A Multi-Omics Analysis of Transcription Control by BRD4

DISSERTATION

zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin.

vorgelegt von

Annkatrin Sarah Bressin

Berlin, 2022

Erstgutachter:	Prof. Dr. Martin Vingron
Zweitgutachter:	Prof. Dr. Rosario M. Piro
Tag der Disputation:	19.12.2022

ABSTRACT

RNA polymerase II (Pol II) regulation during early elongation has emerged as a regulatory hub in the gene expression of multicellular organisms. Prior research links the BRD4 protein to this control point, regulating the release of paused Pol II into productive elongation. However, the exact roles and mechanisms by which BRD4 influences this and potentially other post-initiation regulatory processes remain unknown. This study combines rapid BRD4 protein degradation and multi-omics approaches, including *nascent elongating transcript sequencing* (NET-seq), to uncover BRD4's direct protein functions.

Applying NET-seq in comparative studies required experimental adaptations. First, analyses with spiked-in mouse cells proved essential for reliable normalization. Second, the study identified a disproportional enrichment of a chromatin-associated RNA class as NET-seq's major limitation. Incorporating an additional enrichment step solved this problem and significantly increased Pol II coverage.

The resulting high-sensitivity NET-seq method confirmed BRD4's proposed role in early elongation by revealing a global defect in Pol II pause release upon BRD4 degradation. Observations from proteomics and *chromatin immunoprecipitation followed by sequencing* (ChIP-seq) experiments suggest that the failed recruitment of *Pol II-associated factors* (PAF) causes an assembly defect of a competent elongation complex.

Interestingly, the elongation defect also affected transcribed enhancers. Pol II occupancy increased in a region proximal to the enhancer center, strikingly similar to the impaired Pol II pause release at genes. An integrated multi-omics analysis that included genome-wide 3D genome information revealed reduced interactions between these enhancers and other regulatory regions.

Another unexpected result was the widespread Pol II readthrough transcription quantified by the developed readthrough index, revealing an apparent transcriptional termination defect. The implementation of *long-read nascent RNA-sequencing* (nascONT-seq) combined with a 3′-RNA cleavage efficiency test detected impaired 3′-RNA processing. Notably, those 3′-RNA cleavage defects correlated with the observed termination defects. A potential explanation is the BRD4-dependent recruitment of general 3′-RNA processing factors to the 5′-control region. These observations start to establish regulatory links between 5′ and 3′ control that require further validation. Overall, the results indicate a general BRD4-dependent 5′ elongation control point required for 3′-RNA processing and termination.

PUBLICATION

The results of this work were published in *Molecular Cell* (Arnold* and Bressin* et al. 2021 [7], `https://doi.org/10.1016/j.molcel.2021.06.026`) and *EMBO Molecular Medicine* (Nabavizadeh* and Bressin* et al. 2023 [162], `https://doi.org/10.15252/emmm.202216478`).

ACKNOWLEDGMENTS

First of all, I thank my advisor Andreas Mayer for the opportunity to work as a PhD candidate in his group at the *Max Planck Institute for Molecular Genetics*. I am grateful for all his support and guidance that allowed me to work on this challenging project and led to this thesis.

Furthermore, I want to thank my supervisor Martin Vingron for reviewing this thesis and advising me. I want to thank Tugce Aktas and Rosario Piro for joining and participating in my *thesis advisory committee* meetings, especially Rosario Piro, who agreed to review this work as my external supervisor.

This work would be impossible without the contribution of the experimental biologists from the Mayer laboratory, including Susanne Freier and Nicole Eischer. Further, I want to highlight the outstanding work of Mirjam Arnold and Olga Jasnovidova. Thank you for the countless discussions and the close collaboration, which made this project special to me. I was lucky to have Martyna Gajos as my PhD companion, who supported me with computational analyses during the last years and countless brainstormings.

Working with Nasrinsadat Nabavizadeh and Bruno Reversade on the *osteogenesis imperfecta* project was an educational experience. I want to thank them for this productive collaboration.

Many people in my life helped me personally through my time as a PhD candidate. I want to thank my family members and friends for their emotional support. I can't express how grateful I am for the remarkable support that my partner Benedikt Schielke provided over the last few years. Thank you for all your help, especially over the past few months. Thanks to the institute's employees, which provided the structures that allowed my work, including the *International Max Planck Research School*, the IT and sequencing facility, and PhD coordinator Kirsten Kelleher. Thanks to all colleagues and friends for proofreading parts of my thesis.

* equal contribution

In the end, I want to thank all members of my wonderful group for working with me and helping me understand biology through the eyes of a biologist. It was always a pleasure to work with you.

CONTENTS

1 Introduction 1

2 Biological Background 5

2.1 The Chromatin Template 5

2.2 Regulation of RNA Polymerase II Transcription 7

2.2.1 Transcription Cycle 7

2.2.2 Promoter-proximal Pausing and Release 8

2.2.3 3′-RNA Processing and Termination 9

2.3 Bromodomain and Extraterminal Domain Protein Family 10

2.4 Experimental Techniques in Functional Genomics 12

2.4.1 HTS Technologies 12

2.4.2 Long-read Sequencing 17

2.4.3 Metabolic Labeling of RNA 18

2.4.4 Protein Knockdown Strategies 19

3 Computational Background 21

3.1 Mapping HTS Data 21

3.2 Normalizing HTS Data 22

3.3 Differential Analysis for HTS Count Data 28

3.3.1 Negative Binomial Distribution 28

3.3.2 DEseq2 28

3.4 Gene Ontology Term Enrichment Analysis 31

4 Materials and Contributions 33

I Studying Pol II with High Resolution and Improved Coverage

5 Motivation 37

6 Methods 39

6.1 Identifying Active Transcription Units 39

6.1.1 Active Genes 39

6.1.2 Active Enhancers 40

6.2 Processing HiS-/NET-seq Data 40

6.3 Pol II Indices 42

6.3.1 Promoter-proximal Maximum Pausing Position 43

6.3.2 Pausing Index 44

6.3.3 Termination Zone Length 44
6.3.4 Average Termination Distance 44
6.4 Meta-gene Profiles and Heatmaps 45
7 Results 47
7.1 Extracting Reliable Pol II Occupancy Data 47
7.2 Optimizing Pol II Enrichment Efficiency 47
7.3 Increasing Coverage with the HiS-NET-seq Method 49
7.4 Identifying Method Independent Pol II Occupancy Features . . 52
8 Discussion 57
II Detecting Global Pol II Occupancy Changes with Improved Quantitative Methods
9 Motivation 63
10 Case Studies 65
11 Results 67
11.1 Identifying Genome-wide Pol II Deregulation 67
11.2 Using Reference Cells for New NET-seq Variants 70
11.2.1 Studying a Joint Reference Genome from Human and Mouse 71
11.2.2 SI-NET-seq 73
11.3 Detecting Global Changes with Reference-based Normalization 75
12 Discussion 77
III BRD4 Emerges as Global Regulator of Pol II Transcription
13 Motivation 81
14 Methods 83
14.1 Analyzing HTS Data 83
14.1.1 RNA-seq 83
14.1.2 ChIP-seq and ChIP-Rx 83
14.1.3 nascONT-seq 85
14.1.4 HiChIP 89
14.2 Transcriptional Readthrough Index 92
14.3 Splicing Analysis 93
14.4 3′-RNA Cleavage Efficiency Test 93
15 Results 97
15.1 The BRD4-mediated 5′ Elongation Checkpoint 97
15.1.1 Global Decrease of Nascent RNA Transcription 97

15.1.2 BRD4-degradation Impairs Pol II Pause Release 99
15.1.3 Distinct Role of Other BET Proteins 99
15.2 BET Proteins Regulate 3′-RNA Processing 101
15.2.1 Widespread Pol II Readthrough Transcription 101
15.2.2 Functional Consequences of Termination Defects 103
15.2.3 3′-RNA Cleavage Defects 104
15.2.4 BRD4 Binds 5′ Regions of Readthrough Genes 106
15.3 BRD4-Dependent Recruitment of the 3′-RNA Processing Machinery 108
15.3.1 Recruitment Defects of 3′-Processing Factors 108
15.3.2 Core-Interactome of BRD4 Reveals 5′ and 3′ Regulators . 110
15.3.3 Contribution of Other Elongation Factors on 3′-RNA Processing Defects 110
15.4 Pol II Regulation at Enhancer Regions 113
15.4.1 BRD4 Binds Transcribed Enhancer Regions 113
15.4.2 Pol II Elongation Control at Enhancer Regions 115
15.4.3 Loss of BRD4 Disrupts Regulatory Interactions 117
16 Discussion 121
17 Conclusion 127
18 Other Project Contributions 131
A Supplementary Figures 133
B Supplementary Tables 159

Bibliography 173
List of Figures 197
List of Tables 198
Acronyms 199
Zusammenfassung 203
Selbstständigkeitserklärung 205

1 INTRODUCTION

The proteins that cover most cell functions and shape the identity of all living organisms, including humans, are encoded by genes in the *deoxyribonucleic acid* (DNA) molecules. Extracting this information is a highly regulated multi-step process. This process, called gene expression, produces *ribonucleic acid* (RNA) molecules and synthesizes them into proteins at different rates and compositions depending on the concrete biological condition. *RNA polymerase II* (Pol II) is the protein complex that transcribes the encoded gene information in the DNA into RNA, including all protein-coding and most non-coding RNA genes [43]. Research over the last decades focused on the first step of transcription, transcription initiation, where different factors recruit Pol II to the DNA. After initiation, transcription elongation and termination occur. The fundamental assumption was that gene activity was mainly regulated during the recruitment and initiation of Pol II [183].

With the development of sequencing-based methods revealing Pol II occupancy genome-wide [160, 187, 254], transcription elongation emerged as a rate-limiting step in gene expression control of multicellular organisms [102, 141]. This significant control point occurs after initiation when Pol II pauses in the promoter-proximal region. This process creates an additional step to integrate multiple regulatory signals [1], leading to Pol II pause release and productive elongation. The *bromodomain and extraterminal domain* (BET) protein family is known to regulate this step, including the most studied BRD4 protein. Nevertheless, our knowledge of the detailed steps is still incomplete, including the mechanistic roles of individual BET proteins family members.

Although often overlooked, a successful transcription cycle includes successful termination events at 3′ ends of genes where nascent RNA is cleaved and polyadenylated, triggering Pol II released from the DNA template. This step is essential for creating a functional RNA transcript and avoiding transcriptional interference between gene units [178, 214]. Because Pol II termination is the least studied Pol II transcriptional step, the exact processes in human cells and relevant factors remain poorly characterized.

This thesis aims to enhance the general understanding of the regulatory processes that occur after Pol II initiation and potentially change the productive output of a transcription unit. Furthermore, a central goal includes improving computational and experimental methods that allow this and future studies to investigate Pol II transcription quantitatively and with high sensitivity. Since many factors are involved in Pol II regulation, this work will focus on the regulatory functions of the BET proteins with a focus on BRD4.

This study's main emphasis concerns the direct protein functions using selected *high-throughput sequencing* (HTS) assays that capture immediate consequences of targeted protein degradation after two hours or less.

Identifying these fundamental concepts of gene transcription is critical for understanding the complete regulatory toolbox that allows the complex regulation of gene expression in humans. BET proteins are promising targets for new small molecule inhibitors that are currently in clinical trials to treat different types of cancer [5, 31, 217]. Their essential biological functions *in vivo* will be of general interest to a broad audience.

Structure

Following this general introduction, Chapter 2 **Biological Background** and Chapter 3 **Computational Background** will provide more detailed insights into the notions of molecular biology and the computational models to study them. Chapter 2 **Biological Background** describes fundamental concepts of gene transcription and highlights the proposed functions of BRD4 and other BET proteins. Furthermore, this chapter introduces the published experimental techniques used in this study, including *nascent elongating transcript sequencing* (NET-seq). Chapter 3 **Computational Background** reports established methods in bioinformatics used to analyze HTS data, focusing on normalization and testing methods to detect quantitative changes in count data. Chapter 4 **Materials and Contributions** provides an overview of the materials used in this integrative multi-omics study, including published and unpublished HTS data, software tools, databases, cell lines, and others. Furthermore, this section states contributions from collaborating scientists.

The three main parts summarize and discuss the results of this work. They include individual method sections that describe specific processing steps, analyses, indices, and tests developed for this study. The parts build upon each other, whereas **Part I** and **Part II** provide the methodological basis for **Part III**.

Part I explores the composition of NET-seq data using a refined data processing pipeline and identifies the main limitations. Different optimization steps have led to the new *high-sensitivity NET-seq* (HiS-NET-seq) method. Additionally, this part describes the related benchmark analysis that compared Pol II features between different high-resolution Pol II profiling methods.

Part II adapts established computational methods to identify relevant changes between NET-seq samples and highlights the challenges of detecting them in some transcription studies. The results of the analyses led to the development of a new NET-seq protocol variant, called *spike-in NET-seq* (SI-NET-seq). Furthermore, the section shows the practical application of SI-NET-seq in two case studies.

Part III reports the results of a multi-omics approach to identify post-initiation regulatory functions of BRD4.

This study reveals BRD4's regulatory function at different stages of transcription, including early elongation, 3′-RNA cleavage, Pol II termination, and elongation regulation at enhancer regions. Presented are different computational approaches to detect 3′-RNA cleavage and termination defects.

Chapter 17 **Conclusion** summarizes the main implications of all three parts of this thesis. The final Chapter 18 **Other Project Contributions** reports the key results of an additional project not covered in the central part of this thesis. In the project, the transcriptomic profiling of patient material revealed the genetic cause for a rare type of *osteogenesis imperfecta*.

2 BIOLOGICAL BACKGROUND

This chapter introduces basic molecular biology concepts and techniques used to study them in multicellular organisms. The first section describes the organization of DNA and the encoded elements. In the context of Pol II, the following sections explain the regulation of transcription, with a focus on post-initiation regulatory mechanisms and the contribution of the BET protein family members, concentratig on BRD4. Finally, this section presents essential experimental techniques required for a general understanding of this work.

2.1 THE CHROMATIN TEMPLATE

The basis that holds the information for all living organisms on earth is DNA. Two DNA strands, made of nucleotides, coil around each other and form the double-helix structure [245]. Each nucleotide consists of

- deoxyribose,
- a phosphate group, and
- one of the nucleobases: *adenine* (A), *cytosine* (C), *guanine* (G), and *thymine* (T).

The double-helix structure is assembled by hydrogen bonds formed between the complementary bases G/C and A/T.

Fitting the long DNA molecule into the nucleus of eukaryotes requires a compact and dense chromatin structure formed by DNA and proteins. The repeating unit of the chromatin is the nucleosome [150], which consists of 146 *base pairs* (bp) of DNA wrapped in 1.65 turns around the histone octamer. Heterodimers of H3, H4, H2A, and H2B proteins form the histone protein complex that can be chemically modified, contributing to DNA accessibility [112].

DNA is the static storage of genetic information kept in the cell nucleus and does not produce proteins directly. Transcription is the first intermediate step that produces RNA molecules from the DNA. Besides being primarily single-stranded, RNA is similar to DNA but contains

- ribose with an additional hydroxyl group instead of deoxyribose and
- the demethylated form of thymine named *uracil* (U).

After transcription, the RNA is processed and transported from the cell nucleus to the cytoplasm, where the second intermediate step, the translation from RNA into proteins, occurs.

Promoter

The promoter is a DNA region recognized by factors contributing to transcription initiation. A core promoter element that is required to initiate transcription consists of a *transcription start site* (TSS), a Pol II binding site, and a general transcription factors binding site, such as the AT-rich *TATA box* [127] or the initiator element [219]. Other features, such as histone modifications of the flanking nucleosomes and accessible chromatin, correlate with the promoter's activity. The *histone three lysine twenty-seven acetylation* (H3K27ac) relaxes the chromatin [225] and is generally associated with actively transcribed regions, including promoters. For the other common histone marks, such as *histone three lysine four* (H3K4) *mono-* (H3K4me1) *or trimethylation* (H3K4me3), it is unclear if they are the cause or consequence of transcription [91].

Enhancer

The enhancer is a DNA region recognized by transcription factors, contributing to transcription initiation over large genomic distances, independent of sequence orientation. Transcription factor binding at different enhancer regions regulates the gene expression at one or more promoters and can be highly dynamic between cell types and states [35, 216].

Although early studies [10] discovered the first enhancer more than forty years ago, general enhancer features and how they function remain ambiguous. The lack of a general and easy to apply HTS functional assays forced the field to indirect characterizations using transcription factor binding sites, cofactor binding, or histone modifications.

Enhancers are associated with open chromatin regions flanked by histones with H3K27ac and H3K4 methylation, whereas H3K4me1 is more common at enhancers and H3K4me3 at promoters [44, 84].

More recently, bidirectional enhancer transcription emerged as an additional general enhancer feature [36, 110], which correlates with enhancer activity [85]. It is, however, unclear if the unprocessed and rapidly degraded enhancer RNA contributes to the enhancer function directly [152], indirectly by maintaining open chromatin at the enhancer [74], or if it is an accidental by-product [222]. Overall, how enhancers achieve their enhancing function is still under debate, including the role of spacial enhancer-promoter interactions [207] and enhancer transcription [58]. In the last decade, their clear distinction to promoters blurred as enhancers are likewise transcribed, share most functional chromatin features, and some can act as promoters and vice versa [48, 154].

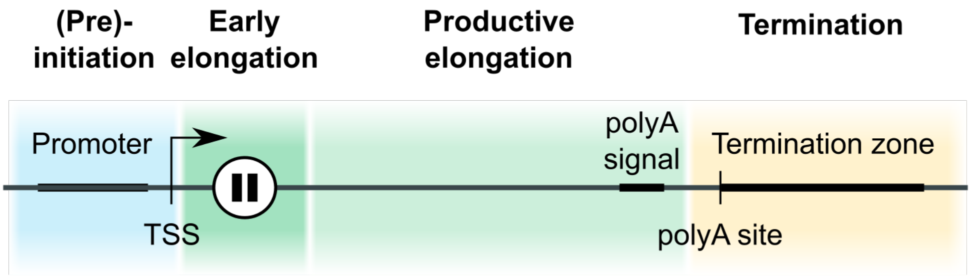


Figure 2.1: Pol II Transcription Cycle. The figure shows the stages of the transcription cycle, including initiation, elongation, termination, and re-initiation (not shown). Pol II is recruited to the promoter and starts transcription at the TSS. Pol II pauses 20-60 nucleotides downstream of the TSS during early elongation in the promoter-proximal region. The productive elongation phase produces nascent RNA. The polyA signal containing nascent RNA is co-transcriptionally cleavaged at the polyA site, triggering Pol II release from the DNA template in the termination zone. The re-initiation phase recycles Pol II for a new transcription cycle.

2.2 regulation of rna polymerase ii transcription

Nuclear transcription in mammalian cells depends on the enzyme complexes RNA polymerases I, II, and III [43], each catalyzes the transcription of specific RNA species. RNA polymerase I synthesizes the highly abundant *ribosomal RNA* that accounts for >85% of cellular RNA in most organisms [116]. RNA polymerase III produces *transfer RNAs*, 5S *ribosomal RNA*, and other small non-coding RNAs, for example, the spliceosomal U6 *small nuclear RNA* (snRNA) [49]. Interestingly, a fourth RNA polymerase exists in the mitochondrion, where the single-subunit protein is exclusively associated with the synthesis of mitochondrial RNA [195]. However, this section focuses on the complex regulation of the twelve subunit protein complex Pol II that transcribes all protein-coding and most non-coding RNA genes, including *enhancer RNA*, snRNA, *small nucleolar RNA* (snoRNA), and *micro RNA*.

2.2.1 *Transcription Cycle*

The Pol II transcription cycle divides into different stages, including initiation, elongation, termination, and re-initiation [213] (Figure 2.1).

Transcription starts with the initiation step. Pol II does not bind the DNA sequences of a promoter directly. Therefore, it requires general transcription factors [121] that recognize the promoter sequence and assemble stepwise into the pre-initiation complex [42]. This complex recruits Pol II to the promoter region and opens the DNA duplex. The first eight or nine nucleotides of the nascent RNA and the DNA template form a DNA-RNA hybrid known as the transcription bubble at the core of elongating Pol II [144].

Initiation is abortive and repeats until Pol II forms this DNA-RNA hybrid of critical length and escapes from the promoter [121].

The elongation stage synthesizes the nascent RNA transcript from the DNA template. In this procedure, the Pol II elongation complex unzips the double-stranded DNA and adds a nucleotide to the growing nascent RNA. In multicellular organisms, an additional step occurs during early elongation, where the interplay of different factors regulates the pausing and release of Pol II in the promoter-proximal region. For a detailed description of this tightly regulated process, see Section 2.2.2. If Pol II escapes from this checkpoint and potentially premature termination [19, 105], it is in an active form bound by elongation factors [232] that stabilize the processive elongation [165, 262].

Pol II termination occurs downstream of the gene's *polyadenylation* (polyA) site, in a termination window or zone [211]. General 3′-RNA processing factors bind to the nascent RNA, guided by conserved DNA signals, and cleave the transcript at the cleavage site, corresponding to the polyA site. Cleavage of the nascent RNA is an essential step that triggers processes removing Pol II from the DNA template and allowing re-initiation of a new transcription cycle. As the termination process is not fully understood, Section 2.2.3 discusses the steps of 3′-RNA cleavage and different termination models.

2.2.2 *Promoter-proximal Pausing and Release*

Promoter-proximal Pol II pausing was first described in *Drosophila melanogaster* (fly) at heat-shock genes [78]. The cell activates this group of genes in response to stressful conditions, suggesting a specific purpose of Pol II pausing, for example, to transcribe genes rapidly.

Instead, the process appeared as a general feature of early elongation in multicellular organisms [160, 187, 254]. Following transcription initiation, Pol II pauses in the promoter-proximal region between 20 and 60 nucleotides downstream of the TSS [40, 168, 188].

Structurally, Pol II pausing involves tilting the DNA-RNA hybrid [234], which impairs nucleotide addition and pause escape. The *DRB sensitivity-inducing factor* (DSIF), composed of SPT4 and SPT5, and the *negative elongation factor* (NELF) complex stabilize the pause [234, 236, 251]. If these factors equally contribute to establishing the Pol II pause is unclear.

The *positive transcription elongation factor b* (P-TEFb) complex, formed by the kinase CDK9 and *cyclin T1*, is required for Pol II pause release [173]. CDK9 triggers this process by phosphorylation of Pol II, SPT5, and NELF [142, 204]. Consequently, NELF dissociates from the complex, supporting the formation of an activated elongation complex [232, 233], where the elongation factor SPT6 and the *Pol II-associated factors* (PAF) complex bind. The PAF complex consists of PAF1, CDC73, CTR9, WDR61, LEO1, and RTF1.

Complex	Module	Factors	Function
CPSF	specificity factor	CPSF160, CPSF30, WDR33, FIP1	polyA signal recognition [209] and polyA polymerase recruitment [223]
	cleavage factor	CPSF73, CPSF100, symplekin	RNA cleavage [140, 202]
CstF	-	CstF50, CstF77, CstF64	U/GU-rich region recognition [22]
Cleavage factors	Im	CFIm25, CFIm59, CFIm68	polyA site selection [20]
	IIm	Pcf11, Clp1	Pol II binding, premature termination [105, 259]

Table 2.1: 3′-RNA Processing Complexes and Associated Factors.

An activated elongation complex results in an active conformation with a free nucleoside triphosphate binding site that can resume transcription. Pol II pause release regulation emerged as a significant control point in most genes [1, 40].

Close to the promoter-proximal pausing sites, a second distinct class of pausing events occurs further downstream of the TSS associated with the entry side of the first nucleosome, named *+1 nucleosome* [168]. Nucleosomal depletion at promoters results in the positioning of the *+1 nucleosome* center, named dyad, around 214 nucleotides downstream of the TSS [262]. The *+1 nucleosome* represents a potential obstacle during early elongation that requires the combined function of elongation factors.

Interestingly, nucleosome-induced pausing of Pol II at the *+1 nucleosome* occurs at the entry site and not directly upstream of the dyad as described for gene-body Pol II pausing [117]. The exact reason for the additional pausing site at the entry side and the difference between gene-body pausing is unclear.

2.2.3 *3′-RNA Processing and Termination*

In metazoa, co-transcriptional 3′-RNA processing links polyadenylation and Pol II termination at *messenger RNAs* and *long non-coding RNAs* [88, 193]. Conserved sequence elements in the nascent RNA and a large protein processing machinery, referred to as 3′-RNA processing factors, induce these processes.

Table 2.1 summarizes the protein complexes involved during 3′-RNA processing, including the two sub-modules of the *cleavage and polyA specificity factor* (CPSF) [257] and the *cleavage stimulation factor* (CstF). The first CPSF sub-module recognizes an upstream polyA signal [179, 180, 209], whereas the CstF complex binds to a U/GU-rich region downstream of the nascent RNA cleavage site [22]. Following binding, the second CPSF sub-module, which contains the endonuclease CPSF73, carries out the cleavage reaction [140, 202]. The *cleavage factor Im* complex regulates the polyA site selection [20]. In contrast, the *cleavage factor IIm* complex links the 3′-RNA processing machinery with Pol II [259] and is associated with the regulation of premature termination [105]. Finally, the CPSF complex recruits the polyA polymerase required for the polyA tail synthesis [223].

Two non-mutually exclusive Pol II termination models, proposed almost thirty-five years ago, rely on 3′-RNA processing [34, 131]. Although it is still unclear how Pol II terminates exactly, evidence suggests a combination of both models [53, 106, 135, 248]. According to the allosteric model, the transcription through the polyA site leads to changes in the elongation complex, where elongation factors dissociate, termination factors associate, or both [131].

The torpedo model depends on 3′-RNA cleavage, which creates an entry site for the *5′-3′ exoribonuclease 2* (XRN2) at the uncapped nascent RNA. XRN2 degrades the RNA until it catches up with the continuously transcribing Pol II, which leads to the release of Pol II from the DNA template by an unknown mechanism [34].

2.3 BROMODOMAIN AND EXTRATERMINAL DOMAIN PROTEIN FAMILY

In humans, the BET protein family consists of the ubiquitously expressed BRD2, BRD3, BRD4, and the testis-specific BRDT proteins [101]. Functionally they are involved in many cellular processes, including cell cycle progression, DNA replication, DNA repair, and transcriptional regulation [55]. This section focuses on the role of BET proteins, specially BRD4, during transcription regulation.

PROTEIN STRUCTURE All BET protein family members consist of two tandem bromodomains and a unique extraterminal domain at the N-terminus [60]. BRD4 expresses a short and long protein isoform, where the latter harbors an additional C-terminal domain that is also present in BRDT [55]. On the one hand, the two bromodomains bind acetylated lysine residues of proteins, such as H3K27ac histones [60, 255] or transcription factors [70]. This ability allows BET proteins to function as "readers" of the acetylated histone code. On the other side, the extraterminal domain binds other transcription factors and chromatin regulators [113, 189], serving as a scaffold protein. This architecture suggests that BET proteins link chromatin and transcriptional regulation.

TRANSCRIPTIONAL REGULATION In 2006, Peterlin and Price proposed an essential role of BRD4 during the regulation of early elongation [173]. *In vitro* studies [98, 252] suggested that BRD4 recruits the P-TEFb complex to pausing Pol II to influence elongation positively (Section 2.2.2). This model was later confirmed in several studies using small molecular inhibitors that mimic lysine acetylation [3, 133]. Upon inhibition, the BET protein bromodomains recognize and bind the small molecules, which blocks the bromodomain-dependent functions resulting in dissociation from the chromatin after six hours [59, 167]. However, in recent studies with a higher temporal resolution due to rapid protein degradation (Section 2.4.4), recruitment of P-TEFb was independent of BET proteins [249] and BRD4 [158, 260]. Therefore, it remains unclear how BET proteins, specifically BRD4, influence elongation positively. Different models suggest BRD4-dependent activation of P-TEFb [96, 130, 247] or independent mechanisms [11].

ENHANCER BET proteins occupy beside promoters also acetylated enhancer regions genome-wide [45, 64, 260] (Section 2.1). Furthermore, binding correlates with the production of *enhancer RNAs* [107, 164]. Inhibition of the BET proteins is associated with reduced *enhancer RNA* levels, which suggests that they positively influence their synthesis [107, 164]. A recent study suggests that the absence of the BET protein family member BRD2 is causing this observation [260]. The same study observed no effect on enhancer transcription upon specific BRD3 or BRD4 protein degradation experiments. However, Lee et al., 2017 [120] detected reduced enhancer transcription in BRD4 knockout cells at selected loci.

Furthermore, BRD4 forms transcriptional condensates, which harbor high levels of Pol II, transcription factors, and co-activators [17, 203]. Enhancers that form these condensates are called super-enhancers and regulate a well-defined set of genes [89]. The disruption of the condensates reduces transcription at the respective target genes [45, 203]. Many important regulators contain low complexity domains, named intrinsically disordered regions, which can undergo liquid-liquid phase separation with other proteins, DNA, and RNA. It is unclear if these droplet-like condensates contribute, besides transcriptional regulation, to the formation or maintenance of promoter-enhancer contacts. The evidence argues against a structural function [45]. However, several studies assign also architectural function to BET proteins [89], including BRD4 [128, 242].

DISEASE-ASSOCIATION Diseases critical genes, such as proto-oncogenes [89] and immunoregulatory genes [241], are sensitive toward BET protein inhibition. Therefore, BET proteins emerged as potential therapeutic targets for diseases [31], including different cancer types [5, 217] and immune-related disorders [241].

2.4 EXPERIMENTAL TECHNIQUES IN FUNCTIONAL GENOMICS

This section reports current techniques used in functional genomics. A short introduction covers the most popular HTS assays that measure transcriptome subsets, protein interactions with DNA, and three-dimensional DNA-DNA interactions. Complementary approaches to HTS are new long-read sequencing methods providing DNA sequence information of whole molecules instead of small fragments. After a short introduction to long-read sequencing technologies, other experimental techniques, including metabolic labeling and protein knockout strategies, are described.

2.4.1 *HTS Technologies*

In the twenty-first century's first decade, new technologies revolutionized the sequencing of short DNA fragments, allowing massive parallelization. The new technology quickly transformed also RNA research, which requires the translation of RNA into *complementary DNA* (cDNA) before being sequenced.

Machines from the *Illumina, Inc.* company dominate the market and produce hundreds of millions of short sequencing reads between 50-300 bps in each run [192]. The process involves cluster generation, sequencing-by-synthesis, and data analysis. The essential steps are the following.

1. The DNA fragment binds to the glass flow cell and is amplified.

2. Sequencing occurs via serial rounds of fluorescently-labeled base incorporation, washing, and imaging. The end of each round removes the 3′ block that paused the reaction, and the process repeats.

3. The sequencing read interprets the wavelengths and signal intensities in the images into a sequence of base pairs.

This section contains an overview of relevant methods that rely on HTS technologies.

RNA-seq

RNA sequencing (RNA-seq) measures the gene expression present in a population of cells using HTS. The essential steps of an RNA-seq experiment are the

1. purification of RNA,

2. fragmentation,

3. reverse transcription into cDNA,

4. *polymerase chain reaction* (PCR) amplification,
5. and HTS.

Different RNA-seq experiments vary considerably depending on the research question and the performed RNA purification and depletion steps. This study analyzed polyA-enriched RNA-seq, total RNA-seq, and nuclei RNA-seq data.

POLYA-ENRICHED RNA-SEQ Sequencing polyadenylated RNA is the most common application of RNA-seq [92]. The purification step selects processed RNA by using thymine oligonucleotide stretches. Those complementary oligonucleotides bind to polyadenine stretches and separate them from the remaining RNA.

TOTAL RNA-SEQ AND NUCLEI RNA-SEQ Total RNA sequencing measures the whole-cell lysates that mainly capture fully processed, stable, and more abundant cytoplasmic RNA. Sequencing RNA fractions from the nuclei enriches newly produced, nascent, and unstable RNA [147]. Therefore, nuclei-RNA-seq performs a cell fractionation to isolate the RNA from the nuclei. Both methods require the depletion of the highly abundant *ribosomal RNAs* from the libraries.

ChIP-seq and ChIP-Rx

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a method to study protein interactions with DNA genome-wide. Different studies started using the method in 2007 to profile transcription factors [100, 197] and histone modifications [12, 155] *in vivo*. The method combines immunoprecipitation and HTS to extract and identify the DNA regions bound by a protein of interest.

The crosslinking step fixes protein-DNA contacts using formaldehyde. Next, sonification leads to the fragmentation of the DNA. The immunoprecipitation step enriches exclusively for DNA fragments that bind the protein of interest using a protein-specific antibody. Before sequencing, reversing the crosslinking isolates the selected DNA fragments and enables PCR amplification. The computational analysis maps the resulting sequencing reads back to the reference genome and identifies the individual binding sites.

Unspecific antibody binding and chromatin accessibility bias the experiments considerably. For this reason, most ChIP-seq applications require a control experiment, which lacks the enrichment with the antibody or uses an unspecific antibody [118], referred to as matched input control.

CHIP-RX *Chromatin immunoprecipitation with reference exogenous genome spike-in followed by sequencing* (ChIP-Rx) is a ChIP-seq variant that adds defined quantities of an exogenous reference genome to the initial sample [9, 170].

The sequencing reads mapping to the spiked-in reference genome are identical among samples and used for between-sample normalization (Section 3.2.6). Previous studies used either material from fly [170] or NIH3T3 mouse cells [9] for combined applications with human cells.

Pol II Profiling

The first requirement to study Pol II across the human genome is an effective Pol II tracking method. A standard method to identify those protein-DNA interactions genome-wide is ChIP-seq. Although ChIP-seq is the most popular method to track protein-DNA interactions [35], using ChIP-seq for Pol II transcription studies fails to discriminate between different stages of the transcription cycle due to the low signal-to-noise ratio and resolution [137].

More recently, sequencing-based methods track Pol II at single-nucleotide resolution across the human genome [136, 148, 169]. All of these methods

- isolate RNA transcripts,
- ligate 3′-RNA adapters,
- produce cDNA and
- perform HTS with high sequencing depth.

As a result, 3′ ends of sequenced RNA are extracted and reveal Pol II active sites at nucleotide resolution genome-wide. The main difference is the enrichment strategy used to purify nascent RNA transcripts.

NET-SEQ Human NET-seq, developed in 2015 [146, 148], isolates chromatin-associated RNA enriched in nascent RNA fragments (Figure 2.2), as described below.

1. Detergents, salt, and urea perform a cell fractionation that isolates the chromatin and the stable RNA-DNA-Pol II elongation complex. The cell fractionation step uses α-amanitin, an inhibitor of Pol II elongation [129], to avoid transcriptional run-on of Pol II.

2. For the purification of RNA, deoxyribonuclease degrades the remaining DNA.

3. The library preparation step ligates a DNA linker with a six nucleotides long *unique molecular identifier* (UMI) to the 3′-hydroxyl group of nascent RNA molecules.

4. The reverse transcription step generates cDNA after RNA fragmentation and size selection (35 - 100 nucleotides). As the DNA linker adds to the overall length, the selected RNA fragments have a size of twelve to seventy-seven nucleotides.

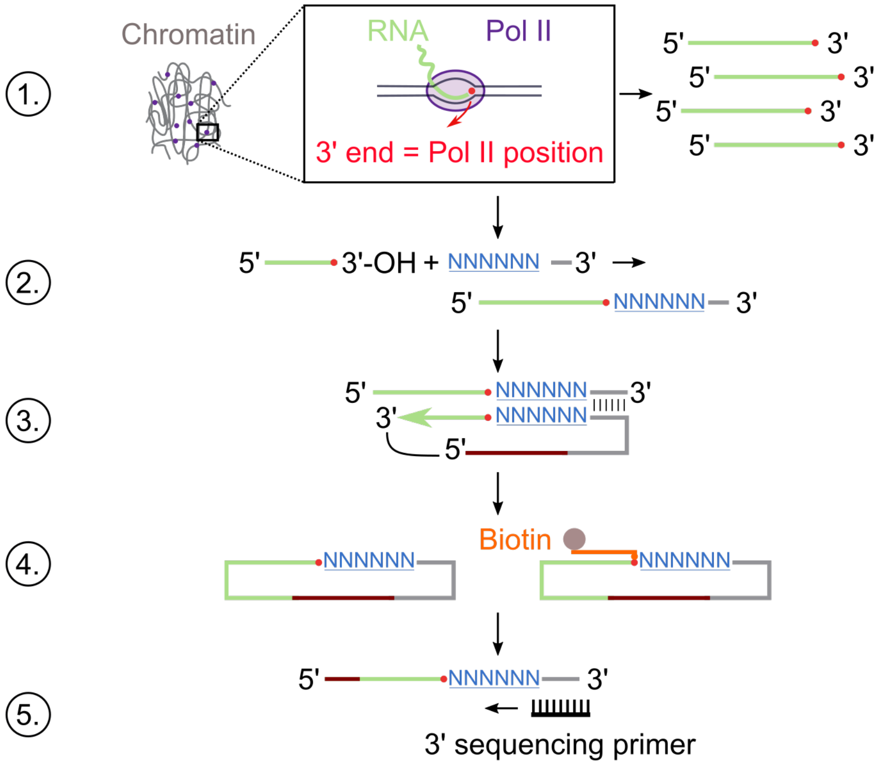


Figure 2.2: Overview of the Human NET-seq Method. 1. Chromatin isolation and RNA purification. 2. UMI/DNA linker ligation, RNA fragmentation (not shown), and size selection (not shown). 3. cDNA synthesis and circularization. 4. Depletion of mature RNAs. 5. PCR amplification (not shown) and 3′ sequencing.

5. Besides nascent RNA, the listed steps co-purify chromatin-associated mature RNA, such as snRNA, snoRNA, *ribosomal RNA*, and *transfer RNA*. Complementary hybridization oligonucleotides deplete the cDNA fragments from the twenty most abundant RNA species. This step includes biotinylated oligonucleotides annealing to the complement 3′ ends of the targets and streptavidin-coupled magnetic beads that remove the bound cDNA.

6. The last steps of the protocol include PCR amplification and sequencing using a single-end HTS technology.

MNET-SEQ The *mammalian NET-seq* (mNET-seq) method, developed in 2015 [169], is the adaption of the original yeast NET-seq protocol [27] for mammalian cells.

Following chromatin isolation, the *micrococcal nuclease* digests all accessible DNA and RNA not protected by a protein. The immunoprecipitation step uses a Pol II-specific antibody to enrich Pol II and associated nascent RNA. Before adapter ligation, a kinase reaction phosphorylates the 5′ ends of the nascent RNA fragments to gain strand-specific libraries.

The library preparation step selects RNA fragments between 35 and 100 nucleotides, ligates the 5′ and 3′ sequencing adapter, produces cDNA, amplifies the library, and performs sequencing using a paired-end HTS technology.

The standard protocol does not include UMI sequences or control measurements to enable the removal of PCR duplicates and unspecific antibody binding bias.

PRO-SEQ AND GRO-SEQ The *precision nuclear run-on sequencing* (PRO-seq) method, developed in 2016 [136], is the successor of the *global run-on sequencing* (GRO-seq) method [39] with increased resolution.

PRO-seq isolates the nuclei and prevents Pol II from continuing transcription by washing native nucleotides away. The stepwise addition of single biotin-labeled nucleotides allows Pol II to pursue elongation with a single or few labeled nucleotides. Next, nascent RNA is extracted, fragmented, and several streptavidin pull-down steps purify labeled RNA. The 5′-cap is removed and replaced by 5′ phosphorylation. The library preparation ligates the 5′ and 3′ sequencing adapter, produces cDNA, amplifies the library, and performs HTS.

GRO-seq performs the same steps with few differences. Pol II pursues elongation in the presence of bromouridine. In the following enrichment steps, antibodies directed against the bromouridine analog enrich the transcripts of interest. Because elongation continues, the resolution is in the order of tens of bases in contrast to the single-nucleotide resolution of PRO-seq [136]. Neither PRO-seq nor GRO-seq incorporates UMI sequences into their library preparation.

HiChIP

If completely stretched, the DNA would be 2 m long [174], but it fits into the tiny cell nucleus of human cells. Achieving such a dense structure without losing functionality requires a specific chromatin structure where DNA is associated with structural proteins that eventually form chromosomes. This 3D genome organization is tightly regulated [23].

To study the 3D genome, different experimental methods [126, 229] that measure contact frequencies between different genomic loci exist. Most methods create hybrid molecules, which contain the genetic information of regions with physical interaction. First, the 3D chromatin structure is stabilized by *in vivo* crosslinking, followed by DNA digestion with restriction enzymes. The DNA fragments near others are re-ligated and form hybrid molecules with the genetic information of genomic loci that were in physical proximity at the beginning of the experiment. The computational analysis identifies a putative long-range interaction if DNA fragments occur from non-adjacent loci of the original linear genomic sequence.

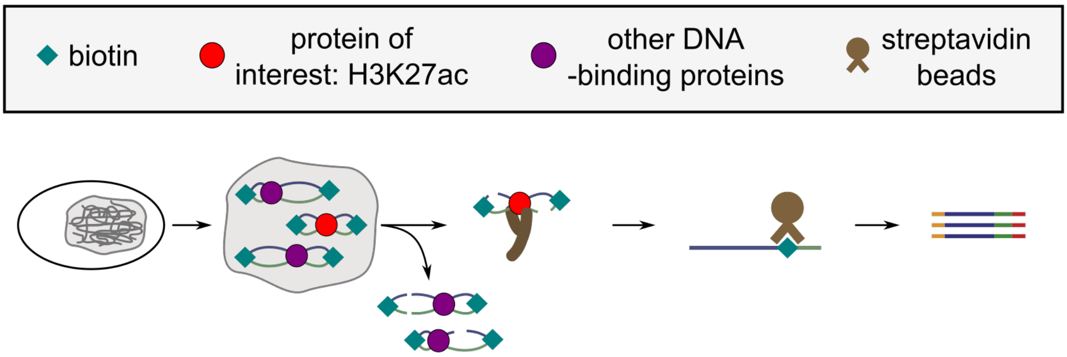


Figure 2.3: Overview of the HiChIP Method. 1. Crosslinking of cells using formaldehyde. 2. Nuclei isolation and generation of *in situ* Hi-C contacts. Generation of Hi-C contacts requires two steps. First, DNA digestion with a restriction enzyme leaves a 5′ overhang. Second, the 5′ overhang is filled with a biotinylated nucleotide residue before re-ligation. 3. Dissolving the nuclei. 4. ChIP and streptavidin beads sequentially enrich for Hi-C contacts marked with H3K27ac and biotin. 5. Transposase-mediated on-bead library construction (not shown). 6. PCR amplification and paired-end sequencing.

Hi-C [126] is the most popular and comprehensive method, quantifying all pairwise contacts in the genome using an enrichment strategy for hybrid molecules, known as Hi-C contacts. Hi-C contacts incorporate biotin-linked nucleotides during re-ligation. Those biotin-linked ligation junctions have a high affinity towards streptavidin, which is used for their enrichment, followed by paired-end HTS. Although widely used, Hi-C requires high sequencing depths to understand the genome's architecture at higher resolution.

Other methods focus on factor-directed [68, 159] or locus-specific [94] interactions. HiChIP [159] combines *in situ* Hi-C together with chromatin immunoprecipitation (Figure 2.3) to enrich 3D contacts that are associated with a protein of interest, such as CTCF, cohesin, or YY1 [159, 246]. Antibodies directed against more general factors, such as Pol II and H3K27ac, enrich contacts between actively transcribed regions, including regulatory interactions between promoters and enhancers. In contrast to Hi-C, the HiChIP method performs the sequencing library preparation with Tn5 transposase.

2.4.2 *Long-read Sequencing*

The major limitation of the HTS technologies is the short sequencing read length between 50 to 300 bp. In 2011 and 2014, the companies *Pacific Biosciences of California, Inc.* and *Oxford Nanopore Technologies Limited* (ONT) released their first sequencers allowing long-read sequencing [2]. Although both methods have become increasingly popular, this section focuses on ONT applied in this study.

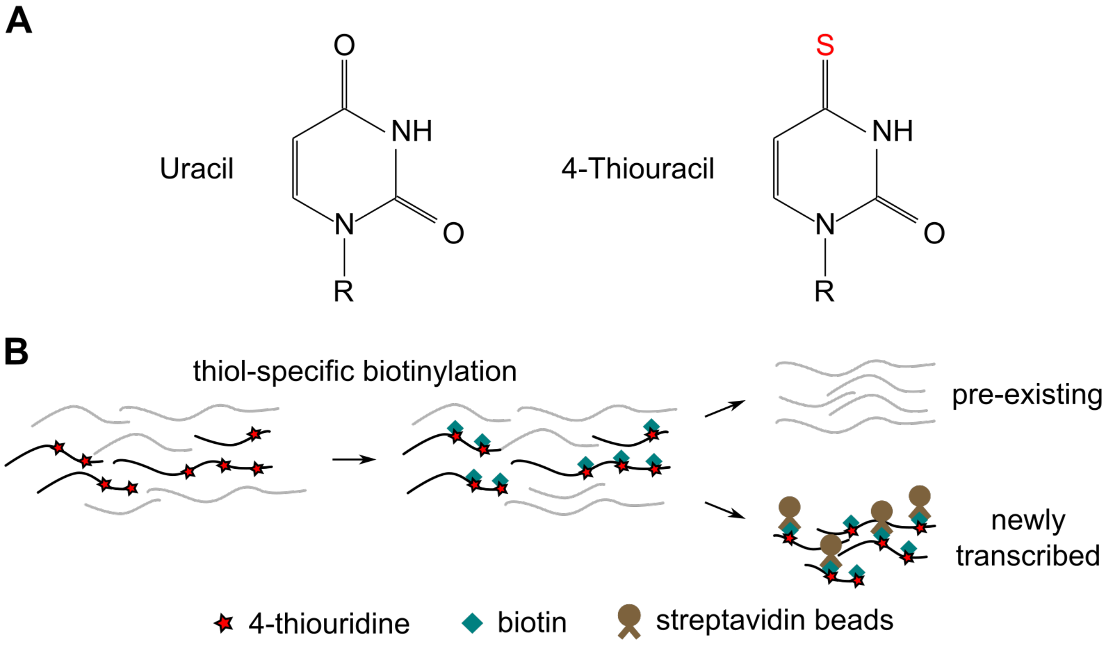


Figure 2.4: Metabolic Labeling using 4-thiouridine. (A) The structural formula (drawn with chemfig [226]) of uracil and 4-thiouracil, which are attached to a ribose ring (residual R, not shown), known as uridine and 4-thiouridine. **(B)** Metabolic labeling incorporates 4sU instead of uridine into RNA for newly transcribed RNAs. The thiol-specific biotinylation, followed by magnetic pull-down, separates pre-existing RNA from the labeled newly transcribed RNA.

ONT's flow cells contain two compartments of ionic solutions separated by a membrane with individual nanopores [97]. The constant voltage difference between both compartments produces an ionic current measured by a sensor. The sensor continuously monitors the changes in the ionic current produced by the controlled passage of DNA or RNA molecules. Advanced machine learning algorithms characterize and translate the current changes into long sequences of nucleotides (Figure 14.1A).

The long-read length between 500 bp and 2.3 megabases [2] can facilitate some applications, including assemblies, structural variants, and isoform identification [230].

2.4.3 *Metabolic Labeling of RNA*

Metabolic labeling of RNA, first applied in 2005 [28], allows the study of RNA metabolism rates and transient transcriptomes genome-wide [87, 211, 235]. The integration of a tag during transcription, mostly the sulfur-containing uridine analog *4-thiouridine* (4sU, Figure 2.4A), enables the differentiation between newly transcribed and pre-existing RNAs. Labeling duration varies and depends on the specific application.

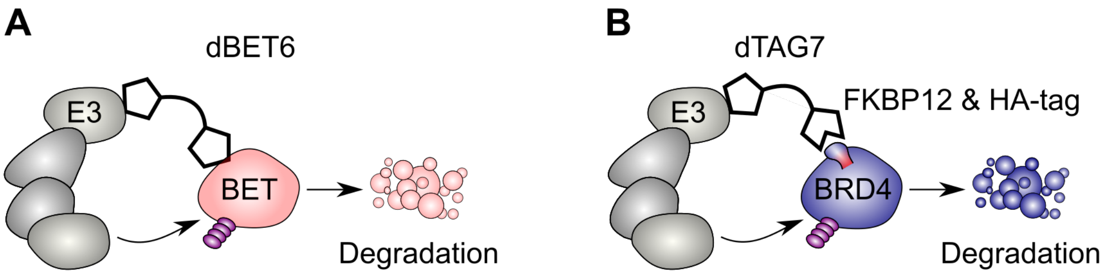


Figure 2.5: Protein Degradation Strategies. The target protein is ubiquitinated (purple) and proteasomal degraded. **(A)** The dBET6 drug induces pan-BET protein degradation. **(B)** *Proteolysis-targeting chimeras* compound dTAG7 induces BRD4 degradation in the K562 dTAG-BRD4 cell line. Degron tag consists of $FKBP12^{F36V}$ and *human influenza hemagglutinin-tag* (HA-tag).

For the isolation of 4sU-labeled RNA, the thiol group is biotinylated and separated with magnetic beads (Figure 2.4B). Other applications [87, 194, 208] detect point mutations caused by the chemical conversion of 4sU into a cytosine analog.

2.4.4 *Protein Knockdown Strategies*

The temporal or permanent loss of a protein is enforced in functional genomics to identify protein functions. Standard techniques are gene knockout or knockdown experiments that affect the respective protein-coding gene. Knockout strategies produce genetically modified ineffective gene versions at the DNA level. In contrast, most knockdown experiments reduce the gene's RNA expression or translation, for example, using short complementary oligonucleotides that block the gene transcription or RNA translation into a functional protein.

Although knockdowns are perceived compared to knockouts as transient, effective protein reductions require treatment times of many hours or days. Cells are dynamic systems capable of reacting to different environmental changes, leading to cell adaptation and compensation effects that can mask the direct protein functions.

Systems that avoid the shortcomings of long treatment times act at the protein level within hours, such as small molecular inhibitors and targeted protein degradation strategies [181], such as *proteolysis-targeting chimeras* [163]. Small molecular inhibitors bind competitively to the protein's domain and block domain-specific functions of a protein [59].

Proteolysis-targeting chimeras induce targeted protein degradation by the cellular ubiquitin-proteasome system. The key idea is to use small bifunctional molecules that transiently bring the target protein and an E3 ligase into spatial proximity.

The proximity leads to target ubiquitination and proteasomal degradation. The following paragraphs introduce two degrader types applied in this study.

dBET6

The dBET6 drug contains two active domains and a linker [249], inducing pan-BET protein degradation upon treatment (Figure 2.5A). On the one hand, the degron binds BET proteins, including BRD2, BRD3, BRD4, and BRDT, using an active domain structurally similar to JQ1 [59]. JQ1 is a small molecule inhibitor that binds the bromodomains of BET proteins. On the other hand, the active domain is structurally similar to thalidomide which binds the E3 ligase cereblon [205].

dTAG7

This study applies dTAG7 treatment for BRD4-specific degradation [163] in a K562 dTAG-BRD4 cell line (Figure 2.5B). In contrast to dBET6, the dTAG system requires the insertion of the degron tag $FKBP12^{F36V}$ in-frame with BRD4 using a *clustered regularly interspaced short palindromic repeats* (CRISPR)-Cas9-mediated locus-specific knock-in. The compound binds BRD4's degron tag and induces specific proteasomal degradation.

3 COMPUTATIONAL BACKGROUND

This chapter introduces basic computational methods used to analyze HTS data. The first part focuses on mapping and standard normalization strategies essential for quantitative comparisons within or between samples. The second part introduces the prerequisites for differential analysis of HTS data between biological conditions. Finally, a method used for functional interpretation of the results is presented.

3.1 MAPPING HTS DATA

Section 2.4.1 describes short-read DNA sequencing which infers a sequence of base pairs from a DNA fragment, known as sequencing read. Mapping a sequencing library, which consists of millions of short sequencing reads, to the reference genome and determining their pairwise sequence alignments is crucial for processing HTS data. For a review on sequencing alignments, see [52].

The main task of an alignment tool is to align a large set of relatively small sequences (sequencing reads) to one large sequence (reference genome) with high sensitivity and manageable computational resources. Different factors make this process computationally intense. Typical for all HTS assays are mismatches introduced by genetic variations and sequencing errors. Furthermore, a more specific challenge exists for transcriptomic data, where most sequencing reads map to non-contiguous genomic regions caused by splicing of the RNA [86].

Many different alignment tools emerged in the last decade [50, 119, 123]. The essential steps of most applications are listed below.

1. Building a reference genome index.
2. Searching for substrings (seeds) of the sequencing read in the reference genome.
3. Performing a pairwise sequence alignment.

The step that mainly distinguishes memory and runtime usage is the data structure used to build a reference genome index. A reference genome index allows a fast lookup to considerably reduce the list of candidate alignment locations. Modern tools such as Bowtie2 [119] or STAR [50] use either a *FM-index* [57], which is based on the *Burrows-Wheeler Transform* [21], or a suffix-array [139].

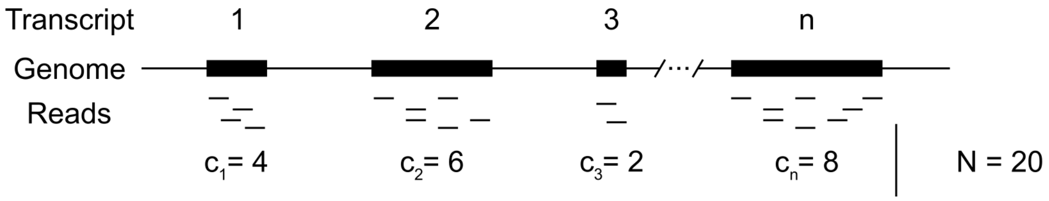


Figure 3.1: HTS Data Example. Schematic representation of HTS data, where N sequencing reads map to $i \in \{1, \ldots, n\}$ transcription units. The c_i value represents the sum of sequencing reads that map to transcription unit i.

No single alignment tool fits all applications. Choosing the right software tool depends on individual conditions such as data type, library size, memory resources, and available time. However, a general trend shows that STAR is commonly applied for transcriptomic data, such as RNA-seq. In contrast, Bowtie2 is more popular for genomic data, for example, from ChIP-seq experiments. Both applications are respectively applied by the ENCODE consortium pipelines [35] and performed well in a recent benchmark study [161].

3.2 NORMALIZING HTS DATA

After mapping, the sequencing reads are counted for regions of interest, such as transcription units or the binned genome. For simplicity, the following paragraphs will primarily refer to transcription units. However, all approaches apply to the binned genome or other regions of interest. Next, the normalization step corrects the HTS data for biases introduced by different sequencing depths, region lengths, or both.

As schematically depicted in Figure 3.1 for an HTS sample, N sequencing reads are distributed across n transcription units. The raw count c_i reports the sum of the sequencing reads assigned to transcription unit $i \in \{1, \ldots, n\}$. Previous studies introduced strategies to correct differences between HTS samples because the length l_i of the transcription unit i and the number of sequencing reads vary substantially within or between samples. Each strategy calculates a scaling factor to correct the raw count value of c_i. This section explains common normalization strategies used in this study and highlights their limitations and application situations using three mock replicate measurements for demonstration purposes (Figure 3.2A). The transcription units 1, 2, and 3 are equally expressed within- and between samples but different sequencing depths, transcript lengths, and outlier measurements mask this ground truth. The raw counts of *Sample B* vary in sequencing depth, whereas *Sample C* contains an outlier measurement for transcription unit 4. All types of variations are typical differences in HTS replicate measurements.

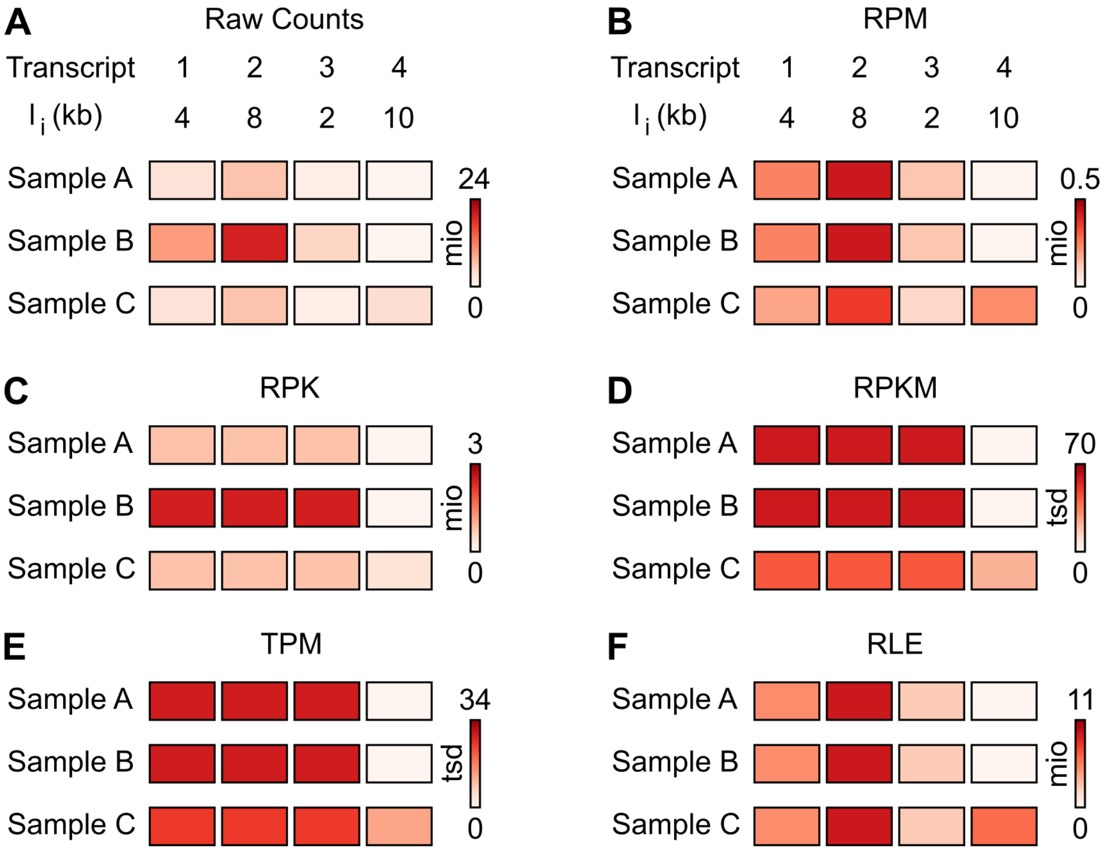


Figure 3.2: HTS Data Normalization. (A) The example contains the three mock replicate measurements labeled as *Sample A*, *Sample B*, and *Sample C* from four transcription units $i \in \{1, \ldots, 4\}$ with different lengths (l_i in *kilobases* (kb)). **(B-F)** Normalized count values obtained after **(B)** RPM, **(C)** RPK, **(D)** RPKM, **(E)** TPM, or **(F)** RLE normalization. Counts are reported in *millions* (mio) or *thousands* (tsd).

3.2.1 *Reads per Million*

The most common normalization strategy corrects a sample for different sequencing depths as the expected raw count of a transcription unit increases with sequencing depth. The *reads per million* (RPM) normalization strategy computes a *"per million"* scaling factor

$$\alpha_{RPM} = \frac{1}{N} \cdot 10^6, \tag{3.1}$$

which depends on the number of mapped sequencing reads $N = \sum_{k=1}^{n} c_k$. The normalized RPM value c_{RPM_i} of transcription unit i is calculated by $c_{RPM_i} = c_i \cdot \alpha_{RPM}$. Figure 3.2B shows that the RPM normalized values of all transcription units across *Sample A* and *Sample B* are equal because the observed differences in the raw count data (Figure 3.2A) are proportional to the sequencing depth differences. However, this strategy neglects the high variability between transcript lengths and is unsuited for within-sample comparisons.

3.2.2 *Reads per Kilobase*

During the library preparation of most HTS experiments, each transcript is fractionated into small pieces for sequencing. As long transcripts produce more small pieces, the number of expected sequencing reads increases with transcript length. In order to enable comparisons between transcription units within a sample, the *reads per kilobase* (RPK) normalization strategy computes a *"per kilobase"* scaling factor

$$\alpha_{RPK_i} = \frac{1}{l_i} \cdot 10^3 \tag{3.2}$$

for each transcription unit $i \in \{1, \dots, n\}$, using the transcription unit length l_i. The normalized RPK value c_{RPK_i} of transcription unit i is calculated by $c_{RPK_i} = c_i \cdot \alpha_{RPK_i}$. In Figure 3.2C, the RPK normalized values of the transcription units 1, 2, and 3 within a sample are equal because the observed differences in the raw count data were proportional to the transcript length differences. Although this normalization strategy accounts for different lengths among the transcription units, it neglects sequencing depth differences and is not suitable for comparisons between samples.

3.2.3 *Reads per Kilobase Million*

The *reads per kilobase million* (RPKM) normalization [157] strategy combined the previously discussed methods, correcting differences in sequencing depth and transcript length.

The scaling factor

$$\alpha_{RPKM_i} = \frac{1}{l_i \cdot N} \cdot 10^9 \tag{3.3}$$

is computed for each transcription unit $i \in \{1, \ldots, n\}$, where l_i reports the transcription unit length and N the total number of sequencing reads that map uniquely to the reference genome. The normalized RPKM value c_{RPKM_i} of transcription unit i is calculated by $c_{RPKM_i} = c_i \cdot \alpha_{RPKM_i}$.

Previous work from Wagner et al., 2012 [237] showed that the sum of normalized values $\sum_{k=1}^{n} c_{RPKM_k}$ is not relative to the RNA molar concentrations. This problem is caused by the denominator N, the total number of mapped sequencing reads, which has no biological interpretation but characterizes a specific sequencing run. In the example (Figure 3.2D), the sum of normalized counts in *Sample C* varies considerably compared to the remaining samples, introducing inconsistencies that could cause inflated statistical significance values in between-sample comparisons. A closely related alternative that is not biased in this way is presented in the following paragraph.

3.2.4 *Transcripts per Kilobase Million*

The *transcripts per kilobase million* (TPM) normalization [237] strategy computes the scaling factor

$$\alpha_{TPM_i} = \alpha_{RPK_i} \cdot \frac{1}{\sum_{k=1}^{n} c_{RPK_k}} \cdot 10^6 \tag{3.4}$$

for each transcription unit $i \in \{1, \ldots, n\}$, using the scaling factor α_{RPK_i} (Formula 3.2) and the sum of all RPK normalized values. The normalized TPM value c_{TPM_i} of transcription unit i is calculated by $c_{TPM_i} = c_i \cdot \alpha_{TPM_i}$.

Like the RPKM normalization strategy, the TPM values account for differences between transcription unit lengths and sequencing depths. However, the sum of all normalized values $\sum_{k=1}^{n} c_{TPM_k}$ is proportional to the relative RNA concentrations [237]. In the example (Figure 3.2E), the sum of the TPM normalized values is constant among all samples increasing reliability for between-sample comparisons.

Notably, as shown for *Sample C*, one outlier measurement influences the RPKM and TPM values across the sample (Figures 3.2D and 3.2E). For this reason, methods that specialize in detecting significant changes between samples apply more robust normalization strategies that are less sensitive to outlier measurements, as described in the following paragraph.

3.2.5 *Median-of-ratios*

The *median-of-ratios* method, also referred to as *relative logarithmic expression* (RLE), is implemented by DEseq2 [132].

This method performs a joint normalization, considering n transcription unit measurements from m samples. Therefore, the C_{ij} value reports the sum of the sequencing reads assigned to the transcription unit $i \in \{1, \ldots, n\}$ in sample $j \in \{1, \ldots, m\}$. The estimated scaling factor α_j, known as the *size factor*, is used to correct the count value $C_{RLE_{ij}} = C_{ij} \cdot \frac{1}{\alpha_j}$ and is calculated for each sample as described below.

1. The normalization method computes a pseudo-reference sample

$$C_i^{geom} = \left(\prod_{j=1}^{m} C_{ij} \right)^{1/m}, \tag{3.5}$$

which contains the geometric mean values from the raw counts C_{ij} of each transcription unit across all samples.

2. For each sample j, the scaling factor

$$\alpha_j = \underset{\mathrm{i}}{\mathrm{median}} \frac{C_{ij}}{C_i^{geom}} \tag{3.6}$$

represents the median of the ratios between observed data C_{ij} and the pseudo-reference sample C_i^{geom} for all $C_i^{geom} \neq 0$.

In the depicted example in Figure 3.2F, the outlier measurement has no impact on the calculated scaling factors, leading to a more robust normalization strategy that correctly identifies no differences between all three replicate measurements for most transcription units. This approach is unsuitable for within-sample comparisons as no gene length correction is applied.

3.2.6 *Reference-based Normalization*

All previously described methods assume that the RNA molar concentrations are constant across all samples. In practical application, this assumption does not always apply. For example, perturbation experiments potentially influence the total amount of RNA produced by the cells. In these cases, no normalization strategy can detect global changes. Some HTS assays addressed this limitation in the past by incorporating spiked-in controls into the experimental protocol and analysis, for example, using material from another species [9, 170] or synthetic ERCC spike-ins [99]. This alternative reference set is likewise biased by the sequencing depth but contains the same RNA molar concentrations in each sample, fulfilling the assumption of the normalization strategies.

As described in Section 3.2.5, the reference-based normalization calculates the scaling factor α'_j for each sample $j \in \{1, \ldots, m\}$ considering an alternative transcription unit set $i' \in \{1, \ldots, n'\}$ and the corresponding count value $C'_{i'j}$.

The resulting normalized value for transcription unit i in sample j is calculated by $C'_{RLE_{ij}} = C_{ij} \cdot \frac{1}{\alpha'_j}$.

The DEseq2's *estimateSizeFactors* function performs reference-based normalization if the *controlGenes* parameter specifies the alternative transcription unit set.

3.2.7 *Normalization of ChIP-seq data*

ChIP-seq is an HTS assay that measures protein-DNA interactions and requires dedicated control measurements for normalization. A more detailed explanation of the experimental steps is available in Section 2.4.1. This paragraph focuses on the *fold-enrichment over matched input control* and the *Pol II-based normalization* strategies.

For a reference genome, divided into G equally sized buckets, c_g and b_g report the sum of sequencing reads that map to the bin $g \in \{1, \ldots, G\}$ in the sample and control experiment, respectively. The scaling factors α_{RPM} and α'_{RPM} are derived from the sample and control experiment for RPM normalization as described in Section 3.2.1.

FOLD-ENRICHMENT OVER MATCHED INPUT CONTROL The experimental immunoprecipitation step introduces the classical bias of ChIP-seq experiments. Systematic enrichment occurs from the unspecific binding of the applied antibody. One of the first ChIP-seq studies [100] showed that this background noise is not uniformly distributed across the genome but accumulates locally. Therefore, a matched control measurement (input control) using either no or an unspecific antibody is necessary to locally normalize the data and distinguish the signal from background noise [118].

The *fold-enrichment over matched input control* (FE) normalization strategy [258] is commonly applied if an input control is available. For each bin $g \in \{1, \ldots, G\}$ in the reference genome,

$$FE_g = \frac{\alpha_{RPM} \cdot c_g}{\alpha'_{RPM} \cdot b_g} \tag{3.7}$$

reports the ratio of the RPM normalized read counts from the sample and the input control experiment. This strategy is implemented in MAC2's *bdgcmp* function [258] using the *-m FE* parameter.

POL II-BASED NORMALIZATION Another more specialized normalization strategy is required if the studied protein of interest depends on the occurrence of another protein that is not constant across experiments.

One example is a Pol II-associated elongation factor that binds to the Pol II elongation complex. Significant Pol II occupancy changes between experiments influence the binding profile of the Pol II-associated elongation factor dramatically.

Therefore, a matched control measurement using an antibody directed against total Pol II is necessary to locally normalize the data and distinguish the sample signal from the underlying Pol II background signal.

If the factor is Pol II-associated and a matched Pol II data set is available, the *Pol II-based normalization* was applied. For each bin $g \in \{1, \ldots, G\}$ in the reference genome and the pseudo count of $pc = 1$,

$$logFE_g = \log \frac{\alpha_{RPM} \cdot c_g + pc}{\alpha'_{RPM} \cdot b_g + pc} \tag{3.8}$$

reports the logarithmic ratio of RPM normalized read counts from the sample and the Pol II control experiment. MAC2 [258] performed the normalization using the *bdgcmp* function with the parameter *-m logFE -p 1*.

3.3 DIFFERENTIAL ANALYSIS FOR HTS COUNT DATA

Identifying changes between two biological conditions requires suitable statistical models that approximate the underlying data, estimate parameters robustly, and detect relevant differences. Many methods use the negative binomial distribution to approximate HTS count data. This section defines the negative binomial distribution and provides an overview of an application that tests for significant differences between conditions.

3.3.1 *Negative Binomial Distribution*

The negative binomial distribution is a discrete probability distribution that models the number of k failures before the r-th success occurs in a number of independent Bernoulli trials [145]. The probability of a successful event is denoted by p and must be $0 < p < 1$. For a negative binomial distributed random variable $X \sim NB(r, p)$, the probability mass function is defined as

$$Pr(X = k) = \binom{k + r - 1}{r - 1} (1 - p)^k p^r \tag{3.9}$$

for the non-negative integers k and r.

3.3.2 *DEseq2*

The differential analysis identifies significant changes in count data across biological conditions. The simplest model compares two conditions, such as control vs. treatment. The two most popular methods that handle small replicate numbers are DEseq2 [132] and edgeR [199].

Both methods assume that the observed count data C_{ij} for transcription unit $i \in \{1, \ldots, n\}$ and sample $j \in \{1, \ldots, m\}$ was sampled from an underlying negative binomial distribution [73] with

$$\begin{aligned} C_{ij} &\sim NB(\mu_{ij}, \delta_i^2) \\ \mu_{ij} &= \alpha_j q_{ij} \\ \delta_i^2 &= \mu_{ij} + dispersion_i \cdot \mu_{ij}^2. \end{aligned} \tag{3.10}$$

The mean μ_{ij} depends on the sample scaling factor α_j (Section 3.2.5) and the quantity q_{ij}, which is proportional to the mean number of sequenced fragments. The dispersion parameter $dispersion_i$ and the mean μ_{ij} are modeling the variance δ_i^2. Based on these assumptions, both methods fit a *generalized linear model* with a logarithmic link

$$\log_2 q_{ij} = \beta_o + x_j \beta_T, \tag{3.11}$$

using the design matrix

$$x_j = \begin{cases} 0 & \text{if j is a control sample} \\ 1 & \text{if j is a treated sample} \end{cases} \tag{3.12}$$

and the coefficients β_0 and β_T. The coefficient β_0 reports the estimated expression strength and β_T the logarithmic fold change of gene i between conditions. The main difference between DEseq2 and edgeR are estimations of the scaling factors and dispersion values. Although both methods show a good performance in benchmark studies [210], this section focuses on approaches implemented by DEseq2, which are more popular than implementations in edgeR (DEseq2: 33,352; edgeR: 12,904; Pubmed citation February 1, 2022).

SCALING FACTOR ESTIMATION The *median-of-the-ratios* method is applied to identify the scaling factor α_j for each sample j as described in Section 3.2.5.

DISPERSION ESTIMATION Estimating of the dispersion parameter $dispersion_i$ is often unreliable due to the small number of replicate measurements for one transcription unit i. Therefore, DEseq2 shares information and assumes that transcription units with similar mean expressions have a similar dispersion. First, the dispersion of a gene i is estimated using the maximum likelihood estimation. Second, the approach fits a curve to the dispersion estimates of all transcription units, referred to as the trend curve. Finally, the method uses an empirical *Bayes* approach that shrinks the dispersion values to the trend curve. The shrinkage strength is optimized and depends on the sample size and the distance of the dispersion value to the trend curve.

DEseq2 fits the parametric trend curve

$$dispersion_{tr}(\overline{\mu}) = \frac{a_1}{\overline{\mu}} + a_0, \tag{3.13}$$

depending on the parameters a_0, a_1, and the normalized mean counts from all transcription units, using the formula

$$\overline{\mu}_i = \frac{1}{m}\sum_{j=1}^{m}\frac{C_{ij}}{\alpha_j}. \tag{3.14}$$

WALD TEST For the estimated logarithmic fold changes β_T, two hypotheses are formulated for each transcription unit:

- $H_0 : \beta_T = 0$, the null hypothesis, states that the logarithmic fold change is equal to zero.
- $H_1 : \beta_T \neq 0$, the alternative hypothesis, is the alternative to the null hypothesis.

DEseq2 performs the *Wald* test [238], which calculates

$$W = \frac{\beta_T}{SE\left(\beta_T\right)} \tag{3.15}$$

where SE estimates the standard error of β_T. The resulting z-statistics are compared to a standard normal distribution, resulting in P values. The *Wald* test's P values that pass the independent filtering step are adjusted for multiple testing. Both approaches are described in the following paragraphs.

MULTIPLE TEST CORRECTION If a calculated P value is below a statistical significance level, the null hypothesis is rejected, and the result is statistically significant. After rejecting or accepting the null hypothesis, two types of errors could appear. A type I error occurs if the null hypothesis is rejected when it is actually true (false positive), whereas the type II error accepts a wrong null hypothesis (false negative). The probability of a type I error is equal to the significance level. Therefore, significance levels of 5% or lower are the standard.

Although this probability is reasonably low for one test, the probability of observing a type I error increases with the number of tests performed. For example, if 10,000 transcription units are tested, 500 false positives are expected. Multiple test correction is required instead of further decreasing the significance level, which increases the type II error. DEseq2 implements an interpretation of the *Benjamini and Hochberg* procedure [16] which controls the *false discovery rate* (FDR) by adjusting the calculated P values for the number of tests performed [83].

1. All P values from n tests are sorted in ascending order, where P_i denotes the P value of rank $i \in \{1, \ldots, n\}$, with $i - 1$ of the P values being smaller or equal to P_i.
2. The *FDR adjusted p-value* (padj) is defined as

$$padj_i = \frac{n}{i} P_i.$$

3. The FDR threshold fdr, with $padj_i \leqslant fdr$, controls the number of expected false positive classifications to the total number of tests with a rejected null hypothesis.

INDEPENDENT FILTERING The independent filtering step aims to decrease the number of performed tests by excluding transcription units with low mean sequencing read counts. Therefore, DEseq2 considers only transcription units with a normalized mean count above an identified cutoff. First, independent filtering estimates a function that reports the number of significant hits depending on potential cutoffs. In this approach, the quantiles of the mean sequencing read counts. Second, the approach fits a curve to the data. Third, the cutoff is selected, maximizing the number of significant hits within one residual standard deviation [71]. This approach is valid only if the filter criterion is independent of the actual test statistic, in this case, because the average expression over all samples does not consider the biological conditions.

3.4 GENE ONTOLOGY TERM ENRICHMENT ANALYSIS

Typically, the differential analysis results in a list of transcription units reporting significant changes between conditions. The following step aims to link the results with prior knowledge and potentially gain new insights, such as an unknown process or function. This downstream analysis may include a *gene ontology* (GO) enrichment analysis which characterizes the identified set of transcription units. For this purpose, the GO consortium [37] developed a comprehensive database collecting knowledge regarding the functions of genes and their protein products. Each GO term contains a list of genes associated with the corresponding molecular function, cellular component, or biological process. The aim is to test if the list of deregulated genes from an omics experiment showed an over-representation for one or more of these GO terms. Different tools exist [72, 153], where most

- build a contingency table for each GO term,
- perform a *Fisher's* exact test [61], and
- correct the calculated P values for multiple testing, for example, by using the described *Benjamini and Hochberg* procedure (Section 3.3.2).

For a differential analysis that tested n genes, $A + C$ were differentially expressed. Out of n genes, $A + B$ are associated with the GO term of interest. D denotes the number of tested genes that are neither differentially expressed nor associated with the respective GO term. The corresponding contingency table (Table 3.1) summarizes these observations.

	differential	not differential	total
not GO term	A	B	$A+B$
GO term	C	D	$C+D$
total	$A+C$	$B+D$	$n = A+B+C+D$

Table 3.1: Contingency Table for GO Enrichment Analysis.

The one-tailed *Fisher's* exact test calculates the exact P-value to identify an over-represented GO term, with

$$P = \sum_{j=0}^{A} H\left(j, A+B-j, A+C-j, D-A+j\right), \tag{3.16}$$

using the hypergeometric distribution

$$H\left(a,b,c,d\right) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}, \tag{3.17}$$

for all $a, b, c, d \geqslant 0$ with non-negative integers.

PANTHER [153] is one of many tools performing over-representation analysis. The web application curates an additional GO-slim annotation database (GO-slim), containing only broader parent terms to facilitate the interpretability.

4 MATERIALS AND CONTRIBUTIONS

This multi-omics study analyses material obtained from public databases, unpublished data from the Mayer laboratory, and external collaborations. This section provides an overview of all materials used and collaborators' contributions.

HTS Data

This study contains various published and unpublished HTS assays, planned and conducted by

- Mirjam Arnold (MRA),
- Nicole Eischer (NE),
- Susanne Freier (SF),
- Dr. Olga Jasnovidova (OJ),
- Dr. Andreas Mayer (AM),
- and Bruno Reversade's laboratory.

Table B.1 lists all datasets and the corresponding experimental biologist that created the dataset using the previously listed acronym. The *Max Planck Institute for Molecular Genetics* sequencing facility performed the sequencing. Dr. Olga Jasnovidova developed the HiS-NET-seq and nascONT-seq methods. Dr. Andreas Mayer developed SI-NET-seq with input from Mirjam Arnold. Furthermore, the study reanalyzed published ENCODE [35] and GEO [54] data from external laboratories, including Bernstein, Bradner, Farnham, Graveley, Lis, Pavri, Schwalb, Shilatifard, Snyder, and Stamatoyannopoulos.

Databases and Software

Table B.2 and Table B.3 list databases and software applications used for this work. Not listed were application dependencies. The data processing pipelines for NET-seq and nascONT-seq were established and updated in collaboration with Martyna Gajos.

Cell Line and Degradation

This study analyzed data from different cell lines, including K562, K562 dTAG-BRD4, NIH 3T3, MOLT4, HCT116, THP-1, *mouse primary activated splenic B lymphocytes*, and primary fibroblast cells from patients.

Mirjam Arnold generated the K562 dTAG-BRD4 cell line from human K562 cells. The cell line expresses a tagged (dTAG) [163] version of BRD4 from its endogenous locus (Section 2.4.4). A more detailed description of the CRISPR/-Cas9 genome editing experiment is available in the corresponding publication [7]. Members of the Reversade laboratory collected the primary fibroblasts cells from patients.

Proteomic Data

Mirjam Arnold planned, conducted, and analyzed all mass spectrometry experiments. Related figures in this study emerged from combining and visualizing the already processed data tables from Arnold* and Bressin* et al. 2021 [7].

Western Blot

Mirjam Arnold planned, conducted, and created all western blot experiments and corresponding figures.

Illustrations

The Figures 2.5B, 11.3A, 15.3B, 15.6A, A.5A-A.5B, 15.4A, 15.4B-A.11B, A.13A-A.13B, A.13C, A.18, and A.19A-A.19B were used from the *Molecular Cell* Arnold* and Bressin* et al. 2021 [7] publication with minimal re-arrangements, font type and size adaptations. The Figures A.24A-A.24C and A.25A-A.25B were used from the *EMBO Molecular Medicine* Nabavizadeh* and Bressin* et al. 2023 [162] publication with minimal re-arrangements, font type and size adaptations.

* equal contribution

Part I

STUDYING POL II WITH HIGH RESOLUTION AND IMPROVED COVERAGE

5 MOTIVATION

The human NET-seq approach is a high-resolution Pol II profiling method that purifies nascent RNA transcripts and performs ultra-deep sequencing of 3′ ends with 100-200 million sequencing reads [146]. However, the published Pol II profile [148], investigating lowly transcribed regions, was derived from a library with approximately 766 million sequencing reads. Studying these lowly transcribed regions is of increasing interest for some research questions, but the required sequencing depth is unsuitable for most potential applications in functional genomics.

Example analyses that would benefit from an improved Pol II coverage are listed below.

1. Sample intense comparative studies with several replicate measurements between different conditions.

2. Identification, description, and comparison of lowly transcribed regions, such as enhancers and 3′ ends of active genes.

3. Elongation rate calculations of Pol II at individual genes.

These studies demand adjustments to the human NET-seq method that enriches nascent RNA more effectively and results in higher Pol II coverage. Therefore, a systematic investigation of NET-seq library compositions is required. Identifying potential limitations helps the collaborating scientists in the laboratory to implement the necessary steps required to improve the protocol.

HTS assays, including NET-seq, contain multiple experimental steps that systematically enrich or deplete library fragments for technical reasons, referred to as biases. These biases, if unrecognized, skew the analysis and interpretation of the data. Previous publications started to identify biases introduced during the library preparation steps of the human NET-seq method [69, 148] (Section 2.4.1). Establishing a data processing pipeline that systematically removes all potential biases and summarizes our current knowledge is essential for studying Pol II transcription reliably.

Independent of optimization efforts in data processing, less obvious biases could escape detection. For NET-seq, where no suitable control experiments exist, the systematic comparison with other high-resolution Pol II profiling methods potentially enables additional insights. Distinguishing between general Pol II features and NET-seq-specific observations allows the identification of potential artifacts that require caution when being interpreted.

Additionally, the systematic comparisons between those approaches require suitable methods. Because the Pol II distribution across genes is not uniform, previous studies developed Pol II-describing indices that compare Pol II occupancy during different elongation stages. However, new indices are needed to investigate and characterize less studied transcriptional stages, such as Pol II termination.

Complying with these specifications requires appropriate adjustments in the experimental NET-seq protocol, processing steps, and the development of complementary approaches to describe Pol II distributions. This part summarizes the corresponding methods and results derived from this motivation.

6 METHODS

This chapter describes how this study defines basic genome features such as actively transcribed genes and enhancer regions. Furthermore, the section presents the newly implemented and improved human NET-seq data processing pipeline to extract Pol II occupancy profiles. Finally, indices that describe Pol II occupancy are characterized.

6.1 IDENTIFYING ACTIVE TRANSCRIPTION UNITS

6.1.1 *Active Genes*

Actively transcribed genes were identified for each cell type and represented a subset of all genes from either human v28 or mouse M18 GENCODE annotations [67]. A gene was classified as active if transcript levels appeared in the corresponding RNA-seq experiments above a defined threshold. Table B.4 summarized the analyzed cell types and their corresponding RNA-seq data. The following steps describe the approach in more detail.

1. RSEM v1.3.1 [122] quantified the number of transcripts produced by each gene and isoform in single-end or paired-end mode using the STAR v2.5.3a [50] alignment tool.

2. The genes with a $TPM \geq 1$ (Section 3.2.4) were selected.

3. The last step refines GENCODE's annotation based on active gene isoforms by identifying the first and last active TSS and polyA site. An active gene isoform contributed at least 10% to the overall gene activity.

Gene Types

GENCODE's v28 biotype annotation classifies the protein-coding, *micro RNA*, snRNA, and snoRNA gene classes for human genes. Genes with lincRNA or antisense biotype annotation were merged and renamed *long non-coding RNAs* (lncRNAs). Genes coding for histone proteins were defined as a subset of protein-coding genes and identified using the HUGO Gene Nomenclature (ID: 864) [177].

6.1.2 *Active Enhancers*

Actively transcribed enhancer units were identified for the K562 and MOLT4 cell lines from annotated FANTOM5 enhancers [36]. Data sets were extracted from the HACER [239] database, which reported cell-type-specific FANTOM5 enhancer units and initiation sites identified by NRSA [240]. Because FANTOM5 did not list the MOLT4 cell line, a similar cell type Jurkat was selected instead. Both cell lines are from *immortalized human T lymphocyte* cells to study *acute T cell leukemia*.

6.2 PROCESSING HIS-/NET-SEQ DATA

The human NET-seq method captures chromatin-associated RNA and requires computational steps to derive a genome-wide quantitative Pol II occupancy track. For a detailed description of the corresponding experimental steps, see Section 2.4.1. This section focuses on a new implementation and refined pipeline version [7, 69] to obtain Pol II occupancy tracks from human NET-seq data, which was first described in Mayer et al., 2015 [148].

If not stated otherwise, data processing steps were implemented in Python using Snakemake v6.8.0 [156], Biopython v1.78 [32], pysam v0.16.0.1 [77] and NumPy v1.20.2 [82]. Table B.5 reports the parameter settings of the applied tools.

After sequencing, the obtained sequencing reads consisted of the UMI sequence (six or ten nucleotides) followed by the RNA fragment (Figure 6.1A). The first nucleotide after the UMI corresponded to the 3′ end of the purified RNA. For small RNA fragments, sequencing reads may harbor segments of the reverse transcriptase primer, which were identified and trimmed using cutadapt v3.4 [143].

Starcode v1.1 [261] collapsed identical fragments sharing the same UMI sequence to one consensus read, removing PCR amplified sequencing reads. Next, the 5′ read ends, corresponding to the UMI sequence, were trimmed, but the sequence information remained associated with the sequencing read. The obtained sequencing read fragments were aligned to the human reference genome (GRCh38.p12) [67] using the STAR aligner v2.7.3a [50] (Figure 6.1B).

Potential artifacts occur if the reverse transcriptase primer binds to a complementary DNA region instead of the DNA linker (Figure 6.1C). The received reverse transcriptase artifacts contained no DNA linker, and, hence, no random UMI was sequenced. The custom python script identified and removed a sequencing read when the UMI corresponded to the genomic sequence adjacent to the aligned sequencing read.

As NET-seq purifies all chromatin-associated RNAs with a 3′-hydroxyl group, computational data processing includes *in silico* masking of loci from abundant non-nascent RNA species.

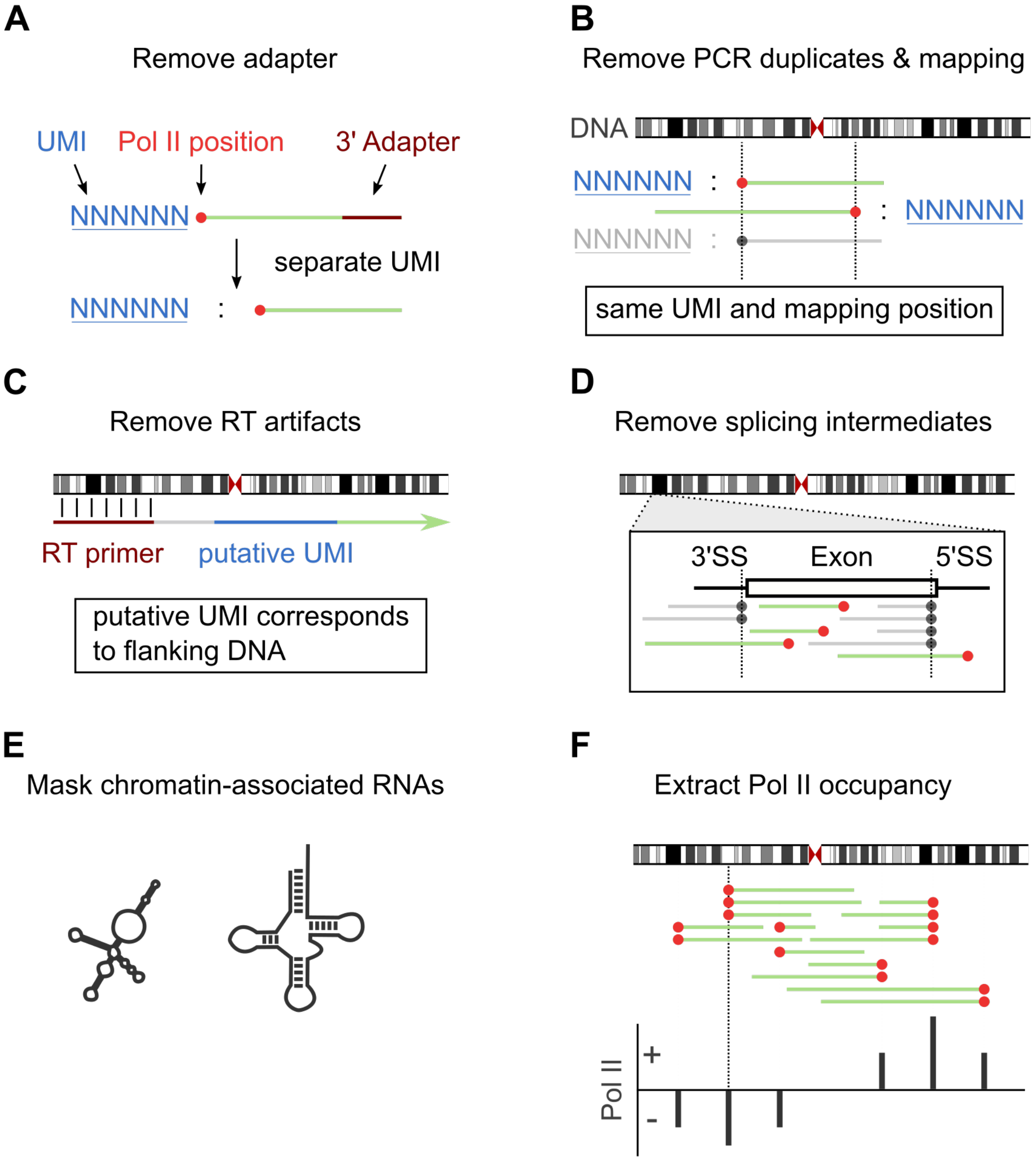


Figure 6.1: Scheme of NET-seq Data Processing.

Systematic bias occurred at active splice sites where sequencing reads mapped to the 3′ most nucleotide position of annotated introns and exons (GENCODE v28 [67], Figure 6.1D). Therefore, the next step masked these regions to exclude RNA intermediates produced during RNA splicing, such as the intron lariat, a by-product of splicing [86].

Furthermore, Bedtools v2.29.2 [185] was applied to mask suspect regions originating from other sources than Pol II (Figure 6.1E), including:

- transcribed regions of RNA polymerase I and RNA Polymerase III,
- chromatin-associated RNA species (Table B.6), and
- blacklisted regions from ENCODE [35].

Finally, the pipeline records the positions corresponding to the 3′ ends of purified RNA fragments for uniquely mapped sequencing reads (Figure 6.1F). The 3′ end mapping position is associated with the occupancy of one Pol II molecule because the enzyme's active center was catalyzing transcription at the corresponding nucleotide before transcription was interrupted.

A NET-seq experiment results in genomic tracks covering the positive and negative strands of the entire genome. The orientation of the mapped sequencing read determines if Pol II originates from the positive or negative strand (Figure 6.1F). Each genomic track reports a series of count data corresponding to the number of polymerases observed at a nucleotide position for a certain number of cells. Because transcription initiation occurs at different rates and elongation is a discontinuous process, Pol II occupancy signals vary substantially among nucleotides. The highest signals are observable at nucleotide positions where Pol II spends more time on average. These nucleotide positions are also known as Pol II pausing sites.

6.3 POL II INDICES

This section defines indices and approaches to compare Pol II occupancy systematically at specific transcriptional stages measured between conditions or different high-resolution Pol II profiling methods. The indices provide region-specific information simplifying the interpretation of potentially observed variations between measurements compared to standard correlation analyses. The *promoter-proximal maximum pausing position* and *pausing index* characterize features of promoter-proximal pausing, whereas the *termination zone length* and the *average transcription termination distance* describe the Pol II termination process (Figures 6.2A-6.2D).

These indices are defined based on an exemplary transcription unit, with the genomic coordinates of the transcription start site tss and the polyA site pa. The transcription unit localizes on the positive strand for notation simplicity, with $tss < pa$.

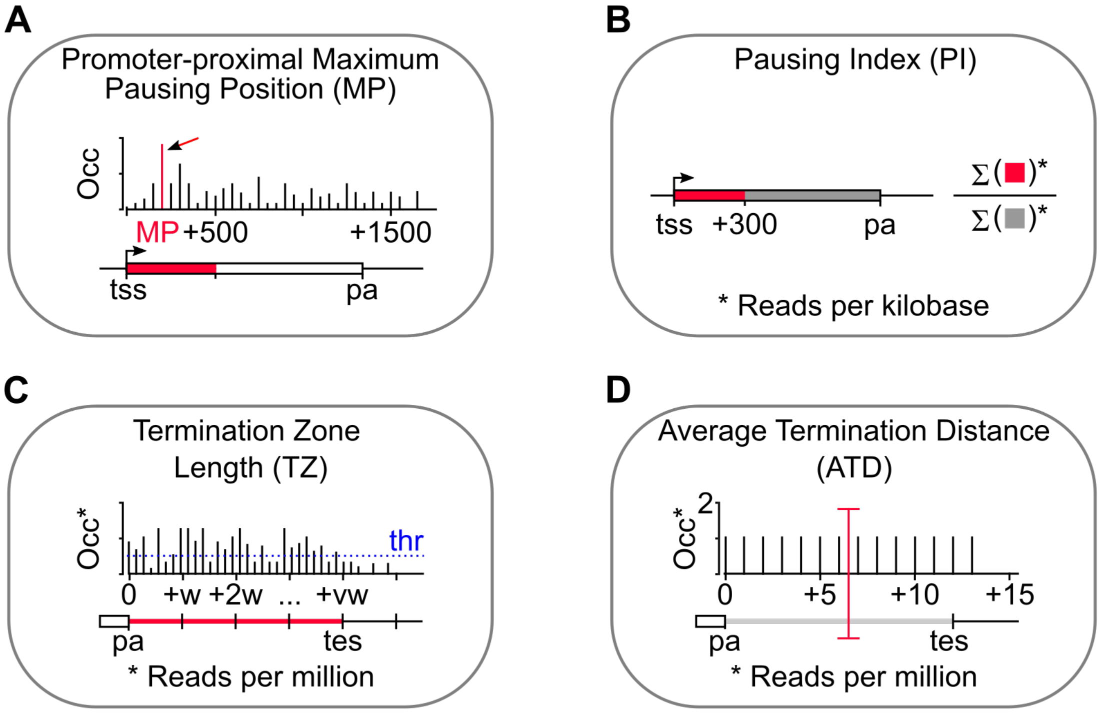


Figure 6.2: Definition of Pol II Indices. Visual representation of the **(A)** *promoter-proximal maximum pausing position*, **(B)** *pausing index*, **(C)** *termination zone length*, and **(D)** *average transcription termination distance*.

However, indices are likewise calculated for transcription units on the negative strand with few adaptations. Furthermore, each score depends on the Pol II occupancy track *Occ* generated in Section 6.2 for human NET-seq. The Occ_g value reports the Pol II occupancy measurement from the positive strand at the genomic position g, defined for all reference genome positions.

6.3.1 *Promoter-proximal Maximum Pausing Position*

Section 2.2.2 of this thesis introduces the concept of Pol II promoter-proximal pausing. Localization of the maximum Pol II signal in the promoter-proximal region can be used with other measurements to compare the properties of promoter-proximal pausing between samples. Therefore, the *promoter-proximal maximum pausing position* (MP) is defined as

$$MP = \arg\max \left\{ Occ_{tss+1}, Occ_{tss+2}, \ldots, Occ_{tss+a_0} \right\}, \tag{6.1}$$

where arg max reports the position with maximum signal in the region of size $a_0 = 500$, resulting in a score between $1 \leqslant MP \leqslant a_o$ (Figure 6.2A).

6.3.2 *Pausing Index*

Various indices were introduced in previous publications [160, 191, 254] to describe the ratio of Pol II in the promoter-proximal and gene-body regions. This study defines the *pausing index* (PI) as

$$PI = \frac{\sum_{k=tss}^{tss+a_1} Occ_k \cdot \frac{1000}{a_1}}{\sum_{k=tss+a_1}^{pa} Occ_k \cdot \frac{1000}{pa-tss-a_1}}, \tag{6.2}$$

with a promoter-proximal region length of $a_1 = 300$ (Figure 6.2B).

The numerator and denominator report the RPK (Section 3.2.2) normalized Pol II occupancy in promoter-proximal and gene-body regions. Transcription units with $PI > 1$ harbored more Pol II per nucleotide in the promoter-proximal compared to the gene-body region. Therefore, a high PI index indicated distinct promoter-proximal pausing.

6.3.3 *Termination Zone Length*

The termination zone of a transcription unit, also known as the termination window [211], informs about the region where Pol II termination likely occurs. This study defines the termination zone as the genomic region between the polyA site *pa* and the *termination end site* (TES) $tes = pa + TZ$, which depends on the *termination zone length* (TZ). The *termination zone length* is calculated by

$$TZ = v \cdot w, \tag{6.3}$$

using the bin size $w = 1,000$ and bin number v, which is derived from the number of consecutive bins downstream of the pa with RPM normalized Pol II occupancy above the threshold $thr = 0.2$ (Figure 6.2C). In other words, the length of the termination zone describes the distance between the polyA site and the genomic coordinate where the Pol II signal dropped below the threshold for the first time. The maximum value of v is selected, with

$$\sum_{k=(z-1)\cdot w}^{zw} Occ_{pa+k} \cdot \alpha_{RPM} \geqslant thr, \forall z \in \{1, \ldots, v\}, \tag{6.4}$$

using the calculated scaling factor α_{RPM} (Section 3.2.1). In order to avoid spillover effects, the termination zone could not span other transcription units and ended at least 1.5 kb upstream of the following active TSS.

6.3.4 *Average Termination Distance*

The *termination zone length* gives only limited insights into the distribution of Pol II in this region.

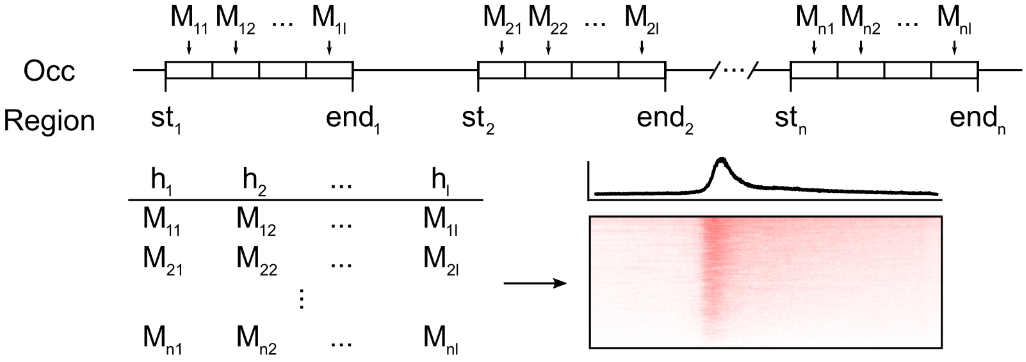


Figure 6.3: Calculation of Meta-gene Profiles and Heatmaps. The scheme shows heatmap visualization of RPM normalized Pol II occupancy measurements at n regions of length l. Each row reports measurements of one region. In contrast, each column reports the signals at one position across all regions. The profile represents the average measurements across all regions at the individual positions.

An additional score is required to summarize Pol II occupancy between the polyA site and the TES. The *average transcription termination distance* (ATD) calculates the weighted distance of Pol II to the polyA site, with

$$ATD = \frac{\sum_{k=0}^{TZ} \alpha_{RPM} \cdot Occ_{pa+k} \cdot k}{\sum_{k=0}^{TZ} \alpha_{RPM} \cdot Occ_{pa+k}}, \tag{6.5}$$

using the scaling factor α_{RPM} (Section 3.2.1) and the previously defined *termination zone length* TZ (Figures 6.2C and 6.2D). However, the score was only calculated if no other transcription unit overlapped the region up to 5 kb downstream of the polyA site.

6.4 meta-gene profiles and heatmaps

Alternative to the previously defined indices, a popular method to study Pol II occupancy in regions of interest are meta-gene profiles and heatmaps [15, 105, 227]. The meta-gene profile is computed for a set of n regions with constant regions length $l = end_i - st_i$ and the genomic start (st_i) and end (end_i) coordinates, where $st_i < end_i$ for all $i \in \{1, \dots, n\}$. For this set,

$$M_{ip} = Occ_{st_i+p} \cdot \alpha_{RPM} \tag{6.6}$$

reports the RPM normalized Pol II occupancy at position $p \in \{1, \dots, l\}$ in region i. The meta-gene profile at position p reports with

$$h_p = \frac{1}{n} \sum_{i=1}^{n} M_{ip} \tag{6.7}$$

the average signal of all regions.

In this work, the regions are derived from actively transcribed genes of the human GENCODE annotation (v28) [67] as described in Section 6.1.1. DeepTools2 [190] calculated and visualized the values of h_p and M_{ip} as metagene profiles and heatmaps, as shown in Figure 6.3.

As outliers skewed the calculated value of h_p considerably, deepTools2 removed all regions i that showed

$$M_{ip} \geqslant maxThr \tag{6.8}$$

for one or more $p \in \{1, \ldots, l\}$ using the parameter *--maxThreshold maxThr*. The outlier threshold ($maxThr$) was calculated for each sample and represented the 99.99-th percentile value of the signals in Occ in most analyses.

7 RESULTS

7.1 EXTRACTING RELIABLE POL II OCCUPANCY DATA

Human NET-seq requires different computational data processing steps to extract a genome-wide Pol II occupancy track. This study implemented an advanced version of the data processing pipeline based on previous publications [146, 148]. First, the pipeline mapped the sequenced reads of a NET-seq experiment to the human reference genome. The mapping followed the sequential removal of PCR duplicates, reverse transcription artifacts, and splicing intermediates (Figure 6.1 and Section 6.2). To correct the presence of the sequencing reads from RNA polymerase I, RNA polymerase III, and abundant RNA in the sample, the pipeline excluded sequencing reads mapped to loci transcribed by other RNA polymerases or chromatin-associated RNAs (Table B.6). Then, the genomic coordinate of each sequencing read's last 3′ end nucleotide was extracted and summed up in a Pol II occupancy profile.

This work tested the NET-seq processing pipeline using published NET-seq data from a human cell line, HeLa S3 (GSE123980 [148]). The resulting tracks (Figure 7.1A) showed strand-specific Pol II coverage over the whole genome at single-nucleotide resolution. The stringent removal of PCR duplicates (8%), splicing intermediates (<1%), and reverse transcriptase artifacts (not shown < 1%) reduced the number of informative Pol II sequencing reads only marginally (Figure 7.1B). A high fraction of sequencing reads were not aligned to the human reference genome (17%) or originated from sn/snoRNA genes (35%). The fraction of sequenced sn/snoRNAs increased by 16% after changing to another human cell line, K562 (OJ01, unpublished, Figure 7.1B). A critical step in removing the abundant sn/snoRNAs from NET-seq libraries was the subtractive hybridization performed during library preparations (Section 2.4.1). These observations indicated a less effective depletion of sn/snoRNAs without cell line-specific design of hybridization oligonucleotides and hence less effective enrichment of nascent RNAs.

7.2 OPTIMIZING POL II ENRICHMENT EFFICIENCY

The Pol II enrichment efficiency describes the fraction of sequencing reads with a unique mapping position at Pol II transcribed loci. Previous observations reported a low Pol II enrichment efficiency of 7-8% in HeLa S3 and 3% in K562 (Figure 7.1B). The limited enrichment efficiency decreased the measured Pol II coverage and forced deep sequencing of target samples.

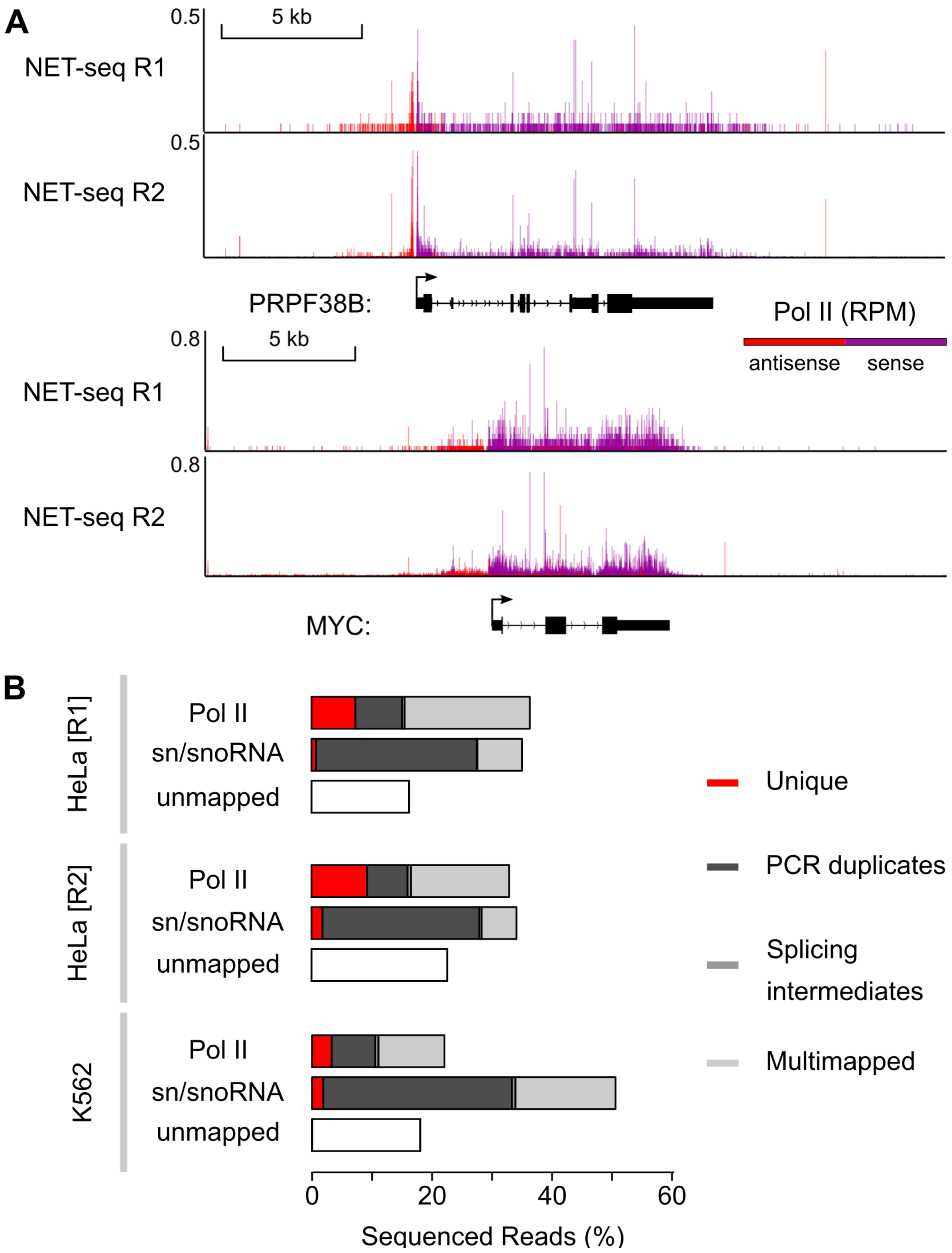


Figure 7.1: Adaptation of Human NET-seq in the K562 Cell Line. The figure shows the human NET-seq method in the HeLa S3 cell line for two biological replicates (R1 and R2). **(A)** Visual representation of sense (purple) and antisense (red) Pol II occupancy at single-gene examples. Data is RPM normalized (Section 3.2.1). **(B)** Fraction of sequencing reads mapping to Pol II transcribed regions (chromosomal and not listed in Table B.6), sn/snoRNA genes, or no locus (unmapped). Sequencing reads are, if possible, further classified into uniquely mapped (red), PCR duplicates (dark gray), splicing intermediates (gray), or without unique mapping position (light gray). The statistics show data from HeLa S3 and K562.

Comparative studies that require several replicates for different biological conditions would lack coverage under these circumstances.

Different optimizations improved the enrichment efficiency sequentially. The analysis showed no unique mapping position in the human reference genome for most small sequencing read fragments (Figure A.1A). Future experiments adjusted the size selection steps of the NET-seq protocol, and only RNA fragments with a minimum size of twenty nucleotides were selected for sequencing to decrease the number of unmapped sequencing reads.

Furthermore, short UMI sequence lengths could lead to collisions, where two RNA fragments obtain the same UMI sequence by chance despite being independent observations. AmpUMI [29], a software that calculates the expected number of UMI collisions, reported an expected 1-2% loss of informative Pol II sequencing reads due to the insufficient six nucleotide UMI sequence length (Figure A.1B). Implementation of a more extended ten nucleotide UMI sequence decreased the number of expected collisions to $< 0.6\%$ among informative Pol II sequencing reads.

New NET-seq data (OJ08 and OJ26, unpublished), which implemented these optimizations, reduced unmapped sequencing reads fractions from 18% to 5% (Figure A.1C). PCR duplicates decreased overall by 21% and among the Pol II sequencing reads by 4%. Although the optimizations doubled informative Pol II sequencing reads in K562 cells from 3% to 6%, the overall Pol II enrichment efficiency remained low. A possible explanation was the increased sequencing of sn/snoRNA transcripts by 8%. This analysis identified extensive sequencing of sn/snoRNA as the primary challenge for efficient Pol II enrichment in the human NET-seq protocol.

7.3 increasing coverage with the his-net-seq method

Neither cell line-specific subtractive hybridization strategies in HeLa S3 nor their implementations in K562 removed chromatin-bound sn/snoRNAs effectively. For this reason, a new NET-seq method incorporated an additional enrichment step that combined the original cell fractionation approach with metabolic labeling. This new high-sensitivity NET-seq approach, named HiS-NET-seq, was based on the original NET-seq protocol [146].

First, HiS-NET-seq (Figure 7.2A) labeled newly synthesized RNAs using the uridine analog 4sU (Section 2.4.3). Next, two RNA purification steps were performed, including cell fractionation and enrichment of 4sU-labeled RNAs. The cell fractionation isolates chromatin and associated RNA, whereas 4sU-labeled RNA enrichment excluded unlabeled mature RNAs. The 4sU labeling approach replaced the subtractive hybridization step of the original protocol. Finally, the method performs library preparation and 3'-end sequencing.

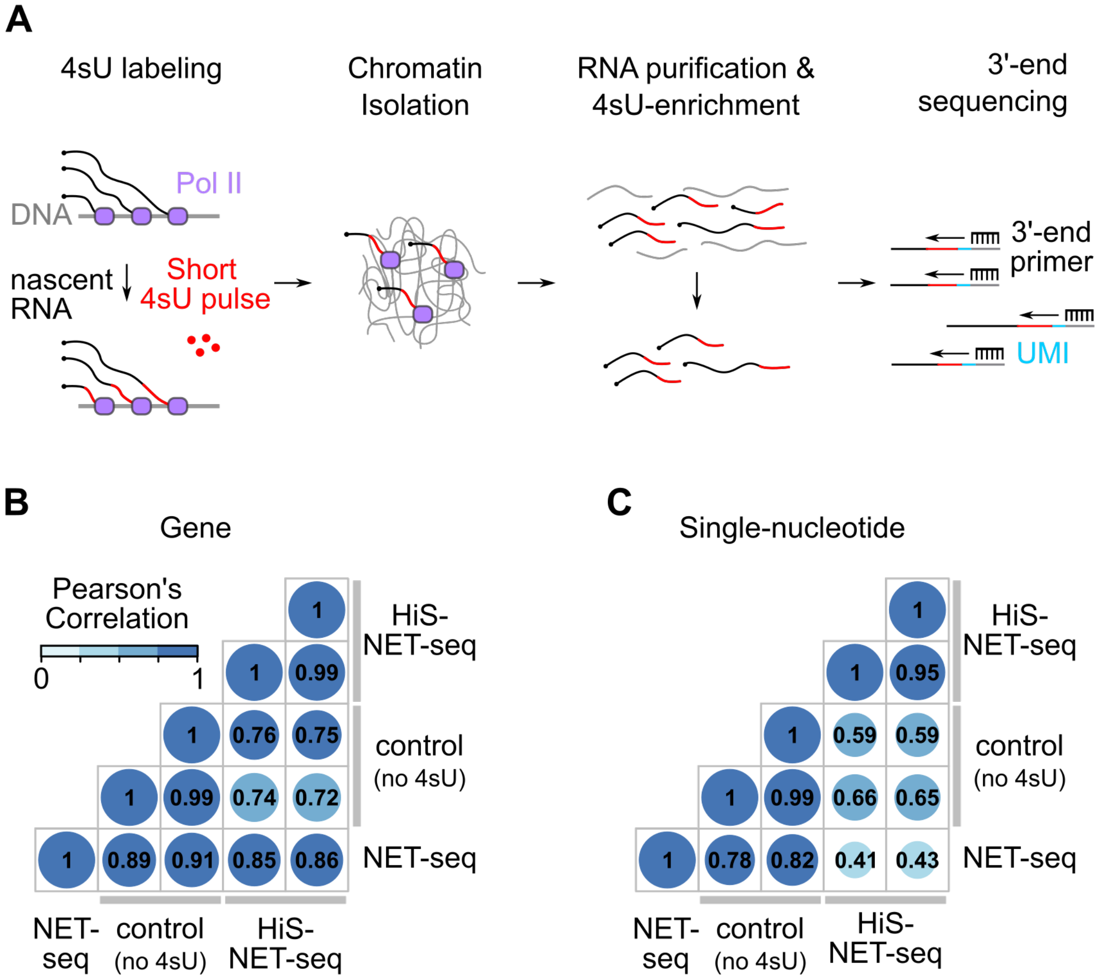


Figure 7.2: Overview of the HiS-NET-seq Method. (A) The uridine analog, 4-thiouridine (4sU), labels newly synthesized RNAs for 10 minutes. Rapid chromatin isolation and 4sU selection purify engaged Pol II and the associated RNAs. The 3′ ends are ligated to a DNA linker containing a mixed random sequence (10 nucleotides, blue) which serves as a UMI. **(B-C)** Pairwise comparisons of *Pearsons's* correlation between NET-seq, HiS-NET-seq, and the corresponding control experiment without 4sU labeling calculated for Pol II occupancy across **(B)** actively transcribed genes (n=11,149) and at **(C)** individual nucleotides (n=98,208,993). Data is *median-of-ratios* normalized (Section 3.2.5).

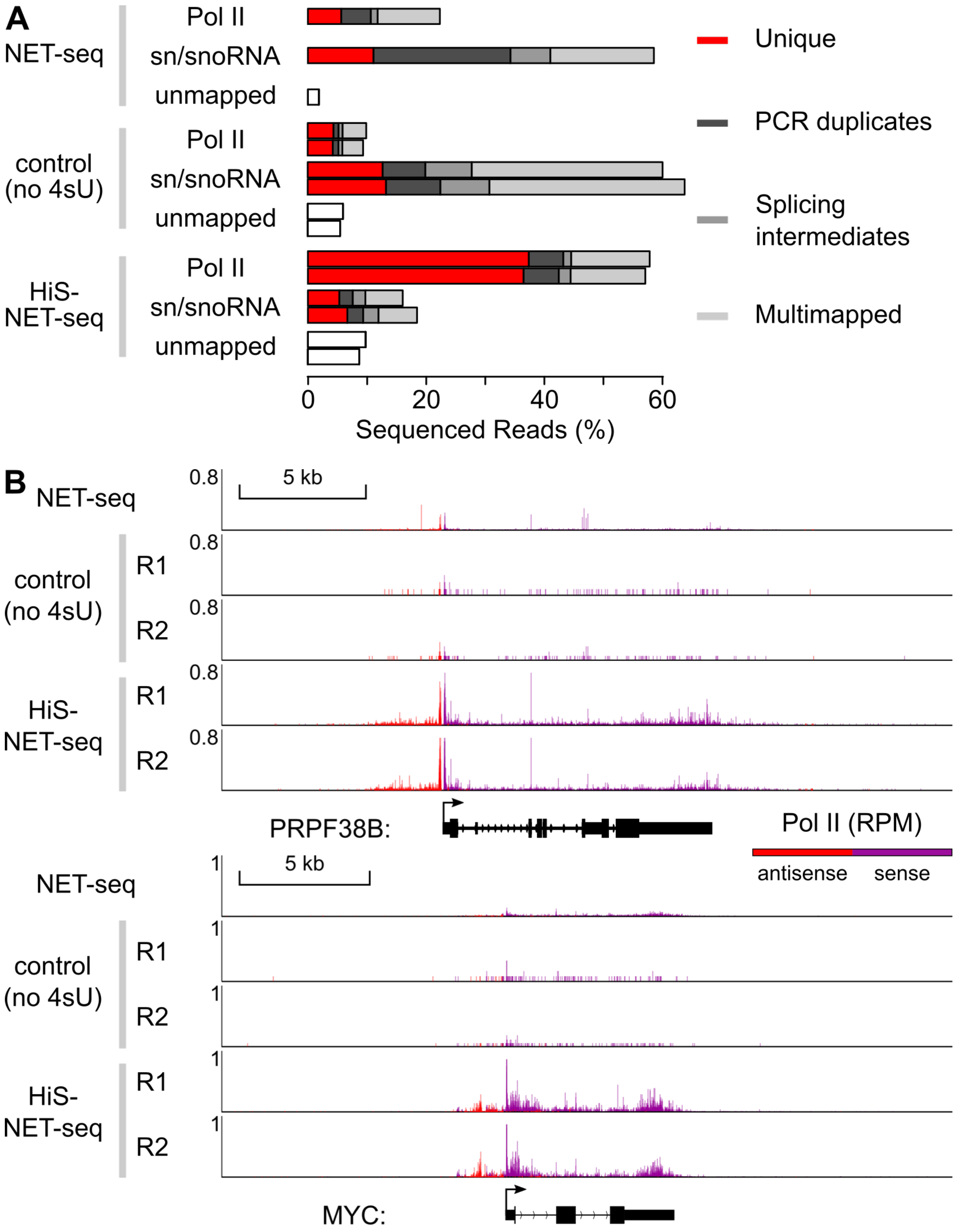


Figure 7.3: Pol II Coverage Gain using 4sU Labeling. Depicted are optimized NET-seq, HiS-NET-seq, and the respective control experiments without 4sU labeling in the human cell line K562. **(A)** Fraction of sequencing reads mapping to Pol II transcribed regions (chromosomal and not listed in Table B.6), sn/snoRNA genes, or no locus in the human reference genome (unmapped). Sequencing reads are, if possible, further classified into uniquely mapped (red), PCR duplicates (dark gray), splicing intermediates (gray), or without unique mapping position (light gray). **(B)** Visual representation of RPM normalized (Section 3.2.1) sense (purple) and antisense (red) Pol II occupancy at single-gene examples.

HiS-NET-seq used adjusted reverse transcriptase and PCR primers [75] to avoid PCR amplification of reverse transcriptase artifacts during the library preparation as tested and described in a previous NET-seq protocol variant [69].

The new method was tested in K562, comparing HiS-NET-seq experiments with 10 minutes of 4sU labeling and purification with control experiments without metabolic labeling (0 min: OJ90, OJ91; 10 min: OJ92, OJ93, unpublished). Compared to standard NET-seq, the 4sU labeling replaced the step of subtractive hybridization. The Pol II signal intensities measured at active genes and single-nucleotides across biological replicates were highly reproducible, as indicated by *Pearson's* correlation coefficients of at least 0.95 (Figures 7.2B and 7.2C). Notably, the correlation analysis revealed high correlations between standard NET-seq (OJ26) and HiS-NET-seq at actively transcribed genes (*Pearson's* correlation coefficient: r=0.85-0.86, Figure 7.2B). However, the Pol II distribution at the individual nucleotides was more distinct in HiS-NET-seq compared to NET-seq approaches without metabolic labeling (Figure 7.2C). A possible explanation was the 19-fold increase of nucleotides, covered by at least one Pol II molecule.

After 10 minutes of 4sU labeling, enrichment efficiency increased from 4% to 37%, mainly due to pronounced reductions of sequenced transcripts from sn/snoRNAs by 45% (Figure 7.3A). Pol II occupancy coverage gain was observable at single-gene examples such as *PRPF38B* and *MYC* (Figure 7.3B). HiS-NET-seq purified Pol II-related nascent RNA more effectively than previous NET-seq protocols. Improved enrichment led to a better signal-to-noise ratio and, depending on the sequencing depth, resulted in better Pol II coverage, decreased sequencing costs, or both.

4sU Selection Bias

Which biases were introduced by 4sU labeling and enrichment? The number of unique sequenced fragments, known as library complexity, increased by 1.7% (Figure A.2A), and the median length of mapped RNA fragments decreased by two nucleotides (Figure A.2B). 4sU selection led to a slight decrease in the uracil frequency of 1% (Figure A.2C). A potential explanation for a decline in sequenced uracil nucleotides was an increased conversion rate of uracil to cytosine (Figure A.2D), likely caused during PCR amplification by the 4sU analog. The high conversion rate (A>G and T>C) of 14% was similar to another study [87]. Labeling and purification of RNA with 4sU had only minor impacts on overall RNA library features.

7.4 IDENTIFYING METHOD INDEPENDENT POL II OCCUPANCY FEATURES

Pol II tracking at high resolution across the human genome has been a central objective for decades to study Pol II transcription and regulation.

However, the mostly applied ChIP-seq method fails to identify the fine structure of Pol II genome transcription due to limited resolution, low coverage, and a low signal-to-noise ratio. Furthermore, ChIP-seq does not differentiate between sense and antisense transcription. New methods addressed these problems and provided strand-specific and high-resolution genome-wide Pol II data. The critical differences between methods were enrichment strategies, bias control, and resolution (Table B.7). This section compares Pol II ChIP-Rx (GSE158965 [7]) with the more recently developed Pol II tracking methods, including HiS-NET-seq (OJ92, OJ93), human NET-seq (OJ26), PRO-seq (GSM1480327 [38], Section 2.4.1), and mNETseq (GSE123980 [80], Section 2.4.1).

New methods, such as qPRO-seq [103] and SNU-seq [151], were excluded from the benchmark because the preprints were neither complete nor peer-reviewed. Other methods, such as GRO-seq [39] and TT-seq [211], were not considered for the following reasons. GRO-seq is the precursor of PRO-seq but does not provide single-nucleotide resolution. TT-seq [211] performed no 3′ end sequencing but measured transcripts produced in a given time. The resulting data does not correspond to Pol II occupancy. This section compares the similarities and differences of Pol II distribution measured by different methods, focusing on HiS-NET-seq.

Correlation

Pol II occupancy at actively transcribed genes showed high Person's correlation coefficients among PRO-seq, HiS-NET-seq, and human NET-seq ($r \geqslant 0.81$, Figure 7.4A). For those methods, the measured transcriptional activity of Pol II was comparable and reproducible. Pol II ChIP-Rx signal likewise correlates with the previously listed methods, although correlation coefficients were lower ($r \geqslant 0.6$, Figure 7.4A). Surprisingly, mNET-seq data sets were distinct, showing low correlation with other high-resolution methods ($r = 0.42$-0.57) and Pol II ChIP-Rx ($r < 0.18$). All methods showed strikingly low correlations between methods considering individual nucleotides (Figure 7.4B). Notably, mNET-seq correlation values were additionally low among replicate measurements ($r = 0.42$). The results imply method-dependent differences in measured Pol II occupancy at single-nucleotide resolution.

Pol II Distribution

What are the systematic differences between the discussed methods? To answer the question and identify features of Pol II transcription, several Pol II-describing indices were applied or developed (Section 6.3), including *promoter-proximal maximum pausing position*, *pausing index*, *termination zone length*, and *average termination distance*.

A pronounced property of Pol II transcription during early elongation is promoter-proximal pausing (Section 2.2.2).

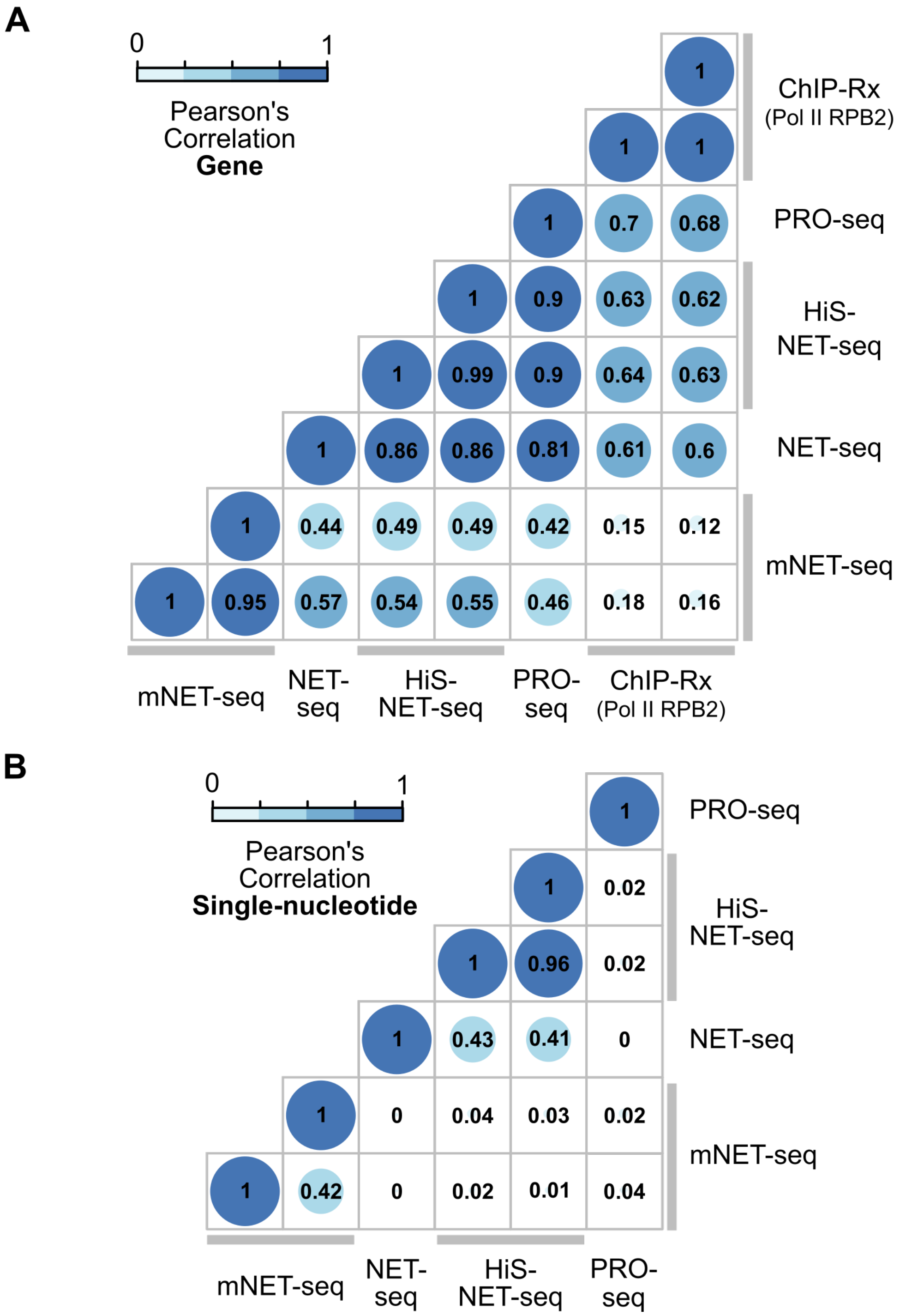


Figure 7.4: Correlation Between Pol II Profiling Methods. Pairwise comparisons of *Pearson's* correlation between ChIP-Rx, PRO-seq, HiS-NET-seq, NET-seq, and mNET-seq, calculated for Pol II occupancy across **(A)** actively transcribed genes (n=11,149) and at **(B)** individual nucleotides (n=98,208,993). ChIP-Rx data was measured using a *Pol II subunit 2* (RPB2) antibody and excluded for the single-nucleotide analysis. Data is *median-of-ratios* normalized (Section 3.2.5).

One feature describing this stage is the *promoter-proximal maximum pausing position* (MP), revealing the distance between TSS and the strongest Pol II signal in the promoter-proximal region. The median maximum signal occurred 80-81 (HiS-NET-seq), 87 (NET-seq), 106 (PRO-seq), and 151-166 (mNET-seq) nucleotides downstream of the annotated TSS (Figure 7.5A). Notably, measurements by mNET-seq are not consistent with the trend observed by other methods (Figure 7.5A).

The *pausing index* (PI) measured the transition from early to productive elongation [160, 191, 254], which describes the proportion of Pol II in the promoter-proximal vs. gene-body region. A high *pausing index* indicates proportionally stronger signals in the promoter-proximal areas. HiS-NET-seq and human NET-seq report higher median indices for active genes than PRO-seq and mNET-seq (Figure 7.5A, PI: 9-13 (HiS-NET-seq), 11 (NET-seq), 5 (PRO-seq), and 6-9 (mNET-seq)). The same trend was observable in meta-gene visualizations where average profiles summarize the Pol II occupancy over thousands of genes (Figure 7.5B and Section 6.4). Interestingly, PRO-seq, which does not measure stalled or arrested Pol II [136], showed the weakest promoter-proximal signal indicated by a comparably low median *pausing index* and Pol II occupancy in the respective area (Figures 7.5A and 7.5B).

The average Pol II occupancy around exon splice sites in gene-body regions was also method-dependent. HiS-NET-seq and NET-seq identified pausing around 5′ and 3′ splice sites. In contrast, PRO-seq shows no such pausing patterns (Figure 7.5B). The most intense signal occurred at 5′ splice sites detected by mNET-seq.

This study developed two new indices for the identification of termination features. Both the *termination zone length* (TZ) and *average termination distance* (ATD) described Pol II occupancy downstream of the polyA site. The TZ describes the region length with Pol II coverage where termination potentially occurs, in contrast to the ATD, which considers the distribution of Pol II in the respective region relative to the polyA site. All methods showed variable TZ with a median length of 4.8-4.9 (HiS-NET-seq), 3.8 (NET-seq), 7 (PRO-seq), and 3.9 kb (Figure 7.5A). A previous study identified a similar termination window with a median length of 3.3 kb using TT-seq [211]. However, the average distance between Pol II and the polyA site was considerably closer (Figure 7.5A, ATD: 1.7-1.9 (HiS-NET-seq), 1.6 (NET-seq), 2.3 (PRO-seq), and 1.4 kb (mNET-seq)).

Finally, all methods show bidirectional transcription at K562 enhancers annotated by FANTOM5 [36] (Figure 7.5B and Section 6.1.2).

Overall, most Pol II profiling methods show similar trends and features of Pol II transcription. However, this analysis also identified considerable discrepancies at single-nucleotide positions, including pausing site positions. These differences are highly relevant and should be considered in models and interpretations that concern, for example, pausing positions.

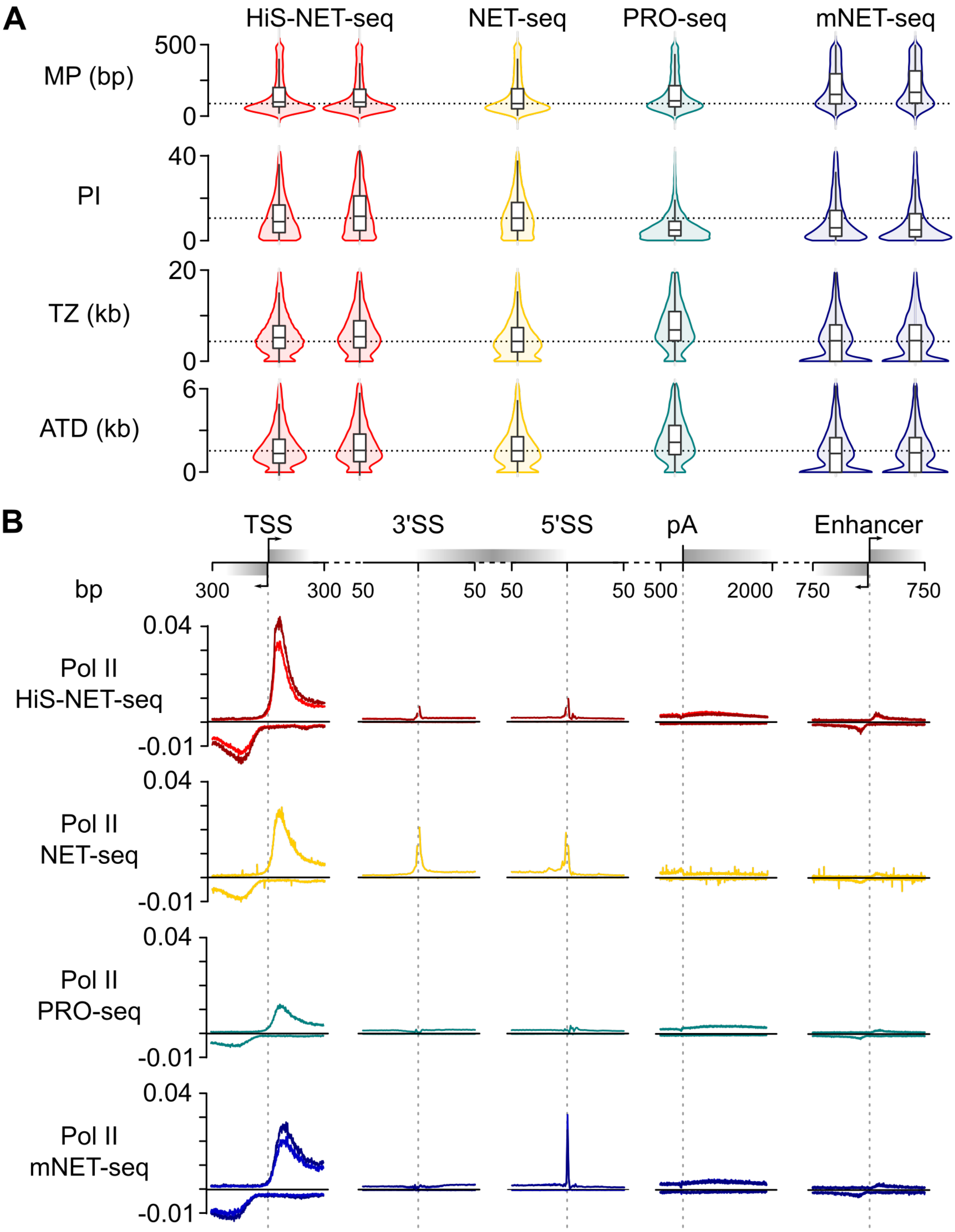


Figure 7.5: Comparison of Pol II Profiling Methods. Considered are actively transcribed non-overlapping (TSS to pA + 5 kb) protein-coding/lncRNA genes (n=8,124) with a minimum gene length of 1 kb and FANTOM5 [36] annotated enhancers (n=6,313) in human K562 cells. **(A)** Distributions of several indices describing Pol II transcription at genes (Section 6.3). **(B)** Mean Pol II occupancy for individual nucleotides at indicated regions, including exon regions (n=53,575). Excluded were exons that appeared first or last in a transcript and regions with signal outliers above the 99.99-quantile. Data is RPM normalized (Section 3.2.1). Masked were TSS, 3′ *splice site* (SS), 5′ SS, and polyA sites for HiS-NET-seq and NET-seq.

8 DISCUSSION

The human NET-seq method belongs to a group of experimental procedures that emerged over the last decade to track Pol II occupancy with high resolution across the human genome. These methods revealed new insights into Pol II distribution and regulation at different transcriptional stages [250].

The results presented in this part summarized the computational data processing steps and identified the advantages and limitations of human NET-seq. Investigation and optimization efforts resulted in HiS-NET-seq, a new method that resolved NET-seq's limitations by combining metabolic labeling and cell fractionation. Furthermore, a comprehensive benchmark analysis systematically compared Pol II distribution across established methods, revealing HiS-NET-seq as an alternative to standard NET-seq and other high-resolution Pol II profiling methods.

Confirming earlier findings [148, 249], the updated human NET-seq data processing pipeline identified and removed potential biases introduced during library preparation, including PCR duplicates, splicing intermediates, and reverse transcriptase artifacts. A new data processing step masks chromatin-associated RNAs and exposes to which extent NET-seq purifies these RNAs, especially sn/snoRNAs (Figure 7.1B). Previous work [148, 249] classified most of these transcripts as PCR duplicates and masked their source of origin. This observation emerged as a critical finding, provoking adjustments and changes in the experimental procedures.

The resulting HiS-NET-seq approach offers higher Pol II coverage by enriching recently synthesized 4sU-tagged RNAs, increasing Pol II-associated nascent RNA levels obtained after cell fractionation (Figure 7.3A). Sequenced libraries showed no composition or characteristic differences from standard NET-seq libraries (Figure A.2). The new method shared most Pol II occupancy features with NET-seq, PRO-seq, and mNET-seq (Figures 7.5A and 7.5B).

In a preprint [151], another research group recently described a similar method named SNU-seq, which combines 3′-RNA sequencing and 4sU labeling. In contrast to HiS-NET-seq, SNU-seq does not perform cell fractionation, limiting the purification of nascent RNA. Instead, SNU-seq enriches for polyA sites of mature RNAs, which will be of interest to other types of studies.

Furthermore, a direct comparison of HiS-NET-seq and other Pol II profiling methods revealed three advantages.

1. Computational processing steps and library preparation prevent or remove biases more effectively than other methods. The improved library preparation, described in Gajos et al. [69], reduced the number of sequenced artifacts produced by reverse transcriptase. Consistent with the results from published simulations [29], extending UMI sequences from six to ten nucleotides improved *in silico* depletion of the remaining artifacts (Figure A.1C). Most of the currently available and analyzed high-resolution Pol II data sets [39, 136, 169] incorporate no UMI sequences for bias correction, limiting their application in quantitative studies (Table B.7). Furthermore, HiS-NET-seq lacks artifacts introduced by unspecific antibody binding during immunoprecipitation, which is often present but overlooked by other Pol II profiling methods [27, 169, 182].

2. HiS-NET-seq showed the best reproducibility at single-nucleotide resolution for methods providing replicate measurements (Figure 7.4B). Low variability among replicates is advantageous for comparative studies, improving sensitivity and decreasing the demand for many replicate measurements [210].

3. The improvement of Pol II coverage reveals transcription at lowly transcribed loci. For example, HiS-NET-seq data revealed more pronounced bidirectional transcription at enhancers (Figure 7.5B) than the other methods, despite its 5-10 times lower sequencing depth.

Together, the listed advantages present HiS-NET-seq as a quantitative, reproducible, and sensitive method that will serve as a promising complementary approach to measure Pol II occupancy in the future.

The main difference between the considered methods was the Pol II pausing signal intensity and position in the promoter-proximal region and near splice sites.

Notably, the *promoter-proximal maximum pausing position* varied for most methods between 80-106 nucleotides downstream of the TSS (Figure 7.5A), which was further downstream than the initially observed 30-60 nucleotides in fly cells [117] or the 20-60 nucleotides observed in K562 [228]. Interestingly, the latter performed single-molecule nascent RNA sequencing, allowing the estimation of the individual distance to the transcription initiation site for each molecule. In contrast, the analysis presented in this study used a constant reference point for each gene defined as the first annotated TSS above a threshold (Section 6.1.1). Therefore, this study applied the index for comparisons between the methods but likely overestimated the actual pausing position distances to the initiation sites.

The Pol II pausing signal around intron and exon boundaries was absent in PRO-seq data but observable for HiS-NET-seq, NET-seq, and mNET-seq (Figure 7.5B) as likewise described in the respective publications [148, 169]. mNET-seq initially reported these trends in datasets where the C-terminal domain of Pol II was serine five phosphorylated but not for the total population of Pol II. Further investigations revealed that the applied antibody was biased towards specific modifications, including Pol II with serine five phosphorylation [6]. Therefore it remained unclear whether pausing at splice sites was a technical artifact of the NET-seq approaches or a natural biological phenomenon. Further experiments and analyses are necessary to address this question.

HiS-NET-seq successfully addressed the main limitations of the original protocol and maintained the initially observed Pol II transcription features of NET-seq. Nevertheless, the labeling and selection of 4sU added additional steps to the NET-seq protocol, which increased the time investment and material costs. Additionally, it remains questionable if the HiS-NET-seq method captures arrested Pol II. Following up on this question could be of interest for future studies.

Part II

DETECTING GLOBAL POL II OCCUPANCY CHANGES WITH IMPROVED QUANTITATIVE METHODS

9 MOTIVATION

One central aim of functional genomics is to identify transcription factors and their regulatory impact on gene regulation. The regulatory function can be revealed when the system is challenged, for example, by an induced perturbation from a knockout, knockdown, mutation, treatment, or disease model. Combined with HTS methods, such as human NET-seq, the comparative analysis identifies the relevant biological differences between the measurements of a control experiment and the perturbation.

Most studies identify Pol II deregulation by comparing meta-gene profiles and Pol II describing parameters across experimental conditions [15, 105, 227, 249]. However, these methods are unsuitable for systematic comparisons, which require robust statistical methods to differentiate between technical and biological differences.

For extensively studied HTS assays, such as RNA-seq [132, 199] or ChIP-seq [221], robust and broadly tested methods emerged that are commonly applied in differential studies [210]. Applying these methods to other HTS assays requires careful adaptation considering method-specific characteristics and potential limitations. Essential for a successful and reliable differential analysis is to compare the data with the underlying model and test if assumptions used for normalizations and parameter estimations can be transferred or require adaptations. The following part introduces two NET-seq case studies to test comparative approaches, intending to identify a suitable strategy for NET-seq.

Recently, increasing attention was drawn to approaches that allow, in contrast to standard HTS methods, the detection of genome-wide uniform changes between conditions [26, 99, 134, 170]. The importance of detecting these uniform changes further emerged with newly developed protein degradation systems that rapidly degrade essential proteins of cells (Section 2.4.4), likely violating the normalization assumptions of most comparative analyses.

New RNA-seq protocols incorporate commercially available references that can be used for normalization to address this limitation [99]. However, the company explicitly developed the references for RNA-seq experiments and did not recommend their adaption for other HTS assays. The following part demonstrates the limitations of comparative analyses and investigates strategies that enable NET-seq to detect global changes between conditions.

10 CASE STUDIES

The NET-seq method was applied in two studies to measure Pol II occupancy changes between different conditions. The first study was a collaboration with the laboratory of Bruno Reversade, director of the *A*STARs Genome Institute of Singapore* and the *Institute of Molecular and Cell Biology*. The second study evolved from collaborating with Georg Winter, principal investigator at the *Research Center for Molecular Medicine of the Austrian Academy of Sciences*. This chapter introduces the datasets and the respective projects briefly.

The NET-seq measurements for both studies were variations of the original NET-seq method, named *spike-in NET-seq* (SI-NET-seq), used explicitly for quantitative comparisons between samples. Section 11.2.2 describes the new variant in detail. However, to highlight the difference between NET-seq and SI-NET-seq, this section keeps referring to the method as NET-seq if the spiked-in references were not considered for data normalization.

OSTEOGENESIS IMPERFECTA The clinical case study investigates five children from two distantly related families in Jordan with a congenital syndrome of *osteogenesis imperfecta* (OI), severe developmental delay, and neonatal progeria. OI affects one in 15,000-20,000 births, caused by a collagen protein type I mutation in 85-90% of all cases [66]. Other genetic causes involve proteins that interact with collagen and affect post-translational modification or folding [65]. Collagen proteins of type I are the most abundant proteins in bones, skin, and extracellular matrices [66]. The five severely-affected children suffered from a complex phenotype, with growth retardation, short stature, multiple bone deformities, and lipodystrophy. Lipodystrophy is a disorder in which the body cannot produce and maintain healthy fat tissue.

Furthermore, the patient displayed neonatal progeria with translucent and wrinkled skin. Other symptoms were acrogeria, premature depigmentation of sparse hair, and pediatric cataract. The genetic cause of the severe syndrome was unknown. The study performed polyA-enriched RNA-seq and SI-NET-seq (GSE197118 and GSE197119, unpublished) to gain insights into the functional consequences of the unknown genetic variant. For this study, the collaborating laboratory extracted primary fibroblast cells from two homozygous patients (V1, V5), one heterozygous healthy parent (IV2), and unrelated healthy individuals (polyA-enriched RNA-seq: WT1 and WT2; NET-seq: WT). The cells from each individual were processed according to the respective experimental protocol, with one and two replicate measurements for RNA-seq and SI-NET-seq, respectively.

PAN-BET PROTEIN DEGRADATION The second data set evolved from a previous study [249] which investigated the regulatory function of BET proteins in the MOLT4 cell line, a human T cell line from a 19-year-old man with *acute lymphoblastic leukemia*. The study developed the drug dBET6 (Section 2.4.4), an optimized chemical degrader that rapidly eliminates BET proteins in two hours or less, used for potential clinical applications. Furthermore, the degrader allows studying the regulatory role of BET proteins on Pol II genome transcription. Different HTS assays suggested that BET proteins act as master regulators of productive transcription elongation. Published data show global reductions in mRNA levels and serine-2 phosphorylated Pol II over gene-body regions. The result of the NET-seq assay was ambiguous as the increased pausing index suggested either increased pausing in promoter-proximal areas, the reduction of productive elongation at gene-body regions, or both. To investigate quantitative changes in these regions, Pol II changes between control (DMSO) and pan-BET protein degradation (dBET6) samples were compared after two hours of treatment using two replicate measurements of SI-NET-seq, respectively (GSE158963, [7]).

11 RESULTS

11.1 IDENTIFYING GENOME-WIDE POL II DEREGULATION

The statistical comparison of NET-seq data measured at different genomic regions needs to account for the characteristics of the count data and the low number of replicate measurements. NET-seq count data at actively transcribed genes (Section 6.1.1) showed classical overdispersion where variance exceeds the mean values (Figures 11.1A and A.3A) in the two example datasets introduced in the last section. Previous studies showed that the negative binomial distribution (Section 3.3.1), which allows the variance of a gene to depend on the mean and the dispersion parameter, performs well with HTS count data [73]. Established tools for the differential analysis of count data build negative binomial *generalized linear models* with a logarithmic link (Equation 3.11) to identify logarithmic fold changes between conditions. DEseq2 [132], the most popular tool, applies such a model to count data, in this case to NET-seq Pol II counts from gene regions. A considerable advantage of using this tool was the more robust estimation of scaling factors and gene-wise dispersion values to solve the *generalized linear models*. The combined application of DEseq2 and NET-seq data with the proposed adaptations discussed in this section is referenced as *differential Pol II occupancy* (DPO) analysis.

Normalization

Independent of the underlying biology, sequencing depths vary between NET-seq samples, making a direct comparison impossible. Different normalization strategies (Section 3.2) exist to remove sequencing depth bias. All normalization strategies assume that the absolute amount of total fragments in each cell was similar across the conditions [132, 157, 199, 237]. The simple RPM library size normalization was unsuited for NET-seq data (Figures 11.1B and A.3B) as few highly abundant genes biased the normalization factor considerably (Figures 11.1C and A.3C). DEseq2 computes scaling factors for each sample, referred to as *median-of-ratios* (Section 3.2.5), which were more robust towards outliers. This approach successfully removed bias introduced by different sequencing depths from the OI data set (Figure A.3D). However, the degrader treatment of the second data set violates the normalization assumption of DEseq2 as pan-BET protein degradation potentially influences the overall amount of nascent RNA produced in the cell. Hence, the normalization did not have the desired effect (Figure 11.1D) and potentially led to unreliable results, as discussed and shown later.

Dispersion Estimation

The *generalized linear model* assumes that the measured count value of a gene was sampled from a negative binomial distribution, modeled by the mean and variance parameters. However, reliable estimation of both parameters requires more than two replicate measurements for each condition. DEseq2 shares information across genes with similar mean expressions to estimate the dispersion parameter directly related to the variance. As described in Section 3.3.2, a curve representing the trend of the estimated dispersion values to the mean expression was fitted to the data. The shrinkage step fitted the estimated dispersion values to the trend curve. DEseq2 uses a parametric approach for the estimation, which results in a good fit from the observed dispersion values in the OI NET-seq study (Figure A.3E). The trend curve overestimates the dispersion for many genes in the pan-BET protein degradation experiment (Figure 11.1E). However, the more flexible local regression [30] visually improved the trend curve calculated for the pan-BET protein degradation experiment (Figure 11.1F). Besides the visual inspection, the *median absolute deviation* [200]

$$MAD = \text{median}\left[\left|X_i - \tilde{X}\right|\right] \tag{11.1}$$

quantifies the performance between the estimated dispersion values X_i and the trend curve $\tilde{X}$, which should be close to zero. The local regression method decreased the MAD by 38% for the pan-BET protein degradation experiment (Figure 11.1F). However, it had only marginal effects in the OI study (Figure A.3F). This result suggests that the parametric approach was unsuitable for some NET-seq data sets where local regression performed better, as shown for the pan-BET protein degradation data that favored the local regression methods to estimate the trend curve. Adjusting the method for dispersion estimation was a critical optimization step that positively influenced the sensitivity of the differential analysis as shown in Section 11.3.

log2FC and p-values

The remaining steps of the differential analysis, including estimating the *generalized linear model* coefficients, hypothesis testing using the *Wald* test, independent filtering, and multiple test correction, were performed as recommended and described in Section 3.3.2. Finally, this analysis considered genes with padj below 0.05 as significantly deregulated.

The number of genes with Pol II occupancy changes varied from 3.5% in the OI patients to 37.1% after pan-BET protein degradation (Table 11.1). For pan-BET protein degradation, the observed results were unexpected and questioned. Previous work [249] observed a global reduction of transcription levels in the same cell line and identified BET proteins as major transcriptional regulators.

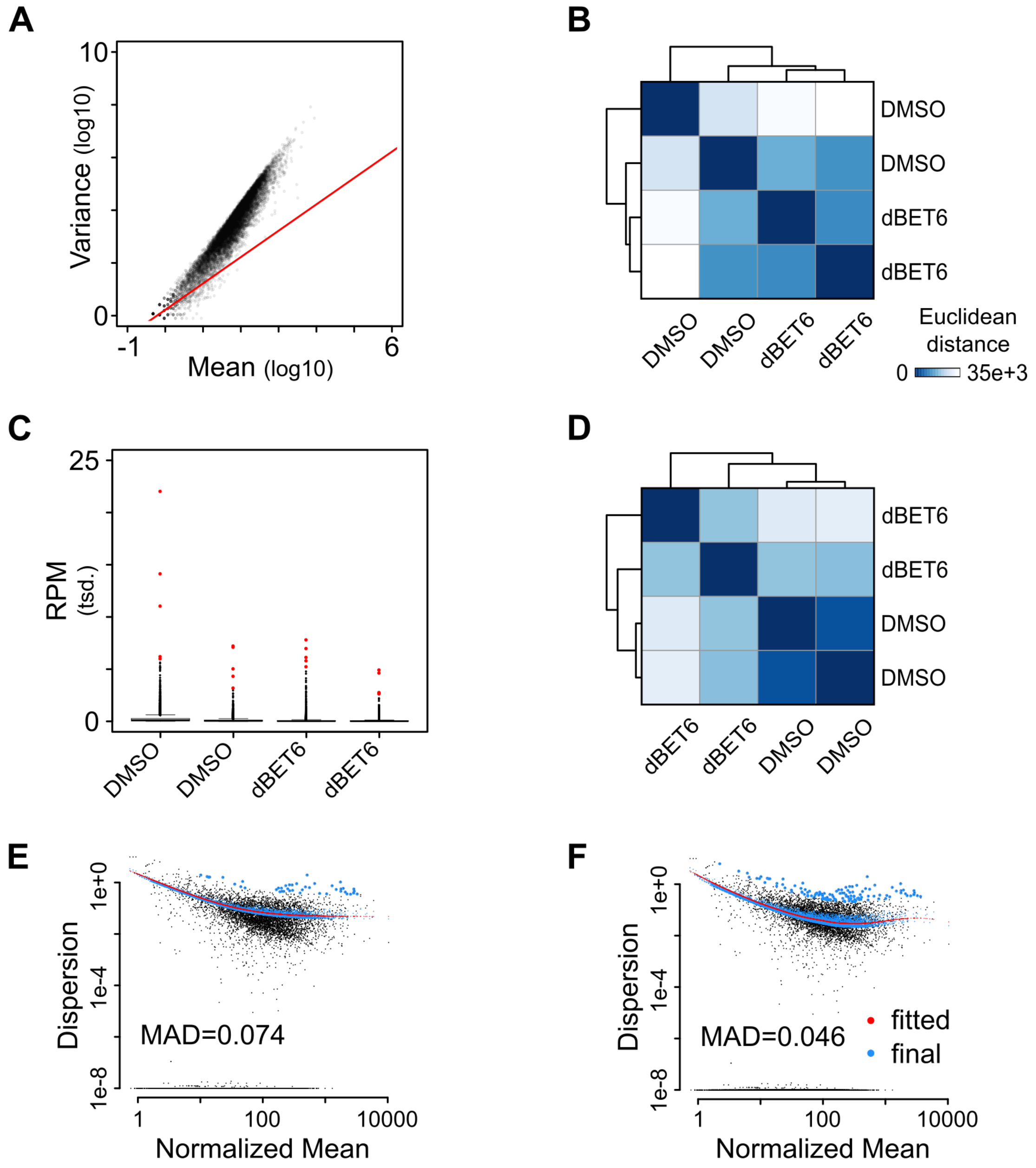


Figure 11.1: Parameter Estimation for Differential Analysis using Pan-BET Protein Degradation Data. Considered are non-overlapping genes (n=11,799, TPM > 1). The experimental design includes the control experiment (DMSO) and two hours of BET proteins degradation (dBET6) measured by NET-seq with two replicate measurements in MOLT4 cells. **(A)** Scatter-plot shows the log10 transformed dependency between variance and gene average. The diagonal is marked in red. **(B, D)** Hierarchical clustering of *Euclidean* distance measured between **(B)** RPM and **(D)** *median-of-ratios* normalized samples. **(C)** RPM normalized Pol II occupancy at genes. Marked are the top 5 expressed genes. **(E, F)** Scatter-plot shows the dependency of the estimated log-transformed dispersion and the normalized mean expression (black). Marked are the trend curve (red) and final dispersion parameters after shrinkage (blue), using **(E)** the parametric or **(F)** local regression method.

	Osteogenesis imperfecta (OI)	pan-BET protein degradation
Genes*	11,185	11,112
Up-regulated	147 (1.3%)	2,290 (20.6%)
Down-regulated	239 (2.1%)	1,833 (16.5%)
total	386 (3.5%)	4,123 (37.1%)

Table 11.1: Differential Analysis with Default Normalization. * Includes all genes that pass the independent filtering step of DEseq2.

The observed trend with NET-seq, where most genes showed no or raised Pol II occupancy changes, did not align with prior knowledge. A potential explanation was the applied normalization of DEseq2. The scaling factors were potentially not meaningful because the experiment violated the normalization assumption. As described in Section 3.2.6, DEseq2's normalization strategy assumes that each cell's absolute amount of total fragments was similar across the conditions. Therefore, DEseq2 generally provides no reliable default normalization in disease models or treatments that potentially influence global levels of genome transcription.

11.2 USING REFERENCE CELLS FOR NEW NET-SEQ VARIANTS

Are global changes detectable with genome-wide approaches? In practice, uniform changes can be detected in RNA-seq [134, 249] or ChIP-seq data [9, 26, 170], if experimental designs incorporate external controls, also known as spike-ins. For RNA-seq, synthetic ERCC RNA spike-ins [99] create standard baseline measurements with identical concentrations across samples. Due to a lack of reliability and complexity, these baseline measurements are not optimal for normalization [196]. However, only spike-in controls detect global changes if they are precisely incorporated and considered in the computational analysis [26, 134]. Unfortunately, the ERCC spike-ins were unsuited for NET-seq experiments, which perform 3'-end sequencing.

ChIP-Rx [170] solved this problem by incorporating an exogenous reference genome from fly cells into each sample for normalization. The ratios of untreated fly cells in each sample were identical, which allowed the reference-based normalization on fly observations exclusively. As humans and flies are distant evolutionary species, cross-mapping was low. However, processing cells from two distantly related species in parallel is challenging. Experimental conditions optimized for one species, here human cells, are often unsuitable for other species, here fly. For example, antibodies with ChIP quality rarely recognize proteins in human and fly cells, except for highly conserved proteins, such as histones.

Adapting the human NET-seq protocol to a generalized NET-seq protocol, suitable for human and fly cells, would be challenging due to significant experimental differences, such as cell culture conditions, media, temperature, cell fractionation, and other steps. In contrast, the NET-seq protocol can process cells from humans and *Mus musculus* (mouse) in parallel without protocol adaptations. This section discusses the advantages and disadvantages of the NET-seq adaptation, SI-NET-seq, which applies whole-cell spike-ins from the mouse for data normalization.

11.2.1 *Studying a Joint Reference Genome from Human and Mouse*

Is a differentiation between sequencing reads from mouse and human cells possible? The last common ancestor of humans and mice lived approximately 90 million years ago. Since then, independent genetic changes (mutations) have accumulated, resulting in about 60% nucleotide divergence [244]. However, protein-coding regions share higher similarities of 36-100%, with an average value of 85% [138]. A valid question that emerged from these numbers was if applications that map sequencing reads to genomic loci could separate the pooled transcripts from human and mouse cells after sequencing.

Most important for an independent reference normalization was low cross-contamination of sequenced reads from human cells, mapping to the mouse genome. To quantify these events, available NET-seq (OJ26) and HiS-NET-seq (OJ92, OJ93) datasets from human K562 cells were sequentially mapped to different references. First, to the human reference genome, and second, to a joint reference genome which consists of the human and mouse reference genomes (changed chromosome names, Figure 11.2A). On average, 0.6% of all sequenced human sequencing reads were incorrectly assigned to a unique mapping position in the mouse genome using the joint reference genome (Table B.8 and Figure 11.2A). This number decreased on average to 0.38% after the NET-seq pipeline performed standard filter processing steps. Data normalization accounts for observations within actively transcribed regions which excluded sequencing reads mapping to the extragenic regions. These steps resulted in 0.07% and 0.2% incorrectly assigned human sequencing reads to the mouse genome using NET-seq and HiS-NET-seq data, respectively (Table B.8).

Furthermore, ambiguous sequencing reads, defined as reads with more than one potential mapping position, likely increase due to the high similarity of orthologous sequence regions. Between 29-47% of sequencing reads mapped uniquely to the human reference genome contained at least one valid mapping position in the mouse genome (Figure 11.2B). The alignment tool, STAR (v2.7.3a) [50], correctly assigned most sequencing reads to the human genome using a joint reference genome with a negligible fraction wrongly assigned to the mouse genome (Figures 11.2A and 11.2C).

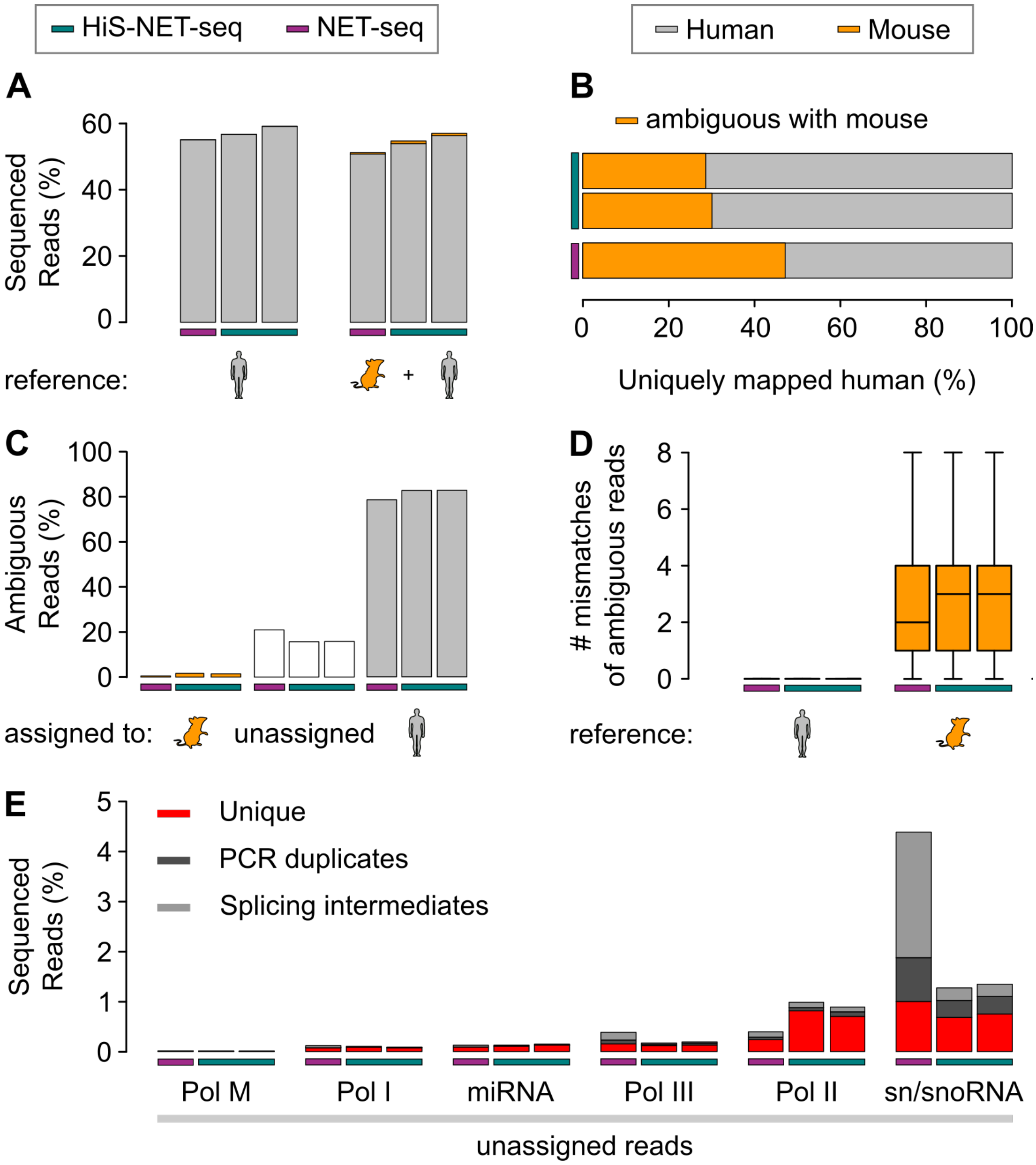


Figure 11.2: Cross-contamination of a Joint Reference Genome from Human and Mouse. The figure shows sample statistics for NET-seq (purple) and HiS-NET-seq (green) measured in the human cell line K562. **(A)** The fraction of uniquely mapped sequencing reads for the human reference genome and the combined human and mouse reference genomes. Colors indicate the fractions mapping to mouse (orange) or human (gray). **(B)** Set of sequencing reads with unique mapping positions in humans and at least one valid mapping position in the mouse genome. **(C)** The alignment tool assigned ambiguous sequencing reads to the mouse (orange) or human (gray) reference genome. Unassigned sequencing reads (white) have no unique mapping position due to the similarity of the two genomes. **(D)** Number (#) of mismatches in alignments of ambiguous sequencing reads mapping to human (gray) and mouse (orange). **(E)** Mapping position of unassigned ambiguous sequencing reads.

Although the alignment tool mapped many sequencing reads to both species, the average sequence alignment with the human reference genome contained 2.7 mismatches less compared to the mean mouse alignment (Figure 11.2D). This difference made a correct assignment feasible for most sequencing reads. Overall, only a fraction (17.5%) of potentially ambiguous sequencing reads did not map to a unique position, resulting in an average loss of 3.3% from all sequenced reads (Figure 11.2C). Most of these sequencing reads originated from sn/snoRNA transcripts (Figure 11.2E), an evolutionarily ancient group of non-coding RNAs with conserved functions [111]. Sn/snoRNA transcripts were highly abundant at the chromatin and excluded during processing steps of the HiS-/NET-seq pipeline. Subsequently, the joint reference genome decreased the number of informative sequencing reads marginally by 0.24% and 0.76% for NET-seq and HiS-NET-seq, respectively. The respective analyses suggest low cross-contamination from human to mouse or ambiguity effects introduced by the combined processing of sequenced reads from both species.

11.2.2 *SI-NET-seq*

For comparative analyses where uniform changes are possible or likely, a new NET-seq variant with mouse cell spike-ins was developed, referred to as SI-NET-seq (Figure 11.3A). This method added NIH 3T3 mouse cells in a specific ratio to the samples (6:1) before continuing with cell fractionation. Adding the exact proportion of cells was critical for normalization and required precision. The method assumes that Pol II occupancy at spiked-in mouse cells was identical across samples. Next, the new protocol variant processed the cell mixes from both species according to the initial NET-seq protocol [146], including cell fractionation, 3′ adapter ligation, production of cDNA, PCR amplification, and sequencing. Hence, data normalization using observations from spike-in controls does not only correct for different sequencing depths and global RNA composition changes but potentially for more complicated technical variations that may occur during the library preparation.

Processing Data with Spike-In Controls

Section 6.2 describes the data processing pipeline of NET-seq that was likewise used for SI-NET-seq with a few adjustments that consider the mouse genome. The mapping step uses a joint reference from the human (GRCh38.p12) and mouse (GRCm38.p6) [67] genomes. Furthermore, the adaptation removes splicing intermediates from both species using the GENCODE v28 and M18 [67] annotations. Next, contaminating RNA species from Table B.6 and blacklisted regions from ENCODE [35] were masked in the human and mouse genomes. Eventually, the human and mouse observations are separated, resulting in a human and mouse Pol II occupancy data set for each sample.

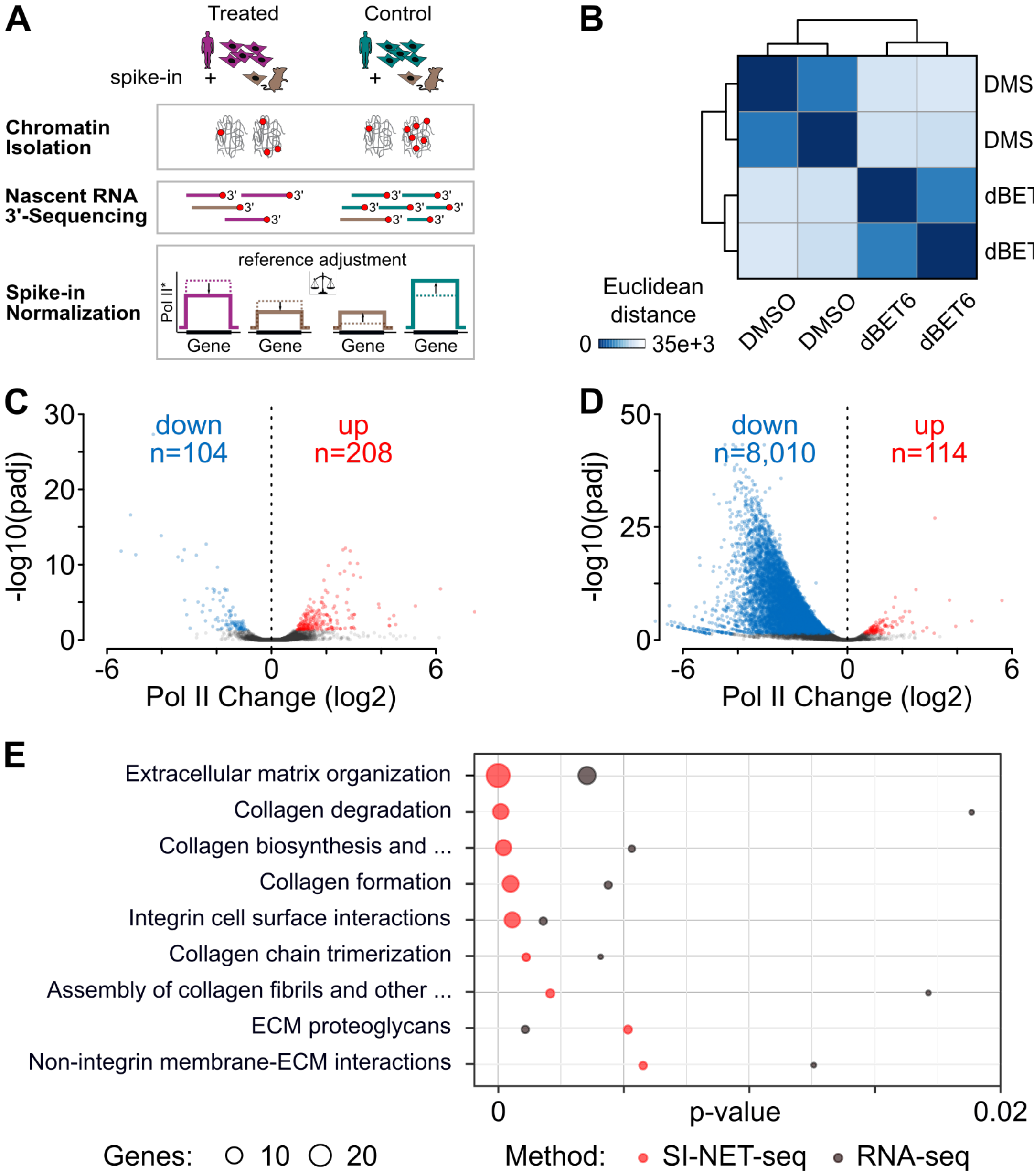


Figure 11.3: Results of DPO Analysis using SI-NET-seq Data. (A) Schematic view of SI-NET-seq. *: The y axis shows the Pol II occupancy before (dashed) and after (solid) normalization. **(B)** Hierarchical clustering of *Euclidean* distance measured between non-overlapping spike-in normalized genes (n=11,799) for the control experiment (DMSO) and after two hours of pan-BET protein degradation (dBET6) in MOLT4 cells. The scaling factors were calculated from mouse gene observations (n=12,997). **(C, D)** Pol II occupancy changes (log2) identified with SI-NET-seq in **(C)** OI data from primary fibroblast cells of patients (n=9,807) and **(D)** after two hours of pan-BET protein degradation in MOLT4 cells (n=11,095). Significant occupancy changes (padj < 0.05) are depicted in blue and red. **(E)** Integrated analysis of significant (p-value < 0.05) REACTOME pathways [76] identified by deregulated genes of OI patients from polyA-enriched RNA-seq and SI-NET-seq data sets.

The Pol II occupancy in untreated mouse cells should be similar across all samples and was used for data normalization.

ADAPTATIONS TO HIS-NET-SEQ The HiS-NET-seq protocol implemented the same adaptations to allow comparative studies between conditions. Due to handling difficulties with whole-cell spike-ins, the new protocol variation pools the samples after cell fractionation with labeled nascent RNA from NIH 3T3 mouse cells (8:1). For this reason, HiS-NET-seq with spiked-in mouse material can not correct for different efficiencies in cell fractionation.

The following part of this study analyzed HiS-NET-seq data with spiked-in mouse material. If not stated otherwise, this protocol variant, incorporating mouse material for normalization, is the default.

11.3 DETECTING GLOBAL CHANGES WITH REFERENCE-BASED NORMALIZATION

With one substantial modification, both SI-NET-seq data sets were re-analyzed with the DPO analysis, as described in Section 11.1. The DPO analysis calculated the scaling factors based on gene count measurements from mouse observations for each sample as described in Section 3.2.6. This approach led to a correct clustering of normalized count data for both data sets, which was not observed with other normalization strategies considering human genes in the pan-BET protein degradation experiment (Figures 11.3B, 11.1B, and 11.1D). Active mouse genes had roughly the same complexity as human genes, including many actively transcribed genes with variations in gene length and nucleotide composition. This complexity was a considerable advantage for normalization over synthetic ERCC spike-ins that lack complexity and variation used in RNA-seq experiments.

With an padj of 0.05 or less, deregulated genes slightly decreased to 3.2% for OI data (n=312, Figures 11.3C and A.4A) but increased to 73.2% after pan-BET protein degradation (n=8,124, Figures 11.3D and A.4B). Interestingly, the local regression strategy for dispersion estimation improved the number of detected genes by 5.4%. The quality check showed only a few genes in the mouse genome with significant changes, which revealed an average specificity of 99.8% (Figures A.4C and A.4D). Therefore, the reference-based normalization strategy improved the test considerably for the pan-BET protein degradation experiment.

The number of significant genes identified with traditional normalization and spike-in controls did not change remarkably for OI data. This observation implies that the unknown disease-causing mutation did not affect global Pol II distribution but a specific set of genes.

The high fraction of genes identified with both normalization methods (54%) and high correlation of estimated fold changes (r=0.998) suggested that spike-in normalization can be applied in different situations, including the standard application without an underlying uniform change. In this case, the spike-in normalization was more conservative than the standard normalization.

The next step compared the SI-NET-seq results with differentially expressed genes measured by polyA-enriched RNA-seq data. Surprisingly, the two methods identified only a small number of twenty-one genes in both differential analyses. Although different deregulated gene sets were detected, both methods detected similar biological pathways affected by the syndrome. Deregulated genes affected pathways of the extracellular matrix and collagen synthesis significantly (Figure 11.3E). This result explained the similarities of the patient's phenotype with the described OI syndrome caused by a direct collagen mutation.

For pan-BET protein degradation, the identified global reduction of Pol II gene transcription (Figures 11.3D and A.4B) was consistent with the results from the previous study [249] and confirmed the essential regulatory role of BET proteins on Pol II transcription. Although pan-BET protein degradation showed a general genome-wide reduction of Pol II, different features such as gene length or gene classes were associated with a more pronounced response (Figures A.4E and A.4F).

Overall, the proposed method provides a practical approach for comparing high-resolution Pol II data across conditions concerning individual gene regions or changes in other regions of interest.

12 DISCUSSION

Several comparative genome-wide transcription studies investigating protein functions, such as BET proteins [45, 249], failed in the past because they did not integrate experimental procedures that allow reliable and quantitative data analyses.

In order to avoid the same problems, this part implemented computational and experimental steps to perform comparisons between Pol II occupancy data. The DPO analysis enables the robust and quantitative analyses of NET-seq data and reports significant changes at individual genomic regions genome-wide. Two new protocol variants, SI-NET-seq and HiS-NET-seq, used spike-ins from mouse cells, which proved essential for comparative studies that violated traditional normalization assumptions. The spike-in controls and the adjusted testing strategy were successfully applied in two studies and revealed new insights into a disease mechanism and Pol II elongation control.

The application of human NET-seq in Winter et al., 2017 showed that previous optimization efforts to gain quantitative data, such as incorporating UMI sequences [148], remained ineffective when a biological condition caused global amplification or depletion of Pol II transcription. The underlying assumption of most commonly applied normalization strategies [132, 157, 199, 237] requires that most observations do not change between conditions. In practice, missing potential uniform changes is a considerable limitation of most transcription studies [26, 170] that requires adjusting the experimental design and the computational analysis of the respective methods.

Previous publications solved this problem by using cross-species references for normalization but avoided cells from closely related species [170, 262]. This study shows that despite a high genetic similarity between the human and mouse genome, cross-mapping from one to the other species was neglectable (Section 11.2.1). Baluapuri et al., 2017 [9] came to a similar conclusion and used mouse cells for cross-species normalization in an adjusted ChIP-Rx protocol.

Unfortunately, this study could only assess cross-contamination from human to mouse but not vice versa due to missing NET-seq data originating from mouse cells. However, bias potentially introduced from mouse transcripts did not change between conditions and had no impact on the results or conclusions from comparative analyses of this work. As a result, the spike-in approach identified global changes between conditions, which was essential to avoid misinterpretation in one of the two case studies.

The similar behaviors of mouse NIH 3T3 and human K562 cells during lysis allowed parallel processing without intense optimizations of the NET-seq protocol. Overall, this study highlights the benefits of this approach, justifying the increased computational and experimental complexity introduced by the spiked-in mouse cells.

Next, changes in Pol II occupancy were identified at individual gene regions using the DEseq2-based DPO analysis. Previous studies performed similar tests with other Pol II profiling methods but not NET-seq [18, 262]. Adaptation of dispersion and normalization estimation were essential to achieve meaningful results and improve test sensitivity.

Why was the region-wise comparison between conditions beneficial? Most studies with high-resolution Pol II occupancy data and different treatment conditions report their results with meta-gene profiles of samples at the regions of interest [15, 105, 227]. Although this approach has several advantages, meta-gene plots fail to identify specific deregulated genes. This lack of resolution challenges the data interpretation if no global changes occur, as shown in Figure A.4A vs. Figure A.4B. In contrast, the DPO analysis considered each gene separately and allowed downstream enrichment- or correlation analyses. Many developed methods and databases that interpret these types of results and identify enrichments prove the effectiveness of this general approach (Section 3.4) [72, 76, 153].

Furthermore, meta-gene profiles can be disproportionally biased by outliers. The DPO analysis avoided this by performing robust variable estimations implemented by DEseq2 (Section 3.3.2).

In the clinical case study, DPO identified deregulated genes and the affected pathways (Figures 11.3C and 11.3E). The study investigated the molecular genetic cause of a rare *osteogenesis imperfecta* syndrome. Deregulated Pol II genes of OI patients affected extracellular matrix and collagen-related pathways, which led to a comparable phenotype of patients with a mutation in one of the collagen genes.

For pan-BET protein degradation, the reference-based normalization strategy identified a gene-class and gene-length specific productive elongation collapse (Figures A.4E and A.4F).

As shown in the two case study examples, DPO analysis successfully used spiked-in mouse cells for the comparative analysis of high-resolution Pol II data. This approach allowed downstream investigation of affected gene sets which improved the interpretability of the results.

Part III

BRD4 EMERGES AS GLOBAL REGULATOR OF POL II TRANSCRIPTION

13 MOTIVATION

The Pol II transcription cycle is a highly regulated multi-step process involving many regulators to produce functional RNA transcripts translated into proteins in living cells. Although this process is essential, several regulatory steps are still insufficiently understood, especially steps after Pol II initiation.

One example is the Pol II early elongation stage, which emerged as an essential regulatory step in multicellular organisms. After a successful transcription initiation event, Pol II pauses proximal to the promoter and requires the active P-TEFb complex to pursue transcription (Section 2.2.2). BET proteins positively influence Pol II elongation [3, 13, 133], where BRD4, the most prominent protein family member, was proposed to recruit P-TEFb [173]. However, recent studies showed that P-TEFb recruitment is independent of BET proteins [158, 249, 260], leaving the open question of the underlying mechanism of BET-dependent Pol II elongation.

Furthermore, it is unclear which BET protein regulates productive Pol II elongation as current treatments are not selective. Recent developments of new degradation technologies allow the selective degradation of target proteins in a few hours (Section 2.4.4), including BRD4 in the human dTAG-BRD4 K562 cell line. Although most studies ascribed the regulatory functions of BET proteins to BRD4, this claim was rarely supported by BRD4-specific perturbation experiments when this project started. The specific degradation of BRD4 can validate this longstanding claim. In the meantime, other studies investigated BRD4-specific functions likewise [7, 158, 260].

Because BET proteins are proposed master regulators of Pol II transcription, loss of BRD4 could potentially influence transcription levels globally. Applying suitable experimental and computational methods is essential to avoid misinterpretation and reveal actual protein functions. Part II of this study developed new NET-seq versions allowing the detection of global changes between conditions.

Of particular interest are potential changes in less studied lowly transcribed regions, for example, during Pol II termination or enhancer regions. Part I of this study implements a new high-resolution NET-seq method with genome-wide Pol II occupancy coverage improvement. The optimized methods potentially reveal additional insights into these lowly transcribed and less studied transcriptional processes upon BRD4 loss.

Emerging evidence links BRD4 to transcriptional condensates and phase separation (Section 2.3), which opens the question of BRD4's role in establishing functional promoter-enhancer contacts. Several studies showed BET protein-dependent reductions of transcriptional activity at enhancers [107, 164]. Nevertheless, it remains unclear which BET proteins are involved, including their underlying function. Surprisingly, a recent study did not detect any changes in enhancer transcription upon BRD4 degradation [260]. However, more studies are required to validate these findings and to characterize the function of BRD4 further. Finally, the BRD4-specific degradation system could reveal new insights into the role of BRD4 in the formation of promoter-enhancer 3D contacts.

The following sections investigate the phenotype observed upon BRD4 loss using multi-omics approaches. Furthermore, the corresponding analyses required the development of novel computational approaches that quantified the impact of the treatment on Pol II genome transcription. Detailed explanations of these approaches are available in the following methods section.

14 METHODS

The following section describes data processing and analysis of HTS methods, including RNA-seq, ChIP-Rx, *long-read nascent RNA-sequencing* (nascONT-seq), and Hi-ChIP. Next, different approaches are proposed to study RNA splicing, 3′-RNA processing, and Pol II termination, using the transcriptional readthrough index, the RNA splicing analysis, and a *3′-RNA cleavage efficiency* test.

14.1 ANALYZING HTS DATA

The FastQC software [4] tested all sequenced HTS data sets to ensure high sequencing quality before applying method-specific processing steps.

14.1.1 *RNA-seq*

RNA-seq data processing steps followed general literature recommendations [33]. Sequencing reads were aligned to the GRCh38.p12 human reference genome using STAR aligner v2.7.3a [50] with default parameters in paired-end or single-end mode. HTSeq v0.13.5 [184] quantified annotated genes from GENCODE v28 [67] in *union* mode. Sections 3.2.5 and 3.3 describe the steps of normalization and differential gene expression analysis.

A combined reference genome that contained additional sequences from synthetic ERCC spike-ins [99] was used for some experiments to reveal global changes between conditions. The adjusted normalization strategy is described in Section 3.2.6.

14.1.2 *ChIP-seq and ChIP-Rx*

The following section describes steps for data processing and differential binding site analysis using ChIP-seq and ChIP-Rx data. An overview of the ChIP-seq approach and the applied normalization strategies are reported in Sections 2.4.1 and 3.2.7.

Data Processing

CHIP-SEQ FROM ENCODE AND GEO ChIP-seq extracted from the ENCODE project [35] was already pre-processed and normalized.

For individual replicates and pooled data, the alignment and FE-normalized files (Section 3.2.7) were downloaded and used in downstream analysis.

For the re-analysis of publicly available ChIP-seq data [253] from GEO, Bowtie2 v2.3.5.1 [119] aligned the sequenced reads to the human reference genome (GRCh38.p12) and reported one sequence alignment for each sequencing read using the *single-end* mode and the parameter *-k 1*. Section 3.2.7 describes the steps for FE data normalization.

CHIP-RX For the ChIP-Rx analysis, the following steps were performed.

1. Bowtie2 v2.3.5.1 [119] aligned the sequencing reads to a joint reference genome which consisted of the human (GRCh38.p12) and mouse (GRCm38.p6) reference genomes using the *paired-end* mode with the parameter *-k 1*.

2. PCR duplicates were marked with PICARD's v2.24.2 [95] *markDuplicates* function.

3. Next, MACS2 v2.2.7.1 [258] identified binding sites on the separated human and mouse tracks with an enriched sample signal compared to the matched input control. This step, called peak calling, was performed separately for human and mouse tracks.

Section 3.2.7 describes how the data was FE normalized.

Differential Binding Site Analysis

DiffBind v3.0.15 [221] performed data normalization and comparisons between binding sites from ChIP-Rx data between conditions. The approach aims to identify differentially-bound binding sites. Table B.9 lists the applied parameter settings and the corresponding functions.

First, DiffBind's *dba.blacklist* function removed the signal from blacklisted regions of the human and mouse reference genomes. Second, peaks from all samples were summarized in consensus peaks if they appeared in at least two samples. The *dba.count* function identified the consensus peaks by detecting the highest sequencing read coverage region (+/- 300 nucleotides). Furthermore, the function quantified the binding strength of each consensus peak in all samples, excluding sequencing read counts from PCR duplicates and the input control experiment. Third, DiffBind's *dba.normalize* function adjusted the human consensus peaks based on the observed mouse data using the *median-of-ratios* normalization strategy (Section 3.2.5). Notably, normalization uses the whole binned mouse genome. Finally, the *dba.contrast* and *dba.analyze* functions performed the differential binding site analysis using the DEseq2 package [132] (Section 3.3.2).

14.1.3 *nascONT-seq*

In the context of this study, the nascONT-seq method was developed. The method quantitatively captures full-length nascent RNA molecules and performs nanopore sequencing. Section 15.2.3 explains the motivation for developing this method, including a description of the library preparation. This part focuses on the computational steps applied to process long-reads from ONT (Section 2.4.2). All software tools and applied parameters are listed in Table B.10.

Figure 14.1A schematically shows ONT sequencing, which relies on the controlled passage of one strand of the cDNA molecule through the nanopore, causing electric current changes recorded in fast5 format files. ONT's corporate software Guppy v3.2.4 interpreted these raw files into nucleotide sequences using the kit and flow cell-specific parameters. After this base-calling step, standard processing includes a filtering step to remove reads harboring incomplete sequences of the nascent RNA molecules. The following section describes the standard approach to identify if a sequencing read is complete, the challenge that emerged with the presented data, and an alternative method applied in this study.

Identification of Full-Length Transcripts

Library preparation introduces the strand-switching primer (SSP) and VN primer (VNP) at the cDNA molecule's 5′ and 3′ ends, which ONT sequences altogether. As depicted in Figure 14.1B, a sequencing read is considered complete or full-length only if the primers flank the fragment in the proper orientation. The only publicly available software application that identifies the full-length transcripts is ONT's Pychopper software [218]. Pychopper detects and removes the primers, reporting only fragments considered full-length molecules. However, this standard approach was not suitable for this study's obtained long-read sequencing data.

The main reason was the usage of the *direct cDNA sequencing kit*. This kit's library preparation steps do not perform PCR amplification, allowing quantitative comparisons between samples. However, PCR amplification is a crucial step for enriching correctly assembled fragments with both primers. Data received without this step contain considerably fewer full-length transcripts.

The developers do not recommend applying the Pychopper software with *direct cDNA sequencing kit* data. Consistent with Pychopper's recommendation, the tool performed poorly and discarded most sequencing reads.

For this reason, this study developed a custom approach to identify full-length transcripts for the *direct cDNA sequencing kit*. Closer inspection of the libraries revealed different read types. Besides full-length transcripts (Figure 14.1B), mainly sequencing reads with only one primer (5′ and 3′ truncated), no primer or fused reads occurred (Figure 14.1C).

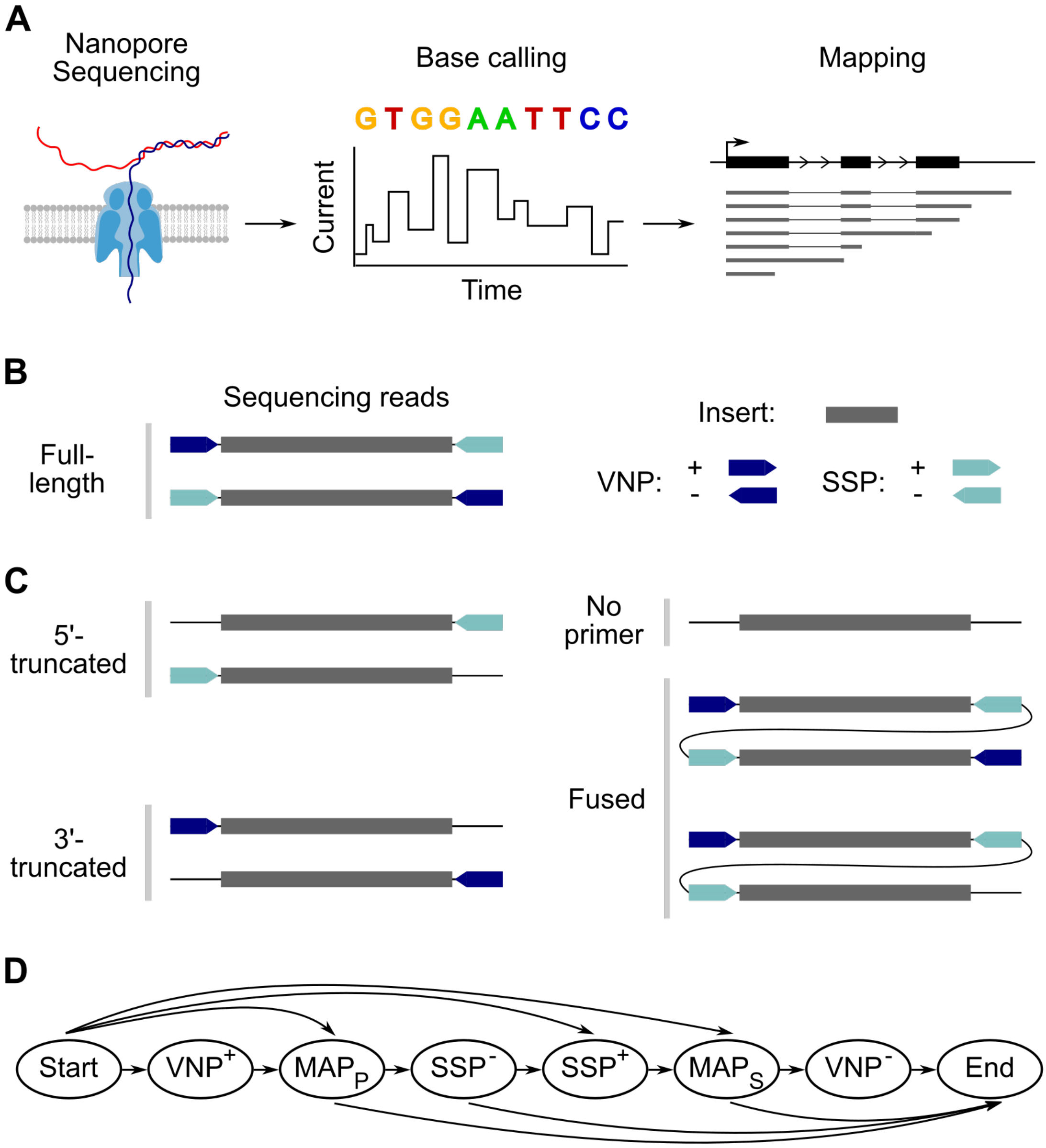


Figure 14.1: Read Types from ONT Data. (A) The schematic shows long-read sequencing with nanopores, creating electric current changes that are monitored and translated into nucleotide sequences (base calling). The mapping step aligns the sequence to the reference genome. **(B-C)** Schematic overview of sequencing reads that contain SSP and VNP primer sequences. The insert represents the nascent RNA molecule and maps to the reference genome. **(B)** Only sequencing reads containing both flanking primers with correct orientation are considered full-length transcripts. **(C)** The remaining read types, including 5′-truncated, 3′-truncated, no primer, or fused read, are defined as depicted. **(D)** Visualization of all valid state combinations from Table B.11 encoded in the matrix T.

Fused sequencing reads, also known as chimeric reads, are hybrids that contain sequence information from two nascent RNA molecules sequenced sequentially. The presented approach is less conservative and classifies each sequencing read into one of the read types, rescuing fused or incomplete reads for downstream analysis, such as the *3′-RNA cleavage efficiency* test (Section 14.4).

The following three steps were applied to classify each sequencing read into one read type.

1. Determining the primer positions and orientations.
2. Identifying mapping regions of the sequencing read.
3. Classifying each read into a sequencing read type.

DETERMINING PRIMER POSITION AND ORIENTATION Two Hidden Markov Models [114] were built from the primer sequences to identify the SSP and VNP primers. The data-driven approach improved the models by including hits from the long-read sequencing data. HMMER v3.3 [176] identified high confidence primer sequences with an E-value < 0.1 and added those to the Hidden Markov Models in two iterations using the HMMER functions (Table B.10). The optimized models were applied for the final primer search, reporting all hits with an E-value < 10.

IDENTIFYING MAPPING REGIONS OF THE READ Minimap2 v2.17 [124] aligned the sequencing reads to the human reference genome (GRCh38.p12) using the parameters listed in Table B.10. Integration of prior knowledge from the human gene annotation (GENCODE v28) improved the mapping. Minimap2 is a splice-aware long-read mapping tool that handles higher error rates in the sequencing reads and reports unique sequencing alignments with the applied parameter settings. Furthermore, to detect fused reads (Figure 14.1C), a read could be broken into pieces and mapped to different loci, referred to as a supplementary mapping position.

CLASSIFYING EACH READ INTO A READ TYPE Each sequencing read with at least one valid mapping position in the human reference genome was classified into a read type, including full-length, 5′-truncated, 3′-truncated, no primer, or fused read. The classification depended on the identified primers relative to the mapping position visualized in Figures 14.1B and 14.1C. The following states describe the occurrence of the respective feature in the sequencing read, namely

- beginning or end of the read ($Start/End$),
- sense or antisense VNP primer (VNP^{+}/VNP^{-}),
- sense or antisense SSP primer (SSP^{+}/SSP^{-}),

- mapping position (MAP_P), or
- a supplementary mapping position (MAP_S).

A sequencing read contains nS feature states sequentially ordered by their occurrence from the 5′ to 3′ end. For the classification of the sequencing read, the subset of feature states was considered that

1. preserved the sequential order of feature states in the sequencing read,
2. occurred in Table B.11, and
3. maximized the scoring function (described below).

For a selected subset with nS' elements that preserved the sequential order of feature states, $fs_i \in \{Start, VNP^+, VNP^-, SSP^+, SSP^-, MAP_P, MAP_S, End\}$ reports the i-th occurring feature state with $i \in \{1, \ldots, nS'\}$, where $fs_1 = Start$, $fs_{nS'} = End$, and $3 \leqslant nS' \leqslant nS$. Each feature state of fs_i is associated with aS_i, assigning the calculated alignment score from HMMER or minimap2 to the respective state. The score ranges between $0 \leqslant aS_i \leqslant 100$ and reports zero for the $Start$ and End states, $aS_1, aS_{nS'} = 0$. Furthermore, the $d_{fs_i, fs_{i+1}}$ value is defined for the states fs_i and fs_{i+1}, reporting the distance between both states in percent of the sequencing read length.

Figure 14.1D visualizes the valid state combinations listed in Table B.11 that are encoded in the matrix

$$T = \begin{array}{c} \\ Start \\ VNP^+ \\ VNP^- \\ SSP^+ \\ SSP^- \\ MAP_P \\ MAP_S \\ End \end{array} \begin{array}{c} \begin{array}{cccccccc} Start & VNP^+ & VNP^- & SSP^+ & SSP^- & MAP_P & MAP_S & End \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{14.1}$$

The value $t_{fs_i, fs_{i+1}} \in \{0, 1\}$ reports the entry of T in row fs_i and column fs_{i+1}. All valid state transitions depicted in Figure 14.1D are set to 1, whereas invalid transitions report 0 entries.

For a subset of state features, the scoring function was defined as

$$score = \left(\sum_{i=1}^{nS'} weight\left(fs_i\right) \cdot aS_i - \sum_{i=1}^{nS'-1} 0.1 \cdot d_{fs_i, fs_{i+1}} \right) \cdot \prod_{i=1}^{nS'-1} t_{fs_i, fs_{i+1}}, \quad (14.2)$$

using the weight function

$$weight\left(fs_i\right) = \begin{cases} 0.5 & , if\ fs_i \in \{MAP_P, MAP_S\} \\ 0.4 & , otherwise \end{cases}, \quad (14.3)$$

the alignment scores aS_i, the distance measurements $d_{fs_i, fs_{i+1}}$, and state transition values $t_{fs_i, fs_{i+1}}$. Finally, the subset with the maximum score specifies the read type listed in the Table B.11. Figure 14.2 shows the approach with an example sequencing read for further clarification.

Most analyses used all sequencing read types, except for the transcript length analysis depicted in Figure 15.4D, which was calculated for full-length transcripts and rescued fused reads.

14.1.4 *HiChIP*

The following section describes data processing steps and the approach performed for differential analyses of contact frequencies using HiChIP data. Section 2.4.1 contains a general description of the HiChIP method, and Table B.12 provides an overview of the applied software and the selected parameters.

Data Processing

HiChIP data processing was based on the HiC-Pro v3.0.0 [212] pipeline and implementation. Briefly, HiC-Pro performed a two-step approach to map the sequencing reads to the human reference genome (GRCh38.p12) [67]. First, paired-end sequencing reads were separately aligned using Bowtie2 v2.3.5.1 [119]. The fraction of reads without a unique mapping position likely spanned the sequenced hybrid molecule created during library preparation described in Section 2.4.1. Next, the tool divided the sequencing reads without unique mapping positions at the putative ligation site, marked by the cut site of the applied restriction enzyme (MboI: ^GATC). HiC-Pro performed a second mapping iteration with slightly changed parameter settings to identify unique mapping positions for the remaining reads. The quality assessment step discarded interaction pairs without unique mapping positions or low mapping quality. As a valid interaction required the ligation of two restriction fragments, this step discarded interaction pairs assigned to a single restriction fragment. Finally, HiC-Pro removed PCR duplicates and generated a binned interaction matrix with 10 kb resolution. The matrix reports the pairwise interaction frequencies between all bins in the genome.

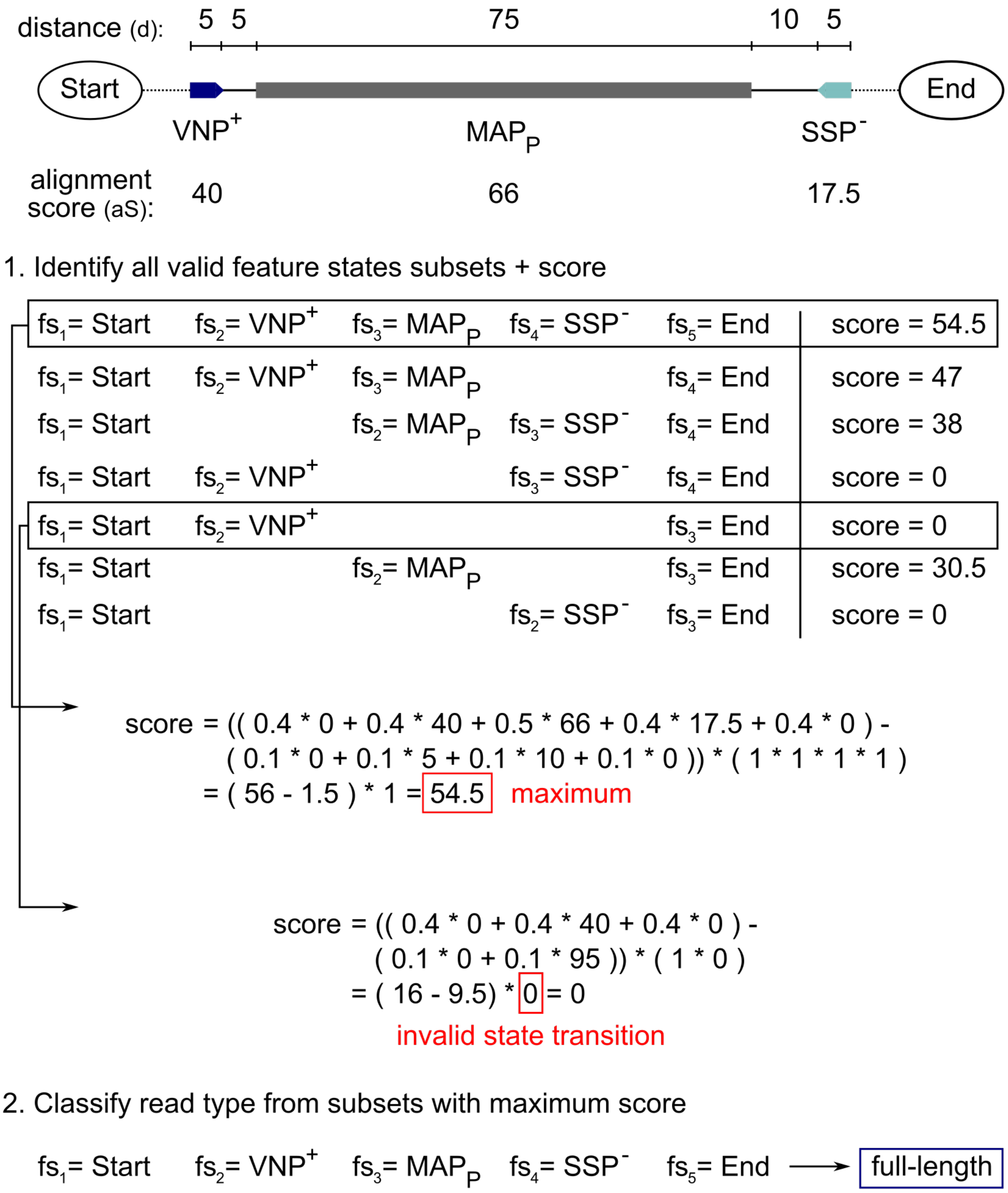

distance (d):
5 5
75
10 5
Start
End
VNP^+
MAP_P
SSP^-
alignment score (aS):
40
66
17.5
1. Identify all valid feature states subsets + score
fs_1= Start fs_2= VNP^+ fs_3= MAP_P fs_4= SSP^- fs_5= End score = 54.5
fs_1= Start fs_2= VNP^+ fs_3= MAP_P fs_4= End score = 47
fs_1= Start fs_2= MAP_P fs_3= SSP^- fs_4= End score = 38
fs_1= Start fs_2= VNP^+ fs_3= SSP^- fs_4= End score = 0
fs_1= Start fs_2= VNP^+ fs_3= End score = 0
fs_1= Start fs_2= MAP_P fs_3= End score = 30.5
fs_1= Start fs_2= SSP^- fs_3= End score = 0
score = ((0.4 * 0 + 0.4 * 40 + 0.5 * 66 + 0.4 * 17.5 + 0.4 * 0) -
(0.1 * 0 + 0.1 * 5 + 0.1 * 10 + 0.1 * 0)) * (1 * 1 * 1 * 1)
= (56 - 1.5) * 1 = 54.5
maximum
score = ((0.4 * 0 + 0.4 * 40 + 0.4 * 0) -
(0.1 * 0 + 0.1 * 95)) * (1 * 0)
= (16 - 9.5) * 0 = 0
invalid state transition
2. Classify read type from subsets with maximum score
fs_1= Start fs_2= VNP^+ fs_3= MAP_P fs_4= SSP^- fs_5= End
full-length

Figure 14.2: Read Type Classification Example. The schematic depicts the classification process of an example sequencing read with the identified feature states *Start*, VNP^+, MAP_P, SSP^-, and *End* ($nS = 5$). Furthermore, the example depicts the alignment scores (aS) and distances (d) in percent. 1. All valid subsets are identified, preserving the sequential occurrence of the feature states in the sequencing read, containing at least $nS' = 3$ feature states, with $fs_1 = Start$ and $fs_{nS'} = End$. Next, Equation 14.2 calculates the *score* for each subset. The visualization depicts the score calculation for two subsets in greater detail. 2. The subset with the highest score is used for classification.

Normalization and Comparison

HiCcompare [220] performed data normalization and comparison between two pooled HiChIP samples. First, QDNAseq's [206] build-in *get_CNV* function identified copy number variations in the K562 dTAG-BRD4 cell line and excluded the respective regions. Second, blacklisted regions from ENCODE [35] were excluded. Third, HiCcompare's *hic_loess* function corrected all interactions using *locally estimated scatterplot smoothing* (loess) normalization [30] on the MD-plot for each chromosome. The MD plot showed the logarithmic changes of the interaction frequencies between two samples on the y-axis and the distances between the interactions on the x-axis. The loess procedure fits a local regression model and jointly removes biases between both datasets.

Finally, the *hic_compare* function converted approximately normally distributed logarithmic changes of the interaction frequencies [206] into z-scores and p-values, followed by multiple test correction (Section 3.3.2). Furthermore, the tool removed interactions with an average expression below nine. Interactions with a minimum distance of 10 kb and a padj value of 0.05 were identified as significant changes between conditions.

Interpretation

HiCcompare reported significant interaction frequency changes at a resolution of 10 kb. The generation of a human genome annotation at the same resolution was required to enable interpretation of the results. Therefore, chromHMM [56] was applied to K562-specific chromatin marks from ENCODE [35], including

- H3K27ac,
- H3K4me1,
- H3K4me3,
- *histone three lysine twenty-seven trimethylation* (H3K27me3),
- *histone three lysine thirty-six trimethylation* (H3K36me3), and
- *histone three lysine seventy-nine dimethylation* (H3K79me2).

The corresponding ENCODE identifiers are listed in the Table B.1.

In the first step, the chromHMM [56] application divided the human genome into 10 kb bins using the *BinarizeBam* function. In the second step, chromHMM's *LearnModel* function trained a multivariate Hidden Markov Model with ten states and reported the most likely states for each bin. In the last step, the resulting states were manually annotated (Figure A.23B) into *promoter*, *enhancer* (*repressed*, *intra-*, and *extragenic*), *3′ end*, *repressed*, and *low-signal* states.

14.2 TRANSCRIPTIONAL READTHROUGH INDEX

Experimental conditions, such as treatments, knockouts, infections, or mutations, potentially influence Pol II termination efficiency, resulting in Pol II readthrough transcription. Pol II readthrough transcription describes a situation where Pol II occurs downstream of the observed termination zone from the control experiment. In mammals, the termination zone occurs in a region downstream of the polyA site and is highly variable for each transcription unit [211]. However, previous studies [8, 14, 18, 109] introducing termination-related indices only considered changes in fixed regions relative to annotated polyA sites. Therefore, this study developed the *transcriptional readthrough index* to quantify Pol II occupancy changes in the individual termination zones calculated for each transcription unit (Figure 15.3B).

The index is defined based on an exemplary transcription unit, with the genomic coordinates of the transcription start site tss and the polyA site pa. For simplicity, the transcription unit localizes on the positive strand with $tss < pa$. However, indices are likewise calculated for transcription units on the negative strand with few adaptations. Furthermore, the index depends on the Pol II occupancy track of the control experiment Occ^C and upon treatment Occ^T. The following steps describe the multi-step procedure.

1. The Pol II occupancy tracks Occ^C and Occ^T were pooled into a pseudo-sample Occ^{CT}.

2. The *extended termination zone length* (TZ) is calculated as described in Section 6.3.3, using the pooled pseudo-sample Occ^{CT} and a bin size of $w = 5,000$. The resulting parameter TZ indicates the length of the zone where RPM normalized Pol II occupancy occurs above the threshold of $thr = 0.2$ in the pooled experiments. However, each sample had to have a calculated RPKM value (Section 3.2.3) of > 0.01 in the extended termination zone, avoiding outliers from influencing the next step disproportionately.

3. Next, the ATD^C and ATD^T values were calculated as described in Section 6.3.4 for both samples Occ^C and Occ^T separately, using the previously identified parameter TZ. The respective value is considerably higher in samples where Pol II readthrough transcription occurs than in the control experiment.

4. Finally, the *transcriptional readthrough index* (RTI) is defined as

$$RTI = ATD^T - ATD^C. \qquad (14.4)$$

This study calculated RTI values from NET-seq, HiS-NET-seq, and GRO-seq data.

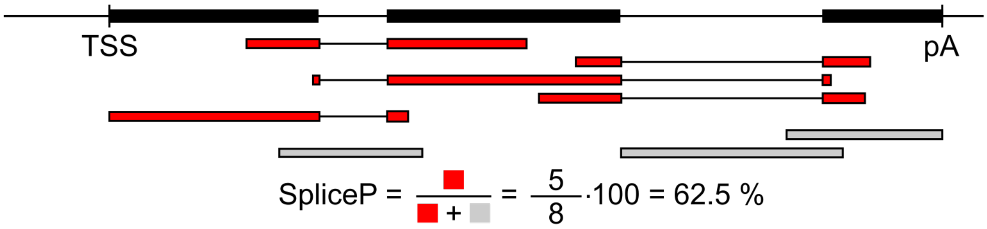


Figure 14.3: Splicing Analysis Example. The schematic depicts the calculation of *SpliceP* for an example gene. The percent of spliced RNA molecules is obtained by the ratio of spliced reads (red) vs. all reads that span the gene and at least one splice site.

Readthrough Index for Antisense Transcription

Antisense transcription systematically occurred on the opposite strand upstream of the TSS (Figures 7.1A, 7.3B, and 7.5B). The calculation of the TZ' and RTI' values for antisense transcription was similar to the calculation of TZ and RTI values for sense transcription with few adjustments. First, the Occ'^C and Occ'^T values report the Pol II occupancy measurements from the **negative strand** in reverse order for a reference genome of size *genomeSize*. Next, a new pseudo transcript was generated, where $tss' = genomeSize - pa$ and $pa' = genomeSize - tss$. Subsequently, the steps in the previous section were performed using tss', pa', Occ'^C, and Occ'^T.

14.3 SPLICING ANALYSIS

Many different approaches exist for alternative splicing analysis [33]. For the context of this study, the definition of a simple splicing score was sufficient.

Let mR be the number of sequencing reads mapping to a gene of interest and overlapping at least one of the annotated splice sites from GENCODE [67]. The value $spliced_i \in \{0, 1\}$ is defined for each sequencing read $i \in \{1, \ldots, mR\}$, reporting 1 if the corresponding sequencing read is spliced and 0 otherwise. The *percent of spliced RNA molecules* (*SpliceP*) of the gene is defined by

$$SpliceP = \frac{\sum_{i=1}^{mR} spliced_i}{mR} \cdot 100. \tag{14.5}$$

This study calculated *SpliceP* from processed RNA-seq data (Section 14.1.1) without PCR duplicates. PCR duplicates were marked and removed using PICARD's v2.24.2 [95] *markDuplicates* function.

14.4 3′-RNA CLEAVAGE EFFICIENCY TEST

Cleavage of nascent RNA at the 3′ end of transcription units is an essential step for Pol II termination and is described in Section 2.2.3.

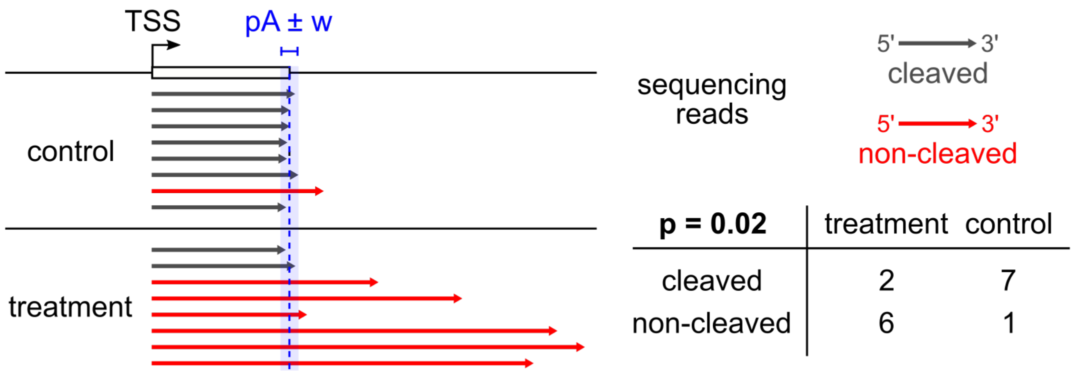


Figure 14.4: 3′-RNA Cleavage Efficiency Test Example. The schematic depicts the *3′-RNA cleavage efficiency* test for an example polyA site. Non-cleaved sequencing reads (red) spanning the polyA site region (blue) increased compared to cleaved sequencing reads (grey) that map with their 3′ ends to the polyA site region upon treatment.

This study developed a *3′-RNA cleavage efficiency* test for nascONT-seq data (Section 14.1.3) to identify significant changes in the number of sequencing reads that span a polyA site between a control and treatment experiment. Briefly, the approach identified active polyA sites, quantified the amount of cleaved and non-cleaved sequencing reads, and performed a *Fisher's* exact test. A changing ratio of non-cleaved vs. cleaved sequencing reads at a polyA site indicates significant deregulation upon treatment.

1. The approach defined up to two active polyA sites per active gene (Section 6.1.1), showing the highest polyA_DB v3.2 [243] database score or the most number of cleaved sequencing reads in the control experiment. The latter is identified by extracting all annotated human polyA sites from the database and identifying the number of cleaved sequencing reads for each polyA site. The corresponding sequencing reads mapped precisely with their 3′ end to the annotated polyA site nucleotide position.

2. Next, the number of cleaved and non-cleaved sequencing reads was quantified for both conditions at each active polyA site in a region +/- w nucleotides upstream and downstream, where $w = 20$. Cleaved sequencing reads mapped with their 3′ ends to the polyA site region, whereas non-cleaved reads spanned the polyA site region. Each polyA site was summarized in a contingency table, as shown in Table 14.1.

	treatment	control	total
cleaved	A	B	$A+B$
non-cleaved	C	D	$C+D$
total	$A+C$	$B+D$	$A+B+C+D$

Table 14.1: 3′-RNA Cleavage Efficiency Table.

3. The *Fisher's* exact test was performed as described in Section 3.4. The one-tailed *Fisher's* exact test calculated a P-value from the contingency table, using Equation 3.16. Cleavage efficiency is significantly reduced at the polyA site if $P \leqslant 0.05$.

Figure 14.4 schematically visualizes steps two and three of the test, identifying a significant reduction of cleavage efficiency in the example.

15 RESULTS

15.1 THE BRD4-MEDIATED 5′ ELONGATION CHECKPOINT

15.1.1 *Global Decrease of Nascent RNA Transcription*

This study used the targeted protein degradation system to investigate BRD4's protein function in the K562 dTAG-BRD4 cell line (Section 2.4.4). The western blot and mass spectrometry experiments revealed a robust degradation of both BRD4 isoforms after two hours of dTAG7 treatment (Figures A.5A and A.5B). First, the gold standard experiment for transcription studies, total RNA-seq (Section 2.4.1), was performed (MRA101-MRA102, MRA105-MRA106, unpublished). Total RNA-seq experiments measure changes in RNA levels, primarily from mature RNA in the cytoplasm. Surprisingly, the differential gene expression analysis identified a significant change in expression levels for 26% of all genes (Figures A.5C and A.5D). The number of identified genes was unexpectedly low compared to the dramatic reductions of transcript levels measured in previous work upon pan-BET protein degradation [249]. A possible explanation for the weak response was the limited treatment time of two hours. Instead of increasing the treatment time and introducing cell compensation and adaptation effects, the more sensitive nuclei-RNA-seq was used (NE04-NE09, unpublished). The approach measures recently produced RNA in the nuclei. Nuclei-RNA-seq revealed a more global decrease of RNA transcripts with significant reductions at 51% of genes (Figures A.6A and A.6B).

Next, different NET-seq protocols with whole-cell mouse spike-ins were applied to gain insights into deregulated Pol II transcription mechanisms. The spike-in strategy, introduced in Section 11.2.2, was successfully applied to HiS-NET-seq upon 40 and 120 minutes of BRD4-specific degradation (OJ94-OJ97, MRA125-MRA128, MRA144-MRA147, unpublished, Figure A.6C). HiS-NET-seq reported significant reductions at 29% and 89% of actively transcribed genes after 40 and 120 minutes, respectively (Figures 15.1A and 15.1B). Upon 40 minutes of treatment, the absence of BRD4 led to a reduction of Pol II at short genes. After 120 min, longer genes were also affected (Figure 15.1C). BRD4-resistant genes without a significant Pol II occupancy change were short and encoded mainly for histone and ribosomal proteins (PANTHER Protein Class [153]; FDR: 2.98E-10 and 1.95E-02). Measurements of SI-NET-seq (GSE158963 [7]) upon 120 minutes of BRD4 loss revealed similar trends.

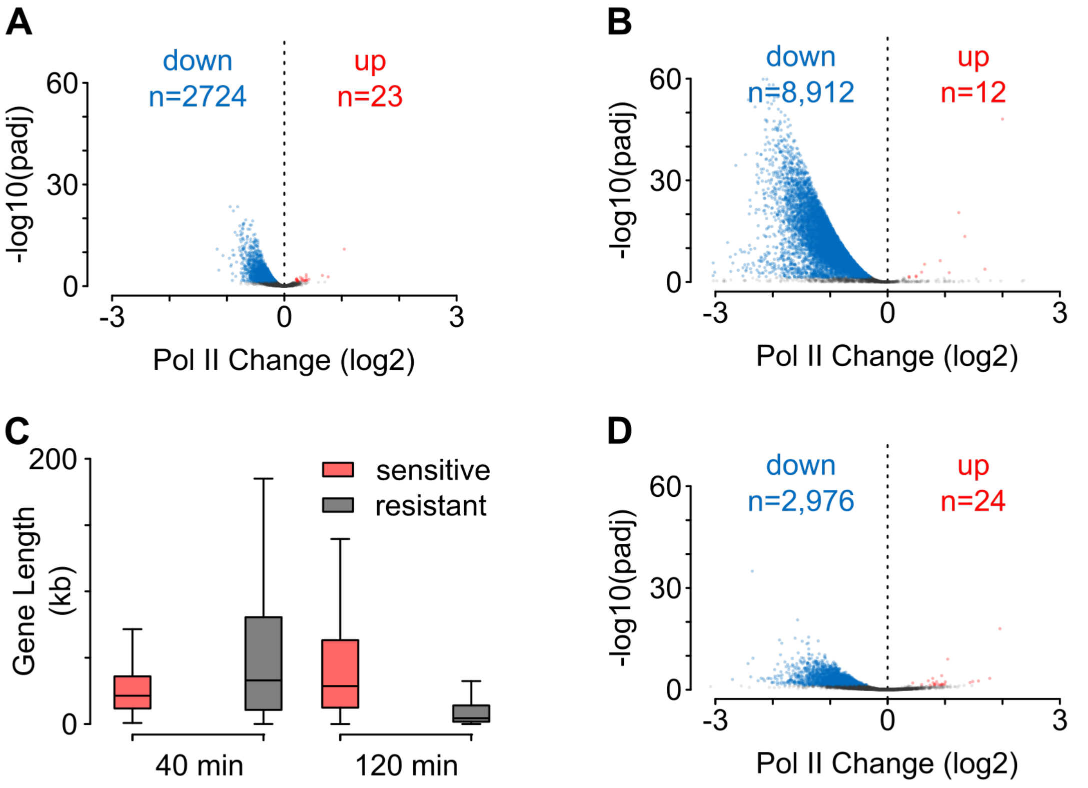


Figure 15.1: BRD4-specific Degradation Decreases Nascent RNA Transcription. Considered are non-overlapping actively transcribed genes from human K562 dTAG-BRD4 cells. **(A-B)** Pol II occupancy changes (log2) identified with HiS-NET-seq upon **(A)** 40 minutes (n=9,358) and **(B)** 120 minutes (n=10,058) of treatment. **(C)** Gene length distribution at BRD4 sensitive and resistant genes after indicated treatment times. BRD4 sensitive genes show significant reduction of Pol II (padj < 0.05, n=2,724 and n=8,912), whereas BRD4 resistant genes show no change (padj > 0.05, n=6,634 and n=1,146). **(D)** Pol II occupancy changes (log2) identified with SI-NET-seq upon 120 minutes of BRD4 degradation (n=9,199). **(A, B, D)** Significant occupancy changes (padj < 0.05) are labeled in blue and red.

However, the method identified a significant Pol II reduction for 32% of the genes, considerably less than 89% determined by HiS-NET-seq (Figures 15.1D and A.6D). This difference demonstrates the considerable sensitivity gain of the HiS-NET-seq method.

Overall, depending on their sensitivity, different transcriptional assays identified a significant role of BRD4 for productive Pol II gene transcription.

15.1.2 *BRD4-degradation Impairs Pol II Pause Release*

Gene transcription and regulation are divided into different stages (Section 2.2.1), which can be resolved by high-resolution Pol II profiling methods, such as NET-seq and protocol variants. The following analysis focused on differences in the regions associated with promoter-proximal pausing and productive elongation. Early elongation, the transcriptional stage where Pol II promoter-proximal pausing and release occurs, spans a few hundred nucleotides downstream of the TSS (Figure 15.2A).

The gene-body region is associated with the productive elongation stage, which starts after the promoter-proximal region and ends at the polyA site (Figure 15.2A). Visual inspections of HiS-NET-seq data indicated, on the one hand, increased Pol II occupancy levels at promoter-proximal regions. On the other hand, the trend shows decreased coverage across gene-body regions after 120 minutes of treatment (Figures 15.2B). Likewise, these trends were observable for individual gene examples (Figures 15.2C and A.7A).

The DPO analysis was applied to the respective regions and summarized in pausing matrices to confirm these changes at the individual gene level (Figures 15.2D-15.2E and A.7B-A.7C). After 40 minutes of BRD4-specific degradation, Pol II occupancy increased in the promoter-proximal region and decreased in gene-body regions (Figure 15.2D). The changes occurred equally in the respective regions, which suggested a dysfunctional release from promoter-proximal pausing.

With an increased treatment time of 120 minutes, the reduction of productive elongation intensified, affecting almost all genes (94%). In contrast, signals in the promoter-proximal regions did not increase further (Figure 15.2E). SI-NET-seq confirmed the results (Figures A.8A-A.8C), however, identified fewer genes with significant gene-body reduction ($padj < 0.05$, 42%). The results suggested that rapid BRD4-specific degradation impairs the Pol II pause release.

15.1.3 *Distinct Role of Other BET Proteins*

Previous studies using pan-BET degradation in MOLT4 cells [249], and this study, which removes specifically BRD4 in K562 dTAG-BRD4 cells, revealed extensive transcriptional defects.

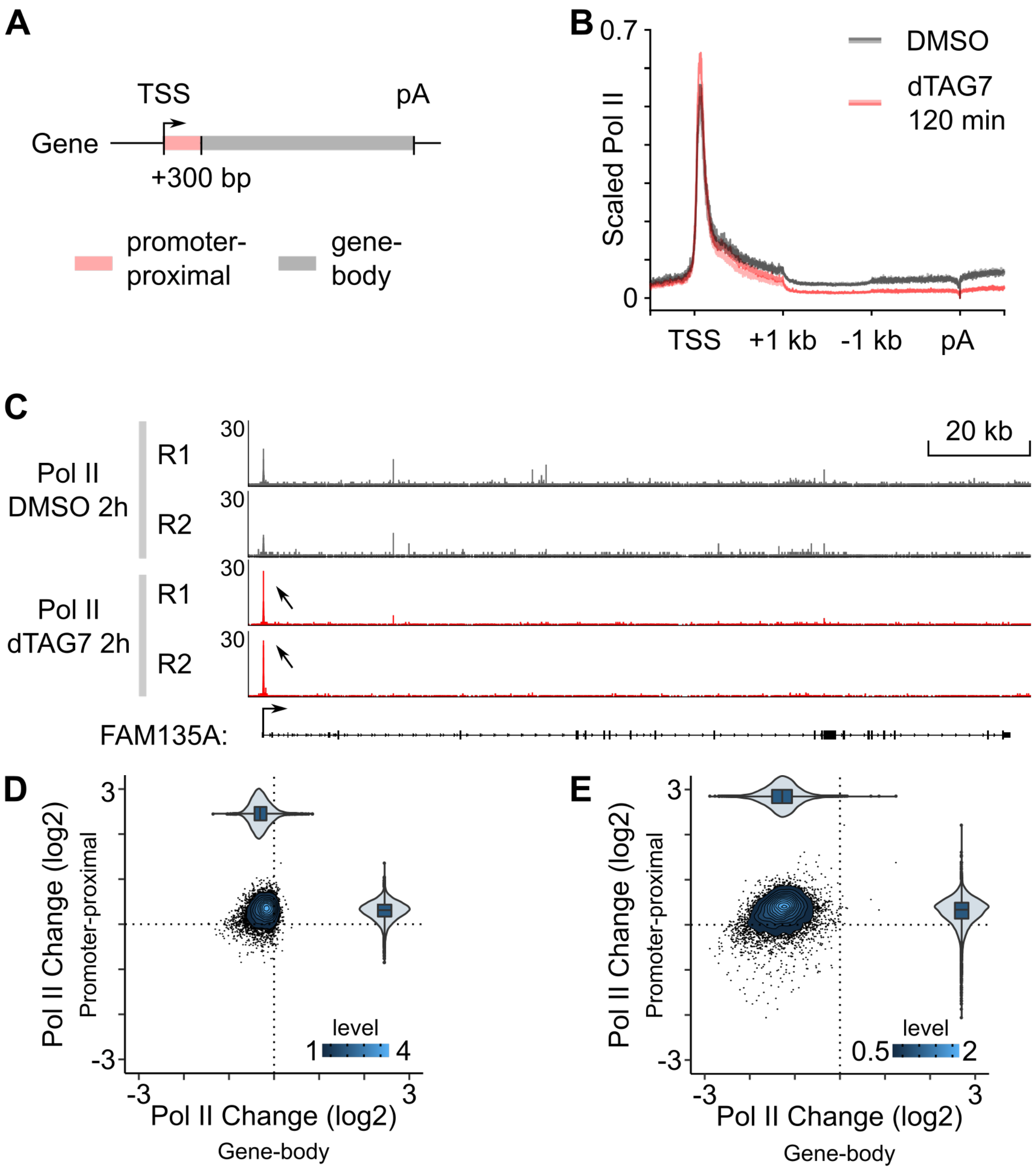


Figure 15.2: Disruption of Pol II Pause Release by BRD4. The figure depicts comparisons between Pol II data at actively transcribed non-overlapping genes from the control experiment (DMSO) and two hours of BRD4-protein degradation (dTAG7) with two HiS-NET-seq replicates measurements in human K562 dTAG-BRD4 cells. **(A)** Definition of promoter-proximal and gene-body regions. **(B)** Meta-gene profile of reference normalized Pol II (n=9,255, Section 3.2.6). Excluded were regions with signal outliers above the 99.90-quantile and the TSS. **(C)** Reference-normalized (Section 3.2.6) Pol II occupancy at an example gene. **(D, E)** Pol II occupancy changes at promoter-proximal (y-axis) and gene-body regions (x-axis) upon **(D)** 40 minutes (n=7,303, four replicate measurements) and **(E)** 120 minutes (n=7,641) of BRD4-specific degradation.

The investigations focused next on a direct comparison of pan-BET and BRD4-specific protein degradation to dissolve the roles of BRD4 and other BET proteins, namely BRD2 and BRD3. Experiments had comparable conditions, such as treatment time, cell line, and method. HiS-NET-seq was still under development when experiments were performed. For this reason, Pol II changes upon pan-BET degradation were measured in K562 dTAG-BRD4 cells with SI-NET-seq (GSE158963 [7]; Figure A.8A). The comparison identified two main differences after two hours of treatment. First, BRD4-specific degradation results in enhanced promoter-proximal pausing, whereas the removal of pan-BET proteins revealed a contrary trend (Figures A.8B-A.8E). Second, pan-BET protein degradation induced a more vigorous response, which reduced Pol II levels by 62% in gene-body regions. The median reduction after BRD4 loss was with 33% smaller. The results ascribe a considerable role in Pol II regulation to other BET proteins, BRD2, BRD3, or both.

15.2 BET PROTEINS REGULATE 3′-RNA PROCESSING

15.2.1 *Widespread Pol II Readthrough Transcription*

Next, the study explored whether BRD4 serves additional roles in Pol II transcription using HiS-NET-seq. Strikingly, this analysis uncovered that acute loss of BRD4 proteins induced Pol II readthrough transcription at the 3′ ends of genes (Figures 15.3A). The already introduced *average termination distance* was applied to quantify the impact of treatments on Pol II termination (ATD, Section 6.3.4). The ATD difference of a gene between conditions, referred to as the *transcriptional readthrough index* (RTI, Figure 15.3B and Section 14.2), describes the average changes of the Pol II distribution in the termination zone relative to the pA site. Hence, a positive RTI reveals a downstream shift of Pol II distribution away from the pA site and indicates readthrough transcription.

BRD4 loss induced a downstream shift of Pol II occupancy with a median RTI of 1.7 kb after 120 minutes of treatment. In contrast, the short degradation time showed no increase (Figure 15.3C). A similar shift of 1.4 kb was observed with SI-NET-seq (Figure A.9A). The most apparent termination defects were observed for protein-coding and lncRNA genes (Figure 15.3D). Both gene classes rely on a polyA signal-dependent termination pathway. No readthrough was observed for gene classes with non-canonical 3′-RNA processing, including

- *antisense transcription units* (Figures A.9B),
- histone genes,
- sn/snoRNA, and
- *micro RNA* genes (Figure 15.3D).

Together, these analyses indicate that BRD4 is required for transcription termination at a group of protein-coding and lncRNA genes.

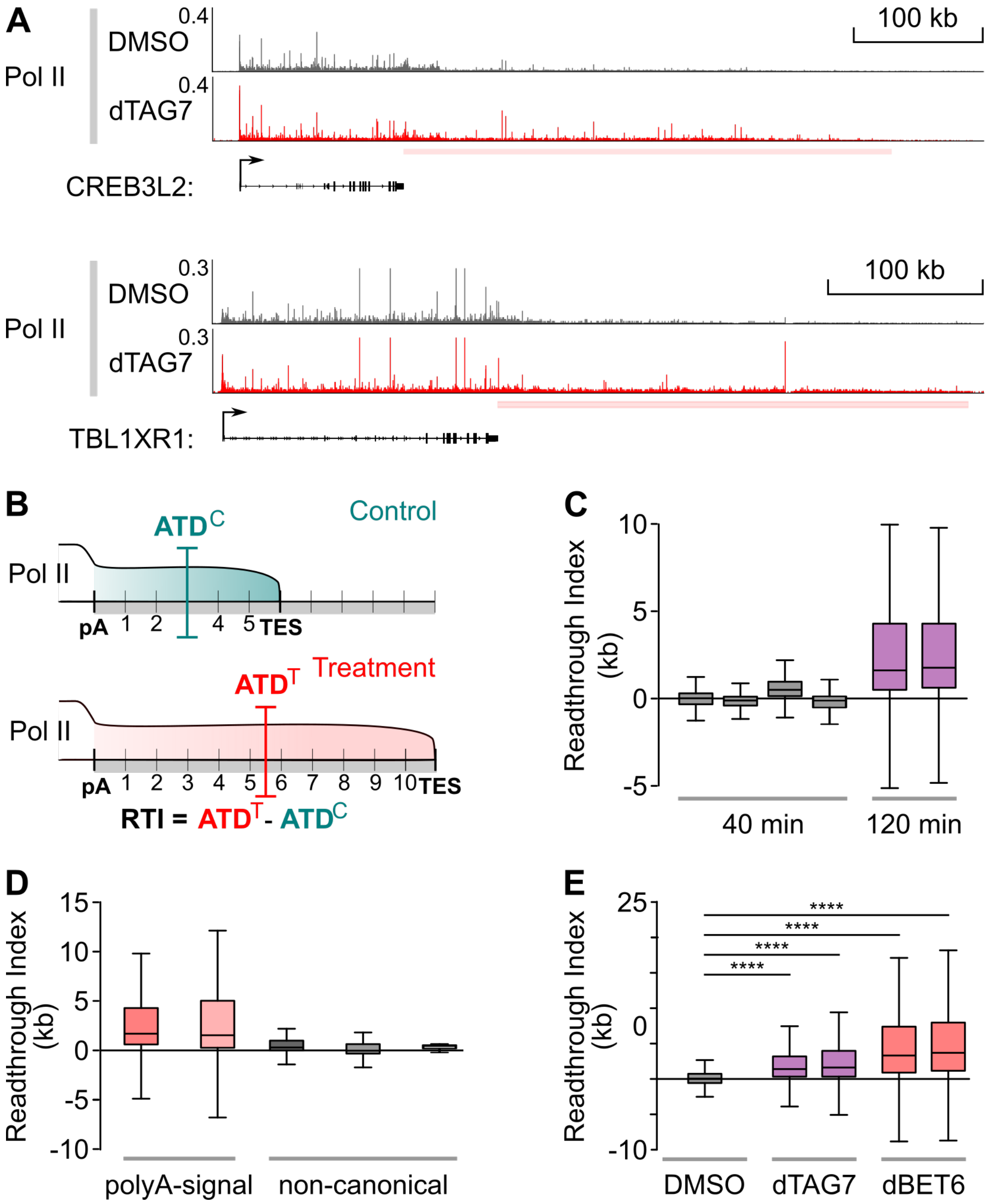


Figure 15.3: BRD4-specific Degradation Induces Pol II Readthrough Transcription. Pol II comparison at actively transcribed genes from human K562 dTAG-BRD4 cells measured by HiS-NET-seq upon 120 minutes of BRD4 loss. **(A, D)** Replicates are pooled for visualization. **(A)** RPM normalized Pol II occupancy at two example genes. **(B)** Schematic view of the RTI calculation based on the ATD (Section 14.2). **(C-E)** Boxplot quantification of RTI values for the indicated **(C)** treatment times (40 minutes: n=9,581-9,619; 120 minutes: n=9,608-9,646), **(D)** gene-classes (polyA-signal: protein-coding (n=8,652), lncRNA (n=549); non-canonical: histone (n=47), sn/snoRNA (n=94), *micro RNA* (n=7)), and **(E)** between pan-BET (dBET6) and BRD4-specific degradation (dTAG7) using SI-NET-seq (DMSO: n=9,446; dTAG7: n=9,418-9,439; dBET6: n=9,114-9,292; one tailed *Wilcoxon* rank sum test ****: p < 2.2e-16).

PAN-BET DEGRADATION AMPLIFIES TERMINATION DEFECTS The level of deregulation on termination caused by BRD4 was compared to pan-BET degradation using comparable SI-NET-seq data. Total loss of BET proteins, which removes BRD2, BRD3, and BRD4, intensified the observed termination defects, with a median RTI of 3.4 kb compared to 1.4 kb for BRD4-specific degradation (Figures 15.3E). Although pan-BET loss results in more pronounced termination defects, the correlation of RTI values was high between both treatments (r = 0.79; Figure A.9C). In the following, genes with a RTI of 5 kb or higher are referred to as readthrough genes. This conservative threshold is required due to the high variance of RTI values calculated between control measurements (Figure A.9D). With this threshold, 37% and 14% of genes were classified as readthrough genes upon pan-BET and BRD4-specific degradation, respectively. Overall, 90% of BRD4-related readthrough genes overlapped with those identified upon pan-BET degradation (Figure A.9E). These results suggest that BRD4 acts together with other BET proteins to regulate Pol II termination.

15.2.2 *Functional Consequences of Termination Defects*

The observed termination defects with uncontrolled Pol II transcription suggested potential consequences, such as increased transcription downstream of readthrough genes. Because transcriptional defects of BRD4 degradation after two hours were less pronounced at mature RNA levels (Figure A.5D), this part focused on published total RNA-seq data obtained upon two and six hours of pan-BET protein degradation in the MOLT4 cell line (GSE79253, [249]). Besides the global reduction of mature RNA levels, which the authors highlighted, further inspections of the differential gene expression results revealed a sub-group of genes with increased transcript levels. Overall, 596 (3%) and 1,689 (7%) genes were up-regulated upon two and six hours of pan-BET degradation.

Most of those genes were located within or downstream of readthrough genes (81%, TSS to TES + 100 kb). However, it was unclear if the increase in transcription at readthrough-associated genes produced functional RNA transcripts that could be translated. Therefore, the analysis focused on up-regulated genes activated by readthrough transcription to investigate if transcripts were spliced. For this group of genes, no RNA products were measured in the respective control experiments (TPM < 1).

The two example genes *TRIM72* and *ITGAM*, located downstream of the readthrough gene *FUS* (RTI = 17.7 kb), were activated after six hours of treatment (Figure A.10A).

Interestingly, transcript levels did not increase uniformly across the genes but specifically at exon regions. Visualization of spliced reads in this region confirmed that novel splice events appeared (Figure A.10B). The same trend was observed globally for activated genes, showing increased spliced RNA molecules with treatment time (Figure A.10C, Section 14.3).

Although the results suggest *de novo* activation and processing of genes by readthrough transcription, these experiments and analyses could not answer finally whether the produced transcripts were functional. Together, these data show that the transcriptional readthrough correlated with the enhanced and activated expression state of neighboring genes.

15.2.3 *3′-RNA Cleavage Defects*

Since Pol II readthrough transcription was primarily observed at polyA signal-containing genes, subsequent analyses addressed whether pan-BET or BRD4 protein degradation affected 3′-RNA cleavage. However, the available RNA-seq data sets, collected after short treatment times, were unsuitable for this analysis. Most measured transcripts were produced when no termination defect was observed (Figure 15.3C). Furthermore, only few sequencing reads overlap with active cleavage sites.

The limitations could be addressed by tracking newly produced whole nascent RNA transcripts with long-read nanopore sequencing [97]. This combination led to the new method nascONT-seq, which combined metabolic labeling (4sU), chromatin fractionation to enrich nascent RNA transcripts, and ONT's long-read sequencing (Figure 15.4A). However, the ONT technology was not designed for sequencing nascent RNA transcripts that lack polyA tails. Therefore, nascONT-seq adapted a recently published approach [51] that adds polyA tails to nascent RNA transcripts, enabling the sequencing through nanopores.

As transcription termination defects were more pronounced after pan-BET degradation, the new method was tested in human K562 cells after two hours of pan-BET protein degradation (GSE158964, [7], Figure A.11A). The new method confirmed Pol II readthrough transcription at selected genes (Figure 15.4B) and genome-wide (Figure A.11B). Next, the *3′-RNA cleavage efficiency* test was developed to investigate whether 3′-RNA cleavage defects directly caused the observed readthrough transcription (Section 14.4). In the first step, sequencing reads at individual polyA sites were classified into cleaved and non-cleaved. Cleaved sequencing reads mapped to the respective polyA site, whereas non-cleaved reads spanned the region. This classification, combined with the one-tailed *Fisher's* exact test, identified deregulated 3′-RNA cleavage sites between conditions ($p\text{-value} < 0.05$). 3′-RNA cleavage efficiency significantly decreased at 296 (14%) polyA sites when replicate measurements were pooled. Genes with 3′-RNA cleavage defect showed considerably higher RTI values than genes without 3′-RNA cleavage efficiency changes (Figure 15.4C).

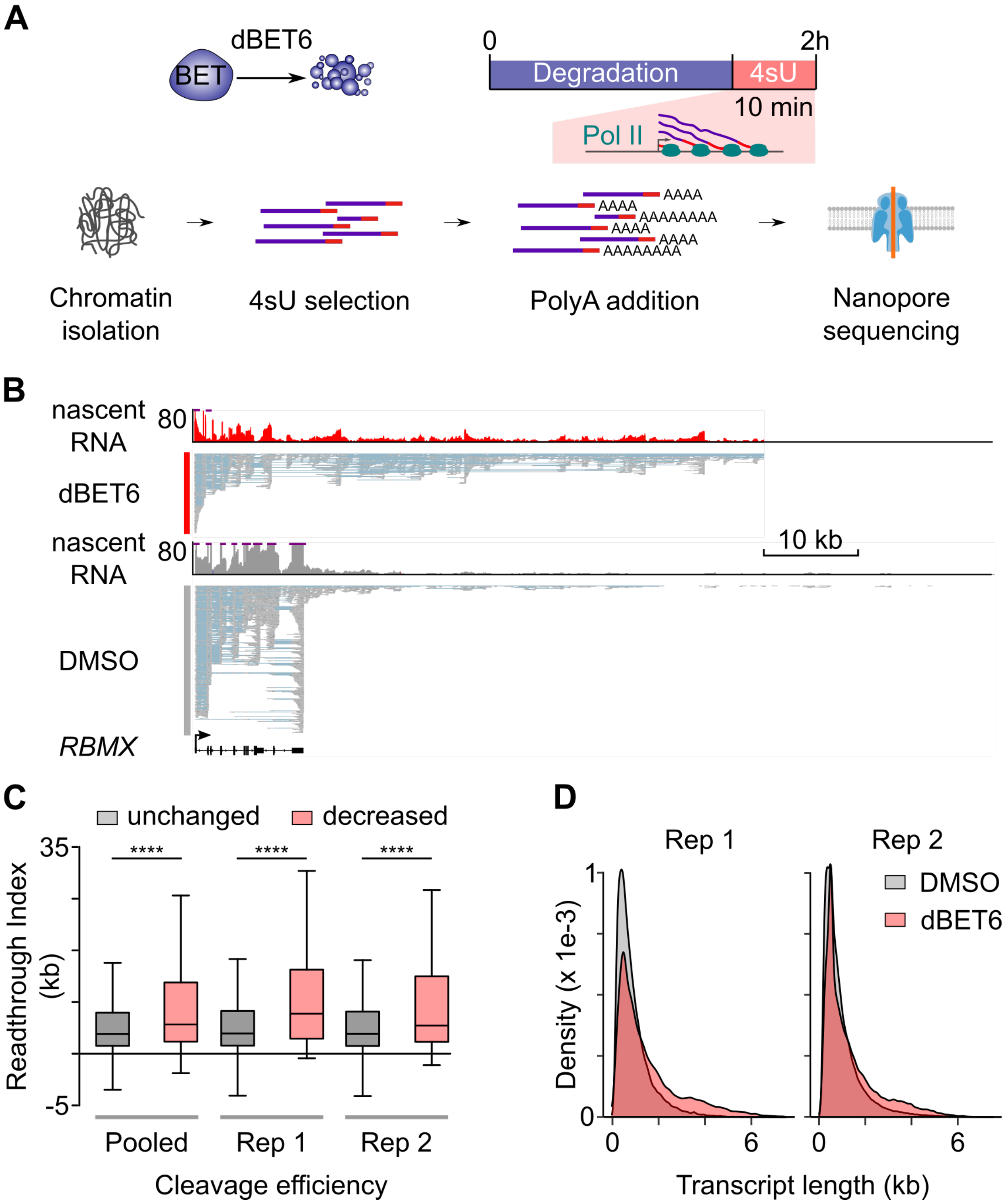


Figure 15.4: BET Proteins are Required for 3′-RNA Cleavage. The figure shows nascONT-seq data upon two hours of pan-BET degradation (dBET6) and the respective controls (DMSO) in human K562 cells. **(A)** Scheme of nascONT-seq. **(B)** Gene track of nascent RNA levels and of individual transcripts. The y-axis depicts the nascent RNA level (reads/bp). Biological replicates were pooled for visualization. **(C)** Boxplot quantification for RTI values at genes with no change or decreased 3′-RNA cleavage efficiency (pooled: n = 1,766, n = 284; rep 1: n = 2,006, n = 85; rep 2: n = 1,867, n = 207; *Wilcoxon* rank sum test ****: p < 3.8e-05). **(D)** Length distribution of nascent transcripts for DMSO (n = 627,995, n = 1,482,187) and treatment (n = 579,228, n = 1,441,147). Only full-length transcripts are depicted (Section 14.1.3).

Finally, nascONT-seq revealed extended transcripts upon treatment with an increased median transcript length of 938 bp compared to 689 in the control experiment (Figure 15.4D, Section 14.1.3). These findings suggest that BET proteins are required for 3′-RNA processing of polyA signal-containing genes.

15.2.4 *BRD4 Binds 5′ Regions of Readthrough Genes*

Transcription defects at the 3′ ends of genes were unexpected because neither BRD4 nor other BET proteins are known to bind near this region [260]. For clarification, BRD4 binding was tested in K562 dTAG-BRD4 cells using ChIP-Rx [170] (MRA111-MRA112, MRA115-MRA116, unpublished). The resulting profiles show BRD4 binding primarily at the 5′ region of non-overlapping genes (Figure 15.5A). Overall, 85% and 4% of non-overlapping genes had at least one BRD4 peak in the 5′ and 3′ regions. These observations excluded the possibility that BRD4 acts directly at polyA sites to regulate 3′-RNA cleavage and prevent Pol II readthrough transcription.

Next, other potential features were investigated that could characterize affected genes. Readthrough genes were significantly longer and revealed higher steady-state gene expression levels (Figure A.12A). Additionally, readthrough genes had a higher AT content, defined as the percentage of A and T downstream of the polyA site. In this region, genes without transcriptional readthrough had a higher GC content, defined as the percentage of G and C (Figure A.12B). Although BRD4 does not bind polyA sites, readthrough genes had significantly more BRD4 bound at 5′ gene regions (Figure 15.5B) and showed more pronounced elongation defects (Figure 15.5C). This observation leads to the hypothesis that BRD4 binding at 5′ ends of genes impacts 3′-RNA processing and termination.

Next, BRD4 reduction across the genome was tracked at readthrough genes upon 40 and 120 minutes of BRD4 degradation with ChIP-Rx (MRA109-MRA110, MRA113-MRA114, unpublished; Figure A.12C). ChIP-Rx [170] is a ChIP-seq variant, including an exogenous reference genome for each sample. Like SI-NET-seq, this additional step allows genome-wide and quantitative comparisons between conditions. The results confirmed previous western blot and mass spectrometry experiments (Figures A.5A and A.5B) and revealed a global reduction of BRD4 (Figure 15.5D). Notably, the comprehensive BRD4 loss was already pronounced after 40 minutes of treatment (Figure A.12D) and was most potent at readthrough genes (Figure 15.5C). The data shows that the termination defect correlated with elongation defects and BRD4 binding in 5′ gene regions.

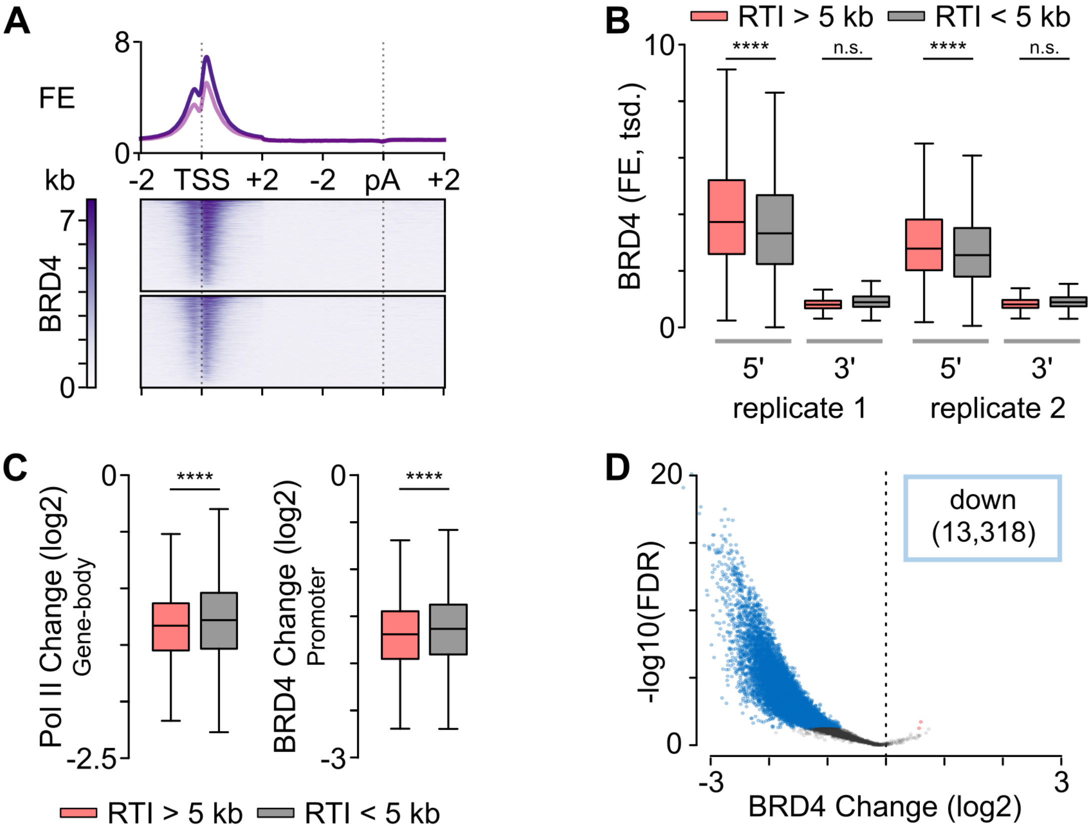


Figure 15.5: Features of BRD4 Binding and Readthrough Genes. Figure presents FE normalized (Section 3.2.7) BRD4 ChIP-Rx data for the control experiment and upon two hours of BRD4 degradation. Considered are actively transcribed non-overlapping genes (TSS to pA site + 5 kb) and BRD4 peaks in human K562 dTAG-BRD4 cells. **(A)** Meta-gene profile of BRD4 occupancy at genes (DMSO, n=7,332, minimum gene length 6 kb). **(B)** BRD4 occupancy at 5′ (TSS +/- 1 kb) and 3′ regions (pA site +/- 1 kb) of readthrough (red, n=1,824) and non-readthrough genes (gray; n=6,618, n.s: p-value = 1, ****: p < 7.7e-12) **(C)** Comparison of of readthrough and non-readthrough genes. Depicted are Pol II occupancy changes (log2) at gene-body regions identified by HiS-NET-seq upon two hours of BRD4-protein degradation (n=1,789, n=6,284; ****: p = 3.9e-09). Furthermore, BRD4 ChIP-Rx occupancy changes (log2) are identified with DiffBind [221] as described in Section 14.1.4. Depicted are changes for peaks that overlapped 5′ gene regions (TSS +/- 1 kb, n=1,589, n=5,361; ****: p = 3.8e-08) and **(D)** all detected BRD4 peaks (n=16,233). Significant occupancy changes (FDR < 0.05) are labeled in blue and red. **(B, C)** One tailed *Wilcoxon* rank sum tests were performed.

15.3 BRD4-DEPENDENT RECRUITMENT OF THE 3′-RNA PROCESSING MACHINERY

15.3.1 *Recruitment Defects of 3′-Processing Factors*

It is known that the disruption of 3′-RNA processing and termination factors can cause 3′-RNA cleavage defects and readthrough transcription [175, 193]. However, a direct regulatory function of BRD4 on 3′-RNA cleavage was unlikely due to its localization at the 5′ ends of genes (Figure 15.5A). At this point, the experiments investigated whether BRD4 loss indirectly perturbed the chromatin localization of relevant 3′-RNA processing factors. *Chromatin mass spectrometry* (chromatin-MS, [7]), which measures protein composition changes at the chromatin upon treatment, identified 76 proteins that were immediately displaced from the chromatin upon acute loss of BRD4 (p-value < 0.05; Figures A.13A). Corresponding to the observed phenotype, the GO term analysis identified dysfunctional biological processes related to 3′-RNA processing and transcription elongation (Figure 15.6A). Overall, ten factors were implicated in 3′-RNA processing, most were CPSF and CstF sub-units (Figure A.13C). Another identified class consisted of elongation factors (Figure A.13A). Similar results were obtained upon pan-BET degradation (Figure A.13B).

To investigate whether the recruitment of the 3′-RNA processing machinery was perturbed at readthrough genes, different 3′-RNA processing factors from the CPSF (FIP1 and CPSF73) and CstF (CstF64) complex were selected for ChIP-Rx experiments (GSE158965, [7]). The mean binding profile of these factors revealed peak occupancy levels at 5′ and 3′ ends of active genes (Figure 15.6B). The abundance of these factors at 5′ gene regions varied. In total, 24% (FIP1), 4% (CPSF73), and 15% (CstF64) of all binding sites overlapped with at least one active TSS (+/- 300 bp). The results suggested that some 3′-RNA processing factors could be recruited during an early transcription phase, presumably during Pol II initiation or early elongation.

BRD4 loss reduced the occupancy of these factors at both 5′ and 3′ regions of genes (Figures 15.6C, A.14A, and A.14B). Additional Pol II ChIP-Rx experiments were performed (GSE158965, [7]) to exclude the possibility that the observed trends were a result of an overall reduced Pol II level at the chromatin (Figure 15.1B). The Pol II-normalized data (Section 3.2.7) revealed that fewer 3′-RNA processing factors were recruited to Pol II in the 5′ regions of genes upon BRD4 degradation (Figures A.15-A.17). This BRD4-dependent recruitment defect was more pronounced for gene regions with perturbed 3′-RNA cleavage than unaffected (Figures A.15-A.17). These findings present impaired recruitment of 3′-RNA processing factors as a plausible cause for the 3′-RNA cleavage defect and the transcriptional readthrough.

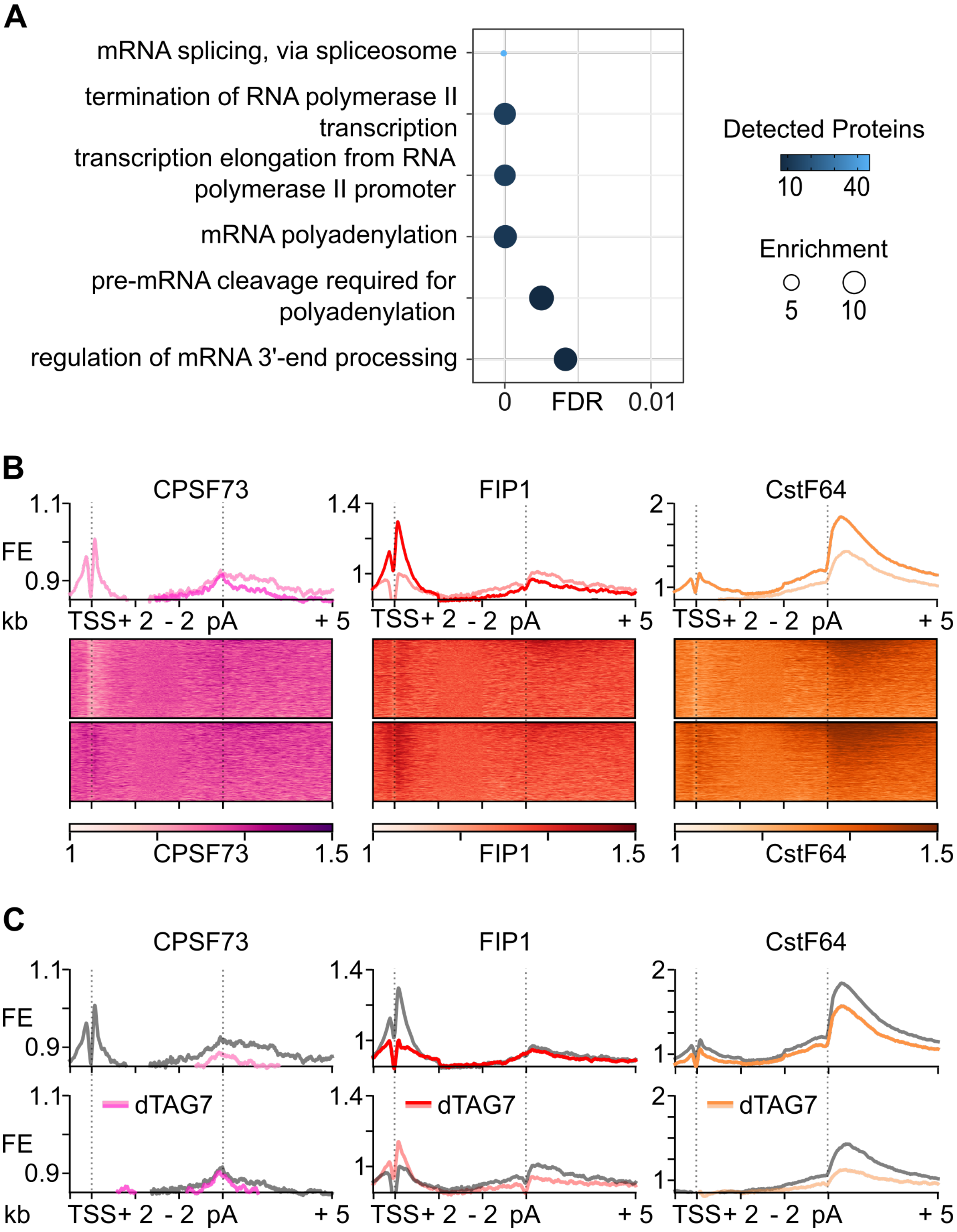


Figure 15.6: BRD4 Recruits 3′-RNA Processing Factors. (A) Significant (FDR < 0.01) GO terms for proteins that were depleted from the chromatin after two hours of BRD4 degradation (n=76, p-value < 0.05). Excluded were GO sub-terms for visualization. **(B-C)** Meta-gene profiles of occupancy levels measured by ChIP-Rx for 3′-RNA processing factors at non-overlapping actively transcribed genes (n=7,331; gene length > 6 kb). Depicted are FE-normalized occupancy values of the respective protein of interest (Section 3.2.7) in the **(B)** control experiment and **(C)** upon two hours of BRD4 degradation (dTAG7).

15.3.2 *Core-Interactome of BRD4 Reveals 5′ and 3′ Regulators*

Potential BRD4 interaction partners were identified using BRD4 *immunoprecipitation followed by mass spectrometry* (IP-MS, [7]) to understand the underlying mechanism that caused the BRD4-dependent elongation and termination defects. The experiment provided a comprehensive list of 379 significant BRD4 interactors (FDR < 0.05, Figure A.18). An integrated analysis of the two proteomic data sets provided insights into the BRD4 core-interactome. This concept summarizes the 29 significant interactors with immediate displacement from the chromatin upon BRD4 loss ($p < 0.05$). The respective GO term analysis reported the enrichment of 5′ elongation and 3′-RNA processing factors (Figure 15.7A). Among the top-ranking candidates were factors of the CPSF (CPSF160, FIP1), *cleavage factor Im* (CFIm25), CstF (CstF77), DSIF (SPT5), and PAF (PAF1, CDC73) complexes (Figure 15.7B). The DSIF and PAF complexes are known regulators of Pol II elongation [232, 233].

These observations lead to the hypothesis that BRD4 underlies a general 5′ elongation control point that primes transcribing Pol II for 3′-RNA processing and termination. Subsequently, ChIP-Rx was performed for some detected elongation factors, namely PAF1 and SPT5 (GSE158965, [7]). As a result, BRD4 interestingly co-localized with the 3′-RNA processing (CPSF and CstF) and the elongation factor PAF1 downstream of the TSS in the promoter-proximal region (Figure 15.7C). In contrast, SPT5 occupancy accumulated 90 nucleotides upstream of the identified control point. These data suggest that BRD4 regulates 5′-elongation and 3′-RNA processing through functional interactions at a 5′ control point.

15.3.3 *Contribution of Other Elongation Factors on 3′-RNA Processing Defects*

Examining the PAF1 and SPT5 binding profiles at TSSs revealed co-localization with BRD4 and 3′-RNA processing factors. Their presence at 5′ active gene regions was expected because both factors are described elongation factors.

However, the binding profile of PAF1 at active genes revealed a more unexpected profile with a substantial accumulation at active 3′ gene regions (Figure 15.8A). Although SPT5 was present in the 3′ regions, the overall binding of PAF1 was more similar to the CstF complex (CstF64) than the elongation factor SPT5 (Figure 15.8A). The re-analysis of other PAF sub-units in an *acute monocytic leukemia* cell line (THP-1) confirmed this finding (GSE62171, [253], Figure A.19A). The co-localization of PAF and 3′-RNA processing factors suggested potential interactions verified in native immunoprecipitation experiments for several PAF sub-units (Figure A.19B).

Further investigations focused on the relative changes of these factors compared to Pol II occupancy upon BRD4 loss using ChIP-Rx (GSE158965, [7]). Interestingly, Pol II-normalized PAF1 data showed similar trends as 3′-RNA processing factors.

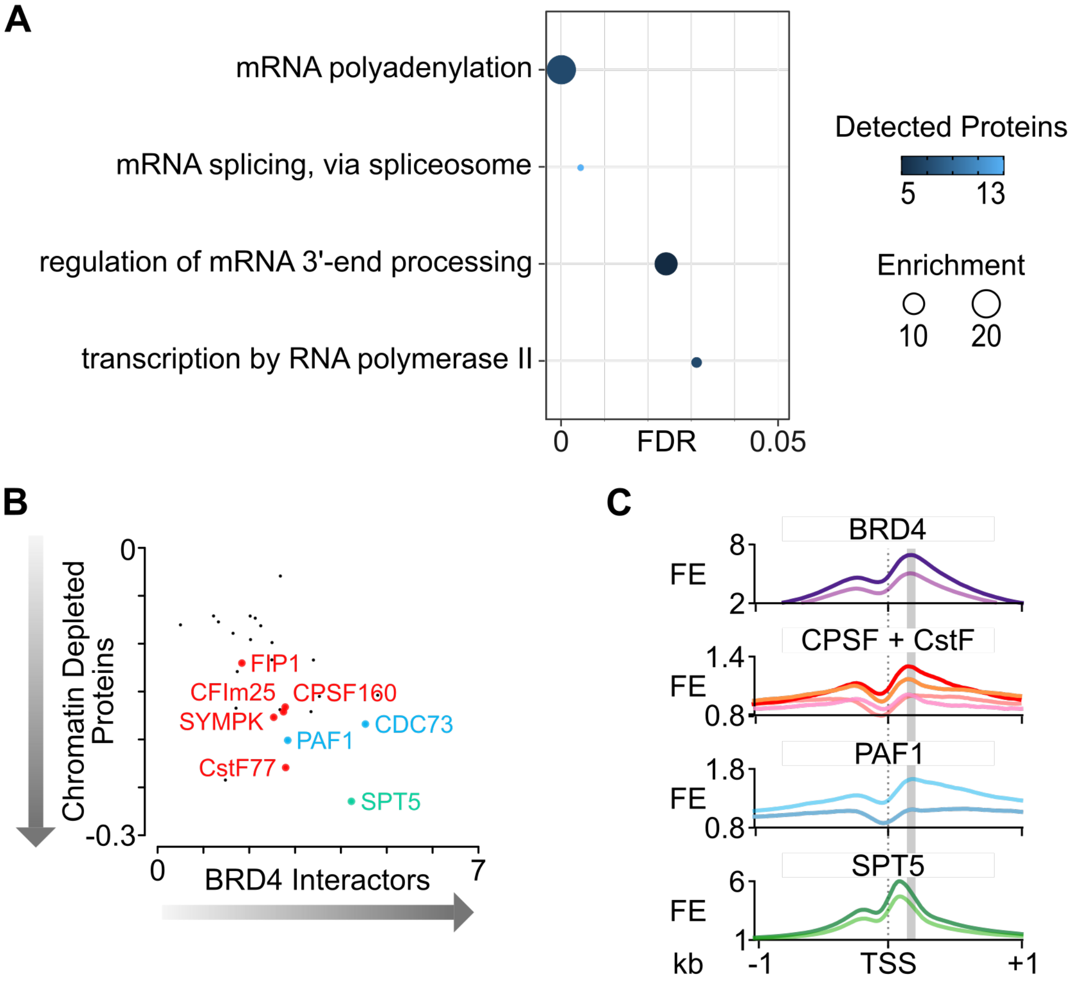


Figure 15.7: Functional Interactions of BRD4. (A-B) Depicted are factors of the core-interactome (n=29) which consists of significant BRD4 interactors (MS-IP, fold change (log2) > 0, FDR > 0.05) that were immediately displaced from the chromatin upon BRD4 degradation (chromatin-MS, fold change (log2) < 0, p-value < 0.05). **(A)** Significant (FDR < 0.05) GO terms of the core-interactome. GO sub-terms were excluded for visualization. **(B)** Scatter-plot highlights top-ranking BRD4 core interactors. The x-axis shows fold changes (log2) from IP-MS experiments, and the y-axis depicts the fold changes (log2) measured by chromatin-MS. Highlighted are 3′-RNA processing factors (red), PAF (blue), and DSIF (green). **(C)** Meta-gene profiles of FE (Section 3.2.7) normalized ChIP-Rx profiles. Presented are occupancy levels for BRD4, CPSF (FIP1, CSTF73), CstF (CstF64), DSIF (SPT5) and PAF (PAF1) at 5′ regions of actively transcribed genes (TSS +/- 1 kb). A gray box marks the peak occupancy location of BRD4 between 170-180 nucleotides downstream of the TSS.

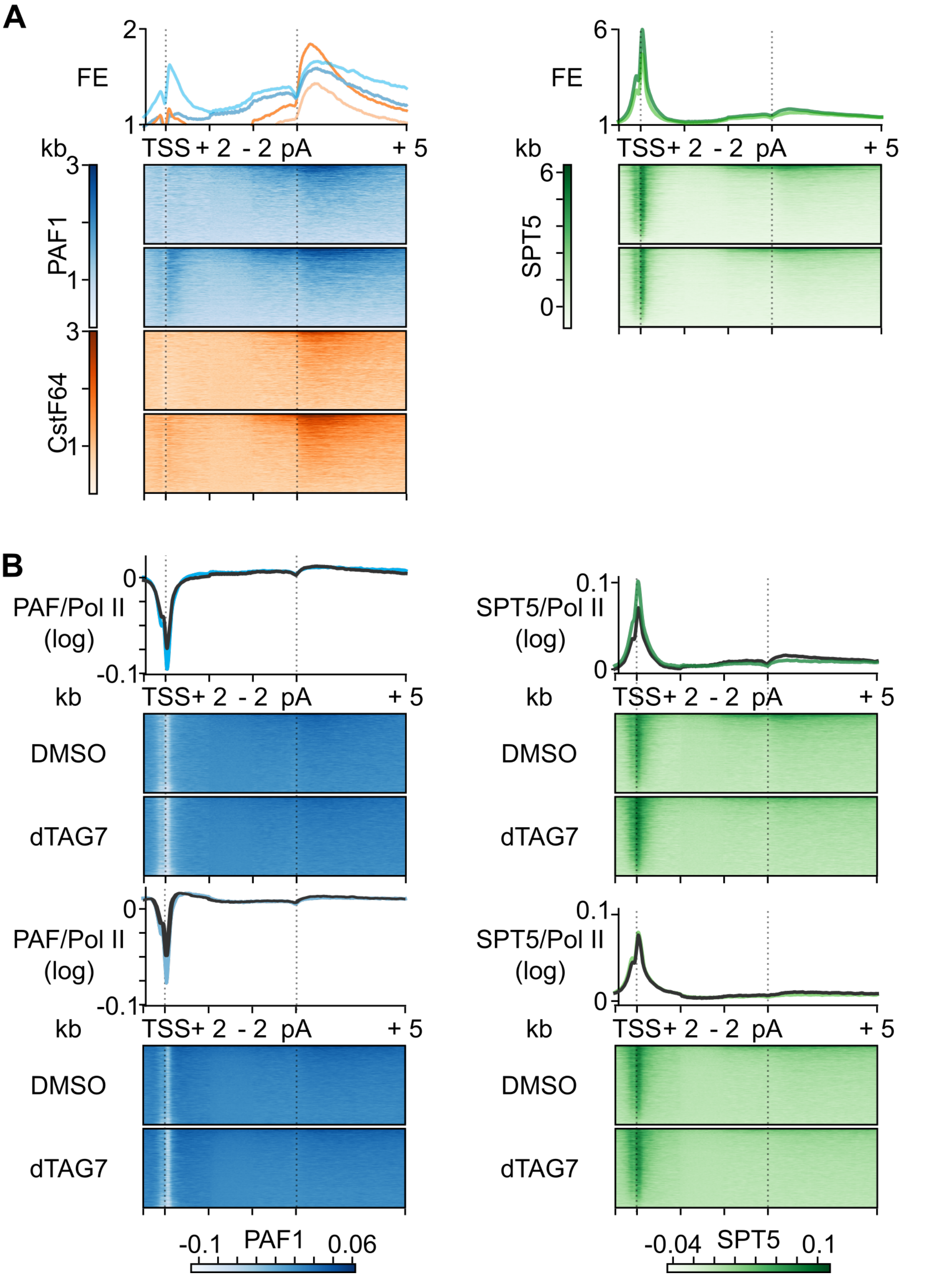


Figure 15.8: PAF1 Recruitment Defect. The figure presents ChIP-Rx data at actively transcribed non-overlapping genes (TSS to pA site + 5 kb, minimum gene length 6 kb) in human K562 dTAG-BRD4 cells. Meta-gene profiles of **(A)** FE-normalized (Section 3.2.7) PAF1, CstF-64, and SPT5 at genes (DMSO, n=7,331), and **(B)** Pol II-normalized SPT5 and PAF1 occupancy levels at genes after two hours of BRD4 degradation (dTAG7) and for the control (DMSO).

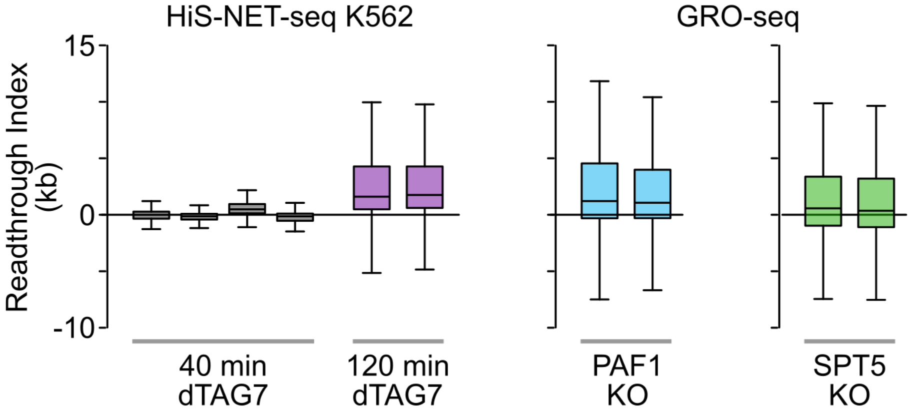


Figure 15.9: BRD4 Interactors Contribute to Transcriptional Defects. Boxplot quantification of RTI values calculated for indicated Pol II profiling methods and treatments using two replicates if not stated otherwise (four replicates 40 min dTAG7: n=9,581-9,619; 120 min dTAG7: n=9,608-9,646; PAF1 KO: n=3,782-4,723; SPT5 KO: n=9,085-9,269). Performed were PAF1 [24] and SPT5 [62] depletions in human HCT116 and *mouse primary activated splenic B lymphocytes*.

PAF1 was depleted from the proposed 5′ control region, suggesting that the recruitment was perturbed (Figure 15.8B and A.19C). In contrast, a relative increase was detected for SPT5 at the 5′ control region (Figure 15.8B and A.19D). SPT5 and PAF1 were slightly depleted from the 3′ region (Figure 15.8B, A.19C and A.19D).

Concerning the collected evidence, the question arose whether PAF1 and SPT5 could induce readthrough transcription, independent of BRD4 perturbations. Published Pol II occupancy data for PAF1 or SPT5 knockdown experiments were collected and reanalyzed to address this question (GSE70408 [24] and GSE132029 [62]). Strikingly, the RTI value calculations for both knockout data sets revealed Pol II readthrough transcription that escaped detection in the original publications (Figure 15.9). Pol II readthrough transcription induced by the PAF1 knockout (median RTI = 1.2 kb) was more pronounced than readthrough transcription induced by SPT5 loss (median RTI = 0.5 kb). The findings suggest that the 3′ defects were partially mediated through the BRD4 interactors PAF1 and SPT5.

15.4 POL II REGULATION AT ENHANCER REGIONS

15.4.1 *BRD4 Binds Transcribed Enhancer Regions*

BRD4 binding was a general feature of actively transcribed genes with peak occupancy profiles at promoter-proximal regions (Figures 15.7C and 15.5A). However, only 59% of all detected BRD4 binding sites overlapped with a promoter or promoter-proximal region (Figure 15.10A).

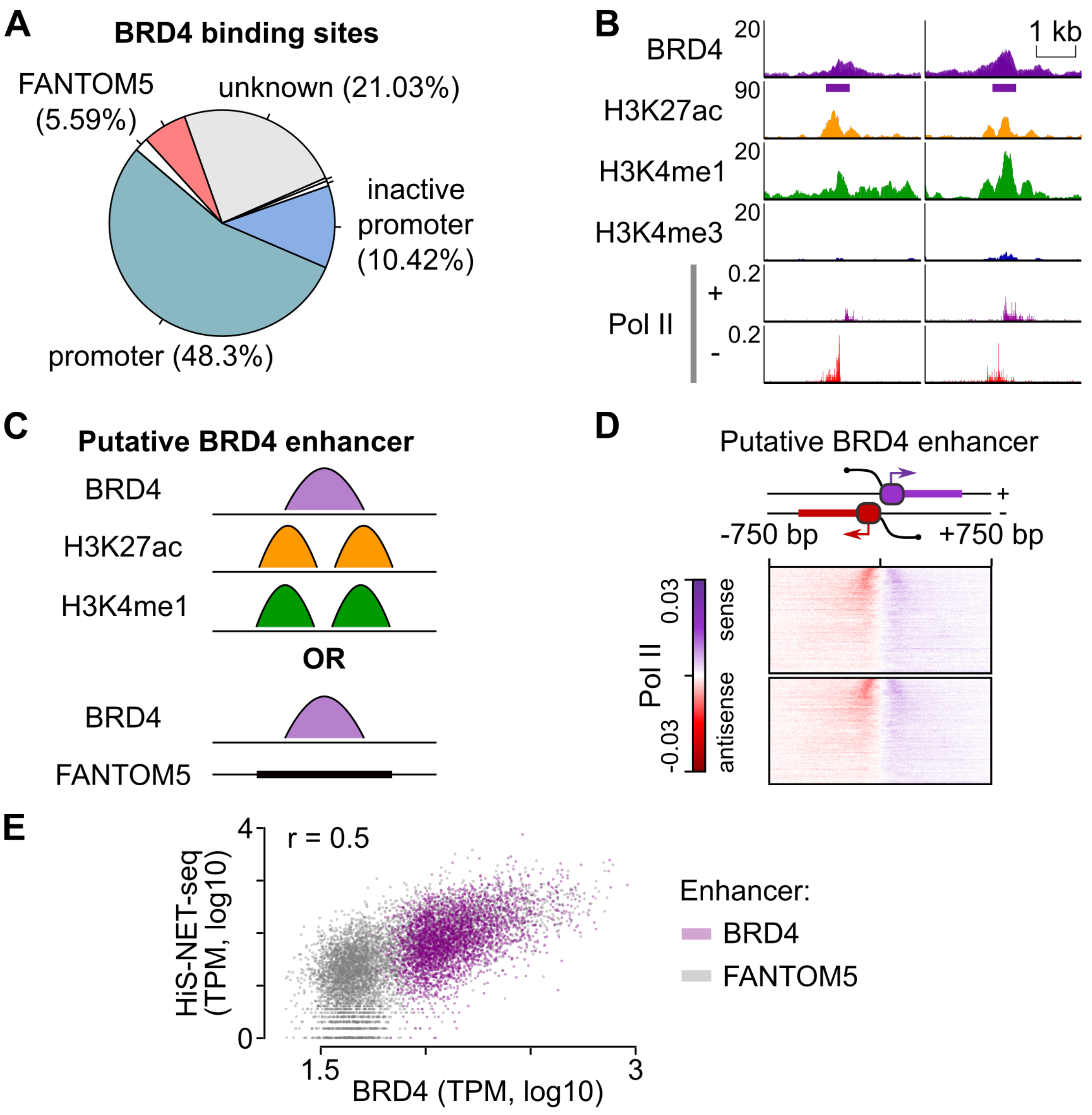


Figure 15.10: BRD4-associated Enhancer Regions. (A) The pie chart shows BRD4 consensus peaks (Section 14.1.2). Classification considers annotated GENCODE [67] promoter and FANTOM5 [36] enhancer regions. BRD4 binding sites at RNA Polymerase I (1.77%), sn/snoRNAs (0.4%), and *micro RNAs* (0.64%) are not labeled. **(B)** Two putative extragenic BRD4 enhancers (chr1:31,171,044 and chr6:37,187,434) overlap H3K27ac, H3K4me1, and bi-directional Pol II transcription but no FANTOM5 enhancer annotation. **(C)** The schematic visualizes the definition of putative BRD4 enhancers, including a BRD4 peak co-localizing with H3K27ac and H3K4me1 or a FANTOM5 annotated enhancer region (Section 6.1.2). **(D)** Heatmaps show bidirectional Pol II occupancy measured by two replicate measurements of HiS-NET-seq at putative BRD4 enhancers. The enhancer center selects the position that maximizes the Pol II signal downstream on the positive strand and upstream on the negative strand. **(E)** Correlation between mean TPM values (log10) of two replicate measurements from HiS-NET-seq and BRD4 (*Pearson's* correlation coefficient: r=0.5). Depicted are FANTOM5 enhancer (n=6,313, grey) and putative BRD4 enhancer (n=3,404, purple).

The tandem bromodomains of BET proteins bind to acetylated histones, such as H3K27ac [60, 255]. The correlation between BRD4 binding and H3K27ac at BRD4 binding sites confirmed this general presumption (r = 0.57-0.6, Figure A.20A). Additionally, BRD4 binds acetylated regulatory regions known to enhance gene expression of target genes, referred to as enhancers [45, 64, 260] (Section 2.1). For this reason, the remaining BRD4 binding sites were compared to annotated enhancer regions listed by FANTOM5 in the K562 cell line [36]. Although BRD4 overlapped to a certain extent with those regions (5.6%, Figure 15.10A), a considerable fraction of BRD4 binding sites remained undefined. Visual inspections of those regions revealed bidirectionally transcribed loci that harbored classical features of enhancers, such as H3K27ac, H3K4me1, and a lack of H3K4me3 (Figure 15.10B). H3K4me3 is a marker associated with promoter regions [44, 84]. Due to the conservative annotation of FANTOM5, a data-driven approach was applied to identify putative enhancer regions in human K562 cells. In this study, putative BRD4-associated enhancers were defined as genomic loci with BRD4 binding sites that co-localized with

- H3K27ac and H3K4me1 peaks or
- FANTOM5 annotated enhancer regions (Figure 15.10C).

BRD4 binding sites near annotated active or inactive promoter regions (+/-100 kb) were excluded. This approach identified 4,308 putative BRD4 enhancer regions with high H3K27ac and H3K4me1 levels (Figure A.20B). Consistent with the current knowledge, these regions lack H3K4me3 compared to BRD4 binding sites at promoter regions (Figure A.20B). The BRD4 abundance was similar between enhancers localized within genes and those in the extragenic areas but less pronounced than promoter-associated binding sites (Figure A.20C). Furthermore, a prominent feature of the putative enhancer regions was the bi-directionally transcribed Pol II (Figure 15.10D and Figure A.20D), which generally correlated with BRD4 binding (Figure 15.10E). The presented results highlight that BRD4 binding correlates with the transcriptional activity of putative enhancer regions.

15.4.2 *Pol II Elongation Control at Enhancer Regions*

The next question was whether BRD4 loss impacted the transcriptional activity at enhancer regions considering the profound effect of BRD4 degradation on productive gene transcription. The investigations focused on changes at previously defined putative BRD4 enhancers measured by HiS-NET-seq. A region 2 kb upstream and downstream of the enhancer center was considered for the corresponding DPO analysis. HiS-NET-seq identified significant Pol II reductions at 15% and 69% of the enhancers upon 40 and 120 minutes of treatment.

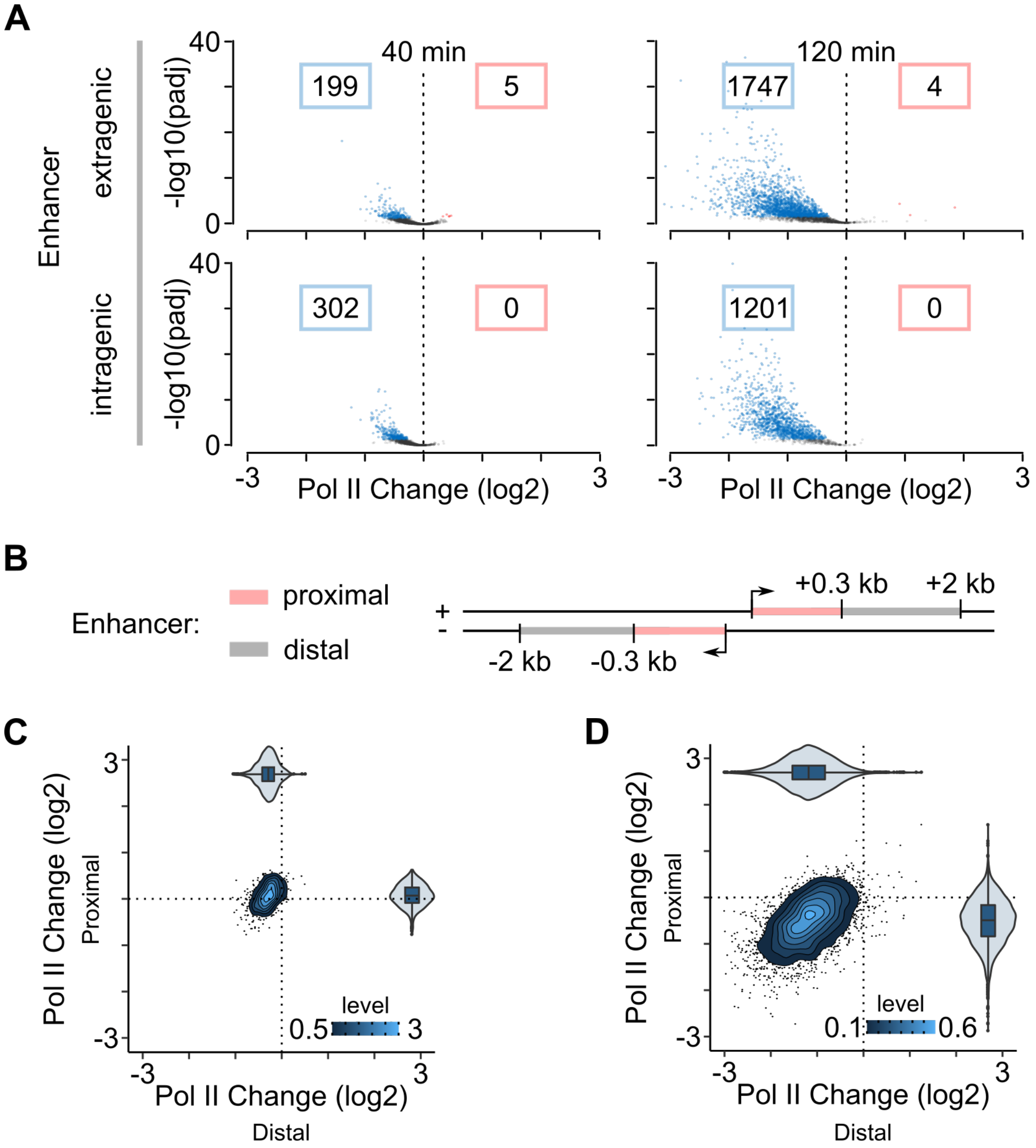


Figure 15.11: Disruption of Productive Enhancer Transcription. The figure presents analyses upon BRD4 degradation in human K562 dTAG-BRD4 cells measured by HiS-NET-seq **(A)** Pol II occupancy changes (log2) upon indicated treatment times (40 min: n=1,983, n=1,361; 120 min: n=2,814, n=1,478). Significant occupancy changes (padj < 0.05) are labeled in blue and red. **(B)** Schematic view of proximal and distal enhancer regions. **(C-D)** Pol II occupancy changes at enhancer proximal (y-axis) and distal regions (x-axis) upon **(C)** 40 minutes (n=793, four replicate measurements) and **(D)** 120 minutes (n=3,751) of BRD4-specific degradation.

Interestingly, the reduction was similar in intra- and extragenic enhancer regions (Figure 15.11A) and was visible for individual examples (Figure A.21A). This observation at extragenic enhancer regions proved that Pol II reductions at gene-body regions did not indirectly cause the observed decreases. The less sensitive SI-NET-seq method did not detect these changes (Figure A.21B).

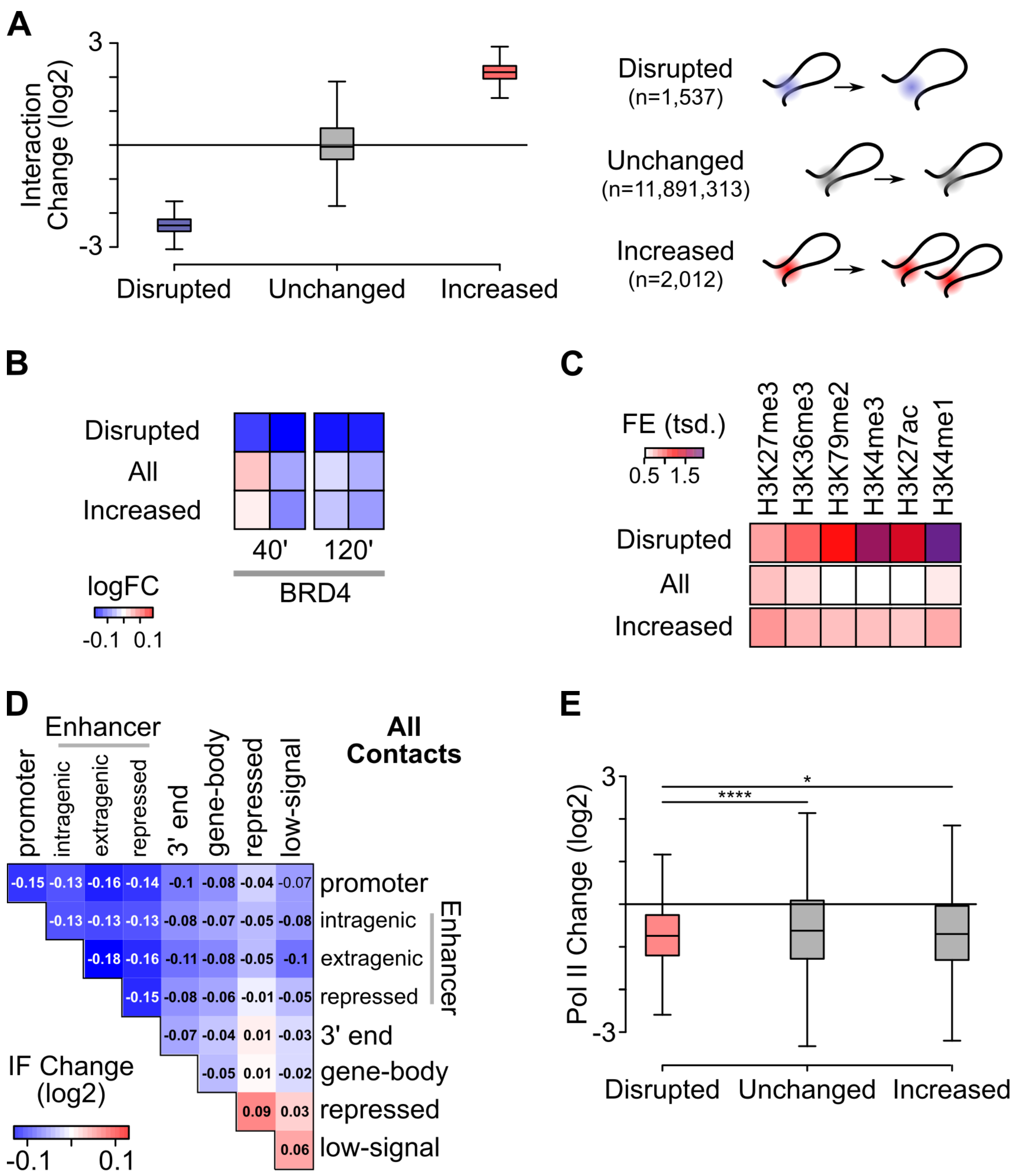
What caused the Pol II occupancy reduction at enhancers? Different publications suggest remarkable similarities between transcription at enhancers and protein-coding genes [36, 58, 85, 110], which opened the question of whether BRD4 might regulate Pol II elongation control at genes and enhancers. Enhancer regions were segmented into proximal and distal areas based on this idea. Proximal regions started at the enhancer center and covered the regions 300 bp upstream and downstream at the corresponding strand (Figure 15.11B). The distal enhancer region covered the area between 300 bp and 2 kb away from the enhancer center. Pol II occupancy increased in the proximal enhancer region but decreased in the distal region after 40 minutes of BRD4 loss (Figure 15.11C). This trend was similar to initial observations at transcription units from genes (Figure 15.2D). After 120 minutes of treatment, the reduction covered proximal and distal regions of enhancers. However, the decrease was more pronounced in distal enhancer regions (Figure 15.11D).

In contrast to, for example, the transcription initiation factors TBP (ENCSR-000EHA [35]), the peak occupancy of BRD4 co-localized downstream of the transcription initiation site together with Pol II and the elongation factor PAF1 in the promoter-proximal and enhancer-proximal regions (Figure A.22). The results suggest a regulatory role of BRD4 in elongation control at some actively transcribed enhancers.

15.4.3 *Loss of BRD4 Disrupts Regulatory Interactions*

The time-dependent reduction of Pol II at enhancers opened the question of whether the loss of BRD4 also perturbed transcription initiation at enhancers after two hours of treatment. A high abundance near promoter and enhancer regions (Figure A.20C) suggests a potential stabilizing role of BRD4 at regulatory contacts.

Therefore, HiChIP [159] experiments upon 120 minutes of BRD4 degradation were performed (MRA10-MRA15, unpublished). HiChIP is a protein-centric chromatin conformation method, similar to ChIA-PET [68]. Those methods enrich DNA-DNA contacts associated with a specific protein or modification of interest. This experiment investigated regulatory contacts between DNA regions with H3K27ac. The software packages HiC-Pro [212] and HiCcompare [220] were applied for data processing, quality checks, and comparative analysis between conditions at 10 kb resolution (Figure A.23A). In the corresponding analysis, significant reduction in the interaction frequency of 1,537 contacts were identified ($padj < 0.05$, Figure 15.12A).



	promoter	intragenic	extragenic	repressed	3' end	gene-body	repressed	low-signal
promoter	-0.15	-0.13	-0.16	-0.14	-0.1	-0.08	-0.04	-0.07
intragenic		-0.13	-0.13	-0.13	-0.08	-0.07	-0.05	-0.08
extragenic			-0.18	-0.16	-0.11	-0.08	-0.05	-0.1
repressed				-0.15	-0.08	-0.06	-0.01	-0.05
3' end					-0.07	-0.04	0.01	-0.03
gene-body						-0.05	0.01	-0.02
repressed							0.09	0.03
low-signal								0.06

Figure 15.12: Reduction of Regulatory Contacts upon BRD4 Loss. Presentation of measurements from human K562 dTAG-BRD4 cells upon 120 minutes of BRD4 degradation. **(A)** Interaction frequency changes (log2) at contacts with reduced (disrupted), increased or without changes (unchanged). Comparison was performed by HiCcompare [220] as described in Section 14.1.4. **(B-C)** Heatmaps show mean values from disrupted (n=2,959), increased (n=3,825) and unchanged (n=241,897) contact anchor regions. **(B)** Occupancy changes (log2) of BRD4 and **(C)** FE of indicated chromatin marks. **(D)** Mean interaction frequency changes (log2) at pairwise contacts between all indicated regions. **(E)** Pol II occupancy changes (log2) measured by HiS-NET-seq at disrupted (n=2959), increased (n=3825) and unchanged (n=241,897) contact anchor regions (one tailed *Wilcoxon* rank sum test *: p = 0.03; ****: p = 1.9e-10).

These contact losses occurred in regions with the highest reductions of BRD4 after 40 and 120 minutes of treatment (Figure 15.12B). Different publicly available epigenomic datasets for K562 cells were analyzed from ENCODE [35] (ENCSR000AKP, ENCSR000EWB, ENCSR000AKR, ENCSR000APD, ENCSR000EWC, ENCSR000EWA) to gain a deeper understanding which types of contacts were affected. Disrupted contacts were highly enriched in chromatin marks associated with actively transcribed genes and enhancers (Figure 15.12C). Next, the epigenomic datasets and chromHMM [56] were applied to gain a genome segmentation with comparable resolution to HiChIP results. The software, which implements an unsupervised Hidden Markov Model, assigns a state to each genome segment associated with a different combination of chromatin marks. For this purpose, the genome was divided into ten states that were manually annotated as

- *promoter,*
- *enhancer (intragenic, extragenic, and repressed),*
- *3′ gene ends,*
- *gene-body,*
- *repressed*, or
- *low-signal states* (Figure A.23B and Section 14.1.4).

In contrast to contacts with increased interaction frequency, disrupted contacts were enriched in regulatory interactions between promoter-promoter, enhancer-promoter, and enhancer-enhancer regions (Figure A.23C). The global trend also showed a general loss of interactions among these regulatory interactions (Figure 15.12D). Furthermore, Pol II occupancy loss in regions with significantly fewer interactions exceeded the global trend (Figure 15.12E). However, the resolution of the applied method does not distinguish between the promoter and promoter-proximal regions. In conclusion, the integrated analysis revealed the disruption of regulatory interaction upon BRD4 loss, which correlated with Pol II occupancy reductions.

16 DISCUSSION

Research over the last decades primarily focused on Pol II initiation mechanisms which occurred at the first step of transcription [183]. In recent years, the methods to study Pol II occupancy improved and revealed regulatory steps during early elongation [160, 187, 254]. These regulatory steps emerged as general [1] and rate-limiting in the expression of genes [102, 141]. This study investigated post-initiation regulatory mechanisms of Pol II that emerged upon rapid BRD4 and pan-BET protein degradation using an integrated multi-omics approach. Results show that BRD4 underlies a general 5'-regulatory hub that controls productive transcription elongation and primes the Pol II elongation complex for 3'-RNA processing and termination. Furthermore, this work suggests that BRD4 actively regulates enhancer transcription with a potentially similar mechanism as observed in promoter-proximal gene regions.

Rapid protein degradation within two hours or less enables the identification of direct protein functions independent of widespread cellular compensation and adaptation effects. The challenge with short-time treatments is the selection of appropriate methods which detect transcriptional changes that occurred in the respective periods. Efforts to identify the consequences of BRD4 loss on mature RNA levels revealed only limited insights into the overall impact of this treatment after two hours (Figure A.5D). Changing the focus to newly synthesized or nascent RNAs, using nuclei-RNA-seq and HiS-NET-seq, revealed the widespread consequences of BRD4 degradation. Both assays showed a global reduction of transcript levels using reference-based normalization strategies (Figure A.6B, 15.1A and 15.1B). These observations were consistent with results from Muhar et al., 2018 [158], showing minor changes at total RNA levels after 90 minutes of BRD4 degradation, in contrast to global reductions revealed using specialized methods that capture newly synthesized RNAs.

The region-specific analysis revealed an impaired pause release where Pol II accumulated in the promoter-proximal regions, with inefficient release into productive elongation over gene-body regions (Figures 15.2D and 15.2E). This observation suggested an essential function of BRD4 during Pol II pause release, which was recently also verified by other BRD4-specific degradation experiments [158, 260].

Interestingly, Pol II accumulation and reduction in both regions were initially balanced, consistent with the deregulated Pol II pausing release hypothesis (Figure 15.2D).

After two hours of BRD4 loss, Pol II remained at the same levels at promoter-proximal regions, whereas elongating Pol II further decreased (Figure 15.2E). This observation suggested that BRD4 potentially affected directly or indirectly other aspects of Pol II transcription, such as initiation, premature termination, elongation rate, or processivity, which could explain the imbalance of changes between both regions.

Pan-BET ablation compared to BRD4-selective degradation showed a general decrease of Pol II independent of the respective region (Figures A.8D and A.8E). However, reduction of Pol II was not uniformly distributed but had a more substantial impact on productive elongation than promoter-proximal pausing. More evident than for BRD4-specific degradation, these results imply a distinct role of BRD2 and BRD3 in the 5′ control region. This hypothesis is consistent with the substantial Pol II depletion at 5′ gene regions upon BRD2 and BRD3 specific degradation experiments observed in another study [260]. Clarifying the protein-specific functions could be an interesting subject for future investigations. In this context, the next step could be an integrated analysis of HiS-NET-seq and TT-seq data to identify potential changes in elongation rates and Pol II processivity [262]. However, new methods are required to identify genome-wide initiation or premature termination defects.

The proposed mechanism for the BRD4-dependent release of promoter-proximal paused Pol II was the recruitment of P-TEFb [98, 173, 252]. Although sub-units of this complex (CDK9 and *cyclin T1*) interacted with BRD4 (Figure A.18), neither pan-BET nor BRD4-specific degradation decreased their abundance at the chromatin (Figure A.13A and A.13B). This observation was consistent with other studies using either pan-BET [249] or BRD4-specific degradation strategies [158, 260], excluding the BRD4-dependent recruitment hypothesis of P-TEFb.

The integrative analysis of proteomic datasets identified other potential candidates. This study focused on candidates from the BRD4 core-interactome, which consisted of BRD4 interactors immediately displaced from the chromatin upon BRD4 loss (Figure 15.7B). The DSIF (SPT5) and PAF complexes (PAF1, CDC73) were top-ranking candidates. Interactions were validated in co-precipitation experiments either in this study (Figure A.19B) or by others [249, 253]. Another interesting candidate, SPT6, was depleted from the chromatin (Figures A.13A and A.13B) and co-precipitated with BRD4 [7] but was not detected as interactor in the mass spectrometry experiment (Figure A.18).

All these factors are components of the transcription elongation complex [232, 233] with different putative roles during transcription. Briefly, the PAF complex is associated with Pol II pause release efficiency [253] and velocity [90, 262], whereas SPT5 [62, 93] and SPT6 [165, 262] are primarily important for Pol II processivity. ChIP-Rx experiments investigated the abundance of these factors upon BRD4 loss.

However, their strong association with Pol II required normalization to remove changes mirroring the general Pol II occupancy trend. The following observations likely contributed to the observed BRD4 phenotype.

1. PAF1 co-localized with BRD4 in the promoter-proximal region (Figure 15.7C). This 5′ control region failed to recruit PAF1 to the Pol II elongation complex upon BRD4 loss (Figure 15.8B). The failed recruitment of the PAF complex could explain increased Pol II levels in the promoter-proximal region due to PAF's function in Pol II pause release [253, 262]. However, a previously observed decrease in Pol II velocity [90, 262] would increase gene-body Pol II occupancy levels. Therefore, loss of the PAF complex can not explain the observed Pol II reduction in the gene-body regions observed upon BRD4 loss.

2. SPT5 peak occupancy was located upstream of the 5′ control region and showed no recruitment defect in this region (Figures 15.8B and 15.7C). The depletion of SPT5 in the regions downstream of the 5′ control point (Figures 15.8B and A.19D) potentially decreased Pol II processivity, which could contribute to the observed reduction of Pol II in the gene-body region [62, 93].

3. SPT6 did not appear in the BRD4 core-interactome. Therefore, subsequent experiments did not consider SPT6. However, recent publications showing SPT6's role during Pol II processivity [262] and the observed interactions between SPT6 and BRD4 [7] hint toward a potential link. If SPT6 recruitment is BRD4-dependent is currently unknown.

Overall, these findings suggested that BRD4 was required to assemble a functional Pol II complex capable of productive elongation.

Unexpectedly, pan-BET and BRD4 degradation directly impacted 3′-RNA processing and termination. The main consequence of this defect was the widespread Pol II readthrough transcription downstream of the termination zone that led to uncontrolled activation and RNA processing of some transcription units (Figures 15.3E, A.10A and A.10C). It was unclear if those newly activated units were functional transcripts that could undergo translation.

The transcriptional readthrough index, developed in this study, identified termination defects genome-wide for individual genes (Figure 15.3B). However, the index systematically underestimates the impact on lowly expressed genes and in gene hubs where many actively transcribed genes occurred with limited distance to each other [7]. Despite these limitations, termination defects occurred primarily at polyA signal-containing genes (Figure 15.3D). Notable, degradation of pan-BET proteins increased the impact on termination significantly (Figure 15.3E), which suggests a collaborating role of BET proteins.

The proposed index successfully identified and compared readthrough genes between different treatments and can be applied in future studies to study HiS-NET-seq or other Pol II profiling data, including PRO-seq and mNET-seq.

Next, 3′-RNA cleavage defects and extended transcripts were identified upon pan-BET protein degradation using long-read sequencing of nascent RNA units (Figures 15.4C and 15.4D). Consistent with previous studies and models [53, 169], the results verified 3′-RNA cleavage as an essential step required to trigger processes that release Pol II from the DNA template, such as allosteric changes, an entry side for the XRN2-dependent termination, or both.

Nanopore sequencing was a powerful approach to identify 3′-RNA cleavage defects and extended transcripts. However, the low enrichment of valid full-length transcripts (Section 14.1.3) was an explicit limitation. To obtain quantitative data for the comparisons, libraries were not PCR amplified. PCR amplification is essential for enriching full-length transcripts with correct primer orientation. Addressing this problem in future experiments requires alternative approaches, such as direct RNA sequencing [51] or PCR amplified libraries combined with UMI sequences. The latter is challenging due to the high error rates of the nanopore sequencing technology [2] but was applied recently in a pioneer study [108]. This study performed no additional nascONT-seq experiments due to the listed problems, including no BRD4-specific degradation experiment. Whether BRD4 specific degradation affected 3′-RNA cleavage globally remained unaddressed. However, at individual genes [7], 3′-RNA cleavage defects were verified upon BRD4 loss using an RT-qPCR-based assay.

A potential underlying mechanism was the perturbed recruitment of some CPSF and CstF factors to the 5′ control region. There were several lines of evidence that supported this hypothesis. First, 3′-RNA processing factors accumulate in 5′ regions and co-localize with BRD4 (Figures 15.6B and 15.7C). Other studies also observed the early recruitment of these factors [47, 79, 105]. Second, CPSF and CstF factors interacted with BRD4 (Figure A.18) and dissociated from the chromatin upon BRD4 loss (Figure A.13A and 15.6C). Finally, the deprivation of 3′-RNA processing factors at the 5′ control region was more pronounced than the general BRD4 induced reduction of Pol II (Figure A.15-A.17). These results suggested BRD4-dependent recruitment of CPSF and CstF factors during an early Pol II elongation phase.

Furthermore, the BRD4 interactors SPT5 [41, 93, 149, 172] and SPT6 [165] are associated with termination defects and potentially contributed to the observed defect.

Of particular interest for this study was the contribution of the PAF complex, which showed significant recruitment defects to the 5′ control point upon BRD4 loss (Figure 15.8B). Interestingly, PAF1 interacted and co-localized with 3′-RNA processing factors across transcription units with peak occupancy at 3′ regions (Figure 15.8A, A.19A and A.19B).

PAF1 loss induced readthrough transcription in another cellular system (Figure 15.9) and appears essential for the recruitment of CPSF and CstF factors at individual gene examples [201]. However, the exact role of the PAF complex in 3′ end processing remains unclear and requires further genome-wide studies.

Additionally, this study described BRD4-binding at transcribed enhancer regions (Figures 15.10A and 15.10B) as presented in many previous studies [45, 64, 260]. Subsequently, putative BRD4 enhancers were defined and used for analyses (Figure 15.10C and A.20B). The most prominent feature of these BRD4-associated enhancers was the high transcriptional activity compared to FANTOM5 enhancers (Figure 15.10E).

Unexpectedly, BRD4 loss significantly reduced enhancer transcription at most putative BRD4 enhancers (Figure 15.11A). SI-NET-seq, which significantly lacked Pol II coverage in this region, detected no significant changes (Figure A.21B). Previous studies showed that BET proteins regulate enhancer transcription [45, 249, 260]. However, Zheng et al., 2021 [260] ascribed the regulatory function of BET proteins to BRD2 and did not detect changes upon BRD4 degradation. Interestingly, their degradation strategy had no impact on BRD4's short protein isoform, which was the main difference compared to BRD4 loss in this study. Differences between both studies suggested isoform-specific functions of the short BRD4 isoform, as also suggested by others [81]. Of particular interest would be a direct comparison of enhancer transcription between pan-BET degradation, BRD4-specific degradation, and the specific loss of BRD4's short protein isoform.

How does BRD4 regulate enhancer transcription? Some studies in recent years suggested that elongation control at enhancers might be similar to promoters, implied by Pol II enhancer pausing [39, 40, 63, 85]. Different observations supported this idea. First, Pol II accumulated at the sense and antisense strand close to the enhancer center, similar to promoter-proximal paused Pol II at genes (Figure 15.10D and 7.5B). Second, BRD4 loss impaired Pol II pause release at enhancer regions (Figure 15.11C). A short treatment time of 40 minutes revealed increased levels of Pol II proximal to the initiation site of actively transcribed enhancers. At the same time, elongation in distal enhancer regions decreased to a similar extent. Third, BRD4 co-localized with PAF1 and Pol II at enhancer control regions (Figure A.22). Unfortunately, the signal coverage of Pol II and PAF ChIP-Rx data was insufficient at enhancers to reliably answer whether BRD4 degradation caused an assembly defect of a competent elongation complex, similar to observations at promoter-proximal regions.

In contrast, two hours of BRD4 loss revealed Pol II reduction at proximal and distal enhancer sub-regions. However, the distal enhancer regions were more substantially impacted (Figure 15.11D).

The overall reduction at enhancer regions suggested that BRD4 had an additional function in Pol II initiation at enhancers besides elongation control. The observed interaction frequency changes of 3D regulatory contacts among promoters and enhancers supported this hypothesis (Figure 15.12D). Interestingly, the Pol II reduction in the identified regions was more potent than the global trend (Figure 15.12E), which was not observed in a previous BRD4 knockout experiment [128].

Overall the results remain correlative, as the selected methods could not answer whether

- reduced 3D interactions caused the reduction of Pol II transcription, or
- reduced Pol II transcription caused the reduction of 3D interactions.

The results generally argued against a global function of BRD4 in stabilizing enhancer-promoter interactions, which decreased at a subset of 1,537 interactions and not genome-wide. Furthermore, the selective response potentially explained why previous studies did not detect changes in enhancer-promoter interactions upon BRD4 loss at selected loci [45]. BRD4's proposed function to form liquid-like condensates [17, 81, 203], which conjunct the transcription apparatus, could mechanistically explain the loss of 3D interactions and transcription. Future experiments could address if and how the formation of liquid-like condensates contributed to the observed defects, for example, using complementary microscopy techniques.

17 CONCLUSION

This study aimed to improve the general understanding of Pol II post-initiation regulatory mechanisms and develop new approaches to uncover them. In this context, this work primarily focused on BRD4-dependent regulatory steps by combining multi-omics technologies and rapid BRD4 protein degradation. Strikingly, the obtained results validated the proposed protein function during early Pol II elongation and discovered unknown BRD4-dependent regulatory steps essential for Pol II transcription during termination and enhancer transcription.

New computational and experimental approaches were co-developed to study Pol II and nascent RNA transcripts quantitatively between conditions, which was essential in uncovering the diverse roles of BRD4. The following paragraphs summarize the significant discoveries of this study.

First, this work strengthened the evidence of a 5′ elongation control point that occurs after initiation and assembles the Pol II elongation complex by recruiting elongation factors to this region. BRD4-dependent regulation of Pol II pause release was identified by the global collapse of productive elongation, which co-occurred with the global increase in promoter-proximal paused Pol II upon BRD4 loss.

New NET-seq-based protocol variants identified the global changes by adding mouse control cells for normalization. This work showed that sequencing data from human and mouse cells with similar genomes could be computationally distinguished, allowing reference-based normalization to identify uniform changes between conditions. The integration of mouse control cells and the adjustment of computational processing and normalization steps allowed the discovery of uniform changes in different gene regions upon treatment.

Second, the study revealed an unexpected role of BRD4 in 3′-RNA processing and termination. BET proteins, specifically BRD4, were linked to termination control for the first time despite being widely studied.

Often transcription studies neglect potential effects on transcription termination. Screenings for initiation or elongation defects focus on well-annotated gene regions, starting at the TSS and ending at the polyA site. Standard reference annotations provide no information on Pol II transcription termination zones. In the past, most studies using termination-related indices considered changes only in fixed regions relative to annotated polyA sites [8, 14, 18, 109].

However, this and other studies showed that Pol II termination occurs in highly variable gene-specific termination zones in humans, which could be distant from the last annotated polyA site.

For this reason, the newly designed readthrough index identifies Pol II distribution changes at variable gene-specific termination zones. The application of the index successfully identified genome-wide termination defects that escaped detection in previous studies caused by knockdowns of BET proteins, BRD4, SPT5, and PAF1.

Third, the data suggest the failed recruitment of 3′-RNA processing factors to the BRD4-dependent 5′ control region. This recruitment defect is a possible explanation for 3′-RNA cleavage defects, extended transcripts, and Pol II readthrough transcription. The defects were identified using the developed *3′-RNA cleavage efficiency* test and the nascONT-seq approach. The latter performed ONT's long-read sequencing. Combining both approaches identified significant reductions of cleavage efficiency upon pan-BET loss and extended transcripts.

Unfortunately, the quantitative but amplification-free ONT-seq technology used for nascONT-seq provided only insufficient coverage of full-length transcripts. Future experiments should consider other published methods [51] that avoid these problems. However, the *3′-RNA cleavage efficiency* test can be universally applied to all transcriptomic HTS assays. Furthermore, the identified link between 5′ elongation control and the recruitment of 3′-RNA processing factors established an exciting concept to study in the future. More follow-up studies could help identify the implications of early 3′-RNA processing factor recruitment on co-transcriptional processing and polyA site selection.

Fourth, this study proposes an elongation control region at actively transcribed enhancers similar to promoter-proximal pausing and release at genes. The Pol II pause release defect at enhancer regions was identified immediately after BRD4 loss. The reduction of Po II in distal enhancer regions co-occurred with the increase in enhancer proximal regions. An essential step to reveal this elongation control point at enhancers was the significant improvement of Pol II coverage by the HiS-NET-seq method. This study identified critical limitations of the human NET-seq approach that were addressed in optimization efforts leading to this new experimental approach.

How and if *enhancer RNAs* contribute to increased target gene expression remains unclear, but the Pol II elongation regulation by BRD4 suggested a direct biological function. This study adds to a few publications that propose post-initiation Pol II regulation at enhancers similar to gene regions. The proposed elongation control at enhancers raises fundamental questions about the general enhancer transcription function. Additional studies are required to understand the potential role of pausing at enhancers and their contribution to gene regulation.

Overall, this study optimized existing and developed new methods to study BRD4-dependent post-initiation regulatory mechanisms of Pol II. The general steps and methods applied in this work can be transferred to other projects and contribute to identifying Pol II deregulation in other cellular systems or experimental designs. This study proposed new exciting concepts of post-initiation regulatory Pol II mechanisms at genes and enhancers, opening many new questions for future studies.

18 OTHER PROJECT CONTRIBUTIONS

Part II of this study introduced the collaboration with the laboratory of Bruno Reversade, director of the *A*STARs Genome Institute of Singapore* and the *Institute of Molecular and Cell Biology*. His group investigates the genetic cause of a congenital syndrome of *osteogenesis imperfecta*. Section 10 describes the patient's phenotype and material collection resulting in polyA-enriched RNA-seq and SI-NET-seq data (GSE197118 and GSE197119, unpublished). Section 11.3 showed as a proof-of-concept the successful application of SI-NET-seq in this clinical case study and identified deregulated genes in patients (Figure 11.3C). The integrated analysis showed the deregulated pathways of the patients identified by both assays (Figure 11.3E), which corresponded to the general disease phenotype and mainly affected extracellular matrix and collagen-related pathways.

However, the genetic founder mutation that caused the syndrome remained unknown despite the following investigations. Homozygosity mapping from *single nucleotide polymorphism* genotyping data identified the Identical-by-Descent region that harbored the potential founder mutation. The area spanned 8.4 megabases on chromosome four with 39 possible candidate genes. *Whole-exome sequencing* reported no compelling recessive mutation in the coding regions of the respective regions, suggesting a mutation in the non-coding DNA.

Interestingly, the RNA-seq analysis uncovered one differentially expressed gene in the Identical-by-Descent region, the *transmembrane anterior posterior transformation 1* (*TAPT1*) gene (Figure A.24A). A previous study [224] reported homozygous mutations in *TAPT1* that caused a complex *osteochondrodysplasia* with clinical overlap to the patient's syndrome. Western blot experiments confirmed a complete loss of the TAPT1 protein in the patient's homozygous fibroblast cells (data not shown).

Because the SI-NET-seq analysis showed no significant Pol II occupancy change in this gene region, the mutation likely affected the processing of the RNA transcript or translation of the protein. Strikingly, the rMATS v3.1.0 [215] splicing analysis on polyA-enriched RNA-seq data revealed a significant alternative splice event in patients, leading to the exclusion of exon twelve in the *TAPT1* gene (Figure A.24B). Skipping of exon twelve led to a frameshift that created a premature stop codon, which likely resulted in *nonsense-mediated decay* (Figure A.24C).

Based on this observation, our collaborator identified the founder mutation segregating with the disease in the family. The disease-causing point mutation replaced guanine with an adenine, twenty-two nucleotides upstream of the 3′ splice site of exon twelve (c.1237-52 G>A, Figure A.25A).

The critical question remained: How could a deep intronic mutation cause an exon skipping event? A plausible hypothesis was that the mutation affected the recognition of the branchpoint, which is essential for the splicing process. Different prediction tools, including RNABPS [166], LaBranchoR [171], and BPP [256], were applied to the corresponding locus to address this hypothesis. Surprisingly, the *in silico* mutation experiment created a new competing branchpoint position seven nucleotides downstream of the predicted wildtype branchpoint (Figure A.25B). The spacial proximity suggested that the mutation either directly created a non-functional competing branchpoint or disrupted the branchpoint recognition by changing the DNA context. Unfortunately, this model derives exclusively from *in silico* tests. If the mutation disrupts the sensing of the branchpoint *in vivo* remains unclear. Future experiments could perform *crosslinking and immunoprecipitation followed by HTS experiments* to see if branchpoint-recognizing factors, for example, the *branchpoint bridging protein*, bind differentially to nascent RNA transcripts of patients.

A

SUPPLEMENTARY FIGURES

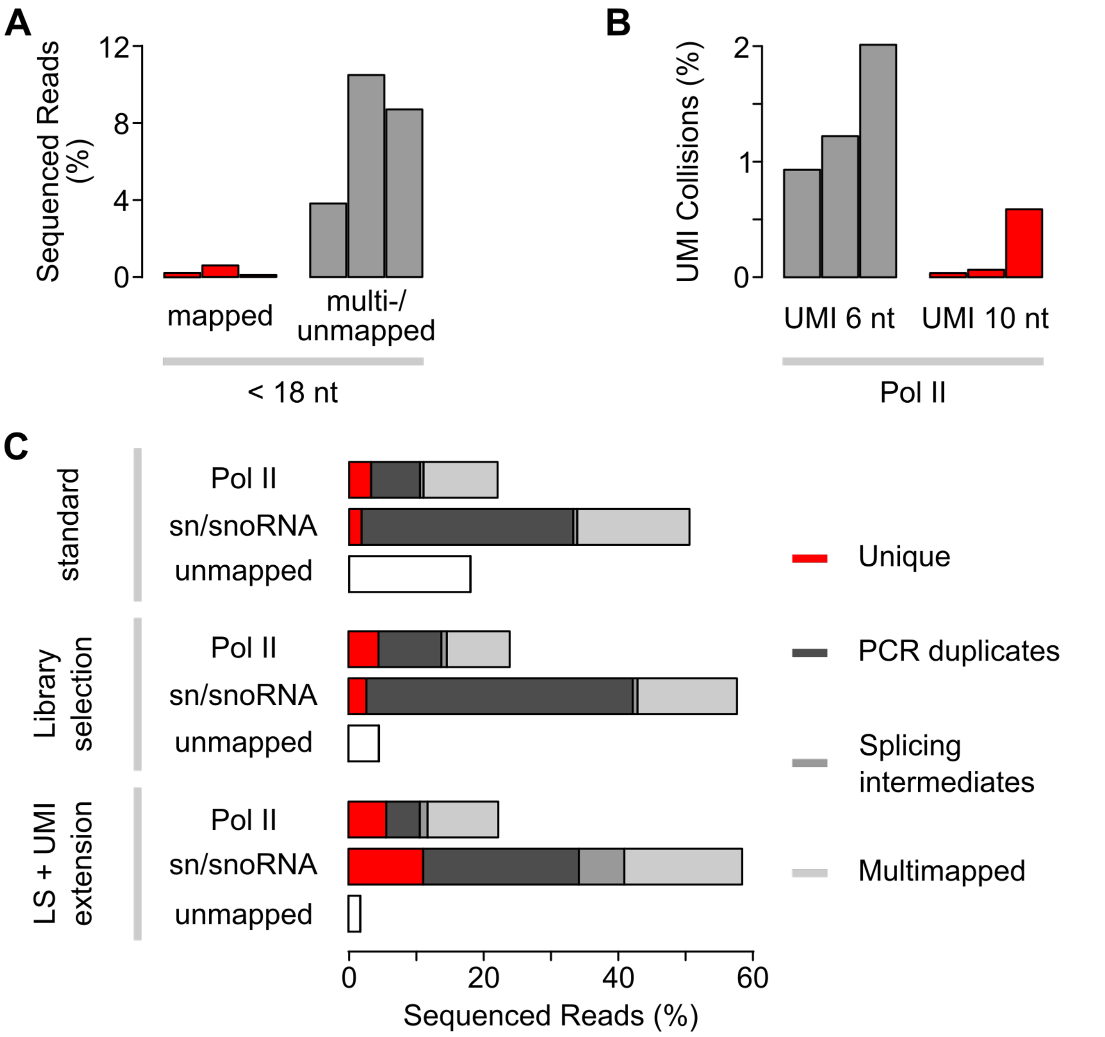


Figure A.1: Optimization of Human NET-seq. Related to Section 7.2. (A-B) Three human NET-seq samples derived from HeLa S3 and K562 show **(A)** the fraction of short sequencing reads aligned to the human reference genome (mapped) compared to reads without unique sequence alignment (unmapped) and **(B)** the expected number of UMI collisions among sequencing reads mapping to Pol II transcribed regions. **(C)** Fraction of sequencing reads mapping to Pol II transcribed regions (chromosomal and not listed in Table B.6), sn/snoRNA genes, or no locus in the human reference genome (unmapped). Sequencing reads are, if possible, further classified into uniquely mapped (red), PCR duplicates (dark grey), splicing intermediates (grey), or without unique mapping position (light grey).

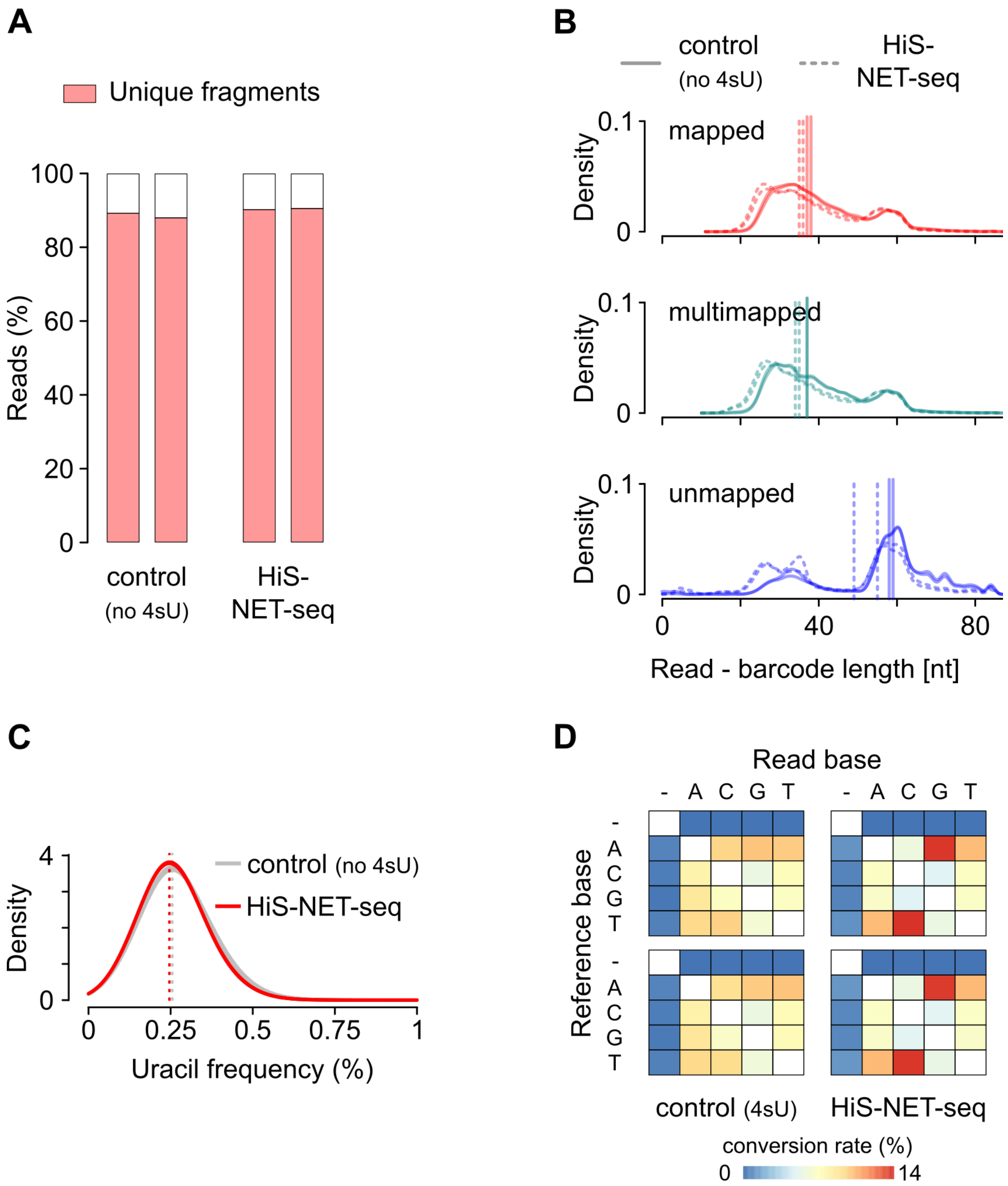


Figure A.2: Impact of 4sU Labeling on Purified RNA Libraries. Related to Section 7.3. The analysis includes HiS-NET-seq libraries with 10 minutes of 4sU labeling and the respective control experiment without metabolic labeling in K562. The comparison includes **(A)** library complexity, **(B)** length (excluding barcode length), **(C)** uracil frequency, and **(D)** conversion rates of sequenced reads. The conversion rate is calculated based on the reference annotation (GRCh38.p12) in humans without considering reported *single-nucleotide polymorphism* or PCR duplicates.

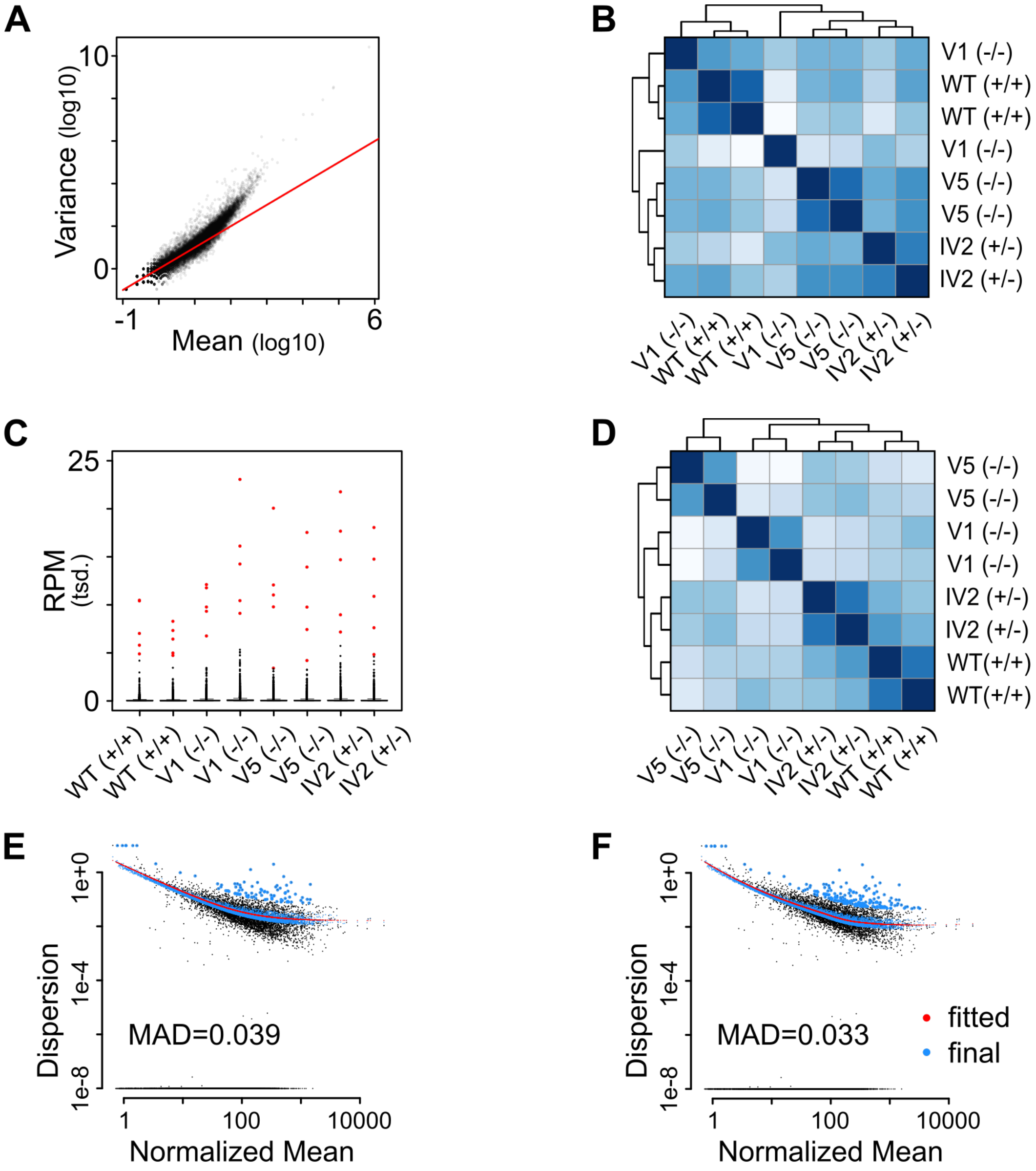


Figure A.3: Parameter Estimation for Differential Analysis using *Osteogenesis Imperfecta* Data. Related to Section 11.1. Considered are active genes (n=11,424) from patient-derived fibroblast cells, including two homozygous patients (-/-; V1, V5), one heterozygous (+/-; IV2) parent, and one unrelated healthy individual (++; WT) measured by NET-seq with two replicate measurements, respectively. **(A)** Scatter-plot shows the log10 transformed dependency between variance and means for genes. The diagonal is marked in red. **(B, D)** Hierarchical clustering of *Euclidean* distance measured between **(B)** RPM and **(D)** *median-of-ratios* normalized samples. **(C)** RPM normalized Pol II occupancy at genes. The top five expressed genes are marked in red. **(E, F)** Scatter-plot shows the dependency of the estimated log-transformed dispersion and the normalized mean expression (black). Marked are the trend curve (red) and final dispersion parameters after shrinkage (blue), using **(E)** the parametric or **(F)** local regression method.

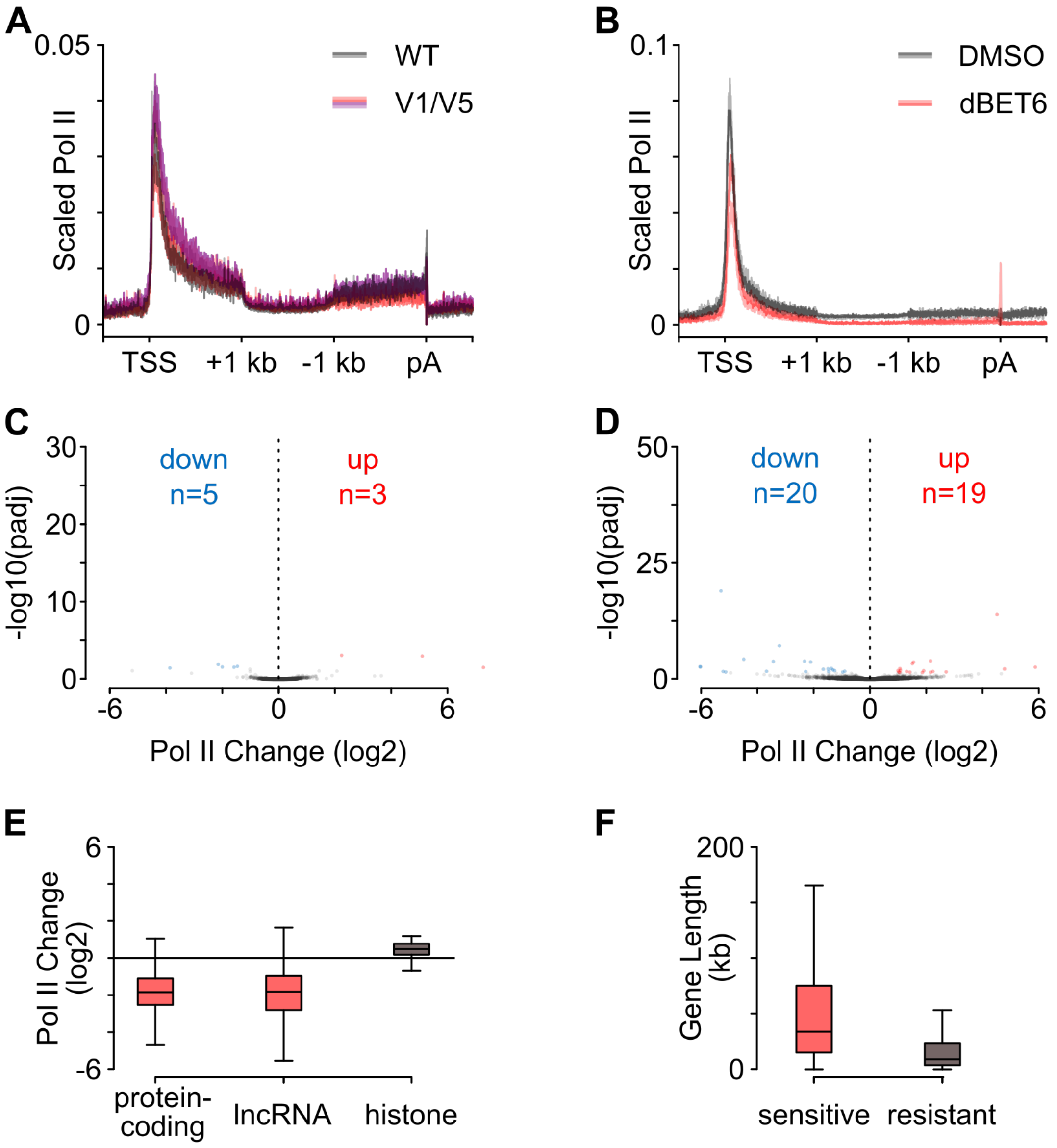


Figure A.4: Data Visualization and Interpretation. Related to Section 11.3. Meta-gene profiles of reference normalized Pol II (Section 3.2.6) at actively transcribed genes in **(A)** primary fibroblast cells (n=13,049) from patients with OI (V1, V5) and **(B)** MOLT4 cells (n=10,171) upon two hours of pan-BET protein degradation (dBET6). Both analyses were compared to the respective controls, DMSO or wild-type (WT). Nucleotides at the TSSs or signal outliers above the 99.90-quantile were excluded. **(C, D)** Pol II occupancy changes (log2) were identified with SI-NET-seq in the reference mouse cells (NIH 3T3) from the **(C)** OI (n=5,794) and **(D)** pan-BET protein degradation (n=11,554) data sets. Significant occupancy changes (padj < 0.05) are labeled in blue and red. **(E, F)** Data were obtained with SI-NET-seq after 2h of pan-BET protein degradation in MOLT4. **(E)** Pol II occupancy changes (log2) for different gene classes (protein-coding: n=9,729, lncRNA: n=911, histone: n=55). **(F)** Gene length distributions at sensitive (padj < 0.05, n=7,799) and resistant (padj > 0.05, n=2,896) genes.

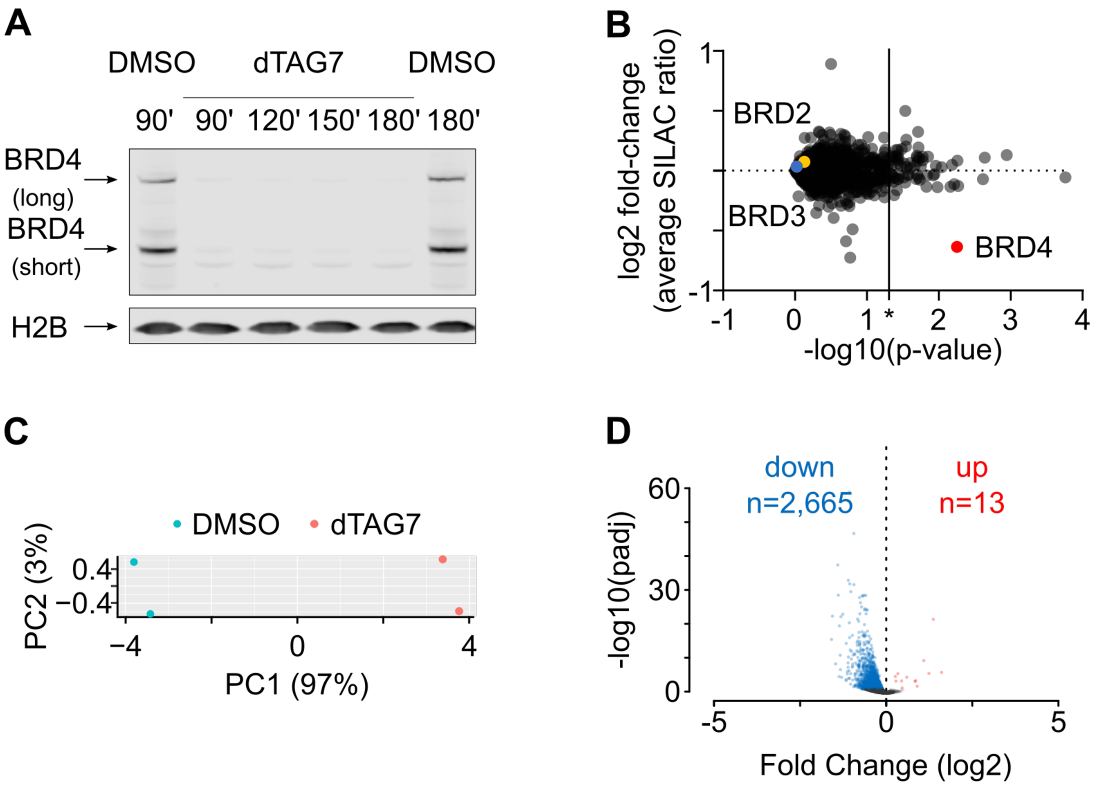


Figure A.5: BRD4-dependent Deregulation of Mature RNA Levels. Related to Section 15.1.1. If not stated otherwise, the experimental designs include control experiments (DMSO) and two hours of BRD4-protein degradation (dTAG7) in human K562 dTAG-BRD4 cells. **(A)** Immunoblot after treatment for the indicated time points. Arrows indicate the expressed long and short BRD4 isoforms. H2B served as a loading control. **(B)** Mass spectrometry quantifies protein levels (n=2,882, four replicate measurements) from SILAC-labelled whole-cell lysates. **(C)** Principal component analysis for batch corrected and ERCC spike-in normalized genes (n=32,136) measured by total RNA-seq with two replicates for each condition. **(D)** Changes of total RNA levels (log2, n=10,296) identified with total RNA-seq using ERCC spike-in controls. Differentially expressed genes (padj < 0.05) are labeled in blue and red.

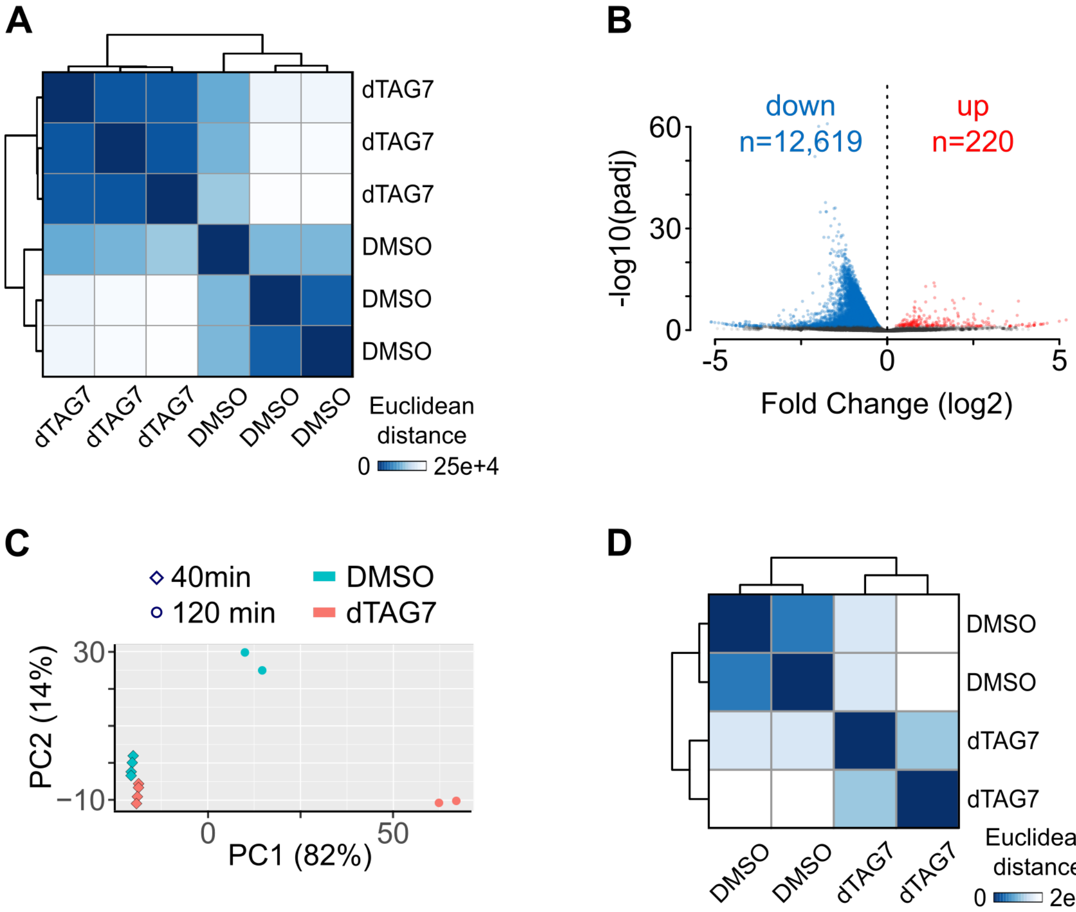


Figure A.6: BRD4-dependent Deregulation of Nascent Transcription. Related to Section 15.1.1. The experimental designs include control experiments (DMSO) and two hours of BRD4-protein degradation (dTAG7) in human K562 dTAG-BRD4 cells. **(A)** Hierarchical clustering of *Euclidean* distance measured by nuclei-RNA-seq between ERCC spike-in normalized genes (n=32,972) for three replicate measurements. **(B)** Changes of RNA levels (log2, n=24,660) identified with nuclei-RNA-seq using ERCC spike-in controls. Differentially expressed genes (padj < 0.05) are labeled in blue and red. **(C)** Principal component analysis for batch corrected and spike-in normalized genes (n=39,617) measured by HiS-NET-seq for indicated time points (40 minutes: four replicates; 120 minutes: two replicates). **(D)** Hierarchical clustering of *Euclidean* distance measured by SI-NET-seq between spike-in normalized genes (n=10,069) for two replicate measurements for each condition.

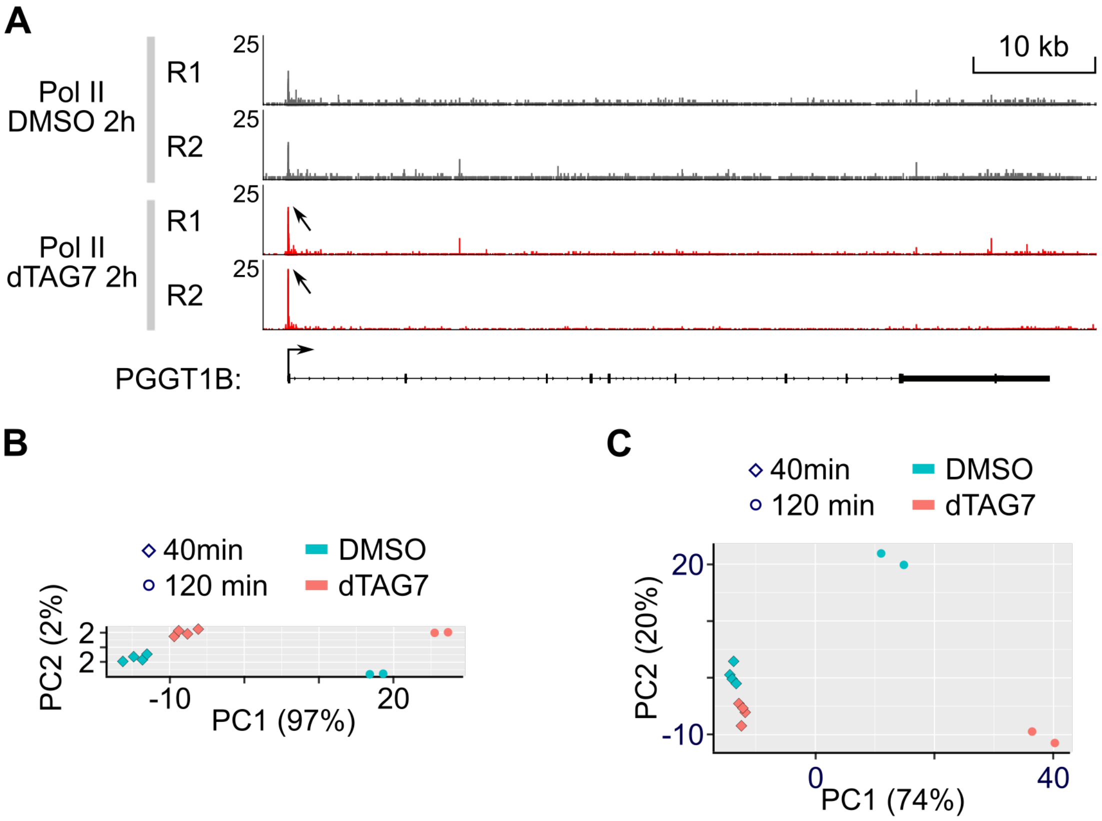


Figure A.7: Regulation of Pol II Pause Release by BRD4. Related to Section 15.1.2 and 15.1.3. The experimental design includes non-overlapping actively transcribed genes from human K562 dTAG-BRD4 cells in the control experiment (DMSO) and two hours of BRD4-protein degradation (dTAG7) with two HiS-NET-seq replicate measurements. **(A)** Pol II occupancy at *PGGT1B* gene. Data is reference-based normalized (Section 3.2.6). **(B, C)** Principal component analysis for batch corrected and reference normalized **(B)** promoter-proximal (n=24,872) and **(C)** gene-body regions (n=34,541) for indicated time points (40 minutes: four replicates).

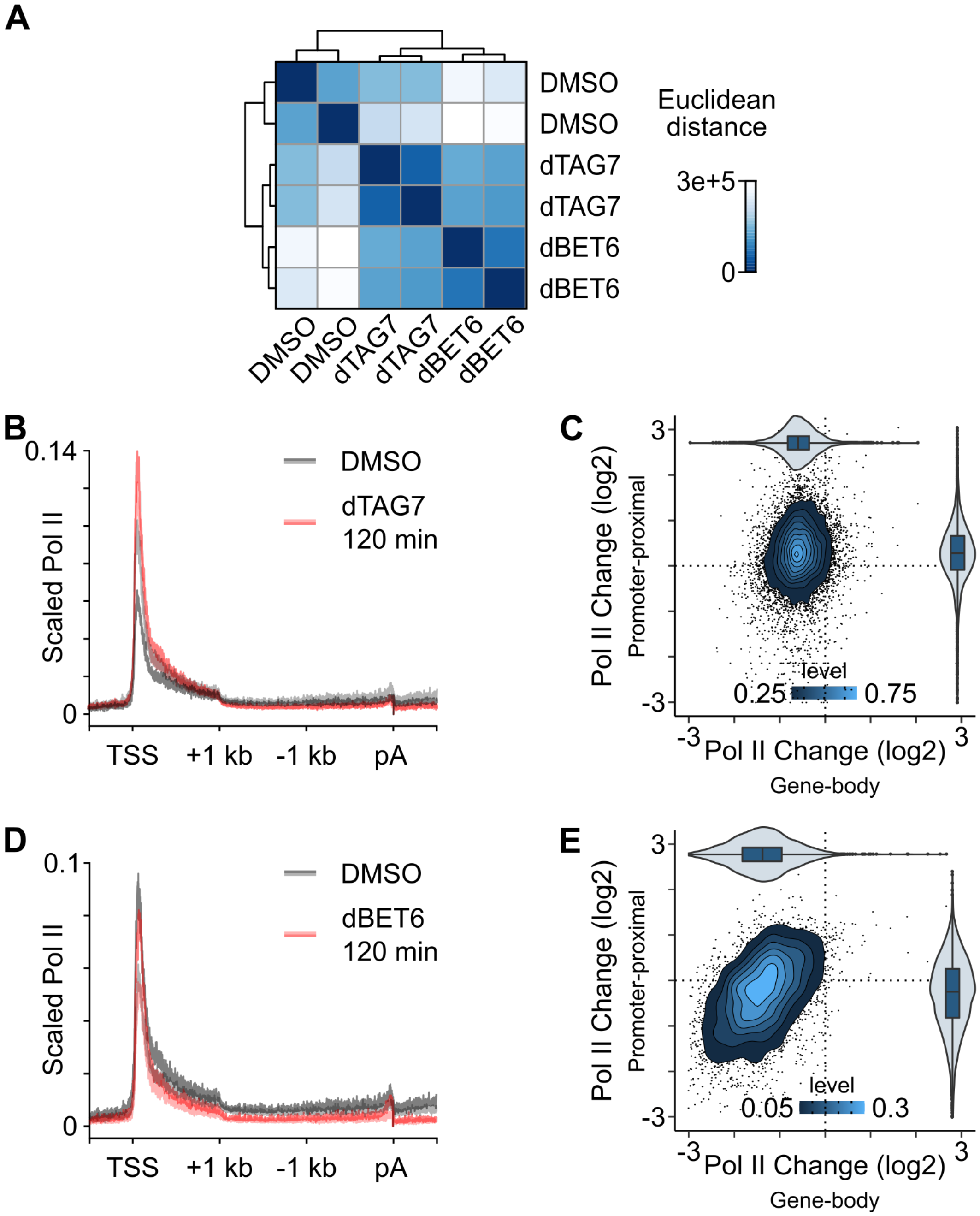


Figure A.8: Elongation Defects Compared Across Treatments. Related to Section 15.1.2 and 15.1.3. SI-NET-seq data for two replicate measurements of controls (DMSO), two hours of **(B, C)** BRD4-specific degradation (dTAG7), and **(D, E)** pan-BET protein degradation (dBET6) in human K562 dTAG-BRD4 cells. **(A)** Hierarchical clustering of *Euclidean* distance between non-overlapping spike-in normalized genes (n=10,099). **(B, D)** Meta-gene profiles of reference -based normalized (Section 3.2.6) Pol II occupancy (dTAG7: n=10,976; dBET6: n=11,014). Regions with signal outliers above the 99.90-quantile and TSSs were masked. **(C, E)** Pol II occupancy changes at promoter-proximal (y-axis) and gene-body regions (x-axis) (dTAG7: n=8,265; dBET6: n=4,293).

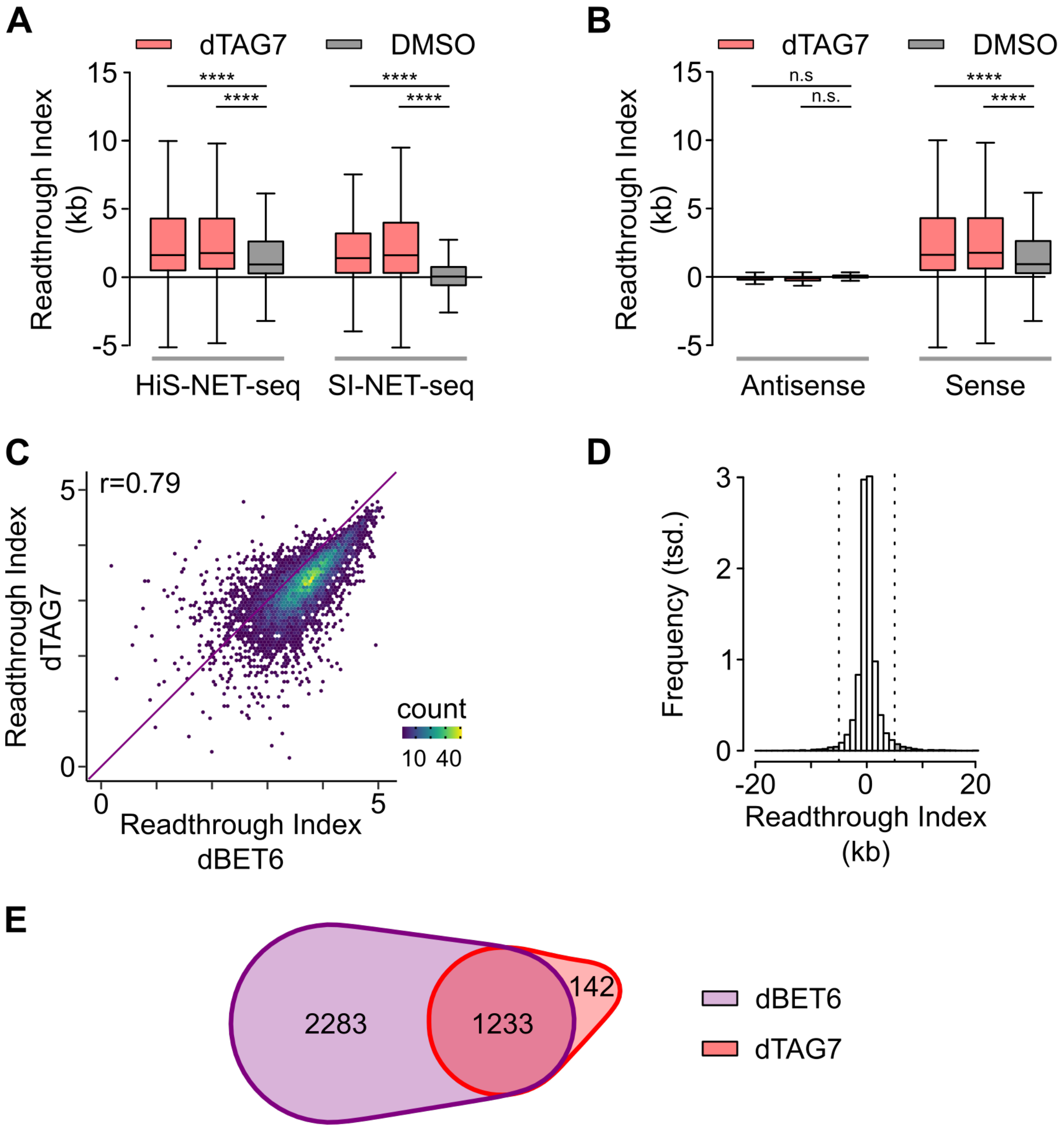


Figure A.9: RTI Comparison Across Treatments. Related to Section 15.2.1. The Figure depicts readthrough transcription at actively transcribed genes upon indicated treatments in human K562 dTAG-BRD4 cells, with two replicate measurements each. **(A-C)** Boxplot quantifications of RTI calculations upon BRD4 degradation (dTAG7) for **(A)** HiS-NET-seq and SI-NET-seq (HiS-NET-seq: n=9608-9,646; SI-NET-seq: n=9,418-9,446; one sided *Wilcoxon* rank sum test ****: p-value < 2.2e-16) and **(B)** antisense transcription units (antisense: n=4,251-4,344; one sided *Wilcoxon* rank sum test n.s: p-value = 1, ****: p <2.2e-16) measured by HiS-NET-seq. **(C-E)** Comparison of pan-BET protein degradation (dBET6) and BRD4 degradation (dTAG7). **(C)** Scatterplot comparison of the computed RTI values (*Pearson's* correlation, r = 0.79). **(D)** Histogram of the RTI distribution between two control measurements. **(E)** Venn diagram of transcriptional readthrough genes (RTI > 5 kb).

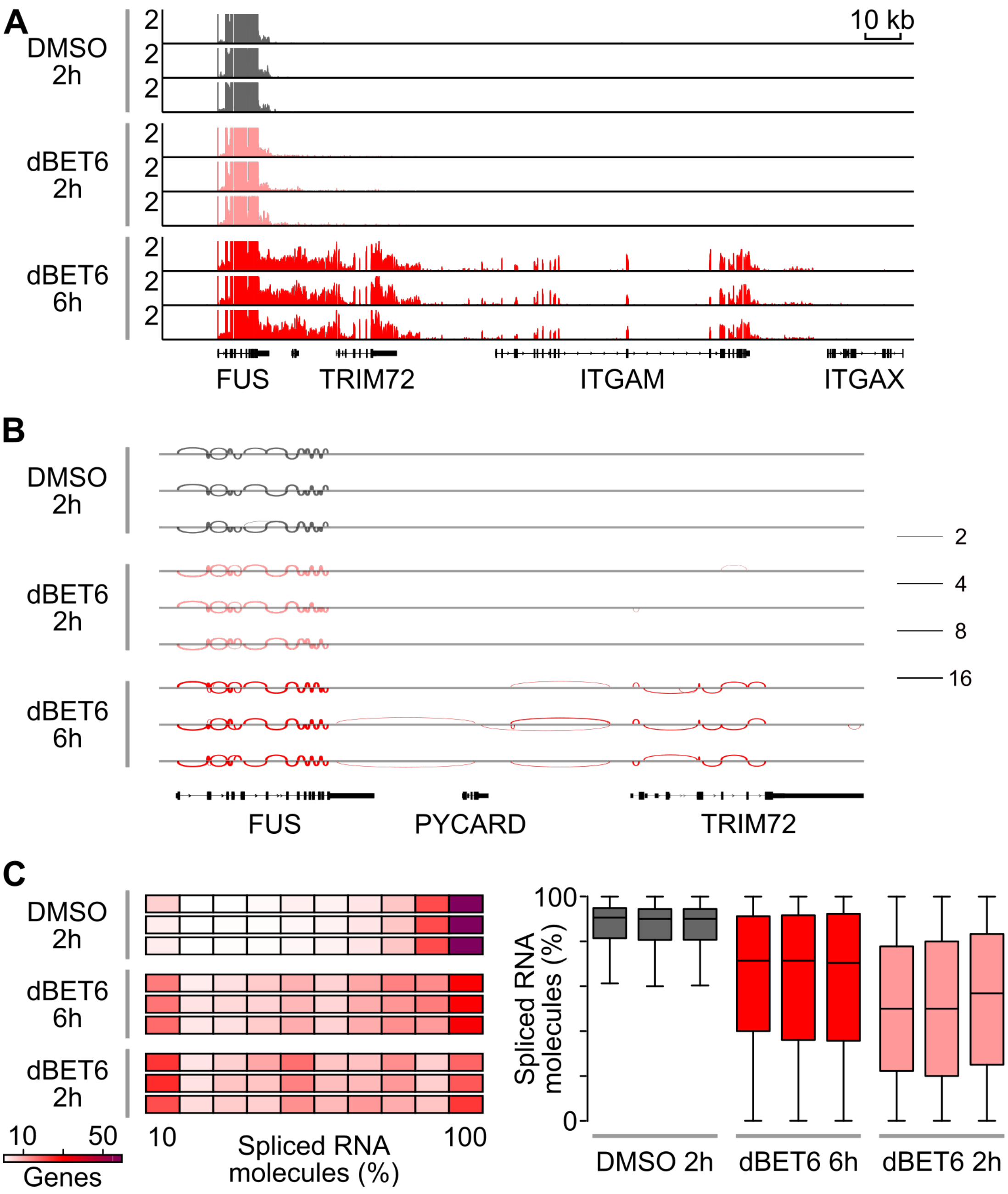


Figure A.10: Readthrough-induced Gene Activation. Related to Section 15.2.2. The figure depicts total RNA-seq data upon two and six hours of pan-BET degradation (dBET6) and the control experiment (DMSO) in MOLT4 for three replicate measurements. **(A)** Gene track of RPM normalized total RNA levels upon indicated treatments. **(B)** Sashimi plot highlights new splicing events downstream of *FUS* (>1 read). **(C)** Distribution heatmap and boxplot of spliced RNA molecules per gene (Section 14.3) across all active genes (n=12,581, DMSO 2h) and at readthrough activated genes (dBET6 2h: n=472; dBET6 6h: n=875) upon indicated treatments.

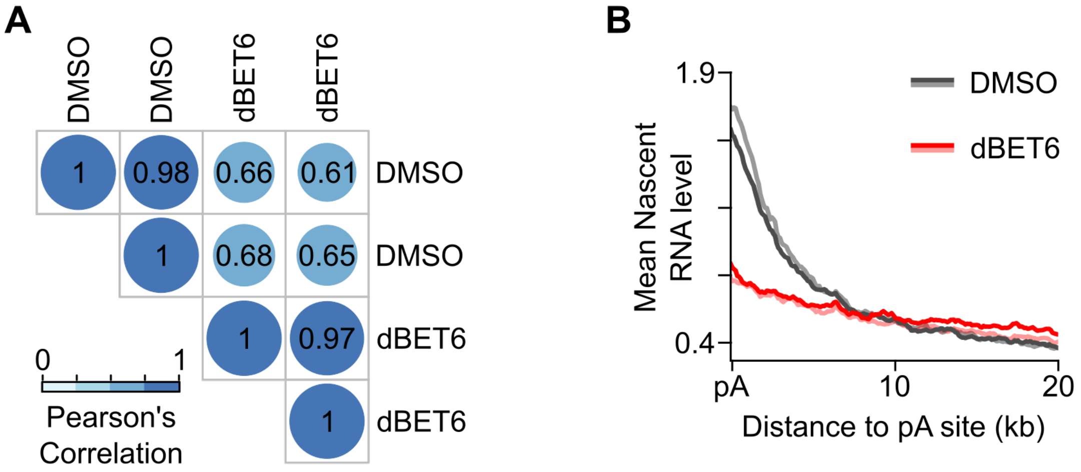


Figure A.11: Nascent RNA Sequencing using ONT. Related to Section 15.2.3. (A) Pairwise comparisons of *Pearsons's* correlation between *median-of-ratios* normalized nascONT-seq samples at actively transcribed genes (n=9,610). FeatureCounts v2.0.0 [125] identified the abundance of transcripts at active genes (Section 6.1.1) using the long-read mode *-L*. **(B)** Mean nascent RNA levels downstream of active genes after 120 minutes dBET6 treatment and DMSO using nascONT-seq data.

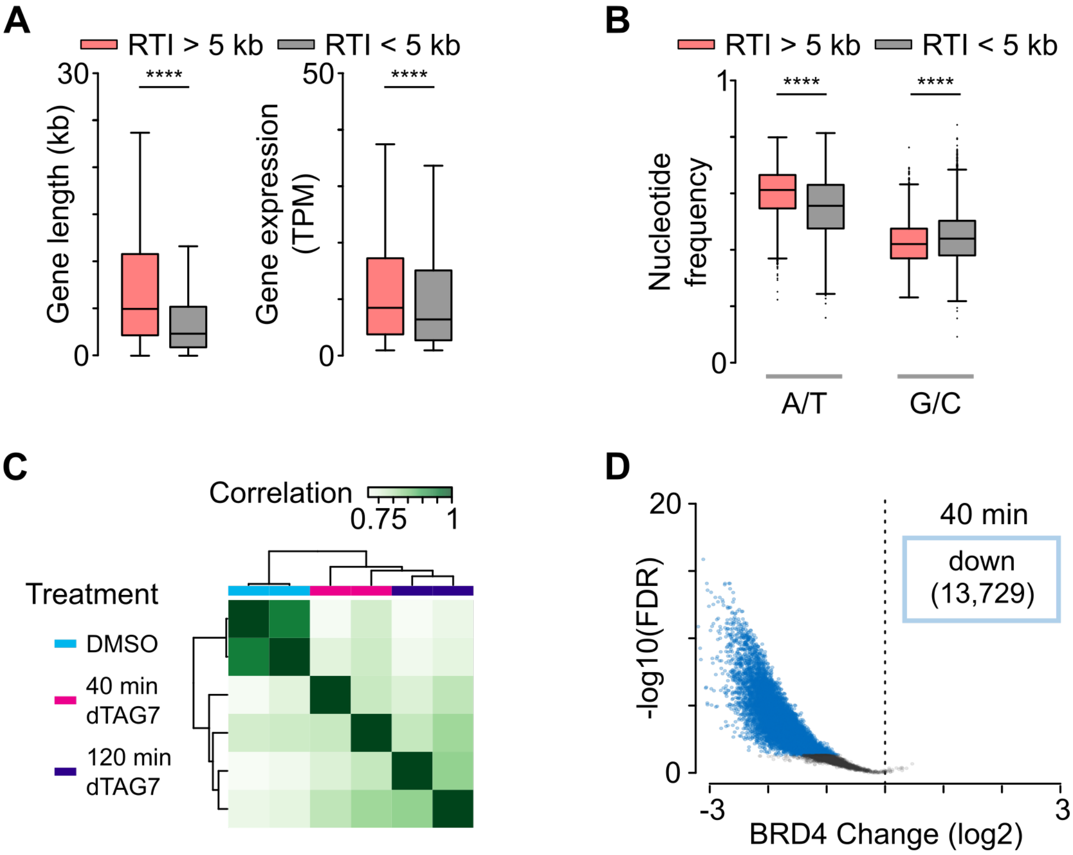


Figure A.12: Features of Readthrough Transcription. Related to Section 15.2.4. Boxplot quantification of **(A)** gene lengths (****: p < 2.2e-16), steady state RNA levels (****: p = 1.9e-14), and **(B)** nucleotide frequencies (****: p < 1.5e-15) at non-overlapping readthrough (n=1,824) and non-readthrough genes (n=6,618) from human K562 dTAG-BRD4 cells. One tailed *Wilcoxon* rank sum tests were performed. **(C)** Hierarchical clustering of cross-correlation between spike-in normalized BRD4 ChIP-Rx samples. **(D)** BRD4 ChIP-Rx occupancy changes (log2) at BRD4 peaks (n=16,233) upon 40 minutes of BRD4-protein degradation. Data was analyzed with DiffBind [221] using spike-in normalization. Significant occupancy changes (FDR < 0.05) are labeled in blue and red.

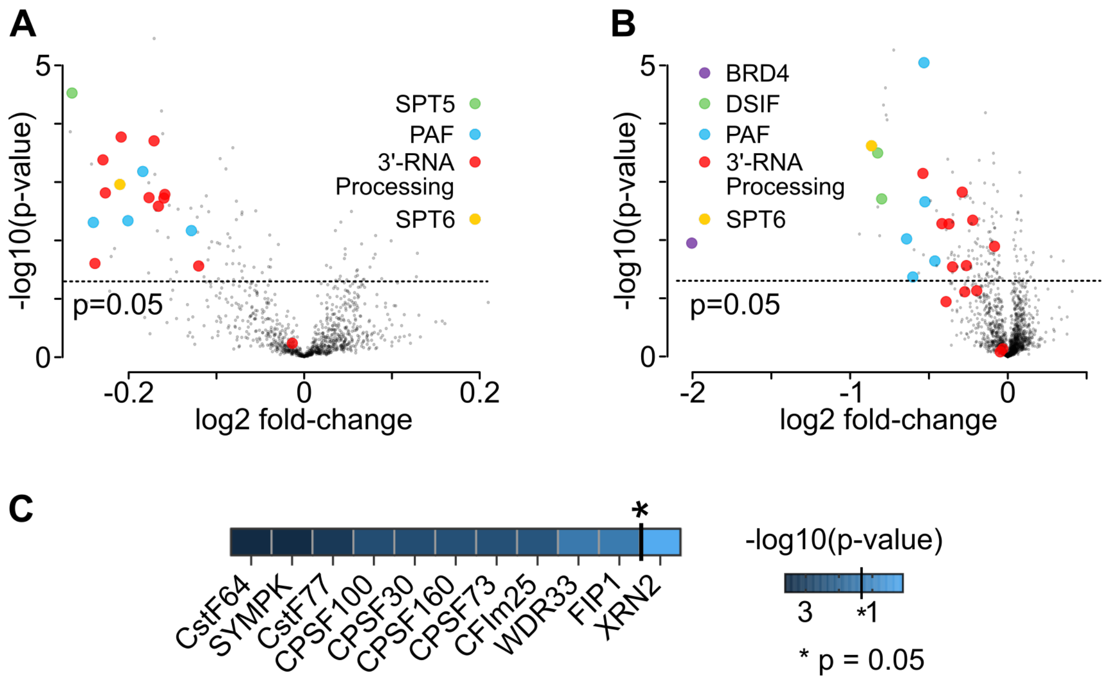


Figure A.13: Depletion of Elongation and 3′-RNA Processing Factors from the Chromatin. Related to Section 15.3.1. If not stated otherwise, data was collected upon 120 minutes of BRD4 degradation in K562 dTAG-BRD4 cells. **(A, B)** Changes in protein chromatin composition upon 120 minutes of **(A)** BRD4 (n=964) and **(B)** BET protein (n=1,563) degradation as determined by quantitative chromatin-MS. Processed data was extracted from [7]. **(C)** Heatmap of 3′-RNA processing and termination factors detected by chromatin-MS upon treatment ranked by p-value.

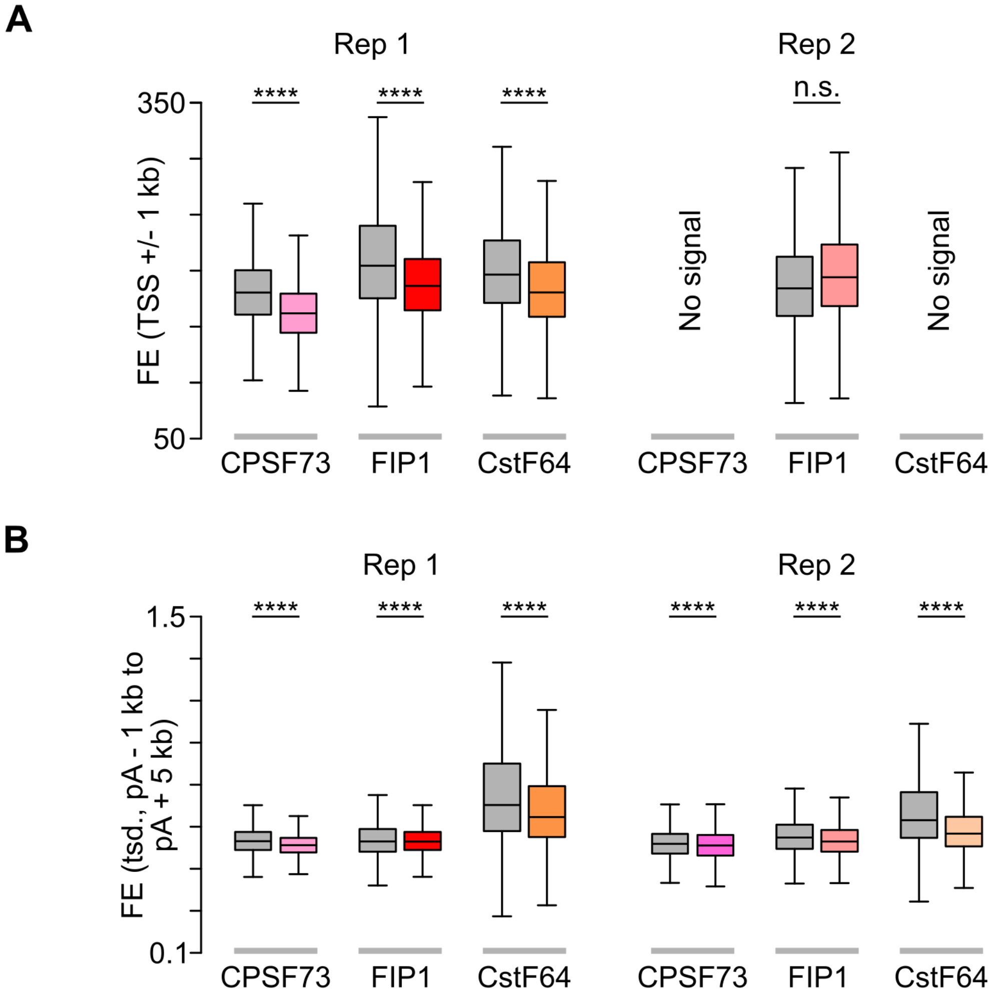


Figure A.14: Recruitment Defect of 3′-RNA Processing Factors to 5′ Gene Regions. Related to Section 15.3.1. Data was collected upon 120 minutes of BRD4 degradation in K562 dTAG-BRD4 cells. Box plot quantification of FE of 3′-RNA-processing factors (CPSF73, FIP1, CstF64) upon treatment and for the DMSO around gene's (n=7,331) **(A)** 5′ (TSS +/- 1 kb) and **(B)** 3′ ends (pA site -1 kb to pA site +5 kb). One tailed *Wilcoxon* signed-rank tests were performed (****: p < 1.72e-12). Two replicate measurements show no enrichment at the 5′ end (No signal).

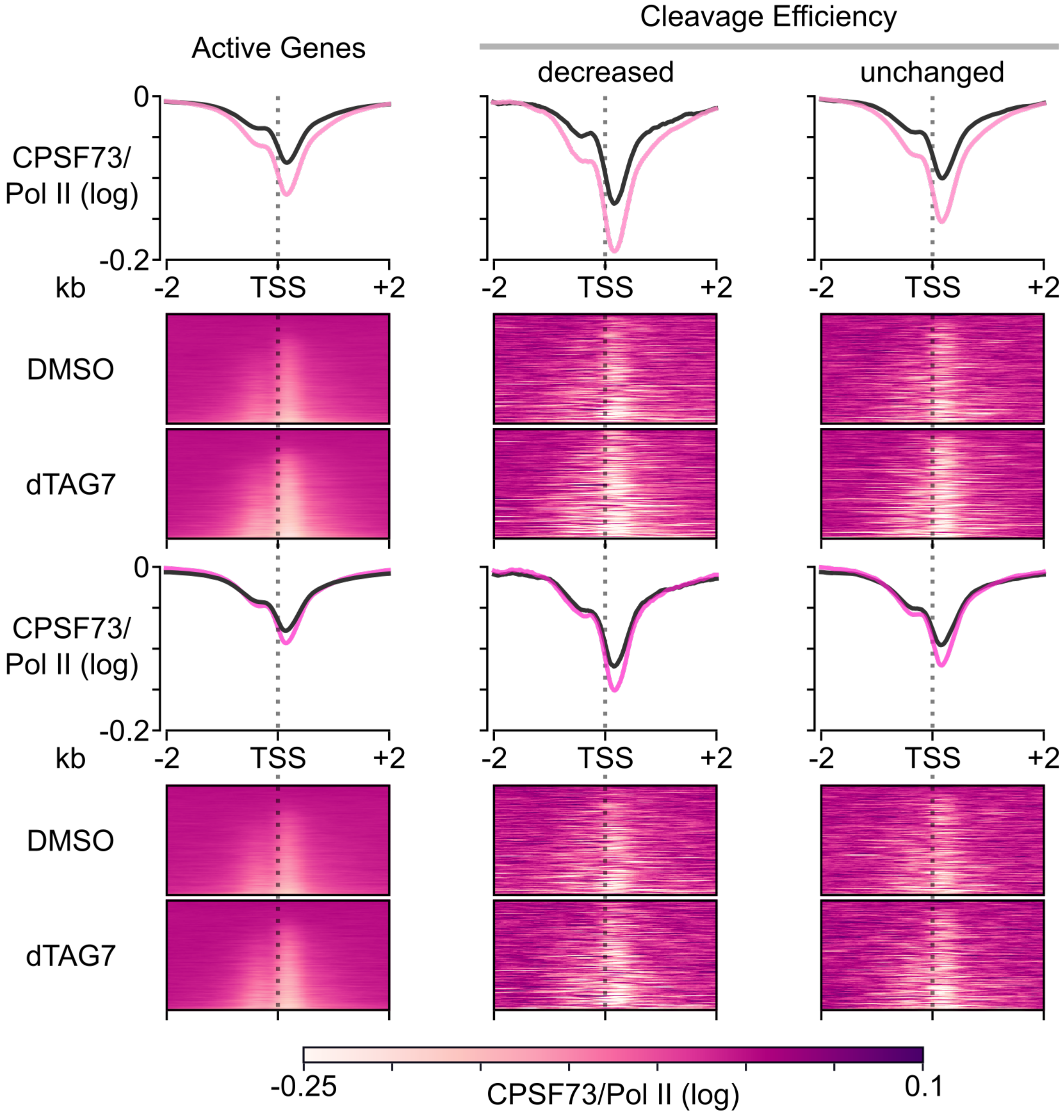


Figure A.15: Recruitment Defects of CPSF73. Related to Section 15.3.1. Meta-gene profiles and heatmaps of Pol II-normalized occupancy levels for CPSF73 upon 120 minutes of BRD4 degradation in K562 dTAG-BRD4 cells at gene regions (TSS +/- 2 kb) of different gene sets (active genes: n=12,364; decreased cleavage efficiency (dBET6): n=282, unchanged cleavage efficiency (dBET6): n=1,794).

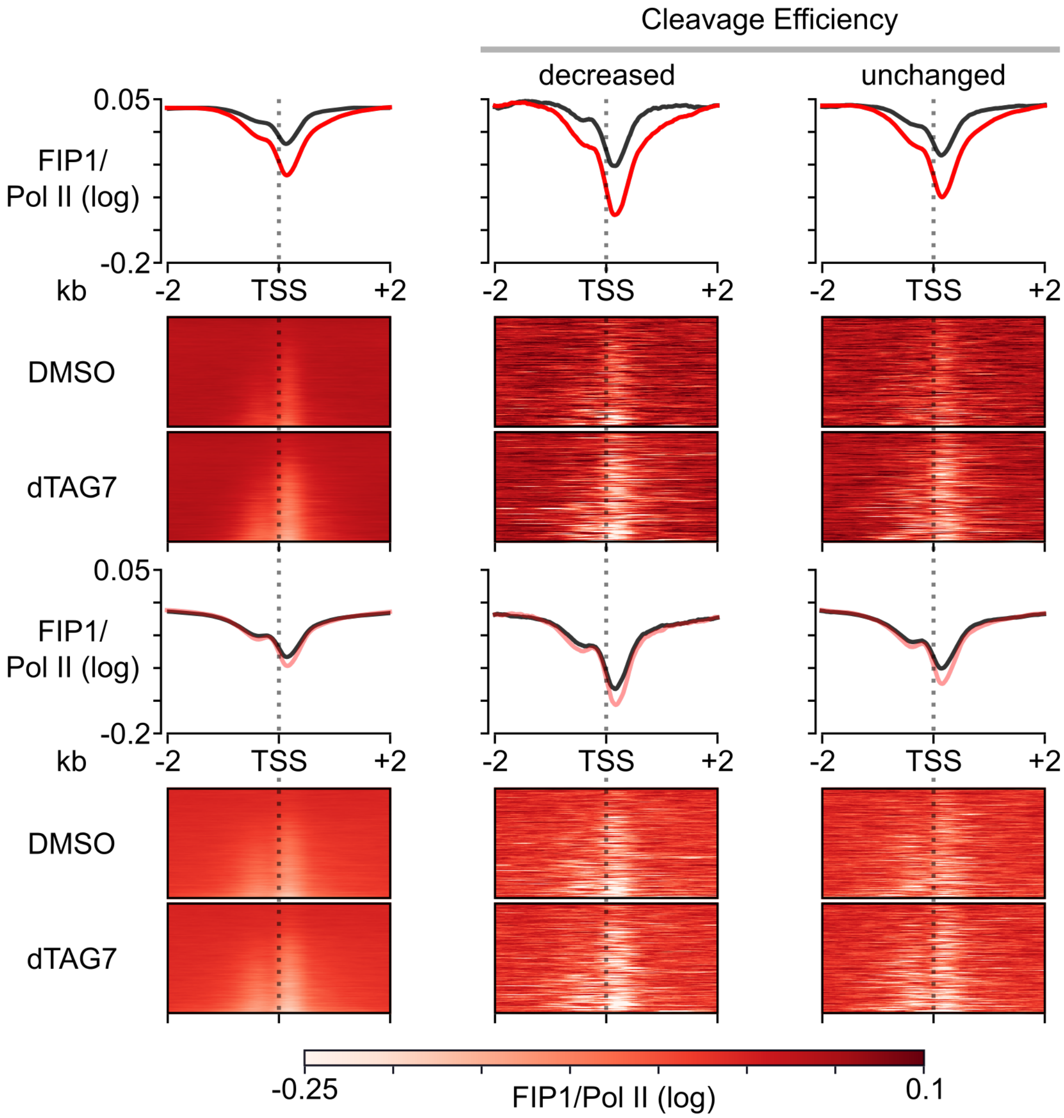


Figure A.16: Recruitment Defects of FIP1. Related to Section 15.3.1. Meta-gene profiles and heatmaps of Pol II-normalized occupancy levels for FIP1 upon 120 minutes of BRD4 degradation in K562 dTAG-BRD4 cells at gene regions (TSS +/- 2 kb) of different gene sets (active genes: n=12,364; decreased cleavage efficiency (dBET6): n=282, unchanged cleavage efficiency (dBET6): n=1,794).

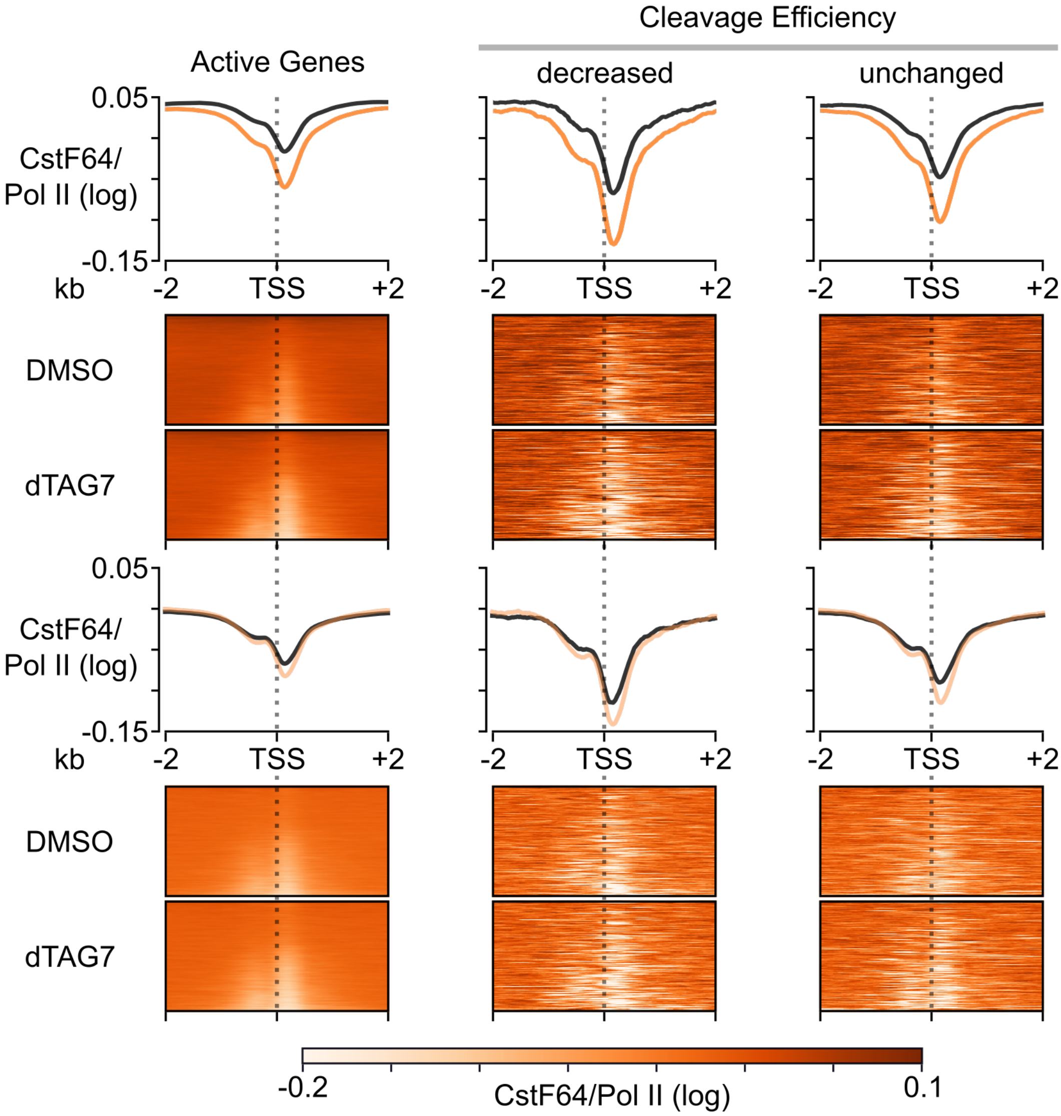


Figure A.17: Recruitment Defects of CstF64. Related to Section 15.3.1. Meta-gene profiles and heatmaps of Pol II-normalized occupancy levels for CstF64 upon 120 minutes of BRD4 degradation in K562 dTAG-BRD4 cells at gene regions (TSS +/- 2 kb) of different gene sets (active genes: n=12,364; decreased cleavage efficiency (dBET6): n=282, unchanged cleavage efficiency (dBET6): n=1,794).

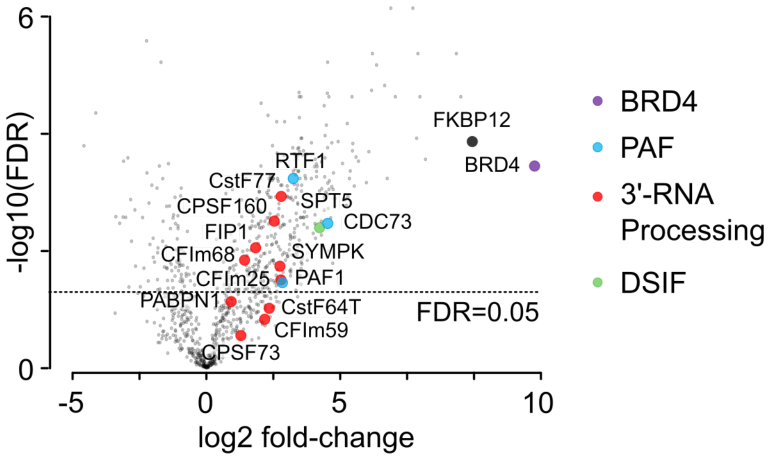


Figure A.18: BRD4 Interactors. Related to Section 15.3.2. Interactome of BRD4 as identified by IP-MS in K562 dTAG-BRD4 cells. The dashed line indicates the significance threshold (FDR < 0.05). Processed data was extracted from [7].

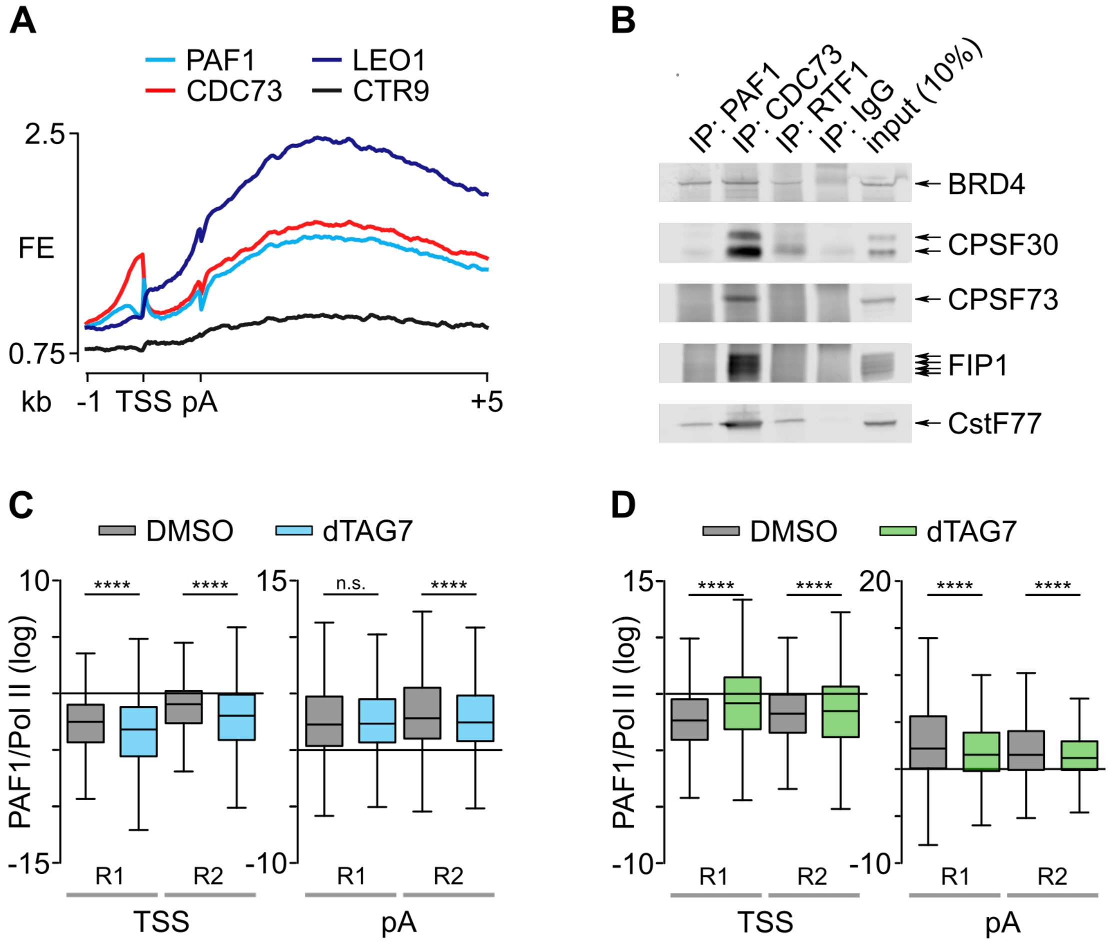


Figure A.19: Contribution of PAF1 to 3′-RNA Processing. Related to Section 15.3.3. **(A)** Meta-gene profiles of FE-normalized PAF subunits in human THP1 cells (GSE62171 [253]) at actively transcribed genes (TSS -1 kb to pA site +5 kb). **(B)** Immunoblot for HA-tagged BRD4 (Section 2.4.4) and several subunits of the CPSF and CstF after native immunoprecipitation of the PAF subunits PAF1, CDC73, and RTF1 in K562 dTAG-BRD4 cells. IgG was included to estimate the specificity of the immunoprecipitation. Input shows the general detectability of the proteins in the lysate. **(C-D)** Box plot quantification of Pol II-normalized **(C)** PAF1 and **(D)** SPT5 upon 120 minutes of dTAG7 treatment and for the DMSO control around gene's (n=12,386) 5′ (TSS +/- 300 kb) and 3′ ends (pA site +3 kb) in K562 dTAG-BRD4 cells. One tailed *Wilcoxon* signed-rank tests were performed (****: p < 2.2e-16; n.s.: p = 0.94).

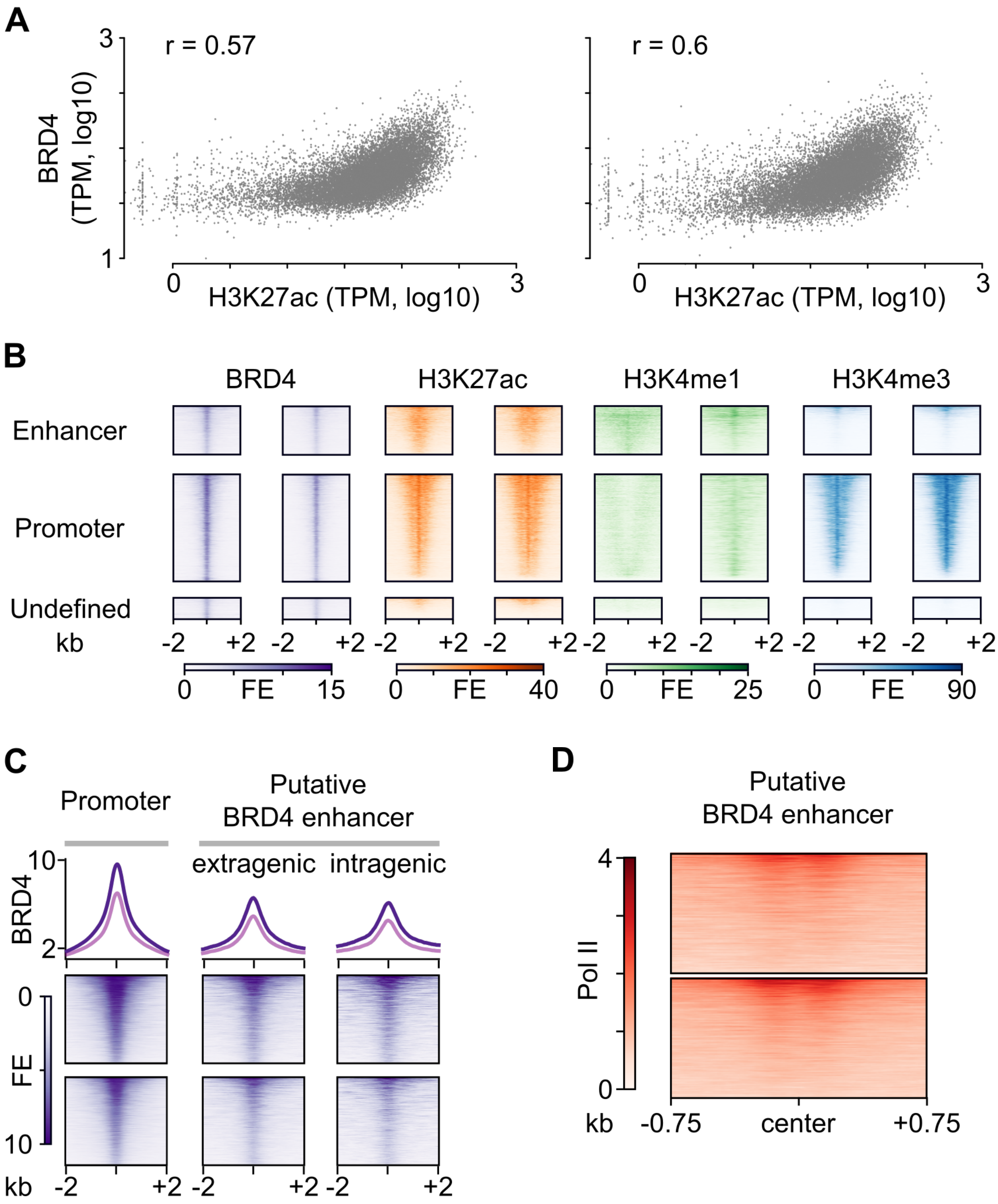


Figure A.20: Putative BRD4 Enhancer Regions. Related to Section 15.4.1. (A) Correlation between TPM values (log10) of H3K27ac and BRD4 occupancy at BRD4 consensus peaks (n=16,184, *Pearson's* correlation coefficient: r=0.56 and r=0.6) for two BRD4 ChIP-Rx replicate measurements. A pooled H3K27ac data set from ENCODE was used. **(B-D)** Heatmaps visualize FE-normalized signals at BRD4 binding sites for indicated proteins or modifications of interest. Depicted are **(B)** putative enhancers (n=4,308), promoter (n=9,504) and undefined (n=1,917) regions, **(C)** promoter (n=9,504), putative extragenic (n=2,829) and intragenic enhancer (n=1,479) regions, and **(D)** putative enhancer regions (n=3,408).

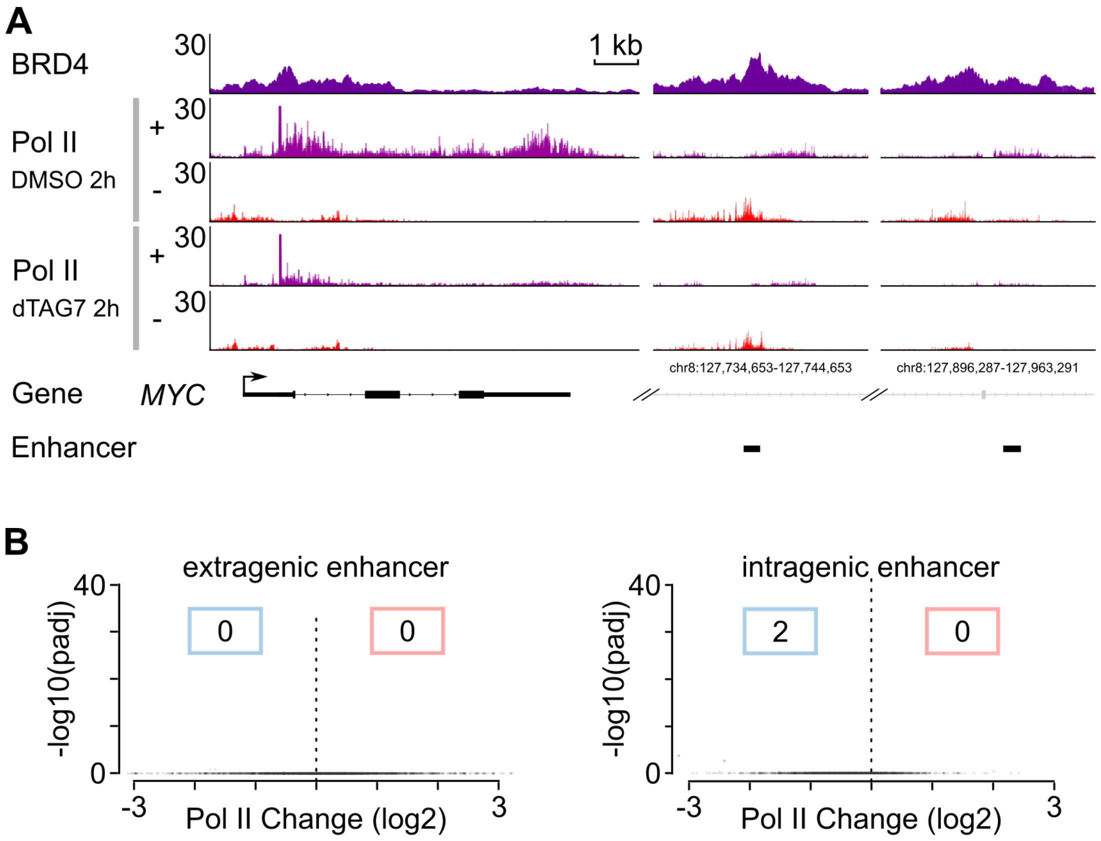


Figure A.21: Elongation Control at Enhancer Regions. Related to Section 15.4.2. The figure depicts Pol II occupancy changes after two hours of BRD4-specific degradation (dTAG7) and the control experiment (DMSO) measured by HiS-NET-seq. **(A)** Reference-normalized (Section 3.2.6) Pol II occupancy at the *MYC* gene and associated annotated FANTOM5 enhancers. **(B)** Pol II occupancy changes (log2) at extragenic and intragenic enhancer regions using SI-NET-seq (n=2,824, n=1,478). Significant occupancy changes (padj < 0.05) are labeled blue and red.

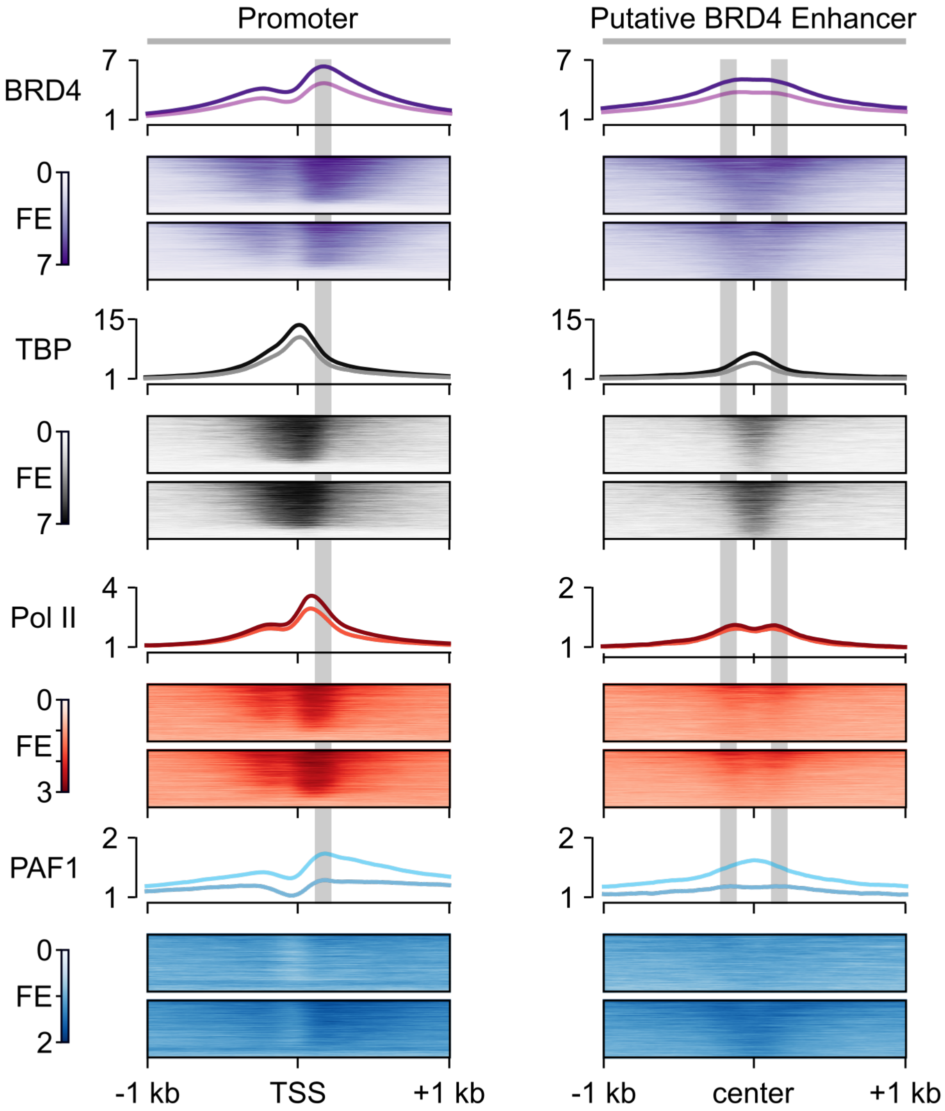


Figure A.22: Protein-binding Landscape at Enhancer Regions. Related to Section 15.4.2. Meta-gene profiles and heatmaps of FE normalized (Section 3.2.7) ChIP-Rx profiles. Presented are occupancy levels for BRD4, TBP (ENCSR000EHA [35]), Pol II, and PAF1 at 5′ regions of actively transcribed genes (n=12,374) and putative BRD4-enhancer centers (n=4,308) for two replicate measurements. A gray box marks the peak occupancy locations of BRD4.

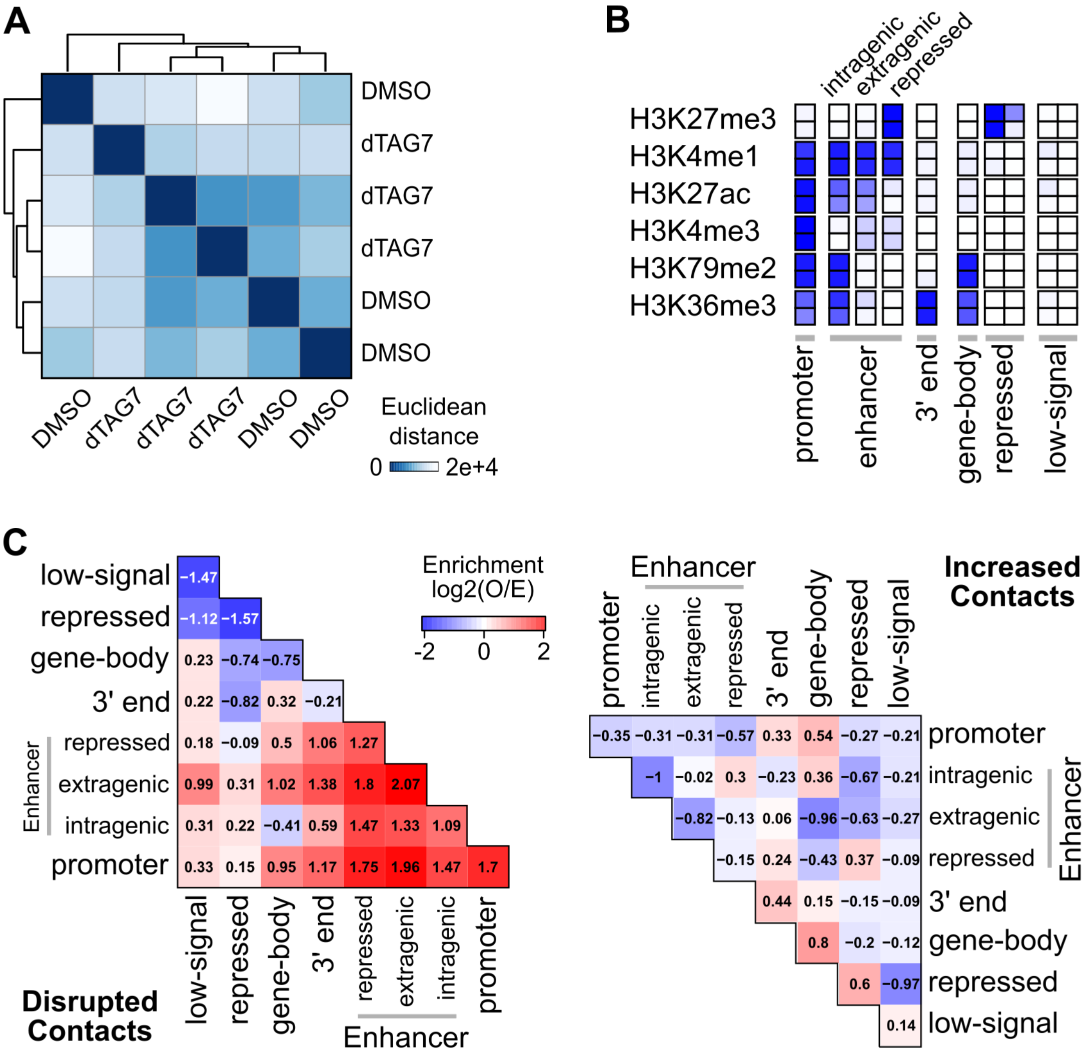


Figure A.23: Characterization of Changed Interactions. Related to Section 15.4.3. Presentation of measurements from human K562 dTAG-BRD4 cells upon 120 minutes of BRD4 degradation. **(A)** Hierarchical clustering of *Euclidean* distance between pairwise normalized MD interaction frequencies for controls (DMSO) and two hours of BRD4-specific degradation (dTAG7) for three replicate measurements. **(B)** Emission probability of chromatin marks in different states of the chromHMM [56] model. Annotations were assigned manually. **(C)** Enrichment of pairwise interactions between different genomic regions among contacts with disrupted and increased interaction frequency changes. Enrichment was calculated as the logarithmic ratio of observed contacts (O) and expected (E). Expected counts were derived from all pairwise contacts in the genome.

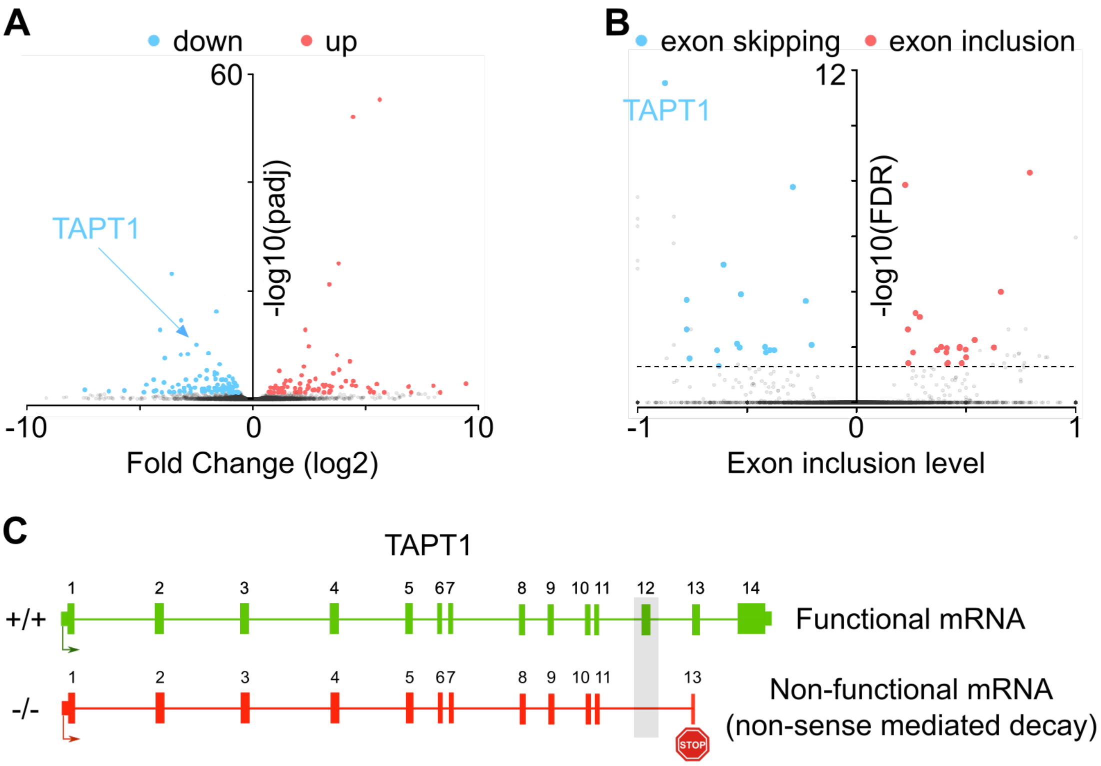


Figure A.24: Intronic TAPT1 Mutation Leads to Exon Twelve Skipping. Related to Section 18. The figure compares data derived from WT (WT1 and WT2) and patient (V.1 (F1), V.5 (F1)) *primary dermal fibroblasts* cells measured by polyA-enriched RNA-seq. **(A)** The volcano plot shows differentially expressed transcript levels (n=17,675). **(B)** The figure summarizes the results of the alternative splicing analysis showing differences in exon usage (x-axis) between patients identified by rMATs [215] (n=19,128). **(A, B)** Significant changes with a FDR adjusted p-value (y-axis) below 0.05 are labeled in blue and red. *TAPT1* is in both analyses among the top-ranking deregulated genes. **(C)** Schematic representation of exon twelve loss in the TAPT1 transcript.

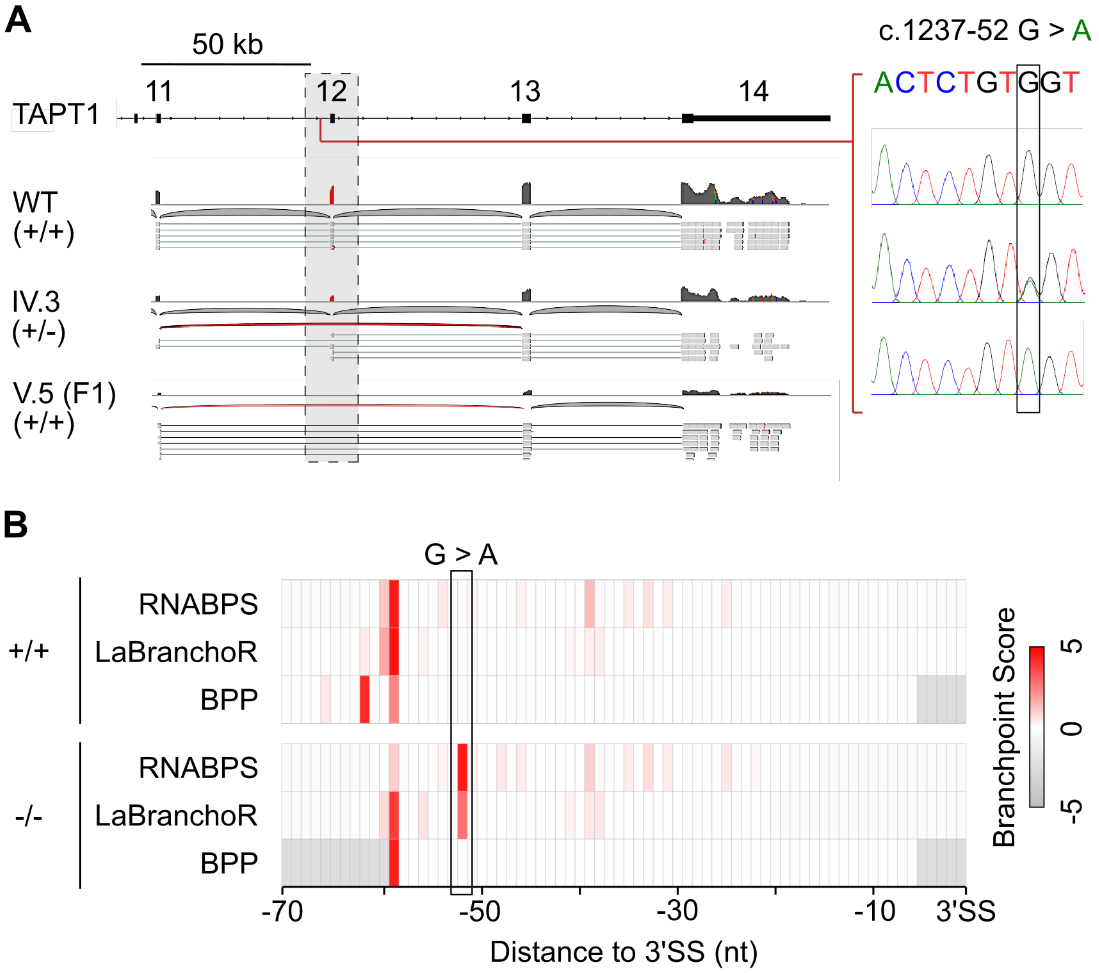


Figure A.25: Mutation Creates New Putative Branchpoint Position. Related to Section 18. The figure compares data derived from WT (WT1 and WT2) and patient (V.1 (F1), V.5 (F1)) *primary dermal fibroblasts* cells measured by polyA-enriched RNA-seq. **(A)** Transcript fragments from patients at the *TAPT1* gene skip the exon twelve compared to WT. The chromatogram shows the intronic mutation (c.1237-52 G>A) of the targeted *Sanger* sequencing in WT, IV.3 (F1), and V.5 (F1). **(B)** Predicted branchpoint scores seventy nucleotides upstream of the 3′ splice site of exon twelve in the *TAPT1* gene for WT and patient (mutation of G>A). In this region, the mutation shifts the predicted branchpoint position from fifty-nine nucleotides to fifty-two nucleotides downstream of the respective 3′ splice site.

B
SUPPLEMENTARY TABLES

Table B.1: High Throughput Sequencing Data Overview. Replicate - Rep

Resource	Cell line	Replicate	Source	Identifier
		ChIP-Rx and ChIP-seq		
ChIP-Rx BRD4 - 120 min DMSO - 120 min dTAG7 - 40 min dTAG7 - input	K562 dTAG-BRD4 + NIH 3T3 (5:1)	1, 2	MRA, unpublished	MRA109-MRA116
ChIP-Rx CPSF73, 120 min - DMSO - dTAG7 - input	K562 dTAG-BRD4 + NIH 3T3 (5:1)	1, 2	MRA [7]	GEO: GSE158965
ChIP-Rx CstF64, 120 min- DMSO- dTAG7 - input	K562 dTAG-BRD4 + NIH 3T3 (5:1)	1, 2	MRA [7]	GEO: GSE158965
ChIP-Rx FIP, 120 min- DMSO- dTAG7 - input	K562 dTAG-BRD4 + NIH 3T3 (5:1)	1, 2	MRA [7]	GEO: GSE158965
ChIP-Rx Pol II Subunit 2, - 120 min DMSO - input	K562 dTAG-BRD4 + NIH 3T3 (5:1)	1, 2	MRA [7]	GEO: GSE158965input: MRA47, MRA67 (unpublished)

High Throughput Sequencing Data. Replicate - Rep

Resource	Cell line	Replicate	Source	Identifier
ChIP-seq K562 - H3K27ac - H3K36me3 - H3K79me2	K562	1, 2	ENCODE, Bernstein laboratory [35]	ENCSR000AKP, ENCSR000AKR, ENCSR000APD
ChIP-seq K562 - H3K27me3 - H3K4me1 - H3K4me3	K562	1, 2	ENCODE, Farnham laboratory [35]	ENCSR000EWB, ENCSR000EWC, ENCSR000EWA
ChIP-seq K562 - TBP	K562	1, 2	ENCODE, Snyder laboratory [35]	ENCSR000EHA
ChIP-seq THP-1 - LEO1 - CDC73 - PAF1 - CTR9	THP-1	1	Roeader laboratory [253]	GEO: GSE62171
		GRO-seq		
GRO-seq HCT116 - shSCR - shPAF1	HCT116	1, 2	Shilatifard laboratory [24]	GEO: GSE70408
GRO-seq primary activated splenic B lymphocytes - SPT5 WT - SPT5 KO	primary activated splenic B lymphocytes	1, 2	Pavri laboratory [62]	GEO: GSE132029
		HiChIP		
HiChIP, H3K27ac, 120 min - DMSO - dTAG7	K562	1-3	MRA, unpublished	MRA10- MRA15

High Throughput Sequencing Data. Replicate - Rep

Resource	Cell line	Replicate	Source	Identifier
		HiS-NET-seq		
HiS-NET-seq - 0 min 4sU (control) - 10 min 4sU	K562	1, 2	OJ, unpublished	OJ90, OJ91OJ92, OJ93
HiS-NET-seq, 120 min - DMSO - dTAG7	K562 dTAG-BRD4 + NIH 3T3 (8:1)	1, 2	OJ, unpublished	OJ94-OJ97
HiS-NET-seq, 40 min - DMSO - dTAG7	K562 dTAG-BRD4 + NIH 3T3 (8:1)	1-4	MRA, unpublished	MRA125-MRA147
		mNET-seq		
mNET-seq	K562	1, 2	Schwalb laboratory [80]	GEO: GSE123980
		nascONT-seq		
nascONT-seq, 120 min - DMSO - dTAG7	K562 dTAG-BRD4	1, 2	OJ [7]	GEO: GSE158965
		NET-seq		
NET-seq HeLa	HeLa S3	1, 2	AM [148]	GEO: GSE123980
NET-seq K562 (standard)	K562	1	OJ, unpublished	OJ01
NET-seq K562 (size selection)	K562	1	OJ, unpublished	OJ08
NET-seq K562 (optimized)	K562	1	OJ, unpublished	OJ26

High Throughput Sequencing Data. Replicate - Rep

Resource	Cell line	Replicate	Source	Identifier
		PRO-seq		
PRO-seq	K562	1	Lis laboratory [38]	GEO: GSM1480327
		RNA-seq		
RNA-seq, nuclei, K562 dTAG-BRD4, 120 min - DMSO - dTAG7	K562 dTAG-BRD4 + ERCC	1-3	NE, unpublished	NE04-NE09
RNA-seq, polyA, mouse primary activated splenic B lymphocytes	primary activated splenic B lymphocytes	1-3	Pavri laboratory [62]	GEO: GSE132029
RNA-seq, polyA, NIH 3T3	NIH 3T3	1, 2	ENCODE, Stamatoyannopoulos laboratory [35]	ENCSR000CLW
RNA-seq, polyA, primary cutaneous fibroblasts - WT 1/2 - HMZ 1/2 - HTZ	primary fibroblasts	1	Reversade laboratory, unpublished	GEO: GSE197120
RNA-seq, total, HCT116	HCT116	1-4	Shilatifard laboratory [25]	GEO: GSE97527
RNA-seq, total, K562	K562	1, 2	ENCODE, Graveley laboratory [35]	ENCSR109IQO

High Throughput Sequencing Data. Replicate - Rep

Resource	Cell line	Replicate	Source	Identifier
RNA-seq, total, K562 dTAG-BRD4, 120 min - DMSO - dTAG7	K562 dTAG-BRD4 + ERCC	1, 2	MRA, unpublished	MRA101, MRA102,MRA105, MRA106
RNA-seq, total, MOLT4, 120 min - DMSO - dBET6	MOLT4 + ERCC	1-3	Bradner laboratory [249]	GEO: GSE79253
RNA-seq, total, MOLT4, 360 min - DMSO - dBET6	MOLT4 + ERCC	1-3	Bradner laboratory [249]	GEO: GSE79253
RNA-seq, total, THP-1	THP-1	1, 2	Roeader laboratory [253]	GEO: GSE62171
		SI-NET-seq		
SI-NET-seq K562 dTAG-BRD4, 120 min - DMSO - dBET6 - dTAG7	K562 dTAG-BRD4 + NIH 3T3 (6:1)	1, 2	MRA [7]	GEO: GSE158963
SI-NET-seq MOLT4, 120 min - DMSO - dBET6	MOLT4 + NIH 3T3 (6:1)	1, 2	AM [7]	GEO: GSE158963
SI-NET-seq primary fibroblasts, - WT - HMZ 1/2 - HTZ	primary fibroblasts + NIH 3T3 (6:1)	1, 2	SF, unpublished	GEO: GSE197120

Resource	Identifier/ Version	Source	URL
FANTOM5	v5	[36]	https://fantom.gsc.riken.jp/
GENCODE	v28, v29, M18, M22	[67]	https://www.gencodegenes.org/
GENCODE	GRCh38.p12, GRCm38.p6	[67]	https://www.gencodegenes.org/
HUGO Gene Nomenclature	864	[177]	https://www.genenames.org/
polyA_DB	v3.2	[243]	http://exon.umdnj.edu/

Table B.2: Annotation Databases Overview.

Table B.3: Software and Algorithms Overview.

Resource	Version	Source	Identifier/URL
chemfig	1.6b	[226]	https://ctan.org/pkg/chemfig?lang=en
chromHMM	1.19	[56]	http://compbio.mit.edu/ChromHMM/
bedtools	2.29.2	[185]	https://bedtools.readthedocs.io/en/latest/
Biopython	1.78	[32]	https://biopython.org/
Bowtie2	2.3.5.1	[119]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
cutadapt	3.4	[143]	https://cutadapt.readthedocs.io/en/stable/
deepTools2	3.2.1	[190]	https://deeptools.readthedocs.io/en/develop/index.html
DEseq2	1.25.4	[132]	http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
DiffBind	3.0.15	[221]	https://bioconductor.org/packages/release/bioc/html/DiffBind.html

Software and Algorithms Overview.

Resource	Version	Source	Identifier/URL
FastQC	0.11.5	[4]	https://github.com/s-andrews/FastQC
grammarly	-	-	https://www.grammarly.com/
Guppy	3.2.4	ONT	https://community.nanoporetech.com
HiC-Pro	3.0.0	[212]	https://github.com/nservant/HiC-Pro
HiCcompare	1.8.0	[220]	https://www.bioconductor.org/packages/release/bioc/html/HiCcompare.html
HMMER	3.3	[176]	http://hmmer.org/
HTSeq	0.13.5	[184]	https://htseq.readthedocs.io/en/master/
IGV	2.8.0	[198]	https://software.broadinstitute.org/software/igv/
MACS2	2.2.7.1	[258]	https://github.com/jsh58/MACS
miRBase	v22.1	[115]	https://www.mirbase.org/
minimap2	2.17	[124]	https://github.com/lh3/minimap2
NumPy	1.20.2	[82]	https://numpy.org/
pysam	0.16.0.1	[77]	https://pandas.pydata.org/
PANTHER	15	[153]	http://geneontology.org/
PICARD	2.24.2	[95]	https://broadinstitute.github.io/picard/
Python	2.7.16 and 3.8.9	[231]	https://www.python.org/
QDNAseq	1.22.0	[206]	https://www.bioconductor.org/packages/release/bioc/html/QDNAseq.html
R	3.6.3	[186]	https://www.r-project.org/

Software and Algorithms Overview.

Resource	Version	Source	Identifier/URL
reactome	75	[76]	https://reactome.org/
RepeatMasker (UCSC)	human (1/3/19) and mouse (3/7/12)	[104]	http://repeatmasker.org/ human: https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=rmsk mouse: https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=mm39&c=chr12&g=rmsk
rMATS	3.1.0	[215]	http://rnaseq-mats.sourceforge.net/
RSEM	1.3.1	[122]	https://github.com/deweylab/RSEM
SAMtools	1.13	[46]	https://www.htslib.org/
Snakemake	6.8.0	[156]	https://snakemake.readthedocs.io/en/stable/
STAR	2.7.3a	[50]	https://github.com/alexdobin/STAR
Starcode	1.1	[261]	https://github.com/gui11aume/starcode
subreads	2.0.0	[125]	http://subread.sourceforge.net/

Cell line or cell type	Source	ID
HCT116	Shilatifard laboratory [25]	GSE97527
K562	ENCODE, Graveley laboratory[35]	ENCSR109IQO
K562 dTAG-BRD4	ENCODE, Graveley laboratory [35]	ENCSR109IQO
MOLT4	Bradner laboratory [249]	GSE79253
NIH 3T3	ENCODE [35]	ENCSR000CLW
primary activated splenic B lymphocytes	Pavri laboratory [62]	GSE132029
primary fibrobast cells from patients	Reversade laboratory, unpublished	GSE197120
THP-1	Roeader laboratory [253]	GEO: GSE62171

Table B.4: RNA-seq Data Defining Actively Transcribed Genes in Cell Lines.

Software	Parameters
cutadapt	-a ATCTCGTATGCCGTCTTCTGCTTG -a AAAAAAAAAAGGGGGGGGGGGGGG -a GGGGGGGGGGGGGGGGGGGGGG -e 0.2 -q 5 --max-n 0.9
Starcode	-d 0
STAR	-clip3pAdapterSeq ATCTCGTATGCCGTCTTCTGCTTG -clip3pAdapterMMp 0.21 -clip3pAfterAdapterNbases 1 -outFilterMultimapNmax 1 -outSJfilterOverhangMin 3 1 1 1- outSJfilterDistToOtherSJmin 0 0 0 0 -alignIntronMin 11 -alignEndsType EndToEnd

Table B.5: NET-seq Pipeline: Software and Parameters.

Type	Source	Keyword	RNA Polymerase
microRNA	[115]	*miRNA, miRNA_primary_transcript*	II
	[67]	*miRNA*	
miscellaneous RNA	[67]	*misc_RNA*	Unknown
ribosomal RNA	[104]	*SSU-rRNA_Hsa, LSU-rRNA_Hsa, 5S (III)*	I
	[67]	*rRNA, rRNA_pseudogene*	
sn/snoRNA	[104]	*U1, U2, U3, U4, U5, U6, U7, U8, U13, U14, U17, 7SK*	II and III
	[67]	*sRNA, snRNA, snoRNA, scaRNA*	
transfer RNA	[104]	*tRNA*	III
vault RNA	[67]	*vaultRNA*	III
Y RNA	[104]	*HY1, HY3, HY4, HY5*	III

Table B.6: *In silico* Masking of Chromatin-associated Mature RNA

Method	Enrichment				Bias control		Single-nucleotid resolution
	Chromatin isolation	Pol II - IP	Run-on	Metabolic labeling	PCR	IP	
GRO-seq [39]	-	-	X	-	-	N/A	-
HiS-NET-seq	X	-	-	X	X	N/A	X
mNET-seq [169]	X	X	-	-	-	-	X
NET-seq [148]	X	-	-	-	X	N/A	X
Pol II ChIP-seq [12, 100]	-	X	-	-	-	X	-
PRO-seq [136]	-	-	X	-	-	N/A	X
qPRO-seq [103]	-	-	-	-	X	N/A	X
SNU-seq [151]	-	-	-	X	-	N/A	X
TT-seq * [211]	-	-	-	X	-	N/A	-

Table B.7: Comparison of Pol II Profiling Methods. (X) Feature applies for the respective method, (N/A) is not available or (-) not. * TT-seq does not perform 3′-end sequencing, and hence provides no Pol II occupancy information.

	NET-seq	HiS-NET-seq	
% sequenced reads	R1	R1	R2
uniquely mapped to mouse	0.43%	0.8%	0.7%
PCR duplicates	- 0.13%	- 0.09%	- 0.08%
splicing intermediates, reverse transcriptase mispriming			
masked regions	- 0.14%	- 0.17%	- 0.18%
extragenic regions	- 0.09%	- 0.32%	- 0.26%
cross-contamination bias	**0.07%**	**0.22%**	**0.18%**

Table B.8: Cross-Mapping of Human Reads to the Mouse Genome. Sample statistics for NET-seq (R1) and HiS-NET-seq (R1 and R2) measured in the human cell line K562.

Function	Parameters

dba.blacklist	blacklist=DBA_BLACKLIST_HG38
	blacklist=DBA_BLACKLIST_MM10

dba.count	minOverlap=2 summits=300
	bRemoveDuplicates=true
	bSubControl=true

dba.normalize	parameters spikein=true normalize=DBA_NORM_RLE
dba.contrast	-
dba.analyze	-

Table B.9: Differential Binding Analysis: Functions and Parameters.

Software	Parameters
Guppy	–flowcell FLO-MIN106–kit SQK-DCS109
HMMER hmmbuild, hmmpress	-
HMMER hmmalign	--trim
HMMER nhmmscan (iteration 1)	--noali --notextw -max -E 0.1
HMMER nhmmscan (iteration 2)	--notextw -max -E 10
minimap2	-ax splice -ub -k14–secondary = no -O 12,32–junc-bonus = 19 --junc-bed

Table B.10: NascONT-seq Pipeline: Software and Parameters.

Read type	Valid state combinations
full-length	$(Start, VNP^+, MAP_P, SSP^-, End)$
	$(Start, SSP^+, MAP_P, VNP^-, End)$
	$(Start, VNP^+, MAP_S, SSP^-, End)$
	$(Start, SSP^+, MAP_S, VNP^-, End)$
5′-truncated	$(Start, SSP^+, MAP_P, End)$
	$(Start, MAP_P, SSP^-, End)$
	$(Start, SSP^+, MAP_S, End)$
	$(Start, MAP_S, SSP^-, End)$
3′-truncated	$(Start, VNP^+, MAP_P, End)$
	$(Start, MAP_P, VNP^-, End)$
	$(Start, VNP^+, MAP_S, End)$
	$(Start, MAP_S, VNP^-, End)$
no primer	$(Start, MAP_P, End)$
	$(Start, MAP_S, End)$
fused	$(Start, VNP^+, MAP_P, SSP^-, SSP^+, MAP_S, VNP^-, End)$
	$(Start, VNP^+, MAP_P, SSP^-, SSP^+, MAP_S, End)$
	$(Start, VNP^+, MAP_S, SSP^-, SSP^+, MAP_P, VNP^-, End)$
	$(Start, VNP^+, MAP_S, SSP^-, SSP^+, MAP_P, End)$

Table B.11: Read Classification of ONT Data. Primary mapping position - MAP_S; Supplementary mapping position -MAP_S; VNP primer sense - VNP^+; VNP primer antisense - VNP^-; SSP primer sense - SSP^+; SSP primer antisense - SSP^-

Software	Parameters
bowtie2 (iteration 1)	--sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end
bowtie2 (iteration 2)	-sensitive -L 20 --score-min L,-0.6,-0.2 --end-to-end
QDNAseq	CNV.level = 2 bin.size = 10,000
HiCcompare hic_loess	min.A = 9
chromHMM BinarizeBam	-b 10000
chromHMM LearnModel	-s 0 -b 10000

Table B.12: HiChIP Pipeline: Software and Parameters.

BIBLIOGRAPHY

[1] Karen Adelman and John T Lis. "Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans." In: *Nature Reviews Genetics* 13.10 (2012), pp. 720–731 (cit. on pp. 1, 9, 121).

[2] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. "Opportunities and challenges in long-read sequencing data analysis." In: *Genome biology* 21.1 (2020), pp. 1–16 (cit. on pp. 17, 18, 124).

[3] Priti Anand et al. "BET bromodomains mediate transcriptional pause release in heart failure." In: *Cell* 154.3 (2013), pp. 569–582 (cit. on pp. 11, 81).

[4] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010 (cit. on pp. 83, 165).

[5] Angeliki Andrikopoulou, Michalis Liontos, Konstantinos Koutsoukos, Meletios-Athanasios Dimopoulos, and Flora Zagouri. "Clinical perspectives of BET inhibition in ovarian cancer." In: *Cellular Oncology* 44.2 (2021), pp. 237–249 (cit. on pp. 2, 11).

[6] *Anti RNA polymerase II CTD MAB (Clone Mabi 0601)*. https://www.cosmobiousa.com/products/anti-rna-polymerase-ii-ctd-mab-clone-mabi-0601. Accessed: 2022-04-08 (cit. on p. 59).

[7] Mirjam Arnold, Annkatrin Bressin, Olga Jasnovidova, David Meierhofer, and Andreas Mayer. "A BRD4-mediated elongation control point primes transcribing RNA polymerase II for 3′-processing and termination." In: *Molecular Cell* (2021) (cit. on pp. v, 34, 40, 53, 66, 81, 97, 101, 104, 108, 110, 122–124, 145, 150, 159, 161, 163).

[8] Carlo Baejen et al. "Genome-wide analysis of RNA polymerase II termination at protein-coding genes." In: *Molecular cell* 66.1 (2017), pp. 38–49 (cit. on pp. 92, 127).

[9] Apoorva Baluapuri et al. "MYC recruits SPT5 to RNA polymerase II to promote processive transcription elongation." In: *Molecular cell* 74.4 (2019), pp. 674–687 (cit. on pp. 13, 14, 26, 70, 77).

[10] Julian Banerji, Sandro Rusconi, and Walter Schaffner. "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences." In: *Cell* 27.2 (1981), pp. 299–308 (cit. on p. 6).

[11] Laura Baranello et al. "RNA polymerase II regulates topoisomerase 1 activity to favor efficient transcription." In: *Cell* 165.2 (2016), pp. 357–371 (cit. on p. 11).

[12] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. "High-resolution profiling of histone methylations in the human genome." In: *Cell* 129.4 (2007), pp. 823–837 (cit. on pp. 13, 169).

[13] Caroline R Bartman, Nicole Hamagami, Cheryl A Keller, Belinda Giardine, Ross C Hardison, Gerd A Blobel, and Arjun Raj. "Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation." In: *Molecular cell* 73.3 (2019), pp. 519–532 (cit. on p. 81).

[14] David LV Bauer, Michael Tellier, Mónica Martínez-Alonso, Takayuki Nojima, Nick J Proudfoot, Shona Murphy, and Ervin Fodor. "Influenza virus mounts a two-pronged attack on host RNA polymerase II transcription." In: *Cell reports* 23.7 (2018), pp. 2119–2129 (cit. on pp. 92, 127).

[15] Felipe Beckedorff et al. "The human integrator complex facilitates transcriptional elongation by endonucleolytic cleavage of nascent transcripts." In: *Cell reports* 32.3 (2020), p. 107917 (cit. on pp. 45, 63, 78).

[16] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300 (cit. on p. 30).

[17] Prashant Bhat, Drew Honson, and Mitchell Guttman. "Nuclear compartmentalization as a mechanism of quantitative control of gene expression." In: *Nature Reviews Molecular Cell Biology* 22.10 (2021), pp. 653–670 (cit. on pp. 11, 126).

[18] Gregory T Booth, Isabel X Wang, Vivian G Cheung, and John T Lis. "Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast." In: *Genome research* 26.6 (2016), pp. 799–811 (cit. on pp. 78, 92, 127).

[19] Kris Brannan et al. "mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription." In: *Molecular cell* 46.3 (2012), pp. 311–324 (cit. on p. 8).

[20] Kirk M Brown and Gregory M Gilmartin. "A mechanism for the regulation of pre-mRNA 3′ processing by human cleavage factor Im." In: *Molecular cell* 12.6 (2003), pp. 1467–1476 (cit. on pp. 9, 10).

[21] Michael Burrows and David Wheeler. "A block-sorting lossless data compression algorithm." In: *Digital SRC Research Report*. Citeseer. 1994 (cit. on p. 21).

[22] José Manuel Pérez Cañadillas and Gabriele Varani. "Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein." In: *The EMBO journal* 22.11 (2003), pp. 2821–2830 (cit. on pp. 9, 10).

[23] Giacomo Cavalli and Tom Misteli. "Functional implications of genome topology." In: *Nature structural & molecular biology* 20.3 (2013), pp. 290–299 (cit. on p. 16).

[24] Fei Xavier Chen, Ashley R Woodfin, Alessandro Gardini, Ryan A Rickels, Stacy A Marshall, Edwin R Smith, Ramin Shiekhattar, and Ali Shilatifard. "PAF1, a molecular regulator of promoter-proximal pausing by RNA polymerase II." In: *Cell* 162.5 (2015), pp. 1003–1015 (cit. on pp. 113, 160).

[25] Fei Xavier Chen et al. "PAF1 regulation of promoter-proximal pause release via enhancer activation." In: *Science* 357.6357 (2017), pp. 1294–1298 (cit. on pp. 162, 167).

[26] Kaifu Chen, Zheng Hu, Zheng Xia, Dongyu Zhao, Wei Li, and Jessica K Tyler. "The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses." In: *Molecular and cellular biology* 36.5 (2015), pp. 662–667 (cit. on pp. 63, 70, 77).

[27] L Stirling Churchman and Jonathan S Weissman. "Nascent transcript sequencing visualizes transcription at nucleotide resolution." In: *Nature* 469.7330 (2011), pp. 368–373 (cit. on pp. 15, 58).

[28] Michael D Cleary, Christopher D Meiering, Eric Jan, Rebecca Guymon, and John C Boothroyd. "Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay." In: *Nature biotechnology* 23.2 (2005), pp. 232–237 (cit. on p. 18).

[29] Kendell Clement, Rick Farouni, Daniel E Bauer, and Luca Pinello. "AmpUMI: design and analysis of unique molecular identifiers for deep amplicon sequencing." In: *Bioinformatics* 34.13 (2018), pp. i202–i210 (cit. on pp. 49, 58).

[30] William S Cleveland and Eric Grosse. "Computational methods for local regression." In: *Statistics and computing* 1.1 (1991), pp. 47–62 (cit. on pp. 68, 91).

[31] Andrea G Cochran, Andrew R Conery, and Robert J Sims. "Bromodomains: a new target class for drug development." In: *Nature Reviews Drug Discovery* 18.8 (2019), pp. 609–628 (cit. on pp. 2, 11).

[32] Peter JA Cock et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics." In: *Bioinformatics* 25.11 (2009), pp. 1422–1423 (cit. on pp. 40, 164).

[33] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis." In: *Genome biology* 17.1 (2016), pp. 1–19 (cit. on pp. 83, 93).

[34] Sheila Connelly and James L Manley. "A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II." In: *Genes & development* 2.4 (1988), pp. 440–452 (cit. on p. 10).

[35] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489.7414 (2012), p. 57 (cit. on pp. 6, 14, 22, 33, 42, 73, 83, 91, 117, 119, 154, 160, 162, 167).

[36] Fantom Consortium et al. "A promoter-level mammalian expression atlas." In: *Nature* 507.7493 (2014), p. 462 (cit. on pp. 6, 40, 55, 56, 114, 115, 117, 164).

[37] Gene Ontology Consortium. "Expansion of the Gene Ontology knowledgebase and resources." In: *Nucleic acids research* 45.D1 (2017), pp. D331–D338 (cit. on p. 31).

[38] Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers." In: *Nature genetics* 46.12 (2014), pp. 1311–1320 (cit. on pp. 53, 162).

[39] Leighton J Core, Joshua J Waterfall, and John T Lis. "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." In: *Science* 322.5909 (2008), pp. 1845–1848 (cit. on pp. 16, 53, 58, 125, 169).

[40] Leighton Core and Karen Adelman. "Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation." In: *Genes & development* 33.15-16 (2019), pp. 960–982 (cit. on pp. 8, 9, 125).

[41] Michael A Cortazar, Ryan M Sheridan, Benjamin Erickson, Nova Fong, Kira Glover-Cutter, Kristopher Brannan, and David L Bentley. "Control of RNA Pol II speed by PNUTS-PP1 and Spt5 dephosphorylation facilitates termination by a "sitting duck torpedo" mechanism." In: *Molecular cell* 76.6 (2019), pp. 896–908 (cit. on p. 124).

[42] Patrick Cramer. "Organization and regulation of gene transcription." In: *Nature* 573.7772 (2019), pp. 45–54 (cit. on p. 7).

[43] Patrick Cramer et al. "Structure of eukaryotic RNA polymerases." In: *Annu. Rev. Biophys.* 37 (2008), pp. 337–352 (cit. on pp. 1, 7).

[44] Menno P Creyghton et al. "Histone H3K27ac separates active from poised enhancers and predicts developmental state." In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21931–21936 (cit. on pp. 6, 115).

[45] Nicholas T Crump et al. "BET inhibition disrupts transcription but retains enhancer-promoter contact." In: *Nature communications* 12.1 (2021), pp. 1–15 (cit. on pp. 11, 77, 115, 125, 126).

[46] Petr Danecek et al. "Twelve years of SAMtools and BCFtools." In: *Gigascience* 10.2 (2021), giab008 (cit. on p. 166).

[47] Lee Davidson, Lisa Muniz, and Steven West. "3′ end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells." In: *Genes & development* 28.4 (2014), pp. 342–356 (cit. on p. 124).

[48] Yarui Diao et al. "A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells." In: *Nature methods* 14.6 (2017), pp. 629–635 (cit. on p. 6).

[49] Giorgio Dieci, Gloria Fiorino, Manuele Castelnuovo, Martin Teichmann, and Aldo Pagano. "The expanding RNA polymerase III transcriptome." In: *TRENDS in Genetics* 23.12 (2007), pp. 614–622 (cit. on p. 7).

[50] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics* 29.1 (2013), pp. 15–21 (cit. on pp. 21, 39, 40, 71, 83, 166).

[51] Heather L Drexler, Karine Choquet, and L Stirling Churchman. "Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores." In: *Molecular cell* 77.5 (2020), pp. 985–998 (cit. on pp. 104, 124, 128).

[52] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998 (cit. on p. 21).

[53] Joshua D Eaton, Laura Francis, Lee Davidson, and Steven West. "A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes." In: *Genes & development* 34.1-2 (2020), pp. 132–145 (cit. on pp. 10, 124).

[54] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic acids research* 30.1 (2002), pp. 207–210 (cit. on p. 33).

[55] Nicole Eischer, Mirjam Arnold, and Andreas Mayer. "Emerging roles of BET proteins in transcription and co-transcriptional RNA processing." In: *Wiley Interdisciplinary Reviews: RNA* (2022), e1734 (cit. on p. 10).

[56] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization." In: *Nature methods* 9.3 (2012), pp. 215–216 (cit. on pp. 91, 119, 155, 164).

[57] Paolo Ferragina and Giovanni Manzini. "Opportunistic data structures with applications." In: *Proceedings 41st annual symposium on foundations of computer science*. IEEE. 2000, pp. 390–398 (cit. on p. 21).

[58] Andrew Field and Karen Adelman. "Evaluating enhancer function and transcription." In: *Annual Review of Biochemistry* 89 (2020), pp. 213–234 (cit. on pp. 6, 117).

[59] Panagis Filippakopoulos et al. "Selective inhibition of BET bromodomains." In: *Nature* 468.7327 (2010), pp. 1067–1073 (cit. on pp. 11, 19, 20).

[60] Panagis Filippakopoulos et al. "Histone recognition and large-scale structural analysis of the human bromodomain family." In: *Cell* 149.1 (2012), pp. 214–231 (cit. on pp. 10, 115).

[61] Ronald A Fisher. "On the interpretation of χ 2 from contingency tables, and the calculation of P." In: *Journal of the Royal Statistical Society* 85.1 (1922), pp. 87–94 (cit. on p. 31).

[62] Johanna Fitz, Tobias Neumann, Monika Steininger, Eva-Maria Wiedemann, Adriana Cantoran Garcia, Alexander Athanasiadis, Ursula E Schoeberl, and Rushad Pavri. "Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction." In: *Nature Genetics* 52.5 (2020), pp. 505–515 (cit. on pp. 113, 122, 123, 160, 162, 167).

[63] Ryan A Flynn et al. "7SK-BAF axis controls pervasive transcription at enhancers." In: *Nature structural & molecular biology* 23.3 (2016), pp. 231–238 (cit. on p. 125).

[64] Barbara Fontanals-Cirera et al. "Harnessing BET inhibitor sensitivity reveals AMIGO2 as a melanoma survival gene." In: *Molecular cell* 68.4 (2017), pp. 731–744 (cit. on pp. 11, 115, 125).

[65] Antonella Forlino, Wayne A Cabral, Aileen M Barnes, and Joan C Marini. "New perspectives on osteogenesis imperfecta." In: *Nature Reviews Endocrinology* 7.9 (2011), pp. 540–557 (cit. on p. 65).

[66] Antonella Forlino and Joan C Marini. "Osteogenesis imperfecta." In: *The Lancet* 387.10028 (2016), pp. 1657–1671 (cit. on p. 65).

[67] Adam Frankish et al. "GENCODE reference annotation for the human and mouse genomes." In: *Nucleic acids research* 47.D1 (2019), pp. D766–D773 (cit. on pp. 39, 40, 42, 46, 73, 83, 89, 93, 114, 164, 168).

[68] Melissa J Fullwood et al. "An oestrogen-receptor-α-bound human chromatin interactome." In: *Nature* 462.7269 (2009), pp. 58–64 (cit. on pp. 17, 117).

[69] Martyna Gajos, Olga Jasnovidova, Alena van Bömmel, Susanne Freier, Martin Vingron, and Andreas Mayer. "Conserved DNA sequence features underlie pervasive RNA polymerase pausing." In: *Nucleic acids research* 49.8 (2021), pp. 4402–4420 (cit. on pp. 37, 40, 52, 58).

[70] Roland Gamsjaeger, Sarah R Webb, Janine M Lamonica, Andrew Billin, Gerd A Blobel, and Joel P Mackay. "Structural basis and specificity of acetylated transcription factor GATA1 recognition by BET family bromodomain protein Brd3." In: *Molecular and cellular biology* 31.13 (2011), pp. 2632–2640 (cit. on p. 10).

[71] Laurent Gatto. *Omics Data Analysis*. `https://uclouvain-cbio.github.io/WSBIM2122/sec-rnaseq.html`. Accessed: 2022-04-09. 2021 (cit. on p. 31).

[72] Ludwig Geistlinger et al. "Toward a gold standard for benchmarking gene set enrichment analysis." In: *Briefings in bioinformatics* 22.1 (2021), pp. 545–556 (cit. on pp. 31, 78).

[73] Marek Gierliński et al. "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment." In: *Bioinformatics* 31.22 (2015), pp. 3625–3630 (cit. on pp. 29, 67).

[74] Daniel A Gilchrist, Gilberto Dos Santos, David C Fargo, Bin Xie, Yuan Gao, Leping Li, and Karen Adelman. "Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation." In: *Cell* 143.4 (2010), pp. 540–551 (cit. on p. 6).

[75] Austin E Gillen, Tomomi M Yamamoto, Enos Kline, Jay R Hesselberth, and Peter Kabos. "Improvements to the HITS-CLIP protocol eliminate widespread mispriming artifacts." In: *BMC genomics* 17.1 (2016), pp. 1–11 (cit. on p. 52).

[76] Marc Gillespie et al. "The reactome pathway knowledgebase 2022." In: *Nucleic acids research* 50.D1 (2022), pp. D687–D692 (cit. on pp. 74, 78, 166).

[77] Paul Gilman, Steven Janzou, Darice Guittet, Janine Freeman, Nicholas DiOrio, Nathan Blair, Matthew Boyd, Ty Neises, Michael Wagner, et al. *PySAM (Python Wrapper for System Advisor Model" SAM")*. Tech. rep. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2019 (cit. on pp. 40, 165).

[78] DAVID S Gilmour and JOHN T Lis. "RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells." In: *Molecular and cellular biology* 6.11 (1986), pp. 3984–3989 (cit. on p. 8).

[79] Kira Glover-Cutter, Soojin Kim, Joaquin Espinosa, and David L Bentley. "RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes." In: *Nature structural & molecular biology* 15.1 (2008), pp. 71–78 (cit. on p. 124).

[80] Saskia Gressel, Björn Schwalb, and Patrick Cramer. "The pause-initiation limit restricts transcription activation in human cells." In: *Nature communications* 10.1 (2019), pp. 1–12 (cit. on pp. 53, 161).

[81] Xinye Han et al. “Roles of the BRD4 short isoform in phase separation and active gene transcription.” In: *Nature Structural & Molecular Biology* 27.4 (2020), pp. 333–341 (cit. on pp. 125, 126).

[82] Charles R Harris et al. “Array programming with NumPy.” In: *Nature* 585.7825 (2020), pp. 357–362 (cit. on pp. 40, 165).

[83] Harvard Chan Bioinformatics Core (HBC). *Introduction to DGE - ARCHIVED.* `https://hbctraining.github.io/DGE_workshop/lessons/05_DGE_DESeq2_analysis2.html`. Accessed: 2022-04-09. 2018 (cit. on p. 30).

[84] Nathaniel D Heintzman et al. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.” In: *Nature genetics* 39.3 (2007), pp. 311–318 (cit. on pp. 6, 115).

[85] Telmo Henriques, Benjamin S Scruggs, Michiko O Inouye, Ginger W Muse, Lucy H Williams, Adam B Burkholder, Christopher A Lavender, David C Fargo, and Karen Adelman. “Widespread transcriptional pausing and elongation control at enhancers.” In: *Genes & development* 32.1 (2018), pp. 26–41 (cit. on pp. 6, 117, 125).

[86] Lydia Herzel, Diana SM Ottoz, Tara Alpert, and Karla M Neugebauer. “Splicing and transcription touch base: co-transcriptional spliceosome assembly and function.” In: *Nature reviews Molecular cell biology* 18.10 (2017), pp. 637–650 (cit. on pp. 21, 42).

[87] Veronika A Herzog et al. “Thiol-linked alkylation of RNA to assess expression dynamics.” In: *Nature methods* 14.12 (2017), pp. 1198–1204 (cit. on pp. 18, 19, 52).

[88] Yutaka Hirose and James L Manley. “RNA polymerase II and the integration of nuclear events.” In: *Genes & development* 14.12 (2000), pp. 1415–1429 (cit. on p. 9).

[89] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. “Super-enhancers in the control of cell identity and disease.” In: *Cell* 155.4 (2013), pp. 934–947 (cit. on p. 11).

[90] Liming Hou, Yating Wang, Yu Liu, Nan Zhang, Ilya Shamovsky, Evgeny Nudler, Bin Tian, and Brian David Dynlacht. “Paf1C regulates RNA polymerase II progression by modulating elongation rate.” In: *Proceedings of the National Academy of Sciences* 116.29 (2019), pp. 14583–14592 (cit. on pp. 122, 123).

[91] Françoise S Howe, Harry Fischl, Struan C Murray, and Jane Mellor. “Is H3K4me3 instructive for transcription activation?” In: *Bioessays* 39.1 (2017), pp. 1–12 (cit. on p. 6).

[92] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. "RNA-Seq methods for transcriptome analysis." In: *Wiley Interdisciplinary Reviews: RNA* 8.1 (2017), e1364 (cit. on p. 13).

[93] Shibin Hu, Linna Peng, Congling Xu, Zhenning Wang, Aixia Song, and Fei Xavier Chen. "SPT5 stabilizes RNA polymerase II, orchestrates transcription cycles, and maintains the enhancer landscape." In: *Molecular Cell* 81.21 (2021), pp. 4425–4439 (cit. on pp. 122–124).

[94] Jim R Hughes et al. "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment." In: *Nature genetics* 46.2 (2014), pp. 205–212 (cit. on p. 17).

[95] Broad Institute. "Picard tools." In: (2016) (cit. on pp. 84, 93, 165).

[96] Friederike Itzen, Ann Katrin Greifenberg, Christian A Bösken, and Matthias Geyer. "Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation." In: *Nucleic acids research* 42.12 (2014), pp. 7577–7590 (cit. on p. 11).

[97] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community." In: *Genome biology* 17.1 (2016), pp. 1–11 (cit. on pp. 18, 104).

[98] Moon Kyoo Jang, Kazuki Mochizuki, Meisheng Zhou, Ho-Sang Jeong, John N Brady, and Keiko Ozato. "The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription." In: *Molecular cell* 19.4 (2005), pp. 523–534 (cit. on pp. 11, 122).

[99] Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. "Synthetic spike-in standards for RNA-seq experiments." In: *Genome research* 21.9 (2011), pp. 1543–1551 (cit. on pp. 26, 63, 70, 83).

[100] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. "Genome-wide mapping of in vivo protein-DNA interactions." In: *Science* 316.5830 (2007), pp. 1497–1502 (cit. on pp. 13, 27, 169).

[101] Michael H Jones, Mariko Numata, and Miyuki Shimane. "Identification and Characterization of BRDT: A Testis-Specific Gene Related to the Bromodomain Genes RING3 andDrosophila fsh." In: *Genomics* 45.3 (1997), pp. 529–534 (cit. on p. 10).

[102] Iris Jonkers and John T Lis. "Getting up to speed with transcription elongation by RNA polymerase II." In: *Nature reviews Molecular cell biology* 16.3 (2015), pp. 167–177 (cit. on pp. 1, 121).

[103] Julius Judd et al. "A rapid, sensitive, scalable method for precision run-on sequencing (PRO-seq)." In: *bioRxiv* (2020) (cit. on pp. 53, 169).

[104] Jerzy Jurka, Vladimir V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. "Repbase Update, a database of eukaryotic repetitive elements." In: *Cytogenetic and genome research* 110.1-4 (2005), pp. 462–467 (cit. on pp. 166, 168).

[105] Kinga Kamieniarz-Gdula et al. "Selective roles of vertebrate PCF11 in premature and full-length transcript termination." In: *Molecular cell* 74.1 (2019), pp. 158–172 (cit. on pp. 8–10, 45, 63, 78, 124).

[106] Syuzo Kaneko, Orit Rozenblatt-Rosen, Matthew Meyerson, and James L Manley. "The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3′ processing and transcription termination." In: *Genes & development* 21.14 (2007), pp. 1779–1789 (cit. on p. 10).

[107] Tomohiko Kanno et al. "BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones." In: *Nature structural & molecular biology* 21.12 (2014), pp. 1047–1057 (cit. on pp. 11, 82).

[108] Søren M Karst, Ryan M Ziels, Rasmus H Kirkegaard, Emil A Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. "High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing." In: *Nature methods* 18.2 (2021), pp. 165–169 (cit. on p. 124).

[109] Tea Kecman et al. "Elongation/termination factor exchange mediated by PP1 phosphatase orchestrates transcription termination." In: *Cell reports* 25.1 (2018), pp. 259–269 (cit. on pp. 92, 127).

[110] Tae-Kyung Kim et al. "Widespread transcription at neuronal activity-regulated enhancers." In: *Nature* 465.7295 (2010), pp. 182–187 (cit. on pp. 6, 117).

[111] Tamás Kiss. "Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions." In: *Cell* 109.2 (2002), pp. 145–148 (cit. on p. 73).

[112] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. "Chromatin accessibility and the regulatory epigenome." In: *Nature Reviews Genetics* 20.4 (2019), pp. 207–220 (cit. on p. 5).

[113] Tsuyoshi Konuma, Di Yu, Chengcheng Zhao, Ying Ju, Rajal Sharma, Chunyan Ren, Qiang Zhang, Ming-Ming Zhou, and Lei Zeng. "Structural mechanism of the oxygenase JMJD6 recognition by the extraterminal (ET) domain of BRD4." In: *Scientific reports* 7.1 (2017), pp. 1–10 (cit. on p. 10).

[114] Timo Koski. *Hidden Markov models for bioinformatics*. Vol. 2. Springer Science & Business Media, 2001 (cit. on p. 87).

[115] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. "miRBase: from microRNA sequences to function." In: *Nucleic acids research* 47.D1 (2019), pp. D155–D162 (cit. on pp. 165, 168).

[116] Amelie J Kraus, Benedikt G Brink, and T Nicolai Siegel. "Efficient and specific oligo-based depletion of rRNA." In: *Scientific reports* 9.1 (2019), pp. 1–8 (cit. on p. 7).

[117] Hojoong Kwak, Nicholas J Fuda, Leighton J Core, and John T Lis. "Precise maps of RNA polymerase reveal how promoters direct initiation and pausing." In: *Science* 339.6122 (2013), pp. 950–953 (cit. on pp. 9, 58).

[118] Stephen G Landt et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." In: *Genome research* 22.9 (2012), pp. 1813–1831 (cit. on pp. 13, 27).

[119] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." In: *Nature methods* 9.4 (2012), pp. 357–359 (cit. on pp. 21, 84, 89, 164).

[120] Ji-Eun Lee et al. "Brd4 binds to active enhancers to control cell identity gene induction in adipogenesis and myogenesis." In: *Nature communications* 8.1 (2017), pp. 1–12 (cit. on p. 11).

[121] Tong Ihn Lee and Richard A Young. "Transcription of eukaryotic protein-coding genes." In: *Annual review of genetics* 34.1 (2000), pp. 77–137 (cit. on pp. 7, 8).

[122] Bo Li and Colin N Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." In: *BMC bioinformatics* 12.1 (2011), pp. 1–16 (cit. on pp. 39, 166).

[123] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." In: *arXiv preprint arXiv:1303.3997* (2013) (cit. on p. 21).

[124] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences." In: *Bioinformatics* 34.18 (2018), pp. 3094–3100 (cit. on pp. 87, 165).

[125] Yang Liao, Gordon K Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." In: *Bioinformatics* 30.7 (2014), pp. 923–930 (cit. on pp. 143, 166).

[126] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *science* 326.5950 (2009), pp. 289–293 (cit. on pp. 16, 17).

[127] RP Lifton, ML Goldberg, RW Karp, and DS Hogness. "The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications." In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 42. Cold Spring Harbor Laboratory Press. 1978, pp. 1047–1051 (cit. on p. 6).

[128] Ricardo Linares-Saldana et al. "BRD4 orchestrates genome folding to promote neural crest differentiation." In: *Nature Genetics* 53.10 (2021), pp. 1480–1492 (cit. on pp. 11, 126).

[129] Thomas J Lindell, Fanyela Weinberg, Paul W Morris, Robert G Roeder, and William J Rutter. "Specific inhibition of nuclear RNA polymerase II by α-amanitin." In: *Science* 170.3956 (1970), pp. 447–449 (cit. on p. 14).

[130] Wen Liu, Qi Ma, Kaki Wong, Wenbo Li, Kenny Ohgi, Jie Zhang, Aneel K Aggarwal, and Michael G Rosenfeld. "Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release." In: *Cell* 155.7 (2013), pp. 1581–1595 (cit. on p. 11).

[131] John Logan, Erik Falck-Pedersen, James E Darnell, and Thomas Shenk. "A poly (A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene." In: *Proceedings of the National Academy of Sciences* 84.23 (1987), pp. 8306–8310 (cit. on p. 10).

[132] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome biology* 15.12 (2014), pp. 1–21 (cit. on pp. 25, 28, 63, 67, 77, 84, 164).

[133] Jakob Lovén, Heather A Hoke, Charles Y Lin, Ashley Lau, David A Orlando, Christopher R Vakoc, James E Bradner, Tong Ihn Lee, and Richard A Young. "Selective inhibition of tumor oncogenes by disruption of super-enhancers." In: *Cell* 153.2 (2013), pp. 320–334 (cit. on pp. 11, 81).

[134] Jakob Lovén, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. "Revisiting global gene expression analysis." In: *Cell* 151.3 (2012), pp. 476–482 (cit. on pp. 63, 70).

[135] Weifei Luo, Arlen W Johnson, and David L Bentley. "The role of Rat1 in coupling mRNA 3′-end processing to transcription termination: implications for a unified allosteric–torpedo model." In: *Genes & development* 20.8 (2006), pp. 954–965 (cit. on p. 10).

[136] Dig Bijay Mahat et al. "Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq)." In: *Nature protocols* 11.8 (2016), pp. 1455–1476 (cit. on pp. 14, 16, 55, 58, 169).

[137] Shaun Mahony and B Franklin Pugh. "Protein–DNA binding in high-resolution." In: *Critical reviews in biochemistry and molecular biology* 50.4 (2015), pp. 269–283 (cit. on p. 14).

[138] W Makałowski, Jinghui Zhang, and Mark S Boguski. "Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences." In: *Genome research* 6.9 (1996), pp. 846–857 (cit. on p. 71).

[139] Udi Manber and Gene Myers. "Suffix arrays: a new method for on-line string searches." In: *siam Journal on Computing* 22.5 (1993), pp. 935–948 (cit. on p. 21).

[140] Corey R Mandel, Syuzo Kaneko, Hailong Zhang, Damara Gebauer, Vasupradha Vethantham, James L Manley, and Liang Tong. "Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease." In: *Nature* 444.7121 (2006), pp. 953–956 (cit. on pp. 9, 10).

[141] Thanasis Margaritis and Frank CP Holstege. "Poised RNA polymerase II gives pause for thought." In: *Cell* 133.4 (2008), pp. 581–584 (cit. on pp. 1, 121).

[142] Nick F Marshall and David H Price. "Purification of P-TEFb, a Transcription Factor Required for the Transition into Productive Elongation (*)." In: *Journal of Biological Chemistry* 270.21 (1995), pp. 12335–12338 (cit. on p. 8).

[143] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads." In: *EMBnet. journal* 17.1 (2011), pp. 10–12 (cit. on pp. 40, 164).

[144] Fuensanta W Martinez-Rucobo and Patrick Cramer. "Structural basis of transcription elongation." In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1829.1 (2013), pp. 9–19 (cit. on p. 7).

[145] Wolfram MathWorld. *Negative Binomial Distribution*. `https://mathworld.wolfram.com/NegativeBinomialDistribution.html`. Accessed: 2022-04-08. 2021 (cit. on p. 28).

[146] Andreas Mayer and L Stirling Churchman. "Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing." In: *Nature protocols* 11.4 (2016), pp. 813–833 (cit. on pp. 14, 37, 47, 49, 73).

[147] Andreas Mayer and L Stirling Churchman. "A Detailed Protocol for Subcellular RNA Sequencing (subRNA-seq)." In: *Current protocols in molecular biology* 120.1 (2017), pp. 4–29 (cit. on p. 13).

[148] Andreas Mayer, Julia Di Iulio, Seth Maleri, Umut Eser, Jeff Vierstra, Alex Reynolds, Richard Sandstrom, John A Stamatoyannopoulos, and L Stirling Churchman. "Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution." In: *Cell* 161.3 (2015), pp. 541–554 (cit. on pp. 14, 37, 40, 47, 57, 59, 77, 161, 169).

[149] Andreas Mayer, Amelie Schreieck, Michael Lidschreiber, Kristin Leike, Dietmar E Martin, and Patrick Cramer. "The spt5 C-terminal region recruits yeast 3′ RNA cleavage factor I." In: *Molecular and cellular biology* 32.7 (2012), pp. 1321–1331 (cit. on p. 124).

[150] James D McGhee and Gary Felsenfeld. "Nucleosome structure." In: *Annual review of biochemistry* 49.1 (1980), pp. 1115–1156 (cit. on p. 5).

[151] Jane Mellor et al. "Mapping Human Transient Transcriptomes Using Single Nucleotide Resolution 4sU Sequencing (SNU-Seq)." In: *bioRxiv* (2021) (cit. on pp. 53, 57, 169).

[152] Carlos A Melo et al. "eRNAs are required for p53-dependent enhancer activity and gene transcription." In: *Molecular cell* 49.3 (2013), pp. 524–535 (cit. on p. 6).

[153] Huaiyu Mi, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools." In: *Nucleic acids research* 47.D1 (2019), pp. D419–D426 (cit. on pp. 31, 32, 78, 97, 165).

[154] Olga Mikhaylichenko, Vladyslav Bondarenko, Dermot Harnett, Ignacio E Schor, Matilda Males, Rebecca R Viales, and Eileen EM Furlong. "The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription." In: *Genes & development* 32.1 (2018), pp. 42–57 (cit. on p. 6).

[155] Tarjei S Mikkelsen et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." In: *Nature* 448.7153 (2007), pp. 553–560 (cit. on p. 13).

[156] Felix Mölder et al. "Sustainable data analysis with Snakemake." In: *F1000Research* 10 (2021) (cit. on pp. 40, 166).

[157] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." In: *Nature methods* 5.7 (2008), pp. 621–628 (cit. on pp. 24, 67, 77).

[158] Matthias Muhar et al. "SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis." In: *Science* 360.6390 (2018), pp. 800–805 (cit. on pp. 11, 81, 121, 122).

[159] Maxwell R Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavari, William J Greenleaf, and Howard Y Chang. "HiChIP: efficient and sensitive analysis of protein-directed genome architecture." In: *Nature methods* 13.11 (2016), pp. 919–922 (cit. on pp. 17, 117).

[160] Ginger W Muse, Daniel A Gilchrist, Sergei Nechaev, Ruchir Shah, Joel S Parker, Sherry F Grissom, Julia Zeitlinger, and Karen Adelman. "RNA polymerase is poised for activation across the genome." In: *Nature genetics* 39.12 (2007), pp. 1507–1511 (cit. on pp. 1, 8, 44, 55, 121).

[161] Ryan Musich, Lance Cadle-Davidson, and Michael V Osier. "Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider." In: *Frontiers in Plant Science* 12 (2021) (cit. on p. 22).

[162] Nasrinsadat Nabavizadeh et al. "A progeroid syndrome caused by a deep intronic variant in TAPT1 is revealed by RNA/SI-NET sequencing." In: *EMBO Molecular Medicine* (2023), e16478 (cit. on pp. v, 34).

[163] Behnam Nabet et al. "The dTAG system for immediate and target-specific protein degradation." In: *Nature chemical biology* 14.5 (2018), pp. 431–441 (cit. on pp. 19, 20, 34).

[164] Sankari Nagarajan et al. "Bromodomain protein BRD4 is required for estrogen receptor-dependent enhancer activation and gene transcription." In: *Cell reports* 8.2 (2014), pp. 460–469 (cit. on pp. 11, 82).

[165] Ashwin Narain et al. "Targeted protein degradation reveals a direct role of SPT6 in RNAPII elongation and termination." In: *Molecular cell* 81.15 (2021), pp. 3110–3127 (cit. on pp. 8, 122, 124).

[166] Iman Nazari, Hilal Tayara, and Kil To Chong. "Branch point selection in RNA splicing using deep learning." In: *IEEE Access* 7 (2018), pp. 1800–1807 (cit. on p. 132).

[167] Edwige Nicodeme et al. "Suppression of inflammation by a synthetic histone mimic." In: *Nature* 468.7327 (2010), pp. 1119–1123 (cit. on p. 11).

[168] Melvin Noe Gonzalez, Daniel Blears, and Jesper Q Svejstrup. "Causes and consequences of RNA polymerase II stalling during transcript elongation." In: *Nature Reviews Molecular Cell Biology* 22.1 (2021), pp. 3–21 (cit. on pp. 8, 9).

[169] Takayuki Nojima, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J Proudfoot. "Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing." In: *Cell* 161.3 (2015), pp. 526–540 (cit. on pp. 14, 15, 58, 59, 124, 169).

[170] David A Orlando, Mei Wei Chen, Victoria E Brown, Snehakumari Solanki, Yoon J Choi, Eric R Olson, Christian C Fritz, James E Bradner, and Matthew G Guenther. "Quantitative ChIP-Seq normalization reveals global modulation of the epigenome." In: *Cell reports* 9.3 (2014), pp. 1163–1170 (cit. on pp. 13, 14, 26, 63, 70, 77, 106).

[171] Joseph M Paggi and Gill Bejerano. "A sequence-based, deep learning model accurately predicts RNA splicing branchpoints." In: *RnA* 24.12 (2018), pp. 1647–1658 (cit. on p. 132).

[172] Pabitra K Parua, Sampada Kalan, Bradley Benjamin, Miriam Sansó, and Robert P Fisher. "Distinct Cdk9-phosphatase switches act at the beginning and end of elongation by RNA polymerase II." In: *Nature communications* 11.1 (2020), pp. 1–13 (cit. on p. 124).

[173] B Matija Peterlin and David H Price. "Controlling the elongation phase of transcription with P-TEFb." In: *Molecular cell* 23.3 (2006), pp. 297–305 (cit. on pp. 8, 11, 81, 122).

[174] Allison Piovesan, Maria Chiara Pelleri, Francesca Antonaros, Pierluigi Strippoli, Maria Caracausi, and Lorenza Vitale. "On the length, weight and GC content of the human genome." In: *BMC research notes* 12.1 (2019), pp. 1–7 (cit. on p. 16).

[175] Odil Porrua and Domenico Libri. "Transcription termination and the control of the transcriptome: why, where and how to stop." In: *Nature reviews Molecular cell biology* 16.3 (2015), pp. 190–202 (cit. on p. 108).

[176] Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, and Robert D Finn. "HMMER web server: 2018 update." In: *Nucleic acids research* 46.W1 (2018), W200–W204 (cit. on pp. 87, 165).

[177] Sue Povey, Ruth Lovering, Elspeth Bruford, Mathew Wright, Michael Lush, and Hester Wain. "The HUGO gene nomenclature committee (HGNC)." In: *Human genetics* 109.6 (2001), pp. 678–680 (cit. on pp. 39, 164).

[178] Nick J Proudfoot. "Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut." In: *Science* 352.6291 (2016), aad9926 (cit. on p. 1).

[179] NJ Proudfoot and GG Brownlee. "Sequence at the 3′ end of globin mRNA shows homology with immunoglobulin light chain mRNA." In: *Nature* 252.5482 (1974), pp. 359–362 (cit. on p. 10).

[180] NJ Proudfoot and GG Brownlee. "3′ non-coding region sequences in eukaryotic messenger RNA." In: *Nature* 263.5574 (1976), pp. 211–214 (cit. on p. 10).

[181] Yuri Prozzillo, Gaia Fattorini, Maria Virginia Santopietro, Luigi Suglia, Alessandra Ruggiero, Diego Ferreri, and Giovanni Messina. "Targeted protein degradation tools: overview and future perspectives." In: *Biology* 9.12 (2020), p. 421 (cit. on p. 19).

[182] Pedro Prudêncio, Kenny Rebelo, Ana Rita Grosso, Rui Gonçalo Martinho, and Maria Carmo-Fonseca. "Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data." In: *Methods* 178 (2020), pp. 89–95 (cit. on p. 58).

[183] Mark Ptashne and Alexander Gann. "Transcriptional activation by recruitment." In: *Nature* 386.6625 (1997), pp. 569–577 (cit. on pp. 1, 121).

[184] Givanna H Putri, Simon Anders, Paul Theodor Pyl, John E Pimanda, and Fabio Zanini. "Analysing high-throughput sequencing data in Python with HTSeq 2.0." In: *arXiv preprint arXiv:2112.00939* (2021) (cit. on pp. 83, 165).

[185] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." In: *Bioinformatics* 26.6 (2010), pp. 841–842 (cit. on pp. 42, 164).

[186] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/` (cit. on p. 165).

[187] Marijana Radonjic, Jean-Christophe Andrau, Philip Lijnzaad, Patrick Kemmeren, Thessa TJP Kockelkorn, Dik van Leenen, Nynke L van Berkum, and Frank CP Holstege. "Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit." In: *Molecular cell* 18.2 (2005), pp. 171–183 (cit. on pp. 1, 8, 121).

[188] Peter B Rahl, Charles Y Lin, Amy C Seila, Ryan A Flynn, Scott McCuine, Christopher B Burge, Phillip A Sharp, and Richard A Young. "c-Myc regulates transcriptional pause release." In: *Cell* 141.3 (2010), pp. 432–445 (cit. on p. 8).

[189] Shaila Rahman, Mathew E Sowa, Matthias Ottinger, Jennifer A Smith, Yang Shi, J Wade Harper, and Peter M Howley. "The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3." In: *Molecular and cellular biology* 31.13 (2011), pp. 2641–2652 (cit. on p. 10).

[190] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. "deepTools2: a next generation web server for deep-sequencing data analysis." In: *Nucleic acids research* 44.W1 (2016), W160–W165 (cit. on pp. 46, 164).

[191] Nikos B Reppas, Joseph T Wade, George M Church, and Kevin Struhl. "The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting." In: *Molecular cell* 24.5 (2006), pp. 747–757 (cit. on pp. 44, 55).

[192] Jason A Reuter, Damek V Spacek, and Michael P Snyder. "High-throughput sequencing technologies." In: *Molecular cell* 58.4 (2015), pp. 586–597 (cit. on p. 12).

[193] Patricia Richard and James L Manley. “Transcription termination by nuclear RNA polymerases.” In: *Genes & development* 23.11 (2009), pp. 1247–1269 (cit. on pp. 9, 108).

[194] Christian Riml, Thomas Amort, Dietmar Rieder, Catherina Gasser, Alexandra Lusser, and Ronald Micura. “Osmium-mediated transformation of 4-thiouridine to cytidine as key to study RNA dynamics by sequencing.” In: *Angewandte Chemie International Edition* 56.43 (2017), pp. 13479–13483 (cit. on p. 19).

[195] Rieke Ringel, Marina Sologub, Yaroslav I Morozov, Dmitry Litonin, Patrick Cramer, and Dmitry Temiakov. “Structure of human mitochondrial RNA polymerase.” In: *Nature* 478.7368 (2011), pp. 269–273 (cit. on p. 7).

[196] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. “The role of spike-in standards in the normalization of RNA-seq.” In: *Statistical Analysis of Next Generation Sequencing Data*. Springer, 2014, pp. 169–190 (cit. on p. 70).

[197] Gordon Robertson et al. “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.” In: *Nature methods* 4.8 (2007), pp. 651–657 (cit. on p. 13).

[198] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. “Integrative genomics viewer.” In: *Nature biotechnology* 29.1 (2011), pp. 24–26 (cit. on p. 165).

[199] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” In: *Bioinformatics* 26.1 (2010), pp. 139–140 (cit. on pp. 28, 63, 67, 77).

[200] Peter J Rousseeuw and Christophe Croux. “Alternatives to the median absolute deviation.” In: *Journal of the American Statistical association* 88.424 (1993), pp. 1273–1283 (cit. on p. 68).

[201] Orit Rozenblatt-Rosen, Takashi Nagaike, Joshua M Francis, Syuzo Kaneko, Karen A Glatt, Christina M Hughes, Thomas LaFramboise, James L Manley, and Matthew Meyerson. “The tumor suppressor Cdc73 functionally associates with CPSF and CstF 3′ mRNA processing factors.” In: *Proceedings of the National Academy of Sciences* 106.3 (2009), pp. 755–760 (cit. on p. 125).

[202] Kevin Ryan, Olga Calvo, and James L Manley. “Evidence that polyadenylation factor CPSF-73 is the mRNA 3′ processing endonuclease.” In: *Rna* 10.4 (2004), pp. 565–573 (cit. on pp. 9, 10).

[203] Benjamin R Sabari et al. "Coactivator condensation at super-enhancers links phase separation and gene control." In: *Science* 361.6400 (2018), eaar3958 (cit. on pp. 11, 126).

[204] Miriam Sansó et al. "P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates." In: *Genes & development* 30.1 (2016), pp. 117–131 (cit. on p. 8).

[205] Matthieu Schapira, Matthew F Calabrese, Alex N Bullock, and Craig M Crews. "Targeted protein degradation: expanding the toolbox." In: *Nature reviews Drug discovery* 18.12 (2019), pp. 949–963 (cit. on p. 20).

[206] Ilari Scheinin et al. "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly." In: *Genome research* 24.12 (2014), pp. 2022–2032 (cit. on pp. 91, 165).

[207] Stefan Schoenfelder and Peter Fraser. "Long-range enhancer–promoter contacts in gene expression control." In: *Nature Reviews Genetics* 20.8 (2019), pp. 437–455 (cit. on p. 6).

[208] Jeremy A Schofield, Erin E Duffy, Lea Kiefer, Meaghan C Sullivan, and Matthew D Simon. "TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding." In: *Nature methods* 15.3 (2018), pp. 221–225 (cit. on p. 19).

[209] Lars Schönemann, Uwe Kühn, Georges Martin, Peter Schäfer, Andreas R Gruber, Walter Keller, Mihaela Zavolan, and Elmar Wahle. "Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33." In: *Genes & development* 28.21 (2014), pp. 2381–2393 (cit. on pp. 9, 10).

[210] Nicholas J Schurch et al. "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" In: *Rna* 22.6 (2016), pp. 839–851 (cit. on pp. 29, 58, 63).

[211] Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. "TT-seq maps the human transient transcriptome." In: *Science* 352.6290 (2016), pp. 1225–1228 (cit. on pp. 8, 18, 44, 53, 55, 92, 169).

[212] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing." In: *Genome biology* 16.1 (2015), pp. 1–11 (cit. on pp. 89, 117, 165).

[213] Jayasha Shandilya and Stefan GE Roberts. "The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling." In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1819.5 (2012), pp. 391–400 (cit. on p. 7).

[214] Keith E Shearwin, Benjamin P Callen, and J Barry Egan. "Transcriptional interference–a crash course." In: *TRENDS in Genetics* 21.6 (2005), pp. 339–345 (cit. on p. 1).

[215] Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data." In: *Proceedings of the National Academy of Sciences* 111.51 (2014), E5593–E5601 (cit. on pp. 131, 156, 166).

[216] Yin Shen et al. "A map of the cis-regulatory sequences in the mouse genome." In: *Nature* 488.7409 (2012), pp. 116–120 (cit. on p. 6).

[217] Tatiana Shorstova, William D Foulkes, and Michael Witcher. "Achieving clinical success with BET inhibitors as anti-cancer agents." In: *British Journal of Cancer* 124.9 (2021), pp. 1478–1490 (cit. on pp. 2, 11).

[218] Botond Sipos, Philipp Rescheneder, and Stephen Rudd. *Pychopper*. `https://github.com/nanoporetech/pychopper`. Accessed: 2020-03-18. 2021 (cit. on p. 85).

[219] Stephen T Smale and David Baltimore. "The "initiator" as a transcription control element." In: *Cell* 57.1 (1989), pp. 103–113 (cit. on p. 6).

[220] John C Stansfield, Kellen G Cresswell, Vladimir I Vladimirov, and Mikhail G Dozmorov. "HiCcompare: an R-package for joint normalization and comparison of HI-C datasets." In: *BMC bioinformatics* 19.1 (2018), pp. 1–10 (cit. on pp. 91, 117, 118, 165).

[221] Rory Stark, Gordon Brown, et al. "DiffBind: differential binding analysis of ChIP-Seq peak data." In: *R package version* 100.4.3 (2011) (cit. on pp. 63, 84, 107, 144, 164).

[222] Kevin Struhl. "Transcriptional noise and the fidelity of initiation by RNA polymerase II." In: *Nature structural & molecular biology* 14.2 (2007), pp. 103–105 (cit. on p. 6).

[223] Yadong Sun, Keith Hamilton, and Liang Tong. "Recent molecular insights into canonical pre-mRNA 3′-end processing." In: *Transcription* 11.2 (2020), pp. 83–96 (cit. on pp. 9, 10).

[224] Sofie Symoens et al. "Genetic defects in TAPT1 disrupt ciliogenesis and cause a complex lethal osteochondrodysplasia." In: *The American Journal of Human Genetics* 97.4 (2015), pp. 521–534 (cit. on p. 131).

[225] Paul B Talbert and Steven Henikoff. "The Yin and Yang of Histone Marks in Transcription." In: *Annual review of genomics and human genetics* 22 (2021), pp. 147–170 (cit. on p. 6).

[226] C Tellechea. "Chemfig: A TEX package for drawing molecules, version 1.6b, Aug. 1, 2021." In: *Available:(visited on 12/22/2021)* () (cit. on pp. 18, 164).

[227] Michael Tellier et al. "CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation." In: *Nucleic acids research* 48.14 (2020), pp. 7712–7727 (cit. on pp. 45, 63, 78).

[228] Jacob M Tome, Nathaniel D Tippens, and John T Lis. "Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers." In: *Nature genetics* 50.11 (2018), pp. 1533–1541 (cit. on p. 58).

[229] Harmen JG Van De Werken et al. "Robust 4C-seq data analysis to screen for regulatory DNA interactions." In: *Nature methods* 9.10 (2012), pp. 969–972 (cit. on p. 16).

[230] Erwin L Van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. "The third revolution in sequencing technology." In: *Trends in Genetics* 34.9 (2018), pp. 666–681 (cit. on p. 18).

[231] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020 (cit. on p. 165).

[232] Seychelle M Vos, Lucas Farnung, Marc Boehning, Christoph Wigge, Andreas Linden, Henning Urlaub, and Patrick Cramer. "Structure of activated transcription complex Pol II–DSIF–PAF–SPT6." In: *Nature* 560.7720 (2018), pp. 607–612 (cit. on pp. 8, 110, 122).

[233] Seychelle M Vos, Lucas Farnung, Andreas Linden, Henning Urlaub, and Patrick Cramer. "Structure of complete Pol II–DSIF–PAF–SPT6 transcription complex reveals RTF1 allosteric activation." In: *Nature Structural & Molecular Biology* 27.7 (2020), pp. 668–677 (cit. on pp. 8, 110, 122).

[234] Seychelle M Vos, Lucas Farnung, Henning Urlaub, and Patrick Cramer. "Structure of paused transcription complex Pol II–DSIF–NELF." In: *Nature* 560.7720 (2018), pp. 601–606 (cit. on p. 8).

[235] Leonhard Wachutka and Julien Gagneur. "Measures of RNA metabolism rates: Toward a definition at the level of single bonds." In: *Transcription* 8.2 (2017), pp. 75–80 (cit. on p. 18).

[236] Tadashi Wada et al. "DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs." In: *Genes & development* 12.3 (1998), pp. 343–356 (cit. on p. 8).

[237] Günter P Wagner, Koryu Kin, and Vincent J Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." In: *Theory in biosciences* 131.4 (2012), pp. 281–285 (cit. on pp. 25, 67, 77).

[238] Abraham Wald. “Contributions to the theory of statistical estimation and testing hypotheses.” In: *The Annals of Mathematical Statistics* 10.4 (1939), pp. 299–326 (cit. on p. 30).

[239] Jing Wang, Xizhen Dai, Lynne D Berry, Joy D Cogan, Qi Liu, and Yu Shyr. “HACER: an atlas of human active enhancers to interpret regulatory variants.” In: *Nucleic acids research* 47.D1 (2019), pp. D106–D112 (cit. on p. 40).

[240] Jing Wang, Yue Zhao, Xiaofan Zhou, Scott W Hiebert, Qi Liu, and Yu Shyr. “Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation.” In: *BMC genomics* 19.1 (2018), pp. 1–18 (cit. on p. 40).

[241] Nian Wang, Runliu Wu, Daolin Tang, and Rui Kang. “The BET family in immunity and disease.” In: *Signal transduction and targeted therapy* 6.1 (2021), pp. 1–22 (cit. on p. 11).

[242] Ranran Wang, Qing Li, Christine M Helfer, Jing Jiao, and Jianxin You. “Bromodomain protein Brd4 associated with acetylated chromatin is important for maintenance of higher-order chromatin structure.” In: *Journal of Biological Chemistry* 287.14 (2012), pp. 10738–10752 (cit. on p. 11).

[243] Ruijia Wang, Ram Nambiar, Dinghai Zheng, and Bin Tian. “PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes.” In: *Nucleic acids research* 46.D1 (2018), pp. D315–D319 (cit. on pp. 94, 164).

[244] Robert H Waterston and Lior Pachter. “Initial sequencing and comparative analysis of the mouse genome.” In: *Nature* 420.6915 (2002), pp. 520–562 (cit. on p. 71).

[245] James D Watson and Francis HC Crick. “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid.” In: *Nature* 171.4356 (1953), pp. 737–738 (cit. on p. 5).

[246] Abraham S Weintraub et al. “YY1 is a structural regulator of enhancer-promoter loops.” In: *Cell* 171.7 (2017), pp. 1573–1588 (cit. on p. 17).

[247] Jocelyn D Weissman, Amit K Singh, Ballachanda N Devaiah, Peter Schuck, Ross C LaRue, and Dinah S Singer. “The intrinsic kinase activity of BRD4 spans its BD2-B-BID domains.” In: *Journal of Biological Chemistry* 297.5 (2021) (cit. on p. 11).

[248] Steven West, Nicholas J Proudfoot, and Michael J Dye. “Molecular dissection of mammalian RNA polymerase II transcriptional termination.” In: *Molecular cell* 29.5 (2008), pp. 600–610 (cit. on p. 10).

[249] Georg E Winter et al. "BET bromodomain proteins function as master transcription elongation factors independent of CDK9 recruitment." In: *Molecular cell* 67.1 (2017), pp. 5–18 (cit. on pp. 11, 20, 57, 63, 66, 68, 70, 76, 77, 81, 97, 99, 103, 122, 125, 163, 167).

[250] Erin M Wissink, Anniina Vihervaara, Nathaniel D Tippens, and John T Lis. "Nascent RNA analyses: tracking transcription and its regulation." In: *Nature Reviews Genetics* 20.12 (2019), pp. 705–723 (cit. on p. 57).

[251] Yuki Yamaguchi, Toshiyuki Takagi, Tadashi Wada, Keiichi Yano, Akiko Furuya, Seiji Sugimoto, Jun Hasegawa, and Hiroshi Handa. "NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation." In: *Cell* 97.1 (1999), pp. 41–51 (cit. on p. 8).

[252] Zhiyuan Yang, Jasper HN Yik, Ruichuan Chen, Nanhai He, Moon Kyoo Jang, Keiko Ozato, and Qiang Zhou. "Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4." In: *Molecular cell* 19.4 (2005), pp. 535–545 (cit. on pp. 11, 122).

[253] Ming Yu, Wenjing Yang, Ting Ni, Zhanyun Tang, Tomoyoshi Nakadai, Jun Zhu, and Robert G Roeder. "RNA polymerase II–associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II." In: *Science* 350.6266 (2015), pp. 1383–1386 (cit. on pp. 84, 110, 122, 123, 151, 160, 163, 167).

[254] Julia Zeitlinger, Alexander Stark, Manolis Kellis, Joung-Woo Hong, Sergei Nechaev, Karen Adelman, Michael Levine, and Richard A Young. "RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo." In: *Nature genetics* 39.12 (2007), pp. 1512–1516 (cit. on pp. 1, 8, 44, 55, 121).

[255] Lei Zeng and Ming-Ming Zhou. "Bromodomain: an acetyl-lysine binding domain." In: *FEBS letters* 513.1 (2002), pp. 124–128 (cit. on pp. 10, 115).

[256] Qing Zhang, Xiaodan Fan, Yejun Wang, Ming-an Sun, Jianlin Shao, and Dianjing Guo. "BPP: a sequence-based algorithm for branch point prediction." In: *Bioinformatics* 33.20 (2017), pp. 3166–3172 (cit. on p. 132).

[257] Yixiao Zhang, Yadong Sun, Yongsheng Shi, Thomas Walz, and Liang Tong. "Structural insights into the human pre-mRNA 3′-end processing machinery." In: *Molecular cell* 77.4 (2020), pp. 800–809 (cit. on p. 10).

[258] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)." In: *Genome biology* 9.9 (2008), pp. 1–9 (cit. on pp. 27, 28, 84, 165).

[259] Zhiqiang Zhang and David S Gilmour. "Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript." In: *Molecular cell* 21.1 (2006), pp. 65–74 (cit. on pp. 9, 10).

[260] Bin Zheng et al. "Acute perturbation strategies in interrogating RNA polymerase II elongation factor function in gene expression." In: *Genes & development* 35.3-4 (2021), pp. 273–285 (cit. on pp. 11, 81, 82, 106, 115, 121, 122, 125).

[261] Eduard Zorita, Pol Cusco, and Guillaume J Filion. "Starcode: sequence clustering based on all-pairs search." In: *Bioinformatics* 31.12 (2015), pp. 1913–1919 (cit. on pp. 40, 166).

[262] Kristina Žumer, Kerstin C Maier, Lucas Farnung, Martin G Jaeger, Petra Rus, Georg Winter, and Patrick Cramer. "Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo." In: *Molecular Cell* 81.15 (2021), pp. 3096–3109 (cit. on pp. 8, 9, 77, 78, 122, 123).

LIST OF FIGURES

Figure 2.1 Pol II Transcription Cycle 7
Figure 2.2 Overview of the Human NET-seq Method 15
Figure 2.3 Overview of the HiChIP Method 17
Figure 2.4 Metabolic Labeling using 4-thiouridine 18
Figure 2.5 Protein Degradation Strategies 19
Figure 3.1 HTS Data Example 22
Figure 3.2 HTS Data Normalization 23
Figure 6.1 Scheme of NET-seq Data Processing 41
Figure 6.2 Definition of Pol II Indices 43
Figure 6.3 Calculation of Meta-gene Profiles and Heatmaps 45
Figure 7.1 Adaptation of Human NET-seq in the K562 Cell Line. . 48
Figure 7.2 Overview of the HiS-NET-seq Method 50
Figure 7.3 Pol II Coverage Gain using 4sU Labeling 51
Figure 7.4 Correlation Between Pol II Profiling Methods 54
Figure 7.5 Comparison of Pol II Profiling Methods 56
Figure 11.1 Parameter Estimation for Differential Analysis (dBET6) 69
Figure 11.2 Cross-contamination of a Joint Reference Genome . . . 72
Figure 11.3 Results of DPO Analysis using SI-NET-seq Data 74
Figure 14.1 Read Types from ONT Data 86
Figure 14.2 Read Type Classification Example 90
Figure 14.3 Splicing Analysis Example 93
Figure 14.4 3′-RNA Cleavage Efficiency Test Example 94
Figure 15.1 BRD4 Degradation Decreases Nascent Transcription . . 98
Figure 15.2 Disruption of Pol II Pause Release by BRD4 100
Figure 15.3 BRD4 Degradation Induces Readthrough Transcription 102
Figure 15.4 BET Proteins are Required for 3′-RNA Cleavage 105
Figure 15.5 Features of BRD4 Binding and Readthrough Genes . . 107
Figure 15.6 BRD4 Recruits 3′-RNA Processing Factors 109
Figure 15.7 Functional Interactions of BRD4 111
Figure 15.8 PAF1 Recruitment Defect 112
Figure 15.9 BRD4 Interactors Contribute to Transcriptional Defects 113
Figure 15.10 BRD4-associated Enhancer Regions 114
Figure 15.11 Disruption of Productive Enhancer Transcription 116
Figure 15.12 Reduction of Regulatory Contacts upon BRD4 Loss . . 118
Figure A.1 Optimization of Human NET-seq 133
Figure A.2 Impact of 4sU Labeling on Purified RNA Libraries . . . 134
Figure A.3 Parameter Estimation for Differential Analysis (OI) . . 135
Figure A.4 Data Visualization and Interpretation 136
Figure A.5 BRD4-dependent Deregulation of Mature RNA Levels . 137

Figure A.6 BRD4-dependent Deregulation of Nascent Transcription 138
Figure A.7 Regulation of Pol II Pause Release by BRD4 139
Figure A.8 Elongation Defects Compared Across Treatments 140
Figure A.9 RTI Comparison Across Treatments 141
Figure A.10 Readthrough-induced Gene Activation 142
Figure A.11 Nascent RNA Sequencing using ONT 143
Figure A.12 Features of Readthrough Transcription 144
Figure A.13 Depletion of Elongation and 3′-RNA Processing Factors from the Chromatin 145
Figure A.14 Recruitment Defect of 3′-RNA Processing Factors to 5′ Gene Regions 146
Figure A.15 Recruitment Defect of CPSF73 147
Figure A.16 Recruitment Defects of FIP1 148
Figure A.17 Recruitment Defects of CstF64 149
Figure A.18 BRD4 Interactors 150
Figure A.19 Contribution of PAF1 to 3′-RNA Processing 151
Figure A.20 Putative BRD4 Enhancer Regions 152
Figure A.21 Elongation Control at Enhancer Regions 153
Figure A.22 Protein Landscape at Enhancer Regions 154
Figure A.23 Characterization of Changed Interactions 155
Figure A.24 Intronic TAPT1 Mutation Leads to Exon Skipping . . . 156
Figure A.25 Mutation Creates New Putative Branchpoint Position . 157

LIST OF TABLES

Table 2.1 3′-RNA Processing Complexes and Associated Factors 9
Table 3.1 Contingency Table for GO Enrichment Analysis 32
Table 11.1 Differential Analysis with Default Normalization 70
Table 14.1 3′-RNA Cleavage Efficiency Table 94
Table B.1 High Throughput Sequencing Data Overview 159
Table B.2 Annotation Databases Overview 164
Table B.3 Software and Algorithms Overview 164
Table B.4 RNA-seq Data Defining Actively Transcribed Genes . . 167
Table B.5 NET-seq Pipeline: Software and Parameters 168
Table B.6 *In silico* Masking of Chromatin-associated Mature RNA 168
Table B.7 Comparison of Pol II Profiling Methods 169
Table B.8 Cross-mapping of Human Reads to the Mouse Genome 169
Table B.9 Differential Binding Analysis: Functions and Parameters 170
Table B.10 NascONT-seq Pipeline: Software and Parameters 170
Table B.11 Read Classification of ONT Data 171
Table B.12 HiChIP Pipeline: Software and Parameters 172

ACRONYMS

4sU	4-thiouridine
A	adenine
ATD	average termination distance
BET	bromodomain and extraterminal domain
bp	base pair
C	cytosine
cDNA	complementary DNA
ChIP-seq	chromatin immunoprecipitation followed by sequencing
ChIP-Rx	ChIP with reference exogenous genome spike-in followed by sequencing
chromatin-MS	chromatin mass spectrometry
CPSF	cleavage and polyA specificity factor
CRISPR	clustered regularly interspaced short palindromic repeats
CstF	cleavage stimulation factor
DNA	deoxyribonucleic acid
DPO	differential Pol II occupancy
DSIF	5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole sensitivity-inducing factor
FDR	false discovery rate
FE	fold-enrichment over matched input control
fly	drosophila melanogaster
G	guanine
GO	gene ontology
GRO-seq	global run-on sequencing
H3K27ac	histone three lysine twenty-seven acetylation

H3K27me3	histone three lysine twenty-seven trimethylation
H3K36me3	histone three lysine thirty-six trimethylation
H3K4me1	histone three lysine four monomethylation
H3K4me3	histone three lysine four trimethylation
H3K79me2	histone three lysine seventy-nine dimethylation
HA-tag	human influenza hemagglutinin-tag
HiS-NET-seq	high sensitivity NET-seq
HTS	high-throughput sequencing
IP-MS	immunoprecipitation followed by mass spectrometry
kb	kilobase
loess	locally estimated scatterplot smoothing
lncRNA	long-noncoding RNA
MAD	median absolute deviation
mio	million
mNET-seq	mammalian NET-seq
mouse	mus musculus
MP	promoter-proximal maximum pausing position
nascONT-seq	long-read nascent RNA-sequencing
NELF	negative elongation factor
NET-seq	native elongating transcript sequencing
OI	osteogenesis imperfecta
ONT	oxford nanopore technologies
padj	FDR adjusted p-value
PI	pausing index
PAF	Pol II-associated factors
PCR	polymerase chain reaction

Pol II	RNA polymerase II
polyA	polyadenylation
PRO-seq	precision nuclear run-on sequencing
P-TEFb	positive transcription elongation factor b
RLE	relative logarithmic expression
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RPB2	Pol II subunit 2
RPK	reads per kilobase
RPKM	reads per kilobase million
RPM	reads per million
RTI	readthrough index
SI-NET-seq	spike-in NET-seq
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
SS	splice site
SSP	strand-switching primer
T	thymine
TES	termination end site
tsd	thousands
TPM	transcripts per kilobase million
TSS	transcription start site
TZ	termination zone length
U	uracil
UMI	unique molecular identifier
VNP	VN primer
XRN2	5′-3′ exoribonuclease 2

ZUSAMMENFASSUNG

In der Genexpression mehrzelliger Organismen hat sich die frühe Elongationsphase als ein Kontrollpunkt für die RNA-Polymerase II (Pol II) herausgestellt. Frühere Studien bringen den Kontrollpunkt, der den Übergang von pausierender zu produktiver Pol II reguliert, mit dem BRD4 Protein in Verbindung. Die genaue Rolle und der Mechanismus, durch den BRD4 diesen und andere regulatorische Prozesse nach der Initiation beeinflusst, bleiben jedoch unbekannt. In dieser Studie werden die unmittelbaren Proteinfunktionen von BRD4 durch schnellen Proteinabbau und den Einsatz verschiedener Omik-Ansätze ermittelt. Dies schließt die hochauflösende NET-seq-Methode ein, welche die Pol II-Verteilung misst.

Zunächst wird die NET-seq-Methode angepasst um quantitative Vergleiche zwischen einzelnen Proben zu ermöglichen. Mit der Hinzugabe von Referenzproben aus Mauszellen wird eine zuverlässige Normalisierung der Signalstärke sichergestellt. Ein zusätzlicher experimenteller Anreicherungsschritt entfernt darüber hinaus unerwünschte Chromatin-assoziierte RNA von der Probe was zu einer deutlichen Verbesserung der Signalabdeckung führt.

Nach dem schnellen Abbau von BRD4 identifizieren die verbesserten Methoden einen globalen Defekt bei der Freisetzung von pausierender Pol II und bestätigen damit die Rolle von BRD4 während der frühen Elongationsphase. Verschiedene Beobachtungen zeigen eine fehlgeschlagene Rekrutierung von Pol II-assoziierten Faktoren, die zu einer fehlerhaften Zusammensetzung des aktiven Elongationskomplexes führen. Interessanterweise treten Elongationsdefekte nicht nur an Genregionen, sondern auch an transkribierten Enhancer-Regionen auf.

Ein unerwartetes Ergebnis ist die beobachtete Transkriptionsaktivität, welche weit über das eigentliche Genende hinausreicht und auf einen Terminationsdefekt hinweist. Dieser Defekt lässt sich mit Hilfe eines neu eingeführten Index erfassen, der die relative Terminationsaktivität zwischen zwei Proben bestimmt. Die Studie entwickelt weitere Methoden zur Identifizierung naszierender RNA-Produkte und deren Effizienz bei der Spaltung der RNA am Genende. Die Ergebnisse weisen auf fehlerhafte Verarbeitungen an einigen Genen hin, welche mit dem Terminationsdefekt korrelierten. Eine mögliche mechanistische Erklärung ist eine BRD4-abhängige und am Genanfang stattfindende Rekrutierung von Faktoren, die für die Spaltung von RNA am Genende benötigt werden. Diese Beobachtung zeigt eine regulatorische Verbindung zwischen Prozessen am Anfang und am Ende von Genen und erfordert weiterer Validierung in zukünftigen Studien. Insgesamt deuten die Ergebnisse auf einen allgemeinen von BRD4 abhängigen Kontrollpunkt hin, der für die erfolgreiche Elongation und Termination der Pol II von Bedeutung ist.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Name:	Bressin
Vorname:	Annkatrin Sarah

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Berlin, 16.05.2022

Annkatrin Sarah Bressin