the resolution of the intended layers of co-accessibility. For example, single cell accessibility profiles from the nine scTurboATAC-seq samples from TNFα treatment of HUVECs showed strong differences between treatment time points as well as cell cycle states in their low-dimensional embeddings (**Figures 2.21A-C**). Since the aim of the co-accessibility analysis was to resolve the effects of TNFα treatment, effects of cell cycle states were removed by selecting only cells in the G1 cell cycle state. Additionally, differences in data quality of cells between samples were reduced, as described in the previous section. For example, in Rep1, this resulted in 1,000 cells per sample with accessibility variation mostly driven by treatment time point (**Figure 2.21D**). Cells were aggregated into metacells with similar chromatin accessibility profiles to reduce both stochastic accessibility variation and data sparsity (**Figure 2.21E**). First, an initial cell was selected for each metacell, maximizing the distances between initial cells in the low-dimensional embedding. Next, for each metacell, the closest 9 cells to the initial cell were

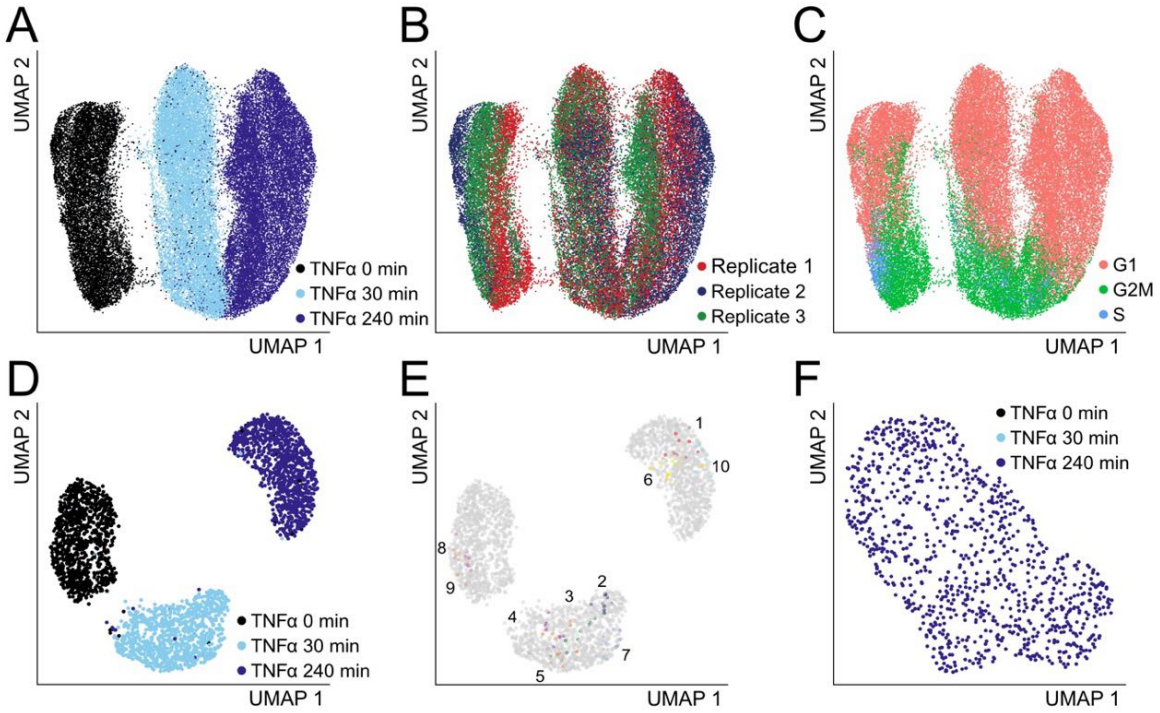


Figure 2.21 Cell populations for co-accessibility analysis of scTurboATAC-seq data from untreated, 30 min, and 240 min TNFα-treated HUVECs. A Low-dimensional embedding of cells from three biological replicates. Color of cells reflects the treatment time point. **B** Same as panel A with color of cells reflecting biological replicate. **C** Same as panel A with color of cells reflecting cell cycle state. **D** Same as panel A for 1,000 selected cells per treatment time point of replicate 1. **E** Same as panel D with color of cells reflecting assignment to metacells 1-10. **F** Same as panel D for only 240 min TNFα treatment time point. Adapted from Seufert *et al.* (2024).

selected based on the low-dimensional embedding, consecutively leaving out the already aggregated cells from further metacell formation. By this, metacells were formed from unique sets of 10 cells each, not using the final 10 % of cells to prevent the forced aggregation of dissimilar cells. These metacells from multiple TNFα treatment time points were used for the *metacell workflow*.

For the *single cell workflow*, homogeneous cells from a single treatment time point were used, selecting only cells in the G1 cell cycle state and selecting 1,000 cells to compensate for technical biases (**Figure 2.21F**). The workflow uses single cells to preserve information on stochastic accessibility changes in the snapshot across many cells. Accessibility of the single cells is quantified in ATAC peaks of high genomic resolution (1 kb for high-quality scTurboATAC-seq data, 2 kb for lower-quality scATAC-seq data), which likely reflect individual CREs. The resulting continuous count matrix from the nine scTurboATAC-seq samples of TNFα treatment in HUVECs shows accessibility counts between 0 and 52 with an inflation of 0 counts (**Figure 2.22A**). In contrast, accessibility is quantified in genomic tiles of lower resolution in the metacell workflow (10 kb for high-quality scTurboATAC-seq data, 20 kb for lower-quality scATAC-seq data). The genomic tiles facilitate unbiased investigation of broader chromatin regions, potentially also reflecting the chromatin state of the regions. The resulting continuous count matrix of genomic tiles shows accessibility counts between 0 and 120 for single cells (**Figure 2.22B**) and 0 and 420 for metacells (**Figure 2.22C**). Genomic tiles and metacells both reduce the number of 0 counts in the count matrix. Co-accessibility of two exemplary peaks from the *single cell workflow* shows the high proportion of co-inaccessibility (simultaneous 0 counts at both peaks, **Figure 2.22D**) in the stochastic accessibility data. In contrast, lower data sparsity in the *metacell workflow* results in co-accessibility from a wider distribution of accessibility counts (**Figure 2.22E**).

In conclusion, the two co-accessibility workflows enrich for different layers of variability in the accessibility signal. They vary in the selection of input cells, where the *single cell workflow* requires homogeneous cells and the *metacell workflow* requires heterogeneous cells, in respect to the variation in cell state or perturbation of interest. The *single cell workflow* uses single cells, while the *metacell workflow* uses aggregated profiles of multiple cells. Finally, the *single cell* and *metacell workflows* quantify accessibility in high-resolution ATAC peaks and lower resolution genomic tiles, respectively. Consequently, the *single cell workflow* is designed to resolve co-accessibility between specific genomic sites from

stochastic accessibility changes, while the *metacell workflow* enriches for co-accessibility in broader genomic regions driven by more global differences in molecular cell states.

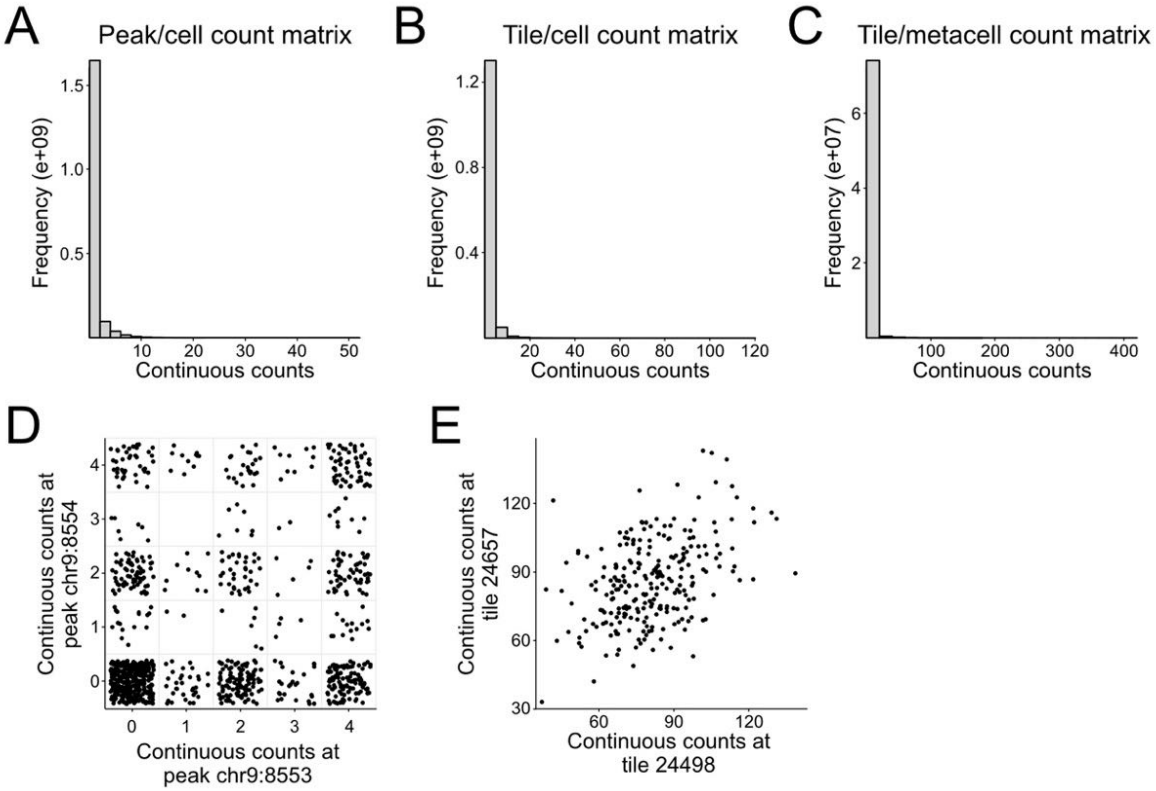


Figure 2.22 Continuous count matrices of peaks and genomic tiles in scTurboATAC-seq data from untreated, 30 min, and 240 min TNFα-treated HUVECs. A Distribution of accessibility counts in ATAC peaks and single cells. **B** Same as panel A for genomic tiles. **C** Same as panel B for metacells. **D** Co-accessibility of exemplary peaks across single cells from untreated HUVECs in replicate 3. Overlaying integers are visualized by jitter plot. **E** Co-accessibility of exemplary genomic tiles across metacells from all treatment time points of replicate 1.

Assessment of background co-accessibility in the *single cell workflow*

In the analysis of TNFα treatment in HUVECs, the *single cell co-accessibility workflow* identified 23,269,378 co-accessible links using the ATAC peak set within a 1 Mb window. To differentiate true co-accessible links from randomly occurring co-accessibility in the background, a local background model was applied. Background co-accessibility was computed by shuffling accessibility count matrices per chromosome by cells and ATAC peaks, following the *single cell workflow* (**Figure 2.23A**). The distributions of background co-accessibility scores were very similar across different samples (**Figure 2.23B**). Potential variations in background co-accessibility scores may arise from differences in cell numbers or data sparsity within a sample. For all samples, background co-accessibility exhibited mean scores of 0.0, with an extended right tail. For each sample, the 99th percentile of background co-accessibility scores was used as the lower cutoff for identifying true co-accessible links. This resulted in 273,813 to 329,820 co-accessible links

above background co-accessibility per sample, termed autonomous links of co-accessibility (ACs).

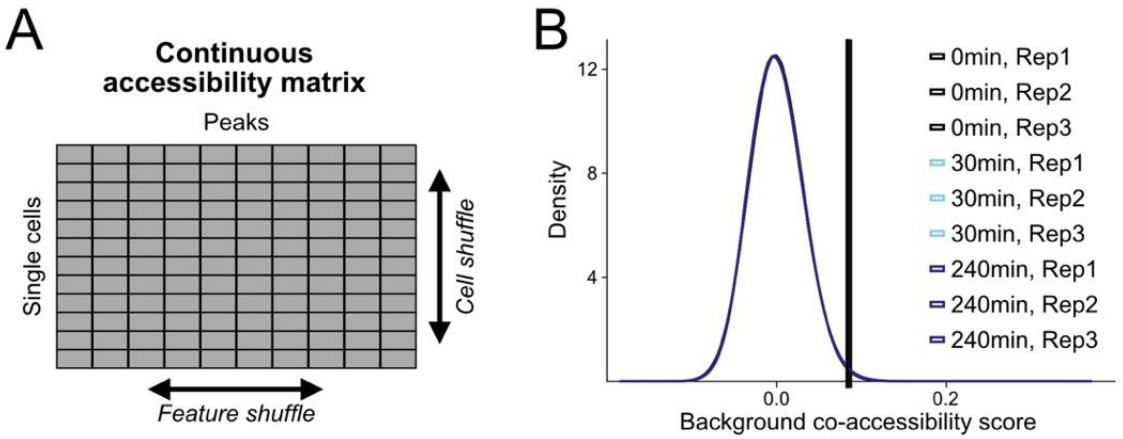


Figure 2.23 Assessment of background co-accessibility to filter for true-positive co-accessible links in the *single cell workflow*. A Scheme of the approach to compute background co-accessibility. Background co-accessibility is determined from continuous accessibility counts in peaks and single cells shuffled by cells and peaks. Pearson correlation coefficients of the shuffled matrices are computed. **B** Distribution of background co-accessibility scores in scTurboATAC-seq data from untreated, 30 min, and 240 min TNFα-treated HUVECs. The 99th percentiles of background co-accessibility scores are marked.

Detection rate of ACs in the single cell population

Previously, continuous quantification of scATAC-seq data revealed strong variation in the number of accessible cells per ATAC peak (see **Section 2.1.4**). Consequently, ACs exhibited differences not only in their co-accessibility scores, but also in their detection rate within the single cell population. The detection rate of an AC among single cells was determined by the average fraction of accessible cells at the linked ATAC peaks, referred to as percent accessible cells (PAC). PAC indicates the prevalence of an AC among the single cells.

The PAC can serve as an additional confidence measure for ACs, alongside the co-accessibility score. *Single cell co-accessibility* analysis of scTurboATAC-seq data from 6 h IFNβ-stimulated MEFs identified ACs with PAC values above 75, which were not detected in the corresponding scATAC-seq data (**Figure 2.24A**, red rectangles). Higher data sparsity in scATAC-seq data increased drop-outs, thereby reducing PAC values. Thus, accessibility signals (>0 counts) are more reliable than inaccessibility signals (0 counts), which could result from true inaccessibility or technical drop-outs, making ACs with higher PAC values more reliable. Additionally, the PAC might reflect an additional layer of biological information in co-accessibility analysis, as higher PAC values may indicate greater AC persistence/stability or higher frequency of AC formation. *Single cell co-accessibility* analysis of scTurboATAC-seq data from TNFα treatment of HUVECs

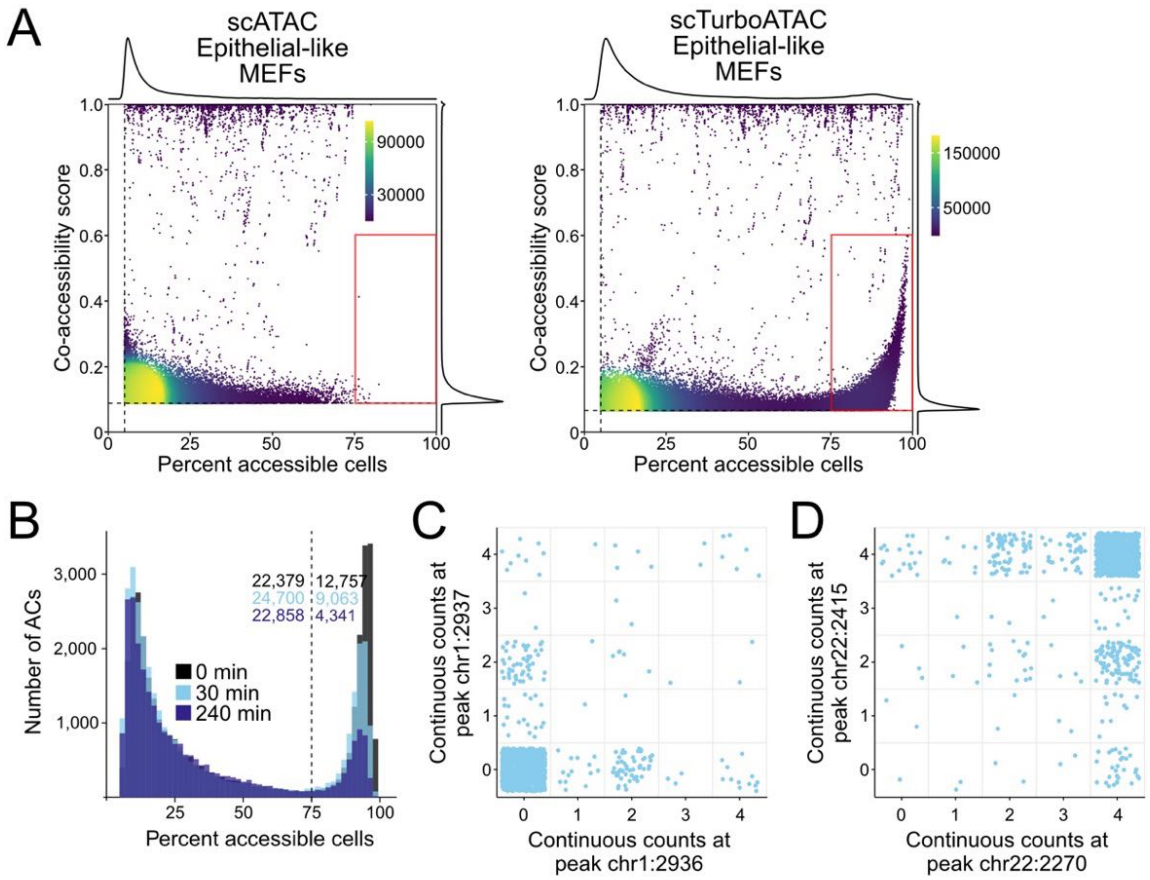


Figure 2.24 Detection rate of autonomous links of co-accessibility (ACs) in single cells. **A** Co-accessibility scores and percent accessible cells (PAC) of ACs in scATAC- (left) and scTurboATAC-seq (right) data from 6 h IFNβ-treated epithelial-like MEFs. The color of points reflects the density of ACs. Density curves of co-accessibility scores and PAC are provided. Dashed lines indicate background co-accessibility cutoff on co-accessibility scores and a minimal 5 % cutoff on PAC. The red rectangles highlight ACs with high PAC. **B** PAC distribution of ACs from scTurboATAC-seq data from untreated, 30 min, and 240 min TNFα-treated HUVECs. **C** Co-accessibility of exemplary ATAC peaks with low PAC from 30 min TNFα-treated HUVECs in replicate 2. Overlaying integers are visualized by jitter plot. **D** Same as panel C for exemplary ATAC peaks with high PAC. Adapted from Seufert *et al.* (2023) and Seufert *et al.* (2024).

revealed a bimodal PAC distribution (**Figure 2.24B**), where most ACs had PAC values either below 25 or above 75. Co-accessibility of ATAC peaks with low PAC values was dominated by simultaneous inaccessibility of the two peaks (**Figure 2.24C**), while co-accessibility of ATAC peaks with high PAC values was driven by simultaneous accessibility (**Figure 2.24D**).

Identification of DCs from *metacell co-accessibility*

The *metacell co-accessibility workflow* revealed broader domains of contiguously enriched co-accessibility using accessibility in 10 kb tiles within a 2 Mb window in the analysis of TNFα treatment in HUVECs (**Figure 2.25A**). These domains were identified genome-wide employing the computational method SpectralTAD, a TAD-calling tool initially designed for HiC-seq data (Cresswell *et al.*, 2020). Instead of using chromatin contact count matrices

from HiC-seq data as intended for SpectralTAD, I applied SpectralTAD to co-accessibility matrices from the *metacell workflow*. To employ SpectralTAD for co-accessibility data, only-positive *metacell co-accessibility* matrices were used, setting all negative co-accessibility scores to zero. SpectralTAD was run to identify small and large domains, separately (**Figure 2.25A**, black and grey regions). While small domains were defined as having a minimal size of 20 kb in a 200 kb window, large domains were defined as at least 200 kb in size within a 2 Mb window. Domain co-accessibility scores were calculated from the average co-accessibility scores within each domain. Most domains showed low average co-accessibility scores below 0.1 (**Figure 2.25B**). However, some domains exhibited high co-accessibility scores of up to 0.8, forming an extended right tail in the distribution. To identify domains with exceptionally high local co-accessibility, the 90th percentile of domain co-accessibility scores was used as a lower cutoff, computed separately for small and large domains. This approach identified approximately 3,500 small and 1,050 large domains with high local co-accessibility, termed domains of contiguous co-accessibility (DCs) (**Figure 2.25A**, red regions). Across all replicates, about 80 % of large DCs overlapped with at least one small DC (**Figure 2.25C**), whereas only roughly 45 % of small DCs overlapped with a large DC.

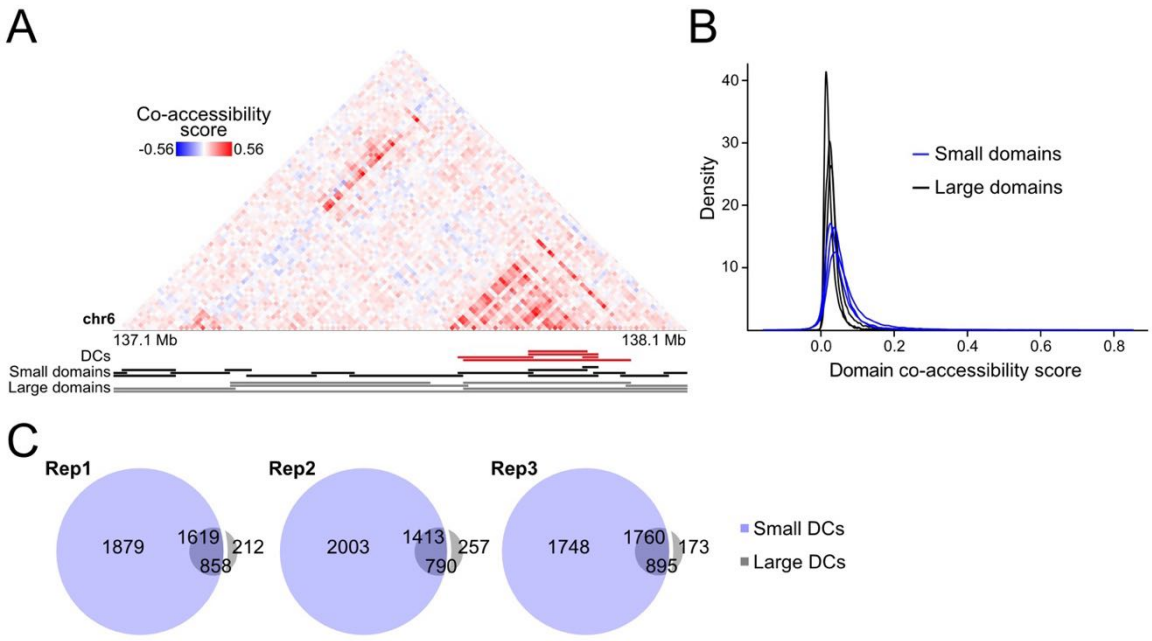


Figure 2.25 Identification of domains of contiguous co-accessibility (DCs) in *metacell co-accessibility* of scTurboATAC-seq data from untreated, 30 min, and 240 min TNFα-treated HUVECs. A *Metacell co-accessibility* in an exemplary region. Small domains (black), large domains (grey), and DCs (red) are annotated. **B** Distribution of average co-accessibility scores in small (blue) and large (black) domains from three biological replicates. **C** Genomic overlap of small and large DCs in three biological replicates.

In summary, RWireX infers chromatin co-accessibility using two workflows that enrich for different layers of variation in chromatin accessibility. The *single cell co-accessibility workflow* identifies autonomous links of co-accessibility from stochastic accessibility fluctuations at distant genomic sites, while the *metacell co-accessibility workflow* resolves broad domains of contiguous co-accessibility from cell state-driven accessibility changes. The workflows differ in their count matrix design, input cell population, genomic and cellular resolution, and the methods applied to identify co-accessibility features. I identified limitations in co-accessibility analysis by data sparsity and technical differences between samples, and addressed these by reducing data sparsity with the TurboATAC protocol and introducing an approach to compensate for technical biases.

2.2.2. Reproducibility of co-accessibility analyses

After developing a computational framework for co-accessibility analysis, which resolves different layers of accessibility variation in chromatin co-accessibility, I aimed to assess the reproducibility of its results. I used three replicates of scTurboATAC-seq data from HUVECs to investigate the robustness of ACs from the *single cell co-accessibility workflow* for each separate sample, as well as DCs from the *metacell co-accessibility workflow* for each replicate.

ACs are reproducible across replicates for TNFα-regulated genes

Single cell co-accessibility was computed for each scTurboATAC-seq sample from HUVECs separately to investigate the reproducibility of ACs. A genomic region around the two TNFα-regulated genes *KLF10* and *GASAL1* was selected as an example to visually inspect the reproducibility of ACs (**Figure 2.26A**), since the two genes showed high numbers of ACs at their promoters. Pseudo-bulk chromatin accessibility profiles were highly similar between replicates (**Figure 2.26A**, top). However, differences in ACs were detected between replicates (**Figure 2.26A**, bottom). Especially high variability was observed for ACs with PACs below 50 and low co-accessibility scores. In contrast, ACs between the genes' promoters were highly reproducible among all samples, showing almost 100 % accessible cells and high co-accessibility scores (around 0.2 for samples from untreated and 240 min TNFα treatment time point; above 0.3 for samples from 30 min TNFα treatment time point). Additionally, ACs between the genes' promoters and two distal H3K27ac peaks were present in 80 % of samples, showing high PAC values between 75-100 (**Figure 2.26A**, bottom: green H3K27ac peaks and blue gene promoters).

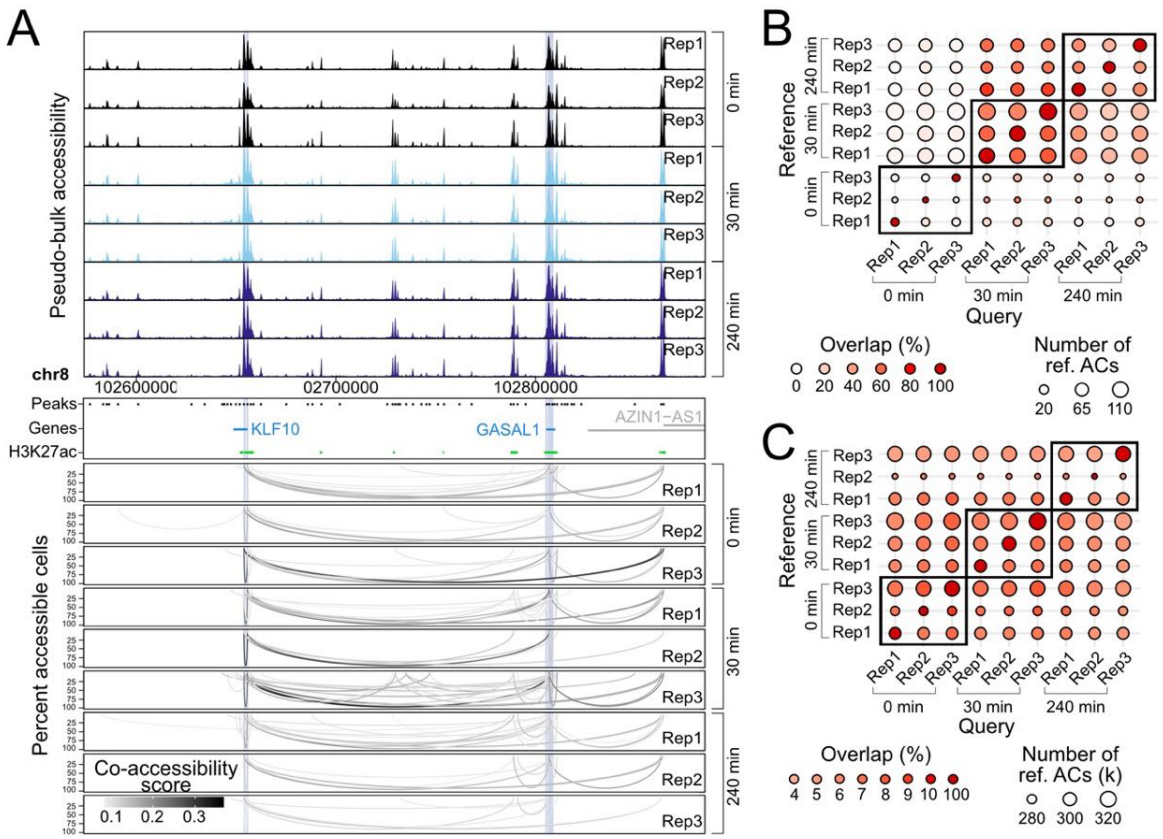


Figure 2.26 Reproducibility of ACs from *single cell co-accessibility* of scTurboATAC-seq replicates from untreated, 30 min, and 240 min TNFα-treated HUVECs. A Chromatin accessibility and ACs from replicates and time points in an exemplary region of TNFα-regulated genes *KLF10* and *GASAL1*. Top: Pseudo-bulk chromatin accessibility tracks; Middle: ATAC peaks (black, 1 kb extended), genes (grey), TNFα-regulated genes (blue) and 1 kb regions around their TSSs (light blue), H3K27ac peaks from ChIP-seq at 30 min time point (green). Bottom: ACs at TNFα-regulated gene promoters. The grayscale and height of loops reflect co-accessibility scores and percent accessible cells of ACs. **B** Comparison of ACs at ten most differential TNFα-regulated genes after 30 min of treatment from replicates and time points. The size and color of the dots reflect the total number of ACs detected in the reference sample and the percent overlap between the samples. **C** Same as panel B for all genome-wide ACs. Adapted from Seufert *et al.* (2024).

To quantify the reproducibility of ACs, the proportion of according ACs between two samples was determined for all combinations of samples. Here, ACs from two samples were classified as consistent if they linked the same two ATAC peaks in both samples, irrespective of their co-accessibility scores and PACs. ACs at the ten most differentially expressed genes after 30 min of TNFα treatment confirmed the previous visual observations of AC reproducibility (**Figure 2.26B**). Only 14-19 ACs were detected in untreated samples, showing overlap below 10 % between replicates (**Figure 2.26B**, black rectangle on bottom left). However, samples after 30 min of TNFα treatment showed strongly increased numbers of ACs (84-107) with approximately 75 % overlap between replicates (**Figure 2.26B**, black rectangle in middle). In samples after 240 min of TNFα treatment, the numbers of ACs (41-54) and their overlap between replicates were again reduced (**Figure 2.26B**, black rectangle at top right). When assessing all genome-wide

ACs, the overlap between replicates was considerably lower (below 10 %) across all treatment time points (**Figure 2.26C**, black rectangles). The overlap between all samples irrespective of treatment time point was equally low.

ACs showed consistent reproducibility across all three replicates, with the lowest number of ACs being present in all three at the same time (**Figure 2.27A**). ACs detected in at least two replicates were used to compile consensus lists of ACs for each treatment time point. These consensus ACs showed less differences between treatment time points in the exemplary region around the two TNFα-regulated genes *KLF10* and *GASAL1* shown before (**Figure 2.27B**). The previously observed ACs between the genes' promoters and from promoters to distal H3K27ac peaks were preserved in the consensus list, becoming more evident with fewer scattered ACs in the vicinity. Additionally, the consensus ACs resolved a distinct increase in co-accessibility scores at the 30 min treatment time point, which were reduced again at the 240 min treatment time point.

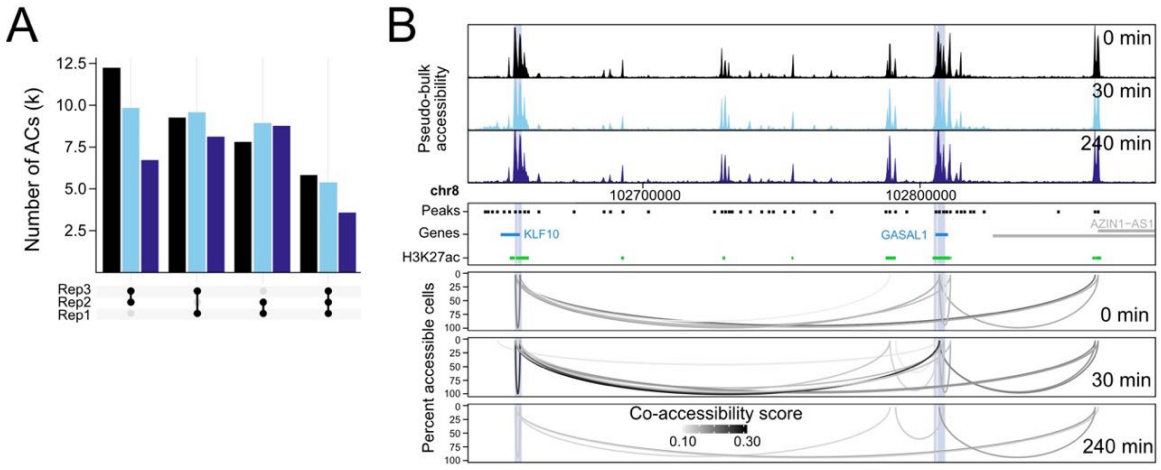


Figure 2.27 Consensus ACs of scTurboATAC-seq replicates from untreated, 30 min, and 240 min TNFα-treated HUVECs. A Number of reproducible ACs in at least two replicates. **B** Chromatin accessibility and consensus ACs at time points in an exemplary region of TNFα-regulated genes *KLF10* and *GASAL1*. Top: Pseudo-bulk chromatin accessibility tracks; Middle: ATAC peaks (black, 1 kb extended), genes (grey), TNFα-regulated genes (blue) and 1 kb regions around their TSSs (light blue), H3K27ac peaks from ChIP-seq at 30 min time point (green). Bottom: ACs at TNFα-regulated gene promoters. The grayscale and height of loops reflect co-accessibility scores and percent accessible cells of ACs. Adapted from Seufert *et al.* (2024).

DCs are reproducible across replicates

Metacell co-accessibility was computed for each replicate of TNFα treatment time point samples from scTurboATAC-seq separately to investigate the reproducibility of DCs. A genomic region around the three TNFα-regulated genes *TNFAIP3*, *WAKMAR2,* and *IFNGR1* was selected as an example to visually inspect the reproducibility of DCs (**Figure 2.28A**), since the three genes located in broader domains of locally increased co-

accessibility scores. Across all three replicates, a domain with strongly enriched co-accessibility scores was apparent around *TNFAIP3* and *WAKMAR2*. Multiple DCs were called in that region for all replicates (**Figure 2.28A**, red regions). In Rep2 and Rep3, comparable DCs were additionally called around *IFNGR1*, where a small domain of enriched co-accessibility scores was visible. The *TNFAIP3/WAKMAR2* and *IFNGR1* DCs were distant but linked by enriched co-accessibility between them in Rep1 and Rep2. Additionally, the *TNFAIP3/WAKMAR2* DC showed enriched co-accessibility to a downstream H3K27ac peak in Rep1 and Rep2. These distal links of the *TNFAIP3/WAKMAR2* DC were not clearly apparent in Rep3, as it showed generally higher co-accessibility with less distinct enrichment in specific regions. This was potentially caused by higher variability in data quality between the samples of Rep3 (see **Figures 2.20B+C**). These differences between the replicates were confirmed when investigating the base pair overlap of their called DCs genome-wide (**Figure 2.28B**). DCs from Rep1 and Rep2 showed 60 % overlap, while both showed only 40 % overlap with Rep3. To obtain consensus DCs across all replicates, consensus *metacell co-accessibility* matrices from all replicates were computed by averaging the replicate *metacell co-accessibility*

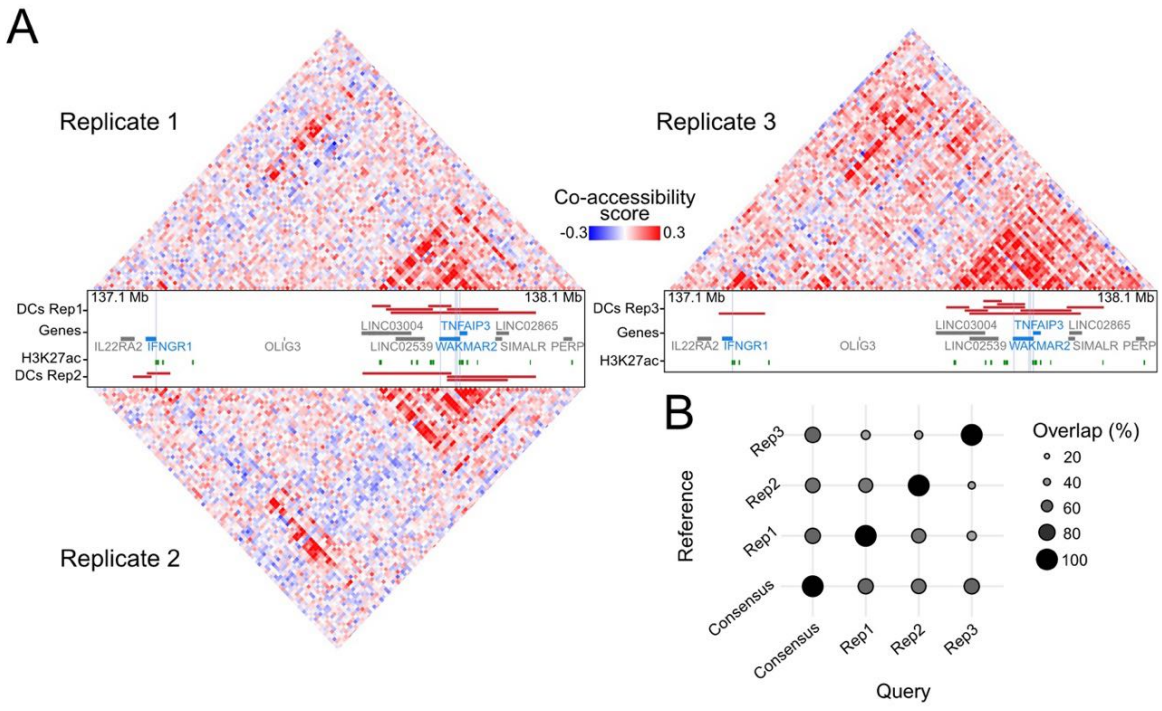


Figure 2.28 Reproducibility of DCs from *metacell co-accessibility* of scTurboATAC-seq replicates from untreated, 30 min, and 240 min TNFα-treated HUVECs. A Chromatin co-accessibility maps and DCs from replicates in an exemplary region of TNFα-regulated genes *TNFAIP3*, *IFNGR1* and *WAKMAR2*. DCs from *metacell co-accessibility* of replicates (red), H3K27ac peaks from ChIP-seq at the 30 min time point (green), genes (grey), TNFα-regulated genes (blue) and 1 kb regions around their TSSs (light blue) are indicated. Limits of the color scale bars are set to -0.3 and 0.3. **B** Comparison of genome-wide DCs between replicates and consensus from average *metacell co-accessibility* of replicates. The size and color of the dots reflect the percent of base pair overlap between the DCs. Adapted from Seufert *et al.* (2024).

matrices. Afterwards, SpectralTAD was used to newly identify domains from this consensus *metacell co-accessibility* matrix and DCs with exceptionally high co-accessibility scores were identified. These consensus DCs showed a generally high overlap (above 60 %) with all replicates (**Figure 2.28B**).

In summary, both ACs and DCs showed high concordance between the replicates of HUVEC scTurboATAC-seq data. For ACs, the overlap between replicates was exceptionally high at TNFα-regulated genes (approximately 75 %), but less on a genome-wide scale (below 10 %). The high reproducibility in TNFα-induced chromatin regions might indicate that some ACs resolve co-accessibility from targeted molecular processes, while others reflect randomly occurring co-accessible events. The presence of these random co-accessible events aligns with my previous observations on the stochastic nature of chromatin accessibility (see **Section 2.1.3**). However, investigating the underlying molecular processes of ACs might elucidate their varying reproducibility. For DCs, the overlap between consensus and replicates was high (above 60 %). Here, differences in reproducibility were mostly driven by varying data quality among samples of one replicate (see **Section 2.1.1**).

2.2.3. Molecular mechanisms driving chromatin co-accessibility

The previous sections demonstrated that ACs and DCs are fundamentally different features of chromatin co-accessibility. However, the underlying molecular mechanisms driving these distinct features remain unclear. To investigate the various molecular aspects of ACs and DCs and identify their underlying biological processes, I utilized scTurboATAC-, bulk HiC-, and bulk H3K27ac ChIP-seq data from HUVECs (see **Table 2.9**). Specifically, I analyzed the relationship of ACs and DCs with TADs, representing higher-order chromatin structures. The TADs were identified from bulk HiC-seq data of untreated HUVECs and reflect regions of high chromatin contact frequencies in the unperturbed baseline condition. In HiC-seq data, the contact frequency of two genomic regions reflects how often they were found in spatial vicinity within the bulk population of cells, where high frequencies suggest that the loci are frequently proximal in 3D space. Additionally, for ACs, I examined these chromatin contact frequencies and potential differences between ACs with high and low detection rates among the single cells. For DCs, I assessed their response to TNFα treatment and local variations in TF binding activity.

ACs show increased contact frequencies in HiC-seq data

Chromatin contact frequencies were studied in the exemplary region around the two TNFα-regulated genes, *KLF10* and *GASAL1*, which previously showed highly reproducible ACs between their promoters and distal H3K27ac peaks (see **Figures 2.26A**, **2.27B**). The two genes are located at the opposite boundaries of the same TAD, indicating that their associated ACs emerged within a single TAD (**Figure 2.29A**). When examining the location of ACs in relation to TADs genome-wide, most ACs (45-52 %) were found within the same TAD (**Figure 2.29B**). However, significant fractions of ACs crossed TAD boundaries (26-33 %) or were located outside of TADs (all 23 %).

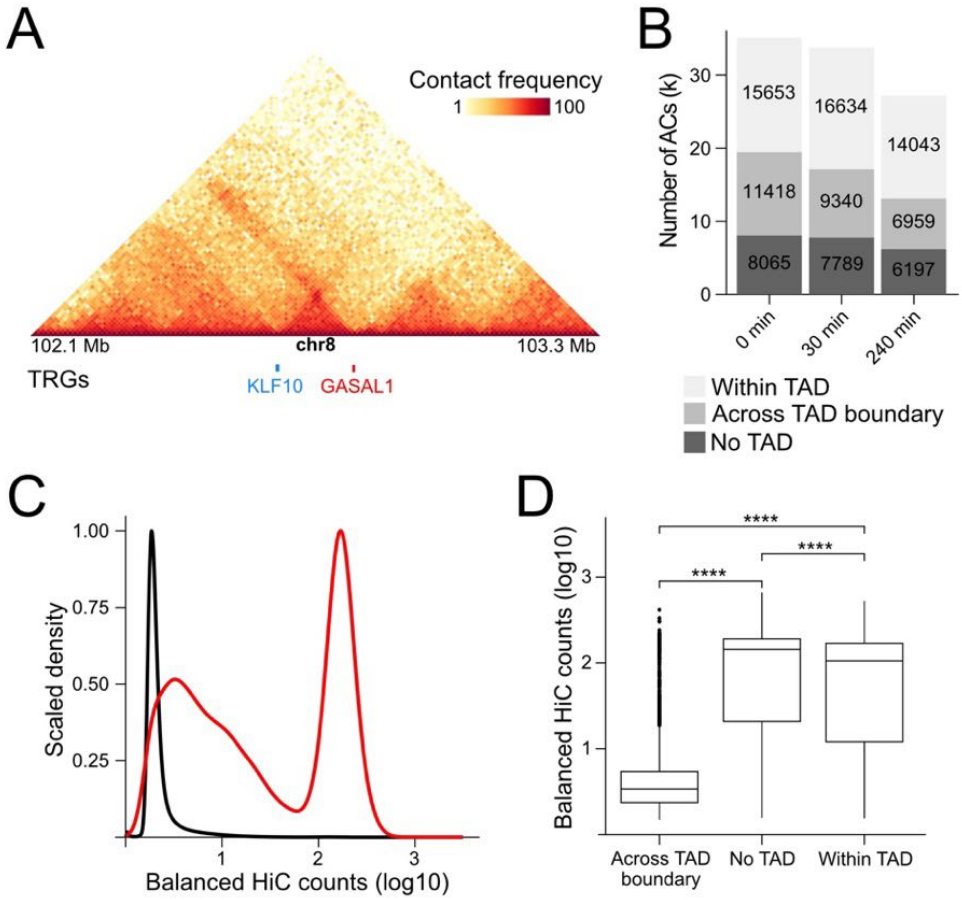


Figure 2.29 Chromatin contact frequencies and TADs at ACs from HiC-seq data of unstimulated HUVECs. A Chromatin contact map in an exemplary region of TNFα-regulated genes *KLF10* and *GASAL1*. The upper limit of the color scale bar is set to 100. **B** Genomic location of ACs in relation to TADs. ACs were classified as within one TAD, across a TAD boundary, and without TAD overlap. **C** Chromatin contacts genome-wide (black) and between AC-linked peaks (red). **D** Chromatin contacts between AC-linked peaks within one TADs, across a TAD boundary, and outside of TADs. P-values < 2.22e-16 from Wilcoxon test are indicated by ****. Adapted from Seufert *et al.* (2024).

Next, I compared the chromatin contact frequencies of AC-linked ATAC peaks to the genome-wide distribution of chromatin contact frequencies (**Figure 2.29C**). The chromatin contact frequencies of AC-linked ATAC peaks showed a bimodal distribution. A fraction of

ACs exhibited approximately 50-fold higher contact frequencies compared to the genome-wide background, while the remaining ACs showed only moderately higher contact frequencies. This bimodal distribution coincided with the ACs' location in relation to TADs (**Figure 2.29D**). ACs within and outside of TADs displayed very high contact frequencies, whereas ACs crossing TAD boundaries showed significantly lower contact frequencies. Nonetheless, all ACs demonstrated contact frequencies above background, indicating that ACs reflect interactions between co-accessible, and thereby co-active, distal chromatin sites.

Frequent and *rare ACs* originate from different molecular processes

Previously, the investigation of the detection rate revealed two types of ACs, as demonstrated by the bimodal PAC distribution of ACs (see **Figure 2.24B**). ACs with PAC above 75 were considered frequent, while ACs with PAC below 75 were considered rare. Chromatin contact frequencies showed different enrichment patterns for these two types of ACs (**Figure 2.30A**). *Rare ACs* with low PACs exhibited a strong enrichment of chromatin contacts in their entire vicinity. In contrast, *frequent ACs* with high PACs displayed enriched chromatin contacts between the linked peaks but depleted chromatin contacts beyond the linked peaks. This observation suggested that *rare ACs* might result from stochastic or random interactions of active chromatin sites in a highly dynamic and transient chromatin environment with generally high contact frequencies. Conversely, *frequent ACs* potentially originate from architectural chromatin interactions, creating diverse chromatin environments of enriched or reduced chromatin interactions between and outside their linked peaks.

To investigate the underlying mechanism of *frequent ACs* in more detail, *single cell* and *metacell co-accessibility* were examined in an exemplary region around the four TNFα-regulated genes *GBP1, GBP2, GBP3* and *GBP4* (**Figure 2.30B**), since these *GBP* genes showed multiple *frequent ACs* (PACs of almost 100) in their surroundings. These *frequent ACs* from *single cell co-accessibility analysis* occurred between multiple H3K27ac peaks surrounding the *GBP* genes (**Figure 2.30B**, bottom: H3K27ac peaks in green). Furthermore, all of them originated from genomic loci at or near gene promoters. The *metacell co-accessibility* analysis revealed a specific pattern at these *frequent ACs* (**Figure 2.30B**, top). The AC-linked ATAC peaks showed high *metacell co-accessibility* between them (**Figure 2.30B**, top: black rectangles), but anti-correlated accessibility with all other genomic loci in the region (**Figure 2.30B,** top: anti-correlated accessibility in blue). Chromatin contact frequencies in this region indicated that the origins of these so-called *blue stripes* did not correspond to TAD boundaries (**Figure 2.30C**, bottom). However, the

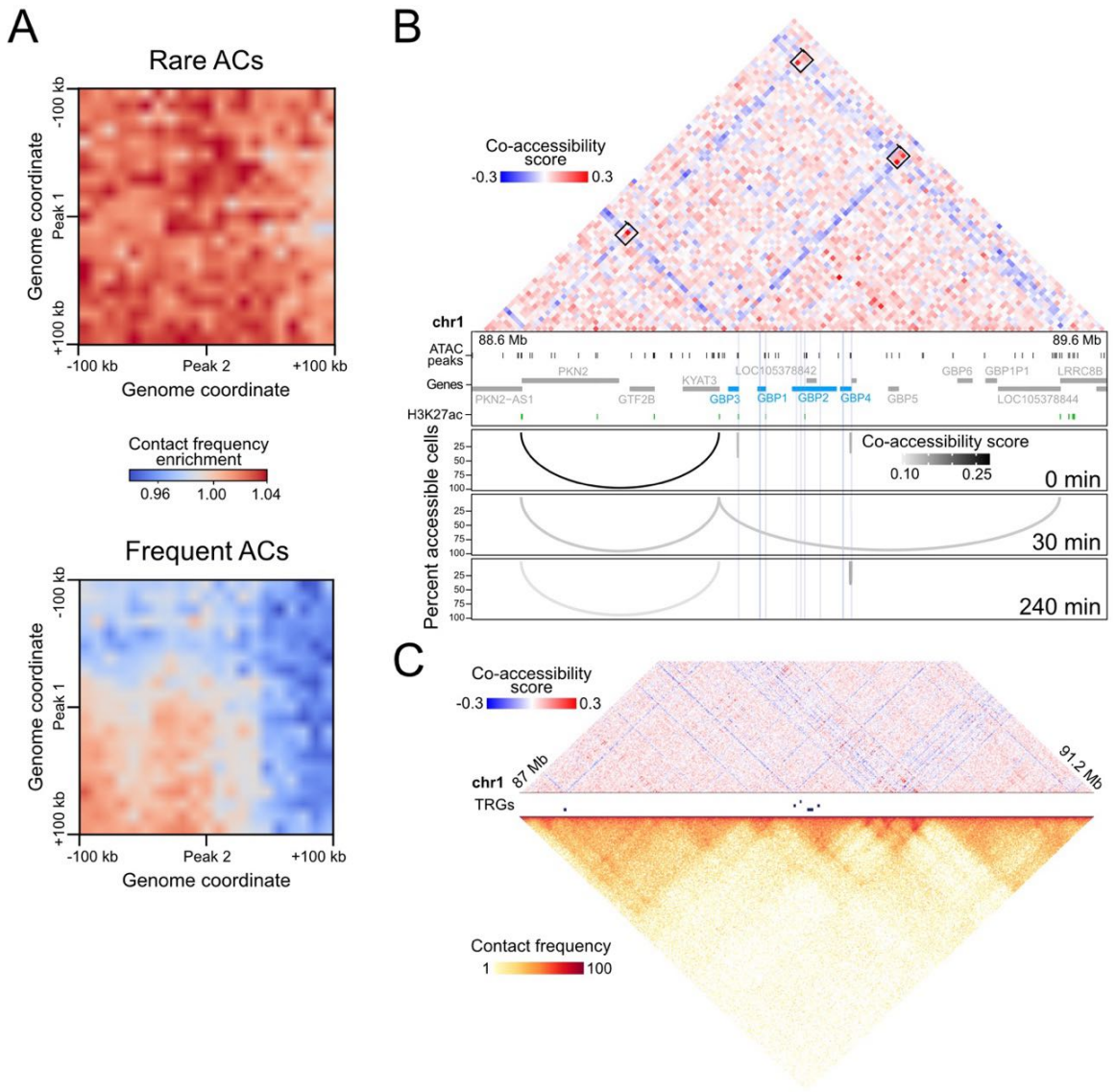


Figure 2.30 Chromatin contacts and *metacell co-accessibility* at *rare* and *frequent ACs*. **A** Chromatin contact frequency enrichment at rare (top) and *frequent ACs* (bottom). Centered, scaled, and averaged pileups of ACs from untreated HUVECs are shown. **B** *Metacell co-accessibility* map (top) and *single cell co-accessibility* consensus ACs (bottom) in an exemplary region of TNFα-regulated genes *GBP1, GBP2, GBP3,* and *GBP4*. The annotation in the middle shows ATAC peaks (black, 1 kb extended), H3K27ac peaks from ChIP-seq after 30 min TNFα treatment (green), genes (grey), TNFα-regulated genes (blue), and 1 kb regions around their TSSs (light blue). Limits of the *metacell co-accessibility* color scale bar are set to -0.3 and 0.3. The grayscale and height of loops reflect co-accessibility score and percent accessible cells of ACs. **C** Zoom out from panel B with maps of *metacell co-accessibility* (top) and chromatin contacts (bottom). Limits of the co-accessibility color scale bar are set to -0.3 and 0.3. The upper limit of the chromatin contact color scale bar is set to 100. Adapted from Seufert *et al.* (2024).

blue stripes coincided with increased chromatin contact frequencies proximal to TAD boundaries, visible as *red stripes* in the HiC map. *Metacell co-accessibility* in the extended genomic region showed that the *blue stripes* were not only observed between neighboring AC-linked H3K27ac peaks but extended further to subsequent linked H3K27ac peaks (**Figure 2.30C**, top). These findings supported the hypothesis that *frequent ACs* and their related *blue stripes* reflect architectural chromatin interactions, e.g. chromatin loops.

Furthermore, their proximity to TAD boundaries, coincidence with chromatin contact stripes, and extension of *blue stripes* beyond their immediate neighborhood suggested that these might represent molecular processes of loop or TAD boundary stacking.

DCs show significant accessibility changes upon TNFα treatment

Chromatin contact frequencies were studied in the exemplary region around the three TNFα-regulated genes, *TNFAIP3, WAKMAR2*, and *IFNGR1*, which previously showed highly reproducible DCs across replicates (see **Figure 2.28A**). The consensus *metacell co-accessibility*, computed by averaging the replicate *metacell co-accessibility* matrices, showed local enrichment of *metacell co-accessibility* around these genes (**Figure 2.31A**, top). Multiple DCs around *TNFAIP3* and *WAKMAR2* were identified (**Figure 2.31A**, red regions), while no DC at *IFNGR1* was detected. The HiC map indicated that the *TNFAIP3*/*WAKMAR2* DC coincided with enriched chromatin contact frequencies in the

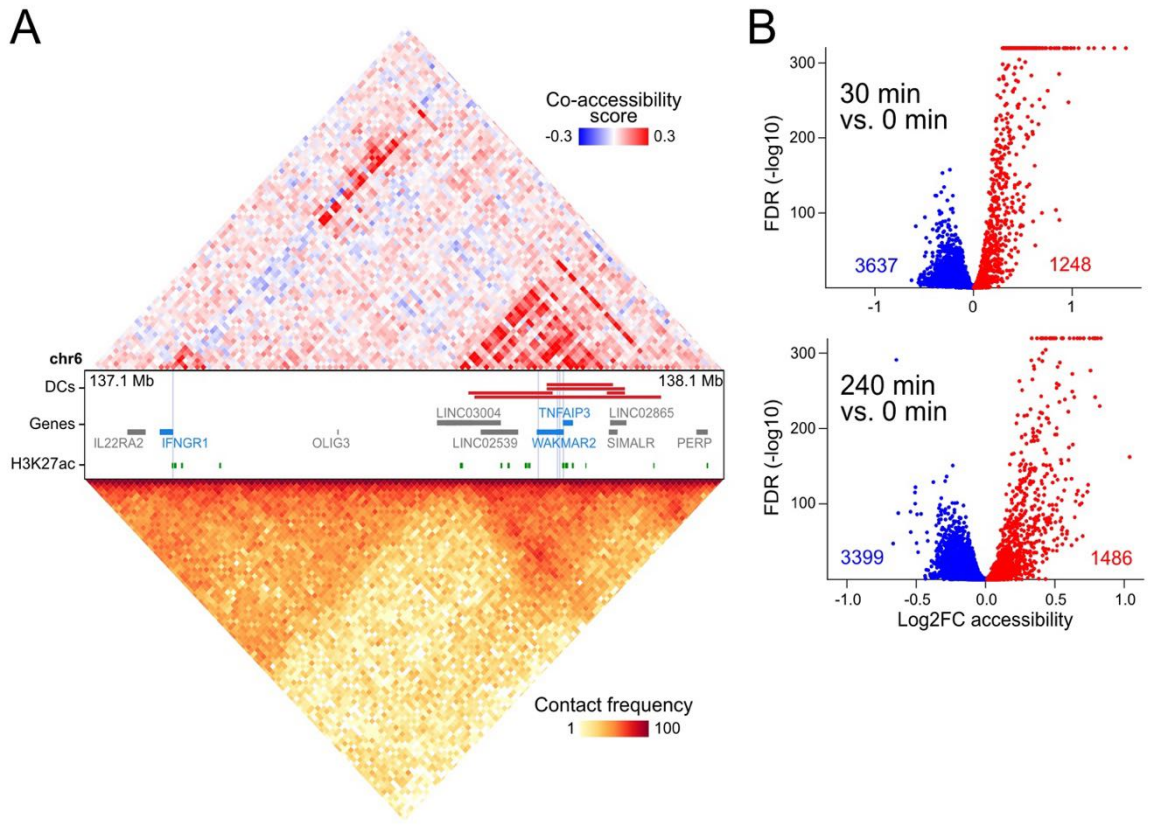


Figure 2.31 Chromatin contacts and accessibility at DCs from scTurboATAC-seq data of untreated, 30 min, and 240 min TNFα-treated HUVECs. A Average *metacell co-accessibility* map from replicates and chromatin contact map from HiC-seq data in an exemplary region of TNFα-regulated genes *TNFAIP3*, *IFNGR1* and *WAKMAR2*. DCs from average *metacell co-accessibility* (red), H3K27ac peaks from ChIP-seq at the 30 min time point (green), genes (grey), TNFα-regulated genes (blue) and 1 kb regions around their TSSs (light blue) are indicated. Limits of the co-accessibility color scale bar are set to -0.3 and 0.3. The upper limit of the chromatin contact color scale bar is set to 100. **B** Differential accessibility in DCs after TNFα treatment of HUVECs across three biological replicates visualized by log2FC and FDR. DCs with FDR below 0.05 are considered significant and marked in red (upregulated) and blue (downregulated). Adapted from Seufert *et al.* (2024).

same region (**Figure 2.31A**, bottom). *IFNGR1* was linked by a broader region of long-distance co-accessibility that spanned across a TAD boundary between the three genes. Nearly all (95 %) of the 4,885 identified DCs showed significant accessibility changes upon TNFα treatment (**Figure 2.31B**). This indicates regulatory chromatin changes throughout the entire DC region rather than specific variations at CREs within the DCs. Accessibility was reduced in most DCs (74 % at 30 min; 70 % at 240 min), while only 26 % and 30 % showed increased accessibility in response to TNFα treatment. However, the opposite pattern was observed for the 683 DCs containing a TNFα-regulated gene: 75 % of these showed increased accessibility upon TNFα treatment, suggesting specific activation and increased activity of DCs with TNFα-regulated genes.

DCs are independent TAD sub-structures

The genome-wide comparison of DCs and TADs showed that DCs have, on average, a tenfold smaller genomic size than TADs (**Figure 2.32A**). Furthermore, the majority of DCs (64 %) were within a single TAD, but approximately 35 % of DCs either overlapped with a TAD boundary or were located outside of TADs (**Figure 2.32B**). These observations suggest that DCs are TAD-independent structures. To understand how contiguous DC form across TAD boundaries, I examined chromatin contact frequencies at TADs with DCs inside, DCs across their boundaries, and TADs without DCs (**Figure 2.32C**). For all groups, chromatin contact frequencies were enriched within TADs. However, only TADs, with DCs crossing their boundaries, showed enriched chromatin contact frequencies with neighboring TADs. In contrast, these frequencies were lower for TADs with DCs inside and for TADs without DCs. This suggests that some TAD boundaries allow DCs to extend beyond them, increasing their contact frequencies with surrounding regions.

DCs show local enrichment of TF binding activity

To investigate the biological mechanisms causing these TAD-independent structures with high co-accessibility and TNFα-induced activity changes, I studied the local TF binding activity in DCs (**Figure 2.33**). TF binding can be studied using pseudo-bulks of scATAC-seq data, as TFs protect their binding motifs from Tn5 insertions, making them inaccessible (see **Section 2.1.5**). I computed accessibility footprints for TFs that showed a genome-wide increase in binding upon TNFα treatment. This comprised 16 TFs from the families of NF-κB, IRF, and CCAAT enhancer binding proteins (CEBP), and others. Each of these TFs possesses an individual binding motif, which was used to predict all potential binding sites of this TF genome-wide. The accessibility footprints were used to infer TF binding scores for each binding site individually (**Figure 2.33**, middle). To investigate local differences in TF binding between genomic regions, I compared the TF footprints and

scores of binding sites within DCs to those in local and genome-wide non-DC background regions (**Figure 2.33**, right).

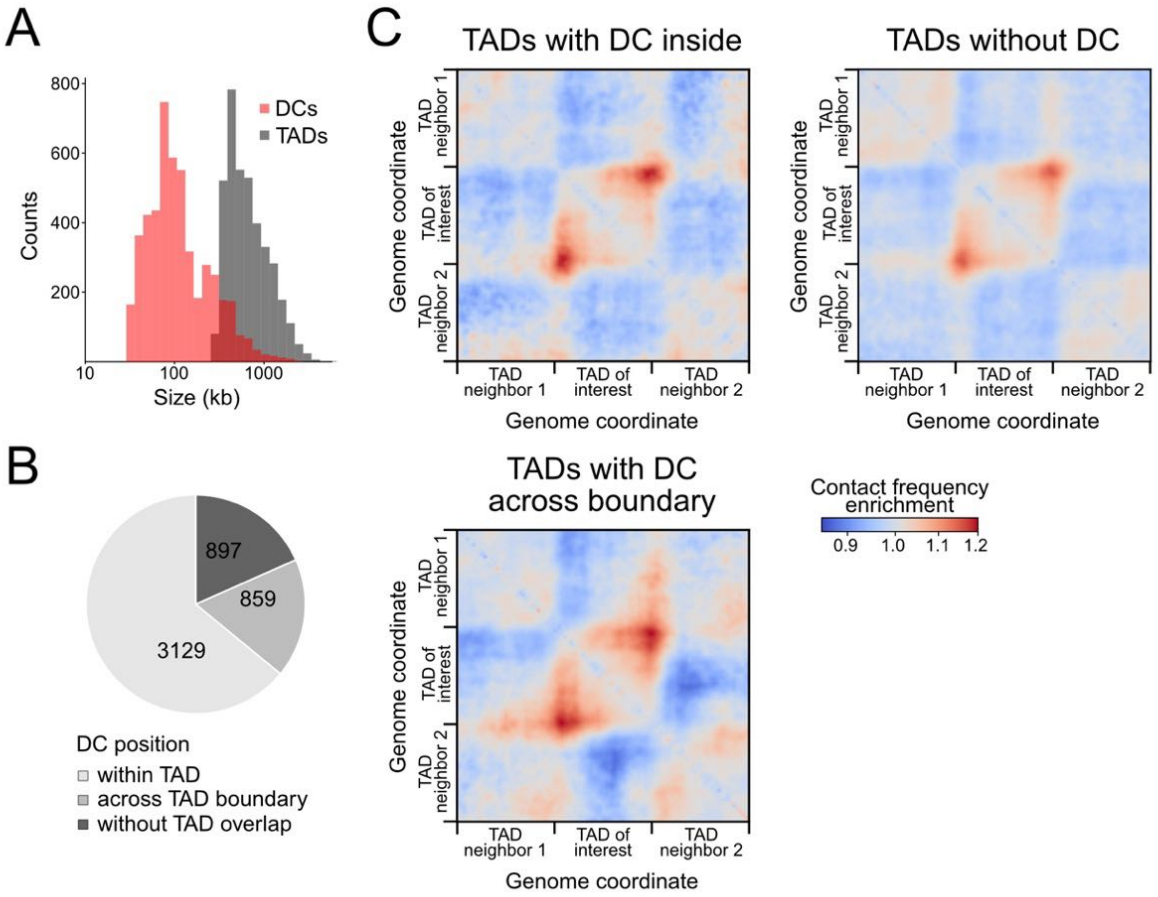


Figure 2.32 Chromatin contacts and TADs at DCs from HiC-seq data of unstimulated HUVECs. A Genomic sizes of DCs and TADs. **B** Genomic location of DCs in relation to TADs. DCs were classified as within one TAD, across a TAD boundary, and without TAD overlap. **C** Chromatin contact frequency enrichment at TADs with DC within (top left), DC across TAD boundary (bottom), and without DC (top right). Centered, scaled and averaged pileups of TADs from untreated HUVECs are shown. Adapted from Seufert *et al.* (2024).

NF-κB/p65 accessibility footprints for binding sites in the previously studied *TNFAIP3/WAKMAR2* DC (see **Figures 2.28A**, **2.31A**) revealed low accessibility at the motif centers and high accessibility in the surrounding regions, indicating high binding activity of NF-κB/p65 at all treatment time points (**Figure 2.34A**, red). Footprints from binding sites in the surrounding non-DC regions of the same size showed distinctly less NF-κB/p65 binding for all time points (**Figure 2.34A**, black). In untreated HUVECs, the footprints revealed low accessibility at both the motif center and its flanking regions indicated little to no NF-κB/p65 binding in the non-DC background. Upon TNFα treatment, accessibility increased in the binding site flanking regions of the background, but there was no depletion of accessibility at the center. This suggests that NF-κB/p65 molecules in the local background of the *TNFAIP3/WAKMAR2* DC rarely and briefly bind to its potential

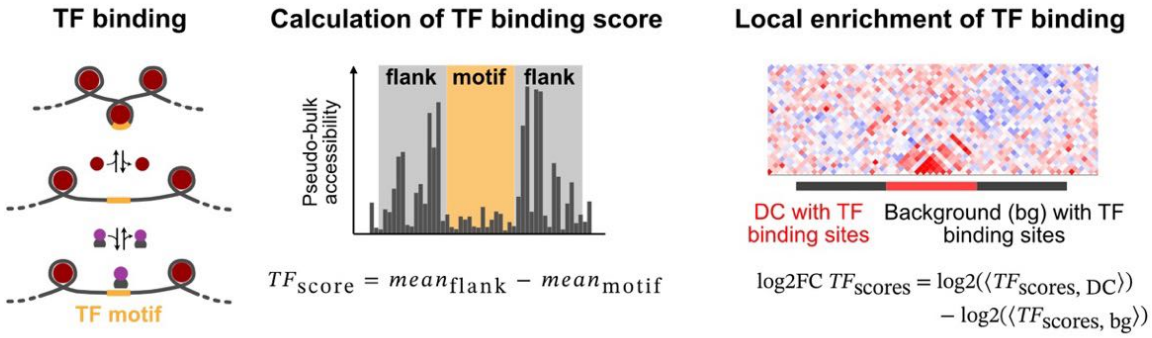


Figure 2.33 Investigation of local alterations in TF binding at DCs using pseudo-bulks of scATAC-seq data. TF binding makes chromatin inaccessible at its binding motifs, while simultaneously increasing the probability of Tn5 insertions in its surrounding accessible regions. These patterns are called TF footprints, from which TF binding scores can be inferred. Local differences in TF binding are identified by comparing TF binding scores in DCs to those in local or genome-wide non-DC background regions. Adapted from Seufert *et al.* (2024).

binding sites, making them generally more accessible. Consequently, potentially low NF-κB/p65 concentration or binding activity in the background did not saturate TF occupancy at the motifs. In contrast, the strongly reduced accessibility at the DC's motif centers indicated high NF-κB/p65 concentration or binding activity, resulting in a saturated TF occupancy of the binding sites. Quantifying this with NF-κB/p65 binding scores showed a distinct increase in NF-κB/p65 binding in the *TNFAIP3/WAKMAR2* DC upon TNFα treatment (**Figure 2.34B**, left). In comparison, NF-κB/p65 binding scores in the genome-wide non-DC background were lower across all TNFα treatment time points (**Figure 2.34B**, right).

Furthermore, differential TF binding between the *TNFAIP3/WAKMAR2* DC and the genome-wide non-DC background revealed local differences in TF binding scores at the exemplary DC (**Figure 2.34C**). Across all TNFα treatment time points, NF-κB family TFs demonstrated significantly higher binding scores in the *TNFAIP3/WAKMAR2* DC than in the non-DC background. In contrast, other TFs showed no difference or even lower binding scores in the *TNFAIP3/WAKMAR2* DC than in the background for untreated HUVECs. PR domain zinc finger protein 1 (PRDM1), type 1 interferon response element (T1ISRE), CEBP, and PU.1-IRF8 exhibited higher binding scores in *TNFAIP3/WAKMAR2* DC upon TNFα treatment. Additionally, IRF-BATF (basic leucine zipper ATF-like transcription factor) and IRF4 showed higher binding scores in the *TNFAIP3/WAKMAR2* DC after 240 min of treatment. These findings were consistently reproducible across replicates at all treatment time points (**Figures 2.34A-C**). A meta-analysis of replicates revealed significantly higher binding scores of NF-κB/p65 and NF-κB/p65/Rel in the *TNFAIP3/WAKMAR2* DC than the non-DC background after TNFα treatment (**Figure 2.34D**). Higher binding scores of PRDM1, T1ISRE, and IRF-BATF were also statistically significant, though less strong.

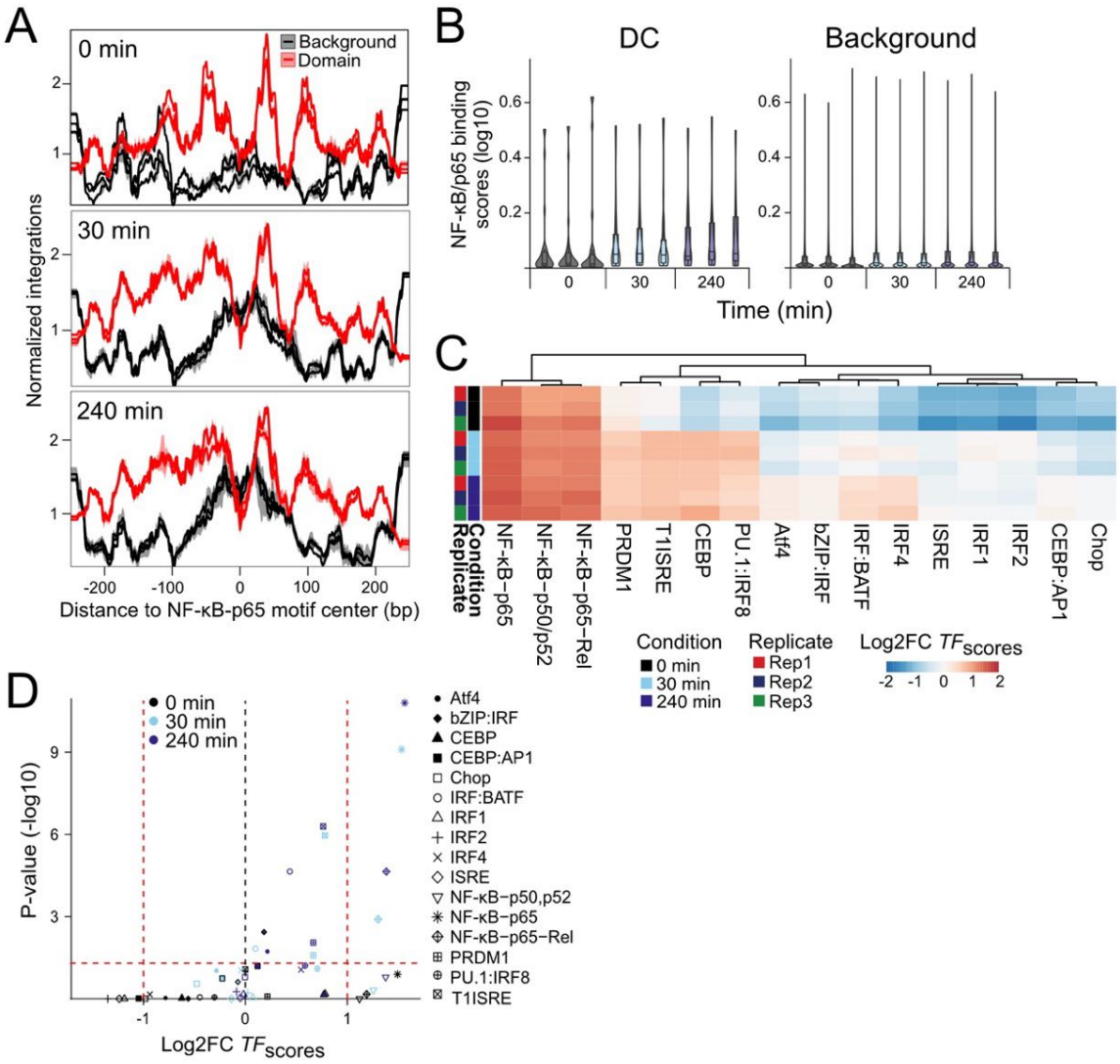


Figure 2.34 Local enrichment of TF binding at the *TNFAIP3/WAKMAR2* DC from scTurboATAC-seq data of untreated, 30 min, and 240 min TNFα-treated HUVECs. A Pseudo-bulk accessibility footprints at NF-κB/p65 motifs in the *TNFAIP3/WAKMAR2* DC (red) and the surrounding non-DC regions of the same size (local background; black). Each line shows the pseudo-bulk accessibility of one biological replicate in untreated (top), 30 min (middle), and 240 min (bottom) TNFα-treated HUVECs. **B** NF-κB/p65 binding scores (log10) of accessible motifs in the *TNFAIP3/WAKMAR2* DC (left) and genome-wide non-DC regions (genome-wide background; right). **C** Differential TF binding in the *TNFAIP3/WAKMAR2* DC vs. the genome-wide background in individual replicates. Color scale limits are set to -2 and 2. **D** Differential TF binding in the *TNFAIP3/WAKMAR2* DC vs. the genome-wide background visualized by average log2FC and p-values from meta-analysis of replicates. TFs with absolute log2FC above 1 and FDR below 0.05 are considered significant. Adapted from Seufert *et al.* (2024).

Significantly higher binding scores in DCs than in the non-DC genome-wide background were observed in 44 % of the 4,885 DCs for at least one of the studied TFs. The DCs displayed varying patterns of locally enriched TFs (**Figure 2.35**). Some DCs showed significantly higher binding scores of a single TF, while others exhibited higher binding scores of an entire TF family or multiple families. Additionally, responses to TNFα treatment varied: Some DCs showed higher binding scores regardless of treatment time

point, while others exhibited increased or decreased local enrichment in response to TNFα treatment. Overall, these results suggest that DCs arise from local differences in TF concentration or binding activity, leading to the distinct contiguous co-accessibility patterns within broad domains.

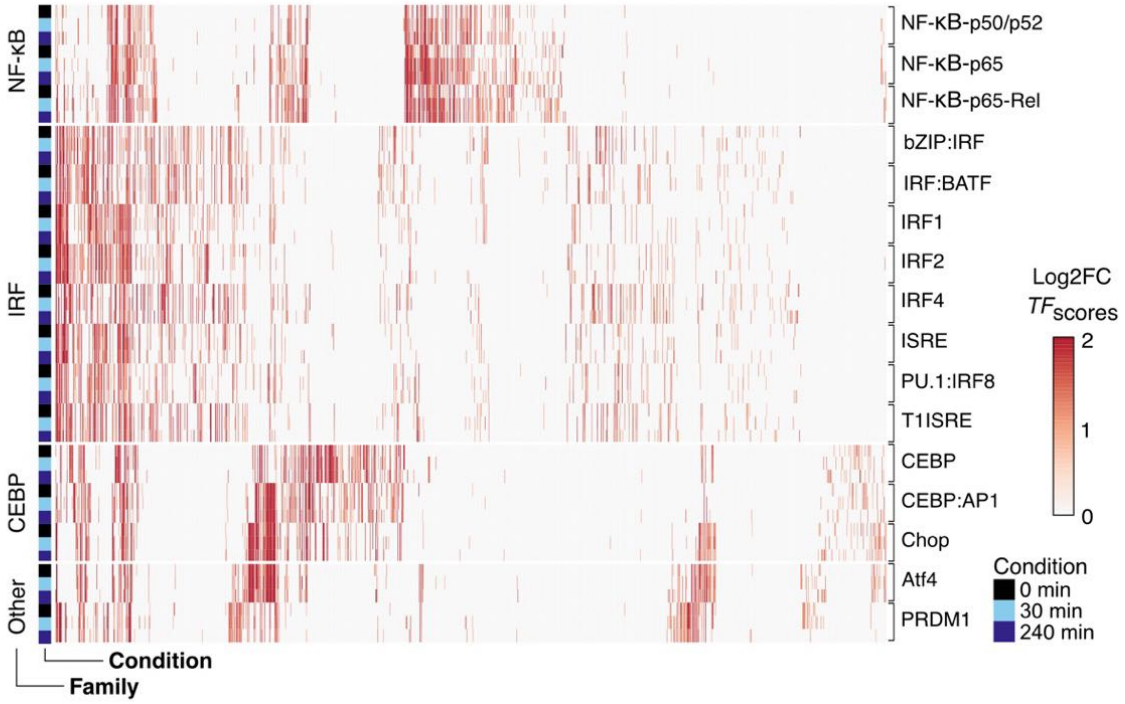


Figure 2.35 Local enrichment of TF binding activity in DCs from scTurboATAC-seq data of untreated, 30 min, and 240 min TNFα-treated HUVECs. Differential TF binding in DCs vs. the genome-wide background in unstimulated and TNFα-stimulated HUVECs. Average log2FC of replicates for all DCs with significant local enrichment of TF binding are shown. TFs are grouped by family. DCs are clustered by summed family enrichment. Color scale limits are set to 0 and 2. Adapted from Seufert *et al.* (2024).

ACs and DCs originate from different molecular processes

In conclusion, the co-accessibility patterns of ACs and DCs arise from distinct molecular mechanism. ACs represent spatial contacts between co-active distal chromatin sites (**Figure 2.36**, left). Here, *rare* and *frequent ACs* originate from chromatin contacts induced by different biological processes. *Rare ACs* are likely the result of random interactions within dynamic chromatin regions with generally high contact frequencies. In contrast, *frequent ACs* seem to reflect architectural interactions, leading to higher contact frequencies of the chromatin within than beyond the AC. These *frequent ACs* potentially arise from targeted chromatin loops or stacking of TAD boundaries. Despite their differences, both AC types represent chromatin topology-mediated contacts between distal sites. In contrast, DCs are contiguous chromatin regions that function as TAD-independent structures with TNFα-dependent accessibility changes (**Figure 2.36**, right).

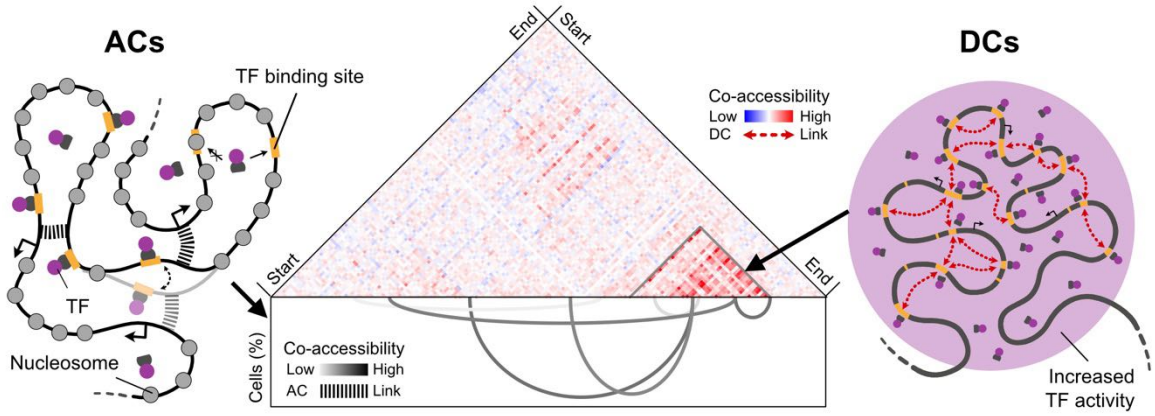


Figure 2.36 Biological mechanisms causing co-accessibility patterns of ACs and DCs. ACs resolve contacts of co-accessible, and thereby co-active, distal chromatin sites. They form two groups with low and high detection rates, each driven by different molecular mechanisms inducing chromatin contact. *Rare ACs* likely reflect random interactions in dynamic chromatin regions with generally high contact frequencies, while *frequent ACs* represent targeted architectural interactions, potentially caused by loops or stacking of TAD boundaries. DCs are TAD-independent structures of enriched co-accessibility that exhibit TNFα-induced accessibility changes. These structures originate from local alterations in TF concentration or binding activity, which render the entire local domain co-accessible. Adapted from Seufert *et al.* (2024).

The distinct co-accessibility pattern of DCs likely results from enrichment of TF binding activity in local subcompartments or hubs. However, the data does not allow to determine whether these nuclear subcompartments are formed by liquid-liquid phase separation of TFs and co-factors or by local TF confinement due to chromatin compaction. Nevertheless, unlike chromatin-mediated ACs, DCs appear to represent TF-mediated chromatin hubs.

In summary, in this chapter I designed and developed a computational framework for chromatin co-accessibility analysis that resolves different layers of co-accessibility by enriching for varying dimensions of accessibility changes (**Section 2.2.1**). By introducing the R software package RWireX, the computational framework was made available to the scientific community. Additionally, I identified and addressed experimental and computational limitations in co-accessibility analysis by assessing its reproducibility for data sets of equal as well as varying quality (**Sections 2.2.1**, **2.2.2**). Finally, I investigated the underlying biological mechanisms causing the distinct patterns in co-accessibility, which I observed before (**Section 2.2.3**).

2.3. Identifying the structure-function relationship between regulatory mechanisms and their transcriptional output

In this chapter, I address the third aim of this thesis: Identifying the structure-function relationship between regulatory mechanisms and their transcriptional output. By applying the computational framework for co-accessibility analysis to various human and mouse systems under perturbation (see **Section 1.3**), I aimed to study genome-wide mechanisms of transcription regulation using chromatin co-accessibility. To achieve this, I used three different model systems under perturbation: (i) Two mouse cell types, namely ESCs and MEFs, untreated and treated with IFNβ, (ii) the TCL1 mouse model for CLL with *Tbx21* wild-type or double knock-out, and (iii) HUVECs untreated and treated with TNFα. The perturbation studies enabled me to investigate the regulatory mechanisms underlying specifically induced changes in transcription. The broad range of mammalian systems, along with internal and external perturbations, allowed me to explore general principles of transcription regulation beyond molecular responses linked to specific pathways and stimuli, such as cytokine treatment.

In the first project, I investigated ESCs and MEFs in an unstimulated condition and after IFNβ treatment for 1 h and 6 h (see **Section 1.3.1**), which was published in Muckenhuber *et al.* (2023). Here, I used scRNA-seq and scATAC-seq data to study the gene regulatory response to IFNβ treatment (**Table 2.10**, project 1). The analyses built upon findings from bulk RNA- and STAT1/STAT2 ChIP-seq data on IFNβ induced changes in TF binding and gene expression. As described in **Chapter 2.1**, sequencing data acquisition and the analysis of bulk and scRNA sequencing from ESCs and MEFs were performed by Markus Muckenhuber. My contribution involved the computational analysis of scATAC-seq data from ESCs and MEFs. In addition, these results were complemented by my bulk RNA-seq data analysis of natural killer (NK) cells co-cultured with non-infected and Hepatitis D virus (HDV)-infected hepatocytes (**Table 2.10**, project 1; **Figure 2.37A**), which were published in Groth *et al.* (2023). Markus Muckenhuber conducted bulk RNA-seq of NK cell co-cultures with hepatocytes, for which samples were prepared by Christopher Groth (Department of Immunobiochemistry, Mannheim Institute for Innate Immunoscience and Medical Faculty Mannheim, Heidelberg University, Germany). Differential expression analysis of HDV-related bulk RNA-seq data was performed by Carsten Sticht (Medical

Faculty Mannheim, Heidelberg University, Germany). My contribution comprised the computational preprocessing of the bulk RNA-seq data.

Table 2.10 Overview of data sets used for identifying the structure-function relationship between regulatory mechanisms and their transcriptional output.

Project	Sample	Perturbation	Method	Sequencing assay
1	ESC	IFNβ treatment	Single-cell seq.	RNA, ATAC
			Bulk seq.	RNA, TF ChIP (STAT1, STAT2)
1	MEF	IFNβ treatment	Single-cell seq.	RNA, ATAC
			Bulk seq.	RNA, TF ChIP (STAT1, STAT2)
1	NK cells	HDV infection	Bulk seq.	RNA
2	TCL1 cells	*Tbx21* knock-out	Single-cell seq.	RNA, TurboATAC
			Bulk seq.	RNA
			Mass spectrometry	
2	CLL patient samples	High/low *TBX21* expression	Bulk seq.	RNA, ATAC
			Mass spectrometry	
		Malignancy	Bulk seq.	RNA
3	HUVEC	TNFα treatment	Single-cell seq.	RNA, TurboATAC, Nuclear RNA
			Bulk seq.	HiC, Histone ChIP (H3K27ac)
			Spatial transcriptomics of nuclear RNA	

In the second project, I investigated TCL1 mouse models for CLL in wild-type conditions (*Tbx21*$^{+/+}$) and with *Tbx21* double knock-out (*Tbx21*$^{-/-}$). The results were published in Roessner *et al.* (2024). As described earlier in **Section 1.3.3**, the *Tbx21* gene encodes the transcription factor T-bet, whose knock-out affects the transcription of its specific target genes as well as secondary targets. I used scRNA-seq and scTurboATAC-seq data from two biological replicates of both TCL1 wild type and *Tbx21*$^{-/-}$ to study T-bet-dependent gene regulation (**Table 2.10**, project 2). The *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ TCL1 mouse models were generated by Philipp Roessner (formerly Division of Molecular Genetics, German Cancer Research Center, Heidelberg, Germany). The scRNA-seq and scTurboATAC-seq data from TCL1 cells were acquired by Markus Muckenhuber. My contribution comprised the computational analysis of these data. The analyses were supported by bulk RNA-seq,

bulk ATAC-seq, and mass spectrometry (MS) data of both *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ TCL1 as well as CLL patient samples (**Table 2.10**, project 2). Bulk TCL1 data were generated by Philipp Roessner. Computational analyses of bulk sequencing data were conducted by Marc Zapatka (Division of Molecular Genetics, German Cancer Research Center, Heidelberg, Germany) and Vincente Chapaprieta (Instituto de Investigaciones Biomédicas August Pi i Sunyer, Barcelona, Spain). MS data analysis was conducted by Pavle Boskovic (Division of Molecular Genetics, German Cancer Research Center, Heidelberg, Germany).

In the third project, I studied HUVECs under untreated condition and after treatment with TNFα and the presented results were published in Seufert *et al.* (2024). I used scRNA-seq and scTurboATAC-seq data to study the gene regulatory response of primary human cells to TNFα treatment (**Table 2.10**, project 3; see **Section 1.3.2**). Single-cell sequencing data were complemented by bulk HiC-seq and H3K27ac ChIP-seq data, as well as data from nuclear RNA obtained through single-nucleus sequencing and multiplexed smFISH imaging, implemented via the so-called padFISH protocol. In addition to the previously described contributions to the HUVEC data set in **Chapter 2.2**, experimental work for the scRNA-seq data of HUVECs was performed by Irene Gerosa and Sabrina Schumacher. Nuclear RNA sequencing in single cells was conducted by Katharina Bauer and Jan-Philipp Mallm. PadFISH imaging of nuclear RNA and its analysis were performed by Irene Gerosa. My contribution was the preprocessing and computational downstream analysis of scRNA-seq, scTurboATAC-seq and nuclear RNA-seq data. Furthermore, I integrated results from all sequencing and imaging data sets.

In addition to the three presented projects, I contributed to studying transcription deregulation in subclones of multiple myeloma (MM) using my computational framework for co-accessibility analysis, published in Poos *et al.* (2023). The project comprised data on 15 patients with relapsed or refractory MM, which were studied at two time points during their treatment regime. The aim of the study was to investigate intratumor heterogeneity among the multidrug-resistant subclones and identify their underlying regulatory mechanisms as potential molecular targets. Here, *single cell co-accessibility* was computed from scATAC-seq data of patient-specific MM subclones. Subclone- or time point-specific ACs at genes that encode for known drug resistance proteins were detected and coincided with upregulated expression of these genes, potentially contributing to the observed drug resistance. The analysis was conducted in collaboration with Alexandra Poos (Department of Internal Medicine V, University Hospital Heidelberg and Clinical Cooperation Unit Molecular Hematology/Oncology, German Cancer Research Center,

Heidelberg, Germany), where I provided scripts for co-accessibility analysis and visualized results. This project is not further shown in this thesis.

2.3.1. Proximal and distal transcription regulation of IFNβ-stimulated genes in mouse cells

In mammalian organisms, IFN signaling is a key component of the antiviral response. It precisely regulates the expression of IFNβ-stimulated genes, so-called ISGs, that protect host organisms from viral infections. For example, we found that upon HDV infection of hepatocytes, IFN gamma (IFNγ) is significantly upregulated in co-cultured NK cells (**Figures 2.37A+B**). Subsequently, the increased IFNγ levels enhance the expression of ISGs, such as *IFIT5* and *ISG20*, in NK cells (**Figure 2.37C**), demonstrating the important role of IFN-dependent transcription regulation in the defense against viral infections. Similarly, IFNs play a crucial role in the immune response against the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), for which we monitored the epidemiological spreading in the Rhine-Neckar/Heidelberg region by sequencing in 2021 (Bundschuh *et al.*, 2024).

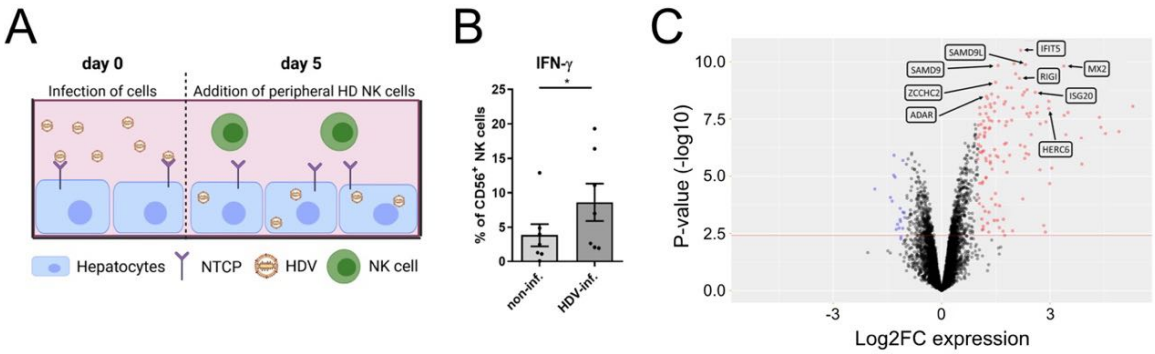


Figure 2.37 NK cell activity is induced by IFNγ after HDV infection. A Scheme of the in-vitro co-culture system of peripheral NK cells and HDV-infected hepatocytes. **B** Fraction of IFNγ expressing peripheral NK cells cultured with supernatant from non-infected and HDV-infected HepG2-hNTCP cells. P-value from Student's t-test is indicated as *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. **C** Differential expression between peripheral NK cells in co-culture with non-infected and HDV-infected HepG2-hNTCP cells. Differential expression is visualized by log2 fold changes and negative log10 p-values. Genes with absolute log2FC above 1 and p-value below $10^{-2.5}$ are considered significant and marked in red (upregulated) and blue (downregulated). Co-cultures and bulk RNA-seq data were prepared by Christopher Groth and Markus Muckenhuber, respectively. Differential expression analysis was conducted by Carsten Sticht. Adapted from Groth *et al.* (2023).

In the following paragraphs, I utilize the comprehensive single-cell and bulk sequencing data set from mouse ESCs and MEFs (see **Table 2.10**, project 1) to examine how IFNβ induces transcription of specific target genes. The genetically identical ESCs and MEFs provide a model to study both the common and cell type-specific induction of ISG expression. Using these ISGs, I aimed to investigate different mechanisms of proximal

and distal transcription regulation and their effects on the transcriptional response to IFNβ. Additionally, the comparison of the two cell types allowed me to assess the regulatory effects of different epigenetic modifications and consequent chromatin states at CREs in an otherwise genetically identical environment.

IFNβ induces cell type-specific gene expression changes

To characterize the transcriptional response of ESCs and MEFs to IFNβ treatment, we performed differential expression analysis of bulk RNA-seq data between unstimulated and 1 h as well as 6 h IFNβ-stimulated ESCs and MEFs. This analysis identified between 57 to 452 upregulated ISGs per cell type and time point (**Figure 2.38A**, red). In contrast, both cell types showed only a few genes with significantly reduced expression after IFNβ treatment. Furthermore, both cell types exhibited a lower expression response after 1 h (30 % of ISGs for ESCs, 25 % for MEFs) compared to 6 h of IFNβ treatment. At both time points, MEFs showed three- to fourfold more ISGs than ESCs (in total 191 ISGs in ESCs, 463 in MEFs). Additionally, the magnitude of the expression increase was generally higher in MEFs. Among the two cell types, the majority of ISGs detected in ESCs were also detected in MEFs (**Figure 2.38B**, orange), while only 17 % of ISGs were specific to ESCs (**Figure 2.38B**, green). In contrast, 66 % of ISGs were specific to MEFs (**Figure 2.38B**,

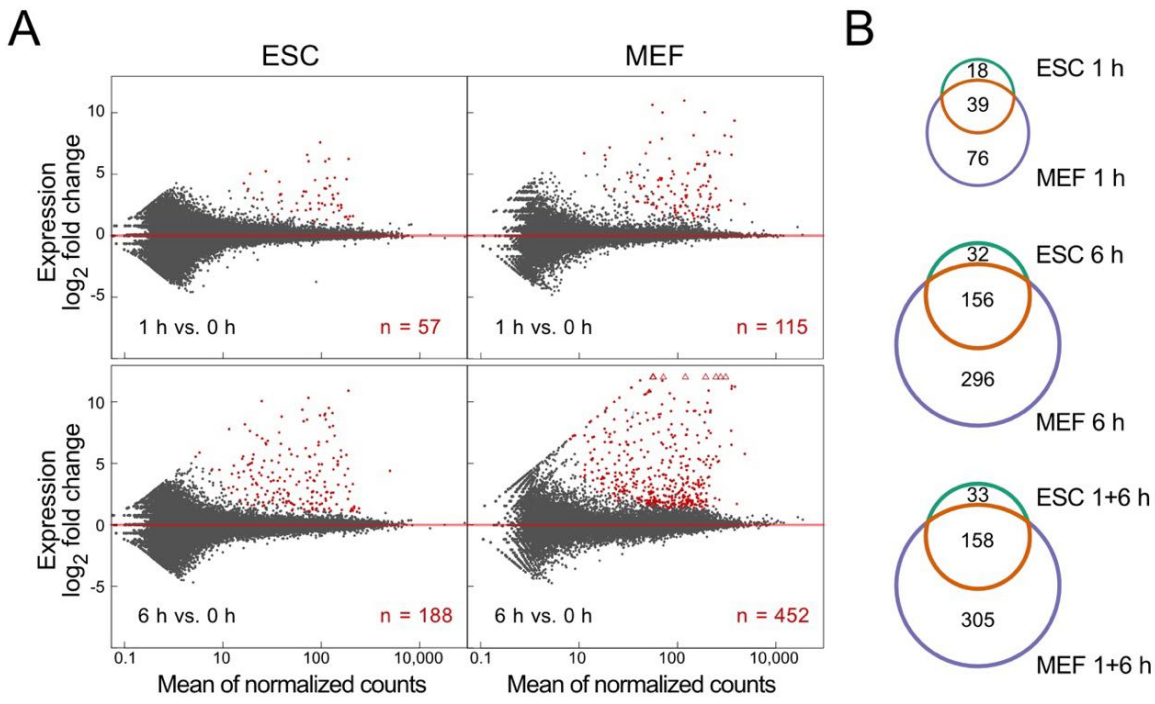


Figure 2.38 IFN-stimulated genes (ISGs) in ESCs and MEFs after 1 h and 6 h of IFNβ treatment from bulk RNA-seq data. Data from four biological replicates of ESCs and two biological replicates of MEFs are shown. **A** Gene expression changes after 1 h (top) and 6 h (bottom) of IFNβ stimulation in ESCs (left) and MEFs (right). ISGs with a log2FC ≥ 1.5 and adjusted p-value < 0.05 are marked in red. **B** Overlap of ISGs between ESCs and MEFs. ISGs after 1 h IFNβ (top), 6 h IFNβ (middle) and both 1 h and 6 h IFNβ (bottom) are shown. Analysis was performed by Markus Muckenhuber. Adapted from Muckenhuber *et al.* (2023).

purple). Overall, both ESCs and MEFs show upregulated expression of specific ISGs upon IFNβ treatment. Notably, many ISGs were common among the cell types and only few ESC-specific ISGs were detected. This suggests that the transcriptional response is predominantly regulated by the IFNβ-induced TFs in both cell types, rather than by their potentially different chromatin states of the CREs.

Single cell expression profiles show uniform responses to IFNβ

Next, we studied gene expression of ESCs and MEFs upon IFNβ stimulation at single-cell resolution to determine whether the upregulation of ISG expression was homogeneous among the cells. Low-dimensional embedding of scRNA-seq data from ESCs revealed a homogeneous distribution of cells for each IFNβ treatment time point (**Figure 2.39A**, left). ESCs treated with IFNβ for 6 h formed a distinct cluster, referred to as C0, which was separate from unstimulated and 1 h stimulated ESCs, that were mixed within cluster C1. In contrast, the low-dimensional embedding of scRNA-seq data from MEFs showed two subpopulations that were distinct across all IFNβ treatment time points (**Figure 2.39A**, right). Within these subpopulations, MEFs were uniformly distributed at each IFNβ

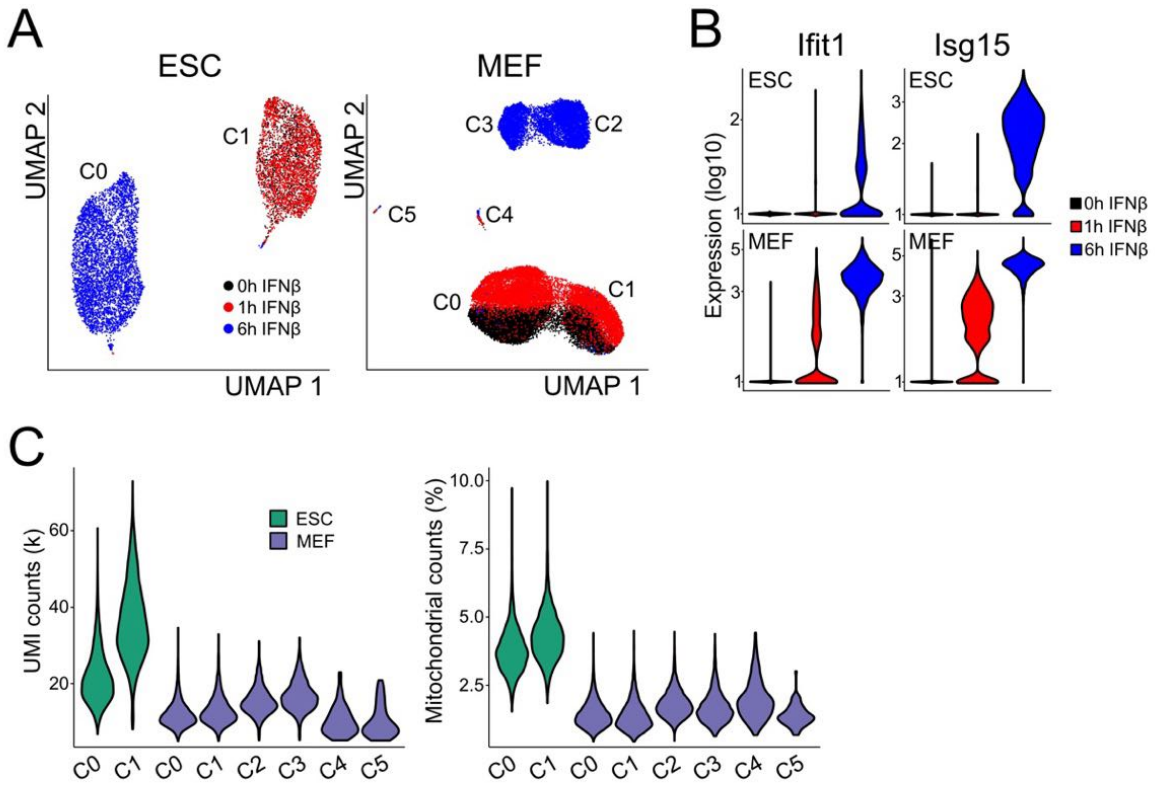


Figure 2.39 Gene expression response to IFNβ treatment in ESCs and MEFs at single-cell resolution from scRNA-seq data. Data from one biological replicate of each ESCs and MEFs are shown. **A** Low-dimensional embedding of ESCs (left) and MEFs (right). Cells are colored according to their IFNβ treatment time points. K-nearest neighbor clusters are labeled by number. **B** Single-cell expression levels of ISGs *Ifit1* and *Isg15* in ESCs (top) and MEFs (bottom). **C** Single-cell UMI counts (left) and percent mitochondrial counts (right) for ESC and MEF clusters. Analysis was performed by Markus Muckenhuber. Adapted from Muckenhuber *et al.* (2023).

treatment time point. Similar to ESCs, MEFs treated with IFNβ for 6 h (clusters C2 and C3) were separate from unstimulated and 1 h stimulated MEFs (clusters C0 and C1). Additionally, unstimulated and 1 h stimulated MEFs were distinct within clusters C0 and C1, but did not form separate clusters. The more pronounced separation of unstimulated and 1 h stimulated MEFs, compared to ESCs, was consistent with the stronger ISG induction observed in MEFs after 1 h of IFNβ stimulation in bulk RNA-seq (see **Figures 2.38A+B**). Accordingly, the expression of well-characterized ISGs such as *Ifit1* and *Isg15* significantly increased in both ESCs and MEFs after 6 h of IFNβ stimulation (**Figure 2.39B**). However, MEFs also showed upregulated expression of these ISGs after just 1 h of IFNβ stimulation. The increase in expression for both ISGs was stronger in MEFs than in ESCs, despite ESCs having generally higher numbers of total and percent mitochondrial UMIs, indicating higher data quality for ESCs (**Figure 2.39C**).

For MEFs, clusters C4 and C5 were excluded due to their low UMI counts (**Figure 2.39C**, left) and the low number of cells originating from all three IFNβ treatment time points (**Figure 2.39A**, right). Within the remaining MEF clusters, the IFNβ response appeared to be overall similar for well-studied ISGs (**Figure 2.40A**). Minor differences in expression were observed for *Ifi27* in the unstimulated and 1 h IFNβ-stimulated MEF clusters C0 and

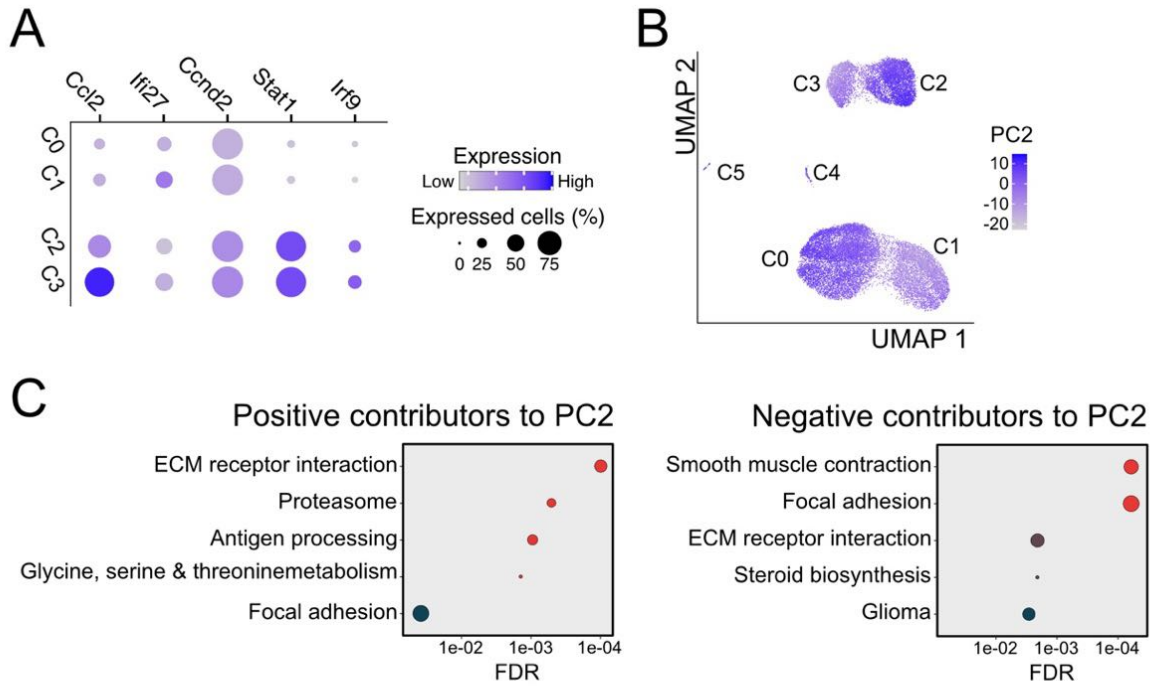


Figure 2.40 Differences between MEF subpopulations untreated and after 1 h and 6 h of IFNβ treatment from scRNA-seq data. Data from one biological replicate of MEFs are shown. **A** Expression levels of ISGs *Irf9*, *Stat1*, *Ccnd2*, *Ifi27* and *Ccl2* in MEF clusters 0-3. **B** Low-dimensional embedding of MEFs. Cells are colored according to their PC2 signal. **C** KEGG pathway enrichment for genes with positive (left) and negative (right) contributions to PC2. The size of the points reflects the number of genes in each KEGG pathway, and the color reflects the FDR. Analysis was performed by Markus Muckenhuber. Adapted from Muckenhuber *et al.* (2023).

C1. Similarly, *Ccl2* showed varying levels of expression in the 6 h IFNβ-stimulated MEF clusters C2 and C3. However, for all exemplary ISGs, the upregulated expression in bulk RNA-seq data was confirmed for both MEF subpopulations in the scRNA-seq analysis, suggesting that the IFNβ stimulation was not driving their separation into distinct clusters across all treatment time points. Instead, we observed that the separation of the two MEF subpopulations was captured by PC2 (**Figure 2.40B**). A KEGG pathway analysis of the genes that contributed positively to PC2 showed the strongest enrichment in processes related to extracellular matrix receptor interaction, while the genes contributing negatively to PC2 were significant enriched in processes associated with smooth muscle contraction and focal adhesion (**Figure 2.40C**). This suggested that the two MEF subpopulations differed in their states during the epithelial-to-mesenchymal transition. MEF clusters C0 and C2, showing high PC2 scores, were annotated as “mesenchymal-like MEFs”, while clusters C1 and C3, with low PC2 scores, were annotated as “epithelial-like MEFs”.

STAT1/2 binding at ISG promoters correlates with increased expression

To investigate the mechanisms regulating increased ISG expression, we investigated genome-wide binding sites of the TFs STAT1 and STAT2 using bulk ChIP-seq data of ESCs and MEFs. STAT1 and STAT2 peaks were called from the ChIP-seq data of ESCs and MEFs at the IFNβ treatment time points separately, indicating STAT1 and STAT2 binding sites. A subset of these STAT1 and STAT2 binding sites showed significantly higher signal after 1 h and 6 h of IFNβ treatment compared to the unstimulated condition. Consequently, these binding sites with increased signal were considered IFNβ-induced STAT1 and STAT2 binding sites. In ESCs, STAT1 binding was induced at 1,133 genomic sites, while only 236 genomic sites showed an increase in STAT2 binding (**Figure 2.41A**, top). Notably, 88 % of these induced STAT2 binding sites also showed simultaneous STAT1 binding, indicating co-binding of STAT1 and STAT2. In the following, these co-bound sites are termed induced STAT1/2 binding sites. In contrast, in MEFs, STAT1 binding was induced at fewer genomic sites (426), while STAT2 binding was induced at more genomic sites (574) compared to ESCs (**Figure 2.41A**, bottom). Similar to ESCs, a high proportion of the induced binding sites in MEFs showed co-binding of STAT1 and STAT2 (65 % of STAT1 binding sites and 48 % of STAT2 binding sites). When comparing STAT1 and STAT2 binding sites between ESCs and MEFs, most individual STAT1 and STAT2 binging sites (96 % and 99 %, respectively) were cell type-specific (**Figure 2.41B**). In contrast, 44 % and 33 % of the induced STAT1/2 binding sites overlapped between ESCs and MEFs, respectively.

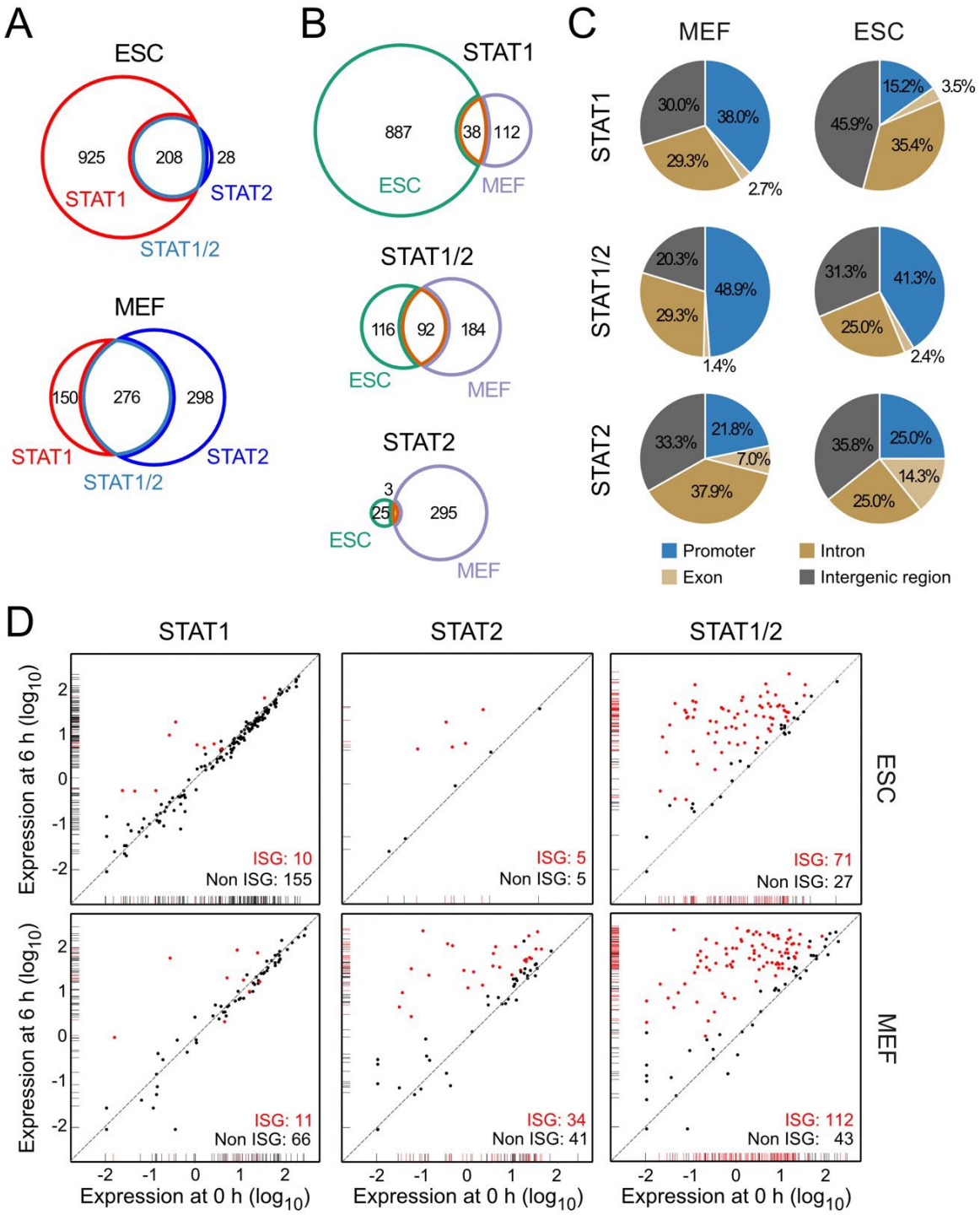


Figure 2.41 STAT1 and STAT2 binding in ESCs and MEFs after 1 h and 6 h of IFNβ treatment from bulk ChIP-seq data. Bulk ChIP-seq of STAT1 and STAT2 as well as bulk RNA-seq data from four biological replicates of ESCs and two biological replicates of MEFs are shown. **A** Overlap of STAT1 and STAT2 binding sites after 1 h and 6 h of IFNβ treatment in ESCs (top) and MEFs (bottom). **B** Overlap of STAT binding sites between ESCs and MEFs. STAT1 (top), STAT1/2 (middle) and STAT2 (bottom) binding sites are shown. **C** Genomic location of STAT1, STAT1/2 and STAT2 binding sites in MEFs (left) and ESCs (right). **D** Gene expression after 0 h and 6 h of IFNβ stimulation in ESCs (top) and MEFs (bottom) for genes with STAT1, STAT2 and STAT1/2 peak at the promoter. ISGs are labeled in red. Analysis was performed by Markus Muckenhuber. Adapted from Muckenhuber *et al.* (2023).

For both ESCs and MEFs, the genomic locations of the induced STAT1 and STAT2 binding sites showed a diverse distribution across promoters, gene bodies, and intergenic regions (**Figure 2.41C**, top and bottom). Only a small proportion of induced STAT1 and STAT2 binding sites were located in promoter and exonic regions (19-41 %), while the majority were found in intronic and intergenic regions (59-81 %). Conversely, approx. 40-50 % of induced STAT1/2 binding sites in ESCs and MEFs were located at promoters (**Figure 2.41C**, middle). This fraction was even higher for induced STAT1/2 binding sites shared between ESCs and MEFs, with 76 % located at gene promoters.

Next, we investigated the relationship between induced STAT1, STAT2, and STAT1/2 binding sites at promoters and their respective gene expression changes after 6 h of IFNβ treatment. For induced STAT1 binding sites at promoters, only 6 % and 14 % of genes showed significantly upregulated expression in ESCs and MEFs, respectively (**Figure 2.41D**, left). In contrast, 50 % and 45 % of genes with induced STAT2 binding sites at the promoter showed significantly increased gene expression in ESCs and MEFs (**Figure 2.41D**, middle). The number of ISGs was even higher for induced STAT1/2 binding sites at promoters, where 73 % and 72 % of genes showed significantly higher expression after 6 h of IFNβ treatment in ESCs and MEFs (**Figure 2.41D**, right). In summary, the strong induction of STAT1, STAT2 and STAT1/2 binding supports the previous observation that the common transcriptional response between the cell types is predominantly regulated by IFNβ-induced TFs. In this context, simultaneous STAT1/2 binding was the most prominent driver of ISG activation in both ESCs and MEFs, particularly through direct binding to gene promoters. In contrast, STAT1 and STAT2 binding sites were cell type-specific and more often found in promoter-distal genomic regions.

IFNβ treatment increases accessibility at STAT1/2 binding sites

Simultaneous STAT1/2 binding at promoters was identified as a key activator of ISG expression. However, the role of non-promoter STAT1/2 binding sites (accounting for 51-59 %) in regulating ISG expression remains unclear. To further characterize the chromatin state at ISGs and STAT1/2 binding sites, I used scATAC-seq data of unstimulated and IFNβ-stimulated ESCs and MEFs. The low-dimensional embedding of chromatin accessibility profiles of ESCs revealed a homogeneous distribution of cells (**Figure 2.42A**, left). Furthermore, there was no separation between unstimulated and 6 h IFNβ-stimulated ESCs in the low-dimensional embedding, indicating no global changes in chromatin accessibility profiles following IFNβ treatment. Similarly, the low-dimensional embedding of scATAC-seq data for MEFs showed a homogeneous distribution of cells across all IFNβ treatment time points (**Figure 2.42A**, right). Consistent with the scRNA-seq data for MEFs,

the low-dimensional embedding revealed two MEF subpopulations that were present across all IFNβ treatment time points (clusters C2 and C3). While there was no distinct separation based on treatment condition for either subpopulation, MEFs treated with IFNβ for 6 h showed local enrichment within the MEF clusters (**Figure 2.42A**, right; note the increased blue intensity in the top right of clusters C2 and C3). However, it is unclear whether this enrichment reflects a biological effect of global changes in chromatin accessibility due to IFNβ treatment or is a result of lower numbers of unique fragments in the 6 h IFNβ-stimulated MEFs (**Figure 2.42B**). The MEF subpopulations were classified as epithelial-like (cluster C2) and mesenchymal-like (cluster C3) based on integration with previously annotated scRNA-seq data (**Figure 2.42C**). I excluded ESC cluster C1 and MEF cluster C1 from further analyses, due to their low UMI counts, low cell numbers, and the unassigned MEF subtype (**Figures 2.42A+C**).

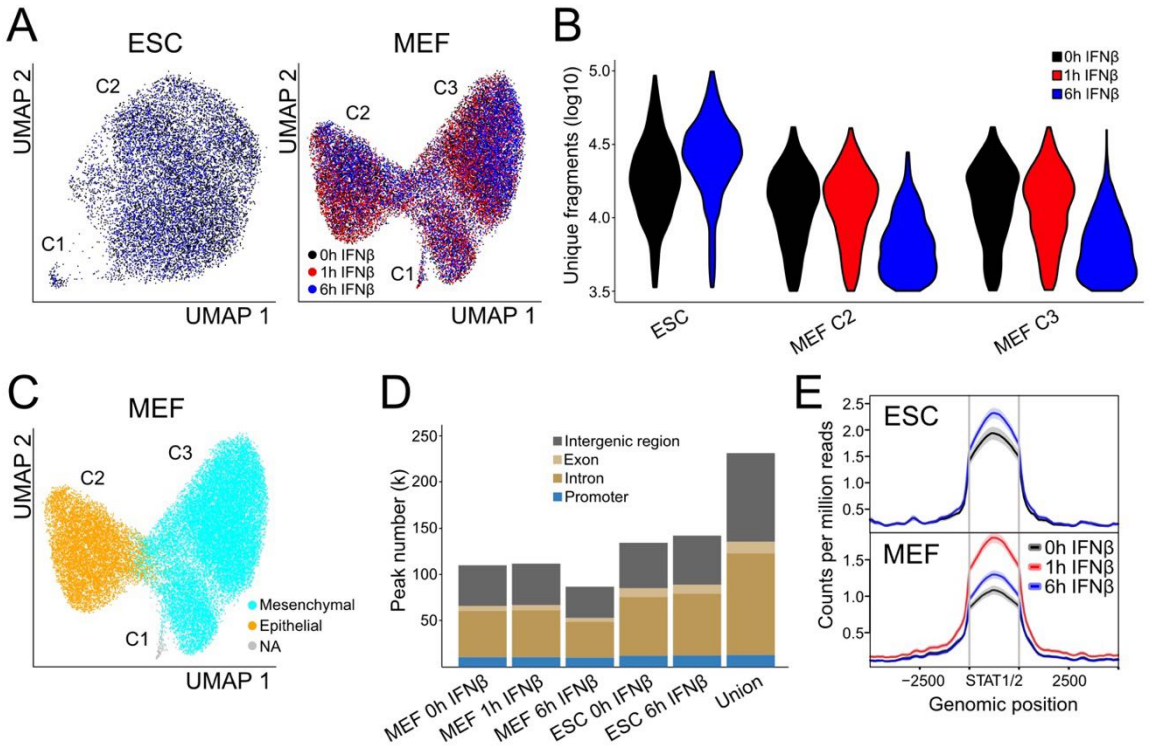


Figure 2.42 Chromatin accessibility in unstimulated and IFNβ-treated ESCs and MEFs at single-cell resolution from scATAC-seq data. Data from one biological replicate of each ESCs and MEFs are shown. **A** Low-dimensional embedding of ESCs (left) and MEFs (right). Cells are colored according to the IFNβ treatment time point. K-nearest neighbor clusters are labeled by number. **B** Single-cell unique fragment counts per time point for ESCs and MEF clusters. **C** Low-dimensional embedding of MEFs. Cells are colored according to their MEF subtypes, derived from integrated scRNA-seq data. **D** Pseudo-bulk ATAC peaks from separate IFNβ treatment time points and cell types, as well as the union of merged ATAC peaks. Peaks are annotated by their genomic position. **E** Chromatin accessibility at STAT1/2 binding sites per time point in ESCs (top) and MEFs (bottom). Adapted from Muckenhuber *et al.* (2023).

Next, I performed ATAC peak calling on pseudo-bulk samples across cell types and all treatment conditions. This identified 231,170 genomic sites with high accessibility signals (**Figure 2.42D**). In contrast, the pseudo-bulks of the individual samples yielded only roughly 75,000-140,000 ATAC peaks. The majority of these ATAC peaks were located within genes, with around 10 % specifically at gene promoters. In ESCs, 244 ATAC peaks (0.1 %) exhibited significantly higher accessibility after 6 h of IFNβ treatment, while only 5 ATAC peaks (0.002 %) showed significantly decreased accessibility (**Table 2.11**). In both epithelial- and mesenchymal-like MEFs, only 70 and 49 ATAC peaks (0.03 % and 0.02 %) demonstrated significantly increased accessibility after 6 h of IFNβ treatment, respectively, while no ATAC peaks showed a significant reduction of accessibility. After 1 h of IFNβ treatment in MEF subtypes, a more ATAC peaks (0.1 % in epithelial-like and 0.1 % in mesenchymal-like MEFs) showed significantly increased accessibility, while only 1 and 4 ATAC peaks exhibited significantly reduced accessibility, respectively.

Table 2.11 Differential accessibility analysis of pseudo-bulk ATAC peaks between untreated and 1 h or 6 h IFNβ-treated ESCs and MEFs.

Number of ATAC peaks with	ESCs	Epithelial-like MEFs	Mesenchymal-like MEFs
Increased accessibility at 1 h vs. 0 h	-	271	273
Decreased accessibility at 1 h vs. 0 h	-	1	4
Increased accessibility at 6 h vs. 0 h	244	70	49
Decreased accessibility at 6 h vs. 0 h	5	0	0

Similar to the scRNA-seq data (see **Figure 2.40A)**, changes in chromatin accessibility in response to IFNβ treatment were highly similar between MEF subtypes. The comparison between MEFs and ESCs revealed a strong early response in MEFs, whereas ESCs exhibited a higher number of significantly differential ATAC peaks after 6 h of IFNβ treatment. The overall low number of significantly differential ATAC peaks in both ESCs and MEFs suggests that chromatin accessibility changes in response to IFNβ treatment were highly specific. This specificity was likely driven by the targeted induction of STAT1/2 binding events, as pseudo-bulk chromatin accessibility at previously identified induced STAT1/2 binding sites specifically increased after IFNβ treatment (**Figure 2.42E**).

STAT1/2 activates ISG expression by binding to distal enhancers

Since the non-promoter STAT1/2 binding events did not induce global alterations in chromatin accessibility, the question remained as to how these STAT1/2 co-bound sites regulate ISG expression. To address this, I performed *single cell co-accessibility* analysis

between ATAC peaks within 1 Mb windows around STAT1/2 binding sites. For ESCs and MEFs, 37 % and 24 % of ISGs had induced STAT1/2 binding sites at their promoters, respectively (**Figure 2.43A**, blue). The chromatin co-accessibility analysis revealed that approximately 25 % of ISGs without promoter STAT1/2 binding showed autonomous links of co-accessibility (ACs, see **Chapter 2.2**) between their promoter and a distal STAT1/2 binding site. The majority of these ISGs (73-86 %) gained their AC between promoter and distal STAT1/2 binding site upon IFNβ treatment (**Figure 2.43A**, green), while a small fraction of ISGs lost their AC after IFNβ stimulation (**Figure 2.43A**, red). The remaining ISGs (44-54 %) showed neither a STAT1/2 binding site at their promoter, nor an AC to a distal STAT1/2 binding site (**Figure 2.43A**, grey), suggesting STAT1/2-independent

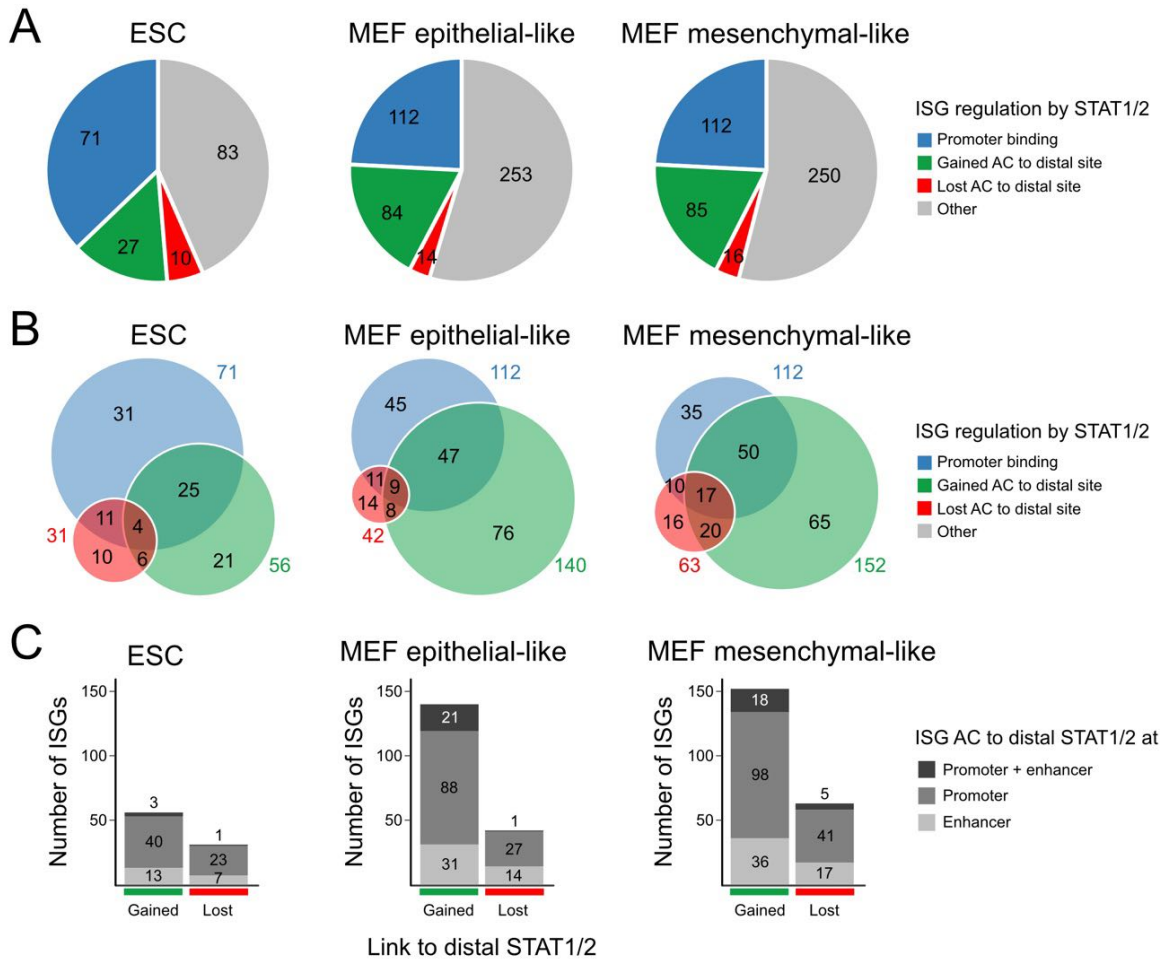


Figure 2.43 Distal STAT1/2 regulation of ISG expression in ESCs and MEFs from *single cell co-accessibility* analysis of scATAC-seq data. Data from one biological replicate of ESCs (left), epithelial-like (middle) and mesenchymal-like MEFs (right) are shown. **A** Regulation of ISGs by STAT1/2 in the respective cell types**.** ISGs are categorized successively based on the presence of a STAT1/2 binding site at the promoter (blue), gained (green) or lost (red) AC between the promoter and a distal STAT1/2 binding site after IFNβ treatment, and other mechanisms of regulation (grey). **B** Overlap of ISG regulation mechanisms by STAT1/2 in the respective cell types. **C** ISGs with gained or lost AC between the promoter and a distal STAT1/2 binding site after IFNβ treatment in the respective cell types. ISGs are annotated by the genomic position of the linked site, which is classified as either another distal STAT1/2-bound ISG promoter and/or a potential enhancer in gene bodies or intergenic regions. Adapted from Muckenhuber *et al.* (2023).

regulation of transcription. Furthermore, roughly 30 % of ISGs showed multiple connections to induced STAT1/2 binding sites, involving both proximal and AC-linked distal STAT1/2 binding (**Figure 2.43B**). Notably, most STAT1/2 binding sites (75 %) with ACs to distal ISG promoters were located in other gene promoters (**Figure 2.43C**), with only 25 % found in intergenic regions or gene bodies. This suggests that induced STAT1/2 binding sites can act as potential *distal CREs*, regardless of their genomic location, whether at promoters, coding or non-coding regions.

One example of ISG regulation by a putative *distal CRE* is the ISG *Uba7* in ESCs. *Uba7* expression is significantly upregulated after 6 h of IFNβ treatment (**Figure 2.44A**, bottom) and does not show induced STAT1/2 binding at its promoter. No ACs were detected prior to IFNβ stimulation, but a distinct AC between the *Uba7* promoter and a distal STAT1/2 binding site appeared after 6 h of IFNβ treatment (**Figure 2.44A**, middle). Simultaneously with the formation of the AC, the pseudo-bulk accessibility signal at both the promoter and the distal enhancer increased after 6 h of IFNβ stimulation (**Figure 2.44A**, top). This indicates that the increase of *Uba7* expression was regulated by an emerging chromatin contact between its promoter and a *distal CRE* with induced STAT1/2 binding.

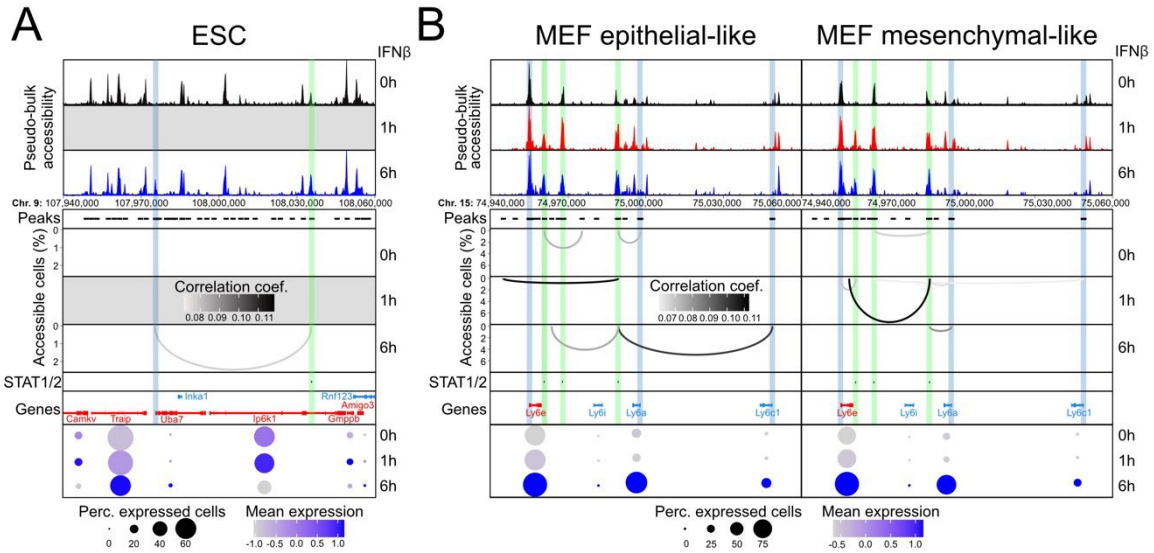


Figure 2.44 Examples of distal STAT1/2 regulation of ISG expression in ESCs and MEFs from *single cell co-accessibility* analysis of scATAC-seq data. Data from one biological replicate of each ESCs and MEFs are shown. **A** Pseudo-bulk accessibility signal (top), ACs (middle), and gene expression (bottom) for the ISG *Uba7* in ESCs. The ISG promoter is marked in blue. All ACs from STAT1/2 binding sites (green) are shown. Log10 normalized gene expression levels from scRNA-seq data are shown. **B** Same as panel A but for the ISGs *Ly6a*, *Ly6c1* and *Ly6e* in MEFs. scRNA-seq analysis was performed by Markus Muckenhuber. Adapted from Muckenhuber *et al.* (2023).

Additionally, I investigated another example of ISG regulation by a whole cluster of *distal CREs* at the *Ly6* gene cluster in MEFs. Here, the expression of *Ly6a*, *Ly6c1*, and *Ly6e* was significantly upregulated after 6 h of IFNβ treatment in both MEF subtypes (**Figure**

2.44B, bottom). Again, no STAT1/2 binding sites were present at the ISG promoters. However, three distal STAT1/2 binding sites between *Ly6e* and *Ly6a* were induced by IFNβ, forming a potential cluster of *distal CREs*. The pseudo-bulk accessibility revealed consistently high accessibility at the *Ly6e* promoter and low accessibility at the *Ly6a* and *Ly6c1* promoters, independent of IFNβ treatment, suggesting that transcription induction was not regulated at the promoters directly (**Figure 2.44B**, top). In contrast, the pseudo-bulk accessibility at all three STAT1/2 binding sites strongly increased upon IFNβ stimulation, which indicated their regulatory activity in the transcriptional response to IFNβ. Between these induced STAT1/2 binding sites and the ISG promoters, multiple ACs were detected (**Figure 2.44B**, middle). Most of these ACs changed upon IFNβ stimulation, including the formation of new ACs and the loss of existing ones. Notably, epithelial- and mesenchymal-like MEFs showed differences in both pseudo-bulk accessibility profiles and ACs, indicating regulatory differences between these MEF subtypes.

Remarkably, the depicted ACs in both ESCs and MEFs exhibited very low PAC values (**Figures 2.44A+B**, middle; note height of loops), which represent the detection rate of an AC among the single cells (see **Section 2.2.1**). This finding was consistent with the previously observed PAC distribution of genome-wide ACs in 6 h IFNβ-treated epithelial-like MEFs, where approximately 90 % of ACs had PAC values below 25 (see **Figure 2.24A**, left). As discussed in **Section 2.2.1**, these low PAC values were likely due to the lower data quality of scATAC-seq data compared to scTurboATAC-seq data, with mean unique fragments per cell ranging from $10^{3.8}$ to $10^{4.5}$ for both ESCs and MEFs (see **Figure 2.42B**). In contrast, scTurboATAC-seq data from MEFs had a higher mean of $10^{4.8}$ unique fragments per cell (see **Table 2.6**), and scTurboATAC-seq data from HUVECs had even higher means ranging from $10^{4.8}$ to $10^{5.3}$ unique fragments per cell (see **Figure 2.20B**). Consequently, the scATAC-seq data of MEFs and ESCs from this analysis had insufficient quality to accurately differentiate between *rare* and *frequent ACs*, as was possible before with higher-quality data.

In summary, *single cell co-accessibility* analysis revealed that non-promoter STAT1/2 binding events regulate ISG expression through specific long-range chromatin interactions, as detected by ACs between ISG promoters and distal STAT1/2 binding sites. Additionally, the analysis showed that induced STAT1/2 binding sites can act as potential *distal CREs*, regardless of their genomic location, whether at promoters, coding or non-coding regions. Interestingly, the regulatory interactions both emerged and disappeared upon IFNβ stimulation, suggesting both either activating or repressive effects. Furthermore, STAT1/2-independent regulation of transcription was observed for

approximately 50 % of ISGs. These ISGs might be regulated by other TFs among the initial STAT1/2-induced ISGs, so-called secondary targets of IFNβ. When comparing these regulatory mechanisms between ESCs and MEFs, ESCs showed a higher fraction of promoter mediated ISG regulation by STAT1/2 compared to MEFs. In contrast, MEFs revealed a higher fraction of STAT1/2-independent ISG regulation compared to ESCs. These findings align with the observed higher number of MEF-specific ISGs, likely having a more diverse secondary response in MEFs than ESCs.

Proximal and distal ISG regulation varies in strength of expression induction

After identifying these different mechanisms of distal ISG regulation by induced STAT1/2 binding sites, we investigated their impact on expression induction. Overall, the mean expression increase of ISGs was lower after 1 h of IFNβ treatment (below 0.25) compared to their mean expression increase above 0.5 after 6 h (**Figure 2.45A**), consistent with previous observations of less pronounced expression induction after 1 h of IFNβ treatment (see **Figure 2.38A).** For ESCs and both MEF subtypes, ISGs with STAT1/2 promoter binding exhibited significantly higher expression induction after 1 h and 6 h of IFNβ treatment compared to ISGs with AC-linked distal STAT1/2 binding sites or STAT1/2-independent regulation (**Figure 2.45A**, blue vs. all others). Furthermore, ISGs that gained an AC to a distal STAT1/2 binding site demonstrated significantly stronger expression increases compared to STAT1/2-independent ISGs (**Figure 2.45A**, green vs. grey). Overall, induced STAT/1/2 binding at promoters caused the strongest induction of ISG expression, followed by ISGs showing activating chromatin interactions with distal STAT1/2 binding sites upon IFNβ stimulation. In contrast, ISGs with both preexisting repressive chromatin interactions with distal STAT1/2 binding sites as well as STAT1/2-independent regulation showed only moderate induction of expression. Notably, only the induction of STAT1/2 binding at promoters exhibited a fast upregulation of ISG expression at the 1 h IFNβ treatment time point. All other regulatory mechanisms by distal STAT1/2 binding and STAT1/2 independency showed slower transcriptional responses with significant upregulation only at the 6 h IFNβ treatment time point.

Additionally, I assessed the expression levels under unstimulated conditions for the differently regulated ISGs (**Figure 2.45B**). In ESCs and mesenchymal-like MEFs, ISGs with preexisting and subsequently lost ACs between their promoter and a distal STAT1/2 binding site showed a significantly lower expression compared to ISGs with STAT1/2-bound promoters. This confirmed that these preexisting ACs, which were lost upon IFNβ treatment, repressed the expression of ISGs in the unstimulated condition. The IFNβ-

induced STAT1/2 binding resolved these ACs and facilitated the upregulation of distal ISG expression.

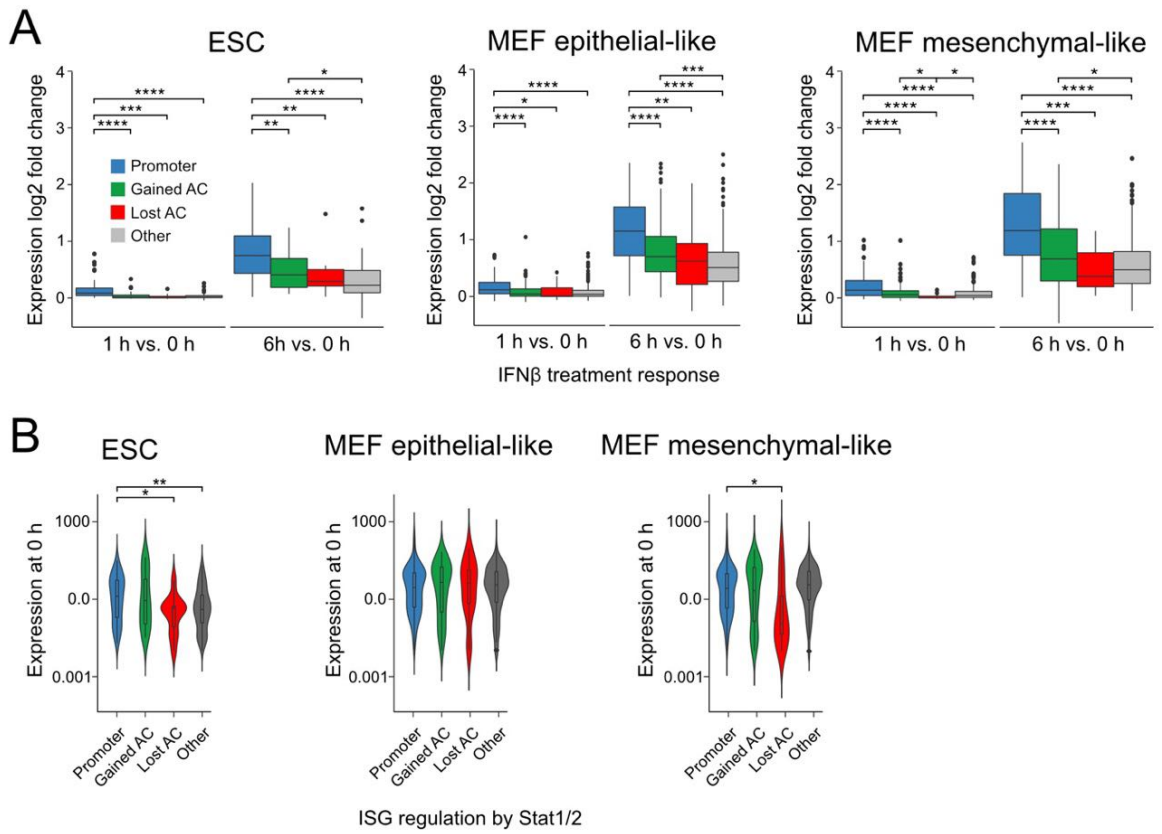


Figure 2.45 ISG expression for varying STAT1/2 regulation mechanisms in ESCs and MEFs from bulk RNA-seq data. Bulk RNA-seq data from four biological replicates of ESCs and two biological replicates of MEFs are shown. **A** ISG expression changes after 1 h and 6 h of IFNβ treatment in ESCs (left), epithelial-like (middle), and mesenchymal-like MEFs (right). Log2FCs are shown for different STAT1/2 regulation mechanisms. Significant p-values from Wilcoxon test are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001. **B** ISG expression prior to IFNβ treatment in ESCs (left), epithelial-like (middle), and mesenchymal-like MEFs (right). TPM values are shown for different STAT1/2 regulation mechanisms. Significant p-values from Wilcoxon test are indicated as *, P < 0.05; **, P < 0.01. Analysis was performed by Markus Muckenhuber. Adapted from Muckenhuber *et al.* (2023).

IFNβ induces domains of increased co-accessibility at ISGs

Next, I conducted *metacell co-accessibility* analysis for both ESCs and MEFs to determine whether ISGs are additionally regulated by broad domains of increased co-accessibility, previously described as DCs that represent nuclear subcompartments with locally increased TF binding activity (see **Chapter 2.2**). Visual inspection of *metacell co-accessibility* maps at ISGs revealed only few regions with domains of locally increased co-accessibility. In general, the co-accessibility scores appeared more variable, with both positive and negative co-accessibility scores showing less distinct patterns. This variability was likely due to the lower quality of this dataset, as previously discussed for ACs.

Nonetheless, I identified two exemplary ISG regions with domains of enriched co-accessibility, one in ESCs and one in epithelial-like MEFs.

In ESCs, a broad domain of approximately 400 kb showed enriched co-accessibility around three ISGs: *Psme1*, *Psme2*, and *Irf9* (**Figure 2.46A**). Within this domain, only one STAT1/2 binding site, located at the promoter of *Irf9*, was induced upon IFNβ treatment. This induced STAT1/2 binding site at the *Irf9* promoter likely led to a fast expression induction, since *Irf9* expression was significantly upregulated after both 1 h and 6 h of IFNβ treatment in bulk RNA-seq data. However, the single induced STAT1/2 binding site in the observed domain of high co-accessibility suggests that the domain was not driven by increased local STAT1/2 binding activity. In contrast to *Irf9*, *Psme1* and *Psme2* showed significantly increased expression only after 6 h of IFNβ treatment, potentially induced by the local enrichment of secondary TF targets of IFNβ in the domain at this later time point. In epithelial-like MEFs, a domain of approximately 200 kb at the ISG *Rnf213* displayed enriched co-accessibility (**Figure 2.46B**). This smaller domain included only the *Rnf213* gene body and roughly 100 kb downstream of the gene. Again, only one IFNβ-induced STAT1/2 binding site was present, located at the upstream promoter of *Rnf213*. In bulk RNA-seq data, *Rnf213* showed a four-fold increase in expression after 1 h of IFNβ treatment, with a further increase by more than ten-fold at the 6 h treatment time point. The proximal STAT1/2 binding at the promoter likely caused fast *Rnf213* expression upregulation, while the subsequent upregulation might have been further accellerated by local enrichment of secondary TF targets of IFNβ in the domain at the later treatment time point.

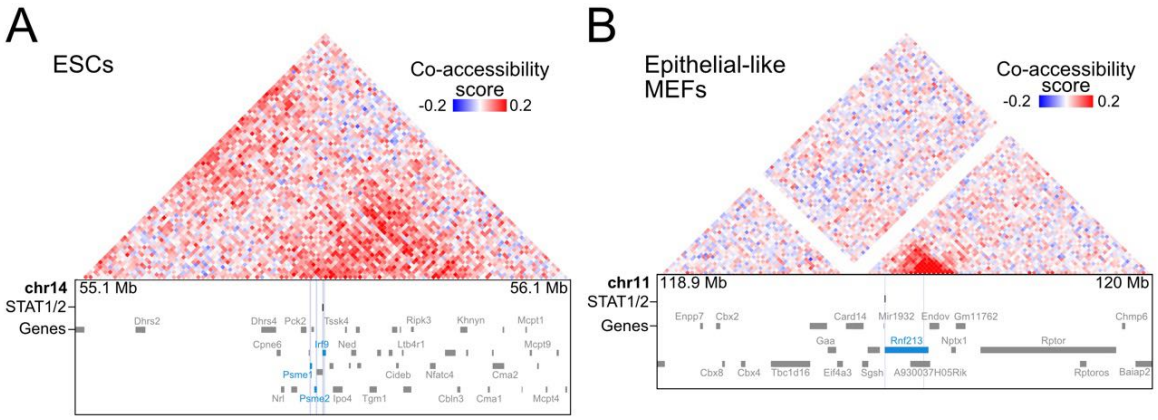


Figure 2.46 Exemplary IFNβ-induced subcompartments at ISGs in ESCs and MEFs from metacell co-accessibility analysis of scATAC-seq data. Data from one biological replicate of each ESCs and MEFs are shown. **A** *Metacell co-accessibility* maps from unstimulated and 6 h IFNβ-stimulated ESCs for a cluster of ISGs that includes *Psme1*, *Psme2*, and *Irf9*. Induced STAT1/2 binding sites (black), gene annotations (grey), ISGs (blue) and 1 kb regions around ISG TSSs (light blue) are annotated. The color scale bar is set between -0.2 and 0.2. **B** Same as panel A for unstimulated, 1 h, and 6 h IFNβ-stimulated epithelial-like MEFs for ISG *Rnf213*. Adapted from Seufert *et al.* (2024).

Overall, the observed domains of locally increased co-accessibility presumably represent IFNβ-induced nuclear subcompartments with locally increased TF binding activity (see **Section 2.2.3**). In addition to the previously described regulatory mechanisms via STAT1/2 binding to promoters and *distal CREs*, they reveal an additional mechanism of ISG regulation. This additional layer of the IFNβ response is likely mediated by secondary TF targets of IFNβ at the 6 h treatment time point, causing an attenuated, STAT1/2-independent expression induction of ISGs.

In summary, in this project I studied how IFNβ induces transcription of specific target genes, investigating different mechanisms of proximal and distal transcription regulation and their varying effects on the transcriptional response to IFNβ. In ESCs and MEFs, IFNβ treatment induces the upregulation of common and specific ISGs. This upregulation is primarily mediated by the induction of STAT1, STAT2 and STAT1/2 binding sites, rather than by global changes in chromatin state. Here, the strongest and fastest expression induction was observed for ISGs with STAT1/2 binding sites at their promoters. However, most IFNβ-induced STAT1/2 binding sites were at non-promoter regions. *Single cell co-accessibility* analysis revealed that many of these binding sites regulate ISG expression through long-range chromatin interactions. Additionally, the analysis showed that induced STAT1/2 binding sites can act as potential *distal CREs*, regardless of their genomic location at promoters, coding or non-coding regions. Interestingly, the long-range chromatin interactions both emerged and disappeared upon IFNβ stimulation. The emerging chromatin interactions between ISGs and distal STAT1/2 binding sites showed strong upregulation of expression upon IFNβ stimulation and were consequently considered to represent activating chromatin interactions. In contrast, disappearing chromatin interactions between ISGs and distal STAT1/2 binding sites showed only moderate expression increases upon IFNβ stimulation and lower basal expression levels under unstimulated condition. This suggests that they represent repressive chromatin interactions and that the IFNβ-induced STAT1/2 binding resolved the chromatin interaction and facilitated the upregulation of distal ISG expression. Additionally, *metacell co-accessibility* analysis identified domains with locally increased co-accessibility following IFNβ treatment. Interestingly, these domains did not appear to be driven by increased local STAT1/2 activity directly but rather by secondary TF targets of IFNβ at the 6 h treatment time point, causing a slower secondary expression induction of ISGs.

2.3.2. Transcription factor T-bet dependent regulation of malignant B cells in chronic lymphocytic leukemia

In addition to investigating IFNβ signaling in mouse ESCs and MEFs, I also examined transcriptional regulation in the mouse TCL1 cancer model for CLL using chromatin co-accessibility analysis (see **Section 1.3.3**). In this model system, transcription was perturbed by knocking out the *Tbx21* gene, which encodes the TF T-bet. Building on the initial characterization of *TBX21* expression and T-bet protein levels in human CLL patients by others (see **Figure 1.10**), I explored its role in tumor suppression and transcription regulation using scRNA-seq and scTurboATAC-seq data (see **Table 2.10**, project 2). The goal was to determine whether the previously identified AC and DC regulatory mechanisms of transcription do not only apply to the transcriptional response to cytokine stimulation but also to internal perturbations by TF knock-out in a cancer context.

T-bet represses cellular proliferation in malignant B cells

To investigate the molecular mechanisms underlying the tumor-suppressive role of T-bet, we acquired scRNA-seq data from TCL1 cells in *Tbx21* double knock-out, *Tbx21*$^{-/-}$, and wild type, *Tbx21*$^{+/+}$, conditions. I analyzed the combined data from 2 biological replicates for each condition, demonstrating a distinct separation of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells in their low-dimensional embedding (**Figure 2.47A**). This indicates generally different transcriptomic profiles between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. Furthermore, k-nearest neighbor clustering identified 2 major clusters (C0 and C1) that contained more than 95 % of the cells (**Figure 2.47B**). These clusters C0 and C1 represented *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ TCL1 cells, respectively. Two additional small clusters (C2 and C3) contained both *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells and exhibited lower levels of *Cd5*, and in the case of C2, lower *Cd19* expression compared to clusters C0 and C1 (**Figures 2.47C+D**). The lower

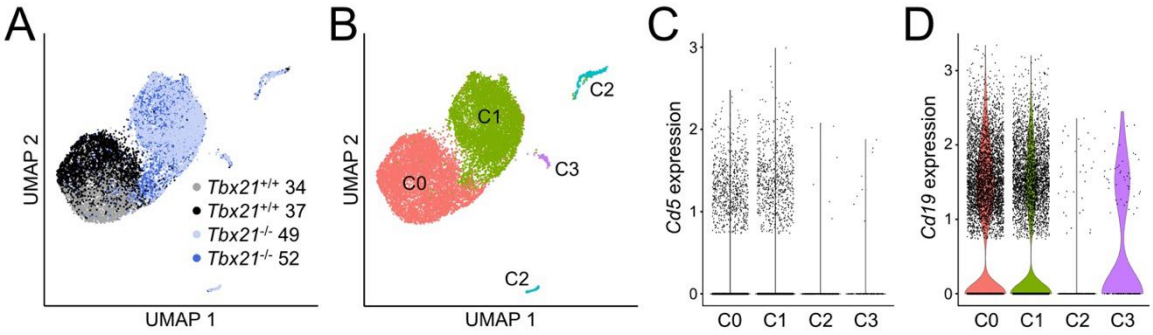


Figure 2.47 Transcriptomic profiles of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells from scRNA-seq data. Data from 2 replicates per condition are shown. **A** Low-dimensional embedding of TCL1 cells, colored according to sample. **B** Same as panel A with coloring according to k-nearest neighbor cluster. **C** *Cd5* expression per single cell in k-nearest neighbor clusters. **D** Same as panel C for *Cd19* expression. Adapted from Roessner *et al.* (2024).

expression of these malignant B cell markers, typically associated with CLL cells, suggests that clusters C2 and C3 likely represent healthy cells. Consequently, clusters C2 and C3 were excluded from further analyses.

I inferred cell cycle states of the investigated TCL1 cells, where *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells showed notable differences (**Figure 2.48A**). Approximately 80 % of *Tbx21*$^{+/+}$ TCL1 cells were in the G1 cell cycle phase in both replicates, whereas 40-50 % of *Tbx21*$^{-/-}$ TCL1 cells were in either the S or G2/M phases. The higher proportion of cells in synthesis and G2/mitotic phases indicates that *Tbx21*$^{-/-}$ TCL1 cells progress faster through the cell cycle, resulting in increased cellular proliferation. We validated this observation using phospho-specific MS data. The data revealed elevated activity of important regulators of cell cycle progression, such as cyclin-dependent kinases (CDKs) and mitogen-activated protein kinases (MAPKs), in Tbx21$^{-/-}$ versus Tbx21$^{+/+}$ TCL1 cells (**Figure 2.48B**). Finally, we measured effects of T-bet levels on cellular proliferation in cultures of CLL-like MEC-1 cell lines, where T-bet or green fluorescent protein (GFP, as control) overexpression was induced. MEC-1 cells with overexpression of T-bet showed significantly lower proliferation rates compared to those overexpressing GFP (**Figure 2.48C**). Collectively, these findings suggest that T-bet plays an important tumor-suppressive role by reducing the cellular proliferation of malignant B cells.

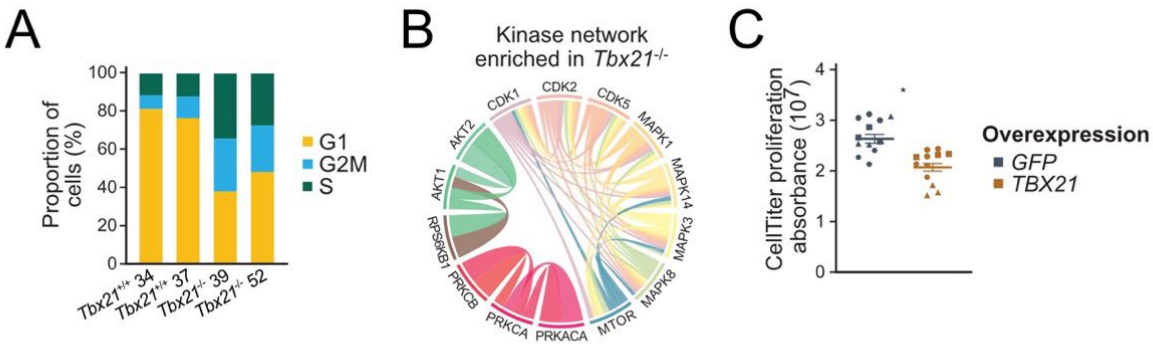


Figure 2.48 Characterization of cellular proliferation for *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells from scRNA-seq and phospho-specific MS data. A Proportion of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells in G1, G2/M, and S cell cycle state**.** ScRNA-seq data from 2 replicates per condition are shown. **B** Kinase network enriched in *Tbx21*$^{-/-}$ versus *Tbx21*$^{+/+}$ TCL1 cells. Phospho-specific MS data from 8 replicates per condition are shown. **C** CellTiter-Glo proliferation assay of MEC-1 cell lines with inducible overexpression of *TBX21* or *GFP*. Data from 4 biological replicates, each with 3 technical replicates, are shown for each cell line. P-value from unpaired t-test is indicated as *, P < 0.05. MS and CellTiter experiments were performed and analyzed by Philipp Roessner and Pavle Boskovic. Adapted from Roessner *et al.* (2024).

T-bet is a silencing TF in malignant B cells

Subsequently, I aimed to study the effects of the TF T-bet on chromatin accessibility using scTurboATAC-seq data from *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. Similar to the corresponding scRNA-seq data, the low-dimensional embedding of single-cell chromatin accessibility profiles revealed a clear separation of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells (**Figure 2.49A**). *Tbx21*$^{-/-}$ TCL1 cells were primarily located in k-nearest neighbor clusters C2 and C3, while cluster C4 contained most *Tbx21*$^{+/+}$ TCL1 cells (**Figure 2.49B**). In addition, three smaller clusters (C1, C5, and C6) contained both *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells and exhibited lower activity scores for *Cd5* and *Cd19*, as computed from the chromatin accessibility signal at the genes (**Figures 2.49C+D**). Consequently, I excluded clusters C1, C5, and C6 from further analyses, as they were considered to represent healthy cells.

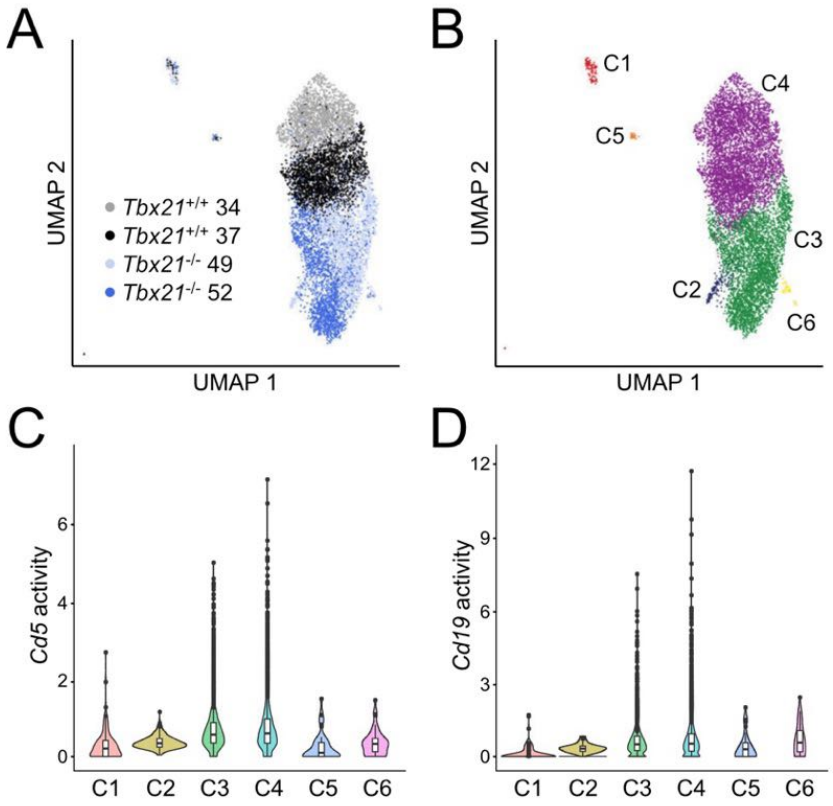


Figure 2.49 Single-cell chromatin accessibility profiles of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells from scTurboATAC-seq data. Data from 2 replicates per condition are shown. **A** Low-dimensional embedding of TCL1 cells, colored according to sample. **B** Same as panel A with coloring according to k-nearest neighbor cluster. **C** *Cd5* gene activity scores per single cell in k-nearest neighbor clusters. **D** Same as panel C for *Cd19* gene activity scores. Adapted from Roessner *et al.* (2024).

To investigate T-bet dependent changes in chromatin accessibility profiles in more detail, I performed pseudo-bulk analysis of chromatin accessibility from the scTurboATAC-seq samples and identified 133,668 ATAC peaks in *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. Of these putative regulatory regions, 4,772 ATAC peaks showed significantly higher accessibility in

Tbx21$^{-/-}$ compared to *Tbx21*$^{+/+}$ TCL1 cells (**Figure 2.50A**). Conversely, only 879 ATAC peaks exhibited reduced accessibility in *Tbx21*$^{-/-}$ TCL1 cells. Notably, these differentially accessible ATAC peaks between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells represented only about 4 % of the total ATAC peaks, suggesting that T-bet does not have a global effect on chromatin accessibility. Additionally, we performed differential accessibility analysis between *TBX21*^{high} and *TBX21*^{low} CLL cells from bulk ATAC-seq of patient samples. For the CLL patient samples, this analysis identified 1,312 ATAC peaks with significantly higher accessibility and 651 ATAC peaks with significantly lower accessibility in *TBX21*^{low} CLL cells, respectively (**Figure 2.50B**). These combined findings suggest that higher T-bet levels predominantly decrease chromatin accessibility at specific sites both in TCL1 and CLL cells.

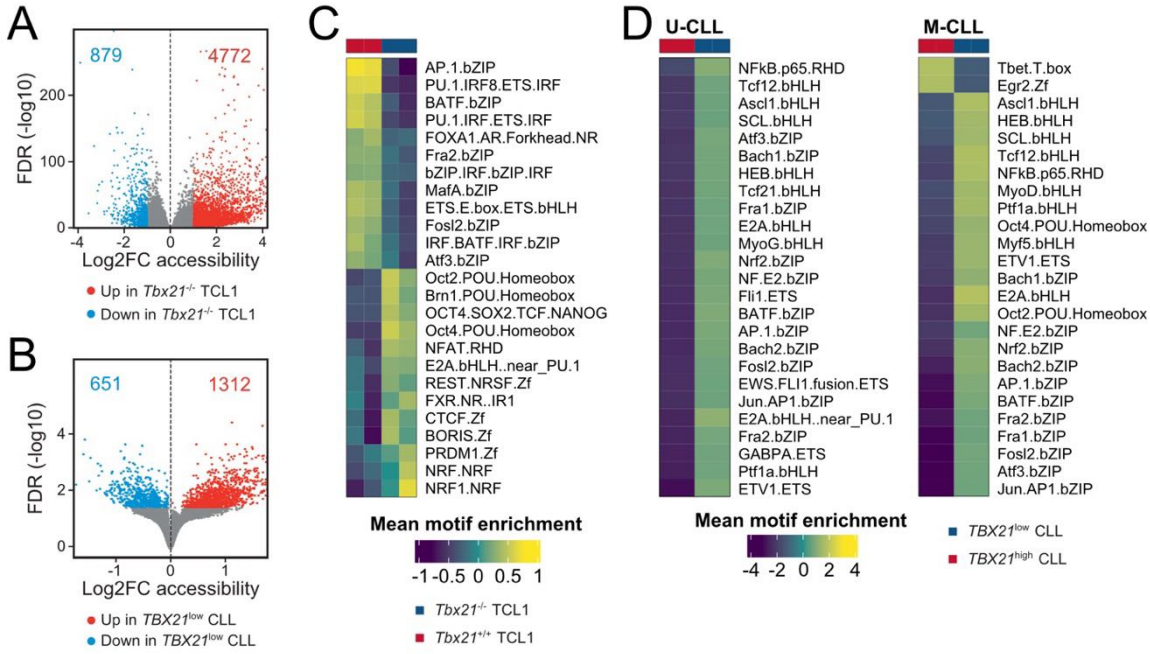


Figure 2.50 Genome-wide chromatin accessibility response to *Tbx21* knock-out in TCL1 cells from scTurboATAC-seq data. A Differential chromatin accessibility in *Tbx21*$^{-/-}$ versus *Tbx21*$^{+/+}$ TCL1 cells from scTurboATAC-seq data. Data from 2 replicates per condition are shown. The numbers of significantly differential ATAC peaks with FDR below 0.05 and absolute log2FC above 1 are indicated. **B** Same as panel A for *TBX21*^{low} versus *TBX21*^{high} CLL cells from bulk ATAC-seq data. **C** TF binding motif enrichment in differentially accessible ATAC peaks between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells from panel A. The top 25 enriched motifs are shown. **D** Mean motif deviation scores in accessible ATAC peaks between *TBX21*^{low} versus *TBX21*^{high} CLL cells with unmutated (U-CLL) and mutated (M-CLL) *IGHV* genes. Analysis of bulk ATAC-seq of CLL cells was performed by Vincente Chapaprieta. Adapted from Roessner *et al.* (2024).

To further investigate the effect of T-bet on these differential ATAC peaks, I analyzed the enrichment of TF binding motifs in the differentially accessible ATAC peaks from *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. This enrichment analysis revealed a consistent enrichment of multiple TF binding motifs across replicates (**Figure 2.50C**). The binding motifs of five TFs, associated with enhanced cellular proliferation, such as E2A and members of the PIT-

OCT-UNC (POU) family, were enriched in *Tbx21*$^{-/-}$ TCL1 cells. This observation further supports T-bet's role in inhibiting cell cycle progression and proliferation. Similarly, the enrichment analysis of TF binding motifs in the differentially accessible ATAC peaks from *TBX21*^{high} and *TBX21*^{low} CLL cells demonstrated an almost exclusive enrichment of TF binding motifs in *TBX21*^{low} CLL cells (**Figure 2.50D**). In this analysis, CLL cells from patients with mutated *IGHV* genes (M-CLL) and unmutated *IGHV* genes (U-CLL) were investigated separately. Notably, only two TFs showed an enrichment of binding motifs in *TBX21*^{high} M-CLL cells, while no TF binding motifs were enriched in *TBX21*^{high} U-CLL cells. Overall, these results suggest that higher T-bet levels decrease chromatin accessibility at specific genomic sites, predominantly reducing the accessibility of TF binding motifs associated with cell cycle progression and proliferation.

T-bet regulates transcription by orchestrating long-range chromatin interactions

In the previous section, I examined the effects of T-bet on chromatin accessibility and the chromatin binding of other TFs. Next, I aimed to investigate the gene regulatory role of T-bet and its direct effects on transcription. First, we identified differentially expressed genes from bulk RNA-seq data of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells, as well as *TBX21*^{low} and *TBX21*^{high} CLL cells. Additionally, we determined the correlation of T-bet protein levels with the levels of other proteins using MS data from the same samples. The comparison of the T-bet dependent gene and protein data revealed a higher number of differentially expressed genes than correlated proteins in TCL1 *Tbx21*$^{-/-}$ vs. *Tbx21*$^{+/+}$ cells and CLL *TBX21*^{low} vs. *TBX21*^{high} cells, respectively (**Figure 2.51A**). Across all datasets, 104 genes/proteins were commonly differential. Consequently, they were considered T-bet dependent with high confidence. This high-confidence set of 104 T-bet associated differential genes/proteins will be referred to as *differential genes* for simplicity. Among the *differential genes*, approximately 50 % exhibited upregulation or downregulation, respectively, when T-bet was present (**Figure 2.51B**). Notably, the expression levels of these *differential genes* distinguished CLL cells from all HC B cell subtypes (**Figure 2.51C**, annotation on top). While some HC B cell subtypes, such as CD5^{+} B cells or IgM^{+}/IgD^{+}/CD27^{+} B cells, exhibited partially similar expression levels to CLL cells, other HC B cell subtypes, such as IgM-only B cells, showed opposite expression levels for all *differential genes*. Overall, this high-confidence set of *differential genes* was not only associated with T-bet but also effectively differentiated malignant B cells from healthy B cells. Thus, these *differential genes* serve as suitable examples for studying T-bet dependent mechanisms of transcription regulation.

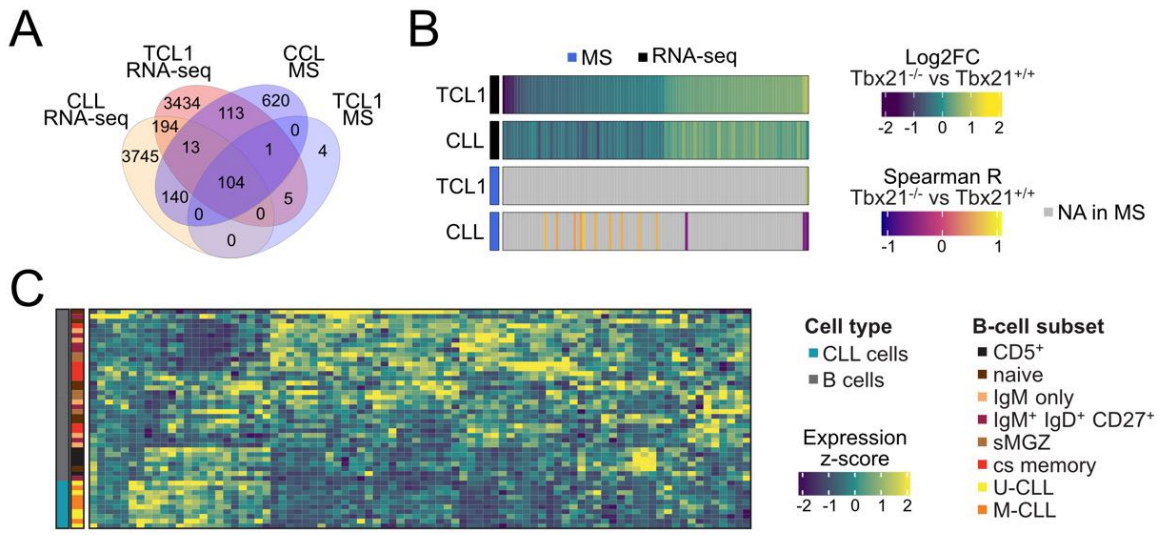


Figure 2.51 Expression and protein levels of T-bet dependent genes from bulk RNA-seq and MS data of TCL1 cells and CLL patient samples. A Overlap of significantly differential genes and correlated proteins between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells, as well as *TBX21*^{high} and *TBX21*^{low} CLL cells. Gene expression was measured by bulk RNA-seq, and protein levels by MS. **B** Differential gene expression and T-bet correlated protein levels in *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells, as well as *TBX21*^{high} and *TBX21*^{low} CLL cells for 104 overlapping *differential genes*. **C** Gene expression levels of 104 overlapping *differential genes* in HC B cells and CLL cells. Experiments and analysis were performed by Philipp Roessner and colleagues. Adapted from Roessner *et al.* (2024).

Following, I analyzed T-bet dependent transcription regulation of the *differential genes* by performing *single cell co-accessibility* analysis of the scTurboATAC-seq data from *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. To investigate T-bet specific regulation, I predicted potential T-bet binding sites within the 133,668 ATAC peaks using the T-bet specific binding motif. This analysis identified 23 % of ATAC peaks as containing at least one potential T-bet binding site, following referred to as *T-bet peaks*. When assessing the genomic positions of the *T-bet peaks* relative to the previously identified *differential genes*, only 25 % of the *differential genes* had a *T-bet peak* at their promoter (**Figure 2.52A**, blue). However, most of the *differential genes* (60 %) demonstrated an AC between their promoter and a distal *T-bet peak* in the *single cell co-accessibility* analysis (**Figure 2.52A**, yellow), indicating that T-bet predominantly regulates transcription via long-range chromatin contacts. The remaining 15 % of *differential genes* showed no link to a *T-bet peak* (**Figure 2.52A**, grey) and might be regulated, as so-called secondary targets of T-bet, by other TFs among the *differential genes*. Interestingly, *differential genes* with a promoter *T-bet peak* exhibited significantly more promoter ACs upon *Tbx21* knock-out (**Figure 2.52B**, right). In contrast, I observed no significant difference in total AC numbers for *differential genes* without a promoter *T-bet peak* (**Figure 2.52B**, left). Overall, *Tbx21*$^{-/-}$ TCL1 cells showed more ACs between *differential genes* and *T-bet peaks* compared to *Tbx21*$^{+/+}$ TCL1 cells (170 and 151, respectively) (**Figure 2.52C**). Notably, only 15 % of

these ACs were detected in both *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells, indicating a rewiring of ACs upon *Tbx21* knock-out. In summary, these findings suggest that T-bet primarily regulates transcription through distal chromatin interactions and that its binding to chromatin may inhibit the formation of long-range chromatin contacts.

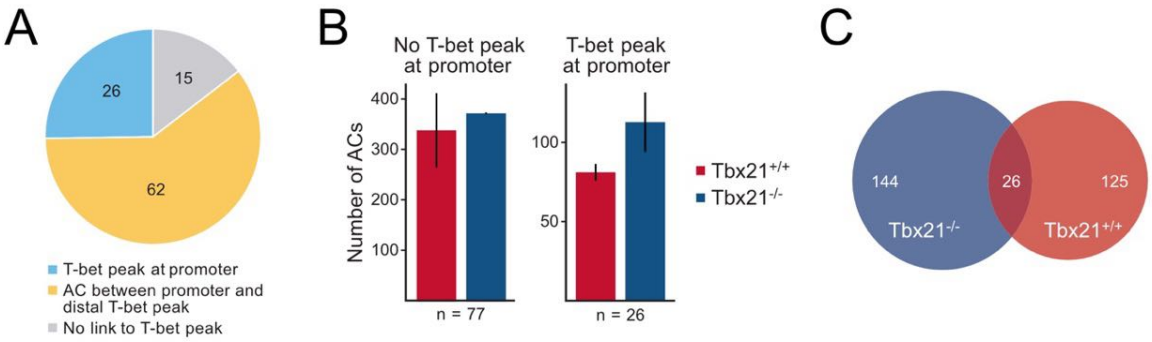


Figure 2.52 *Single cell co-accessibility* analysis in scTurboATAC-seq data of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. A T-bet dependent regulation of 104 *differential genes*. Genes are categorized successively based on the presence of a *T-bet peak* at the promoter (blue), an AC between the promoter and a distal *T-bet peak* (yellow), and no link to a *T-bet peak* (grey). **B** Number of ACs at *differential gene* promoters without (left) and with (right) a *T-bet peak*. ACs within a 1 Mb window are shown. Whiskers represent the standard error of 2 biological replicates. **C** Overlap of ACs from *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells. Merged ACs from biological replicates between *differential gene* promoters and *T-bet peaks* in a 100 kb window are shown. Adapted from Roessner *et al.* (2024).

One example of T-bet dependent transcription regulation is the *differential gene Nos1*. *Nos1* exhibited significantly higher expression levels in bulk RNA-seq data of *Tbx21*$^{-/-}$ compared to *Tbx21*$^{+/+}$ TCL1 cells (log2FC of 0.26, adjusted p-value of 0.0095; **Figure 2.53A)**. The genomic annotation of *Nos1* indicated three TSSs (**Figure 2.53B**, red annotation), each showing different regulatory mechanisms of T-bet dependent transcription repression, as described in the following. TSS1 was inaccessible in both *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells, lacked potential T-bet binding sites, and did not display any ACs. Therefore, it was considered inactive in transcription. TSS2 was moderately accessible in both *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells (**Figure 2.53B**, top). Although there were no potential T-bet binding sites proximal to TSS2, its accessibility was significantly reduced upon *Tbx21* knock-out. The lower accessibility coincided with fewer ACs between TSS2 and the surrounding, distal *T-bet peaks* (**Figure 2.53B**, bottom). Consequently, the higher accessibility and greater number of ACs suggest that T-bet negatively regulates transcription of *Nos1* at TSS2 via repressive chromatin interactions with distal regulatory *T-bet peaks*. Finally, TSS3 contained a potential T-bet binding site and displayed significantly higher accessibility in *Tbx21*$^{-/-}$ compared to *Tbx21*$^{+/+}$ TCL1 cells (**Figure 2.53B**, top). Upon *Tbx21* knock-out, TSS3 additionally showed an AC with a downstream

ATAC peak (**Figure 2.53B**, bottom), suggesting that the loss of repressive T-bet binding at TSS3 permitted the formation of activating chromatin contacts.

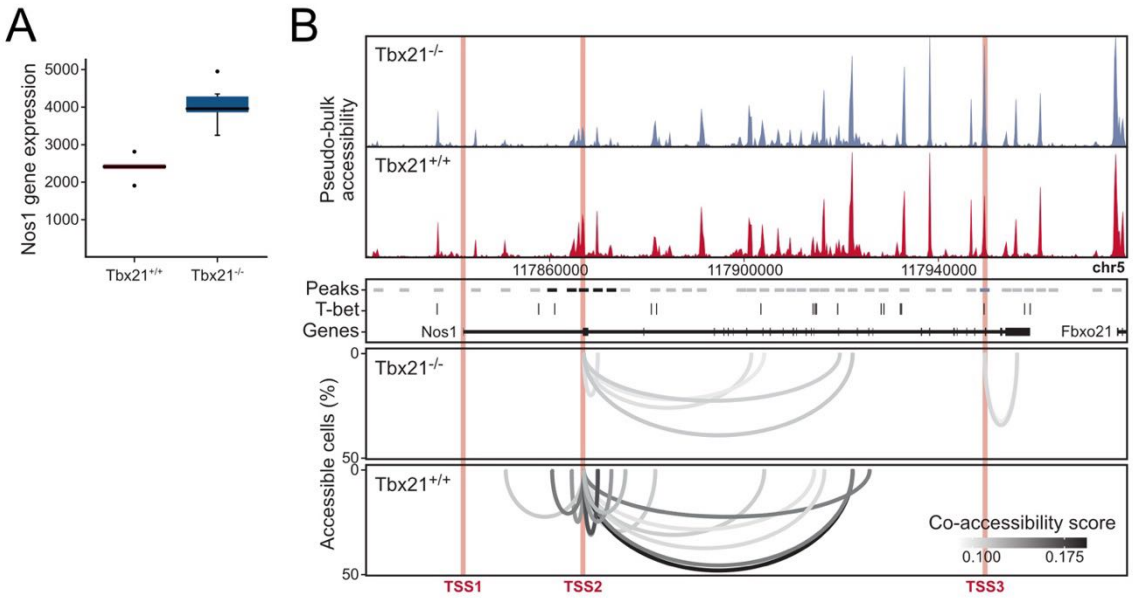


Figure 2.53 T-bet dependent transcription regulation of the *differential gene Nos1* from *single cell co-accessibility* analysis of scTurboATAC-seq data. A Gene expression from 6 biological replicates of bulk RNA-seq data per condition. **B** Pseudo-bulk accessibility signal (top), genomic annotation (middle), and ACs (bottom) for the *differential gene Nos1*. 2 kb regions around ATAC peaks (grey), more accessible ATAC peaks in Tbx21$^{+/+}$ (black), more accessible ATAC peaks in Tbx21$^{-/-}$ (blue), potential T-bet binding sites, and gene annotation are shown. 1 kb around TSSs of *Nos1* are marked in red. Pseudo-bulk accessibility and ACs at *Nos1* TSSs from the biological replicates were merged. Bulk RNA-seq of TCL1 cells was performed by Philipp Roessner and analyzed by Marc Zapatka. Adapted from Roessner *et al.* (2024).

Secondary targets of T-bet regulate chromatin organization

Subsequently, I conducted a *metacell co-accessibility* analysis for *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells to investigate whether T-bet knock-out induces more global, cell state-driven patterns in co-accessibility. Visual inspection of the *metacell co-accessibility* maps at *differential genes* revealed multiple domains with locally increased co-accessibility. One example, showing a clearly defined domain of enriched co-accessibility, is the *differential gene Gimap6* (**Figure 2.54A**). The domain contained several *Gimap* genes, including *Gimap3*, *Gimap4*, *Gimap5*, and *Gimap6*. However, only *Gimap6* was differentially expressed upon *Tbx21* knock-out, showing a significantly reduced expression in bulk RNA-seq data (log2FC of -0.2 for *Tbx21*$^{+/+}$ versus *Tbx21*$^{-/-}$, adjusted p-value of 0.033). Interestingly, the lower expression under T-bet absence and thus the higher expression under T-bet presence contrasted with previous findings regarding the repressive and silencing role of T-bet. Further investigation revealed that the reduction in expression corresponded with a decrease in chromatin accessibility within the domain of enriched co-accessibility in *Tbx21*$^{-/-}$ TCL1 cells, where 3 ATAC peaks exhibited significantly reduced

chromatin accessibility. Consequently, the domain of enriched co-accessibility appeared to be driven by higher activity in *Tbx21*$^{+/+}$ TCL1 cells. Notably, the *Gimap6* promoter did not contain a potential T-bet binding site and no ACs between the promoter and distal *T-bet peaks* were detected. This suggests that the domain formation and expression induction in *Tbx21*$^{+/+}$ TCL1 cells are not directly mediated by T-bet, but rather its upregulated target genes (see **Figure 2.50C**).

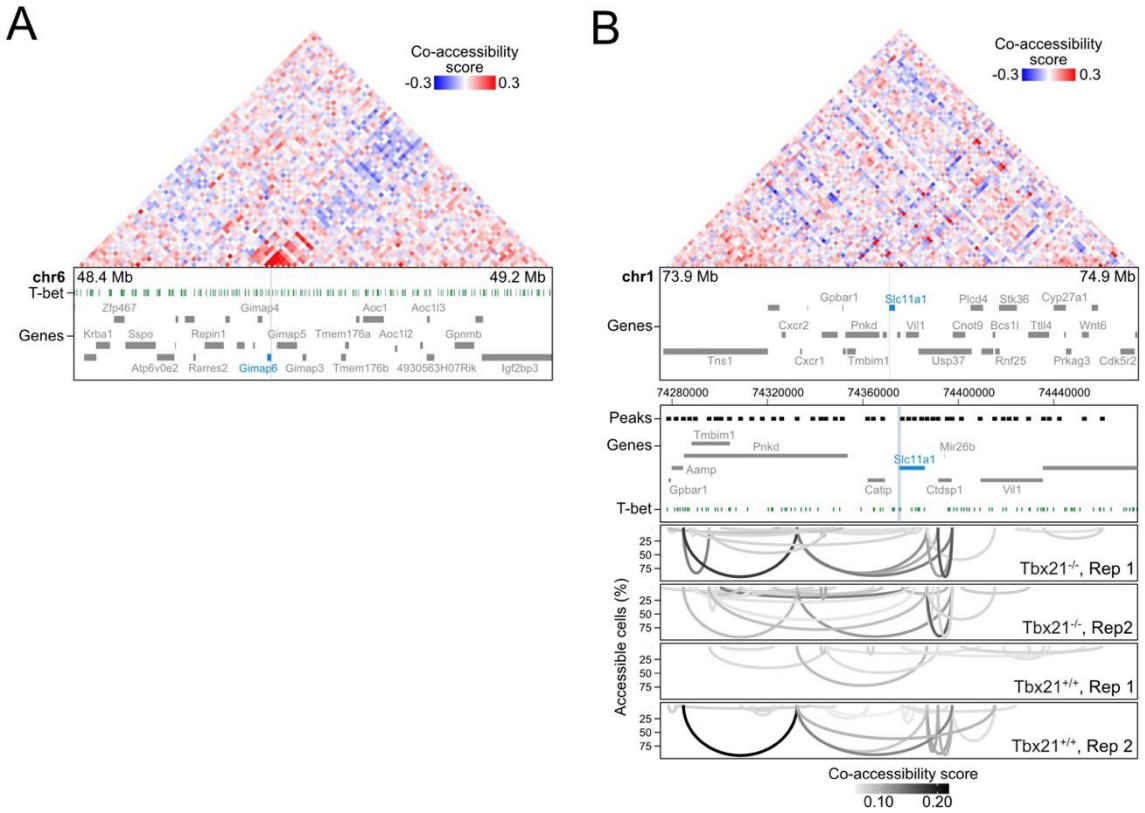


Figure 2.54 *Metacell co-accessibility* analysis at *differential genes* between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells from scTurboATAC-seq data. A *Metacell co-accessibility* map (top) and genomic annotation (bottom) for the *differential gene Gimap6*. Potential T-bet binding sites (green), gene annotations (grey), *differential genes* (blue) and 1 kb regions around *differential gene* TSSs (light blue) are shown. The color scale bar limits of *metacell co-accessibility* scores are set to -0.3 and 0.3. **B** *Metacell co-accessibility* map (top), genomic annotation (middle), and ACs for the *differential gene Slc1a1*. ATAC peaks (black), potential T-bet binding sites (green), gene annotations (grey), *differential genes* (blue) and 1 kb regions around *differential gene* TSSs (light blue) are shown. The color scale bar limits of *metacell co-accessibility* scores are set to -0.3 and 0.3. A zoom-in view of *Slc11a1* ACs is shown. Adapted from Seufert *et al.* (2024).

Interestingly, the *metacell co-accessibility* maps from the scTurboATAC-seq data of *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells revealed the same *blue stripes* with high co-accessibility scores at their intersections as observed in the HUVEC data previously (see **Section 2.2.3**). One example, where these *blue stripes* extended across an entire 1 Mb region, was the *differential gene Slc11a1* (**Figure 2.54B**, top). *Slc11a1* was located between the *blue stripes*, and its expression was significantly reduced in *Tbx21*$^{-/-}$ TCL1 cells (log2FC

of -0.3 for *Tbx21*$^{+/+}$ versus *Tbx21*$^{-/-}$, adjusted p-value of 0.031). Similar to HUVECs, the *blue stripes* coincided with *frequent ACs*, showing PAC values above 90 (**Figure 2.54B**, bottom). These presumably architectural ACs were present in both *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ conditions but were more reproducible and had higher co-accessibility scores for *Tbx21*$^{-/-}$ TCL1 samples. This observation aligned with the previously identified enrichment of CTCF and BORIS binding motifs in accessible ATAC peaks of *Tbx21*$^{-/-}$ TCL1 cells (see **Figure 2.50C**), which both are key factors in chromatin loop formation. The presence of these structural ACs might therefore alter the chromatin organization of the region, leading to reduced transcription of the *differential gene Slc11a1*.

In summary, in this project I studied transcriptional regulation in the mouse TCL1 cancer model for CLL to determine whether the previously identified principles of transcription regulation via promoters or AC- and DC-mediated *distal CREs* hold true beyond the transcriptional response to cytokine stimulation. CLL cells exhibit higher expression and protein levels of T-bet compared to healthy B cells. These elevated levels are induced by inflammatory signals, such as cytokine stimulation or interaction with activated T cells. Along with the longer overall survival observed for CLL patients with high *TBX21* expression, this suggests a tumor-suppressive role for T-bet. This role is primarily driven by T-bet's capacity to repress cell cycle progression and cellular proliferation in malignant B cells. Specifically, the TF T-bet exerts its function by decreasing chromatin accessibility at specific genomic sites, mainly reducing the accessibility of TF binding motifs associated with cell cycle progression and proliferation. Interestingly, T-bet itself primarily regulates transcription through distal chromatin interactions, rather than direct binding to gene promoters. In this context, its binding to chromatin might repress the formation of long-range chromatin contacts, orchestrating transcription regulation between distal regulatory sites. Additionally, T-bet represses the formation of architectural chromatin loops or contacts, likely by indirectly reducing the accessibility of specific sites with CTCF and BORIS binding motifs. Secondary TFs induced by T-bet potentially contribute to broader domains of enriched co-accessibility, which in turn increased expression of *differential genes* within. Overall, this suggests that T-bet itself functions as a transcriptional silencer and represses chromatin contacts and loop formation, while its upregulated target genes might also activate transcription, either through direct DNA binding or by promoting domains of locally increased TF activity.

2.3.3. Molecular mechanisms of TNFα-induced transcriptional co-regulation in human endothelial cells

In the next project, I investigated the transcriptional consequences of the previously identified AC and DC co-accessibility features in HUVECs (see **Chapter 2.2**). Specifically, I aimed to explore their potential regulatory roles in modulating various transcriptional bursting parameters, using both sequencing data and fluorescence microscopy of nuclear RNA (**Figure 2.55**). Additionally, I examined how ACs and DCs, both long-range regulatory mechanisms, might co-regulate the transcription of multiple genes. By analyzing the high-quality scTurboATAC-seq data at different treatment time points, I was able to distinguish AC and DC regulatory effects on time and intensity of the transcriptional response to TNFα.

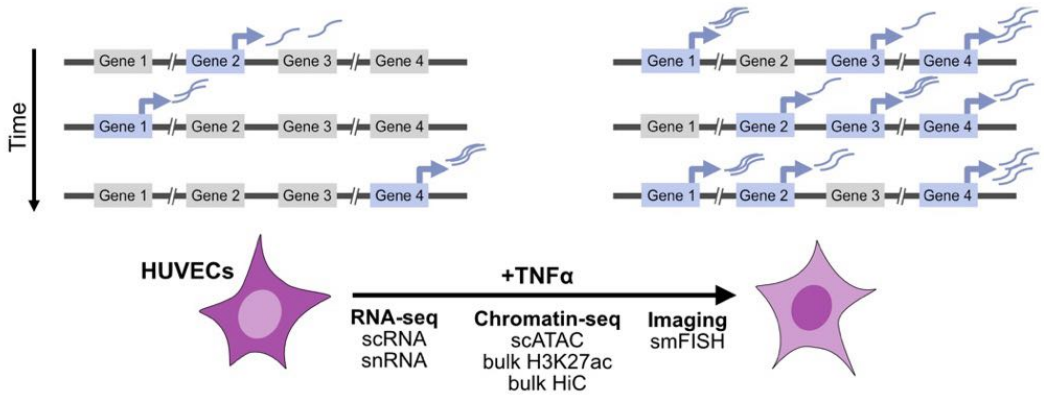


Figure 2.55 Studying TNFα-induced transcription co-regulation in HUVECs. Data from bulk and single-cell sequencing of cellular and nuclear RNA, chromatin accessibility, H3K27ac modifications, and chromatin contacts were analyzed. Additionally, fluorescence microscopy of nuclear RNA was performed. Adapted from Seufert *et al.* (2024).

TNFα induces differential expression of approximately 1,500 genes

scRNA-seq was performed on three biological replicates of untreated and TNFα-treated HUVECs to identify robust transcriptional changes induced by TNFα treatment. The transcriptomic profiles of single cells showed strong differences between treatment time points (**Figure 2.56A**) and cell cycle states (**Figure 2.56C**). In contrast, biological replicates revealed only minimal variation in the low-dimensional embedding (**Figure 2.56B**). For further analysis, only cells in the G1 cell cycle state were selected (**Figure 2.56D**), allowing to specifically resolve TNFα-dependent transcriptional differences. Differential gene expression analysis of G1 cells from untreated versus TNFα-treated conditions revealed differential expression of 1,499 genes in total (**Figure 2.56E**), termed TNFα-regulated genes (TRGs). After 30 min of TNFα treatment, 386 TRGs were identified, with the majority (94 %) showing upregulated expression. In contrast, after 240 min of TNFα treatment, 1,280 TRGs were identified, but only 56 % of them exhibited upregulation,

indicating a more diversified transcriptional response at this later time point. Notably, in the early response to TNFα, only 26 % of TRGs were detected, whereas in the late response, 85 % of TRGs were detected. Comparing all TRGs between the different treatment time points revealed that only 11 % were differentially expressed at both time points, with almost all of these showing upregulated expression in response to TNFα (**Figure 2.56F**, time point annotation on top). Additionally, while the majority of TRGs were protein-coding, about 30 % were lncRNAs, which in turn might play a role in regulating the

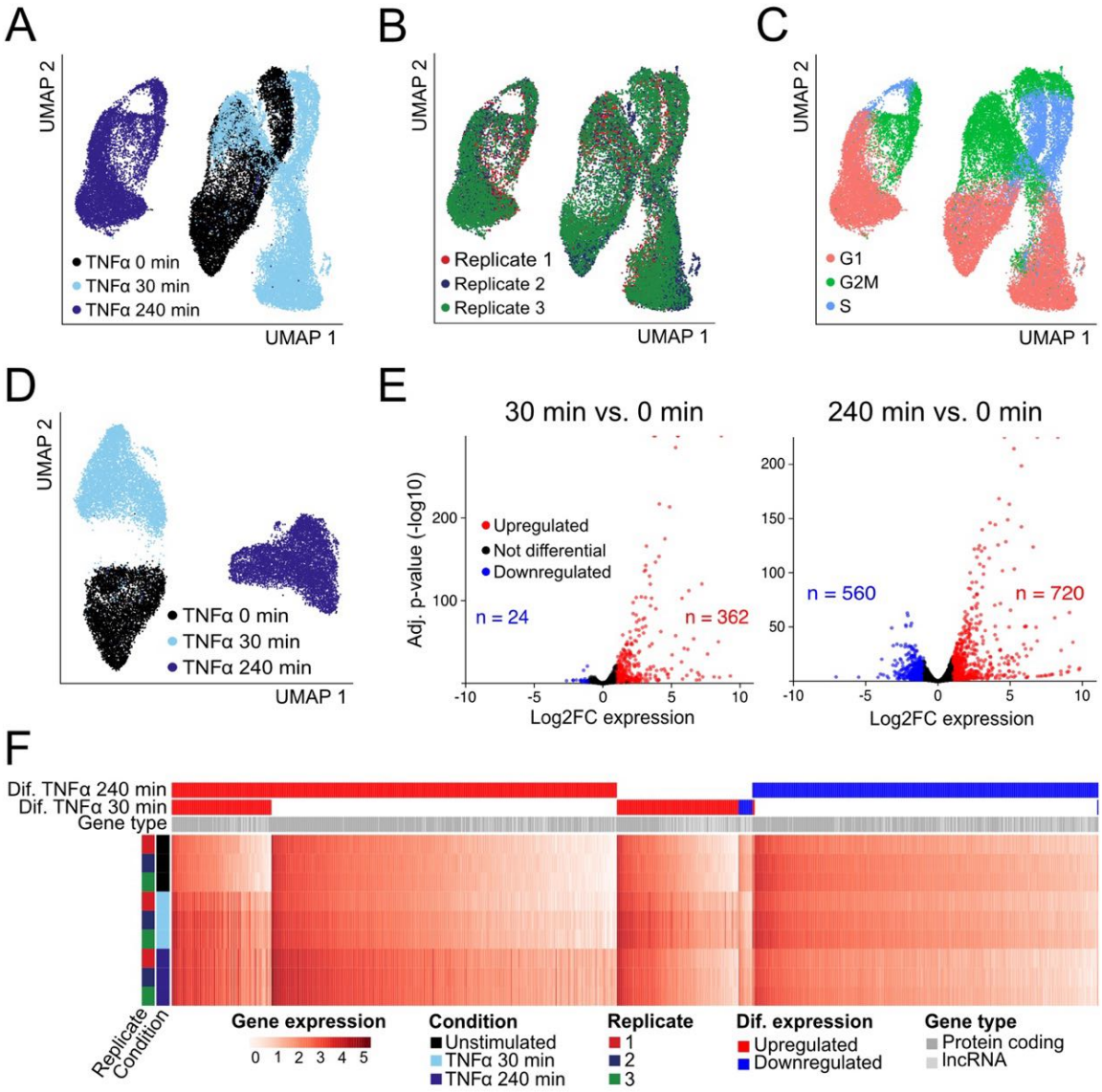


Figure 2.56 Transcriptomic profiles of untreated and TNFα-treated HUVECs from scRNA-seq data. Data from 3 biological replicates per condition are shown. **A** Low-dimensional embedding of HUVECs, colored by treatment time point. **B** Same as panel A, colored by biological replicate. **C** Same as panel A, colored by cell cycle state. **D** Same as panel A for HUVECs only in the G1 cell cycle state. **E** Differential gene expression between untreated and TNFα-treated HUVECs at 30 min and 240 min. Significantly upregulated genes (log2FC ≥ 1; adjusted p-value < 0.05) are shown in red and significantly downregulated genes (log2FC ≤ -1, adjusted p-value < 0.05) are shown in blue. **F** Expression (UMI counts, log10 transformed and scaled) of TNFα-regulated genes (TRGs). Gene type and direction of differential regulation after 30 min and 240 min of TNFα treatment are annotated. Adapted from Seufert *et al.* (2024).

transcriptional response. Overall, the transcriptomic response of HUVECs to TNFα treatment was strong at both time points, showing both sustained and time point-specific transcriptional changes.

TNFα treatment increases the accessibility of specific TF binding sites

To investigate the regulatory drivers and their impact on chromatin accessibility during the transcriptional response to TNFα treatment, I analyzed single-cell chromatin accessibility profiles from scTurboATAC-seq data of the same three biological replicates from untreated and TNFα-treated HUVECs. Similar to the scRNA-seq data, the chromatin accessibility profiles of single cells exhibited strong differences between treatment time points and cell cycle states (see **Figure 2.21A-C**). As a result, only G1 cells were selected for further analysis (**Figure 2.57A**), which showed a clear separation of treatment time points across all replicates in low-dimensional embedding. Pseudo-bulk analysis of these G1 cells identified 201,329 ATAC peaks, with 9 % located at promoters, 61 % in gene bodies, and 30 % in intergenic regions (**Figure 2.57B**). Subsequent analysis of chromatin accessibility levels at these ATAC peaks revealed that only 2 % and 3 % were differentially accessible after 30 min and 240 min of TNFα treatment, respectively (**Figure 2.57C**). The vast majority of these differential ATAC peaks (94-97 %) gained accessibility upon TNFα treatment. Notably, differential ATAC peaks were less frequently located at promoters compared to all ATAC peaks (2 % versus 9 %; **Figures 2.57B+D**). Most differential ATAC peaks were found in intronic or intergenic regions (91-93 %), indicating that TNFα primarily regulates transcription via regulatory elements at intronic or intergenic sites.

Next, I calculated TF binding scores from accessibility footprints of accessible TF binding sites, which were predicted genome-wide from annotated TF binding motifs (see **Figure 2.33**). Differential TF binding analysis revealed increased binding of several TFs after TNFα treatment (**Figure 2.57E**). Notably, members of the NF-κB family showed a strong increase in binding scores at both time points (**Figure 2.57E**). In contrast, binding scores of PRDM1, activating transcription factor 4 (ATF4), and members of the IRF and CEBP families were enhanced only after 240 min of TNFα treatment (**Figure 2.57E**, dark blue). This highlights the fast and sustained activation of known TNFα-regulated NF-κB complexes, while the other TFs likely reflect a secondary and indirect response to TNFα treatment. In summary, TNFα treatment increased chromatin accessibility at specific, primarily non-promoter sites, suggesting the activation of individual promoter-distal regulatory elements rather than a broad reorganization of chromatin conformation. These changes were likely mediated by the increased binding activity of known TNFα-responsive TFs.

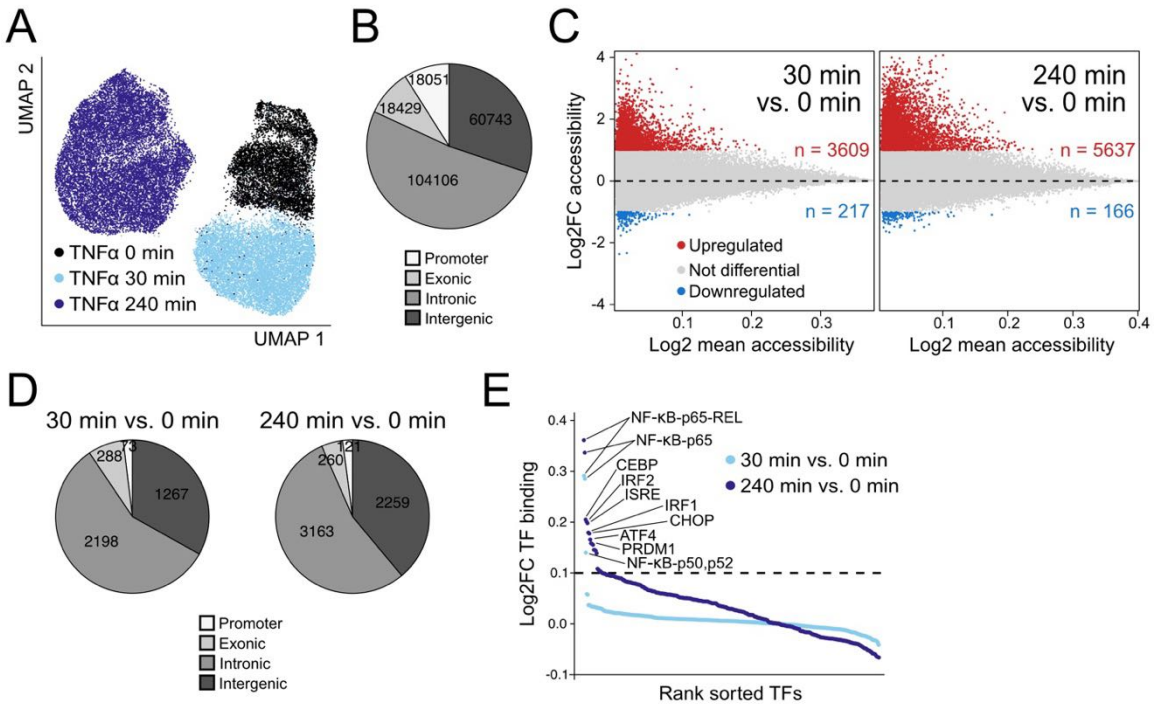


Figure 2.57 Chromatin accessibility profiles of untreated and TNFα-treated HUVECs from scTurboATAC-seq data. Data from 3 biological replicates per condition are shown. **A** Low-dimensional embedding of HUVECs in the G1 cell cycle state, colored by treatment time point. **B** Genomic location of pseudo-bulk ATAC peaks. **C** Differential chromatin accessibility between untreated and TNFα-treated HUVECs at 30 min and 240 min. Significantly upregulated ATAC peaks (log2FC ≥ 1; adjusted p-value < 0.05) are shown in red and significantly downregulated ATAC peaks (log2FC ≤ -1, adjusted p-value < 0.05) are shown in blue. **D** Genomic location of differential ATAC peaks after 30 min (left) and 240 min (right) of TNFα treatment. **E** Differential TF binding between untreated and TNFα-treated HUVECs at 30 min and 240 min. The dashed line represents the cutoff for upregulated TFs (log2FC ≥ 0.1). Top 10 differential TFs are annotated. Adapted from Seufert *et al.* (2024).

TRGs cluster in the genome

To better understand how the well-defined TRGs are collectively regulated by NF-κB and its secondary TFs, I investigated their genomic position in relation to each other, as well as with respect to ATAC peaks and TADs. Interestingly, 91 % of all TRGs showed an ATAC peak at their promoters (**Figure 2.58A**, inner circle). However, only 9 % of these were differentially accessible upon TNFα treatment (**Figure 2.58A**, outer circle). This suggests that although most TRG promoters are active, differential transcription regulation following TNFα treatment is not driven by promoter accessibility itself. Therefore, as other, non-proximal regulatory mechanisms are likely responsible for the TNFα-regulated transcriptional changes, I further examined the genomic location of TRGs relative to each other. In this context, the majority of TRGs (1,008; 67 %) were located within 500 kb of another TRG, while the remaining 33 % of TRGs (491) were isolated in the genome (**Figure 2.58B**, edges between TRG data points). If at least two TRGs were located within 500 kb, they were considered proximal to each other and defined as a *TRG cluster*. These

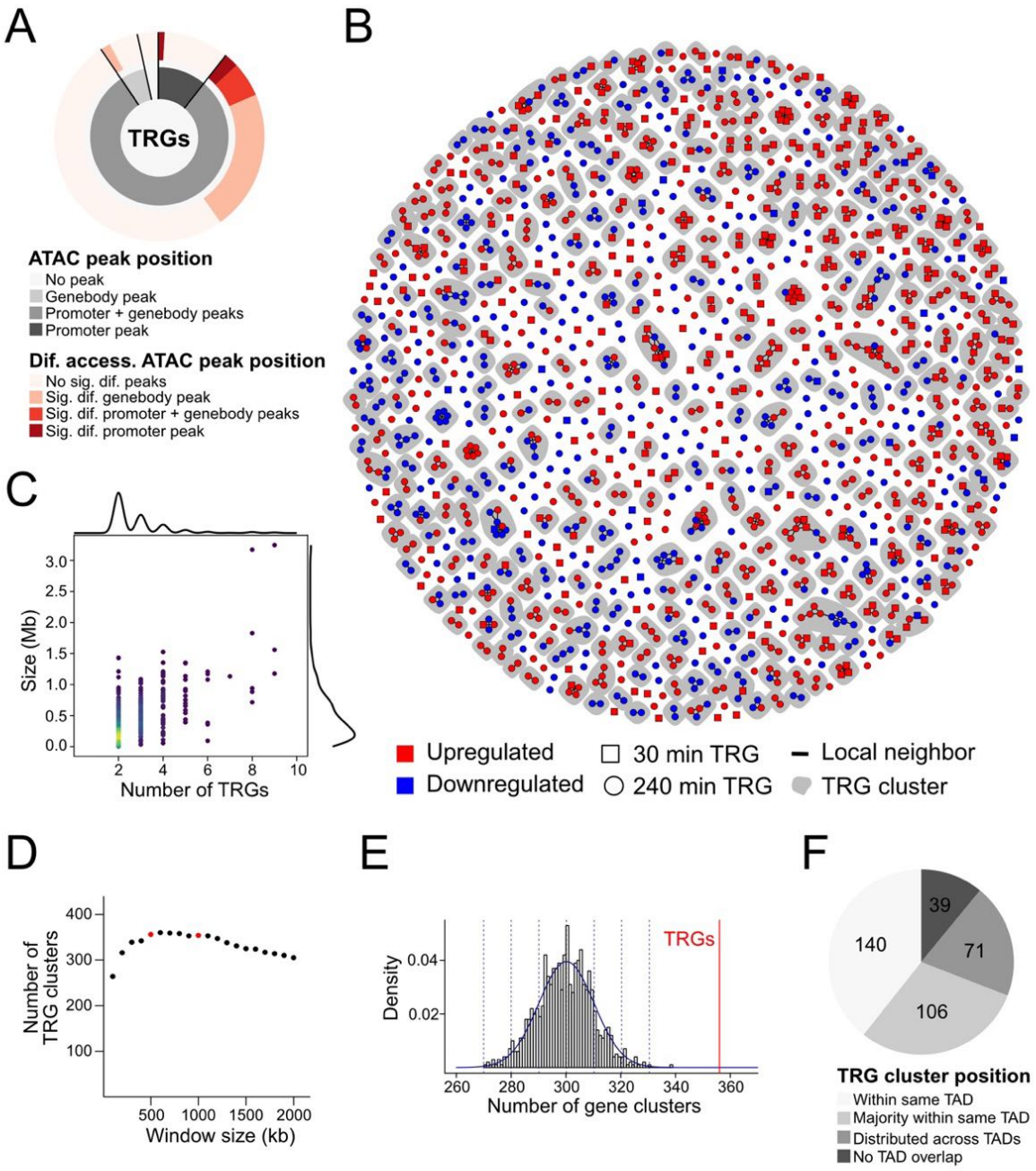


Figure 2.58 Genomic location of TNFα-regulated genes (TRGs) from scRNA-seq data. **A** ATAC peaks at TRGs. The inner circle illustrates the position of all ATAC peaks located either at the promoter or gene body of a TRG. The outer circle highlights the differential ATAC peaks among them. **B** Genomic proximity of TRGs. Each data point represents one TRG, with shape indicating the time point of differential expression and color denoting the direction of regulation. For TRGs differentially expressed at both time points, the 30 min results are annotated. Edges are drawn between TRG neighbors below 500 kb distance. *TRG clusters* are highlighted with a grey background. **C** *TRG cluster* size over TRG number per *TRG cluster*. The colors of points reflect the density of *TRG clusters*. Density curves of *TRG cluster* size and TRG number are shown. **D** Number of *TRG clusters* across different window sizes for defining TRG proximity. 500 kb and 1 Mb are marked in red. **E** Distribution of number of gene clusters detected in 1,499 randomly selected genes for 1,000 times. A window size of 500 kb for TRG proximity was used. The number of *TRG clusters* is marked in red. **F** *TRG clusters* in relation to TADs. TADs from bulk HiC-seq data of untreated HUVECs at 25 kb resolution were used. *TRG clusters* are classified as all TRGs within the same TAD, the majority of TRGs within the same TAD, TRGs distributed across multiple TADs, and no TAD overlap of TRGs. Adapted from Seufert *et al.* (2024).

1,008 proximal TRGs formed 356 *TRG clusters* (**Figure 2.58B**, grey shapes), which varied in the number of TRGs, upregulation or downregulation of transcription upon TNFα treatment, and their degree of proximity (ranging from “pearls on a string” to “fully interlinked” proximal TRGs). Most *TRG clusters* contained only two TRGs, but some *TRG clusters* contained as many as nine TRGs (**Figure 2.58C**, TRG number on the x-axis). The sizes of *TRG clusters* ranged from 100 kb to 3 Mb, with more than 90 % of *TRG clusters* being smaller than 1 Mb (**Figure 2.58C**, *TRG cluster* size on the y-axis).

To identify these *TRG clusters*, I tested different window sizes for defining TRG proximity (**Figure 2.58D**). For window sizes below 500 kb, the number of identified *TRG clusters* increased as the window size grew, as more TRGs were considered proximal. The number of *TRG clusters* plateaued for window sizes between 500 kb and 1 Mb. For window sizes above 1 Mb, the number of *TRG cluster* decreased, likely due to the merging of previously distinct *TRG clusters*. I selected a window size of 500 kb as it represented the inflection point where the number of robustly detected *TRG clusters* stabilized. Next, I compared these *TRG clusters* with the genomic clustering of random selected genes (**Figure 2.58E**) to determine whether the observed *TRG clusters* were a significant feature of TNFα-regulated transcription, or if it simply reflected general gene clustering patterns in the genome. For this, I randomly sampled 1,499 genes (matching the number of TRGs) 1,000 times and calculated the numbers of gene clusters using a 500 kb window for gene proximity. The distribution revealed a mean of 300 gene clusters, with a maximum of 348 gene clusters (**Figure 2.58E**). The 356 *TRG clusters* exceeded the third standard deviation from the mean of random gene clusters, suggesting that the clustering of TRGs is not random but may be important for regulating the TNFα-induced transcriptional response. Finally, I explored the relationship between *TRG clusters* and TADs. For 69 % of *TRG clusters*, all or the majority of TRGs were located within the same TAD (**Figure 2.58F**). Only 20 % of *TRG clusters* had TRGs distributed across multiple TADs.

Clustered TRGs are co-expressed in the same cells

To further investigate whether *TRG clusters* are functionally relevant in TNFα-regulated transcription, I analyzed the co-expression of TRGs in the same cells. Co-expression for each pair-wise TRG combination was calculated using Spearman correlation on UMI counts across cells (**Figure 2.59A**). The average Spearman correlation coefficients from replicates were used to evaluate co-expression between TRG pairs. Using this approach, I computed co-expression of both clustered and isolated TRGs for each TNFα treatment time point. For clustered TRGs, only TRG pairs within the same *TRG cluster* were considered. In untreated condition, both clustered and isolated TRGs displayed low levels

of co-expression (**Figure 2.59B**). Upon TNFα treatment, however, co-expression within *TRG clusters* increased, whereas co-expression among isolated TRGs remained unchanged. A direct comparison of upregulated TRG co-expression at 30 min and 240 min of TNFα treatment revealed significantly higher co-expression within clustered TRGs compared to isolated TRGs (**Figure 2.59C**). Notably, the distribution of co-expression within *TRG clusters* exhibited an extended right tail at the 240 min time point and showed a bimodal distribution at 30 min. Overall, these findings suggest that higher co-expression within *TRG clusters* indicate at least partial co-regulation of transcription within these clusters following TNFα treatment.

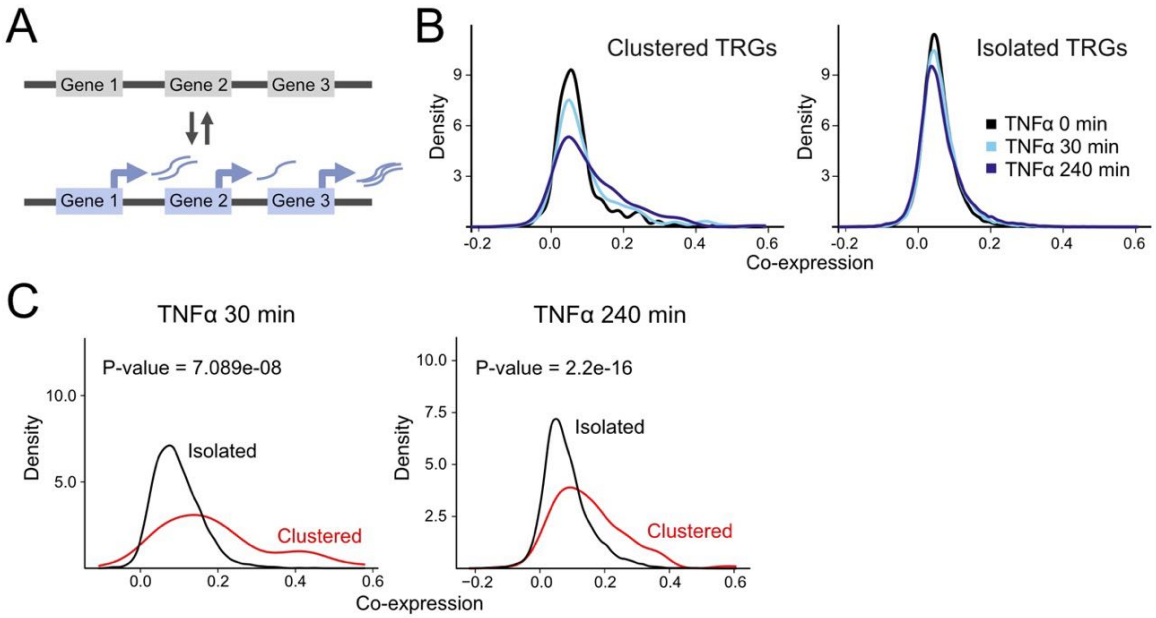


Figure 2.59 Co-expression of clustered and isolated TRGs in untreated and TNFα-treated HUVECs from scRNA-seq data. Data from scRNA-seq of three biological replicates per condition are shown. **A** Co-expression between two TRGs is computed using Spearman correlation of their UMI counts across all cells from one sample. **B** Co-expression of clustered (left) and isolated (right) TRGs at the 0 min, 30 min, and 240 min TNFα treatment time points. Average co-expression of replicates is shown per condition. **C** Co-expression of clustered and isolated TRGs that were upregulated at 30 min (left) and 240 min (right) TNFα treatment time points. Average co-expression of replicates is shown per condition. P-values from two-sided, unpaired Wilcoxon tests are shown. Adapted from Seufert *et al.* (2024).

To validate the observed co-expression from scRNA-seq data, we performed padFISH imaging of nuclear RNA, an imaging protocol for multiplexed smFISH. The padFISH protocol resolves the spatial location of transcription loci, allowing the resolution of allele-specific co-expression. It enables sequential imaging of multiple RNAs but does not provide genome-wide transcriptomic coverage like sequencing techniques. We selected a *TRG cluster* consisting of *CXCL1*, *CXCL2*, *CXCL3*, and *CXCL8* to compare spatially resolved co-expression from padFISH imaging with cell-average co-expression from scRNA-seq data. This so-called *CXCL TRG cluster* spans approximately 350 kb and

includes two additional TRGs, *CXCL5* and *CXCL6*, which were only weakly induced by TNFα. Following TNFα treatment, RNA molecules from all studied *CXCL* genes were detected throughout the nucleus (**Figure 2.60A**, top). Most cells contained zero to two loci with co-localized RNA molecules from at least two genes, appearing white in the microscopy images. These co-localization loci were interpreted as the nuclear positions of *CXCL TRG clusters*, and co-expression at these loci was further analyzed (**Figure 2.60A**, bottom). Different combinations of co-expressed *CXCL* genes were present at the co-localization loci. Quantifying these combinations across 120-290 cells per TNFα treatment time point and replicate revealed a distinct abundance of co-expression patterns (**Figure 2.60B**, dark grey). At both time points, the most frequent co-expression combinations

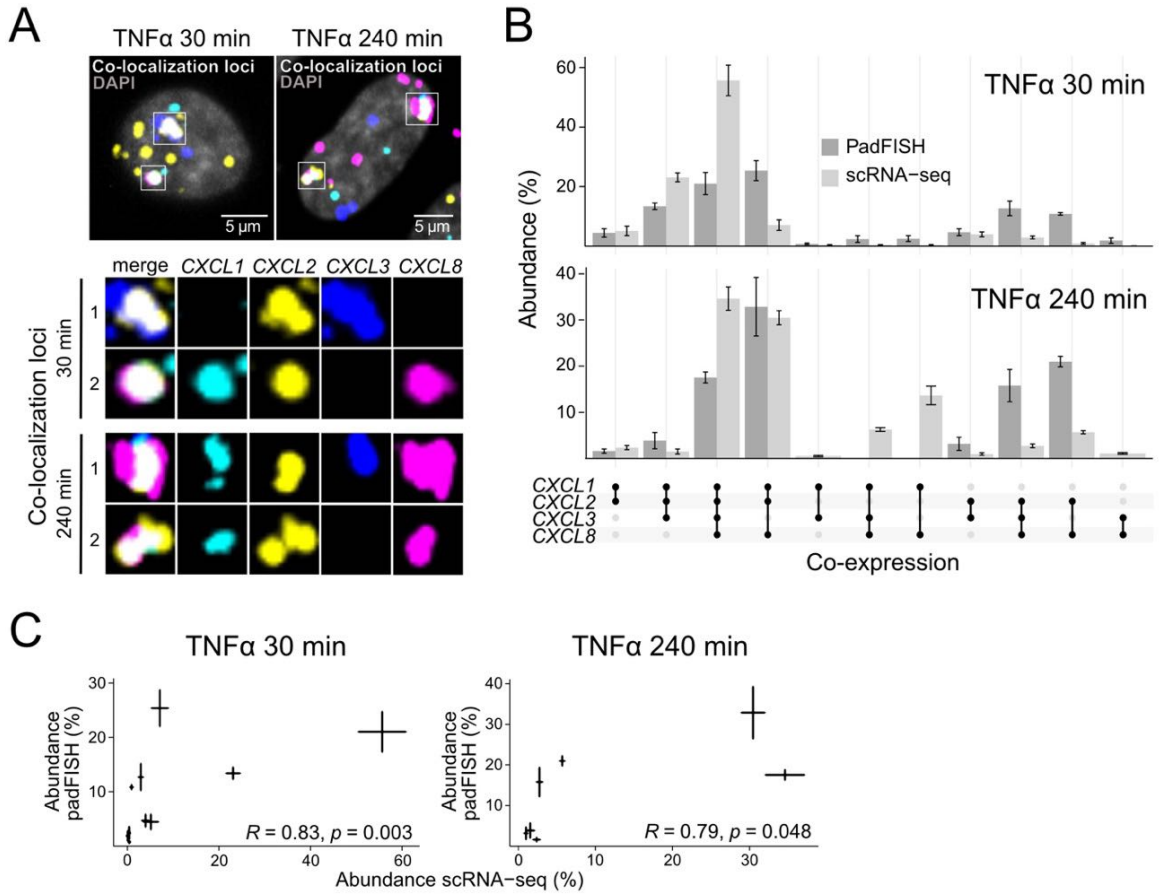


Figure 2.60 Co-expression in exemplary *TRG cluster* with *CXCL1*, *CXCL2*, *CXCL3*, and *CXCL8* in TNFα-treated HUVECs from scRNA-seq and spatial transcriptomics data. Data from padFISH imaging and scRNA-seq of three biological replicates per condition are shown. **A** Exemplary cells from padFISH images of nascent RNA from *CXCL1* (cyan), *CXCL2* (yellow), *CXCL3* (blue), and *CXCL8* (magenta) at 30 min and 240 min of TNFα treatment. Zoom-ins to co-expression loci are shown. **B** Co-expression of *CXCL* TRGs at 30 min (top) and 240 min (bottom) of TNFα treatment. The proportions of *CXCL* co-expression combinations at co-expression loci from padFISH imaging and in cells from scRNA-seq data are shown. The error bars represent the standard errors of replicates. **C** Proportion of *CXCL* co-expression combinations in padFISH imaging over scRNA-seq at 30 min (left) and 240 min (right) of TNFα treatment. Spearman correlation is shown. The error bars represent the standard errors of replicates. Image acquisition and processing were performed by Irene Gerosa. Adapted from Seufert *et al.* (2024).

included *CXCL1/2/3/8*, *CXCL1/2/3*, *CXCL1/2/8*, *CXL2/3/8*, and *CXCL2/8*, while combinations lacking *CXCL2* expression were less common. The same analysis of scRNA-seq data revealed similar co-expression patterns across cells (**Figure 2.60B**, light grey). Furthermore, the proportions of co-expression combinations observed at co-localization loci and in cells showed high correlation (Spearman correlation coefficients of 0.83 and 0.79) between the two data types for both TNFα treatment time points (**Figure 2.60C**). This suggests that scRNA-seq data, despite lacking spatial subcellular resolution, accurately captures the co-expression of multiple genes at the same locus. Consequently, these findings support the conclusion that the previously observed higher co-expression within *TRG clusters* is indicative of local molecular mechanisms co-regulating transcription upon TNFα treatment.

TRGs reveal variable AC and DC features

To investigate these molecular mechanisms underlying transcription co-regulation, I performed *single cell and metacell co-accessibility* analysis on the scTurboATAC-seq data using the RWireX framework. The workflow designs, results, reproducibility among replicates, and mechanistic interpretations were extensively discussed in **Chapter 2.2**, using TNFα treatment in HUVECs as a model system. To shortly summarize, the *single cell co-accessibility workflow* identified autonomous links of co-accessibility, termed ACs, resulting from stochastic accessibility fluctuations in HUVECs of homogeneous cell states. The analysis of chromatin contacts from bulk HiC-seq data showed that these ACs displayed significantly higher chromatin contacts compared to the genome-wide background, suggesting that ACs represent chromatin contacts between active genomic regions. In contrast, the *metacell co-accessibility workflow* identified broad domains of contiguous co-accessibility, termed DCs, defined from cell state-driven accessibility changes in response to TNFα treatment. These DCs were independent of TADs and revealed TNFα-dependent accessibility changes, indicating activation or deactivation upon TNFα treatment. Analysis of TF binding activity using pseudo-bulk scTurboATAC-seq data showed significantly higher TF binding at binding sites in the DCs compared to the genome-wide non-DC background, suggesting that DCs are local nuclear subcompartments with altered concentrations of specific TFs that create unique transcription-regulatory environments.

Here, I utilized the previously identified ACs and DCs from **Chapter 2.2** to study transcriptional regulation of TRGs. For the three TNFα treatment conditions, approximately 27,000 to 35,000 consensus ACs were detected across the biological replicates (**Figures 2.61A**, **2.27A**). Interestingly, 12 % of these ACs were located at TRGs, with around 45 %

of these located at promoters, while the rest were in exonic or intronic regions (**Figure 2.61B**). This was a notable enrichment compared to the 9 % of total ATAC peaks located at promoters (see **Figure 2.57B**), indicating a specific regulatory role of these ACs at TRGs. For DCs, 4,885 consensus DCs were identified from the averaged *metacell co-accessibility* matrices of replicates (**Figures 2.61C**, **2.28B**), with 12 % of DCs also located at TRGs, specifically comprising a TRG promoter.

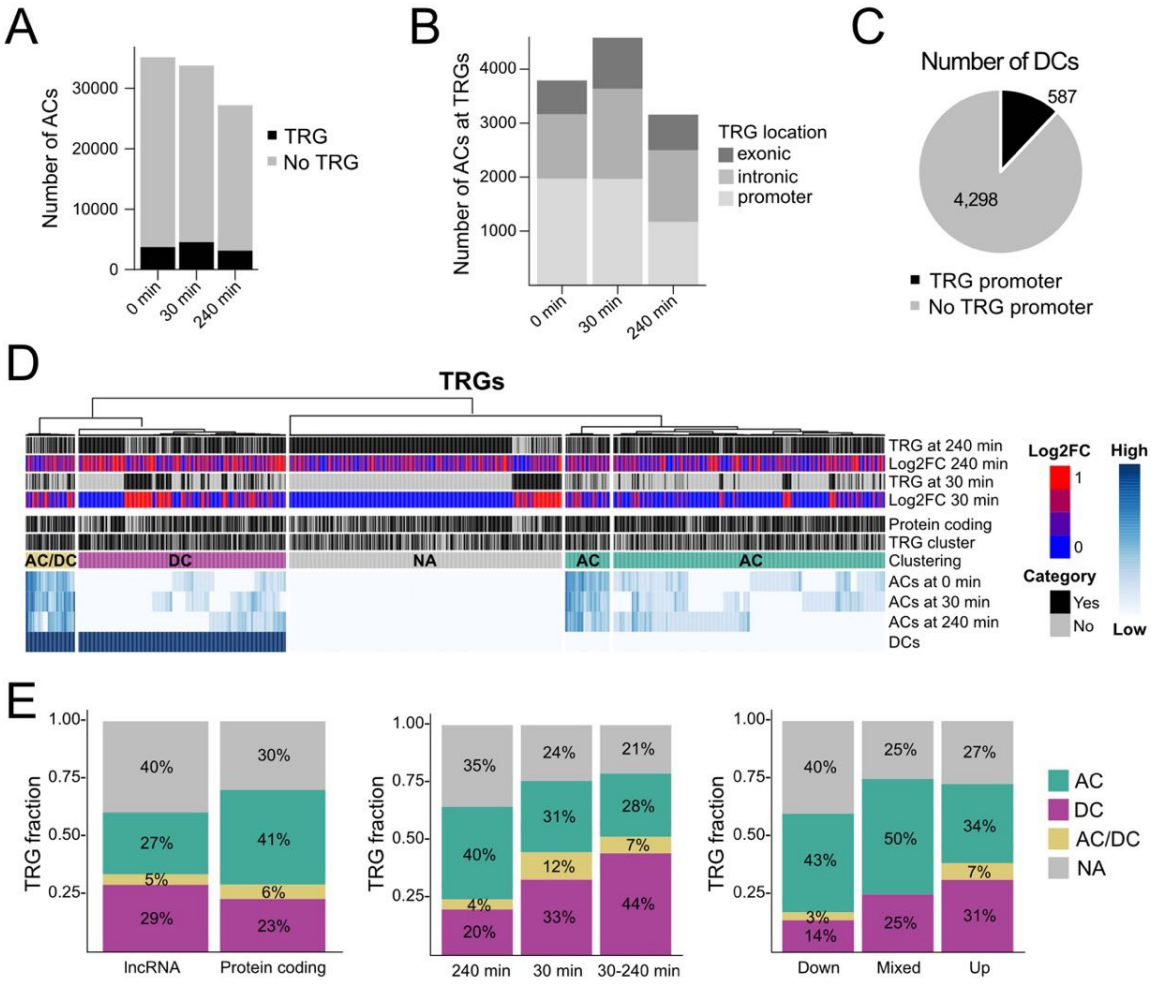


Figure 2.61 AC and DC features of chromatin co-accessibility at TRGs in untreated and TNFα-treated HUVECs from scTurboATAC-seq data. A Number of ACs at TRGs and non-TRG regions for TNFα treatment time points. Replicate consensus of TNFα treatment time points is shown. **B** Number of ACs at TRG promoters, exons, and introns for TNFα treatment time points. Replicate consensus of TNFα treatment time points is shown. **C** Proportion of DCs at TRG promoters and non-TRG promoter regions. Replicate consensus DCs are shown. **D** Quantification of ACs and DCs at TRG promoters per TRG. TRGs are clustered by Ward's method into predominantly AC-driven, DC-driven, AD/DC-driven, and NA groups. Differential TRGs after 30 and 240 min of TNFα treatment, gene type (protein-coding or lncRNA), and *TRG cluster* affiliation are annotated. **E** Proportion of AC-driven, DC-driven, AC/DC-driven, and NA TRGs among protein-coding and lncRNA TRGs (left), early, late and persistently differential TRGs (middle), and upregulated, downregulated and mixed regulated TRGs (right). Adapted from Seufert *et al.* (2024).

After quantifying ACs and DCs at TRG promoters, I used these co-accessibility features to classify TRGs by their regulatory mechanisms (**Figure 2.61D**). Hierarchical clustering based on the number of ACs and DCs at TRG promoters revealed five distinct TRG groups. A large fraction of TRGs (368) were located within DCs but exhibited few or no ACs (**Figure 2.61D**, purple color in “clustering” annotation). Consequently, these TRGs were classified as being predominantly DC-driven. In contrast, two other TRG groups, comprising 561 TRGs, exhibited high or medium numbers of ACs without DCs (**Figure 2.61D**, green color in “clustering” annotation). These TRGs were combinedly classified as being predominantly AC-driven. Additionally, some TRGs (87) were located within DCs and showed high numbers of ACs (**Figure 2.61D**, yellow color in “clustering” annotation), making them simultaneously driven by both ACs and DCs, termed AC/DC-driven. Lastly, a considerable number of TRGs (483) showed neither AC nor DC features and were classified as NA (**Figures 2.61D**, grey color in “clustering” annotation).

Interestingly, protein-coding TRGs were more frequently AC-driven (41 %) compared to DC-driven or NA (23 % and 30 %, respectively), while lncRNA TRGs were more frequently DC-driven or NA (29 % and 40 %, respectively; **Figures 2.61D**, “protein coding” annotation; **2.61E**, left). This suggests that ACs may represent a more targeted regulation of protein-coding TRGs, whereas DCs (and NA) might potentially regulate lncRNAs as “by-products” of broader regulatory events. Additionally, early and persistently differential TRGs were more often DC- and AC/DC-driven (33-44 % and 7-12 %, respectively), while late differential TRGs were predominantly regulated by ACs and NA (40 % and 35 %, respectively; **Figures 2.61D**, “TRG at 30 min” and “TRG at 240 min” annotation; **2.61E**, middle). Moreover, 31 % of upregulated TRGs were DC-driven, whereas only 14 % of downregulated TRGs were DC-driven (**Figures 2.61D**, “Log2FC 30 min” and “Log2FC 240 min” annotation; **2.61E**, right). In contrast, downregulated TRGs were mainly AC-driven or NA (43 % and 40 %, respectively; **Figures 2.61D**, “Log2FC 30 min” and “Log2FC 240 min” annotation; **2.61E**, right).

In summary, the 1,499 TRGs identified from scRNA-seq data demonstrated different mechanisms of transcriptional regulation, which were inferred using RWireX’s co-accessibility workflows on scTurboATAC-seq data. The TRGs were classified into DC-driven, AC-driven, AC/DC-driven, and NA groups based on their predominant regulatory mechanisms. This analysis revealed that DCs are more often involved in early and upregulated lncRNA TRG regulation, suggesting a fast but potentially less specific regulatory role of DCs. Conversely, ACs are more frequently associated with the regulation

of late and downregulated protein-coding TRGs, indicating a more targeted but slower regulatory role.

AC and DC features classify *TRG clusters*

After identifying different groups of TRGs with varying regulatory mechanisms, I aimed to investigate how these regulatory TRG groups relate to the previously identified *TRG clusters* in the genome. To achieve this, I calculated AC and DC scores for each *TRG cluster*. These scores reflect the proportion of TRGs in a cluster that were AC+AC/DC-driven or DC+AC/DC-driven, respectively. The AC and DC scores of *TRG clusters* were not normally distributed, as shown by the significant p-values from the Shapiro-Wilk test for normality (**Figure 2.62A**, histograms). This indicated that TRGs within the same cluster were enriched for either AC- or DC-related features, and that both molecular mechanisms co-regulate local transcription. When comparing the AC and DC scores with the co-expression of TRGs in a cluster, AC scores showed a negative correlation with *TRG cluster* co-expression (**Figure 2.62B**, top). In contrast, DC scores showed a positive correlation with co-expression (**Figure 2.62B**, bottom). Overall, these results suggest that ACs and DCs both co-regulate transcription in *TRG clusters*. Interestingly, while DCs promote simultaneous co-expression of TRGs in local clusters, AC-driven *TRG clusters* display alternating expression of their TRGs.

Next, I used these AC and DC scores of *TRG clusters* to classify the predominant regulatory mechanism within entire clusters. Specifically, I assigned *TRG clusters* as: (i) predominantly AC-driven if AC scores were ≥ 0.5 and DC scores were < 0.5, (ii) predominantly DC-driven if AC scores were < 0.5 and DC scores were ≥ 0.5, (iii) predominantly AC/DC-driven if both AC and DC scores were ≥ 0.5, and (iv) NA if both AC and DC scores were < 0.5 (**Figure 2.62A**, scatter plot). This classification resulted in 147 AC-driven (41 %), 86 DC-driven (24 %), 51 AC/DC-driven (14 %), and 72 NA (20 %) *TRG clusters* (**Figure 2.62C**). AC-driven *TRG clusters* consisted of at least 50 % AC-driven TRGs, with the remaining TRGs mostly NA and rarely DC- or AC/DC-driven TRGs (**Figure 2.62D**, top). Similarly, DC-driven *TRG clusters* contained at least 50 % DC-driven TRGs, with the rest primarily being NA, but also some AC-driven or AC/DC-driven TRGs (**Figure 2.62D**, second from bottom). Approximately 60 % of AC/DC-driven *TRG clusters* contained mostly AC/DC-driven TRGs, while the rest comprised a 50:50 mix of solely AC-driven and DC-driven TRGs, respectively (**Figure 2.62D**, second from top). AC/DC-driven *TRG clusters* comprised only few NA TRGs. In contrast, NA *TRG clusters* were composed mostly of NA TRGs, with only a few containing AC-, DC-, or AD/DC-driven TRGs (**Figure 2.62D**, bottom).

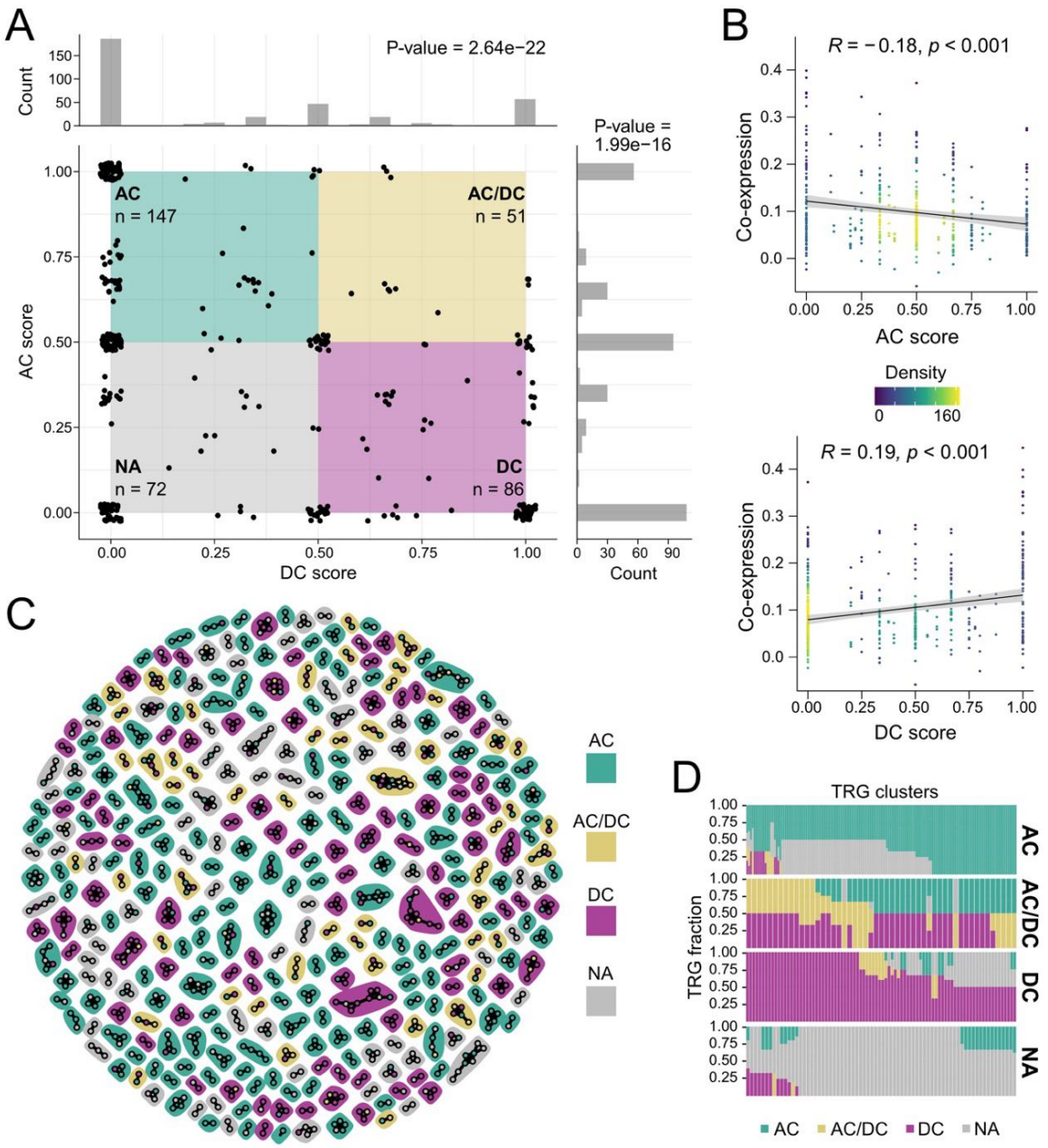


Figure 2.62 AC and DC features of chromatin co-accessibility at *TRG clusters* in untreated and TNFα-treated HUVECs from scTurboATAC-seq data. A AC scores and DC scores of *TRG clusters*. Overlaying data points are visualized by jitter plot. The background color represents the areas classified as predominantly AC-driven, DC-driven, AD/DC-driven, and NA *TRG clusters*. Histograms of AC scores and DC scores are shown. P-values from Shapiro-Wilk tests are indicated. **B** Co-expression over AC score (top) and DC score (bottom) in *TRG clusters*. The colors of points reflect the density of *TRG clusters*. Linear regression line and Spearman correlation are shown. **C** Regulatory mechanisms of clustered TRGs. Each data point represents one TRG, where color indicates the predominant regulatory mechanism of AC, DC, AC/DC, or NA. Edges between TRG neighbors below 500 kb distance are drawn. The background color of *TRG clusters* reflects the predominant regulatory mechanisms of AC, DC, AC/DC, or NA. **D** Composition of AC-driven (top), AC/DC-driven (second from top), DC-driven (second from bottom), and NA (bottom) *TRG clusters*. The proportion of AC-driven, DC-driven, AC/DC-driven, and NA TRGs in the *TRG clusters* is shown. Adapted from Seufert *et al.* (2024).

Having identified *TRG clusters* regulated by different molecular mechanisms, I next examined how these mechanisms govern the expression of multiple TRGs. Therefore, examples of different regulatory mechanisms for AC-, DC-, and AC/DC-driven *TRG clusters* are shown in **Figure 2.63**. The *TRG cluster* of *KLF10* and *GASAL1* has been previously shown in **Chapter 2.2** to showcase AC reproducibility. It is a suitable representative of AC-driven *TRG clusters*, with an AC score of 1.0 and a DC score of 0.0. As described previously, ACs between the TRG promoters were detected and present in almost all cells (indicated by PAC values above 95), making them *frequent ACs* (**Figure 2.63A**). Additional *frequent ACs* between the TRG promoters and two distal H3K27ac peaks were detected. In the untreated condition, the TRG promoters and a downstream H3K27ac peak were linked by these *frequent ACs*. After 30 min of TNFα treatment, the co-accessibility scores of these preexisting ACs increased (from around 0.2 in the untreated condition to 0.3 at 30 min). Moreover, *frequent ACs* connecting the TRG promoters to an intermediate H3K27ac peak emerged. This coincided with a significant increase in *KLF10* and *GASAL1* expression at the 30 min treatment time point (*KLF10* log2FC of 3.04; *GASAL1* log2FC of 1.3). After 240 min of TNFα treatment, the *frequent AC* between the TRG promoters disappeared, while those with distal H3K27ac peaks persisted. Furthermore, the co-accessibility scores of the remaining ACs fell below 0.2, aligning with a return of *KLF10* and *GASAL1* expression to initial levels (no significant differential expression for *KLF10* and *GASAL1* between 0 min and 240 min). *Metacell co-accessibility* analysis revealed no DCs in this *TRG cluster* but showed *blue stripes* of anti-correlated accessibility with high co-accessibility at their intersections for the previously observed sites of *frequent ACs* (**Figure 2.63B**, top). Furthermore, the *frequent ACs* and *blue stripes* corresponded with increased chromatin contact frequencies from HiC-seq data, visible as *red stripes* extending beyond TAD boundaries (**Figure 2.63B**, bottom). Altogether, these findings suggest that early upregulation of *KLF10* and *GASAL1* expression is co-regulated by targeted loop formation between the TRG promoters and two distal H3K27ac peaks of potential *distal CREs*.

In contrast, the *TRG cluster* of *WAKMAR2* and *TNFAIP3* was previously shown in **Chapter 2.2** to illustrate DC reproducibility and its underlying molecular mechanism of local TF enrichment. It is a suitable example of DC-driven *TRG clusters*, as it possesses a DC score of 1.0 and an AC score of 0.0. For both *WAKMAR2* and *TNFAIP3*, no ACs were detected at their promoters (**Figure 2.63C**). Their extended genomic region showed only a few scattered ACs, that possessed low co-accessibility scores and PAC values. On the contrary, *metacell co-accessibility* revealed multiple DCs around the two TRGs, which were in total approximately 200 kb in size and collectively termed the merged *WAKMAR2*/

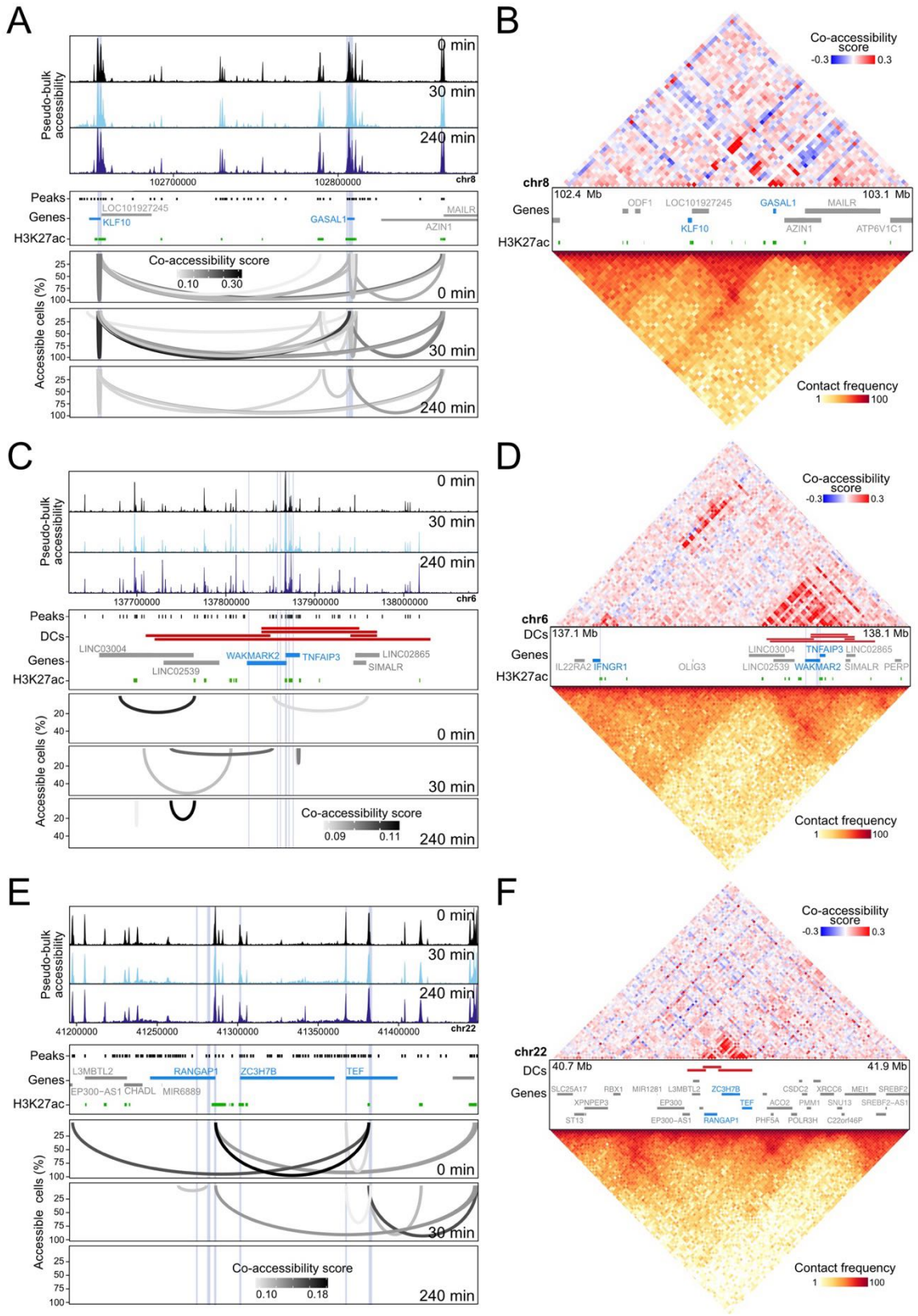


Figure 2.63 Examples of AC-driven, DC-driven, and AC/DC-driven *TRG clusters* in untreated and TNFα-treated HUVECs from scTurboATAC-seq data. A Pseudo-bulk chromatin accessibility (top) and ACs (bottom) at TNFα treatment time points at AC-driven *TRG cluster* of *KLF10* and *GASAL1*. ATAC peaks (black, 1 kb extended), genes (grey), TRGs (blue) and 1 kb regions around their TSSs (light blue), and H3K27ac peaks from ChIP-seq at 30 min time point (green) are indicated. Consensus ACs from replicates at TRG promoters are shown. The grayscale and height of loops reflect co-accessibility scores and percent accessible cells of ACs.

Figure 2.63 (continued) B Average *metacell co-accessibility* map from replicates (top) and chromatin contact map from HiC-seq data of untreated HUVECs (bottom) at AC-driven *TRG cluster* of *KLF10* and *GASAL1*. DCs from *average metacell co-accessibility* (red), genes (grey), TRGs (blue) and 1 kb regions around their TSSs (light blue), and H3K27ac peaks from ChIP-seq at the 30 min time point (green) are indicated. Limits of the co-accessibility score color scale bar are set to -0.3 and 0.3. The upper limit of the chromatin contact color scale bar is set to 100. **C** Same as panel A for DC-driven *TRG cluster* of *WAKMAR2* and *TNFAIP3*. All consensus ACs from replicates are shown. **D** Same as panel B for DC-driven *TRG cluster* of *WAKMAR2* and *TNFAIP3*. **E** Same as panel A for AC/DC-driven *TRG cluster* of *RANGAP1*, *ZC3H7B*, and *TEF*. **F** Same as panel E for AC/DC-driven *TRG cluster* of *RANGAP1*, *ZC3H7B*, and *TEF*. Adapted from Seufert *et al.* (2024).

TNFAIP3 DC (**Figure 2.63D**, top). Accessibility in all these DCs was significantly upregulated after 30 min (log2FCs between 0.54-0.6) as well as 240 min (log2FCs between 0.55-0.79) of TNFα treatment, which was likely caused by the increased local activity of multiple TFs in the merged *WAKMAR2*/*TNFAIP3* DC as shown before (see **Figure 2.34**). This suggested that the increase in accessibility of the merged *WAKMAR2*/*TNFAIP3* DC reflects its higher activity upon TNFα treatment. The enhanced DC activity coincided with the significantly increased expression of *WAKMAR2* (log2FC of 2.1 at 30 min; log2FC of 2.7 at 240 min) and *TNFAIP3* (log2FC of 5.9 at 30min; log2FC of 4.8 at 240min) at both TNFα treatment time points. Interestingly, the merged *WAKMAR2*/*TNFAIP3* DC was connected by high *metacell co-accessibility* scores to the TRG *IFNGR1*, which was 700 kb upstream of the *WAKMAR2* and *TNFAIP3 TRG cluster*. *IFNGR1* was located in a roughly 50 kb region of increased *metacell co-accessibility*, but no DC was identified. The region around *IFNGR1* showed increased accessibility after 30 min of TNFα treatment, which further increased after 240 min. Consistently, *IFNGR1* showed significantly increased expression after 240 min of TNFα treatment (not significant at 30 min; log2FC of 1.4 at 240 min). Notably, the high *metacell co-accessibility* between the merged *WAKMAR2*/*TNFAIP3* DC and the distal *IFNGR1* region spanned across a TAD boundary (**Figure 2.63D**, bottom). Nevertheless, the visually higher *metacell co-accessibility* indicates that this distal *IFNGR1* region might spatially co-assemble with the merged *WAKMAR2*/*TNFAIP3* DC into a combined local subcompartment of increased TF activity. These findings indicate that this subcompartment co-regulates the increase in expression of the *TRG cluster WAKMAR2* and *TNFAIP3* as well as the distal TRG *IFNGR1.*

Lastly, a suitable example for AC/DC-driven *TRG clusters* was the *TRG cluster* of *RANGAP1*, *ZC3H7B*, and *TEF*, which showed both a high AC score of 0.66 and a high DC score of 1.0. In untreated and 30 min TNFα treated HUVECs, multiple *frequent ACs* between promoters of *RANGAP1* and *TEF* were present as well as with distal H3K27ac peaks (**Figure 2.63E**). These ACs disappeared in 240 min TNFα treated HUVECs. Simultaneously, expression of all three TRGs increased after 240 min of TNFα treatment

(*RANGAP1* log2FC of 1.7; *ZC3H7B* log2FC of 2.2; *TEF* log2FC of 1.4), while no significant differences were detected at the 30 min TNFα treatment time point. Moreover, three DCs at the *TRG cluster* were detected in *metacell co-accessibility* (**Figure 2.63F**, top). All showed significantly upregulated accessibility upon TNFα treatment (log2FCs between 0.19-0.41 at 30 min and 0.16-0.39 at 240 min), indicating increasing activity of the subcompartment. Additionally, the metacell co-accessibility map revealed *blue stripes* of anti-correlated accessibility with high co-accessibility at their intersections, that extended from beyond the merged *RANGAP1*/*ZC3H7B*/*TEF* DC. Again, the *blue stripes* coincided with *red stripes* of increased chromatin contact frequencies from HiC-seq data (**Figure 2.63F**, bottom). Taken together, these findings suggest that TNFα treatment induces the early formation of an activating local subcompartment at these three TRGs. This subcompartment does not result in immediate expression induction, as preexisting chromatin loops or interactions of the TRG promoters might block their regulation. Upon loss of these loops or interactions, the TRGs show late upregulation in expression potentially by the local subcompartment. This shows that ACs and DCs are not mutually exclusive but might coordinately co-regulate time-specific expression induction at *TRG clusters*. Furthermore, this example shows that DCs can also form in chromatin regions with preexisting architectural chromatin loops or interactions.

AC-, DC- and AC/DC-driven TRGs show varying transcriptional bursting kinetics

Next, I aimed to examine differences in the transcriptional bursting kinetics of AC-driven, DC-driven and AC/DC-driven TRGs. To do so, I inferred transcriptional bursting kinetics from snap-shots in time of nuclear RNA content in multiple hundred cells (**Figure 2.64**, top). On the one hand, the burst frequency of a gene, described as the on rate (k_{on}) in the two-state model of transcription, was computed from the number of nuclei that contained intronic RNA of the respective gene (**Figure 2.64**, bottom). On the other hand, the burst size of a gene, described as the ratio of synthesis and off rates (k_{syn} and k_{off}, respectively), was computed from the number of the gene's intronic RNAs per nucleus. The snap-shots of nuclear RNA per cell were acquired either by quantification of intronic RNA from snRNA-seq data or direct measurement of intronic RNA by padFISH imaging.

First, I used snRNA-seq data from all TNFα treatment time points to quantify intronic RNAs genome-wide. Following, I used these to infer time point-specific burst frequencies and burst sizes of AC-driven, DC-driven, and AC/DC-driven TRGs. In untreated HUVECs, burst sizes of the three TRG groups were highly similar, showing medians between 2.3 and 2.7 RNA molecules per burst (**Figure 2.65A**, density plot on top). In contrast, burst

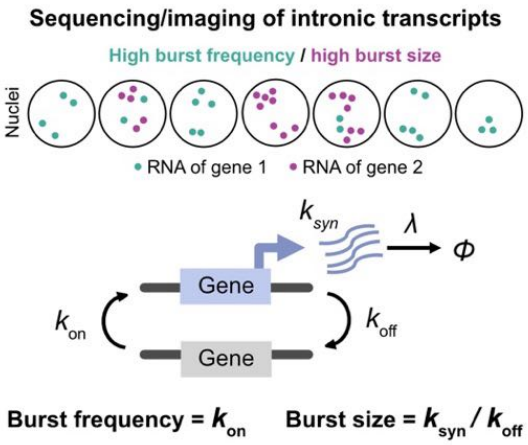


Figure 2.64 Inference of transcriptional bursting kinetics from transcriptomic snap-shots in time of multiple hundred cells. Intronic RNA was quantified in individual nuclei from sequencing or imaging data. Burst frequency and burst size were computed after the two-state model of transcription. Genes switch between their active and inactive state by on rate (k_{on}) and off rate (k_{off}). In active state, RNA is transcribed at synthesis rate (k_{syn}). These transcripts are again degraded at degradation rate (λ). Adapted from Seufert *et al.* (2024).

frequencies varied between the TRG groups in the untreated condition (**Figure 2.65A**, density plot on the right). While DC-driven TRGs only showed a median of 1.1 bursts per hour, AC-driven and AC/DC-driven TRGs showed medians of 1.7 and 2.2 bursts per hour, respectively. Similarly, no differences in burst sizes were present between TRG groups after 30 min of TNFα treatment (2.3-2.8 median RNA molecules per burst; **Figure 2.65B**, density plot on top), whereas AC/DC-driven TRGs showed again more than two-fold higher burst frequencies than DC-driven TRGs (2.9 and 1.3 median bursts per hour, respectively; **Figure 2.65B**, density plot on the right). After 240 min of TNFα treatment, burst sizes of DC-driven and AC/DC-driven TRGs (medians of 4.8 and 4.4 median RNA molecules per burst, respectively) were approximately two-fold higher than AC-driven TRGs (2.8 median RNA molecules per burst; **Figure 2.65C**, density plot on top). For burst frequencies, AC/DC-driven TRGs showed highest median of 2.2 bursts per hour compared to AC-driven and DC-driven TRGs (1.3 and 1.1 median bursts per hour, respectively; **Figure 2.65C**, density plot on the right). To confirm these observations, I performed differential analysis of bursting kinetics between the three TRG groups. Indeed, DC-driven TRGs revealed significantly higher burst sizes compared to AC-driven TRGs (**Figure 2.65D**, shape outlines). Conversely, AC-driven and AC/DC-driven TRGs revealed significantly higher burst frequencies compared to DC-driven TRGs (**Figure 2.65D**, filled shapes). Additionally, AC/DC-driven TRGs showed significantly higher burst frequencies compared to the only AC-driven TRGs.

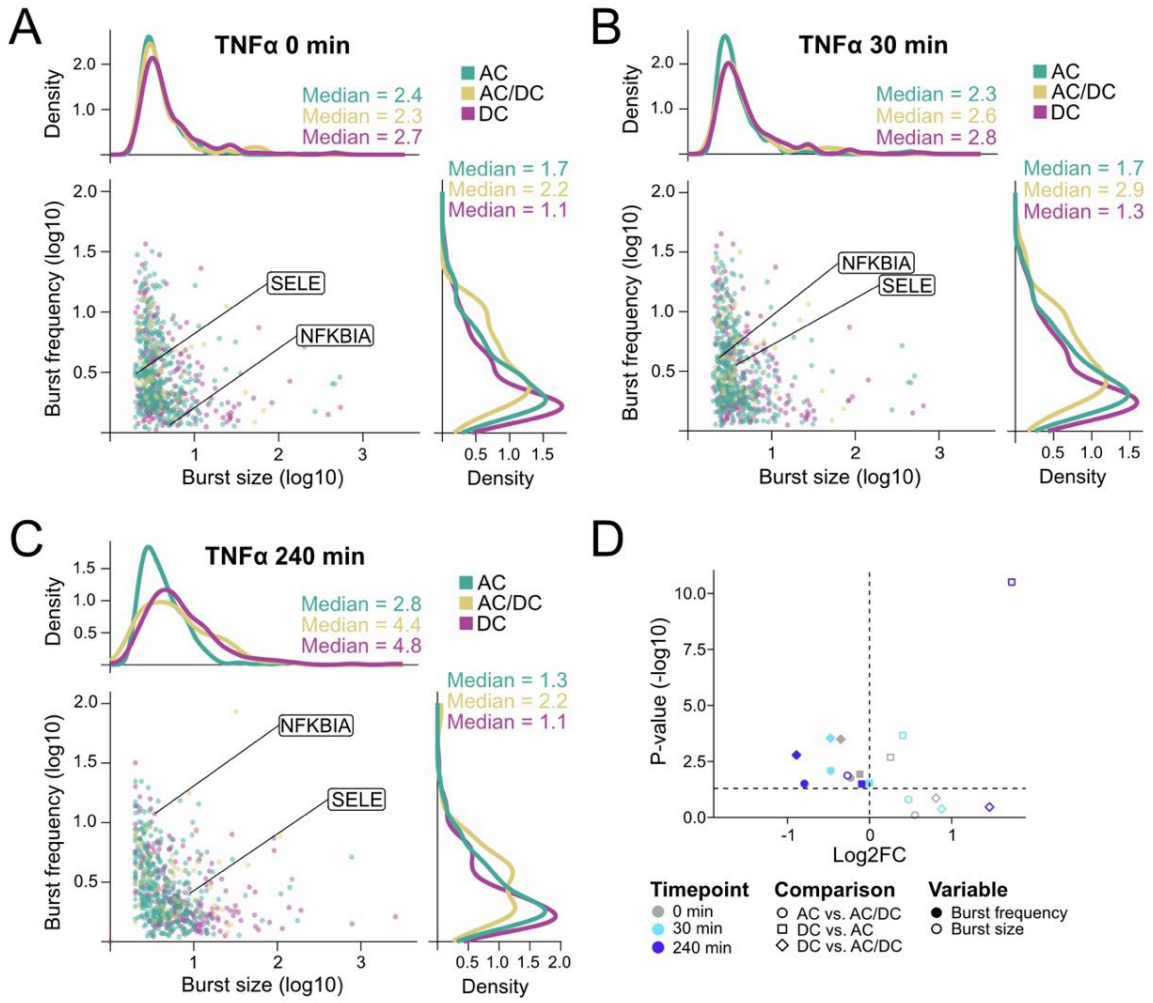


Figure 2.65 Transcriptional bursting kinetics of AC-driven, DC-driven, and AC/DC-driven TRGs in untreated and TNFα-treated HUVECs from snRNA-seq data. Data of nuclear transcripts from snRNA-seq of approximately 300 cells per condition are shown. **A** Burst frequency over burst size of AC-driven (green), DC-driven (purple), and AC/DC-driven (yellow) TRGs in untreated HUVECs. Density curves and medians of burst frequency and burst size are shown. Exemplary TRGs *NFKBIA* and *SELE* are highlighted. **B** Same as panel A for 30 min TNFα treated HUVECs. **C** Same as panel A for 240 min TNFα treated HUVECs. **D** Differential bursting kinetics between AC-driven TRGs and AC/DC-driven TRGs (circle), DC-driven TRGs and AC-driven TRGs (rectangle), as well as DC-driven TRGs and AC/DC-driven TRGs (rhombus). Differential burst frequency (filled shapes) and burst size (shape outlines) are shown. The color indicates the TNFα treatment time point. The dashed line represents the significance cutoff (p-value < 0.05 from two-sided Wilcoxon test). Adapted from Seufert *et al.* (2024).

We validated the transcriptional bursting kinetics inferred from snRNA-seq data for two exemplary TRGs using padFISH imaging of intronic transcripts. The TRG *NFKBIA* showed a strong increase in burst frequency upon TNFα treatment in snRNA-seq data, while burst size remained comparably low (**Figures 2.65A-C**, annotation of *NFKBIA* in scatter plots; **2.66A+B**). Similarly, padFISH imaging detected intronic *NFKBIA* transcripts in many nuclei of the TNFα-treated HUVECs (**Figure 2.66A**) and the data confirmed burst frequency increases and constant burst size of *NFKBIA* upon TNFα treatment (**Figure 2.66B**). In contrast to *NFKBIA*, the TRG *SELE* showed a strong increase in burst size upon TNFα

treatment in snRNA-seq data, while burst frequency remained comparably low (**Figures 2.65A-C**, annotation of *SELE* in scatter plots; **2.66C+D**). This was confirmed by padFISH imaging, which detected intronic *SELE* transcript in few nuclei but at high abundance (**Figure 2.66C**). This resulted in constant burst frequency and TNFα-induced burst size of *SELE* inferred from padFISH data (**Figure 2.66D**). Overall, the padFISH data validated the previously inferred transcriptional bursting kinetics from snRNA-seq data. Here, DC-driven TRGs revealed significantly higher burst sizes compared to AC-driven TRGs. Vice versa, AC-driven TRGs revealed significantly higher burst frequencies compared to DC-driven TRGs. Interestingly, AC/DC-driven TRGs showed both increased burst sizes and burst frequencies.

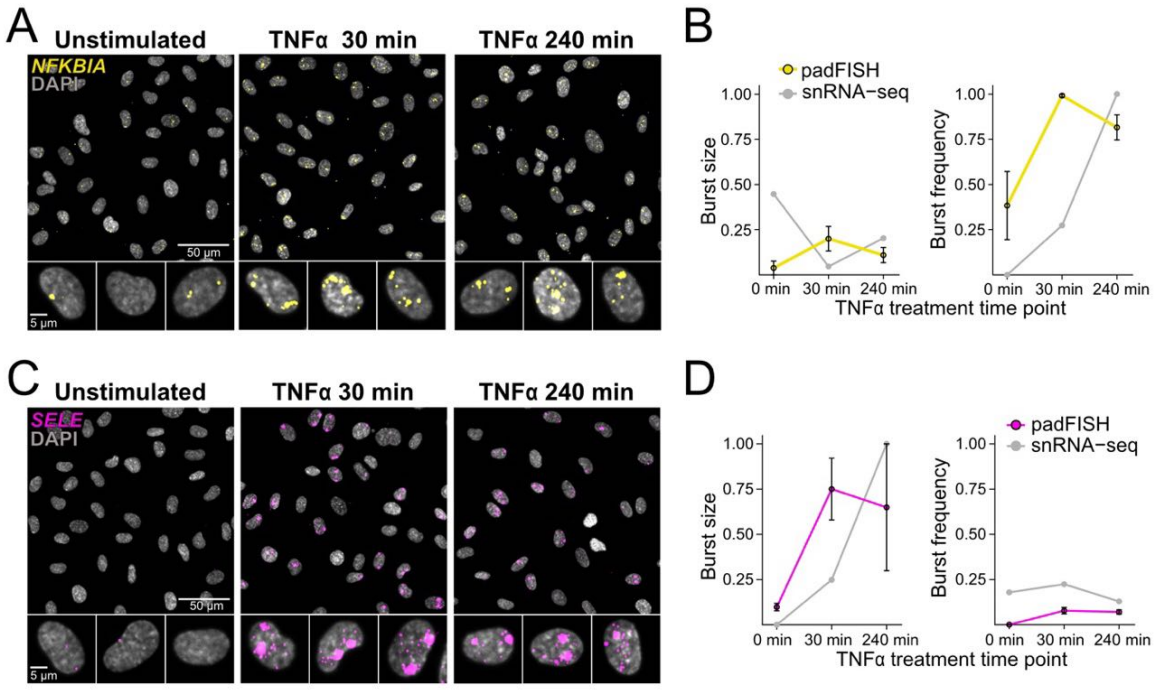


Figure 2.66 Transcriptional bursting kinetics of *NFKBIA* and *SELE* in untreated and TNFα-treated HUVECs from spatial transcriptomics and snRNA-seq data. Data from 2 biological replicates per condition are shown. **A** padFISH images of nascent RNA from *NFKBIA* (yellow) in untreated (left), 30 min (middle) and 240 min (right) of TNFα treatment. Zoom-ins to exemplary cells are shown. **B** Burst size (left) and frequency (right) of *NFKBIA* from padFISH imaging and snRNA-seq at TNFα treatment time points. The error bars represent the standard errors of replicates. **C** Same as panel A for *SELE* in magenta. **D** Same as panel B for *SELE*. Image acquisition and processing were performed by Irene Gerosa. Adapted from Seufert *et al.* (2024).

Promoter-mediated regulation of TRGs shows strongest transcriptional response

Lastly, I aimed to investigate the role of promoter-mediated regulation compared to AC- and DC-mediated regulation of TRG expression. Only 9 % of TRGs contained a differentially accessible ATAC peak upon TNFα treatment at their promoter (see **Figure 2.58A**). These TRGs were classified as promoter-regulated. Notably, the majority of

promoter-regulated TRGs (81 %) showed additional regulation via ACs and / or DCs (**Figure 2.67A**). The transcriptional response of promoter-regulated TRGs was significantly stronger compared to AC-regulated TRGs after 30 min of TNFα treatment (**Figure 2.67B**, left). After 240 min, promoter-regulated TRGs showed a significantly stronger transcriptional response compared to all AC-, DC-, and AC/DC-regulated TRGs (**Figure 2.67B**, right). Additionally, DC- and AC/DC-regulated TRGs showed a significantly stronger early transcriptional response compared to AC-regulated TRGs (**Figure 2.67B**, left).

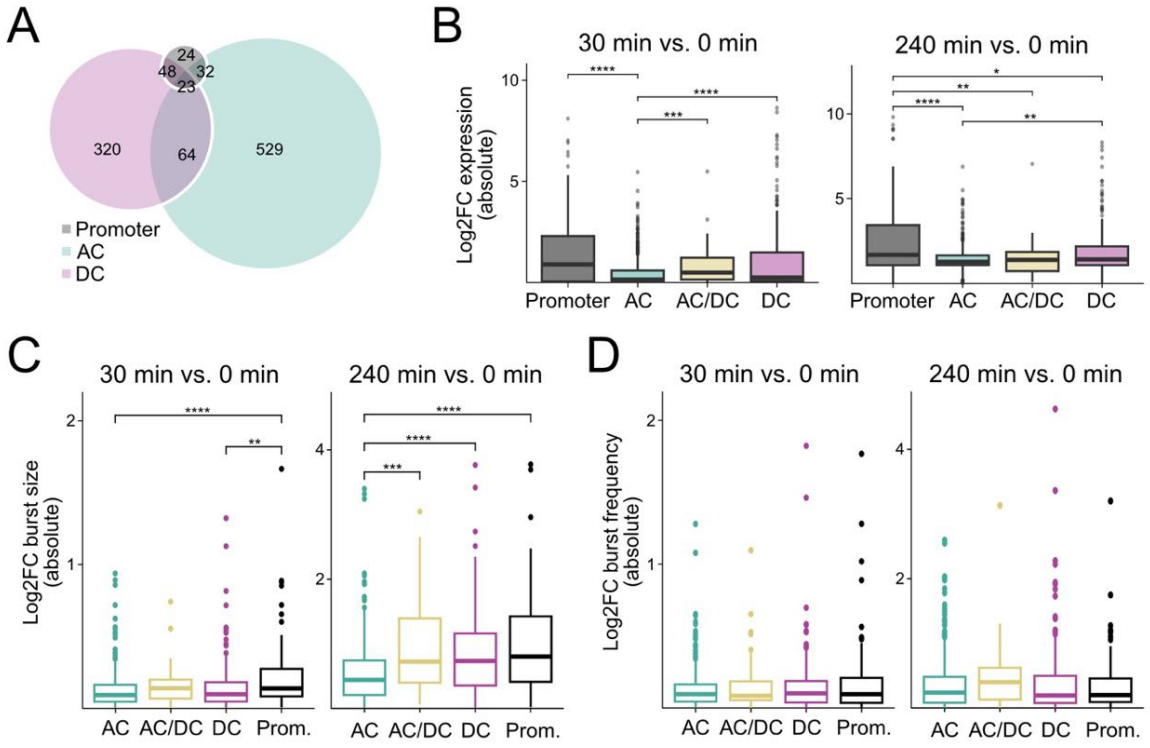


Figure 2.67 Promoter-driven regulation of TRG expression in untreated and TNFα-treated HUVECs from scRNA-seq and snRNA-seq data. A Number of TRGs regulated via their promoters, ACs, and/or DCs. **B** Absolute TRG expression changes after 30 min (left) and 240 min (right) of TNFα treatment in HUVECs. Log2FCs are shown for TRGs regulated via their promoter, ACs, DCs, or both ACs and DCs. Significant p-values from Wilcoxon test are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001. **C** Absolute TRG burst size changes after 30 min (left) and 240 min (right) of TNFα treatment in HUVECs. Log2FCs are shown for TRGs regulated via their promoter, ACs, DCs, or both ACs and DCs. Significant p-values from Wilcoxon test are indicated as *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001. **D** Same as panel C for burst frequency changes.

The transcriptional bursting kinetics of promoter-regulated TRGs revealed a strong differential regulation via burst size (**Figure 2.67C**, black). In contrast, promoter-regulated TRGs showed only low differences in burst frequency upon TNFα treatment (**Figure 2.67D**, black). Promoter-regulated TRGs revealed significantly higher burst size differences compared to AC- and DC-regulated TRGs after 30 min of TNFα treatment (**Figure 2.67C**, left). After 240 min, promoter-, DC-, and AC/DC-regulated TRGs showed

significantly stronger burst size differences compared to AC-regulated TRGs (**Figure 2.67C**, right). For burst frequencies, no TRG regulation group showed exceptionally strong differences at any time point (**Figure 2.67D**). Overall, this suggests that promoter-regulated TRGs show a fast, yet persistent transcriptional response that is stronger compared to AC-, DC-, and AC/DC-driven regulation. Furthermore, promoter-driven TRGs are predominantly regulated via burst size, potentially driven by higher promoter accessibility to TFs and the transcriptional machinery.

In summary, in this project I studied how ACs and DCs, both long-range regulatory mechanisms, co-regulate the simultaneous transcriptional response of multiple genes to TNFα treatment in HUVECs. Additionally, I explored their regulatory effects on transcriptional bursting kinetics. In HUVECs, the TNFα treatment induced differential expression of specific genes, so-called TNFα-regulated genes. These TRGs clustered along the genomic coordinate, forming *TRG clusters*. TRGs within clusters showed higher co-expression across single cells compared to isolated TRGs, which indicates that transcription is locally co-regulated by shared molecular mechanisms within *TRG clusters* upon TNFα treatment. On chromatin level, the TNFα treatment induced changes in chromatin accessibility only at specific loci, which were driven by increased TF binding activity of NF-κB family members and its secondary TF targets. Notably, these differential chromatin loci were mainly located in gene body or intergenic regions and less so at the promoters of TRGs, suggesting that differential expression is primarily regulated by *distal CREs* and long-range regulatory mechanisms.

TRGs showed both specific long-range chromatin contacts or interactions, represented by ACs, and broader, local subcompartments of increased TF activity, represented by DCs. These were used to classify TRGs by their predominant regulatory mechanism into AC-driven, DC-driven, and AC/DC-driven TRGs. Interestingly, DCs more often regulated early-responsive and upregulated long non-coding RNA (lncRNA) TRGs, indicating a fast, activating and potentially less specific regulatory role of nuclear subcompartments. In contrast, ACs more frequently regulated late-responsive and downregulated protein-coding TRGs, suggesting a highly specific but slower transcription regulation via specific chromatin interactions. The previously identified *TRG clusters* showed significant enrichment for either or both AC and DC features, indicating that both specific chromatin interactions and nuclear subcompartments can co-regulate transcription at local *TRG clusters*. In DC-driven *TRG clusters*, this resulted in simultaneous co-expression of TRGs, while AC-driven *TRG clusters* showed alternating expression of their TRGs. This suggested that specific chromatin interactions do not co-regulate transcription by forming

multi-way transcription factories, but rather form contacts with on gene at a time causing alternating transcriptional bursts between the TRGs. Finally, DC-driven TRGs revealed significantly higher burst sizes compared to AC-driven TRGs and vice versa, AC-driven TRGs revealed significantly higher burst frequencies compared to DC-driven TRGs. Interestingly, AC/DC-driven TRGs showed both, increased burst sizes and burst frequencies. This indicated that nuclear subcompartment and specific chromatin contacts regulate different steps during the transcriptional bursting. Moreover, promoter-regulated TRGs showed a significantly stronger transcriptional response at both the early and late time point compared to AC-, and DC-regulated TRGs. They were primarily regulated via burst size. These observed differences between promoter-, AC-, and DC-driven regulation constitute multi-layered regulatory networks in the nuclei of HUVECs that can tightly regulate the temporal complexity and intensity of the transcriptional response to TNFα treatment.

In this chapter, I applied my computational framework for chromatin co-accessibility analysis to various mammalian systems under perturbation. Building on the insights from the previous chapter on the underlying molecular mechanisms of observed chromatin co-accessibility features, I was able to study regulatory mechanisms genome-wide and evaluate their potential interplay in creating a multi-layered network of transcription regulation. Additionally, I assessed their functional impact transcriptional activity. In the discussion, I will apply these findings to derive a model of transcription regulation that integrates genome-wide and simultaneous information on both chromatin-centric and protein-centric transcription regulation.

3. Discussion

In this thesis, a model of transcription regulation was developed that integrates genome-wide information on chromatin tropology- and TF-mediated mechanisms using chromatin accessibility sequencing data at single-cell resolution. I successfully addressed three specific objectives: (i) The experimental and computational analysis of scATAC-seq was advanced. (ii) I developed a computational framework termed RWireX to dissect molecular mechanisms underlying chromatin co-accessibility. (iii) By applying RWireX, I revealed the structure-function relationship between different regulatory mechanisms and their transcriptional output. In the following sections, I discuss my findings of these three parts. Afterwards, I integrate my findings into the AC/DC model of transcription regulation and discuss its generalizability and implications for inflammatory response, differentiation and cancer.

3.1. Advancing the experimental and computational analysis of scATAC-seq

TurboATAC protocol reduces data sparsity in scATAC-seq data

A general challenge in scATAC-seq experiments is to have a large number of cells with sufficiently high coverage per cell to characterize their chromatin accessibility landscapes (De Rop *et al.*, 2024). I encountered this challenge when comparing two biological replicates of IFNβ-treated MEFs: One replicate contained only 5,000 cells with roughly 35,000 unique ATAC fragments per cell on average (see **Section 2.1.1**). At the same time, the other replicate comprised approximately 20,000 cells but with lower unique ATAC fragments of roughly 7,500 per cell on average. The latter replicate displayed significant limitations in its downstream analysis of single cells.

To reduce scATAC-seq data sparsity, we developed the TurboATAC protocol, which enhances coverage by optimizing transposase reaction efficiency in both single- and multi-omic experiments (see **Section 2.1.2**). This optimization yielded significantly higher numbers of unique fragments per 10,000 sequenced reads and cell compared to the commercial 10x Genomics scATAC-seq and Multiome (scRNA-seq and scATAC-seq) protocols (see **Figures 2.7C**, **2.8B**, **2.9C**). The TurboATAC protocol reduced the number of duplicate sequencing reads, likely increasing scATAC library complexity. Thus, our TurboATAC protocol reduced data sparsity without increasing sequencing costs, while maintaining high cell numbers.

Moreover, our TurboATAC protocol outperformed other scATAC-seq approaches regarding sequencing coverage, since several tested methods performed worse than the 10x Genomics scATAC-seq v2 protocol (TXGv2) (De Rop *et al.*, 2024). In this benchmarking study of single-cell ATAC sequencing protocols, TXGv2 performed best with approximately 4,000 unique fragments in peaks at a sequencing depth of 10,000 reads per cell (see Figure 2A in De Rop *et al.* (2024)). Together with an average FRIP score of around 0.7 in TXGv2 (see Figure 1H in De Rop *et al.* (2024)), this resulted in roughly 5,700 unique fragments per 10,000 reads and cell. In our experiments, TXGv2 only yielded about 3,200 unique fragments per 10,000 reads and cell, while our TurboATAC protocol resulted in approximately 5,600 unique fragments per 10,000 reads and cell (**Figure 2.8B**). This indicates high variability in scATAC-seq coverage between data sets for identical protocols, likely due to different experimental set-ups, experimenters and model systems. Nevertheless, our TurboATAC protocol demonstrated comparable or superior performance to all scATAC-seq protocols tested in the benchmarking study.

Another recent experimental approach improves the single-cell combinatorial indexing (sci)-based scATAC-seq protocol by utilizing the small molecule inhibitor Pitstop 2 (Mulqueen *et al.*, 2019). The inhibitor increases the ability of Tn5 to enter nuclei, improving the efficiency of the ATAC reaction. The improved sci assay resulted in an increased library complexity with 90 % of unique reads per cell on average (see Extended Data Figure 2A in Mulqueen *et al.* (2019)). This was significantly higher compared to 43-64 % unique reads with the TurboATAC protocol. observed for varying biological systems and single-cell sequencing depths (**Tables 2.5**, **2.7**, **2.8**). However, single-cell sequencing depth strongly influences duplication rate in addition to library complexity (**Table 2.5**, note lower duplication rate for down-sampled scTurboATAC-seq sample). Indeed, single-cell sequencing depth was significantly lower for the sci assay with 40,000 unique reads per cell (see Figure 1E in Mulqueen *et al.* (2019)) compared to 130,000 unique reads per cell

for our scTurboATAC-seq data (**Table 2.6**, two times the mean number of unique fragments per cell). Moreover, the experimental procedure of the improved sci-based assay is more complex to the TurboATAC protocol, as additional treatment and incubation steps with the Pitstop 2 inhibitor are necessary. Taken together, enhanced sciATAC-seq represents an alternative to the TurboATAC protocol to improve library complexity and reduce data sparsity. However, its experimental procedure is more complicated and available data do not indicate superior performance.

Chromatin accessibility is a stochastic event in single cells

When assessing this sequencing coverage in single cells, the question arises as to what the expected coverage of accessible chromatin regions in individual cells is. Initial ATAC-seq data from bulk experiments of 50,000 cells identified approximately 75,000 peaks, highly comparable to those found in DNase HS-seq data sets (Buenrostro *et al.*, 2013). More recent bulk ATAC-seq data from IFNβ-stimulated MEFs identified an even higher number of peaks, ranging from 140,000 to 230,000 per replicate (**Figure 2.11A**) (Muckenhuber *et al.*, 2023). Pseudo-bulk scATAC-seq data from 10x Genomics and TurboATAC protocols yielded comparable peak numbers, with approximately 150,000 peaks each.

However, at single-cell resolution, only a fraction of these pseudo-bulk peaks was accessible in individual cells. Specifically, only 7.5 % of the pseudo-bulk peaks were accessible in individual cells using the 10x Genomic protocol, while TurboATAC detected 15 % of the peaks per single cell (**Figure 2.11B**). These findings are consistent with previous studies, which reported that only 9.4 % of promoters were accessible in a single cell (Buenrostro *et al.*, 2015). Interestingly, the relationship between the number of accessible peaks per cell and the unique fragments per cell followed a logistic, rather than linear, trend, approaching a plateau at roughly 35,000 accessible peaks per cell (**Figure 2.11C**). This indicates a saturation of accessible peaks detected per cell independent of single-cell sequencing depth.

Furthermore, allele-aware quantification of scATAC-seq data revealed that only 0.11 % of accessibility counts in TurboATAC data were bi-allelic (**Figure 2.12B**). Among the single cells, bi-allelic accessibility did increase only moderately with the number of unique fragments per cell (**Figure 2.12D**). These data suggest that chromatin accessibility is a stochastic process, exhibiting high variability among cells of the same type and state. As a result, the lower number of detected peaks in single cells compared to bulk or pseudo-bulk data likely reflects true biological variation between individual cells rather than purely derived from data sparsity. Nevertheless, we found that the sequencing of the

scTurboATAC data was not at saturation. Increasing sequencing depth further enhanced chromatin accessibility coverage, as seen in the scTurboATAC-seq data from HUVECs (**Figure 2.20B**).

Quantification and normalization strongly impact analyses of scATAC-seq data

Further analyses of the scATAC-seq data revealed distinct patterns of chromatin accessibility signals in pseudo-bulk ATAC peaks across various genomic regions. Peaks with high GC content exhibited significantly elevated Tn5 insertion counts per single cell (**Figure 2.13C**), consistent with a global association of GC-rich regions and accessible euchromatin (Bouwman *et al.*, 2023). Additionally, promoter peaks displayed higher insertion counts per single cell compared to exonic, intronic, and distal peaks (**Figure 2.13B**), indicating distinct patterns of accessibility in different genomic regions. Notably, bi-allelic accessibility was predominantly observed at promoter peaks (**Figure 2.14C**). These findings suggest that continuous and allelic quantification not only differentiate accessible from inaccessible regions in single cells but may also capture varying levels of accessibility. This aligns with previous studies that identified meaningful quantitative information in scATAC-seq data (Martens *et al.*, 2024; Miao & Kim, 2024). Considering the nucleosome positioning information inferred from ATAC-seq data (**Figure 1.5B**) (Buenrostro *et al.*, 2013), higher mono-allelic insertion counts per cell and peak likely reflect nucleosome-depleted chromatin regions. In contrast, lower insertion counts might indicate loosely packed, nucleosome-containing regions.

Despite the value of this quantitative information, it is often lost in downstream analysis, as most methods utilize binarized counts (Luo *et al.*, 2024). Binarization has also been suggested as a normalization technique to reduce technical variation between samples (Heumos *et al.*, 2023). However, the results presented here emphasize the need to preserve the quantitative information of scATAC-seq data, while reducing technical biases across samples and data sets. I selected equal cell numbers with the most comparable unique fragment numbers to minimize technical variation across samples within a data set (see **Section 2.2.1**, Compensation of biases between samples). Subsequently, I utilized continuous count matrices of these bias-compensated cells from multiple samples. In the future, the computational analysis of scATAC-seq data might be further improved by developing robust normalization methods that reduce biases among cells, samples, and data sets, while preserving the quantitative accessibility information.

3.2 Developing a computational framework to dissect the molecular mechanisms underlying chromatin co-accessibility

Several models have been proposed to explain how *distal CREs* might interact with promoters and TSSs to exert their regulatory effects (**Figure 1.2**) (Grosveld *et al.*, 2021; Karr *et al.*, 2022). On the one hand, targeted chromatin interactions via stable or transient chromatin contacts might directly transmit the regulatory TF signal from the *distal CRE* to the promoter (looping and kiss-and-run models). On the other hand, nuclear subcompartments, bringing both the promoter and *distal CRE* in spatial proximity, might facilitate fast diffusion of TF signal from the *distal CRE* to the promoter (proximity model). However, it remains unclear which exact molecular mechanisms, or a combination thereof, are present in eukaryotic nuclei (Ibrahim, 2024). I applied chromatin co-accessibility analysis to study these 3D interactions along the linear genomic coordinate and their potential regulatory mechanisms (**Figure 1.7D**).

Co-accessibility analysis of scATAC-seq data

In the past years, several methods have been developed to identify these co-accessible regions in single cells along the linear genomic coordinate (**Table 1.2**). The original scATAC-seq analysis by Buenrostro *et al.* (2015) observed an association between chromatin co-accessibility and higher-order chromatin organization. In this study, chromatin co-accessibility was inferred by correlating the combined accessibility in genomic windows of 25 peaks across single cells. In contrast, chromatin contacts, as determined by 3C methods, predicted global chromatin organization (Buenrostro *et al.*, 2015). In contrast, more recent approaches have predominantly applied co-accessibility analysis to link individual CREs with their target promoters (Pliner *et al.*, 2018; Mallm *et al.*, 2019; Granja *et al.*, 2021). These methods differ in their genomic and cellular resolution, heterogeneity of used cell populations, and the strategies used to calculate co-accessibility (**Table 1.2**). All approaches report co-accessible links between distal genomic regions as potential regulatory interactions, even though co-accessibility may also arise from cell type-specific accessible peaks (Shi *et al.*, 2022).

Indeed, an association between co-accessible links detected by Cicero and chromatin contacts inferred from 3C methods was reported (see Figure 4 in Pliner *et al.* (2018)). Additionally, the authors observed preferential links between promoters and regions marked by H3K27ac (see Figure 5 in Pliner *et al.* (2018)). Mallm *et al.* (2019) demonstrated that co-accessible link rewiring occurred independently of constitutively bound CTCF sites,

using the CTCF ChIP-seq signal to indicate chromatin loops mediated via CTCF and cohesin (see Figure 5G in Mallm *et al.* (2019)). In contrast, ArchR reports multiple co-accessible links, which appear to be dominated by cell type-specific accessible peaks rather than their distal interactions (see Figure 3I in Granja *et al.* (2021)). These findings underscore the importance of considering the implications of each method's design when interpreting the resulting co-accessible links (**Table 1.2**).

To address this need for a more precise determination of chromatin co-accessibility, I developed RWireX, a co-accessibility framework designed to resolve distinct layers of variation in single-cell accessibility signals (see **Section 2.2.1**). RWireX is based on RWire, that was initially developed in the Division of Chromatin Networks at the German Cancer Research Center in Heidelberg for the analysis of low-throughput scATAC-seq data acquired with the Fluidigm C1 microfluidics platform (Mallm *et al.*, 2019). While RWire efficiently computes co-accessible links for data sets up to 1,000 cells, the growing scale of single-cell data required an extension capable of handling larger datasets. This led to the development of RWireX as an extension of ArchR, leveraging ArchR's functionalities for efficiently storing large scATAC-seq data sets and computing Pearson correlations across large numbers of cells.

RWireX offers two different workflows of co-accessibility analysis. On the one hand, the *single cell co-accessibility workflow* implements and extends RWire's original concept by resolving stochastic variation in chromatin accessibility among homogeneous single cells. This workflow identifies autonomous links of co-accessibility, termed ACs, against a local background model and a lower detection rate threshold. The homogeneous cell population and single-cell resolution, applied in the *single cell co-accessibility workflow*, conceptually align with the design of a method for scRNA-seq data that derives stochastic variation in gene expression (Grun, 2020).

On the other hand, RWireX's *metacell co-accessibility workflow* captures chromatin accessibility variation among metacells from multiple cell types or states, similar to the original ArchR workflow. Unlike ArchR's co-accessibility method, which uses 1 kb ATAC peaks, I quantified the accessibility signal in larger 10 kb genomic tiles to resolve more global events of co-accessibility, similar to the 25-peak windows used by Buenrostro *et al.* (2015). Additionally, I developed a novel method for metacell aggregation. ArchR applies Cicero's metacell method, which allows each cell to contribute to multiple metacells (Pliner *et al.*, 2018). The maximum overlap of 80 % between distinct metacells inflates co-accessibility scores due to shared accessibility information. To mitigate this issue, my method forms unique metacells using fewer, non-overlapping cells with similar chromatin

accessibility profiles. This *metacell co-accessibility workflow* identifies broader patterns of depleted or enriched co-accessibility, such as the domains of contiguous co-accessibility, termed DCs. I found that the RWireX workflows resolve different layers of co-accessibility, likely arising from distinct molecular mechanisms, as discussed in the following sections.

ACs represent spatial chromatin contacts between active sites

RWireX's *single cell co-accessibility workflow* reveals two types of ACs based on their bimodal detection rate distribution across single cells: *Frequent ACs*, which are detected in more than 75 % of cells, and *rare ACs*, which are detected in fewer than 75 % (**Figure 2.24C**). These differences in detection rates likely reflect true variations in the prevalence of chromatin interactions, considering the high coverage of the HUVEC scTurboATAC-seq data set. It is unclear whether these differences arise from the frequency or stability of these chromatin interactions (or both), since scATAC-seq data only captures a snap-shot of cellular states. Interestingly, a previous study found that CTCF- and cohesin-mediated enhancer-promoter loops are rare and dynamic, with loop stability ranging from 10 to 30 min (Gabriele *et al.*, 2022). Additionally, Mach *et al.* (2022) observed that CTCF-anchored loops persist for only about 10 min, stabilizing highly dynamic chromatin environments with transient and frequent chromatin contacts within. Furthermore, these findings suggest that chromatin interactions indeed vary in both frequency and stability. They propose different chromatin interactions with structural chromatin loops, which are relatively rare and more stable, while other chromatin contacts are more transient and dynamic. This is in line with my observation of *frequent* and *rare ACs*.

Both types of ACs show significantly enriched chromatin contacts (**Figure 2.29**), though they exhibit distinct patterns of contact frequencies in their surrounding regions (**Figure 2.30A**). *Rare ACs* show a strong enrichment of chromatin contacts throughout their vicinity, suggesting they may emerge from stochastic interactions of active chromatin sites in a highly dynamic and transient environment (**Figure 3.1**, left) (Sood & Misteli, 2022; Bruckner *et al.*, 2023). In the context of promoter-*distal CRE* interaction models, these *rare ACs* may support the "kiss-and-run" model (**Figure 1.2E**) (Karr *et al.*, 2022). In contrast, *frequent ACs* show enriched chromatin contacts between the linked peaks but depleted contacts beyond them (**Figure 3.1**, right). This pattern indicates that *frequent ACs* might arise from architectural chromatin loops, which create distinct chromatin environments with enriched or reduced chromatin interactions within and outside their linked sites (Mach *et al.*, 2022; Chan & Rubinstein, 2023). In addition, these *frequent ACs* may also include stable TF-mediated chromatin contacts, as they are observed between promoters of actively transcribed genes and distal H3K27ac-marked sites (**Figures 2.26**, **2.27**). These

findings are consistent with the looping model of transcription regulation (**Figure 1.2D**) (Karr *et al.*, 2022).

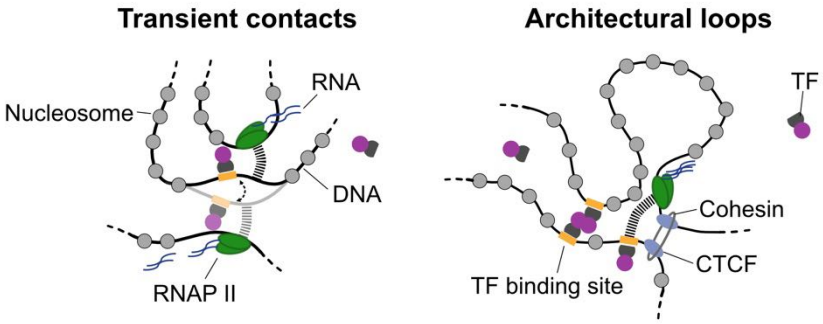


Figure 3.1 ACs resolve different types of chromatin contacts. *Rare ACs* likely reflect random and stochastic chromatin contacts. *Frequent ACs* potentially arise from targeted chromatin interactions, such as architectural chromatin loops or TF-mediated contacts.

Furthermore, some *frequent ACs* were also detected in the *metacell co-accessibility workflow*, identified by the so-called *blue stripes* of anti-correlated accessibility with high co-accessibility at their intersection (**Figure 2.30B**). These patterns have not been reported before in co-accessibility analysis. The *blue stripes* were located near gene promoters and predominantly at H3K27ac-marked sites, extending across multiple consecutively linked sites (**Figure 2.30C**). Remarkably, they coincided with *red stripes* of enriched chromatin contact frequencies in HiC-seq data. A recent study observed similar structures using optical reconstruction of chromatin architecture (ORCA) via fluorescence microscopy and proposed a loop stacking mechanism, where hubs of multiple CTCF-cohesin loops form (Hafner *et al.*, 2023). This loop stacking, later shown to regulate promoters near the stacked hubs by bringing them into proximity with *distal CREs* (Hung *et al.*, 2024), could explain an additional regulatory transcriptional mechanism of the *frequent ACs*.

Comparing the *single cell co-accessibility* results to existing co-accessibility methods suggests that *frequent ACs* correspond to the co-accessible links detected by Cicero, which also show enrichment of chromatin contacts and H3K27ac (Pliner *et al.*, 2018). However, since Cicero aggregates single cells into metacells, it likely does not capture the stochastic and transient interactions represented by *rare ACs*. Moreover, the co-accessible links driven by cell type-specific accessible peaks, as reported by ArchR, are absent in the homogeneous cell population of the *single cell co-accessibility workflow* and vice versa (Granja *et al.*, 2021).

Overall, RWireX's *single cell co-accessibility workflow* provides genome-wide insights into co-accessibility events predominantly driven by chromatin topology. Furthermore, it

differentiates between different mechanisms of chromatin contacts. However, further investigation is required to understand the molecular mechanisms underlying *frequent* and *rare ACs* fully. In this context, validation of ACs could be further improved by generating corresponding Micro-C sequencing data sets, which detect high-resolution chromatin contacts between accessible sites (Hsieh *et al.*, 2015). Additionally, DNA-FISH of AC-predicted peaks in single cells could explore the variability of AC detection rates (Mota *et al.*, 2022). Live-cell imaging as conducted by Gabriele *et al.* (2022) could clarify whether AC stability or frequency drives the observed differences in detection rates. Finally, inferring chromatin conformation in single cells from scHiC-seq data or ORCA would further reveal the local chromatin structure and its variability at predicted ACs (Mateo *et al.*, 2019; Rothorl *et al.*, 2023).

DCs are nuclear subcompartments of enriched TF binding activity

RWireX's *metacell co-accessibility workflow* identifies DCs, which show significantly altered accessibility in HUVECs following TNFα treatment (**Figure 2.31B**). These DCs have not been described before in co-accessibility analysis. I observed that they were independent of TADs, while being on average 10 times smaller (**Figure 2.32**). This suggests an additional layer of global chromatin organization independent of TADs and A/B compartments (Gholamalamdari *et al.*, 2024). To explore the molecular mechanisms underlying these DCs, I assessed their local TF binding activity in the pseudo-bulk scTurboATAC-seq data. This analysis revealed significant local enrichment of TF binding activity within DCs (**Figures 2.34**, **2.35**), suggesting that higher local concentrations of TFs within these nuclear subcompartments facilitate simultaneous, enriched binding to DNA (Garcia *et al.*, 2021; Mazzocca *et al.*, 2023; Mukherjee *et al.*, 2024). Although the TurboATAC data can predict TF enrichment within individual DCs, it does not reveal the molecular mechanisms potentially driving the formation of these nuclear subcompartments (**Figures 1.4C-E**). In the context of promoter-*distal CRE* interaction models, these DCs might support the proximity model, in which nuclear subcompartments surrounding promoters and *distal CREs* facilitate rapid diffusion of TF signals between them (**Figure 1.2C**) (Karr *et al.*, 2022). Additionally, a recent study proposed that condensates, distal super-enhancers, and gene loci follow a "three-way kissing" model, where all three components interact transiently (Du *et al.*, 2024).

A separate study introduced a variation of bulk ATAC-seq, the so-called assay for chromatin-associated condensate sequencing (ACC-seq), where nuclear condensates are crosslinked by fixation, limiting Tn5 access to DNA (He *et al.*, 2024). Here, identifying genome-wide condensates relies on differential accessibility patterns between three

different sequencing experiments of the same sample. He *et al.* (2024) identify condensates by comparing regions that show a signal in normal ATAC-seq and ACC-seq with prior 1,6-hexandiol treatment, but no signal in ACC-seq. Similar to DCs identified through RWireX, the ATAC-seq data in this study can also predict enriched TFs in these condensates. Additionally, the method determines whether a condensate is formed via phase separation from the ACC-seq after 1,6-hexandiol treatment. However, the ACC-seq approach lacks single-cell resolution and requires three distinct experimental readouts, increasing costs and complexity.

Overall, RWireX's *metacell co-accessibility workflow* provides genome-wide insights into co-accessible events dependent on the cell type or state, such as co-differential chromatin states or simultaneous TF binding. The workflow identifies broader domains of enriched co-accessibility, which likely represent nuclear subcompartments. Comparing these *metacell co-accessibility* results with existing co-accessibility methods reveals similarities between the broad DCs identified by RWireX and the broader clusters of co-accessible links detected by ArchR, which are driven by cell type-specific accessible peaks (Granja *et al.*, 2021). However, ArchR treats these as distinct regulatory links, overlooking their higher-order organization into local clusters. In contrast, the TAD-independent DCs identified by RWireX differ from the findings of Buenrostro *et al.* (2015), who reported an association between chromatin co-accessibility and global chromatin organization. While RWireX's *metacell co-accessibility* method may not directly infer global chromatin organization, it could provide insights into structural loops through the previously mentioned *blue stripes*. However, further studies are needed to validate the connection of co-accessibility patterns with nuclear subcompartments and global chromatin organization and to understand the molecular mechanisms behind these observations. In this regard, simultaneous DNA-FISH of DC-predicted regions and TF immunofluorescence could validate nuclear subcompartments at specific loci (Chaumeil *et al.*, 2013; Mota *et al.*, 2022). Additionally, sequencing-based methods such as region capture Micro-C sequencing or ACC-seq could be employed to confirm genome-wide DCs (Goel *et al.*, 2023; He *et al.*, 2024).

3.3 Identifying the structure-function relationship between regulatory mechanisms and their transcriptional output

RWireX identifies various mechanisms of distal transcription regulation from scATAC-seq data, such as transient and stable chromatin contacts, and nuclear subcompartments enriched in TF activity, as discussed in the previous sections. Additionally, proximal transcription regulation can be studied through chromatin accessibility at gene promoters. However, the distinct impact of these proximal and distal mechanisms on transcription has yet to be investigated and compared. In the following section, I discuss their diverse roles in transcription regulation across the three studied model systems under perturbation. These model systems comprise (i) two genetically identical mouse cell types, namely ESCs and MEFs, untreated and treated with IFNβ, (ii) the TCL1 mouse model for CLL with transcription factor T-bet wild-type or double knock-out, and (iii) HUVECs untreated and treated with TNFα.

STAT1/2 strongly activates transcription in mouse cells

The first model system studied here analyzed the upregulation of transcription upon IFNβ treatment in ESCs and MEFs (**Figure 2.38**). The results show that activated STAT TFs act as strong transcriptional activators, as a total of 191 and 463 ISGs were identified in ESCs and MEFs, respectively. This observation is consistent with previous studies, where 200 to 1,000 genes were upregulated by IFN treatment in various model systems (Der *et al.*, 1998; de Veer *et al.*, 2001; Mostafavi *et al.*, 2016). Notably, ESCs showed a weaker response at both time points examined, which aligns with earlier reports on an attenuated IFN response in stem cells (Wang *et al.*, 2013; Guo *et al.*, 2015). This also confirms the previously described complex temporal hierarchy of ISG expression in this model system (Bolen *et al.*, 2014). The transcriptional response was homogeneous at the single-cell level in ESCs and MEFs (**Figure 2.39**). This was in line with a previous study showing a homogeneous response to IFNβ treatment among fibroblasts from individual donors, although this response varied between donors carrying different genetic variants (Kumasaka *et al.*, 2023).

Simultaneous STAT1 and STAT2 binding at promoters led to the strongest upregulation of ISG expression (**Figure 2.41D**). This co-binding likely represents transcription regulation via the ISGF3 complex, formed with IRF9, which primarily controls antiviral ISGs (Stark & Darnell, 2012; Ivashkiv & Donlin, 2014; Platanitis *et al.*, 2019). Notably, IFNβ treatment did not induce global changes in chromatin accessibility but rather specific

increases at STAT1/2 binding sites (**Figure 2.42**). In the temporal hierarchy of ISG expression, simultaneous STAT1 and STAT2 binding at promoters triggered the fastest and most persistent expression response in ESCs and MEFs (**Figure 2.45A**). Supporting this, a recent study observed that chromatin accessibility contributes to the temporal control of ISG expression following IFNβ or IFNγ treatment in bone marrow-derived macrophages (Ravi Sundar Jose Geetha *et al.*, 2024).

More than 50 % of simultaneous STAT1 and STAT2 binding events occurred at *distal CREs* in intronic or intergenic regions (**Figure 2.41C**). These *distal CREs* were linked to ISG promoters by ACs (**Figure 2.43**), suggesting a mechanism of distal regulation via distinct chromatin contacts. This observation is consistent with previous reports on chromatin reorganization around ISG loci upon expression induction (Platanitis *et al.*, 2022). I found that distal regulation occurred between ISG promoters and intergenic *distal CREs* and also between ISG promoters themselves (**Figure 2.43C**). This suggests that CREs may exert both proximal and distal regulation simultaneously. This finding is supported by previous research showing that ISG promoters can function as enhancers to drive ISG expression (Santiago-Algarra *et al.*, 2021).

Interestingly, *distal CREs* exhibited two modes of transcription induction upon co-binding of STAT1/2: either through a gain of activating ACs or the loss of repressive ACs (**Figures 2.43**, **2.45B**). A similar mechanism of gene silencing through chromatin contacts with distal inhibitory CREs has been observed in mouse erythroblasts (Vermunt *et al.*, 2023). ISGs regulated by chromatin contacts with *distal CREs* exhibited a slower temporal response to IFNβ treatment compared to ISGs regulated directly at their promoters (**Figure 2.45A**). Additionally, *distal CREs* with activating ACs induced a stronger ISG expression response than those with repressive ACs. This suggests that activating ACs have a more direct effect on transcription induction, while removing inhibitory ACs may require additional TFs to fully activate the transcription machinery.

Furthermore, *metacell co-accessibility* analysis identified nuclear subcompartments at ISG loci following IFNβ treatment in ESCs and MEFs (**Figure 2.46**), despite the relative data sparsity in the scATAC-seq data set of this study. These nuclear subcompartments did not exhibit increased local STAT1 and STAT2 binding activity after IFNβ treatment. Instead, they may be driven by secondary TF targets at later treatment time points, explaining the delayed temporal induction of ISG expression. Consistent with this, phase separation has not been reported so far for STAT1 and STAT2. However, STAT3 – another member of the same TF family – has been shown to enter and concentrate within mediator condensates at super-enhancers via its IDR (Zamudio *et al.*, 2019). STAT3 is associated

with different signaling pathways compared to STAT1 and STAT2, primarily involving pro-survival and oncogenic functions (Wang *et al.*, 2023). However, all STATs contain a structurally similar IDR in their transactivation domains (Levy & Darnell, 2002), suggesting that STAT1 and STAT2 may also accumulate in preexisting condensates.

T-bet suppresses transcription and reduces tumor proliferation in CLL

The second application of the RWireX framework dissected the association of high T-bet expression with an increased overall survival in patients with CLL (**Figure 1.10**) (Roessner *et al.*, 2024). In both CLL and the TCL1 mouse model of CLL, the transcription factor T-bet functions as a transcriptional repressor (**Figures 2.50, 2.53**). This observation aligns with previous studies reporting T-bet as a repressor of gene expression in B and T cells (Oestreich & Weinmann, 2012; Stone *et al.*, 2019). Based on a scATAC-seq (co-)accessibility analysis, I found that T-bet's repressive effects are linked to enhanced IFN signaling and decreased activity of POU family TFs (**Figure 2.50C**). Interestingly, IFN signaling has been associated with growth arrest in low-risk CLL patients (Tomic *et al.*, 2011), a group likely to exhibit elevated T-bet activity. Conversely, POU family TFs promote cell survival, cell cycle progression and proliferation (Hodson *et al.*, 2016; Lu *et al.*, 2019). This suggests that T-bet suppresses tumor growth by enhancing IFN signaling while simultaneously reducing POU family activity (**Figure 2.48**), explaining its tumor-suppressive role in CLL.

The knock-out of T-bet in the TCL1 mouse model does not allow for studying the temporal hierarchy of transcription regulation in this model system. However, it does enable investigation of the diverse molecular mechanisms governing transcription regulation following the internal perturbation of a specific TF. I found that knocking out T-bet disrupts transcriptional regulation through several proximal and distal mechanisms, including promoter regulation, long-range chromatin contacts, and nuclear subcompartment formation (**Figures 2.52**, **2.54**). These findings are in line with previous studies suggesting that T-bet regulates transcription repression through multiple direct and indirect mechanisms (Oestreich & Weinmann, 2012). Notably, T-bet binding itself appears to inhibit the formation of chromatin contacts with *distal CREs*. This disruption affected both transient and architectural chromatin contacts, as identified by *rare* and *frequent ACs*. Additionally, T-bet's repressive effects were associated with reduced activity of both CTCF and BORIS (**Figure 2.50C**), with BORIS being a paralog of CTCF (Klenova *et al.*, 2002). These findings align with prior research showing that T-bet and CTCF jointly regulate chromatin accessibility, distal contacts, and broader chromatin reorganization in T cells (Sekimata *et al.*, 2009; Liu *et al.*, 2023). However, further validation of these observations

is necessary, potentially using 3C-based methods to investigate T-bet-dependent chromatin reorganization.

TNFα induces transcriptional co-regulation of proximal genes in human cells

Finally, I studied transcription regulation during the TNFα-mediated proinflammatory response in HUVECs, applying the RWireX framework for co-accessibility analysis. TNFα induced a fast increase in activity of NF-κB family TFs within 30 min of treatment (**Figure 2.57**). Subsequently, NF-κB activated secondary TFs, such as PRDM1, ATF4, and members of the IRF and CEBP families, which showed higher activity after 240 min of TNFα treatment. Notably, early NF-κB activity predominantly led to gene upregulation (**Figure 2.56**). However, at the later time point, both upregulation and downregulation of gene expression was present, suggesting that secondary TFs exert more varied regulatory effects. These findings are consistent with numerous studies that report NF-κB as an activator of gene expression (Liu *et al.*, 2017). In contrast, IRF and CEBP family members have been shown to either activate or repress gene expression (Ramji & Foka, 2002; Zhao *et al.*, 2015). Moreover, PRDM1 functions as a master transcriptional repressor (Ren *et al.*, 1999), while the activating transcription factor 4, ATF4, is primarily known as an activator (Neill & Masson, 2023). Overall, these results suggest that the regulatory direction, whether gene expression is induced or repressed, is largely determined by the specific TF involved.

These TNFα-regulated genes, referred to as TRGs, were found to cluster within the genome (**Figure 2.58**), hinting at regulatory mechanisms beyond global TF activity that co-regulate multiple genes in close proximity. This observation aligns with previous studies showing that chromatin reorganization coincides with local co-induction of genes following TNFα or interleukin 1 alpha treatment in HUVECs (Diermeier *et al.*, 2014; Weiterer *et al.*, 2020). TRGs within these clusters exhibited significantly higher co-expression in single cells after TNFα treatment compared to isolated TRGs (**Figure 2.59**). This finding is supported by another study that identified local co-expression as a general principle for regulating functionally related genes, which share distal regulatory elements (Ribeiro *et al.*, 2022).

In addition to promoter-driven transcriptional regulation, I found that TFs induced by TNFα regulate transcription through long-range chromatin contacts with *distal CREs*, as well as the formation of nuclear subcompartments, as shown by co-accessibility analysis using RWireX (**Figure 2.61**). Both ACs and DCs were detected throughout the genome. TRGs were regulated by ACs, DCs, or a combination of both (AC/DC). Interestingly, ACs exhibited three modes of TRG regulation upon TNFα treatment: pre-existing, emerging, or

dissolving contacts. A previous study also demonstrated that pre-existing chromatin contacts between promoters and enhancers are associated with strong gene induction upon TNFα treatment (Jin *et al.*, 2013). Additionally, emerging chromatin contacts have been observed at TNFα-induced genes (Papantonis *et al.*, 2012). Finally, dissolving chromatin contacts were involved in both the upregulation and downregulation of TRGs, confirming the existence of both activating and inhibitory chromatin interactions. I had previously observed these inhibitory interactions in mouse cells treated with IFNβ, and others have also shown them during mouse erythroblast differentiation (Vermunt *et al.*, 2023).

Almost all DCs exhibited significantly altered accessibility following TNFα treatment (**Figure 2.31**), suggesting that TNFα induces both emerging and dissolving nuclear subcompartments. Notably, around 75 % of DCs associated with TRGs showed increased activity, despite most DCs showing reduced activity overall (more than 70 %). While most DC-regulated TRGs were upregulated, the reduction of DC activity could readily explain the downregulation of some TRGs. This observation aligns with a recent finding on forming liquid-liquid phase-separated NF-κB condensates at super-enhancers, which induce transcription following anti-IgM stimulation in B cells (Wibisana *et al.*, 2022). Interestingly, DCs were more frequently involved in regulating lncRNA TRGs compared to promoter- and AC-driven regulation (**Figure 2.61**). This is consistent with a recent study that describes the role of lncRNAs in promoting condensate formation at proximal genes and fine-tuning transcriptional activity at these genes (Natarajan *et al.*, 2023).

The transcriptional changes of TRGs showed varying temporal dynamics for promoter-, DC-, AC-, or AC/DC-mediated regulation. Promoter-regulated TRGs exhibited the fastest, yet most sustained transcriptional responses (**Figure 2.67**). Additionally, promoters induced the strongest response, showing almost exclusively upregulation. In contrast, TRGs regulated by ACs displayed a significantly slower and less strong transcriptional response. In contrast, DC- and AC/DC-regulated TRGs responded equally fast but with significantly lower transcriptional activity.

Interestingly, these regulatory mechanisms also influenced varying bursting kinetics of TRG transcription (**Figures 2.65**, **2.67**). AC-regulated TRGs displayed higher burst frequencies, indicating that transcriptional bursts are initiated when promoters get in contact with *distal CREs*. This finding corroborates earlier reports that enhancer-promoter interactions preferentially regulate burst frequency (Larsson *et al.*, 2019; Wang *et al.*, 2024). Furthermore, Han *et al.* (2024) used cryo-electron microscopy to reveal the molecular mechanism of transcription activation via DNA looping between a promoter and

a distal upstream element in bacteria. In contrast, promoter- and DC-regulated TRGs were primarily modulated by burst size, suggesting that burst duration and production rate may be increased due to higher TF occupancy and RNAP II availability. Accordingly, previous studies have shown that TF binding sites at promoters are key determinants of burst size (Larsson *et al.*, 2019). Additionally, local confinement of TF mobility caused increased TF occupancy at promoters, resulting in prolonged burst duration (Stavreva *et al.*, 2019; Pomp *et al.*, 2024). Lastly, I found that AC/DC-driven TRGs exhibited regulation through both burst frequency and size, implying that the combination of regulatory mechanisms further extends the complexity of transcriptional regulation. Interestingly, a recent study observed burst frequency and size increases upon three-way proximity of a gene, distal super-enhancer and transcriptional condensate (Du *et al.*, 2024). However, the authors conclude that burst frequency and size are mainly enhanced by condensate proximity, although stating its dependence on *distal CREs* by cohesin-mediated chromatin loops.

TRGs in local clusters exhibited enriched co-regulation via AC-, DC-, or AC/DC-mediated mechanisms (**Figure 2.62**). Interestingly, AC-regulated *TRG clusters* displayed anticorrelated expression patterns, which could suggest sequential chromatin contacts between a *distal CRE* and multiple TRG promoters. This finding contradicts previous studies that proposed the formation of so-called NF-κB factories where numerous genes are co-induced through simultaneous contact with a shared transcription factory (Papantonis *et al.*, 2012). In contrast, DCs promoted co-expression in *TRG clusters*, supporting the general assumption that nuclear subcompartments drive the co-induction of gene expression (Ryu *et al.*, 2024). Notably, these opposite effects of AC and DC regulation on co-expression explain the bi-modal distribution of co-expression in *TRG clusters* (**Figure 2.59**).

3.4. The AC/DC model of transcription regulation

In this thesis, I advanced the experimental and computational analysis of scATAC-seq data, developed a computational framework to dissect the molecular mechanisms underlying chromatin co-accessibility, and identified the structure-function relationship between different regulatory mechanisms and their transcriptional output. Building on these findings, I propose the AC/DC model of transcription regulation, which integrates genome-wide information on chromatin topology-mediated and TF-mediated regulatory mechanisms. The AC/DC model describes how a multilayered network of soluble TFs and

various regulatory mechanisms at proximal and *distal CREs* regulates transcription. Initially, I outlined five criteria that characterize a transcriptional response to external or internal stimuli (**Section 1.3**). I found that different layers of the regulatory network determine varying parameters of transcriptional responses. Additionally, I identified that various combinations of TFs and multiple regulatory mechanisms further enhance the versatility of transcriptional responses.

The transcriptional response is shaped by the specific TF involved and the regulatory mechanism in action, including regulation via proximal promoters, architectural or stochastic chromatin contacts, and nuclear subcompartments (**Figure 3.1**). The specific TF primarily determines whether gene expression is induced or repressed, the so-called regulatory direction. The TF's effect can be mediated through direct promoter binding, binding to a *distal CRE*, or enrichment within a nuclear subcompartment. Importantly, the regulatory direction of a TF can be additionally modulated by the dynamics of the regulatory mechanism, which can be pre-existing, emerging, or dissolving in response to perturbation.

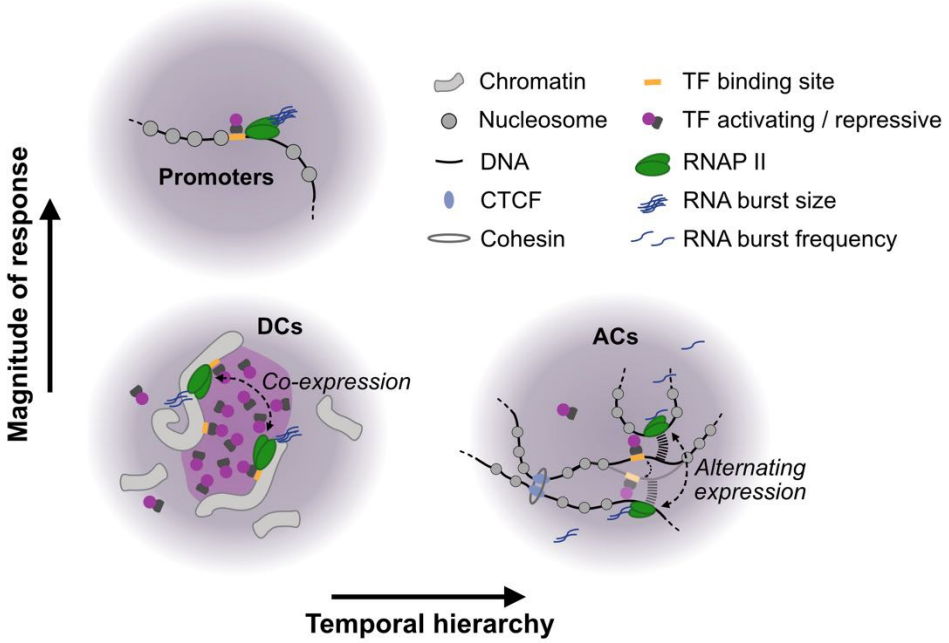


Figure 3.2 Regulatory mechanisms in the AC/DC model of transcription regulation. Mechanisms are arranged by their magnitude and temporal hierarchy of the transcriptional of response. Induced transcriptional bursting kinetics are indicated. For ACs and DCs, co-regulation potential is depicted.

Beyond the direction of the transcriptional response, the regulatory mechanisms also shape the magnitude and temporal hierarchy of the response (**Table 3.1**). Regulation via proximal promoters leads to the strongest, fastest, yet sustained transcriptional response.

Similarly, nuclear subcompartments induce a rapid response, but with a lower magnitude of transcriptional change. In contrast, chromatin contacts mediate significantly slower transcriptional responses compared to promoters and nuclear subcompartments. However, the magnitude of the transcriptional changes is comparable to that of nuclear subcompartments.

Proximal promoters have a limited capacity for co-regulation, only simultaneously inducing genes with a shared promoter. However, promoters can also function as *distal CREs* for other genes when additional regulatory mechanisms, such as long-range chromatin contacts or nuclear subcompartment formation, are involved. Both chromatin contacts or nuclear subcompartments have a high potential for co-regulating transcriptional responses across multiple genes. Chromatin contacts induce alternating gene expression, while nuclear subcompartments promote simultaneous co-expression. Furthermore, all three regulatory mechanisms affect different aspects of transcriptional bursting kinetics. Regulation via proximal promoters and nuclear subcompartments primarily modulates burst size, while chromatin contacts predominantly influence burst frequency. Consequently, various combinations of these regulatory mechanisms can modulate both burst size and frequency, further enhancing the flexibility of transcriptional regulation.

Table 3.1 Regulatory mechanisms in the AC/DC model of transcription regulation. Transcriptional response parameters induced by different regulatory mechanisms.

	Proximal promoter	**Distal chromatin contact**	**Local nuclear subcompartment**
Inferred accessibility feature	Differential promoter peak	AC	DC
Mechanism of regulation	Burst size	Burst frequency	Burst size
Magnitude of regulation	Strong	Medium	Medium
Temporal hierarchy	Fast, persistent	Slow	Fast, persistent
Co-regulation potential	Low	High (alternating expression)	High (simultaneous co-expression)

The AC/DC model depicted above could be extended by incorporating additional regulatory layers, such as the presence of RNAP II and the transcription machinery (Malik & Roeder, 2023), more detailed local chromatin states (Jenuwein & Allis, 2001), or mechanisms of post-transcriptional regulation (Carpenter *et al.*, 2014). Moreover, targeted knock-out experiments of TFs, proximal promoters, or *distal CREs* could further validate the identified regulatory layers (Metzner *et al.*, 2024). The role of chromatin contacts and

nuclear subcompartments could be confirmed by knocking out critical structural proteins such as CTCF or cohesion, disrupting architectural chromatin contacts, or through 1,6-hexanediol treatment, which perturbs phase-separated condensates. Finally, multiplexed fluorescence microscopy approaches combining DNA-FISH, nascent RNA-FISH, and TF immunofluorescence could simultaneously map chromatin contacts, nuclear subcompartments, and their respective transcriptional outputs at subnuclear resolution in single cells (Chaumeil *et al.*, 2013; Kishi *et al.*, 2019; Mota *et al.*, 2022).

The AC/DC model in inflammation, differentiation, and cancer

I identified promoter-, AC-, and DC-mediated transcription regulation in all the studied model systems, suggesting that the proposed AC/DC model of transcription regulation may represent general principles of mammalian transcription regulation. In inflammatory signaling, the primary transcriptional responses to external stimuli are often driven by a limited number of specific TFs (Lawrence, 2009; Ivashkiv & Donlin, 2014). Furthermore, the temporal hierarchy of the transcriptional response is tightly regulated (Bhatt *et al.*, 2012; Bolen *et al.*, 2014). The AC/DC model addresses this complexity by describing different response dynamics for regulatory mechanisms (**Table 3.1**). I found that these temporal characteristics are consistent across the inflammatory stimuli studied, IFNβ and TNFα, in both mouse and human primary cells (Muckenhuber *et al.*, 2023; Seufert *et al.*, 2024).

The AC/DC model also holds for different stages of mouse cell differentiation, when investigating transcription regulation in embryonic stem cells and embryonic fibroblasts at various stages of the epithelial-to-mesenchymal transition. Interestingly, while the differentiation stages exhibited distinct primary TF responses to external IFNβ perturbation, they shared a common set of regulatory mechanisms through proximal promoters, long-range chromatin contacts, and nuclear subcompartments. Additionally, less differentiated cells demonstrated a lower transcriptional response, suggesting that changes in local chromatin states during differentiation potentially influence the likelihood of transcriptional responses (Muckenhuber *et al.*, 2023).

Lastly, the AC/DC model also is applicable to transcription deregulation in cancer, where accumulating mutations progressively perturb the transcriptomic profile. In the TCL1 mouse model for CLL, I found that the transcriptional response to the knock-out of the TF T-bet was mediated through proximal promoters, long-range chromatin contacts, and nuclear subcompartments (Roessner *et al.*, 2024). We also showed that rewiring chromatin contacts during MM subclone evolution enhanced the expression of genes encoding known drug-resistance proteins (Poos *et al.*, 2023). Furthermore, all the

regulatory mechanisms described, including TF activity, chromatin architecture, enhancer-promoter chromatin contacts, and nuclear condensates, have been implicated in cancer emergence and progression (Vishnoi *et al.*, 2020; Suzuki & Onimaru, 2022; Wang *et al.*, 2022; Perlman *et al.*, 2024). Cancer is commonly characterized by many genomic alterations, making temporal regulatory hierarchies challenging to discern. However, the AC/DC model of transcription regulation provides a robust framework for explaining the multi-layered deregulation of transcription in cancer.

3.5. Conclusion

The findings of this thesis provide a novel, integrative approach to transcription regulation, bridging two previously separate fields of research: chromatin topology-mediated and TF-mediated regulation. Through the development and application of new experimental and computational methods, I have demonstrated the co-existence and complex yet coordinated interplay of these regulatory mechanisms on a genome-wide scale across diverse biological systems. The proposed AC/DC model of transcription regulation elucidates the regulatory principles of multiple critical elements in transcriptional control: regulatory direction, magnitude, temporal hierarchy, and local co-regulation potential. This multilayered network comprises information on TFs and various proximal and distal regulatory mechanisms. It precisely governs transcriptional responses to external and internal stimuli at individual genomic loci.

The experimental and computational methods developed in this thesis provide new tools to enhance genome-wide studies of transcription regulation. The scTurboATAC-seq protocol addresses a major challenge in scATAC-seq studies by significantly reducing data sparsity without increasing costs, allowing to distinguish biological variation from technical noise more reliably. The scTurboATAC-seq data can be leveraged to study proximal and distal mechanisms of transcription regulation using my developed co-accessibility framework RWireX. It resolves the molecular mechanisms underlying different layers of co-accessibility, providing a unique tool for analyzing transcriptional regulation. The experimental and computational advancements in scATAC-seq analysis will empower researchers to explore complex transcriptional landscapes across various biological systems. It will enable them to simultaneously investigate both chromatin topology and TF activity in a unified framework and comprehensively investigate their impact on transcription.

My findings from various human and mouse systems upon external or internal perturbations suggest that the AC/DC model of transcription regulation is highly versatile. It applies for inflammatory transcriptional programs, different stages of cell differentiation, and genetic aberrations in cancer across multiple human and mouse model systems. The experimental and computational analyses framework could be readily utilized to study various biological processes, such as developmental programs, disease progression, or drug response in different cell types and organisms. Future work could explore whether the discovered principles of transcription regulation can be generally extended to other biological systems. Furthermore, pooled genetic perturbation screens using for example Perturb-seq or Perturb-ATAC (Dixit *et al.*, 2016; Rubin *et al.*, 2019) could be used to test the applicability of the AC/DC model to the transcriptional responses to a multitude of different perturbations. These findings might expand the utility of the AC/DC model as a general concept of complex transcriptional control dynamics.

The developed approach could be extended by additional single-cell sequencing data sets, such as scChIP-seq of specific TFs or histone modifications (Barcenas-Walls *et al.*, 2024), scHiC-seq for chromatin contacts and the prediction of chromatin conformation (Rothorl *et al.*, 2023), or single-cell global run-on sequencing (GRO-seq) to directly measure nascent RNA (Mahat *et al.*, 2024). These additional data could improve the comprehensive understanding of the regulatory mechanisms in the AC/DC model, providing information on transient and stochastic processes by their single-cell resolution snap-shots. Additionally, they might explain the observed specificity of the regulatory mechanisms, identifying the distinct molecular environments which determine the locus-specific regulatory mechanism, such as specific chromatin states, higher-order chromatin organization, genomic sequence, or TF availability. Furthermore, integrated spatial omics approaches could further elucidate the spatial and temporal interplay of chromatin accessibility, TF localization, and transcriptional output (Kishi *et al.*, 2019; Dang *et al.*, 2023).

Looking forward, the findings and methodologies presented here hold significant promise for applications beyond basic research: The AC/DC model’s capacity to predict transcriptional responses upon perturbation could be applied to dissect the transcriptional impact of complex mutational profiles in cancer. The improved understanding of these deregulated transcriptional profiles might be used for drug discovery, targeting the expression of specific TFs or other genes. Additionally, the AC/DC model could be used to design targeted manipulation of specific regulatory modules in precision medicine, offering opportunities to engineer transcriptional programs for therapeutic purposes.

4. Materials and Methods

4.1. List of data sets

In this thesis, multiple biological systems with perturbed transcriptional states were investigated. Diverse experimental methods were applied to these systems, providing comprehensive data sets from single-cell and bulk sequencing, spatial transcriptomics, and mass spectrometry. An overview of the data sets used in this thesis and references to their respective Results sections is provided in **Table 4.1**.

Table 4.1 Overview of data sets used in this thesis. Organism, biological sample, observed perturbation of transcription, applied method, number of biological replicates, and respective Results section are indicated for all data sets. The number of biological replicates per perturbation condition is provided, except for data sets with *, which indicates the number of total samples.

Organism	Sample	Perturbation	Method	Rep. num.	Results section
Mouse	MEFs	6 h IFNβ treatment	scATAC-seq	2	2.1.1-5, 2.2.1
			scTurboATAC-seq	2	2.1.2-5, 2.2.1
			Bulk ATAC-seq	2	2.1.3
Human	PBMCs	Cell types	scATAC-seq	1*	2.1.2, 2.1.5
			scTurboATAC-seq	1*	2.1.2, 2.1.5
			Multiome scRNA-seq + scATAC-seq	1*	2.1.2
			Multiome scRNA-seq + scTurboATAC-seq	2*	2.1.2
Human	AML patients	Genomic aberrations	scRNA-seq	5*	2.1.5

Human	HUVECs	0 min, 30 min, 240 min TNFα treatment	scTurboATAC-seq	3	2.2.1-3, 2.3.3
			scRNA-seq	3	2.3.3
			snRNA-seq	1	2.3.3
			Spatial transcriptomics	2/3	2.3.3
		30 min TNFα treatment	Bulk ChIP-seq H3K27ac	2	2.2.2-3, 2.3.3
		-	Bulk HiC-seq	1	2.2.3, 2.3.3
Human	NK cells	HDV infection, no infection	Bulk RNA-seq	5	2.3.1
Mouse	ESCs	0 h, 1 h, 6 h IFNβ treatment	Bulk RNA-seq	4	2.3.1
			Bulk ChIP-seq STAT1 + STAT2	4	2.3.1
			scRNA-seq	1	2.3.1
		0 h, 6 h IFNβ treatment	scATAC-seq	1	2.3.1
	MEFs	0 h, 1 h, 6 h IFNβ treatment	Bulk RNA-seq	2	2.3.1
			Bulk ChIP-seq STAT1 + STAT2	2	2.3.1
			scRNA-seq	1	2.3.1
			scATAC-seq	1	2.3.1
Mouse	TCL1 cells	*Tbx21* knock-out, *Tbx21* wild type	scRNA-seq	2	2.3.2
			scTurboATAC-seq	2	2.3.2
			Bulk RNA-seq	6	2.3.2
			Mass spectrometry	8	2.3.2
			Phospho-specific mass spectrometry	8	2.3.2
Human	CLL patients	High *TBX21* / T-bet levels, low *TBX21* / T-bet levels	Bulk RNA-seq	260*	2.3.2
			Bulk ATAC-seq	99*	2.3.2
			Mass spectrometry	68*	2.3.2
		Malignant, non-malignant	Bulk RNA-seq	7/15*	2.3.2

4.2. scATAC-seq and scTurboATAC-seq of MEFs and PBMCs

4.2.1. Sequencing data acquisition

The scTurboATAC protocol was developed by Jan-Philipp Mallm (Single Cell Open Lab, German Cancer Research Center, Germany) as described in Seufert *et al.* (2023). In short, purified Tn5 from the EMBL Protein Expression and Purification Core Facility (European Molecular Biotechnology Laboratory, Heidelberg, Germany) was used (Hennig *et al.*, 2018). Tn5 molecules were assembled with one read 1 and read 2 NGS adapter each. NGS adapters varied for the single-omic and multi-omic ATAC reactions (**Table 4.2**). Both read 1 and read 2 NGS adapters were annealed to a mosaic end (ME) sequence, containing a Phos-ME-phosphorothioate (Phos-ME-PTO) backbone to avoid self-dimerization. Subsequently, Tn5 was loaded with read 1 and read 2 NGS adapters by incubation for 30 min at room temperature. This in-house loaded Tn5 (Tn5-H) was used for the TurboATAC protocol.

Table 4.2 Oligonucleotide sequences of NGS adapters used for Tn5 loading. Modifications are indicated as phos, 5′-phosphate; PTO, phosphorothioate backbone indicated by the *; 23ddC, 2′,3′-dideoxycytidine; ME, mosaic end sequence. Adapted from Seufert *et al.* (2023).

	Single-omic ATAC	**Multi-omic ATAC**
Read 1	5′-TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG-3′	5′-[phos]TCGTCGGCAGCGTCA GATGTGTATAAGAGACAG-3′
Read 2	5′-[phos]GTCTCGTGGGCTCGG AGATGTGTATAAGAGACAG-3′	5′-GTCTCGTGGGCTCGGAG ATGTGTATAAGAGACAG-3′
Phos-ME-PTO	5′-[phos]C*T*G*T*C*T*C*T* T*A*T*A*C*A*[23ddC]-3′	5′-[phos]C*T*G*T*C*T*C*T*T* A*T*A*C*A*C*A*T*C*T-3′

Different Tn5 preparations were incubated with 10 ng lambda DNA in standard tagmentation buffer (Tag buffer) or tagmentation buffer from 10x Genomics (TXG buffer; 10x Genomics, Pleasanton, USA) for 10 min at 55 °C. Identical volumes of Tn5-H at different concentrations, 10x Genomics Tn5 (Tn5-TXG; 10x Genomics, Pleasanton, USA), and Illumina TDE1 enzyme (Tn5-ILMN; Illumina, San Diego, USA) were used. Tagmentation was stopped and tagmented lambda DNA was measured by qPCR on a StepOnePlus machine (Applied Biosystems, Waltham, USA). Ct values were used to obtain relative activities of the different Tn5 preparations.

Experiments and data acquisition were performed by Jan-Philipp Mallm and Katharina Bauer (both Single Cell Open Lab, German Cancer Research Center, Germany) as described here and in Seufert *et al.* (2023). MEFs were cultured and treated with IFNβ as described earlier (Muckenhuber *et al.*, 2023). Human PBMCs were isolated and stored viably frozen in DMSO containing serum as described earlier (Mallm *et al.*, 2019). Libraries were prepared using the Chromium Single Cell ATAC v1.1 kit for MEFs and the Chromium Single Cell ATAC v2.0 and Chromium Single Cell Multiome ATAC + Gene Expression v1.0 kits for PBMCs according to the manufacturer's protocol (10x Genomics, Pleasanton, USA). In each experiment, 10,000 nuclei were used for the Tn5 reaction with standard 10x Genomics Tn5 or Tn5-H and loaded on to the Chromium Next GEM Chip H (PN-1000161). Library concentrations were measured, and fragment size distributions were determined. ATAC libraries were sequenced paired end with 50 bp for read 1 and read 2 each on a NovaSeq 6000 system (Illumina, San Diego, USA) with a sequencing depth of at least 25 k read pairs per nucleus. Single-omic ATAC libraries were sequenced with 8 bp for index 7 and 16 bp for index 5. Multi-omic ATAC libraries were sequenced with 8 bp for index 7 and 24 bp for index 5. Multiome RNA libraries were sequenced paired end with 28 bp and 90 bp for read 1 and read 2 on a NovaSeq 6000 system (Illumina, San Diego, USA). Sequencing was conducted by the DKFZ NGS Core Facility (German Cancer Research Center, Heidelberg, Germany).

4.2.2. Analysis of scATAC-seq and scTurboATAC-seq data from MEFs

I performed the analysis of scATAC-seq and scTurboATAC-seq data from 6 h IFNβ-treated MEFs as described here and for most parts in Seufert *et al.* (2023). ATAC data were processed with Cell Ranger ATAC count (10x Genomics, Pleasanton, USA) using the provided mouse mm10 reference. Quality metrics are provided in **Tables 2.2** and **2.5**. Single-cell data was further analyzed with ArchR and visualized using ggplot2 in R. Empty barcodes were removed using (i) a minimal threshold for number of unique fragments of $10^{3.1}$ for ATAC Rep1 with Tn5-TXGv1.1, $10^{3.8}$ for ATAC Rep2 with Tn5-TXGv1.1, $10^{3.9}$ for ATAC with Tn5-H30, and $10^{4.3}$ for ATAC with Tn5-H100, and (ii) a minimal threshold for TSS enrichment score of 5 for ATAC replicates with Tn5-TXGv1.1 and 12 for ATAC with Tn5-H30 and Tn5-H100. Cell numbers are provided in **Table 2.3** and **2.6**. Barcodes containing multiple cells were removed using Amulet from scDblFinder. Single cells from each experiment were embedded in two-dimensional space separately using an accessibility count matrix in 500 bp genomic tiles, iterative LSI (default parameters, except

clustering resolution of 0.2) and UMAP (default parameters, except LSI components 2-10 for ATAC replicates with Tn5-TXGv1.1 and 2-12 for ATAC with Tn5-H100, minimal distance of points in embedding of 0.5, 30 nearest neighbors). Cell clusters were computed by shared nearest neighbor (SNN) modularity optimization (default parameters, except LSI components 2-10 for ATAC replicates with Tn5-TXGv1.1 and 2-12 for ATAC with Tn5-H100, resolution of 0.1).

ATAC samples with Tn5-H30 and Tn5-H100 were downsampled to equal read numbers compared to ATAC Rep1 with Tn5-TXGv1.1 using Cell Ranger ATAC count (10x Genomics, Pleasanton, USA) with a subsampling rate of 0.6771 for Tn5-H30 and 0.6343 for Tn5-H100 (**Table 2.5**). For downsampled Tn5-H100, empty barcodes were removed using (i) a minimal threshold for number of unique fragments of $10^{4.3}$, and (ii) a minimal threshold for TSS enrichment score of 12. This yielded 5,550 cell barcodes for downsampled ATAC with Tn5-H100. Barcodes containing multiple cells were removed using Amulet from scDblFinder. Single cells were embedded in two-dimensional space using an accessibility count matrix in 500 bp genomic tiles, iterative LSI (default parameters, except clustering resolution of 0.2) and UMAP (default parameters, except LSI components 2-10, minimal distance of points in embedding of 0.5, 30 nearest neighbors). Cell clusters were computed by SNN modularity optimization (default parameters, except LSI components 2-10, resolution of 0.15).

Module scores of apoptosis marker gene activities were calculated for each cell and (Galluzzi *et al.*, 2018). Cells clusters with high apoptosis module scores were removed from further analysis (C1 and C2 from ATAC Rep2 with Tn5-TXGv1.1, C1 from ATAC with Tn5-H100). Sample-specific peaks from pseudo-bulk chromatin accessibility data were called using MACS2 in ArchR (default parameters, except peak summit extension by 1000 bp to each side; reproducibility of 1). Sample-specific peak sets from ATAC Rep2 with Tn5-TXGv1.1, ATAC with Tn5-H30, and ATAC with Tn5-H100 were merged. Transcription factor binding sites in sample-specific peaks were predicted using Homer mm10 motifs from chromVARmotifs. TF footprints at genome-wide STAT1 and CTCF binding sites were calculated for ATAC Rep2 with Tn5-TXGv1.1 and ATAC with Tn5-H100 using ArchR (default parameters, except smoothing window of 5 bp; no Tn5 bias normalization).

Different accessibility count matrices of single cells in merged peaks were generated for ATAC Rep2 with Tn5-TXGv1.1 and ATAC with Tn5-H100. Accessibility signal per single cell and ATAC peak was quantified by insertion-based counting using ArchR (no binarization, maximum counts of 100). The resulting continuous count matrices were binarized to obtain binary counts. Allele counts were inferred using the binary count

matrices and overlapping fragments per cell identified by Amulet. Accessible binary counts that showed overlapping fragments in the respective cell and peak were assigned as bi-allelic accessibility (2 count), while accessible binary counts without overlapping fragments were considered mono-allelic accessibility (1 count). Ternary plots were generated using ggtern. *Single cell co-accessibility* analysis was performed with RWireX using cells from homogeneous clusters C3 and C4 for ATAC Rep2 with Tn5-TXGv1.1 and C2 and C3 for ATAC with Tn5-H100. Technical biases from varying numbers of cells in clusters were compensated by randomly selecting 1,617 cells each. To compare the impact of counting, *single cell co-accessibility* was computed using binary, continuous, and allele count matrices in merged peaks within a 10 kb window using clusters C4 for ATAC Rep2 with Tn5-TXGv1.1 and C2 for ATAC with Tn5-H100. To compare ATAC with Tn5-TXGv1.1 to Tn5-H100, *single cell co-accessibility* was computed using binary count matrices in merged peaks within a 1 Mb window using clusters C3 for ATAC Rep2 with Tn5-TXGv1.1 and Tn5-H100. The resulting co-accessible links were filtered removing all links with negative co-accessibility scores, below 5 % of accessible cells, and with co-accessibility scores below cluster-specific background co-accessibility cutoff.

4.2.3. Analysis of scATAC-seq and scTurboATAC-seq data from PBMCs

I performed the analysis of scATAC-seq and scTurboATAC-seq data from PBMCs as described here and in Seufert *et al.* (2023). ATAC data were processed with Cell Ranger ATAC count (10x Genomics, Pleasanton, USA) using the provided human GRCh38 reference. Single-cell data was further analyzed with ArchR and visualized using ggplot2 in R. Empty barcodes were removed using (i) a minimal threshold of $10^{4.1}$ for ATAC with Tn5-TXGv2 and $10^{4.4}$ for ATAC with Tn5-H100 for number of unique fragments, and (ii) a minimal threshold of 10 for ATAC with Tn5-TXGv2 and 8 for ATAC with Tn5-H100 for TSS enrichment score. Cell numbers are provided in **Table 2.7**. Barcodes containing multiple cells were removed using Amulet from scDblFinder. Single cells from each experiment were embedded in two-dimensional space separately using an accessibility count matrix in 500 bp genomic tiles, iterative LSI (default parameters, except clustering resolution of 0.2) and UMAP (default parameters, except LSI components 2-21, minimal distance of points in embedding of 0.5, 30 nearest neighbors). Cell clusters were computed by SNN modularity optimization (default parameters, except LSI components 2-21, resolution of 0.2).

Sample-specific peaks from pseudo-bulk chromatin accessibility data were called using MACS2 in ArchR (default parameters, except peak summit extension by 1000 bp to each side; reproducibility of 1). Sample-specific peak sets were merged. Cell types were assigned by module scores from gene activity scores of cell type marker genes (**Table 4.3**). B cells from each experiment were embedded in two-dimensional space separately using an accessibility count matrix in 500 bp genomic tiles, iterative LSI (default parameters, except clustering resolution of 0.2) and UMAP (default parameters, except LSI components 2-8, minimal distance of points in embedding of 0.5, 30 nearest neighbors). Cell clusters were computed by SNN modularity optimization (default parame-

Table 4.3 Marker genes of human hematopoietic cell types. Adapted from Seufert *et al.* (2023).

Cell type	Marker genes	Reference
Basophils	*ENPP3*, *CD69*, *CCR3*, *PTGDR2*	Monaco *et al.* (2019); Uhlen *et al.* (2019)
B cells	*MS4A1*, *CD79A*, *CD74*, *CD19*	Uhlen *et al.* (2019); Karlsson *et al.* (2021)
Classical monocytes	*CD14*, *LYZ*, *CST3*	Uhlen *et al.* (2019)
Dentritic cells	*FCER1A*, *CST3*, *CD74*	Uhlen *et al.* (2019)
Eosinophils	*EPX*, *RNASE2*, *CLC*, *SIGLEC8*	Uhlen *et al.* (2019)
Erythrocytes	*HBA2*, *HBB*, *HBA1*, *SLC4A1*, *EPB41*	Karlsson *et al.* (2021)
Mature $CD4^+$ T cells	*CD4*, *IL7R*, *S100A4*, *CD3D*, *CD2*, *IL2*, *TNF*, *IL21*	Uhlen *et al.* (2019)
Mature $CD8^+$ T cells	*CD8A*, *GZMB*, *CD3D*, *CD2*	Uhlen *et al.* (2019)
Naïve $CD4^+$ T cells	*CD4*, *IL7R*, *CCR7*, *CD3D*, *CD34*, *LEF1*	Terstappen *et al.* (1992); Uhlen *et al.* (2019)
Naïve $CD8^+$ T cells	*CD8A*, *GZMB*, *CD3D*, *CD34*, *LEF1*	Terstappen *et al.* (1992); Uhlen *et al.* (2019)
Neutrophils	*CEACAM8*, *ITGAM*, *FCGR3B*, *PTPRC*	Monaco *et al.* (2019); Uhlen *et al.* (2019)
NK cells	*GNLY*, *NKG7*, *GZMB*, *NCAM1*, *FCGR3A*	Uhlen *et al.* (2019); Karlsson *et al.* (2021)
Non-classical monocytes	*FCGR3A*, *MS4A7*	Uhlen *et al.* (2019)
Platelets	*PPBP*, *ITGA2B*, *PF4*	Shattil *et al.* (1985); Poncz *et al.* (1987); Majumdar *et al.* (1991)
Progenitor cells	*CD34*, *KIT*, *PROM1*, *PTPRC* (negative marker), *CD38* (positive marker)	Terstappen *et al.* (1992); Monaco *et al.* (2019)

ters, except LSI components 2-8, resolution of 0.2). B cells from ATAC with Tn5- TXGv2 and ATAC with Tn5-H100 were integrated using Harmony (default parameters) and combined two-dimensional embedding was computed as for sample-specific embeddings. Transcription factor binding sites in sample-specific peaks were predicted using Homer hg38 motifs from chromVARmotifs. Transcription factor motif deviations were computed within 200 bp windows for top 1000 motifs per transcription factor (defined by highest scores) using chromVAR in ArchR.

4.2.4. Analysis of Multiome scRNA-seq, scATAC-seq and scTurboATAC-seq data from PBMCs

I performed the analysis of Multiome scRNA-seq, scATAC-seq and scTurboATAC-seq data from PBMCs as described here and in Seufert *et al.* (2023). Multiome data were processed with Cell Ranger ARC count (10x Genomics, Pleasanton, USA) using the provided human GRCh38 reference. Quality metrics from Multiome data are provided in **Table 2.8**. Single-cell data was further analyzed with ArchR for ATAC, Seurat for RNA, Signac for combined multi-omic data, and visualized using ggplot2 in R. For ATAC data, empty barcodes were removed using (i) a minimal threshold of $10^{3.8}$ for number of unique fragments and (ii) a minimal threshold of 10^{3} for number of reads in TSSs. Cell numbers from ATAC data are provided in **Table 2.8**. Barcodes containing multiple cells were removed using Amulet from scDblFinder. High-quality cells were selected using (i) a maximal threshold of 0.01 for the ratio of reads in blacklisted genomic regions and (ii) a maximal threshold of 3 for nucleosome ratio. Single cells from each experiment were embedded in two-dimensional space separately using an accessibility count matrix in 500 bp genomic tiles, iterative LSI (default parameters, except clustering resolution of 0.2) and UMAP (default parameters, except LSI components 2-20, minimal distance of points in embedding of 0.5, 30 nearest neighbors). Cell clusters were computed by SNN modularity optimization (default parameters, except LSI components 2-20, resolution of 0.3).

For RNA data, empty barcodes were removed using (i) a minimal threshold for number of detected genes of 10^{2}, (ii) a minimal threshold for UMI counts of 500, and (iii) a maximal threshold for percentage of mitochondrial UMI counts of 40. Cell numbers from RNA data are provided in **Table 2.8**. Barcodes containing multiple cells were removed using Scrublet with a cutoff of 0.15 in Python. Single cells from each experiment were embedded in two-dimensional space separately using SCTransform (default parameters), principle component analysis (PCA, default parameters) and UMAP (default parameters, except PCs 1-20). Cell clusters were computed by SNN modularity optimization (default

parameters, except PCs 1-20, resolution of 0.5). For combined multi-omic data, only high-quality cells from both ATAC and RNA were used. High-quality cell numbers from combined data are provided in **Table 2.8**. Single cells from each experiment were co-embedded in two-dimensional space separately using iterative LSI (default parameters, except clustering resolution of 0.2, LSI components 2–25) for ATAC, SCTransform (default parameters) and PCA (default parameters, except PCs 1–30) for RNA, WNN graph and UMAP. Cell clusters were computed by SNN modularity optimization (default parameters, except LSI components 2-25, PCs 1-30, resolution of 0.8).

4.2.5. Data and code availability

Data of scATAC-seq, scTurboATAC-seq and scRNA-seq from MEFs and PBMCs are available at Gene Expression Omnibus (GEO) as described in **Table 4.4**. Supplementary data with intermediate results are available at Seufert *et al.* (2023). My scripts for the computational analyses of scATAC-seq, scTurboATAC-seq and scRNA-seq data from MEFs and PBMCs are provided at https://github.com/RippeLab/TurboATAC. Co-accessibility analysis of scATAC-seq and scTurboATAC-seq data from MEFs was conducted with RWireX (v0.2.05, https://github.com/RippeLab/RWire-IFN).

Table 4.4 Data availability of scATAC-seq, scTurboATAC-seq and scRNA-seq from MEFs and PBMCs. Data are available at Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo) as part of the series GSE235506.

Data set	Samples	Method	GEO ID
MEF, 6 h IFNβ treatment	Tn5-TXGv1.1	scATAC-seq	GSM7504029
	Tn5-H30	scATAC-seq	GSM7504030
	Tn5-H100	scTurboATAC-seq	GSM7504031
PBMC	Tn5-TXGv2	scATAC-seq	GSM7504072
	Tn5-H100	scTurboATAC-seq	GSM7504073
PBMC	Multiome, Tn5-TXG	scATAC-seq	GSM7504032
	Multiome, Tn5-H50	scATAC-seq	GSM7504033
	Multiome, Tn5-H100	scTurboATAC-seq	GSM7504034
	Multiome, Tn5-TXG	scRNA-seq	GSM7504069
	Multiome, Tn5-H50	scRNA-seq	GSM7504070
	Multiome, Tn5-H100	scRNA-seq	GSM7504071

4.2.6. List of applied software packages

Utilized software for the computational analyses of scATAC-seq, scTurboATAC-seq and scRNA-seq data from MEFs and PBMCs is provided in **Table 4.5**.

Table 4.5 Software used for the analysis of scATAC-seq, scTurboATAC-seq and scRNA-seq data of MEFs and PBMCs.

Software	Version	Reference
ArchR	v1.0.3	Granja *et al.* (2021)
Cell Ranger ARC	v2.0.2	Zheng *et al.* (2017); Satpathy *et al.* (2019)
Cell Ranger ATAC	v2.1.0 (PBMC); v2.0.0 (MEF)	Satpathy *et al.* (2019)
ChromVAR	v1.20.0	Schep *et al.* (2017)
ChromVARmotifs	v0.2.0	Schep *et al.* (2017)
Ggplot2	v3.4.2	Wickham (2016)
Ggtern	v3.4.1	Hamilton & Ferry (2018)
Harmony	v0.1.1	Korsunsky *et al.* (2019)
MACS2	v2.1.2	Zhang *et al.* (2008)
Pandas	v1.4.3	McKinney (2010)
Python	v3.10.4	Python Software Foundation (1991)
R	v4.2.2	R Core Team (1993)
scDblFinder	v1.12.0	Thibodeau *et al.* (2021)
Scipy	v1.8.1	Virtanen *et al.* (2020)
Scrublet	v0.2.3	Wolock *et al.* (2019)
SCTransform	v0.3.5	Hafemeister & Satija (2019)
Seurat	v4.3.0	Stuart *et al.* (2019)
Signac	v1.9.0	Stuart *et al.* (2021)

4.3. scRNA-seq of AML patient samples

4.3.1. Sequencing data acquisition

Experiments and data acquisition were performed by Linda Schuster (formerly Division of Chromatin Networks, German Cancer Research Center, Germany) as described in Schuster *et al.* (2023). In short, PBMCs and bone marrow mononuclear cells (BMNCs) from AML patients were depleted from $CD3^+$ cells and enriched for mononuclear cells as described in Stosch *et al.* (2018). Single-cell RNA libraries were prepared using the Chromium Single Cell 3' v2 kit according to the manufacturer's protocol using 8,000 cells per sample (10x Genomics, Pleasanton, USA). Library fragment sizes were cleaned up using SPRI-select beads (Beckman Coulter, Brea, USA). Libraries were sequenced paired end with 26 bp and 96 bp for read 1 and read 2 on NovaSeq 6000 and HiSeq 4000 systems (Illumina, San Diego, USA). Sequencing was conducted by the DKFZ NGS Core Facility.

4.3.2. Analysis of scRNA-seq data

Analysis of scRNA-seq data was performed in collaboration with Linda Schuster as described in Schuster *et al.* (2023), where I supported coding and conceptualization. The analysis of TF activity was carried out using my scripts. Briefly, data was processed using Cell Ranger count (10x Genomics, Pleasanton, USA) with the provided human GRCh38-1.20-premrna reference. scRNA-seq data was further analyzed with Seurat in R. Low-quality cells were removed using (i) minimal and maximal thresholds for number of detected genes of 500 and 3,000, respectively; (ii) a maximal threshold for mitochondrial UMI counts of 15 %. Predicted cell duplicate barcodes with a doublet score above 0.4 from Scrublet in Python were discarded (default parameters, except simulated doublet ratio of 2, 30 neighbors, expected doublet rate of 0.1). Barcodes containing multiple cells were removed using Scrublet with a cutoff of 0.15 in Python. This yielded roughly 17,600 cells in total.

UMI counts were normalized using SCTransform (default parameters) and technical biases by total number of UMI counts and percentage of mitochondrial UMI counts per cell were regressed out. Samples were integrated using canonical correlation analysis (CCA) in Seurat. Single cells from all samples were embedded in two-dimensional space using PCA (default parameters, except using top 3,000 variable genes) and UMAP (default parameters, except PCs 1-15). Cell clusters were computed by Louvain method on k-

nearest neighbor (KNN) graph using Seurat (default parameters, except PCs 1-15). Non-malignant cell types were assigned to cell clusters by expression of cell type marker genes: *CD3D* for T cells and NK cells, *MS4A1* for B cells, *NKG7* and *GNLY* for NK cells, *HBB* for Erythroblasts, and *CD14* for Monocytes. Cell clusters without cell type marker gene expression were assigned as malignant cell clusters. Differential expression analysis was performed by Wilcoxon test in Seurat (default parameters, significance threshold: adjusted p-value < 0.05, log2 fold change > 0.1). Transcription factor activities were calculated with virtual inference of protein-activity by enriched regulons (Viper, default parameters) using human regulons of confidence levels A and B from Dorothea.

4.3.3. List of applied software packages

Utilized software for the computational analyses of scRNA-seq data of AML patient samples is provided in **Table 4.6**.

Table 4.6 Software used for the analysis of scRNA-seq data of AML patient samples.

Software	Version	Reference
Cell Ranger	v3.1.0	Zheng *et al.* (2017)
Dorothea	v1.10.0	Garcia-Alonso *et al.* (2019)
Pandas	v2.0.0	McKinney (2010)
Python	v3.8.16	Python Software Foundation (1991)
R	v4.0.2	R Core Team (1993)
Scipy	v1.9.3	Virtanen *et al.* (2020)
Scrublet	v0.2.3	Wolock *et al.* (2019)
SCTransform	v0.4.1	Hafemeister & Satija (2019)
Seurat	v4.0.0	Stuart *et al.* (2019)
Viper	v1.32.0	Alvarez *et al.* (2016)

4.4. Sequencing and spatial transcriptomics of HUVECs

4.4.1. Sequencing and imaging data acquisition

Experiments and data acquisition were performed by Irene Gerosa, Sabrina Schumacher (both Division of Chromatin Networks, German Cancer Research Center, Germany), Jan-Philipp Mallm and Katharina Bauer as described in Seufert *et al.* (2024). Briefly, HUVECs from pooled donors (Lonza, Basel, Switzerland) were starved for 20-24 h before treatment with human TNFα for 30 and 240 min. Biological replicates were acquired from different aliquots of HUVECs. scRNA libraries were prepared using the Chromium Next GEM Single Cell 5' (dual index) kit v2 with 10,000 cells according to the manufacturer's protocol (10x Genomics, Pleasanton, USA). scTurboATAC libraries were prepared using the Chromium Next GEM Single Cell ATAC kit v2 (10x Genomics, Pleasanton, USA) with 10,000 cells according to the TurboATAC protocol (Seufert *et al.*, 2023). snRNA libraries were prepared using the SMART-seq 2.5 protocol on 384 well plates as described previously in Ghasemi *et al.* (2024). scRNA and scTurboATAC libraries from different treatment conditions were separately pooled and sequenced paired-end on a NovaSeq 6000 system (Illumina, San Diego, USA) using S4 flow cells. Pooled scRNA libraries were sequenced with 100 bp for both read 1 and read 2, while pooled scTurboATAC libraries were sequenced with 50 bp for both read 1 and read 2. snRNA libraries were sequenced paired-end with 25 bp and 50 bp for read 1 and read 2 on a NextSeq 550 system (Illumina, San Diego, USA). Read 1 contained the UMI sequences at the first eight bp. Sequencing was conducted by the DKFZ NGS Core Facility.

Bulk ChIP libraries were prepared from 15-30 million 30 min TNFα-treated HUVECs with IgG control antibody (53017, Active Motif, Carlsbad, USA) and two different H3K27ac antibodies (ab4729, Abcam, Cambridge, UK; C15210016, Diagenode, Liège, Belgium) as described previously in Kolovos *et al.* (2016). Bulk ChIP libraries were sequenced single-end on a HiSeq 2000 system (Illumina, San Diego, USA).

Spatial transcriptomic images were acquired for untreated, 30 min and 240 min TNFα-treated HUVECs using the padFISH protocol (Seufert *et al.*, 2024). Probes against nascent RNA from intronic regions of *CXCL1*, *CXCL2*, *CXCL3*, *CXCL8*, *NFKBIA*, and *SELE* were used. DAPI staining and detection oligonucleotides labeled with Alexa Fluor 488, ATTO 550, Alexa Fluor 647, and Alexa Fluor 750 were used for fluorescence imaging on an Andor Dragonfly 505 spinning disk confocal unit equipped with a Nikon Ti2-E inverted

microscope and a Plan Apo 60x/1.40 oil objective or a 100x CFI SR HP Plan Apochromat Lambda S silicone immersion objective.

4.4.2. Analysis of scRNA-seq data

I performed the analysis of scRNA-seq data from three biological replicates of HUVECs as described here and in Seufert *et al.* (2024). Data were processed with Cell Ranger count (10x Genomics, Pleasanton, USA) using the provided human GRCh38-2020-A reference (default parameters, except including introns). Quality metrics of scRNA-seq data are provided in **Table 4.7**. Data were further analyzed with Seurat and visualized with ggplot2 in R. Empty barcodes were removed using (i) a minimal threshold for number of detected genes of 10^2, (ii) a minimal threshold for UMI counts of 5,000, and (iii) a maximal threshold for percentage of mitochondrial UMI counts of 5. Next, cells were removed that contained (i) UMI counts above the sample's mean plus twice the standard deviation and (ii) mitochondrial UMI counts above or below the sample's mean plus or minus thrice the standard deviation. High-quality cell numbers are provided in **Table 4.7**. Samples were merged, log normalized, and scaled (default parameters, except regressing out UMI counts per cell). Low-dimensional single-cell embedding was computed using PCA (default parameters) and UMAP (default parameters, except PCs 1-16). Cell cycle phases were inferred per single cell by module scores from the expression of cell cycle marker genes (Kowalczyk *et al.*, 2015). Only cells in G1 cell cycle phase were considered for further analysis. Low-dimensional embedding of G1 cells was generated using PCA (default parameters) and UMAP (default parameters, except PCs 1-20).

Differential expression analysis was performed on pseudo-bulk per sample between untreated and TNFα-treated conditions across all replicates using DESeq2 (default parameters, significance threshold: adjusted p-value < 0.05, absolute log2 fold change > 1). Genomic distances between differentially expressed genes on the same chromosome were inferred using GenomicRanges and gUtils with the genome reference arc-GRCh38-2020-A-2.0.0 (10x Genomics, Pleasanton, USA). Differentially expressed genes within 500 kb were considered proximal. If at least two differentially expressed genes were proximal, they were classified as a differentially expressed gene cluster. Consequently, the differentially expressed genes were categorized as clustered or isolated. Gene clusters were visualized as network graph using igraph and qgraph.

Table 4.7 Quality metrics of scRNA-seq data from three replicates of HUVECs. Adapted from Seufert *et al.* (2024).

		Sample (TNFα treatment)		
		0 min	**30 min**	**240 min**
Replicate 1	Sequenced read pairs	311,628,705	265,034,290	302,252,205
	Confidently mapped read pairs (%)	82.9	81.1	81.8
	Number of high-quality cells	3,891	3,098	3,121
	UMI counts per cell (median)	27,718	26,934	29,433
	Genes per cell (median)	6,078	5,920	6,017
	Mitochondrial UMI counts (%, median)	2	2	2
Replicate 2	Sequenced read pairs	226,548,223	252,208,026	265,004,092
	Confidently mapped read pairs (%)	85.5	85.6	85.2
	Number of high-quality cells	3,707	4,326	3,164
	UMI counts per cell (median)	21,272	20,112	27,630
	Genes per cell (median)	5,483	5,179	5,969
	Mitochondrial UMI counts (%, median)	2	2	2
Replicate 3	Sequenced read pairs	290,895,239	325,433,791	271,413,583
	Confidently mapped read pairs (%)	80.4	83.5	81.9
	Number of high-quality cells	4,750	5,341	3,822
	UMI counts per cell (median)	20,068	21,511	22,283
	Genes per cell (median)	5,356	5,457	5,439
	Mitochondrial UMI counts (%, median)	2	2	2

Simultaneous co-expression of multiple genes in the same cell was inferred from Spearman correlation of UMI counts across cells for each sample. Only differentially expressed genes with expression in at least 10 % of cells were considered. Spearman correlation coefficients of replicates were averaged. Co-expression between all isolated differentially expressed genes was compared to co-expression between clustered differentially expressed genes within the same cluster. Overall co-expression per cluster was calculated by averaging the Spearman correlation coefficients of gene combinations within. Additionally, co-expression patterns of *CXCL1*, *CXCL2*, *CXCL3*, and *CXCL8* in the

CXCL gene cluster were determined for each cell. Active transcription was defined per cell that contained at least one UMI count.

4.4.3. Analysis of scTurboATAC-seq data

I performed the analysis of scTurboATAC-seq data from three biological replicates of HUVECs as described here and in Seufert *et al.* (2024). Data were processed with Cell Ranger ATAC count (10x Genomics, Pleasanton, USA) using the provided human GRCh38-2020-A-2.0.0 reference (default parameters). Quality metrics of scTurboATAC-seq data are provided in **Table 4.8**. Data were further analyzed with ArchR and visualized with ggplot2 in R. Empty barcodes were removed using (i) a minimal threshold for number of unique fragments of $10^{4.5}$ and (ii) a minimal threshold for TSS enrichment score of 7. Barcodes containing multiple cells were removed using Amulet with a 5^{th} percentile cutoff for significant q-values from scDblFinder. Next, cell outliers were removed that contained blacklist ratios above the overall mean plus twice the standard deviation. High-quality cell numbers are provided in **Table 4.8**. Low-dimensional single-cell embedding was computed using an accessibility count matrix of 500 bp genomic tiles, iterative LSI (default parameters, except seed of 42) and UMAP (default parameters, except LSI components 2-14, seed of 1). Cell cycle phases were predicted per single cell by integration with scRNA-seq samples using ATAC gene activity scores in ArchR (default parameters, except constrained integration within corresponding samples). Only cells in G1 cell cycle phase were considered for further analysis. Low-dimensional single-cell embedding was computed using an accessibility count matrix of 500 bp genomic tiles, iterative LSI (default parameters, except seed of 42) and UMAP (default parameters, except LSI components 2-8, seed of 1).

Sample-specific peaks from pseudo-bulk chromatin accessibility data were called using MACS2 in ArchR (default parameters, except peak summit extension by 500 bp to each side; reproducibility of 2). A union ATAC peak set across all samples was generated. Differential accessibility analysis for ATAC peaks was performed between untreated and TNFα-treated conditions across all replicates using Wilcoxon test in ArchR (default parameters, except maximum of 6,000 cells per sample, bias correction by TSS enrichment score and number of unique fragments (log10), normalization by number of unique fragments; significance threshold: false discovery rate (FDR) < 0.05, absolute log2 fold change > 1).

Table 4.8 Quality metrics of scTurboATAC-seq data from three replicates of HUVECs. Adapted from Seufert *et al.* (2024).

		Sample (TNFα treatment)		
		0 min	**30 min**	**240 min**
Replicate 1	Sequenced read pairs	2,655,922,656	2,230,005,333	2,428,973,921
	Duplicates (%)	61.5	55.2	58.0
	Confidently mapped read pairs (%)	87.3	88.2	88.0
	Number of high-quality cells	4,902	5,202	5,480
	Unique fragments/cell (median)	128,825	134,896	125,893
	TSS enrichment score (median)	10.34	10.43	10.45
Replicate 2	Sequenced read pairs	2,151,183,819	2,136,266,937	2,387,966,429
	Duplicates (%)	61.9	61.2	53.5
	Confidently mapped read pairs (%)	88.2	89.8	88.6
	Number of high-quality cells	2,425	4,424	6,965
	Unique fragments/cell (median)	204,174	141,254	117,490
	TSS enrichment score (median)	9.87	10.51	10.47
Replicate 3	Sequenced read pairs	2,615,116,880	2,659,573,577	2,032,193,581
	Duplicates (%)	59.1	58.6	71.8
	Confidently mapped read pairs (%)	88.1	88.1	92.0
	Number of high-quality cells	5,884	5,645	5,244
	Unique fragments/cell (median)	123,027	131,826	66,069
	TSS enrichment score (median)	13.5	10.81	11.22

Technical biases between samples from varying numbers of cells and unique fragments per cell were compensated by selecting 1,000 most similar cells from each sample compared to a reference of 1,000 randomly selected cells from the 240 min TNFα-treated HUVEC Rep1 sample. Low-dimensional embedding of bias-compensated cells from Rep1 was generated using an accessibility count matrix of 500 bp genomic tiles, iterative LSI (default parameters, except seed of 42) and UMAP (default parameters, except LSI components 2-9, seed of 1). Low-dimensional embedding of bias-compensated 240 min TNFα-treated cells from Rep1 was generated using an accessibility count matrix of 500

bp genomic tiles, iterative LSI (default parameters, except seed of 42) and UMAP (default parameters, except LSI components 2-6, seed of 1).

Single cell co-accessibility analysis was performed with 1,000 bias-compensated cells for each sample separately using RWireX. Co-accessibility scores were computed within 1 Mb using a continuous accessibility count matrix of ATAC peaks and single cells. The resulting co-accessible links were filtered removing all links with negative co-accessibility scores, percent accessible cells (PAC) below 5, and co-accessibility scores below sample-specific background co-accessibility cutoffs. The remaining links were considered the autonomous links of co-accessibility (ACs). ACs were compared between replicates by determining the percent of identical ACs between two samples. Two-replicate consensus ACs were obtained for each treatment time point by selecting ACs that were detected in at least two replicates. The average co-accessibility scores and PACs of replicates were used for consensus ACs. ACs were visualized in exemplary regions by loop tracks with genomic annotations and number of unique fragment-normalized pseudo-bulk chromatin accessibility tracks using ArchR.

Metacell co-accessibility analysis was performed with 3,000 bias-compensated cells from all treatment time points for each replicate separately using RWireX. Metacells were formed from unique sets of 10 cells each, not using the final 10 % of cells to prevent the forced aggregation of dissimilar cells. Co-accessibility scores were computed within 2 Mb using a continuous accessibility count matrix of 10 kb genomic tiles and metacells. Three-replicate consensus *metacell co-accessibility* was computed by averaging co-accessibility scores across replicates. Domains were called from only positive replicate and consensus co-accessibility score matrices using SpectralTAD (default parameters, except 3 hierarchical levels, run twice for (i) small domains with minimal domain size of 20 kb, window size of 200 kb; (ii) large domains with minimal domain size of 200 kb, window size of 2 Mb). Overall co-accessibility scores per domain were calculated by averaging the co-accessibility scores within the respective domain. The resulting small and large domains were filtered separately by lower cutoffs from 90^{th} percentile domain co-accessibility scores. The remaining domains were considered the domains of contiguous co-accessibility (DCs). Replicate and Consensus DCs were compared by determining the percent of bp overlap between the DCs. The *metacell co-accessibility matrices* were visualized in exemplary regions with genomic annotation of DCs using plotgardener. Differential accessibility analysis for DCs was performed between untreated and TNFα-treated conditions across all replicates using Wilcoxon test in ArchR (default parameters, except maximum of 6,000 cells per sample, bias correction by TSS enrichment score and

number of unique fragments (log10), normalization by number of unique fragments; significance threshold: FDR < 0.05).

The overlap of AC-linked ATAC peaks and DCs with differentially expressed genes from scRNA-seq data was determined using GenomicRanges. ACs and DCs at TSSs ± 500 bp of differentially expressed genes were quantified, deriving so-called AC and DC features. The DC feature was binarized, whereas AC features were log10 transformed with a pseudo-count of 1. Following, both AC and DC features were min-max normalized. Differentially expressed genes were clustered (ward.D clustering, 5 clusters selected) and visualized by heatmap using pheatmap. Differentially expressed genes were classified by their predominant feature into AC-driven, DC-driven, AC/DC-driven or not assigned (NA) genes. Additionally, the differentially expressed genes were classified as promoter-regulated or non-promoter-regulated genes by presence of a significantly differential ATAC peak at any of their TSS ± 500 bp regions. Next, AC score and DC score were calculated for each differentially expressed gene cluster as following:

$$AC\ score_{gene\ cluster} = \frac{N\ genes\ AC_{gene\ cluster} + N\ genes\ AC/DC_{gene\ cluster}}{N\ genes_{gene\ cluster}}$$

$$DC\ score_{gene\ cluster} = \frac{N\ genes\ DC_{gene\ cluster} + N\ genes\ AC/DC_{gene\ cluster}}{N\ genes_{gene\ cluster}}$$

I assigned differentially expressed gene clusters as (i) predominantly AC-driven if AC scores were ≥ 0.5 and DC scores were < 0.5, (ii) predominantly DC-driven if AC scores were < 0.5 and DC scores were ≥ 0.5, (iii) predominantly AC/DC-driven if both AC scores and DC scores were ≥ 0.5, and (iv) NA if both AC scores and DC scores were < 0.5. Annotated gene clusters were visualized as network graph using igraph and qgraph.

Locus-specific TF binding activity scores and TF-bound sites were inferred from pseudo-bulks of individual scTurboATAC-seq samples and Homer motifs from chromVARmotifs in ATAC peaks using Tobias in Python. Number of bound sites per TF were calculated by averaging the total bound sites across replicates. Genome-wide differential TF binding was identified between untreated and TNFα-treated conditions calculating the log2 fold change for number of bound sites per TF (significance threshold: log2 fold change > 0.1). TF footprints were visualized per sample in exemplary DC and non-DC regions using ArchR (default parameters, except smoothing window of 20, no normalization, maximum of 1,000 cells per sample). Differential TF binding between individual DC regions and the global non-DC background regions was calculated within each sample for the genome-wide differential TFs. Average log2 fold change and one-sided Wilcoxon test was used to

compare TF binding activity scores in a DC to the global non-DC background. Meta-analysis with Fisher's method using poolr and averaging of log2 fold changes was performed to integrate results from replicates (significance threshold: combined p-value < 0.05, log2 fold change > 1). DCs with significantly increased TF binding activity were visualized by heatmap (default parameters, except ward.D2 clustering) using pheatmap.

4.4.4. Analysis of snRNA-seq data

Processing of snRNA-seq data was performed by Ezgi Sen (Division of Chromatin Networks, German Cancer Research Center, Germany), while I performed further downstream analysis as described here and in Seufert *et al.* (2024). Data were processed using the nf-core rnaseq pipeline in Nextflow (default parameters, except UMI extraction from read 1, read 2 alignment using Star, quantification of UMI counts in exons for genes and introns for transcripts using Salmon) with the human GRCh38-2020-A reference from Cell Ranger (10x Genomics, Pleasanton, USA). Quality metrics of snRNA-seq data are provided in **Table 4.9**. Data were further analyzed with Seurat and visualized with ggplot2 in R. Cells were removed that contained (i) less than 10^2 detected genes from exon counting and (ii) more than 5 % of mitochondrial UMI counts from exon counting. Next, cells that contained exonic UMI counts above or below the sample's mean plus or minus thrice the standard deviation were removed. High-quality cell numbers are provided in **Table 4.9**. Data was log normalized and scaled (default parameters). Module scores for S and G2M cell cycle phases were inferred per cell from the expression of cell cycle marker genes (Kowalczyk *et al.*, 2015). Cells with S and G2M module scores below the sample's

Table 4.9 Quality metrics of snRNA-seq data of HUVECs. Adapted from Seufert *et al.* (2024).

	Sample (TNFα treatment)		
	0 min	**30 min**	**240 min**
Sequenced read pairs	372,894	557,968	208,211
Confidently mapped read pairs (%)	27.6	20.2	32
Number of high-quality cells	378	380	376
Exonic UMI counts/cell (median)	20,730	24,647	11,129
Exonic genes/cell (median)	8,587	9,594	4,356
Intronic UMI counts/cell (median)	70,800	87,122	50,333
Intronic genes/cell (median)	22,362	25,525	13,896
Mitochondrial RNA UMI counts (%)	0.1	0.06	0.38

mean plus the standard deviation were assigned to G1 cell cycle phase. Only cells in G1 cell cycle phase were considered for further analysis.

Transcriptional bursting kinetics were inferred following the two-state model of transcription as in Mahat *et al.* (2024). Intronic UMI counts for transcripts of differentially expressed genes from scRNA-seq data were used. Only transcripts with the same direction of regulation across conditions compared to the scRNA-seq data were used. Additionally, transcripts with no intronic UMI counts in 95 % of cells from all samples were removed. The capture efficiency per sample was estimated from total UMI counts, expecting 20 % of 500,000 mRNA molecules/cell in the nucleus (0.33 for 0 min; 0.36 for 30 min; 0.21 for 240 min). The transcription time per transcript was estimated from the transcript length, assuming a transcription rate of 150 kb/h. Per transcript and sample, the average intronic UMI counts above 0 ($\overline{intr.UMIs}$) and the number of cells with intronic UMI counts above 0 ($N\ tran.cells$) were calculated. Burst size and burst frequency were calculated for each transcript and sample as following:

$$Burst\ size_{transcript,sample} = 1 + \frac{\overline{intr.UMIs}_{transcript,sample} - 1}{Capture\ efficiency_{sample}}$$

$$Burst\ frequency_{transcript,sample} = \frac{N\ tran.cells_{\ sample} / N\ cells_{sample} / Transcription\ time_{transcript}}{\min\left(Burst\ size_{transcript,sample} * Capture\ efficiency_{sample}, 1\right)}$$

Overall burst size and burst frequency per differentially expressed gene were calculated by weighted averages of transcript-level burst sizes and burst frequencies. For comparison with padFISH, burst sizes and burst frequencies were scaled from zero to one across replicates and both *NFKBIA* and *SELE*.

4.4.5. Analysis of spatial transcriptomics data

Processing and image analysis were performed by Irene Gerosa, while I performed further downstream analysis as described here and in Seufert *et al.* (2024). Briefly, raw image stacks in Imaris format were converted to maximum projected files in TIF format using FIJI. Flatfield correction, chromatic aberration correction, and stitching were performed using FIJI. Nuclei were segmented based on the 4′,6-Diamidin-2-phenylindol (DAPI) fluorescence intensity using Cellpose with a pretrained cyto model (default parameters,

except diameter of 150 for 60x and 200 for 100x objectives). Removal of nuclei at image borders or with overexposure of padFISH fluorescence intensities was conducted in R.

For the co-expression analysis, Z projection of *CXCL1*, *CXCL2*, *CXCL3*, and *CXCL8* fluorescence intensity was performed to identify loci with co-expression of at least two genes. Gaussian Blur 2.0 was used to filter the Z-projected image in FIJI, which was subsequently segmented using ilastik with the pixel classification workflow. Segmented co-expression loci were binarized and filtered using Gaussian blur 1.5 in FIJI. Area, mean gray value, and center of mass were measured for each co-expression locus across all padFISH and DAPI fluorescence intensities using FIJI. Co-expression loci were filtered by a minimal fluorescence intensity threshold and assigned to their respective nuclei. The sum of *CXCL1*, *CXCL2*, *CXCL3*, and *CXCL8* fluorescence intensity per co-expression locus was quantified in R. Active or inactive transcription were defined per co-expression locus by fluorescence intensity above or below a minimum threshold from the bimodal fluorescence intensity distribution per gene and replicate. Only nuclei with one to two co-expression loci were used for the analysis. Additionally, nuclei with exceptionally big or small co-expression loci and high or low fluorescence intensities were removed. Co-expression patterns were determined at all co-expression loci in R and visualized using ggplot2.

For the transcriptional bursting analysis, the sum of *NFKBIA* and *SELE* fluorescence intensity per nucleus was quantified in R. Active or inactive transcription were defined per nucleus by fluorescence intensity above or below a minimum threshold from the bimodal fluorescence intensity distribution per gene and replicate. Transcriptional bursting kinetics were inferred with the same model as for snRNA-seq data in R. The transcript detection efficiency was previously estimated as 0.35 for padFISH (Rademacher *et al.*, 2024). Average transcript lengths per gene were assumed. Burst sizes and frequencies were scaled from zero to one across genes and replicates and visualized using ggplot2.

4.4.6. Analysis of bulk HiC-seq data

Processing of bulk HiC-seq data was performed and pile-up plots were generated by Vassiliki Varamogianni-Mamatsi (Institute of Pathology, University Medical Center Göttingen, Göttingen, Germany), while I performed further downstream analysis as described here and in Seufert *et al.* (2024). Bulk Hi-C-seq data of unstimulated HUVECs were obtained from Rao *et al.* (2014). The genomic annotation of the contact matrices was converted to the hg38 reference using HiCLift. Arrowhead was used to call TADs on the

25 kb contact matrix. Contact frequency enrichment at TADs and between ACs was visualized by pile-up plots using coolpup.py. Balanced contact counts were obtained from cooler (default parameters, except 64 bin balancing). Downstream analysis was conducted in R with visualization by ggplot2. The contact matrix at 10 kb resolution was visualized in exemplary regions using plotgardener. Overlap of TADs with ACs, DCs, and differentially expressed gene clusters was determined using GenomicRanges. Additionally, the overlap of 10 kb HiC bins with ATAC peaks was determined using Genomic Ranges to extract the balanced contact counts of AC-linked peaks.

4.4.7. Analysis of bulk H3K27ac ChIP-seq data

Processing of bulk H3K27ac ChIP-seq data was performed by Panagiotis Liakopoulos (Department of Molecular Biology and Genetics, Democritus University of Thrace, Greece), while I performed further downstream analysis as described here and in Seufert *et al.* (2024). Bulk ChIP-seq data were aligned using Bowtie2 with the human hg38 reference. Genomic H3K27ac coverage files were obtained using ShortRead. H3K27ac peaks were identified as previously described in Stadhouders *et al.* (2015) using IgG ChIP-seq signal as background and a significance threshold with FDR below 0.001 and at least 20 reads per peak. Downstream analysis was conducted in R with visualization by ggplot2. H3K27ac peaks from both antibodies were merged. Overlap of merged H3K27ac peaks with ATAC peaks, ACs, DCs, and differentially expressed genes was determined using GenomicRanges.

4.4.8. Data and code availability

Data of scTurboATAC-seq, scRNA-seq, snRNA-seq and bulk ChIP-seq from HUVECs are available at GEO as described in **Table 4.10**. Spatial transcriptomics data from HUVECs are available at BioImage Archive (https://www.ebi.ac.uk/bioimage-archive) with the accession number S-BIAD1294. Supplementary data with intermediate results are available at Seufert *et al.* (2024). My scripts for the computational analyses of scTurboATAC-seq, scRNA-seq, snRNA-seq and spatial transcriptomics data from HUVECs are provided at https://doi.org/10.5281/zenodo.13221210. Co-accessibility analysis of scTurboATAC-seq data from HUVECs was conducted with RWireX (v1.1.06, https://github.com/RippeLab/RWireX).

Table 4.10 Data availability of scTurboATAC-seq, scRNA-seq, snRNA-seq and bulk ChIP-seq from HUVECs. Data are available at GEO (https://www.ncbi.nlm.nih.gov/geo) as part of the series GSE273430.

Method	Samples	GEO ID
scTurboATAC-seq	HUVEC, replicate 1, untreated	GSM8428166
	HUVEC, replicate 1, 30 min TNFα treatment	GSM8428167
	HUVEC, replicate 1, 240 min TNFα treatment	GSM8428168
	HUVEC, replicate 2, untreated	GSM8428169
	HUVEC, replicate 2, 30 min TNFα treatment	GSM8428170
	HUVEC, replicate 2, 240 min TNFα treatment	GSM8428171
	HUVEC, replicate 3, untreated	GSM8428172
	HUVEC, replicate 3, 30 min TNFα treatment	GSM8428173
	HUVEC, replicate 3, 240 min TNFα treatment	GSM8428174
scRNA-seq	HUVEC, replicate 1, untreated	GSM8428154
	HUVEC, replicate 1, 30 min TNFα treatment	GSM8428155
	HUVEC, replicate 1, 240 min TNFα treatment	GSM8428156
	HUVEC, replicate 2, untreated	GSM8428157
	HUVEC, replicate 2, 30 min TNFα treatment	GSM8428158
	HUVEC, replicate 2, 240 min TNFα treatment	GSM8428159
	HUVEC, replicate 3, untreated	GSM8428160
	HUVEC, replicate 3, 30 min TNFα treatment	GSM8428161
	HUVEC, replicate 3, 240 min TNFα treatment	GSM8428162
snRNA-seq	HUVEC, untreated	GSM8428163
	HUVEC, 30 min TNFα treatment	GSM8428164
	HUVEC, 240 min TNFα treatment	GSM8428165

Bulk ChIP-seq	HUVEC, 30 min TNFα treatment, H3K27ac replicate 1	GSM8449517
	HUVEC, 30 min TNFα treatment, H3K27ac replicate 2	GSM8449518
	HUVEC, 30 min TNFα treatment, Input	GSM8449519

4.4.9. List of applied software packages

Utilized software for the computational analyses of sequencing and spatial transcriptomics data of HUVECs is provided in **Table 4.11**.

Table 4.11 Software used for the analysis of sequencing and spatial transcriptomics data of HUVECs. Not available software versions are indicated as n.a..

Software	Version	Reference
ArchR	v1.0.3	Granja *et al.* (2021)
Arrowhead	n.a.	Durand *et al.* (2016)
Bowtie2	v2.3.3	Langmead & Salzberg (2012)
Cellpose2	n.a.	Pachitariu & Stringer (2022)
Cell Ranger	v7.1.0	Zheng *et al.* (2017)
Cell Ranger ATAC	v2.1.0	Satpathy *et al.* (2019)
ChromVARmotifs	v0.2.0	Schep *et al.* (2017)
Coolpup.py	n.a.	Flyamer *et al.* (2020)
DESeq2	v1.40.2	Love *et al.* (2014)
FIJI	v2.14.0	Schindelin *et al.* (2012)
GenomicRanges	v1.52.0	Lawrence *et al.* (2013)
Ggplot2	v3.4.3	Wickham (2016)
GUtils	v0.2.0	Wala & Imielinski (2023)
HiCLift	n.a.	Wang & Yue (2023)
Igraph	v1.5.1	Csárdi *et al.* (2024)
MACS2	v2.1.2	Zhang *et al.* (2008)
Nextflow	v22.10.6	Ewels *et al.* (2020)
Nf-core rnaseq	v3.9.0	Patel *et al.* (2018)

Pheatmap	v1.0.12	Kolde (2018)
Plotgardener	v1.6.2	Kramer *et al.* (2022)
Poolr	v1.1-1	Cinar & Viechtbauer (2022)
Python	v3.10.12	Python Software Foundation (1991)
Qgraph	v1.9.8	Epskamp *et al.* (2012)
R	v4.3.1 (Seq.) v4.3.2 (Imag.)	R Core Team (1993)
scDblFinder	v1.14.0	Thibodeau *et al.* (2021)
Seurat	v4.3.0.1	Stuart *et al.* (2019)
ShortRead	n.a.	Morgan *et al.* (2009)
SpectralTAD	v1.16.1	Cresswell *et al.* (2020)
Tobias	v0.15.1	Bentsen *et al.* (2020)

4.5. Bulk and single cell sequencing of ESCs and MEFs

4.5.1. Sequencing data acquisition

Experiments and data acquisition were performed by Markus Muckenhuber as described in Muckenhuber *et al.* (2023). Briefly, mouse 129/Ola ESCs and MEFs were cultured for two days before treatment with IFNβ for 1 and 6 h. Biological replicates were acquired from different aliquots of ESCs and MEFs. scRNA libraries were prepared using the Chromium Single Cell 3' kit v2.0 (10x Genomics, Pleasanton, USA) according to the manufacturer's protocol. scATAC libraries were prepared using the Chromium Single Cell ATAC kit v1.0 (10x Genomics, Pleasanton, USA) according to the manufacturer's protocol. scRNA libraries were sequenced paired-end with 26 bp and 74 bp for read 1 and read 2 on a HiSeq 4000 system (Illumina, San Diego, USA). scATAC libraries were sequenced paired-end with 50 bp for both read 1 and read 2 on a NovaSeq 6000 system (Illumina, San Diego, USA). Sequencing was conducted by the DKFZ NGS Core Facility.

RNA was isolated using the NucleoSpin RNA kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's protocol, except eluting twice with RNase-free water within the same tube. rRNAs were removed using the Ribo-Zero rRNA Removal kit (Illumina, San Diego, USA) according to the manufacturer's protocol. Bulk RNA libraries were prepared using the NEB Next Ultra II directional RNA library preparation kit (Illumina,

San Diego, USA) according to the manufacturer's protocol. The ATAC reaction was conducted using Tn5-ILMN (Illumina, San Diego, USA) in Tag buffer. Tagmented DNA was purified using the MinElute PCR Purification kit (Qiagen, Hilden, Germany) and amplified by PCR. Next, the bulk ATAC libraries were purified with AMPure beads (Beckman Coulter, Brea, USA). Immunoprecipitation was conducted using the ChIP enzymatic chromatin IP kit (Cell Signaling Technology, Danvers, USA) according to manufacturer's protocol with antibodies against STAT1 phosphorylated at tyrosine position 701 (#7640, Cell Signaling Technology, Danvers, USA), STAT2 (#72604, Cell Signaling Technology, Danvers, USA), and IgG control (#2729, Cell Signaling Technology, Danvers, USA). Bulk ChIP libraries of STAT1 and STAT2 were prepared using the NEB Next Ultra II DNA library preparation kit (Illumina, San Diego, USA) according to the manufacturer's protocol. Bulk RNA libraries were sequenced single-end with 50 bp on a HiSeq 4000 system (Illumina, San Diego, USA). Bulk ATAC libraries were sequenced paired-end with 50 bp for both read 1 and read 2 on a HiSeq 2000 and 4000 system (Illumina, San Diego, USA). Bulk ChIP libraries of STAT1 and STAT2 were sequenced single-end with 50 bp on a HiSeq 4000 system (Illumina, San Diego, USA). Sequencing was conducted by the DKFZ NGS Core Facility.

4.5.2. Analysis of bulk sequencing data

Analysis of bulk sequencing data was performed by Markus Muckenhuber as described in Muckenhuber *et al.* (2023). Briefly, bulk RNA-seq data were mapped to the mouse mm10 reference and transcript counts were quantified using Star. Transcript counts were normalized to TPMs using RSEM. Differential expression analysis was performed between untreated and IFNβ-treated conditions across all replicates using DESeq2 (significance threshold: p-value < 0.05, log2 fold change > 1.5). Bulk ATAC-seq and STAT1 and STAT2 ChIP-seq data were mapped to the mouse mm10 reference using Bowtie2. Duplicated reads and reads that mapped to mitochondrial and blacklisted regions (Encode Project Consortium, 2012) were removed. Bulk ATAC-seq peak calling was performed for each replicate separately using MACS2. Bulk STAT1 and STAT2 ChIP-seq peak calling was performed across all replicates using MACS2. A consensus peak set was generated by intersecting peak sets of all IFNβ treatment conditions. Differential STAT1 and STAT2 binding analysis for consensus peaks was performed between untreated and IFNβ-treated conditions across all replicates using DiffBind (significance threshold: FDR < 0.05, log2 fold change > 4). Overlap between differential STAT1 and STAT2 peaks was determined using GenomicRanges.

4.5.3. Analysis of scRNA-seq data

Analysis of scRNA-seq data was performed by Markus Muckenhuber as described in Muckenhuber *et al.* (2023). In short, data were processed with Cell Ranger count (10x Genomics, Pleasanton, USA) using the provided mouse mm10 reference (default parameters). Data were further analyzed with Seurat in R. High-quality cells were selected by (i) a minimal threshold for percentage of mitochondrial reads of 2.5 % for ESCs and 0.5 % for MEFs, (ii) a maximal threshold for percentage of mitochondrial reads of 7.5 %, (iii) a minimal threshold for number of detected genes of 2,000 for ESCs and 1,250 for MEFs, and (iv) a maximal threshold for number of detected genes of 6,500. This resulted in 1,332 high-quality ESCs for 0 h, 2,085 for 1 h and 4,825 for 6 h of IFNβ stimulation as well as 9,771 high-quality MEFs for 0 h, 10,186 for 1 h and 7,579 for 6 h of IFNβ stimulation. Samples were merged, normalized (log10), and scaled (default parameters). Single cells were embedded in two-dimensional space using PCA and UMAP.

4.5.4. Analysis of scATAC-seq data

I performed the analysis of scATAC-seq data from ESCs and MEFs as described here and in Muckenhuber *et al.* (2023). Data were processed with Cell Ranger ATAC count (10x Genomics, Pleasanton, USA) using the provided mouse mm10 reference (default parameters). Quality metrics of scATAC-seq data are provided in **Table 4.12**. Data were further analyzed with ArchR and visualized with ggplot2 in R. High-quality cells were selected using (i) a minimal threshold for number of unique fragments of $10^{3.5}$, (ii) a maximal threshold for number of unique fragments of 10^{5}, (iii) a minimal threshold for TSS enrichment score of 4, and (iv) a maximal threshold for ratio of reads in blacklisted genomic regions of 0.0225 for ESCs and 0.016 for MEFs. High-quality cell numbers are provided in **Table 4.12**. Single ESCs and MEFs were embedded in two-dimensional space separately using an accessibility count matrix of 500 bp genomic tiles, iterative LSI (default parameters, except clustering resolution of 0.2) and UMAP (default parameters, except LSI components 1-30 for ESCs and 2-12 for MEFs, minimal distance of points in embedding of 0.5, 30 nearest neighbors). MEF subtypes were predicted per single cell by integration with scRNA-seq samples using ATAC gene activity scores in ArchR (default parameters). Cells in clusters C1 were removed for both ESCs and MEFs.

Table 4.12 Quality metrics of scATAC-seq data from ESCs and MEFs. Cell number is the number of cell-positive barcodes as defined by Cell Ranger ATAC. High-quality cell number is the number of remaining cells after quality filtering. Adapted from Muckenhuber *et al.* (2023).

		Sample (IFNβ treatment)		
		0 h	**1 h**	**6 h**
ESCs	Cell number	8,925	-	5,596
	Number of high-quality cells	7,390	-	4,548
	Unique fragments/cell (median)	16,397	-	24,512
	Fraction of reads in peaks (median)	0.65	-	0.65
MEFs	Cell number	11,656	12,272	19,403
	Number of high-quality cells	8,052	8,395	9,409
	Unique fragments/cell (median)	12,799	12,095	4,816
	Fraction of reads in peaks (median)	0.68	0.64	0.68

Sample-specific peaks from pseudo-bulk chromatin accessibility data were called using MACS2 in ArchR (default parameters, except peak summit extension by 1000 bp to each side; reproducibility of 1). A union ATAC peak set across all samples was generated. Differential accessibility analysis for union ATAC peaks was performed between untreated and IFNβ-treated conditions using Wilcoxon test in ArchR (default parameters, except bias correction by TSS enrichment score and number of unique fragments (log10); significance threshold: FDR ≤ 0.05, absolute log2 fold change ≥ 1). The union ATAC peaks were extended by regions of interest (2 kb around the midpoints of STAT1/2 peaks and differentially expressed gene TSSs).

Technical biases between samples from varying numbers of cells and unique fragments per cell were compensated by selecting 2,700 most similar cells from each sample or MEF subtype compared to a reference of 2,700 randomly selected high-quality cells from the 0 h IFNβ-treated epithelial-like MEFs (**Table 4.13**). *Single cell co-accessibility analysis* was performed with 2,700 bias-compensated cells for each sample and MEF subtype separately using RWireX. Co-accessibility scores were computed within 1 Mb using a continuous accessibility count matrix of extended union ATAC peaks and single cells. The resulting co-accessible links were filtered removing all links with negative co-accessibility scores, p-values below 0.01, and co-accessibility scores below sample-specific background co-accessibility cutoffs. The remaining links were considered the autonomous links of co-accessibility (ACs). Only ACs of STAT1/2 peaks were considered. ACs were

visualized in exemplary regions by loop tracks with genomic annotations and number of unique fragment-normalized pseudo-bulk chromatin accessibility tracks using ArchR. ACs between differentially expressed genes and STAT1/2 peaks were quantified and genes were classified successively by presence of (i) a STAT1/2 peak at the promoter, (ii) gained or (iii) lost AC between the promoter and a distal STAT1/2 peak after IFNβ treatment, and (iv) no link to STAT1/2 peak.

Table 4.13 Cells used for co-accessibility analysis of ESCs and MEFs. Epithelial-like MEF subtype is abbreviated as epi-like and mesenchymal-like MEF subtype as mes-like. Adapted from Muckenhuber *et al.* (2023).

		Sample (IFNβ treatment)		
		0 h	**1 h**	**6 h**
ESCs	Selected cell number	2,700	-	2,700
	Unique fragments/cell (median)	13,443	-	20,143
Epi-like MEFs	Selected cell number	2,700	2,700	2,700
	Unique fragments/cell (median)	13,452	13,463	6,557
Mes-like MEFs	Selected cell number	2,700	2,700	2,700
	Unique fragments/cell (median)	13,454	13,460	10,544

Metacell co-accessibility analysis was performed with 5,400 bias-compensated cells for ESCs and 8,100 for MEFs from all treatment time points of ESCs and MEF subtypes combined using RWireX. Metacells were formed from unique sets of 10 cells each, not using the final 10 % of cells to prevent the forced aggregation of dissimilar cells. Co-accessibility scores were computed within 2 Mb using a continuous accessibility count matrix of 10 kb genomic tiles and metacells. The *metacell co-accessibility matrices* were visualized in exemplary regions with annotated differentially expressed genes and STAT1/2 peaks using plotgardener.

4.5.5. Data and code availability

Data of scATAC-seq and scRNA-seq from ESCs and MEFs are available at GEO as described in **Table 4.14**. Supplementary data with intermediate results are available at Muckenhuber *et al.* (2023). My scripts for the computational analyses of scATAC-seq data from ESCs and MEFs are provided at https://github.com/RippeLab/RWire-IFN. *Single cell co-accessibility* analysis of scATAC-seq data from ESCs and MEFs was conducted with

RWireX (v0.2.05, https://github.com/RippeLab/RWire-IFN). *Metacell co-accessibility* analysis of scATAC-seq data from ESCs and MEFs was conducted with RWireX (v1.1.06, https://github.com/RippeLab/RWireX).

Table 4.14 Data availability of scATAC-seq and scRNA-seq from ESCs and MEFs. Data are available at GEO (https://www.ncbi.nlm.nih.gov/geo) as part of the series GSE160764.

Method	Samples	GEO ID
scATAC-seq	ESC, untreated	GSM4878888
	ESC, 6 h IFNβ treatment	GSM4878889
	MEF, untreated	GSM5852360
	MEF, 1 h IFNβ treatment	GSM5852361
	MEF, 6 h IFNβ treatment	GSM5852362
scRNA-seq	ESC, untreated	GSM4878890
	ESC, 1 h IFNβ treatment	GSM4878891
	ESC, 6 h IFNβ treatment	GSM4878892
	MEF, untreated	GSM8428156
	MEF, 1 h IFNβ treatment	GSM5852363
	MEF, 6 h IFNβ treatment	GSM5852364
	HUVEC, replicate 2, 240 min TNFα treatment	GSM5852365

4.5.6. List of applied software packages

Utilized software for the computational analyses of sequencing data of ESCs and MEFs is provided in **Table 4.15**.

Table 4.15 Software used for the analysis of bulk and single-cell sequencing data of ESCs and MEFs.

Software	Version	Reference
ArchR	v0.9.5	Granja *et al.* (2021)
Bowtie2	v2.3.3	Langmead & Salzberg (2012)
Cell Ranger	v3.0.2	Zheng *et al.* (2017)
Cell Ranger ATAC	v1.1.0	Satpathy *et al.* (2019)
DESeq2	v1.24.0	Love *et al.* (2014)
DiffBind	v2.12.0	Ross-Innes *et al.* (2012)
GenomicRanges	v1.36.4	Lawrence *et al.* (2013)
Ggplot2	v3.3.5	Wickham (2016)
MACS2	v2.1.2	Zhang *et al.* (2008)
Plotgardener	v1.6.2	Kramer *et al.* (2022)
R	v3.6.3 (bulk, scRNA) v4.0.2 (scATAC)	R Core Team (1993)
Rsem	v1.3.0	Li & Dewey (2011)
Seurat	v4.0.1	Stuart *et al.* (2019)
Star	v2.5.3a	Dobin *et al.* (2013)

4.6. Bulk RNA-seq of NK cells after co-culture with HDV-infected hepatocytes

4.6.1. Sequencing data acquisition

Experiments and data acquisition were performed by Christopher Groth (Department of Immunobiochemistry, Mannheim Institute for Innate Immunoscience and Medical Faculty Mannheim, Heidelberg University, Germany) and Markus Muckenhuber as described in Groth *et al.* (2023). Briefly, HepG2 cells overexpressing NTCP (HepG2-hNTCP; Lempp *et al.* (2019)) were cultured in transfection medium with or without HDV for 24 h. HepG2-hNTCP cells were washed and further cultured for 4 days. Following, HepG2-hNTCP cells were either cultured alone for 24 h and supernatant was collected or co-cultured with NK cells in a 4:1 ratio for 48 h. IFNγ expression frequency of NK cells treated with supernatant from non-infected or HDV-infected HepG2-hNTCP cells for 24h was measured by flow

cytometry analysis. NK cells from HepG2-hNTCP cell co-culture were isolated using magnetic beads. RNA was isolated using the RNeasy Mini kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. Bulk RNA libraries were prepared using the TruSeq Stranded Total RNA Gold with Ribo-Zero Plus kit according to the manufacturer's protocol (Illumina, San Diego, USA). Libraries were sequenced paired end with 50 bp for both read 1 and read 2 on NovaSeq 6000 system (Illumina, San Diego, USA). Sequencing was conducted by the DKFZ NGS Core Facility.

4.6.2. Analysis of bulk RNA-seq data

I processed bulk RNA-seq data and Carsten Sticht (Medical Faculty Mannheim, Heidelberg University, Germany) performed differential expression analysis as described here and in Groth *et al.* (2023). Sequencing reads were aligned with Star (default parameters, except sjbdOverhang of 200) using a Star index from the 1000 genomes assembly. Duplicate reads were identified using Sambamba and quality control was performed with Samtools. Reads 1 and 2 were both used for strand-unspecific quantification over exon features from gencode 19 gene models using FeatureCounts (default parameters, except quality threshold of 255). Further data analysis was conducted with systempipeR and visualized with ggplot2 in R. Count data was normalized using voom. Differential expression analysis was performed using limma (default parameters, significance threshold: FDR < 0.05).

4.6.3. List of applied software packages

Utilized software for the computational analyses of bulk RNA-seq data of NK cells is provided in **Table 4.16**.

Table 4.16 Software used for the analysis of bulk RNA-seq data of NK cells. Not available software versions are indicated as n.a..

Software	Version	Reference
FeatureCounts	n.a.	Liao *et al.* (2014)
Ggplot2	v2.2.1	Wickham (2016)
Limma	n.a.	Ritchie *et al.* (2015)
R	n.a.	R Core Team (1993)
Sambamba	v0.6.5	Tarasov *et al.* (2015)

Samtools	v1.6	Danecek *et al.* (2021)
Star	v2.5.3a	Dobin *et al.* (2013)
SystempipeR	n.a.	TW & Girke (2016)
Voom	n.a.	Law *et al.* (2014)

4.7. Sequencing of TCL1 cells and CLL patient samples

4.7.1. Sequencing data acquisition

Experiments and data acquisition were performed by Philipp Roessner (formerly Division of Molecular Genetics, German Cancer Research Center, Heidelberg, Germany) and Markus Muckenhuber as described in Roessner *et al.* (2024). Briefly, *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells were generated as described previously in Chakraborty *et al.* (2021) and transplanted in immunodeficient NOD scid gamma mice. Peripheral blood was drawn and sorted for TCL1 cells using the EasySep mouse pan-B cell isolation kit (Stemcell Technologies, Vancouver, Canada). Biological replicates were acquired from TCL1 cell transplantations into different mice. scRNA libraries were prepared using the Chromium Single Cell Multiome kit v1 (10x Genomics, Pleasanton, USA) according to the manufacturer's protocol. scTurboATAC libraries were prepared using the Chromium Single Cell ATAC kit v1.1 (10x Genomics, Pleasanton, USA) according to the TurboATAC protocol (Seufert *et al.*, 2023). scRNA libraries were sequenced paired-end with 28 bp and 90 bp for read 1 and read 2 on a NovaSeq 6000 system (Illumina, San Diego, USA) using SP flow cells. scATAC libraries were sequenced paired-end with 50 bp for both read 1 and read 2 on a NovaSeq 6000 system (Illumina, San Diego, USA) using S1 flow cells. Sequencing was conducted by the DKFZ NGS Core Facility.

RNA was isolated using the QIAshredder and RNeasy Mini kit (both Qiagen, Hilden, Germany) according to the manufacturer's protocol. Bulk RNA libraries were prepared using the TruSeq Stranded mRNA kit (Illumina, San Diego, USA) according to the manufacturer's protocol. Libraries were sequenced paired-end with 51 bp for both read 1 and read 2 on a NovaSeq 6000 system (Illumina, San Diego, USA). Sequencing was conducted by the DKFZ NGS Core Facility.

4.7.2. Analysis of bulk sequencing data

The analysis of bulk RNA-seq data was performed by Marc Zapatka (Division of Molecular Genetics, German Cancer Research Center, Heidelberg, Germany), while the analysis of bulk ATAC-seq and bulk RNA-seq data from Beekman *et al.* (2018) was performed by Vincente Chapaprieta (Instituto de Investigaciones Biomédicas August Pi i Sunyer, Barcelona, Spain) as described in Roessner *et al.* (2024). In short, bulk RNA-seq data of TCL1 samples were mapped to the mouse GRCm38/mm10 reference including the Illumina spike-in PhiX174 sequence using Star. Expression was quantified in transcripts using Subread. Transcripts with less than 10 counts in total were removed. Differential expression analysis was performed between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 samples using DESeq2 (significance threshold: FDR < 0.05). Log2 fold changes were obtained using apeglm (LFC shrinkage). Sample 804 was identified as outlier from clustering of raw counts (rlog transformed).

Bulk RNA-seq data of 260 CLL patient samples were obtained from Puente *et al.* (2015) and Nadeu *et al.* (2021). Gene counts were normalized using DESeq2 (variance stabilization transformation). Genes with more than 9 counts and more than 0 transcript counts per million kb in at least 22 samples were included for further analysis. CLL patient samples were classified by their *TBX21* expression into *TBX21*^{low} (bottom quartile, n = 65) and *TBX21*^{high} (top quartile, n = 65) groups. Differential expression analysis was performed between *TBX21*^{low} and *TBX21*^{high} CLL patient samples using DESeq2 and including intermediate cases (n = 130; significance threshold: FDR < 0.05). Effect sizes were reduced using apeglm.

Bulk RNA-seq data of 7 CLL patient samples and 15 non-malignant B cell samples were obtained from Beekman *et al.* (2018). The non-malignant samples contained B cell subpopulations from peripheral blood and tonsils, such as naïve B cells, germinal center B cells, memory B cells, and plasma cells. Gene counts were normalized using DESeq2 (variance stabilization transformation).

Bulk ATAC-seq data of 99 CLL patient samples were obtained from Beekman *et al.* (2018). CLL patient samples were classified by their H3K27ac signal in the TBX21 promoter (chr17:47731961-47738364) from paired H3K27ac ChIP-seq data into *TBX21*^{low} (bottom quartile, n = 25) and *TBX21*^{high} (top quartile, n = 25) groups. ATAC peaks with accessibility signal in at least 5 samples from *TBX21*^{low} and *TBX21*^{high} groups each were included for further analysis. Differential accessibility analysis was performed between *TBX21*^{low} and *TBX21*^{high} CLL patient samples using DESeq2 (including IGHV mutational status as co-

factor and 49 intermediate *TBX21* cases; significance threshold: FDR < 0.05). Effect sizes were reduced using apeglm. TF binding sites in ATAC peaks were annotated using Homer motifs from chromVARmotifs. The difference of binding site deviations between *TBX21*^{low} and *TBX21*^{high} CLL patient samples were calculated for each TF using chromVAR.

4.7.3. Identification of T-bet dependent genes

Mass spectrometry data of TCL1 were acquired by Philipp Roessner as described in Roessner *et al.* (2024). Mass spectrometry data of 68 CLL patient samples were obtained from Herbst *et al.* (2022). The analysis of mass spectrometry data was performed by Pavle Boskovic (Division of Molecular Genetics, German Cancer Research Center, Heidelberg, Germany) as described in Roessner *et al.* (2024). T-bet protein levels were correlated to all other proteins (Pearson and Spearman correlation). P-values were adjusted for multiple testing by Benjamini-Hochberg method (significance threshold: adjusted p-value < 0.05).

Integrated analysis of bulk RNA-seq and mass spectrometry data from TCL1 samples and CLL patient samples was performed by Philipp Roessner as described here and in Roessner *et al.* (2024). The overlap between significantly differential genes and significantly correlating proteins from both human and murine data sets was determined. Not detected or quantified proteins were considered as overlapping with bulk RNA-seq data. The overlapping genes of all four data sets were termed T-bet dependent genes.

4.7.4. Analysis of scRNA-seq data

I performed the analysis of scRNA-seq data from TCL1 samples as described here and in Roessner *et al.* (2024). Data were processed with Cell Ranger count (10x Genomics, Pleasanton, USA) using the provided mouse mm10 reference (default parameters, except including introns, ARC-v1 chemistry). Quality metrics of scRNA-seq data are provided in **Table 4.17**. Data were further analyzed with Seurat and visualized with ggplot2 in R. High-quality barcodes were selected using (i) a minimal and maximal threshold for number of detected genes from the 5th and 99th percentiles (140 and 3,491, respectively), (ii) a minimal and maximal threshold for UMI counts of 1,000 and 9,244 (99th percentile), and (iii) a maximal threshold for percentage of mitochondrial UMI counts of 40 and 50 for *Tbx21*$^{+/+}$ and *Tbx21*$^{-/-}$ TCL1 samples, respectively. Barcodes containing multiple cells were removed using DoubletFinder (default parameters). High-quality cell numbers are provided in **Table 4.17**.

Table 4.17 Quality metrics of scRNA-seq data from TCL1 cells. Barcode number is the number of detected cell barcodes by Cell Ranger. High-quality cell number is the number of remaining cells after quality filtering.

		Rep1	Rep2
Tbx21$^{+/+}$ TCL1	Barcode number	55,126	55,018
	Number of high-quality cells	2,242	3,824
	UMI counts per cell (median)	4,510	3,244
	Genes per cell (median)	2,004	1,682
	Mitochondrial UMI counts (%, median)	21	16
Tbx21$^{-/-}$ TCL1	Barcode number	33,224	43,680
	Number of high-quality cells	3,814	3,336
	UMI counts per cell (median)	2,842	3,385
	Genes per cell (median)	1,403	1,615
	Mitochondrial UMI counts (%, median)	28	24

Samples were merged, log normalized, and scaled (default parameters, except regressing out number of detected genes per cell). Cell cycle phases were inferred per single cell by module scores from the expression of cell cycle marker genes (Kowalczyk *et al.*, 2015). Low-dimensional single-cell embedding was computed using scaling (default parameters, except regressing out number of detected genes and cell cycle phase per cell), PCA (default parameters), and UMAP (default parameters, except PCs 1-17). Cell clusters were computed by SNN modularity optimization (default parameters, except PCs 1-17, resolution of 0.2). *Cd5* and *Cd19* marker gene expression was used to identify malignant cell clusters. Non-malignant cell clusters were removed from further analysis.

4.7.5. Analysis of scTurboATAC-seq data

I performed the analysis of scTurboATAC-seq data from TCL1 samples as described here and in Roessner *et al.* (2024). Data were processed with Cell Ranger ATAC count (10x Genomics, Pleasanton, USA) using the provided mouse mm10 reference (default parameters). Quality metrics of scATAC-seq data are provided in **Table 4.18**. Data were further analyzed with ArchR and visualized with ggplot2 in R. High-quality cell barcodes were selected using (i) a minimal threshold for number of unique fragments of $10^{4.25}$ and (ii) a minimal threshold for TSS enrichment score of 5. High-quality cell numbers are

provided in **Table 4.18**. Samples were merged and a low-dimensional single-cell embedding was computed using an accessibility count matrix of 500 bp genomic tiles, iterative LSI (default parameters) and UMAP (default parameters, except LSI components 2, 4-17). Cell clusters were computed by SNN modularity optimization (default parameters, except LSI components 2, 4-17; resolution of 0.15). ATAC gene activity scores of *Cd5* and *Cd19* markers were used to identify malignant cell clusters. Non-malignant cell clusters were removed from further analysis.

Table 4.18 Quality metrics of scTurboATAC-seq data from TCL1 cells. Barcode number is the number of detected cell barcodes by Cell Ranger. High-quality cell number is the number of remaining cells after quality filtering.

		Rep1	Rep2
***Tbx21*$^{+/+}$ TCL1**	Barcode number	15,940	31,155
	Number of high-quality cells	1,966	2,569
	Unique fragments/cell (median)	48,977	58,884
	TSS enrichment score (median)	16.16	22.01
***Tbx21*$^{-/-}$ TCL1**	Barcode number	37,616	23,206
	Number of high-quality cells	2,299	2,260
	Unique fragments/cell (median)	54,954	69,183
	TSS enrichment score (median)	24.39	18.9

Sample-specific peaks from pseudo-bulk chromatin accessibility data were called using MACS2 in ArchR (default parameters, except peak summit extension by 1,000 bp to each side; reproducibility of 1). A union ATAC peak set across all samples was generated. Differential accessibility analysis for union ATAC peaks was performed between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 cells across all replicates using Wilcoxon test in ArchR (default parameters, except a maximum of 4,000 cells per sample, bias correction by TSS enrichment score and number of unique fragments (log10), normalization by number of unique fragments; significance threshold: FDR ≤ 0.05, absolute log2 fold change ≥ 1). TF binding sites in union ATAC peaks were annotated using Homer motifs from chromVARmotifs. Union ATAC peaks with T-bet binding sites were termed *T-bet peaks*. Enrichment of TF binding sites in significantly differential accessible ATAC peaks between *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 were determined by hypergeometric test in ArchR (significance threshold: FDR ≤ 0.1, log2 fold change ≥ 0.5).

Cell barcodes containing multiple cells were removed using Amulet from scDblFinder. Technical biases between samples from varying numbers of cells and unique fragments per cell were compensated by selecting 1,000 most similar cells from each sample compared to a reference of 1,000 randomly selected high-quality cells from the *Tbx21*$^{-/-}$ TCL1 Rep1 sample. *Single cell co-accessibility analysis* was performed with 1,000 bias-compensated cells for each sample separately using RWireX. Co-accessibility scores were computed within 1 Mb using a continuous accessibility count matrix of union ATAC peaks and single cells. The resulting co-accessible links were filtered removing all links with negative co-accessibility scores, percent accessible cells (PAC) below 5, and co-accessibility scores below sample-specific background co-accessibility cutoffs. The remaining links were considered the autonomous links of co-accessibility (ACs). Only ACs of union ATAC peaks at T-bet dependent gene promoters were considered. ACs from biological replicates were merged and visualized in exemplary regions by loop tracks with genomic annotations and number of unique fragment-normalized pseudo-bulk chromatin accessibility tracks using ArchR. ACs between T-bet dependent genes and *T-bet peaks* were quantified and genes were classified successively by presence of (i) a *T-bet peak* at the promoter, (ii) AC between the promoter and a distal *T-bet peak*, and (iii) no link to *T-bet peak*.

Metacell co-accessibility analysis was performed with 4,000 bias-compensated cells from all *Tbx21*$^{-/-}$ and *Tbx21*$^{+/+}$ TCL1 samples combined using RWireX. Metacells were formed from unique sets of 10 cells each, not using the final 10 % of cells to prevent the forced aggregation of dissimilar cells. Co-accessibility scores were computed within 2 Mb using a continuous accessibility count matrix of 10 kb genomic tiles and metacells. The *metacell co-accessibility matrices* were visualized in exemplary regions with annotated T-bet dependent genes and T-bet motifs using plotgardener.

4.7.6. Data and code availability

Data of scTurboATAC-seq and scRNA-seq from TCL1 cells are available at GEO as described in **Table 4.19**. Supplementary data with intermediate results are available at Roessner *et al.* (2024). *Single cell co-accessibility* analysis of scTurboATAC-seq data from TCL1 cells was conducted with RWireX (v0.2.05, https://github.com/RippeLab/RWire-IFN). *Metacell co-accessibility* analysis of scTurboATAC-seq data from TCL1 cells was conducted with RWireX (v1.1.06, https://github.com/RippeLab/RWireX).

Table 4.19 Data availability of scTurboATAC-seq and scRNA-seq from TCL1 cells. Data are available at GEO (https://www.ncbi.nlm.nih.gov/geo) as part of the series GSE234226.

Method	Samples	GEO ID
scTurboATAC-seq	TCL1, *Tbx21*$^{+/+}$, replicate 1	GSM7457615
	TCL1, *Tbx21*$^{+/+}$, replicate 2	GSM7457616
	TCL1, *Tbx21*$^{-/-}$, replicate 1	GSM7457617
	TCL1, *Tbx21*$^{-/-}$, replicate 2	GSM7457618
scRNA-seq	TCL1, *Tbx21*$^{+/+}$, replicate 1	GSM7457619
	TCL1, *Tbx21*$^{+/+}$, replicate 2	GSM7457620
	TCL1, *Tbx21*$^{-/-}$, replicate 1	GSM7457621
	TCL1, *Tbx21*$^{-/-}$, replicate 2	GSM7457622

4.7.7. List of applied software packages

Utilized software for the computational analyses of sequencing data of TCL1 samples and CLL patient samples is provided in **Table 4.20**.

Table 4.20 Software used for the analysis of sequencing data of TCL1 cells and CLL patient samples. Not available software versions are indicated as n.a..

Software	Version	Reference
Apeglm	n.a.	Zhu *et al.* (2019)
ArchR	v1.0.3	Granja *et al.* (2021)
Cell Ranger	v5.0.0	Zheng *et al.* (2017)
Cell Ranger ATAC	v2.0.0	Satpathy *et al.* (2019)
ChromVar	v1.18.0	Schep *et al.* (2017)
ChromVarMotifs	v0.2.0	Schep *et al.* (2017)
DESeq2	v1.24.0	Love *et al.* (2014)
DoubletFinder	v2.0.3	McGinnis *et al.* (2019)

Ggplot2	v3.3.6 (bulk data) v3.4.0 (sc data)	Wickham (2016)
MACS2	v2.1.2	Zhang *et al.* (2008)
Plotgardener	v1.6.2	Kramer *et al.* (2022)
R	v4.1.3	R Core Team (1993)
scDblFinder	v1.12.0	Germain *et al.* (2021)
Seurat	v4.2.0	Stuart *et al.* (2019)
Star	v2.5.3a	Dobin *et al.* (2013)
Subread	v1.5.1	Liao *et al.* (2013)

4.8. Thesis writing

The thesis was written in Microsoft Word. Microsoft Word and ChatGPT were used to correct grammar and spelling, suggest synonyms for repetitive words, and simplify complex and long sentences.

Bibliography

Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* 48: 838-847

Au-Yeung N, Horvath CM (2018) Transcriptional and chromatin regulation in interferon and innate antiviral gene expression. *Cytokine Growth Factor Rev* 44: 11-17

Badia IMP, Wessels L, Muller-Dott S, Trimbour R, Ramirez Flores RO, Argelaguet R, Saez-Rodriguez J (2023) Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 24: 739-754

Barcenas-Walls JR, Ansaloni F, Herve B, Strandback E, Nyman T, Castelo-Branco G, Bartosovic M (2024) Nano-CUT&Tag for multimodal chromatin profiling at single-cell resolution. *Nat Protoc* 19: 791-830

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823-837

Bartman CR, Hamagami N, Keller CA, Giardine B, Hardison RC, Blobel GA, Raj A (2019) Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Mol Cell* 73: 519-532 e514

Baysoy A, Bai Z, Satija R, Fan R (2023) The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 24: 695-713

Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 37: 38-44

Beekman R, Chapaprieta V, Russinol N, Vilarrasa-Blasi R, Verdaguer-Dot N, Martens JHA, Duran-Ferrer M, Kulis M, Serra F, Javierre BM *et al* (2018) The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat Med* 24: 868-880

Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T *et al* (2020) ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* 11: 4267

Berest I, Tangherloni A (2023) Integration of scATAC-Seq with scRNA-Seq Data. *Methods Mol Biol* 2584: 293-310

Berg OG, von Hippel PH (1985) Diffusion-controlled macromolecular interactions. *Annu Rev Biophys Biophys Chem* 14: 131-160

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR *et al* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120: 169-181

Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST (2012) Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* 150: 279-290

Bolen CR, Ding S, Robek MD, Kleinstein SH (2014) Dynamic expression profiling of type I and type III interferon-stimulated hepatocytes reveals a stable hierarchy of gene expression. *Hepatology* 59: 1262-1272

Bouhedda F, Autour A, Ryckelynck M (2017) Light-Up RNA Aptamers and Their Cognate Fluorogens: From Their Development to Their Applications. *Int J Mol Sci* 19

Bouwman BA, Crosetto N, Bienko M (2023) A GC-centered view of 3D genome organization. *Curr Opin Genet Dev* 78: 102020

Bresin A, D'Abundo L, Narducci MG, Fiorenza MT, Croce CM, Negrini M, Russo G (2016) TCL1 transgenic mouse model as a tool for the study of therapeutic targets and microenvironment in human B-cell chronic lymphocytic leukemia. *Cell Death Dis* 7: e2071

Brodsky S, Jana T, Mittelman K, Chapal M, Kumar DK, Carmi M, Barkai N (2020) Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol Cell* 79: 459-471 e454

Bruckner DB, Chen H, Barinov L, Zoller B, Gregor T (2023) Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome. *Science* 380: 1357-1362

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10: 1213-1218

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523: 486-490

Bulger M, Groudine M (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144: 327-339

Bundschuh C, Weidner N, Klein J, Rausch T, Azevedo N, Telzerow A, Mallm JP, Kim H, Steiger S, Seufert I *et al* (2024) Evolution of SARS-CoV-2 in the Rhine-Neckar/Heidelberg Region 01/2021 - 07/2023. *Infect Genet Evol*: 105577

Carpenter S, Ricci EP, Mercier BC, Moore MJ, Fitzgerald KA (2014) Post-transcriptional regulation of gene expression in innate immunity. *Nat Rev Immunol* 14: 361-376

Chakraborty S, Martines C, Porro F, Fortunati I, Bonato A, Dimishkovska M, Piazza S, Yadav BS, Innocenti I, Fazio R *et al* (2021) B-cell receptor signaling and genetic lesions in TP53 and CDKN2A/CDKN2B cooperate in Richter transformation. *Blood* 138: 1053-1066

Chan B, Rubinstein M (2023) Theory of chromatin organization maintained by active loop extrusion. *P Natl Acad Sci USA* 120: e2222078120

Chaumeil J, Micsinai M, Skok JA (2013) Combined immunofluorescence and DNA FISH on 3D-preserved interphase nuclei to study changes in 3D nuclear organization. *J Vis Exp*: e50087

Chen Y, Cattoglio C, Dailey GM, Zhu Q, Tjian R, Darzacq X (2022) Mechanisms governing target search and binding dynamics of hypoxia-inducible factors. *Elife* 11: e75064

Choi KJ, Quan MD, Qi C, Lee JH, Tsoi PS, Zahabiyon M, Bajic A, Hu L, Prasad BVV, Liao SJ *et al* (2022) NANOG prion-like assembly mediates DNA bridging to facilitate chromatin reorganization and activation of pluripotency. *Nat Cell Biol* 24: 737-747

Chong S, Graham TGW, Dugast-Darzacq C, Dailey GM, Darzacq X, Tjian R (2022) Tuning levels of low-complexity domain interactions to modulate endogenous oncogenic transcription. *Mol Cell* 82: 2084-2097 e2085

Cinar O, Viechtbauer W (2022) The poolr Package for Combining Independent and Dependent p Values. *Journal of Statistical Software* 101

Cramer P (2019) Organization and regulation of gene transcription. *Nature* 573: 45-54

Cresswell KG, Stansfield JC, Dozmorov MG (2020) SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics* 21: 319

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA *et al* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *P Natl Acad Sci USA* 107: 21931-21936

Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, Müller K (2024) igraph: Network Analysis and Visualization in R. URL https://CRAN.R-project.org/package=igraph

Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348: 910-914

Dahal L, Walther N, Tjian R, Darzacq X, Graham TGW (2023) Single-molecule tracking (SMT): a window into live-cell transcription biochemistry. *Biochem Soc Trans* 51: 557-569

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM *et al* (2021) Twelve years of SAMtools and BCFtools. *Gigascience* 10

Dang Y, Yadav RP, Chen X (2023) ATAC-See: A Tn5 Transposase-Mediated Assay for Detection of Chromatin Accessibility with Imaging. *Methods Mol Biol* 2611: 285-291

de Laat W, Grosveld F (2003) Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* 11: 447-459

De Rop FV, Hulselmans G, Flerin C, Soler-Vila P, Rafels A, Christiaens V, Gonzalez-Blas CB, Marchese D, Caratu G, Poovathingal S *et al* (2024) Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nat Biotechnol* 42: 916-926

de Veer MJ, Holko M, Frevel M, Walker E, Der S, Paranjape JM, Silverman RH, Williams BRG (2001) Functional classification of interferon-stimulated genes identified using microarrays. *Journal of Leukocyte Biology* 69: 912-920

de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. *Genes & development* 26: 11-24

Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295: 1306-1311

Der SD, Zhou A, Williams BR, Silverman RH (1998) Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *P Natl Acad Sci USA* 95: 15623-15628

Devin A, Cook A, Lin Y, Rodriguez Y, Kelliher M, Liu Z (2000) The distinct roles of TRAF2 and RIP in IKK activation by TNF-R1: TRAF2 recruits IKK to TNF-R1 while RIP mediates IKK activation. *Immunity* 12: 419-429

Diermeier S, Kolovos P, Heizinger L, Schwartz U, Georgomanolis T, Zirkel A, Wedemann G, Grosveld F, Knoch TA, Merkl R *et al* (2014) TNFalpha signalling primes chromatin for NF-kappaB binding and induces rapid and widespread nucleosome repositioning. *Genome Biology* 15: 536

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R *et al* (2016) Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167: 1853-1866 e1817

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485: 376-380

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21

Doench JG (2018) Am I ready for CRISPR? A user's guide to genetic screens. *Nat Rev Genet* 19: 67-80

Du M, Stitzinger SH, Spille JH, Cho WK, Lee C, Hijaz M, Quintana A, Cisse, II (2024) Direct observation of a condensate effect on super-enhancer controlled gene bursting. *Cell* 187: 331-344 e317

Duan Z, Xu S, Sai Srinivasan S, Hwang A, Lee CY, Yue F, Gerstein M, Luan Y, Girgenti M, Zhang J (2024) scENCORE: leveraging single-cell epigenetic data to predict chromatin conformation using graph embedding. *Brief Bioinform* 25

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* 3: 95-98

Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74

Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D (2012) qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software* 48: 1-18

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43-49

Esadze A, Stivers JT (2018) Facilitated Diffusion Mechanisms in DNA Base Excision Repair and Transcriptional Activation. *Chem Rev* 118: 11298-11323

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 38: 276-278

Ferrie JJ, Karr JP, Tjian R, Darzacq X (2022) "Structure"-function relationships in eukaryotic transcription factors: The role of intrinsically disordered regions in gene regulation. *Mol Cell* 82: 3970-3984

Flyamer IM, Illingworth RS, Bickmore WA (2020) Coolpup.py: versatile pile-up analysis of Hi-C data. *Bioinformatics* 36: 2980-2985

Foundation PS, 1991. Python Language Reference. p. URL http://www.python.org.

Fraga MF, Esteller M (2002) DNA methylation: a profile of methods and applications. *Biotechniques* 33: 632, 634, 636-649

Friedman MJ, Wagner T, Lee H, Rosenfeld MG, Oh S (2024) Enhancer-promoter specificity in gene transcription: molecular mechanisms and disease associations. *Exp Mol Med* 56: 772-787

Fukaya T, Lim B, Levine M (2016) Enhancer Control of Transcriptional Bursting. *Cell* 166: 358-368

Gabriele M, Brandao HB, Grosse-Holz S, Jha A, Dailey GM, Cattoglio C, Hsieh TS, Mirny L, Zechner C, Hansen AS (2022) Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science* 376: 496-501

Galluzzi L, Vitale I, Aaronson SA, Abrams JM, Adam D, Agostinis P, Alnemri ES, Altucci L, Amelio I, Andrews DW *et al* (2018) Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018. *Cell Death Differ* 25: 486-541

Gamarra N, Narlikar GJ (2021) Collaboration through chromatin: motors of transcription and chromatin structure. *J Mol Biol* 433: 166876

Garcia DA, Johnson TA, Presman DM, Fettweis G, Wagh K, Rinaldi L, Stavreva DA, Paakinaho V, Jensen RAM, Mandrup S *et al* (2021) An intrinsically disordered region-mediated confinement state contributes to the dynamics and function of transcription factors. *Mol Cell* 81: 1484-1498 e1486

Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* 29: 1363-1375

Germain PL, Lun A, Garcia Meixide C, Macnair W, Robinson MD (2021) Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res* 10: 979

Ghasemi DR, Okonechnikov K, Rademacher A, Tirier S, Maass KK, Schumacher H, Joshi P, Gold MP, Sundheimer J, Statz B *et al* (2024) Compartments in medulloblastoma with extensive nodularity are connected through differentiation along the granular precursor lineage. *Nat Commun* 15: 269

Gholamalamdari O, van Schaik T, Wang Y, Kumar P, Zhang L, Zhang Y, Gonzalez GAH, Vouzas AE, Zhao PA, Gilbert DM *et al* (2024) Beyond A and B Compartments: how major nuclear locales define nuclear genome organization and function. *bioRxiv*

Goel VY, Huseyin MK, Hansen AS (2023) Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nat Genet* 55: 1048-1056

Gourisankar S, Krokhotin A, Wenderski W, Crabtree GR (2024) Context-specific functions of chromatin remodellers in development and disease. *Nat Rev Genet* 25: 340-361

Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, Greenleaf WJ (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* 53: 403-411

Greenberg MVC, Bourc'his D (2019) The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 20: 590-607

Grosveld F, van Staalduinen J, Stadhouders R (2021) Transcriptional Regulation by (Super)Enhancers: From Discovery to Mechanisms. *Annu Rev Genomics Hum Genet* 22: 127-146

Groth C, Maric J, Garcés Lázaro I, Hofman T, Zhang Z, Ni Y, Keller F, Seufert I, Hofmann M, Neumann-Haefelin C *et al* (2023) Hepatitis D infection induces IFN-β-mediated NK cell activation and TRAIL-dependent cytotoxicity. *Front Immunol* 14: 1287367

Grun D (2020) Revealing dynamics of gene expression variability in cell state space. *Nat Methods* 17: 45-49

Guo YL, Carmichael GG, Wang R, Hong X, Acharya D, Huang F, Bai F (2015) Attenuated Innate Immunity in Embryonic Stem Cells and Its Implications in Developmental Biology and Regenerative Medicine. *Stem Cells* 33: 3165-3173

Haberle V, Stark A (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* 19: 621-637

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

Hafner A, Park M, Berger SE, Murphy SE, Nora EP, Boettiger AN (2023) Loop stacking organizes genome folding from TADs to chromosomes. *Mol Cell* 83: 1377-1392 e1376

Hamilton NE, Ferry M (2018) ggtern: Ternary Diagrams Using ggplot2. *Journal of Statistical Software* 87

Han SJ, Jiang YL, You LL, Shen LQ, Wu X, Yang F, Cui N, Kong WW, Sun H, Zhou K *et al* (2024) DNA looping mediates cooperative transcription activation. *Nat Struct Mol Biol* 31: 293-299

Hansen AS, Cattoglio C, Darzacq X, Tjian R (2018) Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* 9: 20-32

Hayashi-Takanaka Y, Yamagata K, Nozaki N, Kimura H (2009) Visualizing histone modifications in living cells: spatiotemporal dynamics of H3 phosphorylation during interphase. *The Journal of Cell Biology* 187: 781-790

He J, Huo X, Pei G, Jia Z, Yan Y, Yu J, Qu H, Xie Y, Yuan J, Zheng Y *et al* (2024) Dual-role transcription factors stabilize intermediate expression levels. *Cell* 187: 2746-2766 e2725

Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, Besir H, Remans K, Steinmetz LM (2018) Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3 (Bethesda)* 8: 79-89

Herbst SA, Vesterlund M, Helmboldt AJ, Jafari R, Siavelis I, Stahl M, Schitter EC, Liebers N, Brinkmann BJ, Czernilofsky F *et al* (2022) Proteogenomics refines the molecular classification of chronic lymphocytic leukemia. *Nat Commun* 13: 6226

Hess JF, Kohl TA, Kotrova M, Ronsch K, Paprotka T, Mohr V, Hutzenlaub T, Bruggemann M, Zengerle R, Niemann S *et al* (2020) Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol Adv* 41: 107537

Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, Lucken MD, Strobl DC, Henao J, Curion F *et al* (2023) Best practices for single-cell analysis across modalities. *Nat Rev Genet* 24: 550-572

Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA (2017) A Phase Separation Model for Transcriptional Control. *Cell* 169: 13-23

Hodson DJ, Shaffer AL, Xiao W, Wright GW, Schmitz R, Phelan JD, Yang Y, Webster DE, Rui L, Kohlhammer H *et al* (2016) Regulation of normal B-cell differentiation and malignant B-cell survival by OCT2. *P Natl Acad Sci USA* 113: E2039-2046

Hsieh TH, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* 162: 108-119

Hsu H, Xiong J, Goeddel DV (1995) The TNF receptor 1-associated protein TRADD signals cell death and NF-kappa B activation. *Cell* 81: 495-504

Hubner MR, Eckersley-Maslin MA, Spector DL (2013) Chromatin organization and transcriptional regulation. *Curr Opin Genet Dev* 23: 89-95

Hung TC, Kingsley DM, Boettiger AN (2024) Boundary stacking interactions enable cross-TAD enhancer-promoter communication during limb development. *Nat Genet* 56: 306-314

Hwang DW, Maekiniemi A, Singer RH, Sato H (2024) Real-time single-molecule imaging of transcriptional regulatory networks in living cells. *Nat Rev Genet* 25: 272-285

Ibrahim DM (2024) Enhancer contacts during embryonic development show diverse interaction modes and modest yet significant increases upon gene activation. *Nat Genet* 56: 558-560

Ivashkiv LB, Donlin LT (2014) Regulation of type I interferon responses. *Nat Rev Immunol* 14: 36-49

Jana T, Brodsky S, Barkai N (2021) Speed-Specificity Trade-Offs in the Transcription Factors Search for Their Genomic Binding Sites. *Trends Genet* 37: 421-432

Jenuwein T, Allis CD (2001) Translating the histone code. *Science* 293: 1074-1080

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503: 290-294

Juven-Gershon T, Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* 339: 225-229

Kang Y, Kim YW, Kang J, Kim A (2021) Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. *FASEB J* 35: e21781

Karin M, Ben-Neriah Y (2000) Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity. *Annu Rev Immunol* 18: 621-663

Karlsson M, Zhang C, Mear L, Zhong W, Digre A, Katona B, Sjostedt E, Butler L, Odeberg J, Dusart P *et al* (2021) A single-cell type transcriptomics map of human tissues. *Sci Adv* 7

Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L (2012) H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* 13: 424

Karr JP, Ferrie JJ, Tjian R, Darzacq X (2022) The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer-promoter communication. *Genes Dev* 36: 7-16

Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* 10: 1930

Kent S, Brown K, Yang CH, Alsaihati N, Tian C, Wang H, Ren X (2020) Phase-Separated Transcriptional Condensates Accelerate Target-Search Process Revealed by Live-Cell Single-Molecule Imaging. *Cell Rep* 33: 108248

Kim S, Wysocka J (2023) Deciphering the multi-scale, quantitative cis-regulatory code. *Mol Cell* 83: 373-392

Kishi JY, Lapan SW, Beliveau BJ, West ER, Zhu A, Sasaki HM, Saka SK, Wang Y, Cepko CL, Yin P (2019) SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat Methods* 16: 533-544

Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 9: 72-74

Klenova EM, Morse HC, 3rd, Ohlsson R, Lobanenkov VV (2002) The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer. *Semin Cancer Biol* 12: 399-414

Ko MS (1991) A stochastic model for gene induction. *J Theor Biol* 153: 181-194

Kolde R (2018) pheatmap: Pretty Heatmaps. URL https://github.com/raivokolde/pheatmap

Kolovos P, Georgomanolis T, Koeferle A, Larkin JD, Brant L, Nikolicc M, Gusmao EG, Zirkel A, Knoch TA, van Ijcken WF *et al* (2016) Binding of nuclear factor kappaB to noncanonical consensus sites reveals its multimodal role during the early inflammatory response. *Genome Res* 26: 1478-1489

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16: 1289-1296

Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128: 693-705

Kowalczyk MS, Tirosh I, Heckl D, Nageswara Rao T, Dixit A, Haas BJ, Schneider R, Wagers AJ, Ebert BL, Regev A (2015) Single cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research*: gr.192237.192115

Kramer NE, Davis ES, Wenger CD, Deoudes EM, Parker SM, Love MI, Phanstiel DH (2022) Plotgardener: cultivating precise multi-panel figures in R. *Bioinformatics* 38: 2042-2045

Kumasaka N, Rostom R, Huang N, Polanski K, Meyer KB, Patel S, Boyd R, Gomez C, Barnett SN, Panousis NI *et al* (2023) Mapping interindividual dynamics of innate immune response at single-cell resolution. *Nat Genet* 55: 1066-1075

Kwak H, Lis JT (2013) Control of transcriptional elongation. *Annu Rev Genet* 47: 483-508

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357-359

Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, Segerstolpe A, Rivera CM, Ren B, Sandberg R (2019) Genomic encoding of transcriptional burst kinetics. *Nature* 565: 251-254

Law CW, Chen Y, Shi W, Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15: R29

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol* 9: e1003118

Lawrence T (2009) The nuclear factor NF-kappaB pathway in inflammation. *Cold Spring Harb Perspect Biol* 1: a001651

Legler DF, Micheau O, Doucey MA, Tschopp J, Bron C (2003) Recruitment of TNF receptor 1 to lipid rafts is essential for TNFalpha-mediated NF-kappaB activation. *Immunity* 18: 655-664

Lempp FA, Schlund F, Rieble L, Nussbaum L, Link C, Zhang Z, Ni Y, Urban S (2019) Recapitulation of HDV infection in a fully permissive hepatoma cell line allows efficient drug evaluation. *Nat Commun* 10: 2265

Levy DE, Darnell JE, Jr. (2002) Stats: transcriptional control and biological impact. *Nat Rev Mol Cell Biol* 3: 651-662

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323

Li J, Hsu A, Hua Y, Wang G, Cheng L, Ochiai H, Yamamoto T, Pertsinidis A (2020) Single-gene imaging links genome topology, promoter-enhancer communication and transcription control. *Nat Struct Mol Biol* 27: 1032-1040

Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* 41: e108

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923-930

Lim PS, Shannon MF, Hardy K (2010) Epigenetic control of inducible gene expression in the immune system. *Epigenomics* 2: 775-795

Lin D, Hong P, Zhang S, Xu W, Jamal M, Yan K, Lei Y, Li L, Ruan Y, Fu ZF *et al* (2018) Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat Genet* 50: 754-763

Liu J, Zhu S, Hu W, Zhao X, Shan Q, Peng W, Xue HH (2023) CTCF mediates CD8+ effector differentiation through dynamic redistribution and genomic reorganization. *J Exp Med* 220

Liu L, Jin G, Zhou X (2015) Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res* 43: 3873-3885

Liu T, Zhang L, Joo D, Sun SC (2017) NF-kappaB signaling in inflammation. *Signal Transduct Target Ther* 2: 17023-

Lotfollahi M, Yuhan H, Theis FJ, Satija R (2024) The future of rapid and automated single-cell data analysis using reference mapping. *Cell* 187: 2343-2358

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15: 550

Lu F, Lionnet T (2021) Transcription Factor Dynamics. *Cold Spring Harb Perspect Biol* 13: a040949

Lu Y, Qu H, Qi D, Xu W, Liu S, Jin X, Song P, Guo Y, Jia Y, Wang X *et al* (2019) OCT4 maintains self-renewal and reverses senescence in human hair follicle mesenchymal stem cells through the downregulation of p21 by DNA methyltransferases. *Stem Cell Res Ther* 10: 28

Luo S, Germain PL, Robinson MD, von Meyenn F (2024) Benchmarking computational methods for single-cell chromatin data analysis. *Genome Biology* 25: 225

Mach P, Kos PI, Zhan Y, Cramard J, Gaudin S, Tunnermann J, Marchi E, Eglinger J, Zuin J, Kryzhanovska M *et al* (2022) Cohesin and CTCF control the dynamics of chromosome folding. *Nat Genet* 54: 1907-1918

Mahat DB, Tippens ND, Martin-Rufino JD, Waterton SK, Fu J, Blatt SE, Sharp PA (2024) Single-cell nascent RNA sequencing unveils coordinated global transcription. *Nature* 631: 216–223

Majumdar S, Gonder D, Koutsis B, Poncz M (1991) Characterization of the human beta-thromboglobulin gene. Comparison with the gene for platelet factor 4. *Journal of Biological Chemistry* 266: 5785-5789

Malik S, Roeder RG (2023) Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. *Nat Rev Genet* 24: 767-782

Mallm JP, Iskar M, Ishaque N, Klett LC, Kugler SJ, Muino JM, Teif VB, Poos AM, Grossmann S, Erdel F *et al* (2019) Linking aberrant chromatin features in chronic lymphocytic leukemia to transcription factor networks. *Mol Syst Biol* 15: e8339

Martens LD, Fischer DS, Yepez VA, Theis FJ, Gagneur J (2024) Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat Methods* 21: 28-31

Mateo LJ, Murphy SE, Hafner A, Cinquini IS, Walker CA, Boettiger AN (2019) Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* 568: 49-54

Mazzocca M, Fillot T, Loffreda A, Gnani D, Mazza D (2021) The needle and the haystack: single molecule tracking to probe the transcription factor search in eukaryotes. *Biochem Soc Trans* 49: 1121-1132

Mazzocca M, Loffreda A, Colombo E, Fillot T, Gnani D, Falletta P, Monteleone E, Capozi S, Bertrand E, Legube G *et al* (2023) Chromatin organization drives the search mechanism of nuclear factors. *Nat Commun* 14: 6433

McGinnis CS, Murrow LM, Gartner ZJ (2019) DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems* 8: 329-337 e324

McGranahan N, Swanton C (2017) Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* 168: 613-628

McKinney W (2010) Data Structures for Statistical Computing in {P}ython. *Proceedings of the 9th Python in Science Conference*: 56 - 61

Meeussen JVW, Pomp W, Brouwer I, de Jonge WJ, Patel HP, Lenstra TL (2023) Transcription factor clusters enable target search but do not contribute to target gene activation. *Nucleic Acids Res* 51: 5449-5468

Metzner E, Southard KM, Norman TM (2024) Multiome Perturb-seq unlocks scalable discovery of integrated perturbation effects on the transcriptome and epigenome. *bioRxiv*

Miao Z, Kim J (2024) Uniform quantification of single-nucleus ATAC-seq data with Paired-Insertion Counting (PIC) and a model-based insertion rate estimator. *Nat Methods* 21: 32-36

Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carre C, Burdin N, Visan L, Ceccarelli M, Poidinger M *et al* (2019) RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep* 26: 1627-1640 e1627

Morey L, Helin K (2010) Polycomb group protein-mediated repression of transcription. *Trends Biochem Sci* 35: 323-332

Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25: 2607-2608

Morrison O, Thakur J (2021) Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin. *Int J Mol Sci* 22

Mostafavi S, Yoshida H, Moodley D, LeBoite H, Rothamel K, Raj T, Ye CJ, Chevrier N, Zhang SY, Feng T *et al* (2016) Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* 164: 564-578

Mota A, Schweitzer M, Wernersson E, Crosetto N, Bienko M (2022) Simultaneous visualization of DNA loci in single cells by combinatorial multi-color iFISH. *Sci Data* 9: 47

Muckenhuber M, Seufert I, Muller-Ott K, Mallm JP, Klett LC, Knotz C, Hechler J, Kepper N, Erdel F, Rippe K (2023) Epigenetic signals that direct cell type-specific interferon beta response in mouse cells. *Life Sci Alliance* 6: e202201823

Mukherjee A, Fallacaro S, Ratchasanmuang P, Zinski J, Boka A, Shankta K, Mir M (2024) A fine kinetic balance of interactions directs transcription factor hubs to genes. *bioRxiv*

Mulqueen RM, DeRosa BA, Thornton CA, Sayar Z, Torkenczy KA, Fields AJ, Wright KM, Nan X, Ramji R, Steemers FJ *et al* (2019) Improved single-cell ATAC-seq reveals chromatin dynamics of in vitro corticogenesis. *bioRxiv*

Nadeu F, Royo R, Clot G, Duran-Ferrer M, Navarro A, Martin S, Lu J, Zenz T, Baumann T, Jares P *et al* (2021) IGLV3-21R110 identifies an aggressive biological subtype of chronic lymphocytic leukemia with intermediate epigenetics. *Blood* 137: 2935-2946

Natarajan P, Shrinivas K, Chakraborty AK (2023) A model for cis-regulation of transcriptional condensates and gene expression by proximal lncRNAs. *Biophys J* 122: 2757-2772

Neill G, Masson GR (2023) A stay of execution: ATF4 regulation and potential outcomes for the integrated stress response. *Front Mol Neurosci* 16: 1112253

Nunes VS, Moretti NS (2017) Nuclear subcompartments: an overview. *Cell Biol Int* 41: 2-7

Oestreich KJ, Weinmann AS (2012) T-bet employs diverse regulatory mechanisms to repress transcription. *Trends Immunol* 33: 78-83

Pachitariu M, Stringer C (2022) Cellpose 2.0: how to train your own model. *Nat Methods* 19: 1634-1641

Papantonis A, Cook PR (2013) Transcription factories: genome organization and gene regulation. *Chem Rev* 113: 8683-8705

Papantonis A, Kohro T, Baboo S, Larkin JD, Deng B, Short P, Tsutsumi S, Taylor S, Kanki Y, Kobayashi M *et al* (2012) TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed. *EMBO J* 31: 4404-4414

Patel H, Ewels P, Manning J, Garcia MU, Peltzer A, Hammarén R, Botvinnik O, Talbot A, Sturm G, bot n-c *et al* (2018) nf-core/rnaseq. *Zenodo*: URL https://nf-co.re/rnaseq

Perlman BS, Burget N, Zhou Y, Schwartz GW, Petrovic J, Modrusan Z, Faryabi RB (2024) Enhancer-promoter hubs organize transcriptional networks promoting oncogenesis and drug resistance. *Nat Commun* 15: 8070

Persad S, Choo ZN, Dien C, Sohail N, Masilionis I, Chaligne R, Nawy T, Brown CC, Sharma R, Pe'er I *et al* (2023) SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat Biotechnol* 41: 1746-1757

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. *Genomics* 93: 105-111

Platanitis E, Demiroz D, Schneller A, Fischer K, Capelle C, Hartl M, Gossenreiter T, Muller M, Novatchkova M, Decker T (2019) A molecular switch from STAT2-IRF9 to ISGF3 underlies interferon-induced gene transcription. *Nat Commun* 10: 2921

Platanitis E, Gruener S, Ravi Sundar Jose Geetha A, Boccuni L, Vogt A, Novatchkova M, Sommer A, Barozzi I, Muller M, Decker T (2022) Interferons reshape the 3D conformation and accessibility of macrophage chromatin. *iScience* 25: 103840

Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A *et al* (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* 71: 858-871 e858

Pomp W, Meeussen JVW, Lenstra TL (2024) Transcription factor exchange enables prolonged transcriptional bursts. *Mol Cell* 84: 1036-1048 e1039

Poncz M, Surrey S, LaRocco P, Weiss MJ, Rappaport EF, Conway TM, Schwartz E (1987) Cloning and characterization of platelet factor 4 cDNA derived from a human erythroleukemic cell line. *Blood* 69: 219-223

Poos AM, Prokoph N, Przybilla MJ, Mallm JP, Steiger S, Seufert I, John L, Tirier SM, Bauer K, Baumann A *et al* (2023) Resolving therapy resistance mechanisms in multiple myeloma by multiomics subclone analysis. *Blood* 142: 1633-1646

Popp AP, Hettich J, Gebhardt JCM (2021) Altering transcription factor binding reveals comprehensive transcriptional kinetics of a basic gene. *Nucleic Acids Res* 49: 6249-6266

Porrua O, Libri D (2015) Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* 16: 190-202

Preissl S, Gaulton KJ, Ren B (2023) Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet* 24: 21-43

Proudfoot NJ (2016) Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* 352: aad9926

Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, Munar M, Rubio-Perez C, Jares P, Aymerich M *et al* (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526: 519-524

Rademacher A, Huseynov A, Bortolomeazzi M, Wille SJ, Schumacher S, Sant P, Keitel D, Okonechnikov K, Ghasemi D, Pajtler K *et al* (2024) Comparison of spatial transcriptomics technologies used for tumor cryosections. *bioRxiv*: 2024.2004.2003.586404

Rahman MM, McFadden G (2006) Modulation of tumor necrosis factor by microbial pathogens. *PLoS Pathog* 2: e4

Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4: e309

Ramasamy S, Aljahani A, Karpinska MA, Cao TBN, Velychko T, Cruz JN, Lidschreiber M, Oudelaar AM (2023) The Mediator complex regulates enhancer-promoter interactions. *Nat Struct Mol Biol* 30: 991-1000

Ramji DP, Foka P (2002) CCAAT/enhancer-binding proteins: structure, function and regulation. *Biochem J* 365: 561-575

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES *et al* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159: 1665-1680

Ravi Sundar Jose Geetha A, Fischer K, Babadei O, Smesnik G, Vogt A, Platanitis E, Müller M, Farlik M, Decker T (2024) Dynamic control of gene expression by ISGF3 and IRF1 during IFNβ and IFNγ signaling. *The EMBO Journal* 43: 2233-2263

Ren B, Chee KJ, Kim TH, Maniatis T (1999) PRDI-BF1/Blimp-1 repression is mediated by corepressors of the Groucho family of proteins. *Genes Dev* 13: 125-137

Ribeiro MD, Ziyani C, Delaneau O (2022) Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis. *Commun Biol* 5: 876

Rippe K (2022) Liquid-Liquid Phase Separation in Chromatin. *Cold Spring Harb Perspect Biol* 14: a040683

Rippe K, Papantonis A (2021) RNA polymerase II transcription compartments: from multivalent chromatin binding to liquid droplet formation? *Nat Rev Mol Cell Biol* 22: 645-646

Rippe K, Papantonis A (2022) Functional organization of RNA polymerase II in nuclear subcompartments. *Curr Opin Cell Biol* 74: 88-96

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47

Rodriguez J, Larson DR (2020) Transcription in Living Cells: Molecular Mechanisms of Bursting. *Annu Rev Biochem* 89: 189-212

Roessner PM, Seufert I, Chapaprieta V, Jayabalan R, Briesch H, Massoni-Badosa R, Boskovic P, Beckendorff J, Roider T, Arseni L *et al* (2024) T-bet suppresses proliferation of malignant B cells in chronic lymphocytic leukemia. *Blood* 144: 510-524

Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR *et al* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 481: 389-393

Rothorl J, Brems MA, Stevens TJ, Virnau P (2023) Reconstructing diploid 3D chromatin structures from single cell Hi-C data with a polymer-based approach. *Front Bioinform* 3: 1284484

Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, Mumbach MR, Ji AL, Kim DS, Cho SW *et al* (2019) Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* 176: 361-376 e317

Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, Chapman R, Hertzog PJ (2013) Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res* 41: D1040-1046

Ryu K, Park G, Cho WK (2024) Emerging insights into transcriptional condensates. *Exp Mol Med* 56: 820-826

Santiago-Algarra D, Souaid C, Singh H, Dao LTM, Hussain S, Medina-Rivera A, Ramirez-Navarro L, Castro-Mondragon JA, Sadouni N, Charbonnier G *et al* (2021) Epromoters function as a hub to recruit key transcription factors required for the inflammatory response. *Nat Commun* 12: 6660

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T (2002) Active genes are tri-methylated at K4 of histone H3. *Nature* 419: 407-411

Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR *et al* (2019) Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* 37: 925-936

Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, Gibrat JF (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *J Comput Biol* 19: 796-813

Schep AN, Wu B, Buenrostro JD, Greenleaf WJ (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 14: 975-978

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B *et al* (2012) Fiji: an open-source platform for biological-image analysis. *Nat Methods* 9: 676-682

Schoggins JW, Wilson SJ, Panis M, Murphy MY, Jones CT, Bieniasz P, Rice CM (2011) A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature* 472: 481-485

Schuster LC, Syed AP, Tirier SM, Steiger S, Seufert I, Becker H, Duque-Afonso J, Ma T, Ogawa S, Mallm JP *et al* (2023) Progenitor like cell type of an MLL-EDC4 fusion in acute myeloid leukemia. *Blood Adv* 7: 7079-7083

Sekimata M, Perez-Melgosa M, Miller SA, Weinmann AS, Sabo PJ, Sandstrom R, Dorschner MO, Stamatoyannopoulos JA, Wilson CB (2009) CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. *Immunity* 31: 551-564

Seufert I, Gerosa I, Varamogianni-Mamatsi V, Vladimirova A, Sen E, Mantz S, Rademacher A, Schumacher S, Liakopoulos P, Kolovos P *et al* (2024) Two distinct chromatin modules regulate proinflammatory gene expression. *bioRxiv*

Seufert I, Sant P, Bauer K, Syed AP, Rippe K, Mallm J-P (2023) Enhancing sensitivity and versatility of Tn5-based single cell omics. *Front Epigenet Epigenom* 1: 1245879

Shannon AE, Boos CE, Hummon AB (2021) Co-culturing multicellular tumor models: Modeling the tumor microenvironment and analysis techniques. *Proteomics* 21: e2000103

Shattil SJ, Hoxie JA, Cunningham M, Brass LF (1985) Changes in the platelet membrane glycoprotein IIb.IIIa complex during platelet activation. *J Biol Chem* 260: 11107-11114

Shi P, Nie Y, Yang J, Zhang W, Tang Z, Xu J (2022) Fundamental and practical approaches for single-cell ATAC-seq analysis. *aBIOTECH* 3: 212-223

Simon JM, Giresi PG, Davis IJ, Lieb JD (2012) Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 7: 256-267

Smale ST (2010) Selective transcription in response to an inflammatory stimulus. *Cell* 140: 833-844

Song L, Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010: pdb prot5384

Sood V, Misteli T (2022) The stochastic nature of genome organization and function. *Curr Opin Genet Dev* 72: 45-52

Soto LF, Li Z, Santoso CS, Berenson A, Ho I, Shen VX, Yuan S, Fuxman Bass JI (2022) Compendium of human transcription factor effector domains. *Mol Cell* 82: 514-526

Stadhouders R, Cico A, Stephen T, Thongjuea S, Kolovos P, Baymaz HI, Yu X, Demmers J, Bezstarosti K, Maas A *et al* (2015) Control of developmentally primed erythroid genes by combinatorial co-repressor actions. *Nat Commun* 6: 8893

Stark GR, Darnell JE, Jr. (2012) The JAK-STAT pathway at twenty. *Immunity* 36: 503-514

Stavreva DA, Garcia DA, Fettweis G, Gudla PR, Zaki GF, Soni V, McGowan A, Williams G, Huynh A, Palangat M *et al* (2019) Transcriptional Bursting and Co-bursting Regulation by Steroid Hormone Release Pattern and Transcription Factor Mobility. *Mol Cell* 75: 1161-1177 e1111

Stone SL, Peel JN, Scharer CD, Risley CA, Chisolm DA, Schultz MD, Yu B, Ballesteros-Tato A, Wojciechowski W, Mousseau B *et al* (2019) T-bet Transcription Factor Promotes Antibody-Secreting Cell Differentiation by Limiting the Inflammatory Effects of IFN-gamma on B Cells. *Immunity* 50: 1172-1187 e1177

Stormo GD (2013) Modeling the specificity of protein-DNA interactions. *Quant Biol* 1: 115-130

Stosch JM, Heumuller A, Niemoller C, Bleul S, Rothenberg-Thurley M, Riba J, Renz N, Szarc Vel Szic K, Pfeifer D, Follo M *et al* (2018) Gene mutations and clonal architecture in myelodysplastic syndromes and changes upon progression to acute myeloid leukaemia and under treatment. *Br J Haematol* 182: 830-842

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R (2019) Comprehensive Integration of Single-Cell Data. *Cell* 177: 1888-1902 e1821

Stuart T, Srivastava A, Madad S, Lareau CA, Satija R (2021) Single-cell chromatin state analysis with Signac. *Nat Methods* 18: 1333-1341

Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science* 332: 472-474

Suzuki HI, Onimaru K (2022) Biomolecular condensates in cancer biology. *Cancer Sci* 113: 382-391

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31: 2032-2034

Team RC, 1993. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. pp. URL https://www.R-project.org/.

Terstappen LW, Huang S, Picker LJ (1992) Flow cytometric assessment of human T-cell differentiation in thymus and bone marrow. *Blood* 79: 666-677

Thibodeau A, Eroglu A, McGinnis CS, Lawlor N, Nehar-Belaid D, Kursawe R, Marches R, Conrad DN, Kuchel GA, Gartner ZJ *et al* (2021) AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biology* 22: 252

Tomic J, Lichty B, Spaner DE (2011) Aberrant interferon-signaling is associated with aggressive chronic lymphocytic leukemia. *Blood* 117: 2668-2680

Trojanowski J, Frank L, Rademacher A, Mucke N, Grigaitis P, Rippe K (2022) Transcription activation is enhanced by multivalent interactions independent of phase separation. *Mol Cell* 82: 1878-1893 e1810

Trojanowski J, Rippe K (2022) Transcription factor binding and activity on chromatin. *Current Opinion in Systems Biology* 31: 100438

TW HB, Girke T (2016) systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics* 17: 388

Uhlen M, Karlsson MJ, Zhong W, Tebani A, Pou C, Mikes J, Lakshmikanth T, Forsstrom B, Edfors F, Odeberg J *et al* (2019) A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* 366

Uyehara CM, Apostolou E (2023) 3D enhancer-promoter interactions and multi-connected hubs: Organizational principles and functional roles. *Cell Rep* 42: 112068

van Mierlo G, Pushkarev O, Kribelbauer JF, Deplancke B (2023) Chromatin modules and their implication in genomic organization and gene regulation. *Trends Genet* 39: 140-153

Vermunt MW, Luan J, Zhang Z, Thrasher AJ, Huang A, Saari MS, Khandros E, Beagrie RA, Zhang S, Vemulamada P *et al* (2023) Gene silencing dynamics are modulated by transiently active regulatory elements. *Mol Cell* 83: 715-730 e716

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J *et al* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17: 261-272

Vishnoi K, Viswakarma N, Rana A, Rana B (2020) Transcription Factors in Cancer Development and Therapy. *Cancers (Basel)* 12

Wagner EJ, Carpenter PB (2012) Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol* 13: 115-126

Wala J, Imielinski M (2023) *GitHub*: URL https://github.com/mskilab-org/gUtils

Wang H, Han M, Qi LS (2021) Engineering 3D genome organization. *Nat Rev Genet* 22: 343-360

Wang M, Sunkel BD, Ray WC, Stanton BZ (2022) Chromatin structure in cancer. *BMC Mol Cell Biol* 23: 35

Wang R, Wang J, Paul AM, Acharya D, Bai F, Huang F, Guo YL (2013) Mouse embryonic stem cells are deficient in type I interferon expression in response to viral infections and double-stranded RNA. *J Biol Chem* 288: 15926-15936

Wang W, Lopez McDonald MC, Kim C, Ma M, Pan ZT, Kaufmann C, Frank DA (2023a) The complementary roles of STAT3 and STAT1 in cancer biology: insights into tumor pathogenesis and therapeutic strategies. *Front Immunol* 14: 1265818

Wang X, Fan Y, Wu Q (2023b) The regulation of transcription elongation in embryonic stem cells. *Front Cell Dev Biol* 11: 1145611

Wang X, Yue F (2023) HiCLift: a fast and efficient tool for converting chromatin interaction data between genome assemblies. *Bioinformatics* 39

Wang Y, Ni T, Wang W, Liu F (2019) Gene transcription in bursting: a unified mode for realizing accuracy and stochasticity. *Biol Rev Camb Philos Soc* 94: 248-258

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63

Wang Z, Zhang Z, Luo S, Zhou T, Zhang J (2024) Power-law behavior of transcriptional bursting regulated by enhancer-promoter communication. *Genome Res* 34: 106-118

Wei MT, Chang YC, Shimobayashi SF, Shin Y, Strom AR, Brangwynne CP (2020) Nucleated transcriptional condensates amplify gene expression. *Nat Cell Biol* 22: 1187-1196

Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL *et al* (2017) YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 171: 1573-1588 e1528

Weiterer SS, Meier-Soelch J, Georgomanolis T, Mizi A, Beyerlein A, Weiser H, Brant L, Mayr-Buro C, Jurida L, Beuerlein K *et al* (2020) Distinct IL-1alpha-responsive enhancers promote acute and coordinated changes in chromatin topology in a hierarchical manner. *EMBO J* 39: e101533

Wibisana JN, Inaba T, Shinohara H, Yumoto N, Hayashi T, Umeda M, Ebisawa M, Nikaido I, Sako Y, Okada M (2022) Enhanced transcriptional heterogeneity mediated by NF-kappaB super-enhancers. *PLoS Genet* 18: e1010235

Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York

Wiens M, Farahani H, Scott RW, Underhill TM, Bashashati A (2024) Benchmarking bulk and single-cell variant-calling approaches on Chromium scRNA-seq and scATAC-seq libraries. *Genome Res*

Wolock SL, Lopez R, Klein AM (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* 8: 281-291 e289

Woringer M, Darzacq X (2018) Protein motion in the nucleus: from anomalous diffusion to weak interactions. *Biochem Soc Trans* 46: 945-956

Wu Y, Seufert I, Al-Shaheri FN, Kurilov R, Bauer AS, Manoochehri M, Moskalev EA, Brors B, Tjaden C, Giese NA *et al* (2023) DNA-methylation signature accurately differentiates pancreatic cancer from chronic pancreatitis in tissue and plasma. *Gut* 72: 2344-2353

Xin B, Rohs R (2018) Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res* 28: 321-333

Young AP, Jackson DJ, Wyeth RC (2020) A technical review and guide to RNA fluorescence in situ hybridization. *PeerJ* 8: e8806

Zamudio AV, Dall'Agnese A, Henninger JE, Manteiga JC, Afeyan LK, Hannett NM, Coffey EL, Li CH, Oksuz O, Sabari BR *et al* (2019) Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes. *Mol Cell* 76: 753-766 e756

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 9: R137

Zhao F, Ma X, Yao B, Lu Q, Chen L (2024) scaDA: A novel statistical method for differential analysis of single-cell chromatin accessibility sequencing data. *PLoS Comput Biol* 20: e1011854

Zhao GN, Jiang DS, Li H (2015) Interferon regulatory factors: at the crossroads of immunity, metabolism, and disease. *Biochim Biophys Acta* 1852: 365-378

Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J *et al* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8: 14049

Zhu A, Ibrahim JG, Love MI (2019) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* 35: 2084-2092